



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2016-0057003
(43) 공개일자 2016년05월23일

(51) 국제특허분류(Int. Cl.)
G06F 19/18 (2011.01) C12Q 1/68 (2006.01)
G06F 19/22 (2011.01) G06F 19/24 (2011.01)
(21) 출원번호 10-2014-0157084
(22) 출원일자 2014년11월12일
심사청구일자 2014년11월12일

(71) 출원인
인하대학교 산학협력단
인천광역시 남구 인하로 100, 인하대학교 (용현동)
(72) 발명자
한경숙
서울특별시 양천구 목동동로 258 금호1차 아파트 101동 1802호
투스전자르갈 나랑후
인천 남구 인하로 100, 4동 제1생활관 520호 (용현동, 인하대학교)
(뒷면에 계속)
(74) 대리인
이원희

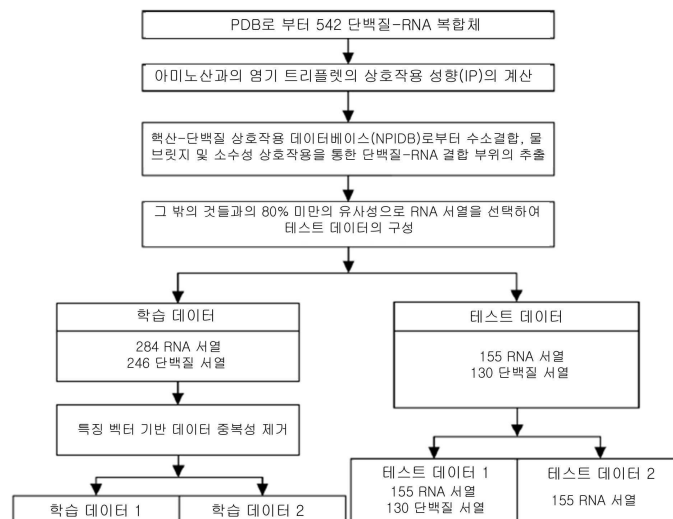
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 결합 상대방을 고려하여 RNA 서열에서 단백질과 결합하는 부위를 예측하는 방법

(57) 요약

본 발명은 RNA 서열에서 단백질과 결합하는 염기 예측 방법에 관한 것으로, RNA 결합 상대방을 고려하여, RNA 서열에서 단백질 결합 부위를 예측하는 새로운 방법을 개발하였다. RNA 및 단백질 서열 모두를 사용하여 단백질이 결합하는 RNA 염기를 예측하는 SVM(support vector machine) 모델을 개발하였고, 비교를 위해 RNA 서열만을 사용하여 단백질이 결합하는 염기를 예측하는 SVM 모델을 개발하여, RNA 서열만을 사용하여 예측하는 것보다 RNA 및 단백질 서열 모두를 사용하여 예측하는 것이 대부분의 성과 측정에서 우수함을 확인하였고, 이러한 새로운 모델은 결합 상대방이 변경될 때 주어진 RNA 서열에 대해 서로 다른 결합 부위를 예측할 수 있음을 확인하였다. 이는 구조를 알 수 없더라도 서열 데이터만으로 단백질과 결합하는 RNA의 염기를 예측할 수 있고, 이는 생화학적 실험에 유용하게 사용될 수 있다.

대표도 - 도1



(72) 발명자

박병규

인천광역시 남동구 인주대로 676번길 22 동아아파트 1동 408호

임진용

인천 남구 주승로 148, 103동 904호 (주안동, 신비마을아파트)

이욱

충남 서산시 남부순환로 876, 3동 605호 (죽성동, 동아1차아파트)

이 발명을 지원한 국가연구개발사업

과제고유번호 2012R1A1A3011982

부처명 미래창조과학부

연구관리전문기관 한국연구재단

연구사업명 여성과학자

연구과제명 단백질과 핵산의 상호작용 예측 모델 개발

기 여 율 1/1

주관기관 인하대학교

연구기간 2012.05.01 ~ 2015.04.30

명세서

청구범위

청구항 1

- 1) RNA 서열에서 단백질과 결합할 가능성이 높은 RNA 염기를 예측하기 위해, RNA 및 단백질의 다양한 특징을 특징 벡터(feature vector)로 표현하는 단계;
- 2) 상기 단계 1)의 표현된 특징 벡터로 표현한 데이터에서 특징 벡터를 기반으로 중복 데이터를 제거하여 학습 데이터를 생성하는 단계; 및
- 3) 상기 단계 2)의 학습 데이터를 이용하여 SVM(support vector machine) 모델을 학습시키고, 학습된 SVM 모델을 이용하여 단백질과 결합하는 RNA 염기를 예측하는 단계를 포함하는, RNA 서열에서 단백질과 결합하는 염기 예측방법.

청구항 2

제 1항에 있어서, 상기 단계 1)의 특징 벡터는 단백질 및 RNA 서열 데이터만 주어졌을 때, 결합부위 예측에 효과적인 특징을 표현하는 것을 특징으로 하는 RNA 서열에서 단백질과 결합하는 염기 예측방법.

청구항 3

제 1항에 있어서, 상기 단계 1)의 특징 벡터(feature vector)의 구성요소로서 표현되는 단백질 및 RNA 서열의 특징은,

전체 RNA 서열에 대한 정보를 표현하는 RNA 전체 특징(global features);

RNA 개별 염기 또는 RNA 염기 트리플렛(nucleotide triplets)에 대한 정보를 표현하는 RNA 국소적 특징(local features); 및

RNA에 결합하는 단백질의 특징을 표현하는 결합 상대방 특징(partner features)을 포함하는 것을 특징으로 하는 RNA 서열에서 단백질과 결합하는 염기 예측방법.

청구항 4

제 3항에 있어서, RNA 전체 특징은 RNA 서열 길이 및 염기 구성이 특징 벡터 표현에 구성요소로 사용되는 것을 특징으로 하는 RNA 서열에서 단백질과 결합하는 염기 예측방법.

청구항 5

제 3항에 있어서, 상기 RNA 국소적 특징은 염기의 분자량(molecular mass), 염기의 pKa 값, 염기 트리플렛과 20 개의 아미노산의 상호작용 성향(interaction propensity; IP)값 및 염기 트리플렛 구성이 특징 벡터 표현에 구성요소로 사용되는 것을 특징으로 하는 RNA 서열에서 단백질과 결합하는 염기 예측방법.

청구항 6

제 3항에 있어서, 상기 결합 상대방 특징은 아미노산의 정규화된 위치(normalized) 및 단백질의 다이펩타이드 구성(dipeptide composition)이 특징 벡터 표현에 구성요소로 사용되는 것을 특징으로 하는 RNA 서열에서 단백질과 결합하는 염기 예측방법.

청구항 7

제 1항에 있어서, 상기 단계 2)는

- 1) 단백질-RNA 상호 작용 쌍들에 존재하는 모든 RNA 및 단백질 서열에서 슬라이딩 윈도우(sliding window) 기법을 이용한 서열 조각을 생성하는 공정;
- 2) 생성된 서열 조각을 상기 구성요소들을 이용하여 특정 벡터에 표현하는 공정; 및
- 3) 표현된 특정 벡터에서 특정 벡터 기반의 중복 제거 기법을 통한 학습 데이터를 구축하는 것을 특징으로 하는 RNA 서열에서 단백질과 결합하는 염기 예측방법.

청구항 8

제 1항에 있어서, 상기 방법은,

- 1) 결합 상대방 단백질 서열을 지정하지 않은 경우, 상대방을 고려하지 않는 모델로 단백질 결합 RNA 염기 예측; 및
- 2) 결합 상대방 단백질 서열이 주어지는 경우, 상대방을 고려한 모델로 단백질 결합 RNA 염기를 예측하는 것을 특징으로 하는 RNA 서열에서 단백질과 결합하는 염기 예측방법.

청구항 9

- 1) 단백질과 상호작용하는 단백질 결합 RNA 염기를 결정하여 단백질-RNA 결합 부위를 결정하는 단계;
- 2) RNA 서열에 있는 상기 단백질 결합 RNA 염기를 예측하기 위해, RNA 및 단백질의 다양한 특징을 특정 벡터로 표현하고, 테스트 데이터를 생성하는 단계; 및
- 3) 상기 단계 2)의 테스트 데이터를 복수의 척도를 이용하여, 학습된 예측 모델의 예측성도를 평가하는 단계를 포함하는 상호작용 단백질을 고려한 RNA 서열에서 단백질과 결합하는 RNA 염기 예측 방법.

청구항 10

제 9항에 있어서, 상기 단계 3)의 복수의 척도는 민감도(Sensitivity, Sn), 특이도(Specificity, Sp), 정확도(Accuracy, Acc), 양성예측도(Positive predictive value, PPV), 음성예측도(Negative predictive value, NPV) 및 매튜 상관관계 계수(Matthews correlation coefficient, MCC)인 것을 특징으로 하는 상호작용 단백질을 고려한 RNA 서열에서 단백질과 결합하는 RNA 염기 예측 방법.

발명의 설명

기술 분야

[0001] 본 발명은 RNA 서열에서 단백질과 결합하는 염기를 예측하는 방법에 관한 것으로, 구체적으로 RNA 결합 상대방인 단백질을 고려하여, RNA와 단백질의 서열 데이터로부터 단백질과 결합하는 RNA 염기를 예측하는 방법에 관한 것이다.

배경 기술

[0002] 단백질과 RNA 사이의 상호작용은 전사 후 유전자 조절(post-transcriptional gene regulation), 선택적 스플라이싱(alternative splicing), 번역(translation) 및 RNA 바이러스의 감염 등의 다양한 생물학적 활성에서 중요

한 역할을 한다(Terribilini M. et al., Prediction of RNA binding sites in proteins from amino acid sequence. *RNA* 2006, 12:1450-1462). 분자의 상호작용을 규명하기 위한 실험 기술의 발달에도 불구하고, RNA-단백질 상호작용의 메커니즘(mechanism)은 완전히 알려져 있지 않다. 따라서, RNA-단백질 상호작용과 상호작용 부위의 연구는 상호작용 메커니즘을 이해하고, 합리적 약물 설계 분야에서 최상의 활성제 및 억제제를 선택하는데 도움을 준다. 최근에 여러 방법이 단백질에서 RNA 결합 부위를 예측하기 위해 적용되었다. 예를 들어, RNABindR 방법은 나이브 베이즈 분류기(Naive Bayes classifier)를 이용하여 단백질 서열에서 RNA 결합 아미노산을 예측한다(Terribilini M. et al., RNABindR: a server for analysing and predicting RNA-binding sites in proteins. *Nucleic Acids Research* 2007, 35:W578-W584). catRAPID는 이차구조(secondary structure), 수소결합(hydrogen bonding) 및 반 데르 발스 결합(van der Waals interaction)으로 단백질 및 RNA 분자의 결합 성향을 추정한다(Bellucci M. et al., Predicting protein associations with long noncoding RNAs. *Nature Methods* 2011, 8:444-445). 또한, 서포트 벡터 머신(support vector machine, SVM)을 이용하여 단백질 서열에서 RNA 결합 부위를 예측한다(Chen YC. et al., Predicting RNA-binding sites from the protein structure based on electrostatics and geometry. *Nucleic Acids Research* 2008, 36:e29, Kumar M. et al., Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins* 2008, 71:189-194). 상기 방법들은 RNA 서열에서 단백질 결합 부위가 아닌 단백질 서열에서 RNA 결합 부위를 예측하고 있다. 또한, RNA 결합 아미노산을 예측할 때, 단백질의 상호작용 RNA는 고려하고 있지 않다. 따라서, 단백질이 다른 RNA에 결합하는 경우에도 단백질 서열에서 같은 결합 부위를 예측하기도 한다.

[0003] 단백질에 결합하는 염기를 예측하는 것은 RNA 결합 아미노산을 예측하는 것보다 어렵다. 그 이유 중 하나는 동일한 길이의 서열에서 단백질 서열 패턴보다 RNA 서열 패턴이 더 적기 때문이다. 또 다른 이유는 단백질-RNA 상호작용의 아미노산보다 염기의 상호작용 성향이 훨씬 낮기 때문이다. 따라서, 결합 염기와 비결합 염기를 구분하기가 매우 어렵다.

[0004] 본 발명자들은 이전 연구에서 서열 유사성 대신 서열 특징에 기초하여 학습 데이터에서 중복된 데이터를 제거하는 방법(redundancy removal method)을 개발하였다(Choi S. and Han K., Prediction of RNA-binding amino acids from protein and RNA sequences. *BMC Bioinformatics* 2011, 12(Suppl 13):S7, Choi S. and Han K., Predicting protein-binding RNA nucleotides using the feature-based removal of data redundancy and the interaction propensity of nucleotide triplets. *Computers in Biology and Medicine* 2013, 43:1687-1697). 특징 기반의 중복 제거 방법은 통상적으로 많이 사용하는 서열 기반의 중복 제거 방법보다 중복이 없고, 대표 서열 조각을 가진 학습 데이터를 구축한다. 특징 기반의 중복 데이터 제거 방법에 의해 구성된 데이터로 학습된 예측 모델은 높은 정확도로 단백질 결합 염기를 찾아냈다(5배 교차 유효성 검사 86.4%, 독립적인 테스트 68.1%).

[0005] 이에, 본 발명자들은 이전 연구의 확장으로, RNA의 결합 상대방을 고려하여 RNA에서 단백질 결합 부위의 예측 방법을 제안한다. 본 발명은 결합 상대방을 고려하고, 서열 데이터로부터 단백질 결합 염기를 예측하는 두 가지 방법을 비교 분석하여, 단백질 결합 염기를 예측하는 개선된 방법을 제공한다.

발명의 내용

해결하려는 과제

[0006] 본 발명의 목적은 RNA 및 단백질 서열을 사용하여 RNA 서열에서 단백질과 결합하는 염기를 예측하는 방법을 제공하는 것이다.

과제의 해결 수단

[0007] 상기 목적을 달성하기 위하여, 본 발명은,

- [0008] 1) RNA 서열에서 단백질과 결합할 가능성이 높은 RNA 염기를 예측하기 위해, RNA 및 단백질의 다양한 특징을 특징 벡터(feature vector)로 표현하는 단계;
- [0009] 2) 상기 단계 1)의 표현된 특징 벡터로 표현한 데이터에서 특징 벡터를 기반으로 중복 데이터를 제거하여 학습 데이터를 생성하는 단계; 및
- [0010] 3) 상기 단계 2)의 학습 데이터를 이용하여 서포트 벡터 머신(support vector machine; SVM) 모델을 학습시키고, 학습된 SVM 모델을 이용하여 단백질과 결합하는 RNA 염기를 예측하는 단계를 포함하는, RNA 서열에서 단백질과 결합하는 염기 예측방법을 제공한다.
- [0011] 또한, 본 발명은,
- [0012] 1) 단백질과 상호작용하는 단백질 결합 RNA 염기를 결정하여 단백질-RNA 결합 부위를 결정하는 단계;
- [0013] 2) RNA 서열에 있는 상기 단백질 결합 RNA 염기를 예측하기 위해, RNA 및 단백질의 다양한 특징을 특징 벡터(feature vector)로 표현하고, 테스트 데이터를 생성하는 단계; 및
- [0014] 3) 상기 단계 2)의 테스트 데이터를 복수의 척도를 이용하여, 학습된 예측 모델의 예측성능을 평가하는 단계를 포함하는 상호작용 단백질을 고려한 단백질 결합 RNA 염기 예측 방법을 제공한다.

발명의 효과

- [0015] 본 발명은 RNA 서열에서 단백질과 결합하는 염기 예측 방법에 관한 것으로, RNA 결합 상대방을 고려하여, RNA 서열에서 단백질 결합 부위를 예측하는 새로운 방법을 개발하였다. RNA 및 단백질 서열 모두를 사용하여 단백질이 결합하는 염기를 예측하는 SVM 모델을 개발하였고, 비교를 위해 RNA 서열만을 사용하여 단백질이 결합하는 염기를 예측하는 SVM 모델을 개발하여, RNA 서열만을 사용하여 예측하는 것보다 RNA 및 단백질 서열 모두를 사용하여 예측하는 것이 대부분의 성과 측정에서 우수함을 확인하였고, 이러한 새로운 모델은 결합 상대방이 변경될 때 주어진 RNA 서열에 대해 서로 다른 결합 부위를 예측할 수 있음을 확인하였다. 이는 구조를 알 수 없더라도 서열 데이터만으로 단백질과 결합하는 RNA의 염기를 예측할 수 있고, 이는 생화학적 실험에 유용하게 사용될 수 있다.

도면의 간단한 설명

- [0016] 도 1은 본 발명의 실시 예에 따른 학습 및 테스트 데이터로 이루어진 두 개의 예측 모델을 구성하는 과정을 나타낸 도이다.
- 도 2는 수소 결합(H-bonds), 물 브릿지(water bridges) 및 소수성 상호작용(hydrophobic interactions)으로 구성된 3가지 타입의 RNA-단백질 상호작용과 단백질-RNA 복합체의 결합 부위를 나타낸 도이다:
- + : 결합 염기; 및
- : 비결합 염기.
- 도 3은 윈도우 크기가 7일 경우 특징 벡터의 구조를 나타내는 도이다.
- 도 4는 본 발명의 실시 예에 따른 두 가지 예측 모델의 수신자 조작 특성(Receiver operating characteristic, ROC) 곡선을 나타낸 도이다.
- 도 5는 서로 다른 단백질과 상호작용하는 RNA의 단백질 결합 부위를 예측하는 예시를 나타낸 도이다.
- 도 6은 RNA 서열 데이터만을 가지고 RNA에서 단백질 결합 부위를 예측하는 예시를 나타낸 도이다.

발명을 실시하기 위한 구체적인 내용

- [0017] 이하, 첨부된 도면을 참조하여 본 발명의 실시 예를 상세히 설명한다. 본 발명을 설명함에 있어서 관련된 공지 기능 혹은 구성에 대한 구체적인 설명은 본 발명의 요지를 모호하게 하지 않기 위해 생략한다.

- [0018] 본 발명은
- [0019] 1) RNA 서열에서 단백질과 결합할 가능성이 높은 RNA 염기를 예측하기 위해, RNA 및 단백질의 다양한 특징을 특징 벡터로 표현하는 단계;
- [0020] 2) 상기 단계 1)의 표현된 특징 벡터로 표현한 데이터에서 특징 벡터를 기반으로 중복 데이터를 제거하여 학습 데이터를 생성하는 단계; 및
- [0021] 3) 상기 단계 2)의 학습 데이터를 이용하여 SVM 모델을 학습시키고, 학습된 SVM 모델을 이용하여 단백질과 결합하는 RNA 염기를 예측하는 단계를 포함하는, RNA 서열에서 단백질과 결합하는 염기 예측방법을 제공한다.
- [0022] 상기 단계 1)은 RNA 서열에서 단백질과 결합할 가능성이 높은 RNA 염기를 예측하기 위해, RNA 및 단백질의 다양한 특징을 특징 벡터에 표현하는 단계로 특징 벡터 안에 RNA 및 단백질 서열들의 특징들을 표현할 수 있으며, (1) RNA의 전체 특징, (2) RNA의 국소적(local) 특징 및 (3) 결합 상대방의 특징으로 세 가지로 분류할 수 있다.
- [0023] 다양한 서열 패턴을 생성하기 위해, 슬라이딩 윈도우 방법을 이용하여 RNA 및 단백질 서열의 특징을 특징 벡터로 표현했다.
- [0024] (1) RNA의 전체 특징;
- [0025] (2) RNA의 국소적 특징; 및
- [0026] (3) 결합 상대방의 특징으로 세 가지로 분류할 수 있다.
- [0027] RNA의 전체 특징은 전체 RNA 서열에 대한 정보이고, RNA의 국소적 특징은 RNA의 개별 염기 또는 염기 트리플렛에 대한 정보이며, 결합 상대방의 특징은 RNA의 결합 상대방인 단백질에 대한 정보이다.
- [0028] 구체적으로, RNA의 전체 특징은 서열 길이와 염기 구성요소를 포함하고, 상기 염기 구성요소는 입력된 RNA 서열에서 네 가지 염기의 빈도수로 나타낸다(수학식 1). RNA의 전체 특징은 서열 길이(L)와 네 가지 염기 구성요소(C_A, C_C, C_G, C_U)인 특징 벡터의 다섯 가지 요소가 요구된다.

수학식 1

$$\text{염기 구성 요소}(n)_{n \in \{A, C, G, U\}} = \sum_{i, n_i = n}^{\text{RNA 서열 길이}} \text{빈도}(n)$$

- [0029]
- [0030] 상기 i는 RNA 서열에서 염기의 위치번호, n은 4가지 RNA 염기(A, C, G, U) 중 하나, n_i 는 RNA 서열의 i번째 염기를 나타냄.
- [0031] 또한, RNA의 국소적 특징은 염기의 분자량(M)과 pKa 값(value), 염기 트리플렛(서열에서 연속한 염기 3개)과 20개의 아미노산의 상호작용 성향(interaction propensity, IP) 및 염기 트리플렛 구성요소를 포함한다.
- [0032] 염기 트리플렛의 상호작용 성향은 이전 연구에서 개발된 기능을 이용하여 단백질-RNA 복합체의 구조 데이터로부터 계산되었다(S. Choi and K. Han, Predicting protein-binding RNA nucleotides using the feature-based removal of data redundancy and the interaction propensity of nucleotide triplets, Computers in Biology and Medicine, Vol. 43, No. 11, 1687-1697, 2013). 각 서열 조각에서 염기 트리플렛 구성요소는 서열 조각에서 나타나는 64개($=4 \times 4 \times 4$)의 염기 트리플렛의 빈도수로 나타낸다(수학식 2). 또한, 염기 트리플렛 구성요소 외에도 국소적 특징을 나타내는 특징 벡터는 $22 \times (\text{윈도우 크기})$ 개의 요소로 구성된다. 여기서 하나의 구성요소는 분자량에 관한 것이고, 하나의 구성요소는 pKa 값에 관한 것이며, 20개의 구성요소는 20개의 아미노산과 염기 트리플렛의 상호작용 성향에 관한 것이다.

[0033] 즉 윈도우 크기가 7일 경우 $22 \times 7 + 64 = 218$ 개의 구성 요소를 포함한다.

수학식 2

$$\text{트리플렛 구성 요소}(t)_{t \in \{64 \text{개 트리플렛}\}} = \sum_{i, t_i=t}^{\text{RNA서열길이}} \text{빈도}(t)$$

[0034]

[0035] 상기 i 는 RNA 서열에서 염기의 위치번호, t_i 는 RNA 서열에서 i 번째 염기 트리플렛을 나타냄.

[0036] 결합 상대방의 특징은 아미노산의 정규화된 위치(normalized position, NP)와 단백질의 다이펩타이드 구성요소(dipeptide composition, DC)의 합을 포함한다. 서열 길이 N 내 위치 i 에서 아미노산의 정규화된 위치는 서열에서의 상대적인 위치(i/N)이다. 각 아미노산을 위해, 하기 수학식 3을 이용하여 아미노산의 정규화된 위치의 합을 계산하였다.

수학식 3

$$\text{정규화된 위치}(a)_{a \in \{20 \text{개 아미노산}\}} = \sum_{j, a_j=a}^{\text{단백질서열길이}} \frac{\text{위치}(a)}{\text{서열 길이}}$$

[0037]

[0038] 상기 j 는 단백질 서열에서 아미노산의 위치 번호, a_j 는 단백질 서열에서 j 번째 아미노산을 나타냄.

[0039] 또한, 다이펩타이드 구성요소(dipeptide composition, DC)는 단백질 서열에서 두 개의 연속된 아미노산의 빈도수를 나타낸다(수학식 4).

수학식 4

$$\text{다이펩타이드 구성 요소}(d)_{d \in \{40 \text{개 다이펩타이드}\}} = \sum_{j, d_j=d}^{\text{단백질서열길이}} \text{빈도}(d)$$

[0040]

[0041] 상기 j 는 단백질 서열에서 아미노산의 위치 번호, d_j 는 단백질 서열에서 j 번째 다이펩타이드를 나타냄.

[0042] 결합 상대방 특징은 20개의 정규화된 위치와 $400 (= 20 \times 20)$ 개의 다이펩타이드 구성요소로, 총 420개의 구성요소가 필요하다. 결합 상대방 특징으로 인해, 동일한 RNA 서열은 다른 단백질 서열과 상호작용할 경우, 다른 특징 벡터로 표현될 수 있다.

[0043] RNA 서열에 슬라이딩 윈도우를 적용하여, 윈도우 중앙에 있는 염기 하나를 표현하기 위해 특징 벡터 하나가 생성된다. w 개의 염기를 포함하는 윈도우는 $w-2$ 개의 염기 트리플렛에 해당한다. 윈도우의 처음과 마지막 염기는 트리플렛을 형성하지 않기 때문에 이의 IP 값은 0으로 설정하였다. 도 3은 윈도우 크기가 7일 경우 특징 벡터의 구조를 나타낸다.

- [0044] 또한, 상기 단계 2)는 표현된 특징 벡터를 기반으로 중복 데이터를 제거하여 학습 데이터를 생성하는 단계로, 이전 연구(Choi S. and Han K. Prediction of RNA-binding amino acids from protein and RNA sequences. *BMC Bioinformatics* 2011, 12(Suppl 13):S7, Choi S. and Han K. Predicting protein-binding RNA nucleotides using the feature-based removal of data redundancy and the interaction propensity of nucleotide triplets. *Computers in Biology and Medicine* 2013, 43:1687-1697)를 기초로, RNA 및 단백질 서열은 특징 벡터로 표현되고, 특징 벡터 기반의 중복 데이터 제거 방법에 의해 중복된 특징 벡터는 제거되었다. 이러한 방법에 의해 중복이 없는 서열 조각과 각각의 특징 벡터로 구성된 학습 데이터를 생성하였다.
- [0045] 또한, 본 발명자들은 레이디얼 기초 함수(radial basis function, RBF) 커널과 SVM의 라이브러리(LIBSVM)를 이용하여 SVM 모델을 구축하였다. RBF 커널과 함께 SVM 모델을 학습시킬 때, C와 γ 두 가지 파라미터 값을 조정해야 한다. 파라미터 C는 결정 표면의 단순성에 대해 학습 데이터의 오분류의 균형을 위한 것이고, 파라미터 γ 는 RBF의 폭을 나타낸다. 파라미터 최적화에 의해 C와 γ 의 최상의 조합을 선택하였고, 본 명세서에 기재한 모든 결과는 C=1, $\gamma=0.03125$ 를 사용하여 실행한 결과이다. 학습 데이터에서 비결합 염기와 결합 염기의 비율의 균형을 맞추기 위해, 결합 염기에 4의 가중치를 할당하였다.
- [0046] 또한, 본 발명은
- [0047] 1) 단백질과 상호작용하는 단백질 결합 RNA 염기를 결정하여 단백질-RNA 결합 부위를 결정하는 단계;
- [0048] 2) RNA 서열에 있는 상기 단백질 결합 RNA 염기를 예측하기 위해, RNA 및 단백질의 다양한 특징을 특징 벡터(feature vector)로 표현하고, 테스트 데이터를 생성하는 단계; 및
- [0049] 3) 상기 단계 2)의 테스트 데이터를 복수의 척도를 이용하여, 학습된 예측 모델의 예측성도를 평가하는 단계를 포함하는 상호작용 단백질을 고려한 단백질 결합 RNA 염기 예측 방법을 제공한다.
- [0050] 상기 단계 1)의 단백질과 상호작용은 수소 결합, 물 브릿지 및 소수성 상호작용으로 구성된 세 가지 유형의 상호작용을 고려하였다. 상호작용 중 적어도 하나에 관련된 염기는 RNA에서 단백질 결합 부위로 분류하였다. 542개의 단백질-RNA 복합체로부터 각각 핵산-단백질 상호작용 데이터베이스(nucleic acid-protein interaction database, NPIDB)를 이용하여 상호작용의 세 가지 유형을 수득하고, 단백질-RNA 복합체의 서열로 통합하였다.
- [0051] 예를 들면, 도 2에 나타난 바와 같이, 3LQX의 RNA 사슬 B에서 7개의 염기(U139, A140, A148, G149, G150, U151 및 C163)는 수소 결합 상호작용을 통해 단백질과 상호작용하고, 6개의 염기(A140, C141, G149, G150, G162 및 A164)는 물 브릿지 상호작용을 통해 단백질과 상호작용하며, 2개의 염기(A140 및 A148)는 소수성 상호작용을 통해 단백질과 상호작용한다. 이러한 세 가지 유형의 상호작용을 RNA 사슬 B에 병합하여 나타내었다(도 2).
- [0052] 아울러, 상기 단계 3)는 테스트 데이터를 복수의 척도를 사용하여 평가하여 상호작용 단백질을 고려한 단백질 결합 RNA 염기를 예측하는 단계로, 예측 모델의 성능은 수학적 5 내지 수학적 10의 민감도(sensitivity, Sn), 특이도(specificity, Sp), 정확도(accuracy, Acc), 양성예측도(positive predictive value, PPV), 음성예측도(negative predictive value, NPV) 및 매튜 상관관계 계수(Matthews correlation coefficient, MCC)의 6가지로 평가하였다. 다음 방정식에서, 참된 양성(true positive, TP)은 결합 염기로 예측된 결합 염기, 참된 음성(true negative, TN)은 비결합 염기로 예측된 비결합 염기, 거짓 양성(false positive, FP)은 결합 염기로 예측된 비결합 염기 및 거짓 음성(false negative, FN)은 비결합 염기로 예측된 결합 염기를 나타낸다.
- [0053]

수학식 5

$$\text{민감도} = \frac{TP}{TP+FN}$$

[0054]

수학식 6

$$\text{특이도} = \frac{TN}{TN+FP}$$

[0055]

수학식 7

$$\text{정확도} = \frac{TP+TN}{TP+TN+FP+FN}$$

[0056]

수학식 8

$$\text{양성예측도} = \frac{TP}{TP+FP}$$

[0057]

수학식 9

$$\text{음성예측도} = \frac{TN}{TN+FN}$$

[0058]

수학식 10

$$\text{매튜 상관관계 계수} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

[0059]

[0060]

민감도는 올바르게 예측된 결합 염기와 실제 결합 염기의 비율이고, 특이도는 올바르게 예측된 비결합 염기와 실제 비결합 염기의 비율이며, 정확도는 올바르게 예측된 염기와 모든 염기의 비율이고, 양성예측도는 올바르게

예측된 결합 염기와 결합으로 예측되는 모든 염기의 비율이며, 음성예측도는 올바르게 예측된 비결합 염기와 비결합으로 예측되는 모든 염기의 비율이고, 매튜 상관관계 계수는 실제와 예측 분류 사이의 상관관계를 나타내며, -1과 1 사이의 점수로 나타낸다. -1은 실제와 예측 사이의 총 불일치를 나타내고, +1은 일치를 나타내며, 0은 완전한 임의의 예측을 나타낸다.

[0061] 이하, 본 발명의 방법의 구체적인 실험 예를 설명한다.

[0062] 도 1에 도시된 바와 같이, 단백질-RNA 복합체는 PDB(protein data bank)에서 획득되었는데, 2013년 9월을 기준으로 했을 때, X-선 결정학(X-ray crystallography)에 의해 3.0 Å 또는 그 이상의 해상도를 갖는 구조가 밝혀진 단백질-RNA 복합체 중에서, 668개의 단백질-RNA 복합체 구조 데이터들이 수집되었고, 학습 데이터를 구성하기 위해 668개의 단백질-RNA 복합체에서 염기의 극히 일부분(약 2%)만 단백질에 결합하는 리보솜 RNA(ribosomal RNA)를 포함하는 복합체를 제거하여, 542개의 단백질-RNA 복합체들이 남겨졌다. 542개의 단백질-RNA 복합체들로부터 439개의 RNA 서열과 376개의 단백질 서열로 구성된 546개의 단백질-RNA 서열 쌍을 추출하였다.

[0063] 또한, 학습 및 테스트 데이터 사이의 중첩을 줄이기 위해, 다른 서열과 80% 또는 그 이하의 유사성을 가지는 RNA 서열을 선택하고, 상기 RNA 서열로 테스트 데이터를 구성하였다. 테스트 데이터는 155개의 RNA 서열과 RNA 서열의 상호작용 상대방(interacting partner)인 130개의 단백질 서열을 포함하고, 나머지 284개의 RNA 서열과 246개의 단백질 서열은 학습 데이터를 구성하는데 사용하였다(도 1).

[0064] 본 발명자들은 비교를 위하여, 두 개의 예측 모델을 생성하였고, 각각 M1 및 M2로 기재하였다. M1은 RNA 서열과 단백질 서열 두 가지 모두 사용하는 예측 모델이고, M2는 RNA 서열만을 사용하는 예측 모델이다.

[0065] M1 모델을 시험하기 위한 테스트 데이터는 155개의 RNA 서열에 1,848개의 단백질-결합 염기 및 4,631개의 비결합 염기를 포함하고, M2 모델은 1,795개의 단백질-결합 염기 및 4,235개의 비결합 염기를 포함하며, 이는 다른 단백질 상대방을 갖는 RNA 서열 결합 부위가 통합되므로, 개수의 차이를 갖는다. 또한, M1을 위한 학습 데이터는 2,716개의 단백질-결합 염기 및 6,432개의 비결합 염기를 포함하고, M2는 2,189개의 단백질-결합 염기 및 4,588개의 비결합 염기를 포함한다. M1 및 M2는 동일한 RNA 서열로부터 구축되었으나, M1은 단백질 서열의 추가적인 특징으로 인해 M2보다 결합 또는 비 결합 염기를 더 포함한다.

[0066] 본 발명의 방법을 이용하여, 다양한 윈도우 크기로의 10배 교차 검증에서 두 모델의 성능을 확인한 결과, 하기 표 1에 나타난 바와 같이, 두 가지 모델 모두 33개의 염기의 윈도우에서 최상의 성능을 보였다. M1 모델은 RNA와 단백질 서열 모두를 사용하여, 민감도는 86.5%, 특이도는 86.2%, 정확도는 86.3%, 양성예측도는 72.6%, 음성예측도는 93.8% 및 매튜 상관관계 계수는 0.69를 나타내었다. 또한, M2 모델은 RNA 서열만을 사용하여, 민감도는 85.7%, 특이도는 80.5%, 정확도는 82.2%, 양성예측도는 67.7%, 음성예측도는 92.2% 및 매튜 상관관계 계수는 0.63을 나타내었다(표 1).

표 1

ws	M1						M2					
	Sn (%)	Sp (%)	Acc (%)	PPV (%)	NPV (%)	MCC	Sn (%)	Sp (%)	Acc (%)	PPV (%)	NPV (%)	MCC
3	92.3	44.0	61.4	48.2	91.0	0.38	99.8	1.9	39.7	39.0	95.0	0.08
5	90.1	72.6	77.8	58.4	94.5	0.58	87.1	53.9	64.7	47.8	89.7	0.39
7	90.2	76.9	80.8	62.0	94.9	0.62	86.2	60.0	68.4	50.5	90.2	0.43
9	90.4	79.5	82.7	64.9	95.2	0.65	87.2	62.6	70.5	52.5	91.2	0.47
11	89.9	81.4	83.9	67.1	95.0	0.67	87.8	65.7	72.8	55.0	91.9	0.50
13	89.8	82.5	84.7	68.4	95.0	0.68	87.8	68.9	75.0	57.3	92.2	0.53
15	89.0	82.4	84.4	68.1	94.7	0.67	87.1	71.3	76.4	59.1	92.1	0.55
17	89.3	83.3	85.1	69.3	94.8	0.68	88.3	72.4	77.5	60.4	92.8	0.57
19	87.8	84.1	85.2	69.9	94.2	0.68	88.3	73.8	78.5	61.6	93.0	0.58
21	88.1	84.2	85.3	70.1	94.4	0.68	88.0	75.0	79.2	62.6	92.9	0.59
23	87.5	85.1	85.8	71.2	94.2	0.69	87.1	76.1	79.7	63.5	92.5	0.59
25	86.3	85.2	85.5	71.1	93.7	0.68	87.0	77.2	80.4	64.6	92.6	0.61
27	86.2	85.4	85.7	71.4	93.6	0.68	86.7	78.4	81.1	65.7	92.5	0.62
29	86.0	85.9	85.9	71.9	93.6	0.69	86.4	78.8	81.2	66.0	92.4	0.62
31	86.1	86.0	86.0	72.2	93.6	0.69	85.6	79.9	81.8	67.0	92.1	0.62
33	86.5	86.2	86.3	72.6	93.8	0.69	85.7	80.5	82.2	67.7	92.2	0.63
35	85.9	86.3	86.2	72.7	93.5	0.69	85.3	80.5	82.0	67.7	91.9	0.63
37	85.1	86.4	86.0	72.6	93.2	0.69	84.8	81.4	82.5	68.6	91.8	0.63
39	85.5	86.4	86.1	72.7	93.3	0.69	84.6	81.4	82.4	68.6	91.6	0.63
41	84.8	86.5	86.0	72.8	93.0	0.69	84.5	81.9	82.7	69.1	91.6	0.63
43	85.1	86.7	86.2	73.1	93.2	0.69	83.5	82.3	82.7	69.3	91.2	0.63
45	85.0	86.8	86.3	73.3	93.2	0.69	83.7	82.1	82.6	69.1	91.3	0.63

[0067]

[0068]

보다 정확한 평가를 위해, 모델의 학습에 사용되지 않은 독립적인 데이터로 SVM 모델로 테스트한 결과, 하기 표 2에 나타낸 바와 같이, 33개의 염기의 윈도우에서 RNA와 단백질 서열 모두를 사용한 경우에는 민감도는 58.8%, 특이도는 87.4%, 정확도는 79.2%, 양성예측도는 65.1%, 음성예측도는 84.2% 및 매튜 상관관계 계수는 0.48을 나타내었고, RNA 서열만을 사용한 경우에는 민감도는 67.7%, 특이도는 78.8%, 정확도는 75.5%, 양성예측도는 57.6%, 음성예측도는 85.2% 및 매튜 상관관계 계수는 0.45를 나타내었다(표 2).

표 2

ws	M1						M2					
	Sn (%)	Sp (%)	Acc (%)	PPV (%)	NPV (%)	MCC	Sn (%)	Sp (%)	Acc (%)	PPV (%)	NPV (%)	MCC
3	92.7	29.5	47.5	34.4	91.0	0.24	98.8	4.7	32.7	30.5	90.1	0.09
5	81.4	58.0	64.7	43.6	88.7	0.36	77.8	52.8	60.2	41.1	84.9	0.28
7	80.7	62.6	67.8	46.3	89.1	0.39	76.3	57.5	63.1	43.2	85.2	0.31
9	81.0	66.2	70.4	48.9	89.7	0.43	76.3	59.0	64.2	44.1	85.4	0.32
11	81.1	68.9	72.4	51.0	90.1	0.45	77.3	62.0	66.6	46.3	86.6	0.36
13	78.8	70.3	72.7	51.4	89.3	0.45	79.3	63.8	68.5	48.2	87.9	0.39
15	77.2	72.9	74.1	53.1	88.9	0.46	77.3	65.2	68.8	48.5	87.2	0.39
17	73.5	75.8	75.1	54.8	87.8	0.46	77.3	67.1	70.1	49.9	87.5	0.41
19	70.6	78.4	76.2	56.6	87.0	0.46	75.5	69.2	71.1	51.0	87.0	0.41
21	68.8	80.2	77.0	58.2	86.6	0.47	74.7	70.8	72.0	52.0	86.8	0.42
23	67.2	81.6	77.5	59.3	86.2	0.47	72.6	73.4	73.2	53.6	86.4	0.43
25	65.6	82.8	77.9	60.4	85.8	0.47	69.3	75.4	73.6	54.4	85.3	0.42
27	63.5	84.1	78.2	61.5	85.2	0.47	68.4	76.5	74.1	55.2	85.1	0.43
29	61.6	85.5	78.7	62.9	84.8	0.47	68.2	77.3	74.6	56.0	85.1	0.43
31	60.8	86.5	79.2	64.3	84.7	0.48	68.0	78.0	75.0	56.7	85.2	0.44
33	58.8	87.4	79.2	65.1	84.2	0.48	67.7	78.8	75.5	57.6	85.2	0.45
35	57.6	88.0	79.3	65.6	83.9	0.47	66.8	79.7	75.9	58.3	85.0	0.45
37	56.8	88.4	79.4	66.1	83.7	0.47	65.7	80.6	76.1	58.9	84.7	0.45
39	54.8	88.7	79.0	66.0	83.1	0.46	65.1	81.0	76.3	59.2	84.5	0.45
41	54.7	88.9	79.1	66.2	83.1	0.46	64.3	81.5	76.4	59.5	84.3	0.45
43	54.3	89.0	79.1	66.3	83.0	0.46	63.3	82.0	76.4	59.8	84.0	0.45
45	54.2	88.9	79.0	66.2	82.9	0.46	62.7	82.1	76.4	59.8	83.9	0.44

[0069]

[0070]

교차 검증과 독립적인 테스트 두 가지에서 단백질과 RNA 서열 데이터를 모두 사용한 M1 모델은 RNA 서열 데이터만을 사용한 M2 모델보다 더 나은 성능을 보였다. 이는 RNA 결합 파트너에 대한 부가 정보의 중요성을 보여준다. 도 4는 33개 염기의 윈도우에서 두 예측 모델의 수신자 조작 특성(Receiver operating characteristic,

ROC)과 커브 아래 면적(area under the curve, AUC)을 보여준다.

[0071] 도 5는 서로 다른 단백질과 상호작용하는 RNA의 단백질 결합 부위를 예측하는 예시를 나타낸다. 도 5는 RNA 서열과 단백질 서열 데이터를 모두 사용한 단백질 결합 염기의 예측을 나타내고 있다. 단백질-RNA 복합체 3HJW는 다른 아미노산 서열을 가진 두 개의 단백질 사슬을 포함하고 있다. 예측 모델은 RNA 및 단백질 서열을 모두 사용하는데, RNA 결합 상대방이 다르기 때문에 같은 RNA 서열이라도 다른 결합 부위를 예측하였다. 예측 모델의 실행에서 RNA 결합 상대방으로 사용된 단백질은 옐로우로 나타내었다(도 5A에서 사슬 A, 도 5B에서 사슬 B)(도 5).

[0072] 도 6은 RNA 서열 데이터만을 가지고 동일한 단백질-RNA 복합체 3HJW에서 단백질 결합 염기를 예측하는 또 다른 예시를 나타내고 있다. RNA의 결합 상대방이 정해지지 않았기 때문에, RNA 서열에서 단백질과 결합하는 부위는 단백질에 대한 정보를 이용하여 예측한 결과인 도 5와 다르게 예측되었다(도 6).

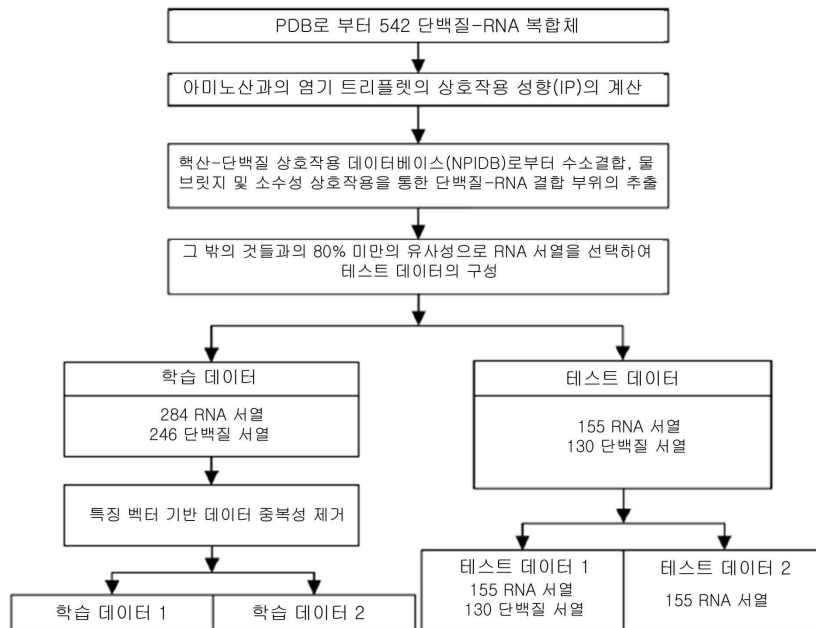
[0073] 상술한 바와 같이, 본 발명자들은 RNA 결합 상대방을 고려하여, RNA 서열에서 단백질 결합 부위를 예측하는 새로운 방법을 개발하였다. RNA 및 단백질 서열 모두를 사용하여 단백질에 결합하는 염기를 예측하는 SVM 모델을 개발하였고, 비교를 위해 RNA 서열만을 사용하여 단백질에 결합하는 염기를 예측하는 SVM 모델을 개발하였다. RNA 서열과 단백질 서열을 모두 사용하는 모델의 경우, 10배 교차 검증에서 86.3%의 정확도와 0.69의 매튜 상관관계 계수를 달성하였고, 독립적인 테스트에서 79.2%의 정확도와 0.48의 매튜 상관관계 계수를 달성하였다. 또한, RNA 서열만을 사용하는 모델의 경우에는 10배 교차 검증에서 82.2%의 정확도와 0.63의 매튜 상관관계 계수를 달성하였고, 독립적인 테스트에서 75.5%의 정확도와 0.45의 매튜 상관관계 계수를 달성하였다. 이러한 검증 결과를 통해, RNA 서열만을 사용하여 예측하는 것보다 RNA 및 단백질 서열 모두를 사용하여 예측하는 것이 대부분의 성과 측정에서 우수함을 확인하였고, 이러한 새로운 모델은 결합 상대방이 변경될 때 주어진 RNA 서열에 대해 서로 다른 결합 부위를 예측할 수 있음을 확인하였다. 이는 구조를 알 수 없더라도 서열 데이터만으로 단백질 결합 RNA의 염기를 예측할 수 있고, 이는 생화학적 실험에 유용하게 사용될 수 있다.

[0074] 이상과 같이 본 발명에 따른 RNA 서열에서 단백질과 결합하는 염기를 예측하는 방법을 예시한 도면을 참조하여 설명하였으나, 본 명세서에 개시된 실시 예와 도면에 의해 본 발명이 한정되는 것은 아니며, 본 발명의 기술 사상 범위 내에서 당업자에 의해 다양한 변형이 이루어질 수 있음은 물론이다.

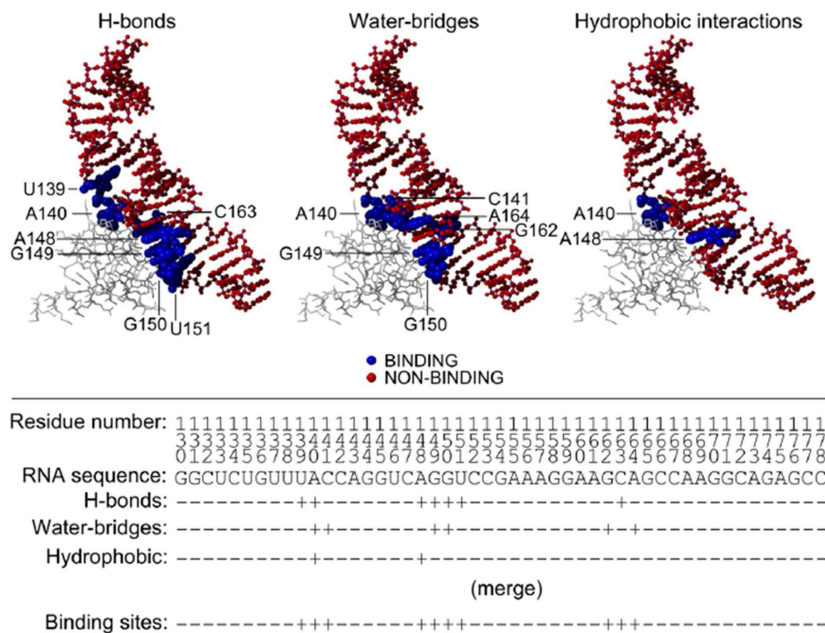
[0075]

도면

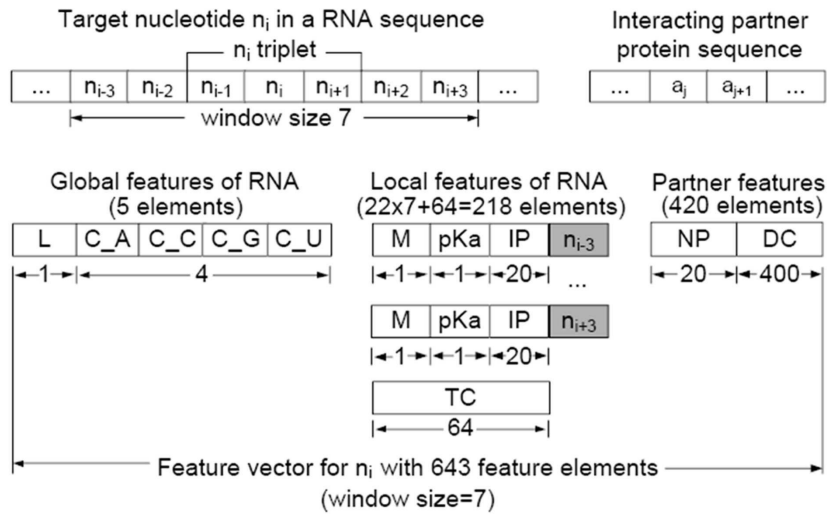
도면1



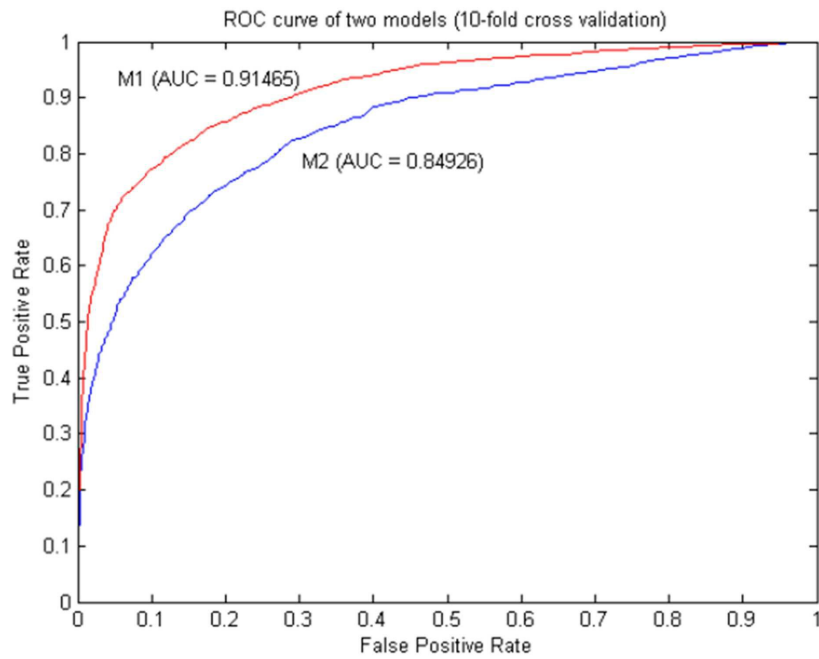
도면2



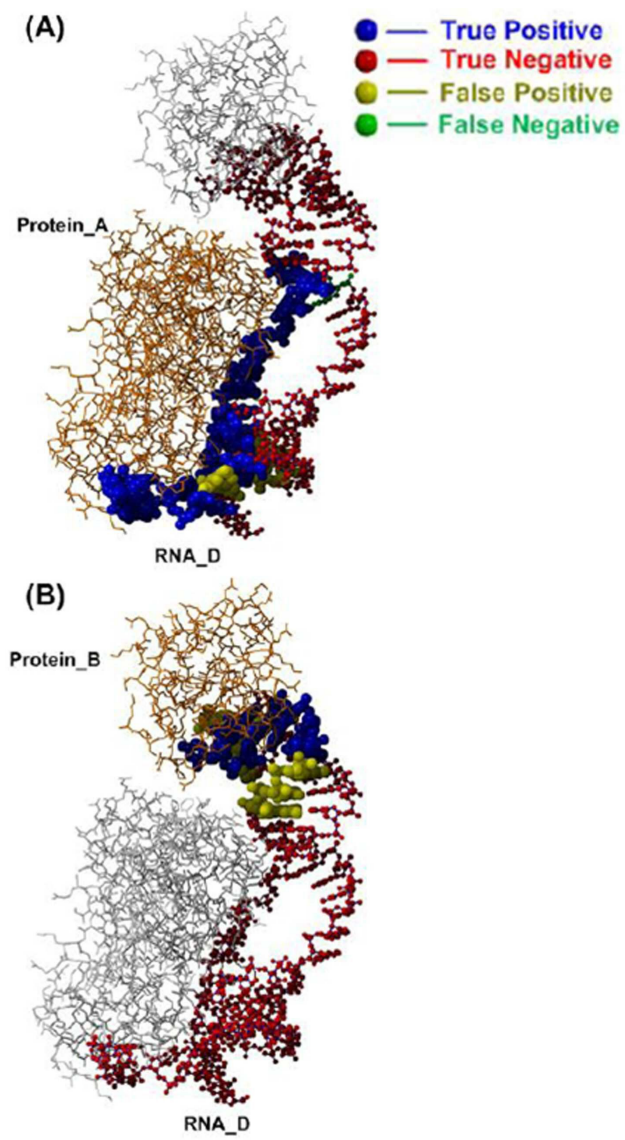
도면3



도면4



도면5



도면6

