遺伝子配置順序データベースの作成と色素体ゲノムの進化への応用

Development of gene order database for plastid genomes and inference of genome evolution

○国沢 隆[1]、ディヴィド サンコフ[2]; Takashi Kunisawa[1], David Sankoff[2]

As genomic sequence data accumulate, methods for comparing gene arrangements on genomes are becoming more important. We had previously developed a combinatorial algorithm to compute an approximate optimal number of recombinations (inversions or transpositions) necessary to convert the gene order in one genome into that of the other. We applied this algorithm for signed permutation to the evolution of plastid genomes. The incomparability of gene labeling is a major problem associated with the use of database annotations. We have constructed a normalized labeling system across a total of 11 plastid genomes, which have so far been completely sequenced. It was shown that the fraction of recombinations to the number of homologous genes compared is an appropriate measure of evolutionary distance for phylogenetic inference. Biological significances of the derived phylogeny were also discussed.

生物の形態を比較して生物の系統一起源と進化一を論じるという生物学の伝統的方法に対して、ＤＮＡやこれにコードされた蛋白質などの分子の情報から生物の歴史を推察しようとする分子系統学が近年急速に発展してきている。この新しい方法は、個々の遺伝子の塩基配列または蛋白質のアミノ酸配列の比較に基づくものと、ゲノム（ＤＮＡ全体をさす）の遺伝子配置順序の比較によるものとに、分けられる。ここでは、遺伝子の並び順の比較からゲノムの進化を推論するときに有用なデータベース構造を議論する。次に、二つの異なる遺伝子順序を同一にするために必要な組み替え数の推定は自然数の整列という古くからよく知られた問題に帰着することを紹介する。また、こうした数理的解析結果の生物学的意味についても、色素体ゲノムの進化を例にして、述べる。

遺伝子配置データベース

近年の分子生物学の発展に伴い、ＤＮＡの塩基配列データが急速に蓄積しており、これらのデータはＧｅｎＢａｎｋ／ＥＭＢＬ／ＤＤＢＪ等のデータベースから利用者に提供されている。色素体を例にすると、表１に示すように、１０種類のゲノムの塩基配列データが１０個の独立したエントリーとして登録されている（表１の *Chlorella vulgaris* は未発表）。塩基配列データが発表されている論文の情報と共に何番目の塩基から何番目の塩基までというように同

表１ 配列決定された色素体ゲノム

ｹﾞﾉﾑ	ﾃﾞｰﾀﾍﾞｰｽ ｱｸｾｯｼｮﾝ	ｻｲｽﾞ (Kbp)	遺伝子数	包膜	色素
緑色植物					
陸上植物					
Oryza sativa	X15091	135	158	2	chl a, b
Zea mays	X86563	140	152	2	chl a, b
Nicotiana tabacum	Z00044	156	136	2	chl a, b
Epifagus virginiana	M81884	70	55	2	chl a, b
Pinus thunbergii	D17510	120	141	2	chl a, b
Marchantia polymorpha	X04465	121	137	2	chl a, b
緑藻					
Chlorella vulgaris C-27 (M.Sugiura, unpublished results)	-	151	210	2	chl a, b
ユーグレナ藻					
Euglena gracilis	X70810	143	103	3	chl a, b
珪藻					
Odontella sinensis	Z67757	118	175	4	chl a, c
紅藻					
Porphyra purpurea	U38804	191	252	2	chl a, phycobilin
灰色植物					
Cyanophora paradoxa	U30821	136	192	2?	chl a, phycobilins

定された遺伝子や蛋白質をコードする可能性のある領域（ＯＲＦという）が注記されている。これらの注釈は、利用者にとって大変便利なものであるが、ゲノムの遺伝子順序を比較したい時などには、異なるエントリー間で遺伝子名等が統一されていないことが問題になる。このincomparability はデータ解析にとって非常に不便なもので、解析者毎にデータを修正することが必要になる。データが増加するにつれて、より高い信頼度で注釈することが可能であり、実際新しいエントリーでは遺伝子名等はより統一的につけられているが、古いエントリーが同時に修正されていないことが実態である。知識はデータの増加とともに変わっていくことを前提にして、データベース全体を更新していくシステムの開発が強く望まれる。ここでは、表１に示した１１種類のゲノムに対して、配列の類似性に基づいて、統一的な遺伝子（またはＯＲＦ）名を付けて、遺伝子順序が簡単に比較できるようにした。現在、様々なゲノムサイズを持ったゲノムの全塩基配列を決定しようとするプロジェクトが進行中であり、将来的には、例えば、１１種類全体にわたって保存した遺伝子配置や、ある３つのゲノムで共通に認められる順序を検出可能なデータベースソフト開発が必要であろう。

最小組み替え数の計算

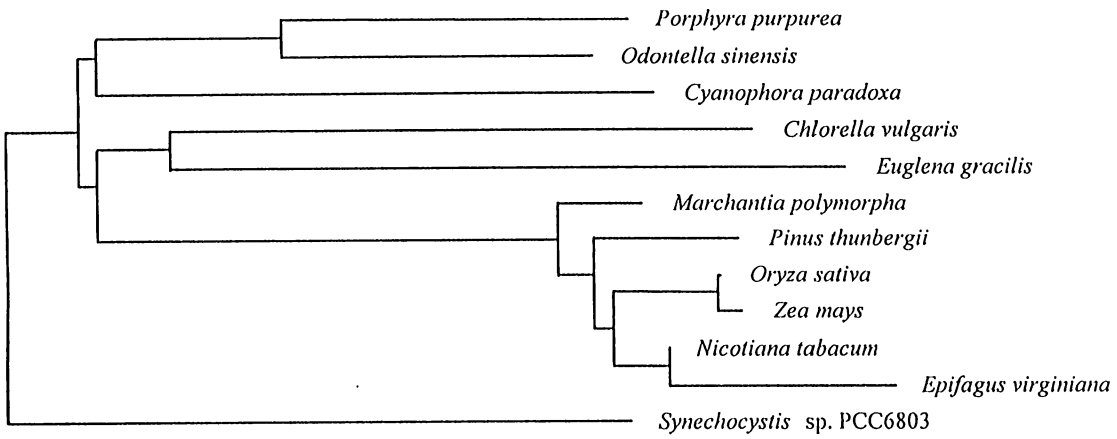
一組のゲノム対の遺伝子配置順序を比較すると、片方には存在するが、他方には見つからない遺伝子がある。こうした遺伝子（またはＯＲＦ）は進化の過程で失われたか、色素体の場合には核ゲノムへ移動したものとの両方の可能性が考えられる。消失や核への移動にも系統に関する情報も含まれている筈であるが、ここでは、片方のゲノムにしか存在しない遺伝子を除外して、両方のゲノムで共通して存在する相同遺伝子順序の比較から系統関係を考察する。遺伝子順序間の進化的距離として、二つの異なる遺伝子順序を同一にするために必要な最小の組み替え数を定義するのが最も自然であろう。種が分岐する以前には、２つのゲノムは同一の遺伝子順序を持っていた筈であり、種の分岐後に起こった遺伝子群の逆位や転座によって配置順序が変更されたと考えて良い。一方のゲノムの相同遺伝子に配置順序に従って通し番号を付けると、このゲノムの遺伝子順序は（1,2,3,4,...,n-1,n）と表すことができる。ここで相同遺伝子の総数を n とした。こうすると、もう一方のゲノムの遺伝子順序は、例えば（1,-3,-2,4,n-1,n,....,n-2）となろう。ここで、-3,-2 は遺伝子３と２はＤＮＡの２本鎖の反対側に移動し、これを逆位という、転写方向も逆向きになったことを表す。また、順序 4,n-1,n, は２番目のゲノムでは２つの遺伝子 n-1 と n が遺伝子４の後に位置が替わった、つまり転座したことを示す。このように、遺伝子順序間の進化的距離を求めることは、符号を持った数列の整列という古くから知られた問題に帰着し、同一の順列にするためには最小何回の逆位や転座というＤＮＡの組み替えが必要かを計算すればよいことになる。

ところが、最小の組み替え数を求めることは、実際には極めてむずかしい。これは、色素体ゲノムの場合では表１からわかるように相同遺伝子数は１００よりも多く、可能な組み合わせの数が処理できないほど多いことに起因する。この事態に対処するために、数理計画法を用いたアルゴリズムを開発してきた（Proc. Nat. Acad. Sci. 89:6575-6579, 1992; Gene 172:GC 11-17, 1996）。このアルゴリズムを用いて、近似的な極少組み替え数を、色素体ゲノムについて計算した。色素体の起源と進化は後に述べる共生説との関係から多くの関心を集め、その系統関係についてはかなり詳細に調べられており、あるていどのコンセンサスが得られている。こうした生物学上の知識と得られた結果を比較することによって、このアルゴリズムの妥当性が、後述するように、検討できる。ともかく、陸上植物の色素体間では組み替え数は１０以下であり、よく似た遺伝子順序であることが分かる。これに対して、陸上植物以外の色素体間での組み替え数は３０以上に達し、進化の過程でかなりかき混ぜられていることになる。

色素体ゲノムの系統関係

求めた組み替え数の相同遺伝子数に対する割合を進化的距離と定義して、図１に系統関係を図示した。図１の系統樹では、シアノバクテリアの一種 *Synechocystis* sp. の遺伝子順序をアウトグループとみなして根をつけて、このシアノバクテリアの仲間から色素体が派生したように描写している。また、系統樹は、Neighbor-Joining Method と呼ばれる方法を用いて作成した。図１に示した系統樹は、最初に非緑色系の色素体（*Porpyra purpurea*, *Odontella sinensis*, *Cyanophora paradoxa*）とその他の緑色系の色素体の分岐が起こり、珪藻 *Odontella sinensis* は紅藻 *Porpyra purpurea* と近縁であり、また、ユーグレナ藻 *Euglena gracilis* は

図1 色素体ゲノムの系統樹



緑藻 *Chlorella vulgaris* と近縁であることを示している。この分岐の様子は全体として生物学的知見と一致している。このことは、逆に、用いたアルゴリズムの妥当性を示していると考えてよかろう。得られた系統樹は、これまで進化上の位置が明確でなかった灰色植物 *Cyanophora paradoxa* の色素体（シアネラ）は非緑色系の色素体の中で分岐していることを示唆している。

色素体ゲノムではリボゾーム蛋白をコードする一群の遺伝子は散在するのではなく集まってクラスターとして存在している。そして、この遺伝子の並び順は真性細菌なかでもシアノバクテリアと呼ばれる酸素発生型光合成細菌の並び順とよく似ている。このことは、あるシアノバクテリアが核を持った真核細胞に食べられて、その後、共生体となったという共生説を支持する。また、表1に示した様々な色素体ゲノムで本質的に同一の並び順であることは、共生が一個のシアノバクテリアから始まったことを示唆しよう。さらに、色素体を包む膜の数と図1の系統樹から、紅藻 *Porpyra purpurea* が2回目の共生をして珪藻 *Odontella sinensis* が生じ、同様に、緑藻 *Chlorella vulgaris* が2回目の共生をしてユーグレナ藻 *Euglena gracilis* の起源となったというシナリオを示唆する。

[1]東京理科大学理工学部応用生物科学科、 [2]モントリオール大学数学研究所
[1]Department of Applied Biological Sciences, Science University of Tokyo, [2]Centre de recherches mathematiques, Universite de Montreal