딥러닝 기반 타겟 및 화합물 초고성능 가상탐색 웹플랫폼 개발 연구

김우연[1], 송길태[2], 남호정[3], 김선[4], 임재창[5], 김종찬[6]

[1]KAIST, [2]부산대, [3]GIST,[4]서울대, [5]HITS, [6]KISTI

wooyoun@kaist.ac.kr, gsong@pusan.ac.kr, hjnam@gist.ac.kr, sunkim.bioinfo@snu.ac.kr, jaechang@hits.ai, jckim1201@kisti.re.kr

A Study on the development of a web-platform for deep learning-based high-performance drug and target discovery

Kim Woo Youn[1]*, Song Giltae[2], Hojung Nam[3], Kim Sun[4], Jaechang Lim[5], Kim Jongchan[6]

[1]KAIST, [2]PNU, [3]GIST, [4]SNU, [5]HITS, [6]KISTI

요 약

초고성능 컴퓨팅(HPC)과 딥러닝의 융합을 통해 신약 개발의 시간·비용 한계를 극복하고, 단백질-화합물 3D 구조 기반 예측 모델로 정확도와 속도를 동시에 개선하였다. 생성형 AI를 활용한 초거대 화합물 라이브러리 가상탐색 최적화 기술을 개발해 기존 대비 처리 효율을 획기적으로 향상시켰으며, 멀티오믹스 데이터 통합 분석을 통해 세포주 단위 약물 반응 예측 정밀도를 높였다. 합성치사 타깃 예측 모델 및 단백질 상호작용 네트워크 기반 약물 병용효과 분석 AI를 구축하는 동시에, 실험 연구자 맞춤형 웹 플랫폼을 기획하여 기술 접근성을 개선하고 있다. 본 연구는 딥러닝 기반의 타깃 발굴, 구조-활성 관계 해석부터 약물 가상탐색까지 신약 개발 초기 과정을 혁신하고 있다.

Ⅰ. 서 론

초고성능 컴퓨팅(HPC)과 딥러닝 기술의 융합은 신약 개발 분야에 혁신적인 변화를 가져올 것으로 기대된다. 본 과제에서는 기존 실험 기반 신약 개발의 시간적·경제적 한계를 극복하고, 신속하고 정확한 후보물질 발굴을 실현하기 위해HPC 기반 딥러닝 가상탐색 및 예측 기술을 개발하고자 한다. 단백질-화합물 복합체의3차원 구조 데이터를 활용한 구조 기반 예측 모델을 통해, 기존 물리 기반 계산 방법의 한계였던 정확도와 속도 간의 상충 문제를 해결하고, 구조-활성 상관관계에 기반한 후보물질의 최적화가 가능해졌다. 또한, 유전체, 전사체, 단백체 등 멀티오믹스 데이터를 통합 분석하는 딥러닝 모델을 도입함으로써, 단순 활성 예측을 넘어 세포주 수준에서의 약물 반응을 정밀하게 예측할 수 있게 되었으며, 이는 전임상 및 임상 성공률을 높이는 데 크게 기여할 것으로 기대된다.

본 연구의 대표적인 성과물로 단백질-약물 결합 잔기 및 결합력 예측 모델을 개발[1]하였다. 또한 수조 개의 화합물로 구성된 초거대 라이브러리에 대한 가상탐색을 위해 기존의 무작위 탐색 대신 생성형AI 기반 최적화 문제로 변형하였고, 이를 통해 기존 대비 획기적으로 향상된 처리 속도와 확장성을 확보하였다. 오믹스 정보를 활용하여 합성치사 타깃 물질을 예측하는AI 모델을 개발하였으며, 단백질 상호작용 네트워크 분석을 통해 약물 병용투여 효과 예측 및 약물 반응 유사도 분석을 위한 AI 모델을 개발하고 있다. 아울러, 실험 연구자들이 최신 인공지능 기반 신약개발 기술을 손쉽게 활용할 수 있도록 사용자 친화적인 웹 기반 플랫폼을 기획하고 있다. 본 논문에서는 현재까지 진행된 연구성과를 소개하고, 향후 계획에 대해 공유하고자 한다.

Ⅱ. 본론

본 연구의 핵심 목표에 따라 유도적합 예측을 위하여 자유 상태 단백질과 특정 화합물이 주어졌을 때 단백질을 구조 변화를 예측하는 모델의 개발을 위하여 자유 상태 단백질이 갖는 구조적 특징 및 정보를 유지하면서 화합물과의 결합을 위한 구조변화를 예측하여 빠른 추론 속도를 가지는 생성 모델을 개발 하였고, 물리 기반 생성 모델링을 적용 및 탬색 공간을 효과적으로 제한으로 높은 정확도를 도모하였다. 주어진 물리적 제약을 만족하는 최단 경로 탐색 방법을 개발하였으며, 이를 통해 자유 상태 단백질 분포와 단백질-화합물 결합 구조 간의 구조 변화 경로 데이터를 구축하고, 단백질과 화합물 사이의 복잡한 상호작용을 원자 단위가 아닌 비공유 결합성 상호작용으로 단순화하여 속도 향상과 동시에 정확도를 유지할 수 있도록 하였다.

구조 기반 가상탐색의 초기 단계에서 탐색 과정의 효율성과 정확성을 확보하기 위하여 단백질 내에서 결합하는 핵심 잔기에 대한 정보를 제공하기 위해 서열기반 단백질-화합물 결합 잔기 예측 모델을 개발하였으며, 화합물 정보를 활용하지 않는 기존 방식과는 달리 화합물에 대한 임베딩을 생성하과 단백질 내 각 잔기에 대한 임베딩을 결합하여 화합물 결합 잔기 예측 모델을 완성하였으며 이를 성과지표 검증을 통해 잔기 임베딩을 형성할 때 주변 잔기에 대한 정보를 반영하는 것과 화합물 정보를 활용하는 것이 예측 정확성을 높이는데 중요함을 확인하였다.

본 연구에서는 주어진 단백질과 높은 상보성을 가지면서 합성 가능성이 고려된 분자를 설계하기 위한 방법론으로 구매 가능한 분자 빌딩블록과 합성 규칙을 이용해 순차적으로 분자를 설계하는 분자 생성 모델 RxnFlow[2]를 개발하였으며, 이는 최신 생성 모델 아키텍처인 GFlowNet[3]을 활용하여 주어진 보상 함수에 대한 다양한 고보상 분자를 생성함과 동시에 해당 분자의 합성 경로를 제시하고 있으며, 이는 단일 CPU 코어에서 초단 20개의 분자를 설계할 수 있어 손쉽게 대규모 라이브러리를 생성할 수 있다. 이는 동일한 타겟에 대해서 사용자에 따라 서로 다른 약물 후보물질을 제안한다.

초거대 가상 라이브러리를 구축하기 위하여, 멀티오믹스 데이터 기반 지식 그래프를 구축하여 대규모 지식그래프에 세포주별 특이적으로 나타나는 멀티오믹스 데이터를 이용하여 합성치사 기반 약물 표적 예측모델을 개발 하였다. 이는 기존 모델들과의 성능 비교 통해 지식그래프와 멀티오믹스 데이터를 함께 사용하였을 때 예측 성능이 향상되는 것을 확인하였

다. 암세포와 단일 항암제 조합에 대한 약물 농도-반응 곡선을 예측하는 단일 항암제 모델과 해당 모델의 결과를 이용하여 복합 항암제의 효능을 예측하는 복합 항암제 모델을 개발하였으며, 맞춤형 정밀의료가이드라인을 예측 및 제안하는 인공지능 프레임워크 DD-PRiSM[4]을 개발하였다. 세포주를 대상으로 한 약물 투여와 실제 환자 체내의 암조직을 대상으로 한 약물 투여는 그 결과가 상이한 것이 알려져 있는데, 이를 해결하기 위해 암세포 혹은 환자의 오믹스 데이터로부터 환경과 무관한, 세포의 상태를 대변할 수 있는 부분을 추출하는 방법을 통해 해당 문제를 해결하고, 이를 위해 도메인 적응 기법을 사용하여 암세포의 유전자 발현량으로부터 암 특성을 추출하여 훈련에 사용된 암세포 뿐만 아니라 환자에서도 정확한 예측을 진행할 수 있도록 복합 항암제 효능 예측 모델의 고도화를 진행중이다. 나아가, 도메인 적응 (Domain adaptation) 기술을 사용하여 세포주를 포함하는 생체외 환경뿐만 아니라 환자와 같은 생체내 환경 데이터에서도 정확한 예측을 진행할 수 있도록 진행하려 한다. 세포주와 환자의 유전자 발현량을 생체외/생체내 환경에 의한 유전자 발현과 해당 세포의 자체 특성 (암 등)에 의한 유전자 발현으로 분리, 실제 해당 세포의 자체 특성 기반 특징을 추출하고자 한다. 최신 인공지능 기반 신약개발 기술을 손쉽게 활용할 수 있도록 사용자 친화적인 웹 기반 플랫폼을 서비스화를 위하여, 국가 슈퍼컴퓨터 자원- 플랫폼 연계 기술 개발 및 서비스 자원-플랫폼 연계 방안을 설계하고 있다.

본 연구는 단백질-화합물 구조 기반 분석 및 멀티오믹스 데이터 분석을 활용하는 신약 개발 딥러닝 요소기술 개발 및 초고성능컴퓨팅 기반 가상탐색 플랫폼 구축하기 위해 개발된 모델들을 지속적으로 개선하고 고도화 위한 연구를 진행 중이다.

Ⅲ. 결론

최근 알파폴드3와 같이 바이오, 제약 분야의 문제를 풀기 위한 딥러닝 기술에 많은 진보가 있었다. 하지만 기술적 진보에 반하여 속도의 한계 및 기존 물리 기반 방법론 대비 낮은 정확도와 낮은 신뢰성으로 인해 대규모 가상탐색 및 실제 신약 개발에 실증된 사례는 많지 않다. 본 과제의 연구 범위는 화합물의 단백질 유도적합과 화합물의 세포 내 영향인 멀티오믹스를 모두 고려한 초거대 라이브러리 가상탐색으로, 기존 방법 대비 높은 신뢰성을 가질 것으로 예상된다. 본 연구의 결과물의 웹서비스 제공을 통해 컴퓨터 비전문가들에게도 초고성능컴퓨팅 신약개발 방법들에 대한 높은 접근 가능성을 제시하고자 한다. 또한, 본 연구의 결과물을 실제 실험적으로 검증함으로써 신약 개발자의 신뢰성을 확보, 실제 신약 개발 산업 분야에서 효율적인 신약개발 도구로서 활용하고자 한다.

ACKNOWLEDGMENT

This work was supported by Basic Science Research Programs through the National Research Foundation of Korea (NRF), grant-funded by the Ministry of Science and ICT (RS-2023-00257479).

참 고 문 헌

[1] Moon, Seokhyun, et al. "PIGNet2: a versatile deep learning-based protein-ligand interaction prediction model for binding affinity scoring and virtual screening." Digital Discovery 3.2 (2024): 287-299.

[2] Seo, Seonghwan, et al. "Generative Flows on Synthetic Pathway for Drug Design." arXiv preprint arXiv:2410.04542 (2024).

[3] Y*oshua, Bengio et al.*, GFlowNet Foundations,*Journal of Machine Learning Research* 24 (2023) 1-55

[4] Jin, Iljung et al., “DD-PRiSM: a deep learning framework for decomposition and prediction of synergistic drug combinations” *Briefings in Bioinformatics*, Volume 26, Issue 1, January 2025,