■S2 群（ナノ・量子・バイオ）- 6 編（バイオインフォマティックス）

5 章 ゲノムデータの解析

（執筆者：稲岡秀檢）［2018 年 4 月 受領］

■概要■

測定技術の向上により大量のゲノムデータが利用可能になっている．本章では大規模なデータ解析で行われている各種解析手法について解説・紹介を行う．

【本章の構成】

本章では，ゲノムデータの解析手法について説明する．5-1 節では，配列解析の手法について説明する．5-2 節では，発現解析の手法について説明する．5-3 節では，タンパク質の結合予測に関する手法について紹介する．5-4 節では，深層学習を用いた解析手法を紹介する．

■S2 群 - 6 編 - 5 章

5-1 配列解析

（執筆者：稲岡秀檢）［2018 年 2 月 受領］

バイオデータベースには塩基配列情報や遺伝子発現情報，DNA メチル化情報などがある．これらのデータは特定のフォーマットで提供されることが多い．例えば塩基配列情報は FASTA 形式[1]や，GenBank 形式[2]などが使用されている．これらのデータを取り扱うときは，既存のファイル形式の読み書きや他形式への変換が必要となる．また，塩基配列の相同性検索などを行うためには，塩基配列の部分配列の置換・挿入といった作業が要求される．こういったデータ操作を効率良く行うために，バイオインフォマティクスに特化した様々なライブラリ群が提供されている．

これらのライブラリでは塩基配列の相同性解析のために広く使用されている BLAST プログラム[3]をプログラム内部から呼び出したり，多岐にわたる検索結果項目の効率的な取扱いなどの高度な操作も提供されている．塩基配列やアミノ酸配列情報など，バイオデータベースには数値データ以外にテキスト形式のデータも多く存在する．そのためテキストデータの取扱いが簡便で，プログラム開発も容易なスクリプト言語である Perl や Ruby などが広く使用されている．上記のライブラリーも Perl や Ruby のために開発されている（BioPerl [4]，BioRuby [5]）．

また，バイオインフォマティクスでは，結果のグラフィカルな表示なども重要な要素となってくる．更に，解析された結果を解釈するために統計的なデータ処理も多用される．こういった要求から統計計算のための言語環境であり，グラフィック表示のためのライブラリなどが充実している統合開発環境である R 言語[6]が解析に利用されることも多い．R 言語では，塩基配列データやマイクロアレイ遺伝子発現データを効率良く取り扱うために Bioconductor [7] というパッケージが提供されている．

■参考文献

1) http://blast.ncbi.nlm.nih.gov/blastcgihelp.shtml
2) http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.htm
3) https://blast.ncbi.nlm.nih.gov/Blast.cgi
4) http://www.bioperl.org/
5) http://bioruby.org/
6) http://www.r-project.org/
7) http://www.bioconductor.org/

■S2 群 - 6 編 - 5 章

5-2 発現解析

（執筆者：稲岡秀檢）[2018 年 2 月 受領]

バイオインフォマティクスなどで広く使われているデータ解析技術として，クラスタリング（Clustering）がある．クラスタリングではデータの類似性に従ってデータのグループ分けを行う．例えば，ある疾病が原因となる遺伝子の発現量に依存して発症する場合について考える．この場合，複数の被験者から得られた網羅的な遺伝子発現データを遺伝子発現量でクラスタリングすることで，このような原因遺伝子を特定できる可能性がある．

クラスタリングアルゴリズムは 1) 分割に基づくクラスタリング，2) 階層的クラスタリング，3) その他（統計モデルやニューラルネットワークに基づくものなど）に大別できる．

5-2-1 分割に基づくクラスタリング

分割に基づくクラスタリングでは，データは以下の条件を満たすものとする．データはオブジェクトからなり，各オブジェクトは必ず 1 つのクラスタ（グループ）に属する．また，オブジェクトを含まないクラスタは存在しない．分割に基づく方法の代表的なものとして k-means 法がある[1)]．k-means 法では，クラスタの類似度は，クラスタ内のオブジェクトの平均（Mean）として測定される．同じクラスタ内の類似度は大きく，異なるクラスタ間のオブジェクトの類似度は小さくなるようにクラスタリングを行うことで，クラスタの分類を行う．以下にそのアルゴリズムを簡単に説明する．

1. k 個の任意のオブジェクトを選び，k 個のクラスタを作成し，その重心の値を初期値とする．
2. 各オブジェクトを最も近い重心を持つクラスタに割り振り直す．
3. 各クラスタの重心を新しく割り当てられたオブジェクトに基づいて再計算する．
4. 各クラスタの重心が移動しなくなるまで，2〜3 の操作を繰り返す．

5-2-2 階層的クラスタリング

階層的クラスタリングは，階層構造（木構造）のクラスタを作成する方法である．階層構造を作成するとき，葉の方向から根の方向にボトムアップに木構造を作成する Hierarchical Agglomerative Clustering（HAC）と，根の方向から葉の方向にボトムダウンに木構造を作成する Hierarchical Divisive Clustering（HDC）がある．HAC では 1 クラスタに 1 オブジェクトのみが含まれる状態から，類似したクラスタをまとめていく方式であり，クラスタ総数は減少する．一方，HDC では，すべてのオブジェクトを含む 1 つのクラスタから，クラスタを徐々に分割していく方式であり，クラスタ総数は増加する．

まず，HAC のアルゴリズムを簡単に解説する．

1. 各オブジェクトをただ 1 つ含むクラスタの集合を作る．
2. クラスタ集合の中から互いに最も類似した 2 つのクラスタを選んで，集合から削除する．

3. 削除した 2 つのクラスタをマージして 1 つのクラスタとしてクラスタ集合に追加する．

4. 設定した条件（クラスタ総数の下限など）が成立するまで，2〜3 の操作を繰り返す．

ここで，クラスタの類似度としては，クラスタ間の最短距離などを用いることが多い．
次に，HDC のアルゴリズムを簡単に解説する．

1. すべてのオブジェクトを含むクラスタ 1 つからなるクラスタ集合を作成する．

2. 所定の規則（例えば，最も類似しているオブジェクトの距離が最大となるところ）に従って集合を分割して新たなクラスタを作成する．

3. 作成したクラスタをクラスタ集合に追加する．

4. 設定した条件（クラスタ総数の上限など）が成立するまで，2〜3 の操作を繰り返す．

■参考文献

1) J. Qi, Y. Yu, L. Wang, J. Liu, and Y. Wang：“ An effective and efficient hierarchical K-means clustering algorith, ” Int J Distributed Sensor Networks, 2017, 13(8), 2017.

■S2 群 - 6 編 - 5 章

5-3 タンパク質-タンパク質結合体の予測

（執筆者：稲岡秀檢）[2018 年 2 月 受領]

タンパク質-タンパク質複合体は，シグナル伝達，分子スイッチング及びユビキチン化などの多様な機能を持つ．X 線結晶学，NMR 分光法及び電子顕微鏡を用いてタンパク質-タンパク質複合体の三次元構造を決定し，表面プラズモン共鳴（SPR），等温滴定熱量測定（ITC），蛍光分光法，分光光度アッセイ，ラジオリガンド結合，停止流蛍光測定などを用いて，結合の熱力学的パラメータ及び速度論的パラメータを得る[1]．こうした実験データから，界面での残基部位に結合するタンパク質-タンパク質複合体の親和性や，相互作用，分子レベルでのタンパク質-タンパク質認識のメカニズムを理解するための熱力学的パラメータなどが得られる．このようにして得られた情報を利用して，以下に示すタンパク質-タンパク質複合体の様々な局面を予測する計算アルゴリズムが開発さている[2]．

・タンパク質-タンパク質相互作用の予測
・非結合タンパク質の既知の三次元構造からの結合部位の予測
・一本鎖のアミノ酸配列からの結合部位の予測
・非結合タンパク質の構造を用いたタンパク質-タンパク質複合体の三次元構造の予測
・三次元構造を用いた複合体の結合親和性の予測
・相互作用するタンパク質の配列からの複合体の結合親和性の予測
・一塩基突然変異における複合体の結合親和性変化の予測

結合していないタンパク質を相互作用させることによって形成された複合体の三次元構造を予測することは，原子レベルでの生体分子認識の原理の理解や，分子機能及び構造に基づく薬剤設計にとって重要である．これらの予測には，剛体ドッキング法あるいはフレキシブルドッキング法が利用される[3]．

剛体ドッキング法では，双方の相互作用するタンパク質は完全に剛性であると考え，ドッキング問題の複雑さを 6 自由度における最適な方向の探索にまで減少する．この手法では，2 つのタンパク質の化学的・幾何学的な適合を最適化することを計算の目的としている．代表的な例としては，酵素阻害剤複合体がある．

剛体ドッキング法では計算が効率化されており，数分程度で計算を行うことができるため，巨大なデータベースを短時間で自動的にスクリーニングできる（ただし，配座の再配置と界面溶媒和に関しては考慮しない）．

フレキシブルドッキング法では，アミノ酸側鎖の回転，タンパク質ドメインの相対運動，及びすべての原子に自由度を持たせ，タンパク質の配座の持つ柔軟性を考慮して計算を行う．この手法では，形状ベースの指標の最適化，エネルギーの最小化，原子レベルのシミュレーションを目的としているため，現在のコンピュータハードウェアでは，単一のドッキングについて計算した場合でも，数時間から数日間の大規模な計算が必要となる．フレキシブルドッキング法の計算において問題点となる主な部分はスコアリング関数と立体構造検索である．以下に各種アルゴリズムやツールの概要を示す．

5-3-1 スコアリング関数

一般に，非共有相互作用，溶媒効果，統計及び接触電位，予測される結合部位，形状相補

性，幾何学，知識ベースのアプローチ，物理的原理及び経験的方法などがスコアリング関数を開発するために使用されている．

タンパク質の構造から自然な形として同定するために，傾向値，剛体を用いた最適化，インタフェースの柔軟性，進化的情報及び，エネルギー，保存性，インタフェース性を組み合わせたコンセンサススコアリング関数も提案されている．

スコアリング関数として，距離に依存する知識ベースのポテンシャルを利用した方法[4]や，データ駆動ドッキングを利用した方法[5]，形状相補性及び物理化学的特性を利用した方法[6]，結合長，結合角，二面角，静電相互作用，ファンデルワールス力，極性溶媒和，非極性溶媒和及びエントロピー，ならびに非極性，極性及び荷電残基の異なる誘電率値，アラニンスキャニング突然変異誘発から得られた実験的結合自由エネルギーなどの追加因子が含まれるエネルギー寄与に基づく方法[7]，分子間界面相互作用を，原子力学・分子力学を用いて解明するために，エネルギーの最小化によるドッキングを行うときに，インタフェースの柔軟性と剛体の最適化を組み合わせ，複雑な構造を自然なタンパク質構造に近づける方法[8]，形状相補性，静電相互作用親和性関数及び知識ベースの界面傾向を用いてドッキングプロトコルを改良し，構造を再順位付けするために溶媒和エネルギー（GBSA）を用いる方法[9]，界面でのいくつかの構造的特徴（面積，短絡，保存，空間的クラスタリング，正に荷電した疎水性残留物の存在など）を計算し，これらの機能を利用してドッキングポーズをランク付けし，ドッキングアルゴリズムで得られた最適なドッキングポーズを同定する方法[10]，界面近傍の水分を含めるることで，より自然に近い構造を同定するため，再ランキングアルゴリズムを実装し，更に構造的特徴を利用した機械学習による方法[11]などが開発されている．

5-3-2 タンパク質-タンパク質複合体の構造予測

非結合タンパク質を用いたタンパク質-タンパク質複合体の三次元構造予測の計算で用いられるアルゴリズムは，

- 形状相補性
- 経験的自由エネルギー推定
- 非回転体側鎖最適化の勾配ベース最小化
- 非結合構造からの情報
- 静電及び脱溶媒和エネルギー
- 構造骨格及び側鎖の柔軟性
- 階層的アプローチ
- 界面サイズ
- 化学的架橋
- 進化的情報
- 知識に基づく推論
- 配列類似性
- 物理化学的性質

などがある．

5-3-3 タンパク質-タンパク質複合体の結合親和性

(1) 結合親和性のデータベース

結合親和性に影響を及ぼす因子や，効率的な予測方法を設計する因子を理解するために，実験的に決定されたタンパク質-タンパク質複合体の熱力学的データをデータベース化することが重要となる．

ASEdb（The Alanine Scanning Energetics database）[12]は，アラニンスキャニング（タンパク質のアミノ酸残基を一つずつアラニンに置換する実験法）のデータベースであり，突然変異複合体の熱力学データを取り扱う．

PINT（The Protein-protein Interactions Thermodynamic Database）[13]は，タンパク質-タンパク質相互作用熱力学データベースであり，解離定数（Kd），結合自由エネルギー（DG），エンタルピー及び熱容量変化などの熱力学的データを，実験条件，アミノ酸配列，複合体構造，関連文献情報と組み合わせて管理している．Protein-Protein Interaction Affinity Database [14]は，複合体の結合親和性と遊離タンパク質と複合体の構造に関するデータベース，PDBBind データベース[15]は，既知の構造の複合体に対する実験的結合親和性測定値に関するデータベースである．SKEMPI（Structural database of Kinetics and Energetics of Mutant Protein Interactions）[16]は，PDB 構造が利用可能である突然変異タンパク質-タンパク質複合体の熱力学データのデータベースである．SKEMPI は，タンパク質-タンパク質複合体の突然変異による結合親和性または自由エネルギーの変化を予測する様々な方法のためのトレーニング及び/または試験データセットとしてしばしば利用されている．

(2) 結合親和性に関連するパラメータ

結合親和性に関連するパラメータとしては，界面情報と配座変化[17]や，水素結合[18]や，突然変異時に 2 kcal/mol 以上の結合自由エネルギー変化を引き起こす比較的少数の界面残基（ホットスポット）がある．結合親和性を定量的に推定するために，接近可能な表面積の変化を用いて最小限の溶媒和に基づいて定式化されたモデル[19]も提案されている．アロステリック効果（モジュレータ及び翻訳後修飾によるタンパク質の構造または動態の変化）も，結合親和性にとって重要であると考えられている．

5-3-4 構造ベースパラメータを用いた結合時の自由エネルギー変化予測

配座変化，原子ペアポテンシャル，タンパク質界面情報や，知識ベースのエネルギー関数を用いた経験的スコア関数などの構造ベースパラメータがタンパク質-タンパク質複合体の結合親和性を予測するために使用されている．タンパク質-タンパク質複合体の結合親和性を予測するアルゴリズムでは，知識ベースのアプローチ，配座変化，重回帰技術などが利用されている．現在までに，タンパク質-タンパク質複合体の結合親和性と構造記述子を関連付けるための定量的構造-活性関係（QSAR）モデル[20]や，残基接触と非相互作用表面に由来する計算モデル[21]などが提案されている．

これらの方法は，タンパク質-タンパク質親和性予測の分野において著しい進歩を示しているが，トレーニングセットでの性能は良好であるが，テストセットにおける実験結果から得られた親和性と予測された親和性との間の相関が低いことや，抗原-抗体複合体の結合親和性をほとんど予測できないといった問題点もある．

5-3-5 結合親和性の配列ベース予測

構造ベースのアルゴリズムのほかに，タンパク質-タンパク質複合体をその結合親和性に基づいて分類する方法[22]や，機能情報を用いて親和性の絶対値を予測する配列ベースの方法[23]も提案されている．タンパク質相互作用ネットワークを構築し[24]，様々な生物から得られた大規模なタンパク質-タンパク質相互作用データを分析するために二項分類モデル[22]が使用されている．タンパク質-タンパク質複合体の結合親和性は，生物系でその複合体が行う機能に依存するという仮説に基づき，結合親和性の実際の値を予測するための回帰モデル[23]も開発されている．

配列ベース法は，相互作用するタンパク質の異なる結合ポーズについての予測ができないことや，配座変化の説明ができないといった手法上の限界があるが，大量の高品質実験データを利用し，複合体をグループ化する方法論を最適化することで，配列ベース法を改善することが可能である．

5-3-6 突然変異時の結合親和性の予測

タンパク質中のアミノ酸残基の置換は，その構造，安定性，結合親和性及び機能を変化させる．そのため，疾患につながる置換も存在する．タンパク質-タンパク質複合体では，結合親和性の変化は重要な因子であり，突然変異による結合自由エネルギー変化の予測は重要である．

既知のタンパク質複合体構造に由来する統計的情報（構造骨格の捻れ角，溶媒の接近可能性，アミノ酸のタイプ，残基間距離）に基づいた予測法[25]が開発されている．この手法の利点は，1 回の操作でタンパク質-タンパク質複合体中のすべての突然変異体の結合親和性を予測できることである．界面構造プロファイルから結合自由エネルギーの変化を予測する方法[26]や，ポアソン-ボルツマン表面積連続溶媒和（MM-PBSA）と組み合わせた，構造最小化，統計的エネルギースコアリング関数と分子力学を利用する方法[27]，複合体タンパク質構造に突然変異をマッピングし，突然変異に関連する変化を計算することで突然変異の有害な影響を予測する分子力学，側鎖最適化アルゴリズムを用いた方法[28]，半経験的エネルギー項，分子内及び分子間接触，溶媒接触可能表面積及び配列保存を利用した機械学習を用いる方法[29]が開発されている．

突然変異時の結合親和性の予測法の主な問題点は，精度と計算速度の双方を達成する方法がないことである．BeAtMuSiC という計算方法[30]は，数秒以内に突然変異の際の結合自由エネルギー変化を予測可能であるが，大きな誤差が生じる．SAAMBE[31]，ELASPIC[29]，MutaBind[32]などの計算方法では高い精度が得られるが，長い計算時間が必要とする．

■参考文献

1) G. Sudha, R. Nussinov, and N. Srinivasan：“ An overview of recent advances in structural bioinformatics of protein-protein interactions and a guide to their principles, ”Prog Biophys Mol Biol, 116, pp.141-150, 2014.
2) M.M. Gromiha, K. Yugandhar, and S. Jemimah：“ Protein-protein interactions: scoring schemes and binding affinity, ” Current Opinion in Structural Biology, 44, pp.31-38, 2017.
3) S.J. de Vries, C.E. Schindler, I.C. de Beauchene, and M. Zacharias：“ A web interface for easy flexible protein-protein docking with ATTRACT, ” Biophys J., 108, pp.462-465, 2015.

4) D.M. Krüger, J.I. Garzón, P. Chacón, and H. Gohlke：“ DrugScorePPI knowledge-based potentials used as scoring and objective function in protein-protein docking, ” PLoS One, 9, e89466, 2014.
5) J. Segura, M.A. Marín-López, P.F. Jones, B. Oliva, and N. Fernandez-Fuentes：“ VORFFIP-driven dock: V-D2OCK, a fast and accurate protein docking strategy, ” PLoS One, 10, e0118107, 2015.
6) M. Ohue, T. Shimoda, S. Suzuki, Y. Matsuzaki, T. Ishida, and Y. Akiyama：“ MEGADOCK 4.0: an ultra-high-performance protein-protein docking software for heterogeneous supercomputers, ”Bioinformatics, 30, pp.3281-3283, 2014.
7) I.S. Moreira, J.M. Martins, J.T. Coimbra, M.J. Ramos, and P.A. Fernandes：“ A new scoring function for protein-protein docking that identifies native structures with unprecedented accuracy, ” Phys Chem Chem Phys., 17, pp.2378-2387, 2015.
8) C.E. Schindler, S.J. de Vries, and M. Zacharias：“ iATTRACT: simultaneous global and local interface optimization for protein-protein docking refinement, ” Proteins, 83, pp.248-258, 2015.
9) R. Chowdhury, M. Rasheed, D. Keidel, M. Moussalem, A. Olson, M. Sanner, and C. Bajaj：“ Protein-protein docking with F(2)Dock 2.0 and GB-rerank, ” PLoS One, 8, e51307, 2015.
10) S. Malhotra, K. Sankar, and R. Sowdhamini：“ Structural interface parameters are discriminatory in recognising near-native poses of protein-protein interactions, ” PLoS One, 9, e80255, 2014.
11) C.T.T. Su, T.D. Nguyen, J. Zheng, and C.K. Kwoh：“ IFACEwat: the interfacial water-implemented re-ranking algorithm to improve the discrimination of near native structures for protein rigid docking, ” BMC Bioinformatics, 15:S9, 2014.
12) K.S. Thorn and A.A. Bogan：“ ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions, ” Bioinformatics, 17(3):284-285, 2001.
13) M.D. Kumar and M.M. Gromiha：“ PINT: Protein-protein Interactions Thermodynamic Database, ”Nucleic Acids Res., 34, D195-D198, 2006.
14) T. Vreven, I.H. Moal, A. Vangone, B.G. Pierce, P.L. Kastritis, M. Torchala, R. Chaleil, B. Jiménez-García, P.A. Bates, J. Fernandez-Recio, A.M. Bonvin, and Z. Weng：“ Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2, ” J Mol Biol., 427(19), pp.3031-3041, 2015
15) Z. Liu, Y. Li, L. Han, J. Li, J. Liu, Z. Zhao, W. Nie, Y. Liu, and R. Wang：“ PDB-wide collection of binding data: current status of the PDBbind database, ” Bioinformatics, 31(3), pp.405-412, 2014.
16) I.H. Moal and J. Fernández-Recio：“ SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models, ” Bioinformatics, 28(20), pp.2600-2607, 2012.
17) J. Janin：“ A minimal model of protein-protein binding affinities, ” Protein Sci., 23(12), pp.1813-1817, 2014.
18) A. Erijman, E. Rosenthal, and J.M. Shifman：“ How Structure Defines Affinity in Protein-Protein Interactions, ” PLoS One, 9(10), e110085, 2014.
19) J.M. Choi, A.W. Serohijos, S. Murphy, D. Lucarelli, L.L. Lofranco, A. Feldman, and E.I. Shakhnovich：“ Minimalistic predictor of protein binding energy: contribution of solvation factor to protein binding, ” Biophys J., 108(4), pp.795-798, 2015.
20) P. Zhou, C. Wang, F. Tian, Y. Ren, C. Yang, and J. Huang：“ Biomacromolecular quantitative structure-activity relationship (BioQSAR): a proof-of-concept study on the modeling, prediction and interpretation of protein-protein binding affinity, ” J Comput Aided Mol Des., 27(1), pp.67-78, 2013.
21) A. Vangone and A.M. Bonvin：“ Contacts-based prediction of binding affinity in protein-protein complexes, ” eLife, 4, e07454, 2015.
22) K. Yugandhar, M.M. Gromiha：“ Feature selection and classification of protein-protein complexes based on their binding affinities using machine learning approaches, ” Proteins, 82(9), pp.2088-2096, 2014.
23) K. Yugandhar and M.M. Gromiha：“ Protein-protein binding affinity prediction from amino acid se-

quence, ” Bioinformatics, 30(24), pp.3583-3589, 2014.
24) K. Yugandhar and M.M. Gromiha :“ Analysis of protein-protein interaction networks based on binding affinity, ” Curr Protein Pept Sci., 17(1), pp.72-81, 2016.
25) Y. Dehouck, J.M. Kwasigroch, M. Rooman, and D. Gilis :“ BeAtMuSiC: Prediction of changes in protein-protein binding affinity on mutations, ” Nucleic Acids Res., 41, pp.W333-W339, 2013.
26) J.R. Brender and Y. Zhang :“ Predicting the Effect of Mutations on Protein-Protein Binding Interactions through Structure-Based Interface Profiles, ” PLoS Comput Biol., 11(10), e1004494, 2015.
27) M. Petukh, M. Li, and E. Alexov :“ Predicting Binding Free Energy Change Caused by Point Mutations with Knowledge-Modified MM/PBSA Method, ” PLoS Comput Biol., 11(7), e1004276, 2015.
28) M. Li, F.L. Simonetti, A. Goncearenco, and A.R. Panchenko :“ MutaBind estimates and interprets the effects of sequence variants on protein-protein interactions, ”Nucleic Acids Res., 44(W1), W494-W501, 2016.
29) N. Berliner, J. Teyra, R. Colak, S.G. Lopez, and P.M. Kim :“ Combining structural modeling with ensemble machine learning to accurately predict protein fold stability and binding affinity effects upon mutation, ” PLoS One, 9(9), e107353, 2014.
30) Y. Dehouck, J.M. Kwasigroch, M. Rooman, and D. Gilis :“ BeAtMuSiC: Prediction of changes in protein-protein binding affinity on mutations, ” Nucleic Acids Res., 41, pp.W333-W339, 2013.
31) M. Petukh, L. Dai, and E. Alexov :“ SAAMBE: Webserver to Predict the Charge of Binding Free Energy Caused by Amino Acids Mutations, ” Int J Mol Sci., 17(4), E547, 2016.
32) M. Li, F.L. Simonetti, A. Goncearenco, and A.R. Panchenko :“ MutaBind estimates and interprets the effects of sequence variants on protein-protein interactions, ” Nucleic Acids Res., 44(W1), pp.W494-W501, 2016.

■S2 群 - 6 編 - 5 章

5-4 深層学習による解析

（執筆者：稲岡秀檢）［2018 年 2 月 受領］

5-4-1 タンパク質-核酸相互作用

遺伝子調節制御の基本単位は，調節タンパク質とその標的 DNA または RNA 分子との間の接触である．これらの相互作用を直接的に予測する生物物理学的モデルは，いまだ不完全であり，特定のタイプの構造に限定される．しかし，大規模な実験データセットによる計算解析は，標的配列における過剰発現によって制御モチーフを同定することを可能にする．

多様な実験データセットからタンパク質-核酸相互作用を計算するための深層学習として DeepBind [1]と呼ばれるアルゴリズムが広範囲に適用可能である．従来の手法に比べて予測力が向上し，モチーフの予測や，RNA 編集や選択的スプライシングについての予測も可能となっている．

DeepBind は，*in vivo* での DNA 結合（クロマチン免疫沈降）及び *in vitro*（タンパク質マイクロアレイ）及び *in vitro* での RNA 結合（ハイスループットスクリーニング）を含む，およそ 1000 の公に利用可能なデータセットを用いて性能が評価された．DeepBind は，タンパク質マイクロアレイスコアをほぼ完全に正確に予測し，高い精度で ChIP-seq の結果を予測し（ROC 曲線下面積 = 0.7），複製実験と同様に RNA 結合を予測した．

DeepBind は，モチーフ発見と結合エネルギー予測を含む既存のバイオインフォマティクス手法と似ているが，モデルパラメータと複雑さが自動的に選択されるという利点がある．近年のトレーニング方法の進歩により，高速な学習アルゴリズムが開発されたため，可能な限り多くの構造の組合せを考慮しても，DeepBind のトレーニングは実用的になる．ENCODE や Roadmap pigenomics など，ますます大きくなっているデータセットのマイニングに広く役立つ．

5-4-2 単一細胞のメチル化解析

DNA メチル化は，最も広範に研究されているエピジェネティックマークの一つであり，染色体安定性，X 染色体不活性化，細胞分化，がん進行及び遺伝子調節を含む広範囲の生物学的プロセスに関与することが知られている．

最近の技術的進歩により，ゲノムワイドバイサルファイト塩配列決定プロトコル（scBS-seq），または，解析部位を限定したバイサルファイト塩基配列決定プロトコル（scRRBS-seq）のいずれかを用いて，単一細胞分解能で DNA メチル化をプロファイリングすることが可能になった．

単一細胞のメチル化状態を予測するための深層ニューラルネットワークに基づく計算方法として DeepCpG [2]がある．DeepCpG は，DNA 配列パターンとメチル化状態の間，ならびに個々の細胞内及び細胞間の隣接 CpG 部位間の関連を活用する．従来の方法と異なる点は，特徴量の抽出とモデルのトレーニングを分離せず同時に行うことである．DeepCpG はモジュラアーキテクチャに基づいており，データ駆動方式で予測 DNA 配列とメチル化パターンを学習する．全ゲノム単一細胞メチル化プロファイリング（scBS-seq）を用いてプロファイリングしたマウス胚性幹細胞ならびに還元表現プロトコル（scRRBS-seq）を用いてプロファイ

リングしたヒト及びマウス細胞について DeepCpG を評価したところ，2 つの細胞タイプで，DeepCpG は以前のアプローチよりもメチル化状態のより正確な予測が可能であった．更に，DeepCpG は，メチル化の変化，及び細胞間のメチル化の変動性に関連した既知及び新規の配列モチーフの両方を明らかにした．

5-4-3 網羅的メチル化解析

DNA メチル化は，シトシンまたはアデニンの 5 番目の炭素にメチル基を付加することを表す．DNA メチル化は，配列中のグアニンがシトシンに続く CpG 部位でより頻繁に生じる．幾つかの領域では，CpG 部位の頻度は平均値の 10 倍となる．これらの領域は CpG アイランド（CGI）と呼ばれている．CpG アイランドは少なくとも 200 塩基対の長さで 50 %以上の GC 含有量を有する．一般に，CGI の外側の CpG 部位はほとんどメチル化されているが，CGI の CpG 部位はほとんどメチル化されていない．この相違は，CGI が通常，区別されたメチル化のパターンを有することを意味し，これは遺伝子調節または遺伝子突然変異において重要であり得る．

DNA メチル化は，遺伝子の発現及びタンパク質の機能調節に影響を及ぼすことが見出されている．DNA メチル化は，様々ながん及び複雑な疾患の発症及び進行に影響を及ぼす可能性がある．異常な細胞株ではメチル化されたプロモータとサプレッサが多く見られる．DNA メチル化の異常は，急性骨髄性白血病などのがんの典型的な特徴の一つである．しかし，DNA メチル化の異常と白血病との間のメカニズムについてはよく分かっていない．乳がんなどの様々ながんにおける DNA メチル化を調べた結果は，異常な DNA メチル化が，通常，幾つかの特定のゲノム位置で生じることを示している[3)]．

メチル化配列決定技術の最近の進歩により，DNA 中のゲノムワイドなメチル化部位の同定が可能になった[4)]．DNA のメチル化パターンをプロファイリングする一つの方法は，DNA のバイサルファイト処理とそれに続くバイサルファイト塩基配列決定と呼ばれる次世代シーケンシングの使用によるものである．現在のバイサルファイト塩基配列決定法は，ゲノムワイドなバイサルファイト塩基配列決定（Whole-genome Bisulfite Sequencing：WGBS）及び解析部位を限定したバイサルファイト塩基配列決定（Reduced Representation Bisulfite Sequencing：RRBS）を含む．WGBS と比較して，RRBS はゲノムの代表的な分画を用いて配列決定の量を減少させる．したがって，RRBS は高い CpG 含有量を有する領域のメチル化パターンを特異的にプロファイリング及び分析する．

ゲノムのウィンドウまたはセグメントにおける CpG 部位のメチル化状態を予測する方法が開発されている[5)]．メチル化予測の現在の方法の大部分は，メチル化状態がバイナリクラス，すなわち CpG 部位またはウィンドウがメチル化またはメチル化されていない（メチル化耐性）ものであると仮定する．しかしながら，ほかのいくつかの方法では，メチル化レベルを 2 段階以上のクラスに分類している．これらの方法のなかで，予測は通常，CGI のような特定の領域に限定されていた．これらの方法によって使用される予測機能には，DNA 組成，GC 含量，配列パターン，及び隣接領域のメチル化状態が含まれている．最近の方法では，ゲノムのメチル化部位を予測するために疑似ヌクレオチド組成を使用する[6)]．連続した領域の DNA 組成及びメチル化状態は，これらの方法で用いられる特徴量で最も一般的なものである．

DNA メチル化の予測に使用されていない特徴の一つは，染色体相互作用である．Hi-C（染

色体高次構造把握，Chromosome Conformation Capture：3C の拡張）は，ゲノム内の染色体内及び染色体間の接触の両方の調査を可能にする[8]．1～1000 キロベースの分解能でゲノムを分析すると，ゲノム全体の立体配座が捕捉される．1 キロベースの解像度は，ゲノム内の遺伝子間の接触を更に捕捉することができる．Hi-C 実験は架橋 DNA を制限酵素で切断し，分子間ライゲーションに有利な非常に希薄な条件下でそれらを連結する．次いで，連結された DNA セグメントを精製して剪断し，対の末端の読み取る．対になった Hi-C 配列は，参照ゲノムにマッピングされる．マッピング後，データはビニング（いくつかの配列を一つの配列として見立てること）され，Hi-C コンタクトライブラリーに正規化される．これは，特定の位置が三次元空間で空間的に接近していることを示す．

特定の領域のメチル化状態を予測するための多くの方法が開発されているが，長い非コード RNA（long non-coding RNAs：lncRNA）の遺伝子座における CpG 部位のメチル化状態の予測はほとんど注目されていない．lncRNA は，200 塩基から 100 キロ塩基（kb）の範囲の非コード遺伝子の転写物であるが，ヒト疾患におけるそれらの潜在的活性はほとんど明らかにされていない．近年の研究結果から，lncRNA が DNA と特異的なクロマチンリモデリング活性との間のコネクタとして機能し[9]，lncRNA の発現レベルが通常，タンパク質コード遺伝子の発現レベルよりも低いことが示されている[10]．更に，lncRNA 発現は発癌の主要な要因であるかもしれない．lncRNA がどのようにがんに影響を及ぼすかについての正確なメカニズムは不明であるが，異常な lncRNA 発現は，主要な遺伝子プロセスに影響を及ぼすことによってがんを引き起こす要因となりうる．

CpG 部位の DNA メチル化状態を予測するための，積み重ねノイズ除去オートエンコーダ（Stacked denoising Autoencoders：SdAs）を適用した深層機械学習手法として DeepMethyl[11]がある．従来の学習アルゴリズムとは異なり，SdAs のトレーニングには，ラベル無しデータを使用した教師無しの前学習ステージと，ラベル付きデータ（既知の目標値を持つデータ）を使用した教師有り微調整ステージの 2 つのステージが含まれている．特徴量としてはゲノムのウィンドウ内で生成された配列特徴量と，Hi-C 実験によって示されたゲノムの三次元トポロジーから生成された特徴量を用いる．

5-4-4 エピジェネティクス

エピジェネティックスは，「DNA 配列とは無関係にゲノム活性を制御する DNA の周辺の分子因子であり，有糸分裂的に安定である」と定義される[12]．各細胞の種類には，細胞の特異的な分化を可能にするユニークなエピゲノムがある．単一の遺伝子型が多くの表現型と関連し得るので，単一のゲノム配列に対して，無限のエピゲノムが存在し得ると考えられる．主なエピジェネティクス機構の一つは DNA メチル化であり，これは DNA 配列を変化させることなく遺伝子発現に影響を及ぼし得る．付加的なエピジェネティック機構には，ヒストン修飾，非コード RNA（ncRNA），及びクロマチン構造が含まれる．

DNA メチル化は，雄性生殖系列を介した世代的遺伝を媒介することが示されており，いままでに多くの研究が行われたエピジェネティクス機構の一つである[13]．多くの研究は，エピジェネティックな変化が，発生過程（例えば，組織形成，器官形成，性決定）に必須であることを示している．エピジェネティックな変化は，また遺伝子発現の変化したパターンをもたらし，肥満，アレルギー，がん，統合失調症，またはアルツハイマー病などの有害な臨床

転帰につながり得る．最近のエピジェネティックな研究は，環境化合物または曝露がどのようにして世代を通じて伝達されるエピジェネティックな疾患状態を促進できるかに焦点を当てている[12]．エピジェネティクス，生物学，及び疾患を理解するためには，疾患に関連するエピジェネティックな変化に対する感受性の領域を予測することが重要である．

この領域における研究の主要な目標は，エピジェネティックな修飾の影響を受けやすいゲノム内の領域を同定することである．これは，DNA メチル化変化（例えば，CpG），ヒストン修飾，ncRNA 発現，またはクロマチン構造変化（例えば，ヌクレオソーム配置）を含み得る．

エピジェネティックな現象を実験から再現することが困難であること，また実験から生物学的データの抽出及び分析は，実験費用が高価であること，計算時間がかかることが問題となる．また，生物学的データセットは高い次元を有するが，関心のある症例（例えば，疾患状態）は比較的まれである．エピジェネティックなデータセット，例えば DNA メチル化データでは，多数の DNA 配列及びゲノム特徴量で記述されており，データとしては十分に高次元化されているが，抽出したい差分的にメチル化された DNA 領域（Differentially Methylated DNA Regions：DMR）は極めて少なく，大部分は非 DMR 部位がある．

これらの課題に対処するためには，エピジェネティックデータセットに特徴量の自動生成，特徴量の自動選択，機械学習を組み合わせた統合アプローチが必要である．

この統合エピジェネティックデータを生成するための能動学習（Active Learning：ACL），データの後天的突然変異の発生率が比較的低いことに対処するための不均衡なクラス学習（Imbalanced Class Learning：ICL），関連するゲノム特徴を手動で定義することの難しさに対処するための深層学習（Deep Learning：DL）の組合せを含む代替アプローチが想定されている[14]．

ACL 及び ICL は，手動で生成された特徴から効率的に学習するために使用される．DL は，ACL / ICL のための特徴を自動的に生成するために使用される．

■参考文献

1) 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.
2) C. Angermueller, H.J. Lee, W. Reik, and O. Stegle :“ DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning, ” Genome Biology, 18:67, 2017.
3) Cancer Genome Atlas Research Network :“ Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia, ” N Engl J Med., 368(22), pp.2059-2074, 2013.
4) Z.D. Smith, H. Gu, C. Bock, A. Gnirke, and A. Meissner :“ High-throughput bisulfite sequencing in mammalian genomes, ” Methods, 48(3), pp.226-232, 2009.
5) R. Das, N. Dimitrova, Z. Xuan, R.A. Rollins, F. Haghighi, J.R. Edwards, J. Ju, T.H. Bestor, and M.Q. Zhang :“ Computational prediction of methylation status in human genomic sequences, ” Proc. Natl. Acad. Sci. USA, 103(28), pp.10713-10716, 2006.
6) Z. Liu, X. Xiao, W.R. Qiu, and K.C. Chou :“ iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide composition, ” Anal Biochem., 474, pp.69-77, 2015.
7) L. Harewood, K. Kishore, M.D. Eldridge, S. Wingett, D. Pearson, S. Schoenfelder, V.P. Collins, and P. Fraser :“ Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours, ” Genome Biology, 8:125, 2017.

8) Z. Wang, R. Cao, K. Taylor, A. Briley, C. Caldwell, and J. Cheng：“ The properties of genome conformation and spatial gene interaction and regulation networks of normal and malignant human cell types, ” PLoS One, 8(3), e58793, 2013.
9) E.A. Gibb, C.J. Brown, and W.L. Lam：“ The functional role of long non-coding RNA in human carcinomas, ” Mol. Cancer, 10, pp.38-55, 2011.
10) D. Ramsköld, E.T. Wang, C.B. Burge, and R. Sandberg：“ An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data, ” PLoS Comput Biol., 5(12), e1000598, 2009.
11) Y. Wang, T. Liu, D. Xu, H. Shi, C. Zhang, Y.Y. Mo, and Z. Wang：“ Predicting DNA Methylation State of CpG Dinucleotide Using Genome Topological Features and Deep Networks, ” Sci Rep., 6:19598, 2016.
12) M.K. Skinner：“ Endocrine disruptor induction of epigenetic transgenerational inheritance of disease, ” Mol Cell Endocrinol., 398(1-2), pp.4-12, 2014.
13) M. Manikkam, M.M. Haque, C. Guerrero-Bosagna, E.E. Nilsson, and M.K. Skinner：“ Pesticide methoxychlor promotes the epigenetic transgenerational inheritance of adult-onset disease through the female germline, ” PLoS One, 9(7), e102091, 2014.
14) L.B. Holder, M.M. Haque, and M.K. Skinner：“ Machine learning for epigenetics and future medical applications, ” Epigenetics, 12(7), pp.505-514, 2017.