どう活かす他人のデータ
―バイオインフォマティクス活用法―

髙橋　広夫[1]*・岩川　秀和[2]・尾之内　均[3]・小島　晶子[4]・町田千代子[5]

近年のマイクロアレイ技術や次世代シーケンサーなどの急速な技術革新により，一度に観測できる生物情報は膨大なものとなり，生物情報データベースはものすごい勢いで成長を続けている．このような膨大なデータの山の中から重要な情報（つまり宝）を見つけ出すための手段が，バイオインフォマティクスという学問である．また，有用な情報を抽出する過程をデータマイニングとも呼ぶ．

生物工学会の会員に限らないが，バイオインフォマティクスと統計学の関係について誤解している人が多くいるようなので，まず，ここから整理したい．バイオインフォマティクスを日本語に直訳すると生物情報学（あるいは，生物情報科学）という言葉になる．それでは，情報学と統計学は違うのであろうか．まず，情報学とは，情報をどのように扱うかについて考え研究する学問である．しかし，どこまでが情報学であるのかの境界は非常に曖昧である．一方，統計学とは，確率を基礎にして，不確実性を含むデータから，一定の確実さをもった判断を下す学問である．情報学の中で統計学的な考え方は多用されているので，筆者から見ると情報学に統計学が含まれるように思うのだが，生粋の統計学者から見ると統計学と情報学は相容れぬ考え方だそうだ．

前置きが長くなったが，筆者はバイオインフォマティクスとは，統計学や情報学を使って膨大なデータを読み解く学問であると考えている．本編では，生物情報データベースやバイオインフォマティクスの成り立ちなど基本的なところから，膨大なデータをバイオインフォマティクスで解析するときに注意すべきこと，どのような成果が得られるのかまで，筆者の経験から例をあげて説明したい．

ウェブサイトに公開された生物情報データベース

データベースとは，特定のテーマに沿ったデータを集めて管理しているもので，容易に検索・抽出などの再利用が可能なものであり，狭義にはコンピュータによって実現されたものを指す．生物情報データベースは，おもにウェブサイト上で構築されたもので，歴史的には各国の各研究機関がそれぞれ構築したものが多かったが，現在では統合された大規模なデータベースとなっている．生物情報データベースには，塩基配列情報データベース，アミノ酸配列情報データベース，タンパク質構造情報データベース，遺伝子発現情報データベース，タンパク質機能データベース，代謝ネットワークデータベースなどがある．

塩基配列情報データベースとしては，欧州バイオインフォマティクス研究所（European Bioinformatics Institute：EBI）が管理するEuropean Molecular Biology Laboratory（EMBL）[1]，米国国立バイオテクノロジー情報センター（National Center for Biotechnology Information：NCBI）が管理するGenBank[2]，日本の国立遺伝学研究所（National Institute of Genetics：NIG）が管理するDDBJ（DNA Data Bank of Japan）[3]が存在するが，これらは三大国際塩基配列データバンクと呼ばれ，相互にデータがリンクされたInternational Nucleotide Sequence Database Collaboration（INSDC）という一つのデータベースなっているため，日本人が利用する場合は，国立遺伝学研究所の公開しているDDBJを使うのが一番便利であろう．DDBJの登録総塩基数の推移を調べたところ，図1のように指数関数的というほどではないが，年々かなり速いペースでデータベースが肥大していることが分かる．

アミノ酸配列情報データベースでは，歴史的に，Protein Information Resource（PIR），SWISS-PROT，TrEMBLが存在していた．しかし，PIRはSWISS-PROTに吸収され，さらに，SWISS-PROTとTrEMBLが合併し，Universal Protein Resource（UniProt）[4]になった．SWISS-PROTは，研究者が人手でハイレベルのアノテーション（付加情報）をつけたアミノ酸配列データベース

*著者紹介　[1]千葉大学大学院園芸学研究科応用生命化学領域（准教授）　E-mail: hiro.takahashi@chiba-u.jp
[2]奈良先端科学技術大学院大学バイオサイエンス研究科（博士研究員）
[3]北海道大学大学院農学研究院応用生命科学部門応用分子生物学分野（准教授）
[4]中部大学応用生物学部環境生物科学科（講師），[5]中部大学応用生物学部応用生物化学科（教授）

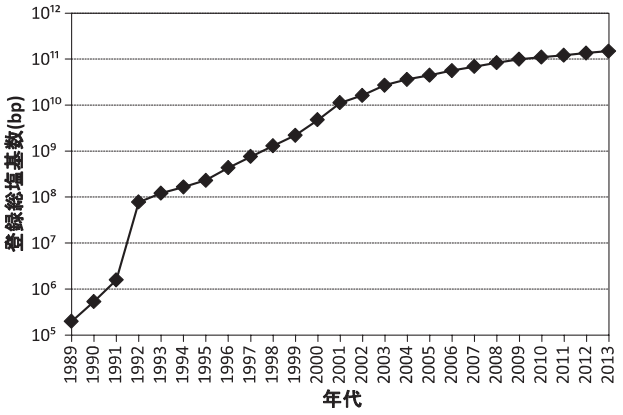

図1．DDBJの登録総塩基数の推移

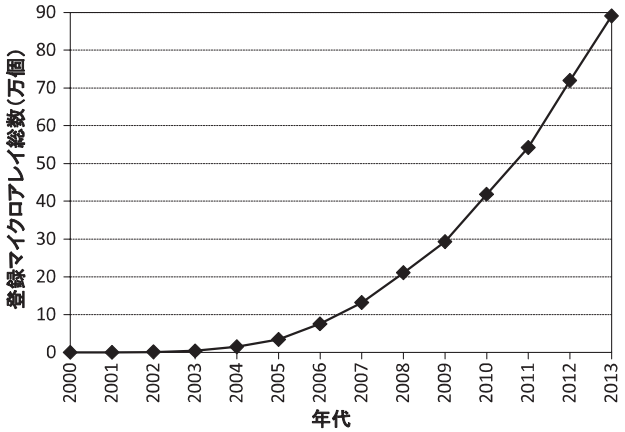

図2．GEOに登録されたマイクロアレイ数の推移

であるが，TrEMBLは，コンピュータにより自動でアノテーション（付加情報）をつけたアミノ酸配列データベースであり，性質が異なることから，現在でもUniProtの中で区別できるようになっている．その他，タンパク質研究奨励会が管理しているProtein Research Foundation（PRF）[5]や，タンパク質構造データベースであるWorldwide Protein Data Bank（PDB）[6]にもアミノ酸配列情報が含まれるために，UniProtと合わせてPRFとPDBも使われることが多い．

遺伝子発現データベースとしては，NCBIのGene Expression Omnibus（GEO）[7]やEBIのArrayExpress[8]が有名である．塩基配列の場合と同様，論文に投稿する場合，マイクロアレイデータのデータベースへの登録が義務づけられている場合が多いので，現在では膨大なマイクロアレイデータがこれらのデータベースに登録され続けている．GEOに関する登録マイクロアレイ総数の推移を調べたところ，図2のように，毎年，リニアにデータベースが肥大していることが分かる．

バイオインフォマティクスの成り立ち

バイオインフォマティクスは，歴史的には1980年代前半に，ヨーロッパのEMBL（1980年），アメリカのGenBank（1982年），日本のDDBJ（1984年）という3大塩基配列データベースが設立されたことに端を発して発達してきた．1980年代後半から1990年にかけて，現在でもよく使われる配列解析アルゴリズムのFASTAアルゴリズム（1985年）やBasic Local Alignment Search Tool（BLAST）アルゴリズム（1990年）が発表された．また，同じく1990年にヒトゲノム計画が発足し，このプロジェクトでバイオインフォマティクスが重要な役割を果たした．そして，今日では，膨大な生物情報を解析するために，バイオインフォマティクスが不可欠の分野となった．

遺伝子発現データの再現性とバイオインフォマティクスによるデータ品質管理

マイクロアレイ技術の進歩によって遺伝子発現データの同時観測が可能となり，年々コストも下がってきて，1回あたり5～10万円で実験ができるような時代となった．「とにかくマイクロアレイの実験をやってみよう．何か見えてくるかもしれない．」という考えで，実験をやってしまったという事例をよく聞くが，まず，マイクロアレイの実験を行う前に，ポジティブコントロールになる遺伝子の発現の検証を，Reverse transcription polymerase chain reaction（RT-PCR）などを用いて行っておいた方がよいだろう．また，バイオインフォマティクスは万能であり，質の悪い遺伝子発現データから質の高いデータに変換できる，と思っている研究者が少なくないが，これは半分迷信である．質の悪いデータからは，良い情報は得られない．最大限努力して良いデータを得た上で，バイオインフォマティクスによるさまざまな手法を駆使して，古典的な方法では抽出できないような有用情報を得ようというのがバイオインフォマティクス流の正しい道である．

他人から得たデータについては，もちろん実験計画の悪さはデータを眺めただけでは見えてこない．しかし，データの質の不均一性というのは，統計学的指標である相関係数を用いた手法などで，ある程度の評価が可能である．この方法について図3のような例で説明したい．野生株，変異株1，変異株2についてマイクロアレイデータを取得したとする．マイクロアレイデータでは，数千から数万以上の遺伝子発現データを同時に取得することが可能である．取得した遺伝子発現データについて，すべてのペアにおける相関係数を計算する．図3Aでは，

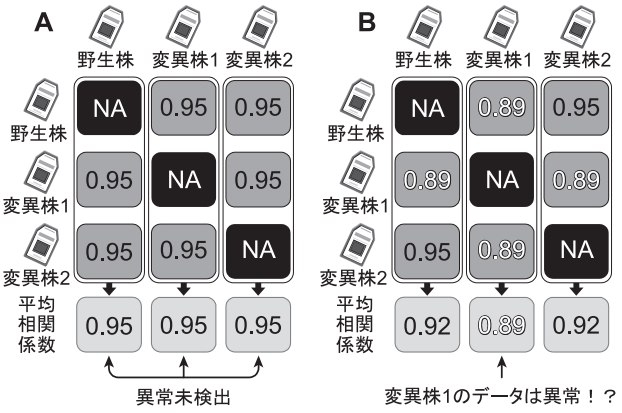

図3. 相関係数法を用いたマイクロアレイデータの品質チェック法

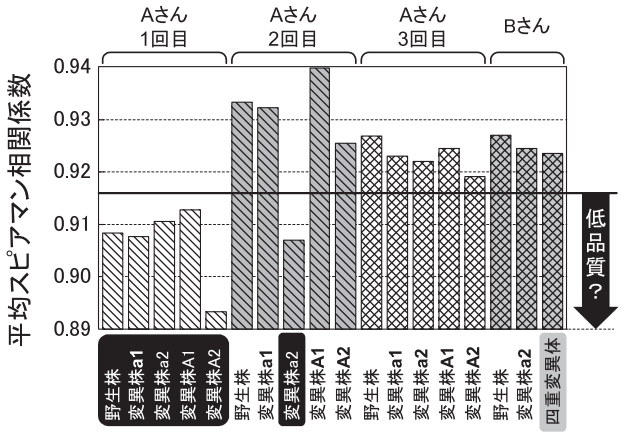

図4. 実験や実験者間におけるアレイデータの平均相関係数の比較

それぞれ計算された相関係数が，0.95になったことを示している．この場合，それぞれのマイクロアレイに対する平均相関係数を計算すると，すべて0.95であり，特にマイクロアレイ間で違いが見られない．一方，図3Bでは，変異株1のデータが異常であった場合を想定している．この場合，野生株，変異株1，変異株2のそれぞれのマイクロアレイの平均相関係数を計算すると0.92，0.89，0.92となり，変異株1に対する平均相関係数だけが低い値をとる．この変異株1のデータのように，特に際だって小さな平均相関係数を示すものがあれば，このデータは他のデータとは，なにか質的に違うデータになっていることが示唆される．ここで注意すべきは，質的な違いが必ずしも実験の失敗を意味するわけではなく，あくまでもデータの質の違いがある可能性が示唆されるだけである．したがって，実験を失敗したと断定する必要はないが，いったい何が違ってこういった質的違いが生じたのかについて，十分な議論がなされるべきであろう．

前述の相関係数を用いた方法でマイクロアレイデータの質を評価している研究例[9]を紹介したい．この研究では，データの品質チェックを行い，3回目のデータのみ用いている．図3に示した方法で，18枚分のマイクロアレイデータのすべてのペアについて，相関係数を計算して，各マイクロアレイデータに対する平均相関係数を計算した（図4）．相関係数を求める手法には，ピアソンの積率相関係数やスピアマンの順位相関係数が有名であるが，今回はパラメトリックでノイズに強いスピアマンの順位相関係数を用いている．また，相関係数の算出に用いた遺伝子は，18枚分のマイクロアレイデータのうち，すべてのマイクロアレイでノイズより高いシグナル値が観測できた遺伝子のみを用いている．左から，1回目のAさんの野生型と4種類の遺伝子の変異体の計5マイクロアレイ分のデータが並んでいる．そして5マイクロアレイ分の実験をさらに2セット行った実験と，Bさんが行った，4つの遺伝子を破壊した四重変異体，a2変異体，および，野生型の3マイクロアレイの結果を示している．左から5つ分がもっとも初期に実験を行ったもので，RNA抽出などのプロトコールが試行錯誤的で未完成だったこともあり，全体的に低い値である．2回目の実験は全体的には非常に高めだが，a2変異体の数値だけが低い．3回目にはプロトコールが洗練され，数値が安定したと解釈できる．この確立されたプロトコールで実験したBさんのマイクロアレイデータは，四重変異体，a2変異体，および，野生型ともに安定して高い平均相関係数を示している．いったいどの数値の平均相関係数だったらダメとするか，あるいは良いとするかについて明確な基準はないが，この結果を見る限り，1回目の5サンプルと2回目のa2変異株のデータは他と比べて異質なデータであり，解析には使わないほうがよいのではないかという判断ができる．ちなみにこの実験では，植物の茎頂部（茎頂分裂組織と葉原基を含む）を切り取ってRNA抽出を行っているが，AさんとBさんのサンプリングに使った植物は，種を蒔いてからの日数が若干異なる．それにもかかわらず，Aさんの3回目とBさんのデータでは，ほぼ同じ相関係数であるので，1回目の5サンプルと2回目のa2変異株のデータは，サンプリングの日数の違い以上に何かが異なることを意味している．

また，ここから，必ずしも野生型同士の比較でないといけないというわけでもないことが言える．ここに示したのは一例であるが，同じプロトコールであっても，ラボが違えば相関係数が低くなることも多いし，同一ラボであっても，サンプリングの時期が1年違うだけで相関係数が低くなることもある．必ず相関係数を使った方法

で検証する必要があるというわけではないが，解析をする前に何らかの方法でデータの質の均一性をチェックし，比較可能なデータであるかを確認するべきであろう．

マイクロアレイデータのアノテーション情報の信頼性

近年のDNAマイクロアレイの市場は，アフィメトリクス，アジレント・テクノロジー，イルミナの3社でおよそ80–90%のシェアを占めている．各メーカーが生産しているマイクロアレイは，それぞれメーカーごとにexpressed sequence tag（EST）データベースなどをもとにしてマイクロアレイのプローブ設計を行い，独自にアノテーション情報を公開・更新している．その結果，マイクロアレイの設計当時は未知遺伝子であった情報が更新されて，機能が明らかになる場合がある．一方で，せっかく注目していた遺伝子（マイクロアレイ上のプローブ）が，アノテーション情報が更新されたために，ゲノム上にマップできない転写物を検出するためのプローブ，もしくは，複数の遺伝子とクロスハイブリダイゼーションするプローブであったということもしばしば起こる．最近，筆者がよく解析していたマイクロアレイのひとつであるアフィメトリックス製GeneChip Arabidopsis ATH1 Genome Arrayについて調べてみたところ，搭載されているプローブセット22,746個のうち，なんと約10%にあたる2177個が，クロスハイブリダイゼーションが生じる，もしくは対応する転写物が存在しないという状況であった．また，Arabidopsis Tiling 1.0R Arrayについては，設計時には320万か所のゲノムに対応するプローブが搭載されているが，ゲノム情報の更新やクロスハイブリダイゼーションにより，10%にあたる32万か所の情報がうまくゲノムにマップできなくなった．つまり，残った288万か所のデータのみが利用可能ということである．このように，生物情報データは日々更新されているものであるため，常に新しい情報に気を配っておく必要がある．

塩基配列データベースを用いたデータマイニング

データマイニングとは，データの山からコンピュータを用いて有用情報を抽出することを指す．生物情報データベースには膨大な量のデータが蓄積されており，多くは無料で利用できる反面，信頼性については誰も保証しない場合が多い．筆者らは，真核生物のmRNAの5′ untranslated region（5′UTR）に存在するupstream open reading frame（uORF）のうち，種間で進化的に保存されているものを同定するために，DDBJ/GenBank/EMBLの配列データを横断的に解析し，進化保存的なuORFを抽出するための新手法，BLAST-based algorithm for the identification of upstream open reading frames with conserved amino acid sequences（BAIUCAS）法を開発し，モデル植物であるシロイヌナズナに応用した．その結果，今までに報告のない新規の進化保存的な18個のuORFペプチドを発見した[10]．この研究では，使われているデータベース間の矛盾を考慮しながら有用情報のマイニングを行っているので，例として説明したい．

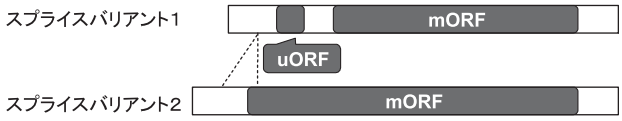

図5．スプライスバリアント間におけるuORFの位置関係

まず，データベースThe Arabidopsis Information Resource（TAIR）[11]から，5′UTRとCDS配列情報，ゲノム上のポジション情報を取得した．ここからさらに5′UTR上に存在する開始コドンと終止コドンのセットを検出することで網羅的にuORFを抽出する．真核生物では，遺伝子の10–50%ほどの5′UTRにuORFが存在すると見積もられている[12–15]．そのうち数十%ほどのuORFが，alternative splicingによって生じたもので，図5のように，別のsplice variantのタンパク質コード領域と，ゲノム上でオーバーラップしている．このようなuORFは，仮に進化的に保存されていても，別のsplice variantにコードされるタンパク質の機能的な重要性のために見かけ上uORFが保存されているだけかもしれない．もちろん，このタイプのuORFの機能的重要性の可能性も否定はできないが，最終的にウエットな実験で確認する目的があったので，より可能性の高いもののみを抽出するために，このタイプのuORFは解析から除外した．

ここでは，詳細な説明を省略するが，BAIUCAS法の6つのステップをパスしたuORF候補が，種間で保存されたuORFの候補として抽出された．しかし，この中のいくつかの候補は，UniProtに登録されたタンパク質の一部と相同性があることが分かった．つまり，UniProtに登録されているタンパク質にもかかわらずTAIRに登録されていないか，もしくは，TAIRに登録されているがゲノムポジションが間違って登録されているというものである．こういった片方のデータベースの不備をカバーするために，BAIUCAS法の最終ステップの後に，手作業でuORFの塩基配列をBLASTxを用いてUniProtになげ，main ORF（mORF）由来のタンパク質がヒットしたuORFは候補から除去した（図6）．さらにBLASTxだけでは，シロイヌナズナのもとの遺伝子由

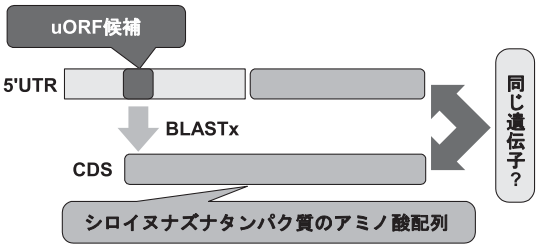

図6．データベース間における矛盾

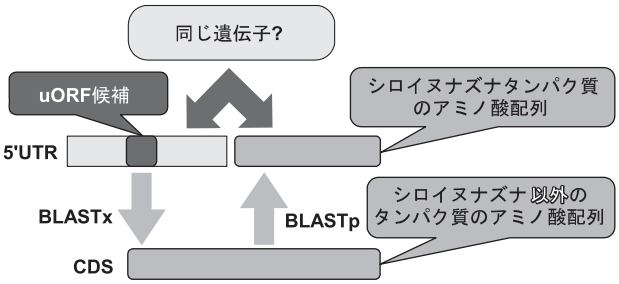

図7．uORFとmORFが他の生物で融合タンパク質を形成する場合

来のタンパク質にヒットしなかったものの，BLASTxでヒットしたタンパク質配列をBLASTpでなげて，もとのシロイヌナズナのタンパク質に戻れたものも候補から除外した．このようなタイプのuORFは，シロイヌナズナではuORFになっていても，シロイヌナズナ以外の植物では，uORFとタンパク質が融合タンパク質になっていることを意味している（図7）．もしかしたら，こういったタンパク質はデータベースに相同なシロイヌナズナのタンパク質が，まだ登録されていないだけであるということも考えられる．

以上のように偽陽性になりうる候補を注意深く除去しながら抽出した結果，筆者らは，前述のように18個の進化的に保存されたuORFペプチドを発見した．さらに，このうちの少なくとも7個は，下流のmORFの翻訳制御機能を有していることが実験的に確認できた（未発表）．残り11個については，mORFの翻訳制御機能の有無について十分検証できていないが，別のタイプの何らかの機能性ペプチドをコードしているuORFなのかもしれない．このように，データベースに登録されている情報の多くは他人のデータであるが，間違いがあることを前提に注意深く使用すれば非常に有用なものも多い．うまく他人のデータを利用することができれば，実験にかかる必要や時間を大幅に減らし，新発見の手助けをしてくることが期待できる．

マイクロアレイデータを用いたデータマイニング

シロイヌナズナの葉の形態形成を制御している*ASYMMETRIC LEAVES1*（*AS1*）と*AS2*遺伝子に関する変異体の茎頂部のマイクロアレイデータを用いた，発現データのデータマイニングに関する成功解析例[16]を紹介したい．この研究では，2種類の発現データセットを用いている．一つ目のデータセットは，*AS1*と*AS2*に関する単一変異体のマイクロアレイデータ（データセットA）である．二つ目のデータは，*AS1*と*AS2*の変異株の表現型を亢進するようなモディファイア遺伝子の変異株との二重変異体に関するマイクロアレイデータ（データセットB）である．どちらのデータセットも，Knowledge-based fuzzy adaptive resonance theory（KB-FuzzyART）[9]を用いてクラスタリング解析している．前述のように，マイクロアレイデータの質は些細なことで大きな影響を受けやすい．相関係数解析を用いても，データセットAとデータセットBの間で大きな質的違いなどは見られなかったが，単一変異体と二重変異体という違いがあるため，結果の統合を行うときにいったんKB-FuzzyARTでクラスタリング解析を行い，その解析結果を統合した．その結果，葉の形態を制御する*AS1*と*AS2*およびモディファイア遺伝子との共通の下流遺伝子の候補として，57遺伝子が抽出された．この中には，すでに*AS1*と*AS2*の下流遺伝子であることが示されている*BREVIPEDICELLUS*（*BP*），*AUXIN RESPONSE FACTOR3*（*ARF3*），*ARF4*，*KANADI2*，*YABBY5*が含まれていた．*BP*はAS1とAS2によって直接抑制される遺伝子であり，茎頂分裂組織の維持に関わるホメオボックス遺伝子である．また，最近，我々は*ARF3*がAS1とAS2のダイレクトターゲットであること，AS1とAS2は低分子RNAを介して*ARF3*と*ARF4*との抑制をしていることも明らかにした[17]．すなわち，57遺伝子の中に，確かにこれらの遺伝子が含まれていたことから，我々はこれらの遺伝子の下流で起こると予測される分化から分裂へのスイッチを担う実行部隊の遺伝子も57遺伝子に含まれると期待した．実際に，57遺伝子中には，細胞周期進行制御における重要な因子CDK inhibitor（Cyclin-dependent protein kinase inhibitor）をコードする*KIP-RELATED PROTEIN 2*（*KRP2*）/*INHIBITOR OF CYCLIN-DEPENDENT KINASE 2*（*ICK2*），*KRP5*/*ICK3*が含まれていた．さらに，*KRP2*/*ICK2*と*KRP5*/*ICK3*は，遺伝学的な実験から，*AS1*と*AS2*の下流遺伝子の中でももっとも重要な遺伝子セットである*ARF3*と*ARF4*のさらに下流にあることを明らかにすることができた．

一般的にバイオインフォマティクスの目的は，有用

情報の濃縮や判定・分類を行うことである．各バイオインフォマティクス手法の性質をよく理解した上で，適切にデータを扱うことにより信頼性の高いデータ抽出が可能となる．

おわりに

10年ほど前は，まだまだバイオインフォマティクスは未成熟で，なおかつ，ドライ研究者とウエットの研究者の連携が不十分であったが，近年は，バイオインフォマティクスはだいぶ成熟してきて，ドライ研究者とウエット研究者がうまく連携して良い成果がだせたという話も聞くようになった．しかしながら，ドライ研究者からウエット研究者へ，ウエット研究者からドライ研究者へといったように，一方通行の連携が多くみられる．将来的には，相互に情報のフィードバックを行いながら，綿密に研究を進めることで，より優れた研究が行われることが期待される．

また，ウエットの研究者が使いたがるようなシンプルで使いやすい，たとえばBLASTのようなソフトはまだ十分充実していないように思われる．さらに，エクソンアレイ，タイリングアレイ，RNA-Seqのように，新しい技術が開発されるたびに，改良したバイオインフォマティクス手法が必要となるというサイクルを繰り返していることから，しばらくはこれからもバイオインフォマティクスの発展が求められ続けるだろうと思われる．

最後に，本編の研究を進めるにあたり，ウエットの研究者の立場から，ドライ研究者との相互連携の重要性について理解を示し，多大な協力頂いた名古屋大学特任教授町田泰則先生と北海道大学教授内藤哲先生の両先生に，感謝の意を述べたいと思います．

文献

1) http://www.ebi.ac.uk/embl/
2) http://www.ncbi.nlm.nih.gov/genbank/
3) http://www.ddbj.nig.ac.jp/
4) http://www.uniprot.org/
5) http://www.prf.or.jp/
6) http://www.wwpdb.org/
7) http://www.ncbi.nlm.nih.gov/geo/
8) http://www.ebi.ac.uk/arrayexpress/
9) Takahashi, H. *et al.*: *J. Biosci. Bioeng.*, **106**, 587 (2008).
10) Takahashi, H. *et al.*: *Bioinformatics*, **28**, 231 (2012).
11) http://www.arabidopsis.org/
12) Churbanov, A. *et al.*, *Nucleic Acids Res.*, **33**, 5512 (2005).
13) Galagan, J. E. *et al.*: *Nature*, **438**, 1105 (2005).
14) Kawaguchi, R. and Bailey-Serres, J.: *Nucleic Acids Res.*, **33**, 955 (2005).
15) Rogozin, I. B. *et al.*: *Bioinformatics*, **17**, 890 (2001).
16) Takahashi, H. *et al.*: *Plant Cell Physiol.*, **54**, 418 (2013).
17) Iwasaki, M. *et al.*: *Development*, **140**, 1958 (2013).