소수성과 치환행렬에 기반한 신호서열 예측
(Signal Sequence Prediction Based on Hydrophobicity and Substitution Matrix)

지 상 문 †
(Sang-Mun Chi)

요 약 본 논문에서는 미지의 아미노산 서열이 신호 펩티다제 I에 의해 절단되는 분비성 단백질인지를 판별하고, 분비성 단백질일 경우에는 절단 위치를 예측하는 방법을 제안한다. 아미노산의 소수성을 이용한 전처리를 수행하여 분비성 단백질의 선도서열인 신호서열의 존재와 절단 위치를 추정한다. 전처리를 통해서 신호서열 아닌 서열을 초기에 제외함으로써 신호서열 예측의 정확도를 높인다. 지지벡터기계를 신호서열의 예측에 효과적으로 적용하기 위해서, 생물학적 정보와 관련된 아미노산 서열간의 거리를 제안한다. 아미노산의 세포내 위치를 예측할 수 있는 소수성 척도와 아미노산의 진화적인 관계를 나타낼 수 있는 치환행렬을 이용하여 아미노산 서열간의 거리를 정의한다. Swiss-Prot release 50 단백질 자료에 대하여 교차타당성 기법을 사용하여 실험한 결과 제안한 방법은 신호서열중에 98.9%를 신호서열로 판별하였고, 88%의 절단위치 예측정확도를 보였다. 기존의 방법과의 비교실험을 통해서 제안한 방법이 신호서열의 예측에 더욱 효과적임을 확인하였다.

키워드 : 신호서열 판별, 절단위치 예측, 소수성, 치환행렬, 아미노산 서열거리

Abstract This paper proposes a method that discriminates signal peptide and predicts the cleavage site of the secretory proteins cleaved by the signal peptidase I. The preprocessing stage uses hydrophobicity scales of amino acids in order to predict the presence of signal sequence and the cleavage site. The preprocessing enhances the performance of the prediction method by eliminating the non-secretory proteins in the early stage of prediction. For the effective use of support vector machine for the signal sequence prediction, the biologically relevant distance between the amino acid sequences is defined by using the hydrophobicity and substitution matrix; the hydrophobicity can be used to predict the location of amino acid in a cell and the substitution matrix represents the evolutionary relationships of amino acids. The proposed method showed 98.9% discrimination rates from signal sequences and 88% correct rate of the cleavage site prediction on Swiss-Prot release 50 protein database using the 5-fold-cross-validation. In the comparison tests, the proposed method has performed significantly better than other prediction methods.

Key words : Signal Sequence Discrimination, Cleavage Site Prediction, Hydrophobicity, Substitution Matrix, Distance between Amino Acid Sequences

1. 서 론

노벨 생리의학상(1999년)을 받은 Blobel박사에 의해 발견된 신호펩티드(signal peptide) 또는 신호서열(signal sequence)은 분비성 단백질, 수용성 단백질, 막단백질과 같은 소포체에서 합성되는 단백질의 아미노 말단의 선도 펩티드이다. 신호펩티드의 역할은 신호인지입자와 결합하여 소포체막으로 이동하여 합성된 펩티드사슬이 소포체 내강으로 이동하게 하는 것이다. 신호 펩티다제(signal peptidase)에 의해 신호펩티드가 절단되면 단백질이 완성되어 소포체 내강으로 들어가게 된다.

폭발적으로 증가하는 유전체 및 단백질자료에서 자동으로 분비단백질을 찾기 위한 방법으로, 아미노산 서열정보만으로 신호펩티드의 존재유무와 신호 펩티다제에 의해 절단되는 위치를 예측하는 방법이 연구되고 있다. 이러한 연구는 초기의 가중치 행렬방법에서 시작하여 현재에는 신경망(NN: neural network), 은닉마르코프모델(HMM: hidden markov model), 지지벡터기계

· 이 논문은 2007학년도 경성대학교 학술연구비지원에 의하여 연구되었음
† 정 회 원 : 경성대학교 컴퓨터과학과 교수
smchiks@ks.ac.kr
논문접수 : 2007년 1월 22일
심사완료 : 2007년 5월 31일

(SVM: support vector machine)를 사용하여 연구되고 있다[1-7].

1980년대부터 신호서열에 대한 연구를 선도해온 von Heijne그룹의 SignalP는 NN과 HMM을 사용하는 방법으로 널리 사용되고 있으며 가장 성능이 높다[5,8]. 이 방법은 신호서열 예측을 위한 정보로 아미노산의 종류와 위치를 사용한다. SVM을 사용하는 방법은 아미노산의 종류를 이용하는 방법[6]과, 아미노산 서열간의 공통된 아미노산의 개수로서 두 서열의 거리를 정의하는 방법[7]이 있다. 본 연구에서는 아미노산의 종류대신에 화학적 특성인 소수성과 생물학적인 특성인 아미노산간의 진화적 관계를 사용하여 아미노산의 거리를 정의한다. 신호서열간의 거리는 절단위치를 기준으로 대응되는 아미노산간의 소수성과 치환행렬의 값의 차이로 정의하였다. 정의된 거리는 수학적으로 거리의 정의를 만족하며, 아미노산 서열상의 위치특이적인 특성을 거리의 계산에 반영할 수 있다. 또한, 아미노산의 소수성에 기반한 전처리를 사용하여 미지의 아미노산 서열에 신호서열이 존재하는지의 여부와 존재한다면 절단위치가 속하는 구간을 추정한다. 전처리를 통해서 신호펩티드가 존재할 가능성이 적은 아미노산 서열을 제거하고, 절단위치를 탐색할 범위를 축소함으로써, 신호서열 예측 성능을 높이고, 학습과 평가에 소요되는 시간을 단축한다.

2장에서는 전처리 방법을 설명하고, 3장에서는 제안한 아미노산 서열간의 거리와 이를 SVM에 적용하는 방법을 설명한다. 4장에서는 제안한 방법을 사용하여 신호서열 예측실험을 하고, 5장에서 결론을 맺는다.

2. 소수성을 이용한 전처리 방법

신호서열간에는 아미노산 서열상의 유사성이 크지는 않으나, 3개의 보존된 영역인 n, h, c영역이 존재 한다[그림 1]. 아미노말단에 양전하를 띤 아미노산들로 구성된 n-영역은 그 길이가 1-12로 매우 가변적이다. 7-15개의 소수성 잔기로 이루어진 h-영역은 신호서열이 보이는 가장 강한 특징 중의 하나이다. 절단부위와 h-영역사이에 3-8개의 극성 또는 전하가 없는 아미노산이 존재하는 c-영역이 나타난다. 특히, 명백한 모티프로서 절단위치이전의 -3과 -1위치에는 작고 중성의 잔기를 가진 아미노산이 나타나고 그림 1에는 대표적인 알라닌을 나타내었다[9,10].

본 연구의 전처리 과정에서는 신호서열의 확연한 구조적 특징인 -1위치에 나타날 수 있는 아미노산의 종류와 h-영역의 소수성을 이용하여 신호서열의 유무와 절단위치가 속하는 구간을 추정한다.

2.1 절단위치의 아마노산의 종류에 의한 전처리

절단위치에 존재하는 특정 아미노산들은 효과적인 절단을 막으며[5,9], 절단위치 이전의 -1위치에 빈번히 나타나는 아미노산은 아미노산 종류중의 일부분으로 한정되어 있다. 절단 위치인 -1위치에 나타나는 아미노산의 종류를 SignalP 3.0[5]에서는 진핵세균의 경우에는 A, C, G, L, P, Q, S, T만이 나타나고, 그람-양성 및 그람-음성 세균의 경우에는 A, G, S, T만이 나타난다고 가정하였고, 이러한 아미노산이외의 아미노산이 나타난 자료는 실험적인 오류인 경우가 많음을 보였다[5]. 본 논문에서도 같은 방법으로 위의 아미노산이 절단위치에 나타난 자료만을 신호서열 예측 알고리즘의 학습과 평가에서 사용한다.

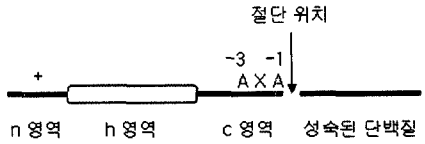

그림 1 신호서열의 구조

2.2 소수성을 이용한 절단위치의 구간 추정

소수성(hydrophobicity)은 개별 아미노산의 친수성/소수성 정도를 수치화한 것으로 미지의 단백질의 형태와 생체내 위치를 예측할 수 있게 해준다. 신호펩티드는 단백질이 합성되는 동안에는 소포체막에 결합되어 있으므로 소수성이 높은 아미노산으로 구성되어 있고, 신호서열의 구조에서 h-영역의 강한 소수성이 이를 반영한다.

그림 2의 실선은 Swiss-Prot 단백질자료[11]중의 하나로서 ID가 5NTD_HUMAN인 신호서열을 포함한 단백질자료의 소수성도를 보여준다. 소수성도로 Engelman등이 제안한 방법을 사용하였다[12]. 그림 2에서 1번부터 26번까지의 아미노산이 신호서열에 해당한다. 소수성이 큰 아미노산이 10번부터 21번까지 연속으로 나타나는 신호펩티드의 전형적인 특징을 보인다.

본 논문에서는 신호서열마다 길이가 다르고 소수성의 변화가 크므로, 소수성도를 직접 사용하지 않고 신호서열의 예측에 유용하도록 변환하는 방법을 사용한다. 소

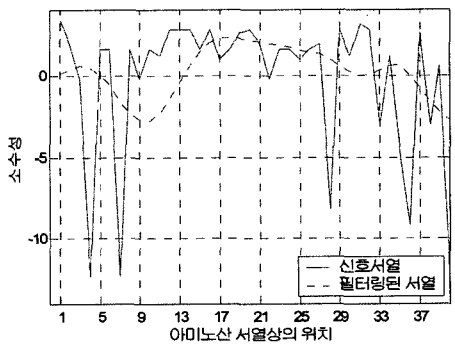

그림 2 신호서열의 소수성도

수성 정보를 이용하는 연구로서, 막통과 단백질에서 막에 내재되는 부분서열을 예측하는 연구가 있다[13]. 이 연구에서는 각 아미노산을 중심으로 주변 19개의 아미노산의 평균 소수성을 사용하였다. 이러한 이동 평균을 사용하는 방법은 저대역통과 필터링의 일종으로 높은 주파수대역을 감쇠시켜 소수성도의 급격한 변동을 평활화한다. 그러나, 이동 평균을 사용하는 저대역통과 필터링은 주파수 응답특성면에서 보면 고주파 대역의 감쇠가 미흡하다. 본 연구에서는 소수성을 신호서열 예측에 적용하기 위하여, 고주파 대역의 감쇠특성이 우수하고 감쇠 대역의 선택이 용이한 필터를 사용한다. 신호서열 예측에 효과적인 통과대역을 찾기 위해서 여러 통과대역을 시험하였고, 본 논문에서는 식 (1)의 시스템 함수와 그림 3의 주파수 응답을 가지는 IIR(Infinite-duration impulse response) 필터를 사용하였다[14]. 식 (1)의 필터는 그림 3의 주파수응답에서 보듯이 정규화된 주파수(normalized frequency) 0.1이하를 통과시키는 저대역통과필터로 고주파 대역을 효과적으로 감쇠한다.

$$H(z)=\frac{0.0337+0.0224z^{-1}+0.0221z^{-2}+0.0338z^{-3}}{1-1.7369z^{-1}+1.2005z^{-2}-0.2896z^{-3}} \quad (1)$$

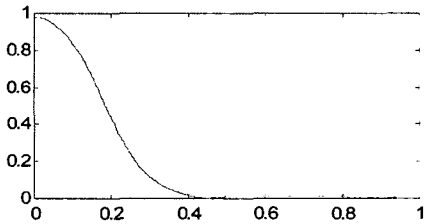

그림 3 저대역통과 필터의 주파수 응답

소수성을 이용하여 신호서열을 예측하기 위한 첫 단계에서는 저대역통과 필터링으로 소수성을 평활화한다. 두 번째 단계에서는 그림 2의 점선으로 표시된 것과 같은 필터링된 소수성을 이용하여 소수성이 문턱치 0.7 보다 큰 아미노산이 연속적으로 3개 이상 나타날 때 신호서열을 포함한 서열로 판정하고, 소수성이 0.7보다 큰 연속적인 구간을 h-영역으로 가정한다. 세 번째 단계에서는 신호서열의 절단위치는 h-영역 이후와 소포체막 외부에 놓이는 소수성이 작은 아미노산서열사이에 존재하므로, 절단위치는 폐구간 [h-영역의 중앙위치, h-영역 이후에 소수성이 최소인 위치+D]내에 위치하는 것으로 예측한다. 여기서, D(진핵생물은 6, 원핵생물은 9)는 폐구간이 절단위치를 포함하는 확률을 높이기 위해서, 2종 오류(false positive)가 많더라도 1종 오류(false negative)가 작아지도록 선택하였다.

절단위치 부근의 아미노산의 위치특이적인 분포를 이용한 2.1절의 전처리와 h-영역의 소수성을 이용한 2.2절의 전처리를 통하여 신호서열이 존재하지 않는 자료는 SVM의 학습 및 신호서열의 예측에서 제외한다. 학습과 예측에 신호서열과 관련성 적은 자료를 전처리에서 제외하므로 보다 효과적인 학습과 예측이 가능하다. 또한, 전처리를 통해 예측된 절단위치가 속하는 구간내의 아미노산을 중심으로 주위의 일정한 길이의 아미노산 서열만을 학습과 평가에 이용하므로, 여러 파라미터의 조건하에서 실험이 가능하도록 계산량을 감소시킬 수 있다.

3. 아미노산 서열간의 거리

신호서열을 예측하는 방법들이 사용하는 정보는 아미노산의 종류와 서열상의 위치이다. 가중 행렬방법에서는 절단위치에서 상대적인 각 위치의 아미노산 분포를 구하고, 미지의 아미노산 서열이 이 분포와 유사하면 신호서열로 예측한다[1-4]. 신경망과 HMM을 사용하는 SignalP[5]는 아미노산을 20차원의 벡터로 나타내고, 아미노산의 종류에 따라 20차원 중 한 요소만 1이고 나머지는 0으로 하는 희소코딩(sparse coding)을 사용한다. SVM 기반의 방법에서는 희소코딩을 이용하는 방법[6]과 두 서열에 공통으로 존재하는 아미노산의 부분서열의 개수로 거리를 정의하는 방법[7]이 있다.

위에서 알아본 기존의 방법들에서 두 아미노산 간의 거리는 두 아미노산이 같은지 다른지에 따라서만 결정된다. 따라서 상이한 아미노산간의 거리는 아미노산의 종류와 상관없이 같은 값을 갖는다. 본 연구에서는 아미노산마다 고유의 생물화학적인 특성을 반영하여 아미노산의 종류에 종속적인 거리를 정의한다. 아미노산의 소수성과 치환행렬을 이용하여 아미노산간의 유사성을 정의하는데, 소수성은 신호서열의 뚜렷한 구조적 특징인 소수성 영역의 특징을 나타내기에 유용하고, 치환행렬은 아미노산간의 진화적인 유사성을 나타낸다.

3.1 아미노산 서열간의 거리

아미노산 서열간의 거리를 정의하기에 앞서 아미노산간의 거리를 정의한다. 아미노산 a와 b사이의 소수성의 차이를 이용한 거리

$$d_h(a,b)=\sqrt{|h(a)-h(b)|} \quad (2)$$

와 치환행렬을 이용한 거리를 제안한다.

$$d_s(a,b)=\sqrt{s(a,a)+s(b,b)-2s(a,b)} \quad (3)$$

여기서, 소수성 $h(a)$는 논문[13]에 정의된 아미노산의 a의 소수성을 나타내고, $s(a,b)$는 BLOSUM50 행렬[15]에서 정의된 아미노산 a와 b의 유사도를 나타낸다. 신호서열은 서열상의 보존도가 작으므로 상동성이 작은 단백질간의 진화적 거리를 계산하기에 유용한 BLOSUM50을 사용한다.

소수성을 이용한 거리 d_h는 정의로부터 거리(metric)가 되기 위한 수학적 조건들을 만족함을 쉽게 확인할 수 있다. 하지만, 치환행렬을 이용한 거리 d_s의 경우에는 모든 아미노산간의 거리는 양수이고, 삼각부등식을 만족하는 지가 정의로부터 명백하지 않다. 이를 증명하기 위해서, 가능한 모든 아미노산의 조합 a, b에 대하여 식 (3)의 거리가 양수이고, 가능한 모든 아미노산의 조합 a, b, c에 대해서 $d_s(a,b) \le d_s(a,c)+d_s(c,b)$을 만족함을 조사를 통해서 확인하였다.

길이가 다른 서열간의 거리는 일반적으로 동적프로그램을 사용하여 비선형 정합을 수행한 후에 최종거리를 구한다. HMM을 사용하는 경우에도 비선형적으로 아미노산 서열을 각 상태에 정합한 후에, 각 상태에 속하는 아미노산의 확률을 구한다. SVM기반의 방법[7]에서는 서열상의 공통된 부분서열의 개수를 이용하므로 비선형 정합의 과정과 유사하다. 그러나, HMM으로 구현된 신호서열예측[5]에서도 절단위치 이전의 위치특이적이 아미노산을 고려하기 위해서는 절단위치 이전의 몇개의 위치에서는 비선형정합을 하지 않고, 신경망을 사용하는 경우에는 비선형정합 없이 고정적인 길이를 가지는 아미노산 서열을 이용하였다. 본 논문에서도 일정한 길이를 가지는 아미노산 서열을 사용하고 비선형정합을 하지 않는다. 예비실험을 수행한 결과 비선형정합을 수행한 후에 대응되는 아미노산간의 거리를 구하는 것보다 절단위치를 기준으로 상대적으로 같은 위치에 존재하는 아미노산간의 거리를 사용하는 것이 성능이 높음을 확인하였다. 이러한 방법은 계산량의 감소뿐 아니라 아미노산의 위치특이적인 성질을 나타낼 수 있는 장점이 있다. 최종적으로 아미노산 서열 $x=(x_1,x_2,\ldots,x_n)$, $y=(y_1,y_2,\ldots,y_n)$사이의 거리의 제곱을 대응되는 아미노산 거리의 제곱의 합으로 정의하였다.

$$d_h^2(x,y)=\sum_{k=1}^{n} d_h^2(x_k,y_k)=\sum_{k=1}^{n}|h(x_k)-h(y_k)| \quad (4)$$

$$d_s^2(x,y)=\sum_{k=1}^{n} d_s^2(x_k,y_k) \quad (5)$$
$$=\sum_{k=1}^{n}[s(x_k,x_k)+s(y_k,y_k)-2s(x_k,y_k)]$$

3.2 SVM에서 아미노산간의 거리 이용

신호서열의 판별과 절단위치의 예측을 위하여 SVM을 사용하였다. SVM은 분류오류를 최소화하기 위해 두 부류의 최대여백 초평면(maximal margin hyperplane)을 구하는 기계학습 방법이다[16-18]. 본 연구에서 사용하는 학습자료를 (x_i,y_i), $x_i \in R^n$, $y_i=1$ 또는 -1로 표시하면 x_i는 아미노산 서열이고, y_i는 신호서열 판별에서는 신호서열은 1, 신호서열이 아닌 서열은 -1이고, 절단위치 예측에서는 절단위치가 일정한 위치(서열의 끝에서 세번째)에 존재하는 서열을 1, 그렇지 않은 서열을 -1로 하였다. SVM의 학습은 두 부류의 최대여백 초평면

$$f(x)=w^T x+b \quad (6)$$

에서 파라미터 w,b를 구하는 것이고, x는 아미노산 서열이다. 최대여백을 갖기 위해서는 식 (7)의 최적화 문제를 풀어야 한다.

$$\text{최소화} \quad \frac{1}{2}w^T w+c\sum_{i=1}^{l}\xi_i \quad (7)$$

$$\text{제약조건} \quad y_i(w^T\phi(x_i)+b)\ge 1-\xi_i,\ \xi_i \ge 0$$

여기서, l은 학습자료의 개수이고, $\phi()$는 비선형적인 분류를 위해서 자료를 변환하는 함수이고, ξ_i는 잡음이나 이상치로 인해서 두 개의 부류로 정확히 나누어 지지 않는 자료에 대한 분류문제를 처리하기 위해 도입한 변수이고, c 는 페널티 파라미터이다.

최적화 문제를 풀기위해 식 (7)의 쌍대 라그랑지안 함수(dual Lagrangian function)를 최적화하는 방법이 쓰인다. 또한, 비선형 분류를 위한 $\phi()$함수의 도입은 실제 계산상에서는 $\phi(x_i)^T\phi(x_j)$의 형태로 나타나고, 이것은 변환된 공간에서 내적의 조건을 만족하는 수정된 내적(inner product)인 커널 $K(x_i,x_j)$의 계산으로 대체된다. 가우시안 커널 $K(x,y)=\exp(-\gamma\, d^2(x,y))$을 본 연구에서는 사용하였고, 최종적으로 사용한 커널은 식 (4)와 (5)의 두 개의 거리를 결합한 커널을 사용하였다.

$$K(x,y)=K_s(x,y)K_h(x,y) \quad (8)$$
$$=\exp(-\gamma_s d_s^2(x,y)\cdot\exp(-\gamma_h d_h^2(x,y))$$
$$=\exp(-\gamma_s d_s^2(x,y)-\gamma_h d_h^2(x,y))$$

여기서, γ_h,γ_s는 양의 값을 가지는 파라미터로서 커널의 형태를 결정한다. 커널과 커널의 곱은 역시 커널이므로 식 (8)은 커널의 조건을 만족한다.

4. 신호서열 예측 실험

4.1 단백질 자료 및 평가기준

신호서열 예측 방법의 성능을 알아보기 위해서는 생화학적 방법에 의해 구조가 밝혀진 신호서열을 포함하는 단백질자료가 필요하다. Swiss-Prot release 50 단백질 자료[11]를 사용하여 논문[8]의 방법으로 핵에서 인코드되는 진핵세균의 분비 단백질 1687개와 신호 펩티다제 1에 의해 절단되는 원핵세균의 분비단백질 483개를 얻었다. 이 자료에 2.1절의 절단위치의 아미노산의 종류를 이용한 전처리를 하여 1626개의 진핵생물 단백질과 463개(그람-음성 334개, 그람-양성 129개)의 원핵생물 단백질을 얻었고, 이를 실험자료로 사용하였다. 신호서열을 포함하고 있지 않은 단백질 자료는 1142개의

진핵생물 단백질과 426개(그람-음성 297개, 그람-양성 129개)의 원핵생물 단백질로 구성된 논문[8]의 자료를 사용하였다.

신호서열 판별 방법들의 성능을 평가하기 위한 기준[5]을 정의하기 위해서, 신호서열을 신호서열로 예측한 것은 tp(true positive), 신호서열이 아닌 것을 신호서열이 아닌 것으로 예측한 것은 tn(true negative), 신호서열이 아닌 것을 신호서열로 예측한 것은 fp(false positive), 신호서열을 신호서열이 아닌 것으로 예측한 것을 fn(false negative)이라 하자. 정확도(ACC: accuracy)는 신호서열과 신호서열이 아닌 단백질이 각각 신호서열과 신호서열이 아닌 것으로 정확히 예측된 비율이다.

$$ACC = \frac{tp+tn}{tp+tn+fp+fn} \tag{9}$$

민감도(SEN: sensitivity)는 신호서열이 신호서열로 올바르게 판별된 비율이다.

$$SEN = \frac{tp}{tp+fn} \tag{10}$$

특이도(SP: specificity)는 신호서열로 예측된 단백질 중에서 실제로 신호서열인 비율이다.

$$SP = \frac{tp}{tp+fp} \tag{11}$$

가양성도(FP: false positive rate)는 신호서열이 아닌 단백질이 신호서열로 예측된 비율이다.

$$FP = \frac{fp}{fp+tn} \tag{12}$$

4.2 신호서열 판별 실험

미지의 단백질 자료가 신호서열을 포함하는지를 판별하는 실험을 하였다. 신호서열을 포함하고 있는 자료와 신호서열을 포함하고 있지 않은 자료를 각각 5개의 부분자료로 균등하게 나누고, 4개의 부분자료로 학습하고 나머지 한 개의 자료에 대해서 예측하는 교차 타당성(5-fold-cross-validation)을 수행하였다. 자료를 나누는 방법에 따라 실험결과가 다르므로 무작위로 자료를 나누어 실험하는 과정을 30번 반복하여 평균을 구하였다.

학습단계에 사용되는 신호서열을 포함한 자료는 절단위치를 기준으로 절단위치 이전의 n_b개와 이후의 n_a개의 아미노산을 학습에 사용한다. 학습에 사용되는 신호서열이 아닌 아미노산 서열은 전처리과정에서 신호서열로 예측된 서열로서 신호서열과 유사한 자료이다. 전처리에서 추정된 구간내의 모든 위치를 절단위치로 가정하여 절단위치 이전의 n_b개와 이후의 n_a개의 아미노산으로 이루어진 아미노산 서열을 사용하였다.

본 논문에서는 SVM을 학습하기 위해서 mySVM[19]을 사용하였다. 학습을 위해서는 아미노산 서열의 길이를 결정하는 n_b, n_a, 식 (7)의 c, 식 (8)의 γ_h, γ_s값이 주어져야 한다. 절단위치 이후의 아미노산들은 아미노산 서열상의 특성이 적으므로 $n_a = 2$로 고정값을 사용 하였다. 다른 파라미터는 Swiss-Prot release 40 단백질 자료와 신호서열이 아닌 단백질 자료를 사용하여 교차 타당성(cross validation)을 통하여 선정하였다. 예비 실험을 통해서 학습이 제대로 되지 않거나(under training), 과도하게 학습되는(over training) 범위를 제외하고 $n_b = 15, 17, 19$, $c = 1, 2, 3$, $\gamma_h = 0, 0, 0.02, 0.04, 0.06, 0.08, 0.1$, $\gamma_s = 0.02, 0.04, 0.06, 0.08, 0.1$의 모든 조합에 대해 조사하였다. 실험마다의 편차를 고려하여 최적의 n_b 값은 평균값을 이용하여 선택하였다.

$$n_b = \underset{n_b=15,17,19}{\arg max} \sum_{c} \sum_{\gamma_h} \sum_{\gamma_s} ACC(n_b, c, \gamma_h, \gamma_s) \tag{13}$$

여기서, $ACC(n_b, c, \gamma_h, \gamma_s)$은 파라미터가 $n_b, c, \gamma_h, \gamma_s$일 때의 식 (9)의 값이다. 최적의 n_b를 먼저 구한 후에 식 (14)에 따라 c를 구하였고, 같은 방법으로 표 1의 γ_h, γ_s를 구하였다 .

$$c = \underset{c=1,2,3}{\arg max} \sum_{\gamma_h} \sum_{\gamma_s} ACC(n_b, c, \gamma_h, \gamma_s) \tag{14}$$

신호서열 판별의 첫 단계는 2.2장의 소수성을 이용한 전처리이다. 표 2의 결과에서 보듯이 그람양성균만이 1개의 신호서열이 신호서열이 아닌 것으로 판별되는 오류가 발생하였고, 진핵생물과 그람-음성균의 경우에는 모든 신호서열이 신호서열로 판별되었다. 판별오류가 발생한 서열은 Swiss-Prot자료의 ID가 MTCY_LEUME로서 소수성이 매우 낮았다. 전처리단계에서 신호서열의 판별오류는 0.05%로 작게 유지하면서, 신호서열이 아닌 서열을 41%정도 신호서열이 아닌 것으로 판별하였다. 신호서열로 예측된 서열들은 두 번째 단계의 처리를 통해서 최종적으로 신호서열인지가 판별된다.

신호서열 판별의 두 번째 단계에서는 첫 단계의 전처

표 1 신호서열 판별에 사용된 파라미터

	진핵생물	그람-음성균	그람-양성균
n_b	17	17	17
c	3	2	2
γ_s	0.006	0.006	0.004
γ_h	0.008	0.006	0.004

표 2 전처리 결과 (아미노산 서열의 수)

	진핵생물	그람-음성균	그람-양성균
tp	1626	334	128
fn	0	0	1
tn	477	108	54
fp	665	189	75

표 3 전처리를 통한 추정된 구간의 평균 길이

	진핵생물	그람-음성균	그람-양성균
신호서열	28	29	28
신호서열이 아닌 서열	13	15	14

리과정에서 신호서열로 판별된 서열을 대상으로 식 (6)의 값이 특정 문턱치보다 큰지를 비교한다. 전처리를 하지 않을 경우 모든 단백질 자료를 대상으로 모든 위치를 절단위치로 가정하고 SVM을 학습하고 판별실험을 하여야 하므로, 계산량 증가와 성능저하가 발생한다. 전처리를 통하여 추정된 비교적 짧은 구간만을(표 3) 절단위치로 가정하여 학습과 판별실험에 사용되는 서열의 개수를 감소시켰다. 이 방법의 단점은 추정된 구간이 실제 절단위치를 포함하지 않을 수 있다는 것이다. 실험결과를 조사하여 보니, 추정된 구간은 절단위치의 99.7%를 포함하여 전처리에 의한 판별성능의 저하가 작았다.

교차 타당성(5-fold-cross-validation)기법을 사용하여 신호서열 판별 실험을 수행하였다. 표 4와 5에 나타났듯이 제안한 방법은 효과적으로 신호서열을 판별함을 알 수 있다. 본 논문에서 사용한 신호서열이 아닌 단백질 자료는 신호서열과 유사한 막횡단 단백질이 포함되어 있어서 판별이 어려워 FP가 크지만[8], 제안한 방법은 진핵생물과 원핵생물 모두 판별성능이 높았다.

문턱치에 따라 신호서열의 판별이 영향을 받는다. -1에 가까운 문턱치를 사용할수록 보다 많은 서열이 신호서열로 판정되어, 신호서열이 아닌 단백질이 신호서열로 판별되는 빈도도 많아진다. 그림 4는 문턱치를 -0.6부터 0.1까지 0.05씩 증가시키면서 판별 실험한 결과를 보여준다. 작은 가양성도(false positive)에서도 높은 정확도로 판별이 이루어짐을 볼 수 있다.

파라미터의 값의 변화가 판별성능에 미치는 영향을 알아보았다. 먼저, 각기 다른 아미노산 서열의 길이에서

표 4 신호서열 판별 성능 (%), 문턱치 0.0

	진핵생물	그람-음성균	그람-양성균
ACC	97.2	97.7	96.8
SEN	97.4	97.8	96.6
SP	97.8	98.0	96.5
FP	3.1	2.4	3.0

표 5 신호서열 판별 성능 (%), 문턱치 -0.2

	진핵생물	그람-음성균	그람-양성균
ACC	97.0	96.5	96.1
SEN	99.1	98.6	98.8
SP	95.9	95.2	93.4
FP	6.0	6.1	6.0

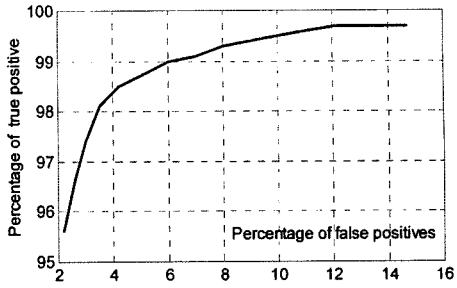

그림 4 문턱치에 따른 판별성능

표 6 아미노산의 길이에 따른 ACC (%)

	진핵생물	그람-음성균	그람-양성균
$n_b=15$	96.3	97.1	94.9
$n_b=17$	96.6	97.0	95.1
$n_b=19$	96.5	96.7	94.6

표 7 파라미터 c에 따른 ACC (%)

	진핵생물	그람-음성균	그람-양성균
$c=1$	96.4	96.7	94.4
$c=2$	96.7	97.1	95.5
$c=3$	96.8	97.2	94.4

의 c, γ_s, γ_h의 모든 조합에 대한 평균을 조사하였는데, 서열의 길이에 따른 성능의 차이는 크지 않았다[표 6].

아미노산 서열의 길이 $n_b=17$로 고정하고 각각의 c의 값에서 γ_s, γ_h의 모든 조합에 대해서 실험하여 평균을 조사하였다[표 7]. 길이와 마찬가지로 c값에 따라 성능이 크게 변하지는 않았다.

그림 5에 $n_b=17, c=3$일 때 γ_s, γ_h의 값에 따른 진핵생물의 판별 성능을 나타내었다. γ_s, γ_h값은 학습이 제대로 되지 않거나(under training), 과도하게 학습되는(over training) 범위를 제외한 0과 0.01사이에 대해서 실험하였다. 다른 파라미터보다는 값에 따라 성능의 차이가 커서, $\gamma_s=0.002, \gamma_h=0.0$일 때 95.8%이고, γ_s, γ_h가 0.004에서 0.008사이의 값들은 약 97% 정도로 비교적 균일한 성능을 보였다.

4.3 신호서열의 절단위치 예측

신호서열을 포함하고 있는 자료를 5개의 부분자료로 균등하게 나누고, 4개의 부분자료로 학습하고 나머지 한 개의 자료에 대해서 예측하는 교차 타당성(5-fold-cross-validation) 기법으로 30번을 반복 실험하여 평균을 구하였다. 학습자료로 절단위치를 기준으로 절단위치 이전의 n_b개와 이후의 n_a개의 아미노산을 사용하였고, 전처리로 추정된 구간에서 실제 절단위치를 제외한 모든 위

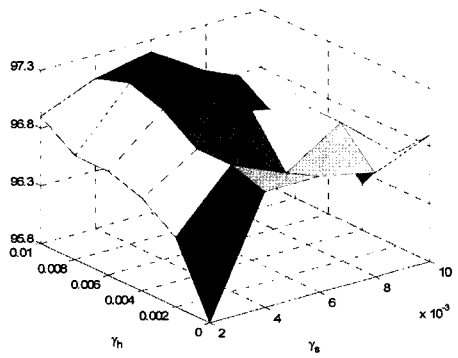

그림 5 γ_s, γ_h의 값에 따른 판별 성능 (%)

표 8 절단위치 예측에 사용된 파라미터

	진핵생물	그람-음성균	그람-양성균
n_b	17	15	19
c	2	1	1
γ_s	0.01	0.008	0.008
γ_h	0.006	0.01	0.01

표 9 절단위치 예측 정확도 (%)

진핵생물	그람-음성균	그람-양성균
88.0	93.1	87.0

치를 기준으로 이전의 n_b개와 이후의 n_a개의 아미노산을 대립 부류를 학습하기 위해 사용하였다. 절단위치는 전처리를 통해 추정된 구간내의 모든 위치에 대해서 식 (6)을 계산하여 가장 큰 위치로 예측하였다.

신호서열 판별과 마찬가지로 Swiss-Prot release 40 단백질자료를 사용하여 절단위치 예측을 위한 파라미터를 구하였다[표 8].

표 9는 Swiss-Prot release 50에 대하여 교차타당성을 사용한 절단위치 예측실험의 결과이다. 진핵생물과 원핵생물 자료수를 고려하여 평균하면 88%의 예측정확도이다. 대부분의 신호서열 예측 방법에서처럼 신호서열의 길이가 긴 그람-양성균의 성능이 진핵생물이나 그람-음성균보다 낮았다.

4.4 기존 신호서열예측 방법과 비교

제안한 방법을 여러 논문에 발표된 결과들과 비교하였고, SignalP 웹서버를 이용하여 비교실험 하였다. 논문의 결과들은 서로 다른 실험자료를 사용하므로 정확한 비교대상이 아니지만, 대략적 경향을 알아 볼 수 있다. 가중행렬 방법[1-3]이나 PSORT[4], SubLoc[6]의 방법은 신호서열의 예측성능에서 SignalP[5]나 SVM[7]을 사용하는 방법보다 저조하다[5,7,8]. SVM을 사용하는 방법[7]은 신호서열의 판별만을 수행하는데 6%의 가양성도(false positive)일 때 80%이하의 정확도, 8%의 가양성도(false positive)일 때 80%정도의 정확도를 보인다. 표 5에서 보듯이 가양성도가 6%일 때, SEN은 98.9%이므로 제안한 방법의 성능이 높음을 알 수 있다. SignalP V2-NN은 신호서열자료에서 절단위치를 예측하는 정확도가 학습자료와 평가자료가 약간의 중복이 있는 경우는 84.6%이었고, 학습자료와 평가자료가 겹치지 않는 경우에는 79.8%이었고, 가양성도는 18.3%이었다[8]. 본 논문의 방법에서 절단위치를 예측하는 실험인 표 9를 보면 평균적으로 88%의 정확도를 나타내었다. 교차타당성을 사용하였으므로 학습자료와 평가자료가 겹치지 않으므로 본 논문의 방법이 SignalP V2-NN보다 성능보다 높다.

웹(http://www.cbs.dtu.dk/services/SignalP/)을 통하여 비교평가를 수행하였다. SignalP 3.0[5]은 Swiss-Prot release 40.0이하의 자료로 학습되었다. 본 논문의 방법을 Spase-HS(Signal PeptidASE based on Hydrophobici ty and Substitution matrix)라고 명명하고 Swiss-Prot release 40.0으로 학습하였다. 학습자료의 개수는 진핵생물은 1215개, 그람-음성균은 325개, 그람-양성균은 129개이다. 학습자료와 독립적인 평가자료를 얻기 위해서 논문[8]의 방법으로 Swiss-Prot release 50.0에서 신호서열 자료를 추출하고, Swiss-Prot release 40.0 자료와 중복된 자료는 제거하였다. 또한, 2.1절의 절단위치의 아미노산의 종류에 의한 전처리를 수행하여 최종적인 평가자료를 마련하였다. 평가자료의 개수는 진핵생물은 949개, 그람-음성균은 113개, 그람-양성균은 41개이다.

SignalP는 신호서열 여부와 절단위치를 예측한 결과를 함께 출력한다. Spase-HS도 신호서열 판별분석을 수행한 후에 신호서열로 예측된 서열에 대하여 절단위치를 예측하였고, 학습을 위하여 신호서열이 아닌 자료에서 무작위로 80%를 사용하였고, 나머지 20%를 평가에 사용하였다. 학습과 평가를 30회 반복하여 평균을 나타내었다. SignalP3-NN이 SignalP3-HMM보다 모든 기준에서 성능이 높으므로, SignalP3-NN와 비교하였고, 신호판별시에는 SignalP에서 추천하는 D-score를 사용하였다. Spase-HS는 문턱치로 -0.2를 사용하였다.

표 10결과에서 보듯이 본 논문의 방법이 진핵생물의 SEN을 제외하고는 SignalP3-NN보다 신호서열 판별성능이 높았다. 그러나, FP의 경우에는 제안한 방법이 원핵생물자료에서 성능이 좋지 않았는데, 원핵생물의 실험자료가 부족하므로 학습이 충분하지 않았다고 여겨진다. 표 11의 결과는 신호서열로 정확하게 판별되고, 절단위치도 정확하게 예측된 비율이다. 제안한 방법은 진핵생물과 원핵생물 자료 모두에서 월등한 성능향상을 보였다.

표 10 신호서열 판별 성능 (%)
Spase-HS (SignalP3-NN)

	진핵생물	그람-음성균	그람-양성균
ACC	98.0 (94.5)	96.5 (95.1)	97.8 (96.5)
SEN	98.8 (99.4)	99.1 (92.9)	100.0(92.7)
SP	98.7 (89.6)	95.8 (89.7)	96.5 (92.7)
FP	5.5 (9.6)	8.7 (4.0)	5.5 (2.3)

표 11 절단위치 예측 정확도 (%)
Spase-HS (SignalP3-NN)

진핵생물	그람-음성균	그람-양성균
89.7 (85.4)	96.5 (85.0)	90.2 (80.5)

5. 결론 및 향후연구

본 논문에서는 미지의 아미노산 서열이 신호서열을 포함하고 있는지를 예측하고, 신호서열을 포함하는 것으로 예측된 경우에는 절단위치를 예측하는 방법을 제안하였다. 기존의 신호서열 예측방법에서는 아미노산의 종류와 서열상의 위치 정보를 사용하지만, 제안한 방법에서는 아미노산의 생물/화학적인 특성을 이용하였다. 소수성을 사용하여 아미노산의 생물체내에 가능한 위치에 대한 정보와 치환행렬을 사용하여 아미노산간의 진화적인 유사성을 이용하여 아미노산 서열간의 거리를 정의하였다.

제안한 방법을 기존의 방법들과 비교한 결과 예측정확도가 크게 향상된 것을 확인할 수 있었다. 패턴인식에서 높은 성능을 보이는 SVM을 효과적으로 적용하기 위한 전처리과정과 아미노산의 생물/화학적인 특성을 이용하였기 때문이라고 판단된다.

본 논문에서는 소수성을 이용하여 신호서열의 존재여부와 절단위치를 추정하기 위하여 저대역 통과 필터링에 의한 소수성의 평활화와 휴리스틱에 의한 절단위치 탐색방법을 사용하였다. 추후에는 신호처리 기법을 더욱 체계화 하고, 아미노산이 생체막에 내재되는지를 알아볼 수 있는 용매 모델(solvent model)을 이용한 접힘 에너지(folding energy)를 적용할 계획이다. 또한, 웹을 통해서 신호서열 예측이 가능하도록 웹 인터페이스를 개발할 예정이다.

참 고 문 헌

[1] von Heijne, G., "A new method for predicting signal sequence cleavage sites," Nucl. Acids Res., 14, pp. 4683-4690, 1986.

[2] von Heijne, G., "Sequence analysis in Molecular biology: Treasure trove or trivial pursuit," Academic, pp. 429-436, 1987.

[3] McGeoch, D.,J., "On the predictive recognition of signal peptide sequence," Virus Res., 3, pp. 271-286, 1985.

[4] Nakai, K., Horton, P., "PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization," Trends Biochem. Sci., 24, pp. 34-36, 1999.

[5] Bendtsen, J.,D., Nielsen, H., von Heijne, G., Brunak, S., "Improved prediction of signal peptides: SignalP 3.0," J. Mol. Biol., 340, pp. 783-795, 2004.

[6] Hua, S., Sun, Z., "Support vector machine approach for protein subcellular localization prediction," Bioinformatics, 17, pp. 721-728, 2001.

[7] Vert, J.P., "Support vector machine prediction of signal peptide cleavage site using a new class of kernels for strings," proc. pacific symposium on biocomputing, pp. 649-660, 2002.

[8] Menne, K.M., Hermjakob, H., Apweiler, R., "A comparison of signal sequence prediction methods using a test set of signal peptides," Bioinformatics, 16, pp. 741-742, 2000.

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

[10] Käll, L., Krogh, A., Sonnhammer, E.,L.,L., "A combined transmembrane topology and signal peptide prediction method," J. Mol. Biol., 338, pp. 1027-1036, 2004.

[11] http://www.expasy.org/sprot/download.html

[12] Engelman, D.M., Steitz, T.A., Goldman, A., "Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins," Annu. Rev. Biophys. Biophys. Chem., 15, 321-353, 1986.

[13] Kyte, J., Doolittle, R.F., "A simple method for displaying the hydropathic character of a protein," J. Mol. Biol., 157, pp. 105-132, 1982.

[14] Oppenheim, A.V., Schafer, R.W,. Discrete-time signal processing, Prentice-Hall, New Jersey, 1989.

[15] Henikoff, S., Henikoff, J.G., "Amino acid substitution matrices from protein blocks," proc. natl. acad. sci., 89, pp. 11915-11919, 1992.

[16] Boser, B., Guyon, I., Vapnik, V., "A training algorithm for optimal margin classifiers," proc. workshop, computational learning theory, pp. 144-152, 1992.

[17] Cortes, C., Vapnik, V., "Support-vector network," Machine learning, 20, pp. 273-297, 1995.

[18] Vapnik, V., The nature of statistical learning theory, Springer-Verlag, New York, NY, 1995.

[19] Rüping, S., mySVM-Manual, http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/

지 상 문

1991년 서울대학교 수학교육과(학사). 1993년 한국과학기술원 수학과(석사). 1998년 한국과학기술원 전산학과(박사). 1993년-2000년 삼성전자 정보통신. 2001년-현재 경성대학교 컴퓨터과학과 조교수. 관심분야는 기계학습, 생물정보학