한국지능시스템학회 논문지 2009, Vol. 19, No. 5, pp. 730-736

N-Block substring 가중 선형모형을 이용한 단백질 CDS의 특징 추출 및 분류

Feature Selection and Classification of Protein CDS Using n-Block substring weighted Linear Model

최성용* · 김진수* · 한승진** · 최준혁*** · 임기욱**** · 이정현*

Seong-Yong Choi*, Jin-Su Kim*, Seung-Jin Han**, Jun-Hyeog Choi***, Kee-Wook Rim**** and Jung-Hyun Lee*

* 인하대학교 컴퓨터공학과

** 경인여자대학 정보미디어학부

*** 김포대학 e-비지니스과

**** 선문대학교 컴퓨터정보학부

요 약

방대한 유전 정보를 분석, 가공하는 생명정보학의 중요성은 더욱 높아지고 있다. 본 논문에서는 단백질의 1차 구조만으로 단백질의 구조와 기능을 예측하는 새로운 데이터마이닝 방법을 제안한다. 단백질 서열만으로 특징 추출시 발생할 수 있는 문제점인 방대한 탐색공간을 효과적으로 축소하기 위해 n-Block substring 탐색 알고리즘을 제안한다. 또한 선별된 각 substring의 도메인 연관도를 결정하는 가중치를 구하여 가중 선형모형을 구축함으로써 구조와 기능에 관련이 있을 것으로 예상되는 단백질 도메인의 특징을 추출하고 분류에 효과적임을 보인다. 도메인에 포함되는 각각의 CDS(coding sequence)에 대해 모형으로부터 구한 점수를 통해 해당 도메인과의 연관성의 정도를 추정하며, 분류 효율을 더욱 향상시킬 수 있음을 보인다.

키워드 : 바이오 인포매틱스, 데이터 마이닝, 가중선형모형, n-Block 서브스트링

Abstract

It is more important to analysis of huge gemonics data in Bioinformatics. Here we present a novel datamining approach to predict structure and function using protein's primnary structure only. We propose not also to develope n-Block substring search algorithm in reducing enormous search space effectively in relation to feature selection, but to formulate weighted linear algorithm in a prediction of structure and function of a protein using primary structure. And we show efficient in protein domain characterization and classification by calculation weight value in determining domain association in each selected substring, and also reveal that more efficient results are acquired through claculated model score result in an inference about degree of association with each CDS(coding sequence) in domain.

Key Words : Bioinformatics, datamining, weighted linear model, n-Block substring

1. 서 론

게놈 프로젝트의 완성으로 방대한 양의 기초적인 유전 정보들이 여러 데이터베이스에 저장되어 있다. 그러나 현재까지 440억개 이상의 염기 서열과 4천만개 이상의 아미노산 서열이 밝혀진 것에 비해 비록 그 수가 빠른 속도로 증가하고는 있지만 단백질의 구조와 기능에 밀접한 연관성이 있는 단백질의 3차 구조는 전체의 1%도 되지 않는 약 3만 여개 밖에 밝혀지지 않았다 [1-3].

X-ray crystallography 또는 NMR spectroscopy와 같은 실험적 방법을 이용해서 단백질의 구조와 특성들을 결정할 수 있지만 진행 속도가 느리고 많은 비용이 요구되는 문제점이 발생한다. 이러한 상황에서 단백질의 순수 아미노산 서열인 1차 구조 정보로부터 단백질의 분자 구조와 생물학적 기능을 예측하려는 연구가 활발하다 [4-7]. 그러나 많은 단백질들의 구조와 특성이 밝혀짐에 따라 단백질의 서열 유사도(similarity)가 높지 않은 경우에도 유사한 구조와 기능을 가지는 동종 단백질(homology protein)들이 많이 존재한다는 사실이 알려져 있으며, 그와 같은 이유로 순수 단백질 서열정보만으로 단백질의 구조와 기능을 예측하는 경우에 어려움이 발생할 수 있다. 또한 서열로부터 얻을 수 있는 정보들은 경우에 따라 다르게 나타나며, 이에 따른 정보

접수일자 : 2008년 7월 31일
완료일자 : 2009년 9월 10일
본 연구는 지식경제부 및 정보통신산업진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음 (NIPA-2009-C1090-0902-0020)

의 부족으로 기존의 서열 정보에 추가하여 부가적인 정보들을 결합하여 유사한 구조와 기능을 갖는 군집으로 분류하는 방법들이 있다 [8-10].

이러한 시도들은 구조와 기능이 알려진 단백질 서열 군집의 상동성(homology)에 기반하여 새로이 발견된 단백질 서열의 구조와 기능을 예측하고, 밝혀진 구조와 기능 정보를 이용하여 진화 과정을 예측할 수 있으며 실험적 방법을 이용할 수 없는 특수한 상황에서도 단백질의 구조 및 특성 정보를 얻을 수 있는 유일한 대안이 될 수 있다. 이와 같이 유사한 구조와 기능을 갖는 단백질들로부터 공통된 부가적인 정보들을 결합하여 단백질의 구조와 기능을 분류한 PROSITE, PRINTS, Pfam, BLOCK, SBASE와 같은 2차 단백질 데이터베이스들이 있다. 각각의 데이터베이스들은 서로 다른 방법들과 다양한 형태의 생물학적 정보들을 사용하여 분류하였기 때문에 그 결과는 모든 데이터 베이스마다 조금씩 다르게 나타난다. 이러한 2차 단백질 서열 데이터베이스는 새롭게 발견된 단백질 서열에 대해 구조와 기능을 추측할 수 있는 중요한 도구가 될 수 있지만 각각의 데이터베이스가 제공하는 조금씩 다른 결과로 인해 사용자의 혼돈이 발생할 수 있는 문제가 있다 [11].

또한 PROSITE는 같은 구조와 기능을 갖는다고 알려진 단백질 집합에서 공통된 특성이 될 수 있는 모티프와 다양한 형태의 생물학적 정보들을 사용하여 family, domain, repeat, zinc finger의 4 종류의 계층적 분류를 수행하여 각 계층에 밝혀진 단백질들을 분류해 놓았다 [3]. 모티프는 공통의 구조 또는 기능을 공유하는 특정한 단백질 서열의 일부 지역 혹은 부분을 가리키는 용어로써 단백질에서 다양한 기능을 수행하고 조절한다고 알려져 있으며, 이러한 하나 혹은 그 이상의 모티프가 단백질들을 특성화 하는데 사용될 수 있고, 이로부터 임의의 단백질 서열에서 특정한 모티프를 찾아내어 특정 단백질 군집으로의 분류가 가능하다 [12]. 그러나 진화 혹은 변이에 의한 서열과 구조의 변화시, 기존의 밝혀진 모티프로는 군집을 특성화 할 수 없는 문제점이 발생 할 수 있으며 또한 데이터베이스에서 분류된 단백질 서열들을 모티프만으로는 완벽하게 분류할 수 없는 단점이 있다.

본 논문에서는 분류된 단백질 데이터베이스의 단백질들의 군집을 순수 아미노산 서열만으로 해당 분류를 예측할 수 있는 상동성 모형화(homology modelling) 방법을 제안한다. 1차 서열만으로 특징 추출시 발생할 수 있는 방대한 탐색공간을 효과적으로 축소하기 위해 n-Block substring 탐색 알고리즘을 제안하고, 2차 단백질 데이터베이스에서 분류된 각 군집에서 탐색된 n-Block substring을 각각의 중요도에 따라 서로 다른 가중치를 부여하여 가중 선형모형을 구축함으로써 새로운 단백질들에 대해 각 도메인들의 상동성을 점수화하여 그 구조 및 기능을 예측한다.

2. 탐색공간 설정 및 n-Block substring 탐색

본 논문에서는 1차 구조 정보만으로 단백질 군집의 특징을 추출하기 위해 2차 단백질 데이터베이스에서 분류된 단백질 군집에 속하는 각각의 순수 단백질 아미노산 서열들로부터 연속적이고 다양한 크기의 모든 아미노산 조합을 고려하여 해당 군집의 공통된 특징으로 결정한다. 이러한 시도에서의 문제점은 각 아미노산 서열들로부터 탐색해야 하는 탐색공간의 규모가 너무 방대하다는 것이다. 예를 들어, 두 개의 아미노산으로 구성되는 2-Block substring의 아미노산 조합은 총 400개(20×20)에 한정되지만 3-Block substring의 아미노산 조합은 8,000개($20\times20\times20$), 4-Block substring의 아미노산 조합은 160,000개($20\times20\times20\times20$), n-Block substring의 아미노산 조합은 20^n개로 탐색공간이 지수적으로 증가한다. 이에 대한 해결 방안으로 방대한 전체 탐색공간을 탐색하지 않고 탐색된 2-Block substring 으로부터 만들어진 n-Block substring만을 탐색하는 n-Block substring 탐색 알고리즘을 제안한다.

표 1은 n-Block substring 탐색 알고리즘을 위해 이미 분류된 특정 단백질 군집에 속하는 모든 단백질들을 대상으로 400개의 가능한 모든 2-Block substring에 대한 빈도 정보를 저장하기 위한 2-Block substring 빈도표이다.

표 1. 2-Block substring 빈도표.
Table 1. 2-Block substring frequency table.

2-Block substring \ 군집내 서열	seq_1	seq_2	…	seq_m	지지도
AA					
AC					
⋮					
YY					

표 2. n-Block substring 탐색 알고리즘.
Table 2. n-Block substring search algorithm.

```
/*특정 단백질 군집에서 공통의 n-Block substring 탐색
2-BS : 지지도를 만족하는 2-Block substring의
저장공간(입력)
n-BS :  지지도를 만족하는 n -Block substring의
저장공간(출력)
CDS : 특정 군집내 CDS들의 저장공간
Temp : 탐색된 substring의 임시 기억공간
tempcnt : 특정 CDS에서 탐색된 substring의 빈도수
supportcnt : 탐색된 substring을 포함하는 CDS 수*/

 For ( n=2 ; count(n-BS)=0 ; n++)  {
  For ( i=1 ; i > count(n-BS) ; i++)  {
   For ( j=1 ; j > count(2-BS) ; j++)   {
    If (Mid(n-BS(i), n, 1) = Mid(2-BS(j), 1, 1) )
    {
        Temp = (Left(n-BS(i), n)+Mid(2-BS(j), 2, 1)
        // 지지도 계산 시작
        For (k = 1 ; k > count(CDS) ; k++)
        {
          For (l = 1 ; l > count(CDS(k))-n+1 ; l++) {
          If (Temp = Mid(CDS(k), l, n+1))
          { tempcnt = tempcnt + 1 }
          }
          If (tempcnt >= 1)
          { supportcnt = supportcnt + 1 }
        }
        // 지지도 계산 종료
          // 지지도를 만족하는 substring을 저장
        If (supportcnt / count(CDS) > 지지도)
        { (n+1)-BS ← Temp }
       tempcnt = 0
       supportcnt = 0
    }
   }
  }
 }
```

$$\text{지지도} = \frac{n-Block\ \ substring \text{ 을 포함하는 군집내 단백질 수}}{\text{군집내 전체 단백질 수}} \quad (1)$$

식 (1)의 지지도에 따라 특정 단백질 군집의 공통된 특징이 될 수 있는 2-Block substring을 탐색한다. 예를 들어 특정 단백질 군집내의 단백질 서열이 모두 10개이고 어떤 2-Block substring을 포함하는 단백질들이 9개라고 한다면 해당 군집에서의 substring 지지도는 90%가 된다. 미리 정한 임계값 이상의 지지도를 나타내는 2-Block substring들의 모든 가능한 조합을 통하여 3-Block substring을 구성한다. 만약 탐색된 2-Block substring이 AA, AC, CA 라고 가정한다면 2-Block substring의 두 번째 아미노산과 2-Block substring의 첫 번째 아미노산이 일치하는 3-Block substring을 구성한다. 즉, 후보 3-Block substring은 AA + AA = AAA, AA + AC = AAC, AC + CA = ACA, CA + AA= CAA, CA + AC= CAC의 5가지로 얻어진다. 이와 같이 얻어진 후보 3-Block substring으로부터 식 (1)의 지지도를 계산하여 미리 정한 임계값 이상의 3-Block substring을 탐색한다. 이와 같은 방법으로 표 2의 n-Block substring 탐색 알고리즘을 사용하여 특정 임계값 이상의 지지도를 만족하는 n-Block substring이 존재하지 않을 때까지 탐색을 반복하고, 탐색된 n-Block substring을 해당 군집의 특징으로 사용한다.

3. n-Block substring의 가중 선형모형

3.1 가중치 결정

n-Block substring 탐색 알고리즘으로 탐색된 지지도를 만족하는 모든 substring들은 해당 단백질 군집의 특징이 될 수 있다. 그러나 같은 n-Block substring이라고 하더라도 각각의 단백질 군집에서의 중요성은 다를 수 있고, 그로 인해 분류 성능을 낮출 수 있으므로 보다 효과적인 단백질 군집 분류를 위해 각 sbustring의 중요도를 나타내는 가중치를 계산한다.

표 3은 특정 단백질 군집에서 군집내 단백질(seq_m)들 각각에 대해 탐색된 n-Block substring($substring_n$)의 빈도수($freq_{nm}$)를 나타낸다.

표 3. 탐색된 n-Block substring 의 단백질별 빈도수.
Table 3. Protein frequency of searched n-Block substring.

군집내 서열 / substring	seq_1	seq_2	⋯	seq_m
$substring_1$	$freq_{11}$	$freq_{12}$		$freq_{1m}$
$substring_2$	$freq_{21}$	$freq_{22}$		$freq_{2m}$
⋮				
$substring_n$	$freq_{n1}$	$freq_{n2}$		$freq_{nm}$

$$\boldsymbol{AVG} = (seqavg_1, seqavg_2, \cdots, seqavg_m)$$
$$, \ seqavg_i = avg(freq_{1i}, freq_{2i}, \cdots, freq_{ni}) \quad (2)$$

$$\boldsymbol{substrin}\, g_i = (freq_{i1}, freq_{i2}, \cdots, freq_{im}) \quad (3)$$

$$distance(\boldsymbol{AVG, substrin}\, g_i)_i = \| \boldsymbol{AVG - substrin}\, g_i \|, i = 1,2,\cdots n \quad (4)$$

표 3으로부터 특정 단백질 군집에서 각각의 substring의 중요도를 나타내는 가중치를 구하기 위해 군집내 각 단백질들에서 탐색된 substring의 평균 빈도를 구해 해당 군집의 평균 벡터(AVG)를 식 (2)를 이용하여 구한다.(식 (2),(3),(4)에서 굵은체는 벡터를 의미한다.)

군집내 모든 서열들이 생성하는 벡터 공간에서 식 (2)로부터 계산된 군집의 평균 벡터와 군집내 각 substring 벡터(식 (3))와의 유클리드 거리(식 (4))를 이용하여 특정 군집의 단백질들이 만들어내는 벡터 공간의 중심으로부터 각 substring들이 떨어져 있는 정도를 가중치로 설정 할 수 있다. 즉, 적은 거리를 나타내는 substring은 특정 군집에서 더욱 중요한 특징이 될 수 있고, 반대로 큰 거리를 나타내는 substring은 특정 군집에서 의미가 없는 특징이라고 볼 수는 없지만 군집의 특성을 나타내는 중요한 특징이라고 볼 수도 없다. 즉, 식 (4)로부터 얻어진 거리 정보가 가중치와는 반대의 의미로 직접 모형에 적용할 수 없기 때문에 식 (5)를 이용하여 벡터 공간의 중심으로부터의 거리가 가까울수록 높은 가중치를 할당하고 그 반대의 경우에는 낮은 가중치를 할당한다. 식에 나타난 $distance_{\max}$는 특정 단백질 군집에서 계산된 모든 거리들 중 최대값을 나타내며 $distance_{\min}$은 최소값을 나타낸다. 이 때, 계산된 가중치의 범위는 $0 \leqq weight_i \leqq 1$이다.

$$weight_i = 1 - \frac{(distance_i - distance_{\min})}{distance_{\max}} \quad (5)$$

3.2 가중 선형모형 및 군집 예측 시스템

모든 단백질 군집으로부터 각 단백질 군집에서 탐색된 substring들과 계산된 가중치들로부터 식 (6)을 이용하여 가중 선형모형을 구축하고 해당 정보를 단백질 군집 모형 데이터베이스에 저장한다.

$$\text{가중 선형모형} = \sum_{i=1}^{n} weight_i \cdot substring_i \quad (6)$$

$$\text{모형점수} = \frac{\sum_{i=1}^{n} weight_i \cdot substring_i}{\sum_{i=1}^{n} weight_i} \quad (7)$$

$$\text{여기서 } substring_i = \begin{cases} 1 & count(substring_i) \neq 0 \\ 0 & count(substring_i) = 0 \end{cases}$$

새로운 단백질 서열을 가중 선형모형을 이용하여 구축된 시스템에 입력시, 각각의 단백질 군집에 속하게 될 가능성을 식 (7)의 모형점수 공식을 이용하여 가장 높은 모형점수를 보이는 군집에 할당함으로써 새로운 단백질의 구조와 기능을 예측하게 된다.

그림 1은 본 논문에서 제안한 n-Block substring과 가중 선형모형을 이용한 단백질 군집 예측 시스템의 구성도를 나타낸다.

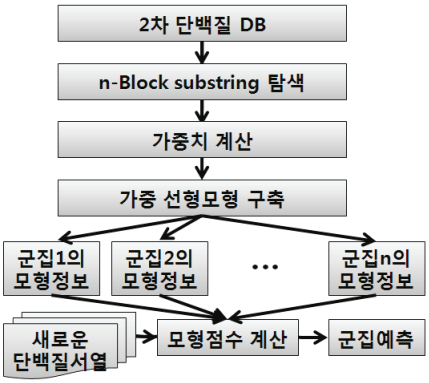

그림 1. 단백질 군집 예측 시스템의 구성도
Fig. 1. Schematic diagram of protein cluster prediction system.

4. 실험 및 성능평가

4.1 평가 척도

단백질 서열 분류에서 동종 단백질 서열로 정확히 분류하는 TP(True Positive)와 그렇지 못한 단백질 서열을 명확하게 구분하는 TN(True Negative)로 분류하는 것이 가장 바람직하지만 실제의 경우에는 어떠한 상동성이 존재하지 않아도 상동성이 존재한다고 결정하는 오류인 FP(False Positive)와 실제로 존재하는 동종 단백질 서열들을 올바르게 분류하지 못하는 오류인 FN(False Negative)를 발생시켜 분류 효율을 저하시키는 문제가 발생한다.

본 논문에서 사용되는 평가 척도로 정보검색의 정확률(Precision)과 재현율(Recall)에 기초한 PPV(Positive Prediction Value), Sensitivity, 그리고 Specificity를 사용한다 [5]. PPV는 모형에 의해 올바르게 분류된 전체 서열중 동종 단백질 군집에 속하는 서열들의 비율을 의미하고, Sensitivity는 동종 단백질 군집에 속하는 서열들 중에서 모형에 의해 올바르게 분류된 서열들의 비율을 말하며, Specificity는 전체 서열 중 동종 단백질 군집에 포함되지 않는 서열들 중 모형에 일치하지 않는 정도를 나타내는 척도로써 각각 식 (8),(9),(10)에 의해 계산된다.

$$PPV = \frac{TP}{TP + FP} \tag{8}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{9}$$

$$Specificity = \frac{TN}{TN + FP} \tag{10}$$

4.2 실험방법 및 결과

본 논문에서는 2차 단백질 데이터베이스 중의 하나인 PROSITE에서 계층적으로 분류된 4개의 단백질 군집(family, domain, repeat, zinc finger)중 도메인에 속하는전체 704개의 단백질 군집에 대해 임의로 추출된 13개의 도메인과, 그에 속하는 324개의 CDS(coding sequence : 단백질로 변환되는 DNA/RNA 서열 영역)를 이용하여 실험하였다.

표 4는 실험에 사용된 13개의 도메인과 324개의 CDS를 나타낸다 [3]. 실험에 사용된 방법은 13개의 도메인 각각에 대해 해당 도메인에 포함된 모든 CDS 데이터를 모형 구축을 위한 훈련 데이터로 정한 후, n-block substring 탐색 알고리즘을 사용하여 얻어진 substring들에 가중치를 구하여 가중 선형모형을 구축한 후, 324개의 전체 데이터를 가지고 도메인 분류 실험을 진행하였다.

표 4. 실험에 사용된 도메인과 CDS.
Table 4. Domain and CDS used in experiment.

번호	도메인	CDS수
1	3′-5′ exonuclease	36
2	A1pp	23
3	A.T hook DNA-binding	48
4	ACBP	35
5	acetyltransferase	8
6	albumin	35
7	alpha NAC	21
8	alpha-type protein kinase	10
9	antistasin	4
10	asparagine synthetase	23
11	autotransporter(TC 1.B.12)	70
12	AWS	4
13	AXH	7
합계		324

표 5. 지지도 100%에서 n-Block substring의 수와 분류 결과.
Table 5. Number of N-Block substring and classification result in support 100%.

도메인	CDS	Sub-string	TP	FP	TN	FN	PPV	Speci-ficity
3′-5′ exonuclease	36	34	36	53	235	0	40.45%	81.60%
A1pp	23	68	23	6	295	0	79.31%	98.01%
A.T hook DNA-binding	48	8	48	30	246	0	61.54%	89.13%
ACBP	35	탐색된 substring 없음						
acetyltransferase	8	122	8	0	316	0	100%	100%
albumin	35	11	35	44	246	0	44.30%	84.78%
alpha NAC	21	28	21	8	295	0	72.41%	97.36%
alpha-type protein kinase	10	159	10	0	314	0	100%	100%
antistasin	4	1467	4	0	320	0	100%	100%
asparagine synthetase	23	91	23	0	301	0	100%	100%
autotransporter (TC 1.B.12)	70	45	70	27	227	0	72.16%	89.37%
AWS	4	323	4	0	320	0	100%	100%
AXH	7	213	7	0	317	0	100%	100%
합계/평균	324						80.85%	95.02%

표 6. 3′-5′ exonuclease domain에서 지지도별 CDS 모형 점수 및 평균.
Table 6. Each CDS model score and average in 3′-5′ exonuclease domain.

No	CDS	90%	80%	70%	60%	50%
1	P09155	0.79	0.64	0.60	0.51	0.40
2	P44442	0.79	0.71	0.68	0.59	0.49
3	O08307	0.96	0.96	0.91	0.88	0.81
4	O09053	0.99	0.98	0.96	0.87	0.74
5	O32801	1.00	0.95	0.92	0.89	0.86
6	O34996	0.98	0.93	0.92	0.90	0.89
7	O51498	0.96	0.89	0.87	0.83	0.74
8	O64235	0.93	0.89	0.85	0.76	0.63
9	O67779	0.94	0.84	0.81	0.73	0.57
10	O93530	0.97	0.97	0.96	0.90	0.75
11	P00582	0.97	0.95	0.94	0.90	0.84
12	P0A550	0.98	0.95	0.92	0.87	0.80
13	P0A551	0.98	0.95	0.92	0.87	0.80
14	P19822	0.99	0.96	0.92	0.83	0.69
15	P30314	0.97	0.94	0.93	0.87	0.72
16	P34603	0.90	0.83	0.81	0.74	0.62
17	934607	0.95	0.91	0.91	0.83	0.70
18	P43741	1.00	0.97	0.95	0.92	0.86
19	P46835	0.99	0.96	0.94	0.89	0.82
20	P52026	0.99	0.96	0.93	0.91	0.87
21	P56105	0.95	0.90	0.88	0.86	0.79
22	P59199	1.00	0.98	0.95	0.91	0.87
23	P59200	1.00	0.98	0.95	0.91	0.87
24	P74933	0.97	0.95	0.93	0.88	0.81
25	Q01780	0.98	0.95	0.91	0.83	0.67
26	Q04957	0.99	0.96	0.92	0.91	0.87
27	Q05254	0.90	0.84	0.80	0.73	0.61
28	Q14191	0.99	0.97	0.96	0.88	0.75
29	Q55971	0.97	0.95	0.93	0.91	0.86
30	Q59156	0.95	0.94	0.92	0.86	0.79
31	Q8VEG4	0.88	0.82	0.79	0.67	0.55
32	Q9CDS1	1.00	0.97	0.94	0.90	0.86
33	Q9F173	0.98	0.96	0.95	0.90	0.84
34	Q9NVH0	0.87	0.80	0.76	0.68	0.55
35	Q9S1G2	1.00	0.94	0.92	0.87	0.85
36	Q9ZJE9	0.95	0.90	0.88	0.86	0.78

No	CDS	평균
1	P09155	0.59
2	P44442	0.65
3	P43741	0.73
4	P52026	0.74
5	Q55971	0.78
6	O51498	0.78
7	Q59156	0.78
8	O64235	0.81
9	P34603	0.86
10	O08307	0.86
11	P00582	0.87
12	P59200	0.88
13	Q14191	0.88
14	P19822	0.88
15	P30314	0.89
16	Q9F173	0.89
17	Q9NVH0	0.90
18	Q9ZJE9	0.90
19	P56105	0.90
20	Q8VEG4	0.91
21	O32801	0.91
22	Q9S1G2	0.91
23	P34607	0.91
24	P59199	0.92
25	Q01780	0.92
26	P0A551	0.92
27	O34996	0.92
28	Q05254	0.93
29	O67779	0.93
30	Q9CDS1	0.93
31	O09053	0.93
32	P46835	0.93
33	Q04957	0.93
34	P0A550	0.94
35	P74933	0.94
36	O93530	0.94

표 5는 단백질 도메인에 속하는 모든 CDS를 만족하도록 지지도의 임계값을 100%로 설정하고 실험한 결과로서 탐색된 n-Block substring의 수, 분류된 결과인 TP, FP, TN, FN, PPV 그리고 Specificity의 값을 보여준다. 이 때의 Sensitivity는 지지도를 100%로 설정하여 모든 도메인에서 100%가 되기 때문에 표에서는 생략하였다. 실험에 사용된 13개의 도메인중 6개의 도메인에서 PPV 100%, Specificity 100%로 전체 서열 데이터에서 해당 도메인을 모두 정확하게 분류했으며, 평균 PPV는 80.85%, 평균 Specificity는 95.02%로 높게 나타났다. 그러나 이는 지지도를 100%로 설정한 경우의 실험 결과로서, 일부 서열이 특정 도메인의 구조와 기능에 연관성이 부족하거나, 또는 어떠한 오류로 인해 잘못 분류가 되어 있는 경우 전체 분류 성능의 효율을 낮추게 할 가능성이 존재한다.

표 6의 왼쪽은 이러한 가능성을 밝혀내고, 전체 분류의 효율을 높이기 위해 실험 결과 중 가장 낮은 PPV(40.45%)의 값을 보이는 3′-5′ exonuclease 도메인만을 가지고 실험한 결과로써, 실험 방법은 지지도를 90%부터 50%까지 10% 간격으로 변화시키며 각 지지도에서 모형을 새롭게 구축하고, 구축된 모형에서의 3′-5′ exonuclease 도메인에 속하는 모든 CDS의 모형 점수를 구한 결과이고, 표 6의 오른쪽은 이들 각 CDS의 지지도별 모형 점수를 평균하여 점수별로 오름차순으로 정렬한 결과이다. 이 때, 점수가 높을수록 구축된 모형에 적합하다고 볼 수 있지만, 반대로 점수가 낮을 경우에는 구축된 모형에는 적합하지 않다고 볼 수 있다.

그림 2는 표 6 오른쪽의 각 CDS의 지지도별 모형 점수 평균을 그래프로 나타낸 것이다. 그래프에 나타난 선들은 3′-5′ exonuclease 도메인내의 CDS들을 나타내며, 그래프 위쪽의 높은 점수의 CDS는 해당 도메인을 잘 설명한다고 볼 수 있으며, 반대로 그래프 아래쪽의 낮은 점수의 CDS는 해당 도메인을 잘 설명한다고 볼 수 없다. 그림 2의 제일 아래쪽의 P09155 : [RND_ECOLI Ribonuclease D]의 평균 점수는 0.59로 도메인내의 CDS중 가장 적은 값을 나타낸다.

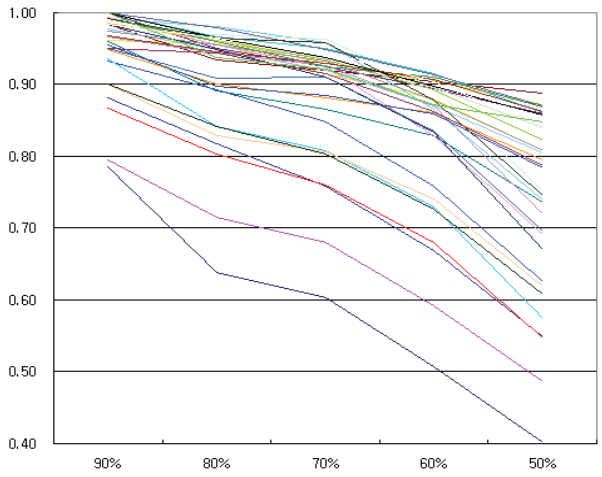

그림 2. 3′-5′ exonuclease domain에서 각 CDS의 모형점수비교.
Fig. 2. Model score comparison of each CDS in 3′-5′ exonuclease domain.

그림 3은 지지도 100%에서 가장 좋지 못한 결과를 보였던 3′-5′ exonuclease 도메인에서, 가장 작은 모형점수를 가지는 CDS순으로 하나씩 제거하여 모형을 구축한 후 얻어진 PPV, Specificity, Sensitivity의 값을 나타낸 것이다. P09155, P44442, P43741, P52026의 4개의 CDS를 삭제한 후의 모형 구축시 40.45%에서 54.10%의 13.65%의 성능 향상을 보였다.

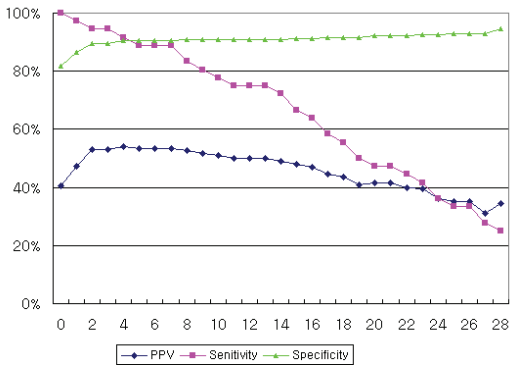

그림 3. Sensitivity 변화시 PPV와 Specificity 변화
Fig. 3. PPV and Specificity variation at Sensitivity change.

5. 결 론

실험에서 ACBP 도메인의 경우 다른 도메인들과는 달리 n-Block substring을 찾아내지 못한 이유는 해당 도메인의 일부 CDS의 길이가 다른 CDS와 비교하여 상당히 작은 것에 그 주요 원인이 있다. 즉, 35개의 CDS중에서 P81624 : [ACBP1_DIGLA]의 서열은 30개의 아미노산으로 이루어져 있으며, 그 길이가 다른 서열과 비교하여 아주 짧기 때문에 substring을 탐색하지 못했다. 이러한 경우에는 지지도를 설정하는 임계값을 조금 낮추어 줌으로써 가중 선형모형에 사용되는 substring을 탐색할 수 있다.

본 논문에서 사용된 방법의 실험 결과는 순수 단백질 서열만을 가지고 Sensitivity 100%에서 평균 PPV 80.85%, Specificity 95.02%이었다. 비록 모든 단백질 도메인을 완벽하게 예측하지는 못했지만 순수 단백질 서열만으로 단백질의 구조와 기능을 예측할 수 있는 가능성을 다시 한번 확인함과 동시에, 특정 도메인내의 CDS에 대해 모형으로부터 구한 모형점수를 통해 다른 도메인과의 연관성의 정도를 추측해 볼 수 있고 이로부터 도메인들의 유사성 추정에도 이용될 수 있을 것으로 보인다.

참 고 문 헌

[1] http://www.ncbi.nlm.nih.gov/Genbank/genbank-stats
[2] http://www.rcsb.org/pdb/holdings.do
[3] http://ca.expasy.org/cgi-bin/get-similar?all=domain
[4] A. Brazma, L. Jonassen, I. Eidhammer, and D. Gilbert, “Approaches to the automatic discovery of patterns in bioseqeucnces,” *Jounal of Computational Biology*, no. 5, pp. 279-305, 1998.
[5] A. Karwath and R. D. King, “Homology Induction: the use of machine learning to improve sequence similarity searches,” *BMC Bioinformatics*, vol. 3, no. 11, 2002.
[6] D. Kell, and R. D. King, “On the Optimization of Classes for the Assignment of Unidentified Reading Frames in Functional Genomics Programmes: The Need for Machine Learning,” *Trends in Biotechnology*, vol. 3, no. 18, pp. 93-98, 2000.
[7] R. D. King, A. Karwath, A. Clare, and L. Dehapse, "Accurate Prediction of Protein Class in the M. tuberculosis and E. coli Genomes Using Data Mining," *Yeast (Comparative and Functional Genomics)*, vol. 4, no. 17, pp. 283-293, 2000.
[8] J. Park, S. A. Teichmann, T. Hubbard, and C. Chothia, "Intermediate sequences increase the detection of homology between sequences," *Journal of Molecular Biology*, vol. 1, no. 273 pp. 349 - 54, 1997.
[9] T. J. P. Hubbard, A, G. Murzin, S. E. Brenner, and C. Chothia, "SCOP: a Structural Classification of Proteins database," *Nucleic Acids Research*, vol. 25, no. 1, pp. 236-239, 1997.
[10] R. D. King, H. W. Paul, A. Clare, "Confirmation of data mining based predictions of protein function," *Bioinformatics* vol. 7, no. 20, pp. 1110-1118, 2004.
[11] F. S. Domingues, W. A. Koppensteiner, M. J. Sippl, "The role of protein structure in genomics," *FEBS Letters*, no. 476, pp. 98-102, 2000
[12] Y. Gao, K. Mathee, G. Narasimhan, X. Wang, "Motif Detection in Protein Sequences," In *Proceedings of SPIRE*, pp. 63-72, 1999.

저 자 소 개

최성용(Seong-Yong Choi)
1993년 : 인하대학교 통계학과 학사
2001년 : 인하대학교 통계학과 석사
2001년~현재 : 인하대학교 컴퓨터정보공학과 박사과정

관심분야 : 유비쿼터스 컴퓨팅, 무선 센서 네트워크, 임베디드 시스템, 데이터마이닝
E-mail : choisymail@gmail.com

김진수(Jin-Su Kim)
1998년 : 인천대학교 전자계산공학과 학사
2001년 : 인하대학교 컴퓨터공학과 석사
2001년~현재 : 인하대학교 컴퓨터정보공학과 박사과정

관심분야 : 유비쿼터스 컴퓨팅, 무선 센서 네트워크, 데이터마이닝, 정보검색
E-mail : kjspace@inha.ac.kr

한승진(Seung-Jin Han)

1990년 : 인하대학교 전자계산공학과 학사
1992년 : 인하대학교 전자계산공학과 석사
2002년 : 인하대학교 전자계산공학과 박사
2002년～2004년 : 인하대학교 컴퓨터공학부 강의조교수
2004년～현재 : 경인여자대학 정보미디어학부 조교수
2007년～현재 : TTA PG103 표준화위원

관심분야 : USN, MANET, Mobile Computing, 임베디드 시스템, Security
E-mail : softman@kic.ac.kr

최준혁(Jun-Hyeog Choi)

1990년 : 경기대학교 전자계산공학과 학사
1995년 : 인하대학교 전자계산공학과 석사
2000년 : 인하대학교 전자계산공학과 박사
1997년～현재 : 김포대학 e-비즈니스과 부교수
2003년～현재 : 김포발전연구소 소장

관심분야 : 정보검색, 유전자 알고리즘, 신경망, USN, 임베디드 시스템, 전자상거래 보안
E-mail : jhchoi@kimpo.ac.kr

임기욱(Kee-Wook Rim)

1977년 : 인하대학교 전자공학과 학사
1987년 : 한양대학교 전자계산학과 석사
1994년 : 인하대학교 전자계산공학과 박사
1989년～1996년 : 한국전자통신연구원 시스템연구부장, 주전산기(타이컴)Ⅲ,Ⅳ 개발사업 책임자
2001년～1999년 : 한국전자통신연구원 컴퓨터소프트웨어 연구소장
2000년～현재 : 선문대학교 컴퓨터정보학부 교수

관심분야 : 실시간데이터베이스시스템, 운영체제, 시스템구조
E-mail : rim@sunmoon.ac.kr

이정현(Jung-Hyun Lee)

1977년 : 인하대학교 전자공학과 학사
1980년 : 인하대학교 전자공학과 석사
1988년 : 인하대학교 전자공학과 박사
1979년～1981년 : 한국전자기술 연구소 시스템 연구원
1984년～1989년 : 경기대학교 전자계산학과 교수
1989년～현재 : 인하대학교 컴퓨터공학부 교수

관심분야 : 자연어처리, HCI, 음성인식, 정보검색, 고성능 컴퓨터구조
E-mail : jhlee@inha.ac.kr