위치 종속 유사도 스펙트럼을 이용한 단백질 서열의 아미노산 조성 추정

(Estimating Amino Acid Composition of Protein Sequences Using Position-Dependent Similarity Spectrum)

지 상 문 †

(Sang-Mun Chi)

요 약 단백질의 아미노산 조성은 생물정보학의 여러 문제를 해결하기 위한 기초적인 정보로 자주 활용된다. 본 논문에서는 아미노산간의 진화적인 연관성을 정의한 BLOSUM 행렬에서 유도한 유사도 함수를 사용하여 아미노산 조성을 결정한다. 이러한 방법은 생물학적인 연관성이 있는 단백질 서열일수록 비슷한 아미노산 조성을 갖도록 한다. 또한 단백질의 구조와 기능에 중요한 역할을 하는 위치-특이적인 아미노산의 분포를 추정하기 위해서 레이더나 음성 신호의 스펙트럼 분석에 사용되는 개념인 시간-종속 분석, 시간 해상도와 주파수 해상도의 개념을 적용하였다. 제안한 방법을 단백질의 세포내 위치예측에 적용하여 기존의 아미노산 조성 추정 방법을 사용하는 것보다 크게 향상된 성능을 보임을 확인하였다.

키워드 : 아미노산 조성, 유사도 함수, 스펙트럼 분석, 단백질의 세포내 위치 예측

Abstract The amino acid composition of a protein provides basic information for solving many problems in bioinformatics. We propose a new method that uses biologically relevant similarity between amino acids to determine the amino acid composition, where the BOLOSUM matrix is exploited to define a similarity measure between amino acids. Futhermore, to extract more information from a protein sequence than conventional methods for determining amino acid composition, we exploit the concepts of spectral analysis of signals such as radar and speech signals—the concepts of time-dependent analysis, time resolution, and frequency resolution. The proposed method was applied to predict subcellular localization of proteins, and showed significantly improved performance over previous methods for amino acid composition estimation.

Key words : Amino Acid Composition, Similarity Measure, Spectral Analysis, Protein Subcellular Localization Prediction

1. 서 론

단백질의 아미노산 조성은 그 단백질을 구성하는 각 아미노산의 종류별 빈도수로 정의된다. 아미노산 조성은 단백질의 기초적인 특징으로, 단백질의 기능과 구조를 알기 위한 기본적인 정보이다. 예를 들면, 세포질, 세포외 기질과 핵, 리소솜 등의 여러 세포소기관에서 같은 세포내 위치에 존재하는 단백질들은 유사한 아미노산의 조성과 기능을 가지며[1], 신호서열과 막통과 단백질에서 막을 관통하는 부분의 아미노산들과 수용성 단백질의 내부에는 소수성이 큰 아미노산 조성으로 구성된다[2-4]. 따라서, 아미노산 조성은 단백질의 기능과 구조를 예측하는 생물정보학의 여러 분야에 이용 되고 있고, 그 예로 단백질의 세포내 위치 예측에 사용된다[5-8].

본 논문에서는 아미노산의 개수에 기초하여 아미노산 조성을 결정하는 방법보다 단백질 서열로부터 보다 많

· 이 연구는 2008학년도 경성대학교의 지원에 의하여 연구되었음

† 정 회 원 : 경성대학교 컴퓨터과학과 교수
smchiks@ks.ac.kr

논문접수 : 2009년 9월 30일
심사완료 : 2009년 11월 3일

은 유용한 정보를 추출하기 위해서, 아미노산간의 진화적인 연관성에 기반한 유사도 함수를 사용한다. 유사도 함수는 단백질 서열간의 유사도를 측정하기 위해서 널리 사용되는 치환행렬인 BLOSUM 행렬을[9] 사용하여 구성한다. 동일한 아미노산이 아니라도 유사도 함수의 값을 아미노산 조성의 추정에 반영하여, 비록 아미노산의 각각의 개수가 다르더라도 진화적으로 유사한 단백질은 비슷한 아미노산 조성을 갖도록 한다. 또한, 단백질의 구조적, 기능적인 특성은 전체서열에서의 아미노산 조성뿐만 아니라, 부분적으로 나타나는 특징적인 아미노산 서열에 큰 영향을 받으므로, 부분적인 아미노산 조성을 효과적으로 추정할 수 있는 방법을 사용 하였다. 부분적인 아미노산을 추정하는 방법으로 단백질 서열의 앞부분인 아미노말단에서 국부적인 아미노산 조성을 추정하는 방법과[6], 아미노말단, 중간부분, 뒷부분인 카르복시말단에서 국부적인 아미노산 조성을 추정하는 방법이 있다[10]. 이 방법들은 고정된 길이의 부분서열을 사용하고, 신호서열과 모티프가 포함됨으로서 나타나는 아미노산 조성의 특성을 이용하는 반면에, 본 논문에서는 신호서열과 모티프를 고려하지 않고, 단백질 서열의 길이에 비례하여 부분서열의 길이를 결정하고, 각 부분서열에서 안정적인 아미노산 조성을 추정하기 위해서 유사도 함수의 파라미터를 조정한다. 이를 위하여 레이더나 음성신호 등의 처리에 사용되는 시간종속 분석, 시간해상도와 주파수 해상도의 개념을[11] 적용하여 아미노산 조성을 추정하였다.

2. 위치 종속 유사도 스펙트럼을 사용한 아미노산의 조성 추정

2.1 아미노산간의 유사도 함수 정의

단백질 서열의 아미노산 조성을 추정하기 위한 기존 방법을 다음과 같다. 아미노산 N개로 구성된 단백질 서열을 $x[n]\,(0 \le n < N)$이라면, 아미노산의 종류는 20가지이므로 아미노산 $a_k\,(1 \le k \le 20)$가 단백질 서열에서 나타나는 빈도는 다음 식으로 계산한다.

$$\sum_{n=0}^{N-1} \langle x[n], a_k \rangle, \tag{1}$$

여기서, $\langle , \rangle$는 $x[n]$과 a_k가 같은 경우에는 1의 값을 갖고, 그렇지 않을 경우에는 0을 갖는다. 본 논문에서는 이러한 동일한 아미노산의 개수에 기반한 기존의 아미노산 조성 추정법을 개수 스펙트럼이라고 부르기로 한다.

식 (1)의 $\langle a, b \rangle$는 a와 b가 동일해야만 0이 아닌 값을 갖는데, 이를 아미노산간에는 물리화학적 특성인 아미노산의 크기, 전하량과 소수성을 참조한 유사도로 대체할 수 있다. 이러한 유사도는 물리화학적 특성이 유사한 아미노산들은 단백질내에서 비슷한 구조와 기능을 한다는 사실을 이용할 수 있게 한다. 본 논문에서는 여러 종의 생물체에 존재하는 비슷한 구조와 기능을 나타내는 단백질을 사용하여 아미노산간의 진화적인 연관성을 정의한 치환행렬이 존재하여 단백질 서열간의 유사도를 계산하는데 널리 이용되므로[9], 아미노산의 조성 추정에 아미노산의 동일성 정보 대신에, 아미노산의 진화적인 유사도를 이용하여 아미노산의 특성에 적합한 정보를 추출한다.

생물학적인 특성과 연관된 유사도 함수를 정의하기 위해서, 첫 번째로 아미노산 a와 b간의 거리를 정의한다. 두 아미노산 a와 b의 거리는 치환행렬의 하나인 BLOSUM50을[9] 사용하여 정의하였다. 치환행렬 값인 $s(a,b)$는 a와 b의 진화적인 연관성이 클수록 큰 값을 주므로, 벡터 x, y가 유사할수록 큰 값을 주는 내적 $<x,y>$과 비슷하다. 내적과 거리의 관계인 $d(x,y)^2 = <x,x>+<y,y>-2<x,y>$에서 내적을 치환행렬 값으로 대치하여 거리를 정의하였다.

$$d(a,b) = \sqrt{s(a,a)+s(b,b)-2s(a,b)}\,. \tag{2}$$

BLOSUM 행렬은 대칭행렬이므로 $d(a,b) = d(b,a)$를 만족한다. 식 (2)의 거리를 이용하여 유사도 함수를 정의하는데, 거리와 유사도는 서로 반비례 하는 특성을 가져야 한다. 본 논문에서는 여러 형태의 식으로 유사도 함수를 시험하여 보고, 다음의 형태의 유사도 함수를 사용하였다. 정규확률분포 $(2\pi)^{-1/2}\exp(-d^2(x,\mu)/2\sigma^2)$에서 평균 μ로부터 거리 $d(x,\mu)$인 x의 확률분포 함수값과 비슷하게 거리가 $d(a,b)$인 아미노산간의 유사도는 다음과 같이 정의한다.

$$\exp(-\sigma\, d^2(a,b)). \tag{3}$$

파라미터 σ는 유사도 함수의 특성을 조절하는 중요한 역할을 한다. 식 (3)에서 두 아미노산 a와 b가 같으면 σ값에 무관하게 1의 값을 갖지만, 두 아미노산이 다를 경우에는 σ값이 작을수록 커다란 값을 가진다. 따라서, 작은 σ값은 유사한 아미노산을 아미노산 조성 추정에 더 크게 반영 한다. 본 논문에서는 식 (3)의 유사도 함수를 개수 스펙트럼에서 사용되는 식 (1)의 $\langle , \rangle$를 대체한 유사도 스펙트럼을 제안한다.

$$S[k] = \sum_{n=0}^{N-1} \exp(-\sigma\, d^2(x[n], a_k)) \tag{4}$$

개수 스펙트럼은 단백질 서열내의 각 아미노산의 개수를 나타내는 반면에, 유사도 스펙트럼은 서열내에서 각 아미노산과 진화적으로 유사한 아미노산의 양을 나타내므로 진화적으로 관련있는 단백질 서열은 비슷한 유사도 스펙트럼을 갖게 된다. 아미노산 서열의 스펙트럼을 이용하기 위해 아미노산 서열을 소수성열로 변환하여

퓨리에 분석으로 서열의 스펙트럼을 구하는 방법이 있다[12]. 단백질의 전체적인 소수성 특징은 모티프나 세포내 단백질 분류 신호 등의 아미노산 종류에 따른 부분서열을 나타내기 어려운 반면, 유사도 스펙트럼은 아미노산의 종류별 분포를 나타내므로 부분서열의 특성을 포함할 수 있다.

2.2 위치 종속 분석

신호처리 분야에서는 신호특징(진폭, 주파수, 위상)이 시간에 따라 변하는 경우에 효과적으로 정보를 추출하기 위해서 시간-종속 스펙트럼 분석이 사용된다[11]. 이 방법에서는 전체 신호서열을 부분 서열로 나누고, 그 부분 서열에서는 비교적 신호의 특징이 변하지 않고 유지된다고 가정 하고, 부분서열마다 신호의 특징을 추출한다. 아미노산의 조성 추정에서는 신호특징이 변하는 것에 대응되는 현상이 단백질 서열의 위치에 따라서 아미노산 조성이 변하는 것이다. 아미노산의 조성은 단백질 서열에서의 위치에 따라 변하는데, 예를 들어, 신호펩티드와 막통과 단백질에서 막을 통과하는 부분은 소수성이 높은 아미노산들로 구성되고[2-4], 단백질은 각각의 세포내 위치마다의 고유의 PH나 이온의 세기 등의 생리화학적 환경의 영향으로 각기 다른 아미노산 조성을 갖고, 특히, 단백질내부에 매몰되는 아미노산서열보다는 표면부위에 존재하는 아미노산들이 더욱 세포내 위치에 종속적인 고유한 조성을 갖는다[1].

시간-종속 분석의 개념과 유사하게 단백질 서열에서의 위치에 따른 아미노산 조성을 추정하기 위해서 위치 종속 분석법을 제안한다. 먼저, 길이가 N인 단백질 서열로부터 길이 $L=N/S$인 부분서열을 만든다. 여기서 S는 부분서열의 길이를 결정하는 파라미터로 S가 클수록 부분 서열의 길이는 짧아진다. 그림 1과 같이 r-번째 부분 서열은 다음의 L개의 아미노산으로 구성된다.

$$x[rR+m], 0 \le m \le L-1, \quad (5)$$

여기서, 본 논문에서는 $R=L/2$이다. 파라미터 S값에 따라서 부분 서열의 길이는 $L=N/S$로 결정되고, 식 (5)에 의해서 $r=0,1,...,2S-2$인 값을 가질 수 있으므로 부분서열의 개수는 $2S-1$이 되고, 부분서열은 r에 따라서 전체 서열내의 위치가 결정된다.

시간-종속 분석에서는 부분 서열의 길이가 짧을수록 시간에 따른 신호의 특징 변화를 잘 표현할 수 있어서 높은 시간 해상도(time resolution)를 얻을 수 있다. 그러나, 부분 서열의 길이가 짧으면 세밀한 주파수 분석이 불가능 하여 낮은 주파수 해상도(frequency resolution)를 갖는 시간 해상도와 주파수 해상도의 상충관계가 존재한다. 이러한 상충관계는 아미노산의 조성을 추정하기 위한 위치 종속 분석법에도 나타난다. 파라미터 S가 클수록 부분 서열의 길이는 짧아지므로 아미노산 조성의 변화를 잘 나타낼 수는 있으나, 각 부분 서열에 속하는 아미노산의 수가 작아서 안정적인 조성추정이 어렵다. 하지만, 파라미터 σ를 조절하여 이러한 상충관계를 완화할 수 있음을 다음 장에서 보인다.

제안한 위치 종속 유사도 스펙트럼으로 각 부분서열마다의 유사도 스펙트럼 $S[k](1 \le k \le 20)$를 얻고, 이를 정규화한 $\tilde{S}[k]=Sk]/\sum_{n=1}^{20} S[k]$를 실험에 사용하였다. 또한 단백질 서열의 길이를 단백질의 특성에 포함하기 위해서, 본 논문에서는 단백질 서열의 길이 N을 수정한 $0.1\times N^{0.3}$을 사용하였다. 이러한 수정의 이유는 길이 N의 크기가 유사도 스펙트럼보다 훨씬 커서, 실험결과에 큰 영향을 미치기 때문에 이를 완화하기 위함이다. 이는 인간의 청각에 의한 소리크기의 인지는 음향 에너지의 0.33 제곱승임을 참조하여 만든 휴리스틱한 방법이다. 아미노산의 조성을 추정하는 것과 유사한 방법으로 두 개의 연속된 아미노산인 아미노산 짝(amino acid pair)의 조성을 추정한다. 즉, 20×20개의 아미노산 짝인 a_k, a_m의 유사도 스펙트럼은 다음식으로 얻는다.

$$\sum_{n=0}^{N-1} \exp(-\sigma d^2(x[n], a_k) - \sigma d^2(x[n+1], a_m))$$

3. 실험 및 분석

3.1 단백질의 세포내 위치 예측

제안한 방법을 단백질의 세포내 위치 예측에 적용하여 기존의 아미노산 조성 추정과 비교한다. 실험에 사용한 PLOC 자료는 평가를 위해서 자주 사용하는 자료로서, Swiss-Prot(release 39)에서 추출한 7579개의 단백질 서열로 구성되어 있고, 이 단백질은 각각 12개의 세포내 위치(chloroplast, cytoplasmic, cytoskeleton, endoplasmic reticulum, extracellular, golgi apparatus, lysosomal, mitochondrial nuclear, peroxisomal, plasma membrane, vacuolar) 중의 하나에 존재하고 아미노산 서열의 유사도가 80%이하이다[5]. PLOC 자료는 단백질의 개수가 거의 균등하게 5개의 부분 자료로 나누어져 있어서 교차 타당성 평가(cross validation test)가 용이하다. 즉, 4개의 부분 자료를 사용하여 단백질의 세포내 위치를 예측하기 위한 패턴분류 방법을 구성하여 나머지 1개의 부분 자료에 대한 평가를 수행한다. 이렇게 학습 자료와 평가 자료를 구성하는 각기 다른 5가지 경우에 대해 실험하여 평균적인 성능을 얻는다.

단백질의 세포내 위치를 예측하기 위한 패턴분류기로 SVM(support vector machine)[13]을 구현한 LIBSVM [14]을 사용하였다. SVM에 사용되는 커널로서는 가우시안 커널,

$$K(x,y)=\exp(-\gamma \| x-y \|^2) \quad (6)$$

을 사용하였다. LIBSVM에 사용되는 파라미터 C는 10으로 고정하였고, 식 (6)의 γ와 식 (3)의 σ는 교차 타당성 평가에서 학습 자료로 이용되는 4개의 부분 자료를 사용하여 최적 파라미터를 구하고, 이를 나머지 1개의 부분자료의 평가에 사용하였다. 구체적으로 $\sigma=1,2,3,4$와 $\gamma=10\times 2^k$, $k=0,...,12$로 이루어진 4×13 가지의 각각의 파라미터 조합에 대해서 학습 자료인 4개의 부분자료 중에서 3개의 부분자료를 사용하여 SVM을 구성하고, 나머지 1개를 평가하였다. 이렇게 부분자료를 3개와 1개로 나누는 것은 4가지 방법이 있으므로 이들 4가지 조합에 대해서 성능을 평균하여 가장 높은 성능을 보이는 파라미터 조합을 찾았다. 찾아진 파라미터를 σ_1, γ_1이라고 하면 σ_1과 $0.75\times\gamma_1\times 1.1^k$, $k=0,1,...,8$에 대해서 위와 같은 방법으로 재차 최적인 파라미터를 탐색하는 과정을 반복하였다.

세포내 위치를 예측하는 방법의 성능비교는 TA(Total Accuracy)$=\sum_{i=1}^{K} T_i/N$와 LA(Local Accuracy) $=\sum_{i=1}^{K} P_i/K$를 사용하였다. 여기서, $K=12$는 세포내 위치의 개수, $N=7579$은 총 단백질 개수, T_i는 세포내 위치 i에 존재하는 단백질 중에서 올바르게 예측된 개수, $P_i=T_i/N_i$로서 세포내 위치 i에 존재하는 단백질의 개수 N_i에서 올바르게 예측된 비율이다. LA는 12개의 세포내 위치에서의 예측정확도를 평균한 값으로, 단백질이 많이 속하는 세포내 위치와 적게 속하는 세포내 위치에서의 예측정확도사이의 균형을 맞추어서 성능을 나타낼 수 있게 한다. 본 논문에서는 최적 파라미터 γ와 σ의 탐색하는 과정에는 LA를 사용하였다.

표 1에서 AA(Amino Aacid)와 AAP(Amino Acid Pair)는 각각 아미노산 조성과 아미노산 짝의 조성을 나타낸다. S가 커질수록 부분서열의 길이는 짧아지고, 부분서열은 개수는 $2S-1$이 된다. 표 1에서 보듯이 S값이 5나 6이 되어 부분 서열의 아미노산 조성을 사용할 때 정확도가 높았다. 대부분의 단백질의 세포내 위치 예측방법에서는 단백질의 각기 다른 정보에 기반한 예측방법을 구성하고 이를 결합하여 성능을 향상시킨다. 표 1에서 $S=4,5,6,7$인 AA와 $S=1,4,5,6,7$인 AAP에 기반한 SVM의 9개의 예측결과를 다수결 투표(majority voting) 방식을 사용하여 예측 정확도를 구한 결과 TA는 84.5%, LA는 70.8%로 향상되었다.

표 2에서 PDSS(Position-Dependant Similarity-Spectrum)는 제안한 방법을 나타내고, 이를 기존의 아미노산 추정법인 개수 스펙트럼을 사용하는 PLOC과[5] 비교하였다. 두 방법 모두 PLOC 자료를 사용하는데, 제안

표 1 위치 종속 유사도 스펙트럼을 사용한 단백질의 세포내 위치 예측 정확도(%)

S	AA(TA:LA)	AAP(TA:LA)
1	76.6 : 60.9	79.5 : 65.6
2	81.0 : 68.3	82.1 : 68.6
3	82.0 : 68.5	82.8 : 67.4
4	82.7 : 69.3	83.4 : 69.1
5	**83.0 : 69.5**	83.1 : 69.4
6	82.4 : 68.5	**83.4 : 70.6**
7	82.8 : 69.4	83.3 : 68.8

표 2 아미노산 조성의 추정 방법간의 비교(%)

	AA(TA:LA)	AAP(TA:LA)	Voting(TA:LA
PDSS	83.0 : 69.5	83.4 : 70.6	84.5 : 70.8
PLOC	72.4 : 56.7	75.9 : 56.8	78.2 : 57.9

한 PDSS는 커다란 성능의 향상을 보였다. AAP 외에도 g-갭 AAP 조성도 사용하였는데, 이는 두 개의 아미노산 짝으로서 두 아미노산사이에 존재하는 g개의 아미노산의 종류는 고려하지 않는 것으로, g-갭 AAP ($g=1,2,3$)는 PLOC과 PDSS에서 모두 각 방법의 AAP보다 성능이 약간 떨어지는 결과를 보였다.

개수 스펙트럼을 사용하여 k개의 아미노산으로 구성된 k-튜플($1\le k\le 7$)의 조성을 사용하는 경우[7]는 PLOC이 $k=1,2$만을 사용하는 방법을 확장한 방법으로 4-튜플이 81.2%의 TA와 64.1%의 LA로 가장 성능이 높았다. AA나 AAP에 비해 성능의 향상을 얻을 수는 있었으나, k-튜플의 종류는 20^k이므로 $k\ge 4$인 경우는 k-튜플의 종류가 크게 증가하여 SVM의 학습과 평가시간이 매우 크게 증가한다. 반면에 PDSS는 $1+(2S-1)\times 20^k$ (AA는 $k=1$, AAP는 $k=2$)의 보다 적은 계산량으로 더 높은 예측 정확도를 얻었다.

아미노산 조성의 사용과 더불어 단백질이 세포내 위치로 이동하는데 필요한 특정한 아미노산 서열들의 정보(N-terminal targeting sequences, internal signal anchors, sorting-sequence motifs)를 사용하는 Multi Loc을[6] PLOC 자료에 대해서 실험한 결과[8], 식물자료는 73.6% TA와 71.3% LA, 동물자료는 76% TA와 73.6% LA, 곰팡이자료는 75.8% TA와 72.5% LA를 얻었다. PDSS에서 다수결 투표를 사용하여 얻은 성능과 비교하면 TA의 경우에는 PDSS가 10% 정도 높고, LA의 경우에는 비슷하므로, 전체적으로 PDSS가 우수하다고 할 수 있다.

3.2 위치 종속 유사도 스펙트럼의 특성

부분서열의 길이를 결정하는 S와 식 (3)의 유사도 함수의 특성을 결정하는 σ에 의해 PDSS는 특징 지워진

다. 파라미터 S값이 결정되면, 이 값에 적합한 최적의 σ값이 학습과정에서 선택된다. 5개의 부분 자료로 나누어진 PLOC 자료에서 4개의 자료를 학습 자료로 사용하여 자동적으로 σ를 결정하고, 나머지 1개의 부분 자료를 평가하는데, 이렇게 학습 자료와 평가 자료를 나누는 5개의 경우에서 결정된 σ의 평균을 표 3에 나타내었다.

표 3 학습과정에서 자동으로 선택된 σ값

S	AA	AAP
1	3.8	4.0
2	2.4	3.0
3	2.0	3.2
4	1.6	2.6
5	1.8	2.0
6	1.4	2.0
7	1.4	2.0

표 3에서 보듯이 파라미터 S가 커짐에 따라 선택되어지는 σ가 점차 작아진다. 이러한 현상은 시간-종속 스펙트럼 분석에서 분석구간이 짧을수록 시간 해상도는 높아지나, 주파수 해상도가 낮아지는 시간 해상도와 주파수 해상도의 상충관계의 개념을 참조하여 이해할 수 있다. PDSS는 S가 커질수록 부분서열의 길이 $L=N/S$가 짧아지고 부분 서열의 총 개수 $2S-1$가 커지므로, 단백질 서열의 국부적인 특징과 아미노산의 조성의 변화를 효과적으로 표현할 수 있는 장점이 있으나, 각 부분서열에 속하는 아미노산의 개수가 작아서 안정적인 조성의 추정이 어렵다. 따라서, 부분서열의 길이가 짧은 경우에도 아미노산 조성을 효과적으로 추정할 수 있도록 σ가 선택되어야 한다. 그런데, 작은 값의 σ를 갖는 유사도 함수일수록 아미노산의 조성을 추정할 때, 유사한 아미노산으로부터의 기여도를 크게 반영한다. 이러한 S와 σ의 관계에 의해서, S가 커질수록 선택되는 σ는 작아진다. 대조적으로 부분서열의 길이가 충분할 경우에는 커다란 σ를 사용하여 유사한 아미노산의 영향을 감소시켜서 아미노산간의 판별력을 높일 때가 최적이다.

표 4는 학습과정에서 σ를 탐색하지 않고, σ값에 따른 성능을 알아 본 결과이다. 여러 가지 S와 σ하에서 AA을 사용한 단백질의 세포내 위치 예측 정확도 중에서 LA를 나타내었다. 표 3과 동일한 경향인 S가 커질수록 최적인 σ는 작아짐을 볼 수 있다. 따라서, PDSS에서는 고정된 σ를 사용하지 않고, 주어진 S에 종속적으로 최적의 σ가 학습과정에서 자동으로 선택되어져서 시간해상도와 주파수해상도의 상충관계를 완화시킴을 확인할 수 있다.

표 4 여러 S와 σ하에서 AA을 사용한 LA(%)

S \ σ	4	3	2	1
1	**62.1**	61.3	59.4	54.2
2	68.4	**68.7**	68.5	65.6
3	67.4	67.9	**68.2**	68.0
4	68.0	69.2	**70.0**	69.4
5	67.4	68.4	**69.4**	69.4
6	67.5	68.7	69.2	**69.4**
7	66.8	68.0	69.1	**70.2**

4. 결론 및 향후연구

본 논문에서는 단백질 서열에서 보다 많은 정보를 포함한 아미노산 조성을 추출하기 위해서 유사도 함수를 사용한 위치 종속 유사도 스펙트럼을 제안하였다. 제안한 방법의 효율성을 알아보기 위해서 단백질의 세포내 위치 예측에 적용하여서 성능 향상을 확인하였다. 제안한 방법은 위치 종속적인 방법을 사용하여 단백질 서열에서 위치 특이적으로 존재하고 단백질의 기능과 구조에 큰 영향을 주는 부분 서열의 정보를 효과적으로 나타낼 수 있다. 또한 유사도 함수의 파라미터를 조절하여 단백질 서열내에서 부분적인 아미노산의 정보를 추출하는 경우에도 성능이 저하되지 않도록 하였다.

향후에는 제안한 방법을 아미노산 조성을 이용하는 여러 생물정보학의 문제에 적용하려 한다. 예를 들면, 단백질의 세포내 위치를 예측하는 방법에서 기존의 아미노산 조성의 추정을 제안한 방법으로 대체하여 예측성능을 개선하려 한다. 많은 방법들은 예측성능을 높이기 위해서 예측하려는 단백질 서열만으로 얻을 수 있는 정보 외에도, 외부적인 정보인 유전자 정보 데이터베이스에서 모티프, 유전자 온톨로지, 단백질의 기능 영역에 대한 정보, 그리고 여러 논문에 나타나는 단백질 정보를 이용하고 있다. 이러한 방법은 계산량이 증가하고, 이미 알려진 정보를 활용하는 방법이므로 새로운 단백질 서열이나 알려진 정보가 작은 경우에는 효과적이지 않지만, 단백질에 대한 알려진 정보가 증가함에 따라 성능이 향상되고 있다. 제안한 아미노산 조성 추정방법을 이러한 방법들에 부가적으로 사용하여 성능을 향상시키는 연구를 수행할 예정이다.

참 고 문 헌

[1] M. A. Andrade, S. I. O'Donoghue, and B. Rost, "Adaption of protein surfaces to subcellular location," *J. Mol. Biol.*, **276**, pp.517-525, 1998.

[2] M. Paetzel, A. Karla, N. C. Strynadka, and R. E. Dalbey, "Signal peptidases," *Chem. Rev.*, 102, pp. 4549-4580, 2002.

[3] V. Goder, and M. Spiess, "Molecular mechanism of signal sequence orientation in the endoplasmic reticulum," *The EMBO Journal.*, 22, pp.3645-3653, 2003.
[4] E. Granseth, G. von Heijne, and A. Elofsson, "A study of the membrane-water interface region of membrane proteins," *J. Mol. Biol.*, 346, pp.377-385, 2005.
[5] K.-J. Park, and M. Kanehisa, "Prediction of protein subcellular location by support vector machines using compositions of amino acids and amino acid pairs," *Bioinformatics*, **19**, pp.1656-1663, 2003.
[6] A. Höglund, P. Dönnes, T, Blum, H.-W. Adolph, and O. Kohlbacher, "Multiloc: prediction of protein localization using n-terminal targeting sequences, sequence motifs and amino acid compositions," *Bioinformatics*, **22**, pp.1158-1165, 2006.
[7] W.-W. Yang, B.-L. Lu, and Y. Yang, "A comparative study on feature extraction from protein sequences for subcellular localization prediction," *IEEE Symposium on CIBCB*, pp.201-208, Toronto, Canada, 2006.
[8] H. Shatkay, A. Höglund, S. Brady, T. Blum, P. Dönnes, and O. Kohlbacher, "Sherloc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data," *Bioinformatics*, **23**, pp.1410-1417. 2007.
[9] S. Henikoff, and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *proc. natl. acad. sci.*, **89**, pp.11915-11919, 1992.
[10] S. Matsuda, J.-P. Vert, H. Saigo, N. Ueda, H. Toh, and T. Akutsu, "A novel representation of protein sequences for prediction of subcellular location using support vector machines," *Protein Sci.*, 14(11), pp.2804-2813, 2005.
[11] A. V. Oppenheim, and R. W. Schafer, *Discrete-time signal processing*. Prentice-Hall, New Jersey, 1989.
[12] K. Gupta, D. Thomas, S. Vidya, K. Venkatesh, and S. Ramakumar, "Detailed protein sequence alignment based on Spectral Similarity Score (SSS)," *BMC Bioinformatics*, **6**(105), 2005.
[13] V. Vapnik, *Statistical learning theory*, John Wiley & Sons, 1998.
[14] C.-C. Chang, and C.-J. Lin, LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

지 상 문
1991년 서울대학교 수학교육과(학사). 1993년 한국과학기술원 수학과(석사). 1998년 한국과학기술원 전산학과(박사). 1993년~2000년 삼성전자 정보통신. 2001년~현재 경성대학교 컴퓨터과학과 부교수. 관심분야는 기계학습, 생물정보학, 분자모델링