

# 5

## ゲノムデータの視覚化による効果的な理解

伊藤武彦

(株) 三菱総合研究所  
takehiko@mri.co.jp

**近**年のシーケンス技術の発展に従って、微生物からヒト、マウスなどにいたるまでさまざまな種類、サイズのゲノムが日々決定され、各種特徴量や遺伝子関連情報などゲノム上に隠された膨大な情報や知見が蓄積されてきている。

これらの情報は相互に密接な関連性を有し、また空間的、時間的にさまざまな広がりを見せており、膨大な情報の中から必要な情報のみをユーザに的確に伝えるために視覚化、可視化が必要不可欠なものとなっている。

本稿では、ゲノムに代表される膨大な情報がどのように効果的に視覚化されているのか実例を取り上げて紹介する。

れることが可能になる。また複数生物種のゲノムを比較することで進化の歴史を知ることにも可能になる。これがヒトを始めとしてさまざまな生物種のゲノム解読が盛んに行われている理由なのである。さらにゲノムには、ACGT という文字の並びとしてデジタルに表現が可能であるという特徴がある。この特徴を活かし、解読される膨大な量のゲノムの解釈には、生物学の範疇を超え、さまざまな情報学のテクニックが用いられており、これが急速なゲノム解読、解釈を可能にしている1つの理由ともなっている。ヒトのゲノムサイズは約30億文字、3,000Mbpで、新聞朝刊に換算すると約25年分にも相当し、線状のゲノムは23本の染色体と呼ばれる構造体

### ゲノムとゲノムに隠されたさまざまな情報

目の色や巻き舌のできるできないといった特徴が親子の間で似ることは遺伝と呼ばれ、親から子へとこれらの特徴は遺伝情報として伝えられる。この仕組みは基本的には全生物共通のものであり、ある個体を持つ遺伝情報全体のことはゲノム、遺伝的な形質を規定している個々の因子は遺伝子と呼ばれ、遺伝子はゲノム内のある領域に埋め込まれている。ゲノムの実体は、DNA（デオキシリボ核酸）という化学物質がフォスフォジエステル結合を作り、分岐のない鎖状の分子構造をとっている生体高分子である。DNAはデオキシリボース（糖）、リン酸、塩基から構成される核酸の一種で、塩基にはアデニン、シトシン、グアニン、チミンの4種類があり、それぞれA,C,G,Tと略される（図-1）。

ゲノムにはある生物個体の遺伝情報が詰まっており、ゲノムを解読することでその生物の設計図全体を手に入

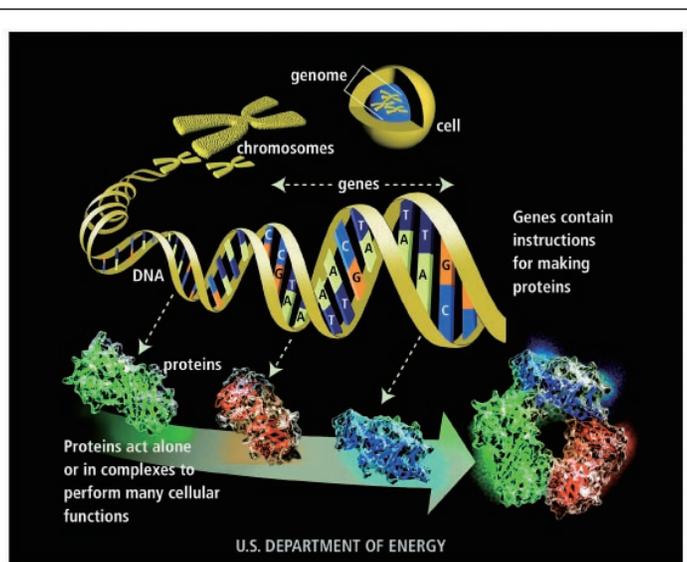


図-1 細胞から染色体、DNA、タンパク質へ  
[http://www.ornl.gov/sci/techresources/Human\\_Genome/education/images.shtml](http://www.ornl.gov/sci/techresources/Human_Genome/education/images.shtml) より転載

に複雑に折り畳まれて格納されている。染色体の本数も生物種によって大きく異なる。一方それに対し下等な原核生物では、ゲノムは環状になっている場合がほとんどである。

では、ゲノムの中に遺伝情報すなわち遺伝子は一体どのような形で詰まっているのであろうか。実はゲノムという暗号文章からの、遺伝子の完全な解釈というのはまだなされていないのである。遺伝子を代表するものとして、生体内でさまざまな働きをするタンパク質の鋳型となっているものが多数存在する(図-1)。さまざまな情報が実験的に得られているタンパク質をコードしている遺伝子でさえもその完全な情報を得るには至っていない。ましてや、その遺伝子からいつでもどこでもタンパク質を作り出すかを指令しているとされる、転写制御領域の暗号解明はまだ先のことである。そのため、ゲノムからの遺伝子発見など暗号解読のために、実験のみならずさまざまなゲノム配列からの情報学的特徴量抽出が行われ、何とか解釈を加えていこうというのが現状なのである。たとえば、タンパク質をコードしている領域の前にはGやCが偏って分布しているCpGアイランドと呼ばれる領域を形成していることが多いとか、進化を考えた場合に、遺伝子など重要な領域にはジャンク部分と比べて変異の蓄積が少ないと考えられるため、生物種間でゲノムを比較すると遺伝子領域は保存度が高いとか、ゲノム配列から得られるさまざまな特徴量に基づき遺伝子領域を絞っていくことが通常行われている。

このようにゲノムの解釈には、さまざまな情報を列挙、比較していくことが必要になるが、対象となるデータは非常に膨大である。たとえばヒトゲノム3,000Mbpに対し得られる情報はゲノムサイズの何倍もの大きさになるため、これを人間が解釈するためにはどうしても可視化による手助けが必要となってくる。ゲノムの未知な領域の役割の研究や、新規遺伝子領域の発見のための研究では、さまざまな多くの擬陽性を含んだデータが大量に出される。これは実験の感度などに起因することが多く、その中から本当に意義の解釈をすることが求められる。ゲノム上に多数観測されるシグナルがあったとして、それが本当のものなのかあるいはある領域を持つ配列のエントロピーに起因するノイズなのかなどは、他の実験結果や配列の特徴量と並べて人間が判断しないといけなことが多い。ある実験結果のデータはどうも遺伝子間にピークが見られることが多いとかいった発見も計算機的手法で実現することは困難で視認によってなされることが多いのである。

また、染色体全体をマクロに俯瞰することによって、染色体の各領域が持つであろう機能的な違いを類推するような研究のサポートになる場合もあれば、逆に数kb

を拡大し塩基ごとに得られる情報からの研究、たとえば転写制御領域予測などのサポートになる場合もある。これら非常にマクロな情報の俯瞰からミクロな領域の詳細なブラウズまでさまざまなスケールの間を自由に行き来が可能で可視化が必要とされることが多い。実験技術の進歩に伴い必要とされる表示したい情報、可視化技術も常に新しいものが要求されている。

可視化を技術的側面から見ると、ゲノム情報の表示そのものが研究対象となるようなアルゴリズムとしての目新しさはほとんどない。ただ、数百GBにも及ぶような膨大な情報をデータベースに格納し、ユーザがある視点に立ったときに該当領域だけの情報を高速に取り出してそのデータから描画するためのテクニックが必要にはなる。それ以外はいかに情報を分かりやすく伝えるかのテクニカルなアイデアが問われるところとなる。

次章以降では、ゲノムに関する膨大な情報がどのように効果的に視覚化されているのか、実例を取り上げて紹介することにする。

## ゲノムブラウザの紹介

前章で述べたようにゲノム情報を効果的に俯瞰するためには可視化が必要となり、微生物からヒトなど高等真核生物に至るまで、さまざまなゲノムブラウザが開発されている。本章ではさまざまなゲノムブラウザによるゲノム情報の視覚化事例を紹介する。GenBank, DDBJといった配列データベースが「1. バイオデータベースの歴史と展望」で紹介されているが、これらのデータベースは研究者が個別にデータを登録することが可能であり、その意味で多少の間違いの存在や統一的なアノテーションが付与されていないなど、いわばデータバンクとなっている。たとえばGenBankでヒトの核酸配列は1,088万本も登録されており(2006年1月現在)、ACGTが記載されたゲノムの断片配列情報と、その由来がヒトである以外に情報がないデータも非常に多く、ユーザの使い勝手は必ずしもよくはない。

それに対し各ゲノムデータベース・ブラウザは、1次データベースのデータを精査して作成されたゲノム配列に対して、遺伝子情報などを付与した2次データベースとなっており、ユーザの使い勝手を考慮された作りとなっている。ゲノムブラウザとしてバクテリアのものとヒトのものを紹介するが、バクテリアではゲノムの各領域が果たす役割はヒトと比べてかなり解明されており、種間の違いによる進化の研究などに図示化が多く用いられる。ヒトではいまだゲノムの果たす役割の未知な領域が圧倒的に多く、その解明のために図示化が多く用いられており、両者では表示対象が違うもののその技術に差

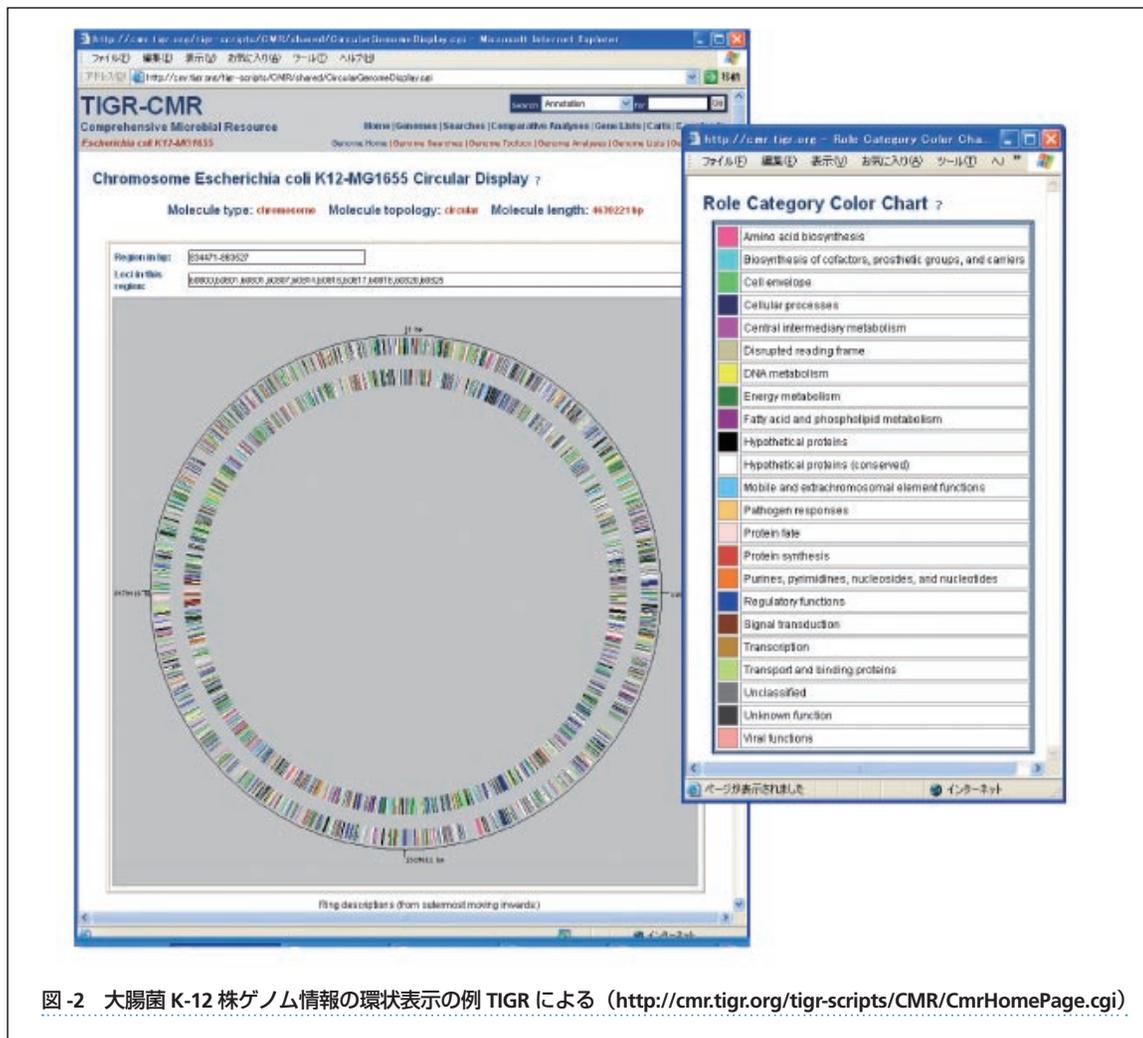


図-2 大腸菌 K-12 株ゲノム情報の環状表示の例 TIGR による (<http://cmr.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi>)

があるわけではない。

これらゲノムブラウザで表示されるデータは、大別するとゲノムの位置情報を X 軸にとった場合に各ポジションに対して得られる Y 軸方向の浮動小数点情報か、ゲノム上の位置をどこからどこまでが何というかたちで規定するタイプのデータに集約される。後者は GFF 形式と呼ばれるゲノム上でのスタート位置、エンド位置などを TAB 区切りで記載したフォーマットのデータである。タンパク質の立体構造表示のように 3 次元データを表示したりすることはまずない。また、ゲノムブラウザに付随する機能として表示される遺伝子の詳細情報などは遺伝子の名前をキーとしたリレーショナルデータベースに格納され、呼び出しに応じて必要な HTML や図を cgi で作成する。

### バクテリアゲノムブラウザ

バクテリアは、1995 年にインフルエンザ菌の全ゲノムが決定されたのを皮切りに数多くのゲノムが決定されている（2006 年 1 月現在 268 種解読完了：Genomes OnLine Database による）。バクテリアではゲノムに対する遺伝子の予測は容易である。しかしながら、その機能

はまだ分かっていないものが数多い。図-2 に TIGR から提供されている大腸菌 K-12 株のゲノムに対するアノテーション情報のページを示したが、このようにバクテリアのゲノム情報はその形状から環状に表現されることが多い。この図では機能ごとに色分けがなされている。

一方図-3 には NCBI gMap を用いた大腸菌の仲間同士の比較の結果図を示した。各行がそれぞれ大腸菌の個々のゲノム情報に対応し、その下に色つきの矢印で示された領域が種類間での配列の類似性が高い相同領域になる。このような表現方法を用いるとゲノム上での領域の入れ替わりや欠損、挿入が一目で分かる。たとえば図中の赤四角で囲んだ領域 2 は上の 2 つの種類にしか存在しない。2 つの種類はいずれも病原性大腸菌 O157 であり、この挿入された領域に病原性をもたらす遺伝子群が実際に含まれていたのである。K12 と O157 とをゲノムレベルで図示化して比べることにより、挿入された領域の存在が明らかになるとともに、病原性という実際にその菌が持つ機能とゲノム情報とが結びつくのである。

### ヒトゲノムブラウザ

ヒトを始めとした高等真核生物では、バクテリアのゲ

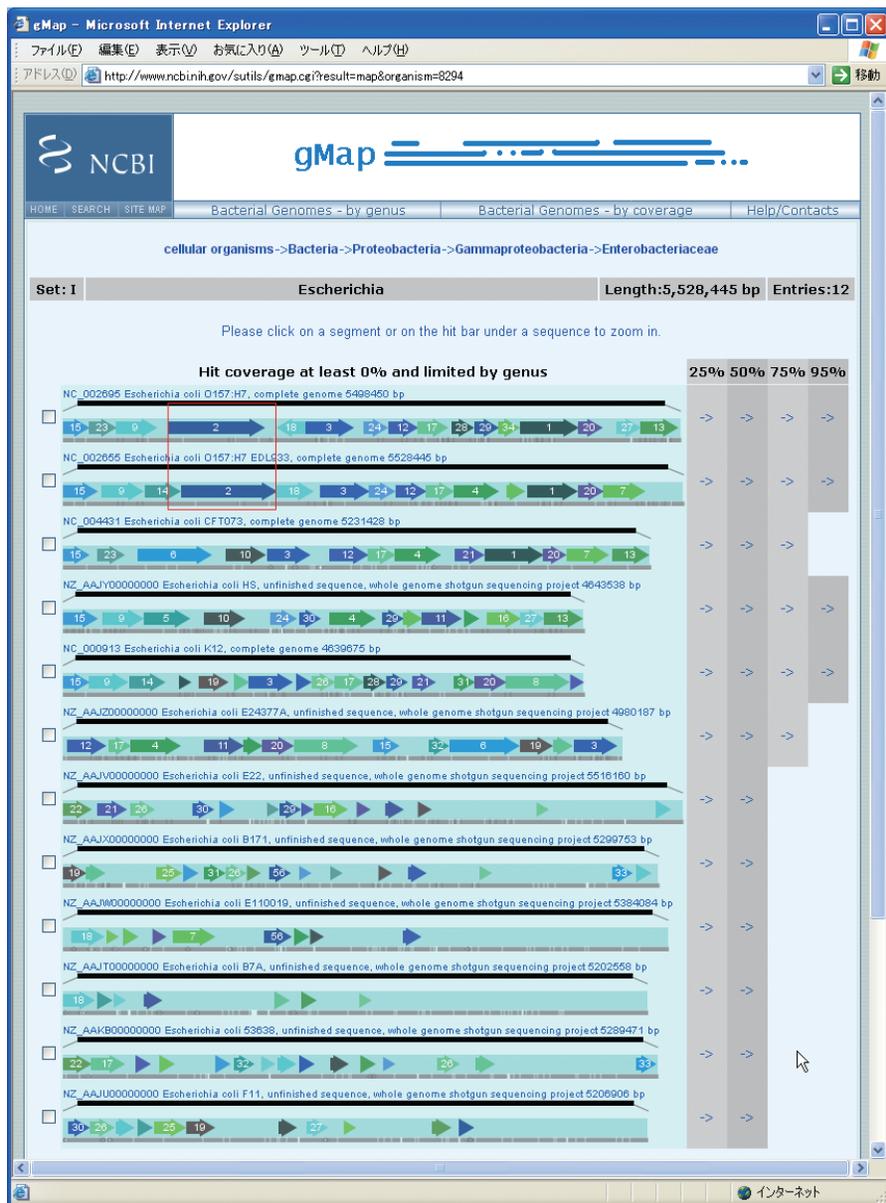


図-3 NCBI gMap による大腸菌同士の比較表示例  
上から2つが O157 で上から5つ目が通常実験で用いられている種類である。  
<http://www.ncbi.nlm.nih.gov/sutils/gmap.cgi>

ノムと比べて圧倒的に遺伝子密度が低く、遺伝子構造もエクソンと呼ばれるタンパク質をコードしている領域が、イントロンと呼ばれるゲノム領域で分断されているため複雑であり、遺伝子領域予測は困難である。このため、ゲノムブラウザでは遺伝子情報のみならず、各種特徴量や多少の擬陽性を含んだ予測情報なども合わせて表示し、ユーザが表示される情報を取捨選択した上で実験などの足掛かりとするような使い方も多い。ヒトゲノムのブラウザとしては、NCBI, University of California から提供されているものなど著名なものはいくつか存在するが、本節では京都大学から提供されている HAL データベース (<http://hal.genome.ist.i.kyoto-u.ac.jp/>) を中心に紹介する。

HAL データベースでは染色体の一領域に関し、その領域に含まれる遺伝子を中心としたさまざまなアノテ

ション情報を提供している。ヒトゲノムに対する遺伝子アノテーションではまだ完全なものは存在しないため、複数のプロジェクトが予測した遺伝子を並べて表示することで、その結果を比較しユーザがその情報の信頼度を判断した上で利用できるようにデザインされている。また、各遺伝子に関する詳細情報を提供する機能や、各種キーワードによる検索、ユーザが手持ちの DNA 配列および遺伝子情報を HAL 上で見る機能を有している。まず、トップページからある染色体をクリックなどで指定すると、図-4 のような各染色体のトップページへと移動する。この画面では指定した染色体に関し、各手法で予測された遺伝子の数などの統計情報や染色体全体に対する遺伝子密度①、GC 含量の分布②、さらにはマウスゲノムとの類似領域に基づいて進化的に保存されている

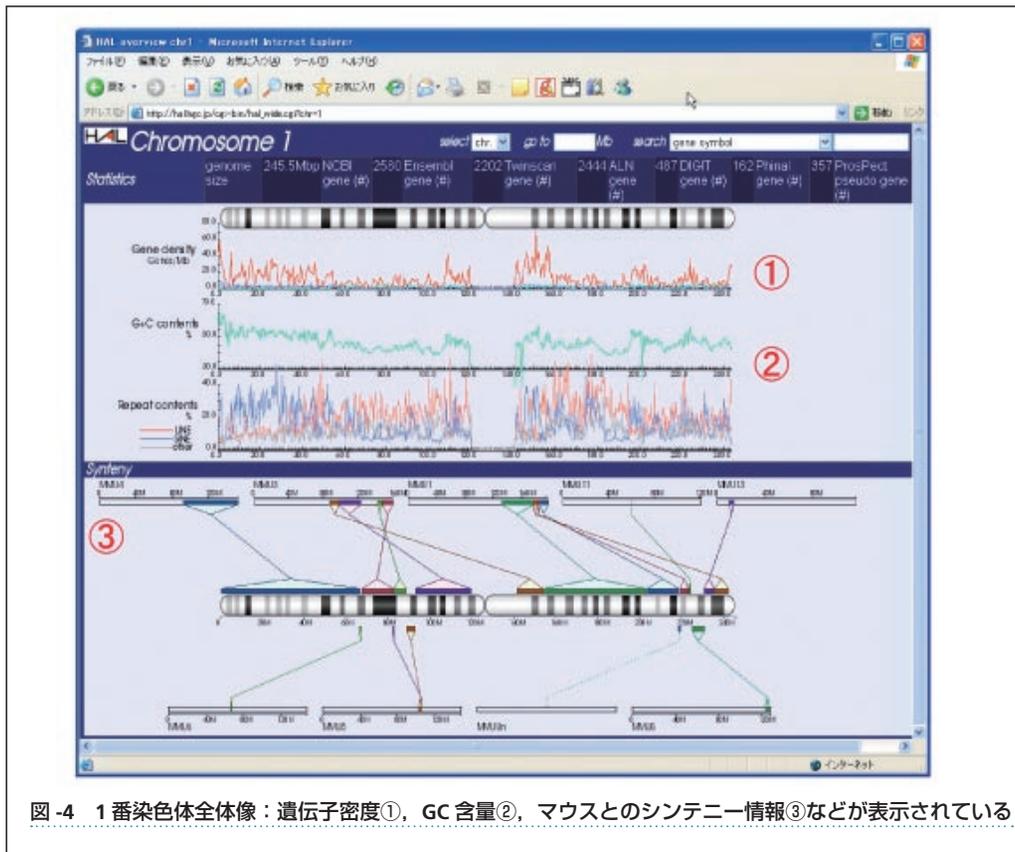


図-4 1番染色体全体像：遺伝子密度①，GC含量②，マウスとのシンテニー情報③などが表示されている

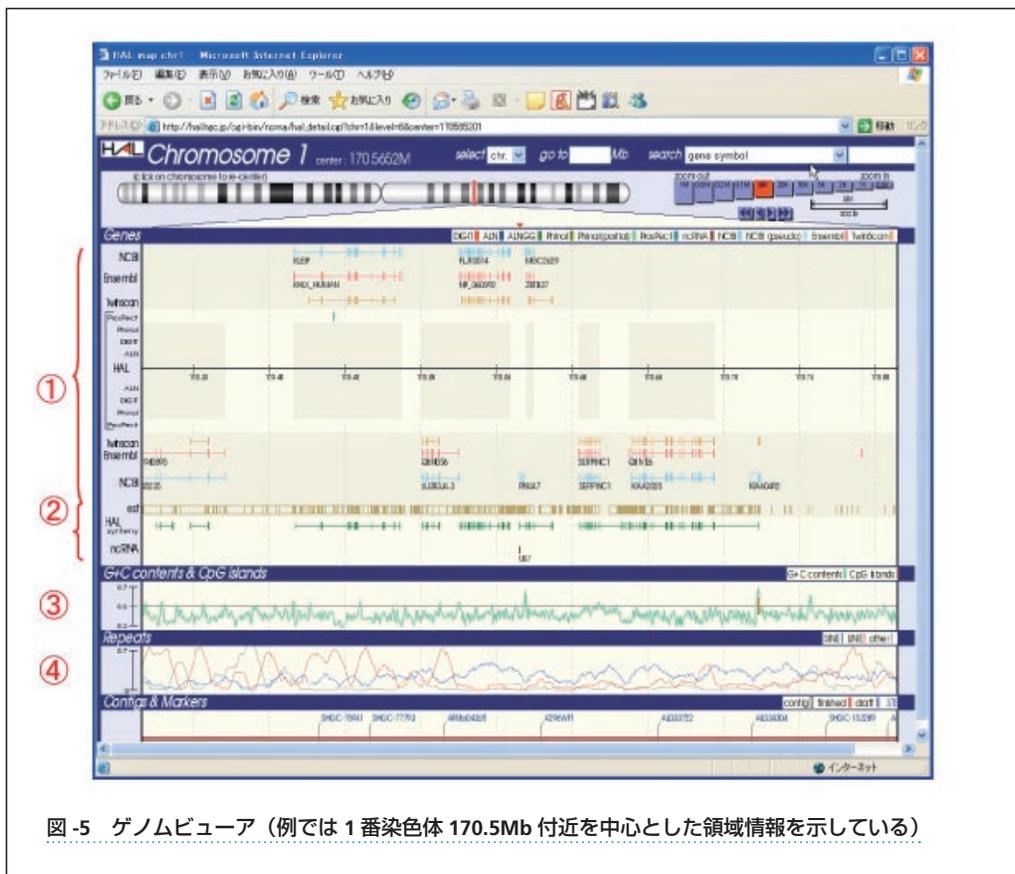


図-5 ゲノムビューア（例では1番染色体170.5Mb付近を中心とした領域情報を示している）

領域を示すシンテニー情報③など染色体全体が持つ特徴を俯瞰することが可能である。

次にこの画面上の染色体の図をクリックすると、クリックした領域を中心としたゲノム領域のアノテーション情報が見られるページ（ゲノムビューア）へと移動

する（図-5）。ゲノムビューアでは、染色体のある領域（図-5では170.3Mbから170.8Mbの0.5Mb領域）に関するアノテーション情報を表示している。この際に裏では各情報が格納されたデータベースから該当領域に含まれる情報を抽出し、列ごとにcgiによる描画が行われ

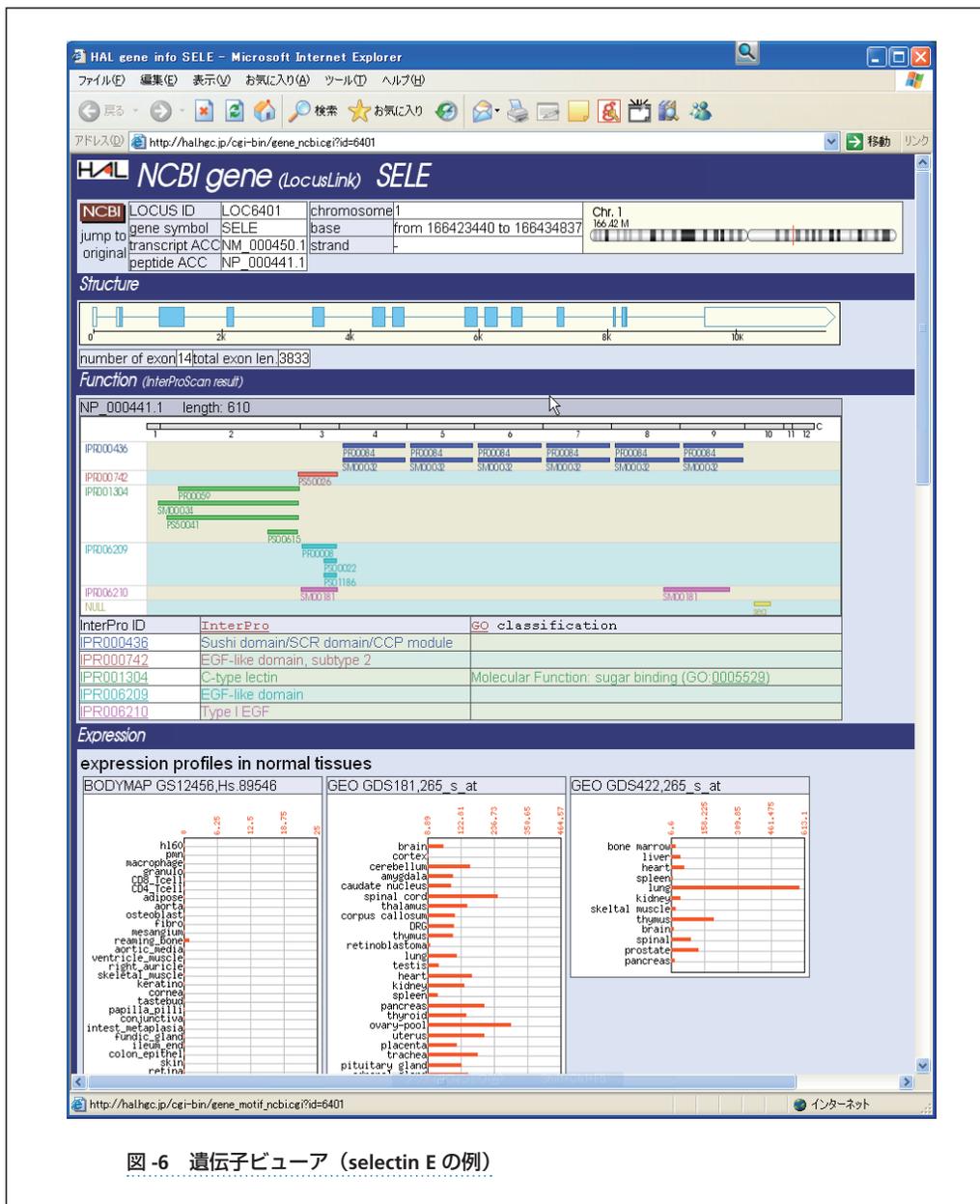


図-6 遺伝子ビューア (selectin E の例)

ている。画面中 Genes の領域 (①) では各予測手法により予測された遺伝子を示している。下部②では、遺伝子の断片配列である EST のヒット情報および、マウスとの保存領域情報が示されている。その下の領域③では GC 含量および CpG アイランドの位置を、さらに下の領域④ではゲノム中繰り返し配列の分布などが示されている。これらの情報は個別のデータベースに格納されており、それをゲノムの位置情報をキーとして取り出したものである。

表示させたい領域の移動や表示領域の拡大・縮小には画面上部のアイコンを用いる。イデオグラム上をクリックするあるいはテキストボックス内に移動したい場所を指定することで、希望する領域へと移動することが可能となる。拡大・縮小時にはその解像度に応じて折れ線グラフのタイプのデータは各ポジションでの値を計算しなおしている。

図-5 で示されたゲノムビューアからある遺伝子をク

リックすると図-6 で示すような遺伝子ビューアへと移動する。この遺伝子ビューアでは、遺伝子に関する外部データベースでのアクセス番号、ゲノム上での位置などの情報、エクソン・イントロン構造の模式図、mRNA 配列、アミノ酸配列のほか、同様な機能を有した遺伝子間に共通に見られるモチーフ情報の予測結果、あるいは NCBI GEO に登録が見られる遺伝子に関しては、その遺伝子がどの臓器でよく用いられているかを示した発現プロファイルも同時に見られるようになっている。このようにヒトゲノムに関するブラウザでは、染色体全体を俯瞰するようなマクロな視点から、個々の遺伝子、配列に関するまでのミクロな視点までを自由に行き来することが求められる。

### そのほかのゲノム情報の可視化

エクソン-イントロン構造や、モチーフの位置情報に関しては、文字による情報よりも図示化による方が直感

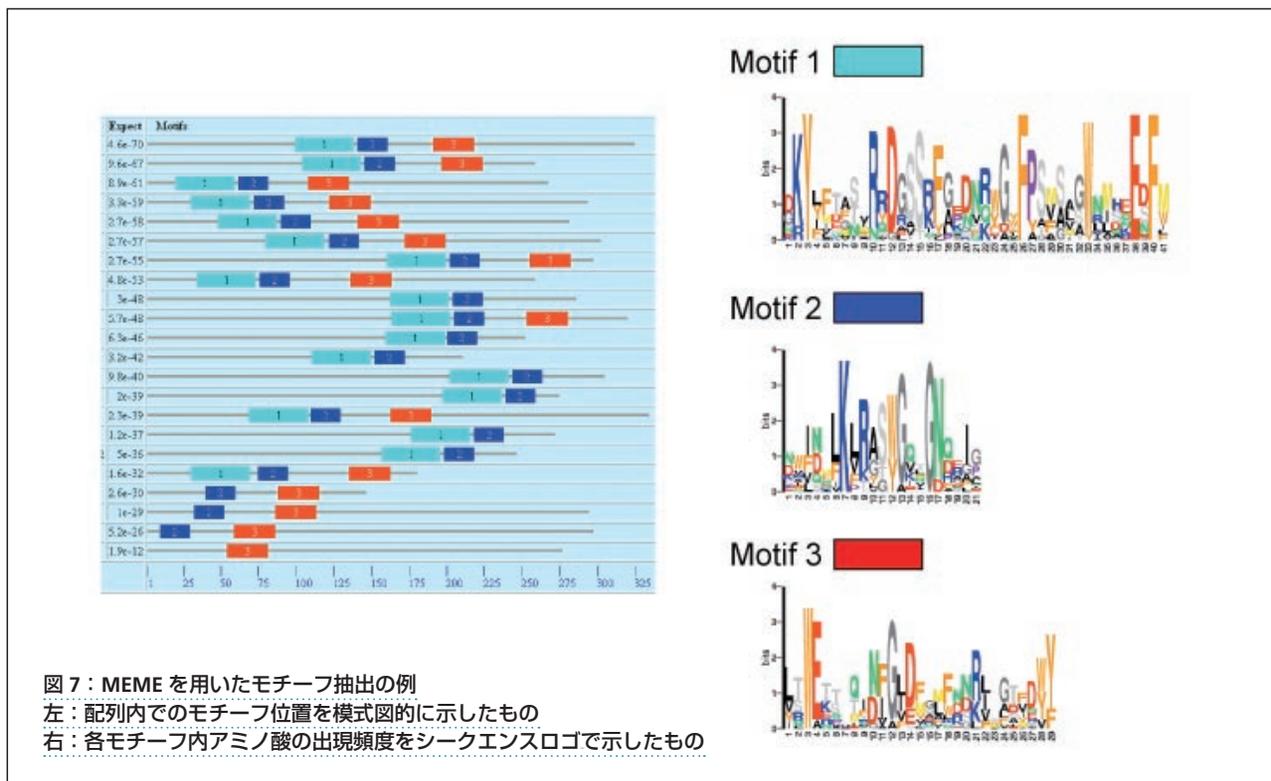


図7: MEMEを用いたモチーフ抽出の例  
 左: 配列内でのモチーフ位置を模式図的に示したもの  
 右: 各モチーフ内アミノ酸の出現頻度をシークエンスロゴで示したもの

的にも理解が容易である。図-7には、機能が未知ではあるけれども互いに相同性を示すアミノ酸配列群に対して、MEMEと呼ばれるモチーフ抽出プログラムを用いた結果を示す。モチーフとは複数のアミノ酸配列内に共通に見つかる配列の領域で、それぞれがある特徴を持った働きをしているとされている領域である。図左では各行が個々のアミノ酸配列に対応しており、水色、青、赤で示されている四角がモチーフになる。結果より明らかにこれらの配列群には3つのモチーフ領域が同じ順序で現れることが見てとれる。右に示した図はシークエンスロゴと呼ばれるもので、3つの各モチーフ内の各配列の位置でどのようなアミノ酸がとられやすいかを示している。各文字はアミノ酸配列で大きいほどその場所でその配列がとられやすいことを示している。よく保存されている配列部分は機能的に重要であると考えられ、さらには構造などと絡めて機能の類推への発展も期待できる。

図-8に示したものは、JST(科学技術振興機構)より提供されている、ヒトゲノムのある領域に対する他生物種ゲノムの類似度をグラフにした比較ゲノムブラウザである(<http://www-btln.jst.go.jp/ComparativeGenomics/>)。各行が生物種に対応しており、縦軸に類似度を示す。上から順にチンパンジー、牛、イヌ、マウス、ラット、オポッサム、ニワトリ、フグゲノムのヒトゲノムとの類似度が示されている。図中黄色で示されているのが遺伝子領域であり、赤でエクソン領域が示されている。図より遺伝子領域、特にエクソン領域が生物種の間でよく保存されていることが確認できる。このブラウザも拡大、縮小が自由に行え、保存されている領域の配列取得も可能

である。既知の遺伝子領域以外にも生物種間で高度に保存されている領域も散見でき、そのような領域は何かしらの生物学的な意味を持つ領域である可能性が高い。

最後に図-9にヒトゲノム21番染色体同士を比較した図を示す。これはドットプロットもしくはハープロットと呼ばれるもので、縦軸、横軸にそれぞれ配列をとり、共通の配列が認められるところにドットを打っていくことで、視覚的に配列間の関係を捉えるために用いられる手法である。この図では、さらにマクロに捉えるために配列の類似度に応じてドットを色分けしており、赤くなるほどその類似度が高いことを示す。縦軸、横軸に同じヒトゲノム21番染色体全体を用いた自分同士の比較解析であるため、100%の一致を見た結果が対角線上に赤く示されている。興味深いことに、それ以外にも4Mbほどに渡って青色の線が対角線に平行に見て取ることができる。これは実際には、非常に相同性の低い領域が島状に、しかしある曲線上に点在していることに由来しているもので、マクロに見ているためにその存在に気づくことがようやくできるスケールの事象である。この図によって初めて、ヒト21番染色体体内に非常に太古に起こったゲノム重複の存在が知られることとなった。

### ゲノム情報可視化の与えるインパクト

以上、ゲノム情報の可視化事例をいくつか駆け足ではあるが、紹介してきた。いずれもゲノム配列から得られる膨大な情報をより効果的、直感的に示すためにさまざまな工夫がなされたものである。これらの可視化結果を生物学的見地に基づいて解釈することで、新しい知見の

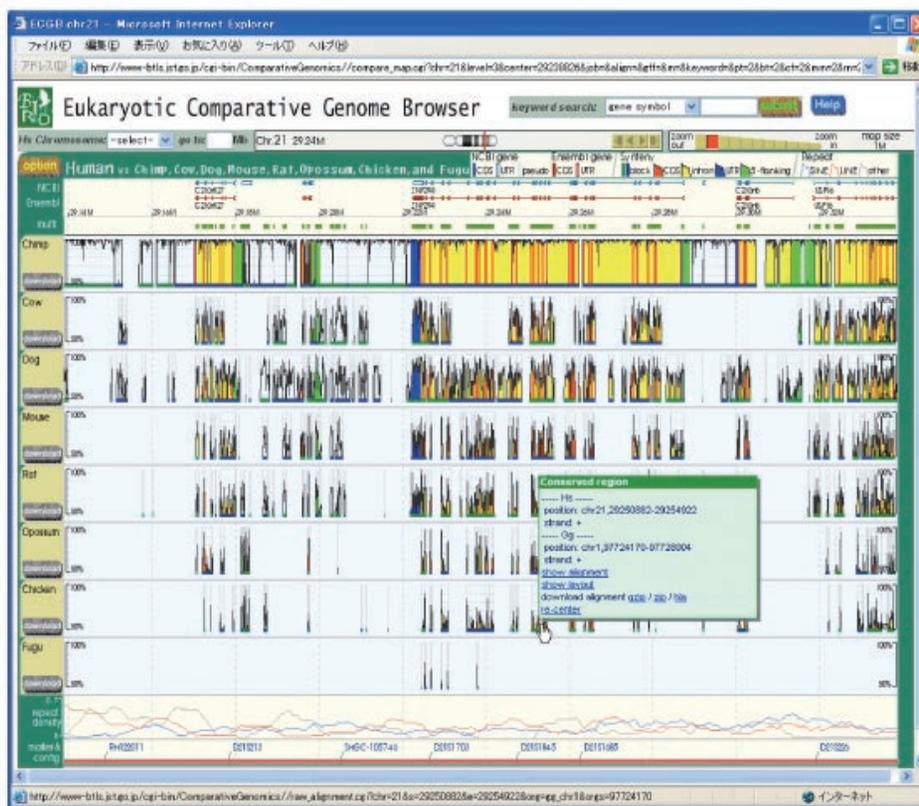


図-8 比較ゲノムブラウザの例：ヒト21番染色体29Mb付近を表示したもの

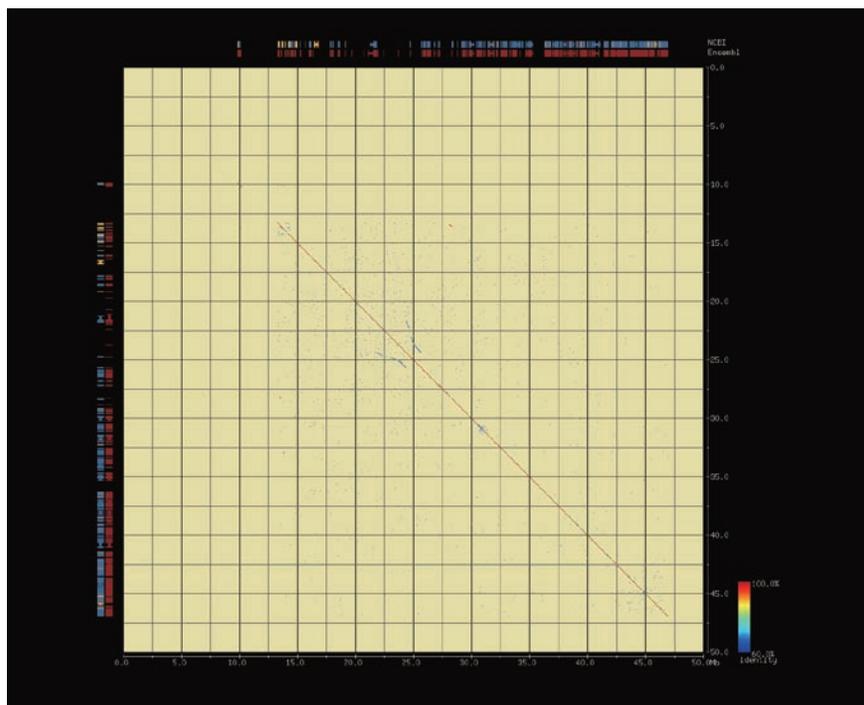


図-9 ヒト21番染色体同士のドットプロット例

発見につながっていく可能性について少しは触れることができたのではないかと思います。可視化の技術自体は何も目新しいものではない。しかしゲノムに対する解析は、実験的にも、情報学的にもさまざまな新規手法が用いられ、生み出されるデータはますます膨大なものになってきており、可視化なくしてこれらのデータを理解することはほぼ不可能になってきている。今後もこれら解析に

よって生み出されるデータを無駄にせず、新たな生物学的知見発見の効果的なサポートのために可視化技術の重要性が失われることはないであろう。

#### 参考文献

- 1) Andrews, B. D. et al.:Ensembl 2006, Nucl. Acids Res. 2006 34: D556-D561.
- 2) Benson, D. A. and Karsch-Mizrachi, I. et al. : GenBank, Nucl. Acids Res. 2006 34: D16-D20.

(平成 18年 2月 3日 受付)