■S2 群（ナノ・量子・バイオ）- 6 編（バイオインフォマティックス）

3 章 ハイスループット計測

（執筆者：福岡　豊）［2018 年 2 月 受領］

■概要■

様々な生物のゲノム配列の解読（シーケンシング）が完了し，遺伝子機能の解析及び疾患に関連する遺伝子群の同定などの研究が盛んに行われている．特定の細胞におけるタンパク質の量（発現量）を調べることは重要である．これは細胞の種類や状態（正常／疾患）により合成されるタンパク質が異なるからである．

細胞中のタンパク質の発現変動，相互作用などの全体をプロテオーム（Proteome）という．この言葉は Protein（タンパク質）と ome（ラテン語で「全体」を表す接尾語）との組合せからできている．ちなみに、ゲノム（Genome）も Gene（遺伝子）と ome からなる．ほかにも，全体を意味する言葉として，トランスクリプトーム（Taranscriptome = Transcription 転写 + ome），メタボローム（Metabolome = Metabolism 代謝 + ome），フィジオーム（Physiome = Physiology 生理機能 + ome）などがある．しかし，実際には網羅的というよりは大規模という意味合いで使われていることが多い．学問領域を表す場合には，Genomics や Proteomics のように語尾が omics に変化する．そこで，遺伝子やタンパク質などを大規模に調べる学問や研究領域を総称してオミックス（Omics）と呼んでいる．

このように大規模なオミックス解析が必要とされるのは，生命をシステムとして理解しようとする試みが始まっているからである．このような学問をシステムバイオロジーという（6 章参照）．本章では，DNA マイクロアレイなど，オミックスで使われるハイスループットな計測法について説明する．近年，公的資金を用いて得た研究成果の共有が世界的に進んでおり，ハイスループット計測によるデータが公的データベース（4 章参照）に蓄積されている．そのようなデータは誰でもダウンロードして利用することができる．データを処理するうえで，データが生成される原理やデータの特徴を知っておくことが望ましい．本章では測定原理や特徴の説明を中心として，データ解析法については 5 章で説明する．

オミックスの特徴を表 3.1 にまとめた[?]．大規模な解析により個々の要素間の相互作用が明らかになって，新しい知見が得られるのは当然であるが，そのほかにも多くの長所がある．新しい計測方法や信号処理法が開発されるであろうし，知識・研究成果の共有によって想定外の知見が得られることが期待される．一方で，網羅的なデータの取得のためには，従来は計測していなかったデータも計測することが必要になり，効率が低下するとの懸念もある．新しい技術の必要性は，裏を返せば，計測方法や信号処理方法が未確立であることを意味している．多くのデータを同時に計測する場合，コストの制約からデータの質の低下は避けられない．

【本章の構成】

本章では，ゲノム解析（ゲノミクス）やプロテオーム解析（プロテオミクス）の基礎となるハイスループットな計測技術について説明する．3-1 節では，遺伝子の発現を計測する技術である DNA マイクロアレイについて，その原理や遺伝子発現以外への適用例を述べる．3-2 節では，遺伝子の塩基配列を高速に決定する次世代シーケンサ（Next Generation Sequencer：

表 3.1　オミックスの特徴

長所	
	新しい知見
	新技術（計測・処理法）の創生
	知識・研究成果の共有
	組合せによる高価（予想外の知見）
短所	
	非効率性
	解析法が未確立
	データの質の低下

NGS）を用いた DNA 及び RNA の解析について説明する．3-3 節では，タンパク質の大規模な発現解析のための二次元電気泳動法について説明する．なお，本章では計測法の原理についてのみ説明する．ハイスループット計測で得られるデータの解析については 5 章で述べる．

■参考文献

1)　田中　博 (編) :“ 先制医療と創薬のための疾患システムバイオロジー, ” 培風館, 2012.
2)　福岡　豊 :“ 姿勢制御系のモデリングとフィジオーム, ” 信学技法, MBE2002-149, vol.102, no.728, pp.9-12, 2003.

■S2 群 - 6 編 - 3 章

3-1 DNA マイクロアレイ

（執筆者：福岡　豊）[2018 年 2 月 受領]

DNA マイクロアレイは[1, 2]，相補な配列を有する一本鎖 DNA 同士，あるいは一本鎖 DNA と RNA が結合（ハイブリダイズ）する性質を利用し，細胞中に存在する mRNA の種類と量を評価する技術である（図 3.1）．少量のサンプルで数千から数万程度の遺伝子の発現量を調べることができる．このような情報から細胞の状態（疾患の有無など）を推定できるものと考えられている．また，様々な条件での発現パターンを調べることにより，キーとなる遺伝子群を同定できるものと期待されている．

3-1-1 mRNA 量の測定

タンパク質の発現量を大規模かつ効率的に計測することは難しい（具体的な計測法については後述）．そこで，タンパク質の代わりに細胞中の mRNA の量（遺伝子の発現量）が計られている（トランスクリプトーム）．これは，細胞中の特定のタンパク質の量とそのタンパク質をコードする mRNA の量がほぼ比例しているからである．

DNA マイクロアレイによる解析の概要は以下の通りである．基板上に多くのスポットを作り，それぞれのスポットに特定の配列を持つ一本鎖 DNA を固定する．基板に固定される一本鎖 DNA としては，cDNA（mRNA から逆転写酵素を用いて作成）や合成オリゴヌクレオチドが用いられる．解析の対象とする細胞から抽出した mRNA を cDNA（または cRNA）に変換し，蛍光物質でラベリングする．ラベリングされた cDNA を含む溶液を基板上の DNA とハイブリダイズさせる（ハイブリダイゼーション，図 3.1A）．一定の時間が経過した後に，ハイブリダイズしていない cDNA を洗い流す．次に，マイクロアレイ上のスポット全体をイメージスキャナで蛍光画像に変換する．どのスポットがどのプローブ（遺伝子）に対応するかは分かっているので，各スポットの蛍光強度（図 3.1B）からサンプル中の mRNA の量が評価できる．

DNA マイクロアレイには，大きく分けて 2 通りの方式がある．一つはスタンフォード大学で開発された方式で，ガラス基板上に cDNA を貼り付けることによりマイクロアレイを作成する．また，この方式では 2 種類のサンプル（例えば、がん細胞と正常細胞）から得られる cDNA を異なる波長の蛍光色素でラベリングする．もう一方は Affymetrix 社によって開発された方法で，シリコン基板上にフォトリソグラフィを使って，オリゴヌクレオチドを合成することによりマイクロアレイを作成する．その際，1 プローブに積層できるのは 25 塩基程度であるので，1 つの遺伝子に対して複数のプローブを用意する必要がある．現在では，Affymetrix 社以外にも合成ヌクレオチドを用いたマイクロアレイを販売するメーカがあり，この方法が主流となっている．2 つの方式では，発現量として得られる数値の意味合いが異なる．スタンフォード方式では 2 種類のサンプルにおける発現量の比が得られるのに対し，Affymetrix 方式ではそれぞれのスポットに対応する遺伝子の発現量が得られるので注意が必要である．

3-1-2 mRNA 量以外の測定

近年では DNA マイクロアレイを mRNA 量以外の測定に用いる方法が開発されている．こ

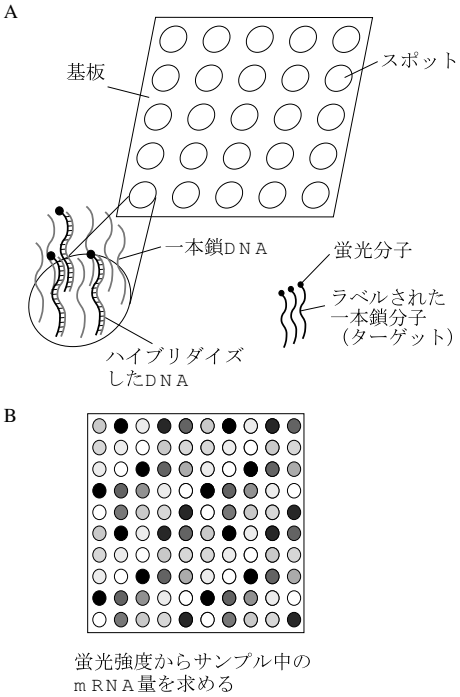

図 3.1　DNA **マイクロアレイの原理**

こでは，そのような方法について説明する．

SNP

一塩基多型（SNP）の変異が生じている場所から上流側の配列を使い，多型の塩基を含めて 2 種類のプライマーを作成する．塩基は ACGT の 4 種類があるため，厳密には 3 種類すべての多型に対してプライマーを用意する必要があるが，頻出する多型について検討することが多いため，ここでは 2 種類の多型について述べる．2 種類のプライマーをそれぞれ Cy3, Cy5 で標識する．多型の塩基配列の上流側から PCR を行い，変位が生じている場所までの塩基配列を得る．この配列は一塩基多型の情報を含む配列であるので「アレル配列」と呼ぶことにする．

同時に多型による変異が生じている場所から塩基配列の下流の配列を用いてプライマーを作成し，多形の下流側から PCR を行い，変位が生じている場所の直前までの塩基配列を得る．この配列は遺伝子配列の染色体上の部位特異的な配列を得るための作業であるため「場所配列」と呼ぶことにする．

この一連の操作により「アレル配列」と「場所配列」が結合した一塩基のみ配列が異なる 2 つの塩基配列が得られることになる[3)]．「場所配列」の部分に相補なプローブを持ったアレイを用意すれば，蛍光標識された配列がハイブリダイゼーションされることになる．一塩基

多型が生じていれば異なる色の蛍光色素で標識されているので，発光色の違いにより一塩基多型が生じている部位とその頻度を測定することができる．

DNA メチル化

一塩基多型の測定技術を応用することで DNA メチル化の度合いを測定することができる[4]．

DNA 配列にバイサルファイト処理を行うと，シトシン (C) がウラシル (U) に置換される．一方，メチル化されたシトシンは置換されずそのままメチル化シトシンとして存在する．もし DNA メチル化を測定したい CpG 部位がメチル化されていれば，その部位の CpG はそのままの状態であり，もしメチル化されていなければその部位の CpG は UG となる．つまり，バイサルファイト処理によりメチル化シトシン部位のみの塩基が異なる配列ができることになる．

一塩基多型の測定方法と同様の処理を行うことで，DNA メチル化が生じていれば異なる色素で標識された配列を得ることができる．この手法を用いて発光色の違いにより DNA メチル化が生じている部位とその頻度を測定することができる．

■参考文献

1) D. Stekel：“ Microarray Bioinformatics, ” Cambridge University Press, 2003.
2) 佐々木博己, 青柳一彦 (編)：“ DNA チップ実験まるわかり, ” 羊土社, 2004.
3) R. Shen, J.B. Fan, D. Campbell, W. Chang, J. Chen, D. Doucet, J. Yeakley, M. Bibikova, E.W. Garcia, C. McBride, F. Steemers, F. Garcia, B.G. Kermani, K. Gunderson, and A. Oliphant：“ High-throughput SNP genotyping on universal bead arrays, ” Mutation Research, vol.573, pp.70-82, 2005.
4) M. Bibikova, Z. Lin, L. Zhou, E. Chudin, E.W. Garcia, B. Wu, D. Doucet, N.J. Thomas, Y. Wang, E. Vollmer, T. Goldmann, C. Seifart, W. Jiang, D.L. Barker, M.S. Chee, J. Floros, and J.B. Fan：“ High-throughput DNA methylation profiling using universal bead arrays, ”Genome Research, vol.16, pp.383-393, 2006.

■S2 群 - 6 編 - 3 章

3-2 次世代シーケンサ

（執筆者：福岡　豊）[2018 年 2 月 受領]

3-1 節で述べたように，遺伝子の発現量は細胞の状態を反映するので，疾患の診断に利用できると考えられている．一方，ゲノムの配列は個体（個人）ごとに異なる．ある疾患の患者群と非罹患者の群のゲノム配列を比較すれば，その疾患に関連した変異が見つかる可能性が高い．そこで，個人の変異を調べれば疾患の罹りやすさが推定できると考えられている．

ゲノム配列の解析を行う装置をシーケンサという．初期のシーケンサでは，1 塩基ずつ配列を決定していたが，現在ではシーケンシングを高度に並列化して高速に行う装置が実用化されている．このような装置を次世代シーケンサ（NGS：Next Generation Sequencer）[1]という．NGS では，数日程度で個人のゲノム解析が実現可能であり，がんをはじめとする多くの疾患についてマーカとなる変異が探索されている．

3-2-1 DNA 配列解読の原理

NGS について説明する前に，従来の配列解読の原理について説明する．初期のシーケンサは 1970 年代後半に開発され，基礎を確立した研究者の名前を冠してサンガー法と呼ばれている[2]．NGS には幾つかの方式が存在するが，いずれもサンガー法の発展型であると考えることができる．

サンガー法では，長い DNA をそのまま解析することはできず，断片化する必要がある（NGS でも同様）．断片化には制限酵素と呼ばれる DNA を特定の部位で切断する酵素を用いる．制限酵素は特定の塩基配列を認識して DNA を切断する．制限酵素にも様々な種類があり，それぞれ切断部位（認識する配列）が異なっている．サンガー法では，数百～千塩基程度の断片を解析の対象とするので，そのような長さを持つ断片となるように調整する．

次に，解析したい断片の末端に特異的に結合するプライマーを準備する．プライマーとは，上記の断片とは相補な配列を持つ数塩基程度の一本鎖 DNA である．これを解析する断片にハイブリダイズさせる．DNA 複製酵素（DNA ポリメラーゼ）を用いて伸長反応を開始させることで，解析したい断片の複製ができる．その際に，特定の塩基で伸長反応を停止させる特殊なヌクレオチドを混ぜておく．例えば，A で停止するようヌクレオチドを混ぜると，配列中の A 部位で停止した様々な長さの断片が複製できる．通常のヌクレオチドも存在するので，これを取り込んだ場合は，通常の伸長反応が生じる．上記の操作を A，C，G，T のそれぞれについて行い，各塩基の出現部位で停止した断片を作る．この操作は，4 つの塩基それぞれに対して別個に行う．

最後に，電気泳動法を用いて，各断片の長さを調べる．電気泳動法では，DNA が負に帯電しており，直流電界中を陽極側に移動する性質を利用する．ポリアクリルアミドゲル中で泳動させると，長い断片はゲルの網目に引っかかって移動距離が小さくなる．一方，短い断片ほど遠くまで移動することから，移動距離によって長さを調べることができる．例えば，一番短い断片が C で停止していれば，最初の塩基は C である．二番目に短い断片が G，三番目が A で停止していれば，配列を CGA と決定できる．これを解析したい断片の全体にわたって行えば，全体の配列が決定できる．

高速化・効率化のためにキャピラリー中で電気泳動を行う装置が実用化されており，キャ

ピラリーシンケンサと呼ばれている．ヒトゲノム計画では，主にキャピラリーシーケンサが用いられた．しかし，30 億塩基を持つヒトのゲノム配列を決定するのに，世界中の拠点で並行して解析を進めても数年を要した．

シーケンシングの速度を飛躍的に向上させたのが NGS である[1)]．NGS でも 1 塩基ごとに伸長反応を行いながら配列を決定するが，電気泳動法ではなく，蛍光などの光学的原理を用いる．特定の塩基を用いて伸張させたときに，その塩基が結合した部位が光るように工夫しておく（何種類かの方法あり）．4 種類の塩基についてこれを行うと，ある位置の塩基が決定できる．一連の操作を繰り返すことで，順番に配列を決定できる．サンガー法では塩基ごとに別個に反応槽を準備する必要があったが，NGS では同じ反応槽中で 4 種類の塩基の伸長反応を行う．したがって，1 種類の塩基について反応させた後に，不要な塩基などをウォッシュ（洗浄）する必要がある．このような操作を超並列に行うことで，NGS ではシーケンシング速度の向上を図っている．

読み取り誤差があるので，1 回の測定だけでは不十分である．そこで，同じ部位由来の配列について何回かの読み取りを行う．このときに用いられるのが depth という量で，冗長度を表している．大きな depth は，その部位を何回も読み取っていることを意味する．depth が大きければ，読み取り誤差なのか個体における配列の変異なのかが区別できる．すなわち，変異であれば，何回読んでも同じ塩基になるはずであるが，誤差であれば同じ塩基となる可能性は低い．

NGS で解析できる断片の長さは，百～数百塩基程度とサンガー法よりも短い．このような配列（リードと呼ぶ）が大量に得られる．機種によっては数百 GB のデータを出力するので，リード数は何千万にも及ぶ．NGS で解読する配列をどのように得るかで，様々なアプリケーションに分類できる．ある生物のゲノム全体や，疾患に関連する遺伝子群に対応する DNA を解析の対象とするのが DNA-seq と呼ばれるアプリケーションである．一方，RNA-seq では細胞中の mRNA の配列の解析を行う．また，メチル化されたヒストンの近傍のみを切り出して解析するのが ChIP-seq である．いずれのアプリケーションにおいても，大量のデータから効率的に有用な知見を抽出する方法の開発が望まれている．

3-2-2 DNA 配列の解析

DNA を対象とした NGS のアプリケーションを DNA-seq と呼び，疾患関連遺伝子の変異の探索などを目的として行われる．一方，ゲノム配列が決定されていない生物について，新たに配列を決定するためにも使われる（デノーボシーケンシング）．それらに加えて，ゲノム配列が決定された生物について再度シーケンスするためにも使われる（リシーケンシング）．また，腸内細菌叢など，複数の細菌のゲノム解析に使われることもある．ゲノムサイズが大きい生物の場合，ゲノムの一部分を解析対象とする場合もある．例えば，タンパク質に翻訳される部分のみを対象とすることがあり，これをエクソーム解析という．ヒトゲノムではタンパク質に翻訳されるのは全体の 2%程度なので，効率的な解析が可能となる．

前述のように，NGS の出力は何百万から何千万にも及ぶリードである．一般的にリードの長さは百から二百塩基程度である．デノーボシーケンシング以外では，解析対象の生物の標準的なゲノム配列は決定されているので，それをリファレンス（参照配列）として用いる．各リードがリファレンスのどの部位に由来するかを同定することをマッピングという．リファ

レンスと比較することで，変異の有無も確認できる．デノーボシーケンシングでは，リファレンスが利用できないので，リードをつなぎ合わせて，標準的な配列を決定する必要がある．この作業をアセンブルという．

3-2-3 mRNA 配列の解析

RNA を解析対象とした NGS のアプリケーションを RNA-seq といい，転写産物全体（トランスクリプトーム）の配列解析に加えて，発現の定量にも用いられる[3]．前述の DNA マイクロアレイでは，あらかじめ準備されたプローブにハイブリダイズする mRNA しか解析することができない．一方，RNA-seq では遺伝子配列に変異がある場合や，アレル特異的な発現を調べることも可能である．

DNA-seq と同様にマッピングを行い，リードがリファレンスのどこにマップされるかを調べる．しかし，mRNA はエクソンのみが結合された配列を持つが，リファレンスではイントロンも含まれ．したがって，RNA-seq のマッピングは DNA-seq より複雑である．リファレンスと比較することで，変異の有無が分かる．一方，リファレンスの特定の部位にいくつのリードがマップされるかを調べることによって，その部分の発現量を推定できる．この際，マップされたリードの総数がそのまま発現量に対応するのではないことに注意が必要である．長い転写産物ほど多くのリードが生じるので，単位長さ当たりの量に換算するなどの補正が必要である．

3-2-4 その他の測定

ここでは，特定のタンパク質などが結合した部位の DNA 配列を対象とするアプリケーションについて説明する．代表的なものに，クロマチン免疫沈降法（ChIP）を用いる方法がある．ChIP はタンパク質に対応した抗体を用いて，目的とするタンパク質-DNA の結合体を遠心分離して取り出し，DNA とタンパク質の相互作用を調べる方法である．これを用いたアプリケーションを ChIP-seq と呼び，転写因子やメチル化ヒストンの解析に用いられる．

■参考文献

1) 二階堂愛 (編) :“ 次世代シークエンス解析スタンダード, ” 羊土社, 2014.
2) 中村桂子, 松原謙一 (監訳) :“ 細胞生物学　原書第 4 版, ” 南江堂, 2016.
3) 鈴木　穣 (編) :“ NGS アプリケーション RNA-seq 実験ハンドブック, ” 羊土社, 2016.

■S2 群 - 6 編 - 3 章

3-3 プロテオーム解析

（執筆者：福岡　豊）[2018 年 2 月 受領]

多くの種類のタンパク質の発現量を解析（プロテオーム解析）するには，二次元電気泳動と質量分析を組み合わせた方法が用いられている．二次元電気泳動法では，まず等電点の違いを利用して固定化 pH 勾配等電点電気泳動によりサンプル中のタンパク質を分離し，その後 pH 勾配に垂直な方向に分子量に応じてタンパク質を分離する（図 3.1）．電気泳動の前に，あらかじめタンパク質を蛍光ラベリングしておくことによって，得られた情報は蛍光画像に変換される．1 つのタンパク質は等電点と分子量に応じて，蛍光画像上の一定の位置にスポットを形成するので，サンプルに含まれるタンパク質の種類に応じて複数のスポットからなるパターンが得られる（ただし，同一のサンプルを用いて複数回の測定を行ったときに，スポットの位置が微妙にずれることがある）．また，蛍光強度はタンパク質の発現量に対応している．スポットのパターンを比べることによって，あるサンプルで発現量が増減したタンパク質に対するスポットを抽出することができる．しかし，この段階では当該のスポットに含まれるタンパク質の種類を正確に知ることはできない．そこで，ゲル上の当該スポットからタンパク質を含む資料片を切り出し，質量分析などの手法を用いてタンパク質の種類を同定する．

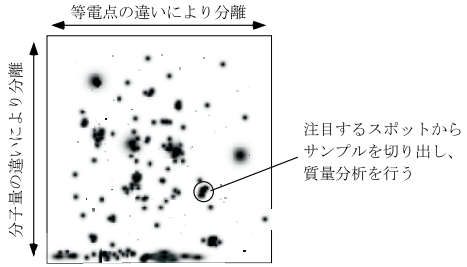

図 3.1　二次元電気泳動

■参考文献

1)　戸田年総：“ プロテオーム解析と疾患解析, ” 実験医学, vol.19, no.11，pp.1443-1448, 2001.