출현 시퀀스 마이닝 기반의 단백질 2 차 구조 예측†

Meijing Li*, 이헌규**, Khalid E.K. Saeed*, 손호선*, 류근호*
*충북대학교 데이터베이스/바이오인포메틱스 연구실
**한국전자통신연구원 우정물류기술연구부
e-mail:{mjlee, abolkog, shon0621, khryu} @dblab.chungbuk.ac. kr, hg_lee@etri.re.kr

Predict Protein Secondary Structure based on Emerging Sequence Mining

Meijing Li*, Heon Gyu Lee**, Khalid E.K. Saeed*, Ho Sun Shon*, Keun Ho Ryu*
*Database/Bioinformatics Laboratory, Chungbuk National University
**Poster Technology Research Center, Electronics & Telecommunications Research Institute

요 약

최근 단백질 기능 예측을 위한 서열비교와 구조비교 기법들은 정확한 분류가 가능한 반면, 새로운 단백질 기능 분류를 함에 있어서 많은 복잡도가 따른다. 따라서 이 논문에서는 보다 빠른 단백질의 구조 분류 및 예측을 위하여 출현 시퀀스(emerging sequence)를 기반으로 하는 분류기법을 제안하였다. 이 기법에서는 먼저, 출현 시퀀스 마이닝 알고리즘을 이용하여 단백질 서열 데이터로부터 4 가지의 단백질 2 차 구조 출현 시퀀스를 발견하고, SVM 을 이용하여 단백질의 출현 시퀀스 속성으로부터 단백질의 2 차 구조를 예측하였다.

1. 서론

최근 단백질 구조와 기능을 예측하고자 하는 연구는 생명정보학에 있어서 중요한 이슈가 되고 있다. 단백질의 구조와 기능에 관한 연구는 현재 생명공학의 여러 분야에서 활발히 이루어지고 있으며, 연구의 분야 또한 데이터마이닝 기법을 비롯하여 통계, 기계학습 등의 접근 방법들이 제시되고 있다.

서열 기반의 단백질 구조 및 기능 예측 하는데 가장 기본이 되는 것은 서열 정렬(sequence alignment)이다. 이는 서열간의 상관관계를 보여주기 위한, 특히 상동성(homology)을 나타내기 위해 핵산이나 단백질의 서열을 정렬하는 것을 말 하며, 몇 개의 서열을 정렬 하는가에 따라 서열 쌍 정렬(pairwise alignment)과 다중정렬(multiple alignment)로 나누어진다. 또한 상동성의 종류에 따라 전역정렬(global alignment)과 국소정렬(local alignment)로 구분된다. 전역정렬은 두 서열을 비교하는 경우, 서로 동일한 종류의 핵산이나 단백질 서열에서 전체적으로 최대의 상동성을 측정하는 정렬방법이며, 국소정렬의 경우는 두 서열의 어떤 부분 서열이 높은 상동성을 가지는가를 위한 정렬방법이다. 또한 동적 프로그래밍은 크게 대표되는 두 가지의 방법이 전역정렬로 쓰이는 니들만-분취 법과 국소정렬로 쓰이는 스미스-워터만 법이다. 이 두 가지 방법은 공통적으로 서열 쌍 정렬의 범주에 속하며, 기본이 되는 최대 공통문자열(LCS: Longest Common Substring)[1]을 사용한다.

이러한 두 가지의 정렬방법을 기본으로 하여 많은 정렬 알고리즘들이 생겨나게 되었다. Lipman Wilbur, Pearson 등이 해싱과 원도우 크기 설정의 방법을 사용하여 고안한 FASTP/FASTN[2]는 국소적 상동성 정렬을 포함한 FASTA 로 발전하여 데이터베이스 대상의 상동성 검색에 사용되었고, Altschul 과 Karlin 등에 의하여 개발된 BLAST[3]는 데이터베이스 대상의 국소적 상동성 검색(local homology search)을 위해 가장 많이 활용되어 왔다. BLAST 는 휴리스틱(heuristic) 알고리즘을 사용하여 매우 효율적이고 빠른 반면 서열유사도가 작은 경우 정확도가 떨어지는 단점이 있다[4][5].

이 논문에서는 서열정열 비교에 의한 단백질의 기능 예측이 아닌, 단백질 서열의 출현 시퀀스(emerging sequence)로부터 분류 모델을 만드는 데이터마이닝 기법을 적용한 단백질 구조 분류 기법을 제안하였다. 기존의 연구에 비해 이 기법은 단백질 구조 예측을 보다 효율적이고 많은 데이터에 대해 동시에 실행할 수 있게 하였다.

이 논문의 구성은 다음과 같다. 2 장에서는 단백질 원본 서열데이터로부터 출현 시퀀스를 추출하는 과정을 기술하였고, 3 장에서는 출현 시퀀스로부터 SVM 분류기를 생성하여 단백질 2 차 구조를 예측하는 마이닝 기법에 대해 설명한다. 4 장에서는 제안한 단백질 구조 분류 모델의 실험 및 결과 분석을 기술하였으며, 5 장에서는 이 논문의 결론을 맺는다.

† 이 논문은 2009 년도 정부(과학기술부)의 재원으로 한국과학재단(R01-2007-000-10926-0)과 2009 년도 정부(교육과학기술부)의 재원으로 한국과학재단(No. R11-2008-014-02002-0)의 지원을 받아 수행된 연구임.

2. 단백질 서열 데이터 추출 및 출현 시퀀스 발견

원시 단백질 서열 및 구조 데이터의 추출과 단백질 서열로부터 suffix tree 구조를 기반으로 출현 시퀀스[6]를 발견한다.

2.1 CATH 데이터베이스 단백질 서열 데이터로부터의 파일 생성 및 클래스 생성

CATH 데이터베이스에서 단백질 서열과 2 차 구조 정보를 추출하여 전처리를 수행한다. CATH 데이터베이스는 <표 1>과 같이 단백질 2 차 구조에 따라 4 개의 계층으로 분류되며 CATH 데이터베이스의 특징인 C-level, A-level, T-level, H-level 중 처음 계층 C-level 에서의 Class1(C1): 주로 Alpha 구조인 시퀀스들, Class2(C2): 주로 Beta 구조인 시퀀스들, Class3(C3): Alpha 와 Beta 구조가 모두 들어 있는 시퀀스들, Class4(C4): 단백질 2 차 구조가 거의 없는 시퀀스들 등 4 가지 클래스 라벨(class label) 따라 단백질 서열 데이터를 분류하였으며 단백질 서열데이터의 ID 와 시퀀스데이터, 그리고 각 단백질 서열이 속해 있는 클래스의 정보로 <표 2>과 같이 데이터를 구성하였다.

<표 1> 각 클래스에 대한 label 처리

Protein ID	***Sequence***	***Class***
1ubaA0	QEKEAIERLKALGFPESLVIQAYFACEKNENLAANFLLSQNFDDE	1.10.8.10 (Class1)
1pkxA1	MKFKTGVAEISNAIDQYVTGTIGEDEDLIKWKALFEEVPEL	1.10.287.440 (Class1)
1gab00	TIDQWLLKNAKEDAIAELKKAGITSKFYNAINKAKTVEEVNALKNEILKAHA	1.10.8.60 (Class1)
1iqjL1	KLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTL	2.10.25.10 (Class2)
…	…	…

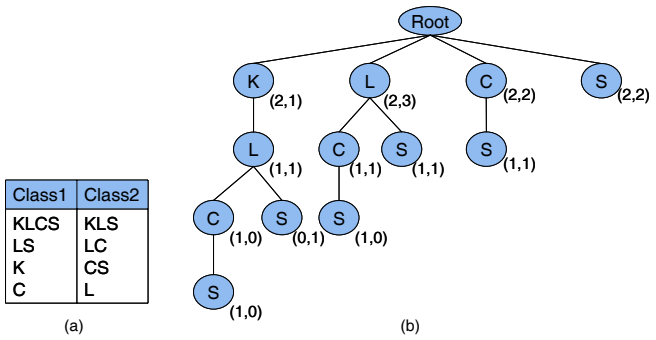
클래스 라벨의 생성 후, 얻어진 각 계층별 서열 데이터로부터 출현 시퀀스를 발견하게 된다.

2.2 단백질 서열 데이터에서의 출현 시퀀스 추출

출현 시퀀스는 KDD(Knowledge Discovery in Database)패턴 중 하나로 새로운 패턴으로서 시퀀스 데이터베이스에서 한 데이터 클래스의 출현 시퀀스는 다른 데이터 클래스와 비교할 때 해당 클래스에서 더 빈발하게 발생하는 부분서열을 말한다[6]. 일반적으로 연관(association)분석에서 자주 발생하는 패턴과는 달리 출현 시퀀스는 높은 구별력(discriminating power)으로 분류 문제에 적용되어 더욱 유용하다고 알려져 있다[6]. 전처리 단계에서 얻은 데이터를 클래스 레벨(class level)에 따라 분류 한 후에 출현 시퀀스 알고리즘의 첫 단계로서 단백질 서열로 MT(merged suffix tree)를 구축한다. MT 구조의 각 노드는 단백질 서열에서의 잔기를 가리키고 각각의 노드는 비교되는 두 개의 클래스 각각의 빈발도(support count)값을 가지며 따라서 이 빈발도 값들로 두 클래스 각각의 성장률(growth rate) 값을 구할 수 있다. 맨 처음, MT 는 root 노드만 가지고 있으며 클래스들의 빈발도 값은 (C1-count, C2-count)=(0, 0) 이다. 단백질 서열 및 그 서열의 부분 서열을 읽으면서 시퀀스 순서에 따라 매개 노드와 비교하여 같을 경우 그 노드의 해당 클래스의 빈발도 값에 1 을 더해주고 다를 경우 그 노드의 다른 부분 트리(subtree)를 만들어준다. 이 과정을 예를 들어 설명하면 (그림 1)과 같다. (그림 1)의 (a)는 단백질 서열데이터이고 (b)는 데이터에 의해 생성된 MT 이다.

<표 2> CATH 데이터베이스 계층별 구성

계층	설명	분류 라벨	설명
C (Class)-level	단백질 총체적인 2 차 구조 상황에 따라 분류	Class1(C1)	주로 α (Alpha)구조로 구성
		Class2(C2)	주로 β (Beta) 구조로 구성
		Class3(C3)	α, β 구조로 구성
		Class4(C4)	2 차 구조가 거의 없음
A (Architecture)-level	도메인 형태에 따라 분류(2 차 구조 연결성 고려하지 않았음)	30 개 클래스 라벨	
T (Topology)-Level	도메인 핵심(core)부분의 폴더(fold)부분의 동일 여부에 따라 분류	379 개 클래스 라벨	
H (Homologous super family)-level	단백질 상동성(homologous)에 따라 분류	637 개 클래스 라벨	



(그림 1) 단백질서열데이터로부터의 MT 구축의 예

다음 단계로 구축된 MT 로부터 출현 시퀀스를 추출한다. 노드들의 각 빈발도(support count) 값은 처음 root 노드로부터 그 노드까지의 선행시퀀스(pre-sequence)의 빈발도 값이다. 트리로부터 생성된 부분 시퀀스에서 두 클래스의 각각에 해당되는 최소 빈발도(minimum support count)값과 최소 성장률(minimum growth rate)을 만족 시키지 못하는 부분 시퀀스를 제거되며, 남은 출현 시퀀스들을 발견한다. 두 개의 서

로 다른 클래스에 해당되는 두 집합 D1, D2 에 대해, 부분 시퀀스(sub-sequence) X 의 D1 에 대한 D2 의 성장률은 (식 1)과 같이 정의 된다.

$$growthRate_{D_1 \to D_2}(S) = \begin{cases} 0 & Supp_{D_1} = 0 \ \ and \ \ Supp_{D_2} = 0 \ \text{일때} \\ \infty & Supp_{D_1} = 0 \ \ and \ \ Supp_{D_2} = 0 \ \text{일때} \\ \dfrac{Supp_{D_1}}{Supp_{D_2}} & otherwise \end{cases} \quad \text{(식 1)}$$

Sub-sequence	Support count		Growth rate of ES	
	Class1	Class2	Class1	Class2
K	2	1	2	
L	2	3		1.5
C	2	2		
S	2	2		
KL	1	1		
LC	1	1		
LS	1	1		
CS	1	1		
KLC	1	0	∞	
KLS	0	1		∞
LCS	1	0	∞	
KLCS	1	0	∞	

$\rho_S = 1$
$\rho_G = 1.5$

Emerging pattern	
Class1	Class2
K	L
KLC	KLS
LCS	
KLCS	

(그림 2) 출현 시퀀스 발견 과정

또한 성장률 임계 값 ρ>1 에 대하여 부분 시퀀스 X 가 growthrate(X)≥ρ 의 성장률을 가질 때, 부분 시퀀스 X 를 ρ-Emerging Sequence(ρ-ES)라고 한다.

(그림 1)에서의 예제로부터 생성된 부분 시퀀스와 해당 클래스의 성장률과 대응하는 출현 시퀀스는 (그림 2) 와 같다.

3. SVM 을 이용한 단백질 구조 예측

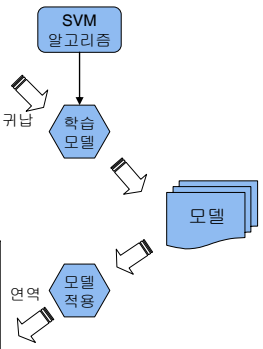
SVM [7]는 최근 주목을 받는 분류 기술 중 하나인데, 이 기술은 통계적 학습 이론에 뿌리를 두고 있으며 숫자 인식에서부터 텍스트 분류에 이르기까지 많은 실제 응용에서 좋은 결과를 보여주었다. 또한, SVM 은 높은 차원을 가지는 생물정보학 데이터의 분석에 적합하며, 기존의 분류기 중에서 성능이 좋은 것으로 알려져 있다.

훈련집합(Training set)

Protein ID	Emerging Sequence	Class
1	$ES_1, ES_2, ES_5, ES_9, ES_{18}, \cdots, ES_{637}$	C1
2	$ES_2, ES_3, ES_{49}, ES_{27}, ES_{79}, \cdots, ES_{637}$	C2
3	$ES_3, ES_4, ES_{30}, ES_{43}, ES_{18}, \cdots, ES_{455}$	C4
4	$ES_1, ES_4, ES_{55}, ES_{39}, ES_{118}, \cdots, ES_{637}$	C2
…	…	…

시험집합(Test set)

Protein ID	Emerging Sequence	Class
1	$ES_3, ES_4, ES_5, ES_{13}, ES_{20}, \cdots, ES_{600}$	?
2	$ES_1, ES_2, ES_3, ES_{26}, ES_{108}, \cdots, ES_{587}$	?
3	$ES_1, ES_2, ES_3, ES_4, ES_{209}, \cdots, ES_{637}$	?
4	$ES_1, ES_2, ES_3, ES_{99}, ES_{318}, \cdots, ES_{633}$	?
…	…	…

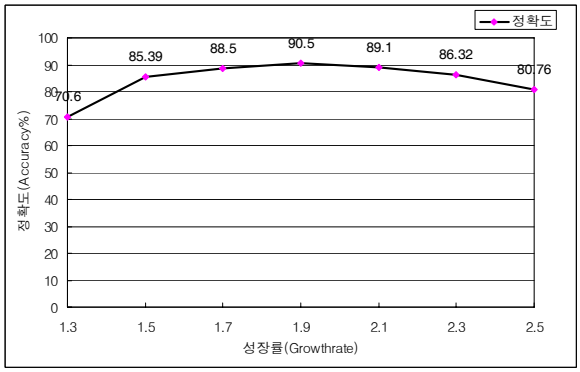


(그림 3) 단백질 데이터로부터의 분류모델 구축 과정

전처리한 단백질 서열데이터에 SVM 분류기법을 적용하기 전에 서열데이터에서 2 장에서 추출한 출현 시퀀스들을 찾아내어 (그림 3)과 같이 단백질 데이터의 분류 속성으로 표현한다. 다음 2 차 구조 클래스 레벨이 알려져 있고 출현 시퀀스로 구성된 단백질 서열 데이터를 훈련 데이터로 하여 분류 모델을 구축하고, 이 분류 모델을 2 차 구조 분류 클래스 레벨을 모르는 출현 시퀀스로 구성된 단백질 서열 데이터로 구성된 테스트 데이터에 적용하여 단백질 서열데이터의 2 차 구조를 예측한다. 단백질 데이터로부터의 분류모델 구축 과정의 예제는 (그림 3)과 같다.

4. 실험 및 평가

이 장에서는 출현 시퀀스로부터 SVM 분류 기법과 기존의 분류 기법 CARM(classification association rule mining)[8]과 결정트리(C4.5)[9], 단순 베이지안 분류기(naive Bayesian)[10]을 적용하여 성능 평가를 하였다. 지표로는 재현률(recall), 정확률(precision)[11]을 측정하였다. 이 실험을 위하여 생성한 데이터 집합은 CATH database 의 총 36,620 개의 단백질 서열 데이터중 20% 에 해당하는 7,324 개의 데이터를 샘플 데이터로 이용하여 실험하였다. 출현 시퀀스 탐사 시 가장 좋은 출현 시퀀스를 찾을 수 있는 성장률 임계값 설정을 위하여 성장률 변화에 따른 분류 결과, (그림 4)와 같이 성장률을 1.9 로 설정 했을 때, 정확도는 90.5 로서 가장 높았다.



(그림 4) 성장률 변화에 따른 정확도

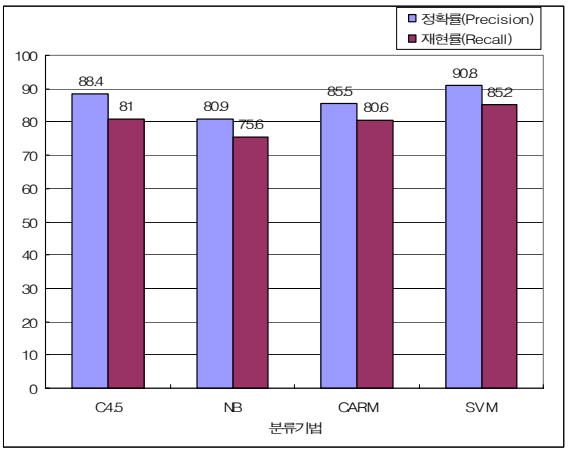
평가 척도로 재현률(recall)과 정확률(precision)을 사용하였는데 표현은 (식 2), (식 3)과 같으며, 재현률은 전체 탐사된 출현 시퀀스 중 목표 항목집합에서 현재 결과 집합이 어느 정도 누락이 없는 출현 시퀀스 집합들로 구성이 되었는지를 나타내 준다.

$$\text{재현률(recall)} : r = \frac{TP}{TP + FN} \quad \text{(식 2)}$$

$$\text{정확률(precision)} : p = \frac{TP}{TP + FP} \quad \text{(식 3)}$$

출현 시퀀스 정확률이 가장 높을 때의 성장률 임계값 1.9 일 때 생성된 출현 시퀀스로부터 SVM, 베이지안 분류기(NB), 결정트리(C4.5), CARM 등 분류 기법

적용에서의 정확률, 재현률에 대해 비교 분석하였으며 (그림 5)와 같다.



(그림 5) 각 분류 모델의 재현률(recall)과 정확률(precision)의 비교

위의 결과, 단백질 시퀀스 데이터의 출현 시퀀스를 분류 속성으로 하여 4 가지 분류 기법을 적용하였는데 SVM 분류기법의 정확률이 90.5%, 재현률이 85%로서 가장 좋은 결과를 보여주었다.

5. 결론

이 논문에서는 CATH 데이터베이스의 단백질 서열 데이터로부터 출현 시퀀스를 생성하고 생성된 출현 시퀀스들로부터 SVM 분류기를 이용하여 분류모델을 구축하여 단백질 2 차 구조 예측을 하는 방법론을 제안하였다. 이 논문에서 제안한 단백질 구조 클래스 예측을 위한 출현 시퀀스 마이닝의 적용은 기존의 서열들을 비교하면서 예측하는 방법과는 다른 접근법으로서 새로운 단백질이 발견 되었을 때, 보다 빠른 분류를 할 수 있는 점에 기여 하였다. CATH database 의 단백질 서열 데이터로 실험을 하였는데 정확률이 90.5%, 재현률이 85% 로서 좋은 결과를 보여주었다.

참고문헌

[1] Michael S. Paterson, Vlado Dancik, "Longest Common Subsequences," Math. Found. Com. Sci., pp. 127-142, 1994

[2] D. J. Lipman, W. R. Pearson, "Rapid and sensitive protein similarity searches", Science, Vol. 227, pp. 1435-1441, 1985.

[3] S. Karlin, S.F. Altschul, "Applications and statistics for multiple high-scoring segments in molecular sequences," Proc. Natl. Acad. Sci, Vol. 90, pp. 5873-5877, 1993.

[4] S. Kimmen, "Phylogenonmic inference of protein molecular function: advances and challenges, " Bioinformatics, Vol. 20, No. 2, pp. 170-179, 2004.

[5] M. Kanehisa & B. Peer, "Bioinformatics in the post-sequences era," Nat. Gene. Supp., Vol. 33, pp. 305-310, 2003.

[6] S. Chan. B. Kao C.L. Yip. "Mining Emerging Substrings", 2003.3

[7] Corimma Cortes and Vladimir Vapnik. "Support-vector networks ", Machine Learning, 20(3):273-297, September 1995.

[8] Yanbo J. Wang, Qin Xin and Frans Coenen, "A Novel Rule Ordering Approach in Classification Association Rule Mining", Machine Learning and Data Mining in Pattern Recognition, pp. 339-348, 2007.

[9] Quinlan, J. R., " C4.5: Programs for Machine Learning", San Mateo, CA: Morgan Kaufmann

[10] N. Friedman, D. Geiger and M. Goldszmidt,"Bayesian Network Classifiers", Machine Learning, pp: 131-163, 2004.

[11] Sergio A. Alvarez, "An exact analytical relation among recall, precision, and classi_cation accuracy in information retrieval", Boston, Technical Report BCCS-02-01, 2002.

[12] N. Friedman, D. Geiger and M. Goldszmidt,"Bayesian Network Classifiers", Machine Learning, pp: 131-163, 2004.

[13] Sergio A. Alvarez, "An exact analytical relation among recall, precision, and classi_cation accuracy in information retrieval", Boston, Technical Report BCCS-02-01, 2002.