千葉中央博自然誌研究報告第2巻 第2号: 83-98. 1993年3月.

組合せ論的視点から見た系統推定

——最節約法と離散数学の接点——

三中信宏

農林水産省農業環境技術研究所環境管理部計測情報科調査計画研究室
〒305 茨城県つくば市観音台 3-1-1

要 旨 最節約原理に基づく系統推定において生じるいくつかの概念的・計算的問題は，離散数学（組合せ論・グラフ理論・半順序理論）を用いて解決することができる．本論文では，生物分類学と系統学における分岐分類の手法が抱える三つの組合せ論的問題を論じる．

第一の問題は，形質状態データ行列からの最節約分岐図の構築に伴って生じる．分類群間の包含関係に基づく半順序集合として分岐図を定義すると (Minaka, 1987)，複数の最節約分岐図の集合を単一の Hasse 図として表現することができる．系統樹は分岐図とは概念的に異なる．順序構造としての系統樹は，分岐図とは異なり，祖先子孫関係という別の順序関係から構築されているからである．また，それらの分岐図の集合はある集合 Boole 代数の部分構造と解釈され，分岐図の直和・直積演算 (Minaka, 1990) および順序情報量を定義できる．分岐図の直和は，各分岐図の OTU 数に等しい次元をもつ Boole 代数の部分構造として網状の Hasse 図を形成する．代数的にこの Hasse 図を解釈すると，もとの分岐図のすべての分類群の包含関係に関する情報はそこに保存されている．分岐図の直積もまた同様の網状の Hasse 図を形成する．しかし，その Hasse 図が埋めこまれる Boole 代数はもとの分岐図よりも高次元である．これらの分岐図演算は，複数の分岐図のもつ情報を簡潔に要約するのに用いることができる．半順序に基づく分岐図の定義を用いるもう一つの利点は，分岐図あるいはその集合のもつ順序情報量を全順序拡大の個数の相対値として定量化できることである．ある分岐図の全順序拡大の個数は，その分岐図の順序イデアルの全体集合の Hasse 図から計算できる．

第二の問題は，分岐図の全長に関する樹長分布の組合せ論的特性である．樹長分布の特性を，同時分布および周辺分布と関係づけながら論じる．分岐図長の同時分布とは，各形質に対する分岐図の長さを軸とする多次元空間内での頻度分布である．この同時分布を全長軸に対して射影することにより，樹長分布が得られる．この射影に伴う次元の減少（多次元→1次元）は，情報の損失をもたらす．実際，樹長分布においては，全長は同じで樹形の異なる分岐図の群の存在に関する情報が失われている．なぜなら，樹長分布を構築するときの射影の方向は，分岐図間の樹形の差を表す方向（全長軸に直交する）と平行しているからである．

第三の問題は，ある分岐図のもとでの仮想的形質状態の最節約的な復元に関係している．Farris (1970) と Swofford and Maddison (1987) は，順序的形質について，HTU 復元のための方法を提案した．最近，Hanazawa, Narushima and Minaka (1992, to appear) は，彼らの方法をより一般化するアルゴリズムを与えた．Hanazawa *et al.* は，Swofford and Maddison (1987) において証明が不備であった命題の完全な証明を与え，さらに多分岐的な分岐図に対する一般化を行ない，与えられた情報のもとですべての可能な最節約的 HTU 形質状態配置を枚挙する再帰的アルゴリズムを開発した．複数の最節約的 HTU 配置が存在するとき，それらはある順序関係に基づくベクトル空間を構成する．ACCTRAN 配置と DELTRAN 配置はこのベクトル空間においてそれぞれ最大値・最小値となるだろう．HTU 形質状態の復元方法は生態学・行動学における種間比較に影響をもたらす．

キーワード: 系統推定，分岐分類学，最節約原理，離散数学，組合せ論，比較法．

1. 系統分類体系をめぐって

生物分類体系を構築することは，生物群に関する形質情報を体系化し，それに基づいて対象生物の分類を行なうことを意味する．その際，生物進化の歴史すなわち系統発生の問題を避けては通れない．生物に関する形質情報を何らかの基準にしたがって要約・貯蔵し必要に応じて検索できて，はじめてそれらの情報は「体系化」されたといえるだろう．問題は，形質情報を体系化するための基準をどうするのかという点である．ここでは，単純な例を挙げて，形質情報の体系化の意味を考えたい．

簡単な例として，{ヒト，イヌ，カエル}という対象生物群を考える．これらの生物は，リンネ式階層分類体系では，図1のように分類されている．図1の階層分類体系を支持する形質情報はいくつもあるが，その

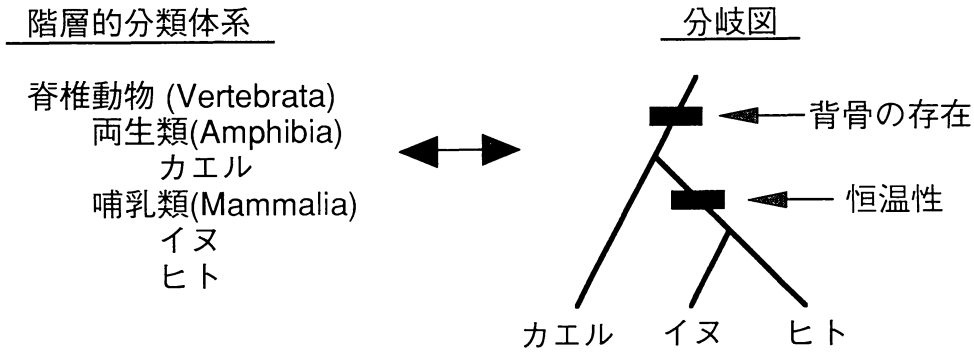


図 1. 階層的分類体系と分岐図の対応．形質情報に基づく最節約分岐図と階層的分類体系を 1 対 1 に対応させるというのが，分岐分類学の考え方である．分岐図においては，3 種の脊椎動物が作る分類群{カエル，イヌ，ヒト}および{イヌ，ヒト}の包含関係だけではなく，それらの分類群を支持する形質情報も明示的に表示できる．しかし，階層的分類体系では，分類群間の包含関係だけが明示的に表示される．

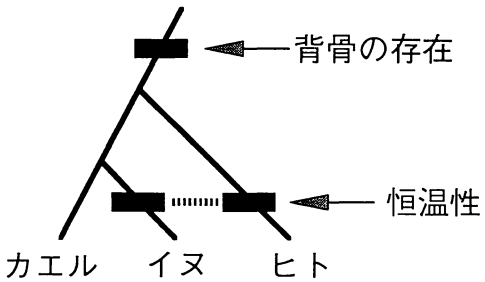


図 2. 非最節約分岐図における形質分布の表現．図 1 と同じ分類群に関する，ある非最節約分岐図上の形質分布の表現．{カエル，イヌ}という分類群を持つこの分岐図上では，恒温性という形質の分布を説明するのに，2 回の形質変化を要求している．

うち「恒温性」ならびに「背骨の存在」という二つの形質を考える．前者はイヌとヒトに限定されているが，後者は 3 種すべてに分布している形質である．このとき，上の階層的分類体系の持っている構造［分類群の間の集合論的包含関係の構造］は，分岐図 (cladogram) というグラフによって表示することができる．この包含関係の構造は集合論でいう “Venn diagram” を用いても表現することができ，{カエル{イヌ，ヒト}}と表される．それでは，この分岐図は，形質の情報をどのように体系化しているのだろうか．次の 2 点を指摘することができる：

1) 形質の「共有性」の表示：それぞれの形質がどの生物に観察されるのか，また，ある形質を共有する生物はどれかが明らかになる．
2) 形質の「普遍性」の表示：共有形質の分布の相対的な広さが明らかになる．例えば，恒温性と背骨の二つの形質を比べると，後者を共有する群が前者を共有する群を真部分集合として包含しているので，背骨の方がより普遍性の高い（分布の広い）共有形質であることがわかる．

共有形質に基づいて生物群を分類することにより，形質の分布に関する情報が効率的に要約されていることに注意されたい．情報の体系化という観点から分類体系の評価を下すならば，もっとも単純に形質情報を要約できる分類体系が最適な体系として選択されるのである．例えば，上の例で{{カエル，イヌ}ヒト}という対立仮説（図 2 を参照）を考えると，この分類体系は背骨の形質分布をうまく説明できても，恒温という特性の分類学的分布を効率的に説明できない．図 2 の対立仮説のもとでは，恒温性という形質の分布をイヌとヒトで別々に記載しなければならないから，形質情報の要約の観点からはこの分類体系は非効率的である．

この単純な例からもわかるように，分類体系の対立仮説がいくつかあるときに，形質情報の要約という定量的観点から仮説間の客観的評価を下すことが可能になる．対立仮説のうち，もっとも単純に形質分布を説明できる仮説を選択するという原理—「最節約性 (parsimony) の原理」という—は，分類体系の構築を考える上できわめて重要な基準であるといえる．最節約原理に基づく系統推定法（「最節約法」parsimony method と呼ばれる）は，もともとは分岐分類学 (cladistics) という系統分類学の一方法に由来するものである (Wiley, 1981; Wiley *et al.*, 1991)．しかし，現在では形態学的形質だけでなく分子レベルの配列データをも解析できる手法として系統学において広く用いられている．

系統発生に基づく生物分類体系—「系統体系」(phylogenetic sysytem) という—を構築するための方法はさまざまな問題を含んでいる．その第一は，上で述べた最節約原理に基づく推定量である最節約分岐図が推定量としてどのような統計的性質を持っているかという問題である．近年，この点については最尤法

(maximum likelihood method) と最節約法の支持者の間で活発な論争が見られる．分岐分類学において実践されている最節約原理に基づく系統推定は，実際の進化プロセスに関する仮定（モデル）を最少にしつつ分岐図の構築を行なおうとする．この点で最節約法は，進化モデルを厳密に規定する最尤法とは異なっている．しかし，Sober (1988) が試みたように，最節約法をある弱い進化モデルに従う系統推定法とみなすならば，最節約法の与える分岐図は最尤推定値であることが証明されるかもしれない．それが成功したならば（現時点ではまだ "open problem" だが），最節約法と最尤法は，異なる進化モデルのもとでの最尤推定を行なうという点で共通しているといえるだろう．

最節約原理の分岐分類学における位置づけに関しては，生物学・科学哲学・統計学などさまざまな観点から議論されている．これらの問題に興味のある読者は，最節約法を支持する Sober (1983, 1985, 1988) と最尤法を支持する Felsenstein (1983, 1988) および誌上論争 (Felsenstein and Sober, 1986) を参照されたい．

2. 組合せ論的問題としての系統推定

最節約法のさまざまな側面は，前節で言及した統計学的問題だけでなく，組合せ論 (combinatorics) と—あるいはもっと一般的に離散数学 (discrete mathematics) と—密接な関係を持つことが知られている．（系統分類学に関連する範囲の離散数学については，三中 (1991) の「付録」を参照されたい．）グラフ理論・組合せ論・順序理論などを包括する離散数学は，もっとも新しい数学の一領域であるが，最節約法に基づく系統推定の根本問題のいくつかは離散数学の応用例として興味深い問題を提起している（三中，1991, 1992a）．以下で「組合せ論的」という言葉を用いるとき，それは解こうとする問題が離散数学的に扱えることを意味する．

組合せ論的にアプローチできる系統推定問題として，具体的に次の二つを議論する：1) 分岐図の樹形（トポロジー）の推定；2) ある分岐図のもとでの仮想的形質状態の復元．第一の問題は従来からあった系統学上の問題だが，形態・分子のデータが急速に蓄積されている今日では，大量の形質情報から迅速に分岐図を作成する手法と得られた分岐図の信頼性の評価方法の開発が大きな課題となっている．第二の問題は，従来は第一の問題に付随する二次的な問題と考えられていた．しかし，仮想的共通祖先での形質状態の復元は，分岐図の樹形の推定とはまた別の独立した組合せ論的問題であることがわかってきた．最近の行動学や生態学における「比較法」(comparative method) が形質進化の時間的順序と相関を議論するようになって，この認識はさらに強まっている（三中，1991）．

3. 分岐図の樹形（トポロジー）の推定

ある形質データのもとでの最節約的な分岐図推定は，グラフの末端点 (OTU: 操作的分類単位) に関する制約条件（形質状態）が与えられたとき，いくつかの仮想的な内部点（HTU: 仮想的分類単位）を適当に構築しながら，全長が最少となるグラフを決定するという問題として定式化できる．このある制約条件のもとでの最短グラフの構築という問題は，グラフ理論では，「Steiner 問題」として有名である．実際，最節約原理に基づく系統推定は一種の Steiner 問題としてみなされる．Sankoff and Rousseau (1975: 240) によれば，系統推定における Steiner 問題とは，OTU の形質状態の情報が与えられたときに，適当な仮想的分類単位 (HTU: Steiner 点とも呼ばれる) を分岐図の内部分岐点に配置することにより，全長が最小となる分岐図の全体的な樹形を決定する問題と定義できる．したがって，樹形と HTU 形質状態の同時推定を行なう必要がある．これは非常に困難な問題であり，計算機科学上の最難度問題群である「NP 完全」(NP-complete) というカテゴリーに属している (Graham and Foulds, 1982)．NP 完全問題は本質的には解決されていない（解決可能かどうかさえよくわかっていない）．しかし，合理的な計算時間の範囲で，最節約分岐図を完全枚挙あるいは発見的探索するためのアルゴリズムはよく研究されている．たとえば，完全枚挙による大域的最節約分岐図を発見するために，分枝限定法 (branch-and-bound method: Hendy and Penny, 1982) が開発されている．また，分枝限定法が使えないほど多くの OTU を含むデータに対しては，枝を交換することにより全長のより短い分岐図を探索するためのアルゴリズムがいくつか開発されている (Swofford and Olsen, 1990)．（最節約分岐図の構築を含む，コンピューターを用いた系統分析ソフトウェアについては，三中・斎藤 (1992) がそのリストを作った．）

しかし，そもそも分岐図とは何か，そのグラフはどのような構造を持ち，どんな情報を表現しているのかについては，必ずしも研究者間で同意が成立しているわけではない (Minaka, 1987)．また，ある形質データのもとで，可能なすべての分岐図の樹形がどのような頻度分布を示すのかについては，まだ研究が始まったばかりである (Hillis, 1991; Huelsenbeck, 1991)．これらの点について，私自身の考えもまじえながら論じる．

3.1 分岐図の枚挙と体系化

分岐図をある種の系統樹であるとみなすならば (Wiley, 1981 のように)，分岐図という概念そのものの存在価値はおそらくなくなるだろう．しかし，私は，分岐図と系統樹はそれぞれ異なる種類の関係を表示し

ていると考えている．分岐図は分類群間の包含関係を表すグラフであるのに対し，系統樹はOTU（種など）の間の祖先子孫関係を示すグラフである．分岐図と系統樹は，そこに図示されている関係の種類こそ異なっているものの，ともにある順序関係の構造を示すグラフであるという点ではまったく違いはない．分岐図や系統樹が順序構造であるという観点に立脚すると，それらの構造に対する半順序理論 (theory of partial orders) 的な定義が可能になる．いま，分類対象であるOTUの集合をSとし，Sの部分集合全体の集合を$p(S)$［Sの冪集合という］と書く．このとき分岐図Cとは次の条件を満たす$p(S)$の部分集合［C構造と呼ぶ］のグラフ（Hasse図）である（Minaka, 1987; 三中, 1991）：

1) $S \in C$ かつ $\phi \notin C$;
2) 任意の X, Y $(\in C)$ に対して $X \cap Y \neq \phi$ ならば，$X \subseteq Y$ または $Y \subseteq X$;
3) 任意の x, y $(\in S)$ に対して $x \neq y$ ならば，次の2条件を同時に満足する X, Y $(\in C)$ が存在する：
   条件1) $x \in X$ かつ $y \notin X$;
   条件2) $x \notin Y$ かつ $y \in Y$.

この定義に従うかぎり，枚挙の対象となる分岐図は，$p(p(S))$という有限集合の要素である．分岐図の半順序理論的モデルとしてのC構造は，集合論的な包含関係を前提として定義されている．この時点で，われわれは「分岐分析において何を枚挙するのか？」という問題に対するある解答を得た．すなわち「有限集合Sのうえで定義された包含関係による半順序集合$p(S)$の部分集合であるC構造全体からなる集合」の中での要素（C構造）の枚挙が焦点になる．

上のように半順序理論的に樹状図を定義すると，従来は明確に理解されていなかったいくつかの点を明らかにすることができる（Minaka, 1987; 三中, 1991）．

なお，亀井 (1992a, b) は，多次元空間内での超立方体（集合Boole代数と順序同型）の作図と解析のためのパソコン用プログラムを開発した．さらに，汎用数式処理ソフト *Mathematica*™ を用いれば，コンピューターを用いた一般の半順序構造の解析が可能だろう（Skiena, 1990 参照）．

3.1.1 分岐図による系統樹集合の分割

分岐図は対象となる生物群の部分集合すべてから成る冪集合のうえで定義された順序構造である（正確には分岐図はある順序構造のHasse図表現である）．一方，祖先子孫関係という別種の順序関係を考えると，その祖先子孫関係が定義される集合のうえでの順序構造が系統樹であると規定できる (Minaka, 1987). 分岐図集合と系統樹集合の間には準同型写像が定義でき，その結果分岐図集合の各要素は系統樹集合を直和分割することが証明された．したがって，分岐図が包含関係を系統樹が祖先子孫関係を表示するグラフであると考えるかぎり，両者の間に1対1の対応は存在しない．これは分岐図が構造的に系統樹よりも包含的なグラフであるという Nelson and Platnick (1981) の主張の論理的支持である．樹状図の定義については別のところに詳しく書いているので（Minaka, 1987; 三中, 1989），それらを参照されたい．

3.1.2 分岐図の集合 Boole 代数への埋め込み

冪集合上で定義される包含関係による可能な順序構造のうち，全体集合を要素として持つ順序集合（半束：semilattice）はその Hasse 図が分岐的な樹状図になる．したがって分岐図はある条件を満たす半束構造といえる．この半束構造にさらに空集合を付加して網状性を付与した構造を束 (lattice) といい，特に冪集合の全要素は集合 Boole 代数という束を形成する (Davey and Priestley, 1990). この集合 Boole 代数という概念は順序構造としての分岐図をさらに一般化したものであり，任意の分岐図はある集合 Boole 代数の部分構造として実現される．集合 Boole 代数はその Hasse 図が多次元空間内の超立方体であるため，分岐図もまたその多次元空間の中に埋め込まれることになる．集合 Boole 代数は分岐図の構造を分析する理論的基礎を与える．

現実のデータの解析においては，複数の分岐図の集合を考慮しなければならない場合が少なくない．例えば，同程度に最節約的な分岐図が得られた場合とか，所属不明の分類群がある場合，あるいは，多分岐を解釈する場合がそれに当たる（3.1.3 節を参照）．さらに，分断生物地理学 (vicariance biogeography) においては，複数の種分岐図 (species cladogram) の比較とそれらの持つ情報をふまえた地域分岐図 (area cladogram) の構築が目標である (Page, 1990). しかし，現在の分岐分類学は単一の分岐図の持つ情報を解析する手段は持っているが，複数の分岐図の集合全体が持つ情報を解析したり統合したりする理論体系をまだ構築していない．複数の分岐図から合意樹 (consensus tree) を計算するという方法もあるが，もとの分岐図集合の持つ情報はかなり失われてしまう（3.1.4 節を参照）．集合 Boole 代数を用いることは，解決への一つの手掛かりになると私は考える．この問題は，今後さらなる研究が期待される（行列を用いた複数の分岐図の同時表現法を開発している Ragan, 1992, in press をも参照されたい）．

3.1.3 分岐図の直和と直積

分岐図を多次元超立方体である集合 Boole 代数の一部分と考える利点の一つに，複数の分岐図に含まれる情報を統合できるという点が挙げられる．分類群が

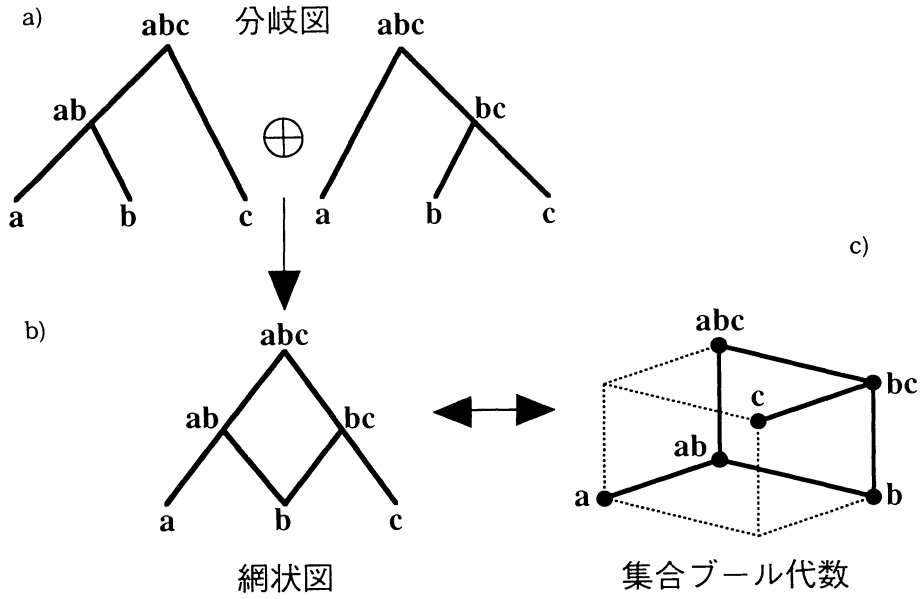
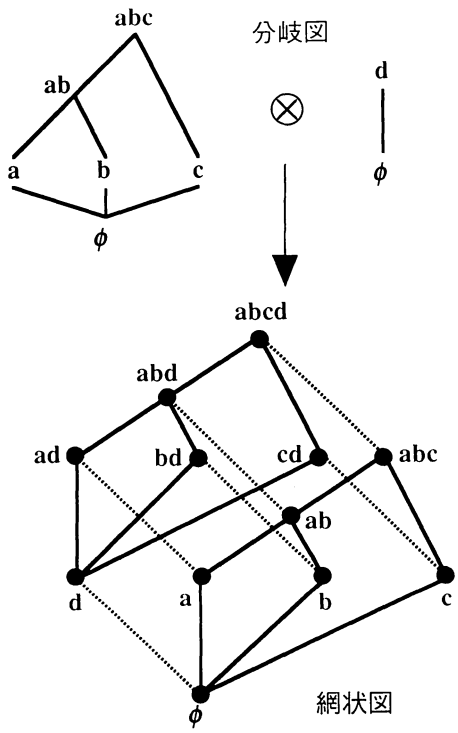


図 3. 分岐図間の直和演算．分岐図の内部分岐点は，末端点 (OTU) から成る部分集合である．たとえば，内部分岐点 ab は部分集合 {a, b} を意味する．a) 二つの異なる分岐図の直和演算．b) 直和演算の結果得られる網状図．c) 網状図を 3 次元集合ブール代数の中に埋め込む．直和演算は，同程度に最節約的な分岐図が複数個存在するとき，それらの持つ情報を統合するときなどに用いる．



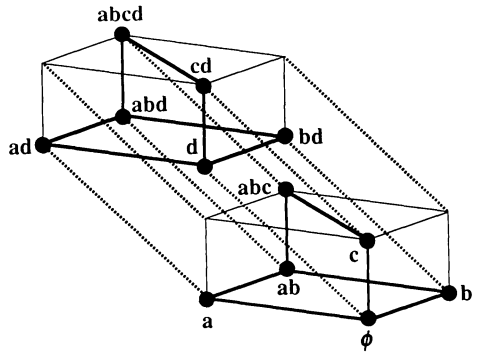


図 4. 分岐図間の直積演算．a, b, c, から成る分岐図と d のみから成る分岐図 (singleton) の間の直積演算．a) まずはじめにそれぞれの分岐図に空集合 (ϕ) を付加する．b) 二つの分岐図の間で包含関係の直積を計算する．c) 得られた結果を 4 次元集合ブール代数に埋め込む．直積演算は所属不明の種を既存の分類体系に組み込むときなどに用いられる．

完全に階層的であるならば，対応する順序構造の Hasse 図は分岐的であるが，非階層的である場合には Hasse 図は網状的になる．この網状 Hasse 図は任意の分類群の集合の順序構造を正確に記述できる図式表現である．しかも，交雑現象など特定の進化モデルを仮定する必要はまったくない．網状 Hasse 図をもとにして，二つの分岐図の構造を統合する直和演算（direct sum: 図 3）と分岐図に不確定要素を組み込む場合の直積演算（direct product: 図 4）を定義する (Minaka,

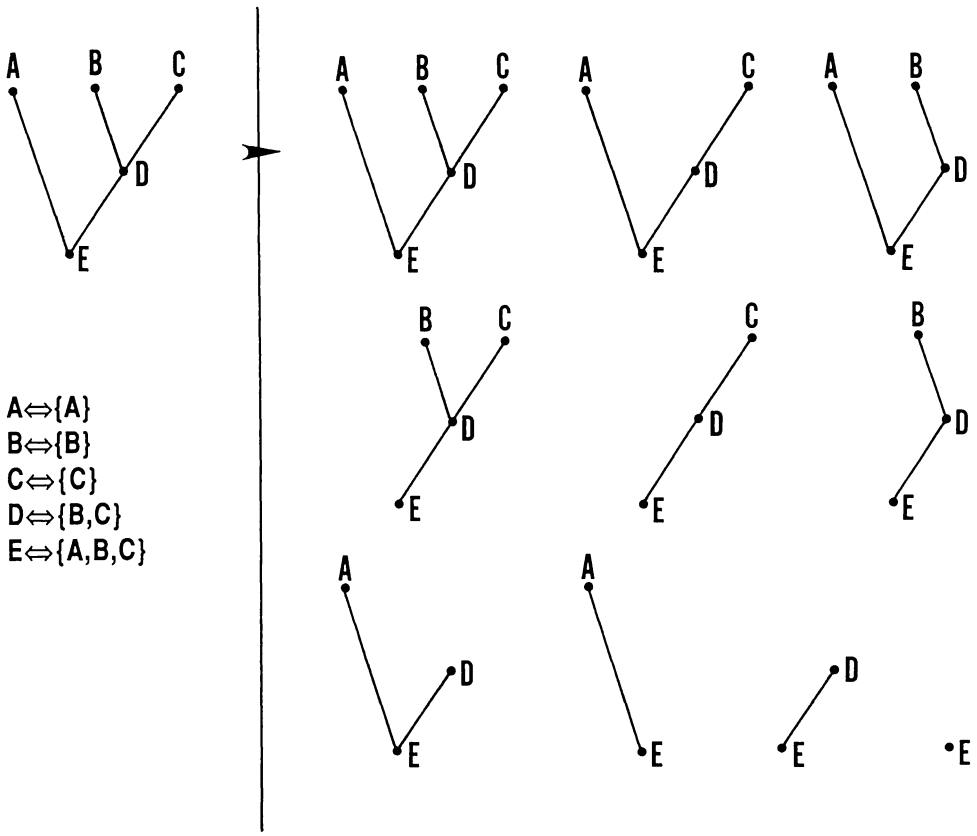


図 5. 分岐図の順序イデアル（上下逆転させた双対として表示している）．図左に示した分岐図の根 (E) を必ず含むすべての部分木（順序イデアル）を図右に示した．

1990)．直和は同程度に最節約的な複数の分岐図の情報を統合する場合に使える．一方，直積は帰属不明の分類対象を分岐図に組み込むときなどに使える．

3.1.4 全順序拡大に基づく分岐図の情報量

半順序理論は，また分岐図の順序情報量を測定することを可能にする．順序構造の情報量の尺度としては，全順序拡大 (linear extension: Stanley, 1986) が適当であるとされている (Atkinson, 1985, 1989)．全順序拡大を求めるには，まずはじめに，ある分岐図の順序イデアルをすべて求める（図 5 にはその双対を示した）．順序イデアルとは，分岐図の根（図 5 の E）を含む部分木である．次いで，得られた順序イデアル集合の Hasse 図を作成する（図 6）．この Hasse 図の最小元から最大元への最短経路の個数が全順序拡大の個数である．n 個の要素（部分集合）から成る順序構造の全順序拡大の最大個数は $n!$（反鎖に対応する）である．したがって，全順序拡大の個数を z とすると，

$$I = -\log_2 \frac{z}{n!}$$

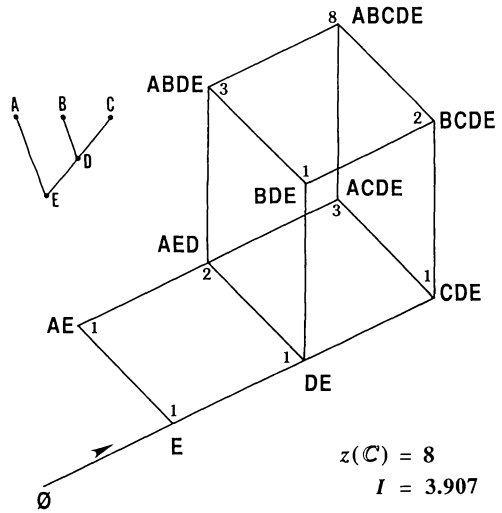
で定義される量 (Atkinson, 1985) は，その分岐図の順序情報量の尺度として用いられる．

この全順序拡大を用いることにより，個々の分岐図の順序情報量のみならず，網状 Hasse 図の形式に統合された任意の分岐図の集合全体の情報量を数値化することができる．例えば，従来から非階層的な分類群の配置の表示に用いられてきた多分岐的分岐図は，全順序拡大の尺度のもとでは，対応する網状 Hasse 図と比べて順序情報量が小さいという結論が得られる（三中，1993a)．

祖先子孫関係の構造を図示する系統樹は，上で定義した C 構造の派生構造として導かれる．また，最近の分子系統学では unrooted tree を作ることが多いが（三中・斎藤，1993)，このタイプのグラフはある条件を満たす rooted tree（例えば分岐図）の集合と見なすことができる．つまり，ある unrooted tree は，それを rooting することにより得られるすべての分岐図の集合であると定義できる．

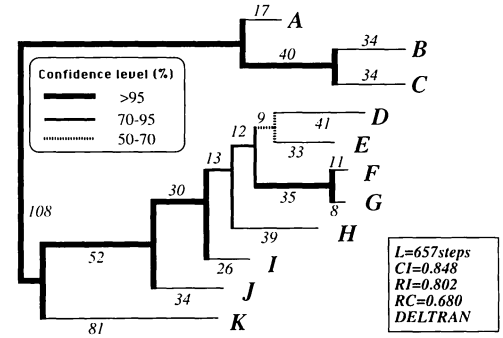
3.2 樹長分布の組合せ論的特性

分岐図の信頼性を統計的に評価する方法には，もとの形質データからの無作為再抽出に基づくブーツスト



$\mathcal{I}(C)$ = {ABCDE,ACDE,ABDE,
BCDE,CDE,BDE,AED,AE,DE,E,Ø}

図 6. 分岐図の全順序拡大と情報量．図 5 で示した順序イデアルの集合 $\mathcal{I}(C)$ に対する Hasse 図を作成する．最小元である空集合 (ϕ) から最大元であるもとの分岐図（"ABCDE" で表示）までの Hasse 図上の最短経路（極大鎖）の個数 $z(C)$ は，その分岐図の全順序拡大の個数に等しい．この図では $z(C)=8$ であるから，この分岐図の順序情報量は $I=-\log_2(8/5!)=3.907$ と計算される．



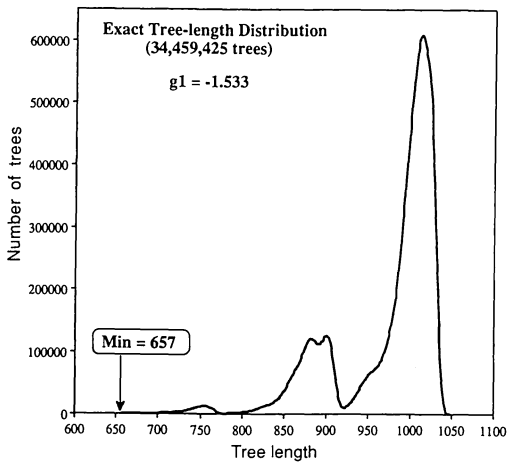


図 7. DNA 塩基配列に基づく藻類の最節約分岐図と樹長分布．a) 光合成の暗反応（Calvin 回路）において CO_2 の取り込みに関与する酵素 Rubisco（リブロース二リン酸カルボキシル化/酸素添加酵素）の大サブユニット (*rbcL*) の DNA 塩基配列に基づく最節約分岐図．データは Fujiwara *et al.* (to appear) を用い，種名は記号 (A～K) で表示した．DNA 塩基配列の大きさは 1491 塩基対であるが，実際の計算では GC 含量のバイアス補正のためアミノ酸配列に変換した配列データ（497 アミノ酸）を用い，分枝限定法により網羅的探索を行なった．分岐図の枝のタイプは 1000 回のブーツストラップに基づく信頼度を表し，各枝に付けた数値は DELTRAN 最適化配置（4.2 節参照）のもとでの各枝の長さ（置換数）である．b) このデータのもとでの樹長分布．OTU 数が 11 のときの可能なすべての分岐図 (unrooted tree で 34,459,425 個）の樹長を数え上げることにより作成した．計算は，PAUP version 3.0 (Swofford, 1990) による．

ラップ法 (Felsenstein, 1985b) や分岐図の枝の長さの頻度分布（樹長分布）などがある．前者は分岐図のある枝がどれほど再現性があるかを調べる方法であるのに対し，後者はもとの形質データにどれほどの系統学的情報が含まれているかを調べる方法である．ブーツストラップ法はこれまでも統計学的方法として広く用いられてきたが，樹長分布の解析はほとんど進んでいなかった．図 7 は，光合成植物の暗反応回路において CO_2 を取り込む反応を触媒する酵素 Rubisco の大サブユニット遺伝子の DNA 塩基配列に基づく分岐図の樹長分布である (Fujiwara *et al.*, to appear)．OTU 数が 10 程度ならば比較的短時間に全数調査に基づく正確な樹長分布曲線を作成することができる．また，OTU がそれよりも大きな場合には，分岐図の全体集合から無作為抽出することにより，樹長分布の形を推定すればよい．ブーツストラップ法は現在多くの系統分析ソフトウェアに組み込まれているが，樹長分布解析も PAUP version 3.0 (Swofford, 1990) を用いれば可能である．

3.2.1 樹長分布・同時分布・周辺分布

これまでこの種の樹長分布から得られる情報としては，分布曲線の歪度 (skewness) が有用であるといわれてきた (Hillis, 1991; Huelsenbeck, 1991)．n 個の分岐図から成る樹長分布の歪度は次の式で示される：

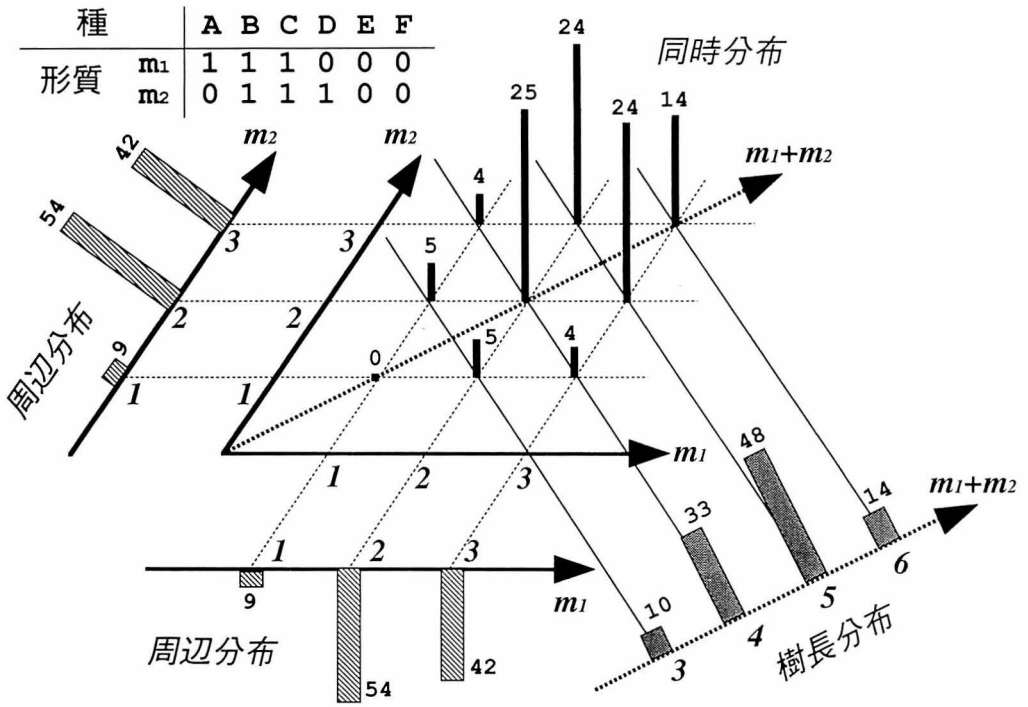


図 8. 樹長分布・同時分布・周辺分布の関係．6 個の OTU (A～F) に関して 2 個の二値的形質 m_1, m_2 が観察されたとする．可能な 105 個すべての分岐図 (unrooted) について，各形質に対する全長（1 以上 3 以下）を 2 次元の頻度分布で示したのが同時分布である．それぞれの形質軸に対してその同時分布を射影することにより周辺分布が得られる．さらに，m_1+m_2 軸に対して同時分布を射影すると樹長分布が得られる．同時分布における m_1-m_2 軸方向は同じ全長を持つ分岐図の構造的な差異を表すと解釈できる．しかし，樹長分布にはこの分岐図間の差異は反映されない．

$$g_1=\frac{\frac{1}{n}\sum_{i=1}^{n}(L_i-\bar{L})^3}{\left\{\frac{1}{n}\sum_{i=1}^{n}(L_i-\bar{L})^2\right\}^{\frac{3}{2}}}$$

L_i: i 番目の分岐図の全長

$\bar{L}$: n 個の分岐図の全長の平均

$g_1<0$ のとき分布は「左に歪んでいる」(left-skewed) といい，$g_1>0$ のとき「右に歪んでいる」(right-skewed) という．もとの形質情報が系統学的な情報を多く含むほど（ノイズが少ないほど），樹長分布が左に歪み，歪度はより大きな負の値を取る．樹長分布のこの性質は，数値的には調べられていたが，解析的研究はまだ始まったばかりである．

いま n 個の二値的形質 $1, 2, \ldots, n$ を考え，各形質のもとでの分岐図の枝の全長を確率変数 $m_1, m_2, \ldots, m_n$ と表す．第 i 形質の二値的状態がそれぞれ a_i と b_i 個の種に分布したとすると（a_i+b_i＝全種数），確率変数 m_i の分布関数 $f_i(a_i, b_i)$ は

$$f_i(a_i, b_i)=\frac{(m_i-1)!(2n-3m_i)(2n-5)!!}{(2n-2m_i-1)!!}\times N(a_i, m_i)N(b_i, m_i)$$

$$\text{ただし } N(x, y):=\frac{(2x-y-1)!}{(x-y)!(y-1)!\,2^{x-y}}$$

によって解析的に表現できる (Carter *et al.*, 1990)．この頻度分布曲線 f_i は n 形質に関するある同時分布 $f_{1,2,\ldots,n}$ の周辺分布であると規定される．つまり，任意の i に対して

$$f_i(a_i, b_i)=\sum_{\text{not } i} f_{1,2,\ldots,n}(a_1, a_2, \ldots, a_n; b_1, b_2, \ldots, b_n)$$

という関係式が導かれる．n 形質に関する分岐図の全長 L は，

$$L:=\sum_{i=1}^{n} m_i$$

と定義されるから，全長 L の確率分布は上の同時分布に従う n 変量 $m_1, m_2, \ldots, m_n$ の和の分布である．もっとも，現時点では同時分布や樹長分布の解析的表現式は求められていない．

樹長分布・同時分布・周辺分布の相互関係については，図 8 に示した仮想例を見られたい．二つの形質 m_1, m_2 についての同時頻度分布のもとで，樹長分布は m_1+m_2 方向の軸に沿った分布である．したがって，その法線方向（m_1-m_2 方向）の情報は樹長分布においては失われていることになる．ところが，この法線

方向の情報とは，同程度の最節約性を持つ分岐図（長さが同じ）の構造的差異を反映する情報である．例えば，図8の樹長分布で全長3という長さを持つ10個の同程度に最節約的な分岐図について，もとの同時分布ではその半数が $(m_1, m_2)=(1, 2)$ に属し，残る半数は $(m_1, m_2)=(2, 1)$ に属している．分岐図の構造のうえからいえば，同時分布上の点 $(1, 2)$ に属している5個の分岐図は (ABC) という枝（または clade）を共有しているのに対し，点 $(2, 1)$ に属している5個の分岐図は (BCD) という枝を共有している．明らかに，もとの同時分布を m_1+m_2 方向に射影して樹長分布を作ると，m_1-m_2 方向の情報は捨てられている．

与えられたデータのもとで同程度に最節約的な分岐図の間の構造的差異については，最近関心が高まっている．例えば，Hendy *et al.* (1988) は分岐図間の類似度 (symmetric difference metric) を用いて最節約分岐図間の表形的距離を測定している．Maddison (1991) は分枝交換の回数によって分岐図間の類似度を規定している．しかし，そのような1次元的な距離変量では分岐図間の類似性は間接的にしか表現できないと思われる．むしろ2次元以上の形質空間での直接的な多変量解析に基づく分岐図の構造分析を目指す必要があると私は考えている（三中，1992e, to appear).

3.2.2 制約樹長分布に基づく分岐図の信頼性評価

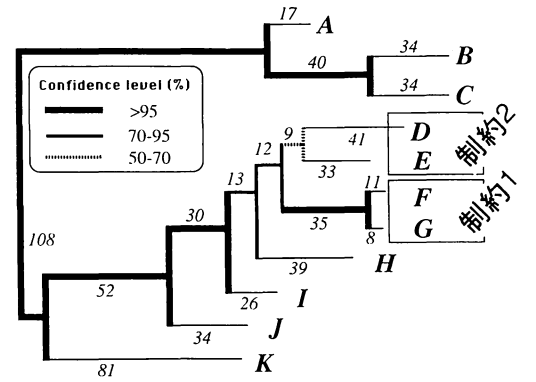
樹長分布に関係するもう一つの問題は，分岐図長の準最節約性である．最節約分岐図に近い全長を持つ分岐図はトポロジー的に互い類似する傾向がある（三中，1992b)．この事実は，最節約および準最節約分岐図の集合には共通の枝があることを示唆している．言いかえれば，それらの分岐図がよい成績をあげたのは最節約性の上で有利な枝を持っていたからであると考えられる．樹長分布を用いれば，ある分岐図の枝（または clade）の存在が，分岐図の全体集合の中で「平均的」に最節約性にどの程度貢献しているかを調べることができる（三中，1992d)．以下の議論では，樹長分布は全数調査によって決定されるものと仮定する．分岐図の全体集合からの無作為抽出の場合についてはここでは論じない．

いま，ある枝 G の存在をトポロジー的な制約条件とする分岐図の集合を考え，その制約付きの樹長分布を「G 制約分布」$f|_G$ と呼ぶことにする．G 制約分布のもとでの樹長分布 $f|_G$ の平均 E と無制約分布 f（もとの樹長分布）の平均を比較することにより，その枝 G の最節約性への貢献度を計ることができる．つまり，その枝の存在を樹形上の制約とする樹長分布の平均値が，もとの無制約樹長分布の平均値よりも小さいならば，与えられたデータのもとでその枝は分岐図の最節約性に対して平均的に正の貢献をしたと判定される．ここで，この最節約性への平均的貢献度を枝 G の「強度」(strength) と呼び，$\Delta(G)$ と表すことにする．この $\Delta(G)$ は上の定義により次式で表現される：

$$\Delta(G) := E_{f|_G}(L) - E_f(L)$$

ただし $E_g(L)$ という表現法は，g という樹長分布のもとでの分岐図の全長 L の平均値を意味する．$\Delta(G)<0$ のとき枝 G は正の貢献，一方 $\Delta(G)>0$ のとき枝 G は負の貢献をしている．

この強度 $\Delta(G)$ を尺度として任意の分岐図に含まれるそれぞれの枝の最節約性に対する平均的貢献度を計算すると，最節約的な分岐図ではどの枝も大きな強度（絶対値の大きな負の値）を持つのに対し，非最節約的な分岐図では各枝の強度は小さい（正の値または絶対



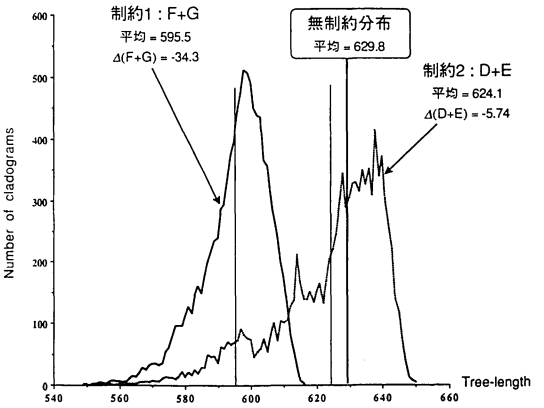


図9. 分岐図に対する構造的制約．a) 図7aの分岐図の二つの枝の強度を調べる．枝 $F+G$ および枝 $D+E$ をそれぞれ制約1および制約2と呼ぶ．ブーツストラップ信頼度（bootstrap replicates における再現率：3.2.2節参照）では，枝 $F+G$ は枝 $D+E$ よりも信頼度が著しく高いことに注意されたい．b) 制約1および制約2のもとでの樹長分布．各枝の強度 Δ を見ると，枝 $F+G$ は，枝 $D+E$ と比較して，平均的に最節約性に対してより大きな正の貢献をしているといえる．

値の小さな負の値）傾向が認められる（三中，1992 d）．図9は，前出の図7の分岐図において，樹形上の二通りの制約を置いた場合の制約樹長分布である．制約1 ($F+G$) および制約2 ($D+E$) のもとでの制約樹長分布の平均値は，それぞれ595.5, 624.1である．したがって，無制約分布での平均値629.8と比べると，枝1と2の強度はそれぞれ $\Delta(F+G)=-34.3$, $\Delta(D+E)=-5.74$ となる．この結果は，$F+G$ という枝（制約1）は最節約性に対して平均的に大きな貢献をしている「強い」枝であるが，$D+E$ という枝（制約2）は平均的貢献度の小さい「弱い」枝であることを意味する．

ここで，上で述べた樹長分布解析とブーツストラップ分析の比較をしておく（図10）．両者はともに分岐図の枝の「妥当性」を量的に評価するという共通の目的を持っている．しかし，その評価の方法が異なっている．第一の違いは，データの扱いである．ブーツストラップ分析では，もとの形質データをある仮想的な形質集合の「代表者」（無作為標本）と仮定し，その形質データを母集団とみなして形質の再抽出を繰り返す．その結果得られた派生的データのそれぞれについて最節約分岐図（bootstrap replicateと呼ぶ）を作成し，それらの分岐図における枝の出現率を計算してその枝の信頼性の尺度とするわけである．一方，樹長分布解析では，もとのデータはまったくいじらない．第二の違いは，調査対象となる分岐図の最節約性の程度である．ブーツストラップ分析では最節約的な分岐図だけを調べる．一方，樹長分布解析では，もとのデータに対して最節約的・非最節約的のいかんを問わずすべての分岐図の樹長を調査する．

両者のこれらの違いを考えるならば，樹長分布解析とブーツストラップ分析は相異なる観点から分岐図の枝の妥当性を評価しているものといえるだろう．それぞれの手法の長所と短所については，今後さらに考察する必要がある（三中，1992f, 1993b）．

4. ある分岐図のもとでの仮想的形質状態の復元

全長最小化という最節約基準を満たす分岐図の樹形が決定されたとき，第二の問題として仮想的分類単位(HTU) での仮想的形質状態の決定という問題がある．すなわち，OTU の形質状態の情報が与えられ，同時に分岐図の樹形も与えられているときに，内部分岐点に位置する HTU の最節約的な仮想的形質状態配置を決定することである．ある分岐図の樹形のもとで，その全長を変化させない，HTU に対する仮想的形質状態の最節約的配置を枚挙することは，樹形の決定それ自体とはまた別のタイプの組合せ論的問題である．もちろん，すでに述べた NP 完全である最節約分岐図の樹形決定の問題と比べれば，この仮想的形質状態の決定問題は難度が低い．それでもなお，グラフ理論のうえでまだ未解決の問題が多く残されている．

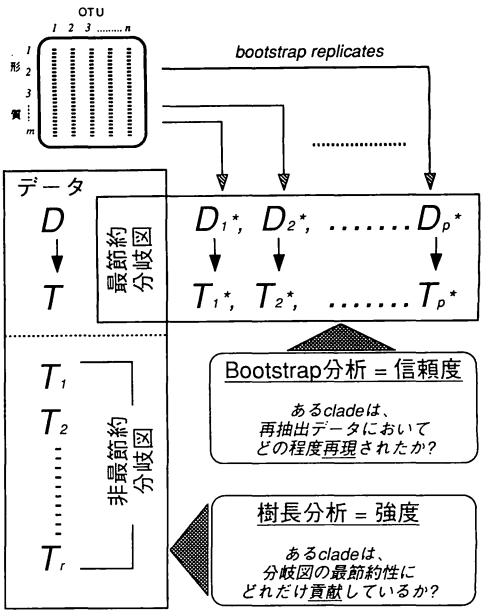


図10. 樹長分布解析とブーツストラップ分析．ブーツストラップ分析では，もとのデータからの再抽出を繰り返し行ない，分岐図の枝の再現率をもって信頼度の尺度とする．樹長分布解析では，もとデータに対する最節約・非最節約分岐図の全長を調べ，ある枝の存在が分岐図集合全体の最節約性に対して平均的にどの程度の貢献をしているかをもって強度の尺度とする．

(Minaka, 1992; Hanazawa *et al.*, 1992, to appear; 成嶋ほか，1993)．

4.1 仮想的形質状態の復元アルゴリズム

初期の分岐分析の理論では，仮想的形質状態の決定と最節約分岐図の決定は単一の問題だった．実際，Hennig (1966) の系統分析理論や Farris (1970) の Wagner tree 理論では，最節約系統樹が決まると同時に HTU での仮想的形質状態も決定されていた．確かに，それらの理論に基づく仮想的形質状態配置は最節約的配置の一つではある．しかし，可能な最節約的配置のすべてを網羅するものではなかった (Swofford and Maddison, 1987)．樹形決定と配置決定を相異なる二つの問題であると認識することが，まずはじめに必要である（三中，1992c）．

以下では，形質として離散的形質状態をとるものだけを考える．しかし，この離散的形質の中には，形態学的形質によく見られる「順序的」(ordered) な形質と DNA の塩基配列データに典型的に見られる「非順序的」(unordered) な形質の二つのタイプがある．たとえば，a→b→c という形質変換系列を持つ順序的形質を考えると，状態 a から c への変化のためには必ず中間状態 b を経由しなければならない．したがって，

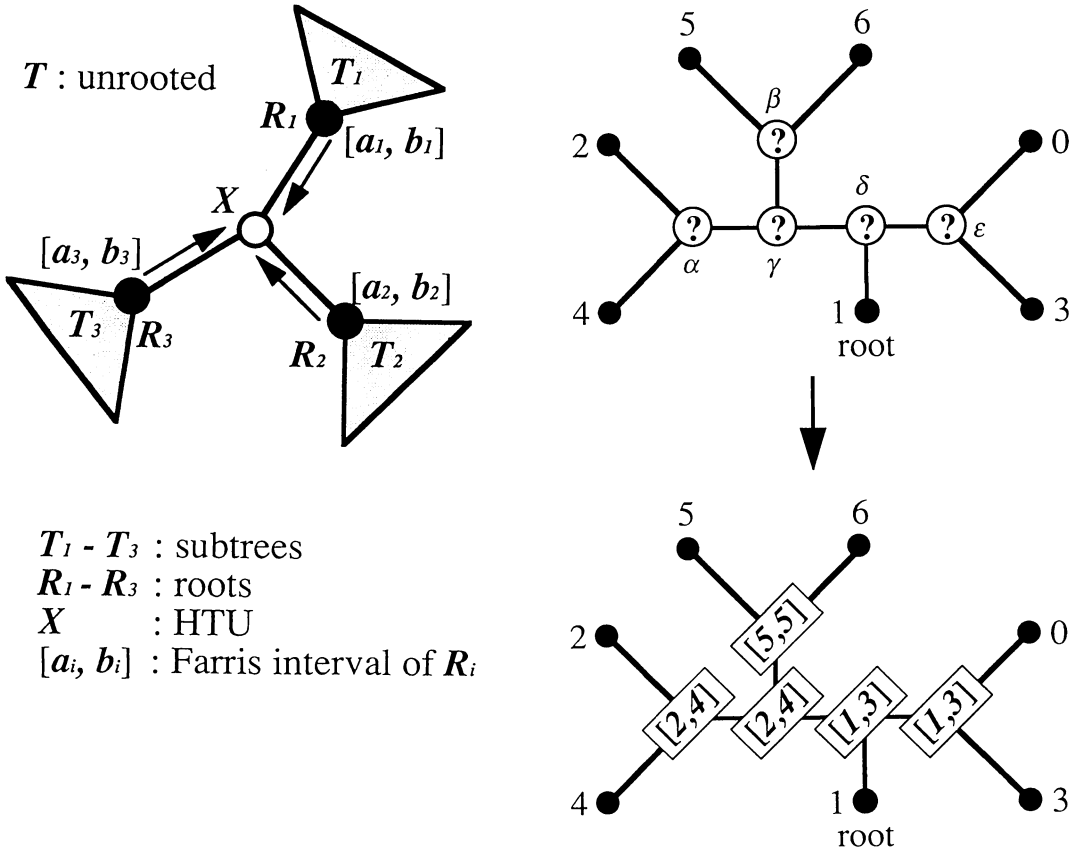


図 11. 仮想的形質状態配置の網羅的枚挙．a）順序的形質に関する Hanazawa *et al.* に基づく HTU 形質状態の復元アルゴリズム．もとの分岐図のある HTU(X) を参照点としてそれに隣接する HTU（または OTU: 図では R_1〜R_3）を起点とする部分木 (T_1〜T_3) を考える．各部分木の末端 (OTU) から開始して逐次的に HTU の Farris 区間を構築していく．X に隣接する R_1〜R_3 の Farris 区間が決定された後，それらの区間のメジアンを計算すれば X における MPR set（可能なすべての仮想的形質状態の集合）が求められる．b) HTU 形質状態の復元の実例（Swofford and Maddison, 1987 の例）．5 個の HTU (α〜ε) に対する MPR set（閉区間として表示した）が復元された．

状態 a から c への変化に要する変化回数は 2 である．一方，その形質が非順序的であるならば，任意の状態は他の任意の状態に 1 ステップで変化できる．

上で指摘したように，Hennig の系統分類理論や Farris の Wagner tree 理論では，可能な最節約的配置を網羅的には発見できない．ここでの問題は，OTU と樹形に関する与えられた情報のもとで，すべての最節約的な形質状態配置パターンを枚挙することである．順序的形質については，Swofford and Maddison (1987) が理論的に追究している．しかし，彼らの議論は二分岐的系統樹に限定されており，また定理の証明も部分的に不完全であった．Hanazawa *et al.* (1992, to appear) は，Swofford and Maddison (1987) の議論を敷衍し，任意の多分岐的分岐図のもとでの HTU 復元に関する一般的定理の証明と簡便な復元アルゴリズムの開発を行なった．

このアルゴリズム（図 11 参照）は，次の 2 ステップから成る (Hanazawa *et al.*, to appear):

ステップ 1）各 HTU に対する特性区間の決定:

ある HTU に関してもとの分岐図の根を取り除き，unrooted tree を作る．次いで，末端 OTU からその指定された HTU にいたるまでのすべての HTU の「Farris 区間」(Swofford and Maddison, 1987) を逐次的に作成する．その結果，指定された HTU に隣接する HTU のそれぞれに対する Farris 区間が決まる．指定された HTU に対して可能な最節約的な形質状態配置の集合（特性区間: Swofford and Maddison, 1987 のいう "MPR set"）は，それに隣接する HTU の Farris 区間すべての中央値（メジアン）である．このステップ 1 は，末端の OTU から指定された HTU に向かう「求心的」

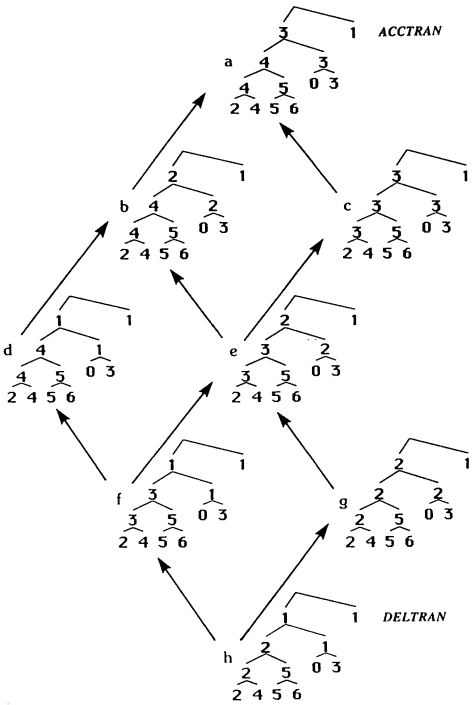


図 12. 仮想的形質状態配置の代数的構造．図 11b の例での特性区間 (MPR set) に基づくすべての HTU 形質状態配置パターンの集合に半順序関係▶を導入する．ACCTRAN 配置と DELTRAN 配置はこの Hasse 図の中でそれぞれ最大元と最小元に相当する．

プロセスである．

ステップ 2) 各 HTU に対する最節約的な形質状態配置の決定:

ある一つの HTU について，その特性区間が決定されたならば，その特性区間に属する任意の形質状態を指定することにより，他のすべての HTU における最節約的な形質状態配置を中央値として再帰的に決定できる．このステップ 2 は，指定された HTU から末端の OTU に向かう「遠心的」プロセスである．

非順序的形質については，Fitch (1971), Hartigan (1973), Rinsma *et al.* (1990) などの研究がすでにある．Hanazawa *et al.* (to appear) のアルゴリズムは，少し変更するだけで，この非順序的形質の最節約的配置をも扱えるものと思われる．

4.2 仮想的形質状態配置の束論的体系化

仮想的形質状態の配置パターンの枚挙の問題では，複数個の分岐点における仮想的形質状態の組である形質状態ベクトルの全体集合（ベクトル空間）が枚挙の舞台となる．Swofford and Maddison (1987) は，HTU 形質状態配置の枚挙の過程で得られた複数の配置パターンの集合において，次の二つの対照的な配置パターンの存在を指摘した（図 12 参照）:

DELTRAN 最適化配置（形質変換遅延最適化配置）

形質状態の変化ができるだけ分岐図の末端近くで生じるように HTU の形質状態を決定する．その結果，形質の収斂回数が最大化される．

ACCTRAN 最適化配置（形質変換促進最適化配置）

形質状態の変化ができるだけ分岐図の根本近くで生じるように HTU の形質状態を決定する．その結果，形質の逆転回数が最大化される．これは，Farris (1970) のアルゴリズムに基づく配置方法である．

Swofford and Maddison (1987) の論文では順序的形質だけが議論されていたが，塩基配列データのような非順序的形質にも同様の ACCTRAN/DELTRAN 的配置が存在する（三中・斎藤，1993）．

しかし，この ACCTRAN/DELTRAN 的配置が，枚挙されたすべての HTU 形質状態配置のベクトル空間の中でどのように特徴づけられるのかという問題はまだ未解決である．この問題へのアプローチとして，このベクトル空間に要素間の大小関係に基づく以下の半順序関係▶を導入する（三中，1991 付録参照）: ある与えられた分岐図の n 個の HTU における仮想的形質状態を要素とする二つの n 次元ベクトル

$$\boldsymbol{a}=(a_1, a_2, \ldots, a_n),\ \boldsymbol{b}=(b_1, b_2, \ldots, b_n)$$

に対して，

$$\forall i,\ a_i \leq b_i \Leftrightarrow \boldsymbol{a} \blacktriangleright \boldsymbol{b}$$

により半順序関係▶を定義する．この半順序関係▶を導入した n 次元ベクトル空間には最大元と最小元の存在し，それはそれぞれ ACCTRAN 最適化ベクトルと DELTRAN 最適化ベクトルに対応することが予想されている．これは，その n 次元ベクトル空間が束を形成しているのではないかという予想である．

仮想的形質状態のベクトル空間が半順序関係▶に関して構造化できたならば，ACCTRAN/DELTRAN 的配置を含むすべての仮想的形質状態の配置パターンを体系化することができるだろう．具体的にいうと，ある分岐図の参照分岐点に対して根に向かう方向の局所的 ACCTRAN/DELTRAN 化および末端に向かう方向の局所的 ACCTRAN/DELTRAN 化という二種類の形質復元方法を考えることができる．図 12 は，図 11 で復元した特性区間 (MPR set) に基づいて，すべての可能な HTU 形質状態配置パターンを表示したものである．図の中の矢印は▶に基づく半順序関係である．

仮想的形質状態配置は，最近の進化生態学で問題になっている「比較法」(comparative method) と密接に関係している（三中，1991）．形質進化の説明を行

なう場合，系統関係の影響は無視できない (Felsenstein, 1985a)．自然選択の観点から形質進化を説明するに当たって系統の情報をどのように組み込むかに関しては，いくつも提案がなされている (Pagel and Harvey, 1988; Harvey and Pagel, 1991)．ここでは，分岐図の樹形を利用する比較法 (Maddison, 1990) に伴う問題点を以下で指摘したい．

分岐図の上である形質がどのように進化したかをトレースするためには，HTU 形質状態配置をまず決定しなければならない．例えば，Maddison (1990) の提唱する形質の相関性の検定は，対象となる二つの形質の分岐図上での仮想的形質状態配置が前もって与えられていることを前提とする．その際，HTU 配置が異なるときに検定結果がどの程度影響を受けるのかという問題は，詳しく調べる必要がある．分岐図を用いた比較法を行なうに当たっては，複数の最節約分岐図および複数の最節約 HTU 形質状態配置という二つの問題に直面する．第一の最節約分岐図の複数解の問題は，例えばそれらの分岐図の構造的類似性に基づく要約 (Maddison, 1991)，あるいは分岐図の信頼性または強度に基づく重みづけによる絞り込み（青木重幸コメント）によって，解決されるかもしれない．また，第二の問題については，形質進化に関する事前情報に基づく HTU 配置の重みづけによって，あるいは分子レベルの配列データの場合には適当な形質進化モデルのもとでの HTU 配置の尤度を計算することにより，より少数の配置パターンを選択できるようになるかもしれない（三中ほか，1993)．また，シミュレーションによる HTU 配置の影響の評価 (Maddison, 1990) も行なう必要があるだろう．これらに関しては，さらに研究すべき課題が山積している．

5. 有限性との格闘

本稿では，分岐分析に関連するいくつかの組合せ論的問題を論じた．分岐図の枚挙あるいは HTU 形質状態配置の枚挙は，本質的に組合せ論的枚挙の問題である．Stanley (1986) が指摘するように，組合せ論的枚挙がある有限集合の要素を枚挙することであるならば，枚挙されるべき対象と枚挙の行なわれる場をつねに明示することが重要だろう．分岐図の枚挙は $p(p(S))$ という有限集合の上で行なわれる．一方，HTU 形質状態配置の枚挙は OTU 形質状態と分岐図の樹形によって制約された形質ベクトル空間の上で行なわれる．分岐図や系統樹は逐次的に構築すべきものであって枚挙すべきものではないという見解もありうるだろうが，それは系統分類学における樹状図が，何らかの順序関係に基づいて構築された離散的な構造であるという本質を見逃している．OTU という離散的な存在を相手にしてその離散構造を追究する以上，枚挙は避けられない．ある分岐図や HTU 形質状態配置が生物学的にありえないといった議論が成立するのは，それらの離散構造の枚挙が完全に行なわれることを前提としている．この点は，特定の分岐図および HTU 形質状態配置を前提にして形質進化を論じる比較法にとりわけよく当てはまる．

有限なものをすべて数え上げるという行為は，ともすれば連続的な量の変化を扱う研究に比べて低く見られているらしい（秋山，1991)．しかし，有限性にまともに取り組むと深刻な問題が生じる．最節約分岐図を枚挙する有効なアルゴリズムの構築が不可能に近いという宣告を下している NP 完全性はその代表である．系統分類学者は，この枚挙の問題をもっと真剣に考えるべきだろう．もちろん，系統分類学に付随する組み合わせ論的問題の解決は，狭義の系統分類学者には荷が重すぎるかもしれない．もっと多くの離散数学研究者がこの分野に関心を向けることを期待したい．離散数学研究者にとって「おいしい問題がそこにも，ここにも」という秋山 (1991) の言葉は，現在の系統分類学の状況を的確に言い表している．

現代の系統分類学は，従来用いられてきた形態的形質ばかりでなく，近年急速に蓄積されてきた DNA など遺伝子レベルの配列データをも同時に考察しなければならない．しかし，たとえ大量の分子データを用いたとしても，それに基づいて系統推定を行なうときには，解決を要するさまざまな組合せ論的問題が横たわっている．系統推定の問題は用いるデータの種類（分子か形態か）によって解決されるのではない．どのような方法論によって分岐図を構築するのか，そしてその方法論に付随する数学的問題をどのように解決するのかという点こそ重要なのである．

謝　辞

本稿は，1991 年 12 月 8 日に千葉県立中央博物館で開催された第 3 回自然誌シンポジウム「生物の進化と生物地理」における私の講演「分岐図の多次元幾何学：分類学および生物地理学における言語としての順序理論」の内容をふまえたものである．本稿をまとめるに当たり，以下の方々の見解を参考にさせていただいた．この場を借りてお礼を申し上げたい（50 音順・敬称略)：青木重幸（立正大・教養・生物）・太田邦昌（松戸市）・粕谷英一（新潟大・教育・生物）・亀井喜久男（岐阜東高校）・斎藤成也（国立遺伝研・進化遺伝）・直海俊一郎（千葉県立中央博）・成嶋　弘（東海大・理・情報数理）・長谷川英祐（都立大・理・動物生態)．

引 用 文 献

秋山　仁．1991．離散数学のすすめ．数理科学 1991(9): 10–13.

Atkinson, M. D. 1985. Partial orders and comparison

problems. Congr. Numer. 47: 77–88.
Atkinson, M. D. 1989. The complexity of orders. *In* Rival, I. (ed.), Algorithms and Order, pp. 195–230. Kluwer, Amsterdam.
Carter, M., M. Hendy, D. Penny, L. A. Székely and N. C. Wormald. 1990. On the dustribution of lengths of evolutionary trees. SIAM J. Discr. Math. 3: 38–47.
Davey, B. A. and H. A. Priestley. 1990. Introduction to Lattices and Order. Cambridge Universtity Press, Cambridge.
Farris, J. S. 1970. Methods of computing Wagner trees. Syst. Zool. 19: 83–92.
Felsenstein, J. 1983. Parsimony in systematics: biological and statistical issues. Ann. Rev. Ecol. Syst. 14: 313–333.
Felsenstein, J. 1985a. Phylogenies and the comparative method. Amer. Nat. 125: 1–15.
Felsenstein, J. 1985b. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39: 783–791.
Felsenstein, J. 1988. Phylogenies from molecular sequences: inference and reliability. Ann. Rev. Genet. 22: 521–565.
Felsenstein, J. and E. Sober. 1986. Parsimony and likelihood: an exchange. Syst. Zool. 35: 617–626.
Fitch, W. M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. 20: 406–416.
Fujiwara, S., N. Minaka, H. Iwahashi, J. Someya and S. Nishikawa. Structure and contranscription of the plastid-encoded *rbcL* and *rbcS* genes of the marine haptophyte, *Pleurochrysis carterae*. Submitted to J. Phycol.
Graham, R. L. and L. R. Foulds. 1982. Unlikelihood that minimal phylogenies for a realistic biological sutdy can be constructed in reasonable computational time. Math. Biosci. 60: 133–142.
Hanazawa, M., H. Narushima and N. Minaka. 1992. A method for generating all most parsimonious reconstructions on a given tree: a problem originated in cladistics. Abstracts of the Fifth Franco—Japanese Days on Combinatorics and Optimization (University of Kyoto, Kyoto). (no pagination)
Hanazawa, M., H. Narushima and N. Minaka. Generating most parsimonious reconstructions on a tree: a generalization of the Farris-Swofford-Maddison method. Submitted to Discr. Appl. Math.
Hartigan, J. A. 1973. Minimum mutation fits to a given tree. Biometrics 29: 53–65.
Harvey, P. H. and M. D. Pagel. 1991. The Comparative Method in Evolutionary Biology. Oxford University Press, Oxford. 〔粕谷英一訳, 進化生物学における比較の方法. 北海道大学図書刊行会, 札幌 (近刊)〕
Hendy, M. D. and D. Penny. 1982. Branch and bound algorithms to determine minimal evolutionary trees. Math. Biosci. 59: 277–290.
Hendy, M. D., M. A. Steel, D. Penny and I. M. Henderson. 1988. Families of trees and consensus. *In* Bock, H. H. (eds.), Classification and related methods of data analysis, pp. 355–362. Elsevier, Amsterdam.
Hillis, D. M. 1991. Discriminating between phylogenetic signal and random noise in DNA sequences. *In* Miyamoto, M. M. and J. Cracraft (eds.), Phylogenetic Analysis of DNA Sequences, pp. 278–294. Oxford University Press, New York.
Huelsenbeck, J. P. 1991. Tree-length distrubution skewness: an indicator of phylogenetic information. Syst. Zool. 40: 257–270.
亀井喜久男. 1992a. 多次元立方体の表現. 数理科学 1992 年 1 月号, pp. 68–73.
亀井喜久男. 1992b. HYPERCUBE EXHIBITION version 5.11. コンピューター・ソフトウェア. 著者からの配布.
Maddison, W. P. 1990. A method for testing the correlated evolution of two binary characters: are gains or losses concentrated on certain branches of a phylogenetic tree? Evolution 44: 539–557.
Maddison, D. R. 1991. The discovery and importance of multiple islands of most-parsimonious trees. Syst. Zool. 40: 315–328.
Minaka, N. 1987. Branching diagrams in cladistics: Their definitions and implications for biogeographic analyses. Bull. Biogeogr. Soc. Japan 42: 65–78.
三中信宏. 1989. 現代進化生物学における分岐分類学: Hennig 以降の理論展開とその積極的評価. 種生物学研究 13: 18–44.
Minaka, N. 1990. Cladograms and reticulated structures: A proposal for graphic representation of cladistic structures. Bull. Biogeogr. Soc. Japan 45: 1–10.
三中信宏. 1991. 分岐図の科学と行動生態学との接点および「真正分類系統学」の誤謬. 昆虫分類学若手懇談会ニュース 60: 1–51.
三中信宏. 1992a. Phylogenetic forest に分け入る: 分岐分析における組合せ論的諸問題. 第 23 回種生物学シンポジウム (松本), 講演要旨・資料集, p. 45.
三中信宏. 1992b. 最節約分岐図間の類似性と情報尺度. 第 47 回日本生物地理学会大会 (東京), 講演要旨集, pp. 11–12.
三中信宏. 1992c. 分岐図上での仮想的形質状態の復元: 最節約原理・形質進化・比較法. シンポジウム「分子系統学と生物分類学との接点」(国立遺伝学研究所, 三島), 講演要旨集 (no pagination).
三中信宏. 1992d. 樹長分布に基づく分岐図の信頼性評価. 第 52 回日本昆虫学会大会・第 36 回日本応用動物昆虫学会大会 (弘前大学, 弘前), 合同大会講演要旨, p. 101.
三中信宏. 1992e. 系統推定法と形質情報の保存. 第 57 回日本植物学会大会シンポジウム「陸上植物の系統: データの総合と展開」(帝塚山短大, 奈良), 研究発表記録, p. 60.
三中信宏. 1992f. 新しい誤差推定法とコンピュータによるその応用. 第 1 回計測と情報解析研究会「コンピュータが拓く情報解析」(農環研, つくば), 講演要旨集, pp. 67–84.
Minaka, N. 1992. Theoretical cladistics and phylogeny reconstruction: Multiple solutions under the principle of parsimony. Abstracts of the 29th International Geological Conference, Kyoto, p. 346.
三中信宏. 1993a. 系統推定の諸問題: 統計学的・組合せ論的視点から. 農環研年報 9: 47–54.
三中信宏. 1993b. 豊穣なる猥雑: 日本植物分類学会との遭遇. 日本植物分類学会ニュースレター 70: 6–19.
Minaka, N. A comment on the properties of tree-length distribution. (to appear)

三中信宏・成嶋　弘・花沢正純. 1993. 系統樹上の仮想的共通祖先の復元について: グラフ理論に基づく再帰的アルゴリズム. 日本計量生物学会 1993 年度年会 (統計数理研究所, 東京), 要旨集.
三中信宏・斎藤成也. 1992. 系統樹作成のためのソフトウェアリスト, 第 1 版. 著者からの配布.
三中信宏・斎藤成也. 1993. 系統樹作成法: 最大節約法. 日本生化学会 (編) 新生化学実験講座第 16 巻　分子進化実験法, pp. 380–395. 東京化学同人, 東京.
成嶋　弘・花沢正純・三中信宏. 1993. Generating most parsimonious reconstructions on a tree: 数理生物学からの問題. 日本数学会 1993 年度年会応用数学分科会講演アブストラクト, pp. 109–112.
Nelson, G. and N. Platnick. 1981. Systematics and Biogeography: Cladistics and Vicariance. Columbia University Press, New York.
Page, R. D. M. 1990. Component analysis: a valiant failure? Cladistics 6: 119–136.
Pagel, M. D. and P. H. Harvey. 1988. Recent developments in the analysis of comparative data. Quart. Rev. Biol. 63: 413–440. 〔農環研分類学研究会訳, 「比較データの分析における最近の進歩」〕
Ragan, M. A. 1992. Phylogenetic inference based on matrix representation of trees. Mol. Phylog. Evol. 1: 53–58.
Ragan, M. A. Matrix representation in reconstructing phylogenetic relationships among the eukaryotes. BioSystems. (in press)
Rinsma, I., M. Hendy and D. Penny. 1990. Minimally colored trees. Math. Biosci. 98: 201–210.
Sankoff, D. and O. Rousseau. 1975. Locating the vertices of a Steiner tree in an arbitrary metric space. Math. Progr. 9: 240–246.
Skiena, S. 1990. Implementing discrete mathematics: combinatorics and graph theory with *Mathematica*™. Addison-Wesley, Redwood.
Sober, E. 1983. Parsimony in systematics: philosophical issues. Ann. Rev. Ecol. Syst. 14: 335–358.
Sober, E. 1985. A likelihood justification of parsimony. Cladistics 1: 209–233.
Sober, E. 1988. Reconstructing the Past: Parsimony, Evolution, and Inference. The MIT Press, Massachusetts. 〔三中信宏訳, 系統分類学の基礎: 最節約原理・進化過程・統計的推論. 蒼樹書房, 東京 (近刊)〕
Stanley, R. P. 1986. Enumerative combinatorics, volume 1. Wadsworth & Brooks/Cole, Monterey. 〔成嶋　弘・山田　浩・渡辺敬一・清水昭信訳, 数え上げ組合せ論 I (1990), 日本評論社, 東京〕
Swofford, D. L. 1990. PAUP: Phylogenetic Analysis Using Parsimony, version 3.0. Computer software and documentation distributed by the author.
Swofford, D. L. and W. M. Maddison. 1987. Reconstructing ancestral character states under Wagner parsimony. Math. Biosci. 87: 199–229.
Swofford, D. L. and G. J .Olsen. 1990. Phylogeny reconstruction. *In* Hillis, D. M. and C. Moritz (eds.), Molecular Systematics. pp. 411–501. Sinauer Ass., Sunderland.
Wiley, E. O. 1981. Phylogenetics: the Theory and Practice of Phylogenetic Systematics. John Wiley & Sons, New York. 〔宮　正樹・西田周平・沖山宗雄共訳, 系統分類学: 分岐分類の理論と実際 (1991), 文一総合出版, 東京〕
Wiley, E. O., D. Siegel-Causey, D. R. Brooks and V. A. Funk. 1991. The Compleat Cladist: a Primer of Phylogenetic Procedures. Univ. Kansas Mus. Nat. Hist., Spec. Publ. 19, Lawrence. 〔宮　正樹訳, 系統分類学入門: 分岐分類の基礎と応用 (1992), 文一総合出版, 東京〕

Parsimony, Phylogeny and Discrete Mathematics: Combinatorial Problems in Phylogenetic Systematics

Nobuhiro Minaka

National Institute of Agro-Environmental Sciences, Kannondai 3-1-1, Tsukuba, Ibaraki 305, Japan

Phylogeny estimation based on the principle of parsimony has several conceptual and computational problems, which can be approached with the aid of discrete mathematics (combinatorics, graph theory and partial order theory). In this paper I will discuss three combinatorial problems in the cladistic method in biological systematics and phylogenetics.

The first problem arises in searching the most parsimonious cladogram(s) under a given character state data matrix. If cladogram is defined to be a particular type of finite set partially ordered by inclusion relation among taxa (Minaka,1987), any set of cladograms can be represented in a single, partially reticulated, Hasse diagram. Phylogenetic tree is conceptually different from cladogram because the former is a partially ordered set defined by another different relation, ancestor-decendent relation. Moreover, any set of cladograms can be embedded in a Boolean algebra. It enables us to define two cladogram operations (direct sum and direct product: Minaka, 1990) and to measure the order information of one or more cladograms. Direct sum of cladograms generates a single, partially reticulated, graph embedded in a Boolean algebra whose dimension is equal to the number of OTUs of each cladogram. Algebraically interpreted, such a reticulated graph includes all component information of the original set of cladograms. Direct product of cladograms also produces a single reticulated graph. But in this case the resultant graph can be embedded in a Boolean algebra whose dimension is higher than that of any of the original cladograms. These cladogram operations can be utilized in combining the information of more than one cladogram succinctly. An additional merit of adopting the partial-order concept of cladogram is that the information content of one or more cladogram can be quantified as the

relative number of linear extensions. The number of linear extensions associated with a cladogram is calculated from the Hasse diagram constructed from the set of all order ideals of the cladogram.

The second problem is associated with tree-length distribution. Some combinatorial properties of tree-length distribution are discussed with reference to the joint and marginal distribution of the lengths of cladograms. The joint distribution of the lengths of cladograms is a frequency distribution in a multidimensional space spanned by character axes. Tree-length distribution is derived from a projection of the joint distribution onto the axis of total length. The reduction of dimensionality associated with the projection implies the loss of information. In fact tree-length distribution is insensitive to the topological difference among equally parsimonious cladograms, because the direction of the projection is orthogonal to the axis of total length such that the topologically separated clusters of cladograms with equal total length are mixed together.

The third problem is concerned with how to reconstruct in a parsimonious way the hypothetical character-states assignable to the hypothetical taxonomic units (HTUs) at the interior points of a cladogram. Farris (1970) and Swofford and Maddison (1987) proposed a method for HTU reconstruction for an ordered character. Recently Hanazawa, Narushima and Minaka (1992, to appear) developed another, more efficient, recursive algorithm for HTU reconstruction. They proved and generalized several propositions about parsimonious HTU reconstruction, some of which were mentioned without full proof in Swofford and Maddison (1987). When there exist two or more equally parsimonious reconstructions of HTU character states, the set of all parsimonious HTU reconstructions can be treated as a vector space (lattice) partially ordered by a binary relation. ACCTRAN and DELTRAN reconstructions are supposed to be the maximal and minimal elements, respectively, in the vector space. The way of reconstructing the character states of HTU may affect the compararive analyses in behavioral ecology and sociobiology.