실수 지수 메트릭으로 구성된 스트링 커널을 이용한 신호펩티드의 절단위치 예측

(Signal Peptide Cleavage Site Prediction Using a String Kernel with Real Exponent Metric)

지 상 문 †
(Sang-Mun Chi)

요 약 지지벡터기계는 자료간의 유사도를 커널함수를 사용하여 계산하고, 이러한 유사도를 이용하여 패턴을 분류하는 최적인 초평면을 구한다. 따라서 자료의 특성을 효과적으로 반영할 수 있는 유사도의 사용이 중요하다. 본 연구에서는 아미노산 서열간의 최적의 유사도를 얻기 위해서, 아미노산의 진화적인 관계와 소수성으로부터 유도된 메트릭을 실수 지수를 가지는 형태로 일반화하였다. 제안한 메트릭이 메트릭의 조건을 만족하고, 아미노산 서열과 DNA 서열의 유사도를 계산하기 위해서 널리 사용되는 스트링 커널 내에서 이용되는 메트릭과의 관련성을 알아본다. 또한, 적용하려는 문제에 보다 효과적인 메트릭을 일반화 메트릭에서 찾을 수 있음을 신호펩티드의 절단위치 예측실험을 통하여 알아본다.

키워드 : 지지벡터기계, 메트릭, 스트링 커널, 신호펩티드 절단위치

Abstract A kernel in support vector machines can be described as a similarity measure between data, and this measure is used to find an optimal hyperplane that classifies patterns. It is therefore important to effectively incorporate the characteristics of data into the similarity measure. To find an optimal similarity between amino acid sequences, we propose a real exponent exponential form of the two metrices, which are derived from the evolutionary relationships of amino acids and the hydrophobicity of amino acids. We prove that the proposed metric satisfies the conditions to be a metric, and we find a relation between the proposed metric and the metrics in the string kernels which are widely used for the processing of amino acid sequences and DNA sequences. In the prediction experiments on the cleavage site of the signal peptide, the optimal metric can be found in the proposed metrics.

Key words : Support Vector Machines, Metric, String Kernels, Signal Peptide Cleavage Site

1. 서 론

패턴 분류에 우수한 성능을 나타내고 있는 지지벡터기계(SVM: support vector machines)는 생물학적 서열(DNA 서열, 아미노산 서열)의 처리에도 성공적으로 적용되고 있다[1-9]. 일반적인 SVM은 수치자료인 벡터를 대상으로 하기 때문에, 수학적으로 정의된 벡터간의 내적을 사용할 수 있고, 보다 효과적인 분류를 위해서 비선형공간상에서 내적의 역할을 수행하는 커널함수를 사용한다. 하지만, 생물학적 서열은 벡터가 아니고 가변길이 문자열로 구성되어 있으므로, 이들 서열간의 유사도를 얻기 위한 방법으로 여러 스트링 커널이 연구되고 있다[6-10].

스트링 커널들은 두 서열을 이루는 문자열에서 일치하는 부분서열의 빈도가 높을수록 큰 유사도를 나타낸다. DNA 서열에서는 서열을 이루는 문자가 같은 경우에만 유사한 것으로 판정하는 방법이 일반적이지만, 아미노산 서열은 문자의 동일성을 유사도 척도로 사용하는 방법과[6-8], 치환행렬로 정의된 아미노산간의 진화적인 유사성을 이용하는 방법[9], 물리화학적인 특성을

· 이 논문은 2008학년도 경성대학교 학술연구비지원에 의하여 연구되었음

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

논문접수 : 2009년 6월 23일
심사완료 : 2009년 8월 20일

정보과학회논문지: 소프트웨어 및 응용 제36권 제10호(2009.10)

이용하는 방법이 있다[10].

본 논문에서는 스트링 커널에서 가장 기본적인 역할을 수행하는 서열을 이루는 문자간의 유사도를 구하는 방법을 연구한다. 아미노산 서열간의 유사도를 정의하기 위해서 널리 사용되는 문자의 동일성을 거리로 이용하는 방법과 치환행렬에서 유도한 거리를 사용하는 방법을 일반화하여 실수 지수를 가지는 메트릭을 제안하다. 제안한 메트릭이 메트릭의 조건들을 만족함을 보이고, 기존의 메트릭은 제안한 메트릭의 극한적인 예임을 보인다. 또한, 이들 두 극한적인 메트릭의 중간 영역에 존재하는 메트릭의 특성과 성능을 조사한다. 메트릭에 아미노산 자료의 특성을 반영하기 위한 방법으로 치환행렬에서 거리를 유도하는 방법과 더불어 소수성(hydrophobicity)에 기반한 거리함수를 사용한다. 소수성은 단백질 접힘(protein folding) 과정과 생체막에 내재하는 막단백질(membrane protein)의 형성에 중요한 역할을 하는 아미노산의 물리적인 특성으로[11,12], 아미노산 서열을 위한 메트릭에 이용될 수 있음을 알아본다.

제안한 메트릭을 신호펩티드의 절단위치 예측에 적용하기 위해서 여러 스트링 커널의 특징을 살펴본다. 스트링 커널이 아미노산 서열의 처리에 적용되는 경우에는 아미노산의 동일성만을 이용하기 보다는 치환행렬값을 사용하는 것이 높은 성능을 나타내고[9], 스플라이스 사이트와 같이 특정위치를 예측하는 문제[6]에 있어서는 서열상의 위치에 종속적인 유사성을 사용하는 것이 효과적이었다. 본 논문의 실험 대상인 신호펩티드 또는 신호서열은 분비단백질의 선도 아미노산서열로서, 이 신호펩티드가 절단되고 난 나머지 부분이 성숙된 단백질이 된다. 따라서, 아미노산 서열을 대상으로 하므로 치환행렬을 사용하는 방법과, 절단위치를 찾아내야 하므로 서열상의 위치에 종속적인 유사도를 계산하는 방법을 커널함수의 구성에 포함한다.

2. SVM과 스트링 커널

실험에 사용되는 파라미터의 의미와 메트릭이 사용되는 커널함수를 설명하기 위해서 SVM에 대해서 간략히 알아보고, 아미노산 서열의 처리에 사용되는 여러 가지 스트링커널을 살펴본다.

2.1 SVM

SVM은 두 부류를 분리하는 초평면(hyperplane)이 가장 가까운 자료와 거리가 최대화 되도록 한다[13-15]. 즉, $y_i \in \{1,-1\}$로 부류가 표시된 자료 x_i로 이루어진 공간에서,

$$f(x_i) = w^T x_i + b \tag{1}$$

가 y_i와 같은 부호를 갖고, 두 부류가 최대여백(maximal margin)을 갖도록 다음 문제에서 w, b를 구한다.

$$\min_{w,b} \ \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i \tag{2}$$
$$\text{with } \ y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

여기서, n은 학습 자료 개수, $\phi(\cdot)$는 비선형적인 분류 문제를 해결하기 위해서 x_i를 비선형공간으로 변환하는 함수이다. ξ_i는 잡음이나 이상치로 인해서 두 개의 부류로 정확히 나누어 지지 않는 자료를 처리할 수 있게 하며, C는 여백 크기와 학습자료 오분류 비율을 조절한다.

식 (2)를 해를 구하기 위해서 라그랑지안 함수를 w, b에 대해서 최소화한 후에, α에 대해서 다시 최대화하는 다음의 쌍대문제를 푼다.

$$\max_{\alpha} \ -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \phi(x_i)^T \phi(x_j) + \sum_{i=1}^{n} \alpha_i \tag{3}$$
$$\text{with } \sum_{j=1}^{n} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \ i = 1, \ldots, n.$$

여기서, $\phi(x_i)^T \phi(x_j)$대신, 커널 $K(x_i, x_j)$을 사용하여 $\phi(\cdot)$를 명시적으로 사용하지 않고 변환된 비선형 공간에서 내적을 계산 한다. 커널함수로 널리 사용되는 가우시안 커널

$$K(x,y) = \exp(-\gamma \, d^2(x,y)) \tag{4}$$

은 두 벡터 x, y의 거리 $d(x,y)$가 클수록 작은 유사성을 가지고, γ는 커널의 형태를 결정하는데, 클수록 폭이 좁은 모양이 되어 보다 복잡한 자료를 분류가 가능하다. 자료가 대규모일 경우에는 (3)을 구하기 위해서 일반적인 최적화 알고리즘을 사용하기 어렵고, 문제를 작은 단위로 나누어 처리하는 방법을 사용한다. 본 논문에서는 LIBSVM[16]을 사용하였다.

자료의 부류 판별은 (3)에서 얻은 α_i와 b를 사용한

$$f(x) = \sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + b \tag{5}$$

의 값에 따라 자료를 분류한다.

2.2 스트링 커널

식 (4)의 가우시안 커널과 같은 커널함수는 자료간의 유사도를 계산하고, 이를 기반으로 최적 초평면을 얻는데 이용된다. 스트링 커널은 문자열간의 유사도를 구하는 커널함수로서, 기본적으로 두 서열간의 공통된 부분서열의 개수가 많을수록 높은 유사도를 나타낸다. 공통된 부분을 계산하는 방법에 따라 여러 접근으로 나눌 수 있다.

스펙트럼 커널과 스펙트럼 커널을 향상시킨 미스매치 커널은 아미노산 서열정보로부터 단백질을 구조별로 분류하는 응용에 성공적으로 사용되었다[7,8]. 스펙트럼 커

널에서 k-스펙트럼은 길이 k인 연속된 부분서열로 정의하고, 두 서열에 포함되는 공통된 k-스펙트럼의 개수를 계산한다. 즉, 아미노산의 알파벳을 A, $\alpha \in A^k$를 k-스펙트럼, $\phi_\alpha(x)$를 서열 x에 포함된 α의 개수라 하고, 서열 x를 특징벡터 $\Phi_k(x) = (\phi_\alpha(x))_{\alpha \in A^k}$로 나타낼 때, 서열 x, y의 k-스펙트럼 커널은 $<\Phi_k(x), \Phi_k(y)>$으로 정의한다. 여기서, <,>은 일반적인 벡터간의 내적이다. 미스매치 커널은 길이 k인 부분서열에서 m개까지는 동일하지 않은 아미노산을 포함해도 일치하는 부분서열로 허용한다.

LA 커널(local alignment kernel)은 단백질 상동성 검출 (protein homology detection)에서 높은 성능을 보이는 Fisher 커널[1], 미스매치 커널[8]보다 향상된 성능을 보였다[9]. LA 커널은 두 서열에서 갭을 포함하여 부분적으로 정합된 점수들을 모두 합하여 유사도를 계산한다. 아미노산 사이의 유사도는 $K_a^{(\beta)}(x,y) = \exp(\beta s(x,y))$로 정의되는데, $s(x,y)$는 아미노산 x, y사이의 치환행렬[17]값이다. $K_{(n)}^{(\beta)}(x,y)$는 n개의 아미노산으로 구성된 서열 x, y간의 유사도로 아미노산 사이에 갭을 허용하고, 갭에 대한 패널티를 고려한 커널이고, 최종적으로 LA 커널은 $K_{LA}^{(\beta)}(x,y) = \sum_{i=0}^{\infty} K_{(i)}^{(\beta)}(x,y)$이다.

WD 커널(weighted degree kernel)은 유전체에서 스플라이스 위치(splice site)를 예측할 때 사용되었다[6]. DNA 서열에서 특정위치를 찾아내야 하므로 WD 커널은 위치에 종속적으로 서열간의 유사성을 계산한다. 서열 x에서 위치 l에서 시작하여 δ개의 DNA로 이루어진 부분서열을 $u_{\delta,l}(x) = x_l x_{l+1} \cdots x_{l+\delta-1}$로 표시할 때 WD 커널은 $K(x,y) = \sum_{\delta=1}^{d} w_\delta \sum_{l=1}^{N-\delta+1} I(u_{\delta,l}(x) = u_{\delta,l}(y))$로 정의된다. 여기서, $w_\delta = d - \delta + 1$인 가중치이고 $I()$는 괄호안이 같을 때만 1이고 나머지는 0인 함수이다. WD 커널은 스펙트럼커널과 유사하지만 위치특이적인 정보를 사용하는 점이 다르다. WDS 커널(weighted degree kernel with shifts)은 WD커널과 같고 부분서열의 이동을 허용한다. $K(x,y) = \sum_{\delta=1}^{d} w_\delta \sum_{l=1}^{N-\delta+1} \sum_{s=0}^{S(l)} \delta_s \mu_{\delta,l,s,x,y}$ 여기서 $S(l)$은 이동길이, $\delta_s = 1/(2(s+1))$이고, $\mu_{\delta,l,s,x,y} = I(u_{\delta,l+s}(x) = u_{\delta,l}(y)) + I(u_{\delta,l}(x) = u_{\delta,l+s}(y))$이다.

스트링 커널들의 성능을 비교해 보면 아미노산 서열을 처리하는 분야에서는 아미노산의 동일성만을 이용하는 스펙트럼 커널과 미스매치 커널보다는 치환행렬을 이용하는 LA 커널이 효과적이었다. 또한, 서열상의 특정위치를 예측하는 분야에서는 위치종속적인 방법인 WD 커널과 WDS 커널이 효과적이다. 따라서, 본 논문의 대상이 되는 신호서열의 절단위치 예측에서는 치환행렬에서 유도된 커널과 위치종속적인 정보를 포함한 커널이 효과적이라 예상된다.

3. 실수 지수를 가지는 메트릭

메트릭(metric) 또는 거리함수 $d(x,y)$는 다음조건을 만족하는 함수이다[18].

($M1$) d는 음이 아닌 유한한 실수값을 가진다.

($M2$) $x = y$일 때만 $d(x,y) = 0$.

($M3$) $d(x,y) = d(y,x)$ (대칭성).

($M4$) $d(x,y) \le d(x,z) + d(z,y)$ (삼각 부등식).

이 조건을 모두 만족하여야 거리로서 의미를 가질 수 있고 두 자료간의 거리를 정의할 수 있다. 거리와 유사도는 서로 반비례하는 관계가 있으므로, 자료간의 거리를 결정할 수 있으면 이를 이용하여 유사도를 정의할 수 있다. 본 논문에서는 아미노산간의 유사도를 계산하는 스트링커널과 여러 가지 메트릭과의 관련성을 이용하여 효과적인 커널을 구성한다. 먼저 이산메트릭(discrete metric) d_d는 다음과 같이 정의된다[18].

$$d_d(x,x) = 0,\ d_d(x,y) = 1 (x \neq y). \tag{6}$$

여기서, x, y는 커널로 계산되는 두 서열에서 대응되는 아미노산이나 DNA이다. 이산메트릭은 대응되는 문자의 동일성만을 이용하여 거리를 정의하고, 메트릭의 조건을 만족한다. 스펙트럼 커널, 미스매치 커널, WD 커널과 WDS 커널을 이루는 기본 단위인 $I(x=y)$는 $1 - d_d(x,y)$와 같으므로 이들 커널은 이산메트릭을 사용하여 구성되었다고 볼 수 있다.

LA 커널의 경우에는 커널의 계산에서 대응되는 두 아미노산 x, y의 유사도를 치환행렬값 $s(x,y)$를 사용하여 나타낸다. 치환행렬값 $s(x,y)$는 x와 y의 유사도를 나타내므로 내적 $<x,y>$과 비슷한 성질을 가진다. 내적으로부터 거리를 유도하는 식인, $d(x,y)^2 = <x-y, x-y> = <x,x> + <y,y> - 2<x,y>$에서 내적을 치환행렬값으로 대치하면 다음의 메트릭 d_s를 정의할 수 있다.

$$d_s(x,y) = \sqrt{s(x,x) + s(y,y) - 2s(x,y)}. \tag{7}$$

메트릭 d_s의 경우에는 ($M2$)와 ($M3$)을 만족하는 것은 명백하다. 다른 두 조건 ($M1$)과 ($M4$)를 만족하는지는 주어진 치환행렬에 대해서 아미노산의 가능한 모든 조합에 대한 조사를 통해서 알 수 있고, 본 논문에서 사용하는 치환행렬에 대해서는 메트릭의 조건을 만족함을 확인하였다. 수학적으로는 아미노산 서열이 벡터공간을 이루지는 않으므로 내적을 정의할 수 없고 치환행렬값과 내적을 직접적으로 대응시킬 수는 없다. 하지만 식

(7)의 값과 LA 커널함수값의 반비례관계는 성립하므로 LA 커널은 메트릭 d_s 값과 관련이 크다.

본 논문에서는 메트릭 $d(x,y)$을 일반화하기 위해서 실수 지수 p를 가지는 형태로 변환한

$$d(x,y)^p,\ 0 \le p \le 1 \tag{8}$$

p-메트릭을 제안한다. 실수 지수 p가 1일 때는 본래의 메트릭 d이고, p가 0일 경우에는 이산 메트릭 d_d가 된다. p-메트릭은 메트릭의 조건을 만족한다.

이유: p-메트릭 $d(x,y)^p$는 정의로부터 $(M1)$, $(M2)$와 $(M3)$을 만족함은 명백하므로, $(M4)$를 만족하는지를 알아본다. d가 메트릭이므로

$$d(x,y) \le d(x,z) + d(z,y), \tag{가}$$

을 만족한다. 또한,

$$[d(x,z)+d(z,y)]^p \le d(x,z)^p + d(z,y)^p,\ 0 \le p \le 1 \tag{나}$$

은 $d(x,z)=d(z,y)=0$이면 양변이 0이 되어 성립한다. $d(x,z) \ne 0$일 때는 양변을 양수인 $d(x,z)^p$로 나눈 형태인 $[1+d(z,y)/d(x,z)]^p \le 1+[d(z,y)/d(x,z)]^p$의 성립과 (나)의 성립은 동치이다. 이는 $s=d(z,y)/d(x,z)$로 놓았을 때 $f(s)=1+s^p-(1+s)^p$, $s \ge 0, 0 \le p \le 1$에서 $f(s) \ge 0$과 동치다. $df(s)/ds = ps^{p-1} - p(1+s)^{p-1} = p(s^{p-1}-(1+s)^{p-1}) \ge 0$ 이고 $f(0)=0$이므로 $f(s)$는 0보다 크다. 부등식 (가)와 (나)에 의해서

$$d(x,y)^p \le d(x,z)^p + d(z,y)^p$$

이므로 $d(x,y)^p$는 $(M4)$를 만족하여 메트릭이 된다.

p가 1일 때는 본래의 메트릭 d가 된다. 또한, $x=y$일 때는 $\lim_{p\to 0} d(x,y)^p = \lim_{p\to 0} 0^p = 0$이고, $x \ne y$일 때는 $d(x,y) \ne 0$이므로 $\lim_{p\to 0} d(x,y)^p = 1$이다. 따라서 p가 0일 경우에는 이산 메트릭 d_d가 된다. ■

$p>1$인 경우에는

$$d(x,y)^p,\ p>1 \tag{9}$$

$(M1)$, $(M2)$와 $(M3)$을 만족함은 명백하고, 주어진 메트릭에 종속적으로 $(M4)$의 만족함이 결정된다. 따라서, 가능한 모든 조합의 x, y, z에 대해서 $(M4)$를 만족하면 메트릭이 된다. 아미노산과 DNA는 종류가 20개와 4개로 한정되어 있으므로 모든 가능한 조합에 대해서 삼각부등식이 만족하는지를 조사하는 것은 가능하므로 $d(x,y)^p, p>1$가 메트릭이 되는지를 판단할 수 있다.

대부분의 관련연구에서는 p-메트릭 $d_s(x,y)^p$에서 p가 1인 치환행렬를 사용한 메트릭 d_s와 p가 0인 이산 메트릭 d_d을 사용하고 있으나, $0<p<1$인 중간 영역도 메트릭이 되고, $p>1$인 경우에는 삼각부등식을 만족하면 메트릭이 되므로 SVM의 응용분야에 따라 적합한 p-메트릭 $d_s(x,y)^p, p \ge 0$을 선택할 수 있어서, 보다 적용범위가 큰 메트릭이 된다. p-메트릭에서 p의 변화에 따라 각기 다른 메트릭이 되는데, 각 아미노산간의 거리의 순서는 바뀌지 않으나, 상대적인 차가 바뀌게 된다. 즉, p가 0일 경우에는 서로 다른 x, y들 간의 거리는 모두 1로 같고, p가 증가함에 따라 상대적이 차이가 커진다.

매트릭을 확장하기 위한 방법으로, 두 개의 각기 다른 메트릭 d_1, d_2를

$$d(x,y) = \sqrt{d_1(x,y)^2 + d_2(x,y)^2} \tag{10}$$

로 결합하여도 메트릭이[18] 된다. 따라서, 서로 다른 메트릭들을 결합하여 메트릭으로 사용할 수 있다. 또한, 두 개의 벡터 $x=(x_1,x_2,\ldots,x_n)$와 $y=(y_1,y_2,\ldots,y_n)$간의 메트릭은 다음을 사용하였다.

$$d_E(x,y) = \sqrt{d(x_1,y_1)^2 + d(x_2,y_2)^2 + \ldots + d(x_n,y_n)^2} \tag{11}$$

본 논문에서는 아미노산 간의 거리 d_s는 BLOSUM 행렬[17]을 사용하여 식 (7)로 계산하였다. 실험대상인 신호서열은 서열상의 보존도가 작으므로 상동성이 작은 단백질간의 진화적 거리를 계산하기에 유용한 BLOSUM50을 사용하였다. 치환행렬 BLOSUM50의 경우에는 아미노산간의 유사도는 -5에서 15사이의 값을 가진다.

또한, 소수성(hydrophobicity)의 차이를 이용한 거리

$$d_h(x,y) = |h(x) - h(y)| \tag{12}$$

를 사용하였다. 여기서, $h(x)$는 아미노산 x에 대응되는 소수성도[12]로서 -12.3에서 3.7까지의 값을 가진다. 소수성이란 무극성분자들(nonpolar molecules)은 물과 수소결합 또는 이온성 상호작용에 참여하지 않으므로, 물분자들끼리의 상호작용처럼 유리하게 일어나지 않기 때문에 무극성분자들은 물에서 응집하려는 경향을 가지는 것을 말한다. 이러한 소수성 효과(hydrophobic effect)가 단백질 접힘(protein folding)과 생체막에 내재하는 막단백질(membrane proteins)의 형성에 중요한 역할을 한다. 신호펩티드는 아미노산 서열의 선도서열로서 물분자들 사이의 지질막인 소포체막에 내재되므로 소수성이 중요한 역할을 수행하므로, 아미노산간의 거리에 소수성을 이용한 메트릭 d_h를 사용하였다.

4. 실험 및 분석

4.1 실험자료와 실험조건

신호펩티드의 절단위치를 예측하는 실험을 위해서 생화학적 방법에 의해 구조가 밝혀진 신호서열을 포함하는 단백질자료가 필요하다. 신호서열 자료에서 전문가들이 오류를 제거하여 자료의 신뢰성을 향상 시킨 SPDB 자료[19]를 실험에 사용하였는데, 1984개의 진핵생물 단

백질과 328개의 그람-음성세균 단백질을 포함하고 있다.

식 (1)의 SVM을 학습하기 위한 $y_i=1$에 해당하는 자료는 절단위치 이전의 17개, 이후의 2개 아미노산으로 구성된 서열을 사용하였고, $y_i=-1$에 해당하는 서열은 아미노산의 서열의 모든 위치에서 시작하여 길이가 19인 부분서열을 추출한 후에 $y_i=1$에 해당하는 서열을 제외하였다. 예비실험을 수행한 결과 서열의 길이와 식 (2)의 C값의 변화에 따라 비슷한 성능을 보이므로 서열의 길이는 19, C는 5로 고정된 값을 사용하였다. 식 (4)의 γ는 최적값을 선택하기 위해서 자료를 무작위로 나누어 80%를 학습에 사용하고 나머지 20%를 평가에 사용하였고, 자료의 분할에 따라 실험결과가 달라지므로 반복하여 실험한 값의 평균을 나타내었다. 절단위치 예측방법의 성능은 평가에 사용되는 신호펩티드 중에서 절단 위치가 정확히 예측된 비율로 나타내었다.

연구[11,20]에 따르면 절단위치 이전에 존재하는 특정 아미노산들은 효과적인 절단을 막으며 절단위치 바로 이전위치(-1위치)에 빈번히 나타나는 아미노산의 종류는 일부분으로 한정되어 있다. 또한, 절단위치 이전의 -21 위치에서 -4위치까지도 아미노산의 종류에 따른 다음의 정보량은 다르다.

$$I_j = \log_2 20 + \sum_{\alpha} n_j(\alpha)/N_j \circ \log(n_j(\alpha)/N_j)$$

여기서 j는 절단위치를 기준으로 하는 서열상의 위치이고, $n_j(\alpha)$는 아미노산 α가 위치 j에 나타나는 횟수이고, N_j는 위치 j에 나타나는 모든 아미노산의 수이다. $\log_2 20$은 아미노산의 종류 20개가 나타내는 최대 엔트로피이므로 I_j는 최대엔트로피와 위치 j에서의 엔트로피의 차이이다. 따라서, 위치 j에 나타나는 아미노산의 종류가 작으면 작을수록 I_j는 커지게 되고, 그 위치에서 신호서열의 특징적인 아미노산의 분포를 나타낸다고 판단할 수 있다. 본 논문에서는 그람-음성균의 I_j값이 크므로 이를 학습자료에서 추정하여 서열 x와 y에서 j번째 아미노산인 $x[j]$와 $y[j]$간의 거리의 제곱은 $I_j \cdot d_p(x[i],y[i])^2$을 사용하여 가중치를 주었다.

4.2 신호서열의 절단위치 예측

아미노산 서열을 대상으로 높은 성능을 보이는 LA 커널을 사용하여 그람-음성균의 절단위치를 예측하였다. 실험을 10회하여 정확도의 평균을 나타내었다. 파라미터 β가 $0.03 \le \beta \le 0.11$인 경우의 성능은 표 1과 같았고, 이외의 범위에서는 정확도가 50%이하였다. LA 커널은 부분 서열의 공통점을 이용하는 스트링커널 중에서 최적의 성능을 가지고 있으나, 서열의 모든 위치에서의 부분서열의 일치도를 계산하므로 절단위치를 찾기 위한 위치특이적인 면을 반영하지 않으므로 신호서열의 절단위치 예측이 있어서는 정확도가 높지 않았다.

표 1 LA 커널의 예측 정확도 (%)

β	0.03	0.05	0.07	0.09	0.11
%	59.9	68.4	70.1	70.0	62.6

LA 커널은 서열 길이의 제곱에 비례한 시간 복잡도인 반면 WD, WDS 커널은 선형의 시간 복잡도이므로 실험의 반복회수를 증가시켜서 100번으로 하였다. 표 2의 정확도를 보면 LA 커널보다 훨씬 향상된 정확도를 보였다. 이는 절단위치를 기준으로 위치특이적인 면을 고려하기 때문이라고 판단된다. 표 2에서 δ는 WDS 커널에서 부분서열의 길이이고, $S(l)$은 이동길이이고 $S(l)=0$이면 WD커널에 해당한다. m은 미스매치 커널과 같이 m개까지는 동일하지 않은 아미노산을 포함해도 일치하는 부분서열로 허용함을 의미하고 $m=0$일 때는 WDS 커널이다. 이밖에 δ가 5일 때 다양한 $S(l)$과 m에 대해서 조사하였으나 성능의 차이가 크지 않았다.

표 2 WDS 커널의 예측 정확도 (%)

δ	$S(l)$	m	정확도
1	0	0	89.5
3	0	0	89.3
3	0	1	89.4
3	0	2	89.7
3	2	0	90.0
3	2	1	89.1
3	2	2	90.1

WDS 커널은 이산 메트릭을 사용하여 아미노산의 동일성 여부만을 이용하므로 아미노산의 유사성을 사용하는 LA 커널에 비해서 사용하는 정보가 한정적이다. 하지만 서열상의 위치가 인접한 아미노산만을 비교하여 절단위치의 예측에는 표 2에서 보듯이 효과적이다. 본 논문에서는 WDS 커널과 같이 서열상의 위치 정보를 사용하면서 LA 커널처럼 아미노산의 유사성 정보를 동시에 사용한다. 또한, $d(x,y)^p$에서 기존에 사용되던 $p=0$인 이산 메트릭과 $p=1$인 메트릭과 더불어, $0<p<1$인 메트릭들과 삼각부등식을 만족하는 $p>1$인 메트릭의 성능을 실험한다. WDS 커널에서 $\delta=1, S(l)=0, m=0$인 가장 간단한 커널에 대해서 $d(x,y)^p$를 적용하였다. 먼저 그람-음성 세균 자료를 무작위로 분할하여 2000회 반복 실험하여 절단위치 예측 정확도의 평균을 구하였다.

표 3의 d_h^p은 식 (12)의 소수성에 기반한 메트릭을 나

표 3 그람-음성균의 절단위치 예측 정확도 (%)

d_h^p	p	1	0.5	0.4	0.3	0.2	0.1
	%	65.7	89.6	90.5	91.2	**91.4**	91.3
d_s^p	p	1.4	1.3	1.2	1.1	1.0	0.9
	%	**91.4**	91.3	91.2	91.3	**91.4**	91.3

타낸다. p가 1일 경우의 소수성 자체로는 효과적인 거리함수의 역할을 하지 못하였지만, p가 0으로 접근하는 제안한 메트릭을 사용함에 따라 성능이 증가하여 $p=0.2$에서 최대가 되었고, p가 0인 이산 메트릭으로 수렴함에 따라 다시 성능이 감소하였다. 즉, 최적의 메트릭이 p가 1인 본래의 메트릭과 p가 0인 이산 메트릭 사이에 존재하였다. 표 3의 d_s^p는 치환행렬에 기반한 식 (7)의 메트릭을 나타낸다. $p=2$일 때 삼각부등식을 만족함을 모든 아미노산의 조합에 대해서 조사를 통하여 확인하였다. 따라서, $0 \le p \le 2$인 d_s^p를 사용하였고, $p=1.4, 1.0$일 때 좋은 성능을 보였다.

표 4는 진핵생물 자료를 사용한 결과이다. 자료의 수가 그람-음성균보다 크므로 실험결과가 실험마다 크게 변하지 않으므로 100회 반복 실험하였다.

표 4 진핵생물의 절단위치 예측 정확도 (%)

d_h^p	p	1	0.5	0.4	0.3	0.2	0.1
	%	53.9	82.0	83.1	84.1	**84.1**	83.9
d_s^p	p	1.4	1.3	1.2	1.1	1.0	0.9
	%	83.2	83.1	84.0	84.1	**84.2**	84.1

진핵생물의 경우에 WDS 커널에서 $\delta=1, S(l)=0, m=0$에 해당하는 $p=0$인 이산 메트릭은 82.9%의 정확도였으므로 소수성에 기반한 메트릭과 치환행렬에 기반한 메트릭이 모두 향상된 성능을 보임을 알 수 있다.

표 5는 식 (10)을 사용하여 소수성에 기반한 메트릭과 치환행렬에 기반한 메트릭중에서 최적인 메트릭을 결합한 메트릭의 성능을 나타내었는데, 표 3과 표 4에 비하여 향상된 성능을 보였다. 표 3에서 $d_h(x,y)^p$는 $p=0.2$에서 최적이었고, $d_s(x,y)^p$는 $p=1.4$일 때 91.38%이었고, $p=1$일 때 91.36%이므로 $p=1.4$가 최적이었다. 따라서, $d(x,y)=\sqrt{(d_h(x,y)^{0.2})^2+(d_s(x,y)^{1.4})^2}$를 그람-음성균에 사용하였다. 표 4에서 $d_h(x,y)^p$는 $p=0.3$일 때 84.07%, $p=0.2$일 때 84.14%로 $p=0.2$가 최적이고, $d_s(x,y)^p$는 $p=1$일 때 최적이었다. 따라서, 진핵생물에는 $d(x,y)=\sqrt{(d_h(x,y)^{0.2})^2+(d_s(x,y)^{1.0})^2}$를 사용하였다.

표 5 결합 p-메트릭의 정확도 (%)

그람-음성균	진핵생물
91.60	85.13

4.3 신호서열 예측 방법과 비교

본 논문에서 제안한 p-메트릭은 이산메트릭과 치환행렬값을 이용하는 메트릭보다 신호펩티드의 절단위치를 예측하는데 효과적임을 실험을 통해 확인하였다. 이절에서는 메트릭과 패턴분류 알고리즘을 포함한 전체적인 절단위치 예측시스템과 비교한다.

신호펩티드의 절단위치를 예측하는 방법 중에서 가장 높은 성능을 보이는 방법은 SignalP로 알려져 있다 [20-22]. 이 방법은 20가지 종류의 아미노산을 나타내기 위해서, 각각의 아미노산을 20차원의 벡터로 나타내고, 아미노산의 종류에 따라 20차원 중 한 요소만 1이고 나머지는 0으로 하는 희소코딩(sparse coding)을 사용한다. 이는 아미노산의 동일성 여부만을 사용하는 이산 메트릭과 유사하다. 패턴분류를 위한 모델로서 신경망과 은닉 마르코프 모델을 사용하는데 신경망을 사용하는 방법이 평균적으로 성능이 높다.

SignalP는 SWISS-PROT 40이전의 자료로 학습되었으므로, 본 논문에서 사용한 자료중 SWISS-PROT 40 이후의 자료만을 사용하여 평가하였다. 그람-음성균은 190개였고, 진핵생물은 1492개이다. SignalP 웹인터페이스를 사용하여 절단위치를 예측하였다[20]. 신경망을 사용하는 SignalP3-NN의 결과를 표 6에 나타내었다. 표 5의 제안한 방법보다 약간 저하된 성능을 보인다.

표 6 SignalP 절단위치 예측 정확도 (%)

그람-음성균	진핵생물
89.47	85.05

5. 결론 및 향후연구

본 논문에서는 여러 스트링 커널에 사용되고 있는 메트릭을 조사하였고, 이를 일반화하여 실수 지수를 가지는 p-메트릭을 제안하였고, 제안한 메트릭이 수학적으로 메트릭의 조건을 만족함을 보였다. 신호펩티드의 절단위치 예측에 제안한 메트릭을 적용한 결과 효과적으로 절단위치을 예측함을 확인하였다. 스트링 커널에서 사용되는 이산메트릭과 치환행렬에서 유도한 메트릭은 p-메트릭에서 $p=0$, $p=1$인 특별한 경우이고, 이들 특별한 경우를 제외한 영역에서 적용하려는 문제에 적합한 메트릭이 존재할 수 있음을 실험을 통해서 알아보았다.

본 논문에서는 WDS 커널에서 $\delta=1$, $S(l)=0$, $m=0$인 가장 간단한 커널만을 사용하였는데, 다양한 부분서

열의 길이, 이동길이, 미스매치 허용을 고려하면 향상된 성능을 얻을 수 있을 것이다. 이를 위해서는 WDS 커널에서 사용되는 이산메트릭을 p-메트릭으로 대치하여야 하고, WDS 커널에서 사용되는 가중치를 p-메트릭에 적합하게 바꾸는 연구가 필요하다.

향후에는 제안한 p-메트릭을 메트릭을 사용하는 다양한 계산생물학의 분야에 적용할 예정이다.

참 고 문 헌

[1] Jaakkola, T., Diekhans, M., Haussler, D., "A discriminative framework for detecting remote protein homologies," *J. Comp. Biol.*, 7, pp.95-114, 2000.

[2] Pavlidis, P., Weston, J., Cai, J. and Noble, W. S., "Learning gene functional classifications from multiple data types," *J. Comp. Biol.*, 9, pp. 401-411, 2002.

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

[4] Zavaljevski, N., Stevens, F. J. and Reifman, J., "Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions," *Bioinformatics*, 18, pp.689-696, 2002.

[5] 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.

[6] Sonnenburg, S., Schweikert, G., Philips, P., Behr, J. and Rätsch, G., "Accurate splice site prediction using support vector machines," *BMC Bioinformatics*, 8(Suppl 10):S7, 2007.

[7] Leslie, C., Eskin, E. and Noble, W. S., "The spectrum kernel: A string kernel for SVM protein classification," *proc. pacific symposium on biocomputing*, pp.566-575, 2002.

[8] Leslie, C., Eskin, E., Cohen, A., Weston, J. and Noble, W. S., "Mismatch string kernels for discriminative protein classification," *Bioinformatics*, 20, pp.467-476, 2004.

[9] Saigo, H. Vert, J.-P., Akutsu, T. and Ueda, N., "Protein homology detection using string alignment kernels," *Bioinformatics*, 20, pp. 1682-1689, 2004.

[10] Kim, J.K., Bang, S.Y., and Choi, S., "Sequence driven features for prediction of subcellular localization of proteins" *Pattern Recognition*, 39(12), pp.2301-2311, 2006.

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

[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, pp.321-353, 1986.

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

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

[15] Vapnik, V., Statistical learning theory, John Wiley & Sons, 1998.

[16] Chang, C-C. and Lin, C-J., LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

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

[18] Kreyszig, E.. Introductory Functional Analysis with Applications, John Wiley & Sons, New York, 1978.

[19] Choo, KH, Tan TW and Ranganathan, S., "SPdb - a signal peptide database," *BMC Bioinformatics*, 6:248, 2005.

[20] 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.

[21] 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.

[22] 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.

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