단백질의 세포내 위치를 예측하기 위한 외부정보의 성능 비교

(Comparison of External Information Performance in Predicting Subcellular Localization of Proteins)

지 상 문 †

(Sang-Mun Chi)

요 약 단백질의 세포내 위치와 단백질의 기능은 연관성이 크므로, 단백질의 세포내 위치 예측을 통해서 그 기능에 대한 정보를 얻을 수 있다. 예측 정확도를 높이기 위해서 아미노산 서열 정보이외의 외부 정보들을 효과적으로 이용하려는 연구가 활발하다. 본 논문에서는 아미노산 서열 유사성, 단백질 프로파일, 유전자 온톨로지, 모티프, 문헌 정보에 내재된 세포내 위치 예측 능력을 비교한다. 단백질간의 서열 유사성이 80% 이하인 PLOC 자료를 사용한 실험에서는 서열 유사성과 유전자 온톨로지를 이용하는 방법이 효과적이며, 94.8%의 예측정확도를 얻었다. 단백질 서열간의 유사성이 30% 이하로서 단백질간의 서열 유사성이 작은 BaCelLo IDS 자료는 유전자 온톨로지를 사용하는 것이 효과적이었고, 동물은 93.2%, 곰팡이는 86.6%의 예측정확도로 크게 향상된 성능을 얻었다.

키워드 : 단백질의 세포내 위치 예측, 아미노산 서열 유사성, 단백질 프로파일, 유전자 온톨로지

Abstract Since protein subcellular location and biological function are highly correlated, the prediction of protein subcellular localization can provide information about the function of a protein. In order to enhance the prediction performance, external information other than amino acids sequence information is actively exploited in many researches. This paper compares the prediction capabilities resided in amino acid sequence similarity, protein profile, gene ontology, motif, and textual information. In the experiments using PLOC dataset which has proteins less than 80% sequence similarity, sequence similarity information and gene ontology are effective information, achieving a classification accuracy of 94.8%. In the experiments using BaCelLo IDS dataset with low sequence similarity less than 30%, using gene ontology gives the best prediction accuracies, 93.2% for animals and 86.6% for fungi.

Key words : Protein Subcellular Localization Prediction, Amino Acid Sequence Similarity, Protein Profile, Gene Ontology

1. 서 론

단백질은 세포와 생물의 구조와 기능에 중심적인 역할을 한다. 전통적인 생화학실험을 사용하지 않고, 컴퓨터과학의 여러 기술을 적용하여 단백질의 구조와 기능을 밝히는 연구가 활발하다. 이는 단백질이 크기, 모양, 전하, 수소결합 능력, 소수성, 화학 반응성과 같은 물리화학적 특성이 제각기 다른 20 종류의 아미노산으로 구성된 선형중합체이고 그 구조와 기능이 아미노산 순서에 의해서 결정되어지므로, 아미노산 서열로부터 구조와 기능을 예측하는 것이 원리적으로는 가능하기 때문이다. 본 논문에서는 패턴분류 방법을 사용하여 아미노산 서열로부터 단백질의 세포내 위치를 예측한다.

세포는 기능적으로 구분된 세포 소기관으로 구성되고, 특정 기능을 수행하기 위해 협력하는 여러 단백질들은 같은 세포내 위치에 존재하므로 단백질의 세포내 위치

· 이 논문은 2010학년도 경성대학교 학술연구비지원에 의하여 연구되었음

† 종신회원 : 경성대학교 컴퓨터학부 교수
smchiks@ks.ac.kr

논문접수 : 2010년 5월 11일
심사완료 : 2010년 9월 3일

정보과학회논문지: 소프트웨어 및 응용 제37권 제11호(2010.11)

는 단백질의 기능에 대한 정보를 준다. 따라서, 단백질의 세포내 위치에 대한 정보는 프로테오믹스, 약물 표적 발견, 시스템 생물학에 이용된다. 예를 들면, 세포외 공간과 세포 표면에 위치하는 단백질은 약물표적으로서 약학적 응용의 대상이 된다. 이러한 단백질은 세포 표면의 호르몬 수용체로서 세포막에 존재하면서 신호전달에 관여하는 부분과 세포 밖 부분에서 리간드와 결합하는 부분으로 구성되는데, 세포 밖 부분만을 가진 단백질을 만들어 세포에 삽입하여 신호 전달을 감소시켜 관절염의 중요한 치료방법으로 사용하는 예가 있다[1].

단백질의 세포내 위치를 계산적으로 예측하는 방법을 분류하면, 첫 번째로 단백질의 앞부분 서열인 아미노 말단 서열만을 사용하는 방법과 아미노 말단 서열과 더불어 추가적 정보를 함께 사용하는 방법이 있다[2-8]. 이 방법은 세포안에서 단백질의 운반과 위치를 결정하는 내재적인 신호가 단백질 서열의 아미노 말단서열에 주로 존재한다는 사실을 이용한다. 두 번째 방법은 아미노산 서열에 포함된 아미노산 조성을 이용하는 방법[3-13]으로, 같은 세포내 위치에 존재하는 단백질들은 유사한 아미노산의 조성과 기능을 가지며, 신호서열과 막통과 단백질에서 막을 관통하는 부분의 아미노산들과 수용성 단백질의 내부에는 소수성이 큰 아미노산들로 구성된다는 사실을 이용한다[14-17].

세 번째로는 외부에 구축된 데이터베이스에서 단백질과 연관된 정보를 추출하여 외부정보로 이용하는 방법들이 있다. 단백질 프로파일의 아미노산 조성을 사용하는 방법은 주어진 단백질 하나에서 정보를 추출하지 않고, 서열상으로 유사한 단백질들을 찾아서 이들 서열들의 평균으로 아미노산 조성을 추정한다[4,18]. 문헌정보를 이용하는 방법은 각각의 단백질 서열을 설명하는 문서에 등장하는 단어에서 단백질의 세포내 위치 정보를 표현하는 중요 용어를 선별하여 예측에 사용한다[19-21]. 유전자 온톨로지 정보를 사용하는 방법은 단백질의 분자적 기능, 생물학적 과정, 세포 요소와 관련된 개념을 특징화한 용어로서 단백질을 표현한 데이터베이스를 세포내 위치 예측에 이용한다[8,12,22,23]. 이러한 아미노산 서열이외의 외부정보를 이용하는 방법들은 계산량이 많고, 이미 알려진 문서 정보나 단백질에 대한 정보를 활용하는 방법이므로 새로운 단백질 서열이나 알려진 정보가 적은 경우에는 효과적이지 않다. 하지만, 단백질에 대한 정보가 증가하고, 서열정보와는 다른 관점에서 단백질을 표현하므로 서열상의 유사성이 작더라도 세포내 위치를 간접적으로 포함하고 있는 정보를 활용할 수 있다. 본 논문에서는 외부정보가 예측 정확도에 미치는 영향을 서열의 중복도가 각기 다른 단백질 자료에 대하여 비교하여, 효과적인 외부정보의 사용을 제안한다.

논문의 구성은 다음과 같다. 2장에서는 단백질의 세포내 위치를 예측하는 방법을 살펴본다. 특히, 3장의 실험에 사용되는 방법을 구체적으로 알아본다. 3장에서는 단백질 서열정보 이외의 외부정보가 단백질의 세포내 위치 예측에 미치는 효과를 비교하고, 4장에서 결론을 맺는다.

2. 단백질의 세포내 위치예측을 위한 정보

2.1 아미노 말단 서열

단백질 서열의 처음 부분인 아미노 말단 서열 정보가 중요한 이유는 분비 단백질임을 알 수 있게 해주는 신호서열이 아미노 말단에 존재하고 서열상의 유사성이 크지는 않으나, 전하의 분포와 아미노산의 소수성이 보존된 3개 영역이 존재하여 다른 세포내 위치의 단백질과 구분되기 때문이다. 또한, 엽록체, 미토콘드리아로 향하는 표적 펩티드도 아미노 말단에 존재하며, 엽록체로의 표적 펩티드는 세린이 풍부하고 산성 잔기는 없으며, 미토콘드리아로의 표적 펩티드는 아르기닌, 세린, 알라닌은 풍부하나 음전하 잔기는 희귀하다는 아미노산 분포의 구별되는 특성을 가지고 있기 때문이다[15-17]. 아미노 말단의 정보를 세포내 위치 예측에 이용하기 위해서, 아미노 말단의 50-100개의 아미노산을 직접 사용하거나, 아미노산 조성을 구한다. 하지만, 현재의 기술로는 유전체에서 단백질을 코딩하는 첫 부분인 시작 코돈을 70% 이하의 정확도로 예측하고, 섬유아세포 성장 인자(fibroblast growth factors)와 같은 분비단백질은 전형적인 분비경로를 따르지 않아 아미노 말단 신호서열이 없으며, 핵 위치 신호(nuclear location signal)는 서열전반에 걸쳐 나타날 수 있으므로, 아미노 말단 서열만을 이용하는 방법은 불완전하다. 따라서 아미노 말단 서열 정보를 다른 정보와 함께 사용하는 방법이 널리 쓰인다[3-8].

본 논문의 3장 실험 방법 중에서 아미노 말단 서열을 사용하는 방법들이 있는데, WoLF PSORT[6]는 신호서열과 표적펩티드를 나타내는 분류 신호를 나타내기 위해 아미노산 서열을 이용하고 BaCelLo[4], MultiLoc[5], SherLoc[7], MultiLoc2[8], PDSS[13], LOCSVMPSI[18], SherLoc2[24]는 아미노 말단의 아미노산 조성을 다른 정보와 함께 사용한다.

2.2 아미노산 조성

아미노산 조성은 단백질을 구성하는 아미노산들의 상대적 빈도이다. 아미노산 조성은 세포내 위치와 연관성이 높다. 예를 들면, 신호펩티드와 막통과 단백질에서 막을 통과하는 부분은 소수성이 높은 아미노산들로 구성되고, 세포내 위치마다의 고유의 PH나 이온의 세기 등의 생화학적 환경의 영향으로 각기 다른 아미노산 조성을 갖고, 단백질 내부에 매몰되는 아미노산 서열보다는 표

면부위에 존재하는 아미노산들이 더욱 세포내 위치에 종속적인 고유한 조성을 갖는다[14-17]. 하지만, 아미노산 조성은 단백질의 특성을 표현하는 중요 정보인 아미노산 순서를 기본적으로는 표현할 수 없다. 따라서 서열상의 아미노산 순서를 포함하기 위하여 여러 개의 아미노산 연속체를 하나의 단위로 사용하거나[9,10], 전체 아미노산 서열을 나누어 여러 부분으로 나누고, 각 부분구획에서 아미노산 조성을 구하는 방법이 있다[3-5,11,13]

본 논문의 3장 실험에서 많은 방법들이 아미노산 조성을 사용한다. PLOC은 아미노산, 아미노산 짝, 그리고 g-갭 아미노산 짝(두 개의 아미노산 사이에 존재하는 각각 $g=1,2,3$개의 아미노산의 종류는 고려하지 않는다.)의 조성을 사용하여 5개의 패턴분류기를 만든 후에, 이들의 예측결과를 다수결 투표하여 최종결과를 얻는다[9]. Yang은 $k(1 \le k \le 5)$개의 연속된 아미노산으로 구성된 k-튜플 조성을 사용하는 각각의 패턴분류기를 만들고 이들의 다수결 투표 결과로서 예측한다[10]. 그러나, k-튜플의 종류는 20^k이므로 $k \ge 4$인 경우는 k-튜플의 종류가 크게 증가하여 패턴분류기의 학습과 평가 시간이 크게 증가한다. Chou는 아미노산 조성이 아니라 일정 간격의 아미노산들간의 소수성과 곁가지 잔기의 무게의 변화를 이용한다[12]. BaCelLo[4], MultiLoc[5], SherLoc[7], MultiLoc2[8], LOCSVMPSI[18], SherLoc2[24]는 아미노산 순서 정보를 포함하기 위하여 아미노산 서열을 나누어 각각의 부분구획에서 아미노산 조성을 구한다.

본 논문의 3장 실험에서는 외부 정보를 구할 수 없는 단백질은 아미노산 조성을 PDSS로 구한다[13]. PDSS는 전체서열을 부분서열로 나누고, 부분서열 각각에서 아미노산의 조성을 추정한다. 이렇게 서열을 부분 서열로 나누는 방법은 일반적으로 널리 사용되고 있는데 단백질 서열의 위치에 따라서 변하는 아미노산 조성이 단백질의 특성을 나타내는 중요한 정보이기 때문이다. PDSS는 부분서열끼리 그림 1과 같이 겹치게 하여 전체 서열을 나누는 과정에서 발생하는 인위적인 아미노산 조성의 불연속을 완화한다.

본 논문에서는 그림 1의 r-번째 부분 서열이 L개의 아미노산으로 구성된다면 $R=L/3$로 하여 부분 서열간에 2/3가 겹치게 하였다.

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

각 부분 서열에서 단백질 프로파일 방법과 비슷하게 아미노산 조성을 구한다. 서열 위치 n의 아미노산 $x[n]$에 연관된 20개 아미노산 $a_k\,(1 \le k \le 20)$의 빈도를 구하는 과정에서 단백질 프로파일은 유사서열을 탐색하여 프로파일을 만들지만, PDSS는 유사서열 탐색과정 없이 다음 식으로 구한다.

$$\exp(-\sigma\, d^2(x[n], a_k)). \quad (2)$$

여기서, $d(a,b)$는 아미노산 a와 b사이의 거리로서 아미노산간의 진화적인 연관성을 나타내는 치환행렬 BLOSUM50[25]을 사용하여 정의되었다[13]. 부분서열 $x[n]\,(0 \le n < L)$에서 아미노산 a_k의 조성은

$$\sum_{n=0}^{L-1} \exp(-\sigma\, d^2(x[n], a_k)) \quad (3)$$

으로 구한 값을, 모든 a_k에 대한 전체합으로 나누어 정규화한 값을 사용한다. 파라미터 σ는 부분서열의 길이가 짧을수록 부분 서열내의 아미노산 개수가 적어 안정적인 조성 추정이 어려우므로 유사한 아미노산으로부터의 기여도를 크게 반영하기 위하여 작은 값을 선택 한다. PDSS에서는 최적인 σ를 학습자료를 사용하여 탐색하였지만, 본 논문에서는 계산량 감소를 위해 휴리스틱하게 $\sigma=\log_{15}L$을 사용하였다.

아미노산 서열에서 세 종류의 정보를 만들고, 이들을 각각 사용하여 예측하고, 예측결과를 다수결 투표하여 최종결과를 얻는다. 첫 번째로 $S=2$를 사용하여 전체 서열의 길이 N에서 부분 서열의 길이 $L=N/S$를 얻고, 각 부분 서열에서 아미노산 조성을 구하고, 아미노산 45개로 이루어진 아미노 말단 서열에 대해서 부분 서열의 길이 $L=45/S$를 얻고, 각 부분 서열에서 아미노산 조성을 구한다. 마찬가지 방법으로 $S=3,4$일 때에도 PDSS로 아미노산 조성을 구하여 패턴 분류기의 입력으로 사용한다.

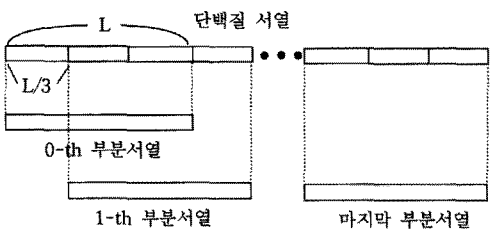

그림 1 전체서열을 부분서열로 분할하는 방법

2.3 서열 유사성

아미노산 서열이 유사하면 그 단백질의 기능과 구조가 유사하기 때문에, 생물정보학에서 미지의 단백질에 대한 정보를 얻기 위해서 가장 많이 사용하는 방법은 서열 유사성이 높은 이미 정보가 알려진 단백질을 찾는 것이다. BLAST(Basic Local Alignment Search Tool)는 유사서열 탐색에 가장 많이 사용되는 방법이다[26]. BLAST 탐색을 통해서 세포내 위치를 알고 있는 가장 가까운 유사성을 가진 단백질을 찾아서 세포내 위치를 찾는 방법이 서열 유사성이 높은 단백질을 찾을 수 있는 경우에는 기계학습보다 예측성능이 높은데, BLAST

E-값(expectation value)이 e^{-80}이하인 단백질들에 대해서는 90%의 정확도를 얻을 수 있고, 서열간의 동일함이 30% 이상이면 기계학습보다 BLAST 탐색에 의한 세포내 위치예측의 정확도가 높다[27,28].

본 논문의 3장 실험에 BLAST로 표시한 방법은 BLAST 탐색으로 세포내 위치를 찾는다. 이 방법은 일반화 능력이 작아서 유사하지 않은 단백질에 대한 예측은 부정확하다. 따라서, 본 연구에서는 BLAST 탐색으로 E-값이 0.001이하인 서열유사성이 높은 단백질을 찾을 수 있는 경우에만 BLAST 탐색으로 예측하였고, 그 밖의 서열에는 2.2절에서 설명한 PDSS를 이용하여 세포내 위치를 예측하였다.

2.4 단백질 프로파일

특정 단백질에 대한 프로파일을 만들기 위해서는, 먼저 서열상 유사한 단백질들을 데이터베이스에서 탐색하여 특정 단백질과 정합한 후에, 아미노산 서열상의 위치별로 아미노산 분포를 만든다. 이러한 아미노산 서열의 위치에 의존하는 점수행렬(position-specific scoring matrix: PSSM)을 사용하여 주어진 단백질과 서열상으로 유사한 단백질들의 정보를 포함할 수 있다. 단백질 프로파일은 단백질의 세포내 위치와 관련된 정보를 직접적으로 사용하지는 않는다. 따라서 세포내 위치정보와 관련된 정보를 사용할 수 없는 단백질의 경우에도 크게 예측성능이 저하되지 않는 장점이 있으나, 세포내 위치와 관련된 외부정보가 존재하는 단백질에 대해서는 직접적으로 외부정보를 사용하는 것보다 높은 정확도를 얻기 어렵다.

본 논문의 3장에서 LOCSVMPSI는 PSI-BLAST[26]를 사용하는데, 반복회수를 지정하는 파라미터 $j=3$, 주어진 E 값 이하의 서열만들 사용하도록 하는 파라미터 $h=0.001$로 만들어진 PSSM을 사용하여 단백질 프로파일을 만든다. 한편, BaCelLo[4]는 두 종류의 아미노산 조성을 각각 단백질 서열과 단백질 프로파일로부터 구한다. 단백질 프로파일은 BLAST를 이용하여 주어진 단백질과 E-값이 10^{-4}이하의 유사성을 갖는 Swiss-Prot 48[29] 내의 단백질들과의 정합을 통해서 각 아미노산의 빈도를 질의된 단백질의 각 위치별로 구한다. 이 방법은 단백질이 세포내 위치로 올바르게 분비되도록 역할을 수행하는 펩티드 정보를 포함하기 위하여 여러 경우를 고려하여, 최종적으로 아미노산 조성을 전체서열, 아미노 말단에서 길이가 20, 40, 60인 세부분과 단백질 서열의 끝부분인 카르복시 말단에서 길이가 20, 50, 100인 세부분에서 각각 구하여 이용하였다. 본 논문에서는 단백질 프로파일을 사용하는 방법을 구현하여 3장의 비교실험에 사용하였다. 이 방법을 Profile이라 명명하였고, $j=2$, $h=0.001$인 PSI-BLAST를 사용하여 대용량 단백질 데이터베이스인 UniRef90[29]을 탐색하여 PSSM을 만들었다. 전체서열을 길이가 같은 두 부분서열로 나누고 각각의 부분서열에 속하는 PSSM에 나타나는 아미노산 빈도를 계산하여 아미노산 조성을 추정하였다.

2.5 문헌 정보와 유전자 온톨로지

문헌 정보를 이용하는 방법은 Swiss-Prot 키워드를 이용하는 방법[19]과 PubMed 초록을 이용하는 방법[7, 20,21,24]이 있다. 이들 방법은 단백질을 설명하는 문서에 나타나는 용어 중에서 빈도가 너무 적은 것과 너무 많은 것을 제외하는 등의 세포내 위치에 대한 정보를 표현하는 중요 용어를 선별하는 문서처리 방법을 사용한다.

유전자 온톨로지(Gene Ontology)는 유전자의 분자적 기능, 생물학적 과정, 세포 요소와 관련된 개념을 공통된 어휘로 나타내어 대량의 생물 관련 자료를 계산적 관점에서 처리하려고 만들었다[30]. 단백질의 유전자 온톨로지를 얻기 위한 방법으로 주어진 단백질을 질의 서열로 사용하여 BLAST 탐색으로 E값이 10^{-9}이하의 5개의 단백질을 Swiss-Prot에서 찾고, 찾아진 단백질의 유전자 온톨로지를 사용하는 방법이 있다[22]. 이 방법은 유전자 온톨로지간의 유사도를 유전자에 대한 설명의 상세함이 증가하는 깊이로 정의하고, 가장 유사한 단백질의 세포내 위치로 주어진 단백질의 세포내 위치를 결정한다. [23]에서는 E값을 $10^{-9}, 10^{-8}, \ldots, 10^{-1}$로 차례로 증가시키면서 BLAST 탐색에서 먼저 찾아지는 정보가 알려진 단백질의 유전자 온톨로지를 이용하는 방법과 예측하려고 주어진 단백질을 미지의 단백질로 가정하지 않고, 그 단백질의 유전자 온톨로지를 직접적으로 사용한다. 이 방법은 전체 유전자 온톨로지를 사용하지 않고 조합적 최적화 방법을 사용하여 세포내 위치 예측에 효과적인 유전자 온톨로지 집합을 만들고 이들 온톨로지만을 세포내 위치 예측에 사용한다. 단백질의 유전자 온톨로지를 GOA 데이터베이스[31]에서 직접 찾는 방법[22,23]과 달리, InterPro[32]를 이용할 수 있다. InterPro에서는 단백질의 영역, 패밀리, 기능, 위치를 표현하는 예측모델과 특징(signature)을 입력된 아미노산 서열을 사용하여 탐색할 수 있는 도구를 제공한다.

본 논문의 비교 방법중 SherLoc은 문서정보로 PubMed 초록을 이용한다. SherLoc2는 또한 유전자 온톨로지를 사용하는데 Chou와 Mutiloc2와 같이 InterPro를 통하여 얻은 엔트리(단백질의 영역, 기능과 위치에 대한 정의)를 InterPro에서 제공하는 방법으로 유전자 온톨로지로 변환하여 패턴 분류기의 입력으로 사용한다. 3장의 실험에서 사용한 GOA는 GOA Uniport 81[31] 데이터베이스를 사용한다. 우선, GOA 데이터베이스에 존재하

는 Swiss-Prot 단백질의 목록을 추출한 후에, 이 목록에 해당하는 아미노산 서열을 Swiss-Prot 57에서 얻어서 아미노산 서열 자료를 만든다. 미지의 단백질을 질의 단백질로 사용하여 BLAST 탐색을 아미노산 서열 자료에 수행하면, 유사도 순서로 단백질 서열이 얻어진다. 질의 단백질이 아미노산 서열 자료에 이미 존재하는 경우에는 이것이 가장 유사한 서열로 탐색되고, 아닌 경우에도 서열상의 유사성은 단백질의 구조와 기능의 유사성과 같은 관련성이 있으므로 가장 유사한 단백질의 유전자 온톨로지를 주어진 단백질의 유전자 온톨로지로 사용한다.

3. 실험 및 분석

3.1 실험 자료, 성능 척도와 패턴 분류기

실험 자료는 아미노산 서열간의 유사도가 80% 이하인 단백질로 구성된 PLOC[9]과 동일한 아미노산이 30% 이하인 서열로 구성된 중복성이 작은 BaCelLo IDS[33]를 사용한다. PLOC 자료는 Swiss-Prot 39에서 추출한 7579개의 단백질 서열로서, 12개의 세포내 위치(chloroplast, cytoplasmic, cytoskeleton, endoplasmic reticulum, extracellular, golgi apparatus, lysosomal, mitochondrial, nuclear, peroxisomal, plasma membrane, vacuolar) 중의 하나에 위치한다. 균등한 개수의 5개 자료로 나누어져 있어서, 실험에서는 4개의 자료를 사용하여 단백질의 세포내 위치를 예측하기 위한 패턴 분류기를 구성하여 나머지 1개 자료에 대한 평가를 수행한다. 학습 자료와 평가 자료를 구성하는 각기 다른 5가지 경우에 대해 실험하여 평균적인 성능을 표시하였다. BaCelLo IDS 자료는 구상 단백질(globular protein)들로 4개의 세포내 위치(cytoplasmic, extracellular, mitochondrial, nuclear) 중의 하나에 단백질이 존재한다. 학습자료는 Swiss-Prot 48에서 동물 단백질은 2579개, 곰팡이 단백질은 1198개를 추출하였다. 평가자료는 Swiss-Prot 54에서 Swiss-Prot 48이외의 자료에서 학습자료와의 서열 동일성이 30% 이하인 것만 추출하였다. 같은 세포내 위치에서도 서열 동일성이 30% 이상인 것은 같은 군집에 속하도록 군집화하여, 동물 자료는 432개의 군집, 곰팡이 자료는 418개의 군집으로 구성된다.

단백질의 세포내 위치를 예측하는 방법의 성능은

$$\text{TA(Total Accuracy)} = \sum_{i=1}^{K} T_i/N,$$

$$\text{LA(Local Accuracy)} = \sum_{i=1}^{K} P_i/K$$

를 척도로 사용한다[9]. K는 세포내 위치의 개수이고, N은 PLOC 자료에서는 단백질의 총 개수이고 BaCelLo IDS 자료에서는 군집의 총 개수이다. T_i는 PLOC 자료에서는 세포내 위치 i에 존재하는 단백질 중에서 올바르게 예측된 개수이고, BaCelLo IDS 자료에서는 군집별로 올바르게 예측된 비율을 먼저 구하고, 이를 합한 값이다. $P_i = T_i/N_i$에서 N_i는 PLOC 자료에서는 세포내 위치 i에 존재하는 단백질의 개수이고 BaCelLo IDS 자료에서는 세포내 위치 i에 존재하는 군집의 개수이다. TA는 전체 단백질 중에서 정확히 예측된 비율이고, LA는 세포내 위치에 속하는 단백질의 수는 고려하지 않고, K개의 각각의 세포내 위치에서의 예측정확도를 평균한 값이다.

2장에서 살펴본 여러 방법으로 추출한 단백질에 대한 정보를 패턴 분류기에 입력하여 단백질의 세포내 위치를 예측한다. PLOC, Yang, PDSS, LOCSVMPSI, MultiLoc, SherLoc, BaCelLo, MultiLoc2, SherLoc2, Profile은 SVM(Support Vector Machines)을 패턴 분류기로 사용하였고, SVM에 이용되는 파라미터는 최적의 정확도를 주는 값을 탐색하여 사용되었다. 본 논문에서 실험한 PDSS와 Profile은 가우시안 커널 $\exp(-\gamma|u-v|^2)$을 사용하여 LIBSVM[34]으로 학습과 평가를 한다. 파라미터로 γ는 앞에서 SVM을 사용하는 방법들과 마찬가지로 그리드 탐색을 사용하여 예측방법과 실험자료에 따라 최적 값을 찾았다; $2^k(k=0,1,...,10)$ 에 대해서 최적인 k를 먼저 찾고, 2^{k-1}을 1.1배씩 증가시켜서 2^{k+1}보다 작은 값 사이에서 다시 최적인 값을 찾았다. LIBSVM의 파라미터 중에 C는 값에 따라 커다란 성능차이를 보이지 않으므로 계산량 감소를 위해 $C=1$로 고정하였고, 많은 단백질이 속하는 세포내 위치를 위주로 학습되지 않도록, 클래스 i를 가중하는 $wi = \sum_{j=1}^{K} N_j/N_i$로 하였다.

서열 유사성을 이용하는 방법에서는 BLAST 탐색을 통해서 가장 유사한 단백질의 세포내 위치로 예측하는 방법을 사용하고, Chou와 GOA는 전체 유전자 온톨로지의 개수를 차원으로 하는 벡터에서 특정 온톨로지가 존재하면 해당하는 요소를 1로 하고 존재하지 않으면 0으로 하는 벡터로 변환한 후에 벡터간의 유사도를 사용하여 가장 유사한 단백질의 세포내 위치로 예측하는 방법을 사용한다. 벡터간의 유사도는 벡터간의 내적을 각 벡터의 크기로 나눈 코사인 유사도를 사용한다. WoLF PSORT는 kNN(k nearest neighbor classifier)를 사용한다.

3.2 외부정보를 사용한 세포내 위치 예측 비교

여러 논문에서 제안된 외부정보들의 유용성을 알아보기 위해서, 동일한 자료에 대한 성능을 비교한다. 서열의 중복도가 다른 단백질로 구성된 자료를 사용하여 중복성의 변화에 따른 성능변화도 알아본다.

표 1은 PLOC 자료에 대해서 세포내 위치별로 올바르게 예측된 비율을 %로 나타내었다. 표 1에서 세포내 위치 ch는 chloroplast, cp는 cytoplasmic, cs는 cytoskeleton, er은 endoplasimc reticulum, ex는 extracellular go는 Gogi aparatus, ly는 lysosomal, mi는 mitochondrial, nu는 nuclear, pe는 peroxisomal, pl은 plasma membrane, va는 vacuolar를 나타낸다. 또한 P는 식물, A는 동물, F는 곰팡이 자료에 대한 실험을 나타낸다. PLOC, Yang과 PDSS는 아미노산 서열 정보만을 사용하기 때문에 주어진 단백질과 유사한 자료를 찾을 수 없거나 외부정보가 없어도 사용할 수 있고, 계산량이 적어서 수행속도가 빠른 장점이 있다. 이들 중에서 PDSS의 성능이 가장 높으므로, 이후의 외부정보를 이용하는 방법에서 외부정보를 찾을 수 없는 단백질 서열에 대해서는 PDSS를 적용하였다. PDSS에 의한 성능향상의 통계적 유의성을 조사하기 위해, p_1과 p_2를 각각 Yang과 PDSS가 정확하게 예측한 확률(TA와 같음) 이고, n_1과 n_2는 Yang과 PDSS의 평가자료 개수(두 경우 모두 7579개)이고, $p=(p_1*n_1+p_2*n2)/(n_1+n_2)$라 하자. 귀무가설 $H_0: p_1 \geq p_2$과 대립가설 $H_a: p_1 < p_2$에 대해서 검정통계량 $z=(p_1-p_2)/\sqrt{p\cdot(1-p)/[1/n_1+1/n_2]}$는 귀무가설 하에서 정규분포를 따른다. 표 1을 사용하여 계산하면 $z=-2.15$이고, P값($z<-2.15$인 확률)은 0.0158 이므로, 유의수준 5%로 귀무가설을 기각할 수 있다. 따라서 PDSS의 성능향상은 통계적 유의성이 있고, PLOC은 $z=-9.29$로서 PDSS의 성능향상은 더욱 통계적 유의성이 있다.

표 1의 LOCSVMPSI와 Profile은 외부정보로서 단백질 프로파일을 사용하는 방법이다. 단백질 프로파일에 기반한 방법이 서열정보만을 사용하는 PLOC, Yang보다 성능이 높다. 이는 단백질 데이터베이스에서 PSI-BLAST로 탐색한 유사한 서열들의 정보를 사용하여 주어진 단백질을 보다 일반화된 서열로 표현하기 때문이다. MultiLoc은 전체서열의 아미노산 조성, 아미노 말단의 아미노산 조성과 더불어 외부정보로서 신호앵커의 유무확률, 43개 서열 모티프의 유무 정보를 사용한다. 여기서, 서열 모티프는 짧은 길이의 특징적인 아미노산 서열을 의미한다. SherLoc은 MultiLoc과 함께 외부정보로서 PubMed 초록을 문헌정보로 사용하는 방법이고 Chou는 외부정보로서 유전자 온톨로지로 사용하고, 유전자 온톨로지를 얻을 수 없는 서열은 아미노산 서열에서 정보를 추출한다. 외부정보로서 문헌정보를 사용하는 SherLoc과 유전자 온톨로지를 사용하는 Chou가 성능을 크게 향상시켰다.

BLAST는 BLAST 탐색을 통해서 가장 가까운 유사성을 가진 단백질의 세포내 위치로 예측한다. TA와 LA가 PLOC, Yang, PDSS보다 높았고, 외부정보로 유사한 단백질 서열들을 사용하는 LOCSVMPSI, Profile 보다 높았고, 외부정보로 모티프를 사용하는 MultiLoc보다도 성능이 높았다. 특히, LA가 크게 향상되어 여러 세포내 위치에서 성능이 균등하다. PLOC 자료와 같이 서열간의 유사성이 비교적 높은 자료에 대해서는 기계학습 방법보다 BLAST 탐색을 이용하는 것이 유리하다. 비슷한 결과로, [4]에서는 NNPSL자료[35]에 대해서 BLAST 탐색으로 세포내 위치를 예측하여 기계학습기

표 1 PLOC 자료에 대한 여러 방법의 세포내 위치 예측 정확도(%)

		ch	cp	cs	er	ex	go	ly	mi	nu	pe	pl	va	TA	LA
PLOC		72.3	72.2	58.5	46.5	78.0	14.6	61.8	57.4	89.6	25.2	92.2	25.0	78.2	57.9
Yang		79.7	77.8	55.9	68.4	84.0	17.3	61.1	58.3	92.6	39.9	95.6	46.5	82.8	64.8
PDSS		78.2	79.9	65.0	67.5	93.7	46.8	53.8	77.4	87.0	35.2	93.2	50.0	84.1	69.0
LOCSVMPSI		76.5	76.4	60.0	61.4	89.7	46.8	62.4	68.2	91.5	41.6	94.7	40.7	83.5	67.5
MultiLoc	P	66	80	60	78	84	55	-	67	75	71	83	65	73.6	71.3
	A	-	75	65	81	83	51	77	69	78	74	81	-	76.0	73.6
	F	-	75	64	78	83	53	-	69	78	74	83	69	75.8	72.5
SherLoc	P	84	85	78	84	87	81	-	85	88	83	89	83	85.3	84.2
	A	-	80	80	88	91	83	81	86	87	81	89	-	86.4	84.5
	F	-	83	78	86	86	81	-	85	88	80	88	83	85.4	83.8
Chou		93.9	91.5	80.0	90.3	90.0	76.6	92.5	83.6	95.3	82.4	95.0	66.7	92.4	86.5
BLAST		85.7	80.7	87.5	84.2	86.3	76.6	81.7	73.7	88.9	81.6	88.2	64.8	84.8	81.7
BLAST+PDSS		89.3	84.1	85.0	83.3	94.3	72.3	83.9	83.1	94.0	80.0	95.3	66.7	90.4	84.3
Profile		76.8	79.3	72.5	80.7	88.3	72.3	78.5	69.7	89.2	76.0	90.9	57.4	84.0	77.6
Profile+PDSS		78.1	79.7	72.5	80.7	88.7	72.3	78.5	71.4	92.2	77.6	92.1	53.7	85.4	78.1
GOA		96.1	90.8	90.0	93.0	97.1	85.1	90.3	91.1	95.0	81.6	97.4	83.3	94.3	90.9
GOA+PDSS		96.1	90.9	90.0	93.0	97.2	85.1	90.3	91.5	96.7	81.6	97.4	83.3	94.8	91.1

반의 방법[18, 36]보다 높거나 유사한 성능을 나타냄을 보였다. NNPSL자료는 유사성이 90% 이하인 단백질로 구성되어 있으므로, 비교적 유사한 자료가 존재하기 때문이다. BLAST+PDSS는 BLAST탐색에서 E값이 0.001이하로 유사한 서열이 탐색될 때만 BLAST로 세포내 위치를 예측하고, 그 외의 서열은 PDSS로 예측하는 방법인데, 서열자료만을 사용하므로 계산량이 작은 장점이 있다. 이 방법은 외부정보로 모티프와 문헌정보를 사용하는 SherLoc보다 성능이 높았다.

본 논문에서 구현한 GOA는 Chou처럼 유전자 온톨로지를 사용하는 방법이다. 유전자 온톨로지 데이터베이스의 양이 증가함에 따라서, 더 많은 단백질에 대한 정보를 이용할 수 있으므로 성능이 향상되었다. GOA와 BLAST 탐색을 이용하는 방법은 기계학습이 학습과정을 통하여 많은 자료가 속한 부류쪽의 정확도가 높아지는 경향과 달리, 단백질간의 유사성만을 기준으로 예측하므로 이러한 경향이 적다. 즉, SVM을 패턴분류기로 사용하는 방법은 많은 단백질이 속한 세포내 위치인 cytoplasmic, nuclear, plasma membrane에서는 높은 성능을 보이지만, 소수의 단백질이 속한 세포내 위치인 cytoskeleton, golgi apparatus, vacuolar 등에서는 크게 성능이 떨어진다.

외부정보를 찾을 수 없는 경우에도 세포내 위치를 예측하여야 하므로, 아미노산 서열만으로 예측이 가능한 PDSS를 적용하였다. 표 1에서 +PDSS로 나타낸 부분은 BLAST, Profile, GOA로 외부 정보를 찾을 수 없는 서열에 PDSS를 적용한 것이다. 서열 중복성이 비교적 높은 PLOC 자료를 사용한 표 1의 실험에서는 유전자 온톨로지와 서열 유사성을 사용하는 것이 가장 효과적이었고, 문헌 정보와 단백질 프로파일도 아미노산 조성보다는 효과적이다.

BaCelLo IDS 자료는 단백질 서열간의 중복성이 작으므로, 자료의 중복성 감소에 따른 외부정보의 효용성 변화를 알아볼 수 있다. 표 2에서 세포내 위치 ch는 chloroplast, mi는 mitochondrial, nu는 nuclear, sp는 secretory pathway를 나타낸다. 또한 A는 동물, F는 곰팡이 자료에 대한 실험을 나타낸다. WoLF PSORT는 분류신호, 아미노산 조성, 모티프를 이용하고, MultiLoc2는 MultiLoc에 사용하는 정보에 유전자 온톨로지와 계통발생(phylogeny) 정보를 함께 사용한다. WoLF PSORT보다 MultiLoc2가 약간 우수한 성능을 보인다. SherLoc2는 MultiLoc2와 문헌정보를 함께 사용하여 MultiLoc2보다 예측정확도를 높였다. 표 2에 나타난 MultiLoc2는 SherLoc2 논문[24]에 나타난 성능으로 MultiLoc 자료[5]를 사용하여 학습한 MultiLoc2-HighRes의 성능이다. BaCelLo IDS 자료로 학습한 MultiLoc2-LowRes에 대

표 2 BaCelLo IDS자료에 대한 여러 방법의 세포내 위치 예측 정확도 비교(%)

		cy	mi	nu	sp	TA	LA
PDSS	A	56.1	87.5	77.5	92.0	76.9	78.3
	F	33.1	85.7	68.4	88.9	56.8	69.0
BaCelLo	A	51	74	57	93	64	69
	F	32	79	72	100	57	71
SherLoc2	A	77	79	62	87	71	76
	F	69	45	52	78	59	61
MultiLoc2	A	71	83	58	87	68	75
	F	56	51	50	78	53	59
WoLF PSORT	A	34	71	77	92	71	69
	F	11	53	93	89	51	62
BLAST	A	29.8	10.4	55.5	39.6	42.7	33.8
	F	15.6	23.4	79.6	33.3	40.7	38.0
BLAST +PDSS	A	56.1	85.4	77.7	92.0	76.8	77.8
	F	34.2	88.3	71.1	88.9	58.7	70.6
Profile	A	48.2	79.2	54.8	91.3	62.6	68.4
	F	40.3	57.8	74.3	33.3	55.7	51.4
Profile +PDSS	A	48.2	79.2	75.8	88.7	73.0	73.0
	F	36.4	69.5	79.0	44.4	58.1	57.3
GOA	A	74.7	91.7	78.8	95.9	82.4	85.3
	F	77.2	94.8	92.8	100	86.6	91.2
GOA +PDSS	A	74.7	91.7	98.2	100	93.2	91.2
	F	77.2	94.8	92.8	100	86.6	91.2

한 논문[8]에서는 동물은 73% TA, 80% LA이고 곰팡이는 60% TA, 66% LA로 향상된 결과를 얻었다. BaCelLo는 아미노산 조성과 단백질 프로파일을 사용하는 방법으로 SherLoc2와 비교하면 동물자료는 성능이 떨어지나 곰팡이 자료는 성능이 높다.

BaCelLo IDS 자료에 대한 실험 결과인 표 2에서 동물과 곰팡이의 TA와 LA를 모두 합한 값을 내림차순으로 표 3에 나타내었다. 표 3에서 보듯이 유전자 온톨로지를 추출하기 위해서 InterPro를 사용하는 MultiLoc2와 SherLoc2보다 GOA 데이터베이스를 사용하는 방법이 성능이 훨씬 높다. PLOC 자료에 대해 InterPro를 사용하는 Chou는 전체자료에 대하여 86.5%만 유전자 온톨로지를 얻었으나, GOA 데이터베이스를 사용할 경우에는 99.5%의 자료에 대해서 유전자 온톨로지를 얻었다. 이는 시간이 지남에 따라 유전자 온톨로지 정보가 증가한 것과 동시에, GOA 데이터베이스는 전문가가 문헌을 통하여 수집하는 유전자 온톨로지와 여러 가지 자동적인 방법을 통합한 더 광범위한 정보를 사용하기 때문이다. GOA 데이터베이스를 사용하여 유전자 온톨로지가 추출되는 자료의 비율은 BaCelLo IDS 동물 자료는 89.8%이고 이중에서 93.4%가 정확히 예측되었고, 곰팡이 자료는 100%이고 이중에서 90.1%가 정확히 예측되었다.

표 3 단백질의 세포내 위치 예측 방법 비교

예측 방법	TA(A)+LA(A)+TA(F)+LA(F)
GOA+PDSS	362.2
GOA	345.5
BLAST+PDSS	283.9
PDSS	281.0
MultiLoc2-LowRes	279.0
SherLoc2	267.0
Profile+PDSS	261.4
BaCelLo	261.0
MultiLoc2	255.0
WoLF PSORT	253.0
Profile	238.1
BLAST	155.2

표 1에서 보듯이 PLOC 자료에서는 BLAST 탐색을 사용하는 방법이 성능이 높았으나, 자료간의 중복성이 작은 BaCelLo IDS 자료에서는 성능이 크게 떨어졌다. 또한, 전체 자료에 대한 BLAST 탐색에서 E 값이 0.001이하인 비율은 PLOC 자료는 84.0%이고 이중 93.5%가 정확히 예측되었다. BaCelLo IDS 동물 자료는 15.5%이고, 이중 83.1%가 정확히 예측되었고, 곰팡이 자료는 9.2%이고 이중 87.5%가 정확히 예측되었다. 하지만, BLAST 탐색의 결과를 E 값이 0.001이하인 유사성이 높은 서열에 대해서만 이용하고, 이외의 서열은 PDSS를 적용하는 BLAST+PDSS는 외부정보를 사용하는 BaCelLo와 SherLoc2보다 성능이 높았다. 프로파일에 기반한 Profile, BaCelLo와 분류신호를 외부정보로 사용하는 WoLF PSORT는 자료를 구성하는 단백질의 유사성이 감소함에 따라 성능이 크게 감소하여 PDSS보다 성능이 떨어졌다. PDSS는 아미노산간의 진화적 유사성을 이용하여 프로파일을 만드므로, 유사한 단백질 서열들이 감소하여도 성능이 크게 저하되지 않았다.

4. 결론 및 향후연구

본 논문에서는 단백질의 세포내 위치를 예측할 때에 외부 정보인 유사 단백질 서열, 유전자 온톨로지, 문헌정보가 예측정확도에 미치는 영향을 비교하였다. 단백질 서열간의 유사성이 80% 이하인 단백질로 구성된 PLOC 자료는 비교적 아미노산 서열간의 유사도가 크며, 이러한 자료에서는 서열의 유사성, 단백질 프로파일과 문헌정보가 세포내 위치 예측에 유용하였으나, 단백질 서열간의 유사성이 30% 이하로 서열유사성이 작은 BaCelLo IDS 자료에서는 효과가 크게 감소하였다. 실험자료 전체에서의 아미노산 서열의 유사성과 관계없이, 예측하려는 단백질과 높은 서열유사성을 가진 단백질을 찾을 수 있는 경우에는 BLAST 탐색을 이용한 예측방법이 정확도가 높았다. 자료의 중복도에 비교적 영향이 적은 유전자 온톨로지 정보를 사용하는 것이 성능이 높았고, 유전자 온톨로지를 GOA 데이터베이스에서 직접 추출하는 것이 InterPro 데이터베이스를 사용하는 것보다 효과적이었다. 유전자 온톨로지와 서열 유사성을 우선적으로 적용한 후에, 이들 방법으로 정보가 추출되지 않는 서열만 PDSS를 적용하는 방법이 가장 효과적이었다.

본 논문에서는 유전자 온톨로지를 추출하여 단백질의 세포내 위치를 예측하였는데, 유전자 온톨로지는 단백질 특징을 계산화된 모델로 표현하기 위한 것이므로 단백질의 세포내 위치 예측에 맞추어서 최적화된 것은 아니다. 따라서, 향후에는 이를 최적화하는 연구를 진행할 예정이며, 문헌정보를 이용하는 방법도 추가할 예정이다. 단백질의 세포내 위치예측에서 외부정보의 활용은 계산량이 크게 증가하고, 이미 알려진 정보진 정보가 없으면 적용할 수 없다는 단점이 있으나, 단백질에 대한 활발한 연구로 점점 정보가 증가하는 추세이므로, 외부정보를 적극적으로 활용하는 방법이 효율적이라 판단된다.

참 고 문 헌

[1] H. Lodish, A. Berk, C.A. Kaiser, et al., Molecular Cell Biology, sixth Ed., p.710, W.H. Freeman and Company, New York, 2007.

[2] O. Emanuelsson, H. Nielson, S. Brunak, G. von Heijne, "Predicting subcellular localization of protein based on their N-terminal amino acid sequence," *J. Mol. Biol.*, 300, pp.1005-1016, 2000.

[3] R. Nair, B. Rost, "Mimicking cellular sorting improves prediction of subcellular localization," *J. Mol. Biol.*, 348, pp.85-100, 2005.

[4] A. Pierleoni, P. L. Martelli, P. Fariselli, R. Casadio, "BaCelLo: a balanced subcellular localization predictor," *Bioinformatics*, 22, e408-e416, 2006.

[5] A. Höglund, P. Dönnes, T, Blum, H.-W. Adolph, O. Kohlbacher, "MultiLoc: prediction of protein localization using n-terminal targeting sequences, sequence motifs and amino acid compositions," *Bioinformatics*, 22, pp.1158-1165, 2006.

[6] P. Horton, et al. "WoLF PSORT: protein localization predictor," *Nucleic Acids Res.*, 35:W585-W587, 2007.

[7] H. Shatkay et al., "SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data," *Bioinformatics*, 23, pp.1410-1417. 2007.

[8] T. Blum, S. Briesemeister, O. Kohlbacher, "MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction," *BMC Bioinformatics*, vol.10, no.274, doi: 10.1186/ 1471-2105-10-274. 2009.

[9] K.-J. Park, 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.

[10] W.-W. Yang, B.-L. Lu, 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.

[11] Q. Cui, T. Jiang, B. Liu, S. Ma, "Esub8: a novel tool to predict protein subcellular localizations in eukaryotic organisms," *BMC Bioinformatics*, vol.5, no.66, 2004.

[12] K. Chou, Y. Cai, "Prediction of protein subcellular locations by GO-FunD-PseAA predictor," *Biochem Biophys Res Commun*, 320, pp.1236-1239, 2004

[13] S.-M. Chi, "Estimating amino acids composition of protein sequences using position-dependent similarity spectrum," *Journal of KIISE : Software and Applications*, vol.37, no.1, pp.74-79, JAN. 2010. (in Korean)

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

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

[16] V. Goder, M. Spiess, "Molecular mechanism of signal sequence orientation in the endoplasmic reticulum," *The EMBO Journal*, 22, pp.3645-3653, 2003.

[17] E. Granseth, G. von Heijne, A. Elofsson, "A study of the membrane-water interface region of membrane proteins," *J. Mol. Biol.*, 346, pp.377-385, 2005.

[18] D. Xie, A. Li, M. Wang, Z. Fan, H. Feng, "LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST," *Nucleic Acids Res.*, 33, W105-W110, 2005.

[19] R. Nair, B. Rost, "Inferring sub-cellular localization through automated lexical analysis," *Bioinformatics*, 18Supple(1), S78-S86, 2002.

[20] S. Brady, H. Shatkay, "EpiLoc: a (working) text-based system for predicting protein subcellular location," *Pac. Symp. Biocomput.*, pp.604-615, 2008.

[21] A. Fyshe, Y. Liu, D. Szafron, R. Greiner, P. Lu, "Improving subcellular localization prediction using text classification and the Gene Ontology," *Bioinformatics*, vol.24, no.21, pp.2512-2517, 2008.

[22] Z. Lei, Y. Dai, "Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction," *BMC Bioinformatics*, vol.7, no.491, 2006.

[23] W.-L. Huang, et al., "ProLoc-GO: Utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization," *BMC Bioinformatics*, vol.9, no.80, 2008.

[24] S. Briesemeister, et al., "SherLoc2: A high-accuracy hybrid method for predicting subcellular localization of proteins," *J. Proteome Research*, vol.8, no.11, pp.5363-5366, 2009.

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

[26] S. F. Altschul, et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res.*, 25, pp.3389-3402, 1997.

[27] R. Nair, B. Rost, "Sequence conserved for subcellular localization," *Protein Sci.*, 11, pp.2836-2847, 2002.

[28] C. S. Yu, Y. C. Chen, C. H. Lu, J. K. Hwang, "Prediction of protein subcellular localization," *Proteins*, 64, pp.643-651, 2006.

[29] A. Bairoch, et al., "The universal protein resource (UniProt) in 2010," *Nucleic Acids Res.*, D142-D148, 2010.

[30] M. Ashburner, et al., "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet*, 25, pp.25-29, 2000.

[31] D. Barrell, et al., "The GOA database in 2009-an integrated gene ontology annotation resource," *Nucleic Acids Res.*, 37, Database issue doi:10.1093/nar/gkn803, 2009.

[32] S. Hunter, et al., "InterPro: the integrative protein signature database," *Nucleic Acids Res.*, 37, Database issue D211-D215, 2009.

[33] R. Casadio, P. L. Martelli, A. Pierleoni, "The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation," *Brief Funct Genomic Proteomics*, 7, pp.63-73, 2008.

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

[35] A. Reinhardt, T. Hubbard, "Using neural networks for prediction of the subcellular location of proteins," *Nucleic Acids Res.*, 26, pp.2230-2236, 1998.

[36] M. Bhasin, G. P. S. Raghava, "ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST," *Nucleic Acids Res.*, 32, W414-W419, 2004.

지 상 문
정보과학회논문지 : 소프트웨어 및 응용
제 37 권 제 1 호 참조