

(12) 특허협력조약에 의하여 공개된 국제출원

(19) 세계지식재산권기구
국제사무국

(43) 국제공개일

2021년 5월 6일 (06.05.2021)



(10) 국제공개번호

WO 2021/085702 A1

- (51) 국제특허분류:
G16B 25/10 (2019.01) G16B 35/10 (2019.01)
G16B 30/10 (2019.01)
- (21) 국제출원번호: PCT/KR2019/014785
- (22) 국제출원일: 2019년 11월 4일 (04.11.2019)
- (25) 출원언어: 한국어
- (26) 공개언어: 한국어
- (30) 우선권정보:
10-2019-0138120 2019년 10월 31일 (31.10.2019) KR
- (71) 출원인: 주식회사 모비스 (MOBIS CO., LTD.) [KR/KR]; 13486 경기도 성남시 분당구 판교로 253, 씨동 203호, Gyeonggi-do (KR). 울산대학교 산학협력단 (UNIVERSITY OF ULSAN FOUNDATION FOR INDUSTRY COOPERATION) [KR/KR]; 44610 울산시 남구 대학로 93 (무거동), Ulsan (KR). 재단법인 아산사회복지재단 (THE ASAN FOUNDATION) [KR/KR]; 05505 서울시 송파구 올림픽로 43길 88 (풍납동), Seoul (KR).
- (72) 발명자: 김지현 (KIM, Jee Hun); 13628 경기도 성남시 분당구 미금로 177, 309동 1102호, Gyeonggi-do (KR). 심재용 (SIM, Jae Yong); 16930 경기도 용인시 수지구 만

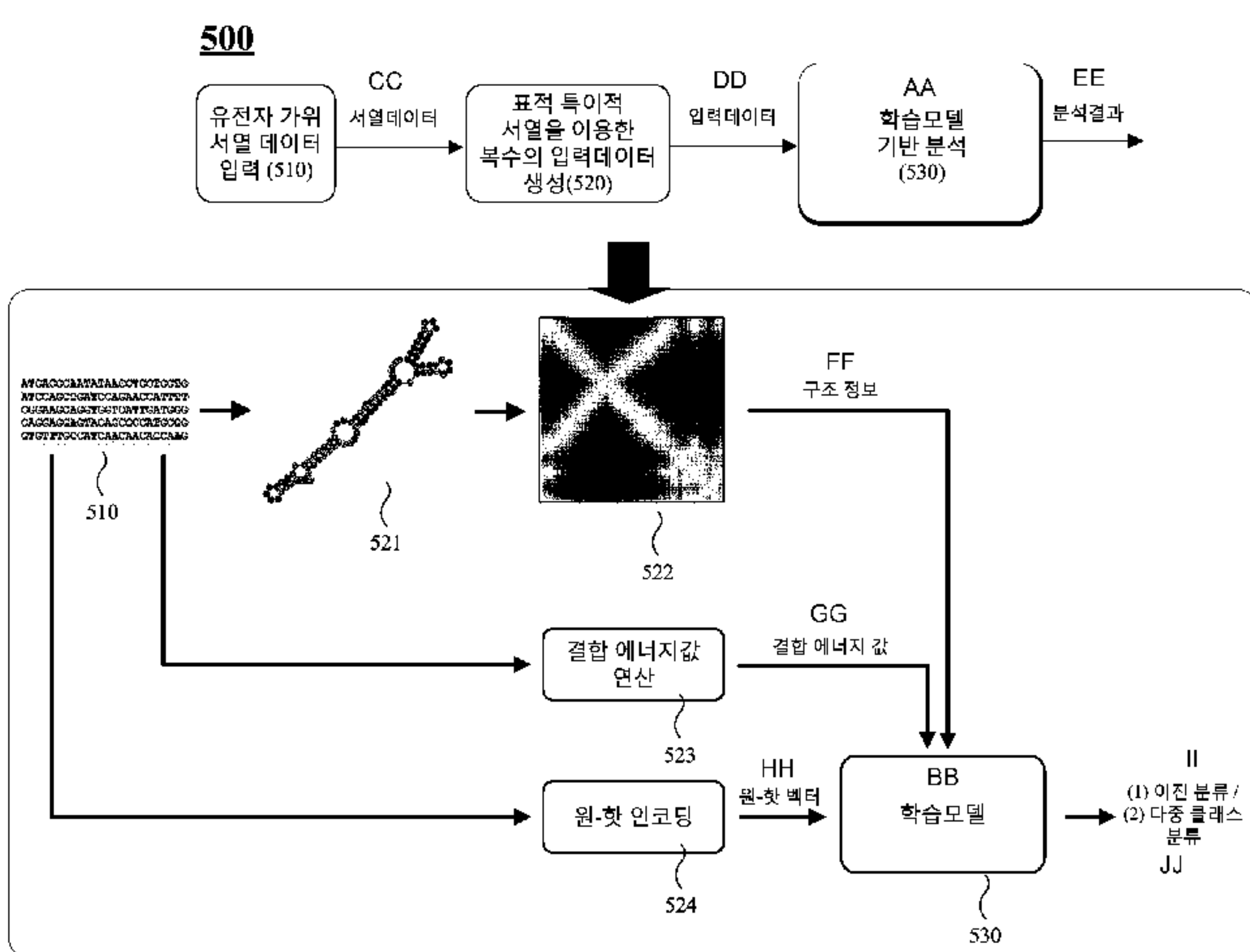
현로 127, 808동 901호, Gyeonggi-do (KR). 이근호 (LEE, Geun Ho); 08360 서울시 구로구 연동로13길 25, 108동 102호, Seoul (KR). 유현 (YOO, Hyun); 07946 서울시 양천구 목동중앙북로7가길 38, 102동 603호, Seoul (KR). 이준현 (LEE, Jun Hyun); 16930 경기도 용인시 수지구 만현로 79, 505동 304호, Gyeonggi-do (KR). 이준엽 (LEE, Joon Yhup); 05400 서울시 강동구 강동대로 143-16, 801호, Seoul (KR). 하지현 (HA, Ji Hyeon); 16362 경기도 수원시 장안구 율전로 64번길 7, 302호, Gyeonggi-do (KR).

(74) 대리인: 특허법인 아이시스 (ISIS IP LAW LLC); 06160 서울시 강남구 테헤란로63길 14, 8층, Seoul (KR).

(81) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 국내 권리의 보호를 위하여): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(54) Title: METHOD AND APPARATUS FOR ANALYZING PROGRAMMABLE NUCLEASE EFFECT

(54) 발명의 명칭: 유전자 가위 효과를 분석하는 방법 및 장치



510 ... Input sequence data of programmable nuclease
520 ... Generate plurality of pieces of input data by using target-specific sequence
523 ... Calculate bond energy value
524 ... One-hot encoding
AA ... Learning model-based analysis
BB ... Learning model
CC ... Sequence data
DD ... Input data
EE ... Analysis result
FF ... Structure information
GG ... Bond energy value
HH ... One-hot vector
II ... (1) Binary classification
JJ ... (2) Multi-class classification

(57) Abstract: A method for designing a programmable nuclease by using a learning model comprises the steps in which: an analysis apparatus receives sequence data of a programmable nuclease; the analysis apparatus uses the sequence data so as to generate structure information about a structure formed by a target-specific sequence of the programmable nuclease; the analysis apparatus inputs the structure information into the learning model; and the analysis apparatus evaluates the effect of the programmable nuclease on the basis of information outputted by the learning model.

(57) 요약서: 학습모델을 이용하여 유전자 가위를 설계하는 방법은 분석장치가 유전자 가위의 서열 데이터를 입력받는 단계, 상기 분석장치가 상기 서열 데이터를 이용하여 유전자 가위의 표적 특이적 서열이 구성하는 구조에 대한 구조 정보를 생성하는 단계, 상기 분석장치가 상기 구조 정보를 학습모델에 입력하는 단계 및 상기 분석장치가 상기 학습모델이 출력하는 정보를 기준으로 상기 유전자 가위의 효과를 평가하는 단계를 포함한다.



(84) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 역
내 권리의 보호를 위하여): ARIPO (BW, GH, GM, KE,
LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM,
ZW), 유라시아 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 유
럽 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI,
FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK,
MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI
(BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML,
MR, NE, SN, TD, TG).

공개:

— 국제조사보고서와 함께 (조약 제21조(3))

명세서

발명의 명칭: 유전자 가위 효과를 분석하는 방법 및 장치 기술분야

- [1] 이하 설명하는 기술은 유전자 가위 서열을 기준으로 유전자 가위 효과를 분석하는 기법에 관한 것이다.

배경기술

- [2] 유전자 가위(programmable nuclease)는 세포 및 개체의 유전자 특성 부위를 자르는 인공 효소이다. 유전자 가위는 인공 효소로 표적 유전자를 제거하는 유전자 편집 (genome editing) 기술을 의미하기도 한다.
- [3] 유전자 가위 분야의 연구 목표는 표적 서열에 대한 효율(on-target efficiency)을 높이고, 동시에 목표하지 않은 다른 서열에 대한 작용(off-target activity)을 줄이는 것이다. 유전자 가위는 가이드 RNA의 서열과 표적 서열이 일부 상호보완이 아닌 경우 경우에도 작용할 수 있다. 이 경우, 유전자 가위가 표적 서열 외에 다른 서열에 접합하여 DNA 돌연변이를 유발할 수 있다.

발명의 상세한 설명

기술적 과제

- [4] 종래 연구는 유전자 가위에서 표적 서열에 특이적인 서열을 변경해가면서 표적 서열에 효과적인지 테스트를 수행하는 방식을 사용한다. 이와 같은 종래 방법은 유전자 서열 중 하나 이상의 뉴클레오타이드를 변경하면서, 표적 서열에 대한 유전자 가위의 효과를 검증하는 실험을 반복한다. 결국, 종래 접근법은 효과적인 유전자 가위를 탐색하는데 매우 많은 시간과 비용이 소요된다는 한계가 있다.
- [5] 이하 설명하는 기술은 학습모델을 사용하여 효과적인 유전자 가위를 검출하고자 한다.

과제 해결 수단

- [6] 학습모델을 이용하여 유전자 가위 효과를 분석하는 방법은 분석장치가 유전자 가위의 서열 데이터를 입력받는 단계, 상기 분석장치가 상기 서열 데이터를 이용하여 유전자 가위의 표적 특이적 서열이 구성하는 구조에 대한 구조 정보를 생성하는 단계, 상기 분석장치가 상기 구조 정보를 학습모델에 입력하는 단계 및 상기 분석장치가 상기 학습모델이 출력하는 정보를 기준으로 상기 유전자 가위의 효과를 평가하는 단계를 포함한다.
- [7] 유전자 가위 효과를 예측하는 분석장치는 유전자 가위의 서열 데이터를 입력받는 입력장치, 뉴클레오타이드 서열 기준으로 해당 서열이 구성하는 구조에 대한 구조 정보를 생성하는 프로그램 및 유전자 가위의 구조 정보를 기준으로 유전자 가위의 효과를 분석하는 학습모델을 저장하는 저장장치 및 상기 프로그램을 이용하여 상기 서열 데이터에서 유전자 가위의 표적 특이적 서열에 대한 구조 정보를 생성하고, 생성된 구조 정보를 상기 학습모델에

입력하여 유전자 가위의 효과를 평가하는 연산장치를 포함한다.

- [8] 유전자 가위 효과를 예측하는 시스템은 뉴클레오타이드 서열 기준으로 해당 서열이 구성하는 구조에 대한 구조 정보를 생성하는 구조 예측 서버 및 유전자 가위의 구조 정보를 기준으로 유전자 가위의 효과를 분석하는 학습모델을 저장하고, 상기 구조 예측 서버로부터 유전자 가위의 서열 데이터에서 유전자 가위의 표적 특이적 서열에 대한 구조 정보를 수신하고, 상기 구조 정보를 상기 학습모델에 입력하여 유전자 가위의 효과를 평가하는 평가 서버를 포함한다.

발명의 효과

- [9] 이하 설명하는 기술은 유전자 가위의 서열 데이터를 기준으로 특정한 표적 서열에 효과적인지 빠르게 예측한다. 이하 설명하는 기술은 사전에 훈련된 기계학습모델을 사용하여 표적 서열에 효과적인 유전자 가위의 서열을 정확하게 예측한다.

도면의 간단한 설명

- [10] 도 1은 CRISPR Cas9 구조에 대한 예이다.
 [11] 도 2는 유전자 가위 효과를 예측하는 시스템에 대한 예이다.
 [12] 도 3은 입력 데이터에 대한 예이다.
 [13] 도 4는 유전자 가위 효과를 예측하는 과정에 대한 예이다.
 [14] 도 5는 유전자 가위 효과를 예측하는 과정에 대한 다른 예이다.
 [15] 도 6은 인공 신경망에 대한 예이다.
 [16] 도 7은 인공 신경망에 대한 다른 예이다.
 [17] 도 8은 학습모델을 훈련하는 과정에 대한 예이다.
 [18] 도 9는 유전자 가위 효과를 예측하는 분석장치에 대한 예이다.
 [19] 도 10은 유전자 가위 효과 예측한 모델 성능 평가에 대한 예이다.
 [20] 도 11은 유전자 가위 효과 예측한 모델 성능 평가에 대한 다른 예이다.
 [21] 도 12는 유전자 가위 효과 예측한 모델 성능 평가에 대한 또 다른 예이다.

발명의 실시를 위한 형태

- [22] 이하 설명하는 기술은 다양한 변경을 가할 수 있고 여러 가지 실시례를 가질 수 있는 바, 특정 실시례들을 도면에 예시하고 상세하게 설명하고자 한다. 그러나, 이는 이하 설명하는 기술을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 이하 설명하는 기술의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.
- [23] 제1, 제2, A, B 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 해당 구성요소들은 상기 용어들에 의해 한정되지는 않으며, 단지 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 이하 설명하는 기술의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. 및/또는 이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의

관련된 기재된 항목들 중의 어느 항목을 포함한다.

[24] 본 명세서에서 사용되는 용어에서 단수의 표현은 문맥상 명백하게 다르게 해석되지 않는 한 복수의 표현을 포함하는 것으로 이해되어야 하고, "포함한다" 등의 용어는 실시된 특징, 개수, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함을 의미하는 것이지, 하나 또는 그 이상의 다른 특징들이나 개수, 단계 동작 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 배제하지 않는 것으로 이해되어야 한다.

[25] 또, 방법 또는 동작 방법을 수행함에 있어서, 상기 방법을 이루는 각 과정들은 문맥상 명백하게 특정 순서를 기재하지 않은 이상 명기된 순서와 다르게 일어날 수 있다. 즉, 각 과정들은 명기된 순서와 동일하게 일어날 수도 있고 실질적으로 동시에 수행될 수도 있으며 반대의 순서대로 수행될 수도 있다.

[26]

[27] 이하 명세서에서 사용하는 기술 내지 용어에 대하여 먼저 설명한다.

[28] 개체(subject)는 세포, 조직 또는 유기체를 포함한다. 개체는 기본적으로 인간, 동물, 식물, 미생물 등을 포함하는 의미이다.

[29] 유전자 가위는 세포 및 개체의 유전자 특성 부위를 자르는 인공 효소이다. 유전자 가위는 인공 효소로 표적 유전자를 제거하는 유전자 편집 (genome editing) 기술을 의미하기도 한다. 예컨대, 유전자 가위를 이용하여 손상된 DNA를 제거하고, 정상 DNA를 삽입하여 질병을 치료할 수 있다.

[30] 대표적인 유전자 가위는 징크핑거 뉴클레이즈(Zinc Finger Nuclease, ZFN), 탈렌(TALENs·Transcriptor Activator-Like Effector Nucleases), 크리스퍼(CRISPR)가 있다. 크리스퍼는 3세대 유전자 가위에 해당하며, 현재 가장 주목을 받는 유전자 가위이다. 이하 설명의 편의를 위하여 크리스퍼를 기준으로 설명한다.

[31] 표적 유전자 내지 표적 서열은 유전자 가위의 편집 대상이 되는 서열을 의미한다.

[32] 크리스퍼 유전자 가위는 유전자 편집의 대상이 되는 DNA의 상보적 염기를 지니는 RNA(크리스퍼 RNA) 및 표적 유전자를 제거하는 효소로 구성된다. 크리스퍼 캐스9(CRISPR Cas9)은 표적 유전자를 찾는 역할을 수행하는 크리스퍼와 표적 유전자를 제거하는 단백질인 캐스9으로 구성된다. 한편, 크리스퍼 유전자 가위 관련한 연구가 진행되면서, 캐스 9이 아닌 다른 단백질(예컨대, Cpf1)을 이용하는 유전자 가위도 연구되고 있다. 이하 설명의 편의를 위하여 가장 많이 연구되는 CRISPR Cas9을 기준으로 설명한다.

[33] CRISPR Cas9은 표적 서열은 찾기 위한 가이드 RNA 및 Cas9으로 구성된다. 가이드 RNA는 sgRNA(single guide RNA)로 구성된다. 가이드 RNA가 표적 서열을 찾는데 핵심적 역할을 한다. 가이드 RNA는 연구자가 특정 서열을 갖도록 합성할 수 있다. 가이드 RNA를 합성하는 기술 및 스크리닝하는 기술은 다양하다.

- [34] 이하 설명하는 기술은 표적 서열에 대한 효과가 좋고, 표적 서열이 아닌 다른 서열에 대한 부작용이 적은 유전자 가위를 검출하기 위한 것이다.
- [35] 기계 학습(machine learning)은 인공 지능의 한 분야로, 컴퓨터가 학습할 수 있도록 알고리즘을 개발하는 분야를 의미한다. 학습모델은 데이터 집합에 대해 특정 유형의 패턴을 파악하고 인식하도록 학습된 모델이고, 결과물은 프로그램 코드를 저장하는 파일 형태이다. 학습모델은 접근 방법에 따라 인공신경망, 결정 트리 등과 같은 다양한 유형의 모델이 있다.
- [36] 이하 설명하는 기술은 유전자 가위의 서열을 기준으로 표적 서열에 효과적인 유전자 가위를 검출한다. 연구자는 이하 설명하는 기술을 이용하여 유전자 가위 후보에 대한 효과를 검증한다. 연구자는 이하 설명하는 기술을 기반으로 효과적인 유전자 가위를 설계할 수 있다.
- [37]
- [38] 도 1은 CRISPR Cas9 구조에 대한 예이다. 크리스퍼 유전자 가위는 목표 서열에 상보성을 갖는 서열을 갖는다. 고유 영역(constant region)은 루프 구조를 포함하는데, 캐스 9 단백질이 결합하는 부위이다. 목표 서열에 특이적인 서열은 crRNA(CRISPR RNA)라고 명명하기도 한다. 캐스 9와 결합하는 부위를 tracrRNA라고 명명하기도 한다. 가이드 RNA는 crRNA와 tracrRNA로 구성되는 복합체라고 할 수 있다. crRNA 일부와 tracrRNA 일부가 결합하여 루프 구조를 형성하기도 한다. 표적 특이적 서열은 20nt(nucleotide) 길이의 스페이서(spacer)를 포함한다.
- [39] 종래 연구 방법은 스페이서 영역의 서열 중 nt를 하나씩 변경해가면서 표적에 대한 결합력을 측정하여 유전자 가위의 효과를 예측하였다.
- [40] 이하 설명하는 기술은 표적 서열에 특이적인 crRNA를 중심으로 표적 서열에 효과적인 유전자 가위 구조를 검출하고자 한다. 이하 설명하는 기술은 표적 서열 특이적이고, 설계 가능한 서열을 검출하고자 한다. 설계 가능한 서열은 sgRNA 중 20nt일 수 있다. 또는 설계 가능한 서열은 crRNA 전체일 수도 있다.
- [41] 이하 설명하는 기술은 학습모델을 이용하여 sgRNA 전체 또는 일부 서열을 분석한다. 이하 학습모델이 입력받는 서열을 표적 특이적 서열이라고 명명한다. 표적 특이적 서열은 표적에 직접 결합하는 서열을 포함한다. 나아가, 표적 특이적 서열은 표적 서열 결합에 간접적으로 영향을 주는 서열을 포함할 수 있다. 따라서, 표적 특이적 서열은 crRNA 전체 또는 일부를 포함할 수도 있다.
- [42]
- [43] 도 2는 유전자 가위 효과를 예측하는 시스템에 대한 예이다. 도 2는 3가지 유형의 시스템 내지 장치를 도시한다. 분석 장치가 유전자 가위의 효과를 예측하는 주체이다. 분석 장치는 유전자 가위의 서열을 기준으로 해당 유전자 가위가 특정 표적 서열에 효과적인지 여부를 분석한다. 도 2에서 분석 장치는 분석 서버(110, 210) 및 컴퓨터 단말(300)의 형태로 도시하였다.
- [44] 도 2(A)는 분석 서버(110) 및 구조 예측 서버(120)를 포함하는 시스템(100)에

대한 예이다. 사용자 단말(10)은 유전자 가위 서열을 분석 서버(110)에 전달한다. 유전자 가위 서열은 디지털 데이터의 형태이다. 분석 서버(110)는 유전자 가위 서열을 구조 예측 서버(120)에 전달한다. 구조 예측 서버(120)는 뉴클레오타이드 서열을 기준으로 해당 서열이 구성하는 2차 구조 또는 3차 구조를 예측한다. 구조 예측 서버(120)는 유전자 가위 서열 중 표적 특이적 서열의 2차 구조 또는 3차 구조를 예측할 수 있다.

- [45] RNA 서열을 기준으로 해당 RNA 서열이 구성하는 2차 구조 또는 3차 구조를 예측하는 상용툴도 있다. 예컨대, RNAdraw, RNAfold, RNAstructure 등이 있다. 따라서, 구조 예측 서버(120)는 이와 같은 상용 프로그램 또는 자체 프로그램을 사용하여 수신하는 서열에 대한 2차 구조 또는 3차 구조를 예측할 수 있다. 구조 예측 서버(120)는 2차 구조 또는 3차 구조에 대한 정보를 분석 서버(110)에 전달한다.
- [46] 이하, 표적 특이적 서열에 대한 2차/3차 구조에 대한 정보 및 2차/3차 구조를 기준으로 생성된 정보를 모두 구조 정보라고 명명한다. 구조 정보는 RNA의 2차 구조 영상, 3차 구조 영상, 2차 구조를 표현하는 매트릭스, 3차 구조를 표현하는 매트릭스, 2차/3차 구조 기준으로 생성되는 거리 맵(distance map), 2차/3차 구조 기준으로 생성되는 접촉 맵(contact map) 등 다양한 형태 중 어느 하나일 수 있다. 구조 정보는 학습모델에 입력되는 입력 데이터에 해당한다.
- [47] 분석 서버(110)는 사전에 훈련된 학습모델을 보유한다. 학습모델은 구조 정보를 기준으로 유전자 가위에 대한 효과에 대한 정보를 출력하는 모델이다. 분석 서버(110)는 수신한 구조 정보를 학습모델에 입력한다. 분석 서버(110)는 학습모델이 출력하는 정보를 기준으로 유전자 가위에 대한 효과를 예측 내지 검증한다. 분석 서버(110)는 분석한 결과를 사용자 단말(10)에 전송할 수 있다.
- [48] 분석 서버(110)는 구조 예측 서버(120)로부터 수신한 표적 특이적 서열에 대한 2차 구조 또는 3차 구조에 대한 정보를 가공하여 거리 맵, 에너지 값 등을 생성할 수도 있다.
- [49] 도 2(B)는 분석 서버(210)를 포함하는 시스템(200)에 대한 예이다. 사용자 단말(20)은 유전자 가위 서열을 분석 서버(210)에 전달한다. 유전자 가위 서열은 디지털 데이터의 형태이다.
- [50] 분석 서버(210)는 유전자 가위의 효과를 예측한다. 분석 서버(210)는 뉴클레오타이드 서열을 기준으로 해당 서열이 구성하는 2차 구조 또는 3차 구조를 예측하는 프로그램 내지 모델을 보유한다. 분석 서버(210)는 유전자 가위 서열 중 표적 특이적 서열의 2차 구조 또는 3차 구조를 예측한다. 분석 서버(210)는 전술한 구조 정보를 생성한다.
- [51] 분석 서버(210)는 사전에 훈련된 학습모델을 보유한다. 분석 서버(210)는 수신한 구조 정보를 학습모델에 입력한다. 분석 서버(210)는 학습모델이 출력하는 정보를 기준으로 유전자 가위에 대한 효과를 예측 내지 검증한다. 분석 서버(210)는 분석한 결과를 사용자 단말(20)에 전송할 수 있다.

- [52] 도 2(C)는 컴퓨터 단말(310) 형태의 분석 장치에 대한 예이다. 컴퓨터 단말(310)은 뉴클레오타이드 서열을 기준으로 해당 서열이 구성하는 2차 구조 또는 3차 구조를 예측하는 프로그램 내지 모델을 보유한다. 또한, 컴퓨터 단말(310)은 구조 정보를 기준으로 유전자 가위에 대한 효과에 대한 정보를 출력하는 학습모델을 보유한다.
- [53] 컴퓨터 단말(310)은 사용자(30)로부터 입력받은 유전자 가위 서열을 기준으로 유전자 가위 서열 중 표적 특이적 서열의 2차 구조 또는 3차 구조를 예측한다. 컴퓨터 단말(310)은 예측되는 2차 구조 또는 3차 구조에 대한 정보 내지 연관 정보를 포함하는 구조 정보를 생성한다. 컴퓨터 단말(310)은 생성한 구조 정보를 학습모델에 입력한다. 컴퓨터 단말(310)은 학습모델이 출력하는 정보를 기준으로 유전자 가위에 대한 효과를 예측 내지 검증한다. 사용자(30)는 분석 결과를 확인할 수 있다.
- [54]
- [55] 도 3은 입력 데이터에 대한 예이다.
- [56] 도 3(A)는 RNA의 2차 구조 영상에 해당한다. 입력 데이터(구조 정보)가 일반적인 영상인 예이다. 도 3(A)는 비교적 많은 서열로 구성된 2차 구조를 예로 도시하였다. 유전자 가위 경우, 가이드 RNA 또는 표적 특이적 서열이 구성하는 2차 구조는 도 3(A)보다 단순할 수도 있다.
- [57] 도 3(B)는 RNA 서열이 구성하는 2차 구조를 기준으로 생성한 거리 맵에 대한 예이다. 거리 맵은 가로축과 세로축을 갖는 2차원 매트릭스이다. 가로축 및 세로축은 일정한 순서에 따라 라벨링되는 뉴클레오타이드 서열이다. 거리 맵은 2차 구조에서 하나의 뉴클레오타이드를 기준으로 다른 뉴클레오타이드와의 거리를 나타낸다. 도 3(B)는 거리를 일정한 색상으로 표시한 예이다. 예컨대, 더 밝은 색일수록 가까운 거리를 나타낼 수 있다.
- [58] 도 3(B)에서 입력 데이터는 2차원 거리 맵 영상이다. 나아가, 도 3(B)의 거리 맵은 가로축과 세로축의 요소가 동일하기 때문에 대각선을 기준으로 대칭된다. 따라서, 입력 데이터는 중복되는 영역을 제외한 삼각형 영역만을 사용할 수도 있다.
- [59] 도 3(A) 및 도 3(B)는 전술한 구조 정보에 대한 예이다. 한편, 학습모델은 구조 정보 외에 다른 형태의 입력 데이터를 사용할 수도 있다. 다른 형태의 입력 데이터도 뉴클레오타이드 서열 특이적인 정보이다. 예컨대, 연구자는 상용 프로그램을 이용하여 뉴클레오타이드 서열로부터 이중 결합을 푸는데 소요되는 에너지 값을 연산할 수 있다.
- [60] 입력 데이터는 뉴클레오타이드 서열 특이적인 에너지 값이 될 수 있다. 도 3(C)에는 유전자 가위 정보 중 하나, 그리고 학습모델에서 사용할 수 있는 변수 중 하나인 에너지 값 (DNA-DNA, DNA-RNA hybrid duplexes energy)을 표현한다. 에너지 값 $E_i^C(\tilde{X}\tilde{Y})$ 는 DNA strand, 즉 DNA-DNA 결합을 푸는 것에 필요한

수치이며, 에너지 값 $E_i^{MM}(X\tilde{Z})$ 는 풀어진 DNA 부분에 RNA strand와 결합 형성후 결합 분리할 시 필요한 에너지 수치이다. 이하, 이중 결합을 푸는데 소요되는 에너지(hybrid duplexes energy)를 결합 에너지라고 명명한다.

- [61] 나아가, 입력 데이터는 뉴클레오타이드 서열을 인코딩한 벡터 형태일 수 있다. 도 3(D)는 유전자 가위 서열 일부분을 원-핫 인코딩 절차 예시다. 원-핫 인코딩한 결과 데이터를 원-핫 벡터라고 명명한다. 서열 내 뉴클레오타이드가 A, C, G, U로 구성이 되어있으면 뉴클레오타이드 당 해당하는 위치에 1로 표현하고 나머지는 0으로 표현한다. 학습모델은 도 3과 같은 다양한 입력 데이터 유형을 이용할 수 있다.
- [62]
- [63] 도 4는 유전자 가위 효과를 예측하는 과정(400)에 대한 예이다. 분석장치(110, 220 또는 310)는 도 4에서 설명하는 예측 과정(400)을 통해 특정 유전자 가위의 효과를 예측할 수 있다.
- [64] 도 4는 구조 정보를 이용하여 유전자 가위 효과를 예측하는 과정에 대한 예이다. 분석 장치는 유전자 가위 서열 데이터를 입력받는다(410). 서열 데이터는 통상적으로 디지털 데이터 형태이다. 서열 데이터는 유전자 가위 전체의 서열, 가이드 RNA의 서열(crRNA+tracrRNA), 가이드 RNA의 서열의 일부 서열, crRNA + tracrRNA 및 crRNA 중 어느 하나일 수 있다.
- [65] 분석 장치는 입력받은 유전자 가위 서열 데이터 전체 또는 유전자 가위 서열 데이터 일부를 기준으로 구조 정보를 생성할 수 있다(420). 예컨대, 분석 장치는 표적 특이적 서열을 기준으로 구조 정보를 생성할 수 있다. 도 4의 하단에서 거리맵을 사용한 예를 도시하였다. 도 4의 거리 맵은 RNA의 2차 구조를 표현하는 영상 데이터에 해당한다.
- [66] 분석장치는 구조 정보를 학습모델에 입력하여 입력된 서열 데이터(특정 유전자 가위)에 대한 효과를 분석한다(430). 도 4의 하단에는 인공신경망 형태의 학습모델을 도시하였다. 분석장치는 학습모델이 출력하는 결과를 기준으로 분석 대상인 유전자 가위의 효과를 판단한다.
- [67] 학습모델은 0 또는 1의 값과 같이 2개의 분류만을 하는 이진 분류 모델일 수 있다. 이 경우 학습모델은 입력된 유전자 가위 서열이 특정 표적 서열에 효과적인지(1) 또는 효과가 없는지(0)에 대한 정보를 출력한다. 나아가 학습모델은 다중 클래스 중 어느 하나로 분류할 수 있다. 이 경우 학습모델은 입력된 유전자 가위 서열이 특정 표적 서열에 효과적인지 여부를 효과적인 클래스, 후보 클래스 및 효과없는 클래스 중 어느 하나로 분류할 수도 있다.
- [68]
- [69] 도 5는 유전자 가위 효과를 예측하는 과정(500)에 대한 다른 예이다. 도 5는 구조 정보 외에 부가 정보를 활용하여 이용하여 유전자 가위 효과를 예측하는 과정에 대한 예이다. 부가 정보는 서열 특이적인 결합 에너지 값 및 원-핫 벡터 중

적어도 하나를 포함할 수 있다.

- [70] 분석 장치는 유전자 가위 서열 데이터를 입력받는다(510). 서열 데이터는 통상적으로 디지털 데이터 형태이다. 서열 데이터는 유전자 가위 전체의 서열, 가이드 RNA의 서열(crRNA+tracrRNA), 가이드 RNA의 서열의 일부 서열, crRNA + tracrRNA 및 crRNA 중 어느 하나일 수 있다.
- [71] 분석 장치는 입력받은 유전자 가위 서열 데이터 전체 또는 유전자 가위 서열 데이터 일부를 기준으로 구조 정보 및 부가 정보를 생성할 수 있다(520).
- [72] 분석 장치는 표적 특이적 서열을 기준으로 구조 정보를 생성할 수 있다. 도 5는 구조 정보로 거리 맵을 사용하는 예를 도시한다. 분석 장치는 2차 구조를 예측하는 툴(tool)을 사용하여 표적 특이적 서열의 2차 구조를 예측한다(521). 분석 장치는 2차 구조를 기준으로 거리 맵을 생성한다(522).
- [73] 분석 장치는 표적 특이적 서열을 기준으로 부가 정보를 생성할 수 있다. 분석 장치는 결합력 예측 툴을 사용하여 표적 특이적 서열의 결합 에너지 값을 연산할 수 있다(523). 또한, 분석 장치는 표적 특이적 서열을 원-핫 인코딩하여 원-핫 벡터를 생성할 수도 있다(524).
- [74] 분석장치는 구조 정보 및 부가 정보를 학습모델에 입력하여 입력된 서열 데이터(특정 유전자 가위)에 대한 효과를 분석한다(530). 분석장치는 학습모델이 출력하는 결과를 기준으로 분석 대상인 유전자 가위의 효과를 판단한다.
- [75] 학습모델은 0 또는 1의 값과 같이 2개의 분류만을 하는 이진 분류 모델일 수 있다. 이 경우 학습모델은 입력된 유전자 가위 서열이 특정 표적 서열에 효과적인지(1) 또는 효과가 없는지(0)에 대한 정보를 출력한다. 나아가 학습모델은 다중 클래스 중 어느 하나로 분류할 수 있다. 이 경우 학습모델은 입력된 유전자 가위 서열이 특정 표적 서열에 효과적인지 여부를 효과적인 클래스, 후보 클래스 및 효과없는 클래스 중 어느 하나로 분류할 수도 있다.
- [76]
- [77] 신경망 모델은 RNN(Recurrent Neural Networks), FFNN(feedforward neural network), CNN(convolutional neural network) 등 다양한 모델이 사용될 수 있다. 이하 설명의 편의를 위하여 CNN 모델을 중심으로 설명한다. CNN 모델은 주로 컴퓨터 비전 분야에 사용된다. 최근 CNN 모델은 영상뿐만 아니라, 자연어처리, 벡터로 구성된 매트릭스 등을 입력받아 처리할 수 있다.
- [78]
- [79] 도 6은 학습모델인 인공 신경망에 대한 예이다. 도 6은 CNN 모델에 대한 예이다. 도 6은 표적 서열 특이적인 구조 정보(영상 정보)를 입력받는 모델에 해당한다.
- [80] CNN은 컨볼루션 계층(convolution layer, Conv), 풀링 계층(pooling layer, Pool) 및 전연결 계층(fully connected layer)을 포함한다. 또한, 계층들은 반복적으로 다수가 배치될 수 있다. 도 6의 상단 CNN은 5개의 컨볼루션 계층, 2개의 풀링 계층, 2개의 전연결 계층(Fully connected layer) 구조를 가질 수 있다.

- [81] 컨볼루션 계층은 입력 이미지에 대한 컨볼루션 연산을 통해 특징맵(feature map)을 출력한다. 이때 컨볼루션 연산을 수행하는 필터(filter)를 커널(kernel)이라고도 부른다. 필터의 크기를 필터 크기 또는 커널 크기라고 한다. 커널을 구성하는 연산 파라미터(parameter)를 커널 파라미터(kernel parameter), 필터 파라미터(filter parameter), 또는 가중치(weight)라고 한다.
- [82] 컨볼루션 계층은 컨볼루션 연산과 비선형 연산을 수행한다.
- [83] 컨볼루션 연산은 일정한 크기의 윈도우에서 수행된다. 윈도우는 영상의 좌측 상단에서 우측 하단까지 한 칸씩 이동할 수 있고, 한 번에 이동하는 이동 크기를 조절할 수 있다. 이동 크기를 스트라이드(stride)라고 한다. 컨볼루션 계층은 입력이미지에서 윈도우를 이동하면서 입력이미지의 모든 영역에 대하여 컨볼루션 연산을 수행한다. 컨볼루션 계층은 영상의 가장 자리에 패딩(padding)을 하여 컨볼루션 연산 후 입력 영상의 차원을 유지할 수 있다.
- [84] 비선형 연산 계층(nonlinear operation layer)은 뉴런(노드)에서 출력값을 결정하는 계층이다. 비선형 연산 계층은 전달 함수(transfer function)를 사용한다. 전달 함수는 Relu, sigmoid 함수 등이 있다.
- [85] 풀링 계층(pooling layer)은 컨볼루션 계층에서의 연산 결과로 얻은 특징맵을 서브 샘플링(sub sampling)한다. 풀링 연산은 최대 풀링(max pooling)과 평균 풀링(average pooling) 등이 있다. 최대 풀링은 윈도우 내에서 가장 큰 샘플 값을 선택한다. 평균 풀링은 윈도우에 포함된 값의 평균 값으로 샘플링한다.
- [86] 전연결 계층은 최종적으로 입력 영상을 분류한다. 전연결 계층은 이전 컨볼루션 계층에서 출력하는 값을 모두 입력받아 최종적인 분류를 한다. 도 4에서 전연결 계층은 소프트맥스(softmax) 함수를 사용하여 분류 결과를 출력한다.
- [87]
- [88] 도 7은 학습모델인 인공신경망에 대한 다른 예이다. 도 7은 표적 서열 특이적인 구조 정보와 부가 정보를 입력받아 분석하는 학습모델에 해당한다. 도 7에서 학습모델은 거리 맵(A), 결합 에너지(B) 및 원-핫 벡터(C)를 입력받아 처리한다. 합성곱 신경망(예컨대, CNN)은 도 6과 유사하게 컨볼루션 계층, 풀링 계층 및 전연결 계층으로 구성될 수 있다.
- [89] 거리 맵은 CNN의 최초 컨볼루션 계층에 입력된다. 거리 맵은 컨볼루션 계층과 풀링 계층을 통하여 특징맵으로 변환된다. 특징맵은 전연결 계층에 입력된다. 결합 에너지는 전연결 계층에 입력될 수 있다. 또한, 원-핫 벡터도 전연결 계층에 입력될 수 있다.
- [90] 전연결 계층은 특징맵, 결합 에너지 및 원-핫 벡터를 기준으로 최종적인 분류를 할 수 있다. 예컨대, 전연결 계층은 2차원 특징 맵을 1차원 매트릭스로 변환 후 연산 처리를 하고, 결합 에너지 및 원-핫 벡터 정보를 1차원 매트릭스로 변환 후 특징 맵 뒤에 연결하며 최종적인 분류를 할 수 있다. 즉, 전연결 계층은 특징맵과 부가 정보(결합 에너지 및 원-핫 벡터 중 적어도 하나)를 하나의 1차원

매트릭스로 변환하고, 변환된 1차원 매트릭스 기반하여 최종 분류를 할 수 있다.

[91]

[92] 도 8은 학습모델을 훈련하는 과정에 대한 예이다. 도 8은 학습모델을 훈련하는 시스템(600) 구성을 도시한다.

[93] 훈련데이터를 마련하는 과정을 먼저 설명한다. 훈련데이터는 {입력데이터, 입력데이터에 대한 분류값}으로 구성된 데이터 세트이다. 훈련 데이터를 G 라고 하자. 훈련 데이터 세트가 전체 n 개인 경우, 훈련 데이터 세트 $G = (x_i, y_i)$ 라고 표현할 수 있다. 이때, $0 \leq i < n$ 이다. x_i 는 i 번째 입력데이터이고, y_i 는 입력데이터 x_i 에 대한 분류값이다.

[94] 컴퓨터 단말(50)은 유전자 가위의 서열 데이터를 입력받거나 생성한다. 서열 데이터를 s 라고 표현한다. s_i 는 i 번째 서열 데이터이다.

[95] 구조 예측 장치(610)는 도 2의 구조 예측 장치(120)와 유사한 장치이다. 구조 예측 장치(610)는 입력되는 뉴클레오타이드 서열을 기준으로 해당 뉴클레오타이드 서열의 2차 구조 또는 3차 구조를 예측한다. 구조 예측 장치(610)는 2차 구조 또는 3차 구조에 해당하는 영상을 구조 정보로 생성할 수 있다. 또는, 구조 예측 장치(610)는 2차 구조 또는 3차 구조를 기준으로 연관된 정보를 구조 정보로 생성할 수도 있다. 예컨대, 구조 예측 장치(610)는 구조를 포함하는 영상, 거리 맵, 접촉 맵 등과 같은 형태의 구조 정보를 생성할 수 있다. 또한, 구조 예측 장치(610)는 서열 특이적인 결합 에너지, 원-핫 벡터 등(부가 정보)을 생성할 수도 있다. 구조 예측 장치(610)는 s_i 를 입력받아 x_i 를 출력한다. 구조 예측 장치(610)는 x_i 를 i 에 대한 식별자와 함께 데이터베이스(DB, 630)에 저장할 수 있다.

[96] 점수 연산 장치(620)는 유전자 가위 서열을 기준으로 특정 표적에 대한 유전자 가위의 효과를 특정 점수로 산출한다. 유전자 가위 효과를 나타내는 점수를 효과 점수(effect score)라고 명명한다. 효과 점수는 공지된 솔루션을 이용하여 연산할 수 있다. 예컨대, GenomeCRISPR_full05112017

(<https://www.dkfz.de/signaling/crispr-downloads/GENOMECRISPR/>)과 같은 솔루션을 이용할 수 있다. 또는 유전자 가위 관련 실험 정보를 관리하는 DB를 활용하여 특정 표적에 대한 특정 서열의 유전자 가위 효과를 추정할 수도 있다. 효과 점수는 특정 표적에 대한 특정 유전자 가위의 효과를 정량적으로 평가한 값이다. 효과 점수는 동일한 표적이라도, 유전자 가위를 구성하는 서열에 따라 달라진다. 점수 연산 장치(620)는 s_i 를 입력받아 y_i 를 출력한다. 점수 연산 장치(620)는 y_i 를 i 에 대한 식별자와 함께 데이터베이스(DB, 630)에 저장할 수 있다.

[97] 구조 예측 장치(610) 및 점수 연산 장치(620)는 일정한 데이터 처리가 가능한 연산장치이다. 예컨대, 구조 예측 장치(610) 및 점수 연산 장치(620)는 서버, PC, 스마트 기기 등과 같은 장치일 수 있다.

[98] 데이터베이스(630)는 n 개의 서열에 대한 훈련데이터 세트를 저장한다고

가정한다.

- [99] 훈련 장치(640)는 훈련데이터를 이용하여 학습모델을 훈련한다. 훈련 장치(640)는 입력데이터 x_i 에 대하여 학습모델이 y_i 라는 값을 출력하도록 학습모델을 반복적으로 훈련한다. CNN 경우 목적 함수를 최소화하는 방향으로 학습된다. 학습 과정은 CNN 모델에서 사용하는 가중치를 최적화하는 과정에 해당한다. 예컨대, 가중치 최적화는 경사 하강법(gradient descent method)을 이용할 수 있다.
- [100] 훈련이 제대로 끝나면, 학습모델은 x_i 라는 구조 정보에 대하여 y_i 라는 값을 출력하게 된다. 효과 점수 y_i 는 0 또는 1 중 어느 하나의 값일 수 있다. 또는 효과 점수 y_i 는 일정한 범위 내의 정수 내지 실수 값일 수도 있다.
- [101] 도 8은 유전자 가위 서열을 기준으로 구조를 예측하는 장치(610) 및 유전자 가위 서열을 기준으로 효과 점수를 연산하는 장치(620)를 별도로 도시하였다. 그러나, 사용자 단말(50)과 같은 하나의 장치가 유전자 가위 서열을 기준으로 구조 정보를 생성하고, 해당 유전자 가위 서열에 대한 효과 점수를 연산하여 훈련 데이터를 생성할 수 있다.
- [102] 예컨대, 사용자 단말(50)은 RNA 구조를 예측하는 웹 솔루션(예컨대, RNAfoldWebServer, <http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi>)을 이용하거나, 설치된 프로그램을 이용하여 유전자 가위 서열 또는 표적 특이적 서열에 대한 2차 구조/3차 구조를 예측할 수 있다. 사용자 단말(50)은 예측된 구조를 기준으로 거리 맵과 같은 구조 정보를 생성할 수 있다. 또한, 사용자 단말(50)은 GenomeCRISPR_full05112017과 같은 솔루션을 이용하여 효과 점수를 연산할 수도 있다.
- [103] 도 8은 사용자 단말(50)과 훈련 장치(640)를 별도의 객체로 도시하였다. 다만, 훈련 데이터를 마련하는 객체와 학습모델을 훈련하는 객체가 동일할 수도 있다.
- [104] 한편, 학습모델은 특정한 표적 서열에 대한 유전자 가위 서열의 효과 내지 적합도를 평가한다. 따라서, 학습모델은 표적 서열마다 서로 다른 모델로 사전에 학습되는 것이 바람직하다.
- [105]
- [106] 도 9는 유전자 가위 효과를 예측하는 분석장치(700)에 대한 예이다. 분석장치(700)는 도 2의 분석 장치(110, 210 또는 310)에 해당하는 장치이다.
- [107] 분석장치(700)는 전술한 학습모델을 이용하여 특정 서열을 갖는 유전자 가위의 효과를 예측한다. 분석장치(700)는 물리적으로 다양한 형태로 구현될 수 있다. 예컨대, 분석장치(700)는 PC와 같은 컴퓨터 장치, 네트워크의 서버, 영상 처리 전용 칩셋 등의 형태를 가질 수 있다. 컴퓨터 장치는 스마트 기기 등과 같은 모바일 기기를 포함할 수 있다.
- [108] 분석장치(700)는 저장장치(710), 메모리(720), 연산장치(730), 인터페이스 장치(740), 통신장치(750) 및 출력장치(760)를 포함한다.
- [109] 저장장치(710)는 유전자 가위의 효과를 예측하는 신경망 모델을 저장한다.

신경망 모델은 사전에 학습되어야 한다. 저장장치(710)는 RNA 서열을 기준으로 2차 구조 또는 3차 구조를 예측하는 프로그램을 저장할 수 있다. 저장장치(710)는 2차 구조 또는 3차 구조를 기준으로 구조 정보를 생성하는 프로그램을 저장할 수도 있다. 나아가 저장장치(710)는 데이터 처리에 필요한 다른 프로그램 내지 소스 코드 등을 저장할 수 있다. 저장장치(710)는 입력되는 유전자 가위 서열 및 분석 결과(유전자 가위 효과)를 저장할 수 있다.

[110] 메모리(720)는 분석장치(700)가 수신한 데이터를 분석하는 과정에서 생성되는 데이터 및 정보 등을 저장할 수 있다.

[111] 인터페이스 장치(740)는 외부로부터 일정한 명령 및 데이터를 입력받는 장치이다. 인터페이스 장치(740)는 물리적으로 연결된 입력 장치 또는 외부 저장장치로부터 유전자 가위 서열 또는 가이드 RNA 서열을 입력받을 수 있다. 인터페이스 장치(740)는 데이터 분석을 위한 학습모델을 입력받을 수 있다. 인터페이스 장치(740)는 학습모델 훈련을 위한 학습데이터, 정보 및 파라미터값을 입력받을 수도 있다.

[112] 통신장치(750)는 유선 또는 무선 네트워크를 통해 일정한 정보를 수신하고 전송하는 구성을 의미한다. 통신장치(750)는 외부 객체로부터 유전자 가위 서열 또는 가이드 RNA 서열을 수신할 수 있다. 통신장치(750)는 모델 학습을 위한 데이터도 수신할 수 있다. 통신장치(750)는 입력된 서열에 대하여 결정된 분석 결과를 외부 객체로 송신할 수 있다.

[113] 통신장치(750) 내지 인터페이스 장치(740)는 외부로부터 일정한 데이터 내지 명령을 전달받는 장치이다. 통신장치(750) 내지 인터페이스 장치(740)를 입력장치라고 명명할 수 있다.

[114] 입력 장치는 분석 대상인 유전자 가위에 대한 구조 정보를 입력 내지 수신받을 수 있다. 예컨대, 입력 장치는 외부 서버나 DB로부터 구조 정보를 수신할 수 있다. 입력 장치는 분석 대상인 유전자 가위에 대한 2차 구조 또는 3차 구조 정보를 수신할 수도 있다.

[115] 출력장치(760)는 일정한 정보를 출력하는 장치이다. 출력장치(760)는 데이터 처리 과정에 필요한 인터페이스, 분석 결과 등을 출력할 수 있다.

[116] 연산 장치(730)는 저장장치(710)에 저장된 프로그램을 이용하여 유전자 가위 서열에 대한 2차 구조 또는 3차 구조를 예측할 수 있다. 연산 장치(730)는 RNA 2차 구조 또는 3차 구조 영상을 생성할 수 있다. 또한, 연산 장치(730)는 RNA 2차 구조 또는 3차 구조를 기준으로 거리 맵과 같은 구조 정보를 생성할 수도 있다. 즉, 연산 장치(730)는 학습모델이 입력받을 수 있는 입력 데이터를 생성할 수 있다. 연산 장치(730)는 필요한 경우 데이터 전처리를 할 수도 있다.

[117] 연산 장치(730)는 생성한 입력 데이터(구조 정보)를 학습모델에 입력하여, 분석 대상인 유전자 가위 서열에 대한 효과를 예측할 수 있다. 연산 장치(730)는 효과에 대한 이진 분류 또는 다중 분류 결과를 기준으로 유전자 가위 서열에 대한 효과를 예측할 수 있다.

[118] 한편, 연산 장치(730)는 주어진 훈련 데이터를 이용하여 유전자 가위 서열의 효과 예측에 사용되는 학습모델을 훈련할 수도 있다.

[119] 연산 장치(730)는 데이터를 처리하고, 일정한 연산을 처리하는 프로세서, AP, 프로그램이 임베디드된 칩과 같은 장치일 수 있다.

[120]

[121] 이하 기술한 학습모델에 대한 성능을 검증하는 실험 결과를 설명한다.

[122] 먼저, 실제 학습모델 훈련에 사용한 데이터에 대하여 설명한다. 훈련 데이터는 GenomeCRISPR_full05112017 데이터 (이하, 데이터 셋)이다. 상기 데이터 셋은 3,800 만개의 유전자 가위 서열들로 구성되어 있다. 데이터 셋 내 유전자 가위 서열들은 고유한 유전자 가위 서열이 아니고, 다수의 세포주에 중복 실험된 유전자 가위 서열들이다. 하나의 유전자 가위는 세포주마다 다른 효과를 보인다.

[123] 데이터 셋은 다음과 같은 통계 처리를 통하여 학습이 용이한 데이터 셋으로 준비하였다. 연구자는 최초 데이터 셋에서 30개 미만의 세포주에 실험된 서열을 제외하였다. 이는 유전자 가위 서열의 고유성을 유지하며 잘림 유무를 예측하고자 일부 데이터를 필터링한 것이다. 또한, 연구자는 80% 이상의 세포주에서 효과 점수 5점 이상인 유전자 가위 서열을 효과 좋음(잘 잘림)으로 정의하고, 80% 이상의 세포주에서 효과 점수가 -5점 이하인 유전자 가위 서열을 효과 낮음(안잘림)으로 정의하였다. 필터링된 데이터 셋은 25,593개의 유전자 가위 서열로 구성되었다. 필터링된 데이터 셋은 효과 좋음으로 라벨링된 데이터들 12,481개 및 효과 낮음으로 라벨링된 데이터들 13,112개로 구성되었다.

[124] 학습모델은 기술한 도 7과 같은 모델을 구현하였다. 학습모델은 이진 분류에 따른 효과 예측을 하는 모델이다. 학습모델은 입력 데이터가 서로 다른 4가지 학습모델을 구현하였다. 4가지 학습모델은 아래의 표 1과 같다.

[125] [Table 1]

모델 구분	입력 데이터
제1 모델	2차원 거리 맵
제2 모델	(i) 2차원 거리 맵, (ii) 결합 에너지
제3 모델	(i) 2차원 거리 맵, (ii) 원-핫 벡터
제4 모델	(i) 2차원 거리 맵, (ii) 결합 에너지, (iii) 원-핫 벡터

[126] 4개의 학습모델은 모두 k-폴드 교차검증(k-fold cross validation)을 수행하였다. k는 5로 설정하였다. 학습 데이터는 랜덤으로 섞어서 5개로 나누었다. 5개 중 4개는 학습용으로 사용하였고, 나머지 1개는 검증용 데이터로 사용하였다.

[127] 다음 폴드에서 검증용 데이터는 이전 폴드와 다른 데이터로 변경하였고, 학습 데이터는 검증용 데이터로 선택된 하나의 데이터를 제외한 나머지 데이터를 사용하였다. 결국, 5번의 검증 과정에서, 학습모델은 서로 다른 조합의 학습 데이터들을 사용하여 학습되었고, 서로 다른 데이터를 사용하여 검증되었다.

나아가, 5-폴드 교차 검증을 30번 반복 수행하였다.

[128] 학습모델의 성능 평가 측도는 정확도 (accuracy), ROC (Receiver Operating Characteristics)의 곡선 아래 면적 (Area Under the Curve, AUC) 등을 사용하였다. 성능은 실험 횟수 (iterations) 당 5-폴드 교차검증에 의해 생성된 5개의 데이터 셋의 평균 성능의 분포를 기준으로 분석하였다.

[129] ROC 곡선을 나타내는 그래프는 x축이 분류 모형의 위양성률 (false positive rate, FPR)이고, y축이 진양성률 (true positive rate, TPR)이다. TPR은 민감도(sensitivity)를 나타내고, FPR은 '1-특이도(specificity)'를 나타낸다.

[130] 이진 분류(binary classification) 문제에 대해, AUC는 분류기의 양성 데이터(positive object)에 대한 예측 값이 음성데이터(negative object)에 대한 예측 값보다 높을 확률을 의미한다. AUC는 ROC 곡선의 아래 면적을 구함으로써 추정 가능하다.

[131] 정확도는 모델 예측 값과 데이터 y값을 비교하며 전체 데이터에 대하여 분류가 일치하는 비율을 나타낸다. 정확도(ACC)는 아래 수학적식으로 표현할 수 있다. TP는 True positive 개수, TN은 True negative 개수, FP는 False positive 개수이고, FN은 False negative 개수이다.

[132]
$$ACC = \frac{TP+TN}{TP+TN+FP+FN}$$

[133] 도 10은 유전자 가위 효과 예측한 모델 성능 평가에 대한 예이다. 도 10은 학습모델에 대하여 에폭(epoch) 당 정확도를 나타낸다. 도 10은 5-폴드 교차 검증을 30번 반복(전체 학습을 150번)한 결과의 평균을 나타낸다. 도 10(A)는 학습 데이터를 이용하여 에폭 당 정확도를 나타낸 결과이고, 도 10(B)는 검증 데이터를 이용하여 에폭 당 정확도를 나타낸 결과이다. 도 10의 결과를 살펴보면, 제4 모델(거리맵 + 결합 에너지 + 원-핫 벡터)이 4개의 학습모델 중 가장 정확도가 높았다. 정확도 순서는 제4 모델 > 제2 모델(거리맵 + 결합 에너지) > 제3 모델(거리맵 + 원-핫 벡터) > 제1 모델(거리맵)이다. 전체적으로 학습모델은 입력데이터로 거리맵과 함께 에너지가 사용될 때 비교적 높은 성능을 보였다.

[134]

[135] 도 11은 유전자 가위 효과 예측한 모델 성능 평가에 대한 다른 예이다. 도 11은 5-폴드의 평균 AUC 값을 30번 반복적으로 구한 분포도를 박스 플롯(box plot)으로 나타낸다. 도 11(A)는 훈련 데이터를 사용하여 분석한 결과이다. 도 11(B)는 검증 데이터를 사용하여 분석한 결과이다. 검증 데이터 기준으로 AUC의 중간 값을 살펴보면, 에너지를 포함하는 학습모델은 약 0.93, 포함하지 않은 학습모델은 약 0.922이다. 따라서, 전체적으로 에너지를 포함한 모델 성능이 포함하지 않은 모델보다 성능이 좋다고 할 수 있다.

[136] 도 12는 유전자 가위 효과 예측한 모델 성능 평가에 대한 또 다른 예이다. 도 12는 학습모델에 대한 TPR과 TNR 기준의 성능을 나타내는 박스 플롯이다. 도

12(A)는 임계값 0.5 기준으로 TPR의 성능을 평가한 결과이고, 도 12(B)는 임계값 0.5 기준으로 TNR의 성능을 평가한 결과이다. 도 12를 살펴보면, 입력 데이터로 에너지를 포함한 모델이 에너지를 포함하지 않은 모델보다 성능이 더 높다는 것을 알 수 있다.

[137]

[138] 또한, 상술한 바와 같은 학습모델 훈련 방법 및 학습모델을 이용한 유전자 가위 효과 분석 방법은 컴퓨터에서 실행될 수 있는 실행가능한 알고리즘을 포함하는 프로그램(또는 어플리케이션)으로 구현될 수 있다. 상기 프로그램은 비일시적 판독 가능 매체(non-transitory computer readable medium)에 저장되어 제공될 수 있다.

[139] 비일시적 판독 가능 매체란 레지스터, 캐쉬, 메모리 등과 같이 짧은 순간 동안 데이터를 저장하는 매체가 아니라 반영구적으로 데이터를 저장하며, 기기에 의해 판독(reading)이 가능한 매체를 의미한다. 구체적으로는, 상술한 다양한 어플리케이션 또는 프로그램들은 CD, DVD, 하드 디스크, 블루레이 디스크, USB, 메모리카드, ROM 등과 같은 비일시적 판독 가능 매체에 저장되어 제공될 수 있다.

[140]

[141] 본 실시례 및 본 명세서에 첨부된 도면은 전술한 기술에 포함되는 기술적 사상의 일부를 명확하게 나타내고 있는 것에 불과하며, 전술한 기술의 명세서 및 도면에 포함된 기술적 사상의 범위 내에서 당업자가 용이하게 유추할 수 있는 변형 예와 구체적인 실시례는 모두 전술한 기술의 권리범위에 포함되는 것이 자명하다고 할 것이다.

[142]

[143]

청구범위

- [청구항 1] 분석장치가 유전자 가위의 서열 데이터를 입력받는 단계;
상기 분석장치가 상기 서열 데이터를 이용하여 유전자 가위의 표적 특이적 서열이 구성하는 구조에 대한 구조 정보를 생성하는 단계;
상기 분석장치가 상기 구조 정보를 학습모델에 입력하는 단계; 및
상기 분석장치가 상기 학습모델이 출력하는 정보를 기준으로 상기 유전자 가위의 효과를 평가하는 단계를 포함하는 학습모델을 이용하여 유전자 가위 효과를 분석하는 방법.
- [청구항 2] 제1항에 있어서,
상기 표적 특이적 서열은 상기 유전자 가위를 구성하는 가이드 RNA 전체 또는 일부 서열인 학습모델을 이용하여 유전자 가위 효과를 분석하는 방법.
- [청구항 3] 제1항에 있어서,
상기 분석장치는 뉴클레오타이드 서열을 기준으로 구조를 예측하는 프로그램을 이용하여 상기 표적 특이적 서열에 대한 2차 구조 또는 3차 구조를 예측하고, 상기 2차 구조 또는 3차 구조에 대한 상기 구조 정보를 생성하는 학습모델을 이용하여 유전자 가위 효과를 분석하는 방법.
- [청구항 4] 제1항에 있어서,
상기 구조 정보는 상기 표적 특이적 서열의 2차 구조 또는 3차 구조에 대한 영상인 학습모델을 이용하여 유전자 가위 효과를 분석하는 방법.
- [청구항 5] 제1항에 있어서,
상기 구조 정보는 상기 표적 특이적 서열의 2차 구조 또는 3차 구조에서 한 쌍의 뉴클레오타이드 사이의 거리 정보를 시각적으로 표현한 거리 맵(distance map)인 학습모델을 이용하여 유전자 가위 효과를 분석하는 방법.
- [청구항 6] 제1항에 있어서,
상기 분석장치는 상기 서열 데이터를 이용하여 결합 에너지(hybrid duplexes energy) 및 원-핫 벡터(one-hot vector) 중 적어도 하나를 포함하는 부가 정보를 생성하고, 상기 학습모델에 상기 부가 정보를 더 입력하는 학습모델을 이용하여 유전자 가위 효과를 분석하는 방법.
- [청구항 7] 제6항에 있어서,
상기 학습모델은 컨볼루션 계층, 풀링 계층 및 전연결 계층을 포함하는 인공신경망 모델이고,

상기 컨볼루션 계층과 상기 풀링 계층은 상기 구조 정보를 특징 맵으로 변환하고,

상기 전연결 계층은 상기 특징 맵과 상기 부가 정보를 입력받아 상기 표적 특이적 서열을 포함하는 유전자 가위에 대한 효과 정보를 출력하는 학습모델을 이용하여 유전자 가위 효과를 분석하는 방법.

[청구항 8]

유전자 가위의 서열 데이터를 입력받는 입력장치;

뉴클레오타이드 서열 기준으로 해당 서열이 구성하는 구조에 대한 구조 정보를 생성하는 프로그램 및 유전자 가위의 구조 정보를 기준으로 유전자 가위의 효과를 분석하는 학습모델을 저장하는 저장장치; 및

상기 프로그램을 이용하여 상기 서열 데이터에서 유전자 가위의 표적 특이적 서열에 대한 구조 정보를 생성하고, 생성된 구조 정보를 상기 학습모델에 입력하여 유전자 가위의 효과를 평가하는 연산장치를 포함하는 유전자 가위 효과를 예측하는 분석장치.

[청구항 9]

.제8항에 있어서,

상기 표적 특이적 서열은 상기 유전자 가위를 구성하는 가이드 RNA 전체 또는 일부 서열인 유전자 가위 효과를 예측하는 분석장치.

[청구항 10]

제8항에 있어서,

상기 구조 정보는 상기 표적 특이적 서열의 2차 구조 또는 3차 구조에서 한 쌍의 뉴클레오타이드 사이의 거리 정보를 시각적으로 표현한 거리 맵(distance map)인 유전자 가위 효과를 예측하는 분석장치.

[청구항 11]

제8항에 있어서,

상기 저장장치는 뉴클레오타이드 서열에 대한 결합 에너지(hybrid duplexes energy)를 연산하는 프로그램을 더 저장하고,

상기 연산장치는 상기 서열 데이터를 이용하여 결합 에너지를 연산하고, 상기 결합 에너지를 상기 학습모델에 더 입력하는 유전자 가위 효과를 예측하는 분석장치.

[청구항 12]

제8항에 있어서,

상기 연산장치는 상기 서열 데이터에 대한 원-핫 인코딩(one-hot encoding)을 하여 원-핫 벡터를 생성하고, 상기 원-핫 벡터를 상기 학습모델에 더 입력하는 유전자 가위 효과를 예측하는 분석장치.

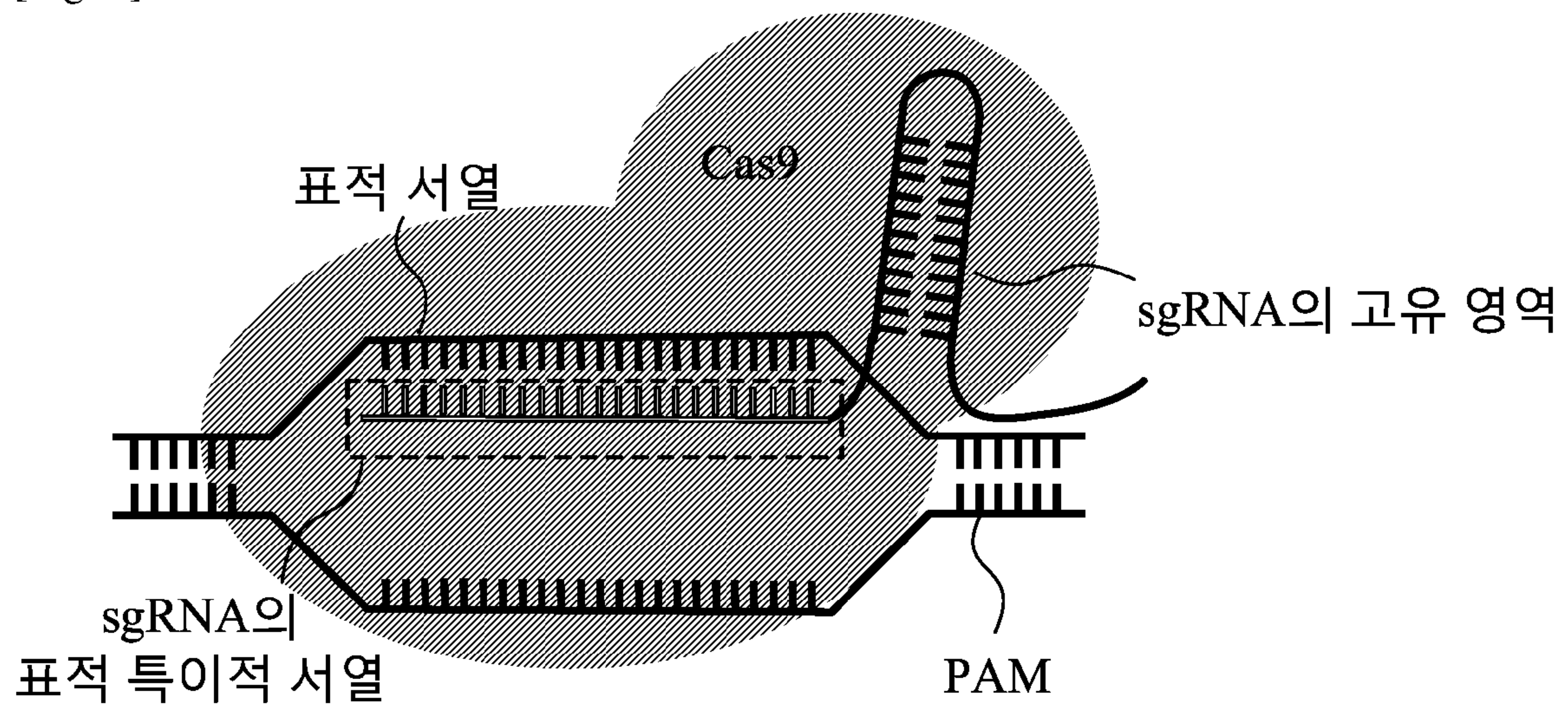
[청구항 13]

제8항에 있어서,

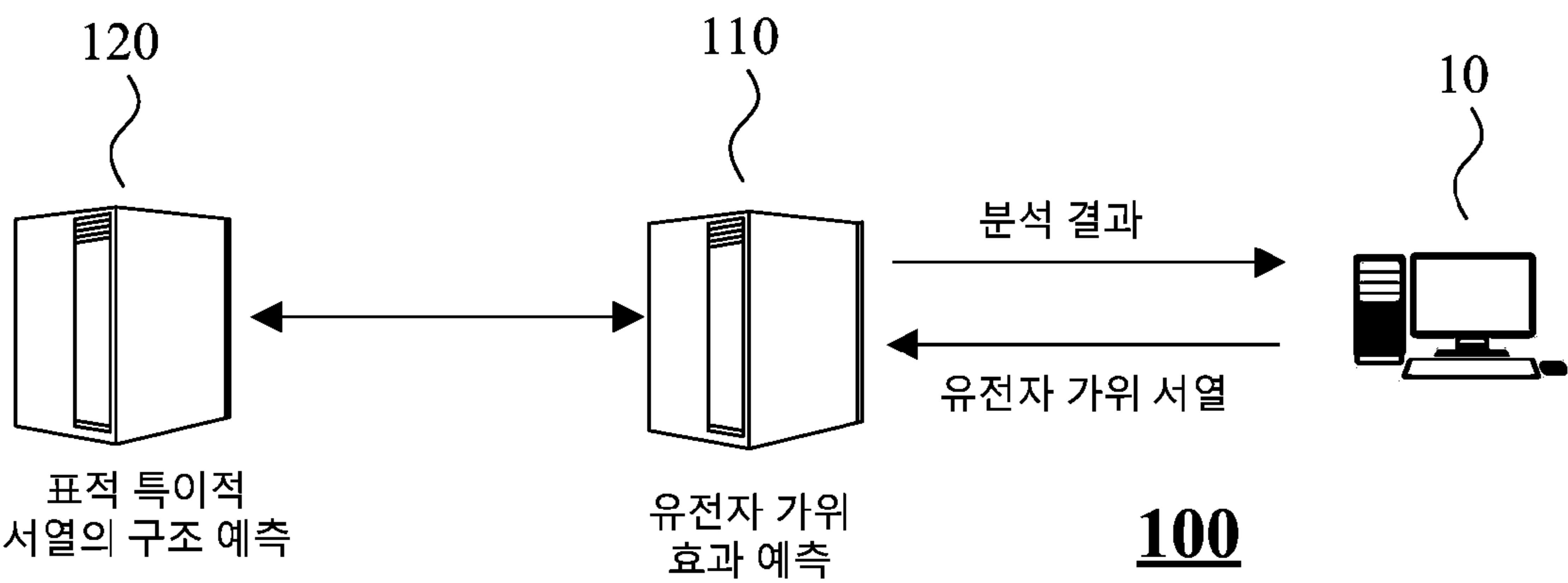
상기 연산장치는 상기 서열 데이터를 이용하여 결합 에너지(hybrid duplexes energy) 및 원-핫 벡터(one-hot vector) 중 적어도 하나를 포함하는 부가 정보를 생성하고, 상기 학습모델에 상기 부가

- [청구항 14] 정보를 더 입력하는 유전자 가위 효과를 예측하는 분석장치.
제13항에 있어서,
상기 학습모델은 컨볼루션 계층, 풀링 계층 및 전연결 계층을 포함하는 인공신경망 모델이고,
상기 컨볼루션 계층과 상기 풀링 계층은 상기 구조 정보를 특징 맵으로 변환하고,
상기 전연결 계층은 상기 특징 맵과 상기 부가 정보를 입력받아 상기 표적 특이적 서열을 포함하는 유전자 가위에 대한 효과 정보를 출력하는 유전자 가위 효과를 예측하는 분석장치.
- [청구항 15] 컴퓨터에서 제1항 내지 제7항 중 어느 하나의 항에 학습모델을 이용하여 유전자 가위를 설계하는 방법을 실행하기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록 매체.
- [청구항 16] 뉴클레오타이드 서열 기준으로 해당 서열이 구성하는 구조에 대한 구조 정보를 생성하는 구조 예측 서버; 및
유전자 가위의 구조 정보를 기준으로 유전자 가위의 효과를 분석하는 학습모델을 저장하고, 상기 구조 예측 서버로부터 유전자 가위의 서열 데이터에서 유전자 가위의 표적 특이적 서열에 대한 구조 정보를 수신하고, 상기 구조 정보를 상기 학습모델에 입력하여 유전자 가위의 효과를 평가하는 평가 서버를 포함하는 유전자 가위 효과를 예측하는 시스템.
- [청구항 17] 제16항에 있어서,
상기 구조 정보는 상기 표적 특이적 서열의 2차 구조 또는 3차 구조에 대한 영상 또는 상기 표적 특이적 서열의 2차 구조 또는 3차 구조에서 한 쌍의 뉴클레오타이드 사이의 거리 정보를 시각적으로 표현한 거리 맵(distance map)인 유전자 가위 효과를 예측하는 시스템.

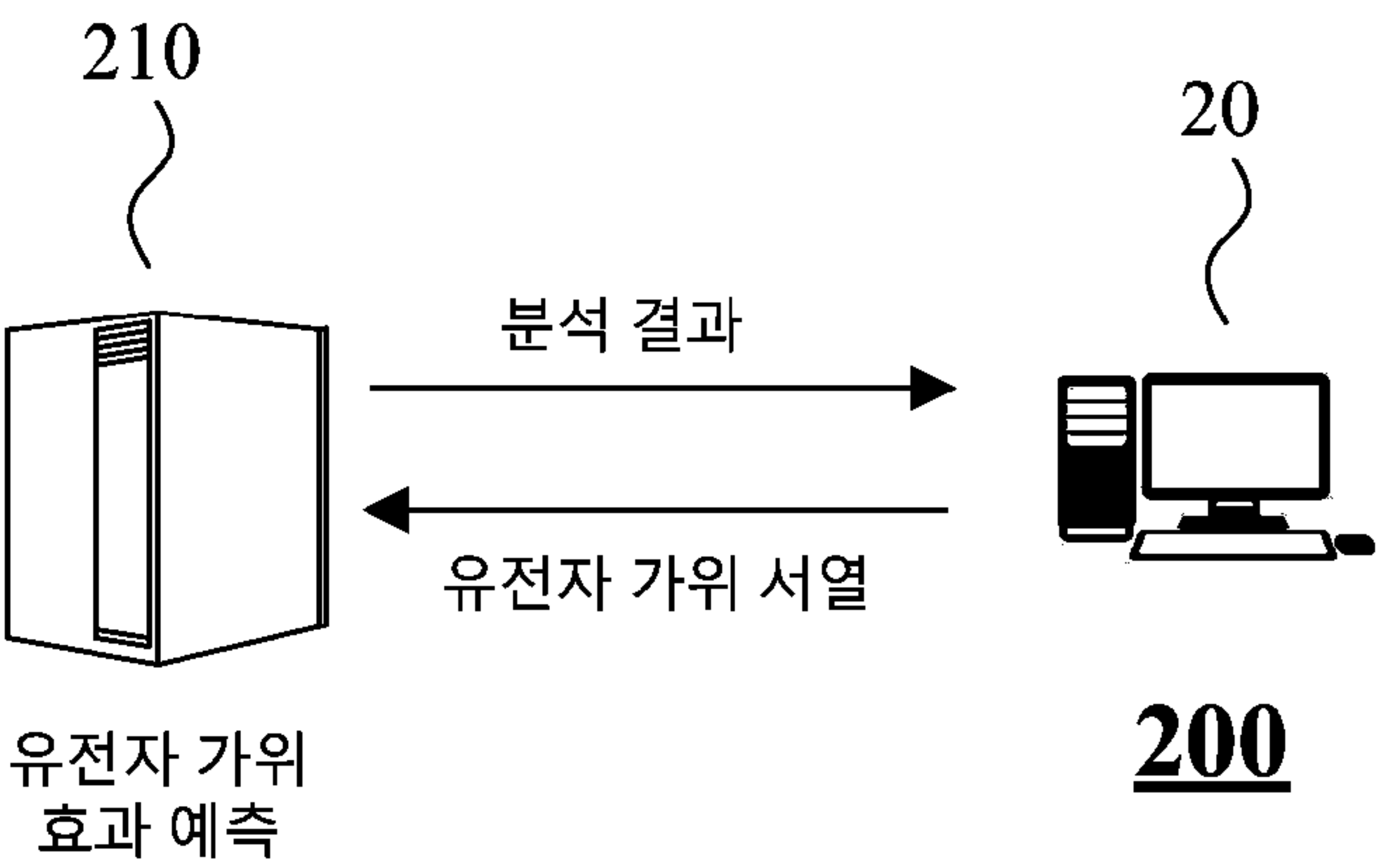
[Fig. 1]



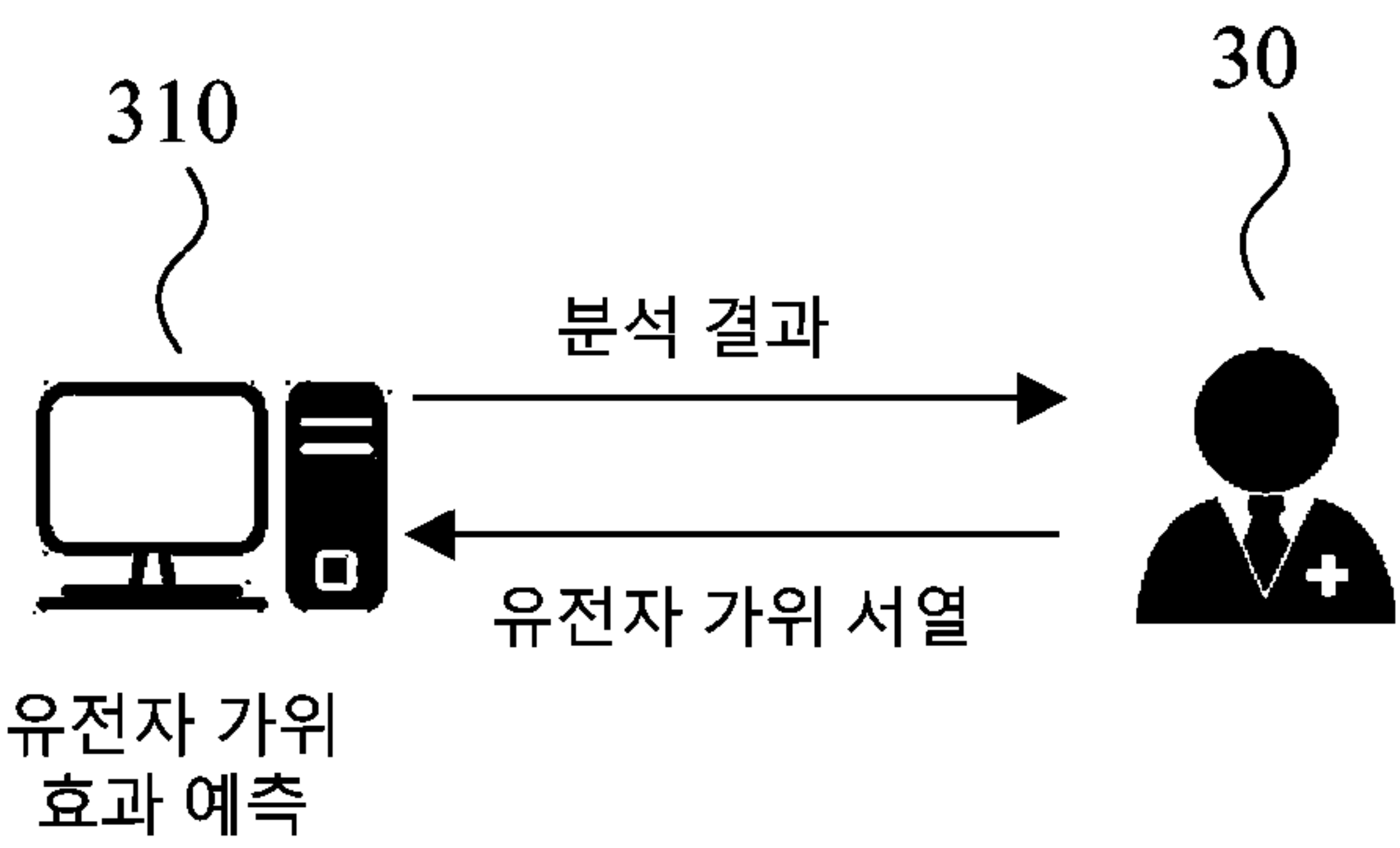
[Fig. 2]



(A)

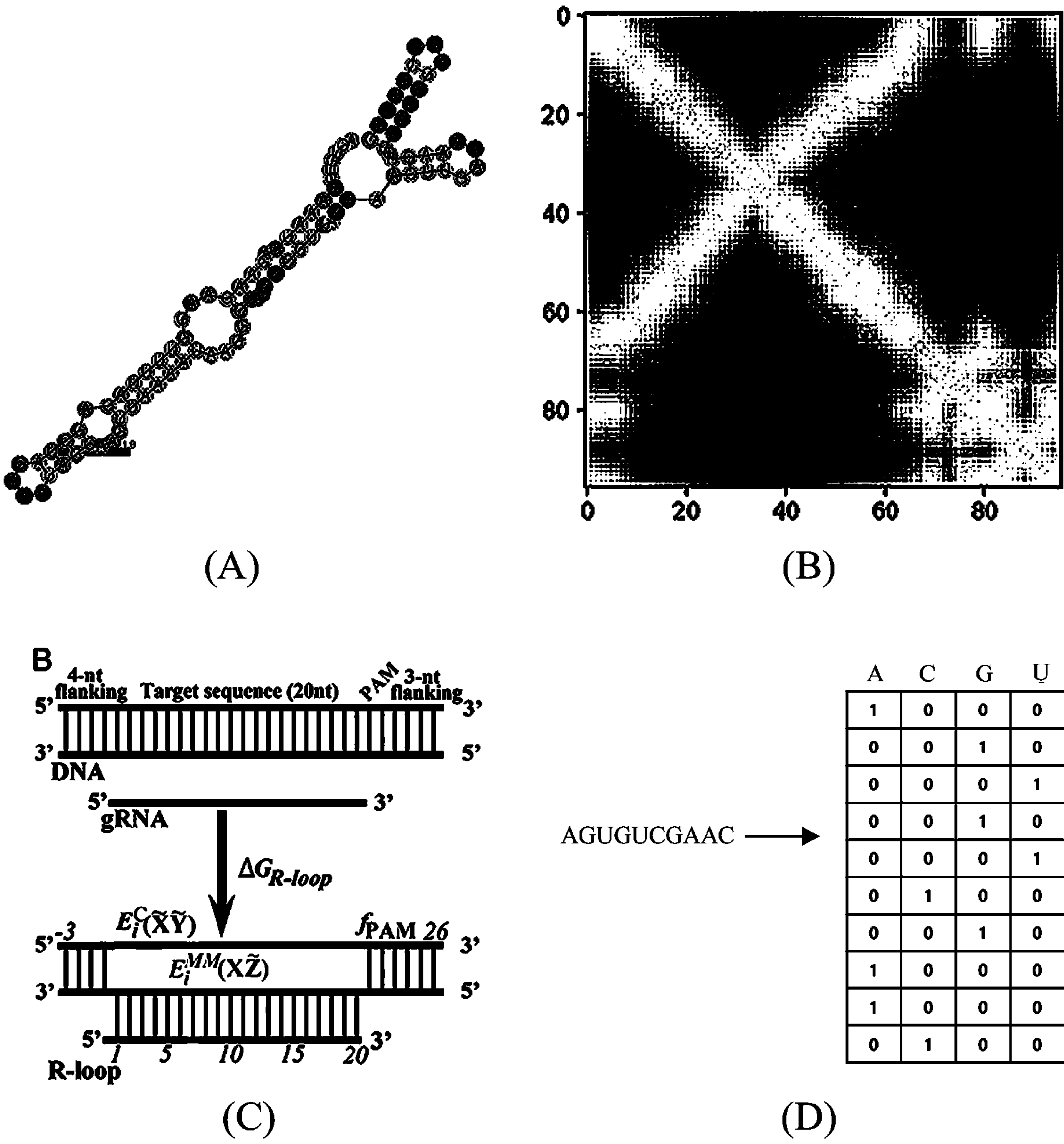


(B)



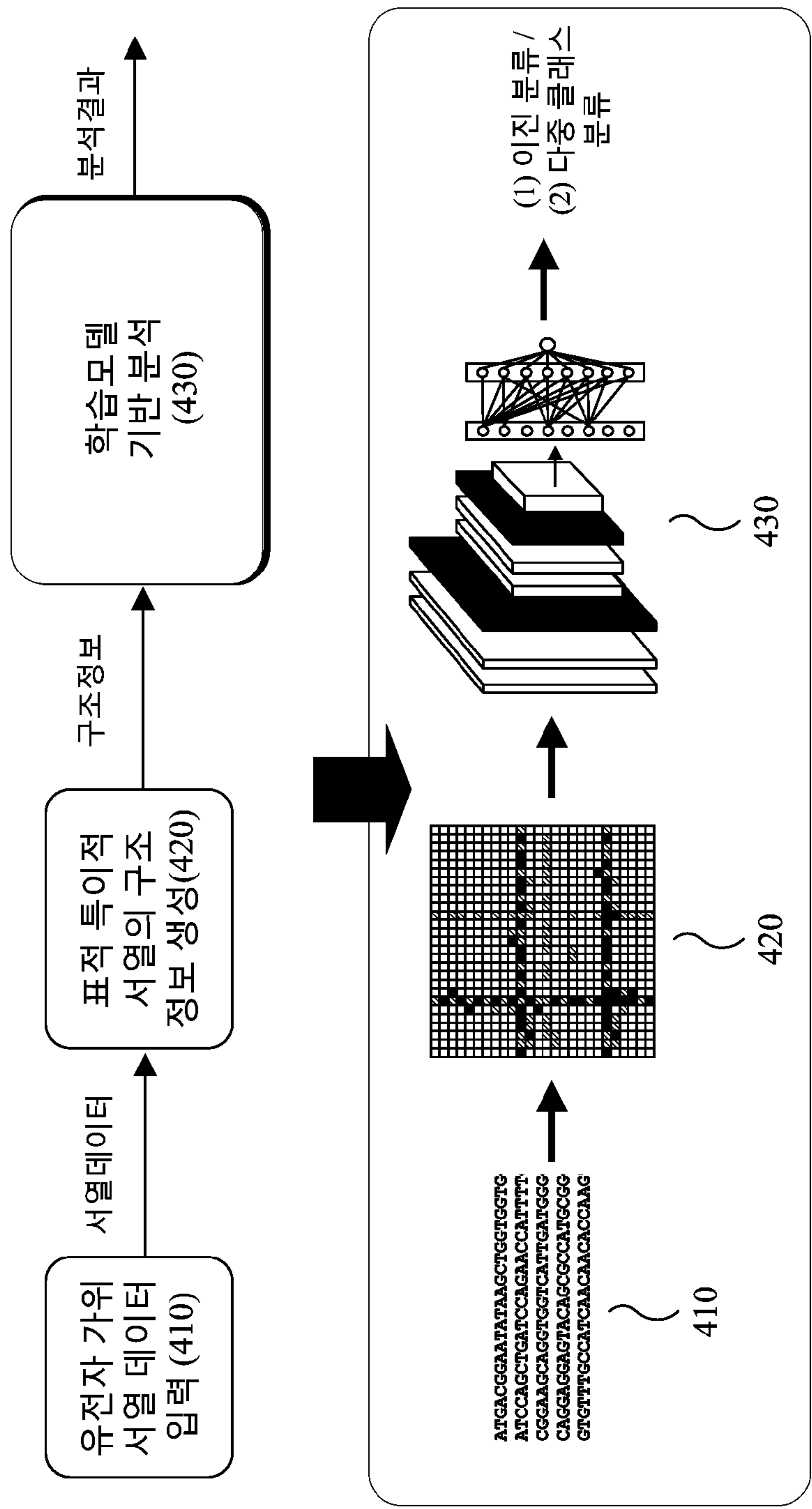
(C)

[Fig. 3]



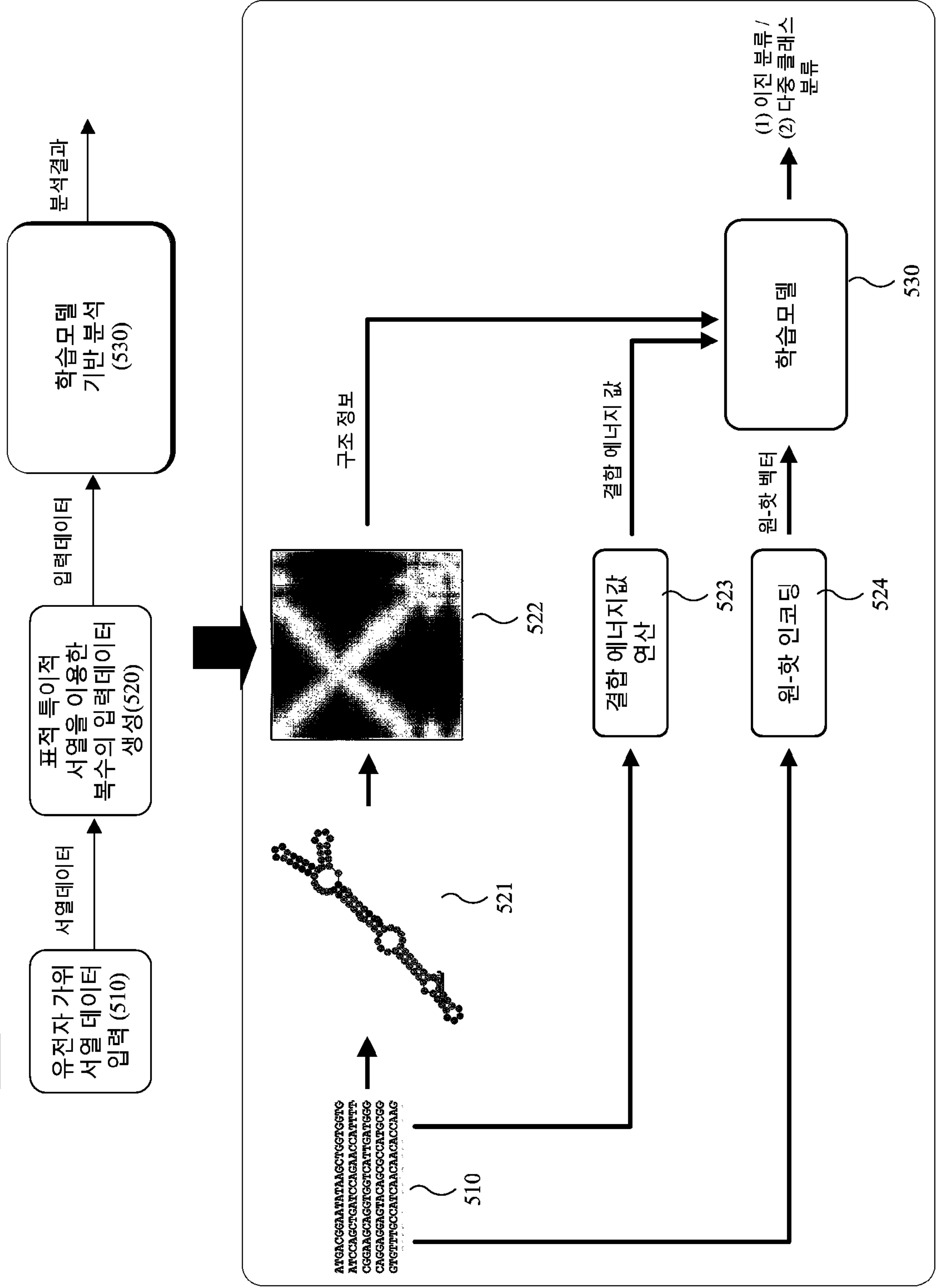
[Fig. 4]

400

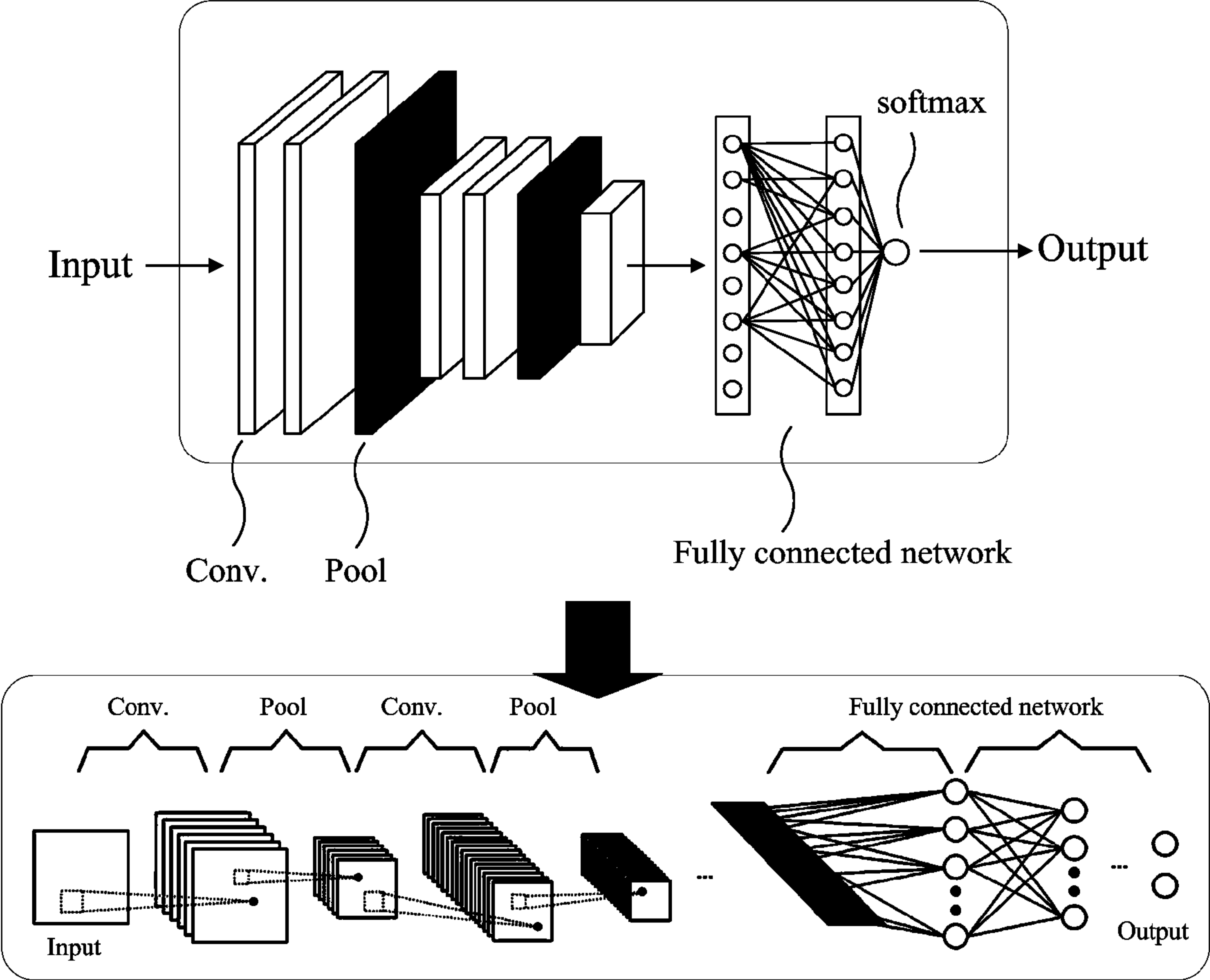


[Fig. 5]

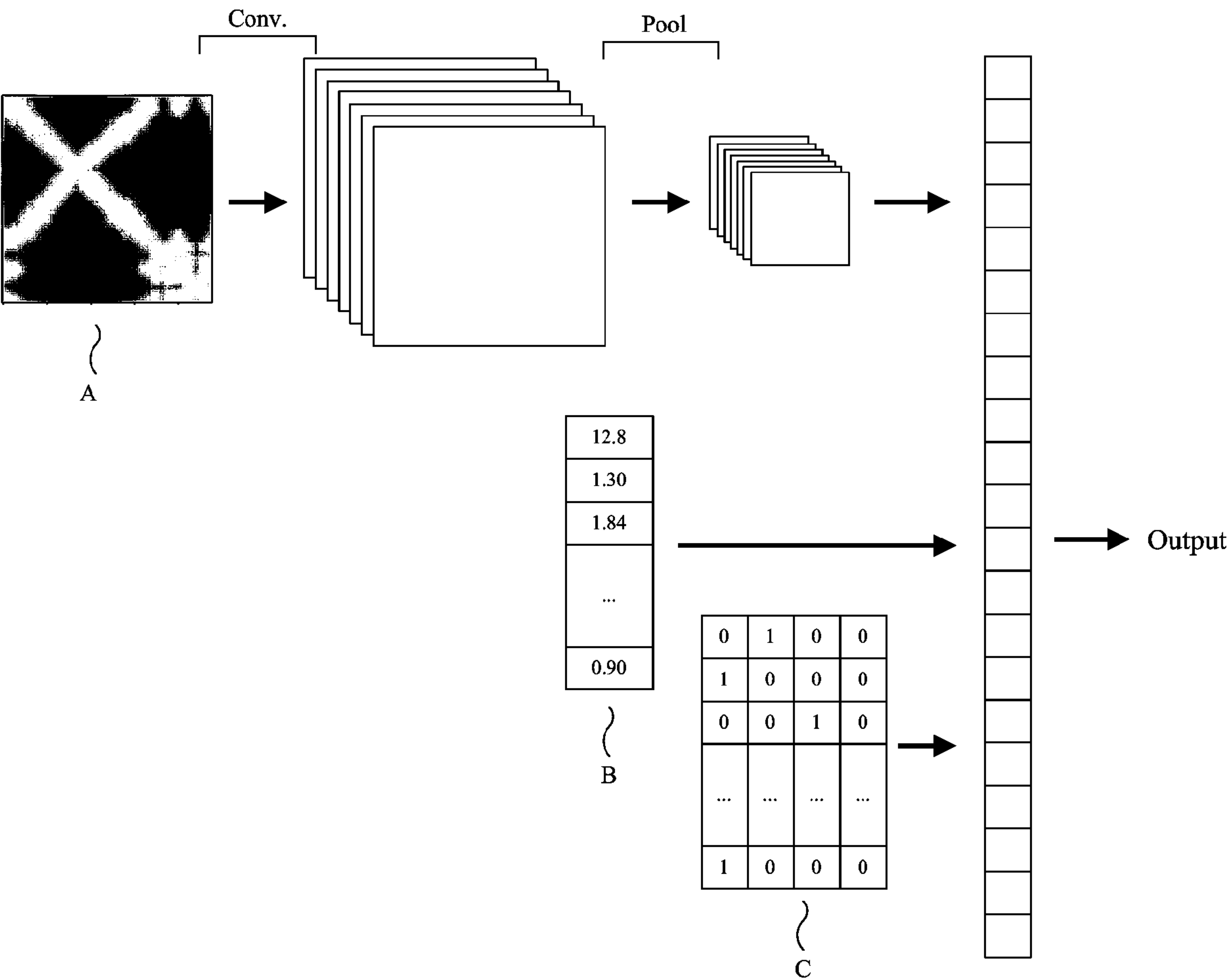
500



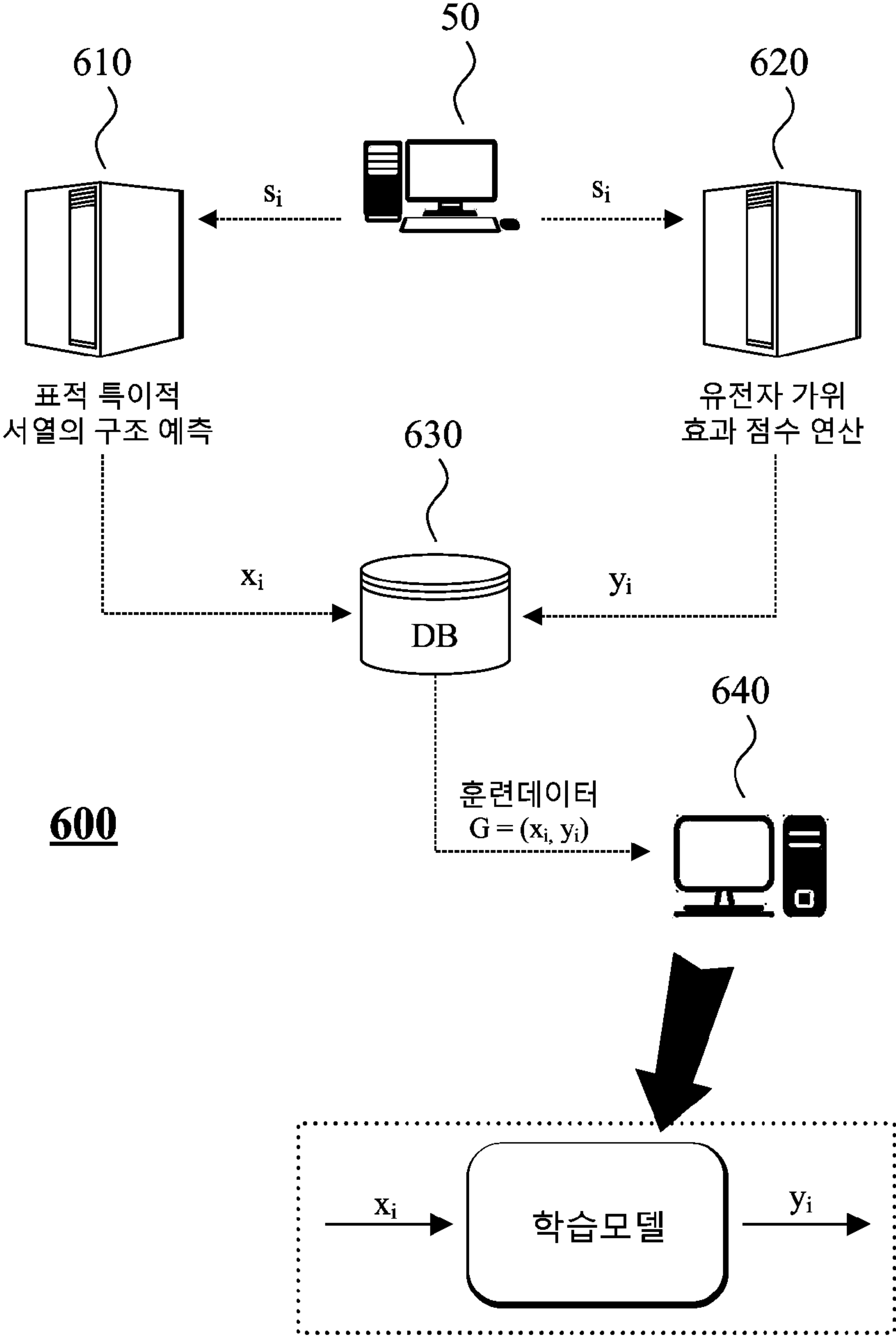
[Fig. 6]



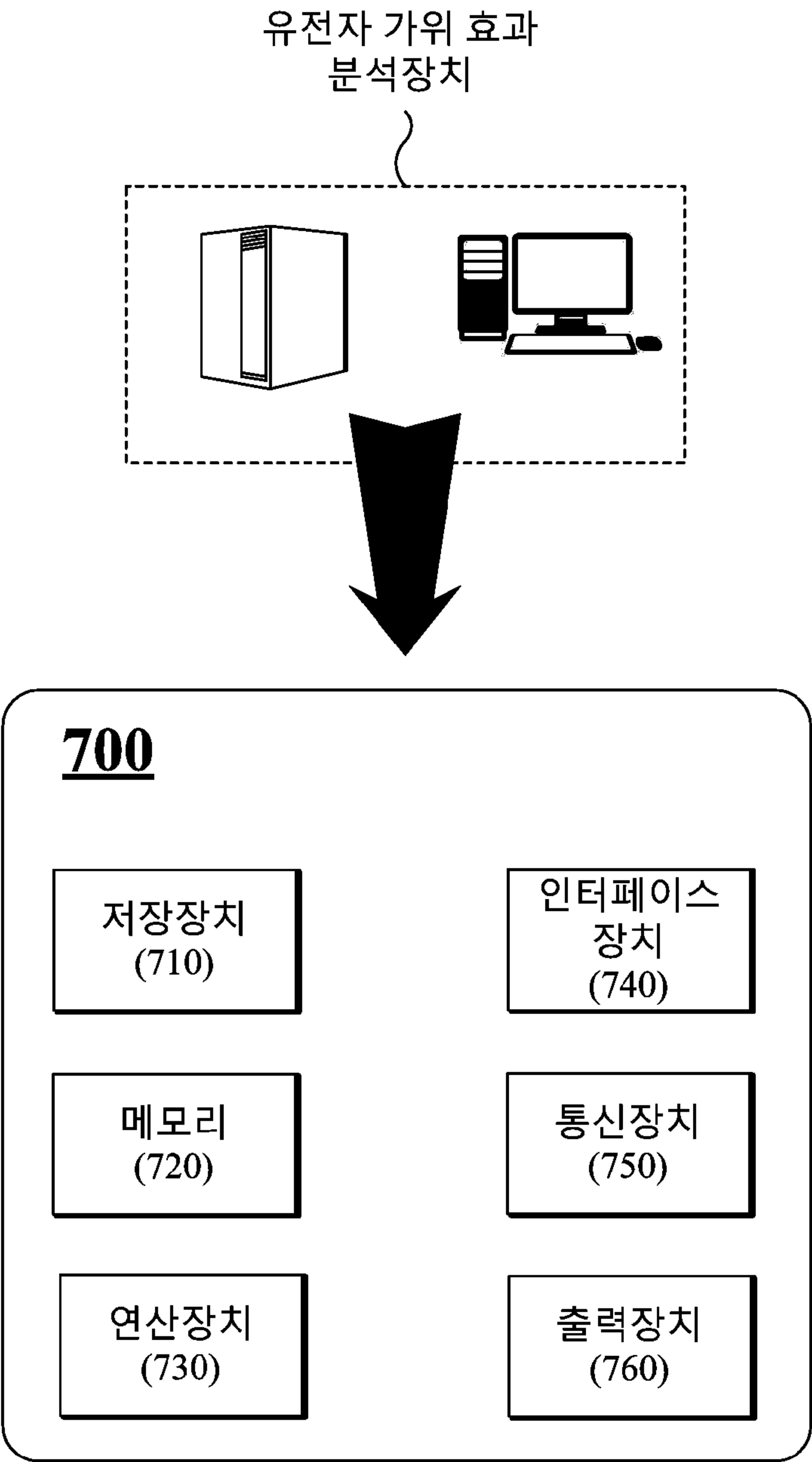
[Fig. 7]



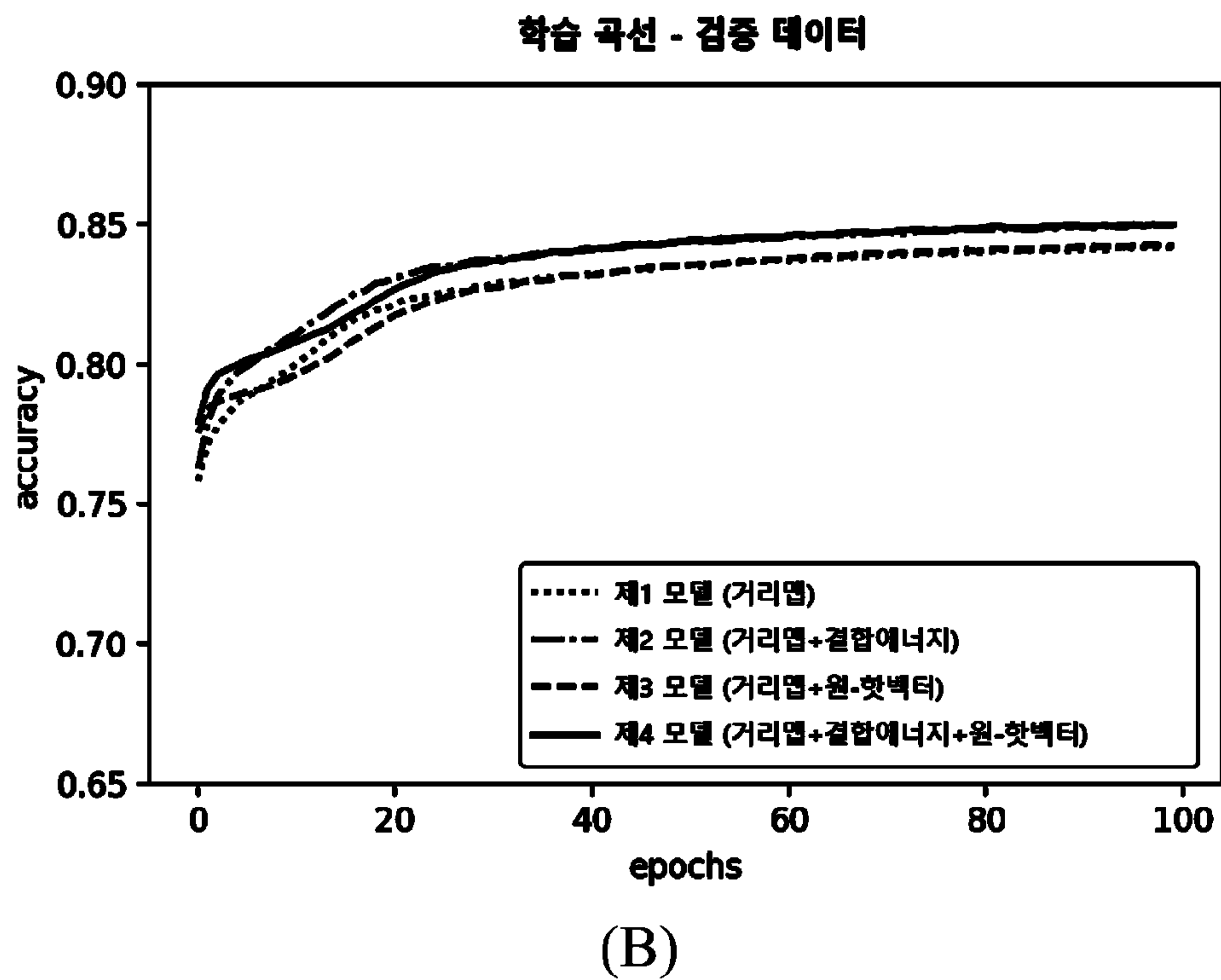
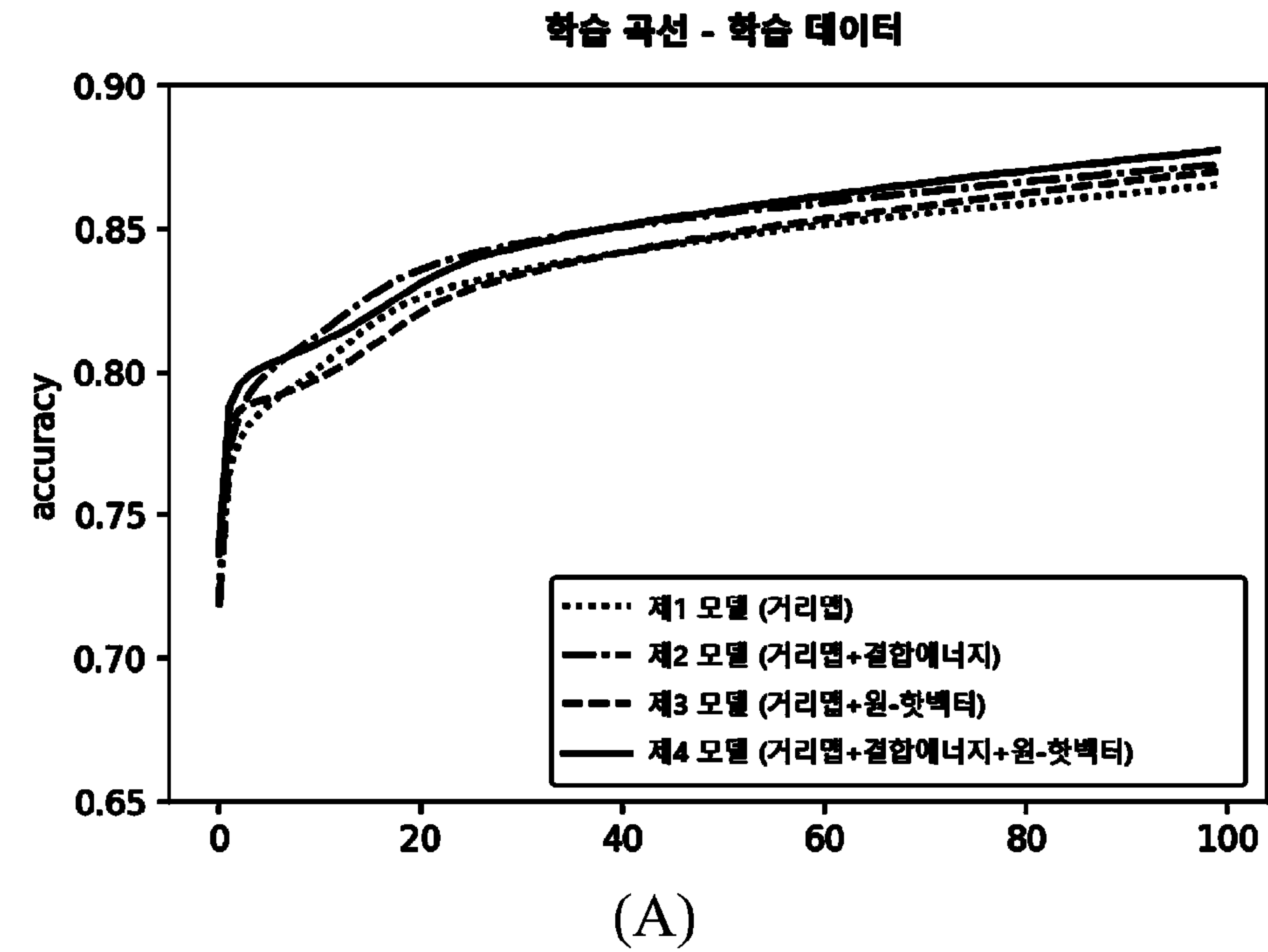
[Fig. 8]



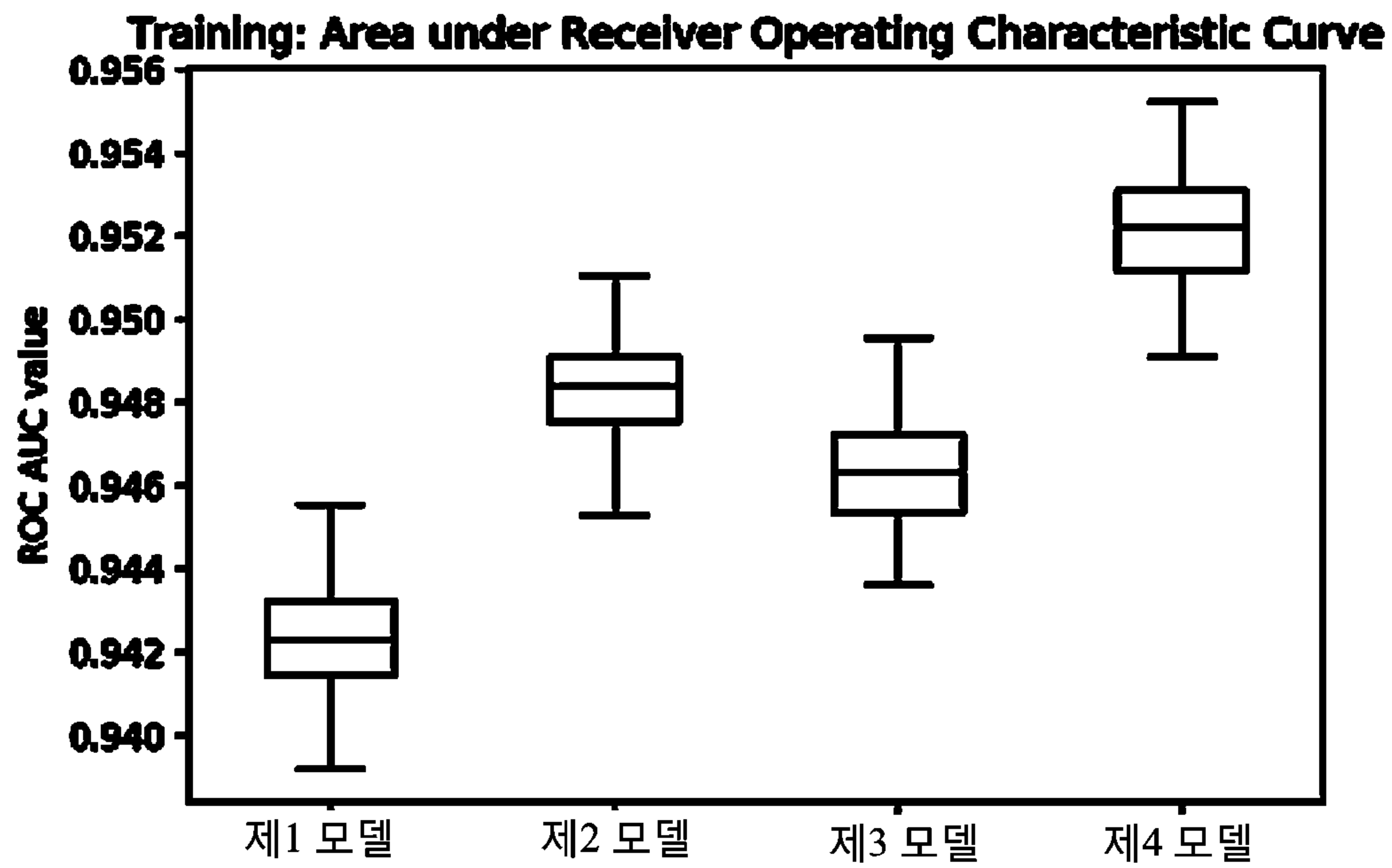
[Fig. 9]



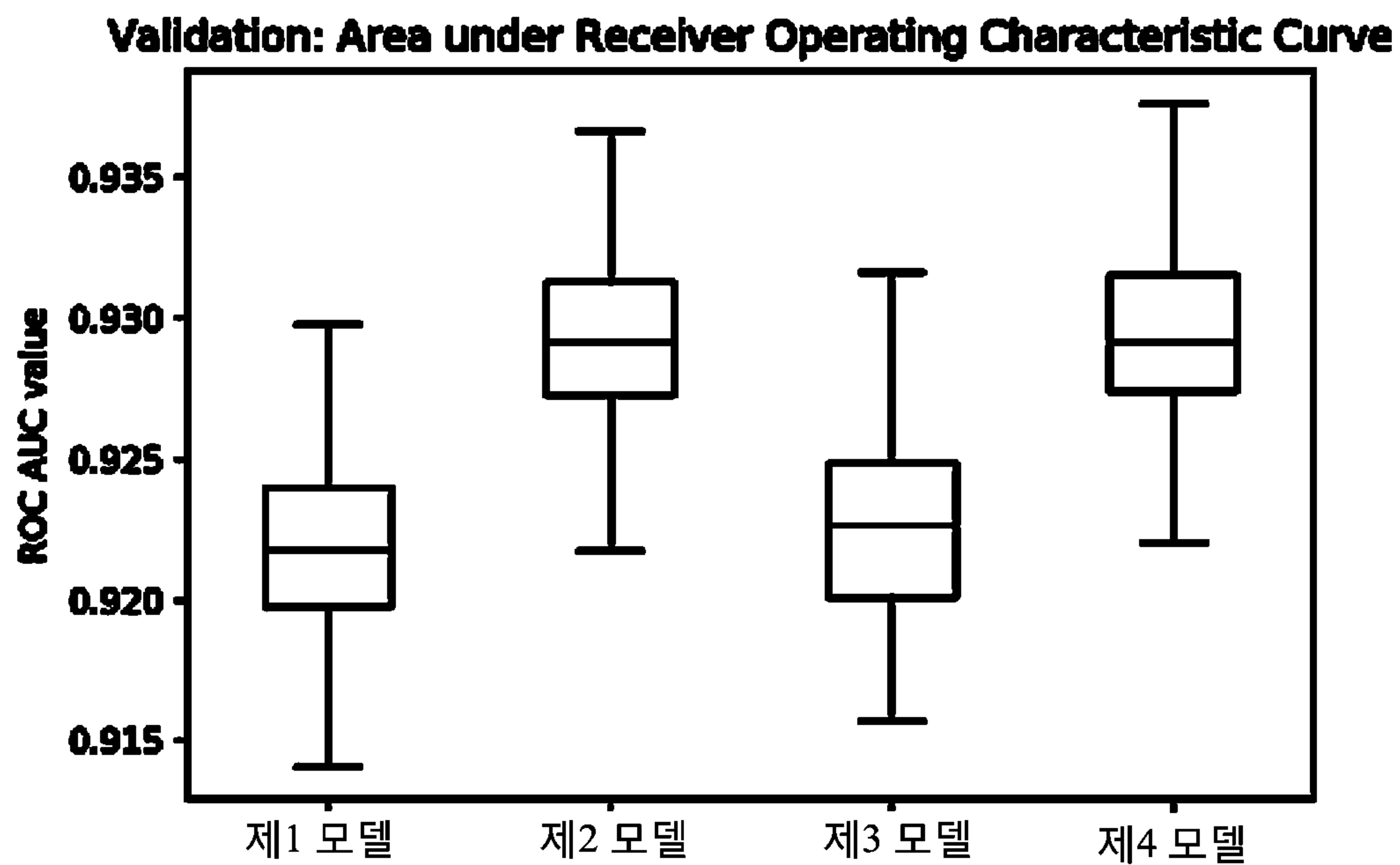
[Fig. 10]



[Fig. 11]

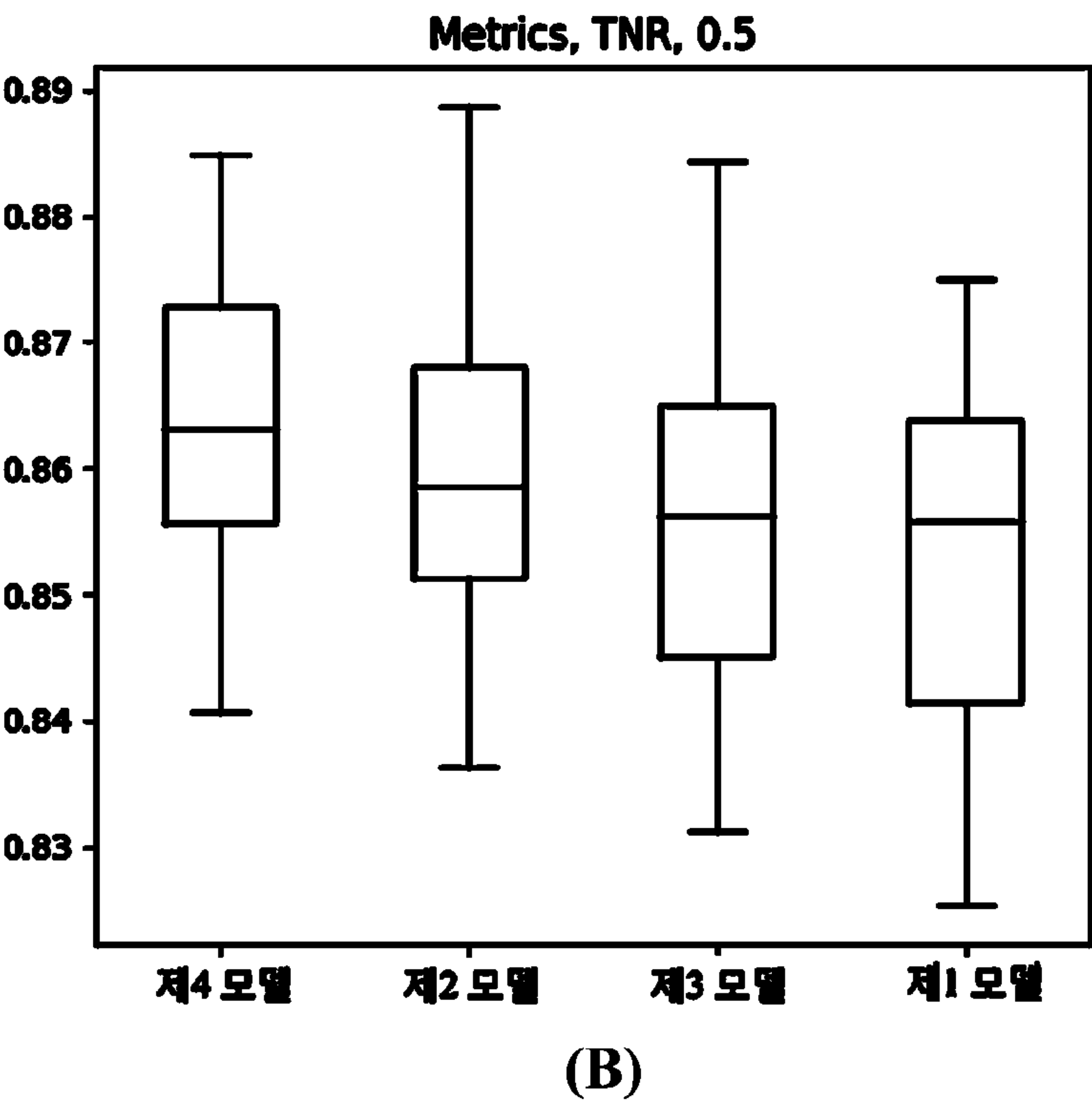
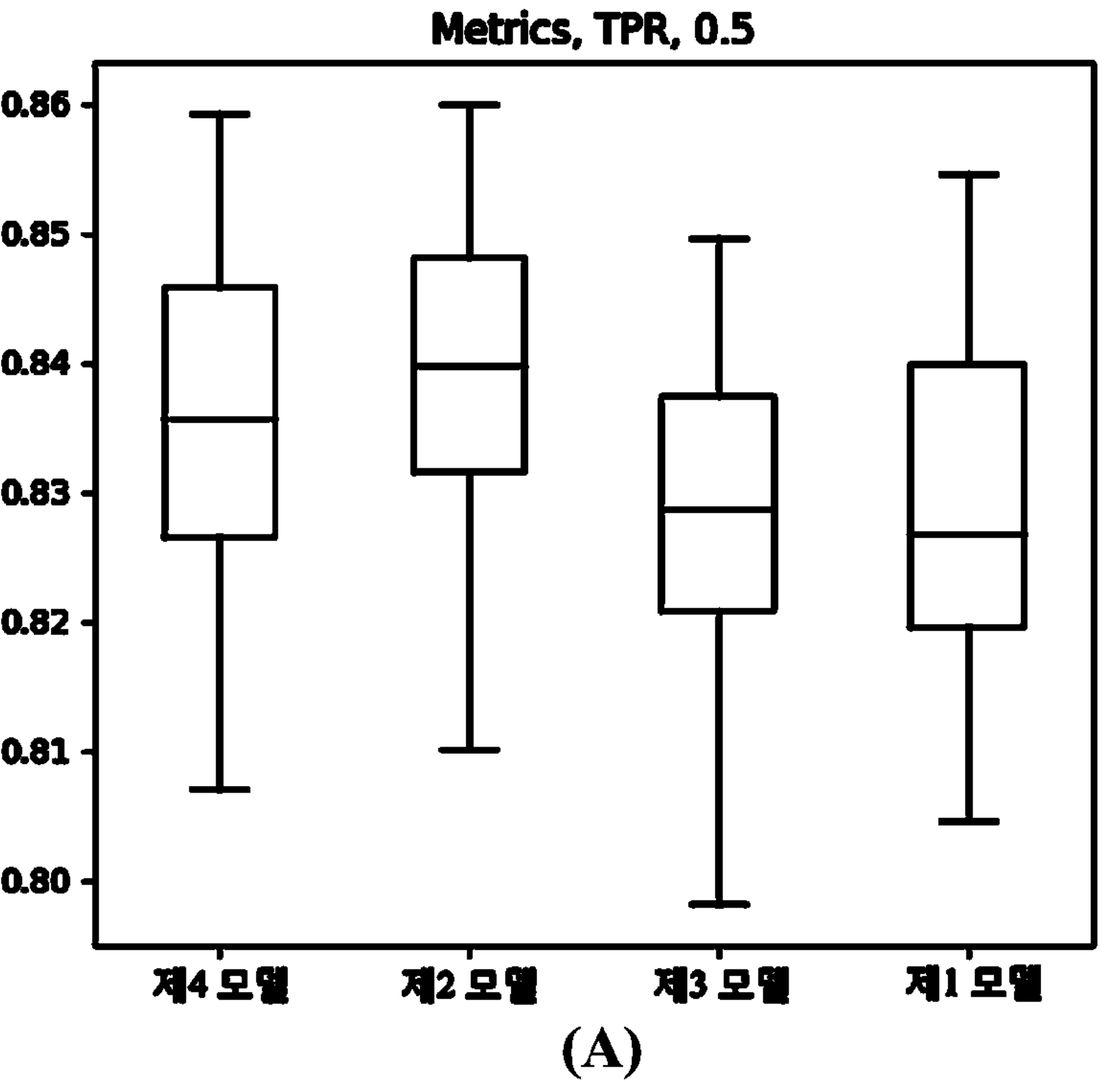


(A)



(B)

[Fig. 12]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/KR2019/014785

A. CLASSIFICATION OF SUBJECT MATTER

G16B 25/10(2019.01)i, G16B 30/10(2019.01)i, G16B 35/10(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G16B 25/10; G16B 25/00; G16B 5/00; G16B 30/10; G16B 35/10

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean utility models and applications for utility models: IPC as above

Japanese utility models and applications for utility models: IPC as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS (KIPO internal) & Keywords: CRISPR/Cas9, sequence, structural information, learning model, deep-learning, assessment

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	KR 10-2019-0048926 A (INDUSTRY-ACADEMIC COOPERATION FOUNDATION, YONSEI UNIVERSITY et al.) 09 May 2019 See claims 1, 4 and 14-15, paragraphs [0042] and [0044] and figure 8.	1-17
Y	FUSI, N. et al. In silico predictive modeling of CRISP/Cas9 guide efficiency. The preprint server for biology. 26 June 2015, pp. 1-31 See page 3, last paragraph, and page 6, third paragraph.	1-17
Y	GHOSAL, S. et al. Computational approaches for designing efficient and specific siRNAs. Bioinformatics. 2012, chapter 11, pp. 261-276 See page 264 and first paragraph-second paragraph.	1-17
A	CHUAI, G. et al. Deep CRISPR: optimized CRISPR guide RNA design by deep learning. Genome Biology. 2018, vol. 19, no. 80, pp. 1-18 See entire document.	1-17
A	WANG, L. et al. Prediction of sgRNA on-target activity in bacteria by deep learning. BMC bioinformatics. 24 October 2019, vol. 20, no. 517, pp. 1-14 See entire document.	1-17



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

31 AUGUST 2020 (31.08.2020)

Date of mailing of the international search report

01 SEPTEMBER 2020 (01.09.2020)

Name and mailing address of the ISA/KR

Korean Intellectual Property Office
Government Complex Daejeon Building 4, 189, Cheongsu-ro, Seo-gu,
Daejeon, 35208, Republic of Korea

Facsimile No. +82-42-481-8578

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/KR2019/014785

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	PIETAL, M. J. et al. RNAmapping 2D - calculation, visualization and analysis of contact and distance maps for RNA and protein-RNA complex structures. BMC Bioinformatics. 2012, vol. 13, no. 333, pp. 1-11 See entire document.	1-17

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/KR2019/014785

Patent document cited in search report	Publication date	Patent family member	Publication date
KR 10-2019-0048926 A	09/05/2019	KR 10-2029197 B1	08/10/2019

A. 발명이 속하는 기술분류(국제특허분류(IPC))
G16B 25/10(2019.01)i, G16B 30/10(2019.01)i, G16B 35/10(2019.01)i

B. 조사된 분야

조사된 최소문헌(국제특허분류를 기재)
G16B 25/10; G16B 25/00; G16B 5/00; G16B 30/10; G16B 35/10

조사된 기술분야에 속하는 최소문헌 이외의 문헌
한국등록실용신안공보 및 한국공개실용신안공보: 조사된 최소문헌란에 기재된 IPC
일본등록실용신안공보 및 일본공개실용신안공보: 조사된 최소문헌란에 기재된 IPC

국제조사에 이용된 전산 데이터베이스(데이터베이스의 명칭 및 검색어(해당하는 경우))
eKOMPASS(특허청 내부 검색시스템) & 키워드: 유전자 가위 (CRISPR/Cas9), 서열(sequence), 구조 정보(structural information), 학습 모델(learning model), 딥러닝(deep-learning), 평가(assessment)

C. 관련 문헌

카테고리*	인용문헌명 및 관련 구절(해당하는 경우)의 기재	관련 청구항
Y	KR 10-2019-0048926 A (연세대학교 산학협력단 등) 2019.05.09 청구항 1, 4, 14-15, 단락 [0042], [0044], 도면 8	1-17
Y	FUSI, N. 등, "In silico predictive modeling of CRISP/Cas9 guide efficiency", The preprint server for biology, 2015.06.26, 페이지 1-31 페이지 3, 마지막 문단, 페이지 6, 세 번째 문단	1-17
Y	GHOSAL, S. 등, "Computational approaches for designing efficient and specific siRNAs", Bioinformatics, 2012, 챕터 11, pp 261-276 페이지 264, 첫 번째-두 번째 문단	1-17
A	CHUAI, G. 등, "Deep CRISPR : optimized CRISPR guide RNA desing by deep learning", Genome Biology, 2018, 제19권, 제80호, 페이지 1-18 전체 문헌	1-17
A	WANG, L. 등, "Prediction of sgRNA on-target activity in bacteria by deep leargning", BMC bioinformatics, 2019.10.24, 제20권, 제517호, 페이지 1-14 전체 문헌	1-17

☒ 추가 문헌이 C(계속)에 기재되어 있습니다.

☒ 대응특허에 관한 별지를 참조하십시오.

* 인용된 문헌의 특별 카테고리:

“A” 특별히 관련이 없는 것으로 보이는 일반적인 기술수준을 정의한 문헌
“D” 본 국제출원에서 출원인이 인용한 문헌
“E” 국제출원일보다 빠른 출원일 또는 우선일을 가지나 국제출원일 이후에 공개된 선출원 또는 특허 문헌
“L” 우선권 주장에 의문을 제기하는 문헌 또는 다른 인용문헌의 공개일 또는 다른 특별한 이유(이유를 명시)를 밝히기 위하여 인용된 문헌
“O” 구두 개시, 사용, 전시 또는 기타 수단을 언급하고 있는 문헌
“P” 우선일 이후에 공개되었으나 국제출원일 이전에 공개된 문헌
“T” 국제출원일 또는 우선일 후에 공개된 문헌으로, 출원과 상충하지 않으며 발명의 기초가 되는 원리나 이론을 이해하기 위해 인용된 문헌
“X” 특별한 관련이 있는 문헌. 해당 문헌 하나만으로 청구된 발명의 신 규정 또는 진보성이 없는 것으로 본다.
“Y” 특별한 관련이 있는 문헌. 해당 문헌이 하나 이상의 다른 문헌과 조합하는 경우로 그 조합이 당업자에게 자명한 경우 청구된 발명은 진보성이 없는 것으로 본다.
“&” 동일한 대응특허문헌에 속하는 문헌

국제조사의 실제 완료일
2020년 08월 31일 (31.08.2020)

국제조사보고서 발송일
2020년 09월 01일 (01.09.2020)

ISA/KR의 명칭 및 우편주소
대한민국 특허청
(35208) 대전광역시 서구 청사로 189,
4동 (둔산동, 정부대전청사)
팩스 번호 +82-42-481-8578

심사관
정다원
전화번호 +82-42-481-5373



C (계속). 관련 문헌		
카테고리*	인용문헌명 및 관련 구절(해당하는 경우)의 기재	관련 청구항
A	PIETAL, M. J. 등, “RNAmapping 2D – calculation, visualization and analysis of contact and distance maps for RNA and protein-RNA complex structures”, BMC Bioinformatics, 2012, 제13권, 제333호, 페이지 1-11 전체 문헌	1-17

국제출원번호
PCT/KR2019/014785

PCT/KR2019/014785

공개일

2019/10/08