

2026

데이터, AI, 자동화를 중심으로 본 합성생물학 융합기술동향



KSBA

KSBA

한국합성생물학발견협회
Korea Synthetic Biology Association

Contents



한국합성생물학발전협의회
Korea Synthetic Biology Association

총론

이도현

융합분과 분과장, KAIST 교수

004

합성생물학을 위한 데이터 기술동향

007

김현욱

융합분과위원, KAIST 교수

008

김하성

융합분과위원, 한국생명공학연구원 책임연구원

022

나도균

융합분과위원, 중앙대학교 교수

042

정철희

융합분과위원, 고려대학교 교수

070

합성생물학을 위한 AI 기술동향

083

이준학

융합분과위원, 한국과학기술정보연구원 디지털바이오컴퓨팅연구단 단장

084

염민선

융합분과위원, 나무아이씨티 연구소장

106

이선재

융합분과위원, 광주과학기술원 생명과학과 교수

126

송자연

융합분과위원, 성균관대학교 송자연 조교수

140

합성생물학을 위한 자동화 및 로봇틱스 기술동향

159

임현의

융합분과위원, 한국기계연구원 바이오파운드리 DX/AI 연구단 단장

160

이대식

융합분과위원, 한국전자통신연구원 책임연구원

174

김진환

융합분과위원, 케이랩 주식회사 연구소장

186

이석재

융합분과위원, 나노종합기술원 나노융합기술개발본부 본부장

204

총론

합성생물학은 유전체·단백질·대사 네트워크를 설계 가능한 공학적 대상으로 바라보며, 생명체의 기능을 정밀하게 조정하고 새로운 생물학적 장치를 구축하려는 시도에서 출발한다. 최근 생명정보의 폭발적 증가, 계산기술의 급격한 발전, 자동화 장비의 고도화가 동시에 이루어지면서 합성생물학은 급속히 “데이터 기반 생명공학”으로 전환되고 있다. 이러한 변화는 단순한 실험 규모 확장을 넘어, 설계—제작—검증—학습(DBTL)이라는 반복적 연구구조를 지능적으로 운영할 수 있는 기반이 마련되었다는 점에서 기술적·사회적 의미가 크다.

현재의 합성생물학이 지향하는 복잡도와 속도를 감안하면, 전통적인 수작업 중심 실험 방식만으로는 산업적 요구를 충족하기 어렵다. 대규모 데이터의 체계적 축적과 활용, 인공지능(AI) 기반 분석·설계 기술, 그리고 자동화·로보틱스 실험 인프라가 결합되어야 미래형 합성생물학 생태계가 완성된다. 본 보고서는 세 가지 기술축을 하나의 연속적 흐름으로 재정리하여, 합성생물학의 기술적 구조와 정책적 시사점을 종합적으로 제시한다.

합성생물학의 목적은 생명 기능을 공학적으로 이해하고 새롭게 설계하는 데 있다. 이를 위해서는 생명 현상이 시간·공간적으로 어떻게 변화하는지를 정밀하게 측정된 데이터가 필요하다. 생명 시스템은 유전자 발현, 단백질 상호작용, 효소 반응, 대사 플렉스, 신호 전달 등의 복잡한 계층으로 구성되며, 이 계층들은 서로 영향을 주고 받으며 동적으로 변화한다. 따라서 단일 차원의 데이터만 확보해서는 생명 현상의 전체 맥락을 이해하기 어렵다.

최근에는 단일세포 전사체, 대규모 단백질체·인산화체, 실시간 생물공정 데이터, 고해상도 영상 기반 세포 표현형 등 다양한 모달리티의 데이터가 통합적으로 생산되고 있다. 그러나 생산 비용이 높고 분석 난도가 커 분산된 연구실 단위에서 구조적 통합이 쉽지 않다. 특히 생물공정 데이터는 실험마다 조건이 다르고 변동성이 크기 때문에 고품질 데이터를 지속적으로 축적하기 어렵다. 이러한 이유로 국가 단위의 데이터 인프라, 표준화된 메타데이터 구조, 연구시설 간 상호운용성이 필수적인 요소로 부상하고 있다.

또한 연구현장에서 생산된 데이터를 효과적으로 공유하는 데에는 기술적·사회적 난관이 존재한다. 기술적 난관은 데이터의 형태와 의미를 규격화 혹은 표준화하여 상호운용 가능하도록 하는 기술과 원활한 공유를 위한 정보인프라가 필요하다는 점이다. 사회적 난관은 전통적인 연구 문화 특성상 자신이 공들여 생산한 데이터를 외부로 공개하는데 심리적, 현실적 장벽이 존재한다는 점이다. 이러한 난관을 극복할 수 있는 기술적, 제도적 체계 마련이 필요하다. 결국 데이터의 통합력은 합성생물학의 예측·설계 능력을 결정하는 가장 중요한 기반이며, 데이터가 충분히 모이고 서로 연결될 때 AI 기반 설계와 자동화 실험은 비로소 높은 효율을 발휘하게 된다.

AI 기술은 합성생물학의 ‘두뇌’ 역할을 수행한다. 생명정보는 고차원적이고 비선형적이며 방대한 경우가 많아 인간이 직접 구조를 파악하거나 최적의 설계를 선택하기 어렵다. AI는 이러한 복잡한 패턴을 학습해 DNA·RNA·단백질 구조와 기능을 예측하고, 특정 목적에 적합한 생물학적 설계를 자동으로 생성할 수 있다.

대규모 언어모델 기반의 핵산 및 아미노산 서열 모델은 단백질의 구조적 안정성을 예측하고, 특정 화학반응을 수행할 수 있는 새로운 효소를 설계하며, 유전자 발현을 조절하는 서열을 자동으로 생성하는 데 활용된다. 이미지 기반 모델은 세포 형태 변화나 세포 간 상호작용을 분석하는 데 유용하며, 시계열 모델은 생물 공정의 상태 변화 패턴을 예측한다.

AI는 설계뿐 아니라 DBTL 전체의 효율을 높이는 역할을 한다. 실험 결과로 생성된 데이터를 분석하여 다음 단계의 설계를 개선하는 ‘지능형 피드백 구조’를 형성하고, 기존의 시행착오 기반 탐색보다 훨씬 빠른 속도로 최적화된 설계안을 도출한다. 이러한 구조는 합성생물학을 예측-설계 중심의 공학 discipline으로 전환시키는 핵심적 계기라 할 수 있다.

합성생물학의 실험량은 매우 방대하며, 한 번의 연구 프로젝트에서도 수백~수천 개의 유전자 조합, 수천 개의 균주, 수십만 개의 배양-측정 공정이 필요할 수 있다. 수작업 기반 접근으로는 시간·노동력·정확도 측면에서 감당하기 어렵다. 따라서 자동화 장비는 DBTL 사이클을 실제로 운영하기 위한 필수적 요소이다.

자동화된 유전자 합성 플랫폼은 고속·대량 DNA 조립을 가능하게 하며, 로봇 배양 시스템은 균주를 안정적으로 성장시킨 뒤 고처리량 분석 장비로 실험 결과를 자동 추출한다. 고해상도 영상 분석, 유세포 분석, 질량분석 기반 대사체 측정 등 다양한 분석 장비는 자동화 소프트웨어와 연동되어 반복적이고 정밀한 실험 수행을 가능하게 한다.

또한 반도체 기술을 활용한 온칩 기반 합성생물학 플랫폼은 초소형·초병렬 실험을 가능하게 하여, 실험 속도를 극대화하고 비용을 낮춘다. 이러한 자동화 기술은 단순히 실험을 빠르게 하는 수준이 아니라, AI가 학습할 수 있는 수준의 안정적이고 재현성 있는 데이터 생성 환경을 만드는 핵심 기반이다.

효율적인 합성생물학 R&D를 위해서는 데이터, AI, 자동화 플랫폼이 각각 독립적으로 존재해서는 안 된다. 세 요소는 데이터로부터 AI 학습으로, AI 예측 및 설계로부터 자동화된 실험으로, 그리고 실험결과는 새로운 데이터로 제시되어 AI 모델을 개선하는 순환구조를 형성하여, 서로의 효율을 극대화할 수 있어야 한다. 이러한 통합 구조는 실험 규모와 속도를 획기적으로 확대하고, 산업 수준에서 합성생물학을 활용할 수 있는 기반을 마련한다. 정책적 측면에서도 이 세 요소의 상호 연동성 확보는 국가 생명공학 경쟁력의 핵심 지표로 작용할 것이다.

합성생물학 생태계를 효과적으로 구축하기 위해서는 기술적 요소뿐 아니라 다음과 같은 정책적 고려가 필수적이다.

- 국가 단위 데이터 레이크 구축 및 연구시설 간 데이터 공유 체계 확립
- 자동화 장비와 생명공학 플랫폼의 국산화 및 고도화
- AI 기반 생명설계를 위한 고성능 컴퓨팅 인프라 확충
- 다중 모달리티 생명정보의 표준화 및 상호운용성 확보
- 안전·보안 가이드라인 마련 및 기술 활용 규제 정비

이러한 제도적 기반이 갖추어질 때, 연구자들은 안정적인 인프라 위에서 혁신 연구를 장기적으로 수행할 수 있으며, 기업은 새로운 합성생물학 기반 제품을 시장에 빠르게 내놓을 수 있다.

결론적으로 합성생물학은 데이터·AI·자동화라는 세 기술축이 유기적으로 결합될 때 비로소 실질적인 산업적 영향력을 발휘하게 된다. 본 보고서에서 제시한 통합적 관점은 이러한 기술적 상호작용을 하나의 일관된 구조로 재해석하고 있으며, 이는 앞으로의 연구 개발 체계 구축과 정책 방향 설정에 중요한 참고가 될 것이다. 앞으로의 합성생물학 R&D는 이 세 기술의 고도화와 통합 능력에서 경쟁력이 판가름 될 것이다.

SECTION
1

합성생물학을 위한 데이터 기술동향



한국합성생물학발전협의회
Korea Synthetic Biology Association

김현욱

융합분과위원, KAIST 교수

김하성

융합분과위원, 한국생명공학연구원 책임연구원

나도균

융합분과위원, 중앙대학교 교수

정철희

융합분과위원, 고려대학교 교수

유전체 데이터 기반 인공지능 기술 동향과 합성생물학적 활용

1 머리말

합성생물학은 생명체의 유전적 구성요소를 설계·재조합하여 새로운 기능을 창출하는 융합 분야로, 그 성장은 유전체 데이터의 생산과 활용 역량에 의해 결정된다. 최근 차세대 시퀀싱(Next-Generation Sequencing, NGS) 기술의 고도화와 인공지능의 발전으로, 유전체 데이터는 단순한 서열 정보에서 벗어나 생물학적 기능과 설계 가능성을 직접 예측할 수 있는 지식 자원으로 전환되고 있다. 본 장에서는 이러한 변화 속에서 국내외 합성생물학 연구가 어떻게 데이터 중심으로 설계되고 발전해가는지를 조망한다. 특히 유전체 데이터의 품질 평가, 발현 및 기능 예측, 변이 효과 분석, 유전자 클러스터 분류에 이르는 전 과정에 인공지능을 접목하여 생물학적 복잡성을 해석하고 설계 효율을 혁신적으로 향상시키는 기술적 흐름을 살펴본다.

2 유전체 데이터의 기술 동향: 품질 평가에서 기능 예측까지

2.1 유전체 데이터(Genomic Data)

유전체 연구의 발전으로 전장 유전체 시퀀싱(Whole-Genome Sequencing, WGS)을 통해 방대한 양의 유전체 데이터가 생성되고 있으며, 이를 효과적으로 관리하고 분석하는 것이 중요한 과제로 부상하고 있다. 차세대 시퀀싱의 발전은 데이터 생성 비용을 크게 낮추고 그 산출량을 늘렸지만, 데이터셋의 막대한 규모와 높은 복잡성으로 인해 데이터 처리와 품질 관리에 어려움을 초래했다. 이를 해결하기 위해 유전체 데이터를 고처리량(high-throughput) 방식으로 처리하고 품질을 관리하는 것이 중요해졌으며, 특히 최근에는 인공지능 기법을 도입하여 방대한 데이터를 효율적으로 처리하는 추세이다.

한편, 이렇게 생성된 방대한 유전체 데이터는 다양한 생물학 데이터베이스의 기반 자원으로 활용되고 있다. 유전자는 단백질, 대사경로, 화합물 등 모든 생명현상의 근간이며 KEGG(Kyoto Encyclopedia of Genes and Genomes), UniProt, PDB(Protein Data Bank) 등과 같은 주요 생물학 데이터베이스들 또한 유전자 서열(유전체)과 연결되어 있다. 즉, 대부분의 주요 생물학 데이터베이스는 유전체 정보를 참조하거나 제공하고 있다(표 1).

표 1> 유전체 데이터와 연관된 주요 생물학 데이터베이스

NP-Atlas	데이터베이스	주요 내용 및 데이터	데이터 규모	최신 업데이트
유전체 (Genomics)	NCBI [1]	• 생물학 전반의 핵심 데이터 허브 • GenBank, RefSeq, PubMed 포함 • 유전체-단백질-논문 분석 도구 제공	수십억 건의 서열 및 주석	2025 (v15.0.0)
	GenBank [2]	• 모든 공개 서열의 저장소 • 원시 DNA/RNA 서열 데이터 보관 • 주석 및 돌연변이 정보 제공	약 59억 건의 서열	2025 (Release 268)
	RefSeq [3]	• 표준화된 비중복 참조 서열 제공 • 검증된 유전자-전사체-단백질 서열 포함 • 품질 보증된 참조 데이터 제공	약 5.6억 건의 서열 레코드	2025 (RefSeq Release 232)
대사 경로 (Metabolic pathway)	KEGG [4]	• 유전자-단백질-대사경로 통합 해석 플랫폼 • 경로맵, 효소, 약물, 질병 등 관련 정보 제공	Genes: 약 6,289만건, Pathway maps: 581개 등	2025 (Release 116.0)
단백질 (Protein)	UniProt [5]	• 단백질 서열 및 기능 주석 • 단백질 구조, 변이 정보 포함	약 2억 개의 단백질 서열(UniProtKB)	2025
	PDB [6]	• 단백질-핵산의 3D 구조 데이터 • X-ray, NMR, Cryo-EM 기반 구조 제공 • 구조 좌표 및 리간드 정보 포함	약 24만 개의 단백질 구조	2025
천연물 (Natural products)	NP-Atlas [7]	• 미생물 유래 천연물 구조의 수집 및 정제 데이터 • 화학 구조 및 생물학-화학 메타데이터 제공	약 36,500개 화합물	2024
	COCONUT [8]	• 공개 천연물 데이터 통합 컬렉션 • 화합물, 기원 생물 종 및 참고문헌 정보 포함	약 45만 개 화합물	2025
생합성 유전자군 (Biosynthetic Gene Clusters, BGCs)	MIBiG [9]	• 실험적으로 검증된 BGC의 표준화-전문 검토 데이터베이스 • BGC 메타데이터 • 생합성 구조 및 생물활성 정보 포함	약 4,200개 BGC	2024 (v4.0)
	antiSMASH D [10]	• 2차 대사 관련 미생물 유전체의 BGC 자동 주석화 데이터베이스 • 미생물 유전체로부터 주석화된 BGC 정보	약 23만 1,500개 BGC, 592개 균주(고 세균-세균-진균 포 합)	2024 (v4.0)
	JGI-SMC [11]	• 미생물 생합성 잠재력의 대규모 예측 및 탐색 지원 • 미생물 유전체 및 메타게놈에서 예측된 BGC • 분류학적 정보, 모듈 구조, 유사도 주석 포함	1.1만 개 BGC 영역, 130만 개 서열	2025
	ClusterCAD [12]	• PKS(Polyketide Synthase)/NRPS(Nonribosomal Peptide Synthetase) 기반 모듈형 생합성 경로 설계 플랫폼 • PKS/NRPS 및 하이브리드 BGC 서열 • 예측 화학 구조	531개 BGC, 그 중 183개 수동 검증	2022 (v2.0)

유전체 분야의 핵심 데이터베이스들은 생명정보의 수집, 검증, 그리고 표준화를 담당하며 생물학 연구의 기반을 이루고 있다. 그 중 NCBI(National Center for Biotechnology Information), GenBank, RefSeq(Reference Sequence)는 서로 유기적으로 연결되어 유전체 데이터를 생성·관리·제공하는 중심 축으로 기능하고 있다.

NCBI는 미국 국립보건원(National Institutes of Health, NIH) 산하 기관으로, 분자 생물학 정보의 국가적 허브이자 생명정보학 연구의 중심 역할을 수행하고 있다. NCBI는 GenBank, RefSeq, PubMed, BLAST(Basic Local Alignment Search Tool) 등 다양한 핵심 생물정보 자원과 도구를 운영하며, 전 세계 연구자들이 생물학적 데이터에 자유롭게

접근하고 분석할 수 있도록 지원한다. 이러한 활동을 통해 NCBI는 유전체 정보의 축적·표준화·활용을 선도하는 국제적 생명정보 플랫폼으로 자리매김하고 있다.

GenBank는 NCBI가 운영하는 세계 최대의 공개 유전자 서열 데이터베이스로, 전 세계에서 생성된 모든 공개 DNA 서열 정보를 수집·보관하고 있다. GenBank는 일본의 DDBJ(DNA Data Bank of Japan) 및 유럽의 ENA(European Nucleotide Archive)와 함께 International Nucleotide Sequence Database Collaboration(INSDC)을 구성하며, 세 기관은 매일 데이터를 상호 교환하여 전 세계적으로 일관된 유전자 정보 접근을 보장한다. 데이터는 연구자들로부터 직접 제출되며, 자동화 및 수동 검증 과정을 거쳐 품질이 보장된 후 공개된다. GenBank는 2개월마다 업데이트되며, Entrez Nucleotide, BLAST, NCBI e-utilities 등을 통해 검색·정렬·다운로드할 수 있다. 데이터 사용에는 별도의 제한이 없으나, 일부 제출자는 지적 재산을 주장할 수 있다. 이러한 개방형 구조를 통해 GenBank는 가장 최신의 유전체 서열 정보를 공유하고 생명과학 연구를 촉진하는 핵심 글로벌 인프라로 기능하고 있다.

RefSeq는 NCBI가 제공하는 유전자 참조 서열 데이터베이스로, 유전체(genome), 전사체(transcriptome), 단백질(protein) 정보를 통합적으로 포함하고 있다. RefSeq는 생명과학 연구에서 표준 역할을 수행하며, 유전체 주석(annotation), 유전자 식별 및 기능 분석, 돌연변이 및 다형성 연구, 발현 분석, 종 간 비교 연구 등에 활용된다. 데이터는 GenBank에서 선택된 서열을 기반으로 계산 분석, 자동화된 유전체 주석 파이프라인(진핵/원핵), 수동 큐레이션 등을 통해 생성되며, 원핵생물·진핵생물·고세균·바이러스 등 다양한 생물 종을 포괄한다. RefSeq의 주요 특징은 비중복성, 뉴클레오타이드-단백질 서열의 명시적 연계, 지속적인 업데이트와 품질 검증, 형식의 일관성 유지, 전문가 및 협력 기관에 의한 큐레이션이다. 모든 RefSeq의 액세스션(accession) 번호는 밑줄(_)을 포함하며, BLAST, Entrez, NCBI Assembly 및 Genome Data Viewer 등을 통해 접근할 수 있다. 이를 통해 RefSeq는 유전체 해석의 안정적인 기준을 제공하며, 생물학적 정보의 신뢰성과 재현성을 보장하는 핵심 데이터 인프라로 자리잡고 있다.

한편, 천연물 데이터베이스는 유전체 데이터에 포함된 생합성 유전자군(Biosynthetic Gene Clusters, BGCs)의 분석을 통해 유전자와 화합물 간의 복잡한 관계를 해석하는 핵심 기반으로 활용되고 있다^[13]. 최근에는 이러한 천연물 데이터베이스를 기반으로 인공지능과 계산생물학 기법을 활용해 유전체 정보로부터 효소 반응과 생합성 경로를 예측하는 연구가 활발히 진행되고 있다. 이러한 접근은 대규모 유전체 및 대사체 데이터를 통합하여 합성생물학의 관점에서 새로운 생합성 경로를 설계하고, 천연물의 재조합최적화를 가능하게 한다.

예를 들어, NP-Atlas는 미생물 유래 2차 대사산물의 구조와 생합성 분류 정보를 포함하고 있어, 특정 화합물이 어떤 균주의 유전체 내에서 생성되는지를 추적하는 데 유용하다. 이러한 정보는 유전적 기원과 대사산물 간의 연관성을 규명하는 데 사용되며, 미생물 대사체 맵핑 및 생합성 경로 해석의 참고 자료로 기능하고 있다.

또한 COCONUT(Collection of Open Natural Products)과 같은 데이터베이스는 화합물 구조에 더해 분류학적 기원과 채집 위치, 외부 유전체 기반 자원(PubChem, ChEBI 등)과의 연결 정보를 제공함으로써, 화합물-유전자-생물 종 간 관계 분석을 가능하게 한다. 이러한 통합적 데이터는 천연물 구조의 다양성을 유전체 수준의 대사 잠재력과 연계해 탐구할 수 있는 기반을 제공하고 있다.

2차 대사산물을 그에 대응하는 생합성 유전자군과 연결하는 과정 또한 기본적으로 유전체 데이터에 기반하고 있다. 이러한 생합성 유전자군 데이터베이스들은 유전 정보와 화학 구조, 실험 데이터를 통합하여, 유전체 내에서 특정 대사 경로가 어떻게 암호화되어 있는지를 해석하고 예측할 수 있도록 한다.

MIBiG(Minimum Information About a Biosynthetic Gene Cluster)는 3,600개 이상의 실험적으로 검증된 생합성 유전자군을 포함하는 표준화된 참조 데이터베이스로, 각 항목에는 유전체 좌표(genomic coordinates), 도메인 구조, 주석 정보, 생성물의 화학 구조 및 생물활성이 기록되어 있다. 이를 통해 연구자는 유전체 상의 특정 유전자 구간이 실제 생화학적으로 특성화된 경로에 해당하는지를 검증할 수 있으며, antiSMASH(Antibiotics and Secondary Metabolite Analysis Shell)나 PRISM(Prediction Informatics for Secondary Metabolomes)과 같은 유전체 마이닝 도구들은 MIBiG을 벤치마크로 사용해 생합성 유전자군 예측 결과를 평가한다.

JGI-SMC(Joint Genome Institute's Secondary Metabolism Collaboratory)는 수백 개의 게놈 및 메타게놈 데이터를 병렬로 처리하여 새로운 생합성 유전자군을 고속 탐색하는 클라우드 기반 분석 플랫폼이다. 각 후보군은 분류학적 기원, 도메인 구조, 그리고 MIBiG와의 유사도 점수를 기반으로 주석화되며, 이렇게 생성된 풍부한 정보는 유전체 데이터와 대사체 데이터를 연결하는 연구에 활용된다.

또한 ClusterCAD 2.0은 유전체에서 유래한 폴리케타이드 및 비리보솜 펩타이드 유전자 정보를 기반으로, 새로운 생합성 경로를 설계하고 효소 조합 모델링 및 경로 플렉스 시뮬레이션을 수행할 수 있는 통합 도구를 제공한다. 이 플랫폼은 유전자 서열에 포함된 도메인 구조와 기질 특이성 정보를 활용하여, 유전체 수준에서 합성 가능한 새로운 대사 경로를 설계할 수 있게 한다.

이러한 천연물 데이터베이스나 생합성 유전자군 데이터베이스들은 유전체 데이터로부터 생합성 경로를 해석하고 설계할 수 있는 기반을 마련함으로써, 복잡한 유전 정보와 대사 기능 간의 연결고리를 명확히 하고 있다. 그러나 유전체 데이터의 규모와 복잡도가 증가함에 따라, 사람이 수작업으로 수집·정제(upstream) 단계를 수행하거나 분석·모델링(downstream) 과정에서 활용하기에는 한계가 있다. 이에 유전체 데이터 분석에 인공지능 기법을 도입하여 방대한 유전체 데이터에 내재한 인간이 파악하기 어려운 패턴과 상관관계를 자율적으로 학습·식별함으로써 생합성 잠재력 예측과 신규 대사 경로 설계의 효율을 크게 높이는 추세이다.

인공지능 기법을 유전체 데이터 처리에 적용하는 것의 핵심 강점은 복잡한 패턴을 자율적으로 학습·식별할 수 있다는 점에 있다. 또한 인공지능은 방대한 유전체 데이터 속에 숨겨진 미세한 연관성을 밝혀내어, 균주 개발에 중요한 유전 요소의 기능적 의미를 예측할 수 있게 하였다. 예를 들어, 합성곱 신경망(Convolutional Neural Network, CNN)과 같은 딥러닝 아키텍처를 미생물 유전체에 학습시키면 대사 경로를 제어하는 프로모터 또는 다른 조절 요소를 식별하도록 할 수 있다. 이와 같이 분석에 소요되는 시간과 비용을 획기적으로 줄임으로써, 인공지능은 연구 효율을 극대화하고 미생물 생물공정의 합리적 설계와 최적화를 위한 핵심 도구로 자리매김하고 있다.

이러한 인공지능 기반 접근법은 유전체 데이터 분석과 합성생물학 전반에 걸쳐 핵심적인 변화를 이끌고 있으며, 이를 현대 생명공학의 주요 연구 동향으로 볼 수 있다. 이어지는 내용에서는 이러한 변화를 중심으로, 1) 유전체 품질 평가, 2) 유전자 발현 예측, 3) 유전자 서열 기반 기능 예측, 4) 유전자 돌연변이 효과 예측, 5) 유전자 클러스터 분류의 다섯 가지 주제를 통해 유전체 데이터 분석이 인공지능 기법을 통해 합성생물학에 보다 실질적이고 효율적으로 활용되는 흐름을 살펴본다^[14].

2.1.1 유전체 품질 평가(Genome Quality Evaluation)

모든 생물정보학 연구와 생명공정의 최적화는 고품질의 유전체 데이터를 기반으로 시작된다. 정확하고 신뢰할 수 있는 유전체 데이터 없이는 이후의 변이 분석, 유전자 발현 해석, 생합성 경로 예측 등 어떤 연구도 신뢰성 있게 수행될 수 없다. 따라서 연구의 첫 단계는 데이터의 양보다 유전체 서열의 품질을 정밀하게 평가하고 검증하는 과정이어야 한다.

유전체 품질 평가의 근본적인 어려움은 실제 생물학적 변이와 무작위 시퀀싱 오류(artifacts)를 구분하는 것에 있다. 딥러닝 기반 변이(variant) 탐지 도구인 DeepVariant는 이 문제를 이미지 분류 과제로 재정의하여 해결한다. 즉, 특정 유전자의 위치를 포함

하는 시퀀싱 데이터를 여러 층으로 쌓은 이미지 형태로 변환한 뒤, 염기서열 정보와 품질 점수, 기타 다양한 속성들을 컴퓨터가 이해할 수 있는 숫자 데이터(텐서)로 바꾼다. 이렇게 변환된 데이터를 합성곱 신경망에 학습시켜, 실제 변이에서 나타나는 미세한 패턴을 구분하고 시퀀싱 과정에서 생기는 단순한 오류 패턴은 걸러낸다. 이러한 방식으로 각 후보 변이를 정확히 분류함으로써, DeepVariant는 거짓 양성 및 거짓 음성 변이 검출을 크게 줄여 유전체 데이터의 정확성과 신뢰도를 향상시킨다. 이와 유사하게 Clairvoyante는 PacBio나 Oxford Nanopore과 같은 롱 리드 시퀀싱(long-read sequencing) 데이터를 분석하기 위한 딥러닝 기반 변이 탐지 도구이다. 롱 리드 시퀀싱은 한 번에 수천에서 수십만 염기를 연속적으로 읽을 수 있어 구조적 변이(structural variant)나 복잡한 반복 서열을 분석하는 데 유리하지만, 염기 판독 과정에서 신호 노이즈와 염기 식별 오류가 자주 발생해 단일 염기 정확도가 낮은 것이 특징이다. Clairvoyante는 이러한 롱 리드의 높은 오류율을 보정하면서, 잡음이 많은 환경에서도 진짜 변이와 시퀀싱 플랫폼 특이적 오류를 구별하도록 설계된 멀티태스크 신경망(Multi-task Neural Network)을 활용한다.

대용량의 차세대 시퀀싱 데이터를 수동으로 품질 관리하는 것은 복잡하고 시간이 많이 소요되지만, 후속 분석(downstream analysis)의 신뢰성을 위해 필수적인 과정이다. 머신러닝 기반 도구인 SeqQscorer는 이러한 과정을 자동화하여, FASTQ 파일의 품질을 ‘고품질’ 또는 ‘저품질’로 분류한다. 구체적으로, SeqQscorer는 우선 NGS 파이프라인의 여러 단계에서 추출된 지표들을 통합하여 각 FASTQ 파일의 종합 특성 벡터를 생성한다. 추출된 지표들에는 원시 리드 분석의 품질 플래그(categorical quality flags), 유전체 매핑 통계, 리드 분포와 같은 수치형 데이터가 포함된다. 이후 랜덤 포레스트(random forest) 기반 머신러닝 모델을 학습시켜, 통합된 특징 집합에 따라 각 파일을 고품질 또는 저품질로 이진 분류한다. 이러한 데이터 기반 접근은 자동화되고 객관적인 파일 품질 지표를 제공하여, 연구자가 문제 있는 샘플을 체계적으로 거를 수 있게 하고 유전체 기반 결과의 신뢰성과 재현성을 높인다.

2.1.2 유전자 발현 예측(Gene Expression Prediction)

유전자 발현은 세포의 특성과 생물학적 반응을 결정하는 핵심 과정이다. 이는 DNA 상의 조절 서열(프로모터, 인핸서 등)과 전사인자(transcription factor)의 복합적인 상호작용에 의해 조절된다. 기존의 머신러닝 모델은 이러한 복잡한 조절 관계를 단순화하여 다루었기 때문에, 실제 발현 수준이나 비암호화 변이의 영향을 정확히 반영하기 어려웠다.

최근에는 딥러닝 기반 Sequence-to-Expression(S2E) 모델이 등장하면서, DNA 서열

데이터만으로 유전자 발현 수준을 정밀하게 예측할 수 있는 가능성이 열렸다^[15]. 이 접근법은 유전자 조절의 문법(regulatory grammar)을 데이터로부터 직접 학습해, 서열 변이 효과 분석이나 인공 조절 서열 설계에도 활용될 수 있다.

초기 연구에서는 합성곱 신경망이 주로 사용되어 전사인자 결합 부위나 염색질 접근성(DNase-seq, ATAC-seq 같은 실험으로 측정되는 DNA 접근성)을 예측하는 데 효과를 보였다. 하지만 합성곱 신경망은 장거리 유전 조절 관계를 학습하기 어렵다는 한계가 있었다.

이를 보완하기 위해 등장한 트랜스포머 기반 모델 Enformer는 원시 DNA 서열로부터 직접 유전자 발현 수준을 예측하며, 인핸서-프로모터(유전자 조절 구간 간)의 장거리 상호작용까지 학습할 수 있다는 점에서 S2E 모델의 대표적인 기반이 되었다. 이후 개발된 모델은 입력 서열 길이를 50만 bp 이상으로 확장해, 전사 개시·스플라이싱·종결 과정을 통합적으로 예측함으로써 모델의 해상도와 범위를 한층 높였다. 이와 같은 대규모 모델의 발전은 성능 면에서 큰 진전을 이뤘지만, 동시에 연산 비용 증가와 모델 해석의 어려움이라는 과제를 남겼다. 이에 따라, 소규모 MPRA(Massively Parallel Reporter Assay) 데이터 등에서 효율적으로 학습할 수 있는 경량 합성곱 신경망 모델도 여전히 중요한 대안으로 연구되고 있다.

유전자 발현 예측에는 RNA-seq, ChIP-seq, ATAC-seq, 히스톤 변형 데이터 등 다양한 유전체 데이터가 사용된다. 이러한 이질적 데이터를 함께 학습시키는 멀티태스크 학습(multi-task learning)은 예측 정확도를 높이고, 염색질 개방성과 히스톤 변형 등 생물학적 특징 간의 연관성을 반영할 수 있다. 하지만 최근 연구에서는 멀티태스크 모델이 특정 세포 유형의 특이적 발현을 충분히 포착하지 못한다는 점이 지적되며, 오히려 작고 특화된 단일 태스크(single-task) 모델이 빠르고 정확할 수 있음을 보여준다. 또한, 대규모 ATAC-seq이나 염색질 데이터를 활용한 전이학습(transfer learning)이 주목받고 있다. 사전학습(pre-training) 후, 소규모 MPRA 데이터로 파인튜닝(fine-tuning)하여 예측력을 향상시키는 방식이다.

유전자 발현 예측 모델의 성능이 향상되면서, 그 예측 과정이 실제로 무엇을 학습하고 있는지 이해하려는 시도도 활발히 진행되고 있다. In silico 돌연변이 분석이나 역전파 기반 특성 지도(attribution map)와 같은 기법을 이용하면, 어떤 염기서열이 발현 변화에 가장 큰 영향을 미치는지를 시각적으로 확인할 수 있다. 이를 통해 전사인자 모티프 간의 결합 패턴이나 배열 간격 등, 유전자 발현을 조절하는 기본 원리(‘유전자 조절 문법’)를 데이터로부터 직접 추출할 수 있다.

또한 복잡한 딥러닝 모델의 내부 구조를 생물학적으로 해석하기 위해, 모델 단순화 및

설명 가능성 향상 연구도 활발하다. 예를 들어, 예측 모델을 단순한 선형 형태로 근사화한 대리 모델(surrogate model) 이나, 내부 작동 과정을 시각적으로 설명할 수 있는 ExplainNN(설명 가능한 신경망) 등이 개발되어, 예측 정확도와 생물학적 해석 가능성을 동시에 확보하려는 노력이 이어지고 있다.

이러한 예측 결과는 대규모 실험 데이터를 통해 검증된다. MPRA나 CRISPR 간섭(CRISPRi) 실험을 활용해 모델의 예측값을 실제 발현 수준과 비교함으로써, 모델의 신뢰성을 평가하고 조절 서열의 기능을 검증한다. 다음 절에서는 원시 서열로부터 유전자 발현뿐 아니라 그 기능을 예측하고 해석하는 사례들을 살펴본다.

2.1.3 유전자 서열 기반 기능 예측(Sequence-Based Functional Prediction)

유전자 기능의 예측과 해석은 생명과학 및 생명공학 연구에서 후속 분석(downstream analysis)의 출발점이다. 전통적인 방법들은 주로 서열 상동성이나 진화적 보존성, 후성유전(epigenomic) 마커 등 외부 주석에 의존했으나, 최근에는 인공지능을 활용해 원시 DNA 서열(raw sequence)로부터 직접 기능적 결과를 모델링하는 방향으로 발전하고 있다. 과거 도구들은 주로 모티프 매칭이나 고정된 유전체 특징에 기반한 단순한 서열 패턴에만 의존했기 때문에, 대립유전자 수준(allele-specific) 또는 단일 염기 수준(single-nucleotide resolution)의 세밀한 기능 변화를 포착하는 데 한계가 있었다.

이에 최근에는 단순히 서열 정보뿐 아니라 유전자 주변의 유전체 맥락(genomic context), 즉 유전자의 위치와 인접한 유전자 구성을 고려하여 기능을 추론하려는 시도를 하고 있다. 이는 인접 유전자가 종종 같은 기능을 수행하는 원핵생물에서 특히 중요하다. 이러한 개념을 바탕으로 개발된 GeNLP는 자연어 처리 기술을 활용해 미생물 유전자의 기능을 예측하는 인공지능 모델이다. GeNLP는 유전자를 하나의 ‘단어’, 주변 유전체 영역을 ‘문장’처럼 바라보고, 이들 사이의 기능적 관계를 데이터로부터 학습한다. 즉, 유전체 데이터를 언어처럼 해석하여 유전자들이 어떤 맥락에서 함께 작동하는지를 파악하고, 그 관계를 숫자 벡터 형태로 표현한다. 수백만 개의 게놈과 메타게놈 데이터를 기반으로 학습된 word2vec 모델을 사용하는 GeNLP는 약 56만 개 이상의 유전자군을 고차원 공간에 배치하며 기존의 주석 정보가 없어도 기능이 유사한 유전자를 자동으로 탐색할 수 있다. 또한 이 플랫폼은 KEGG의 10개 기능 범주(예: 대사, 유전 정보 처리, 세포 과정 등)를 기준으로 유전자 기능을 예측하며, 계통 분포와 유사 기능 유전자의 관계를 시각화하는 웹 인터페이스를 제공한다. 결과적으로 GeNLP는 단순한 서열 비교를 넘어, 유전체 구조 속에 숨겨진 기능적 신호를 해석할 수 있는 새로운 인공지능 기반 접근법을 제시한다.

반면, GeNLP는 유전자 서열과 실제 유전자 발현 결과 간의 정량적 관계를 직접 예측하지는 않는다. 이 한계를 보완하기 위해 개발된 Enformer는 2.1.2에서 언급하였듯이, 특정 분석 목적(유전자 발현 예측)에 맞게 설계된 모델로, 활용 범위가 제한된다는 한계가 있다. 이를 극복하기 위해 Google DeepMind는 AlphaGenome을 개발하여, 대규모 유전체 데이터 기반 사전 학습을 통해 더 넓은 해석 능력을 확보했다. 기존의 모델들이 염색질 특징이나 유전자 발현 등 개별 생물학적 현상에 초점을 맞췄다면, AlphaGenome은 최대 백만 염기 길이의 서열을 처리하며, 다양한 조직과 세포 유형에서 유전형-표현형(genotype-phenotype) 간 관계를 예측한다.

AlphaGenome은 유전체 데이터를 짧은 k-mer 단위(예: 6-mer)로 나누어 벡터로 변환하고, 합성곱 신경망으로 지역적 패턴을, 트랜스포머로 장거리 의존성을 학습한다. 이러한 구조를 통해 유전체 내 조절 요소 간의 문맥적 관계와 기능 패턴을 파악할 수 있으며, 사전 학습된 모델은 변이 효과 예측, 유전자 발현 추론, 기능 요소 분류 등 다양한 생물학적 분석에 맞게 파인튜닝(fine-tuning)할 수 있다. 즉, AlphaGenome은 대규모 유전체 데이터를 활용해 유전형-표현형 연관성을 학습하는 인공지능 기반 모델로, 생물학적 서열 해석의 새로운 데이터 중심 패러다임을 제시하였다.

2.1.4 유전자 돌연변이 효과 예측(Mutation Effect Prediction)

유전체 데이터를 기반으로 유전 변이의 기능적 결과를 정확히 예측하는 것은, 합성생물학과 생물공정공학에서 우수 균주를 설계하기 위한 핵심 과제로 부상하고 있다. 최근에는 자연어 처리 기술의 발전에서 착안하여, DNA 서열을 구조화된 언어 데이터로 해석하고 염기서열에 내재된 국소적·전역적 문맥 정보를 함께 학습하는 접근이 확산되고 있다. 이러한 서열 기반 유전체 데이터 분석 흐름 속에서 합성곱 신경망, 순환 신경망(Recurrent Neural Network, RNN), XGBoost(Extreme Gradient Boosting) 등 다양한 인공지능 모델이 도입되며, 변이 효과 예측의 정밀도와 확장성을 높이고 있다.

예를 들어 Dialogue for Reverse Engineering Assessment and Methods(DREAM) 챌린지에서는 수백만 개의 무작위 프로모터 서열과 이에 따른 효모의 발현 데이터를 이용해 여러 딥러닝 모델을 비교했다. 그 결과, EfficientNet 기반 합성곱 신경망, 양방향 LSTM(Bi-LSTM), 트랜스포머 등 다양한 구조의 모델이 외부 주석 없이도 단일 염기 변이(single-nucleotide variant)가 유전자 발현에 미치는 미세한 조절 효과를 정확히 찾아낼 수 있음을 보여주었다. 이러한 모델들은 DNA 서열을 문장처럼 다루어, 원-핫 인코딩(one-hot encoding)이나 임베딩(embedding) 방식으로 데이터를 수치화한다. 이를 통해 합

성(synthetic)과 자연(native) 서열 모두에서 조절 효과를 자동으로 예측할 수 있다.

또 다른 접근으로, Chromatin Accessibility Enabled Learning Model은 염기 편집 효율(base editing efficiency)을 예측하기 위해 기계학습과 유전체 편집 데이터를 결합했다. 이 모델은 1,000개 이상의 인간 유전자 위치(locus) 데이터를 학습한 XGBoost 회귀 모델로, DNA 임베딩과 염색질 접근성 정보를 함께 사용해 예측의 정확도를 높였다. 특히 실제 실험 데이터를 활용해 학습했기 때문에 결과의 생물학적 신뢰성이 높다.

최근에는 개별 변이 하나하나의 영향뿐 아니라, 유전체 전체의 조절 관계와 유전자 간 상호작용까지 함께 분석하여 변이의 기능적 효과를 총체적으로 파악하려는 연구가 활발하다. 예를 들어, Genomic Pre-trained Network는 자가지도 학습(self-supervised learning)을 통해 순수 서열 데이터만으로 변이 효과를 학습할 수 있음을 보여주었고, Genomic Language Model(gLM)은 유전자 간의 연결 구조인 오페론(operon)과 공조절 관계(co-regulation)까지 포착할 수 있었다. 이런 연구들은 유전체 데이터 속 복잡한 조절 신호를 더 깊이 이해하고, 지역적인 조절 요소(local regulatory elements)와 유전체 전체 구조를 함께 고려하는 새로운 데이터 중심의 유전체 해석 방향을 제시하고 있다.

2.1.5 유전자 클러스터 분류(Gene Cluster Classification)

유전자 클러스터의 기능적 분류는 한 유전체 내에서 물리적으로 인접해 있으며 공통의 생물학적 기능을 수행하는 유전자 집단의 역할을 규명하는 데 초점을 맞춘다. 이 단계는 개별 유전자의 기능 분석을 넘어, 유전자들이 어떻게 협력하여 복잡한 대사 경로, 조절 네트워크, 특정 생물학적 과정을 형성하는지를 이해하는 과정이다. 그러나 유전체 데이터의 폭발적인 증가로 인해 기능이 알려지지 않은 클러스터(orphan clusters)가 대거 발견되고 있으며, 이들의 복잡한 구성과 상호작용 때문에 단순한 서열 유사성 분석만으로는 기능을 정확히 예측하기 어렵다.

이러한 한계를 해결하기 위해 여러 분석 기법과 인공지능 기술을 결합한 다층적 기능 예측 프레임워크(Rodriguez del Rio Framework)가 개발되었다. 이 방법은 먼저 공유 유전자 배열(synteny analysis)을 이용해 서로 가까이 존재하며 함께 보존된 유전자들의 관계를 파악한다. 이어서 AlphaFold2를 사용해 단백질의 새로운 3차원 구조를 예측하고, 그 구조를 이미 기능이 알려진 단백질들과 비교하여 3D 구조적 유사성 기반으로 기능을 추정한다. 이렇게 인공지능 예측과 실험 검증을 결합한 접근법은 지금까지 기능이 밝혀지지 않았던 유전체 영역에 새로운 정보를 추가하고 있다. 그 결과, 생물공정 최적화와 신호소 탐색에 활용할 수 있는 새로운 효소와 대사 경로의 자원이 빠르게 확대되고 있다.

여러 유전자 클러스터 중에서도 생합성 유전자군은 의약품, 농업, 산업용 소재 등에서 활용되는 천연물을 만드는 유전자 집합으로, 현대 생명공학의 핵심 연구 대상이다. DeepBGC는 유전체 데이터만으로 생합성 유전자군을 자동으로 식별하고 분류하는 모델이다. 이 모델은 단백질 도메인 서열(protein domain sequence)을 입력으로 받아, 데이터를 앞뒤 양방향으로 읽어 유전자 배열의 순서와 주변 관계를 동시에 학습하는 양방향 LSTM(Bi-LSTM) 신경망 구조를 활용한다. 이를 통해 각 유전자 클러스터가 가진 특정 패턴과 조합의 특징을 더 정밀하게 파악할 수 있다. 그 결과, DeepBGC는 수천 개의 미생물 유전체에서 대규모 자동화된 생합성 유전자군 탐색과 분류를 수행할 수 있으며, 기존의 규칙 기반 방법으로는 발견하기 어려웠던 새로운 BGC 계통까지 찾아낼 수 있게 되었다. 최근의 연구는 유전체 데이터에서 생합성 유전자군을 단순히 찾아내는 것을 넘어서, 이들이 어떤 방식으로 작용하고 조절되는지를 밝히는 방향으로 발전하고 있다. 이를 위해 연구자들은 BGC 안에 포함된 효소들의 촉매 기능, 특정 기질에 대한 반응 특이성, 단백질의 3차원 구조, 그리고 여러 효소 간의 상호작용을 유전체 데이터와 함께 분석하고 있다. 이러한 접근은 유전자 클러스터가 실제로 어떤 대사산물을 만들어내는지를 분자 수준에서 이해하기 위한 핵심 기반이 된다.

표 2> 유전체 분석을 위한 인공지능 모델 및 사용 데이터 개요

구분	인공지능 모델	사용된 유전체 데이터	주요 목적 및 특징
유전체 품질 평가 (Genome quality evaluation)	DeepVariant [16]	차세대 시퀀싱(NGS) 데이터(염기서열, 품질 점수 등)	CNN 기반 이미지 분류 접근으로 변이 탐지 정확도 향상, 시퀀싱 오류 제거
	Clairvoyante [17]	롱리드 시퀀싱(long-read) 데이터(PacBio, Oxford Nanopore)	멀티태스크 신경망 기반 변이 탐지, 잡음이 많은 롱리드 데이터에서 진짜 변이 식별
	SeqQscorer [18]	NGS 파이프라인 품질 지표(FASTQ 품질 플래그, 매핑 통계, 리드 분포 등)	Random Forest 기반 머신러닝으로 FASTQ 품질 자동 분류(고/저품질)
유전자 발현 예측 (Gene expression prediction)	Sequence-to-Expression(S2E) [19]	DNA 서열, RNA-seq, ChIP-seq, ATAC-seq, 히스톤 변형 데이터	딥러닝(CNN-Transformer) 기반 발현 예측, 멀티태스크·전이학습 적용
	Enformer [20]	원시DNA 서열(최대20만bp)	트랜스포머 기반, 장거리 인핸서-프로모터 상호작용 학습을 통한 유전자 발현 예측
	ExplaiNN [21]	S2E 모델의 예측 결과 데이터	모델 해석 및 설명 가능성 향상, 예측 근거 시각화
유전자 서열 기반 기능 예측 (Sequence-based functional prediction)	GeNLP [22]	대규모 유전체 및 메타게놈 서열 데이터	NLP(word2vec) 기반, 유전자를 '단어', 주변 유전체 영역을 '문장'으로 해석해 기능적 관계 학습
	Enformer [23]	원시 DNA 서열(최대20만bp)	트랜스포머 기반, 장거리 인핸서-프로모터 상호작용 학습을 통한 유전자 발현 예측
	AlphaGenome [24]	대규모 유전체 서열(최대100만bp), 다조작·다세포 데이터	CNN+Transformer 기반 사전학습 모델, 유전형·표현형 연관성 예측, 다양한 생물학적 과정에 파인튜닝 가능

생합성 유전자군 (Biosynthetic Gene Clusters, BGCs)	DREAM Challenge [25]	무작위 프로모터 서열, 효모 발현 데이터	EfficientNet-CNN, Bi-LSTM, Transformer 구조 비교·평가, 단일 염기 변이 효과 학습
	Chromatin Accessibility Enabled Learning Model [26]	인간 유전자 좌위(locus) 데이터(DNA 임베딩, 염색질 접근성)	XGBoost 회귀 기반, 염기 편집 효율(base editing efficiency) 예측
	Genomic Pre-trained Network [27] / Genomic Language Model(gLM) [28]	순수 DNA 서열 데이터	Self-supervised 학습, 유전자 간 상호작용 및 공조절 관계 학습
유전자 클러스터 분류 (Gene cluster classification)	DeepBGC [29]	단백질 도메인 서열 (protein domain sequence)	Bi-LSTM 기반, 미생물 유전체에서 BGC 자동 탐색 및 분류, 새로운 클러스터 계통 발굴
	Rodríguez del Río Framework [30]	유전자 배열(synteny), 단백질 3D 구조 데이터(AlphaFold2 예측 포함)	다층적 분석 프레임워크, 구조 기반 기능 주석, '유전체 압축물질' 영역 기능 예측

3 맺음말

합성생물학 분야는 이제 실험 중심의 연구를 넘어, 대규모 유전체 데이터와 인공지능 기반 해석 기술이 주도하는 데이터 중심의 혁신 단계로 진입하고 있다. 고품질의 유전체 데이터는 설계 정확도를 높이고, 유전자 서열 기반 기능 예측 및 변이 효과 모델은 신속한 생합성 경로 탐색과 균주 최적화를 가능하게 한다. 이러한 기술적 진전은 유전자 수준의 설계에서 시스템 수준의 생명공학으로 확장시키며, 나아가 합성생물학이 산업적 응용과 국가 전략 기술로 자리매김하는 기반을 마련한다. 결국, 유전체 데이터의 효율적 생산과 지능적 활용은 미래 합성생물학의 경쟁력을 결정짓는 핵심 요소로서, 과학적 혁신과 산업적 가치 창출을 동시에 견인할 것이다.

[참고문헌]

- [1] <https://www.ncbi.nlm.nih.gov/home/about/mission/>
- [2] <https://www.ncbi.nlm.nih.gov/genbank/>
- [3] <https://www.ncbi.nlm.nih.gov/refseq/about/>
- [4] <https://www.genome.jp/kegg/docs/relnote.html>
- [5] <https://www.uniprot.org/help/downloads>
- [6] <https://www.rcsb.org/pages/faq>
- [7] <https://doi.org/10.1093/nar/gkae1093>
- [8] <https://doi.org/10.1093/nar/gkae1063>
- [9] <https://doi.org/10.1093/nar/gkae1115>
- [10] <https://doi.org/10.1093/nar/gkad984>
- [11] <https://doi.org/10.1093/nar/gkae1060>
- [12] <https://doi.org/10.1093/nar/gkac1075>
- [13] <https://doi.org/10.1039/D5NP00059A>
- [14] <https://doi.org/10.1016/B978-0-443-24738-5.00003-3>
- [15] <https://doi.org/10.1038/s41576-025-00841-2>
- [16] <https://doi.org/10.1038/nbt.4235>
- [17] <https://doi.org/10.1038/s41467-019-09025-z>
- [18] <https://doi.org/10.1186/s13059-021-02294-2>
- [19] <https://doi.org/10.1038/s41576-025-00841-2>
- [20] <https://doi.org/10.1038/s41592-021-01252-x>
- [21] <https://doi.org/10.1186/s13059-023-02985-y>
- [22] <https://doi.org/10.1093/bioinformatics/btae034>
- [23] <https://doi.org/10.1038/s41592-021-01252-x>
- [24] <https://github.com/google-deepmind/alphagenome>
- [25] <https://doi.org/10.1038/s41587-024-02414-w>
- [26] <https://doi.org/10.1038/s41467-022-35056-0>
- [27] <https://doi.org/10.1073/pnas.2311219120>
- [28] <https://doi.org/10.1038/s41467-024-46947-9>
- [29] <https://doi.org/10.1093/nar/gkz654>
- [30] <https://doi.org/10.1038/s41586-023-06955-z>

[주요 용어 풀이]

- 차세대 시퀀싱(Next-Generation Sequencing, NGS): 대량 병렬 시퀀싱으로 유전체 데이터를 고처리량으로 생산하는 기술.
- 전장 유전체 시퀀싱(Whole-Genome Sequencing, WGS): 전장 유전체를 모두 해독하는 시퀀싱 방식.
- 고처리량(High-throughput): 실험·분석을 단위 시간당 대규모로 처리하는 특성 또는 체계.
- 생합성 유전자군(Biosynthetic Gene Cluster, BGC): 특정 2차 대사산물 합성에 관여하는 인접 유전자 집합.
- 유전체 주석(Genome annotation): 유전자·조절요소의 위치와 기능을 규정·기록하는 과정.
- FASTQ: 염기서열과 각 염기의 품질 점수를 함께 저장하는 표준 텍스트 포맷.
- 롱리드 시퀀싱(Long-read sequencing): 수 kb 이상 연속 서열을 읽어 반복구간·구조변이 해석에 유리한 시퀀싱 방식.
- 구조적 변이(Structural variant): 수십 bp 이상 규모의 삽입/결실/중복/역위/전좌 등 유전체 구조 변화.
- 단일 염기 변이(Single-nucleotide variant): 유전체 염기서열에서 단일 염기 1개가 치환되며 발생하는 서열 변화.
- RNA-seq: 전사체 서열을 분석해 유전자 발현을 정량하는 기술.
- ChIP-seq: 단백질-DNA 결합 부위(전사인자, 히스톤 수정 등)를 전장 수준에서 프로파일링하는 기술.
- ATAC-seq: 염색질 접근성을 측정해 개방된 유전체 영역을 규명하는 기술.
- DNase-seq: DNase I 민감 영역을 기반으로 염색질 개방성을 측정하는 기술.
- 염색질 접근성(Chromatin accessibility): 전사인자·효소 접근이 가능한 유전체의 개방 정도.
- 전사인자(Transcription Factor, TF): 특정 서열에 결합해 유전자 전사를 촉진/억제하는 조절 단백질.
- 프로모터(Promoter): 전사 개시 복합체가 결합하는 유전자 상류의 조절 서열.
- 인핸서(Enhancer): 떨어진 위치에서 전사를 증강시키는 조절 서열(장거리 상호작용 가능).
- 오페론(Operon): 원핵생물에서 인접 유전자들이 하나의 프로모터로 함께 전사되는 단위.
- 신테니(Synteny): 서로 다른 유전체 간 보존된 유전자 배열의 상대적 순서·관계.
- 유전체 마이닝(Genome mining): 유전체 서열에서 대사경로·BGC·조절 신호 등 기능 단서를 체계적으로 탐색·예측하는 분석 접근.

단백질 데이터 기술동향

1 머리말

단백질 기능의 정밀한 이해는 합성생물학의 핵심 중 하나이며, 이를 위해서는 단백질의 구조, 서열, 발현, 상호작용, 리간드 결합 특성 등 복합적인 정보의 통합 분석이 필수적이다. 최근 단백질 서열 데이터베이스의 확장, 고해상도 구조 데이터의 증가, RNA-seq 및 MS 기반 프로테오믹스의 발달, 그리고 광범위한 PPI 및 단백질-리간드 상호작용 데이터의 체계적 축적은 단백질 생물학 연구의 양적·질적 도약을 가능하게 하였다. 또한 AlphaFold2/3, DiffDock, PLM 기반 예측 모델 등 최신 AI 기술의 발전은 단백질 기능 해석과 설계 과정을 근본적으로 변화시키며, 단백질 공학, 대사공학, 약물 개발, 자동화 등의 연구 분야에서 새로운 패러다임을 제시하고 있다. 본 보고서는 이러한 단백질 연구에 필요한 데이터를 체계적으로 통합하여, 그 종류와 데이터베이스, 분석 도구를 정리함으로써 단백질 기반 연구 전반에 필요한 통합적 이해를 제공하고자 한다.

2 단백질 데이터의 기술 동향

2.1 단백질 서열 데이터

단백질 서열 데이터는 아미노산의 선형 배열을 나타내는 일차 구조 정보를 포함하며, 단백질 기능, 구조적 조직, 진화적 관계를 이해하는 데 필수적인 기초 자료이다. 단백질 서열은 일반적으로 FASTA 형식으로 제공되며, 헤더에는 생물 종, 유전자명, 기능 정보 등의 주석이 포함된다. 이러한 서열 정보는 상동성 탐색, 기능 예측, 도메인 분석, 변이 효과 평가, 구조 모델링 및 단백질 언어모델 기반 표현 학습의 기반이 된다.

2.1.1 단백질 서열 분석 도구

단백질 서열 분석 도구는 상동성 탐색, 다중 서열 정렬(MSA), 도메인 및 기능 예측, 딥러닝 기반 표현 분석 등 다양한 범주를 포함하며, 단백질의 기능적 특성 규명과 구조적 해석에 필수적인 역할을 수행한다. BLASTp는 seed-and-extend 전략과 통계적 유의성 평가(E-value)를 기반으로 유사 단백질을 탐색하는 가장 대표적인 알고리즘이다^[1]. PSI-BLAST

는 반복적 PSSM(position-specific scoring matrix) 업데이트를 통해 원거리 상동성 탐지 민감도를 높인 확장된 알고리즘이다.

MSA 도구인 MAFFT, Clustal Omega, MUSCLE은 보존된 잔기, 기능적 motif, 도메인 경계를 분석하는 데 널리 사용된다. MAFFT는 FFT 기반 알고리즘을 통해 고속이면서도 높은 정렬 품질을 제공한다^[2]. InterProScan은 Pfam, SMART, PROSITE 등 여러 기능·도메인 DB를 통합하여 단백질 도메인 및 motif를 예측하는 도구이다^[3]. ESM-2, ProtBERT, ProtT5와 같은 Transformer 기반 단백질 언어 모델(Protein Language Models, PLMs)은 대규모 단백질 서열 학습을 통해 고차원 embedding을 생성하며, 이는 기능 분류, 변이 효과 예측, 구조 모델링 등 다양한 downstream task에 활용된다^[4]. 최신 모델인 ESMFold 및 AlphaFold는 deep learning 기반 구조 예측의 정확도를 크게 향상시켰다.

2.2. 단백질 구조 데이터

단백질의 3차원 구조는 효소 활성 메커니즘, 기질 결합 특성, 단백질-단백질 상호작용, 돌연변이 효과, 그리고 단백질 엔지니어링 설계 등 대부분의 분자적 기능을 규정하는 핵심 요소이다. 이러한 이유로 실험적으로 결정된 구조 데이터와 AI 기반 예측 구조는 생명과학·의생명 공학 전반의 분석과 설계에서 필수적이다.

2.2.1 단백질 구조 데이터의 종류

전통적으로 단백질 구조는 X-ray 결정학, 핵자기공명분광법(NMR), 그리고 Cryo-EM(cryo-electron microscopy)을 통해 결정되어 왔다. X-ray 결정학은 수 Å 수준의 고해상도 구조를 제공하며 가장 널리 사용되는 기법이다. Cryo-EM은 대형 단백질 복합체의 구조 규명에 강점을 가지며, 최근 기술적 발전으로 해상도가 지속적으로 향상되고 있다. NMR은 용액 상에서 유연한 구조 및 저분자량 단백질 연구에 적합하다. 딥러닝의 도입 이후 AlphaFold2는 실험 수준에 근접하는 정확도로 단백질의 3차원 구조를 예측하는 데 성공하였다^[5]. AlphaFold-Multimer는 단백질 복합체 구조 예측을 확장하며, 이후 발표된 AlphaFold3는 단백질-리간드, 단백질-핵산 상호작용까지 모델링할 수 있는 범용 구조 예측 모델로 발전하였다^[6]. Meta AI의 ESMFold는 MSA 없이도 단일 서열만으로 고속 구조 예측이 가능하며^[7], RoseTTAFold 계열은 three-track 네트워크 기반으로 서열·pair3D 좌표 정보를 통합하여 구조를 생성한다^[8].

2.2.2 대표 단백질 구조 데이터베이스

PDB는 X-ray, NMR, Cryo-EM을 포함한 실험적으로 결정된 구조를 저장하는 국제적 저장소이며, 현재 22만 개 이상의 구조를 제공한다. 각 구조 파일은 PDB 또는 mmCIF 형식으로 제공되며, 원자 좌표, 해상도, R-factor, 리간드 정보, 생물학적 assembly 등의 메타데이터를 포함한다^[9]. AlphaFold DB는 2억 개 이상의 단백질에 대해 예측된 고정확도 구조를 제공하며, pLDDT 및 pTM score 기반의 신뢰도 정보를 포함한다^[10]. pLDDT>90는 매우 높은 신뢰도 구조로 간주되고, pLDDT<50 구간은 무질서(disordered) 영역일 가능성이 높다. AlphaFold3는 단백질뿐만 아니라 DNA/RNA, 리간드, 항체-항원 복합체까지 모델링할 수 있는 차세대 범용 구조 예측 모델로 설계되었다. 기존 AlphaFold2가 단백질 단일체 및 일부 복합체를 처리하는 데 중점적이었다면, AlphaFold3는 화학적 그래프 및 다중 modality 입력을 처리하는 데 강점을 가진다. 이 모델은 확률적 diffusion-like refinement 접근을 사용하여 복합체 구조의 예측 정확도를 크게 향상시켰다.

2.2.3 구조 기반 분석을 위한 주요 도구

구조 분석에는 PyMOL, ChimeraX, Rosetta, FoldX 등이 널리 사용된다. PyMOL은 PDB 및 AlphaFold 예측 구조 시각화에서 사실상 표준 도구로 기능하며, 구조적 motif 분석, 리간드 결합 포켓 시각화, 돌연변이 모델링 등 광범위한 분석에 활용된다^[11]. Chimera/ChimeraX는 Cryo-EM density map fitting 및 domain motion 분석에 강점을 갖는다^[12]. FoldX와 Rosetta는 단백질 안정성 변화($\Delta\Delta G$), 결합 친화도($\Delta\Delta G$) 및 돌연변이 설계에 널리 사용된다. 또한 fpocket은 결합 포켓 예측에, AutoDock Vina 및 Glide는 단백질-리간드 docking에, HADDOCK 및 ClusPro는 단백질-단백질 복합체 도킹 분석에 활용된다^[13]. 실험 구조 및 예측 구조의 신뢰도 평가는 여러 지표를 통해 이루어진다. Ramachandran plot: ϕ/ψ backbone angle의 분포를 통해 구조적 타당성을 검증한다. pLDDT / pTM: AlphaFold 구조의 residue-level/local 및 topology-level/global 신뢰도 지표로 사용된다^[14]. MolProbity score: clash, rotamer 품질 등 전체 구조의 기하학적 적합성을 기반으로 점수를 제공한다^[15].

2.2.4 구조 기반 분석전략

미지의 효소에 대해 AlphaFold 구조를 생성한 뒤 fpocket을 이용하여 활성 포켓을 탐색하고, 보존성 분석을 통해 효소적 기작에 중요한 잔기를 도출할 수 있다^[16]. 이후 효소 motif의 보존 여부를 확인하고, 리간드 도킹 분석을 통해 기질 결합 가능성을 평가함

로써 효소 기능을 구조적으로 추론할 수 있다. 이러한 방식은 metagenomic enzyme discovery 분야에서 특히 유용하다. AlphaFold3는 단백질-기질 복합체를 직접 예측할 수 있기 때문에 새로운 효소의 기질 특이성을 in silico에서 평가할 수 있게 한다. 예를 들어 후보 서열 집합을 AlphaFold3로 예측하여 리간드 결합 친화도 surrogate score를 산출하고, 주요 잔기의 mutational scan 결과를 고려하여 최적 후보를 선별한 뒤 소수만 실험 검증하는 방식으로 효율적인 directed evolution pipeline을 구성할 수 있다.

2.3. 단백질 도메인과 모티프 데이터

단백질의 도메인과 모티프는 서열 수준에서 단백질의 기능과 구조적 특징을 추론하는 데 있어 가장 핵심적인 단서 중 하나로, 특히 metagenomic enzyme discovery, 단백질 기능 예측, directed evolution 설계, 돌연변이 hotspot 탐색 등의 연구에서 필수적인 역할을 수행한다.

2.3.1 단백질 도메인(Domain)의 개념

단백질 도메인은 단백질 내에서 비교적 독립적으로 접히는 구조적·기능적 단위로 정의되며, 일반적으로 약 50–200 아미노산 길이를 갖는 경우가 많다. 도메인은 촉매 활성(예: kinase 도메인), 핵산 결합(예: helix–turn–helix), 단백질–단백질 상호작용(예: SH2, SH3), 세포 신호전달(예: PH, PDZ) 등 특정 기능을 수행하며, 하나의 단백질은 보통 여러 도메인 조합으로 전체 기능이 결정된다. 예를 들어 “signal peptide–lipase domain–C-terminal regulatory tail”과 같은 모듈식 구조는 분비, 촉매활성, 조절 기능이 연속된 구조 단위로 배치된 대표적 사례이다. 단백질 모티프는 보통 3–15 아미노산 정도의 짧은 서열 패턴으로, 특정 촉매 메커니즘, 금속 이온 결합, PTM 위치 등 기능적 특이성을 반영하는 시그니처 역할을 한다. 이러한 모티프는 PROSITE와 같은 데이터베이스에서 정규식 기반 패턴(pattern) 또는 위치 특이 점수행렬(PSSM, profile)로 정의되며, HXH(zinc-binding motif), GXSQ(serine hydrolase catalytic motif), N-{P}[ST]{P}(N-glycosylation site), [ST]P(proline-directed kinase 인산화 모티프)와 같은 대표 사례들이 있다^[17].

2.3.3 대표 도메인/모티프 데이터베이스

InterPro는 Pfam, SMART, PROSITE, TIGRFAM, Gene3D 등 여러 개별 단백질 서명(signature) 데이터베이스를 통합하여 단백질 패밀리, 도메인, 보존 부위를 일관성 있

게 주석하는 통합 자원이다^[18]. InterPro 엔트리는 해당 도메인의 명칭, 구성 member DB(Pfam, SMART 등), 3D 구조 패밀리(CATH, SCOP), 기능 관련 GO term을 함께 제공하여, 단일 엔트리만으로도 단백질의 기능적 범주를 비교적 정확하게 추론할 수 있다. 예를 들어, IPR029058(alpha/beta hydrolase fold)와 같이 주석된 엔트리는 “hydrolase activity” GO term 및 Pfam:PF00561 등과 연결되어 있어, 미지의 단백질이 esterase/lipase 계열일 가능성을 강하게 시사한다. Pfam은 프로파일 HMM(profile HMM)에 기반한 단백질 패밀리/도메인 데이터베이스로, UniProt 서열의 대부분을 도메인 단위로 커버할 정도로 광범위하다^[19].

각 Pfam 패밀리는 다중서열정렬(MSA)과 이에 기반한 profile HMM으로 정의되며, HMM 로고를 통해 각 위치별로 보존된 아미노산 패턴을 직관적으로 파악할 수 있다. 예를 들어, 트립신 계열 도메인(PF00089 또는 유사 family)에서는 His-Asp-Ser로 구성된 catalytic triad가 높은 보존성을 보이며, 이는 서열 수준에서 효소 메커니즘을 직접 확인할 수 있게 한다. PROSITE는 생물학적으로 의미 있는 단백질 도메인, 패밀리 및 기능 부위를 정규식 패턴(pattern)과 profile(PSSM)로 기술한 데이터베이스이다^[20]. PATTERN은 N-{P}[ST]{P}와 같은 명시적 규칙 기반 motif를 제공하며, PROFILE은 보다 유연한 통계 모델로 motif 변형을 허용하면서도 특이성을 유지한다. PROSITE는 각 signature에 대한 상세 문서를 포함하고 있어, 기능·구조·분자 메커니즘을 함께 파악하는 데 유용하다. SMART(Simple Modular Architecture Research Tool)는 주로 신호전달 관련 도메인(SH2, SH3, PH, Ras 등)에 강점을 갖는 도메인 데이터베이스로, profile HMM을 사용하여 도메인 식별 및 단백질 도메인 아키텍처 분석을 제공한다^[21]. 특히 signaling protein에서 반복적으로 관찰되는 모듈식 도메인 구성(SH3-SH2-kinase 등)을 시각화하고, 중간 보존 양상을 비교하는 데 빈번히 사용된다. MEME Suite는 관련 서열 집합으로부터 de novo로 motif를 발견하고, motif-모티프 DB, motif-서열 DB 간 검색, motif enrichment 분석 등을 수행할 수 있는 통합 툴킷이다^[22]. MEME는 Expectation-Maximization(EM) 알고리즘을 사용하여 motif 길이, 발생 위치, motif 모델을 데이터에서 직접 추정하며, 결과를 sequence logo 형태로 제공한다. 이는 metagenomic enzyme family에서 공통적인 새로운 catalytic motif를 발굴하는 데 특히 유용하다.

2.3.5 도메인·모티프 탐색 알고리즘

Pfam, SMART, TIGRFAM과 같은 도메인 DB는 profile HMM에 기반하여 도메인을 정의하며, 이를 검색하는 표준 도구가 HMMER이다^[23]. profile HMM은 다중서열정렬에서

각 위치별 아미노산 빈도와 삽입·삭제 패턴을 통계적으로 모델링하여, 원거리 상동체에 대해서도 높은 민감도로 도메인을 검출할 수 있다. HMMER의 hmmscan/hmmsearch는 query 서열을 profile HMM 라이브러리(Pfam-A.hmm 등)와 정렬하여, E-value와 domain score 기반으로 유의한 hit를 보고한다. 이 접근법은 단순 pairwise alignment에 비해 원거리 상동성 탐지에서 우수한 성능을 보이는 것으로 알려져 있다. PROSITE pattern과 같이 정규식으로 정의된 motif는 서열 문자열에 대한 pattern matching을 통해 빠르게 탐색할 수 있다.

예를 들어, protein kinase active site motif [ST]Px[HRK]는 S/T-P-x-H/R/K 패턴을 만족하는 위치를 검색함으로써, 서열 상에서 촉매 중심 근처를 직접 찾을 수 있다. 이 방법은 연산 비용이 매우 낮고 motif 위치를 명확하게 제공하지만, 허용 가능한 변이 폭이 상대적으로 작다는 한계를 가진다. PSSM 또는 PROSITE profile과 같은 위치 특이 scoring 모델은 motif 각 위치에서 20종 아미노산 분포를 확률적으로 정의함으로써, 정규식 패턴보다 훨씬 유연한 motif 인식을 가능하게 한다^[24]. 이는 motif 내 보존성이 낮은 위치, 화학적 유사 아미노산 치환 등을 자연스럽게 허용하면서도 전체 motif의 통계적 유의성을 평가할 수 있다는 장점이 있다.

PSI-BLAST에서 사용되는 PSSM도 이러한 profile 개념의 한 예이며, remote motif/도메인 탐색에 널리 활용된다. MEME는 관련 서열 집합으로부터 알려지지 않은 motif를 de novo로 추정하는 도구로, EM 알고리즘을 이용하여 motif 모델과 배경 모델을 동시에 최대우도 추정한다. 사용자는 motif 개수 및 길이 범위를 설정할 수 있으며, MEME는 enrichment가 높은 motif를 우선적으로 보고하고 sequence logo 형태로 시각화한다. 이는 metagenomic enzyme family에서 새로운 catalytic motif, metal-binding motif 등을 발견하는 데 특히 유용하며, 후속적으로 profile HMM 생성 및 기능 예측으로 연결될 수 있다.

2.3.6 도메인 주석 및 기능 예측 예시

예를 들어, 미지의 단백질 서열 'MGYDSRAGWPFSSGIGLSLGHSSLAALKAL...'에 대해 InterProScan을 수행한 결과 Pfam:PF00561(alpha/beta hydrolase), PROSITE: GxSxG catalytic motif, SMART: Abhydrolase_1 등으로 주석되었다고 가정하면, 이 단백질은 α/β hydrolase fold를 가지며 serine hydrolase형 GX SXG 촉매 모티프를 포함하는 esterase/lipase 계열 효소일 가능성이 높다고 해석할 수 있다. 이러한 정보는 구조 예측(AlphaFold) 및 활성 부위 구조 분석과 결합될 경우, substrate 범위 및 특이성 추

론으로 자연스럽게 연결된다. 다른 기능 예측 예시로 metagenomic contig에서 추출한 새로운 단백질 서열에 대해, 먼저 InterProScan으로 Pfam:PF07859(esterase/lipase), PROSITE: GXSQ catalytic motif 등의 annotation을 얻고, HMMER로 PF00561(α/β hydrolase fold)에 대한 hit를 확인함으로써 esterase 계열 효소임을 거의 확신할 수 있다. 이어서 MEME로 동일 family 서열 집합에서 HGGXH, GXSQ와 같은 추가 motif를 de novo로 추출하고, MAFFT를 이용해 다중서열정렬을 수행하여 catalytic serine 주변의 보존성과 인접 loop의 변이 패턴을 평가한다. 이를 기반으로 AlphaFold2/3 구조 예측을 수행하면 catalytic triad 위치와 기질 결합 포켓 부피를 구조적으로 평가할 수 있고, short-chain esterase로서의 특이성을 가설로 제시할 수 있다. 마지막으로 이러한 도메인·모티프·구조 정보를 바탕으로 active motif 주변, metal-binding site 및 loop region을 중심으로 돌연변이 hotspot을 선정하면 directed evolution 라이브러리 설계에 직접적으로 활용할 수 있다.

2.4. 단백질 발현 데이터

2.4.1

단백질 발현 데이터는 특정 세포 또는 조직이 주어진 조건에서 어떤 단백질을 어느 정도 수준으로 발현하는지를 정량적으로 기술하는 정보로, 기능 이해, 질환 메커니즘 규명, 신호전달 네트워크 분석, 대사경로 조절, 효소 활성 최적화, directed evolution 설계, 그리고 biofoundry 기반 자동화 공정 최적화에 이르기까지 폭넓게 활용된다. 발현 정보는 전사체(transcriptome) 수준의 mRNA abundance와 단백질체(proteome) 수준의 실제 단백질 abundance라는 두 축으로 나뉘며, 두 수준은 translation, 단백질 분해, post-transcriptional/post-translational regulation 때문에 반드시 일대일로 대응하지 않으므로 통합적인 해석이 요구된다. 발현 분석은 크게 RNA 수준의 transcriptomics와 단백질 수준의 proteomics로 구분된다. Transcriptomics는 RNA-seq 또는 microarray를 이용하여 유전자별 mRNA abundance를 측정하며, Proteomics는 주로 LC-MS/MS에 기반한 질량분석을 통해 실제 존재하는 단백질과 그 양을 정량한다. 이 두 데이터는 전사-번역-분해 과정을 포괄하는 계층적 정보를 제공하며, mRNA와 단백질 발현 간 불일치 영역은 translational control, 단백질 안정성, miRNA 조절 등 후성적·후전사적 조절 메커니즘을 탐색하는 중요한 단서가 된다.

2.4.2 Transcriptomics (RNA-seq) 데이터

RNA-seq는 고처리량 시퀀싱 기술을 이용하여 세포 내 전사체를 포괄적으로 프로파일링 하는 방법으로, 현재 transcriptomics의 표준으로 자리잡고 있다. 전형적인 분석 파이프라인은 다음과 같은 단계로 구성된다.

(A) Raw FASTQ 생성: 세포 또는 조직으로부터 추출한 RNA를 cDNA로 역전사한 뒤, 시퀀싱을 통해 수백만~수십억 개의 short read를 획득한다. FASTQ 파일에는 염기서열과 함께 각 염기에 대한 품질 점수(Q-score), 어댑터 서열 등이 포함된다. (B) Read 정렬 (Alignment/Mapping): Bowtie2, STAR, HISAT2 등의 splice-aware aligner를 사용하여 read를 참조 유전체 또는 전사체에 매핑한다. STAR와 HISAT2는 인트론-엑손 경계를 고려하는 알고리즘을 제공하여 스플라이싱 이벤트를 정확히 반영할 수 있다. (C) Count 테이블 생성: featureCounts, HTSeq 등 도구를 이용하여 각 유전자 또는 전사체 단위로 read count를 집계한다.

이 단계에서 중복 read 처리, multi-mapping read 처리 등의 옵션 설정이 중요하다. (D) 정규화(Normalization): 라이브러리 크기와 유전자 길이 차이를 보정하기 위해 TPM, FPKM, RPKM과 같은 지표 또는 DESeq2의 size factor 기반 정규화를 적용한다. (E) 차등 발현(Differential Expression, DEG) 분석: DESeq2, edgeR, limma-voom 등의 통계 패키지를 사용하여 서로 다른 조건 간 발현량 차이를 유의하게 보이는 유전자를 검출한다.

이름	데이터	특징	링크
GEO (NCBI)	RNA-seq Microarray scRNA-seq	세계 최대 발현량 저장소 다양한 조건·질환 포함	https://www.ncbi.nlm.nih.gov/gds
SRA (NCBI)	Raw RNA-seq	원시 데이터 제공	https://www.ncbi.nlm.nih.gov/sra
ENCODE	RNA-seq ChIP-seq ATAC-seq	전사 조절 epigenomics 표준 데이터셋	https://www.encodeproject.org/
GTEx	(조직별) RNA-seq eQTL	인간 조직 발현 비교 표준	https://gtexportal.org/home/
Expression Atlas (EBI)	RNA-seq Microarray	조건별 differential expression 제공	https://www.ebi.ac.uk/gxa/home
ArrayExpress (EBI)	RNA-seq Microarray	GEO와 유사 유럽계 데이터	https://www.ebi.ac.uk/biostudies/arrayexpress
TCGA	(암 조직) RNA-seq mutation	대규모 암 전사체, 유전체, 임상 통합	https://www.cancer.gov/ccg/research/genome-sequencing/tcga
CCLE	(암 세포주) RNA-seq	Drug response 연계	https://sites.broadinstitute.org/ccle/

2.4.3 Proteomics (MS 기반) 데이터

Proteomics는 실제로 존재하는 단백질 및 그 양을 측정하는 기술로, 기능적 결과를 가장 직접적으로 반영한다. 일반적인 bottom-up proteomics 워크플로우는 다음 과정을 따른다. (A) 세포/조직으로부터 단백질 추출 (B) trypsin 등 protease를 이용한 단백질의 peptide 수준 소화 (C) LC-MS/MS로 peptide 분리 및 질량분석 (D) MS/MS fragmentation 패턴을 데이터베이스(in silico digest)와 매칭하여 peptide 동정 (E) peptide 정보를 기반으로 한 protein inference (F) label-free(LFQ, iBAQ) 또는 표지 기반(TMT, iTRAQ 등) 정량.

이름	데이터	특징	링크
PRIDE (EBI)	Raw MS peptide ID Quant	세계 최대 MS 기반 proteomics 저장소	https://www.ebi.ac.uk/pride
PeptideAtlas	Spectral evidence Reprocessed proteomics	HQ spectral library Species-specific proteome map	https://peptideatlas.org
MassIVE (UCSD)	Raw MS Spectral libraries	다양한 분석엔진 결과 포함	https://massive.ucsd.edu/ProteoSAFe
ProteomicsDB	Protein abundance PTM	조직-세포주별 정량, protein turnover 제공	https://www.proteomicsdb.org/
Human Protein Atlas (HPA)	단백질-IHC Localization	항체 기반 단백질 발현 지도	https://www.proteinatlas.org/
CPTAC	암 조직 proteomics PTM	암 중심 proteogenomics 핵심	https://dctd.cancer.gov/programs/occpr
SWATHAtlas (HPA)	DIA/SWATH spectral library	targeted proteomics 설계 필수 자원	https://swathatlas.org/
PhosphoSitePlus	Phosphorylation Acetylation PTM	문헌 기반 PTM site-level DB	https://www.phosphosite.org/

2.4.4 Transcriptomics (RNA-seq) 데이터 분석 알고리즘

Transcriptomics 데이터 분석을 위해서 Alignment 알고리즘이 사용될 수 있다. STAR는 splice-aware seed-and-extend 방식과 효율적인 인덱싱을 통해 대규모 RNA-seq 데이터도 빠르게 정렬할 수 있는 알고리즘이다^[25]. HISAT2는 graph-based 인덱스를 활용하여 복잡한 스플라이싱 패턴과 다형성을 효율적으로 처리하며, 메모리 사용량이 상대적으로 낮다^[26].

이러한 도구들은 short read가 인트론-엑손 경계를 가로지르는 경우를 정확히 처리하여 전사체 구조를 복원하는데 기여할 수 있다. 전사체의 차등발현을 분석하는 도구로

DESeq2가 대표적이다^[27]. Negative binomial 분포를 기반으로 count 데이터의 과산포(overdispersion)를 모델링하고, size factor 기반 정규화 및 dispersion 추정 후 Wald test 또는 LRT를 통해 DEG를 검출한다. edgeR은 negative binomial GLM 및 quasi-likelihood 접근을 통해 low count 영역에서도 안정적인 추론을 제공한다^[28]. Limma-voom은 voom으로 count 데이터를 log-CPM과 mean-variance 관계로 변환하여 가중치가 부여된 선형모형을 적용하는 방식으로, 소규모 샘플에서도 잘 작동한다^[29]. DEG 리스트에 대해 GO enrichment, KEGG pathway 분석, GSEA(Gene Set Enrichment Analysis) 등을 수행함으로써, 발현 변화가 집중되는 생물학적 프로세스와 경로를 체계적으로 도출할 수 있다.

2.4.5 Proteomics 데이터 분석 알고리즘

MaxQuant/Andromeda은 DDA(data-dependent acquisition) 기반 proteomics 분석에서 사실상 표준 도구로, peptide-spectrum 매칭, FDR 제어, LFQ 정량 등을 통합적으로 수행한다^[30]. MSFragger는 초고속 open search 기능을 제공하는 검색 엔진으로, PTM 탐색 및 대규모 데이터셋 분석에 유리하다^[31]. Proteome Discoverer (Sequest 등)는 다양한 검색 엔진과 통합 워크플로우를 제공하는 상용 분석 플랫폼이다. 이들 알고리즘은 실험 MS/MS spectrum과 이론적 fragment ion mass 간의 일치 정도를 scoring 하여 peptide를 동정하고, 이후 protein inference 및 FDR 제어를 통해 신뢰도 높은 단백질 리스트를 생성한다. 정량적으로 차등 발현 분석을 수행하기 위해서는 Label-free 정량(LFQ) 피크 면적 또는 intensity를 기반으로 샘플 간 상대 발현량을 비교한다. 또한 TMT, iTRAQ 등 isobaric mass tag를 이용해 여러 샘플을 multiplexing하여 동일 러닝에서 정량함으로써 실험 간 변동을 줄이는 방식도 활용이 가능하다. 이 후 MSstats이나 Perseus, Limma 등의 통계 도구를 사용해서 통계적으로 의미있는 변화인지를 검증할 수 있다.

2.4.6 스트레스 조건에서의 발현 변화 분석 예시

새로운 열충격(heat stress) 조건에서 세포의 반응을 규명하고자 할 때, 우선 RNA-seq를 수행하여 DESeq2로 분석하면 HSP70, HSP90, Catalase 등의 유전자가 높은 log2FC와 유의한 p-value로 검출될 수 있다. 그 후 동일 조건에서 수행한 LC-MS/MS 데이터를 MaxQuant와 Perseus로 분석하면 HSP90과 Catalase 단백질 수준의 유의한 증가, 일부 대사효소의 감소 등이 관찰될 수 있다.

이러한 통합 분석은 heat shock protein 시스템과 ROS 처리 기작이 동시에 활성화 되고, 비필수 대사 경로가 억제되는 전형적인 stress adaptation 패턴을 보여준다. 나아가 directed evolution 실험에서 스트레스 조건 하에서 선별되는 변이체 라이브러리의 발현 패턴을 비교함으로써, 적응에 중요한 경로와 표적 단백질을 규명할 수 있다. 특히, transcriptomics와 proteomics를 genomics 및 Post-Translational Modification analysis 분석과 통합하여 proteogenomics 분석을 수행함으로써, 기존 주석에 없는 novel ORF도 발굴이 가능하다.

2.5. 단백질 번역 후 변형 데이터 (Post-translational Modification, PTM)

단백질 번역 후 변형(PTM)은 단백질이 리보솜에서 번역된 이후 일어나는 다양한 화학적 변형으로, 단백질의 활성 조절, 안정성, 구조적 구성, 세포 내 위치(localization), 상호작용 네트워크 및 신호전달 경로를 결정하는 핵심 생물학적 과정이다. PTM은 kinase·glycosyltransferase와 같은 효소 또는 비효소적 반응에 의해 부여되며, 단백질 기능 조절의 주요 스위치로 작동한다^[32]. 대표적인 PTM은 phosphorylation, glycosylation, ubiquitination, acetylation, methylation, lipidation, SUMOylation, proteolytic cleavage 등이 있으며, 이들은 단백질의 기능적 다양성을 크게 확장시키는 역할을 수행한다. 예를 들어 phosphorylation은 신호전달의 빠른 on/off 조절 메커니즘이며, glycosylation은 단백질 안정성과 분비경로를 결정하고, ubiquitination은 proteasome 경로를 통한 단백질 분해를 유발한다 [Jensen, Ole Nørregaard. “Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry.” Current opinion in chemical biology 8.1 (2004): 33-41.].

PTM종류	데이터
Phosphorylation	kinase 신호전달, 효소 활성 조절
Glycosylation	단백질 안정성 증가, 분비 및 막 단백질의 기능 유지
Ubiquitination	26S proteasome 분해 신호 제공
Acetylation	히스톤 구조 조절 및 전사 조절
Methylation	chromatin regulation 및 유전자 발현 제어
Lipidation	막 결합(membrane anchoring) 기능
SUMOylation	핵 내 이동 및 전사 억제
Proteolytic cleavage	전구단백질의 활성형 변환

2.5.1 주요 PTM 데이터베이스

PhosphoSitePlus는 phosphorylation, acetylation, ubiquitination 등 다양한 PTM에 대한 실험 기반(high-confidence) site-level 정보를 제공하는 가장 확립된 데이터베이스 중 하나이다^[33]. 이 데이터는 특정 신호경로 내에서 PTM의 역할을 규명하고 kinase-substrate 네트워크를 추론하는 데 중요한 근거가 된다. dbPTM은 다양한 PTM 유형을 통합적으로 제공하는 데이터베이스로, 실험적으로 검증된 site뿐 아니라 motif 기반 예측값도 함께 포함한다^[34]. GlyGen / UniCarbKB은 GlyGen과 UniCarbKB는 glycan 구조, glycosylation site 정보 및 glycoprotein annotation을 제공하는 glycosylation 특화 자원이다^[35]. PLMD (Protein Lysine Modification Database)은 lysine 기반 PTM(acetylation, ubiquitination, methylation 등)에 특화된 데이터베이스로, K residue 중심의 사이트 분석에 유용하다^[36]. Human Protein Atlas는 조직별 PTM 발현 정보(특히 phosphoproteomics 기반)를 제공한다^[37].

2.5.2 PTM 예측 알고리즘 및 분석 원리

PTM 분석은 단백질 단위가 아닌 개별 잔기(residue) 수준에서 수행되며, 규칙 기반 모델(motif), 확률 모델, 딥러닝 모델 등 다양한 계산 접근법이 사용된다. PROSITE, ELM 등에서는 Kinase 특이적 motif를 기반으로 phosphorylation site를 인식한다. MusiteDeep, DeepPhos, GPS와 같은 머신러닝 모델은 validated phosphosite를 학습하여, residue 주변 window sequence를 입력으로 site-specific 예측을 수행한다^[38]. Glycosylation 예측은 규칙 기반 N-glycosylation motif: N-{P}[ST]{P}으로 NetNGlyc, GlyNet, DeepGlyco 등 ML 모델을 사용 가능함^[39]. Ubiquitination/Acetylation/Methylation 예측에는 UbPred, GPS-Ubi 등 local sequence 및 surface accessibility를 활용하며^[40] DeepAcet 등 딥러닝 기반 모델은 히스톤 조절 연구에 활용되고 있다^[41].

2.5.3. PTM 분석 도구

InterProScan은 PROSITE motif 기반으로 glycosylation, phosphorylation motif 등을 자동 검출할 수 있다^[42]. ScanProsite 정규식 기반 motif를 직접 검색하는 도구로, 특정 PTM motif를 빠르게 스캔할 수 있다. MoMo (MEME Suite)은 PTM이 풍부한 단백질 집합에서 새로운 PTM motif를 탐색하는 de novo discovery 도구이다^[43]. PTM-specific 플랫폼으로 PhosphoSitePlus, GlyGen, NetPhos, GPS 등 다양한 웹서버 기반 분석 도구가 존재한다. 도구 활용 예시로 새로운 metagenomic hydrolase가 특정 조건에서만

활성 증가를 보이는 경우, PTM 관여 여부를 확인하기 위해 다음과 같은 접근이 가능하다. InterProScan으로 motif 검사하여 catalytic motif 및 phosphorylation motif 존재 여부를 파악하고 MusiteDeep을 사용한 결과를 분석한다. 구조기반 검증을 위해 AlphaFold를 활용한 잔기 분석을 수행하며 실제 활성 실험 검증을 수행으로 메커니즘을 확인함.

2.6. 단백질-단백질 상호작용 데이터 (Protein-Protein Interaction, PPI)

단백질-단백질 상호작용(PPI)은 세포 내에서 단백질이 독립적으로 기능하기보다는 복합체 또는 네트워크 형태로 상호작용하며 기능을 수행한다는 개념에 기반하며, 이는 생물학적 기능의 거의 대부분을 조직하는 기본 단위이다. PPI 분석은 단백질 기능 추정, 신호전달 경로 해독, 치료표적 발굴, directed evolution에서의 상호작용 히트존(hot-spot) 설계, 복합체 구조 모델링(예: AlphaFold-Multimer, AlphaFold3) 및 대사경로 최적화 등에 필수적이다. PPI는 기능적·물리적 관점에서 여러 유형으로 구분될 수 있다. 물리적 상호작용(binary interaction)은 두 단백질이 직접 결합하는 형태이며, 예컨대 p53-MDM2 결합이 이에 해당한다. 복합체 기반 상호작용(complex-level interaction)은 다수 단백질이 집합해 리보솜이나 프로테아좀 같은 거대 복합체를 구성하는 경우이다. 기능적 상호작용(functional association)은 직접 결합이 없더라도 동일 경로 내 기능적 연결된 단백질들 간의 관계를 뜻하며, 예컨대 키네신-튜불린(kinesin-tubulin) 신호체계가 이에 해당한다. 조절적 상호작용(regulatory interaction)은 kinase → substrate, E3 ligase → substrate 등과 같이 단백질 간 결합 및 수정(PTM)을 통해 기능을 조절하는 형태이다. 마지막으로 문헌 기반 상호작용(literature-based interaction)은 공동 발현(co-expression), 공언급(co-mention) 등 통계·텍스트마이닝 기반으로 유도된 관계이다. 이러한 유형의 다양성은 PPI 분석을 통해 단백질이 단독으로 작동하지 않고 네트워크 안에서 협업하며 기능을 수행함을 보여준다.

2.6.1. 주요 PPI 데이터베이스

BioGRID는 세계적으로 가장 널리 참조되는 PPI 실험 데이터를 수집·큐레이션하는 저장소로, 이형(Y2H), 공동면역침전(Co-IP), 풀다운(pull-down), MS 기반 정제 등 다양한 실험 기법을 포함한다. 인간, 마우스, 효모(yeast), 식물 등 수십 종을 대상으로 하며, 물리적 상호작용 뿐 아니라 유전적 상호작용(genetic interaction)까지 제공한다. 최신 버전에서는 2 900 000건 이상의 단백질 및 유전 상호작용을 포함하고 있다^[44]. IntAct는 EMBL-EBI가 운영하는 PPI 데이터베이스로, PSI-MI 표준 포맷을 적용하며 상호작용 유

형(type), 실험 방법(method), 신뢰도(confidence) 및 메타데이터를 체계적으로 제공한다. 연구자들이 신뢰도 높은 상호작용 데이터를 확보하는 데 자주 사용된다^[45]. STRING은 실험 기반 상호작용뿐 아니라 예측 기반, 공동 발현, 진화적 공출현(phylogenetic co-occurrence), 유전자 인접(gene neighborhood) 등의 다양한 데이터를 통합하여 단백질-단백질 연관들을 점수화하고 네트워크 형태로 제공하는 대표적 자원이다. 최신 버전(2025)에서는 물리적 상호작용(physical), 기능적 상호작용(functional), 조절적 상호작용(regulatory)을 분리하여 제공한다^[46]. HuRI는 인간 단백질을 대상으로 체계적인 Yeast Two-Hybrid 스크리닝을 통해 구축된 이진(binary) 상호작용 참조 지도이다. 약 53 000개의 고품질 PPI를 포함하며, 이는 과거 소규모 연구에서 수집된 것보다 약 4배 많은 수이다^[47]. 이 지도는 유전체-전사체-단백체 통합 분석을 가능하게 하여 유전자 변이(genomic variation)와 표현형(phenotype) 간의 연결고리를 제공한다.

2.6.2 PPI 검출 및 예측 알고리즘

실험 기반 방법으로 Yeast Two-Hybrid (Y2H)는 핵 내 reporter 시스템을 이용해 두 단백질이 직접 결합하면 reporter가 활성화되는 방식이다^[48]. Co-Immunoprecipitation (Co-IP)/Pull-down 방법은 항체 또는 tag-기반으로 단백질 또는 복합체를 정제하고, 그 내부 상호작용 파트너를 동정한다^[49]. Affinity Purification-Mass Spectrometry (AP-MS)는 bait 단백질을 정제하고 LC-MS/MS로 복합체의 구성을 파악하여 복합체 기반 PPI를 탐색한다^[50]. 계산 기반 방법으로는 동시 진화(co-evolution) 해석(DCA, EV-couplings), domain-domain 패턴 인식 등을 통해 상호작용 가능성 있는 단백질 쌍을 예측한다. 또한 HADDOCK, ClusPro 등의 docking 알고리즘 또는 AlphaFold-Multimer/AlphaFold3와 같은 딥러닝 모델로 단백질 복합체 구조를 예측하고, 상호작용 인터페이스(interface)를 분석한다. STRING은 gene neighborhood, co-expression, phylogenetic profile 등 여러 증거를 통합해 네트워크 형태의 상호작용을 예측한다.

2.6.5 PPI 기반 기능 추론 및 응용 사례

예를 들어 미지의 metagenomic hydrolase 후보 단백질이 있을 경우, 다음과 같은 분석 흐름을 적용할 수 있다. (A). STRING에서 후보 단백질로 네트워크를 검색하여 연관 단백질(예: ABC transporter subunit, 막 외 단백질, 지질 대사 효소 등)을 확인한다. (B) 네트워크 클러스터링 결과에서 hydrolase/esterase 클러스터, 막단백질 클러스터 등이 관찰된다면 후보 효소의 기질 특성이 지질 유사 물질(lipid-like substrate)일 가능성을 제시할 수 있다.

(C) AlphaFold-Multimer를 사용해 hydrolase + transporter 복합체 구조를 예측하여 인터페이스 안정성을 평가한다. (D) 해당 결과를 바탕으로 “이 효소는 세포 외부 지질 유사 기질을 분해하고 ABC transporter가 그 산물을 수송할 가능성이 있다”는 가설을 제안할 수 있다. (E) 이와 같은 분석은 고효율 metagenomic enzyme discovery 파이프라인 또는 대사 경로 엔지니어링 워크플로우에서 효율적인 후보 선별 및 기능 추정에 기여한다.

2.7. 단백질-리간드 상호작용 데이터

소분자(ligand)라 함은 약물(drugs), 대사체(metabolites), 효소 기질(substrates), 조효소(cofactors), 자연물(natural products) 등을 포함하며, 이들이 단백질과 상호작용함으로써 생물학적 기능을 조절한다. 단백질-리간드 상호작용은 약물 개발, 효소 공학, 센서 개발, 대사 경로 설계, 비표적(off-target) 독성 연구 등에서 핵심적인 역할을 하며, 특히 효소 공학 분야에서는 binding pocket, substrate fit, catalytic geometry 조절 순서의 기본 워크플로우를 따른다. 단백질-리간드 상호작용은 다음과 같은 활용이 가능하다. 리간드(약물)가 단백질 타깃과 얼마나 강하게 결합하는지(binding affinity), 어떤 결합 모드를 갖는지(binding mode)를 분석하여 inhibitor 또는 agonist 설계를 수행한다. 효소공학 분야에서는 효소가 기질과 결합하는 방식(binding pocket)의 구조와 특이성(specificity)을 분석하고, 이를 바탕으로 효소의 활성/선택성/안정성을 조절한다. 리간드 유도 구조변화(ligand-induced conformational change)를 이용해 생물센서(biosensor)를 설계하고, 리간드 존재 유무에 따른 단백질 구조 변화나 발현 변화를 탐지한다. 대사공학 분야에서는 리간드-단백질 결합 특성이 경로(flux) 제어의 중심이 되며, 효소의 기질 결합 특성을 조절함으로써 대사 경로의 효율을 최적화할 수 있다. 이처럼 소분자-단백질 상호작용 연구는 기초생물학부터 응용생명공학까지 광범위하게 활용이 가능하다. 단백질-리간드 상호작용 데이터 표기 방식은 화합물의 화학구조를 컴퓨터로 처리하고, 구조 기반 결합 예측·가상 스크리닝에 활용 가능하게 한다.

SMILES (Simplified Molecular Input Line Entry System):

예: `CC1=CC(=O)NC(=O)N1`

InChI (International Chemical Identifier):

예: `InChI=1S/C14H19N3O3/c1-8-5-6-12(14(17)18)10(9(8)2)...`

Molfile / SDF (Structure Data File):

원자 좌표와 결합 정보를 포함하며 docking·시뮬레이션에 필수.

2.7.1. 단백질-리간드 데이터베이스

ChEMBL은 약 300 만 건 이상의 bioactivity 데이터를 포함하는 대규모 바이오액티비티 데이터베이스이다. 여기에는 IC, EC, Ki, Kd 등의 리간드-단백질 결합 지표, 타겟-화합물 매핑, assay 조건, drug-likeness 정보 등이 포함되어 있다^[51].

예를 들어, Target: EGFR, Compound: Gefitinib(DB00317), IC = 0.033 μ M 등의 데이터가 있다. 이러한 데이터는 EGFR 억제제의 효능 평가에 활용될 수 있다. BindingDB는 특히 Kd, Ki, IC 등 리간드-단백질 결합 친화도(affinity)에 특화된 데이터베이스로, 실험 기반(binding assay) 데이터를 중심으로 한다^[52]. 타겟 단백질과 소분자 화합물 간의 결합기(affinity) 자료가 풍부하며, 데이터 수집 및 갱신이 활발하다. PDBbind는 3차원 단백질-리간드 복합체 구조(PDB)와 해당 복합체의 결합 친화도 데이터를 연결한 특수한 DB이다^[53]. 구조 기반 머신러닝/딥러닝 모델 학습에 널리 사용되며, ΔG , Ki, Kd 등의 결합 지표가 포함된다. STITCH은 단백질-화합물 상호작용 네트워크 데이터베이스로, 실험 데이터, 예측 데이터, 공동 등장(co-occurrence), 문헌(text mining) 기반 상관관계를 통합 제공한다^[54]. DrugBank은 단백질-약물 관계(target-drug)를 체계적으로 정리한 데이터베이스로, 예를 들어 Target: ACE, Drug: Enalapril, Mechanism: Competitive inhibitor 등의 정보가 제공된다^[55]. PubChem BioAssay는 수많은 생화학적 assay 결과 데이터를 포함하며, 리간드-단백질 상호작용 연구의 기초 데이터로 활용된다^[56]. 이들 데이터베이스는 서로 중복되기도 하고 상보적 역할을 수행하며, 리간드 설계, 단백질 기능 추정, 머신러닝 모델 학습 데이터로서 핵심 자원이다.

2.7.4 단백질-리간드 상호작용 분석 알고리즘

단백질-리간드 상호작용 분석 알고리즘은 Structure-based docking (물리 기반), AlphaFold를 포함한 Deep learning 기반 알고리즘으로 크게 나눌 수 있고 이들 알고리즘은 PyMOL (구조 및 상호작용 분석), ChimeraX (고급 구조/표면/궤적 시각화), Fpocket (포켓 탐지 및 특성 분석), RDKit (리간드 처리 특화), OpenMM (MD 기반 에너지 분석) 등 다양한 분석 도구와 함께 활용되고 있다.

이름	내용	특징	링크
구조기반 AutoDock Vina	최적화 기반 리간드 위치와 점수 계산	비교적 빠르고 정확한 무료 오픈소스	https://www.ebi.ac.uk/pride
구조기반 Glide	HTVS, SP 등 기반 도킹 정확도와 속도 조절	상용, 고정밀모드 가능	https://peptideatlas.org

구조기반 GOLD	Genetic algorithm 기반 리간드 유연성 고려 도킹	유연한 리간드 도킹에 강점	https://massive.ucsd.edu/ProteoSAFe
DL 기반 GNINA	Autodock Vina에 3D CNN 접목 성능향상	딥러닝 + 전통 docking 하이브리드	https://www.proteomicsdb.org/
DL 기반 DeepDock	PDBbind 등 학습 기반 ΔG /affinity를 예측	구조 기반 binding affinity 회귀 모델	https://www.proteinatlas.org/
DL 기반 DiffDock	3D diffusion 기반 리간드 pose 생성	높은 정확도	https://dctd.cancer.gov/programs/occpr
DL 기반 AlphaFold3	단백질 리간드 복합체 직접 예측	docking의 여러 단계를 대체 보완 가능	https://swathatlas.org/

단백질 생물학 데이터와 분석 기술은 빠르게 진화하고 있으며, 특히 AI 기반 구조 예측, 리간드 결합 예측, 기능 예측 모델의 등장은 기존의 생물정보학적 접근을 뛰어넘는 정밀성과 범위를 제공하고 있다. 본 문서에서 다룬 서열, 구조, 발현, PPI, PTM, 리간드 상호작용 데이터는 서로 독립된 정보가 아니라, 단백질 기능을 다층적으로 해석하기 위해 반드시 통합되어야 하는 상보적 자원들이다. 또한 BioGRID, UniProt, ChEMBL, AlphaFold DB 등 대규모 데이터베이스와 PyMOL, MAFFT, DiffDock, AlphaFold3 같은 분석 도구는 단백질 기능 연구, 효소 엔지니어링, 신약 개발, directed evolution, biofoundry 자동화 연구 등을 위한 핵심 기반을 제공한다. 본 정리는 이러한 흐름 속에서 연구자가 단백질 분석의 전체 지형을 빠르게 이해하고, 필요한 데이터와 방법론을 효율적으로 선택할 수 있도록 돕는 지침서가 되기를 기대한다.

[참고문헌]

- [1] Altschul SF, et al. Basic Local Alignment Search Tool. J Mol Biol. 215:403–410 (1990)
- [2] Katoh K, et al. MAFFT multiple sequence alignment software. Mol Biol Evol. 30:772–780 (2013)
- [3] Finn RD, et al. InterPro in 2022. Nucleic Acids Res. 50:D712–D720 (2022)
- [4] Lin, Zeming, et al. “Language models of protein sequences at the scale of evolution enable accurate structure prediction.” BioRxiv 2022 (2022): 500902.
- [5] Jumper J, et al. AlphaFold: highly accurate protein structure prediction. Nature 596:583–589 (2021)
- [6] Abramson, Josh, et al. “Accurate structure prediction of biomolecular interactions with AlphaFold3.” Nature 630.8016 (2024): 493-500.

- [7] Lin, Zeming, et al. "Evolutionary-scale prediction of atomic-level protein structure with a language model." *Science* 379.6637 (2023): 1123-1130
- [8] Baek M, et al. Accurate prediction of protein structures using a three-track neural network (RoseTTAFold). *Science* 373:871–876 (2021)
- [9] <https://www.rcsb.org/>
- [10] <https://alphafold.ebi.ac.uk/>
- [11] <https://www.pymol.org/>
- [12] <https://www.cgl.ucsf.edu/chimerax/>
- [13] <https://autodocksuite.scripps.edu/>, <https://rascar.science.uu.nl/haddock2.4/>
- [14] Mariani, Valerio, et al. "IDDT: a local superposition-free score for comparing protein structures and models using distance difference tests." *Bioinformatics* 29.21 (2013): 2722-2728., Abramson, Josh, et al. "Accurate structure prediction of biomolecular interactions with AlphaFold 3." *Nature* 630.8016 (2024): 493-500.
- [15] Williams, Christopher J., et al. "MolProbity: more and better reference data for improved all - atom structure validation." *Protein Science* 27.1 (2018): 293-315.
- [16] Le Guilloux, Vincent, Peter Schmidtke, and Pierre Tuffery. "Fpocket: an open source platform for ligand pocket detection." *BMC bioinformatics* 10.1 (2009): 168.
- [17] <https://prosite.expasy.org>
- [18] <https://www.ebi.ac.uk/interpro/>
- [19] El-Gebali, Sara, et al. "The Pfam protein families database in 2019." *Nucleic acids research* 47.D1 (2019): D427-D432.
- [20] <https://prosite.expasy.org/>
- [21] <https://smart.embl.de/>
- [22] <https://meme-suite.org/>
- [23] Eddy SR. Profile hidden Markov models. *Bioinformatics* 14:755–763 (1998)
- [24] Sigrist, Christian JA, et al. "PROSITE, a protein domain database for functional characterization and annotation." *Nucleic acids research* 38.suppl_1 (2010): D161-D166.
- [25] Dobin, Alexander, et al. "STAR: ultrafast universal RNA-seq aligner." *Bioinformatics* 29.1 (2013): 15-21.
- [26] Kim, Daehwan, et al. "Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype." *Nature biotechnology* 37.8 (2019): 907-915.
- [27] Love, Michael I., Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome biology* 15.12 (2014): 550.

- [28] Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *bioinformatics* 26.1 (2010): 139-140.
- [29] Law, Charity W., et al. "voom: Precision weights unlock linear model analysis tools for RNA-seq read counts." *Genome biology* 15.2 (2014): R29.
- [30] <https://maxquant.org/>, Cox, Jurgen, et al. "Andromeda: a peptide search engine integrated into the MaxQuant environment." *Journal of proteome research* 10.4 (2011): 1794-1805.
- [31] Kong, Andy T., et al. "MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics." *Nature methods* 14.5 (2017): 513-520.
- [32] Walsh, Christopher T., Sylvie Garneau -Tsodikova, and Gregory J. Gatto Jr. "Protein posttranslational modifications: the chemistry of proteome diversifications." *Angewandte Chemie International Edition* 44.45 (2005): 7342-7372.
- [33] Hornbeck, Peter V., et al. "PhosphoSitePlus, 2014: mutations, PTMs and recalibrations." *Nucleic acids research* 43.D1 (2015): D512-D520.
- [34] Chung, Chia-Ru, et al. "dbPTM 2025 update: comprehensive integration of PTMs and proteomic data for advanced insights into cancer research." *Nucleic acids research* 53.D1 (2025): D377-D386.
- [35] Toukach, Philip V., and Ksenia S. Egorova. "Carbohydrate structure database merged from bacterial, archaeal, plant and fungal parts." *Nucleic acids research* 44.D1 (2016): D1229-D1236.
- [36] Zhang, Weizhi, et al. "CPLM 4.0: an updated database with rich annotations for protein lysine modifications." *Nucleic Acids Research* 50.D1 (2022): D451-D459.
- [37] Uhlen, Mathias, et al. "Towards a knowledge-based human protein atlas." *Nature biotechnology* 28.12 (2010): 1248-1250.
- [38] Wang, Duolin, et al. "MusiteDeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization." *Nucleic Acids Research* 48.W1 (2020): W140-W146.
- [39] Julenius, Karin, et al. "Prediction of glycosylation sites in proteins." *Bioinformatics for glycobiology and glycomics: an introduction* (2009): 163-192.
- [40] Radivojac, Predrag, et al. "Identification, analysis, and prediction of protein ubiquitination sites." *Proteins: Structure, Function, and Bioinformatics* 78.2 (2010): 365-380
- [41] <https://github.com/Lab-Xu/DeepAcet>
- [42] Jones, Philip, et al. "InterProScan 5: genome-scale protein function classification." *Bioinformatics* 30.9 (2014): 1236-1240.

- [43] Bailey, Timothy L., et al. "MEME SUITE: tools for motif discovery and searching." *Nucleic acids research* 37.suppl_2 (2009): W202-W208.
- [44] <https://thebiogrid.org>
- [45] <https://www.ebi.ac.uk/intact/home>
- [46] <https://string-db.org/>
- [47] <https://interactome-atlas.org/>
- [48] Fields, Stanley, and Ok-kyu Song. "A novel genetic system to detect protein-protein interactions." *Nature* 340.6230 (1989): 245-246
- [49] Lin, Jer-Sheng, and Erh-Min Lai. "Protein-protein interactions: co-immunoprecipitation." *Bacterial protein secretion systems: methods and protocols*. New York, NY: Springer New York, 2017. 211-219.
- [50] Vasilescu, Julian, and Daniel Figey. "Mapping protein-protein interactions by mass spectrometry." *Current opinion in biotechnology* 17.4 (2006): 394-399
- [51] Gaulton, Anna, et al. "ChEMBL: a large-scale bioactivity database for drug discovery." *Nucleic acids research* 40.D1 (2012): D1100-D1107.
- [52] Liu, Tiqing, et al. "BindingDB in 2024: a FAIR knowledgebase of protein-small molecule binding data." *Nucleic acids research* 53.D1 (2025): D1633-D1644.
- [53] Wang, Renxiao, et al. "The PDBbind database: methodologies and updates." *Journal of medicinal chemistry* 48.12 (2005): 4111-4119.
- [54] Szklarczyk, Damian, et al. "STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data." *Nucleic acids research* 44.D1 (2016): D380-D384.
- [55] Knox, Craig, et al. "DrugBank 6.0: the DrugBank knowledgebase for 2024." *Nucleic acids research* 52.D1 (2024): D1265-D1275
- [56] Kim, Sunghwan, et al. "PubChem 2023 update." *Nucleic acids research* 51.D1 (2023): D1373-D1380.

신호·유전자조절·대사 네트워크 데이터

1 머리말

세포 내에는 수 많은 분자(유전자, 단백질, 대사물질 등)들이 복잡한 상호작용을 하고 있다. 이 모든 분자를 실험적으로 측정하기에는 어려움이 있기에 그 동안에는 주로 소수의 유전자나 단백질, 대사산물 등에 집중된 연구를 수행했었다. 그러나, 최근 고속/대량 실험 기술의 발전으로 이들을 총체적으로 관찰하는 것이 가능해졌으며, 이로 인해 생물학적 데이터가 폭발적으로 증가하고 있다.

이에 따라, 시스템생물학과 합성생물학은 기존의 단일 유전자 분석 및 조작 수준을 넘어, 최근에는 이들 간의 상호작용을 총체적으로 분석하고 재설계하는, 네트워크 중심의 패러다임으로 전환되고 있다. 이는 세포의 특징이나 기능이 단일 유전자가 아닌, 여러 분자적 사건이 통합되어 시스템 전체 수준에서 결정되기 때문이다. 따라서, 이러한 시스템 수준에서의 통합적 이해는 합성생물학과 시스템생물학 분야의 발전에 필수적이라고 할 수 있다.

본 보고서는 세포에서 측정되는 여러 계층의 데이터 중에서도 주로 많이 생산되는 네트워크 데이터인 다음의 세 가지에 초점을 맞추려 한다. (1) 세포 외 신호를 세포 내로 전달하는 과정인 신호 전달 네트워크(signaling network), (2) 전사(transcription) 프로그램을 조절하는 유전자 조절 네트워크(gene regulatory network), (3) 생화학적 효소 반응을 통해 세포의 여러 대사활동을 수행하는 대사 네트워크(metabolic network)이다. 이 세 계층은 신호 감지(signal sensing) → 신호 전달(signaling) → 유전자 조절(genetic regulation) → 대사반응(metabolism)의 일반적인 방향성을 가지지만, 실제로는 상황 의존적이고 복잡하며 반복적인 계층간 상호작용을 통해 연결되는 특징을 가지고 있다.

일반적으로, 각 네트워크는 독립적인 데이터 모달리티를 기반으로 분석되는데, 신호 전달 네트워크는 인산화 단백질체(phosphoproteome) 및 시계열 인산화 단백질체(time-series phosphoproteome), 유전자 조절 네트워크(regulatory network)는 전사체(transcriptome), 대사 네트워크(metabolic network)는 대사 반응 네트워크(reaction-metabolite network) 및 대사체/대사 흐름체(metabolome/fluxome)를 이용한다. 인산화 단백질체(phosphoproteome and time-series phosphoproteome) 데이터는 신호 전달 단백질인 인산화 효소(kinase)와 이의 기질 단백질(substrate protein)의 관계에 관한 정보를 제공하며, 단백질의 인산화(phosphorylation) 정보를

기반으로 하고 있다. 유전자 조절 네트워크는 보통 전사인자(transcription factor)와 이의 조절을 받는 유전자가 연결된 형태이며, 주로 유전자 조절의 인과적 관계를 규명하기 위해 실험 조건을 교란(perturbation) 시키거나 다양한 환경 변화 조건 아래에서 전사체 데이터(transcriptomic data)를 수집하고 분석하여 유전자 조절 네트워크를 추론한다. 혹은, 기존의 유전자 조절 네트워크가 세포의 변화에 어떻게 영향을 미치는지를 규명하기 위해 네트워크를 전사체 데이터로부터 분석하기도 한다. 대사 네트워크는 대사 산물(metabolite)과 효소 반응(enzymatic reaction)을 기반으로 한 그래프 또는 하이퍼그래프로 표현되며, 유전체(genome) 분석을 통해 효소-반응을 추론하여 대사 네트워크를 구축하기도 하고, 대사체나 대사 흐름체 데이터를 이용하여 대사 네트워크가 어떻게 활성화되고 변화하는지를 분석하기도 한다.

이런 데이터 규모의 확장에도 불구하고 실험 방법의 한계로 인해 여전히 제한된 데이터만 측정(효소 활성 측정 등은 여전히 고속/대량 실험이 어려움) 가능한 경우가 많아 분석에 어려움이 있다. 또한 각 네트워크 내의 요소와 그들 간의 연결이 서로 다른 의미(유전자 발현 조절 혹은 효소 반응 등)를 갖고 있기에, 여러 계층의 네트워크를 통합 분석하는 것 역시 쉽지 않다. 이런 이유로 현재의 연구는 주로 개별 네트워크의 분석과 설계에 집중되어 있다.

이러한 배경에서 이 보고서는 다음과 같은 내용을 다루고자 한다. 먼저 각 계층(앞서 설명한 3가지 네트워크)과 관련된 데이터 유형을 정의한 후, 데이터 생산을 위한 실험방법, 현재까지 구축된 데이터베이스(database) 및 이를 이용한 분석 사례를 소개하고자 한다.

2 신호 전달 네트워크(Signaling network)

2.1 신호 전달 네트워크 데이터

신호전달 네트워크는 외부 자극을 수용체가 인식하고, 연쇄적인 인산화/탈인산화 과정을 통해 세포 내부의 상태 변화를 유도하는 핵심 시스템이다. 그러나 이 신호 전달 과정은 단순한 직렬 구조가 아니며, 세포의 조건과 상태에 따라 경로(pathway)의 활성 정도와 방향성이 복잡하게 달라지기 때문에 그 전체적인 동적 특성을 분석하는 데 큰 어려움이 있다. 최근 인공지능 기술이 급격히 발전하면서 이러한 복잡성을 해결할 새로운 가능성이 열리고 있는데, 특히, 단백질 서열과 대규모 시계열 인산화 단백질체 데이터를 분석하여,

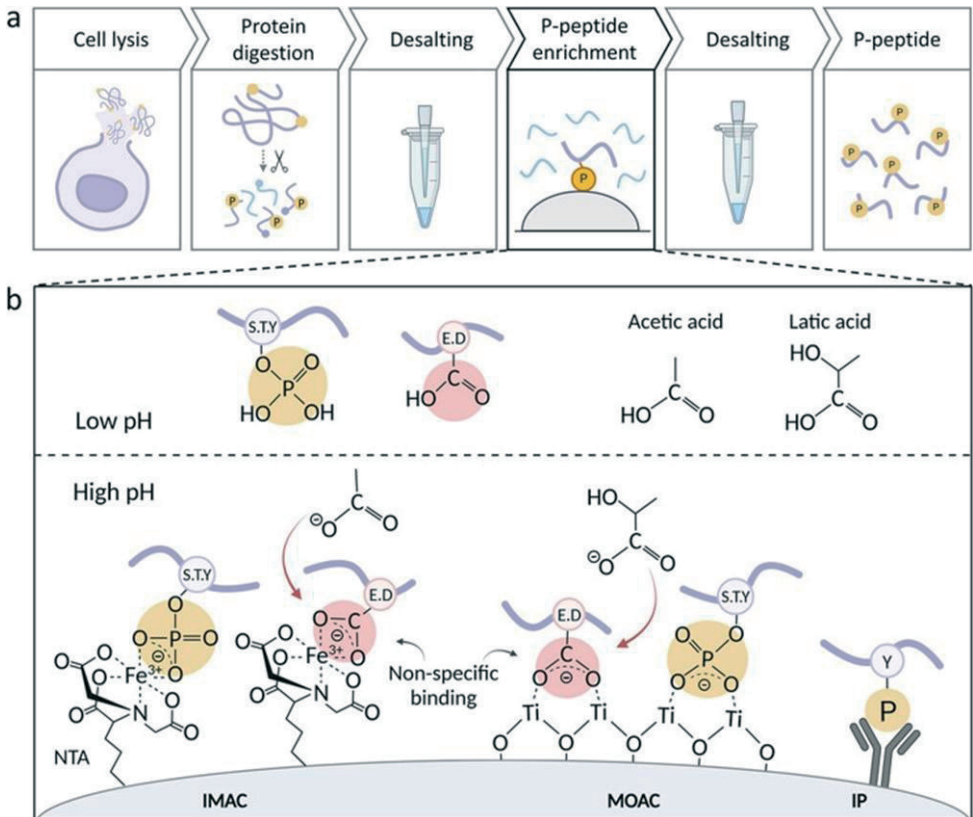
기존에 알려지지 않았던 인산화 효소-기질 상호작용과 신호 전달 경로를 추론하는 연구 결과가 새롭게 보고되고 있다. 여기에서는 인공지능 기술을 활용한 최신 연구 사례를 살펴보고, 이 과정에서 인산화 단백질체 데이터가 어떻게 네트워크 재구성의 핵심 정보로 활용되는지 소개하고자 한다.

2.2 데이터의 생산

2.2.1 인산화 단백질체 데이터(phosphoproteome)

인산화 단백질체 데이터는 세포를 용해 후 단백질을 단백질 분해효소(protease, 주로 trypsin/Lys-C)로 분해하고, 조각난 펩타이드(peptide) 중 인산화된 펩타이드(phosphopeptide)만 선택적으로 분리한 뒤 액체 크로마토그래피-탠덤 질량 분석법(LC-MS/MS)로 동정/정량하는 과정을 통해 생산된다(그림 1).

그림 1> 일반적인 인산화 단백질체 시료 전처리 과정



Muneer et al (2025). Proteomics, 25(1-2), e202400087.

인산화 펩타이드를 분리하는 가장 널리 쓰이는 방식은 IMAC(Fe^3/Ga^3)^[1], MOAC($\text{TiO}/\text{Ti-IMAC}$, Zr-IMAC)^[2], 그리고 phosphotyrosine(pTyr) immunoprecipitation^[3]이다. IMAC/MOAC의 경우, 산성 펩타이드(acidic peptide)가 비특이적으로 결합하는 것을 막기 위해 낮은 pH(2-2.5)에서 acetic/lactic acid 등의 유기산을 첨가하는 조건을 사용하며, 이는 phosphate와 산성 아미노산(acidic amino acid)의 pKa 차이를 이용해 선택성을 높이는 방법이다^[4, 5]. 이 방법은 아미노산 서열 뿐만 아니라 정확히 어느 아미노산에 인산기(phosphate group)가 붙어 있는지 그 위치까지 특정할 수 있는 장점이 있다. 인산화된 타이로신(pTyr, phosphorylated tyrosine)을 이용한 분리 방법의 경우, 인산화된 타이로신을 인식하는 anti-pTyr 항체(antibody)를 이용하므로 인산화된 타이로신만 인식하는 제한이 있지만 IMAC/MOAC에 비해 상대적으로 높은 민감도를 보이는 것이 특징이다^[6].

2.2.2 시계열 인산화 단백질체 데이터(time-series phosphoproteome)

시계열 인산화 단백질체 데이터는 리간드 등 외부 조건에 변화를 준 시점을 기준으로 짧게는 수 초 혹은 수 분, 길게는 수 시간대의 다중 시점에서 인산화 단백질체 샘플을 수집하고 분석하여 얻게 된다. 각 시간 지점에서 얻은 인산화 변화를 상대적으로 비교하여 인산화 효소-기질의 반응이 어떻게 흘러갔는지를 시간에 따라 재구성할 수 있게 된다.

2.3 관련 데이터베이스

신호 네트워크 분석을 위해서는 인산화 부위가 어디인지에 대한 사전정보, 어떤 인산화 효소에 의해 어떤 기질 단백질이 인산화 되는지에 대한 사전정보, 그리고 주요 신호 전달 단백질의 경로(pathway) 정보를 이용하여 실험 데이터를 분석하게 된다. 즉, 분석을 위해서는 여러 정보가 필요하므로 많은 연구자들이 이들 정보를 수집하여 통합한 여러 데이터베이스를 구축해서 다양한 분석에 활용할 수 있도록 제공하고 있다(표 1).

PhosphoSitePlus는 실험으로 관측된 인산화가 일어나는 아미노산 위치(phosphosite)와 검증된 인산화 효소(kinase)-기질 관계를 체계적으로 제공하고 있다^[7]. SIGNOR(The SIGNaling Network Open Resource)는 신호전달 효소가 기질 단백질의 활성을 증가시키는지 아니면 억제시키는지에 대한 명확한 정보를 제공하고 있기에, 신호전달의 방향성과 효과 분석을 위한 네트워크 기반 분석에 많이 활용된다^[8]. Reactome Pathway Database는 세포 내에서 일어나는 반응(reaction)을 기본 단위로하여 검증된 다양한 경

로(pathway) 정보를 제공하는 데이터베이스이다^[9]. 이 데이터베이스를 이용하면 실험으로 측정된 인산화 정보를 경로라는 맥락(pathway context)에서 분석하고 해석할 수가 있다. OmniPath는 160개 이상의 데이터베이스를 통합하여 구축한 대규모 분자간 상호작용 및 경로 데이터베이스이며, 이 안에는 신호전달 효소-기질 네트워크가 포함되어 있다^[10]. 이때, 단순히 인산화뿐만 아니라 다양한 번역 후 변형(post-translational modification, PTM) 정보가 포함되어 있어 신호전달의 과정 뿐만 아니라 구체적인 효소반응을 이용한 다양한 분석에 OmniPath 데이터베이스를 활용할 수 있다. 또한, 인간(human) 뿐만 아니라 생쥐(mouse)와 쥐(rat)에서의 네트워크 정보를 포함하고 있어, 인간 중심의 지식을 다른 종으로 확장하여 분석하거나, 모든 종을 포함하는 광범위한 네트워크를 구축하는 데 활용될 수도 있다. PRIDE는 ProteomeXchange 컨소시엄의 핵심 공개 저장소로, 전 세계의 LC-MS/MS 데이터와 정량 결과를 제공하고 있다^[11]. 연구자들은 이 방대한 공용 데이터를 활용하여, 기존 인산화 단백질체 데이터를 재분석하거나 여러 데이터를 통합 분석하여 새로운 지식을 발굴하는 등의 연구를 수행할 수 있다^[11,12]. 이 외에도 신호전달 네트워크와 관련된 다양한 데이터베이스가 존재하여 다양한 분석에 활용되고 있다(표 1).

표 1> 신호전달 네트워크 분석에 활용되는 다양한 데이터베이스

구분	데이터베이스	주요 내용 및 데이터	데이터 규모	주요 용도
PTM 위치정보 및 효소-기질 관계	PhosphoSite Plus ^[13]	• 실험 기반의 인산화 위치(phosphosites) 정보와 • 검증된 인산화 효소-기질(kinase-substrate) 관계	약 50만건의 건의 PTM 정보	인산화 효소(kinase)-기질 네트워크의 기본 구조를 만드는 데 사용
	DEPOD ^[14]	• 검증된 인간의 탈인산화 효소-기질 관계(phosphatase-substrate) 정보	200여 개의 탈인산화효소 정보	탈인산화 효소(phosphatase)-기질 네트워크의 기본 구조를 만드는 데 사용
	dbPTM ^[15]	• 실험으로 검증된 PTM과 위치정보 및 예측된 PTM의 통합된 데이터베이스	실험으로 확인된 약 300만건의 PTM 정보	다양한 PTM 정보를 네트워크에 통합하는데 사용
신호전달 경로 정보 (Signal in gpathway)	SIGNOR ^[16]	• 검증된 단백질-화합물·외부 자극-세포표현형 간의 인과관계와 경로 정보 제공	4만건의 관계 정보 및 1.4만건의 논문 정보	신호전달의 방향성과 효과 분석을 위해 사용
	OmniPath ^[17]	• 160개 이상의 데이터 자원이 통합된 대규모 신호전달 네트워크	1백만개 이상의 효소-PTM/단백질-단백질/단백질 복합체 관계 정보	인간 중심의 신호전달 네트워크를 다른 종으로 확장하여 분석하는데 사용
	Reactome ^[18]	• 검증된 세포 내 단백질 기반 반응정보 및 경로정보 데이터 베이스	인간을 포함한 15종, 총 9만개의 단백질, 10만개의 단백질 복합체, 9만개의 반응(reaction) 정보, 2만개의 경로 정보	다양한 효소반응 및 경로 분석에 활용
인산화 단백질체 데이터	PRIDE ^[19]	• 인산화 단백질체를 포함한 MS 기반 단백질체 실험데이터	4만건의 실험 데이터	기존 인산화 단백질체 데이터를 재분석하거나 여러 데이터를 추가해 분석하는 용도로 사용
	MassIVE ^[20]	• MS기반 실험 데이터	20만개의 단백질, 1000만 개의 펩타이드 등이 포함된 2만건의 실험데이터	

2.4 신호전달 네트워크의 분석

2.4.1 인공지능을 활용한 비모델 종의 신호전달 네트워크 재구성

대부분의 생물학 연구는 인간(human)이나 생쥐(mouse)와 같은 주요 모델 생물 종(model species)에 집중되어 있다. 그 결과, 이들 종에서는 어떤 단백질이 신호전달에 관여하고 그들의 기질이 무엇인지에 대한 정보가 많으며, 이를 바탕으로 한 신호전달 네트워크 또한 잘 구축되어 있다. 하지만 연구가 상대적으로 부족한 비모델 종(non-model species)의 경우, 이러한 정보가 거의 알려져 있지 않다. 따라서 비모델 종 연구의 첫 번째 과제는, 신호전달 네트워크를 구성하는 핵심 단백질, 즉 노드(node)를 먼저 식별하고, 이들 간의 상호작용인 엣지(edge)를 규명하여 전체 네트워크를 재구성하는 것이다. 최근 이러한 문제를 해결하기 위해 인공지능 모델이 활발히 도입되고 있다.

Van den Broeck 등(2023)은 단백질의 아미노산 서열 정보만을 이용하여 그 기능을 예측하는 PF-NET 모델을 개발했으며, 기능의 종류에는 인산화/탈인산화가 포함되어 있어 주어진 단백질이 신호전달 요소로서 네트워크를 구성하는데 포함될 수 있을지를 예측했다^[21]. 구체적으로, PF-NET은 단백질 서열을 입력받아 1D-Convolutional Neural Network(CNN)으로 서열의 주요 모티프(motif)를 추출하고, 어텐션(attention) 및 biLSTM 메커니즘을 통해 서열 전반의 장거리 의존성(long-range dependencies)을 통합했다. 이 모델을 통해 사전 정보 없이 단백질 서열만으로도 해당 단백질이 인산화나 탈인산화 효소로 기능할지 여부를 예측하여, 네트워크의 잠재적 구성 요소를 새롭게 발굴할 수 있었다.

네트워크의 ‘노드’가 식별되면, 다음 단계는 이들 간의 총체적인 관계, 즉 ‘엣지’를 규명하여 네트워크를 구성하는 것이다. 비모델 종의 경우, 이 관계(예: 누가 누구를 인산화시키는지)에 대한 정보 역시 없으므로, 실험조건이 교란(perturbation)된 후 시스템의 반응을 관찰하여 얻는 시계열 인산화 단백질체 데이터를 먼저 확보하고, 단백질 간의 동적 관계를 추론하는데 이 데이터를 활용할 수 있다. Van den Broeck 등(2023)의 연구에서는, 이 시계열 데이터를 동적 베이저안 네트워크(Dynamic Bayesian Network)와 같은 통계 모델로 분석하여, 단백질 간의 인과 관계를 추론하고 전체 신호전달 네트워크를 재구성하였다.

요약하자면, 비모델 종에 대한 사전 지식이 부족한 상황일지라도, (1) 인공지능으로 단백질 서열에서 기능(노드)을 예측하고, (2) 시계열 인산화 단백질체 데이터로 관계(엣지)

를 추론하는 두 단계의 접근 방식을 통해, 새로운 신호전달 네트워크를 재구성하는 것이 가능하다.

2.4.2 사전 지식을 결합한 RNN 기반 리간드-전사조절인자(TF) 관계 예측

합성생물학 분야에서 세포 기능을 네트워크 수준에서 설계하기 위해서는, 외부 자극(예: 리간드, pH 등)의 조합이 복잡한 내부 신호망을 거쳐 최종적으로 어떤 전사인자(transcription factor, TF) 활성 패턴으로 수렴하는지를 예측할 수 있는 도구가 필수적이다. 이런 도구 중 하나가 Large-scale knowledge-EMBedded Artificial Signaling-networks(LEMBAS) 모델이다^[22]. LEMBAS는 순환신경망(Recurrent Neural Network, RNN)을 활용하여 대규모 단백질체 데이터로부터 세포 내의 동적 신호 전달 과정을 모델링하고 재구성할 수 있다. LEMBAS의 핵심 전략은, 일반적인 기계학습의 ‘블랙박스’ 모델을 사용하는 대신 이미 알려진 생물학적 사전 지식을 RNN 구조에 직접 내장하는 것이다. 구체적으로, Michaelis-Menten 거동을 근사하는 전용 활성화 함수를 만들고, 두 노드 간의 관계가 활성인지 억제인지에 따라서 + 및 -로 부호를 부여하도록 하여 RNN의 가중치를 줌으로써 실제 신호전달 네트워크를 모사하는 방법이다.

이 덕분에, 신호가 음수가 되거나 비현실적으로 높게 되는 과포화 상태가 되는 것을 방지하여 모델을 해석 가능한 상태로 유지할 수 있는 장점이 있다. 이러한 생체 시스템의 모사를 통해 신호 흐름을 RNN으로 정확하게 모사할 수 있었으며, 결국 리간드 농도에서 전사인자(TF) 활성에 이르는 과정을 예측할 수 있었다.

3 유전자 조절 네트워크(Gene regulatory network) 분석

3.1 유전자 조절 네트워크 데이터

유전자 조절 네트워크는 외부 혹은 내부의 신호를 감지하고, 이에 맞는 반응을 위해 필요한 단백질을 생산하도록 제어하는 핵심 중간 계층의 역할을 한다. 즉, 외부 신호(예: 리간드)가 신호전달 네트워크를 통해 전달되면, 이 신호는 최종적으로 유전자 조절 네트워크와 연결되며, 이 네트워크는 특정 유전자의 발현을 조절하여 필요한 단백질을 새롭게 생성하거나 불필요한 단백질의 생산을 중단시켜 궁극적으로 세포 내 변화를 유도하게 된다.

그러므로 유전자 조절 네트워크의 분석을 위해서는 네트워크의 구성요소인 유전자가 현재 어떤 상태에 있는지를 파악하는 것이 매우 중요하며, 이 ‘상태’는 유전자 발현으로 이어지므로, 발현의 최종 산물인 단백질이나 중간 단계인 mRNA를 측정하여 알아낼 수 있다. 다만, 단백질체의 경우 실험이 복잡하고 정량화할 수 있는 대상이 제한되는 단점이 존재하기에^[23] 상대적으로 고속 및 대량 분석이 가능한 RNA 기반의 실험을 통해 전사체 데이터를 확보하고 이를 활용하는 것이 일반적이다.

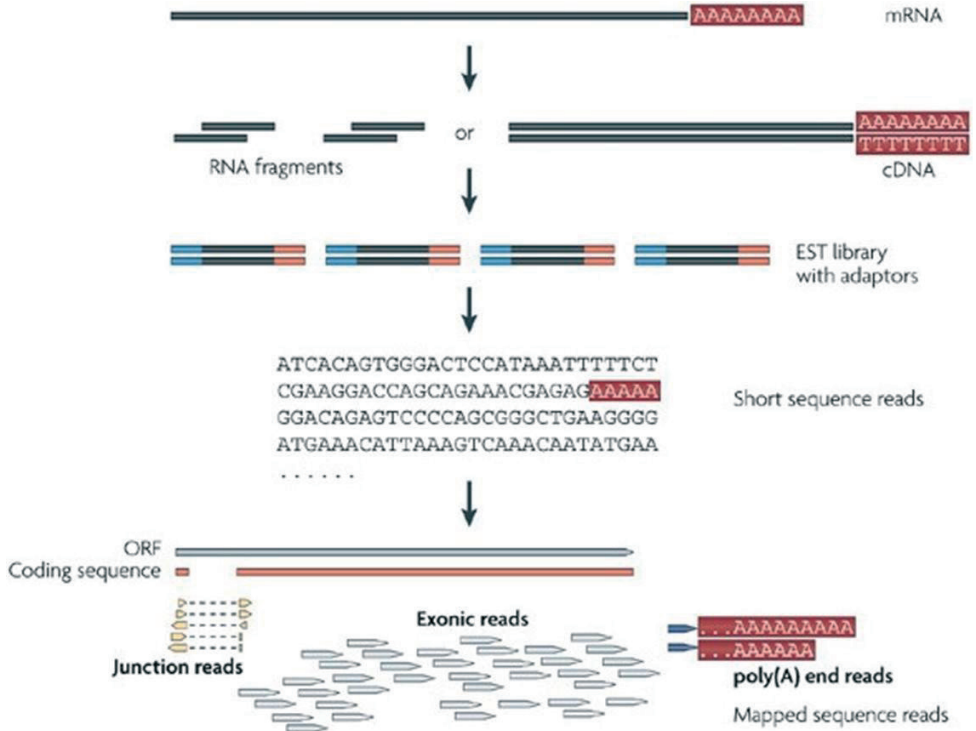
3.2 데이터의 생산

유전자 조절 네트워크 분석을 위해 가장 광범위하게 생산되는 데이터는 전사체이며, 최근 전사체 데이터 생산을 위해 주로 이용되는 실험 방법은 RNA-Sequencing(RNA-seq)이다. RNA-seq은 세포나 조직 내에 존재하는 RNA 분자를 차세대 시퀀싱 기술로 분석하여, 유전자의 발현 수준을 총체적으로 파악하는 핵심 기술이다. 전체 실험 과정은 크게 (1) 라이브러리 제작, (2) 시퀀싱, (3) 데이터 분석의 세 단계로 나눌 수 있다. 첫 번째 단계인 라이브러리 제작에서는, 분석하고자 하는 세포나 조직에서 전체 RNA를 분리 및 정제하고, rRNA와 같이 분석에 불필요한 RNA를 제거하는 과정을 거치게 된다. 이렇게 선별된 RNA는 시퀀싱에 적합하도록 수백 bp 정도의 작은 조각으로 만들어지고, 역전사(reverse transcription) 과정을 통해 각 RNA 조각을 안정적인 상보적 DNA(cDNA)로 변환된다. 합성된 cDNA 조각들의 양쪽 말단에는 특정 서열인 어댑터(adapter)를 붙임으로써, PCR을 통해 라이브러리의 양을 시퀀싱에 충분하도록 증폭하게 된다. 어댑터는 다음 단계인 시퀀싱에서 시퀀싱 기계가 이를 인식하고 시퀀싱을 수행하는데도 이용된다.

두 번째 단계는 시퀀싱이며, 각 cDNA 조각의 서열을 대량으로 읽어낸다. 시퀀싱 결과, 수천만 개에서 수억 개에 달하는 짧은 염기 서열 조각(short reads) 데이터가 생성된다.

세 번째 단계는 시퀀싱으로 얻은 대량의 서열 조각(reads)들을 기준(reference)이 되는 유전체(genome) 또는 전사체(transcriptome) 서열과 비교하여 어떤 유전자의 어느 부분에 위치하는지를 알아내게 된다. 그리고, 해당되는 유전자와 매핑이 되는 서열 조각(read)이 몇 개나 되는지를 파악하여 그 유전자가 얼마나 발현되고 있는지를 알아내게 된다. 유전자 발현 정도 외에도, 서열 조각(read)이 엑손(exon)에서 왔는지, 엑손-엑손 접합부(junction)에 걸쳐 있는지 등을 파악하여, 유전자의 구조와 다양한 대체 스플라이싱(alternative splicing) 양상을 알아낼 수도 있다.

그림 4> 일반적인 RNA-seq 워크플로우: Wang et al (2009). Nature Reviews Genetics, 10(1), 57-63.



3.3 관련 데이터베이스

유전자 조절 네트워크를 추론하고 분석하기 위해서는 RNA-seq 데이터뿐만 아니라 여러 유형의 데이터베이스를 통합적으로 활용하는 것이 필요하다. 먼저, 유전자 조절 네트워크 분석의 기본 입력이 되는 RNA-seq 발현 데이터가 필요한데, RNA-seq을 포함한 많은 전사체 데이터는 GEO와 ArrayExpress 같은 공용 데이터베이스에 통합 수집되어 있다^[24, 25]. 하지만 이 데이터베이스들은 주로 실험으로 얻은 원시 데이터(raw data)를 저장하며, 원시 데이터를 생산한 연구자들이 각자의 방법과 도구로 데이터를 분석하기 때문에, 이렇게 가공된 데이터는 모두 제각각이라는 문제점이 있다.

recount3나 ARCHS4와 같은 데이터베이스는 이러한 문제를 해결하기 위해서, 여러 저장소의 원시 전사체 데이터를 수집한 뒤, 모두 일관된 방법으로 원시데이터의 값을 재처리한 가공 데이터를 제공한다. 이 덕분에 연구자들은 서로 다른 실험의 데이터를 직접 비교

하며 통합 분석을 수행할 수 있다.

유전자 조절 네트워크를 추론하거나 분석하기 위해서는 전사체 데이터만으로는 부족하며, 네트워크의 구조와 조절 관계를 규명하기 위한 보조 데이터가 필수적이다. 이러한 보조 데이터는 다음과 같은 것들이 있는데, 첫째, 전사인자가 DNA의 어느 위치에 결합하는지에 대한 정보(TF-DNA 결합 위치)가 있다. 이 정보는 ENCODE, ReMap, ChIP-Atlas 등에서 제공하며, 이는 주로 ChIP-Seq 또는 ATAC-seq 실험 데이터로부터 확보된 것이다^[26, 27, 28]. 둘째, 전사인자가 인식하고 결합하는 특정 DNA 서열, 즉 모티프(motif) 정보가 필요하다. 이 정보는 JASPAR나 HOCOMOCO와 같은 데이터베이스에서 제공한다^[29, 30]. 셋째, 문헌이나 실험을 통해 이미 검증된 전사인자와 타겟 유전자 간의 관계(예: 활성화 또는 억제) 정보가 필요하다. 이러한 사전 지식은 TRRUST나 DoRothEA와 같은 데이터베이스에서 제공하고 있다^[31, 32].

표 2> 유전자 조절 네트워크 분석을 위한 데이터베이스

구분	데이터베이스	주요 내용 및 데이터	데이터 규모	주요 용도
전사체 실험 데이터	GEO [33]	• 전사체 실험 데이터	8백만건 이상의 실험 데이터	전사체 데이터를 재분석하거나 여러 데이터를 통합 분석하는데 사용
	ArrayExpress [34]	• 전사체 실험 데이터	30,000건 이상의 실험 데이터	
	recount3 [35]	• GEO/ArrayExpress에 존재하는 실험데이터가 일관 방법으로 처리되어 가공된 전사체 데이터	인간 30만건, 생쥐 40만건의 재처리된 전사체 데이터	
	ARCHS4 [36]		총 35,000개의 재처리된 전사체 데이터	
전사인자 - 결합DNA 위치 정보	ENCODE [37]	• TF ChIP-seq/ATAC-seq 데이터를 포함한 다양한 유전체 데이터	3만여 건의 실험 데이터	전사인자 결합부위 또는 타겟 유전자 탐색에 활용
	ReMap [38]	• 여러 프로젝트에서 생산된 TF ChIP-seq 통합 데이터	3천개의 전사인자, 14,000건의 ChIP-seq 데이터	
전사인자 - 결합DNA 서열 정보	JASPAR [39]	• 중복이 제거된 고품질의 전사인자 결합 모티프 서열 데이터	2천여개의 모티프 프로파일(profile) 정보	전사인자의 DNA결합서열을 분석, 예측하는 데이터로 활용
	HOCOMOCO [40]	• 인간과 생쥐(mouse)에 특화된 고품질 전사인자 결합 서열 정보	약 천 개의 인간 전사인자와 800여개의 생쥐 전사인자 정보	
전사인자-타겟 유전자 정보	TRRUST [41]	• 문헌 기반으로 검증된 전사인자-타겟유전자 관계 및 조절 정보(활성/억제)	800개의 인간 전사인자에 대한 약 8000건의 전사인자-타겟유전자 상호작용 데이터	전사인자와 타겟유전자 정보를 이용한 네트워크 구축 및 분석에 활용
	DoRothEA [42]	• 신뢰도 정보가 추가된 전사인자-타겟유전자 정보	20,244개의 gene들을 타겟으로 하는 1395개의 TF 데이터	

3.4 유전자 조절 네트워크의 분석

3.4.1 인공지능을 이용한 유전자 조절 네트워크의 재구성

유전자 조절 네트워크 역시 신호전달 네트워크와 마찬가지로 노드(node, 전사인자 혹은 타겟유전자)와 엣지(edge, 전사 조절 관계)로 구성되어 있다. 이와 유사하게, 컴퓨터 공학 분야에서 다루는 그래프 역시 유사하게 노드와 엣지로 구성되어 있기에 기존의 그래프를 분석하는 다양한 기술이 동일하게 세포의 네트워크를 분석하는데도 자주 활용된다. 최근에는, 인공지능 기술이 발전함에 따라 유전자 조절 네트워크 모델링 역시 그래프 신경망(Graph Neural Network, GNN)을 많이 활용하고 있는 추세이다. 이 보고서에는 GNN을 이용하여 사전정보 없이도 유전자 조절 네트워크를 재구성하는 인공지능 모델인 MTLGRN을 소개하고자 한다.^[49].

MTLGRN은 유전자 조절 네트워크에서 엣지, 즉 전사인자와 표적유전자의 관계를 찾도록 학습된 모델인데, 이를 위해 노드 특성(feature)으로 프로모터 서열(promoter sequence)과 유전자 패밀리 정보 등 유전자에 대한 정보를 사용하게 된다. 이 모델의 학습 과정에서는 기존의 방법과 다른 점은 여러 과제를 학습하도록(multi-task learning)한 점이며, 엣지 예측 뿐만 아니라 원래의 유전자 발현을 예측하도록 학습시켰으며, 또한 특정 유전자의 결실(knockout)이 일어났을 때 다른 유전자의 발현이 어떻게 변하는지 예측하도록 학습시켰다. 이러한 multi-task learning 전략의 정점은, 기존의 예측 모델에서 충분히 활용하지 못하던 유전자 결실(knockout)이라는 직접적 인과 정보와 프로모터(promoter) 서열이라는 전사(transcription) 조절의 직접적 정보를 통합하여 높은 성능으로 유전자 조절 네트워크를 구성할 수 있다는 점이다. 또한, 이 방법에서는 전사인자-타겟유전자라는 사전정보가 없이도 맥락에 맞게 유전자 조절을 추론할 수 있었다는 점이다. 이런 장점 때문에, MTLGRN은 사전 정보가 없는 비모델 종에 대해서도 효과적으로 유전자 조절 네트워크를 재구성할 수 있다.

3.4.2 유전자 조절의 종류와 방향성 분석

유전자 네트워크를 추론하거나 재구성할 때, 단순히 두 유전자가 전사조절 관계에 있다는 것을 아는 것뿐만 아니라, 어떤 유전자가 전사인자이고 어떤 유전자가 타겟 유전자인지(방향성), 그리고 이 관계가 활성화인지 억제인지(유형)를 명확히 아는 것이 중요하다. 이 모든 정보가 통합된 네트워크를 정교한 유전자 조절 네트워크(fine gene regulatory network, FGRN)라고 하며, FGRN을 정확히 재구성하는 것은 유전자 네트워크 분석의

핵심적인 난제 중 하나이다.

FGRN 재구성을 위해 개발된 인공지능 모델 중 하나가 DeepFGRN인데^[50], 이 모델은 두 가지 접근 방식을 결합하여 전사조절인자와 타겟 유전자의 관계성을 예측한다. 첫째, 두 유전자의 발현 패턴을 분석하여 발현 상관관계를 학습한다. 둘째, 생성적 적대 신경망(Generative Adversarial Network, GAN)을 이용하여 전사인자가 될 유전자와 표적이 될 유전자의 고유한 임베딩(embedding)을 각각 생성하여 학습한다.

DeepFGRN은 이처럼 GAN 기반의 적대적 학습을 활용하므로, 사전 정보가 부족한 경우에도 복잡한 유전자 조절 네트워크를 효과적으로 분석할 수 있는 장점을 갖는다. 특히 이 모델은 유전자 조절의 인과관계와 상관관계를 모두 추론할 수 있는데, 이 능력은 질병 관련 바이오마커와 같은 핵심 유전자를 발굴하거나, 미생물 균주를 목적에 맞게 유전적으로 조작하는 데 필요한 핵심 단서를 발굴하는데 활용될 수 있다.

4 대사 네트워크(Metabolic network)

4.1 대사 네트워크 분석을 위한 데이터

대사 네트워크는 세포의 특징이 실제적으로 표현되는 최종 실행 계층이며, 구체적으로는 효소에 의해 기질이 에너지와 다양한 생체물질로 전환되는 반응들의 연결망이다. 다른 네트워크와 마찬가지로 대사 네트워크 역시 노드와 엣지로 구성되는데, 일반적으로 노드는 대사산물(metabolite)을, 엣지는 이들을 연결하는 효소 반응(enzymatic reaction)을 나타낸다^[51].

대사 네트워크의 분석은 다른 네트워크에 비해 매우 복잡한데, 이는 엣지(반응)가 단순한 연결 여부뿐만 아니라, 반응의 화학량론(stoichiometry), 열역학적 특징, 그리고 효소 동역학적 특징(kcat, Km 등)을 모두 포함하기 때문이다. 또한, 이 반응을 촉진하거나 억제하는 여러 대사산물이나 보조인자(cofactor) 역시 네트워크의 복잡한 구성요소가 되므로, 전체 시스템을 분석하기는 쉽지 않다.

이러한 대사 네트워크를 재구성하는 첫 번째 방법은 상향식(bottom-up) 접근법인 유전체 기반 재구성이^[52]. 이 방법은 먼저 유전체(genome)에서 효소 기능을 하는 유전자를 식별하고, 이 유전자들이 매개하는 효소 반응을 기존 데이터베이스를 통해 확인한 뒤, 이

관계를 모두 연결하여 네트워크를 완성한다. 이 방식은 해당 유전체가 가진 모든 잠재적 대사 능력을 포괄하는 네트워크를 재구성할 수 있다는 장점이 있다. 그러나 기능이 식별되지 않은 효소 유전자로 인해 일부 반응이 누락되는 불완전성 문제가 발생할 수 있으며, 이 방법으로는 반응의 여러 동력학적 특징이 누락되는 단점이 있다. 또한, 이 네트워크는 세포의 모든 잠재적 대사 과정을 포함하는 정적(static)인 모델이므로, 특정 조건에서 실제로 활성화되는 주요 반응이나 경로(하위 네트워크)가 무엇인지 직접 알기 어려운 단점이 있다.

두 번째 방법은 하향식(top-down) 접근법으로, 실험으로 측정된 데이터로부터 네트워크를 직접 추론하는 것이다^[53]. 이는 특정 조건이 변화된 교란(perturbation) 상태나 시간의 흐름에 따른 대사체 농도(metabolomics) 또는 대사 흐름(fluxomics) 데이터를 대량으로 측정한 뒤, 데이터 간의 상관관계나 인과관계를 통계적으로 추론하는 방식이다. 예를 들어, A 물질이 증가한 직후 B 물질이 항상 증가한다면 $A \rightarrow B$ 의 반응 관계를 가정하여 네트워크를 구축할 수 있다. 이 방법은 조건 특이적이며 동적인(dynamic) 네트워크를 얻을 수 있다는 장점이 있다. 그러나 대사체 측정 실험의 기술적 한계로 인해, 측정이 용이한 대사산물 중심의 불완전한 네트워크가 만들어지기 쉬우며, 측정되지 않은 대사산물 관련 경로는 누락될 수 있다.

최근에는 이 두 방법의 단점을 극복하기 위한 통합적(hybrid) 접근법이 널리 사용된다. 이는 유전체 기반 재구성으로 전체 대사 네트워크(구조적 골격)를 먼저 구축하고, 여기에 전사체, 단백질체, 대사체 데이터를 통합하여 특정 조건에서 실제로 활성화되는 하위 네트워크가 무엇인지 정교하게 모델링하는 연구 방식이다.

4.2 데이터의 생산

4.2.1 대사체 데이터(Metabolome)

대사체(metabolome) 데이터는 세포, 조직, 배양액 등 주어진 시료 내에 존재하는 대사산물들을 총체적으로 측정하여 얻을 수 있는데, 실험 목적에 따라 접근법은 크게 두 가지로 나뉜다. 첫째, ‘비표적(untargeted)’ 방식은 광범위한 커버리지를 통해 가능한 한 많은 대사산물을 탐색하고 신규 물질을 발굴하는 데 적합하며^[54], 둘째, ‘표적(targeted)’ 방식(MRM/SRM 기반)은 특정 경로(pathway)의 대사산물을 정밀하게 정량하고 검증하는 데 이용된다^[55].

일반적인 실험의 과정은, 먼저 대사 과정을 급속히 중단(quenching)하고 내부에 농도

를 알고 있는 표준물질을 투입한 뒤, 용매를 이용해 대사체를 추출하는 것으로 시작한다 [56, 57]. 표준물질은, 실험 과정에서 얼마나 많은 시료가 손실되었는지를 확인할 수 있으므로, 이를 이용해 다른 대사산물의 농도를 보정하는 목적으로 활용된다. 이후, 액체 크로마토그래피(LC-MS, RP/HILIC 방식) 또는 기체 크로마토그래피(GC-MS)로 대사산물을 분리하여 측정하는데, GC-MS의 경우, 1차 대사체와 같이 휘발성이 낮은 물질을 분석하기 위해 유도체화(derivatization) 과정을 거친다. 이런 대사체 데이터의 신뢰도는 여러 방법을 이용해서 측정할 수 있는데, 앞서 설명한 표준물질을 분석 전-중-후의 농도를 확인함으로써 실험의 오류, 측정치의 노이즈 등을 확인할 수 있다.

4.2.2 대사 흐름 데이터(Fluxomics)

대사 흐름체(fluxome) 데이터는 안정 동위원소 추적(stable-isotope tracing)을 통해 얻는데, 대표적으로 ^{13}C -포도당(glucose)과 같은 라벨링된 기질을 세포에 투입하고, 이 기질이 대사 경로를 거치며 변화하는 동위원소 패턴(isotopologue/isotopomer)을 GC-MS나 LC-MS로 측정해서 얻을 수 있다 [58, 59]. 이 측정된 패턴을 수학적 모델에 적용하여 세포 내부의 대사 속도(intracellular flux)를 추정할 수 있다.

이를 위한 실험 설계는 크게 두 가지로 나뉘는데, ^{13}C -Metabolic Flux Analysis(동위원소 정상평형상태 MFA)는 라벨링된 기질이 투입된 후 충분한 시간이 지나 대사 흐름과 동위원소 비율이 평형(steady-state)에 도달했을 때 시료를 채취하는 방법이다 [58, 60]. 따라서 이 방법은 해당과정(glycolysis)이나 TCA 회로를 포함하는 핵심 대사 경로의 전반적인 흐름(flux)을 측정하는 데 유리하다. INST-MFA(동위원소 비정상상태 MFA)는 라벨링 기질 투입 직후부터 짧은 시간 간격으로 샘플을 여러 번 채취하여 동적인(dynamic) 변화를 포착하는 방법이다. 따라서 대사 과정이 매우 빠르거나 동적으로 변하는 대사 경로를 추적하는 데 유리하다 [61].

고품질의 대사 흐름체 데이터를 생성하기 위해서는 다음의 요건이 중요한데 [62, 63], (1) $[1-^{13}\text{C}]$ 또는 $[\text{U}-^{13}\text{C}]$ 포도당과 같이 목적에 맞는 적절한 동위원소 추적자(tracer)의 선택과 조합, (2) 배지, 세포 성장 상태 등 배양 조건의 표준화, (3) 빠른 대사 중단(quenching) 및 대사산물 추출, 그리고 (4) MS/MS 측정을 통한 대사산물 내 위치별 라벨링 데이터 확보다. 특히 마지막의 MS/MS를 통한 분석은, 동위원소가 대사산물의 어느 위치(탄소)에 결합했는지에 대한 정밀한 정보를 제공하기에, 이 위치 정보(isotopomer data)는 해당 대사산물이 형성되기까지 여러 대사 경로가 각각 얼마나 기여했는지를 추적할 수 있는 핵심 단서가 되므로, 정교한 대사 흐름 분석에 필수적이다.

4.3 관련 데이터베이스

대사 네트워크를 구축하고 분석하기 위해서는 각기 다른 전문화된 정보를 제공하는 여러 데이터베이스를 유기적으로 통합해서 사용해야 한다. 이들 데이터베이스는 어떤 정보가 들어 있느냐에 따라, (1) 유전체 규모 대사 네트워크 모델, (2) 대사경로 및 효소반응 정보, (3) 실험 데이터 저장소, (4) 분석을 위한 표준화 데이터베이스로 나눌 수 있다(표 3).

유전체 규모의 대사 네트워크 모델을 제공하는 데이터베이스에는 BiGG Models와 ModelSEED가 있다^[64, 65]. BiGG Models는 문헌을 통해 검증된 고품질의 유전체 규모 대사 모델(Genome-Scale Models)을 제공하며, 이는 Flux Balance Analysis(FBA)와 같은 시뮬레이션에 바로 활용될 수 있는 장점이 있다. ModelSEED는 유전체 서열로부터 대사 네트워크 초안을 자동으로 재구성하는 플랫폼과 그 결과물인 네트워크를 제공한다. 이 플랫폼은 유전체가 시퀀싱된 새로운 종에서 대사 네트워크를 신속하게 구축하는 데 유용하다.

대사경로 및 효소반응 정보가 수록된 데이터베이스에는 KEGG, MetaCyc/BioCyc, Rhea 등이 있다^[66, 67, 68]. KEGG는 대사 경로(metabolic pathway), 효소반응과 대사산물 간의 관계 및 화학량론(stoichiometry) 정보를 제공하며, 이는 대사 네트워크를 재구성할 때 네트워크의 기본적인 정보로 활용될 수 있다. MetaCyc과 BioCyc은 전문가가 문헌을 기반으로 검증한 상세한 개별 효소반응 및 종(species)별 대사 경로 정보를 제공한다. 이 데이터베이스는 종별 대사 네트워크 모델을 정교하게 다듬거나, 재구성된 대사 네트워크에 누락된 반응이 없는지 탐색하는 데 활용될 수 있다. Rhea는 생화학 반응 자체를 표준화하는 데 중점을 둔 데이터베이스이다. 이는 효소반응의 방향성과 전하 균형(charge-balance)이 검증된 반응 데이터를 제공하여, 모델의 물리화학적 정확성을 검증하는 데 활용될 수 있다.

실험 데이터 저장소에는 MetaboLights와 Metabolomics Workbench가 있다^[69, 70]. 이 두 데이터베이스는 연구자들이 공개한 원시(raw) 및 가공된(processed) LC-MS/GC-MS 대사체 데이터와 실험 정보를 제공한다. 그리고, MassBank와 GNPS는 MS/MS 스펙트럼 라이브러리를 제공하는데, 이 스펙트럼은 실험에서 측정된 물질이 어떤 대사산물인지 동정하는 데 활용될 수 있다.

하나의 대사산물이라도 이를 명명하는 다양한 방법이 존재하므로, 데이터베이스마다 동일 대사산물이 서로 다른 이름으로 표시되어 있을 수 있다. 이런 문제를 해결하기 위해 화합물에 대한 통일된 정보를 제공하는 표준화 데이터베이스가 존재한다. 예를 들어, ChEBI와 HMDB는 화합물의 표준 명명법, 동의어, 분자 구조 및 물리화학적 특성 정보를 제공하고 있으며, MetaNetX는 KEGG, BiGG, Rhea 등 여러 데이터베이스 간에 사용되

는 대사산물 및 반응의 ID를 상호 매핑(cross-reference)해줌으로써, 서로 다른 출처의 데이터를 병합할 때 발생하는 문제를 해결하는 데 필요한 정보를 제공하고 있다.

대사 네트워크를 효과적이며 신뢰성 높게 재구성하거나 분석하기 위해서는 이러한 다양한 목적의 데이터베이스가 체계적으로 통합되는 게 필요하다.

표 3> 유전자 조절 네트워크 분석을 위한 데이터베이스

구분	데이터베이스	주요 내용 및 데이터	데이터 규모	주요 용도
대사 네트워크	BiGG Models [74]	• 검증된 고품질의 유전체 규모 대사 모델	70개 이상의 유전체 규모의 대사네트워크	FBA 등의 계산에 활용
	ModelSEED [75]	• 화합물 및 효소반응 정보 • 유전체 정보로부터 대사 네트워크를 재구성하는 플랫폼 제공	3만개의 화합물과 3만개의 효소반응	유전체 정보로부터 대사네트워크 재구성하는데 이용
대사경로/ 효소반응	KEGG [76]	• 효소반응과 대사산물 데이터 • 대사경로 정보 • 화학량론(stoichiometry) 정보	약 600종류의 경로 정보와 2만여개의 화합물	대사 네트워크 분석이나 재구성에 활용
	MetaCyc [77]	• 검증된 종별 대사경로와 효소 반응 정보	3천 종류의 경로 정보, 2만개의 반응정보, 2만개의 대사산물 정보	
	Rhea [78]	• 효소반응의 방향성과 전하 균형(charge-balance)이 검증된 반응 데이터	약 2만개의 반응정보 및 1.5만 개의 화합물 정보	
실험 데이터 저장소	MetaboLights [79]	• 대사체 LC-MS/GC-MS 실험 데이터	총 1.5만개의 데이터세트	실험데이터의 재분석이나 여러 실험데이터의 통합 분석에 활용
	Metabolomics Workbench [80]	• 대사체 실험 데이터	약 4천건의 데이터세트	
부가 데이터 베이스	MetaNetX [81]	• 데이터베이스 간의 화합물 매핑 정보	2,5만개의 반응과 3만개의 화합물 ID 매핑 정보	다른게 명명된 동일 화합물을 찾는 데 활용

4.4 대사 네트워크의 분석

4.4.1 대사 네트워크에서 누락된 효소 반응 예측

비모델 생물에서 대사 네트워크를 재구성할 때 가장 큰 어려움은 누락된 효소 반응으로 인해 대사 경로가 끊어지고, 이로 인해 재구성된 대사 네트워크 모델이 세포의 실제 표현형을 예측하는 정확도가 낮아질 수 있다는 점이다. CHESHIRE(CHEbyshev Spectral Hyperlink pREdictor)는 이 문제를 하이퍼그래프(Hypergraph)의 링크 예측 문제로 접근하여 해결하고자 한 딥러닝 방법이다.[82]. 특이한 점은 CHESHIRE는 실험 데이터 없이 오직 네트워크의

위상(topology) 정보만으로 누락된 효소 반응(하이퍼링크)을 예측할 수 있었다는 점이다.

대사 네트워크는 노드(대사산물)와 엣지(효소 반응)로 나타낼 수 있는데, 이때 하나의 효소 반응(엣지)은 여러 대사산물(노드)을 기질이나 생성물로 동시에 연결하므로, 대사 네트워크는 하나의 엣지에 여러 노드가 결합되는 하이퍼그래프 형태로 정확하게 표현된다. 먼저, 학습을 위해 네트워크 상에서 특정 대사산물에 인접한 다른 대사산물의 정보를 기반으로 인접행렬을 생성하고, 이를 체비쇼프 스펙트럼 그래프 컨볼루션 네트워크(Chebyshev Spectrum Graph Convolutional Network)를 사용하여 같은 반응에 참여하는 대사산물 간의 상호작용을 학습하여 모델을 개발했다. 이 과정을 통해 모델은 현재 네트워크에는 없지만 위상적으로 존재할 가능성이 높은 잠재적 반응(하이퍼링크)을 학습하고 예측하도록 개발되었으며, 이렇게 만들어진 CHESHIRE를 통해 49개의 유전체 기반 대사네트워크에서 200개의 누락된 효소반응을 찾아낼 수 있었다.

CHESHIRE는 이처럼 네트워크의 위상 구조 정보만으로 누락된 반응을 예측하므로, 다른 사전 지식이 없는 상태에서도 ModelSEED 등으로 얻은 초안 수준의 유전체 규모 대사 모델을 정교하게 다듬고 완성도를 높이는 데 유용하게 활용될 수 있다.

4.4.2 언어 모델을 이용한 데이터 마이닝과 대사네트워크 정교화

그래프 기반 딥러닝이 네트워크의 위상(topology) 정보에서 누락된 반응을 추론하는 반면, 자연어 처리와 대규모 언어 모델(Large Language Model, LLM)은 또 다른 핵심 정보 원인 ‘문헌’에서 네트워크 구조를 추론하는 강력한 도구로 부상하고 있다.

KEGG나 MetaCyc과 같은 정형화된 데이터베이스는 전문가의 직접적인 검증에 의존하므로 최신 연구 결과를 데이터베이스에 반영하기까지는 시간이 걸린다. 반면, 최신 논문에는 새로운 효소 반응이나 조절 관계가 텍스트, 표, 그림의 형태로 이미 존재하는데, LLM은 이러한 비정형 데이터에 포함된 지식을 자동으로 추출하여, 기존 대사 네트워크에 누락된 효소 반응이나 대사산물을 보강하는 역할을 할 수 있다.

NEKO(network for knowledge organization)는 이러한 문헌 마이닝을 위해 LLM을 활용하는 대표적인 사례이다^[83]. NEKO는 Retrieval-Augmented Generation(RAG) 기술을 사용해서 사용자의 질문이나 주제와 관련된 가장 관련성 높은 문헌을 방대한 문헌 데이터베이스에서 찾고, LLM이 해당 문헌 속의 효소, 대사체, 유전자, 배양 조건뿐만 아니라 대사 경로 이미지 내의 정보까지 포함하여 이들 간의 관계 정보를 추출한다. 이렇게 추출된 정보는 해석 가능한 지식 그래프(knowledge graph) 형태로 체계화되며, 이는 기존 대사 네트워크 모델에서 누락된 연결을 보강하거나 새로운 경로를 탐색하는 근거 자료로

활용될 수 있다. NEKO는 GPT와 같은 범용 모델의 제로샷(zero-shot) Q&A 방식보다 더 구체적인 정보를 제공하는 것으로 알려져 있으며, 이 모델 역시 공개되어 있기에 개별 연구자들이 이를 활용하여 새로운 정보를 발굴할 수 있다.

5 맺음말

본 보고서는 신호전달, 유전자 조절, 대사 네트워크라는 세 가지 핵심 계층을 소개했다. 합성생물학 분야에서 새로운 기능의 세포를 정밀하게 설계하기 위해서는 이들 네트워크에 대한 정확한 이해가 선행되어야 하며, 이는 대규모 데이터와 이를 분석할 인공지능 기술을 필수적으로 요구한다. 이에 따라, 본 보고서는 각 계층의 데이터 종류, 생산 방법, 관련 데이터베이스, 그리고 최신 인공지능 분석 기법을 순차적으로 다루었다.

최근 고속/대량 실험 기술의 발전으로 새로운 유형의 대규모 데이터가 축적되고 있으나, 해결할 과제 역시 분명하다. 여전히 실험 기술의 한계로 인해 데이터가 부정확하거나 누락되는 경우가 발생하며, 실험 자체의 높은 비용 때문에 인공지능 모델 학습에 충분한 대규모 데이터를 생산하기 어려운 것도 현실이다. 이러한 데이터의 양과 질의 문제는 세포 현상을 정교하게 이해하고 분석, 설계하는 데 걸림돌이 되고 있다.

그러나 최근 급격히 발전하는 인공지능 기술은 이러한 대규모 데이터의 정확한 분석뿐만 아니라, 불완전하거나 누락된 실험 데이터를 보완하는 방향으로까지 개발되고 있기에, 인공지능 기술의 발전은 생명 현상의 복잡성을 극복하고, 궁극적으로 네트워크 수준의 정밀한 제어 및 설계를 가속화할 것으로 기대된다.

[참고문헌]

- [1] Arribas Diez I, Govender I, Naicker P, Stoychev S, Jordaan J, Jensen ON. Zirconium(IV)-IMAC Revisited: Improved Performance and Phosphoproteome Coverage by Magnetic Microparticles for Phosphopeptide Affinity Enrichment. *J Proteome Res.* 2021 Jan 1;20(1):453-462. doi: 10.1021/acs.jproteome.0c00508. Epub 2020 Nov 23. PMID: 33226818.
- [2] Low TY, Mohtar MA, Lee PY, Omar N, Zhou H, Ye M. WIDENING THE BOTTLENECK OF PHOSPHOPROTEOMICS: EVOLVING STRATEGIES FOR PHOSPHOPEPTIDE ENRICHMENT. *Mass Spectrom Rev.* 2021 Jul;40(4):309-333. doi: 10.1002/mas.21636. Epub 2020 Jun 3. PMID: 32491218.
- [3] Grønborg M, Kristiansen TZ, Stensballe A, Andersen JS, Ohara O, Mann M, Jensen ON, Pandey A. A mass spectrometry-based proteomic approach for identification of serine/threonine-phosphorylated proteins by enrichment with phospho-specific antibodies: identification of a novel protein, Frigg, as a protein kinase A substrate. *Mol Cell Proteomics.* 2002 Jul;1(7):517-27. doi: 10.1074/mcp.m200010-mcp200. PMID: 12239280.
- [4] Fila J, Honys D. Enrichment techniques employed in phosphoproteomics. *Amino Acids.* 2012 Sep;43(3):1025-47. doi: 10.1007/s00726-011-1111-z. Epub 2011 Oct 15. PMID: 22002794; PMCID: PMC3418503.
- [5] Block H, Maertens B, Spriestersbach A, Brinker N, Kubicek J, Fabis R, Labahn J, Schäfer F. Immobilized-metal affinity chromatography (IMAC): a review. *Methods Enzymol.* 2009;463:439-73. doi: 10.1016/S0076-6879(09)63027-5. PMID: 19892187.
- [6] Zoumaro-Djayoon AD, Heck AJ, Muñoz J. Targeted analysis of tyrosine phosphorylation by immuno-affinity enrichment of tyrosine phosphorylated peptides prior to mass spectrometric analysis. *Methods.* 2012 Feb;56(2):268-74. doi: 10.1016/j.jymeth.2011.09.003. Epub 2011 Sep 16. PMID: 21945579.
- [7] Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, Latham V, Sullivan M. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.* 2012 Jan;40(Database issue):D261-70. doi: 10.1093/nar/gkr1122. Epub 2011 Dec 1. PMID: 22135298; PMCID: PMC3245126.
- [8] Lo Surdo P, Iannuccelli M, Contino S, Castagnoli L, Licata L, Cesareni G, Perfetto L. SIGNOR 3.0, the SIGNaling network open resource 3.0: 2022 update. *Nucleic Acids Res.* 2023 Jan 6;51(D1):D631-D637. doi: 10.1093/nar/gkac883. PMID: 36243968; PMCID: PMC9825604.

- [9] Milacic M, Beavers D, Conley P, Gong C, Gillespie M, Griss J, Haw R, Jassal B, Matthews L, May B, Petryszak R, Ragueneau E, Rothfels K, Sevilla C, Shamovsky V, Stephan R, Tiwari K, Varusai T, Weiser J, Wright A, Wu G, Stein L, Hermjakob H, D'Eustachio P. The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Res.* 2024 Jan 5;52(D1):D672-D678. doi: 10.1093/nar/gkad1025. PMID: 37941124; PMCID: PMC10767911.
- [10] Türei D, Valdeolivas A, Gul L, Palacio-Escat N, Klein M, Ivanova O, Ölbei M, Gábor A, Theis F, Módos D, Korcsmáros T, Saez-Rodriguez J. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Mol Syst Biol.* 2021 Mar;17(3):e9923. doi: 10.15252/msb.20209923. PMID: 33749993; PMCID: PMC7983032.
- [11] Perez-Riverol Y, Bandla C, Kundu DJ, Kamatchinathan S, Bai J, Hewapathirana S, John NS, Prakash A, Walzer M, Wang S, Vizcaíno JA. The PRIDE database at 20 years: 2025 update. *Nucleic Acids Res.* 2025 Jan 6;53(D1):D543-D553. doi: 10.1093/nar/gkae1011. PMID: 39494541; PMCID: PMC11701690.
- [12] Ochoa D, Jarnuczak AF, Viéitez C, Gehre M, Soucheray M, Mateus A, Kleefeldt AA, Hill A, Garcia-Alonso L, Stein F, Krogan NJ, Savitski MM, Swaney DL, Vizcaíno JA, Noh KM, Beltrao P. The functional landscape of the human phosphoproteome. *Nat Biotechnol.* 2020 Mar;38(3):365-373. doi: 10.1038/s41587-019-0344-3. Epub 2019 Dec 9. PMID: 31819260; PMCID: PMC7100915.
- [13] <https://www.phosphosite.org/homeAction.action>
- [14] <https://depod.bioss.uni-freiburg.de/>
- [15] <https://biomics.lab.nycu.edu.tw/dbPTM/>
- [16] <https://signor.uniroma2.it/>
- [17] <https://omnipathdb.org/>
- [18] <https://reactome.org/>
- [19] <https://www.ebi.ac.uk/pride/>
- [20] <https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp?redirect=auth>
- [21] Van den Broeck L, Bhosale DK, Song K, Fonseca de Lima CF, Ashley M, Zhu T, Zhu S, Van De Cotte B, Neyt P, Ortiz AC, Sikes TR, Aper J, Lootens P, Locke AM, De Smet I, Sozzani R. Functional annotation of proteins for signaling network inference in non-model species. *Nat Commun.* 2023 Aug 3;14(1):4654. doi: 10.1038/s41467-023-40365-z. PMID: 37537196; PMCID: PMC10400656.
- [22] Nilsson A, Peters JM, Meimetis N, Bryson B, Lauffenburger DA. Artificial neural networks enable genome-scale simulations of intracellular signaling. *Nat Commun.* 2022 Jun 2;13(1):3069. doi: 10.1038/s41467-022-30684-y. PMID: 35654811; PMCID: PMC9163072.

- [23] Chandramouli K, Qian PY. Proteomics: challenges, techniques and possibilities to overcome biological sample complexity. *Hum Genomics Proteomics*. 2009 Dec 8;2009:239204. doi: 10.4061/2009/239204. PMID: 20948568; PMCID: PMC2950283.
- [24] Clough E, Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Lee H, Zhang N, Serova N, Wagner L, Zalunin V, Kochergin A, Soboleva A. NCBI GEO: archive for gene expression and epigenomics data sets: 23-year update. *Nucleic Acids Res*. 2024 Jan 5;52(D1):D138-D144. doi: 10.1093/nar/gkad965. PMID: 37933855; PMCID: PMC10767856.
- [25] Sarkans U, Füllgrabe A, Ali A, Athar A, Behrangi E, Diaz N, Fexova S, George N, Iqbal H, Kurri S, Munoz J, Rada J, Papatheodorou I, Brazma A. From ArrayExpress to BioStudies. *Nucleic Acids Res*. 2021 Jan 8;49(D1):D1502-D1506. doi: 10.1093/nar/gkaa1062. PMID: 33211879; PMCID: PMC7778911.
- [26] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012 Sep 6;489(7414):57-74. doi: 10.1038/nature11247. PMID: 22955616; PMCID: PMC3439153.
- [27] Hammal F, de Langen P, Bergon A, Lopez F, Ballester B. ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Res*. 2022 Jan 7;50(D1):D316-D325. doi: 10.1093/nar/gkab996. PMID: 34751401; PMCID: PMC8728178.
- [28] Zou Z, Ohta T, Oki S. ChIP-Atlas 3.0: a data-mining suite to explore chromosome architecture together with large-scale regulome data. *Nucleic Acids Res*. 2024 Jul 5;52(W1):W45-W53. doi: 10.1093/nar/gkae358. PMID: 38749504; PMCID: PMC11223792.
- [29] Rauluseviciute I, Riudavets-Puig R, Blanc-Mathieu R, Castro-Mondragon JA, Ferenc K, Kumar V, Lemma RB, Lucas J, Chèneby J, Baranasic D, Khan A, Fornes O, Gundersen S, Johansen M, Hovig E, Lenhard B, Sandelin A, Wasserman WW, Parcy F, Mathelier A. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res*. 2024 Jan 5;52(D1):D174-D182. doi: 10.1093/nar/gkad1059. PMID: 37962376; PMCID: PMC10767809.
- [30] Vorontsov IE, Eliseeva IA, Zinkevich A, Nikonov M, Abramov S, Boytsov A, Kamenets V, Kasianova A, Kolmykov S, Yevshin IS, Favorov A, Medvedeva YA, Jolma A, Kolpakov F, Makeev VJ, Kulakovskiy IV. HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors. *Nucleic Acids Res*. 2024 Jan 5;52(D1):D154-D163. doi: 10.1093/nar/gkad1077. PMID: 37971293; PMCID: PMC10767914.
- [31] Han H, Cho JW, Lee S, Yun A, Kim H, Bae D, Yang S, Kim CY, Lee M, Kim E, Lee S, Kang B, Jeong D, Kim Y, Jeon HN, Jung H, Nam S, Chung M, Kim JH, Lee I. TRRUST v2:

- an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.* 2018 Jan 4;46(D1):D380-D386. doi: 10.1093/nar/gkx1013. PMID: 29087512; PMCID: PMC5753191.
- [32] Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.* 2019 Aug;29(8):1363-1375. doi: 10.1101/gr.240663.118. Epub 2019 Jul 24. Erratum in: *Genome Res.* 2021 Apr;31(4):745. doi: 10.1101/gr.275408.121. PMID: 31340985; PMCID: PMC6673718.
- [33] <https://www.ncbi.nlm.nih.gov/geo/>
- [34] <https://www.ebi.ac.uk/biostudies/arrayexpress>
- [35] <https://rna.recount.bio/>
- [36] <https://archs4.org/>
- [37] <https://www.encodeproject.org/>
- [38] <https://remap2022.univ-amu.fr/>
- [39] <https://jaspar.elixir.no/>
- [40] <https://hocomoco13.autosome.org/>
- [41] <https://www.grnpedia.org/trrust/>
- [42] <https://saezlab.github.io/dorothea/>
- [43] Khalil AS, Collins JJ. Synthetic biology: applications come of age. *Nat Rev Genet.* 2010 May;11(5):367-79. doi: 10.1038/nrg2775. PMID: 20395970; PMCID: PMC2896386.
- [44] Cai Y, Wang Y, Hu S. Synthetic Gene Circuits Enable Sensing in Engineered Living Materials. *Biosensors (Basel).* 2025 Aug 22;15(9):556. doi: 10.3390/bios15090556. PMID: 41002296; PMCID: PMC12467368.
- [45] Velazquez JJ, LeGraw R, Moghadam F, Tan Y, Kilbourne J, Maggiore JC, Hislop J, Liu S, Cats D, Chuva de Sousa Lopes SM, Plaisier C, Cahan P, Kiani S, Ebrahimkhani MR. Gene Regulatory Network Analysis and Engineering Directs Development and Vascularization of Multilineage Human Liver Organoids. *Cell Syst.* 2021 Jan 20;12(1):41-55.e11. doi: 10.1016/j.cels.2020.11.002. Epub 2020 Dec 7. PMID: 33290741; PMCID: PMC8164844.
- [46] He F, Murabito E, Westerhoff HV. Synthetic biology and regulatory networks: where metabolic systems biology meets control engineering. *J R Soc Interface.* 2016 Apr;13(117):20151046. doi: 10.1098/rsif.2015.1046. Epub 2016 Apr 13. PMID: 27075000; PMCID: PMC4874428.
- [47] Schlitt T, Brazma A. Current approaches to gene regulatory network modelling. *BMC Bioinformatics.* 2007 Sep 27;8 Suppl 6(Suppl 6):S9. doi: 10.1186/1471-2105-8-S6-S9. PMID: 17903290; PMCID: PMC1995542.

- [48] Mochida K, Koda S, Inoue K, Nishii R. Statistical and Machine Learning Approaches to Predict Gene Regulatory Networks From Transcriptome Datasets. *Front Plant Sci.* 2018 Nov 29;9:1770. doi: 10.3389/fpls.2018.01770. PMID: 30555503; PMCID: PMC6281826.
- [49] Cui W, Long Q, Xiao M, Wang X, Feng G, Li X, Wang P, Zhou Y. Refining computational inference of gene regulatory networks: integrating knockout data within a multi-task framework. *Brief Bioinform.* 2024 Jul 25;25(5):bbae361. doi: 10.1093/bib/bbae361. PMID: 39082651; PMCID: PMC11289685.
- [50] Gao Z, Su Y, Xia J, Cao RF, Ding Y, Zheng CH, Wei PJ. DeepFGRN: inference of gene regulatory network with regulation type based on directed graph embedding. *Brief Bioinform.* 2024 Mar 27;25(3):bbae143. doi: 10.1093/bib/bbae143. PMID: 38581416; PMCID: PMC10998536.
- [51] Amara A, Frainay C, Jourdan F, Naake T, Neumann S, Novoa-Del-Toro EM, Salek RM, Salzer L, Scharfenberg S, Witting M. Networks and Graphs Discovery in Metabolomics Data Analysis and Interpretation. *Front Mol Biosci.* 2022 Mar 8;9:841373. doi: 10.3389/fmolb.2022.841373. PMID: 35350714; PMCID: PMC8957799.
- [52] Thiele I, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc.* 2010 Jan;5(1):93-121. doi: 10.1038/nprot.2009.203. Epub 2010 Jan 7. PMID: 20057383; PMCID: PMC3125167.
- [53] Cak r T, Khatibipour MJ. Metabolic network discovery by top-down and bottom-up approaches and paths for reconciliation. *Front Bioeng Biotechnol.* 2014 Dec 3;2:62. doi: 10.3389/fbioe.2014.00062. PMID: 25520953; PMCID: PMC4253960.
- [54] Grim CM, Luu GT, Sanchez LM. Staring into the void: demystifying microbial metabolomics. *FEMS Microbiol Lett.* 2019 Jun 1;366(11):fnz135. doi: 10.1093/femsle/fnz135. PMID: 31210257; PMCID: PMC6597752.
- [55] Zhou J, Yin Y. Strategies for large-scale targeted metabolomics quantification by liquid chromatography-mass spectrometry. *Analyst.* 2016 Nov 14;141(23):6362-6373. doi: 10.1039/c6an01753c. PMID: 27722450.
- [56] Lu W, Su X, Klein MS, Lewis IA, Fiehn O, Rabinowitz JD. Metabolite Measurement: Pitfalls to Avoid and Practices to Follow. *Annu Rev Biochem.* 2017 Jun 20;86:277-304. doi: 10.1146/annurev-biochem-061516-044952. PMID: 28654323; PMCID: PMC5734093.
- [57] Van Gulik WM, Canelas AB, Taymaz-Nikerel H, Douma RD, de Jonge LP, Heijnen JJ. Fast sampling of the cellular metabolome. *Methods Mol Biol.* 2012;881:279-306. doi: 10.1007/978-1-61779-827-6_10. PMID: 22639217.
- [58] Zamboni N, Fendt SM, Rühl M, Sauer U. (13)C-based metabolic flux analysis. *Nat Protoc.* 2009;4(6):878-92. doi: 10.1038/nprot.2009.58. Epub 2009 May 21. PMID:

- 19478804.
- [59] Buescher JM, Antoniewicz MR, Boros LG, Burgess SC, Brunengraber H, Clish CB, DeBerardinis RJ, Feron O, Frezza C, Ghesquiere B, Gottlieb E, Hiller K, Jones RG, Kamphorst JJ, Kibbey RG, Kimmelman AC, Locasale JW, Lunt SY, Maddocks OD, Malloy C, Metallo CM, Meuillet EJ, Munger J, Nöh K, Rabinowitz JD, Ralser M, Sauer U, Stephanopoulos G, St-Pierre J, Tennant DA, Wittmann C, Vander Heiden MG, Vazquez A, Vousden K, Young JD, Zamboni N, Fendt SM. A roadmap for interpreting (13)C metabolite labeling patterns from cells. *Curr Opin Biotechnol*. 2015 Aug;34:189-201. doi: 10.1016/j.copbio.2015.02.003. Epub 2015 Feb 28. PMID: 25731751; PMCID: PMC4552607.
 - [60] Sauer U. Metabolic networks in motion: 13C-based flux analysis. *Mol Syst Biol*. 2006;2:62. doi: 10.1038/msb4100109. Epub 2006 Nov 14. PMID: 17102807; PMCID: PMC1682028.
 - [61] Jazmin LJ, Young JD. Isotopically nonstationary 13C metabolic flux analysis. *Methods Mol Biol*. 2013;985:367-90. doi: 10.1007/978-1-62703-299-5_18. PMID: 23417813.
 - [62] de Falco B, Giannino F, Carteni F, Mazzoleni S, Kim DH. Metabolic flux analysis: a comprehensive review on sample preparation, analytical techniques, data analysis, computational modelling, and main application areas. *RSC Adv*. 2022 Sep 7;12(39):25528-25548. doi: 10.1039/d2ra03326g. PMID: 36199351; PMCID: PMC9449821.
 - [63] Long CP, Antoniewicz MR. High-resolution 13C metabolic flux analysis. *Nat Protoc*. 2019 Oct;14(10):2856-2877. doi: 10.1038/s41596-019-0204-0. Epub 2019 Aug 30. PMID: 31471597.
 - [64] King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, Ebrahim A, Palsson BO, Lewis NE. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res*. 2016 Jan 4;44(D1):D515-22. doi: 10.1093/nar/gkv1049. Epub 2015 Oct 17. PMID: 26476456; PMCID: PMC4702785.
 - [65] Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotechnol*. 2010 Sep;28(9):977-82. doi: 10.1038/nbt.1672. Epub 2010 Aug 29. PMID: 20802497.
 - [66] Kanehisa M. The KEGG database. *Novartis Found Symp*. 2002;247:91-101; discussion 101-3, 119-28, 244-52. PMID: 12539951.
 - [67] Karp PD, Billington R, Caspi R, Fulcher CA, Latendresse M, Kothari A, Keseler IM, Krummenacker M, Midford PE, Ong Q, Ong WK, Paley SM, Subhraveti P. The BioCyc collection of microbial genomes and metabolic pathways. *Brief Bioinform*. 2019 Jul

- 19;20(4):1085-1093. doi: 10.1093/bib/bbx085. PMID: 29447345; PMCID: PMC6781571.
- [68] Bansal P, Morgat A, Axelsen KB, Muthukrishnan V, Coudert E, Aimo L, Hyka-Nouspikel N, Gasteiger E, Kerhornou A, Neto TB, Pozzato M, Blatter MC, Ignatchenko A, Redaschi N, Bridge A. Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Res.* 2022 Jan 7;50(D1):D693-D700. doi: 10.1093/nar/gkab1016. PMID: 34755880; PMCID: PMC8728268.
- [69] Haug K, Cochrane K, Nainala VC, Williams M, Chang J, Jayaseelan KV, O'Donovan C. MetaboLights: a resource evolving in response to the needs of its scientific community. *Nucleic Acids Res.* 2020 Jan 8;48(D1):D440-D444. doi: 10.1093/nar/gkz1019. PMID: 31691833; PMCID: PMC7145518.
- [70] Sud M, Fahy E, Cotter D, Azam K, Vadivelu I, Burant C, Edison A, Fiehn O, Higashi R, Nair KS, Sumner S, Subramaniam S. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* 2016 Jan 4;44(D1):D463-70. doi: 10.1093/nar/gkv1042. Epub 2015 Oct 13. PMID: 26467476; PMCID: PMC4702780.
- [71] Brown M, Dunn WB, Dobson P, Patel Y, Winder CL, Francis-McIntyre S, Begley P, Carroll K, Broadhurst D, Tseng A, Swainston N, Spasic I, Goodacre R, Kell DB. Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. *Analyst.* 2009 Jul;134(7):1322-32. doi: 10.1039/b901179j. Epub 2009 Apr 9. PMID: 19562197.
- [72] Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.* 2016 Jan 4;44(D1):D1214-9. doi: 10.1093/nar/gkv1031. Epub 2015 Oct 13. PMID: 26467479; PMCID: PMC4702775.
- [73] Moretti S, Tran VDT, Mehl F, Ibberson M, Pagni M. MetaNetX/MNXref: unified namespace for metabolites and biochemical reactions in the context of metabolic models. *Nucleic Acids Res.* 2021 Jan 8;49(D1):D570-D574. doi: 10.1093/nar/gkaa992. PMID: 33156326; PMCID: PMC7778905.
- [74] <http://bigg.ucsd.edu/>
- [75] <https://modelseed.org/>
- [76] <https://www.genome.jp/kegg/>
- [77] <https://metacyc.org/>
- [78] <https://www.rhea-db.org/>
- [79] <https://www.ebi.ac.uk/metabolights/>
- [80] <https://www.metabolomicsworkbench.org/>

- [81] <https://www.metanetx.org/>
- [82] Chen C, Liao C, Liu YY. Teasing out missing reactions in genome-scale metabolic networks through hypergraph learning. Nat Commun. 2023 Apr 25;14(1):2375. doi: 10.1038/s41467-023-38110-7. PMID: 37185345; PMCID: PMC10130184.
- [83] Xiao Z, Pakrasi HB, Chen Y, Tang YJ. Network for knowledge Organization (NEKO): An AI knowledge mining workflow for synthetic biology research. Metab Eng. 2025 Jan;87:60-67. doi: 10.1016/j.jymben.2024.11.006. Epub 2024 Nov 21. PMID: 39580108.

[주요용어풀이]

- 시스템생물학(Systems Biology): 생물학적 구성요소 간의 상호작용과 그로 인해 발생하는 시스템 전체의 특성을 총체적으로 연구하는 학문 분야.
- 합성생물학(Synthetic Biology): 생명 현상의 원리를 바탕으로 자연에 존재하지 않는 생명체의 구성요소나 시스템을 설계하고 제작하는 학문 분야.
- 데이터 모달리티(Data Modality): 데이터가 표현되는 방식이나 유형을 의미하며, 텍스트, 이미지, 소리 등 서로 다른 형태의 데이터를 지칭.
- 교란(Perturbation): 안정된 상태에 있는 시스템에 의도적인 방해나 변화를 가하여 그 반응을 관찰하는 실험적 행위.
- 위상(Topology): 기하학적 대상의 연결성이나 구조적 관계.
- 신호 전달 네트워크(Signaling Network): 세포가 외부의 신호를 감지하고 이를 세포 내부로 전달하여 특정 반응을 유도하는 분자(주로 단백질) 간의 상호작용 연결.
- 유전자 조절 네트워크(Gene Regulatory Network): 전사인자(Transcription Factor)와 같은 조절 인자가 특정 유전자의 발현을 켜거나 끄는 상호작용을 총체적으로 나타낸 네트워크.
- 대사 네트워크(Metabolic Network): 생물체 내에서 일어나는 모든 생화학 반응(효소 반응)과 그 반응에 관여하는 대사산물 간의 복잡한 연결망.
- 인산화 단백질체(Phosphoproteome): 특정 시점, 특정 조건의 세포나 조직 내에 존재하는 인산화(phosphorylation)된 단백질의 총합을 의미.
- 시계열 인산화 단백질체(Time-series Phosphoproteome): 시간에 따른 인산화 단백질체의 변화 양상을 연속적으로 측정한 데이터 세트.

- 전사체(Transcriptome): 특정 세포나 조직에서 발현되는 모든 RNA 분자(mRNA, tRNA, rRNA 등)의 총합을 의미.
- 대사체(Metabolome): 특정 세포나 조직 내에 존재하는 포도당, 아미노산, 지방산 등 모든 저분자 대사산물의 총합을 의미.
- 대사 흐름체(Fluxome): 특정 조건에서 대사 네트워크를 통과하는 각 생화학 반응의 실제 속도(flux) 값들의 총합.
- 인산화 효소(Kinase): ATP와 같은 분자에서 인산기(phosphate group)를 떼어내어 다른 분자(주로 단백질)에 옮겨 붙이는 ‘인산화’ 과정을 촉매하는 효소의 총칭.
- 탈인산화 효소(Phosphatase): 인산화 효소와 달리, 단백질에 붙은 인산기를 제거하는 효소의 총칭.
- 인산화(Phosphorylation): 특정 분자에 인산기(phosphate group)가 화학적으로 결합하는 반응으로, 특히 단백질의 활성을 조절하는 핵심 기전.
- LC-MS/MS: 액체 크로마토그래피(LC)로 혼합물을 분리하고, 탠덤 질량 분석기(MS/MS)로 각 성분을 식별하고 정량하는 고감도 분석 기술.
- 인산화 부위(Phosphosite): 단백질의 아미노산 서열 중 인산기가 결합하는 특정 위치(주로 Serine, Threonine, Tyrosine)를 의미.
- 번역 후 변형(Post-Translational Modification, PTM): 단백질이 리보솜에서 합성(번역)된 후, 인산화나 당화(Glycosylation) 등 추가적인 화학적 변형을 거쳐 기능이 조절되는 과정.
- 전사인자(Transcription Factor, TF): DNA의 특정 부위에 결합하여 해당 유전자의 RNA 전사(transcription) 과정을 촉진하거나 억제하는 조절 단백질.
- RNA-Sequencing(RNA-seq): 차세대 시퀀싱(NGS) 기술을 이용해 전사체(RNA)를 cDNA로 변환한 뒤, 그 염기 서열을 대량으로 읽어내어 유전자 발현을 정량하는 기술.
- 원시 데이터(Raw data): 실험 장비에서 생성된 직후 아무런 처리나 가공을 거치지 않은 원본 데이터 파일(예: 시퀀싱의 FASTQ 파일, 질량 분석기의 .raw 파일).
- 데이터 재처리(Re-processing): 이미 공개되거나 수집된 여러 원시 데이터를 일관된 표준 파이프라인(분석 방법)을 사용해 처음부터 다시 분석하고 가공하는 과정.
- 모티프(Motif): 생물학적 서열(DNA, RNA, 단백질) 내에서 특정 기능(예: TF 결합)을 수행하며 반복적으로 나타나는 보존된 패턴이나 염기서열을 의미.
- 정교한 유전자 조절 네트워크(FGRN): 유전자 간의 조절 관계뿐만 아니라, 조절의 방향성(누가 누구를)과 유형(활성인지 억제인지)까지 포함하는 상세한 네트워크.
- 하이퍼그래프(Hypergraph): 일반 그래프의 엣지가 2개의 노드만 연결하는 것과 달리, 하나의 ‘하이퍼엣지’가 2개 이상의 노드를 동시에 그룹으로 묶거나 연결할 수 있는 일반화된 그래프 구조.
- 화학량론(Stoichiometry): 화학 반응식에서 반응물과 생성물 간의 정량적인 몰(mole) 비율 관계를 나타내는 것.
- 유전체 기반 재구성(Genome-scale Reconstruction): 한 생명체의 전체 유전체(genome) 서열 정보를 바탕으로, 해당 생명체가 가질 수 있는 모든 대사 반응을 추론하여 총체적인 대사

- 네트워크 모델을 구축하는 과정.
- 비표적/표적 대사체 분석(Untargeted/Targeted Metabolomics): 비표적 분석은 시료 내의 가능한 한 많은 대사산물을 탐색하는 방식이며, 표적 분석은 미리 정해진 특정 대사산물 그룹만을 정확하게 정량하는 방식.
 - 급속 중단(Quenching): 대사체나 대사 흐름 분석 시, 샘플 채취 순간의 대사 활동을 즉시(초 단위) 중단시키기 위해 초저온 용매(예: 액체 질소, 콜드 메탄올)로 세포를 급속 냉각시키는 실험 기법.
 - 안정 동위원소 추적(Stable-isotope tracing): ^{13}C 나 ^{15}N 와 같이 방사선이 없는 무거운 동위원소(안정 동위원소)로 표지된 기질을 세포에 공급하여, 이 동위원소가 대사 경로를 따라 이동하는 것을 추적하는 실험 기법.
 - 동위원소 패턴(Isotopologue/Isotopomer): Isotopologue는 분자 전체에 포함된 동위원소의 총 개수 차이(예: $\text{M}+0$, $\text{M}+1$, $\text{M}+2$)이며, Isotopomer는 동위원소가 분자 내 특정 위치에 결합한 차이(위치 이성질체)를 의미.
 - ^{13}C -MFA (Metabolic Flux Analysis): ^{13}C 로 표지된 기질을 공급한 후, 시스템이 동위원소 정상상태(steady-state)에 도달했을 때의 라벨링 패턴을 분석하여 대사 흐름을 추정하는 기법.
 - INST-MFA (Isotopically Non-Stationary MFA): ^{13}C 기질 투입 직후, 동위원소 비율이 정상상태에 도달하기 전의 비정상상태(non-stationary)에서 시간대별로 샘플을 채취하여 대사 흐름을 추정하는 기법.
 - 그래프 신경망(Graph Neural Network, GNN): 노드와 엣지로 구성된 그래프 구조의 데이터를 직접 학습하고 처리하기 위해 설계된 딥러닝 모델의 한 종류.
 - 다중 과제 학습(Multi-Task Learning): 하나의 인공지능 모델이 서로 연관된 여러 개의 작업을 동시에 학습함으로써, 개별 작업을 따로 학습하는 것보다 높은 일반화 성능을 달성하려는 기계 학습 방식.
 - 생성적 적대 신경망(Generative Adversarial Network, GAN): 진짜 같은 가짜 데이터를 생성하는 ‘생성자’와 그것을 감별하는 ‘판별자’가 서로 경쟁하며 학습(적대적 학습)하는 생성형 인공지능 모델.
 - RAG(Retrieval-Augmented Generation): 대규모 언어 모델(LLM)이 답변을 생성할 때, 내부 지식에만 의존하지 않고 외부의 신뢰할 수 있는 문서(예: 논문)를 먼저 검색(Retrieval)하여 그 정보를 기반으로 답변을 생성(Generation)하는 인공지능 기술.

생물공정 데이터 기반 스마트 발효 기술동향

1 머리말

생물공정(bioprocess)은 미생물, 동물세포, 식물세포 등 살아있는 생물체를 이용하여 의약품, 바이오연료, 식품, 화학물질 등 고부가가치 제품을 생산하는 핵심 산업 기술이다. 특히 발효(fermentation) 공정은 생물공정의 중심에 있으며, 페니실린과 같은 항생제부터 단일클론항체(monoclonal antibody)와 같은 바이오의약품에 이르기까지 광범위한 제품 생산에 활용되고 있다. 그러나 생물학적 시스템은 본질적으로 비선형적이고 시간 가변적이며 예측 불가능한 특성을 지니고 있어, 공정의 실시간 모니터링과 정밀 제어가 매우 어렵다.

생물공정 분야의 디지털 전환은 4차 산업혁명의 핵심 동력으로서, 전통적인 발효 공정을 데이터 중심의 지능형 시스템으로 변화시키고 있다. 특히 바이오의약품, 바이오연료, 그리고 지속가능한 화학소재에 대한 수요가 급증하면서, 발효 공정의 효율성과 안정성을 획기적으로 개선할 수 있는 스마트 발효 기술의 중요성이 더욱 부각되고 있다^[1].

전통적인 발효 공정 관리는 온도, pH, 용존산소(dissolved oxygen, DO) 등 물리화학적 파라미터에 의존하며, 생체량(biomass), 대사산물 농도, 제품 농도 등 핵심 공정 변수는 주로 오프라인 분석을 통해 측정된다. 이러한 오프라인 측정은 시료 채취, 전처리, 분석 과정에서 상당한 시간이 소요되며, 샘플링 빈도가 제한적이어서 공정의 동적 변화를 실시간으로 파악하기 어렵다. 이러한 한계를 극복하기 위해 소프트 센서(soft sensor) 기술이 주목받고 있다.

소프트 센서는 실시간으로 측정 가능한 온라인 센서 데이터와 수학적 모델을 결합하여 직접 측정이 어려운 변수를 간접적으로 추정하는 기술이다^[2]. 최근에는 인공지능과 머신러닝 기술의 발전으로 데이터 기반 소프트 센서의 성능이 크게 향상되었으며, 발효 시계열 데이터(fermentation time series)를 효과적으로 활용하여 공정 상태를 예측하고 제어하는 것이 가능해졌다^[3].

그러나 생물공정 데이터는 본질적으로 희소하고 비용이 많이 든다는 문제가 있다. 각 배치 실험은 수일에서 수주가 소요되며, 대규모 데이터셋 구축에 막대한 시간과 비용이 필요하다. 이러한 소규모 데이터(small data) 문제를 해결하기 위해 합성 시계열 데이터(synthetic time series) 생성 기술이 새로운 해결책으로 부상하고 있다. 변분 오토

인코더(Variational Autoencoder, VAE), 생성적 적대 신경망(Generative Adversarial Network, GAN) 등의 생성 모델을 활용하여 실제 발효 데이터와 유사한 특성을 가진 합성 데이터를 생성함으로써, 모델 학습에 필요한 데이터셋을 확장하고 예측 성능과 강건성을 향상시킬 수 있다^[1].

본 보고서는 생물공정 데이터 기반 스마트 발효 기술의 최신 동향을 종합적으로 분석한다. 특히 소프트 센서 기술의 발전, 발효 시계열 데이터의 특성과 분석 방법, 합성 시계열 데이터 생성 기술, 그리고 이들의 융합을 통한 차세대 스마트 발효 시스템의 가능성을 탐구한다.

2 생물공정 데이터 기반 스마트 발효의 기술 동향

2.1 스마트 발효 기술 개요

유스마트 발효는 전통적인 발효 공정에 IoT(Internet of Things), 인공지능, 빅데이터 분석 기술을 융합하여 자율적이고 지능적인 발효 공정 관리를 실현하는 차세대 생물공정 기술이다. 이는 Industry 4.0의 개념을 생물공정 분야에 적용한 것으로, 디지털화된 공정 데이터를 기반으로 실시간 의사결정과 자동화된 공정 제어를 가능하게 한다^[2].

Lee 등(2025)은 스마트 농업 기술로서의 센서와 머신러닝 기반 정밀 발효 시스템을 다루었다. 이 연구는 발효 공정의 실시간 모니터링과 제어를 위한 IoT 센서, 데이터 분석 플랫폼, 머신러닝 알고리즘의 통합 시스템을 제시하였으며, 특히 소규모 발효 시설에서도 경제적으로 구현 가능한 스마트 발효 솔루션을 강조하였다^[2].

표 1> 스마트 발효 핵심 기술과 연관 데이터 분석 도구

기술분야	핵심 기술	주요 데이터 유형	분석 도구/플랫폼
실시간 데이터 수집	IoT 센서 네트워크, PAT 시스템	pH, DO, 온도, 교반속도, 분광데이터(NIR, Raman)	LabVIEW, MATLAB, PI System, OSIsoft
소프트 센서	딥러닝 기반 예측, MLOps 파이프라인	시계열 공정 데이터, 바이오매스, 산물농도	TensorFlow, PyTorch, MLflow, Kubeflow
합성 데이터 생성	VAE, GAN 기반 데이터 증강	발효 프로파일, 이상 상황 시나리오	TSGM, TimeGAN, Pytorch Lightning
지능형 제어	Model Predictive Control, 강화학습	제어 신호, 최적화 파라미터	MATLAB Simulink, OpenAI Gym

2.2 소프트 센서 기술의 개요와 발전

소프트 센서는 하드웨어 센서로부터 수집된 데이터와 수학적 모델을 결합하여 직접 측정하기 어려운 변수를 간접적으로 추정하는 기술이다. 생물공정에서 소프트 센서는 온도, pH, 용존산소, 교반 속도, 배기가스 조성 등 실시간으로 측정 가능한 변수들을 입력으로 받아, 생체량, 기질 농도, 대사산물 농도, 제품 농도, 비성장속도(specific growth rate) 등 측정이 어렵거나 시간이 많이 소요되는 변수들을 예측한다.

Luttmann 등(2012)은 생물공정에서 소프트 센서의 현황과 권고사항을 종합적으로 정리한 보고서를 발표하였다^[7]. 이 보고서는 유럽생명공학연맹(European Federation of Biotechnology)의 전문가 패널에 의해 작성되었으며, 소프트 센서 기술의 산업적 적용을 촉진하기 위한 핵심 과제와 발전 방향을 제시하였다. 보고서는 소프트 센서가 공정분석기술(Process Analytical Technology, PAT) 구현의 핵심 도구이며, 특히 바이오의약품 제조에서 공정 이해도와 제어 성능을 향상시킬 수 있는 잠재력이 크다고 강조하였다.

소프트 센서의 접근법은 크게 메커니즘 기반(mechanistic), 데이터 기반(data-driven), 하이브리드(hybrid) 모델로 분류된다. 메커니즘 기반 소프트 센서는 물리화학적 원리와 생물학적 지식에 기반한 수학적 모델을 사용한다. Golabgir 등(2015)은 생화학 공정 모델의 관측가능성 분석(observability analysis)이 메커니즘 기반 소프트 센서 개발에 유용한 도구임을 입증하였다^[14]. 이를 통해 소프트 센서 설계 과정에서 적절한 센서 조합을 선택하고, 추정의 신뢰성을 사전에 평가할 수 있다.

데이터 기반 소프트 센서는 공정의 메커니즘에 대한 명시적 지식 없이 입력-출력 데이터의 통계적 관계를 학습하는 접근법이다. Kadlec 등(2009)은 공정산업에서 데이터 기반 소프트 센서에 대한 포괄적인 리뷰를 제공하였다^[12]. Zhu 등(2020)은 발효 공정을 위한 소프트 센서의 현대적 모델링 방법을 종합적으로 리뷰하였으며, 특히 심층신경망(Deep Neural Network), 퍼지논리(Fuzzy Logic), 확률적 잠재변수 모델(Probabilistic Latent Variable Model) 등 최신 기법들의 발효 공정 적용 사례를 소개하였다^[11].

하이브리드 소프트 센서는 메커니즘 모델과 데이터 기반 모델의 장점을 결합한 접근법이다. Brunner 등(2020)은 *Pichia pastoris* 유가배양 공정을 위한 생체량 소프트 센서를 개발하면서 단계 감지(phase detection) 알고리즘과 하이브리드 모델링을 결합하였다^[8]. 메커니즘 모델(탄소 수치)과 데이터 기반 모델(다중선형회귀)을 결합한 이 하이브리드 접근법은 전체 공정에 걸쳐 평균 상대 예측 오차 5.52%, 결정계수(R^2) 0.96의 우수한 성능을 보였다.

표 2> 소프트 센서 개발 단계별 주요 기법과 도구

개발 단계	주요 기법	사용 도구	핵심 고려사항
데이터 수집 및 전처리	시계열 보간법, 이상치 제거, 정규화	Pandas, NumPy, Scikit-learn	센서 드리프트 보정, 결측값 처리 전략
변수 선택	상관분석, PCA, ICA, 상호정보량	Feature-engine, mlxtend	다중공선성 제거, 차원 감소 최적화
모델 훈련	LSTM, GRU, Transformer, Hybrid Models	TensorFlow, PyTorch	과적합 방지, 하이퍼파라미터 최적화
성능 평가	교차검증, 외부 검증, 드리프트 감지	MLflow, WandB	모델 안정성, 일반화 성능

2.3 발효 시계열 데이터의 특성과 분석

발효 공정은 본질적으로 시간에 따라 변화하는 동적 시스템이다. 배치(batch) 또는 유가 배양(fed-batch) 모드에서 운전되는 발효 공정은 시작부터 종료까지 세포 성장, 기질 소비, 대사산물 축적, 환경 조건 변화 등 다양한 생물학적 및 물리화학적 변화를 겪는다. 발효 시계열 데이터는 비정상성(non-stationarity), 단단계 특성(multi-phase behavior), 가변 배치 길이(variable batch length), 높은 차원성과 상관관계 등 몇 가지 독특한 특성을 갖는다.

효과적인 소프트 센서 개발을 위해서는 원시 시계열 데이터를 적절히 전처리하고 유용한 특징(feature)을 추출하는 것이 중요하다. Paulsson 등(2014)은 대사열(metabolic heat) 신호의 순차적 필터링을 통해 노이즈를 효과적으로 제거하는 방법을 제시하였다^[15]. 이 연구에서는 3단계 순차 필터(지수 필터, 중간값 필터, 지수 필터)를 적용하여 저열 생산 배양에서도 안정적인 비성장속도 추정이 가능함을 보였다.

Brunner 등(2021)은 생물공정을 위한 소프트 센서 개발의 과제를 비판적으로 검토하였다^[9]. 이 리뷰는 가변 공정 길이(variable process length), 다중 공정 단계(multiple process phases), 센서 오류로 인한 잘못된 모델 입력 등 생물공정 소프트 센서 개발의 세 가지 주요 과제와 이에 대한 해결 방안을 논의하였다.

Kiviharju 등(2008)은 생체 열량계(biocalorimetry)가 생체량 소프트 센서로 유용하게 활용될 수 있음을 보였다^[13]. 생체 열량계는 미생물 대사에서 발생하는 열을 측정하여 대사 활성과 성장 속도를 모니터링하는 기술로, 비침습적이고 실시간 측정이 가능하다는 장점이 있다.

Golabgir 등(2016)은 라만 분광법(Raman spectroscopy) 데이터와 배기가스 분석기 데이터를 메커니즘 모델과 결합한 소프트 센서를 개발하였다^[19]. 이 연구는 Penicillium

chrysogenum 유가배양 공정의 생리학적 특성을 온라인으로 모니터링하는 것을 목표로 하였다. 라만 분광법을 통해 생물반응기 내 페니실린 농도를 정확하고 빈번하게 측정할 수 있었으며, 이를 호흡속도 측정과 결합하여 비성장속도와 비생산속도를 실시간으로 추정하였다.

Wang 등(2023)은 다중 센서 데이터 융합(multi-sensor data fusion)을 통해 소프트웨어 센서의 성능을 향상시키는 연구를 수행하였다^[5]. 근적외선(NIR) 분광법과 라만 분광법 데이터를 결합하고, 직교 방법(orthogonal method)을 사용하여 중복 특징을 제거함으로써 단일 센서 모델 대비 R^2 값을 9.2-100.4% 향상시켰다. 이러한 다중 센서 융합 접근법은 각 센서의 장점을 활용하고 단점을 보완하여 더 강건하고 정확한 공정 모니터링을 가능하게 한다.

2.4 합성 시계열 데이터 생성 기술

생물공정 데이터의 가장 큰 한계 중 하나는 데이터의 희소성이다. 각 발효 실험은 수일에서 수주가 소요되며, 대규모 생물반응기를 사용하는 경우 실험 비용이 매우 높다. 또한 생물학적 변동성으로 인해 각 배치는 고유한 특성을 가지며, 충분한 통계적 다양성을 확보하기 위해서는 많은 수의 배치 실험이 필요하다. 이러한 소규모 데이터(small data) 문제는 특히 딥러닝과 같은 데이터 집약적 머신러닝 기법의 적용을 어렵게 만든다.

데이터 증강(data augmentation)은 기존 데이터로부터 새로운 학습 샘플을 생성하여 학습 데이터셋의 크기와 다양성을 증가시키는 기법이다. 최근에는 딥러닝 기반 생성 모델(generative model)을 활용한 합성 시계열 데이터 생성이 주목받고 있다. 주요 생성 모델로는 생성적 적대 신경망(GAN), 변분 오토인코더(VAE), 확산 모델(Diffusion Model) 등이 있다.

Kwon 등(2024)은 VAE를 활용하여 에탄올 발효 공정의 합성 시계열 데이터를 생성하는 연구를 수행하였다^[11]. 이 연구는 소프트웨어 센서의 예측 성능과 강건성을 향상시키기 위한 데이터 증강 전략으로 VAE를 제안하였다. 구체적으로, 11개의 발효 실험 데이터를 사용하여 VAE를 학습시킨 후, 잠재 공간에서 샘플링하여 합성 데이터를 생성하였다. Leave-One-Out Cross-Validation(LOOCV)을 사용한 평가 결과, 합성 데이터를 추가한 경우 소프트웨어 센서의 예측 성능이 크게 향상되었다. 원본 데이터만 사용한 모델에 비해 합성 데이터를 추가한 모델은 R^2 점수가 34% 향상되었고, 변동성(표준편차)이 82% 감소하였다.

이 연구의 핵심 통찰은 VAE가 실험 데이터의 본질적인 통계적 분포를 학습하여, 생물학

적으로 타당하고 다양한 합성 시계열을 생성할 수 있다는 것이다. 생성된 합성 데이터는 원본 데이터의 특성(예: 에탄올 생산 곡선, 기질 소비 패턴, pH 변화)을 유지하면서도 배치 간 변동성을 반영하였다. 이를 통해 학습 데이터의 해 공간(solution space)을 더 완전하게 탐색할 수 있게 되어, 모델이 과적합되지 않고 새로운 조건에서도 정확한 예측을 제공할 수 있게 되었다.

Acosta-Pavas 등(2024)은 합성 데이터와 전이 학습을 활용한 대규모 발효 공정의 디지털 트윈 개발을 연구하였다^[4]. 이 논문은 제한된 실제 데이터 상황에서 메커니즘 모델을 통해 생성된 합성 데이터로 머신러닝 모델을 사전 학습하고, 소량의 실제 데이터로 미세 조정하는 전략을 제시하였다. 이를 통해 데이터 수집 비용을 크게 줄이면서도 정확한 디지털 트윈을 구축할 수 있었다.

표 3> 소프트 센서 개발 단계별 주요 기법과 도구

생성 기법	장점	단점	적용 분야
VAE 기반	안정적 훈련, 확률적 해석 가능	생성 품질 제한, 모드 붕괴 위험	연속 공정 데이터, 정상 상태 시뮬레이션
GAN 기반	고품질 합성 데이터, 다양한 패턴 생성	훈련 불안정성, 모드 붕괴 문제	이상 상황 시뮬레이션, 극값 조건 데이터
TimeGAN	시간 의존성 보존, 동적 패턴 학습	계산 복잡도 높음, 대용량 메모리 필요	배치 발효 프로파일, 공정 전환 시나리오
하이브리드 접근법	각 기법 장점 결합, 강건한 성능	복잡한 아키텍처, 하이퍼파라미터 조정	통합 공정 최적화, 다목적 시뮬레이션

2.5 최신 연구 동향 및 산업 적용

2.5.1 전이 학습과 디지털 트윈

전이 학습(transfer learning)은 한 도메인에서 학습한 지식을 다른 관련 도메인으로 전이하여 소규모 데이터 문제를 해결하는 기법이다. Rogers 등(2022)은 전이 학습을 생물 공정의 예측 모델링에 적용하는 연구를 수행하였다^[17]. 이 연구는 *Desmodesmus* sp.와 *Chlorella sorokiniana*라는 두 가지 미세조류의 루테인(lutein) 생산 공정을 대상으로 하였다. 먼저 *Desmodesmus* sp.의 풍부한 데이터를 사용하여 딥러닝 모델을 학습시킨 후, 이 모델의 지식을 *Chlorella sorokiniana* 모델로 전이하였다. *Chlorella sorokiniana*의 학습 데이터가 매우 제한적인 상황에서도, 전이 학습을 통해 구축된 모델은 처음부터 학습한 모델이나 메커니즘 모델보다 우수한 예측 성능을 보였다.

디지털 트윈(digital twin)은 물리적 시스템의 가상 복제본으로, 실시간 데이터를 사용하여 물리적 시스템의 상태를 모니터링하고 미래 거동을 예측하며 최적화 전략을 시뮬레이션할 수 있다. Rydal 등(2024)은 대규모 곰팡이 발효 공정에 적용 가능한 확장 가능한 적응형 하이브리드 모델링 프레임워크를 개발하였다^[18]. 이 연구는 Novonesis(구 Novozymes)의 파일럿 규모 복잡한 곰팡이 발효 공정을 대상으로 하였다.

연구팀은 먼저 데이터 기반 소프트 센서를 개발하여 주요 기질인 글루코스 농도를 추정하였다. 파일럿 규모 데이터를 사용하여 학습된 이 소프트 센서는 정규화 제곱평균제곱근 오차(N-RMSE) 2%의 높은 정확도를 달성하였다. 더욱 주목할 만한 점은 이 소프트 센서가 실험실 규모(파일럿 규모의 1/20 부피)에서도 N-RMSE 7.8%의 우수한 성능을 보였다. 이는 데이터 기반 모델의 규모 간 일반화 가능성을 입증한다.

2.5.2 강화 학습 기반 공정 최적화

강화 학습(reinforcement learning)은 환경과의 상호작용을 통해 보상을 최대화하는 행동 정책(policy)을 학습하는 머신러닝 기법이다. Petsagkourakis 등(2020)은 배치 생물 공정 최적화를 위한 강화 학습 방법론을 제안하였다^[20]. 이 연구는 불확실하고 확률적인 생물공정의 배치 간 학습(batch-to-batch learning)에 정책 경사(policy gradient) 기법을 적용하였다.

이 방법론의 핵심은 순환신경망(RNN)으로 제어 정책을 파라미터화한다는 것이다. RNN은 공정 상태를 입력으로 받아 다음 제어 조치를 출력으로 제공한다. 세 가지 사례 연구(미세조류 루테인 생산, 페니실린 생산, 비평할 공정)를 통해 이 접근법의 효과를 검증하였으며, 불확실성과 공정-모델 불일치가 있는 상황에서도 거의 최적에 가까운 정책을 얻을 수 있음을 보였다.

2.5.3 산업 적용 사례

Reid 등(2024)은 산업 미생물 및 세포 배양 공정에서 PAT와 소프트 센서의 적용 현황을 리뷰하였다^[3]. 이 논문은 제약 및 바이오의약품 산업에서 품질기반설계(Quality by Design, QbD)와 PAT의 채택이 증가하고 있으며, 특히 라만 분광법, NIR 분광법, 용량 센서 등의 첨단 센서와 머신러닝 기반 소프트 센서의 통합이 공정 이해도와 제어 성능을 크게 향상시키고 있다고 보고하였다.

Mandenius와 Gustavsson(2015)은 PAT와 소프트 센서, QbD의 통합에 대해 논의하였다^[10]. 이 논문은 소프트 센서가 QbD 원칙을 실천하는 핵심 도구이며, 특히 공정 이

해(process understanding), 설계 공간(design space) 정의, 공정 제어 전략(control strategy) 개발에 중요한 역할을 한다고 강조하였다.

2.5.4 실용화를 위한 과제

소프트 센서 기술이 실험실에서 산업 현장으로 성공적으로 전환되기 위해서는 여러 과제를 해결해야 한다. Siegl 등(2023)은 유사성 분석과 단계별 재보정을 통해 소프트 센서의 일반화 가능성을 향상시키는 연구를 수행하였다[6]. 이 연구는 원료 변동성, 스케일업 효과, 공정 교란 등에 대한 강건성을 확보하는 방법을 제시하였다.

Zhang 등(2019)은 하이브리드 물리 기반 및 데이터 기반 모델링을 생물공정의 온라인 시뮬레이션 및 최적화에 적용하는 혁신적인 프레임워크를 제안하였다^[16]. 이 프레임워크는 먼저 물리 기반 노이즈 필터(간단한 동역학 모델)를 사용하여 원시 공정 측정 데이터를 교정하여 고품질 데이터를 생성한다. 주요 과제로는 모델의 강건성과 일반화, 모델 유지보수와 적응, 규제 승인, 데이터 인프라와 통합, 인력과 교육 등이 있다.

3 맺음말

생물공정 데이터 기반 스마트 발효 기술은 센서 기술, 데이터 과학, 인공지능의 융합을 통해 발효 공정의 실시간 모니터링, 예측, 최적화를 실현하는 차세대 패러다임이다. 본 보고서에서는 이 분야의 핵심 기술인 소프트 센서, 발효 시계열 데이터 분석, 합성 시계열 데이터 생성 기술의 최신 동향을 종합적으로 분석하였다.

소프트 센서 기술은 메커니즘 기반, 데이터 기반, 하이브리드 접근법을 통해 빠르게 발전하고 있다. 특히 딥러닝 기반 데이터 모델과 메커니즘 지식의 통합은 정확도와 해석 가능성을 동시에 향상시키는 유망한 방향이다. 라만 분광법, NIR 분광법, 생체 열량계 등 첨단 센서 기술의 통합은 소프트 센서의 성능을 더욱 높이고 있다.

발효 시계열 데이터의 고유한 특성인 비정상성, 다단계 거동, 가변 배치 길이는 소프트 센서 개발의 주요 과제이다. 그러나 적절한 전처리, 특징 추출, 시계열 모델링 기법을 통해 이러한 과제를 효과적으로 해결할 수 있다. LSTM, GRU 등 순환신경망 기반 모델은 시계열의 장기 의존성을 포착하는 데 특히 효과적이다.

합성 시계열 데이터 생성은 생물공정의 고질적인 소규모 데이터 문제를 해결하는 혁신적인 접근법이다. VAE, GAN 등 생성 모델을 활용하면 실제 발효 데이터와 유사한 특성을

가진 합성 데이터를 생성하여 학습 데이터셋을 확장할 수 있다. 이는 소프트 센서의 예측 성능과 강건성을 크게 향상시킬 수 있음이 입증되었다.

향후 전망으로는 디지털 트윈의 성숙화, 자율 생물공정의 실현, 멀티모달 데이터 통합, 연합 학습과 지식 공유, 설명 가능한 AI의 발전, 그리고 지속가능성과 순환경제의 관점에서 스마트 발효 기술의 중요성 증대 등을 예상할 수 있다.

결론적으로, 생물공정 데이터 기반 스마트 발효 기술은 생물공학과 데이터 과학의 융합을 통해 발효 산업의 패러다임을 근본적으로 변화시키고 있다. 소프트 센서, 시계열 분석, 합성 데이터 생성, 디지털 트윈, 강화 학습 등의 기술이 통합적으로 발전하면서, 더욱 효율적이고 지속가능하며 고품질의 생물공정이 실현될 것이다. 이러한 기술 발전은 바이오의약품, 바이오연료, 바이오화학, 식품 등 광범위한 산업 분야에 혁신을 가져올 것으로 기대된다.

[참고문헌]

- [1] <https://doi.org/10.3390/bioengineering11080803>
- [2] <https://doi.org/10.3390/s25175251>
- [3] <https://doi.org/10.1093/jimb/kuae019>
- [4] <https://doi.org/10.1016/j.compchemeng.2024.108736>
- [5] <https://doi.org/10.3390/s23136014>
- [6] <https://doi.org/10.3390/s23042178>
- [7] <https://doi.org/10.1002/biot.201100506>
- [8] <https://doi.org/10.1002/bit.27454>
- [9] <https://doi.org/10.3389/fbioe.2021.722202>
- [10] <https://doi.org/10.1002/jctb.4477>
- [11] <https://doi.org/10.3390/s20061771>
- [12] <https://doi.org/10.1016/j.compchemeng.2008.12.012>
- [13] <https://doi.org/10.1007/s10295-008-0346-5>
- [14] <https://doi.org/10.1002/btpr.2176>
- [15] <https://doi.org/10.3390/s141017864>
- [16] <https://doi.org/10.1002/bit.27120>
- [17] <https://doi.org/10.1002/bit.27980>
- [18] <https://doi.org/10.1002/bit.28670>
- [19] <https://doi.org/10.1002/cite.201500101>
- [20] <https://doi.org/10.1016/j.compchemeng.2019.106649>

[주요용어풀이]

- 생물공정(Bioprocess): 미생물, 동물세포, 식물세포 등 살아있는 생물체를 이용하여 의약품, 바이오연료, 식품, 화학물질 등을 생산하는 공정
- 발효(Fermentation): 미생물을 이용하여 유기물을 분해하고 특정 물질을 생산하는 생물학적 공정
- 소프트 센서(Soft Sensor): 실시간으로 측정 가능한 온라인 센서 데이터와 수학적 모델을 결합하여 직접 측정이 어려운 변수를 간접적으로 추정하는 기술
- 시계열 데이터(Time Series Data): 시간의 흐름에 따라 순차적으로 관측되는 데이터
- 합성 데이터(Synthetic Data): 실제 데이터의 특성을 모방하여 인공적으로 생성한 데이터
- 변분 오토인코더(VAE, Variational Autoencoder): 데이터의 잠재적 표현을 학습하여 새로운 데이터를 생성할 수 있는 생성 모델
- 생성적 적대 신경망(GAN, Generative Adversarial Network): 생성자와 판별자가 경쟁적으로 학습하여 실제와 유사한 데이터를 생성하는 딥러닝 모델
- 공정분석기술(PAT, Process Analytical Technology): 공정의 실시간 모니터링과 제어를 통해 제품 품질을 보장하는 기술적 접근법
- 품질기반설계(QbD, Quality by Design): 제품 및 공정의 설계 단계부터 품질을 고려하여 개발하는 체계적 접근법
- 생체량(Biomass): 생물반응기 내 살아있는 세포의 총량
- 배치(Batch): 원료를 한 번에 투입하고 반응이 완료될 때까지 운전하는 공정 방식
- 유가배양(Fed-batch): 배치 공정 중 필요에 따라 기질이나 영양소를 추가로 공급하는 공정 방식
- 디지털 트윈(Digital Twin): 물리적 시스템의 가상 복제본으로, 실시간 데이터를 사용하여 시스템의 상태를 모니터링하고 미래 거동을 예측하는 기술
- 전이 학습(Transfer Learning): 한 도메인에서 학습한 지식을 다른 관련 도메인으로 전이하여 학습 효율을 높이는 머신러닝 기법
- 강화 학습(Reinforcement Learning): 환경과의 상호작용을 통해 보상을 최대화하는 행동을 학습하는 머신러닝 기법

- 라만 분광법(Raman Spectroscopy): 레이저 광을 시료에 조사하여 산란광을 분석함으로써 분자 구조와 화학 조성을 파악하는 분석 기술
- 근적외선 분광법(NIR, Near-Infrared Spectroscopy): 근적외선 영역의 빛을 이용하여 물질의 화학적 조성을 분석하는 기술
- LSTM(Long Short-Term Memory): 장기 의존성을 학습할 수 있는 순환신경망의 일종
- 하이브리드 모델(Hybrid Model): 메커니즘 기반 모델과 데이터 기반 모델을 결합한 모델
- 데이터 증강(Data Augmentation): 기존 데이터를 변형하거나 새로운 데이터를 생성하여 학습 데이터셋을 확장하는 기법

SECTION
2

합성생물학을 위한 AI 기술동향



한국합성생물학발전협의회
Korea Synthetic Biology Association

이준학

융합분과위원, 한국과학기술정보연구원 디지털바이오컴퓨팅연구단 단장

염민선

융합분과위원, 나무아이씨티 연구소장

이선재

융합분과위원, 광주과학기술원 생명과학과 교수

송자연

융합분과위원, 성균관대학교 송자연 조교수

합성생물학 시개발 활용을 위한 컴퓨팅 인프라 및 플랫폼 현황

1 머리말

백신·치료제·바이오제조·환경·보건안보로 이어지는 과학기술 어젠다가 동시다발적으로 가속화되면서, 합성생물학 연구의 DBTL(Design-Build-Test-Learn) 사이클은 더 빠르고 더 큰 규모로, 그리고 더 표준화된 방식으로 발전되고 있다. 특히 AI의 발전에 따라 AI를 이용한 설계·실험계획·해석 과정이 중요해짐에 따라, 데이터·컴퓨팅 인프라와 실험 자동화 플랫폼은 ad-hoc 도구의 조합이 아니라 상호운용 표준과 보안·거버넌스를 기반으로 한 체계적인 설계가 중요성이 부각되고 있다. 합성생물학 연구는 더 이상 ‘실험실 중심’만으로는 속도와 규모를 확보하기 어렵다. 특히 생성형 AI와 기계학습 기반 설계가 주류로 자리 잡으면서, 대용량 데이터 레이크·표준화된 메타데이터·하이브리드 HPC/클라우드·재현 가능한 워크플로우가 결합된 ‘디지털 파운드리 스택’이 핵심 인프라로 부상했다.

본 보고서는 이러한 필요성을 전제로, 합성생물학 AI 개발·활용을 가능하게 하는 글로벌/국내 인프라의 현황을 계층적으로 정리하고, 향후 합성생물학 분야에서의 AI활용을 위한 아키텍처 및 주요 고려사항들을 정리하고 제시한다.

합성생물학을 위한 컴퓨팅인프라의 스택은 대략 네 층으로 요약된다. 첫째, 설계·부품 레지스트리 계층에서는 SBOL(Synthetic Biology Open Language)과 이를 구현한 SynBioHub, JBEI-ICE가 설계 교환·버전관리의 사실상 표준을 형성하고 있다. 둘째, 자동화·파운드리 운영(DBTL) 계층에서는 Benchling·Synthace·TeselaGen과 같은 상용 플랫폼과 Aquarium·Opentrons 등 오픈 생태계가 실험 설계·장비 제어·데이터 캡처를 소프트웨어 정의 방식으로 통합한다. 셋째, 오믹스 분석·워크플로 계층에서는 AWS HealthOmics, Google Cloud Life Sciences/Terra, DNAnexus, Seven Bridges, Galaxy가 WDL/Nextflow/CWL을 중심으로 재현 가능한 파이프라인 실행을 지원한다. 넷째, 개방 데이터·지식 리소스 계층에서는 INSDC(ENA/NCBI/DBJ), UniProt, AlphaFold DB, ESM Atlas, DOE KBase·ELIXIR·GA4GH 등 대규모 데이터와 상호운용 표준이 AI 학습·평가의 기반을 제공한다. 이 네 층이 결합될 때, 설계 데이터·실험 메타데이터·오믹스 결과·모델 산출물이 단일 계보(provenance)로 엮이며, 이는 대규모 모델의 신뢰 가능한 학습과 규제 환경에서의 재현성을 동시에 담보한다.

국내의 여건은 강점과 과제가 공존한다. K-BDS는 연구데이터의 수집·품질·분석을 위한

기반을 제공하고, NURION(국가 슈퍼컴)과 KREONET(연구망)은 고성능 계산·데이터 전송의 토대를 제공하고 있다. 더불어 국가 바이오파운드리 구축이 본격화되면서, 연구-산업 간 DBTL 연계와 표준화의 제도적 창구가 열렸다. 컴퓨팅 인프라와 데이터 플랫폼 관점에서 WDL/Nextflow/CWL등의 워크플로 표준 운영과 Dockstore 등록, 온프레미스 HPC와 퍼블릭 클라우드를 잇는 하이브리드 아키텍처 및 비용·보안 거버넌스의 제도화 등의 고민이 필요한 시점이다. 이러한 컴퓨팅 인프라 및 데이터 플랫폼의 체계 수립은 국가 파운드리-민간/학계 실험실-클라우드랩 간 프로토콜·데이터의 상호운용을 현실화하고, 대규모 모델 개발의 데이터 주권과 국제 협력을 동시에 달성하는 길이 된다.

본 보고서는 데이터 플랫폼-연구 자동화-컴퓨팅 인프라의 통합을 중심축으로, 글로벌 사례의 IT 관점 핵심, 국내 현황, 그리고 국제적 표준 기반의 실행형 레퍼런스 아키텍처와 단계별 도입 전략을 제시한다.

2 글로벌 바이오파운드리 인프라

글로벌 선도 바이오파운드리는 고처리량 로봇틱스와 LIMS/ELN, 자동화 워크플로우 오케스트레이션, 고속 분석 파이프라인, 데이터 레이크/웨어하우스, AI 모델 서빙까지 하나의 플랫폼으로 엮어 ‘생물학형 소프트웨어 공급망’을 운영하고 있다.

2.1 Ginkgo Bioworks (미국)

Ginkgo Bioworks는 세계 최대 규모의 상업용 바이오파운드리 플랫폼을 운영하는 기업으로, 합성생물학 분야의 혁신을 주도하고 있다. 이 회사의 파운드리 시설은 10억 달러 이상의 투자를 통해 구축되었으며, 고도로 자동화된 실험실 환경에서 주당 10,000개 이상의 DNA 구조물을 제작할 수 있는 능력을 갖추고 있다.

시설 내부에는 6대의 Illumina NovaSeq을 포함한 최첨단 차세대 시퀀싱 장비가 구비되어 있으며, 144개의 ambr® 250 바이오리액터를 통해 정밀한 발효 최적화 작업을 수행한다. 295,000 평방피트에 달하는 거대한 시설에서 650명 이상의 전문 인력이 근무하며, 주당 5,000개 이상의 질량분석 데이터 포인트를 생성하는 등 대규모 데이터 생산 능력을 보유하고 있다.

Ginkgo의 핵심 자산 중 하나인 Codebase는 약 20억 개의 단백질 서열을 포함하는 방

대한 유전자 데이터베이스다. 이 데이터베이스에는 500만 개 이상의 효소 디자인이 구축되고 테스트된 결과와 함께 100,000개 이상의 AAV 캡시드 특성화 데이터, 메타지놈 효소 라이브러리, 그리고 최적화된 대사 경로 데이터가 축적되어 있다.

특히 주목할 만한 점은 Google Cloud와의 전략적 파트너십을 통해 생성형 AI 플랫폼을 구축했다는 것이다. 이를 통해 DBTL 사이클을 완전히 자동화하고, 바이오인포매틱스, 시스템 생물학, 딥러닝 분석 기능을 통합하여 운영하고 있다. 또한 AWS Batch를 활용함으로써 스토리지 비용을 90%까지 절감하는 성과를 달성했다.

2.2 Zymergen (미국)

Zymergen은 머신러닝과 로봇 자동화를 결합한 독특한 합성생물학 플랫폼을 제시하였다. 이 회사는 75,000개 이상의 생체분자로 구성된 메타지놈 데이터베이스를 구축했으며, AI 기반 유전자 편집 알고리즘을 개발하여 로봇 자동화 플랫폼을 통한 고처리량 실험을 수행했다.

Zymergen의 주요 성과 중 하나는 과학자들이 일반적으로 생각하지 못한 게놈 영역의 70%에서 표현형을 발견했다는 점이다. 이는 AI와 자동화의 결합이 인간의 직관을 넘어서는 발견을 가능하게 한다는 것을 보여주는 사례로 볼 수 있다.

2.3 DOE 시스템 생물학 인프라 (미국)

미국 에너지부(DOE)는 시스템 생물학 연구를 위한 세계적 수준의 공공 인프라를 구축하여 운영하고 있다. 그 중심에는 KBase(Systems Biology Knowledgebase), JGI(Joint Genome Institute), 그리고 NMDC(National Microbiome Data Collaborative)가 있다.

KBase는 DOE가 운영하는 통합 시스템 생물학 플랫폼으로, 시스템 생물학의 예측과 설계라는 도전과제를 해결하기 위해 설계되었다. 현재 28,000명 이상의 전 세계 사용자가 이 플랫폼을 활용하고 있으며, 500TB 이상의 사용자 데이터와 12TB의 참조 데이터가 축적되어 있다. KBase는 90,000개 이상의 미생물 게놈, 50개 이상의 식물 게놈, 300개 이상의 Biolog 배지 포물레이션, 그리고 KEGG, BIGG, MetaCyc 등에서 수집한 30,000개 이상의 반응 및 화합물 데이터를 통합하여 제공한다. 특히 재현 가능한 워크플로우를 위한 Narratives 시스템은 데이터, 시각화, 설명을 하나의 디지털 노트북에서 통합 관리할 수 있게 해준다.

JGI는 로렌스 버클리 국립연구소에 위치한 DOE의 통합 게노믹 과학 사용자 시설로, 연간

200테라베이스 이상의 DNA를 시퀀싱하고 분석하는 능력을 갖추고 있다. 약250명의 과학자와 기술자가 2,180명의 연구 사용자를 지원하며, 15,000개 이상의 DOE JGI 프로젝트가 공개적으로 접근 가능하다. JGI는 17,000개 이상의 주석이 달린 게놈과 메타게놈 데이터를 제공하며, Genome Portal, Phytozome, IMG, Genomes OnLine Database 등의 분석 시스템을 운영한다. 최근에는 KBase팀과 협력하여 “BER Data Lakehouse”라는 데이터 레이크하우스를 구축 중이며, 이는 AI 지원 연구 분석을 위한 통합 인프라 개발의 일환이다.

NMDC는 DOE가 주도하는 국가 마이크로바이옴 데이터 협업 플랫폼으로, FAIR (Findable, Accessible, Interoperable, Reusable) 원칙에 따른 통합 프레임워크를 개발하고 있다. 이 플랫폼은 Submission Portal과 Field Notes 모바일 앱, NMDC EDGE(생물정보학 워크플로우 실행 인터페이스), 그리고 Data Portal과 API를 통해 연구자들에게 서비스를 제공한다. 로렌스 버클리 국립연구소가 주도하고 Oak Ridge, Los Alamos, Pacific Northwest 국립연구소가 파트너로 참여하며, DOE의 고성능 컴퓨팅 시스템을 활용하여 대규모 마이크로바이옴 데이터 분석을 지원한다.

2.4 글로벌 바이오파운드리 인프라 시사점

글로벌 합성생물학 관련 인프라의 사례들에서 알 수 있듯이, 합성생물학 전 과정에서 생산되는 데이터의 통합적인 관리와 저장 활용을 위한 데이터 기반 연구 환경이 필수적으로 구축되고 활용되고 있다. 특히 유전체 및 유전자 서열을 중심으로 한 데이터들과 이를 기반으로 한 유전자 편집 알고리즘 및 AI를 이용할 수 있는 연구 환경 등을 다수의 관련 연구자들이 협력을 통해 목표를 달성할 수 있는 체계의 마련이 중심이 되고 있다. 이러한 합성생물학 연구 수행을 위한 컴퓨팅 인프라 및 연구 환경에서 고려되어야할 점은 아래와 같이 정리해볼 수 있다.

- 데이터 레이크: 연구 수행을 위한 기초 데이터베이스들의 구축 및 연계, 원본 불변 저장(Immutable) + 메타데이터 카탈로그(프로비넌스, 계보/라인리지) + 거버넌스(RBAC/ABAC)
- 워크플로우: WDL/CWL/Nextflow 등의 워크플로우 표준을 이용한 데이터 분석 과정의 표준화, 온프레미스 인프라 및 클라우드 인프라 등을 이용한 분석을 위한 컴퓨팅 파워 제공
- MLOps: 실험 추적(특성/하이퍼파라미터/데이터 버전), 모델 레지스트리(승인·스테이지), 특성 저장소(Feature Store)

3 클라우드 기반 합성생물학 플랫폼

3.1 퍼블릭 클라우드 인프라

주요 클라우드 서비스 제공업체들은 생명과학 연구를 위한 특화된 서비스를 경쟁적으로 출시하며 합성생물학 연구의 디지털 전환을 가속화하고 있다.

3.1.1 AWS (Amazon Web Services)

AWS는 생명과학 연구를 위한 포괄적인 클라우드 서비스를 제공하며, 특히 AWS HealthOmics라는 관리형 서비스를 통해 유전체 데이터의 저장, 질의, 분석을 통합적으로 지원한다. 이 서비스는 HIPAA 적격 인증을 받았으며, WDL, Nextflow, CWL 등 주요 워크플로우 언어를 지원하고, 필요에 따라 10만 개 이상의 vCPU로 확장할 수 있는 능력을 갖추고 있다.

AWS Batch는 Nextflow, Cromwell 등의 바이오인포매틱스 파이프라인을 자동화하는 데 활용되며, Amazon SageMaker는 단백질 구조 예측과 의료 이미지 분할을 위한 AI/ML 모델 훈련에 사용된다. Amazon S3는 내구성 99.999999999%를 제공하는 유전체 데이터 레이크로 활용되고 있다. 실제로 Ginkgo Bioworks는 AWS를 활용하여 컴퓨팅(스토리지) 비용을 최대 90%까지 절감하는 성과를 달성했다.

3.1.2 Google Cloud Platform

Google Cloud Platform은 Cloud Life Sciences API(구 Google Genomics)를 통해 유전체 분석 파이프라인 실행을 지원하며, BigQuery를 통해 대규모 유전체 및 표현형 데이터의 통합 분석을 가능하게 한다. 특히 Ginkgo Bioworks와의 파트너십을 통해 생성형 AI 플랫폼을 구축한 것은 합성생물학과 AI의 융합을 보여주는 대표적인 사례다.

Terra(AnVIL) 플랫폼은 보안 워크스페이스에서 Dockstore, Bioconductor, Galaxy와 연계하여 대형 공공 데이터 분석에 최적화된 환경을 제공한다. Preemptible VM 기술을 활용하여 비용 효율적인 대규모 분석을 지원하며, 연구자들이 제한된 예산으로도 대규모 컴퓨팅 리소스를 활용할 수 있도록 한다.

3.2 퍼블릭 클라우드 인프라 시사점

AWS와 Google Cloud 등 퍼블릭 클라우드 인프라는 여러 가지 장단점이 존재한다. 우선 퍼블릭 클라우드는 다양한 수준에서의 서비스들을 풍부하게 제공하고 있기 때문에 합성생물학 연구의 각 단계별로 또한 용도별로 원하는 기능들을 선택적으로 이용하여 빠른 시간안에 연구 환경 구성이 가능하다. 일 예로 AWS에서는 SageMaker, Bedrock, Comprehend, Kendra 등등 다양한 AI 기반 서비스들을 제공하여 바로 필요한 서비스를 구성하는 것이 가능할 뿐 아니라 AWS Marketplace 등에 등록되어있는 다양한 생명정보 관련 분석 도구 및 서비스들을 이용하여 바로 연구에 활용할 수 있는 체계를 제공하고 있다. 이러한 점은 빠른 시간안에 필요한 서비스 및 환경을 구성하고 연구 활용에 집중할 수 있다는 큰 장점을 제공한다.

다만, 퍼블릭 클라우드 인프라를 중심으로 한 연구 환경 구성시 다음과 같은 단점 역시 존재한다. 퍼블릭 클라우드는 사용량·egress 비용으로 장기·대용량 워크로드의 TCO 예측이 어렵고 높아질 수 있으며, 데이터 주권/규제 준수 비용과 벤더 종속 위험이 있다. 또한 초고성능 GPU/HPC·저지연 스토리지·전용망이 필요한 경우 온프레미 전용 시스템 대비 비용·성능 비효율이 발생할 수 있고, 서비스 수명주기(종료·정책 변경) 리스크가 상존한다. 따라서 국내 현황·규제를 고려한 하이브리드 아키텍처와 데이터 거버넌스 설계가 필수적이다.

4 데이터 관리 및 설계 플랫폼

합성생물학 연구의 효율성과 재현성을 높이기 위해서는 체계적인 데이터 관리와 표준화된 설계 플랫폼이 필수적이다. 이 분야에서는 상업적 플랫폼과 오픈소스 도구들이 상호 보완적으로 발전하고 있다.

4.1 설계·부품 레지스트리 & 표준

합성생물학 분야의 표준화는 SBOL(Synthetic Biology Open Language)을 중심으로 진행되고 있다. SBOL v3는 유전자, 회로, 측정 데이터의 기계 가독 설계 교환 표준으로 자리잡았으며, SBOL Visual을 통해 시각적 표현 방법도 표준화되었다. 이러한 표준은 서로 다른 연구 그룹과 소프트웨어 도구 간의 데이터 교환을 원활하게 만들어 협업과 재현성을 크게 향상시켰다.

SynBioHub는 커뮤니티가 운영하는 SBOL 기반 설계 저장소로, SBOL과 GenBank 형식의 설계를 저장, 공유, 검색할 수 있는 플랫폼이다. REST API와 플러그인 시스템을 통해 다른 도구들과 연계되며, iGEM Registry나 JBEI-ICE와도 통합 운영된다. 공개 허브뿐만 아니라 기관별 사설 인스턴스 구축도 가능하여 지적재산권 보호와 공유의 균형을 맞출 수 있다.

JBEI-ICE는 Joint BioEnergy Institute가 DOE의 지원으로 개발한 오픈소스 부품 레지스트리 플랫폼이다. 플라스미드, 미생물 균주, Arabidopsis 종자 등 다양한 생물학적 부품을 관리할 수 있으며, Web of Registries 기능을 통해 분산된 레지스트리들을 연결하여 운영할 수 있다. 웹 인터페이스와 REST API를 통한 접근이 가능하며, BLAST 검색과 전체 텍스트 검색 기능을 제공한다.

iGEM Registry는 국제 유전공학 머신 대회(iGEM)에서 운영하는 표준부품 카탈로그로, 수만 개의 바이오브릭(BioBrick) 부품을 보유하고 있다. 교육과 연구 생태계에서 가장 널리 사용되는 부품 소스 중 하나로, 특히 학부생과 대학원생들의 합성생물학 입문에 중요한 역할을 한다.

Addgene은 비영리 조직이 운영하는 플라스미드 유통 플랫폼으로, 실제 물질 유통과 함께 디지털 서열 정보를 제공한다. 공식 Developers API와 일일 JSON 덤프를 제공하여 머신러닝 파이프라인과의 연계가 용이하며, 이는 AI 기반 설계 도구 개발에 중요한 학습 데이터 소스가 되고 있다.

4.2 자동화·파운드리 운영 플랫폼

4.2.1 Benchling

Benchling은 생명과학 R&D를 위한 통합 클라우드 플랫폼으로, 분자생물학 연구부터 산업 바이오테크놀로지까지 폭넓은 영역을 지원한다. Molecular Biology 2.0이라 불리는 이 플랫폼의 핵심 기능은 배치 서열 설계, CRISPR 설계, 클로닝 소프트웨어를 포함하며, 고처리량 분자생물학 작업을 효율적으로 수행할 수 있도록 한다.

Bioregistry 기능을 통해 플라스미드, 단백질 서열, 부품 등을 체계적으로 추적 관리할 수 있으며, 전자 실험 노트는 협업과 워크플로우 자동화를 지원한다. 최근에는 AI 기반 도구들을 통합하여 생산성을 크게 향상시켰으며, Benchling API를 통해 발효 데이터의 자동 시각화와 외부 도구와의 통합을 지원한다.

실제 사용 사례로는 Constructive Bio가 메가베이스 규모의 합성 게놈을 설계한 프로젝트와 Synlogic이 발효 데이터를 자동화하고 실시간 진행 상황을 추적한 사례가 있다. 이러한 성공 사례들은 통합 플랫폼의 중요성과 효과를 잘 보여준다.

4.2.2 TeselaGen

TeselaGen은 합성생물학 연구의 전체 DBTL 사이클을 지원하는 통합 플랫폼으로, 각 단계별로 특화된 툴킷을 제공한다. Design Toolkit은 조합 라이브러리 설계와 SBOL 지원을 통해 복잡한 유전자 회로 설계를 가능하게 하며, Build Toolkit은 DNA 합성 및 조립 프로토콜을 자동으로 생성하고 기존 인벤토리를 활용하여 비용을 최적화한다.

Test Toolkit은 서열 검증, NGS 리드 매핑, 변이 식별 등의 기능을 제공하며, Learn Toolkit은 AI 모델 배포와 딥러닝 기술을 통합하여 실험 결과로부터 통찰을 도출한다. 워크플로우 관리 시스템은 실험 프로세스를 자동화하고 오케스트레이션하며, 랩 자동화 모듈은 로봇 시스템과의 원활한 통합을 지원한다.

Technical University of Denmark, Lawrence Berkeley National Laboratory와의 협력을 통해 플랫폼의 기능을 지속적으로 개선하고 있으며, Nature Communications에 머신러닝 모델과 메커니즘 모델을 결합한 연구를 발표하는 등 학술적 성과도 거두고 있다.

4.2.3 오픈소스 플랫폼

오픈소스 생태계에서도 다양한 프로젝트들을 통해 합성생물학 도구의 민주화와 혁신에 중요한 역할들이 추진되고 있다. Aquarium은 워싱턴 대학에서 개발한 오픈소스 LIMS(Laboratory Information Management System)로, 설계-실행-데이터 캡처를 일체형으로 지원한다. GitHub에서 소스코드를 공개하고 컨테이너 기반 배포를 지원하여, 연구실에서 쉽게 도입할 수 있다. 재현 가능한 프로토콜과 완전한 실행 기록을 제공하여 연구의 투명성과 재현성을 보장한다.

Opentrons와 Automata는 저비용 자동화 솔루션을 제공하는 플랫폼들이다. Opentrons는 Python API 기반으로 로봇 액체 핸들링을 제어할 수 있으며, Automata는 LINQ SDK를 통해 워크셀 자동화를 지원한다. 이들은 HTTP API를 통해 외부 시스템과 통합될 수 있으며, SDK를 통해 생성되는 데이터의 구조화가 용이하여 후속 분석과 머신러닝 모델 학습에 활용하기 좋다.

4.3 데이터 관리 및 설계 플랫폼 시사점

글로벌 합성생물학 플랫폼은 SBOL 기반 설계 표준, 자동화 로봇, LIMS/ELN, 오믹스 분석, AI 기반 최적화를 하나의 디지털 DBTL 스택으로 통합하며 연구 생산성과 재현성을 빠르게 끌어올리고 있다. 반면 국내는 일부 기능에 국한된 개별 요소 플랫폼들은 존재하지만, 설계-빌드-실험-데이터-AI를 연결하는 Benchling-TeselaGen 수준의 통합형 DBTL 플랫폼은 아직 부재하다. 이는 한국이 물리적 장비 중심의 파운드리 단계에 머물러 있음을 의미하며, 앞으로는 자동화-워크플로-HPC/클라우드-MLOps를 통합한 디지털 파운드리 구축이 국가 경쟁력의 핵심 과제가 될 것으로 판단된다. 현재 국제적으로도 미국 등을 제외하고는 DBTL 전과정에 대한 소프트웨어 스택은 과도기이며, 한국이 전략적으로 이런 역할을 할 수 있는 플랫폼의 주요 기능들을 설계하고 통합 활용할 수 있는 디지털 합성생물학 생태계 구축에 적기라고 말할 수 있다.

5 AI 기반 단백질 설계 및 예측 플랫폼

5.1 AlphaFold (DeepMind/Google)

AlphaFold는 DeepMind가 개발한 딥러닝 기반 단백질 구조 예측 시스템으로, 단백질 구조 예측 분야의 패러다임을 완전히 바꾼 혁신적인 기술이다. AlphaFold 2는 CASP14 대회에서 실험적 구조와 경쟁할 수 있는 수준의 정확도를 달성했으며, 중위 GDT_TS 92.4(전체 타겟)라는 놀라운 성과를 기록했다. 이 시스템은 진화적, 물리적, 기하학적 제약을 통합하고, Evoformer 블록과 구조 모듈을 통한 반복 정제 과정을 거쳐 고정밀 구조를 예측한다.

AlphaFold Protein Structure Database는 이 기술의 실용적 활용을 극대화한 결과물이다. 2억 개 이상의 단백질 구조 예측을 무료로 제공하며, 대화형 시각화 도구와 함께 잔기별 신뢰도 추정(pLDDT) 정보를 제공한다. 이 데이터베이스는 UniRef90 데이터셋의 대부분을 커버하며, UniProt 항목의 85% 이상과 통합되어 있어 연구자들이 쉽게 접근하고 활용할 수 있다. 최신 AlphaFold Server는 AlphaFold 3를 기반으로 하여 단백질-DNA/RNA/리간드 상호작용까지 예측할 수 있도록 기능이 확장되었다.

5.2 Rosetta (University of Washington)

워싱턴 대학에서 개발한 Rosetta Software Suite는 단백질 구조의 계산 모델링과 분석을 위한 포괄적인 알고리즘 모음이다. 이 시스템은 고정 백본 단백질 설계, 측쇄 형태 예측, 단백질-단백질 도킹 등 다양한 기능을 제공하며, Monte Carlo 최적화와 담금질 시뮬레이션을 통해 에너지 최소화 구조를 찾는다. Rosetta의 가장 큰 강점은 실험적으로 검증된 수많은 설계 성공 사례를 보유하고 있다는 점이다.

Rosetta@home은 전 세계 약55,000대의 자원봉사 컴퓨터를 활용한 분산 컴퓨팅 프로젝트로, 일반 시민들의 컴퓨터 자원을 활용하여 대규모 단백질 구조 예측과 설계를 수행한다. 이 프로젝트는 CASP와CAPRI 실험에서 지속적으로 우수한 성능을 입증해왔다. RosettaDesign Server는 웹 기반 인터페이스를 통해 단백질 설계 시뮬레이션을 제공하며, 20개 아미노산 모두를 고려한 전체 단백질 재설계를 지원한다. 이는 단백질 안정화, 결합 친화도 향상, 새로운 단백질 구조 설계 등 다양한 응용 분야에 활용되고 있다.

5.3 Deep Genomics (캐나다)

Deep Genomics는 RNA 생물학에 특화된 AI 플랫폼을 개발한 캐나다 기업으로, BigRNA 플랫폼을 통해 RNA 기반 치료제 개발을 혁신하고 있다. 이 플랫폼은 1조 개의 유전체 신호와 18억 개의 조정 가능한 매개변수를 가진 파운데이션 모델로, 조직별 RNA 발현 조절 메커니즘을 정확하게 예측할 수 있다.

BigRNA는 단백질 및microRNA 결합 부위를 식별하고, 스테릭 블로킹 올리고뉴클레오타이드 설계를 자동화하여 RNA 기반 치료제 개발 과정을 크게 단축시킨다. AI Workbench는 10억 개 이상의 분자를 100만 개 이상의 표적에 대해 평가할 수 있는 능력을 갖추고 있으며, 70%의 성공률로 신규 표적에서 후보 물질을 도출한다. 일반적으로 18개월 내에 후보 발견이 가능하며, 온타겟 효과와 게놈 전체 오프타겟 효과 데이터를 동시에 생성하여 안전성과 효능을 종합적으로 평가한다.

5.4 ESM (Evolutionary Scale Modeling)

Meta AI가 개발한 ESM(Evolutionary Scale Modeling)은 대규모 언어 모델 기술을 단백질 서열 분석에 적용한 혁신적인 접근법이다. ESM Metagenomic Atlas는 6.17억 개의 메타지노믹 단백질 구조 예측을 제공하는 방대한 데이터베이스로, 이는 자연계에 존재하는 단백질 다양성의 상당 부분을 포함한다.

이 모델은 특히 효소 설계와 도메인 스캐폴드 탐색에 유용하며, 기존에 알려지지 않은 새로운 단백질 구조와 기능을 발견하는 데 기여하고 있다. ESM의 접근법은 진화적 정보를 대규모로 학습하여 단백질의 구조와 기능을 예측하며, 이는 합성생물학에서 새로운 효소나 단백질 부품을 설계하는 데 중요한 자원이 되고 있다.

6 국내 합성생물학 인프라 현황

우리나라는 합성생물학을 국가 전략기술로 지정하고 체계적인 인프라 구축을 추진하고 있다. 연구 기술 수준은 세계 4위로 평가되지만, 바이오파운드리와 같은 핵심 인프라 부족으로 산업 생태계 형성에 어려움을 겪어왔다.

6.1 국가 바이오파운드리 구축 사업

2024년 1월, ‘바이오파운드리 인프라 및 활용기반 구축 사업’이 예비타당성 조사를 통과하면서 국내 합성생물학 인프라 구축이 본격화되었다. 이 사업은 2025년부터 2029년까지 5년간 총 1,263억원이 투입되는 대규모 국가 프로젝트로, 과학기술정보통신부와 산업통상자원부가 공동으로 주관하며, 한국생명공학연구원의 국가바이오파운드리사업단이 주관기관 역할을 수행한다.

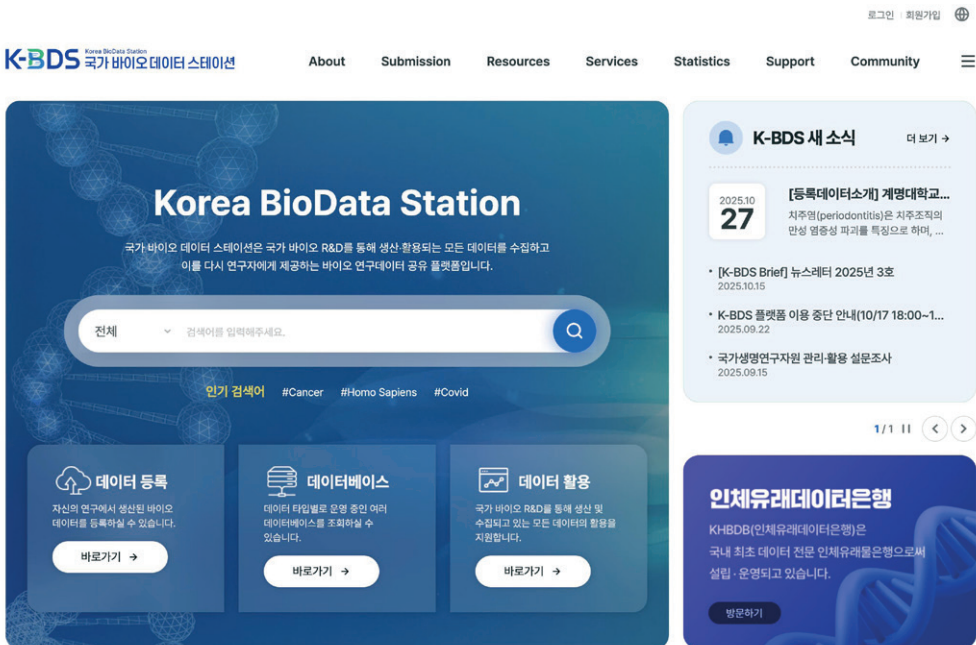
사업의 핵심 내용은 크게 네 가지로 구성된다. 첫째, 바이오파운드리 전용 센터를 건립하는 것으로, 2025년에 설계를 시작하여 2027년 완공을 목표로 한다. 둘째, 합성생물학 연구 과정의 각 단계별 자동화와 표준화 작업을 수행할 수 있는 핵심 워크플로우를 개발한다. 셋째, 바이오파운드리 인프라의 효율적 운영을 위한 통합 플랫폼을 2025년부터 2029년까지 단계적으로 구축한다. 넷째, 자동화, 지능화, 기기 간 연결 기술 등 디지털화 기술을 개발하여 핵심 기기 및 장비를 고도화한다.

이 사업을 통해 우리나라는 합성생물학 기술경쟁력을 제고하고, 바이오제조 가속화 기반을 마련하며, 연구개발 및 사업화 기간을 단축할 것으로 기대된다. 특히 AI와 연계한 혁신 기술 발전의 토대가 될 것으로 전망된다.

다만, 앞절의 해외 사례와 같은 데이터 기반 디지털 합성생물학 연구 생태계 조성을 위한 컴퓨팅인프라 및 플랫폼 등에 대한 고민이 추가적으로 필요한 것으로 보인다.

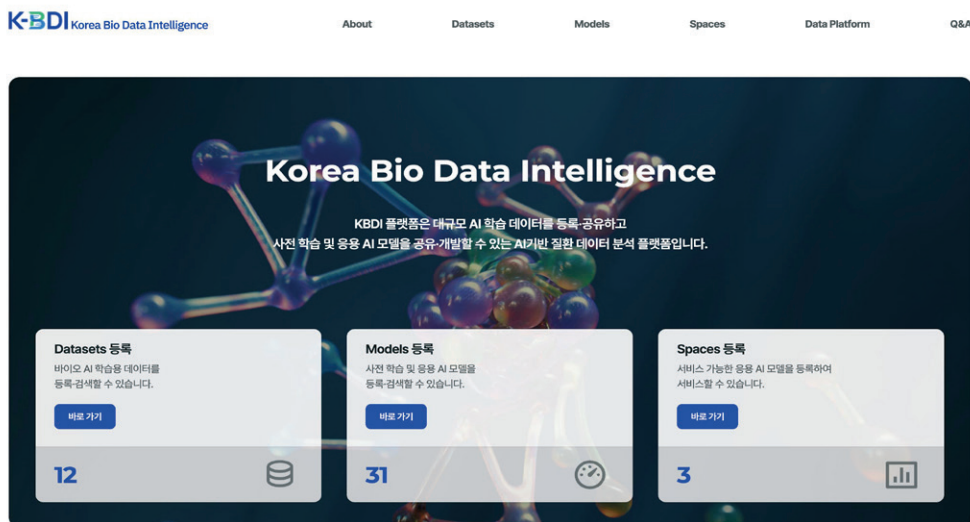
6.2 바이오 데이터 스테이션

국내 바이오 데이터 인프라는 KOBIC과 한국과학기술정보연구원(KISTI)을 중심으로 구축되어 있다. K-BDS(Korea BioData Station)는 범부처 바이오 R&D 데이터 허브로, 데이터 수집, 품질 관리, 제공을 통합적으로 수행한다. 데이터 타입별 표준 등록 양식을 제공하여 품질과 상호운용성을 강화하며, 국내 생성 데이터의 FAIR 원칙 준수와 데이터 주권 확보에 기여하고 있다.



BioProject, BioSample, KRA(차세대 시퀀싱), KArray(마이크로어레이), KNA(염기서열), KPOP(단백체 MS), KMAP(대사체 MS/NMR), KCODE(화합물/활성), KBI(바이오 이미징), KPC(감염병 전임상), GeNA(기타)와 같은 데이터 분류를 통해 바이오 분야의 다양한 데이터를 포괄적으로 수집하고 활용할 수 있는 체계를 제공중이다.

이와 더불어 바이오 데이터를 이용한 AI연구 등을 포함한 데이터 활용 연구 지원을 위한 K-BDS Cloud(cloud.kbds.re.kr)의 데이터 활용지원 프로그램 서비스를 운영중이며 연간 70개 이상의 연구과제에 NVIDIA A100 GPU 등을 활용할 수 있도록 지원하고 있다. 또한, 연구자들 간의 데이터 기반 연구 협력을 활성화하기 위한 K-BDS Marketplace 서비스를 제공하고 있다.



또한 K-BDS 사업의 일환으로 한국연구재단에서는 ‘데이터 기반 디지털바이오 선도시스템 개발 사업’을 통해 Korea Bio Data Intelligence(K-BDI, www.kbdi.re.kr)를 구축하고 서비스를 제공하고 있다. K-BDI를 통해 국내 바이오 R&D 사업을 통해 구축되고 공유 활용 될 수 있는 바이오 분야 AI 학습용 데이터를 등록하고 공유할 수 있도록 기능을 제공할 뿐 아니라, 바이오 분야 AI 모델들을 등록하고 공유할 수 있는 AI 모델 저장소의 기능을 함께 제공하고 있다. 특히, 주요 분야의 데이터들을 이용하여 학습된 pre-trained model을 제공하고 있으며 연구자들은 K-BDI에서 제공하는 컴퓨팅 인프라를 이용하여 개인의 AI모델을 직접 개발하거나 개발된 모델을 서비스하고 이를 이용할 수 있는 바이오 분야 AI 생태계를 조성하고자하는 노력을 함께 추진하고 있다.

이러한 국가R&D사업을 통해 생산되는 다양한 바이오 데이터의 수집/활용 체계를 국가적으로 마련하고 제공하고 있지만, 설계·실험 자동화·DBTL 워크플로우와는 연결되기 위한 방안이 마련되어있지 않으며, SynBioHub/ICE 같은 합성생물학 설계 표준 스택도 부재한 상태이다. 향후 K-BDS의 주요 서비스들을 연계 활용하여 국가적 디지털 합성생물학 생태계를 구축하는 과정이 필요하다. K-BDS는 OpenAPI/클라우드 지원·Marketplace를 통해 데이터→분석→공유의 사용자 여정을 제공하며, 향후 DRS/htsget/refget 수용과 자동화/워크플로 표준 연계가 이루어지면 국가 디지털 파운드리 중추가 될 수 있다.

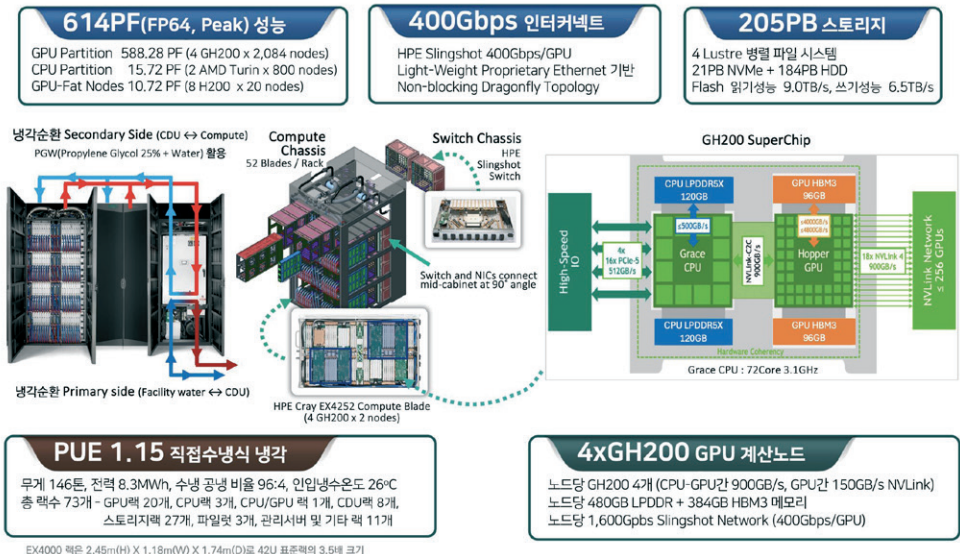
6.3 국가슈퍼컴퓨팅센터

KISTI가 운영하는 NURION은 국가 5호 슈퍼컴퓨터로, 25.7 페타플롭스급 성능을 보유하고 있다. 2018년 Top500 리스트에서 11위를 기록한 이 시스템은 대용량 병렬 파일시스

템과 버스트 버퍼를 갖추고 있어, 단백질 구조 예측이나 단백질 설계와 같은 대규모 학습, 튜닝, 시뮬레이션에 활용되고 있다. 5호기는 26년도까지 운영이 될 예정이며 이후 슈퍼컴 6호기를 도입하여 시를 비롯한 다양한 대규모 연산이 필요한 분야에 활용될 예정이다.



슈퍼컴퓨터 6호기는 614PFlops의 성능을 제공할 예정이며, 이는 기존 5호기 대비 약 23배 이상의 성능을 제공하는 것으로 최대한의 성능을 이끌어내기 위해 NVIDIA H200 GPU를 약 8,500장을 도입하여 제공한다. 이와 더불어 다양한 데이터 기반 연구 지원을 위해 총 약 200PB 이상의 데이터 저장소를 제공하는 국내 최대 규모 컴퓨팅 인프라로 26년도 하반기 서비스가 제공될 계획이다. 특히, 대규모 GPU 및 스토리지 인프라를 통해 합성생물학과 관련되어있는 다양한 바이오 분야의 거대 문제 해결 및 AI 모델 개발 및 시를 이용한 DBTL 사이클 연구 수행에 다방면으로 활용 될 수 있을 것으로 기대된다.



KREONET은 KISTI가 운영하는 국가 연구망으로, 테라급 백본(100-1,200 Gbps/세그먼트)을 제공한다. 국제 연구망과 연동되어 있어 대용량 데이터의 국제적 교류를 지원하며, 클라우드 다이렉트 연결의 필요성이 제기되고 있어 향후 클라우드 기반 연구 인프라와의 통합이 기대된다.

구분	5호기 누리온 (NURIION)	6호기 (도입 중)
연산 성능(이론)	25.7 PF	약 600 PF (GPU 가속)
주요 HW	Cray CS500, Intel Xeon Phi 7250 (KNL) + 일부 Skylake, Omni-Path	NVIDIA GH200 포함 GPU 8,496장, 초고속 네트워크 >400 Gbps
스토리지	병렬 파일시스템 + 버스트버퍼	약 205 PB(계획) + 초고속 병렬 FS
네트워크	Omni-Path	>400 Gbps급 전용 네트워크
예상 효과	대규모 시뮬레이션/정적 모델 학습	대형 생성형 모델·멀티모달 생물학 모델 학습 가속, 대용량 파이프라인 처리량 증대
상태/일정	'18 서비스 개시 '26년 운영 종료 예정	'26 상반기 구축 목표(계약 체결 완료)

KISTI의 국가슈퍼컴 및 과학기술연구망 등 합성생물학을 비롯한 다양한 데이터 기반 대규모 연구 지원을 위한 컴퓨팅 인프라가 도입되어 이용되고 있다. 다만, 국가슈퍼컴은 계산과학 분야에서 공유활용을 위해 구축된 장비로, 특정 연구 서비스, 특히 바이오파운드리와의 연계성을 통한 DBTL의 통합 구현을 위한 요소 장비로 활용될 수 있는지에 대한 고민이 필요하며, 국가적으로 구축된 인프라들의 연계성을 통해 더욱 효과적인 합성생물학 생태계 구축이 가능할 것으로 판단된다.

7 권장 아키텍처 및 도입 전략

성공적인 합성생물학 AI 인프라 구축을 위해서는 체계적인 아키텍처 설계와 단계적 도입 전략이 필요하다.

7.1 레퍼런스 아키텍처

효과적인 합성생물학 AI 인프라는 계층화된 구조로 설계되어야 한다. 데이터 계층에서는 다양한 소스로부터 데이터를 수집하고 관리한다. 원천 데이터는 SBOL/SynBioHub를 통한 설계 데이터, ELN/LIMS를 통한 실험 및 로트 데이터, 그리고 FASTQ/BAM/VCF 형식의 오믹스 데이터를 포함한다.

데이터 수집과 식별은 GA4GH의 표준들을 중심으로 표준화되어 저장/전송될 수 있도록 구성되어야한다. 저장은 하이브리드 방식을 채택하여 온프레미스 HPC, 오브젝트 스토리지, 퍼블릭 클라우드를 상황에 맞게 활용할 수 있도록 구성하여 국내 법/제도적인 문제와 참여 기관들의 데이터 보안 이슈 등을 해결하며 연구가 추진될 수 있는 체계 마련이 필요하다.

분석과 오케스트레이션 계층에서는 WDL, Nextflow, CWL 등의 표준 워크플로우 언어를 사용한다. DNAnexus, AWS HealthOmics, Galaxy 등과 같은 데이터 분석 플랫폼을 활용하여 워크플로우를 실행하며, 모델 학습과 추론 파이프라인은 Docker나 Singularity를 통해 컨테이너화하고 Dockstore에 등록하여 관리한다.

자동화 연동 계층에서는 실험 자동화 플랫폼을 통합한다. Synthace, TeselaGen, Aquarium, Opentrons, Automata 과 같은 플랫폼을 이용하여 기관의 요구사항에 맞게 환경을 구성하고 제공할 필요가 있다.

7.2 도입 체크리스트

성공적인 인프라 도입을 위해서는 체계적인 접근이 필요하다. 첫째, 데이터 표준화가 기본이 되어야 한다. SBOL, ISA, MixS와GA4GH 표준을 기본으로, 설계-실험-오믹스 데이터 전주기에 걸친 메타데이터 스키마를 수립하여야 한다.

둘째, 워크플로우 표준화가 중요하다. 신규 파이프라인은 WDL, Nextflow, CWL 중 하나로 작성하고, Dockstore에 등록하며, GitOps 방식으로 운영한다. 이는 재현성과 버전 관리를 보장한다.

셋째, 자동화 통합 전략이 필요하다. 실험 자동화 플랫폼을 1-2종 선정하고, SBOL 설계에서 프로토콜 변환까지의 경로를 확보한다. 예를 들어 Synthace나 TeselaGen과 같은 상업 플랫폼과 Opentrons나 Automata 같은 오픈 플랫폼을 조합하여 사용할 수 있다.

넷째, 하이브리드 컴퓨팅 전략을 수립한다. 대규모 학습은 NURION과 같은 HPC나 GPU 클러스터에서 수행하고, 운영과 공유는 퍼블릭 클라우드를 활용한다. AWS나 Google Cloud Platform 등과 같은 클라우드 서비스를 이용하여 인프라의 확장성과 활용성을 높일 필요가 있다.

8

맺음말

합성생물학 AI 개발을 위한 컴퓨팅 인프라는 바이오파운드리, 클라우드 랩, AI 플랫폼, 대규모 데이터베이스 등이 유기적으로 연결된 복잡한 생태계를 형성하고 있다. Ginkgo Bioworks와 같은 선도 기업들은 이미 자동화, AI, 클라우드 컴퓨팅을 완전히 통합한 혁신적 플랫폼을 운영하고 있으며, AWS, Google Cloud 등 주요 클라우드 제공업체들도 생명과학 특화 서비스를 경쟁적으로 강화하고 있다.

국제적으로는 Global Biofoundries Alliance를 중심으로 표준화와 협력이 활발히 진행되고 있으며, 미국의 DOE KBase, NMDC, JGI 등 공공 인프라는 FAIR 데이터 원칙에 따라 대규모 데이터를 통합 관리하고 있으며, ELIXIR, GA4GH 등 국제 협력체는 데이터 공유와 상호운용성을 위한 표준을 지속적으로 개발하고 있다.

국내에서는 1,263억원 규모의 국가 바이오파운드리 구축 사업이 2025년부터 본격 추진되어, 한국생명공학연구원을 중심으로 국제 협력 네트워크가 구축되고 있으며, K-BDS, KISTI 국가슈퍼컴퓨터 등 기존 인프라를 활용한 통합 생태계 구축이 진행 중이다. 이는 우리나라가 합성생물학 분야의 국제적 위상을 더욱 공고히 하고, 산업 생태계를 활성화하는 계기가 될 것이다.

향후 합성생물학 분야는 AI 기반 설계, 클라우드 기반 자동화, 자가구동 실험실 등의 기술 혁신을 통해 더욱 빠르게 발전할 전망이다. 특히 생성형 AI와 대규모 언어 모델의 도입은 생물학적 설계와 실험 계획 수립 방식을 근본적으로 변화시킬 것이다. 이러한 기술 발전은 기후변화, 식량 안보, 맞춤형 의료 등 인류가 직면한 주요 과제 해결에 중요한 기여를 할 것으로 기대된다.

성공적인 합성생물학 AI 인프라 구축을 위해서는 표준화, 상호운용성, 보안, 지속가능성을 종합적으로 고려해야 한다. SBOL, GA4GH 등 국제 표준의 적극적 채택과 함께 국내 실정에 맞는 인프라 구축 전략 수립이 중요하다. 또한 산업계와 학계의 적극적인 참여를 통한 생태계 조성이 성공의 핵심 요인이 될 것이다. 이를 통해 우리나라가 합성생물학과 AI 융합 분야의 글로벌 리더로 도약할 수 있을 것으로 기대한다.

[참고문헌]

- [1] Arkin, A.P., et al. (2018). KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nature Biotechnology*, 36, 566-569. <https://doi.org/10.1038/nbt.4163>
- [1] Beal, J., et al. (2020). The Synthetic Biology Open Language (SBOL) Version 3.0.0. *Journal of Integrative Bioinformatics*, 17(2-3), 20200017.
- [2] Carbonell, P., et al. (2019). An automated Design-Build-Test-Learn pipeline for enhanced microbial production of fine chemicals. *Communications Biology*, 2, 147.
- [3] Chao, R., Mishra, S., Si, T., & Zhao, H. (2017). Engineering biological systems using automated biofoundries. *Metabolic Engineering*, 42, 98-108.
- [4] Deep Genomics. (2024). AI Platform for RNA Therapeutics. <https://www.deepgenomics.com/AI-Platform/>
- [5] Dixon, T.A., Freemont, P.S., Johnson, R.A., & Pretorius, I.S. (2022). A global forum on synthetic biology: the need for international engagement. *Nature Communications*, 13, 3516.
- [6] Eloë-Fadrosh, E.A., et al. (2022). The National Microbiome Data Collaborative Data Portal: an integrated multi-omics microbiome data resource. *Nucleic Acids Research*, 50(D1), D828-D836.
- [7] Freemont, P.S. (2019). Synthetic biology industry: data-driven design is creating new opportunities in biotechnology. *Emerging Topics in Life Sciences*, 3(5), 651-657.
- [8] GA4GH. (2024). Genomic Data Toolkit. <https://www.ga4gh.org/genomic-data-toolkit/>
- [9] Gardner, T.S., et al. (2022). Synthetic biology: from hype to impact. *Trends in Biotechnology*, 40(12), 1425-1436.
- [10] Ginkgo Bioworks. (2024). Platform Overview. <https://www.ginkgobioworks>.

com/platform/

- [11] Google and Ginkgo. (2024). Foundry-Scale Data Meets AI. <https://ai.ginkgo.bio/resources/blog/google-and-ginkgo-foundry-scale-data-meets-ai>
- [12] Grigoriev, I.V., et al. (2014). The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Research*, 42(D1), D26-D31.
- [13] Hillson, N., et al. (2019). Building a global alliance of biofoundries. *Nature Communications*, 10, 2040.
- [14] Holowko, M.B., Frow, E.K., Reid, J.C., Rourke, M., & Vickers, C.E. (2021). Building a biofoundry. *Synthetic Biology*, 6(1), ysaa026.
- [15] Hutchison, C.A., et al. (2016). Design and synthesis of a minimal bacterial genome. *Science*, 351(6280), aad6253.
- [16] Jumper, J., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583-589.
- [17] Kelwick, R., et al. (2020). Developments in the tools and methodologies of synthetic biology. *Frontiers in Bioengineering and Biotechnology*, 8, 834.
- [18] Kitney, R., & Freemont, P. (2012). Synthetic biology – the state of play. *FEBS Letters*, 586(15), 2029-2036.
- [19] McLaughlin, J.A., et al. (2018). SynBioHub: A standards-enabled design repository for synthetic biology. *ACS Synthetic Biology*, 7(2), 682-688.
- [20] Nielsen, A.A., et al. (2016). Genetic circuit design automation. *Science*, 352(6281), aac7341.
- [21] Purnick, P.E., & Weiss, R. (2009). The second wave of synthetic biology: from modules to systems. *Nature Reviews Molecular Cell Biology*, 10(6), 410-422.
- [22] Richardson, S.M., et al. (2017). Design of a synthetic yeast genome. *Science*, 355(6329), 1040-1044.
- [23] Rosano, G.L., et al. (2019). New tools for recombinant protein production in *Escherichia coli*: A 5-year update. *Protein Science*, 28(8), 1412-1422.

- [24] Smanski, M.J., et al. (2016). Synthetic biology to access and expand nature's chemical diversity. *Nature Reviews Microbiology*, 14(3), 135-149.
- [25] Strateos. (2024). Cloud Laboratory Platform. <https://strateos.com/>
- [26] Tellechea-Luzardo, J., et al. (2022). Fast biofoundries: coping with the challenges of biomanufacturing. *Trends in Biotechnology*, 40(7), 831-842.
- [27] TeselaGen. (2024). Operating System Overview. <https://teselagen.com/teselagen-operating-system-overview/>
- [28] The UniProt Consortium. (2023). UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1), D523-D531.
- [29] Varadi, M., et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1), D439-D444.
- [30] Vickers, C.E., & Freemont, P.S. (2022). Pandemic preparedness: synthetic biology and publicly funded biofoundries can rapidly accelerate response time. *Nature Communications*, 13, 753.
- [31] Walsh, D.I., et al. (2019). Standardized, automated microfluidically generated DNA assembly. *ACS Synthetic Biology*, 8(1), 97-106.
- [32] Wood-Charlson, E.M., et al. (2020). The National Microbiome Data Collaborative: enabling microbiome science. *Nature Reviews Microbiology*, 18, 313-314.
- [33] Woolston, B.M., Edgar, S., & Stephanopoulos, G. (2013). Metabolic engineering: past and future. *Annual Review of Chemical and Biomolecular Engineering*, 4, 259-288.
- [34] Young, E.M., et al. (2022). Iterative algorithm-guided design of massive strain libraries, applied to itaconic acid production in yeast. *Metabolic Engineering*, 70, 45-58.
- [35] DOE KBase. <https://www.kbase.us/>
- [36] ELIXIR. <https://elixir-europe.org/>

- [37] Emerald Cloud Lab. <https://www.emeraldcloudlab.com>
- [38] K-BDS. <https://www.kobic.re.kr/kbds/>
- [39] KREONET. <https://www.kreonet.net/>
- [40] NURION. <https://www.ksc.re.kr/>
- [41] Terra. <https://terra.bio/>

[주요용어풀이]

- 데이터 레이크 / 레이크하우스 (Data Lake / Lakehouse): 원시 데이터(파일·객체)를 포맷 제한 없이 모으는 저장소가 데이터 레이크, 여기에 테이블 형식(거버넌스·트랜잭션)까지 결합해 분석·운영을 통합한 형태가 레이크하우스.
- 메타데이터 카탈로그·프로비던스: 데이터의 출처·가공 이력·사용 내역을 구조화해 추적하는 체계. 누가(Who)·언제(When)·무엇을(What)·어떻게(How) 만들었는지 기록해 재현성과 감사 가능성을 높임.
- 워크플로 표준 (WDL/CWL/Nextflow DSL): 분석 파이프라인을 선언적으로 기술하는 언어/DSL. 실행 환경(온프레/HPC/클라우드)과 도구를 분리해 이식성과 재현성을 확보함.
- 컨테이너·이미지 레지스트리 (Docker/OCI, Singularity/Apptainer): 실행 환경(의존성·라이브러리)을 ‘이미지’로 고정해 어디서나 동일하게 실행하도록 하는 기술과, 그 이미지를 보관·배포하는 저장소.
- 실행 오케스트레이션 (Slurm, Cromwell/Nextflow Tower, Kubernetes/Batch): 다양한 자원(GPU·CPU등)을 조율해 파이프라인을 대규모로 안정 실행하는 스케줄러/플랫폼. 재시작·스케일링·모니터링을 표준화함.
- GA4GH 데이터 액세스 표준 (DRS / htsgget / refget): DRS는 데이터 객체를 위치와 무관한 표준 ID로 식별·획득, htsgget은 BAM/CRAM/VCF의 부분 범위 스트리밍, refget은 참조서열을 해시 ID로 일관 관리해 상호운용·재현성을 보장.
- 객체 스토리지 수명주기·계층화 (Lifecycle & Tiering): S3류 오브젝트 저장소의 핫/웜/콜드 계층과 보존정책을 자동 적용해 비용을 최적화하는 기법(버전관리·아카이브·삭제 규칙 포함).
- MLOps (실험 추적·모델 레지스트리/CD·가드레일): 데이터/모델 버전관리, 하이퍼파라미터·지표 기록, 승인된 모델만 배포하는 레지스트리, 자동 검증·보안 정책을 포함한 ML 운영 체계.

AI 기반 유전자 및 단백질 설계 기술 동향

1 머리말

합성생물학과 AI의 융합은 생명공학 분야에 혁신적인 변화를 일으키고 있다. 특히 AlphaFold와 같은 AI 모델의 등장은 50년 간 이어진 단백질 구조 예측 문제를 해결하여, 모든 알려진 단백질의 3차원 구조를 거의 실험적 정확도로 예측하는 쾌거를 이루었다. 실제로 DeepMind의 AlphaFold2는 2021년 Nature Methods에서 단백질 구조 예측 분야의 올해의 방법으로 선정될 만큼 큰 반향을 일으켰으며, 수분~수시간 내에 단백질 접힘 구조를 정밀하게 예측함으로써 시간과 비용을 획기적으로 절감했다. 그 결과 AlphaFold2는 전 세계 단백질 서열을 대상으로 2억건이 넘는 3차원 구조 예측 모델을 공개하였으며, 이는 기존 실험 기반 구조생물학이 접근하기 어려웠던 대규모 구조 정보를 단기간에 확보한 획기적 전환점이 되었다.

AlphaFold의 성공 이후 RoseTTAFold, ESMFold, OmegaFold 등 다양한 구조예측 모델이 잇따라 등장하면서, 단백질 구조 예측이 데이터 기반 예측 과학에서 AI-생성 기반 생명 설계로 전환되고 있다. 이러한 변화는 생물학을 실험 중심에서 계산 중심으로 이끄는 거대한 패러다임 전환이라 할 수 있다. 이제 연구자들은 AlphaFold가 예측한 거대 규모의 단백질 구조 데이터베이스를 자유롭게 활용하여 단백질 기능을 이해하고 상호작용을 모델링하고 있다.

이처럼 AI 기반 예측 기술이 생명과학의 난제를 풀어내자, 이제는 한 걸음 더 나아가 생성형 AI로 새로운 생물학적 설계를 하는 단계로 진입하고 있다. 딥러닝의 발전, 특히 트랜스포머 기반 대규모 언어 모델(LLM)의 부상으로, DNA나 단백질 서열로부터 물리적 결과나 기능을 예측하는 복잡한 작업까지 수행할 수 있게 되었다.

더 나아가 이러한 생성형 AI는 단순 예측을 넘어, 생물학적 설계 자체를 자동화하는 방향으로 발전하고 있다. 예를 들어 AI가 주어진 목적에 맞는 최적의 유전자 서열이나 단백질 서열을 스스로 설계 및 제안함으로써, 인간의 직관과 시행착오에 의존하던 생물설계 과정이 설계-구현-실험-학습의 자동화 파이프라인으로 전환되고 있다. 이러한 변화는 합성생물학의 설계 사이클을 비약적으로 가속화할 뿐 아니라, 사람이 쉽게 고안하기 어려운 복잡한 새로운 생물 시스템까지도 AI가 모색할 수 있도록 한다. 실제로 일부 기업들은 자사의 AI 플랫폼으로 특정 표적에 결합하는 신규 단백질 서열을 생성해내는 등 AI를 활용

한 de novo 단백질 디자인에 뛰어 들고 있다.

합성생물학 전문가들에게 AI와의 융합은 이제 선택이 아닌 필수 도구로 자리매김하고 있으며, 본 보고서에서는 AI 기반 유전자 및 단백질 설계의 최신 기술 동향을 살펴보고자 한다.

2 생성형 AI를 활용한 DNA/RNA 서열 설계

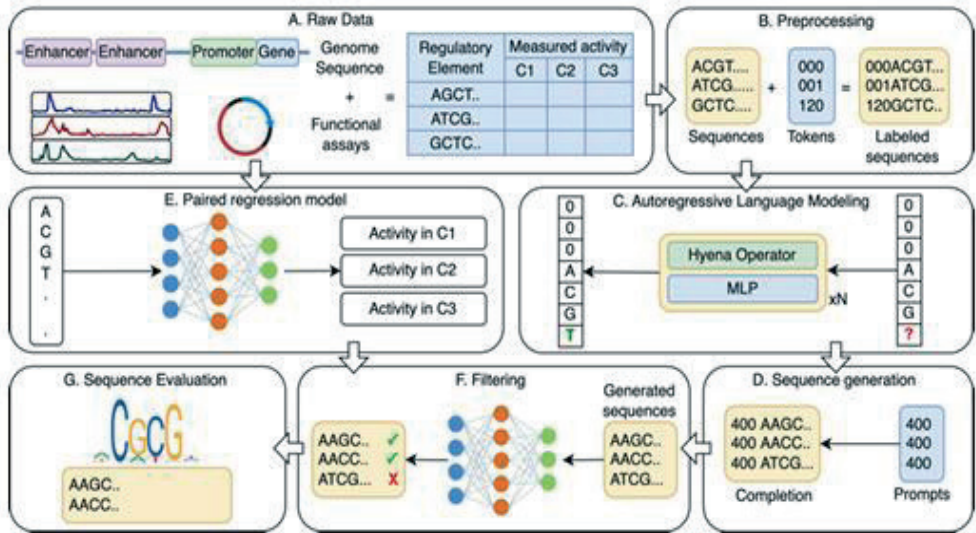
생성형 AI의 도입으로, 인간이 미처 고안하지 못했던 새로운 유전자 서열을 목표 기능에 맞게 설계하는 일이 현실화되고 있다. 과거에는 유전자 발현을 높이거나 조절하기 위해 코돈 최적화 등 제한된 범위의 수작업 개량이 이루어졌다면, 이제 딥러닝 생성 모델이 방대한 서열 공간을 탐색하며 지능적으로 서열을 창출해낼 수 있다. 특히 프로모터, 인핸서 등의 비부호화 조절 DNA 서열은 전사인자 결합 모티프의 조합으로 유전자 발현 수준을 결정하는데, 이러한 시스-조절요소의 최적 설계에 생성형 AI가 큰 강점을 보이고 있다. 생성 모델은 자연 서열의 통계와 규칙을 학습하여 특정 세포 타입에서 강하게 혹은 약하게 발현되는 합성 프로모터/인핸서 서열을 직접 만들어낼 수 있으며, 이렇게 디자인된 서열은 기존 자연 서열과 유사한 특성 분포를 가져 세포 내에서 예측 불가능한 이상 현상 없이 원하는 발현 조절 기능을 수행할 것으로 기대된다.

2.1 특정 기능 및 발현을 최적화된 유전자 서열 디자인

합성생물학에서 흔히 요구되는 것 중 하나는 특정 기능을 수행하면서 발현이 잘 되는 유전자를 만드는 것이다. 생성형 AI는 이러한 요구에 부응하여, 목표 단백질의 발현량을 극대화하거나 원하는 동작을 하도록 DNA 서열을 디자인할 수 있다. Autoregressive Language Model을 DNA 서열에 적용한 대표적인 사례로 Genentech 연구진의 regLM을 들 수 있다.

regLM은 HyenaDNA 등의 트랜스포머 기반 자연어 처리 기법을 DNA 생성에 응용하여, 높은 발현 활성이나 낮은 활성, 세포 유형 특이적 활성 등 지정한 활성을 보이는 합성 프로모터/증폭자 서열을 자동으로 생성해낸다^[1].

Schematic of regLM



이 프레임워크로 연구진은 효모와 인간 세포에서 동작하는 다양한 합성 조절 서열들을 설계하였고, AI가 만들어낸 서열들이 실제 실험적으로 검증된 조절요소들과 유사한 motifs 특징을 지니면서도 원하는 활성 수준을 보일 것으로 예측되었다. 흥미롭게도, regLM 처럼 생성적 접근법은 기존의 강화학습 기반 서열 최적화나 지도학습 예측 모델을 이용한 염기치환 방식과 달리, 자연 서열의 분포를 학습해 현실성 높은 서열을 만들어낸다는 장점이 있다. 따라서 단순히 예측 점수를 높이는 변이 조합을 찍어내는 이전 방식보다, 생성 모델이 설계한 서열은 세포 내 맥락에서 안정적으로 작동할 가능성이 높다^[2].

오토회귀 언어 모델 (Autoregressive Language Models, ARLM)

원리 및 학습 방식

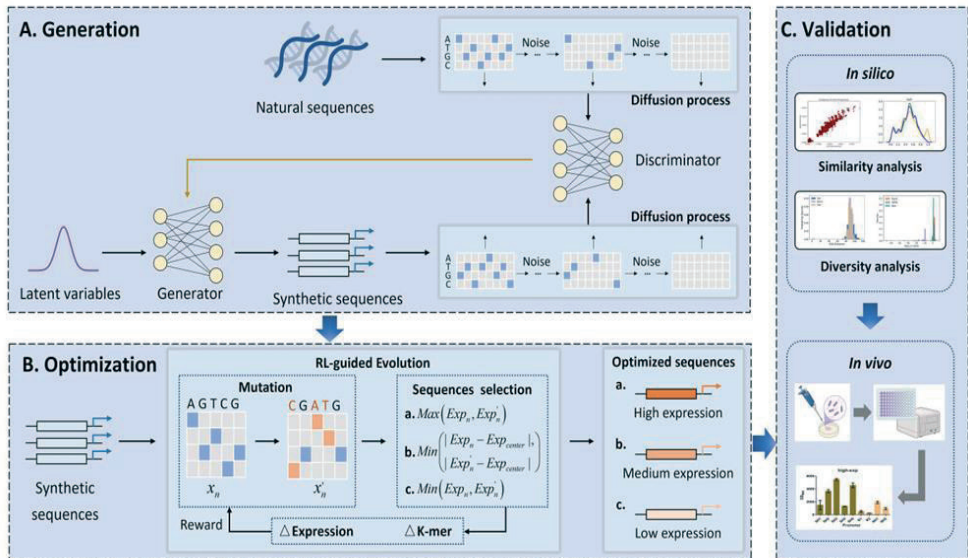
대규모 데이터와 모델 파라미터를 활용한 스케일 업이 ARLM 발전을 주도했다. 예컨대 세일즈포스 연구진이 발표한 ProGen 모델은 2023년 Nature Biotechnology에 보고된 초기 사례로, 2억 개 이상의 아미노산으로 구성된 방대한 단백질 서열 코퍼스를 학습하여 새롭고 기능적인 단백질 서열들을 생성할 수 있음을 입증했다. 후속 연구에서는 모델 크기와 데이터 규모를 크게 늘린 ProGen2와 ProGen3가 등장했다. 2025년 공개된 ProGen3는 혼합전문가(MoE) 구조를 도입해 파라미터 수를 46억 개까지 확장하고, 3.4억 개의 단백질(1.1조 아미노산 토큰)에 달하는 데이터로 학습되었다^[3]. 그 결과 이전 세대보다 12배 이상 많은 데이터를 학습한 이 모델은 서열 생성뿐 아니라 중간 삽입/보충 생성까지 가능해졌다. 규모 확장을 통해 모델은 서열 상의 제약 학습 능력과 표현력이 비약적으로 향상되었으며, 실험실 데이터로 후처리 정렬하면 안정성 예측 등에서 추가 성능 향상이 있음을 보였다.

최근
성능 향상

대규모 데이터와 모델 파라미터를 활용한 스케일 업이 ARLM 발전을 주도했다. 예컨대 세일즈포스 연구진이 발표한 ProGen 모델은 2023년 Nature Biotechnology에 보고된 초기 사례로, 2억 개 이상의 아미노산으로 구성된 방대한 단백질 서열 코퍼스를 학습하여 새롭고 기능적인 단백질 서열들을 생성할 수 있음을 입증했다. 후속 연구에서는 모델 크기와 데이터 규모를 크게 늘린 ProGen2와 ProGen3가 등장했다. 2025년 공개된 ProGen3는 혼합전문가(MoE) 구조를 도입해 파라미터 수를 46억 개까지 확장하고, 3.4억 개의 단백질(1.1조 아미노산 토큰)에 달하는 데이터로 학습되었다^[3]. 그 결과 이전 세대보다 12배 이상 많은 데이터를 학습한 이 모델은 서열 생성뿐 아니라 중간 삽입/보충 생성까지 가능해졌다. 규모 확장을 통해 모델은 서열 상의 제약 학습 능력과 표현력이 비약적으로 향상되었으며, 실험실 데이터로 후처리 정렬하면 안정성 예측 등에서 추가 성능 향상이 있음을 보였다.

생성형 AI 기법 중 생성적 적대신경망(GAN)이나 확산(diffusion) 모델도 유전자 서열 디자인에 응용되고 있다. 2025년 보고된 PromoDGDE 기법은 Diffusion-GAN 모델을 기반으로 한 프로모터 생성 방법으로, 강력한 프로모터부터 약한 프로모터까지 다양한 발현 강도의 합성 프로모터를 목표 수치에 맞춰 설계하는 것을 목표로 한다.

Promoter design process based on PromoDGDE



이 방법은 확산모델로 자연 프로모터 서열의 분포를 학습하여 일차 후보를 만든 뒤, 강화 학습과 유전 알고리즘으로 해당 서열들을 발현 강도 예측값이 원하는 수준에 수렴하도록 동적으로 최적화한다. 그 결과 E. coli와 효모를 대상으로 한 in vivo 검증 실험에서,

PromoDGDE가 생성한 합성 프로모터들의 60% 이상이 실제로 예상했던 발현 조절 효과를 보임으로써 설계 성공률을 크게 높였다. 이는 이전에 보고된 GAN 기반 프로모터 설계 성공률(~70% 기능적 프로모터 생성)과 비교해서도 손색없는 수준으로, 생성모델과 최적화 알고리즘 결합 접근법의 높은 효율을 보여준다^[4].

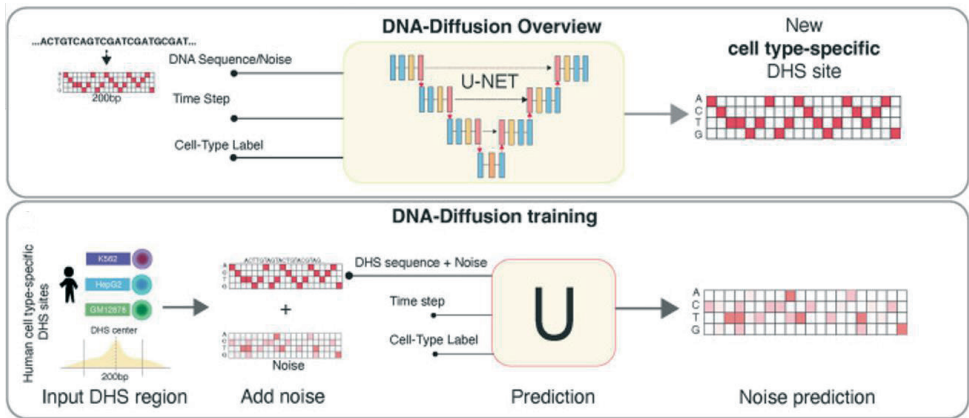
생성적 적대 신경망 (Generative Adversarial Networks, GAN)

원리 및 특징

GAN은 생성자(Generator)와 판별자(Discriminator)라는 두 신경망이 경쟁적으로 학습하는 생성 모델 구조이다. 생성자는 무작위 노이즈로부터 가짜 데이터를 만들어내고, 판별자는 입력 데이터가 진짜인지 가짜)인지 판별한다. 두 네트워크는 미니맥스 게임처럼 서로를 속이고 판별하는 과정에서 동시에 향상되어, 훈련이 잘 진행되면 생성자는 판별자가 구별할 수 없을 만큼 진짜와 흡사한 데이터를 만들어내게 된다. 2014년 Goodfellow 등이 GAN을 처음 제안한 이후, 딥컨볼루션 GAN(DCGAN), 조건부 GAN(cGAN), CycleGAN, StyleGAN 등 다양한 변형이 등장하여 이미지 생성 분야를 중심으로 혁신을 일으켰다. 특히 StyleGAN 계열은 고해상도의 포토리얼리스틱 얼굴 이미지를 생성할 만큼 품질을 끌어올렸고, CycleGAN은 이미지 간 변환(예: 사진을 그림으로)에서 준수한 성과를 보였습니다. 다만 GAN 훈련은 매우 불안정하기 쉽고, 모드 붕괴나 훈련 불발 등의 문제가 흔히 발생한다. 이를 해결하기 위해 여러 기술적 개선이 도입되었는데, Wasserstein GAN(WGAN)은 손실 함수를 변경하여 안정적인 학습을 유도했고, WGAN-GP는 그래디언트 페널티로 판별자의 Lipschitz 조건을 확보해 훈련을 더욱 안정화했다. 또한 Self-Attention GAN(SAGAN)은 셀프 어텐션으로 이미지의 장기 의존성을 잘 포착하여 일관성 있는 세부 묘사를 가능하게 했다. 이 외에도 스펙트럼 정규화, 프로그레시브 러닝 등 기법들이 GAN의 생성 품질과 다양성을 향상시켰다.

한편 확산 모델(diffusion model)을 직접 유전체 서열 생성에 적용한 사례로 Harvard 의대 Pinello 교수팀의 DNA-Diffusion 연구가 주목된다. 이 접근법은 이미지 생성 확산 모델을 DNA 시퀀스에 응용한 것으로, 모델에 간암 세포에서 활성화되는 강화인자 서열과 같은 조건을 주면 그에 맞는 맞춤형 합성 조절 서열을 만들어낼 수 있다. 놀라운 점은, DNA-Diffusion 모델이 동일한 표적 유전자에 대해서도 다양한 세기의 활성 조절 서열을 생성할 수 있다는 것이다. 마치 유전자 발현의 볼륨 스위치를 조절하듯, 완전히 켜거나 완전히 켜 상태만이 아니라 중간 단계의 발현율까지 세밀하게 달성하는 서열들을 만들어냈다는 보고가 있다^[5].

Schematic of DNA-Diffusion



이는 향후 정밀한 유전자 발현 제어가 필요한 유전자 치료나 대사공학에 유용하게 쓰일 수 있을 것이다. 현재 DNA-Diffusion에서 생성한 서열들의 실험적 효율성 검증이 진행 중이지만 이러한 확산 모델 기반 설계는 기존의 GAN, RNN 등과 더불어 합성 생물학적 서열 디자인의 새로운 도구로 부상하고 있다. 이렇듯 생성형 AI를 통해 특정 기능에 최적화된 유전자 서열 디자인이 빠르게 발전하고 있다. 이러한 AI 기반 서열 설계 기술은 점차 합성 생물학 실험 자동화 플랫폼과 결합되어, 모델이 제안한 서열을 로봇이 합성·측정하고 그 결과가 다시 모델 학습으로 피드백되는 순환적 설계 체계로 확장되고 있다. 이러한 AI-실험 통합 사이클은 생물설계의 속도와 정밀도를 비약적으로 향상시킬 것으로 기대된다(6).

요약하면, 합성 프로모터 분야에서 언어모델 기반 regLM이 세포 맞춤형 조절 서열을 현실감 있게 설계하고, 확산과 강화학습 복합기법인 PromoDGDE가 원하는 발현 강도의 프로모터를 높은 성공률로 만들어내며, 이미지 확산모델 응용까지 등장하여 유전자 발현 조절의 연속적 튜닝도 가능해지고 있다. 이러한 기술들은 향후 유전자회로 설계나 대규모 발현 조절 네트워크 합성에도 응용되어, 인간 공학적 통찰과 AI의 학습능력이 결합된 새로운 생물설계 패러다임을 이룰 것으로 기대된다.

확산 모델 (Diffusion Model)

원리 및 특성

확산 모델은 데이터에 점진적으로 노이즈를 추가하는 과정과 그 반대과정(denoising)을 학습하여 새로운 데이터를 생성하는 확률적 생성 모델이다. 학습 시에는 깨끗한 데이터에 단계적으로 잡음을 섞어 나가며 노이즈 첨가 과정을 정의하고, 모델은 이 과정을 역으로 수행하여 임의의 노이즈로부터 원본같은 데이터를 복원하도록 훈련된다.

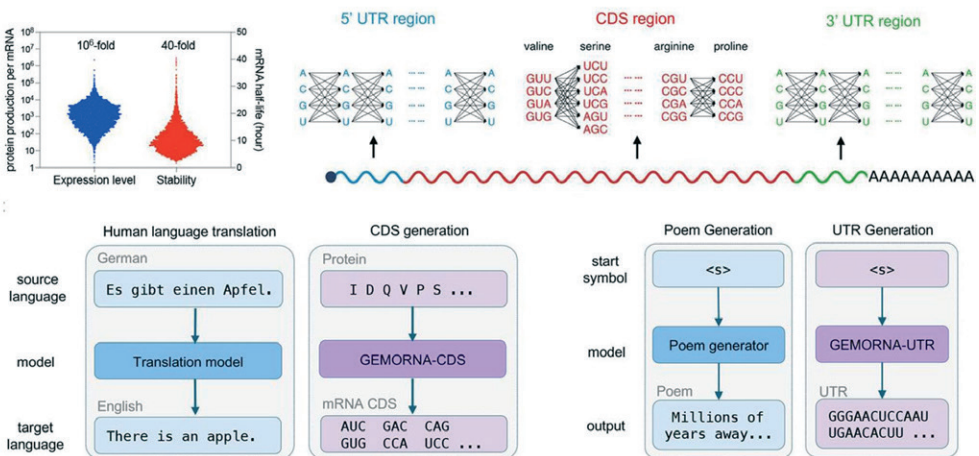
원리 및 특성

생성 시에는 아무 정보가 없는 가우시안 노이즈에서 시작해 모델이 예측한대로 잡음을 여러 단계에 걸쳐 제거하면, 마치 사진이 점차 뚜렷해지듯 의미 있는 최종 샘플이 얻어진다. 이러한 노이즈 제거 생성 접근은 이론적으로 안정적이고 다양성을 보장하며, 대규모 U-Net과 같은 딥러닝 아키텍처의 결합으로 이미지 생성 등에서 GAN을 능가하는 성능을 보여주었다. 예컨대 Stable Diffusion이나 DALL-E 2 같은 모델들은 확산 모델 기반으로 텍스트 조건부 이미지를 고품질로 생성해낸다. 확산 모델의 단점이었던 느린 생성 속도를 개선하기 위해 최근에는 샘플링 단계 수를 줄이는 고속 알고리즘(예: DDIM, DPM-Solver), 잠재 확산 모델(Latent Diffusion: 입력을 저차원 잠재공간으로 압축하여 연산량 감소) 등이 도입되어 실용성을 높였다. 그 결과 확산 모델들은 소분자 생성, 영상 및 음성 생성 등 다양한 영역에서 새로운 SOTA를 달성하고 있으며, 안정적인 학습과 높은 생성 품질 덕분에 범용 생성 AI의 주류로 자리잡고 있다.

2.2 mRNA 백신 서열 최적화 및 개발 가속화 사례

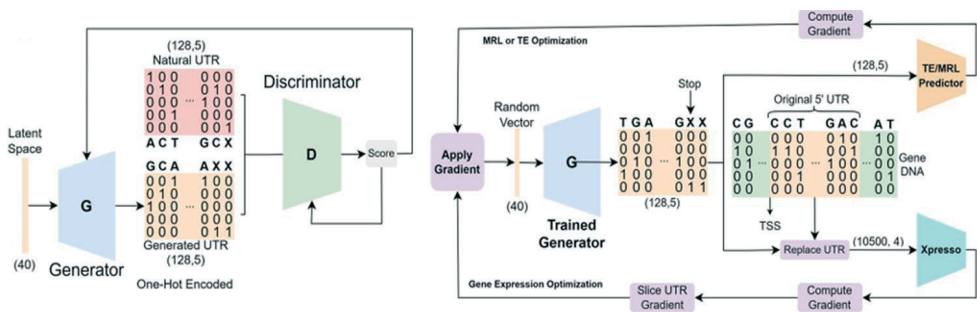
mRNA 백신 및 치료용 mRNA 분야는 생성형 AI의 활약이 두드러진 영역 중 하나이다. mRNA 기반 치료제는 개발 속도가 빠르고 변이에 유연하지만, 낮은 안정성과 번역효율이 한계로 지적되어 왔다. AI는 이러한 문제를 서열 설계 단계에서부터 해결하고 나아가 개발 기간 자체를 단축하는 데 활용되고 있다. 2025년 Science에 보고된 GEMORNA 플랫폼은 생성형 AI를 이용한 mRNA 서열 최적화의 획기적인 성과를 보여주었다. 이 연구에서 제시된 GEMORNA는 트랜스포머 계열의 생성 모델로, mRNA 코딩서열(CDS)과 번역 비조직(UTR)을 모두 설계 대상으로 삼아 단백질 발현량과 mRNA 안정성을 극대화하는 새로운 mRNA 서열들을 생성했다.

Overview of designing therapeutic mRNAs with GEMORNA



그 결과 GEMORNA가 설계한 mRNA는 기존에 최적화된 표준 서열 대비 in vitro에서 루시퍼레이스 발현을 최대 41배까지 높였고, 치료 단백질인 에리트로포이에틴(EPO)의 발현은 15배 향상시켰다. 나아가 동일한 플랫폼으로 디자인한 COVID-19 백신용 mRNA를 생쥐에 면역한 경우, 기존 서열로 유도한 면역반응보다 현저히 높은 항체 역가를 얻는 데 성공했다. 흥미로운 점은 이 AI가 circular RNA 서열에도 응용되어, 고안된 순환형 mRNA가 EPO를 높은 수준으로 발현함은 물론 CAR-T 세포의 항암 살해능을 증진시키는 사례도 확인되었다는 것이다. 이러한 연구는 딥러닝 생성 모델이 mRNA 백신과 치료제 개발에 있어 전례 없는 성능 개선을 이끌어낼 수 있음을 증명한다^[6]. 실제로 GEMORNA 플랫폼을 개발한 Raina Bio에서는 생성 AI를 활용하여 신약 개발 타임라인을 획기적으로 단축하고 mRNA 의약품의 새로운 적용 분야를 열겠다고 발표하고 있어, AI가 백신/신약 개발의 게임체인저로 자리잡을 가능성을 시사한다. mRNA의 번역 효율을 높이기 위해서는 5' 및 3' UTR 서열, 코돈 구성, mRNA 2차 구조 등 여러 요소를 최적화해야 한다. 이러한 다차원 최적화 문제 역시 AI가 해결하고 있는데, 2025년 보고된 UTRGAN 모델이 한 예이다. UTRGAN은 5' UTR 서열 생성에 특화된 GAN 기반 모델로, 생성된 UTR 서열의 번역 효율을 예측하는 판별자와의 경쟁을 통해 고효율 UTR을 만들어낸다. 이 모델은 자연 UTR의 다양한 통계적 특성을 모사하면서도 표적 유전자에 대한 리보솜 결합능과 안정성을 극대화하도록 서열을 진화시켰다. 그 결과 UTRGAN이 생성한 5' UTR은 초기 서열들에 비해 예측된 단백질 발현량을 평균 5배까지, 리보솜 부하율을 2배까지 향상시켰으며, 번역 효율 지표는 무려 34배 상승하는 놀라운 성능을 보였다. 더 중요한 것은, 단순 계산상 도출된 서열이 아닌 AI 생성 UTR 서열을 실제 실험 검증한 결과, UTRGAN이 설계한 서열이 인간 TNF- α 단백질을 기존 강력한 5' UTR(β -globin UTR)보다도 더 높은 번역률로 발현시켰다는 점이다^[7]. 이는 AI가 mRNA 서열의 미세한 부분까지 설계하여 in vitro 실험에서 유의미한 개선을 이루었음을 보여주며, 향후 mRNA 백신 개발시 기존에 인위적으로 조합하던 UTR 라이브러리를 AI로 자동 설계하는 방향을 제시한다.

Generative and optimization architectures of UTRGAN



또 다른 관점에서, AI 알고리즘을 활용한 mRNA 서열 최적화도 주목할 만하다. 딥러닝 기반이 아니더라도, 방대한 서열 공간을 탐색하는 알고리즘은 생성형 AI와 유사한 효과를 낼 수 있다. 예를 들어 2023년 Nature 논문에서는 LinearDesign이라는 알고리즘을 발표하여, SARS-CoV-2 스파이크 단백질을 코딩하는 최적 mRNA 서열을 단 11분 만에 계산해냈다. 이 알고리즘은 전통적인 언어처리 기법을 응용하여 mRNA 2차 구조의 안정성과 코돈 최적화를 동시에 달성하도록 서열을 탐색한 것으로, 그 결과물이 가진 성능은 놀라웠다. LinearDesign으로 생성한 mRNA 백신 서열은 기존 코돈 최적화만 거친 서열보다 mRNA 반감기와 단백질 발현량을 크게 향상시켰고, 특히 생쥐 실험에서 항체 생성량이 기존 대비 최대 128배까지 증가하는 등 백신 효능을 비약적으로 높였다. 이는 AI를 활용해 mRNA 서열의 구조적 안정성까지 고려한 설계가 얼마나 중요한지 보여주는 결과로서, 해당 연구는 원리 기반 mRNA 설계(principled mRNA design)의 잠재력을 입증했다라고 강조하고 있다. mRNA 설계에는 코돈 최적화뿐 아니라 GC 비율, UTR 안정성, 이차구조 접힘, 면역 자극 회피 등 물리화학적 제약이 복합적으로 작용한다. LinearDesign은 이러한 요소를 동적계획법 기반의 알고리즘으로 통합해 최적 서열을 탐색함으로써, AI와 계산생물학의 경계를 허문 대표적 사례로 평가된다. LinearDesign은 특정 생성형 신경망을 쓴 것은 아니지만, 전통 알고리즘과 AI의 경계가 모호할 정도로 복잡한 최적화 문제를 해결하여 mRNA 백신 개발을 가속화한 사례라 할 수 있다. 특히 Baidu 연구진은 LinearDesign을 상용화 단계로 발전시켜, LinearDesign-P를 통해 단백질 수준의 조건부 설계까지 구현하고 있다. 이러한 접근은 mRNA 구조의 안정성과 발현 효율을 동시에 최적화할 수 있는 차세대 백신 설계 전략으로 주목받고 있다.

LinearDesign

Linear Design의 핵심 원리

LinearDesign은 긴 RNA 염기서열을 목표한 2차 구조에 맞게 안정적으로 설계하기 위해 개발된 효율적인 알고리즘이다. 이 알고리즘은 특히 mRNA 백신이나 RNA 기반 치료제의 서열 최적화에서 핵심적인 역할을 한다. RNA 설계의 본질적인 목표는 주어진 목표 구조를 가장 안정적으로 형성할 수 있는 염기서열을 찾는 것이지만, RNA의 길이가 늘어남에 따라 가능한 염기쌍 조합의 수가 지수적으로 증가하기 때문에 계산량이 매우 크다. 기존의 RNA 구조 예측 알고리즘들은 구조 예측에는 유용하지만, 역으로 서열을 설계하는 데에는 속도와 효율성의 한계가 있었다. LinearDesign의 핵심 아이디어는 이러한 계산 복잡도를 극복하기 위해 동적계획법을 기반으로 한 점수 최적화 접근법을 사용하는 것이다. 즉, 주어진 목표 구조 내의 각 염기쌍에 대해 안정성을 나타내는 점수를 정의하고, 전체 서열의 총 점수를 최대화하는 조합을 탐색한다.

Linear
Design의
핵심 원리

이때 RNA 전체 에너지를 지역적으로 더할 수 있는 구조로 모델링함으로써, 동적계획법을 통해 효율적인 최적화가 가능해진다. 그러나 단순한 DP만으로는 여전히 계산량이 많기 때문에, LinearDesign은 Beam Search 기법을 결합하여 탐색 공간을 획기적으로 줄였다. Beam Search는 모든 가능한 경로를 탐색하는 대신, 각 단계에서 상위 b개의 유망한 후보만 남기고 나머지는 가지치기하는 방식이다. 이 덕분에 기존 $O(n^3)$ 복잡도를 가지던 RNA 구조 탐색 문제를 $O(n \cdot b)$ 의 준선형 시간으로 해결할 수 있게 되었다. 이러한 접근은 Baidu가 개발한 RNA 구조 예측 알고리즘 LinearFold에서 영감을 받은 것이다. 또한 mRNA 설계에서는 단백질의 아미노산 서열이 이미 고정되어 있기 때문에, 각 코돈이 선택할 수 있는 염기 조합이 제한되어 있다. LinearDesign은 동적계획 과정에서도 이러한 코돈 제약을 직접 반영하여, 가능한 코돈 조합 중에서 구조적으로 가장 안정적인 서열을 찾는다. 동시에 mRNA의 안정성(ΔG)과 코돈 적합도(Codon Adaptation Index)를 함께 고려함으로써, 단백질 발현 효율과 RNA 구조 안정성을 동시에 최적화할 수 있다.

결국 LinearDesign은 목표 RNA 구조에 맞는 최적 서열을 빠르고 정확하게 찾는 알고리즘으로, 동적계획법 기반의 점수 최적화, Beam Search를 통한 선형화 탐색, 그리고 코돈 제약 통합이라는 세 가지 핵심 요소를 결합하고 있다. 이러한 혁신적인 접근 덕분에 기존 방법으로는 불가능했던 수천 염기 길이의 mRNA 서열도 효율적으로 설계할 수 있게 되었으며, 실제로 Baidu 연구진이 2020년 발표한 LinearDesign은 mRNA의 반감기와 단백질 발현 효율을 크게 향상시키는 데 성공했다. 이후 LinearDesign은 LinearDesign-P 등으로 확장되어 단백질 수준의 설계까지 고려할 수 있게 되었으며, 오늘날 RNA 기반 치료제와 백신 개발 분야에서 가장 주목받는 서열 설계 기술 중 하나로 자리 잡았다.

요약하면, 생성형 AI 및 알고리즘을 통한 mRNA 서열 설계 분야에서는 GEMORNA 같은 딥러닝 기반 플랫폼이 백신/단백질 발현을 획기적으로 향상시켰고, UTRGAN 등 특화 모델이 UTR을 정밀 설계하여 번역효율을 높였으며, LinearDesign 알고리즘 등은 2차구조 안정성과 코돈 사용을 모두 만족하는 최적 서열을 빠르게 찾아냈다. 이러한 기술들은 코로나19 팬데믹을 계기로 꽃핀 mRNA 백신 기술을 한층 더 발전시켜, 변이에 강인하고 고효율인 차세대 백신 개발을 앞당기고 있다. 나아가 AI가 제안한 서열을 토대로 실제로 백신을 제작하고 시험하는 설계-실험 주기가 빨라지면서, 향후 새로운 전염병 출현 시 신속 대응 백신 개발에도 AI가 핵심 역할을 할 것으로 전망된다.

3 AI 기반 단백질 구조 예측 및 de novo 디자인 (AlphaFold 이후의 혁신)

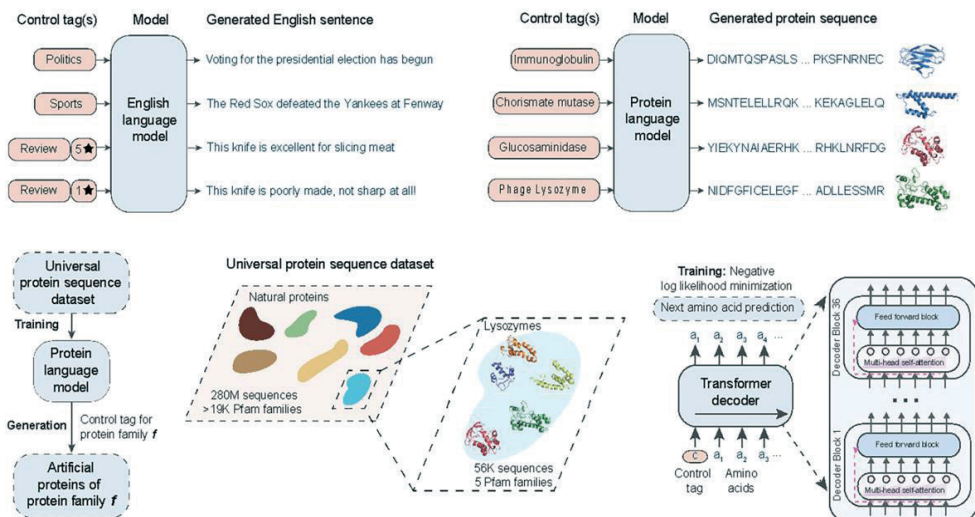
AlphaFold2의 성공 이후, 단백질 분야에서는 구조 예측을 넘어 새로운 단백질 설계로 초점이 이동하고 있다. AlphaFold2는 CASP14 대회에서 경쟁자들을 압도하며 평균 원자 오차 1Å 내외의 정확도로 단백질 입체구조를 예측해 구조생물학계에 충격을 주었고, 인간 단백질 포함 전 종단위 프로테옴의 구조까지 예측함으로써 사실상 단일 폴딩 단백질 구조 예측 문제를 해결했다는 평가를 받았다^[8]. 실제로 방대한 단백질 프로테옴에 대한 예측 3D 구조 모델이 공개되어, 생물학 연구자들은 새로운 실험 없이도 관심 단백질의 구조 모델을 즉시 활용할 수 있게 되었다. AlphaFold의 이러한 획기적 업적은 단백질 공학 분야에도 큰 영향을 주었는데, 이제 연구자들은 AI 예측 구조를 기반으로 단백질 기능을 해석하거나 리간드 결합 부위를 예측하며, 더 나아가 신규 단백질 디자인의 정확성 검증에도 AlphaFold를 활용하고 있다. 예를 들어 de novo 단백질을 설계한 후 AlphaFold로 접힘 여부를 확인하는 접근법은 최근 몇 년간 표준적인 검증 단계로 자리잡았다.

AlphaFold 이후 남은 구조생물학 과제로는 다중 단백질 상호작용 복합체, 구조 유연성이 큰 단백질, 작은 화합물/리간드 결합 등 복잡한 문제들이 있다. 이를 위해 AlphaFold를 확장한 AlphaFold-Multimer나, 2024년 발표된 AlphaFold 3 모델이 단백질-단백질 및 기타 분자 상호작용까지 보다 정확히 예측하는 방향으로 진화하고 있다. AlphaFold 3는 특히 리간드, RNA, DNA 등 비단백질 분자와의 상호작용을 동시에 모델링할 수 있도록 개선되어, 생체 복합체 수준의 예측을 가능하게 했다. 이러한 다중 상호작용 예측은 약물 결합 해석 및 단백질 네트워크 분석에서도 큰 파급효과를 보이고 있다. DeepMind는 2023년 말 AlphaFold의 차기 버전에서 구조 예측 정확도를 더욱 높이고 리간드와 복합체까지 다룰 수 있게 개선했음을 예고하였고, 2024년에는 AlphaFold 3가 발표되어 복합체-리간드 상호작용 예측이 확장되었고, 연구자 접근 채널이 제공되고 있다. 이처럼 AI 기반 구조예측은 지속적으로 고도화되어 이제는 모든 생체 분자들의 구조와 상호작용을 예측하는 방향으로 나아가고 있다^[9].

한편, 구조 예측 기술의 성숙은 단백질 설계 분야의 혁신적인 도약을 촉발했다. 더 이상 자연계에 존재하는 단백질의 구조를 맞추는 데에만 머물지 않고, AI를 이용해 전혀 새로운 구조와 기능을 지닌 단백질을 설계하는 시도가 활발해진 것이다. 과거에는 단백질 설계가 주로 경험적 지식과 제한적 컴퓨터 설계에 의존했다면, 이제 생성형 AI가 방대한 단백질 서열/구조 데이터를 학습하여 기존 제약에 맞는 완전히 새로운 단백질 서열을 창출하고 있다. 예를 들어 Absci와 같은 기업들은 항체 분야에 AI를 도입해, 특정 표적에 강하

게 결합하면서도 인간에게 면역원성이 낮은 신규 항체 서열을 생성해내고 있다고 알려져 있다. 또한 Meta(페이스북) AI 연구진은 ESMFold라는 대형 언어모델 기반 구조예측을 선보이고 이를 활용한 단백질 생성 실험도 진행하였으며, Salesforce 연구진은 ProGen이라는 대규모 언어 모델을 훈련시켜 완전히 새로운 효소 서열들을 만들어내는 데 성공했습니다^[10]. 이러한 움직임은 단백질 공학에 있어 AI의 위상이 구조 예측에서 설계로 전환되고 있음을 의미한다. 최근에는 ProGen 외에도 ProtGPT2, ProteinMPNN, Chroma 등의 모델이 등장하여 구조-서열 동시 설계를 구현하고 있다. 이들 모델은 AlphaFold로 예측된 구조를 즉시 검증에 활용함으로써, 설계-검증의 폐루프를 완성하고 있다.

Artificial protein generation with ProGen



특히 2023년 Nature Biotechnology에 보고된 ProGen 연구는 생성형 AI로 기능적 단백질을 설계한 대표적 사례이다. 연구진은 2억 8천만 개에 달하는 자연 단백질 서열 데이터를 대형 언어모델에 학습시킨 후, 라이소자임과 같은 단백질 유형 태그를 조건으로 주어 해당 계열의 새로운 단백질 서열을 생성했다. ProGen이 만들어낸 수백만 개의 후보 중 일부를 합성하여 시험한 결과, 몇몇 인공 효소들이 자연 효소와 동등한 수준의 활성을 나타냈다.

놀랍게도, 이들 AI-설계 효소는 서열적으로 자연계와 상당히 달랐음에도 기능을 수행했는데, 예를 들어 두 개의 생성 효소는 서로 서열 유사성이 18% 정도에 불과했지만 모두 닭의 흰자위 리소자임에 견줄만한 세균 세포벽 분해 활성을 보였고, 각각의 서열은 알려

진 어떤 자연 단백질과도 90%, 70% 정도만 일치하는 매우 새로운 서열이었다. 더 나아가 추가 선별에서는 생성 효소 서열의 단 31%만이 알려진 단백질과 유사해도 활성을 나타내는 경우가 확인되어, 자연 단백질이라면 아미노산 하나만 변해도 기능을 잃을 수 있는 것에 비해 AI 생성 단백질은 기존 진화 경로와 다른 새로운 해법으로 기능을 구현하고 있음을 보여주었다. 또한 흥미로운 점은, 이러한 인공 단백질들의 3차 구조를 X선 결정학으로 분석한 결과 예상된 접힘 구조와 일치하는 올바른 구조를 형성하고 있었다는 것이다. 이는 ProGen 모델이 아미노산 서열의 문법만 배운 것이 아니라, 단백질 입체구조 형성에 필요한 규칙까지도 어느 정도 학습했음을 시사한다. 요약하면, ProGen 연구는 대규모 언어모델이 진화의 제약을 넘어 완전히 새로운 단백질을 설계할 수 있으며, 그 결과물도 충분히 기능하고 구조를 이룰 수 있다는 것을 입증하였다.

전문가들은 이를 두고 단백질 설계 분야가 새로운 시대에 진입했다고 평가하며, 향후 약물 개발이나 산업효소 분야 등에서 AI로부터 창발한 혁신 설계가 속속 등장할 것으로 기대하고 있다. Diffusion 모델 역시 단백질 설계 분야에서 큰 성공을 거두고 있다. 2023년 David Baker 교수팀은 RFdiffusion이라는 Rotationally Equivariant diffusion 모델을 발표하여, 사용자가 지정한 구조/기능 요구사항에 부합하는 새로운 단백질 구조를 생성할 수 있음을 보여주었다^[11]. RFdiffusion은 단백질의 3D 좌표를 점진적으로 생성하도록 훈련된 확산 모델로, 조건으로 대칭 조건, 목표 단백질과의 결합 표적, 특정 기능 모티프 삽입 등을 줄 수 있다.

예를 들어 이 모델을 이용하면 표적 바이러스 단백질에 강하게 결합하는 단백질을 설계하도록 조건을 걸고 새로운 결합 단백질의 구조와 서열을 생성할 수 있다. 연구팀은 실제로 RFdiffusion으로 인플루엔자 바이러스의 헤마글루티닌(HA)에 결합하는 신규 단백질을 설계하고 실험 검증하였다. 그 결과 생성된 단백질 중 하나는 HA에 높은 친화도로 결합하였으며, 극저온 전자현미경(cryo-EM)으로 구조를 분석해 보니 설계한 모델과 거의 동일한 형태로 표적에 결합하고 있음이 확인되었다.

다시 말해, AI가 머릿속에서 그려낸 단백질 설계도가 실제 물리적 현실에서도 그대로 구현된 것이다. 이 밖에도 RFdiffusion은 단순한 단일 구조 생성뿐 아니라, 이종 단백질 사이의 결합계면 디자인, 금속 이온 결합 부위 갖는 새로운 단백질, 다중 서브유닛 대칭 올리고머 등 다양한 문제에 적용되어 실험적으로 기능하는 디자인들을 다수 만들어냈다. RFdiffusion의 공개로 전 세계 연구자들이 해당 모델을 활용할 수 있게 되었고, 벌써부터 이를 응용한 신약 개발용 단백질 저분자 결합부위 디자인, 촉매 효소의 활성 부위 설계 등의 후속 연구들이 쏟아지고 있다.

RFdiffusion (단백질 설계를 위한 확산 모델)

RFdiffusion은 워싱턴 대학 데이비드 베이커 연구팀이 개발한 단백질 설계용 생성 모델로서, 이미지 생성 AI에서 성공을 거둔 확산 알고리즘을 단백질 분야에 적용한 것이다. 이 접근법은 최첨단 단백질 구조 예측 신경망과 생성 모델을 결합함으로써, 사용자가 제시하는 조건에 부합하는 새로운 단백질 구조를 만들어낸다. 간단히 말해, 무작위 초기 구조에서 시작하여 점진적으로 현실적인 단백질 접힘 구조를 형성하도록 학습된 확률적 AI 모델이라고 볼 수 있다.

RFdiffusion의 핵심 엔진은 베이커 연구진이 2021년에 발표한 단백질 구조예측 모델 RoseTTAFold의 개량판이다. RoseTTAFold는 알파폴드와 유사하게 서열로부터 3차원 구조를 예측하는 딥러닝 모델로, 잔기들 간의 거리/각도 제약과 좌표 정보를 동시에 처리하는 3-트랙 아키텍처를 갖는다. 이 모델은 회전 및 병진 변환에 등변한 (rotationally equivariant) 잔기 좌표를 표현을 사용하여, 전역 좌표계에 구애받지 않고 분자의 3D 구조를 다룰 수 있다. RFdiffusion에서는 이러한 RoseTTAFold 모델을 기반으로 확산 모델의 디노이징 네트워크로 활용하도록 파인튜닝하였다. 즉, 원래는 아미노산 서열을 입력받아 최종 구조를 예측하던 네트워크를 이제는 잡음이 섞인 구조를 입력받아 원래의 올바른 구조로 복원해내도록 재훈련한 것이다. 이러한 최소한의 아키텍처 변경을 통해 RFdiffusion은 RoseTTAFold의 강력한 구조 예측 능력을 계승하면서도, 생성적 설계에 활용될 수 있게 되었다.

확산 모델은 데이터에 점진적으로 잡음(noise)을 더했다가 이를 제거하도록 학습된 확률적 생성 모델이다. RFdiffusion에서는 단백질 구조의 주사슬(backbone)을 각 잔기의 Ca 원자 좌표와 국소 방향으로 표현하고, 여기에 무작위 변형(좌표에는 가우시안 잡음, 회전에는 랜덤 회전 노이즈)을 추가하여 학습용 손상 데이터를 만든다. 모델은 이렇게 왜곡된 구조로부터 원래의 단백질 구조를 복원하도록 훈련되며, 이를 통해 임의의 노이즈 패턴을 단백질 형태로 되돌리는 역과정을 학습한다.

생성 시에는 처음에 완전히 무질서한 구조(잡음 상태)로 시작한 뒤, RFdiffusion 네트워크를 수백 회의 반복 단계로 적용하여 점차 실제 단백질과 유사한 구조로 수렴시킨다. 초반에는 출력 구조가 단백질답지 않다가도 반복이 진행될수록 구조적 제약이 누적되어 자연스럽게 설계 가능한 단백질 골격으로 정련된다. 이 과정에서 사용자는 원하는 설계 목표에 따라 다양한 조건을 부여할 수 있다. 예를 들어, 대칭적인 올리고머를 만들고자 하면 대칭 축 정보를, 특정 단백질에 결합하는 바인더를 설계하려면 표적 단백질 표면의 결합 핫스팟 지점을, 특정 기능성 모티프를 포함한 단백질을 만들고자 하면 그 모티프의 좌표를 각각 조건으로 입력하여 생성 과정을 유도할 수 있다.

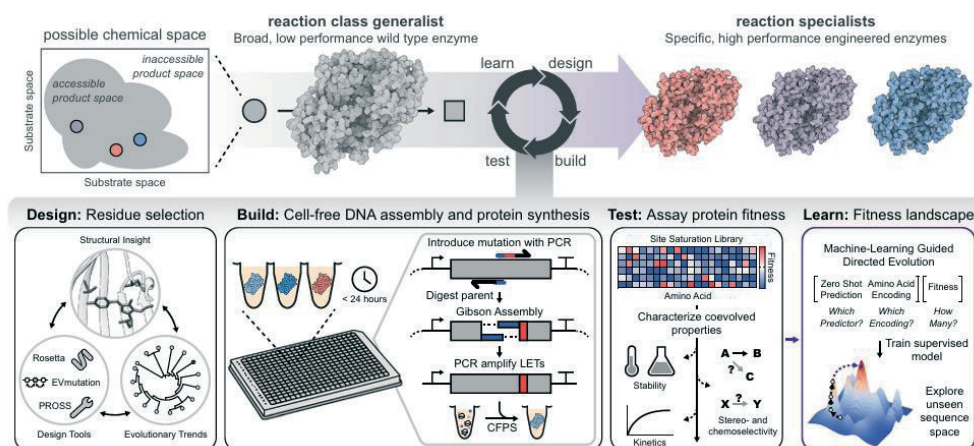
정리하면, AlphaFold 이후의 단백질 분야에서는 구조 예측의 완성과 함께 de novo 단백질 디자인의 시대가 본격 개막했다. AlphaFold로 축적된 방대한 구조 지식을 토대로, 언어모델(예: ProGen)과 확산모델(RFdiffusion) 등이 새로운 단백질을 창조해내고 있으며, 그 성능은 인간의 상상력을 뛰어넘는 수준으로 입증되고 있다. 또한 이러한 AI 설계 단백질들을 실제로 합성해 실험함으로써, AI 디자인, 실험 검증 및 피드백 학습의 순환 고리가 형성되어 단백질 엔지니어링의 설계-구현-테스트-학습(DBTL) 사이클이 가속화되고 있다. AlphaFold 이후 혁신의 키워드는 단연 생성이라 할 수 있으며, AI는 더 이상 자연의 단백질을 분석하는 도구가 아니라 새로운 생물학적 기능을 창조하는 창작자로 활약하고 있다.

4 효소 디자인 및 촉매 활성 예측을 위한 AI 기반 진화 공학

효소 공학은 합성생물학의 핵심 응용 분야 중 하나로, 기존 효소의 성능을 개량하거나 새로운 화학 반응을 수행하는 효소를 개발하는 일을 포함한다. 전통적으로 효소 엔지니어링은 지도된 진화 방법론에 크게 의존해왔다. 이는 자연에서 우수한 효소를 출발점으로 다수의 돌연변이 라이브러리를 만들고, 실험적으로 고활성 변종을 선별하여 다시 변이시키는 돌연변이-선별의 반복 과정을 거치는 방식이다. 하지만 이러한 기존 접근은 변이 도서관의 크기와 선별 속도에 한계가 있어 서열 공간의 극히 일부분만 탐색한다는 문제가 있다. 한 두 군데 자리만 바꾼 국소 탐색으로는, 때로 개별 돌연변이로 효과가 나타나지 않는 에피스타틱 상호작용을 놓치기 쉽고, 새로운 화학반응을 촉매하는 효소를 찾기에는 방대한 조합을 다 시도해볼 수 없다. 또한 전통적 진화 공학은 특정 기질 하나의 승자 효소만을 찾는 경향이 있어, 그 과정에서 쌓인 데이터를 인근 반응이나 유사 기질에 활용하기 어려운 단점도 있다. AI 기반 진화 공학은 이러한 한계를 극복하고자, 머신러닝을 이용해 효소의 돌연변이-활성 관계를 학습하고 이를 토대로 유망 변종을 예측함으로써 탐색 효율을 높였다. 기본 아이디어는 비교적 적은 수의 변이 샘플에 대해 실험으로 활성을 측정한 데이터를 학습 세트로 삼아, 서열로 활성을 예측하는 모델(일명 fitness predictor)을 만들고, 이 모델이 점수가 높다고 예측하는 변이를 선택하여 시험해보는 것이다. 이러한 ML 보조 탐색은 사람이 일일이 라이브러리를 디자인하기보다 훨씬 광범위한 서열 공간을 지능적으로 탐색할 수 있으며, 연구자가 관심 갖는 다중 목표도 함께 고려하기 쉽다. 나아가 최근에는 사전학습된 단백질 언어모델을 활용하여, 실험 데이터가 거의 없는 경우에도 진화적 문맥상 중요한 변이를 예측하는 Zero-shot 예측까지 시도되고 있다. 즉, 방대한 단백질 서열의 패턴을 학습한 언어모델이 특정 효소의 서열만 보고도 어떤 변이가 기능에 유리할지 점칠 수 있다는 것이다. 이러한 기법은 실험 데이터 축적 이전에 탐색 범위를 줄여주는 보조 수단으로 각광받고 있다.

ML-guided 진화 공학의 한 성공 사례로, 2025년 보고된 Landwehr 등의 연구를 들 수 있다. 이 연구에서는 세포 추출물 기반 고속 합성/분석 플랫폼을 구축하고, 1,217개의 효소 변이체를 한꺼번에 발현시켜 10,953개에 달하는 다양한 기질-효소 조합에 대한 활성 데이터를 얻었다. 여기서 타깃 효소는 아미드 결합 형성 반응을 촉매하는 아미드 합성효소였는데, 다양한 의약품 전구체 분자 9종을 기질로 투입해 어떤 변이가 어떤 기질에 활성을 보이는지 측정한 것이다. 이렇게 얻은 대규모 서열-기능 데이터셋을 기반으로 능형 회귀 등의 ML 모델을 훈련하여, 각 기질에 대해 높은 활성을 가질 것으로 예측되는 새로운 효소 서열들을 제시했다.

An ML-guided, cell-free enzyme engineering platform

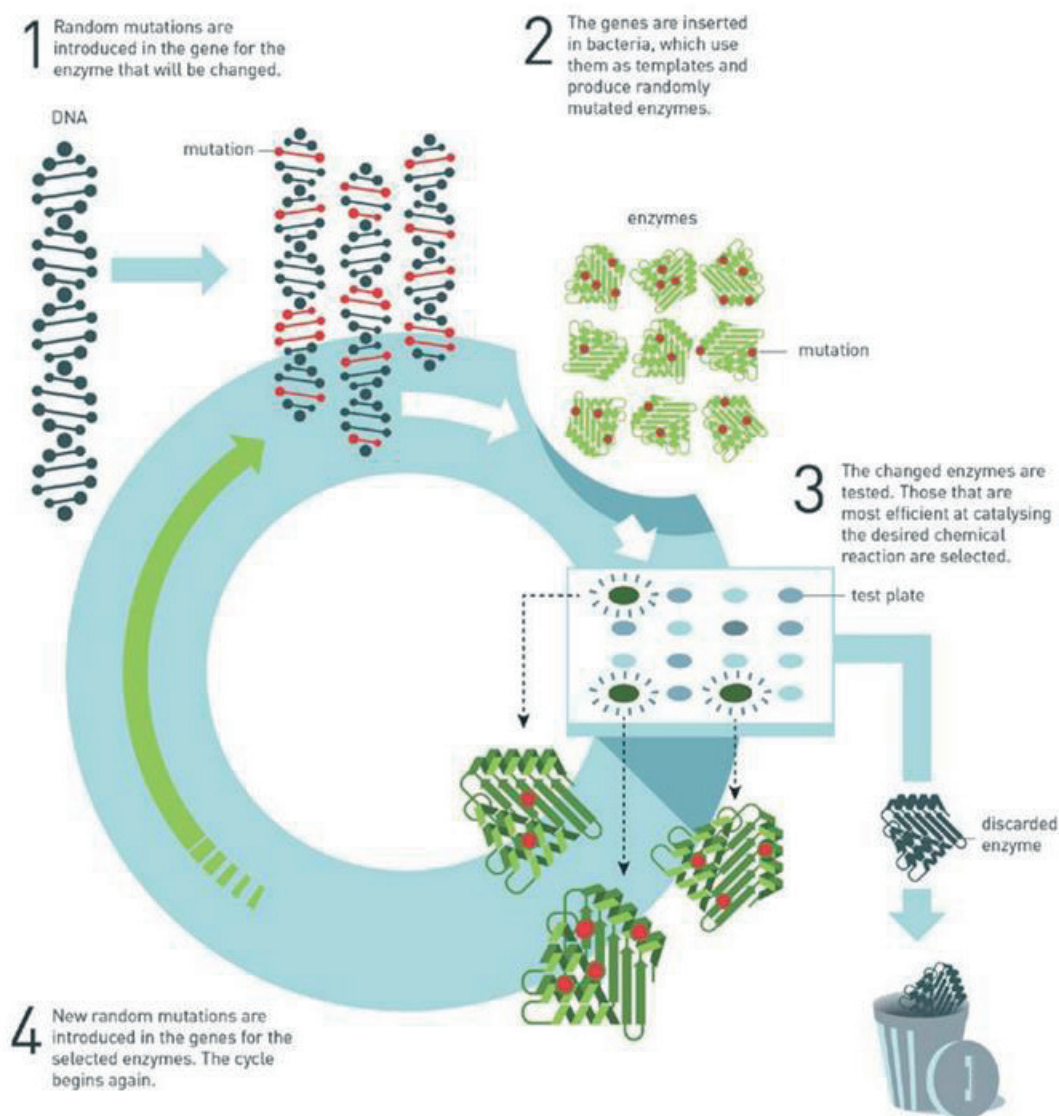


그 결과 ML 모델이 추천한 효소 변이들은 원래 효소 대비 1.6배에서 최대 42배까지 향상된 활성으로 목표 아마이드 합성 반응을 수행하였다. 이러한 연구들은 Bayesian optimization, 강화학습 기반 탐색 등 다양한 AI 기법을 결합해, 실험 효율을 극대화하고 있다. 특히 cell-free 자동화 시스템과 결합된 고속 실험 플랫폼은 AI 예측-실험 검증-모델 재학습의 피드백 루프를 형성하여 효소 개발 주기를 획기적으로 단축시킨다. 다시 말해, AI가 효소의 활성 지도를 학습한 덕분에, 실험적으로 탐색하지 않았던 조합의 돌연변이라도 바로 고효율 변이를 찾아낼 수 있었다^[12]. 이 플랫폼은 또한 한 번에 여러 다른 화학 반응에 대한 최적 효소를 병렬적으로 찾을 수 있음을 보여주었는데, 저자들은 이를 멀티태스킹 효소 엔지니어링의 개념으로 제시하였다. 해당 연구의 ML 가이드 프레임워크는 일반적인 CPU 상에서도 작동할 만큼 경량이어서, 특별한 장비 없이도 연속적인 설계-학습 사이클을 실행할 수 있었으며, 연구진은 이러한 AI+고속실험 통합 접근법이 특정 효소 기능을 신속히 강화하거나 완전히 새로운 기질로 범위를 넓히는 데 크게 기여할 것으로 전망했다. 실제로 논문 제목에서도 accelerated enzyme engineering이라는 표현을 사용하여, AI가 접목된 효소 엔지니어링이 기존 방법 대비 얼마나 가속화될 수 있는지 강조하고 있다.

ML 보조 접근법은 전례 없는 신규 기능 부여에도 활용되고 있다. 예컨대 과거 자연계에 존재하지 않는 화학결합을 효소로 생성하는 연구에서, 탄소-규소 결합 형성 효소를 진화시킨 사례가 유명하다. 2020년대 초반 한 연구팀은 기존의 directed evolution과 머신러닝 순차 적용을 통해, 세린돌핵효소를 돌연변이시켜 선택적인 C-Si 결합 형성 반응을 촉매하는 변이효소를 얻었다. 이때 Bayesian Optimization 등 ML 기법이 진화 과정에도

입되어 상대적으로 적은 실험으로 최적 변이를 찾는 데 도움을 주었다. 그 결과 연구팀은 입체선택적 탄소-규소 결합을 생성하는 효소를 설계할 수 있었고, 이는 머신러닝이 아니었다면 방대한 돌연변이 라이브러리를 필요로 했을 도전적인 목표였다. 이처럼 시는 탐색 공간을 좁혀주고 이상적인 지점을 예측함으로써, 인공 효소로 새 분자결합/반응을 창조하는 과정도 가속화하고 있다.

The work flow for the directed evolution of enzymes



나아가 Active Learning과 같은 기법을 통해 자동화된 실험 설계도 가능해지고 있다. 이는 실험을 통해 얻은 정보로 ML 모델을 업데이트하고, 모델이 아직 불확실성이 큰 영역의 실험 조건을 제안하여 추가 실험을 수행, 다시 모델을 개선하는 순환을 말한다. 효소 엔지니어링에 이 방식을 적용한 2024년 연구에서는, 효소 활성 뿐 아니라 서열 다양성까지 고려한 라이브러리 설계를 ML로 수행하여, 소수의 라운드만으로도 폭넓은 변이 탐색과 최적화에 성공했다. 이러한 능동학습 기반 라이브러리 최적화는 사람의 직관에 의존했던 돌연변이 라이브러리 디자인을 데이터에 기반한 과학적 방식으로 전환함으로써, 더 높은 적중률로 유용 변이를 찾아주는 역할을 했다.

AI 기반 효소 디자인은 두 갈래로 요약할 수 있다. 첫째, 앞서 말한 생성모델을 통한 새로운 효소의 창조이다. 이는 특정 반응이나 기능을 수행하는 완전히 새로운 서열을 만들어내는 것으로, ProGen이나 RFdiffusion 같은 생성형 AI가 이에 해당한다. 둘째는 본 절에서 논한 ML-guided 진화 공학으로, 기존 효소를 빠르게 개량하거나 다중 목적 최적화를 달성하는 접근이다. 궁극적으로는 이 두 흐름이 합쳐져, 생성 AI로 초기 효소 설계 → 실험 및 ML로 미세 최적화라는 하이브리드 전략이 이상적인 효소 엔지니어링 파이프라인으로 확립될 것으로 예상된다. 실제로 AlphaFold로 구조를 예측하고, 생성모델로 주요 골격을 만든 뒤, 세부 돌연변이는 ML 예측으로 손질하는 식의 워크플로우가 가능하며, 이미 일부 연구자들은 이러한 통합적 접근을 시도하고 있다.

5

맺음말

AI와 합성생물학의 융합은 기하급수적 혁신을 이끌며, 설계 자동화와 탐색 지능화라는 양대 축에서 생명공학의 지평을 넓히고 있다. 머리말에서 언급한 것처럼, AI는 이제 생물학에서 새로운 아이디어의 원천이자 창의적 설계자의 역할을 수행하고 있다. 유전자 서열 디자인 영역에서는 Generative AI가 인간이 상상하기 힘든 서열 조합을 제시함으로써 세포 공장을 최적화하고 유전자 치료 효율을 높이고 있다. 단백질 디자인 영역에서는 AlphaFold로 대변되는 예측능력을 넘어, 완전히 새로운 단백질 세계를 창조하고 그 기능을 실험으로 입증하는 단계에 이르렀다. 효소 공학에서는 AI가 진화 공정을 가속하고 있으며, 이에 따라 인류는 원하는 화학반응을 촉매하는 맞춤형 효소를 빠르게 얻는 시대를 맞이하고 있다.

물론 해결해야 할 도전도 남아있다. AI-생물학 융합이 본격화됨에 따라, 연구 표준화와

데이터 공유, 모델의 해석 가능성, 생명윤리 및 연구거버넌스 강화가 필수적으로 요구된다. 또한 이러한 기술들을 다룰 수 있는 차세대 생명정보학 인력 양성과 다학제적 협력체계 구축이 병행되어야 한다. 생성형 AI가 내놓은 설계안이 실제 생물 시스템에서 항상 안정적으로 작동하는가에 대한 검증, AI 모델의 해석 가능성 확보와 신뢰성 평가, 그리고 AI 기술의 이중용도 위험 관리 등이 그 예이다. 특히 생물학적 지식이 부족한 상태에서 AI가 제시한 위험한 병원체 서열 등이 악용되지 않도록 하는 안전장치와 정책적 대비도 중요하다. 그럼에도 불구하고 분명한 사실은, 합성생물학 분야에서 AI의 역할은 갈수록 커지고 필수화될 것이라는 점이다. 전문가로서 우리는 최신 AI 기술 동향을 적극적으로 습득하고 활용하여, 보다 효율적이고 혁신적인 생물 설계를 이루어내야 할 것이다. 동시에 이러한 첨단 기술을 윤리적이고 안전하게 활용하는 가이드라인 수립에도 힘써야 할 것이다. AI와 합성생물학의 조화는 현대 생명공학의 새로운 물결이며, 그 속도와 임팩트는 우리가 익숙했던 연구 방식을 근본적으로 재편하고 있다. 앞으로 다가올 10년은, AI가 창조한 생명 시스템들이 연구실과 산업현장에서 현실이 되는 시대가 될 것이며, 합성생물학자들은 이 변화의 선두에서 그 잠재력을 최대한 실현해야 할 것이다.

[참고문헌]

- [1] Lal, A. et al. Designing realistic regulatory DNA with autoregressive language models (regLM). *Genome Research*, 34, 1–10 (2024).
- [2] Gu, Y. et al. De novo promoter design method based on deep generative and dynamic evolution algorithm (PromoDGDE). *Nucleic Acids Research*, 51, 9048–9066 (2023).
- [3] Profluent. Scaling unlocks broader generation and deeper functional design in protein language models. Preprint on bioRxiv (2025). DOI: 10.1101/2025.04.15.649055v1
- [4] Groff-Vindman, C. S. et al. The convergence of AI and synthetic biology: the looming deluge. *npj Biomedical Innovations*, 2, 20 (2025).
- [5] Pinello, L. et al. DNA-Diffusion: Leveraging generative models for controlling

chromatin accessibility and gene expression via synthetic regulatory elements. (Preprint, 2024).

- [6] Zhang, H. et al. Deep generative models design mRNA sequences with enhanced translational capacity and stability. *Science*, 381, eadr8470 (2025).
- [7] Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596, 583-589 (2021).
- [8] Barazandeh, S. et al. UTRGAN: learning to generate 5' UTR sequences for optimized translation efficiency and gene expression. *Bioinformatics Advances*, 5, vbaf134 (2025).
- [9] Zhang, H. et al. Algorithm for optimized mRNA design improves stability and immunogenicity. *Nature*, 621, 396–403 (2023).
- [10] Madani, A. et al. Large language models generate functional protein sequences across diverse protein families. *Nature Biotechnology*, 41, 142–151 (2023).
- [11] Watson, J. L. et al. De novo design of protein structure and function with RFdiffusion. *Nature*, 620, 1089–1100 (2023).
- [12] Landwehr, G. M. et al. Accelerated enzyme engineering by machine-learning guided cell-free expression. *Nature Communications*, 16, 865 (2025).
- [13] Yang, K. K. et al. Machine learning-assisted directed protein evolution with combinatorial libraries. *PNAS*, 117(45), 27945–27953 (2020).

합성생물학 분야의 바이오 파운데이션 모델 기술 활용 동향

1 머리말

합성생물학 분야에서 새로운 기능성을 가지는 DNA, RNA, 단백질 설계를 위하여 생물 시스템을 공학적으로 설계 및 제작하는 과정을 거치며, 이를 위해 표준화된 유전자 부품을 조합하고, 전사/번역 조절에 관여하는 RNA 스위치 설계 또는 인공단백질들을 디자인하고 합성하게 된다. 이러한 과정들은 Design-Build-Test-Learn (DBTL) 주기를 통한 공학적인 설계 과정을 통해서 최적화 과정을 거치게 되면서, 기존 생물학 연구의 한계인 느린 개발 속도와 작은 규모에 한정되는 등의 문제를 해결하게 되고, 예측 가능한 시스템을 제작할 수 있게 되며, 표준화된 생명 시스템을 구축할 수 있게 된다.

최근에는 DBTL의 공학적 설계에 있어서 “인공지능 파운데이션 모델”을 도입하게 됨에 따라서, 각 주기에 다양한 기능성을 효과적으로 탐색하게 하고, 자동화를 가속화하며, 고용량의 데이터를 처리하게 할 수 있게 하며, 새로운 지식 추론들을 가능할 수 있게 함으로써 그 역할이 크게 주목받고 있다. 특히 DBTL주기에서 핵심적인 과정인 trial-and-error 단계를 데이터 기반의 지능적인 설계 방식으로 전환을 가속화시킴에 따라서, 예측 불가능하고, 높은 복잡도를 가지는 생물학적 시스템을 효과적으로 설계할 수 있는 핵심 동력이 되어줄 것으로 예상된다.

파운데이션 모델은 2017년 트랜스포머 (transformer) 아키텍처가 도입됨에 따라서 Large Language Model (LLM) 개발이 폭발적으로 성장하면서 등장하기 시작하기 하였으며, 특히, 대규모의 레이블링 되지 않은 데이터를 학습시키고 광범위한 다운스트림 작업에 활용할 수 있는 범용적 AI 모델 개발이 가능함이 확인되어짐에 따라서, BERT, GPT와 같은 초기 파운데이션 모델의 개발들이 이어졌다. 이러한 혁신적인 AI모델의 개발이 텍스트 데이터 분야에서 뿐만 아니라, 대규모 바이오메디컬 데이터들이 축적되어지고 활용되어지고 있는 바이오 분야에서도 적용되기 시작하게 되었는데, 특히 생명현상을 예측하고, 새로운 원리나 지식을 추론할 수 있는 파운데이션 모델의 장점을 바탕으로 다양한 분야, 특히 합성생물학 분야에서의 효소 기능 최적화 또는 새로운 단백질 구조 설계 등의 분야에 있어서 큰 두각을 나타낼 수 있을 것으로 예상하고 있다.

2 바이오 파운데이션 모델 개발 근황

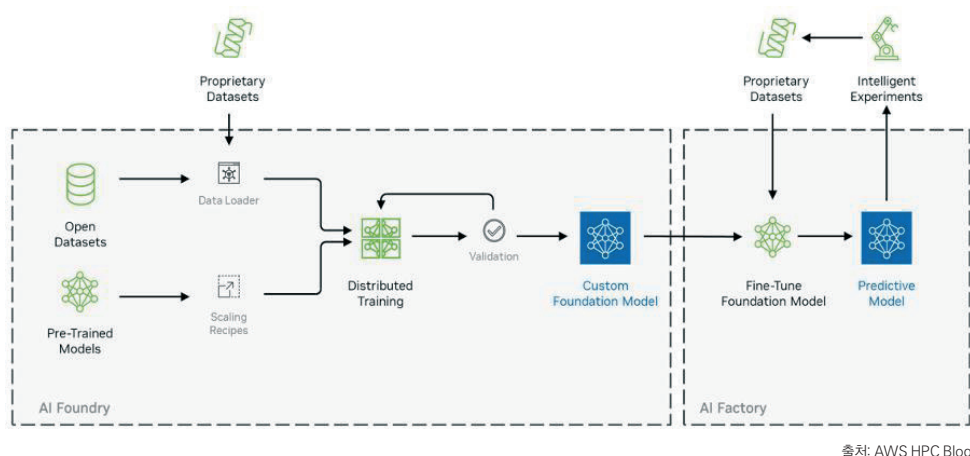
현재 바이오 파운데이션 모델이 가장 활발하게 적용되고 있는 분야는 생체 분자 설계 및 신약개발 분야, 그리고 정밀 의학 분야 등에 해당하며, 대규모의 생물학적 데이터들(DNA, RNA, 단백질, 구조, 생체 분자 데이터 등)을 학습하여 새로운 생명현상 조성 원리를 예측하고 새로운 기능성의 분자를 설계하는 데에 활용된다. 주로 공공 생물정보학 데이터베이스를 공유되고 있는 기초 연구 데이터(유전체 서열, 단백질 구조 및 기능 데이터 또는 화학/약물 데이터 등)를 바탕으로 다양한 바이오 파운데이션 모델들이 개발되어지고 있으며, 그 외에도 임상연구기관에서 환자 진료 및 임상 연구를 통해서 생성하는 전자 의무기록(EMR) 데이터, 의료영상 데이터, 대규모 코호트 데이터들을 활용하거나, 혹은 실험실 또는 연구소에 생산하고 있는 다양한 단일세포 유전체 데이터들을 활용하여서, 이질적인 분자 수준 데이터 또는 세포/임상 수준 데이터들을 활용하여서 생명현상 원리를 예측하고, 새로운 지식을 추론하는 모델을 개발하게 된다.

바이오 파운데이션 모델은 특히 생물학적 데이터가 가지고 있는 이질성(heterogeneity)을 다루기 위해서는 멀티모달(multi-modal) 모델 개발이 필수적이다. 예를 들어서 가장 기본적인 유전정보가 코딩되는 DNA/RNA/단백질 서열 데이터들과 특정 기능을 나타내게 되는 단백질 및 약물 구조라는 3차원 구조 정보, 그리고 해당 유전자들이 발현하여 나타내는 기능 또는 표현형을 설명할 수 있는 유전자 발현량, 세포 표현형, 임상 데이터들을 동시에 통합하여 학습할 수 있는 멀티모달 모델이 필수적이며, 이러한 모델의 설계를 바탕으로 단백질 구조를 예측하고 설계하거나, 특정 질병을 진단하고 치료 타겟을 발굴하는 다양한 테스크들을 수행하게 된다.

예를 들어서, 2022년도에 개발된 Prot-GPT2 모델은 GPT 아키텍처를 바탕으로 단백질 서열에 대한 거대 언어 모델(Large Language Model, LLM)을 개발하여서, 새로운 기능을 갖는 단백질 서열을 설계하고 합성할 수 있게 돕고 있으며^[a], MIT CSAIL 연구소와 빅테크 기업인 Recursion에서 개발한 생체분자 파운데이션 모델인 Boltz-2 모델의 경우^[a], 단백질 구조와 이와 결합하는 화합물의 결합력을 동시에 예측할 수 있는 코폴딩(co-folding) 모델로, 오픈소스로 공개되어 대규모의 virtual drug screening을 가능하게 만들어, 기존의 단백질 구조 예측 인공지능 모델인 알파폴드 모델을 뛰어넘었다는 평가를 받고 있다. 이러한 흐름들은 다양한 빅테크 기업들에서도 관련 플랫폼 개발에 박차를 가하고 있으며, 예를 들어서 NVIDIA社에서는 DNABERT, ESM-2, OpenFold 등의 DNA, 단백질, 화합물 구조, 기능 예측 바이오 파운데이션 모델 컬렉션을 BioNeMo 프레임워크를 제공하

고, 특히 자사의 NVIDIA GPU 클러스터 (DGX)에 탑재가능한 형태로 제공함으로써 암젠 (Amgen) 등의 대형 제약회사의 AI 신약개발을 가속화하도록 돕고 있다[그림 1].

그림 1> NVIDIA BioNeMo 워크플로우



본 동향 보고서에서는 최근에 개발된 다양한 형태의 바이오 파운데이션 모델에 대한 동향을 추가적으로 기술해보고, 예를 들어서 유전체, 전사체 데이터에 기반 한 파운데이션 모델, 단백질 구조 파운데이션 모델 및 다양한 형태의 멀티 모달 파운데이션 모델들을 살펴보고, 해당 모델들이 합성생물학 분야에서의 역할들을 조사해보고자 한다.

2.1 유전체 및 전사체 파운데이션 모델 개발 근황

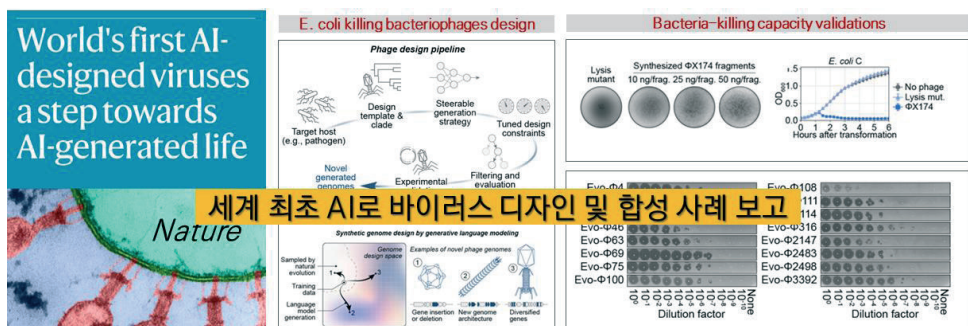
유전체 및 전사체 데이터는 유전정보를 담고 있는 가장 기본적인 생명정보 데이터 중 하나로서, DNA 염기서열과 유전자 발현 데이터를 바탕으로 발현되는 유전자의 기능을 예측하고 추론하는데 활용가능하기 때문에, 다양한 형태의 생명원리 추론을 가능케 하는 다양한 유전체/전사체 바이오 파운데이션 모델들이 초기부터 개발이 이어져 오고 있다. 가장 대표적인 예시로서 DNA 염기서열을 바탕으로 트랜스포머 아키텍처를 통해서 학습된 유전체 바이오 파운데이션 모델을 들 수 있으며, Nucleotide Transformer, DNABERT, HyenaDNA등의 DNA언어 모델을 바탕으로 1차적으로는 특정 유전자 서열 지역의 기능 예측, 예를 들어서 프로모터 예측, DNA 메틸화, 예측, 유전변이에 따른 기능적 영향을 예측하는 등의 추론들을 가능케 하며, 또한 특정 기능을 갖거나 높은 전사율을 갖는 DNA, RNA서열을 설계하고 생성하는 데에도 활용될 수 있다. 이러한 DNA, RNA 설계 능력을 바탕으로 합성생물학 분야에서의 새로운 분자 설계를 가능케 하며, 백신 개발

이나 유전자 치료제 설계 역시 가능케 할 것으로 예상된다. 예를 들어서 DNABERT의 경우 DNA 서열을 k-mer조각으로 변환한 후 학습한 모델로써, 미세 조정(fine-tuning)을 통해서 transcriptional start site (TSS) 주변 promoter 지역이나 transcription factor binding site (TFBS) 지역 예측을 가능케 하는 모델로 알려져 있고, Enformer 모델의 경우, DNA 서열 정보만으로도 해당 유전자의 발현을 예측하는 모델로써, 유전체 기능 연구 등에 활용할 수 있는 모델로써 개발되어졌다.

빅테크 기업인 구글 딥마인드社에서도 유전체 바이오 파운데이션 모델에 뛰어들면서, 유전체 기능 예측이 가능한 AlphaGenome 모델을 개발하여, 다양한 유전체 변이의 기능 연구를 가능케 하고 있으며, 앞으로 관련 분야의 개발 경쟁이 가속화될 것으로 예상된다.

최근에는 스탠포드대학교와 Arc Institute에서 공동으로 개발한 Evo2모델의 경우[3], OpenGenome2 데이터셋으로 학습된 400억개(40B)의 파라미터로 구성된, 현재까지 가장 진보적인 형태의 유전체 바이오파운데이션 모델이다. Evo2는 새롭게 개발된 StripedHyena2 아키텍처를 바탕으로 long range interaction을 역시 잘 포함시킬 수 있는 모델로 개발되어서 유전변이의 기능적 영향 예측, 생체 분자 서열 생성 및 설계뿐만 아니라 제로샷 학습 (zero-shot learning, ZSL)이 가능한 모델이다. 최근에는 Evo2모델을 바탕으로 미세조정 (fine-tuning)을 통해서 세계 최초로 특정 박테리아를 타겟할 수 있는 새로운 박테리오파지 바이러스 유전체를 디자인하여 합성하는 사례가 발표된바 있어, 앞으로도 합성생물학 분야의 유전체 설계 분야에 있어서 다양한 파격적인 혁신 연구 사례들이 나올 것으로 기대 된다[4, 5].

그림 2> Evo2모델을 바탕으로 세계 최초로 새로운 바이러스 유전체 설계 및 합성 성공



출처: Katie Kavanagh, Nature News (2025) & Samuel H King et al., BioRxiv (2025)

전사체 바이오 파운데이션 모델의 경우, 특히 최근에 대규모로 생성되고 있는 단일세포 전사체 데이터를 바탕으로 폭발적으로 성장하고 있는 추세이며, 단일세포 전사체 데이터 생성 시에 확인된 다양한 세포 유형, 질병상태, 조직 미세 환경 정보를 학습하여, 유전자 발현 정보로부터 유전자 조절 관계, 질병상태, 약물 반응성을 예측하는 모델로써 개발되어져, 거대 세포 모델 (Large Cellular Models, LCM)이라고 불릴 정도로, 다양한 세포 표현형을 예측할 수 있는 파운데이션 모델로 주목받고 있다. 특히 Chan-Zukerberg Initiative (CZI)에서 지원하고 있는 Human Cell Atlas 등의 대규모 이니셔티브를 바탕으로 수억개 이상의 단일세포 시퀀싱 데이터셋이 이미 공개적으로 공유되고 있기 때문에, 높은 데이터 접근성과 다양한 분야의 응용 가능성, 신약 개발과의 연계성 등으로 인해서, 다른 바이오 파운데이션 모델 분야 보다 더욱 폭발적인 성장이 예상되고 있다.

전사체 파운데이션 모델의 대표적인 예시로는, 토론토 대학교 Bo Wang 교수팀에서 개발한 scGPT 모델을 들 수 있으며, 인체의 다양한 조직에서 유래한 3천만개 이상의 단일세포 전사체 데이터를 학습하여, 세포 유형 발굴 (cell annotation), 유전자 타겟팅에 따른 유전자 발현 변화 예측 (perturbation response), 배치 교정, 멀티오믹스 통합 뿐만 아니라 유전자 조절 네트워크 추론 등의 다양한 테스크 수행이 가능하다. scGPT외에도 geneformer, scBERT, scFoundation 등의 파운데이션 모델들이 개발되었고, 단일세포 멀티오믹스 데이터 분석 및 새로운 세포 유형 확인 등의 다양한 테스크에 폭넓게 활용되어지고 있다. 최근에는 perturb-seq과 같은 다양한 유전자를 타겟팅한 유전자 제어에 따른 단일세포 데이터들이 다양하게 생성되고 있어서, 더욱 정교한 파운데이션 모델들 개발되어질 것으로 예상된다.

전사체 바이오파운데이션 모델은 기본적으로 유전자 발현을 바탕으로 유전자 조절 관계를 예측하고 설계할 수 있는 모델이기 때문에, 합성생물학 분야에서 핵심적인 유전자 회로 설계 및 최적화 분야 (Design)에 직접적으로 활용될 수 있으며, 합성된 회로 (Build)에 따른 표현형-전사체 관계를 디코딩하고 해석해줌으로써 새로운 유전자 회로 설계 결과를 해석하고 (Test) 이를 바탕으로 새롭게 디자인 (Design)하는 데에 핵심적인 역할을 할 수 있다. 추가적으로 맞춤형 RNA 설계 역시 가능하기 때문에, 인공 RNA서열 디자인을 바탕으로 새로운 합성 미생물 디자인 역시 가능케 할 것으로 예상된다.

2.2 단백질 구조 파운데이션 모델 개발 현황

단백질 구조 예측 인공지능 모델은 최근 가장 급속적으로 발전되고 있는 분야 중 하나이며, 알파폴드 2 (AlphaFold2) 모델이 2020년 CASP14 구조 예측 대회 최고치를 경신했

에 따라서, 기존의 인실리코 모델들의 성능을 훨씬 뛰어넘는 심층학습 모델들의 개발이 가속화 되었고, 그로 인해 더 이상 실험 관측 없이도 단백질 구조를 정교하게 예측하게 되면서, 패러다임의 근본적인 대전환이 시작되었음을 알리게 되었다. 알파폴드2 모델은 아미노산 서열로부터 단백질 구조 3차원 좌표를 예측하는 엔드-투-엔드 예측 방식으로, 어텐션 (attention) 메커니즘을 바탕으로 다중 서열 정렬 (multiple sequence alignment, MSA) 정보 및 아미노산 잔기 쌍 (pair)간의 공간적 거리에 기반 한 공진화적인 정보를 학습하고 구조정보를 예측하는 방식으로 진행되게 된다. MSA 정보는 정확한 구조 정보 예측을 가능하게 하였으나, MSA를 위해서 높은 양의 계산 비용이 수반되어, 학습 및 추론에 많은 시간이 소요되게 되며, 동등 서열이 존재하지 않아 MSA 결과를 얻을 수 없는 orphan protein의 구조 예측이 한계성을 노출하고 있다. 특히 합성생물학 분야에서 새롭게 디자인하여 합성하는 인공단백질 또는 미니 바인더 단백질 등은 기존에 진화적으로 유사한 서열을 가지고 있지 않기 때문에, 해당 어플리케이션에 적용하기 어려운 한계점을 가지고 있다.

최근에는 대규모 언어 모델 (Large Language Model, LLM)에서 사용하는 트랜스포머 아키텍처를 바탕으로 차세대 단백질 언어 모델 (Protein Language Model)이 개발되어 구조 예측에 활용되고 있다. 비록 MSA에 기반 한 알파폴드 모델 또는 로제타폴드 모델에비해서는 성능이 다소 떨어지지만, MSA 과정 없이도 구조 예측이 가능하며, 특히 단백질 설계 과정에 효과적인 성능을 보이는 것으로 알려져 있다. 단백질 언어 모델의 대표적인 예시로서, 메타사에서 개발한 ESMFold를 들 수 있다. ESMFold는 마스크 언어 모델링 (masked language modeling, MLM)을 바탕으로 한 사전 학습을 통해서 암묵적인 공진화 패턴을 학습하였고, 이를 바탕으로 단백질 구조를 추론함으로써, MSA과정을 우회하면서도 효율성과 높은 적용범위를 극대화 시킬 수 있는 모델로서 자리 잡게 되었다. 특히 알파폴드와 로제타폴드의 경우는 동등서열이 존재하지 않는 인공단백질의 구조 예측에는 한계를 보이고 있지만, ESMFold는 그와 달리 인공단백질 또는 동등서열이 극히 드문 단백질에도 효과적인 성능을 보이는 것으로 확인되고 있다. 뿐만 아니라 ESM3, ESMDiff 모델을 바탕으로 원하는 구조를 생성하는 서열을 새롭게 디자인하는 것이 가능해지고 있으며, ProtGPT2, ProteinMPNN 등의 단백질 언어 모델들을 활용해서도 원하는 스캐폴드에 해당하는 단백질 서열 디자인 역시 가능해지고 있다.

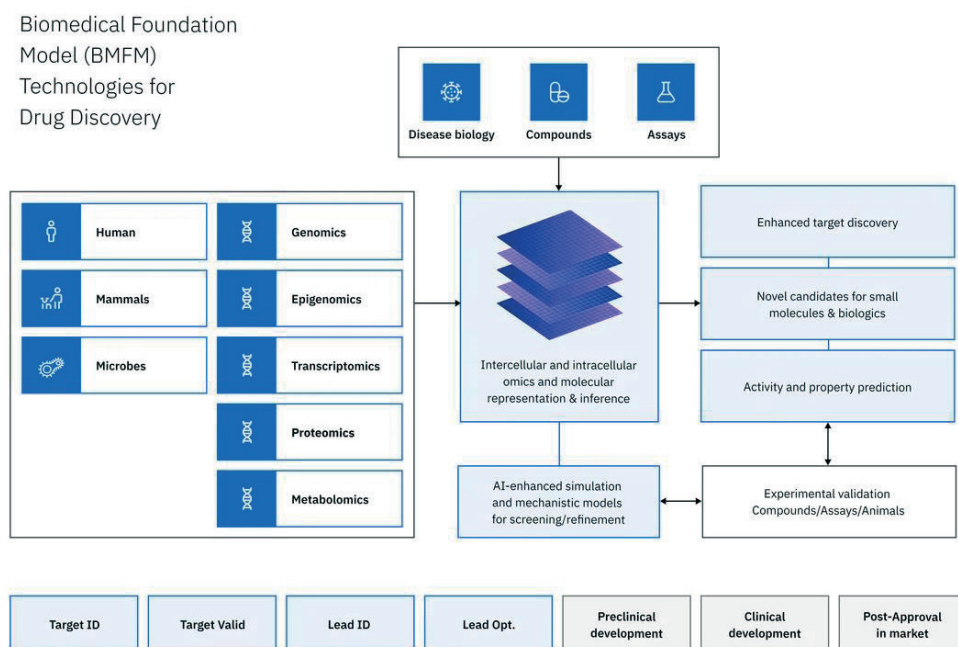
앞으로 단백질 구조 예측 모델의 성능 개선과 함께 합성생물학 분야에 있어서 특정 기능성을 가지는 다양한 인공 단백질, 효소 엔지니어링, 항체 디자인 등의 다양한 응용 분야들이 계속적으로 발전될 것으로 예상된다. 아쉽게도 단백질 구조 예측 인공지능 모델들이

아직까지는 동역학 모델링, 단백질 복합체 또는 상호작용 지도 예측에 있어서는 한계점을 보이고 있으며, 대규모의 단백질의 구조 예측에 소요되는 높은 계산량을 감당하기 위하여 계산 효율성을 극대화시켜야 하는 과제들이 남아있지만, 이미 몇 차례의 성공 사례를 바탕으로 더욱 정교하고 효율적인 모델 개발들과 그와 관련된 지속적인 연구개발 투자가 이뤄질 것으로 예상된다. 이러한 급진적인 발전 속도와 함께, 해당 모델들이 신약개발 및 합성생물학 분야의 발전을 함께 이뤄나갈 것으로 예상된다.

2.3 멀티모달 바이오파운데이션 모델 개발 근황

최근에는 다양한 바이오메디컬 데이터들의 축적과 함께, 다양한 생물학적 컨텍스트를 추론할 수 있는 멀티모달 바이오파운데이션 모델들이 개발되고 있다. 특히 이러한 멀티모달 바이오파운데이션 모델은 특정 질병의 진단 또는 치료 가능 표적을 효과적으로 식별하고, 이를 타겟팅할 수 있는 새로운 생물학적 제제 또는 화합물 디자인을 가능하게 한다. 이러한 예시로써 IBM社에서 개발한 Biomedical Foundation Model (BMFM)을 들 수 있으며, 해당 모델을 바탕으로 치료 타겟 발굴 및 검증, 선도물질 (lead compound) 발굴 등, 신약 개발 단계에서 필요한 다양한 테스트들을 수행할 수 있게 된다.

그림 3> IBM에서 개발한 Biomedical Foundation Model 워크플로우



출처: IBM Research

그림 4> Therapeutics Data Commons 플랫폼의 AI-Ready 데이터셋 및 학습 테스트



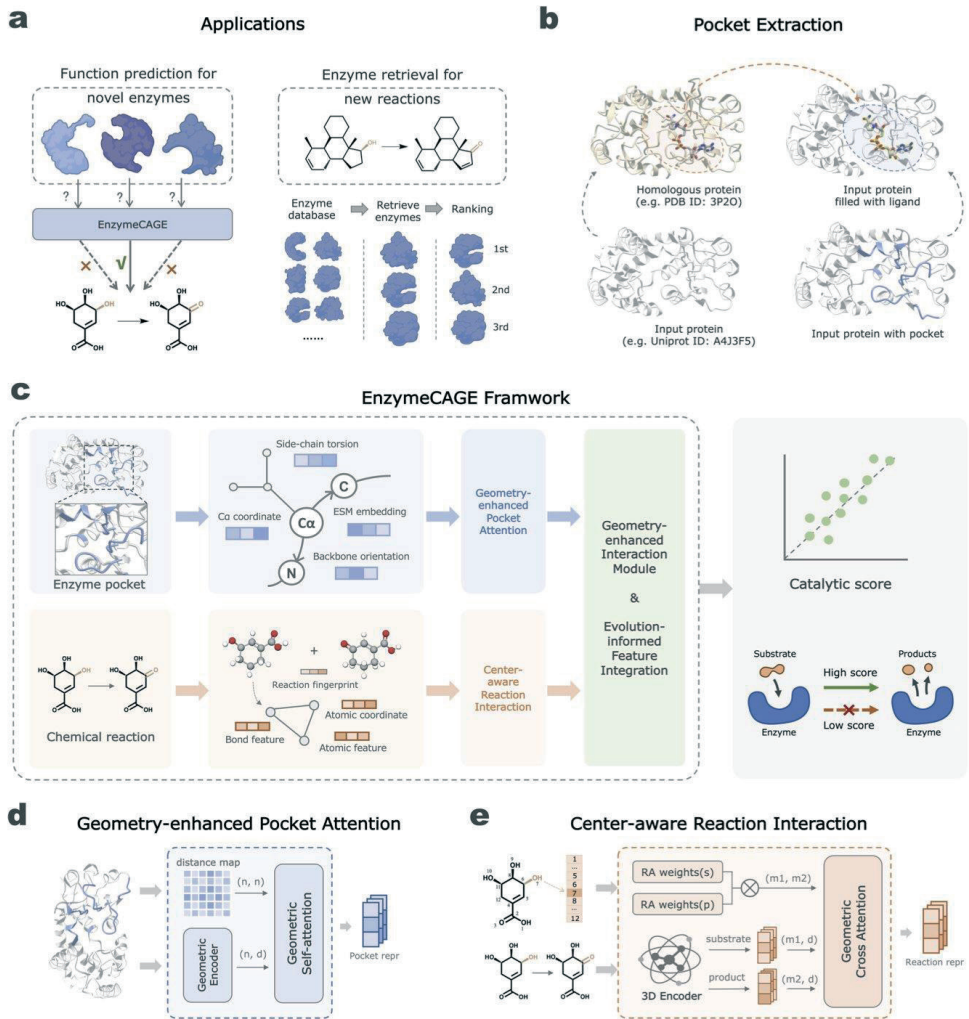
3 바이오 파운데이션 모델의 합성생물학 분야 응용 사례

합성생물학은 자연계에 존재하지 않는 새로운 생물 시스템 설계하고 제작하여, 기존의 생물체 도입으로 해결되지 않았던 다양한 인류 난제들을 해결할 수 있는 초석이 되고 있다. 예를 들어서, 의학, 에너지, 환경, 농업분야의 다양한 부분에서 합성생물학 기술의 응용사례들을 찾아볼 수 있으며, mRNA백신, 맞춤형 치료제, 유전자 가위 기반 유전자, 세포치료제, 바이오연료 생산, 바이오플라스틱 분해 미생물, 바이오화학 소재 생산, 농작물 품질 향상, 배양육, 대체 단백질 생산, 바이오 센서 개발 등의 다양한 분야에서 응용되고 있다.

3.1 기능 최적화 효소 엔지니어링

효소는 합성생물학 분야에서 가장 필수적 핵심 도구 중 하나로서, 새로운 대사 경로를 설계 및 구축하고, 유용한 화학물질을 합성하거나 바이오 연료등을 생산할 때에 가장 필수적인 부품 중에 하나이다. 기존에는 유전공학 (genetic engineering), 유도진화 (directed evolution)기법을 활용하여 효소의 특정 기능 최적화시켰으나, 바이오 파운데이션 모델이 앞으로는 다양한 형태의 효소 기능 최적화 디자인을 설계해줄 것으로 예상된다. 바이오 파운데이션 모델의 생성 능력은 특히 기존에 잘 알려지지 않은 orphan reaction을 위한 새로운 효소 설계 역시 가능케 할 수 있으며, 예를 들어서 enzymeCAGE와 같은 바이오 파운데이션 모델은 새로운 기질 (substrate)의 반응을 촉매해줄 수 있는 새로운 효소를 디자인해줄 수 있으며, Orphan-194 데이터셋을 기존 대비 73% 이상 높은 정확도로 예측 가능성이 확인된 바 있다. 최근에는 비천연 보조인자 (co-factor) 디자인 등을 통한 인공 효소 시스템을 개발하여 성공하였으며, 그 외에도 특정 조건을 만족하는 효소 서열을 생성하는 proGen 모델과 최적의 변이 모델을 생성하는 GFlowNet등의 다양한 인공지능 모델들이 도입되고 있다. 그리고 기존에 개발되었던 알파폴드 모델, 오픈폴드 모델 모두 효소 엔지니어링에 활용되며, 활성부위 발굴 및 시각화, 안정성 개선, 도킹 시뮬레이션 등에 활용되어서 효소의 기능 최적화 엔지니어링에 활용되어질 수 있다.

그림 5> 효소엔지니어링을 위한 enzymeCAGE 파운데이션 모델



출처: Yong Liu et al., BioRxiv (2024)

4 바이오 파운데이션 모델의 한계점 및 바이오 안전성 강화 방안 모색

바이오 파운데이션 모델은 위와 같이 다양한 합성생물학 분야에 응용될 수 있는 핵심 도구로 자리 잡을 것이며, 이를 개발하기 위한 고품질의 합성생물학 연구 데이터 축적도 같이 선행되어야 한다. 특히 바이오 파운데이션 모델은 학습에 사용된 데이터의 품질과 양에 의해서 성능이 좌우될 수밖에 없는데, 특히 바이오메디칼 고품질 데이터 확보 및 레이블링된 데이터는 바이오 파운데이션 모델 분야뿐만 아니라 관련분야 중 하나인 생명정보학 분

야 전반에서도 아직 완전히 해결하지 못한 큰 숙제 중에 하나이다. 특히 생물학적 컨텍스트에 따른 선택적인 데이터 편향으로 인해서 모델 성능에 불균형을 가지고 올 수 있으며, 전사체와 같은 일부데이터들은 원 데이터 (raw data)에서 임베딩 벡터로 압축하는 과정에서 심각한 정보 손실을 겪기도 한다. 특히, 복잡한 생명 현상을 모델링해야하기 때문에, 단순한 멀티모달 데이터 통합만으로는 해결되지 않는 문제점을 노출하고 있으며, 설령 고품질의 바이오 데이터가 구축이 된다고 하더라도, 바이오 파운데이션 모델 자체가 가지는 블랙박스 문제로 인한, 해석의 한계성, 신뢰도 측정 방법론의 부재 등으로 인해서 신뢰성있는 바이오 파운데이션 모델 구축에 대해서 다각적인 논의가 꼭 수반되어야 한다.

특히 바이오 파운데이션 모델이 가지고 있는 De novo 디자인 생성 능력은 합성생물학 분야에서 강력한 “도구”이자, 심각한 생물보안 위험을 초래할 수 있는 “무기”로도 악용되어 사용되어질 “양날의 검”이 될 수도 있다. 기존의 생물보안 안전장치들이 위험 서열 (sequence of concern, SOC)에 대한 서열 유사성을 바탕으로 스크리닝 장치가 마련되고 있기 때문에, 자연단백질과 낮은 서열 유사성을 가지지만, 특정 기능을 동일하게 가지는 인공 단백질들을 바이오 파운데이션 모델을 바탕으로 생성하게 되면, 생물보안 스크리닝을 우회할 수 있는 바이오 보안 격차를 만들어낼 가능성이 존재한다. 특히 기존의 엔지니어링 접근법과 달리, 바이오 파운데이션 모델을 바탕으로 새롭게 생성되는 생명시스템 부품들에 대한 레퍼런스를 찾기 어렵기 때문에, 규제 당국의 통제를 벗어날 수 있는 역설적인 상황도 발생할 수도 있다. 따라서, 이러한 바이오 파운데이션 모델의 적절한 통제와 규제를 위해서는 오히려 오픈소스 인파구축과 투명성 강화, AI 기반 생물보안 스크리닝 툴 개발 등의 다양한 접근법을 새롭게 고안하여, 개발과 보안을 동시에 가져갈 때에, 개발자, 사용자, 규제당국 모두에게 성공적인 결과를 가져올 것으로 예상된다.

5

맺음말

바이오 파운데이션 모델은 현재 신약개발 분야뿐만 아니라 합성생물학 개발 전주기에 있어서 핵심 도구로 자리 잡을 것으로 예상되고 있다. 현재까지 대규모로 생성된 분자, 화합물, 구조, 바이오메디컬 데이터를 활용한 바이오 파운데이션 모델 개발을 통해서, 다양한 추론 모델들이 만들어질 예정이며, 기존 자연계에 존재하지 않았던 새로운 생명 시스템을 구현하는 데에 핵심 축을 담당할 것이다. 예를 들어서, 유전체, 전사체 파운데이션 모델, 단백질 구조 파운데이션 모델 또는 멀티모달 파운데이션 모델을 바탕으로 유전자 조절 회

로를 해석하고 설계하거나, 특정 기능성을 가지는 다양한 인공 단백질들 설계, 합성해내며, 맞춤형 치료제 개발을 위한 새로운 치료 타겟발굴을 가능케 할 것이다. 그러나 바이오 파운데이션 모델은 학습에 사용되는 데이터에 크게 종속적이기 때문에, 고품질 데이터 확보와 데이터 편향 문제를 해결하는 것이 필수적이다. 이를 위해선 바이오 파운데이션 모델 개발에 활용 되어질 데이터 공유에 대한 국가적 지원 체계가 확보되어야 할 것이며, 바이오 데이터 생산, 기탁, 공유, 활용 각 단계에 있어서 전주기적 지원 체계가 잘 확립되어야 할 것이다. 이러한 데이터 공유 체계는 단순히 바이오 파운데이션 모델 성능 고도화를 위해 사용되는 것뿐만 아니라, 이미 기 구축된 바이오 파운데이션 모델을 통해서 새롭게 생성되는 단백질 설계에 대한 바이오 안정성 검증을 위한 검증 데이터로서도 활용될 수 있을 것이다.

[참고문헌]

- [1] Noelia Ferruz et al., “ProtGPT2 is a deep unsupervised language model for protein design”, Nature Communications (2022)
- [2] Saro Passaro et al., “Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction”, BioRxiv (2025)
- [3] Garyk Brixi et al., “Genome modeling and design across all domains of life with Evo 2”, BioRxiv (2025)
- [4] Katie Kavanagh, “World’s first AI-designed viruses a step towards AI-generated life”, Nature News (2025)

- [5] Samuel H King et al., “Generative design of novel bacteriophages with genome language models”, BioRxiv (2025)
- [6] NVIDIA BioNeMo, AWS HPC Blog
- [7] Biomedical Foundation Model, IBM Research
- [8] Therapeutics Data Commons (TDC), <https://tdcommons.ai/>
- [9] Kexin Huang et al., “Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development”, Arxiv (2021)
- [10] Yong Liu et al., “EnzymeCAGE: A Geometric Foundation Model for Enzyme Retrieval with Evolutionary Insights”, BioRxiv (2024)
- [11] Yanrong Ji et al., “DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome”, Bioinformatics (2021)
- [12] Zeming Lin et al., “Evolutionary-scale prediction of atomic-level protein structure with a language model”, Science (2023)
- [13] Gustaf Ahdritz et al., “OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization”, Nature Methods (2024)
- [14] Hugo Dalla-Torre et al., “Nucleotide Transformer: building and evaluating robust foundation models for human genomics”, Nature Methods (2024)
- [15] Eric Nguyen et al., “HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution”, Arxiv (2023)
- [16] Christina V. Theodoris et al., “Transfer learning enables predictions in network biology”, Nature (2023)
- [17] Fan Yang et al., “scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data”, Nature Machine Intelligence (2022)
- [18] Minsheng Hao et al., “Large-scale foundation model on single-cell transcriptomics”, Nature Methods (2024)

[주요용어풀이]

- Foundation Model (파운데이션 모델): 레이블이 지정되지 않은 대량의 원시데이터를 비지도 학습 방식으로 훈련된 모델. 하나의 파운데이션 모델을 미세 조정 (fine-tuning) 하여서 다양한 특정 작업을 수행할 수 있도록 만들 수 있음. 자연어 처리, 컴퓨터 비전, 분자구조 데이터 학습 등 다양한 분야에서 활용 가능함.

합성생물학 기반 지능형(Intelligent) 진단센서

1 머리말

감염병, 암, 항생제 내성(antimicrobial resistance, AMR)과 같은 복합 보건 위협이 전 지구적 차원에서 동시다발적으로 증가하면서, PCR(polymerase chain reaction)과 ELISA(enzyme-linked immunosorbent assay)에 기반한 기존 분자진단 패러다임만으로는 신속하고 유연한 대응에 한계가 있다는 점이 분명해지고 있다. 코로나19 팬데믹은 신종·재출현 병원체, 지속적인 변이의 등장, 지역사회 단위의 산발적 확산이 동시에 일어나는 상황에서, 초고민감도·다중 표적·현장형(point-of-care, POC)·저비용·대량 생산 가능성을 모두 만족하는 차세대 진단 기술이 필요함을 명확히 드러냈다. 여기에 고령화와 만성질환, 희귀질환, 보건안보 차원의 생물위기 대응까지 더해지며 “언제, 어디서, 무엇을, 얼마나 자주” 진단해야 하는지에 대한 요구 조건은 과거보다 훨씬 복잡해지고 있다.

합성생물학(synthetic biology)은 이러한 요구 변화에 대응하는 새로운 진단 기술 축으로 부상하고 있다. 유전자, 프로모터, 리보솜 결합 부위, 센서 단백질 등을 표준화된 모듈로 간주하고, 이를 논리 회로처럼 조합해 특정 핵산·단백질·대사산물에 반응하는 “프로그래머블 바이오 센서”를 구현하는 접근이 바로 합성생물학 기반 바이오센서(synthetic biology-enabled biosensor, SBB)이다 [1]. 합성 유전자 회로는 설계 대상 서열만 교체하면 새로운 병원체, 변이, 암 바이오마커에 비교적 짧은 시간 안에 대응할 수 있어, 변화 속도가 빠른 감염병·암·AMR 환경에서 전략적 유연성을 제공한다. 동시에 이러한 회로는 무세포(cell-free) 시스템, 종이 기반 플랫폼, 미세유체칩(microfluidic chip), 웨어러블 텍스타일 등과 결합되며, “랩 온 페이퍼(lab-on-paper)”와 같은 저비용 현장형 진단 양식으로 빠르게 확장되고 있다.

최근의 흐름은 SBB를 단순히 한 번 측정하고 끝나는 센서가 아니라, 연산(computing)·기억(memory)·자가보정(self-calibration) 기능을 내장한 지능형(intelligent) 진단 시스템으로 고도화하는 방향으로 이어지고 있다. AND/OR/NOT 등의 로직을 구현하는 제어 회로, 토글 스위치와 리콤비나제 기반 메모리 회로, 누설(leakage)을 억제하는 피드백 회로 등이 통합되면서, 센서는 단순한 “검출기”를 넘어 복잡한 분자 정보를 해석하고 상태 변화를 추적하는 “분자 연산 장치”에 가까워지고 있다. 이때 인공지능(artificial intelligence, AI)은 회로 설계·최적화·운영 전 단계에 걸쳐 핵심적인 역할

을 수행한다. 예를 들어 Angenent-Mari 등이 구축한 대규모 토홀드 스위치 라이브러리는 딥뉴럴네트워크(deep neural network, DNN)의 학습 데이터가 되어 기존 열역학·동역학 모델보다 약 10배 높은 성능 예측 정확도를 달성하였고, VIS4Map과 같은 설명가능 AI(explainable AI, XAI) 기법은 누설 신호를 유발하는 RNA 2차 구조 패턴을 시각적으로 드러내어 설계 실패의 원인을 구조 수준에서 규명할 수 있게 했다 [2, 3]. 나아가 나노소재 기반 광학·전기화학 바이오센서에서도 머신러닝·딥러닝을 활용해 복잡한 스펙트럼 신호에서 항생제 내성 패턴을 추출하고, 극미량 병원체 신호를 분리·분류하는 연구들이 축적되고 있다.

이처럼 AI와 합성생물학의 통합은 in silico 단계에서의 서열 설계, 클라우드-연계 자동화 실험실(cloud lab)에서의 고속 검증, 현장 장비에서의 실시간 패턴 분석까지 진단 전주기를 관통하는 새로운 연구·개발 패러다임을 형성하고 있다. 국내 관점에서 감염병, 암, AMR, 환경·식품 등의 분야에서 “AI × 합성생물학 진단 플랫폼”을 선제적으로 확보하는 것은 보건안보와 바이오경제 경쟁력 측면에서 전략적 과제로 간주될 수 있다. RT-RPA-CRISPR 기반 일회용 현장진단 키트, 종이 기반 CFPS-CRISPR 플랫폼, EV mRNA·miRNA·ctDNA를 초고민감도로 검출하고 AI로 패턴을 해석하는 액상생검 진단, 나노소재-CRISPR-AI를 결합한 AMR 바이오센서, CFPS-Cas와 스마트폰-AI 판독을 결합한 법과학·환경·식품 안전용 진단 플랫폼 등은 그 대표적인 예라 할 수 있다 [4]. 본 보고서는 이러한 국제·국내 기술 동향을 바탕으로, “AI × 합성생물학 진단”을 축으로 하는 지능형 진단센서 기술의 구조와 응용, 산업화 가능성을 조망하고자 한다.

2 합성생물학 기반 진단 플랫폼의 개요

2.1 Programmability 기반의 설계 가능성

합성생물학 기반 진단 플랫폼은 DNA·RNA·단백질·세포 등을 기능 단위로 분해한 뒤, 이를 다시 조합해 감지-처리-출력 기능을 수행하도록 설계하는 “프로그래머블 진단 시스템”이다 [5]. PCR(polymerase chain reaction)과 ELISA(enzyme-linked immunosorbent assay)가 이미 완성된 반응 조건과 장비 성능에 의존하는 정적인 기술이라면, 합성생물학 진단은 표적이 바뀔 때마다 회로 자체를 새로운 규칙으로 재프로그래밍할 수 있다는 점에서 본질적으로 다르다. 토홀드 스위치, CRISPR-Cas 인식 모듈, 리보스위치, 아르고노트 단백질과 같은 서열 기반 인식 요소는 목표 서열만 바꾸면 새로운 병

원체·변이·바이오마커에 즉각 대응할 수 있도록 설계할 수 있으며, 감염병 변이, 암 이질성, 항생제 내성 패턴과 같이 빠르게 변화하는 표적 공간에 기민하게 적응할 수 있는 기반을 제공한다.

더 나아가 합성생물학적 프로그래밍은 단순 존재 여부 감지를 넘어, 조건부 활성화, 다중 입력 논리 연산, 피드백 억제 등 “분자 연산” 수준의 기능 구현을 가능하게 한다 [6]. 예를 들어 병원체 RNA와 내성 유전자가 동시에 존재할 때만 반응하는 AND 게이트형 센서, 특정 miRNA 조합이 충족될 때만 작동하는 암 특이 회로는 기존 단일 표적 기반 진단으로는 어렵던 정밀 분자 프로파일링을 가능하게 한다. 이때 AI는 설계 공간 탐색을 가속하는 핵심 도구로 작동한다. 대규모 서열-기능 데이터셋을 학습한 예측 모델은 어떤 서열 조합이 높은 감도·특이도를 낼지 사전에 점수화하고, 생성형(transformer 기반) 모델은 misfolding 가능성이 낮고 누설이 적은 새로운 스위치 구조를 제안함으로써 “직접 설계”에서 “AI 보조 설계”로의 전환을 이끌고 있다. 결과적으로 Programmability는 “하드웨어는 유지한 채 소프트웨어(서열·회로)만 갈아끼우는” 업데이트형 진단 체계를 가능하게 하고, AI는 이 업데이트 과정을 빠르고 체계적으로 수행하게 해주는 설계 엔진 역할을 한다.

2.2 Modularity 및 대규모 설계 확장성(Scalability)

Modularity(모듈성)는 합성생물학 진단의 공학적 핵심이다. 복잡한 생물 반응을 입력(Input)·제어(Control)·출력(Output)과 같은 표준화된 모듈로 분리해 설계·조합·대체할 수 있도록 만드는 것이다 [7]. BioBrick, Golden Gate, Gibson Assembly 등 DNA 조립 표준은 기능 단위 모듈을 USB 규격처럼 다룰 수 있게 해주며, 특히 Golden Gate의 Type IIs 제한효소 기반 scarless 조립은 회로의 정확성과 재현성을 크게 높여준다. 이러한 모듈식 구조 덕분에 동일한 센서 회로가 세포 기반 시스템, 무세포(cell-free) 시스템, 동결건조 패치, microfluidic chip, paper-based μ PAD 등 다양한 형태의 디바이스로 손쉽게 이식될 수 있고, 진단 “하드웨어”와 독립적으로 설계가 가능하다 [8].

Scalability(확장성)는 이 모듈성을 AI와 자동화 기술이 확장하는 방향으로 구현된다. 합성유전자 라이브러리 자동 제작, FACS-NGS 기반 Flow-seq, 액체 취급 로봇, 클라우드 기반 자동화 실험실은 수천~수만 개의 회로 조합을 병렬로 시험할 수 있게 만든다. 이렇게 축적된 대규모 실험 데이터는 딥러닝 모델이 서열-구조-기능 관계를 직접 학습하는 “데이터 기반(data-driven) 설계”의 기반이 되며, 예측 모델은 어떤 조합을 우선 실험해야 하는지 설계 단계에서부터 랭킹을 제시한다. 더 나아가 생성형 모델은 학습된 규칙을 바탕으로 기존에 없던 회로 설계를 제안하고, 클라우드 랩은 이 설계를 자동으로 조립·테스

트·피드백하는 DBTL(Design-Build-Test-Learn) 사이클을 구현한다.

결국 Programmability-Modularity-Scalability는 합성생물학 진단을 “AI가 설계를 주도하고, 자동화 실험이 이를 검증하며, 회로는 여러 디바이스 위에서 재사용되는” 형태의 플랫폼 기술로 진화시키고 있다. 이는 감염병·암·AMR·환경·식품 등 다양한 영역에서 신속·정밀·저비용·현장형 진단 시스템을 구축하는 데 있어 핵심적인 구조적 토대다.

3 주요 합성생물학 진단 플랫폼

3.1 Toehold Switch 기반 RNA 센서

Toehold switch는 합성 RNA로 구성된 구조적 스위치로, 입력 RNA의 존재 여부를 구조 변화로 변환해 단백질 리포터 발현을 유도하는 플랫폼이다 [9]. 기본 구조는 RBS(ribosome binding site)와 Start codon이 이차구조의 스템-루프 내에 숨겨져 있어 기본적으로 번역이 억제된 상태로 유지되며, 표적 RNA가 toehold 영역에 결합하면 RNA 구조 전체가 해체되면서 RBS가 노출되고 번역이 개시되는 방식이다. 이러한 작동 원리는 전적으로 염기서열 상보성에 기반하므로, 신규 병원체나 돌연변이가 등장하더라도 표적 서열만 변경하면 즉시 새로운 센서를 제작할 수 있다는 강력한 프로그래밍성을 제공한다. 또한 Toehold switch는 in vitro(cell-free) 시스템과 매우 높은 호환성을 가지며, 동결건조 후 종이나 필름 위에서 재수화만으로 작동하기 때문에 저비용·원팟(one-pot) POC(point-of-care) 진단이 가능하다. 실제로 Zika, SARS-CoV-2, Ebola 등 다양한 공중보건 병원체에 적용된 바 있으며, RNA 구조 기반이므로 단일 염기 변이 구별(variant typing)에도 높은 잠재력을 가진다. 더불어 대규모 라이브러리를 제작한 후 Flow-seq 기반 고속 스크리닝을 통해 수만 개의 설계 조합에서 최적의 스위치를 찾을 수 있어, AI 기반 예측모델과 결합될 때 플랫폼의 확장성이 극대화된다.

3.2 Riboswitch 기반 분자 감지

Riboswitch는 원래 세포 내에서 특정 대사물질이 결합하면 RNA 구조가 재배열되어 전사 또는 번역 조절이 일어나는 천연 RNA 조절 요소이다. 합성생물학은 이러한 자연 기작을 모방해, 특정 리간드가 결합하면 구조적 변화를 통해 ON/OFF 출력이 제어되는 인공 riboswitch를 설계한다 [10]. 이들 센서는 small molecule(금속 이온, 환경 독소, 약

물, 대사산물 등)을 직접 감지할 수 있다는 점에서, 핵산 기반 플랫폼과 상호보완적인 기능을 가진다. 또한 riboswitch는 전기적 장치 없이도 분자 결합만으로 구조 변화가 일어나기 때문에 높은 안정성과 반응 속도를 제공하며, 세포 내 대사 네트워크와 직접 연동이 가능해 환경 모니터링, 독성 물질 감시, 세포 기반 치료제의 내부 상태 모니터링 등 다양한 적용 가능성을 가진다. 최근에는 aptamer-switch 하이브리드 구조나 multi-input riboswitch도 개발되면서 복잡한 논리 연산이 가능한 RNA 제어 회로로 확장되고 있다.

3.3 전사조절 인자 기반 센서(TF-biosensor)

전사조절인자(Repressor 또는 Activator)를 이용한 TF 기반 센서는 단백질과 리간드의 결합을 분자 신호로 변환하는 방식이다 [11]. 특정 리간드가 TF에 결합하면 DNA 상의 operator 영역과의 결합력이 변화하고, 그 결과 downstream 유전자의 발현이 조절되어 최종 신호가 출력된다. TF 기반 센서는 구조적으로 견고하고 환경 변화에 강인하며, 세포 대사 전반과 깊이 연결되어 있기 때문에 실제 산업 미생물 엔지니어링에서 특정 대사산물 농도를 모니터링하거나 독성 화학물질을 실시간 감지하는 데 매우 유용하다. 또한 감염병 진단뿐 아니라 법과학 현장에서 약물·독소의 현장 분석, 식품 산업에서 부패·오염 감지, 환경 분야에서 중금속 및 농약 검출 등 다양한 실용적 응용을 보인다. 합성 TF 또는 재설계된 DNA-binding motif를 도입하면 타겟 선택성을 크게 확장할 수 있어, 향후 맞춤형 화학물질 감지 시스템 구축의 기반이 된다.

3.4 CRISPR-Cas12/13/14 기반 진단

CRISPR 기반 진단은 Cas12, Cas13, Cas14의 비표적 절단 활성(collateral cleavage)을 이용해 신호 증폭 및 고감도 검출을 구현하는 차세대 핵산 진단 플랫폼이다 [12]. Cas12는 DNA 표적을 인식하면 ssDNA reporter를 자르고, Cas13은 RNA 표적을 인식하면 ssRNA reporter를 절단함으로써 형광·전기화학·색 반응 등으로 출력된다. 이 기술은 단일 염기 변이(SNP)까지 정확히 구별할 수 있을 만큼 특이도가 높아 SHERLOCK, DETECTR 등의 플랫폼으로 발전해 실제 감염병·바이러스 변이 분석·암 표적 검출에 널리 사용되고 있다. 특히 Cas14는 초소형(400~700 aa) 단백질로, 저온·저자원 환경에서도 안정하게 작동해 차세대 휴대형 센서 개발의 핵심 후보로 주목받는다. 또한 Cas의 타겟 인식 메커니즘이 PAM/PFS 서열 특성을 기반으로 하기 때문에, AI 기반 gRNA 최적화와 결합하면 기존 진단 방식에서 불가능했던 초고감도·다중화(multiplex)·원스텝(one-pot) 반

응 시스템 구축이 가능하다.

3.5 Cell-Free Systems 기반 Rapid Prototyping

Cell-free protein synthesis(CFPS)는 살아 있는 세포 없이, 세포에서 추출한 효소·리보솜·tRNA 등 핵심 구성 성분을 시험관 내에서 직접 구동하는 시스템이다 [13]. CFPS 기반 플랫폼은 세포 성장 시간(수 시간~수십 시간)이 필요 없는 장점이 있어, DNA 회로가 의도대로 작동하는지를 수 시간 내 확인할 수 있는 “Rapid Prototyping”에 최적화되어 있다. 또한 세포 내에서 흔히 발생하는 metabolic burden, 독성 물질 축적, 세포 스트레스 반응 등이 제거되므로 합성 회로의 기능이 더 정확하게 반영된다. Freeze-drying(동결건조)을 통해 종이·필름·패치 형태로 휴대하거나 장기 보관할 수 있어, 현장형(Point-of-Care) 진단 플랫폼의 핵심 기술로 자리 잡고 있다. 최근에는 LoC, 섬유(textile), 웨어러블 디바이스에 CFPS·CRISPR 회로를 직접 통합하는 연구가 활발하다.

4 지능형(Intelligent) 합성생물학 진단 시스템의 구조

4.1 개념적 진화: 감지에서 연산·학습·적응으로

초기 바이오센서는 단순히 특정 분자(RNA, 단백질, 독성 물질 등)의 존재 여부를 신호로 변환하는 감지 중심의 구조였다. 그러나 감염병 변이의 급속한 등장, 항생제 내성(AMR)의 복잡한 전파 네트워크, 암에서의 극저농도 희귀 변이 탐지와 같은 고도화된 진단 문제를 해결하려면 단순한 ON/OFF 판별만으로는 충분하지 않다. 새로운 진단 시스템에는 복수의 입력을 조합해 조건부 판단을 수행할 수 있는 연산 능력, 지속적으로 노출된 자극을 보관하는 기억 기능, 환경 변화에 따른 자동 보정(self-calibration) 능력, 그리고 데이터 기반의 지능적 설계 및 성능 최적화가 요구된다.

이러한 필요성에 따라 합성생물학 기반 진단 플랫폼은 감지(Sensing)에서 출발해, 연산(Computing), 기억(Memory), 학습(Learning), 적응(Adaptation) 기능을 단계적으로 통합하는 방향으로 진화하고 있다 [14, 15]. 연산 기능은 AND, OR, NOT, NAND와 같은 Boolean 논리를 유전자 회로 수준에서 구현하여 다중 바이오마커가 특정 조합을 이룰 때만 신호가 출력되도록 만든다. 이는 변이를 구분하거나 암과 같은 복잡 질환을 진단할 때 위양성(false positive)을 억제하는 핵심 기술적 기반을 제공한다. 기억 모듈은 일시

적 자극을 DNA 재조합, 돌연변이 축적, 단백질 상태 변화 등을 통해 장기간 저장할 수 있어, 체내 미생물 기반 라이브 진단처럼 지속적인 모니터링이 필요한 상황에서 특히 중요하게 활용된다.

자가보정 기능은 합성 유전자 회로의 대표적인 문제인 OFF-state leakage를 억제해 낮은 농도에서도 높은 신호대잡음비(SNR)를 유지하도록 하며, 환경 조건 변화에 대한 시스템 안정성을 높인다. 이어서 학습 단계에서는 AI가 서열-구조-기능 간의 복잡한 상관관계를 스스로 학습하여 회로 설계를 자동화하고 성능을 최적화한다. 최근에는 Transformer 기반 생성 모델이 de novo로 새로운 서열을 설계하는 수준까지 발전해, 설계-실험-최적화 과정을 크게 가속하고 있다.

마지막으로 적응 단계에서는 합성생물학 센서를 하이드로겔, 3D 프린팅 구조체, 식물 기반 스마트 소재 등과 융합함으로써 외부의 습도·압력·화학적 자극 변화에 능동적으로 반응하는 “지능형 생체-물질 하이브리드 플랫폼”으로 확장되고 있다. 이러한 플랫폼은 환경 변화에 즉각적으로 적응하면서도 신뢰도 높은 생물학적 감지를 수행할 수 있어, 기존의 전자적 센서가 가지지 못했던 고유한 장점을 발휘한다.

4.2 모듈 설계: BioBrick → Golden Gate → 빠른 프로토타입

합성생물학 진단 시스템의 성능은 모듈을 얼마나 빠르고 정확하게 조립하느냐에 달려 있다. 초기의 BioBrick 표준은 DNA 부품을 정해진 제한효소 인식부로 절단·조립하는 방식이었으나, 인식부가 최종 산물에 “scar”로 남아 유전자 기능을 방해하는 문제가 있었다. 이러한 한계를 극복한 기술이 Type IIS 제한효소를 이용하는 Golden Gate Assembly이다. Golden Gate는 제한효소의 인식부와 절단부가 분리되어 있어, 인식부는 최종 산물에서 제거되고, 조립된 부위에는 사용자가 원하는 4-5 nt의 overhang만 남게 된다. 이를 통해 완전한 scarless assembly가 가능해졌고, 서로 다른 overhang을 설계함으로써 다양한 모듈을 시험관 하나에서(one-pot) 정확한 순서로 조립할 수 있게 되었다.

이러한 모듈화·표준화 패러다임은 이후 MoClo(Modular Cloning), GoldenBraid, Synthetic Promoter Library(SPL) 등으로 확장되며, 합성 회로의 자동 설계·조립을 위한 산업적 기반을 형성하였다. 특히 cell-free 시스템과 Golden Gate 조립을 결합하면, DNA 조립-발현 테스트-성능 평가가 하루 이내에 이루어지는 초고속 DBTL(Design-Build-Test-Learn) 사이클을 구축할 수 있다. 이는 Toehold switch와 CRISPR gRNA 라이브러리처럼 수천~수만 조합을 빠르게 테스트해야 하는 진단 플랫폼 개발에서 필수적인 기술이다.

4.3 연산·기억·피드백 회로 통합

지능형 합성생물학 기반 진단 시스템은 단순한 반응 모듈을 넘어서, 연산(Computing), 기억(Memory), 피드백(Self-calibration) 기능을 동시에 구현하는 복합 제어 구조로 진화하고 있다. 이러한 세 가지 모듈은 독립적으로 움직이기보다는 상호 보완적으로 작동하여, 전체 시스템의 정확도·안정성·적응성을 극대화하는 방향으로 결합된다.

먼저 연산 회로는 여러 개의 입력 신호를 조합하여 특정 조건에서만 출력을 생성하는 기능을 수행한다. 대표적인 접근 방식은 dCas9과 상호 간섭이 없는 직교적 gRNA 조합을 활용해 AND, OR, NAND, NOR와 같은 Boolean 논리를 유전자 회로 수준에서 구현하는 것이다. 예를 들어 암 진단에서는 두 개 이상의 바이오마커가 모두 존재할 때만 신호가 생성되도록 AND 게이트를 설정하여 위양성을 크게 줄일 수 있다. 변이 바이러스 진단에서는 특정 뉴클레오타이드 패턴이 동시에 충족될 때만 반응이 일어나도록 설계함으로써, 기존 RT-PCR 기반 판별보다 높은 정밀도로 variant typing을 수행할 수 있다.

기억 회로는 입력 신호가 사라진 이후에도 그 정보를 저장할 수 있는 기능을 제공하며, 환경 모니터링, 장내 미생물 기반 라이브 진단, 암 미세환경의 이력 추적과 같이 시간 누적 정보가 중요한 영역에서 핵심적이다. Toggle switch 구조는 두 가지 안정한 상태(A/B)를 갖고 특정 자극을 계기로 상태가 영구적으로 전환되기 때문에 장기적인 기록에 유리하다. Recombinase 기반 회로는 DNA 절단과 반전 같은 재조합 반응을 이용해 외부 자극을 서열 수준의 영구적 변형으로 저장한다. 또한 CRISPR 기반 Tape Recorder 시스템은 Cas 단백질이 특정 시점에 순차적으로 돌연변이를 축적해 시간 순서에 따라 기록을 남길 수 있도록 설계되어, 기존 진단 기술이 제공하지 못했던 시간적 해상도의 정보를 획득할 수 있게 한다. 이러한 메모리 회로들은 생체 환경에서 수개월 이상 안정적으로 유지될 수 있어, 장내 염증 신호나 장기적 환경 독성 노출과 같은 사건의 누적 기록에 특히 유용하다.

피드백 회로는 합성 유전자 회로의 고질적인 문제 중 하나인 OFF-state leakage(누설 신호)를 해결하기 위해 도입된다. 누설 신호는 특히 저농도 표적을 검출해야 하는 진단에서 민감도를 감소시키는 주요한 요인으로 작용한다. 이를 극복하기 위해 다양한 자가 보정 구조가 활용되는데, 대표적으로 negative feedback은 출력 신호가 일정 수준 이상 증가하면 이를 다시 억제하는 방식으로 과도한 신호 폭주를 방지한다. Feedforward loop 구조는 입력 신호가 여러 경로를 통해 출력의 활성화와 억제 양쪽에 동시에 영향을 주도록 하여, 잘못된 신호가 발생할 가능성을 구조적으로 차단한다. 더 나아가 dynamic regulation 회로는 외부 환경 변화(pH, 온도, 이온 농도, 대사 상태 등)에 따라 시스템 반응성을 조절함으로써, LoC 칩 또는 종이 기반 센서처럼 실험 조건 변동성이 큰 환경에서

도 높은 신호 안정성을 확보하도록 돕는다.

이와 같이 연산, 기억, 피드백 기능의 통합은 디지털 논리적 판단, 시간 축 기반 정보 기록, 환경 변화에 대한 적응적 반응을 동시에 구현할 수 있게 하여, 기존의 단순한 분자진단 플랫폼을 넘어서는 고도화된 지능형 합성생물학 진단 시스템의 기반 기술로 자리 잡고 있다 [16].

5 AI·클라우드 기반 진단 회로의 자율 설계

5.1 기존 모델의 한계와 데이터 기반 접근

합성생물학 기반 진단 회로의 성능은 RNA·DNA 구조 형성, 단백질-핵산 상호작용, 회로 간 간섭(circuit crosstalk) 등 복잡한 생물학적 요인의 영향을 받는다. 그러나 기존 설계 도구인 NUPACK, ViennaRNA 등의 열역학 기반 모델은 RNA 접힘(folding)의 동역학적 속성, 세포 내 복잡 환경(in vivo context), collateral cleavage와 같은 효소적 비선형성 등을 반영하지 못해 실제 센서 성능과의 상관도가 매우 낮다는 한계점이 있다 [17].

이로 인해 기존 방식은 합리적 설계(rational design)라기보다 연속적인 시행착오(trial-and-error)에 가까웠으며, 특히 Toehold switch·CRISPR gRNA·riboswitch 등 구조의 존적 센서는 설계-테스트-수정의 반복 비용이 매우 컸다.

이 문제를 해결한 대표적 접근이 대규모 실험 데이터 기반의 AI 학습(model training on Big Data)이다. Angenent-Mari et al.(2021)은 9만 개 이상의 Toehold switch 라이브러리를 직접 합성하고 Flow-seq(FACS + NGS) 기반 고속 스크리닝으로 방대한 성능 데이터를 구축했다. 이를 DNN(Deep Neural Network)에 학습시킨 결과, 열역학 모델 대비 10배 이상 높은 예측 정확도($R^2 = 0.43-0.70$)를 달성하였다 [2].

이는 AI 모델이 서열-구조-기능 간의 복잡한 비선형 관계를 스스로 학습할 수 있음을 보여주며, 합성생물학 설계 패러다임이 모델 기반 → 데이터 기반(data-driven)으로 전환되고 있음을 시사한다.

5.2 XAI 기반 설계 통찰 제공

AI 기반 예측 모델이 높은 정확도를 보이더라도, 연구자가 실제로 이를 설계 과정에 활용하기 위해서는 모델이 어떤 구조적 특징을 기반으로 판단을 내렸는지를 이해할 수 있어야 한다. 이러한 요구를 충족하기 위해 도입된 기법이 바로 설명가능한 AI(XAI) 기반의

VIS4Map(Visualizable Secondary Structure Saliency Mapping)이다 [18]. VIS4Map은 딥러닝 모델이 예측에 기여한 RNA 2차 구조 영역을 열지도 형태로 시각화함으로써, 스위치 실패의 원인이나 구조적 취약 영역을 직관적으로 파악하도록 돕는다. 실제 Toehold switch 라이브러리에 해당 기법을 적용한 연구에서는 AI 모델이 OFF-state 누설을 유발하는 스템 구조의 불안정성, trigger RNA와 경쟁 결합이 발생하는 부위, 혹은 잘못된 2차 구조 형성이 일어나는 hotspot을 정확히 강조하는 것으로 나타났다. 이러한 결과는 AI가 단순히 성능을 맞추는 예측기 역할을 넘어, 설계 상의 문제를 구조적 관점에서 설명하고 개선 방향을 제시하는 “설계 조언자(design advisor)”로 기능할 수 있음을 보여준다. 결국 VIS4Map과 같은 XAI 도구는 합성생물학 회로 설계의 불확실성을 줄이고, 고성능 바이오센서를 보다 체계적으로 구축할 수 있는 중요한 기반 기술로 자리잡고 있다.

5.3 Cloud-lab 기반 DBTL 자동화

AI 기반 설계는 단순히 in silico 예측에 머무르지 않고, 클라우드 기반 자동화 실험 환경과 결합되면서 진단 회로 개발의 새로운 패러다임을 열고 있다. 최근 개발된 클라우드 랩(Cloud-lab) 시스템은 설계-구축-테스트-학습(Design-Build-Test-Learn, DBTL)의 반복 과정을 인간의 개입 없이 자동으로 수행하는 자율형 연구 흐름을 구현한다 [19]. 이 과정에서 AI 모델은 먼저 표준화된 Golden Gate 혹은 모듈형 클로닝 규격에 맞추어 새로운 유전자 회로를 설계하고, 로봇 실험 장비는 해당 설계를 기반으로 DNA 조립을 즉시 수행한다. 조립된 회로는 배양이 필요 없는 cell-free 시스템에서 바로 발현·검출 단계를 거치며 수 시간 내 성능 테스트가 이루어진다. 측정된 신호는 클라우드로 다시 전송되어 AI 모델의 추가 학습에 활용되고, 그 결과 모델의 예측 정확도는 반복될수록 점점 향상된다. 이러한 자동화된 DBTL 사이클은 기존 수개월이 걸리던 회로 최적화 과정을 며칠 또는 하루 단위로 단축시킬 수 있는 잠재력을 갖고 있다. 더 나아가 대규모 병렬 테스트가 가능해지면서 수천~수만 개 조합의 센서 설계를 연속적으로 탐색할 수 있어, 인간이 설계하기 어려운 복잡한 진단 회로도 자율적으로 최적화할 수 있는 기반이 마련되고 있다.

5.4 Generative AI 기반 new-to-nature 회로 설계

최근 AI 기술의 진전은 예측 모델을 넘어 생성 모델(GAN, Transformer, Diffusion model 등)로 확장되며 합성생물학 진단 회로 개발의 지평을 크게 넓히고 있다. 생성형 AI는 기존에 존재하던 회로의 단순 변형을 넘어, 학습된 생물학적 규칙과 구조 패턴을 바탕

으로 완전히 새로운 기능을 수행하는 서열을 de novo로 만들어낼 수 있다. 예를 들어, 기존 Toehold switch의 프레임워크를 유지하면서도 misfolding 경향이 줄어든 새로운 스템-루프 구조를 제안하거나, 자연계에는 존재하지 않는 riboswitch의 aptamer 도메인을 재설계해 새로운 대사산물에 반응하는 센싱 기능을 창조하는 것이 가능하다. 또한 생성형 모델은 dCas9 기반 억제 회로에서 orthogonal gRNA 조합을 자동으로 도출해 회로 간 간섭을 최소화하거나, Cas 단백질의 활성 도메인 변이를 예측해 더 높은 효율 또는 더 낮은 오프타겟 특성을 가진 단백질 서열을 직접 설계할 수도 있다. 이러한 접근은 기존의 생물학적 탐색 공간을 크게 확장하며, 인간 연구자가 직관적으로 설계하기 어려운 high-dimensional sequence-structure-function landscape를 AI가 능동적으로 탐색할 수 있게 만든다. 결국 생성형 AI는 단순한 도구를 넘어, 미래 지능형 진단 생태계의 중심 엔진으로 작동할 것으로 예상되며, 조기 감염병 대응, AMR 조기 탐지, 암 액상생검 기반 변이 분석 등 복잡하고 변화가 빠른 질병 환경에 최적화된 새로운 진단 플랫폼 개발을 가능하게 할 것이다.

6 실제 응용 분야

합성생물학 기반 진단 플랫폼은 감염병 대응, 암·액상생검 기반 정밀의학, 환경·식품 안전, 법과학 등 여러 분야에서 실질적 성과를 보이며 빠르게 확장되고 있다. 이 장에서는 지능형 SBB(Intelligent Synthetic Biology-enabled Biosensors)와 AI 기반 시스템이 실제로 적용되고 있는 대표 응용 사례를 분야별로 정리한다 [20-24].

6.1 감염병 및 항생제 내성(AMR) 진단

합성생물학 기반 핵산 진단 기술은 감염병 대응 분야에서 가장 빠른 속도로 실용화되고 있다. 특히 RT-RPA(reverse transcription-recombinase polymerase amplification)와 CRISPR-Cas 시스템을 통합한 one-pot 등은 핵산 진단 기술은 별도의 고가 장비 없이도 수십 분 내 핵산 증폭과 표적 절단, 신호 생성까지 모두 단일 반응에서 수행할 수 있어 현장형(Point-of-Care) 진단에 적합하다. 이러한 기술 흐름을 대표하는 사례가 SHERLOCK(Cas13 기반 RNA 진단), DETECTR(Cas12 기반 DNA 진단)이며, 두 플랫폼은 바이러스 RNA/DNA를 극미량 수준에서도 감지할 수 있는 초고감도(collateral cleavage 기반) 분석을 가능하게 하였다.

최근에는 Toehold-mediated strand displacement(TMSD)와 CRISPR 반응을 결합하여 단일 염기 수준(SNP-resolution)의 변이 분별이 가능한 기술도 등장하였다. TMSD 회로는 구조적으로 매우 예민한 strand displacement 반응을 통해 특정 염기 차이만으로도 반응 활성도를 구분할 수 있어, 변이 바이러스 판별이나 약물 내성 변이 탐지 등 고정밀 진단에 유용하다.

감염병과 항생제 내성(AMR)의 동시 모니터링도 중요한 연구 주제로 떠오르고 있다. CRISPR-Cas 기반 진단 회로는 높은 서열 특이성을 바탕으로, 병원체 유무와 동시에 ARG(antimicrobial resistance gene)의 존재 여부까지 판별하는 동시 검출(multiplex) 플랫폼으로 확장되고 있다. 예컨대 Cas12·Cas13의 orthogonal crRNA 세트를 이용하면, 동일 반응 시스템 내에서 병원체 유전자와 내성 유전자를 독립적으로 검출할 수 있어, 감염병 치료 전략 수립에 필요한 정밀 정보를 신속히 제공할 수 있다.

또한 μ PADs(microfluidic paper-based analytical devices)는 종이 기반 미세유체 구조를 활용해 low-cost, zero-instrumentation 환경에서도 핵산·단백질·대사물질 검출을 가능하게 하며, 최근에는 스마트폰 카메라 기반 AI 판독 시스템과 결합하여 현장에서 진단 결과를 자동 분석하는 플랫폼으로 발전하고 있다. 이러한 이동형·저비용·AI 판독 기반 시스템은 팬데믹 대응, 접경지역·도서지역 감시, 응급현장 모니터링 등 다양한 국가·지역 보건 상황에서 높은 활용 가능성을 보이고 있다.

6.2 암·액상생검 진단

액상생검은 암 진단 패러다임을 조직 기반에서 체액 기반으로 전환시키는 핵심 기술로 자리 잡고 있으며, 합성생물학 기반 진단 기술은 이 발전을 크게 가속하고 있다. 암세포에서 유래한 **EV RNA, miRNA, ctDNA(circulating tumor DNA)**는 조기 암 진단 및 치료 반응 모니터링에 매우 유용한 바이오마커로 알려져 있다. 특히 EV(Extracellular Vesicle)는 RNA·miRNA·단백질 등 기능성 cargo를 보호된 상태로 운반하기 때문에, SBB 기반 진단 회로가 직접 표적을 인식하기에 적합한 형태로 존재한다.

이러한 바이오마커를 고감도로 분석하기 위해 Lab-on-a-Chip(LoC)를 활용한 SERS(surface-enhanced Raman scattering) 기반 기술이 적극 활용되고 있다. 금속 나노입자를 이용해 Raman 신호를 극대화하는 SERS는 miRNA나 돌연변이 ctDNA와 같은 초저농도 표적을 안정적으로 감지할 수 있으며, 미세유체 칩 내부에서 시료 전처리·농축·검출 과정을 자동화함으로써 소량의 혈액만으로도 변이를 정밀하게 분석할 수 있다.

한편, Freeze-dried CFPS(cell-free protein synthesis) 기반 합성생물학 회로는 기존

PCR 대비 훨씬 간단한 절차로 15-30분 내 핵산 신호를 검출할 수 있어, 병원 외부나 자원이 제한된 환경에서도 암 관련 표적을 즉시 확인할 수 있는 실질적 장점을 제공한다. 이와 함께 AI 기반 패턴 분석 기술은 EV RNA·miRNA 발현 패턴이나 ctDNA 변이 패턴에서 신호 노이즈를 제거하고 single-molecule 수준의 차이를 식별하는 데 활용되며, 조기 암 진단의 민감도와 특이도를 크게 향상시키는 방향으로 발전하고 있다.

7 산업화·상용화 전략

합성생물학 기반 진단 플랫폼은 연구실 단계 기술을 넘어 실제 의료·산업 생태계로 빠르게 전환되고 있다. 그러나 상용화를 위해서는 기술적 성숙도(Technology Readiness Level, TRL)의 단계적 향상, 제조 공정의 표준화, 품질관리 체계의 확립, 규제 제도 준수 등 복합적인 고려가 필요하다. 특히 합성생물학은 DNA·RNA 회로, CRISPR 효소, 세포-프리 시스템과 같은 생물학적 요소를 기반으로 하기 때문에 일반 전자센서와 달리 바이오품질의 변동성·안정성·보관 조건이 제조 및 유통 단계에서 중요한 변수로 작용한다.

우선 산업적 전환에서 요구되는 핵심 과제는 Lab-scale 회로를 GMP(우수의약품 제조 및 품질관리 기준) 수준으로 확장하는 과정이다. 현재 대부분의 합성생물학 회로는 분자생물학 실험실에서 소규모 배치로 제작되고 있으나, 상용화를 위해서는 효소 안정성, 배치 간 변동성, 반응 재현성, 동결건조(freeze-drying) 조건의 표준화 등 다양한 공정 파라미터를 체계적으로 관리해야 한다. 이를 위해 생물학적 반응을 정량적으로 보정하는 품질관리(QC) 시스템, 온도·습도·충격에 대한 내성을 확보하는 포장·물류 기술, 장기간 보관을 위한 안정제(stabilizer) 조성 등 기반기술 개발이 필요성이 크다.

산업적 전환을 가속화하는 또 다른 요소는 AI 기반 공정 자동화 및 품질 예측 모델이다. 제조 과정에서 발생하는 변동성을 AI 모델이 실시간으로 분석해 불량률을 조기 예측하거나, 회로 디자인 단계에서 반응 실패 가능성을 사전에 평가함으로써 전체 개발 비용을 크게 절감할 수 있다. 실제로 제조-성능 간 관계를 학습한 예측 모델은 온도·습도·반응버퍼 조성의 변화에 따른 반응 편차를 보정하며, 최적 제조 조건을 도출하는 데 활용되고 있다. 마지막으로 시장 측면에서 합성생물학 기반 진단 기술은 감염병 대응, 항생제 내성 모니터링, 암 조기진단, 식품 안전, 환경 감시, 법과학 등 다양한 수요처에서 빠르게 시장을 형성하고 있다. 특히 “저비용·신속·현장 사용 가능”이라는 특징은 개발도상국 보건 시장, 군사 분야, 재난 대응 산업에서 강력한 경쟁력을 갖는다. 향후 시장 확장을 위해서는 규제 과학

의 정립, 보험 등재 체계 구축, 제조·임상 검증 간소화 지원 등 정책 기반 마련이 필요하다.

8 미래전망

합성생물학과 AI 기반 진단 기술은 국가 보건안보, 감염병 대응 전략, 정밀 의료, 식품·환경 안전, 법과학적 증거 분석 등 여러 분야에서 국가 경쟁력을 좌우할 핵심 기술군으로 부상하고 있다. 따라서 국가 차원의 전략은 단순한 기술 지원을 넘어, 산업 생태계 전체를 포괄하는 장기적 관점에서 추진될 필요가 있다 [25-28].

첫째, 감염병 대비 측면에서는 국가 단위의 합성생물학 기반 신속 진단 플랫폼 구축이 필요하다. 팬데믹을 계기로 확인된 바와 같이, 고가 장비에 의존하는 기존 PCR 체계만으로는 급속히 확산되는 신종 병원체에 대응하기 어렵다. 정부는 범용 CRISPR 기반 진단 키트, one-pot 핵산 증폭 기술, Toehold-Riboswitch 기반 RNA 진단 시스템 등을 비축 가능한 형태로 개발하고, 공항·항만·지역 보건소 등에서 즉시 활용할 수 있는 현장형 POC 인프라를 구축할 필요가 있다.

둘째, AI 기반 진단 회로와 클라우드 실험실(cloud lab) 인프라 구축은 국가 경쟁력 확보에 중요한 요소이다. 미국·유럽 등에서 이미 상용화된 cloud lab 시스템은 자동화된 로봇 실험 플랫폼을 기반으로 수천~수만 개의 회로 조합을 24시간 내 평가할 수 있어, 신속한 진단 기술 개발에 매우 유리하다. 우리나라도 합성생물학 DBTL(Design-Build-Test-Learn) 인프라를 국가 단위로 구축해 진단기술을 신속하게 최적화하고, 국내 산업 전반에 자동화 기반 R&D 생태계를 확산할 필요가 있다.

셋째, 규제·윤리 측면에서는 합성생물학 진단 기술의 특성을 반영한 맞춤형 규제 과학(Regulatory Science)이 요구된다. 예를 들어 CRISPR 기반 진단키트는 유전물질 절단 효소를 포함하기 때문에 안전성·환경 영향 평가가 필요하며, AI 기반 판독 시스템은 알고리즘의 편향성·설명가능성 확보가 필수적이다. 따라서 신속한 인허가를 지원하면서도 안전성·신뢰성을 확보할 수 있는 규제 체계가 필요하다.

마지막으로 미래 전망 측면에서 합성생물학 기반 진단 기술은 단순 진단을 넘어 지능형 헬스케어 플랫폼, 웨어러블·삽입형(in-body) 바이오센서, 자율형 로봇 진단 시스템, 환경 기반 스마트 모니터링 인프라로 확장될 가능성이 크다. 특히 AI가 진단 회로를 자동 설계하고, 3D 프린팅이 센서를 제작하며, CFPS 회로가 현장에서 즉시 가동되는 생태계가 구축되면, 향후 국가 보건·안보·환경 정책의 핵심 기반이 될 것으로 예상된다.

9 맺음말

감염병, 암, 항생제 내성 등 복합적인 보건 위협이 빠르게 증가하는 시대에, 기존 PCR·면역검사 중심의 진단 기술만으로는 국가적·전 지구적 대응에 한계가 뚜렷하다. 본 보고서에서 살펴본 합성생물학 기반 진단 플랫폼은 생물학적 회로를 공학적으로 재구성하여, 분자 수준에서 정보를 감지·변환·처리하는 새로운 형태의 지능형 센서를 구현한다. 특히 Toehold·Riboswitch·TF 기반 센서, CRISPR-Cas 시스템, Cell-Free 시스템 등 다양한 플랫폼은 기존 분자진단의 병목을 극복하고, 신속성·다중성·저비용·현장성 등 차세대 진단이 요구하는 핵심 특성을 충족한다.

여기에 인공지능 기술이 결합되면서 합성생물학 진단은 단순 감지 수준을 넘어, 회로 설계 자동화, 성능 예측, 피드백 기반 최적화, 생성형 설계 등 지능형 진단 시스템으로 도약하고 있다. AI 모델은 서열-기능 관계를 직접 학습함으로써 기존 이론적 설계가 가진 예측 한계를 보완하고, VIS4Map과 같은 설명가능 인공지능(XAI) 기법은 회로 실패 원인을 구조적으로 해석할 수 있게 한다. 더 나아가 클라우드 랩 기반의 자동화 DBTL 시스템은 수천~수만 개의 회로 조합을 실시간으로 평가하여 최적 설계를 도출함으로써 기술 개발 속도를 획기적으로 향상시킨다.

이러한 기술적 진보는 감염병, 암, AMR, 식품·환경 모니터링, 법과학 등 실질적 사회 현장에서 이미 응용 가능성을 입증하고 있다. 특히 합성생물학 진단의 “저비용·즉시성·고정확도” 특성은 국가 보건안보 시스템 구축에 필수적이며, 의료 접근성이 낮은 취약 지역에서의 POC 진단, 원격·자가 진단, 대량 감염 대응 등 다양한 정책적 요구를 충족한다. 산업적으로는 3D 프린팅 기반 POC 장비, CFPS 기반 웨어러블 진단, AI-체인오브커스터디 기반 법과학 플랫폼 등 새로운 시장 분야를 창출하며 고부가가치 산업 생태계를 형성할 것으로 기대된다.

결국 합성생물학과 AI가 결합된 차세대 진단 기술은 단순한 기술 발전을 넘어, 건강·안보·환경·산업 전반의 구조를 재편할 수 있는 전략 기술로 자리 잡을 것이다. 본 보고서가 이러한 변화의 흐름을 조망하고, 향후 국가 차원의 기술 개발 전략 및 연구·산업 로드맵 수립에 기여하기를 기대한다.

[참고문헌]

- [1] Zhang, C., Liu, H., Li, X., Xu, F., & Li, Z. (2023). Modularized synthetic biology enabled intelligent biosensors. *Trends in Biotechnology*, 41(8), 1055-1065.
- [2] Angenent-Mari, N. M. (2021). Designing Next Generation Nucleic Acid Diagnostics Using Synthetic Biology and Artificial Intelligence (Doctoral dissertation, Massachusetts Institute of Technology).
- [3] Angenent-Mari, N. M., Garruss, A. S., Soenksen, L. R., Church, G., & Collins, J. J. (2019). Deep learning for RNA synthetic biology. *Biorxiv*, 872077.
- [4] Jiang, L., Zhao, J., Lian, J., & Xu, Z. (2018). Cell-free protein synthesis enabled rapid prototyping for metabolic engineering and synthetic biology. *Synthetic and Systems Biotechnology*, 3(2), 90-96.
- [5] Jones, K. A., Zinkus-Boltz, J., & Dickinson, B. C. (2019). Recent advances in developing and applying biosensors for synthetic biology. *Nano Futures*, 3(4), 042002.
- [6] Hicks, M., Bachmann, T. T., & Wang, B. (2020). Synthetic biology enables programmable cell - based biosensors. *ChemPhysChem*, 21(2), 132-144.
- [7] Chen, S., Chen, X., Su, H., Guo, M., & Liu, H. (2023). Advances in synthetic-biology-based whole-cell biosensors: principles, genetic modules, and applications in food safety. *International journal of molecular sciences*, 24(9), 7989.
- [8] Gallup, O., Ming, H., & Ellis, T. (2021). Ten future challenges for synthetic biology. *Engineering Biology*, 5(3), 51-59.
- [9] Yarra, S. S., Ashok, G., & Mohan, U. (2023). Toehold switches; a foothold for synthetic biology. *Biotechnology and Bioengineering*, 120(4), 932-952.
- [10] Dong, X., Qi, S., Khan, I. M., Sun, Y., Zhang, Y., & Wang, Z. (2023). Advances in riboswitch - based biosensor as food samples detection tool. *Comprehensive Reviews in Food Science and Food Safety*, 22(1), 451-472.
- [11] Ding, N., Zhou, S., & Deng, Y. (2021). Transcription-factor-based biosensor engineering for applications in synthetic biology. *ACS Synthetic Biology*, 10(5), 911-922.

- [12] Xu, X., & Qi, L. S. (2019). A CRISPR–dCas toolbox for genetic engineering and synthetic biology. *Journal of molecular biology*, 431(1), 34-47.
- [13] Hodgman, C. E., & Jewett, M. C. (2012). Cell-free synthetic biology: thinking outside the cell. *Metabolic engineering*, 14(3), 261-269.
- [14] García Martín, H., Mazurenko, S., & Zhao, H. (2024). Special issue on artificial intelligence for synthetic biology. *ACS synthetic biology*, 13(2), 408-410.
- [15] Iram, A., Dong, Y., & Ignea, C. (2024). Synthetic biology advances towards a bio-based society in the era of artificial intelligence. *Current Opinion in Biotechnology*, 87, 103143.
- [16] Afroz, T., & Beisel, C. L. (2013). Understanding and exploiting feedback in synthetic biology. *Chemical Engineering Science*, 103, 79-90.
- [17] Nshogozabahizi, J. C., Aubrey, K. L., Ross, J. A., & Thakor, N. (2019). Applications and limitations of regulatory RNA elements in synthetic biology and biotechnology. *Journal of Applied Microbiology*, 127(4), 968-984.
- [18] Garruss, A. S. (2021). Accelerating the Understanding and Design of Intracellular Biosensors by Massively Multiplexed Experimentation and Machine Learning. Harvard University.
- [19] Jessop-Fabre, M. M., & Sonnenschein, N. (2019). Improving reproducibility in synthetic biology. *Frontiers in bioengineering and biotechnology*, 7, 18.
- [20] Khalil, A. S., & Collins, J. J. (2010). Synthetic biology: applications come of age. *Nature Reviews Genetics*, 11(5), 367-379.
- [21] Weber, W., & Fussenegger, M. (2012). Emerging biomedical applications of synthetic biology. *Nature Reviews Genetics*, 13(1), 21-35.
- [22] Callura, J. M., Cantor, C. R., & Collins, J. J. (2012). Genetic switchboard for synthetic biology applications. *Proceedings of the National Academy of Sciences*, 109(15), 5850-5855.
- [23] Tyagi, A., Kumar, A., Aparna, S. V., Mallappa, R. H., Grover, S., & Batish, V. K. (2016). Synthetic biology: applications in the food sector. *Critical Reviews in Food Science and Nutrition*, 56(11), 1777-1789.

- [24] Yan, X., Liu, X., Zhao, C., & Chen, G. Q. (2023). Applications of synthetic biology in medical and pharmaceutical fields. *Signal transduction and targeted therapy*, 8(1), 199.
- [25] Angenent-Mari, N. M., Garruss, A. S., Soenksen, L. R., Church, G., & Collins, J. J. (2020). A deep learning approach to programmable RNA switches. *Nature communications*, 11(1), 5057.
- [26] Taha, B. A., Ahmed, N. M., Talreja, R. K., Haider, A. J., Al Mashhadany, Y., Al-Jubouri, Q., ... & Arsad, N. (2024). Synergizing nanomaterials and artificial intelligence in advanced optical biosensors for precision antimicrobial resistance diagnosis. *ACS Synthetic Biology*, 13(6), 1600-1620.
- [27] Scognamiglio, V., Antonacci, A., Lambrea, M. D., Litescu, S. C., & Rea, G. (2015). Synthetic biology and biomimetic chemistry as converging technologies fostering a new generation of smart biosensors. *Biosensors and Bioelectronics*, 74, 1076-1086.
- [28] Zhou, C., Jiang, F., Chen, W., Nugen, S. R., & Huang, C. (2025). Synthetic biology meets diagnostics: Engineering biosensing platforms for rapid and accurate pathogen and viral detection. *Biosensors and Bioelectronics*, 290, 117946.

[주요용어풀이]

- Deep neural network (DNN): 입력층(input layer)과 출력층(output layer) 사이에 다종의 은닉층(hidden layer)을 포함하는 인공 신경망(ANN: Artificial Neural Network)
- eXplainable Artificial Intelligence (XAI): 인공지능(AI)의 판단 결과에 대한 이유를 인간이 이해할 수 있도록 제공하는 기술.
- Modularity: 시스템의 컴포넌트가 분리되거나 재연결되는 정도
- Scalability: 다양한 영상 크기, 프레임률, 화질 등의 확장성을 사용자 수의 증대에 유연하게 대응할 수 있는 정도



합성생물학을 위한 자동화 및 로봇릭스 기술동향



한국합성생물학발전협의회
Korea Synthetic Biology Association

임현의

융합분과위원, 한국기계연구원 바이오파운드리 DX/AI 연구단 단장

이대식

융합분과위원, 한국전자통신연구원 책임연구원

김진환

융합분과위원, 케이랩 주식회사 연구소장

이석재

융합분과위원, 나노종합기술원 나노융합기술개발본부 본부장

합성생물학을 위한 자동화/지능화 발전 방향

1 머리말

전통적으로 이루어지던 바이오 연구에서 인공지능(Artificial Intelligence)의 등장은 바이오연구의 혁신을 불러일으키고 있으며, 자동화와 로봇의 접목을 기반으로 하는 자율 제조는 바이오산업의 변화를 예측케 한다. 바이오자율제조 of 가장 대표적인 플랫폼은 바이오파운드리라고 말할 수 있다. 특히 합성생물학의 연구개발과정은 [설계(Design)-제작(Build)-검증(Test)-학습(Learn)] 순환공정으로 이루어지며, 바이오파운드리(Biofoundry)로 불리는 분석, 배양, 추출, 분리 등의 바이오 공정 기기가 집합된 인프라를 통해 효율적으로 구현될 수 있다. 따라서, 미래 유망기술로서 합성생물학은 부품화·표준화·모듈화라는 공학적 개념을 생물학에 도입하여 생물 시스템의 설계·제작 기간을 비약적으로 단축하여 레드, 그린, 화이트 바이오 전 영역에 혁신을 가져오는 중이다. 본 동향보고서에서는 합성생물학을 위한 자동화/지능화에 관련된 기업들의 연구동향을 살펴보고, 이에 대응해야 하는 우리의 역할과 방안을 생각해 보고자 한다. 더불어 합성생물학 실험의 자동화/지능화를 위한 학계의 연구동향은 한국전자통신연구원 이대식박사님, 분석기술동향은 케이랩의 김진환연구소장님, 그리고 웨이퍼/칩기반으로의 발전방향은 나노종합기술원의 이석재박사님이 소개해 주셨다.

2 합성생물학을 위한 바이오파운드리

2.1 개념

바이오파운드리는 Engineering Biology의 4대 요소인 DBTL(Design-Build-Test-Learn)은 유전자 설계에, 대량의 유전자 테스트-학습결과가 피드백 되도록 하는 시스템으로, 합성생물학 전체 과정에 인공지능 및 로봇 기술을 도입하여 바이오 연구개발 프로세스를 고속화, 고효율화하는 플랫폼 기술이다^[1]. 즉, 바이오파운드리는 합성생물학 응용 및 연구 활성화를 위해 구축한 인공유전체 및 유전자회로의 신속한 설계, 합성, 위탁제조 및 성능시험을 위한 기술 및 시설로써 합성생물학과 빅데이터, AI, 자동화 로봇 등 첨단기술과의 융합으로 불확실성이 높고 생산성이 낮았던 합성생물학의 한계를 극복하고, 바이

오 제조 공정의 속도, 규모, 생산효율 향상이 가능하게 만든다. 바이오파운드리 구현을 위해서는 명확한 대상과 DBTL (Design-Build-Test-Learn) 사이클의 구성 기술이 필요하다. 또한 바이오파운드리에는 독립적인 장비들의 나열이 아닌 유기적으로 연계된 하나의 시스템으로서 높은 효율 구현이 가능하므로 DNA 설계부터 해당 DNA가 도입된 세포 표현형 테스트에 이르는 고도화된 일련의 생물실험과 관련 장비들이 연계되어 순환적 엔지니어링 구현이 가능해진다^[2].

설계(Design)는 소프트웨어 기반 접근 방식으로, 기존 DB화된 유전체 정보를 바탕으로 합성경로 파악, 유전자 합성 설계 시퀀스 등을 포함한 모든 수준의 디자인을 포함한다. DNA, RNA부품(promoter, terminator, rbs, regulator 등), 단백질펩티드효소부품, 조절회로, 대사경로, 인공세포 등 모든 형태의 DNA서열 예측을 통한 설계를 포함하는 바이오정보의 설계화 단계이다.

제작(Build)은 설계 단계에서 최적의 합성 조건을 갖춘 유전체 회로도를 기반으로 실제 유전자를 합성하는 과정으로, 대량의 올리고 DNA를 합성, 조립하여 설계를 통해 추구하는 모든 형태의 실물을 제조하는 바이오정보의 실물화 단계이다.

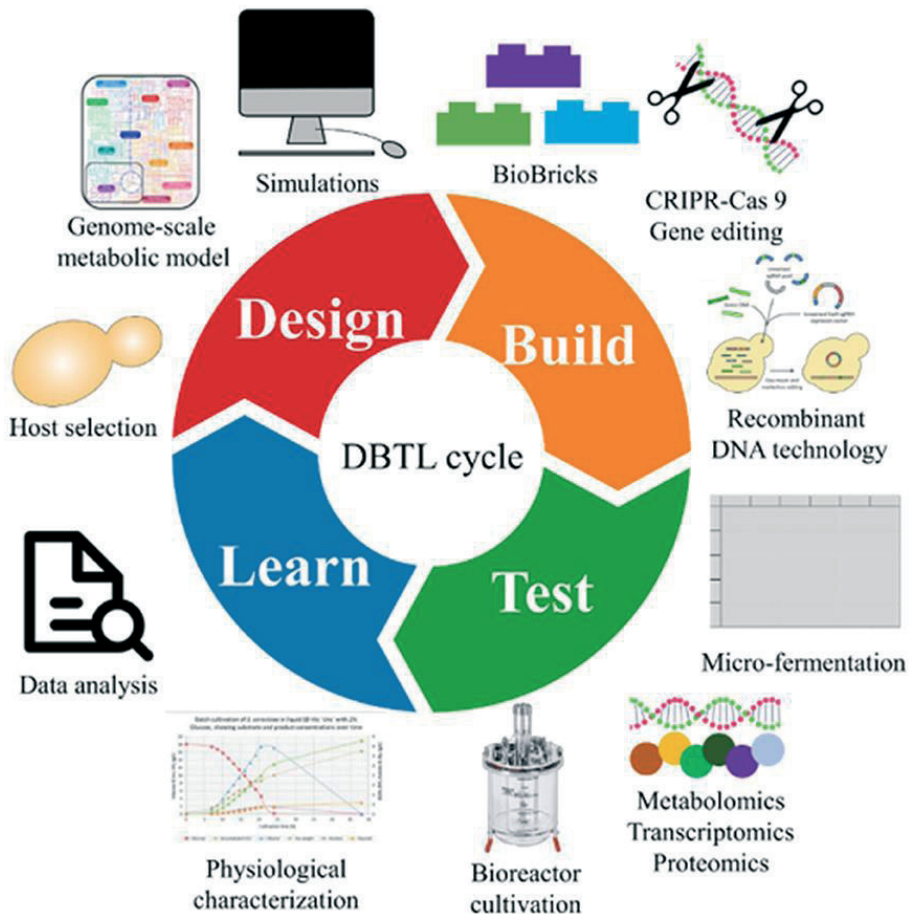
검증(Test)은 합성된 구성 및 균주의 테스트. 일반적으로 가상 셀(cell) 상태에서 실시해서 통과되었던 테스트와 같은 동일한 조건(온도, 압력 등)에서 대규모의 스크리닝 테스트를 실시하고, 효소단백질 활성, 상호작용, 세포기능 분석 등 모든 실물의 표현형(phenotype)에 대한 대량 정보를 확보하는 실물 바이오정보의 검증화 단계이다.

학습(Learn)은 선행 단계 수행과정, 결과를 통계적 방법과 머신러닝을 사용한 예측 모델로 분석한 후, 다음 DBTL 사이클을 위한 Design 단계에서 개선해야 할 부분을 도출하고, 유전자정보(genotype)와 표현형정보(phenotype)의 연관성을 분석하여 DBTL사이클의 예측력, 정밀제조를 가능하게 하는 실물 바이오정보의 최적화 단계라 할 수 있다.

표 1> 바이오 파운드리의 4대 핵심 과정, D-B-T-L (Design-Build-Test-Learn)

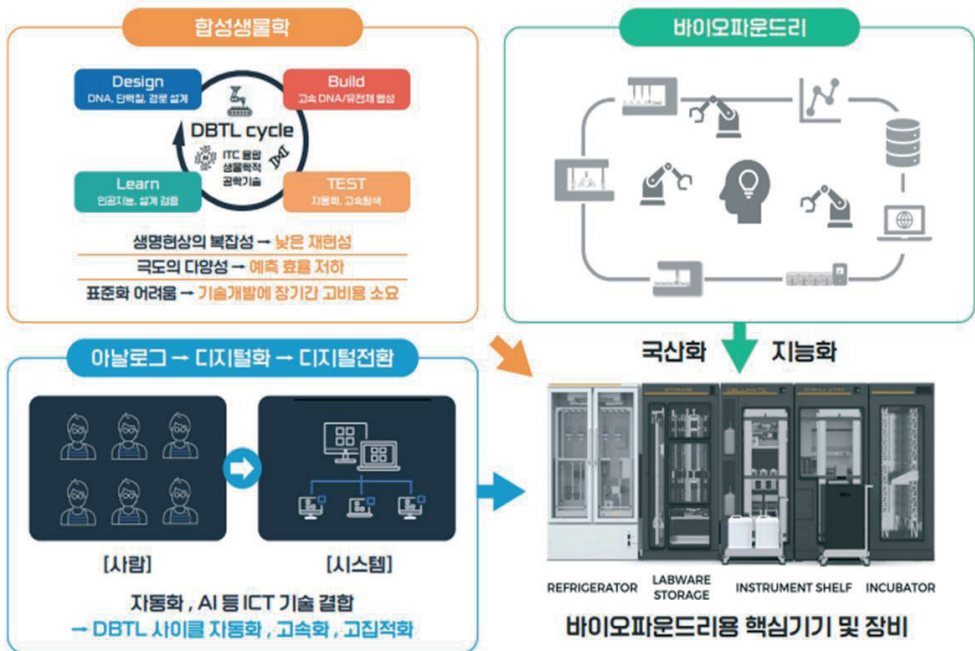
4대 핵심 과정	내용
Design	목표 기능을 정의하고 유전체 및 대사경로를 설계
Build	자동화된 장비를 통한 DNA합성
Test	다양한 자동화장비를 이용한 유전체(자) 성능분석
Learn	DNA서열과 빅데이터(Test)를 이용한 모델링 및 인공지능 도구를 이용한 학습단계

그림 1> Engineering Biology의 Design-Built-Test-Learn 순환 개념^[3]



더불어 공공영역에서의 바이오파운드리 기술은 오픈이노베이션 형태로 다수의 기업이 활용할 수 있는 컨테이너(Container)와 클라우드(Cloud) 기술 등이 있다. Docker로 알려진 컨테이너 기술은 표준화된 컴퓨터 운영환경으로 볼 수 있으며 이를 통해서 각종 하드웨어 의존적인 설정이나 드라이버 응용소프트웨어 설치를 걱정하지 않고 컨테이너 이미지를 내려받는 것으로 원하는 컴퓨팅 환경을 구성할 수 있다. 이러한 기술들이 바이오파운드리에 접목되면서 생산성과 접근성이 향상되고 표준화된 데이터가 대량으로 생산되면서 바이오에서 활용 확장성이 더욱 높아질 수 있다. 미국, 영국 등 해외 주요국에서는 바이오파운드리의 완전 자동화, 빅데이터 및 데이터 분석 활용, 클라우드 플랫폼 도입 등 바이오 및 디지털 첨단 제조 기술 융합을 적극적으로 추진 중으로 바이오연구의 오랜 난제인 속도와 스케일, 불확실성의 한계를 극복하려고 바이오분야에서의 디지털트랜스포메이션 가속화가 활발히 진행 중이다.^[4]

그림 2> 바이오파운드리 디지털트랜스포메이션 구축 개념



바이오파운드리 핵심기기 및 장비는 생명공학 기술이 접목되는 만큼 요구 성능이 엄격하여, 장비 품목별로 글로벌 브랜드 제품이 독과점하는 특성을 보인다. 바이오파운드리 핵심기기 및 장비 산업은 최근 5년간 연평균증가율 33.1%로 빠르게 시장이 확대되고 있으나, 국내기업의 매출은 '21년도 1,910억 원으로 바이오 산업 전체 매출 21조 원의 0.9% 수준이며, 심각하게 글로벌 회사에 의존하고 있어 경제적 측면은 물론 기술 종속과 유출도 우려되는 상황이다. 바이오파운드리를 통해 직접적인 산업화가 진행될 신약, 고기능성 바이오소재, 바이오화학소재, 바이오에너지, 진단기기 등의 제조 생산성, 기능성, 효율성, 품질 등을 향상하기 위해서는 높은 수준의 기기 지능화, 연속공정화 등의 고도화된 장비 기술의 발전이 필연적이다. 바이오파운드리 파이프라인에 사용될 핵심 기기의 국산화 및 공정 지능화, 연속화 등의 고도화 기술 개발을 통해 합성생물학의 연구개발 및 사업화 속도를 제고하고 국내 바이오 장비 및 부품 기업들의 생태계 조성 및 혁신 건인이 시급함을 알 수 있다^[5].

표 2> 국내 바이오장비 및 기기 산업 규모 및 실적 (단위: 억 원)

구분	2017	2018	2019	2020	2021	CAGR(%)
내수 시장 (바이오 산업 내 비중, %)	1,174 (1.8%)	1,240 (1.7%)	2,455 (3.0%)	3,334 (3.5%)	3,685 (2.6%)	33.1
매출액 (바이오 산업 내 비중, %)	1,130 (1.1%)	889 (0.8%)	1,105 (0.9%)	1,721 (1.0%)	1,910 (0.9%)	14.0
수출 (바이오 산업 내 비중, %)	469 (0.9%)	305 (0.6%)	405 (0.6%)	477 (0.5%)	506 (0.4%)	1.9
수입 (바이오 산업 내 비중, %)	514 (3.3%)	655 (3.8%)	1,754 (8.5%)	2,089 (8.6%)	2,282 (4.8%)	45.2
무역수지	△45	△350	△1,349	△1,612	△1,776	-

※ 출처 : 산업통상자원부, 한국바이오협회, 2021년 기준 국내 바이오산업 실태조사 결과보서, 2022.

2.2 기업동향

미국, 영국, 중국, 일본 등 해외 주요 선진국에서는 바이오산업 혁신을 가속하기 위해 합성생물학연구를 위한 공공 바이오파운드리를 설립하고 운영 중이다. 해외 주요 공공 바이오파운드리에서는 핵심기기 및 장비의 디지털트랜스포메이션 및 자동화/연속화에 많은 투자 중이며, 미국에 있는 다수의 민간기업이 바이오파운드리를 구축하여 단기간에 사업화 성과를 창출함으로써 전 세계적인 주목을 받고 있다. 가장 대표적인 미국 Amyris는 효모를 이용한 합성생물학 전문기업으로 유전자변형 효모를 이용한 화장품이나 향료(flavor), 고분자 소재, 윤활유, 연료 등으로 상업화를 성공하였다. 원래 효모가 갖지 않은 대사경로를 가진 유전자변형 효모를 제조하고, 이를 통해 항-말라리아제로 이용되는 썩(mugwort) 속 식물유래 아르테미시닌산, 공업 용도로 이용되는 파네센, 케모마일 유래 비사보롤(정유 등에 포함되는 성분) 등 15개 제품을 개발하였으며, 2017년 기준 매출액 1억 4300만 달러로 발표되었다. 미국 Moernerna Therapeutics는 2018년 생명공학 역사상 가장 높은 6억 4천만 달러의 IPO를 발표한 회사로, mRNA기반 기술 플랫폼을 활용하여 신종 코로나 바이러스 백신 개발에서 최선두 주자이다. 미국의 Twist Bioscience는 2013년 설립된 DNA를 합성하는 생명공학 회사로, 포토리소그래피(Photolithgraphy) 기반의 마이크로 플레이트 칩을 기반으로 특정 DNA를 대량생산하는 기술을 보유하고 있어 성장세가 가파르다.

최근에는 로봇자동화, AI, 클라우드 등의 디지털트랜스포메이션 개념이 도입된 바이오파운드리를 구축한 기업들이 성공적으로 시장에 정착 중인데, 미국 보스턴의 Ginkgo Bioworks는 2009년 MIT 출신 인력들이 공동 설립한 회사로, 2013년 80여개의 실험실 업무를 자동화하고 소프트웨어로 통합한 첫번째 파운드리를 설립하였다. 자체 제품 개발 보다는 충분한 실험설비를 갖추지 못한 소규모 스타트업들의 연구개발을 지원하여 유망

후보물질을 발굴하는 전략을 가지고 있으며 2014년 이후 총 7억 2,600만 달러(약 8,850억 원)의 투자를 유치해왔으며, 가장 최근에는 2019년 9월에 2억 9000만 달러 투자를 유치해, 기업 가치 최대 42억 달러를 기록 중이다. 미국 캘리포니아의 Zymergen는 수천 개의 실험 조건을 병렬 테스트하고 결과를 자동 분석하는 공정 자동화 및 머신 러닝 기술이 도입된 바이오파운드리를 설립하였다. 최근 유기체 발효 공정을 통해 제조 가능한 연성 전자소재 히알린 필름을 개발하여 주목받았고, 2014년 이후 실시된 4차례의 자금 조달 라운드에서 5억 7,400만 달러 조달받았다.

표 3> 국내외 주요 민간 바이오파운드리 기업 현황

국가	바이오파운드리	주요 활동 내용
미국	Amyris	- 아르테미시닌산을 비롯하여 7년간 15개 물질 상용화, NASDAQ 상장
미국	Twist Bioscience	- 합성DNA 최대 공급 기업, NASDAQ 상장
미국	Zymergen	- 바이오파운드리 기술로 히알린 필름 개발
미국	Emerald Cloud Lab	- 바이오 연구실 가상화 기술 기반의 스타트업
미국	Ginkgo Bioworks	- DNA 부품 기반 프로세스에 자동화 로봇/자동화 시스템 도입
미국	Transcriptic	- 클라우드 기반 바이오 연구실 스타트업
미국	Viridos	- 엑손 모빌과 협력 화석연료 대체 목적의 미세 조류 및 광합성 연구
미국	Intrexon	- 말라리아 모기 퇴치 기술 연구
미국	Inscripta	- 디지털 유전공학 기반의 바이오파운드리 기업
미국	Berkeley Lights	- 싱글 Cell 분석 기술 기반의 기업
미국	Moderna	- RNA기반 코로나19 백신 개발
영국	Labman	- 맞춤형 연구실 자동화 시스템 기업
영국	Astechproject	- 로봇 기반의 자동화 시스템 기업
일본	Robotic Biology Consortium	- 로봇 기반의 자동화 시스템 및 데이터 관리 시스템 기업
국내	추진 예정	- 국내의 합성생물학 기반 바이오파운드리 인프라 구축의 경우 이제 겨우 태동기로, 산업계 참여는 지극히 제한적인 실정 - 초고속 DNA설계-합성에 필요한 바이오파운드리 공정개발 지원 및 시험시설은 개별 기관 기업 차원에서 소규모로 운영 중 - 최근 삼성 바이오, CJ제일제당 LG화학, SK 등 일부 국내 기업들이 바이오플라스틱 관련 기술 확보 및 제품 개발에 적극적으로 투자

바이오장비 글로벌시장을 선도하는 대표적인 기업은 Thermo Fisher Scientific, Danaher, Agilent Technologies, Waters, Shimadzu 등으로 5개 사의 시장 점유율은 약 30% 수준으로 추정되며^[6], 선도기업 대부분 매우 다양한 제품군을 보유하고 있다.

그림 3> 글로벌 바이오장비 기업의 제품군^[6]

Sr. No.	Company	Spectroscopy	Chromatography	Polymerase Chain Reaction (PCR)	Liquid Handling Systems	Microscopy	Flow Cytometry	Next-Generation Sequencing (NGS)	Centrifuges	Electrophoresis	Cell Counting	Clinical Chemistry Analyzers	Immunassays	Other Technologies
1	Thermo Fisher Scientific Inc.													
2	Danaher Corporation													
3	Agilent Technologies, Inc.													
4	Waters Corporation													
5	Shimadzu Corporation													
6	PerkinElmer Inc.													
7	Bio-Rad laboratories, Inc.													
8	Bruker													
9	Becton, Dickinson and Company													
10	QIAGEN N.V.													

또 한 외국의 글로벌 장비 기업들은 적극적인 R&D, 협력, M&A를 통해 경쟁력을 확대·유지하고 있는 것으로 파악된다.

표 4> 바이오장비 주요기업 최근 활동^[7]

기업	내용
Thermo Fisher Scientific	▷Binding Site Group 인수(23.1., 28억 달러)로 진단 기술 역량 확대 ▷Celltrio와 전자동 세포 배양 시스템 도입 협력(23.2.) ▷Covid-19 백신 제조를 위해 Moderna와 협력(22.2.) ▷재조합 단백질 제조 기업 PeproTech 인수(22.1., 18.5억 달러)
Danaher	▷Duke대와 유전자 치료를 위한 Danaher Beacon 이니셔티브 협력(22.11.) ▷Aldveron 인수(21.8., 96억 달러)로 mRNA 관련 역량 강화 ▷SVision의 AI 소프트웨어 자산 인수(21.3.)로 광디지탈 솔루션 분야 강화 ▷로봇 무균 충전기 기업 Vanrx Pharmsystems 인수(21.2.)
Agilent Technologies	▷NGS 관련 Avida Biomed 인수(23.1.) ▷샘플 전처리 자동화 관련 이슈에 대해 Mettler Toledo와 협력(22.9.) ▷크로마토그래피 분야 Polymer Standards Service 인수(22.8.) ▷APC와 액체 크로마토그래피 자동 분석 솔루션 협력(22.5.)
Waters	▷ASEAN 바이오분석 아카데미 설립을 위해 한국의 BioInfra와 협력(22.6.) ▷Megadalton Solutions의 CDMS 기술 관련 기술자산 및 지재권 인수(22.2.) ▷Sartorius의 바이오리액터에 Waters 질량분석기를 적용하는 협력(21.10.)
Shimadzu	▷Nissui Pharmaceutical 인수(22.11.)로 임상관련 역량 강화 ▷TetraScience의 R&D 데이터 클라우드를 활용하는 파트너십 체결(22.4.) ▷Total SE, UPPA, Oviedo대와 산화물 분석 시스템 개발 연구 협력(21.1.)
Sartorius	▷세포치료 및 유전자치료 관련 역량 강화를 위해 Polyplus 인수(23.3.) ▷바이오프로세스 관련 역량 강화를 위해 WaterSep BioSeparation 인수(20.12.) ▷재조합 알부민 기반 솔루션 확보를 위해 Alubmedix 인수(22.8.)
Becton, Dickinson and Company	▷Parata Systems 인수(22.7.)로 혁신적 제약 자동화 솔루션 개발(22.7.) ▷Labcorp와 flow cytometry 기반 동반진단키트 개발 및 상용화 협력(22.8.)

Perkinelmer	▷Scitara와 사용자의 작업 유연성을 높이는 인포매틱스 솔루션 협력(22.4.) ▷항체 및 시약 기업 BioLegend 인수(21.9.) ▷바이러스 벡터 기술 기업 Sirion Biotech 인수(21.6.)
Bio-Rad Laboratories	▷헬스케어 및 진단 기업 Curiosity Diagnostics 인수(22.8.) ▷분자진단 기업인 한국의 씨젠과 미국 시장에서의 파트너십 체결(21.6.)b
Bruker	▷Acquifer Imaging 인수(23.1.)로 첨단 형광 현미경 이미징 제품군 확대 ▷뇌과학 연구 강화를 위해 Neurescence 인수(22.12.) ▷미니스코프 기업 Incscopix 인수(22.11.)
Qiagen	▷NGS 관련 디지털 분석 플랫폼을 위한 Sophia Genetics와 파트너십 체결(23.3.) ▷미국 시장에서의 동반진단 서비스의 파트너로 Helix와 협력(23.1.) ▷재조합 효소 제조기업 Blirt 지분 인수(22.5.)
Eppendorf	▷고급 원심분리기 선도가업이 되기 위해 Koki의 원심분리기 사업 인수(20.3.) ▷상하이에 신규 제조 사이트 오픈(23.2.)
Hitachi High Technologies	▷중국 내 분석 시스템 판매 확대를 위해 Techcomp Scientific 지분 인수(20.5.)
Horiba	▷Green Tropism과 협력으로 자동 데이터 해석 역량 강화(21.2.)

3 바이오실험의 자동화/지능화

3.1 하드웨어(로봇 기반 통합 시스템)

바이오파운드리 핵심기기 및 장비 선정을 위해 합성 생물학을 위한 워크플로에 사용되는 Build와 Test에 사용하는 대표적인 장비들을 모두 조사하면 다음과 같다.

바이오파운드리 핵심기기 및 장비관련하여 특허 현황을 살펴보면 대부분의 장비들이 오래전부터 사용해온 장비들이임에도 불구하고, 대체적으로 2000년대부터 활발한 특허출원이 시작하였고, 한국·미국·중국·일본·유럽에서 전체적으로 증가하고 있는 것으로 확인된다.

이러한 다양한 장비들을 연결하여 자동화/지능화를 구현하는 디지털트랜스포메이션을 위해서 2011년 독일에서 제시된 개념인, 인더스트리 4.0을 다시 되집어 보게 된다. 전체 제조 공정과 공장·창고관리·물류에 이르기까지 공급망을 구성하는 모든 요소의 디지털화·스마트화를 의미하는 인더스트리 4.0은 요즘 인공지능의 발전에 힘입어 자율제조(Autonomous Manufacturing)의 개념으로 발전하고 있다. 이러한 디지털트랜스포메이션은 9가지 요소 기술로 나눌 수 있으며, 제조 공정의 속도·유연성·품질·효율성을 향상시켜 제품 경쟁력을 비약적으로 향상하는 것이 목표이다.

그림 4> D-B-T-L 과정에서의 필수 이용 장비 개략도

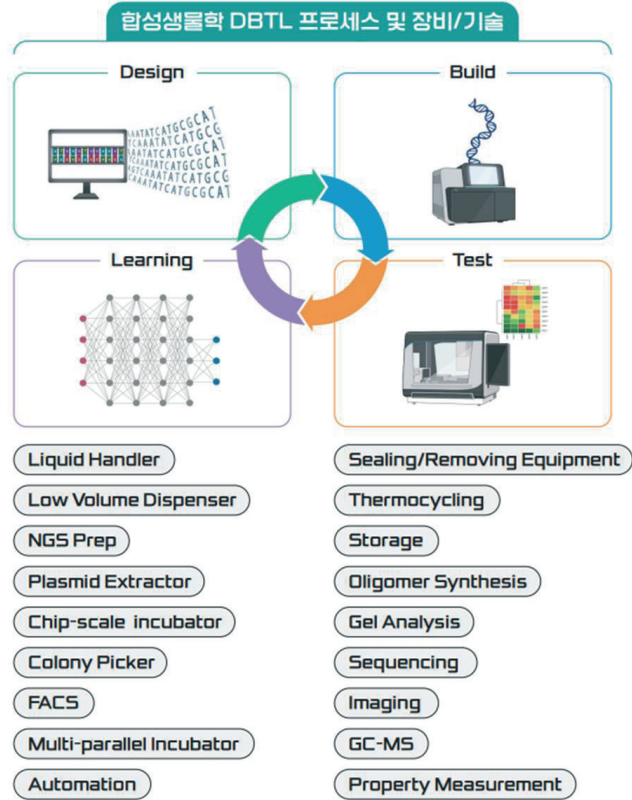
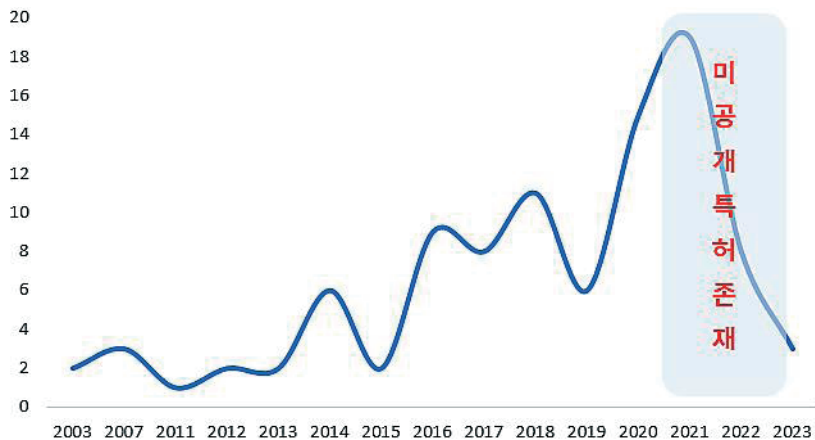


표 5> ‘인더스트리 4.0’의 9가지 구성요소^[8]

요소	내용
수평적·수직적 시스템 통합	- 조직 내 부서간 뿐만 아니라, 가치체인(value chain) 파트너 간 협력 강화 필요
사물인터넷 (IoT)	- 완전한 자동생산 시스템 구축을 위해 생산부품과 기계 간 내장된 센서와 장치를 통해 데이터 송·수신 등 실시간 소통이 필수
사이버보안	- 수십억 개의 장치와 커뮤니케이션 채널이 연계된 새로운 프레임워크에서 신뢰 구축을 위한 사이버보안 기술이 필수
클라우드 (Cloud)	- 데이터를 생성하는 다수의 장치와 센서를 지원하기 위해 실시간 서비스와 확장성(scalability)을 제공하는 클라우드 컴퓨팅 기술이 필수적으로, ERP 시스템 등 기존의 산업 모니터링과 컨트롤 시스템이 향후 클라우드로 이전될 전망
빅데이터 분석	- 제품개발, 시제품 검증 등 전 제조 과정에서 빅데이터 분석을 적용하여 목표지향적인 혁신, 마케팅 및 의사결정을 지원해야 함
시뮬레이션 (Simulation)	- 시뮬레이션을 통한 제품 및 제조 공정의 가상 모델링으로 제품과 공정의 신속한 검증을 실현하여 비용 및 시간을 절감
적층 제조법 (3D 프린팅)	- 적층 제조법은 맞춤형 제품을 생산하기 위해 널리 사용될 전망이며, 적층제조 시스템이 분산화로 운송거리와 재고품을 절감할 수 있음
증강현실 (Augmented reality)	- 증강현실 글래스 등과 같은 장치를 통해 현장 근로자에게 직접 근무지침을 지지하거나 가상의 작업훈련에 활용하여, 의사결정 내용의 신속한 전달과 생산성 제고에 중요한 역할을 할 전망
로봇 (Robots)	- 로봇은 현재 생산조립 라인에서 로봇 암스(arms)의 형태로 사용되고 있으며, 앞으로도 더욱 정교한 과정을 수행하며 널리 활용될 전망

디지털트랜스포메이션은 융합된 개념을 가지고 있어 바이오파운드리를 위한 연결화·자동화해주는 시스템 또는 로봇·장비에 대한 특허동향을 조사해 보면 다른 기술들에 비해 출원 건수가 천천히 증가하고 있는 모습을 보이고 있고, 아직까지는 다른 기술들보다 실질적인 연구활동이 미흡한 것으로 확인된다. 우리나라는 2012년부터 출원을 시작하였으나, 전체적으로 거의 출원 건수가 없는 것으로 보아 보유한 기술은 거의 없는 것으로 파악되고, 주요국과 다수 기업들이 동시다발적으로 연구개발을 실시하고 있으며, 대부분이 아직 연구개발 단계에 머무르고 있는 것으로 추측이 된다. 따라서, 바이오파운드리 장비/자동화/연결화 관련한 기술은 다수 기업들이 적은 출원건수를 보유하고 있는 것으로 보아 태동기에서 성장기로 넘어가는 단계로 파악되고 그 외 장비들과 결합하여 개발을 하게 된다면 신기술로써 폭발적인 기술성장을 이끌 수 있을 것으로 판단된다.

그림 6> 바이오파운드리 장비/자동화/연결화 연도별 특허출원 현황



3.2 소프트웨어(소프트웨어 및 AI 모델 및 응용 기술)

바이오파운드리 각 핵심장비들을 네트워크 및 로봇으로 온/오프라인 전영역에서 연결하여 자동화 및 연속화를 구현하고, 지능형 공정 최적화를 포함하는 클라우드 기반 바이오파운드리 통합 관제 플랫폼 구축할 수 있는 소프트웨어 개발도 매우 중요하다. 바이오파운드리 핵심기기 및 장비를 물리적, 네트워크로 연결하여 합성생물학 공정의 자동화 및 연속화를 구현하며, 클라우드 기반의 지능형 통합 관제 시스템을 통해 바이오파운드리의 공정을 지능적으로 최적 관리하는 합성생물학 공정의 디지털 대전환 수행은 필연적이다. 바이오파운드리 장비간 물리적 연결은 바이오파운드리 공정의 자동화와 공정의 연속화를 위하여 핵심장비 간 용기의 이송을 수행하는 로봇을 통해 실현하여야 한다. 또한 바이

오파운드리 of 모든 장비는 네트워크로 연결되어 각각의 상태를 공유하며 작업을 통합 관리하여야 하며, 실험 공정을 스케줄링 하여 장비의 가동률 등 운용상황을 효율적으로 관리할 수 있도록 하는 지능적 통합 관제 시스템이 구현되어야 한다. 실험에 따른 공정에서의 활용 장비, 활용 순서, 장비당 예상 운용 시간 및 용량 등을 복합적으로 분석하여, 복수의 실험을 수행하는 경우에 바이오오파운드리 장비의 운용 병목 현상을 최소화, 가동률을 최적화하는 공정 지능화 알고리즘도 필요하다. 컨테이너, 도커 등 클라우드 기술을 활용한 바이오오파운드리 관제 플랫폼을 구축하여, 시공간에 제약없는 바이오오파운드리 운용을 가능하게 해야한다.

바이오오파운드리 of 자동화 및 연속화를 위해서 SiLA (Standardization in Lab Automation), LabOP(Laboratory Open Protocol language) 와 같은 바이오오파운드리 친화적 네트워킹 프로토콜을 적용한 협동로봇이 존재하나 그 수와 종류가 제한적이며, 국내에는 전무 한 실정이다. 또한 상용화되어 있는 대부분의 실험실 데이터 및 프로세스 관리를 위한 LIMS(Laboratory Information Management System; 실험실정보관리시스템)는 작업 스케줄링과 워크플로 관리를 지원하나, 배양용기 이송 로봇을 활용하여 연결된 장비 조합에서는 단일 실험 수행 시 일부 장비가 유휴상태임에도 불구하고 추가적인 실험 수행이 불가능하여 바이오오파운드리 전체 장비의 평균 가동률이 제한될 수 있다. 바이오오파운드리 핵심장비 연계를 위한 국제 표준이 없어, 제조사 별 장비의 인터페이스가 상이, 장비의 통합 운용 및 관리 시스템 구축 난이도 및 비용이 높다. 이를 해결하기 위해 OPC, OPC-UA 와 같은 산업용 자동화 어플리케이션을 위한 통신 프로토콜의 적용이 고려되고 있으며, SiLA, LabOP 등과 같은 Lab automation 을 위한 통신 프로토콜이 오픈 소스로 개발, 공개되었으나, 아직 바이오오파운드리 for 위한 국제 표준으로써 활용되는 플랫폼은 없는 상황이다. 바이오오파운드리 is 장비의 다양성 뿐만 아니라 각각의 장비 및 통합 비용이 상당하기 때문에 보급이 쉽지 않은데 대안으로 클라우드기반 바이오오파운드리 관제 시스템을 도입, 누구나 사용할 수 있는 오픈 플랫폼 형태의 바이오오파운드리가 운영될 수 있다.

4 발전방향 및 국산화 제고 방안

앞으로 바이오오파운드리에서의 발전방향은 크게 3가지로 요약될 수 있다. ①장비와 시스템의 소형화, 복합화, ②일회용 소모품, 그리고 ③자동화 및 자율제조화의 활성화이다.

장비들이 차지하는 공간이 적고 휴대성을 높이도록 개발되고 있다. 한 예로 NGS도 원

거리에서 실시간 시퀀싱이 가능한 휴대용 시퀀싱 플랫폼이 신규 진입하고 있다. Illumina의 NGS NovaSeq(80 x 94.5 x 165.6 cm³)부터 소형화된 iSeq 100(42.5 x 30.5 x 33 cm³)까지 판매 중이다. 또한 개별 장비를 각각 사용하는 것이 아니라 여러 기능을 결합한 hyphenated 시스템을 개발하여 성능을 향상시키고 비용을 절감하는 방향으로 기술이 진화 중이다.

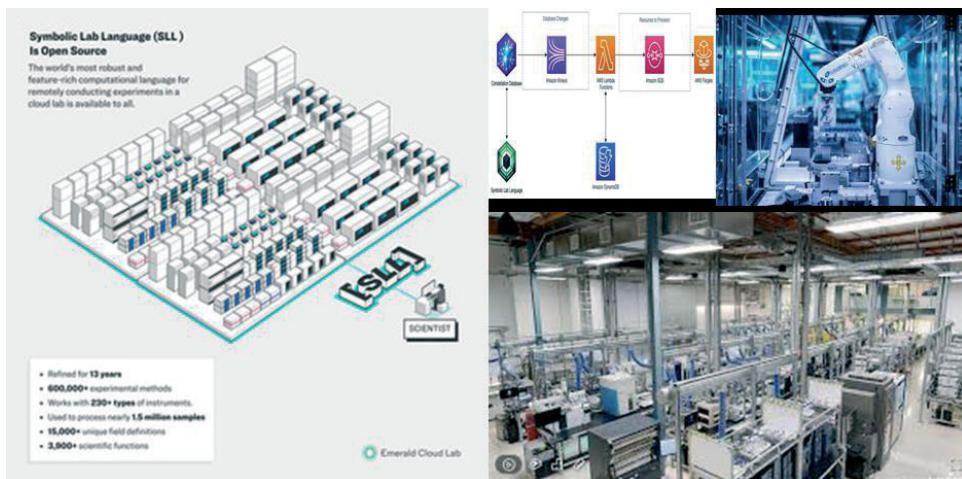
그림 7> Illumina의 NGS 장비들



신속함, 편리함, 효율성 측면에서 상업화 단계에서도 일회용 기반 제조시스템으로 전환이 되어가고 있다. 특히, 고가의 약품이나 독성에 의한 후처리 등을 감안하면 교차오염이 근절되고 자동화 공정 및 설계에 유연성을 부여할 수 있는 싱글유즈백 등 일회용품의 사용은 더 증가될 것으로 예측된다.

바이오장비 및 공정의 정교함이 증가할수록 시료 준비 시간과 비용이 더욱 커지고, 수동 작업에 의한 휴먼 에러 발생 가능성, 인력 부족 등의 문제 해결을 위해 전공정을 자동화·연속화하는 시스템 요구가 증가하고 있으며 이에 대응하는 장비의 개발이 요구되고 있다. 특히, 한 플랫폼에 모든 종류의 데이터를 통합하여 의사결정을 위한 분석이 가능한 새로운 데이터 관리 시스템의 중요성이 증가하므로 핵심 장비들을 서로 연결하여 자동화 및 연속화를 진행하고 클라우드 시스템을 활용, 제어하는 기술이 요구되고 있으며, 바이오장비 분야에서 적용을 시작하고 있다. 가장 대표적인 시스템구축 예는 카네기멜론대학이 만든 Emerald Cloud Lab으로 워크 플로우를 제공하고, 무인 실험실을 바이오 생산 분야에 적용할 수 있도록 사업을 진행 중이다.

그림 8> Emerald Cloud Lab^[9]



2023년 글로벌 바이오의약품 매출은 4,800.3억 달러를 기록하였으며, 향후 6년간(‘23년-’29년) 연평균 9.0%의 증가세로 2029년에는 8,062.9억 달러를 기록할 것으로 전망된다. 더욱이 우리나라가 새로운 성장동력으로 이끄는 2023년 글로벌 바이오의약품 CDMO 매출은 196.8억 달러로 전년 대비(‘22년, 190.1억 달러) 3.5% 증가하였으며, 2029년까지 14.3% 성장하여 438.5억 달러로 성장할 것으로 전망되고 있다^[10]. 점차 바이오 분야의 중요성에 대한 인식이 크게 높아지면서 국가 보건 의료 향상과 바이오헬스 산업 육성에 대한 성장이 가속화될 것으로 전망된다.

하지만 이러한 과정 속에서 사용하게 될 핵심적인 인프라인 바이오장비들의 국산화율은 매우 낮다. 바이오장비의 수입 의존도는 R&D 투자의 해외 유출, 첨단기술 유출 등의 문제를 야기하며, 더 나아가 관련 생태계를 외국에게 종속되게 만들어 국가안보까지 위협할 수도 있다. 이 문제를 해결해야 하는 바이오장비기업들은 대부분 영세한 중소기업으로 수조원대인 해외 글로벌 기업과의 격차가 너무 심하다. 바이오장비는 수요산업의 요구 특성상 매우 보수적이고 경직적인 성격을 갖고 있어 시장에 진입한 선도기업 간 경쟁 강도가 높아 진입장벽이 높은 산업 분야로 주요 품목별로 글로벌 선도 브랜드가 독과점하는 시장으로 국가적 차원의 전략적 대응이 필수적이다.

첨단바이오 기술의 지속 가능한 경쟁력 확보를 위해서는 파이프라인에 사용될 핵심 기기의 국산화 및 공정 지능화, 연속화 등의 고도화 기술 개발을 통해 연구개발 및 사업화 속도를 제고하고, 전문인력 감소, 불확실성 증가, 낮은 생산성 등을 극복하기 위하여 전방위적으로 디지털트랜스포메이션이 이루어져야 한다. 일부 사업들이 기 진행 중이지만 규모

가 적고 산발적이므로, 체계적인 전략과 투자로 국산바이오장비의 신속한 시장 진입과 경쟁력 확보, 독자적인 산업생태계 구축이 가능하도록, 산·학·연이 협력하는 생태계 차원의 대응, 기업 눈높이에 맞는 R&D, 원천기술을 확보할 수 있는 R&D 등이 활발하게 이루어져야 할 것이다.

[참고문헌]

- [1] 2021년도 예비타당성조사 보고서 “바이오파운드리 구축 및 활용기술개발사업” 2021
- [2] 생명공학정책연구센터, 2020 바이오 미래유망기술 2020
- [3] Front. Bioeng. Biotechnol., 30 November 2018
- [4] 2023년도 다부처 공동기획연구 최종보고서 (국가전략기술) “바이오파운드리 핵심기기 및 장비 고도화 기술 개발” 2023
- [5] 기계기술정책지 “바이오장비 산업동향 및 시사점” 2023
- [6] MarketsandMarkets, Life Science Instrumentation Market – Forecast to 2028, 2023.
- [7] BioINpro 146호 “첨단 바이오 R&D 장비 개발동향과 국산화 제고 방안” 2024
- [8] Zdnet.com 2015.05.19.
- [9] <https://www.emeraldcloudlab.com/>
- [10] 한국바이오협회 KBIOIS 브리프 Vol. 72 2024\

[사사표기]

이 연구는 2025년도 산업통상부 및 한국산업기술기획평가원(KEIT) 연구비 지원에 의한 연구임(RS-2025-02218106 & RS-2025-25462399)

합성생물학기반 바이오파운드리 및 공정 자동화 연구동향

1 머리말

합성생물학은 생명체 시스템을 엔지니어링 가능 대상으로 재정의하며, DNA·세포·대사 경로를 설계 가능한 부품으로 다룰 수 있는 새로운 기술 패러다임을 제공한다. 전통적 생물학이 가설 기반 실험 중심이었다면, 현재는 AI 설계·로봇 자동화·디지털 트윈 기반 제어가 결합된 '프로그램 가능한 생명공정' 시대로 전환되고 있다.

특히 바이오파운드리(Biofoundry)는 DBTL(Design-Build-Test-Learn) 자동화를 기반으로 생물학 실험을 산업제조 공정과 유사한 생산 파이프라인으로 전환시킨다. 이러한 기술은 반도체 제조 FAB 모델, MES·QC·AMHS 시스템, 디지털 트윈 기반 예측 제어 기술과 결합하면서 생명공정을 고속·고정밀 플랫폼으로 발전시키고 있다.

기존의 생물학 연구의 전통적인 접근 방식에서 현대적인, 기술 집약적인 접근 방식으로의 전통적인 실험 생물학 연구는 여러가지 구조적 제약을 가지고 있었습니다. 기존에는 실험자 경험을 기반 하고 있어, 연구자의 직관과 경험에 의존하여 실험 방향을 설정했다. 그래서, 반복 실험 및 저속 병목 공정이 되어서, 결과 도출까지 시간이 오래 걸리고 비효율적인 수동 작업이 많았다. 제한된 데이터 기반 해석 위주이었다. 데이터 수집량이 적어 통계적 유의성이나 복잡한 시스템 이해에 한계가 있었다. 최근 기술의 발전은 이러한 한계를 극복하고 생물학을 데이터 기반 공학 분야로 재편하고 있다. 예를들어, DNA 합성 비용 급락으로 DNA 합성이 저렴해지고 빨라져 다양한 유전자 설계 및 실험이 가능해졌다. 초고속 분석 및 데이터 수집을 통해 NGS(차세대 염기서열 분석) 같은 기술로 방대한 양의 생물학적 데이터를 빠르게 얻을 수 있게 되었다. AI 기반 분석 기술로 복잡한 생물학적 데이터를 분석하고 예측하는 데 인공지능이 활용되기 시작했다.

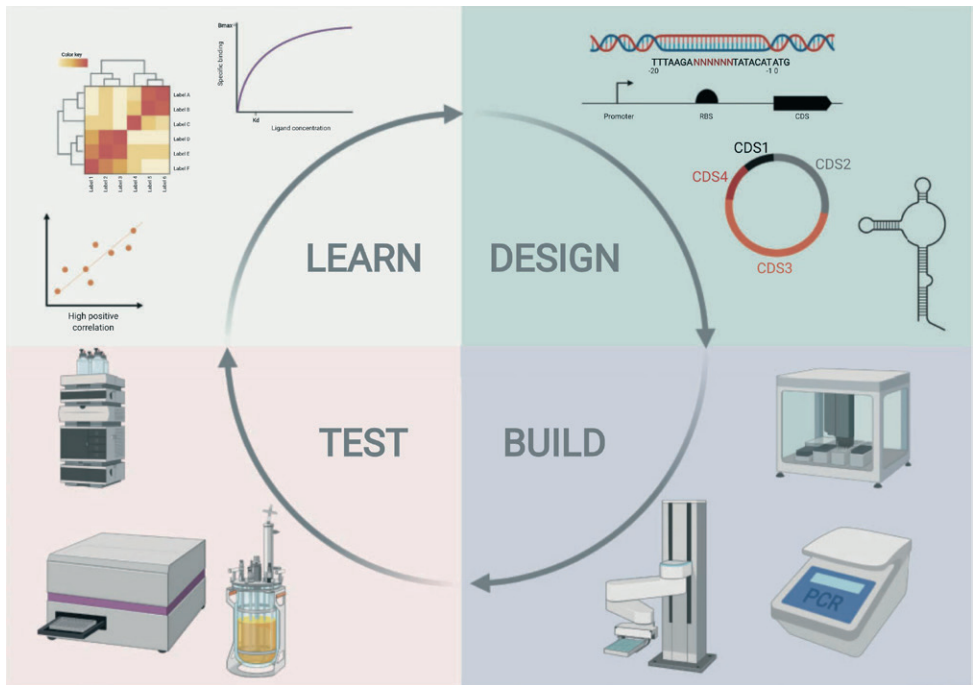
이러한 기술 발전을 바탕으로 현대 합성생물학 연구는 다음과 같이 근본적으로 변화하고 있다. 기존의 논문에서 발표된 것들을 기반하여, 아래와 같이 비교해 보았다. 그래서, 실험 생물학에서 자율적 생명제조인 바이오파운드리로 변화되는 패러다임 전환 면에서 글로벌 동향과 국내에서 국제경쟁력을 갖기위한 전략을 본고에서 제안해 본다.

표 1> 전통적 접근법에 비해 현대 합성생물학연구에서의 변화된 특성들

구분	전통적 접근방식	현대적 접근방식(자율생명제조)
실험 설계	연구자의 가설기반	AI 제안 기반 실험 설계
실험 수행	수작업실험	로봇릭스 및 마이크로플루이드스(자동화)
실험 규모	단일 실험 중심	초병렬 반응 (수천~수만 스케일 동시수행)
최적화 방식	관찰 중심(결과를 보고 판단)	디지털트윈 기반 최적화 (컴퓨터 모델로 예측)
개선 방식	연구자 판단에 의존	Bayesian optimization 및 강화학습(RL) 기반 자동개선

2 DBTL 자동화 기술 동향

DBTL은 현대 생명공학 자동화의 핵심 모델이다.



설계 (Design)-구축(Build)-학습(Learn)-테스트(Test) 사이클. 바이오파운드리 는 일반적으로 사이클의 구축(BUILD) 및 테스트(TEST) 측면에 중점을 두지만, 이는 확실한 규칙은 아님. 예를 들어, 시설은 데이터 분석 및 머신러닝 방법을 개발하여 학습(LEARN) 역량에 더 집중할 수 있음.

(Synthetic Biology, 2020, 6(1): ysaa026)

2.1 DBTL 자동화 기술 동향

2.1.1 합성 생물학의 Design (설계) 단계는 생물학적 시스템을 공학적으로 설계하고 모델링하는 과정으로, 컴퓨터 지원 설계(CAD) 도구, 표준화된 언어, 인공지능 기술을 활용합

니다. 이 단계의 핵심 기술은 다음과 같다. BioCAD 플랫폼은 생물학적 설계를 위한 소프트웨어 환경을 의미한다. 이는 유전자 서열을 편집하고, 구성 요소를 조합하며, 전체 시스템의 작동을 시뮬레이션하는 데 사용된다. 마치 기계 공학에서 CAD 소프트웨어를 사용해 제품을 설계하듯이, BioCAD는 연구자들이 컴퓨터 상에서 유전자 회로, 단백질, 대사 경로 등을 시각화하고 설계할 수 있도록 한다. SBOL (Synthetic Biology Open Language)은 서로 다른 소프트웨어 도구 및 연구자 간에 합성 생물학적 설계를 교환하고 공유하기 위한 개방형 데이터 표준을 말한다. 유전자, 프로모터, 리보솜 결합 부위 등 다양한 생물학적 부품(parts)과 이들의 기능을 표준화된 형식으로 표현할 수 있게 하여, 설계의 재현성과 호환성을 크게 향상시킨다. SBOL 표준을 통해 전 세계 연구자들이 자신의 설계를 쉽게 주고받으며 협업할 수 있다. 생성형 AI 기반 모델 및 자동화 기술도 핵심이라고 할수있다.

예를 들면, 단백질의 기능은 3차원 구조에 의해 결정되지만, 이를 실험적으로 밝히는 것은 매우 어렵고 오랜 시간이 걸렸다. AlphaFold (Google DeepMind 개발)와 ESM (Meta 개발)은 아미노산 서열만으로 단백질의 3차원 구조를 높은 정확도로 예측하는 인공지능(AI) 모델이다. 이러한 모델은 신약 개발, 효소 공학 등 다양한 분야에서 새로운 단백질 설계 및 기존 단백질의 기능 분석에 혁신을 가져왔다.

대사경로 디자인 자동화는 미생물이 특정 유용 물질(예: 바이오 연료, 의약품 원료)을 생산하도록 새로운 대사 경로를 설계하거나 기존 경로를 최적화하는 과정이다. AI와 기계 학습 알고리즘은 대규모의 생물학적 데이터를 분석하여 가능한 경로들을 제안하고 시뮬레이션함으로써, 가장 효율적인 생산 전략을 자동으로 찾아낸다. 이는 실험적 시행착오를 줄이고 개발 속도를 높여준다.

유전자 회로는 세포 내에서 특정 기능을 수행하도록 설계된 유전자들의 조합이다(예: 특정 화학 물질 감지 시 형광 단백질 생성). Cello Framework와 같은 자동 설계 도구는 사용자가 원하는 논리적 기능(예: AND, OR 게이트)을 입력하면, 이를 수행할 수 있는 최적의 유전자 서열과 부품 조합을 자동으로 설계해 준다. 이는 전자 공학에서 집적 회로를 설계하는 것과 유사한 접근 방식으로, 복잡한 생물학적 시스템 설계를 간소화할수있다.

2.1.2

Build 단계의 핵심 기술 및 목표는 대략 아래와 같다. DNA 합성 및 조립 기술으로, 원하는 기능을 수행하는 새로운 유전자 회로 또는 게놈을 설계한 후, 실제로 DNA 조각들을 만들고(합성) 이를 연결하여(조립) 완성된 유전 공학적 구조를 만든다. 초병렬(Massive

Parallel) 변환 및 배양기술으로, 만들어진DNA를 숙주 세포(예: 대장균, 효모)에 도입하여 형질을 전환(Transformation)시킨 후, 이 세포들을 동시에 대규모로 배양한다. 여기서 '초병렬'은 수천, 수만 가지의 서로 다른 유전자 조합을 동시에 테스트할 수 있는 능력을 의미한다. Droplet microfluidics 기반 대량 조립 기술은, 마이크로 단위의 액체 방울(droplet)을 정밀하게 제어하는 기술이다. 이 작은 방울 안에서 DNA 조립, 유전자 도입, 세포 배양 등 다양한 생화학 반응을 수행할 수 있다. 장점으로는, 대량 처리능력으로 기존의 실험 방식보다 훨씬 적은 시약과 공간으로 수많은 반응을 동시에 처리할 수 있다. 효율성 극대화 능력으로, 반응 속도를 높이고 처리량을 획기적으로 증가시켜, 다양한 유전자 조합 라이브러리를 빠르게 구축하고 선별할 수 있다.

Automated strain engineering 플랫폼 기술이 있다. DNA 설계부터 합성, 세포 도입, 배양, 분석에 이르는 전체 과정을 로봇 및 자동화 시스템으로 구현한 것이다. 이는 바이오 파운드리 핵심 구성 요소입니다. 이 기능으로 인간 작업자의 개입을 최소화하고 표준화된 프로토콜에 따라 작업을 수행하여, 실험의 재현성을 높이고 처리 속도를 극대화할 수 있다.

세포공장 구축 비용 최소화 기술 또한 위와 같은 첨단 자동화 기술(Droplet microfluidics, automated platforms)을 통해 세포 공장(Cell factory) 개발에 소요되는 시간과 비용을 획기적으로 줄이는 것으로 중요하다. 효율적인 균주 개발 및 최적화 과정을 통해 바이오 제품 생산의 경제성을 확보하고, 상용화를 촉진할 수 있다는 면에서 의의가 있다. 요약하자면, 이 과정은 자동화된 대규모 병렬 처리 기술(automated platforms, droplet microfluidics)을 활용하여 DNA 조립 및 세포 배양의 효율을 극대화하고, 이를 통해 세포 공장 구축 비용을 최소화하는 것이 가능하다.

2.1.3

Test 단계는 생물학적 시스템(예: 세포 배양, 바이오리액터 공정 등)의 품질 관리(QC) 및 이상 징후 탐지(anomaly detection)를 위해 다양한 최첨단 분석 기술을 통합적으로 활용하는 과정을 의미한다. 이는 주로 바이오의약품 생산, 세포 치료제 개발 또는 정밀 의학 연구와 같은 분야에서 적용된다. 예를 들면, LC-MS (액체 크로마토그래피-질량분석법), GC-MS (기체 크로마토그래피-질량분석법), 프로테오믹스(Proteomics, 단백질체학), high-content imaging과 같은 고도화된 기술을 이용해 세포의 성장, 대사, 건강 상태 등을 실시간 또는 주기적으로 정밀하게 관찰할 수 있다. 품질 관리(QC)로 분석 데이터를 바탕으로 제품 또는 공정이 표준 규격에 부합하는지 확인하고 일관성을 유지할 수 있다. 또한, 이상

징후 탐지(Anomaly Detection)기능으로 정상 범주를 벗어나는 데이터 패턴이나 예상치 못한 변화를 조기에 감지하여 문제 발생을 예방하거나 신속하게 대응할 수 있다.

2.1.4

Learn 단계에서, AI 기반 최적화 엔진은 인공지능을 활용하여 복잡한 시스템이나 프로세스의 성능을 최적으로 개선하는 과정을 의미한다. 이 과정은 베이지안 최적화, 강화학습(RL) 기반 pathway tuning, 실패 데이터 학습 기반 튜닝한 모델링이라는 세 가지 핵심 기술로 구성됩니다.

구체적으로, Bayesian Optimization (베이지안 최적화)은 목적 함수(최적화하려는 대상)의 평가 비용이 비싸거나 시간이 오래 걸릴 때, 효율적으로 최적값을 찾기 위한 확률론적 모델 기반 최적화 기법이다. 평가된 소수의 데이터 포인트를 바탕으로 미지의 목적 함수에 대한 확률적 모델(대리 모델 또는 서로게이트 모델)을 구축하는 원리이다. 이 모델과 획득 함수(Acquisition Function)를 사용하여 다음에 어느 지점을 탐색해야 가장 큰 성능 개선을 얻을 수 있을지 판단하여 불필요한 탐색을 줄이고 최적의 파라미터를 효율적으로 탐색할 수 있다. 주로 AI 모델의 하이퍼파라미터 튜닝이나 실제 실험(예: 신약 개발, 재료 공정 조건 탐색 등)에서 최적의 조건을 찾는 데 활용된다.

강화학습(RL) 기반 Pathway Tuning (경로 튜닝) 기술은 환경과의 상호작용을 통해 최적의 행동 또는 행동 순서(경로)를 학습하여 궁극적인 보상을 최대화하는 데 사용된다. 에이전트(AI)가 현재 상태를 인식하고, 선택 가능한 행동 중 보상이 최대화되는 행동을 선택하도록 학습하는 원리이다. 이 과정에서 명시적인 프로그래밍이나 레이블링된 데이터 없이 시행착오를 통해 스스로 학습한다. “Pathway tuning”이라는 문맥에서는 특정 목표 달성을 위한 일련의 순차적인 결정 과정이나 경로를 최적화하는 데 적용될 수 있다. 예를 들어, 자율 시스템의 의사 결정 과정, 복잡한 공정의 단계별 최적화, 네트워크 라우팅 경로 설정 등에 활용될 수 있다는 것이다.

실패 데이터 학습 기반 튜닝한 모델링 기술은 일반적인 성공 사례 데이터뿐만 아니라, 실패한 사례에 대한 데이터도 학습 과정에 포함시켜 모델의 신뢰성과 견고성을 높이는 역할을 한다. 실패 데이터는 일반적인 데이터셋에 비해 불균형하거나 부족할 수 있지만(데이터 희소성 문제), 이 데이터를 분석하여 실패의 원인이 되는 패턴이나 조건을 식별한다는 원리를 기반한다. 이를 통해 모델은 위험 요소나 예외 상황에 대비하고, 실제 환경에서 발생할 수 있는 오류를 최소화하도록 학습된다. 모델이 특정 조건에서 왜 실패하는지 이해함으로써 예측 불가능한 상황에 대한 대응력을 높이고, 더욱 안정적이고 튜닝한(Robust)

시스템을 구축하는데 활용되고 있다.

이 세 가지 기술은 서로 보완적으로 작동하여 효율적인 탐색(베이지안 최적화), 동적인 의사 결정 경로 최적화(강화학습), 실제 환경에서의 안정성 확보(실패 데이터 학습)를 모두 달성하는 강력한 AI 기반 최적화 엔진을 구성한다.

3 Self-Driving Lab(SDL) 아키텍처

SDL(Self-Driving Lab)은 인공지능(AI)을 활용하여 실험 설계, 수행, 해석, 최적화 과정을 완전히 자동화하는 자율 연구 시스템이다. 인간 과학자의 개입을 최소화하고 연구 속도와 효율성을 획기적으로 높이는 것을 목표로 한다.

SDL의 핵심 요소 및 대표 연구는 대략 다음과 같다.

자연어 기반 목적 해석(NLP/LLM): 사용자가 자연어(일반적인 언어)로 연구 목표를 입력하면, 시스템이 이를 이해하고 구체적인 실험 계획으로 변환한다. 이는 대규모 언어 모델(LLM)과 자연어 처리(NLP) 기술을 활용한다.

프로토콜 컴파일러(AutoProtocol·SiLA2 API): 자연어 명령이나 추상적인 실험 계획을 로봇 및 실험 장비가 실행할 수 있는 구체적인 기계어 명령(프로토콜)으로 변환한다. SiLA2 API는 서로 다른 제조사의 장비 간 통신 표준 역할을 한다.

동적 스케줄링(멀티로봇·장비 오케스트레이션): 여러 대의 로봇과 다양한 실험 장비의 작업을 효율적으로 조율하고 관리한다. 실험 과정 중 발생하는 예상치 못한 상황에 따라 계획을 유동적으로 변경(동적 스케줄링)하여 최적의 실험 진행을 보장한다.

multi-omics + imaging QC 센싱: 실험 과정에서 생성되는 다양한 데이터(유전체학, 단백질체학 등"omics" 데이터 및 이미징 데이터)를 실시간으로 수집하고 품질 관리(QC)를 수행한다. 이는 실험의 신뢰성을 높이는 데 필수적이다.

BO/RL 모델로 공정 지속 최적화: 베이지안 최적화(Bayesian Optimization, BO)나 강화 학습(Reinforcement Learning, RL)과 같은 AI/머신러닝 모델을 사용하여 실험 결과를 분석하고 다음 실험 조건을 제안한다. 이 과정을 반복하며 연구 목표 달성까지 지속적으로 공정을 최적화할 수 있다. 대표 연구 사례로는 아래와 같다.

A-Lab (Nature, 2023): 효소 최적화를 위한 자율 연구소이다. AI가 효소의 활성을 최대화하기 위한 조건을 스스로 탐색하고 실험하며 최적화 과정을 성공적으로 수행하여 네이

처(Nature)지에 발표되었다. [1]

iBioFAB (UIUC): 미국 일리노이 대학교 어바나-샴페인(UIUC)의 자율 생물학 파운드리입니다. DNA 구축 과정을 자동화하여 기존 대비 비용을 97.7% 절감하는 성과를 보였다.

4 글로벌 선도 인프라 및 사례

합성 생물학 및 바이오파운드리 분야에서 자동화, 데이터 기반 접근 방식, 산업 적용에 중점을 둔 핵심 글로벌 대표 인프라 및 사례들을 아래와 같이 정리해 보았다. 각 기관의 특징은 다음과 같습니다:

4.1 미국 DOE Agile Biofoundry (ABF)

대규모 DBTL 자동화: Design-Build-Test-Learn (설계-제작-실험-학습) 주기를 고도로 자동화하고 확장하는 데 중점을 둡니다. 이를 통해 새로운 생물학적 시스템 개발 속도를 획기적으로 높이고 있다.

산업 적용 중심: 미국 에너지부(DOE) 산하 기관으로서, 바이오에너지, 바이오제품, 친환경 화학물질 생산 등 산업적으로 활용 가능한 기술 개발을 목표로 하고 있다.

4.2 Ginkgo Bioworks

대규모 생명 제조 플랫폼(Biofab): 세계 최대 규모의 민간 바이오파운드리를 운영하며, 고객을 대신해 대규모 생물학적 실험 및 생산을 수행하고 있다.

데이터 기반 Cell programming (세포 프로그래밍): 방대한 유전자 코드베이스와 자동화된 워크플로우를 통합하여 미생물(세포)을 설계하고 엔지니어링하고 있다. AI 및 머신러닝을 활용해 대규모 고품질 생물학적 데이터 세트를 생성하고 분석하고 있다.

4.3 SynCTI (Singapore)

병원-공정 통합형 자동화 연구를 수행하는 싱가포르의 합성 생물학 연구소(Synthetic Biology Research Consortium, Singapore-MIT Alliance for Research and Technology) 산하 기관이다. 병원성 미생물 연구부터 산업 공정 적용까지 통합된 자동화 연구 시스템을 구축하고 있다.

4.4 RIKEN (Japan) / AIST

자율 세포 엔지니어링, AI-omics: 일본 이화학연구소(RIKEN)와 산업기술종합연구소(AIST)는 자율적으로 세포를 엔지니어링할 수 있는 시스템과 인공지능(AI) 기반의 옴믹스(omics) 데이터(유전체학, 단백질체학 등) 분석 기술을 결합하여 연구 효율을 극대화하고 있다.

4.5 CAS-TIB (China)

국가 대규모 생명공학 클러스터: 중국과학원 산하 텐진 산업생명공학연구소(CAS-TIB)는 국가 차원의 대규모 투자를 바탕으로 생명공학 공정 연구 및 산업화를 위한 대형 클러스터를 조성하고 있습니다. 이는 중국 바이오경제 발전을 위한 핵심 인프라 역할을 한다.

이 기관들은 모두 자동화된 바이오파운드리를 기반으로 하여, 기존의 수작업 중심 연구 방식에서 벗어나 데이터 및 AI 기반의 효율적이고 신속한 바이오 연구 및 제조를 선도하고 있다는 공통점을 가지고 있다.

5 한국 연구 동향 및 기회

한국은 반도체·ICT·정밀장비 기반을 보유한 국가로서, 대규모의 반도체 파운드리를 다수 보유하고 있다. 반도체 파운드리는 이미 확립된 반도체 대량 생산 시스템인 반면, 바이오 파운드리는 생명공학 분야에 공학적 자동화 개념을 도입하여 연구 속도를 획기적으로 높이고 새로운 바이오 경제를 열어가는 초기 단계의 핵심 인프라라고 볼 수 있습니다. 비즈니스 모델과 생산 개념에서 공통점을 가지면서도, 다루는 대상과 목적에서 차이가 있다.

간단히 비교하면, 바이오파운드리와 반도체 파운드리는 모두 '위탁 생산(파운드리)'이라는 비즈니스 모델과 '자동화' 및 '분업화'라는 개념을 공유하지만, 다루는 대상과 목적에서 큰 차이를 보입니다.

공통점은 아래와 같다.

- 파운드리 개념: 외부의 설계(바이오텍팍리스)를 받아 생산을 전담하는 위탁 생산 방식의 비즈니스 모델을 따른다.
- 분업화: 설계와 생산이 분리되어, 각 전문 분야에 집중할 수 있는 산업 생태계를 지향한다. 반도체는 팍리스-파운드리로, 바이오는 바이오텍-CDMO/CMO로 나뉘는 경향

이 유사합니다.

- 자동화 및 효율 추구: 대규모 시설(팹, 바이오파운드리 센터)에 고가의 장비를 갖추고, 자동화된 공정을 통해 생산 효율성과 속도를 극대화한다.
- 표준화: 공학적인 접근 방식을 도입하여, 설계부터 생산까지의 과정을 표준화하고 정밀도를 높이는 것을 목표로 한다.

표 2> 반도체파운드리와 바이오파운드리 간의 특성 비교

구분	반도체 파운드리	바이오 파운드리
주요 대상	실리콘 웨이퍼 기반의 전자회로(무생물)	유전자(DNA), 단백질, 세포 등 생명체 구성요소(생물)
주요 목적	최종 제품인 반도체 칩의 대량 양산	새로운 바이오제품 생산을 위한 연구개발(R&D) 및 시제품 제작 가속화
공정 특성	고도로 정밀하고 표준화된 제조 공정. 변수 통제가 용이함	생물학적 특성으로 인해 다양한 변수가 발생할 수 있으며, 여전히 사람의 개입이 필요한 경우가 많음
산업 성숙도	이미 고도로 분업화되고 성숙한 주력 산업	합성생물학 기반의 초기 단계 산업으로, 인프라 구축 및 기술 선점 경쟁 중
최종 결과물	CPU, 메모리 등 물리적 반도체 제품	의약품 원료, 바이오 연료, 신소재 등 다양한 형태의 바이오 제품 또는 솔루션

이러한 대규모 제조 시설의 설계, 구축, 및 AI의 융합으로 이루어지는 풍부한 경험이 있어, 바이오파운드리 구축면에서도 공통점, 기회 및 다음과 같은 영역에서 실질적인 강점이 존재한다.

6 맺음말

합성생물학은 생명 시스템을 공학적으로 설계, 제작, 조립하는 기술로, 기존의 실험실 중심 연구에서 인공지능(AI)과 로봇공학, 빅데이터 기술과 결합하여 표준화되고 자동화된 '바이오 제조(Bio-manufacturing)' 산업으로 급격히 발전하고 있습니다. 이는 바이오 제품 생산의 속도와 효율성을 획기적으로 높이고 비용을 절감하는 게임 체인저로 부상하고 있습니다.

미래 바이오 산업의 경쟁력은 연구자의 개별적인 '실험 능력'이 아니라, AI와 자동화 시스템이 주도하는 '자율 생명제조 운전 능력', 즉 스스로 설계하고, 만들고, 테스트하며 학습하는 (DBTL) 자동화된 생산 시스템을 얼마나 효율적으로 운영하느냐에 의해 결정될 것입니다.

대한민국은 세계 최고 수준의 반도체, 정밀 제조, 의료 데이터 기반 기술을 보유하고 있다, 이를 바이오 분야에 접목할 때 큰 시너지 효과를 낼 수 있다. 이러한 강점은 바이오-디

지털(Bio-Digital) 제조 분야의 선두 주자(프런티어)로 도약할 수 있는 기반이 될 수 있다. 따라서, 치열한 국가적 과학기술 경쟁을 고려할 때, 기술선점 위해, 정부의 지속적이면서도 과감한 연구과제 지원이 필요한 상황이다.

[사사표기]

※ 본 연구는 국가과학기술연구회 글로벌 TOP 전략연구단 유전체 기반 개인맞춤형 암유전자 진단기술 개발 사업에서 수행됨[빅데이터 활용 AI-ICT 진단플랫폼 개발 과제]

[참고문헌]

- [1] Nielsen, A.A. et al. Genetic circuit design automation. Science 352, aac7341 (2016).
- [2] Khalil, A.S. & Collins, J.J. Synthetic biology: applications come of age. Nat. Rev. Genet. 11, 367–379 (2021).
- [3] Purnick, P.E. & Weiss, R. The second wave of synthetic biology. Nat. Rev. Mol. Cell Biol. 10, 410–422 (2009).
- [4] Brophy, J.A. & Voigt, C.A. Principles of genetic circuit design. Nat. Methods 11, 508–520 (2014).
- [5] Lee, S.Y. et al. Systems metabolic engineering. Nat. Rev. Microbiol. 17, 665–681 (2019).
- [6] Hillson, N. et al. Building a global alliance of biofoundries. Nat. Commun. 10, 1038 (2019).
- [7] Chao, R. et al. Biofoundries: engineering biology infrastructure. Nat. Rev. Methods Primers 3, 12 (2023).

- [8] Carbonell, P. et al. Bio-CAD: design tools for synthetic biology. Trends Biotechnol. 37, 691–705 (2019).
- [9] Griffiths, R. et al. A-Lab autonomous system. Nature 616, 72–78 (2023).
- [10] Ghanem, M. Digital twins in biotechnology. Nat. Biotechnol. 40, 1450–1460 (2022).

[주요용어풀이]

- MES (Manufacturing Execution System): 작업 현장에서 작업 일정, 작업 지시, 품질 관리, 작업 실적 집계 등 제반 활동을 지원하기 위한 관리 시스템. 생산 계획과 실행의 차이를 줄이기 위한 시스템으로 현장 상태의 실시간 정보 제공을 통하여 관리자와 작업자의 의사 결정을 지원하는 기능을 수행.
- CAD (Computer-Aided Design): 기계, 전기, 건축 등 분야의 설계 작업에 컴퓨터를 이용하는 것. 전문 지원 도구를 사용하여 생산성이 오르고, 품질이 향상되며, 자동화 생산 등 관련 업무와 연동될 수 있는 장점이 있음.
- Digital Twin: 물리적인 사물과 컴퓨터에 동일하게 표현되는 가상 모델. 제너럴 일렉트릭(GE: General Electric)에서 만든 개념.
- RL (Reinforcement Learning): 강화학습, 기계학습 중 컴퓨터가 주어진 상태(State)에 대해 최적의 행동(Action)을 선택하는 학습 방법. 강화형 기계학습은 지도형/비지도형 기계학습에 이용되는 훈련 데이터 대신, 주어진 상태에 맞춘 행동의 결과에 대한 보상(Reward)을 준다. 컴퓨터는 보상을 이용하여 성능을 향상한다. 주로 게임이나 로봇 제어 등에 적용.
- NLP (Natural Language Processing): 사람이 사용하는 언어를 컴퓨터가 인식하여 이를 처리할 수 있도록 하는 인공지능 기술.
- LLM (Large Language Model): 대규모 언어모델, 대규모 텍스트 데이터셋을 활용하여 학습한 파라미터의 수가 매우 큰 모델로 문장을 이해하거나 새로운 텍스트를 생성하는 데 사용.
- API (Application Programming Interface): 응용 프로그래밍 인터페이스, 응용 소프트웨어

트웨어와 소프트웨어 플랫폼 사이에 존재하며, 플랫폼 상에서 실행하는 애플리케이션 개발을 용이하게 하는 소프트웨어 라이브러리 집합.

- ICT (Information and Communication Technology): 정보통신기술, 인터넷, 휴대폰 등의 기술 및 운영 프로그램을 통칭하는 것으로 기존의 정보 기술(IT, Information Technology)의 연장으로 주요 차이점은 정보 기술과 통신 사업간 융합을 강조한 것.
- GC-MS (Gas chromatography mass spectrometry): 가스 크로마토그래피와 질량 분석계가 합쳐진 장비를 의미.
- CMO (Contract Manufacturing Organization) : 의약품 위탁생산 기업, 다른 제약사의 의약품을 대신 생산해주는 전문 기업을 의미
- CDMO (Contract Development and Manufacturing Organization) : 위탁개발생산, 제약·바이오 회사를 대신하여 의약품의 개발부터 생산까지 전 과정을 통합적으로 제공하는 서비스를 말함.

합성생물학 자동화를 위한 분석장비 동향

1 머리말

합성생물학은 생물학적 시스템을 공학적 원리로 설계하고 구축하는 분야로, 전통적인 생명공학이 자연에서 발견된 기능을 경험적으로 개선하는 데 초점을 두었다면, 현대 합성생물학은 설계 가능성과 재현성을 높이는 체계적 엔지니어링을 지향한다. 이러한 패러다임 전환의 중심에는 DBTL(Design-Build-Test-Learn) 사이클, 즉 설계-제작-시험-학습 사이클이 있으며, 최근에는 로봇 자동화 기술과 클라우드 기반 계산 자원의 접목으로 하루 수천 건의 샘플 처리와 대용량 시퀀싱 데이터의 생산·분석이 일상화되면서 개발 주기가 획기적으로 단축되고 있다.

이 과정에서 분석 기술은 단순한 보조 수단을 넘어 DBTL 사이클의 정밀도와 신뢰성을 결정짓는 기반 인프라로 작용한다. 차세대 염기서열 분석(NGS, Next-Generation Sequencing)은 DNA 구조체 검증의 병목을 해소하며 바이오파운드리 운영의 표준이 되고 있으며, 마이크로플레이트 리더와 형광 활성 세포 분리 시스템 (FACS, Fluorescence-Activated Cell Sorting)기술을 중심으로 한 고처리량 스크리닝 기술은 수천 개 균주의 병렬 평가를 가능하게 한다. 질량분석 기반 대사체학은 세포 내 탄소 흐름을 정량화하여 대사 경로 재설계의 과학적 근거를 제공하며, 라만 분광법을 비롯한 실시간 공정 분석 기술(PAT, Process Analytical Technology)은 공정 데이터를 즉각적으로 제어 시스템과 연계하여 무인 적응형 바이오 제조를 실현하고 있다.

본 보고서는 합성생물학 분야에서 활용되는 핵심 분석 솔루션의 최신 기술 동향을 종합적으로 검토한다. 징코 바이오웍스(Ginkgo Bioworks), 아미리스(Amyris), 제노마티카(Genomatica) 등 선도 기업들의 구체적인 기술 구현 사례를 통해, DNA 구조체 검증, 고처리량 표현형 스크리닝, 질량분석 기반 대사 흐름 제어, 실시간 분광 분석 등 각 기술이 합성생물학 자동화에 기여하는 방식과 산업적 적용 현황을 다룬다. 특히 자동화, 표준화, 데이터 연계 관점에서 분석 장비가 어떻게 설계 정보와 실제 공정 반응 간 실시간 피드백 루프를 구성하며, 인공지능 기반 예측 모델 및 디지털 트윈과 결합되어 합성생물학 자동화의 핵심 인프라로 진화하고 있는지 조망한다.

2 합성생물학 핵심 과제와 분석 솔루션

2.1 합성 DNA 구조체 검증과 차세대 염기서열 분석 기술

합성생물학에서 설계된 DNA 구조체가 정확하게 조립되었는지 확인하는 검증 과정은 전체 연구개발 비용과 시간의 상당 부분을 차지하며, 전통적인 생어 시퀀싱(Sanger sequencing) 방식은 샘플당 높은 비용과 낮은 처리량으로 인해 수백~수천 개의 DNA 클론을 동시에 검증해야 하는 바이오파운드리 환경에서 병목 현상을 유발했다. 이러한 한계를 극복하기 위해 차세대 염기서열 분석 기술이 DNA 구조체 검증의 핵심 도구로 부상했으며, 단일 시퀀싱 런에서 수천 개 샘플을 병렬 처리하고 샘플당 비용을 절감하는 것이 가능해졌다.

본 절에서는 징코 바이오웍스의 자동화된 차세대 염기서열 분석 검증 인프라와 클라우드 기반 데이터 파이프라인 구축 사례, 아미리스의 차세대 염기서열 분석 기반 품질관리 시스템이 합성생물학 산업의 표준으로 확산된 과정, 그리고 단문형 시퀀싱의 한계를 극복하기 위해 도입된 팍바이오 에스엠알티(PacBio SMRT, Single Molecule Real-Time sequencing)와 옥스퍼드 나노포어(Oxford Nanopore) 기반 장문형 시퀀싱 기술의 활용 사례를 살펴본다.

2.1.1 자동화된 NGS 검증 인프라 구축

미국의 징코 바이오웍스는 합성생물학의 전주기 자동화를 위해 대규모 바이오파운드리를 구축하고, 이를 기반으로 DNA 합성 및 검증 시스템을 운영하고 있다. 이 플랫폼은 일루미나 노바섹(Illumina NovaSeq) 기반 차세대 염기서열 분석을 중심으로 수천 개 DNA 샘플을 병렬 처리하며, 로봇이 미생물 콜로니의 자동 선별·배양을 수행한다.

이후 에이더블유에스 배치(AWS Batch)와 아파치 에어플로(Apache Airflow)로 구성된 클라우드 기반 파이프라인을 통해 데이터 정제, 정렬, 품질 검증이 이루어진다. 샘플 수에 따라 유연하게 구성되는 자동화 워크플로는 품질 검사와 시퀀싱을 병렬화하여 기존 생어 시퀀싱 방식 대비 처리 속도를 향상시켰다.

징코 바이오웍스의 플랫폼은 시퀀싱 데이터를 품질관리 데이터베이스와 연동하여 오류 패턴을 머신러닝으로 학습하고, 설계 단계에 피드백하는 구조를 갖추고 있다. 이를 통해 합성 오류나 유전적 변이를 탐지하고, 반복적인 학습을 통해 설계 정확도를 개선한다. 이러한 데이터 중심 자동화 시스템은 합성생물학에서 시간이 많이 소요되던 검증 단계를 단축시켰으며, 바이오파운드리 산업의 품질 표준화와 생산성 향상에 기여하고 있다. 징코 바이오웍스 사례는 연구실 수준의 자동화 기술을 산업 수준으로 확장한 사례로 활용되고 있다.

2.1.2 NGS 기반 품질관리 시스템

아미리스는 산업용 합성생물학에서 DNA 조립의 정확성을 확보하기 위해 일루미나 미섹(Illumina MiSeq) 플랫폼과 넥스테라(Nextera) 기반 시퀀싱 기술을 활용한 품질 검증 체계를 개발했다. 2015년 에이씨에스 신세틱 바이올로지(ACS Synthetic Biology)에 발표한 연구에서 한 번의 시퀀싱 런으로 4,000개 이상의 DNA 클론을 확인할 수 있는 자동화 워크플로를 구축했다.

이를 통해 플라스미드 1개당 약 15배 커버리지를 확보하면서도 시료당 3달러 미만 비용으로 전수 검증이 가능해졌으며, 기존 생어 시퀀싱과 비교해 약 20배의 비용 효율성을 달성했다. 아미리스 연구팀은 넥스테라 태그멘테이션(tagmentation) 반응을 100배로 소형화하고 수천 개 샘플을 추적하는 소프트웨어를 개발하여 합성 DNA 제작의 신뢰성과 공정 속도를 향상시켰다.

이러한 접근법은 내부 품질 관리를 넘어 합성생물학 산업의 품질 보증 표준으로 확산되었다. 아미리스는 다르파(DARPA, Defense Advanced Research Projects Agency) 협력 1단계로 400개 분자에 대해 400개 군주를 구축하는 성과를 달성했다. 이후 징코 바이오웍스, 지머겐(Zymergen, 현재 징코에 인수) 등 주요 바이오파운드리 기업들이 자동화된 차세대 염기서열 분석 기반 품질 검증 프로세스를 채택하면서, 데이터 중심 군주 엔지니어링으로 기술 패러다임이 전환되었다.

라이거스(Lygos)는 미국 에너지부 애자일 바이오파운드리(Agile BioFoundry)와 협력하여 50만 개 이상의 데이터 포인트를 생성하는 다중오믹스 데이터셋 프로젝트에 참여하고 있다. 아미리스가 구축한 차세대 염기서열 분석 기반 품질관리 체계는 합성생물학 자동화 분야의 참고 사례로 활용되고 있다.

2.1.3 장문형 시퀀싱의 구조 검증 자동화 적용

합성생물학에서 제작한 복잡한 유전자 구조물이 설계대로 정확히 조립되었는지 확인하는 것은 중요한 과정이다. 기존에 사용되던 일루미나 등의 단문형 시퀀싱 기술은 짧은 DNA 조각만을 읽을 수 있어, 반복 서열이나 큰 규모의 유전자 삽입 부분에서는 검증이 어려웠다. 특히 수천 개 염기 이상의 긴 합성 DNA를 검증할 때는 조립 과정에서 발생한 오류나 누락을 찾아내기 어려운 한계가 있었다.

이러한 문제를 해결하기 위해 긴 DNA 서열을 한 번에 읽을 수 있는 장문형 시퀀싱 기술이 활용되고 있다. 팍바이오의 에스엠알티 시퀀싱은 99.9% 이상의 정확도로 1만~2만 개 염기를 연속으로 읽어낼 수 있고, 옥스퍼드 나노포어의 나노포어 시퀀싱은 100만 개 염기

이상의 초장문형 서열도 분석할 수 있어 복잡한 유전자 구조의 검증이 가능해졌다.

장문형 시퀀싱 기술의 성숙과 함께 자동화된 분석 시스템들이 개발되고 있다. DNA 서열 읽기부터 최종 분석 결과 도출까지 6단계의 과정이 자동으로 처리되는 표준 워크플로가 구축되어, 복잡한 분석 과정을 거치지 않고도 검증 결과를 얻을 수 있다. 이 기술은 DNA를 증폭시키지 않고 원본 상태에서 분석할 수 있어, 구조체의 실제 특성을 확인할 수 있다.

옥스퍼드 나노포어의 미니온(MinION)과 같은 소형 장비는 현장에서 유전자 검증이 가능하며, 실시간으로 결과를 확인할 수 있어 연구 효율을 향상시키고 있다. 장문형 시퀀싱은 합성생물학 연구에서 중요한 검증 도구로 활용되고 있다.

2.2 고처리량 표현형 분석 기술의 진화

합성생물학의 설계-구축-테스트-학습(DBTL) 사이클을 가속화하기 위해서는 대량의 군주 변이체를 신속하게 평가할 수 있는 고처리량 분석 기술이 필수적이다. 질량분석이 세포 내 대사 흐름을 정밀하게 추적하는 기술이라면, 고처리량 표현형 분석은 수백~수만 개 군주의 전반적 성능을 병렬로 평가하는 스크리닝 기술이다.

본 절에서는 마이크로플레이트 기반 자동 스크리닝 시스템과 유세포 분석 기반 단일세포 분석 기술이 바이오파운드리 자동화에 어떻게 기여하고 있는지 살펴본다. 마이크로플레이트 리더를 중심으로 수백~수천 개 군주의 병렬 분석 시스템, 광유전학 적용을 위한 조명 통합형 플랫폼, 머신러닝 기반 최적화를 통한 생산 수율 개선 사례를 검토할 것이다. 또한 단일세포 수준에서는 AI 기반 데이터 분석과 마이크로플루이딕스가 결합된 유세포 분석 시스템의 자동화 처리 능력, 이미징 및 질량 유세포 분석을 통한 복잡한 세포 특성 정량화 방법을 다룰 예정이다. 이를 통해 고처리량 표현형 분석 기술이 유전자 회로 평가, 생합성 회로 분석, 세포주 개발 등 다양한 적용 분야에서 생물학적 부품의 상업화 가능성을 조기에 평가하고, 합성생물학 실험실의 자동화에 기여하는 방향성을 조망하고자 한다.

2.2.1 마이크로플레이트 기반 자동 스크리닝 시스템

마이크로플레이트 리더는 합성생물학 고처리량 스크리닝의 핵심 분석 장비로 자리잡고 있다. 멀티모드 플레이트 리더는 96웰~384웰 포맷에서 형광, 흡광, 발광 측정을 통해 유전자 발현, 세포 성장, 대사산물 농도를 동시에 정량화한다.

광합성 미생물 배양을 위한 고처리량 조명 시스템은 웰별 독립 조명 제어와 실시간 측

정을 결합하여 배양 조건을 최적화한다. 배지 조성 최적화를 통해 성장 속도를 크게 향상시킨 사례가 보고되었으며, 엽록체 합성생물학 연구에서는 384웰 어레이 기반 형광 정량으로 140개 이상의 프로모터, UTR(Untranslated Region), 발현 조절 요소를 특성화하였다. 자동화 워크플로와 통합하여 수천 개 형질전환 균주의 병렬 분석을 구현한 사례도 증가하고 있다.

머신러닝 기반 최적화는 플레이트 리더의 효율성을 극대화하고 있다. 마이크로플레이트 발효 측정 데이터를 학습하여 조합 라이브러리의 3% 미만만 스크리닝하고도 대사산물 생산 수율을 60% 이상 개선하는 성과가 보고되었다. 이는 전체 라이브러리를 스크리닝하지 않고도 최적 균주를 선별할 수 있어, 시간과 비용을 대폭 절감하는 효과를 제공한다.

광유전학 적용을 위한 조명 통합형 플레이트 리더도 발전하고 있다. 96웰 LED 어레이 시스템은 각 웰에서 독립적인 광자극과 흡광 및 형광 측정을 동시 수행하며, 웰 간 교차 조명 없이 광유전학 회로의 고처리량 특성 분석을 가능하게 한다. 로봇 워크스테이션과 통합된 자동화 플랫폼은 조명 장치, 마이크로플레이트 리더, 진탕기를 연계하여 광 입력-리포터 발현 관계를 고처리량으로 스크리닝한다.

이러한 분석 장비 기술은 합성생물학 DBTL 사이클의 테스트 단계를 가속화하고, 유전자 회로 및 대사 경로의 상업화 가능성을 실험실 단계에서 조기 평가할 수 있는 인프라를 제공하고 있다. 특히 형광-흡광 동시 측정 기능과 온도/가스 제어 기능을 갖춘 멀티모드 플레이트 리더는 바이오파운드리 구축의 장비로 자리잡고 있다.

2.2.2 FACS 기반 단일세포 분석 및 선별 시스템

합성생물학과 바이오파운드리에서 단일세포 분석은 세포 집단의 이질성을 이해하고 최적 균주를 선별하는 핵심 기술이다. 마이크로플레이트 기반 분석이 평균적인 세포 특성을 측정한다면, 단일세포 분석은 개별 세포의 유전자 발현, 대사 활성, 생산성 차이를 정밀하게 파악할 수 있다. 동일한 유전자를 가진 세포라도 단백질 발현 수준, 대사 상태, 세포 주기 단계에 따라 표현형이 크게 다를 수 있으며, 이러한 세포 간 변이(cell-to-cell variation)를 정량화하는 것은 안정적인 생산 균주 개발에 필수적이다.

FACS(Fluorescence-Activated Cell Sorting)는 형광 신호를 기반으로 초당 수만 개의 세포를 분석하고 원하는 특성을 가진 세포만 물리적으로 분리하는 장비다. 세포 현탁액이 좁은 노즐을 통과하면서 레이저 빔을 만나면, 각 세포의 형광 신호와 산란광(scatter)이 검출된다. 전방산란광(FSC)은 세포 크기를, 측방산란광(SSC)은 세포 내부 복잡도를 나타내며, 여러 파장의 형광 채널을 통해 10개 이상의 파라미터를 동시에 측정할

수 있다. 원하는 형광 범위에 속하는 세포는 정전기적으로 편향되어 별도의 수집 용기로 분리되며, 이를 통해 수백만 개 규모의 군주 라이브러리에서 상위 0.1% 고성능 군주만 선택적으로 회수할 수 있다.

합성생물학에서 FACS의 핵심 응용은 형광 바이오센서를 활용한 고생산 군주 스크리닝이다. 목적 대사산물의 세포 내 농도에 반응하여 형광단백질을 발현하는 전사인자 기반 바이오센서를 설계하면, 형광 강도가 곧 생산성을 나타내는 지표가 된다.

바이오파운드리 환경에서 FACS는 완전 자동화된 워크플로의 일부로 통합 운영된다. 로봇 액체 핸들러가 96웰 또는 384웰 플레이트에서 배양된 군주 라이브러리를 자동으로 FACS에 주입하고, 선별된 세포는 다시 플레이트 형식으로 회수되어 자동 배양 시스템이나 차세대 시퀀싱(NGS) 분석으로 연결된다. Ginkgo Bioworks의 바이오파운드리기는 수십 대의 FACS 장비를 병렬 운영하며 주당 수백만 개의 군주 변이체를 스크리닝한다. 선별된 세포 데이터는 LIMS(Laboratory Information Management System)에 실시간으로 저장되며, 형광 분포 패턴, 세포 크기, 산란광 프로파일 등의 다차원 데이터가 머신러닝 알고리즘을 통해 분석되어 다음 스크리닝 라운드의 선별 기준 최적화에 활용된다. 최근에는 오픈소스 FACS 자동화 시스템이 개발되어 게이팅 알고리즘이 표준화되고 운영자 간 변이가 제거되면서, 고처리량 세포 공학 워크플로의 접근성이 크게 향상되었다.

최근에는 FACS와 미세유체(microfluidics) 기술을 결합한 차세대 플랫폼이 개발되고 있다. 액적 미세유체 시스템은 단일세포를 피코리터~나노리터 크기의 수용액 액적에 격리시켜, 각 세포가 독립적인 마이크로 배양기 역할을 하도록 한다. 이 액적들은 초당 수천 개씩 생성되며, 형광 검출기를 통과하면서 실시간으로 선별된다. 선별된 액적 내 세포는 용해되어 RNA나 DNA가 방출되고, 바코드 시퀀싱을 통해 수만 개 세포의 유전형과 표현형을 동시에 분석할 수 있다.

2.3 질량분석 기반 대사 흐름 분석과 자동화

합성생물학에서 대사 흐름을 정밀하게 이해하고 설계하는 것은 군주 개발 성공의 핵심이다. 질량분석(MS)은 대사 경로를 정량적으로 추적하고 재설계하는 데 가장 널리 활용되는 분석 기술이다. 특히 탄소-13(^{13}C) 표지 기질을 이용한 대사 플럭스 분석(^{13}C -MFA)은 세포 내 탄소 흐름을 수치 모델로 표현하여 대사 경로 재설계의 근거를 제공한다.

본 절에서는 이러한 질량분석 기술이 군주 설계-제작-검증-학습(DBTL) 자동화 사이클에 어떻게 통합되고 있는지 살펴본다. LC-MS 기반 군주 최적화, GC-MS를 활용한 가스 발

효 공정 최적화, 그리고 고해상도 질량분석과 디지털 트윈의 융합을 통해 질량분석이 바이오파운드리 기술의 핵심 분석 기반으로 진화하는 과정을 확인할 수 있다.

2.3.1 LC-MS 기반 균주 최적화

액체 크로마토그래피-질량분석법(LC-MS)은 세포 내 대사체를 정밀 분석하여 대사 경로의 병목 지점을 규명하고 균주 설계 타겟을 제시하는 핵심 분석 도구이다. 제노마티카(Genomatica)는 LC-MS/MS 기반 ^{13}C 대사 흐름 분석을 통해 대장균 내 1,4-부탄디올 생산 경로의 병목 효소를 규명하고 균주를 최적화하여 상업 규모 생산 공정을 구축하였다. 이후 대규모 동역학 모델(ORACLE 프레임워크)과 LC-MS 데이터를 통합하여 추가적인 대사 엔지니어링 타겟을 예측하는 방식으로 진화하였다.

최근 LC-MS는 초고속 분석 기술과 인공지능 데이터 해석이 결합되어 대규모 균주 스크리닝 플랫폼으로 발전하고 있다. 2024년 개발된 레이저 제거 전기분무 이온화 질량분석법(LAESI-MS) 플랫폼은 배양 플레이트에서 직접 지방산 생산 균주를 스크리닝하는 인시츄(in situ) 분석을 가능하게 한다. 최근 보고된 탈착 전기분무 이온화 질량분석 이미징(DESI-MSI) 기반 워크플로는 샘플 전처리 과정을 최소화하여 샘플당 1분 이내의 속도로 전체 대사 프로파일을 측정한다. 징코 바이오웍스는 고속 LC-MS 시스템과 기계학습을 통합한 플랫폼을 통해 수천 개 규모의 변이체 라이브러리에서 반복적인 설계-제작-시험-학습(DBTL) 사이클을 수행하여 균주 성능을 크게 향상시켰다. LC-MS 기반 분석은 합성생물학 균주 개발 워크플로에서 높은 활용도를 보이는 분석 기술로 자리잡고 있으며, 자동화 바이오파운드리 구축에 중요한 역할을 하고 있다.

2.3.2 GC-MS 및 오프가스 MS 기반 발효 공정 개선

기체 크로마토그래피-질량분석법(GC-MS)은 휘발성 대사산물과 가스상 기질·생성물을 실시간으로 분석할 수 있어, 가스 발효 및 광합성 기반 바이오공정 최적화에 필수적인 분석 도구로 활용되고 있다. 랜자텍(LanzaTech)은 온라인 GC-MS 시스템과 대사 흐름 모델링을 결합하여 일산화탄소(CO) 기반 가스 발효 공정을 최적화하였다. 플렉스 균형 분석(FBA)과 대사체 프로파일링을 통해 CO 공급 조건이 환원 페레독신 소비와 NADH 생성 경로에 미치는 영향을 정량화하였으며, 이를 기반으로 상업 규모 에탄올 생산 공정을 구축하였다. 2024년에는 CO 직접 전환 기반 영양 단백질 생산 공정을 발표하며, GC-MS 기반 공정 분석 플랫폼을 확장하고 있다.

최근 GC-MS는 실시간 오프가스 분석기(Off-gas Analyzer), 라만 분광기, 공정 분석 기

술(PAT) 센서와 통합되어 다중 모드 발효 모니터링 시스템으로 진화하고 있다. 2022년 시네코시스티스(Synechocystis) 광합성 연구에서는 GC-MS 기반 ^{13}C 대사 흐름 분석을 통해 균주 설계 시뮬레이션과 발효 조건 최적화의 피드백 루프를 구성하는 핵심 입력 값으로 활용된다. GC-MS와 오프가스 분석의 결합은 배양 환경 변화에 대한 미생물 대사 반응을 실시간으로 추적할 수 있다는 점에서, 자동화된 공정 제어 및 적응형 발효 전략 구현의 핵심 분석 기술로 활용되고 있다.

2.3.3 고해상도 질량분석과 디지털 트윈 융합

오르비트랩(Orbitrap) 및 푸리에 변환 이온 사이클로트론 공명 질량분석기(FT-ICR MS)는 높은 분해능과 질량 정확도를 바탕으로 ^{13}C -MFA 분석에서 세포 내 대사 흐름을 정밀하게 재구성한다. 동위원소 패턴 분석을 통해 탄소 원자 흐름을 정량화할 수 있으며, 이는 대사 경로 병목 탐지, 대사 모델 보정, 합성 회로 검증 등 균주 설계 전 단계에 활용된다. 2014년 개발된 IsoDesign 소프트웨어는 동위원소 라벨링 조성을 최적화하여 측정 가능한 플럭스 수와 정확도를 극대화하며, 합성생물학 균주 개발에서 대사 경로 재설계의 정량적 근거를 제공하는 핵심 도구로 자리잡고 있다.

이러한 고해상도 질량분석 데이터는 디지털 트윈 및 게놈 규모 대사 모델(Genome-Scale Metabolic Model, GEM)과 결합되어 인실리코 균주 설계 플랫폼으로 진화하고 있다. 징코 바이오웍스는 LC-MS/MS, GC-MS를 포함한 대규모 질량분석 시스템을 갖춘 자동화 바이오파운드리를 통해 대량의 대사체 및 단백질 데이터를 생성하며, 이를 계산생물학 모델과 기계학습 알고리즘에 통합하여 설계-제작-시험-학습(DBTL) 사이클을 가속화하고 있다. 바이오메이드(BioMADE)를 비롯한 글로벌 바이오파운드리 네트워크는 표준화된 질량분석 프로토콜과 대사 플럭스 데이터베이스 구축을 추진하여, 실험실 간 데이터 호환성을 확보하고 AI 기반 균주 설계 정확도를 높이는 방향으로 발전하고 있다. 고해상도 질량분석은 실험 데이터와 예측 모델을 연결하는 핵심 인터페이스로서, 설계-예측-검증이 통합된 자동화 생물학 연구 시스템의 기반 기술로 작동하고 있다.

2.4 실시간 분광 분석 기반 공정 분석 기술(PAT)

합성생물학 기반 바이오 제주는 설계-제작-시험-학습(DBTL) 주기의 반복을 통해 생물학적 회로, 세포주, 백신, 대사물질 등을 신속하게 개발 및 생산하는 것을 목표로 한다. 이러한 공정을 산업 수준으로 확장하기 위해서는 각 단계에서 발생하는 데이터의 실시간 수

집, 품질 변수의 정량적 예측, 공정 자율 제어가 필수적이다. 이를 가능하게 하는 핵심 기술이 공정 분석 기술(PAT)이다. PAT는 실시간 측정을 통해 공정을 이해하고 제어하는 시스템으로, FDA가 제시한 품질보증 전략의 핵심이며, 바이오 공정의 ‘블랙박스’를 ‘투명한 공정’으로 전환하는 역할을 한다.

라만 분광법을 활용한 PAT는 비파괴 인라인 측정으로 즉각적 피드백을 제공하며, 멸균 가능하고 시약을 사용하지 않는 무시약(reagent-free) 실시간 센싱이 가능하여 배양기 내부 주요 대사산물과 세포 특성을 예측하는 데 널리 사용된다. 특히 최근에는 미니 바이오리액터 시스템과의 통합을 통해 고속처리 공정 개발 플랫폼으로 발전하고 있으며, 실시간 분광 데이터를 바이오리액터 제어 시스템과 직접 연계하여 자동 피드백 제어를 구현하는 단계에 이르렀다. 본 절에서는 먼저 미니 바이오리액터에 라만 플로우 셀을 통합하여 다수의 조건을 병렬 실험하고, 구축된 예측 모델을 생산 규모로 전환하는 최신 동향을 살펴본다. 이어서 표준 통신 프로토콜을 통해 실시간 라만 분광 데이터를 바이오프로세스 제어 시스템과 연계하여 주요 대사물 농도를 정밀 유지하는 자동 제어 기술과, 머신러닝 기반 고급 예측 모델을 통한 조기 공정 이상 감지 기술의 발전 동향을 검토한다. 이를 통해 PAT가 합성생물학 자동화를 위한 핵심 인프라로서 설계 정보와 실제 공정 반응 간 피드백 루프를 구성하고, 품질 중심 공정 설계(QbD) 실현과 무인 적응형 공정 제어를 가능하게 하는 최신 기술 동향을 파악하고자 한다.

2.4.1 미니 바이오리액터 라만 분광 분석

라만 PAT는 비파괴 인라인 측정으로 즉각적 피드백을 제공한다는 점에서 차별화된다. 미니 바이오리액터 시스템은 교체 가능한 라만 플로우 셀을 분석 모듈의 샘플 컵에 연결하는 방식으로 통합되며, 액체 핸들러가 각 미니 바이오리액터에서 자동으로 소량의 샘플을 채취하여 실시간 분광 측정을 수행한다. 2022년 연구에서는 미니 바이오리액터 시스템에 프로토타입 라만 플로우 셀과 바이오분석기를 통합하여, 라만 측정과 기준값 측정 간 시간 간격을 5분 이내로 단축함으로써 스펙트럼 변화와 실제 대사물 농도를 직접 상관시킬 수 있는 신뢰성 높은 모델 구축이 가능해졌다. 단일 실험에서 다수의 상이한 바이오리액터 조건이나 공정 궤적으로부터 데이터를 생성할 수 있어, 수백 회의 생산 배치를 무작위 변동으로 실행하는 것보다 훨씬 큰 설계 공간을 확보하고 통계적으로 유의미한 데이터를 획득하여 저비용으로 예측 라만 모델을 신속하게 구축할 수 있다.

이러한 통합 플랫폼의 핵심 장점은 공정 개발 단계에서 구축된 모델을 제조 규모로 원활하게 전환할 수 있다는 점이다. 최근 연구에서는 미니 바이오리액터의 플로우 셀 측정과 교

반 탱크 바이오리액터의 인라인 측정 간 양방향 모델 전환에 성공하여, 포도당과 젖산 등 주요 대사물에 대한 라만 분광 모델이 개발 단계에서 생산 단계까지 효율적으로 적용될 수 있음을 입증했다. 고속처리 라만 분광은 소량 샘플만으로 주요 공정 변수(CPP)와 품질 속성(CQA)을 예측할 수 있어, 여러 분석기에 대한 의존도를 낮추고 자본 및 운영 비용을 절감하면서도 제품 농도 같은 모니터링이 어려운 변수에 대한 실시간 정보를 제공한다.

2.4.2 실시간 라만 분광의 자동 제어 시스템 통합

실시간 라만 분광 데이터를 바이오리액터 제어 시스템과 직접 연계하여 자동 피드백 제어를 구현하는 기술이 합성생물학 공정 자동화의 핵심 기술로 발전하고 있다. 2024년 연구에서는 라만 분석기를 표준 통신 프로토콜을 통해 바이오프로세스 제어 시스템과 통합하여, CHO 세포 배양 중 포도당 농도를 5g/L 목표값에서 오차 $\pm 0.2\text{g/L}$ 이내로 3일간 정밀 유지하는 데 성공했다. 라만 분석기는 30분 간격으로 스펙트럼을 자동 수집하고 편광 최소 자승(PLS) 모델을 통해 포도당 농도를 실시간 예측하며, 이 데이터가 제어 소프트웨어로 전송되어 프로그래밍된 피드백 루프가 펌프 속도를 자동 조정한다. 2025년 최신 연구에서는 라만과 정전용량 분광을 결합하고 랜덤 포레스트, 그래디언트 부스팅 같은 머신러닝 기법을 적용하여, 기존 선형 회귀 방법 대비 세포 생존율 모니터링 정확도를 31.7%, 생존 세포 밀도 예측 정확도를 27.3% 향상시켰으며, 특히 90% 이상의 높은 생존율 구간에서 조기 공정 이상 감지 능력이 개선된 것으로 나타났다.

이러한 자동 제어 통합 기술은 2.3에서 다룬 오프라인 질량분석 기반 정밀 대사체 프로파일링과 상호 보완적인 역할을 수행한다. 질량분석이 배치별 샘플 채취를 통해 수백 개 대사물의 정밀 정량과 대사 경로 병목 구간 발견에 강점을 보이는 반면, 실시간 라만 분광은 비파괴 인라인 측정을 통해 즉각적인 피드백과 공정 제어를 가능하게 한다.

2024년 연구에서는 CNN(합성곱 신경망) 기반 케모메트릭스 모델이 NIR 분광 데이터를 처리하여 대장균 발효 공정에서 20개 이상의 저농도 분석물을 실시간 정량화하는 성과를 보였다. 이는 다중 대사물 동시 모니터링의 자동화 가능성을 제시했다.

전기화학 바이오센서 같은 보조 기술은 특정 대사물(포도당, 젖산)에 대해 저비용 및 고속 응답성을 제공한다. 이러한 센서는 라만 분광과 함께 멀티센서 융합 시스템을 구성할 가능성을 보유하고 있다. 그러나 현재까지는 선택성과 센서 수명 문제로 인해 합성생물학 균주 스크리닝보다는 단순 공정 모니터링에 제한적으로 활용되고 있다.

향후 실시간 분광 기술과 AI 기반 예측 모델의 결합은 합성생물학 자동화 플랫폼에서 무인 적응형 공정 제어를 실현하는 주요 기술로 발전할 것으로 전망된다.

3

맺음말

본 보고서에서는 합성생물학 연구개발의 핵심 단계별로 요구되는 분석 기술과 자동화 솔루션을 종합적으로 검토하였다. 차세대 염기서열 분석(NGS) 기술은 DNA 구조체 검증의 병목을 해소하며 바이오파운드리 운영의 기반이 되고 있으며, 마이크로플레이트 리더와 유세포 분석기를 중심으로 한 고처리량 스크리닝(HTS) 기술은 수천 개 균주의 병렬 평가를 통해 DBTL 사이클을 가속화하고 있다. 질량분석(MS) 기반 대사체학은 세포 내 탄소 흐름을 정량화하여 대사 경로 재설계의 근거를 제공하며, 라만 분광법을 비롯한 실시간 공정분석기술(PAT)은 공정 데이터를 제어 시스템과 연계하여 적응형 바이오제조를 구현하고 있다.

표 1> 합성생물학 자동화를 위한 분석장비별 핵심 응용 및 기술

분석장비	핵심 응용	기술 요약
차세대 염기서열 분석기 (NGS)	DNA 구조체 검증 자동화	Illumina NovaSeq/MiSeq 기반 대량 병렬 검증, AWS 클라우드 파이프라인, ML 기반 오류 학습
	품질관리 시스템	Nextera 기반 고처리량 시퀀싱, 저비용 전수 검증, 바이오파운드리 표준화
장문형 시퀀싱 장비	복잡한 구조 검증	PacBio SMRT 고정확도 장문형 분석, Oxford Nanopore MinION 현장 실시간 검증
멀티모드 플레이트 리더	고처리량 균주 스크리닝	형광/흡광/발광 다중 측정, 고밀도 웰 병렬 분석, 온도/가스 환경 제어
	광유전학 회로 분석	LED 어레이 독립 광자극, 흡광/형광 동시 측정, 웰별 조명 제어
	ML 최적화 스크리닝	최소 라이브러리 스크리닝으로 최적 균주 선별, 시간/비용 절감
형광 활성 세포 분리 시스템 (FACS)	단일세포 수준 균주 선별	초고속 세포 분석 및 물리적 분리, 형광 바이오센서 기반 고생산 균주 스크리닝
	자동화 워크플로 통합	로봇 액체 핸들러 연계, LIMS 데이터 통합, ML 기반 선별 기준 최적화
	액적 미세유체 융합	단일세포 액적 격리, 바코드 시퀀싱, 유전형-표현형 동시 분석
액체 크로마토그래피 질량분석기 (LC-MS)	대사 경로 병목 분석	¹³ C 표지 대사 플럭스 분석, Genomatica 부탄디올 균주 최적화 사례
	초고속 균주 스크리닝	LAESI-MS 인시츄 분석, DESI-MSI 고속 프로파일링, 전처리 최소화
기체 크로마토그래피 질량분석기 (GC-MS)	가스 발효 공정 최적화	휘발성 대사산물 실시간 분석, LanzaTech CO ₂ 발효 균형 분석
	오프가스 통합 분석	¹³ C 대사 흐름 추적, 실시간 대사 반응 모니터링
고해상도 질량분석기	정밀 동위원소 분석	Orbitrap/FT-ICR 기반 ¹³ C-MFA, IsoDesign 실험 설계 최적화
	디지털 트윈 융합	계능 규모 대사 모델 연동, 대규모 오믹스 데이터 생성, 실험-예측 자동화 루프
라만 분광기	실시간 공정 제어	비파괴 인라인 측정, 바이오리액터 자동 피드백 제어, CHO 배양 영양소 정밀 유지
	미니 바이오리액터 통합	고속처리 공정 개발, 라만 플로우 셀 자동 측정, 생산 규모 모델 전환
근적외선 분광기 (NIR)	다중 대사물 정량	CNN 케모메트릭스 모델, 대장균 발효 다중 분석물 실시간 모니터링
멀티센서 융합 시스템	ML 기반 공정 이상 감지	라만+정전용량 분광 결합, Random Forest 기반 세포 생존율 예측 정확도 향상

글로벌 바이오파운드리 선도 기업들은 이러한 분석 기술을 중심으로 자동화 플랫폼을 구축하고 있다. NGS 시스템, 유세포 분석기, 질량분석기, 분광 센서 등을 통합하여 대규모 실험 반복과 데이터 기반 설계를 수행하며, 특히 인공지능 기반 예측 모델 및 디지털 트윈과 결합하여 설계 정보와 실제 공정 반응 간 피드백 루프를 구성함으로써 품질 중심 공정 설계(QbD) 구현과 생산 신뢰성 확보에 기여하고 있다.

국내에서도 합성생물학 연구 인프라 구축이 진행되고 있으나, 핵심 분석 장비 상당수를 해외 제품에 의존하고 있는 것이 현실이다. 일루미나(Illumina)의 NGS 시스템, 써모피셔(Thermo Fisher)와 사이텍스(Cytiva)의 바이오리액터 및 분광 분석 장비, 애질런트(Agilent)와 워터스(Waters)의 질량분석기 등이 국내 바이오파운드리와 연구기관의 주요 분석 인프라를 구성하고 있다. 이는 장비 도입 비용뿐만 아니라 유지보수, 부품 교체, 소프트웨어 업그레이드 등에서 발생하는 지속적인 비용 부담과 기술 지원 지연으로 인한 연구 개발 효율성 저하로 이어지고 있다.

다만 최근 국내에서도 일부 영역에서 의미있는 기술 개발 노력이 진행되고 있다. 분광분석 영역에서는 다중모드 마이크로플레이트 리더, 라만 분광기, NIR 분석 장비의 국산화가 시도되고 있으며, 질량분석 분야에서도 일부 기업은 GC-MS와 LC-MS 시스템 개발 기술을 보유하고 있다.

특히 인공지능 기반 분석 소프트웨어 및 데이터 처리 플랫폼 개발에서는 국내 정보기술 역량을 활용한 경쟁력 확보 가능성이 높다. 유세포 분석, 고처리량 스크리닝 장비 등의 영역에서도 기술 개발이 시도되고 있으나, 글로벌 선도 기업 대비 기술 격차가 존재하며 상용화 및 시장 진입에 어려움을 겪고 있다.

국내 분석장비 산업의 기술 수준 향상은 단순한 수입 대체를 넘어 국내 바이오파운드리 생태계의 경쟁력 강화로 이어질 수 있다. 이를 위해서는 산학연 협력을 통해 실제 현장 요구를 반영한 실증 환경을 조성하고, 바이오파운드리 간 데이터 표준 및 공유 체계를 마련하여 분석장비 제조사-소프트웨어 개발사-최종 사용자가 참여하는 협력 플랫폼을 구축하는 것이 중요하다.

합성생물학이 지속 가능한 바이오 경제의 핵심 기술로 발전하기 위해서는, 본 보고서에서 다룬 분석 기술들의 지속적인 개선과 함께 이를 뒷받침하는 산업 생태계 구축이 병행되어야 한다. 분석 기술은 합성생물학 자동화의 핵심 인프라이자 바이오 신약, 지속가능 화학물질, 차세대 연료 등 전략 산업 발전의 기반 기술이다. 단기적으로는 해외 선도 장비의 효율적 도입과 활용을 통한 연구개발 역량 강화를, 중장기적으로는 국내 분석장비 기술 자립도 향상과 글로벌 경쟁력 확보를 동시에 추진해야 한다. 정부 차원의 전략적 R&D

투자, 바이오파운드리 인프라 구축 사업과 분석장비 개발의 연계, 국제 협력 네트워크 구축이 체계적으로 이루어질 때 국내 합성생물학과 분석장비 산업이 동반 성장하는 선순환 구조를 만들어갈 수 있을 것이다.

[참고문헌]

- [1] Shapland EB et al., (2015). Low-Cost, High-Throughput Sequencing of DNA Assemblies Using a Highly Multiplexed Nextera Process. *ACS Synthetic Biology*, 4(7), 860-866.
- [2] Chin CS et al., (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, 10, 563-569.
- [3] Ardui S et al., (2018). Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Research*, 46(5), 2159-2168.
- [4] Jain M et al., (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, 17, 239.
- [5] Tyler AD et al., (2018). Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. *Scientific Reports*, 8, 10931.
- [6] Currin A et al., (2019). Highly multiplexed, fast and accurate nanopore sequencing for verification of synthetic DNA constructs and sequence libraries. *Synthetic Biology*, 4(1), ysz025.
- [7] Gerhardt KP et al., (2023). High-throughput feedback-enabled optogenetic stimulation and spectroscopy in microwell plates. *Communications Biology*, 6, 1190.
- [8] Gerhardt KP et al., (2016). An open-hardware platform for optogenetics and photobiology. *Scientific Reports*, 6, 35363.
- [9] Baret JC et al., (2009). Fluorescence-activated droplet sorting (FADS): efficient microfluidic cell sorting based on enzymatic activity. *Lab on a Chip*, 9(13), 1850-1858.
- [10] Huynh E et al., (2024). An open-source FACS automation system for high-throughput cell biology. *PLoS One*, 19(3), e0299402.
- [11] Rogers JK & Church GM, (2016). Genetically encoded sensors enable real-time observation of metabolite production. *Proceedings of the National*

- Academy of Sciences, 113(9), 2388-2393.
- [12] Klein AM et al., (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5), 1187-1201.
 - [13] Macosko EZ et al., (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5), 1202-1214.
 - [14] Zhang Z et al., (2023). High-throughput screening of microbial strains in large-scale microfluidic droplets. *Frontiers in Bioengineering and Biotechnology*, 11, 1105277.
 - [15] Long CP et al., (2019). High-resolution ¹³C metabolic flux analysis. *Nature Protocols*, 14, 2856-2877.
 - [16] Zemaitis KJ et al., (2024). Fourier Transform Ion Cyclotron Resonance Mass Spectrometry Applications for Metabolomics. *Biomedicines*, 12(8), 1786.
 - [17] Millard P et al., (2014). IsoDesign: A software for optimizing the design of ¹³C-metabolic flux analysis experiments. *Biotechnology and Bioengineering*, 111(1), 202-208.
 - [18] Graf A et al., (2022). Automated Data Generation for Raman Spectroscopy Calibrations in Multi-Parallel Mini Bioreactors. *Sensors*, 22(9), 3397.
 - [19] Domján J et al., (2022). Real-time amino acid and glucose monitoring system for the automatic control of nutrient feeding in CHO cell culture using Raman spectroscopy. *Biotechnology Journal*, 17(1), 2100395.
 - [20] Banerjee S et al., (2024). NIR spectroscopy-CNN-enabled chemometrics for multianalyte monitoring in microbial fermentation. *Biotechnology and Bioengineering*, 121(6), 1803-1819.
 - [21] Abu-Absi SF et al., (2022). The role of Raman spectroscopy in biopharmaceuticals from development to manufacturing. *Analytical and Bioanalytical Chemistry*, 414(2), 969-991.
 - [22] Yang C et al., (2021). Application of near-infrared spectroscopy technology in the complex fermentation system to achieve high-efficiency production. *Bioresources and Bioprocessing*, 8(1), 96.

[주요용어풀이]

- 합성생물학(Synthetic Biology): 생물학적 시스템을 공학적 원리로 설계하고 구축하는 융합 학문 분야
- DBTL(Design-Build-Test-Learn): 설계·제작·시험·학습의 네 단계를 반복하며 생물학적 시스템을 최적화하는 합성생물학 개발 프레임워크
- 바이오파운드리(Biofoundry): DBTL 자동화를 위해 로봇, 분석장비, 정보시스템을 통합한 합성생물학 전주기 자동화 플랫폼
- NGS(Next-Generation Sequencing, 차세대 염기서열 분석): 수백만~수십억 개의 DNA 서열을 병렬로 고속 분석하는 차세대 염기서열 분석 기술
- PAT(Process Analytical Technology, 공정 분석 기술): 실시간 센서, 데이터 분석, 모델 기반 제어를 결합해 바이오공정을 모니터링·예측·최적화하는 공정 분석·관리 프레임워크
- 생어 시퀀싱(Sanger Sequencing): 디디엔티피(ddNTP, Dideoxynucleotide Triphosphate)를 이용한 전통적인 DNA 염기서열 분석 방법
- 장문형 시퀀싱(Long-Read Sequencing): 수 kb에서 수백 kb 길이의 긴 DNA 서열을 한 번에 읽어 구조 변이와 반복 서열을 분석할 수 있는 시퀀싱 기술
- SMRT(Single Molecule Real-Time sequencing): 팩바이오(PacBio)의 단일분자 실시간 염기서열 분석 기술
- 고처리량 스크리닝(HTS, High-Throughput Screening): 수백~수천 개 조건을 자동화 장비로 병렬 실험해 기능이 우수한 후보를 선별하는 기술
- 마이크로플레이트 리더(Microplate Reader): 96~1536웰 플레이트 내 샘플의 흡광·형광·발광 신호를 자동 측정하는 분석 장비
- 광유전학(Optogenetics): 빛으로 제어 가능한 광수용체 단백질을 이용해 세포 활성을 정밀하게 조절하는 기술
- LED(Light-Emitting Diode, 발광 다이오드): 전류를 빛으로 변환하는 반도체 소자
- 유세포분석기(Flow Cytometer): 흐름 상태의 단일세포에 레이저를 조사해 세포의 물리·형광 특성을 고속 정량 분석하는 장비
- 마이크로플루이드스(Microfluidics): 미세 유체 채널을 이용해 세포·분자를 소량·정밀하

게 조작·제어하는 기술

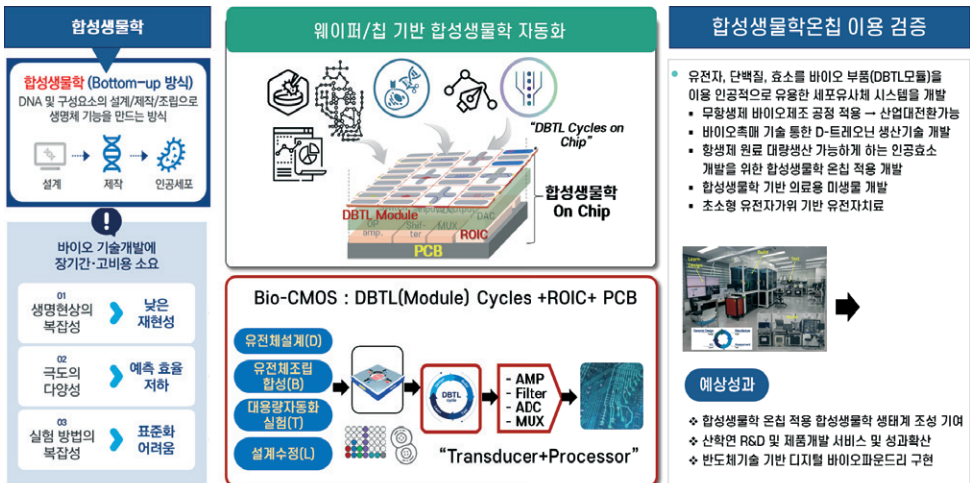
- FACS(Fluorescence-Activated Cell Sorting, 형광 활성화 세포 분류): 형광 표지된 세포를 유세포분석기로 고속 분류·선별하는 기술
- MS(Mass Spectrometry, 질량분석): 이온의 질량-전하비(m/z)를 측정해 화합물의 구조와 농도를 분석하는 기술
- ^{13}C -MFA(^{13}C Metabolic Flux Analysis): 탄소-13 동위원소 라벨 기질을 이용해 세포 내 대사 경로의 탄소 흐름을 정량 모델링하는 분석 기법
- LC-MS(Liquid Chromatography-Mass Spectrometry): 액체 크로마토그래피로 분리한 화합물을 질량분석으로 정성·정량 분석하는 기술
- LAESI-MS(Laser Ablation Electrospray Ionization-Mass Spectrometry): 레이저로 시료를 제거하고 전기분무 이온화 후 질량분석하는 고속 분석 기술
- DESI-MSI(Desorption Electrospray Ionization-Mass Spectrometry Imaging): 탈착 전기분무 이온화를 이용한 질량분석 이미징 기술
- GC-MS(Gas Chromatography-Mass Spectrometry): 기체 크로마토그래피로 분리된 휘발성 화합물을 질량분석으로 정성·정량 분석하는 기술
- FBA(Flux Balance Analysis, 플럭스 균형 분석): 대사 네트워크의 화학양론적 제약 조건을 기반으로 대사 흐름을 예측·최적화하는 계산 기법
- Orbitrap: 이온을 전기장에서 궤도 운동시키며 초고분해능 질량 스펙트럼을 획득하는 질량분석기
- FT-ICR MS(Fourier Transform Ion Cyclotron Resonance Mass Spectrometry): 푸리에 변환 이온 사이클로트론 공명 원리를 이용한 초고분해능 질량분석기
- iMFA(Isotope-assisted Metabolic Flux Analysis): 동위원소 보조 대사 플럭스 분석
- 라만 분광법(Raman Spectroscopy): 레이저 조사 시 발생하는 라만 산란 신호를 이용해 분자의 진동 상태와 화학 조성을 분석하는 분광 기술
- QbD(Quality by Design): 사전 설계와 공정 제어를 통해 제품 품질을 일관되게 확보하는 품질 중심 설계 개념
- 미니 바이옱터(Mini Bioreactor): 소규모 병렬 배양 실험을 위한 자동화된 소형 생물반응기 시스템

- PLS(Partial Least Squares, 부분 최소 자승법): 다변량 데이터 분석에서 예측 모델을 구축하는 통계 기법
- CNN(Convolutional Neural Network, 합성곱 신경망): 이미지나 스펙트럼 데이터 처리에 특화된 딥러닝 신경망 구조
- NIR 분광법(Near-Infrared Spectroscopy): 근적외선 영역의 흡광 스펙트럼을 이용해 용액 내 성분 농도와 물성을 비침습적으로 예측하는 분광 기술
- 디지털 트윈(Digital Twin): 실제 바이오공정의 상태와 거동을 디지털 모델로 실시간 모사·예측해 공정 설계와 제어에 활용하는 기술
- 균주 엔지니어링(Strain Engineering): 유전자 삽입·삭제·발현 조절 등을 통해 미생물 균주의 생산성·안정성·기능을 설계·최적화하는 기술

합성생물학온칩 구현 위한 모듈화기술

1 머리말

바이오산업은 디지털·바이오 전환시대를 맞이하여 기존 아날로그(Try and Error, 경험)중심에서 공학적 부품·모듈 개념을 도입, 디지털 기반으로 생명체 구성요소 및 시스템을 설계 제작하는 방식으로 패러다임이 전환하고 있음. 반도체 기술은 바이오분야의 디지털 산업 전환과 첨단바이오 부품화 적응을 위하여 요구되는 반도체 기반의 소형화, 집적화, 자동화, 일체형이 가능한 웨이퍼/칩 기반으로 생명체의 구성요소 (Design, Build, Test, Learning)을 제작한다면 기존의 낮은 재현성, 예측 효율 저하, 표준화, 모듈화 어려움에 대한 문제를 혁신적으로 해결이 가능함으로 합성생물학온칩(Synthetic Biology On Chip, SBoC) 플랫폼 확보를 통해 중요한 바이오파운드리 설계인자 도출, 결과 검증/예측이 가능한 기술 구현이 가능한 기술로 판단이 된다. 구체적으로는 AI·로봇을 접목하여 합성생물학의 DBTL사이클을 표준화·고속화·자동화 인프라 설계시 SBoC 플랫폼을 통해 설계인자를 도출하고 테스트베드처럼 활용한다면 실험단계 및 기존 워크플로 개발시 각 DBTL을 라이브러리화 된 모듈화 개념을 도입하여 구현한다면 기존 바이오파운드리에서 제한점인 시험부피, 시험개수(이론적으로는 수만~백만 병렬 동시 실험가능), 시접목이 가능한 디지털 데이터 도출 통한 재현/예측력 증가가 가능한 플랫폼으로 판단이 된다. 특히 Build (예, DNA 라이브러리 기반 균주 제작) 모듈, Test (예, 단백질체, 대사체, 유전자발현, 단백질-단백질 상호작용 검사 가능한 센서 모듈) 모듈을 도입한다면 기존의 상용화 시간, 비용절감이 가능하여 기술주도권 확보 가능

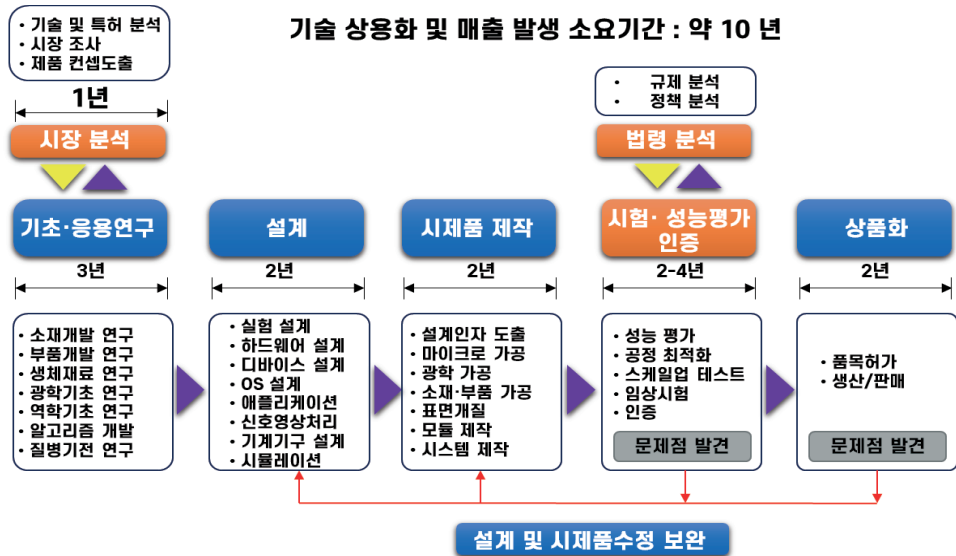


2 모듈화 기술

2.1 모듈화 기술 개요 및 필요성

글로벌 질병진단 시장은 개인 맞춤형 예방·진단으로 변화되고 있으며, 융·복합기술형 제품에 의한 시장 확대 가속화에 따라, 기능/일체형 중심에서 모듈형/조립형으로 패러다임이 변화되고 있으며 기존 기능/일체형 바이오칩은 단위부품으로서, 하나의 기능이 실패할 경우 전체 바이오칩 기능이 실패하게 되어, 제품개발 및 상용화 과정에서 장시간·고비용의 문제점이 발생한다(그림 2).

그림 2> 일반적인 바이오칩 제품 상용화 과정 모식도



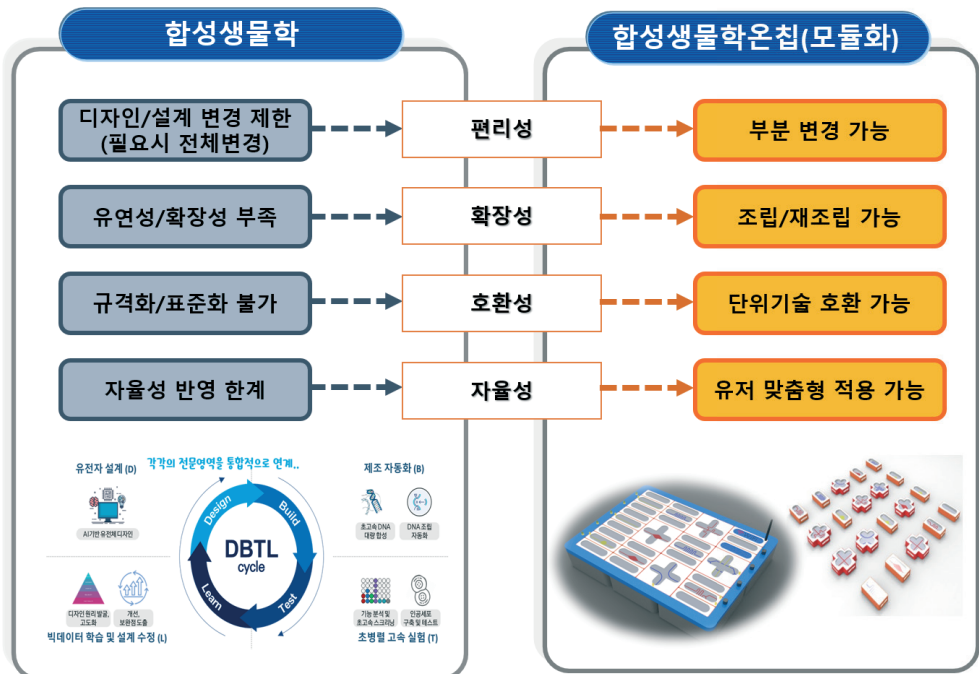
이에, 반도체 기반의 소형화, 집적화, 자동화, 일체형이 가능한 웨이퍼/칩 기반으로 생명체의 구성요소 (Design, Build, Test, Learning)을 제작한다면 기존의 낮은 재현성, 예측 효율 저하, 표준화 어려움에 대한 문제를 혁신적으로 해결이 가능함으로 합성생물학온칩(Synthetic Biology On Chip, SBoC) 구현이 가능한 기술이 될수 있다.

현재 바이오칩의 경우 하나의 칩 위에 다양한 바이오 기능들을 구현할 수 있기 때문에 기존의 랩온어칩(Lab-on-a-chip) 기술에 비해 바이오 콘텐츠 (합성생물학, 질병진단 등) 구현에 효율적으로 사용 될 수 있음. 그러나 ❶ 이들을 작동시키기 위해서는 바이오칩 전문가만이 다룰 수 있는 시린지 펌프와 같은 다양한 칩 밖의 요소들이 필요하며, ❷ 특히,

합성생물학온칩 구현을 위해서는 다양한 사용자의 필요에 따라 전체적인 DBTL 구성을 바꿀수 있어야 하고 최적의 DBTL 사이클을 구현하기 위한 효율성과 경제성이 떨어짐. 따라서, ③ 현재 바이오칩 및 랩온어칩 기술 보다 더 사용자 친화적이고 사용자 맞춤 형태로 발전시키는데 한계가 있음에 따라서 최근 합성생물학의 실험(DNA합성 라이브러리화, 생체측매 유도진화, 인공미생물 개량 및 제작, 발표 최적화, 새로운 프로토콜 개발 등) 최적화에 적용할수 있는 모듈화 기술을 재조명 할 필요가 있는 것으로 판단이 된다. 이는 기존 바이오실험에서 사용되는 올리고머 합성/세포 도입 ▶ 세포배양 ▶ 효과분석 ▶ NGS 및 균주규명 등 전체 실험실을 고속·고효율·최적화 가능한 기

존의 랩온어칩 개념과 다른 새로운 플랫폼으로 확장이 가능할것으로 판단이 된다. 모듈화 기술은 사용자 맞춤형 온칩 형태로 합성생물학 실험이 가능한 기술이며 합성생물학의 DBTL 등의 기능을 단위 모듈로 분할하여 각각의 모듈칩으로 설계하는 것을 특징으로 하며 각 단위 기능별 모듈이 자유롭게 조립 및 연동가능한 장점을 지님. 각 단위 기능별 모듈의 제작(본딩/플랫폼) 기술과 함께 시료 혼합 및 농축, 단위기능 모듈칩 간 인터페이스, 유량 조절 기술 등을 포함하는 모듈칩 제조 기술 등이 포함된다(그림 3).

그림 3> 모듈화기술의 기존 기술 대비 차별성

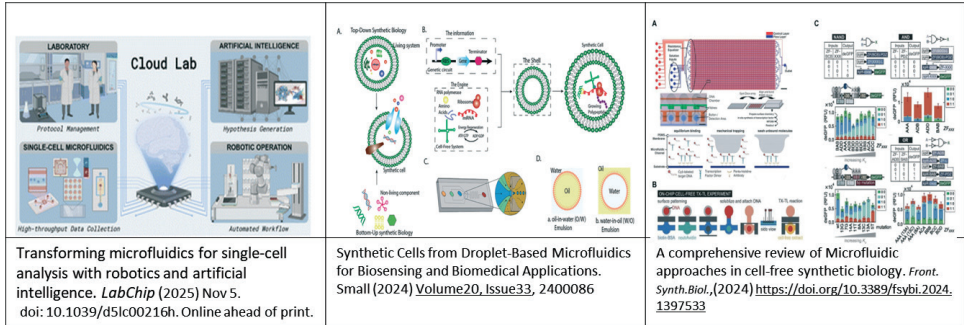


따라서 모듈형 합성생물학온칩은 각 단위기술별 모듈을 자유롭게 조립 및 연동이 가능하여 사용자의 수요에 따라 자유로운 구성이 가능하며, 대량생산을 통해 저렴한 가격으로 제공이 가능한 장점을 갖는다. 다양한 합성생물학 실험자는 새로운 아이디어가 발생했을 경우, 제공되는 모듈칩 라이브러리를 활용하여 신속하고 저렴하게 아이디어를 구현해 볼 수 있다. 따라서, 모듈형 합성생물학온칩은 생명과학에 공학적 기술개념을 도입하여, 인공적으로 생명체의 구성요소, 시스템을 설계·제작·합성하는 기술분야를 연구개발하거나 상용화 하고자 하는 연구자 및 기업에게 비용과 시간을 획기적으로 절감할 수 있는 기회를 제공하여 사업화 성과 창출 확대 가능할것으로 기대된다.

2.1 합성생물학온칩 모듈화 기술 동향

합성생물학 기술역량은 세계 최고 기술수준인 미국과 비교했을 때 2020년 기준 국내 수준은 75% 정도 (2030년 90% 수준 목표)이며 아직 합성생물학온칩 관련 핵심기술은 선도국인 미국에서조차도 거의 발표되고 있지 않은 상황으로 판단이 된다. 기존 바이오연구의 고질적인 문제로 지적되는 낮은 속도, 낮은 효율성의 한계를 극복하기 위해 세포개량 및 대사최적화 연구를 현재 랩온어칩, 미세유체칩 기술을 적용하는 연구 움직임이 활발하며 아직 모듈형 합성생물학온칩 개념의 연구는 미미한 상황으로 분석이 된다 (그림 4). 우리의 강점기술인 대사공학, 발효공학을 모듈형 합성생물학온칩 기술을 통해 세계최고 수준으로 확보한다면 글로벌 난제를 해결하고 신시장을 창출할 수 있는 의료분야 혁신, 고부가 바이오소재 생산 등에 합성생물학온칩 기술을 적용이 가능할것으로 판단이 된다. 합성생물학은 인공적으로 생명 시스템을 설계, 제작(조립), 합성하는 분야다. DNA나 RNA 같은 핵산(유전물질), 유전체(게놈), 단백질 등을 합성해서 자연계에 존재하지 않는 인공 생명체(인공 세포 혹은 미생물)를 만들고 여기서 백신이나 의약품 개발에 필요한 소재를 제조하는 것이다. 현재 단순 생명체의 경우 인공적 제작이 가능한 단계까지 기술이 발전했다. 최근에는 미국에서 세계 최초로 번식이 가능한 진정한 의미의 인공 생명체(세포)가 탄생하기도 했다. 유용한 기능을 하는 인공 생명체를 비교적 자유롭게 설계, 제작해 우리가 원하는 바이오소재 생산에 활용하는 시대가 도래한 것이다.

그림 4> 미세유체칩 활용 합성생물학 연구 최신 동향



특히, 바이오파운드리라는 이런 합성생물학의 전 과정을 표준화, 자동화, 고속화한 시스템이다. 인공지능(AI)과 로봇 기술을 기반으로 설계(Design)-제작(Build)-검증(Test)-학습(Learn)의 과정을 자동화해 생물학 실험과 제조 공정을 수행하는 ‘바이오 자동화 플랫폼’이라 할 수 있다. 지난 ‘스마트 세포공장’이라고 할 수 있다. 이는 최근 엔비디아 CEO 젠슨황 (Jensen Huang)이 소개한 반도체산업의 AI Factory와 비슷한 개념이라 할 수 있다. 이는 기본적으로 AI Factory는 데이터를 수집하고, 모델을 학습시키고, 환경을 시뮬레이션하고, AI를 제품에 배포하도록 설계된 가상 인프라 시스템이다. (혁신 가속화 가능) AI 팩토리를 사용하면 수천 개의 모델을 동시에 빠르게 반복하고 테스트할 수 있으므로 출시 시간을 단축하고 제품 주기를 개선할 수 있다. (운영 효율성) AI 파이프라인을 일상적인 워크플로에 통합하면 수작업이 줄어들고 휴먼 에러가 제거되며 복잡한 의사 결정이 자동화된다. (확장성) AI Factory는 로컬에서 글로벌, 현장에서 클라우드로 워크플로우를 원활하게 확장하여 기업이 민첩성을 유지할 수 있게 한다. 이와 유사하게 바이오파운드리를 활용하면 바이오 연구의 오랜 난제였던 속도, 스케일(규모), 불확실성의 한계를 극복할 수 있다. 짧은 기간에 생산성은 확 끌어올리고 비용은 크게 줄일 수 있는 것이다. 우리는 반도체산업에서 AI Factory 도입과 마찬가지로 바이오산업에서 바이오파운드리 도입을 앞두고 있다. 이 플랫폼은 더 빠른 혁신, 운영 효율성, 확장성 제공하는 동시에 기업은 더 스마트하게 적응력이 뛰어난 향후 AI 기반 바이오셋플 및 서비스를 개발하여 경쟁 우위를 확보할 수 있다고 판단이 된다. 다만 바이오파운드리의 제한점은 고가의 자동화, 연계장치 설치 위한 비용, 복잡성에 제한점을 본 저자는 합성생물학온칩 모듈화기술을 통해 극복한다면 거대기술 기업이 독식하고 있는 바이오산업을 모든 규모의 비즈니스 위한 혁신의 미래 도래가 가능할것으로 판단이 된다.

2.1 모듈화 기술 개발 통한 다양한 바이오응용 결과 소개

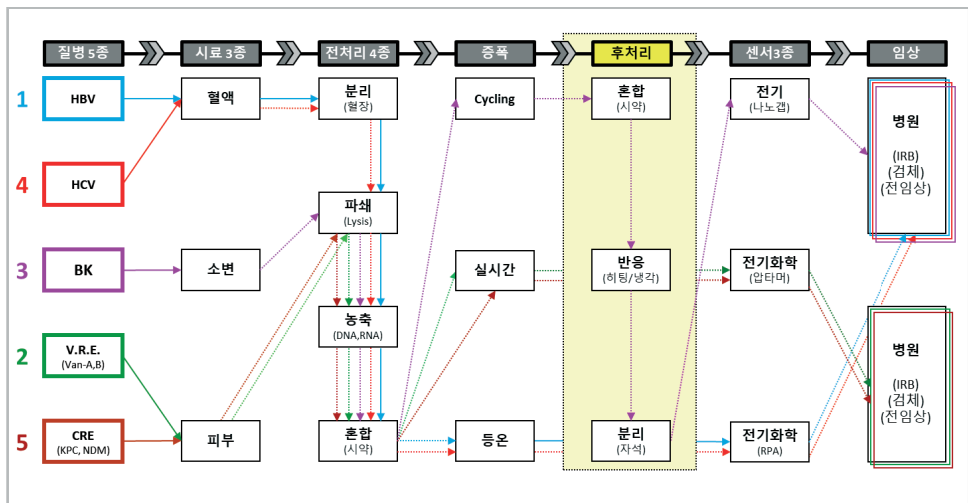
※ 본 지면에서 소개하는 연구결과는 본 저자가 개발 책임자로 수행한 분자진단용 전기 나노바이오센서 모듈화기술 개발(나노소재기술개발사업, 원천기술개발사업, '17~'22))결과 및 기관사업 수행결과를 발췌하여 독자의 합성생물학온칩 개념의 이해를 돕고자 함.

2.1.1 모듈화기술 적용 통한 분자진단키트 개발 사례

2.1.1.1 생체 시료전처리 구성 요소 분석 (그림 5)

5종의 질병(HBV, HCV, CMV, VRE, CPE)에 대하여 3종의 생체시료(혈액, 타액, 피부스왑)를 전처리, 유전자 증폭과 판별을 통하여 질병을 진단하는데 필요한 모든 단위 기능을 모듈화하기 위하여 모듈칩의 구성 요소를 분석하여 모듈화 진행하였음.

그림 5 < 5종의 질병 분자진단 위한 전처리, 증폭, 판별의 전체 구성 요소 분석도



2.1.1.2 분자진단 적용을 위한 단위기능 모듈칩

모듈형 디바이스의 시작품 제작을 위하여 개별 기능 구현이 가능하며 미세채널을 가지는 디바이스와 케이스를 설계하고 모듈칩과 케이스, 그리고 모듈칩간 연결 시 누수를 방지하는 누수 방지 기술을 개발함. 다양한 누수 방지 기술 및 이를 고정화 하는 기술이 존재하나 존재하나 본 연구에서는 연구자와 향후 유저의 손쉬운 디바이스 개발 및 시작품 제작과 성능평가를 위하여 인터페이싱 시 Leakage를 최소화하기 위하여 러버실러를 설계하고 사출하여 제작하는 기술을 개발함. 모듈칩 내 미세유체 채널을 형성하고 모듈간 연결 및 용액의 누수 방지를 위한 세부 요구조건을 도출한 후 누수 방지를 위한 개념을 도출하

고 설계를 진행하였음. 모듈칩의 구동을 위하여 러버실러를 설계하고 모듈칩과 케이스간 누수 방지, 모듈칩과 모듈칩간의 누수 방지를 위하여 3차원 설계를 진행하였음 (그림 6).

그림 6> 단위기능 모듈칩 라이브러리 구축 및 어셈블리 기술

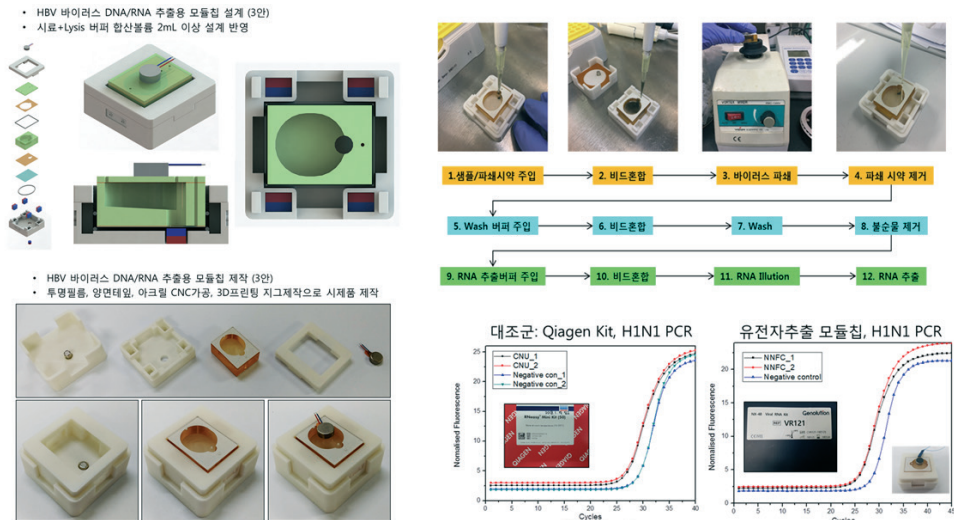


2.1.1.3 분자진단을 위한 생체 시료전처리 모듈 설계 및 제작

시료 파쇄 및 분리를 위한 핵심 구동부 설계 및 제작

혈중 바이러스나 박테리아 시료 파쇄 및 RNA/DNA를 분리하기 위한 유전자추출모듈칩을 설계하고 제작하였음. 투명필름, 양면테이프, 플라스틱으로 구성되어 있으며, 자성입자를 이용하는 방식을 적용함.

그림 7> 유전자 추출 모듈칩 이용한 RNA 추출 및 RT-PCR 결과

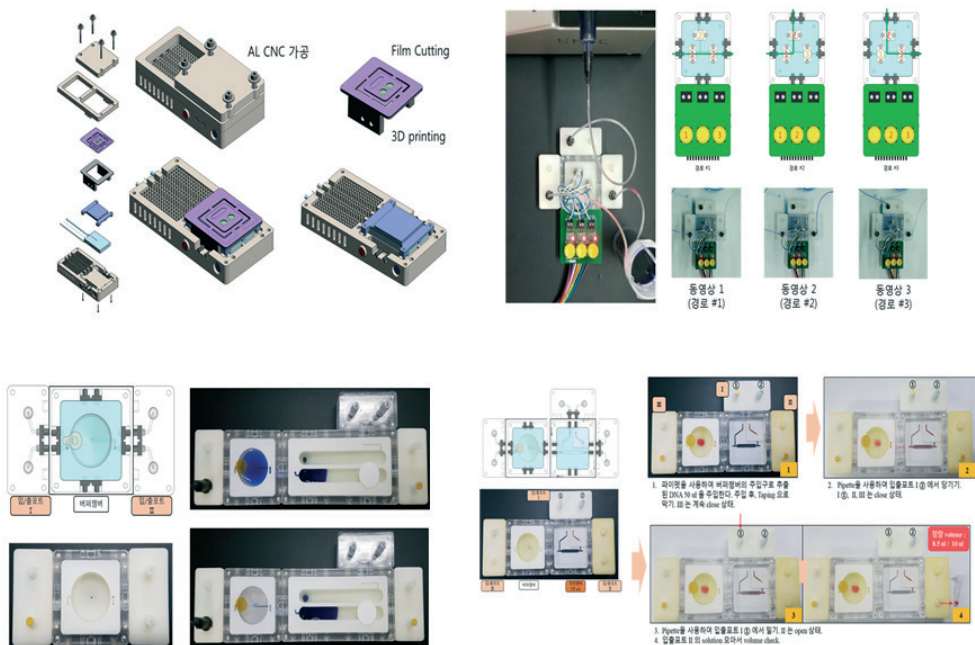


유전자추출 모듈칩을 이용하여 H1N1 바이러스 RNA 추출 실험을 수행하였다. 전체 프로세스는 magnetic Bead를 이용하는 상용화된 전자동 유전자추출기의 프로세스를 참조하였다. 계란의 요수에 포함된 H1N1 바이러스파티클 200 μ L를 취하여 유전자추출 모듈칩에 주입하고, RNA를 추출한 후, Real-time PCR을 수행하여 유전자추출 모듈칩의 RNA 추출 성능을 검증함. 대조군으로 Qiagen RNeasy Mini Kit을 이용하여 추출한 RNA를 이용하여 동일한 조건으로 Real-time PCR을 수행하여 유전자 증폭 정도를 상호 비교함. 대조군과 동일하게 두가지 실험에서 모두 Negative Control보다 높은 값이 확인되어, 유전자추출 모듈칩을 이용한 H1N1 바이러스의 RNA 추출 성능을 검증하였음 (그림 7).

2.1.1.4 모듈칩 적용을 통한 유전자 증폭 성능 검증

4x8형 유전자 증폭 모듈을 설계하여 제작하였음. 3-way 밸브 모듈의 체결구조가 개선되어 밀폐성을 구현하였고 버퍼챔버의 디자인을 변경하여 액체 이송시 잔류량을 최소화 하였음. 정량챔버 디자인을 개선하여 10 μ L 레벨의 정량이 가능해졌음. (그림 8)

그림 8> PCR모듈(좌상), 밸브모듈(우상), 버퍼챔버(좌하), 정량칩(우하)

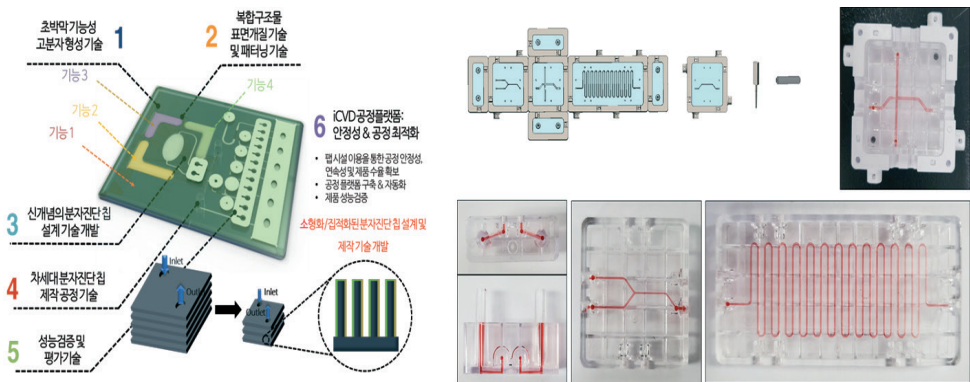


2.1.1.5 단위기능 모듈칩의 준양산형 제조 최적화 기술 개발

플라스틱 사출성형 및 필름라미네이팅 공정을 이용하여 모듈칩 양산 제조기술을 확보하였음. 10만개 이상 제작 가능한 양산수준의 기술을 확보하였음

다양한 소재와 기능으로 구성된 서로 다른 4개의 모듈을 연결(어셈블리)하여 분자진단이 가능한 모듈칩 어셈블리를 구현하였음. 제작된 모듈칩 어셈블리는 4개의 파트로 나뉘어져있음. 첫 번째 PDMS칩(Injection module)은 시료 주입 및 이송 버튼으로 구성되며, 주입된 시료는 버튼을 누름에 따라 다음 파트인 시료 전처리 파트로 이동함. 두 번째 칩(Pre-treatment module)은 필름으로 제작된 칩으로, 생체시료의 전처리가 가능하도록 구성됨. 세 번째 칩(Mixing module)은 PDMS로 된 채널 및 2개의 버튼으로 구성되어 있으며, 전기화학 활성 분석을 위한 전해질과 샘플 시료를 섞어주는 역할을 하며 다음 칩인 electrode module로 시료를 이송하는 역할을 함. 네 번째 (Electrode module)은 전해질과 혼합된 시료로부터 전기화학 활성을 분석하는 모듈임. 이렇게 제작된 통합형 칩 및 PDMS 버튼을 이용하여 시료의 주입부터 분석파트까지 전력 없이 이동이 가능함 (그림 9).

그림 9> 단위기능 모듈 모듈어셈블리 통한 칩 최적화 테스트



2.1.1.6 양산공정으로 제조된 시제품을 이용한 실증평가

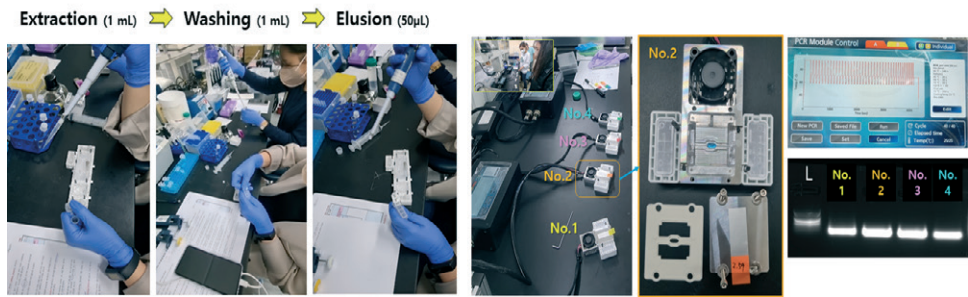
플라스틱 사출성형 공법과 필름 라미네이팅 공법을 적용하여 모듈칩 시제품에 대한 양산 공정을 확보하였으며, 양산공정으로 제작된 모듈칩 시제품을 이용하여 다양한 응용분야에 적용실험을 진행하여 양산공정으로 제작된 시제품의 필드테스트를 수행함으로써, 양산공정의 재현성과 생산 안정성을 확인하였음.

양산공정으로 제조된 유전자추출 모듈칩을 이용하여 연세세브란스병원에서 환자 분변

에서 획득한 수퍼박테리아(VRE) 샘플을 이용하여 유전자 추출 성능평가를 수행한 결과, 기존 병원에서 보유한 상용화 장비와 동일한 성능의 유전자 추출 성능을 확인.

양산공정으로 제작한 유전자증폭 반응칩과 유전자증폭시스템을 이용하여 국내 분자 진단 회사가 보유한 질병유전자를 10^1 copy ~ 10^6 copy 범위에서 증폭한 결과, 모든 범위에서 우수한 유전자 증폭 결과를 확인하였으며, 기존 상용화된 유전자증폭 장비와 동등한 성능을 확인 (그림 10).

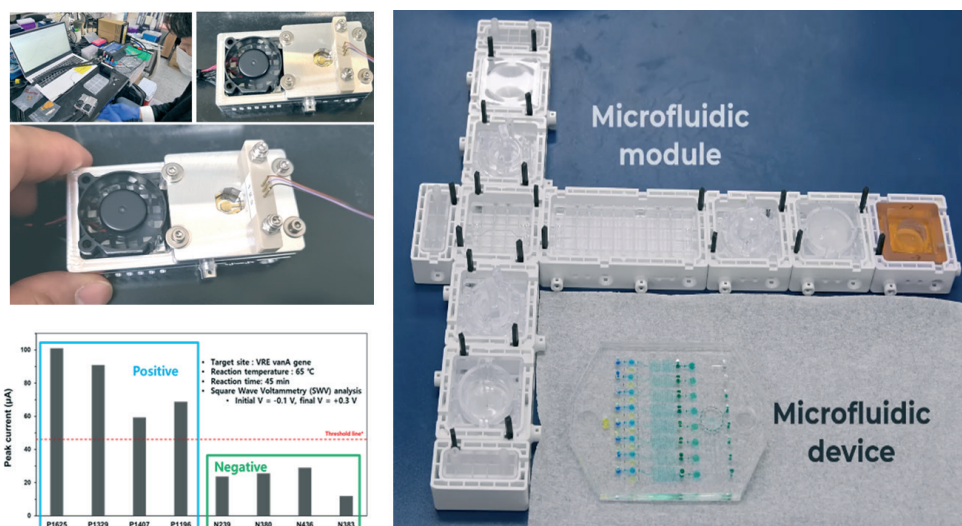
그림 10> 모듈칩 시제품 이용 실증 평가



양산공정으로 제작한 전기화학 등온증폭 반응칩과 등온증폭시스템을 이용하여 수퍼박테리아(VRE) 유전자의 유전자증폭 성능을 평가하여 모든 샘플에서 양성/음성 구별이 완벽히 가능함을 확인하였으며, 병원에서 보유한 유전자증폭 시스템과 동일한 양성/음성 구분 결과를 확인하였음.

양산공정으로 제조된 모듈칩 시제품 라이브러리를 활용하여 무동력펌프 기반 아미노산 분석용 미세유체 소자를 구성하여, 조립된 모듈칩 구성품에서 아미노산 분석이 가능함을 확인하였고, 이를 기반으로 8개의 채널이 집적되어 있는 아미노산 분석용 멀티채널 일체형 미세유체칩을 설계 및 제작하여 아미노산 분석 성능을 확인하였음. 이로써, 모듈칩으로 기능검증을 수행한 후, 집적화된 일체형 미세유체칩을 제작할 수 있음을 실증적으로 확인하였음 (그림 11).

그림 11> 모듈칩 실증 평가 (VRE 진단) 및 일체형 모듈칩 위한 어셈블리



2.1.1 모듈화기술 적용 통한 세포 배양 칩 및 시스템 개발 사례

전통적인 2D 세포 배양 방식은 세포를 플라스틱이나 유리 기판 위에서 단일 층으로 배양하는 기술로, 경제적이고 사용이 간단하며 이미 표준화되어 널리 활용되고 있음

최근에는 미세유체 기반 기술이 세포 배양 기술에 적용되면서 소형화된 장치에서 세포 배양, 분석, 약물 테스트를 통합적으로 수행할 수 있는 솔루션이 등장하고 있으며, 이러한 미세유체 기반 기술은 정밀한 실험 제어가 가능하고 자원을 효율적으로 사용할 수 있는 장점을 가지고 있지만, 초기 설계 변경이나 제작 비용이 높다는 문제점이 존재

세포 배양 연구는 세포 유형에 따라 적합한 온도, 산소 농도, pH 등을 조절해야 하며, 약물 반응 테스트나 대사 분석을 추가하려면 별도의 기능을 통합할 필요성이 있으며, 다양한 조건과 기능을 요구. 기존의 일체형 시스템은 이러한 다양한 요구에 유연하게 대응하기 어려운 반면, 모듈형 시스템은 필요에 따라 기능을 추가하거나 제거할 수 있어 사용자 맞춤형 설계가 가능. 기존 세포 배양 기술은 설계 변경이 어렵고 시간이 오래 걸리는 문제가 존재. 예를 들어, 새로운 세포 유형이나 조건을 테스트하려면 전체 시스템을 다시 설계해야 가능. 그러나 모듈화된 시스템은 단순히 필요한 모듈만 교체하거나 추가함으로써 신속한 프로토타이핑과 설계 변경이 가능. 이로 인해 신약 개발 및 독성 테스트와 같은 분야에서 빠른 실험 반복이 가능.

모듈형 시스템은 높은 유연성과 확장성을 제공. 세포 배양 외에도 약물 처리, 대사 분석

등 다양한 기능을 개별 모듈로 구성할 수 있으며, 이를 통해 연구 목적에 따라 시스템을 쉽게 조합하거나 변경이 가능. 약물 반응성을 테스트하려면 약물 주입 모듈을 추가하고, 대사 분석이 필요하다면 분석 센서를 결합할 수 있는 유연성은 연구자의 작업 효율성을 높이고, 다양한 실험 조건을 만족할 수 있는 환경 조성 가능

모듈화는 표준화와 데이터 통합 측면에서도 큰 이점을 제공. 모듈 간 표준화된 인터페이스를 통해 데이터 수집과 분석이 용이하며, IoT 기술과 결합하면 실시간으로 데이터를 모니터링하고 클라우드 기반으로 분석. 이를 통해 연구 결과의 신뢰성과 재현성을 높이는 데 기여 가능

모듈칩 적용을 통한 다세포 배양 기능 확립하였고 세부적으로는 배지 연속 순환 펌프 시스템에서 배양 안정성, 온도 및 가스 제어 시스템의 안정성, 세포배양 모듈 2개를 연결하고, 동시에 2종 이상의 다세포 배양 안정성 확인하였다. (배양 조건 : 세포 (C2C12, T98G), 배양 조건 (37°C, 5% CO₂, 2일 배양), 배지 (DMEM : MEM=1:1), 순환 속도 (10ul/s)) (그림 12).

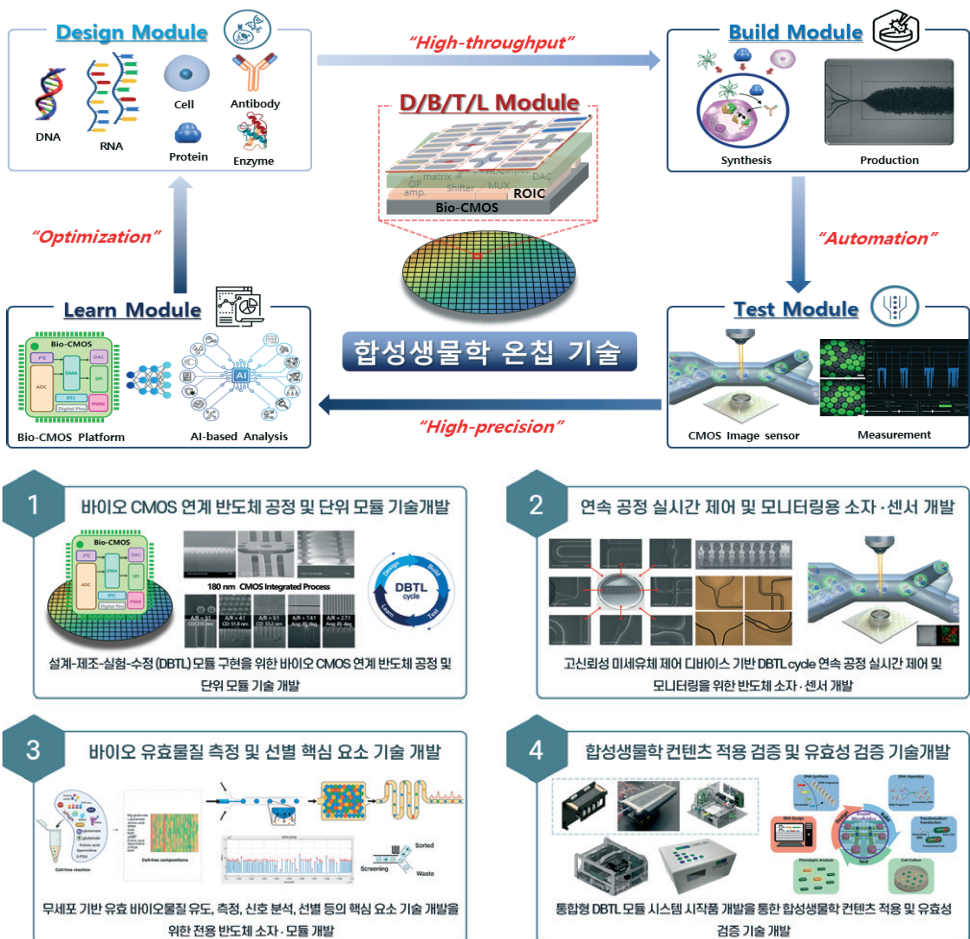
그림 12> 모듈형 세포 배양칩 및 배양환경 조절 시스템 개발



3 맺음말

바이오파운드리라는 이런 합성생물학의 전 과정을 표준화, 자동화, 고속화한 시스템이다. 인공지능(AI)과 로봇 기술을 기반으로 설계(Design)-제작(Build)-검증(Test)-학습(Learn)의 과정을 자동화해 생물학 실험과 제조 공정을 수행하는 ‘바이오 자동화플랫폼’으로 바이오 산업의 새로운 패러다임으로 자리 잡고 있다. 이에 우리는 기존의 아날로그 방식의 연구개발 방식의 한계점을 극복하기 위해 반도체/멤즈 기술 기반의 새로운 합성생물학온칩 구현을 위해 새로운 플랫폼 (모듈화 기술, 바이오-CMOS 기술 등)을 통해 바이오산업 혁신을 가속화할때라고 생각한다. 또한 다양한 분야에서 이런 반도체 플랫폼을 활용하여 합성생물학 콘텐츠 적용 검증 및 유효성 검증이 필요할때라고 생각한다 (그림 12).

그림 13> 합성생물학온칩 기술 및 응용분야



[참고문헌]

- [1] (Nate 뉴스, 2021.10.26) 합성생물학 기반 바이오파운드리, 한국 모더나 키운다
<https://news.nate.com/view/20211026n01379>
- [2] (보고서) 플랫폼 연계 제품상용화 최종보고서 (반도체공정기반 나노메디컬 디바이스 개발, 나노종합기술원, 2025.11.5.)
- [3] (보고서) 분자진단용 전기 나노바이오센서 모듈화기술 개발 최종보고서 (나노소재기술개발사업, 원천기술개발사업, '17~'22))
- [4] (발표자료) 바이오파운드리 기획/사업 추진현황 (생명연 합성생물학연구소 이승구 소장, 2024, 2.)
- [5] (보고서) 제4차 생명공학육성 기본계획: 바이오 대전환을 선도하는 혁신전략 2030 (관계부처 합동)
- [6] (참고문헌) Transforming microfluidics for single-cell analysis with robotics and artificial intelligence. LabChip(2025) Nov 5. doi: 10.1039/d5lc00216h.
- [7] (참고문헌) Synthetic Cells from Droplet-Based Microfluidics for Biosensing and Biomedical Applications. Small (2024) Volume20, Issue33, 2400086
- [8] (참고문헌) A comprehensive review of Microfluidic approaches in cell-free synthetic biology. Front. Synth. Biol.,(2024) <https://doi.org/10.3389/fsybi.2024.1397533>

저자 소개

이도현 KAIST 교수, dhlee@kaist.ac.kr

김현욱 KAIST 교수, ehukim@kaist.ac.kr

김하성 한국생명공학연구원 책임연구원, haseong@kribb.re.kr

나도균 중앙대학교 교수, blisszen@cau.ac.kr

정철희 고려대학교 교수, damo363@korea.ac.kr

이준학 한국과학기술정보연구원 디지털바이오컴퓨팅연구단 단장, june@kisti.re.kr

염민선 나무아이씨티 연구소장, ms.yeom@namuict.co.kr

이선재 광주과학기술원 생명과학과 교수, leesunjae@kaist.ac.kr

송자연 성균관대학교 송자연 조교수, sparkling@skku.edu

임현의 한국기계연구원 바이오파운드리 DX/AI 연구단 단장, helim@kimm.re.kr

이대식 한국전자통신연구원 책임연구원, dslee@etri.re.kr

김진한 케이랩 주식회사 연구소장, jhkim37@klab.im

이석재 나노융합기술원 나노융합기술개발본부 본부장, sjlee@nnfc.re.kr

2026 데이터, AI, 자동화를 중심으로 본 합성생물학 융합기술동향 보고서

발 행 인 한국합성생물학발전협의회 운영위원회

발 행 처 한국합성생물학발전협의회

대전광역시 유성구 과학로 125, KOBIC동 1323호

T. 042)860-4371~81 E. kmyang@kribb.re.kr / wogml919@kribb.re.kr

발 행 일 2026년 2월 23일

책임편집 양경미 사무국장, 김재희 연구원

한국합성생물학발전협의회 2026



대전광역시 유성구 과학로 125, KOBIC동 1323호
042) 860-4371~81
kmyang@kribb.re.kr / wogm1919@kribb.re.kr