다양한 아미노산 속성을 이용한 단백질 상호작용 예측*

최일영[0], 정유진
한국외국어대학교 컴퓨터공학과
bugsoda@hanmail.net[0], chungyj@hufs.ac.kr

Predicting Protein-Protein Interactions Using Various Amino Acid Properties

Illyoung Choi[0], Yoojin Chung
Dept. of Computer Engineering, Hankuk Univ. of Foreign Studies

요 약

이 논문에서는, 단백질의 상호작용을 다양한 아미노산의 속성과 Support Vector Machine(SVM)을 사용하여 예측하였다. SVM을 사용한 단백질 상호작용의 예측 시스템에 단백질 상호작용에 중요한 작용을 하는 아미노산의 속성을 사용하고 있다. 이번 실험은 9가지의 아미노산의 속성의 조합 즉, 511(2^9-1)가지의 아미노산 속성을 SVM 학습데이터로 사용하여 예측시스템의 결과를 비교한다. 실험에는 Database of Interacting Proteins(DIP)를 사용하였다. 실험을 위하여 DIP의 H.pylori를 학습용데이터로 사용하고, E.coli를 예측데이터(검증데이터)로 사용하였다. 실험에 따르면 H.pylori의 학습데이터와 E.coli를 예측데이터의 가공에 '소수성'을 사용한 방법보다 '방향성'을 사용한 방법이 더 높은 수치를 나타냈다.

1. 서 론

최근 Support Vector Machine(SVM)[1]을 단백질 상호작용의 예측에 활용하기 위한 연구가 활발하다[2]. SVM의 학습데이터와 예측데이터에 따라 예측결과가 크게 차이가 나는데, 이 논문에서는 '소수성'이 특정 대상에 국한하여 높은 예측율을 나타냄을 보이고, 9가지 아미노산의 속성을 조합한 511가지 속성(2^9-1)의 조합을 예측시스템에 적용하였다. 이 실험의 결과로 '방향성'을 사용한 예측시스템이 높은 확률을 보였다.

이 논문은 다음과 같이 구성되어 있다. 2절에서는 511가지 아미노산의 속성조합을 적용한 시스템의 설계와 구현에 대해 설명하고 3절에서는 H.pylori와 E.coli를 시스템에 적용한 실험과 결과를 기술한 후 마지막으로 결론을 기술한다.

2. 설계 및 구현

구축할 예측 시스템은 SVM과 Database of Interacting Proteins(DIP)[3]에서 가져온 단백질 상호작용 데이터와 전체 단백질 시퀀스를 사용한다. Fasta format(20개의 아미노산을 알파벳으로 표기하는 방식)으로 된 단백질 시퀀스와 H.pylori의 단백질 상호작용데이터, E.coli의 단백질 상호작용데이터를 사용하고 그 중 H.pylori의 단백질 상호작용데이터를 SVM학습용 데이터로 사용하고, E.coli의 단백질 상호작용데이터는 만들어진 예측 시스템에 적용하여 예측율을 알아보기 위해 사용한다. 상호작용데이터에서 단백질의 DIP고유번호를 전체 단백질 시퀀스에서 찾아 DIP고유번호로 된 단백질 상호작용 쌍을 단백질 시퀀스의 쌍으로 변환한다. 분석시스템에 아미노산의 속성의 511가지 조합을 적용하기 위해 자동으로 아미노산속성의 조합을 만들어 내고 그에 맞는 데이터를 만들어내는 프로그램을 사용했다. <그림 1>은 예측시스템의 데이터를 가공하는 방법을 간단하게 나타낸 그림이다.

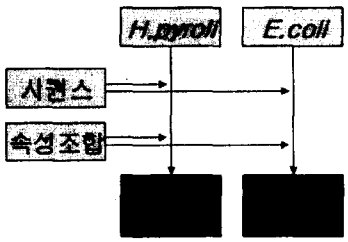

그림 1. 데이터의 가공

2.1 SVM 예측 시스템

이 논문에서 사용된 SVM 프로그램은 'Tiny SVM'[1]을 사용하였고, 예측을 위한 데이터는 종간의 유사도가 높은 것으로 알려진 H.pylori와 E.coli를 DIP에서 가져와 사용하였다. '소수성'을 사용한 예측시스템이 'yeast'를 대상으로한 실험에서 높은 예측율을 보였는데 같은 시스

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

템의 학습과 분석대상을 바꾸고, SVM 학습 및 검증 데이터의 가공시에 511가지의 속성을 적용하였다.

2.2 데이터의 가공

DIP에 존재하는 H.pyroli와 E.coli의 단백질 상호작용 데이터는 단백질 시퀀스를 포함하고 있지 않기 때문에 전체 단백질의 시퀀스를 가지고 있는 데이터가 필요하다. 우선 DIP의 상호작용 데이터에서 단백질 쌍의 DIP고유번호를 가져와 DIP의 시퀀스파일에서 DIP고유번호에 해당하는 시퀀스를 얻어오고, fasta format으로 된 단백질의 시퀀스를 한 문자씩 읽어 해당 아미노산의 특성에 따라 SVM데이터로 가공을 한다. 아미노산의 종류별로 9가지 대표적인 속성을 정리한 자료에 따라 "속성이 존재" 또는 "속성이 없음"으로 나타낸다. <표 1>은 아미노산을 9가지 속성에 따라 분류해 놓은 표이다. 예를 들어 "ACG"의 시퀀스를 갖는 단백질의 경우 Hydrophobicity (소수성)을 적용할 때 "1:1 2:1 3:0"으로 나타 낼 수 있다.

Amino Acid	Hydro-pobicity	positive	negative	polar	charged	small	tiny	aromatic	aliphatic
Ala	X	-	-	-	-	X	X	-	-
Cys	X	-	-	-	-	X	-	-	-
Asp	-	-	X	X	X	X	-	-	-
Glu	-	-	X	X	X	-	-	-	-
Phe	X	-	-	-	-	-	-	X	-
Gly	X	-	-	-	-	X	X	-	-
His	X	X	-	X	X	-	-	X	-
Lys	X	X	-	X	X	-	-	-	-
Ile	X	-	-	-	-	-	-	-	X
Leu	X	-	-	-	-	-	-	-	X
Met	X	-	-	-	-	-	-	-	-
Asn	-	-	-	X	-	X	-	-	-
Pro	-	-	-	-	-	X	-	-	-
Gln	-	-	-	X	-	-	-	-	-
Arg	-	X	-	X	X	-	-	-	-
Ser	-	-	-	X	-	X	X	-	-
Thr	X	-	-	X	-	X	-	-	-
Val	X	-	-	-	-	X	-	-	X
Trp	X	-	-	X	-	-	-	X	-
Tyr	X	-	-	X	-	-	-	X	-

표 1. 아미노산의 속성

기존의 논문들[4]에서는 '소수성'을 사용한 실험이 높은 예측율을 나타냈다. '소수성'을 사용한 가공은 단백질의 상호작용에 중요한 영향을 미친다고 알려져 있는 기존의 지식을 예측데이터의 가공에 적용한 방법이다. 하지만 이번 실험에서는 위 9가지 속성으로 생성가능한 모든 부분집합(Power Set), 총 511가지(2^9-1)의 속성의 조합을 사용하여 예측시스템에 적용하였다. H.pyroli와 E.coli의 상호작용 데이터는 전체 단백질 시퀀스로부터 시퀀스를 얻어오고, 511가지의 속성조합에 따라 가공한 후 각각 SVM 학습 데이터와, 예측 데이터로 사용한다. SVM의 학습을 위해서는 '상호작용 단백질 데이터'와 '상호작용하지 않는 단백질 데이터'가 필요한데, '상호작용하지 않는 단백질 데이터'의 경우는 DIP에서 별도로 제공하지 않기 때문에, Random방식으로 존재하지 않는 단백질 상호작용데이터를 생성하여 사용하였다.

2.3 데이터의 실험과 검증방법

511가지 실험데이터의 학습과 예측을 자동화하기 위한 프로그램을 개발했다.

예측시스템의 학습방법, 단백질 쌍의 조합방법, 상호작용하지 않는 데이터의 생성방법 등에 따라 여러 예측방법이 있을 수 있지만, 기존의 yeast를 대상으로 '소수성'을 사용한 예측시스템에서 높은 예측율을 나타낸 점을 고려하여, 실험대상과 속성만을 변화시켜 이번 실험에 적용 하였다. <그림 2>는 'yeast'의 '소수성'을 사용한 실험의 결과이다.

```
$ svm_classify lw3y.svm modelly3x
Accuracy:   97.40000% (1948/2000)
Precision:  95.40230% (996/1044)
Recall:     99.60000% (996/1000)
System/Answer p/p p/n n/p n/n: 996 48 4 952
```

그림 2. 'yeast'의 '소수성'을 사용한 실험에서의 결과

3 실험과 평가

정확한 예측율을 구하기 위해 상호작용이 알려진 E.coli의 데이터를 예측에 적용하고 실제 상호작용 데이터와 비교하여 확률을 계산한다. <표 2>는 아미노산의 속성에 해당하는 약자를 정리한 표이다. <표 3>은 속성의 조합을 통해 얻어진 511가지 실험의 결과를 정리한 표로써 속성의 조합에 해당하는 Accuracy, Precision, Recall을 나타 낸 표이다. 예를들어 "NST"의 경우 N(Negative), S(Small), T(Tiny)의 조합을 뜻한다.

약자	속성
H	Hydrophobicity (소수성)
P	Positive
N	Negative
p	polar
C	Charged
S	Small
T	Tiny
A	Aromatic (방향성)
a	aliphatic

표 2. 속성별 약자

속성	Accuracy	Precision	Recall
A(방향성)	90.2196	95.4955	84.4621
NTa	72.6547	80	60.5578
H(소수성)	70.2595	62.8141	99.6016

표 3. 속성의 조합에서 높은 수치를 나타낸 속성들

511개의 속성을 조합한 실험에서 기존의 '소수성'을 사용한 데이터의 가공법이 H.pyroli와 E.coli에서는 비교적 낮은 수치를 나타냄을 알 수 있다. '방향성'을 사용한 데이터의 가공이 '소수성'을 사용한 데이터의 가공보다 훨씬 더 높은 예측율을 나타내는데 '방향성'은 단백질 상호작용의 중요한 역할을 한다고 알려지지 않은 속성이다. '방향성'은 yeast를 대상으로 한 별도의 실험에서 '소수성'을 사용한 예측보다 낮은 예측율을 보였다. <그림 3>은 '방향성'을 yeast에 적용한 실험의 결과이다. <그림 4>는 '소수성'을 yeast에 적용한 실험의 결과이다. 여기서 <그림 2>보다 예측율이 떨어지는 이유는 학습데이터의 수와 상호작용이 없는 데이터등의 변수가 있기 때문이다. <그림 2>는 이 실험 후에 다른 방법으로 실험한 결과이다.

```
[bios@bi HorA]$ svm_classify A.svm A.mod
Accuracy:    55.22239% (1105/2001)
Precision:   56.67090% (446/787)
Recall:      44.55544% (446/1001)
System/Answer p/p p/n n/p n/n: 446 341 555 659
```

그림 3. '방향성'을 yeast에 적용한 실험결과

```
[bios@bi HorA]$ svm_classify H.svm H.mod
Accuracy:  77.86107% (1558/2001)
```

그림 4. 위와 동일한 조건으로 '소수성'을 yeast에 적용한 실험결과

이로써 지금까지의 '소수성'을 사용한 데이터의 가공법이 특정 대상(target)에 국한된 높은 예측율을 나타내는 것임을 알 수 있고, '방향성'에 기반한 데이터의 가공역시 특정 대상에 국한된 예측율을 나타내고 있음을 알 수 있다.

4. 결론

이 논문에서는, 아미노산의 속성 중 '소수성'에 기반한 단백질 상호작용 예측 실험을 확장하여 실제 예측에 쓰이게 될 다른 종 간의 상호작용 예측을 다양한 아미노산의 속성조합을 통해 얻어진 511가지의 실험결과를 비교했다. 다양한 속성을 사용한 실험을 통해 SVM을 사용한 단백질 상호작용 예측에 높은 예측율을 보이는 속성을 발견하기 위한 실험이었다. 이 논문을 통해 '방향성'이 H.pyroli와 E.coli간의 예측에서 높은 예측율을 보이는 것을 확인 할 수 있는데, 여기에는 두 상호작용 단백질 쌍의 결합방법, 상호작용하지 않는 데이터의 정의 등과 같이 학습, 검증 데이터의 가공법에 따라 실험결과가 달라질 수 있다. 이번 실험을 통해 얻어진 '방향성'이라는 아미노산의 속성도 H.pyroli, E.coli 간의 예측에서만 높은 확률을 나타낼 수도 있지만, 더 많은 실험을 통해 발견해봐야 할 문제이다.

H.pyroli, E.coli의 실험과 yeast의 실험에서와 같이 각각의 대상에 맞는 속성을 발견하는 것도 앞으로의 상호작용예측에 많은 도움이 될 수 있을 것이며, SVM을 사용한 더 높은 예측율을 보이는 시스템의 구현과 실제 적용을 위하여 다양한 환경요인을 바꿔가며 실험하기 위한 계획을 하고 있다. 앞으로 yeast에서의 511가지 조합속성에 대한 실험을 진행할 예정이며, 집쥐-사람 간의 실험도 진행할 예정이다.

References

[1] Tiny SVM
http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/

[2] Bock,J.R. and Gough,D.A. (2001) Predicting protein-protein interactions from primary structure. Bioinformatics, 17, 455-460.

[3] DIP (Database of Interacting Proteins)
http://dip.doe-mbi.ucla.edu/

[4] Hopp.T.P and Woods.K.R. (1981) Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl Acad. Sci. USA.78,3824-3828