# 주요 서열 구성의 선택에 의한 단백질의 세포내 소기관 위치 예측

김수진[O1,2] 정제균[1,2] 이제근[1,2], 장병탁[1,2,3]
서울대학교 생물정보학 협동과정[1]
서울대학교 바이오정보기술 연구센터[2]
서울대학교 컴퓨터공학부[3]
{sjkim[O], jgjoung, jkrhee, btzhang}@bi.snu.ac.kr

# Predication of Protein Subcelluar Localization by Selecting Significant Sequence Composition

Soo-Jin Kim[01,2] Je-Gun Joung[1,2] Je-Keun Rhee[1,2] Byoung-Tak Zhang[1,2,3]
Graduate Program in Bioinformatics, Seoul National University[1]
Center for Bioinformation Technology, Seoul National University[2]
School of Computer Science and Engineering, Seoul National University[3]

**요 약**

단백질들이 어느 세포내 소기관에 위치하는지에 대한 지식은 그들의 기능을 예측하는데 있어서 중요한 정보를 제공한다. 하지만 실험적으로 세포내 소기관 위치를 분석하는 작업은 많은 비용과 시간을 요구한다. 따라서 지금까지 단백질의 세포내 소기관 위치 예측을 위한 다양한 계산적 방법들이 개발되었으나, 효율적인 학습 데이터의 생성에 있어서 문제점을 가지고 있다. 본 논문은 기계학습 기법을 이용하여 주요 서열 구성을 선택함으로써 예측의 성능을 최대화 하는 방법을 제안하고자 한다. 실험은 효모의 단백질의 세포 내 소기관 위치 예측에 있어서 주요 아미노산 서열들을 선택함으로써 예측의 성능을 향상시키는 결과를 보이고 있다.

## 1. 서 론

대량의 유전체 서열 분석이 완료됨에 따라 대규모의 DNA서열 및 단백질 서열들에 의한 생물학 연구가 가능해졌다. 이제는 포스트 지놈 시대로 핵심적인 조절 네트웍과 그에 관여하는 유전자들의 기능을 밝히는 작업이 주요 연구 주제로 부각되고 있다. 세포내 소기관 위치에 대한 지식은 이러한 단백질의 기능을 밝히는데 있어서 중요한 정보를 제공하고 있다. 그림 1은 세포내 소기관 위치의 개념도를 나타내고 있다.

현재까지 세포내 소기관 위치 예측을 위한 다양한 기계학습기법들이 제시되었다. 지금까지 기계학습을 이용한 세포내 소기관 위치 예측 방법으로 (1) 신경망(Neural network)과 SVM (Support vector machine)을 이용하여 아미노산 구성을 기본으로 한 예측 (2) Bayesian framework와 $K$-nearest neighbor를 이용하여 다양한 단백질의 특성 통합에 의한 예측 (3) Navie Bayesian network를 이용하여 상동관계(homology)를 기본으로 한 예측 등이 개발되었다. 그러나 이들 방법에는 어떤 기작에 의해 단백질이 그 세포내 소기관에 위치하는 지에 대한 명확한 분석이 없으며, 또한 학습 데이터의 전처리 측면에서 어떤 자질(features)들을 사용할 지에 대한 평가가 미미하다.

따라서 본 논문에서는 단백질의 세포내 소기관 위치 예측에 적합한 자질들의 선택과 학습 방법을 제안하고자 한다. 여기서 사용하는 학습 데이터는 아미노산 구성으로써, 세포내 소기관 위치에 결정적인 정보를 제공하다고 알려져 있다. 실험 결과는 특정 자질들이 단백질 위치를 결정하는데 중요한 요소가 됨을 보여 주고 있다.

## 2. 세포내 소기관 위치 예측 방법

본 논문에서는 효모(yeast)의 세포내 소기관을 8개의 구획으로 분류하고, SVM을 이용하여 위치 예측을 해 보았다. 그림 1은 단백질이 세포내 각 소기관으로 이동하는 것을 설명하고 있다. 그림 1과 같은 단백질의 이동 위치를 예측하기 위해 아미노산 구성을 기본으로 학습을 위한 입력 벡터를 만들고, 그 중 단백질이 세포내 소기관으로 이동하는 데에 영향을 많이 미치는 주요 자질을 선택하였다. 이 자질들을 이용하여 학습하고 예측 결과를 얻은 후, 자질 선

택을 하지 않은 예측 결과와 비교 분석하였다.

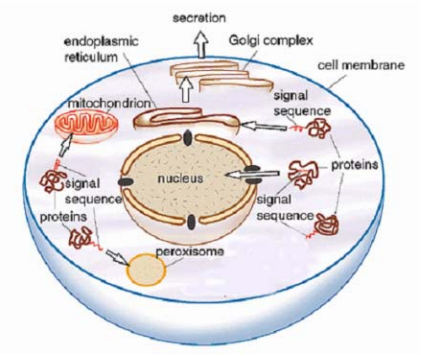


그림 1 단백질의 세포내 소기관으로의 이동

### 2-1. SVM을 이용한 단백질 위치 예측

SVM은 커널(kernel)을 이용하여 자질 공간(feature space)에 입력 데이터를 사상(mapping)시키고, 그 공간에서 분류(classification)를 하는 방법이다. SVM의 학습의 있어서 기본 원리는 2개의 클래스(+1, -1)를 가리키는 $y_i$로 레이블 된 $x$ 벡터($\vec{x}$)가 높은 차원의 자질 공간으로 사상되어 ($\vec{x}_i \in R^d$ ($i$ = 1,2,...,$N$)), 그 공간에서 각 클래스의 가장 가까운 데이터(nearest data)와 초평면(hyperplane) 사이의 거리 한계를 가장 크게(maximize margin) 구성한다. 사상은 커널 함수, $K(\vec{x}, \vec{x}_i)$에 의해 수행된다. 이 때 SVM에 의한 클래스 결정 함수는 식 (1)과 같이 나타난다.

$$f(\vec{x}) = w^T \vec{x} + b = \left( \sum_{i=1}^{N} y_i a_i . K(\vec{x}, \vec{x}_i) + b \right) \quad (1)$$

### 2-2. 자질 선택 방법

각 단백질의 아미노산 구성을 분석하여 입력 벡터를 만들고, SVM을 이용하여 각 자질이 결과를 예측하는데 미치는 영향 정도를 평가하여 특정 자질을 선택한다. 자질 선택은 식 (1)에서 입력 데이터에 대한 가중치 $w$의 값에 의해 결정된다. 가중치 $w$는 (2)의 식으로부터 얻어진다.

$$\vec{w} = \sum_{i=1}^{N} y_i a_i \vec{x} \quad (2)$$

가중치를 표현하는 벡터 $w$는 식 (2)에서와 같이 입력 벡터 $x_i$의 선형 결합(linear combination)이다. 여기서 식 (3)에 의해 최종적으로 $x_i$에 대한 자질 결정 스코어 $S(x_i)$값이 얻어지고 이 순위에 따라 자질이 선택된다.

$$S(x_i) = w_i^2 \quad (3)$$

가중치 $w_i$에 의해 각 자질의 중요도가 결정되고, 그 중 높은 순위에 있는 자질을 선택하게 된다.

### 2-3. 실험 데이터 및 설계

본 논문에서는 NCBI Database에 있는 효모 전체의 단백질 아미노산 서열을 이용하였다. 각 단백질 서열을 1/$n$씩 나누고, 그 각 서열은 ($n$ x 20)차원으로 입력 벡터를 구성한다. 실험에서는 $n$이 1, 2, 3인 각각의 경우에 대해 단백질 위치를 예측해보았다. 각 차원은 20가지 아미노산 각각을 의미한다. 입력 벡터의 값은 다음과 같은 식에 의해서 계산되어 진다.

$$f(a_i) = \frac{N(a_i)}{l} \quad (4)$$

식 (4)에서 $l$은 서열의 길이를 의미하고, $N(a_i)$는 아미노산 $a_i$에 대한 개수를 나타낸다.

효모의 각 단백질의 세포내 소기관에 위치 정보는 SGD (Saccharomyces Genome Database)로부터 얻었다. 이를 기반으로 5 fold cross validation 테스트를 사용해서 정확도를 측정했다.

## 3. 실험 결과

효모의 세포내 소기관 구획을 소포체(ER), 골지체(Golgi), 세포질(Cytosome), 핵(Nucleus), 과산화소체(Peroxisome), 세포막(Plasma membrane), 리소좀(Lysosome), 이렇게 8구획으로 나누어 단백질의 위치를 예측하였다.

| | TP | FP | TN | FN | SP | SN |
|---|---|---|---|---|---|---|
| 소포체 | 0.945 | 0.055 | 0.997 | 0.003 | 0.948 | 0.997 |
| 핵 | 0.882 | 0.118 | 0.997 | 0.003 | 0.894 | 0.996 |
| 골지체 | 0.546 | 0.454 | 0.922 | 0.078 | 0.670 | 0.875 |
| 세포질 | 0.56 | 0.44 | 0.732 | 0.268 | 0.625 | 0.676 |
| 미토콘드리아 | 0.713 | 0.287 | 0.397 | 0.603 | 0.580 | 0.541 |

표 1. 각 단백질의 세포 내 소기관에 위치 예측의 Specificity와 Sensitivity (60차원의 입력 벡터)

Positive 데이터의 개수는 소포체 255개, 핵 524개, 골지체 119개, 세포질 250개, 미토콘드리아 376개로 하였고 negative 데이터의 개수는 positive 데이터 개수에 비례하여 설정하였다. 위의 표 1에서는 자질 선택을 하지 않고 60차원의 벡터로 SVM을 이용하여 학습시킨 결과이다. 소포체와 핵은 specificity 와 sensitivity가 모두 높게 나왔다. 전체적인 정확도 역시 소포체에서는 97.5%, 핵에서는 94.4%로 높게 측정되었다. 그러나 과산화소체, 세포막, 리소좀에 위치하는 단백질은 이와 같은 방법으로는 예측하기가 쉽지 않았다.

다음으로 각 단백질이 특정 세포내 소기관에 위치하게 하는 특성을 알아보기 위해 자질 선택을 해서 학습을 시켜보았다. 2.2절에서 설명한 방법을 통해 자질을 높은 순위에서 10개를 선택하여 예측해 본 결과, 자질을 모두 사용 했

을 때의 결과와 비슷한 정확도를 얻을 수 있었다. 표 2는 10개의 자질만을 이용한 경우 예측된 결과를 보여준다.

| | TP | FP | TN | FN | SP | SN |
|---|---|---|---|---|---|---|
| 소포체 | 0.933 | 0.067 | 1.000 | 0.000 | 0.937 | 1.000 |
| 핵 | 0.882 | 0.118 | 0.998 | 0.002 | 0.894 | 0.998 |
| 골지체 | 0.546 | 0.454 | 0.991 | 0.009 | 0.686 | 0.983 |
| 세포질 | 0.524 | 0.476 | 0.776 | 0.224 | 0.620 | 0.700 |
| 미토콘드리아 | 0.75 | 0.25 | 0.374 | 0.626 | 0.599 | 0.545 |

표 2. 순위가 높은 10개의 자질을 선택하여 각 단백질의 세포내 소기관에 위치 예측의 Specificity와 Sensitivity (10차원의 입력 벡터)

입력 벡터의 차원이 1/6으로 줄었음에도 불구하고 specificity와 sensitivity가 큰 차이를 보이지 않았고, 골지체와 미토콘드리아에서는 자질을 모두 사용하였을 때 보다 더 좋은 결과를 보여주었다. 따라서 특정 몇몇의 자질이 각 단백질을 특정 세포내 소기관으로 위치시키는 데 중요한 역할을 하는 것을 알 수 있었다.

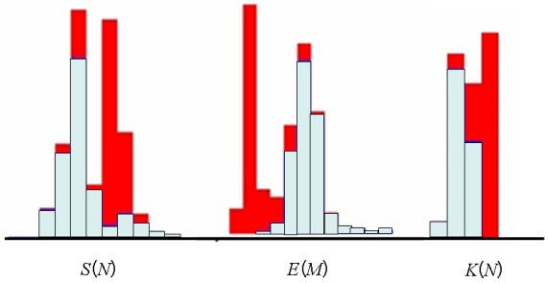


그림 2. 핵에서의 선택된 자질에 따른 데이터 분포도. $S$, $E$, $K$: 각각의 아미노산, $N$ ($N$-터미널 방향의 서열), $M$ (중간서열)

그림 2는 핵에서의 주요 자질에 따른 positive 데이터와 negative 데이터에 대한 분포를 보여주고 있다.

한편 입력 벡터 차원에 따라서도 결과에 차이를 보인다는 것을 알 수 있었다. 표 3은 소포체와 세포질에서 입력 벡터 차원이 다른 경우의 예측 정확도를 보여주고 있다. 세포내 소기관 위치에 따라서 40차원과 60차원 간에는 다소 차이가 있었다.

| | 40차원 | 60차원 |
|---|---|---|
| 소포체 | 57.3% | 97.5% |
| 세포질 | 71.6% | 64.6% |

표 3. 입력 벡터의 차원이 다를 때 특정 세포내 소기관의 단백질 위치 예측 정확도

소포체의 경우 입력 벡터의 차원이 40인 경우 현저하게 정확도가 떨어졌다. 따라서 예측하고자하는 위치에 따라 입력 벡터의 차원을 다르게 하면 더 좋은 결과를 얻을 수 있을 것이다.

## 4. 결론

본 논문에서는 효모 단백질의 아미노산 서열로부터 서열의 길이에 상관없이 일정한 입력 벡터를 가지고 SVM을 이용하여 단백질의 세포내 소기관 특정 위치를 예측해 보았다. 또 이를 기반으로 자질 선택을 하여 순위가 높은 자질만을 선택하여 학습을 해도 결과에 많은 차이를 보이지 않음을 알 수 있었다. 이로써 특정 주요 자질이 단백질을 특정 세포내 소기관으로 위치시키는 데에 결정적인 역할을 수행한다는 것을 알 수 있었다. 결정적인 역할을 하는 주요 자질들에 대한 분석을 통해, 각 단백질의 특정 세포내 소기관으로의 이동에 중요한 요인들을 알 수 있을 것이다. 또한 입력 벡터의 차원 변화와 자질 수의 최적화를 통해 예측 정확도를 더욱 높일 수 있을 것이다.

현 연구는 효모의 세포내 소기관을 8구획으로 나누어 실험하였으나, 앞으로 구획을 보다 세분화한 예측도 수행되어야할 것이다.

## 감사의 글

이 논문은 과학기술부 국가 지정 연구실 사업(NRL)에 의하여 지원되었음.

## 참고 문헌

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

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

[3] Drawid, A. and Gerstein, M., A Bayesian system integrating expression data with sequence patterns for localizing proteins: Comprehensive application to the yeast genome, *J. Mol. Biol.* 301:1059-1075. 2000.

[4] Horton, P. and Nakai, K., A probabilistic classification system for predicting the cellular localization sites of proteins, *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 11:241-247. 1996.

[5] Lu, Z., Szafron, D., Greiner, R., Lu, P., Wishart, D.S., Poulin, B., Anvik, J., Macdonell, C., and Eister, R., Predicting subcellular localization of proteins using machine-learned classifiers, *Bioinformatics* 20:547-556. 2004.