7 ポストゲノム時代の配列情報解析

浅井　潔
東京大学／
（独）産業技術総合研究所
asai@k.u-tokyo.ac.jp

バクテリアに始まりヒトを含む多くの真核生物のゲノム配列が決定されたが，もとよりゲノム配列の決定はその解読ではなく，ゲノム配列情報の意味を解明するためには多くの課題が残されている．

その第1は遺伝子発見と機能アノテーションである．多くのゲノム配列が遺伝子位置，機能が注釈付けされたかたちで公表されているが，その作業は熟練した研究者の人手によるところが多く，自動化された技術としては確立していない．遺伝子制御ネットワーク，代謝パスウェイ，シグナル伝達パスウェイなどを解明し，データベース化することはより高次の課題であるが，ゲノムに存在する遺伝子セットの機能アノテーションが前提となっている．今後は，DNA マイクロアレイによる発現解析，タンパク質相互作用の解析，タンパク質立体構造などの多元的な情報と配列情報を統合した取り組みが主流となっていくであろう．

多くのゲノムが決定されたことにより可能になった「比較ゲノム」の研究により，共通の保存領域の中には，タンパク質コード領域ではない部分（非コード領域）の方が多いことが明らかになった．このうち相当部分が機能を持つ RNA ではないかと考えられている．RNA 干渉と miRNA の発見によっても注目を浴びている RNA ではあるが，実は RNA 配列の情報解析技術は確立された技術とは言いがたい．二次構造と配列相同性の両方を考慮した実用的な配列の比較・検索手法はまだ存在しないが，近年カーネル法や共通二次構造予測など，新しい手法が提案されている．

ゲノム配列の解読

生物配列（DNA 塩基配列とタンパク質アミノ酸配列）の文字列処理（配列情報解析）は，今日に至るまで，バイオインフォマティクスの最重要課題の1つとして研究されてきた．配列情報解析の重要な対象であるゲノム塩基配列は，すでに200 種類以上が決定され，さらに多くが進行中であるといわれている．

膨大なゲノム配列をここまで高速に決定できるようになった要因には，塩基配列の自動読み取り技術と配列情報解析技術の進歩が挙げられる．塩基配列の自動読み取り装置では，塩基ごとに異なる色を用いた蛍光標識を自動的に読み取り，画像処理技術，信号処理技術を用いて4 種類の文字の列に変換される．自動読み取り装置で一気に精度良く読める塩基配列は千塩基以下であり，ゲノム配列は数百万塩基から数十億塩基のサイズがあるから，個々の塩基配列がゲノムのどの位置のものかを注意深く管理しないと，どこがどこだか分からなくなる．従来は，ゲノムのさまざまな位置から，大きな断片から小さな断片に至るまで階層的に位置決めをしながら取り出し，読むべきDNA 配列が注意深く選ばれていた．ところが，配列情報解析技術の進歩で，全ゲノム配列をさまざまな方法で多数の断片に切断し，それぞれを自動読み取り装置で読み取った結果から，読み取りミスを考慮しながら重複部分の情報を用いて断片をつなぎ合わせ，ゲノム塩基配列を一気に復元する「全ゲノムショットガン」が可能になった．

ゲノム配列に「書かれている」情報のうち最も重要なものは，タンパク質をコードしている遺伝子である．ヒトの30 億文字のゲノム配列の中には，このような遺伝子が約3万個あるといわれている．その中のいくつかは，ゲノム配列が決定される前から実験によって存在が確認され，遺伝子部分の塩基配列が決定されていたが，残りの大部分の遺伝子は，ゲノム配列が決定された後で情報技術によって見出されたものである．個々の遺伝子はDNA からmRNA へ転写された後に，mRNA からタンパク質へ翻訳され，互いに相互作用しながら，生体内で代謝・シグナル伝達をはじめさまざまな機能を発揮する．このうち，特にDNA からmRNA への転写に関しては，どの遺伝子がどのような種類の細胞で，どのようなタイミングで行われるかによって，生体の活動が制御

されている．これらの制御に関する情報もまた，ゲノム配列の情報の中に隠されている．また，タンパク質コード遺伝子以外に，DNA から転写されたRNA それ自身が，機能性RNA として重要な機能を担っている場合も多く，近年注目が集まっている．

ゲノム配列の「解読」においては，これらゲノム配列に書かれた情報の「意味」を見出すための配列解析技術が重要である．本稿では，ゲノム配列の「解読」に向けた配列解析技術の課題について，目標と手法の両面から概観する．

近年は，単純な配列解析よりも，DNA マイクロアレイによる多数の遺伝子の発現パターンや，タンパク質間相互作用の網羅的な検出結果を用いた遺伝子制御ネットワークや代謝，シグナル伝達パスウェイの推定などが注目されている．しかし，ゲノム塩基配列が決定された後最初に行わなければならないことは，ゲノム塩基配列上のどの部分にどのような遺伝子が存在するかを同定することである．また，遺伝子の発現を制御している配列の解析，タンパク質コード遺伝子以外の機能性RNA 遺伝子の同定なども，遺伝子ネットワークなどより高次の解析を行う前提として必要とされる解析である．

配列解析の手法としては，整列（アラインメント），クラスタリング，共通パターン（モチーフ）の抽出，類似配列の検索などが必要となるが，そのすべてに共通に重要なのが配列比較である．生物配列には，互いに類似の配列が数多く含まれている．その最大の理由は，生物配列が突然変異による進化の過程を経て多様性を獲得してきたために，互いに共通の祖先を持っていることである．また，生体内において共通の機能を果たすため，立体構造や機能部位の配列が似ているという側面もある．このような類似の配列を互いに比較し，あるいは膨大な配列中から類似の配列を検索することによって，ゲノム中の遺伝子の場所を推定し，種類を分類し，構造・機能を推定し，ゲノムの進化の歴史を読み解くことによってゲノムという膨大な文字列の意味を探る研究が行われてきた．本稿の後半では，配列を比較することの意味について，RNA 配列の解析における問題点を含めて議論する．

遺伝子発見と機能アノテーション

決定されたゲノム塩基配列のどの部分に，どんな遺伝子があるかを配列情報解析によって明らかにすることは，配列情報解析の重要な課題である．筆者は麹菌(Aspergillus oryzae) のゲノム配列解析プロジェクトに参加したので，麹菌の場合を例にとって遺伝子発見と機能アノテーションについて解説しよう．

麹菌はカビの一種（真核生物）で酒，醤油，味噌の製造に広く用いられている．2001 年に始まった麹菌ゲノム配列解析プロジェクトの結果，8 本の染色体の約3 千760 万塩基が決定された．産業技術総合研究所の生命情報科学研究センターでは，2001 年の8 月に麹菌のドラフト配列（大まかに決定されたゲノム塩基配列）の提供を受け，約13,752 個のタンパク質コード遺伝子の位置を特定した[1]．

遺伝子発見で最も信頼できる方法は，既知遺伝子と類似の文字列を検索することである．麹菌のゲノム解析においては，約6,000 本のEST の配列情報が利用可能であった．EST とは，細胞内でゲノムDNA から転写されてできたmRNA の一部をDNA に逆転写して得られるcDNA の一種で，その文字列が細胞内でDNA からRNA へ転写されている証拠である．sim4というソフトウェアを用いてEST に対応するゲノム配列上の位置が見つけ出された．既知遺伝子としては，その生物自身の遺伝子として実験で確認されたもののほかに，他の生物で発見された遺伝子の文字列が使われる．異なった生物の遺伝子であっても，進化によって関連付けられたもの同士は，類似の文字列を持っているからである．公開のデータベースに登録された既知遺伝子と類似の文字列は，BLAST というソフトウェアを用いてゲノム配列中から検索した．

これらの結果は，遺伝子の大まかな位置の推定には役立つが，スプライス部位を含む遺伝子構造を正確に決めるには不十分であることが多い．そこで，ESTとゲノムの対応関係と配列の統計情報を統合して遺伝子構造を決定するために，筆者らの開発した多重出力隠れマルコフモデル（HMM）による遺伝子領域予測ソフトウェアGeneDecoder[2] を用い，BLAST 検索の結果得られた遺伝子候補から正確な遺伝子構造の推定をするためには，後藤修氏（現京都大学）が開発したALN[3] というソフトウェアを用いた．

統計情報や，既知遺伝子による方法のほかに，近年注目されている遺伝子発見に関する第3 の方法が，比較ゲノムによる遺伝子発見である．異なった生物のゲノム配列では，機能的に重要でない文字列は進化の過程で大きく変化し，類似性は低くなっていると考えられる．反対に，機能的に重要な文字列は比較的保存されていると考えるのが自然である．近縁のゲノムを互いに比較し，保存性の高い部分を抽出すれば，そこに遺伝子などの機能的に重要な配列が含まれている確率が高くなる．ヒトゲノムのほか，マウスやラットのゲノムが利用可能となり，比較ゲノムの研究は注目を浴びている．麹菌の場合は，同じAspergillus 属の2 つのゲノムAspergillus fumigatus，Aspergillus nidulansのゲノム解析が完了し

ており，これらの3種のゲノムの比較が行われた．2001年の麹菌の遺伝子発見の段階では，残念ながらこれら2種のゲノム配列は利用可能ではなかったので，比較ゲノムによる遺伝子発見は行われなかった．

遺伝子発見の次に行うことは，遺伝子の機能アノテーションである．既知遺伝子の情報に基づいて発見された遺伝子の場合は，元の既知遺伝子の機能が既知であれば，発見された遺伝子も同様の機能を持っていると推定される．また，遺伝子発見自体はESTや統計情報に基づいて行われたものであっても，発見された遺伝子の文字列に対して，改めて既知遺伝子との類似性の検索を行うことによって機能に関する手がかりを得ることができる．そのほかに，タンパク質にはモチーフと呼ばれる部分文字列のパターンが数多く知られていて，特定のモチーフを持つものは特定の機能を持つと推定できる場合もある．麹菌プロジェクトでは，類似性検索やモチーフの存在に関する計算機による自動解析の結果と推定された機能を麹菌ゲノム解析コンソーシアムのメンバに配布し，研究者が手作業で修正して機能アノテーションの確定を行った．

ゲノム配列中に点在するタンパク質コード遺伝子を発見し，それらの機能をアノテーションすることはゲノム配列決定後に行われる最初の配列情報処理であり，その後のすべての解析の基本である．それぞれの遺伝子が，どのような制御を受け，どのように発現し，どのように協調して生体内で働いているのかを知ることは，その次の課題である．

遺伝子の発現量とその発現のタイミングを制御するメカニズムは複雑だと思われるが，現在までに知られているメカニズムのうち主要なものの1つに，主に遺伝子上流に存在する転写制御配列（ゲノム塩基配列からmRNAへの転写の制御に関係する特徴的な配列）がある．多くの場合，転写制御配列の特徴的な文字列を特殊なタンパク質が認識し，結合することによって転写を開始，促進，抑制する．転写制御配列の研究はこれまでも盛んに行われてきたが，DNAマイクロアレイ技術による遺伝子の種類ごとのmRNA検出（定量性には問題があるが），質量分析によるタンパク質の直接測定などにより，制御配列と遺伝子発現の関係を解析するためのデータが飛躍的に増加している．

転写制御配列は，数塩基～十数塩基程度の配列パターンで，共通の働きを持つ配列同士は「似ている」．既知の転写制御配列は，転写制御に関係するタンパク質が結合することや，配列の一部を人工的に変更すると転写量が増減することが確かめられている．これらの配列をモデル化し，類似の配列を遺伝子の上流から探し出すことはそれほど難しいことではない．より困難なのは，未知の「共通パターン」を探し出すことである．すべての遺伝子が同一の「共通パターン」の配列を持っているわけではない．しかし，同一の「共通パターン」を持っている場合，類似の転写制御を受ける可能性があるから，DNAマイクロアレイなどで類似のタイミングで発現量が増減するような遺伝子の上流から「共通パターン」を探すことには意味がある．

機能性RNAの配列解析

近年，多くの生物のゲノム配列が決定され，それらを互いに比較する「比較ゲノム」研究が盛んになった．これまでゲノム配列解析の中心であったタンパク質コード遺伝子も，これらの保存配列から数多く発見されている．一方，これらの保存配列の中に，タンパク質をコードしているとは思えない部分がかなりの割合に上ることも明らかになった．DNA配列のコード領域は3文字単位のコドンがアミノ酸1種類に対応するが，その統計的偏りや終止コドン（アミノ酸に対応せずに翻訳を止めてしまう特殊なコドン）の現れ方から非コード領域だと推定される配列が，保存配列から多数見つかったのである．また，本来確実に発現している遺伝子の塩基配列を効率的に得るために行われてきたcDNAの配列決定の結果得られたデータからも，多くの非コード領域が見つかっている．

これらの非コード領域には，転写された遺伝子で翻訳されない部分（UTR），遺伝子の転写制御にかかわる制御配列（プロモータなど）のほかに，多くの非コードRNA（ncRNA）が含まれているのではないかと考えられている．非コードRNAとは，ゲノムのDNA配列から転写されてできるRNAで，タンパク質に翻訳されることのない一群のRNAの総称である．その中でも，他の生体分子と相互作用することで一定の機能を発現するRNAまたはその遺伝子を機能性RNAと呼ぶ．翻訳過程においてコドンとアミノ酸に特異的に結びついて仲立ちをする転移RNA（tRNA），翻訳を行うリボゾーム（タンパク質とRNAの複合体）を形成するリボゾームRNA（rRNA），リボヌクレアーゼP（RNAse P），リボザイム（RNA enzyme）などは代表的な非コードRNAである．このほか，rRNAの塩基修飾にかかわるsmall nucleolar RNA（snoRNA），tRNAとmRNAの両方の働きを持つtmRNAなども知られている．

近年の研究で，タンパク質コード遺伝子と相補的な配列を持つRNAが，その遺伝子の発現を阻害するRNA干渉と呼ばれる現象が注目されるようになった．この現象を利用して，従来遺伝子組み換え技術を用いて行っていた特定遺伝子の無効化を，簡単に行うことができる

ようになりつつある．また，似たような機構で遺伝子の発現を阻害するマイクロRNA（miRNA）と呼ばれる機能性RNAがゲノム上に存在することが明らかになった．これらの機能性RNAのデータベースとしては，Rfam[4]が有名である．

RNA遺伝子はタンパク質遺伝子に見られるコドン使用頻度や開始コドンから最初に現れる終止コドンまでの領域（ORF: Open Reading Frame）のような一般的指標で特徴付けることが難しい．1本鎖のRNAは，互いに相補的な塩基（AとU，GとC．時にGとU）が塩基対をつくり，二次構造と呼ばれる局所構造をとる（**図-1**参照）．塩基対が続く部分をステムと呼び，ステムを構成する連続文字列は配列上離れた2カ所に存在する．機能性RNAの配列情報解析を行う場合，その二次構造を考慮することが重要である．

rRNAのような相同性の高いncRNAに対しては通常の類似配列検索（BLASTなど）で既知ncRNAを発見できるが，二次構造を考慮できないため，一般のncRNAに対しては有効でない．相同なRNA遺伝子配列群の保存領域と二次構造に関する情報が正しく得られた場合には，プロファイルSCFG（確率文脈自由文法）などを用いて，ゲノム配列中から，対象となるRNA遺伝子を発見することが可能となる．麹菌プロジェクトでは，tRNAscan-SE[5]というソフトウェアを用いてtRNA遺伝子の発見を行った．tRNAscan-SEは，tRNAの既知の二次構造をモデル化した確率文脈自由文法（SCFG）を用いたソフトウェアである．

rRNAとtRNA以外の機能性RNAについては，その検索手法は確立しておらず，ゲノム配列に対してルーチンで行う作業によって発見することは現状ではむずかしい．また，1本のRNA遺伝子候補配列が与えられたとき，相同なRNA遺伝子を検索するための手法は，現状では皆無に等しい．さらに，シュードノット（二次構造の対応関係が「入れ子」に収まらず，クロスしてしまうような構造）を含む構造を扱う場合は計算量が膨大で，現実的な時間での検索が困難である．

RNA配列から，その二次構造を計算によって求めることを，RNAの二次構造予測と呼んでいる．RNAの二次構造予測は，バイオインフォマティクスにおける古典的問題の1つである．RNAの二次構造は，その相補塩基対の形成によってエネルギーが最小となるような構造をとっているだろうという予想のもとに，最小エネルギー構造を計算するアルゴリズムが考えられている．RNAがとり得る二次構造のすべての可能性は，すべての可能な相補塩基対の組合せの数に匹敵するが，相補塩基対の数が最大になる構造を求めるNussinovアルゴリズム，隣り合う塩基対の積み重ねエネルギーなどより詳細なエネルギー計算を行うZukerアルゴリズムなどが知られている．これらのアルゴリズムは，シュードノット構造を許さないという制約があるにもかかわらず，配列の長さの3乗オーダーの計算時間が必要であり，しかも実際には，得られる二次構造は正しくないことが多い．現状で最も性能のよい二次構造予測を行うには，共通二次構造を持つ複数のRNA配列の正しいマルチプルアラインメントが必要となる．ところが，二次構造予測を行う前に，共通の二次構造を持つことが分かっているRNA配列が得られることは珍しい．しかも，「正しいマルチプルアラインメント」とは，二次構造を反映したマルチプルアラインメントのことを指しており，そのようなアラインメントが二次構造予測に必要だとすると，「鶏と卵」になってしまい，何も予測できない．

機能性RNA遺伝子を発見するための標準的な手法は存在しないが，その基礎となる手法は発展しつつある．その1つが，複数のRNAの共通二次構造を直接求めるアルゴリズムである．Sankoff（1985）は，長さLの配列N本の共通二次構造を$O(L^{3n})$の計算時間，$O(L^{2n})$のメモリで求めるアルゴリズムを提案した．当時の計算機の能力では，明らかに実用的でない手法であったが，現在ではその改良版や，よりヒューリスティックなアルゴリズムが考案されている．

配列の比較

本章の内容は，榊原氏による「1.バイオインフォマティクス概説」の副題「比べることで生命は解明できるか？」と関係が深く，特に「ゲノムレベルの比較」，「木構造データの比較」を参照されたい．配列解析の最も基本的で重要な課題は，配列の比較であるが，基礎になる考え方には大きく分けて2種類ある．その1つは，文字列同士を動的計画法によって整列させ，対応する部分の文字の一致度を主な指標とした文字列の類似度を計算する配列アラインメントである．もう1つは，一定の長さまでの文字列（単語）を数え上げ，それによって長い文字列を特徴付けたり，類似部分の検索に用いたりするものである．これらの配列比較技術は，生物配列の分類，モデル化等に必要であり，遺伝子発見とアノテーション，制御配列の解析，機能性RNAの解析などに共通に必要なものである．

2本の配列の最適な整列とそのスコアは2次元の動的計画法によって配列の長さの積のオーダーの計算時間で求められる．このスコアを配列の類似度となるように置換行列を注意深く選ぶことにより，配列アラインメントによる配列比較は，今日まで配列情報解析の中核的な手

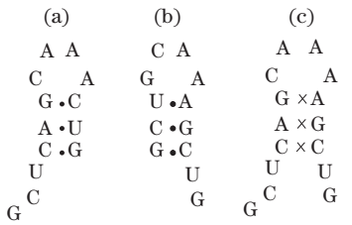

図 -1　二次構造を持つ RNA の比較

図 -2　RNA 配列のアライメント (1) : 二次構造を反映している．四角の枠は対応するステムを示す

図 -3　RNA 配列のアライメント (2) : 単純な動的計画法による．四角の枠で示されたステムの対応がずれてしまっている．

図 -4　RNA 配列のアライメント (3) : (a) と (b) は二次構造を反映している．(a) と (C) の前半，(b) と (c) の後半は同一．

法として用いられてきた．比較的短い 2 本の配列を単に比較するだけであれば，この長さの 2 乗程度の計算時間は問題とならないが，大きな配列データベースに類似の部分配列をゲノム配列全体から検索するような場合には，実行が困難となる．1990 年代には，BLAST や Fasta といった，文字列のインデキシング，有限状態オートマトンなどと配列アラインメントを組み合わせたより高速なソフトウェアが一般的に使われるようになった．1990 年代の初頭からは，HMM や SCFG などの文字列の確率モデル（確率文法）が配列情報解析に広く使われるようになった．これらの確率モデルでは，モデルの隠れ状態（確率文法の非終端記号）と文字とのアラインメントを構文解析によってとることによって，遺伝子発見や配列アラインメント，構造予測などを行うことができる．

近年注目が集まっている，異なった種類のゲノム全体を比較する「比較ゲノム」において，長いゲノム配列全体同士で対応位置を大域アラインメントで単純に求めることは計算量の面でむずかしいため，標準的な手法は確立していない．そこで，比較的保存性の高い対応位置をまず見つけ，残りの部分を後からアラインメントする手法がよく用いられる．麹菌と Aspergillus 属ゲノムとの比較では，遺伝子発見をまず行い，遺伝子同士の対応位置を参考にアラインメントを行うべきブロックを推定する手法が用いられた．

RNA 配列の比較も，単純な動的計画法ではうまくいかない．図 -1 を見てほしい．2 本の配列 (a)，(b) は，ともに 1 組の長さ 3 相補文字列からなるステムを持っている．この 2 本の配列を比較するには，どのようにするのが良いだろうか．2 本の配列は同じような二次構造をとっていることから，二次構造上の対応部分同士を比較するために，**図 -2** のようなアラインメントをとり，何らかのスコアを計算することが妥当だと思われる．ところが，通常の配列比較で用いられるアラインメントを行うと，**図 -3** のようになってしまう．これでは，対応するステムを構成する文字列同士が比較されないから，二次構造をまったく考慮しない類似度を与えることになるだろう．機能性 RNA の比較・検索においては，二次構造と配列類似性の両方を考慮してスコアを定義するべきであるから，図 -2 のようなアラインメントを与えるような手法を用いなければならない．

今度は，**図 -4** を見てほしい．配列 (c) は配列 (a) の後半部分と，配列 (b) の前半部分をつなぎ合わせて作ったものである．図 -1 に示したように，配列 (c) は，相補塩基対によるステムを (a) や (b) のように作ることができないから，RNA 配列としては (a) と (b) が近く，(c) は遠い，という尺度が望ましい．ところが，図 -4 のアラインメントに通常のスコアを用いると，(a) と (c)，(b) と (c) が近いという結論になってしまう．

以上のことから，RNA 配列検索の基礎となる RNA 配列比較に関する 2 つの問題が存在することが分かる．その第 1 は，RNA 配列比較を行うために RNA 配列の整列を行う場合，その共通二次構造を考慮しないで通常のアラインメントを行うと，二次構造上の対応する位置が整列しないような不適切なアラインメントが得られ，その

結果RNAとして正しい配列類似性が得られないということである．したがって，RNA配列比較のためのアラインメントは，二次構造と配列類似性を総合して行う必要がある．第2の問題は，たとえ二次構造と配列類似性を総合したアラインメントが与えられた場合でも，類似度の尺度を正しく選ばないと，二次構造が同一の配列間よりも，二次構造がまったく異なる配列間の類似度の方が高くなってしまうことである．以上から，RNA配列の検索のためには，二次構造を考慮した整列法と類似尺度が必要であることが分かる．二次構造が既知の場合には，榊原氏の概説にあるように，木構造による二次構造の比較が可能であるが，二次構造が未知のRNA配列の比較は簡単ではない．

配列から特徴量を抽出して比較する手法は，一見乱暴なようだが，配列アラインメントを直接求めることがむずかしい場合には有効な手法となり得る．近年になって，サポートベクターマシン（SVM）をはじめとするカーネル法が生物配列の解析においても流行し，文字や単語を数えてそれを特徴量とするようなカーネルが，タンパク質の機能や細胞内での局在性などの予測に用いられた．確率モデルとカーネル法を組み合わせて用いる方法も現れ，筆者らはSCFGとカーネル法（周辺化カーネル）を組み合わせた新しいRNA配列解析の手法を提案した[6]．SCFG上の周辺化カーネルにおいては，文法上同じ非終端記号に対応する塩基もしくは塩基対の確率的な頻度が比較され，配列間の潜在的な共通二次構造に基づく類似度がすべての構造について確率的に平均化した値として得られる．その値はどのような文法を用いるかによって影響を受けるが，文法自体は必ずしも多くの相同性配列群による学習を必要としないのが利点である．

今後の課題

本稿では，ゲノム配列からの遺伝子発見，機能アノテーション，制御配列解析，機能性RNA発見，比較ゲノムなどの配列情報解析の課題と，その根底にある配列比較の問題について述べてきた．バイオインフォマティクスの研究者の間でも，配列情報解析は過去のもので，ポストゲノム時代においてはより高次のデータの解析が重要だと考える人々が多いように思われる．しかし，今日のバイオインフォマティクスの活況の背景には，数多くのゲノム配列が利用可能になったことがある．DNAマイクロアレイやタンパク質相互作用，RNA干渉を用いた遺伝子ノックアウト実験などの結果は，改めてゲノム配列情報と照らし合わせて解釈され，ゲノム配列の「解読」が進むのである．

今後，さらに数多くのゲノム配列が決定され，その際には，遺伝子発見と機能アノテーションをより高速・正確に行う技術が要求される．遺伝子発見の基礎となる統計情報によるコード領域の予測は壁に突き当たった感があるが，転写単位の予測，スプライス機構のより詳細な理解による予測，比較ゲノムによる遺伝子発見などは，まだまだ発展するべき課題である．タンパク質コード遺伝子機能の予測は，既知の遺伝子との類似性が高くない限り，タンパク質立体構造を予測し，その機能を解明する研究と不可分である．従来は，実験による構造決定や機能解析に待たねばならない状況であったが，今後は，配列比較とクラスタリング，相同性による立体構造予測，分子計算による立体構造予測，既知の情報のデータベース化とその利用といった総合的なシステムによって機能アノテーションが行われるようになるであろう．

本稿では，比較ゲノムの課題について深く触れる機会がなかったが，個々の遺伝子の配列比較ではなく，ゲノム配列全体がどのように入れ替わり，組み合わされて現在に至ったのか，ゲノムレベルでの進化そのものを研究の対象とする時代が到来している．麹菌とAspergillus属との比較ゲノムにおいても，個々の遺伝子の進化上の時間的距離と，ゲノム構造の進化上の時間的距離の間にどのような関係があるのか，興味深い現象が見つかっている．

機能性RNAについては，現在大ブームになりつつあるが，その配列解析技術は未熟で，発展途上である．ゲノム配列やcDNAから，効果的に機能性RNAを発見したり，クラスタリングしたりするための確立した手法は存在しない．筆者らも，確率文脈自由文法上の周辺化カーネルによるソフトウェア（Sokos），ステム候補列のアラインメントによるRNA配列比較・検索ソフトウェア（Scarna）などを提案しているが，手法に関する本格的な競争はこれからである．

参考文献
1）Machida, M. et al.: Genome Sequencing and Analysis of Aspergillus Oryzae, Submitted in 2005.
2）Asai, K., Itou, K., Ueno, Y. and Yada, T.: Recognition of Human Genes by Stochastic Parsing, Pacific Symposium on Biocomputing 98, pp.228-239（1998）.
3）Gotoh, O.: Homology-based Gene Structure Prediction: Simplified Matching Algorithm Using a Translated Colon (ton) and Improved Accuracy by Allowing for Long Gaps, Bioinformatics, 16,(3) , pp.190-202（2000）.
4）Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. and Eddy S. R.: Rfam: an RNA Family Database, Nucleic Acids Research, 31: pp.439-441（2003）.
5）Lowe, T. M. and Eddy, S. R.: tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence, Nucleic Acids Research, 25: pp.955-964（1997）.
6）Kin, T., Tsuda, K. and Asai, K. : Marginalized Kernels for RNA Sequence Data Analysis, Genome Informatics, 13, pp.112-122（2002）.

（平成17年2月9日受付）