1 バイオインフォマティクス概説
― 比べることで生命は解明できるか？ ―

榊原 康文
慶應義塾大学理工学部生命情報学科
yasu@bio.keio.ac.jp

Doolittle による相同性検索を用いたがん遺伝子の発見以来，「比べる」という戦略はバイオインフォマティクスの研究の中心であった．配列の比較解析にはじまり，最新の研究においても，発現プロファイルの比較，ネットワークの比較，比較ゲノムなど比べるという研究手法はますます重要さを増している．本稿においては，この比べるという視点から，バイオインフォマティクスの研究概要と発展について述べる．

生体分子配列の比較と解析

生物を形作る設計図は，DNA という生体分子に書かれていることは，今日誰でも知っている事実である．また，DNA に書かれている情報は，親から子へ伝わる遺伝情報であり，細胞が分裂するときに同じものがコピーされることも知っている．DNA 上にどのように情報が書かれているかというと，DNA（デオキシリボ核酸）は燐酸とデオキシリボースという糖が交互につながった長い鎖が骨格になっている生体高分子で，アデニン，チミン，グアニン，シトシンという窒素を含有する4 つの塩基のいずれか1つがそれぞれのデオキシリボースについている．この4 つの塩基の並び方が遺伝情報を表していて，その並び方をDNA配列と呼んだりする．単純に見てしまえば，DNA 配列は塩基を文字と見なすことにより，4 文字アルファベット（4 つの塩基の名前の頭文字を取って，A, T, G, C）上の文字列と考えることができる．2003 年の春に，人間の全DNA配列を決定したヒトゲノムの完全配列が公開され，いまその解析が全世界的に進められている．

分子生物学という，分子のレベルで生命の活動を研究していく分野においてコンピュータが欠かせない道具となっている理由の1 つは，生物のなかで分子によって表されている情報の量が膨大であることによる．たとえば，人間のDNA 配列（ヒトゲノム）の大きさは，A, T, G, C の4 文字の文字列と見なしたときに，長さ約30 億の文字列となる．この大きさの情報は，ディジタル記憶媒体のCD であれば1 枚にほぼ収まる大きさであるが，もっと分かりやすいもので例えると，新聞であれば，朝刊と夕刊を365 日毎日読み続けて約50 年かかることになる．これを，コンピュータの助けを借りることなく人間の目と手だけで処理をするのは，とても不可能である．バイオインフォマティクスの重要な研究の1 つに，DNA配列などの生命情報を保存したり検索するためのデータベースの開発がある．それは，何よりも生命活動の情報というものがこのように巨大であることがその理由である．

鎖状に並んだA, T, G, C からなる文字列としてのDNA配列を解析する最も基本的な方法は，2 つ以上のDNA配列を比較することである．実験から得られたDNA配列のデータを解析するときには，今までに分かっているDNA 配列と似ている（「相同性がある」という）かどうかを調べ，また似ている場合にはDNA配列のどの部分が同じでどの部分が違うかを調べることが重要となる．これは，DNA 配列が似ていれば遺伝子としての機能も似ているという経験的な知識に基づくものである．この並べて似ているかどうかを調べる操作を，アライメント(alignment) と言う．2つの配列をただ並べるのではなく，似ている部分どうしが一致するように並べる．

この配列比較の手法が，最初に威力を発揮した最も有名な研究が，Doolittle による相同性検索を用いたがん遺伝子の発見である[5]．この研究により，サル肉腫ウイルスのがん遺伝子シス（sis）とヒトの血小板由来増殖因子（PDGF）のアミノ酸配列が一致している（そっくりである）ことが発見された．この発見は2つの意味において驚きをもって迎えられた．その1つは，がん遺伝子が正常な細胞の増殖・分化や個体発生を司る遺伝子とほとんど同じものであることが初めて具体的に明ら

かにされ，がん遺伝子の姿が浮かび上がってきたことであり，もう1つはその発見が試験管の中の実験ではなく，コンピュータによる相同性検索の結果得られたことである[14]．Doolittle がそれまでに構築してきたデータベースと相同性検索プログラム，そして総当たりの仕事もいとわないコンピュータの存在が可能にした成果である．バイオインフォマティクスの夜明けと呼んでよいものであり，今日においてもバイオインフォマティクス研究の意義と効力を明確に伝えている優れた研究である．

この「比べる」という戦略はバイオインフォマティクス研究の中心であり続け，最新の研究においても，比べるという研究手法はますます重要さを増している．RNA 2次構造の比較，糖鎖構造の比較，代謝ネットワークやシグナル伝達ネットワークの比較，タンパク質立体構造の比較や発現プロファイルの比較，そして比較ゲノムなど，多種多様なデータの比較がますます盛んに行われている．本稿では，これらの生物学データを表す抽象的な数学的データ構造の観点から分類して，その数学的データ構造を比較する手法を考察しながら，現在のバイオインフォマティクスの技術について概観する．

動的計画法：計算機科学の最大の貢献

たとえば，次の2 つの配列の中で上にあるDNA配列が実験から得られて，これを下にある，すでに機能などが解析されているDNA配列と比較することを考える：

```
GAGGTTATCAAAAGCTACTAGTCCA
GAGGATAACAAGGCTACTATCACA
```

この2 つの配列を同じ文字が一致するように正確に並べると次のようになる：

```
GAGGTTATCAA-AA-GCTACTAGTC-CA
GAGG--AT-AACAAGGCTACTA-TCACA
```

このように2 つの配列を並べる時に，両者が一致しない場所がある場合，この場所を空として飛び越して次の配列をつなげる必要がある．この空の場所のことを，「ギャップ」と呼ぶ（上の図では，“-”という記号を入れて表している）．生物学的にはギャップは，進化の過程において突然変異などによって，その部分が新たに挿入されたまたは欠失したものと解釈される．また他の塩基に置き換わる置換も起こり得る．この挿入，欠失，置換の操作によって，ある配列から別の配列へ変換するのに必要とする操作の最小ステップ数のことを2 つの配列間のエディット距離と呼ぶ．

（大域）ペアワイズアライメントとは，2 つのDNA 配列に対して，適切な位置にギャップ記号を挿入することで，配列中の同じ位置に同じ塩基（あるいは性質がよく似た塩基）が並ぶようにする操作のことである．さらに，マルチプルアラインメントとは，3 本以上の複数の配列に対して，同じ塩基（あるいは性質がよく似た塩基）ができるだけ同じカラムに来るように，適切な位置にギャップ記号を挿入して各配列を並べたものであり，ペアワイズアライメントを3 本以上の配列に拡張したものである．2 つの配列の一致している部分が最大になるような大域ペアワイズアライメントを高速に求めるためのコンピュータプログラムは，動的計画法というアルゴリズムを用いて作成することができる．

今，$X = x_1x_2 \cdots x_m, Y = y_1y_2 \cdots y_n$ を入力列とする．また，$c[0..m, 0..n], b[0..m, 0..n]$ を2次元の配列とする．ここで，$c[0..m, 0..n]$ は2つの配列間の類似度を計算するための配列であり，$b[0..m, 0..n]$ はアライメントを生成するときにバックトレースするための配列である．この配列 $c[0..m, 0..n]$ は，次のように再帰的に定義される：

$$c[i, j] = \max \begin{cases} c[i-1, j-1] + s(x_i, y_j), & -\ \text{case}(1), \\ c[i-1, j] + d, & -\ \text{case}(2), \\ c[i, j-1] + d, & -\ \text{case}(3), \end{cases}$$

初期化： $c[0, 0] = 0,\ c[i, 0] = i \times d,\ c[0, j] = j \times d.$

ここで，$s(x_i, y_j)$ は2つの文字 x_i と y_j の間の置換度を表す．実際には，核酸とアミノ酸に対する統計的な置換行列を用いる．dはギャップを挿入するときのギャップスコアを表す．また配列 $b[1..m, 1..n]$ は，

$$b[i, j] = \begin{cases} \text{“}\nwarrow\text{”} & \text{if case}(1), \\ \text{“}\uparrow\text{”} & \text{if case}(2), \\ \text{“}\leftarrow\text{”} & \text{if case}(3). \end{cases}$$

初期化： $b[i, 0] =$ “↑”，$b[0, j] =$ “←”

のように定義される．このように定義された再帰式は，一般に，動的計画法を用いて効率よく（多項式時間で）計算することができる．

例題として，2 つの短い配列AGCGTAG とGTCAGAに対する大域アライメントと，そのアライメントを動的計画法を用いて計算した時の配列 c と b の値は**図-1** のようになる．

数学的な議論を少しすると，動的計画法は，対象とする問題のクラスが次の2 つの性質を満たすときに効率よく（多項式時間で）その問題を解くことができる：(1) 全体の問題に対する最適解は，その中に部分問題

大域アライメント：

```
AG-C-GTAG
-GTCAG-A-
```

配列　$c(7, 6)$, $b(7, 6)$：

	j	0	1	2	3	4	5	6
i			G	T	C	A	G	A
0		0	0 ←	0 ←	0 ←	0 ←	0 ←	0 ←
1	A	0 ↑	0 ↑	0 ↑	0 ↑	1 ↖	1 ←	1 ↖
2	G	0 ↑	1 ↖	1 ←	1 ←	1 ↑	2 ↖	2 ←
3	C	0 ↑	1 ↑	1 ↑	2 ↖	2 ←	2 ↑	2 ↑
4	G	0 ↑	1 ↖	1 ↑	2 ↑	2 ↑	3 ↖	3 ←
5	T	0 ↑	1 ↑	2 ↖	2 ↑	2 ↑	3 ↑	3 ↑
6	A	0 ↑	1 ↑	2 ↑	2 ↑	3 ↖	3 ↑	4 ↖
7	G	0 ↑	1 ↖	2 ↑	2 ↑	3 ↑	4 ↖	4 ↑

図-1　配列 AGCGTAG と GTCAGA に対する大域アライメント（上）と配列 c と b の値（下）

に対する最適解を含んでいる（optimal substructure），(2) 部分問題の空間が十分小さい（異なる部分問題の数は，入力サイズの多項式くらいの大きさ）（overlapping subproblems）．このような特徴を持つ問題の解を計算する上で，動的計画法が採る戦略は，各部分問題を一度だけ解いて，テーブルに確保して必要になった時に参照することである．配列の大域アライメントを求める問題は，数学的には最長共通部分列（LCS）という問題に定式化され，このLCSは次のように2つの数学的特徴を持っている：

2つの配列 $X = x_1 \cdots x_m$, $Y = y_1 \cdots y_n$ に対して，$Z = z_1 \cdots z_k$ を X と Y のLCSとすると，

1. $x_m = y_n$ ならば，$x_m = y_n = z_k$ であり，かつ Z_{k-1} は X_{m-1} と Y_{n-1} のLCSである．
2. $x_m \neq y_n$ ならば，$z_k \neq x_m$ のとき，Z は X_{m-1} と Y のLCSである．
3. $x_m \neq y_n$ ならば，$z_k \neq y_n$ のとき，Z は X と Y_{n-1} のLCSである．

動的計画法は，大域アライメントの問題から始まって，局所アライメントや繰り返し一致アライメント，アフィンギャップアライメント，より複雑なデータ構造である木構造データやグラフ構造データのアライメント，さらに隠れマルコフモデルや確率文脈自由文法の構文解析などの問題に幅広く応用され，これらの問題を効率よく解くための手法として大活躍している[1)]．

ポストゲノムにおける比較解析

比較解析という戦略はバイオインフォマティクス研究の中心であり続け，最新の研究においては，多種多様なデータの比較がますます盛んに行われている．これらのさまざまな生物学データを，それに対応する数学的データ構造の観点から分類して，比較する手法を考察する．

線形構造データの比較

分岐のない直線状の構造をとる生体分子としては，DNA，RNA，そしてタンパク質がある．いわゆるセントラルドグマにおける最も重要な3つの生体高分子である．まずはじめに，最も単純な線形構造（グラフ理論的には分岐とループを持たない有向グラフ）の比較解析とその手法について見ていく．この中で，RNAとタンパク質は複雑な立体構造をとり，酵素などの触媒作用やさまざまな生体機能をつかさどるので，高次構造の比較も必要となってくる．

遺伝子レベルの比較

「遺伝子」という単位は，タンパク質をコード化する配列として一般に定義されている．遺伝子の情報に関連する配列としては，タンパク質のアミノ酸配列，cDNA（mRNAから逆転写によって得られる相補的なDNA配列），EST（mRNAからDNAシーケンサーによって1回だけ配列決定されたcDNAの断片配列），ゲノム上の遺伝子コード領域やORFなどがある．これらの配列をデータベースから検索する方法としては，BLASTやFASTAなどのソフトウェアが有名である．BLASTの仕組み（Aho-Corasickアルゴリズム）やe-valueなどのスコアについては，数多くある他の文献（たとえば，文献1），15），17），18），など）にゆずりたいが，基本的には相同性検索を実行するために先ほどの動的計画法を用いた配列アライメントを計算している．また，先ほどの配列アライメントの説明では，核酸間の置換度として0, 1の単純なものを用いたが，BLASTなどでは進化上の置換頻度を統計的に求めたBLOSUMやPAMなどのスコア行列が用いられる．

ゲノムレベルの比較

世界的なゲノム配列決定プロジェクトにより，現在までに約200種以上の生物のゲノム（その生物が持つ全DNA塩基配列）の配列が決定され，今後も多くのモデル生物のゲノム配列決定プロジェクトが進行中である．ゲノムの大きさも，微生物の100万塩基対程度からヒトゲノムの30億塩基対，さらに植物ではより長い配列を

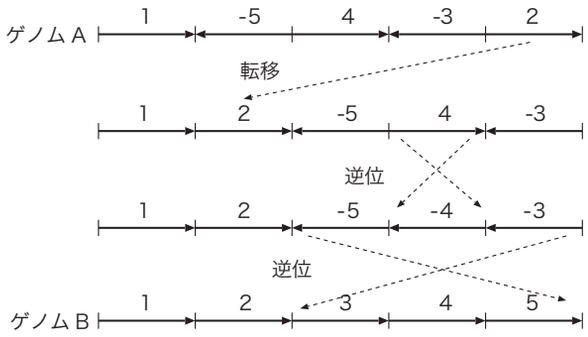

図-2　ゲノムAからゲノムBへ1回の転移と2回の逆位によるゲノム再編成.

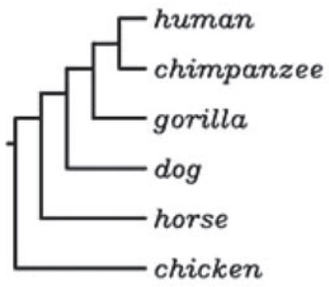

図-3　有根進化系統樹の例

持つなど，さまざまである．このようにゲノム配列が決定されてくると，遺伝子のレベルではなく，もっと大きな単位，たとえば染色体レベルでの比較解析が可能になってくる．比較ゲノムと呼ばれる比較的新しい分野においては，ヒトとマウスのゲノムを染色体レベルで比べたり，シンテニーと呼ばれる遺伝子よりも大きなブロック単位で比較解析を行う研究が盛んである．さらに，遺伝子やシンテニーブロックの位置が移動する転移や逆位，染色体の分裂または融合などのゲノム再編成という解析も可能になってくる（**図-2** 参照）．遺伝子のオーソログ（近縁の生物種間の同一遺伝子）やシンテニーブロックの検出には，相同性検索の手法がおもに用いられるが，転移や逆位をともなうゲノム間距離の計算には非常に複雑なアルゴリズムが必要となる（文献2）などを参照）．また，ゲノム配列に対するアノテーション情報や多生物種間の配列比較に関して，最新かつ精度の高いデータを提供するデータベースであるUCSCゲノムブラウザは，ゲノムレベルでの比較解析において強力な道具となる[25]．

非コード領域の比較

ゲノム配列が決定されると，タンパク質をコード化していない領域の解析も重要な課題となってくる．これらの非コード領域には，非コードRNA（タンパク質をコード化するDNA配列はメッセンジャーRNAに転写されるが，メッセンジャーRNA以外のRNAは非コードRNAまたは機能性RNAと呼ばれる）や遺伝子の転写と発現を調節するプロモータ領域，SINEやLINEなどの反復配列などが含まれる．これらの非コード領域における配列解析は，統一的な手法はなく，それぞれの問題に適した手法やアルゴリズムが開発されている（非コードRNAの比較解析については，次の節で詳しく述べる）．たとえば，プロモータ領域における転写の重要な因子として，転写因子が結合する転写結合部位があり，その検出や同定は重要な課題である．数文字の塩基配列からなる結合部位は，シグナル配列やコンセンサス配列と呼ばれ，同じ機能の遺伝子に対する複数（生物種）のプロモータ領域からその特徴的なコンセンサス配列を探し出す必要がある．コンセンサス配列を表現する手段としては，重み行列や隠れマルコフモデルなどの確率的計算モデルがよく使用され，それを同定する方法としては，ギブスサンプリングなどのヒューリスティックな手法が用いられている[9), 17)]．また，クロマチン免疫沈降とマイクロアレイを組み合わせた最近の実験技術ChIP-chip法により，転写因子とプロモータ領域のDNA配列との結合を直接的かつ網羅的に測定することが可能となり，そのデータも利用可能となっている[10)]．

●木構造データの比較

次に，木構造データの比較解析について見ることにする．木構造は，グラフ理論的にはループを持たない有向グラフのことで，とくに根付き木を考えることにする．木構造によって表される分子生物データの代表的なものとして，進化系統樹やRNAの2次構造，また最近では糖鎖が挙げられる．進化系統樹は，言うまでもなく，共通先祖の生物種（それを根とする）からはじまって，進化の途中において分化して複数の異なる子孫の生物種が創出されていくプロセスを表したものである（**図-3** 参照）．進化系統樹は，一般に枝の長さが進化速度を表すように作られており，生物種間に共通に保存される遺伝子のDNA配列の置換数を数えることによって計算される．進化における置換数を数えるための安定的な遺伝子として，リボゾームRNAがよく用いられる．

一方，機能性のRNA分子はDNAと異なり，2次構造と呼ばれる生物的構造を形成することにより生化学的機能を有する．このRNAの2次構造は，ワトソンクリッ

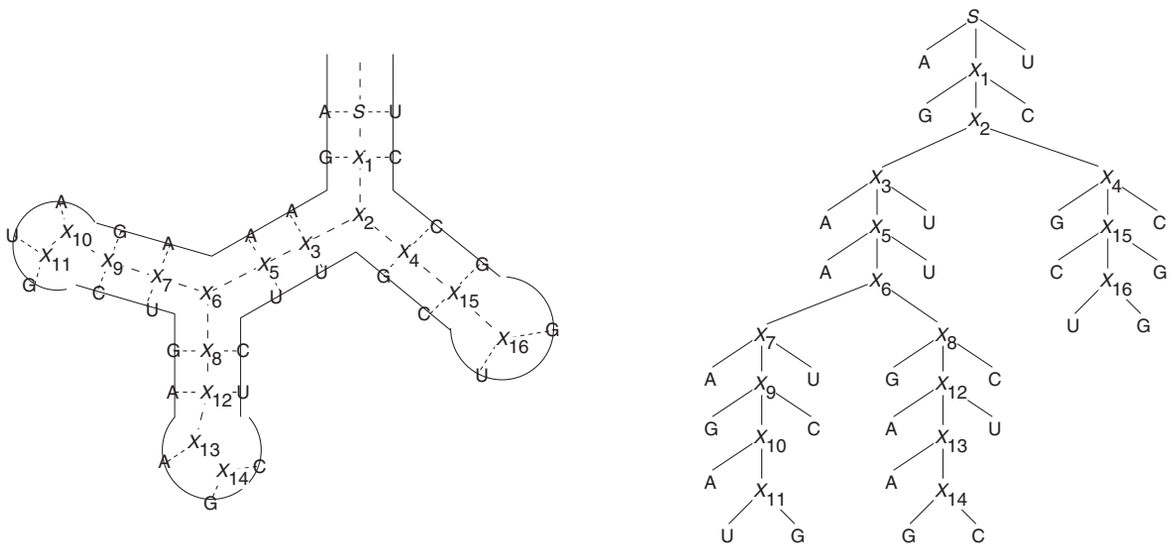

図 -4　RNA 配列 AGAAAGAUGCUGAAGCUCUUGCUGGCCU が形成する 2 次構造（左）とその 2 次構造を表現する木構造（右）

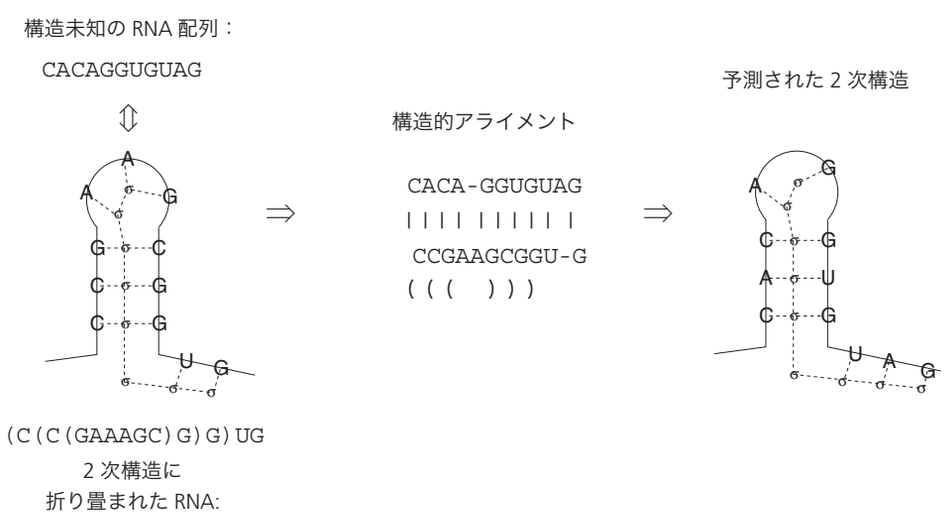

図 -5　RNA 配列の構造的アライメント

ク結合と呼ばれる塩基間の水素結合によって形成され，分岐構造も多く出現することから，木構造データによって表現される（**図 -4** 参照）．

このように木構造で表現されたデータを比較すると，RNA においては，2 次構造の予測や機能性 RNA のゲノムからの探索という分子生物学やゲノムサイエンスの問題を解くことができるようになる[11)～13)]．具体的には，RNA の構造的アライメントと呼ばれる問題は，構造が未知の RNA 配列に対して既知の 2 次構造を当てはめる問題であるが，木構造のアライメントを拡張することで解くことができる（**図 -5** 参照）．そしてこのアルゴリズムを用いてゲノム上の非コード RNA 領域を網羅的に探索する．

細胞内の転写物の 90% 以上が非コード RNA であると推定されており，非コード RNA 領域の探索や機能同定はポストゲノムにおいて重要な課題である．しかし，遺伝子のコード領域と異なり，RNA のコード領域には分子生物的な文法構造や強いシグナルがないため，その同定はより難しい問題である．木構造アライメントを用いた方法は，2 次構造を手がかりに RNA 領域を発見しようとする試みである．

木構造データの比較は，一般に木アライメントという問題に定式化され，線形構造の場合と同様に動的計画法を用いて効率よく解くことが可能である[6)]．

近年，細胞表面上の認識物質として糖鎖が注目されている．細胞間の接着や細胞内に進入しようとするウイルスの標的として，糖鎖は非常に重要な物質であり，それらのメカニズムを解明するためには糖鎖の構造解析

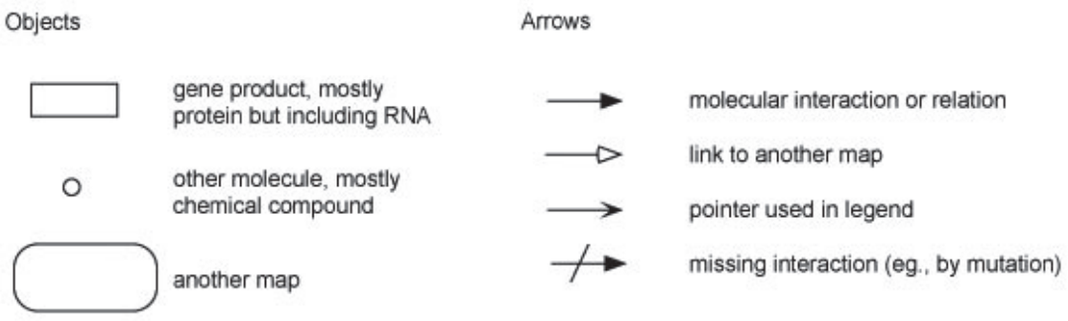

図 -6　KEGG のグラフ表現に関する表記法の一部

が必須である．糖鎖は，結合位置の多様性から複雑な分岐型の構造をとるため，そのデータは木構造によって表現される．糖鎖のデータベースはほとんど存在しないが，KCaM（KEGG GLYCAN）[23] はその中で唯一のデータベースである．KCaM においては，糖鎖構造の比較による検索が可能となっており，その方法は木構造の比較とマッチングが基本となっている[4]．

●グラフ構造データの比較

バイオテクノロジーが進歩して，細胞内のさまざまな物質を高精度に測定したり，発現状態を網羅的に計測することが可能になってくると，細胞内の動的な活動を解析したりモデル化する研究が盛んになってくる．その代表的なものが，代謝やシグナル伝達などのパスウエイの解析である．パスウエイは，分岐やループ，サイクルなどのさまざまな経路形態をとるため，その表現としてグラフ構造が用いられる．このようなパスウエイのデータベースの代表的なものとして，KEGG (Kyoto Encyclopedia of Genes and Genomes) [22] が有名である．KEGG は，遺伝子，ゲノム，リガンド，パスウエイ情報等を含んだ生命情報統合データベースである．特にグラフィカルなパスウエイマップと他のデータベースとの網羅的な関連付けは，KEGG の大きな特徴である．また，データの関係付けをすべてグラフとして考えて，さまざまな関係をグラフとして扱うことも大きな特徴である．たとえば，KEGGのグラフ表現に関する表記法の一部は**図 -6** のように定められていて，また細胞周期における p53 に関連するパスウエイは**図 -7** にあるように表現されている．さらに，検索においても，グラフで表現されたパスウエイを取り出して出力する機能を提供している．このような代謝やシグナル伝達の他にも，マイクロアレイや酵母ツーハイブリッド法などの最近の網羅的解析手法の発展により，遺伝子制御ネットワークやタンパク質相互作用ネットワークのデータも得られるようになってきている．いずれもグラフ構造のデータである．

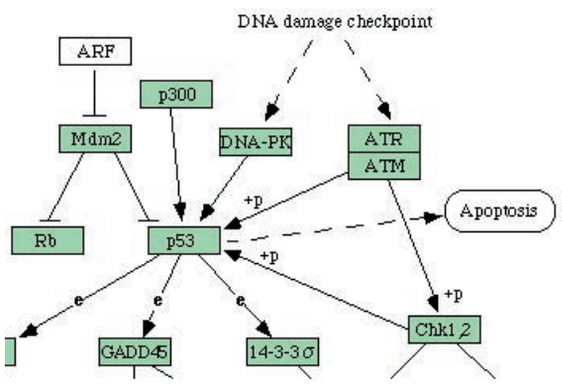

図 -7　細胞周期における p53 に関連するパスウエイ（KEGG データベースから）

これらのパスウエイの解析においては，構造的モチーフと呼ばれるネットワーク上に何度も出現する共通の部分グラフパターンを抽出して，同じような部分構造をネットワーク全体から検出する問題が提案されている．このようなグラフ構造のデータの比較のためには，線形構造や木構造と同様に，やはりアライメントが中心になるが，その計算は単純に動的計画法を適用して解ける問題ではなくなる．たとえば，あるグラフ中に特定の部分グラフが存在するか否かを決定する部分グラフ同型問題や2 つのグラフに共通な最大の部分グラフを求める問題はNP 困難であることが証明されている．したがって，グラフアライメント問題を解くためには，近似的な解法や準最適な解を求めるための何らかのヒューリスティックスが必要となる．

2 つのグラフが与えられた時に，グラフのアライメントは，2 つの部分グラフとその部分グラフ間のノードの対応によって定められる．与えられた2 つのグラフからそれぞれ部分グラフを抜き出して，その2 つの部分グラフの間で1 対1 対応のノード間の対応をとる．2 つのグ

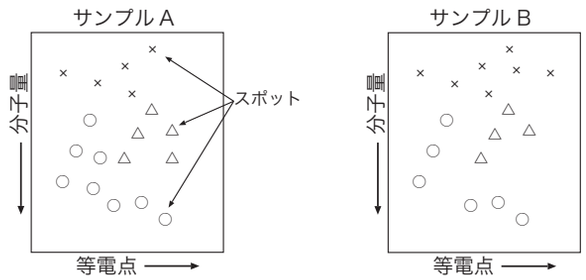

図-8　2次元電気泳動パターンの比較解析

ラフ間にグラフアライメントはいくつも存在するが，それぞれのグラフアライメントのスコアは，ラベル無しグラフの場合にはノード間の辺（リンク）の一致度を数えることにより，またラベル付きグラフの場合にはラベル間の距離も考慮したリンクの一致度とノード間のラベルの距離を加えたもので計算される．2つの部分グラフのノードの数が同じ（n個のノード）であると仮定した場合でも，単純にはノード間の対応の種類は，$n!$通りの数が存在する．しがたって，最大スコアのグラフアライメントを求める問題は，計算量的には難しい問題となる．

一方，グラフの類似度を測る別の方法として，最近発展してきたカーネル関数の使用がある．カーネル関数やサポートベクターマシン（SVM）の解説については別の文献（たとえば，文献21）などを参照）にゆずりたいが，それぞれのデータ構造が持つ特徴をベクトル空間上で表現して，ベクトルの内積として表される2つのデータの類似度をカーネル関数を用いて計算する方法である．2つのグラフの類似度を計算するカーネルにはいくつか提案されている[7),8)]．たとえばDiffusion kernel[8)]では，グラフ上の任意の2つのノード間の類似度を計算する指数カーネルを提案している．グラフのノードに対する隣接行列からある指数関数（指数カーネル）を用いてノード間の類似度を計算する．また，グラフ上のMarginalized kernel[7)]では，2つのグラフ間の類似度を直接計算するカーネルを提案している．基本的なアイディアは，グラフ上に現れるラベル付き経路の出現回数を数えてベクトル化し，その特徴空間上で内積を計算する．ラベル付き経路の種類の数は無限に存在するので，グラフ上のランダムウォークを用いて数える方法を提案している．ただし，このようなカーネル関数を用いて計算されるグラフ間の類似度は，必ずしも元のグラフの（位相）構造を保存するものではない．

2次元（画像）データの比較

細胞内の物質（タンパク質やmRNAなど）を一斉に網羅的に測定する最近の技術として，2次元電気泳動とマイクロアレイが有名である．これらの方法では，いずれもその計測結果としての2次元画像データを処理して解析する必要がある．

タンパク質は大きさと電荷という2つの性質を持っているため，タンパク質の分子量と等電点という2つの性質を利用して，一度にたくさんのタンパク質を同時に分離する方法が2次元電気泳動である．その結果は2次元画像のデータとして得られるので，その中からタンパク質スポットの位置や大きさ，形などを検出して，2次元電気泳動パターンを抽出する．次にデータベース中の他の泳動パターンとの照合や比較解析を行い，スポット中のタンパク質の機能予測や細胞内の状態の計測などが行われる（**図-8**参照）[20)]．

一方，マイクロアレイは，スライドガラス上に数千から数万個の遺伝子またはEST配列のcDNAスポットを作成し，解析するmRNAから調整したターゲットをハイブリダイゼーションさせて，ハイブリッド形成の強度を指標にして，各遺伝子の転写量を測定する方法である．細胞内に発現するすべての遺伝子の動的挙動を効率的かつ定量的に計測することが可能であり，これらの結果を解析することにより，遺伝子ネットワークの推測や遺伝子レベルでの病気の診断などが可能になる．マイクロアレイの計測結果も数万スポット上の蛍光からなる2次元画像として得られる．この2次元画像を処理して，スポットごとの発現量の違いやスポットの発現パターンなどの比較解析が行われる．スポットの発現パターンに基づいて遺伝子を分類することにより機能やネットワークの予測，診断なども行われるため，クラスタリングの手法もよく用いられる[19)]．

図 -9　PDB におけるヘモグロビンのタンパク質 3 次元立体構造情報

●3次元構造データの比較

タンパク質が持つ機能はそのアミノ酸配列が折り畳まって形成される3次元立体構造と密接な関係があるため，ポストゲノムにおけるプロテオミクスの研究においては，3 次元立体構造の比較と予測は重要な課題である．現在までに解明されているタンパク質の3 次元立体構造の数は数千のオーダーであり，たとえばヒトの全遺伝子数3 万弱から比べるとはるかに少ない数である．タンパク質立体構造のデータベースとしては，PDB（Protein Data Bank）[24] が代表的である．タンパク質およびその複合体等の3次元立体構造は，X 線結晶解析や核磁気共鳴（NMR）スペクトルによって構造決定されており，PDB はこれらの立体構造の情報を3次元座標の形で提供している．たとえば，タンパク質を構成する各原子の空間座標は，**図-9** のような形式で記述されている．

タンパク質の3 次元立体構造を比較する問題は，3 次元空間中での2 つの構造の重ね合わせを基本として，重なりが最大限になる位置関係を求めるという問題であり，手法としては立体構造アライメントが知られている．しかし，立体構造アライメントの問題はある定式化のもとでNP 困難であることが知られていて，近似的手法や遺伝的アルゴリズムなどのヒューリスティックスを用いて準最適解を求めている[3), 15)]．

一方，タンパク質の立体構造アライメントを片方が構造未知のアミノ酸配列に応用すると，類似の立体構造に適合させて3 次元立体構造を予測するという問題を解くことができる．タンパク質の立体構造予測は，タンパク質のアミノ酸配列情報だけをもとにして，その立体構造を予測してモデリングする問題である．先ほども述べたように，全遺伝子数に比べて，解明されている3 次元立体構造の数は非常に少ないため，立体構造予測の問題は，バイオインフォマティクスにおいても古くから最も重要な問題の1 つとなっている．既知の立体構造に対してアミノ酸配列を適合させて構造を予測する手法はスレッディングとも呼ばれ，酵素などの機能改変や人工的タンパク質の設計問題には有効な方法である[3)]．

生命情報科学の今後の発展

これからのバイオインフォマティクスを占う上で，用語のレベルは異なるが次の3 つのキーワード，「多種類のデータの統合」，「網羅的な計測データ」，そして「システムバイオロジー」が大きな意味を持ってくると思われる．

本稿で見てきたように，ゲノムなどの配列情報に加えて，分子生物学の技術の進歩により，マイクロアレイの発現プロファイルデータやタンパク質相互作用データ，またタンパク質立体構造データなどのさまざまな種類のデータが比較的簡単に利用可能になってきている．たとえば，モデル生物（かつ真核生物）の1つである出芽酵母（*S. cerevisiae*）では，ゲノム配列，cDNA，発現プロファイル，タンパク質相互作用，ChIP-chip などのほとんどすべての種類のデータがそろっている．このような複数種類のデータを同時に利用して，タンパク質の機能や，シグナル伝達や遺伝子制御などのネットワーク，さらに非コードRNA などのモデリングや予測を精度良く行うことが現実的なレベルになってきている．さらに，網羅的な計測技術により，そのデータの量も指数関数的に増えている．このような技術の進歩を踏まえて，単一の遺伝子を追求するという従来の生物学の方法論から細胞の活動全体をシステム的に捉えて解析したりモデリングして，最終的には計算機上で仮想的にシミュレーションしてしまおうというシステムバイオロジーが盛んになりつつある．システムバイオロジーの詳細については本特集の別記事にゆずりたいが，バイオインフォマティクスとの違いは何かと問われることが増えている．システムバイオロジー自体が広範の研究課題を含んでおり，またその研究アプローチの明確な了解もあるわけではないので，その質問に対する解答はないと思われる．1つ言えることは，情報科学や計算機科学を基礎にして，情報的あるいは計算的なものの見方とセンスで生命現象を解き明かしていくというのが，やはりバイオインフォマティクスのスタンスだと思われる．ただ，システムバイオロジーが分子生物学者と情報科学者の距離を近づけた，あるい

は両者を同じところに巻き込んだという功績は高く評価するべきだと思う．さらに，マイクロアレイなどの網羅的な測定技術を用いた研究においては，ここまでは分子生物でここからは計算機と区切ることはできず，研究の最初の段階からバイオインフォマティクス研究者がかかわっていく必要がある．たとえば，計算機解析をにらんでマイクロアレイのデザインや実験プロトコルを設計することが，マイクロアレイを用いた研究の成功要因でもあるので，分子生物学とバイオインフォマティクスが相互にフィードバックを行うことが必要となる．

もう1つの重要な要素は，未解読の生物のゲノム配列がこれからもますます活発に決定され，それらのデータが利用可能になっていくという状況である．新しいゲノムはやはり新しい発見や知見を与えてくれ，生物の多様性や適合性，そして生命活動メカニズムのすばらしさを教えてくれることである．たとえば，本稿執筆時点で最新のNature（28, Oct., 2004）には，ヒトクリプトスポリジウムのゲノムが解読されたという記事が載っている．920万塩基対のゲノムから，この寄生虫が酸素豊富な汚水中と酸素は少ないが栄養素豊富なヒトの消化管細胞中のどちらでもうまく生きていけるように，異なる種類の代謝遺伝子群が含まれていることが発見された，とある．自然に学ぶことはとても多く，その自然の教科書，ゲノムという教科書を読むためにも，最新のバイオテクノロジーの実験技術とバイオインフォマティクスの情報技術は欠かせないものである．

最後に，バイオインフォマティクスの参考書や教科書は，洋書和書を問わずこの2, 3年の間にものすごい数（これも指数関数的？）が出版されている．筆者が比較的よく参考にしている文献を載せてあるが，興味のある読者は，amazonあたりで「バイオインフォマティクス」や「生命情報」，英語では“Bioinformatics”や“computational biology”というキーワードで検索するとたくさんの本が出てくるので実際に調べていただきたい．

参考文献

1）Durbin, R., Eddy, S., Krogh, A. and Mitchison, G.: Biological Sequence Analysis, Cambridge University Press（1998）．阿久津他訳：バイオインフォマティクス，医学出版（2001）．
2）Pevzner, P. A.: Computational Molecular Biology, MIT Press（2000）．
3）Tsigelny, I. F. (ed.): Protein Structure Prediction, International University Line（2002）．
4）Aoki, K. F., Yamaguchi, A., Okuno, Y., Akutsu, T., Ueda, N., Kanehisa, M. and Mamitsuka, H. : Efficient Tree-matching Methods for Accurate Carbohydrate Database Queries, Genome Informatics , Vol.14, pp.134-143（2003）．
5）Doolittle, R. F., et al: Simian Sarcoma Virus Onc Gene, v-sis, is Derived from the Gene (or genes) Encoding a Platelet-derived Growth Factor, Science, Vol.221, pp.275-277（1983）．
6）Jiang, T., Wang, L. and Zhang, K. : Alignment of Trees - An Alternative to Tree Edit, Theoretical Compututer Science, Vol.143, pp.137-148（1995）．
7）Kashima, H., Tsuda, K. and Inokuchi, A. : Marginalized Kernels between Labeled Graphs, Proceedings of 20th International Conference on Machine Learning (ICML2003), AAAI Press, pp.321-328（2003）．
8）Kondor, R. I. and Lafferty, J. : Diffusion Kernels on Graphs and Other Discrete Input Spaces, Proceedings of 19th International Conference on Machine Learning (ICML2002), AAAI Press, pp.315-322（2002）．
9）Lawrence, C. E., Altschul, S. F., Bogurski, M. S., Liu, J. S., Neuwald, A. F. and Wootton, J. C. : Detecting Subtle Sequence Signals: a Gibbs Sampling Strategy for Multiple Alignment, Science, Vol.262, pp.208-214（1993）．
10）Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I. and et al.: Transcriptional Regulatory Networks in Saccharomyces Cerevisiae, Science, Vol.298, pp.799-804（2002）．
11）Matsui, Y., Sato, K. and Sakakibara, Y. : Pair Stochastic Tree Adjoining Grammars for Aligning and Predicting Pseudoknot RNA Structures, Proceedings of 3rd Computational Systems Bioinformatics Conference (CSB2004), IEEE Computer Society, pp.290-299（2004）．
12）Sakakibara, Y., Brown, M. P., Hughey, R., Mian, I. S., Sjölander, K., Underwood, R. and Haussler, D. : Stochastic Context-free Grammars for tRNA Modeling, Nucleic Acids Research, Vol.22, pp.5112-5120（1994）．
13）Sakakibara, Y.: Pair Hidden Markov Models on Tree Structures, Bioinformatics, Vol.19, pp.i232-i240（2003）．
14）金久 編：ヒューマンゲノム計画，共立出版（1997）．
15）菅原 編：あなたにも役立つバイオインフォマティクス，共立出版（2002）．
16）藤 博幸：タンパク質機能解析のためのバイオインフォマティクス，講談社（2004）．
17）美宅，榊 編：バイオインフォマティクス，東京化学同人（2003）．
18）村上，古谷 編：バイオインフォマティクスの実際，講談社サイエンティフィク（2003）．
19）Kohane, I. S., Butte, A. J. and Kho, A. T. 著，星田有人 訳：統合ゲノミクスのためのマイクロアレイデータアナリシス，シュプリンガーフェアラーク東京（2004）．
20）高橋勝利：プロテオーム解析を支援するインフォマティクス，実験医学増刊「ゲノム医科学と基礎からのバイオインフォマティクス」，Vol.19, No.11, pp.763-770（2001）．
21）津田宏治：カーネル設計の方法，日本神経回路学会誌，Vol.9, No.3, pp.190-195（2002）．
22）KEGG: Kyoto Encyclopedia of Genes and Genomes: http://www.genome.jp/kegg/
23）Glycan Structure Search using KCaM: http://glycan.genome.jp/
24）The RCSB Protein Data Bank: http://www.rcsb.org/pdb/
25）UCSC Genome Browser Home: http://genome.ucsc.edu/

（平成17年1月25日受付）