단백질에서의 RNA 결합 부위 예측

김현우°, 한경숙
인하대학교 전자계산공학과
whytok°@hanmail.net, khan@inha.ac.kr

Prediction of the RNA Binding Sites of Proteins

Hyunwoo Kim°, Kyungsook Han
School of Computer Science and Engineering, Inha University

요 약

PDB로부터 얻은 51개의 단백질-RNA 복합체를 대상으로 기존 연구에서 얻은 단백질과 RNA의 결합 성향성 값과 본 논문에서 새로 구한 단백질의 표면 노출정도에 따른 결합 성향성 값을 이용하여 단백질의 결합 기대치를 구한다. 또한 구한 결합 기대치를 활용하여 새로운 단백질-RNA 복합체를 대상으로 단백질의 결합 부위 예측을 시도하였다. 결합 기대치는 0.240 이상인 경우 결합할 가능성이 높은 것으로 판별하였고, 그 결과 단백질의 결합 후보지를 전체 단백질의 25% 정도로 줄일 수 있었다.

1. 서 론

단백질-단백질 interaction을 밝히는 것은 단백질의 기능을 알아내기 위한 필수적인 과정이기 때문에, 해당 연구는 그 중요성이 인지되어 각지에서 활발히 진행되고 있다. 단백질-단백질 interaction이 bioinformatics의 주요한 이슈이긴 하지만 단백질은 단백질 뿐 아니라 RNA나 DNA와도 interaction 하며 이 또한 중요한 문제이다. 이를 테면 유전자 전사 과정이나 RNA의 효소 작용이 그와 관련된 예이다. 특히 단백질-RNA interaction은 RNA 질병 치료를 위한 약이나 RNA를 이용한 신약 설계에도 응용 될 수 있는 상업적 잠재력 마저 가지고 있어 점차 그 관심은 증가하고 있는 추세이다.

본 논문은 RNA와 interaction 하는 것으로 알려진 단백질을 가지고 해당 단백질의 부위 중 RNA와 interaction 할 가능성이 높은 부위를 찾아내는 기법을 기술한다. 포켓을 이용하여 구조적 접근을 하는 방법으로 interaction 후보지를 구하는 기존 연구 [1]가 있지만, 본 연구에서는 단백질의 각 부위별로 각 아미노산의 종류와 그 접근성을 가지고 interaction 기대치를 계산하는 방법을 이용한다. 본 논문에서는 몇 가지 bioinformatics 분야의 어플리케이션을 이용했으며, 그 결과를 조합하고 분석하기 위해 Python [2]을 이용했다.

2. 데이터 선정 및 구현 방법

2.1 분석데이터 선정

분석 데이터로 PDB [3]에 있는 단백질-RNA complex를 사용했다. 기본적으로 X-ray crystallography 기법에 의해 밝혀진 복합체 중 해상도가 3.0Å 이하인 것을 선정하였으며 PSI-BLAST [4]를 이용하여 상동 단백질을 제외 시켜 같은 복합체의 중복된 분석을 피했다. PSI-BLAST 옵션은 *E* 값을 0.001, *identities* 값은 80% 이상으로 적용하였으며, 그 결과 상동 복합체로 판단되는 것은 제외하고 51개의 단백질-RNA complex를 분석 데이터로 선정했다. 표1은 본 연구의 분석에 사용된 51개 복합체의 PDB ID를 생물학적 기능별로 분류하여 나열한 것이다.

표 1 분석에 사용된 PDB Data

tRNA	15	1B23, 1H4Q, 1C0A, 1H4S, 1EFW 1IL2, 1F7U, 1QF6, 1FFY, 1QTQ 1G59, 1SER, 1GAX, 2FMT, 1K8W
mRNA	1	1B7F
ribosome	6	1DK1, 1DFU, 1FEU, 1HC8, 1I6U, 1MMS
ribozyme	4	1CX0, 1B2M, 1JBR, 1JBS
virus	12	1E7X, 1HDW, 1HE0, 1HE6, 1ZDH 1ZDI, 2BBV, 5MSF, 6MSF, 7MSF 1F8V, 1KNZ
TRAP	3	1C9S, 1GTF, 1GTN
SRP	4	1HQ1, 1JID, 1L9A, 1LNG
others	6	1DI2, 1EC6, 1FXL, 1G2E, 1KQ2 1URN

2.2 Protein-RNA interaction propensity

단백질의 interaction 에는 해당 단백질 내의 잔기의 종류

와 잔기들의 구성 부위에 영향을 받는다. 본 논문에서는 이 두 가지 조건을 이용하여 단백질의 각 아미노산 별로 RNA와의 interaction 기대치를 계산한다.

우리의 이전 연구에서 단백질-RNA 간의 interaction 선호도를 propensity라는 값을 정의하여 계산하였다 [5]. 이 값은 1을 기준으로 1 보다 큰 경우 각각의 아미노산과 핵산 간에 높은 interaction 선호도가 있는 것이고, 1 보다 작은 경우는 낮은 선호도를 갖는 것이다. 그러나 이 값의 경우는 각각의 아미노산과 핵산간에 일반적인 interaction 성향을 구한 것이기 때문에 이 값만을 가지고 interaction 부위를 예측하는 것은 제한적일 수 밖에 없다. 이를 테면 잔기가 표면에 완전히 드러나 있는지 아니면 일부분만 드러나 있는지는 해당 잔기가 외부의 물질과 interaction 하는데 큰 영향을 끼치는 중요한 원인이 되지만 propensity 값만을 본다면 두 경우가 모두 동일한 값을 갖게 된다. 따라서 이번 연구에서는 기존 연구의 propensity 값과 노출된 정도에 따른 성향 값을 이용하여 interaction 기대치를 구하고 그 값을 이용해서 RNA와 interaction 후보지를 선정한다. 표 2는 우리의 기존 연구 [5]에서 계산한 각 아미노산의 propensity value를 보여주고 있다.

표 2 각 아미노산 별 Propensity 값

ARG	LYS	ASN	SER	THR	ASP	TYR	GLU	GLN	HIS
3.18	2.63	2	1.8	1.68	1.43	1.27	1.15	1.11	1
TRP	**PHE**	**GLY**	**MET**	**CYS**	**PRO**	**LEU**	**ALA**	**ILE**	**VAL**
0.65	0.48	0.34	0.31	0.28	0.17	0.15	0.13	0.11	0.02

2.3 노출 정도에 따른 interaction propensity

아미노산의 위치가 반영된 interaction 기대치를 얻기 위해 새로운 값인 노출 정도에 따른 interaction propensity값을 계산해야 한다. 여기에 사용되는 어플리케이션은 NACCESS로, NACCESS 는 PDB 파일을 입력 받아 각 잔기와 원자의 accessibility를 계산하는 어플리케이션이다. Accessibility 값이 클수록 해당 잔기나 원자가 외부에 노출된 정도가 크다는 것을 의미한다 [6]. 본 논문에서는 51개 데이터에 대해 각각의 아미노산의 accessibility를 구하여 interaction 하는 아미노산과 그렇지 않은 아미노산의 개수를 세서 아미노산의 accessibility에 따른 interaction 비율을 계산했다 [표3].

모든 아미노산은 0~249의 accessibility를 갖으며 대체로 값이 클수록 interaction 하는 비율이 높은 경향을 띠지만 정확히 그렇지는 않다. 이를테면 210~219나 240~249의 아미노산은 높은 수준으로 노출 되어있지만, interaction은 제한적으로 일어난다. 가장 높은 비율로 interaction하는 accessibility대는 170~179로 이 수치대의 아미노산은 20%가 interaction을 하고 있다. 반면 접근성이 10 미만인 아미노산의 경우는 거의 RNA와의 interaction이 일어나지 않고 있다.

2.4 Interaction 기대치

각 아미노산의 Interaction 기대치는 아래 수식 (1)을 이용하여 계산한다.

$$E = k \cdot P \cdot EP \qquad (1)$$

기대치 E는 각 아미노산의 propensity 값 P와 accessibility에 따른 interaction propensity EP의 곱으로 얻는다. 여기서 k는 기대치에 영향을 주는 상수로 2.95를 적용한다. 기대치를 계산하는 예를 들면, 175의 accessibility를 갖는 ARG의 interaction 기대치를 구하고자 한다면 2.95 * 3.18 * 0.200 = 1.876 의 기대치를 갖게 되므로 51개 PDB 내의 모든 아미노산의 평균 기대치인 0.126

표 1 아미노산의 노출정도에 따른 propensity 값

접근성	총 잔기 수 (A)	Interaction 잔기 수 (B)	(B) / (A)
0 ~ 9	7769	33	0.004
10 ~ 19	2796	102	0.036
20 ~ 29	2085	99	0.047
30 ~ 39	2080	121	0.058
40 ~ 49	1784	67	0.038
50 ~ 59	1912	59	0.031
60 ~ 69	1767	78	0.044
70 ~ 79	1318	84	0.064
80 ~ 89	1357	56	0.041
90 ~ 99	1131	98	0.087
100 ~ 109	803	39	0.049
110 ~ 119	665	36	0.054
120 ~ 129	678	63	0.093
130 ~ 139	478	46	0.096
140 ~ 149	294	23	0.078
150 ~ 159	211	17	0.081
160 ~ 169	150	17	**0.113**
170 ~ 179	105	21	**0.200**
180 ~ 189	63	8	**0.127**
190 ~ 199	26	3	**0.115**
200 ~ 209	26	3	**0.115**
210 ~ 219	17	1	0.059
220 ~ 229	9	1	**0.111**
230 ~ 239	11	2	**0.182**
240 ~ 249	31	0	0.000
Total	27566	1077	0.039

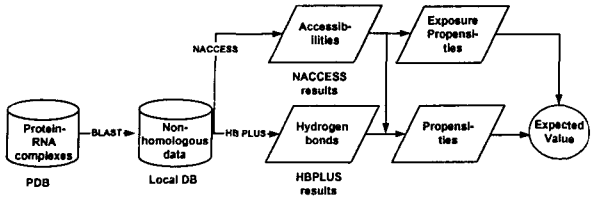

그림 1 기대치를 얻기 위한 개괄적인 과정

과 비교하면 interaction 기대치가 상당히 높은 편이 된다. 그림 1은 기대치를 얻기 위한 전체적인 과정을 개략적으로 보여주고 있다.

3. 평가 및 결론

도출한 식을 이용하여 분석 데이터에 포함 되지 않은 새로운 단백질-RNA 복합체 데이터를 이용해서 1A34, 1BY4, 2A8V를 대상으로 interaction 하는 아미노산을 예측해 보았다. 이 세 복합체는 이미 RNA와 interaction 하고 있는 것을 알고 있기 때문에, 단백질 부위만 따로 떼어서 각 아미노산의 기대 값을 구한 다음 실제 interaction 하는 아미노산과 비교 함으로써 기대치의 신뢰도를 측정할 수 있다.

그림 2는 단백질-RNA 복합체 1BY4를 RASMOL을 이용하여 시각적으로 표현한 것이다. 청색으로 그려진 부분이 기대치가 0.240 이상인 부분이며 spacefill로 표시된 부분은 실제로 결합하는 부분이다.

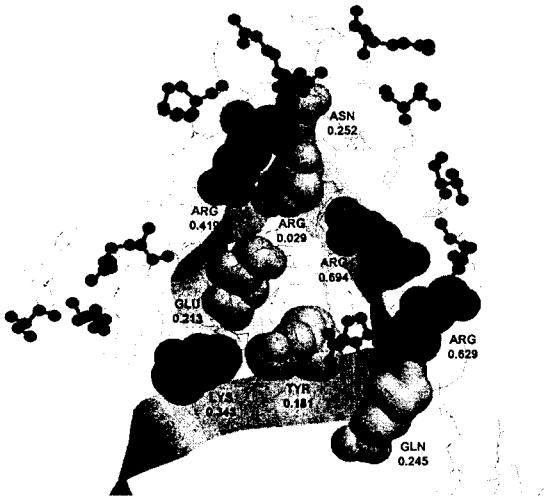

그림 2 단백질과 RNA의 결합 예
황색은 RNA, 청색은 기대치 0.240 이상의 아미노산
Spacefill로 그려진 부분은 실제로 결합하는 아미노산
각 수치는 interaction 기대치를 의미함.

세 복합체는 총 820개의 아미노산으로 구성되어 있으며, 그 중 42개의 아미노산이 RNA와 interaction한다. Interaction 기대치가 0.240 이상인 경우를 interaction 할 가능성이 높은 아미노산으로 분류한다면 820개 중 223개의 아미노산이 interaction 할 가능성이 높은 아미노산으로 분류되고, 그 중 29개 아미노산이 실제로 interaction 하는 아미노산으로 판별되었다. 이 과정에서 전체 단백질의 25% 정도를 결합 후보지로 압축 할 수 있어 관련 연구에 도움을 줄 수 있을 것으로 기대된다. 낮은 기대치를 갖지만 실제로는 interaction하는 아미노산이 13개 있었는데, 그 이유는 실제 interaction은 잔기 단위로 일어나는 것이 아니기 때문에 해당 부근에 interaction이 있다면 인근 잔기들은 낮은 기대치를 갖더라도 interaction 하게 되는 경우가 있기 때문이다. 이러한 문제를 해결하기 위해서는 본 논문에서와 같이 단일 잔기 단위로 기대치를 계산하기 보다는 구역 단위로 계산하여 interaction 후보지를 예측해야 하는데, 이것은 차후 개선시킬 예정이며 보다 효과적인 결과가 기대된다. 또한 본 논문에서는 단백질-RNA interaction의 경우로 한정하여 접근하였지만, 우리의 접근 방식은 단백질-단백질 interaction 이나 단백질-DNA interaction에도 유효한 접근 방식이므로 해당 복합체의 연구에도 쉽게 응용될 수 있다.

후기

본 연구는 정보통신부 정보통신 선도기반기술개발사업 (과제 번호 01-PJ11-PG9-01BT00B-0012)의 지원에 의하여 이루어졌음.

참고문헌

[1] David G. Levitt and Leonard J. Banaszak. POCKET: A computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. J. Mol. Graphics, **10**, 229-234, 1992
[2] http://www.python.org
[3] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. The Protein Data Bank. Nucleic Acids Research, **28**, 235-242, 2000.
[4] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, **25**, 3389-3402, 1997
[5] H. Kim, E. Jeong, S.-W. Lee and K. Han. Computational analysis of hydrogen bonds in protein-RNA complexes for interaction patterns. FEBS Letters, 2003 (in press)
[6] http://wolf.bms.umist.ac.uk/naccess