



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0019404
(43) 공개일자 2020년02월24일

(51) 국제특허분류(Int. Cl.)
G16B 40/00 (2019.01) G16B 30/00 (2019.01)
G16B 5/00 (2019.01)
(52) CPC특허분류
G16B 40/00 (2019.02)
G16B 30/00 (2019.02)
(21) 출원번호 10-2018-0094894
(22) 출원일자 2018년08월14일
심사청구일자 2018년08월14일

(71) 출원인
인하대학교 산학협력단
인천광역시 미추홀구 인하로 100(용현동, 인하대 학교)
(72) 발명자
한경숙
서울특별시 양천구 목동동로 258, 101동 1802호(목동, 금호1차아파트)
이옥
충청남도 서산시 남부순환로 876, 3동 605호 (죽성동, 동아아파트)
(74) 대리인
네이트특허법인

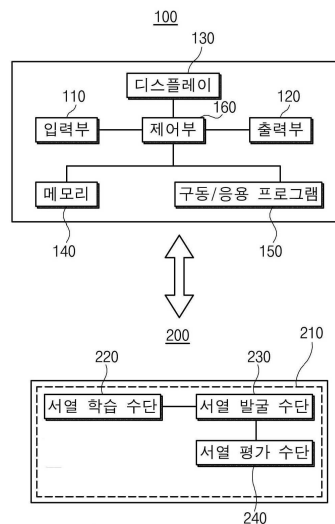
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 기계 학습 알고리즘을 이용하여 표적 단백질과 결합하는 후보 RNA 압타머를 생성하는 프로그램을 기록한 컴퓨터로 판독할 수 있는 기록 매체 및 후보 RNA 압타머를 생성하는 방법

(57) 요약

본 발명은 기계 학습 모델의 하나인 랜덤 포레스트 모델을 통하여 표적 단백질 분자와 결합하는 후보 RNA 압타머 서열을 생성, 검증하는 프로그램을 기록한 컴퓨터로 판독할 수 있는 기록 매체 및 이를 이용하여 표적 단백질 분자와 결합하는 후보 RNA 압타머 서열을 생성, 검증하는 방법에 관한 것이다. 본 발명에 따른 프로그램을 이용하여, 특정 표적 단백질과 결합하는 후보 압타머 서열의 pool이 감소되고, 신속하고 효율적으로 최종 압타머를 선별할 수 있다.

대표도 - 도1



(52) CPC특허분류

G16B 5/00 (2019.02)

이 발명을 지원한 국가연구개발사업

과제고유번호 2015R1A1A3A04001243

부처명 미래창조과학부

연구관리전문기관 한국연구재단

연구사업명 이공분야 기초연구사업 (여성과학자지원사업)

연구과제명 단백질과 핵산의 상호작용 예측 모델 개발

기 여 율 1/2

주관기관 인하대학교

연구기간 2015.05.01 ~ 2018.04.30

이 발명을 지원한 국가연구개발사업

과제고유번호 2017R1E1A1A03069921

부처명 미래창조과학부

연구관리전문기관 한국연구재단

연구사업명 이공분야 기초연구사업 (전략과제 산업수학)

연구과제명 바이오 빅데이터의 수학적 모델링을 통한 암 유전자 발굴과 개인별 유전자 네트워크 추론

기 여 율 1/2

주관기관 인하대학교

연구기간 2017.09.01 ~ 2020.08.31

명세서

청구범위

청구항 1

RNA-단백질 복합체 데이터에 기초하여 RNA 서열의 특징 벡터와 단백질을 구성하는 아미노산 서열의 특징 벡터를 구축하고, 구축된 특징 벡터를 토대로 랜덤 포레스트(random forest) 모델을 적용하여 RNA 서열과 단백질 서열을 훈련시키는 서열 학습 수단; 및

무작위 RNA 서열을 필터링(filtering)하고, 상기 서열 학습 수단을 통해 학습된 랜덤 포레스트 모델을 적용하여, 필터링 된 RNA 서열 중에서 표적 단백질 분자와 결합하는 후보 RNA 압타머를 구축하는 서열 발굴 수단을 포함하고,

상기 서열 학습 수단에서 RNA 특징 벡터는 단백질을 구성하는 아미노산과의 상호작용 경향(interaction propensity), RNA 서열의 모노-뉴클레오타이드(mono-nucleotide; 단일-염기) 조성, 디-뉴클레오타이드(di-nucleotide; 2-염기) 조성 및 유사 트리-뉴클레오타이드(pseudo tri-nucleotide; 유사 3-염기) 조성을 토대로 작성되고,

상기 서열 학습 수단에서 단백질 특징 벡터는 단백질을 구성하는 아미노산의 조성-전하-분포 및 유사 아미노산 조성을 토대로 작성되는 표적 단백질 분자와 결합하는 후보 RNA 압타머를 생성하는 프로그램을 기록한 컴퓨터로 판독할 수 있는 기록 매체.

청구항 2

제 1항에 있어서,

상기 프로그램은, 상기 서열 발굴 수단에 의해 구축된 후보 RNA 압타머를 생성하는 모델을 평가 척도를 사용하여 평가하는 서열 평가 수단을 더욱 포함하는 표적 단백질 분자와 결합하는 후보 RNA 압타머를 생성하는 프로그램을 기록한 컴퓨터로 판독할 수 있는 기록 매체.

청구항 3

제 2항에 있어서,

상기 평가 척도는 민감도(Sensitivity), 특이도(Specificity), 정확도(Accuracy), 양성예측도(Positive predictive value), 음성예측도(Negative predictive value) 및 매튜 상관계수(Matthews correlation coefficient) 중에서 선택되는 적어도 하나인 것을 특징으로 하는 표적 단백질 분자와 결합하는 후보 RNA 압타머를 생성하는 프로그램을 기록한 컴퓨터로 판독할 수 있는 기록 매체.

청구항 4

제 1항 내지 제 3항 중 어느 하나의 항에 있어서,

상기 서열 발굴 수단은, 상기 무작위 RNA 서열의 2차 구조, 자유 에너지, 쌍을 이루지 않는 뉴클레오타이드의 개수 및 풀(pool) 내에서 동일한 2차 구조의 개수를 한정하여 상기 무작위 RNA 서열을 필터링 하는 것을 특징으로 하는 표적 단백질 분자와 결합하는 후보 RNA 압타머를 생성하는 프로그램을 기록한 컴퓨터로 판독할 수 있는 기록 매체.

청구항 5

제 1항 내지 제 3항 중 어느 하나의 항에 있어서,

상기 RNA 특징 벡터를 구성하는 상기 유사 트리-뉴클레오타이드 조성은, 트리-뉴클레오타이드의 소수성(hydrophobicity), 친수성(hydrophilicity) 및 측쇄-중량(side-chain mass)을 포함하고,

상기 단백질 특징 벡터를 구성하는 상기 유사 아미노산 조성은, 아미노산의 소수성, 친수성, 측쇄-중량, 이온화 지수 및 등전점을 포함하는 표적 단백질 분자와 결합하는 후보 RNA 압타머를 생성하는 프로그램을 기록한 컴퓨터로 판독할 수 있는 기록 매체.

청구항 6

컴퓨터에 의해 실현되는 프로그램을 이용하여 표적 단백질 분자와 결합하는 후보 RNA 압타머를 생성하는 방법으로서,

서열 학습 수단에 의하여, RNA-단백질 복합체 데이터에 기초하여 RNA 서열의 특징 벡터와 단백질을 구성하는 아미노산 서열의 특징 벡터를 구축하고, 구축된 특징 벡터를 토대로 랜덤 포레스트(random forest) 모델을 적용하여 RNA 서열과 단백질 서열을 훈련시키는 단계; 및

서열 발굴 수단에 의하여, 무작위 RNA 서열을 필터링(filtering)하고, 상기 서열 학습 수단을 통해 학습된 랜덤 포레스트 모델을 적용하여, 필터링 된 RNA 서열 중에서 표적 단백질 분자와 결합하는 후보 RNA 압타머를 구축하는 단계를 포함하고,

상기 서열 학습 수단에서 RNA 특징 벡터는 단백질을 구성하는 아미노산과의 상호작용 경향(interaction propensity), RNA 서열의 모노-뉴클레오타이드(mono-nucleotide; 단일-염기) 조성, 디-뉴클레오타이드(di-nucleotide; 2-염기) 조성 및 유사 트리-뉴클레오타이드(pseudo tri-nucleotide; 유사 3-염기) 조성을 토대로 작성되고,

상기 서열 학습 수단에서 단백질 특징 벡터는 단백질을 구성하는 아미노산의 조성-전이-분포 및 유사 아미노산 조성을 토대로 작성되는 표적 단백질 분자와 결합하는 후보 RNA 압타머를 생성하는 방법.

청구항 7

제 6항에 있어서,

서열 평가 수단에 의하여, 상기 서열 발굴 수단에 의해 구축된 후보 RNA 압타머를 생성하는 모델을 평가 척도를 사용하여 평가하는 단계를 더욱 포함하는 표적 단백질 분자와 결합하는 후보 RNA 압타머를 생성하는 방법.

청구항 8

제 7항에 있어서,

상기 평가 척도는 민감도(Sensitivity), 특이도(Specificity), 정확도(Accuracy), 양성예측도(Positive predictive value), 음성예측도(Negative predictive value) 및 매튜 상관계수(Matthews correlation coefficient) 중에서 선택되는 적어도 하나인 것을 특징으로 하는 표적 단백질 분자와 결합하는 후보 RNA 압타머를 생성하는 방법.

청구항 9

제 6항 내지 제 8항 중 어느 하나의 항에 있어서,

상기 서열 발굴 수단은, 상기 무작위 RNA 서열의 2차 구조, 자유 에너지, 쌍을 이루지 않는 뉴클레오타이드의 개수 및 풀(pool) 내에서 동일한 2차 구조의 개수를 한정하여 상기 무작위 RNA 서열을 필터링 하는 것을 특징으로 하는 표적 단백질 분자와 결합하는 후보 RNA 압타머를 생성하는 방법.

청구항 10

제 6항 내지 제 8항 중 어느 하나의 항에 있어서,

상기 RNA 특징 벡터를 구성하는 상기 유사 트리-뉴클레오타이드 조성은, 트리-뉴클레오타이드의 소수성(hydrophobicity), 친수성(hydrophilicity) 및 측쇄-중량(side-chain mass)을 포함하고,

상기 단백질 특징 벡터를 구성하는 상기 유사 아미노산 조성은, 아미노산의 소수성, 친수성, 측쇄-중량, 이온화 지수 및 등전점을 포함하는 표적 단백질 분자와 결합하는 후보 RNA 압타머를 생성하는 방법.

발명의 설명

기술 분야

[0001] 본 특허출원은 대한민국 정부 과학기술정보통신부의 “개인기초연구” 연구사업의 일환으로서 “단백질과 핵산의 상호작용 예측 모델 개발” (주관기관: 인하대학교; 과제고유번호: 2015R1A1A3A04001243) 과제와, “바이오 빅데이터의 수학적 모델링을 통한 암 유전자 발굴과 개인별 유전자 네트워크 추론”(주관기관: 인하대학교; 과제고유번호: 2017R1E1A1A03069921) 과제의 수행 결과물에 관한 것이다.

[0002] 본 발명은 표적 물질과 결합하는 분자를 생성하는 프로그램을 기록한 매체에 관한 것으로, 더욱 상세하게는 기계 학습 알고리즘을 이용한 학습 과정을 통하여 표적 단백질과 결합하는 후보 RNA 압타머를 생성, 구축하기 위한 프로그램을 기록한 컴퓨터로 판독할 수 있는 기록 매체와, 이를 이용하여 표적 단백질과 결합하는 후보 RNA 압타머를 생성, 구축하는 방법에 관한 것이다.

배경 기술

[0003] 생물학적 메커니즘과 관련한 분자에 대한 이해가 깊어지면서, 생물학적 분자들 사이의 상호작용에 대한 관심이 높아지고 있다. 이러한 생물학적 분자들 중에서 압타머가 최근 주목을 받고 있다. 압타머(aptamer)는 자체로 안정적인 3차원 구조를 가지면서 표적 분자에 대해 높은 친화성(affinity)과 특이성(specificity)을 가지면서 결합할 수 있는 특징을 가지는 단일가닥 핵산(single stranded nucleic acid)이다. 표적 분자와 결합하는 압타머는 단일클론항체(monoclonal antibody)와 유사하지만, 단일클론항체에 비하여 다음과 같은 장점을 가지고 있다.

[0004] 항체는 대략 150 kDa의 큰 분자 구조를 가지고 있어서 제조 및 변형이 어렵지만, 통상적으로 100개 이하의 염기 서열로 이루어지는 압타머는 작은 분자 구조를 가지고 있기 때문에 변형이 용이하다. 압타머는 항체에 비하여 안정성이 매우 우수하기 때문에, 실온에서 운반 및/또는 보관이 가능하고 멸균 후에도 그 기능을 유지할 수 있다. 뿐만 아니라, 압타머는 변성(denaturation)이 일어나더라도 짧은 시간에 재생(regeneration)이 가능하기 때문에, 장시간 또는 반복 사용이 요구되는 진단용 소재로서 쉽게 응용될 수 있다.

[0005] 더욱이, 항체는 동물이나 세포를 이용하여 제조하기 때문에 많은 시간과 비용이 요구되며, 제조되는 시기나 방법에 따라 기능성이 달라질 가능성이 있다(batch to batch variation). 하지만, 압타머는 화학적 합성방법을 이용하여 제조되기 때문에 단시간에 적은 비용으로 제조될 수 있으며, batch to batch variation이 거의 없고, 고순도의 정제 과정이 매우 용이하다. 또한, 항체나 다른 의약품 단백질의 경우에는 빈번하게 생체내 면역거부반응이 일어나지만, 압타머는 이러한 생체내 면역거부반응이 거의 일어나지 않는 것으로 알려져 있어 치료용 소재의 개발에 있어 장점이 있다. 뿐만 아니라, 항체를 만들기 어려운 독소(toxin), 복잡한 단백질 복합체 또는 당-단백질 복합체에 대해서도 압타머를 제조할 수 있으며, 새로운 표적 물질에 대한 결합 물질로의 변형이 용이하여(flexibility) 신규한 압타머 발굴에 활발하게 이용될 수 있다.

[0006] 새로운 압타머를 선별하는 것과 관련해서, 1990년대 콜로라도 대학의 Larry Gold 연구팀에 의해 개발된 발굴 기술인 SELEX(Systematic Evolution of Ligands by EXponential enrichment)라는 기법이 기본적으로 활용되고 있다. SELEX를 통해 새로운 압타머를 발굴하는 과정을 살펴보면, 1) DNA 합성 및 시험관 전사(*in vitro* transcription, RNA인 경우)을 이용하여 다양한 형태의 핵산 라이브러리($>10^5$)를 제조하고, 2) 핵산 구조체 라이브러리 내의 핵산 구조체들(압타머 후보 분자들)에 대하여 원하는 표적 분자와 결합할 수 있는 핵산 구조체만을 선별(screening)하는 과정을 거친다. 선별 과정은 예를 들어, 친화 크로마토그래피(Affinity chromatography)와 같은 방법을 통해 표적 분자와 결합하지 않은 핵산 구조체를 제거(washing)하고 표적 분자에

결합하는 것만을 선택적으로 얻는다. 마지막으로 표적 분자로부터 핵산 구조체를 분리(elution)하고, 분리된 핵산을 PCR(중합효소연쇄반응, polymerase chain reaction) 등의 방법으로 증폭시킨다. 이어서, 증폭된 핵산 구조체만을 이용한 선별 및 분리 과정을 5 내지 15회 반복하여, 매우 우수한 결합력과 특이성을 보이는 압타머를 발굴할 수 있다.

[0007] 그런데, 압타머 후보 분자를 선별하기 위하여 통상적으로 10^{15} 이상의 무작위 핵산 라이브러리를 제조하고, 그 중에서 극히 일부의 핵산 서열만이 표적 분자에 결합하기 때문에, 최종적으로 표적 분자에 결합하는 압타머를 선별할 때 많은 시간, 노동 및 비용이 요구된다. 뿐만 아니라, SELEX에 의해 확인된 압타머는 생체내(*in vivo*)에서 다양한 항원 밀도, 상호작용 및 미세환경(microenvironment)로 인하여 생체내에서 표적 분자와 결합하지 못할 수 있다.

[0008] 압타머를 선별할 때의 문제점과 관련해서, 미국특허 제9,315,804호에서는 적합도 함수를 고려하여 후보 압타머로부터, 표적 물질에 대한 적어도 하나의 압타머를 식별하는 방법을 제안하고 있다. 하지만, 상기 미국특허에서도 여전히 초기 압타머 후보 물질의 풀(pool)이 대량으로 존재하기 때문에, 표적 물질과 강하게 결합하는 압타머를 선별할 때 적지 않은 시간과 노력이 투입되어야 한다.

발명의 내용

해결하려는 과제

[0009] 본 발명은 전술한 종래 기술의 문제점을 해소하기 위하여 제안된 것으로, 본 발명의 목적은 신속하고 효율적인 방법으로 표적 물질과 결합하는 잠재적인 후보 압타머를 생성, 구축하기 위한 프로그램을 기록한 컴퓨터로 판독할 수 있는 기록 매체 및 표적 물질과 결합하는 잠재적인 후보 압타머를 생성, 구축하는 방법을 제공하고자 하는 것이다.

과제의 해결 수단

[0010] 본 발명의 일 측면에 따르면, 본 발명은 RNA-단백질 복합체 데이터에 기초하여 RNA 서열의 특징 벡터와 단백질을 구성하는 아미노산 서열의 특징 벡터를 구축하고, 구축된 특징 벡터를 토대로 랜덤 포레스트(random forest) 모델을 적용하여 RNA 서열과 단백질 서열을 혼련시키는 서열 학습 수단; 및 무작위 RNA 서열을 필터링(filtering)하고, 상기 서열 학습 수단을 통해 학습된 랜덤 포레스트 모델을 적용하여, 필터링 된 RNA 서열 중에서 표적 단백질 분자와 결합하는 후보 RNA 압타머를 구축하는 서열 발굴 수단을 포함하고, 상기 서열 학습 수단에서 RNA 특징 벡터는 단백질을 구성하는 아미노산과의 상호작용 경향(interaction propensity), RNA 서열의 모노-뉴클레오타이드(mono-nucleotide; 단일-염기) 조성, 디-뉴클레오타이드(di-nucleotide; 2-염기) 조성 및 유사 트리-뉴클레오타이드(pseudo tri-nucleotide; 유사 3-염기) 조성을 토대로 작성되고, 상기 서열 학습 수단에서 단백질 특징 벡터는 단백질을 구성하는 아미노산의 조성-전하-분포 및 유사 아미노산 조성을 토대로 작성되는 표적 단백질 분자와 결합하는 후보 RNA 압타머를 생성하는 프로그램을 기록한 컴퓨터로 판독할 수 있는 기록 매체를 제공한다.

[0011] 선택적인 실시형태에서, 상기 프로그램은, 상기 서열 발굴 수단에 의해 선별된 후보 RNA 압타머를 생성하는 모델을 평가 척도를 사용하여 평가하는 서열 평가 수단을 더욱 포함할 수 있다.

[0012] 예를 들어, 상기 평가 척도는 민감도(Sensitivity), 특이도(Specificity), 정확도(Accuracy), 양성예측도(Positive predictive value), 음성예측도(Negative predictive value) 및 매튜 상관계수(Matthews correlation coefficient) 중에서 선택되는 적어도 하나인 것을 특징으로 한다.

[0013] 일례로, 상기 서열 발굴 수단은, 상기 무작위 RNA 서열의 2차 구조, 자유 에너지, 쌍을 이루지 않는 뉴클레오타이드의 개수 및 풀(pool) 내에서 동일한 2차 구조의 개수를 한정하여 상기 무작위 RNA 서열을 필터링 하는 것을 특징으로 한다.

[0014] 한편, 상기 RNA 특징 벡터를 구성하는 상기 유사 트리-뉴클레오타이드 조성은, 트리-뉴클레오타이드의 소수성(hydrophobicity), 친수성(hydrophilicity) 및 측쇄-중량(side-chain mass)를 포함하고, 상기 단백질 특징 벡터를 구성하는 상기 유사 아미노산 조성은, 아미노산의 소수성, 친수성, 측쇄-중량, 이온화 지수 및 등전점을 포함할 수 있다.

[0015] 본 발명의 다른 측면에 따르면, 본 발명은 컴퓨터에 의해 실현되는 프로그램을 이용하여 표적 단백질 분자와 결

합하는 후보 RNA 압타머를 생성하는 방법으로서, 서열 학습 수단에 의하여, RNA-단백질 복합체 데이터에 기초하여 RNA 서열의 특징 벡터와 단백질을 구성하는 아미노산 서열의 특징 벡터를 구축하고, 구축된 특징 벡터를 토대로 랜덤 포레스트(random forest) 모델을 적용하여 RNA 서열과 단백질 서열을 훈련시키는 단계; 및 서열 발굴 수단에 의하여, 무작위 RNA 서열을 필터링(filtering)하고, 상기 서열 학습 수단을 통해 학습된 랜덤 포레스트 모델을 적용하여, 필터링 된 RNA 서열 중에서 표적 단백질 분자와 결합하는 후보 RNA 압타머를 구축하는 단계를 포함하고, 상기 서열 학습 수단에서 RNA 특징 벡터는 단백질을 구성하는 아미노산과의 상호작용 경향(interaction propensity), RNA 서열의 모노-뉴클레오타이드(mono-nucleotide; 단일-염기) 조성, 디-뉴클레오타이드(di-nucleotide; 2-염기) 조성 및 유사 트리-뉴클레오타이드(pseudo tri-nucleotide; 유사 3-염기) 조성을 토대로 작성되고, 상기 서열 학습 수단에서 단백질 특징 벡터는 단백질을 구성하는 아미노산의 조성-전이-분포 및 유사 아미노산 조성을 토대로 작성되는 표적 단백질 분자와 결합하는 후보 RNA 압타머를 생성하는 방법을 제공한다.

[0016] 필요한 경우, 또한, 서열 평가 수단에 의하여, 상기 서열 발굴 수단에 의해 구축된 후보 RNA 압타머를 생성하는 모델을 평가 척도를 사용하여 평가하는 단계를 더욱 포함할 수도 있다.

발명의 효과

[0017] 본 발명의 프로그램 및 방법에 따르면, 기계 학습 알고리즘으로서, 의사결정 트리의 단점을 개선한 랜덤 포레스트 알고리즘을 활용한 훈련, 학습 모델을 활용하여 특정 표적 분자와 결합할 가능성이 매우 높은 압타머를 생성, 구축할 수 있다. 표적 분자와 결합할 가능성이 높은 잠재적인 압타머 후보 물질만으로 초기 pool을 형성하고, 이와 같이 감소된 pool 내의 핵산 라이브러리에서 출발하여 신속하고 효율적으로 최종 압타머를 선별할 수 있다.

[0018] 특정 표적 분자와 결합할 가능성이 높은 잠재적 압타머 후보 물질만으로 최종 압타머를 신속하고 효율적으로 선별할 수 있기 때문에, 신약 표적 단백질에 대한 압타머와 경쟁적으로 결합할 수 있는 신약 후보 물질을 선별하여 새로운 신약 후보 물질의 개발, 압타머를 바이오마커(biomarker)로 활용하는 진단 시약이나 마이크로어레이(microarray)의 개발, 오염원 등과 같은 환경 유해물질 또는 식품 유해물질 등의 농도를 측정하기 위하여 압타머 또는 압타머-표적 분자의 결합 여부에 따른 신호를 검출하는 바이오 센서 등을 개발할 때 본 발명이 활용될 수 있을 것으로 기대된다.

도면의 간단한 설명

[0019] 도 1은 본 발명의 예시적인 실시형태에 따라, 기계 학습 알고리즘의 하나인 랜덤 포레스트 알고리즘을 활용하여 표적 단백질 분자와 결합하는 후보 RNA 압타머를 구축, 생성하는 프로그램이 탑재된 컴퓨터와, 프로그램이 기록된 컴퓨터로 판독 가능한 기록 매체의 구성을 개략적으로 도시한 도면이다.

도 2는 본 발명의 예시적인 실시예에 따라, 후보 RNA 압타머 서열을 발굴하기 위한 전체 framework를 개략적으로 나타낸 도면이다. mC는 모노-뉴클레오타이드(mono-nucleotide; 단일-염기) 조성, dC는 디-뉴클레오타이드(di-nucleotide; 2-염기) 조성, PseTNC는 유사 트리-뉴클레오타이드 조성(pseudo tri-nucleotide composition, 유사 3-염기 조성), C-T-D는 아미노산 그룹의 조성-전이-분포(composition-transition-distribution), PseAAC는 유사 아미노산 조성(pseudo amino acid composition)을 나타낸다.

도 3은 본 발명의 예시적인 실시형태에 따라, 랜덤 포레스트 알고리즘을 활용하여 표적 단백질 분자와 결합하는 후보 RNA 압타머를 구축, 생성하는 방법을 개략적으로 도시한 플로 차트이다.

도 4는 본 발명의 예시적인 실시예에 따라, RNA 서열에서 27개의 뉴클레오타이드로 이루어진 RNA 서열의 포지티브 윈도우 및 네거티브 윈도우의 일례를 나타낸다. RNA 서열이 27개 뉴클레오타이드의 슬라이딩 윈도우를 사용하여 스캐닝 됨에 따라, 해당 윈도우를 나타내는 특징 벡터(feature vector)가 생성된다. +는 단백질과 결합하는 뉴클레오타이드를 나타내고, -는 단백질과 결합하지 않는 뉴클레오타이드를 나타낸다. 하나의 윈도우에서 중간 뉴클레오타이드가 단백질과 결합하는 뉴클레오타이드라면, 해당 윈도우를 암호화하는 특징 벡터는 포지티브로 간주된다. 반면, 하나의 윈도우가 단백질과 결합하지 않는 뉴클레오타이드만을 포함한다면, 해당 윈도우에 대한 특징 벡터는 네거티브로 간주된다. 포지티브도 아니고 네거티브도 아닌 특징 벡터들은 학습 과정에서 사용되지 않는다.

도 5는 공지된 RNA 압타머의 2차 구조의 예를 나타낸 것이다. 2차 구조는 땀글링 말단(dangling end) 염기를 가지거나 가지지 않으면서 적어도 3개의 염기쌍(base pair)에 의해 종결되는(closed) 구조를 가지며, 붉은 원으로

표시된 표적 단백질 분자와 가장 잘 결합하는 뉴클레오타이드는 해당 RNA 2차 구조의 단일-가닥 부분에서 관찰된다.

발명을 실시하기 위한 구체적인 내용

- [0020] 이하, 필요한 경우에 첨부하는 도면을 참조하면서 본 발명을 설명한다.
- [0021] 다양한 세포 과정에서 핵산 분자와 단백질 분자 사이의 상호작용은 필수적이다. 최근 대용량 데이터 처리 기술 (high-throughput technology)에 의하여 단백질 분자와 핵산 분자 사이의 상호작용에 대한 많은 데이터가 생성되어, 특정 서열에서 결합 부위(binding sites)를 예측하거나 또는 이들 서열 사이의 상호작용을 결정하기 위하여, 컴퓨터 모델을 이용한 방법을 개발할 필요가 있다.
- [0022] 특히, 단순히 하나의 핵산 서열이 표적 단백질 분자에 결합하는지를 결정하는 단순한 분류 모델이나 특정 표적 단백질 분자에 대해서만 적용되는 것을 넘어서, 표적 단백질 분자에 대하여 신규한 압타머 서열을 발굴, 생성할 수 있는 컴퓨터 모델을 개발할 필요가 있다. 본 발명에 따르면 표적 단백질 분자와 결합하는 후보 RNA 압타머 서열(apramer sequences)을 생성, 선별, 평가 및/또는 검증할 수 있도록 랜덤 포레스트(Random Forest; RF)라는 기계 학습 모델을 이용하는데, 이에 대해서 설명한다.
- [0023] 도 1은 발명의 예시적인 실시형태에 따라, 랜덤 포레스트 알고리즘/모델을 활용하여 표적 단백질 분자와 결합하는 후보 RNA 압타머를 구축, 생성하는 프로그램이 탑재된 컴퓨터와, 프로그램이 기록된 컴퓨터로 판독 가능한 기록 매체의 구성을 개략적으로 도시한 도면이고, 도 2는 본 발명의 예시적인 실시형태에 따라, 랜덤 포레스트 모델을 적용하여 표적 단백질 분자와 결합하는 후보 RNA 압타머를 구축, 생성하는 방법을 개략적으로 도시한 플로 차트이다.
- [0024] 도 1에 나타내 바와 같이, 표적 단백질 분자와 결합하는 후보 RNA 압타머 서열을 생성하는 프로그램(210)이 기록된 매체(200)는 적절한 컴퓨터(100)에 탑재될 수 있다. 컴퓨터(100)는 최소한의 데이터 처리(data processing) 능력을 가지는 데이터 처리 장치이다.
- [0025] 예를 들어, 컴퓨터(100)는 데스크톱, 노트북과 같은 컴퓨터 단말기이고/이거나, 스마트폰, 태블릿 PC, PDA 등과 같은 모바일 단말기일 수 있다. 필요한 경우, 컴퓨터(100)는 인터넷이나, 3G, LTE 및/또는 5G 등과 같은 데이터 통신망과 연결될 수 있다. 예를 들어, 컴퓨터(100)는 서버(도시하지 않음)와 통신망을 통하여 연결될 수도 있고, 서버에 로컬(local) 형태로 연결될 수도 있지만, 본 발명에 따른 컴퓨터가 반드시 통신망에 연결되는 것은 아니다.
- [0026] 컴퓨터(100)는 다양한 하드웨어(110 내지 140 및 160)와 이러한 하드웨어를 실행시키기 위한 구동/응용 프로그램(150)과 같은 소프트웨어인 다양한 컴포넌트를 포함할 수 있다.
- [0027] 예시적인 실시형태에서, 컴퓨터(100)는 키보드, 마우스 및/또는 터치 패널과 같이 데이터 또는 정보, 예를 들어 응용 프로그램의 일종인 후보 압타머를 생성하는 프로그램을 이용하여 적절한 후보 압타머를 훈련, 학습하기 위한 초기 입력 값을 작성할 수 있는 입력부(110)를 갖는다. 컴퓨터(100)는 또한 입력 또는 다운로드 된 데이터 또는 정보를 인쇄하기 위한 프린터와 같은 출력부(120)와, 이들 데이터를 화면으로 표시하기 위한 모니터와 같은 디스플레이(130)를 포함한다. 아울러, 컴퓨터(100)는 다양한 데이터 및 정보(예를 들어, 표적 단백질 분자의 아미노산 서열이나, 후보 RNA 압타머의 2차 구조 데이터; 생성된 후보 압타머 서열; 공지된 종래의 RNA 압타머 서열 및/또는 2차 구조 등)를 저장하고 있는 RAM, ROM, 하드디스크 등과 같은 메모리(140)를 가지고 있으며, 컴퓨터(100)에서의 다양한 구현 작업을 위하여 구동/응용 프로그램(150)을 탑재하고 있다.
- [0028] 컴퓨터(100)에서 구현되는 구동/응용 프로그램(150)의 예로서 WINDOWS 등과 같은 구동 소프트웨어는 물론이고, 문서, 이미지, 영상 데이터 등의 콘텐츠를 작성, 편집할 수 있는 콘텐츠 편집 소프트웨어, 인터넷과 같은 통신망에서의 데이터나 정보를 검색하기 위한 브라우저, 인공 신경망을 이용한 훈련, 학습 모델을 구축하기 위한 소프트웨어 등을 들 수 있다. 하지만, 그 외에도 다양한 구동/응용 프로그램이 컴퓨터(100)에 탑재될 수 있을 것이다.
- [0029] 또한, 컴퓨터(100)는 이들 하드웨어와 소프트웨어의 작업을 적절히 제어할 수 있도록 CPU와 같은 제어부(160)를 포함한다. 특히 컴퓨터(100) 중의 메모리(140)와 제어부(160)는 기록 매체(200)에 기록된 프로그램(210) 명령에 따른 일련의 작업을 실현한다는 점에서, 일종의 프로그램 실행부를 구성한다.
- [0030] 한편, 컴퓨터(100)가 판독할 수 있도록 컴퓨터(100)에 탑재될 수 있는 기록 매체(200)는 본 발명에 따라 표적

단백질 분자와 결합하는 후보 압타머 서열을 생성, 선택, 검증하는 프로그램(210)이 기록된다. 후술하는 바와 같이, 본 발명에 따른 프로그램(210)을 사용하여 표적 단백질 분자와 결합하는 후보 RNA 압타머 서열을 효율적으로 생성할 수 있다. 예를 들어, 본 발명에 따라 컴퓨터로 판독 가능한 기록 매체(200)는 표적 단백질 분자와 결합하는 후보 압타머 서열을 생성, 선별, 평가 및 검증하기 위한 일련의 프로그램 명령은 물론이고, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합한 것을 포함할 수 있다.

[0031] 본 발명에 따라 후보 단백질 서열과 결합하는 후보 압타머 서열을 생성하는 프로그램(210) 명령이 기록된 컴퓨터로 판독 가능한 기록 매체(200)의 예는 컴퓨터 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록 장치를 포함한다. 컴퓨터(100)가 판독할 수 있는 기록 매체(200)의 예로는 자기 테이프, 하드 디스크, 플로피 디스크와 같은 자기 매체, CD-ROM, DVD와 같은 광기록 매체, 플롭티컬(floptical) 디스크와 같은 자기-광 매체, 및 ROM, RAM, 플래시 메모리 등과 같은 프로그램 명령을 저장, 수행하도록 특별히 구성된 하드웨어 장치가 포함된다.

[0032] 또한, 캐리어 웨이브(예를 들어, 인터넷을 통해 전송)의 형태로 구현되는 것도 포함되는데, 이러한 기록 매체(200)는 프로그램 명령, 데이터 구조 등을 지정하는 신호를 전송하는 반송파를 포함하는 광 또는 금속선, 도파관 등의 전송 매체일 수도 있다. 이러한 컴퓨터(100)가 판독 가능한 기록 매체(200)는 통신망으로 연결된 컴퓨터 시스템에 분산되어, 분산 방식으로 컴퓨터(100)가 읽을 수 있는 코드가 저장되고 실행될 수 있다. 한편, 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 아래에서 예시적으로 기술하고 있는, 본 발명을 구현하기 위한 기능적인 프로그램, 코드 및 코드 세그먼트들은 본 발명이 속하는 기술분야에서 통상의 프로그래머들에 의해 용이하게 추론될 수 있을 것이다.

[0033] 기록 매체(200)에 기록된 표적 단백질 분자와 결합하는 후보 RNA 압타머를 생성하는 프로그램(이하, RNA 압타머 생성 프로그램이라고 약칭한다, 210)은 적절한 학습, 훈련 모델을 통하여 RNA 서열 및 단백질 아미노산 서열을 훈련하는 서열 학습 수단(220)과, 무작위 생성된 RNA 서열로부터 표적 단백질 분자에 결합하는 적절한 후보 RNA 압타머를 구축하는 서열 발굴 수단(230)을 포함하고, 필요에 따라 서열 발굴 수단(230)에서 구축된 후보 RNA 압타머의 적절성 등을 평가하는 서열 평가 수단(240)을 포함할 수 있다.

[0034] 본 발명에 따라 후보 RNA 압타머를 발굴하는 과정을 도 3을 참조하면서 설명한다. 도 3은 본 발명의 예시적인 실시예에 따라 표적 분자와 결합하는 후보 RNA 압타머를 발굴하기 위한 전체 프레임워크(framework)를 개략적으로 나타낸 도면이다.

[0035] 서열 학습 수단(220)은 단백질-RNA 복합체(complex) 데이터를 가지고 있는 공지된 데이터베이스를 이용하여, RNA 특징 벡터와 단백질 특징 벡터를 구축하고, 구축된 특징 벡터로부터 랜덤 포레스트(random forest; RF) 모델을 이용하여 RNA 서열과 단백질 서열을 훈련시킨다.

[0036] 하나의 예시적인 실시형태에서, 결합면역침전법(Cross-linking and Immunoprecipitation, CLIP; 클립)에 따라 분석, 확인된 단백질-RNA 복합체 데이터 세트를 제공하는 CLIPdb와 같이 생체내(*in vivo*) 분석 방법인 클립 및 이의 변형, 개량 분석 방법을 통하여 확인된 단백질-RNA 복합체 데이터 세트를 제공하는 공지된 데이터베이스를 활용할 수 있다.

[0037] 예를 들어, 클립 및 이의 변형인 생체내 분석 방법은 1) RNA-단백질의 결합 부위를 확인하기 위하여 UV-가교(UV cross-linking)과 면역침전(Immunoprecipitation)을 결합하는 CLIP-seq(CLIP sequencing; HITS(high-throughput sequencing)-CLIP), 2) 세포의 RNA 결합 단백질이나 microRNA를 함유하는 ribonucleoprotein 복합체를 확인하는데 사용되는 PAR-CLIP(photoactivatable ribonucleoside-enhanced cross-linking and Immunoprecipitation), 3) 단백질과 RNA 분자를 공유 결합하기 위하여 UV 광을 이용하는 iCLIP(individual nucleotide-resolution cross-linking and Immunoprecipitation), 4) RNA 양을 감소시키고 면역침전된 RNA의 방사선 표지를 생략하는 sCLIP(simple CLIP) 등을 들 수 있다.

[0038] 다른 예시적인 실시형태에서, 시험관내(*in vitro*) 분석 방법인 SELEX(Systematic Evolution of Ligands by EXponential enrichment) 및 이의 변형, 개량인 시험관내 분석 방법을 통하여 확인된 단백질-RNA 복합체 데이터 세트를 제공하는 다른 공지의 데이터베이스를 활용할 수 있다.

[0039] 일례로, 셀렉스(SELEX)는 1) 핵산 라이브러리 제조, 2) 친화 크로마토그래피 등을 이용하여 표적 단백질 분자와 결합하는 핵산 구조체 선별, 3) 핵산 구조체를 분리, 증폭하는 과정을 반복하여 확인된 단백질-RNA 복합체에 대한 데이터 세트가 제공될 수 있다.

- [0040] 한편, SELEX를 개량, 변형한 다른 방법을 통하여 확인된 단백질-RNA 복합체 데이터 세트가 서열을 학습하기 위하여 활용될 수 있다. SELEX를 개량, 변형한 다른 방법은 1) 핵산 분해효소에 의해 분해되지 않는 L-oligo-nucleotide를 이용한 거울상 압타머(Spiegelmer) 방법, 2) 고순도의 단백질 정제 과정 없이 세포 표면에 존재하는 단백질과 결합하는 cell-to-Aptamer 방식으로 특이적 압타머를 발굴하는 Cell SELEX, 3) 모세관 전기영동을 이용하는 capillary electrophoresis SELEX (CE-SELEX), 4) counter-SELEX, 5) Toggle SELEX 등을 포함할 수 있다. 그 외에도, SELEX를 통해 얻어진 초기의 압타머를 안정적이고 강력한 압타머로 개량하기 위하여, 1) RNA 압타머의 Ribose 2'-OH를 2'-F 나 2'-NH₂, 2'-O-methyl group으로 치환하거나, 압타머를 polyethylene glycol(PEG)과 같은 고분자나 diacylglycerol 혹은 cholesterol을 접합시키는 post-SELEX 과정이 수행될 수도 있다.
- [0041] 서열 학습 수단(220)은 RNA 서열 및 단백질의 아미노산 서열을 훈련, 학습시키기 위하여 이들 서열에 대한 적절한 특징 벡터를 구축한다. 본 발명에서, RNA 서열에 대한 특징 벡터는 RNA의 아미노산과의 상호작용 경향(interaction propensity; IP), RNA 서열의 모노-뉴클레오타이드 조성(mono-nucleotide composition; mC, 단일-염기 조성), 디-뉴클레오타이드 조성(di-nucleotide composition; dC, 2-염기 조성) 및 유사 트리-뉴클레오타이드 조성(pseudo tri-nucleotide composition; PseTNC, 유사 3-염기 조성)을 토대로 작성될 수 있다.
- [0042] RNA-단백질 상호작용을 예측할 때, RNA 특징 벡터의 기초 특징의 하나인 아미노산과의 뉴클레오타이드 트리플렛의 상호작용 경향(IP)은 매우 강력한 특징이 될 수 있다. 다른 특징 역시 RNA 서열의 특징으로서 활용될 수 있다.
- [0043] 한편, 단백질의 아미노산 서열을 훈련시키기 위한 특징 벡터는 단백질을 구성하는 아미노산의 조성-전이-분포(composition-transition-distribution; C-T-D) 및 유사 아미노산 조성(Pseudo Amino Acid composition; PseAAC)을 토대로 작성될 수 있다. 단백질 특징을 추출하기 위하여 단백질을 구성하는 20개의 아미노산은 적절한 기준에 따라 군집화(clustering) 될 수 있다. 일례로, 단백질을 구성하는 아미노산은 쌍극자(dipole) 및 부피(volume)를 기준으로 군집화 될 수 있지만, 본 발명이 이에 한정되지 않는다.
- [0044] 또한, RNA 서열을 훈련시키기 위한 특징의 하나인 유사 3-염기 조성(PseTNC) 및 단백질의 아미노산 서열을 훈련시키기 위한 특징의 하나인 유사 아미노산 조성(PseAAC)과 관련해서는 각각 RNA 서열을 구성하는 뉴클레오타이드의 생리화학적 성질이 추출될 수 있다. 일례로, 유사 3-염기 조성과 관련하여 공지된 단백질-RNA 복합체에서 RNA 서열을 구성하는 뉴클레오타이드의 소수성(hydrophobicity), 친수성(hydrophilicity) 및 뉴클레오타이드를 구성하는 염기의 측쇄-중량(side-chain mass)이 뉴클레오타이드의 특성으로서 고려될 수 있다. 한편, 유사 아미노산 조성과 관련해서, 공지된 단백질-RNA 복합체에서 단백질을 구성하는 아미노산의 소수성, 친수성, 측쇄-중량, 이온화 지수(예를 들어 카르복시 말단 및 아미노기 말단의 이온화 지수), 등전점이 고려될 수 있다. 하지만, 그 외에도 필요에 따라 RNA를 구성하는 뉴클레오타이드 및/또는 단백질을 구성하는 아미노산의 다른 특징이 활용될 수도 있다.
- [0045] 이들 유사 3-염기 조성 및 유사 아미노산 조성을 연산할 때, 가중치 인자(weight factor)와 단백질의 아미노산 서열을 따라 형성되는 상관관계의 개수 순위(counted rank; 티어(tier), 통상 λ 로 표시됨)를 설정할 수 있다. 일례로, 가중치 인자(weight factor)로서는 0.05 내지 0.70 사이의 임의의 값을 설정할 수 있다. 한편, 상관관계의 티어는 1) 입력되는 단백질 서열의 길이보다 작아야 하고, 2) 음의 정수가 아니어야 하며, 3) 티어 값이 0 일 때, 유사 아미노산 조성(PseAAC)의 출력은 통상적인 아미노산 조성으로 지정되도록 설정할 필요가 있다. 하나의 예시적인 실시형태에 따르면, 단백질을 구성하는 아미노산 서열의 상관관계 티어 값은 1 내지 5 사이의 정수로 설정될 수 있다.
- [0046] 한편, SELEX 등의 실험 기법에서 최종적으로 확인된 RNA 압타머가 30개 이하인 경우가 많다. 또한, PAR-CLIP 기법의 경우에 통상적으로 21개 내지 35개의 뉴클레오타이드로 구성되는 단백질 결합 RNA 서열을 확인할 수 있다. 따라서, 이러한 분석 방법을 통해서 확인된 뉴클레오타이드에 대응되는 길이를 가지는 RNA 서열을 훈련 및 학습을 위한 서열로 설계하는 것이 바람직할 수 있다. 따라서 훈련되는 RNA 서열의 뉴클레오타이드 개수는 20 내지 40개, 바람직하게는 21개 내지 35개로 설정될 수 있지만, 본 발명이 이에 한정되지 않는다.
- [0047] 공지된 단백질-RNA 복합체의 데이터 세트에 기초하여, RNA 서열과 단백질의 아미노산 서열의 특징 벡터를 구축하면, 서열 학습 수단(230)은 기계 학습(machine learning) 알고리즘(또는 모델)을 이용하여 해당 RNA 서열과 아미노산 서열을 훈련시킨다. 예를 들어, 서열 학습 수단(230)은 랜덤 포레스트(random forest, RF) 모델을 활용하여 RNA 서열 및 단백질의 아미노산 서열의 특징 벡터를 훈련시킬 수 있다. 랜덤 포레스트 모델은 의사결정

트리의 단점을 개선하기 위한 알고리즘의 하나로써, 변수에 임의성을 더하여 앙상블 이론이 갖는 장점을 극대화하여 예측 및 분류 정확도를 개선하여 안정성을 얻을 수 있는 것으로 알려져 있다.

- [0048] 랜덤 포레스트 모델은 데이터에서 부트스트래핑(bootstrapping) 과정을 통한 배깅(bagging, Bootstrap Aggregation)을 통하여 주어진 데이터 세트에서 무작위로 subset을 N번 샘플링(sampling)하여 N개의 샘플링 데이터 세트(즉, N개의 예측 모델)를 생성할 수 있다. 이어서, 각각의 샘플링 된 데이터 세트에서 임의의 변수를 선택하는 과정을 진행하는데, 변수의 개수를 선택할 때, M개의 총 변수들 중에서 $\sqrt{M}$ 또는 M/3개의 개수만큼 변수들을 무작위로 선택하고 나머지 변수는 모두 제거하는 과정 반복한다. 이와 같이 변수 선택이 진행된 의사결정 트리들을 종합하여 앙상블(Ensemble) 모델을 만들고, OOB 오류(Out-Of-Bag error)를 통해 오-분류를 평가한다.
- [0049] 랜덤 포레스트 모델을 적용하면, 학습 오류를 구성하는 바이어스(bias, 이 값이 높으면 예측 결과가 실제 결과와 비교해서 부정확함)를 낮추면서, 배리언스(variance, 이 값이 높으면 예측 결과가 특정 데이터 세트에서는 잘 맞지만 다른 데이터 세트에서는 잘 맞지 않음)를 줄일 수 있다.
- [0050] 하나의 예시적인 실시형태에서, 윈도우의 중앙 뉴클레오타이드가 단백질과 결합하는 뉴클레오타이드인 경우, 해당 윈도우를 암호화하는 특징 벡터는 포지티브로 정의될 수 있다. 반면, 윈도우가 단백질과 결합하는 뉴클레오타이드를 포함하지 않은 경우, 해당 윈도우를 암호화하는 특징 벡터는 네거티브로 정의될 수 있다. 포지티브도 아니고 네거티브도 아닌 특징 벡터는 제외되어, 훈련 및 학습 과정에서 사용되지 않을 수 있다.
- [0051] 한편, 서열 발굴 수단(230)은 무작위 서열을 필터링(filtering)하고, 전술한 서열 학습 수단(220)을 통해 훈련된 랜덤 포레스트 모델을 적용하여, 필터링 된 RNA 서열 중에서 표적 단백질 분자와 결합하는 후보 RNA 압타머를 구축, 생성한다.
- [0052] 서열 발굴 수단(230)이 무작위 RNA 서열을 필터링 할 때, RNA 서열의 2차 구조의 제약 조건(constraints)으로서 2차 구조, 자유 에너지, 쌍을 이루지 않는 염기의 개수, 풀 내에서 동일한 2차 구조의 개수 등을 설정할 수 있다.
- [0053] RNA 서열의 2차 구조로서 말단 부위(예를 들어 RNA 서열의 1 내지 5번째 염기와, 끝에서 1 내지 5번째 염기)는 당글링 말단(이중 가닥을 형성할 수 있도록 결합이 이루어진 2개의 핵산 가닥 중에서 하나의 핵산 가닥 말단이 다른 핵산 가닥의 말단보다 길거나 짧은 말단) 염기를 가지거나 가지지 않을 수 있지만, 말단 부위는 적어도 3개의 염기 쌍(예를 들어 3 내지 7개, 바람직하게는 3 내지 5개의 염기 쌍)에 의해 종결될 필요가 있다.
- [0054] RNA 서열의 자유 에너지는 -4.0 내지 -7.0 kcal/mol, 바람직하게는 -5.0 내지 -6.0 kcal/mol 미만이어야 한다. 이는 공지된 RNA 압타머의 자유 에너지의 Z-스코어 값을 고려한 것이다.
- [0055] 한편, 필터링을 위한 제약 조건으로서 RNA 서열 중에서 쌍을 이루지 않는(unpaired) 뉴클레오타이드의 개수를 일정 개수 이상 설정할 필요가 있다. 일반적으로 RNA의 2차 구조에서 이중 가닥 부분의 뉴클레오타이드(즉, 염기 쌍을 이루는 부분의 뉴클레오타이드)에 비하여 단일 가닥 부분의 뉴클레오타이드(즉, 염기 쌍을 이루지 않는 부분의 뉴클레오타이드)가 단백질에 빈번하게 결합한다. 단일 가닥 부분의 뉴클레오타이드가 단백질에 주로 결합하는 데는 몇 가지 이유가 있다. 단일 가닥 부분의 뉴클레오타이드는 그들의 입체구조(conformation)를 쉽게 바꿀 수 있기 때문에, 단일 가닥 부분의 뉴클레오타이드(즉, 쌍을 이루지 않는 뉴클레오타이드)는 이중 가닥 부분의 뉴클레오타이드(즉, 염기 쌍을 형성하는 뉴클레오타이드)에 비하여 보다 유연하다. 뿐만 아니라, 쌍을 이루지 않는 뉴클레오타이드는 단백질의 아미노산과 수소결합을 형성할 수 있는 전자주개(donor) 또는 전자받개(acceptor) 원자를 가지고 있다.
- [0056] 따라서, 상대적으로 유연한 구조를 가지면서 단백질의 아미노산과 용이하게 수소결합을 형성할 수 있는 쌍을 이루지 않는 뉴클레오타이드의 개수는 구축하고자 하는 후보 RNA 압타머의 대략 30% 이상으로 설정하는 것이 바람직할 수 있다. 예를 들어, 후보 RNA 압타머의 뉴클레오타이드를 20 내지 40개로 설계한 경우, 쌍을 이루지 않는 뉴클레오타이드의 개수를 대략 5개 내지 20개로 설정할 수 있다.
- [0057] 아울러, 풀(pool) 내에서 동일한 2차 구조가 반복되는 경우, 적절한 후보 RNA 압타머를 구축하기 곤란할 수 있다. 따라서 풀 내에서 동일한 2차 구조의 개수는 300개 미만, 예를 들어 200개 미만, 바람직하게는 150개 미만, 예를 들어 100개 미만으로 설정할 수 있다.
- [0058] 서열 발굴 수단(230)은 서열 학습 수단(230)에서 훈련된 랜덤 포레스트 모델을 필터링 된 RNA 서열에 적용하여 후보 RNA를 예측, 구축, 생성할 수 있다.

- [0059] 이와 같이, 공지된 RNA-단백질 복합체로부터, 이들의 서열 수준 및 구조 수준에서, 상호작용하는 RNA 서열과 표적 단백질 분자의 주요한 특징(feature)을 확인한다. 본 발명에서 상호작용하는 RNA와 표적 단백질 분자의 주요 특징을 사용하여 랜덤 포레스트(RF) 모델을 구축한다. 뒤에서 확인되는 바와 같이, 교차 검증(cross validation)과 본 발명에서 구축된 RF 모델에 대한 독립적인 테스트 수행 결과는, 본 발명의 RF 모델에서 발견된 일련의 잠재적인 RNA 압타머는 강력한 후보 압타머 서열을 포함하고 있어서, 통상적인 SELEX 등의 공정에 의해 요구되는 시간과 비용을 크게 줄일 수 있다.
- [0060] 선택적으로, 도 2에 개략적으로 나타낸 절차에 따라 서열 학습 수단(220) 및 서열 발굴 수단(230)을 통하여 예측, 구축, 선별된 후보 RNA 압타머에 대하여 공지된 구조 분석 소프트웨어를 활용하여 그 구조를 분석할 수 있다. 예를 들어, 구조 분석 소프트웨어를 활용하여, 서열 발굴 수단(230)에 의해 선별된 후보 RNA 압타머의 2차 구조와, 표적 단백질 분자에 도킹된 후보 RNA 압타머의 결합 부위를 중심으로 하는 결합 구조를 분석할 수 있다.
- [0061] 이때, 구조 분석 소프트웨어를 활용하여 서열 발굴 수단(230)에 의해 선별된 후보 RNA 압타머의 2차 구조와 공지된 RNA 압타머의 2차 구조를 비교 분석할 수 있고/있거나, 표적 단백질 분자에 후보 RNA 압타머를 도킹시켜, 3차원 구조의 단백질-RNA 복합체를 이룬 상태에서 후보 RNA 압타머의 표적 단백질 분자에 대한 결합 부위를 중심으로 후보 RNA 압타머의 표적 단백질 분자에 대한 결합 구조와, 공지된 RNA 압타머의 표적 단백질 분자에 대한 결합 구조를 비교할 수 있다.
- [0062] 일례로, 후보 RNA 압타머 및/또는 공지된 RNA 압타머의 2차 구조는 RNAfold (Lorenz et al, 2011, ViennaRNA Package 2.0. Algorithms Mol. Biol. 6(1), 26) 등의 소프트웨어를 사용할 수 있고, PseudoViewer (Byun and Han, 2009, PseudoViewer3: Bioinformatics 25(11), 1435-1437) 등의 소프트웨어를 이용하여 시각화될 수 있다. 한편, 표적 단백질 분자에 도킹한 상태에서 후보 RNA 압타머 및/또는 공지된 RNA 압타머의 결합 구조는 HDOCK (Yan et al., 2017, HDock: a web server for proteinprotein and proteinDNA/RNA docking based on a hybrid strategy, Nucleic Acids Res. 45(W1), W365-W373) 등의 소프트웨어를 활용할 수 있으며, RNA 압타머의 3차 구조는 RNAComposer (Popenda et al., 2012, Automated 3D structure composition for large RNAs, Nucleic Acids Res. 40(14), e112) 등의 소프트웨어를 사용하여 분석할 수 있다.
- [0063] 한편, 후보 RNA 압타머 생성 프로그램(210)은 필요에 따라 서열 발굴 수단(230)에서 후보 RNA 서열을 생성하는데 사용한 랜덤 포레스트 모델을 검증하는 서열 평가 수단(240)을 더욱 포함할 수 있다. 서열 평가 수단(240)은 전술한 서열 학습 수단(220), 서열 발굴 수단(230)의 랜덤 포레스트 모델에서 사용한 포지티브(+) 인스턴스 및 네거티브(-) 인스턴스에 대하여 적절한 평가 척도를 사용하여 평가, 검증한다. 필요에 따라, 독립적인 테스트가 또한 서열 평가 수단(240)에 의해 수행될 수 있다.
- [0064] 예시적인 실시형태에서, 서열 검증 수단(250)은 표준적인 교차 검증, 예를 들어 표준적인 10배 교차 검증이 수행될 수 있다. 선택적으로, 표준적인 10배 교차 검증 이외에도 하나 남기기(leave-one-out; LOO) 교차 검증이 수행될 수 있다. LOO 교차 검증을 수행한 이유는, 통상적인 k-배 교차 검증은 PPI(단백질-단백질 상호작용) 또는 RNA 상호작용과 같은 쌍을 이루는 입력 값에 대하여 예측 성능을 과대평가하는 경향이 있기 때문이다 (Abbasi, W.A., Minhas, F.U.A.A.: Issues in performance evaluation for host-pathogen protein interaction prediction, Journal of Bioinformatics and Computational Biology 14(3):1650011 (2016)).
- [0065] 하나의 예시적인 실시형태에서, 서열 평가 수단(240)은 본 발명에 따라 사용된 랜덤 포레스트 모델에 대하여 평가 척도로서 민감도(sensitivity), 특이도(specificity), 정확도(accuracy), 양성예측도(positive predictive value, PPV), 음성예측도(negative predictive value, NPV) 및 매튜 상관계수(Matthews correlation coefficient, MCC) 등이 사용될 수 있다. 이들 평가 척도는 하기 식 (1) 내지 (6)으로 표현될 수 있다.

[0066]
$$Sn(Sensitivity) = \frac{TP}{TP + FN} \quad (1)$$

[0067]
$$Sp(Specificity) = \frac{TN}{TN + FP} \quad (2)$$

[0068]
$$Accuracy(Acc) = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

$$\text{Positive Predictive Value(PPV)} = \frac{TP}{TP + FP} \quad (4)$$

$$\text{Negative Predictive Value(NPV)} = \frac{TN}{TN + FN} \quad (5)$$

$$\text{Matthews Correlation Coefficient(MCC)} = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (6)$$

위 식 (1) 내지 (6)에서 민감도는 서열 발굴 수단(230)의 실제 포지티브 인스턴스 중에서 랜덤 포레스트 모델에 의해 포지티브로 맞게 예측된 인스턴스의 비율이고, 특이도는 실제 네거티브 인스턴스 중에서 랜덤 포레스트 모델에 의해 네거티브로 맞게 예측된 인스턴스의 비율이며, 정확도는 실제 인스턴스 중에서 랜덤 포레스트 모델에 의해 맞게 예측된 포지티브 인스턴스 및 네거티브 인스턴스의 비율이고, 양성예측도는 랜덤 포레스트 모델에 의해 포지티브로 예측된 인스턴스 중에서 실제 포지티브 인스턴스의 비율이며, 음성예측도는 랜덤 포레스트 모델에 의해 네거티브로 예측된 인스턴스 중에서 실제 네거티브 인스턴스의 비율을 측정하는 것이다.

또한, 식 (1) 내지 (6)에서 TP(true positive, 참된 긍정)는 포지티브로 올바르게 예측된 포지티브 인스턴스, TN(true negative, 참된 부정)은 네거티브로 올바르게 예측된 네거티브 인스턴스, FP(false positive, 거짓 긍정)은 포지티브로 잘못 예측된 네거티브 인스턴스, FN(false negative, 거짓 부정)은 네거티브로 잘못 예측된 포지티브 인스턴스를 의미한다.

필요한 경우, 서열 평가 수단(240)은 선행 연구에서 확인된 데이터 세트를 활용하여, 본 발명에서 확인된 후보 RNA 구축 모델(RF 모델)에 대한 독립적인 테스트를 수행할 수도 있다.

한편, 본 발명은 전술한 RNA 압타머 생성 프로그램(210, 도 1 참조)을 이용하여 표적 단백질 분자와 결합하는 후보 RNA 압타머를 생성하는 방법에 관한 것이다. 도 3은 본 발명의 예시적인 실시형태에 따라 표적 단백질 분자와 결합하는 후보 RNA 압타머를 생성하는 방법을 개략적으로 도시한 플로 차트이다.

도 3에 나타난 바와 같이, 컴퓨터에서 실현되는 프로그램을 이용하여 표적 단백질 분자와 결합하는 후보 RNA 압타머를 생성하는 방법은, 단백질-RNA 복합체 데이터 세트를 토대로, RNA 서열 및 단백질을 구성하는 아미노산 서열을 랜덤 포레스트 모델을 활용하여 훈련, 학습하는 단계(S210 단계)와, 무작위 RNA 서열을 필터링 하고, 필터링 된 무작위 RNA 서열로부터 훈련된 랜덤 포레스트 모델을 활용하여 표적 단백질 분자와 결합하는 후보 RNA 압타머를 예측, 구축, 생성하는 단계(S220 단계)를 포함하고, 필요에 따라, 예측, 선별, 구축된 후보 RNA 압타머에 대하여 척도를 사용하여 평가하는 단계(S230 단계)를 포함할 수 있다.

RNA 서열 및 단백질의 아미노산 서열을 훈련, 학습하는 단계(S210 단계)는 서열 학습 수단(220, 도 1 참조)에 의해 수행될 수 있다. 실험적 분석 방법, 예를 들어, 클립(CLIP) 및/또는 셀렉스(SELEX) 등의 분석 방법을 통하여 확인된 단백질-RNA 복합체 데이터 세트에 기초하여 RNA 서열의 특징 벡터와 단백질을 구성하는 아미노산 서열의 특징 벡터를 구축하고, 구축된 특징 벡터에 랜덤 포레스트 모델을 적용하여 RNA 서열과 단백질 서열을 훈련시킬 수 있다. 일례로서, RNA 서열을 훈련시키기 위한 특징 벡터는 단백질을 구성하는 아미노산과의 상호작용 경향, RNA 서열의 단일-염기 조성, RNA 서열의 2-염기 조성 및 유사 3-염기 조성을 포함할 수 있다. 예를 들어, 유사 3-염기 조성은 RNA 서열을 구성하는 뉴클레오타이드의 소수성, 친수성 및 염기의 측쇄-중량을 포함할 수 있다. 이때, 훈련되는 RNA 서열의 뉴클레오타이드의 개수는 20 내지 40개로 설정될 수 있지만, 본 발명이 이에 한정되지 않는다.

또한, 단백질 서열을 훈련시키기 위한 특징 벡터는 단백질을 구성하는 아미노산의 조성-전하-분포 특성과, 유사 아미노산 조성을 포함할 수 있다. 유사 아미노산 조성은 단백질을 구성하는 아미노산의 소수성, 친수성, 측쇄-중량, 이온화 지수 및 등전점을 포함할 수 있다.

후보 RNA 압타머를 예측, 구축, 생성하는 단계(S220 단계)는 서열 발굴 수단(230, 도 1 참조)에 의해 수행될 수 있다. 본 단계에서, 무작위 RNA 서열 중에서 특정 조건을 충족하는 RNA 서열을 필터링 하고, 서열을 훈련, 학습하는 단계(S210 단계)에서 훈련된 랜덤 포레스트 모델을 필터링 된 RNA 서열에 적용하여 필터링 된 RNA 서열 중에서 표적 단백질 분자와 결합하는 후보 RNA 압타머를 예측, 구축, 생성, 선별할 수 있다.

예시적인 실시형태에서, 무작위 RNA 서열을 필터링 할 때, 무작위 RNA 서열의 2차 구조, 자유 에너지, 쌍을 이루지 않는 뉴클레오타이드의 개수 및 풀(pool) 내에서 동일한 2차 구조의 개수를 한정하여 상기 무작위 RNA 서열을 필터링 할 수 있다.

- [0081] 필요한 경우, 선별, 구축된 후보 RNA 압타머에 대하여 공지된 구조 분석 소프트웨어를 활용하여 2차 구조 및/또는 표적 단백질 분자와 도킹한 상태에서 결합 구조를 분석할 수 있다. 이때, 선별된 후보 RNA 압타머와, 공지된 RNA 압타머의 2차 구조 및 표적 단백질 분자에 대한 결합 구조가 비교, 분석될 수 있다.
- [0082] 한편, 후보 RNA 압타머를 적절한 평가 척도를 사용하여 평가하는 단계(S230 단계)는 서열 평가 수단(240, 도 1 참조)에 의해 수행될 수 있다. S230 단계는, 후보 RNA를 예측, 생성, 선별하기 위해 사용된 랜덤 포레스트 학습 모델을 적절한 평가 척도를 사용하여 평가, 검증한다. 예시적인 실시형태에서, 평가 척도는 민감도(Sensitivity), 특이도(Specificity), 정확도(Accuracy), 양성예측도(Positive predictive value), 음성예측도(Negative predictive value) 및 매튜 상관계수(Matthews correlation coefficient) 중에서 선택되는 적어도 하나일 수 있다. 평가 방법은 특별히 한정되는 것은 아니지만 표준적인 교차 검증, 예를 들어 표준적인 10배 교차 검증이나, 하나 남기기(LOO) 교차 검증이 수행될 수 있다. 필요에 따라, 본 단계에서는 선행 연구의 결과를 활용한 독립적인 테스트가 또한 수행될 수 있다.
- [0083] 이하, 예시적인 실시형태를 통하여 본 발명을 설명하지만, 본 발명이 하기 실시예에 기재된 기술사상으로 한정되지 않는다.
- [0084] 실시예
- [0085] [데이터 세트]
- [0086] 단백질 데이터 뱅크(Protein Data Bank, PDB)에서 5.0 Å 이상의 양호한 해상도를 가지는 X-선 결정학에 의해 해명된 단백질-RNA 복합체를 추출하였다. 10개 미만의 짧은 RNA 서열과, 120개를 초과하는 긴 RNA 서열을 가지는 복합체를 제거하여, 총 696개의 단백질-RNA 복합체만을 남겨두었다. 아미노산과의 뉴클레오타이드의 상호작용 경향(interaction propensity; IP)을 학습하기 위하여 남겨진 696개의 단백질-RNA 복합체를 사용하였다.
- [0087] 아울러, 전술한 696개의 단백질-RNA 복합체에 포함되지 않은, 공지된 RNA 압타머-단백질 복합체에 대한 구조 데이터를 PDB에서 얻었다. PDB에는 35개의 RNA 압타머-단백질 복합체가 존재하였다.
- [0088] 본 실시예에서 사용된 모델을 다른 모델과 비교하기 위하여, 압타머 베이스(Cruz et al., 2012, Aptamer Base: a collaborative knowledge base to describe aptamers and SELEX experiments. Database (Oxford), bas006)로부터 Li 등 (Li et al., 2014, Predictor of Aptamer-Target Interacting Pairs with Pseudo-Amino Acid Composition, PLoS One 9(1), e86729)에 의해 구축된 기준(benchmark) 데이터세트를 획득하였다. 본 실시예에 따른 모델을 테스트하기 위한 포지티브 예로서, 포지티브 데이터로부터 56개의 RNA 압타머-단백질 쌍을 사용하였다.
- [0089] [예측 모델을 학습하기 위한 특징]
- [0090] SELEX 과정은 30-mer 또는 60-mer의 다양한 영역을 가지는 올리고-뉴클레오타이드 라이브러리로부터 압타머를 선택하는데, 선택된 압타머는 통상적으로 30-mer보다 짧다. 따라서 표적 단백질 분자에 결합하는 30-mer보다 짧은 핵산 서열로 이루어진 잠재적인 RNA 압타머를 생성할 수 있도록, 본 실시예에서는 27-mer의 RNA 압타머를 생성하도록 설계하였다. 각 쌍의 RNA와 단백질 서열에 대하여, 다음과 같은 특징을 암호화하였다.
- [0091] RNA 특징: 아미노산과의 뉴클레오타이드 트리플렛(triplets)의 IP, 모노-뉴클레오타이드 조성(mC, 단일-염기 조성), 디-뉴클레오타이드 조성(dC, 2-염기 조성), 유사 트리-뉴클레오타이드 조성(PseTNC, 유사 3-염기 조성)
- [0092] 단백질 특징: 아미노산 그룹의 조성-전이-분포(C-T-D), 유사 아미노산 조성(PseAAC).
- [0093] 아미노산과의 뉴클레오타이드 트리플렛의 IP를 정의하기 위하여, 696개의 단백질-RNA 복합체의 데이터 세트를 사용하여, $4^3=64$ 인 뉴클레오타이드 트리플렛과 20개의 아미노산 사이에서 1,280개(64×20)의 IP를 연산하였다. 아울러, 각각의 RNA 서열에 대하여, 단일-염기 조성(mC)과 2-염기 조성(dC)을 또한 연산하였다.
- [0094] 한편, 단백질 서열을 표현할 수 있도록, 20개의 아미노산을 그들의 쌍극자(dipoles)와 부피(volume)를 기준으로 7개의 그룹으로 군집화(clustering)하였다. {M (methionine), S (Serine), T (Threonine), Y (Tyrosine)}, {F (phenylalanine), I (Isoleucine), L (Leucine), P (Proline)}, {H (Histidine), N (Asparagine), Q (Glutamine), W (Tryptophan)}, {A (Alanine), G (Glycine), V (Valine)}, {K (Lysine), R (Arginine)}, {D (Aspartic acid), E (Glutamic acid)}, {C (Cysteine)}. 단백질 서열 내의 각각의 아미노산 그룹에 대하여, 조성-전이-분포(composition-transition-distribution, C-T-D) 특징을 표현하였는데, 이 특징은 특징 벡터에서 $7 + 21 + 35 = 63$ 요소를 요구한다(Dubchak et al., 1995, Prediction of protein folding class using global

description of amino acid sequence, Proc. Natl. Acad. Sci. U.S.A. 92(19), 8700-8704).

- [0095] 추가적인 특징으로서, 각각의 RNA 서열에 대한 유사 3-염기 조성(PseTNC) (Chen et al., 2014, it is-PseTNC: a sequence-based predictor for identifying initiation site in human genes using pseudo trinucleotide composition. Anal. Biochem. 462, 76-83)과 각각의 단백질 서열에 대한 유사 아미노산 조성(PseAAC) (Shen and Chou, 2008. PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition, Anal. Biochem. 373(2), 386-388)을 연산하였는데, 이들 특징은 서열 내부에서 잔기(residues)의 서열-순서 효과(sequence-order effect)를 고려한다.
- [0096] 구체적으로, PseTNC와 관련해서는 뉴클레오타이드의 3개의 생리화학적 성질, 즉 소수성(hydrophobicity), 친수성(hydrophilicity), 측쇄-중량(side-chain mass)을 사용하여 PseTNC를 연산하였다. 이와 유사하게, PseAAC와 관련해서는 아미노산의 다양한 생리화학적 성질, 즉 소수성, 친수성, 측쇄-중량, 이온화 지수 1(pK1, a-CO₂H), 이온화 지수 2 (pK2, NH₃), 25℃에서의 등전점(isoelectric point, pI)을 이용하여 PseAAC를 연산하였다. PseTNC와 PseAAC를 연산할 때, 가중치(weigh factor, w)을 0.1로, 상관관계의 티어(tier)인 λ 를 1로 설정하였다. PseTNC와 PseAAC는 각각 티어 상관 인자를 가지는 트리-뉴클레오타이드 조성과 아미노산 조성을 나타낸다. 티어 상관 인자의 개수는 파라미터 λ 와 같기 때문에, PseTNC와 PseAAC에 대한 특징 벡터에서 요소의 개수는 각각 65와 21이다. 669개의 요소를 가지는 특징 벡터에서 한 쌍의 RNA 윈도우와 단백질 쌍이 암호화된다: 500 IP, 4 mC, 16 dC, 65 PseTNC, 63 C-T-D, 21 PseAAC.
- [0097] [훈련을 위한 포지티브 및 네거티브 인스턴스들]
- [0098] 랜덤 포레스트(RF) 모델을 구축하고, 단백질-RNA 복합체로부터 생성된 전술한 특징 벡터에서 RF 모델을 훈련시켰다. RF 모델을 구축할 때, 트리(trees)의 개수는 35개로 설정하였으며, 특징의 개수는 특징 요소 개수의 제곱근으로 설정하였다.
- [0099] 30-mer 미만의 잠재적인 RNA 압타머를 생성하는 것이 본 발명의 주요한 목적이기는 하지만, 데이터 세트를 훈련하고 테스트할 때 RNA 서열은 다양한 길이를 가지고 있다. 도 3에 도시한 바와 같이, 27개의 뉴클레오타이드를 사용하여, 포지티브 인스턴스(positive instance)와 네거티브 인스턴스(negative instance)를 정의하였다. 27개 뉴클레오타이드의 윈도우를 사용하여 RNA 서열이 스캐닝 되면서, 해당 윈도우를 나타내는 특징 벡터가 생성된다. 윈도우의 중앙 뉴클레오타이드가 단백질-결합 뉴클레오타이드라면, 해당 윈도우를 암호화하는 특징 벡터는 포지티브로 고려되었다. 만약 어떤 윈도우가 단백질-비결합 뉴클레오타이드만을 포함한다면, 해당 윈도우에 대한 특징 벡터는 네거티브로 고려되었다. 포지티브도 아니고 네거티브도 아닌 특징 벡터들은 훈련 데이터 세트에서 제거되었는데, 이들은 훈련을 위한 매우 불균형한 포지티브/네거티브 인스턴스를 야기하기 때문이다. 다만 예외적으로, 최초와 마지막 윈도우는 해당 윈도우의 임의의 위치에서 단백질 결합 뉴클레오타이드를 포함한다면 모두 포지티브로 고려되었다. 본 실시예에 따른 훈련 데이터세트에서, 네거티브 인스턴스에 대한 포지티브 인스턴스의 비율은 약 3:1이었다.
- [0100] [RNA 압타머의 구조적 제한]
- [0101] 주어진 단백질 서열에 대한 27-mer RNA 압타머를 발굴하기 위하여, 6×10^6 개의 27-mer 무작위 RNA 서열을 생성하고, RNAfold(Lorenz et al., 2011. ViennaRNA Package 2.0. Algorithms Mol. Biol. 6(1), 26)를 이용하여 이들의 2차 구조를 예측하였다. 무작위 RNA 서열 중에서, 2차 구조가 다음 4가지의 제약 조건(constraints)을 만족시키는 것을 필터링 하고, 후보 압타머의 초기 풀(pool)을 구축하였다.
- [0102] 첫째, 2차 구조는 다음의 형태 중의 하나의 형태에서 땀글링 말단(dangling end 말단보다 길거나 짧은 말단) 염기를 가지거나 가지지 않으면서 적어도 3개의 연속적인 염기에 의해 종결되어야 한다. 형태는 (((N21))), (((N20))), (((N20))), and (((N19))). 여기서 Nk는 k-mer 서열 단편이다.
- [0103] 둘째, 자유 에너지는 -5.7 kcal/mol 미만이어야 한다.
- [0104] 셋째, 2차 구조는 적어도 11개의 쌍을 이루지 않는(unpaired) 염기를 함유하여야 한다.
- [0105] 넷째, 풀 내에서 동일한 2차 구조의 개수는 150을 초과해서는 안 된다.
- [0106] 전술한 4개의 제약 조건은 공지된 모든 RNA 압타머-단백질 복합체의 분석에서 유도된다. 도 4에 도시한 바와 같이, 공지된 구조를 가지는 대부분의 RNA 압타머의 2차 구조는 땀글링 말단 염기를 가지거나 가지지 않으면서 적어도 3개의 염기 쌍을 가지는 스템(stem)으로 종결된다.

- [0107] 자유 에너지의 제약 조건은 Chushak 등(Chushak and Stone, 2009, In silico selection of RNA aptamers. Nucleic Acids Res. 37(12), e87)이 수행한 연구에서 얻어졌다. 연구자들은 10,000개의 동일 길이의 무작위 서열의 평균 자유 에너지를 사용하여 Z-스코어(Z-score)와 표준편차를 계산하였다. 공지된 RNA 압타머의 자유 에너지의 Z-스코어는 -0.9 내지 -4.3 사이의 범위이기 때문에, 자유 에너지의 컷-오프 값(cut-off value)은 Z-스코어 -1에 대응하는 -5.7 kcal/mol로 설정하였다.
- [0108] 예를 들어, 도 4에 도시한 바와 같이, 표적 단백질 분자에 결합하는 공지된 RNA 압타머의 구조에서, 단백질에 결합하는 뉴클레오타이드는 RNA 2차 구조의 이중 가닥 부분보다는 단일 가닥 부분에서 빈번하게 관찰된다(Chushak and Stone, 2009; Carothers et al., 2004, Informational complexity and functional activity of RNA structures. J. Am. Chem. Soc. 126(16), 5130-5137).
- [0109] RNA 압타머의 길이를 고려하여, 본 실시예에서 적어도 11개의 쌍을 이루지 않는 뉴클레오타이드를 가지는 후보 압타머를 선별하였다(3번째 제약 조건).
- [0110] 최초에 생성된 6×10^6 개의 무작위 RNA 서열에 의해 형성된 많은 2차 구조는 상호 유사하다. 일부 2차 구조는 10,000 회 이상 관찰된 반면, 다른 2차 구조는 1회만 관찰되었다. 잠재적인 RNA 압타머의 다양성을 유지하기 위하여, 동일한 2차 구조의 개수를 150으로 제한하였다(4번째 제약 조건).
- [0111] 전술한 4개의 구조적 제약 조건을 적용한 뒤에, 최초 6×10^6 개의 무작위 서열로부터 38,327개의 RNA 서열만이 남았다. 각각의 단백질 표적 분자에 대하여, RNA-단백질 서열의 쌍이 전술한 RF 모델에 대해 특징 벡터에서 암호화되었다. 상기 RF 모델은 35개의 서브-트리(subtrees)로 이루어져 있고, 각각의 서브-트리는 입력 특징 벡터에 투표되었다(voted). 전술한 4가지 제약 조건에 의해 선별된(screened) 모든 RNA 서열에 대하여, RNA 서열 및 단백질의 아미노산 서열을 혼련할 때 적용한 RF 모델로 연산하여, 포지티브 투표(positive votes)의 비율로부터 확률 추정(probability estimate)을 얻었다. 자유 에너지 및 확률과 관련하여 제약 조건에서 선별된 38,327개의 RNA 서열에 순위를 부여하고, 최고의 확률을 가지는 서열 중에서도 가장 낮은 자유 에너지를 가지는 탑 10 순위를 가지는 후보 RNA 압타머를 선별하였다.
- [0112] [교차 검증 및 독립 테스트]
- [0113] 상기에서 선별된 후보 RNA를 예측, 생성하기 위한 학습 모델을 10배 교차 검증(10-fold cross validation), 하나 남기기(leave-one-out) 교차 검증 및 독립 테스트를 수행하여 평가하였다. 교차 검증 및 독립 테스트는 전술한 식 (1) 내지 (6)으로 정의되는 민감도(sensitivity, Sn), 특이도(specificity, Sp), 정확도(accuracy, Acc), 양성예측도(positive predictive value, PPV), 음성예측도(negative predictive value, NPV) 및 매튜 상관관계수(MCC)를 사용하여 평가하였다. 표 1은 10배 본 실시예에 따라 표적 단백질 분자와 결합하는 후보 RNA 압타머를 구축, 생성하는 것과 관련한 10배 교차 검증 및 하나 남기기 교차 검증의 결과를 나타낸다. 표 1에 나타난 바와 같이, 2개의 교차 검증에서 모두 우수한 성능을 보여주었으며, 특히 하나 남기기 교차 검증 수행에 비하여 10배 교차 검증에서 보다 양호한 성능을 보여주었다.

표 1

[0114] 후보 RNA 압타머 구축 모델의 검증

검증	Sn	Sp	Acc	PPV	NPV	MCC
10배 교차	95.10%	98.69%	97.79%	96.04%	98.37%	0.941
LOO	94.12%	98.04%	97.06%	94.12%	98.04%	0.922

- [0115] 한편, 본 실시예에 따라 구축된 RF 모델에 대한 독립 테스트에서, 다른 선행 연구(Li et al., 2014, Prediction of Aptamer-Target Interacting Pairs with Pseudo-Amino Acid Composition. PLoS One 9(1), e86729; Zhang et al., Prediction of aptamer-protein interacting pairs using ensemble classifier in combination with various protein sequence attributes. BMC Bioinformatics 17, 225)의 benchmark 데이터 세트로부터 단백질에 결합하는 RNA 압타머가 포지티브 인스턴스로 사용되었다. 선행연구는 동일한 benchmark 데이터 세트를 사용하였다. 선행연구는 포지티브 데이터 세트에서 압타머와 단백질 표적에 대하여 무작위로 쌍을 형성하여 네거티브 인스턴스를 생성하였으나, 선행연구에서 단백질 파트너가 포지티브 데이터 세트 내의 단백질 표적 중 하나로 변환하면 네거티브 인스턴스가 포지티브 인스턴스가 되기 때문에 선행연구의 네거티브 인스턴스를 사용하지는 않았다. 독립 테스트와 관련한 네거티브 인스턴스에 대하여, 무작위 RNA 서열을 생성하고, 이들을 PDB에서 얻어진 35개

의 단백질-RNA 복합체에서 단백질 사슬 중 하나와 쌍을 형성하였다. 독립 테스트 수행 결과를 하기 표 2에 나타낸다.

표 2

후보 RNA 압타머 구축 모델의 독립 테스트

방법	Sn	Sp	Acc	PPV	NPV	MCC
Li et al.	48.3%	87.1%	77.4%	-	-	0.372
Zhang et al.	73.8%	71.3%	71.9%	-	-	0.388
실시예	76.8%	66.1%	71.4%	69.4%	74.0%	0.431

표 2에 나타난 바와 같이, 본 실시예에 따라 표적 단백질 분자와 결합하는 후보 RNA 압타머를 구축하기 위한 실시예의 RF 모델은 대체적으로 선행논문에 비하여 우수한 성능을 보여주었다. 선행논문과 본 실시예에서 사용한 네거티브 인스턴스의 차이로 인하여, 민감도가 비교를 위하여 가장 중요하고 공정한 평가 척도인 것으로 보인다.

[결론]

컴퓨터 연산을 이용하여 표적 단백질 분자와 결합하는 RNA 압타머를 종래의 방법은 표적 단백질 분자에 대한 신규 RNA 압타머를 발굴하는데 사용될 수 없다. 종래 기술에서 채택된 모델은 특정 표적으로만 제한되기 때문에, 주어진 한 쌍의 단백질과 RNA 서열이 상호작용하는지를 결정하는 분류 모델이기 때문이다. 본 실시예에서는 상호작용하는 RNA와 단백질의 다양한 특징과, RNA 2차 구조에 대한 구조적 제한 조건을 사용하여 표적 단백질 분자에 대한 잠재적인 후보 RNA 압타머를 구축하기 위한 컴퓨터 모델을 제안한다. RNA와 단백질 서열의 다양한 특징을 이용하여 랜덤 포레스트 모델을 구축하여, 서열을 훈련, 학습시켰다. 무작위 RNA 서열을 적절한 기준에 따라 필터링 하고, 구축된 랜덤 포레스트 모델을 이용하여 표적 단백질 분자에 대한 후보 RNA 압타머를 예측, 구축할 수 있다. 구축된 후보 RNA 압타머는 교차 검증 및 독립 테스트에서도 우수한 성능을 보여주었다. 따라서, 본 실시예의 모델을 적용하여 표적 단백질 분자와 결합하는 초기 핵산 서열의 풀(pool)의 크기를 크게 줄일 수 있기 때문에, 시험관내 실험에서 요구되는 막대한 시간과 비용을 절감할 수 있을 것으로 기대된다.

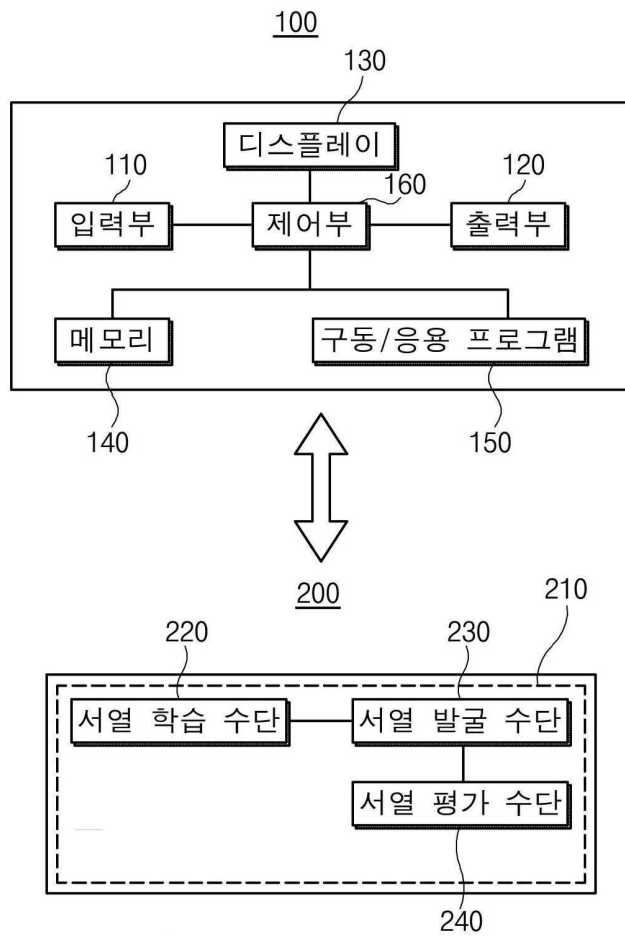
상기에서는 본 발명의 예시적인 실시형태 및 실시예에 기초하여 본 발명을 설명하였으나, 본 발명이 상기 실시형태 및 실시예에 기재된 기술사상으로 한정되는 것은 아니다. 오히려 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자라면 전술한 실시형태 및 실시예를 토대로 다양한 변형과 변경을 용이하게 추고할 수 있다. 하지만, 이러한 변형과 변경은 모두 본 발명의 권리범위에 속한다는 점은, 첨부하는 청구범위에서 분명하다.

부호의 설명

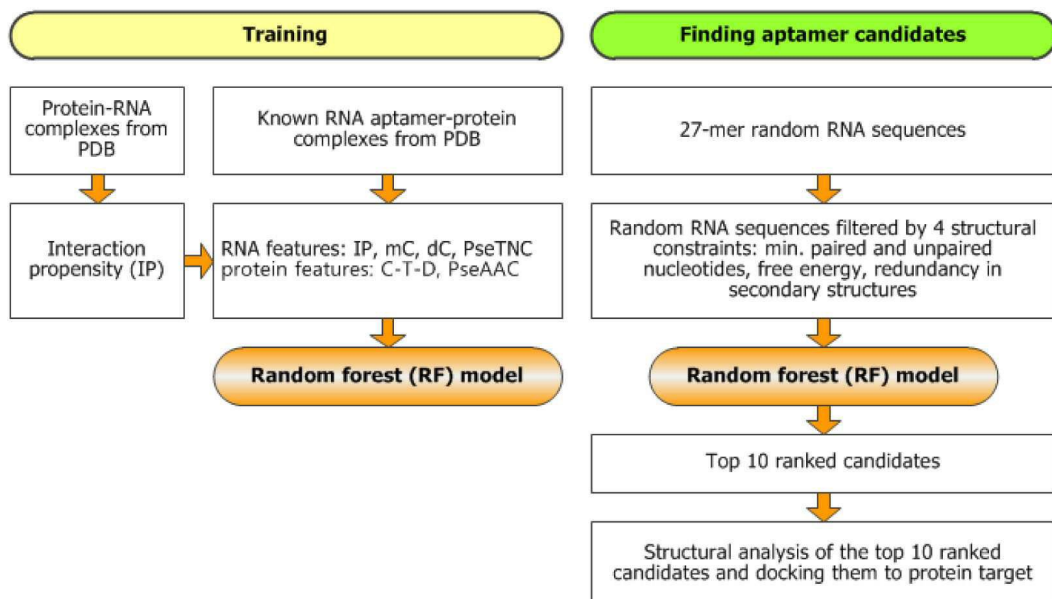
- | | |
|-------------------------|-----------------|
| 100: 컴퓨터 | 110: 입력부 |
| 120: 출력부 | 130: 디스플레이 |
| 140: 메모리 | 150: 구동/응용 프로그램 |
| 160: 제어부 | |
| 200: 기록 매체 | |
| 210: 후보 RNA 압타머 생성 프로그램 | |
| 220: 서열 학습 수단 | 230: 서열 발굴 수단 |
| 240: 서열 평가 수단 | |

도면

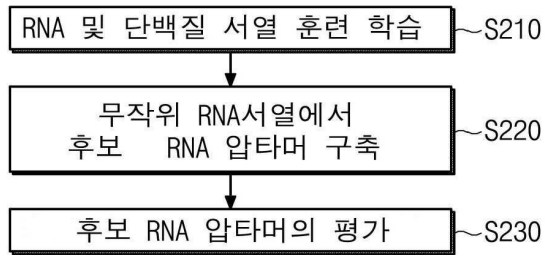
도면1



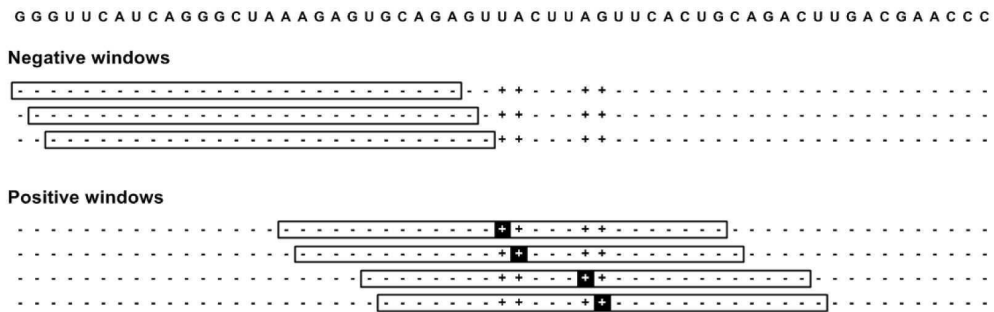
도면2



도면3



도면4



도면5

