(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0064172
(43) 공개일자 2023년05월10일

(51) 국제특허분류(Int. Cl.)
G16B 35/00 (2019.01) *C12Q 1/6806* (2018.01)
C12Q 1/6886 (2018.01) *G16B 30/10* (2019.01)
G16B 40/20 (2019.01) *G16B 5/00* (2019.01)
(52) CPC특허분류
G16B 35/00 (2019.02)
C12Q 1/6806 (2018.05)
(21) 출원번호 10-2021-0149466
(22) 출원일자 2021년11월03일
심사청구일자 없음

(71) 출원인
주식회사 지씨지놈
경기도 용인시 기흥구 이현로30번길 107 (보정동)
(72) 발명자
조은해
경기도 용인시 기흥구 이현로 30번길 107 (보정동)
이태림
경기도 용인시 기흥구 이현로 30번길 107 (보정동)
(74) 대리인
이처영, 장제환

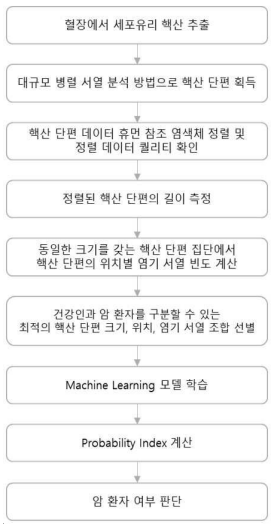
전체 청구항 수 : 총 13 항

(54) 발명의 명칭 **세포유리 핵산단편 위치별 서열 빈도 및 크기를 이용한 암 진단 방법**

(57) 요 약

본 발명은 세포유리 핵산단편 말단 서열 빈도 및 크기를 이용한 암 진단 및 암 종 예측방법에 관한 것으로, 보다 구체적으로는 생체시료에서 핵산을 추출하여, 서열정보를 획득하여 정렬한 리드를 기반으로 핵산단편의 말단 서열 빈도와 핵산단편의 크기를 도출한 다음, 이를 벡터화된 데이터로 생성한 후, 학습된 인공지능 모델에 입력하여 계산된 값을 분석하는 방법을 이용한 암 진단 및 암 종 예측방법에 관한 것이다. 본 발명에 따른 세포유리 핵산단편 말단 서열 빈도 및 크기를 이용한 암 진단 및 암 종 예측방법은 벡터화된 데이터를 생성하여 AI 알고리즘을 이용하여 분석하기 때문에 리드 커버리지가 낮더라도 높은 민감도와 정확도를 나타내어 유용하다.

대 표 도 - 도1



(52) CPC특허분류
C12Q 1/6886 (2022.01)
G16B 30/10 (2019.02)
G16B 40/20 (2019.02)
G16B 5/00 (2019.02)

명 세 서

청구범위

청구항 1

다음의 단계를 포함하는 무세포 핵산을 이용한 암 진단을 위한 정보의제공방법:

(a) 생체시료에서 핵산을 추출하여 서열정보를 획득하는 단계;

(b) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계;

(c) 상기 정렬된 서열정보(reads)를 이용하여 핵산단편(fragments)의 위치별 서열 상대 빈도 및 핵산단편의 크기를 도출하는 단계; 및

(d) 도출된 서열 상대 빈도 및 크기 정보를 암을 진단하도록 학습된 인공지능 모델에 입력하여 분석한 출력 결과값과 기준값(cut-off value)을 비교하여 암 유무를 판정하는 단계에 있어서,

상기 인공지능 모델은 핵산단편의 위치별 서열 상대 빈도 및 핵산단편의 크기 정보를 기반으로 정상 샘플과 암 샘플을 구별하도록 학습된 것을 특징으로 함.

청구항 2

제1항에 있어서, 상기 (a) 단계는 다음의 단계를 포함하는 방법으로 수행되는 것을 특징으로 하는 암 진단을 위한 정보의 제공방법:

(a-i) 혈액, 정액, 질 세포, 모발, 타액, 소변, 구강세포, 태반세포 또는 태아세포를 포함하는 양수, 조직세포 또는 이의 혼합물에서 핵산을 수득하는 단계;

(a-ii) 채취된 핵산에서 솔팅-아웃 방법(salting-out method), 컬럼 크로마토그래피 방법(column chromatography method) 또는 비드 방법(beads method)을 사용하여 단백질, 지방, 및 기타 잔여물을 제거하고 정제된 핵산을 수득하는 단계;

(a-iii) 정제된 핵산 또는 효소적 절단, 분쇄, 수압 절단 방법(hydroshear method)으로 무작위 단편화(random fragmentation)된 핵산에 대하여, 싱글 엔드 시퀀싱(single-end sequencing) 또는 페어 엔드 시퀀싱(pair-end sequencing) 라이브러리(library)를 제작하는 단계;

(a-iv) 제작된 라이브러리를 차세대 유전자서열검사기(next-generation sequencer)에 반응시키는 단계; 및

(a-v) 차세대 유전자서열검사기에서 핵산의 서열정보(reads)를 획득하는 단계.

청구항 3

제1항에 있어서, 상기 (c) 단계의 핵산단편의 크기는 127-129bp, 137-139bp, 148-150bp, 156-158bp 및 181-183bp로 구성된 군에서 선택되는 것을 특징으로 하는 암 진단을 위한 정보의 제공방법.

청구항 4

제1항에 있어서, 상기 (c) 단계의 핵산단편의 위치별 서열 상대 빈도는 동일한 크기의 핵산단편에서, 각각의 위치에서 검출되는 A, T, G, C 염기를 가지는 핵산단편의 수를 전체 핵산 단편 수로 정규화한 값인 것을 특징으로 하는 암 진단을 위한 정보의 제공방법.

청구항 5

제4항에 있어서, 상기 (c) 단계의 핵산단편의 위치는 핵산단편의 5' 말단에서 1 내지 10개 염기인 것을 특징으로 하는 암 진단을 위한 정보의 제공방법.

청구항 6

제4항에 있어서, 상기 (c) 단계의 핵산단편의 위치별 서열 상대 빈도는 핵산단편의 위치는 핵산단편의 5' 말단에서 1 내지 5개 위치에서는 A, T, G 및 C 염기의 빈도이며, 6 내지 10개 위치에서는 A 염기의 빈도인 것을 특징으로 하는 암 진단을 위한 정보의 제공방법.

청구항 7

제1항에 있어서, 상기 (c) 단계의 핵산단편(fragments)의 위치별 서열 상대 빈도 및 핵산단편의 크기는 표 3에 기재된 것에서 선택되는 어느 하나 이상인 것을 특징으로 하는 암 진단을 위한 정보의 제공방법.

청구항 8

제1항에 있어서, 상기 (d) 단계의 인공지능 모델은 AdaBoost, Random forest, Catboost, Light Gradient Boosting Model 및 XGBoost로 구성된 군에서 선택되는 것을 특징으로 하는 암 진단을 위한 정보의 제공방법.

청구항 9

제8항에 있어서, 상기 인공지능 모델이 XGBoost이고, binary classification 을 학습할 경우, 손실함수는 하기 수식 1로 표시되는 것을 특징으로 하는 암 진단을 위한 정보의 제공방법:

수식 1:

$$\text{loss(model(x),y)} = -\frac{1}{n}[\sum_{i=1}^{n}(y_i \log(model(x_i)) + (1-y_i)\log(1-model(x_i)))]$$

$$\text{model}(x_i) = i\text{번째 input에 XGboost model output}$$

y = 실제 label 값

n = input data 수

청구항 10

제1항에 있어서, 상기 (e) 단계의 인공지능 모델이 입력된 서열 상대 빈도 및 크기 정보를 분석하여 출력하는 결과값은 XPI(XGBoost Probability Index)값인 것을 특징으로 하는 암 진단을 위한 정보의 제공방법.

청구항 11

제1항에 있어서, 상기 (d) 단계의 기준값은 0.5이며, 0.5 이상일 경우, 암 인 것으로 판정하는 것을 특징으로 하는 암 진단을 위한 정보의 제공방법.

청구항 12

생체시료에서 핵산을 추출하여 서열정보를 해독하는 해독부;

해독된 서열을 표준 염색체 서열 데이터베이스에 정렬하는 정렬부;

정렬된 서열 기반의 핵산단편의 위치별 서열 상대 빈도 및 핵산단편의 크기를 도출하는 핵산단편 분석부; 및

도출된 핵산단편의 위치별 서열 상대 빈도 및 핵산단편의 크기 정보를 학습된 인공지능 모델에 입력하여 분석하고, 기준값과 비교하여 암 유무를 판정하는 암 진단부;

를 포함하는 암 진단 장치.

청구항 13

컴퓨터 판독 가능한 저장 매체로서, 암 진단을 위한 정보를 제공하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하되,

(a) 생체시료에서 핵산을 추출하여 서열정보를 획득하는 단계;

(b) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계;

(c) 상기 정렬된 서열정보(reads)를 이용하여 핵산단편(fragments)의 위치별 서열 상대 빈도 및 핵산단편의 크기를 도출하는 단계; 및

(d) 도출된 서열 상대 빈도 및 크기 정보를 암을 진단하도록 학습된 인공지능 모델에 입력하여 분석한 출력 결과값과 기준값(cut-off value)을 비교하여 암 유무를 판정하는 단계에 있어서,

상기 (d) 단계의 인공지능 모델은 핵산단편의 위치별 서열 상대 빈도 및 핵산단편의 크기 정보를 기반으로 정상 샘플과 암 샘플을 구별하도록 학습된 것을 특징으로 하는 단계를 통하여, 암 진단을 위한 정보를 제공하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하는 컴퓨터 판독 가능한 저장 매체.

발명의 설명

기 술 분 야

[0001] 본 발명은 세포유리 핵산단편 위치별 서열 상대 빈도 및 크기를 이용한 암 진단 방법에 관한 것으로, 보다 구체적으로는 생체시료에서 핵산을 추출하여, 서열정보를 획득하여 정렬한 리드를 기반으로 핵산단편의 위치별 서열 상대 빈도와 핵산단편의 크기를 도출한 다음, 이를 학습된 인공지능 모델에 입력하여 계산된 값을 분석하는 방법을 이용한 암 진단 방법에 관한 것이다.

배 경 기 술

[0003] 임상에서의 암 진단은 통상적으로 병력 조사, 물리적 검사 및 임상적 평가 후 조직 생검(tissue biopsy)을 수행하여 확인하고 있다. 임상 실험에 의한 암 진단은 암 세포의 수가 10억 개 이상이고 암의 직경이 1cm 이상일 경우에만 가능하다. 이 경우, 암 세포는 이미 전이능력을 가지고 있으며, 적어도 이들 중 반은 이미 전이된 상태이다. 또한, 조직생검은 침습적이어서 환자에게 상당한 불편함을 주고, 암 환자를 치료하다 보면 조직생검을 수행할 수 없는 경우도 자주 있다는 문제점이 있다. 이외에, 암 스크리닝에 있어서 암으로부터 직접 또는 간접적으로 생산되는 물질을 모니터링하기 위한 종양 마커가 사용되고 있지만, 암이 존재하는 경우에도 종양 마커 스크리닝 결과 반 이상이 정상으로 나타나고, 암이 없는 경우에도 자주 양성으로 나타나기 때문에, 그 정확성에 한계가 있다.

[0004] 이와 같은 통상적인 암 진단 방법의 문제점을 보완할 만한 비교적 간편하고 비침습적이며 높은 민감도 및 특이도를 가진 암 진단 방법의 요구에 따라, 최근 암의 진단, 추적 검사로 환자의 체액을 활용하는 액상생검(liquid biopsy)이 많이 이용되고 있다. 액상생검은 비침습적(non-invasive)인 방법으로, 기존의 침습적인 진단 및 검사

방법의 대안으로 주목 받고 있는 진단기술이다.

[0005] 최근에는 액상생검에서 획득한 세포 유리 DNA (cell free DNA)을 이용하여 암 진단 및 암 종 감별을 수행하는 방법이 개발되고 있으며(US 10975431, Zhou, Xionghui et al., bioRxiv, 2020.07.16.201350), 특히, 세포 유리 핵산 말단 서열의 모티프 빈도 정보를 분석하여 암 진단, 산전진단 또는 장기이식 모니터링에 이용하는 방법이 알려져 있다(WO 2020-125709, Peiyong Jiang et al., cancer discovery, Vol. 10, 2020, pp. 664-673).

[0006] 한편, Gradient Boosting Algorithm (GBM)은 회귀분석 또는 분류 분석을 수행할 수 있는 예측모형이며 예측모형의 앙상블 방법론 중 부스팅 계열에 속하는 알고리즘이다. Gradient Boosting Algorithm은 Tabular format 데이터 (엑셀형태와 같이 X-Y Grid로 되어있는 데이터)에 대한 예측에서 엄청난 성능을 보여주고, 머신러닝 알고리즘 중에서도 가장 예측 성능이 높다고 알려진 알고리즘이다.

[0008] 이러한 Gradient Boosting Algorithm을 이용하여 바이오 분야에 활용하는 다양한 문헌(Daping Yu et al., Thoracic Cancer Vol. 11, pp. 95-102. 2020, KR 10-2061800, KR 10- 2108050, KR 10-2021-0081547)이 존재하고 있으나, 혈액 내 무세포 DNA(cell-free DNA, cfDNA)의 서열분석 정보를 기반으로 GBM을 통해 암을 진단하는 방법에 대해서는 연구가 부족한 실정이다.

[0009] 이에, 본 발명자들은 상기 문제점들을 해결하고, 높은 민감도와 정확도의인공지능 기반 암 진단방법을 개발하기 위해 예의 노력한 결과, 무세포 핵산단편의 위치별 서열 상대 빈도와 핵산단편의 크기 정보를 기반으로 최적의 서열 상대 빈도 및 크기 조합을 선별하고, 이를 학습된 인공지능 모델로 분석할 경우, 높은 민감도와 정확도로 암 진단을 수행할 수 있다는 것을 확인하고, 본 발명을 완성하였다.

발명의 내용

해결하려는 과제

[0011] 본 발명의 목적은 세포유리 핵산단편의 위치별 서열 상대 빈도 및 크기를 이용한 암 진단방법을 제공하는 것이다.

[0012] 본 발명의 다른 목적은 세포유리 핵산단편 위치별 서열 상대 빈도 및 크기를 이용한 암 진단 장치를 제공하는 것이다.

[0013] 본 발명의 또 다른 목적은 상기 방법으로 암 진단을 수행하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하는 컴퓨터 판독 가능한 저장 매체를 제공하는 것이다.

과제의 해결 수단

[0015] 상기 목적을 달성하기 위하여, 본 발명은 (a) 생체시료에서 핵산을 추출하여 서열정보를 획득하는 단계; (b) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계; (c) 상기 정렬된 서열정보(reads)를 이용하여 핵산단편(fragments)의 위치별 서열 상대 빈도 및 핵산단편의 크기를 도출하는 단계; 및 (d) 도출된 서열 상대 빈도 및 크기 정보를 암을 진단하도록 학습된 인공지능 모델에 입력하여 분석한 출력 결과값과 기준값(cut-off value)을 비교하여 암 유무를 판정하는 단계에 있어서, 상기 인공지능 모델은 핵산단편의 위치별 서열 상대 빈도 및 핵산단편의 크기 정보를 기반으로 정상 샘플과 암 샘플을 구별하도록 학습된 것을 특징으로 하는 무세포 핵산을 이용한 암 진단을 위한 정보의 제공방법을 제공한다.

[0016] 본 발명은 또한, 생체시료에서 핵산을 추출하여 서열정보를 해독하는 해독부; 해독된 서열을 표준 염색체 서열 데이터베이스에 정렬하는 정렬부; 정렬된 서열 기반의 핵산단편의 위치별 서열 상대 빈도 및 핵산단편의 크기를 도출하는 핵산단편 분석부; 및 도출된 핵산단편의 위치별 서열 상대 빈도 및 핵산단편의 크기 정보를 학습된 인공지능 모델에 입력하여 분석하고, 기준값과 비교하여 암 유무를 판정하는 암 진단부; 를 포함하는 암 진단 장치를 제공한다.

[0017] 본 발명은 또한, 컴퓨터 판독 가능한 저장 매체로서, 암 진단을 위한 정보를 제공하는 프로세서에 의해 실행되

도록 구성되는 명령을 포함하되, (a) 생체시료에서 핵산을 추출하여 서열정보를 획득하는 단계; (b) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계; (c) 상기 정렬된 서열정보(reads)를 이용하여 핵산단편(fragments)의 위치별 서열 상대 빈도 및 핵산단편의 크기를 도출하는 단계; 및 (d) 도출된 서열 상대 빈도 및 크기 정보를 암을 진단하도록 학습된 인공지능 모델에 입력하여 분석한 출력 결과값과 기준값(cut-off value)을 비교하여 암 유무를 판정하는 단계에 있어서, 상기 (d) 단계의 인공지능 모델은 핵산단편의 위치별 서열 상대 빈도 및 핵산단편의 크기 정보를 기반으로 정상 샘플과 암 샘플을 구별하도록 학습된 것을 특징으로 하는 단계를 통하여, 암 진단을 위한 정보를 제공하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하는 컴퓨터 판독 가능한 저장 매체를 제공한다.

발명의 효과

[0019] 본 발명에 따른 세포유리 핵산단편 위치별 서열 상대 빈도 및 크기를 이용한 암 진단 방법은 최적의 핵산단편 위치별 서열 상대 빈도 및 크기 정보를 수득하여 AI 알고리즘을 이용하여 분석하기 때문에 리드 커버리지가 낮더라도 높은 민감도와 정확도를 나타내어 유용하다.

도면의 간단한 설명

[0021] 도 1은 본 발명의 세포유리 핵산단편 위치별 서열 상대 빈도 및 크기를 이용한 암 진단 방법을 수행하기 위한 전체 흐름도이다.

도 2는 본 발명의 일 실시예에서 건강인과 암 환자 사이에서 크기별로 상대 빈도가 통계적으로 유의미하게 차이가 있는 핵산단편 크기를 선별하는 과정의 예시이다.

도 3은 본 발명의 일 실시예에서 확인한 핵산단편들의 크기별 상대 빈도의 통계값과 선별한 핵산단편들의 크기 분포를 확인한 그래프이다.

도 4는 본 발명의 일 실시예에서 제작한 FESS table을 heatmap 형식으로 시각화한 것이다.

도 5의 왼쪽 패널은 도 4의 점선으로 표시된 부분을 확대한 것이고, 오른쪽 두 패널은 위치별 염기 서열의 상대 빈도를 통계적으로 분석한 결과이다.

도 6은 본 발명의 일 실시예에서 선별한 핵산단편의 위치에서 A, T, G, C 각 염기 서열의 상대 빈도를 계산하여 각각의 염기 서열 사이의 유사성을 통계적으로 확인한 결과이다.

도 7의 (A)는 본 발명의 일 실시예에서 구축한 머신러닝 모델의 성능을 Accuracy와 AUC로 확인한 결과이며, (B)는 혼동행렬(confusion matrix)이다.

도 8은 본 발명의 일 실시예에서 구축한 머신러닝 모델에서 예측한 건강인 및 신경모세포종 환자의 확률값이 실제 환자와 얼마나 일치하는 지를 머신러닝 모델이 출력한 XPI 값의 분포를 통해 확인한 결과이다.

도 9는 본 발명의 일 실시예에서 확인한 핵산단편들의 크기별 상대 빈도의 통계값과 선별한 핵산단편들의 크기 분포를 서로 다른 위치와 염기에서 확인한 그래프이다.

도 10은 본 발명의 일 실시예에서 선별한 feature의 중요도에 따라 소수의 feature로 구축한 머신러닝 모델의 성능을 확인한 결과로서, 윗 패널은 정확도(Accuracy)이고, 아래 패널은 AUC(Area Under Curve)이다.

발명을 실시하기 위한 구체적인 내용

[0022] 다른 식으로 정의되지 않는 한, 본 명세서에서 사용된 모든 기술적 및 과학적 용어들은 본 발명이 속하는 기술분야에서 숙련된 전문가에 의해서 통상적으로 이해되는 것과 동일한 의미를 갖는다. 일반적으로 본 명세서에서 사용된 명명법 및 이하에 기술하는 실험 방법은 본 기술 분야에서 잘 알려져 있고 통상적으로 사용되는 것이다.

[0023] 본 발명에서는, 샘플에서 획득한 서열 분석 데이터를 참조 유전체에 정렬한 다음, 정렬된 서열정보를 기반으로 핵산단편의 위치별 서열 상대 빈도 및 핵산단편의 크기를 도출하고, 상기 도출된 핵산단편의 위치별 서열 상대 빈도 및 핵산단편의 크기 정보를 학습된 인공지능 모델 입력한 다음, XPI값을 계산하여 분석할 경우, 높은 민감도와 정확도로 암 진단을 수행할 수 있다는 것을 확인하고자 하였다.

[0024] 즉, 본 발명의 일 실시예에서는, 혈액에서 추출한 DNA를 시퀀싱 한 뒤, 참조 염색체에 정렬한 다음, 이를 이용하여 핵산단편의 위치별 서열 상대 빈도 및 핵산단편의 크기를 도출하고, 최적의 핵산단편의 위치별 서열 상대 빈도 및 핵산단편의 크기 조합을 도출한 다음, 이를 딥러닝 모델에 학습시켜 XPI 값을 계산하였으며, 이를 기준값과 비교하여 암 진단을 수행하는 방법을 개발하였다(도 1)

[0025] 따라서, 본 발명은 일관점에서,

[0026] 다음의 단계를 포함하는 무세포 핵산을 이용한 암 진단을 위한 정보의제공방법에 관한 것이다:

[0027] (a) 생체시료에서 핵산을 추출하여 서열정보를 획득하는 단계;

[0028] (b) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계;

[0029] (c) 상기 정렬된 서열정보(reads)를 이용하여 핵산단편(fragments)의 위치별 서열 상대 빈도 및 핵산단편의 크기를 도출하는 단계; 및

[0030] (d) 도출된 서열 상대 빈도 및 크기 정보를 암을 진단하도록 학습된 인공지능 모델에 입력하여 분석한 출력 결과값과 기준값(cut-off value)을 비교하여 암 유무를 판정하는 단계에 있어서,

[0031] 상기 인공지능 모델은 핵산단편의 위치별 서열 상대 빈도 및 핵산단편의 크기 정보를 기반으로 정상 샘플과 암 샘플을 구별하도록 학습된 것을 특징으로 함.

[0033] 본 발명에 있어서, 상기 핵산 단편은 생체시료에서 추출한 핵산의 조각이면 제한없이 이용할 수 있으며, 바람직하게는 세포 유리 핵산 또는 세포 내 핵산의 조각일 수 있으나, 이에 한정되는 것은 아니다.

[0034] 본 발명에 있어서, 상기 핵산 단편은 통상의 기술자에게 알려진 모든 방법으로 얻을 수 있으며, 바람직하게는 직접 서열분석하거나, 차세대 염기서열 분석을 통해 서열분석하거나 또는 비특이적 전장 유전체 증폭(non-specific whole genome amplification)을 통해 서열분석하여 얻거나, 프로브 기반 서열분석을 통해 얻을 수 있으나, 이에 한정되는 것은 아니다.

[0036] 본 발명에서 상기 암은 고형암 또는 혈액암일 수 있으며, 바람직하게는 비호지킨 림프종 (non-Hodgkin lymphoma), 호지킨 림프종 (non-Hodgkin lymphoma), 급성 골수성 백혈병 (acute-myeloid leukemia), 급성 림프구성 백혈병 (acute-lymphoid leukemia), 다발성 골수종 (multiple myeloma), 경부암 (head and neck cancer), 폐암, 교모세포종 (glioblastoma), 대장/직장암, 췌장암, 유방암, 난소암, 흑색종 (melanoma), 전립선암, 간암, 갑상선암, 위암, 담낭암, 담도암, 방광암, 소장암, 자궁경부암, 원발부위불명암, 신장암, 식도암, 신경모세포종 및 중피종 (mesothelioma)으로 구성된 군에서 선택될 수 있으며, 더욱 바람직하게는 신경모세포종일 수 있으나, 이에 한정되는 것은 아니다.

[0038] 본 발명에 있어서,

[0039] 상기 (a) 단계는

[0040] (a-i) 혈액, 정액, 질 세포, 모발, 타액, 소변, 구강세포, 태반세포 또는 태아세포를 포함하는 양수, 조직세포 또는 이의 혼합물에서 핵산을 수득하는 단계;

[0041] (a-ii) 채취된 핵산에서 솔팅-아웃 방법(salting-out method), 컬럼 크로마토그래피 방법(column chromatography method) 또는 비드 방법(beads method)을 사용하여 단백질, 지방, 및 기타 잔여물을 제거하고 정제된 핵산을 수득하는 단계;

[0042] (a-iii) 정제된 핵산 또는 효소적 절단, 분쇄, 수압 절단 방법(hydroshear method)으로 무작위 단편화(random fragmentation)된 핵산에 대하여, 싱글 엔드 시퀀싱(single-end sequencing) 또는 페어 엔드 시퀀싱(pair-end sequencing) 라이브러리(library)를 제작하는 단계;

[0043] (a-iv) 제작된 라이브러리를 차세대 유전자서열검사기(next-generation sequencer)에 반응시키는 단계; 및

[0044] (a-v) 차세대 유전자서열검사기에서 핵산의 서열정보(reads)를 획득하는 단계;

[0045] 를 포함하는 것을 특징으로 할 수 있다.

[0046] 본 발명에서, 상기 (a) 단계의 서열정보를 획득하는 단계는 분리된 무세포 DNA를 1백만 내지 1억 리드 깊이로 전장 유전체 시퀀싱을 통해 획득하는 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.

[0048] 본 발명에 있어서, 상기 생체시료는 개체로부터 얻어지거나 개체로부터 유래된 임의의 물질, 생물학적 체액, 조직 또는 세포를 의미하는 것으로, 예를 들면, 전혈(whole blood), 백혈구(leukocytes), 말초혈액 단핵 세포(peripheral blood mononuclear cells), 백혈구 연층(buffy coat), (혈장(plasma) 및 혈청(serum)을 포함하는) 혈액, 객담(sputum), 눈물(tears), 점액(mucus), 세비액(nasal washes), 비강 흡인물(nasal aspirate), 호흡(breath), 소변(urine), 정액(semen), 침(saliva), 복강 세척액(peritoneal washings), 골반 내 유체액(pelvic fluids), 낭종액(cystic fluid), 뇌척수막 액(meningeal fluid), 양수(amniotic fluid), 선액(glandular fluid), 췌장액(pancreatic fluid), 림프액(lymph fluid), 흉수(pleural fluid), 유두 흡인물(nipple aspirate), 기관지 흡인물(bronchial aspirate), 활액(synovial fluid), 관절 흡인물(joint aspirate), 기관 분비물(organ secretions), 세포(cell), 세포 추출물(cell extract), 정액, 모발, 타액, 소변, 구강세포, 태반세포, 뇌척수액(cerebrospinal fluid) 및 이의 혼합물을 포함할 수 있으나, 이에 한정되는 것은 아니다.

[0050] 본 발명에서 용어, ”참조집단”은 표준 염기서열 데이터베이스와 같이 비교할 수 있는 기준(reference) 집단으로, 현재 특정 질환 또는 병증이 없는 사람의 집단을 의미한다. 본 발명에 있어서, 상기 참조집단의 표준 염색체 서열 데이터베이스에서 표준 염기서열은 NCBI 등의 공공보건기관에 등록되어 있는 참조 염색체일 수 있다.

[0051] 본 발명에 있어서, 상기 (a) 단계의 핵산은 무세포 DNA 일 수 있으며, 보다 바람직하게는 순환종양세포 DNA(circulating tumor DNA) 일 수 있으나, 이에 한정되는 것은 아니다.

[0053] 본 발명에서, 상기 차세대 유전자서열검사기(next-generation sequencer)는 당업계에 공지된 임의의 시퀀싱 방법으로 사용될 수 있다. 선택 방법에 의해 분리된 핵산의 시퀀싱은 전형적으로는 차세대 시퀀싱(NGS)을 사용하여 수행된다. 차세대 시퀀싱은 개개의 핵산 분자 또는 고도로 유사한 방식으로 개개의 핵산 분자에 대해 클론으로 확장된 프록시 중 하나의 뉴클레오타이드 서열을 결정하는 임의의 시퀀싱 방법을 포함한다(예를 들어, 105개 이상의 분자가 동시에 시퀀싱된다). 일 실시형태에서, 라이브러리 내 핵산 종의 상대적 존재비는 시퀀싱 실험에 의해 만들어진 데이터에서 그것의 동족 서열의 상대적 발생 수를 계측함으로써 추정될 수 있다. 차세대 시퀀싱 방법은 당업계에 공지되어 있고, 예를 들어 본 명세서에 참조로서 포함된 문헌(Metzker, M. (2010) Nature Biotechnology Reviews 11:31-46)에 기재된다.

[0054] 일 실시형태에서, 차세대 시퀀싱은 개개의 핵산 분자의 뉴클레오타이드 서열을 결정하기 위해 한다(예를 들어, 헬리코스 바이오사이언스(Helicos BioSciences)의 헬리스코프 유전자 시퀀싱 시스템(HeliScope Gene Sequencing system) 및 퍼시픽바이오사이언스의 팩바이오 알에스 시스템(PacBio RS system)). 다른 실시형태에서, 시퀀싱, 예를 들어, 더 적지만 더 긴 리드를 만들어내는 다른 시퀀싱 방법보다 시퀀싱 단위 당 서열의 더 많은 염기를 만들어내는 대량병렬의 짧은-리드 시퀀싱(예를 들어, 캘리포니아주 샌디에고에 소재한 일루미나 인코포레이티드(Illumina Inc.) 솔렉사 시퀀서(Solexa sequencer)) 방법은 개개의 핵산 분자에 대해 클론으로 확장된 프록시의 뉴클레오타이드 서열을 결정한다(예를 들어, 캘리포니아주 샌디에고에 소재한 일루미나 인코포레이티드(Illumina Inc.) 솔렉사 시퀀서(Solexa sequencer); 454 라이프 사이언스(Life Sciences)(코네티컷주 브랜포드에 소재) 및 아이온 토렌트(Ion Torrent)). 차세대 시퀀싱을 위한 다른 방법 또는 기계는, 이하에 제한되는 것은 아니지만, 454 라이프 사이언스(Life Sciences)(코네티컷주 브랜포드에 소재), 어플라이드 바이오시스템스(캘리포니아주 포스터 시티에 소재; SOLiD 시퀀서), 헬리코스 바이오사이언스 코포레이션(매사추세츠주 캠브릿지에 소재) 및 에멀젼 및 마이크로 유동 시퀀싱 기법 나노 점적(예를 들어, 지누바이오(GnuBio) 점적)에 의해 제공된다.

[0055] 차세대 시퀀싱을 위한 플랫폼은, 이하에 제한되는 것은 아니지만, 로슈(Roche)/454의 게놈 시퀀서(Genome Sequencer: GS) FLX 시스템, 일루미나(Illumina)/솔렉사(Solexa) 게놈 분석기(Genome Analyzer: GA), 라이프(Life)/APG의 서포트 올리고(Support Oligonucleotide Ligation Detection: SOLiD) 시스템, 폴로네이터

(Polonator)의 G.007 시스템, 헬리코스 바이오사이언스의 헬리스코프 유전자 시퀀싱 시스템(Helicos BioSciences' HeliScope Gene Sequencing system), 옥스포드 나노포어 테크놀로지스(Oxford Nanopore Technologies)의 PromethION, GriION, MinION 시스템 및 퍼시픽 바이오사이언스(Pacific Biosciences)의 팩바이오알에스(PacBio RS) 시스템을 포함한다.

[0056] 본 발명에서, 상기 (b) 단계의 서열 정렬은 컴퓨터 알고리즘으로서 게놈에서 리드 서열(예를 들어, 차세대 시퀀싱으로부터의, 예를 들어 짧은-리드 서열)이 대부분 리드 서열과 기준 서열 사이의 유사성을 평가함으로써 유래될 가능성이 있는 경우로부터 동일성에 대해 사용되는 컴퓨터적 방법 또는 접근을 포함한다. 서열 정렬 문제에 다양한 알고리즘이 적용될 수 있다. 일부 알고리즘은 상대적으로 느리지만, 상대적으로 높은 특이성을 허용한다. 이들은, 예를 들어 역동적 프로그래밍-기반 알고리즘을 포함한다. 역동적 프로그래밍은 그것들이 더 간단한 단계로 나누어짐으로써 복잡한 문제를 해결하는 방법이다. 다른 접근은 상대적으로 더 효율적이지만, 전형적으로 철저하지 않다. 이는, 예를 들어 대량 데이터베이스 검색을 위해 설계된 휴리스틱(heuristic) 알고리즘 및 확률적(probabilistic) 방법을 포함한다.

[0057] 전형적으로, 정렬 과정에 두 단계가 있을 수 있다: 후보자 검사 및 서열 정렬. 후보자 검사는 가능한 정렬 위치의 더 짧은 열거에 대해 전체 게놈으로부터 서열 정렬을 위한 검색 공간을 감소시킨다. 용어가 시사하는 바와 같이 서열 정렬은 후보자 검사 단계에 제공된 서열을 갖는 서열을 정렬시키는 단계를 포함한다. 이는 광역 정렬(예를 들어, 니들만-분쉬(Needleman-Wunsch) 정렬) 또는 국소 정렬(예를 들어, 스미스-워터만 정렬)을 사용하여 수행될 수 있다.

[0058] 대부분의 속성 정렬 알고리즘은 색인 방법에 기반한 3가지 유형 중 하나를 특징으로 할 수 있다: 해쉬 테이블(예를 들어, BLAST, ELAND, SOAP), 접미사트리(예를 들어, Bowtie, BWA) 및 병합 정렬(예를 들어, 슬라이더(Slider))에 기반한 알고리즘. 짧은 리드 서열은 정렬을 위해 전형적으로 사용된다.

[0059] 본 발명에 있어서, 상기 (b) 단계의 정렬단계는 이에 제한되지는 않으나, BWA 알고리즘 및 Hg19 서열을 이용하여 수행되는 것일 수 있다.

[0060] 본 발명에 있어서, 상기 BWA 알고리즘은 BWA-ALN, BWA-SW 또는 Bowtie2 등이 포함될 수 있으나 이에 한정되는 것은 아니다.

[0062] 본 발명에 있어서, 상기 (b) 단계의 서열정보(reads)의 길이는, 5 내지 5000 bp이고, 사용하는 서열정보의 수는 5천 내지 500만개가 될 수 있으나, 이에 한정되는 것은 아니다.

[0064] 본 발명에 있어서, 상기 (c) 단계를 수행하기에 앞서 정렬된 핵산 단편의 정렬 일치도 점수(mapping quality score)가 기준값 이상인 리드를 선별하는 단계를 추가로 포함하는 것을 특징으로 할 수 있으며, 상기 기준값은 정렬된 핵산 단편의 퀄리티를 확인할 수 있는 값이면 제한없이 이용가능하여, 바람직하게는 50 내지 70점, 더욱 바람직하게는 60점인 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.

[0066] 본 발명에서, 상기 (c) 단계의 핵산단편의 크기는 핵산단편의 5' 말단에서 3' 말단까지의 염기 개수이다.

[0067] 본 발명에 있어서, 상기 (c) 단계의 핵산단편의 크기는 건강인과 암 환자를 구분할 수 있는 크기이면 제한없이 사용할 수 있고, 바람직하게는 90 내지 250bp일 수 있으며, 더욱 바람직하게는 127-129bp, 137-139bp, 148-150bp, 156-158bp 및 181-183bp로 구성된 군에서 선택될 수 있으나, 이에 한정되는 것은 아니다.

[0068] 예를 들어, 하기와 같이 페어드-엔드 시퀀싱에 의해 서열 분석된 핵산단편이 있을 시,

[0069] Forward strand: 5`-**TACA**GACTTTGGAAT-3` (서열번호 1)

[0070] Reverse strand: 3`-ATGACTGAAAC**CTTA**-5` (서열번호 2)

[0071] Forward strand 5` 말단에서부터 3' 말단까지의 염기 개수인 15가 상기 핵산단편의 크기 값이 된다.

[0073] 본 발명에 있어서, 상기 (c) 단계의 핵산단편의 위치별 서열 상대 빈도는 동일한 크기의 핵산단편에서, 각각의

위치에서 검출되는 A, T, G, C 염기를 가지는 핵산단편의 수를 전체 핵산 단편 수로 정규화한 값인 것을 특징으로 할 수 있다.

[0074] 본 발명에 있어서, 상기 (c) 단계의 핵산단편의 위치는 핵산단편의 5' 말단에서 1 내지 10개 염기인 것을 특징으로 할 수 있다.

[0075] 본 발명에 있어서, 상기 (c) 단계의 핵산단편의 위치별 서열 상대 빈도는 핵산단편의 위치는 핵산단편의 5' 말단에서 1 내지 5개 위치에서는 A, T, G 및 C 염기의 빈도이며, 6 내지 10개 위치에서는 A 염기의 빈도인 것을 특징으로 할 수 있다.

[0076] 본 발명에 있어서, 상기 (c) 단계의 핵산단편(fragments)의 위치별 서열 상대 빈도 및 핵산단편의 크기는 표 3에 기재된 것에서 선택되는 하나 이상인 것을 특징으로 할 수 있고, 바람직하게는 표 7에 기재된 것에서 Top 1 내지 Top 5까지의 핵산단편의 위치별 서열 상대 빈도 및 핵산단편의 크기일 수 있으며, 더욱 바람직하게는 표 7에 기재된 것에서 Top 50까지의 핵산단편의 위치별 서열 상대 빈도 및 핵산단편의 크기일 수 있고, 가장 바람직하게는 Top 375까지의 핵산단편의 위치별 서열 상대 빈도 및 핵산단편의 크기일 수 있다.

표 3

[0077] Feature List

#	Feature	#	Feature	#	Feature
1	Size127_For1_A	126	Size139_For1_A	251	Size157_For1_A
2	Size127_For1_T	127	Size139_For1_T	252	Size157_For1_T
3	Size127_For1_G	128	Size139_For1_G	253	Size157_For1_G
4	Size127_For1_C	129	Size139_For1_C	254	Size157_For1_C
5	Size127_For2_A	130	Size139_For2_A	255	Size157_For2_A
6	Size127_For2_T	131	Size139_For2_T	256	Size157_For2_T
7	Size127_For2_G	132	Size139_For2_G	257	Size157_For2_G
8	Size127_For2_C	133	Size139_For2_C	258	Size157_For2_C
9	Size127_For3_A	134	Size139_For3_A	259	Size157_For3_A
10	Size127_For3_T	135	Size139_For3_T	260	Size157_For3_T
11	Size127_For3_G	136	Size139_For3_G	261	Size157_For3_G
12	Size127_For3_C	137	Size139_For3_C	262	Size157_For3_C
13	Size127_For4_A	138	Size139_For4_A	263	Size157_For4_A
14	Size127_For4_T	139	Size139_For4_T	264	Size157_For4_T
15	Size127_For4_G	140	Size139_For4_G	265	Size157_For4_G
16	Size127_For4_C	141	Size139_For4_C	266	Size157_For4_C
17	Size127_For5_A	142	Size139_For5_A	267	Size157_For5_A
18	Size127_For5_T	143	Size139_For5_T	268	Size157_For5_T
19	Size127_For5_G	144	Size139_For5_G	269	Size157_For5_G
20	Size127_For5_C	145	Size139_For5_C	270	Size157_For5_C
21	Size127_For6_A	146	Size139_For6_A	271	Size157_For6_A
22	Size127_For7_A	147	Size139_For7_A	272	Size157_For7_A
23	Size127_For8_A	148	Size139_For8_A	273	Size157_For8_A
24	Size127_For9_A	149	Size139_For9_A	274	Size157_For9_A
25	Size127_For10_A	150	Size139_For10_A	275	Size157_For10_A
26	Size128_For1_A	151	Size148_For1_A	276	Size158_For1_A
27	Size128_For1_T	152	Size148_For1_T	277	Size158_For1_T
28	Size128_For1_G	153	Size148_For1_G	278	Size158_For1_G
29	Size128_For1_C	154	Size148_For1_C	279	Size158_For1_C
30	Size128_For2_A	155	Size148_For2_A	280	Size158_For2_A
31	Size128_For2_T	156	Size148_For2_T	281	Size158_For2_T
32	Size128_For2_G	157	Size148_For2_G	282	Size158_For2_G
33	Size128_For2_C	158	Size148_For2_C	283	Size158_For2_C
34	Size128_For3_A	159	Size148_For3_A	284	Size158_For3_A
35	Size128_For3_T	160	Size148_For3_T	285	Size158_For3_T
36	Size128_For3_G	161	Size148_For3_G	286	Size158_For3_G
37	Size128_For3_C	162	Size148_For3_C	287	Size158_For3_C
38	Size128_For4_A	163	Size148_For4_A	288	Size158_For4_A

39	Size128_For4_T	164	Size148_For4_T	289	Size158_For4_T
40	Size128_For4_G	165	Size148_For4_G	290	Size158_For4_G
41	Size128_For4_C	166	Size148_For4_C	291	Size158_For4_C
42	Size128_For5_A	167	Size148_For5_A	292	Size158_For5_A
43	Size128_For5_T	168	Size148_For5_T	293	Size158_For5_T
44	Size128_For5_G	169	Size148_For5_G	294	Size158_For5_G
45	Size128_For5_C	170	Size148_For5_C	295	Size158_For5_C
46	Size128_For6_A	171	Size148_For6_A	296	Size158_For6_A
47	Size128_For7_A	172	Size148_For7_A	297	Size158_For7_A
48	Size128_For8_A	173	Size148_For8_A	298	Size158_For8_A
49	Size128_For9_A	174	Size148_For9_A	299	Size158_For9_A
50	Size128_For10_A	175	Size148_For10_A	300	Size158_For10_A
51	Size129_For1_A	176	Size149_For1_A	301	Size181_For1_A
52	Size129_For1_T	177	Size149_For1_T	302	Size181_For1_T
53	Size129_For1_G	178	Size149_For1_G	303	Size181_For1_G
54	Size129_For1_C	179	Size149_For1_C	304	Size181_For1_C
55	Size129_For2_A	180	Size149_For2_A	305	Size181_For2_A
56	Size129_For2_T	181	Size149_For2_T	306	Size181_For2_T
57	Size129_For2_G	182	Size149_For2_G	307	Size181_For2_G
58	Size129_For2_C	183	Size149_For2_C	308	Size181_For2_C
59	Size129_For3_A	184	Size149_For3_A	309	Size181_For3_A
60	Size129_For3_T	185	Size149_For3_T	310	Size181_For3_T
61	Size129_For3_G	186	Size149_For3_G	311	Size181_For3_G
62	Size129_For3_C	187	Size149_For3_C	312	Size181_For3_C
63	Size129_For4_A	188	Size149_For4_A	313	Size181_For4_A
64	Size129_For4_T	189	Size149_For4_T	314	Size181_For4_T
65	Size129_For4_G	190	Size149_For4_G	315	Size181_For4_G
66	Size129_For4_C	191	Size149_For4_C	316	Size181_For4_C
67	Size129_For5_A	192	Size149_For5_A	317	Size181_For5_A
68	Size129_For5_T	193	Size149_For5_T	318	Size181_For5_T
69	Size129_For5_G	194	Size149_For5_G	319	Size181_For5_G
70	Size129_For5_C	195	Size149_For5_C	320	Size181_For5_C
71	Size129_For6_A	196	Size149_For6_A	321	Size181_For6_A
72	Size129_For7_A	197	Size149_For7_A	322	Size181_For7_A
73	Size129_For8_A	198	Size149_For8_A	323	Size181_For8_A
74	Size129_For9_A	199	Size149_For9_A	324	Size181_For9_A
75	Size129_For10_A	200	Size149_For10_A	325	Size181_For10_A
76	Size137_For1_A	201	Size150_For1_A	326	Size182_For1_A
77	Size137_For1_T	202	Size150_For1_T	327	Size182_For1_T
78	Size137_For1_G	203	Size150_For1_G	328	Size182_For1_G
79	Size137_For1_C	204	Size150_For1_C	329	Size182_For1_C
80	Size137_For2_A	205	Size150_For2_A	330	Size182_For2_A
81	Size137_For2_T	206	Size150_For2_T	331	Size182_For2_T
82	Size137_For2_G	207	Size150_For2_G	332	Size182_For2_G
83	Size137_For2_C	208	Size150_For2_C	333	Size182_For2_C
84	Size137_For3_A	209	Size150_For3_A	334	Size182_For3_A
85	Size137_For3_T	210	Size150_For3_T	335	Size182_For3_T
86	Size137_For3_G	211	Size150_For3_G	336	Size182_For3_G
87	Size137_For3_C	212	Size150_For3_C	337	Size182_For3_C
88	Size137_For4_A	213	Size150_For4_A	338	Size182_For4_A
89	Size137_For4_T	214	Size150_For4_T	339	Size182_For4_T
90	Size137_For4_G	215	Size150_For4_G	340	Size182_For4_G
91	Size137_For4_C	216	Size150_For4_C	341	Size182_For4_C
92	Size137_For5_A	217	Size150_For5_A	342	Size182_For5_A
93	Size137_For5_T	218	Size150_For5_T	343	Size182_For5_T
94	Size137_For5_G	219	Size150_For5_G	344	Size182_For5_G
95	Size137_For5_C	220	Size150_For5_C	345	Size182_For5_C
96	Size137_For6_A	221	Size150_For6_A	346	Size182_For6_A

97	Size137_For7_A	222	Size150_For7_A	347	Size182_For7_A
98	Size137_For8_A	223	Size150_For8_A	348	Size182_For8_A
99	Size137_For9_A	224	Size150_For9_A	349	Size182_For9_A
100	Size137_For10_A	225	Size150_For10_A	350	Size182_For10_A
101	Size138_For1_A	226	Size156_For1_A	351	Size183_For1_A
102	Size138_For1_T	227	Size156_For1_T	352	Size183_For1_T
103	Size138_For1_G	228	Size156_For1_G	353	Size183_For1_G
104	Size138_For1_C	229	Size156_For1_C	354	Size183_For1_C
105	Size138_For2_A	230	Size156_For2_A	355	Size183_For2_A
106	Size138_For2_T	231	Size156_For2_T	356	Size183_For2_T
107	Size138_For2_G	232	Size156_For2_G	357	Size183_For2_G
108	Size138_For2_C	233	Size156_For2_C	358	Size183_For2_C
109	Size138_For3_A	234	Size156_For3_A	359	Size183_For3_A
110	Size138_For3_T	235	Size156_For3_T	360	Size183_For3_T
111	Size138_For3_G	236	Size156_For3_G	361	Size183_For3_G
112	Size138_For3_C	237	Size156_For3_C	362	Size183_For3_C
113	Size138_For4_A	238	Size156_For4_A	363	Size183_For4_A
114	Size138_For4_T	239	Size156_For4_T	364	Size183_For4_T
115	Size138_For4_G	240	Size156_For4_G	365	Size183_For4_G
116	Size138_For4_C	241	Size156_For4_C	366	Size183_For4_C
117	Size138_For5_A	242	Size156_For5_A	367	Size183_For5_A
118	Size138_For5_T	243	Size156_For5_T	368	Size183_For5_T
119	Size138_For5_G	244	Size156_For5_G	369	Size183_For5_G
120	Size138_For5_C	245	Size156_For5_C	370	Size183_For5_C
121	Size138_For6_A	246	Size156_For6_A	371	Size183_For6_A
122	Size138_For7_A	247	Size156_For7_A	372	Size183_For7_A
123	Size138_For8_A	248	Size156_For8_A	373	Size183_For8_A
124	Size138_For9_A	249	Size156_For9_A	374	Size183_For9_A
125	Size138_For10_A	250	Size156_For10_A	375	Size183_For10_A

[0078] 본 발명에서, 핵산단편의 위치는 핵산단편의 5' 말단을 기준으로 정의된다.

[0079] 예를 들어, 상기 서열번호 1의 forward strand의 5' 말단에서부터 핵산단편의 위치는 For1, For2, …For 15의 값을 가질 수 있고, reverse strand도 마찬가지이다. 상기 서열번호 1의 For1 값은 T이고, reverse strand의 Rev1 값은 A이다.

[0081] 본 발명에서, 핵산단편의 위치별 염기서열의 빈도는 하기와 같은 과정으로 계산할 수 있다.

[0082] a) 전체 핵산 단편을 동일한 크기를 갖는 핵산 단편 집단으로 구분하는 단계;

[0083] b) 각 그룹 내에서 핵산 단편 위치 별 A, T, G, C 염기의 개수를 계수하는 단계; 및

[0084] c) 수식 2를 이용하여 핵산 단편 위치 별 염기의 개수를 정규화하는 단계.

[0085] 수식 2: Size120_For1_A = $\frac{\#\ of\ fragments\ with\ size=120,position=For1,base=A}{\#\ of\ all\ reads\ sequenced}$

[0087] 본 발명에 있어서, 상기 수식 2의 size와 position 및 염기는 정규화하고자 하는 크기, 위치 및 염기에 따라 달라진 다는 것은 당업자에게 자명하다.

[0088] 본 발명에 있어서, 상기 (d) 단계의 인공지능 모델은 건강인과 암 환자를 구별할 수 있도록 학습할 수 있는 모델이면 제한없이 사용가능하며, 바람직하게는 머신러닝 모델인 것을 특징으로 할 수 있다.

[0090] 본 발명에 있어서, 상기 인공지능 모델은 AdaBoost, Random forest, Catboost, Light Gradient Boosting Model

및 XGBoost로 구성된 군에서 선택되는 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.

[0092] 본 발명에 있어서, 상기 인공지능 모델이 XGBoost이고, binary classification 을 학습할 경우, 손실함수는 하기 수식 1로 표시되는 것을 특징으로 할 수 있다.

[0093] 수식 1: Binary classification

[0094] $$\text{loss}(\text{model}(x),y) = -\frac{1}{n}[\sum_{i=1}^{n}(y_i \log(model(x_i)) + (1-y_i)\log(1-model(x_i)))]$$

[0095] $\text{model}(x_i) = i\text{번째 input에 XGboost model output}$

[0096] $y =$ 실제 label 값

[0097] $n =$ input data 수

[0099] 본 발명에서, 상기 binary classification은 인공지능 모델이 암 유무를 판별하도록 학습하는 것을 의미한다.

[0101] 본 발명에서, 상기 인공지능 모델이 XGBoost일 경우, 학습은 하기 단계를 포함하여 수행되는 것을 특징으로 할 수 있다:

[0102] i) 핵산단편의 위치별 서열 상대 빈도 및 크기 정보를 training(학습), validation(검증), test(성능평가) 데이터로 분류하는 단계;

[0103] 이 때, Training 데이터는 XGBoost 모델을 학습할 때 사용되고, Validation 데이터는 hyper-parameter tuning 검증에 사용되며, Test 데이터는 최적의 모델 생산 후, 성능 평가로 사용되는 것을 특징으로 함.

[0104] ii) Hyper-parameter tuning 및 학습 과정을 통해서 최적의 XGBoost 모델을 구축하는 단계;

[0105] iii) Hyper-parameter tuning을 통해서 얻어진 여러 모델의 성능을 validation data를 이용하여 비교하여, validation data 성능이 가장 좋은 모델을 최적의 모델로 결정하는 단계;

[0107] 본 발명에서, 상기 Hyper-parameter tuning 과정은 XGBoost 모델을 이루는 여러 parameter(learner tree의 최대 깊이, learner tree의 개수, learning rate 등) 값을 최적화 하는 과정으로 Hyper-parameter tuning 과정으로는 Bayesian optimization 및 grid search 기법을 사용하는 것을 특징으로 할 수 있다.

[0108] 본 발명에서, 상기 학습 과정은 정해진 hyper-parameter들을 이용하여 XGBoost 모델의 내부 parameter(weights)들을 최적화 시켜, Training loss 대비 validation loss가 증가하기 시작하면 모델이 과적합(Overfitting) 되었다 판단하고, 그전에 model 학습을 중단하는 것을 특징으로 할 수 있다.

[0110] 본 발명에 있어서, 상기 d) 단계에서 인공지능 모델이 입력된 핵산단편의 위치별 서열 상대 빈도 및 크기 정보로부터 분석한 결과값은 특정 score 또는 실수이면 제한없이 이용가능하며, 바람직하게는 XPI(XGBoost Probability Index) 값인 것을 특징으로 할 수 있으나 이에 한정되는 것은 아니다.

[0112] 본 발명에서, XGBoost Probability Index는 인공지능 모델의 output을 0 ~ 1 scale로 조정하여 확률값으로 표현한 값을 의미한다.

[0113] Binary classification일 경우에는 sigmoid function을 이용하여 암 일 경우 XPI 값이 1이 되게끔 학습을 하게 된다. 예를 들어, 신경모세포종 샘플과 정상 샘플이 입력되면, 신경모세포종 샘플의 XPI 값이 1에 가깝도록, 그

리고 정상 샘플은 0에 가깝도록 학습하는 것이다.

[0115] 본 발명에서, 상기 인공지능 모델은 학습할 때, 암이 있으면 output 결과가 1에 가깝게 학습하고, 암이 없으면 output 결과가 0에 가깝게 학습을 시켜서, 0.5를 기준으로 0.5 이상이면 암이 있다고 판단하고, 0.5 이하이면 암이 없다고 판단하여 performance 측정을 수행하였다(Training, validation, test accuracy).

[0116] 여기서, 0.5의 기준값은 언제든지 바뀔 수 있는 값이라는 것은 통상의 기술자에게 자명한 것이다. 예를 들어서 False positive(위양성)를 줄이고자 하면, 0.5보다 높은 기준값을 설정하여 암이 있다고 판단되는 기준을 엄격하게 가져 갈 수 있고, False Negative(위음성)를 줄이고자 하면 기준값을 더 낮게 측정하여 암이 있다고 판단되는 기준을 조금 더 약하게 가져 갈 수 있다.

[0117] 가장 바람직하게는 학습된 인공지능 모델을 이용하여 unseen data(학습에 training하지 않은 답을 알고 있는 data)를 적용시켜서, XPI값의 probability를 확인하여 기준값을 정할 수 있다.

[0119] 본 발명은 다른 관점에서, 생체시료에서 핵산을 추출하여 서열정보를 해독하는 해독부;

[0120] 해독된 서열을 표준 염색체 서열 데이터베이스에 정렬하는 정렬부;

[0121] 정렬된 서열 기반의 핵산단편의 위치별 서열 상대 빈도 및 핵산단편의 크기를 도출하는 핵산단편 분석부; 및

[0122] 도출된 핵산단편의 위치별 서열 상대 빈도 및 핵산단편의 크기 정보를 학습된 인공지능 모델에 입력하여 분석하고, 기준값과 비교하여 암 유무를 판정하는 암 진단부;

[0123] 를 포함하는 암 진단 장치에 관한 것이다.

[0125] 본 발명에서, 상기 해독부는 독립된 장치에서 추출된 핵산을 주입하는 핵산 주입부; 및 주입된 핵산의 서열정보를 분석하는 서열정보 분석부를 포함할 수 있으며, 바람직하게는 NGS 분석 장치일 수 있으나, 이에 한정되는 것은 아니다.

[0126] 본 발명에서, 상기 해독부는 독립된 장치에서 생성된 서열정보 데이터를 수신하여 해독하는 것을 특징으로 할 수 있다.

[0128] 본 발명은 또 다른 관점에서, 컴퓨터 판독 가능한 저장 매체로서, 암 진단을 위한 정보를 제공하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하되,

[0129] (a) 생체시료에서 핵산을 추출하여 서열정보를 획득하는 단계;

[0130] (b) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계;

[0131] (c) 상기 정렬된 서열정보(reads)를 이용하여 핵산단편(fragments)의 위치별 서열 상대 빈도 및 핵산단편의 크기를 도출하는 단계; 및

[0132] (d) 도출된 서열 상대 빈도 및 크기 정보를 암을 진단하도록 학습된 인공지능 모델에 입력하여 분석한 출력 결과값과 기준값(cut-off value)을 비교하여 암 유무를 판정하는 단계에 있어서,

[0133] 상기 (d) 단계의 인공지능 모델은 핵산단편의 위치별 서열 상대 빈도 및 핵산단편의 크기 정보를 기반으로 정상 샘플과 암 샘플을 구별하도록 학습된 것을 특징으로 하는 단계를 통하여, 암 진단을 위한 정보를 제공하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하는 컴퓨터 판독 가능한 저장 매체에 관한 것이다.

[0135] 다른 양태에서 본원에 따른 방법은 컴퓨터를 이용하여 구현될 수 있다. 일 구현예에서, 컴퓨터는 칩 세트에 연결된 하나 이상의 프로세서를 포함한다. 또한 칩 세트에는 메모리, 저장 장치, 키보드, 그래픽 어댑터(Graphics Adapter), 포인팅 장치(Pointing Device) 및 네트워크 어댑터(Network Adapter) 등이 연결되어 있다. 일 구현예에서, 상기 칩 세트의 성능은 메모리 컨트롤러 허브(Memory Controller Hub) 및 I/O 컨트롤러 허브에 의하여

가능하다. 다른 구현예에서, 상기 메모리는 칩 세트 대신에 프로세서에 직접 연결되어 사용될 수 있다. 저장 장치는 하드 드라이브, CD-ROM(Compact Disk Read-Only Memory), DVD 또는 기타 메모리 장치를 포함하는 데이터를 유지할 수 있는 임의의 장치이다. 메모리는 프로세서에 의하여 사용된 데이터 및 명령에 관여한다. 상기 포인팅 디바이스는 마우스, 트랙볼 (Track Ball) 또는 다른 유형의 포인팅 디바이스일 수 있고, 키보드와 조합하여 입력 데이터를 컴퓨터 시스템으로 전송하는데 사용된다. 상기 그래픽 어댑터는 디스플레이 상에서 이미지 및 다른 정보를 나타낸다. 상기 네트워크 어댑터는 근거리 또는 장거리 통신망으로 컴퓨터 시스템과 연결된다. 본원에 사용되는 컴퓨터는 하지만 위와 같은 구성으로 제한되는 것은 아니고, 일부 구성이 없거나, 추가의 구성을 포함 할 수 있으며, 또한 저장장치영역네트워크(Storage Area Network, SAN)의 일부일 수 있으며, 본원의 컴퓨터는 본원에 따른 방법의 수행을 위한 프로그램에 모듈의 실행에 적합하도록 구성될 수 있다.

[0137] 본원에서 모듈이라 함은, 본원에 따른 기술적 사상을 수행하기 위한 하드웨어 및 상기 하드웨어를 구동하기 위한 소프트웨어의 기능적, 구조적 결합을 의미할 수 있다. 예컨대, 상기 모듈은 소정의 코드와 상기 소정의 코드가 수행되기 위한 하드웨어 리소스(Resource)의 논리적인 단위를 의미할 수 있으며, 반드시 물리적으로 연결된 코드를 의미하거나, 한 종류의 하드웨어를 의미하는 것은 아님은 본원 기술분야의 당업자에게 자명한 것이다.

[0139] 본원에 따른 방법은 하드웨어, 펌웨어, 또는 소프트웨어 또는 이들의 조합으로 구현될 수 있다. 소트프웨어로 구현되는 경우 저장매체는 컴퓨터와 같은 장치에 의해 판독가능한 형태의 저장 또는 전달하는 임의의 매체를 포함한다. 예를 들면 컴퓨터 판독가능한 매체는 ROM(Read Only Memory); RAM(Random Access Memory); 자기디스크 저장 매체; 광저장 매체; 플래쉬 메모리 장치 및 기타 전기적, 광학적 또는 음향적 신호 전달 매체 등을 포함한다.

[0141] **실시예**

[0142] 이하, 실시예를 통하여 본 발명을 더욱 상세히 설명하고자 한다. 이들 실시예는 오로지 본 발명을 예시하기 위한 것으로서, 본 발명의 범위가 이들 실시예에 의해 제한되는 것으로 해석되지는 않는 것은 당업계에서 통상의 지식을 가진 자에게 있어서 자명할 것이다.

[0144] **실시예 1. 혈액에서 DNA를 추출하여, 차세대 염기서열 분석 수행**

[0145] 건강인 202명 및 신경모세포종 환자 61명의 혈액을 10mL씩 채취하여 EDTA Tube에 보관하였으며, 채취 후 2시간 이내에 1200g, 4℃ 15분의 조건으로 혈장 부분만 1차 원심분리한 다음, 1차 원심분리된 혈장을 16000g, 4℃ 10분의 조건으로 2차 원심분리하여 침전물을 제외한 혈장 상층액을 분리하였다. 분리된 혈장에 대해, Chemagic ccfNA 2K Kit (chemagen)을 사용하여 cell-free DNA를 추출하고, MGIEasy cell-free DNA library prep set kit 를 사용하여 library preparation 과정을 수행한 다음, DNBseq G400 장비 (MGI) 를 100 base Paired end 모드로 sequencing 하였다. 그 결과, 샘플 당 약 170 million 개의 reads가 생산되는 것을 확인 하였다.

[0147] **실시예 2. 최적의 핵산단편 위치별 서열 상대 빈도 및 핵산단편 크기 선별**

[0148] 2-1. 핵산단편 위치와 염기서열의 상대 빈도 정의 및 측정

[0149] 핵산단편의 위치는 핵산단편의 5' 말단을 기준으로 정의하였다.

[0150] 실시예 1에서 수득한 리드는 paired-end sequencing read이고, 100bp 길이이므로, forward strand는 5' 말단에서부터 For1, For2, …For 100까지의 위치를 설정하였고, Reverser strand에서도 5' 말단에서부터 Rev1, Rev2, …Rev 100까지의 위치를 설정하였다. 핵산단편의 조립은 bedtools 프로그램의 bamtobed -bedpe 옵션을 사용하였다.

[0151] 핵산단편의 위치별 염기서열의 상대 빈도를 구하는 과정을 간략히 설명하자면 먼저, 실시예 1에서 생산한 약 170M read 정도의 시퀀싱 데이터에서, 임의로 17M read를 선별하여 downsampling 한 다음, QC filtering를 수행하고, Size, position, base (ex, Size120_For1_A) 조합을 만족하는 fragment 수 계수한 뒤, 위의 3 QC

filtering 후에 남아 있는 전체 시퀀싱 read 수로 나누어 normalization을 수행한 것이다.

[0152] 보다 구체적으로는 하기의 방법으로 수행하였다.

[0153] 1. 전체 핵산 단편을 동일한 크기를 갖는 핵산 단편 집단으로 구분하였다. 예를 들어, 핵산 단편 크기가 101인 그룹, 150인 그룹, …200인 그룹 등.

[0154] 2. 각 그룹 내에서 핵산 단편 위치 별 A, T, G, C 염기의 개수를 계수하였다. 예를 들어, 핵산 단편의 크기가 120인 집단에서의 핵산 단편 위치 별 염기의 수를 계수하면 아래 표 1과 같이 정리할 수 있다.

표 1

[0155]

Size=120	For1	For2	For3	...	For100	Rev100	...	Rev3	Rev2	Rev1
A	5,683	6,999	6,694	...	6,998	6,429	...	6,807	6,942	5,619
T	4,680	4,422	8,283	...	7,825	7,986	...	8,329	4,438	4,716
G	4,194	5,555	2,730	...	3,970	3,952	...	2,772	5,609	4,111
C	8,566	6,153	5,413	...	4,336	4,768	...	5,219	6,142	8,676
N	12	6	15	...	6	-	...	8	4	13
Sum	23,135	23,135	23,135	...	23,135	23,135	...	23,135	23,135	23,135

[0156] 위의 표를 해석해 보면, 크기가 120이었던 핵산 단편은 총 23,135 개가 있었고, 그 중 For1 위치에 A, T, G, C 염기를 갖고 있던 핵산 단편이 각각 5,683개, 4,680개, 4,194개, 8,566개 있다는 것을 의미한다.

[0157] 3. 위의 과정으로 핵산 단편 위치 별 염기의 개수를 계수한 후, 시퀀싱 된 전체 리드 수 (핵산 단편 크기 구분 없이, 생산된 분석 대상의 모든 리드 수. 실시예 1에서는 15,063,130 개)로 나누어 수식 2로 정규화 (Normalization) 하여, 아래 표 2(FESS_Table_120)와 같이 상대 빈도를 계산한 FESS(Fragment End Seuqnece frequency and Size) table을 제작하였다.

[0158] 수식 2: Size120_For1_A = $$\frac{\#\ of\ fragments\ with\ size=120,position=For1,base=A}{\#\ of\ all\ reads\ sequenced}$$

표 2

[0159]

Size=120	For1	For2	For3	...	For100	Rev100	...	Rev3	Rev2	Rev1
A	0.000377	0.000465	0.000444	...	0.000465	0.000427	...	0.000452	0.000461	0.000373
T	0.000311	0.000294	0.000550	...	0.000519	0.000530	...	0.000553	0.000295	0.000313
G	0.000278	0.000369	0.000181	...	0.000264	0.000262	...	0.000184	0.000372	0.000273
C	0.000569	0.000408	0.000359	...	0.000288	0.000317	...	0.000346	0.000408	0.000576

[0161] 4. N (시퀀싱 에러, 낮은 퀄리티 등의 이유로 염기 서열 측정 불가했던 경우) 값의 상대 빈도는 계산하지 않았다.

[0163] 2-2. 최적의 핵산단편 크기 선별

[0164] 분석 대상 핵산 단편의 위치와 염기 서열을 (For1_A)로 고정하고 아래 분석을 진행하였다.

[0165] 1. 핵산 단편 크기를 1씩 변화시켜가면서 건강인과 신경모세포종 환자군 사이에서 (For1_A)의 상대 빈도 분포 차이가 있는지를 Kruskal-Wallis Test를 이용하여 통계적으로 확인하였다. 즉, 도 2에 기재된 바와 같이, 크기가 118인 핵산 단편 집단에서는 (For1_A)의 상대 빈도가 건강인보다 신경포세포종 환자군에서 통계적으로 유의미한 수준으로 높게 분포하는 것을 확인할 수 있다. 같은 방법으로, 크기가 168인 핵산 단편 집단에서는 (For1_A)의 상대 빈도가 두 집단에서 큰 차이 없이 분포하는 것을 확인할 수 있으며, 크기가 185인 핵산 단편 집단에서는 (For1_A)의 상대 빈도가 건강인 에서 신경모세포종 환자군보다 통계적으로 유의미한 수준으로 낮게 분포하는 것을 확인할 수 있다.

[0166] 2. 이러한 방법으로 으로 핵산 단편 크기를 101에서 200까지 변화시켜가면서 건강인과 신경모세포종 사이의 (For1_A) 상대 빈도 차이를 통계적으로(p-value) 확인하였다.

[0167] 그 결과, 도 3의 X축은 핵산 단편의 크기를, Y 축은 $-\log_{10}(p)$ 값을 나타내는데, Y 축 값이 클수록 건강인과 신경포세포종 환자 사이에서 차이가 크다는 것을 의미한다. 도 3에 기재된 바와 같이, 10 정도의 핵산 단편 크기를 주기로 하여 건강인과 신경모세포종 사이에 (For1_A) 빈도 차이가 크게 벌어지는 ($-\log_{10}(p)$ 값이 peak를 찍고 내려가는) 것을 확인하였다.

[0168] 또한, 이러한 패턴이 Training Dataset 뿐만 아니라, 독립된 Validation Dataset에서도 동일하게 반복되는 것으로 보아, 초록색으로 표시한 핵산 단편 크기들이 Training Dataset에 overfitting 된 우연한 패턴이 아닌 것을 확인할 수 있습니다.

[0169] 두 Dataset에서 공통적으로 $-\log_{10}(p)$ 값이 peak를 보이는 핵산 단편 크기를 선택하여(127~129, 137~139, 148~150, 156~158, 181~183), 총 15개의 핵산 단편 크기를 선별하였다.

[0170] 아울러, 다른 위치의 다른 염기에서도 유사한 패턴이 나타나는 것을 확인하였다(도 9).

[0172] 2-3. 최적의 핵산단편 위치 선별

[0173] 실시예 1에서 수득한 데이터는 100 PE 데이터이므로, 분석에 사용 가능한 핵산 단편 위치는 For1~100, 그리고 Rev1~100까지 총 200 가지이다.

[0174] 도 4는 표 2의 *FESS_Table_120*을 Heatmap 형식으로 시각화한 것으로, 점선으로 표시된 양 끝 쪽 (For1~10, Rev1~10) 일부에서만 위치에 따른 A, T, G, C 염기 서열의 상대 빈도 차이가 관찰되고, read의 뒷부분(~100)으로 갈수록 거의 비슷한 A, T, G, C 염기 서열의 상대 빈도가 반복되는 것을 확인할 수 있다.

[0175] 예를 들어, For1의 A, T, G, C 염기 서열 상대 빈도는 For2의 A, T, G, C 상대 빈도와 상당한 차이를 보이지만, For11의 A, T, G, C 염기 서열 상대 빈도와 For99의 A, T, G, C 염기 서열 상대 빈도, 그리고 For100의 A, T, G, C 염기 서열 상대 빈도는 큰 차이 없이 거의 유사한 것을 확인할 수 있다.

[0176] 따라서, 학습 모델의 성능향상을 위해 read 뒷부분 위치를 제외한 For1~10, Rev1~10 위치만 모델 학습 대상 feature로 선별하였다.

[0177] 추가적으로, 도 4의 점선으로 표시된 영역을 확대하면 도 5와 같은데(Rev1~10은 Rev10~1 순으로 역으로 정렬하였다), 가장 왼쪽 패널을 보면, Forward와 Reverse의 같은 위치에 있는 같은 서열의 상대 빈도가 서로 상당히 유사한 것을 확인할 수 있다.

[0178] 예를 들어, (For1_A와 Rev1_A), (For1_T와 Rev1_T), (For1_G와 Rev1_G), 그리고 (For1_C와 Rev1_C)가 서로 유사한 상대 빈도 값을 갖고, 같은 방법으로 (For2_A와 Rev2_A), (For2_T와 Rev2_T), (For2_G와 Rev2_G), 그리고 (For2_C와 Rev2_C)가 서로 유사한 상대 빈도 값을 갖는다.

[0179] 이러한 유사성을 건강인 집단에서의 Pearson's correlation coefficient로 측정해보면 도 5의 오른쪽 두 패널과 같다. 건강인 집단에서 측정된 For1_A의 상대 빈도 값들과 Rev1_A의 상대 빈도 값들 사이의 유사성, For1_T의 상대 빈도 값들과 Rev1_T의 상대 빈도 값들 사이의 유사성, For1_G의 상대 빈도 값들과 Rev1_G의 상대 빈도 값들 사이의 유사성, 및 For1_C의 상대 빈도 값들과 Rev1_C의 상대 빈도 값들 사이의 유사성은 모두 1인 것을 확인하였다.

[0180] 이러한 분석을 통해 핵산 단편의 Forward strand쪽 5` 말단 염기 서열의 상대 빈도와 Reverse strand쪽 5` 말단 염기 서열의 상대 빈도가 유사하다는 것을 확인할 수 있었고, 따라서, Rev1~10 위치를 제외한 For1~10 위치만 모델 학습 대상 feature로 선별하였다.

[0182] 2-4. 최적의 핵산단편 위치별 염기 서열 선별

[0183] 실시예 2-3에서 선별한 10 곳의 위치에서는 각각 A, T, G, C 네 종류 염기 서열의 상대 빈도를 계산할 수 있다. 예를 들어, For1 위치에서는 For1_A, For1_T, For1_G, For1_C의 상대 빈도를 계산할 수 있다. 모델 학습 대상 변수를 줄이기 위해, 같은 위치에 있는 염기 서열들 사이의 유사성을 확인하여 추가적인 선별을 진행하였다. 위치 별 염기 서열 선별은 건강인 집단에서 하기 방법으로 진행하였다.

[0184] 1. For1~10 각 위치에서 A, T, G, C 염기 서열의 상대 빈도를 계산하고,

[0185] 2. (For1_A와 For1_T), (For1_A와 For1_G), (For1_A와 For1_C), (For1_T와 For1_G), (For1_T와 For1_C), 그리고 (For1_G와 For1_C) 사이의 유사성을 Pearson's correlation coefficient로 측정하였다.

[0186] 그 결과, 도 6에 기재된 바와 같이, For1~5 위치에서는 A, T, G, C 네 종류 염기 서열의 상대 빈도 사이의 유사성이 낮은 것을 확인하였으며, For6~10 위치에서는 A, T, G, C 네 종류 염기 서열의 상대 빈도 사이의 유사성이 상당히 높은 것을 확인하였다.

[0187] 따라서, For1~5 위치에서는 A, T, G, C 네 종류의 염기 서열을 모두 선별하고, For6~10 위치에서는 A, T, C, G 중 대표 값으로 A 염기 서열만 선별하였다.

[0188] 결론적으로 최적의 핵산단편 크기 및 위치별 서열 상대 빈도는 하기와 같다:

[0189] 1) 핵산 단편 크기: 127, 128, 129, 137, 138, 139, 148, 149, 150, 156, 157, 158, 181, 182, 183. 총 15개.

[0190] 2) 핵산 단편 위치: For1~10. 총 10개.

[0191] 3) 핵산 단편 위치 별 염기 서열 조합

[0192] For1~5: A, T, G, C For6~10: A

[0193] 15개 크기 * 25개 위치_염기서열 = 375개 Features

[0194] 375개의 Feature 조합은 표 3에 기재하였다.

[0196] **실시예 3. 머신러닝 모델 구축 및 학습 과정**

[0197] 실시예 2에서 선별한 375개 Feature들의 상대 빈도 값을 인풋으로 하여 건강인, 신경모세포종 환자를 구분하는 머신러닝 모델을 학습하였다. 머신러닝 알고리즘은 XGBoost를 사용하였다.

[0198] 전체 샘플을 Training, Validation, Test 데이터 세트로 나누어 Training 데이터 세트는 모델 학습에, Validation 데이터 세트는 hyper-parameter tuning에, Test 데이터 세트는 최종 모델 성능 평가에 사용하였다. 각 세트 별 샘플 수는 아래와 같다.

표 4

[0199]

데이터 세트	건강인	신경모세포종 환자	합계
Train	99	30	129
Valid	42	13	55
Test	61	18	79
Total	202	61	263

[0200] Hyper-parameter tuning 과정은 XGBoost 모델을 이루는 여러 parameter(learner tree의 최대 깊이, learner tree의 개수, learning rate 등) 값을 최적화 하는 과정이다.

[0201] Hyper-parameter tuning 과정에는 Bayesian optimization 및 grid search 기법을 사용하였고, Training loss 대비 validation loss가 증가하기 시작하면 모델이 과적합(Overfitting) 되었다고 판단하여 model 학습을 중단하였다.

[0202] Hyper-parameter tuning을 통해서 수득한 여러 모델의 성능을 Validation 데이터 세트를 이용하여 비교하고, 이 중 Validation 데이터 세트 성능이 가장 좋은 모델을 최적의 모델이라 판단하고, Test 데이터 세트로 최종 성능 평가를 수행하였다.

[0203] 상기 과정을 거쳐서 만들어진 XGBoost 모델에 임의의 샘플에서 계산된 375개 feature의 상대 빈도 값 벡터를 인풋으로 넣어 주면, 해당 샘플의 건강인일 확률, 신경모세포종 환자일 확률이 계산되고, 이 확률 값을 XGBoost Probability Index (XPI)라 정의하였다.

[0204] 임의의 샘플에서 계산된 XPI 값이 0.5 초과이면 신경모세포종 환자로, 0.5 이하이면 건강인으로 판단하였다.

[0206] **실시예 4. 구축한 모델의 성능 확인**

[0207] 4-1 성능 확인

[0208] 실시예 3에서 구축한 머신러닝 모델에서 출력한 XPI 값의 성능을 테스트 하였다. 모든 샘플은 Train, Validation, Test 그룹으로 나눠 진행했고, Train 샘플을 이용하여 model을 구축한 다음 Validation 그룹 및 Test 그룹의 샘플을 이용해서, Train 샘플을 이용해 만든 모델의 성능을 확인하였다.

표 5

[0209]

	Accuracy	AUC
Train	1.000	1.000
Validation	0.945	0.952
Test	0.937	0.987

[0210] 그 결과, 표 5 및 도 7에 기재된 바와 같이, Accuracy 는 Train, Valid, Test 그룹에서 각각 1.000, 0.945, 0.937인 것을 확인하였고, ROC 분석 결과인 AUC 값은 Train, Valid, Test 그룹에서 각각 1.000, 0.952, 0.987인 것을 확인하였다

[0212] 4-2. XPI 분포 확인

[0213] 실시예 3에서 구축한 머신러닝 모델의 출력값인 XPI 값이 실제 환자와 얼마나 일치하는 지를 확인하였다. 도 8의 X 축은 실제 샘플의 그룹 (True label) 정보를 나타내고, Y 축은 왼쪽에서부터 순서대로 머신러닝 모델에서 계산한 건강인(Normal), 신경모세포종 환자(NBT)일 XPI 값을 나타낸다.

[0214] 그 결과, 도 8에 기재된 바와 같이 XPI 분포는 Train, Validation, Test 데이터 세트 모두에서 건강인 샘플들은 건강인일 확률이 가장 높게 분포하는 것을 확인하였으며, 신경모세포종 환자 샘플들은 간암 환자일 확률이 가장 높게 나타나는 것을 확인하였다.

[0216] **실시예 5. Feature 별 모델 성능 확인**

[0217] 5-1 Feature 별 중요도 도출

[0218] 실시예 2에서 선별한 feature를 이용하여 실시예 3에서 학습모델을 구축하였고, 각각의 feature를 사용하여 XGB 모델을 학습했을 때, 각 feature들의 importance 값은 하기 표 6과 같다.

표 6

[0219] Feature 별 Importance

Rank	Feature	Importance	Rank	Feature	Importance
1	Size_149_For_5_G	0.093	189	Size_138_For_5_A	0.000
2	Size_128_For_1_A	0.071	190	Size_138_For_5_T	0.000
3	Size_138_For_1_G	0.054	191	Size_138_For_5_G	0.000
4	Size_182_For_4_T	0.052	192	Size_138_For_5_C	0.000
5	Size_157_For_1_A	0.032	193	Size_138_For_6_A	0.000
6	Size_127_For_1_A	0.031	194	Size_138_For_7_A	0.000
7	Size_158_For_1_A	0.029	195	Size_138_For_8_A	0.000
8	Size_137_For_7_A	0.029	196	Size_138_For_9_A	0.000
9	Size_156_For_5_G	0.024	197	Size_138_For_10_A	0.000
10	Size_182_For_5_T	0.024	198	Size_139_For_1_A	0.000
11	Size_127_For_2_C	0.024	199	Size_139_For_1_T	0.000
12	Size_139_For_2_A	0.023	200	Size_139_For_1_G	0.000

13	Size_183_For_3_G	0.023	201	Size_139_For_1_C	0.000
14	Size_181_For_5_T	0.020	202	Size_139_For_2_T	0.000
15	Size_148_For_1_G	0.020	203	Size_139_For_2_G	0.000
16	Size_156_For_1_G	0.019	204	Size_139_For_2_C	0.000
17	Size_150_For_9_A	0.018	205	Size_139_For_3_A	0.000
18	Size_127_For_5_G	0.018	206	Size_139_For_3_T	0.000
19	Size_183_For_1_T	0.017	207	Size_139_For_3_G	0.000
20	Size_181_For_1_G	0.016	208	Size_139_For_4_A	0.000
21	Size_137_For_2_T	0.016	209	Size_139_For_4_T	0.000
22	Size_182_For_8_A	0.015	210	Size_139_For_4_G	0.000
23	Size_156_For_1_T	0.013	211	Size_139_For_4_C	0.000
24	Size_158_For_3_T	0.012	212	Size_139_For_5_A	0.000
25	Size_157_For_2_T	0.012	213	Size_139_For_5_T	0.000
26	Size_137_For_1_A	0.011	214	Size_139_For_5_G	0.000
27	Size_150_For_2_C	0.010	215	Size_139_For_5_C	0.000
28	Size_181_For_3_G	0.010	216	Size_139_For_6_A	0.000
29	Size_127_For_3_G	0.010	217	Size_139_For_7_A	0.000
30	Size_156_For_1_C	0.010	218	Size_139_For_8_A	0.000
31	Size_156_For_4_G	0.009	219	Size_139_For_9_A	0.000
32	Size_127_For_1_C	0.009	220	Size_139_For_10_A	0.000
33	Size_156_For_2_T	0.009	221	Size_148_For_1_A	0.000
34	Size_138_For_3_G	0.009	222	Size_148_For_1_T	0.000
35	Size_182_For_1_A	0.009	223	Size_148_For_1_C	0.000
36	Size_157_For_3_T	0.009	224	Size_148_For_2_A	0.000
37	Size_156_For_3_T	0.008	225	Size_148_For_2_T	0.000
38	Size_158_For_1_C	0.008	226	Size_148_For_2_G	0.000
39	Size_158_For_2_G	0.007	227	Size_148_For_3_A	0.000
40	Size_158_For_2_C	0.007	228	Size_148_For_3_T	0.000
41	Size_182_For_5_C	0.007	229	Size_148_For_3_G	0.000
42	Size_158_For_1_T	0.006	230	Size_148_For_3_C	0.000
43	Size_149_For_1_G	0.006	231	Size_148_For_4_A	0.000
44	Size_156_For_4_C	0.006	232	Size_148_For_4_T	0.000
45	Size_158_For_1_G	0.006	233	Size_148_For_4_C	0.000
46	Size_157_For_3_A	0.006	234	Size_148_For_5_A	0.000
47	Size_157_For_2_A	0.005	235	Size_148_For_5_T	0.000
48	Size_127_For_2_T	0.005	236	Size_148_For_5_G	0.000
49	Size_158_For_2_T	0.005	237	Size_148_For_5_C	0.000
50	Size_156_For_3_G	0.005	238	Size_148_For_6_A	0.000
51	Size_148_For_2_C	0.004	239	Size_148_For_7_A	0.000
52	Size_137_For_2_C	0.004	240	Size_148_For_8_A	0.000
53	Size_127_For_1_G	0.004	241	Size_148_For_9_A	0.000
54	Size_181_For_1_A	0.004	242	Size_148_For_10_A	0.000
55	Size_129_For_3_T	0.004	243	Size_149_For_1_A	0.000
56	Size_150_For_1_G	0.004	244	Size_149_For_1_C	0.000
57	Size_127_For_2_A	0.004	245	Size_149_For_2_A	0.000
58	Size_137_For_3_G	0.003	246	Size_149_For_2_T	0.000
59	Size_158_For_2_A	0.003	247	Size_149_For_2_G	0.000
60	Size_157_For_1_C	0.003	248	Size_149_For_2_C	0.000
61	Size_181_For_2_T	0.003	249	Size_149_For_3_A	0.000
62	Size_148_For_4_G	0.003	250	Size_149_For_3_T	0.000
63	Size_182_For_2_C	0.003	251	Size_149_For_3_G	0.000
64	Size_149_For_1_T	0.002	252	Size_149_For_3_C	0.000
65	Size_150_For_5_T	0.002	253	Size_149_For_4_A	0.000
66	Size_157_For_3_G	0.002	254	Size_149_For_4_T	0.000
67	Size_127_For_3_C	0.002	255	Size_149_For_4_G	0.000
68	Size_183_For_1_A	0.002	256	Size_149_For_4_C	0.000
69	Size_156_For_5_T	0.002	257	Size_149_For_5_A	0.000
70	Size_139_For_3_C	0.002	258	Size_149_For_5_T	0.000

71	Size_183_For_1_C	0.002	259	Size_149_For_5_C	0.000
72	Size_138_For_2_A	0.002	260	Size_149_For_6_A	0.000
73	Size_158_For_3_A	0.002	261	Size_149_For_7_A	0.000
74	Size_157_For_1_T	0.002	262	Size_149_For_8_A	0.000
75	Size_150_For_3_T	0.002	263	Size_149_For_9_A	0.000
76	Size_128_For_3_G	0.002	264	Size_149_For_10_A	0.000
77	Size_158_For_3_G	0.002	265	Size_150_For_1_A	0.000
78	Size_127_For_3_T	0.002	266	Size_150_For_1_C	0.000
79	Size_127_For_5_C	0.001	267	Size_150_For_2_A	0.000
80	Size_182_For_1_G	0.001	268	Size_150_For_2_T	0.000
81	Size_156_For_2_A	0.001	269	Size_150_For_2_G	0.000
82	Size_158_For_4_T	0.001	270	Size_150_For_3_G	0.000
83	Size_137_For_5_A	0.001	271	Size_150_For_3_C	0.000
84	Size_183_For_1_G	0.001	272	Size_150_For_4_A	0.000
85	Size_137_For_1_C	0.001	273	Size_150_For_4_T	0.000
86	Size_156_For_4_A	0.001	274	Size_150_For_4_G	0.000
87	Size_156_For_3_C	0.001	275	Size_150_For_4_C	0.000
88	Size_182_For_2_A	0.001	276	Size_150_For_5_A	0.000
89	Size_183_For_2_C	0.001	277	Size_150_For_5_G	0.000
90	Size_127_For_4_G	0.001	278	Size_150_For_5_C	0.000
91	Size_137_For_2_A	0.001	279	Size_150_For_6_A	0.000
92	Size_127_For_4_C	0.001	280	Size_150_For_7_A	0.000
93	Size_181_For_3_C	0.001	281	Size_150_For_8_A	0.000
94	Size_129_For_2_T	0.001	282	Size_150_For_10_A	0.000
95	Size_157_For_5_G	0.001	283	Size_156_For_1_A	0.000
96	Size_127_For_1_T	0.001	284	Size_156_For_2_G	0.000
97	Size_150_For_3_A	0.001	285	Size_156_For_2_C	0.000
98	Size_127_For_4_T	0.001	286	Size_156_For_3_A	0.000
99	Size_156_For_7_A	0.001	287	Size_156_For_4_T	0.000
100	Size_182_For_1_C	0.001	288	Size_156_For_5_A	0.000
101	Size_181_For_4_G	0.001	289	Size_156_For_5_C	0.000
102	Size_150_For_1_T	0.001	290	Size_156_For_6_A	0.000
103	Size_127_For_2_G	0.000	291	Size_156_For_8_A	0.000
104	Size_127_For_3_A	0.000	292	Size_156_For_9_A	0.000
105	Size_127_For_4_A	0.000	293	Size_156_For_10_A	0.000
106	Size_127_For_5_A	0.000	294	Size_157_For_1_G	0.000
107	Size_127_For_5_T	0.000	295	Size_157_For_2_G	0.000
108	Size_127_For_6_A	0.000	296	Size_157_For_2_C	0.000
109	Size_127_For_7_A	0.000	297	Size_157_For_3_C	0.000
110	Size_127_For_8_A	0.000	298	Size_157_For_4_A	0.000
111	Size_127_For_9_A	0.000	299	Size_157_For_4_T	0.000
112	Size_127_For_10_A	0.000	300	Size_157_For_4_G	0.000
113	Size_128_For_1_T	0.000	301	Size_157_For_4_C	0.000
114	Size_128_For_1_G	0.000	302	Size_157_For_5_A	0.000
115	Size_128_For_1_C	0.000	303	Size_157_For_5_T	0.000
116	Size_128_For_2_A	0.000	304	Size_157_For_5_C	0.000
117	Size_128_For_2_T	0.000	305	Size_157_For_6_A	0.000
118	Size_128_For_2_G	0.000	306	Size_157_For_7_A	0.000
119	Size_128_For_2_C	0.000	307	Size_157_For_8_A	0.000
120	Size_128_For_3_A	0.000	308	Size_157_For_9_A	0.000
121	Size_128_For_3_T	0.000	309	Size_157_For_10_A	0.000
122	Size_128_For_3_C	0.000	310	Size_158_For_3_C	0.000
123	Size_128_For_4_A	0.000	311	Size_158_For_4_A	0.000
124	Size_128_For_4_T	0.000	312	Size_158_For_4_G	0.000
125	Size_128_For_4_G	0.000	313	Size_158_For_4_C	0.000
126	Size_128_For_4_C	0.000	314	Size_158_For_5_A	0.000
127	Size_128_For_5_A	0.000	315	Size_158_For_5_T	0.000
128	Size_128_For_5_T	0.000	316	Size_158_For_5_G	0.000

129	Size_128_For_5_G	0.000	317	Size_158_For_5_C	0.000
130	Size_128_For_5_C	0.000	318	Size_158_For_6_A	0.000
131	Size_128_For_6_A	0.000	319	Size_158_For_7_A	0.000
132	Size_128_For_7_A	0.000	320	Size_158_For_8_A	0.000
133	Size_128_For_8_A	0.000	321	Size_158_For_9_A	0.000
134	Size_128_For_9_A	0.000	322	Size_158_For_10_A	0.000
135	Size_128_For_10_A	0.000	323	Size_181_For_1_T	0.000
136	Size_129_For_1_A	0.000	324	Size_181_For_1_C	0.000
137	Size_129_For_1_T	0.000	325	Size_181_For_2_A	0.000
138	Size_129_For_1_G	0.000	326	Size_181_For_2_G	0.000
139	Size_129_For_1_C	0.000	327	Size_181_For_2_C	0.000
140	Size_129_For_2_A	0.000	328	Size_181_For_3_A	0.000
141	Size_129_For_2_G	0.000	329	Size_181_For_3_T	0.000
142	Size_129_For_2_C	0.000	330	Size_181_For_4_A	0.000
143	Size_129_For_3_A	0.000	331	Size_181_For_4_T	0.000
144	Size_129_For_3_G	0.000	332	Size_181_For_4_C	0.000
145	Size_129_For_3_C	0.000	333	Size_181_For_5_A	0.000
146	Size_129_For_4_A	0.000	334	Size_181_For_5_G	0.000
147	Size_129_For_4_T	0.000	335	Size_181_For_5_C	0.000
148	Size_129_For_4_G	0.000	336	Size_181_For_6_A	0.000
149	Size_129_For_4_C	0.000	337	Size_181_For_7_A	0.000
150	Size_129_For_5_A	0.000	338	Size_181_For_8_A	0.000
151	Size_129_For_5_T	0.000	339	Size_181_For_9_A	0.000
152	Size_129_For_5_G	0.000	340	Size_181_For_10_A	0.000
153	Size_129_For_5_C	0.000	341	Size_182_For_1_T	0.000
154	Size_129_For_6_A	0.000	342	Size_182_For_2_T	0.000
155	Size_129_For_7_A	0.000	343	Size_182_For_2_G	0.000
156	Size_129_For_8_A	0.000	344	Size_182_For_3_A	0.000
157	Size_129_For_9_A	0.000	345	Size_182_For_3_T	0.000
158	Size_129_For_10_A	0.000	346	Size_182_For_3_G	0.000
159	Size_137_For_1_T	0.000	347	Size_182_For_3_C	0.000
160	Size_137_For_1_G	0.000	348	Size_182_For_4_A	0.000
161	Size_137_For_2_G	0.000	349	Size_182_For_4_G	0.000
162	Size_137_For_3_A	0.000	350	Size_182_For_4_C	0.000
163	Size_137_For_3_T	0.000	351	Size_182_For_5_A	0.000
164	Size_137_For_3_C	0.000	352	Size_182_For_5_G	0.000
165	Size_137_For_4_A	0.000	353	Size_182_For_6_A	0.000
166	Size_137_For_4_T	0.000	354	Size_182_For_7_A	0.000
167	Size_137_For_4_G	0.000	355	Size_182_For_9_A	0.000
168	Size_137_For_4_C	0.000	356	Size_182_For_10_A	0.000
169	Size_137_For_5_T	0.000	357	Size_183_For_2_A	0.000
170	Size_137_For_5_G	0.000	358	Size_183_For_2_T	0.000
171	Size_137_For_5_C	0.000	359	Size_183_For_2_G	0.000
172	Size_137_For_6_A	0.000	360	Size_183_For_3_A	0.000
173	Size_137_For_8_A	0.000	361	Size_183_For_3_T	0.000
174	Size_137_For_9_A	0.000	362	Size_183_For_3_C	0.000
175	Size_137_For_10_A	0.000	363	Size_183_For_4_A	0.000
176	Size_138_For_1_A	0.000	364	Size_183_For_4_T	0.000
177	Size_138_For_1_T	0.000	365	Size_183_For_4_G	0.000
178	Size_138_For_1_C	0.000	366	Size_183_For_4_C	0.000
179	Size_138_For_2_T	0.000	367	Size_183_For_5_A	0.000
180	Size_138_For_2_G	0.000	368	Size_183_For_5_T	0.000
181	Size_138_For_2_C	0.000	369	Size_183_For_5_G	0.000
182	Size_138_For_3_A	0.000	370	Size_183_For_5_C	0.000
183	Size_138_For_3_T	0.000	371	Size_183_For_6_A	0.000
184	Size_138_For_3_C	0.000	372	Size_183_For_7_A	0.000
185	Size_138_For_4_A	0.000	373	Size_183_For_8_A	0.000
186	Size_138_For_4_T	0.000	374	Size_183_For_9_A	0.000

187	Size_138_For_4_G	0.000	375	Size_183_For_10_A	0.000
188	Size_138_For_4_C	0.000			

[0220] 5-2. TopN feature 성능 확인

[0221] 실시예 3의 방법으로 상위 1번 feature 만을 사용하여 구축한 XGB 모델, 2번까지 사용한 모델, 3번, 4번, 5번, 6번, 7번, 8번, 9번, 15번, 20번, 25번, 30번, 35번, 40번, 45번 및 50번까지를 사용하여 구축한 XGB 모델의 성능을 실시예 4의 방법으로 확인한 결과, 표 7 및 도 10에 기재된 바와 같이 5개의 상위 Feature를 사용하더라도 충분한 성능이 발휘되는 것을 확인을 하였다.

표 7

Top N feature performance

Type	Top1	Top2	Top3	Top4	Top5	Top6	Top7	Top8	Top9	Top10	Top15	Top20	Top25	Top30	Top35	Top40	Top45	Top50	ALL
ACC_Train	0.845	0.938	0.977	0.930	0.915	0.977	0.868	0.915	0.891	0.946	0.891	0.891	0.930	0.922	1.000	0.938	0.953	1.000	1.000
ACC_Valid	0.855	0.927	0.909	0.927	0.945	0.927	0.909	0.927	0.909	0.909	0.909	0.909	0.927	0.927	0.945	0.927	0.927	0.945	0.945
ACC_Test	0.911	0.873	0.889	0.934	0.987	0.949	0.886	0.937	0.886	0.911	0.886	0.856	0.861	0.873	0.911	0.859	0.886	0.937	0.937
AUC_Train	0.941	0.991	0.997	0.980	0.988	0.986	0.914	0.980	0.931	0.991	0.906	0.906	0.957	0.940	1.000	0.986	0.993	1.000	1.000
AUC_Valid	0.653	0.870	0.888	0.934	0.916	0.912	0.884	0.883	0.876	0.919	0.874	0.874	0.927	0.910	0.934	0.936	0.912	0.929	0.952
AUC_Test	0.953	0.926	0.931	0.965	0.979	0.971	0.964	0.974	0.952	0.933	0.938	0.938	0.919	0.933	0.974	0.948	0.945	0.979	0.987

[0222]

[0223] 즉, 표 7의 위 3행은 Accuracy (ACC) 방법으로 성능을 측정한 결과이고, 아래 3 행은 AUC 방법으로 성능을 측정한 결과이다. ACC와 AUC 성능을 측정한 Train, valid, test set의 구성은 동일하다. Accuracy (ACC) 는 모델에서 예측된 확률 값이 정해진 cutoff 값 (cutoff = 0.5) 보다 높은지 낮은지를 판단해 측정하는 성능 지표이며, AUC는 ACC와 다르게 특정한 cutoff를 설정하지 않고, 예측된 확률 값의 분포가 정상인 집단과 암 환자 집단에서

얼마만큼 분명하게 차이나는지를 측정하는 성능 지표이다.

[0224] ACC 의 경우 cutoff 값을 어떻게 설정하는지에 따라 결과가 달라질 수 있기 때문에, AUC 값을 기준으로 해석하는 것이 맞다. Test set의 AUC 값을 기준으로 표 7의 결과는

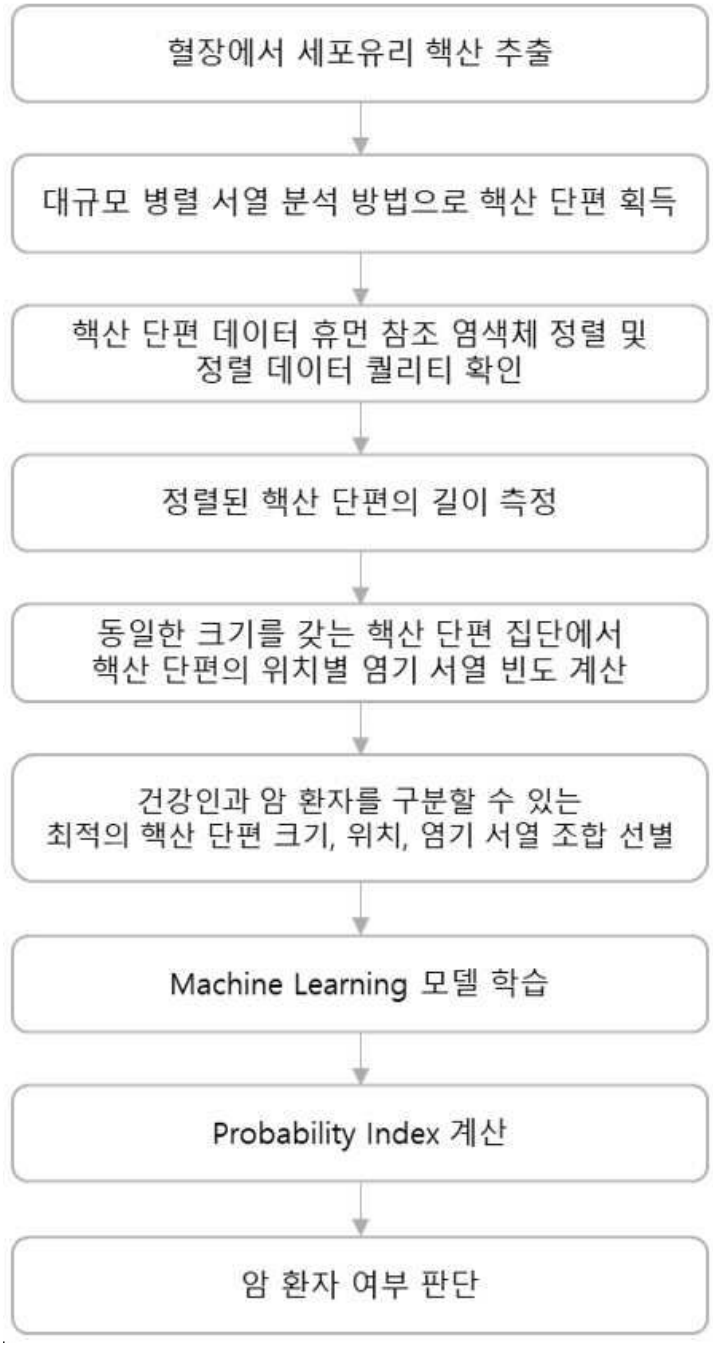
[0225] i) 375 개의 모든 Feature를 사용했을 때 AUC=0.987로, feature들의 일부 부분집합을 사용했을 때와 비교했을 때 가장 높은 성능을 보이고 있다.

[0226] ii) 375 개 feqture를 사용했을 때와 비슷한 Test AUC 성능을 확보할 수 있는 가장 적은 feature의 개수를 찾아보면 TopN = 5 인 것을 확인할 수 있다.

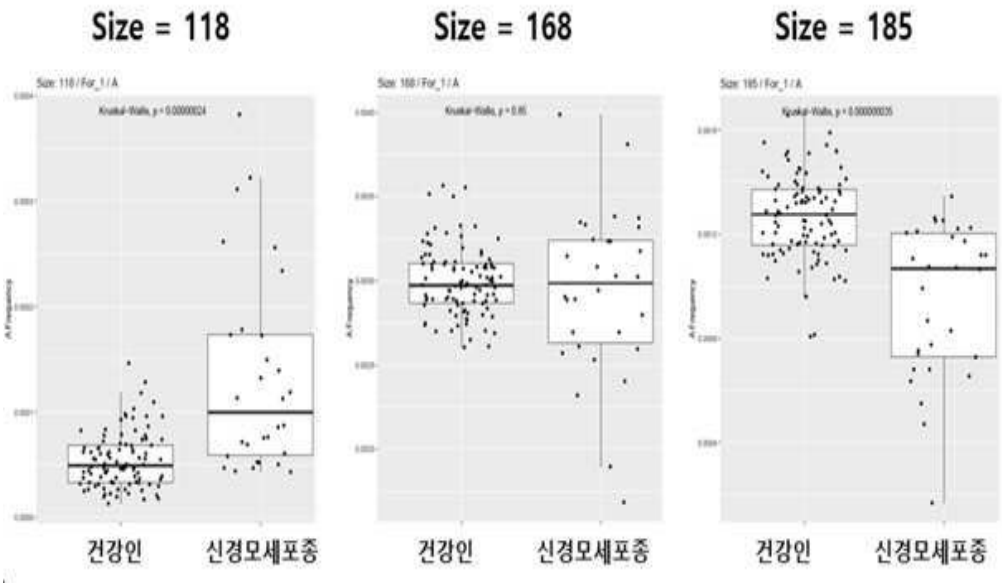
[0228] 이상으로 본 발명 내용의 특정한 부분을 상세히 기술하였는 바, 당업계의 통상의 지식을 가진 자에게 있어서 이러한 구체적 기술은 단지 바람직한 실시 양태일 뿐이며, 이에 의해 본 발명의 범위가 제한되는 것이 아닌 점은 명백할 것이다. 따라서, 본 발명의 실질적인 범위는 첨부된 청구항들과 그것들의 등가물에 의하여 정의된다고 할 것이다.

도면

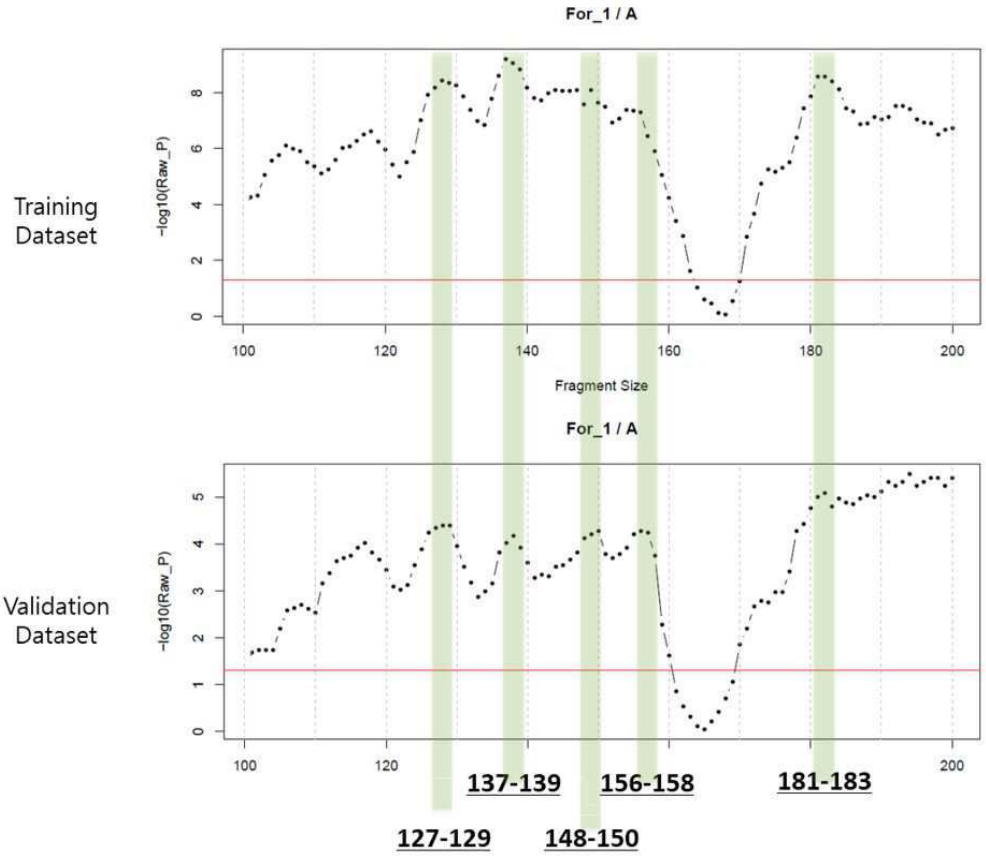
도면1


혈장에서 세포유리 핵산 추출
대규모 병렬 서열 분석 방법으로 핵산 단편 획득
핵산 단편 데이터 휴먼 참조 염색체 정렬 및
정렬 데이터 퀄리티 확인
정렬된 핵산 단편의 길이 측정
동일한 크기를 갖는 핵산 단편 집단에서
핵산 단편의 위치별 염기 서열 빈도 계산
건강인과 암 환자를 구분할 수 있는
최적의 핵산 단편 크기, 위치, 염기 서열 조합 선별
Machine Learning 모델 학습
Probability Index 계산
암 환자 여부 판단

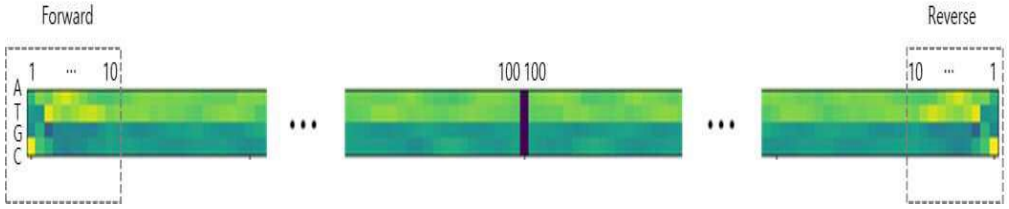
도면2


Size = 118
Size = 168
Size = 185
건강인
신경모세포종
건강인
신경모세포종
건강인
신경모세포종

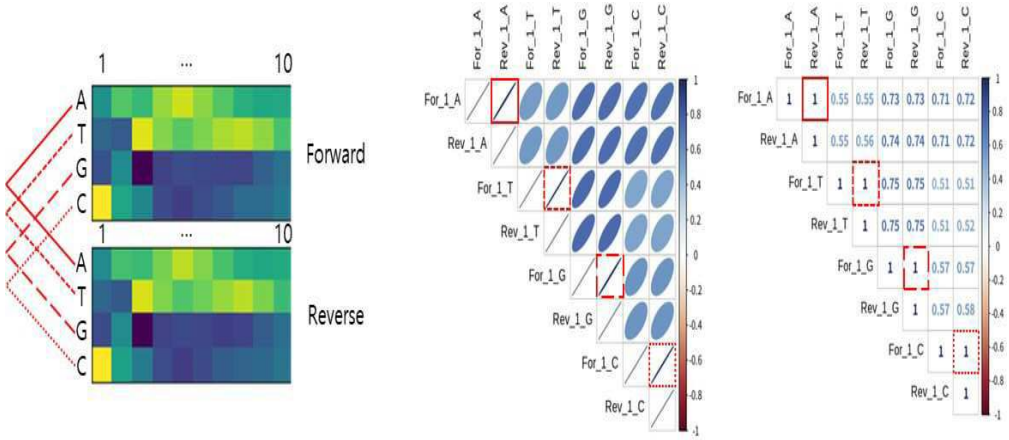
도면3


Training Dataset
Validation Dataset
For_1 / A
-log10(Raw_P)
Fragment Size
For_1 / A
-log10(Raw_P)
127-129
137-139
148-150
156-158
181-183

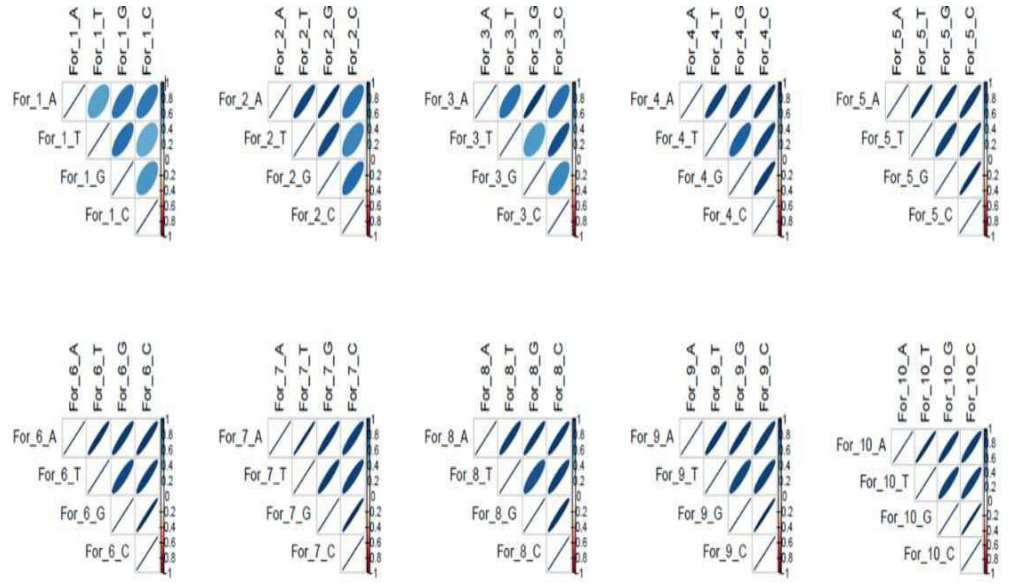
도면4


Forward
Reverse
1 ... 10
100 100
10 ... 1
A
T
G
C

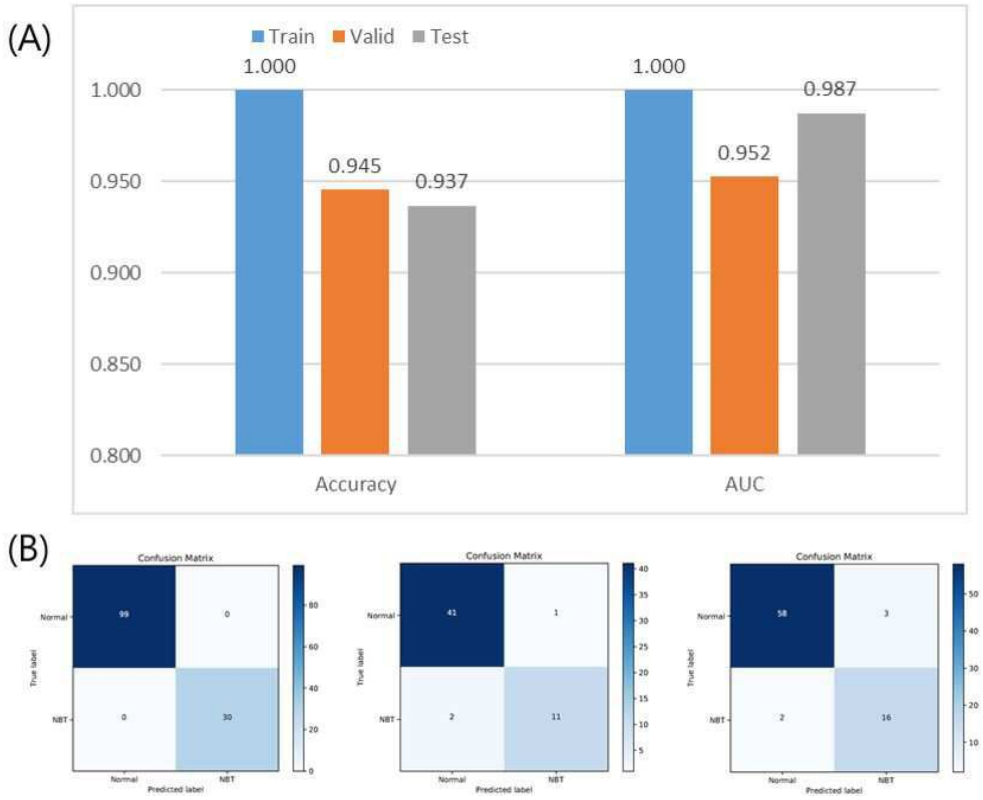
도면5


1 ... 10
Forward
Reverse
A
T
G
C

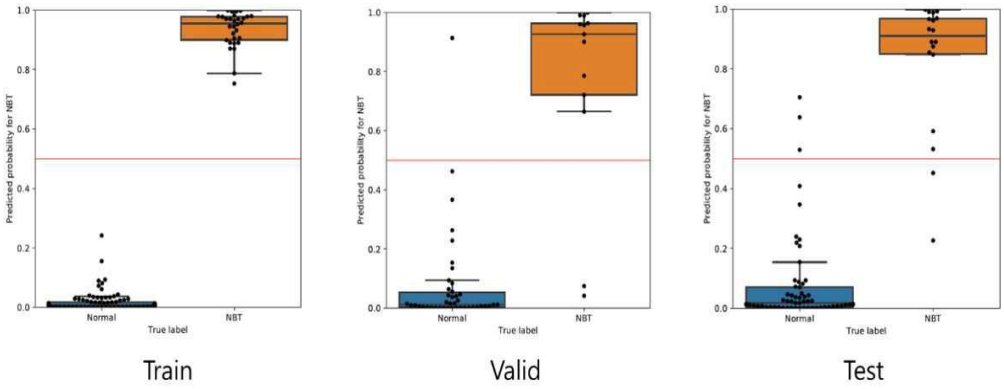
도면6



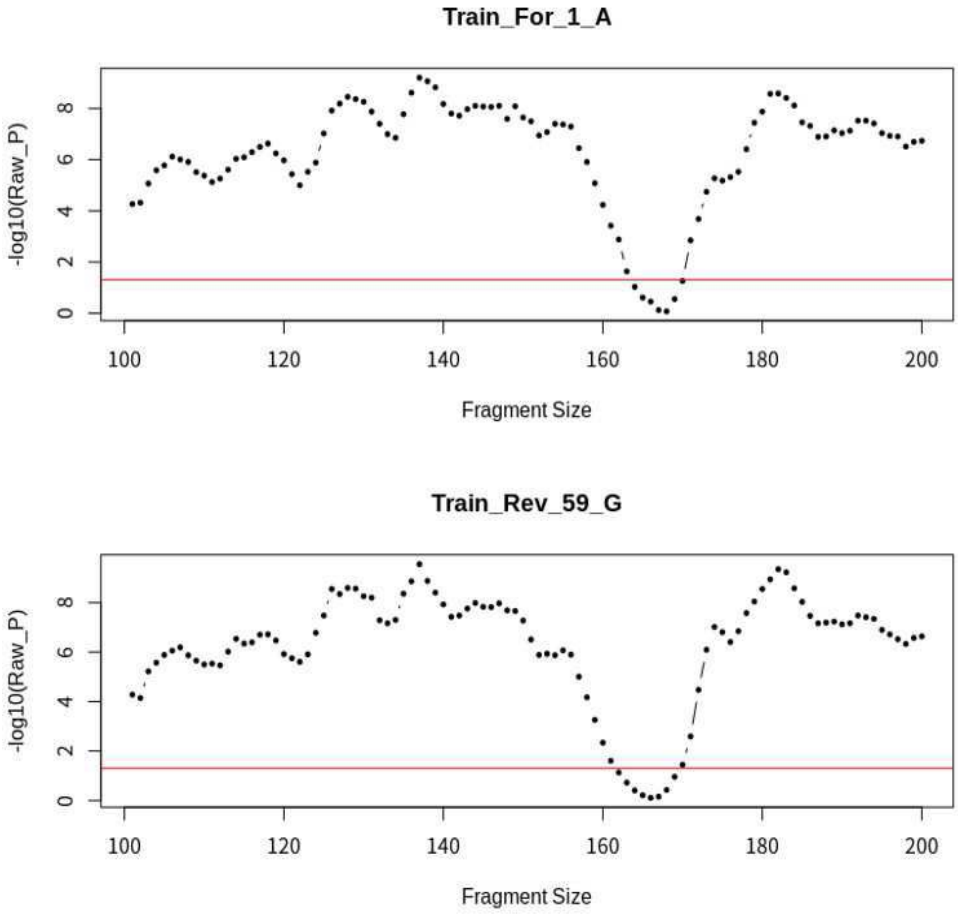
도면7


(A)
Train
Valid
Test
1.000
0.945
0.937
1.000
0.952
0.987
1.000
0.950
0.900
0.850
0.800
Accuracy
AUC
(B)
Confusion Matrix
Normal
NBT
True label
Predicted label
99
0
0
30
41
1
2
11
58
3
2
16

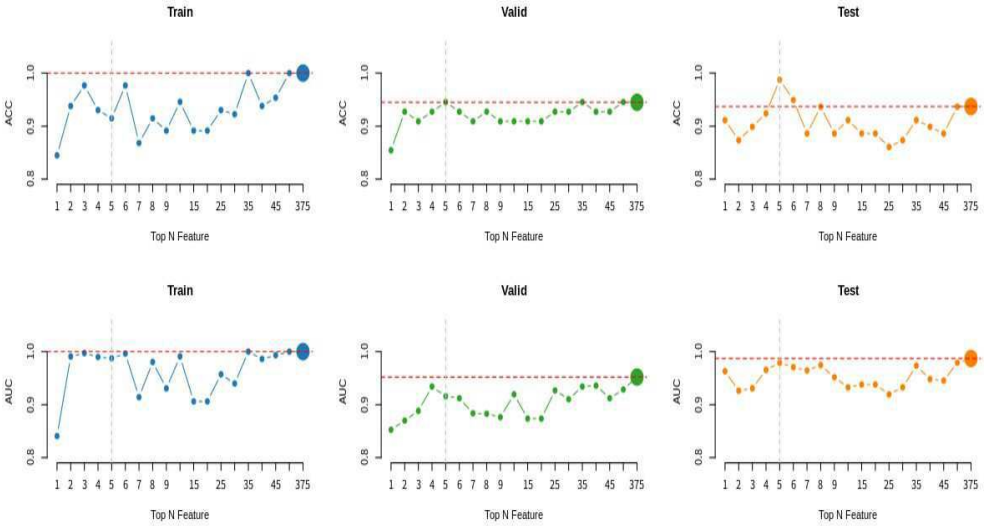
도면8


Predicted probability for NBT
Normal
NBT
True label
Train
Valid
Test

도면9


Train_For_1_A
-log10(Raw_P)
0
2
4
6
8
100
120
140
160
180
200
Fragment Size
Train_Rev_59_G
-log10(Raw_P)
0
2
4
6
8
100
120
140
160
180
200
Fragment Size

도면10


Train
Valid
Test
ACC
AUC
0.8
0.9
1.0
1 2 3 4 5 6 7 8 9 15 25 35 45 375
Top N Feature