ProSLP : Penta-gram 기반 단백질의 세포내 위치 예측 시스템

박호은, 김진숙
{hepark,jinsuk}@kisti.re.kr
한국과학기술정보연구원 바이오인포매틱스센터
대전광역시 유성구 유성우체국 사서함 122호, 305-600

초 록

ProSLP는 펜타 그램(penta-gram) 색인방법과 *k*-nearest neighbor(kNN) 알고리즘을 이용하여 단백질의 세포내 위치를 예측하는 시스템이다. ProSLP는 입력된 서열에 대하여 높은 유사도를 가지는 상위 *k*개의 서열이 가지는 세포내 위치 정보를 이용하여 단백질의 세포내 위치를 예측한다. 최적화된 *k* 값을 결정하고 위치 예측 성능을 평가하기 위하여 훈련데이터집합(Training set)과 시험데이터집합(Test set)으로 구성된 두 개의 테스트컬렉션(Test Collection)을 이용하여 실험을 진행하였다. 성능평가방법은 정확률(Precision), 재현율(Recall) 및 정확률과 재현율의 break-even point를 이용한 성능평가방법을 사용하였다. ProSLP는 두 테스트컬렉션에 대해 각각 93%와 98%이상의 높은 break-even point를 보였고 또한 시스템의 성능은 *k* 값의 변동에 대하여 큰 영향을 받지 않았다. ProSLP는 http://proslp.kisti.re.kr을 통하여 사용할 수 있다.

Abstract

A new prediction system for protein subcellular localization site(s), called ProSLP, has been developed. The ProSLP system classifies a protein sequence to one or more subcellular compartments based on the location of top *k* sequences which show the highest similarity against the input sequence. Currently the ProSLP extracts penta-grams as features of the sequences, computes scores of the potential localization site(s) using *k*-nearest neighbor (kNN) algorithm and finally presents the predicted site(s) and their associated scores. For two test collection tested so far, the ProSLP shows precision-recall break-even points of more than 0.93 and 0.98, respectively. The ProSLP service is available at http://proslp.kisti.re.kr.

1. 서론

인간 게놈 프로젝트를 비롯한 거대 염기서열분석 프로젝트의 결과로, 기능이 아직 규명되지 않은 새로운 단백질의 서열이 급격히 증가하고 있고, 단백질 데이터베이스들의 데이터 용량 또한 아주 빠른 속도로 증가하고 있다. 이러한 생물정보 데이터의 급속한 증가는 생물학자들에게 단백질의 기능 규명이라는 새로 문제를 야기하게 되었다. 단백질의 기능을 규명함에 있어 단백질의 세포내 위치에 대한 정보를 알 수 있다는 것은 중요한 열쇠가 될 수 있다. 따라서 주어진 서열의 기능 규명을 위한 중요한 단서로서 단백질의 세포내 위치를 결정하는 것은 생물학 또는 생물정보학자들에게 아주 중요한 일이다 (Chou and Elrod, 1999). 또한 단백질의 세포내 위치를 예측할 수 있다면 기능 규명을 위한 복잡한 단계들을 대폭 줄여줄 수 있게 될 것이다.

하나의 생리학적 기능에 참여하는 일련의 단백질들은 세포 내에서 동일한 위치에 존재하여야 한다 (Eisenhaber and Bork, 1998). 이는 단백질의 생리학적 기능은 단백질이 세포 내에서 존재하는 위치와 밀접한 관계가 있다는 것을 의미한다. 또한 신진대사 과정, 신호전달 체계, 단백질의 3차원 구조와의 관계와 같은 생리학적 기능 또는 구조적인 기능을 이해하기 위하여 단백질의 세포내 위치를 확인하는 것은 매우 중요한 일이라는 것을 알려주고 있다. 단백질의 세포내 위치 예측은 단백질의 기능에 대한 중요한 단서를 제공해 줄 뿐만 아니라 신약 개발을 위한 후보군 선별, 유전자 생성물에 대한 주석 자동화, 단백질에 대한 보다 심도 있는 연구 등을 가능하게 해 준다 (Gardy et al., 2003).

단백질의 세포내 위치에 대한 예측은 주로 단백질 서열 데이터베이스에 대한 유사도 검색을 통한 방법에 의존해 왔다. 실험실에서 다양한 방법으로 세포내 위치를 확인할 수도 있지만 이러한 과정은 너무 많은 시간과 비용을 필요로 한다. 데이터베이스에 입력되는 단백질 서열의 수가 매우 빠르게 증가하기 때문에 단백질의 세포내 위치를 신속하게 결정하는 단계는 단백질의 기능을 밝히기 위한 과정에 있어 아주 중요한 단계가 될 수 있다.

지난 수년간 온라인을 통하여 이용 가능한 여러 자동화된 세포내 위치 예측 시스템이 개발되어왔다. 이러한 시스템으로는 TargetP (Emanuelsson *et al.*, 2000), PSORT (Nakai and Horton, 1999), NNPSL (Reinhardt and Hubbard, 1998), MitoProt/Predotar (Feng, 2002), PLOC (Park and Kanehisa, 2003) 등이 있다. 이 시스템들은 세포내 위치 예측을 위하여 두 가지 방법을 이용하여 개발되었다. 한가지 방법은 아미노산 조성에만 의존한 예측 방법이다. 예측하고자 하는 서열을 기존의 서열 데이터베이스와 비교하여 유사한 서열을 검색하고 이를 이용하여 입력된 단백질 서열의 세포내 위치를 예측한다. PLOC, NNPSL, SubLoc 등의 시스템에서 아미노산 조성에 의존한 방법을 사용하고 있다. 그러나 단백질 서열들을 아미노산

조성으로만 분석하였을 때 많은 정보를 잃을 수 있다는 사실이 이미 알려져 있다. 두 번째 방법은 아미노산 조성과 함께 시그널 펩타이드(Signal Peptide)를 분석하여 단백질의 세포내 위치를 예측한다. 그러나 많은 경우에 있어서 시그널 펩타이드를 찾을 수 없고 전체 서열에서 부분적으로 지정할 수가 없기 때문에 시그널 펩타이드에 의존하는 방법은 다소 문제를 야기할 수 있다.

문서 범주화(Text categorization) 분야의 연구자들은 문서를 표현하기 위한 적합한 자질(Feature)의 추출이 범주화의 효율성(Effectiveness)에 있어서 매우 중요하다는 것을 지적하였다 (Sebastiani, 2002). 또한 대용량의 문서에 대한 문서범주기의 종류는 kNN 범주기를 제외하고는 실용적이지 못하다는 것이 증명되어 왔다 (Yang, 1994; Sebastiani, 2002). 본 논문에서는 단백질의 세포내 위치 분류를 위한 효율적인 자질추출(Feature extraction) 도구로써 n-그램(gram) 방법을 도입하였다. kNN 범주기와 n-그램(gram) 방법을 이용하여 대용량 세포내 위치 예측 서비스를 구축하였다. 이 예측 시스템은 ProSLP(Protein Subcellular Localization Prediction의 약자)라 명명되었으며 http://proslp.kisti.re.kr 에서 온라인으로 이용가능하다.

2. 시스템 및 방법

2.1 구축 시스템 환경

ProSLP 세포내 위치예측시스템은 2GB 메모리를 가진 펜티엄-III 컴퓨터와 리눅스 (Linux) 운영체제하에서 운영되고 있다. 이 예측 시스템은 KRISTAL-2000[1]이라고 불리는 정보 검색 시스템에 *k*-nearest neighbor(kNN) 문서범주화 알고리즘, n-그램(gram) 자질 추출 방법을 함께 도입하여 만들어졌다. 모든 서열 데이터에 대하여 n-그램(gram) 자질을 추출하고 추출된 자질은 역파일(Inverted File)로 KRISTAL-2000 시스템에 저장된다.

2.2 자질 추출 (Feature Extraction)

자연어로 작성된 문서는 사람이 이해할 수 있는 단어들로 구성되어 있다는 사실은 명백하다. 따라서 문서 범주화에 있어 자질 추출은 문서에 포함되어 있는 단어(Word)나 구(Phrase) 등으로 이루어진다. 하나의 단백질 서열 또한 21개의 아미노산 정보로 이루어진 언어로 쓰여진 문서로 간주할 수 있다. 그러나 아미노산 서열로 이루어진 단순한 문자열로부터 어떻게 의미있는 자질들을 추출하느냐 하는 것은 명확하지 않다. 이 문제를 해결하기 위해 ProSLP는 고정된 길이 n의 간격으로 중복되어지는 n-그램(gram) 자질 추출 방법을 도입하였다. 가령 길이 n의 값이 5

[1] http://giis.kisti.re.kr 을 통하여 KRISTAL-2000 정보 검색 및 관리 시스템에 대한 자세한 정보를 볼 수 있다.

라고 한다면 단백질 서열 "ACDEFLERR"은 "ACDEF", "CDEFL, "DEFLE", "EFLER", "FLERR"로 분할되어진다. 데이터베이스의 모든 서열에 대하여 자질의 추출과 색인작업(Indexting)이 완료되면 일반적인 정보검색 방법을 이용하여 단백질 서열을 검색할 수 있다. 현재 ProSLP는 자질 추출 방법으로 n-그램(gram)의 간격이 5인 펜타 그램(penta-gram) 방식을 사용하고 있다.

2.3 유사도 측정 (Similariy Measures)

세포내 위치를 예측하고자 하는 새로운 단백질 서열이 입력되면 먼저 데이터베이스로부터 가장 유사도가 높은 상위 k 서열을 검색하고, 검색된 상위 k 서열이 가지는 세포내 위치정보로부터 입력 서열의 적합한 위치정보를 예측한다. 질의(Query) 서열과 대상(Target) 서열간의 유사도는 펜타 그램(penta-gram)의 통계정보를 기반으로 하는 벡터 공간 모델 (Vector space model)을 이용하여 계산된다. 질의 서열 q 와 대상 서열 s 간의 유사도의 계산식은 다음과 같다.

$$Sim(q,s) = \frac{1}{W_s} \cdot \sum_{t \in q \wedge s} w_{s,t} w_{q,t}$$

with:

$$W_s = \log(1 + \sum_{t \in s} f_{s,t})$$

$$w_{s,t} = \log(f_{s,t} + 1) \cdot \log(\frac{N}{f_t} + 1)$$

$$w_{q,t} = \log(f_{q,t} + 1) \cdot \log(\frac{N}{f_t} + 1)$$

$f_{s,t}$은 서열 s에 대한 펜타 그램(penta-gram) 토큰 t의 빈도; N은 데이터 집합의 전체 서열의 수; f_t는 펜타 그램(penta-gram) 토큰 t가 한번 또는 그 이상 발생하는 빈도; $w_{s,t}$는 서열 s에 대한 토큰 t의 가중치; W_s는 대상 서열 s의 길이에 대한 표준화 인수이다.

상위 k 서열의 집합, $K = \{ s_1, s_2, s_3, \ldots, s_k\}$가 유사도 측정에 의해 결정되어지면 질의 서열 q에 대한 위치정보 (c)의 가중치 w_c는 아래의 식과 같이 계산된다.

$$w_c = \sum_{i=1}^{k} sim(q, s_i) \cdot (c \wedge C_i)$$

C_i는 서열 $s_i \in K$가 가지는 모든 세포내 위치 정보의 집합이다. 각각의 위치정보에 대한 가중치를 계산한 후, 주어진 임계값(Threshold) 이상의 가중치를 가지는 위치정보가 단백질의 최종 세포내 위치로서 결정되어진다.

2.4 데이터 집합 (Data Set)

세포내 위치 예측 효율과 kNN 문서범주기의 최적화를 위하여 본 연구에서는 두 개의 데이터 집합을 실험에 사용하였다. 첫번째 데이터의 집합은 CELL_LOC (Eisenhaber and Bork, 1998)의 위치정보 분류 데이터를 이용하여 구축하였다. CELL_LOC에서 제공하는 위치정보 데이터는 SWISS-PROT의 단백질 'Access number'와 위치정보만으로 구성되어져 있어 CELL_LOC의 위치정보와 SWISS-PROT 데이터베이스 (Bairoch and APweiler, 2000) release 39.0의 서열 및 세포내 위치정보를 재구성하여 첫번째 데이터 집합을 구성하였다(데이터집합1). 데이터 집합 1은 전체 97,560 개의 서열 엔트리로 구성되어 있으며, 서열정보는 13개의 세포내 지점들을 위치정보로 분류하여 구성하였다(표 1). 이 데이터 집합에서 임의로 10,798개의 서열을 분할하여 이를 시험문서집합(Test Set)으로 삼고, 나머지 서열들을 훈련문서집합(Training Set)으로 구분하였다.

표 1. CELL_LOC정보로부터 구축한 데이터 집합 1에서 사용된 13개의 위치정보.

Symbol	Description
intracellular	Proteins inside the cell
cytoplasmic	Cytoplasm
nuclear	Nucleus
endop_r_golgi	Endoplasmic Reticulum or Gogi Apparatus
mitochondrion	Mitochondrial proteins
chloroplast	Chloroplast
DNAb	DNA-binding proteins
RNAb	RNA-binding proteins
extracellular	Proteins secreted to outside the cell
Membrane	Membrane-floating proteins
Transmem	Trans-membrane proteins
Viral	Proteins of virus origin
hypothetical	Still undefined

두 번째 데이터 집합은 SWISS-PROT 데이터 베이스 release 40.0으로부터 위치정보와 서열 정보 데이터를 직접 추출하여 구성하였다(데이터집합2). 이 데이터집합의 분류체계는 12개의 위치정보로 구성되어 있으며, SWISS-PROT 배포용 데이터의 CC 필드 (Field)에 포함되어 있는 위치정보에 대한 기술부분("-!- SUBCELLULAR LOCATION")으로부터 세포내 위치를 표현하는 주요 키워드를 이

용하여 각 서열을 해당 위치로 분류하였다 (Park and Kanehisa, 2003). 데이터 집합 2는 전체 51,885 엔트리를 포함하고 있으며, 전체 데이터를 시험집합 8,645개와 훈련집합 43,240개로 분할하였다.

표 2. 데이터 집합 2를 구축하기 위하여 SWISS-PROT 데이터베이스로부터 추출하여 분류한 12개의 위치정보 및 위치정보 추출에 사용된 키워드 (Park and Kanehisa, 2003).

Symbol	Keywords
Chloroplast	Chloroplast
Cytoplasmic	Cytoplasmic
Cytoskeleton	Cytoskeleton, Filament, Microtubule
endoplasmic reticulem	Endoplasmic Reticulum
Extracellular	Extracellular, Secreted
golgi apparatus	Golgi
Lysosomal	Lysosomal
Mitochondrial	Mitochondrial
Nuclear	Nuclear
Peroxisomal	Peroxisomal, Microsomes, Gloxysomal, Glycosomal
plasma membrane	Integral membrane
Vacuolar	Vacuolar, Vacuole

구축된 두개의 데이터 집합은 학습용 데이터와 시험용 데이터로 나누어 실험에 사용되었다. 데이터를 나누어 사용한 것은 두 가지의 목적이 있다. 첫째, 테스트 데이터의 서열에 대하여 학습용 데이터에 의해 예측된 위치정보와 사전 분류되어진 시험용 데이터의 위치정보와 비교하여 ProSLP의 예측 효율성을 측정하였다. 두 번째, 나누어진 데이터 집합을 이용하여 가장 좋은 예측 효율성을 위하여 kNN분류기의 k 값을 최적화하였다. 데이터 집합 1은 86,720 단백질 엔트리를 가진 학습용 데이터와 10,840 서열과 위치정보만을 담고 있는 시험용 데이터로 분류되었다. 데이터 집합 2는 43,240 엔트리의 학습용 데이터와 8,645 엔트리의 시험용 데이터로 분류되어 사용되었다.

2.5 성능 측정 (Performance Measure)

세포내 위치예측의 효율성을 측정하기 위하여 기본적인 효율성 측정도구로써 정확도(Precision)와 재현율(Recall)의 표준정의를 사용하였다. 정확도(p)와 재현율(r)은 다음과 같이 정의된다.

$$p = \frac{\text{locations relevant and retrieved}}{\text{locations retrieved}} = \frac{tp}{tp + fp}$$

$$r = \frac{\text{locations relevant and retrieved}}{\text{locations relavant}} = \frac{tp}{tp + fn}$$

tp (true positive)는 전문가에 의해 부여된 범주와 기계에 의해 부여된 범주가 일치하는 것이다. *fp* (false positive)는 전문가는 부여하지 않은 반면 기계는 부여한 범주이다. *fn* (false negative)은 전문가는 부여하였으나 기계는 부여하지 않은 범주를 말한다.

정확도와 재현율에 따라 문서범주화 분야에서의 많은 연구자들이 효율성 측정 도구로써 F_1 지표 (F_1 measure)를 사용해 왔다. 정확도와 재현율을 사용한 F_1 지표의 정의는 다음과 같다.

$$F1 = \frac{2pr}{P + r}$$

정확도와 재현율이 같아질 때의 F_1 지표를 정확도와 재현율의 break-even point (BeP)라고 한다. BeP는 어떤 지점에서든 F_1 값보다는 같거나 작다. BeP는 다른 범주기들과의 효율성을 비교하기 위하여 가장 일반적으로 사용된다 (Yang, 1999; Sebastiani, 2002). 본 연구에서도 ProSLP의 세포내 위치예측 효율성 지표로서 정확도, 재현율, 그리고 BeP를 사용하였다. 단 BeP 지표의 측정이 불가능한 경우에서는 F_1 지표를 제시하였다. 전체 데이터에 대한 세포내 위치정보에 대한 정확도와 재현율의 평균을 구하고 성능을 평가하기 위하여 micro-averaged break-even point를 사용하였다.

3. 결과

k-Nearest Neighbor 범주기에 있어 범주화 효율성을 향상시키기 위한 가장 중요한 단계는 데이터 집합에 대한 최고의 성능을 줄 수 있는 *k* 값을 결정하는 단계이다 (Yang, 1994; Sebastiani, 2002). 그림 1은 ProSLP 시스템에서 변화되는 *k* 값에 따른 Recall-Precision Break-even Point (BeP)를 보여준다. *k* vs. BeP 그래프로부터 사용된 ProSLP 데이터 집합에 대하여 *k* 값의 변화에 대하여 예측 효율성은 크게 영향을 받지 않는 다는 것을 알 수 있다. 데이터 집합 1을 사용하여 실험한 결과인 그림 1의 a 그래프에서는 k 값이 증가할 수록 BeP 지표는 점점 감소하는 것을 알 수 있고 데이터 집합 2를 사용하여 실험한 결과인 b의 그래프는 k 값에 따른 변화가 거의 없다는 것을 보여준다.

실험에 의해 계산된 정확도와 재현율, 그리고 BeP 지표를 표 3과 4에 정리하였다. 데이터 집합 1을 사용한 실험에서는 그래프에서 알 수 있듯이 k=1에서 가장 효율성이 높게 나왔고 k 값이 증가하면 할 수록 효율성이 점점 줄어가는 것을 알 수 있다. 그러나 BeP는 0.987에서 0.998의 범위에서 변화하였다. 그래프는 점차 줄어드는 경향성을 보이긴 하지만 극히 적은 값의 차이라는 것을 확인 할 수 있다. 데이터 집합 2의 결과에서는 k 값에 따른 BeP의 변화가 거의 없다는 것을 알 수 있다.

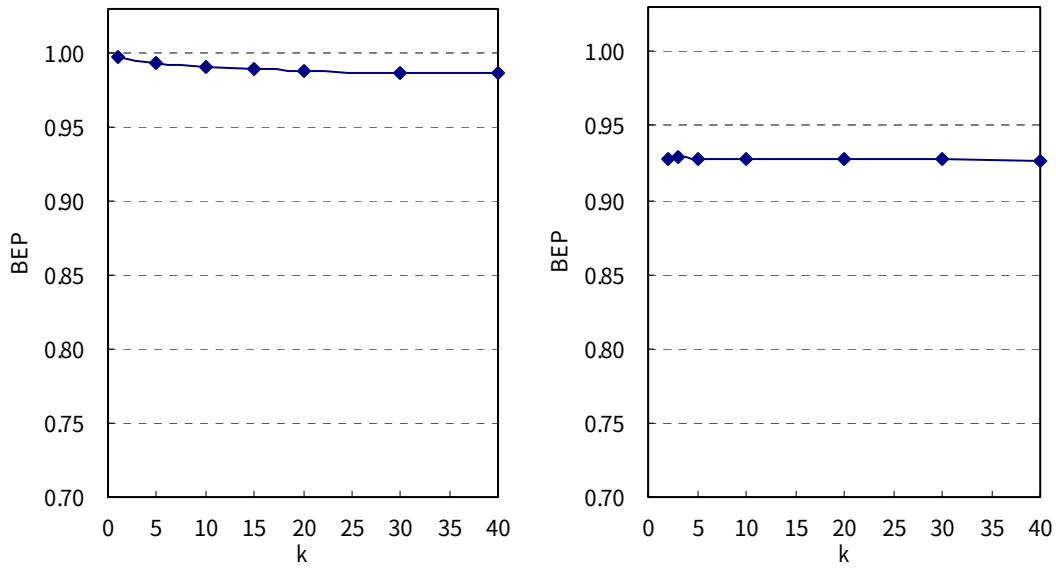

그림 1. 데이터 집합 1(그래프 a)과 데이터 집합 2(그래프 b)를 이용한 ProSLP 시스템의 예측 효율성 실험 결과.

ProSLP는 60%~85%의 정확도를 보이는 다른 시스템과 비교했을 때 93%이상의 굉장히 높은 정확도를 보여준다. 이것은 분류기의 학습에 사용된 데이터의 집합의 크기에 기인한 것으로 판단된다. 다른 시스템은 수백 또는 수천개의 데이터에 의해 학습된 반면 ProSLP는 86,720 및 43,240 개의 데이터 엔트리에 의해 학습되어졌다. 이것은 ProSLP에서 입력된 질의 서열에 대하여 유사 단백질 서열을 발견하고 찾을 가능성 다른 시스템에 비하여 훨씬 높다는 것을 의미한다.

현재 ProSLP는 가장 좋은 성능을 보여주는 k 값인 1과 3이 아닌 10으로 구성되어 있다. 가장 좋은 성능을 보여주는 k 값과 10일 때의 값의 차이가 크지 않고 또한 k=1에서의 가능한 에러를 k 값을 늘임으로써 위치정보의 예측을 위한 대상범위를 넓히고자 함이다.

표 3. 데이터 집합 1을 이용한 k값 변화에 따른 세포내 위치예측 효율성 측정 결과. 데이터 집합 1은 86,720개로 구성된 학습용 데이터와 10,840의 서열을 가진 테스트용 데이터로 구성되었다.

k	Precision	Recall	BeP[a]
1	0.997	0.999	0.998[b]
5	0.994	0.994	0.994
10	0.991	0.991	0.991
15	0.989	0.989	0.989
20	0.988	0.988	0.988
30	0.987	0.987	0.987
40	0.987	0.987	0.987

[a] Micro-averaged break-even points

[b] Micro-averaged F_1 measure

표 4. 데이터 집합 2을 이용한 k값 변화에 따른 세포내 위치예측 효율성 측정 결과. 데이터 집합 2는 43,240개로 구성된 학습용 데이터와 8,645의 서열을 가진 테스트용 데이터로 구성되었다.

k	Precision	Recall	BeP[a]
1	0.9292	0.9302	0.9297[b]
2	0.9282	0.9282	0.9282
3	0.9292	0.9292	0.9292
5	0.9282	0.9282	0.9282
10	0.9279	0.9279	0.9279
20	0.9272	0.9272	0.9272
30	0.9271	0.9271	0.9271
40	0.9269	0.9269	0.9269

[a] Micro-averaged break-even points

[b] Micro-averaged F_1 measure

ProSLP에서 관심을 가지는 또 다른 요소는 예측 속도이다. 그림 2에 쿼리 서열에 위치정보의 예측 시간을 도식화하였다. 48개의 서열을 무작위로 추출하였고 각각 질의 서열에 대하여 세포내 위치 예측 시간을 측정하였다. 질의 서열의 평균길이는 262개의 아미노산으로 이루어져 있고, 평균 예측 시간은 2.28초로 측정되어졌다.

예측 시간은 서열의 길이에 정비례한다는 것을 그림 2를 통하여 알 수가 있다. 이것은 ProSLP 시스템에 적용된 자질 추출방법의 특징이다. 길이 *l*의 서열에 대한 펜타그램(Penta-gram)의 숫자는 *l*-5+1이다. 문서 범주화에서 자질 집합에서 어떤 기준을 가지고 제한된 자질들을 서열소스로부터 추출하는 과정- 이 과정을 자질 선택 (Feature Selection)이라고 한다-은 분류의 효율성과 속도에 많은 향상을 이끌어왔다 (Yang, 1999; Sebastiani, 2002). 현재 ProSLP는 펜타그램(penta-gram)으로 고정되어 있지만 n-그램(gram) 자질에 대한 선택방법이 적용된다면 좀 더 낳은 예측 속도 향상을 기대할 수 있을 것이다.

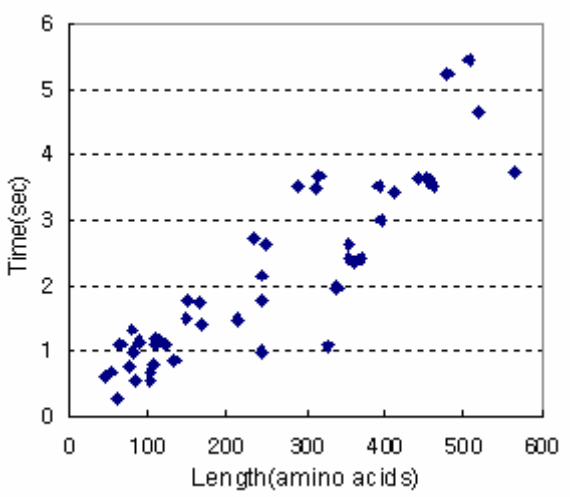

그림 2. 단백질 서열질의의 길이에 따른 위치 정보 예측 속도. 48개의 질의 서열을 사용하였고, 각 서열의 아미노산 길이는 53~600이며 평균길이는 262개이다.

4. 토의

포스트 게놈 시대로 들어섬에 따라 유전자 서열정보로부터 새로운 유전 생성물을 이끌어 내는 것이 중요해지고 있다. 단백질의 세포내 위치를 결정하는 것이 가능해진다면 단백질의 생물학적 기능 또는 구조적 기능에 대한 중요한 단서를 제공하게 될 것이다. 아미노산 조성은 세부적인 단계에서 다른 세포내 위치들의 단백질들을 구분하기 위한 좋은 정보들을 포함하고 있다는 것이 알려져 왔다. 본 연구에서는 단백질의 세포내 위치에 대한 예측 시스템이 단백질 서열에 대한 n-그램(gram) 자질 추출 방법과 kNN 분류기를 이용하여 구축가능하며, 더불어 매우 높은 예측 성능을 보여 줄 수 있다는 것을 알 수 있었다.

ProSLP는 가능한 단백질의 세포내 위치를 예측함에 있어 높은 정확도를 보여준다. 나아가 예측 속도는 낡은 컴퓨터 방식에서조차 매우 실용적인 속도를 보인다.

ProSLP의 세포내 위치예측 서비스 및 시스템에 사용된 데이터 집합은 http://proslp.kisti.re.kr에서 제공받을 수 있다. 또한 ProSLP의 웹페이지는 세포내 위치 예측에 있어 메타 검색기의 기능까지 포함하고 있어 ProSLP 자체의 예측 결과 뿐 아니라 다른 예측 사이트의 결과까지 한번에 보고 비교하는 것이 가능하다.

ProSLP 시스템이 생물학자와 생물정보학자의 단백질 관련 연구에 있어 많은 도움을 줄 수 있게 되기를 희망한다.

5. 참고 문헌

[1] Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TreEMBL in 2000. *Nucleic Acids Res.*, **28**, 45-48

[2] Chou, K. C. and Elrod, D. W. (1999) Protein subcellular locations prediction. *Protein Engineering*, **12**, 107-118

[3] Eisenhaber, F. and Bork, P. (1998) Evaluation of human-redable annotation in biomolecular sequence database with biological rule libraries. *bioinformatics*, **15**, 528-535

[4] Emanuelsson, O., Nielsen, H., Brunak, S., and Heijne, G. V. (2000) Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence. *J. Mol. Biol.*, **300**, 1005-1016

[5] Feng, Z. P. (2002) An overview on predicting the subcellular location of protein. *In silico Biol.*, **2**, 291-303

[6] Gardy, J. L., Spencer, C., Wang, K., Ester, M., Tusnady, G. E., Simon, I., Hua, S., deFays, K., Lambert, c., Nakai, K., and Brinkman, F. S. L. (2003) PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res.*, **31**, 3613-3617

[7] Park, K. J. and Kanehisa, M. (2003) Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics*, **19**, 1656-1663

[8] Nakai, k. and Horton, P. (1999) PSORT: a program for detecting the sorting signals of proteins and predicting their subcellular localization. *Trends Biochem. Sci.*, **24**, 34-35

[9] Reinhardt, A. and Hubbard, T. (1998) Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.*, **26**, 2230-2236

[10] Sebastiani, F. (2002) Machine Learning in Automated Text categorization. *ACM computing Surveys*, **34**, 1-47

van Rijsbergen, C. (1979) Information Retrieval. Butterworths, London, 1979

[11] Yuan, Z. (1999) Prediction of protein subcellular locations using Markov chain models. *FEBS Letters*, **451**, 23-26

[12] Yang, Y. (1994) Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval. *Proceeding of the 17th Annual International ACM/SIGIR conferences on Research and Development in Information Retrieval*, 13-22

[13] Yang, Y (1999) An Evaluation of Statistical Approaches to Text Categorization. *Journal of Information Retrieval*, **1**, 67-88