「生物情報学とは ?」

～生物学と情報学の出会い～

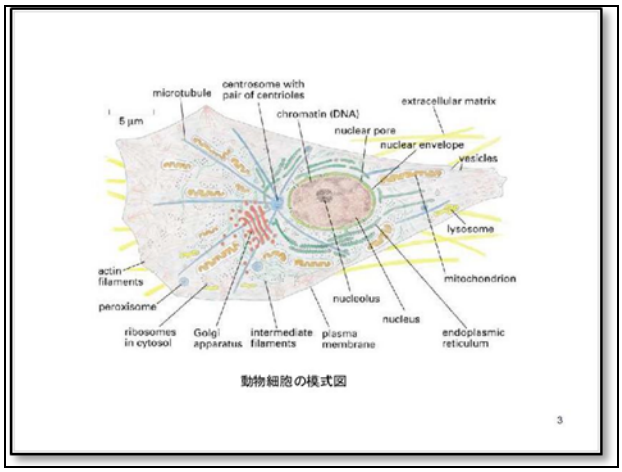
群馬大学大学院工学研究科准教授
宮　澤　三　造

1. はじめに

宮澤でございます。ご丁寧なご紹介を有難うございます。このような機会を与えて下さった関係者の方々に深く御礼を申し上げます。

今日は、わたくしの研究しております生物情報学という分野につきましてご紹介したいと思っております。生物情報学は、20年、30年前であれば、最も遠縁な関係ではなかったかと思われる生物学と情報学という 2 分野のカップリングで、ここ10年から15年の間に、非常に早いスピードで発展してきた分野です。今日はその経緯をご紹介したいと思っております。

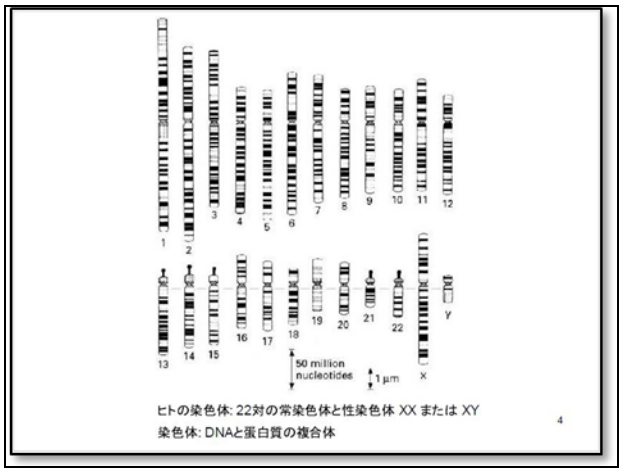
生物情報学の目的は、一口でいいますと、遺伝情報の総体であるゲノムにコードされている生物情報を、情報学の手法で読み解くこと、となるかと思います。生物学と情報学の境界領域という都合上、情報学の知識だけでなく、生物学の知識も必要とする分野となっております。



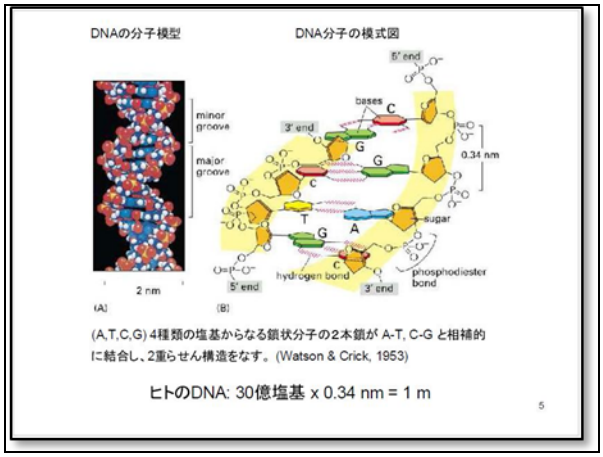
2. 遺伝情報の爆発的増加

右図は動物細胞の写真です。ここに核膜に包まれた核がありまして、この中に生物の持っている全遺伝情報が含まれています。（スライド No. 3）

核内にある遺伝情報の源が、染色体と呼ばれる物質で、その模式図が右下の図です。縞模様が見えますが、実際の顕微鏡下ではもっとぐちゃぐちゃとした格好をしておりますが、大きさの比較ができるように分かりやすく綺麗に並べており

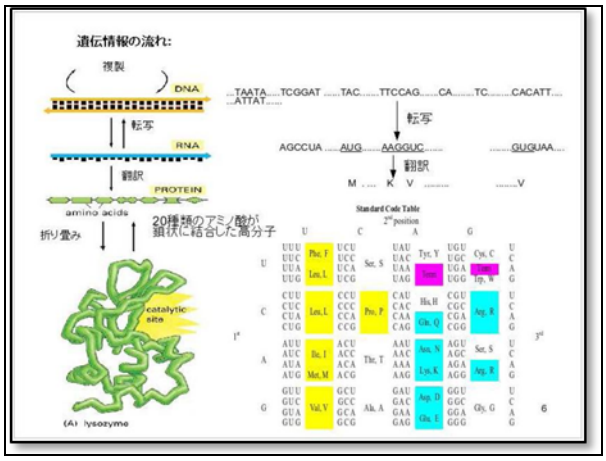


ます。ヒトの場合ですと、ここに示されているような22対の常染色体と、２本の性染色体からなります。性染色体は、Ｘが２本の場合は女性、ＸとＹが１本づつあると男性です。（スライドNo.4）



染色体は、遺伝情報を持ったDNAとタンパク質の複合体で、そのDNA分子の化学構造を示したのが右図です。DNAは、４種類の塩基、Ａ、Ｔ、Ｃ、Ｇという文字で表しますが、その４種類の塩基が鎖状に結合し、それが２本互い違いになって組み合わされ、その間でＡにはＴが、ＧにはＣが互いに相補的に結合して、２重らせん構造をとっています。

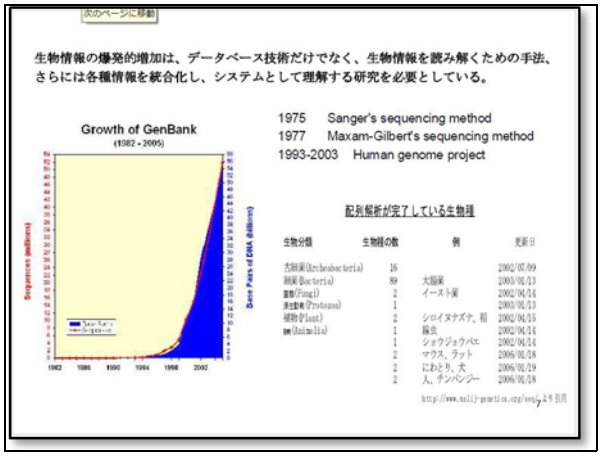
図はその分子構造を示したものですが、赤いところは骨格部分で、その間のブルーのところは塩基対に相当します。塩基間の距離は、約0.34 nmです。（1 nmは10^9 mです。）ヒトのDNAは総延長30億塩基ありますから、全長は30億×0.34nmとなり、真直ぐ伸ばしますと約１メートルに相当します。一つの細胞にはこの量の２倍含まれますので、DNAの総量といたしましては、ヒト１個体当たり、細胞の数は不確定ですが、DNAの総延長は、地球から太陽までの距離の800倍以上に相当します。遺伝情報は、この４種類のATCGの順序にコードされています。（スライドNo.5）



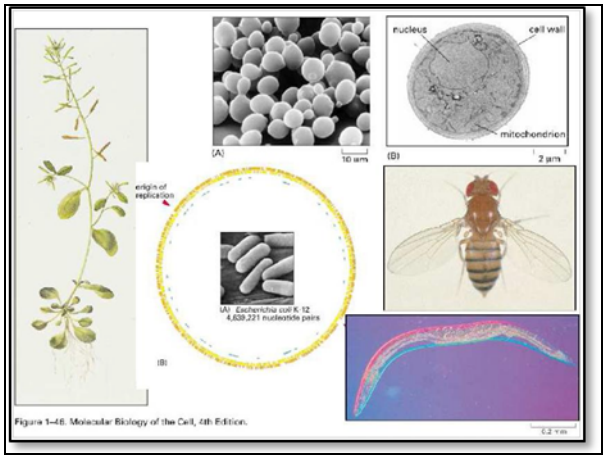
DNAはこのように２本鎖ですが、そのDNAは一旦RNAと呼ばれる１本鎖の、DNAによく似た分子に転写されます。ただ、RNAがDNAと違う点は、１本鎖で存在することの他に、少しだけ化学構造が違い、特に塩基はATCGのＴの代わりにＵという分子が使われています。DNAの２本鎖のうちの１本鎖を鋳型（いがた）とし、DNAのＴにはＡが、ＣにはＧが対応するように転写され、RNAが合成されます。そのRNAは、３塩基（コドン）が一つのアミノ酸を指令しアミノ酸配列（タンパク質）に翻訳されます。４種類の塩基がございますので、３塩基からなる64種類のコドンが、20種類のアミノ酸に対応します。上図にコドン表と名付けられているその対応表を示します。アミノ酸は、３文字又は１文字で略しますが、ここでは長い配列を示す都合上、１文字コードで示すことにします。アミノ酸の正確な名前はここでは必要ありませんので、これがアミノ酸の名前だと思っていただいて結構です。

タンパク質は、地球上の生物では、20種類のアミノ酸が同じく鎖状に結合した高分子で、

生理的条件下では固有の立体的な構造に折り畳まれ機能を果たしています。このコドン表の中で、ピンク色で示したものは特殊なコドンで、これらのコドンはアミノ酸配列の終了を意味し、ストップコドンとか、ターミネーションコドンと呼ばれています。（スライドNo.6）

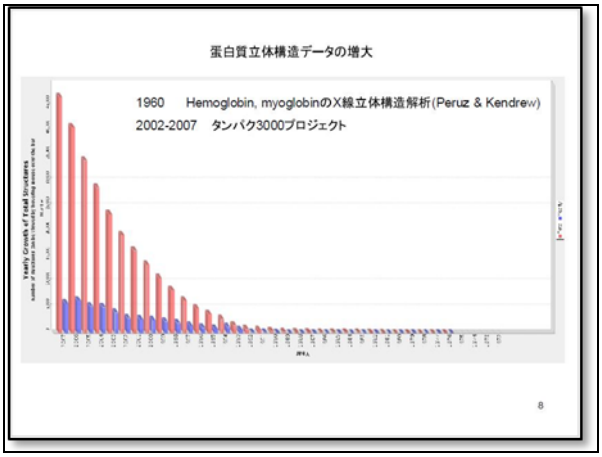


生物情報学は、ここ10年から15年で甚だしく進展したと申し上げましたが、その背景になっているのが、分子生物学における解析技術の進歩です。グラフは、各年度までに解析されたDNA配列の総延長を縦軸に示しています。2005年までに解析された総延長は、約60ビリオン、ビリオンというのは10の9乗ですが、人間のDNAの長さは3×10^9で、約3ビリオンですので、人間のDNAの約20倍程度のDNAが2005年までに解析されたというわけです。さらにこの曲線は、指数関数以上の急激な増加を示しています。このグラフは、生物種によらず解析された配列の全長を示したものですが、ヒトもチンパンジーも既に解析されています。すべて解析されたと言ってもDNAの塩基配列だけですが、ここに書いてあるように、鶏や犬も完了しました。稲は日本が主導して解析を終了いたしました。最近では、病原菌に相当する細菌も多数解析され、細菌も含めると今(2009/09)や1000種類ほどの生物種において全塩基配列の解析が終了しています。（スライドNo.7）



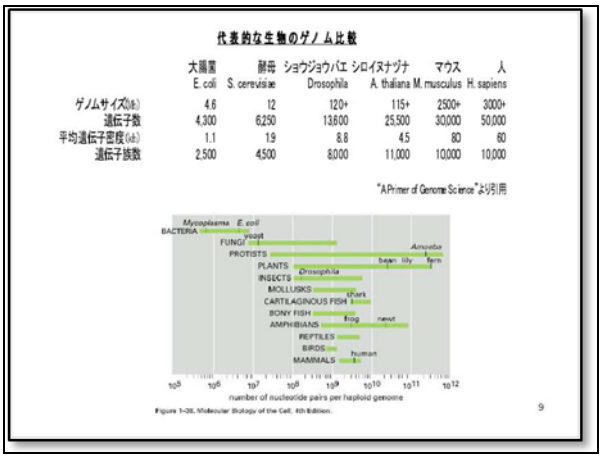
ここで、代表的な生物の写真をお見せしますと、まず図の真中の写真が大腸菌です。大腸菌は、先ほどお見せした動物細胞とは異なり、核を持っておりませんし、DNAも環状です。図中で次に下等なのがイースト菌（右上）です。イースト菌は大腸菌と同じく細胞1個から成る単細胞生物ですが、大腸菌とは違って核膜を持っていて、大腸菌よりはずっと人間に近い生物です。核を持っていない生物を原核生物といい、核を持つ生物を真核生物と名付けています。ここに示した中で次に高等な生物は多細胞の線虫（右下）で、レタスなどについている数ミリの大きさの生物です。遺伝学/発生学的に非常に詳しく研究されています。それからショウジョウバエの写真が図（右中）にあります。数ミリの大きさで、果物によく付着していますが、この生物も遺伝学的に昔から詳細に研究されています。また、植物を代表してシロイヌナズナが図左に示されています。（スライドNo.8）

先ほど、解析された DNA 塩基配列の総延長をご紹介しましたが、他にも多数の遺伝情報が解析されています。たとえば、 DNA にコードされているタンパク質の立体構造における原子 1 つ 1 つの位置座標は X 線を使って解析することが可能ですが、右図は既に解析されている立体構造データの増え方を示しています。ブルーのところは、各年に解析を終了したタンパク質の数です。赤は累積です。2006年度では約6000、 1 日当たり約20近いタンパク質が解析されています。（スライドNo. 9）



現在では、このような解析データは、データベースとして維持され、すべてインターネット上にリリースされています。塩基配列、タンパク質配列、タンパク質の立体構造、さらには生物種ごとのデータベース、それ以外にも多数ございまして今や恐ろしいほどの数のデータベースが作成されています。そのため、データベースのデータベースがありますし、著名なポータルサイトがアメリカ、ヨーロッパ、日本等で、国の予算を得て維持されています。また、各種の解析ツールが同じくアメリカ、ヨーロッパ、日本で提供されています。今では、生物関連学科では情報学の科目は必修です。その理由は、インターネット上の検索 / 解析ツールを使わざるを得ない状況になっているからです。一方、大量のデータを管理するデータベース技術だけでなく、膨大なデータの中から意味ある情報を得るための情報学的な手法も必要とされています。では、生物情報学的な研究要請としてどのような解析が必要とされるようになったか少しお話ししたいと思います。

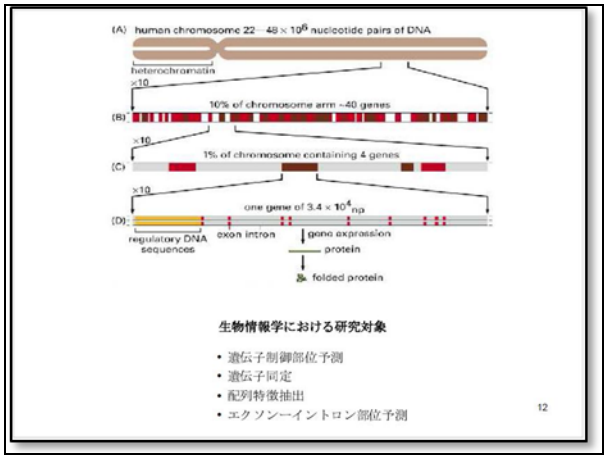
生物における全遺伝情報の一組をゲノムと言いますが、そのゲノムのサイズを、 DNA の長さ（塩基数）で見ますと、先ほど述べましたように、ヒトは3×10^9、一方で大腸菌は5×10^6です。右図で示したサイズの単位は、10の 6 乗、メガベース(Mb)です。概ね高等なほど長いのですが、生物種によっては幅広く分布しています。たとえば、同じ昆虫でも、最小と最大では 100 倍ほどの違いがあります。ヒトよりも長い DNA をもった生物も多々存在します。



一方、興味深いのは、 DNA がコードしているタンパク質の数です。それを遺伝子の数で見ますと、大腸菌は 5,000 ぐらいなのに対し、ヒトはその10倍の 5 万、マウスで 3 万ぐらいです。ここですぐ気付くのは、大腸菌とヒトの遺伝子の数がたった10倍なのに、なぜこれだけの違いがあるのか、その違いはどんなところに反映されるのかということですが、

これに関しては、後で触れることにします。

さて、右上の表には、平均の遺伝子密度の逆数が示してあるのですが、ヒトの遺伝子密度は、大腸菌に比べ約60分の１、つまり、遺伝子の数に比して、塩基配列が非常に長いということです。一遺伝子に対応するDNAの長さは、大腸菌でもヒトでもそれほど違いませんので、この遺伝子密度の差は一重に遺伝子ではない領域がヒトでは非常に長いということを意味します。（スライドNo.11）

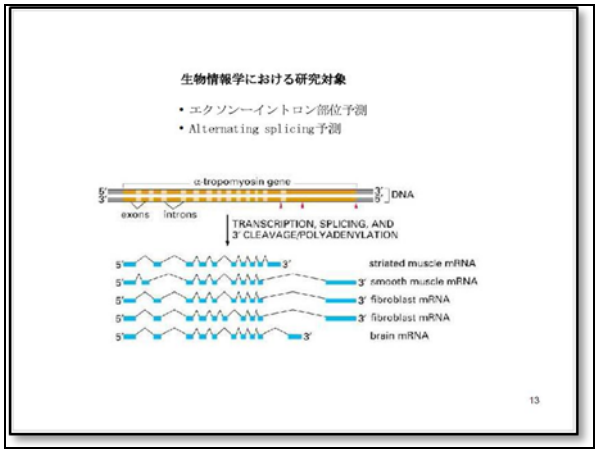


ヒトの場合ですと、DNAの長さの約４％が遺伝子をコードした部分で、それ以外の90%以上の部分は、遺伝子をコードしておりません。ここでいう遺伝子というのは、タンパク質をコードしている部分だと思って下さい。右図上に人間の22番目の染色体が描いてありまして、その染色体の１％に相当する部分を拡大したのがその下の縞模様の部分です。黒い部位は、実験的に確認された遺伝子の部位です。約1%のDNA上に４個の遺伝子が平均的に含まれますが、赤の部位が、統計的な性質を使って計算機プログラムによって予測した遺伝子の部位です。ゲノムの長さを考えると、人間が手作業で遺伝子の所在を同定するのはとうてい無理です。当然ながら、ある種のアルゴリズムに基づいて、計算機プログラムで処理すべき問題です。このように多量のデータは、情報学的な処理を必要としているというわけです。

遺伝子の同定は、各種の生物のDNAが解析されるようになってきますとルーチンワーク的な作業になりますが、生物情報学的な技法が必要です。さて、これらの遺伝子部位の全領域がタンパク質をコードしているわけではありません。タンパク質をコードしている領域は、図中最下部に示したDNA上の赤く塗った部分で、この領域をエクソンといいます。灰色の部分はイントロンと言い、タンパク質をコードしておりません。このように真核生物では遺伝子領域はエクソンとイントロンから構成されています。よって、エクソンとイントロン部位の同定／予測が必要となります。これもまさに生物情報学的な解析です。エクソンとイントロン境界予測の簡単なモデルを後でご紹介します。

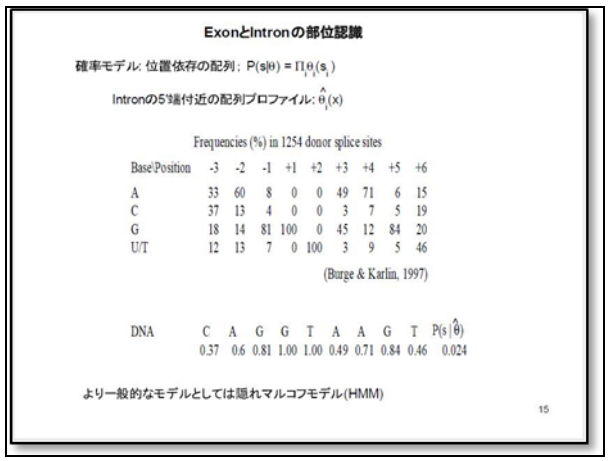
また、遺伝子領域のDNAはRNAに転写されるわけですが、その転写を制御している配列が前の方（上流）にございます。そのような制御部位の予測も必須となります。通常、一定の統計的なルールに基づいて予測しています。（スライドNo.12）

次の図に示すDNA領域は、先ほど示しました遺伝子領域を拡大したものです。遺伝子領域で転写されたRNAは、その後に処理を受けまして、イントロンの部分が切り取られ、エクソン部分だけを繋いだ構造になります。それをスプライシングといいます。このエクソンの繋ぎ方は、組織によって異なる場合があります。つまり、1つの遺伝子領域から、組織により異なるRNAが生成されタンパク質が合成される場合があります。したがって、1つの遺伝子から複数のタンパク質が合成される場合がありまして、先ほどヒトの場合5万の遺伝子と言いましたが、実は5万の遺伝子は、それよりはるかに多種のタンパク質に翻訳されていると考えられます。また、このような異なるエクソンの組合せ（アルタネイテブスプライシング）を予測する必要もありますが、今のところ難解で、ほとんど研究が進展しておりません。（スライドNo.13）

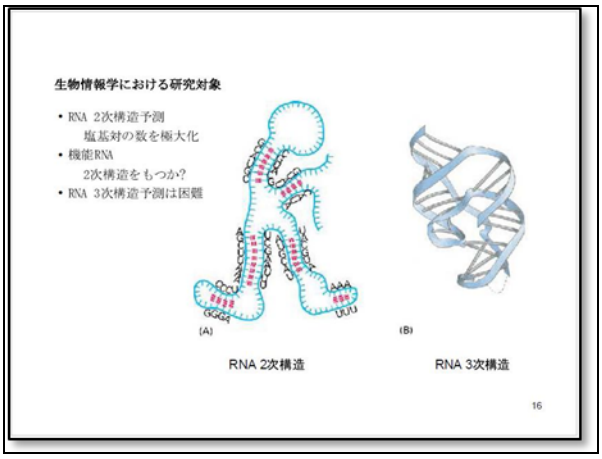


一つの例でございますが、イントロンとエクソンの境界を予測するための簡単な方法をご説明します。

エクソンとイントロンの部位は、簡単なルールである程度予測することが可能です。境界近傍での塩基の頻度分布を調べてみますと、右側がイントロンで左側がエクソンとしますと、イントロンの左端の部位には、常にGという塩基で、その次がU(DNAの場合はT)という塩基が出現します。同様に、それほどは明確ではありませんが、その前後で一定の頻度を持って塩基が並んでいるということが分かります。そういう統計的な知識を使いまして、たとえば、最も単純なモデルとしましては、あるDNA配列が与えられたときに、各塩基が一定の部位に見出される頻度を、Cの場合でしたら37%ですから0.37とし、Aの場合です0.60としまして、すべて部位の頻度を掛け合わせた数が図で示した配列の場合は0.024となりますが、この数が大きければ大きいほど、この配列がこのイントロンとエクソン境界に位置しているであろう確率は高まるわけです。それ故、この値が一定の値を超えた時に、エクソンとイントロンの境界の部位であるとして予測をするというのがこの方法です。実際にはもう少し複雑なモデルを使うわけですが、原理的にはこのような方法で境界を予測いたします。（スライドNo.15）

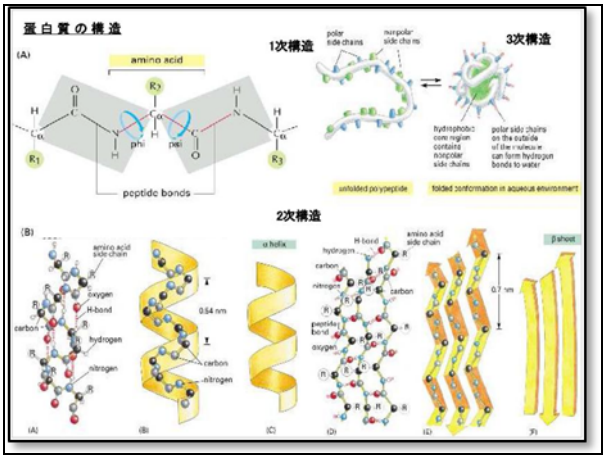


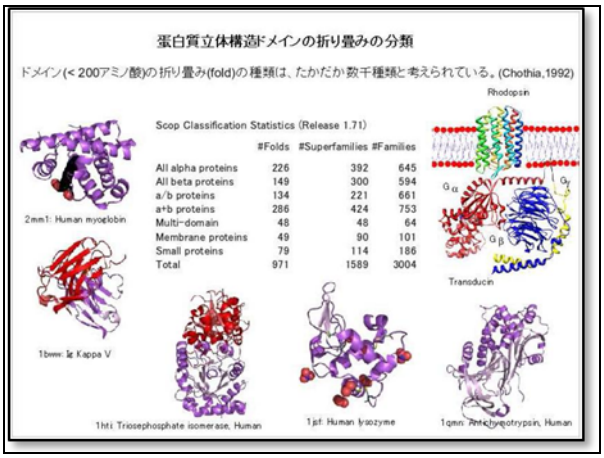
先ほどゲノム DNA はその4%しか遺伝子をコードしていないと申しました。では残りの 96% は何か役割を果たしているのでしょうか。かれこれ 5 年程前までは、多分ジャンクだろうと考えられていました。しかしその後、他の 96% の領域でも遺伝子の発現の制御に関わっている部位も多々あるという実験的な証拠がいくつか出て参りまして、今では発現制御に関わる RNA のことを機能 RNA と称しますが、



機能 RNA の予測が関心を集めております。機能 RNA はひょっとしたら、特異な 2 次構造を形成しているかもしれないということが言われております。 RNA は一本鎖で存在しますが、 DNA と同様に、相補的に G には C 、 A には U が結合して塩基対を成すことができます。この塩基対構造のことを 2 次構造といいます。 2 次構造が機能 RNA に関係あるか興味の持たれるところです。 2 次構造は塩基対が多ければ多いほど安定と考えられますので、塩基対の数が最大であるような構造として予測します。後で紹介する動的計画法という手法で解くことが可能です。実際の RNA は、さらに折り畳まれて 3 次構造を成しています。図でリボンのところは骨格部分、そして塩基対のところは棒状に示された部分です。 3 次構造の予測は、非常に難しくてほとんど手がつけられておりませんが、 RNA の安定性には、 2 次構造以上に 3 次構造が影響していると思われます。（スライドNo.16）

さて、 RNA の指令に基づき生成されるタンパク質は、20種類のアミノ酸からなる鎖状高分子で非常に軟らかい分子です。右図はその化学構造ですが、20種類のアミノ酸の違いは、このR_1と書いてある部位です。すべてのアミノ酸に共通な骨格構造においては、N-Cαと Cα-Cの化学結合軸がほぼ自由に回転できます。その結果、タンパク質は非常に特異な 2 次構造をとることができ、その 1 つは右図に示したらせん構造で、αヘリックス (helix) と言います。もう 1 つはβシート (sheet) といわれるシート状の構造です。こういうビルディングブロックがさらに折り畳まれ 3 次構造を成しています。アミノ酸は大きく 2 種類に分けられまして、 1 つは水を好む親水性アミノ酸、もう 1 つは水を嫌う疎水性アミノ酸です。タンパク質は、疎水性アミノ酸を内側に、表面には親水性アミノ酸が配置されるように折り畳まれます。（スライドNo.17）

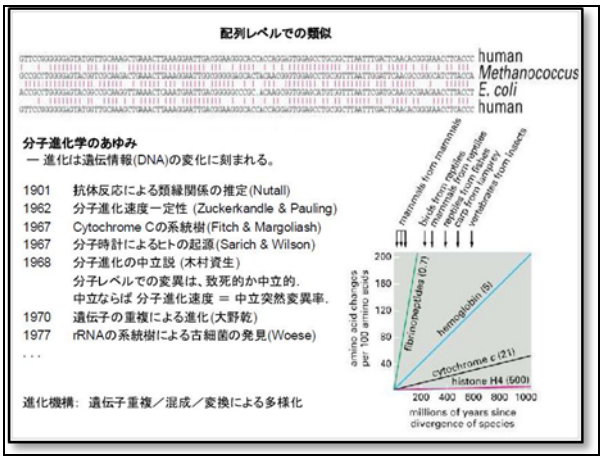




非常に多様な種類のタンパク質立体構造が可能ですが、地球上の生物を構成するタンパク質ではフォールド(fold)(タンパク質の立体構造の折り畳みの形状)は高々数千種類だと考えられています。それは、地球に生命が誕生したときの祖先タンパク質が数千種類であったということの反映であろうと考えられています。高々数千種類なら、それらをすべて解析してしまおうとするプロジェクトが近年組織されまして、日本でも過去5年間で約3千のタンパク質を解析するというプロジェクトが実施されました。現在までのところ判明しているフォールドの分類を、右図に示しておきます。スーパーファミリーという分類がありますが、同じスーパーファミリーに属するタンパク質はすべて同じ祖先に由来すると考えています。これまでにフォールドの比較、分類から約1600のスーパーファミリーが判明しています。(スライドNo.18)

3. 生物情報学的知見

さて、ここまで、DNAの生物情報学的な解析を中心に述べてきましたが、以下では、そのような解析からどういう意味ある知見が得られるのかということをお話ししたいと思います。



多種の生物のDNAの塩基配列を比較すると、生物の種間で非常に似ている部分配列があることに気付きます。右図は、上下にヒト、その間に上からメタン生成古細菌、大腸菌(E. coli)のDNAの部分塩基配列を示します。古細菌は、極限環境、たとえば深海ですとか温泉ですとか、高温、高圧、高酸のところに住んでいる細菌です。ここで、赤い線で結んでいる部位は、同じ塩基が出現している部位です。これを見るとヒトと大腸菌のDNAでこんなにも似ている部位が存在するということに驚かれると思います。この2配列の差異に関して、たとえば2つの配列で異なっている塩基の比率が生物種とどのような関係を示すは興味深い問題です。この類の解析は、1962年に非常に注目を集めました。その当時はまだ塩基配列が解析されていませんで、利用できたのはタンパク質配列のデータです。様々な生物種からの同種のタンパク質のアミノ酸配列を相互に比較し、その差異を縦軸に、横軸にそれら2つの生物種が今から何年前に分岐したかを記したのが図の右下のグラフです。

単位は百万年ですので、注意していただきたいのですが、面白いことに、タンパク質の2配列の差異は、それらの生物種の分岐時間にほぼ比例するということが分かりました。つまりこれは、分子進化速度と言いますが、塩基やアミノ酸の置換速度が一定であるということを意味します。この分子進化速度は、タンパク質ごとで異なります。その理由ももちろん分かっておりますが、ここで問題にしたいのは分子進化速度が一定であるということです。これが何を意味するかと言いますと、配列の比較から得られた差異を時計として使用することができる、つまり、タンパク質に固有の進化速度が分かれば配列間の差異から生物種が分岐した年代が推定できるということです。この性質を分子時計と呼んでおりますが、その発見以降、塩基/アミノ酸配列の比較に基づく進化の研究が可能になりました。

また、重要な発見が日本人によってなされました。木村資生先生の提唱によるもので、このような分子レベルでの変異は、ほとんどは致死的で、致死的というのはこの変異が起こりますとほとんどが死滅してしまうという意味ですが、生き残った変異は例外はあるもののほとんどの変異は淘汰に関して中立的であるという説です。優位なものが生き残ると主張するダーウィンが提唱した進化の淘汰説をご存じの皆様は、淘汰説に反していると思われるかと思います。現在、形質レベルの変異に関してはダーウィンの淘汰説、分子レベルでは木村先生の中立説が正しいと考えられております。この中立説は分子進化速度一定性を基礎づける理論となっております。さて、多分皆さんが疑問に持たれるのは、変異が中立なら多種類の有用な遺伝子がなぜ生じたのかということでしょう。それは遺伝子重複というメカニズムです。遺伝子が2つにコピーされますと、そのうちの1つに変異が起こっても致死変異にはなりません。2つのうち一方の損なわれていない完全な遺伝子が機能することによって生存できるからです。そういうメカニズムによって、他方の遺伝子は多数の変異を蓄積することができ、それがある程度の時間が経過した後に異なる機能を持った遺伝子として発現するというのが、遺伝子重複による進化のメカニズムです。（スライドNo.34）

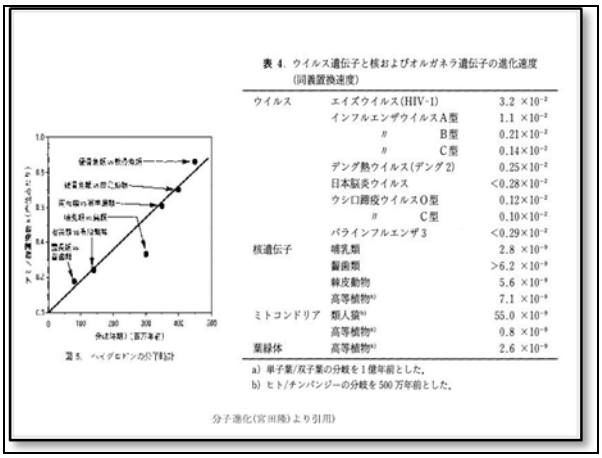


表 4. ウイルス遺伝子と核およびオルガネラ遺伝子の進化速度
(同義置換速度)

ウイルス	エイズウイルス(HIV-1)	3.2×10^{-3}
	インフルエンザウイルスA型	1.1×10^{-3}
	〃 B型	0.21×10^{-3}
	〃 C型	0.14×10^{-3}
	デング熱ウイルス(デング2)	0.25×10^{-3}
	日本脳炎ウイルス	$<0.28 \times 10^{-3}$
	ウシ口蹄疫ウイルスO型	0.12×10^{-3}
	〃 C型	0.10×10^{-3}
	パラインフルエンザ3	$<0.29 \times 10^{-3}$
核遺伝子	哺乳類	2.8×10^{-9}
	鳥類	$>6.2 \times 10^{-9}$
	爬虫動物	5.6×10^{-9}
	高等植物[a]	7.1×10^{-9}
ミトコンドリア	類人猿[b]	55.0×10^{-9}
	高等植物[a]	0.8×10^{-9}
葉緑体	高等植物[a]	2.6×10^{-9}

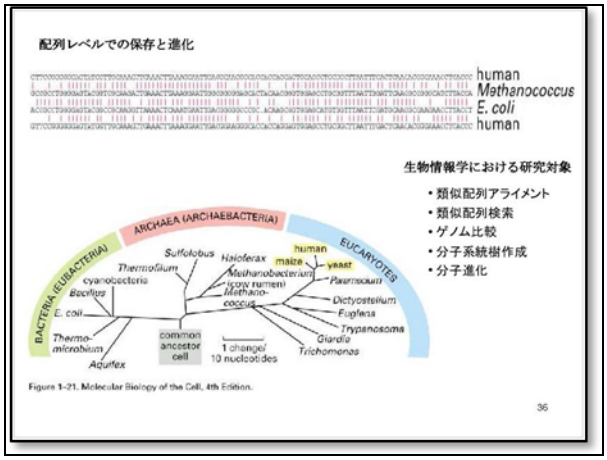
a) 単子葉/双子葉の分岐を1億年前とした。
b) ヒト/チンパンジーの分岐を500万年前とした。

分子進化(宮田隆)より引用

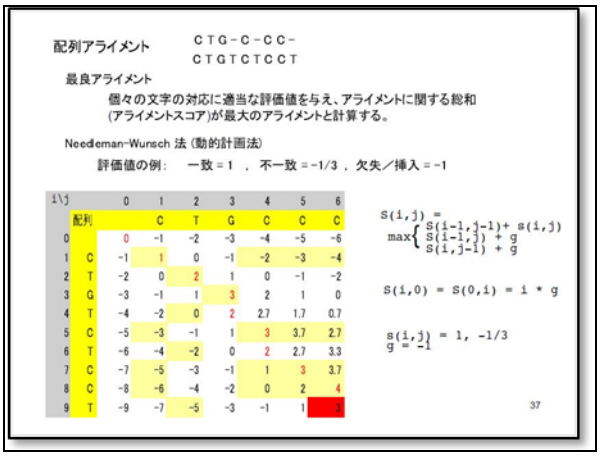
先ほど進化速度はタンパク質ごとで違うと申しましたが、さらに生物種でも随分違います。たとえば、エイズウイルスやインフルエンザウイルスは、私ども人間の約 10^7 倍、進化速度が速くなってございます。つまり、エイズウイルスは、DNA配列に高頻度で変異がおこりカメレオンのごとく自らが変化します。そのため、エイズウイルスに対するワクチンを作ることが非常に困難になっています。言い替えると、ホストの免疫系からの攻撃を避けるために遺伝子を変化させるべく進化速度を高めるという機能を獲得してきたと言えるかもしれません。（スライドNo.35）

さて、塩基/アミノ酸配列を比較することによって進化的な知見が得られるわけですが、

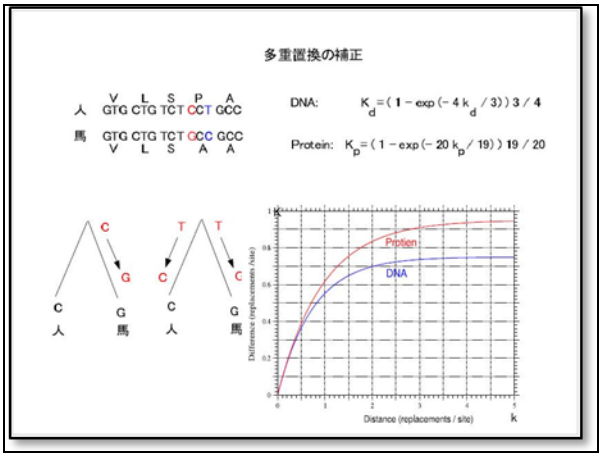
現在、多くの場面で情報学的な手法が必要とされています。たとえば、類似している配列を並べるアルゴリズム、塩基/アミノ酸配列のデータベース上で類似している配列を検索する技法、ゲノムは人間の場合ですと 10^9 もありますので、そういった長さのものを比較するための技術、またそうした比較に基づいて系統樹を作成するアルゴリズム、そしてさらに、得られた系統樹に基づく進化の解析が必要になってくるわけです。（スライドNo.36）



右図に、2つの配列を並べるためのアルゴリズムの最も簡単な例を示します。2つのDNA配列を比較した例です。一定の配列の塩基/アミノ酸をアライメントと定義します。個々の塩基/アミノ酸の対応に、適当な評価値を与え、アライメントに関する総和が最大なアライメントを最良アライメントと定義いたします。評価値の例としては、たとえば、塩基が一致したときは1、不一致のときは-1/3、そして、片方で塩基が挿入され、片方で欠失が起こっている場合には、-1という評価値を与えますと、可能な限り塩基が一致しているようなアライメントが最良アライメントとなります。最良アライメントは動的計画法というアルゴリズムを使って計算することができます。この方法は、物理学における磁石の理論で使われている手法とまったく同じです。（スライドNo.37）

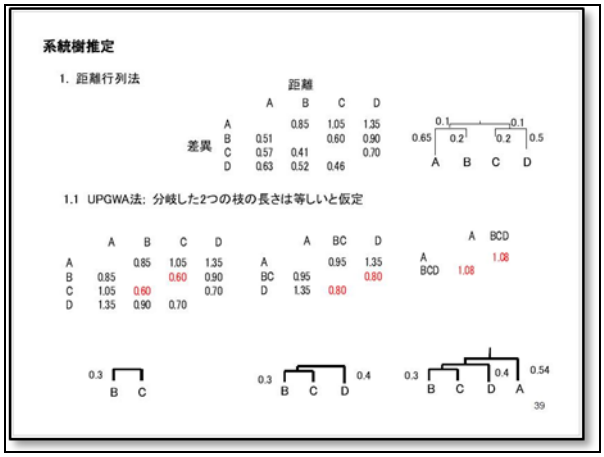


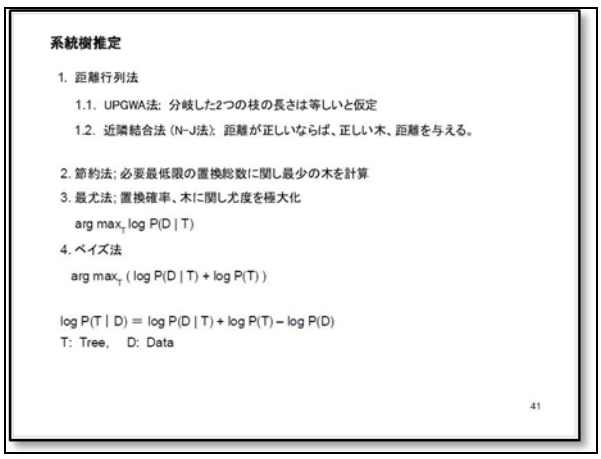
さて、2つの配列があって、ある部位でCとGの違いがあるという場合に、系統樹の一方の枝において祖先配列におけるCがGに変異して現在の配列に至ったのか、実は祖先はTで各々の枝でTからC、TらGに置換が生じたのか不明です。この場合前者では　1回置換が起こり、後者では2回置換が生じていますが、両方とも確率的に可能ですので、確率的にそのような多重置換の補正を加え、2配列の差異からその部位で生じたであろう置換数の期待値を推定します。右図では、横軸に差異が示され、その差異に対応する置換



数の期待値が縦軸に示されています。置換数は0から無限大ですが差異は0から1までの値を取りますので、当然この関数は図のような曲線になります。タンパク質の場合は赤い線で、DNAの場合はブルーの線で、差異から実際に生じたであろう置換の数が推定できます。（スライドNo.38）

差異から計算された置換数（進化距離）からなる行列を距離行列と申しますが、右図では行列の下半分に4種類の配列を互いに比較した差異を示します。この差異から先ほどの関係式を用いまして計算された置換数の期待値が行列の上半分に示されています。このような距離行列から系統樹を推定する方法の中で最も簡単な方法を右図に示してます。この方法は、階層的クラスタリングの類に属し、昔から分類学の分野で用いられている手法です。距離だけを示した行列が図の下部に書いてありますが、まず最初にこの行列で数値の最も小さな要素、つまり最も類似した配列対を選びます。この例では、BとCが最も類似していますので、まずその2つを束ねます。次にこのBCのクラスタと他の配列との距離を計算します。ここでBCのクラスタとAとの距離は、A,Bの距離と、A,Cの距離の平均値として0.95とします。次に、次元が一次元小さいAとBCとDからなる距離行列を作成し、再度、行列の要素中で最も小さな要素0.80を選びます。この例ではBCとDとの距離ですので、BCのクラスタにDを加えます。さらに、BCDとAとの距離を同じくA,B、A,C、A,Dの距離の平均値として計算し、図で示したようなA,B,C,Dの系統樹を推定いたします。この方法では、根元から各配列までの距離は、すべて等しいと仮定いたします。この仮定はいわば進化速度が一定であるという仮定ですから、先ほどの分子時計の仮定に相当いたします。ただ、分子時計の仮定が成立しない場合もございますので、望ましい方法としては、このような仮定をしない方がよろしいわけです。たとえば、図の右上に示した系統樹では各枝で分子速度が大きく異なります。この系統樹に対して先の方法を適用しますと眞の系統樹とは異なる図右下に示した系統樹が得られます。（スライドNo.39）

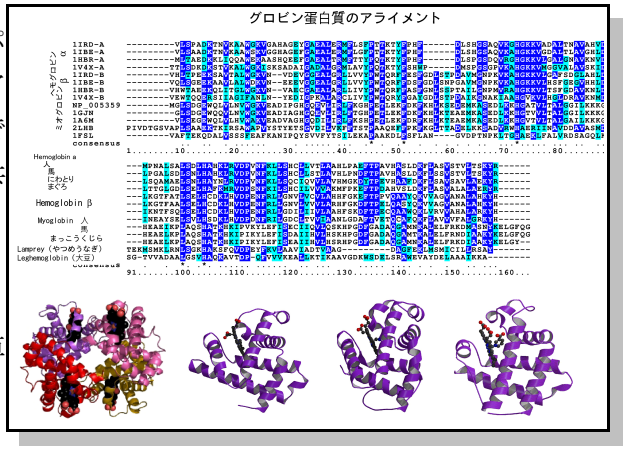




実は、距離が正しければ眞の系統樹を与える近隣結合法という方法があります。さらには系統樹に最小限必要な置換数を最小とするような系統樹を計算する最大節約法という方法もあります。また、最尤法により配列アライメントが得られる確率を最大にするような系統樹を計算す

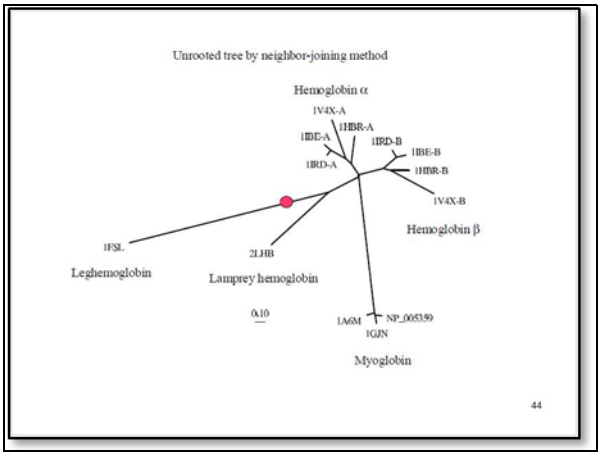
る方法もあります。（スライドNo.41）

右図は、グロビンタンパク質スーパーファミリーの代表的タンパク質のアライメントを示しています。この図では、示した生物種のタンパク質間で共通なアミノ酸を青で示しています。（スライドNo.42）

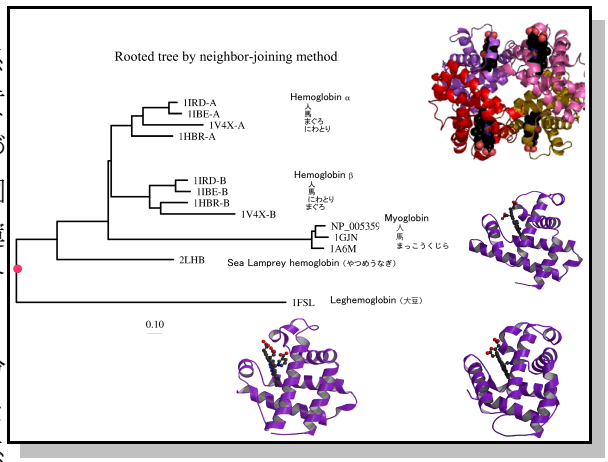


このアライメントから、差異を計算し、その差異から距離を計算します。

得られた距離から、近隣結合法を用い計算された系統樹が右図です。近隣結合法ですと、分子時計は仮定いたしませんので、どこに根が位置するかは不明です。根の位置は他の知見から推定いたします。ここで比較した配列の中で大豆は、他配列との差異は0.8以上で、1アミノ酸当たり約3回置換が生じているような非常に遠縁なタンパク質です。大豆はこれらの配列中では唯一植物で、他は全て動物です。よって、生物分類上の知見から、根は必ず大豆への枝の上にあるということが自明です。根のところを持って持ち上げますと、有根系統樹が得られます。（スライドNo.44）



右図に示されたこの系統樹では、分子時計の仮定は成立いたしません。これには理由がございます。この図に示した系統樹は、ヘモグロビンα、ヘモグロビンβ、ミオグロビンの各ファミリーの系統樹を部分系統樹として含みます。また、各ファミリーの系統樹で一番上の配列がヒトのタンパク質です。ヘモグロビンαとβは、2分子づつ4分子が結合し大きなタンパク質（図右上）を構成し、酸素を体の末端まで運搬します。皆さんの血液中の赤血球に含まれるタンパク質です。ヒト、馬、鶏、マグロのヘモグロビンαとβ 2つの部分系統樹ですが、マグロと鶏の位置がαとβで違っています。αとβ 2つの部分系統樹で樹形が異なりますので矛盾した結果となっています。もっと多数のタンパク質を用

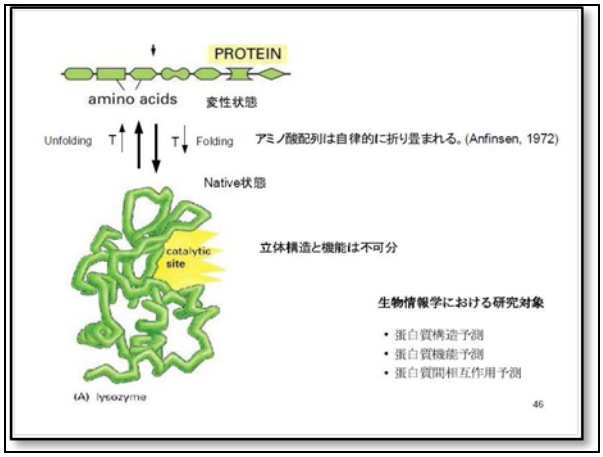


いると、生物種の分岐を反映した正しい系統樹が得られるはずです。グロビン族では何度か遺伝子の重複が生じタンパク質の分化が起こっています。祖先タンパク質が重複を起こしミオグロビンとαとβの祖先に分化し、αとβの祖先はまた重複を起こしαとβに分化しました。ミオグロビンとαとβの祖先への分化はヒトの祖先とマグロの祖先との分岐以前に生じ、その結果、ヒトはミオグロビン、αとβの遺伝子をすべて所持しています。このように、配列の分岐は、生物種の分岐に伴うばかりではなく遺伝子重複によっても生じます。配列比較から、生物のゲノムはダイナミックに変化していることが明らかになっています。

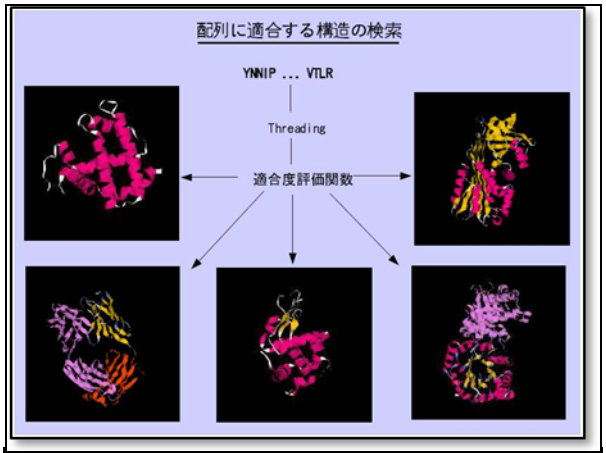
塩基/アミノ酸が置換されますとほとんどの場合はタンパク質の立体構造が維持されずタンパク質はその機能を果たすことができなくなります。そのため突然変移のほとんどは生物にとって致死的です。ところが、遺伝子が重複されますと、その一つの遺伝子上の突然変移は他が機能すれば致死的にはならず、変異を蓄積することができます。その結果、遺伝子は変化し遺伝子の分化が生じます。このような状況下では、進化速度は一定ではなく変化します。よって、タンパク質の分化が生じているここで示したような例では、分子進化速度は一定ではありません。

さて、大豆の場合、タンパク質配列は他と 80%以上も違っています。ところが、立体構造で見ますと非常に良く似ています（図左下）。このように、タンパク質が機能を果たす上で不可分な立体構造は配列より保存されがちであることが知られています。（スライドNo.45）

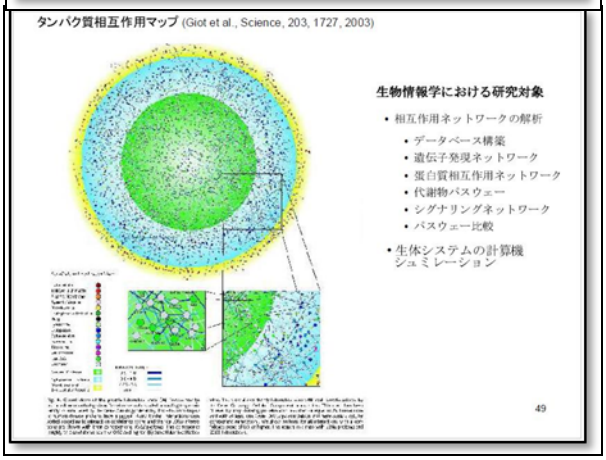
タンパク質が畳まれる立体構造は、温度を上げますと壊れて鎖状になり、温度を下げますとまた各タンパク質に固有の立体構造を形づくります。この事実は、立体構造を形づくる情報はすべてアミノ酸の配列順序にコードされていることを示します。よって原理的には、配列から構造の予測が可能であるはずです。また、タンパク質の機能は構造と不可分ですので、構造に基づいた機能の予測が望まれています。（スライドNo.46）



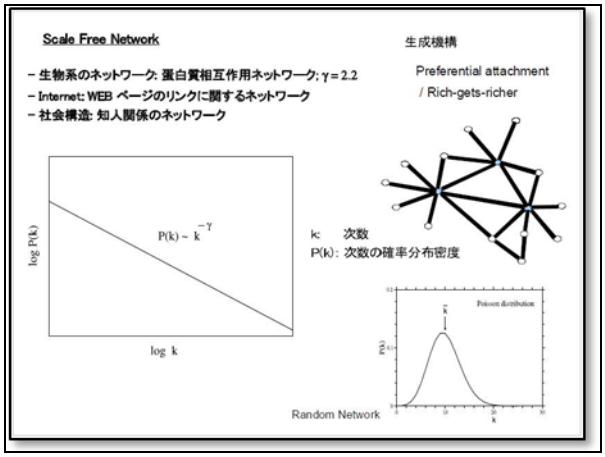
配列の解析は簡単である一方、立体構造解析は非常に手間のかかる作業です。配列からの構造予測が必要とされる由縁です。また、その一環として、既に多数の構造が解析されていますので、データベース中のタンパク質構造の中から構造未知の配列に適合する構造をピックアップして提示してくれるような高速な検索システムも必要となっています。（スライドNo.47）



さて生体では、多数の遺伝子及びタンパク質が相互作用して機能を司っているわけですが、近頃では解析技術の進歩により、様々な細胞における遺伝子発現とその時間変化、相互作用している遺伝子及びタンパク質を決定することができるようになりました。右図では、相互作用している要素を線で結んでいます。遺伝子／タンパク質間の相互作用を通じて生体をシステムとして理解しようとする研究が、近年盛んになりつつあります。（スライドNo.49）



このような相互作用のネットワークは非常に面白い性質をもっています。たとえば一つのタンパク質が相互作用する相手の数をKとしますと、グラフ理論の方では次数と言いますが、その頻度分布は、$K^{-\gamma}$ という関数形をしています。この γ は2から3の間の数ですが、タンパク質の相互作用の場合、約2.2です。このような性質を持つネットワークを、一般にスケールフリーネットワークと言います。一方、一定の確率でランダムに相互作用した場合には、その頻度分布は図で示したようなベル型の形となります。



このようなスケールフリーネットワークは、生物系ネットワークだけではありません。皆さんが使っているWebページのリンクからなるネットワークもこのような性質をもっています。さらには社会構造における知人関係のネットワークなどもまさにスケールフリーネットワークですし、共著者間を結んだネットワークもこのスケールフリーネットワークになっています。もちろん、遺伝子／タンパク質の相互作用ネットワークが何故そのよう

な構造（スケールフリーネットワーク）になるのか面白い問題です。（スライドNo.50）

4. おわりに

生物情報学という分野の成り立ちと、研究の現状をお話しさせていただきました。

また、この分野の将来像としては、各種の情報を統合化し、システムとして生物を理解しようとする方法へ発展しつつあるということでございます。

長らくご静聴どうも有難うございました。

【引用文献】

画像の多くは、Molecular Biology of the Cell, V.4から引用
分子描画は、 MolScript を用いた。

（当日の記録を基に再構成しました。なお、紙面の都合により、一部のスライドの説明部分を省略させていただいたことをご了承下さい。）