Proteome Letters　2016；1：63-80

教育セミナー：プロテオミクス熊の巻 **2015**
総説

# どのデータベースを使うか
## ～データベース検索と配列解析・誤解と難題～

吉　沢　明　康*

*E-mail: acyshzw@kuicr.kyoto-u.ac.jp

京都大学化学研究所バイオインフォマティクスセンター：611-0011 京都府宇治市五ヶ庄

（受付 2016 年 5 月 12 日，改訂 2016 年 6 月 10 日，受理 2016 年 6 月 14 日）

質量分析法によるプロテオミクス解析では，他のオミックス科学と同様，データのコンピュータ解析の過程が必須である．しかしゲノム科学やトランスクリプトーム解析に比べれば，質量分析法やプロテオミクスのためのバイオインフォマティクス，或いは解析手法・ソフトウェアは未だ発展途上であり，未解決の問題が多数残されている．更に，この状況に起因する多くの誤解や，特に実験系の研究者には扱いにくい技術的な問題も生じている．本稿ではこれらの問題を踏まえて，タンパク質の同定過程，特にデータベース検索法とそれに関連する基本的な事項について，プロテオミクス初心者を念頭に解説する．具体的には，*de novo* シークエンシング法とデータベース検索法の対比，PTM 探知のための手法が検索結果に及ぼす影響，生命科学データベースの概観とデータベース解析への応用上での要注意点などについて述べる．

### 1　序　論

10 年ほど前のこと，或るオミックス解析プロジェクトで，得られたプロテオーム・データに対するデータベース検索の結果が大きな問題になった．或るプレカーサーイオン（precursor ion）の質量ピークに対する検索結果が，検索対象にしたデータベースによって「逆に」なっていたのである．即ち，Swiss-Prot に対する検索結果では，タンパク質 A が第 1 位に，タンパク質 B が第 2 位になっているにもかかわらず，NCBI nr に対する検索結果では，その同じタンパク質 B が第 1 位に，そしてタンパク質 A が第 2 位になっていたのだった．これが，プロジェクトの関係者の間で問題になった．「違うタンパク質にヒットした，というのならともかく，同じタンパク質にヒットしているのに順位が逆になるのはおかしい．同一のペプチド配列を“読んだ”結果なのだから，どんなデータベースに入っていようが，より“似ている”タンパク質が高い順位で出る筈だ」というわけである．

このプロジェクトはそのまま迷走してしまったが，考えてみると，プロテオミクスもインフォマティクスも専門ではなかったその関係者たちの疑問はもっともで，この結果からは不思議な印象を受ける．例えば BLAST で異なったデータベースに検索をかけた場合，E-value の絶対値が変わることはあっても，E-value 順に検索結果をソートしたときに，同じアミノ酸配列（或いは塩基配列）の順番が入れ替わることはない．

では，質量ピークのデータベース検索では，どうしてこのようなことが起こり得るのか．それを検討する過程で，プロテオミクスでしばしば見受けられる誤解や，頻繁に遭遇する難題について論じ，注意を喚起するのが本稿の目的である．

本稿で対象にする「質量分析法を用いたプロテオーム解析」は，概ね以下のような段取りで進む：

（1）試料の前処理→（2）質量スペクトルの測定→（3）スペクトル波形処理→（4）質量ピーク探知→（5）同定

質量スペクトルから対象物質イオンの正確な *m/z* を得るためには，分析計のハードウェア特性や測定誤差・ノイズなどによって山型になったスペクトルから，本来のデルタ関数様の（櫛状の）信号，即ちピークを推定する必要があり，そのためのステップが（3）と（4）に該当する．計算機処理が必要なのは（3）以降の 3 ステップで，本稿ではこのうち最後のステップ，「（5）同定」部分，特にデータベース検索法に焦点を絞って論じる．

なお，イオンが 1 価の場合は *m/z* の値は質量の値に一致するが，ピークからはそのイオンが 1 価か多価かは判断できないので，多価イオンだった場合は質量の値を計算する必要がある．しかし本稿では簡単のためその過程は省略し，全て 1 価イオンと仮定して述べる．また用語の表記は原則として文献 1），2）の表記に準拠し，アミノ酸残基のモノアイソトピック質量（monoisotopic mass）も文献 1）掲載の値を用いる．

© 2016 Japanese Proteomics Society

## 2　誤解 1　データベース検索ではアミノ酸配列を「読め」ない

最初の問題は，質量分析データのデータベース検索法によるアミノ酸配列同定を，塩基配列の決定と同じ意味で「読む」と表現し解釈することは適切か，ということである．そこでまず，質量分析法による配列決定の方法について概観する．

塩基配列の決定の場合に用いられる（古典的な）Sanger 法では，「核酸 1 個違いの長さの塩基配列」を生成し，3′-末端の塩基を 1 個ずつ同定することによって塩基配列を決定する（Fig. 1(a)）．質量分析法の場合，同様にアミノ酸を 1 個ずつ同定することによってアミノ酸配列を決定する手法は，*de novo* シークエンシング（*de novo* sequencing）法と呼ばれる．Fig. 1(b) に示すように，「アミノ酸 1 個違いの長さのアミノ酸配列」を生成して質量スペクトルを測定すれば，質量ピークの間隔がちょうど C 末端アミノ酸 1 個の質量に相当するので，それを同定していくことでアミノ酸配列が確定する．これらの手法は共に，「最小単位（塩基またはアミノ酸）」を 1 個ずつ同定していく（最小単位の種類と配列内の位置が 1 個ずつ決定される）ことで，全体の文字列を確定する，という方法であり，そこが「読む」と形容される所以でもある．従って，「質量分析法による *de novo* シークエンシング法でアミノ酸配列を同定する」場合には，「配列を読む」と形容しても，誤解は生じない．

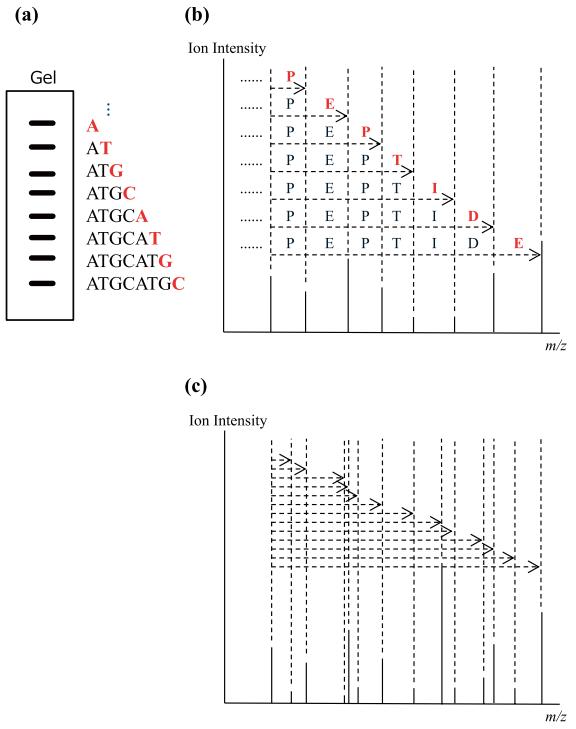


**Fig. 1**　Reading a sequence; Sanger method for nucleic acids and de novo sequencing method for peptides

A simple diagram of the Sanger method. Nucleic acid molecules are separated in the gel, and their 3′-terminals are identified to “read” the whole nucleic acid sequence. A simple diagram of the de novo sequencing method represented on a model mass peaks. The arrow lengths correspond to the m/z values of partial sequences of the amino acid sequence “PEPTIDE.” Each mass peak interval corresponds to the mass of the amino acid at the terminus (in this figure, the C-terminal) of each peptide. Examination of all peak intervals to find which interval is most probable to correspond to the mass of an amino acid.

### 2-1　*de novo* シークエンシング法は実際には困難である

しかし実際には，*de novo* シークエンシング法は巧く機能することが少ない．その理由は以下のようなものである：

第一に，タンパク質には PCR 法がないため，試料を増幅できない．またペプチドによってイオン化効率が異なるため，充分に測定できるほどの量のイオンが，特定のペプチドについては測定されないこともあり得る．このために，「1 個違いの長さの配列」が全て測定されてピークが得られる，という保証がない．

第二に，どれほど注意深く試料を精製しても，夾雑物の混入の可能性は高い．

第三に，*de novo* シークエンシング法は「アミノ酸」を決定する手法ではなく，その「質量」を決定する方法であるため，アミノ酸自体に修飾があった場合には，正しくアミノ酸を同定できる保証がない．

このように「どのピークを採用すれば良いか」判断が難しく，「（翻訳後修飾（PTM）のため）ピークの間隔に非常に多数の可能性がある」複雑な場合には，考え得る全ての場合について“総当たり”で確認することが必要になる．即ち「どのピークとどのピークの間が，アミノ酸 1 個に相当するのか全て試す」「考えられる全ての PTM の組み合わせを試す」などの試行を行う必要がある（Fig. 1(c)）．従って検討すべき場合の数は膨大なものになる．

総当たり問題は，問題を部分問題に単純化してから組み合わせる，例えば「一部分のピークのみをアミノ酸配列に対応させ，その組み合わせとして解く」ことで効率的に解くことができることがある．アミノ酸配列や塩基配列の類似性を検証する場合にも同様の問題が生じるが，この場合には動的計画法（dynamic programming）[3] が用いられることが多い．よく用いられる BLAST[4] の場合も，まず統計的手法を用いて『枝刈り』，即ち「およそ正解になりそうにない」配列をデータベースから除外した上で，最終的には動的計画法を用いて，問い合わせ配列とデータベース中の候補配列間の類似性を求めている．このときに動的計画法の評価関数（結果を“点数化”するための統一的かつ最適な基準）として用いられるのは「置換行列（substitution matrix）」，即ち「アミノ酸（や塩基）が別のアミノ酸（や塩基）に置換される頻度をスコア化したデータ」であり，

BLOSUM や PAM などの行列が用いられる．

しかしながら，この置換マトリックスのような評価関数は，質量スペクトルの検索の場合には存在していない．先述の BLOSUM や PAM は，飽くまでも「進化過程でアミノ酸が置換される場合には，性質の近いものから置換される（置換されたアミノ酸が性質の近いものだった場合，その生物種が生き残る）」という原理を反映しているものであり，「質量ピークの間隔をどのように読み間違える可能性があるか」という観点は反映しておらず，また「ピーク強度」がアミノ酸の量（個数）に比例するわけでもないからである．

これは本質的な問題であるため，現在も多くの工夫が為され，優れたソフトウェアも発表されてはいるが，*de novo* シークエンシング法がアミノ酸配列決定法の主流になるのは，現状ではまだ難しい．

### 2-2　データベースを利用する

そこで一般には，「予めペプチド配列を準備し，その（理論）*m/z* を計算して，測定で得られた試料の *m/z* と比較することによって，『一番もっともらしい』アミノ酸配列を探す」方法が用いられる．しかし仮に，単純にアミノ酸を組み合わせて作った理論ペプチドの *m/z* を計算しようとすると，その場合の数（ペプチドの種類）は膨大な数になる（例えばアミノ酸 7 個長のペプチドは，$20^7$=12 億 8 千万種類存在する）．従って実際には，理論ペプチドの計算ではなく，実在するタンパク質配列をもとに *m/z* の計算を行う．これが通常用いられる「データベース検索」である．単にアミノ酸配列を推定するだけでなく，「その試料が何であるかを確定する」ためには，その生物種のゲノム全体をカバーした配列データベースが必要である．また，遺伝子予測の段階で CDS（Coding Sequence）と認識されていない配列や，ゲノムにコードされていない配列（抗体の可変領域やポリケチドなど）は，配列データベースには通常，収録されていないので，同定することはできない．

「MS スペクトル（プレカーサーイオン）のみを用いる同定」法，即ち PMF（Peptide Mass Fingerprinting）法 [5] を用いると，プレカーサーイオンの *m/z* に合致するペプチドをデータベースから探すことができる（Fig. 2(a)）．ここで注意が必要な問題は以下の 2 点である：

1. *m/z* が一致するペプチドは，通常，複数種類存在する
2. ペプチドは複数のタンパク質に含まれていることがある

問題 1 については，*m/z* の差が一般的な質量分析計の分解能以下しかない，非常に近接した値を持つペプチドは多数存在しており，これらの区別は難しい．例えば（ペプチドではなく単一アミノ酸残基の例であるが）リシン（K）のモノアイソトピック質量（monoisotopic mass）は

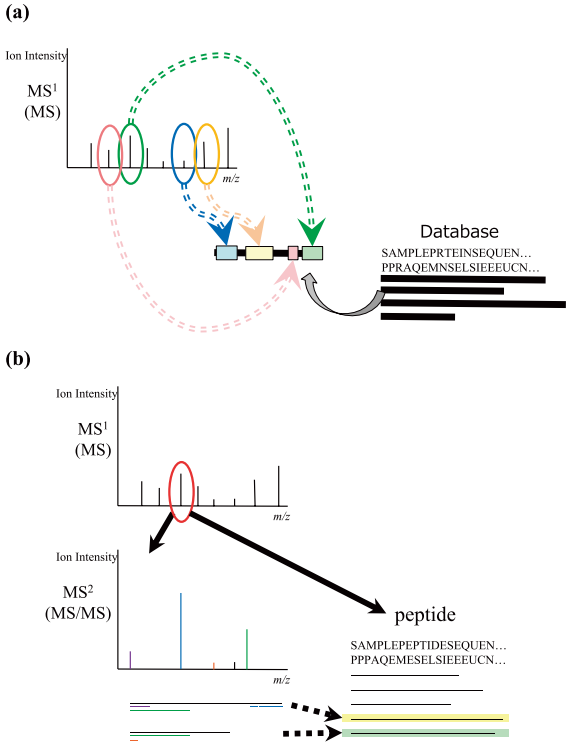


**Fig. 2**　Simple diagrams of Peptide Mass Fingerprinting (PMF) and Peptide Fragmentation Fingerprinting (PFF)

(a) PMF: Simple diagrams of mass peaks and protein sequences. The black bar represents a protein sequence, and the thick colored bars represent peptide sequences within the protein. Each mass peak corresponds to a peptide in a protein sequence stored in the database; in this figure, the peak in the colored circle corresponds to a peptide of the same color. Through the search, one protein that contains all peptides corresponding to the measured *m/z* values from MS spectra is assigned.
(b) PFF: Simple diagrams of mass peaks and peptide sequences. The dashed line represents a trypsin-digested peptide sequence. Peptide sequences corresponding to a MS peak (precursor ion) are extracted from the database, and MS/MS peaks (product ions) derived from that precursor ion are obtained. The b/y-ions are then estimated, and their corresponding peptide fragments are searched. In this figure, a MS/MS peak and its corresponding peptide fragment are marked with the same color. Two peptides that contain these peptide fragments are selected.

128.09496 Da，グルタミン（Q）は 128.05858 Da で，その差は 0.03638 Da しかないが，この差は四重極型（或いはイオントラップ型）質量分析計では区別するのが非常に難しい．

これ以外にも例えば，アラニン（A）とセリン（S）の質量の合計，及びグリシン（G）とトレオニン（T）の質量の合計は，完全に一致している．A+S の 2 個の単離アミノ酸中には H 原子 14 個，C 原子 6 個，N 原子 2 個，O 原子 5 個が含まれるが，G+T でもこの個数は完全に同一だからである．従って両者の理論質量は一致し，*m/z* 値からこの両者（の組み合わせ）を判別することは，原理的に

不可能である．

問題2についてはまず，実際のプロテオーム配列データ中に「一つの生物種の全タンパク質の中で，1種類のタンパク質にしか含まれていないペプチド」（unique peptide または specific peptide，以下「特異的ペプチド」）が何個存在するか確認してみたい．UniProt 2014_2[6] の human proteome dataset に対して，EMBOSS[7] プログラム•スイート ver. 6.2.0-2 に収録された digest プログラムを用い，ミスクリーベージ（missed cleavage）なしの条件で計算すると；

- 同データセットに収録されたヒトのタンパク質は88,665個
- これをトリプシンで消化して生成するペプチドは798,545個
- このうち特異的ペプチドは339,925個（全体の約42.6%）

特異的ペプチドが充分にイオン化されて測定された場合，これを proteotypic peptide と呼び，このペプチドだけでタンパク質を同定することが可能になるが，このデータが示すとおり，ヒトタンパク質からトリプシン消化で生成するペプチドのうち約6割，過半数は特異的ではなく，即ち proteotypic ではあり得ないことになる．

従って PMF では，測定されたペプチドの MS ピークから，可能性のある全てのペプチドをリストアップ，それらを，由来した可能性のあるタンパク質全てにマッピングし，「マップされるペプチドが最も多い」タンパク質を「最も確度が高い」と推定する．なお，「最も多い」という判定には，「タンパク質のカバー率が最も高い」という基準と，「マップされたペプチドの数（＝質量ピークから該当するペプチドを推定（assign）できた回数）が最も多い」という基準の，両方が用いられる．

当然ながら，この手法は混合物試料には適切でない．試料が純品でないならば，或るペプチドが複数のタンパク質から由来した可能性があった場合，どれ由来なのか判定ができないからである．このため，質量分析に「混合物試料を分離しつつ同定する」ことを任せようとしている場合には，PMF は手法として不充分である．2次元電気泳動などで試料を充分に分離した上で，同定のみを質量分析に任せる場合には，この手法で充分なこともある．

### 2-3　MS/MS ピークも利用する

混合物試料の場合には，MS/MS スペクトル（プロダクトイオン，product ion）も利用する．この手法は PMF に対比して Peptide Fragment (Fragmentation) Fingerprinting（PFF），または MS/MS イオンサーチ（MS/MS ion search）と呼ばれるが，その骨子を単純化すると，「プレカーサーイオンだけでなく，そのプロダクトイオンも用いる」，「ペプチドだけでなく，その部分ペプチド（ペプチド断片, peptide fragment）との一致も見る」ことによって，「一番もっともらしいペプチド配列を探す」ということになる（Fig. 2(b)）．

PFF で最初に行われるのは，PMF の場合と同様，「プレカーサーイオンから生じる MS ピークについて，そのピークを生じる可能性のあるペプチド（の候補）を推定する」ことである．言い換えるとこれは，「そのようなピークを生じる可能性のないペプチドを排除する」作業に該当し，計算機処理的には（BLAST の統計的手法に基づく処理と同様）「枝刈り」を意味する．

次に「プロダクトイオンから生じる MS/MS ピーク」を同定する．プレカーサーイオンは CID 等の手法によって開裂し，プロダクトイオンが生じているが，CID の条件下では開裂は「1個のペプチドにつき1カ所」でしか生じない．このため，「プロダクトイオンに対応するペプチド断片」は，元の「プレカーサーイオンに対応するペプチド」の「どちらか一方の末端を含む部分ペプチド」になっている．これらのプロダクトイオンのうち，開裂がペプチド結合の位置で生じ，「N 末端を含む部分ペプチド」がイオン化したものが *b*-ion，「C 末端を含む部分ペプチド」がイオン化したものが *y*-ion であり，これらを候補ペプチドの N-末端側または C- 末端側に揃えて（align して）矛盾が生じないものを絞り込める（プロダクトイオンに対応するペプチド断片は短いことも多く，その位置が任意の場合は候補ペプチドが非常に多数になる可能性があるが，実際には「末端がペプチドの末端と一致する」という“位置情報”によって候補が限定されている）．なお実際の測定結果には，機器の性能による差や測定誤差などが生じるため，“許容誤差”として tolerance を指定する．

この過程は，*de novo* シークエンシング法を用いることができないデメリットを，或る程度カバーしている．前述のように *m/z* が偶然一致する別のペプチドは通常，複数存在する．しかしその部分配列同士でも偶然 *m/z* が一致する可能性は小さくなる．従って可能な限り多数の部分ペプチドを整列（align）して情報を重ねていくことで，偶然の可能性を非常に低く抑えられている．特に「長さ1個違い」のペプチド断片（のイオン）が測定された場合には，その部分については実質的に *de novo* シークエンシングと同等の検証を行っていることになる．但し，この過程でもプレカーサーイオンの対応するペプチドが必ず1個に絞り込めることが保証されるわけではない．

「アミノ酸1個ずつを確定しながら配列を確定（identify）する（＝配列を読む）」*de novo* シークエンシング法と違って，この作業で可能になったのは「検索したデータベースの中で，最も可能性が高い配列を割り当てる（assign）」ことである．この差異は本質的に「違うもの」として扱われて

いる．例えばSwiss-Protは（現在は基本的にはEnsemblデータベースなどからアミノ酸配列を得ているが）歴史的な経緯もあって，実験的に確認されたアミノ酸配列は吟味の上収録することになっている．このため，研究者が自分の同定したアミノ酸配列を登録するための窓口（SPIN）を現在も設けている（https://www.ebi.ac.uk/swissprot/Submissions/spin/）が，ここでのデータ投稿は「エドマン法（Edman degradation），もしくは質量スペクトルを手作業で解析した（*de novo* シークエンシング法を用いた）場合」に限定されており，「検索エンジンを使った場合には，PRIDE[8]（EBIが運営するプロテオーム・レポジトリー）に登録せよ」と明記されている．

なお一般に「スコア」は開発者のロジックに基づいた点数に過ぎず，最適な指標であることが証明されたものではない．そこで例えばBLASTによる検索結果には，スコア（bit score）だけでなく，結果の信頼性を示すためのE-value，即ち「“正解”ではない配列が，“たまたま偶然で”同じスコアを出してしまう頻度を示す期待値」も表示されている．この算出のためには，GenBankに蓄積された核酸配列（と，それを翻訳したアミノ酸配列）の分布がまず調査され，これを正しく記述するために極値分布（extreme value distribution: Karlin-Altschul統計[9]）が考案されて，その分布式に基づいた計算が行われている．質量分析の検索エンジンの場合，「配列の類似性」を評価する分布ではなく，「質量スペクトルからアミノ酸配列を割り当てるときに誤判定する可能性」を評価する分布でなければならないが，そのような統計はまだ作られていない．従って，検索エンジンが示す“Expect（またはE-value）”の値は，BLASTと全く同様の意味でのE-valueではない（但し，同様に利用できるように工夫はされている）．

## 3　誤解2　データベース検索ではタンパク質は決められない

上述のPFFデータベース検索で決定できたのは，飽くまでも（タンパク質配列を切断して生成した）ペプチドの配列である．既に述べたように，ヒトの（トリプシン消化）ペプチドの約6割は特異的ではなく，複数のタンパク質中に存在している．従って，ペプチドを割り当てただけでは，タンパク質を推定したことにはならない（proteotypic peptideが推定できた場合は，それで決定的である）．

従って，最も可能性の高いペプチドの割り当て（assignment）の次の段階は「タンパク質の推定（inference）」である．基本的にはこの過程は単純で，推定されたペプチドを“より多く”含んでいるタンパク質が選ばれる．一つのペプチドを共有するタンパク質が多ければ多いほど，そのペプチドの“重み付け”を軽くするような処理が必要になる．また検索エンジンは，MSスペクトル（プレカーサーイオン）及びそこから得られたMS/MSスペクトル（プロダクトイオン）全体と，割り当てられたペプチド全てを紐付けて，“保守的”にタンパク質を推定する．即ち或るペプチドが，「他のペプチドによっても推定されているタンパク質A」と「他には推定するペプチドがないタンパク質B」の両方に含まれる場合，そのペプチドはタンパク質Aに由来する，と判断する．

この過程を改善するために，結果の信頼性を独自に評価するツールも作成されている．例えばSeattle Proteome Centerが配布しているTrans-Proteomic Pipeline（TPP）プログラム・スイートの中には，この問題に対応するためのProteinProphet[10]というプログラムがある．これは，プロダクトイオンからどのタンパク質が推定されたかというデータを基に，期待値最大化法（EMアルゴリズム）を用いて，タンパク質の推定が正しい確率を推定している．

タンパク質がペプチドを基に推定されている結果，ユーザーが想定していなかった副作用が生じることがある．それが，「測定を繰り返すと，タンパク質の同定結果が変わる」という現象，即ち「同一のサンプルを複数回測定し，その全ての結果について検索エンジンで検索すると，同定されるタンパク質が変わることがある」という事態である．

これは，質量スペクトルの測定の度に得られるスペクトルが完全に一定なわけではない，ということが原因である．スペクトルの形状が僅かに変化したために，或る測定回でのピーク検出時のみ，スペクトルの一部分がピークと認識される，ということは充分にあり得る．この結果，新しいプレカーサーイオンのピークが追加され，例えばこれに対してペプチドAが割り当てられたとする．今まではペプチドB及びCによってタンパク質Zが推定されていたが，仮にペプチドAとペプチドBで別のタンパク質Yが，より高いカバー率で推定可能で，かつペプチドCが由来した可能性のあるタンパク質Xも存在するのであれば，タンパク質の推定結果は「Z」から「XとY」に変わる．

MS/MSピーク（プロダクトイオン）の場合でも，同様のことが起こり得る．例えばFig. 3(a)の例では，ペプチド3と4の割り当ては確定しており，更に赤丸をつけたプレカーサーイオンのMS/MSピークを調べると，ペプチド1と2が候補となっている．しかしペプチド1のほうが，（プロダクトイオンに基づいて割り当てられた）ペプチド断片によるカバー率が高いため，ペプチド1が採用され，この結果，同じくカバー率の高い（短い）タンパク質Bが採用されている．しかしFig. 3(b)に示す測定回ではMS/MSピークが1個増え，これをペプチド2に割り当てることができたため，ペプチド2のカバー率がより高くなり，このプレカーサーイオンが割り当てられたのはペプチド2になった．ペプチド2がタンパク質Aに含まれていたため，推定されるタンパク質もAに変わってしまっている．

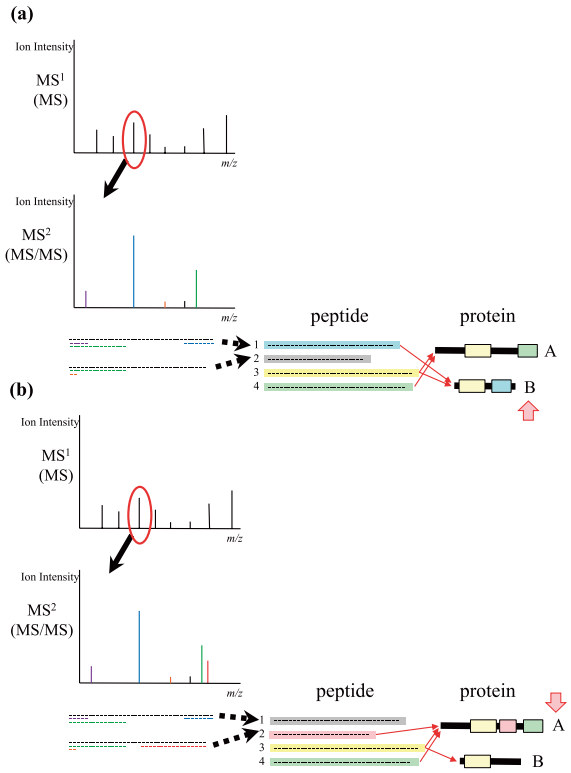


**Fig. 3**　An additional MS/MS peak may change the protein inference result

(a) The diagram is shown in the same manner as Fig. 2. As shown in Fig. 2, appropriate trypsin-digested peptides are extracted by the precursor ion information, and peptide 1, of which longer region is covered by the fragment peptide, is presumed by the product ion information. For peptides 3 and 4, which were identified by other product ions, a protein is inferred; in this figure, protein B is inferred because it is more covered by peptides 1 and 3 than protein A is covered by peptides 3 and 4.
(b) In case that an additional peak (red) is observed in a MS/MS peak list and a corresponding fragment peptide is identified: in this figure, the identified peptide has been changed to peptide 2; as a result, protein A, which is covered by peptides 2, 3, and 4 is changed to be inferred as protein A instead of protein B, which was covered only with by peptide 3.

「質量分析によるタンパク質の同定とは，実際には一番もっともらしいペプチドの割り当て（assign）であり，更にその結果から推定（infer）したタンパク質を出力している」ということを認識していなければ，結果の変動に驚かされることになる．

## 4　難題 1　PTM を指定しないと同定できず，指定すると結果が得られない

タンパク質の同定を更に難しくしているのが PTM である．既に述べてきたように，データベースサーチでは試料ペプチド全体の *m/z* と，データベース配列の *m/z* を比較する．従って「PTM の影響を一旦考慮せずに，まずペプチドのアミノ酸配列だけを確定し，後から PTM の影響を修正する」ようなことは不可能で，最初から「PTM を検討に入れた」形での配列推定しか実施できない．

PTM が *m/z* に与える影響は非常に大きい．例えばグリシン残基のモノアイソトピック質量は 57.02146 Da であるが，リン酸化による質量の増加は 79.99633 Da であって，グリシンよりも大きい．従って，PTM のあるペプチドを同定する場合には，PTM を考慮に入れることはデータベース検索にとって必須である．

ところがこれには大きな敵が存在する．それは『組み合わせ爆発（combinatorial explosion）』である．

データベース検索で PTM を探知する方法は，現状では，基本的には variable modification 法しかない．この方法は即ち，「生じる可能性のある全ての PTM の組み合わせを（メモリ中に）作成し，総当たりで調べる」という手法である．

仮に，以下のようなペプチド（アミノ酸配列）があったとする：

PEPTIDESTYLE（アミノ酸 1 文字記号で表記）

このペプチドにリン酸化が生じているかどうかを検証するためには，「リン酸化が生じたペプチド」の *m/z* を計算し，それを実測値と比較せねばならない．リン酸化が生じるのは S, T, Y の 3 種類のアミノ酸であるから，検索エンジンは，「これらのアミノ酸がリン酸化された場合のアミノ酸配列」（以下，「仮想ペプチド」と表記する）をメモリ中に作り出す（正確には，その *m/z* を計算するだけだが，便宜上ここではアミノ酸配列を書き出して説明する）．

「リン酸化された S」「リン酸化された T」「リン酸化された Y」を，仮にそれぞれ「s」「t」「y」と小文字で表すとすると，メモリ中に新たに作り出される仮想ペプチド（正確には，「*m/z* を計算する対象のペプチド」）は，以下の 15 個である：

| | | |
|---|---|---|
| PEPtIDESTYLE | PEPTIDEsTYLE | PEPTIDEStYLE |
| PEPTIDESTyLE | PEPtIDEsTYLE | PEPtIDEStYLE |
| PEPtIDESTyLE | PEPTIDEstYLE | PEPTIDEsTyLE |
| PEPTIDEStyLE | PEPtIDEstYLE | PEPtIDEsTyLE |
| PEPtIDEStyLE | PEPTIDEstyLE | PEPtIDEstyLE |

元のペプチド配列と合わせて 16 個の場合について，全て「実測値とマッチするかどうか」総当たりで検討することになる．同様に，PTM 化される可能性のあるアミノ酸がペプチド中に *n* 個あるならば，検討すべき場合の数は $2^n$ になる．

これは即ち，「variable modification 法を用いると，*n* 個の PTM 可能部位を持つ 1 個の配列が，メモリ中で合計 $2^n$ 個に増殖する」ことを意味する．またこれらの配列は，元々 1 つの配列だったものを部分的に変更したものであるので，よく似ている．そして，「データベースの配列数が増える」場合，「よく似た配列が多数，データベース中に含まれる」場合は，共に E-value に影響を与える（信頼性が低下する）

可能性が高い．

次に，仮想例でなく実際のデータとして，再び UniProt 2014_2 のヒト・プロテオーム・データを用い，「タンパク質をトリプシン消化して得られたペプチド 1 個の中に存在する，S，T，Y の個数（＝リン酸化可能部位）」を数える．結果を Fig. 4 に示す．横軸は「1 ペプチド中に存在する S または T または Y の個数」，縦軸は「ペプチド（データベース配列のペプチドと仮想ペプチド）の個数」である（このグラフの原データを Supplementary (Table S1(a)) として付す）．この結果が示すように，1 ペプチド中に存在するリン酸化可能部位は 1 個のことが最も多く，その個数（そのようなペプチドの数）は 208,023 個．リン酸化可能部位は，次いで 0 個または 2 個のことが多く，10 個を超えて，非常に少数ながらほぼ 50 個まで分布する．リン酸化可能部位の数が最も多いペプチドの場合，1 個のペプチド中に 47 カ所の S または T または Y が存在している．ここから生成する仮想ペプチド（＝追加された，検討する必要がある場合の数）は $2^{47}$-1，即ち約 140 兆個で，意味のある時間内に計算（検討）を終了することは不可能である．このような「場合の数の爆発的増加」は，一つのペプチド中に多数の修飾部位が存在している場合に生じる．これが『組み合わせ爆発』である．

仮に「1 ペプチド中に 13 個以上の修飾可能部位（＝ S または T または Y）が存在する場合は考慮しない」と仮定し，「1 ペプチド中の修飾部位が 12 個以下」であるようなペプチドのみで場合の数を数えたとしても，それでも合計約 3,371 万個になる（Supplementary (Table S1(b)) 参照）．既に述べたように，このデータセット中のタンパク質配列

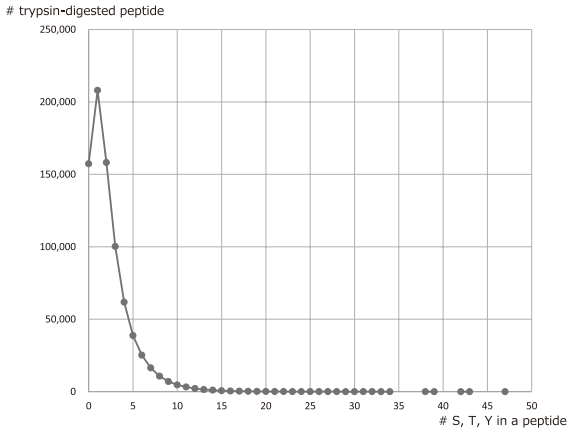


**Fig. 4**　Distribution of the number of possible phosphorylation sites (S or T or Y) in trypsin-digested peptides derived from human proteins stored in UniProt 2014_2, and the peptide number

X axis: The number of possible phosphorylation sites (S or T or Y) in a peptide
Y axis: The number of corresponding trypsin-digested peptides derived from human protein stored in UniProt 2014_2

は 88,665 個で，そこから生成するトリプシン消化ペプチドは 798,545 個である．従って variable modification 法によってメモリ中に生成される仮想ペプチドの数（その *m/z* が検討対象になる場合の数）は，元々トリプシン消化によって生成されていたペプチドの約 40 倍になる．

なお注意が必要だが，トリプシンが切断部位で切断を行わない現象，ミスクリーベージも場合の数を増加させる．切断されないことによってペプチドが長くなるため，「1 ペプチド中に含まれる PTM 可能部位」の数が増加するからである．

以上をまとめると，

1. PTM は「あり」の条件でないと，PTM が探知できないのみならず，そのペプチド自体も同定できない．

これに対しては，PTM「なし」の条件で検索を行い，得られた「PTM なし」のペプチドを含むタンパク質を「結果の候補」として，それらに対してのみ PTM「あり」の検索を行う（“multi-path search”[11]），といった対策が考えられる．PTM が全く生じていないペプチドが存在している可能性は高いので，通常の検索で結果的にこれとほぼ同等のことが行われていることもある．

また次の問題として，

2. PTM 可能部位が多すぎると，検索エンジンが検討すべき場合の数が“組み合わせ爆発”を起こす．

これに対しては，「1 つのペプチド中に，非常に多種類の PTM が混在している可能性は低い」という一般的な考えに基づいて，PTM の種類を 1 種類（高々 2 種類）程度に抑えて検索を実行する．ミスクリーベージ数（missed cleavage number）も高々 1（可能ならば 0）に指定する．仮に「1 ペプチド内の PTM 可能部位」の数が多くなりすぎた場合，非常に長い計算時間がかかることになる．例えば先述の「47 カ所のリン酸化可能部位を含むペプチド」が 1 個含まれているだけで，仮に他のペプチドに一切 PTM がなかったとしても，検討すべき場合の数は 798,545，つまり約 80 万から 140 兆に増加し，計算量・計算時間も 1 億倍以上に増加する．実際には，このような配列に対する仮想ペプチド生成処理は取りやめられるか，或いはメモリ不足でプログラムが異常終了するだろう．

経験的に，検索時間が数時間を超えることは少ないので，検索時間が数時間に達した段階で，一旦終了して PTM の条件を変更し，再検索したほうが効率がいいだろう．

PTM 探知のためには多くの工夫が為されているが[12]，大量のアミノ酸配列に対する推定法としては，variable modification 法を用いて上述のように対応する，というのが主流の対応であろう．

## 5　難題 2　オミックス解析では結果が信頼できない？

E-value の信頼性が低下するのは，前項で述べた PTM

の探知の場合に限らない．オミックス解析自体でも同じことが発生する．検索エンジンが示すE-valueは，ピーク1個にアミノ酸配列1個を対応づける（assign）ことに対して計算される．しかしプロテオーム解析では数百個，場合によっては数千個以上のタンパク質を同時に扱うため，多重検定（multiple testing）を行う必要がある．

検定で特に問題になる結果のカテゴリは，以下の2つである：

・偽陽性（false positive）：αエラーまたは第1種過誤，誤って“正解”とされているが実は“不正解”
・偽陰性（false negative）：βエラーまたは第2種過誤，誤って“不正解”として捨てられてしまった，実は“正解”

なお「擬陽性」という単語もある（一定以上の年齢層には「ツベルクリン反応の結果の一つ」として馴染みがある）が，これは「陽性と陰性の中間程度の値」のことで，意味は全く異なる．

最も“保守的”（手堅い）とされるBonferroni検定を行うと，結果の採用・不採用を決める基準である有意水準を，検定の回数（この場合は，同定を行うピークの数）で割って比較する必要がある．この結果，有意水準が厳しく（小さく）なりすぎて，有意と判定される結果がなくなってしまう．この手法は「p-value（またはE-value）を用いてfamilywise error rateを調節する方法」であるが，一般に「偽陽性（false positive）を減らす処理によって偽陰性（false negative）が増え，偽陰性を減らす処理によって偽陽性が増える」という関係がある．Bonferroni検定は，「αエラーを厳格に減らしたために，βエラー，即ち『正解なのに誤判定され捨てられる（“取りこぼす”）』ケースを増加させてしまっている」ことになる．

そこで用いられるようになったのが，「検索結果には一定程度の偽陽性が含まれることが不可避である」という考えに基づいて，False Discovery Rate（FDR）を計算し，この値を調整することで結果を求める，という手法である．FDRは「陽性の（positiveな）結果の内の偽陽性の比率」，即ち

FDR＝(false positive assignment)/(all positive assignment)

で表される．E-valueが個々の同定結果について計算され，「その値を用いて，その（個々の）結果に信頼性があるかどうかを判断し，信頼性が低いと判断された結果を除く」ために用いられるのに対して，FDRは「その結果全体の中に，何％くらいの偽陽性が含まれているか」を見積もるための指標である．従って個々の 偽陽性の結果を取り除くのには用いることができない．むしろ「結果の何％かはαエラー（偽陽性）である」ことを宣言し，それだけのαエラーを許容する代わりにβエラー（偽陰性）を減少させるための指標である．

この手法は，1995年にヘブライ大学のBenjaminiとHochbergによって発表され[13]，その後2002年頃までに，マイクロアレイの研究者が中心になって応用が進められた（例えば，統計解析用のR言語の，adjusted p-value関数やq-value関数なども，マイクロアレイの研究者によって書かれている）．

質量分析を用いたプロテオミクス研究でFDRが利用されるようになったのは2005年頃で，target-decoy searchという形で，このBenjamini-Hochberg法が導入された[14),15)]．即ち，target searchは通常のデータベースに対する検索を意味し，target配列は通常のデータベース配列を意味する．FDR，即ち「同定された全タンパク質配列」中の「誤って同定された結果」（の比率）を計算するためには，「誤って同定された結果」（偽陽性）の個数を知る必要があるが，そもそも「どれが偽陽性か」を知ることができないからこの問題が生じているわけで，従ってこのままでは計算は不可能である．そこでまず，「明らかに不正解である」と断言できるような配列，即ち「本来の配列データベースには決して含まれない配列」を，検索対象のデータベースに加えておく．これがdecoy配列である．アミノ酸組成によるバイアスを避けるために，殆どの場合decoy配列には，データベース配列のアミノ酸をシャッフルしたrandom配列か，もしくはデータベース配列を完全に“前後逆”にしたreverse配列が用いられる（作成に相対的に手間のかからない，reverse配列が用いられることのほうが多いようであるが，どちらがより適切かは吟味の必要がある）．

なお，これらの配列セットは別々に検索対象にしてもいいが，両者を一つのファイルにして検索することが（処理が簡便なため）多い．decoy配列が同定される場合は，「target配列が同定される（もしくは何も同定されない）筈であったにもかかわらず，誤ってdecoy配列が同定された」場合であり，これは確率的に生じると考えられる．従って当然，逆に「decoy配列のみが同定される（もしくは何も同定されない）筈であったにもかかわらず，誤ってtarget配列が同定された」場合も，等確率で生じる，即ち（母集団が同数なので）同数含まれると考えられる．これで偽陽性の結果の個数を知ることができる．

注意が必要であるが，decoy配列を同定した場合は偽陰性（false negative，即ち“正解”だったのに誤判定されて“不正解”とされた結果）ではない．データベース検索の結果，データは「陽性（positive，配列を同定できた）」と「陰性（negative，配列を同定できなかった）」の2群に分けられるが，そのうち陽性のみがサーチ結果として表示されている．従って，「誤ってdecoy配列が同定されたピーク」と「誤ってtarget配列が同定されたピーク」は共に，「表示されているのだから陽性」であり，かつ「同定の誤り」即ち偽（false）な判定であるので，どちらも偽陽性（false positive）ということになる．従って，target配列とdecoy

配列を同時に検索した場合，偽陽性の数は，decoy 配列が同定された数（decoy assignment (number)）を 2 倍する必要がある．

FDR=(Decoy assignment×2)/{(Target assignment–Decoy assignment)+(Decoy assignment×2)}

=(Decoy assignment×2)/(Target assignment+Decoy assignment)

=(Decoy assignment×2)/All (positive) assignment

繰り返しになるが，FDR は全ての結果が得られた後でなければ計算できないし，誤判定している結果を特定する目的にも使うことができない．また検索結果の E-value 閾値（threshold）を幾らに指定するのが適切か，ということも（事前には）判らない．従って E-value 閾値を色々変えてみて，その全ての場合の検索結果について FDR を計算し，その結果から，意図した FDR になるように E-value 閾値を見積もる必要がある．これでは手間がかかるので，通常は，E-value の閾値を大きく（悪く）とっておいて計算する．検索結果を E-value の良い順からソートしてリスト化し，リストの上位から下位に向かって，検索結果の配列を別の“最終結果”リストに採用していく．1 配列増やす度に“最終結果”全体の FDR を計算し，FDR が事前に決めた値（1%や 5%など）に達したところで採用をやめれば，最終結果全体の FDR を事前に決めた値に調節できる．

## 6　誤解 3　「%（パーセント）ホモロジー」という概念はない

データベース検索の結果から，「同定できたペプチドの，タンパク質全体に対する比率」を「タンパク質全長の $n$% に相当する長さのペプチドが同定された」というように求めることができる．これはちょうど一般の分子生物学研究で，2 つのアミノ酸配列や塩基配列を比較したときの両者の配列の一致度を，「アミノ酸配列全体の $n$%が一致している」として%を用いて表すのと同様である．ところが，これを「$n$%のホモロジーがある」と表現していることがある．

しかし「ホモロジー homology」という語・概念は進化学の用語・概念であって，進化的な類縁関係があるときに「ホモロジー（相同性）がある」と呼ぶ（同様に，「進化的に類縁の遺伝子」のことを homolog と呼び，その日本語訳は「相同遺伝子」である）．類縁関係を%で表すことはできない．例えば親子鑑定をしたときに「血縁がある確率は 10%です」と言われれば，その意味は（確率論的にもそれ以外でも）明らかだが，「血縁は 10%だけあります」と言われたらナンデスカソレハとなるだろう．血縁（遺伝的類縁）は「ある」か「ない」かのどちらかである．パーセンテージはその確率を示すに過ぎない．また，進化的な類縁関係の判定には進化系統樹を描くことが必要であり，配列が似ているかどうかだけでは判定できない．そして，質量分析法による測定のみでそのような結論が得られることはない．従って質量分析の結果を「%ホモロジー」と表現するのは，二重の誤用である．

このような誤用が散見されるようになった理由の一つは，例えば BLAST のような「配列の類似性比較プログラム」が「ホモロジー検索（homology search）」と呼ばれているからではないかと思われる．しかしホモロジー検索とは，「ホモロジーが『ある』か『ない』のどちらであるかを，何らかの類似性（配列の一致度を用いることが多い）を用いて推定する手法」という意味で，それで結果が「70%」というのは，「ホモロジーが存在する確率が 70%（その程度に似ている）」という意味でしかない（しかもこの値は，「配列の一致度合い」の値とは異なる）．別個に系統解析を行っていない限り，プロテオミクスや質量分析の研究で「ホモロジー」という用語が登場する可能性は考えにくい．

## 7　難題 3　データベースが多すぎる

データベース検索過程に於ける最後の難題は，「非常に多くのデータベースがある中で，どのデータベースを（試料ペプチドの同定に）使うのが最も効率的か？」というものである．Table 1 は配列の同定に利用されることの多いデータベース，或いはそれに関連する代表的な配列データベースについてまとめたものであるが，どう違っているのか，極めて判りにくい（なお，少し詳細な説明を加えた日本語記述の表を，Supplementary (Table S2) として付した）．

そこでデータベースの利用にあたっては，以下のようなことに留意することが必要だろう．

○プロテオーム解析のためのデータベース（などのリソース）については，欧 EMBL-EBI が特に力を入れている

米 NCBI も，従来プロテオミクス関連リソースに力を入れてきたが，この数年間は，プロテオーム・データ・レポジトリー Peptidome[16),17)] の閉鎖，検索エンジン OMSSA（Open Mass Spectrometry Search Algorithm）[18)] の開発終了など，プロテオミクスからは少し離れている．EBI は 2000 年代前半から継続的に，UniProt[6)]，neXtProt[19)] やプロテオーム測定の“生”データのレポジトリー PRIDE[8)] の運営のほか，プロテオームに対応する遺伝子データベースとして Ensembl[20)] を整備するなど，多くのリソースをプロテオーム解析に有用な形で結びつけている．

○収録配列数の多さと，アノテーションの品質の高さは，概ね相反する傾向にある

収録配列数が多ければ，それだけ，全てに対して詳細なアノテーションを行うのは難しくなる．通常のデータベースは，網羅性かアノテーションの品質かどちらかのみを追求しており，その点で，UniProt は特徴的である．UniProt は，キュレータによる詳細アノテーションのある Swiss-

**Table 1** Popular public databases for life science research

**Protein sequence database**

| Name | Formal Name | Developer | Description | Reference |
|---|---|---|---|---|
| UniProt | Universal Protein Resource | SIB (Switz.) & EMBL-EBI (EU) | The collective name of protein databases, consisting of UniProtKB, UniRef, and UniParc. | 6) |
| UniProtKB | | | An integrated database for proteins, consisting of Swiss-Prot and TrEMBL. | |
| UniRef | | | Clustered sets of sequences from the UniProtKB and selected UniParc sequences. | |
| UniParc | | | A comprehensive and non-redundant protein sequence database, which archives all past sequences under UniProt. | |
| Swiss-Prot | | | Generated by manual annotation of TrEMBL sequences by curators. High quality annotation to identify isoforms. | |
| TrEMBL | Translated EMBL | | Automated translation of base sequences in ex-EMBL (current ENA) database into amino acid sequences; presumed to be the same as Genpept. With automatic annotation for genes. | |
| neXtProt | | SIB (Switz.) & GENEBIO (Switz.) | Aims for model organism database for *Homo sapiens*; collecting all known information on human sequences and annotations. | 19) |
| GenPept | | NCBI (US) | Automated translation of base sequences in GenBank database into amino acid sequences; presumed to be the same as TrEMBL. | — |
| nr (nr-aa) | | NCBI (US) ICR, Kyoto Univ. (Jpn), etc. | An amino acid sequence collection for the search engine target datasets; collected sequences from multiple databases and redundant sequences removed. | — |
| IPI | International Protein Index | EMBL-EBI (EU) | Project completed; inherited to UniProt | 21) |

**CDS sequence database**

| Name | Formal Name | Developer | Description | Reference |
|---|---|---|---|---|
| RefSeq | The Reference Sequence | NCBI (US) | Manually annotated nucleotide/amino acid sequences by curators; organism specific sequence data files not available. | 28) |
| KEGG GENES | Kyoto Encyclopedia of Genes and Genomes | ICR, Kyoto Univ. (Jpn) | Sequences from RefSeq and other reliable resources are "purified" and classified into organism specific data files with annotations and rich hyperlinks. | 30) |
| Ensembl | | EMBL-EBI (EU) | The database of ORFs (and genes/proteins) directly predicted from the entire genome independently from the genome projects. EBI designates this database as the gene database corresponding to UniProt. | 20) |
| CCDS | Consensus CoDing Sequence | NCBI (US) & Sanger Institute (UK) | A common ID is given to the sequence commonly included in the CDS sets for both human and mouse, predicted by NCBI and the set by Ensembl; aims for "a complete set of protein-coding genes with high quality annotation." | 27) |
| H-inv | H-invitational | AIST & Tokai Univ. Medical School (Jpn) | Human mRNA database with very detailed annotation and hyperlinks. | 29) |
| H-EPD | H-inv Extended Protein Database | | An union set of H-inv, RefSeq and UniProt; entries from these databases are merged and redundant entries are removed. Generated especially for searching for missing proteins. | 31) |

**Nucteotide sequence database**

| Name | Formal Name | Developer | Description | Reference |
|---|---|---|---|---|
| GenBank/ ENA (EMBL)/ DDBJ | GenBank/European Nucleotide Archive/ DNA Databank of Japan | GenBank: NCBI (US)/ENA: EMBL-EBI (EU)/DDBJ: NIG (Jpn) | Nucleotide sequence repositories submitted by the experimental scientists themselves; maintained under the international cooperation (INSDC). | 32)~35) |
| Entrez Gene | | NCBI (US) | The data search/retrieve interface for all data in NCBI; managing data with Gene ID. | — |
| nr/nt (nr-nt) | | NCBI (US) ICR, Kyoto Univ. (Jpn), etc. | A nucleotide sequence collection for the search engine target datasets; collected sequences from multiple databases and redundant sequences removed. | — |

Prot と，網羅性の高い TrEMBL から成り，更新が終了した International Protein Index（IPI）[21] に代わる役割も果たす．生物種ごとのタンパク質データセット（Proteome Dataset）のダウンロードも可能である（但し，配列に重複がないことは保証されていない）．

UniProt で最も紛らわしいのは，「UniProt」と「UniProtKB」と「Swiss-Prot」の違いであろう（Table 1 参照）．プロテオミクス分野では UniProtKB 以外の UniProt データベース（即ち UniRef と UniParc）を使うことは少なく（またこれらの名称を明示するのが普通で），このため UniProtKB は UniProt と省略されることが非常に多い（本稿でも UniRef, UniParc には触れないので，今までもこの後も，特に断りなく，UniProtKB の意味で UniProt と書いている）．

また，「UniProt（またはその Proteome Dataset）に対する検索」では「既知のプロテオーム全体」に対する検索を実現できているが，「Swiss-Prot に対する検索」では「プロテオームの部分集合」に対する検索しかできていないことになる（なおヒト・タンパク質については，基本的な部分のアノテーションは全て Swiss-Prot で完了しているが，アイソフォーム（isoform）情報などは現在も拡充中である）．

○UniProt 以外の選択肢としては，MOD（Model Organism Database; モデル生物データベース）が有用な可能性がある

「特定の生物種（特にモデル生物）専門のデータベース」，特に，その生物種の研究コミュニティが結集して作成した，いわゆるコミュニティ・データベース（community database）は，収録した情報の質が非常に高いことが多い．代表例としては，以下のようなものが挙げられる；

- MGD[22]（マウス），RGD[23]（ラット），FlyBase[24]（ショウジョウバエ），WormBase[25]（センチュウ（線虫）），TAIR[26]（シロイヌナズナ）

○適切なタンパク質データベースが存在しない場合には，遺伝子データベースの配列を利用することになるが，収録配列がタンパク質の配列と異なっていることには注意が必要である

*m/z* 値からは配列が「類似」しているかどうか評価できないため，質量ピークのデータベース検索では「概ね似ているアミノ酸配列」を探知することが難しい．従って配列の網羅性が高いデータベースが望ましく，またアイソフォームや主鎖切断（truncation）などの結果，タンパク質の配列が遺伝子の配列から変化していることもあり得るので，それらの事実がデータベースから判るのが望ましい．遺伝子データベースにはアイソフォームや PTM 情報は含まれていないことが多いので，この意味では利用に向いていない（但し UniProt でもこのような情報が網羅できている保証はない）．例えば CCDS[27] は，CDS 部分のコンセンサスを集めたもので，存在確度の高い配列の集合である．しかしこれは即ち，「少数精鋭の配列」だから，検索対象として「できるだけ多くの配列を網羅する」という条件は満たしていない．

○データベースの目的，特徴，特に配列の重複に注意する

多くのデータベースはエントリに重複がある．この重複によって生じるバイアスは検索結果に影響をもたらす（variable modification 法の場合と同じ現象である）．nr は複数のデータベースを統合しているが，重複を除いているため，重複のある GenBank を翻訳した GenPept，同じく重複のある EMBL を翻訳した TrEMBL よりも，検索の対象には適している．従って，試料タンパク質の由来生物種が不明な場合には，（TrEMBL を含む）UniProt 全体に対して検索をかけるよりも，nr に対して検索をかけたほうが，結果が有意か否か判断しやすい可能性がある．

現在までの開発の歴史を振り返ると，塩基 / アミノ酸配列を収集したデータベースが（場合によっては複数個）作成され，肥大化し始めると，その内容を整理したデータベースが作成されるようになる．例えば遺伝子情報が豊富になった時期には，ゲノム情報を元に RefSeq[28]，Entrez Gene，CCDS といったデータベース（など）が登場したし，マイクロアレイを用いたトランスクリプトの研究が隆盛を極めた時期には，H-inv[29] が登場している．作成の目的を念頭に置くことで，より相応しいデータベースの利用が可能になるだろう．

生命科学データベースは新しいデータベースが次々に開発され，また更新が止まるものも少なくない．日本語で調査するならば，JST NBDC の『Integbio データベースカタログ』（http://integbio.jp/dbcatalog/）で簡単な解説を見ることができる．またデータベース自体も「データベース論文」という形で，多くのジャーナルに掲載されるようになっている（*Nature* や *Cell* の Resource コーナーに載ることも稀にある）．“データベース論文を載せるジャーナル”として最も代表的なものは，*Nucleic Acids Research* 誌の毎年 1 月 1 日号（Database issue）及び 7 月 1 日号（Web server issue）であり，これらの調査は有益であろう．

## 8　結　論

ここまで，質量スペクトルからのアミノ酸配列推定に関わる誤解・難題について駆け足で考察してきた．ここで，序論で述べた問題に戻ってみたい．簡単に言えば，「測定する度にタンパク質同定結果が変わる」という現象と似たことが生じていると考えられる．例えば以下のようなことが起こった可能性がある：

タンパク質 A が B よりも長いならば，同じペプチドによってこれらのタンパク質が同定されている場合，カバー率は B のほうが（短いので）高くなり，スコアも高くなる．

更にもう一つ別のペプチドXも同定されていて，これがAと，Swiss-Protに収録された別のタンパク質Cに含まれている場合，Aを支持するペプチドが1つ増えるため，Aの方が順位が高くなるだろう（XがAのみに含まれている場合は，proteotypic peptideであることになるので，結果はAのみになる）.

しかしSwiss-Protは高品質のアノテーションを手作業で行っているため，nrと比べれば配列数は圧倒的に少ない.ペプチドXが，「nrにのみ含まれる充分に多数のアミノ酸配列」中にも存在していた場合，ペプチドXの“重み付け”は低くなり，判定に殆ど寄与しなくなる可能性がある．或いは，Swiss-Prot中のどのタンパク質にも帰属できなかったイオンとペプチドXが全て，「nrにしか含まれていないタンパク質D」に帰属可能であれば，ペプチドXは「全てD由来」と判断される可能性が高い．いずれの場合でも，配列数の多いnrではペプチドXが同定に寄与せず，短いBのほうが高い順位になるだろう.

質量分析法によるタンパク質同定では，配列を「読んで」いるわけではなく，また直接同定されるのもペプチドであって，タンパク質はそれを基にデータベース中から推定した結果として得られる．それを認識していれば，この，序論で述べたプロジェクトも迷走することはなかったかもしれない.

2015年から，JST NBDC統合化推進プログラムのもとで，日本発のプロテオーム統合データベースjPOST（http://jpost.org/）の構築が始まった．タンパク質同定の方法や，統計的信頼性の確保，プロテオーム解析のためのアノテーションなど，この分野のバイオインフォマティクスには課題が山積であり，この分野へ参入する研究者が強く望まれている.

なお本年2016年から，筆者を含む有志研究者で「質量分析インフォマティクス研究会」を立ち上げた．この会は日本バイオインフォマティクス学会（JSBi）の公募研究会としての活動も行っているので，その一環として定期的にワークショップなどを行う予定である．また中長期的には，質量分析法やプロテオミクス分野の研究者とバイオインフォマティクス研究者の情報交換や交流の場としていきたいと考えている．「インフォマティクスが必要だと思ってはいるが，自分では手が着けられない実験系研究者」から「生物のことは全く知らないが，この分野に関心のある情報系研究者」まで，幅広く交流を行えるよう，会のメンバーはJSBi会員に限定せず，一切の制限を設けていない．会のweb URLはhttp://www.ms-bio.info/である．興味をお持ちの方は，是非ご参加いただきたい.

## 謝　辞

本稿は，2015年7月23日に熊本市で開催された日本プロテオーム学会年会の教育セミナー『プロテオミクス熊の巻2015』で行った講演を基に，加筆したものである．『教育セミナー』という構成上の都合で本稿（本講演）は単著としたが，取り上げる内容の選定から原稿に対するコメントまで，jPOSTプロジェクト（http://jpost.org/）及びMass++ユーザー会（http://www.mspp.ninja/）のメンバー，特に以下の先生方からご指導やご協力を頂いた．厚く御礼申し上げる.

石濱泰（京都大学•薬•製剤機能解析），松本雅記（九州大学•生体防御研•トランスオミクス），五斗進（京都大学•化研•バイオインフォマティクスセンター），荒木令江（熊本大学•医•腫瘍医学），田畑剛（京都大学•薬•製剤機能解析），草野麻衣子（名古屋大学•医•法医•生命倫理学）（敬称略）

また，現在私が所属するバイオインフォマティクスセンター化学生命科学研究領域教授の緒方博之先生はじめ，緒方研究室のメンバーにも有形無形のご援助を頂いた．厚く御礼申し上げる.

jPOSTプロジェクトは，JST NBDC（科学技術振興機構•バイオサイエンスデータベースセンター）「統合化推進プログラム」予算を受けて進められている．また計算リソースは，京都大学化学研究所スーパーコンピュータシステムから提供を受けた.

著者に開示すべき利益相反状態は無い.

## 文　献

1）日本プロテオーム学会．プロテオミクス辞典．東京：講談社；2013.
2）日本バイオインフォマティクス学会．バイオインフォマティクス事典．東京：共立出版；2006.
3）Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147(1):195–197.
4）Altschul SF, Madden TL, Schaffer AA, *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–3402.
5）Pappin DJ, Hojrup P, Bleasby AJ. Rapid identification of proteins by peptide-mass fingerprinting. Curr Biol. 1993;3(6):327–332.
6）UniProt C. UniProt: a hub for protein information. Nucleic Acids Res. 2015;43(Database issue):D204–D212.
7）Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000;16(6):276–277.
8）Vizcaino JA, Csordas A, del-Toro N, *et al.* 2016 update of the PRIDE database and its related tools. Nucleic Acids Res. 2016;44(D1):D447–D456.
9）Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci U S A. 1990;87(6):2264–2268.
10）Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry.

Anal Chem. 2003;75(17):4646–4658.
11) Tharakan R, Edwards N, Graham DR. Data maximization by multipass analysis of protein mass spectra. Proteomics. 2010;10(6):1160–1171.
12) Na S, Paek E. Software eyes for protein post-translational modifications. Mass Spectrom Rev. 2015;34(2):133–147.
13) Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J Royal Stat Soc B. 1995;57(1):289–300.
14) Elias JE, Haas W, Faherty BK, Gygi SP. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. Nat Methods. 2005;2(9):667–675.
15) Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods. 2007;4(3):207–214.
16) Slotta DJ, Barrett T, Edgar R. NCBI Peptidome: a new public repository for mass spectrometry peptide identifications. Nat Biotechnol. 2009;27(7):600–601.
17) Ji L, Barrett T, Ayanbule O, *et al.* NCBI Peptidome: a new repository for mass spectrometry proteomics data. Nucleic Acids Res. 2010;38(Database issue):D731–D735.
18) Geer LY, Markey SP, Kowalak JA, *et al.* Open mass spectrometry search algorithm. J Proteome Res. 2004;3(5):958–964.
19) Gaudet P, Michel PA, Zahn-Zabal M, *et al.* The neXtProt knowledgebase on human proteins: current status. Nucleic Acids Res. 2015;43(Database issue):D764–D770.
20) Yates A, Akanni W, Amode MR, *et al.* Ensembl 2016. Nucleic Acids Res. 2016;44(D1):D710–D716.
21) Kersey PJ, Duarte J, Williams A, *et al.* The International Protein Index: an integrated database for proteomics experiments. Proteomics. 2004;4(7):1985–1988.
22) Bult CJ, Eppig JT, Blake JA, *et al.* Mouse Genome Database G. Mouse genome database 2016. Nucleic Acids Res. 2016;44(D1):D840–D847.
23) Shimoyama M, De Pons J, Hayman GT, *et al.* The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. Nucleic Acids Res. 2015;43(Database issue):D743–D750.
24) Attrill H, Falls K, Goodman JL, *et al.* FlyBase: establishing a Gene Group resource for Drosophila melanogaster. Nucleic Acids Res. 2016;44(D1):D786–D792.
25) Howe KL, Bolt BJ, Cain S, *et al.* WormBase 2016: expanding to enable helminth genomic research. Nucleic Acids Res. 2016;44(D1):D774–D780.
26) Berardini TZ, Reiser L, Li D, *et al.* The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. Genesis. 2015;53(8):474–485.
27) Farrell CM, O'Leary NA, Harte RA, *et al.* Current status and new features of the Consensus Coding Sequence database. Nucleic Acids Res. 2014;42(Database issue):D865–D872.
28) Pruitt KD, Brown GR, Hiatt SM, *et al.* RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. 2014;42(Database issue):D756–D763.
29) Takeda J, Yamasaki C, Murakami K, *et al.* H-InvDB in 2013: an omics study platform for human functional gene and transcript discovery. Nucleic Acids Res. 2013;41(Database issue):D915–D919.
30) Kanehisa M, Sato Y, Kawashima M, *et al.* KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44(D1):D457–D462.
31) Imanishi T, Nagai Y, Habara T, *et al.* Full-length transcriptome-based H-InvDB throws a new light on chromosome-centric proteomics. J Proteome Res. 2013;12(1):62–66.
32) Clark K, Karsch-Mizrachi I, Lipman DJ, *et al.* GenBank. Nucleic Acids Res. 2016;44(D1):D67–D72.
33) Gibson R, Alako B, Amid C, *et al.* Biocuration of functional annotation at the European nucleotide archive. Nucleic Acids Res. 2016;44(D1):D58–D66.
34) Mashima J, Kodama Y, Kosuge T, *et al.* DNA data bank of Japan (DDBJ) progress report. Nucleic Acids Res. 2016;44(D1):D51–D57.
35) Cochrane G, Karsch-Mizrachi I, Takagi T. International Nucleotide Sequence Database C. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res. 2016;44(D1):D48–D50.

Proteome Letters 2016 ; 1 : 76

# Which Database to Use?

## —Confusions and Puzzles in Database Search and Sequence Analysis—

Akiyasu C. Yoshizawa*

*E-mail: acyshzw@kuicr.kyoto-u.ac.jp

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan

(Received: May 12, 2016; Revised: June 10, 2016; Accepted: June 14, 2016)

For mass spectrometry based-proteomics studies, computational analyses of obtained data are indispensable. However, the analysis methodologies and software for mass spectrometry and/or proteomics are currently still under development and many problems thus remain unfixed; consequently, researchers, especially experimental scientists, often suffer from technical issues and popular misinterpretations. Based on these problems, we describe in this review the computational processes for protein identification for proteomics beginners, especially the algorithms of database search and related basic issues: the comparison of *de novo* sequencing method and database search method, the effects of PTM detection on the search results, an overview of life science databases, and tips and cautions for their application to database searches.

***Keywords*:** bioinformatics; computational analysis; database search; identification; mass spectrometry.

**Table S1**

(a) Possible phosphorylation site in a trypsin-digested peptide

| # S, T, Y in a peptide | # peptide | # possible phosphorylated peptides |
|---|---|---|
| 47 | 1 | 140,737,488,355,327 |
| 46 | 0 | 0 |
| 45 | 0 | 0 |
| 44 | 0 | 0 |
| 43 | 1 | 8,796,093,022,207 |
| 42 | 1 | 4,398,046,511,103 |
| 41 | 0 | 0 |
| 40 | 0 | 0 |
| 39 | 1 | 549,755,813,887 |
| 38 | 1 | 274,877,906,943 |
| 37 | 0 | 0 |
| 36 | 0 | 0 |
| 35 | 0 | 0 |
| 34 | 2 | 34,359,738,366 |
| 33 | 2 | 17,179,869,182 |
| 32 | 9 | 38,654,705,655 |
| 31 | 7 | 15,032,385,529 |
| 30 | 5 | 5,368,709,115 |
| 29 | 5 | 2,684,354,555 |
| 28 | 11 | 2,952,790,005 |
| 27 | 19 | 2,550,136,813 |
| 26 | 21 | 1,409,286,123 |
| 25 | 29 | 973,078,499 |
| 24 | 36 | 603,979,740 |
| 23 | 47 | 394,264,529 |
| 22 | 71 | 297,795,513 |
| 21 | 82 | 171,966,382 |
| 20 | 123 | 128,974,725 |
| 19 | 178 | 93,323,086 |
| 18 | 248 | 65,011,464 |
| 17 | 314 | 41,156,294 |
| 16 | 454 | 29,752,890 |
| 15 | 651 | 21,331,317 |
| 14 | 1,054 | 17,267,682 |
| 13 | 1,472 | 12,057,152 |
| 12 | 2,163 | 8,857,485 |
| 11 | 3,227 | 6,605,669 |
| 10 | 4,654 | 4,761,042 |
| 9 | 7,010 | 3,582,110 |
| 8 | 10,683 | 2,724,165 |
| 7 | 16,400 | 2,082,800 |
| 6 | 25,210 | 1,588,230 |
| 5 | 38,736 | 1,200,816 |
| 4 | 61,802 | 927,030 |
| 3 | 100,238 | 701,666 |
| 2 | 158,211 | 474,633 |
| 1 | 208,023 | 208,023 |
| 0 | 157,343 | 0 |
| Total | 798,545 | 154,879,337,257,752 |

(b) Possible phosphorylation site in a trypsin-digested peptide, of which the number of phosphorylated sites are less than 13

| # S, T, Y in a peptide | # peptide | # possible phosphorylated peptides |
|---|---|---|
| 12 | 2,163 | 8,857,485 |
| 11 | 3,227 | 6,605,669 |
| 10 | 4,654 | 4,761,042 |
| 9 | 7,010 | 3,582,110 |
| 8 | 10,683 | 2,724,165 |
| 7 | 16,400 | 2,082,800 |
| 6 | 25,210 | 1,588,230 |
| 5 | 38,736 | 1,200,816 |
| 4 | 61,802 | 927,030 |
| 3 | 100,238 | 701,666 |
| 2 | 158,211 | 474,633 |
| 1 | 208,023 | 208,023 |
| 0 | 157,343 | 0 |
| Total | 793,700 | 33,713,669 |

**Table S2** Popular public databases for life science research and detailed introduction

アミノ酸配列データベース　　2016 年 5 月 12 日確認

| データベース名 | 正式名 | URL | 編纂者 | 配列数（生物種数） | 簡略説明 | 内容 |
|---|---|---|---|---|---|---|
| UniProt | Universal Protein Resource | http://www.uniprot.org/ | スイス・SIB & 欧・EMBL-EBI | — | UniProtKB，UniRef，UniParc の 3 データベースから構成される，タンパク質データベースの総称 | UniProtKB，UniRef，UniParc の 3 データベースから構成される，タンパク質データベースの総称．2002 年に，既存の 3 データベース（Swiss-Prot，TrEMBL，PIR）を統合した UniProtKB を中心に発足した． |
| UniProtKB | | http://www.uniprot.org/uniprot/ | | — | Swiss-Prot と TrEMBL から構成される，タンパク質データベース | 「KB」は Knowledge Base の略．「UniProt」と省略されることが非常に多く，事実 UniProt（全体）のメインコンテンツである．現在は，Swiss-Prot と TrEMBL の 2 データベースから構成される（PIR は実質的に Swiss-Prot に吸収）．各生物種毎の Proteome Dataset を準備しており，ダウンロード可能． |
| UniRef | | http://www.uniprot.org/uniref/ | | 79,568,127 [UniRef100] 41,730,393 [UniRef90] 17,048,127 [UniRef50] | UniProtKB 収録配列を類似性に基づいてクラスタリング | UniProtKB 収録配列に対して，配列類似性に基づいてクラスタリングし，そのクラスターをデータベース化している． |
| UniParc | | http://www.uniprot.org/uniparc/ | | 120,721,825 | UniProtKB の過去の情報を集積 | UniProt Archive の略．UniProtKB の過去の情報を網羅的に蓄積している．既知の全アミノ酸配列の履歴までをも含む． |
| Swiss-Prot | | http://www.uniprot.org/uniprot/?query=*&fil=reviewed%3Ayes | | 551,193 (10,401) | TrEMBL の配列をキュレータがアノテーションして作成．isoform レベルのアノテーション（高品質） | 1986 年に開発が開始された．配列に詳細なアノテーションを付与しており，現在では「TrEMBL の配列を，人間のキュレータが手作業でアノテーションし，重複がなくなるようにその結果を収録」している．また，isoform ごとにアノテーションが付けられている（＝どの配列がどの配列の isoform かを吟味して，情報を一つのエントリにまとめ，その旨分類明記されている）．翻訳後修飾情報も増加し続けており，一般的には充分な量の情報が収録されているが，翻訳後修飾専門のデータベースに比べれば収録件数は少ない． |
| TrEMBL | Translated EMBL | http://www.uniprot.org/uniprot/?query=*&fil=reviewed%3Ano | | 62,148,086 (474,979) | 旧 EMBL（現 ENA）の塩基配列をアミノ酸配列に翻訳．配列は GenPept と同一（の筈）．計算機による自動アノテーションで，遺伝子レベル | 旧 EMBL（現 ENA）データベースの CDS（Coding Sequence）情報を塩基配列からアミノ酸配列に変換したもので，NCBI の GenPept に相当する．アミノ酸配列自体は GenPept と同一（の筈）で重複がある．InterPro を用いた自動アノテーション，原核生物に対する HAMAP 自動アノテーションなど，比較的詳細なアノテーションが行われている．遺伝子対象の自動アノテーションであるため，isoform ごとのアノテーションはない（＝ isoform 配列も，独立した配列として無関係にエントリが作られている）． |
| Proteome Set | | http://www.uniprot.org/proteomes/ | | (49,790) | ゲノム決定済みの生物種のプロテオーム・データ | UniProt に含まれる，「ゲノム決定済みの生物種」のタンパク質の全データを，生物種毎に収録した．プロテオーム解析の場合には決定版のデータと言えるが，TrEMBL の配列もそのまま収録しているので，配列に重複がないことが保証されているわけではない． |
| neXtProt | | http://www.nextprot.org/db/ | スイス・SIB & スイス・GENEBIO | 20,055+ 41,992 isoforms (human) | ヒト版の“モデル生物データベース”を目指し，ヒトの配列・アノテーション全ての点で既知全情報の集約を目指す | 一般に“モデル生物データベース”では，そのモデル生物についての全遺伝子・トランスクリプト・タンパク質情報が網羅されるが，ヒト版のこのようなデータベースの構築を目指している．Swiss-Prot がヒトタンパク質の基本的なセットのアノテーションを既に完了しているので，そのデータを基本に，ゲノム・トランスクリプトーム・プロテオームの各レベルの情報，遺伝子変異や alternative splicing 情報，PTM 情報などを，ArrayExpress，UniGene，PeptideAtlas，COSMIC など多数のデータベースから収集し統合した．更に 80 万を超える配列 ID，CCDS や Affymetrix 社の ID，Illumina 社の DNA probe set ID までを収集し，タンパク質情報と関連づけた．更に個々のアノテーションについて，「タンパク質で実験的に確認」「トランスクリプトで確認」など，“品質ランク”を設定し表示している． |
| GenPept | | http://www.ncbi.nlm.nih.gov/protein | 米・NCBI | 193,739,511 | GanBank の塩基配列をアミノ酸配列に翻訳．配列は TrEMBL と同一の筈 | NCBI が，GenBank に収録された塩基配列の CDS を翻訳してアミノ酸配列に変換したもの．EBI の TrEMBL に相当．基本的には，アミノ酸配列自体は TrEMBL と同一（の筈）． |
| nr (nr-aa) | | http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome など | 米・NCBI<br>京都大化研（GenomeNet(*1)）など | 87,063,583 [NCBI] 87,055,864 [GenomeNet] | 複数のデータベース収録のアミノ酸配列を重複なしに収録した，検索対象用の配列コレクション | 塩基配列の nt (nr-nt) と同様，「複数のデータベースの配列を収録した multi Fasta ファイル」による配列コレクションである．主に「1 回の検索で，複数のデータベースの配列に検索をかけたのと同等の効果を得る」という目的で作成されており，一般的な意味でのデータベースではない．アノテーションは最初の decsription 行（> で始まる行）のみである．<br>NCBI 版は nr と呼ばれ，GenPept，PDB，Swiss-Prot，PIR，PRF の配列を収集し，WGS（Whole Genome Shotgun）データに含まれる“Environmental sample”を取り除いて non-redundant にしたもの．これに対して GenomeNet 版では GenPept，UniProt，Refseq，PDBSTR から配列を集めて non-redundant にしたものだが，実質的に内容はほぼ同一と考えて良い． |
| IPI | International Protein Index | http://www.ebi.ac.uk/IPI | 欧・EMBL-EBI | 327,465（7）［最終版，2011/9/27］ | 更新終了，UniProt に引き継がれる | ヒトゲノム決定の時期に，EBI の UniProt・Ensembl 研究者らが中心になって作成した，既知タンパク質配列の完全セット．UniProt，Ensembl，RefSeq から配列を収集し，重複のないように整理の上，代表的なデータベースへのリンクを付した．既知の（既に他データベースに記載されている）isoform については個別に配列を取得できるようにした．タンパク質を主眼とした（特に質量分析法での配列決定を念頭に置いた）データベースであるため，仮に異なった遺伝子から完全に同一配列のタンパク質が翻訳される場合でも，1 個の（そのタンパク質の）エントリしか作られていない． |
| PIR | Protein Information Resource | — | 米・Georgetown 大 | — | 更新終了，Swiss-Prot に引き継がれる | “最初に作られた配列データベース”NBRF の Atlas of Protein Sequence and Structure を電子化して始まったもので，特に分子進化的な観点からファミリー分類など詳細なアノテーションを付していた．現在は更新停止．研究グループは PIR の名称のままで UniProt などに参加． |

ゲノム情報に基づく **CDS** 対応データベース（塩基＆アミノ酸）　　2016 年 5 月 12 日確認

| データベース名 | 正式名 | URL | 編纂者 | 配列数（生物種数） | 簡略説明 | 内容 |
|---|---|---|---|---|---|---|
| RefSeq | The Reference Sequence | http://www.ncbi.nlm.nih.gov/refseq/ | 米・NCBI | Proteins: 61,034,675 Transcripts: 14,035,988 (58,776) | キュレータによってアノテーションされた塩基・アミノ酸配列 | 2003 年から作成開始．キュレータが（つまり人手で）GenBank 配列から冗長性のない配列を抽出して作成した，「標準配列によるゲノム網羅的なデータセット」を作成している．DNA，RNA，タンパク質それぞれの形でデータセットが準備されている．<br>分子生物学的な解析のために作成，というよりも「（公共財としての）基盤データの整備」を念頭に置いているため，配列の吟味が終わったものから順次エントリが追加されていくようになっており，生物種毎のファイルは提供されていない．<br>なお日本では「レフセック」と（この綴りをそのまま）読む人が多いようだが，筆者の知る限りアメリカのバイオインフォマティクス研究者は「レフシーク」と発音している． |

| | | | | | | |
|---|---|---|---|---|---|---|
| KEGG GENES | Kyoto Encyclopedia of Genes and Genomes | http://www.genome.jp/kegg/genes.html | 京都大化研 | 18,927,971 (5,224) | RefSeq などを元に，配列を生物種毎に整理．データを“精製”し，アノテーションとリンクを多数追加した | KEGG はパスウェイ・データベースとして特に有名だが，実際には単一のデータベースというよりも，複数の分子生物学データベースのスイートであり，KEGG 中で塩基/アミノ酸配列を参照する必要があるときには，全てこの GENES を参照している．配列は「最も信頼できるソースから収集」，アノテーションは「KEGG の編集方針に基づいて，自動及び人手で付与」特に「外部のデータベースへのリンクを豊富に付ける」という方針で編集されており，RefSeq 配列を採用していることが多い．<br>GENES の特徴は，データ整理ではなく分子生物学研究の視点から編纂されていることで，例えば 1 生物種の配列データが 1 ファイルにまとめられている． |
| Ensembl | | http://www.ensembl.org/（日本では http://asia.ensembl.org/ にリダイレクト） | 欧・EMBL-EBI | (86) [Ensembl]<br>(65) [metazoa]<br>(39) [plants]<br>(589) [fungi]<br>(158) [protist]<br>(29,777) [bacteria] | ゲノム全体をカバーする形で独自に遺伝子予測した．UniProt に対応する遺伝子データベースとして EBI が指定 | 各ゲノムプロジェクトから塩基配列データの提供を受け，そのデータから自前で遺伝子予測を行っている．予測方法も，通常と違って ***ab initio*** と呼ばれる「統一的な基準に従って一から予測する」方法を用いており，遺伝子予測の方法論を改良する舞台としても用いられている．RefSeq のような「配列の精度を上げる」という編集方針とは違い，「決定された全ゲノム配列全体から遺伝子を予測する」という方針をとり，カバー率が高いことから，EBI は（UniProt に対応する）公式の標準的遺伝子データベースとしてこの Ensembl を選んでいる．核酸のデータとタンパク質のデータの両方を公開している．<br>現在は生物種の taxonomy ごとに“姉妹”サイト 6 つに分かれている．<br>なお，データベース名の読み方は「アンサンブル」であるが，通常のアンサンブルの綴りは ensemble で，この名称は最後の e が抜いてある（EMBL と掛けた命名）． |
| CCDS | Consensus CoDing Sequence | https://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi | 米・NCBI/英・Sanger 研究所 | 31,292 (human)<br>24,788 (mouse) | NCBI と Ensebml の両方で予測された CDS の共通部分に，共通の ID を付与，「タンパク質コード遺伝子に高品質のアノテーションを付した完全セット」を目指す | ゲノムからの遺伝子予測は，NCBI と EBI/Ensembl において自動で，Sanger 研と NCBI/RefSeq において curator によって行われているが，これらの作業は全て独立に進んでいる．このため遺伝子予測の結果は，よく似ているが完全に一致するわけではない．そこで，「タンパク質コード遺伝子に高品質のアノテーションを付した完全セット」を作成することを目的に，ヒトとマウスの CDS について，NCBI のゲノム・リソースと Ensembl の結果を比較し，共通の ID（CCDS ID）を付けている．また，Sanger 研究所と NCBI の RefSeq グループから，それぞれ curator によるアノテーション情報の提供を受けている． |
| H-inv | H-invitational | http://www.h-invitational.jp/hinv/ahg-db/index.jsp | 産総研 / 東海大・医 | protein coding transcripts: 196,619 (human) | 詳細なアノテーションのある，ヒト mRNA データベース | 2004 年から公開している，ヒトの遺伝子と転写産物を対象としたデータベース．Splicing variant や機能性 RNA，タンパク質の機能ドメイン，細胞内局在等々多数のアノテーションを付加している．外部データベースへのリンクも非常に豊富である．一方，トランスクリプト段階，即ち splicing variant までの収録であるため，isoform や PTM の研究には直接的には向いていない．<br>2002 年に JBIRC が行った，『ヒト完全長 cDNA アノテーション・国際招待会議（H-invitational）』で，“コミュニティ・アノテーション並み”のアノテーションを実現するために，海外から多数の専門家を招聘した，というところからこの独特な名称が生まれている． |
| H-EPD | H-inv Extended Protein Database | http://hinv.jp/hinv/h-epd/ | | 40,367 (human) | H-inv と RefSeq と UniProt を merge して重複を除いたもので，特に missing protein 探索を念頭に置いて作成された | 特に HUPO の C-HPP 計画（Chromosome-centric Human Proteome Project）で利用することを念頭に，“missing protein”探索を一つの大きな目的として作られたもので，H-inv と RefSeq と UniProt を merge して重複を除いている．H-inv に登録された alternative splicing variant の情報や豊富なアノテーションがそのまま利用できる． |

塩基配列データベース　　2016 年 5 月 12 日確認

| データベース名 | 正式名 | URL | 編纂者 | 配列数（生物種数） | 簡略説明 | 内容 |
|---|---|---|---|---|---|---|
| GenBank/ENA (EMBL)/DDBJ | GenBank/ European Nucleotide Archive/DNA Databank of Japan | http://www.ncbi.nlm.nih.gov/genbank/ http://www.ebi.ac.uk/ena http://www.ddbj.nig.ac.jp/index-j.html | GenBank: 米・NCBI/ ENA: 欧・EMBL-EBI/ DDBJ: 遺伝研 | 193,739,511 [GenBank] 724,619,242 [ENA] 191,094,643 [DDBJ] | 研究者が自ら登録する，塩基配列レポジトリ | 国際塩基配列データベース．1970 年代末から 80 年代前半に構築が開始された．発足当初は，データベース・センターが学術ジャーナルからデータを抽出していたため，3 データベースでジャーナルを分担していた．その後（データの増加に伴って）個々の研究者が自分で自分のデータを登録する方式（レポジトリ方式）に変更，その際に地理的な分担に移行した．<br>このため，各データの著作権は登録した個々の研究者にあり，全く同一の遺伝子に対して別個の研究が行われ，それに対応するエントリが生成された場合でも，エントリの編集・統一などができない．この結果，多くの重複データを含んでいる．<br>なお，現在ではインターネット経由で相互コピーを行うため，3 データベースの内容に大差はない．概ね DDBJ は GenBank と同じ構成をとっている（web 版のインターフェースは EBI と同じものを使っている）が，EBI は最近大きく構成を変え，EMBL データベースとして知られてきた塩基配列関係のデータベースをはじめ，NGS の read データなど複数を，ENA の名で統合した．3 機関は INSDC（International Nucleotide Sequence Database Collaboration）として連携している． |
| Entrez Gene | | http://www.ncbi.nlm.nih.gov/ | 米・NCBI | — | NCBI の全データベースの検索インターフェース（データベースとしては，Gene ID に基づいた NCBI の全情報の整理） | Entrez は本来，（データベースの名称ではなく）NCBI の配列検索システム，その後は web インターフェースの名称だったが，2000 年代後半から NCBI の web サイト全体の統一的な検索エンジンに昇格した．RefSeq の編集が進んだ結果，そのデータに基づいて GenBank のデータを整理する作業が開始され，各遺伝子に Gene ID をアサインし，それをキーにして各 GenBank エントリの重複を整理，更に関連するゲノムマップ，発現情報など NCBI データベース中の情報へのリンクを収集したのが Entrez Gene である（従って Entrez というデータベースがあるわけではない）．<br>なおこの名称 Entrez は，フランス語の動詞 entre（英語の enter）を二人称複数に対する命令形にした場合の活用形で，「お入りなさい」「ここから入場」くらいの意味になる．読み方は活用のない場合と同じで「アントレ」． |
| nt (nr-nt) | | http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome など | 米・NCBI<br>京都大化学研（GenomeNet）など | 36,444,514 [NCBI] 185,260,177 [GenomeNet] | 複数のデータベース収録の塩基配列を収録した，検索対象用の配列コレクション | アミノ酸配列の nr (nr-na) と同様，nt は，NCBI や GenomeNet [※1] など“大手”の公共データベースサイトでサービスされている，「複数のデータベースの配列を収録した multi Fasta ファイル」による配列コレクションである．主に「検索の対象」としてのみ作成され，一般的な意味でのデータベースではない．アノテーションは各配列の最初の行（> で始まる description 行）のみである．<br>NCBI 版は nt と呼ばれ（nr/nt と書かれていることもある），GenBank，EMBL，DDBJ，PDB [※2]，RefSeq から配列を集め，EST，STS，GSS，WGS などを除いて non-redundant である．<br>これに対して GenomeNet 版では GenBank，EMBL，DDBJ，RefSeq から配列を集めて non-redundant である．GSS などの巨大容量のデータを含むため，GenomeNet 版のほうが収録配列数は多い． |

注釈

| 番号　名称 | 正式名 | URL | 運営者 | | 簡略説明 | 内容 |
|---|---|---|---|---|---|---|
| (*1) GenomeNet | | http://www.genome.jp/ | 京都大化研 | — | DDBJ と並ぶ，日本の二大生命科学データベース・サイト．生命科学データベースのミラーや，独自コンテンツの KEGG などを提供する（KEGG 側から見れば，「KEGG の入れ物」） | 歴史的経緯もあって「Net」という名称が用いられているが，少なくとも現在の実体は単一のデータベース・サービス・サイト（データベースなどの“入れ物”）である．公共データベースをミラーリングする（原サイトと協力して，コピーを公開することで原サイトへのアクセス集中などを予防する）ほか，自前のデータベースの KEGG などを公開し，多数の web 版バイオインフォマティクス・ツールもサービスしている．<br>なお，検索インターフェース dbget が，独自コンテンツ KEGG やミラーコンテンツ全てに対する統合的なインターフェースとなっており，これはちょうど NCBI の Entrez に相当する． |
| (*2) PDB (wwPDB) | (worldwide) Protein DataBank | http://www.wwpdb.org/ | Rutgers 大学など多数が関与 | — | 主にタンパク質の立体構造データベース | Brookhaven 国立研究所で作成が始まり，その後 Rutgers 大学に引き継がれた RCSB PDB と，大阪大学蛋白研などが支援する PDBj など 4 つのデータベースが連携する，世界最大のタンパク質立体構造のデータベース（現在はタンパク質以外の，例えばウイルス粒子の構造なども収録されている）．立体構造決定に向いている特殊な（生物種の）タンパク質が収録されていることもあり，PDB にのみ収録されている配列も存在する． |