최대 엔트로피 모델을 이용한 막횡단 단백질 예측

윤성희[O1] 차정원[2] 박승수[1]
이화여자대학교[1] 창원대학교[2]
sungs@ewhain.net[O] jcha@changwon.ac.kr sspark@ewha.ac.kr

Maximum Entropy Approach to Transmembrane Protein Prediction

Sung Hee Yoon[O1] Jeong Won Cha[2] Seung Soo Park[1]
[1] *Dept. of Computer Engineering, Ewha Womans University*
[2] *Dept. of Computer Engineering, Changwon University*

요 약

막횡단 단백질(Transmembrane Protein)은 약물표적(drug target)으로 신약개발로 대표되는 바이오 산업에서 중요한 연구대상이 되고 있다. 막횡단 단백질의 구조는 실험적 기법 또는 컴퓨터 모델링 기술을 이용하여 연구되고 있으며 컴퓨터 모델링 방법 중에서는 Hidden Markov Mode(HMM)에 기반한 시스템들이 좋은 성능을 보이고 있다. 그런데 이러한 시스템들은 구조형성에 관여하는 단백질의 다양한 특성에 대한 지식은 많이 고려하고 있지 않다. 만약 이러한 특성들이 고려된다면 구조 예측에 효과적인 보다 지능적인 모델을 만드는데 도움을 줄 수 있을 것이다. 본 논문은 단백질의 특성과 관련한 다양한 정보들을 융합하는데 효율적인 최대엔트로피모델(Maximum Entropy Model)을 이용하여 막횡단 단백질의 서열(sequence)로부터 막횡단 지역을 예측하는 방법을 제시하고자 한다.

1. 서 론

막단백질은 세포 밖의 신호를 세포내로 전달, 대사를 조절하고 세포의 분화를 유도하는 것으로 의약과 농업분야에서 중요한 연구대상이다. 특히 의약품의 80% 이상이 막단백질을 통해 작용하기 때문에 약물 표적이 되고 있다.

막단백질의 구조연구는 X 선 결정학(X-ray Crystallography), 핵자기공명(Nuclear Magnetic Resonance)법 등의 실험적 기법을 이용한 구조규명과 컴퓨터 모델링 기술을 이용한 구조예측 두 가지로 나눌 수 있다. 그런데 실험적 방법은 시간과 비용 면에서 비효율적이고, 결정체 구조를 만들기 어렵다는 단점을 갖는다. 이러한 한계는 컴퓨터 모델링 기술을 이용하여 극복 할 수 있다.

컴퓨터 모델링 기술을 이용한 막횡단 지역 예측 시스템으로는 TMHMM[1], HMMTOP[2], TopPred[3]등이 있다. 이 시스템들은 확률모델인 HMM 에 기반한다. HMM 은 확률통계(statistics)분야에서의 Maximum Likelihood Estimation(MLE)을 응용한 것으로 각 상태에서의 발생확률과, 변이확률을 가지고 구조를 예측한다. 그리고 HMMTOP 과 TOPPRED 은 단백질의 특성 정보도 이용하는데 각각 아미노산의 분포, 소수성 정보와 positive inside rule 을 사용한다. 그런데 단백질은 앞에서 언급한 것 이외에도 구조와 관련하여 다양한 특성을 가지며 이러한 특성정보의 융합은 구조 예측에 효과적인 보다 지능적인 모델을 만드는데 도움을 줄 수 있을 것이다. 그리고 HMM 은 다양한 정보들을 고려하는데 한계가 있기 때문에 단백질의 다양한 특성 정보를 사용하기 위해서는 다른 기계학습(Machine Learning) 방법이 필요하다.

본 논문에서는 단백질의 특성정보들을 효율적으로 융합 할 수 있는 최대 엔트로피 모델을 사용한다. 최대 엔트로피 모델은 단백질의 다양한 특성정보를 자질(feature)로 사용하여 서열로부터 α-helix 막횡단 지역을 예측한다.

2. 관련연구

2.1 아미노산과 단백질

단백질은 아미노산의 중합체 이며, 박테리아로부터 사람에 이르기까지 모든 생물 종에 있는 단백질들은 20 가지의 같은 아미노산들로 구성된다. 각 아미노산은 크기, 모양, 전하, 화학 반응성 등이 다른 곁사슬(side chain)을 가지며 곁사슬에 따라 아미노산의 성질이 결정된다. 아미노산의 성질은 단백질의 구조형성에 관여하며, 단백질의 구조는 단백질이 생물학적으로 작용 할 수 있는 능력을 준다. 따라서 단백질이 매개하는 광범위한 기능들은 이 20 가지 아미노산의 다양성과 융통성에서 나오는 것이라 할 수 있다.

아미노산은 산성, 염기성, 친수성, 소수성, 방향성, 지방성등의 성질을 가지며 이것은 성질의 유무에 따라 이진으로 나타낼 수도 있고, 소수성과 같이 실험적으로 밝혀진 모델에 의해 정규화된 값으로 나타낼 수도 있다.

2.2 막단백질

막단백질은 내재성 단백질과 외재성 단백질로 나눌 수 있다. 내재성 단백질은 액정구조를 가진 인지질의 2 중층 속을 자유롭게 유동할 수 있고, 외재성 단백질은 막의 세포질쪽 표면에 존재하여 내재성 단백질의 분포를 조절한다.

내재성 막 단백질인 막횡단 단백질은 단백질의 대부분은 세포막에 묻혀 있고, 양 끝은 각각 세포내부와 외부로 돌출된 상태로 존재한다. 따라서 단백질 서열은 세포내부(inside), 막(Membrane), 세포외부(outside) 중 한가지 환경에 속하게 된다. 이 세 가지 환경을 위상(topology)이라 하며, 단백질은 막에서 2 차 구조 형태를 갖는다. 따라서 막횡단 단백질의 구조를 예측한다는 것은 2 차 구조를 띄는 막횡단 지역을 예측하는 것이다.

3. 최대 엔트로피 모델

최대 엔트로피 방법은 제약이 주어지지 않을 경우에는 가능한 가정을 자제한다는 원칙에 의해서 확률 값을 추정하는 방법이다. 이러한 가정은 학습 데이터에서 자질과 표현된 결과인 위상과의 관계로부터 추론된다. 이러한 성질을 만족하는 확률분포는 가장 높은 엔트로피를 갖게 되는데 이것은 유일하고, maximum-

likelihood 분포와도 유사하며, 지수승의 형태를 갖는다[4]. 최대 엔트로피 모델은 (식 1)과 같은 형태를 지닌다.

$$P(o \mid h) = \frac{1}{Z(h)} \prod_{i=1}^{k} \alpha_i^{f_i(h,o)} \qquad \text{(식 1)}$$

여기에서 α_i는 $\alpha_i = \exp(\lambda_i)$ 이고, λ_i는 실수 파라미터 이다. o는 표현된 결과를 의미하고 h는 문맥을 의미한다. $Z(h)$는 정규화 함수이고 각 자질 함수 $f(h,o)$는 (식 2)에서 보는 것과 같이 바이너리 함수로 정해놓은 분류조건을 만족하였는지 그렇지 않은지를 구분해 주는 함수이다. 예측을 할 때 만약 하나의 단어가 특정한 범주에 속한다면 o는 true 나 false 값을 가지게 된다.

$$f_i(h,o) = \begin{cases} 1 & if\ o = true\ and\ previousword = the \\ 0 & otherwise \end{cases} \qquad \text{(식 2)}$$

α_i는 Generative Iterative Scaling(GIS)[5]를 이용해서 구할 수 있는데 이것은 반복을 통해서 파라미터를 점진적으로 향상시키는 방법이다.

최대 엔트로피에서 가장 중요한 문제는 자질을 선택하는 것이다. 자질은 주어진 데이터의 특성을 잘 표현하는 것이 중요하기 때문에 데이터나 문제에 대한 충분한 지식이 없을 경우에는 f_i를 만들기가 어렵다. 이러한 문제를 해결하기 위해서 본 논문에서는 고려된 자질을 1~8 개로 조합하여 138 번의 실험을 하였다.

최대 엔트로피는 사용하는 각 자질들이 독립이어야 한다는 제약이 없다. 따라서 아미노산의 성질과 같이 서로 관계(relation)를 갖는 정보들을 융합하여 효율적인 모델을 만들 수 있다.

4. 최대 엔트로피 모델을 위한 자질

아미노산의 특성을 성격에 따라 크게 세가지로 나누었으며 각 특성 에서 고려된 자질들은 모델 학습 시 자질로 사용된다.

4.1 아미노산의 성질

2 장에서 아미노산의 화학 반응성이 단백질의 구조결정에 영향을 끼친다고 하였다. 따라서 아미노산의 성질 중 화학 반응성을 중심으로 자질을 선별하였다. [표 1]은 아미노산 별로 성질을 정리한 것이며, Standford 대학의 Folding@Home[6]에서 제공한 아미노산 속성표를 참고하였다.

표 1. 20개 아미노산의 성질

아미노산	아미노산의 성질
A	aliphatic, hydrophobic, neutral
R	polar, hydrophilic, charged(+), 염기성
N	polar, hydrophilic, neutral, 산성
D	polar, hydrophilic, charged(-), 산성
C	polar, hydrophobic, neutral, 황화
Q	polar, hydrophilic, neutral, 산성
E	polar, hydrophilic, charged(-), 산성
G	aliphatic, neutral
H	aromatic, polar, hydrophilic, charged(+), 염기성
I	aliphatic, hydrophobic, neutral
L	aliphatic, hydrophobic, neutral
K	polar, hydrophilic, charged(+), 염기성
M	hydrophobic, neutral, 황화
F	aromatic, hydrophobic, neutral
P	hydrophobic, neutral
S	polar, hydrophilic, neutral, 수산기가 있는 지방성
T	polar, hydrophilic, neutral, 수산기가 있는 지방성
W	aromatic, hydrophobic, neutral
Y	aromatic, polar, hydrophobic
V	aliphatic, hydrophobic, neutral

4.2 모델에 의한 소수성 값

아미노산은 Eisenberg, Kyte 등 실험적으로 밝혀진 모델에 의해 책정된 소수성 값을 갖는다. 본 논문에서는 실험에 적합한 Eisenberg scale 을 사용한다. Eisenberg scale 이란 Eisenberg model 에 의한 소수성 값이며, 그 값은 표 2 에서 보여준다.

표 2. Eisenberg scale

A	0.62	C	0.29	D	-0.90	E	-0.74	F	1.19
G	0.48	H	-0.40	I	1.38	K	-1.50	L	1.06
M	0.64	N	-0.78	P	0.12	Q	-0.85	R	-2.53
S	-0.18	T	-0.05	V	1.08	W	0.81	Y	0.26

4.3 아미노산의 발생 정확도

아미노산은 각각이 갖는 화학적 성질과 환경 사이의 관계, 인접 아미노산과의 관계 등에 의해 특정 위상에서 높은 출현빈도를 보이거나, 그 반대의 현상을 보인다. 이러한 아미노산의 분포성향은 서열을 분석을 통해 알 수 있으며, 이를 바탕으로 각 위상에서의 아미노산 발생 정확도를 산출하였다.

표 3 은 길이 별로 각 위상에서 높은 발생 정확도를 나타내는 아미노산 또는 아미노산 조각과 그 정확도 값을 2 개씩 보여준다. 정확도 값(A_{st})은 $A_{st} = N_{st}/N_{sw}$ 에 의해 계산되며 이때 N_{st} 는 아미노산 조각 s 가 위상 t 에서 나타나는 빈도, N_{sw} 는 s 가 전체 위상 w 에서 나타나는 빈도이다. 즉, 아미노산 발생 정확도란 아미노산 또는 아미노산 조각이 각 위상에서 나타나는 확률을 의미하는 것이다.

표 3. 아미노산 조각의 정확도

길이	세포내부		막		세포외부	
	Seg	A_{si}	Seg	A_{sM}	Seg	A_{so}
1	R	0.45	I	0.46	D	0.65
	K	0.43	F	0.42	N	0.63
2	RR	0.65	MI	0.69	CE	0.94
	KR	0.62	LI	0.67	EC	0.88
3	MKW	1.00	YWV	1.00	QNW	1.00
	NYW	1.00	YWM	1.00	HEP	1.00
4	QAKK	1.00	VIIG	1.00	MGSG	1.00
	RRRR	1.00	IIAV	1.00	DSLL	1.00
5	LRTPL	1.00	SWVSF	1.00	PKVSY	1.00
	SLRTP	1.00	LGITT	1.00	IWVPD	1.00

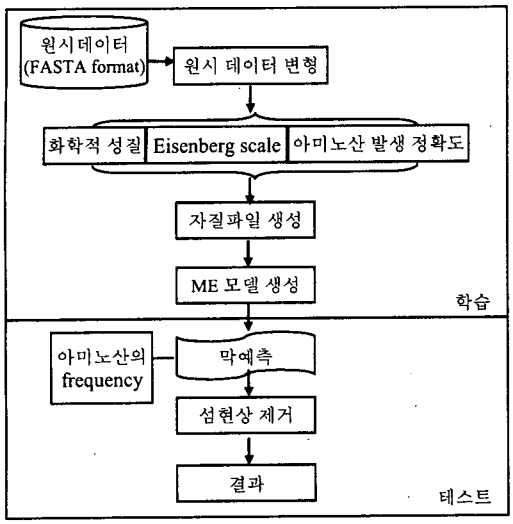
Seg: 아미노산 또는 아미노산 조각

A_{si}, A_{sM}, A_{so} : s가 세포내부, 막, 외부에서의 발생하는 정확도

정확도가 1 인 것은 아미노산 또는 아미노산 조각이 해당 위상에서만 나타났다는 것을 의미한다. 따라서 표 3 으로부터 MKW, NYW, QAKK 등의 아미노산 조각은 세포내부에만 출현했다는 것을 알 수 있다. 전체적으로 보면 조각의 길이가 길수록 A_{st} 값이 높은 반면, N_{sw} 값은 작아진다.

5. 학습 및 평가 과정

실험에서 사용한 원시 데이터는 FASTA format이며, 학습 시 (아미노산-위상)의 일대일 대응 쌍으로 변형한다. 최대 엔트로피를 이용한 학습 및 평가 과정은 그림 1과 같다.



[그림 1] 시스템 구조

실제서열과 예측된 서열간 비교에서 하나의 막횡단 지역에 대해서 아미노산 6개 이상이 매칭되었을 때 정확히 예측된 것으로 간주한다. 그런데 모델 예측 결과 서열 중간에 아미노산 6개 미만의 막 조각들이 섬처럼 떠있는 현상이 발견되었다. 이 현상은 예측 성능을 저하시킨다. 따라서 이러한 경우 앞뒤의 문맥을 고려하여 막 조각을 제거하는 후처리 작업을 실행한 후, 결과를 보인다.

6. 실험

학습 데이터는 TMHMM에서 사용한 것으로 실험적으로 밝혀진 위상이 레이블링 된 160개의 단백질 서열을 사용한다. 하나의 아미노산은 하나의 위상에만 속하지 않고, 아미노산은 앞뒤에 위치한 아미노산들과의 관계에 의해서 일부 특성은 영향력이 축소 또는 확대될 수 있다. 따라서 window size를 3으로 한다. 모델 학습 시 자질은 4장에서 언급한 특성들을 사용하며, 각 자질은 1~8개의 조합으로 만들어져 총 138번의 반복실험을 진행하였다. 조합은 개수에 따라 자질조합 1~8순으로 명명하며 평가는 10-fold cross validation을 사용한다. 결과는 다음과 같다.

표 4는 각 자질이 예측에 끼치는 영향력을 알아보기 위해 각 자질을 단독으로 사용한 결과이고, 표 5는 자질을 2개씩 조합한 결과이다.

표 4. 자질조합 1

자질조합 1	재현율	정확도	Location
Eisenberg scale	71.5%	92.3%	35.0%
그 외	46.5%	50.4%	3.1%

재현율: 정확히 예측된 막 / 실제 막
정확도: 정확히 예측된 막 / 예측 된 막
Location: 모든 막이 정확히 예측된 단백질 수/ 총 단백질 수

표 5. 자질조합 2

자질조합 2		재현율	정확도	Location
Eisenberg scale (E)	소수성	97.52%	81.49%	51.25%
	Polar	97.51%	81.47%	50.63%
	Ali	99.28%	79.18%	52.50%
	Aro	99.43%	78.55%	51.88%
	Ch	98.82%	77.68%	49.38%
	Acc	99.21%	78.92%	50.00%
	all	83.16%	88.48%	40.63%
그 외		46.5%	50.4%	3.1%

Ali: aliphatic, Aro: aromatic, Ch: charged, Acc: 아미노산의 발생 정확도
All: 황화, 산성, 염기성, 수산기가 있는 지방성을 하나로 묶은 것.

두 결과를 통해 Eisenberg scale이 예측에 중요한 역할을 한다는 것을 알 수 있다. 특히 표 5는 각 자질들이 Eisenberg scale과의 관계를 통해서만 재현율과 Location에서 성능 향상을 이룬다는 것을 보여준다. 이러한 결과는 자질조합 3~8에서도 볼 수 있다. 표 5로부터 각 자질이 Eisenberg scale과의 관계를 통해서만 성능향상을 할 수 있다는 것을 발견하였으므로 3개 조합부터는 기본적으로 Eisenberg scale을 포함하도록 한다.

표 6. 자질조합 3~8

자질조합	재현율	정확도	Location
E+aro+ch	99.65%	76.41%	51.25%
E+aro+ch+acc	99.65%	76.37%	49.38%
E+ali+aro	99.63%	77.18%	51.25%
E+ali+aro+acc	99.63%	76.93%	48.13%
E+aro+acc	99.62%	74.91%	48.75%

표 6은 자질조합 3~8까지의 실험 결과 중 재현율이 높은 5개를 보여준 것으로 자질의 융합을 통해 재현율이 향상됨을 보여준다. 실험 결과를 통해 적절한 정보의 융합이 예측에 긍정적인 영향을 끼친다는 것을 증명하였다.

7. 결론 및 향후 연구

본 논문은 최대 엔트로피 모델을 이용하여 막횡단 단백질의 α-helix 막횡단 지역을 예측하는 방법을 제시하였다. 모델 학습 시 자질은 서로 관계가 있거나 이질적인 단백질의 특성 정보를 사용하였고, 자질을 1~8개로 조합하여 138번 반복 실험을 하였다.

실험 결과를 통해 단백질 정보의 융합이 재현율에서 우수한 결과를 나타낸다는 것을 알 수 있었다. 특히 자질을 2개 이하로 조합한 경우 보다 3개 이상 조합한 경우 100%에 근접한 재현율을 보이는 것을 볼 수 있다. 이것은 정보의 융합이 예측에 긍정적인 영향을 끼친다는 것을 증명한 것이다. 하지만 재현율에 비해 정확도 성능이 약간 낮아지는 것은 본 연구가 향후 해결해야 할 과제이다. 이 문제는 재현율이 높게 나타나는 자질조합을 사용하여 후보막을 선정 한 후 그 중에서 실제 막을 찾아내는 2단계 예측 또는 alpha TM[7]과 같이 서로 다른 방법론의 통합을 이용해 해결 할 수 있을 것이다.

참고문헌

[1] Krogh, A., B., Larsson, et al, "Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes." Journal of Molecular Biology, 305(3), pp. 567-580, 2001
[2] Tusnady, G., E., I., Simon, "Principles Governing Amino Acid Composition of Integral Membrane Proteins: Application to Topology Prediction", Journal of Molecular Biology, 283(2), pp. 489-506, 1998
[3] von Heijne G., "Membrane Protein Structure Prediction, Hydrophobicity Analysis and the Positive-inside Rule", Journal of Molecular Biology, 225(2), pp. 487-494, 1992
[4] Della Pietra, S., V., Della Pietra and J., Lafferty, "Inducing Features of Random Fields.", IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(4), pp.380-393, 1997
[5] Darroch, J., N., D., Ratcliff, "Generalized Iterative Scaling for Log-Linear Models." The Annals of Mathematical Statistics, 43(5), pp.1470-1480, 1972
[6] http://folding.stanford.edu/education/AminAcid.html
[7] 송철환, et al, "SVM과 HMM을 이용한 α-Helix 막횡단 단백질 예측", 한국정보과학회, 가을학술발표논문집(2), pp. 817-819, 2003