1-22 生命科学研究のための法令・指針

生命科学研究の実施にあたり遵守すべき法令・研究倫理指針

Keyword 遺伝子組換え，カルタヘナ法，個人情報の保護に関する法律，人を対象とする生命科学・医学系研究に関する倫理指針

ヒトを対象とした研究を実施する際，ヒトから採取した試料由来の解析データ・カルテ情報・検査値・質問票などの臨床情報を用いることがある。2015 年改正の『個人情報の保護に関する法律』において「個人情報」の定義が明確化され，一部のゲノムデータが個人識別符号に，また，臨床情報が要配慮個人情報に該当することとなった。そのため，ヒトを対象とした研究を実施する際には，個人情報の保護に留意しつつ，各研究分野に関係する研究倫理指針に準拠した研究活動を実施する必要がある。

各種オーミクス解析技術[▼6-1]や遺伝子組換え技術[▼1-17]の革新的な進歩により，網羅的で精度の高いデータを大量に得られるようになってきた。それに伴い，人の尊厳や人権に関わる生命倫理の問題，遺伝子組換え技術による生物多様性への影響や安全性の問題等が生じるようになってきた。それらの問題の適正化を図るため法令・指針が新たに整備され，国際的な調和を図りつつ，その時々の課題に即して改正・統廃合されてきた。現在までの生命科学研究に関係する法令・指針の状況を**図 1** に示す。

≫遺伝子組換え・生物多様性の保全

遺伝子組換え研究を実施する際は，1993 年に生物多様性の保全や持続可能な利用のための国際的な枠組みとして採択された『生物多様性条約』（日本は 1993 年に締結）に基づいて定められた『遺伝子組換え生物等の使用等の規制による生物の多様性の確保に関する法律（平成 15 年法律第 97 号）』（通称『カルタヘナ法』）に準じ，対象とする生物種によってそれぞれの主務官庁が定める規制措置を講じる必要がある。また，海外の遺伝資源を研究に用いる場合は，遺伝資源へのアクセスと利益分配（Access and Benefit Sharing; ABS）（通称『名古屋議定書』）についても留意すべきである。

≫ヒトを対象とした研究

2021 年 6 月 30 日に施行された医学系研究やヒト由来試料等を用いる生命科学研究を実施する際に準拠すべき『人を対象とする生命科学・医学系研究に関する倫理指針』（以下『生命・医学指針』とする）と，その根拠法となる『個人情報の保護に関する法律（平成 15 年法律第 57 号）』（以下『個情法』とする）について解説する。

≫研究倫理指針

医学研究とヒトゲノム[▼1-16]研究が同時に実施されるケース

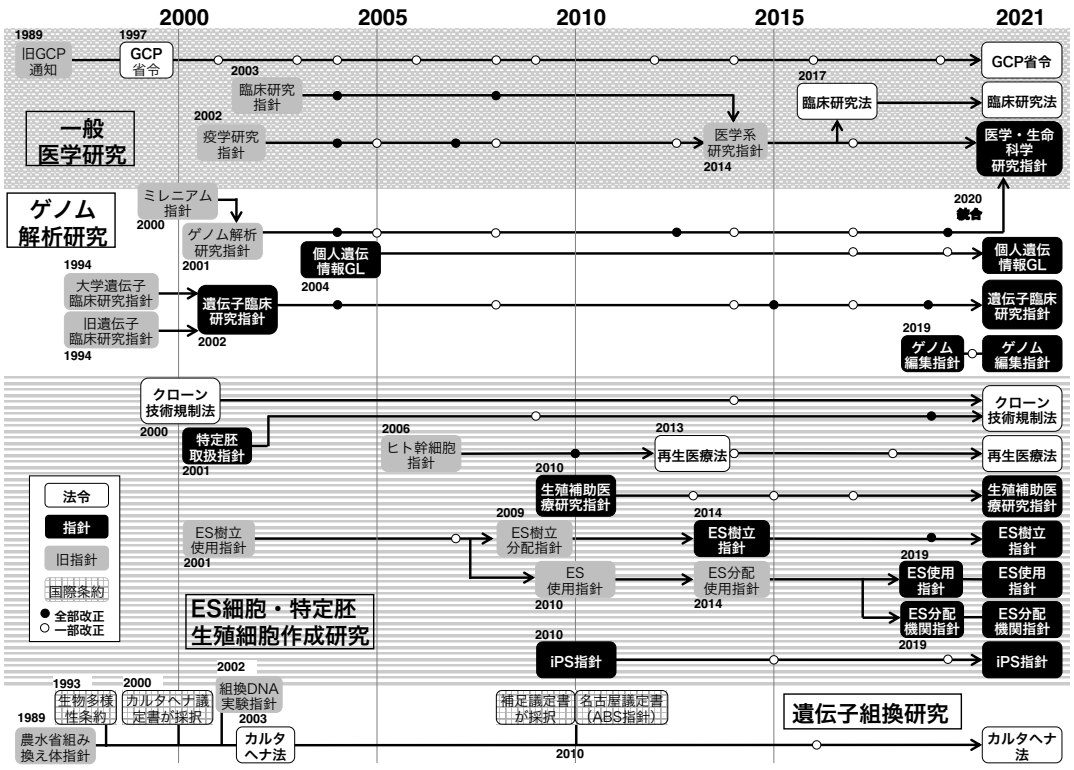


図 1. 生命科学研究に関係する法令・指針

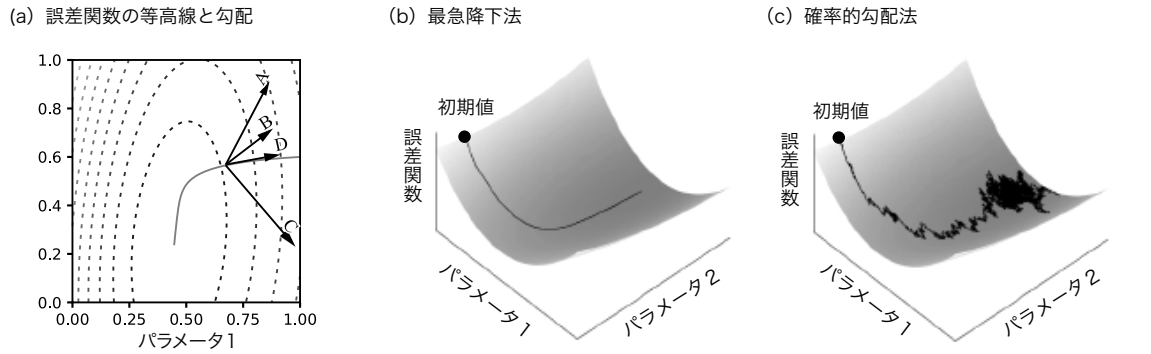


図 2. 誤差関数の曲面と勾配法における解の軌跡

(a)学習データが 3 個で，それぞれの損失の勾配が A, B, C だったとする。3 個の損失の勾配の平均が誤差関数の勾配となる(矢印 D)。(b)最急降下法を実行すると，ボールが転がるように最急降下方向に解が移動する。(c)確率的勾配法では，ランダムな動きをしながらも谷に向かって移動する。

いが，確率的勾配法では学習データ 1 個だけに対して損失関数の勾配を計算すればよいので，1 反復あたりの計算量は約 1/(学習データ数)に抑えられる。

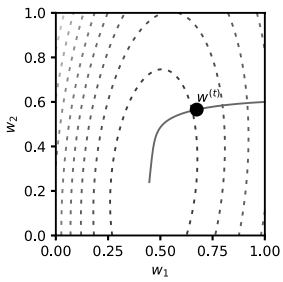
≫ミニバッチ勾配法

最急降下法と確率的勾配法の長所を折衷する方法である。確率的勾配法では各反復で 1 個だけ無作為に学習データを選ぶのに対し，ミニバッチ勾配法ではミニバッチを無作為に選ぶ。ミニバッチとは少数の学習データの集合のことである。ミニバッチ内のデータに対して損失関数の勾配を計算し，その平均で誤差関数の勾配を近似する。確率的勾配法と同じく，その近似勾配の期待値はやはり誤差関数の勾配に一致する。ミニバッチ勾配法の利点は，ミニバッチが大きいほど，近似勾配の分散が小さくなることである。深層学習の場合，GPU のメモリサイズの許す限り大きなミニバッチサイズを設定することが多い。

▼2-3

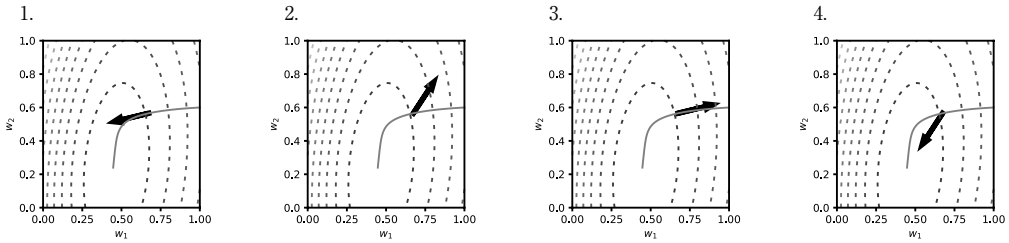
練習問題 **出題▶H31（問 37） 難易度▶B 正解率▶58.2%**

誤差関数 E(w)を最小にするパラメータ $w=(w_1, w_2)$ の値を最急降下法によって求める機械学習問題を考える。図中の点線は等高線を示している。実線は，初期値 $\boldsymbol{w}^{(0)}=(1.00, 0.60)$ から最急降下法を開始し，最小値 $\boldsymbol{w}^{*}=(0.45, 0.24)$ に到達したときの軌跡を表わしている。



最急降下法では，時刻 t において，$\boldsymbol{w}^{(t)}=\boldsymbol{w}^{(t-1)}-\eta\nabla E(\boldsymbol{w}^{(t-1)})$ によって，パラメータの値を更新する。ただし，$\boldsymbol{w}^{(t-1)}$ は時刻 $(t-1)$ におけるパラメータの値，η は正の定数，$\nabla E(\boldsymbol{w}^{(t-1)})$ は誤差関数の勾配を表わす。

ある時刻 t におけるパラメータ値が $w^{(t)}=(0.67, 0.57)$ であった．誤差関数の勾配 $\nabla E(\boldsymbol{w}^{(t-1)})$ としてもっとも適切な方向を選択肢の中から 1 つ選べ。



解説 勾配は，等高線に垂直で，関数の値が増加する向きとなる。よって，正解は選択肢 3 である。

参考文献

1)『確率的最適化』(鈴木大慈著，講談社，2015)
2)『深層学習』(人工知能学会監修，近代科学社，2015)

4-4 構造分類

タンパク質立体構造の分類

Keyword 構造ゲノミクス，構造分類，ファミリー，スーパーフォールド，スーパーファミリー

タンパク質の立体構造解析数が大きく増え始めた 1980 年代後半から，アミノ酸配列の類似性が低く，相同性（進化的な類縁性）はないと考えられていたタンパク質間で，立体構造（フォールド）が類似しているケースが数多く報告されるようになった。やがて統計解析からタンパク質のファミリーは，たかだか 1000 個程度しかないとする説が提唱され，タンパク質立体構造の分類法や構造分類データベースの整備につながった。

≫タンパク質ファミリー 1000 個説

配列相同性の認められないタンパク質間でも，フォールド[4-2]がよく似ている例が多数見つかっている。1992 年に，C・チョーシアは，ある期間内に配列決定されたタンパク質が新規配列ファミリーとなる割合，構造既知タンパク質と相同性がある配列データベース内のタンパク質の割合，構造データベース内の遠縁のファミリーの数から，ファミリーの総数をせいぜい 1000 個程度と見積もった。この「タンパク質ファミリー 1000 個説」が正しければ，フォールドの数も 1000 個以下となるはずであり，全フォールドを構造解析することも不可能ではない。この予想やヒトゲノム計画[1-16]の完了が引き金となって，2000 年代にはタンパク質の網羅的な立体構造解析を目標とした構造ゲノミクス研究が展開された。

≫ファミリー・フォールドの階層性

同一の祖先タンパク質から派生した進化的に関係がある相同なタンパク質をホモログとよぶ[5-8]。ホモログは，お互いに約 30％以上のアミノ酸配列一致度を示し，フォールドがよく保存されている場合が多い。比較的近縁のホモログを集めたグループをファミリーという。ファミリーはアミノ酸配列の類似性[4-8]により，さらに細かなサブファミリーに分けられたり，反対に，いくつかの互いに相同なファミリーが統合され，スーパーファミリーがつくられたりする。前述のように，異なる進化的起源をもつスーパーファミリーが共通のフォールドをもつ例は多い。なかでも非常に多くのスーパーファミリーを含むフォールドのことをスーパーフォールドという。**図 1** には代表的なスーパーフォールドを示した（丸がヘリックス，三角がβストランドを表わし，線はアミノ酸配列上での順番をたどっている）。一般にはフォールドとはタンパク質ドメイン[4-10]に対して定義される。ドメインより小さな構造[4-3]が類似している場合は構造モチーフとよばれる。

≫立体構造分類データベース

タンパク質立体構造を分類したデータベースはいくつかあるが，SCOP（Structural Classification Of Proteins）や CATH（Class, Architecture, Topology, Homology）がその代表である。いずれのデータベースも立体構造を階層的に分類する。SCOP では上位の分類階層からクラス，フォールド，スーパーファミリー，ファミリーに分類される。CATH ではクラス，アーキテクチャー，トポロジー，ホモロジーになる（**図 2**）。SCOP を例にとると，クラスは 2 次構造の種類による分類で all α クラス（主にαヘリックスからできている。**図 1** のグロビンフォールド[5-8]がこれにあたる），all βクラス（主にβシートからできている。免疫グロブリンフォールドがこれにあたり，抗体はこのフォールドの繰り返しでできている[1-12]），$\alpha+\beta$クラス（αヘリックスとβストランドが配列上混在している），α/βクラス（αヘリックスとβストランドが

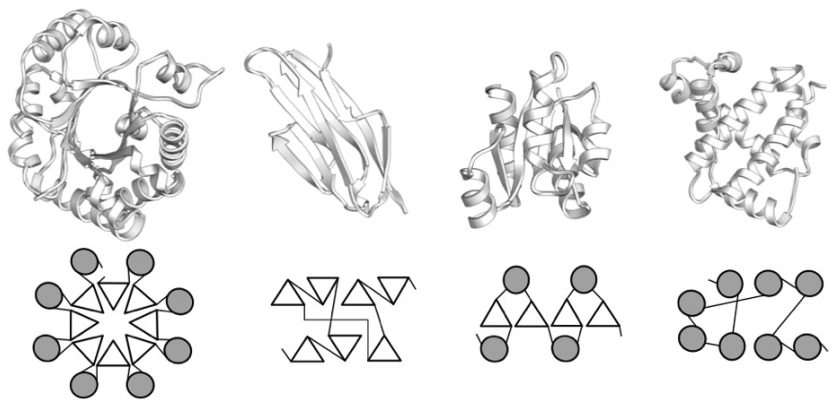


図 1. 代表的なスーパーフォールド

4-10 マップ分析

タンパク質立体構造のマップ分析

Keyword 立体化学，ドメイン，コンタクトマップ，ラマチャンドランマップ

多くのタンパク質は数千～数万の原子からなる複雑な分子であるので，その立体構造のようすを理解するのは難しい。そこで，タンパク質の構造を簡単に把握するために各種のマップ分析法が工夫されている。ここでは分子内の相互作用を解析するためのコンタクトマップと，主鎖構造の概要を把握するためのラマチャンドランマップについて説明する。

≫コンタクトマップ

タンパク質のフォールド[4-2]は，ポリペプチド鎖上の特定のアミノ酸残基同士が接触(コンタクト)することで形成される。そこで**図1**のように，タンパク質内でどのアミノ酸残基が接近しているかコンタクトマップを描いて調べると，フォールドの概要がわかる。コンタクトマップの縦軸と横軸にアミノ酸残基番号を展開し，対応するアミノ酸の組(**図1**の場合はiとj)のC_α原子間距離が一定の値以下の場合に対応するマス目を塗る。しきい値となる原子間距離は5～10Åが目安になる(**図1**では8Å以下を黒，16Å以下を灰色で塗っている)。コンタクトマップからは以下のことを読み取ることができる。黒マスが集中している領域は，アミノ酸残基が密にコンタクトしているので，ドメインとよばれる立体構造[1-12]上ひとまとまりの領域を形成する。図の抗体分子ではN末端とC末端の2つのドメインが存在することがわかる。また2つのドメイン間にはコンタクトがほとんどない(**図1**左の点線内がほとんど塗られていない)ので，これらのドメインは構造上独立性が高いこともわかる。この2つのドメインは構造がよく似ており，免疫グロブリンフォールド[4-4]をもった遺伝子の重複によってできたと推定される。構造上独立性は構造上のドメインの定義の1つであり，コンタクトマップにより容易に同定できる。

このコンタクトマップを配列情報のみから予測することができれば，立体構造予測[4-11]における残基間距離の拘束条件として用いることができる。コンタクトマップ予測にはさまざまな機械学習[2-17]手法が適用され，とくに相同配列から取り出される共進化情報を利用する直接結合分析(Direct-coupling analysis)法の登場によって，その精度は近年飛躍的に発展し，立体構造予測精度の向上に大きく寄与している。

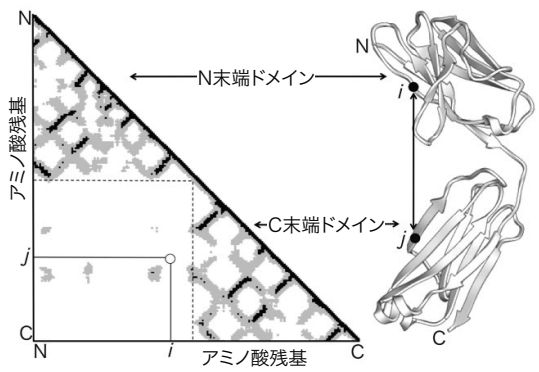


図1. 抗体分子(右)のコンタクトマップ(左)
βシート構造では，逆平行βシートは対角線に垂直に，平行βシートは対角線に平行に，コンタクトの黒線が現われる。

≫ラマチャンドランマップ

タンパク質のフォールドは，ポリペプチド鎖の化学結合の回転によるコンフォメーション変化によって形成される。ポリペプチド鎖ではペプチド結合[4-2]は二重結合性があり自由に回転できないので，C_α原子–アミノ基N原子間の結合の回転角ϕ(ファイ)角とC_α原子–カルボキシ基C原子間の結合の回転角ψ(プサイ)[4-1]角の2つでそのおおよその構造が決定される。R. A. ラマチャンドランによって提案されたラマチャンドランマップは，横軸にϕ角，縦軸にψ角をとり，対応する点にそれぞれのアミノ酸残基をプロットする。このマップは，構造がわかっているタンパク質内のアミノ酸による統計的分布から，もっとも好まれるmost favored領域(90%以上のアミノ酸は

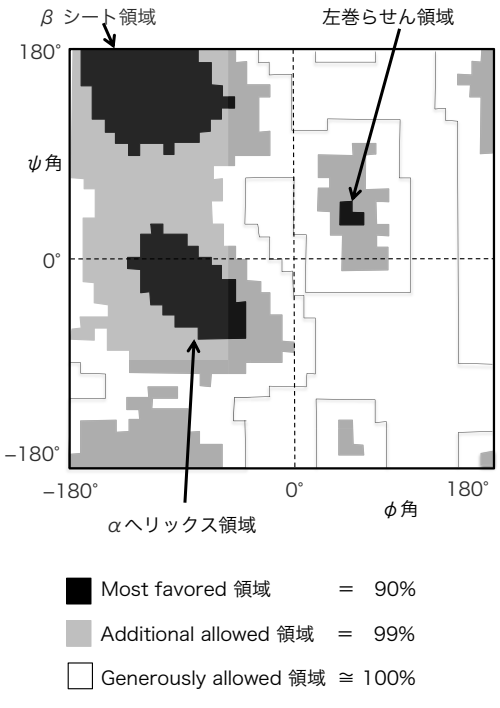


図2. ラマチャンドランマップ

4-11 立体構造予測

タンパク質の立体構造を予測する方法

Keyword 二次構造予測, コンタクトマップ予測, ホモロジーモデリング, フォールド認識, *de novo* 予測法, *ab initio*予測法, 深層学習

計算機によるタンパク質立体構造予測は，立体構造未知タンパク質の機能発現のメカニズムを知る上で重要である．立体構造予測手法には，データベースに登録されている他の立体構造を利用するホモロジーモデリング法やフォールド認識法，既知構造に依存しない *de novo*（デノボ）予測法，*ab initio*（アブイニシオ）予測法などがある。近年では，深層学習の応用が進み，*de novo* 予測法をベースにさまざまな手法を統合した手法が成功を収めている。古くから取り組まれてきた立体構造上のさまざまな特徴の予測とあわせて紹介する。

≫配列・構造特徴の予測

タンパク質の配列・立体構造中に見出される特徴的な部分配列や構造は，立体構造予測に役立つさまざまな情報を提供する。二次構造予測法は，アミノ酸配列中で二次構造[1-9]（αヘリックス，βストランド，それ以外）[4-9]をとるアミノ酸残基を予測する。天然変性領域予測は単量体としては特定の構造をとらない残基を予測する。また，コンタクトマップ[4-10]予測は，配列を構成するどのアミノ酸残基同士が接触するかを予測する。とりわけ相同配列から取り出される共進化情報を利用する直接結合分析（direct-coupling analysis）法により，その精度は近年飛躍的に向上した。

≫鋳型（テンプレート）を用いた立体構造予測

立体構造未知のタンパク質のアミノ酸配列（標的）と有意な配列類似性をもつ構造既知タンパク質がデータベース中に存在すれば，その既知構造を鋳型として，予測標的配列の構造予測を行なうことができる．鋳型構造を用いる予測[3-2]では，鋳型配列と予測標的配列との正確なアラインメントを必要とする。

≫比較モデリング法（comparative modeling）・ホモロジーモデリング法（homology modeling）

相同なタンパク質のフォールドはアミノ酸配列より保存される[4-8]傾向にある。比較モデリング法，またはホモロジーモデリング法は，この性質を利用する。まず，標的配列の類似配列の中から立体構造既知の配列を選び，この配列の立体構造を鋳型とする。次に，鋳型と標的配列のアラインメントを参照しながら，鋳型構造のアミノ酸を標的のアミノ酸に置換する。鋳型に対して標的側への欠失または挿入と見なされる部位は，それぞれ除去あるいは挿入してモデリングする。最後に，原子間の衝突や化学結合のゆがみを分子力学計算[4-12]等で排除し，標的の立体構造を予測する（**図 1** 上）。ホモロジーモデリング法では，標的と鋳型の配列類似度が高いほど予測精度は高い。配列一致度 20% 以上が，鋳型選択時の目安である。

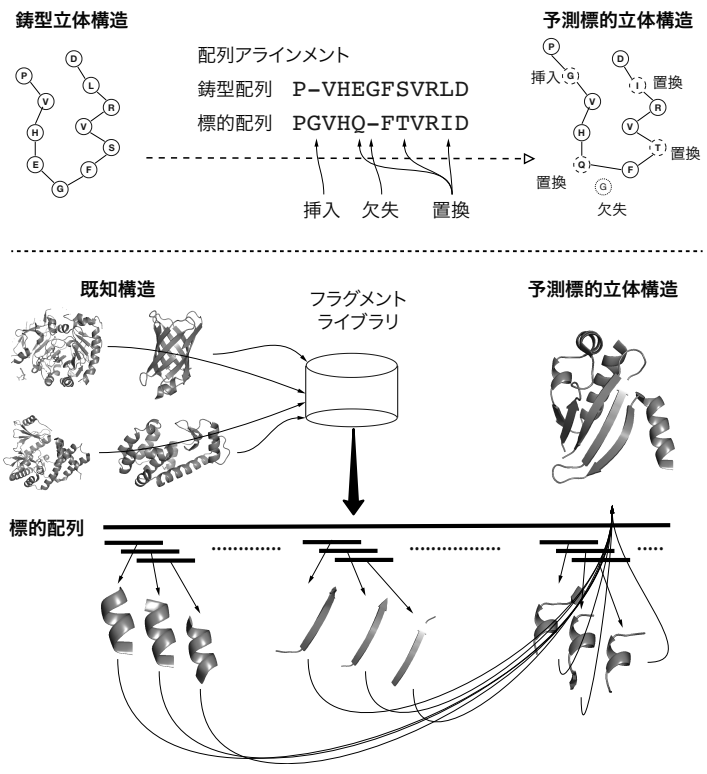


図 1. ホモロジーモデリング法（上），フラグメントアセンブリ法（下）の概略

≫フォールド認識法（fold recognition）

鋳型とする立体構造に相同性を仮定しない手法も存在する。データベース中には配列間の類似性が微弱であるにもかかわらず，立体構造が酷似している例が多く存在する．また，新規フォールドの報告数は年々減少し，自然界に存在するタンパク質フォールド数[4-4]は有限であると考えられている。これらの事実をもとに，フォールド認識法では標的配列を報告されているフォールドのいずれかに当てはめて立体構造を予測する。この方法は，アミノ酸配列をアラインメントするのと同様に立体構造（3D）とアミノ酸配列（1D）をアラインメントするので 3D-1D 法，あるいは，アミノ酸配列を糸（スレッド）として立体構造を裁縫するイメージから，スレッディング法とよばれる。

4-14 電子顕微鏡のインフォマティクス

電子顕微鏡のインフォマティクス

Keyword クライオ電子顕微鏡，単粒子解析，3 次元再構成，フィッティング法，デノボモデリング法

生体高分子の原子レベルの立体構造を決定する第 3 の手法として，クライオ電子顕微鏡を用いた単粒子解析が急速に普及してきた。この手法は，膨大な量の二次元画像を計測・収集し，それらから 3D マップを計算，マップにあてはまるように原子モデルを構築する。X 線や NMR の実験データは，そもそも多数分子の集団から得られた平均信号であるのに対し，単粒子解析の実験データは，1 分子ごとの個別の画像であり，大量の 1 分子画像の統計処理で分子集団の構造情報が抽出される。そのため，必要なデータ量，計算量は他の手法に比べ大きく，情報処理技術に大きく依存する。

≫クライオ電子顕微鏡による単粒子解析の手続き

図 1 に電顕の単粒子解析(single particle analysis)の計測と画像処理の概要を示した。クライオ透過型電子顕微鏡で観察する試料は，多数の高分子を含んだ非晶質(アモルファス)の氷の薄い膜である。高分子が溶けた水溶液を，穴が開いたカーボン膜にのせて急速冷却することで作成する。この試料は真空や電子線照射による損傷を受けにくい。結晶状態と異なり，分子の向きはそろっていないことに注意。最新の電子顕微鏡では，直接電子検出器を用いて，多数の画像を動画として撮影し，動画から電子線照射による試料の動きを補正して，ぶれのない 1 枚の画像を得る計算を行なう(動画補正)。次に，電顕画像に多数写っている分子の粒子のそれぞれの位置を同定する(ピッキング)。まず，数枚の画像に対して手作業でピッキングすることが多い。その後，各粒子画像を切り出し，画像の類似性からグループに分類する(2D 画像分類)。同じ向きの分子が写っているグループの平均画像を計算すると，ホワイトノイズが相殺され，明瞭な分子像が得られる。よって，各グループの平均画像から，そのグループが実際の分子に相当するか(○)，ノイズや不純物によるものか(×)判断できる。ここで○の平均画像を鋳型として，鋳型ベースの自動ピッキングを多数の電顕画像に対して行なうと，良質な粒子画像を多数抽出できる。1000〜100 万個ほどの粒子を用意するのが普通である。この後，後述する 3 次元再構成で 3D マップを推定する。こうした画像処理用のソフトとして，Relion と cryoSPARC がよく使われる。

≫ 3 次元再構成の計算の手続き

多数の 2D 粒子画像とその投影方向がわかれば，3 次元再構成(3D reconstruction)という計算を行なうことで，3D マップの推定ができる。実空間で行なう重み付き逆投影法と周波数空間で行なうフーリエ再構成法がある(**図 2**)。フーリエ再構成法は，投影切断面定理(projection slice theorem)に基づく。この定理は，「2D 投影画像のフーリエ変換は，3D マップのフーリエ変換の投影方向に直交する切断面に等しい」ことを保証する(**図 2**)。よって，十分な種類の投影方向の投影画像があれば，周波数空間の 3D マップを埋めつくすことができ，それを逆フーリエ変換すれば，実空間の 3D マップを得ることができるはずである。

しかし，単粒子解析では各粒子の投影方向はわからない。そこで，**図 3** のような反復的な算法で計算する。まず，何らかの初期 3D マップ(事前知識がなければランダム)を用意して，10^3–10^4 通りの投影方向に対し，そのマップの投影 2D 画像群を生成，2D 粒子画像と比較することで，投影方向を推定する(E- ステップ)。次に，推定された投影方向と粒子画像を使って 3D マップを更新する(M- ステップ)。フーリエ再構成法を用いる場合

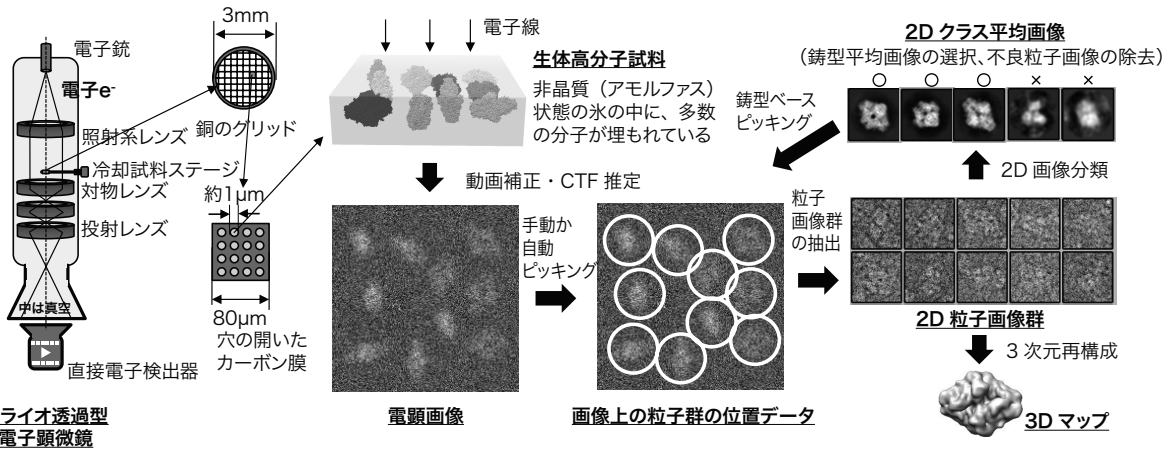


図 1. クライオ電子顕微鏡による単粒子解析の手続き

が多い。この2つのステップを収束するまで反復する。この解法は，最尤法，あるいは最大事後確率推定[▼2-14]の形式であり，投影方向は潜在変数になる。[2D 粒子画像]＝[投影演算子][3D マップ]＋[ノイズ]の生成モデルを用いる。最大事後確率推定では，3D マップの事前確率も考慮される。最初の 3D マップは，乱択した少数の粒子を用いる，確率的勾配降下法を用いて作成される。そのマップを基に，画素数，粒子数を増やし，より詳細なマップへ精密化を行なう。1 組の粒子画像群から数個の 3D マップを同時に再構成する計算法も実装されている。

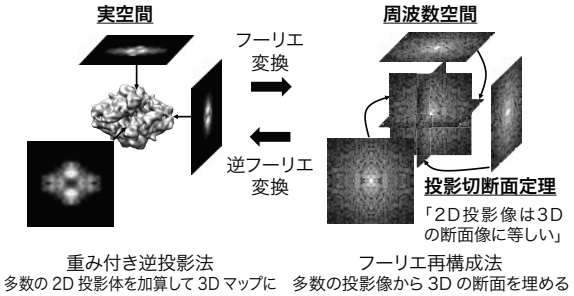


図 2. 実空間と周波数空間の 3 次元再構成

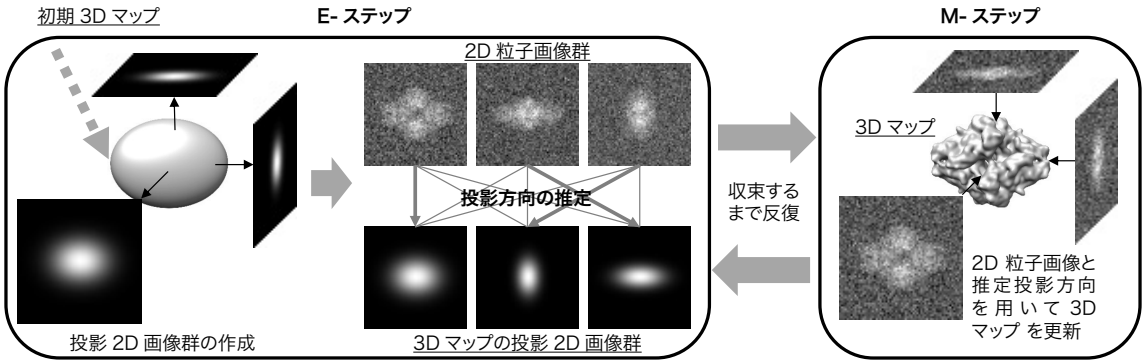


図 3. 典型的な 3 次元再構成のアルゴリズム

≫電顕 3D マップに基づく原子モデリング

フィッティング法とデノボモデリング法の 2 つがある。フィッティング法では，X 線など他の実験手法[▼1-21]や構造予測[▼4-11]による原子モデルを 3D マップに重ねる。デノボモデリング法では，既存の原子モデルを使わずに，最初から原子モデルを構築する。二次構造要素，アミノ酸などの位置をまず粗く同定し，それらを 1 本の配列に接続，さらにアミノ酸配列をあてはめる計算を行なう。フィッティング[▼1-21]より高い分解能が要求され，モデリングする領域をドメイン程度に限定する必要がある。最後に，手修正を行ない，精密化計算を行なうことで，よりよくマップに適合するよう原子モデルを局所変形させる。これら全過程において，手作業が多いが，主観性を排除するため，全自動モデリングの手法開発も試みられている。

電顕で決定された原子モデルを使用する場合，鵜呑みにせず，元にした 3D マップの分解能の値を確認，マップと原子モデルとの重なりを分子ビューアで視認すべきである。一般に分解能が 4 Å より悪い場合，側鎖や結合低分子の正確な構造は期待できない。分解能が悪くてもフィッティング法などで一見詳細なモデルが置いてある場合があるので注意を要する。また，1 つの 3D マップ内の場所による分解能の差が大きい場合も多い。全体の分解能が良くても，柔軟に動く部分が不鮮明なマップになり，その部分の原子モデルの精度が下がることがある。

練習問題 出題▶未出題　難易度▶B　正解率▶ー %

電子顕微鏡の単粒子解析(single particle analysis)に関する以下の記述において，もっとも不適切なものを選択肢の中から 1 つ選べ。

1. 単粒子解析は粒子 1 つ 1 つの 2 次元画像を入力データとして用いる。
2. 単粒子解析において多数の 2 次元粒子画像が必要なのは，平均化によりノイズを相殺するためである。
3. 単粒子解析において多数の 2 次元粒子画像が必要なのは，さまざまな方向の投影像を得るためである。
4. 単粒子解析により，それぞれの粒子ごとに別々の 3D マップを得ることができる。

解説 単粒子解析では多数の 2D 粒子像の統計処理で 1 つの 3D マップを得る。したがって選択肢 4 が正解。試料を傾けながら多数枚の画像を撮影する電子線トモグラフィー法では，粒子ごとの 3D マップを得ることができる。

参考文献

1)「クライオ電子顕微鏡で見えた生命のかたちとしくみ」(佐藤主税他著，『実験医学』5 月号，Vol. 36，No. 8，2018)

4-15 タンパク質デザイン

望みの立体構造や機能を有するタンパク質をデザインする方法

Keyword タンパク質デザイン，タンパク質工学，人工進化

望みの立体構造や機能を有する新しいタンパク質分子を人工的に作成する試みはタンパク質のデザインとよばれる。デザインにおける一連のプロセスは，主鎖構造設計→配列設計→配列最適化という３つの段階に分けられる。どこを出発点とするかは目的によって異なり，たとえば，まったく新しい立体構造を持つタンパク質の設計を目的とするなら主鎖構造の設計が出発点となり，既知のタンパク質の改良を試みる場合は配列最適化のみ行なう場合も多い（**図 1**）。ここでは，主鎖構造設計→配列設計→配列最適化という一連のプロセスに従い，その要素技術を概観する。

≫主鎖構造のデザイン

タンパク質のデザインと聞くと，アミノ酸残基[▼1-8]をどのような順につなげていくかという問題(配列の設計)に注目しがちである。しかし本来は，アミノ酸配列がどのように折りたたまれるのか，目的となる立体構造を先に設定する必要があり(構造の設計)，配列はその立体構造に適したものとなるように設計される。主鎖構造の設計は後に続く配列の設計に大きな影響を与えるため非常に重要である。

≫鋳型構造を用いた主鎖構造デザイン

目的構造の設定において，既知のタンパク質構造を設計の出発点とする場合がある。たとえば，天然タンパク質の構造を部分的に改変しようとする場合や，ターゲット分子の表面形状と相補的な構造をもつ既知構造を鋳型として利用する場合などである。前者はタンパク質の物理化学的特性をコントロールすることを目的とし，後者は特定のターゲット分子表面[▼4-9]へ結合してその活性を阻害するタンパク質の設計を目的することが多い。

≫鋳型構造を用いない主鎖構造デザイン

鋳型として既知構造[▼4-11]を設定しない場合，ゼロから立体構造を構築する必要がある。このように，天然タンパク質を鋳型とせずに新規人工タンパク質を設計する場合は特に *de novo* デザインと呼ばれる。*de novo* デザインでは，まずターゲットとする構造のおおまかな形状(フォールド[▼4-2])を決める必要がある。想定し得るあらゆるフォールドが必ずしもタンパク質分子として実現可能とは限らないため，フォールドの設定には注意を払う必要がある。一般的に，α-バンドル構造のような天然に頻出するパターンが選択されることが多い。

とくに対称性の高いフォールドを設定した場合，主鎖構造の構築は二次構造間[▼1-9]の距離や角度など少数のパラメータを変化させることで機械的に生成できる。この種の方法はパラメトリックデザインと呼ばれ，コイルドコイルや繰り返し構造を網羅的に生成する際に使われる。

一方，より汎用的な構造構築方法としては，立体構造予測[▼4-11]の分野と同様，フラグメントアセンブリ(FA)法が使われる。天然構造由来の断片構造を組み合わせて全体を構築することで，設計構造に含まれる局所的な構造要素の妥当性が保証される。FA 法は柔軟な構造構築が可能であるため，たとえば天然に観測されていないフォールドをもつタンパク質でも設計することが可能であり，正しく折りたたんだ成功例も実際に報告されている。

≫アミノ酸配列のデザイン

目的とする立体構造が設定された後，次に行なうのは配列のデザイン。つまり，目的構造を安定化するアミノ酸配列を設計することである。配列デザインでは，側鎖を含むモデルの作成と安定性の評価を繰り返し行なうことで，目的構造にもっとも適した配列を探索する。モデ

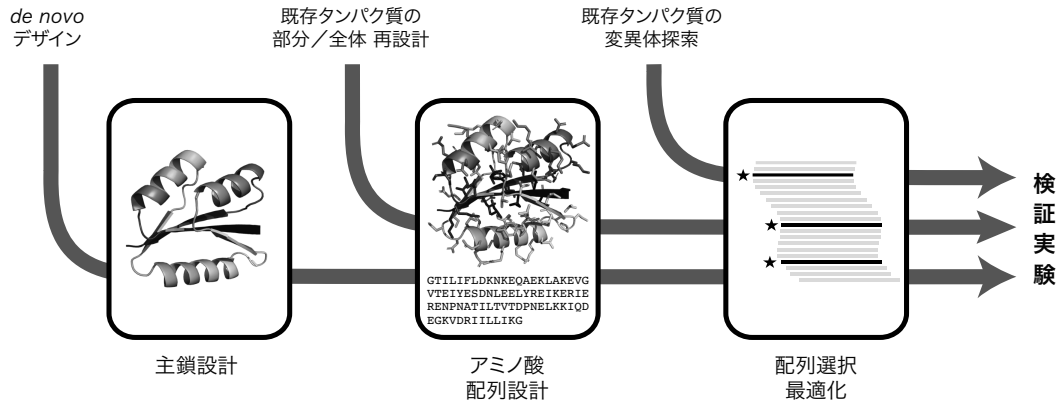


図 1. タンパク質デザインにおける，主鎖設計・アミノ酸配列設計・配列選択と最適化，という３つの段階の概念図
各段階のどこからデザインをスタートさせるべきかは目的に応じて異なる。

6-2 サンガー法・次世代シークエンシング

サンガー法および次世代・超並列シークエンシング技術

Keyword シークエンサ，リード，アセンブリ，N50，マッピング

DNA の塩基配列を読み取る装置を DNA シークエンサ，あるいは単にシークエンサとよぶ。20 世紀はサンガー法による読み取りが主流だったが，2007 年頃に次世代シークエンシングあるいは超並列シークエンシングとよばれる技術が普及した。この手法は遺伝子発現量を測定する DNA マイクロアレイ（DNA チップ）を代替しつつあり，現在も新技術が発表されている。

≫サンガー型シークエンシング

20 世紀のシークエンサはサンガー型とよばれ，1980 年にノーベル化学賞を受けたサンガー（Frederick Sanger）の配列決定法に基づいている（**図 1**）。サンガー法では，鋳型となる DNA[1-5] から DNA ポリメラーゼによって相補鎖を作る際に，基質となる 4 種のデオキシ塩基にジデオキシ塩基[1-7]を 1% ほど加える（デオキシ塩基における 3 位の水酸基が欠けたもので ddNTP と書く。N は ATGC の 4 種類）。加える割合は読み取る塩基長に応じて調節する。伸長の際に ddNTP が取り込まれるとそれ以上伸びず，取り込まれた位置で相補鎖の合成がストップする。たとえば ddATP を加えた場合はアデニンの部位で，ddCTP を加えればシトシンの部位で止まった，さまざまな長さの相補鎖が生じる。

そこで放射性同位元素 ^{32}P 入りの塩基と ddNTP を一種類（ATGC の 4 通り）だけ混ぜて作成した相補鎖を平板型のアクリルアミドゲル電気泳動[1-18] 4 レーンにかけ，泳動位置をオートラジオグラフィーで判定すると ATGC の位置関係がわかる。ヒトゲノム計画が始まった 1990 年頃までは，サンガー法で 300 塩基を読むのにまるまる一日かかっていた。

国際ヒトゲノム計画をきっかけに，多くの作業が自動化・高速化された。まず放射性同位元素の利用は廃止して 4 種の ddNTP を 4 色の蛍光色素で標識した。さらにゲル板の側面からレーザー照射して得た蛍光を，ラインセンサーで判定するようにした。最後に平板ゲルではなくキャピラリーゲル電気泳動[1-19]を採用し，泳動時間を短縮した。複数のキャピラリーから溶出する塩基を迅速かつ正確に判定する技術には日本が大きく貢献している。

≫次世代シークエンシング・超並列シークエンシング

サンガー型シークエンサの欠点は，鋳型となる DNA[1-17] を大腸菌等の微生物を用いて個別に増幅（クローニング）する手間にあった。PCR 産物を直接読み取ることも可能だが，1 配列に 1 キャピラリーを要する。この欠点を解消したのが 2007 年頃から普及した次世代シークエンシング（NGS: Next Generation Sequencing）技術である。

NGS の特徴は反応のミニチュア化にある。エマルジョンとよばれる小さな油滴や，高密度に配置した反応セル毎に鋳型 DNA をトラップし，超並列で PCR 増幅と配列決定を実施する。並列化により 1 日あたりの読み取り塩基数は百ギガ超にまで向上した。各社[1-4]で読み取り原理は異なるが，多くは DNA ポリメラーゼによる伸長反応のシグナルを読み取る。その手法は，4 種の塩基が取り込まれる際に 4 種の異なる蛍光標識を出すようにして CCD カメラで見分けるものと，4 種の塩基を順番に適用して伸長する反応を半導体センサーで検出するものに大別される。

ショートリードの NGS 技術は 100 塩基から 200 塩基

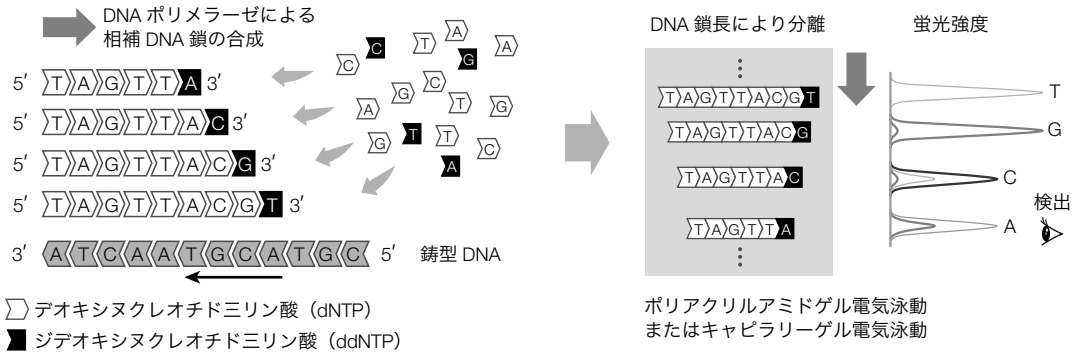


図 1. サンガー法の原理

鋳型となる DNA（図では 5′-CGTACGTAACTA-3′）に相補的な塩基を作成する際に、ddNTP（図では黒いブロック）によってさまざまな長さの相補鎖を作成する。4 種の ddNTP は異なる蛍光色素で標識してあり，どの塩基で停止しているかがわかる。電気泳動では短い配列ほど速く流れるため溶出する順番を記録すれば配列を決定できる。検出される蛍光にはノイズや重複もあり，そこから塩基を特定する作業をベースコールという。ベースコールの信頼性はクオリティスコア Q で記載される。

6-3 DNA マイクロアレイ・NGS の応用解析

DNA マイクロアレイと次世代シークエンシング技術による応用解析

Keyword DNA マイクロアレイ，RNA-Seq, Hi-C, 一細胞計測

NGS が普及する以前，遺伝子発現量は DNA マイクロアレイを用いて測定されていた。NGS が普及してからは，Seq 解析とよばれるさまざまな手法が生まれている。基本は mRNA を逆転写して読み出す RNA-Seq だが，タンパク質が結合あるいはアクセスできる部分だけを取り出す ChIP-Seq など幅広く実施されている。また微小流路を用いた測定系のミニチュア化が進み，一細胞計測も可能になった。

≫ DNA マイクロアレイ（または DNA チップ）

DNA マイクロアレイは，ガラスあるいはシリコン基板上にプローブとよばれる短い核酸配列を稠密に配置した測定装置である。細胞から抽出した mRNA を逆転写して cDNA に変換した試料とアレイをハイブリダイゼーションさせることで，遺伝子の発現量を測定する（**図1**）。国際ヒトゲノム計画が佳境に入った 2000 年前後，DNA マイクロアレイは遺伝子の発現量を測定する際の第一選択肢であった（それ以前はリアルタイム PCR やブロッティング[1-19]を用いた）。

次世代シークエンシング（NGS）[6-2]装置が普及すると，DNA マイクロアレイを未知配列の決定や検出に使うことはほぼ無くなった。アレイはおもに，ヒトを含むモデル生物における遺伝子の発現量（トランスクリプトーム解析）[6-1]や多型[1-16]を判別する目的で利用される。

重要な応用先は医療検査や病気の診断である。ヒトの遺伝子型判別（ジェノタイピング）[5-5]のように確認すべき部位が定まっている場合，安定して迅速かつ低コストに定量・判別できるアレイ技術は，NGS に優る選択肢である。ヒト向けには一塩基多型（SNP）やコピー数多型（CNV）[1-16]を検出するアレイのみならず特定の疾患（たとえばがん）に特化したものなど，広く実用化されている。

≫ RNA-Seq 解析

RNA-Seq[1-18] とは，細胞から抽出した mRNA[1-6] をもとに作成した NGS リードをリファレンス配列にマップして遺伝子領域ごとにカウントする解析法である。リード数は 100 万単位で正規化し，RPKM（Read Per Kilobase per Million mapped reads）あるいは FPKM（Fragment PKM）という単位で集計する。この作業には NGS リードだけでなくゲノム上の遺伝子領域の情報が必要になる。そこで NGS 結果の SAM データ[6-2]とあわせ，GFF（General Feature Format）[1-6]とよばれるエキソンや開始コドン，保存配列などの特徴量（feature）を記載したタブ区切りファイルを利用する。GFF の拡張版が GFF2, GFF3，あるいは GTF である。モデル生物の GFF は Ensembl Genomes やカリフォルニア大学サンタクルーズ校（UCSC）ゲノムブラウザ等から取得できる。

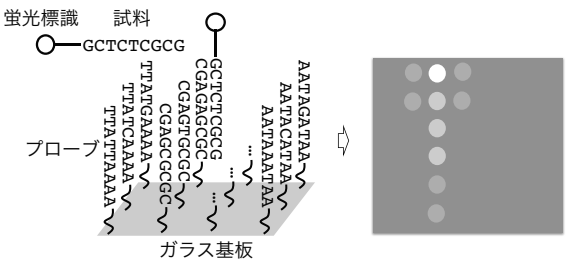


図 1. DNA マイクロアレイ

DNA マイクロアレイでは基板上に短い DNA 配列（プローブ）を位置を特定できるようスポット状に固定する（図左）。測定にはまず試料となる DNA を蛍光物質でラベルする。試料 DNA を相補的なプローブと結合させ，蛍光の強度で定量する（図右）。プローブの塩基長はアレイにより異なる。SNP アレイや DNA メチル化を検出するメチル化アレイでは 1 塩基のちがいでも検出できる

≫さまざまな Seq 解析

NGS 解析には，mRNA の末尾にあるポリ A 鎖[1-6]領域を欠いたもののみを選別して NGS を実施するノンコーディング RNA（ncRNA）-Seq や，ゲノムをバイサルファイト処理[1-20]して非メチル化シトシンをウラシルに変換してから実施するバイサルファイト -Seq がある。

転写制御因子のような DNA 結合タンパク質に対する抗体を用いて，当該タンパク質が結合するゲノム領域だけを切り出して濃縮して解析する手法を ChIP-Seq とよぶ。ChIP とはクロマチン免疫沈降（Chromatin Immuno Precipitation）の略であり，DNA チップの全盛期には ChIP-chip 解析が盛んであった。

医療分野ではヒトゲノムにおけるエキソン[1-5]部分のみに注目したエクソーム -Seq も実施される（エキソームとも書かれる）。ヒトゲノムは全長が既知のため，特別に設計したプライマー集合を用意すれば遺伝子疾患に関連しそうな部分だけを濃縮できる。

さまざまな Seq 解析の名前は，NGS データの公共リポジトリ SRA[6-2] に登録する際のライブラリ名として標準化されている。国内では DDBJ[3-1] による SRA ハンドブックにそのリストがある。中でも数が多いのは，NGS 解析を手がかりにクロマチン構造を探る手法である。

≫クロマチン構造の解析

核内において，DNA は小さく折りたたまれてクロマチン繊維[1-1]を構成する。この構造は発生過程や転写機構に応じて動的に変化（リモデリング）[1-11]している。ヌクレオソームがほどけて露出した部位をオープンクロマチン領域とよび，遺伝子発現を制御する領域と考えられているた

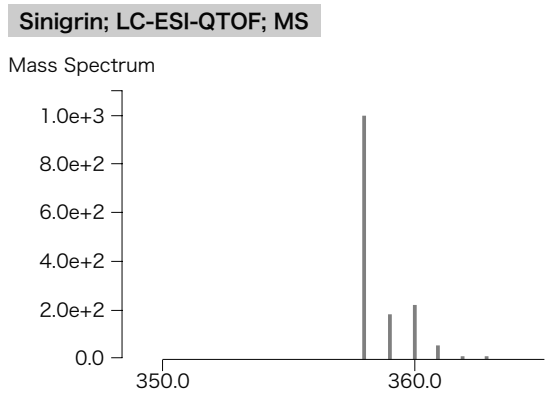


図 1. アブラナ科に特有の代謝物群であるグルコシノレートの一種，シニグリンのマススペクトル(整形後)
縦軸がイオン量(強度)，横軸が m/z になる。シニグリンの組成は $C_{10}H_{16}NO_9S_2$(カリウムが外れた負イオン)で硫黄を 2 つ含むため，質量 358 のピーク(前駆イオン)の隣に同位体 ^{13}C に由来する 359 と ^{34}S に由来する 360 のピークが現れる。

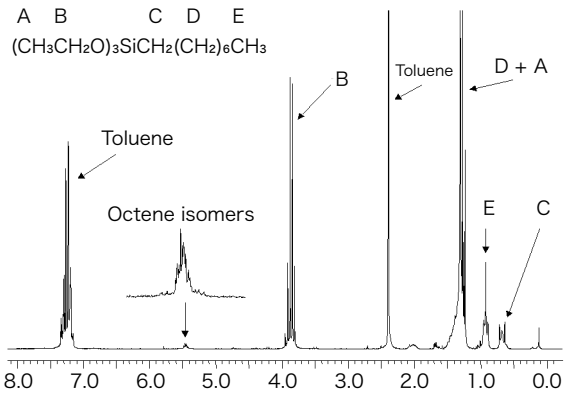


図 2. プロトンの核磁気共鳴を観察した ^{1}H-NMR のスペクトルの例
横軸が化学シフトで縦軸が相対強度。分子構造の中に含まれる水素の種類に対応したピークが生じる。

化合物同定を終えた結果は，物質名と測定量を列挙したデータ行列になる。多くの場合は多変量解析を用いて代謝の動きと表現型[▼1-14]あるいは遺伝型などの関係を見出す。クラスター解析や主成分分析，回帰分析に加えて，PLS(Projection to Latent Structures あるいは Partial Least Squares)も多用される。[▼2-20]

≫ NMR メタボロミクス

NMR 法は正確にいうと NMR 分光法(spectroscopy)である。測定するのは分子中の原子核が共鳴する周波数だが，周波数自体は磁場の強さに依存するためテトラメチルシラン(TMS)が出す基準周波数からのずれ(化学シフト)をスペクトルとして表現する。NMR 装置に書かれる 600 MHz や 900 MHz という値が基準周波数である。

化学シフトは原子の化学結合状態で決まる。よってこの値から近接する分子構造を立体配置を含めて同定できる。一次元 NMR スペクトルでは横軸に化学シフトとその相対強度を縦軸にとり，二次元 NMR スペクトルでは縦軸と横軸にそれぞれ異なる化学シフトを配置する。NMR スペクトルの単位に Hz ではなく ppm を用いるのは基準からのずれ(シフト)を百万分率(parts per million)で表わすためである。これにより測定磁場の強さに依存しないスペクトルとなる。

NMR 法の特徴は測定試料を破壊しない非侵襲法であること，そのため複雑なサンプル調製も不要のまま連続測定でき，再現性が高いことである。ただし質量分析に比べると感度は劣り，試料はマイクロモル量が必要になる(質量分析はナノモル量で済む)。また強磁場の測定装置ほど分解能が上がるが，超電導コイルを液体ヘリウムで冷やす装置およびその維持費は高価になる。

測定したスペクトルをデータベースに照会して物質同定するところは質量分析と共通である。しかし NMR のスペクトルデータベースに収載される物質数は 1000 程度にとどまっておりスケールアップが望まれている。

練習問題 **出題▶H30（問 77） 難易度▶A 正解率▶36.3%**

血液中や尿などの生体サンプル内の代謝物を網羅的に測定するメタボローム解析について，もっとも不適切なものを選択肢の中から 1 つ選べ。

1. 核磁気共鳴(NMR)を用いた解析は，測定のスループットが高く，多くのサンプルを測定するのに適している。
2. 液体クロマトグラフィー質量分析(LC/MS)による解析は，極性が比較的高い成分の分析に有効である。
3. ガスクロマトグラフィー質量分析(GC/MS)による解析は，ガス状または気化する成分の分析に有効である。
4. NMR は LC/MS に比べて感度が高く，より少ないサンプルで濃度の低い化合物を測定することができる。

解説 国際的な化学連合 IUPAC の推奨では，手法としての質量分析(mass spectrometry)を MS と略し，装置(mass spectrometer)を MS と単独では書かない。同様に NMR は現象を表わし，測定機械そのものは NMR 装置と記載する。NMR 法は質量分析に比べると感度は高くない。そのため必要なサンプル量は NMR 法 のほうが多くなる。

参考文献

1)『メタボロミクス実践ガイド』(馬場健史ほか編集，実験医学別冊，羊土社，2021)

6-5 プロテオーム解析

プロテオーム解析・プロテオミクス

Keyword 質量分析，二次元ゲル電気泳動，FDR，フラグメンテーション，ショットガン

プロテオミクスという用語は従来，タンパク質を二次元ゲル電気泳動で分画してペプチド配列を読む手法や，質量分析を用いてタンパク質を個別に同定する手法を意味してきた。近年はショットガン法による網羅的な解析から立体構造解析も含んだ，より幅広い意味で用いられる。とりわけヒトや酵母に関しては多くの情報が公開・共有されており，国際ヒトプロテオーム機構（HUPO）ウェブサイトのリソース欄から主要情報にアクセスできる。

≫プロテオーム解析の変遷

ヒトゲノムにある遺伝子の数はおよそ2万である[▼1-16]。そこから選択的スプライシング[▼1-5]等を経て生成されるタンパク質は，組織のちがいなどを総合すれば数十万は存在する（UniProt データベース[▼3-1]にはヒト・アミノ酸配列が19万以上ある）。また，細胞内で働くタンパク質は翻訳後にさまざまな修飾を受ける[▼1-11]。その多くは，切断，メチル化・リン酸化，脂質や糖鎖の付加だが，列挙していくと修飾の多様性は数十に及ぶ。機能するタンパク質の量とその鋳型となる mRNA の量は必ずしも相関しないため，タンパク質の定量は重要である。

従来型のプロテオーム解析では細胞から抽出した全タンパク質を二次元ゲル電気泳動[▼1-19]で展開し，解析したい部分をトリプシン等のプロテアーゼでペプチドに分解して質量分析する。古くはエドマン分解によりアミノ酸の並びを決めた。得られたマススペクトルは，ゲノムから理論的に予測されるペプチドのデータベースと照合し，もとになる遺伝子を同定する。この手法は指紋照合にちなんでペプチドマスフィンガープリンティング（PMF）法とよばれる。

PMF 法によるタンパク質同定は，今でもヒト病態の解析などで実施される。疾患例とそうでない例など2種のプロテオーム試料をそれぞれ蛍光色素でラベルし，泳動結果を比較して差の大きな部分を解析すれば，疾患に関連するタンパク質がわかる。

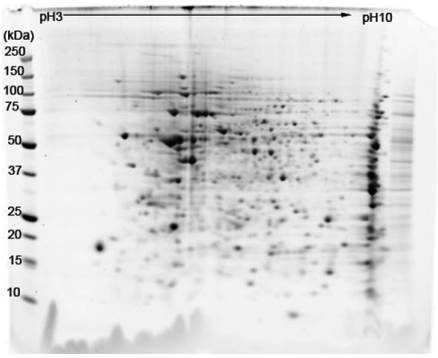


図 1. マウス精巣タンパク質の二次元電気泳動結果
等電点による分離が上部にあたり，軽いタンパク質ほど速く下方に移動する。結果はタンパク質を蛍光染色してからレーザースキャナでデジタル化する。医薬基盤・健康・栄養研究所疾患モデル小動物研究室の許可を得て掲載。

より基礎的な分野では電気泳動で分離せずタンパク質を混合物のまま分解，測定するショットガンプロテオミクスが主流になっている（ノンターゲット分析[▼6-1]の1つ）。混合物のままプロテアーゼ処理したペプチドを必要に応じてイオンクロマトグラフ等で分画し，液体逆相クロマトグラフ質量分析計（LC-MS/MS）にかける。タンパク質のままでは大きさや疎水性もまちまちだが，ペプチドにすることで性質を均質化し，統一された分析法をハイスループットに適用できる。リン酸化[▼1-11]されているペプチドだけを二酸化チタンビーズを用いて選択するなど，特定のペプチドを濃縮したショットガン解析も盛んに行なわれる。

≫二次元ゲル電気泳動の利点と欠点

二次元ゲル電気泳動[▼1-19]ではタンパク質の等電点に基づく分離をアガロースゲルを用いて実施し，その結果をポリアクリルアミド電気泳動（SDS-PAGE）によって分子量に基づいて展開する。この手法により，1000 以上のタンパク質をスポットとして分離，解析できる（**図 1**）。

等電点に基づく電気泳動では 100 K ダルトン（Da）を超える巨大タンパク質は泳動しづらく，10 KDa 以下のものはゲルを SDS-PAGE 用に洗浄する際に抜けおちやすい。また分析が容易な等電点の範囲は酸性側に偏る。上手な展開にも多くの熟練を要するため，最近はショットガンプロテオミクスが主流になってきた。

≫質量分析計によるタンパク質同定

ショットガンプロテオミクスで用いるデータベースとは，対象とする生物種のゲノム上に記載される全タンパク質のアミノ酸配列ライブラリである。ここからペプチドに対応する m/z を自動計算し，翻訳後修飾も考慮して検索に使う。当然のことながら，ゲノムから予測できないタンパク質は同定できない。さらに，同じペプチドあるいはほとんど同じ質量のペプチドが複数のタンパク質から生じうるため，遺伝子の推定は容易ではない。

混合物試料からペプチドを測定するにはタンデム質量分析（MS/MS）を利用するため，データベース検索は MS/MS イオンサーチともよばれる。前駆体イオンとして検出されたペプチドがフラグメントになるとき，ペプチド結合がランダムに1箇所切断されてイオンになる（**図 2**）。理論的にはすべての部分配列がプロダクトイオ