2 ヒトゲノム解読と ヒト遺伝子地図の精緻化

森下 真一　東京大学大学院新領域創成科学研究科／東京大学大学院情報理工学系研究科
moris@gi.k.u-tokyo.ac.jp

小笠原 準　東京大学大学院情報理工学系研究科
jun@gi.k.u-tokyo.ac.jp

本蔵 俊彦　東京大学大学院新領域創成科学研究科
honkura@gi.k.u-tokyo.ac.jp

山田 智之　東京大学大学院新領域創成科学研究科
yamada@gi.k.u-tokyo.ac.jp

ヒトゲノムが解読され，過去収集された遺伝子がヒトゲノム上でどのような構造を持つかが明らかにされると，遺伝子の機能解明に重要な手がかりが得られる．この構造解明のためには，約400万個の遺伝子およびその断片を高速かつ高精度でゲノムに写像する計算技術が有効である．本稿では，写像の分子生物学的な意義と，高速な写像ソフトウェアを実装する技術を解説する．

ヒトという生体の設計図ヒトゲノム，その解読は純粋な好奇心を満足するには十分すぎる目標であったといえる．しかしヒトゲノム解読以前はすべて読むのではなく，遺伝子をコードしている部分だけを読む，というアプローチが主流でありヒトゲノム解読に比べて少ない予算で実行できた．膨大な研究費を投じて解読されたヒトゲノム，解読することによってどのような情報が新たに得られるかを説明することが本解説の目的である．しかしその前に関連する分子生物学的知識[1), 2)]を手短に導入する．

DNAからタンパク質への情報伝達

まず，ヒトに代表される真核生物において，DNAからタンパク質へと情報が伝達する仕組みを図-1に示す．真核生物の細胞内には核があり，その中にDNAという物質が存在し，タンパク質をコードしている．図-1において黒い箱で示された部分がタンパク質をコードしているエキソンと呼ばれる領域である．エキソンは1つだけでタンパク質をコードしていることは稀で，通常は複数のエキソンが連結してタンパク質をコードしてい

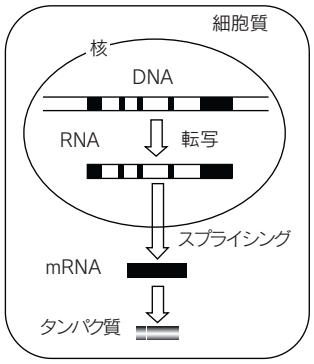

図-1　真核生物におけるDNAからタンパク質への情報伝達

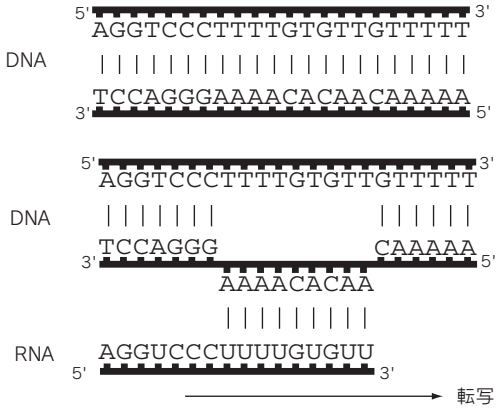

図-2　DNAからRNAの転写

る．図-1では5つのエキソンをコードしている場合を例示している．またエキソンの間を埋める領域はイントロンと呼ばれる領域で，タンパク質をコードしていない．図ではエキソンとイントロンの長さに顕著な差がないが，これは図を簡単化したためである．実際にはイントロンはエキソンに比べ非常に長い場合が多い．DNAはタンパク質をコードするエキソンを網羅して含む物質である．そのため遺伝子（gene）を総称するゲノム（genome）という言葉でも呼ばれる．

転写

エキソンの集まりは，「転写」と呼ばれる仕組みにより，RNAに読み取られる．転写の仕組みをもう少し詳しく説明したのが図-2である．図-2の上に示すように，DNAは，A（アデニン），T（チミン），G（グアニン），C（シトシン）の4種類の塩基が連結した2つの塩基列よりなる．各列には糖とリン酸が結合したバックボーンがあり，図では黒い太線で表現されている．このバックボーン部分に結合しているのが4種類の塩基である．バックボーンには方向性があり，糖の5つの炭素につけられた番号をもちいて，5' 炭素から3'炭素へ至る方向をもって方向性が明示される．

一方，塩基列間では必ずAとTの組およびGとCの組がそれぞれ水素結合により相対する．AとTは互いに相補的な塩基と呼ばれ，GとCも同様である．相補的塩基が連続して結合することにより，2億個を超える塩基の列が安定して形成される．転写時には図-2の下のように一部がほどけ，5' から3' の方向に沿って順番に塩基情報がRNAに写し取られる．この際，若干注意が必要なのは，RNAではT（チミン）の代用としてU（ウラシル）という物質が使われ，Aに対してはUが相補的塩基として情報が伝わる点である．図では2つの塩基列のうち，上側のDNA列の情報をRNAに伝達する際に，下側の列を写真のネガのように使っている．逆の場合もあり，下側の列の情報を，上側をネガとして使いRNAへと情報を伝える．

スプライシング

転写されたRNAは核から細胞質へと出る際に，スプライシングという仕組みによりイントロン部分が除去される．その結果生成されるのがmRNAである．このプロセスで厄介なのは，産物であるmRNAだけを見ても，エキソンの境界がまったく分からないという点である．この情報損失を復元するには，どうすればよいか？全DNA解読以前は手間のかかる作業であった．まずmRNAの配列中で，この配列を認識するのに十分なほどユニークな配列を見つけ出し蛍光し，染色体上に貼り付け，顕微鏡で観測して大体の位置を近似する．これだけではエキソンの構造も分からないので，つづいて染色体での位置が近似された領域のDNA配列をたとえば100万塩基ぐらい読む．最後にストリングマッチによりmRNAをエキソンへと分解するという工程である．

しかしDNA配列の約90％以上が解読された現在は，計算機を使ってmRNAの配列を直接DNA配列へと写像

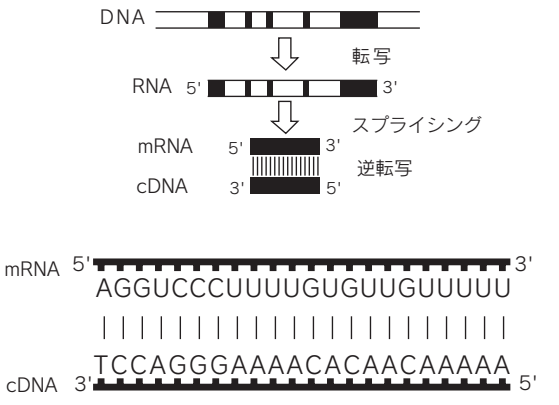

図-3　cDNAの生成

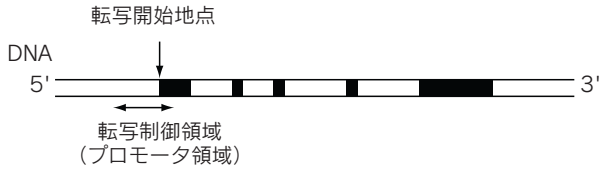

図-4　転写調節領域（プロモータ領域）

して，エキソン情報を復元することが原理的には可能になった．しかし約30億の塩基を持つヒトゲノムへの写像は工夫しないと膨大な計算コストがかかる．この問題を如何に高速に解くかについては，後半で述べる．

タンパク質への翻訳

図-1に示すようにmRNAはさらにタンパク質へと翻訳される．ほとんどのmRNAはタンパク質に翻訳されることにより生体内で機能を発揮するものの，タンパク質の機能は未知である場合が多い．そこで機能を知るために，生体内でのさまざまな種類の細胞や組織でのタンパク質の振る舞いを観測することが重要になってくる．しかしタンパク質を直接観測するのに比べ，タンパク質に対応するmRNAを観測する方が容易であり，とくに近年はマイクロアレーやDNAチップなどmRNAを高速かつ大量に観測する装置が普及している．ただし，mRNAはすべてタンパク質に翻訳されるわけでなく，mRNAと対応するタンパク質の間に量的な線形相関が必ずしもない点には留意する必要があり，しばしば議論にもなる．

cDNA収集

mRNAを観測する際には，mRNAの各塩基と相補的な塩基のDNA配列を生成し，この相補的DNA塩基鎖が元のmRNAと結合する性質をしばしば利用する．この相補的DNA塩基鎖は，DNA配列に存在するイントロンを含まず，エキソン部分を連結した情報だけを含むため，DNAと区別するためにcDNA（complementary DNA）と呼ばれる（図-3参照）．cDNAは実験により逆転写酵素を使いmRNAから生成することが可能である．DNAの全体の約3％程度しかmRNAをコードしていないと見積もられている．そのため全DNA配列の解読よりcDNA収集のほうが効率的であるという主張もある．cDNAは，ヒトゲノム解読と並行して精力的に収集されてきている．

mRNAを観測する目的からすると，mRNAを全長に渡ってcDNAに転写する必要さえない．mRNAを識別するのに十分な長さ（たとえば100程度）の部分列だけを収拾すれば十分ともいえる．このように短いcDNA配列をEST（Expressed Sequence Tag）と呼ぶ．一方mRNAの全長に渡って逆転写した配列を完全長cDNAと呼び区別する．EST収集は，塩基配列読取装置の動作コストを低減する合理的価値がある．dbESTというESTデータベースには2001年9月末現在約380万個のESTが完全長cDNAも含めて登録されている．このようにcDNAやESTを観測・蓄積することは重要であるが，以下ではDNAに存在しcDNAにはない有用な情報について説明する．

ヒトゲノム解読の意義 ー転写制御領域解析

図-4に示すように，遺伝子をコードしているエキソンの転写が開始される場所を転写開始地点と呼ぶ．転写開始地点の近傍は，転写制御領域（プロモータ領域）と呼ばれ，転写を制御する因子が結合する比較的短い塩基配列（たとえばTATAAAやTGACTC）が存在する重要な領域である．プロモータ領域はcDNAやESTには写し取られない．ではプロモータ領域はDNA中でどのようにして同定すればよいか？　この目的のためには，転写開始地点から読まれたことが保証されている完全長

cDNAをDNAへと写像して，転写開始地点を特定すればよい．したがって完全長cDNAの収集は重要だが，収集には困難が伴い価値の高いデータとなっている．ヒト完全長cDNAについては東大医科研の菅野による卓越した研究がある．現在8,060個の完全長cDNAが公開されプロモータ領域の解析に使われている．

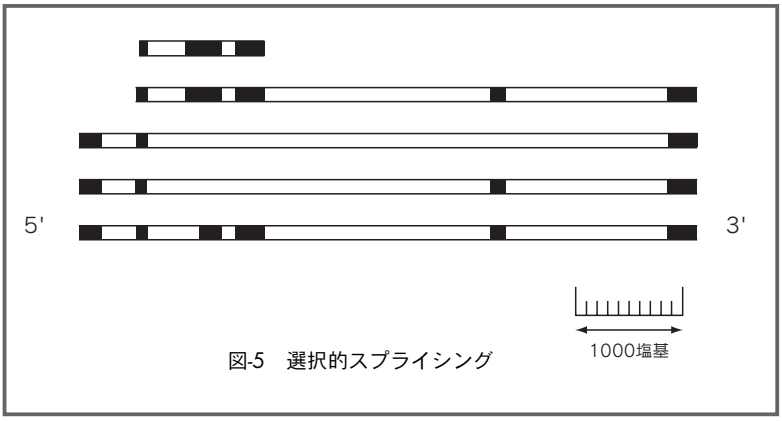

図-5　選択的スプライシング

ヒトゲノム解読の意義 －SNP解析

DNA配列には，先祖から伝承される個性が部分配列の特徴として蓄積されている．最近注目されている個性はSNP (Single Nucleotide Polymorphism) と呼ばれるDNA上の特定の位置における1塩基の変異である．偶然に同一の位置で同じ変異を起こす確率は低い．したがって同一のSNPを持つ集団は先祖を共有している可能性が高い．このため同一の遺伝病集団から共通のSNPが見つかったとすれば，そのSNPは疾患の原因遺伝子に連鎖して先祖から子孫に受け継がれた可能性を示唆する．このような理由から，遺伝病解析でのSNPの重要性は増している．SNPはすでに100万個以上DNA上に見つかっている．SNPの中で特に重要性が高いのはエキソン上のSNPであり，タンパク質のアミノ酸配列に決定的な変化をもたらす場合もある．またプロモータやイントロン中のSNPも遺伝子のきわめて近くに存在するため，原因遺伝子と連鎖して遺伝する可能性が高い．このようにエキソン上には存在しないもののエキソン周辺に実在するSNPは，cDNAではなくゲノムを解読することで初めて得られる情報である．エキソン周辺のSNP位置を確定するためにも，SNP周辺の配列をDNAに写像して，SNP位置を決めることが有効である．

ヒトゲノム解読の意義 －選択的スプライシング

多数のESTや完全長cDNAをDNA上に写像すると，同じ領域に同一エキソンを共有するESTが複数写像される場合が少なくない．たとえば図-5はヒト19番染色体の遺伝子座q13に写像された5つのESTの構造を示している．下から1番目のESTが6つのエキソンを持つのに対して，2番目のESTでは中間のエキソンが2個，3番目のESTでは3個が欠落していることが分かる．さらに4番目のESTでは5' 側から2番目のエキソンの長さが他のエキソンより長い特徴がある．また5番目のESTは4番目のESTが部分的に読まれた断片と考えられる．このようにDNAの同一領域から転写されていても，エキソンを選択的に組み合わせた転写物が複数生成される場合がある．この現象を選択的スプライシングと呼ぶ．いま選択的スプライシングは，にわかに注目されている．というのも，ヒト遺伝子の総数は4万個以下と予想され，線虫遺伝子の総数約19,000個の倍程度であり，遺伝子の総数と生物種の複雑度に相関は少ないと考えられるようになったからである．むしろヒトは選択的スプライシングにより多様な転写物を生成することで生体としての複雑さを獲得しているという見方が浮上している．選択的スプライシングのパターンを枚挙するためにも，ESTのゲノムへの写像が有効である．

ヒトゲノム解読の意義 －mRNA観測精度の向上

先にmRNAの観測を行うために多数のcDNAやESTが収集されてきた背景を述べた．観測の精度を上げるにはプライマーと呼ばれるmRNA中の長さが20程度の配列で，他のmRNA中にはほとんど出現しない，言い換えれば観測対象のmRNAを識別するのに十分なユニークさを持つ配列の設計が重要になる．従来プライマーはESTから設計されてきたが，ESTの読取エラー率が5%程度あり，このノイズが設計の障害となることが少なくない．一方，ヒトゲノムは何重にも読んでいるため，読取エラー率は0.01%以下といわれている．そこで，ESTをDNAへ写像した後に，その写像された領域のDNA配列を使ってESTの読取エラーを補正し正確なプライマーを作成することが可能になる．

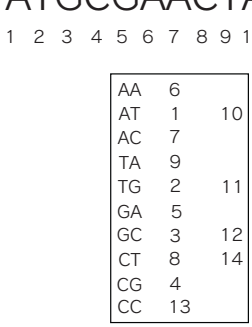

図-6　DNA配列と索引表

ESTとDNAデータの頻繁な更新

以上，ヒトゲノム解読がもたらす有用な情報について説明したが，中でもESTをDNAに写像する技術は基本的な道具となっている．ESTのデータベースであるdbESTには2001年9月末に約380万本のESTが登録され，しかも単調に増加している．一方，DNA配列の方は90％以上の塩基を読んだと見積もられているが，いまだ読み取られていないギャップ，誤った方向を持つ部分列が多数存在している．そのためデータベースは常に更新されており，完備な状態になるには恐らく2003年ごろまでかかるといわれている．したがって当面は，ESTデータベースおよびDNAの更新が続くので，更新されるたびにESTを写像するような体制が必要であろう．1日程度で写像できるソフトウェアはゲノム研究の推進に欠かせない．

ESTのDNA配列への写像技術 ー ESTの位置の近似と索引表

ESTをDNAへの写像は，どのようにすれば高速に実行できるか？　まず問題になるのは，DNAが長大であることである．ヒト1番染色体は全長で約2.6億塩基を含んでおり，常染色体で最も短い22番でさえ約4千万塩基ある．一方ESTの長さは高々数千である．そこでESTがDNA中でのコードされている範囲を近似することが計算時間を短縮するには有効である．範囲の絞込みのためには，まず長さが高々20程度の短い配列が，DNA中でどの位置に出現するかという情報を表にしておくことが有効である．このような表の概念は計算機科学では索引表（look-up table）と呼ばれるが，ゲノム解析用ソフトウェアでは1980年代から愛用されている．以後，索引となる短い配列をキー配列と呼ぶことにする．たとえば図-6では，上に長さ15の短いDNA配列が，下に長さ2のキー配列に対する索引表が示されている．索引表からはキー配列ATがDNA配列中で1番目，および10番目に出現していることが分かる．

索引表を用いると，ESTの最初のたとえば20文字がDNA中のどの位置に出現するか，その候補のリストを短時間で検索できる．一方ESTの最後の20文字についても同様である．このようにするとESTをコードする領域（最初から最後のエキソンまで）を狭めることができ，通常は数万，最長でも約300万塩基程度まで探索範囲を絞り込める．したがって索引表さえ構築できれば，ESTの位置の近似はさほど難しい問題にはならないが，索引表の構築には配慮が必要である．

巨大な索引表の構築

索引表にはキー配列がランダムに検索される．検索スピードを加速するためにも索引表を主記憶に常駐させるべきである．索引表を容易に主記憶に格納できるかといえば微妙である．約30億の塩基列には，長さNのキー部分列も約30億ある．仮にこれらキー配列の位置情報をすべて主記憶に格納するならば，32bitのアドレス空間では足りない．もし十分な主記憶，たとえば64bitのアドレス空間で少し多めに見積もって48GB程度あれば，すべてのキー配列を保存した索引表ができる．しかし2GBを超える計算機システムは高価である．

どうしても32bitのアドレス空間のなかで処理するならば，一歩後退することになるが，索引表を2次記憶のファイルへ展開して，シークする方法をとるのが一案であるが，速度は極度に劣化する．一方，重要なキー配列だけを残して他は間引くことで，索引表を2GB以内に抑える試みもある．まずヒトゲノムには全体に渡って稠密に存在する繰返し配列に着目する．たとえばSINEと呼ばれる繰り返し配列はゲノム上に100万コピー以上存在するので，SINEに含まれるキー配列を使っても，探索空間を狭めるには役に立たない．UC，Santa Cruzのジム・ケントは，出現数の多いキー配列を間引くことで，1GB程度の索引表を構築している．ただし間引かれたキー配列は，索引表での位置検索ができな

いため，DNAを直接走査することになり，計算の負荷を高める結果をまねく．

またNの大きさの選択もEST位置を近似する際に考慮しなければならない．たとえばN＝10の場合を考えると，配列の種類は$4^{10} \fallingdotseq 10^6$である．塩基がランダムに出現すると仮定しても，1つのキー配列が出現する位置の数は平均約3×10^3となり，位置の候補を絞り込むには不十分である．たとえばN=14であれば，配列の種類が$4^{14} \fallingdotseq 2.56 \times 10^8$となり，1つのキー配列あたり平均12カ所の出現となり，探索空間を比較的良好に絞り込めたといえる．

著者の一人小笠原は，索引表をすべて主記憶上に格納することで，1秒間に約100個のESTをDNA上に写像できるか否かを判定し，写像できる場合にはエキソンへの分解を行うソフトウェアを開発した．この結果，公開されている全ESTをDNA上に1日以内で写像することが可能になった．キー配列の長さNは14とし，使用している計算機は富士通社PrimePower 1000である．PrimePowerはSMP型並列計算機であるが，DSM型やクラスター型並列計算機上で索引表を適切に分解して，どの程度の高速化を達成できるか否かを試すことは興味深い．

配列比較

DNA中にESTがコードされている領域を絞り込んだ後は，ESTをDNAへ詳細に写像する配列比較の技術が必要になる．たとえば，短いDNA配列ATGCGATTAGと短いEST配列CGTTを比較することを考える．DNAはCGTTを部分配列として含んでいないが，きわめて近いCGATTを部分列として含んでいる．そこでCGATTのAが読み飛ばされた配列がCGTTであると考えるのが1つの解釈である．この読み飛ばしを「-」という記号で表現し，ギャップと呼ぶ．するとATGCGATTAGがCGTTを含むことを以下のようにギャップを使って表現できる．

```
ATGCGATTAG
||  ||
CG-TT
```

縦棒「|」は塩基が一致したことを示すのに対して，スペースは一致していない，もしくは対応する塩基が読み飛ばされて存在しないことを示す．2つの配列の間で各塩基がどの塩基で対応しているか，もしくは読み飛ばされているかをたとえば上のように記述した結果をアラインメントと呼ぶ．

読み飛ばしは，DNAからmRNAへと転写およびスプライシングが情報伝達する過程で発生するか，もしくはmRNAの読み取りミスである．mRNAの読み取りミスは約5％程度と見積もられており，mRNAの頭部と尾部に多い．ESTとDNAの間で不一致が起こる他の原因として，DNA上の変異も考えられる．たとえばATGCG[A]TTAGの四角で囲ったAがTの変異と仮定すれば，CG[A]TをCGTTと解釈し，この解釈を以下のようなアラインメントで表現できる．

```
ATGCGATTAG
||  |
CGTT
```

AとTが不一致であるため縦棒は引かれていない．

ESTとDNAの塩基配列を比較してみると，上の例のように複数のアラインメントが得られることがしばしばである．そこで各々のアラインメントに点数をつけ，高い点数を獲得したアラインメントを効率的に計算する方法が提案されている．これらのアルゴリズムの概略は本特集の阿久津の記事，詳細はPevznerの成書[3]に譲り，ここでは研究の歩みを述べるにとどめる．DNAの長さをM，ESTの長さをNとすれば，動的計画法を使って最適アラインメントを計算するSmith-Watermanのアルゴリズム[4]は1980年代前半に提案されており，その最悪計算量は$O(MN)$である．しかし，現場で使う際には速度が問題になり，また最悪計算量が大きく改善されたアルゴリズムも過去報告されていない．このような状況のため最適アラインメントを計算するのではなく，発見的方法で比較的良い解を計算するソフトウェアの開発が1980年代後半からは主流になり，現在BLAST，FASTAが最も利用されるようになる．

エキソンへの分解

いままでに述べた配列比較の方法では，エキソンを認識するような配慮がなされていない．たとえば，以下のアラインメントを考えてみよう．

```
AAGCATTGGTCAGCCATGCT
|||  |  |    |||
ATT-G-C---CAT
```

ESTが4つの部分配列に分断されている．5つのギャッ

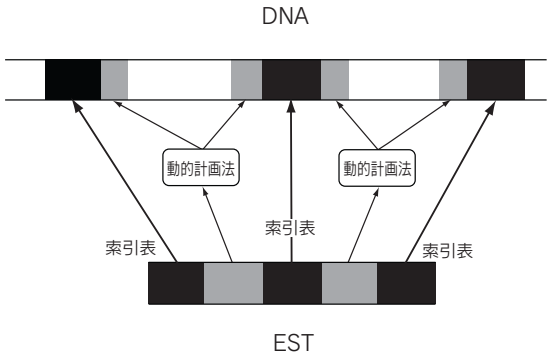

図-7　索引表と動的計画法の組合せによるESTのDNAへの写像

プを，まとめることで以下のような2つの部分配列に分割したアラインメントも生成できる．

```
AAGCATTGGTCAGCCATGCT
||||        ||||
ATTG-----CCAT
```

2つのアラインメントを比較すると，どちらの場合もESTの塩基はすべてDNAの塩基と一致しており，さらに導入されたギャップの数も5で等しい．しかし，エキソンへと分解することを目標にすれば，ギャップはできるだけ連続するように配慮すべきであり，上の例では後者のアラインメントを選択したい．この目的を満たすため1982年に後藤はSmith-Watermanの動的計画法を改良している[5]．後藤の動的計画法もそのままではESTを長大なDNAに高速には写像することは困難であるが，索引表との組合せで写像を効率化できる．図-7に示すように，EST中で索引表を使ってDNAに写像できる部分配列（黒い箱）をまず写像する．通常この計算だけではESTを完全には写像できない．なぜなら読み取りミスやDNA中の変異などにより索引表だけでは写像できなかった配列（灰色の箱）が残るからである．残された部分列は動的計画法によりギャップや塩基の不一致を許しながら丁寧に写像する．

図-7は著者らが試みた方法を示している．索引表を使った写像方式にはさまざまな工夫が可能である．たとえば，先に述べたようにUC，Santa Cruzのジム・ケントは，索引表から頻度の多いキー配列を除くことで，小さな索引表を作成している．最悪の場合，あるEST配列のキー配列が1つも，この小さな索引表に載っていない可能性もある．しかし現実にはどれか1つのキー配列が載っている場合がほとんどであり，このキー配列がDNA上どの位置に出現するかを頼りに，前後に写像を伸長させる方式をとっている．ちなみにジム・ケントのソフトウェアを著者らの計算機（PrimePower1000）にインストールして実行したところ，1秒間に平均して1個のESTを写像することができた．著者らは索引表を間引かずに使って1秒間に平均100個なので，小さな索引表を使いながらも健闘している．

ヒト遺伝子地図の閲覧サイト

ESTをDNAへ写像した結果は，数値やテキスト情報としてではなく，グラフィカルに表示してユーザである医科学・生物学者に提供する必要がある．以下のWWWサイトが写像結果を公開している．

- Human Genome Browser (UC, Santa Cruz, US)
  http://genome.cse.ucsc.edu/
- MapViewer (NCBI, NIH, US)
  http://www.ncbi.nlm.nih.gov/genome/guide/human/
- Ensembl Genome Server (Sanger Center, UK)
  http://www.ensembl.org/
- Gene Resource Locator (University of Tokyo, Japan)[6]
  http://grl.gi.k.u-tokyo.ac.jp/

これらツールを利用する際には，（1）自分が興味を持っている遺伝子が正確に写像されるかどうか？　（2）写像された場合，選択的スプライシング，SNP，転写制御領域の構造について十分な情報が得られるか？　（3）GUIは使い勝手がよいか？　などが評価基準になるであろう．また，これらのサイトが使用している写像方式はまったく異なり，更新の頻度もまちまちなので，併用しながら情報を補うのがよいであろう．各サイト開発者も頻繁に他のサイトを参照して，改良を重ねているようである．互いに切磋琢磨することで，より良いゲノム構造の閲覧サイトができあがってゆくことを期待したい．なお，著者の一人森下が平成13年度冬学期に東京大学大学院にて講義した「バイオインフォマティクス・ソフトウェアの構築」の講義ノートを以下のサイトで公開している．参考になれば幸いである．

http://www.gi.k.u-tokyo.ac.jp/~moris/lecture/VLDB-Genome/

参考文献

1) Watson, J.D., Hopkins, N.H., Roberts, J.W., Steiz, J.A. and Weiner, A. M.: Molecular Biology of the Gene, Fourth Edition. Benjamin/Cummings Publishing Company, Inc. (1988).
2) Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J.D.: Molecular Biology of the Cell, Third Edition. Garland Publishing, Inc., (1994).
3) Pevzner, P.A.: Computational Molecular Biology－An Algorithmic Approach. MIT Press (2000).
4) Smith, T.F. and Waterman, M.S.: Identification of Common Molecular Subsequences, Journal of Molecular Biology, 147, pp.195-197 (1981).
5) Gotoh, O.: An Improved Algorithm for Matching Biological Sequences, Journal of Molecular Biology, 162, pp.705-708 (1982).
6) Honkura, T., Ogasawara, J., Yamada, T. and Morishita, S.: The Gene Resource Locator: Gene Locus Map for Transcriptome Analysis. Nucleic Acids Research, Jan. (2002).

（平成13年10月22日受付）