EP-tree 마이닝을 이용한 단백질 DISORDER/ORDER 지역 분류

박홍규*, 이헌규*, 이미정**
*한국전자통신연구원 우정물류기술연구부
**충북대학교 데이터베이스/바이오인포메틱스 연구실
e-mail: hkpark@etri.re.kr

Classification of Protein DISORDER/ORDER Region Using EP-tree Mining

Hong Kyu Park*, Heon Gyu Lee*, Meijing Li**
*Dept. of Postal&Logistics Technology, Electronics & Telecommunication Research Institute
**Database/Bioinformactics Lab., Chungbuk National University

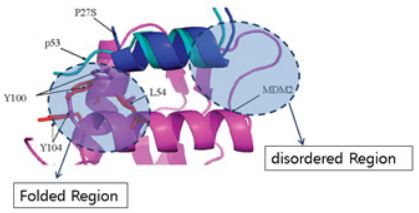
요 약

단백질 1차 서열로부터 DISORDER와 ORDER지역을 예측하기 위해서 이 논문에서는 EP-tree에 기반한 출현패턴 발견 알고리즘을 제안하였다. EP-tree 알고리즘을 적용함으로서 기존의 단백질 특징 추출을 통한 방법과 달리 서열 자체에서 발견되는 출현패턴만을 이용하여 분류 모델을 생성하므로 기존의 신경망이나 SVM 보다 분류모델 생성 및 예측 속도가 빠르다. 또한 Disprot 4.9과 CASP7 테스트 데이터로 DISORDER/ORDER 지역을 예측한 결과, 73.4%의 높은 정확성을 보였다.

1. 서론

생물학분야에서 단백질 구조 및 기능예측에 관한 연구는 매우 중요한 연구이다. 단백질 구조에서 서열형태로 풀려있는 부분을 disorder구역이라 하고 생화학적 반응을 일으켜 접혀져 있는 부분을 order구역이라고 한다(그림 1). 단백질의 disorder구역이 반응을 일으켜 order구조로 변하는 과정에서 단백질은 기능을 하기 때문에 단백질 disorder구역을 찾아내는 것이 단백질 서열로부터 기능예측을 할 수 있는 중요한 부분이다. 또한, 대부분의 hub_protein[1]은 disordered protein보다 쉽게 결합 반응하지 않으므로[1] disorder구역 예측이 단백질 기능 예측에 있어 중요하며 disorder 서열 데이터로부터 단백질 정렬기법으로 시퀀스 상사성 분석을 하게 되는데 단백질 서열에서의 disorder구역과 order구역을 구분하여 주는 것으로 disorder구역과 order구역이 서로 정렬 비교되는 것을 회피하여 분석결과의 더 높은 정확도를 보장할 수 된다. 서열데이터에서는 disorder구역과 order구역을 분리하여 단백질의 3차 구조 또는 3차 구조의 특징 예측을 더 쉽게 진행할 수 있게 한다. 단백질 데이터로부터 disorder 구역 예측을 하는 기존의 예측 프로그램은 주로 sliding window와 feature selection기법을 사용하여 생성된 패턴들에 하나 또는 몇 개의 기계학습을 적용하여 분류한다. 단백질 disorder구역 예측 연구에 많이 쓰이는 기계학습 기법에는 Support Vector Machine (SVM)[2][3], Neural Networks(NNs)[5], Regression[6] 등이 있다. 그밖에 sliding window[7]기법과 random forest machine learning[8] 모델을 이용한 연구 방법도 있다. 많은 예측 프로그램이 현재 발표되어 있고 신뢰도가 높은 예측 프로그램도 많지만, 친수성, B-factor 포함량, 특정위치수(position-specific score) 프로파일 등 단백질의 disorder구역 특성들과 SVM, NNs 등 서로 다른 알고리즘을 사용하기에 예측 모델을 구축하는데 쓰이는 데이터가 다름에 따라 disorder구역도 차이가 있다.

이 논문에서는 기존의 단백질 특징 추출을 통한 기능예측이 아닌 순수 단백질 1차 서열에서 EP-tree 마이닝 기법을 적용하여 각각 disorder 구역과 order 구역에 속한 부분 서열들을 찾아내는 방법을 제안한다. 제안한 서열 기반의 방법은 예측 결과가 단백질 disorder 서열의 어떤 한 특성에 치우치지 않는다. 단백질 disorder 예측을 위해 이 논문에서는 첫째, 기존의 단백질 서열 데이터베이스에서 disorder 서열데이터와 order 서열데이터를 수집 및 분류정리한다. 둘째, 단백질 disorder 구역 예측 모델 구축을 위해 단백질 disorder 출현패턴 서열과 order 출현패턴 서열을 각각 생성한다.



(그림 1) 단백질의 disorder와 order 구역

1) 반응을 일으켜 결합된 단백질 파트너가 많은 단백질

마지막으로 예측하려는 단백질 서열에 sliding window와 트리 기반의 출현패턴 마이닝 기법을 적용하여 disorder 출현패턴과 order 출현패턴으로 구성된 부분 서열을 검색하는 방법으로 disorder 구역과 order 구역을 예측 분류한다.

논문의 구성은 다음과 같다. 2절에서는 데이터 소개로부터 데이터 추출, 출현패턴 생성을 위한 EP-tree 알고리즘, order 출현패턴 가지치기(pruning) 및 단백질 구역 예측 방법에 대해 단계별로 기술하였으며, 3절에서는 기존의 주된 예측 프로그램과의 예측 결과에 대해 비교 평가 하였다. 마지막 4절에서는 연구 내용에 대한 결론을 맺는다.

2. 방법론

이 논문에서는 single_classifier로 제안된 출현 시퀀스 생성 알고리즘을 double-classifier로 변형시키고 단백질 시퀀스 데이터의 성질에 적합하도록 출현 시퀀스 생성과정에서 사용되는 파라미터에 대해 새롭게 정의 하였다. 예측단계에서 sliding window 기법과 출현패턴 마이닝의 분류 단계에서 사용되는 스코어 계산을 적용하여 disorder구역을 예측하였다.

2.1 EP-tree에 의한 출현 패턴 마이닝

단백질 disorder 구역 예측은 이미 알려져 있는 단백질 서열로부터 order 지역보다 target 클래스인 disorder 구역에서 발생 빈도가 높은 부분 서열 패턴들을 찾아내는 것이다. 이때 하나의 단백질 서열에서 disorder 구역이 여러 개 있을 수 있고 하나의 disorder 구역에도 중복된 부분 서열이 있을 수 있기 때문에[9] 지지도(support)에 대한 정의와 같이 측정하려는 부분 서열을 포함한 클래스에 해당된 서열의 개수를 지지도 값으로 하면 disorder/order구역의 서열 개수보다 많을 수 있다. 따라서 출현패턴의 지지도 외에 추가적으로 성장률(GR : growthrate) 임계값을 적용한다.

이 논문에서는 최소지지도(δ) 및 growth rate(ρ)를 만족하고, 중복 패턴이 제거된 필수 출현 패턴 마이닝 알고리즘을 제안한다.

제안된 EP-tree의 구조는 다음과 같다.

① 하나의 루트와 자식 노드인 prefix, subtree, 해더 테이블로 구성된다.

② 트리의 각 노드에서는 item-name, $count_1$, $count_2$, node-link를 가진다. $count_1$은 D_1 트랜잭션에서의 항목의 수이고 $count_2$는 D_2에서의 항목의 수이다.

③ 해더테이블의 각 엔트리에는 item-name, head of node-link, count1, count2 정보를 갖는다. count1, count2는 node-link에 연결된 각 항목들의 모든 개수의 합니다.

④ EP-tree와는 다르게 트리의 운행은 top-down 방식이며, EP-tree는 오름차순 순서를 유지 한다[2].

FP-tree의 경우, 첫 단계에서 1-항목에 대한 내림차순을 기준으로 트리가 구성되나 제안된 EP-tree에서는 오름차순을 기준으로 구성된다. 모든 항목에 대한 트리 구성시, 오름차순으로 구성을 하게 되면 트리의 가지가 더 늘어나며, 느린 트리 운행이 되는 단점을 가진다. 그러나 chi-square 검정을 통해 많은 불필요한 패턴의 제거를 EF-growth 단계에서 동시에 수행할 수 있는 장점을 가지며, 중복 패턴이 될 수 있는 후보 패턴을 생성하지 않는 장점을 가진다.

<표 1>의 D1, D2에서의 출현 패턴 마이닝을 위한 EP-tree의 구성 단계는 다음과 같다.

1단계 : D1, D2의 항목들 중에서 최소지지도, δ을 만족 못하는 모든 항목을 제거한다.

2단계 : 정의 4의 Pattern ranking에 의해 SupportRatio 기준으로 모든 항목들은 오름차순 정렬된다.

3단계 : <표 1> 트랜잭션의 항목을 오름차순 순서를 고려하여 EP-tree를 구성한다.

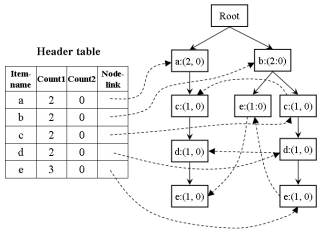
<표 1>의 데이터를 < 에 의해 순서화[3] 한 결과는 <표 2>이고, 생성된 EP-tree는 (그림 3)과 같다. (그림 2)는 D1 데이터 집합의 모든 항목들이 삽입되었을 때까지의 구성된 트리이고 (그림 3)은 D2 데이터 집합의 모든 항목까지 삽입한 완성된 트리이다.

<표 1> 두 클래스를 갖는 데이터의 예

D_1					D_2				
a		c	d	e	a	b			
a							c		e
	b				a	b	c	d	
	b	c	d	e				d	e

<표 2> 오름차순 정렬에 의한 트랜잭션 데이터 집합

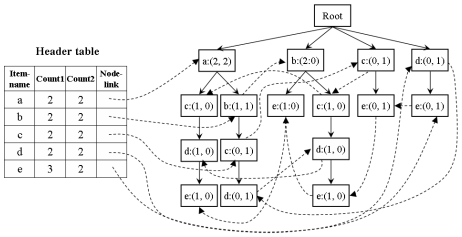
ID	Class	Ordered Itemset
100	D_1	a c d e
200	D_1	a
300	D_1	b e
400	D_1	b c d e
500	D_2	a b
600	D_2	c e
700	D_2	a b c d
800	D_2	d e



(그림 2) D_1의 항목들을 삽입한 트리

2) 노드 R이 노드 N의 부모노드이고, 항목 i, j가 두 노드에 존재한다면, 두 항목은 오름차순 i<j를 유지한다.

3) <표 5>의 모든 항목들에 대한 순서는 e, a, b, c, d이다.



(그림 3) D_1, D_2 데이터의 모든 항목을 삽입한 완성된 EP-tree

2.2 단백질 disorder / order 출현패턴 서열 추출

출현패턴 서열 추출 과정은 다음과 같다.

첫 번째 단계는 먼저 disorder 서열 데이터로부터 EP-tree를 구축한다. 두 번째 단계는 disorder 지역 출현패턴과 order 지역 출현패턴을 추출하는 단계로서 구축된 EP-tree로부터 각각의 disorder와 order 두 클래스에 해당되는 빈발도와 성장률 두 개의 파라미터 값을 기준으로 출현패턴 서열을 추출한다.

2.3 단백질 시퀀스 데이터에서의 disorder 구역 확정

생성된 disorder 출현패턴 서열과 order 출현패턴 서열에 score값을 기준으로 단백질 서열에서 disorder 구역을 예측하는 과정은 disorder 출현 시퀀스를 이용하여 구역 예측하는 단계와 order 출현 시퀀스를 사용하여 1차적으로 예측한 결과에 대해 정제해 주는 두 단계로 구분된다.

Disorder구역 예측에서 disorder와 order 시퀀스 데이터로 추출한 패턴의 개수가 평형 되지 않는 문제점을 해결하게 위해 norm_score를 적용하였다.

$$score(s,C)=\sum_{e\subseteq s, e\in E(C)}^{pat_s}(e)\cdot\frac{growthrate(e)}{growthrate(e)+1} \qquad 식(1)$$

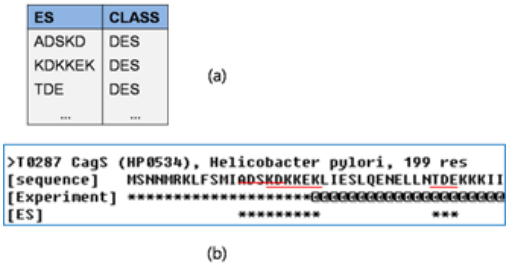
$$norm\_score=\frac{score}{median\_score} \qquad 식(2)$$

예측하려는 단백질 서열에서 sliding window 방법으로 disorder 패턴을 찾아내어 그 단백질 서열의 disorder 구역으로 하고 order 패턴을 찾아내어 order 구역으로 예측한다. 이 과정의 예는 (그림 4)와 같다.

(그림 4)의 (a)는 예측에 사용된 disorder_EX들이고, (b)는 예측 결과로서 "*"는 disorder구역이며 "@"는 order 구역을 나타낸다. Experiment는 생물학 실험으로 얻은 결과이며 패턴은 출현 패턴 서열을 적용하여 예측한 결과를 가리킨다.

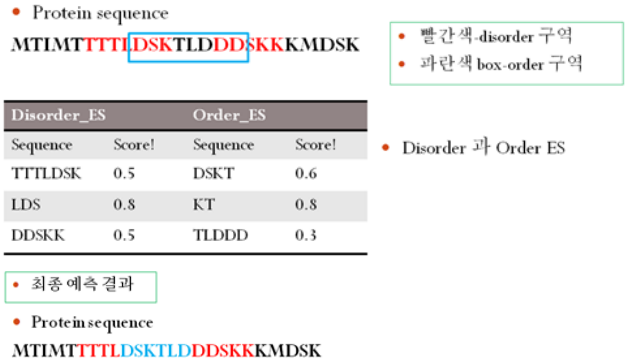
이런 과정에서 disorder 패턴과 order 패턴으로 예측한 부분이 겹치거나 disorder 패턴과 order 패턴을 모두 포함하지 않은 부분이 존재 할 수 있다.

이 논문의 최종 연구 목적은 disorder구역 예측으로 disorder구역 예측의 정확도를 보장하는 것이 가장 중요하다. Disorder 구역 예측의 정확도를 보장하기 위하여 (1) disorder 패턴으로 예측된 부분을 disorder 구역으로 예측하고 (2) 겹치는 부분에 대해서는 그 부분에 포함되는 disorder 패턴, Order 패턴들과 norm_score를 계산, 비교하여 norm_score가 더 큰 클래스로 예측, 분류된다.



(그림 4) disorder구역 예측 예

(그림 5)는 겹친 부분에 대한 간단한 계산, 예측과정이다.



(그림 5) Score값을 적용한 disorder구역 예측의 예

예제에서 disorder와 order 구역이 겹치는 부분이 부분서열은 "DSK"와 "DD" 두 개가 있다. 이 두개의 부분서열에 대해 계산한 score 의 값은 아래와 같다.

· 겹친 부분 1: DSK
  - Score(D_ES)=0.5+0.8=1.3("TTTLDSK"와 "LDS")
  - Score(O_ES)=0.6+0.8=1.4("DSKT"와 "KT")
  DSK는 order 구역이다
· 겹친 부분 2: DD
  - Score(D_ES)=0.5 ("DDSKK")
  - Score(O_ES)=0.3 ("TLDDD")
  DD는 disorder 구역

Disorder 출현 패턴 서열로 1차 예측한 결과를 order 출현 패턴 서열로 정제하여 잘못 예측된 disorder 구역을 제거한다.

3. 실험 및 평가

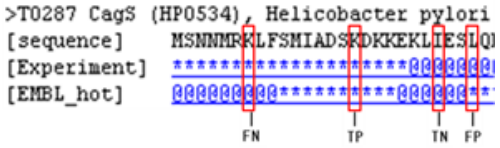
3.1 데이터

단백질 disorder 분류예측 모델의 실험평가를 위한 training dataset로 disorder sequence dataset과 order sequence dataset을 사용하였다. disorder dataset은 Disprot(version 4.9)[10]에서 추출하였고 order dataset은 PDB[11]에서 추출하였다. disorder dataset와 order dataset의 모든 서열 데이터는 중복되지 않고 25% 보다 높은 pair-wise identity을 가지며 disorder sequence dataset은 523개 시퀀스 데이터, 1195개 구역, 67555개 잔

기를 포함하고 있으며, order sequence dataset은 order 구조가 명확하고 길이가 최소 80개 잔기인 시퀀스들을 추출하였고 290개 구역, 67555개 잔기가 포함된다. 데이터 크기의 문제 때문에 생기는 과잉적합을 방지하기 위하여 Test dataset로 CASP 7의 96개 단백질 서열 중 36개 서열과 Disprot (Version 4.9)에서의 10% 샘플데이터인 40개의 단백질 서열을 사용하였다. CASP 7[12]을 단백질 disorder 구역 예측하는 test dataset으로 많이 쓰는데 CASP 7의 단백질 서열에는 disorder 시퀀스가 너무 적게 포함되어 있으므로 긴 disorder 구역을 많이 포함하고 있는 Disprot 에서 샘플 데이터로 40개의 서열 데이터를 추출하여 총 76개 서열이 포함되어 있다.

3.2 실험 및 결과

단백질 disorder구역 예측 기법 평가를 위해 민감도, 특이도, 정확률, 정확도를 비교한다. (그림 6)은 평가항목의 정량적 평가를 위해 선행된 TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative)의 결과의 예를 보여주었다.



(그림 6) disorder 구역 예측에서의 TP, TN, FP, FN

기존에 공개된 예측 프로그램에 같은 test dataset을 이용한 결과와 이 논문에서 제안한 예측 프로그램과의 결과를 비교한 실험평가 결과는 <표 2>와 같다. <표 2>은 매개의 단백질 서열을 기준으로 예측하였을 때의 각 disorder 예측 프로그램의 결과이다.

<표 2> disorder 예측 프로그램의 비교 결과

	Sensitivity	Specificity	Precision	Accuracy
EMBL_hot	50.3	40.9	60.4	61.1
EMBL_coil	69.4	28.5	59.1	63.1
EMBL_remark	25.8	51.1	53	59.8
ES	67.9	32.1	61	73.4

실험 비교 결과로 보면 <표 2>과 같이 특이도를 제외한 모든 평가 항목에 대해 다른 알고리즘보다 성능이 우수하다.

4. 결론

이 논문에서는 EP-tree 구조의 출현패턴 개념을 단백질 서열데이터에 적용하여 order 구역과 disorder 구역을 분류하였다. 이 연구에서 단백질 disorder 구역 예측에 복잡한 생물학적 특징 추출법을 이용하지 않고 단지 단백질 서열만으로 예측 하였음에도 높은 정확도를 보여 주었다. 제안한 방법은 긴 서열의 disorder에도 적용할 수 있으며 분류 모델 구축 과정이 서열 패턴 그 자체를 이용 하므로 대용량의 데이터 분류에도 적합하다.

참고문헌

[1] Maslov S., Sneppen K. “Specificity and stability in topology of protein networks” Science 296 2002, 910-913.

[2] Jones DT, Ward JJ “Prediction of disordered regions in proteins from position specific score matrices” Proteins 2003, 53:573-578.

[3] Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z “Length dependent prediction of protein intrinsic disorder” BMC Bioinformatics 2006, 7:208.

[4] Thomson R, Esnouf R “Prediction of natively disordered regions in proteins using a bio-basis function neural network” LNCS 3177 2004:108-116.

[5] Cheng J, Sweredoski M, Baldi P “Accurate prediction of protein disordered regions by mining protein structure data” Data Mining and Knowledge Discovery 2005:213-222.

[6] Liu J, Tan H, Rost B “Loopy proteins appear conserved in evolution” J Mol Biol 2002, 322:53-64.

[7] Prilusky, J., C. E. Felder, T. Zeev-Ben-Mordehai, E. H. Rydberg, O. Man, J. S. Beckmann, I. Silman, and J. L. Sussman. 2005. FoldIndex “a simple tool to predict whether a given protein sequence is intrinsically unfolded” Bioinformatics. 21:3435 - 3438.

[8] Han P, Zhang X, Feng ZP “Predicting disordered regions in proteins using the profiles using amino acid indices” BMC Bioinformatics 2009.10(Suppl 1):S42

[9] G. Dong, X. Zhang, L. Wong, and J. Li. “Classification by Aggregating Emerging Patterns” Proceedings of the 2nd Int'l Conference on Discovery Science(DS'99), pp.30-42,1999.

[10] Vucetic S, Obradovic Z, Vacic V, Radivojac P, Peng K, Iakoucheva LM, Cortese MS, Lawson JD, Brown CJ, GSikes J, Newton CD, Dunker AK: DisProt: “A Database of Protein Disorder” Bioinformatics 2005, 21:137-140.

[11] Hobohm U, Sander C “Enlarged representative set of protein structures” Protein Sci 1994, 3:522

[12] Moult J., Fidelis K., Zemla A., Hubbard T. “Critical assessment of methods of protein structure prediction (CASP)-round 5” Proteins 2003, 53:334-339.