解　説

大量 DNA データを対象とした遺伝情報のコンピュータ解析†

五條堀　孝††　森山悦子††
内藤公敏†††　河合正人†††

1. はじめに

ヒトの DNA は，直径がわずか 10〜100 ミクロンの細胞の中に，総実長約 1 m の長さで存在する．この長い糸状の分子は，A，T，G，Cという 4 種類の文字（塩基）が約 3×10^9 個も並んでいる長大な文字列のようなものである．このいわば句読点もない長大な文字列上に，生物機能の中心的な役割を担う遺伝子やそれらの発現及び相互作用を制御する領域の情報が刻みこまれている．この 3×10^9 個の塩基からなるヒトの遺伝情報がいかに大量な情報であるかは，情報量を計算して英文に換算してみるとよく分かる．DNA は 4 種類のみの文字から構成されるのに対し，英文はアルファベット 26 文字と語間のスペース 1 字の計 27 文字で構成されている．ここで，ピリオドやカンマなどは一応除く．情報量は普通ビットを単位として表現されるので，1 塩基が占める場所（サイト）は $4=2^2$ で 2 ビットの情報を含むことができる．一方，英文の場合には，1 文字当りの情報量は，$27=2^{4.75}$ なので，4.75 ビットに相当する．したがって，情報量から考えると，英文の 1 文字は DNA の約 2.38 文字分に相当する．このことから，ヒトの DNA が含みうる情報量は，約 13 億字の英文で書かれた文章と同等であることが分かる．また，この DNA の配列データの中には，蛋白質をコードする領域も数多く含まれている．蛋白質は，20 種類のアミノ酸が 100 個から 1,000 個のオーダで並んだものであり，ちょうど 20 種類のアルファベットからなる小文章のようなものである．

† Computer Analyses of Genetic Information for Mass Sequence Data by Takashi GOJOBORI, Etsuko N. MORIYAMA (National Institute of Genetics, Department of Evolutionary Gnetics), Kimitoshi NAITO and Masahito KAWAI (Facom-Hitac Limited).

†† 国立遺伝学研究所進化遺伝研究部門
††† ファコム・ハイタック(株)

このような莫大な塩基配列データやアミノ酸配列データが利用可能になると，これらのデータから生物機能に直結した重要で有意義な情報をどのようにして抽出するかということが，本質的な問題となってくる．実際，DNA の塩基配列を決定するという基本技術は分子生物学的には確立しており，むしろ情報抽出のためのアプローチがこれからの問題として数多く残っているように思われる．本論文では，遺伝情報解析の重要性と分子進化学的方法論の重要性を指摘し，この分野における情報科学的立場からの貢献が生命科学の今後の発展に不可欠であることを述べてみたい．

2. 遺伝情報解析

DNA の塩基配列という遺伝情報から有益な生物情報を抽出する際に，非常に有効な方法は，塩基配列の比較を行うことである．つまり，よく似た塩基配列を集め，相互の類似度が最大になるように，それらを並置することである．これを「アラインメント」という．特に，2 本以上の配列についてアラインメントを行うことを，「マルチプル・アラインメント」という．マルチプル・アラインメントを行うと，比較する配列間のどの部分がよく似ており，どの部分に大きな違いがあるのかが分かってくる．ただし，正確にそれらを知ろうとすると，量的な解析が必要となることは当然である．

先に述べたように，DNA の塩基配列を決定する基本技術は確立しているので，現在のバイオテクノロジを用いれば，遺伝子の機能は未知のまま，その塩基配列を決定することができる．このため，遺伝子機能を知ることなく，遺伝情報としてのデータのみが多く蓄積してきている．このような状況において，機能が未知な塩基配列からその機能を推定することは非常に重要なことである．この機能推定を行うための方法の一つが「相同性探索法」と言われるもので，機能が未知

```
AGGTCTT
:  :  :
AG--CGT
```

図-1 二つの塩基配列のアラインメントの例
":" は相同な部分，"—" はギャップを示す．

の配列データと統計的によく似た配列を，機能が既知の配列データベースから探索する方法である．

2.1 アラインメント操作

たとえば，図-1 にあるような二つの塩基配列があったとしよう．配列1と配列2の両者において，同一の塩基をもつ割合が最大になるように，これらの配列を並列することが可能である．しかし，このアラインメントを行うとき，両者に対応しない場所もあるので，この場合その場所は空として，飛び越して次に配列をつなげていかなければならない．この空の場所のことを一般に「ギャップ (gap)」という．また，塩基やギャップの存在するそれぞれの場所をサイトという．図-1 では，「4つのサイトで塩基が同一であり，二つのサイトにギャップが存在する」．この例からも分かるように，どこにギャップを入れるかや，塩基が異なるときにどのサイトにどの塩基を置くかは，ある仮定を置かないかぎり一意には決まらない．したがって，たとえばギャップと塩基差異にある重みづけを行って，そのスコアの最も少ないアラインメントを採用することになる．ここでは塩基配列のアラインメントについて説明したが，アミノ酸配列の場合にも，まったく同様にアラインメントを行うことができる．

アラインメントには，スコアのつけ方も含めて，いくつかのアルゴリズムが存在するが，一般に最もよいアラインメントは，進化的過程の反映として二つの配列データが並べられることである．つまり，もし二つの配列が共通の祖先配列から由来してきたと考えると，もともとは同一の配列であったはずである．時間が経つにつれて，二つの進化的経路にわかれ，それぞれの経路で塩基やアミノ酸が置き換わったり（これを「塩基置換」あるいは「アミノ酸置換」という），また塩基やアミノ酸が欠けたり付け加わったり（「欠失」や「付加」という）する突然変異を蓄積しながら現在に至ったと考えられる。アラインメント操作によって得られた二つの配列の並び方が，塩基（あるいはアミノ酸）が異なっているサイトでは進化の過程で塩基置換（あるいはアミノ酸置換）が起こったように，ギャップのサイトではそこに欠失や付加が起こったように，これらの配列の進化的な過去が反映された形になっているのが良いのである．もちろん，突然変異の発生やその蓄積は確率的な事象であるので，得られた配列のアラインメントが，それらの進化的な歴史を反映していることを保証するアルゴリズムを作成することは，ときとして難しいこともある．

2.2 相同性探索法

類似性のある二つの塩基（アミノ酸）配列は，「アラインメント操作」によってサイトごとに並列することができる．このアラインメント操作によって並べられた二つの配列の類似性が，統計的に有意なものであるかどうかを判断する必要がある．つまり，塩基（アミノ酸）組成の等しいランダム配列どうしをアラインメントしたときに得られる類似性よりも，いま注目している二つの配列の類似性が有意に高いかどうかを調べなければ，たとえアラインメントが得られたとしてもあまり意味がないのである．この原理を利用して，ある塩基やアミノ酸配列の生物学的機能を予測する非常に有効な方法が「相同性探索法」といわれるものである．

「相同性探索法」を説明する前に，「DNA（蛋白質）データベース」について簡単に説明しよう．現在までに決定された塩基（アミノ酸）配列の一つ一つについて，その遺伝子名，単離された生物種名，その機能や構造に関するコメントなどを一つのエントリとしてデータベース化したものが，「DNA（蛋白質）データベース」である．現在，DNA データベースとしては，GenBank（アメリカ合衆国），EMBL（ヨーロッパ），DDBJ（日本）が，そして蛋白質データベースとしては NBRF/PIR（アメリカ合衆国）や蛋白奨励会（日本）が，国際的に主要なデータベースとしてよく知られている．

これらの DNA（蛋白質）データベースを検索すれば，今までにどのような塩基（アミノ酸）配列が決定されているかなどを即座に知ることが可能である．この検索を機能の分かっていない塩基（アミノ酸）配列を照会配列（プローブ）として行い，その照会配列と類似性の高い塩基（アミノ酸）配列を DNA（蛋白質）データベースの中から探すのが「相同性探索法」である．このとき，類似性は配列間にギャップなどを考慮してアラインメントを行い，統計的に有意なエントリだけをとりだしてくる．つまり，一つの照会配列に対して DNA（蛋白質）データベースの全エントリの一つずつにアラインメント操作をするようなものである．もし，統計的に有意に類似性の高い塩基（アミノ酸）配列が取り出され，そしてその配列の機能が分

かっていたとすれば，機能が分からなかった照会配列は，その類似性の高さから，同様な機能を有することが合理的に推測されるからである．この論理は，分子進化学の中心的な理論である「分子進化の中立説」[5]における一つのパラダイムに裏付けされている．つまり，一つの共通祖先の塩基配列からいくつかの塩基配列が分岐してきたとすると，機能的に重要な領域では，そこに突然変異が生じたものは生体に有害な効果をもたらすことが多いので生物集団中に残りにくく，子孫塩基配列間にはかなりの保存性が存在する．一方，機能的にあまり重要でない領域では，突然変異が生じても生体の生存にあまり影響を与えないので，突然変異を蓄積しやすく，塩基配列間にはあまり保存性がない．したがって，照会配列と高い類似性を示す部分が既知の配列にあるということは，分子進化学的にみても両者が同様な機能を有すると考えてよいことになる．

この方法を最初に用いたのは，カリフォルニア大学サンディエゴ校の R. F. Doolittle 教授のグループであった．がん遺伝子の *sis* が血小板由来の細胞増殖因子（PGDF）と塩基配列上の類似性が高いことを相同性探索法で発見し，がん遺伝子が細胞増殖因子としての機能をもつことを示唆したのは，1983年当時としても非常に衝撃的なことであった[1]．この発見を契機として，がん遺伝子の研究は飛躍的に進んでいった．

2.3 マルチプル・アラインメント

相互に類似性のある二つの塩基（アミノ酸）配列を，その類似度が最大になるようにアラインメントすることができることはすでに述べた．このアラインメントが3本以上の塩基（アミノ酸）配列において必要となることがよくある．特に，「相同性探索法」によって2本以上の塩基（アミノ酸）配列が照会配列と類似性があるとして取り出された場合などは，まさに，マルチプル・アラインメントが必要となる．マルチプル・アラインメントのアルゴリズムは現在いくつか存在するものの，その問題の難しさのためにまだあまり確立しておらず，発見的なアルゴリズムを用いて試行錯誤で行うような状態が続いている．その中でもよく用いられるアルゴリズムは，2本の配列データのアラインメントを基礎におくものである．たとえば，まずある2本の配列のアラインメントを行うとしよう．このとき，塩基（アミノ酸）が異なるサイトでは，その両方の異なる塩基（アミノ酸）が等確率で存在しうるとして，この両配列を平均したような配列を作成する．この平均配列に第3の配列を加え，再び同様の平均操作を行う．この平均操作という意味は，塩基（アミノ酸）差異にスコアを与えるときに塩基（アミノ酸）の存在しうる確率に従って平均をとるという意味である．たとえば，二つの配列のあるサイトに塩基AとTがあったとすると，この二つの配列の平均配列は次のようになる．

配列1　　A

配列2　　T

平均配列(1+2)　　$\frac{1}{2}\mathrm{A}+\frac{1}{2}\mathrm{T}$

この平均配列(1+2)に配列3を加えるには，配列3の対応するサイトがAであれば，配列1と2の平均配列(1+2)と配列3の対応するサイトで塩基が一致するスコアを次のように計算する．

平均配列(1+2)　　$\frac{1}{2}\mathrm{A}+\frac{1}{2}\mathrm{T}$

配列3　　A

一致するスコア　　$\left(\frac{1}{2}\times 1\right)+\left(\frac{1}{2}\times 0\right)=\frac{1}{2}$

このようにして以下第4，第5の配列を順次加えていくことにより，最終的にマルチプル・アラインメントを完成させることができる〔たとえば，文献3)．このほか *Methods in Enzymology* の分子進化学的手法の特集号には，いくつかの方法が紹介されているので参照されたい〕．

2.4 分子進化系統樹作成

塩基配列やアミノ酸配列のアラインメントから，分子進化系統樹を構築することができる（系統樹作成法の詳細に関しては，本特集の“分子進化とシミュレーション”を参照のこと）．分子進化系統樹は，比較した遺伝子や生物種に関する重要な知見をいろいろ与えてくれることが多い．たとえば，ある遺伝子がその生体の機能にとってどのくらい重要な働きをもつかは，その遺伝子が進化的にどのくらい古い生物種に存在するかを知ることによって，大まかに理解することができる．がんに関係した *ras* という遺伝子は，哺乳類などの高等動物にはもちろん，最近ではイースト菌などにも存在することが分かってきており，細胞内の情報伝達系に重要な働きをしていることが報告されている．実際に，いくつかの *ras* 遺伝子の塩基配列から分子進化系統樹を構築してみると，この遺伝子の起源は，細胞内に核を有する「真核生物」と核を有しない「原核生物」とが進化的に分岐した非常に古い時期に

相当することが分かる．このことから，*ras* 遺伝子の機能は，生体機能にとって本質的に重要であることが強く示唆される．

また，これらの分子進化系統樹の作成は，今後大量に発見されてくることが確実な新しい遺伝子や蛋白質を系統的に分類するために必須の方法論となるであろう．特に，その遺伝子のゲノム上の位置や遺伝子構造と分子進化系統樹を対比させながら考えると，その遺伝子構造の進化的変遷が明確に理解されるであろう．現在，分子生物学においては，「構造」を知るためだけの研究目的から「構造」を通して「機能」を解明するという研究目的へと，意識の変革が行われようとしている．分子進化系統樹の作成は，その目的意識に応えることのできる強力な方法論を提供するものと思われる．

3. 分子進化学的解析の応用

ここまで，塩基（アミノ酸）配列データを用いた分子進化学的手法を中心とする遺伝情報解析の方法のいくつかについて紹介してきた．今後，ヒトや各種生物ゲノムプロジェクトが進むに従い，膨大な量の配列データが蓄積されていくことが予想される．このような状況下では，大量の DNA データをどのように処理するか，またどのような方法で分子生物学的に有用な情報を抽出していくか，などという問題が生じてくるであろう．筆者らは，これまでに確立されてきた分子進化学的手法をさらに発展的に用いることにより，新しい視点からこれらの問題に対処することを試みているので，そのいくつかの例をここで紹介しよう．

3.1 進化方向の予測

多数本の塩基（アミノ酸）配列データを互いの類似性が最大になるようにマルチプル・アラインメントを行い，この比較をもとに配列間の塩基（あるいはアミノ酸）置換数を推定すれば，この推定値から分子系統樹を構築することができる．その方法の詳細については，本特集の“分子進化とシミュレーション”（舘野義男），及び文献 4) などに述べられている．これらの分子進化学的手法を応用することにより，遺伝情報がどういう方向に進化しているのか，すなわち「塩基（アミノ酸）置換の方向／パターン」を知ることが可能となる．このことは，遺伝情報から未来進化の予測が原理的に可能であることを意味する．

単に配列データを比較するだけでも，各塩基（アミノ酸）間にどのような頻度で置換が生じたかを知ることはできる．しかし，どの塩基（アミノ酸）からどの塩基（アミノ酸）へ置換したかという「置換の方向」までも知ることは一般に不可能である．そこで筆者らは，配列データを比較する際に，分子系統樹からの情報を取り入れることで，「方向」を知ることを可能にした．まず，多数本の配列データのアラインメントをもとに置換数を計算し，分子進化系統樹を構築する．次に，この分子進化系統樹を念頭におきながら，もう一度アラインメントを各サイトごとに見直していく．

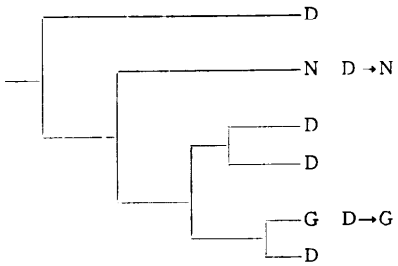

図-2 アミノ酸置換パターンの推定方法
6本のアミノ酸配列のアラインメントの中のある1サイトのアミノ酸の比較を示している．図中の系統樹は，このアラインメント全体から計算された各配列間のアミノ酸置換数をもとに構築されたもの．

図-2 は，あるアミノ酸配列のアラインメントにおいて，一つのサイトに存在するアミノ酸を，配列全体から構築された分子進化系統樹に合わせて示したものである．系統樹の分岐の方向を考え合わせることにより，このサイトではD→NとD→Gの方向のアミノ酸置換がそれぞれ1回ずつ起こったことが推測される（Dはアスパラギン酸，Nはアスパラギン，Gはグリシンというアミノ酸を示す1文字表記である）．同様の操作をアラインメント上に存在する全サイトに関して繰返し行っていくことにより，与えられた配列における置換の方向性やパターンを数量的に知ることができる．

上記のようにして得られたアミノ酸置換パターンは，20 種類のアミノ酸から 20 種類のアミノ酸へ（20 行×20 列）の経験的な遷移確率行列として表される．この遷移確率行列は，アラインメントを行ったアミノ酸配列の過去の進化過程を示していると理解される．同時に，分子進化学的性質が今後もほとんど変化しないとすれば，未来に起こるであろう進化の様態を示すものでもある．すなわち，今後このアミノ酸配列がどのように進化していくかを予測することができるわけである．

筆者らは現在，病原性ウイルス，特に AIDS ウイ

ルスのワクチン開発の新しい方法論として，この「進化方向の予測」を取り入れた研究を行っている．AIDS ウイルスをはじめとする RNA ウイルスの遺伝子では，真核生物遺伝子の約 100 万倍という超高速の塩基置換速度をもつことが知られている．特に，ウイルスが感染した宿主との抗原抗体反応に関与する外被糖タンパク遺伝子 (*env*) の塩基置換速度は高い．このため AIDS ウイルスの免疫学的多様性は非常に大きく，同じ宿主の体内においても免疫学的反応性の異なる変異体が多数存在する．この大きな変異性のため，AIDS ウイルスは宿主の抗体による免疫的攻撃から免れており，これが AIDS ウイルスに対するワクチン開発の大きな障害ともなっている．特に，あるウイルス株に対して有効なワクチンでも，他のウイルス株に対してはほとんど効力をもたず，また，有効であったワクチンも，時間とともにウイルスが変異していくために効力を失うこともある．そこで筆者らは，AIDS ウイルス *env* タンパクのアミノ酸配列のアラインメントから推定されたアミノ酸置換パターンを利用して，アミノ酸置換の方向性を予測したワクチン開発の研究を進めている[2]．この予測の精度をあげるために必要な条件の一つは，アラインメントのできるアミノ酸配列が大量に存在することである．AIDS ウイルスでは，世界的に関心が高いためアミノ酸や塩基配列を実験的に決定する研究が非常に盛んであり，*env* 遺伝子領域だけでも現在すでに 20 本以上の配列データが利用可能となっている．また，インフルエンザウイルスなど他の病原性ウイルスでも，多量の配列データが蓄積されつつある．したがって，大量な遺伝情報の分子進化学的解析は，社会医学的に重要な感染症ウイルスに対する抜本的な臨床対策に重要な貢献を行うことが期待される．今後は，計算機のさまざまな発展とも相まって，AIDS ウイルスだけでなく多くの病原性ウイルスにおいて，各ウイルス特有のアミノ酸置換パターンを推定することが可能になるであろう．このような研究が成功すれば，病原性ウイルスに対する効率のよいワクチン開発への強力な方法論を提供することになると期待される．

3.2 機能予測 (モチーフ探索)

「相同性探索法」が，ある塩基 (アミノ酸) 配列の生物学的機能の予測を行う際にきわめて有効な手段であることはすでに述べた．しかし，最近の分子生物学や分子進化学の研究の進展により，一つの遺伝子はいくつかの部分配列 (これを「ドメイン」という) の集合としてできており，各ドメインの機能の統合として，遺伝子や蛋白質の機能は理解されるべきであるということが明らかになってきた．したがって，このような機能予測としての「相同性探索法」においては，与えられた遺伝子の配列全体に対する相同性よりも，各ドメインに対して「相同性探索」を行うほうが有効であるように思われる．このような「ドメイン探索」に関するアルゴリズムの開発も，欧米では現在盛んに行われている．

これまでに，実際にいくつかの機能ドメインにおいて多くの異なった遺伝子間で共通に保存されている配列，すなわち「共通配列」(保存配列) が見つかってきている．これらは，同じ機能を有する配列のアラインメントにおける共通部分として見つかってきた．図-3 に示したように，「共通配列」と呼ばれるものの中には，ウイルスの遺伝情報の複製酵素である逆転写酵素にみられる "YXDD" (Yはチロシンというアミノ酸，Xはアミノ酸のどれでもよいことを示す) のように連続したものや，Zn フィンガ構造という特殊な蛋白

```
HIV      PDIVIYQ YMDD LYVGS
HTLV−I   PQCTILQ YMDD ILLAS
MMLV     PDLILLQ YVDD LLLAA
RSV      PSLCMLH YMDD LLLAA
<モチーフ>            YXDD
```

(a)

```
1  YI C SFAD C GAAYNKNWKLQA H LC−K H
2  FP C KEEG C EKGFTSLHHLTR H FL−T H
3  FT C DSDF C DLRFTTKANMKK H FNRF H
4  YV C HFEN C GKAFKKHNQLKV H QF−S H
<モチーフ>  C (X4) C (X12) H (X3) H
```

(b)

図-3 アミノ酸配列における「モチーフ」の例
レトロウイルスの逆転写酵素(a) (HIV: AIDS ウイルス，HTLV-I: ヒトT細胞白血病ウイルス，MMLV: ネズミ白血病ウイルス，RSV: トリ肉腫ウイルス)，及びアフリカツメガエルの TFIIIA という蛋白質にみられる Zn フィンガの繰返し構造(b)のアミノ酸配列のアラインメントと，それから得られる各モチーフを示す．各アルファベットは，アミノ酸を1文字表記で表したものである．また，"—" はギャップ，"X" はどのアミノ酸でもよいことを表す．

質構造にみられる "C-X_{2-4}-C-X_{12}-H-X_{3-4}-H"（Cはシステイン，Hはヒスチジンというアミノ酸，X_n は n 個の何でもよいアミノ酸）のように特定のアミノ酸だけが保存されており，その他の部分は長さも含めて変化を許すものもある．筆者らは，このようにある程度の不確実性を含みながらも保存されている配列のパターンを「モチーフ」と呼んでいる．加速度的に蓄積されていくであろう配列データの中からこのように機能的に重要な「モチーフ」を抽出していくことは，遺伝子や蛋白質の機能を予測する上で，今後非常に重要な仕事となっていくであろう．筆者らは，不確実性を含んだ「モチーフ」を配列データベースから抽出するためのアルゴリズムの開発を現在進めている．

すでに機能の分かっている配列データを含む配列データベースから多くの「モチーフ」を抽出することによって，知識データベースとしての「モチーフデータベース」を構築することができる．「モチーフデータベース」は，配列データの未知の機能を予測するために必須のものとなると考えられる．すなわち，機能が未知の配列データに対して，「モチーフデータベース」を用いた「モチーフ探索」を行い，その配列データにどのような「モチーフ」が含まれるかを知ることができれば，そのモチーフに共通にみられる機能から与えられた配列データの機能を推測することができるのである．

今後，配列データの量産体制が確立すれば，まったく機能の分からないDNA断片としての配列データが日ごとに蓄積してくることが予想される．したがって，配列データベースとそれにともなう「モチーフデータベース」の更新は日単位で行われなければならない．さらに，こうして作られた「モチーフデータベース」を用いた機能予測の作業(モチーフ探索も)，かなり頻繁に行っていくことになるであろう．これらの作業をルーチン化して行っていくためには，効率のよい「モチーフ探索」のためのアルゴリズムの開発，及び，巨大化するデータベースを処理するために，スーパ・コンピュータやパラレルコンピュータの利用も含めた解析処理の高速化の問題を解決する必要がある．

4. 遺伝情報解析におけるワークステーションとスーパ・コンピュータの必要性

「相同性探索」や「モチーフ」の抽出，遺伝子機能の推定（「モチーフ探索」）などを効率的かつ精度よく行うためには，優れたアルゴリズムやソフトウェア，及び超高速の情報処理が可能な先端的計算機が必要となろう．特に，「ヒト・ゲノムプロジェクト」のように膨大な塩基配列データが蓄積していく状況においては，ベクトル計算機やパラレル計算機など，超高速の大型電子計算機は必須である．

ここで，生命科学の実験研究者の立場から考えてみると，高度に専門的なソフトウェアを使うことはもちろん，非常に高価な超高速計算機を個人的に使用することはほとんど不可能に近い．しかし，実験研究者が生産したデータを自分の思うように自から解析したいという欲求は切実なものであろうと考えられる．つまり，個人使用が不可能なほどの大量 DNA（蛋白質）データを用いた超高速処理を個人的に行いたいという実験研究者の欲求はきわめて自然であるものの，その実現可能性において矛盾した一面ももつのである．この一見どうしようもない矛盾を解消する方策は，今まさに計算機科学の分野で発展が著しい情報ネットワークの構築である．つまり，実験研究者の各研究室にワークステーションを配し，それらを計算機センタなどの超高速計算機とネットワークで結ぶことである．計算機センタなどの超高速計算機では，DNAや蛋白質の配列データなどの大量遺伝情報の集中管理を行い，また高度なアルゴリズム開発や高速処理を要する計算を行う一方，各実験室のワークステーションでは，これらの大量な遺伝情報の検索やさまざまな解析手法を総合的にメニュ化したようなユーザが利用しやすいソフトウェアの動作する環境を提供する必要があろう．

つまり，各実験研究者は各自の実験室で，ワークステーションを通じて，あたかも大量な遺伝情報のデータベースを個人的に使用し，ある程度高度な解析も超高速で行える場が提供されることになる．

5. 総合的遺伝情報解析システム

配列データベースの巨大化にともなう，ワークステーションとスーパ・コンピュータという二極分化された遺伝情報解析環境を想定して，筆者らはワークステーション上で稼働する操作性の良い遺伝情報解析システムの開発を進めている．本稿の最後に，このシステムの特徴を簡単に紹介しよう．

総合的遺伝情報解析システム（筆者らは「IDEN」と名付けている）は，今まで説明してきたような高度な遺伝情報解析手法を，実験研究者自身がほとんど習

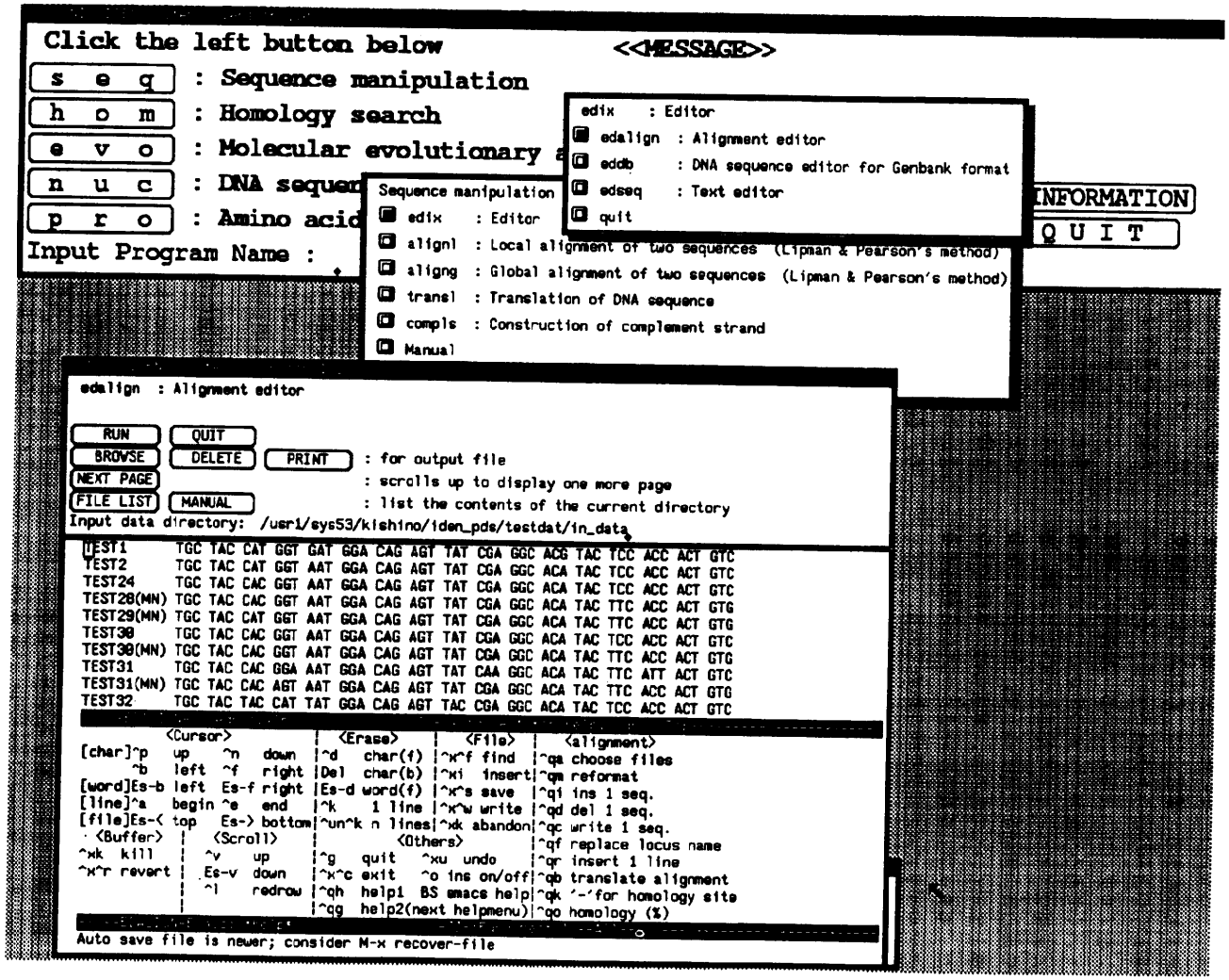

図-4 総合的遺伝情報解析システム「IDEN」の配列アラインメント用専用エディタの画面例

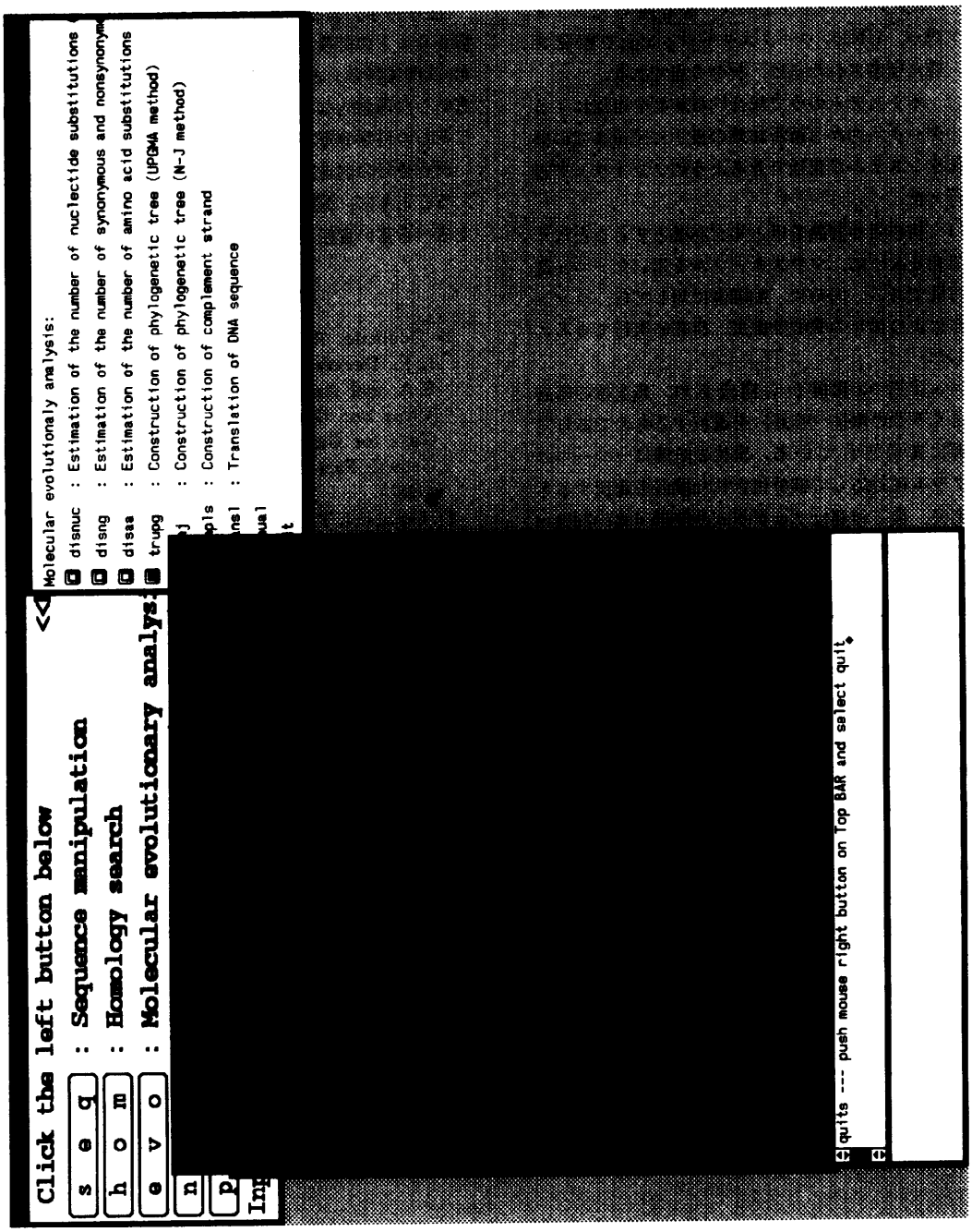

図-5 総合的遺伝情報解析システム「IDEN」で分子進化系統樹を作成した結果の画面例

熟を必要とすることなく手軽に利用できることを目的としている．特に，ユーザが取り扱いやすい MMI（マン・マシン・インタフェース）を考慮したソフトウェア群が，UNIX-ワークステーション上で開発された．特に留意された点は，次の2点である．

（1） 本システムのみで総合的な解析を可能にするため，データ入力から解析結果の表示に至るまでの機能全体をシステムが包括できるようにソフトウェア設計を行った．

（2） 初心者が習熟をほとんど必要とすることなく利用できるように，マウスカーソルを用いたメニュ選択が可能である．同時に，熟練者に対しては，コマンド操作により個々の機能を直接，自在に実行できるようにした．

メニュは三つの階層から構成され，最上位の階層は大きく5つの機能（後述）を選択するものであり常に画面に表示されている．次位の階層は一つ一つのプログラムに対応して細分化された機能を選択するものであり，最下位層はプログラムの起動・実行制御パラメータの選択・データファイルの選択を行うものである．

「IDEN」の機能は，遺伝子配列データの入力機能である「配列操作」，「相同性探索」，「分子進化学的解析」，「塩基配列解析」，「アミノ酸配列解析」，という5つの解析メニュから構成されている．特に，配列アラインメント用の専用エディタ（図-4），分子進化学的解析機能（塩基置換数の推定，分子進化系統樹の作成など；図-5）は本システム特有のものであり，ここまで述べてきたような分子進化学的な遺伝情報解析を，かなり高度なものまで研究室内で手軽に行うことができる．今後は，実験研究者自身の研究室内で，このように分子進化学的な遺伝情報解析手法を含んだ解析ソフトウェアを用いて，実験とその解析を同時に進行させることのできる研究環境が必要とされるようになるであろう．

6. ま と め

「生命科学」あるいは「生物科学」という自然の理解からヒトの健康にまで関係した人類の本質的な科学と，「情報科学」という現代経済国家の将来的発展に直結した基盤的な科学との融合によって重要で先端的な新しい科学分野が誕生しようとしている．この新しい研究分野の創造は，両分野の研究者に大きな夢を与えるとともに，国際的な状況を含む社会的な期待に応えるべき重い責任をも発生させるであろう．

参 考 文 献

1) Doolittle, R. F., Hunkapiller, M. W., Hood, L. E., Devare, S. G., Robbins, K. C., Aaronson, S. A. and Antoniades, H. N.: 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).

2) Gojobori, T. and Moriyama, E. N.: Molecular Phylogeny of AIDS Viruses and Its Application to Vaccine Development, In: Population Biology of Genes and Molecules, (eds., Takahata, N. and Crow, J. F.). Baifukan, Tokyo (1990).

3) Hein, J.: Unified Approach to Alignment and Phylogenies, In: Methods in Enzymology, Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences (ed., Doolittle, R. F.). Vol. 183, Academic Press, Inc., New York (1990).

4) 根井正利：分子進化遺伝学（五條堀孝，斎藤成也共訳），培風館（1990）．

5) 木村資生：分子進化の中立説，紀伊国屋書店（1986）．

（平成2年5月21日受付）