(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2013-0035732
(43) 공개일자 2013년04월09일

(51) 국제특허분류(Int. Cl.)
G06F 19/10 (2011.01)
(21) 출원번호 10-2011-0100228
(22) 출원일자 2011년09월30일
심사청구일자 2011년09월30일

(71) 출원인
인하대학교 산학협력단
인천광역시 남구 인하로 100, 인하대학교 (용현동)
(72) 발명자
한경숙
서울시 강서구 등촌동 아이파크 109동 201호
최성욱
경기도 과천시 별양로 163, 507동 707호 (별양동, 주공아파트)
(74) 대리인
이원희

전체 청구항 수 : 총 12 항

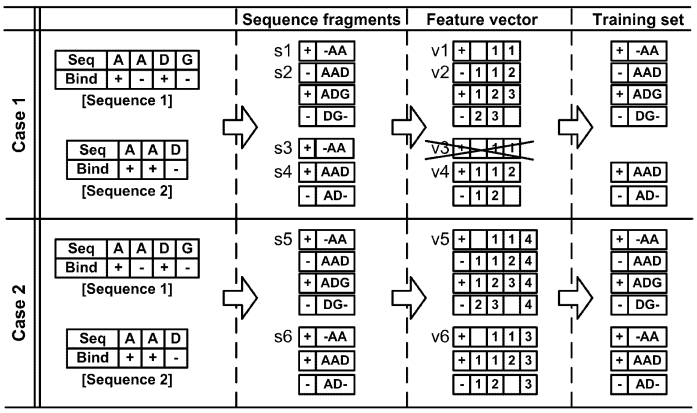
(54) 발명의 명칭 **특징 벡터에 기반한 훈련 데이터의 중복 제거 방법**

(57) 요 약

본 발명은 특징 벡터에 기반하여, 중복 데이터가 제거된 훈련 데이터를 생성하는 방법에 관한 것으로, 해결하고자 하는 기술적 과제는 종래의 서열 유사성을 바탕으로 한 방법보다 불필요한 중복 데이터를 포함하지 않으면서도 더욱 큰 규모의 데이터 세트를 생성할 수 있고, 특징 벡터에 단백질과 RNA 서열 모두의 특징을 포함함으로써 단백질의 특징만을 사용하는 종래의 방법보다 더욱 효과적으로 단백질 서열에 존재하는 RNA-결합 아미노산을 예측할 수 있는 특징 벡터에 기반한 훈련 데이터 생성 방법을 제공하는데 있다.

이를 위해 본 발명에 따른 본 발명에 따른 특징 벡터에 기반한 훈련 데이터 생성 방법은 (1) RNA와 수소결합으로 상호작용하는 RNA-결합 아미노산을 결정하여 단백질-RNA 결합 부위를 결정하는 단계와, (2) 상기 RNA와 상호작용하는 아미노산 트리플릿의 경향성을 계산하는 단계와, (3) 단백질 서열에 있는 상기 RNA-결합 아미노산을 예측하기 위해 단백질 및 RNA 서열의 다양한 특징을 특징 벡터에 코드화하는 단계와, (4) 상기 코드화된 특징 벡터를 기반으로 데이터의 불필요한 중복을 제거한 훈련 데이터 세트를 구축하는 단계를 포함하는 것을 특징으로 한다.

대 표 도 - 도6



이 발명을 지원한 국가연구개발사업

과제고유번호	2008-0058358
부처명	교육과학기술부
연구사업명	중견연구자지원사업
연구과제명	상호작용하는 RNA와 단백질에서의 결합부위의 예측
주관기관	인하대학교 산학협력단
연구기간	2008.05.01 ~ 2011.02.28

특허청구의 범위

청구항 1

(1) RNA와 수소결합으로 상호작용하는 RNA-결합 아미노산을 결정하여 단백질-RNA 결합 부위를 결정하는 단계;

(2) 상기 RNA와 상호작용하는 아미노산 트리플릿의 경향성을 계산하는 단계;

(3) 단백질 서열에 있는 상기 RNA-결합 아미노산을 예측하기 위해 단백질 및 RNA 서열의 다양한 특징을 특징 벡터에 코드화하는 단계;

(4) 상기 코드화된 특징 벡터를 기반으로 데이터의 불필요한 중복을 제거한 훈련 데이터 세트를 구축하는 단계;

를 포함하는 것을 특징으로 하는 특징 벡터에 기반한 훈련 데이터의 중복 제거 방법.

청구항 2

제 1항에 있어서,

상기 (4) 단계 후,

(5) 상기 단백질 서열에 있는 RNA-결합 아미노산의 예측 성과를 복수의 예측 성능 척도를 사용해 평가하는 단계를 더 포함하는 것을 특징으로 하는 특징 벡터에 기반한 훈련 데이터의 중복 제거 방법.

청구항 3

제 1항 또는 제 2항에 있어서,

상기 (1) 단계에서,

단백질과 RNA의 수소원자 (H), 수소 공여자 (D, donor), 수소 수용자 (A, acceptor), acceptor antecedent (AA)에 해당하는 원자들이 다음 기하학적 조건을 만족해야 하는 것을 특징으로 하는 특징 벡터에 기반한 훈련 데이터의 중복 제거 방법.

기하학적 조건 : D-A 거리 < 3.9 Å, H-A 거리 < 2.7 Å, D-H-A 각도 ≥ 90° , H-A-AA 각도 ≥ 90°

청구항 4

제 1항 또는 제 2항에 있어서,

상기 (2) 단계에서,

상기 아미노산 트리플릿과 뉴클레오티드 사이에 존재하는 상호 작용 경향성(IP_{tb})은 하기의 수학식에 의해 결정되는 것을 특징으로 하는 특징 벡터에 기반한 훈련 데이터의 중복 제거 방법.

$$IP_{tb} = \sum_{\substack{t = AA\ triplet \\ b = nucleotide}} \frac{1}{\overline{HA_{tb}}\cos(\angle DAH) \cdot N_{PR}} \Big/ \frac{N_t \cdot N_b}{N_P \cdot N_R}$$

(여기서, $\overline{HA_{tb}}\cos(\angle DAH)$ 는 아미노산 트리플릿 t와 결합하는 뉴클레오티드 b 사이의 D-A에 H-A를 투영한 거리의 역수를 전부 합한 값, N_{PR}은 뉴클레오티드에 결합하는 모든 아미노산 트리플릿의 총 개수, N_t는 아미노산 트리플릿 t의 개수, N_P는 아미노산 트리플릿의 총 개수, N_b는 뉴클레오티드 b의 개수, N_R은 데이터 세트에 있는 뉴클레오티드의 총 개수를 의미한다.)

청구항 5

제 1항 또는 제 2항에 있어서,

상기 (3) 단계에서,

특징 벡터에 구성요소로서 코드화되는 단백질 및 RNA 서열의 특징은,

목표 아미노산의 전체 서열 정보를 표현하는 단백질 서열의 전역적인 특징;

아미노산 개개의 특징을 대표하는 아미노산의 국지적인 특징; 및

상기 단백질과 상호 작용하는 RNA 서열의 특징을 포함하는 것을 특징으로 하는 특징 벡터에 기반한 훈련 데이터의 중복 제거 방법.

청구항 6

제 5항에 있어서,

전역적인 특징으로는 해당 단백질 서열의 길이 및 각 아미노산의 빈도수가 특징 벡터 코드화에 구성요소로 사용되는 것을 특징으로 하는 특징 벡터에 기반한 훈련 데이터의 중복 제거 방법.

청구항 7

제 5항에 있어서,

국지적 특징으로는 해당 서열에서의 목표 아미노산의 정규화된 위치 (Normalized position), 소수성 (Hydropathy), 접근 가능한 표면적 (Accessible surface area), 분자량 (molecular mass), 아미노산 곁사슬의 pKa 값, 계산된 아미노산 트리플릿의 IP값이 특징 벡터 코드화에 구성요소로 사용되는 것을 특징으로 하는 특징 벡터에 기반한 훈련 데이터의 중복 제거 방법.

청구항 8

제 5항에 있어서,

상호 작용 상대방인 RNA 서열의 특징으로는 RNA 서열에서 각각의 뉴클레오티드의 정규화된 위치의 합을 코드화 구성요소로서 사용되는 것을 특징으로 하는 특징 벡터에 기반한 훈련 데이터의 중복 제거 방법.

청구항 9

제 1항 또는 제 2항에 있어서,

상기 (4) 단계는

1) 단백질-RNA 상호 작용 쌍들에 존재하는 모든 단백질 서열에서 슬라이딩 윈도우(sliding window) 기법을 이용한 서열 조각을 생성하는 공정;

2) 생성된 서열 조각을 상기 구성요소들을 이용하여 특정 벡터에 코드화하는 공정; 및

3) 코드화된 특정 벡터에서 특징 벡터 기반의 중복 제거 기법을 통한 훈련 데이터 세트를 구축하는 공정을 포함하는 것을 특징으로 하는 특징 벡터에 기반한 훈련 데이터의 중복 제거 방법.

청구항 10

제 9항에 있어서,

같은 결합 라벨을 지닌 동일한 구성 요소 수치를 갖는 특징 벡터들은 중복 데이터로서 고려되어 이러한 특징 벡터들 중 하나의 특징 벡터만이 학습 데이터에 포함되며, 각각 다른 결합 라벨을 지닌 동일한 구성 요소 수치를 갖는 특징 벡터들이 존재할 경우 각각 하나의 (하나는 결합, 그리고 하나는 비결합) 특징 벡터만이 훈련 데이터 세트에 포함되는 것을 특징으로 하는 특징 벡터에 기반한 훈련 데이터의 중복 제거 방법.

청구항 11

제 10항에 있어서,

상기 라벨은 상기 서열 조각의 중간에 위치한 아미노산이 결합 아미노산이라면 결합 (+)으로 취급되고, 그 외의 경우는 비결합 (-)으로 취급되는 것을 특징으로 하는 특징 벡터에 기반한 훈련 데이터의 중복 제거 방법.

청구항 12

제 2항에 있어서,

상기 (5) 단계에서,

상기 RNA-결합 아미노산의 예측 성과를 평가하기 위해 사용되는 복수의 예측 성능 척도는 민감도 (Sensitivity), 특이도 (Specificity), 정밀도 (Precision rate), 정확도(Accuracy), Net prediction, F-measure 및 상관 계수 (Correlation coefficient)를 포함하는 것을 특징으로 하는 특징 벡터에 기반한 훈련 데이터의 중복 제거 방법.

명 세 서

기 술 분 야

[0001] 본 발명은 특징 벡터에 기반하여 훈련 데이터의 중복을 제거하는 방법에 관한 것으로, 단백질 서열에 존재하는 RNA-결합 아미노산 예측 모델 생성을 위해 사용될 특징 벡터 기반의 학습 데이터 생성 방법에 관한 것이다.

배 경 기 술

[0002] 단백질과 RNA 사이에 존재하는 상호 작용을 연구하기 위해 많은 실험적 연구와 이론적 연구가 있었지만, 상호 작용의 정확한 메커니즘은 완전히 알려져 있지 않고 있다.

[0003] 최근 많은 수의 단백질-RNA 복합체들의 구조데이터가 밝혀짐에 따라, 단백질-RNA 복합체의 서열 정보만을 이용하여 단백질 서열에 존재하는 RNA-결합 아미노산을 예측하기 위해 다양한 시도가 이루어지고 있다. 예를 들어, Wang L, Brown SJ: BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res. 2006, 34:243-248.에 소개된 BindN은 Support Vector Machine (SVM)을 사용하여 아미노산의 화학적 특성을 바탕으로 단백질 서열에 존재하는 RNA 또는 DNA-결합 아미노산을 예측하고, Terribilini M, Lee J, Yan C, Jernigan RL, Honavar V, Dobbs D: Prediction of RNA binding sites in proteins from amino acid sequence. RNA 2006, 12:1450-1462. 및 Terribilini M, Dander JD, Lee JH, Zaback P, Jernigan RL, Honavar V, Dobbs D: RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. Nucleic Acids Res. 2007, 35:578-584.에 소개된 RNABindR은 Naive Bayes 분류기를 이용해 단백질 서열에 존재하는 RNA와 결합하는 아미노산을 예측한다.

[0004] 이와 같이, 서열 정보만을 이용한 단백질 서열에서의 RNA-결합 아미노산 예측을 위해 제안된 다양한 기법들은 서로 다른 기계학습기법과 다양한 종류의 결합 특징들을 사용하여 예측 모델을 생성하고 있지만, 공통적으로 서열 유사도 기반의 학습데이터 생성 방법을 사용하고 있다. 또한, 이들 중 어느 것도 RNA-결합 아미노산을 예측할 때, 주어진 단백질에 대한 상호 작용 상대방인 RNA를 고려하지 않고 있어, 상대방 RNA 분자가 바뀜에 따라

단백질 서열의 RNA-결합 부위가 바뀌는 자연스러운 생물학적 현상을 예측 과정에 반영하지 못한다는 제약사항이 존재한다.

[0005] 본 발명의 발명자들은 단백질과 RNA 사이의 상호 작용에 관한 연구를 수행하고 있으며, 아미노산과 뉴클레오티드의 결합-특이적 특징을 발견하기 위한 노력의 일환으로, 최근 발견된 단백질-RNA 복합체의 구조데이터를 포함하는 확장적인 분석을 통해 아미노산과 뉴클레오티드 사이에 존재하는 상호 작용 경향성을 수치화 할 수 있는 다양한 종류의 IP (Interaction Propensity) 수식들을 개발하였으며 이를 이용한 상호 작용 경향성을 측정하였다.

[0006] 본 발명자들의 분석 결과는 서열에서 연속한 3개의 아미노산 (이하, 아미노산 트리플릿으로 지칭)의 IP가 개별적인 아미노산의 IP나 아미노산의 생화학적 특성 보다 높은 결합 특이성을 보인다. 본 출원의 발명자들은 결합 성향 분석을 위하여 개발한 아미노산 트리플릿 IP의 수식을 최근까지 밝혀진 단백질-RNA 복합체의 구조데이터들에 대하여 적용하여 결합 성향을 계산하였으며 계산된 결합 성향 수치 정보를 단백질 서열에서 RNA-결합 아미노산을 예측하는 예측 모델 생성에 사용하였다.

[0007] 종래의 RNA-결합 아미노산을 예측하기 위해 제안된 다양한 방법들은 공통적으로 단백질 서열의 결합 정보와 부분적인 서열 정보를 고려하지 않은 채, 전체 서열 유사성을 기준으로 훈련 데이터를 구축한다. 하지만 이러한 서열 유사도 기반의 방법으로 구축된 훈련 데이터에서는, 결합 부위 예측에 유용하게 사용될 수 있는 결합 정보가 손실되는 단점이 있다.

[0008] 도 4는 서열 유사도를 기반으로 하는 중복 데이터 제거 과정 도중 일어나는 결합 정보의 손실을 나타내는 도이다.

[0009] 서열 유사도를 바탕으로 군집화 시켜주는 프로그램인 CD-HIT에 의해 묶여진 도 4의 단백질 서열들을 고려하였을 때, CD-HIT은 단백질-RNA 복합체 1F7Y의 단백질 사슬 A(1F7Y:A)를 대표 서열로 선택하여 학습 데이터에 포함하고 있으며, 이는 선택된 서열이 군집에서 가장 길기 때문이다.

[0010] 대표 서열 하나 만을 제외하고 다른 모든 서열들을 제거하는 방식은 작은 훈련 데이터군을 생성하기는 하지만, 대표 서열로 선택되지 못한 서열들의 부분적인 결합 정보들이 중복 데이터를 제거하는 과정에서 함께 제거되는 문제점이 있다.

[0011] 즉, 서열 데이터에 대해 적용되는 중복 데이터 제거 기법의 일반적인 시행에서, 훈련 데이터 생성에 서열 전체가 포함되거나 제외된다. 이것은 전부 아니면 전무 식의 접근법이므로, 원본 서열의 결합 정보 중 부분적으로 존재하는 유용한 결합 정보들이 훈련 데이터에 포함되지 않는 문제점이 있다.

[0012] 이와 같은 문제점을 해결하기 위해, 본 발명자들은 특징 벡터를 기반으로 하는 중복 데이터 제거 방법을 제안한다. 상기 특징 벡터를 기반으로 하는 접근법에서, 같은 결합 라벨을 지닌 동일한 특징 벡터들은 중복 데이터로서 고려되어 이러한 특징 벡터들 중 하나의 특징 벡터만이 학습 데이터에 포함되며, 각각 다른 결합 라벨을 지닌 동일한 특징 벡터들이 존재할 경우에는 결합 특징 벡터와 비결합 특징 벡터를 모두 훈련 데이터 세트에 포함한다. 이와 같이, 특징 벡터를 기반으로 하는 학습 데이터 생성 방법은 중복적인 데이터를 제거함과 동시에 초기 서열들에서 가능한 모든 결합 정보를 학습 데이터에 포함시키고 있어 예측 모델 생성에 있어 강력한 훈련 데이터를 제공하게 된다.

발명의 내용

해결하려는 과제

[0013] 상기한 바와 같은 문제를 해결하기 위하여, 본 발명에서는 종래에 일반적으로 사용되어지는 서열 유사도를 바탕

으로 한 학습 데이터 생성 방법 대신에 불필요한 중복 데이터를 포함하지 않으면서도 더욱 큰 규모의 데이터 세트를 생성할 수 있는 특징 벡터에 기반한 훈련 데이터 생성 방법을 제공하는 것을 목적으로 한다.

[0014] 본 발명의 다른 목적은, 특징 벡터에 단백질과 RNA 서열 모두의 특징을 포함함으로써 단백질의 특징만을 사용하는 종래의 방법보다 더욱 효과적으로 단백질 서열에 존재하는 RNA-결합 아미노산을 예측할 수 있는 특징 벡터에 기반한 훈련 데이터 생성 및 예측 모델을 제공하는 것이다.

과제의 해결 수단

[0015] 상기한 바와 같은 목적을 달성하기 위해 본 발명에 따른 특징 벡터에 기반한 훈련 데이터의 중복 제거 방법은 (1) RNA와 수소결합으로 상호작용하는 RNA-결합 아미노산을 결정하여 단백질-RNA 결합 부위를 결정하는 단계와, (2) 상기 RNA와 상호작용하는 아미노산 트리플릿의 경향성을 계산하는 단계와, (3) 단백질 서열에 있는 상기 RNA-결합 아미노산을 예측하기 위해 단백질 및 RNA 서열의 다양한 특징을 특징 벡터에 코드화하는 단계와, (4) 상기 코드화된 특징 벡터를 기반으로 데이터의 불필요한 중복을 제거한 훈련 데이터 세트를 구축하는 단계를 포함하는 것을 특징으로 한다.

[0016] 또한, 상기 (4) 단계 후, (5) 상기 단백질 서열에 있는 RNA-결합 아미노산의 예측 성과를 복수의 예측 성능 척도를 사용해 평가하는 단계를 더 포함할 수 있다.

[0017] 상기 (1) 단계에서, 단백질과 RNA의 수소원자 (H), 수소 공여자 (D, donor), 수소 수용자 (A, acceptor), acceptor antecedent (AA)에 해당하는 원자들이 다음 기하학적 조건을 만족해야 한다: D-A 거리 < 3.9 Å, H-A 거리 < 2.7 Å, D-H-A 각도 ≥ 90° , H-A-AA 각도 ≥ 90° .

[0018] 상기 (2) 단계에서, 상기 아미노산 트리플릿과 뉴클레오티드 사이에 존재하는 상호 작용 경향성 (IP_{tb})은 하기의 수학식에 의해 결정될 수 있다.

[0019]
$$IP_{tb} = \sum_{\substack{t = AA\ triplet \\ b = nucleotide}} \frac{1}{\overline{HA_{tb}}\cos(\angle DAH) \cdot N_{PR}} \Big/ \frac{N_t \cdot N_b}{N_P \cdot N_R}$$

[0020] 여기서, $\overline{HA_{tb}}\cos(\angle DAH)$ 는 아미노산 트리플릿 t와 결합하는 뉴클레오티드 b 사이의 D-A에 H-A를 투영한 거리의 역수를 전부 합한 값, N_{PR}은 뉴클레오티드에 결합하는 모든 아미노산 트리플릿의 총 개수, N_t는 아미노산 트리플릿 t의 개수, N_P는 아미노산 트리플릿의 총 개수, N_b는 뉴클레오티드 b의 개수, N_R은 데이터 세트에 있는 뉴클레오티드의 총 개수를 의미한다.

[0021] 상기 (3) 단계에서, 특징 벡터에 구성요소로서 코드화되는 단백질 및 RNA 서열의 특징은, 목표 아미노산의 전체 서열 정보를 표현하는 단백질 서열의 전역적인 특징과, 아미노산 개개의 특징을 대표하는 아미노산의 국지적인 특징 및 상기 단백질과 상호 작용하는 RNA 서열의 특징을 포함할 수 있다.

[0022] 전역적인 특징으로 해당 단백질 서열의 길이 및 각 아미노산의 빈도 수가 특징 벡터 코드화에 구성요소로 사용되며, 국지적 특징으로는 해당 서열에서의 목표 아미노산의 정규화된 위치 (Normalized position), 소수성 (Hydropathy), 접근 가능한 표면적 (Accessible surface area), 분자량 (molecular mass), 아미노산 곁사슬의 pKa 값, 계산된 아미노산 트리플릿의 IP값이 특징 벡터 코드화에 구성요소로 사용된다.

[0023] 또한, 상호 작용 상대방인 RNA 서열의 특징으로는 RNA 서열에서 각각의 뉴클레오티드의 정규화된 위치의 합을 코드화 구성요소로서 사용된다.

[0024] 상기 (4) 단계는 1) 단백질-RNA 상호 작용 쌍들에 존재하는 모든 단백질 서열에서 슬라이딩 윈도우 (sliding window) 기법을 이용한 서열 조각을 생성하는 공정과, 2) 생성된 서열 조각을 상기 구성요소들을 이용하여 특정 벡터에 코드화하는 공정 및 3) 코드화된 특정 벡터에서 특징 벡터 기반의 중복 제거 기법을 통한 훈련 데이터를 구축하는 공정을 포함할 수 있다. 이 단계에서, 같은 결합 라벨을 지닌 동일한 구성 요소 수치를 갖는 특징 벡터들은 중복 데이터로서 고려되어 이러한 특징 벡터들 중 하나의 특징 벡터만이 학습 데이터에 포함되며, 각각 다른 결합 라벨을 지닌 동일한 구성 요소 수치를 갖는 특징 벡터들이 존재할 경우 각각 하나의 (하나는 결합, 그리고 하나는 비결합) 특징 벡터만이 훈련 데이터 세트에 포함된다. 상기 라벨은 상기 서열 조각의 중간에 위

치한 아미노산이 결합 아미노산이라면 결합 (+)으로 취급되고, 그 외의 경우는 비결합 (-)으로 취급될 수 있다.

[0025] 상기 (5) 단계에서, 상기 RNA-결합 아미노산의 예측 성능을 평가하기 위해 사용되는 복수의 예측 성능 척도는 민감도 (Sensitivity), 특이도 (Specificity), 정밀도 (Precision rate), 정확도(Accuracy), Net prediction, F-measure 및 상관 계수 (Correlation coefficient)를 포함할 수 있다.

발명의 효과

[0026] 상기한 바와 같이 본 발명에 따른 특징 벡터에 기반한 훈련 데이터의 중복 제거 방법에 의하면, 종래의 일반적인 서열 유사도 기반의 방법보다 중복 데이터를 학습 데이터에 포함시키지 않으면서도 초기 서열의 가능한 모든 결합 정보를 포함하는 학습 데이터를 생성할 수 있고, 특징 벡터에 단백질과 RNA 서열 모두의 특징을 구성 요소로서 포함함으로써 단백질의 특징만을 사용하는 방법보다 더욱 효과적으로 단백질 서열에 존재하는 RNA-결합 아미노산을 예측할 수 있다.

도면의 간단한 설명

[0027] 도 1은 본 발명의 실시 예에 따른 단백질 서열에서의 RNA-결합 부위 예측 모델 생성 방법의 주요 단계를 나타내는 제 1 블럭도.

도 2는 본 발명의 실시 예에 따른 단백질 서열에서의 RNA-결합 부위 예측 모델 생성 및 평가의 주요 단계를 나타내는 제 2 블럭도.

도 3은 본 발명의 실시 예에 따른 특징 벡터에 기반한 훈련 데이터 생성의 일 블럭도.

도 4는 서열 유사도를 기반으로 하는 중복 데이터 제거 과정 도중 일어나는 결합 정보의 손실을 나타내는 도.

도 5는 9개의 아미노산을 포함하는 서열 조각의 코드화된 특징 벡터의 구조를 나타내는 도.

도 6은 특징 벡터 기반의 중복 데이터 제거 방법의 예시를 나타내는 도.

도 7은 중복 데이터를 제거하는 과정에 있어서 서열 유사도 기반의 방법과 특징 벡터 기반의 방법의 비교를 나타내는 도.

도 8은 단백질-RNA 복합체 3OVB의 단백질 사슬 B와 RNA 사슬 D 사이에서의 실제 RNA-결합 부위와 본 발명의 방법으로 생성된 SVM 예측 모델을 사용하여 예측된 결과를 비교한 도.

발명을 실시하기 위한 구체적인 내용

[0028] 이하, 첨부된 도면을 참조하여 본 발명의 실시 예를 상세히 설명한다. 우선, 도면들 중 동일한 구성요소 또는 부품들은 가능한 한 동일한 참조부호를 나타내고 있음에 유의해야 한다. 본 발명을 설명함에 있어서 관련된 공지기능 혹은 구성에 대한 구체적인 설명은 본 발명의 요지를 모호하게 하지 않기 위해 생략한다.

[0029] 도 1은 본 발명의 실시 예에 따른 단백질 서열에서의 RNA-결합 부위 예측 모델 생성 방법의 주요 단계를 나타내는 제 1 블럭도이다.

[0030] 본 발명의 실시 예에 따른 단백질 서열에서의 RNA-결합 부위 예측 모델 생성 방법은 도 1에 도시된 바와 같이, 단백질-RNA 결합 부위 결정 단계 (S10), 아미노산 트리플릿 경향성 계산 단계 (S20), 특징 벡터 코드화 단계 (S30) 및 훈련 데이터 세트 구축 단계 (S40)를 포함한다.

[0031] 상기 단백질-RNA 결합 부위 결정 단계 (S10)는 RNA와 수소결합으로 상호작용하는 RNA-결합 아미노산을 결정하여 단백질-RNA 결합 부위를 결정하는 단계이다.

[0032] 종래의 단백질-RNA 결합 부위에 대한 연구들, 예를 들어 BindN과 RNABindR에서는 단백질의 아미노산의 원자가 상호작용 상대방 RNA에 존재하는 뉴클레오티드의 원자들 중 하나와 5 Å 이내의 거리에 떨어져 있는 경우 해당 아미노산을 RNA-결합 아미노산으로 고려되었다.

[0033] 이에 반해, 본 발명에서는 이 보다 더욱 엄격한 기준이 사용되는데, 본 발명의 단백질-RNA 결합 부위 결정 단계(S10)에서는, 상기 RNA-결합 아미노산이 RNA와의 수소결합 상호작용에 대한 다음과 같은 기하학적 기준을 만족하여야 한다.

[0034] 구체적으로, 상기 기하학적 기준으로 단백질과 RNA의 수소원자 (H), 수소 공여자 (D, donor), 수소 수용자 (A, acceptor), acceptor antecedent (AA)에 해당하는 원자들이 D-A 거리 < 3.9 Å, H-A 거리 < 2.7 Å, D-H-A 각도 ≥ 90° , H-A-AA 각도 ≥ 90° 인 기하학적 조건을 만족해야 한다.

[0035] 상기 아미노산 트리플릿 경향성 계산 단계 (S20)는 상기 RNA와 상호작용하는 아미노산 트리플릿의 결합 성향을 계산하는 단계이다.

[0036] 일반적으로, 동일한 종류의 아미노산들은 각기 다른 이웃 아미노산들 또는 다른 이차 구조 속에서, 서로 다른 상호 작용 경향성을 가질 수 있는데, 연속으로 배열된 아미노산 세 개 (아미노산 트리플릿 또는 삼중 아미노산으로 불린다)의 상호 작용 경향성을 계산하여 그 경향성을 RNA-결합 아미노산을 예측하기 위해 사용할 수 있다.

[0037] 상기 아미노산 트리플릿 t와 뉴클레오티드 b 사이에 존재하는 상호 작용 결합 성향 IP_{tb}은 하기의 [수학식 1]에 의해 결정될 수 있다.

수학식 1

[0038]
$$IP_{tb} = \sum_{\substack{t = AA\ triplet \\ b = nucleotide}} \frac{1}{\overline{HA_{tb}}\cos(\angle DAH) \cdot N_{PR}} \Big/ \frac{N_t \cdot N_b}{N_P \cdot N_R}$$

[0039] 여기서, $\overline{HA_{tb}}\cos(\angle DAH)$ 는 아미노산 트리플릿 t와 결합하는 뉴클레오티드 b 사이의 D-A에 H-A를 투영한 거리의 역수를 전부 합한 값, N_{PR}은 뉴클레오티드에 결합하는 모든 아미노산 트리플릿의 총 개수, N_t는 아미노산 트리플릿 t의 개수, N_P는 아미노산 트리플릿의 총 개수, N_b는 뉴클레오티드 b의 개수, N_R은 데이터 세트에 있는 뉴클레오티드의 총 개수를 의미한다.

[0040] 상기 [수학식 1]에서 IP_{tb}는 개별적인 아미노산과 뉴클레오티드 사이의 수소 결합들을 사용하는 대신 트리플릿의 아미노산 중 하나와 뉴클레오티드 사이의 D-A-H 각도를 이용한 코사인 값을, H-A 거리와 곱한 값의 역수를 취한 것을 사용한다.

[0041] 여기서, 수소 결합에 존재하는 D-A에 투영된 H-A 거리의 값을 사용하는 목적은 H-A 거리를 고려함과 동시에 D-A와의 거리적 관계 또한 고려하기 위함이고, 역수를 사용하는 이유는 멀리 떨어진 공여자-수용자 쌍 보다는, 가까운 공여자-수용자 쌍의 수소 결합에 보다 높은 IP 값을 배치하기 위함이다. 총 8,000개의 아미노산 트리플릿과 4개의 뉴클레오티드가 존재하기 때문에, 아미노산 트리플릿과 뉴클레오티드 사이에서 32,000 개의 IP들을 계

산할 수 있다.

[0042] 상기 특징 벡터 코드화 단계 (S30)는 단백질 서열에 있는 상기 RNA-결합 아미노산을 예측하기 위해 단백질 및 RNA 서열의 특징을 특징 벡터에 코드화하는 단계이다.

[0043] 단백질 서열에 있는 RNA-결합 아미노산들을 예측하기 위해, 특징 벡터 안에 단백질과 RNA 서열들의 몇몇 특징들을 표현할 수 있는데, 이 특징들은 세 개의 다른 특징 종류들로 분류될 수 있다.

[0044] 구체적으로, 상기 특징 벡터에 벡터구성요소로 코드화되는 상기 단백질 및 RNA 서열의 특징은, 목표 아미노산의 전체 서열 정보를 대표하는 단백질 서열의 전역적인 특징과, 아미노산 개개의 특징을 대표하는 아미노산의 국지적 특징 및 상기 단백질과 상호 작용하는 RNA 서열의 특징을 포함할 수 있다.

[0045] 여기서, 상기 단백질과 상호 작용하는 RNA 서열의 특징은 파트너 서열에 따른 단백질 서열에서의 다른 결합 부위들을 예측하기 위해 다른 특징들과 함께 사용될 수 있다.

[0046] 상기 특징 벡터에 구성요소로 코드화되는 단백질 서열의 전역적인 특징은 서열 길이 (L)와 아미노산 조성 (C)을 포함할 수 있다.

[0047] 여기서, 아미노산 조성은 단백질 서열에서 20 종류의 아미노산들의 빈도수를 나타내는데, 상기 단백질 서열의 일반적인 특징은 특징 벡터에서 총 21개의 구성 요소들을 요구하며, 이 때 하나는 서열 길이를 위해, 그리고 20개는 아미노산 조성을 위해 요구된다.

[0048] 상기 특징 벡터에 구성요소로 코드화되는 아미노산의 국지적 특징은 정규화된 위치 (N), 소수성 (H), 접근 가능한 표면적 (A), 분자량 (M), 아미노산 곁사슬의 pKa (P) 값과, 아미노산 트리플릿의 상호 작용 경향성 (IP)을 포함할 수 있다.

[0049] 여기서, IP는 IP_A, IP_C, IP_G 그리고 IP_U 라는 네 개의 구성요소들로 표시되며, 이 때 IP_A는 뉴클레오티드 아데닌 (A)에 대한 아미노산 트리플릿의 상호 작용 경향성을 나타낸다.

[0050] 한편, 서열 내부에서 아미노산의 정규화된 위치는 하기의 [수학식 2]에 의해 계산될 수 있다. 정규화된 위치를 제외하고, 동일한 아미노산 또는 아미노산 트리플릿은 국지적 특징에 대해서는 같은 값을 가진다.

수학식 2

[0051]
$$\text{Normalized Position}(i) = \frac{\text{Position}(i)}{\text{Sequence Length}}$$

[0052] 한편, 상기 특징 벡터에 벡터구성요소로 코드화되는 RNA 서열의 특징들은 RNA 서열에서 각각의 뉴클레오티드의 정규화된 위치의 합을 코드화한 것일 수 있다.

[0053] 구체적으로, 이러한 특징은 각 4개의 뉴클레오티드에 대해, 하기의 [수학식 3]에 의해 계산되어, 4개의 구성요소로 특징 벡터에 나타날 수 있다(R_A, R_C, R_G, R_U). 이러한 구성요소로 인해, 동일한 아미노산 서열들은 그들이 각자 다른 RNA 서열들과 상호 작용 하는 경우 다른 특징 벡터로 코드화될 수 있다.

수학식 3

$$R_{b\in\{A,C,G,U\}} = \sum_{i,b_i=b}^{\text{sequence length}} \text{Normalized Position}(b_i)$$

[0054]

[0055] 도 5는 9개의 아미노산을 포함하는 서열 조각의 코드화된 특징 벡터의 구조를 나타내는 도이다.

[0056] 각각의 벡터구성요소들은 특징 벡터에 나타나는 조건일 때, [0, 1] 범위의 값으로 정규화될 수 있다.

[0057] 구체적으로, 도 5에 도시된 바와 같이, 단백질의 일반적인 특징들(L에 대응하는 하나의 구성 요소와 C에 대응하는 20개의 구성 요소들)과 그 파트너인 RNA 서열의 특징(R에 대응하는 4개의 구성요소들)은 전체 단백질 서열 조각에 대해 한 번 나타나 있지만, 아미노산의 국지적인 특징들은 각각의 내부 아미노산들에서 나타나야 한다(N, H, A, M, P에 대응하는 5개의 구성요소들과 IP에 대응하는 4개의 구성요소들).

[0058] 여기서, IP 값은 창의 말단 아미노산들, 예를 들면 도 5의 a_{I-4}와 a_{I+4}에 대해서는 정의되지 않으므로, 말단 아미노산들에 대해서는 5개의 구성요소들만이 나타난다.

[0059] 한편, 서열을 코드화할 때에 겹쳐지는 트리플릿들을 사용하기 때문에, w개의 아미노산들로 구성된 서열 조각에 대해 w-2개의 트리플릿들이 대응한다.

[0060] w개의 아미노산들로 구성된 서열 조각들이 사용되는 때에, 아미노산 i에 대응하는 특징 벡터는 아미노산 트리플릿 T(i-(w -1)/2)로 시작하여 T(i-(w-1)/2-1), T(i-(w-3)/2-1), …, T (i+(w-3)/2-1)과 T(i+(w-1)/2-1)들을 포함한다.

[0061] 그러므로, w 아미노산들의 서열 조각은 9w + 17의 구성요소들의 특징 벡터로서 코드화되는데, 이러한 구성요소는 21 개의 일반적인 구성 요소들 (1 L과 20 C들), 4 RNA 구성 요소들 (R_A, R_C, R_G 그리고 R_U), w-2 내부 아미노산에 대응하는 9개의 국지적 구성 요소들 (N, H, A, M, P, 그리고 4 개의 IP들), 그리고 2개의 말단 아미노산에 대응하는 5개의 국지적 구성 요소들 (H, H, A, M, P)이다.

[0062] 또한, 특징 벡터는 서열 조각의 중간 아미노산이 결합 아미노산이라면 +1(결합)으로 표시되며, 그 외의 경우 -1(비결합)으로 표시된다.

[0063] 상기 훈련 데이터 세트 구축 단계 (S40)는 상기 코드화된 특징 벡터를 기반으로 데이터의 불필요한 중복을 감소시킨 훈련 데이터 세트를 구축하는 단계이다.

[0064] 도 3은 본 발명의 실시 예에 따른 특징 벡터에 기반한 훈련 데이터 생성에 대한 블럭도이다.

[0065] 상기 훈련 데이터 세트 구축 단계 (S40)는 도 3에 도시된 바와 같이, 서열 조각 생성 공정 (S41)과, 서열 조각 코드화 공정 (S42) 및 훈련 데이터 세트 구축 공정 (S43)을 포함할 수 있다.

[0066] 상기 서열 조각 생성 공정 (S41)은 단백질-RNA 상호 작용 쌍들에 존재하는 모든 단백질 서열에서 중복되는 서열 조각을 생성하는 공정이다.

[0067] 상기 서열 조각 코드화 공정 (S42)은 상기 서열 조각을 상기 특정 벡터에 코드화하는 공정이다.

[0068] 여기서, 같은 결합 라벨을 지닌 동일한 구성 요소 수치를 갖는 특징 벡터들은 중복 데이터로서 고려되어 이러한 특징 벡터들 중 하나의 특징 벡터만이 학습 데이터에 포함되며, 각각 다른 결합 라벨을 지닌 동일한 구성 요소 수치를 갖는 특징 벡터들이 존재할 경우 각각 하나의(하나는 결합, 그리고 하나는 비결합) 특징 벡터만이 훈련 데이터 세트에 포함된다. 상기 라벨은 상기 서열 조각의 중간에 위치한 아미노산이 결합 아미노산이라면 결합 (+)으로 취급되고, 그 외의 경우는 비결합 (-)으로 취급될 수 있다.

[0069] 상기 훈련 데이터 세트 구축 공정 (S43)은 상기 특정 벡터에서 중복 데이터를 제거한 후 상기 훈련 데이터 세트를 구축하는 공정이다.

[0070] 도 6은 특징 벡터 기반의 중복 데이터 제거 방법의 예시를 나타내는 도이다.

[0071] 도 6의 케이스 1에서, s1과 s3 서열 조각들은 같은 아미노산 서열을 가지며, s1과 s3의 중간 아미노산들은 둘 다 결합 부위들이다. 그러므로 s1과 s3 서열 조각들을 대표하는 특징 벡터 v1과 v3들은, 같은 벡터구성요소들과 라벨을 가진다.

[0072] 훈련 데이터 세트에 존재하는 불필요한 중복 데이터를 제거하기 위해, 단 하나의 서열 조각 (이 케이스에서는 s1이 된다)만이 남겨지고 s3은 제거된다.

[0073] 반대로 서열 조각 s2와 s4들은 같은 벡터구성요소들을 가지지만 다른 라벨들을 포함하므로, 그들의 특징 벡터 v2와 v4는 동일하지 않다. 따라서, s2와 s4 모두가 훈련 데이터 세트에 포함되어 있다.

[0074] 한편, 도 6의 케이스 2에서는, 단백질의 전역적인 특징인 서열 길이가 특징 벡터에 추가되어 있으므로, 서열 조각 s5와 s6을 대표하는 특징 벡터 v5와 v6는 더 이상 동일하지 않다.

[0075] 도 7은 중복 데이터를 제거하는 과정에 있어서 서열 유사도 기반의 방법 (Sequence-based method; 이하 S-방법)과 특징 벡터 기반의 방법 (Feature vector-based method; 이하 F-방법)의 비교를 나타내는 도이다.

[0076] 한편, 도 7에 도시된 바와 같이, 특징 벡터를 기반으로 하는 방법은 단백질 서열에 존재하는 모든 가능한 서열 조각들을 사용하여 중복 데이터가 없는 훈련 데이터 세트를 구축하지만, 종래의 서열 유사성을 기반으로 하는 방법은 몇몇 서열 조각들을 제외하여 특징 벡터를 기반으로 하는 방법보다 작은 훈련 데이터 세트를 구축한다.

[0077] 또한, 서열 유사도를 기반으로 하는 방법은 훈련 데이터 세트 내부에 중복된 데이터 (조각 2와 조각 4)를 남기지만, 특징 벡터를 기반으로 하는 방법은 특징 벡터들과 그들의 결합 정보 (라벨)들을 고려함으로써 중복된 데이터를 훈련 데이터 세트에 포함하지 않는다.

[0078] 도 2는 본 발명의 실시 예에 따른 단백질 서열에서의 RNA-결합 부위 예측 모델 생성 및 평가의 주요 단계를 나타내는 제 2 블럭도이다.

[0079] 도 2에 도시된 바와 같이, 상기 훈련 데이터 세트 구축 단계 (S40) 후, RNA-결합 아미노산의 예측 성과 평가 단계 (S50)를 더 포함할 수 있다.

[0080] 상기 RNA-결합 아미노산의 예측 성과 평가 단계 (S50)는 상기 단백질 서열에 있는 RNA-결합 아미노산의 예측 성과를 복수의 척도를 사용해 평가하는 단계이다.

[0081] 상기 RNA-결합 아미노산의 예측 성과를 평가하기 위해 사용되는 복수의 척도는 하기의 수학식들에 의해 계산되는 민감도 (Sensitivity), 특이도 (Specificity), 정밀도 (Precision rate), 정확도(Accuracy), Net prediction, F-measure 및 상관 계수 (Correlation coefficient)를 포함할 수 있다.

수학식 4

[0082] $$\text{Sensitivity (Sn)} = \frac{TP}{TP+FN}$$

수학식 5

[0083] $$\text{Specificity (Sp)} = \frac{TN}{TN+FP}$$

수학식 6

[0084] $$\text{Precision rate (Pr)} = \frac{TP}{TP+FP}$$

수학식 7

[0085] $$\text{Accuracy (Ac)} = \frac{TP+TN}{TP+FP+TN+FN}$$

수학식 8

[0086] $$\text{NetPrediction(NP)} = \frac{\text{Sn}+\text{Sp}}{2}$$

수학식 9

[0087] $$\text{F-measure (Fm)} = \frac{2\times\text{Sn}\times\text{Sp}}{\text{Sn}+\text{Sp}}$$

수학식 10

[0088]
$$\text{Correlation Coefficient (CC)} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}}$$

[0089] 상기 수학식에서, 참 양성 (TP)은 결합 아미노산으로 예측된 결합 아미노산이며, 참 음성 (TN)은 비결합 아미노산으로 예측된 비결합 아미노산이고, 거짓 양성 (FP)는 결합 아미노산으로 예측된 비결합 아미노산이며, 거짓 음성 (FN)는 비결합 아미노산으로 예측된 결합 아미노산이다.

[0090] 여기서, 상기 민감도는 실제 RNA-결합 아미노산이면서 RNA-결합으로 올바르게 예측된 아미노산들의 퍼센트이고, 특이도는 실제 RNA-비결합 아미노산이면서 RNA-비결합으로 올바르게 예측된 아미노산들의 퍼센트이다.

[0091] 또한, 정확도는 올바르게 결합 여부가 예측된 아미노산들의 퍼센트이나, 상기 정확도는 높은 불균형을 띠는 데이터 세트들에서는 오해의 여지가 있다. 예를 들어 10개의 긍정과 90개의 부정 샘플들을 지니는 데이터 세트들에서, 정확도는 모든 샘플들이 부정으로 분류될 경우에는 최대 90%까지 올라갈 수 있다.

[0092] 더불어, Net prediction은 민감도과 특이도의 평균이고, 상관 계수는 각기 다른 방법들의 성과들을 모두 비교하기에 가장 좋은 단일 척도이다.

[0093] 이하, 본 발명에 따른 특징 벡터에 기반한 훈련 데이터 생성 방법의 구체적인 실험 예를 설명한다.

[0094] 단백질-RNA 상호작용의 데이터 세트와 관련하여, 본 발명의 발명자는 세 개의 다른 단백질-RNA 상호 작용 데이터 세트들, 구체적으로 PRI3149, PRI727 및 PRI267을 구축했다.

[0095] 상기 PRI3149 데이터 세트에서, 단백질-RNA 복합체들은 단백질 데이터 은행 (Protein Data Bank)에서 획득되었는데, 2009년 11월을 기준으로 했을 때, 3.0 Å 또는 그 이상의 해상도를 사용한 X-선 결정학 (X-ray crystallography)에 의해 442개의 단백질-RNA 복합체 구조데이터들이 결정되어 있었고, 수소 결합에 대한 기하학적 기준들을 442개의 단백질-RNA 복합체들에 적용한 결과, 3,149개의 상호 작용하는 단백질-RNA 서열 쌍들을 생성하는 429개의 단백질-RNA 복합체들이 그 기준을 충족한 채로 남겨졌다.

[0096] 단백질 p가 두 개의 다른 RNA인 r1과 r2와 상호 작용했다면, p-r1과 p-r2 모두 데이터 세트에 포함되었다. 3,149 개의 단백질-RNA 상호 작용 서열 쌍들은 2,663개의 단백질 서열들과 812개의 RNA 서열들로 형성되었다. PRI3149 데이터 세트로부터, 본 발명자는 특징 벡터에 기반한 훈련 데이터 생성 기법을 적용하여 훈련 데이터를 구축했다.

[0097] 또한, PRI727과 PRI267 데이터 세트들은 PRI3149 데이터 세트로부터 독립적으로, 오직 RNA-결합 아미노산 예측 성능을 시험해보기 위한 용도로만 구축되었다. 본 발명자들은 2009년 11월 이후로 PDB에 저장된, 총 107개의 단백질-RNA 복합체들을 획득했다. 수소 결합에 대한 기하학적 기준들을 107개의 단백질-RNA 복합체들에 적용한 뒤, 592개의 단백질 서열들과 244개의 RNA 서열들을 포함하는, 727개의 서열 쌍을 PRI727 데이터 세트로 구축하였다.

[0098] 보다 더 엄격한 평가를 위해, PRI727 데이터 세트에 있는 단백질과 RNA 서열들 중 PRI3149에 있는 서열들과 60% 이상의 서열 일치도를 보이는 모든 쌍들은 제거되었다. 그 결과로, 192개의 단백질 서열들과 211개의 RNA 서열들을 포함하는 267쌍의 상호 작용하는 단백질-RNA 서열들이 PRI267 데이터 세트를 형성하기 위해 남겨졌다.

[0099] 한편, 특징 벡터를 기반으로 하는 데이터의 불필요한 중복 감소와 관련하여, 3,149쌍의 상호 작용하는 단백질-RNA를 포함하는 PRI3149 데이터 세트들은 초기에 59,398개의 RNA-결합 아미노산과 542,627개의 비결합-아미노산을 포함한다. 만약 불필요한 중복을 포함하는 데이터가 제거되지 않는다면, 양성 서열 조각들의 개수는 결합 아미노산의 개수와 같을 것이고, 음성 서열 조각들의 개수는 비결합 아미노산의 개수와 같을 것이다.

[0100] 본 발명자들은 3,149개의 서열 쌍들을, 그들의 특징들의 두 가지 다른 조합들 (모든 단백질 특징들과 RNA 특징들과 단백질의 국지적 특징)을 사용하여 나타냈으며, 본 발명에 따른 특징 벡터 기반의 중복 데이터 제거 방법을 적용했다.

[0101] 하기의 [표 1]은 PRI3149 데이터 세트에 특징 벡터를 기반으로 하는 데이터의 불필요한 중복 감소 방법을 적용한 뒤에 남아있는 특징 벡터들의 개수를 나타낸다.

표 1

[0102]

Window size	#Positive feature vectors	#Negative feature vectors	#Total vectors	#Common vectors
with 9 features (L, C, N, H, A, M, pK_a, IP, and R)				
1	21,282	198,578	219,860	2,811
3	21,282	198,585	219,867	2,811
5	21,283	198,590	219,873	2,811
7	21,283	198,596	219,879	2,811
9	21,284	198,601	219,885	2,811
11	21,284	198,606	219,890	2,811
13	21,284	198,611	219,895	2,811
15	21,284	198,616	219,900	2,811
with 6 features (N, H, A, M, pK_a, and IP)				
1	6,286	74,829	81,115	3,641
3	6,618	81,390	88,008	3,164
5	6,658	81,729	88,387	3,164
7	6,681	81,891	88,572	3,168
9	6,702	82,010	88,712	3,170
11	6,710	82,129	88,839	3,168
13	6,720	82,242	88,962	3,169
15	6,733	82,349	89,082	3,173

[0103] 구체적으로, [표 1]의 공통 벡터 (Common vector)들은 동일한 벡터 구성요소들을 포함하지만 다른 결합 라벨들 (결합은 +1, 비결합은 -1)을 지니는 특징 벡터들을 나타내는데, 적은 수의 공통 특징 벡터들 보다, 많은 수의 공통 특징 벡터들이 있는 데이터 세트에서 서로 다른 종류의 분류군들로 구분하는 것이 더욱 어렵다.

[0104] [표 1]에서 나타난 것처럼, 모든 특징들 (단백질 서열 길이, 아미노산 조성, 정규화된 위치, 소수성, 접근 가능한 표면적, 분자량, 그리고 아미노산 곁사슬의 pKa, 아미노산 트리플릿의 IP를 사용하는 것이 단백질의 6가지 국지적 특징들을 사용하는 것보다 모든 창 크기에서 더욱 많은 특징 벡터들을 생산했으며 보다 적은 양의 공통 특징 벡터들을 생산했다.

[0105] 서열 조각들의 6가지 국지적 특징들로 코드화하였을 때, 보다 큰 크기의 창을 사용한 특징 벡터를 기반으로 하는 중복 데이터 제거 방법은 보다 큰 데이터 세트를 구축했다. 하지만 9가지 특징들로 코드화하였을 때, 특징 벡터를 기반으로 하는 중복 데이터 제거 방법은 창 크기에 상관 없이 비슷한 크기의 불필요한 중복 데이터가 없는 데이터 세트들을 구축했다.

[0106] 하기의 [표 2]는 PRI727과 PRI267 데이터 세트들을 이용하여, 특징 벡터를 기반으로 하는 중복 데이터 제거 방법을 서열 유사성을 기반으로 하는 중복 데이터 제거 방법과 그 예측 성능을 비교하고 있다.

표 2

Dataset Construction	Sensitivity (%)	Specificity (%)	Accuracy (%)	NP (%)	Fm (%)	CC
PRI727 dataset						
S-method (100%)	84.1	75.8	76.3	80.0	79.7	0.32
S-method (80%)	84.9	74.3	74.9	79.6	79.2	0.31
S-method (60%)	85.4	72.7	73.5	79.1	78.6	0.30
F-method	87.2	81.7	82.1	84.5	84.4	0.40
PRI267 dataset						
S-method (100%)	46.4	86.8	85.9	66.6	60.5	0.14
S-method (80%)	48.4	85.7	84.9	67.2	62.2	0.14
S-method (60%)	49.6	84.5	83.8	67.0	62.5	0.13
F-method	60.7	91.0	90.3	75.8	72.8	0.24

[0107]

[0108] S-방법은 CD-HIT 프로그램을 사용한 서열 유사성을 기반으로 하는 중복 데이터 제거 방법 (Sequence based method)으로 구축된 학습 데이터의 예측 모델을 의미한다. 여기서, 괄호 안의 숫자는 CD-HIT에서 사용되는 군집의 서열 유사도 한계점을 나타내며, 각 한계점에 따라 생성된 군집의 가장 긴 서열들이 대표 서열로서 훈련 데이터 세트에 포함되었다.

[0109] F-방법은 특징 벡터를 기반으로 하는 중복 데이터 제거 방법 (Feature vector-based method)으로 구축된 학습 데이터의 예측 모델을 의미한다. S-방법과 F-방법의 단 한 가지 차이점은 그들의 훈련 데이터 세트들의 구축 방법이다. SVM 예측 모델은 9개의 특징들과 15 크기의 서열 조각들을 사용하여 훈련되어졌고 시험되었다. PRI727과 PRI267 데이터 세트들 모두에서, F-방법이 S-방법보다 모든 척도들에서 나은 결과를 보였다.

[0110] 한편, RNA에 대한 아미노산 트리플릿의 상호 작용 경향성과 관련하여, 본 발명자들은 [수학식 1]에 의해, 8,000개의 아미노산 트리플릿들과 4개의 뉴클레오티드들 사이의 32,000개의 상호 작용 경향성을 PRI3149 데이터 세트를 사용하여 계산했다.

[0111] 총 32,000 IP들 중 18,301개의 IP(57.2%)들이 0이 아닌 값을 가졌으며, 그 범위는 0.0에서 4.789796이었다. (SHK, U) 쌍이 가장 높은 IP 값 (4.79)을 가졌으며 CRR은 모든 뉴클레오티드들에서 평균적으로 높은 IP 값 (2.46)을 가졌다. 반대로, 2,238 IP들이 모든 뉴클레오티드들에서 0인 값을 가졌다.

[0112] 아미노산 트리플릿들의 IP에 이어, 본 발명자들은 [수학식 3]을 이용하여 PRI3149 데이터 세트에 있는 RNA 서열들에서 네 개의 RNA 특징 구성 요소들(R_A, R_C, R_G, R_U)들을 계산했다. PRI3149 데이터 세트는 812개의 RNA 서열들을 포함하며, 이 중 단 312개의 서열들만이 서열 정보만을 이용하였을 때 각각 구분 가능하다. 본 발명자들이 312개의 서열들에 대해 4개의 RNA 특징들을 나타냈을 때, 그들은 고유한 구성 요소들로 이루어진 특징 벡터들로 코드화됨을 확인하였다.

[0113] 단백질 서열에서의 RNA-결합 아미노산 예측에 있어 아미노산의 RNA와의 상호 작용 경향성에 대한 몇몇 정의들이 예측 성능에 끼치는 영향들을 검사하기 위해, 본 발명자들은 3개의 다른 IP 정의들을 사용하여 중복 데이터가 제거된 훈련 데이터 세트를 코드화했다.

[0114] 즉, 단일 아미노산들의 상호 작용 경향성 sIP, 기존 연구에서 사용된 아미노산 트리플릿들의 상호 작용 경향성 prev_tIP 및 본 발명에서 사용된 아미노산 트리플릿들의 상호 작용 경향성 tIP이 그것이다.

[0115] 하기의 [표 3]에서 보여진 결과들은 15 크기의 서열 조각들을 사용한 5-배수 교차 타당도 검증 (5-fold cross validation)을 통해 획득되었다.

표 3

Features	Sn (%)	Sp (%)	Pr (%)	Ac (%)	NP (%)	Fm (%)	CC
sIP	41.3	70.9	13.2	68.1	56.1	52.2	0.08
prev_tIP	57.0	83.5	27.0	80.9	70.3	67.8	0.29
tIP	57.4	83.8	27.5	81.3	70.6	68.2	0.30
sIP + R	57.3	83.1	26.7	80.6	70.2	67.8	0.29
prev_tIP + R	61.5	85.1	30.7	82.8	73.3	71.4	0.34
tIP + R	61.4	85.2	30.8	82.9	73.3	71.4	0.35

[0116]

[0117] 아미노산 트리플릿들의 IP(즉, prev_tIP와 tIP) 정의가 사용된 SVM 예측 모델들이 단일 아미노산들의 IP (sIP) 정의가 사용된 것들보다 나은 예측 성능을 보이고 있다. 단일 특징으로, 아미노산 트리플릿들의 새로운 IP (tIP) 정의가 가장 좋은 성과를 보였고, IP가 RNA 서열 특징들 (R_A, R_C, R_G, R_U)과 함께 사용되었을 때, 예측 성능은 IP를 단독으로 사용한 예측과 비교했을 때 더욱 향상되었다.

[0118] 본 발명자들은 데이터를 위한 특징 벡터를 코드화하고, 몇 가지 다른 조건에서 SVM 예측 모델을 생성하며, PDB 파일들로부터 얻은 다양한 기하학적 기준들을 사용하여 수소 결합들을 분류하기 위한 프로그램들을 개발하였다.

[0119] HBPLUS 라는 이름의 프로그램은 PDB 파일들로부터 수소 결합 부위들을 분류하는 용도로 널리 사용되고 있지만, 이 프로그램은 RNA에 있는 OP1이나 O2'같은 원자 이름들을 가지고 있는 새로운 PDB 파일들은 적용시킬 수 없다.

[0120] 본 발명자에 의해 개발된 수소 결합들을 분류하는 프로그램에서는 RNA의 O2', O3'원자들은 수소 결합의 잠재적인 공여자로서 포함되었다. 같은 맥락에서, RNA의 OP1, OP2, O2', O3', O4', 및 O5'원자들은 수소 원자의 잠재적인 수용자로서 포함되었다.

[0121] RNA-결합 아미노산들을 예측하는 것에 창 크기가 끼치는 영향을 평가하기 위해, PRI3149 데이터 세트에 대해 다양한 창 크기들을 사용하여, 특징 벡터를 기반으로 하는 데이터의 불필요한 중복 감소 방법을 적용한 몇몇 데이터 세트들이 구축되었다.

[0122] 하기의 [표 4]는 PRI3149 데이터 세트로부터 구축된, 불필요한 중복이 없는 8개의 다른 데이터 세트들에 의해 훈련된 SVM 예측 모델에 대한 5-배수 교차 타당도 검증의 예측 성과를 보여주고 있다.

표 4

Window size	Sn (%)	Sp (%)	Pr (%)	Ac (%)	NP (%)	Fm (%)	CC
1	47.3	73.6	16.1	71.1	60.4	57.6	0.14
3	66.5	84.8	30.9	82.7	74.1	72.6	0.36
5	64.9	84.5	30.9	82.6	74.7	73.4	0.36
7	65.6	84.9	31.7	83.0	75.2	74.0	0.37
9	66.8	84.9	32.2	83.2	75.9	74.8	0.38
11	66.8	85.4	32.8	83.6	76.1	74.9	0.39
13	67.1	85.8	33.7	84.0	76.5	75.3	0.40
15	68.4	85.9	34.1	84.2	77.1	76.1	0.41

[0123]

[0124] [표 4]에서 나타난 모든 결과들은 다음의 SVM 매개 변수 값을 사용해 획득되었다.

[0125] 즉, 매개 변수 값은 C = 10, r = 1/특징 벡터의 구성 요소들의 개수, w+ (양성 군의 가중치) = 음성 특징 벡터의 수 / 양성 특징 벡터의 수, 그리고 w- (음성 군의 가중치) = 1이다.

[0126] 모든 성과 척도에서 서열 조각 크기가 최대 15까지 상승함에 따라 예측 성과는 향상되었다. 그러므로 5-배수 교

차 타당도 검증의 최상의 예측 성과(84.2% 정확도, 76.1% f-척도, 그리고 0.41 상관계수)는 조각 크기 15를 사용한 SVM 예측 모델을 생성하였다.

[0127] 한편, 본 발명자들은 단백질 서열에 존재하는 RNA-결합 아미노산들을 예측하는 데에 있어서 사용될 수 있는 기존에 제안된 방법들과 본 발명의 방법을 비교했다.

[0128] 먼저, BindN은 아미노산의 화학적 특성들을 기반으로 SVM 기법을 사용하고, RNABindR은 Naive Bayes 분류기를 이용해 단백질 서열에 존재하는 RNA-결합 아미노산을 예측한다.

[0129] 이 방법들은 특징 벡터를 코드화할 때 각자 다른 특징들을 사용하고, 예측 모델을 구축할 때에도 각자 다른 기계 학습 알고리즘들을 사용하지만, 이들 모두가 훈련 데이터 세트 생성에 관해서는 서열 유사성을 기반으로 한 중복 데이터 제거 방법을 사용하였고 RNA-결합 아미노산을 예측하는 데에 있어서 상호 작용 파트너인 RNA 서열들의 정보들을 고려하지 않았다.

[0130] 이러한 방법들을 객관적으로 비교하기 위해, 본 발명자들은 이들을 각각 PRI727 데이터 세트와 PRI267 데이터 세트에 시험했다. PRI727과 PRI267 데이터 세트들 모두는, 본 발명에서 SVM 예측 모델을 훈련시키기 위해 사용했던 PRI3149 데이터 세트와는 다르게 구축되었다.

[0131] 하기의 [표 5]는 기존 방법들의 다양한 옵션들을 사용한 예측 성능들을 보여준다.

표 5

[0132]

Approach	Sensitivity (%)	Specificity (%)	Accuracy (%)	NP (%)	Fm (%)	CC
PRI727 dataset						
RNABindR (opt)	51.5	90.4	88.0	70.9	65.6	0.31
RNABindR (sp)	29.6	96.6	92.4	63.1	45.3	0.29
RNABindR (sn)	90.2	45.9	48.6	68.0	60.8	0.18
BindN (sn80)	88.4	51.0	53.4	69.7	64.7	0.19
BindN (sp80)	67.2	77.2	76.5	72.2	71.8	0.25
Our method 1	87.2	81.7	82.1	84.5	84.4	0.40
Our method 2	82.1	76.4	76.8	79.2	79.1	0.32
PRI267 dataset						
RNABindR (opt)	4.0	98.4	96.4	51.2	7.6	0.03
RNABindR (sp)	0.8	99.8	97.8	50.3	1.7	0.02
RNABindR (sn)	65.9	57.9	58.1	61.9	61.7	0.07
BindN (sn80)	80.2	53.5	54.0	66.8	64.2	0.10
BindN (sp80)	45.6	80.2	79.5	62.9	58.1	0.09
Our method 1	60.7	91.0	90.3	75.8	72.8	0.24
Our method 2	48.0	85.6	84.8	66.8	61.5	0.13

[0133] 여기서, RNABindR은 세 가지 옵션, 즉, 높은 민감도 (sn), 높은 특이도 (sp) 및 최적성(opt)으로 구동되었고, BindN은 두 가지 옵션, 즉, 80%로 기대되는 민감도 (sn80)과 80%로 기대되는 특이도 (sp80)으로 구동되었다.

[0134] 예측 성능에 대해 각각의 특징을 사용했을 때 나타나는 효과들을 검사하기 위해, 본 발명자들은 각자 다른 특징들을 사용한 두 개의 SVM 예측 모델들을 생성했다.

[0135] [표 5]에 있는 Our method 1은 RNA 서열의 특징들 뿐만 아니라 단백질의 8 가지 특징들(2개의 전역적인 특징들과 6개의 국지적 특징들)을 사용한 SVM 예측 모델을 나타내고, 또한, Our method 2는 8개의 단백질 특징들만을 사용한 SVM 예측 모델을 나타낸다.

[0136] Our method 1과 Our method 2 모두에 대해, 특징 벡터를 기반으로 하는 방법이 사용되어 PRI3149 데이터 세트로부터 중복 데이터가 제거된 훈련 데이터 세트들을 구축했다.

[0137] PRI727과 PRI267 데이터 세트에 대해 모두에서, RNA 특징들뿐만 아니라 단백질 특징들을 사용한 Our method 1이 민감도와 특이도 양쪽 모두에서 보다 높은 값을 보였지만, 다른 방법들은 높은 민감도 또는 특이도 중 하나만을 보였다.

[0138] 높은 민감도와 높은 특이도 뿐만 아니라, Our method 1은 Ner prediction, F-measure 및 상관 계수와 관련해서, 단백질 특징들만을 사용한 Our method 2을 포함하는 다른 방법들보다 높은 값들을 가졌다.

[0139] 단백질 특징들만을 사용한 Our method 2는 기존의 방법들과 비교했을 때 유사한, 또는 보다 나은 예측 성과들을 달성했다. 이 결과는 특징 벡터를 기반으로 하는 방법과 그 특징들이 RNA-결합 아미노산을 예측하는 데에 있어서 매우 정확한 예측 모델을 생성할 때 유용하다는 것을 보여준다.

[0140] 도 8은 단백질-RNA 복합체 3OVB의 단백질 사슬 B와 RNA 사슬 D 사이에서의 실제 RNA-결합 부위와 본 발명의 방법으로 생성된 SVM 예측 모델을 사용하여 예측된 결과를 비교한 도이다.

[0141] 단백질 서열에 존재하는 RNA-결합 아미노산을 예측하기 위한 대부분의 학습 접근법들은 서열 유사성을 기반으로 한 훈련 데이터 세트를 구축함을 통해 이루어진다. 서열 데이터에 존재하는 불필요한 중복 데이터를 제거하기 위한 과정에서, 서열 유사성을 기반으로 한 방법은 훈련 데이터 세트에 전체 서열을 더하거나 제외한다.

[0142] 유사 서열 또는 심지어 동일한 서열들조차, 그들의 상호 작용 파트너가 변화함에 따라 종종 매우 다른 상호 작용 부위를 가진다. 하지만 서열 유사성을 기반으로 중복 데이터 제거 방법을 사용하여 훈련 데이터를 구축했을 때에는, 많은 결합 정보들이 손실된다.

[0143] 상술한 바와 같이, 본 발명자들은 중복 데이터를 제거하기 위해 특징 벡터를 기반으로 하는 방법을 개발했다. 이러한 방법은 일반적으로 사용되는 서열 유사도를 기반으로 하는 감소 방법보다 더욱 큰 크기의 훈련 데이터 세트를, 중복성이 없는 데이터들로 구축할 수 있다.

[0144] 추가적으로, 특징 벡터를 기반으로 하는 방법에 의해 구축된 훈련 데이터 세트는 불필요한 중복 데이터를 포함하지 않지만, 서열 유사성을 기반으로 하는 방법에 의해 구축된 데이터 세트는 한 서열이 그 구조 내부에 유사한 부분서열들을 포함하는 경우 불필요한 중복 데이터를 생성할 가능성이 높다.

[0145] 단백질 서열에 존재하는 RNA-결합 아미노산을 예측하는 기존의 접근법들은 단백질의 상호 작용 파트너, 즉, RNA를 고려하지 않았으므로, 그러한 접근법들은 주어진 단백질 서열에 대해서는 항상 같은 RNA-결합 부위들을 예측했으며, 심지어는 그 단백질이 실제로는 다른 RNA 분자들에 결합할 때에도 그러한 예측을 했다.

[0146] 본 발명자들은 단백질과 RNA 서열들 모두 단백질 서열의 RNA-결합 아미노산을 예측하기 위해 특징 벡터 코드화의 특징들을 고려했다. 본 발명에 따른 SVM 예측 모델은 3,149 개의 단백질-RNA 상호작용 서열 쌍에 5-배수 교차 타당도 검증을 진행했을 때 84.2% 정확도, 76.1% F-measure, 그리고 0.41 상관 계수 값을 보였다.

[0147] 또한, 보다 엄격한 평가를 위해 기존의 예측 모델을 훈련과정에 사용되지 않았던 두 개의 새로운 데이터 세트들에 본 발명에 따른 SVM 예측 모델의 예측 성능을 시험했다. 새로운 데이터 세트들에서, 본 발명에 따른 SVM 예측 모델은 90.3% 정확도, 72.8% F-measure, 및 0.24 상관 계수 값을 보였고, 동일한 데이터 세트들에서 다른 방법들과 비교한 결과는 본 발명의 방법이 다른 방법들보다 뛰어나다는 것을 증명할 수 있었다.

[0148] 이상과 같이 본 발명에 따른 특징 벡터에 기반한 훈련 데이터의 중복 제거 방법을 예시한 도면을 참조로 하여 설명하였으나, 본 명세서에 개시된 실시 예와 도면에 의해 본 발명이 한정되는 것은 아니며, 본 발명의 기술사상 범위 내에서 당업자에 의해 다양한 변형이 이루어질 수 있음은 물론이다.

부호의 설명

[0149] S10:단백질-RNA 결합 부위 결정 단계

S20:아미노산 트리플릿 경향성 계산 단계

S30:특징 벡터 코드화 단계

S40:훈련 데이터 세트 구축 단계

S41:서열 조각 생성 공정

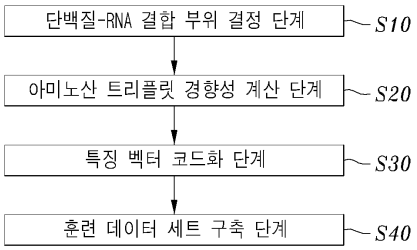
S42:서열 조각 코드화 공정

S43:훈련 데이터 세트 구축 공정

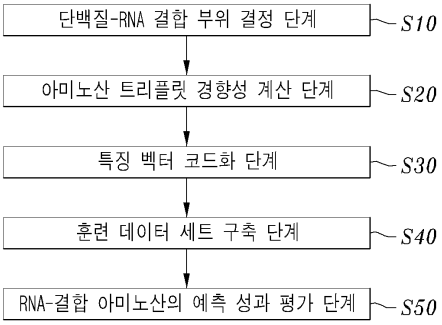
S50:RNA-결합 아미노산의 예측 성과 평가 단계

도면

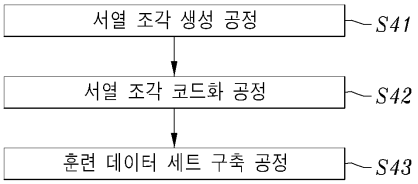
도면1



도면2



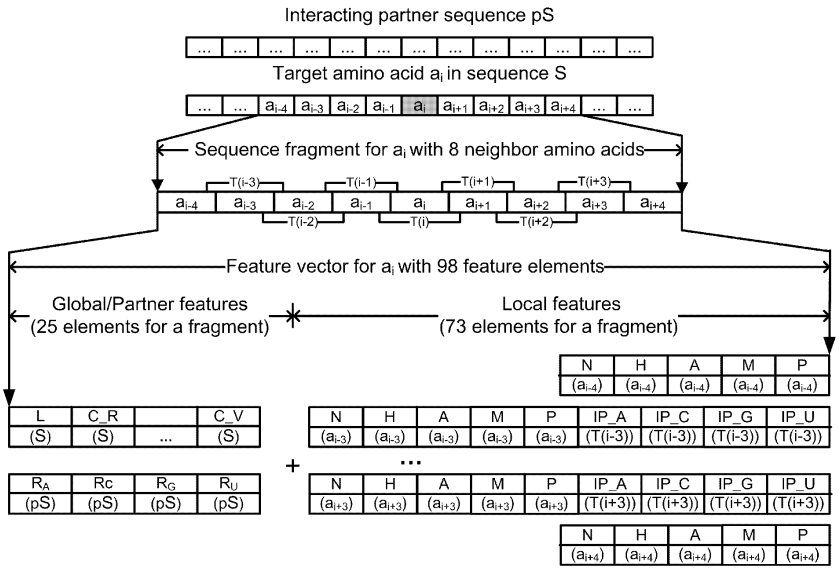
도면3



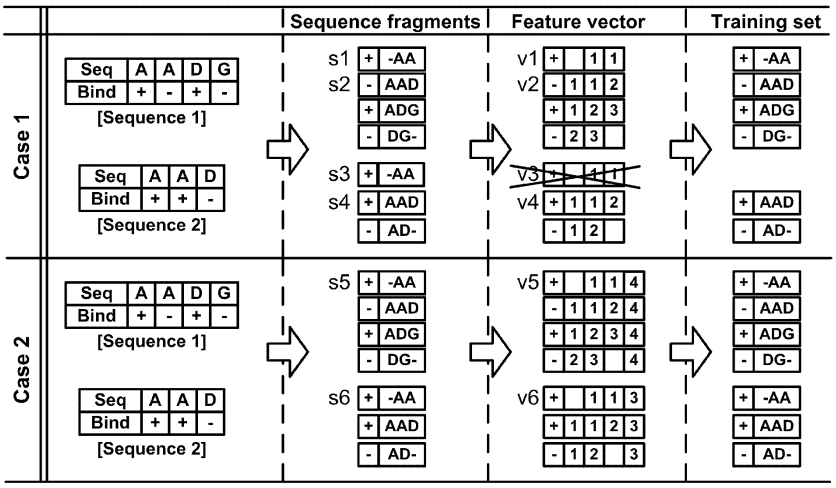
도면4

	PDBID	CHAIN	Identity of CDHIT	Identity of BLASTP	E-value of BLASTP
	1F7Y	A	-	-	-
SEQ	MPITKEEKQKVMQEFARFPGDTGSTEVQVALLTLRINRLSEHLKVHKKDHHSHRGLLMMVGQRRRLLRYLQREDPERYRMLIEKLGIRG				
BIND					
	1G1X	B	94%	100%	3.00E-48
SEQ	PITKEEKQKVIQEFARFPGDTGSTEVQVALLTLRINRLSEHLKVHKKDHHSHRGLLMMVGQRRRLLRYLQREDPERYREIVEKLGLRG				
BIND					
	1G1X	B	94%	100%	3.00E-48
SEQ	PITKEEKQKVIQEFARFPGETGSTEVQVALLTLRINRLSEHLKVHKKDHHSHRGLLMMVGQRRRLLRYLQREDPERYREIVEKLGLRG				
BIND					
	1G1X	G	94%	100%	3.00E-48
SEQ	PITKEEKQKVIQEFARFPGDTGSTEVQVALLTLRINRLSEHLKVHKKDHHSHRGLLMMVGQRRRLLRYLQREDPERYREIVEKLGLRG				
BIND					
	1G1X	G	94%	100%	3.00E-48
SEQ	PITKEEKQKVIQEFARFPGETGSTEVQVALLTLRINRLSEHLKVHKKDHHSHRGLLMMVGQRRRLLRYLQREDPERYREIVEKLGLRG				
BIND					
	1KUQ	A	96%	96%	4.00E-40
SEQ	MPICKEEKQKVIQEFARFPGETGSTEVQVALLTLRINRLSEHLKVHKKDHHSHRGLLMMVGQRRRLLRYLQREDPERYRALIEKLGI				
BIND					
	1N32	O	97%	97%	3.00E-41
SEQ	PITKEEKQKVIQEFARFPGDTGSTEVQVALLTLRINRLSEHLKVHKKDHHSHRGLLMMVGQRRRLLRYLQREDPERYRALIEKLGIRG				
BIND					

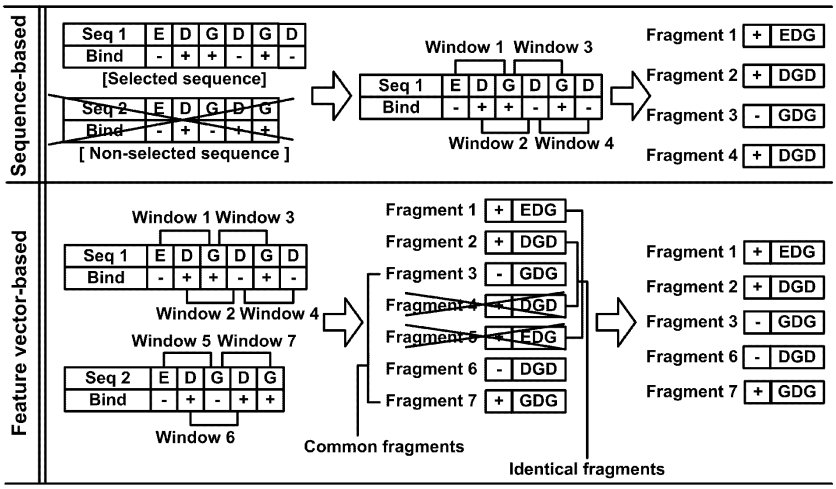
도면5



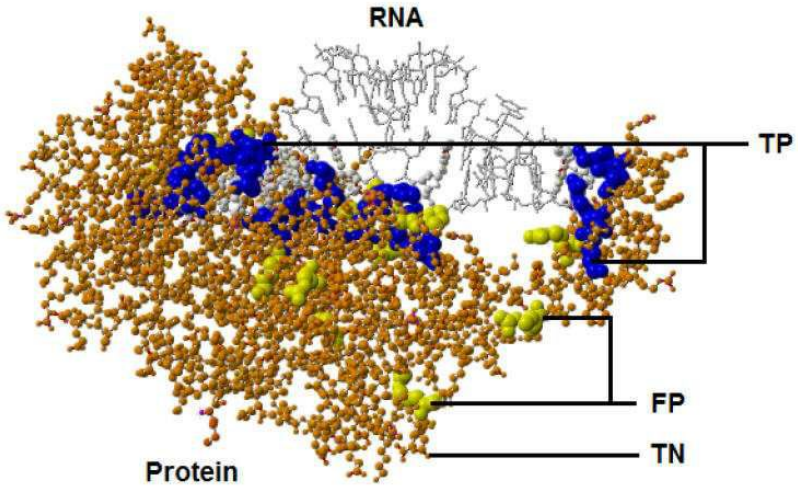
도면6


Sequence fragments
Feature vector
Training set
Case 1
Case 2
[Sequence 1]
[Sequence 2]
s1 s2 s3 s4 s5 s6
v1 v2 v3 v4 v5 v6
-AA
AAD
ADG
DG-
AD-

도면7


Sequence-based
Feature vector-based
[Selected sequence]
[Non-selected sequence]
Window 1 Window 3
Window 2 Window 4
Window 5 Window 7
Window 6
Fragment 1 + EDG
Fragment 2 + DGD
Fragment 3 - GDG
Fragment 4 - DGD
Fragment 5 + EDG
Fragment 6 - DGD
Fragment 7 + GDG
Common fragments
Identical fragments

도면8



RNA chain 3OVB:D GGAAGUAGAUGGUUCAAGUCCAUUUACUUCCACCA

Protein chain 3OVB:B
MKVEEILEKALELVIPDEEEVRKGREAEEELRRRLDELGVEYVFVGSYARNTWLK
GSLEIDVFLLFPEEFSKEELRERGLEIGKAVLDSYEIRYAEHPYVHGVVKGVEVD
VVPCYKLKEPKNIKSAVDRTPFHHKWLEGRIKGKENEVRLLKGFLKANGIYGAEY

Actual binding sites
----------+-----
-----+----------+++-+----------
----------+--+----------+

Predicted binding sites
----------+-----
-----+----------+++-+----------
----------++--++----------+