特集

バイオインフォマティクス研究の今　1

ビッグデータと AI が拓く新時代のバイオインフォマティクス

医療と創薬の AI 新時代

竹本和広　Kazuhiro TAKEMOTO

本誌においてバイオインフォマティクスが特集されるのは 2007 年以来のことである。この間，この研究分野においても大きな進展が多数あった。特に，ビッグデータと人工知能（AI）の融合は，この研究分野に驚異的な進歩をもたらしている。本稿では，近年の研究動向を踏まえ，ビッグデータと AI 技術がどのようにバイオインフォマティクスの新時代を開拓してきたか，そして今後の展開について概説する。

バイオインフォマティクス

バイオインフォマティクス[1]とは生物学と情報学の学際領域分野であり，生命情報学などとも呼ばれる。種々のゲノム計画が特に盛んだった 21 世紀初頭頃まではゲノムデータのような配列データを中心に取り扱っていたが，ポストゲノム時代に入ると，ゲノムから生命現象を理解するという目的の下に，その対象とする分野は急激に拡大していった。タンパク質の立体構造解析はもちろんのこと，トランスクリプトーム（網羅的遺伝子発現）解析，化学と情報学の学際領域分野であるケモインフォマティクス，医用画像解析のようなコンピュータビジョン関連の分野についてもその範疇となってきた。もちろん，これは計測技術の大きな進展により網羅的計測が可能になったことが大きく影響している。計測技術からもたらされるビッグデータを分析し，有用な科学的知見を抽出したり，予測などの側面で産業応用に役立てたりすることが，バイオインフォマティクスの中心的な課題になっていった。そして，このビックデータ分析に関して重大な情報技術革新があった。いわゆる，AI 技術革新である。

たけもと・かずひろ
九州工業大学大学院情報工学研究院生命化学情報工学研究系　教授
〔経歴〕2008 年京都大学大学院情報学研究科博士課程修了，博士（情報学）。科学技術振興機構さきがけ専任研究者などを経て，12 年九州工業大学大学院情報工学研究院助教。15 年同准教授。22 年より現職。〔専門〕バイオインフォマティクス，ネットワーク科学。
E-mail: takemoto@bio.kyutech.ac.jp

深層学習の台頭

近年の AI 技術革新を語る上で深層学習[2]は切っても切り離せない。深層学習は多層からなるニューラルネットワークを用いた機械学習手法であり，AI 技術の中核を担っている。ニューラルネットワークは，その技術的な限界から，かつては過去の遺物と見なされていた時期もあったが，近年の技術的な進展により，深層学習として大きな発展を遂げ，2010 年頃からその隆盛が顕著になった。特に深層学習は，ビッグデータから複雑なパターンや関係性を抽出し，予測，分類，生成などを高性能で行うことを可能とし，これまでは難しいとされてきた画像認識や自然言語処理で華々しい成果をあげてきた。まさに，深層学習はビッグデータ時代を象徴する技術であり，その性能の高さから様々な分野での応用が急速に進んでいった。それは，バイオインフォマティクスが範疇とする研究分野においても，もちろん例外ではなかった。

医用画像診断

最初に顕著な躍進があったのは，医用画像診断の分野であった。深層学習の画像認識能力の高さを応用したのである。特に 2017 年に Esteva らによって発表された皮膚がん診断を行う AI[3]は 1 つのマイルストーンであろう。著者らは深層学習技術を用いて皮膚科医に匹敵するレベルで皮膚がん診断を行うことができることを示してみせた。その他にも，胸部 X 線画像からの肺炎診断や網膜の光干渉断層画像からの糖尿病性網膜症診断を行う AI[4]が開発され，それらについても専門医

と同等のレベルで診断を行うことが示された。なお，このような人間とAIの比較は医療分野でも広く行われており，特に近年のメタ解析[5]からはAIの診断能力は医療現場で実際に活躍する診断医と同等であることが示されている。

専門医に匹敵するようなAIが開発可能であることは，産業的な意味でも大きなインパクトを与えた。大学発を含むベンチャー企業はもちろんのこと大手企業なども医療AIの開発に乗り出し，その開発は加速化している[6]。例えば，新型コロナウイルスの流行に伴い，様々な企業が胸部X線画像などから新型コロナウイルス性肺炎を検出するためのAIを開発したことは記憶に新しい。また，SkinVision[7]に代表されるようなスマートフォンアプリケーションなどの開発も進められ，医療AIと一般市民の距離が急速に縮まったことも特筆すべきことであろう。

AlphaFold2の出現

本誌の読者にとって，より身近なところで言えばやはりAlphaFold2[8]は外せないだろう。AlphaFold2とはDeepMindが開発したタンパク質立体構造予測プログラムであり，わずかな時間でタンパク質のアミノ酸配列からその立体構造を極めて高い精度で予測することができる。これは，これまで明らかにされてきたタンパク立体構造に関する叡智（データベース）と深層学習によって飛躍的に高められた自然言語処理によってなされた。科学的な正確性をやや欠く表現になるが，アミノ酸配列も1つの文章であり，タンパク質立体構造のビッグデータを深層学習で用いることにより立体構造に関連する重要な文法や規則などが抽出できた結果だと言えるだろう。

AlphaFold2は商用目的を含め誰でも利用可能であり，科学界と産業界の両方に大きな影響を与えている。立体構造が不明なタンパク質についてもその立体構造をうかがい知ることができる。すでにAlphaFold2によって予測されたタンパク質立体構造はデータベース[9]にまとめられており，誰でも利用可能である。このことは特に創薬において重大なインパクトになることは想像に難くない。例えば，Alphabet（DeepMindとGoogleの親会社）は予測構造に対して強く結合するリガンド（薬候補）を精度良く設計することを目的とした創薬企業を設立している。

アミノ酸配列からの立体構造予測問題はバイオインフォマティクスにおける積年の課題であったが，AI技術革新によって解決に向けて大きく前進したと言える。

なお，AlphaFold2関しては森脇によるすばらしい日本語総説[10]を参照してほしい。タンパク質立体構造予測の歴史，立体構造予測の原理，AlphaFold2を含む深層学習ベースの立体構造予測プログラムについて詳細に解説されている。この節は所詮その総説を要約したものにすぎない。

余談ではあるが，上記の総説が掲載されているJSBi Bioinformatics Review[11]では，バイオインフォマティクス分野の研究動向に関する良質な日本語総説論文が発信されている。バイオインフォマティクスに少しでも興味を持つ方は，ぜひ参考にしてほしい。

AI創薬

より化学に直接関係するテーマとしては新規化合物合成があげられるだろうか。特に創薬の観点からは，新薬候補の探索と合成が極めて重要である。新薬候補となりうる化合物は10^{23}から10^{60}個程度[12]であると見積もられているが，これまでに合成された化合物はこのうちの10^{8}個程度[13]である。科学的な新規性や産業的な価値を求めて新たな化合物を探索合成することがいかに難しいことかがわかる。この問題を回避するためには情報学的なアプローチ（例えばComputer-Aided Drug Design）が有効である。そして，このアプローチがAI技術革新によって新たな局面を迎えている。具体的に，深層学習が得意とする「生成」を応用したのである。

OpenAI社が開発したChatGPT[14]を使ったことのある読者も多いだろう。ChatGPTは，深層学習に基づいた大規模言語モデルの一種である。大量の文章データを学習することで，まるで人間が書いたような自然な文章を生成することを可能にした。

化合物はSMILES（Simplified Molecular Input Line Entry System）記法を用いて文字列で表現されることが多い。

そこには文法規則があり，SMILES 表記された化合物はある種の文章であるとみなすことができる。このことを利用して，SMILES 表記された化学構造（つまり文章）を生成するような AI を深層学習で開発する試みが精力的に行われている。先駆的なものとしては Gómez-Bombarelli らの研究[15]がある。ただし，この研究で使われた生成モデルは現在から見ると生成能力に限界がある。特に，文法的に正しくない SMILES を生成することもしばしばである。そのため，文字列生成により特化した生成モデルを使ったり[16]，大規模言語モデルの基礎となる最新鋭の生成モデルを使ったりして[17]，より高性能な化学構造生成モデルの開発が進められている[18]。

深層学習に基づく化学構造生成の利点は様々あるが，産業的な観点からは，条件付けを行いながら多様な化学構造生成が可能である点が注目に値する。例えば，創薬の観点からは，ヒトに対する毒性が極めて小さく，また合成しやすいような化合物が求められることが多いだろう。生成モデルは，その内部において SMILES のような離散的な文字列，そして毒性や生成のしやすさといった化合物の特性が連続的に表現されている。そのため，種々の条件を考慮しながら化合物空間をなめらかに探索することができる。このため，所望の性質を持つ多様な新規化学構造を生成することができる。

近年では，この化学構造生成にトランスクリプトームデータを組み合わせるような研究も行われている。化合物の特性だけでなく「生命現象」も考慮して化合物を生成するという試みである。例えば，ある標的遺伝子がノックアウトされたときのトランスクリプトームデータを条件に用いて化学構造生成を行うことで，その標的の阻害剤候補分子を生成することができる[19]。また，患者のトランスクリプトームデータを用い，その遺伝子発現パターンを反転させる（つまり健康状態にするような）条件付けを行うことによって，薬物候補分子を生成することもできる[20]。

まとめと課題

医療，創薬分野に注目し，AI 技術革新によるバイオインフォマティクスの進展について概説した。社会的ニーズの高まりもあり，このような AI 技術はこれからも加速度的に発展していくだろう。このような時代において，バイオインフォマティクスの重要性はますます高まっていくものと期待できる。

ただ，課題も残される。例えば，セキュリティである。敵対者は AI の出力（予測結果など）を簡単に制御することができる[21]。このことは医用画像診断において，誤診断や保険金詐欺などと関連し，大きな社会問題になる可能性がある[22]。また，プライバシー保護を考慮した AI の開発と運用も重要である[23]。AI 技術と社会の関係については今後の大きな課題である。

1）福永津嵩, 岩切淳一, バイオインフォマティクスのための生命科学入門, コロナ社, **2022**.
2）岡谷貴之, 深層学習, 改訂第 2 版, 講談社, **2022**.
3）A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, *Nature* **2017**, *542*, 115.
4）D. S. Kermany et al., *Cell* **2018**, *172*, 1122.
5）X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, *Lancet Digit. Heal.* **2019**, *1*, e271.
6）井川房夫（編著）, 藤田広志（編著）, これだけでわかる！ 医療 AI, 中外医学社, **2021**.
7）A. Udrea, G. D. Mitra, D. Costea, E. C. Noels, M. Wakkee, D. M. Siegel, T. M. de Carvalho, T. E. C. Nijsten, *J. Eur. Acad. Dermatol. Venereol*. **2020**, *34*, 648.
8）J. Jumper et al., *Nature* **2021**, *596*, 583.
9）M. Varadi et al., *Nucleic Acids Res*. **2022**, *50*, D439.
10）森脇由隆, *JSBi Bioinformatics Review* **2022**, *3*, 47.
11）https://www.jstage.jst.go.jp/browse/jsbibr/list/-char/ja
12）P. G. Polishchuk, T. I. Madzhidov, A. Varnek, *J. Comput. Aided Mol. Des*. **2013**, *27*, 675.
13）S. Kim et al., *Nucleic Acids Res*. **2020**, *49*, D439.
14）OpenAI, *OpenAI Blog* **2022**. https://openai.com/blog/chatgpt
15）R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, A. Aspuru-Guzik, *ACS Cent. Sci*. **2018**, *4*, 268.
16）M. Popova, O. Isayev, A. Tropsha, *Sci. Adv*. **2018**, *4*, eaap7885.
17）V. Bagal, R. Aggarwal, P. K. Vinod, U. D. Priyakumar, *J. Chem. Inf. Model* **2022**, *62*, 2064.
18）T. Sousa, J. Correia, V. Pereira, M. Rocha, *J. Chem. Inf. Model* **2021**, *61*, 5343.
19）O. Méndez-Lucio, B. Baillif, D.-A. Clevert, D. Rouquié, J. Wichard, *Nat. Commun*. **2020**, *11*, 10.
20）C. Yamanaka, S. Uki, K. Kaitoh, M. Iwata, Y. Yamanishi, *Mol. Inform*. **2023**, *42*, 2300064.
21）H. Hirano, K. Takemoto, *Algorithms* **2020**, *12*, 268.
22）H. Hirano, A. Minagi, K. Takemoto, *BMC Med. Imaging* **2021**, *21*, 9.
23）G. A. Kaissis, M. R. Makowski, D. Rückert, R. F. Braren, *Nat. Mach. Intell*. **2020**, *2*, 305.