단백질 서열의 n-Gram 자질을 이용한 세포내 위치 예측

Classification Protein Subcellular Locations Using n-Gram Features

김진숙
한국과학기술정보연구원

Jinsuk Kim
Korea Institute of Science & Technology Information (KISTI)

요약

단백질의 기능은 그 기능을 발휘하는 세포내의 위치와 밀접한 연관이 있다. 따라서 새로운 단백질의 서열이 밝혀지면 이 단백질의 세포내 위치를 규명하는 것은 생물학적으로 매우 중요한 일이다. 이 논문에서는 단백질의 *n*-그램과 *k*NN (*k*-Nearest Neighbor) 분류기를 이용한 새로운 세포내 위치예측 방법을 다룬다. 이 방법은 입력 단백질 서열과 가장 유사한 가중치를 가지는 *k*개의 단백질이 가지는 세포내 위치 정보들을 취합하여 입력 단백질의 세포내 위치를 추정한다. 단백질간의 유사도 가중치는 두 단백질서열의 5-그램 자질의 유사도를 비교하여 계산된다. 단백질의 세포내 위치예측 정확도를 검증하기 위해 SWISS-PROT 단백질 데이터베이스로부터 세포내 위치가 알려진 51,885개의 서열을 추출하여 대용량 테스트 컬렉션을 구축하였으며, 다른 연구자들이 제공하는 또 하나의 소용량 테스트 컬렉션을 실험에 사용하였다. 이 논문에서 사용한 예측방법은 대용량 테스트컬렉션에 대해 약 93%의 정확도를 보여주었으며, 소용량 데스트컬렉션을 이용하여 이전 실험과 비교하였을 때도 이 방법이 다른 시스템에 비해 성능이 우월함을 알 수 있었다.

Abstract

The function of a protein is closely co-related with its subcellular location(s). Given a protein sequence, therefore, how to determine its subcellular location is a vitally important problem. We have developed a new prediction method for protein subcellular location(s), which is based on *n*-gram feature extraction and *k*-nearest neighbor (*k*NN) classification algorithm. It classifies a protein sequence to one or more subcellular compartments based on the locations of top *k* sequences which show the highest similarity weights against the input sequence. The similarity weight is a kind of similarity measure which is determined by comparing *n*-gram features between two sequences. Currently our method extract penta-grams as features of protein sequences, computes scores of the potential localization site(s) using *k*NN algorithm, and finally presents the locations and their associated scores. We constructed a large-scale data set of protein sequences with known subcellular locations from the SWISS-PROT database. This data set contains 51,885 entries with one or more known subcellular locations. Our method show very high prediction precision of about 93% for this data set, and compared with other method, it also showed comparable prediction improvement for a test collection used in a previous work.

I. 서 론

인간게놈프로젝트 등 대규모 서열규명작업들이 진행되면서 새로운 DNA 서열들이 수없이 작성되고 있다. 이에 따라 오늘날 단백질 데이터베이스의 규모도 급속하게 커지고 있으며 신규 단백질 서열의 기능을 규명할 수 있는 시스템적 방법론의 개발이 큰 관심을 끌고 있다. 단백질이 세포내에서 어떤 위치에 존재하는 지는 그 기능을 밝히는 데 중요한 단서가 된다[2,3,6]. 예를 들어 바이러스의 병원성은 세포밖으로 배출되는 단백질에 기인한다. 따라서 이 병원성 인자를 찾고자 할 경우에는 세포외로 배출되는 것으로 밝혀지거나 예측된 단백질 서열만을 대상으로 연구를 진행할 수 있으므로 연구의 효율을 극대화할 수 있다.

단백질의 세포내 위치예측은 일반적으로 단백질 서열 데이터베이스를 대상으로 유사도 분석을 통해 이루어진다. 정확한 위치는 실험실에서의 실험결과를 통해 이루어져야 하나 이는 시간과 비용적인 측면에서 매우 소모적이다. 이의 대안으로서 단백질 서열 DB를 바탕으로 세포내 위치를 예측하려는 시도들이 이루어지고 있다. 대표적인 연구들로는 TargetP[4], PSORT[10], NNPSL[11], MitoProt[5], Predotar[5], PLOC

[9] 등이 있다. 대부분의 연구들은 두 가지 자질을 위치예측에 사용하고 있다. 첫째는 단백질 서열의 아미노산 조성을 사용한다(PLOC, NNPSL). 두 번째는 아미노산 조성과 시그널 펩티드를 동시에 이용하는 것이다(TargetP). 그러나 많은 경우 서열에서 시그널 펩티드를 찾을 수 없을 뿐만 아니라 단백질의 아미노산 조성만을 자질로서 사용하게 되면 서열자체의 고유한 정보를 잃어버릴 가능성이 크다. 이 때문에 아미노산 서열 자체를 세포내 위치예측에 직접 활용해야 한다는 비판이 제기되었다[17].

이와는 독자적으로 전산학에서는 문서로부터 적절한 자질을 추출함으로써 문서의 분류를 자동으로 계산할 수 있는 문서범주화 분야의 연구가 성공적으로 진행되고 있다[12]. 본 논문은 단백질 아미노산 서열을 하나의 문서로 간주하여 효과적인 자질로서 아미노산의 5-gram을 선정하였다[7]. 더불어 *k*NN 분류기[15,16,12,8]를 적용하면 대용량 단백질 데이터베이스를 기반으로 효율적인 단백질 예측 시스템을 구현할 수 있음을 알 수 있었다.

II. 실험방법

1. 실험환경

실험은 듀얼 Pentium Xeon 2.8GHz, 2GB 메모리, RAID-5 SCSI 저장장치를 장착한 리눅스장비에서 수행되었다. 문서범주화의 전반적인 과정은 정보검색과 유사하기 때문에 이 실험에서는 *k*NN(*k*-nearest neighbor) 분류기와 *n*-gram 추출 방법을 정보검색관리시스템 KRISTAL-IRMS (http://www.kristalinfo.com) 상에 구현하였다. 단백질 서열로부터 추출한 *n*-gram 자질은 KRISTAL 시스템의 역파일에 저장되며 벡터공간모델을 사용하여 입력서열에 대해 상위 *k*개의 유사서열을 검색한 후 이를 입력서열의 세포내 위치예측에 사용한다.

2. 자질추출(Feature Extraction)

자연어로 씌어진 문서에서 자질을 추출하는 방법은 비교적 명확하다. 문서범주화에서 문서의 자질은 대개 문서내의 단어들로 이루어진다[12]. 반면 단백질은 20개의 아미노산 코드로 이루어진 연속적인 문자열로 표현되므로 의미있는 자질을 추출하는 것은 명확하지 않다. 본 논문에서는 단백질 서열을 서로 중첩하는 *n*-그램으로 분할하여 이를 자질로 사용하였으며, $n = 5$를 선택하였다[7]. 예를 들어 단백질 서열 "ACDEFLERR"은 "ACDEF", "CEDFL", "DEFLE", "EFLER", "FLERR"의 자질을 가지게 된다. $n = 3, 4, 6, 7$인 경우에 대해서도 실험을 수행하였으나 $n = 5$(penta-gram)인 경우가 가장 좋은 성능을 보였다(결과 생략).

3. 유사도측정

본 논문에서는 입력서열이 주어지면 이와 가장 유사한 상위 k개의 서열을 벡터공간모델을 사용하여 검색한다. 여기에 사용된 서열간 유사도 $S(q,s)$는 다음과 같다(q는 입력서열, s는 비교대상 서열):

$$S(q,s) = \sum_{t \in q \wedge s} (w_{s,t} \bullet w_{q,t})$$

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

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

이며 $f_{s,t}$는 5-그램 토큰 t의 서열 s내 빈도, N은 데이터베이스(학습집합) 내에 포함된 모든 서열의 수, f_t는 t를 포함하는 모든 서열의 수이다.

4. 범주가중치 측정

상위 k개의 서열 집합 $K = \{s_1, s_2, s_3, \cdots, s_k\}$가 결정되면 이들로부터 범주 c_i의 적합도 $w(q,c_i)$는 다음과 같이 계산한다:

$$w(q,c_i) = \sum_{n=1}^{k} S(q,s_n) \bullet (s_n \wedge C_i)$$

여기서 C_i는 학습집합내에서 범주 c_i로 미리 할당된 서열의 집합이다. 모든 후보 범주에 대한 범주가중치 계산이 완료되면 이 가중치들은 $(0,1]$로 정규화되고 사용자에 의해 주어진 임계값 이상을 가지는 후보 범주만을 최종 범주로 반환한다.

5. 데이터 집합

본 논문에서는 [9]에서 사용한 7,580개의 서열이 포함된 소규모 데이터집합(PLoc이라 명명)과 SWISS- PROT[1]에서 추출한 51,885개의 서열을 가지는 대규모 데이터집합(SLP라 명명)을 성능평가에 사용하였다. 표 1에서 보여주는 바와 같이 SWISS-PROT에서 세포내 위치 항목에서 지정된 키워드가 출현하는 단백질들을 12개의 세포내 위치로 분류하여 PLoc 및 SLP 데이터집합을 구축하였다.

[표 1] 세포내 위치(Subcellular)를 지정하는 위치별 키워드 목록

Subcellular location	Keywords
Chloroplast	chloroplast
Cytoplasmic	cytoplasmic
Cytoskeleton	cytoskeleton
	filament
	microtubule
Endoplasmic reticulum	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

PLoc의 경우 원핵생물의 서열은 제거하였으며 B, Z, X 등의 불명확한 아미노산 코드가 있는 서열은 제거되었으나 SLP에서는 이런 제약없이 모든 서열을 데이터에 포함하였다. 각 데이터집합은 학습집합(Training Set)과 시험집합(Test Set)으로 분할된다. 학습집합은 kNN 분류기의 학습에 사용되며 시험집합은 학습된 분류기의 성능평가를 위해 사용된다. 표 2에 PLoc과 SLP의 학습집합 및 시험집합의 개수를 표시하였다. PLoc의 경우 이전 실험과의 비교를 위해 5가지로 분할하였다.

[표 2] PLoc과 SLP 데이터집합의 학습집합/시험집합 분할

Data Set	Test Set	Training Set	Total
PLoc1	1,522	6,058	7,580
PLoc2	1,514	6,066	7,580
PLoc3	1,521	6,059	7,580
PLoc4	1,508	6,072	7,580
PLoc5	1,515	6,065	7,580
SLP	8,645	43,240	51,885

6. 성능평가지표

성능평가는 정확도(precision)와 재현율(recall)을 기반으로 마이크로평균(Micro-averaging) 기법과 매크로평균(Macro-averaging) 기법을 모두 사용하였다[12]. 마이크로평균 정확도(P_{mi})와 마이크로평균 재현율(R_{mi})은 다음과 같이 정의된다:

$$P_{mi} = \frac{\text{locations relevant} \wedge \text{retrieved}}{\text{locations retrieved}} = \frac{TP}{TP+FP}$$

$$R_{mi} = \frac{\text{locations relevant} \wedge \text{retrieved}}{\text{locations relevant}} = \frac{TP}{TP+FN}$$

TP는 시험집합에 대해 kNN 분류기를 이용하여 세포내 위치를 예측했을 경우 true positive의 총수, FP는 false positive의 수, FN은 flase negative의 수를 나타낸다[12].

마이크로평균 기법의 성능평가는 학습집합내에서 범주간 할당된 문서수의 편차가 심할 경우에는 할당된 문서가 많은 범주의 평가지표에 전반적인 성능평가에 오류를 범할 가능성이 크다는 보고가 있다[14]. 따라서 본 논문에서는 매크로평균 정확도(macro-averaged precision) P_{ma}와 매크로평균 재현율(macro-averaged recall) R_{ma}를 측정하여 성능평가를 보충하였다.

$$P_{ma} = \frac{\sum_{i=1}^{m} p_i}{m} \quad \text{여기서} \quad p_i = \frac{TP_i}{TP_i + FP_i}$$

$$R_{ma} = \frac{\sum_{i=1}^{m} r_i}{m} \quad \text{여기서} \quad r_i = \frac{TP_i}{TP_i + FN_i}$$

m은 전체 범주의 수이며(여기서는 12), P_{ma}는 각 범주당의 정확도를 평균한 값이며 R_{ma}는 각 범주당 재현율의 평균값이 된다. 마이크로 평가에서는 각 범주의 성능평가가 되지 않으나 매크로 평가에서는 각 범주에 대한 성능이 별도로 평가되어 각 범주가 전체 성능에 미치는 영향을 평가할 수 있는 장점이 있다.

III. 실험결과

본 논문에서는 kNN 분류기의 k값으로 20을 선정하였다. 실험결과 k값은 5이상일 경우에는 성능평가에 큰 영향을 미치지 않았다(결과 생략).

1. 성능평가

표 3은 PLoc과 SLP 데이터집합에 대한 단백질 세포내 위치 예측 성능평가 결과이다. PLoc의 경우는 5개의 학습/시험서열 분할에 대해서 각각의 실험결과를 측정하고 이를 평균한 매크로평균 재현율과 마이크로평균 재현율을 나타냈다.

표 3에서 12개 위치의 매크로평균 재현율은 PLoc의 경우 0.567 ~ 0.926의 범위에 있고, SLP의 경우에는 0.650 ~ 0.976이다. 마이크로평균 재현율의 경우 PLoc과 SLP에 대해 각각 0.814와 0.928의 재현율을 보여주었다(참조: 이 두 값은 모두 F_1 평가지표[13]의 특수지점인 Break-even Point[12,16]를

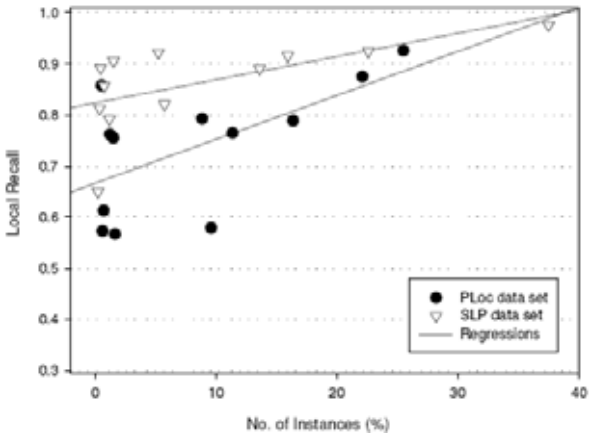
측정한 것이다).

[표 3] PLoc과 SLP 데이터집합에 대한 성능평가 결과

Subcellular locations	PLoc	SLP
Chloroplast	0.793	0.922
Cytoplasmic	0.789	0.976
Cytoskeleton	0.858	0.650
Endoplasmic reticulum	0.755	0.906
Extracellular	0.765	0.891
Golgi apparatus	0.573	0.792
Lysosomal	0.762	0.892
Mitochondrial	0.579	0.821
Nuclear	0.926	0.916
Peroxisomal	0.567	0.857
Plasma membrane	0.875	0.925
Vacuolar	0.613	0.813
Macroaveraged recall (R_{ma})	0.738	0.863
Microaveraged recall (R_{mi})	0.814	0.928

2. 결과분석

표 3에서 대규모 데이터인 SLP에 대해서는 마이크로평균 재현율(실제로는 BeP)이 92.8%, 소규모 데이터인 PLoc에 대해서는 81.4%로 나타났다. 동일한 방법으로 측정하였지만 소형 학습집합과 대형 학습집합의 성능이 크게 차이가 났다. 이는 *k*NN 분류기의 특성상 top-*k* 서열의 정확도가 서열의 수가 늘어남에 따라 증가하는 것을 의미하는 것으로 분석된다. 이와 더불어 그림 1은 12개 범주에 대한 매크로평균 재현율을 학습집합에서 각 범주에 포함된 단백질 수(전체의 백분율로 표현)에 대한 그래프를 보여주고 있다.



▶▶ 그림 1. 범주당 재현율과 범주당 표본 단백질 수의 관계

그림 1은 각 범주가 가지는 학습집합 표본수의 크기와 예측성능이 서로 관계가 있음을 보여주고 있다. 즉, 각 범주에서 표본의 수가 적을수록 재현율이 감소하는 경향을 PLoc과 SLP 데이터 집합 모두에서 보여주고 있다. 그러나 표본수가 전체적으로 풍부한 SLP 데이터집합의 경우 이러한 편차가 PLoc 데이터집합에 비해서 그렇게 크지 않음을 회귀선(regression)을 통해 알 수 있다.

3. 이전 연구와의 성능비교

실험데이터를 공개한 과거 실험중에서 대표적인 세포내 위치예측 연구로서 SVM 알고리즘을 적용한 PLOC[9]을 선정하였다. 표 4에서는 [9]에서 발표된 위치예측 성능과 본 논문의 실험결과를 비교한다. 실험에 사용된 데이터는 [9]에서 공개하는 PLoc 데이터집합을 사용하였다.

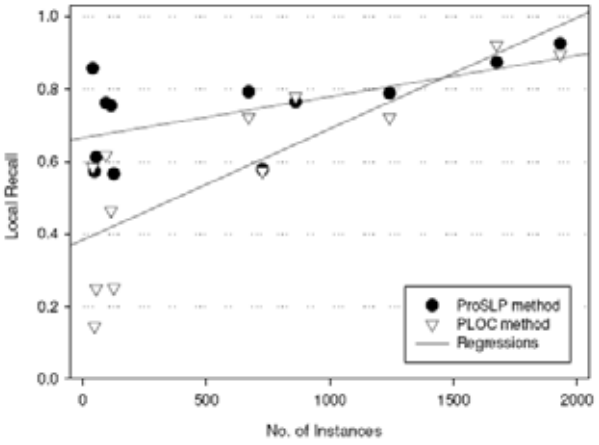
[표 4] PLOC[9] 시스템과 본 논문(ProSLP)의 성능 비교

Subcellular locations	PLOC	ProSLP
Chloroplast	72.3	79.3
Cytoplasmic	72.2	78.9
Cytoskeleton	58.5	85.8
Endoplasmic reticulum	46.5	75.5
Extracellular	78.0	76.5
Golgi apparatus	14.6	57.3
Lysosomal	61.8	76.2
Mitochondrial	57.4	57.9
Nuclear	89.6	92.6
Peroxisomal	25.2	56.7
Plasma membrane	92.2	87.5
Vacuolar	25.0	61.3
Macroaveraged recall (%)	57.9	73.8
Microaveraged recall (%)	78.2	81.4

표 4에서 PLOC 방법의 결과를 보면 범주에 따라 재현율의 차이가 14.6 ~ 92.2%로 그 편차가 매우 심함을 알 수 있다. 그러나 본 논문의 경우 56.7~92.6%로 PLOC 방법에 비해서는 그 편차가 심하지 않음을 알 수 있다. 이러한 특징 때문에 마이크로평균 재현율이 78.2%에서 81.4로 향상되었으며 특히 매크로평균 재현율의 경우 57.9%에서 73.8%로 크게 향상되었음을 알 수 있다.

그림 2는 그림 1과 마찬가지로 PLOC 시스템과 본 논문의 시스템이 PLoc 데이터집합을 대상으로 보여주는 성능평가 결과를 범주당 재현율을 범주당 표본수에 따라서 보여주고 있다. 동일한 데이터집합을 사용하였음에도 PLOC 시스템은 범주당 표본수에 따라서 성능의 편차가 극심함을 알 수 있다. 그러나 본 논문에서 제시하는 방법은 표본수가 적어도 크게 영향을 받지 않음을 알 수 있다. 이러한 결과는 학습에 크게 의존적인(*busy* learning) 인공지능망이나 SVM과 같은 분류 알고리즘에서 특징적으로 나타나는 과최적화(over-fitting)에 기인하는 것이다[12,9]. 그러나 본 논문에서 제시하는 *k*NN 기법은 학습에 크게 의존하지 않기 때문에 과최적화 문제가 발생하지 않음으로 해서 그림 2에서 보여주는 것과 같이 표본수의 편차에 크게 영향을 받지 않고 범주간에 고른 예측성능을 보여주

는 것으로 보인다.



▶▶ 그림 2. PLOC과 본 논문(ProSLP)의 범주당 재현율과 범주당 표본수의 비교

IV. 결론 및 토의

본 논문에서는 수만 건의 단백질 서열을 대상으로 *n*-그램을 자질로 하여 *k*NN 분류 알고리즘을 적용하면 효율적으로 입력 단백질의 세포내 위치를 예측할 수 있음을 보였다. 본문에서 언급되지는 않았으나 PC급 장비에서도 수백 개의 아미노산 서열로 구성된 단백질의 위치예측을 평균 3초 이내에 수행할 수 있었다. 타 온라인 시스템의 경우 분단위로 계산을 수행해야 하는 점을 고려할 때 속도적인 측면에서도 본 논문의 시스템은 단백질의 기능연구에 큰 기여를 할 수 있을 것으로 기대된다.

문서범주화에서 자질선택(feature selection)은 속도뿐만 아니라 성능도 크게 높여주지만[12,16,8], 본 논문에서는 *n*-그램의 특이성 때문에 자질선택을 적용하지 않았다. 또한 Scoring Matrix와 같은 단백질 고유의 특징을 *n*-그램에 적용하면 생물학적 관점을 추가로 활용할 수 있을 것으로 기대된다. 실험결과에서 알 수 있듯이 본 논문이 제시하는 방법은 표본의 수가 커질수록 정확도가 향상되는 특징이 있으므로 단백질의 수가 급증하는 현재의 추세에 적합한 단백질 예측시스템이 될 것으로 기대된다.

▌참 고 문 헌▐

[1] Bairoch, A. and Apweiler, R., “The SWISS-PROT Protein Sequence Database and its Supplement TrEMBL in 2000”, *Nucleic Acids Res.*, Vol. 28. pp. 45-48, 2000.

[2] Chou, K .C. and Elrod, D. W., "Protein Subcellular Location Prediction", *Protein Engineering*, Vol. 12, pp. 107-118, 1999.

[3] Eisenhaber, F. and Bork, P., "Evaluation of Human-readable Annotation in Biomolecular Sequence Database with Biological Rule Libraries", *Bioinformatics*, Vol. 15. pp. 528-535, 1998.

[4] Emannuelsson, O., *et al*, "Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acids Sequence", *J. Mol. Biol.*, Vol. 300. pp. 1005-1016, 2000.

[5] Feng, Z. P., "An Overview on Predicting the Subcellular Location of Protein", *In Silico Biol.*, Vol. 2. pp. 291-303.

[6] Gardy, J. L., *et al.*, "PSORT-B: Improving Protein Subcellular Localization Prediction for Gram-negative Bacteria", *Nucleic Acids Res.*, Vol. 31. pp. 3613-3617, 2003.

[7] Hwang, M.-N. and Kim, J., "Protein Sequence Search Based on N-Gram Indexing", *Bioinformatics and Biosystems*, Vol. 1. No. 1. pp. 46-50, 2006.

[8] Kim, J. and Kim, M. H., "An Evaluation of Passage-based Text Categorization", *J. Intelligent Info. Sys.*, Vol. 23. No. 1. pp. 47-65, 2004.

[9] Park, K. J. and Kanehisa, M., "Prediction of Protein Subcellular Locations by Support Vector Machines Using Compositions of Amino Acids and Amino Acid Pairs", *Bioinformatics*., Vol. 19. pp. 1656-1663, 2003.

[10] Nakai, K. and Hoton, P., "PSORT: A Program for Detecting the Sorting Signals of Proteins and Predicting their Subcellular Localization", *Trends Biochem. Sci.*, Vol. 24. pp. 34-35, 1999.

[11] Reinardt, A. and Hubbard, T., "Using Neural Networks for Prediction of the Subcellular Location of Proteins", *Nucleic Acids Res.*, Vol. 26. pp. 2230-2236.

[12] Sebastiani, F., "Machine Learning in Automated Text Categorization", *ACM Comp. Surv.,* Vol. 34. pp. 1-47.

[13] van Rijsbergen, C., "Information Retrieval", Butterworths, London, 1979.

[14] Wiener E, *et. al.*, "A Neural Network Approach to Topic Spotting", *Proc. SDAIR-95*, pp. 317-332, 1995.

[15] Yang, Y., "Expert Network: Effective and Efficient Learning from Human Decision in Text Categorization", *Proc. ACM SIGIR*, pp. 13-22, 1994.

[16] Yang, Y., "An Evaluation of Statistical Approaches to Text Categorization", *Info. Retrieval*, Vol. 1. pp. 69-90, 1999.

[17] Yuan, Z., "Prediction of Protein Subcellular Locations Using Markov Chain Models", *FEBS Letter*, Vol. 451. pp. 23-26.