情 報 学 基 礎 27－8
（1992. 9. 8）

OHFデータから整列クローン・ライブラリーを求める方法

陶山　明
東京大学教養学部物理学教室

我々は、ヒトゲノムの全塩基配列を決定するために、オリゴヌクレオチド・ハイブリダイゼーション・フィンガープリント(OHF)から、整列されたゲノムDNAライブラリを構築する方法を開発した。MOFと呼ばれる我々の方法は、1 MBという巨大なゲノムDNA断片を含むクローンに適用することが出来る。また、大量のクローンを自動化機械で処理するのにも適している。長さ470 kbの人工的なゲノムを用いたシミュレーションの結果、MOF法により実用上満足できる正確さでクローンの整列が行なえることが示された。また、MOF法はOHFの実験データに常に含まれる実験誤差に対して強い方法であることがわかった。

A METHOD FOR CONSTRUCTION OF ORDERED CLONE LIBRARY FROM OHF DATA

Akira Suyama
Institute of Physics, College of Arts and Sciences, The University of Tokyo
Komaba 3-8-1, Meguro-Ku, Tokyo 153, Japan

We have developed a method for constructing ordered genomic DNA libraries from oligonucleotide hybridization fingerprints (OHF) to accomplish the whole genome sequencing of human genome. Our method, MOF (Mapping by Oligonucleotide Fingerprints), can be applied to large clones containing 1 Mb genomic DNA fragments. It is also suitable for processing numbers of clones on automated systems. The computer simulation for 470 kb artificial genome demonstrated that clones can be actually ordered by MOF with practically satisfactory accuracy. The simulation experiments also demonstrated that MOF is highly resistant to experimental errors always included in actual OHF data.

1 はじめに

遺伝情報は、DNA分子を構成する4種類の塩基、アデニン、チミン、グアニン、シトシン（A, T, G, Cと表記される）の配列に蓄えられている。したがって、ゲノムDNAの塩基配列、すなわち、A, T, G, Cの文字の並びを決定することは、ゲノムの遺伝情報解読の第一歩といえる。

長さ500塩基程度の短いDNA断片の塩基配列を末端から1塩基ずつ決定する技術は、1980年代の前半に確立された。そして、この技術をもちいて、既に多数の塩基配列が決定されている。これらの成果をもとに、現在、30億塩基の長さをもつヒトゲノムをはじめとする巨大ゲノムの全塩基配列を決定するプロジェクトが世界的に進められている。

しかし、巨大ゲノムの全塩基配列を既存の短いDNA断片の塩基配列決定技術のみで決定することは、事実上、不可能である。ゲノムDNAを、互いに重なりをもつ配列決定可能な長さのDNA断片にランダムに分断し、決定されたそれらの塩基配列をもとに全塩基配列を再構成しようとすると、実は、ゲノムの全長の5倍から25倍という長いDNAの塩基配列を決定しなくてはならないからである。

この問題を克服する方法のひとつは、図1に示したように、ゲノムDNA断片を含むクローンをゲノム上にマップし、その位置を決定することである[1]。クローンの位置が決りクローンが整列されると、ゲノム全体を覆いつくすのに必要な最少数のクローンを見つけることができる。それらのクローンについてのみ既存の方法で1塩基ずつ配列決定を行なえば、塩基配列決定の作業が大幅に削減され、巨大ゲノムの全塩基配列を決定することが可能になる。

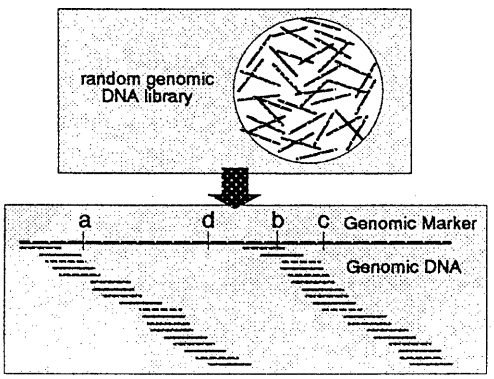

図1　クローンのマッピング

ヒトなどのようにゲノムサイズが非常に大きい場合、クローンの位置決定を行なう方法には、1 MB程度の大きいゲノムDNA断片を含むクローン・ライブラリーに適用できること、大量のクローンを処理するための機械による自動化が可能であることが必要とされる。これらの条件を満たす方法として、我々は現在、MOF (Mapping by Oligonucleotide Fingerprints) 法の開発を行なっている [2]。この方法は、現在のところ、巨大な1 MBという巨大なゲノムDNA断片を含むクローンにも適用可能な、唯一の汎用的な方法である。ここではMOF法の概説を行なう。詳細は投稿予定の論文[2] を参照されたい。

2 MOF法

2.1 方法の概要

MOF法は、実験により得られるオリゴヌクレオチド・ハイブリダイゼーション・フィンガープリント (OHF) から、連続したクローンの集まりであるcontig (contiguous nucleotide sequence) 内でのクローンの位置を決定する方法である。OHFは、DNAプローブ・オリゴマーが結合する塩基配列をクローンがもつかどうかを表すデータであり、DNAオリゴヌクレオチドを用いて調べたクローンの指紋 (fingerprint) といえる。図2に示したように、クローンDNAをナイロンメンブラン上にドットブロットしたのち、DNAプローブ・オリゴマーとハイブリダイズさせることにより求められる。

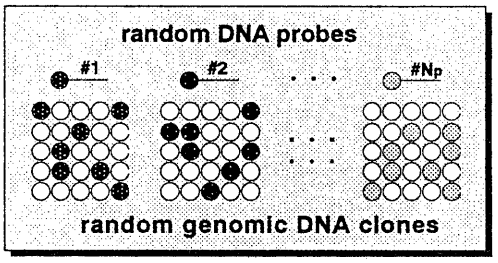

図2　OHF

MOF法では、N_p 個のDNAプローブについて決定したOHFデータを用いてクローンの整列を行なう。N_p の値は、整列の正確度と実験の手間とのトレードオフにより決る。後に示すように、100個程度のDNAプローブを用いれば、実用上十分な正確度でクローンの整列を行なうことができる。プローブの塩基配列としては、ランダムでしかもベクターや既知の繰り返し配列には結合しないものを使用する。

MOF法により、OHFからクローンの整列を行なう解析は大きく三つの過程にわけられる。最初の解析過程では、OHFから各々のクローン対の重なりの大きさを決定する。次に、重なりの大きさを用いてMASOC (MAximum Set of Ovelapping Clones) を求め、MASOC内でのクローンの位置をメトリック多次元尺度法により決定する。MASOCとは互いに重なりをもつクローンの最大集合である。したがって、MASOC内の全てのクローンと重なりをもつクローンは他に存在しない。最後に、MASOC内のクローンの位置とcontig内でのMASOCの位置からcontig内でのクローンの位置が決定され、クローンの整列が行なわれる。

2.2　クローンの重なりの大きさの決定

OHFから i 番目と j 番目のクローンとの重なりの大きさを決定するために、N_p 個のDNAプローブの中で二つのクローンに同時に結合するDNAプローブの総数 h_{ij} をまず算出する。ハイブリダイゼーション・シグナルの強度からクローンに結合するDNAプローブの個数を求めることは大きな実験誤差をともなうので、プローブがクローンに結合したか否かの情報のみを使用する。

クローン間の重なりの大きさを O_{ij} とすると、h_{ij}/N_p は平均値が

$$m(O_{ij}) = \alpha^3 O_{ij}^{\ 3} - \alpha^2(\alpha\beta + \alpha + 1)O_{ij}^{\ 2} + \alpha(\alpha^2\beta + \alpha\beta + \alpha - 1)O_{ij} - \alpha^2\beta,$$

標準偏差が

$$\sigma(O_{ij}) = \sqrt{\frac{m(O_{ij})[1 - m(O_{ij})]}{N_p}}$$

の分布に従う。したがって、重なりの大きさ O_{ij} は、三次方程式

$$\alpha^3 O_{ij}^{\ 3} - \alpha^2(\alpha\beta + \alpha + 1)O_{ij}^{\ 2} + \alpha(\alpha^2\beta + \alpha\beta + \alpha - 1)O_{ij} - \alpha^2\beta - \frac{h_{ij}}{N_p} = 0$$

の解として与えられる。DNAプローブの長さを l_p 塩基、i 番目と j 番目のクローンに含まれるゲノムDNA断片の長さを L_i, L_j とすると、

$$\alpha = \frac{L_i}{4^{l_p}}, \qquad \beta = \frac{L_i}{L_i}$$

である。

三次曲線 $m(O_{ij})$ の形の解析から、重なりの大きさを正確に求めるためには、β が

$$\beta = 1$$

α が

$$0 < \alpha \leq \alpha_c = 0.4142$$

を満たす α_c に最も近い値をとる必要があることがわかる。β の条件から、クローンに含まれるゲノ

ムDNA断片の長さが均一であることが必要とされる。また、αに対する条件から、DNAプローブの最適の長さが決定される。最適の長さはクローンに含まれるゲノムDNA断片の長さにより決る。たとえば、YACベクターのように1 Mbという巨大なDNA断片でも、その長さは高々11塩基である。このように長さは短いので、プローブを調製する上でなんら問題は生じない。

2.3 MASOCの決定

MASOCを構成するクローンとMASOCの順序関係は、図3のように求められる。クローンの重なりの大きさから直接決定した図の左の行列は、クローン同士の間に重なりがあるか否かを表わしている。1は重なりがあることを、0は重なりがないことを示している。クローンの順序を入れ換える行列の変換操作により、この行列は図の右のように右上に0が集まった形に変えることができる。この行列で、要素が全て1の三角形領域に含まれるクローンが各MASOCを構成するクローンとなる。また、三角形領域の位置関係がMASOCの順序関係を表す。

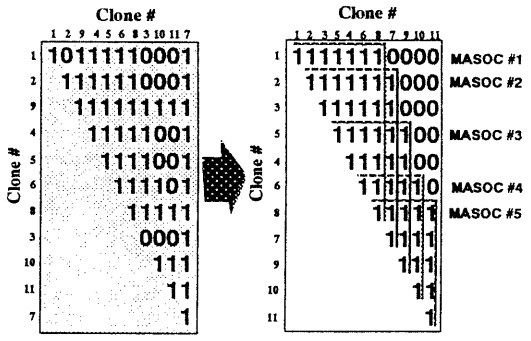

図3 MASOCの決定

2.4 MASOC内のクローンの位置決定

メトリック多次元尺度法は、全てのデータの間の距離が与えられたときに、最小次元で構成される空間の中でデータの配置を決定する多変量解析の一手法である [3]。重なりの大きさが O_{ij} の二つのクローンの間の距離 d_{ij} は、クローンに含まれるゲノムDNA断片の平均長を L とすると、

$$d_{ij} = L(1 - O_{ij})$$

で与えられる。MASOCに含まれる全てのクローン対について重なりの大きさが与えられているので、上式によってMASOC内の全てのクローン対について距離を定義することができる。したがって、これらの距離データをメトリック多次元尺度法で処理することにより、MASOC内でのクローンの位置を決定することができる。

しかし、こうして決定されたクローンの位置は正確ではない。OHFデータから求めたクローンの重なりの大きさ、クローンの距離には、統計的な誤差や実験誤差が含まれているためである。メトリック多次元尺度法で処理する際に、MASOC内の全てのクローンの距離データを用いると、たまたま統計誤差や実験誤差が大きい距離データのために、決定されたクローンの位置が全体的に著しく不正確になる。そこで、MOF法では、その原因となるクローン、すなわち、そのクローンとの重なりの大きさのデータが他のクローンのデータと整合性がよくないクローンを発見し、そのようなクローンを取り除いて位置の決定を行なうようにした。

整合性がよくないクローンを見つけるために、整合性指数 TC

$$TC = \sum_{i<j} \left[\frac{m(O_{ij}^{MDS}) - m(O_{ij}^{EXP})}{\sigma(O_{ij}^{MDS})} \right]^2$$

を用いた。ここで O_{ij}^{EXP} はOHFの実験データから

求めたクローン間の重なりの大きさ、O_{ij}^{MDS}はメトリック多次元尺度法により決定された位置を用いて計算したクローン間の重なりの大きさである。TCは、メトリック多次元尺度法により決定されたクローンの位置を用いて計算したh_{ij}/N_pとOHF実験データから得られるh_{ij}/N_pとのずれを表している。MASOC内の全てのクローン対の重なりの大きさのデータの整合性がよいと、メトリック多次元尺度法によりクローンの位置が正確に決り、TCは小さくなる。

2.5 Contig内のクローンの位置の決定

図4は、MASOC内のクローンの位置とMASOCの順序関係からcontig内でのクローンの位置を決定する方法を示している。まず、隣合うMASOCに共通するクローンの位置の平均位置を一致させることにより、MASOCの重心間距離が求められる。この距離を用いて、contig内でのMASOCの重心位置が決定される。こうして決定されたcontig内でのMASOCの重心位置とMASOC内でのクローンの位置とから、クローンのcontig内での位置を決定することができる。

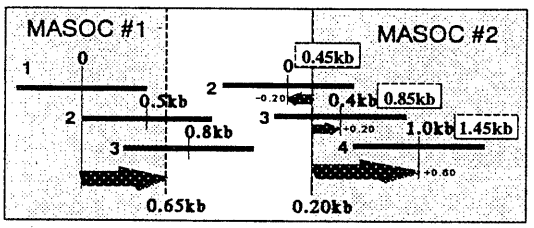

図4 クローンの位置決定

3 MOF法の評価

3.1 シミュレーション実験

MOF法によるクローンの整列の正確度を評価するために、長さが470 kbのランダムな塩基配列をもつ人工のゲノムDNAを用いて、シミュレーション実験を行なった。470 kbという長さは、大腸菌ゲノムの10分の1の大きさである。

ランダム・ゲノムDNAライブラリーを計算機上で作成するために、まず、このゲノムDNAを6塩基認識の制限酵素Sau3Aで部分分解した。そして、得られたDNA断片の中から、長さが15 kb±2 kbの断片を100個ランダムに選び出した。これらの処理は、長さ15 kb±2 kbのゲノムDNA断片を含むランダム・ゲノムDNAライブラリーから、100個のクローンをランダムに選ぶ実験操作に相当する。

次に、こうして選んだ100個のクローンと、長さ9塩基の100個のランダムDNAプローブとのハイブリダイゼーションを計算機の中で行い、OHFを決定した。そして、得られたOHFをMOF法のデータ処理プログラムにより解析し、それぞれのクローンのcontig内での位置を決定した。

3.2 クローンの整列の正確度

図5は、シミュレーション実験の結果の一部を示したものである。正しいクローンの位置（実線）とMOF法で決定された位置（破線）とを比較すると、MOF法によりクローンの位置がかなり正確に決まることがわかる。

OHFから得られるクローン間の重なりの大きさには必ず統計誤差が含まれる。したがって、MOF法で決定されるクローンの位置には真の位置からのずれが存在する。このずれの大きさの分布を示したものが図6である。図の横軸は、MOF法で決定されたクローンの位置が正しい位置からどれだけずれているかを表している。単位はクローンに含まれるDNA断片の平均長である15 kbである。縦軸

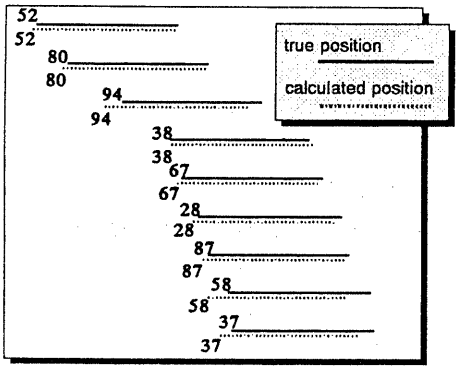

図5　クローンの位置の真値と計算値の比較

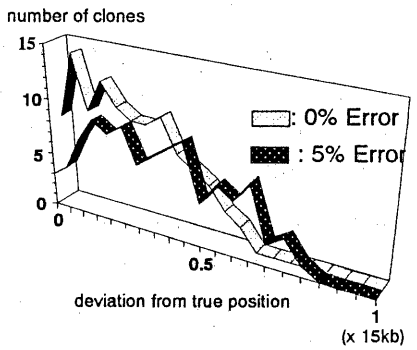

図6　クローンの位置決定の正確度

は、あるずれの大きさをもつクローンの数を表わしている。

図6の▨線から明らかなように、ずれの広がりが存在する。その主な原因は、OHFから得られるクローン間の重なりの大きさに含まれる統計誤差である。したがって、より多くのDNAプローブについてOHFを決定し解析に用いれば、ずれの広がりは小さくすることが出来る。しかし、より多くのDNAプローブについてハイブリダイゼーション実験を行ないOHFを決定することは、実験に要する時間、コストを増やすことになる。したがって、実験の手間とクローンの位置の正確度との間でトレードオフがなされることになる。図6の▨線は、100個程度のDNAプローブについてOHFを決定すれば、相当数のクローンについて、実用上問題のない程度に位置が正しく決定されることを示している。

3.3　実験誤差の影響

実験により得られるOHFには、false positiveあるいはfalse negativeの実験誤差が必ず含まれる。そこで、これらの実験誤差がMOF法によるクローンの整列におよぼす影響を調べた。実験誤差は、OHFの全ドットの中から適当な割合のドットをランダムに選び、それらのハイブリダイゼーション結果を反転させることにより導入した。

5%のデータを反転させたOHFを用いてMOF法によりクローンの整列を行なった結果を図6に■線で示した。誤差が全くない場合に比べて少し正確度が落ちるが、実用上問題がない程度にクローンの整列を行なうことが出来ることがわかる。

OHFに実験誤差が混入してもMOF法により得られるクローンの位置の正確度が大きく減少しないのは、多次元尺度法によりMASOC内のクローンの位置を決定する際に、整合性指数 TC を用いて最適化処理を施しているからである。この処理を行なわないと、少しの実験誤差でも正確度がかなり悪くなる。最適化処理のために一部のクローンのOHFデータは捨てられるのでクローンの数を2、3割増やすことが必要になるが、クローンの整列の正確度を減少を大きく押えることが出来る。

4　参考文献

1) 小原雄治, 蛋白質・核酸・酵素, **35**, 2335-2347 (1990).
2) Suyama, A., Itoh, H., and Wada, A. (in preparation).
3) R.Sibson, Multidimensional Scaling, Wiley (1981).