한양대학교

“SW프로젝트” 제안서 작성 안내

1. 양식의 푸른색 글씨는 작성요령으로 제출 시에는 반드시 삭제

2. 목차별 내용을 1페이지 내외로 작성하되 각 내용별로 새 페이지로 작성할 것
   (예를 들어 1번에 대해 작성한 내용이 1.5 페이지 분량이라면
   2번에 대한 내용은 나머지 0.5페이지를 비워두고 새로운 페이지에서 작성)

3. 공업센터 본관 503호 이유정 선생님께 오프라인 제출

한양대학교

SW프로젝트 제안서

프로젝트명	단백질 3차 구조 예측	
프로젝트 유형	지정 주제 (O) 자유 주제 ( )	졸업 작품 (O) 졸업 논문 ( )
프로젝트 요약	단백질 구조 관련 지식은 응용 분야가 넓고 매우 유용하다. 그중에서도 단백질 접힘 구조에 대한 예측은 아주 중요한 기술이다. 하지만 아직까지 단백질 구조예측에 대해 충분할 만큼의 빠르고 정확한 기술을 가지고 있지 않다. 언어 혹은 아미노산 모델을 학습하여 최종적으로 벤치마크 데이터에 대해서 기존의 개발된 단백질 접힘 구조에 대한 예측 기술의 학습과 재현을 해보고, 보다 효율적으로 접힘 구조의 예측을 위해 개선해본다.	
R&D 산출물	SW (O), HW ( ), 특허 ( ), 논문 ( ), 프로그램등록 ( )	
지도교수	노영균	
예상기간	2021.08.14 – 2021.05.01.	

전공	학번	학년	이름	연락처
컴퓨터소프트웨어	2013011985	4	성범모	qjaah2@gmail.com 010-4267-1248
컴퓨터소프트웨어	2013011866	4	김주호	juho0153@naver.com 010-4800-0161

• 연락처는 이메일과 전화번호를 모두 쓰되 반드시 수신 가능한 것으로 기입
(연락을 받지 못해 불이익을 당할 수 있음)

한양대학교

목 차

1. 프로젝트 배경 및 목표

이 프로젝트는 Transformer 모델을 이용해서 seqeunce 데이터를 분석하여 단백질의 3차원 구조 데이터를 예측하는 것이다.

단백질은 아미노산 결합으로 연결된 3차원 구조를 이루고 있으며, 다양한 구조를 가질 수 있고 그 세부적인 구성에 따라서 다양한 기능을 수행할 수 있는데, 특정 단백질이 어떤 기능을 수행할지 예측하기 위해서는 먼저 단백질의 3차 구조를 알아내는 것이 바탕이 되어야 한다. 이 3차원 구조를 실험적으로 알아내기 위해서는 X선 결정학의 도움을 빌리거나, 극저온 현미경 등을 활용하여야 하지만, 이 방법들로 단백질 접힘구조를 밝히려면 소요되는 시간과 비용이 많이 들어서 짧게는 몇 개월에서 길게는 몇 년이 걸리기도 한다. 어떤 복잡한 단백질 구조는 10년이 걸려도 실마리를 잡지 못하는 경우도 있다.

하지만 인공지능과 알고리즘의 발전으로 인하여 실험을 통한 결과를 확인하기 전에 그 결과를 어느 정도 예측할 수 있게 되었다. 예를 들어 딥마인드사에서 개발된 'AlphaFold'는 복잡한 단백질 구조를 불과 2시간 만에 예측하기도 하였다.

이렇게 단백질의 구조를 예측할 수 있게 되면 이는 신약 개발을 함에 있어서 아주 중요한 역할을 할 수 있다. 또한, 구조 예측을 통해서 기존의 효소보다 더 효과적으로 반응을 하는 효소를 얻을 수도 있게 된다. 예를 들면 플라스틱을 분해하는 효소를 발견하고, 이 효소의 반응을 극대화하게 된다면 현대사회의 큰 문제 중 하나인 쓰레기 문제의 해결에 기여할 수 있다. 그리고 단백질 3차 구조를 이용해서 백신을 개발하는 연구가 진행중인데, 단백질 3차 구조 예측을 통해서 변이된 바이러스를 치료할 수 있는 백신을 더 빠르게 개발할 수 있게 된다.

이러한 이유로 단백질 구조를 예측하기 위하여, 관련된 언어, 사진, 생명 정보의 대용량 데이터 처리를 위한 transformer 모델의 학습과 원리 파악에 관한 연구를 수행하려고 한다.

프로젝트에서 달성하고자 하는 최종목표는 단백질에 관한 대용량 데이터를 입력하여 그 3차 구조에 대해 예측하는 알고리즘을 구현해보는 것이다. 또한 현재 단백질 3차 구조를 가장 빠르고 정확하게 분석하는 기술인 딥마인드의 'AlphaFold' 학습 방법을 재현해본다.

2. 프로젝트 내용

프로젝트의 최종목표를 달성하는 데 필요한 역량은 Transformer 알고리즘에 대한 이해가 필요하고, 대용량의 데이터 처리를 요구하기 때문에 병렬처리 연산에 대한 이해가 필요하다. 또한, 단백질 3차 구조를 예측하고 분석하는 데에 필요한 생명공학적 지식도 요구된다.

단백질은 아미노산이 3차원 구조로 결합되어 있는데, 이 3차 구조를 예측하기 위해서는, 단백질을 구성하고 있는 레지듀(Residue)들 사이의 CB 원자(CA for Glycine)간의 거리 정보(distance hitogram: distogram)를 예측하고 각 레지듀들의 백본 비틀림각(Backbone Torsion Angle)과 2차구조 정보를 예측해야 한다. 단백질의 다중서열정렬(MSA)속에서 서로 공변하는 레지듀들 간의 거리 정보를 통해서 3차원 구조를 모델링하는 것이다.

단백질의 생체내에서의 기능(Function)을 이해하기 위해 그 3차원적 구조(Structure)를 규명하는 것은 현대 분자생물학과 생물물리학에서 가장 기초적인 연구였다. X-ray 결정학, 핵자기 공명법(NMR) 등의 실험적 결정 방법들이 주로 사용되어 왔고, 이론과 컴퓨터를 이용한 계산 방법으로 단백질의 구조를 예측하고자 하는 오랜 시도가 있었다. 일반적으로 단백질의 서열이 유사한 경우(≥30%), 그 3차 구조도 유사하기 때문에 이미 밝혀진 구조가 있으면 그것을 템플릿(Template)으로 사용하여 모델링하는 방법과, 템플릿이 없는 경우 여러 에너지 함수들을 사용하여 모델링하는 방법들이 사용되어왔다. 템플릿이 존재하지 않는 경우, 예측된 모델의 정확도는 대략 40% 미만으로 오랜 기간 정확도의 향상이 정체되고 어려웠다. 한편 기계학습, 곧 인공지능 연구의 최근의 발전은 그동안 불가능해 보이던 여러 영역들에서 인간의 능력을 뛰어 넘는 결과들을 보여 주고 있고, 기존 전통적 방법의 한계를 극복할 수 있는 가능성을 보여 주고 있다. 알파고(AlphaGo) 개발을 통해 세계의 주목을 받았던 구글의 DeepMind팀에서 최근 알파폴드(AlphaFold)를 지난 단백질 구조예측 대회 CASP13 (2018년 12월)에서 소개하며 기대를 뛰어넘는 결과와 함께, 과학계에 신선한 충격을 가져왔다. 이는 템플릿이 없는 서열에 대한 단백질 구조를 예측하는 데 있어서, 인공지능의 가능성을 보여준 사례라고 볼 수 있다.

현재 단백질 3차 구조를 예측하는 프로그램중 가장 뛰어난 프로그램인 'AlphaFold'를 연구하면서 원리를 파악하며, 최종적으로는 아미노산 배열 데이터를 이용해서 딥마인드의 알파폴드 학습 방법을 재현하려고 한다. 또한 알파폴드를 재현하면서 단백질 연구에서 어떤 것을 이용할 수 있는지 연구하고, 가능하다면 알파폴드에서 필요한 개선점에 대해서 찾아본다.

3. 프로젝트 추진 계획

프로젝트 기간 동안 추진할 개괄적인 월단위 계획		
2021.08	교수님 면담 / 단백질 3차구조 예측 관련 세미나(고등과학원, 주기형교수님)	
2021.09	Alphafold 관련 학습	python, keras, numpy등 필요한 프로그래밍 능력 학습
2021.10		GPU를 활용하는 병렬프로그래밍 관련 학습
2021.11		Alphafold 프로그램 작동 및 실험
2021.12		Alphafold 논문 및 단백질 3차 구조 관련 논문 학습
2022.01	Alphafold의 알고리즘과 동일한 Transformer 모델 개발	
2022.02		
2022.03	Alphafold의 알고리즘을 개선한 Transformer 모델 개발	
2022.04		
2022.05	새로 개선한 Transformer 모델의 평가 및 수정	

프로젝트 팀원의 구성과 역할의 분담		
역할	성범모	김주호
	Alphafold 관련 논문 해석 및 정리 프로젝트 관련 미팅 회의록 작성 및 정리 프로젝트 세부 일정 조율	단백질 3차 구조 관련 논문 해석 및 정리 프로젝트 미팅 주선 프로젝트 github 개설 및 관리
	공동	
	python, keras, numpy등 관련 프로그래밍 능력 학습 GPU를 활용하는 병렬프로그래밍 관련 학습 Alphafold 프로그램 구현	

4. 결론

프로젝트 최종 목표는 Transfomer 모델의 학습을 통하여 생명정보에 관한 대용량 데이터를 처리하는 알고리즘을 만들어보는 것이다. 또한 가능하다면 CASP14 대회에서 사용된 벤치마크 프로그램에서 가장 좋은 성능을 내었던 알파폴드의 성능에 근접할 수 있도록 개선해보는 것이다.

단백질 3차 구조 관련한 프로젝트는 단순히 인공지능 알고리즘만을 학습하여 만들 수 있는 것이 아니라, 단백질이 왜 접힘 구조를 갖게 되며, 단백질 분자 간의 어떠한 결합을 통해 어떠한 모양으로 3차 구조를 갖는지 등 전반적인 관련 지식들을 이해하고 있어야 하므로 알고리즘적, 프로그래밍적 능력뿐만 아니라, 단백질 3차 구조 관련한 생명공학적 지식이 많이 필요하고 향상될 것이라고 생각한다. 그리고 생명 정보 sequence 데이터가 매우 방대하기 때문에 병렬 프로그래밍 없이 프로그램을 실행하는 것은 수행비용 면에서 매우 비효율적이다. 따라서 속도를 개선하는 측면에서 필요한 병렬 프로그래밍 관련 지식도 어느 정도 요구될 것이라고 생각한다. 또한 프로젝트 수행에 관련된 지식뿐 아니라, 팀 단위 작업에 대한 이해와 팀원과 효율적으로 분업하는 방법을 배울 수 있을 것으로 생각한다.

이 프로젝트의 응용 분야와 학술적, 산업적 기대효과에 대해서는 3가지로 설명하려고 한다.

첫 번째는 단백질 기능 예측을 통한 이득을 기대해 볼 수 있다. 단백질 3차 구조는 단백질의 기능을 이해하는 데 중요한 역할을 하며, 특히 단백질-리간드 상호작용 (protein-ligand interaction)에 대한 연구는 단백질 기능연구의 시작점일 뿐 아니라, 궁극적으로 신약 개발에 있어 중요한 연구이다.

두 번째는 구조예측과 리간드 도킹 연구를 통해 제시된 3가지 단백질에 대하여 실험을 수행하여 기존의 효소보다 30~60배 더 효율적으로 원하는 방향으로 반응을 진행하게 하는 효소를 얻을 수 있다. 예를 들면, 플라스틱을 분해하는 효소의 반응성을 극대화하게 된다면, 현대사회의 가장 큰 문제라고 할 수 있는 쓰레기 문제를 해결하는데 기여하는 효과를 기대할 수 있다.

세 번째는 학술적인 자연과학 분야의 발전을 기대할 수 있다. 생체내의 모든 단백질의 구조와 기능을 밝히는 일은 생명현상에 대한 규명을 더 넓게 가능하게 한다. 따라서 이런 정보를 기반으로 응용된 자연과학의 발전을 기대해 볼 수 있다.

5. 참고 문헌

1. 오민아, 주기형, 이주영 / 단백질 3차 구조 모델링에 기반한 리간드 결합부위 예측 / 한국산업응용수학회 / 2009

2. 오민아, 주기형, 이인호, 이주영 / 단백질 구조 예측 왜, 어떻게 하나? / 생화학분자생물학회 / 2010

3. 주기형 / 알파폴드(AlphaFold) : 단백질 3차 구조 예측 / 한국생물공학회 / 2019

4. Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick 등 18인 Improved protein structure prediction using potentials from deep learning / nature / 2020.01.15