Random forest 를 이용한 RNA 에서의 단백질 결합 영역 예측

최대식, 박병규, 채한주, 이욱, 한경숙
인하대학교 컴퓨터공학과
e-mail : khan@inha.ac.kr

Prediction of protein binding regions in RNA using random forest

Daesik Choi, Byungkyu Park, Hanju Chae, Wook Lee, Kyungsook Han,
Dept. of Computer Engineering, Inha University

요 약

단백질과 RNA 의 상호작용 데이터가 대량으로 늘어남에 따라, 단백질과 RNA 의 결합부위를 예측하는 계산학적인 방법들이 많이 개발되고 있다. 하지만, 많은 계산학적인 방법들은 단백질에서 단백질과 RNA 결합부위를 예측한다는 한계점이 있었다. 본 논문에서는 RNA 와 단백질의 서열정보를 모두 사용하여, 단백질과 결합하는 RNA 결합부위를 예측하는 기법과 그 결과를 논한다. WEKA random forest(http://www.cs.waikato.ac.nz/ml/weka/)를 이용하여 예측 모델을 개발하였고, RNA 서열의 서열 프로파일, 서열 composition, 결합 상대방의 단백질의 특성 등을 특징으로 표현하였다. Random forest 기법을 사용한 cross validation 의 결과로서 1:1 모델에서 제일 높은 성능인 92.4% sensitivity, 92.0% specificity, 92.2% accuracy 를 보였고, independent test 에서는 72.5% sensitivity, 90.0% specificity, 92.1% accuracy 를 보였다.

1. 서론

차세대 서열 결정 기법과 결합 CLIP(cross-linking and immunoprecipitation)과 같은 대량신속처리 실험 기법에서 최근 기술의 진보에 따라, RNA 와 결합하는 단백질 및 이들의 표적 RNA 에 대한 발견이 가속화되고 있다. 이에 따라 컴퓨터 모델링 기법들이 많이 나오고 있지만, 현재 사용되고 있는 대부분의 컴퓨터 모델링 기법은 RNA 에서 단백질 결합 영역을 찾기보다는 단백질에서 RNA-결합 영역을 찾는 것으로 주로 제한되고 있다. 예를 들어, BindN 기법[1]의 업그레이드 버전인 BindN+ 기법[2]은 단백질 서열의 생물학적 특징 및 진화 정보로부터 RNA- 또는 DNA-결합 잔기를 예측하기 위하여 support vector machine (SVM)을 이용한다. 또한 RNABindRPlus 기법[3]은 최적화 SVM 으로부터 예측 및 서열 유사도 기법으로부터 예측을 조합하여 단백질 서열에서 RNA-결합 잔기를 예측한다.

RNA 에서 단백질과의 결합부위를 예측하는 것은, 단백질에서 RNA 와 결합하는 아미노산을 예측하는 것보다 훨씬 어려운데, 그 이유는 다음과 같다. 20 종류의 아미노산으로 구성되는 단백질의 경우, w 개 아미노산으로 구성되는 단백질 서열의 패턴이 20^w 개 존재하지만, 4 종류의 염기로 구성되는 RNA 의 경우 w 개의 염기를 갖는 RNA 서열 패턴이 4^w 개 존재한다. RNA 서열 패턴의 다양성이 단백질 서열 패턴에 비해 현저히 낮기 때문에, 서열 정보만을 사용하여 RNA 에서 단백질과의 결합부위를 예측하는 것은 단백질에서 핵산과 결합하는 영역을 예측하는 기술에 비하여 훨씬 어렵다. 이러한 이유로, 최근까지도 서열정보만을 이용하여 RNA 에서 단백질과의 결합부위를 성공적으로 개발된 기술이 별로 없고, 단백질에서 RNA 와 결합하는 부위를 예측하는 연구들이 주로 진행되었다. catRAPID 기법[4]은 2 차 구조, 수소 결합 및 반데르발스 힘의 기여도를 조합하여 RNA 와 단백질 분자에서의 결합 성향을 측정한다. catRAPID 기법은 50 개 이상의 염기 또는 핵산을 가지는 RNA 서열에 대해서만 단백질 결합 영역을 예측할 수 있다. DeepBind 기법[5]은 RNA 에서 단백질과 RNA 의 결합을 예측하는 기법이다. 이 기법은 대량신속처리 실험으로부터 막대한 양의 데이터를 심층 신경망을 이용하여 학습 모델을 만든다.

RNA 서열에서 단백질-결합 영역을 예측하는 문제와 관련해서, DeepBind 기법은 RNAcompete, CLIP-서열 등으로 얻어진 데이터로 학습된다. 이 기법은 결합 단백질별로 별도의 예측 모델이 포함된다. 별도의 모델로 구성되어 있어서 사용자가 결합 단백질에 대한 사전 정보가 없다면, 각각의 모델을 모두 시험해야 한다는 단점이 있다. 추가로 DeepBind 기법은 출력 결과로서 입력된 RNA 서열에서의 단백질과의 결합 영역을 제공하지 않고 결합 스코어만을 제공한다.

이에 따라, 본 논문에서는 실제 생화학적 실험을 수행하기 전 단백질과 상호작용하는 RNA 영역을 신속

하고 정확하게 예측할 수 있는 컴퓨터 모델링 시스템을 개발하였다.

2. 데이터 및 특징

2.1 RNA와 단백질 데이터

단백질과 RNA 의 결합부위 데이터는 CLIPdb[6]에서 서열길이가 25, 결합 친화도가 0.9 이상이며 PARalyzer 실험으로 나온 데이터를 추출하였다. 그리하여 포지티브 데이터로서 단백질과 결합하는 RNA 의 결합부위 데이터로 RNA 서열 5,145 개를 뽑았으며, RNA 와 결합하는 단백질의 수는 총 14 개이다. RNA 와 결합하는 단백질 서열은 NCBI GEO(http://www.ncbi.nlm.nih.gov/geo/)의 사이트에서 얻었다. 네거티브 데이터로서, 참조 인간 유전체 GRCh37/hg19 에서 25 개 염기에 대하여 51,450 개 (포지티브 데이터의 10 배)의 RNA 와 비-결합하는 영역을 선택하였다. 참조로 사용된 인간 유전체는 단백질과 결합하는 영역보다 많은 비-결합 영역을 포함하고 있으므로, 비-결합 영역에 대한 결합 영역의 비율을 다르게 하여 여러 개의 데이터 세트를 구성하였다. 구성된 포지티브와 네거티브 서열데이터의 비율이 1:1, 1:2, 1:4, 1:6, 1:8 그리고 1:10 으로 이루어져 있다.

또한 데이터세트의 중복데이터 제거를 위해 CD-HIT-EST[7]을 이용하여 서열 유사도가 80%이상인 데이터는 전부 삭제하였고, 학습을 위한 훈련 데이터와 실제 테스트를 위한 테스트 데이터로 나누었는데, 단백질과 결합하는 RNA 의 결합부위 데이터를 중복서열 제거 후 남은 4,372 개의 서열의 약 70%를 학습 데이터로 사용하였고, 나머지 30%인 약 1000 개의 서열을 테스트 데이터로 사용하였기 때문에 학습 데이터와 테스트 데이터의 사이에는 서열 중복이 존재하지 않는다. 다음 <표 1>에서는 정확한 학습 데이터의 서열 개수를 확인 가능하다.

<표 1> 중복 제거 후 남은 학습 데이터의 서열 개수.
P: positive data, N: negative data

	1:1	1:2	1:4	1:6	1:8	1:10
P	3,372	3,372	3,372	3,372	3,372	3,372
N	3,679	7,200	13,610	19,065	22,826	26,212
Total	7,051	10,572	16,982	22,473	26,198	29,584

<표 2>에서는 중복 제거 후 남은 테스트 데이터의 개수를 확인 가능하다.

<표 2> 중복 제거 후 남은 테스트 데이터의 서열 개수.
P: positive data, N: negative data

	1:1	1:2	1:4	1:6	1:8	1:10
P	1,000	1,000	1,000	1,000	1,000	1,000
N	1,000	2,000	3,998	5,998	7,998	9,998
Total	2,000	3,000	4,998	6,998	8,998	10,998

2.2 특징의 선택 및 표현

본 연구에서 RNA 서열에서 단백질과의 결합부위를 예측하기 위하여 사용한 특징들은 다음과 같다. RNA 서열 프로파일 정보로서 추출된 RNA 서열의 염기에 대한 위치가중행렬(PWM)[8]을 연산하고, 추출된 RNA 서열 중에서 훈련 데이터로 활용되는 포지티브 데이터와 네거티브 데이터가 활용되는데, 상기 위치가중행렬은 추출된 RNA 서열을 구성하는 각각의 단일 염기에 대한 로그-오즈 스코어에 따른 단일 염기 위치가중행렬(mPWM) 및 추출된 RNA 서열에서 중첩으로 배열되는 2-염기 위치가중행렬(dPWM)을 포함한다.

단일 염기 위치가중행렬은 추출된 RNA 서열을 구성하는 단일 염기 위치가중행렬은 각각의 단일 염기에 대한 로그-오즈 스코어를 연산하여 얻어지는데, (1)을 통하여 연산이 가능하다.

$$\mathrm{mPWM}(i,j) = \left(\frac{\mathrm{frequency}^{+}(i,j)}{\mathrm{frequency}^{-}(i,j)}\right) \quad (1)$$

(1)에서 단일 염기 위치행렬 mPWM(i,j)의 행을 구성하는 i 는 RNA 서열을 구성하는 단일 염기인 아데닌, 사이토신, 구아닌 및 우라실을 각기 나타낸다. j 는 j-번째 위치를 나타내고, $\mathrm{frequency}^{+}$는 특정 단일 염기가 단백질과 결합하는 빈도수이고, $\mathrm{frequency}^{-}$ 는 특정 단일 염기가 단백질과 결합하지 않는 빈도수를 나타낸다. (2)에서는 2-염기 위치가중행렬에 대한 식을 표현한다. 서열의 염기가 n 개라면 단일 염기 위치가중행렬의 특징은 n 개이고, 2-염기 위치가중행렬의 특징은 n-1 개이다.

$$\mathrm{dPWM}(i,j) = \left(\frac{\mathrm{frequency}^{+}(i,j)}{\mathrm{frequency}^{-}(i,j)}\right) \quad (2)$$

서열 composition 의 정보는 서열을 구성하는 각각의 단일 염기, 2-염기 및 3-염기의 빈도수를 포함하여 계산된다. 단일 염기 4 개에 대한 composition, 2-염기 16 개에 대한 composition, 3-염기 64 개에 대한 composition 의 개수로 구성되어 있다.

결합 상대방 단백질의 서열 정보를 이용한 특징으로는 단백질 서열에서의 아미노산 그룹의 composition, 아미노산 그룹 간의 transition, 아미노산의 distribution 3 가지를 사용하였다. 단백질 서열의 특징을 표현하기 위하여 20 가지의 아미노산을 다음 7 개의 그룹으로 분류를 하였다.

그룹 1 = {A, G, V}, 그룹 2 = {C}, 그룹 3 = {M, S, T, Y}, 그룹 4 = {F, I, L, P}, 그룹 5 = {H, N, Q, W}, 그룹 6 = {K, R}, 그룹 7 = {D, E}. 아미노산 그룹 간의 transition 은 단백질 서열에 인접한 아미노산이 아미노산 그룹 1 에서 그룹 2 로 변하거나 또는 그룹 2 에서 그룹 1 로 transition 되는 경우, 그룹 1 에서 그룹 3 으로 또는 그룹 3 에서 그룹 1 로 변하는 경우, 다른 경우까지 포함하여 최종적으로 서로 다른 아미노산 그룹 간의 transition 의 수에 대한 normalized 된 빈도수를 표현한다. 아미노산 distribution 는 단백질 서열에서 각 그룹에 속한 아미노산이 서열에서 첫 번째, 25%, 50%, 75%, 100%에 해당하는 위치를 표현한다. 단백질 서열 정보의 특징의 수는 composition 7 개, transition 21 개, distribution 35 개로 이루어져 있고, RNA 서열 및 단백질 서열의 최종 특징의 개수는 2n + 146 개이다.

본 연구에서는 서열 길이 25 인 RNA 서열을 기준으로 특징을 표현하였기 때문에 RNA 의 특징 133 개와 결합 단백질의 특징 63 개를 더하여 총 196 개의 특징을 사용하였다.

3. 예측 모델의 성능 평가

3.1 성능 지표

본 연구에서는 정확한 성능을 측정하기 위하여 총 6 가지의 성능 지표를 사용하였다. 그 성능은 sensitivity (SN), specificity (SP), accuracy (ACC), positive predictive value (PPV), negative predictive value (NPV), Matthews correlation coefficient (MCC)로 식 (3)~(8)로 정의된다.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (4)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (5)$$

$$PPV = \frac{TP}{TP + FP} \quad (6)$$

$$NPV = \frac{TN}{TN + FN} \quad (7)$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (8)$$

위 수식에서 true positive (TP)는 결합한다고 예측된 서열이 실제로 결합하는 경우의 수, true negative (TN)는 결합하지 않는다고 예측된 서열이 실제로도 결합하지 않는 경우의 수, false positive (FP)는 결합한다고 예측된 서열이 실제로는 결합하지 않는 경우의 수, false negative (FN)는 결합하지 않는 다고 예측된 서열이 실제로는 결합하는 경우의 수를 의미한다.

3.2 10-fold cross validation

10-fold cross validation 을 통한 학습 모델의 성능을 확인하기 위하여 학습 데이터의 특징들을 인코딩하여 학습 모델을 만들었으며, 학습 모델을 이용하여 10-fold cross validation 을 수행하였다. <표 3>는 WEKA 의 random forest 기법을 이용하여 10-fold cross validation 을 나타낸 결과이며, random forest 의 최적의 성능을 보여주기 위한 파라미터로 60 개의 결정 트리(decision tree)와 25 개의 random feature selection 을 지정하여 10-fold cross validation 을 실행하였다. positive 및 negative 클래스의 개수의 비율이 1:1, 1:2, 1:4, 1:6, 1:8, 1:10 의 성능에서 제일 좋은 성능을 보인 것은 positive 및 negative 데이터의 비율이 1:1 인 모델이 제일 좋은 성능을 보였으며, 그 결과값으로는 92.41%의 SN, 92.04%의 SP, 92.21%의 ACC, 91.41%의 PPV, 92.97%의 NPV, 0.844 의 MCC 였다. 데이터의 비율이 비슷할수록 좋은 성능이 나온 것을 확인할 수 있었다.

<표 3> WEKA random forest 를 이용한 10-fold cross validation 결과
P: positive data, N: negative data

P:N	SN(%)	SP(%)	ACC(%)	PPV(%)	NPV(%)	MCC
1:1	92.41	92.04	92.21	91.41	92.97	0.844
1:2	89.65	95.28	93.48	89.89	95.16	0.850
1:4	84.16	97.13	94.55	87.89	96.12	0.826
1:6	81.58	97.83	95.39	86.95	96.78	0.815
1:8	78.97	98.01	95.56	85.43	96.93	0.796
1:10	77.46	98.34	95.96	85.72	97.14	0.793

3.3 Independent test

학습 모델의 공정한 평가를 위하여 학습 데이터와 중복이 되지 않는 테스트 데이터를 가지고 Weka 의 random forest 기법을 이용하여 independent test 를 진행하였다.

<표 4>는 independent test 의 결과를 보여주는데 이 결과 역시 positive 및 negative 의 데이터 비율이 1:1 인 성능이 제일 좋은 결과를 보여주는데, 그 성능은 72.50%의 SN, 90.00%의 SP, 81.25%의 ACC, 87.88%의 PPV, 76.60%의 NPV, 0.635 의 MCC 이다. 또한, 테스트 데이터의 단백질과 결합하는 RNA 서열 데이터가 각 데이터셋마다 다르기 때문에 서로 다른 결과 양상을 보인다.

<표 4> WEKA random forest 를 이용한 Independent test 결과
P: positive data, N: negative data

P:N	SN(%)	SP(%)	ACC(%)	PPV(%)	NPV(%)	MCC
1:1	72.50	90.00	81.25	87.88	76.60	0.635
1:2	56.10	85.90	75.97	66.55	79.65	0.440
1:4	52.70	93.77	85.55	67.91	88.80	0.513
1:6	47.10	95.48	88.57	63.48	91.54	0.484
1:8	50.90	96.07	91.05	61.85	93.99	0.512
1:10	49.60	96.32	92.07	57.41	95.03	0.491

3.4 기존 연구와의 예측성능 비교

본 연구에서 개발한 random forest 모델을 기존의 RNA 에서 단백과 RNA 의 결합 예측 방법인 DeepBind[5]와 비교하기 위한 실험을 하였다. 공평한 성능 비교를 위하여 기존에 사용하였던 데이터와 중복되지 않으며, DeepBind 와 같은 단백질을 가지는 RNA 서열을 각 100 개씩 7 개의 단백질에 대한 서열 길이 25 인 서열을 추출하였고 각 단백질에 대한 실험 데이터는 기존 학습 데이터와의 중복데이터를 제거한 후 실험을 진행하였다.

<표 5>는 7 개의 단백질과 결합하는 RNA 서열을 각

100 개씩 뽑아서 실험한 결과를 DeepBind 의 모델에 넣고 실험한 결과 표이다. DeepBind 는 실험 결과로서 결합력에 대한 스코어를 보여주기 때문에 표준편차를 이용하여 Z-score 를 구하였고 Z-score 가 0 보다 크다면 결합하는 경우의 수로 보았으며 Z-score 가 0 보다 작을 경우 결합하지 않는 경우의 수로 보고 성능을 계산한 결과를 보여준다.

Our model						
protein	SN(%)	SP(%)	ACC(%)	PPV(%)	NPV(%)	MCC
FUS	92.19	96.00	94.51	93.65	95.05	0.884
FXR1	100.00	90.00	94.01	87.01	100.00	0.885
FXR2	86.25	88.00	87.22	85.19	88.89	0.742
IGF2BP2	84.81	89.00	87.15	85.90	88.12	0.739
LIN28A	86.59	90.00	88.46	87.65	89.11	0.767
QKI	83.12	92.00	88.14	88.89	87.62	0.758
TARDBP	17.02	93.00	56.19	69.57	54.39	0.155
weighted average	76.06	91.14	84.55	86.95	83.07	0.686
DeepBind						
FUS	32.00	33.00	32.50	32.32	32.67	-0.350
FXR1	32.99	42.00	37.56	35.56	39.25	-0.251
FXR2	43.01	73.00	58.55	59.70	57.94	0.168
IGF2BP2	48.94	59.00	54.12	52.87	55.14	0.080
LIN28A	36.46	53.00	44.90	42.68	46.49	-0.107
QKI	82.00	81.00	81.50	81.19	81.82	0.630
TARDBP	50.00	86.00	68.00	78.12	63.24	0.386
weighted average	46.62	61.00	53.97	53.73	54.05	0.077

위 결과 weighted average 로 각각의 TP, TN, FP, FN 등 4 가지 개별 지표를 각기 더하여서 평균 성능을 구하였는데, 본 연구에서 만든 모델의 성능이 DeepBind 의 성능보다 좋은 것을 볼 수 있었다. 세부 성능으로서 DeepBind 의 weighted average 가 46.62%의 SN, 61.00%의 SP, 53.97%의 ACC, 53.73%의 PPV, 54.05%의 NPV, 0.077 의 MCC 값을 가지는 반면, 우리 모델의 성능은 76.06%의 SN, 91.14%의 SP, 84.55%의 ACC, 86.95%의 PPV, 83.07%의 NPV, 0.686 의 MCC 가 나와 더 좋은 성능을 나타냄을 확인할 수 있었다.

4. 결론

본 연구에서는 RNA 서열 및 결합 단백질 정보를 모두 고려하여, RNA 서열에서 단백질과 결합 가능한 영역을 찾는 새로운 기법을 소개하였다. 예측 모델을 위하여 WEKA random forest 모델을 구현하였으며, 사용한 특징으로는 단일 염기 위치가중행렬, 2-염기 위치가중행렬, composition, 결합 상대방인 단백질의 7 그룹 composition, transition, distribution 을 사용하였다.

데이터로는 positive 및 negative 의 비율이 1:1, 1:2, 1:4, 1:6, 1:8, 1:10 인 6 개의 데이터셋을 만들어서 성능평가로서 10-fold cross validation 및 independent test 를 진행하였다. 결과적으로 데이터셋의 positive 및 negative 의 비율이 1:1 인 데이터셋의 평가 결과가 제일 높은 성능을 보였으며, 10-fold cross validation 의 성능이 92.41%의 SN, 92.04%의 SP, 92.21%의 ACC, 91.41%의 PPV, 92.97%의 NPV, 0.844 의 MCC 이고, independent test 의 성능은 다음과 같이 72.50%의 SN, 90.00%의 SP, 81.25%의 ACC, 87.88%의 PPV, 76.60%의 NPV, 0.635 의 MCC 의 성능을 보였다.

본 연구 결과는 최근 쏟아져 나오는 대량의 데이터를 사용하여 단백질과 RNA 의 상호작용 연구의 결합 영역을 탐색할 경우, 직접적으로 도움이 되는 정보를 제공해주는 역할을 할 수 있기 때문에, 향후 관련 분야의 연구에 상당한 도움을 줄 수 있을 것으로 보인다.

5. 감사의 글

이 논문은 정부 (미래창조과학부)의 재원 (2015R1A1A3A04001243)과 정부 (교육부)의 재원 (2010-0020163)으로 한국연구재단의 지원을 받아 수행된 기초연구사업임.

참고문헌

[1] Wang, L., Brown, S.J., “BindN: a web-based tool for ecient prediction of DNA and RNA binding sites in amino acid sequence.” *Nucleic Acids Res.* 34:243-247 (2006)

[2] Wang, L., Huang, C., Yang, M. Q., Yang, J.Y. “BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features.” BMC Systems Biology 4 (Suppl 1):S3 (2010)

[3] Walia, R.R., Xue, L.C., Wilkins, K., El-Manzalawy, Y., Dobbs, D., Honavar, V. “RNABindRPlus: A predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins.” PLOS One 9(5):e97725 (2014)

[4] Bellucci, M., Agostini, F., Masin, M., Tartaglia, G.G. “Predicting protein associations with long noncoding RNAs.” Nature methods 8(6):444-446 (2011)

[5] Alipanahi, B., Delong, A., Weirauch, M.T., Frey, B.J.” Predicting the sequence specicities of DNA- and RNA-binding proteins by deep learning.” Nature Biotechnology 33:831-838 (2015)

[6] Yang, Y.-C.T., Di,C., Hu, B., Zhou, M., Liu, Y., Song, N., Li, Y., Umetsu, J., Lu, Z,J. “CLIPdb: A CLIP-seq database for protein-RNA interactions.” BMC Genomics 16:51 (2015)

[7] Huang, Y., Niu, B., Gao, Y., Fu, L., Li, W. “Cd-hit suite: A web server for clustering and comparing biological sequences.” Bioinformatics 26(5):680-682 (2010)

[8] Ahmad, S., Sarai, A. “PSSM-based prediction of DNA binding sites in proteins.” BMC Bioinformatics 6(33):6 (2005)