■S2 群（ナノ・量子・バイオ）- 6 編（バイオインフォマティックス）

6 章 システムバイオロジー

（執筆者：稲岡秀檢，福岡　豊，有阪直哉）［2018 年 2 月 受領］

■概要■

生命現象はタンパク質や RNA などの多数の分子が相互作用することで営まれている．分子生物学の発展に伴い，個々の分子の働きが解明されてきた[1]．しかし，生命現象の全体像を明らかにするには至っていない．これまでに述べてきたように，ハイスループットな計測技術が開発され，多くの遺伝子の発現量などの大規模なデータが蓄積されている．一方で，コンピュータの計算速度は年々，高まっている．また，シミュレーション技術の発展も目覚ましいものがあり，非線形システムについても計算論的に解析できるようになっている．

このような背景から，生命現象をシステムとして理解しようという学問分野が生まれた[2]．それがシステムバイオロジーである．この分野のパイオニアの 1 人である北野は，システムバイオロジーを生物学の 1 分野であると位置づけている．システムとしての生命を理解するためには，生物学・生化学的な実験と計算論的解析を組み合わせる必要があるとしている．北野は，バイオインフォマティックスは情報学の 1 分野とし，システムバイオロジーとは異なる分野だとしている．しかし，その境界は曖昧で，両者をほぼ同じ分野と考えることも多い．本章では，システムバイオロジーについて，幾つかの事例を通して考え方を説明する．

【本章の構成】

本章では，システムバイオロジーとその応用事例について紹介する．6-1 節では，システムバイオロジーの概要について説明する．6-2 節では，システムバイオロジーで用いられる研究モデルについて簡単に紹介する．6-3 節では，システムバイオロジーの応用例を研究事例から紹介する．6-4 節では，がんについての応用例を事例を挙げて紹介する．6-5 節では，最近活発に研究が行われている深層学習を応用したシステムバイオロジーの事例について紹介する．

■参考文献

1) B. Vogelstein and K.W. Kinzler：“ p53 Function and Dysfunction, ” Cell, 70, pp.523-526, 1992.
2) H. Kitano：“ Systems biology: a brief overview, ” Science, 295(5560), pp.1662-1664, 2002.

■S2 群 - 6 編 - 6 章

6-1 システムバイオロジーの概要

（執筆者：稲岡秀檢）[2018 年 2 月 受領]

システムレベルで生物学的システムを理解するには，1) システムの構造，2) システムのダイナミックス，3) システムの制御方法，4) システムの設計方法の 4 つについての理解が必要となる[1)].

6-1-1 システムの構造

遺伝子の相互作用と遺伝子の産生物であるタンパク質や RNA などの生化学的ネットワークを同定することは，生物学的システムの構造を理解するための大きな課題である．ネットワークモデルを作成するために，実験により特定の遺伝子の相互作用を同定することと，遺伝子機能に関する広範な文献調査が行われてきた．

既知の機能を有する遺伝子と同時発現する遺伝子を同定するため，クラスタリング分析が用いられているが，得られる情報は，遺伝子と生物学的現象との「相関」であり「因果関係」については明らかとはならない．

6-1-2 システムのダイナミックス

ネットワーク構造が判明すれば，ネットワークのダイナミックスを調べることができる．刺激に応答するシステムの動作を理解し，次の動作を予測するためにはシステムモデルを定義する必要がある．

システムの定常状態における速度定数についての知識をモデルに取り込むことができれば，ダイナミクスを反映するようにモデルを修正しシステムの挙動の変化を予想できる．動的シミュレータと解析ツールの組合せは，生物学的シミュレーションに関する多くの研究で既に使用されている．

6-1-3 システムの制御方法

ロバスト性は生物系の本質的な特性である．生物学的ロバスト性の根底にあるメカニズムと原則を理解することは，システムレベルで生物を理解するために必要となる．ロバスト性は以下の 3 つの領域に分類することができる．

1. **適応性**：環境変化に対処する能力
2. **パラメータ非感受性**：特定の動態パラメータに対するシステムの相対的な非感受性
3. **ゆるやかな劣化**：壊滅的な故障ではなく，損傷後のシステムの機能の特徴的な緩慢な劣化

工学的システムでは，1. は負帰還やフィードフォワード制御で，2. は同等の機能を持つ複数のコンポーネントによる冗長性で，3. は構造的な安定性で実現されている．

同様の性質が生物学的システムにも見られる．細菌の走化性は負帰還の一例であり，冗長性は，細胞周期及び概日リズムの制御において機能する遺伝子に見られる[2)].

6-1-4 システムの分析方法

システムレベルの分析を行うには，網羅的な一連の量的データが必要となる．測定における網羅性は，次の 3 つの側面を考慮する必要がある．

1. **因子の網羅性**： mRNA 転写物及び一度に測定され得るタンパク質の数
2. **タイムラインの網羅性**：測定が行われる時間枠
3. **項目の網羅性**：mRNA 及びタンパク質濃度，リン酸化，低カロリー化などの複数項目の同時測定

6-1-5 知識発見・データマイニング

知識発見あるいはデータマイニングは，膨大な数の実験データから隠されたパターンを抽出し，仮説を形成する手法である．知識発見は，配列からのエキソン-イントロン及びタンパク質構造の予測や，発現プロファイルからの遺伝子調節ネットワークの推論などで広範に使用されている．これらの方法は，ヒューリスティックに基づく予測や，隠れマルコフモデルなどのアプローチを含む統計的弁別法や，ほかの言語ベースのアルゴリズムを使用することが多い．

6-1-6 シミュレーション解析

シミュレーション解析は，*in silico* 実験で仮説を検証し，*in vitro* や *in vivo* 研究によって検証されるべき予測を提供する手法である．

シミュレーションは，基礎となる仮定の妥当性を検証するために，実験結果と比較される．この段階で不一致が生じた場合，検討中のシステムが不完全であることを意味する．実験により検証されたモデルを使用して，別の条件での実験結果を予測することができる．また，実験的な検証に適していない研究についても利用することができる．

6-1-7 ツール

ソフトウェアのインフラストラクチャは，システムバイオロジーのもう一つの重要な要素である．一般的な目的のために設計された分析ツールを利用してシミュレーションソフトウェアが構築されているが，これらのリソースの統合を可能にする共通のインフラストラクチャや，標準な手法はまだ存在しない．Systems Biology Mark-Up Language（SBML）[3, 2]は，CellML [5, 6]とともに，ソフトウェアツール間でモデルを交換できるようにする XML ベースのコンピュータ可読モデル定義の標準を提案している．Systems Biology Workbench（SBW）[7]は SBML 上に構築され，システムバイオロジー研究用のモジュラオープンソースソフトウェアのフレームワークを提供する．

これらは，Kyoto Encyclopedia of Genes and Genomes（KEGG）[8, 9]，The UCSD Signaling Gateway [10, 11]及び Science Signaling [12, 13]など人間が読める形式だけではなく，機械で実行可能なモデルも作成することができ，生物学的パスウェイに関係する新世代のデータベースの価値を大幅に高めた．

■参考文献

1) H. Kitano：“ Systems biology: a brief overview, ” Science, 295(5560), pp.1662-1664, 2002.
2) M. von Schantz and S.N. Archer：“ Clocks, genes and sleep, ” J R Soc Med., 96(10), pp.486-489, 2003.
3) http://www.sbml.org/
4) C. Chaouiya, D. Berenguier, S.M. Keating, A. Naldi, M. P. van Iersel, N. Rodriguez, A. Dräger, F. Büchel, T. Cokelaer, B. Kowal, B. Wicks, E. Gonçalves, J. Dorier, M. Page, P.T. Monteiro, A. von Kamp, I. Xenarios, H. de Jong, M. Hucka, S. Klamt, D. Thieffry, N. Le Novére, J. Saez-Rodriguez, and T. Helikar：“ SBML Qualitative Models: a model representation format and infrastructure to foster interactions between qualitative modelling formalisms and tools, ” BMC Systems Biology, 7:135, 2013.
5) http://www.cellml.org/
6) A.K. Miller, J. Marsh, A. Reeve, A. Garny, R. Britten, M. Halstead, J. Cooper, D.P. Nickerson, and P.F. Nielsen：“ An overview of the CellML API and its implementation, ” BMC Bioinformatics, 11:178, 2010.
7) H. Kitano：“ Standards for modeling, ” Nat Biotechnol., 20(4), 337, 2002.
8) http://www.genome.jp/kegg/
9) M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima：“ KEGG: new perspectives on genomes, pathways, diseases and drugs, ” Nucleic Acids Res., 45(D1), pp.D353-D361, 2017.
10) http://www.signaling-gateway.org/molecule/
11) A.G. Gilman, M.I. Simon, H.R. Bourne, B.A. Harris, R. Long, E.M. Ross, J.T. Stull, R. Taussig, H.R. Bourne, A.P. Arkin, M.H. Cobb, J.G. Cyster, P.N. Devreotes, J.E. Ferrell, D. Fruman, M. Gold, A. Weiss, J.T. Stull, M.J. Berridge, L.C. Cantley, W.A. Catterall, S.R. Coughlin, E.N. Olson, T.F. Smith, J.S. Brugge, D. Botstein, J.E. Dixon, T. Hunter, R.J. Lefkowitz, A.J. Pawson, P.W. Sternberg, H. Varmus, S. Subramaniam, R.S. Sinkovits, J. Li, D. Mock, Y. Ning, B. Saunders, P.C. Sternweis, D. Hilgemann, R.H. Scheuermann, D. DeCamp, R. Hsueh, K.M. Lin, Y. Ni, W.E. Seaman, P.C. Simpson, T.D. O’Connell, T. Roach, M.I. Simon, S. Choi, P. Eversole-Cire, I. Fraser, M.C. Mumby, Y. Zhao, D. Brekken, H. Shu, T. Meyer, G. Chandy, W.D. Heo, J. Liou, N. O’Rourke, M. Verghese, S.M. Mumby, H. Han, H.A. Brown, J.S. Forrester, P. Ivanova, S.B. Milne, P.J. Casey, T.K. Harden, A.P. Arkin, J. Doyle, M.L. Gray, T. Meyer, S. Michnick, M.A. Schmidt, M. Toner, R.Y. Tsien, M. Natarajan, R. Ranganathan, and G.R. Sambrano：“ Overview of the Alliance for Cellular Signaling, ” Nature, 420(6916), pp.703-706, 2002
12) http://stke.sciencemag.org/
13) N.R. Gough：“ Science’s signal transduction knowledge environment: the connections maps database, ” Ann N Y Acad Sci., 971, pp.585-587, 2002.

■S2 群 - 6 編 - 6 章

6-2 システムバイオロジーで利用される計算モデル

（執筆者：稲岡秀檢）[2018 年 2 月 受領]

システムバイオロジーにおける土台となっている主な計算モデリング手法の重要な特徴について説明する[1].

6-2-1 プロセス代数

生物システムのモデリングと解析のために，プロセス代数（並行システムを形式的にモデリングする手法）がシステムバイオロジー研究で利用されている．プロセス代数の表現力は，並行プロセスの相互作用，通信，及び同期に関する数式による記述を可能にする．システムバイオロジーにおいてプロセス代数が利用される理由は，生物学的システムが並行反応システムとして考えられるためである．システムバイオロジーにおけるプロセス代数の主要な例には，Beta-Binders / BlenX [2]（生物システムを解析，シミュレーションするための計算機上のワークベンチ），SPiM [3]（生物学的プロセスモデルのデザイン，シミュレーションのためのプログラミング言語），Bio-PEPA [4]（生物学的システムの解析とモデリングのためのフレームワーク），sCCP [5]（生物学的システムのモデリングのための確率的同時制約プログラミング環境），BioShape [6]（生物学的システムの空間的な形状を基礎としたシミュレーション環境）などがある．

6-2-2 ルールベースシステム

ルールベースのモデリングは，生物の生化学的相互作用をモデル化するためにシステムバイオロジーで使用される化学反応表現と非常に類似している．例えば，酵素（E）が基質（S）に結合し，酵素（E）を放出することによって生成物（P）を生成する古典的な酵素反応を考えてみる．これは，以下のような 2 つの単純なルールを使用することによって，非常にコンパクトで簡潔な記述で表現できる．

$$E + S \leftrightarrow ES$$
$$ES \rightarrow E + P$$

例えば，アミラーゼ（酵素）にデンプン（基質）が結合し，加水分解することで，マルトース（生成物）が放出される．

この手法の重要な特徴の一つは，ルールは独立した単位であり，簡単に交換または変形できることにある．更にルールベースのモデルの単純な構文は，テキスト形式で保存することができ，グラフ表現を使用して編集・視覚化することができる．BIOCHAM [7]，Kappa [8]，BioNetGen [9]などの多くのルールベースモデリング言語及びツールが，システムバイオロジーで利用されている．

6-2-3 ペトリネット

ペトリネットは 2 つのノードを持つ有向グラフである（図 6･1）．バーで表されるトランジションと呼ばれるノードの集合と，円で表されるプレースと呼ばれるノードの集合からな

る．トランジションは起こりうる事象（反応）を意味し，プレースは反応が生じる条件を意味する．矢印で表されるアークは，これらのノードを相互に接続し，流れの方向を示す．プレースはトランジションにのみ接続することができ，その逆も可能である．データは，一般に黒マークによって示される「トークン」として表される．トークンは，トランジションを経て出力プレースで作成され，入力プレースで消費される．トランジションは，直接接続されているプレースの一つに幾つかのトークンが存在することによって使用可能になり，「発火」する．ペトリネットは化学プロセスの記述の目的で開発されたが，並行分散システムを解析・特定するためにコンピュータサイエンスにおいても集中的に利用された．直感的かつグラフィカルなモデリングスタイルを持つため，システムバイオロジーにおいても普及している．ペトリネットは，定性的（ペトリネットの静的構造トポロジーによって与えられる）と定量的（トークン分布の時間発展によって与えられる）分析の両方が緊密に統合されたフレームワークを提供する．システムバイオロジーで使用されるペトリネットによるツールとしては，Snoopy [10]，MARCIE [11]，GreatSPN [12]，Pathway Logic Assistant [13]がある．

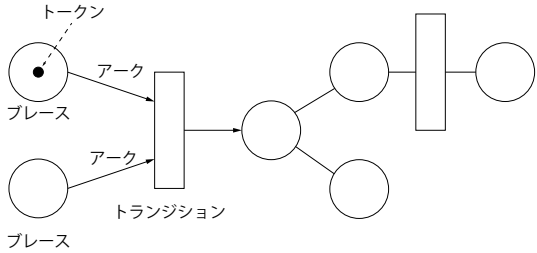

図 6･1　ペトリネットの一例

6-2-4 ブーリアンネットワーク／質的ネットワーク

ブーリアンネットワークは，活性化（真の状態）または失活（偽の状態）のいずれかを考慮することによって，遺伝的調節ネットワークのダイナミクスを近似するために使用されることが多い．ブーリアンネットワークは，ブール型変数によって定義される．ブーリアンネットワークのモデリング技術は，遺伝子制御ネットワークの堅牢性と安定性を分析するために広く利用されている．システムバイオロジーにおけるブーリアンネットワーク解析の関連ツールとしては，GINsim [14]，BoolNet [15]及び BNS [16]がある．

新たに提案された定性的ネットワークとして，ネットワークの要素が 2 値ではなく，有限個の値を想定したブーリアンネットワークの拡張がある．この拡張は，単なるブール値よりも高い柔軟性を提供する．定性的ネットワークのモデリングと解析のためのツールはとしては，Bio Model Analyzer（BMA）[17]がある．

6-2-5 ステートチャート

生物システムのダイナミクスをモデル化するための方法の一つに，その行動を特徴づける状態のシーケンスを特定するものがある．例えば，タンパク質にリン酸基が付加されると，その機能的挙動がリン酸化状態に変化し，タンパク質-タンパク質相互作用が可能になる．シ

ステムは，何らかの事象（例えば，遺伝子の活性化または阻害）の発生が，その内部の挙動をある状態から別の状態へ移動させるまでは，現在の状態にとどまる．この特徴により，生物学的システムは，異なるレベル（分子，細胞，組織，器官，または個体レベル）あるいは同じレベル間で生じる，もしくは異なるタイミング及び順序で発生し，行動を決定するような，イベント駆動の並行相互作用である多スケールの反応系と考えることができる．

ステートチャートは，生物学的システムのモデリングの複雑さを扱う適切な構造（遷移，事象，条件，直行領域などの状態の階層）を提供できるため，システムバイオロジーで広く利用されている（図 6･2）．ただし，古典的なステートチャートの表記法では，可能な組合せのパラメータを別個の状態として指定する必要があり，状態の数が爆発的に増加する．システムバイオロジーに最も関連があるステートチャート用のツールとしては IBM Rational Rhapsody [18]がある．

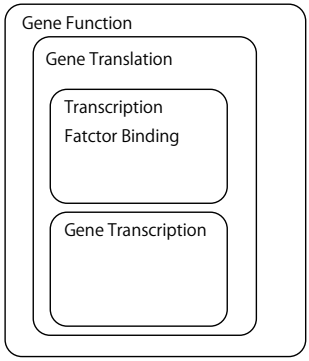

図 6･2　ステートチャートの一例

6-2-6 ハイブリッドシステム

ハイブリッドシステムは，各状態において連続的なダイナミクス（一般には常微分方程式）を用いて状態ベースの離散表現（一般には定性値）を拡張する．ハイブリッドモデリング技術は，明確なスイッチング特性を示す生物学的システムの挙動の研究に関して注目を集めている．一般的にハイブリッドモデリングは，定性的情報（離散状態によって与えられる）と定量的情報（連続的な力学によって与えられる）を結合することに適している．幾つかのハイブリッド系同定法が提案されている．これらの同定法は，区分的線形または区分的多相アフィン関数を用いて複雑な非線形動力学を近似し，非線形モデルを数理解析できるようにする目的や，多細胞の大規模シミュレーションを改善するために提案されている．生物学におけるハイブリッドシステムモデリングツールとしては，Rovergene [19]，BioDivine [20]，Breach [?]，dReach [21], S-TaLiRo [22]がある．

6-2-7 時空間モデル

連続状態決定論的時空間システムは，一般に半線形放物型偏微分方程式の形をとる「反応拡

散」システムの観点から定式化される．離散状態設定では，細胞構造の集合的挙動をシミュレートするために，コンパートメントモデル（細胞膜計算モデル），エージェントベースモデル，及び格子モデル（セルオートマトン，セルラーポッツモデル）が用いられる．これらのモデルはすべて，局所及び非局所の相互作用から生じる様々な行動を表現する．

(1) コンパートメントモデル

生物学的システムは，一般に，コンパートメント（細胞膜，細胞核，細胞小器官）に組織化され，一定の規則に従ってそれらの間で分子を交換する．コンパートメントモデルは，コンパートメントの動的再配列（ミトコンドリアで観察される典型的な挙動），及びそれらの間の分子の輸送などの幾つかの生物学的特徴を捕捉するように特化されている．

コンパートメントモデルに関連するモデリングフレームワークとして BioAmbients [23]がある．BioAmbients はコンパートメント間の統合，分割，及び通信を指定できる特別な演算子を持つプロセス代数である．BAM [24]は，統計的な BioAmbients を実行するためのツールである．

(2) エージェントベースモデル

エージェントベースモデルでは，エージェントと呼ばれる自律的な意思決定エンティティ（一つの物事を表すひとまとまりのデータ）の集合を考える．エージェントは，環境を個別に感知し，一連のルールに基づいて意思決定を行う．エージェントベースモデルは，エージェントとエージェントの間の関係を記述する必要がある．エージェント間の関係は，環境への対応あるいは近隣のエージェントへの行動（競争や協力）といった反応におけるエージェントの変化と適応の行動パターンを示す．

集団内のすべてのエージェントは識別できるので，エージェント固有の履歴や行動を持つことができる．より複雑なエージェントベースモデルでは，ニューラルネットワーク，進化的アルゴリズムなどの技術に基づく学習及び適応ルールをモデルに組み込む．

単一細胞ベースモデルは，エージェントベースモデルの最も有望なものの一つである．このモデルでは，エージェントは多くの細胞の機能的及び構造的特徴，及び，より現実的な挙動を有し，生物システムの異なる中間スケールで現象の検出を可能にする．

細胞ベースモデルは，複製のダイナミクス及びその発達の各段階（細胞の形状，サイズ及び機械的特性）に関する情報など細胞の重要な行動特性を表すことができる．

単一細胞ベースのモデルは，FLAME [25]及び REPAST [26]を使用して実装することが可能である．

(3) 格子モデル

格子は，同一の n 次元閉鎖グリッドサイトによって形成され，各方向の周期的または固定の境界条件によって特徴付けられるグラフが，規則的に繰り返されることで定義される．この格子によるモデルは，分子レベル，細胞レベル，組織レベルまたは器官レベルの相互接続プロセスのシステム記述に特に適している．

セルラーオートマトン[27]は，空間，時間，状態で離散化された動的システムである．細胞パターン形成は，近距離（接着力及び細胞-細胞シグナル伝達など）及び遠距離（機械的ストレス場または拡散化学物質など）の相互作用から生じるものとみなすことができる．

Bethe 格子[28]は，免疫学的ネットワークに適用された，末端を持たない階層的に配置された周期に依存しないネットワークである．

マルチスケール格子モデルは，生物全体から分子レベルまでほとんどすべてのスケールで何が起こるかを観察することができる．しかし，すべてのレベルをまとめることは非常に困難である．まとめ上げるためには，複数の空間スケールにわたるモデルのスケールアップと均一化，及び複数の時間スケールをすり合わせる技術が必要となる．

この問題は，規則的な格子上の確率的モンテカルロ法に基づくセルラーポッツモデルのようにエネルギーを考慮することによって克服することができる可能性がある．このモデルでは，単細胞生物，細胞のクラスタ，個々の細胞のような一般化された個別の細胞や，栄養素あるいは低分子の勾配のような連続的な場における，細胞-細胞接着または細胞-栄養素相互作用などのプロセスのエネルギー記述に注目している．

セルラーポッツモデルの開発のためのフレームワークとしては，細胞，組織及び器官レベルで様々な解剖学的及び病理学的条件をモデル化するために使用されている CompuCell3D [29]がある．このフレームワークでは，直感的な生物学的記述を用いて，プロセスの厳密なエネルギー的及び機械的処理の両方を組み合わせることができる．

6-2-8 形式分析

計算モデルは，生物学的プロセスをシミュレートするコンピュータプログラムに変換される．生成されたプログラムは，特定の初期条件を持つシステムの行動を予測するために使用できる．その結果，コストのかかる実験の数を減らすことができ，仕組みを明らかにすることが期待される対象にのみ，すべての努力と資源を集中させることができる．

計算モデルの別の利点は，プログラムの挙動の正当性を正式にチェックできることである．システムバイオロジーにおいて，これらの手法は，推論及びモデルの分析，実験結果の検証，関心のある挙動の自動的な確認，及びシステムの入力またはパラメータの同定に非常に有用である．

プログラムの形式分析を行う，すなわちプログラムを形式的に検証するとは，プログラムの実行の結果が，そのプログラムから予想されうる仕様を満たすことを証明することである．

以下に，プログラムを検証するために設計されたモデル検査，実行時検証，及び静的解析という生物学的モデルを分析するために広く使用されている形式検証手法について概説する．

(1) モデル検査

モデル検査は，生物学的モデルにおける特定の挙動の出現を確認することができる自動的な形式的検証技術である．この手法は，Kripke 構造[30]と呼ばれる有限数の状態を持つ離散時間モデルで動作する．Kripke 構造は，特別なラベル付きグラフであり，ノードは生物学モデルを実行することによって生成される到達可能な状態を表し，エッジは状態遷移を表す（図 6･3）．ラベリング機能は，各ノードを，対応する到達可能な状態に保持されている命題集合に対応させる．遷移関係は，各状態について可能な継承者の集合を指定する．

NuSMV [31]や CADP [32]などのツールは，生物学的モデルを，非常に効率的なモデルチェッカで分析できる Kripke 構造の表現に変換することができる．この手法の主な欠点は，モデルの状態数がそのパラメータ数で指数関数的に増加し，状態爆発問題を引き起こすことである．

モデル検査技術は，連続及び離散時間マルコフ連鎖（CTMC や DTMC），ペトリネット，ハイブリッドシステム，空間格子モデルのような，Kripke 構造とみなすことができるほかの多くの計算モデルに拡張されている．CTMC 及び DTMC の場合，PRISM [33]のような確率

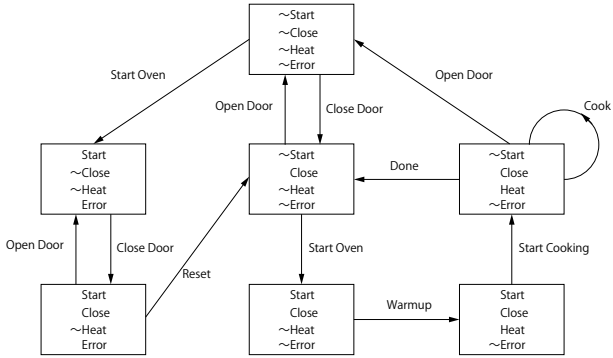

図 6･3　Kripke 構造の一例（オーブンの動作）

的なモデル検査を使用することができる．

（2）実行時検証／監視

モデル検査の状態爆発の問題を克服する方法は，全体的な流れの検証を行う代わりに，現在の単一の実行トレースに分析を集中させることである．実行時検証は，プログラムの現在の実行（例えば，遺伝子調節ネットワークシミュレーション中のタンパク質発現の濃度の時系列）が関心のある性質を満たしているか違反しているかどうかを確認することを目的とした軽量で強力な検証手法である．

（3）静的解析

「静的」という用語が示すように，この分析は実際にはモデルを動作させずにモデルの静的な記述に対して実行される．モデル検査は，一般にモデルのセマンティクスを実行することによって発生したすべての状態を探索する必要がある．しかし，静的解析は，仕様の構文レベル，あるいは実行可能なモデルの有限近似に対する抽象的な解釈に対して行うことができる．静的分析は，基礎となる具体的な計算をすべて行うことなく，モデル仕様に関する重要な情報（例えば，制御構造，種濃度の流れ，種間の相互作用）を明らかにすることができる．静的分析は生物学的モデルを分析するための有用な技術となった．生物学的パスウエイを解析するために使用され，成功したこともある[34]．

静的解析は，実際のシステムの複雑さに対処するために不可欠である．モデル検査では，状態の爆発の問題のために，すべての反応シーケンスのチェックは失敗するからである．

本節で説明した内容は，6-1 節で述べた分類のうち，「分析方法の検討 」にあたる．

■参考文献

1) E. Bartocci and P. Lió :“ Computational Modeling, Formal Analysis, and Tools for Systems Biology, ” PLoS Comput Biol., 12(1), e1004591, 2016.
2) L. Dematté, C. Priami, and A. Romanel :“ The Beta Workbench: a computational tool to study the dynamics of biological systems, ” Brief Bioinform., 9(5), pp.437-449, 2008.
3) A. Phillips and L. Cardelli :“ Efficient, Correct Simulation of Biological Processes in the Stochastic Pi-calculus, ” I. Proc. of CMSB 2007, The 6th Conference on Computational Methods in Systems Biology, 4695, pp.184-199, 2007.

4) F. Ciocchetta and J. Hillston：“ Bio-PEPA: A framework for the modelling and analysis of biological systems, ” Theoretical Computer Science, 410(33-34), pp.3065-3084, 2009.
5) L. Bortolussi and A. Policriti：“ Modeling Biological Systems in Stochastic Concurrent Constraint Programming, ” Constraints, 13(1-2), pp.66-90, 2008.
6) E. Bartocci, F. Corradini, B. Di, R. Maria, E. Merelli, and L. Tesei：“ A Spatial Mobile Calculus for 3D Shapes, ” Scientific Annals of Computer Science, 20(1), 2010.
7) N. Chabrier-Rivier, F. Fages, and S. Soliman：“ The Biochemical Abstract Machine BIOCHAM, ”International Conference on Computational Methods in Systems Biology, 3082, pp.172-191, 2005.
8) V. Danos, J. Feret, W. Fontana, R. Harmer, and J. Krivine：“ Rule-Based Modelling of Cellular Signalling, ” International Conference on Concurrency Theory, 4703, pp.17-41, 2007.
9) E.M. Clarke, J.R. Faeder, C.J. Langmead, L.A. Harris, S.K. Jha, and A. Legay：“ Statistical model checking in biolab: Applications to the automated analysis of t-cell receptor signaling pathway, ”International Conference on Computational Methods in Systems Biology, 5307, pp.231-250, 2008.
10) M. Heiner,M. Herajy,F. Liu, C. Rohr, and M. Schwarick：“ Snoopy - A Unifying Petri Net Tool, ” International Conference on Application and Theory of Petri Nets, 7347, pp.398-407, 2012.
11) M. Heiner, C. Rohr, and M. Schwarick：“ MARCIE-Model Checking and Reachability Analysis Done Efficiently, ” International Conference on Application and Theory of Petri Nets and Concurrency, 7927, pp.389-399, 2013.
12) S. Baarir, M. Beccuti, D. Cerotti, M. De Pierro, S. Donatelli, and G. Franceschinis：“ The GreatSPN tool: recent enhancements, ” ACM SIGMETRICS Perform Eval Rev., 36(4), pp.4-9, 2009.
13) C. Talcott nad D.L. Dill：“ The Pathway Logic Assistant, ” International Conference on Proceedings of the Workshop Computational Methods in Systems Biology (CMSB), pp.228-239, 2005.
14) C. Chaouiya, A. Naldi, and D. Thieffry：“ Logical modelling of gene regulatory networks with GINsim, ” Methods Mol Biol., 804, pp.463-79, 2012.
15) C. Müssel, M. Hopfensitz, and H.A. Kestler：“ BoolNet - an R package for generation, reconstruction and analysis of Boolean networks, ” Bioinformatics, 26(10), pp.1378-1380, 2010.
16) E. Dubrova, M. Teslenko：“ A SAT-based algorithm for finding attractors in synchronous Boolean networks, ” IEEE/ACM Trans Comput Biol Bioinform., 8(5), pp.1393-1399, 2011.
17) D. Benque, S. Bourton, C. Cockerton, B. Cook, J. Fisher, S. Ishtiaq, N. Piterman, A. Taylor, and M.Y. Vardi：“ BMA: Visual Tool for Modeling and Analyzing Biological Networks, ”International Conference on Computer Aided Verification, 7358, pp.686-692, 2012.
18) N. Kam, I.R. Cohen, and D. Harel：“ The immune system as a reactive system: Modeling t cell activation with statecharts, ” Proceedings IEEE Symposia on Human-Centric Computing Languages and Environments, pp.15-22, 2001.
19) G. Batt, B. Yordanov, R. Weiss, and C. Belta：“ Robustness analysis and tuning of synthetic gene networks, ” Bioinformatics, 23(18), pp.2415-2422, 2007.
20) J. Barnat, L. Brim, I. Černá, S. Dražan, J. Fabriková, J. Láník, D. Šafránek, and H. Ma：“ BioDiVinE: A Framework for Parallel Analysis of Biological Models, ” International Workshop on Computational Models for Cell Processes 6, pp.31-45, 2009.
21) S. Kong, S. Gao, W. Chen, and E. Clarke：“ dReach: δ-Reachability Analysis for Hybrid Systems, ” International Conference on Tools and Algorithms for the Construction and Analysis of Systems, 9035, pp.200-205, 2015.
22) Y. Annpureddy, C. Liu, G. Fainekos, and S. Sankaranarayanan：“ S-TaLiRo: A Tool for Temporal Logic Falsification for Hybrid Systems, ” International Conference on Tools and Algorithms for the Construction and Analysis of Systems, 6605, pp.254-257, 2011.
23) A. Regev, E.M.Panina, W. Silverman, L. Cardelli, and E. Shapiro：“ BioAmbients: an abstraction for

biological compartments, ” Theoretical Computer Science, 325(1), pp.141-167, 2004.
24) V.A. Muganthan, A. Phillips, and M.G. Vigliotti：“ BAM: BioAmbient machine, ”International Conference on Application of Concurrency to System Design, pp.45-49, 2008.
25) P. Richmond, D. Walker, S. Coakley, and D. Romano：“ High performance cellular level agent-based simulation with FLAME for the GPU, ” Brief Bioinform., 11(3), pp.334-347, 2010.
26) M.J. North, N. Collier, and J.R. Vos：“ Experiences creating three implementations of the REPAST agent modeling toolkit, ” ACM Trans Model Comput Simul., 16(1), pp.1-25, 2006.
27) A. Deutsch and S. Dormann：“ Cellular Automaton Modeling of Biological Pattern Formation: Characterization, Applications, and Analysis, ” Genetic Programming and Evolvable Machines, 8(1), pp.105-106, 2007.
28) H.A. Bethe：“ Statistical Theory of Superlattices, ” Proceedings of the Royal Society of London Series A, Mathematical and Physical Sciences, 150(871), pp.552-575, 1935.
29) J.A. Izaguirre, R. Chaturvedi, C. Huang, T. Cickovski, J. Coffland, G. Thomas, G. Forgacs, M. Alber, G. Hentschel, S.A. Newman, and J.A. Glazier：“ COMPUCELL, a multi-model framework for simulation of morphogenesis, ” Bioinformatics, 20(7), pp.1129-1137, 2004.
30) S.A. Kripk：“ Semantical Considerations on Modal Logic, ” Acta Philosophica Fennica, 16, pp.83-94, 1963.
31) A. Cimatti, E. Clarke, F. Giunchiglia, and M. Roveri：“ NuSMV: A New Symbolic Model Checker, ” International Journal on Software Tools for Technology Transfer, 2(4), pp.410-425, 2000.
32) H. Garavel, F. Lang, R. Mateescu, and W. Serwe：“ CADP 2010: A Toolbox for the Construction and Analysis of Distributed Processes, ”International Conference on Tools and Algorithms for the Construction and Analysis of Systems, 6605, pp.372-387, 20011.
33) M. Kwiatkowska, G.Norman, and D. Parker：“ Probabilistic Model Checking for Systems Biology, ” Symbolic Systems Biology, Jones and Bartlett, pp.31-59, 2010.
34) H. Pilegaard, F. Nielson, and H.R. Nielson：“ Pathway analysis for BioAmbients, ”The Journal of Logic and Algebraic Programming, 77, pp.92-130, 2008.

■S2 群 - 6 編 - 6 章

6-3 システムバイオロジーの応用例

（執筆者：稲岡秀檢，福岡　豊）［2018 年 2 月 受領］

6-3-1 創薬及びシステム治療

計算，実験，観測の組合せによるシステムバイオロジーの手法は，創薬や個々の患者の治療体制の最適化に非常に関連している[1)]．個々の一塩基多型（SNP）の分析は，あらゆる形態の病理学的状態に対する個々の遺伝的感受性を明らかにすることが期待されるが，複雑な相互作用が関与する場合，このような関係を特定することは不可能である．

例えば，ある遺伝子 A の変異が特定の疾患を誘発する例を考えてみる．可変性の影響を補償する回路が存在する場合，感受性の関係は明らかではないかもしれない．これらの補償回路が何らかの理由で故障した場合に限り，遺伝子 A の多型は，疾患感受性に関連する．創薬と治療の最適化のための新たな機会を創出する可能性のある複数の遺伝子を含む，より複雑な関係を解明するには，より機械的なシステムベースの解析が必要である．

伝統的なバイオインフォマティクスアプローチに加えて，コンピュータシミュレーションと分析は，創薬効率を大幅に向上させるために頻繁に用いられている．経験的 ADME/Tox（吸収分布代謝排泄/毒性）及び薬物動態予測は，ある程度成功して使用されている．例えば，300 を超える化合物の受動透過測定と，水素結合供与体，水素結合受容体，及び分子量などの既知の構造的特徴との間の相関に基づくヒト腸管吸収モデルが，新規化合物の吸収を予測するために使用されている[2)]．

本節は，6-1 節で述べた分類のうち，「シミュレーション解析」にあたる．

6-3-2 スケールアップ

これまで，システムバイオロジーにおけるほとんどのシミュレーションは，細菌の走化性，概日リズム，シグナル伝達経路の一部，細胞周期の単純化モデル及び赤血球のフィードバック回路など，細胞内の比較的小さなサブネットワークを対象とする傾向があった．最近では，大規模なシミュレーションの研究が始まっている．生化学ネットワークのレベルでは，上皮成長因子（EGF）シグナル伝達カスケードのシミュレーションが行われている．このシミュレーションは，100 以上の方程式とパラメータを含み，経路の複雑な挙動を予測し，外部及び内部の EGF 受容体の役割を同定するために使用される．このフィジオームプロジェクトは，*in silico* の器官の本質的な特徴を表す仮想的な器官を作り出すための試みである[3)]．

心臓のシミュレーションは，遺伝学から生理学へのモデルの複数のスケールを統合するこの方向の初期の試みの一つである[4)]．疾患の発症及び創薬の予測のために，肥満及び糖尿病などの特定の疾患に対する全身モデルでさえも開発されている．シミュレーションでは，スケールや質的特性（遺伝子規制，生化学的ネットワーク，細胞間コミュニケーション，組織，器官，患者など）の点で異なるオーダーのモデルの複数の階層を統合する必要がある．プロセスは，確率的計算または微分方程式のいずれかによってモデル化することができるが，多くの場合，両方の方法の組合せが必要である．

しかし，生化学的プロセスは 1 ミリ秒以内に起こるが，ほかのものは数時間または数日かかることがある．従来，生物学的プロセスは，タンパク質輸送，染色体動態，細胞移動また

は組織の形態変化に結合した生化学的ネットワークなど，異なるタイプのプロセスの相互作用を含むことが多い．生化学的ネットワークは，微分方程式と確率的シミュレーションを用いて合理的にモデル化することができるが，多くの細胞生物学的現象は，構造力学，弾性体の変形，ばね質量モデル，及びほかの物理的過程の計算を必要とする．

本節は，6-1 節で述べた分類のうち，「ダイナミクスの解析，シミュレーション解析」にあたる．

6-3-3 バイオマーカ発見への応用例

3 章で述べたように，DNA マイクロアレイや次世代シーケンサ（NGS）などのハイスループット計測法が発展し，遺伝子に関する網羅的なデータが蓄積されている．そのようななかで，様々な疾患と関連する遺伝子の変化（発現量や配列の変異）を調べる研究が進められている．このようなスクリーニングを通じて，疾患のマーカとなる変化が見出されることが期待されている[5, 6]．

このようなアプローチでは，DNA マイクロアレイ[5, 7]や NGS[6, 8]を用いて，特定の疾患患者群と対象群の間で有意に異なる変化を探索する．DNA マイクロアレイでは発現の変化を対象とするのに対し，NGS では配列の変異（DNA-seq）あるいは発現と配列の両方の変化（RNA-seq）を対象とする．いずれの場合も，多くの候補のなかから何らかの方法で絞込みを行う必要がある．最終的には，生物学的な検証が必要である．したがって，信頼性の低い多数のマーカ候補が得られてもあまり意味がなく，少数であっても信頼性の高い候補が得られるのが望ましい．

本節は，6-1 節で述べた分類のうち，「知識発見」にあたる．

6-3-4 創薬への応用例

新薬の開発には膨大な時間とコストを要することや，臨床試験における副作用の発生も十分に把握できないことから，新医薬品の発売数は減少している．この事態を打開する方法として，既存薬を現在使われている疾患以外にも適応拡大を行うドラッグリポジショニング[9, 10]（Drug Repositioning または Drug Repurposing）の研究が進められている．一方，特定のシグナル伝達系など少数の遺伝子の挙動について詳細なモデルを作り，疾患と関連した変化を示すものを探すことも試みられている[11]．また，薬剤のターゲットとなるタンパク質の立体構造を予測し，その構造にうまく作用する化合物を探索することも試みられている[12]．

このように，システムバイオロジーの創薬への応用には様々な方法がある．ここでは，必要な化学的・生物学的知識が比較的少なく，情報科学的側面が強いとの理由で，ドラッグリポジショニングについて詳しく説明する．ドラッグリポジショニングには，以下の利点がある[1]．

1. 市販実績があることで臨床レベルにおける安全性と体内動態がヒトで確認されているという確実性

2. 多くの既存のデータを利用することができる低コスト性

3. 特許を既に取得していること，蓄積されたノウハウと材料が存在することによる優位性

増毛剤のミノキシジルをはじめ，これまでに多くのドラッグリポジショニングの事例がある[13]．臨床試験の過程や使用中に，想定したものとは別の効果が偶然見出されるケースが一般的である．最近では薬剤の化学構造や標的タンパク質などといった特徴に基づき，まだ発見されていない作用や効果を予測するドラッグリプロファイリングが注目されている[14]．蓄積されている膨大な生命情報データを活用するために，この方法には人間が行っている判別や予測といった処理をコンピュータで実現させる機械学習が用いられることが多い．

Wang らは薬剤の化学構造，副作用，標的遺伝子のデータによる薬剤間の類似度と薬剤の標的疾患による疾患間の類似度から，ドラッグリポジショニングの候補となる薬剤を機械学習で予測するシステムを提案している[15]．

以下で方法の概要を説明する．

1. 薬剤データベース PubChem [16]から 888 の薬剤の化学構造のデータを抽出
2. 881 の部分構造が存在するかを 0/1 で表し，888 × 881 次元のマトリックスで表現
3. 薬剤間の化学構造の類似度を重み付きコサイン類似度で定義（Chem）
4. 薬剤データベース KEGG BRITE [17], BRENDA [18], SuperTarget [19], DrugBank [20]の情報を統合し，薬剤の標的タンパク質を抽出
5. 標的タンパク質間の類似度を配列の類似度で定義（Inter）
6. 副作用データベース SIDER [21]から 1. の 888 の薬剤の副作用データを抽出
7. 1450 の副作用の有無を 0/1 で表し，888 × 1450 次元のマトリックスで表現
8. 薬剤間の副作用の類似度を重み付きコサイン類似度で定義（Side-effect）
9. OMIM [22]のデータに基づき疾患の類似度を計算
10. サポートベクターマシン（SVM，Kronecker Product Kernel を使用）で薬剤と疾患のペアの関連を予測

上記の方法で構築した SVM を既知の薬剤-疾患ペアデータで検証した．このデータには，593 の薬剤と 313 の疾患が含まれている．その際，上記の Chem，Inter，Side-effect の各データとそれらを統合した Comb を用いて SVM の学習を行い，10 分割交差検証法を用いて比較した．その結果，Chem での AUC（Area Under the Curve）は 0.834，Inert と Side-effect ではそれぞれ 0.889 と 0.894 となった．また，すべてのデータを統合した Comb では 0.902 となり，最も予測性能が良かった．この方法で構築した SVM で薬剤-疾患の関連性を予測して，新規の組合せが得られれば，リポジショニングの候補となる．

多くのグループが同様な方法を研究しているが，本質的な相違点は少ない．データの選択や類似度の定義が異なる程度である．欧米の製薬メーカを中心に，ドラッグリポジショニングをはじめとしたシステムバイオロジーが応用されはじめている．

本節は，6-1 節で述べた分類のうち，「構造の解析，シミュレーション解析」にあたる．

■参考文献

1) J.E. Bailey：“ Reflections on the Scope and the Future of Metabolic Engineering and Its Connections to Functional Genomics and Drug Discovery, ” Metabolic Engineering, 3, pp.111-114, 2001.
2) H. Kitano：“ Computational systems biology, ” Nature 420(6912), pp.206-210, 2002.
3) J.B. Bassingthwaighte：“ Strategies for the physiome project, ”Ann. Biomed. Eng., 28(8), pp.1043-1058, 2000.
4) D. Noble：“ Modeling the Heart–from Genes to Cells to the Whole Organ, ”Science, 295(5560), pp.1678-1682, 2002.
5) Z.Su, H. Fang, H. Hong, L. Shi, W. Zhang, W. Zhang, Y. Zhang, Z. Dong, L.J¿ Lancashire, M. Bessarabova, X. Yang, B. Ning, B. Gong, J. Meehan, J. Xu, W. Ge, R. Perkins, M. Fischer, and W. Tong：“ An investigation of biomarkers derived from legacy microarray data for their utility in the RNA-seq era, ” Genome Biol., 15, 523, 2014.
6) J.P. Lopez, A. Diallo, C. Cruceanu, L.M. Fiori, S. Laboissiere, I. Guillet, J. Fontaine, J. Ragoussis, V. Benes, G. Turecki, and C. Ernst：“ Biomarker discovery: quantification of microRNAs and other small non-coding RNAs using next generation sequencing, ” BMC Med Genomics, 8, 35, 2015.
7) M. Wang, S. Ruan, J. Ming, and F. Dong：“ Nuclear expression of XBPIs is correlated with breast cancer survival: a retrospective analysis based on tissue microarray, ” Onco Targets Ther., 10, pp.5927-5934, 2017.
8) S.G. Pai, B.A. Carneiro, Y.K. Chae, R.L. Costa, A. Kalyan, H.A. Shah, I. Helenowski, A.W. Rademaker, D. Mahalingam, and F.J. Giles：“ Correlation of tumor mutational burden and treatment outcomes in patients with colorectal cancer, ” J Gastrointest Oncol., 8, pp.858-866, 2017.
9) 沼田　稔：“ DR 研究（既存薬剤再開発）への期待と課題, ”医薬剤ジャーナル, 46, pp.1337-1339, 2010.
10) 山西芳裕：“ データ駆動型の創薬―統計的手法を用いて, ” 実験医学増刊, 35, pp.891-894, 2017.
11) 北野宏明 (編)：“ Dr. 北野の 0 から始めるシステムバイオロジー, ” 羊土社, 2015.
12) 岡崎　進，岡本祐幸 (編)：“ 生体系のコンピュータシミュレーション―タンパク質の構造をどこまで予測できるか, ” 化学同人，2002.
13) 水島　徹：“ 創薬剤が危ない　早く・安く・安全な薬剤を届けるドラッグ・リポジショニングのすすめ, ” 講談社, 2015.
14) 水島　徹，難波卓司：“ ドラッグリプロファイリング研究, ” Drug Delivery System, 21, pp.106-112, 2011.
15) Y.Wang, S. Chen, D. Naiyang, and W. Yong：“ Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data, ” PLoS One., 8, e78518, doi:10.1371/journal.pone.0078518.
16) B. Chen, D. Wild, and R. Guha：“ PubChem as a Source of Polypharmacology, ” J Chem Inform Model, 49, pp.2044-2055, 2009.
17) M. Kanehisa, S. Goto, M. Hattori, K.F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, M. Hirakawa：“ From genomics to chemical genomics: new developments in KEGG, ” Nucleic Acids Res., 34, D354-D357, 2006.
18) I. Schomburg, A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn, and D. Schomburg：“ BRENDA, the enzyme database: updates and major new developments, ” Nucleic Acids Res., 32, pp.D431-D433, 2004.
19) S. Günther, M. Kuhn, M. Dunkel, M. Campillos, C. Senger, E. Petsalaki, J. Ahmed, E.G. Urdiales, A. Gewiess, L.J. Jensen, R. Schneider, R. Skoblo, R.B. Russell, P.E. Bourne, P. Bork, and R. Preissner：“ SuperTarget and Matador: resources for exploring drug-target relationships, ” Nucleic Acids Res., 36, pp.D919-D922, 2008.

20) D.S. Wishart, C. Knox, A.C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali：“ DrugBank: a knowledgebase for drugs, drug actions and drug targets, ”Nucleic Acids Res., 36, pp.D901-D906, 2008.
21) M. Kuhn, M. Campillos, I. Letunic, L.J. Jensen, and P. Bork：“ A side effect resource to capture phenotypic effects of drugs, ” Mol Syst Biol., 6, #343, 2010.
22) A. Hamosh, A.F. Scott , J.S. Amberger, C.A. Bocchini, and V.A. McKusick：“ Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders, ” Nucleic Acids Res., 30, pp.52-55, 2002.

■S2 群 - 6 編 - 6 章

6-4 がんのシステムバイオロジー

（執筆者：稲岡秀檢）［2018 年 2 月 受領］

6-4-1 がんとプロテインキナーゼ阻害剤

毎年 1270 万件の新たながんが報告され，その数は 2030 年までに 2170 万人に増加し，およそ 1300 万人の死亡をもたらすと予想される[1]．がんでは，制限なく細胞が分裂し，細胞内のシグナル伝達プロセスも制御されていない状態にある．

細胞内のシグナル伝達プロセスは多くのプロテインキナーゼ（タンパク質リン酸化酵素）によって制御されている．細胞が正常な状態では，プロテインキナーゼは，細胞増殖，分化及びシグナル伝達プロセスを厳密に制御している．しかし，がんでは細胞分裂は制御されず，細胞周期も通常の進行状況とは異なっている．このような細胞分裂の異常はアポトーシスシグナルが正常に動作していないからである．したがって，アポトーシスシグナルの動作に異常を生じさせるプロテインキナーゼを阻害するメカニズムの理解は，薬剤開発と密接に関係してくる．

6-4-2 がんマーカーとしての PI3KCA

ホスホイノシチド 3-キナーゼ（PI3K）タンパク質の突然変異は，細胞分裂，分化，調節，運動性，生存，及び細胞内輸送の制御を乱し，細胞の増殖が生じ，結果としてがんとなる．PI3K ファミリータンパク質である PIK3CA はよく突然変異を起こすタンパク質であり，腫瘍形成タンパク質として知られ，非小細胞肺がん[2]，胃がん[3]，乳がん[4]，卵巣がん[5]などの多くのがんで報告されている．

PIK3CA は多くのがんにおいて重要な役割を有し，異常な細胞増殖を阻害する標的分子として報告されている．PIK3CA は，サブユニットとして p110a を持っており，p110a の突然変異とがんとの関連性が強いことから，p110a はがん阻害剤の有望な薬剤標的であると考えらる．

6-4-3 システムバイオロジーの応用

Cell Designer 4.4 [6]を使用して構築したパスウェイを用いて，PI3K タンパク質に作用する薬物が，酵素の動態的挙動に関して分析された．このソフトウェアは，構造化 XML フォーマットデータを使用し，離散事象シミュレーションフレームワークを容易に取り扱うことができる．このソフトには SBML Squeezer が組み込まれており，このプラグインにより，タンパク質の活性化，阻害や，可逆性または不可逆性を含む生化学パスウェイの酵素動態についてデザインすることができる．

AutoDock Vina [7]を用いて，すべての分子間の結合について調べた．Ligplot + v.1.4.5 ソフトウェア[8]を用いて，タンパク質とリガンドとの間の分子相互作用を予測した．PyMOL ソフトウェア[9]を用いて分子の解釈を行った．毒性については，身体系内の薬物の経時変化に対する吸着，分布，代謝，及び排泄特性を決定するオンラインサーバ mcule.com [10]によって予測された．

薬物が存在する状態，及び薬物がない状態での酵素動態を検証するための時間経過シミュ

レーションを，常微分方程式を解くことで検証した．システムバイオロジー分析の目的は，標的薬剤が細胞に入り，細胞に影響を及ぼす場合に，タンパク質の酵素動態挙動を理解するための最も単純なモデルを構築することであった．常微分方程式を解くためには分子の初期濃度に関する知識を細胞実験から得る必要があるが，シミュレーションでは濃度の最大値の半分の濃度から数値を連続的に変化させて計算を行った．その結果，がん細胞を死に至らせるのに役立つ主要タンパク質の濃度の急激な低下を観察した．

シミュレーションの結果を受け，PI3K の重要な役割，及びタンパク質構造をタンパク質データバンクから検索した．薬物設計アプローチを用いて，PIK3CA 発がんタンパク質をコードする PI3K ファミリーキナーゼタンパク質 p110a とイソキノリン誘導体による分子ドッキングに関する研究を行い，これを抗がん剤治療のために FDA によって既に承認されている薬剤と比較した．

イソキノリン誘導体は，薬物動態学及び毒性予測の結果，すべてのパラメータが人間の使用のために定義された許容範囲内であり，イソキノリン誘導体 6-(4-エチルフェニル)-10-メトキシインドロ [2,1-a] イソキノリンは毒性のリスクが低かった．また，薬物動態学的プロファイルのような薬物を保証するのに必須である良好な薬物動態特性を示した．分子ドッキング研究は，イソキノリン誘導体がほかの薬物と比較して発がん性タンパク質との結合の高い親和性を有することを示した．以上の結果から，キナーゼファミリーのタンパク質を標的とする抗がん剤としてのイソキノリン誘導体化合物が利用可能であるこが示された[11]．

本節は，6-1 節で述べた分類のうち，「制御法の考察，ダイナミクスの解析，シミュレーション解析」にあたる．

■参考文献

1) American Cancer Society：“ Cancer facts & Figures, ” American Cancer Society, Atlanta, 2016.
2) A.L. Richer, J.M. Friel, V.M. Carson, L.J. Inge, and T.G. Whitsett：“ Genomic profiling toward precision medicine in non-small cell lung cancer: getting beyond EGFR, ” Pharmgenom Pers Med., 8, pp.63-79, 2005.
3) J. Shi, D. Yao, W. Liu, N. Wang, H. Lv, G. Zhang, M. Ji, L. Xu, N. He, B. Shi, and P. Hou：“ Highly frequent PIK3CA amplification is associated with poor prognosis in gastric cancer, ”BMC Cancer, 12(1):50, 2012.
4) K.E. Bachman, P. Argani, Y. Samuels, N. Silliman, J. Ptak, S. Szabo, H. Konishi, B. Karakas, B.G. Blair, C. Lin, and B.A. Peters：“ The PIK3CA gene is mutated with high frequency in human breast cancers, ” Cancer Biol Ther., 8, pp.772-775, 2004.
5) L. Shayesteh, Y. Lu, W.L. Kuo, R. Baldocchi, T. Godfrey, C. Collins, D. Pinkel, B. Powell, G.B. Mills, and J.W. Gray：“ PIK3CA is implicated as an oncogene in ovarian cancer, ”Nat Genet, 21(1), pp.99-102, 1999.
6) A. Funahashi,Y. Matsuoka, A. Jouraku, M. Morohashi, N. Kikuchi, and H. Kitano：“ Cell designer 3.5: a versatile modeling tool for biochemical networks, ” Proc IEEE, 96(8), pp.1254-1265, 2008.
7) O. Trott and A.J. Olson：“ AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading, ”J Comput Chem., 31(2), pp.455-461, 2010.
8) R.A. Laskowski and M.B. Swindells：“ LigPlot+: multiple ligand-protein interaction diagrams for drug discovery, ” J Chem Inf Model., 51(10), pp.2778-2786, 2011.
9) D.A. Smith, C. Allerton, A.S. Kalgutkar, H. van de Waterbeemd, D.K. Walker, R. Mannhold, H. Kubinyi, and G. Folkers：“ Pharmacokinetics and metabolism in drug design, ” Wiley, Hoboken, 2012.

10) R. Kiss, M. Sandor, and F.A. Szalai：“ http://Mcule.com: a public web service for drug discovery, ” J Cheminform., 4(1), P17, 2012.
11) D. Arora, R. Chaudhary, A. Singh：“ System Biology Approach to Identify Potential Receptor for Targeting Cancer and Biomolecular Interaction Studies of Indole [2, 1-a] Isoquinoline Derivative as Anticancerous Drug Candidate Against it, ” Interdisciplinary Sciences: Computational Life Sciences, pp.1-10, 2017.

■S2 群 - 6 編 - 6 章

6-5 機械学習／人工知能の応用

（執筆者：有阪直哉）[2018 年 2 月 受領]

機械学習の一種である深層学習（Deep Learning）について，パーセプトロンや畳み込みニューラルネットワークの基礎理論と実装について，また現代で用いられている学習テクニックの一部を説明する．

6-5-1 形式ニューロン

形式ニューロン（Threshold Logic Unit）は 1943 年にマッカロとピッツ（McCulloch-Pitts）により提案された最初の人工ニューロンである[1]．本物のニューロンの重要な特徴のみを反映し，多くの細かな点を無視したモデルである．そのため，神経活動の複雑な動作を記述することはできないし，実際のニューロンのような複雑な特性も表現できない．あくまでコンピュータで実行可能なよう単純化されたモデルである（図 6･1）．

(1) ニューロンのモデル化

ニューロンの重要な動きは樹状突起からシナプスを通じて得たすべての入力を加算し，その値がある閾値を越えると軸索を通じて出力を出すことである．シナプスは結合強度の違いから，それぞれ信号伝達効率が異なる．これらの特徴をモデル化したものが形式ニューロンであり，1) モデルの出力は ON/OFF，2) 入力によって出力が一意に決まる，3) 閾値を超える入力によって発火する，以上 3 つの点からモデル化されている．各シナプスの結合強度は，重み係数により表現できる．なお，出力と同様に，入力も ON/OFF をとる．

n 個の入力があるとき，それぞれの入力ラインに対応する n 個の重み係数がある．形式ニューロンは入力の重み付き総和を計算する演算は

入力の総和 = 入力ライン 1 の重み係数×入力 1 ＋ 入力ライン 2 の重み係数×入力 2 ＋ ... ＋ 入力ライン n の重み係数×入力 n

$$\begin{aligned} u &= w_1 x_1 + w_2 x_2 + \cdots + w_n x_n \\ &= \sum_{i=1}^{n} w_i x_i \end{aligned} \tag{6･1}$$

である．

この入力の総和を閾値と比較する．閾値は 0 に固定し，常に 1 を値として持つ入力 x_0 を新たに加え，重み付き和 $w_0 x_0$ として表されることが多い．このとき，$w_0 x_0$ はバイアス（bias）と呼ばれる．閾値との比較には

$$H(x) = \begin{cases} 1 & (x > 0) \\ 0 & (x \leq 0) \end{cases} \tag{6･2}$$

と表されるヘヴィサイドの階段関数（ステップ関数）が用いられる．以上から，バイアスを用いた形式ニューロンの出力 y は

$$y = H\left(\sum_{i=0}^{n} w_i x_i \right) \tag{6･3}$$

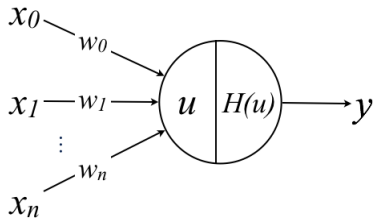

図 6･1　形式ニューロン

となる．入力 x_0 は常に 1 となる．

(2) パーセプトロン

形式ニューロンを結合して作られた 1 層の簡単な人工ニューロンネットワークがパーセプトロン（Perceptron）である．1962 年にフランク・ローゼンブラット（Frank Rosenblatt）によって提案された[2]．形式ニューロンからなるパーセプトロンも，実際の神経系の幾つかの特徴を強調し，ほかの特性は無視されたものであり，神経システムのコピーを意図したものではないとされる．

発火したニューロンの結合が強まるというヘブの学習則を元に，強めるだけでなく弱める場合もあるように修正した学習規則を用いる．集合 A の元を入力したときに 1，集合 B の元を入力したときに 0 を出力したいとする．重みの初期値はランダムである．A の元を判断する際は入力の総和が大きいほうが良いので，ニューロンが誤って 0 と判断した場合は，重みを増加させるよう修正する．一方，B の元を判断する際は入力の総和が小さいほうが良いので，ニューロンが誤って 1 と判断した場合は，重みを減少させる．これは望ましい結果が既知であることが前提であり，この学習方法を教師あり学習と呼ぶ．

限　界

パーセプトロンは学習によってクラスを分類する直線を見つけるが，直線では分類できない場合も少なくない．しかしながら，一層のパーセプトロンは直線によって分類することしかできないため，線形分離が不可能な問題を解くことができない．原理的には，パーセプトロンを多層化すれば線形分離不可能な問題にも適用可能だが，多層化したパーセプトロンではパーセプトロン学習規則で学習することができない．

6-5-2 順伝搬型ネットワーク

(1) 多層パーセプトロン

パーセプトロンでは閾値判定に階段関数を用いられる．しかしこれによって，学習の際にどの程度重みを変更すればよいかが不明となる．そこで，階段関数を変形させたシグモイド関数などの非線形関数を使用することが考えられる．形式ニューロンの階段関数をシグモイド関数に変更したユニットを結合し，入力層，1 層以上の隠れ層，出力層の 3 層以上からなる多層のネットワークを多層パーセプトロン（Multilayer Perceptron：MLP）と呼ぶ（図 6･2）．

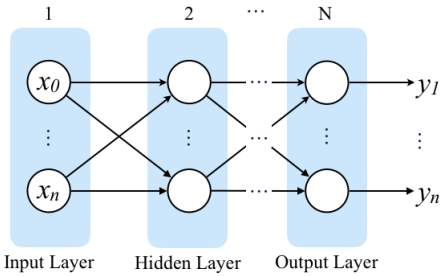

図 6･2　多層パーセプトロン

多層パーセプトロンと名が付いているが，実際に使用されるユニットはパーセプトロンではない．

限　界

入力層を含む，4 層以上の“深い”ネットワークを構成する多層パーセプトロンは，勾配消失などによって学習がうまくいかず，過学習などが発生するため期待した結果が得られない．

(2) 全結合の順伝搬型ニューラルネットワーク

2006 年 Hinton らの Deep Belif Network [5]では，各層を切り離し，それぞれ学習したのち結合し，バックプロパゲーションで学習を行うという事前学習と呼ばれる手法により，ディープなネットワークを学習させることに成功した．第 3 世代 AI ブームのきっかけと考えられている事前学習だが，現在は重みの初期値や活性化関数などを工夫することでディープなネットワークを学習させることができるようになりつつあるため，あまり用いられなくなっている．

6-5-3 誤差逆伝搬法

誤差逆伝搬法（Back Propagation）は重みを効率良く修正する計算方法である[3]．ネットワークの現在の出力と，正しい出力の差異を表す誤差関数 $E(w)$ を定義し，この値を小さくするように重みを調整する．

誤差関数 $E(w)$

勾配法を用いて最小値もしくは極小値を探す目的関数である．二乗誤差や交差エントロピーなどが用いられる．解決すべき問題に応じて変更する．ある入力における二乗誤差は

$$E(w) = \frac{1}{2}\sum_j (t_j - y_j)^2 \tag{6･4}$$

と表される．t が正しい出力，y は現在の出力である．y はネットワークの重み w で決定するため，誤差関数は重み w に関して陰に定義された関数となる．

(1) 勾配降下法（最急降下法）

目的関数の最小値もしくは極小値を探すため計算方法である．誤差関数 $E(w)$ の勾配ベクトルは

$$\Delta w = \frac{\partial E}{\partial \mathbf{w}} \tag{6･5}$$

となり，現在の重み $\mathbf{w}^{(t)}$ それぞれの勾配から，重みを更新する．

$$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \mu \Delta w \tag{6･6}$$

$\mathbf{w}^{(t+1)}$ は更新後の重みである．また，μ を学習係数（Learning Rate）と呼ぶ．μ はチューニングを行う必要がある重要なハイパーパラメータの一つである．勾配降下法では，すべての入力から勾配ベクトルを求める．すなわち，入力サンプル $\mathbf{x_n}(n = 1, \ldots, N)$ がある時誤差 $E(x)$ は

$$E(x) = \sum_{n=1}^{N} E_n(w) \tag{6･7}$$

と表される．ここで，入力サンプルのうち一部をランダムに選択し，その誤差 $E_n(w)$ を使用して重みを更新することを考える．このときの勾配を Δw_n とすると

$$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \mu \Delta w_n \tag{6･8}$$

のように重みを更新する．これを確率的勾配降下法（Stochastic Gradient Descent）と呼ぶ．確率的勾配降下法は勾配降下法と比べ，計算効率が良く早く学習ができる．また，更新反復のたびに目的関数が変わるので，ローカルミニマムにトラップされるリスクを小さくできる．

現在は Optimizer としてモーメンタムや，Adam [6]，Eve [7]などの重み更新の方法が提案されている．

(2) 誤差逆伝搬法

誤差逆伝搬法（Back Propagation）は，前述の勾配降下法を利用して，順伝搬型ネットワークの勾配を効率良く計算する方法である．自動微分の一種である．第 l 層への入力を $u_j^{(l)}$，出力を $z_j^{(l)}$ とすると，入力に関する微分 $\delta_j^{(l)}$ を

$$\delta_j^{(l)} = \frac{\partial E_n}{\partial u_j^{(l)}} = \sum_k \delta_k^{(l+1)} \left(w_{kj}^{(l+1)} f'(u_j^{(l)}) \right) \tag{6･9}$$

と表す．関数 $f(x)$ は活性化関数である．すなわち，第 l 層の $\delta_j^{(l)}$ は上位の第 l+1 層の $\delta_j^{(l+1)}$ から求めることができる．また，求めたい勾配 $\frac{\partial E_n^{(l)}}{\partial w_{ji}^{(l)}}$ は

$$\frac{\partial E_n^{(l)}}{\partial w_{ji}^{(l)}} = \delta_j^{(l)} z_j^{(l-1)} \tag{6･10}$$

から求めることができる．出力層 L の $\delta_j^{(L)}$ は

$$\delta_j^{(L)} = \frac{\partial E_n}{\partial u_j^{(L)}} \tag{6･11}$$

から求められ，出力層から入力層に向けて逆向きの伝搬を行うことで各ユニットの勾配を効

率良く求めることができる．

6-5-4 ディープニューラルネットワーク

(1) 畳み込みニューラルネットワーク

主に画像認識に応用される順伝播型のネットワークである．畳み込みニューラルネットワークは，多層パーセプトロンのように各層間のユニットがすべて結合している全結合と呼ばれる層と，一部のユニット同士が結合した特殊な層から構成される（図 6･3）．畳み込みニューラルネットワークは事前学習が提案される以前から，5 層の深いネットワークの学習に成功していた．Visual Geometry Group が ILSVRC2014 で用いた VGGnet [4]がよく利用される．多層パーセプトロンと同じように誤差逆伝搬法を用いて学習を行う．

(a) 畳み込み層（Convolution）

畳み込み層ではフィルタと呼ばれる小さいサイズの行列を用いて画像の画素値と畳み込み演算を行う．特徴を抽出するためフィルタのパラメータは学習によって更新される．全結合層ではデータの形状は全く考慮されなかったが，畳み込み層では近傍のデータとの形状を加味する．$i \times j$ 画素の入力画像 x，$p \times q$ 画素のフィルタ h として，それぞれの画素値を x_{ij}，h_{pq} とすると，画像の畳み込み $u_{i,j}$ は

$$u_{ij} = \sum_{p} \sum_{q} x_{i+p,j+q} h_{pq} \tag{6･12}$$

となる．畳み込み層の出力 $y_m n$ ではこれにバイアス b が加えられ，活性化関数 $f(x)$ を適用し

$$y_{mn} = f(u_{ij} + b) \tag{6･13}$$

となる．パディングやストライドの設定によって出力画像の画素数は変化する．入力画像からフィルタがはみ出す位置での畳み込みはできないため，画像の端の特徴がとらえにくい．そこで画像にふちに値を追加し，サイズを大きくすることで本来の画像サイズを超えた位置での畳み込みを行うことができる．畳み込みニューラルネットワークではゼロをふちに追加するゼロ・パディングがよく用いられる．また，画像との畳み込みの際に，フィルタをずらすピクセル数をストライドという．通常は特徴を取りこぼしにくい 1 が使用される．

(b) プーリング層（Pooling）

ある範囲の画素値から 1 つの値を求めることで，畳み込み層で得たフィルタの位置感度を低下させる目的で用いられる．画素値のうち最も大きな値を求めるマックスプーリングがよく用いられる．$i \times j$ 画素の入力画像 x のうち，ある領域 $p \times q$ に含まれる画素の集合を P_{ij} とする．P_{ij} の元から 1 つの画素値 u_{ij} を求めるのがプーロングである．P_{ij} の最大値を u_{ij} とするのがマックスプーロングである．

(c) 全結合層（Fully Connectecd）

畳み込み層とプーリング層から得た画像の特徴から，主に分類などを行うためのニューロン同士がすべて結合した層である．

6-5-5 畳み込みニューラルネットワーク（Convolutional Neural Network：CNN）

機械学習で画像の分類といえば畳み込みニューラルネットワークである．階層型の神経回

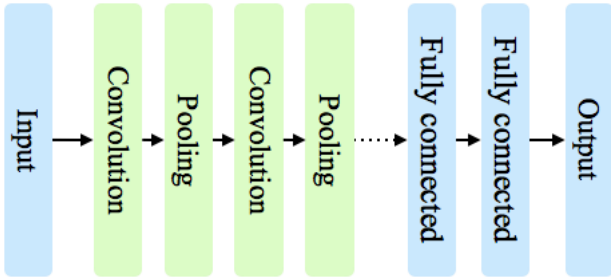

図 6·3　畳み込みニューラルネットワークの一例

路モデルであるネオコグニトロンがルーツであり，ある画像認識コンテストでは人間を超える認識精度に 2015 年に到達した．第三次人工知能ブームの火付け役となった．また，2016～2017 年には Alpha Go にも用いられた．

CNN は主に画像の認識などに用いられるモデルである．複数のピクセルによる構造を巧く情報として扱うことができ，適切に扱えば画像の識別精度は非常に高くなることが知られている．また，比較的容易に学習が進む性質があり，1980 年代に 5 層の畳み込み層を持つ CNN の学習に成功していた．画像以外の構造を持つデータ，例えば自然言語や音などの時系列データに対しても用いられる．

(1) 順伝播型ニューラルネットワークとの違い

層間のニューロンが互いにすべて結合した，全結合層いわゆる順伝播型ニューラルネットワークは，ベクトル（1 次元配列）を入力する．また，入力されるベクトルの成分は順序や，構造を考慮しない．そのため，順伝播型ニューラルネットワークで画像を扱うには，行列をベクトルに変形しなければならず，またピクセルの集合として画像を扱うのではなく，各ピクセルの値を扱うのみである．

一方 CNN は，画像を行列としてそのまま入力できる．更に，畳み込み層などを用いて，データの構造を考慮することができる．順伝播型ニューラルネットワークでも画像を扱うことはできるが，CNN はより自然に画像を画像として扱うことができる．畳み込み層などを重ねて特徴を縮小し，最終的に全結合層を用いて判断をする．すなわち，複数の畳み込み層などと全結合層から構成されるニューラルネットワークである．

(2) CNN に用いられる層

一般に，CNN の画像の構造を考慮した特徴の抽出は，複数の計算処理から構成される．畳み込み，活性化，プーリングである．畳み込みでは学習によって獲得した重み（フィルタ）との一致度となるスカラーを得る．フィルタは，3 × 3 程度の小さな画像であり，入力画像の端から端まで，数ピクセルずつ位置をずらしながら画像上のあらゆる場所でフィルタとの一致度を計算する．ずらすピクセルの数をストライドと呼ぶ．これによって複数の畳み込みから得た一致度を再配置した特徴マップを得る．その際，画像の端の構造を捉えるために，画像の端にゼロを追加して画像サイズを大きくするゼロパディングがよく用いられる．次に，

活性化は活性化関数で特徴マップの写像を得る計算である．現代では活性化関数に Rectified Linear Units（ReLU）が用いられることがほとんどである．

ReLU は，非常に単純な関数ながら，学習を困難にする勾配消失や発散を防ぐといった重要な機能を持つ．順伝播型ニューラルネットワークにおいても，活性化関数には ReLU や派生系の leakly ReLU を用いるのが一般的である．ディープなニューラルネットワークにおいて事前学習なく学習が可能となったのも，ReLU（と上手な初期値）が大きく寄与している．

最後に，プーリングは，重要な情報を残しつつも画像を縮小する手法で，通常は畳み込みや活性化の後，特徴マップに対して行われる．マックスプーリングは画像のある範囲で区切り，その範囲ごとの最大値のみをとり，ダウンサンプリングする．この処理では，フィルタと一致した厳密な位置ではなく，ある範囲内のどこかで一致しているとみなす．すなわち，画像の小さな位置の変化に対しても対応することができる．

これらの層は，組み合わせることで画像の特徴を抽出することができ，CNN が画像の構造を扱ううえで重要な役割を担っている．更に，これらの層を繰り返すことで，浅い層ではエッジなどの単純な特徴を抽出し，深い層では顔や動物などより複雑な特徴を抽出できる．これらのフィルタは自動的に獲得できる．

(3) 全結合層

全結合層は，畳み込み，活性化，プーリングを複数行って，特徴を抽出した後，その特徴から分類や回帰，すなわち判断を行う層である．特徴マップは行列からベクトルに変換され，構造は関係なく全結合層に入力される．画像の識別を行うのであれば，出力は識別結果とその確率となる．通常は全結合層も多層の構造を持たせる．

本節は，6-1 節で述べた分類のうち，「ツールの開発」にあたる．

■参考文献

1) W.S. McCulloch and W. Pitts：“ A logical calculus of the ideas immanent in nervous activity, ” The Bulletin of Mathematical Biophysics, 5(4), pp.115-133, 1943.
2) F. Rosenblatt：“ The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, ” Psychological Review, 65(6), pp.386-408, 1958.
3) D.E. Rumelhart, G.E. Hinton, and R.J. Williams：“ Learning representations by back-propagating errors, ” Nature, 323, pp.533-536, 1986.
4) K. Simonyan and A. Zisserman：“ Very Deep Convolutional Networks for Large-Scale Image Recognition, ” arXiv, 1409.1556, 2014.
5) G. E. Hinton, S. Osindero, and Y. Teh：“ A fast learning algorithm for deep belief nets, ” Neural Computation, 18, pp.1527-1544, 2006.
6) D. P. Kingma and J.L. Ba：“ Adam: A Method for Stochastic Optimization, ” arXiv, 1412.6980, 2014.
7) J. Koushik and H. Hayashi：“ Improving Stochastic Gradient Descent with Feedback, ”arXiv, 1611.01505, 2016.

■S2 群 - 6 編 - 6 章

6-6 システムバイオロジーと深層学習

（執筆者：稲岡秀檢）[2018 年 2 月 受領]

機械学習は，機能的関係をあらかじめ定義することなく，機能的関係をデータのみから学習するためのアプローチである．システムバイオロジーにおいて，機能的関係のメカニズムについては，未知であるか，あるいは不十分な定義しかないことが多い．機械学習はこのような条件で，機能的関係の予測モデルを導出する能力があるという点でシステムバイオロジーにおいて期待されている．遺伝子発現レベルの最も正確な予測は，広範なエピジェネティックな特徴量を用いたスパース（疎）な線形モデル[1)]，またはランダムフォレスト[2)]を使用する方法によって達成される．選択された特徴量がどのように転写物レベルを決定するかについてはまだ研究が必要である．

ゲノミクス，プロテオミクス，メタボロミクス，化合物に対する感受性の予測はすべて，機械学習のアプローチを重要な要素としている[3, 4, 5, 6)]．

標準的な機械学習の流れは，以下の 4 つのステップで記述することができる．

1. データクリーニング
2. 前処理
3. 特徴量の抽出
4. モデルフィッティングとその評価

教師有り機械学習モデルでは，すべての共変量と特徴量を入力データとして記述し（通常はベクトル量），出力値（通常はスカラー値）をラベル付けするのが一般的である．データが記録されたトレーニング対のリスト（ベクトル x_1, ラベル y_1），（ベクトル x_2, ラベル y_2），…，（ベクトル x_n，ラベル y_n）から関数 $y = f(x)$ を学習することを目的とする．

システムバイオロジーにおける代表的な応用例としては，がん細胞が特定の薬物に曝露された場合の生存率を予測することが挙げられる．このとき入力の特徴量（x）は，体細胞の塩基配列の変異体，薬剤の化学的構成，薬剤の濃度であり，ラベルは生存率となる．トレーニングされたモデルは，新しいラベル無しデータサンプルが与えられると，学習された関数 f からその生存率を予測する．このとき関数はブラックボックスのままなので，なぜ特定の突然変異の組合せが細胞増殖に影響を与えるかについては簡単には解釈できない．この教師有り機械学習モデルとしては「回帰」と「分類」が挙げられる．

教師無し機械学習モデルは，出力ラベルを必要とせずに，データ自体からパターンを発見することを目的としている．教師無し機械学習モデルとしては「クラスタ化」，「主成分分析」，「外れ値検出」などが挙げられる．

機械学習モデルにおいて，生データから作られる特徴量を含む入力データは，モデルと環境の関係を表すものとなる．有益な特徴量を引き出すことが機械学習の目的である．その過程は多大な労力を必要とし，またその問題の領域に関する知識も必要となる．これは高次元のデータを利用するときに大きな制限になる．コンピュータによる特徴量の選択を用いても，

膨大な数の可能な入力組合せの有用性を評価することが困難なためである．

最近の機械学習の主な進歩は，データを多層のニューラルネットワークを用いて学習することで，この特徴量抽出のステップを自動化することである[7, 8]．このアプローチは深層学習あるいはディープラーニングと呼ばれる．深層学習は，最下位（入力）層で生データを取り出し，多層で構成されるネットワークにおいて，直前の層からの出力をデータ駆動で連続的に組み合わせることで，抽象的な特徴量表現に変換する複雑な機能をカプセル化している．深層学習は機械学習における活発な研究分野であり，画像認識，音声認識[9, 10]，自然言語理解[11]で広く使われている．最近ではシステムバイオロジーの分野[12]でも利用されている．

ハイスループットな生物学における深層学習の可能性は大きい．大きくて高次元のデータセット（例えば，DNA，RNA の測定，フローサイトメトリ，または自動顕微鏡検査）を使用し，その内部構造を解析するための多層を有する複雑なネットワークを学習する．学習されたネットワークは，新たな機能を発見したり，従来のモデルと比較してパフォーマンスが向上することが期待される．また，結果の解釈能力を高め，生物学的データの構造についての更なる理解を提供することが期待されている．

以下に，転写制御に関するゲノミクスと生物学的画像分析について，深層学習の応用例を紹介する．

6-6-1 転写制御

転写制御ゲノムに関する従来のアプローチとしては，配列の変異を分子の形質変化に関連付けるものがある．これらのなかで，遺伝的に多様な個体間の変異を利用して，定量的形質遺伝子座（quantitative trait loci：QTL）をマッピングするものがある．この手法は，遺伝子発現レベル[13]，DNA メチル化[14]，ヒストン修飾[15]，及びプロテオーム変異[16]に影響を及ぼす転写制御変異体を同定するために用いられている．しかし，このマッピング手法は，本質的に学習用のデータに存在する変異に限定される．したがって，稀な変異の影響を解析するためには，非常に大きなサンプルサイズのデータセットが必要となる．

ゲノム内の領域間の変異を利用してモデルを訓練する方法もある．目的の形質を中心としたウインドウに遺伝子配列を分割することで，単一の個体のデータを利用する場合でも．ほとんどの分子形質について数万の訓練用データを作成することができる．大規模なデータセットがあっても，広範な塩基配列の状況や，遠位の転写制御要素との相互作用に対する分子形質の依存性を解析するなどのように，DNA 配列からの分子形質を予測することは，依然として挑戦に値する課題である．

転写制御に関する深層ニューラルネットワークの価値は 2 つある．

第一に，古典的な機械学習方法は，塩基配列をそのまま入力データとすることができない．そのため，分子生物学に関する事前の知識に基づいて塩基配列から抽出することができる事前定義特徴量（例えば，一塩基変異体（SNV）の存在または欠損，k-mer 頻度，モチーフの出現，配列の保存，既知の調節変異体）を必要とする．深層ニューラルネットワークは，塩基配列データから直接学習することにより，特徴量の手動抽出を回避するのに利用できる．

第二に，深層ニューラルネットは塩基配列をそのまま取り扱うことができるため，配列と相互作用の影響の非線形依存性を捕捉し，複数のゲノムスケールでより広い塩基配列にまたがることができる．深層ニューラルネットワークの実用性は，スプライシング活性[17]，DNA

結合，RNA 結合タンパク質[18]，エピジェネティックマーカの特異性の予測とその DNA 配列変化への影響[19]の研究結果から示されている．

（1）転写制御ゲノミクスにおけるニューラルネットワークの初期応用

転写制御ゲノミクスにおけるニューラルネットワークの最初の成功した応用は，入力の特徴量を変更することなく，古典的な機械学習アプローチを深層ニューラルネットワークに置き換えたものである．例えば，個々のエキソンのスプライシング活性を予測するために，全接続フィードフォワードニューラルネットワークが考えられた[20]．このモデルは，候補エキソン及び隣接するイントロンから抽出された 1000 を超えるあらかじめ定義された特徴量を用いて訓練された．トレーニングサンプルが 10700 個と比較的少数であったにも関わらず，モデルの複雑さと組み合わされることで，この方法はより簡単なアプローチに比べてスプライシング活性の予測精度が大幅に高くなった．また，スプライシングの誤調節に関与する稀な変異を同定することができた．

（2）畳み込みデザイン

畳み込みニューラルネットワーク（Convolutional Neural Networks：CNN）を用いた最近の研究では，特徴量を定義する必要なしに，DNA 配列から直接学習することができた．CNN アーキテクチャは，入力空間の小さな領域のみに畳み込み演算を適用し，領域間でパラメータを共有することによって，全接続ネットワークに比べてモデルパラメータの数を大幅に削減することを可能にする．このアプローチから得られる主な利点は，より大きな配列ウィンドウでモデルを直接学習できることである．

DNA 及び RNA 結合タンパク質の特異性を予測するために畳み込みネットワーク構造が検討された．DeepBind モデル[18]は既存の方法を凌駕し，既知及び新規の配列モチーフを検出することができ，配列変化の影響を定量化し，機能的な SNV を同定することができた．

（3）突然変異の影響の *in sillico* 予測

生の DNA 配列で学習された深層ニューラルネットの重要な応用は，*in sillico* における突然変異の影響を予測することである．そのような配列変化の影響のモデルベースの評価は，QTL マッピングに基づく方法を補完し，特に希少な SNV の制御効果を明らかにしたり，可能性のある遺伝子を細かくマッピングするのに役立つ．そのような予測された制御効果を視覚化するための直観的なアプローチは突然変異マップであり，与えられた入力配列に対するすべての可能な突然変異の影響をマトリックス図で表す[18]．このマップにより，野生型及び突然変異体配列についての予測結合スコアを有する追加のニューラルネットワークを訓練することによって，有害な SNV を更に確実に同定することができた．

（4）複数の形質の共同予測と更なる拡張

畳み込みアーキテクチャは拡張され，転写制御ゲノミクスの一連の研究に適用された．例えば，DNA 配列からのクロマチンマーカの予測にこのアーキテクチャが検討された[19]．第 2 の革新は，複数の出力変数を有するニューラルネットワークアーキテクチャ（いわゆるマルチタスクニューラルネットワーク）を使用して，複数のクロマチン状態を並行して予測することであった[21]．マルチタスクアーキテクチャは出力間で共有された特徴を学習することができるため，汎化性能を向上させ，各形質の独立したモデルを学習する場合と比較してモデル学習の計算コストを大幅に削減する．

同様の観点から，複数の細胞型にわたる DNase I 過敏症を予測し，SNV のクロマチン接近

可能性への影響を定量化するためのオープンソースの深層学習フレームワーク Basset が開発された[12]．このモデルもまた，従来の方法と比較して予測性能を改善し，DNaseI 過敏症に関連する既知及び新規の配列モチーフの両方を検索することができた．

マルチタスクアーキテクチャは，単一細胞バイサルファイト配列決定研究における DNA メチル化状態を予測すると考えられている．このアプローチは，畳み込みアーキテクチャを組み合わせて，有益な DNA 配列モチーフを隣接する CpG 部位から誘導される追加の特徴で検出し，それによってメチル化の状況を説明する[22]．異なるクロマチン標識についてのより正確な有病率推定を得るために，染色体免疫沈降とそれに続く配列決定データのノイズを除去するために CNN が応用された．

現在，CNN は，固定サイズの DNA 配列ウィンドウから特徴を抽出するために最も広く使用されているアーキテクチャの一つである．しかしながら，代替的なアーキテクチャも考えられる．例えば，リカレントニューラルネットワーク（RNN）は，逐次データをモデル化するのに適しており[23]，自然言語及び音声[11]，タンパク質配列[24]，臨床医学データ[23]，及び限定された範囲の DNA 配列のモデリングに適用されている[25]．RNN は，可変長のモデル化配列を可能にし，配列内及び複数の出力にわたって長距離相互作用を検出することができるため，転写制御ゲノミクスのアプリケーションにとって魅力的である．しかし，現時点では，RNN は CNN よりも学習が困難であり，より良い結果を得るためには設定をより良く理解するための更なる作業が必要である．

教師有り学習方法と相補的である，教師無しの深層学習アーキテクチャは，古典的な主成分分析または因子分析と同様に，高次元の非ラベルデータから低次元の特徴表現を学習するが，非線形なモデルを使用する．このようなアプローチの例は，積み重ねオートエンコーダ[26]，制限付きボルツマンマシン（RBM）と Deep Belief Network（DBN）[27]である．学習された特徴量は，データを視覚化するために，または教師有り学習のための入力として使用することができる．例えば，遺伝子発現プロファイルを用いてがん症例を分類するため，あるいはタンパク質の主鎖を予測するために，スパース（疎）オートエンコーダが適用されている．RBM は，タンパク質二次構造，不規則なタンパク質領域，アミノ酸接触の教師有りモデルをその後に訓練するための，深層ネットワークの教師無し事前学習にも使用することができる．

一般に，複雑なモデルを事前に学習するためのラベルがない大量のデータが利用可能な場合，教師無しモデルは，強力なアプローチとなる．一度学習されると，このモデルは，より少ない数のラベル化されたデータが利用可能であるとき，分類タスクのパフォーマンスを向上させるのに役立つ．

6-6-2 画像解析

深層ニューラルネットワークの最も重要な成功例は画像分析である．数百万の写真で訓練された深層アーキテクチャは，人間が行うよりも写真の中の物体をうまく検出することができる[28]．画像分類，物体検出，画像検索，及び意味論的セグメンテーションにおけるすべての最新モデルは，ニューラルネットワークを利用する．

畳み込みニューラルネットワーク（CNN）は，画像解析のための最も一般的なネットワークアーキテクチャである．簡単に言うと，CNN はパターンマッチング（畳み込み）とプーリング（集約）操作を実行する．ピクセルレベルで，畳み込み演算は，与えられたパターンで画

像をスキャンし，すべての位置についてマッチの強さを計算する．プーリングは，領域内のパターンの存在を判定する．例えば，より小さいパッチでの最大パターン一致（max-pooling）を計算し，領域情報を単一の数値に集約する．畳み込み演算とプーリング演算を連続して適用することは，画像解析に使用されるほとんどのネットワークアーキテクチャの中核となる．

（1）計算生物学における最初のアプリケーション-ピクセルレベルの分類

生物学的画像のための深層ネットワークの初期の応用は，ネットワーク出力を構築する付加的なモデルを用いて，ピクセルレベルのタスクに焦点を合わせた．例えば，線虫の胚画像を処理して異常発生を予測する研究で畳み込みニューラルネットワーク（CNN）が利用された[29]．3 つの畳み込み及びプーリング層，全連結出力層を用いて 40 × 40 ピクセルのパッチで中央のピクセルを細胞壁，細胞質，核膜，核または外部培地に分類するように CNN を学習させた．その後，モデル予測は，更なる分析のためにエネルギーベースのモデルに入力された．雑音の多い神経回路画像を復元するなどの生データ解析において，CNN はマルコフランダムフィールドや条件付きランダムフィールドなどの標準的な手法を上回っている．

レイヤを追加することで，ピクセルノイズの除去からより抽象的な画像特徴量のモデリングに移行することができる．5 つの畳み込み及びプーリング層，続いて 2 つの全連結出力層を使用して，乳房組織像における有糸分裂が発見された．

このモデルは，International Conference of Pattern Recognition 2012 での有糸分裂検出の課題獲得において，競合者を大幅に上回った．同様の手法を用いて，各ピクセルを膜または非膜として分類する電子顕微鏡画像で神経構造をセグメント化した．

（2）全細胞，細胞集団及び組織の分析

多くの場合，ピクセルレベルの予測は必要ない．例えば，結腸組織病理画像をがん性及び非がん性に直接分類し，深層ネットワークによる特徴量の教師有り学習が，手作業で特徴量を作成するよりも優れていることが分かった[30]．蛍光タンパク質を有する個々の酵母細胞のあらかじめセグメント化された画像パッチを，異なる細胞下局在パターンに分類するために CNN が使用された．ここでも，深層ネットワークは従来の特徴量に基づく手法を上回った．

（3）学習されたモデルの再利用

CNN の学習には大きなデータセットが必要であるが，生物学的データの取得は高価になることが多い．しかし，数百万の画像が利用できない場合に，深層ニューラルネットワークを使用できないということではない．元の画像の種類に関わらず，ネットワークのレベルが低い（層が少ない）ほど，一般的な画像に発生する同様の信号（エッジ，ブロブ）を捕捉する傾向がある．したがって，CNN は学習に役立つ類似の領域からの画像を再利用することができる．あるいは，ほかのデータによって事前に訓練されたデータを利用して，より関心のあるタスクのモデルを少ない画像で微調整することができる．実際に，オブジェクトを分類するのための何百万もの画像から学んだ特徴量が，ラベル付けされていない，数百の新しい画像を検索，検出，分類することに成功した[31, 32]．このようなアプローチの有効性は，訓練データと新しい領域との間の類似性に依存する．

モデルパラメータを伝達する概念もまた，生物学的画像分析において成功している．例えば，自然画像から学んだ特徴を生物学的データに移すことができ，*in situ* ハイブリダイゼーション画像からのショウジョウバエの発生段階の予測を改善することが示された[33]．このモデルは，様々なスケールで豊富な特徴量を抽出するために，100 万以上の様々な画像を持つ

オープンコーパスである ImageNet[34]のデータを最初に事前にトレーニングした．更に，顕微鏡画像の自動細胞計数のために CNN を訓練するために合成画像が使用された．

(4) 深層学習フレームワーク

深層学習フレームワークは，既存のモジュールから高レベルでニューラルネットワークを容易に構築するために開発された．最も人気のあるものは Caffe，Theano，Torch7，TensorFlow であり，これらはモジュール性，使いやすさ，モデルの定義と学習方法が異なる．

Caffe [35]は，Berkeley Vision and Learning Center によって開発され，C++で書かれている．ネットワークアーキテクチャは構成ファイルで指定され，モデルを訓練し，コードを書くことなくコマンドラインで使用することができる．更に，Python と MATLAB のインタフェースも利用できる．Caffe は，CNN のための最も効率的な実装の一つを提供し，画像認識のための複数の事前訓練されたモデルを提供し，コンピュータビジョンタスクに適している．欠点として，カスタムモデルは C++で実装する必要があるが，これは難しい場合がある．更に，Caffe は再帰的なアーキテクチャに最適化されていない．

Theano [36]は，モントリオール大学によって開発・維持され，Python と C++で書かれている．モデル定義は，プログラミングの代わりに宣言による．つまり，ユーザは成されるべきことを指定するが，どの順序で行うかは指定しない．ニューラルネットワークは計算グラフとして宣言され，ネイティブコードにコンパイルされて実行される．この設計により，Theano は計算ステップを最適化し，自動的に勾配を導出することができる．したがって，Theano はカスタムモデルの構築に適しており，特に RNN の効率的な実装を提供する．Theano の大きな欠点は，より大きなモデルを構築するときにしばしば長いコンパイル時間が必要となることである．

Torch 7[37]は最初，ニューヨーク大学で開発され，スクリプト言語 LuaJIT に基づいている．ネットワークは既存のモジュールを積み重ねることで簡単に構築でき，コンパイルされないため，Theano よりも素早いプロトタイプ作成に適している．Torch7 は，効率的な CNN の実装と，事前にトレーニングされた一連のモデルへのアクセスを提供する．可能性のある欠点は，ユーザが LuaJIT スクリプト言語に精通している必要があることである．また，LuaJIT はカスタム再帰ネットワークを構築するのにはあまり適していない．

TensorFlow [38]は，Google が開発した最新の深層学習フレームワークである．このソフトウェアは C++で書かれており，Python へのインタフェースを提供している．Theano と同様に，ニューラルネットワークはコンピュータグラフとして宣言され，コンパイル時に最適化される．しかし，短いコンパイル時間から TensorFlow はプロトタイピングとして適したものになっている．TensorFlow の強みは，CPU，GPU を含む様々なデバイス間の並列化をネイティブにサポートすることであり，クラスタ上で複数の計算ノードを使用することができる．付属のツール TensorBoard を使用すると，Web ブラウザでネットワークを視覚化したり，学習曲線やパラメータの更新などのトレーニング進捗状況を監視することができる．現在，TensorFlow は RNN の最も効率的な実装を提供している．このソフトウェアは最近開発されたものであるため，現在，わずかな事前訓練モデルしか利用できない．

(5) データ準備

トレーニングデータは，あらゆる機械学習にとって重要である．情報のある特徴量が多いデータほどパフォーマンスが向上するため，データの収集，ラベル付け，クリーニング，正

規化に時間を費やす必要がある．

(6) 必要なデータセットのサイズ

深層学習の成功例の大半は，複雑なモデルに適合するのに十分なラベル付きのトレーニングサンプルが利用可能な教師有り学習の設定にある．遺伝子型からの分子形質を予測するなど，転写制御ゲノミクスの中心的な問題は，訓練事例の数に制限があることである．関心のある形質（例えば，スプライス部位，転写因子結合部位またはエピジェネティックマーカ）を中心とする配列ウィンドウを考慮する戦略は，現在広く使用されており，単一の個体からの入出力ペアの数を増加させるのに役立つ．

画像解析では，データは豊富にあることが多いが，手作業でキュレーションされ，ラベル付けされたトレーニング例は一般的に入手が困難である．そのような場合，トレーニングセットは，既存の画像をスケーリング，回転することによって補強することができる．もう一つの戦略は，画像認識のために大きなデータセットで事前にトレーニングされたネットワークを再利用することである． AlexNet [39]，VGG [40]，GoogleNet [41]または ResNet [28]を使用して，関心のあるデータセット，例えば，特定のセグメンテーションタスクの顕微鏡画像によってパラメータを微調整することができる．このようなアプローチでは，異なるデータセットがエッジやカーブなどの重要な特性や特徴を共有し，それらの間で転送できるという事実を利用している．

(7) 生データの正規化

データ正規化のための適切な選択は，トレーニングを促進し，適切な局所最小化を識別するのに役立つ．

DNA ヌクレオチドのようなカテゴリの特徴は，まず数値的にコード化する必要がある．例えば，DNA ヌクレオチドは A =（1000），G =（0100），C =（0010）及び T =（0001）として一般にコードされる．DNA 配列は，コード化されたヌクレオチドを連結することでバイナリストリングとして表すことができる．また，DNA 配列は各ヌクレオチドをニューラルネットワークの独立した入力特徴量として処理される．CNN では，各エンコードされた塩基の 4 ビットは，一般にヌクレオチドの実体を保存するために画像のカラーチャネルと同様に考えられる．

数値的な特徴量は，通常，その平均値を減算することによってゼロセンタリングされる．画像ピクセルは，通常，個別にゼロセンタリングされるのではなく，カラーチャンネルごとに平均ピクセル強度を減算することによって共同でゼロセンタリングされる．その他の共通の正規化手順としては，特徴量を単位分散で標準化する方法である．また，極端な値のために特徴量の分布が歪んでいる場合は，対数変換や同様の処理手順が適切な場合がある．

検証データとテストデータは，トレーニングデータと一貫して正規化する必要がある．例えば，検証データの特徴量は，検証データ自体の平均値ではなく，訓練データ上で計算された平均値を引くことでゼロセンタリングする必要がある．

本節は，6-1 節で述べた分類のうち，「分析方法の検討，知識発見，ツールの開発」にあたる．

■参考文献

1) C. Cheng, K.K. Yan, K.Y. Yip, J. Rozowsky, R. Alexander, C. Shou, and M. Gerstein：“ A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets, ” Genome Biol., 12(2), R15, 2011.
2) J. Li, T. Ching, S. Huang, and L.X. Garmire：“ Using epigenomics data to predict gene expression in lung cancer, ” BMC Bioinformatics, 16(Suppl.5), S10, 2015.
3) N.W. Libbrecht and W.S. Noble：“ Machine learning applications in genetics and genomics, ” Nat Rev Genet., 16(6), pp.321-332, 2015.
4) A.L. Swan, A. Mobasheri, D. Allaway, S. Liddell, and J. Bacardit：“ Application of Machine Learning to Proteomics Data: Classification and Biomarker Identification in Postgenomics Biology, ” OMICS, 17(12), pp.595-610, 2013.
5) D.B. Kell：“ Metabolomics, machine learning and modelling: towards an understanding of the language of cells, ” Biochem Soc Trans., 33(Pt 3), pp.520-524, 2005.
6) F. Eduati, L.M. Mangravite, T. Wang, H. Tang, J.C Bare, R. Huang, T. Norman, M. Kellen, M.P. Menden, J. Yang, X. Zhan, R. Zhong, G. Xiao, M. Xia, N. Abdo, O. Kosyk, NIEHS-NCATS-UNC DREAM Toxicogenetics Collaboration, S. Friend, A. Dearry, A. Simeonov, R.R. Tice, I. Rusyn, F.A. Wright, G. Stolovitzky, Y. Xie, and J. Saez-Rodriguez：“ Prediction of human population responses to toxic compounds by a collaborative competition, ” Nat Biotechnol., 33(9), pp.933-940, 2015.
7) Y. LeCun, Y. Bengio, and G. Hinton：“ Deep learning, ” Nature, 521(7553), pp.436-444, 2015.
8) J. Schmidhuber：“ Deep learning in neural networks: an overview, ” Neural Netw, 61, arXiv:1404.7828, pp.85-117, 2015.
9) M.D. Zeiler and R. Fergus：“ Visualizing and understanding convolutional networks, ” European Conference on Computer Vision, arXiv:1311.2901, pp.818-833, 2014.
10) D. Li and T. Roberto：“ Deep dynamic models for learning hidden representations of speech features, ” Speech and Audio Processing for Coding, Enhancement and Recognition, pp.153-195, 2015.
11) C. Xiong, S. Merity, and R. Socher：“ Dynamic memory networks for visual and textual question answering, ” arXiv:1603.01417, 2016.
12) D.R. Kelley, J. Snoek, and J. Rinn：“ Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks, ” Genome Res., 26(7), pp.990-999, 2016.
13) S.B. Montgomery, M. Sammeth, M. Gutierrez-Arcelus, R.P. Lach, C. Ingle, J. Nisbett, R. Guigo, and E.T. Dermitzakis：“ Transcriptome genetics using second generation sequencing in a Caucasian population, ” Nature, 464, pp.773-777, 2010.
14) J.T. Bell, A.A. Pai, J.K. Pickrell, D.J. Gaffney, R. Pique-Regi, J.F. Degner, Y. Gilad, and J.K. Pritchard：“ DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines, ” Genome Biol., 12:R10, 2011.
15) F. Grubert, J.B. Zaugg, M. Kasowski, O. Ursu, D.V. Spacek, A.R. Martin, P. Greenside, R. Srivas, D.H. Phanstiel, A. Pekowska, N. Heidari, G. Euskirchen, W. Huber, J.K. Pritchard, C.D. Bustamante, L.M. Steinmetz, A. Kundaje, and M. Snyder：“ Genetic control of chromatin states in humans involves local and distal chromosomal interactions, ” Cell, 162, pp.1051-1065, 2015.
16) A. Battle, Z. Khan, S.H. Wang, A. Mitrano, M.J. Ford, J.K. Pritchard, and Y. Gilad：“ Genomic variation. Impact of regulatory variation from RNA to protein, ” Science, 347, pp.664-667, 2015.
17) M.K.K. Leung, H.Y. Xiong, L.J. Lee, and B.J. Frey：“ Deep learning of the tissue-regulated splicing code, ” Bioinformatics, 30(12), pp.i121-i129, 2014.
18) B. Alipanahi, A. Delong, M.T. Weirauch, and B.J. Frey：“ Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning, ” Nat Biotechnol., 33(8), pp.831-838, 2015.
19) J. Zhou and O.G. Troyanskaya：“ Predicting effects of noncoding variants with deep learning-based sequence model, ” Nat Methods., 12(10), pp.931-934, 2015.

20) H.Y. Xiong, B. Alipanahi, L.J. Lee, H. Bretschneider, D. Merico, R.K.C. Yuen, Y. Hua, S. Gueroussov, H.S. Najafabadi, T.R. Hughes, Q. Morris Y. Barash, A.R. Krainer, N. Jojic, S.W. Scherer, B.J. Blencowe, and B.J. Frey：" The human splicing code reveals new insights into the genetic determinants of disease, " Science, 347(6218), 1254806, 2015.
21) G.E. Dahl, N. Jaitly, and R. Salakhutdinov：" Multi-task neural networks for QSAR predictions, " arXiv:1406.1231, 2014.
22) C. Angermueller, H.J. Lee, W. Reik, and O. Stegle：" DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning, " Genome Biol., 18:67, 2017.
23) Z.C. Lipton, J. Berkowitz, and C. Elkan：" A critical review of recurrent neural networks for sequence learning, " arXiv:1506.00019, 2015.
24) M. Agathocleous, G. Christodoulou, V. Promponas, C. Christodoulou, V. Vassiliades, and A. Antoniou：" Protein secondary structure prediction with bidirectional recurrent neural nets: can weight updating for each residue enhance performance?, " Artificial Intelligence Applications and Innovations, 339, pp.128-137, 2010.
25) B. Lee, T. Lee, B. Na, and S. Yoon：" DNA-level splice junction prediction using deep recurrent neural networks, " arXiv:1512.05135, 2015.
26) P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.A. Manzagol：" Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion, " Journal of Machine Learning Research, 11, pp.3371-3408, 2010.
27) G.E. Hinton, S. Osindero, and Y.W. Teh：" A fast learning algorithm for deep belief nets, " Neural Comput., 18(7), pp.1527-1554, 2006.
28) K. He, X. Zhang, S. Ren, and J. Sun：" Deep residual learning for image recognition, "arXiv:1512.03385, 2015.
29) F. Ning, D. Delhomme, Y. LeCun, F. Piano, L. Bottou, and P.E. Barbano：" Toward automatic phenotyping of developing embryos from videos, " IEEE Trans Image Process., 14(9), pp.1360-1371, 2005.
30) Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, and E.I.C. Chang：" Deep learning of feature representation with multiple instance learning for medical image analysis, " IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.1626-1630, 2014.
31) J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N.Zhang, E. Tzeng, and T. Darrell：" Decaf: a deep convolutional activation feature for generic visual recognition, " arXiv:1310.1531, 2013.
32) A.S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson：" CNN features off-theshelf: an astounding baseline for recognition, " Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp.806-813, 2014.
33) W. Zhang, R. Li, T. Zeng, Q. Sun, S. Kumar, J. Ye, and S. Ji：" Deep model based transfer and multi-task learning for biological image analysis, " Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.1475-1484, 2015.
34) O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, and L. Fei-Fei：" Imagenet large scale visual recognition challenge, "International Journal of Computer Vision, 115(3), pp.211-252, 2015.
35) Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell：" Caffe: convolutional architecture for fast feature embedding, " Proceedings of the 22nd ACM international conference on Multimedia, pp.675-678, 2014.
36) F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. Goodfellow, A. Bergeron, N. Bouchard, D. Warde-Farley, and Y. Bengio：" Theano: new features and speed improvements, " arXiv:1211.5590, 2012.
37) R. Collobert, K. Kavukcuoglu, and C. Farabet：" Torch7: a matlab-like environment for machine learning, " BigLearn NIPS Workshop, 2011.

38) M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng：“ TensorFlow: large-scale machine learning on heterogeneous distributed systems, ” arXiv:1603.04467, 2016.

39) A. Krizhevsky, I. Sutskever, and G.E. Hinton：“ Imagenet classification with deep convolutional neural networks, ”Proceedings of the 25th International Conference on Neural Information Processing Systems, 1, pp.1097-1105, 2012.

40) K. Simonyan and A. Zisserman：“ Very deep convolutional networks for largescale image recognition, ” arXiv:1409.1556, 2014.

41) C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna：“ Rethinking the inception architecture for computer vision, ” arXiv:1512.00567, 2015.