〔教育講演〕

電気泳動, ゲノム, プロテオーム

村 松 正 實

SUMMARY

The development of modern molecular biology strongly depends on the progress of “electrophoresis” techniques. Since Tiselius, the resolution of electrophoretic separation of proteins had been elevated dramatically, especially after the advent of SDS-polyacrylamide gel by Laemmli, until it became two-dimensional by O’Farrell. On the other hand, nucleic acids became also the target of this technique ended up with the establishment of DNA-sequencing gel by Maxam-Gilbert and Sanger. This technique aided by gene cloning technology has determined a great number of genes with respect to the nucleotide sequence. This also led to the elucidation of the relationship between the structure and the function of genes in eukaryotes as well as prokaryotes. The emerging issue was then to determine the whole genomic as well as mRNA (cDNA) sequences of a species including human in order to understand the riddle of life on the molecular level. The Human Genome Project began in 1989/1990 by the leadership of NIH, USA as a consortium of several countries including U. K., Japan, France and Germany. A venture business Celera Genomics came in the competition using another procedure named “whole genome shot-gun”. They reconciled in June 2000, by shaking hands at both sides of the President Clinton, announcing that draft sequence of the human genome had been completed. The results were published in February issues of Nature and Science this year (2001). The results were mostly similar confirming their efforts were correct. But, some unexpected results were also obtained; e. g. the unexpectedly low number of human genes, very high rate of appearance of trasposon-type repetitive sequences etc. The author will interpret the results carefully and discuss further on interesting points leading to proteomes.

Key words: DNA-sequencing, human genome project, single nucleotide polymorphism (SNP), alternative splicing, domain.

分子生物学における電気泳動

20 世紀の特に後半は、分子生物学が大きな発展を遂げた時代であった。それは、タンパク質と核酸の分析技術に大きな進歩があったことと軌を一にしている。

電気泳動は、その中でも中核を占める大きな技術であり、これ無しには現在の分子生物学も遺伝子工学も有り得なかったであろう。

まず、1930 年代に Tiselius が移動界面法で血清タンパク質を分離することに成功したが、その分解能は極めて低いものであった。

その後、濾紙や他の担体を用いることによってしだいに性能は向上し、1960 年代には Laemmli がポリアクリルアミドと SDS (ドデシル硫酸ソーダ) 溶液を使って分子量によるタンパク質の分離を可能にしたことから、一挙にタ

ンパク質などの高分子の分離が可能となった。そして、1975 年には O’Farrell の 2 次元電気泳動 (1 次元をタンパク質の pI で、2 次元目を SDS ポリアクリルアミドによる分子量で分ける) が成功して、細胞抽出液のような多数のタンパク質を含む検体から 1 度に単一タンパク質を 1 スポットとして取り出すことさえ可能となった。

一方、1970 年代から同様のアクリルアミドやアガロースを使った核酸の分離が Peacock や Loening によって開発され、それまで用いられてきた超遠心機による蔗糖密度勾配遠心法を駆逐してしまったのである。こうして、2 大情報高分子であるタンパク質と核酸は共に電気泳動によって単一化できるようになった。この他、逆層クロマトグラフィーやアフィニティクロマトグラフィーも大いに貢献したことは言うまでもない。

これは、分子遺伝学による遺伝子クローニング法と相俟

Electrophoresis, genome and proteome.

Masami Muramatsu; 埼玉医科大学ゲノム医学研究センター

Correspondence address: Masami Muramatsu, Research Center for Genomic Medicine, Saitama Medical School, 1397-1, Yamane, Hidaka 350-1241, Japan.

第 51 回日本電気泳動学会春季大会・教育講演

って mRNA や遺伝子 DNA の同定に繋がった。そして、遂に 1975 年、Maxam-Gilbert と Sanger らにより、独立に DNA 塩基配列決定法が完成する。これは、数百塩基対の DNA の一塩基差をポリアクリルアミド電気泳動で分けることができるいう、極限的な性能を生かした原理に基づいていたが、これが、その後の遺伝子解析から現在のゲノム計画へと繋がった革命的方法となったのである。両者が同時にノーベル化学賞を受賞したのも宜なる哉といえよう。

遺伝子とゲノム

生物の形質(形と性質, **Phenotype**)の担い手はタンパク質であるが、タンパク質の性質はそれを構成する 20 種類のアミノ酸の並び方(配列)によって決まる。そして、そのアミノ酸の並び方を指定しているのが遺伝子であり、それは DNA という化合物でできている。DNA は鎖状の細長い分子であり、通常、2 重鎖を形成している。その成分は、A, G, C, T の 4 つのヌクレオチド(または塩基対)から成るが、この並び方がタンパク質のアミノ酸の並び方を決めている。したがって、DNA はタンパク質の青写真であり、その遺伝情報は DNA の必要な部分が核の中で mRNA (メッセンジャー RNA) として写し取られ(転写され)て後、細胞質でタンパク質に翻訳されるのである。この間の詳しい機構やそこで使われる**遺伝暗号**の正体が 20 世紀後半に入るとやがて明らかにされた。そして、早くから唱えられた一遺伝子・一酵素(タンパク質)仮説(**Beadle と Tatum**)もほぼ正しいことが証明された。ただし、後になってこれが厳密には成立しないことが明らかになる(後述)。

遺伝子は、タンパク質をコードするという機能のみならず、自分自身を何時、何処で発現させるかという調節機能も持っており、それが、プロモーター、エンハンサーと名

付けられた配列で構成されていることも明らかにされた(図 1)。このようにして、**分子生物学のセントラルドグマ**と言われる。

DNA → RNA → タンパク質
という遺伝情報の流れが確立した。

しかし、自然界にも必ず例外というものがある。RNA を鋳型にして DNA を作る酵素がレトロウイルス中に見つかり、**逆転酵素**と名付けられた(Temin, Baltimore, 1970)。DNA 自身の複製(DNA ポリメラーゼによる)も加えると上述の図式は

と書かれることになる。

さて、遺伝子の構造をよく調べると、特に真核生物ではタンパク質をコードしている部分がほとんど無意味の、時にはとてつもなく長い DNA 配列で分断されていることがわかった。

遺伝子 DNA は、タンパク質をコードするのに必要な配列(3 つのヌクレオチドで 1 つのアミノ酸がコードされるから、例えば 300 アミノ酸のタンパク質の場合 900 ヌクレオチドが必要である)の 10 倍から 50 倍も大きい(長い)。

これは、1 つには先に述べた調節領域の存在もあるが、実際に mRNA として読み取られる部分が極めて長いのである。すなわち、その部分は mRNA になっていく**エクソン**と呼ばれる部分と一旦 mRNA 前駆体として転写されるが、直ちに切り取られ捨て去られる**イントロン**と呼ばれる部分が交互に存在し、**スプライシング**と呼ばれる機構により、エクソン同士が継ぎ合わされ、イントロンは切り離されて分解される(Fig. 1)。この時、繋ぎ合わされるエキ

Fig. 1. 遺伝子の構造と発現。

ソンを選択できる**選択的スプライシング (alternative splicing)** ので、1つの遺伝子から複数のタンパク質を作ることができる。

このような遺伝子発現の分子機構が**遺伝子クローニング技術**の発達によって、かなりよくわかってきたにもかかわらず、複雑な生命現象は依然として謎に包まれている。この謎を解くには更に多くの遺伝子の構造と機能およびそれらの相互作用について知らねばならない。

それはつまり、ゲノムを知らなければならないことを意味する。**ゲノム**とは「ある生物種が持つ遺伝子の総体」という程の意味であるが、いくつかの生物のゲノムサイズ(総塩基対数)を Table 1 に示した。この表からわかるように、細胞に寄生して増殖するウイルス(細菌の場合はファージ)では、小さいもので数千塩基対の大きさがあり、数個の遺伝子を持つ。自然界でほぼ独立に生きていける生物としての細菌では 100 万台の塩基対が必要であり、遺伝子数も 1,000 台(大腸菌で約 4,000)となる。そして、真核生物となるとこれらが格段と大きくなることが気づかれよう。最も原始的な真核生物と言われる酵母は 1,200 万塩基対、遺伝子数にして約 6,000 であるが、線虫では 1 億塩基対で 1 万 9,000 遺伝子となる。長い間、遺伝学で使われてきたショウジョウバエは比較的小さいゲノムを持つことで知られていたが、1 億 4,000 万塩基対(遺伝子約 1 万 2,000)である。脊椎動物になるとゲノムサイズは格段に大きくなる。**ホメオボックス遺伝子群**が昆虫では 1 組なのに、脊椎動物では 4 組あることから、進化上、どこかで少なくとも 2 度に亘る遺伝子重複が起こったと考えられているが、実際はもっと差がある。哺乳動物になると、それらの間での差は極めて小さい。これは、進化的にも哺乳動物の出現が極めて新しいことを示している。

「ヒトゲノム計画」の展開

これらのゲノムの塩基配列は 1989 年から米国の **NIH**

Table 1. いろいろな生物のゲノムサイズと遺伝子数.

種 名	ゲノムサイズ (塩基対)*1	遺伝子数
SV 40 (サルのウイルス)	5200	6
陰部マイコプラズマ (<i>Mycoplasma genitalium</i>)	0.58×10 ⁶	800
インフルエンザ菌 (<i>H. influenzae</i>)	1.83×10 ⁶	1,700
大腸菌 (<i>E. coli</i>)	4.6×10 ⁶	4,300
酵母 (<i>S. cerevisiae</i>)	12×10 ⁶	6,300
線虫 (<i>C. elegans</i>)	100×10 ⁶	19,000
ショウジョウバエ (<i>Drosophila</i>)	140×10 ⁶	12,000
マウス (<i>M. musculus</i>)	3.3×10 ⁹	3~40,000*2
ヒト (<i>H. sapiens</i>)	3.0×10 ⁹	3~40,000*2

* 1: 正確にわかっているものでも概略値のみを示した。
* 2: 未だかなり概算値。上方修正される可能性あり。

の **Francis Collins** を中心に各国が協力して始まった「**ヒトゲノム計画**」というコンソーシアム(研究連合体)の流れの中で次々と決められていったものである。それまでは、英国のサンガーセンターがヒトの DNA 断片を片端からシークエンスして次々とインターネットに載せ、しかも cDNA から得られたものは **est** (expressed sequence tag) と名付けて染色体上の位置情報 **sts** (sequence tagged site) と共に発表し、先端を切っていたが、この計画の開始と共に米国のホワイトヘッド研究所 (MIT), ワシントン大学, エネルギー省 (DOE) 研究所等が各染色体(大きな染色体はその内部の分担を決める)を別々のグループが受け持つ形で、日本やフランス, ドイツも加わって大きな流れとなった。一方、途中から米国のベンチャー企業, セレラジノミクス (**Celera Genomics**, これは **Craig Venter** の率いる TIGR がパーキンエルマーなどの大資本と結合してできたもの) が染色体を分けずに核の全 DNA から出発するという全ゲノムショットガン方式によって参加し、熾烈な先陣争いとなった。

しかし、結局 2000 年 6 月、両者の総帥がクリントン大統領を中心ヒトゲノムの**ドラフトシークエンス**(概要の意味)が完成したと称して握手する幕切れとなった。そして、その結果は 2001 年の『Nature』と『Science』誌の 2 月号に発表された。

これらを分析するに、コンソーシアムとセレラ共、データはまさにドラフト(概要)の段階であって完成品に程遠いということである。第 1 にヒト全ゲノム DNA 配列が大体繋がったとは言っても実は各所に不明の部分(ギャップ)があり、本当に全てがシークエンスされたとは言い難い。それから、配列がわかっても遺伝子がそこに存在するかどうか、が判然としない部分が多い。これは主として配列から遺伝子を見出すソフトが未完成であることによると思われる。遺伝子は先にも述べたようにエキソンとイントロンからなり、多くの場合イントロンの方が遙かに長い(実はこのようなことがわかったことも今回のヒトゲノム計画の成果の 1 つなのである)。例えば、十数アミノ酸をコードするエキソンがあつて、その後に数千塩基から十数千塩基対のイントロンがあつて再び短いエキソンになる場合など、途中に紛らわしいエキソン様の読み枠が出てくことも稀ではない。これがエキソンであるかどうかは実際の mRNA(またはこれを逆転写した cDNA)が存在するかどうかによってしか確かめられないのである。このようなわけで全 cDNA 計画が極めて重要であることが浮かび上がってくる。実際、理化学研究所のゲノム科学総合研究センター(GSC)では、マウスの **cDNA エンサイクロペディア計画**(**林崎良英**ら)が進行中であり、2001 年の『Nature』2 月号には先のヒトゲノム計画特集号の 1 週前にマウス cDNA 2 万 1,000 個の全長コレクションが発表された。こ

の annotation (註釈付け) には世界各国から 70 人の若者が参加した (**FANTOM: Functional Annotation of Mouse**). これらは現在進行中のマウスゲノムの完全配列決定と相俟ってヒトを含む哺乳動物のゲノム構造と、その機能の解析に大きな役割を果たすことになる。

ヒトゲノムに戻って、国際コンソーシアムの結果と Celera の結果とは基本的に一致しており、相互に confirm されていると言えよう。予想されていたことであるが、ヒトの反復配列は全 DNA 配列の約半分 (50%) もあり、その大部分がトランスポゾンであることも確実になった。その中でもレトロトランスポゾンによるものが多く、有名な Alu 配列は 30~50 万から 100 万コピーへと上方修正された。

これら自体は AT に富んだ配列 (約 300 塩基対) であるが、ゲノム全体に散らばっており、特に GC に富む領域に好んで挿入されている。そして、そのことがこのいわゆる **利己的 DNA (selfish DNA)** 自身のみならず、ヒトという種にとっても進化上有利であったと考えられる節がある。全体としてイントロンはエキソンの 10~20 倍もあり、一方、ゲノムが遺伝子によって占められる部分は 10% に過ぎない。残る 90% 以上の DNA が何をしているのか。謎は深まるばかりである。現時点では遺伝子の機能はごく 1 部しかわかっていないが、配列から予想される限りでは、核酸関連酵素、転写因子、各種受容体、プロトンコージン (これまた曖昧な表現であるが)、細胞骨格蛋白質が、この順序で多く見出されている。これらの構造から推察するに 1 つの始原的な遺伝子が重複と組み換え (特に各ドメイン間の) を繰り返しながら大きなファミリーへと成長し、多様な環境や刺激に対応し得る生命体へ進化するという構図が見えてくるようだ。

一方、ゲノム DNA の配列は個人により 0.1% 程異なることが言われてきたが、この値もほぼ正しいことがわかった。このような **単塩基多形 (Single nucleotide polymorphism, SNP)** は特に変異遺伝子の近くに存在する場合、連鎖して動くために、遺伝子診断に利用することができる。国際 SNP マップ・ワーキンググループは先の『Nature』誌上で 142 万の SNP をマップしたと公表したが、Celera は既に 210 万の SNP を見出したという。いずれにしても、それら SNP と病気との **連関 (Association)** を見極めることが必要であり、今後の大きな課題として残っている。さらに言えば、このような SNP のセット (ハプロタイプ **Haplotype**) と疾病との関連を早急に解明する必要がある。

プロテオームとその後

プロテオーム (Proteome) はゲノムの概念を DNA か

らタンパク質へ拡張して唱えられているものである。

O'Farrell の 2 次元電気泳動で数千というタンパク質が大腸菌抽出液から分離された時、これが大腸菌の全タンパク質であるならば、それぞれの 1 次構造 (アミノ酸配列) の全部を決めることも将来は可能となるであろうと囁かれた。事実、大腸菌のゲノムが明らかになってみると、約 4,300 の遺伝子が存在することがわかり、O'Farrell の方法の有用性が改めて浮き彫りにされた。ある生物の持つタンパク質の全てを知ることは、生命活動自体がタンパク質の機能によって支えられていることを考えても極めて重要であることが理解されよう。

1 つのタンパク質の構造や機能をとことん追究し、理解することも重要であるが、“生命”を理解するためには、ゲノムが作り出す全てのタンパク質=プロテオームの構造と機能を知らねばなるまい。それは一朝一夕にして成る事業ではないが、いずれ人類が解かなければならない問題である。

通常、1 つのタンパク質はいくつかの **ドメイン (領域, Domain)** から成り立っている。例えば、立体構造を維持するドメイン、他のタンパク質と結合するドメイン、低分子のリガンドと結合するドメイン、酵素活性を持つドメイン等々である。しかも、これらドメインにも、それぞれの用途に応じた特徴があって多くの種類に分けられる。1 つの細胞が例えば 5 万種のタンパク質を持つとすると、それらが持つドメインの種類は数千に達すると考えられている。ゲノム計画がほぼ終わり、全ての cDNA の構造が明らかになると、細胞内の全てのタンパク質の 1 次構造 (アミノ酸配列) が決まることになる。すると、これらからどのような 3 次元構造 (立体構造) を持ったドメインが存在するのか、がわかってくるであろう。それには各ドメイン単位の立体構造の NMR や X 線結晶解析のデータが必要となろう。これに対する計画は世界的にも始まっており、日本では、理化学研究所のゲノム科学総合研究センターで 20 台の高性能 NMR 装置と、播磨にある放射光装置 Spring 8 でそれぞれを強力に推進することになっている。

いろいろなドメインの基本構造が 1,000 のオーダーでわかるようになると、その組み合わせによる高分子タンパク質の機能解析は、より容易となり、タンパク質間相互作用などもより合理的な説明が可能となるであろう。それにより、遺伝子ネットワークの時空的变化がタンパク質のレベルで解明されることになり、生命活動の本質へより一歩近づいた認識が得られるようになるであろう。

これらの研究が相俟って“生命”の理解とその医療への応用を含め、科学が真に人類の福祉に役立つ時代が到来するであろう。