計算システム生物学による創薬： 分子，構造からネットワークへ

夏目やよい，水口 賢司

要約：創薬研究における時間，労力，費用といった様々なコストを下げ，革新的な創薬シーズを効率よく探索する試みの一つとして，コンピュータ解析（①データベース，②統計的モデリング，③数理モデリング）が積極的に利用されつつある．年々増加の一途を辿るデータベースを有効に利用するために，データベースの統合や，格納されたデータの解析を支援するプラットフォームの構築といった試みが需要を増している．また，収集された大量のデータから生物学的に意味のある情報・知識を引き出す技術が必要となることから，機械学習の手法の重要性は高い．一方，利用できるデータ量が不十分である場合などにおいても，理論計算によってシミュレーションをおこなうことにより観測している現象の本質を推定するアプローチも有効であり，これらの手法の特徴を理解した上で目的に応じたコンピュータ解析をおこなうことが肝要である．

1. はじめに

製薬企業における研究開発費が占める割合は他業界と比較して大きく，さらに近年ではその生産性が1990年代後期から2000年代初期と比較して低いことが指摘されている(1)．その一方で，1つの薬が実際に市場に出回るまでに10年かかると言われていた上に，予期せぬ副作用などの理由で販売停止となることもある．創薬研究には膨大な時間，労力，費用といった様々なコストが重くのしかかっており，この状況は間接的には患者数の少ない疾患に対する創薬の遅れなどを招いてきた．創薬研究におけるこれらのコストを下げ，革新的な創薬シーズを効率よく探索する試みの一つとして，コンピュータ解析が積極的に利用されつつある．

近年の技術進歩に伴って低分子化合物や生体分子などの詳細な情報がデータベース化され全世界で共有されるようになっただけでなく，ハイスループットの実験によって各製薬企業や研究施設単位でも大規模なデータを生み出すことが可能となった．これらのデータを活用することにより，これまでの既知情報のみに囚われずより多くの化合物を対象に創薬シーズの候補化合物をスクリーニングすることができる．特定の少数の分子間の物理的相互作用の理解のみならず，コンピュータ解析を通して生物のあり方をシステムとして理解するアプローチは計算システム生物学と位置づけられ，創薬研究においても近年このようなアプローチが注目を集めている．そこで，創薬研究における計算システム生物学の基本要素を①データベース，②統計的モデリング，③数理モデリングの3つに分類し，それぞれについて具体例を挙げながら紹介する．

2. データベース

生物の体内では多様な生体分子が相互作用しながら一つの大きなシステムを形成している．そのため，これらの生体分子に関する生物学データも多岐にわたり，これらを格納するデータベースの数は年々増加している．研究開発費が占める割合が大きい医薬品産業において，これらの既に公開されている情報を有効に活用することで新規の仮説を導き出す試みは，増大し続ける研究開発費に歯止めをかけるだけではなく，既存の医薬品とは異なる作用機構を有する医薬品の創製にもつながると期待される．このような効果を実現するためには，個々の研究グループが独自に取得した実験データなどのフォーマットを統一化するなど，同じ実験から得られたデータ・同じ項目に関するデータが共

キーワード：データベース，統計的モデリング，数理モデリング，機械学習
国立研究開発法人医薬基盤・健康・栄養研究所 バイオインフォマティクスプロジェクト
(〒567-0085 大阪府茨木市彩都あさぎ7-6-8)

E-mail: kenji@nibiohn.go.jp 原稿受領日：2016年11月1日，依頼原稿

Title: Computational systems biology for drug discovery: from molecules, structures to networks

Author: Yayoi Natsume-Kitatani, Kenji Mizuguchi

通の枠組みで記述され同時使用が可能である必要がある。しかしこれまでも論じられている通り(2), データ統合には様々な困難が伴う。そのため, Chen らは多様な公共データベースを統合した「統合データウェアハウス」である TargetMine(3)を開発したほか, 近年ではコンソーシアムによりあらかじめ実験条件や数値の単位などを一定にすることによって量と質の両方を重視したデータベース構築を計画的に進める動きが盛んになりつつある。

そのような例の一つとして, トキシコゲノミクスプロジェクト(4)が挙げられよう。これは医薬基盤研究所(現, 医薬基盤・健康・栄養研究所), 国立医薬品食品衛生研究所と日本国内の主要製薬企業15社が参加したコンソーシアム(第1期:2002~2006年度, 第2期:2007~2011年度)であり, 肝毒性が既に報告されている化合物を中心に約200化合物について, ラット(in vivo/in vitro)やヒト(in vitro)の複数の臓器に対する遺伝子発現プロファイルへの影響をマイクロアレイによってデータ収集をおこない, 投薬後のラットにおける生理学的情報と紐づけてデータベース化(Open TG-GATEs(5))したものである。マイクロアレイの実験データはおおよそ24,000枚分にも及び, かつ均一の実験条件化で同一研究施設において収集されたことから量・質ともに世界的にも類を見ないデータベースの構築に成功しており, Open TG-GATEsは現在に於いてもトキシコゲノミクスの分野において世界最大の公共データベースの一つとして広く利用されている。さらに, このデータベースの活用を幅広げるため, バイオインフォマティクスの技術を必要とせず

にオンラインでデータ解析をおこなうためのインターフェイスとしてToxygates(6)が開発された。通常, データベースに格納された情報を用いてデータを解析する場合, 必要なデータをダウンロードし, データ解析のためにプログラミングをおこなう必要がある。バイオインフォマティクスの技術がある場合においても, 大容量データのダウンロードには長い時間を要するほかデータプロセッシングにかかる手間も無視することができない。また, データ解析をおこなうウェブツールが使用可能な状況にあっても, 実際にアップロード・解析できるデータの容量には限度がある。これらの問題点を打破するため, Toxygatesではウェブ上で興味のある(解析対象である)実験データを選択することにより簡単に遺伝子発現量を表示することができるほか, 選択した遺伝子リストを前述のTargetMineに転送することによって, ある特定の遺伝子機能注釈(Gene Ontology term)やパスウェイなどが遺伝子リスト内で有意に共有されているかを網羅的に解析する機能を搭載している(図1)。

3. 統計的モデリング

統計的モデリングとは, 観測可能なデータに基づいてその母集団の特徴を推定することを可能にするモデルを構築することを指す(例えば, 既に観測されている化合物の分子量とある物性値データから, それらの2つが直線的な関係にあると考えるのは, もっとも簡単なモデリングの例と言える)。データ収集・格納により構築されたデータベースの恩恵を受けるためには, 格納されたデータから生物学的に意味のある情報・

図1 公開実験データのデータベースとその閲覧・解析プラットフォーム

トキシコゲノミクスプロジェクトで取得されたマイクロアレイデータ、病理学的データなどはデータベース(Open TG-GATEs)に格納された形で公開されている。これらの実験データをウェブ上で閲覧・解析をするプラットフォーム(Toxygates)が開発された。

知識を引き出す技術が必要となる。例えば、手元にマイクロアレイの実験結果があったとしても、数値を人の目で見て何らかの遺伝子発現パターンを見つけ出すことは多くの場合において困難であり、そのようなアプローチは解析者の持つ事前知識などに大きく左右される。機械学習は、多くのデータが使用可能である場合において、そのデータから上記のようなパターン、傾向などを抽出する技術を指す。創薬研究に於いても機械学習の手法は有効であり、多くの成果につながっている。そこで、創薬研究に於ける機械学習の応用例として、筆者らのグループで開発した予測モデルをいくつか紹介したい。

PSIVER (Protein-protein interaction Sites prediction serVER) は、タンパク質のアミノ酸配列から別のタンパク質と結合する部位を予測する手法である(7)。タンパク質の構造決定には未だ多くの労力を要するため、構造情報に頼ることなくアミノ酸配列のみを使ってタンパク質結合部位を予測することができる点で特に優れた手法であると言える。この手法を用いて創薬シーズの探索をおこなった例として、乳がん細胞の増殖に関与する2つのタンパク質間の相互作用部位の予測が挙げられよう。新規エストロゲンシグナル制御分子 BIG3 は機能が不明の巨大なタンパク質である。BIG3 が既知のエストロゲン受容体抑制タンパク質である PHB2 と相互作用することにより、エストロゲンシグナルが下流に伝播すると乳がん細胞の増殖につながることから、この2つのタンパク質間の相互作用部位の予測をおこなったところ、この相互作用を阻害するペプチドが有望な創薬シーズであることが確認された(8) (図2)。

EFPrf (Enzyme Function Prediction method by random forests) は、酵素の機能を予測する手法である(9)。一般的に酵素がどのような化学反応を触媒するかといった特徴は EC (Enzyme Commission) 番号と呼ばれる階層的な四桁の数字で表される。そのため、酵素の EC 番号を予測することは酵素の機能を基質や反応形式を含め詳細まで予測することを意味する。アミノ酸配列というごく限られた情報から詳細な酵素機能を予測することができれば生化学的実験のコストを低減することにつながり、創薬研究の効率化に寄与することが期待される。

化合物の化学構造情報から標的タンパク質との結合様式や親和性を予測する手法は各種開発されている。しかし、それぞれの標的について、どのような情報が入力として利用可能かを取捨選択し、どのように適切なアルゴリズムを選択するかといったノウハウは十分に一般化されてはいない。Prathipati らは、既知活性化合物との類似性に基づくアプローチに標的タンパク質の構造情報を加味して独自のスコアリングをおこなうプロトコルを提唱し、コミュニティによるブラインドテストで好成績を取めた(10)。この例は、個別の標的タンパク質について文献等から得られた専門家による知識を用いることで、単純な機械学習の適用以上の性能を引き出せることを示しているが、将来的には、人工知能などの利用により、専門家と類似の行動判断を支援するシステムの構築が可能になるかもしれない。

そのほか、創薬研究分野で需要が高く盛んに機械学習の手法を用いた予測がおこなわれている分野としては、化合物の物性予測が挙げられる。化合物の物性とは、化合物が有する分子量や溶解度などの物理的な特

図2 機械学習によるタンパク質結合部位予測と創薬への応用

a) 予測した結合部位と同じアミノ酸配列を有するペプチドが BIG3-PHB2 の結合を阻害、乳がん細胞の増殖阻害活性を呈することを実験で証明した。b) PHB2 はエストロゲンレセプターの転写活性を阻害するが、BIG3 過剰発現化においては BIG3 が PHB2 と結合して核移行を阻害するためエストロゲンシグナルの恒常的活性化が起こる。

性を指し、これらの値は投与された薬物が体内でどのように吸収・分布・代謝・排泄されるかといった薬物動態予測で必要となる。そのため、化合物の物性を高精度で予測することができれば、実験で物性値を取得する負担を軽減するだけでなく薬物動態予測に利用できるデータの拡張ひいては薬物動態の予測精度の向上につながる事が期待される。

このように機械学習の手法は創薬研究への応用に有用であり、既に多くの成果が報告されている一方で、いくつかの問題点も存在する。①機械学習の手法は前述の通り「多くのデータが使用可能である場合において、そのデータから上記のようなパターン、傾向などを抽出する技術」であることから、データ量が不十分である場合において望む結果を得ることができない。②予測精度は学習に用いたデータと予測したいデータの関係性に強く影響を受ける。例えば化合物の物性を予測する場合において、学習に用いた化合物群が形成する特徴から大きく外れる化合物の物性予測は精度が悪くなる傾向がある。③機械学習で予測をおこなった場合、どのような特徴がその予測結果に寄与したのかがブラックボックスであるケースが多い。機械学習の手法はデータから知識を抽出する強力なツールであると言えるが、Prathipati らの研究にも示されている通り、どのような目的を持ってどのようなデータを解析に用いるかといった意思決定の重要性も同様に予測精度や解析結果に多大な影響を与える。これらの問題点も加味して機械学習の手法を創薬研究に生かすことが肝要である。

4. 数理モデリング

前述のように、十分な質と量のデータが利用可能な状況においては統計的モデリングが強力なツールとなりうる。一方、そのようなデータが入手困難な場合においては、何らかの一般的な法則に基づき観測可能な現象の時間変化を微分方程式などの数式で表現することによってその本質を簡略化した形で捉える数理モデリングが有効である。数理モデリングは、実験データなどを利用せず理論計算によってシミュレーションをおこなうことから、広義における第一原理に基づくモデリングにあたる。創薬研究における数理モデリングには、扱う生物学的階層や原理によって方法および統計的モデリングとの組み合わせも様々な場合がありうる。

物理学的原理に基づくシミュレーション：タンパク質などの分子の構造変化や物性を実験的に求めるには多くの困難を伴う。そこで、分子を古典力学に従う原子

の集合体として近似する分子動力学シミュレーションなどが広くおこなわれている。この方法は、(例えば一定の構造を保持しないため) X線結晶構造解析ではデータを得ることができない対象に対して、モデリングを実行できる可能性を持っている。土屋らは、自然免疫応答において必須のアダプタータンパク質である STING (stimulator of interferon genes) がそのリガンドである cGAMP (cyclic GMP-AMP) によってどのように活性化されるのかを分子動力学シミュレーションによって示した(11)。細胞が病原体などに感染すると、病原体や傷ついた細胞由来の二本鎖 DNA 断片から STING リガンドとして機能する cGAMP が産生される。STING はリガンド結合によって活性化して IFN- β 産生を誘導することから、STING リガンドは抗がん薬あるいはワクチンアジュバントの候補として有望であるとして世界的に探索の競争が激化している。そのため、STING とそのリガンドの結合様式を理解することは STING リガンド探索に有効な戦略をとるために有益な情報となりうる。STING の活性化にはその C 末端 (CTT) のリン酸化が必要であることが既に知られていたが、CTT は一定の構造を取らず、X線結晶構造解析ではその振る舞いが全く観測されていなかった。シミュレーションによって STING に cGAMP が結合するとリガンド結合ポケットの蓋 (Lid) が閉じて CTT が固定されリン酸化以降の反応が起こるといった分子機構が推定された (図 3)。数理モデリングと統計的モデリングとの別の形での連携として、原理的には実験的にデータ取得が可能であっても、可能性が多くて実験が困難な場合 (例えば、各種塩基配列に対応する DNA 二重鎖の立体構造など) に、シミュレーションでデータを生成して、機械学習の入力にするとということもおこなわれている。

システムレベルの数理モデル：質量保存則や化学反応の原理に基づいて、物質の経時的な量的変化などを微分方程式で表現しシミュレーションをおこなうアプローチであり、薬物動態やシステム生物学における代謝反応のモデリングなどに利用されている。係数に機械学習で得られたパラメータを用いる形での連携があり得る。またその際、微分方程式を用いて現象を定量的に表現するのではなく、現象を「あり」「なし」というように定性的に表現したモデリングも可能である (例えば、シグナル伝達の系で各タンパク質が活性化状態か、非活性化状態かという二値を取るとして、最終的な生体反応をシミュレーションするなど)。

このように数理モデリングは使用可能な情報が限られている場合においても有効であり、創薬研究で幅広

図3 分子動力学シミュレーションによる受容体-リガンド結合の分子機構推定

a) STING の活性化に寄与する CTT は一定の構造をとらない。Lid はリガンド結合ポケットの蓋として機能する領域を指す。b) STING にリガンドが結合すると、Lid が閉じて CTT が固定される。これにより CTT のリン酸化以降の反応が起こることが推定された。

く利用されている。しかし、数理モデリングには①構築したモデルの前提となる仮定が妥当か否かの判断が難しい、②計算時間が長くなる傾向がある、などの問題点があり、目的に応じたモデリング手法の選択が必要である。

5. まとめと展望

以上に示した通り、データベース、統計的モデリング、数理モデリングは互いにその強みと弱みを補い合う関係にあり、コンピュータ解析の目的や使用可能なリソースなどに応じて使い分けることが望ましい。中でもデータベースを活用したコンピュータ解析ではデータベースに格納されたデータ量やその質による影響が大きい。多岐にわたるハイスループット解析などによるデータの蓄積は今後も加速することが予想されることからデータベースの管理やデータ入力に関するルール整備などはより重要となると考えられる。さらに、ヒトのデータなど秘匿性の高い情報については、データ取得自体が容易でないことから共有することができればその有用性は高いが公共データベース化は依然として進んでいるとは言い難い。創薬研究におい

ては臨床データがこれにあたり、データベースの形態では入手困難であるのが現状である。

このような背景の中、創薬研究におけるコンピュータ解析は創薬ターゲット探索や候補化合物の構造最適化、薬物動態や毒性 (ADMET) の予測・評価と各創薬研究段階でますますその存在感を増しており、今後のブレークスルーの起爆剤となることが期待されている。

著者の利益相反：夏目やよい、水口賢司 (大日本住友製薬株式会社)。

文 献

- 1) Schulze U, et al. Nat Rev Drug Discov. 2014;13:331-332.
- 2) 水口 賢司. SAR News. 2013;24:2-6. (http://bukai.pharm.or.jp/bukai_kozo/SARNews/SARNews_24.pdf).
- 3) Chen YA, et al. PLoS One. 2011;6:e17844.
- 4) Uehara T, et al. Mol Nutr Food Res. 2010;54:218-227.
- 5) Igarashi Y, et al. Nucleic Acids Res. 2015;43:D921-D927.
- 6) Nystrom-Persson J, et al. Bioinformatics. 2013;29:3080-3086.
- 7) Murakami Y, et al. Bioinformatics. 2010;26:1841-1848.
- 8) 水口賢司. 生命科学から創薬へのイノベーション. 南山堂; 2014. p. 86-87.
- 9) Nagao C, et al. PLoS One. 2014;9:e84623.
- 10) Prathipati P, et al. J Comput Aided Mol Des. 2016;30:817-828.
- 11) Tsuchiya Y, et al. EBioMedicine. 2016;9:87-96.