〔生物工学会誌　第100巻　第11号　607–610．2022〕
DOI: 10.34565/seibutsukogaku.100.11_607

文献・公開データからの学習による「次の一手」の代謝設計提案

中澤　志織[1]・伊藤　潔人[1]*

はじめに

持続可能なものづくりへの転換は時代の要請であり，生物の機能を利用した化合物生産はますます注目と期待を集めている．遺伝子改変生物による化合物生産で多様な物質生産需要に対応するためには，目的化合物が内在性であるか新規代謝経路を要するかにかかわらず，目的化合物を高効率に合成させる代謝設計が不可欠である．

代謝設計の従来技術として，特定の細胞に存在する代謝ネットワークをモデル化し，代謝をシミュレートするアプローチがある．化学量論式で代謝ネットワークを表現したゲノムスケールモデルを用い，ネットワークの代謝フラックスを予測するflux balance analysis[1)]が代表例である．代謝をシミュレートすることで，たとえば，特定の目的化合物の生産量を最大化したい時に，モデル中の代謝反応のいずれを促進し，いずれを抑制すべきかを知ることができる．これらの手法はモデルの中で化学量論的な整合性を確保しながら最適化を行ううえで大変有用である．しかしその長所の裏返しで，モデルの中からしか答えが出ないという短所を持つ．つまり，代謝設計を革新させるには，モデルに含まれていない新たな着想，すなわち「次の一手」を提案する技術が必要である．

人間の研究者が次の一手を探索する時，論文を読む，公開データベース（DB）を閲覧するなど，文献･公開データから新たな着想を探ることは多い．PubMedに収録された文献や，Elixir，KEGGなどの公開DBに集約された遺伝子・タンパク質・代謝反応・代謝経路など，化合物生産向け代謝改変設計に有用な知見は，文献あるいは文献情報が人手により要約された各種公開DBで一定量利用可能であり，先人が積み上げた知識を得ることができる．しかし文献・公開データからの有用情報抽出は，研究者のノウハウに基づいた文献検索・DB検索に頼っているのが実情である．そこで本稿では，機械学習を含む近年の情報処理技術により新しい代謝設計を導く「次の一手」を提案する技術について紹介したい．

代謝設計と機械学習

近年，機械学習技術は著しい発達を遂げ，学習データ量の増加と深層学習の進歩により，入力から出力までのすべての処理をただ1つのモデルで学習するend-to-end学習が可能になった．End-to-end学習は，決まった形式の入力と出力のデータを大量に用意できるタスクにおいて優れた結果を出し，大きな話題を呼んだ．一方，代謝設計というタスクとend-to-end学習とは，実は相性が良くないことが知られている．まず，代謝設計には，新規代謝経路の探索，経路上および経路に影響する各因子の強化・抑制，フィードバック調節の解除など，さまざまな検討事項があること，つまり，目的関数が多様であることがあげられる．さらに，入出力も化合物，代謝反応，酵素番号から，遺伝子の選択，配列の最適化など，多様なパターンがありうる．結果として，さまざまな実験を繰り返すわりに訓練データが大量には集まりにくい．ここが目的変数，入出力が統一されたデータを比較的集めやすい酵素エンジニアリング，タンパク質設計との大きな違いである．

ここで，多様なデータを代謝設計のために利活用する機械学習的アプローチには，大別して下記2種類があると筆者らは考えている．

学習モデルの段階利用　目的や入出力の多様さゆえにend-to-end学習が難しいのであれば，問題を段階に区切って，まとまったデータを集めやすい問題の組合せとし，各問題を解くモデルの組合せで解けばよい，というのは自然なアプローチである．すなわち，公開データを用いて学習させたモデルの予測結果を，分野の知識（ドメイン知識）に基づくルールによる処理や他の学習モデルとうまく組み合わせることで，目的に沿った入出力を実現する．たとえば，非天然化合物合成のための代謝経路設計では，代謝経路を構成する中間体と酵素の設計に加えて，中間体の毒性，新規導入反応のもっともらしさ，外来遺伝子が宿主中で機能する見込みなどを考慮することで，より動作の見込みが高い代謝設計が可能になる[2)]．高精度なタンパク質立体構造予測で話題になったAlphaFold[3)]も，multiple sequence alignment，コンタク

*著者紹介　株式会社日立製作所 研究開発グループ（主任研究員）　E-mail: kiyoto.ito.kp@hitachi.com
[1]株式会社日立製作所 研究開発グループ

トマップのような従来の生物情報学および構造生物学のドメイン知識に基づく処理，さらには自然言語処理分野で発達したtransformerを組み合わせたモデルで構成されている．このアプローチで良い結果を出すためには，データ確保や公開データを用いた学習のしやすさを鑑みて問題を適切に分割し組み合わせる勘所と，考慮すべきポイントや既存の生物学的な知見に関するドメイン知識を活かすことが鍵である．したがって，情報処理・機械学習と解くべき問題の各構成要素のドメイン知識とに通じたメンバーが，分野横断的に協力するチーム編成が重要である．

ここで，公開データで学習しやすい問題とは，たとえば公開データから多数の「正解」例が得られる問題である．代謝経路設計という問題を解くために，実在する化合物構造を正解例とできる「化合物らしさ」や，報告されている代謝反応を正解例とできる「代謝反応らしさ」を公開データから学習したモデルを段階的に組み合わせた例を，事例1で紹介する．

自然言語処理の利用　もう一つのアプローチは，自然言語で公開されている情報，すなわち文献からの有用情報獲得である．代謝設計に有用な知識の多くは論文などに自然言語の形で公表され，その一部だけがDBに要約されて登録される．しかし，文献情報の要約は人手により多くの労力をかけてなされており，増え続ける論文の有用情報を漏れなくDBに登録することは非現実的である．そこで，人手に依存しない自然言語処理による情報抽出技術が重要となる．

自然言語処理の利点の一つは，多様な目的や，入出力パターンに適用できることである．さらに，検索エンジンの使い勝手から日常生活でも実感されるように，近年の機械学習の発展で著しく進歩した分野の一つでもあり，単語やフレーズの抽出精度は格段に向上している．一方，単なるキーワードの一致検索では，得られる答えが与えたキーワードに直結する範囲に限られ，代謝設計として新しい知見「次の一手」の抽出が難しい．そこで有効なアプローチの一つが，連想検索である．連想検索とは，検索のタネとして与えた言葉（または言葉の集合）を手掛かりに，タネに近い文書を探し集める方法であり，欲しい情報を芋づる式に釣り上げることができる．一方，タネを手掛かりにする以上，連想検索させるタネが重要である．これまでにユーザーが試した代謝設計をタネに次の一手を抽出した例を事例2で紹介する．

事例1：学習モデルの段階利用による代謝経路探索

学習モデルの段階利用を，化合物合成を目的とした新規代謝経路設計に適用した筆者らの試み[4]について紹介する．代謝経路は，経路を構成する化合物と，化合物間をつなぐ反応との組合せと捉えることができる．化合物構造は，ElixirのChEMBLやNCBIのPubChemから10^6～10^7の規模で，代謝反応の公開データもKEGG，Rheaなどから10^4～10^5の規模で取得可能である．そこで，図1のように化合物構造のデータを利用して化合物らしさの学習を行い，化合物構造を数値（化合物ベクトル）で表現，次いで，反応前後の化合物構造の変化を化合物ベクトルの差分（反応ベクトル）で表現する手段を採った．このベクトルという表現手法により，公開データ中の代謝反応の反応ベクトルを，その反応とは直接関係しない化合物ベクトルに加算（または減算）することで，文献報告やDB収載がなされていない代謝反応であっても類推できる．

この手法を用いて，出発化合物および目的化合物の化合物ベクトル差分（経路ベクトル）にもっとも近い和を持つ反応ベクトルの組合せを探索することで，出発化合物と目的化合物を結ぶ新規代謝経路の構成要素（反応セット）を得ることができる．ただし，KEGGに収載された反応の全組合せを探索することは膨大な計算量を要求するため，進化計算を利用し効率化を図った．具体的には，KEGG収載反応を反応ベクトルに変換して格納したDBから，反応セットをランダムに複数選択し，その妥当性評価を後述の方法で行う．妥当性が低い反応セットは切り捨て，妥当性が高い反応セットだけを残しつつ妥当性が高い反応セット間で反応を入れ替える処理

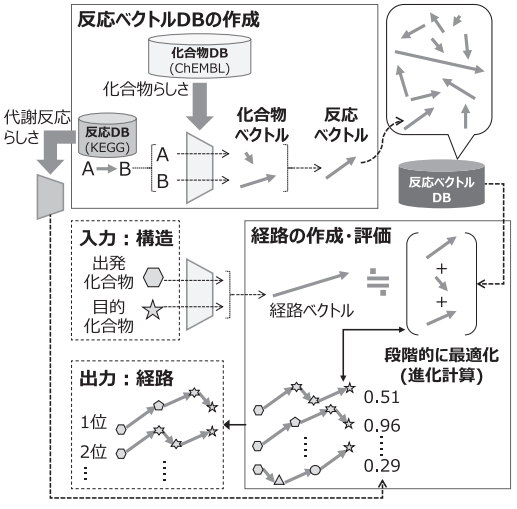

図1．モデルを段階的に組み合わせた代謝経路探索の流れ

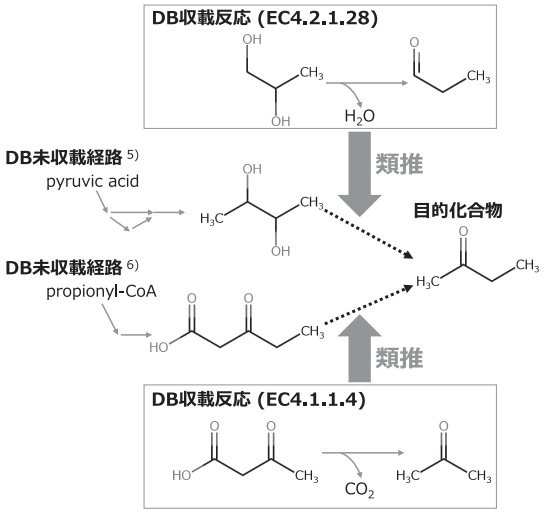

図2. 2-Butanone合成課題に対してDB収載反応からの類推で提案されたDB未収載反応.

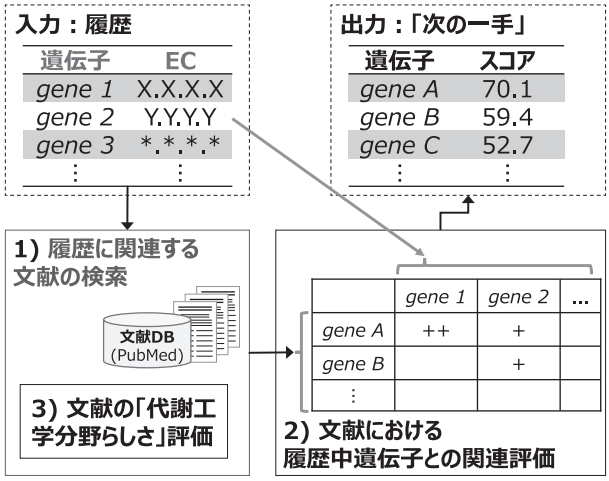

図3. 連想検索を利用した遺伝子改変履歴に基づく次の一手提案の流れ.

を繰り返すことで,段階的に反応セットの最適化を行う.代謝経路の妥当性評価のためには，公開データ中の反応と，そこから生成した架空の反応で代謝反応らしさを学習させたモデルを用意し，代謝経路を構成する各反応につき反応の妥当性を評価し，構成反応の妥当性評価値の積を代謝経路の妥当性とした.

以上のようにモデルを多段階で組み合わせたことで,実際に2-butanone合成に対して，KEGG収載の代謝反応のみを事前に与えたシステムから，KEGG未収載だが現実の動作報告[5,6]がある代謝経路を類推で提案することができた（図2）.

事例2：自然言語処理を用いた次の改変遺伝子提案

次に，自然言語処理を用いて公開文献より次の一手となる遺伝子を提案した事例[7]について紹介する．本システム（図3）は，特定の化合物を特定の宿主生物で生産するという一つの目的に対して，これまでに改変を試みた遺伝子のリスト（以下,「履歴」と称する）を検索のタネとする連想検索で，文献から，履歴と関連する遺伝子を次の一手として提案し，次の着想を支援する.

具体的には，PubMed全件からオープンに利用可能な題名および要旨を探索範囲として文献検索し，得られた文献集合内で同一文献中において履歴中遺伝子とともに言及された実績（共起）が多い遺伝子を，履歴と関連する遺伝子として提案する．ここで履歴および代謝改変という目的に関連しつつ，着想を一歩拡げる次の一手を取るために，下記3段階で連想検索を行っている：1）文献検索の際に，タネとする履歴中遺伝子と機能が似ている（同じ酵素番号を持つ）遺伝子名でも検索，2）検索で得られた文献における，履歴中遺伝子との共起に基づく履歴外遺伝子の評価，3）検索で得られた文献の「代謝工学分野らしさ」を代謝工学分野の文献に対する語彙の類似性で評価，である（図3）.

上記1）のクエリ拡張および2）の共起に基づく検索により，履歴に関連する範囲で着想を一歩拡げた次の一手を提示することができるとともに，人間が見落としなく調べることが不可能な量の文献からでも関連知識を抽出することができる．さらに，PubMed収載文献には基礎生物学や医学などの，学術的に重要だが化合物生産向け代謝設計においてはノイズとなる分野の文献も多く，代謝工学分野らしい文献から抽出された遺伝子に重みを付けることで，より代謝設計への関連の深い提案が期待できる．そこで，代謝設計に有用な知識を重点的に抽出する狙いで，文献の「代謝工学分野らしさ」を評価する仕組みである上記3）を組み合わせた．3）は，事前に作成した，PubMed収載文献全体に比べて代謝工学分野の文献における出現頻度が特に高い単語（分野特徴語）のリストを内部に持ち，分野特徴語の出現頻度によって各文献の「代謝工学分野らしさ」を評価する.

上記システムをシキミ酸生産コリネ菌[8]の代謝設計の履歴に適用した結果を次に示す．18遺伝子の改変からなる履歴をタネに連想検索を行い，26,522本の文献において言及された4,052遺伝子のランキングを次の一手候補として得た．このランキングをシキミ酸生産量世界記録コリネ菌株（以下，元株）の樹立者らが確認し,「次の一手として実験を行う価値がある」として注目した7遺伝子の順位を表1に示す．（“Shikimic acid” OR shikimate）をクエリとする一致検索で得られた文献中の遺伝子名を出

表1．次の一手提案における注目遺伝子の順位

	一致検索	連想検索
文献数	1,772	26,522
遺伝子数	565	4,052
順位		
ppsA	31	17
galP	68	12
pykF	61	10
pfkA	不検出	27
pgi	198	19
ppc	106	22
shiA	31	103

現回数でランキングした場合に比べ，履歴中の18遺伝子をタネとする連想検索では，得られた文献数と遺伝子数が増加し，かつ注目遺伝子のほとんどで順位が上昇した．

すなわち，連想検索によって，知識抽出の範囲が課題のキーワードから容易に検索可能な範囲より拡大され，かつ熟練者の感覚で有望そうな次の一手がより上位に提示された．さらに，元株樹立者らが注目遺伝子から着想を得て10種類の遺伝子改変を設計し，それぞれ元株に追加した結果，うち4種類で元株より最大19.5 %高いシキミ酸生産濃度が見られた[7]．以上より，代謝設計の履歴をタネとする連想検索による，次の一手となる改変遺伝子提案の有効性が示された．

おわりに

以上の事例のように，学習モデルの段階利用，自然言語処理の各アプローチにより，直接的に導くことが難しい，代謝設計を革新させる「次の一手」を導くことができることを示した．これらのアプローチによる新たな着想が，シミュレーションやend-to-end学習と相互補完的に代謝設計を加速するものと考えている．

機械学習技術の発達と公開データの充実は，公開データに基づく機械学習を活用した代謝設計にも追い風である．一方，公開データでは得ることができない実験データを学習して代謝設計にフィードバックする技術の開発も盛んであり，実験データと公開データの利活用が相互補完的に化合物合成微生物の創製における代謝設計を支援すると期待している．

一方，現在の文献・公開DBから得られる情報の短所として，ネガティブな結果が報告されにくい，扱いやすい酵素や，よく知られている遺伝子に報告数や被引用数が偏る，などの点がある．また，配列情報のみが既知である遺伝子の活用への期待も大きい．これらの解決策として，あえて情報の少ない設計を重点的に探索する実験を設計する，配列から機能や性質を予測して代謝設計に活用する，などのアプローチがあり得る．また，未発見・未報告の情報の収集，モデルの構築改良のためのデータ収集，設計の有効性検証などの実験も依然不可欠である．実験データおよび公開データの利活用技術に加え，実験を効率化するラボラトリーオートメーション，データを利活用しやすいよう体系的に整理・共有する仕組みなどのさらなる発達と，それらが相互に発展を促すことが望まれる．

現在，生物工学分野において機械学習は一部タスクに対して圧倒的な性能を示し，代謝設計への情報処理技術の活用は着実に進んでいるとの感がある．一方，段階的にモデルを組み合わせるにも，情報処理技術の出力を具体案に変え実装に移すにも，未だモデル化されていない専門家のドメイン知識や感覚に頼る部分が少なくないのが現状である．たとえば，タンパク質立体構造予測技術の目覚ましい進歩を生物工学にどう活かすかは，少なくとも現在は，人間の知恵次第である．ラボラトリーオートメーションにおいても，たとえば，現行の実験操作をロボットが実施可能なプロトコルに変換する過程は，ウェット実験とロボティクスの知識を要する．したがって，今後の生物の機能を利用した化合物生産の発展には，生物工学・情報処理・ロボティクスなどの各分野を深く学んだ人材間の協力関係が不可欠と考える．同時に，複数分野に通じ分野間を橋渡しできる人材の存在も今後ますます重要である．分野融合的な議論と協働のさらなる活性化によって，生物の機能を利用した持続可能なものづくりがさらに推進されると期待している．

謝　辞

本稿で紹介した研究の一部は新エネルギー・産業技術総合開発機構（NEDO）「植物等の生物を用いた高機能品生産技術の開発」事業により得られたものであり，国立健康・栄養研究所荒木通啓先生，理化学研究所白井智量先生，地球環境産業技術研究機構（RITE）乾将行博士，須田雅子博士，豊田晃一博士，久保田健博士，小暮高久博士，日立製作所 藤大樹博士，今一修博士らと共に行った．この場を借りて感謝する．

文　献

1) 松田史生ら：生物工学会誌，**92**, 593 (2014).
2) Carbonell, P. *et al.*: *BMC Syst. Biol.*, **5**, 122 (2011).
3) Jumper, J. *et al.*: *Nature*, **596**, 583 (2021).
4) Fuji, T. *et al.*: *Bioinformatics*, **36**, i770 (2020).
5) Chen, Z. *et al.*: *PLoS One*, **10**, e0140508 (2015).
6) Srirangan, K. *et al.*: *Appl. Environ. Microbiol.*, **82**, 2574 (2016).
7) Nakazawa, S. *et al.*: *ACS Synth. Biol.*, **10**, 2308 (2021).
8) Kogure, T. *et al.*: *Metab. Eng.*, **38**, 204 (2016).