2022년 글로벌 핵심인재 양성지원 사업

최 종 보 고 서

2022. 9. 22.

유형구분	글로벌핵심인재양성지원
주관연구개발기관명	선문대학교
과 제 명	미생물 게놈 빅데이터 분석을 위한 AI 기반 환경 ICT 융합 글로벌 인재양성
연구책임자	김 정 동
과제기간	2021.05.01.~2022.08.31

■ 국가연구개발혁신법 시행규칙 [별지 제5호서식] <개정 2022. 1. 4.>

최종보고서

보안등급
일반[V], 보안[ ]

중앙행정기관명	과학기술정보통신부	사업명	사업명	글로벌핵심인재양성지원	
전문기관명(해당 시 작성)	정보통신기획평가원		내역사업명	글로벌핵심인재양성지원	
공고번호	제2021-0115호	총괄연구개발 식별번호			
		연구개발과제번호	2021-0-01581		

기술분류	국가과학기술 표준분류	EE0108 (인공지능)	50%	LA0706 (생물정보학)	30%	EH1099 (일리 분류되지 않는 환경처측/감시/분기)	20%	
	부처기술분류 (해당 시 작성)	1순위 소분류 코드명	%	2순위 소분류 코드명	%	3순위 소분류 코드명	%	

총괄연구개발명 (해당 시 작성)	국문				
	영문				
연구개발과제명	국문	미생물 게놈 빅데이터 분석을 위한 AI 기반 환경 ICT 융합 글로벌 인재양성			
	영문	Cultivation of innovative ICT-based interdisciplinary global talents for AI-based big data research in microorganism genome			
주관연구개발기관	기관명	선문대학교 산학협력단	사업자등록번호	312-82-10144	
	주소	(31460) 충청남도 아산시 탕정면 선문로221번길 70	법인등록번호	164871-0004430	
연구책임자	성명	김정동	직위	부교수	
	연락처	직장전화 041-530-2221	휴대전화	010-8806-2156	
		전자우편 kjd4u@sunmoon.ac.kr	국가연구자번호	10166644	
연구개발기간	전체	2021. 5. 1. ~ 2022. 8. 31.(1년 4개월)			
	단계 (해당 시 작성)	1단계 YYYY. MM. DD - YYYY. MM. DD(년 개월)			
		n단계 YYYY. MM. DD - YYYY. MM. DD(년 개월)			

연구개발비 (단위: 천원)		정부지원 연구개발비	기관부담 연구개발비		그 외 기관 등의 지원금				합계			연구개발비 외 지원금
					지방자치단체		기타( )					
		현금	현금	현물	현금	현물	현금	현물	현금	현물	합계	
총계		500,000							500,000		500,000	
1단계	1년차											
	n년차											
n단계	1년차											
	n년차											

공동연구개발기관 등 (해당 시 작성)	기관명	책임자	직위	휴대전화	전자우편	비고	
						역할	기관유형
공동연구개발기관	극지연구소	이준혁	책임연구원	010-4739-7686	junhyucklee@kopri.re.kr	공동	정부출연연
위탁연구개발기관							
연구개발기관 외 기관							

연구개발담당자 실무담당자	성명	박주연	직위	연구원
	연락처	직장전화 041-530-2218	휴대전화	01057899480
		전자우편 a99899480z@gmail.com	국가연구자번호	12462089

이 최종보고서에 기재된 내용이 사실임을 확인하며, 만약 사실이 아닌 경우 관련 법령 및 규정에 따라 제재처분 등의 불이익도 감수하겠습니다.

2022 년 9 월 22 일

연구책임자: 김 정 동 (서명)

주관연구개발기관의 장: 김 종 해 (직인)
공동연구개발기관의 장: 강 성 호 (직인)
위탁연구개발기관의 장: (직인)

중앙행정기관의 장 귀하

■ 국가연구개발혁신법 시행규칙 [별지 제5호서식] <개정 2022. 1. 4.>

최종보고서

보안등급
일반[V], 보안[ ]

중앙행정기관명	과학기술정보통신부	사업명	사업명	글로벌핵심인재양성지원
전문기관명(해당 시 작성)	정보통신기획평가원		내역사업명	글로벌핵심인재양성지원
공고번호	제2021-0115호	총괄연구개발 식별번호		
		연구개발과제번호	2021-0-01581	

기술분류	국가과학기술 표준분류	EE0108 (인공지능)	50%	LA0706 (생물정보학)	30%	EH1099 (달리 분류되지 않는 환경예측/감시/평가)	20%
	부처기술분류 (해당 시 작성)	1순위 소분류 코드명	%	2순위 소분류 코드명	%	3순위 소분류 코드명	%

총괄연구개발명 (해당 시 작성)	국문	
	영문	
연구개발과제명	국문	미생물 게놈 빅데이터 분석을 위한 AI 기반 환경 ICT 융합 글로벌 인재양성
	영문	Cultivation of innovative ICT-based interdisciplinary global talents for AI-based big data research in microorganism genome

주관연구개발기관	기관명	선문대학교 산학협력단	사업자등록번호	312-82-10144
	주소	(31460) 충청남도 아산시 탕정면 선문로221번길 70	법인등록번호	164871-0004430

연구책임자	성명	김정동	직위	부교수
	연락처 직장전화	041-530-2221	휴대전화	010-8806-2156
	연락처 전자우편	kjd4u@sunmoon.ac.kr	국가연구자번호	10166644

연구개발기간	전체	2021. 5. 1. ~ 2022. 8. 31.(1년 4개월)
	단계 (해당 시 작성) 1단계	YYYY. MM. DD - YYYY. MM. DD(년 개월)
	n단계	YYYY. MM. DD - YYYY. MM. DD(년 개월)

연구개발비 (단위: 천원)		정부지원 연구개발비	기관부담 연구개발비		그 외 기관 등의 지원금				합계			연구개발비 외 지원금
					지방자치단체		기타( )					
		현금	현금	현물	현금	현물	현금	현물	현금	현물	합계	
총계		500,000							500,000		500,000	
1단계	1년차											
	n년차											
n단계	1년차											
	n년차											

공동연구개발기관 등 (해당 시 작성)	기관명	책임자	직위	휴대전화	전자우편	비고	
						역할	기관유형
공동연구개발기관	극지연구소	이준혁	책임연구원	010-4739-7686	junhyucklee@kopri.re.kr	공동	정부출연연
위탁연구개발기관							
연구개발기관 외 기관							

연구개발담당자 실무담당자	성명	박주연	직위	연구원
	연락처 직장전화	041-530-2218	휴대전화	01057899480
	연락처 전자우편	a99899480z@gmail.com	국가연구자번호	12462089

이 최종보고서에 기재된 내용이 사실임을 확인하며, 만약 사실이 아닌 경우 관련 법령 및 규정에 따라 제재처분 등의 불이익도 감수하겠습니다.

2022 년 9 월 22 일

연구책임자: 김 정 동 (인)

주관연구개발기관의 장: 김 종 해 (직인)
공동연구개발기관의 장: 강 성 호 (직인)
위탁연구개발기관의 장: (직인)

중앙행정기관의 장 귀하

< 요 약 문 >

※ 요약문은 5쪽 이내로 작성합니다.

<table>
<tr><td colspan="2">사업명</td><td colspan="3">글로벌핵심인재양성지원</td><td>총괄연구개발 식별번호
(해당 시 작성)</td><td colspan="2"></td></tr>
<tr><td colspan="2">내역사업명
(해당 시 작성)</td><td colspan="3">글로벌핵심인재양성지원</td><td>연구개발과제번호</td><td colspan="2">2021-0-01581</td></tr>
<tr><td rowspan="2">기술분류</td><td>국가과학기술
표준분류</td><td>EE0108
(인공지능)</td><td>50
%</td><td>LA0706
(생물정보학)</td><td>30
%</td><td>EH1099
(달리 분류되지 않는 환경에 측/감시/평가)</td><td>20%</td></tr>
<tr><td>부처기술분류
(해당 시 작성)</td><td>1순위 소분류 코드명</td><td>%</td><td>2순위 소분류 코드명</td><td>%</td><td>3순위 소분류 코드명</td><td>%</td></tr>
<tr><td colspan="2">총괄연구개발명
(해당 시 작성)</td><td colspan="6"></td></tr>
<tr><td colspan="2">연구개발과제명</td><td colspan="6">미생물 게놈 빅데이터 분석을 위한 AI 기반 환경 ICT 융합 글로벌 인재양성</td></tr>
<tr><td colspan="2">전체 연구개발기간</td><td colspan="6">2021. 5. 1. ~ 2022. 8. 31.</td></tr>
<tr><td colspan="2">총 연구개발비</td><td colspan="6">총 500,000 천원
(정부지원연구개발비: 500,000 천원, 기관부담연구개발비 : 0 천원,
지방자치단체: 0 천원, 그 외 지원금: 0 천원)</td></tr>
<tr><td colspan="2">연구개발단계</td><td colspan="3">기초[ ] 응용[V] 개발[ ]
기타(위 3가지에 해당되지 않는 경우)[ ]</td><td>기술성숙도
(해당 시 기재)</td><td colspan="2">착수시점 기준( )
종료시점 목표( )</td></tr>
<tr><td colspan="2">연구개발과제 유형
(해당 시 작성)</td><td colspan="6"></td></tr>
<tr><td colspan="2">연구개발과제 특성
(해당 시 작성)</td><td colspan="6"></td></tr>
<tr><td rowspan="2">연구개발
목표 및 내용</td><td>최종 목표</td><td colspan="6">○ 인력양성

1. 인공지능 기반 환경정화 관련 미생물 게놈 빅데이터 유전체 분석에 능통한 글로벌 융합인재 양성
2. 환경정화 ICT 융합 기술 차세대 인재양성
3. 유전체 및 효소 패턴 분석 딥러닝 모델 전문가 인력 양성

○ 연구목표

1. biLSTM 기반 유전자를 통한 정화 미생물 예측 기술개발
- 환경오염관련 미생물 유전자확보 및 환경정화 관련 유전자 클러스터을 예측할 수 있는 모델을 개발하고자 함
2. 정화 관련 효소 데이터베이스 구축 및 특정 효소군을 위한 분류 기술개발
- 환경오염의 원인 중 하나인 고분자 폴리머 관련 분해 효소들의 유전자 데이터베이스 구축하고 특정 효소의 기질 상호작용에 따라 분류할 수 있는 머신러닝 모델을 개발하고자 함
3. 청정지역과 오염지역의 미생물 유전체 패턴 분석
- 청정 및 오염지역 미생물 유전체 데이터베이스 구축하고 청정지역과 오염지역에 따른 미생물의 진화 패턴을 유전체정보를 이용하여 연구하고자 함</td></tr>
<tr><td>전체 내용</td><td colspan="6">○ 정화 미생물 유전체를 이용한 Genomic island 예측
- 환경오염 정화에 관련된 분야에서는 국내의 경우 생명공학과 인공지능이 융합하여 환경정화를 미생물 유전체 빅데이터를</td></tr>
</table>

			기반으로 해결한 사례가 미비한 실적임 - 이에 본 연구진은 환경정화와 인공지능의 융합 분야에서 인재양성이 필요하며, 인공지능과 생명공학의 융합으로 기존 생물학적 환경 문제해결을 위해 미생물의 유전체 정보를 사용하여 인공지능기반으로 환경정화 미생물에 대한 종합적인 분석을 함 ○ 정화 관련 효소 데이터베이스 구축 및 특정 효소군을 위한 분류 기술 개발 - Cytochrome P450 단백질에 대한 시퀀스 정보와 해당 단백질과 상호작용하는 기질에 대한 InChI, InChIKey, SMILES등에 대한 정보를 수집하고 다양한 Protein, Compound와 관련된 데이터베이스를 조사, 분석 및 수집함 - 다양한 기존 모델을 확인하여 최적화된 모델을 설계하고 세팅된 데이터 셋을 활용하여 모델구축 및 최적화를 진행 중에 있음 ○ 플라스틱 분해 가능효소 예측 기술 개발 - 기존의 알려져 있는 데이터베이스 (PMBD 및 PAZy)와 논문에서 수집하였으며 플라스틱 분해와 관련된 큰 분류의 효소 2가지(Hydrolase 및 Oxidoreductase) 중 hydrolase 계열의 유전자 정보를 수집하였음 - 플라스틱 분해 효소의 정보가 적어 검증된 유전자를 이용해서 Hidden Markov model pattern을 추출하고 이를 활용하여 유사유전자를 확보함 - 플라스틱 분해 효소의 패턴에 대한 정보가 적기 때문에 기존 인코딩 방법이 아닌 다양한 인코딩 방법을 적용하여 최적의 인코딩 방법을 찾고자 하였음 ○ 청정지역과 오염지역의 미생물 유전체 패턴 분석 - 남극대륙 바톤반도 내 남극세종과학기지 인근의 rhizosphere에 해당하는 토양으로부터 미생물을 분리 동정하였음. 유전체 분석 정보를 NCBI에 등록하였으며, 전장 유전체 정보를 활용하여 항생제 내성 관련 Gene cluster와 항생물질 생산 Pathway를 함께 분석하였음 - 알려진 유전체 정보를 활용하여 환경 유래 미생물과 오염지역 유래 미생물의 항생제 내성 유전자의 패턴분석을 실시함
	1단계 (해당 시 작성)	목표	
		내용	
	n단계 (해당 시 작성)	목표	
		내용	

연구개발성과	- 환경오염을 정화하기 위한 생물학적 방법 중 미생물을 이용한 방법에서 기능을 수

	행하는 단백질, 미생물, 오염물질, 생합성과정에 대한 데이터베이스를 구축하였으며, 기존의 알려진 프로그램을 이용하여 환경오염정화관련 Genomic islands를 case study에서 찾아 시각화하였음 - 환경오염에 관한 유전자 중 플라스틱 분해 관련 효소, Cytochrome P450에 대한 단백질 및 이외 기능이 알려지지 않은 단백질 기능을 예측하기 위해 CNN 기반 분류 모델을 개발함 - Cytochrome P450에 대한 단백질 정보와 해당 단백질과 상호작용이 있는 기질 정보를 수집하여 데이터베이스를 구축함 - 중요 플라스틱 분해 효소 시퀀스 데이터를 수집하였고 이를 분해가능한 플라스틱류로 분류하였음 - 수집한 플라스틱 분해 효소 시퀀스를 바탕으로 BLAST 및 HMM을 이용하여 플라스틱 분해 가능성을 가지는 효소군을 수집할 수 있었으며 이를 이용하여 플라스틱별 분해효소를 판별하는 모델을 개발함 - 논문을 통해 플라스틱 분해 효소를 수집하여 데이터 증폭 및 데이터 베이스를 구축하였고, 패턴 파악을 위해 머신러닝, 딥러닝 기술인 CNN, BRNN, RandomForest를 통해 모델을 구축하였음. - 극지방에서 분리한 미생물의 유전체의 해독 및 분석을 완료하였고, 이를 포함한 청정지역과 오염지역의 유전체 패턴분석을 확인하였음.
연구개발성과 활용계획 및 기대 효과	- 미생물 유전체 데이터베이스 및 유전자 분석 기술 확보에 따라, 연구기관 및 기업과의 연계 연구 개발 과제로서, 유용 유전자 예측 도구를 활용한 2차대사물질 대량생산 체계 설립, 유용미생물 개발 및 환경정화 관련 효소 개발을 통해 국내에서 많은 연구가 되지 않은 부분에 대한 융합연구가 가능함 - 기능이 알려지지 않은 단백질 시퀀스를 이용한 기능 예측을 위하여 CNN기반 분류 모델을 활용한 새로운 방법론을 제시하고 융합 연구에 대한 후속 연구를 설계할 수 있음 - 극지 유래 미생물에 대한 항생제 내성 연구를 통해 진화론적, 지리적 특성에 대한 연구를 진행하여 새로운 결과를 도출 할 수 있으며, 비교 유전체 연구를 통해 융합연구를 심화적으로 분석할 수 있음 - 본 프로젝트에 참여하는 국내,외 기관들과 협업을 통한 미생물 유전체 정보를 이용한 기술연구 협업 능력 증진 및 전문 인력 양성과 취업 연계 지원이 가능함
연구개발성과의 비공개여부 및 사유	

연구개발성과의 등록·기탁 건수	논문	특허	보고서 원문	연구시설·장비	기술요약정보	소프트웨어	표준	생명자원		화합물	신품종	
								생명정보	생물자원		정보	실물
	10								1			

연구시설·장비 종합정보시스템 등록 현황	구입기관	연구시설·장비명	규격(모델명)	수량	구입연월일	구입가격(천원)	구입처(전화)	비고(설치장소)	ZEUS 등록번호

국문핵심어 (5개 이내)	미생물 게놈	인공지능	환경정화	빅데이터 분석	효소
영문핵심어 (5개 이내)	Microorganism Genome	Artificial Intelligence	Bioremediation	Bigdata Analysis	Enzyme

< 과제 실적 요약 >

□ 추진실적(요약)

ㅇ 주요 연구 실적

ㅇ 세부프로젝트 1 : 환경정화 미생물 예측

- 환경오염물질의 정화 관련 생합성과정, 미생물, 반응, 효소정보 등을 데이터베이스로 구축함
- 오염물질 정화 관련 Genomic islands 분석을 위한 Pipeline을 구축함
- 알려진 7종의 미생물 유전체 정보를 활용하여 구축한 Pipeline의 타당성을 확인하였고, 결과를 통해 새로운 Genomic islands를 제안함
- 또한 실험으로 증명된 본 연구실에서 보유하는 미생물 유전체 정보를 활용하여 Pipeline의 실제 작동여부를 확인하였음

ㅇ 세부프로젝트 2 : 정화관련 효소군 분류

- CYP관련 효소 데이터 셋 수집을 위한 Positive data, Negative data에 대한 정의 및 데이터 셋을 확보하였으며, 관련 interaction compound 등의 정보도 추가로 수집하였음
- 기존의 효소기능을 예측하는 논문 탐색 및 모델을 비교하여 본 프로젝트의 모델 설계를 디자인함
- 플라스틱 생분해 효소 관련 데이터 셋을 구축함으로써 기존에 알려진 데이터베이스의 한계를 극복하였고, 단점을 보안한 데이터베이스를 추가 구축함으로써 분석에 용이한 데이터 셋을 구축함
- 기존 CNN 모델을 활용하여 제작한 데이터 셋으로 실험을 진행하였고, 추가 최적화 과정을 진행중에 있음
- 플라스틱 생분해 효소 관련 데이터 수집 및 데이터베이스 구축하였고, 데이터 수가 부족하여 추가 분석을 통해 데이터 증폭을 완료함
- Protein sequence 기반 19가지 Encoding 기법을 구현하였으며, 분류 모델을 3가지로 구현하여 결과를 도출함

ㅇ 세부프로젝트 3 : 환경관련 미생물 패턴 분석

- 극한지역에서 분리된 미생물에 대한 동정과 더불어 유전체 분석을 진행하였고, 유전체 정보를 활용하여 극한지방 유래 미생물의 항생제 내성 유전자 관련 연구를 진행하였음
- 알려져 있는 항생제 내성 유전자 데이터베이스를 확인하여 다양한 분석을 시도하였으며, 이는 차후 후속 연구로 계획 중에 있음
- 또한, 기존 알려진 미생물 유전체 정보를 활용하여, 환경유래 미생물과 오염지역유래 미생물의 항생제 내성 유전자를 비교하였고, 패턴 분석을 통해 정량적 분석을 완료하였음

ㅇ 국제 협력교류 연구 실적

ㅇ수업 및 연구회의를 통한 실험설계
- 데이터 수집, 타당성, 구체화 등 데이터 처리에 관한 것뿐만 아니라 전처리에 따른 모델의 차이에 대해 습득하였고, 프로그래밍 언어, 머신러닝과 딥러닝에 대한 이론적 이해에 도움을 받음
- 또한 머신러닝에 대한 기초 연구를 통해 모델 구현 이해도를 증가하였으며, 소속된 연구팀에 있는 학생들과 연구 교류를 통해 피드백을 받았고, 머신러닝과 딥러닝을 활용한 연구에서 문제 해결 능력이 향상되었음
- 세부적인 모델 설계 등 파견 대학에서 수강한 머신러닝 수업을 통해 각 연구 분야에 적용시키기 위한 필수 역량을 증진하였고, 결과적으로 모델 구현 및 이해에 많은 도움이 되었음
- 또한 유전체 관련 특강을 통해 연구의 이해도를 향상시켰으며, 파견 학생간의 활발한 토의를 통해 생물학적 기본지식과 컴퓨터 공학적 기본지식을 나눌 수 있는 기회가 되었고, 이는 연구의 다각적 이해를 가능하게 하였음

ㅇ 인력 양성 추진 실적

ㅇ 생명공학 지식과 컴퓨터 기술을 융합하여 글로벌 인재 양성을 도모하였으며, 다른 연구원들과 원할한 의사소통을 통해 문제 해결능력, 문제 이해도, 협업 능력을 증진하였음

ㅇ 또한 해외에서 진행하는 다양한 행사에 참여하여 폭넓은 이해와 앞으로 연구 가능성, 방향 등을 고려할 수 있는 좋은 기회가 되었으며 이는 국가적 차원에서도 환경 문제 분석 파악에 많은 도움이 될 것으로 사료됨

ㅇ 융합 연구는 융합적 사고를 증진시켜줄 뿐만 아니라 융합연구를 하지 않는 사람과도 폭넓은 의사소통이 가능한 인재 양성에 도움이 될 것임

ㅇ 또한 본 프로젝트의 사후관리 시스템 등을 이용하여 더 다양한 융합 연구를 접해 볼 수 있을 것으로 사료되며, 환경오염을 생물학적 관점에서 더 나아가 컴퓨터 공학적 관점에서 시사 하여 환경문제에 대한 다양한 관점의 해결방안을 도출하는 인력을 양성할 수 있음

□ 주요성과

(단위 : 건, 명, 백만원 등)

구분				목표	실적
인력양성 성과	수혜인력	석사	남	-	-
			여	-	-
		박사	남	-	-
			여	-	-
		합계		-	-
	배출인력	석사	남	2	2
			여	1	1
		박사	남	3	3
			여	3	3
		합계		9	9
연구성과	논문	SCI급		1	1
		비SCI	KCI급	-	-
			기타	-	-
		합계		1	1
	학술대회	국제		1	9
		국내		-	-
		합계		1	9
특허성과	특허출원	국제		-	-
		국내		-	-
		합계		-	-
	특허등록	국제		-	-
		국내		-	-
		합계		-	-
기술이전				-	-
상용화(백만원)				-	-
기술료(백만원)				-	-
성과홍보				-	-
시제품 제작실적				-	-
S/W 등록실적				-	-
기술문서				-	-
기타				-	-

* 당초 사업계획서에 기재하지 않은 성과가 있는 경우에도 건수를 작성하고, 해당 성과가 없는 경우 "-" 표기
* 인력양성 성과
  - 수혜인력은 파견 개시일을 기준으로 성과작성, 배출인력은 파견종료일을 기준으로 성과 기재
  - 수혜인력 : 본 과제에 참여하여 파견을 개시한 인력
  - 배출인력 : 본 과제에 참여하여 6개월 이상 해외파견을 수행한 인력
* 연구성과 계산시 과제 참여연구원 여러명이 공동으로 게재한 경우 1건으로 계산
* 논문실적
  - 글로벌 핵심인재 양성지원 사업 사사문구(단독, 공동)가 있는 것만 기재
  - 본 과제의 참여연구원이 공동저자인 경우에만 인정
* 특허
  - 주관연구개발기관 : 주관연구개발기관 단독 또는 공동명의로 출원/등록
  - 과제 참여인력이 발명인으로 등재되어야 함

- 본 과제로 인해 발생된 특허이어야 함(무관한 특허 제출 시 감점대상)

* 기술이전은 기술이전 실시계약서가 있는 경우에 한해서만 기재

□ 장비.기자재 취득목록

순번	장비등록번호	장비명	모델명	사용용도	취득금액	취득일자	제작사	제작국가
1								
2								
3								

* 취득가격이 3천만원 이상인 장비 또는 취득가격이 3천만원 미만이라도 공동 활용이 가능한 장비로, 취득 후 30일 이내에 국가과학기술종합정보시스템장비등록서비스에 등록하고 '국가연구시설장비정보등록증'을 발급받은 장비 기재

목 차

붙임

1. 연구개발과제의 개요

1) 추진 배경

(1) 환경시장 및 인공지능

◦ 2019년 환경산업기술원에서 조사된 세계 환경시장 규모는 1조 2,443억 달러로 최근 7년간 연평균 3.6%씩 증가하는 추세임 (한국환경산업기술원, 2018)

◦ 빅데이터 및 데이터 분석 시장도 2018년 1,688억 달러로 성장하였으며, 2022년에는 2,743억 달러까지 증가할 것으로 전망됨 (그림 1)

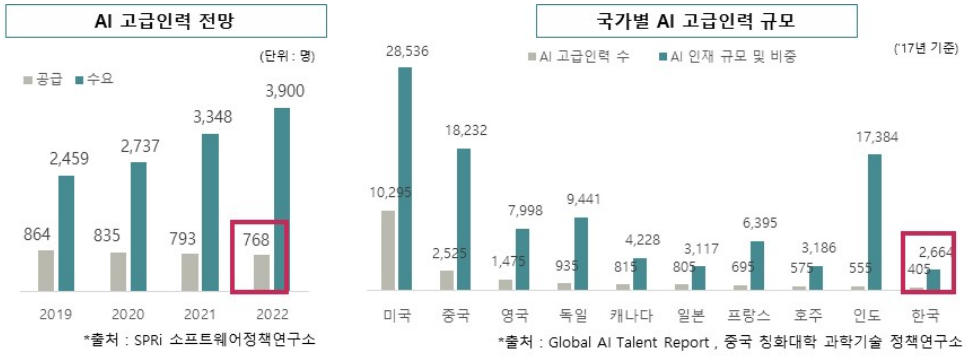
◦ 인공지능을 활용한 플랫폼은 전자, IT 기술뿐만 아니라 환경, 의료 등 다양한 분야에 활용될 수 있는 분야임. 세계 인공지능 시장 규모는 17년도부터 매년 연평균 23%로 성장할 것으로 보여 25년에는 약 66조 원까지 도달할 것이라 예상되었음 (출처 : Statista) 이렇듯 성장하는 시장의 규모에 따라 실질적으로 이를 수행할 수 있는 각 필드 내 전문 인력의 충분한 공급이 필요함

<그림 1> 환경시장, 인공지능시장, 빅데이터시장 규모 (출처 : 산업연구원, Statista)

(2) 관련 산업 인력 통계

◦ 국내 바이오산업 실태조사 결과보고서에 따르면, 국내 바이오산업 종사자는 19년도 기준 학·석사가 증가하는 추세이지만 2018년도 조사된 산업기술인력 수급 실태조사에서는 직무수행을 위한 자질·근로 조건에 맞는 인력이 부족하다고 함

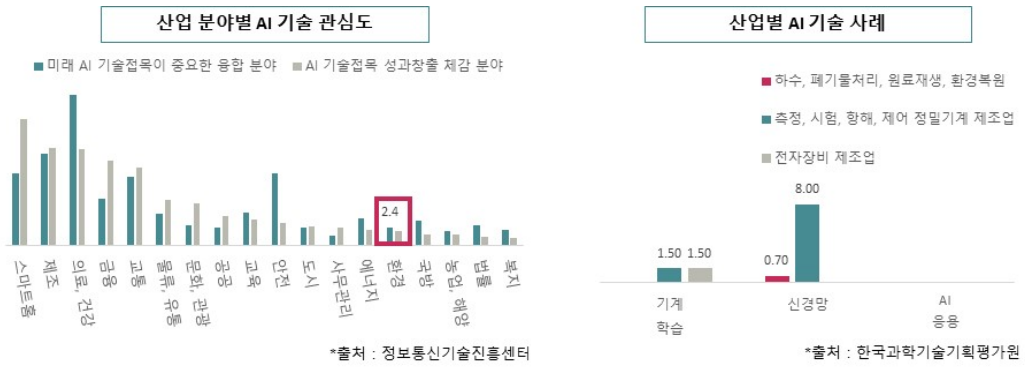
◦ AI 핵심 인재의 경우 대부분 미국, 중국 등 소수 국가에 집중되어 있는데, 국내의 경우 1.3%인 2,664명으로 세계에서 15위 수준임. 격차 해소를 위한 석·박사급 고급 인력 양성에 대한 체계적인 대책이 매우 시급한 실정임 (그림 2)



<그림 2> AI 관련 인력 전망 및 규모

(3) 연구수행 필요성

◦ 산업 분야별 AI 기술 관심도를 살펴보면, 20.1%의 높은 관심도를 갖는 의료·건강 분야에 비해 2.4%의 환경·에너지 분야는 비교적 주목받지 못하고 있음. 관심도가 적은 만큼 그에 따른 성과 창출 체감도 또한 현저히 적은 것으로 조사됨. 또한 대체로 정밀기계 제조업으로 쏠려있는 AI 기술 적용 분야로 인해 선행 사례가 매우 적으므로, 이에 대한 기반과 기술 도입이 시급함



<그림 3> 산업별 AI 기술 관심도 및 사례

◦ 그리하여 본 연구팀은 비선호 항목인 환경 분야에 접근하여 이와 AI를 접목한 새로운 기술을 만들어 보고자 함. 융합형 인재를 배출하게 되면, 다음과 같은 융합형 기업에 취업을 지원하는 것이 가능하고 더 다양한 시각의 인재들을 성장시킬 수 있을 것이라 사료됨

<표 1> 바이오산업과 AI 융합 관련 회사

주요 내용	관련 기업
컴퓨터 비전 기술과 AI를 적용한 작물 생육 모니터링 및 영양분 결핍 탐지, 병해충 감지 기술의 산업화	이스라엘 기업 Prospera
드론 및 인공위성 영상정보를 AI와 결합한 사업화	FarmLogs
AI 로봇을 이용한 농작물 수확 및 농작업(적엽, 잡초 제거, 농약 살포 등)에 적용	Priva, BOSCH, Blue River Technology
살충제와 제초제를 최소한으로 사용할 수 있도록 분당 5천 개의 어린 식물을 촬영하여 새싹과 잡초를 식별하여 잡초에만 제초제를 적용하는 딥러닝 솔루션	Blue River Technology(미국)

◦ 급격한 도시 성장에 따른 부작용으로 무분별한 개발 및 사용이 증가하였고, 현재는 COVID-19 사태와 격리 음식으로 인한 플라스틱의 폐기가 급증하여 다양한 환경오염 문제가 지속적으로 대두됨

◦ 본 연구팀은 미생물 유전체 관련 빅데이터와 인공지능을 융합하여 환경오염에 대한 해결법을 연구하고 더 나아가 지식과 컴퓨터 기술을 다룰 수 있는 글로벌 인재를 양성하는 것이 목표임

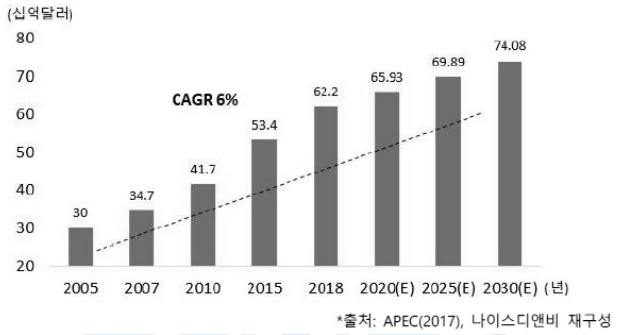
2) 정부 지원 필요성

(1) 산업계 수요 조사

◦ 환경산업기술원에 따르면 2018년 기준 세계 환경산업 시장규모는 1조 2,443억 달러에 달하며, 이는 2020년도 정부 예산안(513조 5,000억 원)의 3배 수준의 규모로 전망

◦ 세계 환경산업(서비스) 시장에서도 상·하수도 및 폐기물 분야가 전체의 77%를 차지하며, 토양 및 지하수 복원사업은 전체 시장 중 5~8%를 차지하는 수준임. 지역별 시장규모는 미국 33%, EU(서유럽) 24%, 일본이 13%, 그리고 중동시장이 11% 정도를 차지함

◦ 특히, 환경오염 정화 관련 토양지하수분야 시장규모는 2005년 300억 달러에서 2018년 622억 달러로 성장하였으며, 동일 추세를 가정할 때 2030년에는 741억 달러 시장을 형성할 것으로 전망



<그림 4> 세계 토양지하수분야 환경산업 시장 규모

◦ 2018년 환경산업통계조사에 따르면 2017년 토양지하수 환경산업 시장규모는 전체 환경시장 규모인 98조 8,188억 원의 1%가 조금 넘는 9,940억 원이었으며, 글로벌 경기침체 등의 영향으로 2017년도 이후의 정화사업 시장은 마이너스 성장을 할 것으로 예상

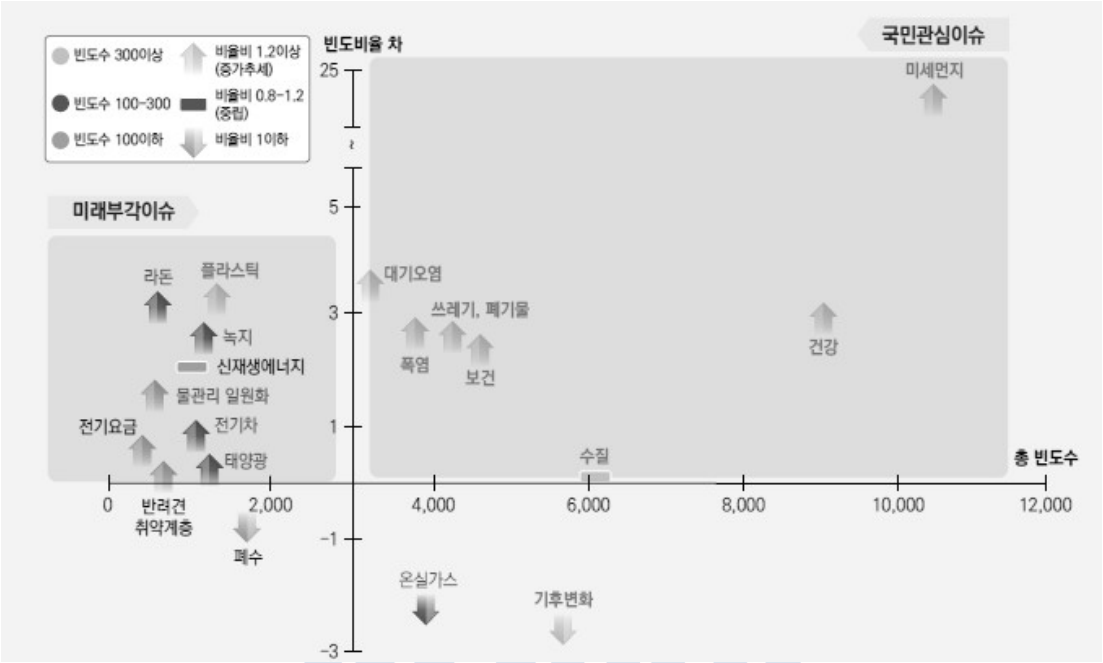
◦ 환경오염 관련 산업은 이미 발생한 오염을 정화하는 기술 분야에 연구개발이 치중되어 있음을 알 수 있으나, 오염조사기술, 오염관리 기술 등 오염을 미리 방지하거나, 모니터링, 관리, 및 판단하는 기술이 미미함

◦ 상대적으로 정화가 어려운 난분해성, 복합오염물질을 정화하는 오염정화기술(난분해성, 복합오염) 분야의 경우 중금속이나 유기화학 물질을 정화하는 기술 분야보다 기술 발전 속도가 상대적으로 느린 것으로 예상함

(2) 과제 추진 및 정부 지원 필요성

◦ 환경산업은 일반적으로 국내 환경정책이나 국제 환경규제 등 법적 및 제도적 요

인에 의한 수요가 창출되며, 환경오염의 발생 과정, 오염물질 특성, 처리방법 등에 따른 다양한 현장에 대응하는 산업적 특성에 따라 연구 방향이 설정됨에 따라 정부 주도적인 연구 및 지원이 필요한 실정임



<그림 5> 빅데이터 분석을 통한 환경 이슈의 변화

◦ 환경보전을 위한 활동과 시설에 따른 공공 복지적 사회 간접 자본의 성격이 강하며, 일반 대중과 광범위한 지역 환경보전을 위한 공공재적 특성을 강하게 띠는 산업으로 이에 따른 연구 방향 및 지원이 필요한 산업임

◦ 화이트 바이오산업의 경우, 친환경을 기반으로 성장하는 산업인 만큼, 정부의 적극적인 시장 창출 지원으로 민간 투자 견인이 매우 중요하며, 플라스틱 사용 규제, 유전자변형기술 규제가 엄격한 가운데, 화이트 바이오 제품 보급 · 확산을 위한 제도, R&D 등의 지원이 부족한 실정임

◦ 바이오 분야 R&D 투자비율에서는 '19년도, 레드(66.3%), 그린(26.7&), 화이트(7.0%) 순으로 정부 지원이 이루어지고 있음, 표 8은 최근 환경정화와 관련하여 정부에서 발주한 연구과제 정보를 나타냄

◦ 2010년 이후 환경 분야 주요 이슈 분석을 위해 상위 20개 언론사 기사를 바탕으로 분야별 · 연도별 키워드에 대한 시계열 빈도 수 분석 수행결과, 모든 환경 분야에서 기사 수는 증가 추세를 보였으며, 이는 환경에 관한 관심 증대를 의미함 (그림 5)

3) 수행기관 추진역량

가. 국내 수행기관 전문성

◦ SW중심대학 등 융합관련 과제를 수주함 (그림 6)

[SW중심대학, 104억원] [4차산업혁명 혁신선도대학] [LINC/LINC+사업, 300억원] [BK21 4단계, 35억원]

<그림 6> 주관연구기관인 선문대학교의 전문성

◦ 2018년 유전체 기반 바이오-IT 융합 연구소와 2019년 1학기 일반대학원「바이오빅데이터융합전공」을 개설하여 바이오와 IT를 융합된 전반적인 연구를 수행중임

◦ 선문대학교의 SW 중심대학산업 산학공동 과제를 진행하였고, 공동연구를 위한 업무 체결을 확약함. 또한, UNLV와 「이미지를 사용한 질병유무 예측 알고리즘」, 「DpreEC: Interpretable deep learning enhances prediction of EC numbers」와 같은 협력 연구를 진행함

<표 2> 주관연구개발기관의 전문성

이름	교육적역량	연구적역량	국제적역량
김정동 교수	• 바이오 빅데이터의 실시간 분석 기법 및 이상 탐지 알고리즘 개발 • 바이오 데이터의 실시간 분석을 통한 패턴 추출 및 시각화 연구	• ‘LLAD: Life-Log Anomaly Detection Based on Recurrent Neural Network LSTM’ 외 23편의 SCI급 및 KCI급 연구논문 게재	• 국제학술대회 홍보의장 - International conference on digital contents (‘19~) • 국제학술대회 프로그램 위원회 - IMCETI(‘14~), ACM RACS(‘18~)
이 현 교수	• 다중체학 데이터를 유전체 기반의 생물학적 딥러닝 모델 표현 • 효과적이며, 해석 가능한 통합 모델 및 학습 방법을 연구 개발	• ‘심층 신경망을 이용한 보행자 검출 방법’ 외 18편의 KCI급 연구논문 게재 및 30개 의 특허 보유	• 재미과학기술자협회 편집위원 활동 • 해외대학 및 연구소와 공동연구 - 브리지포트대학, 네바다 주립대학
오태진 교수	• 극지 희귀박테리아 유래 생리활성물질 탐색 및 게놈 분석 시스템 구축 • 미생물 유래 구조변형효소 관련 연구 및 신규 물질 생산 시스템 구축	• ‘(SCI) A computational approach to identify CRISPR-Cas loci in the complete genomes of the lichen-associated Burkholderia sp. PAMC28687 and PAMC26561’ 외 다수 SCI 급 논문 게재 및 10개의 특허 보유	• 국제전문학술지 편집위원 활동 - journal of Applied Biological Chemistry (‘19~) • 해외 연구소 공동 연구 - 네바다 주립대학, 네바다 사막연구소

나. 공동연구개발기관 전문성

◦ 극지연구소는 극지 지식 창출 · 활용을 통한 기후변화 등과 같은 글로벌 이슈를 해결하고 국가적, 세계적 현안을 해결하고 국익을 확보하여, 극지에 대한 국내외 영향력을 확대하기 위해 연구를 하는 전문기관임 (그림 7)

◦ 1년 내내 빙설의 영향 아래에 극고기압의 지배를 강하게 받는 극한환경, 기상관측소, 군사기지, 어업기지 등과 같은 황량한 무인 지대와 얼음 바다가 이어지는 환경에 관한 연구를 주도

◦ 청정지역에서 분리된 생물자원, 유전체 분석, 항생물질 변형 효소의 생화학적 특성 분석, 기존 상용화 효소와의 비교, 항생물질 변형체 정보, 미생물 유전체의 종 분화와 같은 진화관련 데이터 등을 활용하여 오염된 환경에서 분리된 생물자원 등의 정보가 포함된 데이터베이스를 통해 비교 데이터를 얻을 수 있는 가장 적절한 환경 데이터를 연구하는 국책 연구기관으로 특화되어 있음

[극지연구소 외부 전경]

[극지방 생물 연구 기지]

KOPRI
극지연구소
Korea Polar Research Institute

[극지연구소 로고]

<극림 7> 공동연구 개발기관 주요 전경

4) 해외기관 추진 전문성

가. 기관명: University of Nevada, Las Vegas (UNLV)

◦ 네바다, 라스베가스 대학교는 (이하 네바다 주립대학교)는 네바다 주에서 가장 큰 도시인 라스베가스에 위치해 있으며, 법학, 의대, 치대 전문대학교와 15개의 단과대학교를 포함하는 네바다 주를 대표하는 대학교로 여겨짐

◦ 최근 라스베가스의 의료 및 환경 관련 연구의 중요성이 대두됨에 따라, 2015년에만 500억 이상의 연구비를 투자한 바 있으며, 최근 컴퓨터 관련 학과 또한 급성장 중에 있음. 2018년 카네기 분류에서 미국 내 4천여 개의 대학교 중 상위 131개의 연구학교 (R1: Doctoral Universities – Very high research activity) 중 하나로 선정되었고, 이는 전미 상위 3%에 해당함

◦ 네바다 주립대학교의 National Supercomputing Center는 여러 대의 Supercomputer를 운용하고 있음

◦ 최근 최상위 글로벌 IT 기업인 Google이 2차례에 걸친 대규모 투자를 통해서 네바다의 Henderson (2019년), Storey County (2021년) 지역에 세계적인 규모의 데이터 센터를 설립, 운용하면서 네바다 지역 내에서 경제적인 효과를 낳을 뿐만 아니라, IT 분야에서의 일자리까지 창출되는 기대효과를 실현함

나. 강민곤 교수팀 (DataX Lab)

◦ 네바다 주립대학교에 위치해있는 생물정보학 연구실로, 해석 가능한 딥러닝 연구(Interpretable Deep Learning)을 선도하는 연구팀으로 알려짐. 생물정보학 내에서 보여지는 적은 샘플 수와 딥러닝 모델 해석의 한계를 극복하기 위해, 효과적이며 생물학적으로 해석 가능한 새로운 딥러닝 모델을 개발하는 연구를 주도함

◦ DataX 연구실은 유전체, 단백체를 비롯한 다중생물 유기체 정보 뿐 아니라, 의료용 이미지 데이터와 전자차트 데이터에 이르기까지 다양한 데이터 형태를 종합적으로 분석하여 통합적인 해석을 도출하는 딥러닝 연구를 선도 중에 있음

◦ 미국 내에서는 하버드 의대, Memorial Sloan Kettering Cancer Center, Cincinnati Children's Hospital Medical Center, 그리고 사막연구소와 협업 중에 있으며 국내에서는 선문대학교, 극지연구소, 경상대학교 병원과의 활발한 공동연구를 통해서 딥러닝을 활용한 생물정보학 분야를 선도하고 있음

◦ 최근 3년 이내에 International Journal of Molecular Sciences (IJMS, IF: 5.9), Briefings in Bioinformatics (IF: 11.62), Science Advances (IF: 13.116), Proceedings of the National Academy of Sciences of the United States of America (PNAS) (IF: 9.412), Monthly Notices of the Royal Astronomical Society (MNRAS) (IF: 5.356) 등의 세계적으로 영향력 있는 학술지에 연구결과를 발표하였고, 높은 인용지수를 보이는 것으로 보아 세계적으로 빅데이터 분석과 딥러닝을 활용한 생물정보학을 주도하고 있는 것으로 확인됨

다. 본 연구에서 해외기관 추진의 필요성

◦ 생물관련 빅데이터 분석과 딥러닝 연구를 수행에 필수적인 장비로서, DataX lab에 15대의 GPU 서버와 Intel Xeon Silver 급의 4대의 CPU 서버 보유

◦ 빅데이터와 딥러닝을 활용한 생물정보학 주제의 논문 작성과정에서 해외 지도인력과 해외 기관 연구원들과의 토론과 첨삭을 받을 기회가 마련됨. 논문 구성 및 작성에 대한 비교 분석을 통해서 효과적인 연구 성과를 발표하는 역량을 기를 수 있는 기회를 마련함

◦ 선문대학교와 UNLV 강민곤 규수 연구팀은 2019년부터 IITP에서 주관한 글로벌 핵심인재 양성 프로그램을 시작으로 연구 협업을 진행 중에 있고, Enzyme Commission number prediction 이 가능한 tool 개발 연구가 진행 중이며, 월등한 성능을 보이는 실험 결과를 토대로 학술지 투고를 앞두고 있음

2. 연구개발과제의 수행 과정 및 수행내용

1) 연구 수행 개요

가. 연구 수행 개요

통합프로젝트	세부프로젝트	연구분야	연구 목표	협력 연구기관	석사	박사
환경정화 관련 미생물 유전체 종합 분석	환경정화 미생물 예측	게놈 유전체 분석, 딥러닝	biLSTM기반 미생물 정화 유전자 예측 기술개발	선문대학교 - UNLV	2	2
	정화관련 효소군 분류		데이터베이스 구축 및 정화관련 효소 예측	선문대학교 - UNLV	2	2
	환경관련 미생물 패턴 분석		환경정화 미생물탐지를 위한 청정지역 미생물과 오염지역 미생물의 유전체 패턴 분석	극지연구소 - UNLV - 네바다 사막연구소	-	1
합 계					9	

2) 연구 내용

나-1. 프로젝트 추진일정

순번	추진내용	2022년								2023년						비고
		5월	6월	7월	8월	9월	10월	11월	12월	1월	2월	3월	4월	5월	6월	
1	biLSTM 기반 유전자를 통한 정화 미생물 예측 기술개발															
2	정화 관련 효소 DB 구축 및 특정 효소군을 위한 분류 기술개발															
3	청정지역과 오염지역의 미생물 유전체 패턴 분석															
주요 Milestone 완성점에서의 수행결과		biLSTM 기반 유전자를 통한 정화 미생물 예측 기술개발					정화 관련 효소 데이터베이스 구축 및 특정 효소 군을 위한 분류 기술개발			청정지역과 오염지역의 미생물 유전체 패턴 분석						

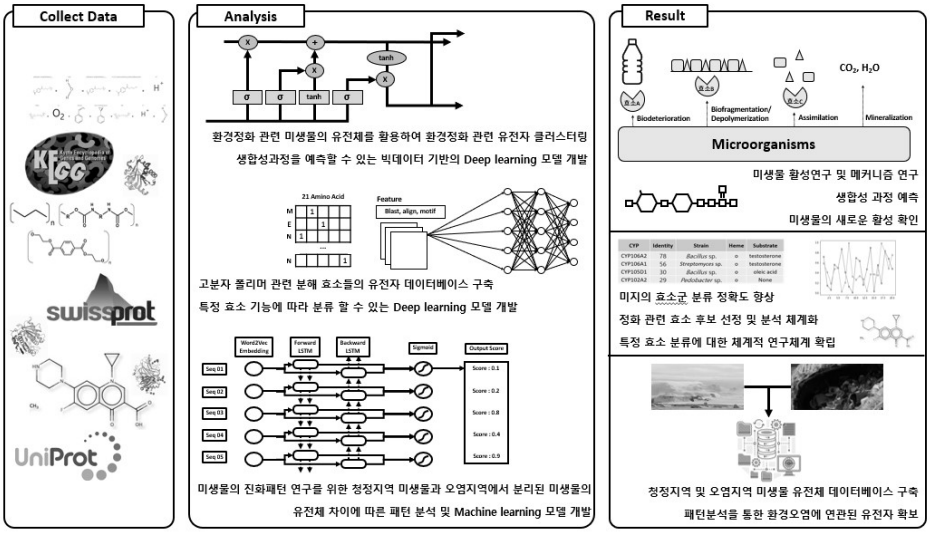
나-2. 통합프로젝트

(1) 인공지능 기반 환경정화 관련 미생물 게놈 빅데이터 유전체 분석 (그림8)

◦ 환경정화 관련 미생물의 유전체를 활용하여 환경정화 관련 유전자 클러스터링 및 생합성과정을 예측할 수 있는 빅데이터 기반의 딥러닝 모델을 개발

◦ 환경오염에 따른 미생물의 진화 패턴을 연구하기 위해, 청정지역 미생물과 오염지역에서 분리된 미생물의 유전체 차이에 따른 패턴을 분석

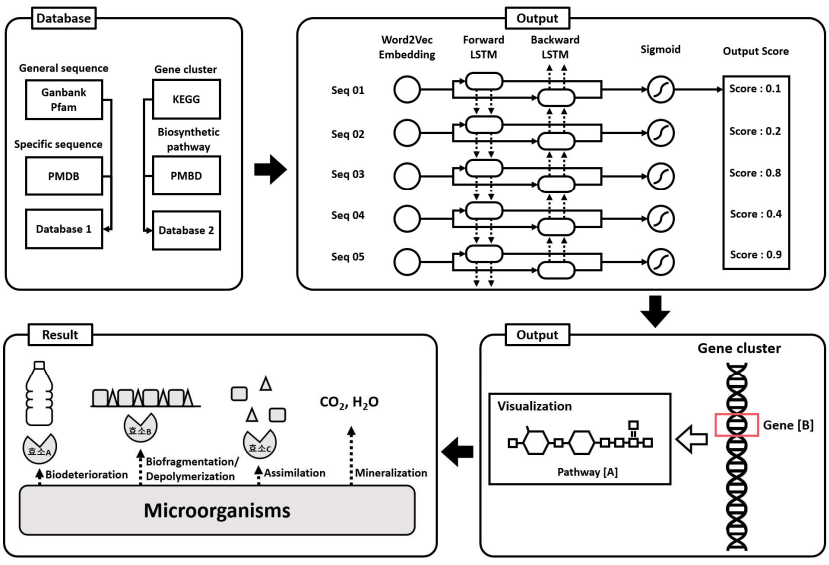
◦ 환경오염의 원인 중 하나인 고분자 폴리머 관련 분해 효소들의 유전자 데이터베이스 구축하고 특정 효소 기능에 따라 분류할 수 있는 딥러닝 모델을 개발



<그림 8> AI 기반 미생물 게놈 빅데이터 유전체 분석의 전체 개념도

나-3. 세부프로젝트

(1) 세부프로젝트1 : biLSTM 기반 유전자를 통한 정화 미생물 예측 기술개발



<그림 9> biLSTM 기반 환경정화 미생물 예측을 위한 개념 모델

○ 특성 환경 미생물의 군집 혹은 총합에 대한 중요성이 집중됨

◦ 유용 미생물 은행 등 빅데이터 활용 인프라를 지원, 미생물 수집, 유전체 분석 및 정보 제공, 미생물 빅데이터 서비스 제공

◦ 인체・작물 등과 미생물 군집간의 상호작용 분석이 가능

◦ 기존의 방대한 공개된 바이오 데이터와 선문대 대사공학연구실에서 보유한 데이터가 존재하지만, 빅데이터 분석과 새로운 딥러닝 모델 개발을 통해서 플라스틱 생분해 과정에 대해 인간이 쉽게 찾기 힘든 생물학적 메카니즘을 효율적으로 찾을 수 있음

○ 수질오염・폐기물 등 환경개선 관련 국내 산업 규모 확대

◦ 미생물 상호작용 기술을 적극적으로 활용해 수질 개선제, 난분해성 폐기물(폐비닐 등) 처리제, 화학살균・소독 대체재 등에 관한 기술개발 강화

◦ 국내 산업 규모 [2019] 2.9조 원 -> [2020] 7.3조 원으로 연평균 8.7% 증가

○ 선문대학교 대사공학연구실에서 보유 중인 플라스틱 및 polyhydroxyalkanoates 분해 관련 미생물을 이용하여 테스트 데이터셋으로 사용할 예정이며, 해당 데이터셋이 정화능력이 있는 미생물을 예측하고자 함

<표 3> 선문대학교 대사공학연구실 보유 플라스틱 및 polyhydroxyalkanoates 관련 미생물

종 이름	분해예상 오염물질	레퍼런스
Arthrobacter	HDPE	Zhiqiang et al., 2019
Bacillus	polyhydroxyalkanoates	Nielsen et al., 2017 Zhiqiang et al., 2019
Brevundimonas	HIPS	Zhiqiang et al., 2019
Brevibacterium	DEHP	Zhiqiang et al., 2019
Burkholderia	polyhydroxyalkanoates	Nielsen et al., 2017 Zhiqiang et al., 2019
Flavobacterium	Nylon, DEP, DETP, PA	Zhiqiang et al., 2019
Janthinobacterium	PVA	Zhiqiang et al., 2019
Klebsiella	HDPE, DMI	Zhiqiang et al., 2019
Microbacterium	PVA, MMP, DEHP, DEP	Zhiqiang et al., 2019
Ochrobactrum	PVC	Zhiqiang et al., 2019
Pseudomonas	PE, PVC, P3HP, PEA	Zhiqiang et al., 2019
Paenibacillus	PE, PLA, PCA	Zhiqiang et al., 2019
Pantoea	LDPE	Zhiqiang et al., 2019
Psychrobacter	PLC	Zhiqiang et al., 2019
Rhodococcus	PE, PCL, PS, DBP	Zhiqiang et al., 2019
Streptomyces	PHB, PHBV, PVA	Zhiqiang et al., 2019
Stenotrophomonas	PVA, PHB, PVA	Zhiqiang et al., 2019
Sphingomonas	PVA, DPP, MMP, DMP	Zhiqiang et al., 2019
Sphingopyxis	PVA, PEG	Zhiqiang et al., 2019
Sphingobium	DBP, DINP, DMP	Zhiqiang et al., 2019
Staphylococcus	PU, PE	Zhiqiang et al., 2019

○ 플라스틱 생분해 과정

◦ 생분해 연관 효소와 생합성과정은 미생물 종에 따라 더욱 복잡할 것으로 예상

◦ 플라스틱 생분해과정은 역할에 따라 크게 4가지(Biodegerioration, Biofragmentation, Assimilation, Minerallization)로 플라스틱을 변질시킨 후, 노출된 표면부터 잘게 분해함

○ 데이터베이스

◦ Genbank : 공개적으로 사용 가능한 모든 주석이 달린 DNA 데이터베이스

◦ Pfam : 다중 서열 정렬, HMM으로 표현되는 단백질 패밀리 모음으로 단백질이 가지고 있는 도메인을 기준으로 단백질 패밀리를 분류함

◦ KEGG pathway : 분자 상호작용, 반응 및 관계 네트워크에 대한 지식을 나타내는 manually pathway

◦ PMBD : a Plastics Microbial Biodegradation Database

○ 정화 관련 유전자 클러스터링

◦ 전처리 : Database 1을 이용하여 단백질 시퀀스의 도메인을 예측하고 분석 결과를 시각화하기 위하여 전처리 데이터에 EC number를 labeling

◦ 모델 : 시퀀스를 Embedding Matrix로 변환하기 위하여 word2vec를 기반으로 제작된 Pfam2vec를 활용하며, Gene cluster를 분석하기 위하여 biLSTM을 사용

◦ 시각화 : biLSTM 모델의 output 결과로 얻은 예측값과 Labeling된 데이터 결과를 이용하여 기존 알려진 생합성과정에 관련된 EC number를 찾아 pathway를 시각화하고자 함

◦ 예측 모델을 해석하여, 플라스틱 생분해에 관련된 단백질 시퀀스의 패턴 이해

○ 1세부 프로젝트의 최종 목표

◦ 국내에서 사용 가능한 정화 관련 미생물에 대한 기초 연구인 활성 연구 및 메커니즘 연구

◦ IT와 융합을 통해 기존에 알려지지 않은 생합성과정을 예측

◦ 미생물의 알려지지 않은 새로운 활성을 알아내어 상업화할 수 있는 미생물 확인

(2) 세부프로젝트2 : 정화 관련 효소 데이터베이스 구축 및 특정 효소군을 위한 분류 기술개발

○ 환경 정화기능이 우수한 미생물 및 효소의 기존연구

◦ 자주 사용되는 염화비닐, 스티로폼 또는 플라스틱은 생활 패턴에 따라 재활용하기가 어려우며, 자연분해가 어려워 환경오염의 주요 원인임

◦ 현재까지 연구되어 온 미생물 및 효소를 통해 다양한 오염물질 분해가 가능 (표 4)

<표 4> 정화 관련 미생물 및 효소의 기존연구

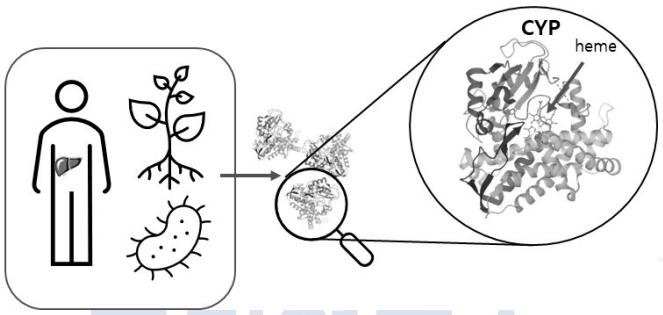
Organism	효소	기능	특허 및 논문
Rhodococcus sp. 24 (미생물)	-	페놀 분해, BTEX 분해	10-2029-0590000, 2019
Microbacterium sp. 28	-	페놀 분해, BTEX 분해	10-2029-0590000, 2020

(미생물)			
Pseudomonas sp. GM1 (미생물)	-	페놀 분해, BTEX 분해	10-2029-0590000, 2021
Novosphingobium fluvii (미생물)	-	환경호르몬 프탈레이트 분해	10-2018-0153685, 2018
Candidatus Methanoliparia (미생물)	canonical methyl-coenzyme M reductase (MCR)	탄화수소의 긴 체인을 산화	Rafael Laso-Pérez et al, 2019
꿀벌부채명나방 (곤충)	esterase, lipase, cytochrome P450	폴리에틸렌 분해	Hyun et al, 2019

○ ‘cytochrome P450 (CYP)’ 소화효소 (그림 10)

◦ 동물, 식물, 균류 등 모든 생명체에 존재하고 있으며, 주요 기능은 물질에 수산기를 붙임. 철이온이 함유된 헴 보조인자를 포함하는 구조적인 특성이 있으며, 촉매 반응의 핵심임

◦ 각 효소군의 촉매 작용 기전 및 기질 특이성이 다르며, 고유의 유전자 보존 서열이 존재



<그림 10> CYP 효소의 구조적 특성

○ ‘Cytochrome P450 (CYP)’ 분류에 대한 비효율적 측면

◦ CYP 분류법에서, Super family는 숫자, 다음 sub family는 알파벳과 숫자가 차례로 나오는 형식임. 타 연구실 저명한 박사를 통하여 분류 번호를 부여받으나 비효율적임. CYP homepage 웹사이트가 운영되어 CYP 데이터가 있으나, 2004년 이후 업데이트되지 않음

◦ 명확한 sub family 지정 없이, super family만을 증명하여 연구되기도 하지만 신뢰할만한 결과를 도출하기까지 단계적인 과정이 필요하여 편리성이 부족함

○ 데이터베이스

◦ UniProt/Swiss-Prot에서 정형화된 패턴의 CYP Sequence와 Label를 데이터베이스화

○ CYP 예측 모델

◦ 전처리 : Uniprot에서 정형화된 패턴의 CYP Sequence와 Non-CYP Sequence데이터를 Biopython 라이브러리를 활용하여 Embedding Matrix로 변환 후 예측모델에서

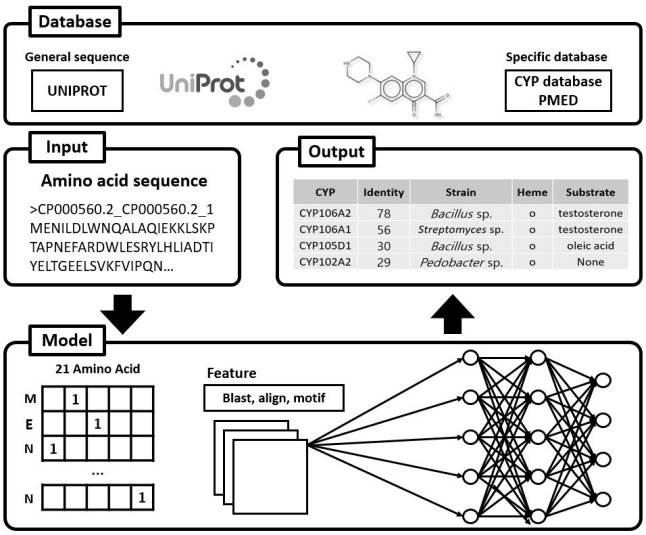
학습, 평가, 검증 데이터로 사용

◦ 1차 모델 & 2차 모델 : Embedding Matrix로 변환된 Sequence데이터를 CNN을 활용하여 CYP 여부 예측 모델 개발 및 CYP 슈퍼 패밀리 분류 모델 개발

○ 2세부 프로젝트의 최종 목표

◦ 기존 복잡한 분석 프로세스를 단순화하여 더 정확한 미지의 효소군 분류

◦ 정화 관련 효소 후보 선정 및 분석 체계화 및 자동화된 특정 효소 분류를 통해 빠른 후속적 연구 진행



<그림 11> 특정효소군을 위한 분류 기술 구조도

(3) 세부프로젝트3 : 청정지역과 오염지역의 미생물 유전체 패턴 분석

○ 환경오염에 따른 미생물의 변화

◦ 오염된 환경에 적응하기 위하여 미생물은 유전자의 추가 · 제거 및 변화를 겪고 이에 따라 특정 유전자들은 상이한 발현 패턴 보유

◦ 환경에 적응한 미생물의 유전체 분석은 유용 유전자 수집 및 진화론적 의미 보유

◦ 또한, 이와 유사하게 북극 · 남극과 같은 극지방은 매우 춥고 건조하는 등 특수한 환경조건에서도 미생물 특유의 종 특이성을 보유

◦ 여러 차례의 빙하기를 거치면서 생긴 동토 및 얼음층은 고대 지구의 환경을 그대로 간직하고 있어 과거 환경에 대한 정보를 얻을 수 있는 좋은 지역임

◦ 극지방 미생물 유전체에 대한 데이터베이스는 NCBi genome database에 대해 일부 수록된 상황이나 일부 분석만 실행된 실정임

◦ 선문대학교 대사공학연구실과 극지연구소는 지속적인 극지 샘플 확보 및 미생물 분리 · 확보, 중요 미생물의 유전체 분석을 통해, 해당 데이터셋 확보

○ 청정지역 및 오염지역의 유전체 데이터베이스

◦ 미생물의 유전체 데이터베이스는 NCBI genome database가 있으나, 미생물 한정, 특히 지역에 따른 분류 기준을 가지고 있는 데이터베이스는 미비함

◦ 대사공학연구실 및 극지연구소가 보유하고 있는 유전체들 정보와 NCBI genome database 및 문헌을 통한 유전체들 정보를 확보하여 데이터베이스화할 예정임

○ 청정지역 및 오염지역의 유전체 내 유전자 패턴 분석

◦ 유전체-유전체간 패턴 분석은 단일 객체간 패턴 분석만이 있으며, 분석 조건에 따라 미비한 결과가 있으며, 특히 청정지역 및 오염지역 유전체 데이터를 바탕으로 한 유전자 패턴 분석은 실시된 적이 없음

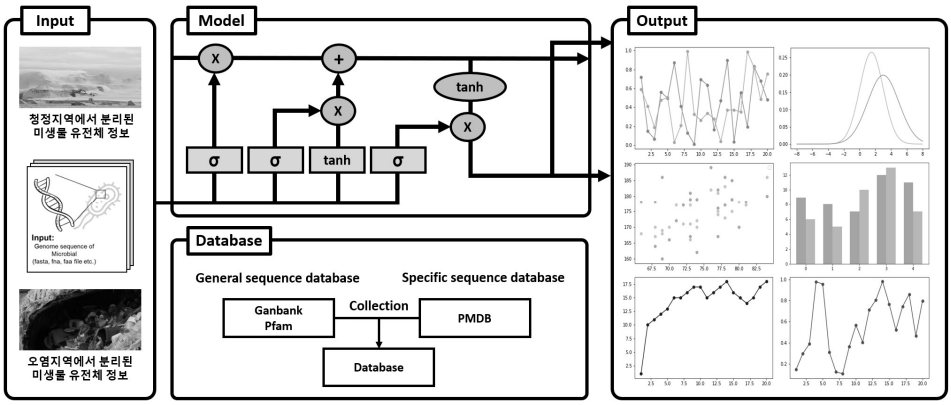
◦ 확보한 청정지역 및 오염지역의 유전체 데이터셋을 바탕으로, 각 유전체 내 Gene Set을 만들 예정이며, 알려진 다양한 패턴 인식 알고리즘을 바탕으로 유전체 내 유전자들의 발현 패턴을 분석할 예정임

○ 유전체 패턴 분석을 위한 딥러닝 모델 개발

◦ Input : 청정지역에서 분리된 미생물 유전체 정보와 오염지역에서 분리된 미생물 유전체 정보의 패턴을 비교하기 위하여 두 가지의 유전체 정보를 사용

◦ 딥러닝 모델: 유전자 간 차이를 분석하기 위하여 LSTM과 CNN으로 모델을 설계하고, 이에 대한 결과 값을 모델 검증을 통하여 연관성을 중심 분석

◦ 딥러닝 모델의 최종 결과물 : 패턴 인식의 결괏값이 정수화 되어있으며, 환경 요소에 대한 값을 해당 패턴 정보로 변환 후 반환한 뒤 그래프로 시각화하고자 함



<그림 12> 청정지역과 오염지역의 미생물 유전체 패턴 분석을 위한 시스템 구조도

○ 세부 프로젝트3의 최종 목표

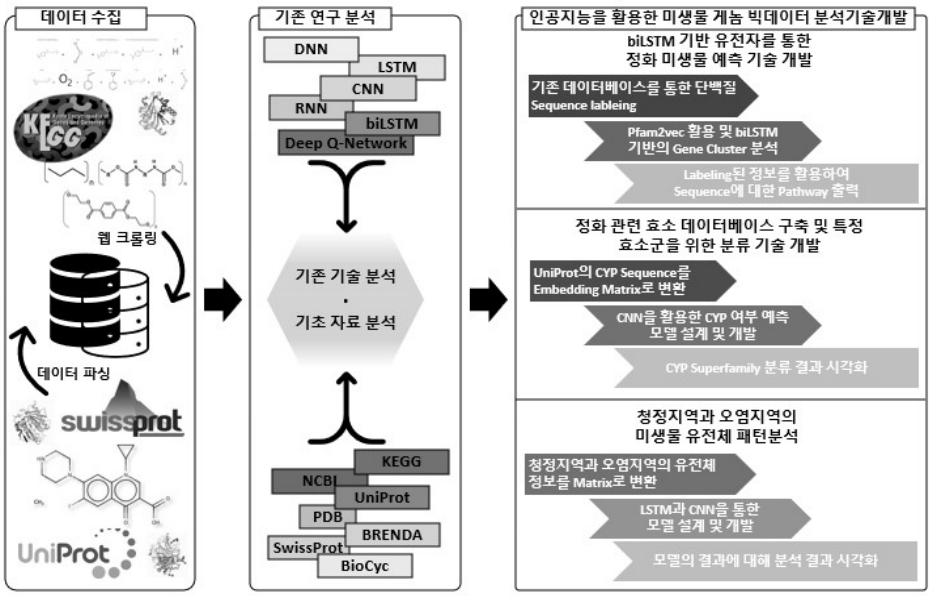
◦ 미생물 분리 지역 구분을 통한 청정, 오염지역 유전체 데이터베이스 구축

◦ 기계학습 분야와의 융합을 통해 정의된 규칙 이외의 새로운 패턴 결과 도출

◦ 각 지역에서 확보한 미생물 유전체들이 보유한 유전자들의 패턴 분석을 통해 환경오염에 연관된 유용 유전자들 확보

2) 추진전략 및 체계

가. 추진전략



<그림 13> 본 연구과제의 추진 전략

○ 과학적 연구 성과관리를 통한 연구개발 품질향상

◦ 파견인력은 연구 진행 상황을 연구 노트에 기록, 추후 연구 관련 문제 발생 시 연구 노트를 제시하여 문제를 해결

◦ 파견인력이 수료, 졸업 등의 사유로 연구중단 시 연구 노트를 활용하여 연속적으로 연구를 수행할 수 있도록 기반 마련

○ 안정적 연구수행을 위한 협업 의사소통 체계 확보

◦ 요구사항 검토 및 의사결정을 위한 협업체계를 마련하여, 연구개발 지연을 방지

◦ Slack 메신저 및 온라인 실시간 플랫폼을 활용하여 국내기관과의 의사소통 진행

○ 연구 협력 강화를 위한 주기적인 국내기관과의 화상회의

◦ 해외 파견인력은 국내와의 연구 협력을 위해 온라인 화상회의 시스템을 활용하여 수시로 국내기관과 회의를 진행

◦ 모든 연구원이 참여하는 화상회의를 국내기관과 월 4회 이상 개최

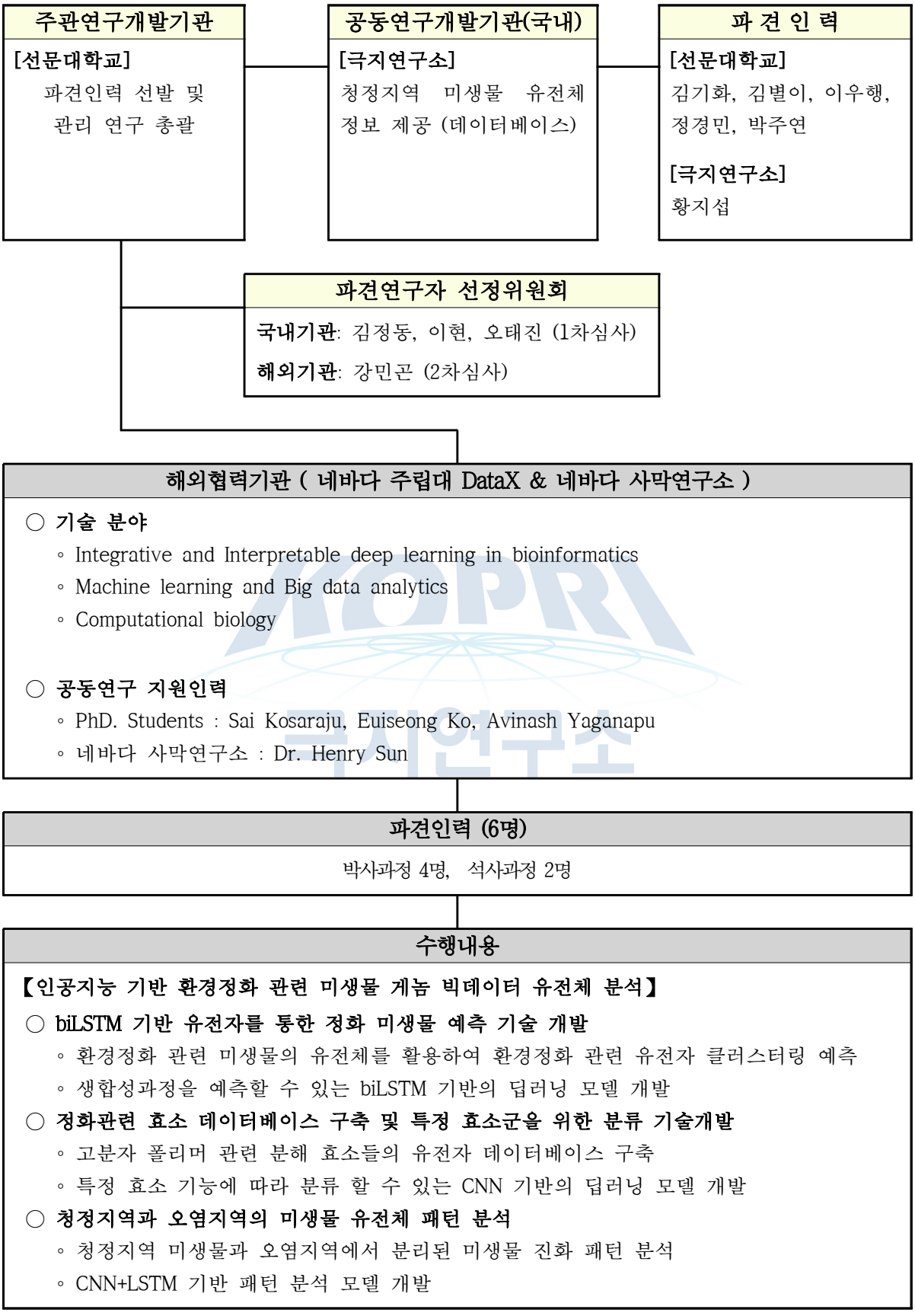
○ 성공적인 논문작성을 위한 Peer review 시스템 마련

◦ 논문 게재 준비 단계에서 파견 인력 사이에서 리뷰 및 토론

◦ 파견 학생들 간의 상호 첨삭 외에도 해외기관 연구원들에게도 첨삭 요청

◦ 연구내용 외에도 논문 구성 및 작성에 대한 비교분석을 통해 논문작성 실력을 향상할 수 있는 시스템을 마련

나. 추진체계


주관연구개발기관
[선문대학교]
파견인력 선발 및
관리 연구 총괄
공동연구개발기관(국내)
[극지연구소]
청정지역 미생물 유전체
정보 제공 (데이터베이스)
파 견 인 력
[선문대학교]
김기화, 김별이, 이우행,
정경민, 박주연
[극지연구소]
황지섭
파견연구자 선정위원회
국내기관: 김정동, 이현, 오태진 (1차심사)
해외기관: 강민곤 (2차심사)
해외협력기관 (네바다 주립대 DataX & 네바다 사막연구소)
○ 기술 분야
◦ Integrative and Interpretable deep learning in bioinformatics
◦ Machine learning and Big data analytics
◦ Computational biology
○ 공동연구 지원인력
◦ PhD. Students : Sai Kosaraju, Euiseong Ko, Avinash Yaganapu
◦ 네바다 사막연구소 : Dr. Henry Sun
파견인력 (6명)
박사과정 4명, 석사과정 2명
수행내용
【인공지능 기반 환경정화 관련 미생물 게놈 빅데이터 유전체 분석】
○ biLSTM 기반 유전자를 통한 정화 미생물 예측 기술 개발
◦ 환경정화 관련 미생물의 유전체를 활용하여 환경정화 관련 유전자 클러스터링 예측
◦ 생합성과정을 예측할 수 있는 biLSTM 기반의 딥러닝 모델 개발
○ 정화관련 효소 데이터베이스 구축 및 특정 효소군을 위한 분류 기술개발
◦ 고분자 폴리머 관련 분해 효소들의 유전자 데이터베이스 구축
◦ 특정 효소 기능에 따라 분류 할 수 있는 CNN 기반의 딥러닝 모델 개발
○ 청정지역과 오염지역의 미생물 유전체 패턴 분석
◦ 청정지역 미생물과 오염지역에서 분리된 미생물 진화 패턴 분석
◦ CNN+LSTM 기반 패턴 분석 모델 개발

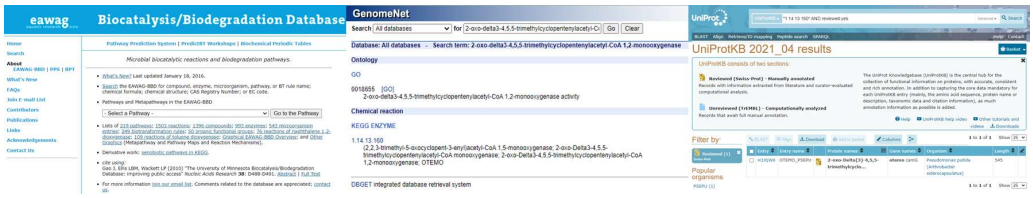
3) 연구 수행 내용

가. 주요 연구 내용

(1) 세부 프로젝트 1 : 환경정화 관련 미생물 유전체를 이용한 생합성 과정 예측

○ 새로운 Database 구축

◦ Biocatalysis/Biodegradation Database는 219개의 pathway 정보와 1503개의 reaction이 포함되어 있고, 이외에도 Genomenet, Uniprot 등이 단백질 정보나 서열 등을 데이터를 포함 (그림 14)



<그림 14> 본 연구에서 사용된 Database

◦ Database를 pathway, enzyme, compound로 class를 나누어 새로운 형식의 Database를 구축하였음 (표 5).

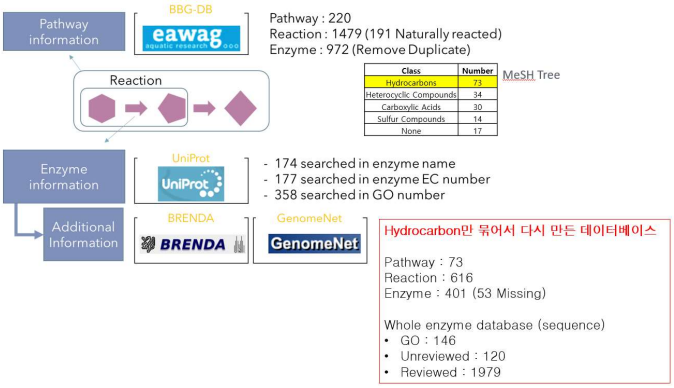
◦ Complete genome sequence로부터 pathway 정보를 기반으로 Biosynthetic gene cluster를 찾기 위해, EAWAG-BBD Database에 있는 정보(Pathway, Enzyme, Bacteria, Compound, Reaction, EC number)를 crawling하고 data를 frame에 맞춰 정형화함

◦ Database의 한계로 enzyme sequence는 Database에 있지 않아서 따로 수집하는 과정을 진행하였음.

◦ UniProt은 게놈 시퀀싱 프로젝트에서 파생 된 단백질 서열 및 기능 정보에 대한 자유롭게 액세스 할 수 있는 데이터베이스임. 효소번호/ 유전자 번호를 검색하여 해당하는 단백질 서열을 크롤링함. 확보한 enzyme sequence는 전체 351개, 추가로 약 1000여 개 이상을 다른 Database에서 수집함

◦ 확보된 enzyme sequence에 대한 이름이 길고 복잡하여 EC number로도 정리하는 과정을 진행함. Protein sequence를 Database화하기 위하여 GenomeNet, Uniprot Database를 활용하여 본 Database를 구축하였음

◦ GenomeNet 데이터베이스를 사용하여 효소 이름을 통해 각 효소에 대한 효소번호(EC number)와 유전자 번호(GO number)를 크롤링(Crawling)함. 구축한 데이터베이스에 대한 통계는 그림 15에 정리함



<그림 15> 구축한 데이터베이스 전체 자료 정리

○ Pipeline 구축

◦ 전체 진행하고자 하는 pipeline을 시각화하였음 (그림 16)

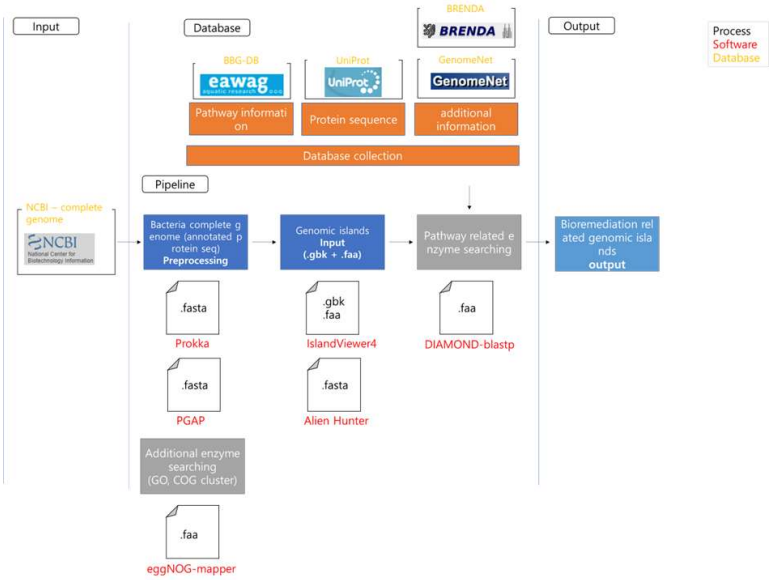
◦ 구축한 Database를 통하여 Bioremediation관련 미생물을 선별하기 위하여 기존의 tools를 사용하였음

◦ Validation 할 때 사용한 균주는 기존 Database에 활성이 있다고 알려진 균주 중 NCBI에 complete genome sequence가 등록되어있는 9종을 선정하였음

◦ 분석에 앞서, complete genome sequence를 annotation 하기위하여 prokka와 bakta 두 가지 tools를 활용하였음

◦ annotation 된 complete genome sequence와 Database에 있는 enzyme sequence 유사도를 확인하기 위하여 DIAMOND-blast를 이용하였음

◦ 최종 요약 및 진행하고 있는 Pipeline과 더불어 사용한 software, process 등을 정리하여 그림으로 나타내었음



<그림 16> 정화 Genomic island 예측 파이프라인

◦ Gene cluster mapping 결과와 논문에 보고된 gene cluster와 비교하였음 (표 5)

<표 5> Blast 결과를 통해 mapping한 3-Phenylpropionate 예상 gene cluster

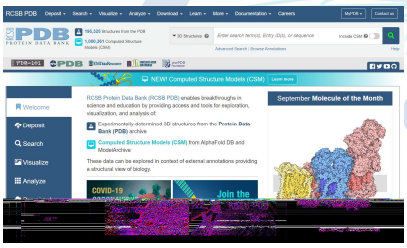
Query definition	Gene name	Query ID	E-value
BMIABBIL_00347 3-(3-hydroxy-phenyl)propionate/3-hydroxycinnamic acid hydroxylase	MhpA	Query_338	2.50E-18
BMIABBIL_00348 2,3-dihydroxyphenylpropionate/2,3-dihydroxicinnamic acid 1,2-dioxygenase	MhpB	Query_339	8.90E-183
BMIABBIL_00349 2-hydroxy-6-oxononadienedioate/2-hydroxy-6-oxononatrienedioate hydrolase	MhpC	Query_340	1.80E-89
BMIABBIL_00350 2-keto-4-pentenoate hydratase	MhpD	Query_341	4.80E-153
BMIABBIL_00351 Acetaldehyde dehydrogenase		Query_342	5.70E-137
BMIABBIL_00352 4-hydroxy-2-oxovalerate aldolase	MhpE	Query_343	9.20E-194

(2) 세부프로젝트2-1 : 정화 관련 효소 데이터베이스 구축 및 특정 효소군을 위한 분류 기술개발

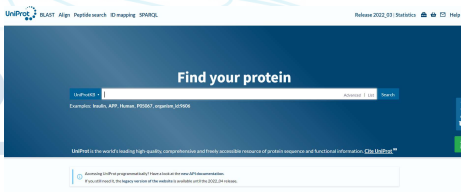
○ 데이터셋 구축

◦ 먼저, 유용하게 사용할 수 있는 데이터베이스가 드물어, 여러 개 데이터베이스에서 필요한 부분을 크롤링하고, 이들을 하나로 모아 데이터 셋으로 구성함.

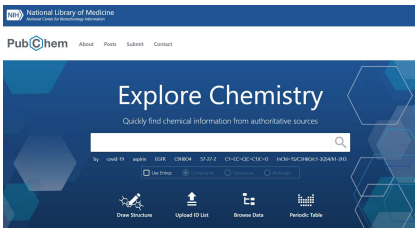
◦ Cytochrome P450 단백질에 대한 시퀀스 정보와 해당 단백질과 상호작용하는 기질에 대한 InChI, InChIKeym SMILES 등에 대한 정보가 필요함. 이를 얻기 위해 다양한 Protein, Compound 관련 데이터베이스를 조사, 분석 및 수집함



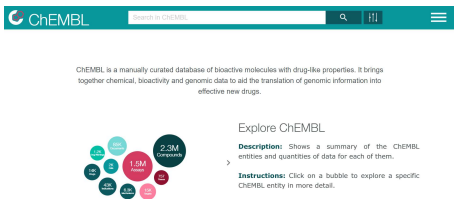
PDB



UniProt



PubChem



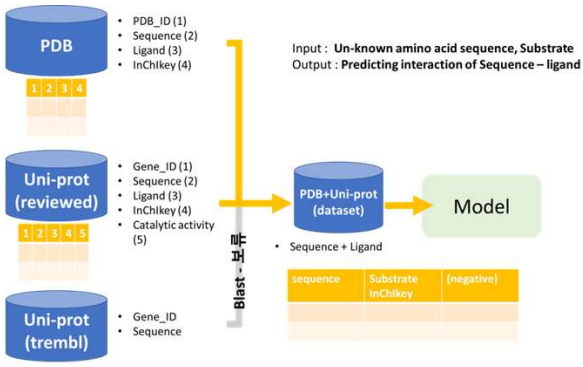
ChEMBL

<그림 17> CYP데이터 수집을 위해 사용된 데이터베이스

◦ Input으로 아미노산 시퀀스와 기질의 분자 구조가 들어가고 output으로 상호작용 결과가 나올 수 있도록 구성하고자 함.

◦ 아미노산 시퀀스와 분자 구조는 각각 알파벳의 나열, 그림 형태를 지니고 있으므로 이들을 descriptor을 이용하여 컴퓨터가 인식할 수 있는 형태로 만들어줌.

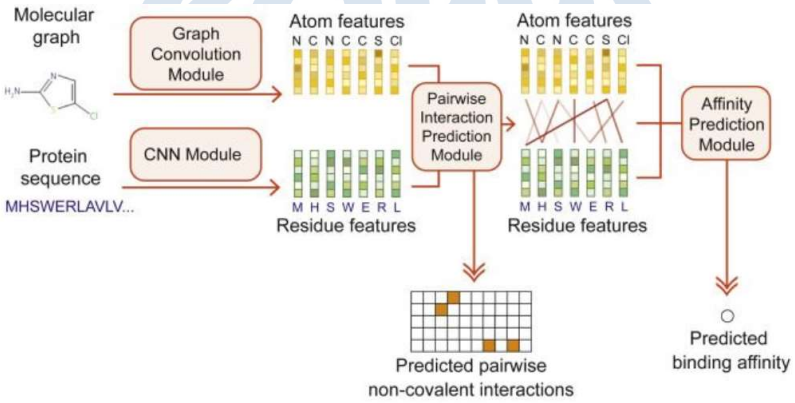
◦ 이때, 분자 구조는 RDKit을 사용하고자 하였으며, 이를 위해 분자 구조, InChI Key, SMILES, InChI의 정보를 담아 데이터 셋을 구성함.



<그림 18> 데이터 전처리 모식도

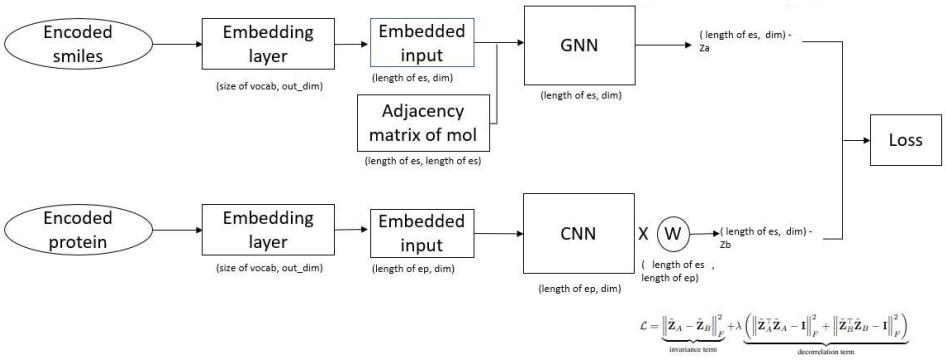
◯ 모델 설계

◦ 기존 CPI 예측 모델 중 BACPI는 단백질과 화합물 간의 결합도를 찾는 것은 단순한 이진 분류 문제가 아니라 연속 값이라고 말함. BACPI에서 제안하는 모델은 이진 분류 문제인 CPI 상호작용을 예측하고 Regression 문제인 결합 활동도 예측함.



<그림 19> BACPI Workfolw

◦ 현재 이용 가능한 대규모 비표지 화합물 및 단백질 데이터로부터 잠재된 특징을 탐색하기를 원하고(기존 방법은 라벨이 붙은 데이터로부터 특징의 단순하고 직접적인 표현을 사용하고 미지의 CPI를 추론하기 위해 이를 사용함) CPI 예측을 위한 강력한 딥러닝 기반의 특징 embedding을 사용하려고 함.



<그림 20> Compound-Protein Interaction 예측 모델 아이디어 시각화

◦ Embedding 된 데이터를 가지고 학습을 할 때 는 GNN(Graph Neural Network)을 활용함. GNN은 그래프에 직접 적용할 수 있는 신경망으로, 점 레벨, 선 레벨, 그래프 레벨에서의 예측 작업에 쓰임. GNN의 핵심은 점이 이웃과의 연결에 의해 정의된다는 것인데, 이를 염두에 두면 GNN이 compound 데이터를 다루는데 적합하다는 것을 알 수 있음. 따라서 이를 통한 compound 예측 모델을 구현할 계획임.

◦ 그리고 Protein을 예측 하는 방법으로는 CNN을 활용할 계획임. Protein 데이터 같은 경우에는 sequence로 이루어져 있기 때문에 graph데이터를 처리하는 방법처럼 처리할 수 없음. 이에 Sequence를 Image matrix처럼 사용한 전례들을 활용하여 sequence를 One-hot encoding을 통해 embedding 하여 처리하려고 함.

◦ 이렇게 각 embedding 방법과 처리 algorithm을 통해 추출된 feature map을 가지고 threshold 값을 조정하여 Compound-Protein Interaction 예측을 위한 모델을 개발 할 계획임.

(2) 세부프로젝트2-2 : 플라스틱 분해 가능효소 예측 기술 개발

◦ 세부 프로젝트 1에서 진행하고자 했던 주요 프로젝트는 플라스틱 생분해 과정에 대한 생합성과정 예측이나 현재 플라스틱 관련된 정보가 많이 없어 세부 프로젝트 2에서 플라스틱 분해 효소에 대한 예측으로 추가 세부 연구를 진행하였음

○ 데이터 수집

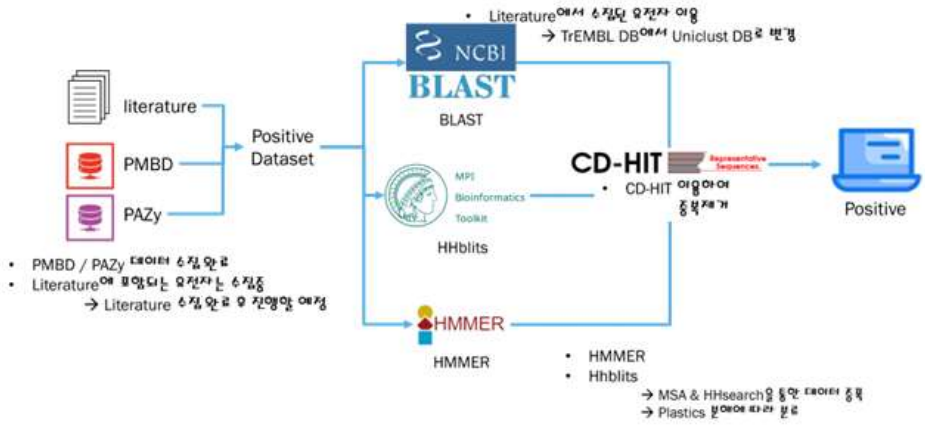
◦ 기존 PMBD (Plastics Microbial Biodegradation Database)와 PAZy (The Plastics-Active Enzyme Database) 데이터베이스의 한계로 논문을 검색하여 수동적으로 효소에 대한 정보를 추가 추출하였음.

◦ 또한, 리뷰 논문 27편과 일반 논문 119편에 따라 플라스틱 분해효소를 포함하고 있는 α/β-hydrolase 효소군의 유전자들을 선별하여 수집함. 결과적으로, weight loss 실험 포함하여 관련 실험을 통해 검증된 유전자는 총 147종의 관련 유전자를 확보함 (그림 21)

PBS	22
PET	45
PCL	23
PHA	42
PU	13
PBF	3
PA	8
PE	3
PES	2
제외	42

<그림 21> 플라스틱별 데이터 수집 정리

◦ 추가적인 데이터 수집을 위해, 수집된 147종의 유전자들을 이용하여 서열 기반의 pattern 추출을 위해 multi-sequence alignment 및 HMM pattern extract를 실시하였고, 기존 예측 데이터베이스인 TrEMBL에서 blast, HMMER 및 HHblits를 통해 유사성이 높은 유전자들을 추출함 (그림 22)



<그림 22> 데이터세트 구성에 대한 전반적인 데이터 수집 방법

◦ LED (The Lipase Engineering Database)에 수립되어 있는 α/β-hydrolase 효소군의 core dataset에서 positive dataset과 동일한 유전자들을 제외하고 CD-HIT 프로그램을 통해 non-redundant negative dataset을 구성하였음.

◦ 딥러닝 모델을 학습 · 구성하기 위해서, 분해가능한 플라스틱별로 정리 · 수집한 효소들을 이용하여 training set과 validation set으로 구성함. 또한, 수집한 데이터들은 위양성 문제를 해결하기 위해서 이미 비슷한 분야에서 동일한 접근방식으로 위양성 문제를 해결한 논문들을 참고하였으며 이를 해결 후 세트를 구성하고자 하였음.

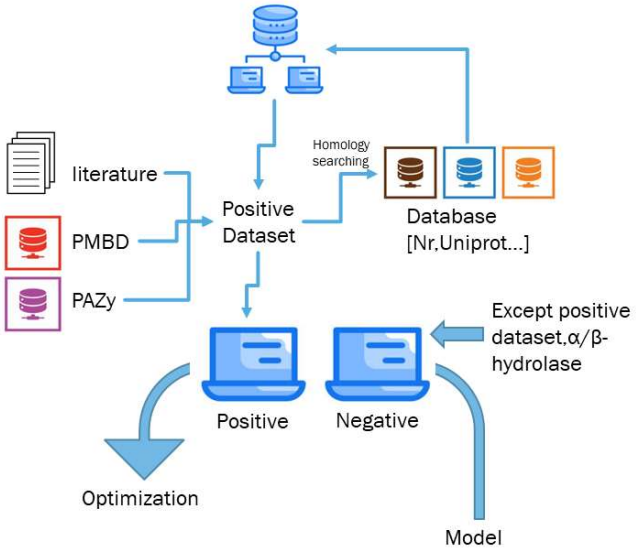
○ 데이터 전처리 연구

◦ 플라스틱 생분해 효소와 다르게 패턴이 알려진 데이터는 기존 One-Hot-Encoding을 통해 데이터 전처리를 진행함

◦ 하지만 플라스틱 생분해 효소의 경우 알려진 패턴이 적기 때문에 전처리 과정에서 더 많은 정보를 담아야하는 문제 발생

◦ 따라서 Jing, Xiaoyang가 발표한 논문, Xu, Yuting 가 발표한 논문, 두 개의 논문을 참고하여 amino acid 의 encoding 방식을 참고하여 19개의 encoding 방법을 구현하였음.

◦ 플라스틱 생분해 효소 데이터는 최근 연구되어 밝혀진 데이터, 패턴이 적었으며 기존 Protein Sequence Encoding 방법으로 많이 사용하는 One-Hot-Encoding을 사용하기에는 패턴 파악에 어려움이 있음. 따라서 프로젝트에서는 기존에 사용한 Protein Sequence Encoding의 19개 기법을 활용하여 딥러닝 모델 데이터 전처리를 진행하였음.



<그림 23> 데이터 전처리 및 학습에 대한 전반적인 내용

○ 플라스틱 생분해 효소를 위한 분류 모델 연구

◦ 플라스틱 생분해 효소 모델은 총 2단계로 구성하였음. 1단계 모델은 플라스틱 생분해 효소 여부에 따른 이진분류, 2단계 모델은 플라스틱 생분해 효소별 분류를 하는 다중 분류 모델로 사용자가 입력한 데이터를 분류함.

◦ 플라스틱 생분해 효소 분류 모델은 Protein sequence를 활용하여 분해 효소를 분류하는 모델로 Embedding Layer, Classification Model, 2가지의 구성으로 이루어져 있음.

◦ Classitifcation model의 경우 총 3가지 방법을 사용하였는데, RandomForest, BRNN의 경우 논문에서 사용한 hyperparameter를 동일하게 사용하여 구현 하였고, CNN의 경우 DeepEC와 동일한 구조를 가지고 사용하였으나, encoding 방법에 따라 Convolutional Layer 값을 다르게 사용하였음.

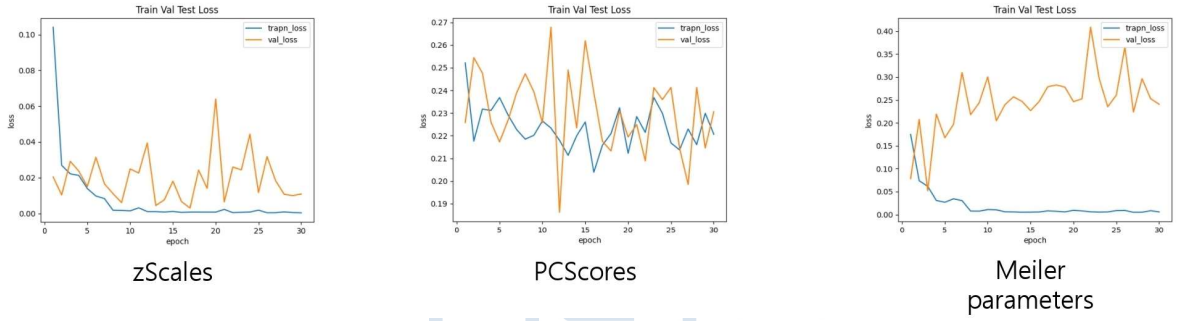
<표6> 플라스틱 생분해 효소 모델 CNN, BRNN 에 대한 107 Genes에 대한 결과

8 / 1e-4		zScales	VHSE	PCscores	sScales	One Hot Encoding	Five Bit Encoding	Six Bit Ecoding	Hydrohobicity matrix	Meiler Parameter	Atchley Factors	PAM 250	Blosum 62	Cristian	Tanaka Sheraga	Miyazawa Energies	Michletti Potentials	AESNN3	ANN4D
CNN	AUC	0.994084	0.994513	0.50	0.50	0.996227	0.948641	0.979336	0.910915	0.951816	0.956443	**0.996828**	0.995370	0.911601	0.691675	0.988082	0.900797	0.923776	0.985510
	w/o F1 Score	**0.873684**	0.811111	0.0	0.0	0.122807	0.851064	0.534247	0.471429	0.139130	0.851064	0.804469	0.824176	0.679012	0.0	0.544218	0.106195	0.853211	0.636943
	w/ F1 Score	**0.886422**	0.837007	0.333333	0.333333	0.402168	0.867283	0.646556	0.607242	0.411418	0.867283	0.831957	0.847047	0.741762	0.333333	0.652892	0.392780	0.851967	0.713305
BRNN	AUC	0.845066	0.747664	0.50	0.50	0.808197	0.779731	0.739347	0.773386	0.756409	0.770471	0.837863	0.873875	0.748521	0.828175	**0.859642**	0.806825	0.789505	0.804939
	w/o F1 Score	0.628205	**0.679012**	0.0	0.0	0.606452	0.569536	0.628205	0.513889	0.482270	0.619355	0.592105	0.592105	0.578947	0.601307	0.610390	0.632911	0.50	0.402958
	w/ F1 Score	0.707485	**0.741762**	0.333333	0.333333	0.691581	0.667515	0.707485	0.633705	0.613957	0.701619	0.683734	0.683734	0.673608	0.689745	0.695706	0.709126	0.623312	0.565438

◦ CNN, BRNN의 Grid Search 를 통하여 Batch size = 8, 16, 32, Learning Rate = 1e-4, 1e-5, 1e-6 으로 설정 하였고, 데이터는 학습:검증:평가 8:1:1로 구분 하여 학습하였음.

◦ 모델의 결과에서 알 수 있듯이 각 encoding 에 따라 결과가 다른 것을 확인 할 수 있으며, 생물학적인 정보를 많이 담고 있는 PAM250, Blosum62와 같은 encoding 이 결과가 좋은 것으로 드러남

◦ Learning Curve의 경우 학습되는 값들에 따라 값들이 달라지는 것을 볼 수 있으며, 결과를 통해 알아 봤을때 Learning rate는 1e-5, Batch size는 16이 적당한 것으로 확인되었음.

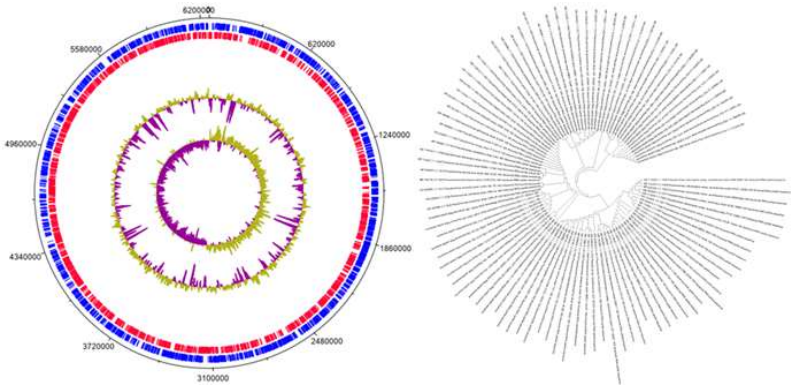


<그림 24> Learning rate 1e-5, Batch size 16의 Hyperparameter에 대한 CNN의 Learning Curve

(3) 세부프로젝트 3 : 청정지역과 오염지역의 미생물 유전체 패턴 분석

○ 미생물 유전체 분석

◦ 남극대륙 바톤반도 내 남극세종과학기지 인근에서 채집된 남극이끼 (Sanionia uncinata) 가 서식하는 지역의 rhizosphere에 해당하는 층의 토양으로부터 미생물을 분리 동정하였음. 16s rRNA 마커 유전자를 활용하여 종 동정을 진행하였고, 가장 가까운 유연관계에 Pseudomonas fluorescens CCM 2115 strain 이 존재하는 것으로 확인됨 (그림 25)



<그림 25> *P. fluorescens* Ant01의 유전체 분석 결과 및 16s 계통수

◦ 청정지역에서 분리한 미생물의 유전체 패턴을 확인하기 위해 전장 유전체 분석을 진행하고, 미국 국립생물공학정보센터 (National Center for Biotechnology Information, NCBI) 에 전장유전체 정보를 등록 중에 있음. Rapid prokaryotic genome annotation (Prokka)와 Gene ontology annotation in InterPro tool (Functional Domatin Prediction)을 사용하여 annotation을 진행함. 전장유전체 분석을 통해서 활용하여 청정지역에서의 유전체 패턴 분석을 위한 기초자료를 확보할 수 있었음. 해당 전장유전체 자료를 기준으로 청정지역과 오염지역에서 발견되는 미생물들의 유전체 패턴을 분석하기 위한 비교 분석 데이터를 수집 할 수 있었음.

○ 분리지역에 따른 항생제 내성유전자 패턴 분석

◦ 항생제 오염을 기준으로 청정지역과 오염지역에서 발견되는 bacteria 의 전장유전체 서열을 사용하여 항생제 내성유전자 패턴 분석을 수행하고자함.

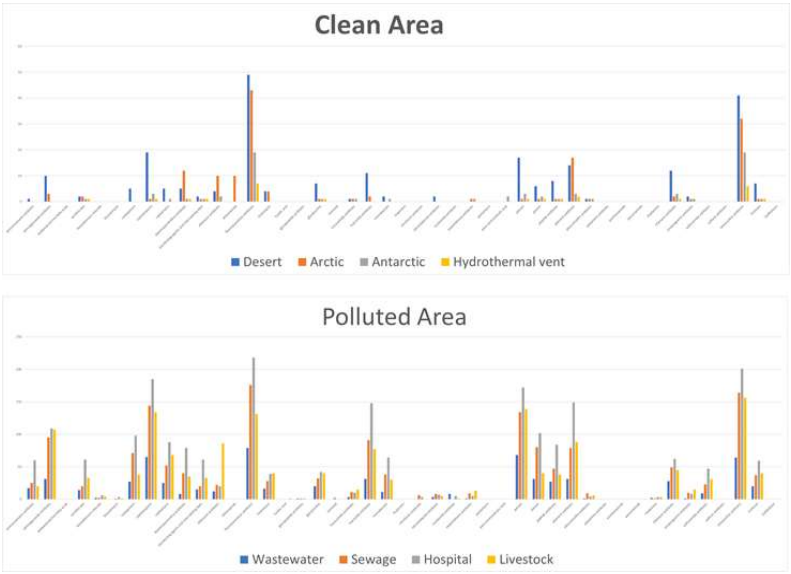
◦ NCBI-Biosample database에서 청정지역으로는 해당하는 환경으로는, 북극, 남극, 사막, 열수구를 선정하였고, 오염지역에 해당하는 환경으로는, 오염수, 하수처리장, 가축, 병원을 선정하였음.

◦ 각각의 환경에 따라 분리된 미생물들의 전장유전체 서열을 수집하였고, 그 중에서 complete genome, chromosome level의 유전체 정보만을 선별하여 항생제 내성유전자 검출 정확도를 높였음. 항생제 내성 유전자 분석은 CARD-RGI software를 사용하였고, 청정지역에서 120개, 오염지역에서 260개의 선별된 전장유전체를 사용하였음 (표 7)

<표 7> ARG 예측 결과

Clean Env. (Non-anthropogenic area)					Polluted Env. (Anthropogenic area)					
	Antarctic	Arctic	Desert	Hydro-thermal vent	Waste-water	Sewage	Livestock	Agri-culture	Aqua-culture	Hospital
Complete level/total assembly	33/180	48/526	30/863	9/795	15/903	38/576	78/552	4/977	51/201	75/998
Total	120/2,364				260/4,007					

◦ 청정지역과 오염지역에서 발견되는 항생제 내성유전자 예측 결과, 총 45종의 항생제 종류에 따른 검출 빈도수를 나타낸 그래프임. 오염된 지역으로 설정한 4개의 환경에서 상대적으로 높은 빈도수의 항생제 내성 유전자가 검출되는 것을 확인할 수 있었음.



<그림 26> 45종의 항생제에 따른 ARG 예측 결과

◦ 극한지 미생물의 전장유전체 서열을 사용하여 청정지역에서 발견되는 특이적인 유전자패턴을 확인하기 위한 유전체 분석 연구 과정을 다음과 같이 설계하였음. 전장유전체 서열을 활용하여 분석할 수 있는 Tool을 선정하여 분석 과정을 설계하였고, 이를 통해서 새로운 미생물의 전장유전체 서열을 활용한 병원성 인자 및 항생제 내성유전자 분석에 활용될 수 있을 것으로 예상됨.

나. 주요 국제 협력교류 연구 내용

나-1. 국제 협력 교류 연구 내용

○ 머신러닝 수업을 통한 실험 설계

◦ 데이터 수집, 데이터 타당성, 데이터 구체화 등 데이터에 따른 모델 설계 구체화

◦ 생물정보학 분야에서의 분석 및 개발을 위해 프로그래밍 언어 습득, 머신러닝과 딥러닝에 대한 이론적 이해에 도움

◦ 데이터 전처리에 따른 모델의 일반성 차이, 데이터 전처리의 중요성 및 딥러닝 모델의 세부적인 연구를 통해 모델의 중요성, 타당성 등 연구 이해도 증가

◦ 머신러닝에 대한 기초 연구를 통해 모델 구현 이해도 증가

◦ 파견 대학의 지도인력과의 개별 연구회의 뿐 아니라, 소속된 연구팀에 있는 박사과정 학생들과의 활발한 연구 교류를 통하여 적극적인 피드백을 받을 수 있었고, 머신러닝과 딥러닝을 활용한 연구에서 문제 해결 능력을 기를 수 있었음

○ 연구회의, 세미나, 컨퍼런스 참여를 통한 실험 구현 구체화

◦ 파견 대학의 지도인력과의 개별 연구회의 뿐 아니라, 소속된 연구팀에 있는 박사과정 학생들과의 활발한 연구 회의 및 교류를 통하여 적극적인 피드백을 주고 받을 수 있었고, 머신러닝과 딥러닝을 활용한 연구에서 문제 해결 능력을 기를 수 있었음

◦ 파견 대학교의 연구팀에서 진행하는 주간 연구미팅에 참석하여 머신러닝과 딥러닝을 활용한 최신 연구 동향을 파악할 수 있었고 생물정보학 분야에 접목시킬 수 있는 알고리즘 또는 연구 방향을 설정하는데 도움을 받을 수 있었음

◦ 파견 대학교에서 진행하는 생물학, 생물정보학과 관련된 세미나에서의 연사발표 그리고 컨퍼런스에 참여함. 세미나 연사발표를 통해 연구 방향을 설정하고 연구를 진행함에 있어 통합적인 시각을 기를 수 있는 기회가 되었음

◦ 컨퍼런스의 파견대학의 박사과정 대학원생과, 박사 후 연구원들의 포스터 발표를 통하여 생물정보학적인 접근이 가능한 다양한 분야에 대해 시야를 확대시킬 수 있는 기회가 되었음

다. 인력파견 연구 내용

다-1. 파견인력 선발 기준

◦ 총 선발 인원 6명 (박사 4명, 석사 2명)

◦ 연구계획서(60점), 연구실적(20점), 대학원성적 (백분율 점수 20점으로 환산)

◦ 어학능력 가산점: TOEIC Speacking L7 이상 보유자

다-2. 파견인력 선발 경과

◦ 2/1 : 파견인력 모집 공고

◦ 2/8 : 지원 서류 마감 (총 12명 지원)

◦ 2/9 : 선발위원회 구성 (연구책임자 외 국내연구진 3명, 해외 연구진 1명)

◦ 2/15 : 선발결과 통보 및 참여기간 협의 (총 6명 선발 완료, 6개월 6명)

다-3. 파견인력 선발 결과

구분	기 관 명	석사생	박사생	계
주관	선문대학교	2	3	5
공동	극지연구소	-	1	1
합계		2	4	6

다-4. 인력 파견 현황

순번	이름 (성별)	파견기관 파견기간 (YYMMDD~ YYMMDD)	과학기술인 등록번호 생년월일 (YYMMDD)	소속	학과	세부 전공명	석 박사 /학기	입학 시기	예상 졸업 시기	군필 여부
1	김기화 (여)	University of Nevada, Las Vegas 2021.08.26.~2022.04.23.	11517962 931020	선문대 학교	생명 공학과	바이오빅데이터융합전공	박사 /8학기	'17. 9.	'22. 2.	여
2	김별이 (여)	University of Nevada, Las Vegas 2021.08.26.~2022.04.23.	11706447 940823	선문대 학교	생명 공학과	바이오빅데이터융합전공	박사 /4학기	'19. 9.	'23. 8.	여
3	이우행 (남)	University of Nevada, Las Vegas 2021.10.01.~2022.06.22.	11830278 930310	선문대 학교	생명 공학과	바이오빅데이터융합전공	박사 /3학기	'20. 3.	'24. 2.	군필
4	정경민 (남)	University of Nevada, Las Vegas 2021.11.01.~2022.06.22.	12697382 950131	선문대 학교	컴퓨터융합전자공학과	바이오빅데이터융합전공	석사 /1학기	'21. 3.	'23. 2.	군필
5	박주연 (여)	University of Nevada, Las Vegas 2021.08.26.~2022.02.24.	12462089 971208	선문대 학교	컴퓨터융합전자공학과	바이오빅데이터융합전공	석사 /1학기	'21. 3.	'23. 2.	여
6	황지섭 (남)	University of Nevada, Las Vegas 2021.10.01~2022.04.27	11811633 960729	극지연 구소	극지 과학	생화학	석박사 /5학기	'19. 3	'24. 3	미필

다-5. 인력 파견 추진 일정

순번	파견인력	2021년								2022년								비고
		5월	6월	7월	8월	9월	10월	11월	12월	1월	2월	3월	4월	5월	6월	7월	8월	
1	김기화				■	■	■	■	■	■	■	■	■					
2	김별이				■	■	■	■	■	■	■	■	■					
3	이우행						■	■	■	■	■	■	■	■	■			
4	정경민							■	■	■	■	■	■	■	■			
5	박주연				■	■	■	■	■	■	■							
6	황지섭						■	■	■	■	■	■	■					
주요 Milestone 완성점에서의 수행결과																		

다-6. 인력양성 주요 내용

1) 김기화 파견연구 개요

<table>
<tr><td rowspan="5">파견개요</td><td>이름</td><td>김기화</td><td>대학</td><td>선문대학교</td></tr>
<tr><td>학과</td><td>생명공학과</td><td>세부전공</td><td>바이오빅데이터융합</td></tr>
<tr><td>파견국가
(도시)</td><td>U.S.A (Nevada)</td><td>파견기관명</td><td>University of Nevada, Las Vegas</td></tr>
<tr><td>해외기관
지도인력</td><td>Mingon Kang. Ph. D., Assistant Professor of Computer Science</td><td>총연구기간
(파견기간)</td><td>210501~220831
(210826~220430)</td></tr>
<tr><td>참여
프로젝트명</td><td colspan="3">환경정화 관련 미생물 유전체 종합 분석</td></tr>
<tr><td rowspan="4">주요내용</td><td>연구주제</td><td colspan="3">정화관련 효소군 분류</td></tr>
<tr><td>참여
프로젝트내
수행역할</td><td colspan="3">환경정화 관련 효소 중 Cytochrome P450에 대한 데이터 수집 및 생명공학 지식 전달</td></tr>
<tr><td>주요 수행
내용</td><td colspan="3">• Cytochrome P450 효소를 주제로 알고리즘을 만들기에 앞서, 이 효소의 역할과 기초적인 생명공학 지식을 전달함.
• 유전체를 뜻하는 DNA, 효소의 구성단위인 아미노산, 단백질의 구조, 그리고 기질의 정의 등등 주제와 연관되는 모든 생명공학 지식들을 나누고, 질의 응답하여 더 명확한 주제를 만들고자 함.
• Data set을 만들기 위한 데이터베이스를 검색.
• PDB, UniProt, ChEMBL 등의 데이터베이스를 검색하고 이 중 필요한 부분만 크롤링 할 수 있도록 토의.
• 기질을 인배딩하기 위한 방법으로 ‘RDKit’을 선정하였으나, 오류가 생겨 이를 해결하기 위한 방법을 모색.
• InChI Key로 데이터를 추합하고 SMILE로 변환하였으나, 오류가 생기게 되어 이를 InChI로 바꾸기로 함.
• Negative data를 만들기 위한 방안 모색.
• 특히, 박테리아 유래 CYP에 대한 방안으로 ‘Journal of biological chemistry’ 저널을 선택하였으며, 이 저널에서 출판된 논문을 보고 데이터를 모으기로 함.</td></tr>
<tr><td>기대효과</td><td colspan="3">• 현재, 박테리아 유래 CYP를 다루는 알고리즘은 드물게 연구되었기 때문에, 미생물 CYP를 아우르는 최초의 알고리즘을 만들 수 있음.
• 기질과 단백질의 상호작용은 최근까지도 계속 연구되고 있는 핫한 주제이므로, 앞으로의 개발에 도움을 줄 수 있음.
• 인간부터 미생물까지 모은 데이터셋을 이용해 다른 알고리즘 개발에 응용해 볼 수 있음.</td></tr>
</table>

2) 김별이 파견연구 개요

<table>
<tr><td rowspan="5">파견개요</td><td>이름</td><td>김별이</td><td>대학</td><td>선문대학교</td></tr>
<tr><td>학과</td><td>생명공학과</td><td>세부전공</td><td>바이오빅데이터융합전공</td></tr>
<tr><td>파견국가
(도시)</td><td>U.S.A (Nevada)</td><td>파견기관명</td><td>University of Nevada, Las Vegas</td></tr>
<tr><td>해외기관
지도인력</td><td>Mingon Kang. Ph. D.,
Assistant Professor of
Computer Science</td><td>총연구기간
(파견기간)</td><td>210501~220831
(210826~220422)</td></tr>
<tr><td>참여
프로젝트명</td><td colspan="3">환경정화 관련 미생물 유전체 종합 분석</td></tr>
<tr><td rowspan="4">주요내용</td><td>연구주제</td><td colspan="3">환경정화 미생물 예측</td></tr>
<tr><td>참여
프로젝트내
수행역할</td><td colspan="3">프로젝트 디자인 및 수행</td></tr>
<tr><td>주요 수행
내용</td><td colspan="3">◦ 정화관련 미생물 종의 유전체의 분석을 위해 정화관련 생합성에 관련된 데이터베이스와 Uniprot을 합쳐 새로운 데이터베이스를 구축하였고, 현재 Gene cluster와 생합성과정을 합쳐서 분석 할 수 있도록 데이터베이스의 포맷을 설정함
◦ 정화 관련 Genomic island 예측을 하기 위한 Pipeline을 구축하였음. 예측에 최적화 하기 위하여 Annotation tools을 비교하였고, 최종적으로 Prokka와 Bakta를 사용함
◦ 정화미생물로 알려진 9종을 이용하여 Pipeline을 통한 분석을 진행하고 결과를 확인함으로써 정화관련 Genomic island 예측을 진행함
◦ Genomic island 분석을 통해 나온 결과 분석을 위해 구축한 데이터베이스를 이용하여 Diamond-blast 분석을 진행함
◦ 최총적으로 기존의 데이터베이스를 활용하여 본 연구에 적합한 데이터베이스를 구축하였고, 기존의 정보들을 최대한 활용하여 분석에 적합한 형태로 데이터를 수정하며 미생물정보학 관련 프로그램 결과를 분석하였음. 본 결과를 이용하여 시각화하여 생합성 과정 및 Genomic islands 분석을 용이하게 하였음</td></tr>
<tr><td>기대효과</td><td colspan="3">◦ 현재 박사주제로 미생물, 식물 유전체 분석을 주로 하고 있는데, 이번 과제를 통해 알고리즘이나 데이터 처리 등에 대해 배움으로써 다양한 활용이 가능해짐
◦ 생물학적 문제를 다각적으로 바라볼 수 있는 시각을 겸비하였고, 프로그래밍 언어를 사용하여 시각화 및 데이터처리를 더욱더 용이하게 할 수 있게되었음 이는 차후 미생물 유전체 빅데이터를 이용한 분석도 가능해질 것으로 사료됨</td></tr>
</table>

3) 이우행 파견연구 개요

<table>
<tr><td rowspan="5">파견개요</td><td>이름</td><td>이우행</td><td>대학</td><td>일반대학원</td></tr>
<tr><td>학과</td><td>생명공학과</td><td>세부전공</td><td>바이오빅데이터융합</td></tr>
<tr><td>파견국가
(도시)</td><td>U.S.A (Nevada)</td><td>파견기관명</td><td>University of Nevada, Las Vegas</td></tr>
<tr><td>해외기관
지도인력</td><td>Mingon Kang. Ph. D., Assistant Professor of Computer Science</td><td>총연구기간
(파견기간)</td><td>210501~220831
(211223~220622)</td></tr>
<tr><td>참여
프로젝트명</td><td colspan="3">환경정화 관련 미생물 유전체 종합 분석</td></tr>
<tr><td rowspan="4">주요내용</td><td>연구주제</td><td colspan="3">정화관련 효소군 분류</td></tr>
<tr><td>참여
프로젝트내
수행역할</td><td colspan="3">플라스틱 분해 가능 효소 데이터 수집 및 데이터세트 구축</td></tr>
<tr><td>주요 수행
내용</td><td colspan="3">◦ 플라스틱 분해 가능 미생물 및 효소에 대한 정보 수집, 분류와 더불어 논문 수집을 통해 플라스틱 분해 가능에 대한 메커니즘 정리 및 관련 데이터베이스를 정리함
◦ 플라스틱 분해 가능 효소에 대해 분해가능한 플라스틱별로 분류하였으며 이를 이용하여 딥러닝 기반 분석을 위한 데이터세트를 구성함
- 플라스틱 분해 가능 효소의 검증 연구에서 사용된 기질들이 제각각 달라 유사한 구조의 플라스틱들을 분류함
- 플라스틱 분해 가능 효소와 증폭된 유전자들간의 유사성, 특이 아미노산 서열들의 분석을 실시하였으며 이를 바탕으로 위양성 문제를 해결하고자 하였음
- 위양성 문제를 해결하기 위해, 비슷한 연구에 접근 방법들을 확인하였으며, 이를 바탕으로 training 및 validation set에 들어갈 정보들을 분류 · 선별함
◦ 부족한 데이터세트를 위해 플라스틱 분해 가능 효소와 유사한 유전자들의 확보 및 정리, 위양성 문제 해결을 시도함</td></tr>
<tr><td>기대효과</td><td colspan="3">◦ 미생물 (박테리아, 곰팡이, 조류 등) 내 플라스틱 분해 가능 효소 및 특이 유전자들의 발굴에 용이하며, 다양한 연구에 도움을 줄 것으로 예상됨
◦ 외부 탄소원을 이용하여 유용물질을 생성하는 화이트팩토리 연구의 일종으로, 플라스틱 폐기물을 이용하여 중요 유용물질을 생성할 수 있음</td></tr>
</table>

4) 정경민 파견연구 개요

<table>
<tr><td rowspan="5">파견개요</td><td>이름</td><td>정경민</td><td>대학</td><td>선문대학교</td></tr>
<tr><td>학과</td><td>컴퓨터융합전자공학과</td><td>세부전공</td><td>바이오빅데이터융합전공</td></tr>
<tr><td>파견국가
(도시)</td><td>U.S.A (Nevada)</td><td>파견기관명</td><td>University of Nevada, Las Vegas</td></tr>
<tr><td>해외기관
지도인력</td><td>Mingon Kang. Ph. D.,
Assistant Professor of
Computer Science</td><td>총연구기간
(파견기간)</td><td>210501~220831
(211223~220622)</td></tr>
<tr><td>참여
프로젝트명</td><td colspan="3">환경정화 관련 미생물 유전체 종합 분석</td></tr>
<tr><td rowspan="4">주요내용</td><td>연구주제</td><td colspan="3">정화관련 효소군 분류</td></tr>
<tr><td>참여
프로젝트내
수행역할</td><td colspan="3">플라스틱 생분해 효소 데이터베이스 구축 및 분류 모델 개발</td></tr>
<tr><td>주요 수행
내용</td><td colspan="3">- 기존 플라스틱의 자연 분해는 최소 20년이 걸림. 불에 태우는 경우 환경호르몬이 배출돼 환경을 위협하는 요소임. 생분해성 플라스틱은 특정 효소와 만나 저절로 썩어서 사라지는 플라스틱으로 가수분해가 가능한 구조를 가지고 있음.
- 현재 알려진 플라스틱 생분해 효소는 PA, PBF, PBS, PCL, PE, PES, PET, PHA, PU 등 다양하게 존재함. 하지만 플라스틱 생분해 효소에 대한 밝혀진 패턴은 적음. 또한 플라스틱 생분해 효소는 별개의 패턴을 가지고있음.
- 새로운 패턴 확보 및 분류를 위해 딥러닝 기술과 융합하여 해결하고자함
- Protein Sequence 데이터와 딥러닝 기술중 하나인 CNN을 통해 분류 및, 예측하는 모델은 현재 많이 나와있음, 또한 새로운 패턴을 찾기 위해서는 인코딩 또한 다양한 방법을 통해 어느 인코딩 방법이 좋은지 판별 하고자 함</td></tr>
<tr><td>기대효과</td><td colspan="3">- 현재 연구를 수행 중인 플라스틱 생분해 효소 데이터 베이스 구축을 통해 컴퓨터 분야와의 융합 연구를 통해 특정 환경 분야의 전문가를 양성 할 수 있음
- 환경 분야 문제에 대한 융합 연구를 통해 컴퓨터 분야의 다양한 융합연구 가능성에 대한 시야를 키울 수 있으며, 더 나아가 실제 환경 문제를 해결 할 수 있는 융합인재를 양성 할 수 있음.
- 데이터베이스 구축 및 딥러닝 분류 모델을 개발함으로써 새롭게 등장하는 플라스틱 분해 효소에 대한 새로운 패턴 발견을 할 수 있으며, 또한 새로운 분야에 대한 기술을 정립 할 수 있음</td></tr>
</table>

5) 박주연 파견연구 개요

파견개요	이름	박주연	대학	선문대학교
	학과	컴퓨터융합전자공학과	세부전공	바이오빅데이터융합전공
	파견국가(도시)	U.S.A (Nevada)	파견기관명	University of Nevada, Las Vegas
	해외기관 지도인력	Mingon Kang. Ph. D., Assistant Professor of Computer Science	총연구기간(파견기간)	210501~220831 (210826~220224)
	참여 프로젝트명	환경정화 관련 미생물 유전체 종합 분석		
주요내용	연구주제	정화관련 효소군 분류		
	참여 프로젝트내 수행역할	데이터 수집 및 전처리, 딥러닝을 통한 정화관련 효소와 기질 간의 상호작용 예측 연구		
	주요 수행 내용	◦ Genbank, Pfam, KEGG, PMBD 등의 데이터베이스를 통해 미생물 정화 유전자 예측을 위한 데이터셋을 구축함. 데이터셋은 fasta 형식의 Sequence를 사용하며 추후 시각화를 위해 EC number를 Labeling 함 ◦ BiLSTM 기반의 Gene Cluster 분석을 위한 모델 설계 및 구현을 진행하고 그 결과를 시각화하는 방법으로 KEGG의 Pathway 데이터베이스를 활용함. ◦ 청정지역에서 분리된 미생물의 유전체와 오염지역에서 분리된 미생물의 유전체를 비교하기 위해 딥러닝 기반의 패턴분석 기법을 연구함. 딥러닝 알고리즘의 대부분이 패턴분석을 수행할 수 있다고 판단하여 목표를 수행하기 위한 최적의 알고리즘을 찾고 구현함. ◦ 정화 관련 효소인 Cytochrome P450 효소 데이터 수집 및 해당 효소와 결합하는 기질 데이터 수집 ◦ 수집한 데이터 정형화 및 전처리 ◦ Cytochrome P450 효소와 기질 간의 상호작용 예측을 위한 관련 연구 동향 파악과 예측 모델 개발을 위한 연구 진행		
	기대효과	◦ 연구를 수행하면서 바이오 빅데이터를 활용한 분석과 예측 모델 개발을 통한 생물학적 지식을 쌓고 융합연구를 통해 각 분야에 대한 다양한 시각을 얻을 수 있음 ◦ 다양한 딥러닝 알고리즘을 구현하고 연구하면서 각 알고리즘의 적합한 사용 및 개발 방법을 익힐 수 있음		

6) 황지섭 파견연구 개요

<table>
<tr><td rowspan="5">파견개요</td><td>이름</td><td>황지섭</td><td>대학</td><td>극지연구소(UST)</td></tr>
<tr><td>학과</td><td>극지과학</td><td>세부전공</td><td>생화학, 구조생물학</td></tr>
<tr><td>파견국가
(도시)</td><td>U.S.A (Nevada)</td><td>파견기관명</td><td>University of Nevada, Las Vegas</td></tr>
<tr><td>해외기관
지도인력</td><td>Mingon Kang. Ph. D., Assistant Professor of Computer Science</td><td>총연구기간
(파견기간)</td><td>210501~220831
(211213~220429)</td></tr>
<tr><td>참여
프로젝트명</td><td colspan="3">환경정화 관련 미생물 유전체 종합 분석</td></tr>
<tr><td rowspan="4">주요내용</td><td>연구주제</td><td colspan="3">환경관련 미생물 패턴 분석</td></tr>
<tr><td>참여
프로젝트내
수행역할</td><td colspan="3">극한지 미생물 유전체 분석, 항생제 오염에 대한 청정지역과 오염지역의 미생물 유전체 패턴분석</td></tr>
<tr><td>주요 수행
내용</td><td colspan="3">◦ 남극대륙 바톤반도 내 남극세종과학기지 인근에서 채집된 남극이끼(Sanionia uncinata) 가 서식하는 지역의 rhizosphere에 해당하는 층의 토양으로부터 미생물을 분리 동정 및 유전체 분석, 미국 국립생물공학정보센터 (National Center for Biotechnology Information, NCBI) 에 전장유전체 정보를 등록 (Genebank accession number: CP104408)
◦ ‘Reactome DB’ 내에서 blast를 통해서 발견되며, 전장유전체 염기서열을 사용하여 총 4개의 Reactome이 확인함
◦ KEGG Pathway 분석 결과 arthrofactin-type cyclic lipopeptide synthetase B 중심의 합성경로가 확인되었고, P. fluorescens Ant01의 전장유전체 상에 존재하는 antibiotics and secondary metabolite biosynthesis gene cluster를 확인하기 위해서 AntiSMASH v,6.0 software를 사용하여 lipopeptide 계열의 항균물질을 생산할 것으로 예측함
◦ 청정지역과 오염지역 환경에 따라 분리된 미생물들의 전장유전체 서열을 수집하였고, 그 중에서 complete genome, chromosome level의 유전체 정보만을 선별하여 항생제 내성유전자 검출 정확도를 높였음. 항생제 내성 유전자 분석은 CARD-RGI software를 사용하였고, 청정지역에서 120개, 오염지역에서 260개의 선별된 전장유전체를 사용하여 분석을 진행하였음</td></tr>
<tr><td>기대효과</td><td colspan="3">◦ 머신러닝과 딥러닝 기반의 유전체 분석 프로그램을 사용과 결과 데이터를 해석할 수 있는 시각을 키울 수 있음.
◦ 다양한 유형의 생물학 분야 빅데이터를 다룰 수 있고 시각화 할 수 있는 역량을 키울 수 있음.</td></tr>
</table>

다-7. 파견인력 관리

1) 출국 지원

○ 비자 및 체재비

◦ University of Nevada, Las Vegas를 통한 허가서(DS-2019) 발급 지원 및 파견인력 J-1 교환방문 비자발급 및 프로그램 등록료 (1학기 $1,000 지원)

◦ 파견인력 1인당 150만원의 왕복 항공료 지원

◦ 만일의 사고에 대비한 유학생 보험료 지원

◦ 현지 생활 지원을 위한 체재비(1인당 월 190만원) 보장

○ 학적관리

◦ 파견 학생은 기존 소속기관에서의 학적 유지 후 파견

2) 파견 기간 중 현지 지도 등 관리

○ 과제수행 관리 및 연구지도

◦ 해외기관 지도교수가 세부 프로젝트 관리, 세부 프로젝트 담당 인력 지원

◦ 현지 연구인력과 공동연구 프로그램 구성

◦ 주 1회 미팅을 통한 각 세부 프로젝트별 연구 주간 보고 및 연구지도

◦ 월 1회 과제 진도 보고를 위한 통합 세미나

○ 과제수행 관리 및 연구지도

◦ 연구 코어 타임 외 탄력적 연구 지향으로 파견인력 연구 환경관리

◦ 주기적인 현지 지도교수와의 개별 면담을 통한 피드백

◦ 세부 프로젝트 간 협업을 통한 현지 부적응 상호개선

3) 비상시 대처 노력 (코로나19 대비 노력)

가. 중도 포기, 미복귀 등의 대처 방안

◦ 파견인력이 정당한 사유 외 과제 중도 포기 시, 과제의 연속적 수행을 위한 파견인력 충원 계획 수립 후 승인, 파견인력 변경으로 인한 항공료 페널티 부과

◦ 파견인력이 정당한 사유 외 미복귀할 경우, 지원금액 환수, 항공료 페널티 부과

나. 부적응, 건강악화 등 대처 방안

◦ 파견인력이 부적응으로 과제수행이 어려울 시, 파견 기간 중 1회 2주간 국내 복귀 기회 제공. 단, 항공료는 자부담

◦ 파견인력이 건강상의 문제로 불가피한 복귀 시에는 복귀하되, 국내 과제 참여 인력 등을 활용하여 파견인력을 충원

다. 기타 상황 발생 시 대처 방안

◦ 파견인력이 예상할 수 없는 만일의 사고에 대비하여, 전원 유학생 보험에 가입하도록 하며, J-1 및 University of Nevada, Las Vegas에서 권장하는 금액으로 가입 유도

4) 취업 등 진로 모니터링

◦ 연구관련 취업이나 진로 등에 관련한 사항 조사 및 관련 업체 연결 시도

5) 과제 관리 전담인력 운영

◦ 각 세부 프로젝트에서는 파견인력과 국내 과제 참여 인력 간 원활한 과제수행을 위해 주기적인 예산 집행 및 연구실적 등을 교류하며 과제 관리

3. 연구개발과제의 수행 결과 및 목표 달성 정도

1) 연구수행 결과

(1) 정성적 연구개발성과

세부프로젝트 1 : 환경정화 미생물 예측

가. 환경오염 관련 데이터베이스 구축 (물질, 반응, 효소, 미생물, 시퀀스 정보 포함)

나. Hydrocarbon degrading bacteria 예측 가능한 Pipeline구축

다. Hydrocarbon degrading 관련 Genomic islands 탐색

라. 알려진 미생물과 본 연구실의 미생물 유전체 정보를 이용한 Case study 진행

세부프로젝트 2-1 : 정화관련 효소군 분류

가. Cytochrome P450 관련 데이터베이스 구축(시퀀스, InChI, InChIKey, SMILES 정보 포함)

나. Compound-Protein Interaction 관련 모델 조사 및 분석 진행

다. Molecular descriptor를 사용한 분자 구조 문자열 변환 및 데이터셋화 실시

세부프로젝트 2-2 : 정화관련 효소군 분류

가. 플라스틱 생분해 효소 관련 데이터 수집 및 데이터베이스 구축

나. 생분해별 분류 및 데이터 증폭 완료

다. Protein Sequence기반 19가지 Encoding 기법 구현

라. 플라스틱 생분해 효소 분류 모델 3가지 구현

세부프로젝트 3 : 환경관련 미생물 패턴 분석

가. 남극 토양 미생물의 전장유전체 해독 및 생명 자원 등록 (NCBI)

나. 미생물 유전체 내 항생물질 및 병원성 인자 관련 pathway 분석

다. 항생제 내성 및 병원성 인자의 정량적인 분석 및 유전체 패턴 분석

라. 프로그램을 사용한 예측결과를 뒷받침하는 표현형적인 검증 실험 진행

(2) 정량적 연구개발성과

< 정량적 연구개발성과표 >

(단위 : 건, 천원)

성과지표명 \ 연도			1단계 (YYYY~YYYY)	n단계 (YYYY~YYYY)	계	가중치 (%)
전담기관 등록·기탁 지표[1]	수혜인력	목표(단계별)				
		실적(누적)				
	배출인력	목표(단계별)	6			
		실적(누적)	6			
	파견인력 만족도	목표(단계별)	6			
		실적(누적)	6			
	파견인력 역량향상도	목표(단계별)	6			
		실적(누적)	6			
	파견결과 보고서	목표(단계별)	6			
		실적(누적)	6			
	성과공모전	목표(단계별)				
		실적(누적)				
연구개발과제 특성 반영 지표[2]		목표(단계별)				
		실적(누적)				
		목표(단계별)				
		실적(누적)				
계						

* 1」전담기관 등록·기탁 지표: 논문[에스시아이 Expanded(SCIE), 비SCIE, 평균Impact Factor(IF)], 특허, 보고서원문, 연구시설·장비, 기술요약정보, 저작권(소프트웨어, 서적 등), 생명자원(생명정보, 생물자원), 표준화(국내, 국제), 화합물, 신품종 등을 말하며, 논문, 학술발표, 특허의 경우 목표 대비 실적은 기재하지 않아도 됩니다.
* 2」연구개발과제 특성 반영 지표: 기술실시(이전), 기술료, 사업화(투자실적, 제품화, 매출액, 수출액, 고용창출, 고용효과, 투자유치), 비용 절감, 기술(제품)인증, 시험제품 제작 및 인증, 신기술지정, 무역수지개선, 경제적 파급효과, 산업지원(기술지도), 교육지도, 인력양성(전문 연구인력, 산업연구인력, 졸업자수, 취업, 연수프로그램 등), 법령 반영, 정책활용, 설계 기준 반영, 타 연구개발사업에의 활용, 기술무역, 홍보(전시), 국제화 협력, 포상 및 수상, 기타 연구개발 활용 중 선택하여 기재합니다 (연구개발과제 특성별로 고유한 성과지표를 추가할 수 있습니다).

(3) 세부 정량적 연구개발성과

[과학적 성과]

□ 논문(국내외 전문 학술지) 게재

번호	논문명	학술지명	주저자명	호	국명	발행기관	SCIE 여부 (SCIE/비SCIE)	게재일	등록번호 (ISSN)	기여율
1	MMDCP: Multi-Modal Dental Caries Prediction for Decision Support System Using Deep Learning	Int J Environ Res Public Health.	Ngnamsie Njimbouom S	19(17)			SCIE	2022 Sep 1	1660-4601	100

□ 국내 및 국제 학술회의 발표

번호	회의 명칭	발표자	발표 일시	장소	국명
1	US Korea Conference	Ki-Hwa Kim	2021.12.15 - 12.18	Hyatt Regency Orange County, Garden Grove, CA	USA
2	US Korea Conference	Byeollee Kim	2021.12.15 - 12.18	Hyatt Regency Orange County, Garden Grove, CA	USA
3	US Korea Conference	Juyeon Park	2021.12.15 - 12.18	Hyatt Regency Orange County, Garden Grove, CA	USA
4	EEECS2022 Kor	Candra Zonyfar	2022.07.20 - 07.22	Jeju Island	Korea
5	EEECS2022 Kor	Prince Delator Gidiglo	2022.07.20 - 07.22	Jeju Island	Korea
6	EEECS2022 Kor	Prince Delator Gidiglo	2022.07.20 - 07.22	Jeju Island	Korea
7	EEECS2022 Kor	Soualihou Ngnamsie Njimbouom	2022.07.20 - 07.22	Jeju Island	Korea
8	EEECS2022 Kor	İnae Kang	2022.07.20 - 07.22	Jeju Island	Korea
9	EEECS2022 Kor	Gloria Geine	2022.07.20 - 07.22	Jeju Island	Korea

□ 기술 요약 정보

연도	기술명	요약 내용	기술 완성도	등록 번호	활용 여부	미활용사유	연구개발기관 외 활용여부	허용방식

□ 보고서 원문

연도	보고서 구분	발간일	등록 번호

□ 생명자원(생물자원, 생명정보)/화합물

번호	생명자원(생물자원, 생명정보)/화합물 명	등록/기탁 번호	등록/기탁 기관	발생 연도
1	*Psudedomonas fluorescens* Ant01	CP104408	NCBI(Biosample)	2022

[기술적 성과]

□ 지식재산권(특허, 실용신안, 디자인, 상표, 규격, 신품종, 프로그램)

번호	지식재산권 등 명칭 (건별 각각 기재)	국명	출원				등록			기여율	활용 여부
			출원인	출원일	출원 번호	등록 번호	등록인	등록일	등록 번호		

o 지식재산권 활용 유형

※ 활용의 경우 현재 활용 유형에 √ 표시, 미활용의 경우 향후 활용 예정 유형에 √ 표시합니다(최대 3개 중복선택 가능).

번호	제품화	방어	전용실시	통상실시	무상실시	매매/양도	상호실시	담보대출	투자	기타

□ 저작권(소프트웨어, 서적 등)

번호	저작권명	창작일	저작자명	등록일	등록 번호	저작권자명	기여율

□ 신기술 지정

번호	명칭	출원일	고시일	보호 기간	지정 번호

□ 기술 및 제품 인증

번호	인증 분야	인증 기관	인증 내용		인증 획득일	국가명
			인증명	인증 번호		

□ 표준화

o 국내표준

번호	인증구분[1]」	인증여부[2]」	표준명	표준인증기구명	제안주체	표준종류[3]」	제안/인증일자

* 1」 한국산업규격(KS) 표준, 단체규격 등에서 해당하는 사항을 기재합니다.
* 2」 제안 또는 인증 중 해당하는 사항을 기재합니다.
* 3」 신규 또는 개정 중 해당하는 사항을 기재합니다.

o 국제표준

번호	표준화단계구분[1]」	표준명	표준기구명[2]」	표준분과명	의장단 활동여부	표준특허 추진여부	표준개발 방식[3]」	제안자	표준화 번호	제안일자

* 1」 국제표준 단계 중 신규 작업항목 제안(NP), 국제표준초안(WD), 위원회안(CD), 국제표준안(DIS), 최종국제표준안(FDIS), 국제표준(IS) 중 해당하는 사항을 기재합니다.
* 2」 국제표준화기구(ISO), 국제전기기술위원회(IEC), 공동기술위원회1(JTC1) 중 해당하는 사항을 기재합니다.
* 3」 국제표준(IS), 기술시방서(TS), 기술보고서(TR), 공개활용규격(PAS), 기타 중 해당하는 사항을 기재합니다.

[경제적 성과]

□ 시험제품 제작

번호	시험제품명	출시/제작일	제작 업체명	설치 장소	이용 분야	사업화 소요 기간	인증기관 (해당 시)	인증일 (해당 시)

□ 기술 실시(이전)

번호	기술 이전 유형	기술 실시 계약명	기술 실시 대상 기관	기술 실시 발생일	기술료 (해당 연도 발생액)	누적 징수 현황

* 내부 자금, 신용 대출, 담보 대출, 투자 유치, 기타 등

□ 사업화 투자실적

번호	추가 연구개발 투자	설비 투자	기타 투자	합계	투자 자금 성격*

□ 사업화 현황

번호	사업화 방식[1]」	사업화 형태[2]」	지역[3]」	사업화명	내용	업체명	매출액		매출 발생 연도	기술 수명
							국내 (천원)	국외 (달러)		

* 1」 기술이전 또는 자기실시
* 2」 신제품 개발, 기존 제품 개선, 신공정 개발, 기존 공정 개선 등
* 3」 국내 또는 국외

□ 매출 실적(누적)

사업화명	발생 연도	매출액		합계	산정 방법
		국내(천원)	국외(달러)		
합계					

□ 사업화 계획 및 무역 수지 개선 효과

성과					
사업화 계획	사업화 소요기간(년)				
	소요예산(천원)				
	예상 매출규모(천원)		현재까지	3년 후	5년 후
	시장 점유율	단위(%)	현재까지	3년 후	5년 후
		국내			
		국외			
	향후 관련기술, 제품을 응용한 타 모델, 제품 개발계획				
무역 수지 개선 효과(천원)	수입대체(내수)		현재	3년 후	5년 후
	수출				

□ 고용 창출

순번	사업화명	사업화 업체	고용창출 인원(명)		합계
			yyyy년	yyyy년	
합계					

□ 고용 효과

구분			고용 효과(명)
고용 효과	개발 전	연구인력	
		생산인력	
	개발 후	연구인력	
		생산인력	

□ 비용 절감(누적)

순번	사업화명	발생연도	산정 방법	비용 절감액(천원)
합계				

□ 경제적 파급 효과

(단위: 천원/년)

구분	사업화명	수입 대체	수출 증대	매출 증대	생산성 향상	고용 창출 (인력 양성 수)	기타
해당 연도							
기대 목표							

□ 산업 지원(기술지도)

순번	내용	기간	참석 대상	장소	인원

□ 기술 무역

(단위: 천원)

번호	계약 연월	계약 기술명	계약 업체명	계약업체 국가	기 징수액	총 계약액	해당 연도 징수액	향후 예정액	수출/ 수입

[사회적 성과]

□ 법령 반영

번호	구분 (법률/시행령)	활용 구분 (제정/개정)	명 칭	해당 조항	시행일	관리 부처	제정/개정 내용

□ 정책활용 내용

번호	구분 (제안/채택)	정책명	관련 기관 (담당 부서)	활용 연도	채택 내용

□ 설계 기준/설명서(시방서)/지침/안내서에 반영

번호	구 분 (설계 기준/설명서/지침/안내서)	활용 구분 (신규/개선)	설계 기준/설명서/ 지침/안내서 명칭	반영일	반영 내용

□ 전문 연구 인력 양성

번호	분류	기준 연도	현황									
			학위별				성별		지역별				
			박사	석사	학사	기타	남	여	수도권	충청권	영남권	호남권	기타

□ 산업 기술 인력 양성

번호	프로그램명	프로그램 내용	교육 기관	교육 개최 횟수	총 교육 시간	총 교육 인원

□ 다른 국가연구개발사업에의 활용

번호	중앙행정기관명	사업명	연구개발과제명	연구책임자	연구개발비

□ 국제화 협력성과

번호	구분 (유치/파견)	기간	국가	학위	전공	내용

□ 홍보 실적

번호	홍보 유형	매체명	제목	홍보일

□ 포상 및 수상 실적

번호	종류	포상명	포상 내용	포상 대상	포상일	포상 기관

[인프라 성과]

□ 연구시설·장비

구축기관	연구시설/ 연구장비명	규격 (모델명)	개발여부 (○/×)	연구시설·장비 종합정보시스템* 등록여부	연구시설·장비 종합정보시스템* 등록번호	구축일자 (YY.MM.DD)	구축비용 (천원)	비고 (설치 장소)

*「과학기술기본법 시행령」 제42조제4항제2호에 따른 연구시설·장비 종합정보시스템을 의미합니다.

[그 밖의 성과]

(4) 계획하지 않은 성과 및 관련 분야 기여사항(해당 시 작성합니다)

세부프로젝트의 세부2의 경우, 계획에 없던 내용을 진행하는 사항으로 플라스틱 분해 관련 효소 정보의 데이터베이스 구축을 완료하였고, 플라스틱 분해 관련된 효소 패밀리인 alpha/beta-hydrolase 효소군 예측 모델 개발이 완료 단계임

2) 목표 달성 수준

추 진 목 표	달 성 내 용	달성도(%)
○ 정화관련 Pathway 정보 데이테베이스 구축	○ Pathway, Protein, 미생물 정보 등 관련 데이터베이스 구축	○ 100
○ Pipeline 구축	○ Annotation, Genomic islands 예측 및 Diamond-blastp 분석 pipeline 구축	○ 90
○ Case study 분석	○ 9종의 알려진 균주 case와 본 연구실의 정화미생물 균주의 case study	○ 80
○ CYP관련 효소 데이터 셋 수집	○ CYP 정보 수집과 기질 바인딩 여부에 따른 positive, negative 데이터 셋 확보	○ 80
○ CYP 분류 모델 개발	○ CYP-기질 상호작용 관련 분류 모델 개발	○ 60
○ 플라스틱 생분해 효소 데이터 셋 구축	○ 알려진 데이터베이스에 Manual 수집을 통한 데이터 셋 구축	○ 100
○ 플라스틱 분해 예측 모델 개발	○ CNN을 이용한 플라스틱 분해 예측 모델 개발	○ 50
○ 극한지역 미생물 유전체 분석	○ 남극 토양유래 미생물 1종의 전장유전체 해독 및 분석	○ 100
○ 청정지역과 오염지역의 미생물 유전체 패턴 분석 모델 개발	○ 청정지역과 오염지역에서 발견되는 항생제 내성 유전자에 대한 정량적 유전체 패턴 분석을 수행함	○ 50

4. 목표 미달 시 원인분석

1) 목표 미달 원인(사유) 자체분석 내용

○ 세부 프로젝트 1 :

◦ BiLSTM을 이용한 유전자 클러스터 분석을 진행하지 못함. 기존에 있는 데이터를 이용하여 분석을 완료한 뒤 단점보안을 하고자 biLSTM을 이용하고자 하였으나, 본 연구의 방향이 biLSTM을 이용한 분석과 맞지 않다는 것을 알게됨

○ 세부 프로젝트 2-1 :

◦ Data set을 구성하는데 있어 정리된 데이터베이스가 없고, 수집된 데이터가 100여개 정도로 추가 데이터를 수집하는 과정이 필요하였음. 또한 데이터를 통일시키는 변환 프로그램이 필요하여 데이터 셋 구축 과정에 시간이 소요됨

◦ negative 데이터를 찾는 것에 문제가 생김. Negative data는 인간 유래 CYP 외에 전혀 찾아볼 수 없었기 때문에 Manual로 논문을 검색해야함

○ 세부 프로젝트 2-2 :

◦ 실질적으로 플라스틱을 분해하는지에 대한 기준 실험이 미비하고 알려진 정보가 많지 않아 관련 논문들의 진위성을 검증할 판단기준을 찾는데 시간이 많이 소모됨. 수집한 논문 유래 유전자 정보들은 수동적으로 추출함

◦ 또한, 수집한 데이터가 약 100여 개 남짓으로 머신러닝 또는 딥러닝 모델에 적용하기 어려워 기존에 보고된 TrEMBL 데이터베이스에서 유사성 있는 유전자들을 수집하고자 하였으며, 위 과정 중 발생할 수 있는 위양성 문제에 대응하고자 많은 실험이 필요하였음.

○ 세부 프로젝트 3 :

◦ 항생제 내성 유전자와 병원성 인자는 수평적 유전자 이동을 통해서 빠르게 진화함으로 특정 지역의 특정 종이 보유하는 고유의 유전체 패턴이 다른 균주에서도 쉽게 관찰되기 때문에, 항생제 노출 오염으로부터 구분된 청정지역과 오염지역에서 발견되는 미생물의 유전체 패턴 분석 개발을 위한 데이터베이스를 구축하는데 한계가 존재함.

2) 자체 보완활동

○ 세부프로젝트 1 : 시퀀스 정보다 없는 데이터베이스에 시퀀스 정보를 포함하여 데이터베이스를 구축하였으며, 알려져 있는 Genomic islands 분석을 할 수 있는 프로그램을 실행하여 Gene clsuter보다 큰 범위의 Hydrocarbon degrading genomic islands를 찾고자 하였음

○ 세부프로젝트 2-1 : 더 많은 데이터를 모으기 위해 계속 이용할 수 있는 데이터

베이스를 찾고, 추합하여 하나의 data set으로 만들었음. 이때 통일되지 않은 칼럼이 있으면, 이를 변환하기 위해 ChEMBL 같은 다른 데이터베이스를 사용하기도 함. Negative data의 경우, 인간 유래 CYP는 이전에 출판된 논문에서 정리된 data set을 참고하여 만들기로 하였음. 박테리아 유래 CYP는 데이터가 없어 수작업으로 논문을 보고 모으기로 함

○ 세부프로젝트 2-2: 기재되어있는 내용 이외의 내용을 수집하기 위해, 비슷한 주제의 논문들과 유사 유전자를 수집하였음. Hidden Markov Model을 통해 수집하고자 하는 유전자들의 유사성 및 진정성을 높이고자 하였으며, 비록 적은 양의 데이터일지라도 높은 수준의 limit을 둠으로서 위양성을 줄이고자 하였음

○ 세부 프로젝트 3 : 항생제 내성 유전자를 기준으로 청정지역과 오염지역의 유전체 패턴에서 큰 차이를 관찰하는데 어려움이 있다는 사실을 인지하고, 청정지역으로 생각되는 남극에서 유래한 미생물 유전체 해독과 분석을 진행함. 또한, 청정지역에 서식하는 미생물 유전체 내에서 항생제 내성 유전자와 병원성 인자를 분석하고, 이를 토대로 항생제 오염에 대한 넓은 전파 위험성을 알리고 내성 극복을 위한 유용 유전자 후보 발굴을 진행함

3) 연구개발 과정의 성실성

○ 세부 프로젝트 1 : 실험디자인을 통해 하고자 하는 연구의 방향을 설정하고자 하였는데, 환경정화로 알려져 있는 7가지 미생물 종의 유전체 정보를 활용하여 Case 분석을 하고자 하였으며, 본 연구실에서 실험 결과가 확인된 미생물 유전체 정보를 이용하여 본 연구의 실제 적용 가능성도 실험하였음. 현재까지 이러한 방향으로 연구한 논문은 보고된 바가 없음

○ 세부 프로젝트 2 : 데이터베이스를 찾기 위해 끊임없이 검색하고, 추합 하는 과정을 거쳤음. 최종적으로 PDB와 UniProt을 주 데이터베이스로 선정하였으며, 그 외에 data set의 캄럼을 계속 추가하였음. 기질 부분도 모델 트레이닝을 위하여 input 형태를 계속 바꿔주었음. 또한, 최근 InChI의 형태로 다시 바꿔 시도 중임. 또한, 부족한 데이터를 위해 지속적으로 논문 및 정보를 수집 중에 있으며, 수집된 일부 유전자들에 대해서도 실질적인 실험들을 통해서 문제를 해결하고자 함. 또한, 위양성 문제를 해결하기 위해 상기 문제를 다루고 있는 일부 논문들을 바탕으로 이를 해결하고자 여러 차례 분석 중에 있음

○ 세부 프로젝트 3 : 실험을 통한 항생제 내성 표현형을 확인하고, 해독된 미생물의 전장유전체를 활용하여 유전형 분석을 진행함. 기존의 알려진 항생제 내성유전자 데이터베이스를 활용하여 오염지역, 환경에서 분리된 샘플을 이용하여 항생제 내성에 대한 비교 분석을 진행함. 추가적으로 항생제 내성과 더불어 이차대사산물 생산경로 분석을 통해 항생물질 생산 후보균주로서의 가능성을 탐색함

5. 연구개발성과의 관련 분야에 대한 기여 정도

◦ 한국에서는 다양한 미생물을 이용한 환경오염에 대한 실험은 뒤떨어지고 있는 상황임. 이러한 상황에서 연구를 진행하여 알려진 미생물의 잠재적 가능성을 확인 할 뿐 아니라 유사기능을 할 것으로 예상되는 Genomic Islands를 탐색함으로써 다양한 환경오염물질 정화 미생물 예측에 도움이 될 것으로 사료됨

◦ 또한 본 연구가 융합연구로써 인공지능을 활용하기 위한 전 과정을 진행하였으므로, 인공지능분석 할 때 필요한 환경오염관련 단백질, 혹은 유전체 데이터를 데이터베이스화 하여 접근성을 높임

◦ 미생물을 포함한 효소 연구자료 및 데이터베이스가 충분하지 않고 개별적으로 다뤄지는 부분이 있어 전체적인 유전적, 기능적 파악이 쉽지 않음. 이 부분을 공략함으로써 새로운 다각적 시각으로 연구에 접근할 수 있도록 도움을 줄 것으로 사료됨

◦ 플라스틱 생분해 연구는 많은 연구단체에서 진행 중에 있으나, PET나 일부 플라스틱에만 집중적으로 실시되어지고 있음. 이에 다른 플라스틱 생분해 연구에 밑바탕이 될 수 있는 유전자 데이터들의 마이닝하고 기재함으로서 연구바탕을 구축하고 데이터베이스 및 플라스틱 분해 가능성 유전자의 예측 및 분류 모델 구축을 통해, 관련 연구에 기여할 수 있을 것으로 예상됨

◦ 본 연구에서 분리 동정하여 전장유전체 해독을 마친 미생물 유전체는 향후 청정지역에서의 오염정도를 모니터링하거나, 새로운 항생물질을 포함한 다양한 이차대사산물을 생산 할 수 있는 후보 균주 탐색에 있어 높은 가치를 가짐.

◦ 유전체 기반의 머신러닝, 딥러닝을 활용한 타겟 유전자 예측 소프트웨어는 활발히 개발되고 있으나, 이를 실제 활용한 연구사례는 상대적으로 매우 적은 실정임. 따라서 개발되어 있는 다양한 예측프로그램들을 검증하며 사용하는 과정이 필수적임.

6. 연구개발성과의 관리 및 활용 계획

○ 세부프로젝트 1 : 본 연구에서 구축한 데이터베이스와 관련 프로그램의 단점을 보안한 후속 연구를 진행하고자 함. Genomic Island안의 패턴 및 Metaboilite 관련 Genomic island 예측을 인공지능을 통해 진행하고자 함. 구축된 데이터베이스는 본교 서버컴퓨터에 보관하여 본 연구실에서 실험적으로도 활용하고자 함

○ 세부 프로젝트 2: 본 연구에서 설계 및 구현한 데이터베이스, 분류 모델을 활용하여 PMBD와 같은 웹사이트로 제공하고자함. 사용자들이 입력한 데이터를 통해 데이터베이스를 구축하는 하나의 데이터로 활용 하여 모델의 성능을 향상할 예정임

○ 세부프로젝트 3 : 추가 연구를 통해 극지연구소에 보관되어있는 미생물을 활용

하여, 예측결과에 맞는 항생 물질을 생산여부를 확인하는 추가실험을 진행하고자함. 또한, 청정지역과 오염지역에서 분리된 미생물과 해독한 전장유전체를 활용하여, 핵심 유전자를 발굴 및 구조와 기능에 대한 패턴 분석을 하고자함

< 연구개발성과 활용계획표 >

구분(정량 및 정성적 성과 항목)			연구개발 종료 후 5년 이내
국외논문	SCIE		2
	비SCIE		0
	계		2
국내논문	SCIE		
	비SCIE		
	계		
특허출원	국내		
	국외		
	계		
특허등록	국내		
	국외		
	계		
인력양성	학사		
	석사		
	박사		
	계		
사업화	상품출시		
	기술이전		
	공정개발		
제품개발	시험제품개발		
비임상시험 실시			
임상시험 실시 (IND 승인)	의약품	1상	
		2상	
		3상	
	의료기기		
진료지침개발			
신의료기술개발			
성과홍보			
포상 및 수상실적			
정성적 성과 주요 내용			

< 별첨 자료 >

중앙행정기관 요구사항	별첨 자료
1. 실적 증빙자료	1) 주관연구개발기관 자체평가 의견서
	2) 자체보안관리 진단서
	3) 파견인력별 프로젝트 결과 보고서
	4) 파견인력별 출입국 증명서
	5) 성과증빙자료 (건수별로 증빙 첨부 필수)

7. 기대효과

1) 협력연구 기대 효과

◦ 다각적 문제 확인 및 토의를 통해 아이디어를 공유하고 각 관점에 따른 문제를 제기하여 보충하여 시너지 효과를 내고 심화된 결과를 도출할 수 있음. 또한 유사 연구자들과 교류를 활성화하고, 지속적인 연구를 통한 글로벌 인재양성이 가능함

◦ 현재 연구를 수행 중인 미생물 유전체를 이용한 분석에서 빅데이터를 분석할 수 있도록 컴퓨터 분야와 융합하여 다양한 생물학적 문제를 바이오 빅데이터를 이용하여 생물학적 문제를 해결할 수 있는 인재를 양성할 수 있음

◦ 또한, 해외 대학교에서 비슷한 연구를 진행하는 학생들과의 의사소통을 통해 기존 접근법에 대해서 새로운 시각으로 바라볼 수 있으며, 다양한 시각의 의견을 주고받아 생각의 폭이 넓어진 융합인재를 양성할 수 있음

◦ 연구적으로는 미생물 유전체를 이용한 복잡한 생물학적 시스템 중 정화관련 Gene cluster와 생합성과정 데이터 셋을 구축하고 다양한 데이터베이스를 활용할 수 있음. 컴퓨터와 융합적인 측면에서 각 분야의 의사소통 문제를 해결하고 다양한 시각에서 문제를 바라볼 수 있음. 기존연구에서의 한계 및 새로운 아이디어를 융합한 측면에서 새로운 시각으로 결과를 확인하고 좋은 결과를 얻을 수 있도록 문제 해결능력이 증진됨

2) 연구성과 활용 방안

◦ NGS의 발달로 다양한 유기체의 유전체 연구가 가속화되어 많은 유전자 정보가 데이터베이스화 되고 있지만, 융합연구에는 알맞지 않아 본 연구에서 구축한 데이터베이스가 많이 활용될 것으로 예상됨. Convolutional Neural Network (CNN) 기반 분류 모델을 활용하여 기능적으로 분류하는 새로운 방법들은 후속 연구에 큰 도움을 줄 수 있음.

◦ 환경문제가 전 세계적으로 크게 대두되어 환경관련 산업에 대한 지속적인 연구와 투자가 진행되고 있어 관련된 효소 연구 또한 진행되고 있음. 이에 따라 특정 정화 관련 효소에 대한 예측 모델의 개발은 환경관련 산업의 기반 기술 확보와 환경관련 산업을 선도할 수 있음. 또한, 특정 효소군의 분류체계 확립 및 분류 · 예측 모델의 개발 등의 새로운 접근 방법은 후속 연구에 큰 도움을 줄 수 있음.

◦ 생명공학 지식과 컴퓨터 기술을 융합할 수 있는 글로벌 인재의 양성은 해당 분야의 전문 인력 양성뿐만 아니라 해당 연구인원들의 폭 넓은 문제 해결능력 및 협업 능력 증진할 수 있음.

3) 파견에 따른 글로벌 인재양성 효과

◦ 해외에서의 박사, 박사 후 연구원으로 연구를 지속 가능성 및 관련 기회 창출을 통해 AI기술 분야와 환경・생물 분야에 기여할 수 있는 핵심 연구인력 양성

◦ 국내 및 해외에서의 환경문제 분석파악과 고도화를 통한 국가 경쟁력 확보

◦ 영어, 전문기술을 습득함으로써 인재 개개인의 실무 능력 및 연구 기술 고도화

4) 후속 연구계획 및 성과관리 방안

◦ 본 과제에서 수행한 환경정화 효소 관련 데이터를 통해 딥러닝 모델을 활용하여 효소-기질 간의 상호작용을 예측하는 연구를 진행할 계획임.

◦ 또한, 본 과제를 진행하면서 수집한 데이터를 통합한 환경정화 관련 바이오 데이터베이스를 구축할 계획임.

◦ 추가적으로 데이터베이스를 연구자가 쉽게 접근할 수 있는 웹 서비스를 제공하여 데이터 공유를 활성화 할 예정임.

◦ 무분별한 데이터 유출과 출처 없는 사용을 방지하기 위한 오픈소스 라이센스 관리에 대한 연구와 데이터 보안을 위한 데이터베이스 시스템 구축에 대한 연구를 진행할 계획임.

◦ 또한 플라스틱 생분해와 관련된 효소에 대한 예측 모델의 연구 결과를 통해 실제 플라스틱 분해 활성을 보이는 효소에 대한 실험과 연구를 진행할 계획임.

5) 파견인력 등 진로 추적 등 사후관리

가. 주기적 취업 여부 조사 실시

◦ 파견 연구 이후, 6 개월 단위로 2회 해당분야 취업 또는 진학 여부 조사

나. 파견 후 진학, 취업 등 진로 설계를 위한 멘토링 프로그램 참여 및 관련 정보 제공

◦ 참여대학과 공동연구기관의 멘토링 프로그램 운영 및 진학, 취업 등을 위한 컨설팅 기회 제공, 파견 연구 경험과 관련된 연구 기관 정보 제공

◦ 학생의 진로 방향과 목표에 따른 멘토링 전문기관 프로그램 정보 및 참여 제공

◦ 해외 대학원 진학 또는 박사후 연구원 등으로 연구를 지속하고 싶은 학생의 경우 파견연구대학의 교수님들과 유관기관 연구진들의 대학 및 연구소에 관한 정보 제공 및 기회 마련 (Georgia State University, University of Texas at San Antonio, Fordham University, University of Florida, University of Texas at Dallas, George Mason University, Desert Research Institute, University of Nevada, Las Vegas 등)

다. 성과관리 및 경험 공유

◦ 파견연구 후 참여자의 역량측정, 대학 자체평가, 성과지표관리 등을 통한 양적 또는 질적 평가를 진행

◦ 참여 학생들의 연구노트 및 결과보고서를 보관 및 관리

◦ 신진 연구 인력에게 파견인력들의 연구노트 및 결과보고서를 공유하여 파견 경험 및 해외 연구 경험 공유

◦ 본 과제 파견 인력과 향후 신진과제의 연구 인력들과의 공동연구를 통해 파견 연구 경험을 활용한 협동 연구 진행

8. 예산집행실적

1) 예산 총괄표

연구비 (천원)	정부출연금 (a)	민간부담금			합계 (a+b)
		현물	현금	소계(b)	
500,000	500,000				500,000

2) 파견인력 지원비

- 주관연구개발기관

구분	1차년도(원)	2차년도(원)
o 파견인력 지원금		
- 학생인건비	44,490,000	53,026,470
- 체재비	23,940,960	42,909,540
- 교육비	-	-
- 출국준비금(항공료, 비자, 여행자보험 등)	11,716,991	
합 계	**80,147,951**	**95,936,010**

- 공동연구개발기관

구분	1차년도(원)	2차년도(원)
o 파견인력 지원금		
- 학생인건비	11,448,668	15,340,820
- 체재비	24,897,687	
- 교육비		
- 출국준비금(항공료, 비자, 여행자보험 등)	2,755,428	147,000
합 계	39,101,783	15,487,820

3) 비목/세목별 예산 집행실적*(주관·공동연구개발기관별로 작성)*

- 주관연구개발기관

(단위 : 천원, %)

비 목 별			정부출연금	구성비
1. 직 접 비			355,739	90.81
1.1 인건비				
1.2. 학생인건비			107,971	27.56
	- 파견인력 인건비*		99,811	25.48
	- 참여연구원 인건비		8,160	2.08
1.3. 연구시설 · 장비비			11,925	3.04
1.4. 연구재료비				
1.5. 연구활동비			219,404	56.01
	- 파견인력 체재비		63,904	16.31
	- 교육비(자율과정)			
	- 연구인력활용비(인턴수당)			
	- 파견인력 출국준비금		11,717	2.99
	- 프로젝트 연구수행비			
		- 지식재산 창출 활동비	0	
		- 외부전문기술 활용비**	120,000	30.63
		- 회의비	7,227	1.84
		- 출장비	1,838	0.47
		- 소프트웨어 활용비		
		- 연구실 운영비	5,657	1.44
		- 연구인력지원비	3,000	0.77
		- 종합사업관리비		
		- 위탁정산수수료	1,812	0.46
		- 논문게재비	4,249	1.08
1.6. 연구수당			16,436	4.20
2. 간 접 비			36,000	9.19
총 계			**391,736**	**100.00%**

- 공동연구개발기관

(단위 : 천원, %)

비 목 별			정부출연금	구성비
1. 직 접 비				
1.1 인건비				
1.2. 학생인건비				
	- 파견인력 인건비*		26,789.488	26.79%
	- 참여연구원 인건비			
1.3. 연구시설 · 장비비				
1.4. 연구재료비				
1.5. 연구활동비				
	- 파견인력 체재비		24,898.687	24.90%
	- 교육비(자율과정)			
	- 연구인력활용비(인턴수당)			
	- 파견인력 출국준비금		348.470	0.35%
	- 프로젝트 연구수행비			
		- 지식재산 창출 활동비		
		- 외부전문기술 활용비**	30,174.397	30.18%
		- 회의비		
		- 출장비	2,573.958	2.57%
		- 소프트웨어 활용비		
		- 연구실 운영비	2,992.212	2.99%
		- 연구인력지원비		
		- 종합사업관리비		
		- 그밖의 비용***		
1.6. 연구수당			3,221	3.22%
2. 간 접 비			9,000	9.00%
총 계			99,997.212	100.0%

부 록 [실적 증빙자료]

* 성과별 증빙은 필수 (제본제출 없이 온라인으로 최종보고서와 함께 제출)

* 파견인력 별 프로젝트 결과보고서 및 출입국증명서, 여권사본(출입국 도장) 제출 필수

【별첨 1】주관연구개발기관 자체평가 의견서

【별첨 2】자체보안관리 진단서

【별첨 3】파견인력 별 프로젝트 결과보고서

【별첨 4】파견인력 출입국 증명서

【별첨 5】성과 증빙자료 (건수별로 증빙 첨부 필수)

【별첨 1】주관연구개발기관 자체평가 의견서

주관연구개발기관	선문대학교
사 업 명	2021년 글로벌 핵심인재 양성지원 사업
과제명	미생물 게놈 빅데이터 분석을 위한 AI 기반 환경 ICT 융합 글로벌 인재양성

자 체 평 가 의 견

1. 연구 수행 및 성과

○ 연구수행 측면

- 각 프로젝트별로 필요한 데이터를 수집하여 데이터베이스 및 데이터셋을 구축하여 활용하였음. 본 연구에 필요한 데이터는 많이 알려지지 않아 본 연구에서 구축한 데이터베이스가 향후 연구에 많은 도움이 될 것으로 사료됨

- 세부 프로젝트 1의 경우 기존 프로그램을 활용한 Genome islands를 Hydrocarbon분해와 관련된 부분을 찾는 연구를 알려진 7개의 Genome 유전체 정보를 활용하여 Case study를 진행하였고, 또한 본 연구실에 보유하고 있는 활성 미생물 분석을 통해 본 연구의 적용가능성을 시험하고자 하였음. 이 과정은 시행된 바가 없으며 환경 정화 관련된 미생물의 유전체 분석시 유용할 것으로 사료됨. 또한 본 연구는 지금 논문 투고중에 있으며, 향후 1년 이내 투고 될 것임

- 세부 프로젝트 2의 경우 계획보다 세부적으로 크게 두가지 효소군에 대한 연구가 진행되었으며, 크게 CYP와 플라스틱 분해관련 효소에 대한 미지 단백질 기능 예측을 시도하고 자 하였음. 본 과제가 미생물 데이터를 활용하는 것으로 CYP 데이터 셋 구축에 많은 어려움을 겪었으나, 다양한 데이터베이스의 활용과 연구자들의 다각적 문제해결 제안을 통해 데이터 셋을 구축하였고, 인공지능 모델을 설계하고자 하였음. 기존 논문에 보고된 모델을 분석하여 모델 설계에 도움이 되고자 하였으며, 이는 차후 데이터 전처리 과정이나 모델 세부 설계에 도움이 될 것으로 사료됨. 이러한 연구는 기존에 예측이 되지 않았던 부분에 대한 예측이 가능할 것임. 또한 플라스틱 분해관련 효소 예측에는 기존 알려진 모델을 활용하여 예측을 시도함으로써 적절한 모델 및 파라미터를 설정할 수 있었으며, 최적화를 통해 모델을 개발하고자 함

세부 프로젝트 3의 경우 기존 알려진 환경유래 미생물과 오염지역 미생물의 항생제 내성 유전자 패턴을 분석할 뿐 아니라 환경유래. 미생물의 유전체 분석 및 세부 Gene cluster, Pathway, Genomic island 분석을 통해 환경유래 미생물의 항생제 내성 활성을 확인 할 수 있었음. 이 결과와 더불어 실제 실험을 통해 확인하는 과정을 진행 중에 있으며, 이는 진화론적, 지리학적으로 해석되어 후속 연구에 도움을 줄 수 있음

○ 국제협력 연구 측면

- 컴퓨터공학적인 수업을 통한 생물학 데이터의 접근 및 데이터베이스 설계에 도움을 주었으며, 해외연구기관의 머신러닝 수업을 통해 실험 설계를 구체화 하고 프로그래밍 언어를 통해 모델의 중요성 및 타당성을 논의 할 수 있음

- 해외협력기관에 소속된 학생들과의 토의를 통해 본 연구의 문제 설계 및 실험 디자인에 도움을 받을 수 있었고, 이는 데이터 전처리, 모델 세부 연구의 깊이있는 탐색

이 가능해짐

○ 인력양성 측면

- 파견된 인력들이 융합연구에 대한 폭 넓은 이해가 가능해지면서, 생물학 문제를 컴퓨터공학 적으로 문제 설계를 할 수 있게 됨으로써 융합연구의 전문 인력양성 가능

2. 추진일정
수정된 추진일정에 따라 진행되었음.

3. 연구비 집행 현황
코로나로 인한 교수진의 해외방문이 집행되지 않아 수정, 집행되었음

4. 기타 종합의견
파견연구를 위한 파견 계획의 차질로 인한 수정이 있었으나, 그 외 파견이 수정 계획대로 진행되었으며, 프로젝트의 내용상 환경에 대한 다양한 관점에서 프로젝트가 진행되어 인력양성과 더불어 환경오염 관련 문제 해결을 위한 연구의 기반이 될 것으로 사료됨

2022 년 9 월 22 일

주관연구개발기관장	직위 : 단장	성명 :	김종해	(인)
연구책임자	직위 : 부교수	성명 :	김정동	(직인)

이 가능해짐

○ 인력양성 측면

- 파견된 인력들이 융합연구에 대한 폭 넓은 이해가 가능해지면서, 생물학 문제를 컴퓨터공학 적으로 문제 설계를 할 수 있게 됨으로써 융합연구의 전문 인력양성 가능

2. 추진일정
수정된 추진일정에 따라 진행되었음.

3. 연구비 집행 현황
코로나로 인한 교수진의 해외방문이 집행되지 않아 수정, 집행되었음

4. 기타 종합의견
파견연구를 위한 파견 계획의 차질로 인한 수정이 있었으나, 그 외 파견이 수정 계획대로 진행되었으며, 프로젝트의 내용상 환경에 대한 다양한 관점에서 프로젝트가 진행되어 인력양성과 더불어 환경오염 관련 문제 해결을 위한 연구의 기반이 될 것으로 사료됨

2022 년 9 월 22 일

주관연구개발기관장	직위 : 단장	성명 :	김종해	(직인)
연구책임자	직위 : 부교수	성명 :	김정동	(인)

【별첨 2】자체보안관리 진단표

자체보안관리 진단표

□ 과제현황

사 업 명	2021년 글로벌 핵심인재 양성지원 사업		
과 제 명	미생물 게놈 빅데이터 분석을 위한 AI 기반 환경 ICT 융합 글로벌 인재양성		
주관연구개발기관	선문대학교	연구책임자	김 정 동
총 협약기간	2021. 5. 1. ~ 2022. 8. 31. (16개월)		

□ 진단항목

구분	체크항목	결과 체크(√표)	비고 (미실시 사유)
보안관리 체계	o 기관 내 보안관리규정을 제정/적용하고 있다	O (√), X ( )	
	o 보안관리 조직이 있으며, 자체 보안점검실시 등 잘 운영되고 있다	O (√), X ( )	
	o 보안교육을 정기적(1회이상/연)으로 실시하고 있다	O (√), X ( )	
	o 보안사고에 대한 방지대책 및 비상시 대응계획이 준비되어 있다	O (√), X ( )	
참여연구원 관리	o 참여연구원에 대하여 보안서약서를 받았다	O (√), X ( )	
	o 참여연구원에게 보안관리의 중요성 등을 인식시키고 있다	O (√), X ( )	
연구개발 내용/결과 관리	o 주요 연구자료 및 성과물의 무단유출 방지대책을 수립하고 있다	O (√), X ( )	
	o 보안성 검토 방법 및 절차를 이행하고 있다	O (√), X ( )	
	o 기술이전 관련 내부규정 및 절차를 준수하고 있다	O (√), X ( )	
연구시설 관리	o 연구시설 보안관련 내부규정 또는 지침을 이행하고 있다	O (√), X ( )	
	o 주요 시설에는 보안장비가 설치되어 있다	O (√), X ( )	
	o 보호구역이 지정되어 있다	O (√), X ( )	
정보통신망 관리	o 정보통신망 보안관련 내부규정 또는 지침이 구비되어 있다	O (√), X ( )	
	o 보안관리책임자의 승인 항목이 구분되어 있다	O (√), X ( )	
	o 주요 데이터에 대해 백업을 실시하고 있다	O (√), X ( )	
	o 개인용 정보통신장비(노트북, USB메모리)에 대하여 인가/관리중이다	O (√), X ( )	
	o 전산망 보호를 위한 HW 및 SW 등을 도입하여 적용하고 있다	O (√), X ( )	
	o 직책, 임무별 열람 권한을 차등화하여 부여하고 있다	O (√), X ( )	

위 과제수행과 관련하여 보안관리 자체진단 결과가 위의 사실과 틀림없음을 확인합니다.

2022년 9월 22일

주관연구개발기관명 : 선문대학교 산학협력단 (인)

연구책임자 : 김정동 (인)

정보통신기획평가원장 귀하

【별첨 2】자체보안관리 진단표

자체보안관리 진단표

□ 과제현황

사 업 명	2021년 글로벌 핵심인재 양성지원 사업		
과 제 명	미생물 게놈 빅데이터 분석을 위한 AI 기반 환경 ICT 융합 글로벌 인재양성		
주관연구개발기관	선문대학교	연구책임자	김 정 동
총 협약기간	2021. 5. 1. ~ 2022. 8. 31. (16개월)		

□ 진단항목

구분	체크항목	결과 체크(√표)	비고 (미실시 사유)
보안관리 체계	o 기관 내 보안관리규정을 제정/적용하고 있다	O(√), X( )	
	o 보안관리 조직이 있으며, 자체 보안점검실시 등 잘 운영되고 있다	O(√), X( )	
	o 보안교육을 정기적(1회이상/연)으로 실시하고 있다	O(√), X( )	
	o 보안사고에 대한 방지대책 및 비상시 대응계획이 준비되어 있다	O(√), X( )	
참여연구원 관리	o 참여연구원에 대하여 보안서약서를 받았다	O(√), X( )	
	o 참여연구원에게 보안관리의 중요성 등을 인식시키고 있다	O(√), X( )	
연구개발 내용/결과 관리	o 주요 연구자료 및 성과물의 무단유출 방지대책을 수립하고 있다	O(√), X( )	
	o 보안성 검토 방법 및 절차를 이행하고 있다	O(√), X( )	
	o 기술이전 관련 내부규정 및 절차를 준수하고 있다	O(√), X( )	
연구시설 관리	o 연구시설 보안관련 내부규정 또는 지침을 이행하고 있다	O(√), X( )	
	o 주요 시설에는 보안장비가 설치되어 있다	O(√), X( )	
	o 보호구역이 지정되어 있다	O(√), X( )	
정보통신망 관리	o 정보통신망 보안관련 내부규정 또는 지침이 구비되어 있다	O(√), X( )	
	o 보안관리책임자의 승인 항목이 구분되어 있다	O(√), X( )	
	o 주요 데이터에 대해 백업을 실시하고 있다	O(√), X( )	
	o 개인용 정보통신장비(노트북, USB메모리)에 대하여 인가/관리중이다	O(√), X( )	
	o 전산망 보호를 위한 HW 및 SW 등을 도입하여 적용하고 있다	O(√), X( )	
	o 직책, 임무별 열람 권한을 차등화하여 부여하고 있다	O(√), X( )	

위 과제수행과 관련하여 보안관리 자체진단 결과가 위의 사실과 틀림없음을 확인합니다.

2022년 9월 22일

주관연구개발기관명 : 선문대학교 산학협력단 (인)

연구책임자 : 김정동 (인)

정보통신기획평가원장 귀하

【별첨 3】파견인력 별 프로젝트 결과보고서

2022년 글로벌 핵심인재 양성지원 사업
파견인력 프로젝트 결과보고서

2022. 9. 22.

유형구분	공동연구
주관연구개발기관명	미생물 게놈 빅데이터 분석을 위한 AI 기반 환경 ICT 융합 글로벌 인재양성
과제명	선문대학교
연구책임자명	김정동
과제기간	2021. 5. 1. ~ 2022. 8. 31.
파견인력수	6

프로젝트 결과보고서-(김기화)

<table>
<tr><td rowspan="5">파견개요</td><td>이름</td><td>김기화</td><td>대학</td><td>선문대학교</td></tr>
<tr><td>학과</td><td>생명공학과</td><td>세부전공</td><td>바이오빅데이터융합</td></tr>
<tr><td>파견국가
(도시)</td><td>U.S.A (Nevada)</td><td>파견기관명</td><td>University of Nevada, Las Vegas</td></tr>
<tr><td>해외기관
지도인력</td><td>Mingon Kang. Ph. D.,
Assistant Professor of
Computer Science</td><td>총연구기간
(파견기간)</td><td>210501~220831
(210826~220430)</td></tr>
<tr><td>참여
프로젝트명</td><td colspan="3">환경정화 관련 미생물 유전체 종합 분석</td></tr>
<tr><td rowspan="3">프로젝트결과</td><td>연구주제</td><td colspan="3">정화관련 효소군 분류</td></tr>
<tr><td>수행역할</td><td colspan="3">환경정화 관련 효소 중 Cytochrome P450에 대한 데이터 수집 및
생명공학 지식 전달</td></tr>
<tr><td>연구수행
결과</td><td colspan="3">환경 정화 관련 효소를 주제로 산화환원 효소인 Cytochrome P450(CYP)을 선정하였음. 이 효소는 기질과 반응하여 주로 -OH 치환기로 전환하는 역할을 수행하는데, 이는 인간의 간에서 하는 해독 작용을 들 수 있음. 이 효소는 인간 뿐 만 아니라 살아있는 생명체 모두에게서 존재하고, 다양한 기질 스펙트럼을 가지고 있음. 그러나, 모든 효소가 그렇듯 기질 특이성이 있기 때문에 각 효소마다 작용할 수 있는 기질들이 차이가 있음. 같은 이름으로 불리는 효소조차 유전적으로 차이가 있으므로 기질과의 작용은 더욱 예측하기 어려움. 그러므로 각 효소의 시퀀스 정보만으로 기질의 특이성은 예측하는 방법을 개발하고자 함.

이미 효소의 구조를 밝히고, 구조에 기질을 닥킹하여 상화작용을 알아보는 인실리코 연구는 과거부터 꾸준히 연구되어왔음. 더 나아가 움직임을 부여한 효소에 기질을 서서히 움직여서 시간당 바인딩하는 모습을 보는 molecular dynamics 기법까지 활발하게 적용되고 있음. 이들은 가상의 환경에서 가상의 데이터를 이용해 시뮬레이션하는 것이지만, 높은 정확도를 보여 많이 사용되는 스크리닝 기법임. 이때, 더욱 정확한 결과를 얻기 위해서는 높은 해상도를 가진 단백질 3차 구조를 사용해야 하는데, 그러기 위해서는 많은 시간과 실험을 위한 버퍼 조성 등이 맞아야 함.</td></tr>
</table>

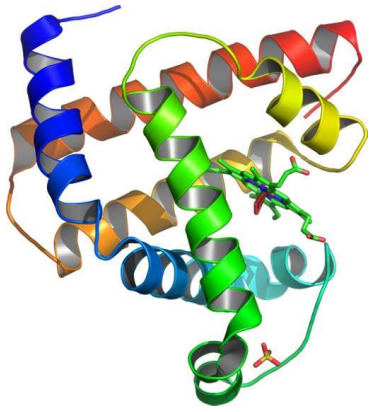


그림. 단백질 3차 구조 예시

이러한 실험적 단계의 소모를 줄여보고자 단백질의 1차 구조인 시퀀스에서 기질과의 상호작용을 예측할 수 있는 알고리즘을 개발하고자 함.

먼저 dataset을 만들기 위해 다양한 데이터베이스를 검색함. 인간 유래 CYP는 많은 논문들과 연구 결과가 정리된 데이터베이스가 존재하고 이들을 사용한 많은 알고리즘이 개발되었음. 특히, drug discovery 관련한 논문들이 많았음. 이는 인간 유래 CYP가 80% 이상의 생리활성에 작용함으로써 중요한 효소로 연구되고 있기 때문임. 최근까지도 인간 유래 CYP에 대한 기질과의 상호작용 연구는 활발히 진행 중에 있으며, 더 나아가 기질 및 저해제를 구별하여 예측할 수 있는 알고리즘도 다양하게 알려짐. 이에 반해 박테리아 유래 CYP는 데이터베이스를 찾기 어려움. 시퀀스, ID, 효소의 명명 등이 설명된 데이터베이스 외에 활성이 있는 기질 및 저해제에 대한 활성 정보를 포함하는 경우는 찾을 수 없었음. 소수의 데이터베이스는 이를 포함하고 있었지만, CYP로 제한했을 때, 너무나 적은 데이터를 포함하고 있었음.

기질과 저해제라는 명확한 구분이 되지 않더라도 박테리아 유래 CYP와 기질이 결합한다는 정보를 알 수 있는 데이터베이스를 선정하기 위해 단백질의 3차 구조를 모아놓은 'Protein Data Bank(PDB)'를 사용함.

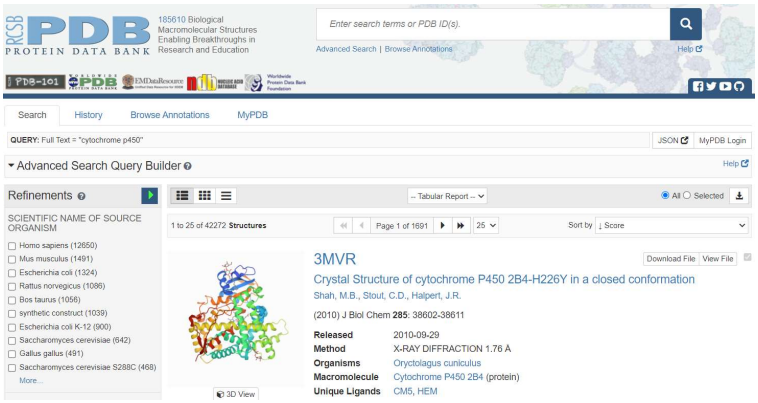


그림. Data Crawling을 위한 PDB 사이트의 Cytochrome P450 데이터 예시

PDB에는 실험적으로 단백질의 3차 구조와 함께 기원, 아미노산 시퀀스, 리간드 등의 다양한 정보들이 포함되어 있음. 특히, 리간드에는 단백질의 구조를 밝힐 때 같이 들어간 화합물들이 정리되어있으므로 이 부분을 크롤링하여 데이터셋을 구성함. 하지만 이 리간드 정보는 단백질의 활성부위에 들어간 기질 뿐만 아니라, 주변에 위치한 버퍼 성분의 화합물, 그리고 저해제들이 구분없이 포함되어 있었기 때문에 이들을 최소한으로 걸러주기 위한 장치가 필요 했음. 그러므로 저분자인 탄소 5개 이하의 화합물들은 제외하고 데이터셋을 모으기로 함.

기질과 효소의 상호작용을 얻을 수 있는 다른 데이터베이스로 Uniprot를 추가함. Uniprot은 광범위한 유전체 정보를 포함하고 있으며, 이들의 활성 정보도 포함되어 있음. 단백질과 기질의 상호작용을 예측하기 위해서는 한 시퀀스와 다양한 기질들의 상호작용을 트레이닝 시키는 것이 중요하므로, Uniprot과 PDB의 시퀀스를 통합하여 이들의 기질 또한 통합함.

이렇게 모은 dataset은 크게 아미노산 시퀀스와 기질, 그리고 그들의 상호작용으로 이루어짐. Input 넣고자하는 정보는 아미노산 시퀀스와 기질이며, 그로 인해 얻을 Output은 그들의 상호작용 여부 예측임. 이때 아미노산 시퀀스와 기질의 정보는 각각 알파벳, 그래프의 형태를 띠고 있기때문에 이들을 인배딩하기 위한 장치가 필요함.

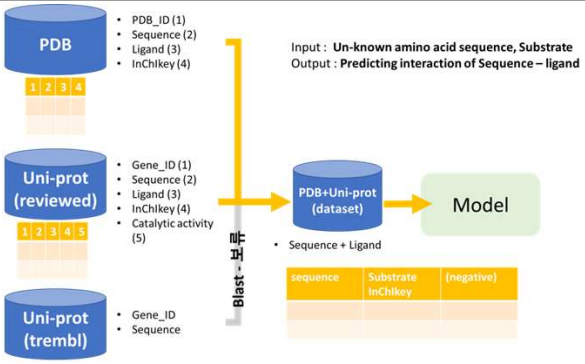
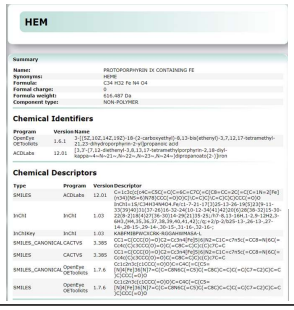


그림. 모델을 만들기전 데이터 전처리 모식도

그래프로 존재하는 기질 정보를 매트릭스화 하기 위한 다양한 descriptor가 있지만, 그 중 RDkit을 사용하기로 함. RDkit은 화합물의 SMILE이나 InChI를 이용하여 컴퓨터가 인식할 수 있는 문자열로 변환해주는 역할을 함. PDB와 Uniprot에서 얻은 SMILE을 추합하여 RDkit을 돌리고자 하였으나, 에러가 생겨서 넘어가지 못하였음. 이는 CYP 효소에 필수적으로 들어있는 heme 화합물 때문인데, heme에는 금속 이온은 Fe가 비공유결합으로 연결되어 있음. 이 때문에 SMILE에는 비공유결합을 나타내는 '.' 표시가 들어가게 됨. 이 '.' 표시가 에러를 내는 주범이나, 다른 논문에는 전처리를 통해 이러한 항목들을 제거하고 진행하였음. 그러나, 지금 우리의 data set에서 이 화합물은 큰 부분을 차지하고 있어 함부로 삭제할 수 없었음. 그리하여 이들을 제거하지 않는 다른 방법을 찾고자 함.

SMILE은 기본적으로 문자열을 형성하는데 규칙이 정해져 있지만, 화합물을 읽는 방향이나 표시하는데 있어 좀 자유롭기 때문에 같은 화합물이어도 조금씩 다른 문자열로 구성되기도 함. 이들은 문자열의 배열, 구성, 기호가 차이가 있더라도 특수하게 허용되는 경우가 많음. 그러므로 한가지 화합물을 넣어도 사용된 알고리즘에 따라 조금씩 변형된 문자열을 얻을 수 있음.



- Canonical Smiles in RDKit in Hem
  - C=Cc1c(C)c2cc3nc(cc4nc(cc5[nH]c(cc1[n-]2)c(C)c5C=C)C(C)=C4CCC(=O)O)C(CCC(=O)[O-])=C3C.[Fe+2]
- Isomeric Smiles in RDKit in Hem
  - C=Cc1c(C)c2cc3nc(cc4nc(cc5c(C=C)c(C)c(cc1[n-]2)[nH]5)C(C)=C4CCC(=O)O)C(CCC(=O)[O-])=C3C.[Fe+2]
- OpenEye in PDB in Hem
  - Cc1c2n3c(c1CCC(=O)O)C=C4C(=C(C5=[N]4[Fe]36[N]7=C(C=C8N6C(=C5)C(=C8C)C=C)C(=C(C7=C2)C)C=C)CCC(=O)O
- OpenEye with stereo in PDB in Hem
  - None

그림. PDBJ와 각기 다른 프로그램으로 확인한 SMILE 표기법

이것은 각기 다른 데이터베이스에서 데이터를 추합한 결과, 전처리하기 어려운 부분이었기 때문에 또 다른 방법으로 InChI를 사용하기로 함. InChI는 승인된 IUPAC 명명 규칙인 화학 물질의 정식 명명 시스템이기 때문에 혼돈의 상황이 오지 않을 것이라 생각됨.

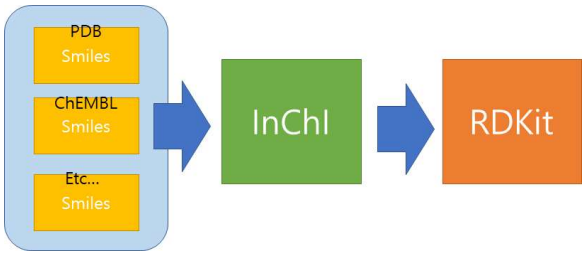


그림. RDKit을 사용하기 전 전처리 단계

현재 dataset에는 positive data만 존재할 뿐, negative 상호작용 data가 있지 않음. 특히, 박테리아 유래 CYP에 대한 상호작용 없는 기질을 찾을 방법이 없었음. 그리하여 선택한 방법은 논문에서 검색하는 방법임. 키워드는 'P450, bacteria' 두 가지로 PubMed에서 검색하였고, 총 4915개의 논문이 검색되었음. 저널별로 검색하여 가장 많은 논문 수를 출판한 저널을 선택하였음. 현재 211개의 논문을 보여 negative 결과를 수집 중에 있음.

	Journal name	Amount	Journal's full name	I.F.
1	J Biol Chem	211	Journal of Biological Chemistry	5.486
2	Arch Biochem Biophys	169	Archives of Biochemistry and Biophysics	4.013
3	Biochemistry	143	Biochemistry	3.162
4	Appl Microbiol Biotechnol	119	Applied Microbiology and Biotechnology	3.53
5	Mutat Res	115	Mutation Research	5.657
6	Biochem Biophys Res Commun	108	Biochemical and Biophysical Research Communications	3.322

그림. Nrgative data 검색을 위한 저널 탐색

연구성과	
프로젝트 기술분야	생명공학, 머신러닝

프로젝트 결과보고서-(김별이)

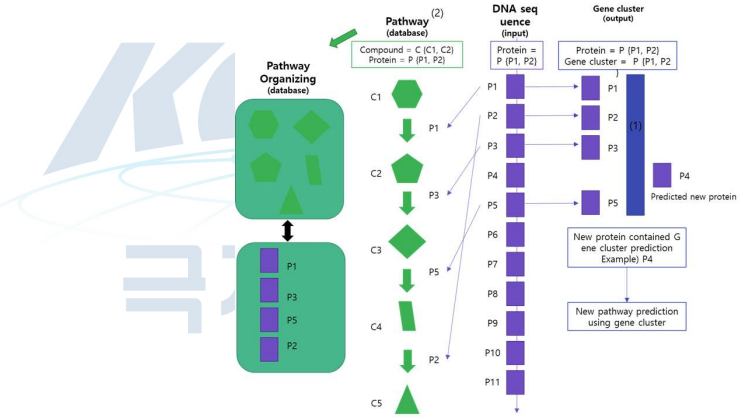
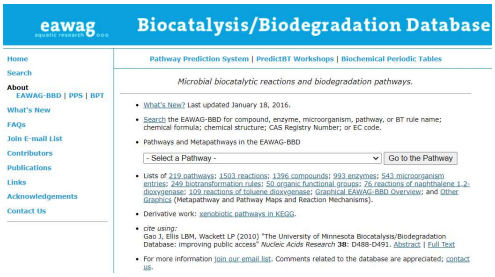
<table>
<tr><td rowspan="5">파견개요</td><td>이름</td><td>김별이</td><td>대학</td><td>선문대학교</td></tr>
<tr><td>학과</td><td>생명공학과</td><td>세부전공</td><td>바이오빅데이터융합전공</td></tr>
<tr><td>파견국가
(도시)</td><td>U.S.A (Nevada)</td><td>파견기관명</td><td>University of Nevada, Las Vegas</td></tr>
<tr><td>해외기관
지도인력</td><td>Mingon Kang. Ph. D.,
Assistant Professor of
Computer Science</td><td>총연구기간
(파견기간)</td><td>210501~220831
(210826~220422)</td></tr>
<tr><td>참여
프로젝트명</td><td colspan="3">환경정화 관련 미생물 유전체 종합 분석</td></tr>
<tr><td rowspan="3">프로젝트결과</td><td>연구주제</td><td colspan="3">환경정화 미생물 예측</td></tr>
<tr><td>수행역할</td><td colspan="3">프로젝트 디자인 및 수행</td></tr>
<tr><td>연구수행
결과</td><td colspan="3">○ Gene cluster와 Pathway 정리 (그림)
 

그림. Pathway와 Gene cluster
○ 새로운 Database 구축
- 기존계획의 문제점으로 plastics에 대한 data만 가지고 있어 제한적이고, 데이터의 양 부족
- Biocatalysis/Biodegradation Database는 219개의 pathway 정보와 1503개의 reaction이 포함되어 있음 (그림)
 

그림, Biocatalysis/biodegradation Database</td></tr>
</table>

- 위 Database를 pathway, enzyme, compound로 class를 나누어 새로운 형식의 Database를 구축하였음 (표 1).

표 1. 구축된 데이터베이스

pathway_name	microorganisms	reaction_from	reaction_to	reaction_id	enzyme_name
Citronellol	Pseudomonas aeruginosa	(2E)-5-Methylhexa-2,4-dienoyl-CoA	3-Hydroxy-5-methylhex-4-enoyl-CoA	r1304	enoyl-CoA hydratase
Geraniol	Pseudomonas aeruginosa	Geranial	Geranylate	r1164	geranial dehydrogenase
Geraniol	Pseudomonas aeruginosa	Geraniol	Geranial	r1163	geraniol dehydrogenase
m-Cresol	Pseudomonas alcaligenes NCIB 9867	3-Hydroxybenzoate	Gentisate	r0402	3-hydroxybenzoate 6-hydroxylas
m-Cresol	Pseudomonas alcaligenes NCIB 9867	3-Hydroxybenzaldehyde	3-Hydroxybenzoate	r0401	benzaldehyde dehydrogenase
Acrylonitrile	Pseudomonas chlororaphis B23	Acrylonitrile	Acrylate	r0622	aliphatic nitrilase
Acrylonitrile	Pseudomonas chlororaphis B23	Acrylate	Acrylyl-CoA	r0085	glutaconate CoA-transferase
Acrylonitrile	Pseudomonas chlororaphis B23	Acrylyl-CoA	Lactoyl-CoA	r0086	lactoyl-CoA dehydratase
Acrylonitrile	Pseudomonas chlororaphis B23	Acrylonitrile	Acrylamide	r0083	nitrile hydratase
Tetrachlorobenzene	Pseudomonas chlororaphis RW71	Tetrachlorocatechol	Tetrachloro-cis,cis-muconate	r0801	catechol 1,2-dioxygenase
Tetrachlorobenzene	Pseudomonas chlororaphis RW71	Tetrachloro-cis,cis-muconate	2,3,5-Trichlorodienelactone	r0802	chloromuconate cycloisomerase
Tetrachlorobenzene	Pseudomonas chlororaphis RW71	2,3,5-Trichloromaleylacetic acid	2,4-Dichloro-3-oxoadipate	r0804	maleylacetate reductase
1,3-Dichloropropene	Pseudomonas cichorii 170	cis-3-Chloro-2-propene-1-ol	cis-3-Chloroallyl aldehyde	r0691	alcohol dehydrogenase
1,3-Dichloropropene	Pseudomonas cichorii 170	trans-3-Chloro-2-propene-1-ol	trans-3-Chloroallyl aldehyde	r0690	alcohol dehydrogenase
1,3-Dichloropropene	Pseudomonas cichorii 170	cis-3-Chloroallyl aldehyde	cis-3-Chloroacrylic acid	r0693	aldehyde dehydrogenase
1,3-Dichloropropene	Pseudomonas cichorii 170	trans-3-Chloroallyl aldehyde	trans-3-Chloroacrylic acid	r0692	aldehyde dehydrogenase
1,3-Dichloropropene	Pseudomonas cichorii 170	cis-1,3-Dichloropropene	cis-3-Chloro-2-propene-1-ol	r0687	haloalkane dehalogenase
1,3-Dichloropropene	Pseudomonas cichorii 170	trans-1,3-Dichloropropene	trans-3-Chloro-2-propene-1-ol	r0686	haloalkane dehalogenase
Dodecyl sulfate	Pseudomonas oleovorans	Lauric acid	Lauroyl-CoA	r0604	acyl-CoA synthetase
Dodecyl sulfate	Pseudomonas oleovorans	1-Dodecanol	Dodecanal	r0603	alcohol dehydrogenase
Dodecyl sulfate	Pseudomonas oleovorans	Dodecanal	Lauric acid	r0605	aldehyde dehydrogenase
n-Octane	Pseudomonas oleovorans	Octanoate	Octanoyl-CoA	r0024	acyl-CoA synthetase
n-Octane	Pseudomonas oleovorans	1-Octanol	1-Octanal	r0022	alcohol dehydrogenase
n-Octane	Pseudomonas oleovorans	1-Octanal	Octanoate	r0023	aldehyde dehydrogenase
n-Octane	Pseudomonas oleovorans	n-Octane	1-Octanol	r0021	alkane 1-monooxygenase
n-Octane	Pseudomonas oleovorans	Octane hydroperoxide	1-Octanol	r0684	alkyl hydroperoxide reductase
4-Chlorobiphenyl	Pseudomonas pseudoalcaligenes KF707	2-Hydroxy-6-oxo-6-(4'-chlorophenyl)-hexa-2,4-dienoate	cis-2-Hydroxypenta-2,4-dienoate and 4-Chlorobenzoate	r0136	2-hydroxy-6-oxo-6-phenylhexa-2
4-Chlorobiphenyl	Pseudomonas pseudoalcaligenes KF707	cis-2-Hydroxypenta-2,4-dienoate	4-Hydroxy-2-oxovalerate	r0101	2-oxopent-4-enoate hydratase
4-Chlorobiphenyl	Pseudomonas pseudoalcaligenes KF707	4-Hydroxy-2-oxovalerate	Pyruvate and Acetaldehyde	r0100	4-hydroxy-2-oxovalerate aldolas
4-Chlorobiphenyl	Pseudomonas pseudoalcaligenes KF707	4-Chlorobiphenyl	cis-2,3-Dihydro-2,3-dihydroxy-4'-chlorobiphenyl	r0133	biphenyl dioxygenase
Dibenzothiophene Degradation	Pseudomonas putida	cis-1,2-Dihydroxy-1,2-dihydrodibenzothiophene	1,2-Dihydroxydibenzothiophene	r0161	1,2-dihydroxy-1,2-dihydronapht

- Complete genome sequence로부터 pathway 정보를 기반으로 Biosynthetic gene cluster를 찾기 위해, EAWAG-BBD Database에 있는 정보(Pathway, Enzyme, Bacteria, Compound, Reaction, EC number)를 crawling하고 data를 frame에 맞춰 정형화함
- 위 Database의 한계로 enzyme sequence는 Database에 있지 않아서 따로 수집하는 과정을 진행하였음
- 확보한 enzyme sequence는 전체 351개, 추가로 약 1000여개 이상을 다른 Database에서 수집 중임
- 확보된 enzyme sequence에 대한 이름이 길고 복잡하여 EC number로도 정리하는 과정을 진행
- 구축된 데이터베이스 통계는 다음 그림에서 나타냄

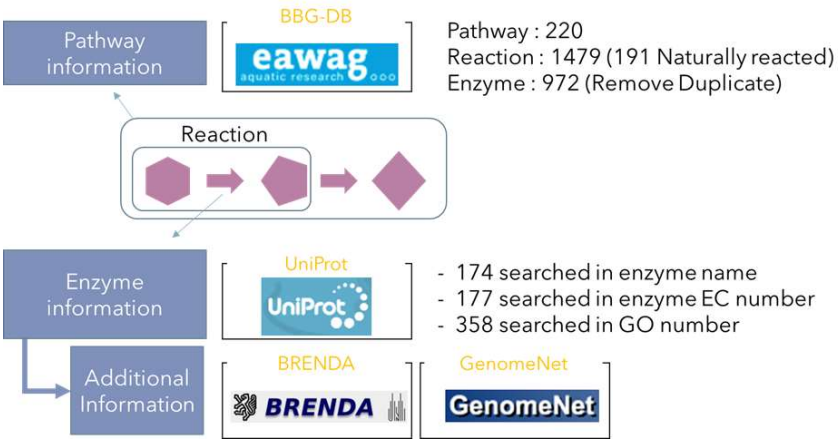


그림. 구축된 데이터베이스 통계

○ 기존연구 현황 분석

- 기존 pathway와 gene cluster 관련 연구에 대해 정리하였음 (표 3)
- 현재 진행하려고 하는 Bioremediation 관련 computational approach는 없으므로 현재 Bioinformatics에서 ML로 가장 많이 연구되고 있는 이차 대사 관련 pathway와 gene cluster 관련 연구에 대한 computational approach를 정리하였음

표. 기존연구 현황

논문	Target	Input	Output	Model
Geoffrey et al., (2019)	gene cluster	complete genome sequence	예측된 gene cluster	biLSTM
Abdur et al., (2020)	metabolic pathway	synthetic dataset. experimental Dataset	예측된 생합성과정	logistic regression
Alexander et al., (2020)	biosynthetic space	Protein	예측된 gene cluster	SVM classifiers
Baranwal et al., (2020)	metabolic pathway	1-radius subgraph 화학구조의 한 부분	예측된 생합성과정	Graph convolution
Rahman et al., (2021)	metabolic pathway	Heterogeneous information network	예측된 생합성과정	Unsupervised feature learning using Skip-gram
Allison et al., (2021)	Biosynthetic gene cluster	Matrix (BGC에 포함된 Protein)	새로운 BGC와 관련된 Protein	Machine learning (train classification algorithm)
Snorre et al., (2021)	metabolic pathway	BGC annotation	다시 설계된 생합성과정	- (Pipeline)

- Bioremediation 관련한 gene cluster, pathway는 보고가 많이 되어있지 않고, computational approach인 논문이 적게 보고되었음
- 2019년부터 2021년 보고된 gene cluster와 pathway 관련 예측 논문을 통해 관련된 data의 형태와 Input, Output data를 확인하였음

○ Pipeline 구축

- 전체 진행하고자 하는 pipeline을 시각화하였음 (그림)

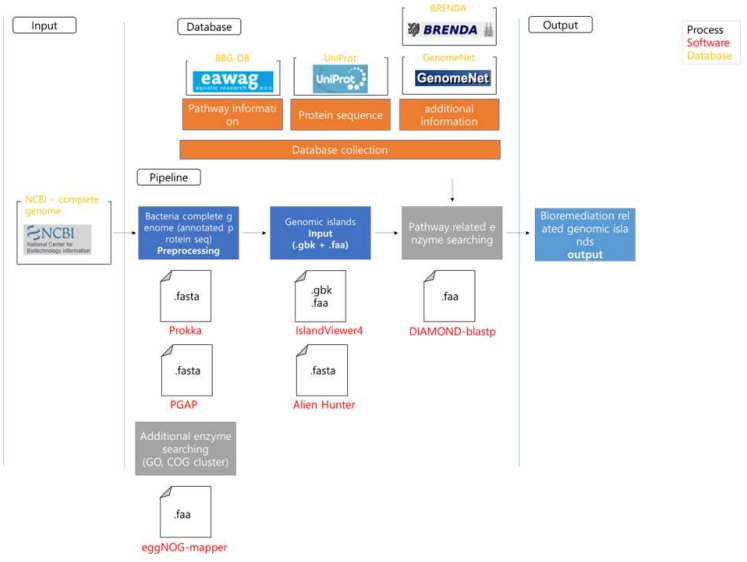


그림. 구축된 Pipeline

- 구축한 Database를 통하여 Bioremediation 관련 미생물을 선별하기 위하여 기존의 tools를 사용하였음
- Validation 할 때 사용한 균주는 기존 Database에 활성이 있다고 알려진 균주 중 NCBI에 complete genome sequence가 등록되어있는 9종을 선정하였음
- 분석에 앞서, complete genome sequence를 annotation 하기 위하여 prokka와 Bakta두 가지 tools를 활용하였음

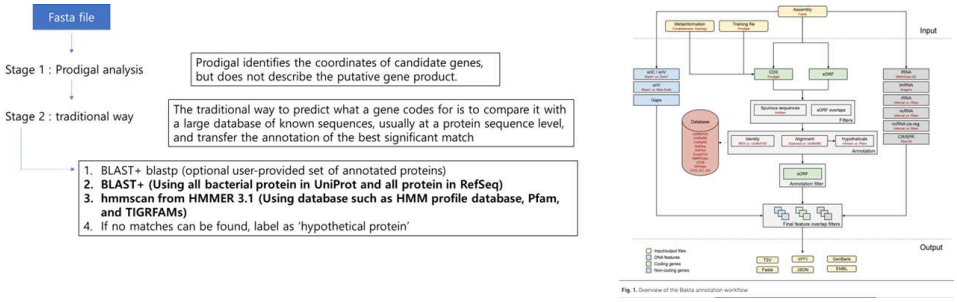


그림. 사용된 annotation tools 분석 방법

- 최근 보고된 Genomic islands분석 review논문을 통해 관련 프로그램을 선정하였음 시도한 프로그램을 표기함 (그림)

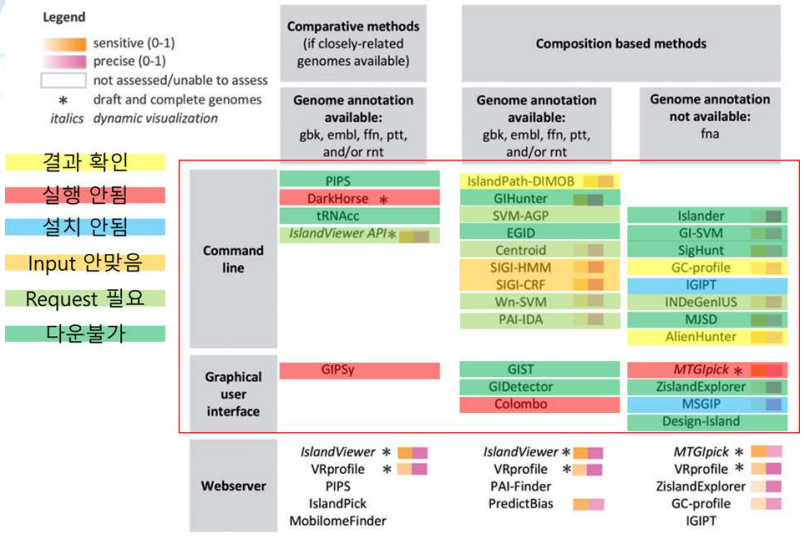


Figure 2. Summary of GI predictor characteristics with a multi-entry decision table. GI predictors were classified according to their interface and the requirement
Sensitivity (recall) and precision are color coded using a gradient from low-to-high as assessed using the comparative genomics data set. Methods that were n
assessed (comparative genomics) or that we were unable to assess are shown in white.

그림. 논문에서 보고된 분석 프로그램 리스트와 사용 시도 결과 요약

- Hydrocarbon 분해 Genomic islands 분석을 위해 Alien hunter와 IslandPath를 이용하여 분석하였음. 각 case study 별로 Genomic islands를 분석하여 표로 표기함 (표)

표. Genomic islands 분석 결과 (Case study)

Species	Pathway	Islandpath-DIMOB	Alien Hunter
Aromatoleum aromaticum EbN1	Ethylbenzene (anaerobic)	19	97
Bacillus sp. OxB-1	Limonene, Testosterone, Atropine	8	87
Deinococcus radiodurans R1	Limonene, Testosterone, Atropine, Menthol (anaerobic), Cyclohexane, Benzoate (anaerobic), alpha-Pinene, Ethylbenzene (anaerobic), Syringate, 2,4,6-Trinitrotoluene (anaerobic), Osmium Reduction, m-Cresol, m-Xylene	2	15
Escherichia coli C600	Limonene, Testosterone, Atropine, Menthol (anaerobic), Cyclohexane	19	66
Escherichia coli K-12	Limonene, Testosterone, Atropine, Menthol (anaerobic), Cyclohexane, Benzoate (anaerobic), alpha-Pinene, Ethylbenzene (anaerobic), Syringate, 2,4,6-Trinitrotoluene (anaerobic), Osmium Reduction, m-Cresol, m-Xylene, o-Xylene	15	73
Sulfurospirillum barnesii SES-3	Limonene, Testosterone, Atropine, Menthol (anaerobic)	9	86
Thauera aromatica K172	Limonene, Testosterone, Atropine, Menthol (anaerobic), Cyclohexane, Benzoate (anaerobic), alpha-Pinene, Ethylbenzene (anaerobic), Syringate, 2,4,6-Trinitrotoluene (anaerobic), Osmium Reduction, m-Cresol, m-Xylene, o-Xylene, p-Xylene, Toluene, Pyridoxine, Anthracene (fungal), Isoniazid	12	72

- annotation 된 complete genome sequence와 Database에 있는 enzyme sequence 유사도를 확인하기 위하여 DIAMOND-blast를 이용하였음 (표)

표. DIAMOND-blastp 결과

Gis_tool	Bacteria_ID Location	results	pident	length	mismatch	gapopen	qstart	qend	sstart	send	evalue	bitscore
IP	GCA_0000 KMBEICPM_04528|174910~176833	Q46444|E128	21.9	544	295	19	161	623	78	572	4.03E-12	65.5
IP	GCA_0002 FEIGHCAE_01056|1023659~1025084	O05204|E157	22.9	336	211	13	127	444	178	483	3.48E-05	42.4
IP	GCA_0008 CGKJGFMB_02885|2857868~2858819	Q9ZI56|E268	23.8	286	181	8	9	283	7	266	1.05E-17	79
IP	GCA_0008 CGKJGFMB_02885|2857868~2858819	A4IT42|E268	24.2	285	179	8	9	283	6	263	1.02E-17	79
IP	GCA_0002 FEIGHCAE_01056|1023659~1025084	P42974|E157	24.4	356	213	16	125	458	179	500	1.40E-04	40.4
IP	GCA_0002 FEIGHCAE_01050|1018816~1019530	E0X9C7|E214	24.6	118	86	2	15	130	453	569	1.95E-04	38.5
IP	GCA_0008 CGKJGFMB_02870|2839519~2840521	Q02972|E170	24.8	311	213	11	19	310	27	335	1.41E-24	99
IP	GCA_0008 CGKJGFMB_03247|3150297~3151893	Q9D0K2|E321	24.8	525	309	21	19	514	48	515	4.28E-23	99
IP	GCA_0008 CGKJGFMB_02875|2845960~2847145	Q9KWL3|E143	24.9	177	114	3	6	181	3	161	4.74E-11	59.7
IP	GCA_0008 CGKJGFMB_03243|3146079~3147045	Q9S7E4|E231	25.4	244	179	3	58	299	109	351	1.47E-28	110
IP	GCA_0030 CNNGEBIO_00405|414559~414949	E0X9C7|E068	25.6	117	85	2	7	123	452	566	3.28E-09	50.1
IP	GCA_0030 NBLBEJPE_03270|3429345~3429789	I7CA98|E214	26.5	113	80	1	30	142	113	222	2.59E-07	44.7
IP	GCA_0033 NEABLLGL_00887|933146~933590	I7CA98|E214	26.5	113	80	1	30	142	113	222	2.59E-07	44.7
IP	GCA_0098 DMAOBAEL_02638|2758563~2759007	I7CA98|E214	26.5	113	80	1	30	142	113	222	2.59E-07	44.7
IP	GCA_0002 FEIGHCAE_01057|1025070~1025433	P0A703|E240	26.8	112	82	0	1	112	1	112	3.94E-13	57.8
IP	GCA_0030 NBLBEJPE_00876|947014~947647	I7CA98|E214	26.8	149	103	2	53	201	74	216	1.26E-14	66.6
IP	GCA_0033 NEABLLGL_03215|3357015~3357648	I7CA98|E214	26.8	149	103	2	53	201	74	216	1.26E-14	66.6
IP	GCA_0098 DMAOBAEL_00615|642312~642945	I7CA98|E214	26.8	149	103	2	53	201	74	216	1.26E-14	66.6
IP	GCA_0008 CGKJGFMB_03246|3148549~3150283	T2G6Z9|E239	26.9	301	162	13	10	270	26	308	1.56E-07	50.4
IP	GCA_0000 KMBEICPM_04561|206903~208067	P07857|E331	27.3	388	252	12	3	370	10	387	2.41E-25	103
IP	GCA_0030 NBLBEJPE_00613|664106~665015	Q79EM8|E181	27.4	248	163	6	9	245	16	257	3.22E-13	65.5
IP	GCA_0033 NEABLLGL_03437|3583417~3584326	Q79EM8|E181	27.4	248	163	6	9	245	16	257	3.22E-13	65.5
IP	GCA_0008 CGKJGFMB_02861|2828579~2829008	O32472|E052	27.5	131	90	3	9	139	7	132	1.21E-11	55.1
IP	GCA_0000 KMBEICPM_04561|206903~208067	O62742|E331	27.7	382	247	12	9	370	16	388	6.02E-25	102

- Gene cluster mapping 결과와 기존 논문에 보고된 gene cluster와 비교하였음 (표)

표. Blast 결과를 통해 mapping 한 3-Phenylpropionate 예상 gene cluster

Query definition	Gene name	Query ID	E-value
BMIABBIL_00347 3-(3-hydroxy-phenyl)propionate/3-hydroxycinnamic acid hydroxylase	MhpA	Query_338	2.50E-18
BMIABBIL_00348 2,3-dihydroxyphenylpropionate/2,3-dihydroxicinnamic acid 1,2-dioxygenase	MhpB	Query_339	8.90E-183
BMIABBIL_00349 2-hydroxy-6-oxononadienedioate/2-hydroxy-6-oxononatrienedioate hydrolase	MhpC	Query_340	1.80E-89
BMIABBIL_00350 2-keto-4-pentenoate hydratase	MhpD	Query_341	4.80E-153
BMIABBIL_00351 Acetaldehyde dehydrogenase		Query_342	5.70E-137
BMIABBIL_00352 4-hydroxy-2-oxovalerate aldolase	MhpE	Query_343	9.20E-194

Blast결과와 모든 정보를 융합하여 Hydrocarbon 분해 Genomic islands를 예측하고 시각화 하는 과정을 진행중에 있음.

연구성과	학회 참여 논문 투고 진행중
프로젝트 기술분야	생물정보학, 프로그램 활용, 유전체 분석, 실험 설계 및 수행

프로젝트 결과보고서-(이우행)

<table>
<tr><td rowspan="5">파견개요</td><td>이름</td><td>이우행</td><td>대학</td><td>일반대학원</td></tr>
<tr><td>학과</td><td>생명공학과</td><td>세부전공</td><td>바이오빅데이터융합</td></tr>
<tr><td>파견국가
(도시)</td><td>미국 (Las Vegas)</td><td>파견기관명</td><td>University of Nevada,
Las Vegas</td></tr>
<tr><td>해외기관
지도인력</td><td>강민곤 교수
(Mingon Kang)</td><td>총연구기간
(파견기간)</td><td>210501~220831
(211223~220622)</td></tr>
<tr><td>참여
프로젝트명</td><td colspan="3">환경정화 관련 미생물 유전체 종합 분석</td></tr>
<tr><td rowspan="3">프로젝트결과</td><td>연구주제</td><td colspan="3">정화관련 효소군 분류</td></tr>
<tr><td>수행역할</td><td colspan="3">플라스틱 분해 가능 효소 데이터 수집 및 데이터세트 구축</td></tr>
<tr><td>연구수행
결과</td><td colspan="3">■ 정화관련 효소 데이터베이스 구축 및 특정 효소군을 위한 분류 기술 개발
- α/β-hydrolase계열 플라스틱 분해 가능 효소의 데이터베이스 구축 및 분류 기술 개발

1. 연구 목적
환경오염의 원인 중 하나인 고분자 폴리머 관련 분해 효소들의 유전자 데이터베이스를 구축하고 특정 효소의 기능에 따라 이를 분류할 수 있는 딥러닝 모델을 개발하고자 하였음.
특히, 전세계적으로 환경문제로 대두되고 플라스틱의 처리의 대안으로서 미생물을 이용한 처리에 대한 연구가 진행됨에 따라, 다양한 종류의 플라스틱의 생분해에 관여하는 단백질에 대한 정보를 데이터베이스화하고, 이를 활용하여 관련 유사단백질들을 이용한 딥러닝 학습으로 플라스틱 분해 가능성을 가진 다른 단백질들을 발굴하고 분해 가능 플라스틱류별로 분류하는 기술을 개발하고자 함.

2. 연구 내용
2.1. 추진 배경
지구 환경 문제에 큰 부분을 차지하고 있는 고분자 폴리머 중 하나인 플라스틱은 고분자 탄소원으로, 플라스틱 고유의 내구성으로 인해 환경에서 지속성이 있어 자연 분해가 어려운 물질임. 다양하고 폭넓은 분야에서 사용되고 있어 매년 그 수요가 증가하지만 부적절한 폐기물 관리로 인해 큰 문제로 대두되고 있음. 조사에 따르면, 이미 축적된 1억 5천만 톤에 이어 매년 9백만톤 이상의 플라스틱이 매립되거나 바다로 유입되고 있음.
플라스틱은 화학적으로 고분자량의 긴 탄화수소 사슬 폴리머로, 주로 석유화학제품에서 파생되어 긴 사슬형태의 폴리머로서 생산되어짐. 플라스틱 폐기물을 처리하는 방법은 매립, 소각, 재활용, 3가지 방법이 있으나, 매립 및 소각은 플라스틱의 고유의 내구성으로 인해 많은 문제를 야기하고 있어 꺼려지는 방법이고,</td></tr>
</table>

재활용 또한 일반적으로 선호되는 방법이지만 비용이 많이 들고 농업용 덮개 필름과 같이 특정 용도의 생분해성 또는 일부 열가소성 플라스틱에만 효과를 볼 수 있음.

한편, 한때 생분해되지 않는 것으로 생각되어왔던 플라스틱이 미생물에 의해 분해될 수 있음이 밝혀짐에 따라, 플라스틱 폐기물을 폴리머 합성의 기초 또는 발효를 위한 탄소원으로서 사용하고자 하는 시도가 증가하고 있음. 미생물에 의한 생분해가 점차 알려짐에 따라 플라스틱 분해 가능성이 있는 미생물의 분리, 식별 및 특성 분석이 수중환경, 폐기물 처리 매립지 또는 플라스틱 정제소와 같은 플라스틱과 직접 접촉이 많은 장소에서 자주 수행되고, 이에 따라 많은 문헌이 발표되고 있지만 기존의 접근 방식으로 자연환경 밖에서 미생물을 성장시키는 것은 매우 어렵고 배양 및 연구할 수 있는 분리종의 양은 1% 이하로 제한되는 있고 미생물의 유전체 시퀀싱 기술의 고도화에도 불구하고 아직 플라스틱 생분해에 대한 세밀한 메카니즘 및 유전자 발굴에 제한적인 데이터만 존재함. 플라스틱의 생분해와 연관되어 있는 많은 미생물들이 보고됨에도 그에 따른 유전체 연구와 대사 메커니즘에 대한 분석, 유전자에 대한 데이터가 미비함.

이에 따라, 플라스틱 생분해에 관련되어 있는 미생물의 정보 및 관련 유전자들, 생분해 메커니즘에 대한 정보를 수집하고 이를 접근 가능한 플랫폼 또는 데이터화하는 것이 이 분야의 연구에 필요하다고 판단함.

플라스틱류의 분해 메커니즘은 PETase와 같은 PET 분해효소의 메커니즘 이외에 자세히 알려진 바 없으나, 대체로 크게 가수분해 과정을 통해 ester기를 절단하여 중합체의 크기를 줄이는 반응과 산화환원 반응을 통해 첨가제를 분해하는 과정, 2가지로 분류가 가능하며, 특히 가수분해 효소 부분에서는 Proteases, Esterase, Glicosidase 등과 같은 α/β-hydrolase 계열 효소군이 주역할을 하는 것으로 알려져있어 산화환원효소를 제외하고 가수분해 효소 중 α/β-hydrolase 계열 효소군에 대한 플라스틱 분해 가능성 효소의 예측 및 분류 모델을 구성하고자 하였음 (그림 1).

2.2. 접근방법

2.2.1. 기존 데이터베이스

- Plastics Microbial Biodegradation Database (PMBD)

플라스틱 생분해에 관련된 데이터베이스는 연구시작단계에서 Plastics Microbial Biodegradation Database (PMBD, 그림 2)이 유일하였으며, 940여 가지의 플라스틱-미생물별 관련 정보, 플라스틱별 생분해가 검증된 79종의 유전자와 이를 바탕으로 TrEMBL 유전자데이터베이스에서 보충한 8,000여종의 유전자, Sequence alignment tool과 Convolutional neural network 기반의 예측 tool를 발표하였음.

플라스틱의 미생물 생분해에 대한 정보를 모으기 위해, 'Biodegradation', 'Bioremediation', 'Depolymerization', 'Enzyme' 등의 keyword 검색을 통해 National Center for Biotechnology Information (NCBI) 데이터베이스에서 플라스틱 생분해에 대한

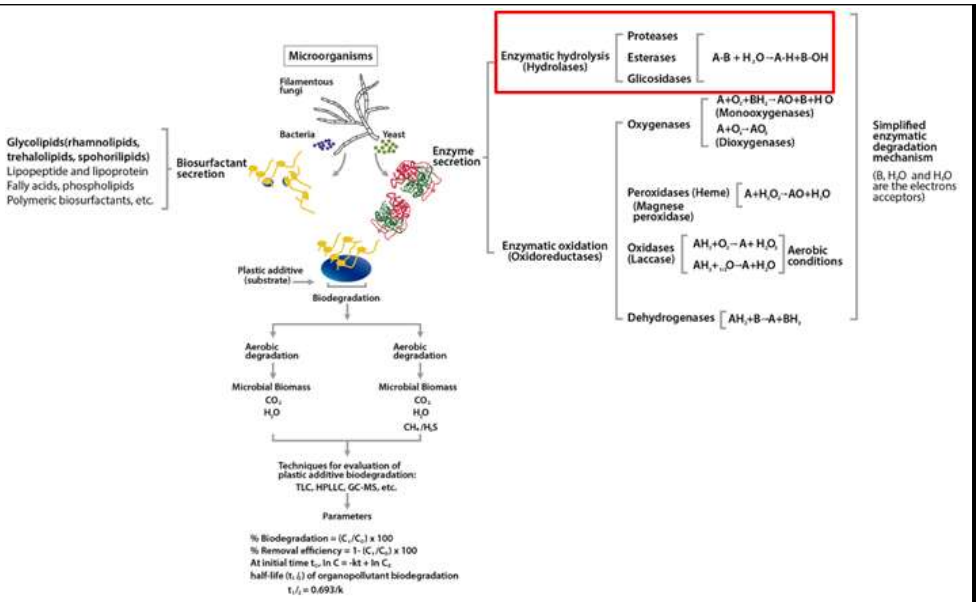


그림. 전반적인 플라스틱 분해 과정 및 관련 효소군 개요

정보를 검색하고 유용 정보들을 수동으로 추출하였음. 또한, UniProt 데이터베이스에서 자동으로 주석이 달린 효소 서열을 keyword를 이용해 수집하였음. 79종의 검증된 유전자들은 추후 연구에 큰 도움이 될 것으로 예상하고 수집하였지만, keyword를 이용해 수집된 8,000여 종의 유전자들은 실제로 플라스틱 생분해에 연관이 되어있는지에 대한 검증절차가 없었으며, 이 데이터를 이용한 CNN기반 tool 또한 false positive를 줄이는 부분에 대한 고찰이 없어 추후 연구에는 사용하지 않고자 하였음.

그림. PMBD 데이터베이스

- The Plastics-Active Enzyme database (PAZy)

PAZy는 연구과정 중 발표된 데이터베이스로서, 단백질에 더 초점이 맞추어져있는 데이터베이스임. 플라스틱이 자연적으로 잘 분해되지 않으며, 발효과정을 통해 직접 사용할 수 없다는 점을 바탕으로 미생물 및 효소 분해 이전에 기계적 처리, UV광에 의한 광분해 등이 선행됨에 따라 분해가 된다고 추측하는 것이 합리적이며, 첨가제로 사용되는 phthalate가 중합체보다 미생물의 생체 이용률이 더 높다는 점을 들어, 검증방법 중 하나인 weight loss 실험을 바탕으로 검증되어진 일부 유전자들에 대해 데이터

		첨삭함에 따라 60개 미만의 플라스틱 활성 효소에 대한 정보를 수집하고 데이터베이스화함. 이에 따라 데이터 수집 및 데이터세트를 만들 때 참고함. 2.2.2 데이터수집 데이터 수집 방법으로, 일부 논문이 차용한 것과 동일하게 NCBI 데이터베이스과 Google scholar 등의 검색엔진을 이용하여 keyword ('Biodegradation', 'Bioremediation' 등)을 바탕으로 논문을 수집하였으며, 플라스틱 분해 효소의 아미노산 서열 및 검증 결과들은 수동으로 추출하여 데이터화함. 기존에 알려져있는 PMBD 및 PAZy 내 플라스틱 분해 효소의 수가 매우 적어 데이터 확장이 필요함. 이에 따라, 시퀀스 유사성을 바탕으로 플라스틱을 분해할 수 있는 가능성이 높은 유사 유전자들을 수집할 필요성이 있음. 이에, PMBD 및 PAZy 내 데이터와 수동으로 추출한 플라스틱 분해 효소의 아미노산 서열들과 reference[4] 논문을 바탕으로, TrEMBL database에서 BLAST를 통해 유사 유전자들을 수집하였으며, 분해가능한 플라스틱별로 데이터화 실시하였으며, 동일 분해가능 플라스틱별로 Alignment tool로 잘 알려져있는 MUSCLE을 이용, Multialignment sequence file을 구성하였으며, 이를 바탕으로 Uniref30 시퀀스 데이터베이스에서 HMMER 및 HHblits을 이용하여 분해가능성 있는 유전자들을 확보하고자 하였음. 또한, false-positive 오류를 최소화하기 위해, 검증된 유전자들 서열과의 유사성, HMM에 대한 Score 및 E-value에 따라 정리하여 필요한 데이터를 수집하고자 하였음. 2.2.3 데이터세트 구성 2.2.3.1 전체 데이터 세트 전반적인 데이터 수집 방법을 그림3과 동일함. PMBD 및 PAZy 내 기재되어 있는 플라스틱 분해 효소의 아미노산 서열 이외에, 수집한 논문에 기재되어 있는 플라스틱 분해 효소의 아미노산 서열을 포함하여 검증된 데이터세트를 구성하며, 이를 바탕으로 BLAST, HMMER, HHblits에서 확인한 α/β-hydrolase 효소군를 predicted set으로 설정함. 동일 서열을 제거하기 위해 CD-HIT 프로그램을 이용하였음.

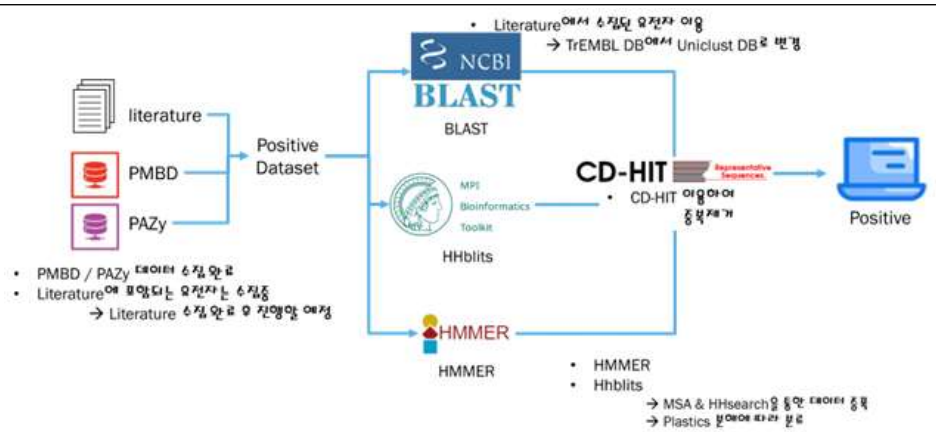


그림. 데이터세트 구성에 대한 전반적인 데이터 수집 방법

2.2.3.2 Positive dataset

분류 모델을 구성하기 위한 positive dataset은 1차적으로 논문 및 기존데이터베이스에서 수집한 검증된 효소 데이터세트와 이를 바탕으로 BLAST, HMMER, HHblits에서 수집된 서열들을 구성함.

동일 서열을 제거하기 위해 CD-HIT 프로그램을 이용하였음.

2.2.3.3 Negative dataset

Positive dataset과 명확하게 구분되면서 분해가능/불가능 효소군을 분류하기 위해, The Lipase Engineering Database (LED) 내에서 α/β-hydrolase core dataset을 수집하였으며, positive dataset과 동일성 검사를 통해 CD-HIT 프로그램을 이용하여 불필요한 서열을 제거하였음.

3. 연구 결과

PMBD 및 PAZy를 바탕으로 수집된 검증된 유전자 서열은 총 158종으로, 대부분이 Cutinase, lipase, carboxylic ester hydrolase, Dioxygenase, dehydrogenase 등이 포함되어 있었음. 특히, Cutinase, lipase, carboxylic ester hydrolase는 α/β-hydrolase 계열의 단백질들로서 따로 수집하였으며, dioxygenase나 dehydrogenase와 같은 산화환원효소들을 제외시켰음.

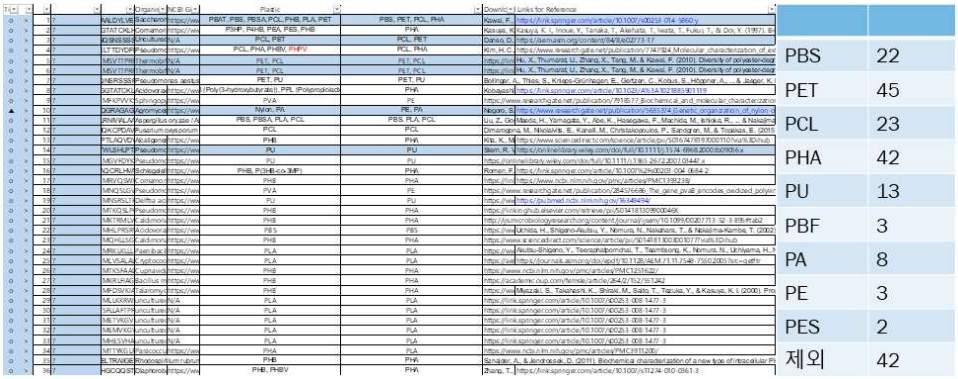
또한, 2021년 2월 14일을 기준으로 Google Scholar와 NCBI 등에서 keyword ("plastics", "bioremediation", "degradation", "enzyme" 등)를 통해 Review 논문 27편, Article 119편을 수집하였으며 수동으로 효소의 아미노산 서열을 수집하여 기존 데이터들과 통합한 결과, 총 189종의 유전자들을 확보할 수 있었음. 하지만, 수집된 유전자들은 oxygenase, peroxidase, dehydrogenase 등과 같이 산화환원반응 효소군을 제외하여 총 147종의 유전자를 확보할 수 있었음.

또한, 플라스틱 분해 가능 효소의 검증과정에서 사용된 기질들이 다양하여 이를 쉽게 분류하기 위해 구조적으로 유사한 중합체들로 구성된 플라스틱별로 정리하였으며, 이를 바탕으로 분해가능한 플라스틱별 효소군으로 분류하였음. 이에 따라,

	GenbankID	Genes names			UniProt links	AA_seq	Organism	Plastic	Download
1	AJ810119	bta1	BTA-hydrolase 1	Q6A0I4	https://www.un	MAVMTPR	Thermobifida fusca (Thermomonospora fusca)	PET	https://www
2	AJ810119	bta2	BTA-hydrolase 2	Q6A0I3	https://www.un	MAVMTPR	Thermobifida fusca (Thermomonospora fusca)	PET	https://www
3	K02640	CUT1	Cutinase 1	P00590	https://www.un	MKFFALTT	Fusarium vanettenii (Neocosmospora pisi)	PET	https://www
4	BBYR01000074	I3F6_4831	Poly(ethylene terephtha	A0A0K8F	https://www.un	MNFPRAS	Ideonella sakaiensis (strain NBRC 110686 / TISTR	PET, PEF,	https://www
5	AEV21261	Leaf-branch compost cutinase	LCC	G9BY57	https://www.un	MDGVLW	Unknown prokaryotic organism	PET, ester	https://www
6	EU004197	LIP	Lipase	O59952	https://www.un	MRSSLVLF	Thermomyces lanuginosus (Humicola lanuginosa	PCL	https://reac
7	XP_003656017.1	CutA	cutinase	G2RAE6	https://www.un	MKFLPILC	Thielavia terrestris CAU709 (Thielavia terrestris NF	PCL	https://reac
8	AAB05922.1	CutA	Cutinase	Q99174	https://www.un	MKFFALTT	Fusarium solani subsp. cucurbitae (Neocosmospo	PBS	https://reac
9			Alpha/Beta-hydrolase	N6VY44	https://www.un	MPLSMNK	Marinobacter nanhaticus D15-8W	PET	https://jour
10			Polyurethanase A	Q4KB56	https://www.un	MGVFDYK	Pseudomonas fluorescens (strain ATCC BAA-477	PU	https://www
11			Triacylglycerol lipase	P0DUB9	https://www.un	ADTYAAT	Pseudarthrobacter phenanthrenivorans (Arthrobac	PBS	https://reac
12			Poly(Hydroxyalkanoate)	Q46334	https://www.un	MRVQSW	Comamonas sp.	PHB	https://reac
13		phaZ	Poly(3-hydroxybutyrate	O24719	https://www.un	MRVQSW	Comamonas testosteroni (Pseudomonas testoster	PHB	https://www
14			Cutinase-like enzyme	S6BC01	https://www.un	MQFKSTF	Pseudozyma antarctica (Yeast) (Candida antarctic	BP films	https://www
15			Cutinase-like enzyme	Q874E9	https://www.un	MLVSALA	Cryptococcus sp. S-2	PLA	https://jour
									https://reac
									https://reac
16	AFUA_4G03560	phaZ	Carboxylic ester hydrol	Q4WBV8	https://www.un	MRGVWR	Neosartorya fumigata (strain ATCC MYA-4609 /	PHB, PES	https://reac
17	BAA07428.1	cutL	Cutinase 1	P52956	https://www.un	MHLRNIV	Aspergillus oryzae (strain ATCC 42149 / RIB 40)	PBS, PBSA	https://pub
									https://link
18	BAL22280.1	lipA	Lipase	G9M5R3	https://www.un	MFSGRFG	Aspergillus niger MTCC 2594	PLA, PCL	https://reac
19	4OYY		Cutinase	A0A075E	https://www.un	QLGAIEN	Humicola insolens (Soft-rot fungus)	PBTF, PBF	https://reac
			Cutinase 1	E9LVH8	https://www.un	MANPYER	Thermobifida cellulosilytica	PBTF, PBF	Enzymatic
20	CAA32819.1	PROK	Proteinase K	P06873	https://www.un	MRLSVLL	Parengyodontium album (Tritirachium album)	PBS, PLA	https://pdf.
21	QEO33334.1		Carboxylesterase	A0A5C1	https://www.un	MDGKQIN	Bacillus velezensis SYBC H47	DBP	https://www
22	AAZ54921.1	Tfu_0883	Triacylglycerol lipase	Q47RJ6	https://www.un	MAVMTPR	Thermobifida fusca (strain YX)	PET	https://pate
23	AAE13316.1	Sequence 2 from patent US 582	HIC	-	-	MKFFTTIL	Humicola insolens (Soft-rot fungus)	PET	https://che

그림. PMBD, PAZY 및 논문 수집을 통한 플라스틱 분해 효소 아미노산 서열 수집

Polybutylene succinate (PBS) 22종, Polyethylene terephthalate (PET) 45종, Polycaprolactone (PCL) 23종, Polyhydroxyalkanoate (PHA) 42종, Polyurethane (PU) 13종, Poly(1,4-butylene 2,5-furandicarboxylate) (PBF) 3종, Polyamide (PA) 8종, Polyethylene (PE) 3종, Polyethersulfone 2종으로 확인할 수 있었음.



PBS	22
PET	45
PCL	23
PHA	42
PU	13
PBF	3
PA	8
PE	3
PES	2
제외	42

그림. 분해가능한 플라스틱별 분해효소 정리

현재는 reference 4 논문을 바탕으로 HMMER 및 HHblits를 이용한 데이터세트를 증폭시키는 중이며, 위양성 문제를 해결하기 위해 적절한 value를 확인하는 작업을 진행중에 있음.

4. 향후 계획

현재, 검증된 플라스틱 분해 가능 유전자들의 phylogenetic tree 분석 및 아미노산 패턴 분석을 실시중에 있으며, HMMER 및 HHblits를 이용하여 UniRef30에서 유사 유전자들을 수집중에 있어 위양성 문제를 해결하는 중에 있음.

이를 바탕으로 Training set과 Validation set을 구성할 예정이며, CNN 및 BRNN 또는 Random-Forest 기법을 이용한 플라스틱 분해가능 효소의 발굴 및 분류 프로그램을 구성할 계획임.

5. Reference

[1] Gan Z, Zhang H. PMBD: a Comprehensive Plastics Microbial Biodegradation Database. Database (Oxford). 2019 Jan 1;2019:baz119. doi:

10.1093/database/baz119. PMID: 31738435; PMCID: PMC6859810.
[2] Buchholz PCF, Feuerriegel G, Zhang H, Perez-Garcia P, Nover LL, Chow J, Streit WR, Pleiss J. Plastics degradation by hydrolytic enzymes: The plastics-active enzymes database-PAZy. Proteins. 2022 Jul;90(7):1443-1456. doi: 10.1002/prot.26325. Epub 2022 Feb 25. PMID: 35175626.
[3] Gado JE, Harrison BE, Sandgren M, Ståhlberg J, Beckham GT, Payne CM. Machine learning reveals sequence-function relationships in family 7 glycoside hydrolases. J Biol Chem. 2021 Aug;297(2):100931. doi: 10.1016/j.jbc.2021.100931. Epub 2021 Jul 1. PMID: 34216620; PMCID: PMC8329511.
[4] Zrimec J, Kokina M, Jonasson S, Zorrilla F, Zelezniak A. Plastic-Degrading Potential across the Global Microbiome Correlates with Recent Pollution Trends. mBio. 2021 Oct 26;12(5):e0215521. doi: 10.1128/mBio.02155-21. Epub 2021 Oct 26. PMID: 34700384; PMCID: PMC8546865.

연구성과

구분	연구성과				특허				기술이전
	논문		학술대회		국제		국내		
	SCI(E)	비 SCI	국외	국내	출원	등록	출원	등록	
실적 (건)									

프로젝트 기술분야

Biodegradation, Bioremediation, Plastic degrading enzyme

프로젝트 결과보고서-(정경민)

<table>
<tr><td rowspan="6">파견개요</td><td>이름</td><td>정경민</td><td>대학</td><td>선문대학교</td></tr>
<tr><td>학과</td><td>컴퓨터융합전자공학과</td><td>세부전공</td><td>바이오빅데이터융합전공</td></tr>
<tr><td>파견국가
(도시)</td><td>U.S.A (Nevada)</td><td>파견기관명</td><td>University of Nevada, Las Vegas</td></tr>
<tr><td>해외기관
지도인력</td><td>Mingon Kang. Ph. D.,
Assistant Professor of
Computer Science</td><td>총연구기간
(파견기간)</td><td>210501~220831
(211223~220622)</td></tr>
<tr><td>참여
프로젝트명</td><td colspan="3">환경정화 관련 미생물 유전체 종합 분석</td></tr>
<tr><td rowspan="3">프로젝트결과</td><td>연구주제</td><td colspan="3">정화관련 효소군 분류</td></tr>
<tr><td>수행역할</td><td colspan="3">플라스틱 생분해 효소 데이터베이스 구축 및 분류 모델 개발</td></tr>
<tr><td>연구수행
결과</td><td colspan="3">1. 플라스틱 미생물 분해 효소에 대한 연구
- 기존 플라스틱의 자연 분해는 최소 20년이 걸림. 불에 태우는 경우 환경호르몬이 배출돼 환경을 위협하는 요소임.
- 생분해성 플라스틱은 특정 효소와 만나 저절로 썩어서 사라지는 플라스틱으로 가수분해가 가능한 구조를 가지고 있음.
- 현재 알려진 플라스틱 생분해 효소는 PA, PBF, PBS, PCL, PE, PES, PET, PHA, PU 등 다양하게 존재함
- 하지만 플라스틱 생분해 효소에 대한 밝혀진 패턴은 적음. 또한 플라스틱 생분해 효소는 별개의 패턴을 가지고있음.
2. 플라스틱 미생물 분해 효소에 대한 기존의 연구
- Plastic Microbial Biodegradation Database (PMBD)는 기존에 알려진 플라스틱 생분해 효소 데이터와 딥러닝 모델을 결합하여 분류 모델 및 데이터베이스 구축 함
- 논문에서 실험을 통해 밝혀진 플라스틱 미생물 분해 효소 데이터를 수집하였음.
- 하지만 데이터 부족 문제로 인하여 수집된 데이터를 기준 데이터로 사용하여, Hidden Markov Model(HMM) 기반의 생명공학 툴인 HMMER을 통해 데이터를 증폭함.
- 딥러닝 기법중 하나인 CNN을 통한 7개의 클래스에 대한 분류를 진행 하였으며, 데이터베이스, 분류 모델을 활용하여 웹사이트로 제공하였음.</td></tr>
</table>

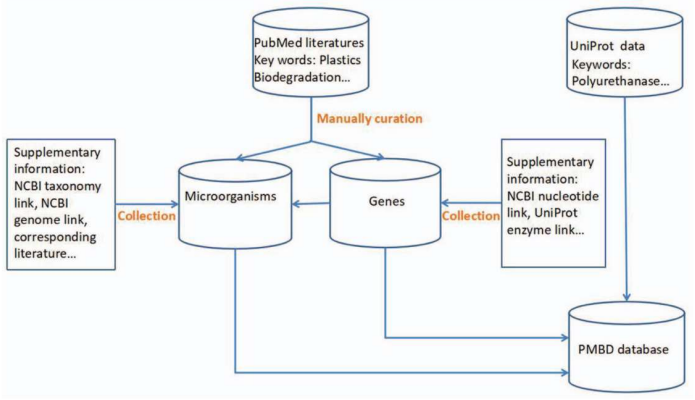


그림. PMBD에서 데이터 수집의 Flow chart.

- Gan, Zhiqiang, and Houjin Zhang. “PMBD: a comprehensive plastics microbial biodegradation database.“ Database 2019 (2019).

1.1. PMBD의 데이터 위양성 문제

- 생명공학에서 HMM을 통한 데이터 증폭은 잔기들의 빈도수를 통해 확률을 계산하고 각 위치별 잔기에 따른 생성될 확률을 통해 Protein Sequence를 생성함
- 하지만 이렇게 생성된 데이터는 실존하지 않고, 만약 Machine Learning, Deep Learning 에 사용하는 경우 부정확한 데이터로 인해 모델의 일반성이 떨어짐.

3. 플라스틱 생분해 데이터 수집

- 플라스틱 생분해 분류 모델 개발을 위해 Positive Data, Negative Data를 각각 다른 방식으로 수집하였음

1.2. Positive Data

- 데이터 위양성 문제 해결을 위해 그림15와 같은 과정을 진행하였음
- 플라스틱 생분해 효소에 대한 정확한 데이터 수집을 위해 실험 바탕의 논문에서 증명된 효소를 수집하였음.
- 수집된 효소는 데이터 부족 문제로 인해 BLAST, HHBlits, HMMER 세 개의 Tools을 통해 데이터를 증폭하였음.
- 3가지 Tools을 사용하여 증폭된 데이터는 유사성 80% 이상의 데이터로만 구성하였음.
- 중복되는 데이터가 있을 경우 분류 모델의 과적합이 생길 수 있기 때문에 CD-HIT를 통해 100% 일치하는 데이터는 중복 제거를 하였음.
- 최종적으로 수집된 데이터는 논문에서 수집된 효소데이터와 100% 일치하는 데이터를 삭제하여 구성 하였음.
- 논문에서 수집된 데이터는 분류모델의 일반화 확인을 위해 사용하였고, 그외 나머지 데이터는 분류모델의 학습, 검증, 평가 데이터로 사용하였음.

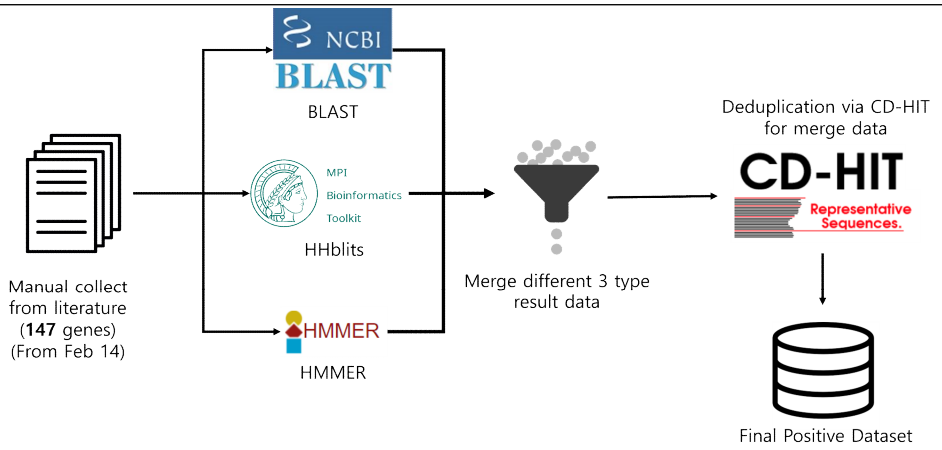


그림. 플라스틱 생분해 효소 데이터 수집에 대한 전체 흐름도

1.3. Negative Data

- The Lipase Engineering Database(LED)의 Core 데이터를 사용하여 Negative Dataset을 구성하였음.
- Positive Dataset과 100퍼센트 일치하는 데이터를 제거한 후 CD-HIT를 통해 100퍼센트 일치하는 데이터는 중복제거 하였음.
- 추가적으로 논문에서 수집된 효소 데이터와 100% 일치하는 데이터 삭제를 통해 최종 Negative Dataset을 생성하였음.

4. 플라스틱 생분해 효소 분류 모델

1.4. 플라스틱 생분해 효소를 위한 데이터 전처리 연구

- 플라스틱 생분해 효소와 다르게 패턴이 알려진 데이터는 기존 One-Hot-Encoding을 통해 데이터 전처리를 진행함
- 하지만 플라스틱 생분해 효소의 경우 알려진 패턴이 적기 때문에 전처리 과정에서 더 많은 정보를 담아야하는 문제 발생
- 따라서 Jing, Xiaoyang가 발표한 논문, Xu, Yuting 가 발표한 논문, 두 개의 논문을 참고하여 Amino Acid 의 Encoding 방식을 참고하여 19개의 Encoding 방법을 개발하였음.
- Jing, Xiaoyang, et al. “Amino acid encoding methods for protein sequences: a comprehensive review and assessment.“ IEEE/ACM transactions on computational biology and bioinformatics 17.6 (2019): 1918-1931.
- Xu, Yuting, et al. “Deep dive into machine learning models for protein engineering.“ Journal of chemical information and modeling 60.6 (2020): 2773-2790.

1.5. 플라스틱 생분해 효소를 위한 분류 모델 연구

- 플라스틱 생분해 효소 분류 모델의 경우 Jing, Xiaoyang가 발표한 논문에서 사용한 BRNN, RandomForest와 DeepEC에서 사용한 CNN을 구조를 사용하여 초기 분류 모델로 구축하였음.
- 플라스틱 생분해 효소 모델은 총 2단계로 구성하였음. 1단계 모델은 플라스틱 생분

해 효소 여부에 따른 이진분류, 2단계 모델은 플라스틱 생분해 효소별 분류를 하는 다중 분류 모델로 사용자가 입력한 데이터를 분류함.

- 1,2단계 모델에서는 CNN, BRNN, RamdomForest 모델을 사용하여 연구를 진행함.
- 플라스틱 생분해 효소 분류 모델은 Protein Sequence를 활용하여 분해 효소를 분류하는 모델로 Embedding Layer, Classification Model, 2가지의 구성으로 이루어져 있음.
- Embedding Layer의 경우 Protein Sequence의 데이터를 전처리 하는 과정으로, 앞서 설명한 19가지의 Encoding 방법을 통해 Protein Sequence를 Matrix로 변환하는 과정이다.
- Classitifcation Model의 경우 총 3가지 방법을 사용하였는데, RandomForest, BRNN의 경우 논문에서 사용한 Hyperparameter를 동일 하게 사용하여 구현 하였고, CNN의 경우 DeepEC와 동일한 구조를 가지고 사용하였으나, Encoding 방법에 따라 Convolutional Layer 값을 다르게 사용하였음.
- Wang, Han, et al. “DeepEC: An error correction framework for dose prediction and organ segmentation using deep neural networks.“ International Journal of Intelligent Systems 35.12 (2020): 1987-2008.

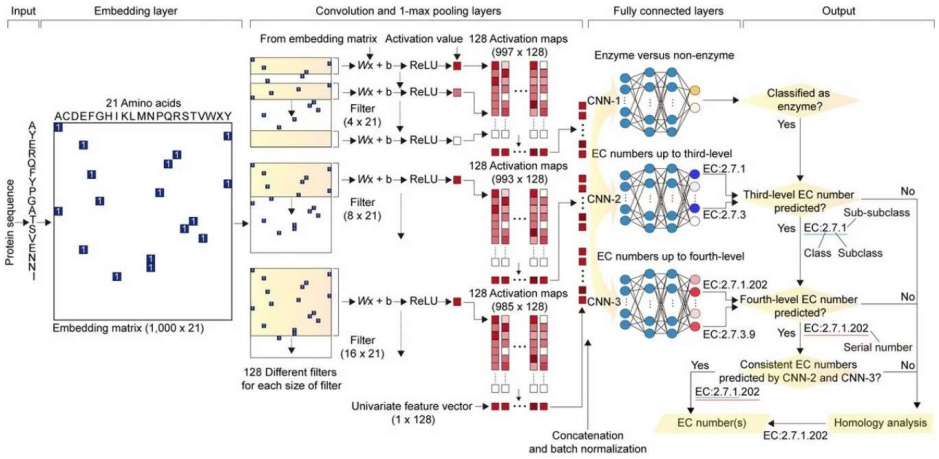


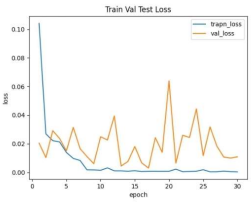
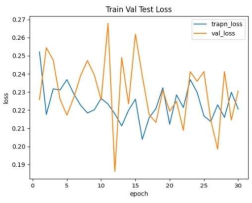
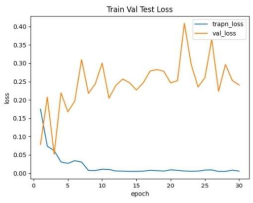
그림. DeepEC의 전체 흐름도

5. 실험 및 결과

B / 1e-4		zScales	VHSE	PCscores	sScales	One Hot Encoding	Five Bit Encoding	Six Bit Ecoding	Hydrohobicity matrix	Meiler Parameter	Atchley Factors	PAM 250	Blosum 62	Cristian	Tanaka Sheraga	Miyazawa Energies	Michaelis Potentials	AESNN3	ANNAD
CNN	AUC	0.994084	0.994513	0.50	0.50	0.996227	0.948641	0.979336	0.910915	0.951816	0.956443	**0.996828**	0.995370	0.911601	0.691675	0.988082	0.900797	0.923776	0.985510
	w/o F1 Score	**0.873684**	0.811111	0.0	0.0	0.122807	0.851064	0.534247	0.471429	0.139130	0.851064	0.804469	0.824176	0.679012	0.0	0.544218	0.106195	0.853211	0.636943
	w/ F1 Score	**0.886422**	0.837007	0.333333	0.333333	0.402168	0.867283	0.646556	0.607242	0.411418	0.867283	0.831957	0.847047	0.741762	0.333333	0.652892	0.392780	0.851967	0.713305
BRNN	AUC	0.845066	0.747664	0.50	0.50	0.808197	0.779731	0.739347	0.773386	0.756409	0.770471	0.837863	0.873875	0.748521	0.828175	**0.859642**	0.806825	0.789505	0.804939
	w/o F1 Score	0.628205	**0.679012**	0.0	0.0	0.606452	0.569536	0.628205	0.513889	0.482270	0.619355	0.592105	0.592105	0.578947	0.601307	0.610390	0.632911	0.50	0.402958
	w/ F1 Score	0.707485	**0.741762**	0.333333	0.333333	0.691581	0.667515	0.707485	0.633705	0.613957	0.701619	0.683734	0.683734	0.673608	0.689745	0.695706	0.709126	0.623312	0.565438

그림. 플라스틱 생분해 효소 모델에 대해서 107 Genes에 대한 CNN, BRNN에 대한 결과

- CNN, BRNN의 Grid Search 를 통하여 Batch size = 8, 16, 32, Learning Rate = 1e-4, 1e-5, 1e-6 으로 설정 하였고, 데이터는 학습:검증:평가 8:1:1로 구분 하여 학습하였음.
- 모델의 결과에서 알 수 있듯이 각 Encoding 에 따라 결과가 다른것을 확인 할 수 있으며, 생물학적인 정보를 많이 담고 있는 PAM250, Blosum62와 같은 Encoding이 결과가 좋은 것을 볼 수 있다.

		- Learning Curve의 경우 학습되는 값들에 따라 값들이 달라지는 것을 볼수 있으며, 결과를 통해 알아 봤을때 Learning rate는 1e-5, Batch size는 16이 적당한것으로 확인 되었음.  zScales  PCScores  Meiler parameters 그림. Learning rate 1e-5, Batch size 16의 Hyperparameter에 대한 CNN의 Learning Curve
	연구성과	*논문 투고 진행중*
	프로젝트 기술분야	정화 관련 효소 데이터베이스 구축 및 특정 효소군을 위한 분류 기술개발

프로젝트 결과보고서-(박주연)

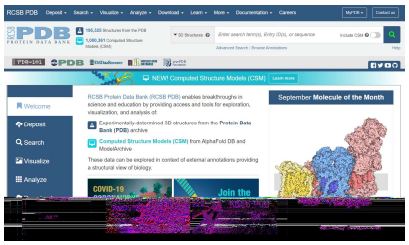
<table>
<tr><td rowspan="5">파견개요</td><td>이름</td><td>박주연</td><td>대학</td><td>선문대학교</td></tr>
<tr><td>학과</td><td>컴퓨터융합전자공학과</td><td>세부전공</td><td>바이오빅데이터융합전공</td></tr>
<tr><td>파견국가
(도시)</td><td>U.S.A (Nevada)</td><td>파견기관명</td><td>University of Nevada, Las Vegas</td></tr>
<tr><td>해외기관
지도인력</td><td>Mingon Kang. Ph. D.,
Assistant Professor of
Computer Science</td><td>총연구기간
(파견기간)</td><td>210501~220831
(210826~220224)</td></tr>
<tr><td>참여
프로젝트명</td><td colspan="3">환경정화 관련 미생물 유전체 종합 분석</td></tr>
<tr><td rowspan="3">프로젝트결과</td><td>연구주제</td><td colspan="3">정화관련 효소군 분류</td></tr>
<tr><td>수행역할</td><td colspan="3">데이터 수집 및 전처리, 딥러닝을 통한 정화관련 효소와 기질 간의 상호작용 예측 연구</td></tr>
<tr><td>연구수행
결과</td><td colspan="3">• 데이터 수집 - Cytochrome P450 단백질에 대한 시퀀스 정보와 해당 단백질과 상호작용하는 기질에 대한 InChI, InChIKeym SMILES 등에 대한 정보가 필요함. 이를 얻기 위해 다양한 Protein, Compound 관련 데이터베이스를 조사, 분석 및 수집함.

- PDB(Protein Data Bank) : PDB는 PDB 아카이브에서 실험적으로 결정된 3D 구조 및 AlphaFold DB 및 ModelArchive의 CSM(Computed Structure Models)에 대한 탐색, 시각화 및 분석을 위한 액세스 및 도구를 제공하여 과학 및 교육의 혁신을 가능하게 하는 데이터베이스임. PDB를 활용하면 생물학의 구조적 관점을 제공하는 외부 주석의 맥락에서 탐색할 수 있기 때문에 실험적으로 검증된 CYP 효소 정보를 얻기에 적절함.

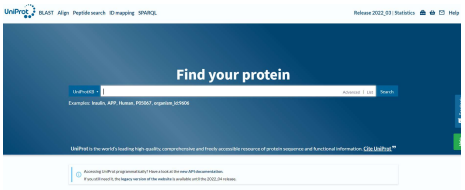
- UniProt : UniProt은 단백질 서열 및 기능 정보를 제공하는 세계 최고의 고품질 종합 리소스로, 실험적으로 검증 된 Swiss-Prot을 보유하는 데이터베이스임. 단백질 서열과 Interaction을 보이는 Compound에 대한 정보가 같이 있기 때문에 CYP 효소를 키워드로 수집하기에 적절한 데이터베이스임. 그리고, TrEMBL은 UniProt에서 파생된 데이터베이스로, Swiss-Prot을 기반으로 Computer-annotated 단백질 서열 데이터베이스임. 이는 실험적 데이터를 기반으로 한 컴퓨터방법론을 통해 자동화된 주석이 달린 데이터기 때문에 후처리를 통해 일정 similarity 값을 정해 데이터를 추출하기에 적합함.

- PubChem : PubChem은 자유롭게 접근할 수 있는 화학 정보를 담은 세계 최대 규모의 데이터베이스임. 이름, 분자식, 구조 및 기타 식별자들로 화학물질을 검색할 수 있고, 화학적 및 물리적 특성, 생물학적 활성, 안전 및 독성 정보, 특허, 문헌 인용 등의 정보를 알 수 있기 때문에 세부화된 검색 조건을 통해 CYP 효소를 수집하기에 적절함.</td></tr>
</table>

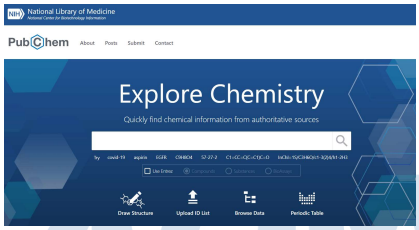
- ChEMBL : ChEMBL은 약물과 유사한 특성을 가진 생체 활성 분자의 수동 큐레이트 데이터베이스임. Genome 정보를 효과적인 신약으로 번역하는데 도움이 되도록 화학적, 생물학적 활성 및 Genome 데이터를 함께 제공하기 때문에 각 sequence나 compound 데이터를 통합하는 InChI, InChIKey,SMILES 등의 정보를 얻기에 적절함.



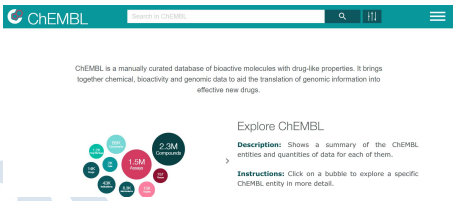
PDB



UniProt



PubChem



ChEMBL

그림 1. CYP데이터 수집을 위해 사용된 데이터베이스

- 데이터 전처리

- 중복 데이터 제거

먼저, 수집된 데이터들에 대한 효과적인 처리를 위해 Python 내장 라이브러리인 Pandas를 사용하였음. Pandas는 Python 프로그래밍 언어를 기반으로 구축된 빠르고 강력하며 유연하고 사용하기 쉬운 오픈 소스 데이터 분석 및 조작 도구임. 이를 활용하여 수집한 데이터들을 DataFrame으로 변환하고 중복 데이터를 처리함. 그 방법으로 Pandas 내장 메소드인 DataFrame.duplicated(), DataFrame.drop_duplicates()를 활용함.

- Human Negative data 추가

본 과제를 진행하면서 지금까지 겪고 있는 가장 큰 문제는 Cytochrome P450 단백질과 연관된 기질 정보는 있지만, 특정 단백질과 특정 기질은 연관되지 않고, 확실히 결합하지 않는다는 내용의 데이터는 찾지 못함. 이에 관련하여 우리가 타겟으로 잡고 있는 Bacteria 데이터가 아닌 Human 데이터를 집중하여 보았고, Human 데이터베이스에는 binding은 하지만 Activation은 없다고 실험적으로 검증된 Human Inactivate 데이터를 찾을 수 있었음 (PubChem, Cytochrome panel assay with activity outcomes.). 따라서 Bacteria negative data를 대체할 수 있다는 가정을 세워 Human negative 데이터를 수집함. 이에 활용 가능성이 보일 경우 추후 예측 모델의 학습 데이터로 사용할 계획임.

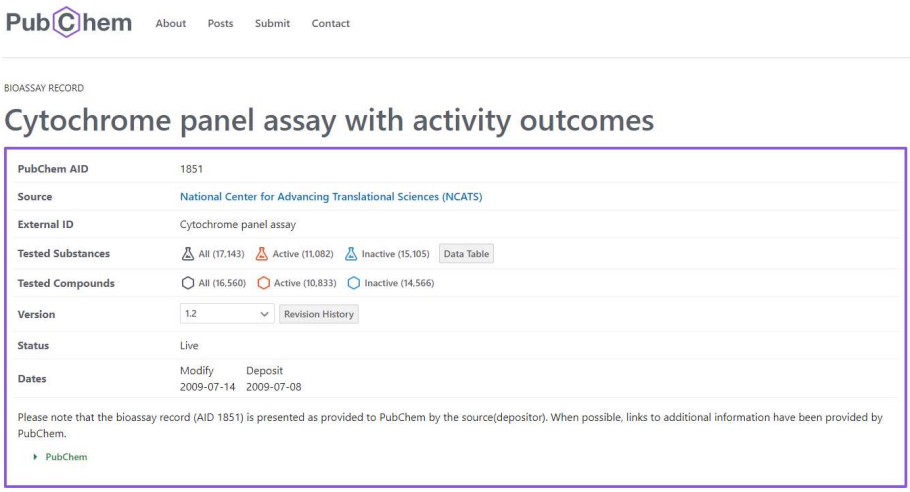


그림 2 PubChem, Cytochrome panel assay with activity outcomes

- Compound < - > Protein을 나타내는 DataFrame 생성

수집한 데이터를 전처리하는 단계를 거쳐 효과적인 데이터저장과 데이터베이스 구축을 위해 DataFrame을 통해 정형화 함. DataFrame의 Colum 으로는 각 데이터를 구분할 수 있는 ID, 각 CYP에 대한 Sequence, 해당 Sequence를 가지는 단백질에 대한 기질을 나타내는 값으로 InChI, InChIKey, 그리고 SMILES 값을 기준으로 한 DataFrame을 생성하였다.

- Compound-Protein Interaction 예측 모델 연구

- Title: Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences.

Objective: 본 논문의 목표는 화합물 및 단백질에 대한 end-to-end representation 학습을 수행하고 representation을 통합하여 화합물에 대해서는 GNN을, 단백질에 대해서는 CNN을 사용하여 Protein Compound Interaction 예측을 개발함.

- Title: A multi-objective neural network for predicting compound-protein interactions and affinities.

Objective: 본 논문에서는 가상 스크리닝에 사용되는 분자 도킹 및 분자 역학은 단백질과 화합물 간의 결합 친화도를 찾는 데 도움이 되지만 3D 구조화된 데이터에 의존한다는 한 가지 제한 사항이 있으므로 본 논문에서는 이러한 한계를 극복하고 그래프만 취하는 구조 없는 모델을 생성하고자 함. 그래서 화합물 및 단백질의 1차 서열을 입력으로 표현하고 CPI 및 친화도를 예측함.

- Title: BACPI: a bi-directional attention neural network for compound-protein interaction and binding affinity prediction.

Objective: 본 논문에서는 CPI 예측으로 단백질과 화합물 간의 결합도를 찾는 것은

단순한 이진 분류 문제가 아니라 연속 값이라고 말함. 본 논문에서 제안하는 BACPI는 이진 분류 문제인 CPI 상호작용을 예측하고 Regression 문제인 결합 활동도 예측함.

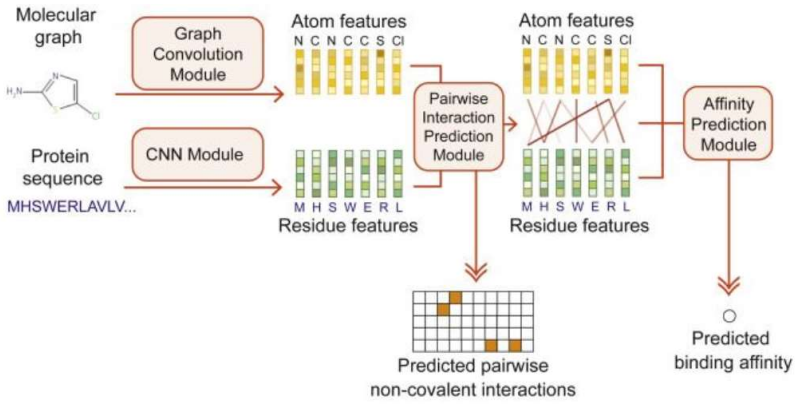


그림 3. BACPI Workfolw

- Title: DeepCPI: A deep learning-based framework for large-scale in silico drug screening. Genomics.

Objective: 본 논문에서는 현재 이용 가능한 대규모 비표지 화합물 및 단백질 데이터로부터 잠재된 특징을 탐색하기를 원하고(기존 방법은 라벨이 붙은 데이터로부터 특징의 단순하고 직접적인 표현을 사용하고 미지의 CPI를 추론하기 위해 이를 사용함) CPI 예측을 위한 강력한 딥 러닝 기반의 특징 임베딩을 사용하려고 함.

- Compound-Protein Interaction 예측 모델 아이디어

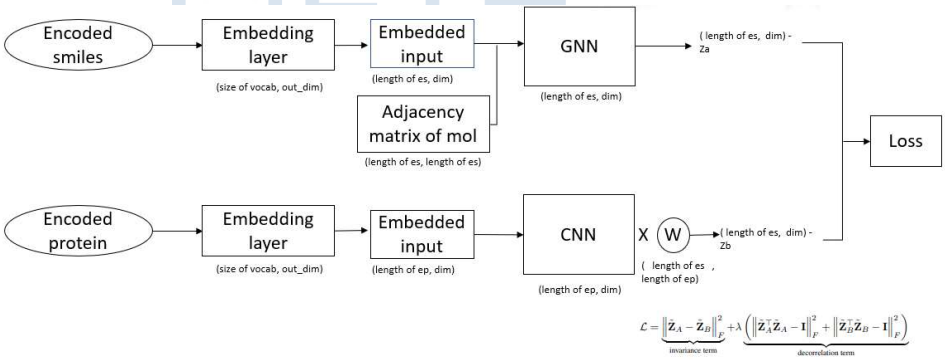


그림 4. Compound-Protein Interaction 예측 모델 아이디어 시각화

- 본 과제에서는 Compound-Protein Interaction 예측을 위한 방법으로 Compound와 Protein을 따로 임베딩하여 학습하고 Feature을 추출하여 통합한 값으로 예측 하는 방법을 제안함. 여기서 Compound는 GNN, Protein은 CNN으로 학습함. Compound는 결국 3D구조로 이루어져있기 때문에 이 특성을 잘 나타낼 수 있는 데이터 형태는 Graph 형태가 있음. Graph는 관계나 상호작용과 같은 추상적인 개념을 다루기에 적합하며 복잡한 문제를 간단한 표현으로 단순화 하기에 용이함. Graph structure을 matrix로 표현하는 방법으로는 Adjacency matrix와 Feature Matrix가 있는데 graph

		의 각 node와 edge 값으로 compound를 Embedding 할 수 있음. $$Adjacency\ matrix\ A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$ $$Feature\ matrix\ F = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$ 그림 5. Adjacency matrix와 Feature Matrix - Embedding 된 데이터를 가지고 학습을 할 때 는 GNN(Graph Neural Network)을 활용함. GNN은 그래프에 직접 적용할 수 있는 신경망으로, 점 레벨, 선 레벨, 그래프 레벨에서의 예측 작업에 쓰임. GNN의 핵심은 점이 이웃과의 연결에 의해 정의된다는 것인데, 이를 염두하면 GNN이 Compound 데이터를 다루는데 적합하다는 것을 알 수 있음. 따라서 이를 통한 Compound 예측 모델을 구현할 계획임. - 그리고 Protein을 예측 하는 방법으로는 CNN을 활용할 계획임. Protein 데이터 같은 경우에는 Sequence로 이루어져 있기 때문에 Graph데이터를 처리하는 방법처럼 처리할 수 없음. 이에 Sequence를 Image matrix처럼 사용한 전례들을 활용하여 Sequence를 One-hot encoding을 통해 embedding 하여 처리하려고 함. Sequence 데이터는 문자열 데이터로 자칫 RNN이나 LSTM 같은 언어처리 알고리즘을 생각하기 쉽지만, 실제 sequence는 자연어 처리에 적합하지 않다고 생각할 정도로 sequence 자체의 특징을 찾기 힘듬. CNN(Convolution Neural Network)은 앞서 말한대로 Image 처리에 효과적이지만 이를 활용하기 위해 Sequence를 Image matrix로 변환함. 이에 선례를 따라 본 과제에서도 CNN을 통한 Protein sequence 처리를 할 계획임. - 이렇게 각 Embedding 방법과 처리 Alogorithm을 통해 추출된 feature map을 가지고 threshold 값을 조정하여 Compound-Protein Interaction 예측을 위한 모델을 개발할 계획임.
	연구성과	논문 투고 예정
	프로젝트 기술분야	Data processing, Data analysis, Deep learning, Compound-Protein Interaction

프로젝트 결과보고서-(황지섭)

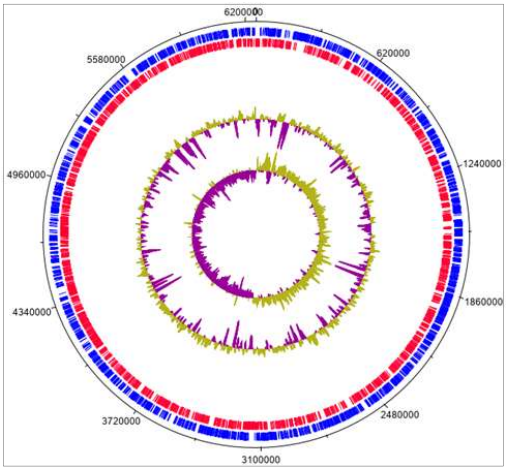
<table>
<tr><td rowspan="5">파견개요</td><td>이름</td><td>황지섭</td><td>대학</td><td>과학기술연합대학원대학교
(극지연구소)</td></tr>
<tr><td>학과</td><td>극지과학</td><td>세부전공</td><td>생화학, 구조생물학</td></tr>
<tr><td>파견국가
(도시)</td><td>U.S.A (Nevada)</td><td>파견기관명</td><td>University of Nevada
,Las Vegas</td></tr>
<tr><td>해외기관
지도인력</td><td>강민곤</td><td>총연구기간
(파견기간)</td><td>210501~220430
(211001~220429)
원격(211001~211212)</td></tr>
<tr><td>참여
프로젝트명</td><td colspan="3">극한환경 미생물 비교연구</td></tr>
<tr><td rowspan="3">프로젝트결과</td><td>연구주제</td><td colspan="3">청정지역과 오염지역의 미생물 유전체 패턴 분석</td></tr>
<tr><td>수행역할</td><td colspan="3">극한지 미생물 유전체 분석, 항생제 오염에 대한 청정지역과 오염지역의
미생물 유전체 패턴분석</td></tr>
<tr><td>연구수행
결과</td><td colspan="3">1. 청정지역인 남극 토양에 서식하는 미생물 분리 동정 및 전장 유전체 분석 (Whole genome sequencing, WGS)
◦ 남극대륙 바톤반도 내 남극세종과학기지 인근에서 채집된 남극이끼 (Sanionia uncinata) 가 서식하는 지역의 rhizosphere에 해당하는 층의 토양으로부터 미생물을 분리 동정하였음.
◦ 남극 토양에서 분리한 미생물의 경우, 16s rRNA 마커 유전자를 활용하여 종 동정을 진행하였고, Pseudomonas fluorescens CCM 2115 strain 으로 동정이 완료됨.
◦ 분리한 균주에 Pseudomonas fluorescens Ant01 strain number를 부여함.
◦ 청정지역에서 분리한 미생물의 유전체 패턴을 확인하기 위해 전장 유전체 분석을 진행하고, 미국 국립생물공학정보센터 (National Center for Biotechnology Information, NCBI) 에 전장유전체 정보를 등록 중에 있음.
◦ Rapid prokaryotic genome annotation (Prokka)와 Gene ontology annotation in InterPro tool (Functional Domatin Prediction)을 사용하여 annotation을 진행함 (CDS: 5,616 개, tRNA : 69 개, rRNA : 19 개
 
</td></tr>
</table>

Figure 1. Circular map of *Pseudomonas fluorescens* Ant01 isolated from Antaractic rhizosphere

2. *Pseudomonas fluorescens* 균주에 대한 계통분석

◦ *Pseudomonas* sp. 에 속하는 균주들은 크게 *P. aeruginosa*, *P. stutzeri*, *P. putida*, *P. syringae*, *P. fluorescens* complex 로 분류되고, 각 complex 마다 subgroup 으로 분류된다는 특징을 갖고 있음. *P. fluorescens* complex 내에도 *P. fluorescens* subgroup을 포함하여 15종 이상의 학명을 갖는 균주로 분류가 됨.

◦ 따라서, 16s rRNA 만으로는 정확한 종 동정을 하는데 어려움이 있음. 최근 연구에서, 다수의 Housekeeping gene을 마커유전자로 사용한 Multilocus sequence typing (MLST) 분석을 사용하여 계통 분석의 정확도를 높였고, 본 연구에서는 이를 벤치마킹하여 계통분석을 진행하고자함.

2-1. 16s rRNA 마커유전자를 활용한 종 동정 및 계통분석

◦ *Pseudomonas* genus 내에 속하는 균주들에 대하여, reference 16s rRNA sequence를 NCBI database에서 추출하였고, 이를 통해 Phylogenetic analysis를 진행하였음. 본 연구에서 분리한 균주는 *P. fluorescens* complex 에 속하는 것으로 밝혀졌으나, 16 rRNA 만으로는 subgroup을 판별할 수 없는 한계점이 존재함.

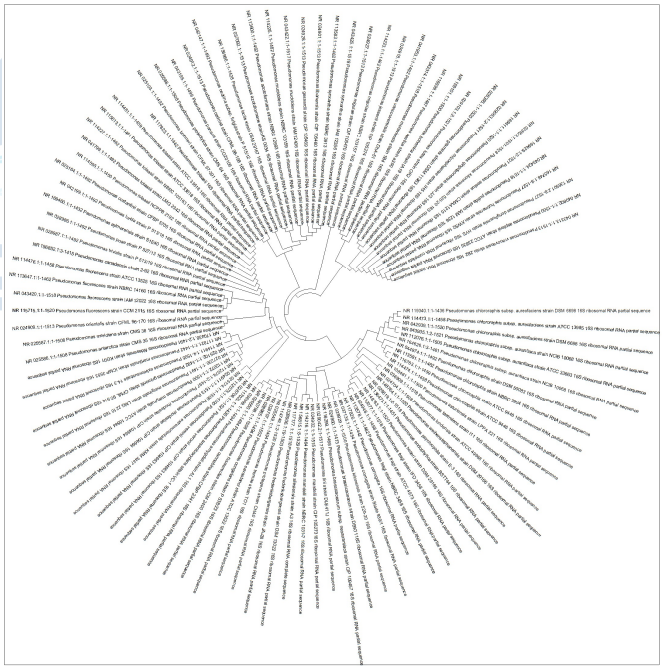


Figure 2. Phylogenetic tree using 16s rRNA sequence of *Pseudomonas* sp.

2-2. Multilocus Sequence Typing (MLST) 분석을 통한 Sequence type 확인 및 계통분석

◦ 해상도 높은 계통분석을 진행하고자 MLST 분석을 시도하게 되었고, 가장 널리 사용되고 있는 'PubMLST' (2010, BMC Bioinformatics)라는 MLST 분석 tool을 사용하여 sequence type을 결정을 시도함. 본 연구에서 분리한 *P. fluorescens* Ant01 의 경우 100% 일치하는 sequence type 이 없는 것으로 확인되었고, 가장 가까운 sequence type 은 22번으로 확인되었음. PubMLST 에 전장유전체와 sequence type 정보를 등록하여 새로운 sequence type number를 발급 중에 있음.

◦ MLST 분석에 사용된 Housekeeping gene : ileS, gyrB, rpoB, recA, rpoD, glnS, nuoD 총 6개의 유전자를 기준으로 탐색되었으며, *P. fluorescens* Ant01 균주에서는 gyrB 유전자는 검출되지 않아, 5개의 유전자를 통한 sequence type 분석이 진행되었음.

Nearest Sequence Type: 22

Locus	Identity	Coverage	Alignment Length	Allele Length	Gaps	Allele
ileS	97.283	100	552	552	0	ileS_39*
gyrB	0	0	0	0	0	No hit found
rpoB	98.742	100	477	477	0	rpoB_80*
recA	97.701	100	435	435	0	recA_22*
rpoD	98.75	100	480	480	0	rpoD_22*
glnS	96.806	100	501	501	0	glnS_22*
nuoD	96.699	99.806	515	516	0	nuoD_22*

Table 1. MLST result and the nearest sequence type of *P. fluorescens* Ant01

2-3. 전장 유전체 서열 비교를 통한 계통분석 (Average Nucleotide Identity, ANI)

◦ MLST 분석 외에도 전장 유전체를 활용한 계통분석을 진행하였다. 분석에는 complete genome level 로 assembly 가 완료된 *P. fluorescens* strain이 사용되었으며, Kostas lab (2016, PeerJ Preprints)에서 제공하는 ANI matrix analysis tool을 사용하였음. 같은 clade에 속하는 strain과 95% 이상의 similarity index를 보이는 것으로 보아, 같은 *P. fluorescens* species 에 속하는 것으로 확인됨. 하지만 *P. fluorescens* reference 균주로 알려진 SBW25 와 PF08 strain 과는 각각 87%, 85% 의 identity를 보이는 것으로 보아, 같은 종 내에서도 염기서열 변화가 상대적으로 크게 발생한 것을 확인할 수 있음.

◦ ANI 분석을 통해서 본 연구에서 분리한 *P. fluorescens* Ant01 strain 에 독립적인 strain number를 부여하기 충분한 염기서열 다양성을 보유하고 있다는 것이 확인되었고, 수직적인 진화의 과정보다는 외래 유전자 도입을 통한 수직적 진화를 통해 빠르게 종내 다양성이 증가하고 있다는 사실을 유추할 수 있음. 향후 전장 유전체 비교분석 연구를 통해서, 외부로부터 도입된 병원성 인자와 항생제 내성 유전자 분석을 수행하여, 청정지역에서 분리한 *P. fluorescens* Ant01 균주가 opportunistic pathogen 으로서의 가능성을 판별하고자함.

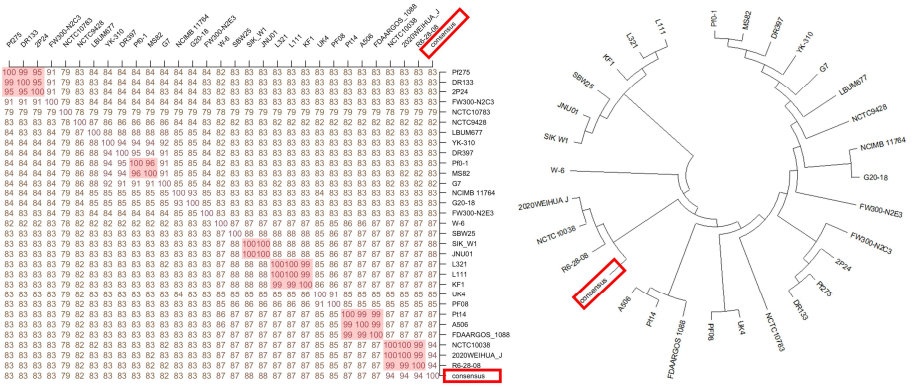


Figure 3. ANI matrix (left) and phylogenetic tree (right) of *P. fluorescens* group strains including Ant01 (marked in red box as named consensus)

3. 전장 유전체 정보 내 Pathway 분석

Orthology relationship 을 갖는 protein sequence database인 EggNOG와 pipeline을 사용하여 전장유전체 상의 protein coding genes 에 대한 functional annotation 을 진행함. EggNOG description 상에서 ‘COG1020, Non-ribosonal peptide synthetase modules and related proteins’로 annotation 되었고, 이를 기반으로 KEGG database를 활용하여 분석하였을 때, ‘Arthrofactin-type cyclic lipopeptide synthetase B, ofaB/arfB’ (KEGG number, K15659) 로 확인되었음. 따라서 본 연구를 통해서 남극에서 분리한 *P. fluorescens* 균주가 cyclic lipopeptide-type 항생제인 anthrofactin 을 합성하는 pathway를 갖고, 신규 항생물질 생산 균주 후보 발굴로서의 의미를 가짐.

3-1 Reactome pathway analysis (2022, Nucleic Acid Research)

◦ 생물학적 pathway database인 ‘Reactome DB’ 내에서 blast를 통해서 발견되며, 전장유전체 염기서열을 사용하여 결과를 얻음. Reactome 내에 protein coding region 은 UniprotKB database 내에서 blast를 통해서 reactome 내의 protein annotation을 진행하였음. 분석 결과 RHOH GTPase cycle (Reactome ID : R-DDI-9013407), Methylation (Reactome ID : (Reactome ID : (Reactome ID : R-DDI-156581), Cobalamin (Cbl, vitamin B12) transport and metabolism (Reactome ID : R-DDI-196741), Sulfur amino acid metabolism (Reactome ID: R-DDI-1614635) 총 4개의 Reactome이 확인되었음. 4개의 Reactome 내에서 공통적으로 발견되는 단백질은 Methionine synthase (Uniprot ID: Q54P92) 로 다수의 생물학적 pathway에 관여하는 단백질로 예상됨.

3-2. KEGG pathway analysis (2021, Nucleic Acid Research)

◦ EggNog mapper를 통해 얻은 target ortholog sequence를 사용하여 KEGG database 내에서 KEGG pathway를 확인함. Pathway 분석 결과 arthrofactin-type cyclic lipopeptide synthetase B (orthology number: K15659) 중심의 합성경로가 확인되었음. 이는 metabolism protein family 중, polyketide biosynthesis pawthway (KO number: 01008) 에 관여하는 핵심 단백질 중 하나로 생각되며, 신규 항생물질 발굴 가능성이 존재함.

3-3. AntiSMASH-metabolic pathway analysis (2021, Nucleic Acid Research)

◦ *P. fluorescens* Ant01 의 전장유전체 상에 존재하는 antibiotics and secondary metabolite biosynthesis gene cluster를 확인하기 위해서 AntiSMASH v,6.0 software를 사용하였음. 68%의 sequence identity 로 ‘viscosin biosynthetic gene cluster from *Pseudomonas fluorescens* SBW25’ pathway (2,788,013-2,831,449)가 발견되었음. 두 번째로 높게 예측된 gene cluster는 60% 의 sequence identity 로, ‘tolaasin I biosynthetic gene cluster from *Pseudomonas costantinii*’ pathway (1-69,124)가 발견됨. 합성이 예측되는 compound 로는 tolaasin I, F 가 존재함. tolaasin 은 toxin 의 일종으로 lipoleptide class에 속함. 이는 KEGG pattway와 유사한 예측 결과를 보여줌에 따라, *P. fluorescens* Ant01에 non-ribosomal peptide (NRP) synthesis 의 일종으로 lipopeptide synthesis pathway가 있을 것으로 생각됨.

4. 타겟 유전자 예측 소프트웨어를 사용한 *Pseudomonas fluorescens* Ant01 균주에 대한 유전체 분석

4-1. 이동성 유전자 부위 (Mobile gene element, MGE) 예측

◦ 항생제 내성유전자와 병원성 인자는 species(종) 또는 strain 사이에서 이동이 가능하

며 수평적 진화를 일으키는 하나의 원인 인자로 작용한다고 알려져있음. 따라서 이동성 유전자 부위 예측을 통해서 함께 이동한 항생제 내성유전자 또는 병원성인자의 출처를 추적함으로써 외래 도입 유전자를 추적하고 모니터링할 수 있음.

◦ ISFinder (2000, J Comput Bio) blast tool을 사용하여, *P. fluorescens* Ant01 균주 내에 존재하는 insertion sequence (IS) 를 찾고, IS 가 유래한 생물종 확인을 통해서, 외래 도입 유전자를 추적할 수 있음.

Sequences producing significant alignments	IS Family	Group	Origin	Score (bits)	E. value
ISPsy24	IS3	IS3	Pseudomonas syringae	442	3e-120
ISPa42	Tn3		Pseudomonas aeruginosa	214	1e-51
ISPsy13	IS3	IS3	Pseudomonas syringae	210	2e-50
ISPpu19	IS66		Pseudomonas putida	190	2e-44
ISPsfu1	IS5	IS5	Pseudomonas fulva	176	3e-40
ISPa126	IS3	IS3	Pseudomonas aeruginosa	157	3e-34
ISPssy	IS5	IS5	Pseudomonas syringae	137	3e-28
ISPa127	IS3	IS3	Pseudomonas aeruginosa	133	4e-27
ISPsy45	IS5	IS5	Pseudomonas syringae	127	2e-25
IS222	IS3	IS3	Pseudomonas aeruginosa	127	2e-25
ISShes11	Tn3		Shewanella sp.	107	2e-19
ISSlo2	IS3	IS3	Shewanella loihica	97.6	2e-16

Table 2. Insertion sequence of *P. fluorescens* Ant01 (Score > 97)

4-2. Genomic island analysis (IslandViewer 4)3-2. Reactome analysis (Omicsbox)

◦ Genomic island 는 8 kb 이상 길이의 genomic region으로 수평적 유전자 이동을 통해 도입된 것으로 여겨지는 유전자 부위 말함. 보통 외래 도입 항생제 내성 유전자와 병원성인자들이 이러한 방식으로 이동함. Genomic island 분석을 통해서 남극 토양에서 유래한 *P. fluorescens* Ant01 균주가 수평적 진화를 통해 얻게 된 항생제 내성 또는 병원성 인자의 분석을 수행하고자함. 이를 통해서 청정지역으로 구분되는 극지역에서의 항생제 오염과 병원성 균주에 대한 노출 정도를 측정할 수 있을 것으로 기대됨.

◦ 분석에는 IslandViewer 4 (2017, Nucleic Acids Research) 소프트웨어가 사용됨.

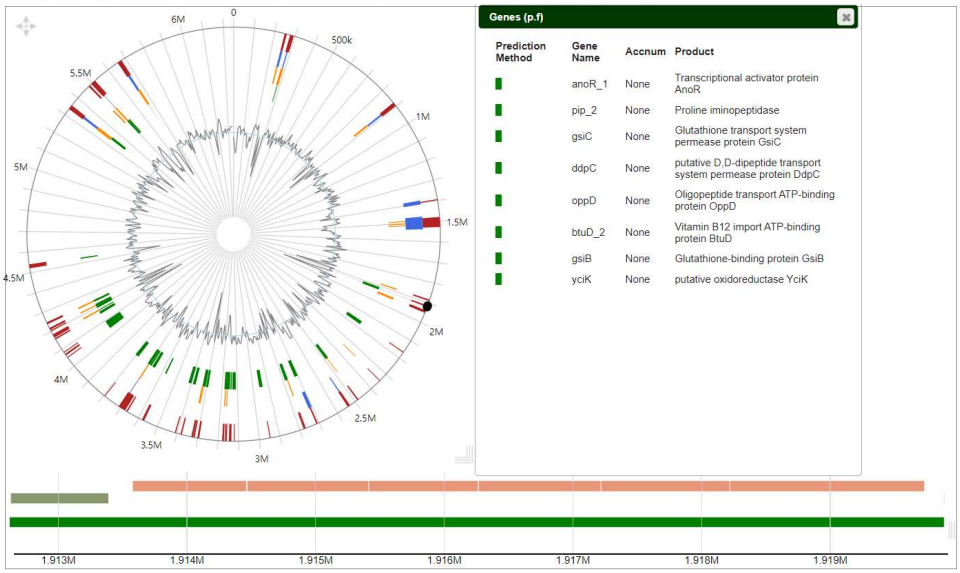


Figure 4. Peptide modification and transport related genomic island (IslandViewer 4)

5—병원성 인자 및 항생제 내성 유전자 패턴 분석 -opportunistic pathogen 가능성 탐색

5.-1. 병원성 인자 유전자 (Virulence Factor, VF) 예측 (2022, Nucleic Acids)

병원성인자 데이터베이스인 'Virulence Factor Database, VFDB' 에서 제공하는 분석 tool 인 'VFanalzer' 를 통해서 전장 유전체 상에 존재하는 병원성인자 (=병원성 유전자) 예측을 가능하게 하며, VFDB 상에 등록되어있는 species와 비교분석을 수행함.

VFDB 상에서 *P. fluorescens* 뿐만 아닌, 병원성 균주로 알려진 *P. aeruginosa*, 식물 병원성 균주로 알려진 *P. syringae*, PGPB 균주로 알려진 *P. putida* 균주들의 reference genome 상에 존재하는 병원성인자들과 비교 분석이 가능함. 분석 결과, 총 5,616 개의 CDS 유전자 중 144개의 병원성 인자 관련 유전자가 검출되었고, 그 중 *P. aeruginosa* 또는 *P. syringae* 균주 그룹으로부터 유래한 것으로 추정되는 유전자들이 함께 발견되었음. 해당 결과와 이동성 유전자 (mobile gene element) 분석을 함께 적용하면 병원성인자가 유래한 균주 확인이 가능할 것으로 생각됨.

VFclass	Virulence factors	Related genes	*P. fluorescens* Ant01	*P. fluorescens*	*P. aeruginosa*	*P. putida*	*P. syringae*
Adherence	LPS O-antigen (P. aeruginosa)	Undetermined	01798, 01801	X	O	X	X
	Type IV pili twitching motility related proteins	chpD, chpE	02990, 02991	△	O	X	X
Biosurfactant	Rhamnolipid biosynthesis	rhlA	05261	△	O	X	X
Iron uptake	Pyoverdine	pvdF, pvdJ	03213, 03216	△	O	X	X
Secretion system	Hcp secretion island-1 encoded type VI secretion system (H-T6SS)	Undetermined	05602, 05603	△	O	X	X
	P. aeruginosa TTSS	pcrD, pscN	03439, 03429	X	O	X	X
	P. syringae TTSS	hrcN, hrcV	00756, 00754	X	X	X	O
Protease	Alkaline protease	aprA	00210, 00211 , 00212, 02775	O	O	X	O

Table 3. Comparative analysis of virulence factor between *Pseudomonas* spp.

5-2. 항생제 내성 유전자 (Antibiotic Resistance Gene, ARG) 예측

◦항생제 내성 유전자 데이터베이스인 'Comprehensive Antibiotic Resistance gene Database, CARD' (2020, Nucleic Acids) 에서 제공하는 분석 tool 인 'Antibiotic Resistance Gene Identifier, RGI' 를 통해서 전장 유전체 상에서 항생제 내성 유전자를 예측할 수 있음. 항생제 종류와 항생제 내성 기작에 따른 분류 결과도 함께 제시함. 따라서 청정지역과 오염지역으로 분리된 지역에서 발견되는 *P. fluorescens* 종에서 예측되는 항생제 내성유전자에 대한 정성분석과 정량분석을 수행함.

RGI Criteria	ARO Term	SNP	Detection Criteria	AMR Gene Family	Drug Class	Resistance Mechanism	% Identity of Matching Region	% Length of Reference Sequence
Strict	vanG		protein homolog model	glycopeptide resistance gene cluster, Van ligase	glycopeptide antibiotic	antibiotic target alteration	37.47	104.30
Strict	Pseudomonas aeruginosa soxR		protein homolog model	ATP-binding cassette (ABC) antibiotic efflux pump, major facilitator superfamily (MFS) antibiotic efflux pump, resistance-nodulation-cell division (RND) antibiotic efflux pump	fluoroquinolone antibiotic, cephalosporin, glycylcycline, penam, tetracycline antibiotic, rifamycin antibiotic, phenicol antibiotic, disinfecting agents and antiseptics	antibiotic target alteration, antibiotic efflux	70.42	95.51
Strict	adeF		protein homolog model	resistance-nodulation-cell division (RND) antibiotic efflux pump	fluoroquinolone antibiotic, tetracycline antibiotic	antibiotic efflux	43.47	97.17
Strict	adeF		protein homolog model	resistance-nodulation-cell division (RND) antibiotic efflux pump	fluoroquinolone antibiotic, tetracycline antibiotic	antibiotic efflux	67.17	100.00
Strict	Acinetobacter baumannii AbaQ		protein homolog model	major facilitator superfamily (MFS) antibiotic efflux pump	fluoroquinolone antibiotic	antibiotic efflux	72.49	101.38

Table 4. ARG prediction results of *P. fluorescens* Ant01 using Resistance Gene Identifier (CARD-RGI) software

5-2. 청정지역과 오염지역에서의 항생제 내성 유전자 패턴 분석

◦ 항생제 오염을 기준으로 청정지역과 오염지역에서 발견되는 bacteria 의 전장유전체 서열을 사용하여 항생제 내성유전자 패턴 분석을 수행하고자함.

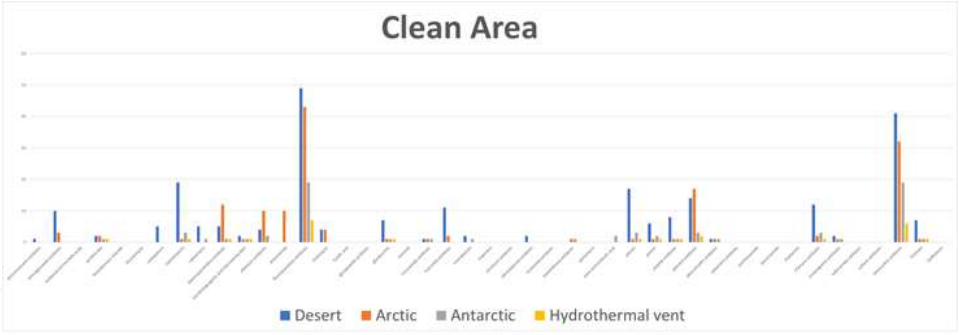
◦ NCBI-Biosample database에서 '채집지 환경' 을 키워드로 검색하여 전장유전체 데이터를 수집하였음. 청정지역에 해당하는 환경으로는, 북극, 남극, 사막, 열수구를 선정

하였고, 오염지역에 해당하는 환경으로는, 오염수, 하수처리장, 가축, 농업지, 수산양식지, 마지막으로 병원을 선정하였음.

◦ 각각의 환경에 따라 분리된 미생물들의 전장유전체 서열을 수집하였고, 그 중에서 complete genome, chromosome level의 유전체 정보만을 선별하여 항생제 내성유전자 검출 정확도를 높였음. 항생제 내성 유전자 분석은 CARD-RGI software를 사용하였고, 청정지역에서 120개, 오염지역에서 260개의 선별된 전장유전체를 사용하였음.

Clean Env. (Non-anthropogenic area)					Polluted Env. (Anthropogenic area)					
	Antarctic	Arctic	Desert	Hydro-thermal vent	Waste-water	Sewage	Livestock	Agri-culture	Aqua-culture	Hospital
Complete level/total assembly	33/180	48/526	30/863	9/795	15/903	38/576	78/552	4/977	51/201	75/998
Total	120/2,364				260/4,007					

Table 4. ARG prediction results using bacteria whole genome assembly deposited NCBI database. The clean and polluted area is separated by the standard of antibiotics pollution.



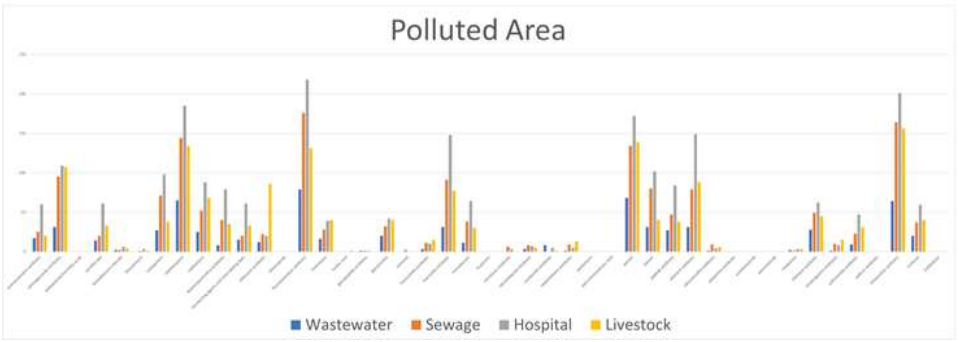


Figure 5. ARG prediction results are quantified with 45 types of antibiotics classification.

6. 항생제 내성 유전자 예측결과에 따른 표현형적 검증

◦ 극지에서 분리한 균주의 항생제 내성에 대한 표현형적인 특징을 확인하기 위해 다양한 종류의 항생제에 대한 감수성 또는 내성을 Disk diffusion assay 실험을 통해 확인함. (Ampicillin, chloramphenicol, Spectinomycin, Tobramycin, Streptomycin, Gentamycin, Vancomycin, Paromomycin, Sisomicin, Apramycin, Kanamycin, Amikacin, Meropenem, Tebipenem, Doripenem, Biapenem, Imipenem, Tetracyclin).

◦ 극지역에서 발견되는 미생물의 경우 다양한 항생제에 대한 감수성이 뛰어나기 때문에, 항생제 오염에 상대적으로 노출빈도가 적은 것으로 확인이 됨.

◦ 항생제 노출에 대한 유전체 패턴을 분석하기 위해서, 전장유전체 데이터가 풍부하며 극지역에서 분리한 미생물 중에서 전장유전체 분석 (Whole genome seqeuncing, WGS) 이 완료되지 않은 균주를 선별하였음.

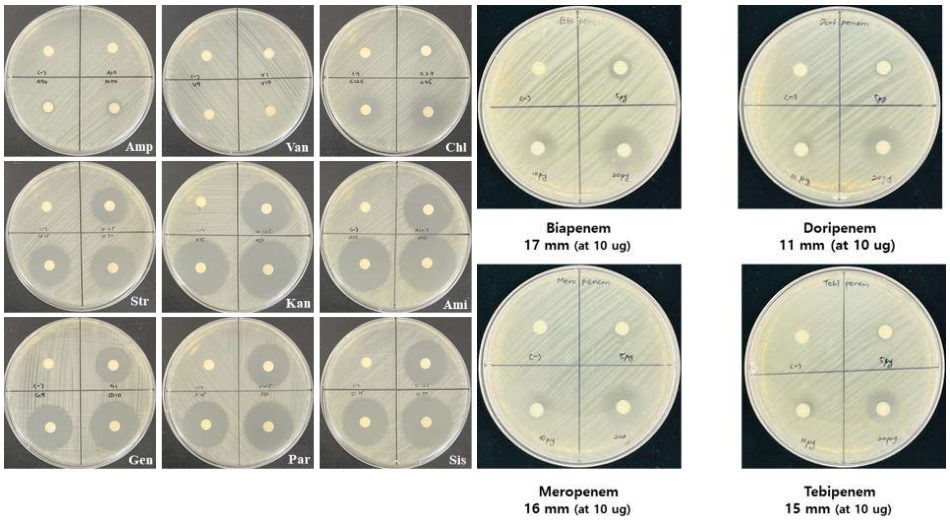


Figure 6. 다양한 종류의 항생제에 대한 감수성 (사용된 항생제 : Ampicilllin, Kanamycin, Chloramphenicol, Streptomycin, Vancomycin, Amikacin, Gentamicin, Paromomycin, Sisomicin, Biapenem, Doripenem, Meropenem, Tebipenem

Figure 7. Imipenem 항생제 내성을 보이는 *Pseudomonas fluorescens Ant01*

연구성과	*논문 투고 진행중*
프로젝트 기술분야	극한지역 미생물 유전체 비교분석, 항생제 내성 유전자 패턴 분석