大規模なゲノム情報の活用

町田　雅之[1]*・浅井　潔[2]・梅村　舞子[1]

大規模なゲノム情報とは

近年は，「生物情報」という言葉に対して違和感を覚えることはほとんど無くなったが，大規模と言えるような生物情報が得られるようになったのは，シークエンサーが登場してからではないかと推測する．生物情報が大量に得られるようになるまでは，「理論生物学」といっても現実的でないと考える人も多かったが，現在では，蓄積の進む生物情報を利用して，生理状態などの生命現象を計算上から理論的に推論することは当然ともいえるようになっている．大規模生物情報の代表格であるDNA塩基配列解析のスループットの向上は凄まじく，わずか20年あまりの間に100万倍以上に達している（図1）．筆者らはこの間，各世代のシークエンシング方式を用いて自ら解析するとともに，生物学の変革を目の当たりにするというスリリングな状況に遭遇する幸運に恵まれた．しかもそれはいまだに終わっていないのである．

DNA塩基配列に始まった生物情報解析は，現在では，発現情報，タンパク質情報，代謝物情報など，いわゆるオミックス情報といわれるように，情報量の増大だけでなく多様性も増している．また，細胞のイメージングとその情報処理も進みつつあり，診断技術への応用も進むと期待されている．これらの中で，情報の質と量，情報処理技術のすべてにおいてもっとも急速に進展しているのがDNA塩基配列であろう．細胞のイメージングに代表される生体画像処理も，撮像素子，情報処理速度，メモリ容量など，技術革新が激しい現在のデジタル技術を用いれば，細胞などの立体のイメージングと基本的な解析はPCでも可能である．しかし，得られた情報から生物学的な意味を取り出すことは，現代の計算機技術をもってしても容易ではない．一方，DNA塩基配列は，最初からATGCという4種文字にデジタル化された情報であり，文字列処理や統計解析に関しては計算機科学の分野で多くの理論，アルゴリズムが存在していた．さらに近年は，接尾辞配列に代表される効果的なデータ構造を活用した超高速アルゴリズムが適用されるようになっている．多様な生物情報の中でもDNAシークエンスに関する技術が際だって進歩したことは，これを可能にする実験的な解析技術が存在したことも理由のひとつと考えられるが，データ取得後の超高速な計算処理が可能であること，これにより意味のある重要な結果を導き出しやすいことが最大の理由ではないかと考える．

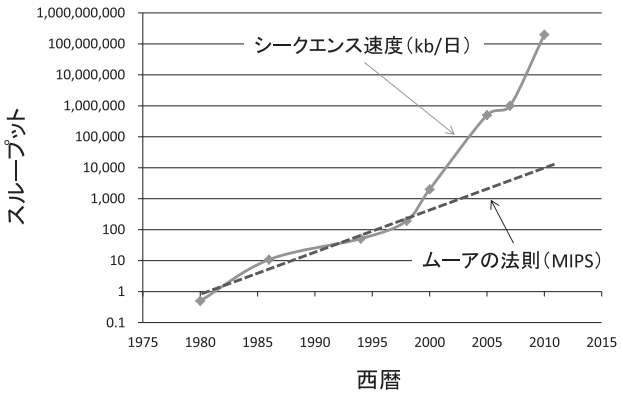

図1．DNAシークエンサーの解析速度の向上．比較としてコンピューターの計算性能の向上を示した[1]．

このことは，DNAシークエンサーに必要な要素技術を考えればはっきりする．最初にDNAシークエンサーが開発されたのは1980年代である．この頃には，Sanger法によるシークエンス反応，電気泳動，蛍光検出などの要素技術はほぼすべて揃っていたといえるが，実用的なシークエンスを行うためには，1本鎖DNAの調製技術，蛍光色素を取込みやすい高性能DNAポリメラーゼ，くせのある波形データからの正確なベースコール，大量のシークエンスを可能にするダイターミネータ法など，多くの技術開発を待たなければならなかった．この結果，Sanger法の最終形といわれるABI社製の3730型キャピラリーシークエンサーが2000年に登場した．この時点で，1980年代末に比較して既に1000倍以上のスループットの向上が得られていたが，2005年にはさらに100倍以上にスループットが向上した，いわゆる次世代型のシークエンサー（NGS：Next Generation Sequencer）が開発された．塩基配列に関する情報解析分野はソフト・ハード両面からもっとも充実しているといえるが，454，Solexa，SOLiDなどの実用化によって，既存の処理能力をもボトルネックとする生産速度を持つに至った．それにも関わらず，シークエンサーの技術開発はとどまるところを知らない．これはとりもなおさず，「情

著者紹介　[1]（独）産業技術総合研究所生物プロセス研究部門（総括研究主幹）　E-mail: m.machida@aist.go.jp
[2]（独）産業技術総合研究所生命情報工学研究センター（センター長）　E-mail: asai-cbrc@aist.go.jp

報が生物学の鍵を握る」という絶対的なモチベーションがあるからに他ならない．このモチベーションの高さによって，塩基配列解析と情報解析技術の開発速度が決定されているのである．

2001年にヒトのゲノム解析が完了したとの報告が為された．一部には，これをもって「DNAシークエンス（ゲノム解析）は終わった」との論調も見られたが，これが完全に誤った状況認識であったことは現状を見れば明らかであろう．生物種に応じたゲノムの多様性に加えて，遺伝子領域以外の無意味と考えられていた箇所に生物学的な意味が見いだされつつあるなど，ゲノム解析の分野は終息するどころかますます拡大しつつある．しかし一方で，この大容量の生物情報を使いこなせていないことも事実である．この理由のひとつとして，生物情報の生産速度の向上が，計算機の情報処理速度のそれを上回っていることがあげられる．近年では，「ビッグデータの衝撃」とも呼ばれるように[2]，大規模なデータをいかに効果的にマイニングするかが問われている．誇張された面があるかもしれないが，現在の生物情報は，その量だけでなく，取扱いの難しさにおいても代表格といって差し支えないと考える．

なお，本稿においては，「ゲノム情報」を生物情報およびオミックス情報の代わりに用いる．一般にオミックスと称される情報の中には，スループットが十分でないことや有効なマイニング手段が存在しない情報が多く含まれると思われるからである．また，生物情報は生物をシステム的・網羅的に扱う以外の情報も含まれる．本稿でのゲノム情報は，「生物が有する遺伝子の全体」であるゲノムの塩基配列，およびこれと密接したトランスクリプトームやプロテオーム情報を含む．これらは，高いスループットと情報解析の方法論の開発が進んでいる．メタボロームは未だに取り扱いが難しい情報と考えられるが，トランスクリプトームやプロテオームとの関係性を含めた解析が重要性を増しつつあり，これを含めることの意義は大きい．

比較ゲノム解析のポテンシャル

効果的なマイニングが可能で，大規模かつ計算機の処理速度が要求される課題として，比較ゲノム解析が考えられる．代表的な比較ゲノム解析としては，オーソログ解析やシンテニー解析がよく知られている．大腸菌，枯草菌，酵母，線虫など，1990年代にモデル生物が解析され，これと並行して，網羅的な遺伝子破壊を代表とする，システム的な遺伝子機能の解析が精力的に行われてきた．そこで，新たに解析された生物種など，着目する生物が有する遺伝子の機能をできるだけ多く予測する手段として，オーソログ解析は現状でもっとも有効な手段である．また，シンテニー解析は，比較的近縁種間での進化などを論じる際などに用いられてきたが，筆者らは，比較ゲノム解析が遺伝子の機能を予測するための有効な手段であることを見いだした．筆者らは他の多くの研究者と協力して，2001年に日本の代表的産業微生物である麹菌（*Aspergillus oryzae*）のゲノム解析を開始した．この頃，ほぼ同じタイミングで，学術的に重要な糸状菌である*Aspergillus nidulans*とヒトへの感染性を有する*Aspergillus fumigatus*のゲノム解析が同時に行われ，これらのグループとの連携によって詳細な比較ゲノム解析を行った．

麹菌のゲノムサイズは約37 Mbで，他の2種（約30 Mb）に比較して25％程度大きく，これら2種とのシンテニー解析によって，シンテニー領域（SB：Syntenic Block）と非シンテニー領域（NSB：Non-Syntenic Block）がゲノム全体にわたってモザイク状に分布していることが明らかとなった[3]．これらの各領域上での遺伝子の分布について，生化学的機能の観点から調べたところ，NSB上には，二次代謝系遺伝子と分泌性の加水分解酵素が集積されていることが明らかとなった．また，NSB上の遺伝子は，相同性解析による機能の予測が困難な遺伝子が多いことも特徴であった．二次代謝系の遺伝子は，最初の抗生物質であるペニシリンや血圧降下剤として知られるモナコリンに代表されるように，医薬品のリード化合物をはじめとする生理活性物質を生産するなど，産業上有用であるにも関わらず，遺伝子レベルでの研究は遅れているといわざるをえない．その理由として，これらの遺伝子の発現は一般に低レベルであることが多く，ESTなどの解析でも，主要な転写・翻訳，一次代謝系の遺伝子などに比較して捉えにくいことがあげられる．また，自然界から分離して実験室で培養していると，生産性が失われる（遺伝子の発現が見られなくなる）といった問題も存在する．

二次代謝系の遺伝子が改めて注目されるようになったのは，1990年代に糸状菌や放線菌などの微生物のゲノム解析が行われるようになってからである．この頃までには，キャピラリーシークエンサーなどのハイスループットな解析装置が整備され，情報処理技術の開発によってホールゲノムショットガンが有効に機能するようになった．これにより，半年以下という比較的短期間で全ゲノムの概要を解析できるようになったことから，モデル生物以外の多様な生物種のゲノム解析が動機づけされた．この結果，糸状菌や放線菌には，ゲノム解析前に

予測していたよりもはるかに多くの二次代謝系遺伝子が存在することが明らかになった．当初は主に，ポリケチド合成酵素（PKS：Polyketide Synthetase）や非リボソームペプチド合成酵素（NRPS：Non-Ribosomal Peptide Synthase）など，アミノ酸配列の相同性によって比較的容易に遺伝子機能を予測できる遺伝子に関するものが中心であった．しかし，筆者らが同定に成功したコウジ酸生合成遺伝子クラスタなど，これまでは二次代謝系遺伝子と予測することができなかった遺伝子も存在することが明らかとなった[4]．さらに，この遺伝子クラスタは前述のNSB上に存在していた．このことは，NSB上に存在する機能未知の遺伝子は，二次代謝系遺伝子である可能性が高いことを裏づけている．筆者らは，この性質に基づいて，比較ゲノム解析によって二次代謝系遺伝子クラスタを予測する方法を開発しているが，これまでに既知の二次代謝系遺伝子クラスタのほぼすべてを予測することに成功し，それ以外にも二次代謝系の可能性が高い多数の遺伝子クラスタを検出している．しかしながら，先にも述べたように，二次代謝系遺伝子は一般的に発現レベルが低いこと，培養ごとに発現が安定しないことなどにより，予測された遺伝子クラスタの実験的なバリデーションは困難であるといわざるをえない．さらに，これらの遺伝子によって生合成される化合物の性質も多様であり，必ずしもODSカラムを用いたLC/MS解析などで網羅的かつ統一的に検出できるとは限らない．セスキテルペン類などの揮発性物質の測定には専用の器具を用いなければならない場合や，イオン化が難しいためにGC/MSを含むMSに基づいた検出が困難な場合も存在する．現在，世界中で，二次代謝系遺伝子の予測ソフトと遺伝子破壊などのバリデーションによる解析が進められているが，ゲノム情報の生産速度に比較してスループットが圧倒的に不足しているといわざるをえない．

比較ゲノム解析による興味深い発見として，二次代謝系遺伝子や一部の細胞外加水分解酵素などの遺伝子の多様性が，他の基本的な遺伝子に比較して，有意に多様性が高いことが明らかになった点があげられる．これは，2種類の麹菌株のゲノム塩基配列を比較することで発見された．前述のように，麹菌のゲノムはSBとNSBの2つの異なる性格の領域がゲノム全体にわたってモザイク状に分散して存在しているが，NSBはSBに比較して明らかに変異率が高いことがわかったのである[5]．これまでにも，染色体末端付近（テロメア近傍）の遺伝子の変異率が高いことが報告されていたが，本解析では，染色体上の位置によらず，NSBか否かにのみ依存して変異率が上昇していることが突き止められた．NSB上には二次代謝系の遺伝子が集中していると考えられ，一部の細胞外加水分解酵素をコードする遺伝子も局在している．したがって，これらの遺伝子は他の遺伝子に比較して明らかに変異率が高いといえる．

表1．麹菌の株間に見られる変異の種類

	遺伝子数	同義置換	非同義置換	同義／非同義置換
SB	9,345	14,229	10,180	1.4
NSB	3,784	8,328	7,228	1.2

*表内の数値は，文献5から抜粋して示した．

麹菌のゲノム解析が行われた当時から，NSB上の遺伝子はSB上のそれに比較して，有意に発現強度が低いことがわかっていた．このことは，高い変異率と合わせて，NSB上の遺伝子の多くが偽遺伝子や不要な遺伝子であることを想像させる．しかし，塩基配列の変異の種類をアミノ酸の置換の観点から見た場合，SB上とNSB上の遺伝子のいずれについても，非同義置換に対する同義置換の比率は1.2～1.4倍でほとんど違いはなかった（表1）．また，ランダムな変異を仮定した場合の同義置換の比率は0.054にすぎないことから，SB，NSBともに同義置換の割合が非常に高く，いずれの領域に存在する遺伝子にも高い選択圧がかかっていることを物語っている[5]．したがって，多くのNSB上の遺伝子は，これまでに試された生育条件では発現が非常に弱いか認められないにも関わらず，何らかの生物学的に重要な意味を持っていると考える必要がある．

図2は，麹菌のゲノム上の各遺伝子について，酵母や他の糸状菌に対応する遺伝子がどの程度存在するかを示したものである[6]．ゲノムサイズが麹菌の約1/2.6と小さい酵母の場合にはSB上の遺伝子も約30%であり，NSB上の遺伝子では15%程度と，存在割合が非常に低い．これは，出芽酵母が二次代謝系をあまり持たないことと一致する．ゲノムサイズが麹菌よりも若干大きな*Neurospora crassa*や*Magnaporthe grisea*を比較対象とした場合には，SB上の遺伝子は60%程度にまで上昇するが，NSB上の遺伝子は30%～35%とそれほど大きくは改善されない．実際，NSB上の遺伝子は近い種同士であってもオーソログを組めない場合が多く，二次代謝系などのNSB上の遺伝子がきわめて高い多様性を有することが明らかとなった．これらのことは，二次代謝系の重要性も相まって，近縁種であってもゲノム解析が有効かつ重要であることを示している．現在では，ゲノム塩基配列が公開されているものだけでも，バクテリアで1000種以上，真菌で100種類以上であり，未公開分は

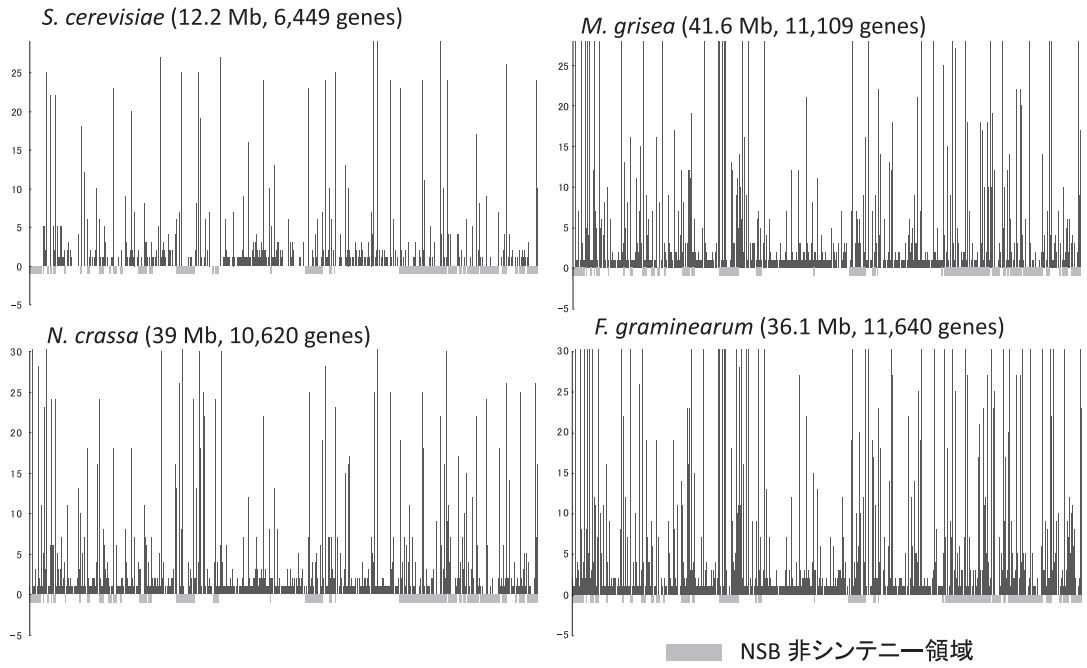

図2. NSB上の遺伝子の多様性．麹菌の第6番染色体上の各遺伝子について，他の真菌のゲノム中に存在する相同性他の赤い遺伝子の数をプロットした[6].

その10倍程度以上は存在すると予想される．遺伝子の対応関係や分布の解析には，multiple alignmentによるクラスタリングや種ごとの全遺伝子のpair wiseな相同解析が一般的な方法であり，種の数の2乗倍の計算時間が必要である．すでに世界中で利用されているNGSの台数を考えれば，150万種とも予測されるすべての菌類のゲノム塩基配列を明らかにすることも十分に現実的な議論であり，上記のような基本的な解析だけでも計算機の能力をはるかに超えるであろうことは容易に想像できる．

大規模なゲノム情報の活用技術

大規模なゲノム情報が，迅速かつ容易に手に入る現在では，これを有効に使いこなす方法論が重要である．周知のように，ゲノム科学の発展とともに，膨大な情報処理技術開発やさまざまな試みがなされてきた．NGSの登場によって，これまで研究が難しかった生物種でも比較的短時間で結果が出せるようになった．しかし，もっとも大きな変革は，特定の種に着目した研究から生物学的な現象や機能に直結した研究方法への加速であろう．これまでは，着目する機能を有する生物種のなかでも，遺伝解析，遺伝子組換技術，ゲノム塩基配列などが利用可能な生物種を中心として研究が進められてきた．こういった生物実験上の利便性は現在でも重要な要素であることに変わりはないが，これらが使えなかったとしても，現在では培養さえできればトランスクリプトームなどのゲノム情報を用いて研究することができる．これにより，着目する生物機能がもっともはっきりした生物種を用いて解析を行うことや，その生物機能を持たないような生物種を用いた比較機能ゲノム解析が有効な手段となった．このような研究方法は，複数種のゲノム情報を同時に取り扱うばかりでなく，ある程度広範な生物種にまで対象を広げて解析･比較を行うことを特徴とする．また，生物学的な答えを導くために，多様なゲノム情報を柔軟に処理することが求められる．これらを使いこなすことができれば，大量のゲノム情報に支えられて，生物実験に必要な時間と労力の大幅な軽減が可能になる．情報量に比較して生物実験に必要な時間と労力が圧倒的に不足している現状では，これを大幅に節約する方法を作らなければならない．このための方策として，生物実験のスループットの向上や，生物実験の情報処理による置き換えが考えられる．

現状では，自らの研究対象に関わるかどうかは別にして，ゲノム情報は確実かつ加速度的に増加する．上述のように，自らの研究対象と直接関係のない生物種のゲノム情報であったとしても，着目する生物機能の研究に有効に利用できる可能性がある．したがって，これを使いこなす術を持たない場合，致命的に競争力を失うリスクを背負わなければならない．生物学的な意味を解析する

ための生物実験は，専門的な知識，複雑な実験，時間と労力が必要であり，このために必要な限られたリソースを有効に利用して研究速度を向上させるためには，その負荷を軽減するための情報処理技術とそのための基盤整備が必要不可欠である．しかも，ゲノム情報の加速度的増加への対応だけでなく，次々に開発される最新の解析技術を利用した多様な情報への柔軟な対応も考えなければならない．

生物実験の負荷を軽減するためには，生物実験系と情報解析系のメリットとデメリットを熟知した上で，この二つの技術のもっとも有効な組合せ解を得る必要がある．たとえば，新規な化合物の生産を標的とするのであれば，これに関わる生合成遺伝子の高精度な予測技術が求められるであろう．このための情報解析の鍵は，多様な解析を可能にする基盤情報の整備，さまざまな状況に柔軟に対応できるソフトウェア開発である．基盤整備には，大容量で高速な計算機と，計算量の爆発を避けるためのアルゴリズムと実装技術の開発が必要である．さらに，生物実験と情報解析の高度な連携においては，生物系実験者に対して，多様な情報を分かりやすく整理して提示するだけでなく，その場で部分的な処理を簡単に実行できるインターフェースなどにより，情報処理結果を生物実験に迅速かつ効果的に利用するための技術が求められる．筆者らは，バイオインフォマティクスだけでなく，ハードウェアやこれに密着したソフトウェアの開発を主要な研究課題とする研究ユニットと密な連携を築くことにより，この難題に挑戦している．ゲノム科学による生物学の変革期はまだ始まったばかりであり，好むと好まざるとに関わらず，このスリリングな状況は少なくとも当分続くことは間違いない．

文　献

1) 経済産業省：2001年度版 通商白書要旨, p.13 (2001).
2) 城田真琴：ビッグデータの衝撃，東洋経済新報社 (2012).
3) Machida, M. *et al.*: *Nature*, **438**, 1157 (2005).
4) Terabayashi, Y. *et al.*: *Fungal Genet. Biol.*, **46**, 953 (2010).
5) Umemura, M. *et al.*: *DNA Res.*, **19**, 375 (2012).
6) Machida, M. *et al.*: *Food Addit. Contam.*, **25**, 1147 (1998).