단백질 상호 작용 예측을 위한 SVM의 부정예제 생성방법론*

김철환, 정유진
한국 외국어 대학교 컴퓨터 공학과
{redfoot, chungyj}@hufs.ac.kr

Negative example generation methods of SVM for predicting protein-protein interactions

Chul-hwan Kim, Yoojin Chung
Dept. of Computer Eng. Hankuk Univ. of Foreign Studies

요 약

생명체의 기본 정보가 저장된 DNA에서 생성되는 단백질은 생명 현상의 중요한 기능적 역할을 수행하기 때문에 단백질과 관련된 다양한 연구가 진행되고 있다. 본 논문에서는 단백질간 상호작용(protein-protein interaction)을 예측하기 위해 시스템을 통계학적 모델인 Support Vector Machine(SVM)을 사용하였다. SVM 시스템은 상호작용이 있는 데이터(긍정예제)와 상호작용이 없는 데이터(부정예제)를 입력으로 하여 모델링 생성과 테스트를 하는데, 상호작용이 있는 데이터는 DIP에 있는 interaction list로 해결이 가능하지만 상호작용이 없는 데이터는 현재 존재하지 않기 때문에 이를 생성하기 위한 생성방법이 필요하다. 이 논문에서는 shuffling, non-interaction list, 그리고 앞의 두 방법을 보완하는 non-interaction list + shuffling이라는 방법을 제시하고 기존의 실험 결과를 상회하는 부정예제 생성방법을 제시한다.

1. 서론

생명체의 기본 정보가 저장된 DNA에서 생성되는 단백질은 생명 현상의 중요한 기능적 역할을 수행하기 때문에 단백질과 관련된 다양한 연구가 진행되고 있다. 물론 이에 따라 단백질 관련 DB가 구축되고 있으며(PDB, PIR, DIP, Swiss-prot[1-4]) 단백질 구조와 기능 예측, 상호작용 예측 등 다양한 실험에서 이러한 데이터가 활용 중이다.

그 중에서도 단백질간 상호작용은 세포외부의 호르몬 결합에서부터 신호전달 과정을 통한 특정 수용체로 이르기까지, 단백질 분자들은 유전자 전사와 대사와 같은 세포 내부의 변화에 영향을 끼치는 세포 외적 정보를 전달하는 커다란 역할을 하고 있다. 또한, 단백질 간 상호작용의 결과는 단백질 분자들이 어떻게 신호를 주고 받는지를 이해하는데 필요한 기본자료를 제공해줄 뿐 아니라, 제약이나 의학적으로 중요한 정보의 제공으로 이어지게 된다. 우리는 이러한 단백질 간 상호작용의 중요성의 이해와 함께 축적되어 가는 단백질 정보들간 상호작용을 예측할 수 있도록 예측시스템에 대한 실험을 진행했다. 예측시스템에 적용된 SVM[5,6]시스템은 training 및 classifying을 할 때 입력하는 데이터가 긍정예제와 부정예제로 나뉘게 되고 긍정예제는 DIP에 있는 interaction list로 해결이 가능하지만, 부정예제는 데이터가 현재 존재하지 않기 때문에 이를 생성하기 위한 방법이 필요하다.

* 본 연구는 한국과학재단 목적기초연구 (R01-2003-000-10860-0) 지원으로 수행되었음

기존의 Bock과 Gough[7]의 연구에서는 부정예제를 단백질 시퀀스의 randomization으로 작성하였지만, 좀 더 다양한 테스트 데이터와 좋은 결과를 얻기 위해 부정예제 생성 방법을 제시한다.

2. 실험 개론

2.1. Support Vector Machine

상호작용 예측 시스템에 적용된 SVM은 기본적으로 두 범주를 갖는 객체들을 분류하는 방법이다. 이는 우리가 목적으로 하고 있는 상호작용 예측 시스템의 상호작용이 " 있는가" , " 없는가" 를 구분하는데 있어 적절한 모델이며 최근에 들어 그 성능을 인정받아 다양한 분야의 예측 시스템에서 적용이 되고 있는 개념이다[5,6].

2.2 실험 과정

기존의 Bock과 Gough의 연구와 같이 우리의 예측 시스템도 단백질의 1차 구조로부터 상호작용을 예측한다. 단백질의 1차 구조는 단백질을 구성하는 아미노산 서열을 말하는데, 우리는 아미노산이 갖고 있는 여러 가지 특성 중 소수성(hydrophobicity)만을 실험에 적용 하였다.

단백질의 시퀀스에서 아미노산이 소수성을 띠게 되면 시퀀스에서의 아미노산 위치에 1을 표시한다. 예를 들어 [그림 1]의 각 행과 같은 리스트를 구성하게 된다.

실험에 적용할 데이터는 DIP에서 제공하는 yeast종의 단백질을 사용 하는데, 예측 시스템에서 사용한 TinySVM[8]은 실험 데이터로 긍정예제와 부정예제를 입력데이터로 사용한다. 상호작용을 예측할 두 단백질의 소

수성을 표현한 시퀀스에 대해 concatenation한 다음, [그림 1]과 같이 단백질간 상호작용이 있는 긍정예제는 concatenation 된 시퀀스 앞에 +1을 붙이고 상호작용이 없는 부정예제는 시퀀스 앞에 -1을 붙여서 TinySVM의 입력데이터로 사용한다.

```
+1 1:1 2:1 3:1 4:1 6:1 9:1 11:1 12:1 14:1 15:1 16:1 1
+1 1:1 2:1 3:1 4:1 6:1 9:1 11:1 12:1 14:1 15:1 16:1 1
+1 1:1 4:1 5:1 6:1 7:1 8:1 9:1 10:1 11:1 12:1 15:1 17
+1 1:1 4:1 5:1 6:1 7:1 8:1 9:1 10:1 11:1 12:1 15:1 17
+1 1:1 4:1 8:1 9:1 10:1 11:1 15:1 16:1 17:1 20:1 22:1
-1 2:1 3:1 4:1 6:1 9:1 11:1 12:1 13:1 14:1 16:1 17:1
-1 2:1 3:1 4:1 6:1 9:1 11:1 12:1 13:1 14:1 16:1 17:1
-1 2:1 3:1 4:1 6:1 9:1 11:1 12:1 13:1 14:1 16:1 17:1
-1 1:1 2:1 3:1 4:1 5:1 6:1 7:1 8:1 9:1 12:1 13:1 14:1
-1 1:1 2:1 3:1 4:1 5:1 6:1 7:1 8:1 9:1 12:1 13:1 14:1
-1 1:1 2:1 3:1 4:1 5:1 6:1 7:1 8:1 9:1 12:1 13:1 14:1
-1 1:1 2:1 3:1 4:1 5:1 6:1 7:1 8:1 9:1 12:1 13:1 14:1
```

[그림 1] 입력 데이터

여기서 긍정예제는 DIP의 interaction list를 사용하고, 부정예제는 현재 상호작용이 없는 데이터가 존재하지 않기 때문에 여러 가지 부정예제 생성방법을 제시하고 실험하였다.

3. 생성 방법

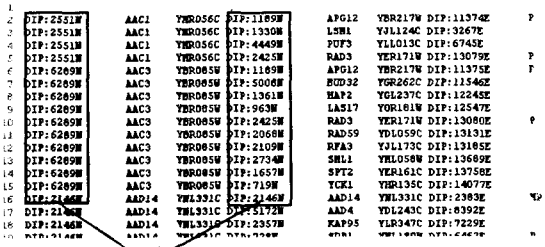
3.1 Shuffling

Shuffling은 돌연변이의 법칙을 적용하는데, 돌연변이는 DNA시퀀스에서 하나 또는 그 이상의 아미노산의 위치가 바뀌게 되는 유전자 변형의 과정으로, 새로운 단백질이 생성되게 되는데 이는 단백질 시퀀스의 하나, 또는 그 이상의 아미노산 서열이 바뀌는 것을 의미한다. 실험에 적용할 단백질의 시퀀스에서 하나의 아미노산을 선택하여 시퀀스 내의 다른 아미노산과 바꾸는 방법으로 local shuffling을 적용하였다. 여기에 덧붙여 특정 아미노산 하나가 아닌 2개, 3개 이런 식으로 shuffling을 적용할 아미노산의 개수를 늘리며, 추가 실험을 수행하였다.

두 번째로 사용한 global shuffling 방법은 단백질이 갖고 있는 전체 시퀀스에 대해 구성하고 있는 아미노산의 개수 및 전체 시퀀스 길이는 그대로 유지한 채, 완전히 새로운 시퀀스 서열이 구성되도록 전체 시퀀스를 재배열하는 방법으로 새로운 단백질 시퀀스를 생성하였다.

3.2 Non-interaction list

Non-interaction list방법은 DIP에 있는 interaction list에 없는 쌍을 부정예제로 활용하는 방법이다. DIP의 interaction list([그림 2])에서 서로 상호작용하는 단백질을 나타내는 첫 번째 열과 세 번째 열의 단백질 번호에서 존재하지 않는 상호작용의 쌍들을 non-interaction list로 작성, 이를 실험 데이터로 활용하는 방법을 사용하였다. 이는 현재 존재하는 단백질간의 interaction의 쌍의 여집합을 예측 시스템의 부정예제로 활용함으로써 실제의 단백질 쌍을 SVM모델에 적용 가능하게 했다. 그리고 단백질의 시퀀스는 현재 존재하는 yeast의 데이터를 그대로 사용하였다.



존재 하지 않는 쌍들의 리스트를 실험 부정예제로 활용

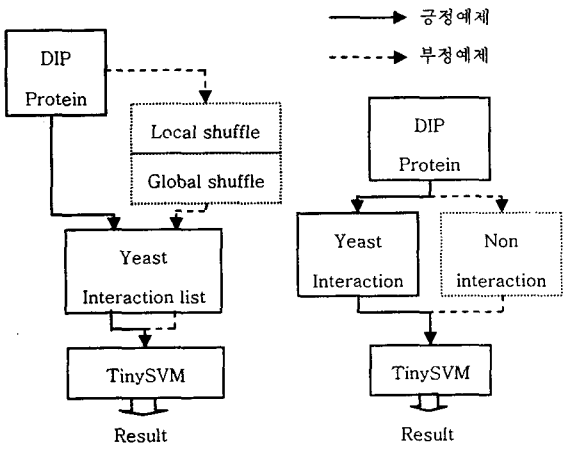
[그림 2] Yeast의 상호작용 리스트

3.3 Non-interaction list + shuffling

마지막으로 고안한 방법은 3.2에서 만든 non-interaction list를 이용하여 3.1에서 사용했던 global shuffling 방법을 적용, 양 방법에서 부족할 수 있는 부분을 보완할 수 있도록 두 가지 방법을 혼합하여 실험에 적용하였다.

4. 실험 절차 요약

앞 절에서 설명한 shuffling과 non-interaction list의 실험절차 [그림 3]를 통해 도식화 하였다.



Shuffling 실험 절차, Non-interaction list 실험 절차

[그림 3] 다양한 부정예제 생성 방법

[그림 3]에서 shuffling은 DIP의 단백질 시퀀스의 서열을 조작하는 방법으로 부정예제를 생성했으며, non-interaction list를 이용한 방법에서는 DIP의 단백질 시퀀스 정보는 그대로 사용하고 interaction list에 없는 non-interaction list를 작성하여 부정예제를 구한다.

5. 실험 및 평가

실험은 DIP에서 제공하는 yeast종의 상호작용 리스트를 2000개씩 4개의 set으로 나누고 각 생성방법에 따라 생성한 데이터 역시 2000개씩 4개의 set으로 나누었다.

그래서 각각의 생성 방법에 따라 4000개씩 4개의 set, 16000개씩 총 64000개의 실험 데이터에 대한 실험이 이

루어 졌다.

실험 결과는 [표 1]과 같다. 즉, 한 개에서 다섯 개까지의 아미노산을 shuffling하여 생성한 데이터를 이용한 local shuffling의 실험결과, 전체 시퀀스에 대해 shuffling하여 생성한 데이터를 이용한 global shuffling 실험결과, interaction쌍에 없는 non-interaction list의 실험결과, 그리고 non-interaction list + shuffling의 실험 결과의 평균을 각각 구하였다.

[표 1]에서 Accuracy, Precision, Recall의 정의는 다음과 같다.

- Accuracy = (pp+nn) / (pp+pn+np+nn)
- Precision = pp / (pp+pn)
- Recall = pp / (pp+np)

여기서 pp와 nn은 각각 true positive와 true negative를 나타내고 시스템의 상호작용 예측인 " 있다" 와 " 없다" 가 각각 맞는 개수를 나타내며, pn과 np는 false positive와 false negative로서 시스템의 상호작용 예측인 " 있다" 와 " 없다" 가 각각 틀린 개수를 나타낸다.

[표 1] 각 생성 방법론에 따른 실험 결과

생성 방법	항목	결과(평균 %)
Local Shuffling	Accuracy	85.29
	Precision	98.59
	Recall	72.01
Global Shuffling	Accuracy	92.93
	Precision	95.04
	Recall	90.65
Non-interaction List	Accuracy	51.70
	Precision	53.84
	Recall	22.49
Non-interaction list + Global Shuffling	Accuracy	94.36
	Precision	98.48
	Recall	90.10

6. 결론 및 향후 과제

[표 1]에서 shuffling을 이용하여 생성한 실험 데이터가 좋은 수행결과를 보여주고 있으며, 하나 또는 여러 개의 아미노산 서열을 변경하는 Local shuffling보다는 전체 아미노산 서열을 다시 배열하는 Global shuffling 방법이 더욱 뛰어난 결과를 보여 주었다. 이는 기존의 Bock과 Gough의 실험에서 사용한 randomization을 통한 shuffling방법의 결과인 accuracy 80%정도의 수치보다 높다.

하지만, interaction list에 존재하지 않는 non-interaction list의 데이터로 상호작용 예측을 하였을 때는 그리 만족할 만한 결과가 나오지 않았다. 이는 현재 interaction list에 없는 단백질 쌍이라고 해서 꼭 list에 없는 단백질 쌍간에 상호작용이 없다고 할 수 없기 때문에 SVM에서 training이 제대로 되지 않았다고 추측된다. 위의 두 방법의 문제점을 보완하기 위해 non-interaction list와 global shuffling방법을 같이 사용한 부정예제 생성 방법이 가장 좋은 실험 결과를 보여 주었다.

그리고 앞으로는 좀 더 신뢰적인 상호작용 예측 시스템의 구현을 위해 정확한 non-interaction list의 확보가 요구되며, shuffling의 방법을 단백질의 고유한 특성이 유지될 수 있도록 domain단위 등을 기반을 두는 방법도 연구해야 할 것이다

References

[1] PDB (Protein Data Bank)
http://www.rcsb.org/pdb/
[2] PIR (Protein Information Resource)
http://pir.georgetown.edu/
[3] DIP (Database of Interaction Proteins)
http://dip.doe-mbi.ucla.edu/
[4] Swiss-Prot (Protein knowledgebase)
http://us.expasy.org/sprot/
[5] Vapnik, V., The Nature of Statistical Learning Theory, Springer-Verlag, New York. (1995)
[6] Burges, C.J.C., " A tutorial on support vector machines for pattern recognition" , Data Mining and Knowledge Discovery, 2, 121-167, (1998).
[7] Joel R. Bock and David A. Gough, " Predicting protein-protein interactions from primary structure" , Bioinformatics vol.17 455-460, (2001).
[8] TinySVM
http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM