1 ゲノム情報科学における情報科学的諸問題

阿久津 達也
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／
京都大学大学院情報学研究科知能情報学専攻
takutsu@kuicr.kyoto-u.ac.jp

情報科学としてのゲノム情報科学

ゲノム情報科学は生物学と情報科学の境界領域の学問であるが，はたして情報科学として価値があるのであろうか？　情報科学から見れば単なるアプリケーションの1つに過ぎないのではないか？　この問いに対する筆者なりの答えは最後の章で述べることにして，まず，生物学と情報科学の違いについて，研究内容ではなく，学会や論文などの観点から見ていこう．なお，本稿では特集タイトルに合わせてゲノム情報科学という用語を使用するが，本稿ではバイオインフォマティクスの同義語として使用する．

もともと情報系の学会を発表の場としていた筆者にとって，ゲノム情報科学分野で研究を始めて驚いたことがあった．それは，情報系では，研究成果の発表は研究会や国際会議で発表した後，論文誌に投稿するというのが一般的であるのに対し，生物系では，いきなり論文投稿するのが一般的であるということであった．もちろん生物系でも国内外で会議は開催されるが，口頭発表できるのは招待講演者のみで，その他の参加者はポスター発表のみというスタイルが一般的である．それに対して情報系の国際会議では10ページ程度のfull paperに近い論文を投稿し，採択された論文の口頭発表が会議の中心となるのが一般的である．この理由はいろいろとあるのだろうが，学会誌や論文誌などにおける論文投稿から採録決定，掲載までの期間によるところが大きいと思われる．また，生物系では，full paperに近いものをいったんどこかで発表してしまうと，それを論文誌に掲載できない，という制約があるらしい．情報系では，国際会議で発表したものを，後に論文誌で発表するのはむしろ奨励されることである．

それでは，ゲノム情報科学はどちらのスタイルなのだろうか？　これがまさに折衷スタイルなのである．ゲノム情報科学にはISMB，RECOMB，PSBといった主要な国際会議があるが，これらの会議では，full paperに近い投稿論文から審査を行って採択された論文の口頭発表もあり，審査のほとんどないポスター発表もある，というスタイルなのである．生物系出身の研究者と情報系出身の研究者がいるため，どちらの分野からでも参加できるようにとの配慮からこのようなスタイルになったようである．

それでは，ゲノム情報科学の研究室は情報系学科，生物系学科のどちらに属していることが多いであろうか？　多くを知っているわけではないが，国内外のどちらにおいても，生物系の学科に属していることが多いようである．特に情報系学科では，ゲノム情報科学のみの研究を行っている研究室は非常に少ないのが現状である．情報系学科においては，アプリケーション

の1つと認識されていることが多いようである.

さて，これまでのことを総合すると，「ゲノム情報科学は，まだ，情報科学の一分野にはなっていない」ということになるであろう．それでは，「ゲノム情報科学は，将来，本当の情報科学になり得るのか」というと，その答えは筆者には分からない．しかし，「情報科学という観点から価値がある問題があるか？」と聞かれれば，少なくとも筆者にとってはYESと答えることができる．本稿では筆者が情報科学として面白いと感じている問題について紹介していく.

配列比較における未解決問題

理論計算機科学や数学などの分野においては未解決問題 (Open Problem) が研究の進展のための大きな原動力となってきた．ゲノム情報科学におけるアルゴリズム的側面を扱った教科書[1] によると，PevznerとWatermanは1995年にゲノム情報科学における57個の未解決問題を提示したが，その約1/4が2000年までに解決されたと書いてある．また，この本には117個の問題が列挙されている．残念なことに，そのうち，いくつが解決済みかは明記されていないが，半分程度は未解決なのではないかと思われる.

ところで，ゲノム情報科学において，最適アライメント問題は古典的かつ最重要なアルゴリズム的問題である．最適アライメント問題は一言でいうと，2個 (以上の) 文字列が与えられた時に，文字間の対応関係を求める問題，つまり，文字列パターンマッチング問題の一種である．2個のDNA配列もしくはアミノ酸配列の類似性を判定するには，このアライメントを計算することが必須であり，理論的観点からも実用的観点からもさまざまな研究が行われてきた．しかしながら，長く研究されてきたわりには未解決の (と思われる) 問題が多い．そこで，この問題を例として何が解決され，何が解決されていないかを説明していこう[1].

DNA配列を例にとり，2個の文字列のアライメント問題を定義しよう．ここではDNA配列を考えているので，各文字列はA,C,G,Tの4種類の文字から構成される．この4種類の文字の各ペアの間にはスコア (類似度) が定義されるが，ここでは簡単のために同じ文字 (X) 間のスコアは1とし ($s(X,X)=1$)，異なる文字 (X,Y) 間のスコアは-1としよう ($s(X,Y)=-1$)．ここで，2個の文字列が与えられた時に，各文字列にギャップ記号 (-) を挿入して，2個の文字列の長さが同じになるようにする．たとえば，AGCTとTCGCTの場合には,

```
A-GCT
TCGCT
```

というようにギャップ記号を入れれば，長さが同じになる．このようにギャップを入れて長さを同じにしたものは「アライメント」と呼ばれ，同じ列の文字どうしが対応しているものとみなされる．そして，列ごとのスコアを加算していくことにより，アライメント全体のスコアを計算することができる．上の例だと，$s(A,T)+s(-,C)+s(G,G)+s(C,C)+s(T,T)$がスコアであり，その和は1となる．ただし，ここでは簡単のため，ギャップ文字と他の文字の間のスコアは常に-1であるものとした ($s(X,-)=s(-,X)=-1$).

ところで，上の例では1個のアライメントを考えたが，以下のようにほかにもさまざまなアライメントを考えることができる.

```
-AGCT      --AGCT     AGC--T
TCGCT      TC-GCT     -TCGCT
```

なお，アライメントにおいてはギャップ記号どうしが同じ列に並んではいけないものとする．前の例では，スコアの和は1であったが，この3個の例ではスコアの和はそれぞれ 1, 0, -2となる．最適アライメント問題は，2個の文字列が与えられた時に，スコアの和が最大となるアライメントを計算する問題である.

最適アライメント問題を解くためのアルゴリズムとして，すべてのアライメントを作り，それぞれのスコアを計算し，最大スコアのアライメントを求める，というアルゴリズムが考えられる．残念ながら，このアルゴリズムは非常に効率の悪い (指数オーダーの) アルゴリズムとなってしまう．というのも，アライメントの個数は指数オーダーであることが知られているからである.

そこで，効率の良いアルゴリズムが必要となるが，20年以上昔より，動的計画法という一般的な手法に基づくアルゴリズムが知られている．ここでは，そのアルゴリズムを簡単に説明しよう．このアルゴリズムでは，アライメント問題を図-1のような2次元のメッシュ状のダイアグラム (有向グラフ) における最長経路問題に変換する．この図では，横方向は1個目の入力文字列に対応しており，縦方向が2個目の文字列に対応している．また，右下に向かう各矢印には，対応する文字間

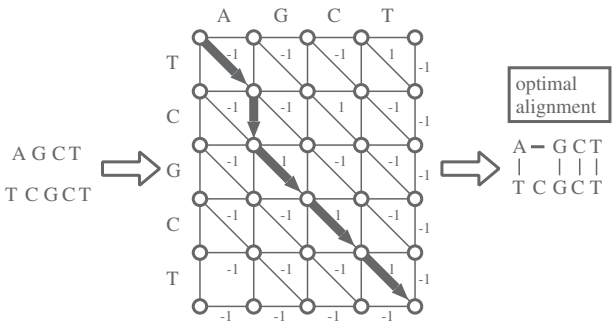

図-1　最長パス問題への変換による最適アライメントの計算

のスコアが対応付けられている．たとえば，左上の矢印には，AとTとのスコア -1 が対応付けられる．また，横方向の矢印および縦方向の矢印にはギャップ文字1文字分のスコアが対応付けられている．すると，アライメントと，左上の点から右下の点までのパス（経路）には，1対1対応が存在し，また，パス上のスコアの和は対応するアライメントのスコアに等しくなる．よって，長さ（スコアの和）が最長となる経路を求めればよい．一般に最長経路問題は難しい問題であるが，図-1のグラフは規則的で単純な形をしているため，動的計画法という手法により比較的効率良く計算することができる．

この計算時間や必要メモリについて簡単な計算をしてみよう．2個の入力文字列の長さをそれぞれ，n，mとしよう．すると，図-1のように作られるグラフには $(n+1)\times(m+1)$ 個の点が存在することになるため，計算時間やメモリも $(n+1)\times(m+1)$ 程度となる（$O(nm)$時間，$O(mn)$スペース）．nとmが同じくらいとすると，これは $O(n^2)$時間，$O(n^2)$ スペースのアルゴリズムとなる．

実はこれまでの議論はそれほど難しいことではなく，アルゴリズムについての基礎を勉強した学生にとっては，期末試験や大学院入試レベルの問題と言えるであろう．ところが，もっと効率の良いアルゴリズムはあるか？ということになると，とたんに難しい未解決問題となってしまうのである．筆者の知る限り，$O(n^2)$時間より本質的に効率の良いアルゴリズムは知られておらず，また，$O(n^2)$時間が最適である，という結果も知られていない．ただし，メモリ効率に関しては，$O(n)$のアルゴリズムがかなり昔より知られている．

これまでは，2個の文字列の比較を考えてきたが，実際には，1個の配列を入力し，データベース中の数百万の配列とのアライメントを計算し，最適なものを1個（もしくは数十個程度）求めるというように利用される場合が多い．この場合，各配列の長さをn程度とし，データベース中の配列の個数をNとし，上記の動的計画法に基づくアルゴリズムを繰り返し適用すると$O(n^2N)$時間かかることになる．この問題に関しては，FASTAやBLASTなどのハッシュ法に基づく実用的なアルゴリズムが開発され，生物学者らによって日常的に使用されるに至っている．しかしながら，本質的に$O(n^2N)$より効率の良いアルゴリズムは知られておらず，この改善は未解決問題であると思われる．また，ギャップを許したマッチのためのハッシュ関数で理論的保証があり，かつ，実用的なものは知られていないと思われる．

2個の文字列に対するアライメントは$O(n^2)$時間でできたが，k個の文字列に対するアライメントはk次元のメッシュ状グラフ上の最長経路問題に変換することにより，$O(n^k)$時間で計算できることが知られている．このアルゴリズム自体は大学院入試レベルのアルゴリズムである．しかしながら，この計算量はkがちょっとでも増えると現実的でなくなるので，より効率の良いアルゴリズムを開発することが求められる．すると，それはとたんに難しい未解決問題となり，現在のところ本質的に$O(n^k)$時間より効率の良いアルゴリズムは知られていないということになる．また，kが固定されていない場合には，一般に，NP困難という難しいクラスの問題になることも知られている．NP困難な最適化問題があった場合に精度保証のある近似アルゴリズムを研究開発するというのは情報科学における一般的な研究スタイルであるが，この問題については，ある定式化のもとで近似率2☆1というアルゴリズムが1993年に発表された．しかしながら，ささいな改良はあるものの，これまで本質的に2という近似率は改良されていない．これも有名な未解決問題である．なお，2000年に，特殊な場合に限っては近似率を1（最適解）にいくらでも近づけることのできるアルゴリズム（専門用語ではPTASという）が，コロモゴルフ計算量などでも有名なMing Liらにより開発された[2]．

このように，これまで多くの研究が行われてきたアライメントに限っても，情報科学的観点から見ればさまざまな未解決問題があることが分かる．

☆1　最大化問題の場合には，近似率は $\frac{\text{最適解の値}}{\text{近似解の値}}$ の上限となる（最小化問題の場合は分子，分母を入れ替える）．

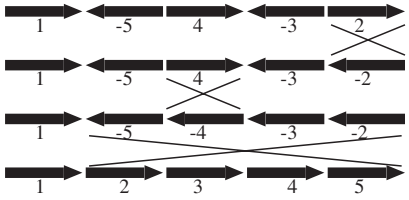

図-2　ゲノムリアレンジメント問題．この図には，一番上の列（キャベツのミトコンドリアDNA配列）から一番下の列（カブのミトコンドリアDNA配列）へ最小回数で移るための反転系列が示されている．

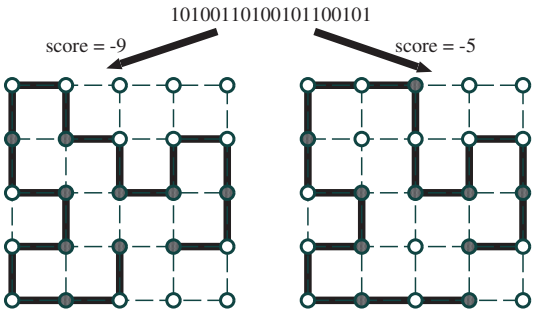

図-3　ストリングフォールディング問題．左側が最適（最小）スコアの埋め込みとなっている．なお，図中では白丸が0を表し，黒丸が1を表す．

アルゴリズム論的未解決問題

配列比較以外にも，ゲノム情報科学にはアルゴリズム論的な未解決問題は数多く残されている．その中から，筆者がある程度なじみのあるものを中心にいくつか紹介する．なお，ページ数の関係があるため，問題の厳密な定義は省略するので，詳細については文献を参照してほしい．

(1) 進化系統樹推定

これはクラスタリングなどにも関係する問題であり，複数の生物種の遺伝子配列が与えられた時に，その配列を比較し，進化の過程を表す木を再構成するという問題である．この問題に関しては，配列間の距離がある制約を満たす場合には，近隣結合法というアルゴリズム（本特集の斎藤氏はその開発者の一人である！）により木が正しく再構成されることが知られている．しかしながら，より一般の距離や，また，最節約法といった基準を用いた場合には，再構成が困難になることが知られている[3]．そこで，最適解を計算する実用的なアルゴリズム，もしくは，精度保証のある近似アルゴリズムを開発することが望まれるが，決定版と呼べるようなアルゴリズムは知られていない．

(2) ゲノムリアレンジメント

複数の生物種のDNA配列を比較すると，進化の過程で，DNAの一部分が反転したり，他の部分に移動した形跡があることが分かる．特に，近い生物種の場合に，遺伝子配列はほとんど同じで，遺伝子の順番や向きが違うだけであるような例がいくつも知られている（図-2[1]）．そこで，2種類の生物種の配列が与えられた時，一方の配列から他方の配列へ変換するための最小回数もしくは最小コストの反転，移動系列を見つけるという問題が考えられる．この問題については，1990年代によく研究され，いくつかの興味深い結果が得られたが，近似精度などに関してはまだ未解決の問題が多い[1]．なお，（ゲノム情報ということを意識していたわけではないが）ビル・ゲイツがこの問題に関連したテーマについての学術論文を書いたことがある，というのは知る人ぞ知る話である[1]．

(3) ストリングフォールディング問題

これは梅山氏の解説にあるタンパク質立体構造予測問題を極端に単純化した問題であり，二進文字列（0と1からなる文字列）をスコアが最小になるように平面格子(もしくは立方格子)上に埋め込むという問題である[4]．スコアは以下のように計算される（図-3）．1と1が格子状で隣り合う位置にある時は -1 であり，それ以外の時には0とし，それらをすべての文字のペアについて足したものが全体のスコアとなる．ただし，もとの文字列上で隣り合う位置にある文字どうしのスコアは0とする．この問題も1990年代によく研究され，NP完全性の証明や近似アルゴリズムの開発などが行われた．しかしながら，最初の近似アルゴリズムが発表されて以来，近似精度の本質的な改善がなされておらず，未解決問題となっている．

統計的情報処理に関する課題

ゲノム情報科学においては，HMM（隠れマルコフモデル）やクラスタリング，混合分布など，さまざまな確

率モデルや統計的手法が応用されている[5]．しかしながら，単なる応用ではうまくいかない場合も多い．そこで，ゲノム情報科学において解決すべき（と思われる）統計的情報処理に関する課題を紹介する．

(1) HMMの学習

HMMは，もともと音声認識などの分野で発展してきた確率モデルであるが，1990年代に入り，ゲノム情報科学にも盛んに応用されるようになった（図-4）．しかしながら，パラメータ推定に利用されるアルゴリズムは，局所最適解しか探索できないEMアルゴリズムがほとんんどであり，推定されるパラメータが初期値に大きく依存してしまう場合が少なくない[5]．その改善のためにさまざまな研究が行われているものの，EMアルゴリズムにとってかわるアルゴリズムはいまだ開発されていない．また，HMMの構造の学習に関しては，いくつかの発見的なアルゴリズムが知られているのみで，実用レベルに達しているアルゴリズムは開発されていないように思われる．また，隠れマルコフモデルの拡張である確率文脈自由文法もRNAの構造推定などに応用されているが，効率や局所最適解に陥りやすいなどの問題点は解決されていない．

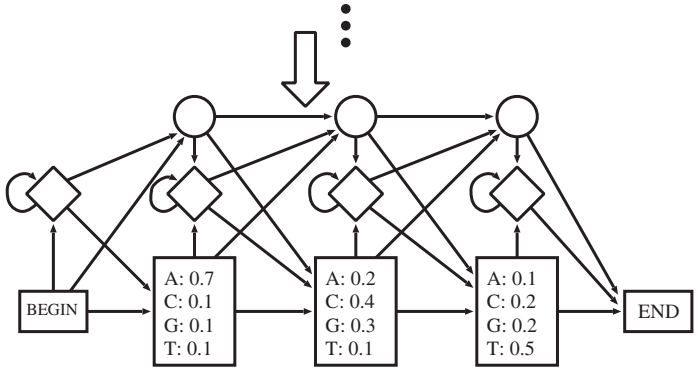

図-4　配列からの隠れマルコフモデル（HMM）の学習．ゲノム情報科学で使用されるHMMの構造は，この図のような形式をしていることが多い．

(2) 誤分類率の最小化

正負の学習データ（たとえば，ある共通の性質を持った遺伝子配列の集合と，それ以外の遺伝子配列の集合）が与えられて，それらを区別できるような配列パターンやパラメータセットを学習する問題はゲノム情報科学において頻繁に現れる．しかしながら，（パターンやパラメータに関する制約のもとで）完全に分類できる場合は非常に少ない．その場合には，誤って分類されるデータの個数を最小化することが望まれるが，残念ながらある種の定式化のもとで困難（NP困難）であることが知られている[6]．近似アルゴリズムなども研究されているものの実用レベルには達していない．サポートベクタマシンを用いた学習でも同種の困難に直面することがあり，ある種の発見的な手法により解決しているが，本質的な解決にはなっていないように思われる．

(3) 確率的アルゴリズムの収束性および精度

混合分布推定，クラスタリング，HMMのパラメータ学習のためのEMアルゴリズム，共通部分配列検出のためのGIBBSサンプリングアルゴリズムなど，ゲノム情報科学においてはいくつもの統計的，確率的なアルゴリズムが開発，もしくは，応用されている[5]．しかしながら，それらのアルゴリズムの収束性（たとえば，多項式回で収束するか？）や，求まる局所最適解の精度などに関しては何の理論的保証もないように思われる．

シミュレーションにおける課題

原子や分子レベルの分子動力学から，代謝反応経路，さらには細胞や個体に至るまでと，ゲノム情報科学はシミュレーションの対象の宝庫である．しかしながら，分子動力学法以外に関してはアドホックなアプローチが多く，課題だらけである．

ここでは細胞のシミュレーションなどに利用される方法を中心に課題を述べる．細胞のシミュレーションなどでは，細胞内部のタンパク質や化学物質の量や濃度の関係を微分方程式などを用いて記述し，差分法などによる数値解法を適用して時間変化を計算するということが行われる[7]．方程式としては，Michaelis-Menten式や質量作用則などが用いられるが，これらを細胞のシミュレーションに用いることの妥当性が（実験データなどにより）証明されていない．

また，用いる方程式が妥当であったとしても，数値解法の妥当性（誤差，ロバスト性）などについての考察

もほとんど行われていない．その結果，「パラメータをうまく調整したら，現実の変化に近いシミュレーション結果が得られました」といった成果はよく得られるものの，安心して予測に使えるレベルには達していない．また，シミュレーションに使うパラメータの調整もトライアンドエラーによるものがほとんどで，自動調節のための確立した手法はないというのが現状である．

よって，上に挙げた課題を克服し，生物のシミュレーション技術を，航空機や自動車，半導体など実用レベルに達しているシミュレーション技術にいかに近づけるかは情報科学的観点からも重要であると思われる．この課題に対してはERATO/SONY CSLの北野氏がシステムバイオロジーの名のもとに精力的な研究を展開しようとしている．北野氏が書いた意欲的かつ詳細な本があるので，興味のある読者は，その本をぜひ参照されたい[7]．

人間のデータ圧縮

60兆個といわれる細胞の集合体である人間を再構成するのに必要な情報が，たった30億文字（CD-ROM1枚分ちょっと）に格納されているのは，情報科学的観点からも驚きである．さらに，最近の解析結果では，遺伝子（DNA配列中で主にタンパク質に翻訳される部分）の個数はたった3万～4万程度ではないかと言われている．

そこで，「どのようにすれば，30億文字に人間を再構成できる情報を格納できるか？」という問題を考えることができる．この問題の最終的な目標は，60兆個の細胞について，種類，空間配置，接続関係を30億文字の中に収め，かつ，個人の違いなどの多様性なども一部の文字を置き換えることにより，再現できるような，情報圧縮・情報表現形式（文法のようなもの）を考えることである．そして，その文法のようなものが生物学的に意味があるかどうかは問わないものとし，発生の過程などは現実のものと違ってよいものとする．情報圧縮が可能かどうかだけを研究の目標とするのである．このような問題に取り組む利点の1つは，（細胞の配置など以外には）実験結果などを必要とせず，生物学者との共同研究も必須ではなく，純粋に情報科学的問題として取り組むことができることである（ただし，生物学者との共同研究を否定するわけではない．もしろ，積極的に行いたいと思っている）．もちろん，いきなり60兆個の細胞を相手にするのではデータ収集だけでも不可能に近いので，まずは，循環器や神経系の配線図を多様性まで含めて再構成できるかような表現形式に取り組むのは面白いのでは，と考えている．もちろん，これまでに研究されてきた，グラフ文法，L-system，人工生命なども似たような動機に基づいているのかもしれないが，それらの分野では，いつの間にか応用，人工進化，理論的詳細などが研究の主流となってしまって，上に挙げた目標とはあまり関係がなくなってしまっているように思われる．

少々脱線しマッドサイエンティストぎみなことを書いてしまったが，最後に，ゲノム情報科学は情報科学として価値があるかについて，（ひとりよがりな答えではあるが）筆者なりの答えを書いてみたい．この解説であげたように，ゲノム情報科学には，情報科学として価値のある問題は少なからずあると思われる．しかしながら，「統計情報処理」の節までに示したような，きちんと定式化でき，よく研究されている問題については，（最適解を出力できるかどうかは別として）ある程度実用的なプログラムがすでに開発されてしまっており，少なくとも筆者にとって，あまり面白い問題ではない．それに対し，後半に示した問題は，定式化もできていないし，どう取り組めばよいかも定かではなく，情報科学として筋の良い問題ではないかもしれない．でも，情報科学として価値があるかないか分からないような問題，情報科学として筋の悪そうな問題に，情報科学的センスを持って取り組むことこそ，ゲノム情報科学の面白さがあるのでは，と考えている．

参考文献

1) Pevzner, P. A.: Computational Molecular Biology. An Algorithmic Approach, The MIT Press (2000).
2) Li, M., Ma, B. and Wang, L.: Near Optimal Multiple Alignment within a Band in Polynomial Time, Proc. 32th ACM Symp., Theory of Computing, pp.425-434 (2000).
3) Gusfield, D.: Algorithms on Strings,Trees, and Sequences, Cambridge University Press (1997).
4) Hart, W. E. and Istrail, S. C.: Fast Protein Folding in the Hydrophobic-hydrophilic Model within Three-eights of Optimal, Journal of Computational Biology, 3, pp.53-96 (1996).
5) 阿久津，浅井，矢田（訳）:バイオインフォマティクスー確率モデルによる遺伝子配列解析ー，医学出版 (2001).
6) Amaldi, E. and Kann, V.: On the Approximability of Minimizing Nonzero Variables or Unsatisfied Relations in Linear Systems, Theoretical Computer Science, 209, pp.237-260 (1998).
7) 北野宏明: システムバイオロジー，ー生命をシステムとして理解するー，秀潤社 (2001).

（平成13年11月15日受付）