Vol. 40 No. SIG 5(TOD 2) 情報処理学会論文誌 May. 1999

異種系統樹間の調停のためのゼロ交差制約の充足

北上 始[†] 森 康真[†] 太田聡史[+] 斎藤成也[+]

二つの異種系統樹データベースから解析に必要な二つの部分木を検索し，それらの部分木から一つの調停木を作成することは，生物種の分子進化学的な研究を行うために有用である．この調停木の作成・利用を効果的に行うために，それらの二つの部分木に対して木の葉節点列がお互いに一致する順序木を見つけることが大変重要である．それらの二つの順序木はゼロ交差制約を満足しており，木構造同士を比較研究する上で便利な順序づけになっている．本論文では，検索された二つの部分木（異種系統樹）からゼロ交差制約を満足する二つの順序木を探索する方法が提案されている．ゼロ交差制約の充足は，葉節点列間に結合行列を定義し，その結合行列に対して，あるヒューリスティックを用いた木探索を行うことにより達成されている．この木探索では，木の枝の間に交差が発生するのを回避するために，葉節点に関するクラスタを定義し，クラスタ同士の交換に基づいた葉節点列の順序づけを行っている．

Satisfying the Zero-Crossover Constraints for Reconciliation across Heterogeneous Trees

HAJIME KITAKAMI,[†] YASUMA MORI,[†] SATOSHI OOTA[+] and NARUYA SAITOU[+]

After searching two subtrees from two heterogeneous tree databases, a reconciled tree found from two heterogeneous subtrees is useful for understanding biological diversity, researching gene duplications, reconstructing taxonomic trees, and assigning a taxonomic name to each branch node of gene trees. It is very important to find two ordered trees with the same sequence of leaf nodes in order to achieve an effective reconciliation. For the reconciliation, two ordered trees that satisfy the zero-crossover constraint are useful for comparing the two heterogeneous trees. This paper proposes a new method for searching for two ordered trees that satisfy the zero-crossover constraint. This is achieved using a heuristic tree search for an interconnection matrix, which is defined by the leaf sequences (layers) of the two trees. To avoid crossovers between the branches of either tree, the order of each leaf sequence is modified in the tree search. The search orders the leaf sequence using swap operations between two clusters with respect to leaf nodes. The method is implemented in Prolog and the implementation results also are presented.

1. はじめに

１９５３年にWatsonとCrickによりDNAの二重螺旋構造が発見されて以来，生命現象の基盤はDNAにたたみこまれた遺伝情報であることが明らかにされた．DNA（Deoxyribonucleic acid）は，RNAウイルスを除く全ての生物がもっている遺伝子情報物質である．DNAは，長い紐状の分子で，その中にA（アデニン），T（チミン），G（グアニン），C（シトシン）という４種類の塩基が並んでいる．１９７０年代中頃の「組み換えDNA技術」は，ヒトを含む多様な生物のDNA解析を可能にするという道を原理的にひらいた．その後，ヒトゲノム解析という新しい科学技術的挑戦の時代が到来し，今から１０年前にヒトゲノム解析計画[1]が始動するに至った．ヒトゲノム解析計画では生物学や医学全般の発展に多大な貢献を果たすとともに，「人間あるいは生物とは何か」にこれまでとまったく違う知見が得られるものと期待されている．ヒトゲノム解析計画を推進するために，ヒトを含む動植物やバクテリアなどの多くの生物に対するDNAデータベース[2),3),4),5)]等が地球規模で整備されてきている．そのようなDNAデータベースを利用すれば，少ない実験データである特定の生物種についての解析を行うような伝統的な生物学研究スタイルから脱却し，体系的かつ網羅的にゲノム解析を行うことができる．DNAデータベースには，付録1[6)]にも示されているように，多くの生物種に関する生物の設計図ともいえるDNA配列データが蓄積され

†広島市立大学情報科学部知能情報システム工学科
Department of Intelligent Systems, Faculty of Information Sciences, Hiroshima City University
＋国立遺伝学研究所進化遺伝学研究部門
Laboratory of Evolutionary Genetics, National Institute of Genetics

ているだけでなく生物種の分類情報（生物分類樹）なども蓄積されている．生物分類樹[7),8)]は，多様な生物種を形態学的に分類することによって決定された情報である．このような状況の中で，DNA データベースを用いたゲノム解析の研究では情報科学の重要性が益々高まっており，有用なソフトウェアの研究開発が期待されている．

DNA 配列データには，（1）過去に起きた生物の進化に関する情報や（2）病気に関する遺伝子情報などが含まれている．それらを含む DNA データベースを解析することにより，生物の進化や起源をはじめとして，病気の原因遺伝子に対する同定や特定の形質との結びつきなども解明することができる．例えばこれまでに，カンブリア紀頃に多細胞動物が爆発的に多様化した理由，ヒトで代表される高等動物の脳の起源，現代人の起源（どこから来たのか），ヒトと病原性ウイルスの進化的かかわり，アルツハイマー症で代表される老人病，ベーチェット病や慢性関節リウマチなどの生活習慣病などについて，多くの知見が得られるようになってきた[9),10),11)]．

著者らは，DNA データベースに含まれている生物分類樹データベースおよび DNA 配列データから計算・収集された分子進化系統樹データベースの両者[12),13),14),15)]を用いて，両者から検索された二つの部分木の調停[16),17),18),19)]に着目している．以後，両者のどちらも系統樹と呼び，木構造として扱う．分子進化系統樹は，DNA データベースの DNA 塩基配列の違いから推測された木であり，生物種の進化的な繋がりが表現されている[20)]．従来の分子進化系統樹を作成するアルゴリズムには，進化の過程で発生する遺伝子重複[注1)]の現象が十分に反映されておらず，生物分類樹などの異なった立場で構築された系統樹との比較検討が重要であると言われている．この他に，宿主と寄生生物の間の進化的かかわりを調べるために，両者の分子進化系統樹間の調停に関する研究も行われている[19),21)]．以上から，調停処理においては，生物分類樹データベースと分子進化系統樹データベースの組み合わせで代表されるように，お互いに構造が異なる二種類の系統樹データベースが用意される．そして，予め利用者が定めた複数の種名を検索キーとして，それらが葉節点になる二つの部分木を検索し，二つの部分木に対する調停が行われる．調停前は一方の部分木から他方の部分木への対応づけが1対nであるのに対して，調停により両者は1対1に対応づけられる（詳しくは付録2を参照されたい）．これにより，進化の過程で遺伝子重複が何時発生したのかを推定することができると同時に，それに基づいてさらに精度の高い分子進化系統樹を作成することができる．これにより，DNA データベースの従来の解析から得られていた知見よりもさらに進んだ知見が得られるものと期待されている．最近の面白い進化学研究では，陸上の哺乳類からずっと以前に分かれたと信じられていたクジラが，実はもっと新しい年代に分かれており，カバに近そうだということが報告されている[10)]．また，調停に関する他の研究として，宿主（例えば多種の鼠）と寄生生物（例えば多種の蚤）の二つの分子進化系統樹を対象にした研究があるが，そのような研究により，ある寄生生物がある年代でその宿主を突然変えてしまうなどの状況が報告されている[21)]．計算機上で調停処理を行うアルゴリズムがいくつか紹介されている．しかし，そのアルゴリズムに生物学的知識が十分に反映されていないこともあり，調停処理を専門家の経験に基づいて手作業で行う場合もある．特に，手作業で調停を行う場合，検索された二つの部分木がともに交差がないと同時に葉節点列同士が同じであれば，混乱なく調停作業を行うことができる．これは交差のない二つの順序木の作成を意味する．そのような両順序木が作成されれば，調停結果を生物学的に考察する時にも，理解が大変容易になる．

以上から本論文では，お互いに構造が異なる生物分類樹データベースと分子進化系統樹データベースから調停用に検索された二つの部分木を対象に，利便性のある二つの順序木を作成する方法について提案する．調停のために利便性のある二つの順序木は，それ自身に交差がないと同時に，二つの木が有する葉節点列の間にも交差がないという制約を満足している．ここでは，そのような交差がないという制約をゼロ交差制約と呼び，主にその制約が充足可能な二つの順序木に着目する．

以下，2章では，異種系統樹データベースとして生物分類樹データベースと分子進化系統樹データベースの両者について概観する．3章では，本論文で扱う順序木に関するデータモデルを明らかにするため，順序木の構造，ゼロ交差制約，基本操作について述べる．4章では，両順序木に対するゼロ交差制約を充足させる方法について明らかにし，5章で Prolog による実装結果について述べる．6章で関連研究について述べた後，7章で本研究の応用例を紹介する．また，8章で

注1) これには次の2種類がある[20)]．(1)進化の過程で，生物種が分化することによって，分化した種間で同じ遺伝子をもつような自然な場合．(2)進化の過程で，同じ生物種の中で，ある遺伝子だけが何度も複製されてしまう場合．

本論文のまとめを行う．

2. 異種系統樹データベース

二つの異種系統樹の組合せには，(1)生物分類樹と分子進化系統樹，(2)異なる分子進化系統樹などがある．前者の組合せは，生物分類樹の視点による分子進化系統樹の有効利用や分子進化系統樹の視点による生物分類樹の見直し[17]に重要であり，後者は，宿主と寄生生物の間の進化的関係を調べるのに重要である．後者の例としては，多種類の鼠を対象にした分子進化系統樹とそれらに寄生する蚤の分子進化系統樹とを分子進化学的に比較検討する研究がある[21]．

生物分類樹は生物種の形態学的分析の結果として得られ，分子進化系統樹はDNA塩基配列の進化学的分析の結果得られる．生物種の分子進化学的な変化が形態変化に直接反映されると仮定すれば，生物分類樹と分子進化系統樹の木構造は一致するはずだが，実際には遺伝子重複などがあるため，一般には両者の木構造は一致しない．両者の葉節点集合は全く同じ集合であるが，木のトポロジーは異なる．以下，生物分類樹データベースと分子進化系統樹データベースの両者について概観する．

図１の例[22),23)]に示されるように，生物分類樹データベースは生物種を形態学的な観点から分類することによって作成されてきた木構造のデータベースである．各葉節点で１つの生物種が表現され，生物種の分類は段階的に葉節点から上位の非葉節点で行われている．葉節点の深さは２５～３０段程度あり，非葉節点には分類名及び深さの目安となる階級（レベル）名が付与されている．代表的な階級名を上から順に示すと，界・門・綱・目・科・属・種などがある．近年の分子進化学の発達もあって，生物分類樹データがかなり見直されるようになってきている．現在，最新の生物分類樹データベースは米国を中心とした国際 DNA データバンク[24),25),26),27)]で管理されている．

分子進化系統樹データベースは，複数の生物種について，それらの DNA 塩基配列間を比較することによって計算された類縁関係を収集したものである．これによってDNA塩基の時間的変化をみることができる．その計算方法には，距離行列法，最大節約法，最尤法などが提案されている[20),28)]．図２は，距離行列法によって計算された分子進化系統樹の例である．分子進化系統樹は木構造データであり，葉節点に生物の種名が表現されている．分子進化系統樹では，非葉節点で種の進化による分岐を表現し，節点間を結ぶ枝の長さで進化の時間的な変化が表現されている．図２において，左側の共通祖先から右側の現存生物へ木をたどることは，過去から現在へ生物進化の流れを追跡することに対応する．

最近，研究者が着目する生物種の集合に対する生物分類樹や分子進化系統樹をインターネット経由で容易に参照することができるようになってきている[12),13),14),15),23)]．

3. データモデル

本章では，二つの異種系統樹データベースおよびそれらから検索される二つの部分木はいずれも同じデータモデルで表現されるとする．また，異種系統樹データベースを順序木とみなしたときの順序は，データの格納順によって決められているとする．別の見方をすると，この順序により，検索時にデータが取り出される順番が決定される．さらに，ここでは，ゼロ交差にできる問題だけを扱うものとする．以下，順序木の構造を定義した後，交差のない順序木を解くためのゼロ交差制約充足問題を定義し，その制約充足問題を解くために重要な基本操作について述べる．

3.1 順序木の構造

本質的な部分を浮き彫りにするために，高さ $n-1$ の木は正規化されているものとする．すなわち，ある

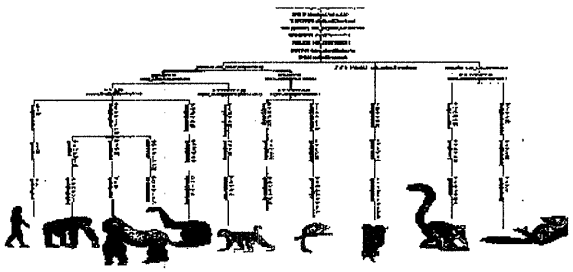


図１ 生物分類樹の例．
Fig. 1 An example of a species tree

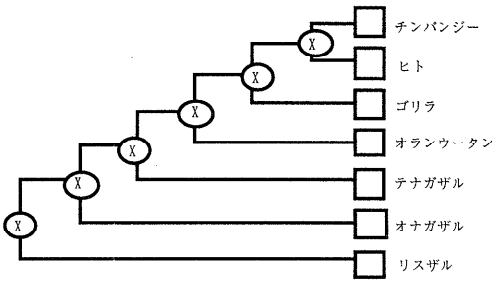


図２ 分子進化系統樹の例．
Fig. 2 An example of a gene tree

葉の深さが$n-1$よりも小さいとき，深さが$n-1$になるように途中にダミーの節点が追加されているものとする．そのような木を定義するために，あるnレベル階層（$n \geq 2$）の有向グラフ$T = g(V, R, n)$[29),30),31)]を考える．この有向グラフが下記の条件を満足するとき，その有向グラフTを順序木と呼ぶ．ただし，Vはある節点の集合であり，RはV上の2項関係$V \times V$の部分集合として定義される辺の集合である．

【条件1】Vは次のようにn個の部分集合に分割される．

$V = N_1 \cup N_2 \cup \cdots \cup N_n$ $(N_i \cap N_j = \phi,\ i \neq j,\ 1 \leq i \leq n,\ 1 \leq j \leq n)$．

以下，N_iをレベルiの節点集合と呼び，nを木の高さと呼ぶ．

【条件2】N_1の要素数$|N_1|$は1である．以下，この節点要素をTの根と呼ぶ．

【条件3】Rは，次のように$n-1$の部分集合に分割される．

$R = B_1 \cup B_2 \cup \cdots \cup B_{n-1}$，$(B_i \cap B_j = \phi,\ i \neq j)$，$B_i \subset N_i \times N_{i+1}$，$1 \leq i \leq n-1$．

このとき，同じ終点を有する任意の2つの辺(d_1, e)，$(d_2, e) \in R$に対して，$d_1 = d_2$を満足する．以下では辺(d, e)において，始点dを親節点と呼び，終点eを子節点と呼ぶことにする．

【条件4】入次数がゼロなる節点の集合はN_1だけである．また，出次数がゼロなる節点集合はN_nだけであり，N_nの各要素を葉節点と呼ぶことにする．

【条件5】レベルiの節点集合N_iに存在する全ての節点に対して，ある順序列が与えられている．すなわち，節点を$d^{(i)}_p \in N_i$で表現すると，順序列は$\sigma_i = d^{(i)}_1 d^{(i)}_2 \cdots d^{(i)}_\alpha$で表現される．ただし，$1 \leq p \leq \alpha$，$\alpha$は$N_i$の節点数を意味する．以後，この順序列の順序関係を2項関係$d^{(i)}_1 <_B d^{(i)}_2$，$d^{(i)}_2 <_B d^{(i)}_3$，…，$d^{(i)}_{p-1} <_B d^{(i)}_p$で表現し，$n$レベル階層グラフを$g(V, R, n, \Sigma)$で表現する（$\Sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$）．

3.2 ゼロ交差制約

葉節点数が同じ二つの順序木$T_1 = g(V_1, R_1, n_1, \Sigma_1)$，$T_2 = g(V_2, R_2, n_2, \Sigma_2)$において，一方の木の葉節点集合から他方の木の葉節点集合への全単射（1対1）が与えられているものとする．例えば，生物分類樹データから分子進化系統樹データへの全単射においては，同じ名前の葉節点名（生物の種名）同士が対応付けられている．ここでは，木T_jの枝同士が交差しない条件をT_jのゼロ交差制約と呼ぶ．また，二つの木T_1, T_2の葉節点レベル間を対応させることにより定義される辺集合を考えたとき，その中の辺同士が交差しない条件をT_1, T_2間のゼロ交差制約と呼ぶ．以下，これらの二種類の制約を総称して，ゼロ交差制約と呼ぶ．

順序木T_kのiレベルにおける節点集合及び順序列を$N_{k,i}$，$\sigma_{k,i}$で表現し，ゼロ交差制約の充足問題（CSP）[32),33)]を定義すると，次のようになる．

(1) 変数　：$X_{k,i} = (x^{(i)}_{k,1}, x^{(i)}_{k,2}, \ldots, x^{(i)}_{k,\gamma})$

(2) 変数の領域：

$D_{k,i} = \{ (d^{(i)}_{k,1}, d^{(i)}_{k,2}, \ldots, d^{(i)}_{k,\gamma}) \in N_{k,i} \times N_{k,i} \times \ldots \times N_{k,i} \mid d^{(i)}_{k,p} \neq d^{(i)}_{k,q},\ 1 \leq p < q \leq \gamma \}$

これは，$X_{k,i}$の取り得る値であり，iレベルにおける可能な順序列$\sigma_{k,i}$の集合でもある．

(3) 制約　：

$P^k_{i,i+1}(X_{k,i}, X_{k,i+1}) \wedge P^{1,2}_{leaf}(X_{k,\alpha}, X_{k,\beta})$

前者はT_kのゼロ交差制約であり，後者はT_1, T_2間のゼロ交差制約である．

ただし，$1 \leq k \leq 2$，$1 \leq i \leq n_k - 1$である．また，α, βは各々n_1, n_2であり，γは$N_{k,i}$の要素数である．さらに，節点の順序列$\sigma_{k,i} = d_1^{(i)} d^{(i)}_2 \ldots d^{(i)}_\gamma$を決める順序関係は，$\gamma$項関係$(d^{(i)}_1, d^{(i)}_2, \ldots, d^{(i)}_\gamma) \in D_{k,i}$の並びの順番により定義されているとする．上記(3)の二種類の制約は以下のとおりである．

【T_kのゼロ交差制約$P^k_{i,i+1}(X_{k,i}, X_{k,i+1})$】

順序木T_kのどのレベルiにおいても，$X_{k,i} = (x^{(i)}_{k,1}, x^{(i)}_{k,2}, \ldots, x^{(i)}_{k,\gamma})$，$X_{k,i+1} = (x^{(i+1)}_{k,1}, x^{(i+1)}_{k,2}, \ldots, x^{(i+1)}_{k,\delta})$に対して，$(x^{(i)}_{k,p}, x^{(i+1)}_{k,r}), (x^{(i)}_{k,q}, x^{(i+1)}_{k,s}) \in B_i$の二要素が，以下を満足する．

$x^{(i)}_{k,p} <_B x^{(i)}_{k,q} \Rightarrow x^{(i+1)}_{k,r} \leq_B x^{(i+1)}_{k,s}$．　(CS_1)

ただし，$1 \leq p < q \leq \gamma$，$1 \leq r < s \leq \delta$であり，γ，δは$N_{k,i}$，$N_{k,i+1}$の要素数である．

【T_1, T_2間のゼロ交差制約$P^{1,2}_{leaf}(X_{k,\alpha}, X_{k,\beta})$】

葉節点レベルの$X_{1,\alpha} = (x^{(\alpha)}_{1,1}, x^{(\alpha)}_{1,2}, \ldots, x^{(\alpha)}_{1,\eta})$，$X_{2,\beta} = (x^{(\beta)}_{2,1}, x^{(\beta)}_{2,2}, \ldots, x^{(\beta)}_{2,\xi})$に対して，以下を満足する．

$X_{1,\alpha} = X_{2,\beta}$．　(CS_2)

ただし，η, ξはT_1, T_2の葉節点数を表わす．

図3に制約ネットワークを示す．ゼロ交差制約の充足問題は，図の三つの表を結合することにより解くことができる．図中の三つの表について左から順に説明しよう．左の表中の各タプルは，T_1に関する複数の順序木の中で(CS_1)を満たすような順序木を表わす．すなわち，それは制約$P^1_{12} \wedge P^1_{23} \wedge \cdots \wedge P^1_{\alpha-1,\alpha}$を満たす集合である．ここでは，そのような順序木が多数存在するので，そのような集合を表で表現している．中央の表は，(CS_2)を満たすT_1とT_2の葉の組合せ集合を表わしている．すなわち，それは制約$P^{1,2}_{leaf}$を満たす集合である．右の表は，(CS_1)を満たすT_2の組合

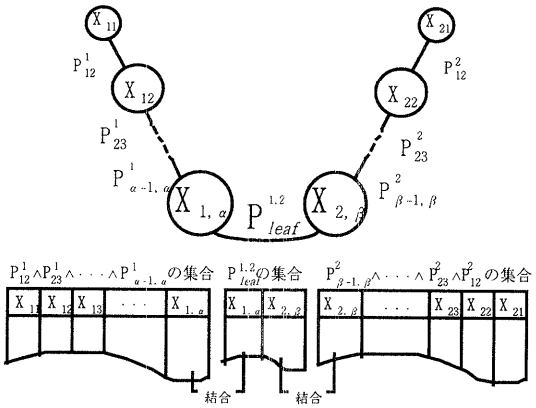


図3 制約ネットワーク

Fig. 3 An example of constraint network

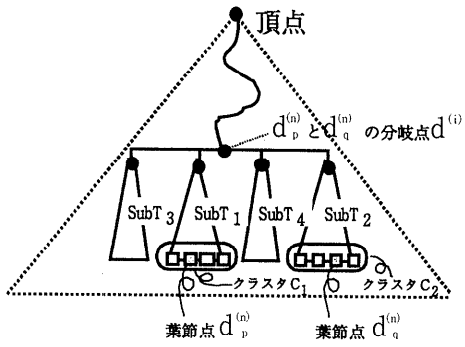


図4 葉節点に関するクラスタの探索

Fig. 4 An example of searching two leaf clusters

せ集合を表わしている．すなわち，それは制約 $P^2_{\beta-1,\beta} \wedge \cdots \wedge P^2_{23} \wedge P^2_{12}$ を満たす集合である．三つの表は与えられた木 T_1 と T_2 から計算されるが，三つの表は巨大なデータになるため，三つの表を各々計算することは明らかに非現実的である．

このゼロ交差制約の充足においては，(CS_1) と (CS_2) の間の制約伝藩が重要であると考える．制約伝藩は，(1) 変数 $X_{1,\alpha}$ と $X_{2,\beta}$ に関する探索により制約 (CS_2) を充足する解の構成及び，(2)制約 (CS_1) を満たすように順序木の更新操作を平行に行うことで達成され得る．これには，(CS_1) を満たす各順序木の初期値が必要だが，初期値は根節点から順に幅優先探索 [34]を行うことで計算できる．

3.3 木構造に対する基本操作

ここでは，ゼロ交差制約の充足に有用な二つの基本操作について述べる．第一は，ある系統樹データベースからゼロ交差の順序木を見つけるための幅優先探索であり，第二は，葉節点列において，ある葉節点を含む部分木を考え，与えられた二つの葉節点を区別するのに有用な二つの最大部分木の探索である．以下では，そのような最大部分木が有する葉節点集合をクラスタと呼ぶ．

ある順序木からゼロ交差制約 (CS_1) を満たす順序木を見つけるための幅優先探索は次のように定義される．前節 3.1 の条件5で格納順により節点の順序 Σ' が定められた順序木 $T' = g(V,R,n,\Sigma')$ を考えよう．このとき，Rの部分集合Sの各要素を起点としたRに関する子検索 $R \triangledown S$ を，$R \triangledown S = \{(b, c) \mid (b, c) \in R, \exists (a, b) \in S\}$ で定義すると，関係Sを起点としたRに関する下方探索の最小不動点を $Z^* = S \cup (R \triangledown Z^*)$ で定義することができる．起点をSとするRでの下方探索の最小不動点 Z^* を計算する方法は以下のとおりである [23]（これとは反対に，上方探索も容易に定義可能である）．

```
Z* := S ;  Z' := S ;
while ( Z' ≠ φ ) do
    Z' := R ▽ Z' ;
    Z* := Z* ∪ Z' ;
end ;
output (Z*) ;
```

ここで，Sを木の根節点を始点とする辺の集合とすると，Z^* として木の辺集合が得られる．この計算過程において，同じ親を始点とする複数の辺が順序木T'から検索されるが，その検索により取り出される順番により順序列 Σ を定めると，ゼロ交差制約を満足する順序木 $T = g(V,R,n,\Sigma)$ を得ることができる．

次に，図4を用いて，二つの葉節点 $d^{(n)}_p, d^{(n)}_q \in N_n$ を区別する二つのクラスタを探索する方法について述べる．葉節点の集合 N_n を有する順序木 $T = g(V,R,n,\Sigma)$ は，ゼロ交差制約 (CS_1) を満足しているとしよう．また，探索される二つのクラスタは，二つの葉節点 $d^{(n)}_p, d^{(n)}_q \in N_n$ を区別する最大クラスタ C_1, C_2 であるとする．ただし，$d^{(n)}_p \in C_1$, $d^{(n)}_q \in C_2$ であり，$C_1 \cap C_2 = \phi$ である．葉節点 $d^{(n)}_p$ のクラスタとは，$d^{(n)}_p$ を含む部分木の葉節点集合を指す．以上の性質を有する二つの最大クラスタ C_1, C_2 は，以下の手順により探索することができる．

(1) 上方探索により，二つの葉節点 $d^{(n)}_p, d^{(n)}_q \in N_n$ の分岐点 $d^{(i)}$ を見つけ，それに連結される複数の部分木から葉節点 $d^{(n)}_p$ を含む部分木 $SubT_1$ 及び葉節点 $d^{(n)}_q$ を含む部分木 $SubT_2$ を探索する．ただし，$i < n$ であり，$SubT_1$ と $SubT_2$ の節点

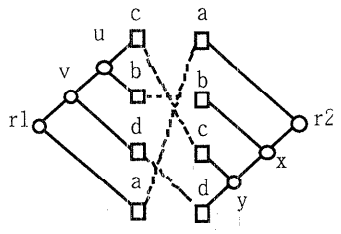


図5 ある順序づけされた異種系統樹

Fig. 5 An example of ordered trees with crossovers

には共通節点がないとする．また，SubT_1 と SubT_2 は，各々，考えられる部分木の中で最大の部分木であるとする．

（2）下方探索により，二つの部分木 SubT_1 と SubT_2 の各々に対する葉節点の集合C_1, C_2を計算する．ただし，$d^{(n)}_p \in C_1$, $d^{(n)}_q \in C_2$, $C_1 \cap C_2 = \phi$ を満足する．

4. ゼロ交差制約の充足

一般に，二つの異種系統樹データベースはゼロ交差制約を満足しない．本章では，それらのデータベースから調停処理に必要な二つの部分木を検索し，検索結果として得られる二つの順序木に対するゼロ交差制約の充足方法を提案する．二つの順序木の葉節点列同士が接続された有向グラフを交差のない状態にするには，各木が(CS_1)を満足し，葉節点列同士の接続において(CS_2)を満足することが重要である．(CS_2) を充足させる前に，(CS_1)を満足するような順序木（初期値）を取り敢えず求めておくことは，前節 3.3 の幅優先探索により達成可能である．しかし，この段階では，図5の点線で示されるように双方の系統樹の葉節点間を結ぶことにより定義される辺同士に交差が生じてしまう．

この状態から交差をなくするために，葉節点の順序列において葉節点の順序を交換することになるが，旨く交換しなければ，木の枝の間に交差ができてしまう．ここでは，両系統樹の葉節点間にだけ結合行列を考え，ゼロ交差制約(CS_1)に違反しないような行列変換を行うことにより，ゼロ交差制約(CS_2)を充足させる方法を提案する．ゼロ交差制約(CS_1)に違反しないような行列変換は，前節で探索されるクラスタ同士の交換により保証される．この行列変換により単位行列を導くことができれば，ゼロ交差制約(CS_2)が充足される．以下，本章では，有向グラフの交差を表現するのに有用な結合行列について述べ，その結合行列に基づいてゼロ交差制約を充足する方式を提案する．また，この方式の計算量についても簡単に触れる．

$$\text{Mat} = \begin{array}{c} \\ c \\ b \\ d \\ a \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix} \end{array} \quad \begin{array}{l} OL_1 = [c, b, d, a] \\ \\ OL_2 = [a, b, c, d] \end{array}$$

図6 結合行列の例

Fig. 6 An example of an interconnection matrix

4.1 結合行列

図5の葉節点列間の結合関係を表わす結合行列を図6に示す．図の左側系統樹及び右側系統樹に関する葉の順序は，各々，OL_1=[c, b, d, a]及びOL_2=[a, b, c, d]で表現されている．左側系統樹の葉 c と右側系統樹の葉の順序集合{a, b, c, d}との間の結合関係は，結合行列Matの1行目の行ベクトル(0 0 1 0)の中に示されている．左側系統樹の葉 c は右側系統樹の葉{a, b, d}と接続関係をもたないので，行ベクトルの第一，第二，第四の要素は0で表現されているが，第三要素は両系統樹間の葉 c 同士で接続関係をもつので1で表現されている．

ゼロ交差制約（CS_1），（CS_2）を満たす葉節点列を計算するためには，図6の結合行列に対して行（及び列）同士を旨く交換することが重要である．明らかにゼロ交差制約（CS_2）を満たす結合行列は単位行列に等しい．単位行列を導くためには，ゼロ交差制約（CS_1）を維持しながら結合行列の対角線上の要素 $Mat_{i,i}$ を1にする必要がある（$1 \leqq i \leqq$ 行または列のサイズ）．そのためには，結合行列の行同士（及び列同士）の交換操作を工夫することが重要である．ここでは，この操作を結合行列の左上から対角線に沿って右下へと順に実施する．操作の基準となる対角線上の要素をピボット要素と呼ぶ．以後，ピボット要素を含む行（または列）をピボット行（または列）と呼ぶ．一般に，要素 $Mat_{i,i}$ の値を1にするような行同士または列同士の交換候補は数多く存在するが，可能な交換操作をまとめると次のとおりである．

（1）ピボット列中で値が1の要素を含む行とピボット行とを交換する．ただし，ピボット要素 $Mat_{Pivot, Pivot}$ の値が1のときは，交換は不要である．

（2）Pivot+j 列中で値が1の要素を含む行とピボット行とを交換する．ただし，$Mat_{Pivot, Pivot+j}$ の値が1のとき，行同士の交換は不要である．行同士の交換を実施後，ピボット列と Pivot+j 列を交換する．

ただし，Pivot は行列の左上要素を1番とするピボット

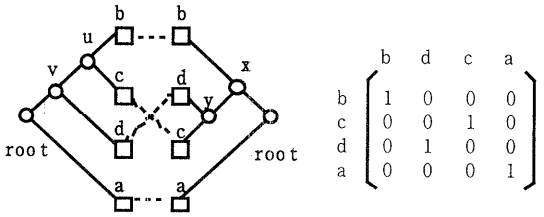


図7 結合行列の例

Fig. 7 An example of constructing a unit matrix

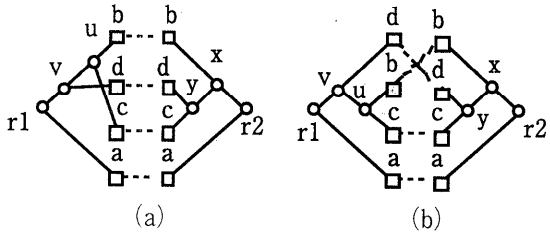


図8 交換できない例

Fig. 8 Examples of impossible swaps

番号であり，1≦j，1≦Pivot+j≦“行（または列）のサイズ”．

上記の交換操作に，ゼロ交差制約（CS_1）の充足処理機能を追加することは容易である．ここでは，結合行列上で葉節点同士を交換する代わりに，前述のクラスタ同士を結合行列上で交換している．この交換により，ゼロ交差制約（CS_1）に違反する状態が作られるのが回避される．しかし，この交換操作だけにとどめておくと，バックトラックにより別解を求める処理ではないにもかかわらず，既に決定済みの葉節点（ピボット処理が終了した対角要素の葉節点）が移動する場合が発生する．その場合は，再計算を意味するので，本方式ではクラスタ同士の交換をあきらめ，ほかの可能な交換候補の探索を進めている．

4.2 処理手順

図7の例を用いて，結合行列の行及び列を交換し，対角化を進める処理の説明から始めることにする．行と列のどちらのサイズ N_{max} も4である．対角化処理は，結合行列の左上から右下へと行われる．図7では，1番目のピボット要素(行が1で列が1の行列要素)$Mat_{1,1}$ の処理が終り，2番目のピボット要素 $Mat_{2,2}$ の処理を開始するときの状態を示している．

前節4.1の交換操作（1）を用いて，ピボット列中で1の値が位置する3行目とピボット行の2行目とを単純に交換すると，図8の(a)に示されるように枝の間に交差が発生する．すなわち，（CS_1）に違反する．これを回避するために，前節3.3で述べたように，cとdを区別するクラスタを見つける．その結果，cを含むクラスタ{c, b}とdを含むクラスタ{d}を見つけることができる．図7において，クラスタ{c, b}とクラスタ{d}を交換すると，図8の(b)が得られるが，既に $Mat_{1,1}$ に設定された値が変化してしまう．これは，再計算を引き起こす原因になるので，このような交換は禁止される．その結果，前節4.1の交換操作（2）を用いて，図7で次の交換候補を見つける処理に移行する．それは，ピボット列の右隣の列（3列目）中に存在する1の値に着目することで達成される．行列の3列目中をみると，既にピボット行に1が存在するので，行同士の交換は不要である．さらに，前節4.1の交換操作（2）を進めると，ピボット列(Pivot)とその隣の列（Pivot+j, j=1）との交換が行われる．この時点では，左側の系統樹の葉節点dとcが交換対象であり，前節3.3で述べたdとcを区別するクラスタは{d}と{c}である．これにより，ゼロ交差制約（CS_1），（CS_2）を満たす結合行列が得られる．

以上の例を踏まえると，ゼロ交差制約を充足するための処理プログラムは以下のように整理することができる．

```
1.  procedure zero_crossovers( Mat,OL1,OL2,NewMat,NewOL1,NewOl2);
2.  begin
      /* Mat: 結合行列（Interconnection Matrix） */
      /* OL1:行側（左側系統樹）の葉節点列 */
      /* OL2:列側（右側系統樹）の葉節点列 */
3.    MatOL      := {Mat,OL1,OL2};
4.    NewMatOL := {NewMat,NewOL1,NewOL2};
5.    FixedLeves := φ;  Pivot    := 1;
6.    while( Pivot ≦N_max) do   /* N_max: 葉節点数 */
7.      J := Pivot;
8.      gen_swap( FixedLeves, Pivot, J, MatOL, NewJ, NewMatOL);
9.      while(gen_swap が異常終了) do /* バックトラック処理 */
10.       if Pivot =1 then return(異常);
          /* （Pivot –1）番目のピボット要素の実行環境を取得 */
11.       pop( stack, {Pivot, FixedLeves, J, MatOL} );
12.       gen_swap( FixedLeves, Pivot, J, MatOL, NewJ, NewMatOL);
13.     end;
        /* 実行環境を保存    */
14.     push( stack, {Pivot, FixedLeves, NewJ, MatOL} );
15.     FixedLeves:= FixedLeves∪{NewOL1(Pivot)に対する葉節点};
16.     Pivot      := Pivot + 1;
17.     MatOL     := NewMatOL ;
18.   end;
19.   return(正常);
20. end;
```

処理プログラムの8行目と12行目にある gen_swap は，第二引数で指定されたピボットの番号 Pivot において交換処理を行う手続きである．この gen_swap で，

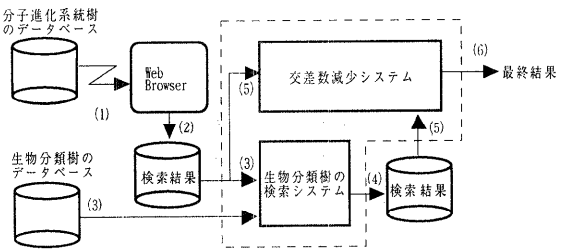


図9 システム構成図

Fig. 9 System configuration

i＋1番目のピボットにおける交換処理を考えてみよう．そこでは，ゼロ交差制約（CS_1）への違反を回避するために，二つの葉節点から両者を区別する二つのクラスタを探索する．そして，その二つのクラスタC_1, C_2間の交換が行われる．どちらかのクラスタに，既に処理済みの1～i行（または1～i列）に対する葉節点が含まれると，それらの葉節点の移動はこれまで対角要素に設定された値を解除することになるので，そのクラスタ間の交換を無効にしている．gen_swap では，既に交換を終えた葉節点を gen_swap の第一引数 FixedLeaves で管理させ，gen_swap 内でクラスタ間の交換時にそのような葉節点が移動することがないようにしている．

i＋1番目のピボットにおいて，どんな交換を考えても禁止される場合がある．その場合，処理プログラムの11行目で直前のi番目のピボット位置にバックトラックし，ほかの交換候補を見つけた後，i＋1番目のピボット位置で新しい交換候補を見つける処理を実施する．これに対処するために，14行目では，スタックを用いて，次のピボット位置に移動するたびに，直前の計算環境を保存している．push 及び pop の第一引数には，スタック領域の名前 stack が格納されている．push の第二引き数により，現在の計算環境が保存され，pop の第二引数により，その直前の計算環境が取り出される．

以上，gen_swap の交換操作は 4.1 で述べたとおりだが，参考までにその詳細を付録3に示しておく．付録3に示されている gen_swap の手続きでは，与えられたピボット要素において，交換候補の集合を生成するだけでなく，交換候補の集合から順に交換行または列の選定と交換処理の実行などが行われている．交換処理の実行は，gen_swap 中の swap 処理にて行われる．

4.3 計算量

ここでは，枝の間に交差が生じないようにするために，結合行列を対角化する処理（ゼロ交差制約（CS_1）及び（CS_2）を充足させる処理）の時間計算量Ωについて考察する．ただし，簡単のため二つの系統樹の構造は同一とし，各木の階層数をnとする．i番目の階層レベルに存在する節点数をM_iで表わすと，各木の葉節点数はM_nである．したがって，二つの系統樹を葉節列間で接続した有向グラフの階層数は$2n+1$である．結合行列を対角化する際にバックトラックがないとすれば，計算量は次のとおりである．

$$O(h_1 \times M_n^2) \quad (1)$$

また，可能なバックトラックを全て実施したとすれば，計算量は次のとおりである．

$$O(h_2 \times M_n^2 \times M_n!) \quad (2)$$

ここで，$h_1 = h_2 = n-1$である．これを式(1)～(2)に代入すると，次の時間計算量が得られる．

$$\Omega = O(n \times M_n^2) \sim O(n \times M_n^2 \times M_n!) \quad (3)$$

ここでは，結合行列上で1の値を対角線上の左上から右下へと順に乗せていくというヒューリスティクスを使っている．したがって，クラスタ交換が無効になることが原因で生ずるバックトラックの回数が多くなければ，計算量は$O(n \times M_n^2)$程度と考えられる．

従来の方式にしたがい，全ての階層間に結合行列を作成する場合について考えてみよう．$i-1$レベル階層とiレベル階層との間の結合行列に着目すると，解の候補数は，$M_i! \times {}_1C_m$個あり，1つの解を計算するための計算量は，$O(M_i \times M_{i-1})$である（$I = M_h = M_{i-1}$, ${}_1C_m$は二項係数）．これにより，上記のh_1, h_2に相当する部分は，以下のようになる．

$$h_1 = O(\pi_{[2 \le i \le n]} M_{i-1} \times M_i) > O(n) \quad (4)$$

$$h_2 = O(\pi_{[2 \le i \le h]} M_{i-1} \times M_i \times M_{i-1}! \times {}_1C_m) > O(n) \quad (5)$$

ただし，πはiに関する乗積を表わす．式(4)～(5)を式(1)～(2)に代入すると，次式が得られる．

$$\Omega = O((\pi_{[2 \le i \le n]} M_{i-1} \times M_i) \times M_n^2) \sim O((\pi_{[2 \le i \le n]} M_{i-1} \times M_i \times M_{i-1}! \times {}_1C_m) \times M_n^2 \times M_n!) \quad (6)$$

以上により，従来の方式では，各結合行列の計算量の積をとった大きな計算量をもつことになり，本提案方式はそのような悪影響を回避することができる．

5. 実装

図9に，ゼロ交差制約を満たす順序木を作成するアルゴリズムを実装したシステム構成図を示す．図中の交差数減少システムは，約500行程度の Prolog 言語により試作されている．生物分類樹の検索システムは，筆者らによって既に作成されているシステム[23]であり，これは，関係データベースシステム SYBASE の SQL や制御フロー言語 CFL 等により実現されている．処理の流れを，図中の括弧つき番号で示す．

アルゴリズムの実装に際しては，少し手を加え，計

算途中で既に得られた交差状態よりも小さな交差数をもつ両順序木を保持する仕組みを追加した．これにより，ゼロ交差制約を満たす解が存在しない場合，バックトラックによる全解探索を行えば，最小の交差数をもつ解を探索することができる．別の見方をすれば，全解探索中の適当な時点（目標交差数に達した場合や交差数の減少が認められなくなった場合など）で処理を打ち切れば，それまでに得られた処理結果の中で最小の交差数をもつ両順序木が得られるようにもなっている．また，実装において，行列 Mat の各列のデータ構造として，列中の値1が存在する要素番号を用いている．例えば，交差数ゼロを表現する4×4の単位行列は，リスト[1, 2, 3, 4]で表現され，リスト要素は昇順に並べられている．また，交差数はリスト要素間の大小関係が反転する回数として計算している．

ここでは，これまでの図7の例題と4つの実際のデータを用いて，Fujitsu S-4/1000E 上でその動作確認を行った．表1にその結果を示す．表中のケース1は図7の例題を使用し，ケース2～ケース5は国立遺伝学研究所の斎藤ら[14]が作成した分子進化系統樹データベース（通称，ジャングルデータベースと呼ばれている）等から検索したものを利用している．ケース4及びケース5は，交差が解けない問題であるが，ここでは，あらかじめ，目標交差数を与え本アルゴリズムの実行中に交差数とそれを比較することにより，アルゴリズムを停止するようにした．本提案方式では，総節点数が100から200個程度でも5分以内に解が求まった．また，ピボットを結合行列の右下に進める中で，ほぼ単調に交差数が減少していくことも確認できた．表中の単純な探索方式とは，結合行列を用いないで全ての階層レベルについて可能な順序を生成する素朴な方法を指すが，これからも提案方式が極めて優れていることがわかる．表中の＊印は，3時間実行しても，希望する解が得られなかったことを示している．

表1 測定結果

Table 1 The execution time

項目 / データ名	生物分類樹の節点数	分子進化系統樹の節点数	最小交差数	初期交差数	CPU時間(秒)	
					単純な探索方式	提案方式
ケース1	7	7	0	1	0. 11	0. 06
ケース2	25	13	0	20	1. 20	0. 20
ケース3	20	25	0	36	229. 52	0. 30
ケース4	117	42	32	162	＊	1. 60
ケース5	241	79	10	411	＊	192. 40

6. 関連研究

異種系統樹の調停の中心概念は二つの木の間の写像概念であるが，この面白いアイデアは Goodman らによって紹介された[18]．その後，Page らを中心とする多くの分子進化学者によって，調停に関する応用研究が行われ，数学的にも整理されてくるようになった[16),17),19)]．従来は，二つの木から作成される調停木は，(1)二つの木の間の写像，(2) 写像結果を利用した部分木の複製・追加により達成されている．調停処理の中では，二つの非順序木から順序木を作成するような煩わしい手作業の仕事があるが，小規模な問題が多かったこともあり，それをも自動化する研究がなされていなかった．本論文では，その手作業の部分をゼロ交差制約の充足問題としてとらえ，それを解決する方法を提案した．

さて，制約充足問題を解く代表的方法には，弛緩法，併合法，状態空間法，木探索法などある[32),33)]．弛緩法は，制約変数の値域からあらかじめ無駄な要素を削除することによって探索空間を絞り込む方法である．併合法は，複数個の制約を一つの制約に併合することにより解をもとめる方法である．状態空間法は，ＳＡ，ＧＡ，ニューラルネット等により状態空間の地形を考慮しながら解を大域的に探索する方法である．木探索法は，本論文で採用した方法である．これは，変数に順次制約に矛盾しない値を割り当てて行く方法である．この木探索には，深さ優先探索や幅優先探索があるが，バックトラックを伴う深さ優先探索の効率を上げるために，いろいろなヒューリスティックスが考えられている．筆者らは，葉節点レベル間の対応関係を表わす結合行列からゼロ交差の結合行列を探すために，深さ優先探索法を採用している．この探索において非葉節点の交差を防止するために，クラスタ同士の交換操作を行っている．これに利用したヒューリスティックな探索は，つぎのような交差数を減らす方法の研究を基礎にしている．

その研究は，ｎレベル階層（ｎ≧2）の一般的な有向グラフに対して行われいる[30),31)]．ｎレベル階層の問題は，二レベル階層の問題の延長として捉えられている．不幸にして，二レベル階層の問題はＮＰ完全[35),36)]であり，重心法や中央値法などいくつかの効果的なヒューリスティックスが研究されてきた[30),31),36)]．重心法は，結合行列上で二レベル階層間の結合状態を表わす各行（及び列）の重心を計算し，重心値が単調増加になるように行（及び列）を交互に並び換える方法である．中央値法は，重心の替わりに中央値（最大値と

最小値の平均値）を計算する方法であり，基本的な考え方は重心法の延長線上にある．

筆者らは，重心法を基礎にしたヒューリスティックな探索を用いているが，二つの木から構成される特殊化された有向グラフを対象にしているため，筆者らの方法には以下のような特徴があると言える．

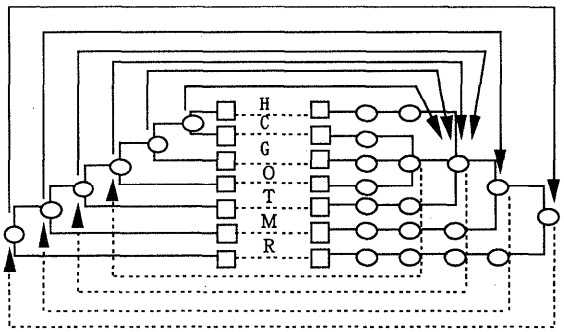
（1）値1が複数個存在する行（または列）が存在しないため，行（及び列）の重心の計算は不要である．すなわち，ヒューリスティックな探索は，対角線上に値1が配置されるように，対角線上の左上から右下へと順に行同士の交換と列同士の交換を交互に行っている．また，本方式では，重心法[30),31)]で行えなかった全解探索もバックトラックにより行えるようになっている．

（2）葉節点レベル同士で対応付けられている二つの木は，いずれも2レベルよりもかなり多い階層をもつが，ゼロ交差制約を充足させるために定義される結合行列は葉節点レベル同士だけである．すなわち，ほかの階層間には結合行列を定義しないので，階層が多くなることによる組み合わせ爆発を防止する可能性をもっている．

（3）行（または列）同士の交換操作においては，葉節点列間以外の階層間がゼロ交差であることを保証するために，その行（または列）を含む行クラスタ（または列クラスタ）をみつけ，クラスタ同士の交換操作を実施している．

7. 応用例

ここでは，二種類の系統樹データベースから霊長類における代表的な生物種を検索し，その検索結果得られた二つの順序木に対して，ゼロ交差制約を満足する順序木を作成した．図10にその結果が示されている．図では2種類の矢印がある．点線の矢印は，生物分類樹から分子進化系統樹への写像結果であるが，実線の矢印は，その反対の写像結果である．ここで，1：nの写像が存在するのは，分子進化系統樹から生物分類樹への写像の場合である．以上により，その場合の写像を1：1にするには，生物分類樹側で部分木の複製・追加が必要であることがわかる．これにより，ヒト，チンパンジー，ゴリラ，オランウータン，テナガザルの生物種間について，生物分類樹の見直しが必要であることがわかる．従来の生物分類樹では，ヒトが他の動物とは少し離れた生物とみなされていたが，実はそれほど離れた生物ではないということを示唆している．その他，最近，ゼロ交差制約を満足する順序木により，調停を行うことにより，動物に数ある筋肉の中で，胃や腸に見られる平滑筋は，骨や心臓の動きに係わる骨筋や心筋とは独立に進化してきてるようだという知見が得られている[37)]．



H:ヒト，C:チンパンジー，G:ゴリラ，O:オランウータン
T:テナガザル，M:オナガザル，R:リスザル

図10 ゼロ交差制約を満足する順序木

Fig. 10 An example of two ordered trees satisfying zero-crossover constraints

8. おわりに

本論文では，生物分類樹データベースと分子進化系統樹データベースから検索された二つの部分木に対して，ゼロ交差制約を満たす二つの順序木の計算方法を提案した．データベース中に格納されている木は，格納順として定められる順序木であるので，そこから調停に必要な葉節点集合を有する部分木を検索するだけでは交差数が多い．即ち，二つの部分木は調停処理に都合のいい順序木ではない．ゼロ交差充足問題は，そのような二つの順序木から葉節点列が同じでかつ枝に交差がない順序木を探索する問題として定義された．そのような二つの順序木は，交差のないn階層グラフと見なすことができる．一般に，交差を有するn階層グラフから交差のないn階層グラフの探索は，各階層間に$n-1$個の結合行列を定義し，それらの結合行列を単位行列に近づけることによって達成される．提案方式では，両系統樹の葉の階層間に唯一の結合行列を定義し，クラスタ交換により木の枝に交差が生じないような行列変換を行った．これにより，単位行列を効果的に見つけることができた．計算量の評価や実装による測定により，本提案方式の有効性を確認することができた．

以下，今後の課題について述べる．これまでに扱った分子進化系統樹は1生物種に対して1遺伝子から構成されていた．今後の分子進化学の発達に伴い重要になると考えられているのは，1生物種に対して複数遺

伝子を含む分子進化系統樹と生物分類樹から一つの調停木を作成する研究である．このような状況において交差制約を充足させる方法や利用者インタフェースの研究は今後の課題である．また，そのような調停木から生物分類樹を効果的に見直す方法の研究も今後の課題である．

謝辞 インターネット経由で，関係フォーマットの生物分類樹データベースを快く提供して頂いた日本DNAデータバンクDDBJをはじめとする国際DNAデータバンクのスタッフに深く感謝致します．また，プログラムの作成にあたり，部分的に作成を協力してくれた広島市立大学情報科学部4年生の西本美都子さんに感謝致します．なお，本研究の一部は，重点領域研究(略称：ゲノムサイエンス，課題番号：08283104)」および広島市立大学特定研究費の支援により行われた．

参考文献

1) 渡部各監修・伊藤敏雄訳, Newton special issue ヒトゲノム解析計画(遺伝情報を解読する巨大プロジェクトの全容), 教育社 (1990).
2) 日本DNAデータバンク(DDBJ) : http://www.ddbj.nig.ac.jp/.
3) 欧州DNAデータバンク(EMBL) : http://www.ebi.ac.uk/.
4) 米国DNAデータバンク(GenBank) : http://www.ncbi.nlm.nih.gov/.
5) ナショナルゲノム資源センター(GSDB) : http://www.ncgr.org/gsdb/.
6) DDBJ News Letter, 日本DNAデータバンク発行(静岡県三島市谷田1111 国立遺伝学研究所 生命情報研究センター), No. 17, March (1997).
7) 馬渡峻輔：動物分類学の論理, 東大出版会, 1996年.
8) 岩槻邦夫, 馬渡峻輔 編集：生物の種多様性, 裳華房 (1996).
9) 文部省特定領域研究「ゲノムサイセンス」公開シンポジウム,「ゲノムサイセンスの新展開」講演要旨集(または, ゲノムサイエンスニュースレター, Vol.3, No.3), 東京大学医科学研究所 (1999).
10) 遺伝子で生物の進化を考える, 第13回「大学と科学」公開シンポジウム, クバプロ (1998).
11) R. ルーウィン著(斎藤成也監訳, 太田聡史ら訳) : DNAから見た生物進化, 別刷日経サイエンス, 日経サイエンス社 (1998).
12) Virginia, M.: Web-Crawling Up the Tree of Life, News & Comment, *Science,* Vol. 273, August (1996).
13) TreeBASE: http://herbaria.harvard.edu/treebase/.
14) Phylogenetic Tree Database (JUNGLE) : http://smiler.lab.nig.ac.jp/jungle/jungle.html.
15) Palaeobiology Research Group : University of Glasgow, http://www.geology.gla.ac.uk/palaeo/dbase.html.
16) Page, R.D.M., and Charleston, M.A.: Reconciled Trees and Incongruent Gene and Species Trees, In: B Mirkin, F R McMorris, F S Roberts and A Rzhetsky (eds), *Mathematical Hierarchies in Biology, DIMACS Series in Discrete Mathematics and Theoretical Computer Science,* American Mathematical Society, Vol. 37, pp.57-70 (1997).
17) Page, R.D.M., and Charleston, M.A.: From Gene to Organismal Phylogeny: Reconciled Trees and the Gene/Species Tree Problem", *Molecular Phylogenetics and Evolution,* Vol.7, pp231-240 (1997).
18) Goodman, M., Czelusniak, J., Romero-Herrera, A. E., and Matsuda, G.: Fitting the Gene Lineage into its Species Lineage: A parsimony strategu illustrated by Cladograms Constructed from Globin Sequences, *Systematic Zoology,* Vol. 28, pp.132-168 (1979).
19) Page, R.D.M.: Maps between Trees and Cladistic Analysis of Historical Associations among Genes, Organisms, and Areas, *Systematics Biology,* Vol. 43, No. 1, pp.58-77 (1994).
20) Nei, M.: *Molecular Evolutionary Genetics,* Published by Columbia University Press, New York, U.S.A. (1990). この本の和訳：根井正利著(五條堀孝, 斎藤成也 共訳)：分子進化遺伝学, 培風館 (1990).
21) Page, R.D.M.: Temporal Congruence Revisited: Comparison of Mitochondral DNA Sequence Divergence in Cospeciating Pocket Gophers and Their Chewing Lice, *Systematic Biology,* Vol. 45, No. 2, pp.151-167 (1996).
22) Napier, J.R., and Napier, P.H.: *The Natural History of Primates,* British Museum (London) (1985).
23) 北上 始, 森 康真, 有川正俊, 佐藤 聡：意味的な異種性を有する生物分類樹データベースの統合化方式, 電子情報通信学会論文誌, Vol. J- 82-D1, No.1, pp.303-314 January (1999).
24) Kitakami, H., Tateno, Y., Gojobori, T. et al.: YAMATO and ASUKA: DNA Database Management System, *Proc. of the 28th Annual Hawaii International Conference on System Sciences,* IEEE Computer Society Press, Vol.5, pp.72-80, January (1995).
25) Tateno, Y. and Gojobori, T.: DNA Data Bank of Japan in the age of information biology, *Nucleic Acids Research,* Vol. 25, No.1, pp. 14-17 (1997).
26) Tateno, Y., Fukami-Kobayashi, K., Miyazaki, S., Sugawara, H. and Gojobori, T.: DNA Data Bank of Japan at work on genome sequence data, *Nucleic Acids Research,* Vol. 26, No.1, pp. 16-20 (1998).
27) The NCBI Taxonomy Homepage : http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html.
28) Tateno, Y.: Is Molecular Evolution Parsimonious ? A Theoretical Approach to the Problem, In *Population Biology of Genes and Molecules,* Baifukan, Tokyo/ Macmillan, London (1990).
29) Warfield, John N.: Crossing Theory and Hierarchy Mapping, *IEEE Transaction on Systems, Man, and Cybernetics,* pp.505-523 (1977).
30) Sugiyama, K., Tanaka, S., and Toda, M.: Methods for Visual Understanding of Hierarchical System Structures, *IEEE Transaction on Systems, Man, and Cybernetics,* pp. 109-125 (1981).

31) 杉山公造:グラフ自動描画法とその応用，計測自動制御学会 (1993).
32) Alan, L. Mackworth: Consistency in Networks of Relations, *Artificial Intelligence*, No. 8, North-Holland Publishing Co., pp.99-118 (1977).
33) 石塚満 著: 知識の表現と高速推論，丸善株式会社 (1996).
34) Kitakami, H., Mori, Y., Oota. S., and Saitou, N.: A Tree Matching Method for Mapping Species Tree to Gene Trees, サイバースペースのためのデータベース第3回ワークショップ，大阪大学待兼山会館，October (1998).
35) Garey, M. R. and Jhonson, D. S.: Crossing Number is NP-complete, *SIAM J. of Algebraic and Discrete Methods*, Vol. 4 No. 3, pp.312-316 (1983).
36) Eades, P., McKay, B. D., and Wormald, N. C.: On an Edge Crossing Problem, *Proc. 9th Australian Computer science Conference*, pp.327-334 (1986).
37) Oota S.: Development of an Integrated System for Molecular Evolutionary Study and It's Application, Department of Genetics, *Ph.D. Thesis*, School of Life Science, Graduate University Advanced Studies, September (1998).

付録1 DNAデータベースのデータ形式[6)]

DNA塩基配列データは，下図のORIGIN行（下部）に記載されている．それ以外の部分には，そのDNA塩基配列に関する付属情報が記載されている．DNA塩基配列に関する生物分類樹の情報は，図の上から6行目～9行目の間のSOURCE行中のORGANISM行に記載されている．この情報は，生物分類樹の葉から根に至る経路上の全節点が生物学的分類情報であり，各データバンクが独自に構築した生物分類樹データベースを用いて作成される．

```
LOCUS       HUMTCAL8        407 bp ds-DNA            PRI       08-APR-1992
DEFINITION  Human cone transducine alpha subunit (fc-alpha) gene, exon8.
ACCESSION   D10384 D90445
KEYWORDS    cone transducin alpha subunit.
SEGMENT     8 of 8
SOURCE      Human DNA, clone lambda-HTC-78.
  ORGANISM  Homo sapiens
            Eukaryota; Animalia; Metazoa; Chordata; Vertebrata; Mammalia;
            Theria; Eutheria; Primates; Haplorhini; Catarrhini; Hominidae.
REFERENCE   1  (sites)
  AUTHORS   Kubo,M., Hirano,T. and Kakimuma,M.
  TITLE     Molecular cloning and sequence analysis of cDNA and genomic DNA for
            the human cone transducin alpha subunit
  JOURNAL   FEBS Lett. 291, 245-248 (1991)
  STANDARD  full staff_review
REFERENCE   2  (bases 1 to 407)
  AUTHORS   Kubo,M.
  JOURNAL   Unpublished (1992)
  STANDARD  full staff_review
COMMENT     Data kindly submitted in computer readable form by:
            Mitsunasa Kubo
            Section of Bacterial Infection
            Institute of Immunological Science
            Hokkaido University
            15-Kita, 7-Nishi, Kita-ku
            Sapporo 060
            Japan
            Phone:  011-716-2111 x5521
            Fax:    011-758-7568
FEATURES            Location/Qualifiers
     intron         <1..30
                    /number=7
     CDS            join(D10377:120..237,D10378:31..73,D10379:31..172,
                    D10380:31..188,D10381:31..159,D10382:31..160,
                    D10383:3...184,31..221)
                    /product="cone transducin alpha subunit"
                    /codon_start=1
     exon           31..>407
                    /number=8
BASE COUNT      140 a     76 c     73 g     118 t
ORIGIN
        1 acttttttcac tatttttctg gaaaacccag gtaacaactc ctatgatgat gcggggaatt
       61 acataaagag ccagttcctt gacctcaata tgcgaaaaga tgtcaaagaa atctacagtc
      121 acatgacctg tgctacagat acacagaatg tcaaatttgt atttgatgca gttacagata
      181 ttatcatcaa agaaaacctc aaggtctgcg gcctctteta atcctcacca ttcctcaggt
      241 ataagttcta taaacaggct tggaattctgg gtaattaaaa acagaaaatt atagtcaata
      301 taccatgaca tgaagaatga atccattctt tggagatgga gtatacatga ctgcaactgt
      361 attcatacg ttcttttcaa agtgtgatag tattgcacgt taaagag
//
```

付録2 異種系統樹間の調停処理[16),19)]

従来の異種系統樹間の調停手順をまとめると，以下のように整理することができる．

（1）両系統樹を上下に探索することにより，ある系統樹の節点集合（定義域）からほかの系統樹の節点集合（値域）への対応付けを行う．
（2）その写像が単写でなければ，複数個の節点から対応付けられる節点について，値域側でそれを頂点とする部分木を複製・追加し，その対応付けが1対1になるようにする．

図11に（1）の処理を図示する．これは，生物分類樹から分子進化系統樹への写像例である．この場合，両者の葉節点の名前が同一である．宿主と寄生生物に関する写像の場合は，両者の葉節点の名前が異なるが，葉同士の対応関係は事前に観測データとして与えられている．図11の例では，同じ葉節点名同士及び双方の根節点 r1, r2 同士は自然に対応付けられている．それらは点線で示されている．ただし図では，ゼロ交差制約が充足していると仮定されている．すなわち，葉節点の並びは双方ともに，a, b, c, d の順である．この写像において，系統樹探索による節点間の対応付けされる節点は，根及び葉節点以外の節点 x, y, u, v である．左側の系統樹における節点 u に対応する右側の系統樹の節点 x は，次のように探索されている．最初に，左側の系統樹において節点 u の葉節点 b, c を探す．次に右側の系統樹において葉節点 b, c の共通の先祖 x を探す．これらにより，節点 u が節点 x に対応付けられる[16),19)]．同様に，節点 v は節点 x に対応付けられることが判る．

図12に（2）の処理を図示する．定義域側の二つの節点に対応付けられた左側の系統樹の節点 x について，根 x の部分木をコピーし，両部分木の親を作成する．図では，□で囲まれた部分木がコピーにより作成された部分であり，●印の節点が新たに作成された親節点である．このコピー後行われる左右の葉節点同士の対応付けは，●印の節点を根とする部分木と節点 v を根とる部分木から容易に決めることができる．斜線の入った□印に着目すると，左側系統樹における既存の節点 d の他に，コピーにより作成された節点 b, c は，絶滅した生物種かまたは未だ観測されていない生物種かのどちらかであると解釈されている．

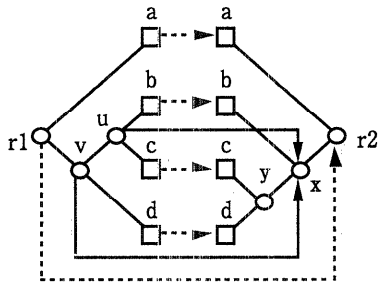


図11 異種系統樹間の写像の例

Fig. 11 An example of a mapping between heterogeneous trees

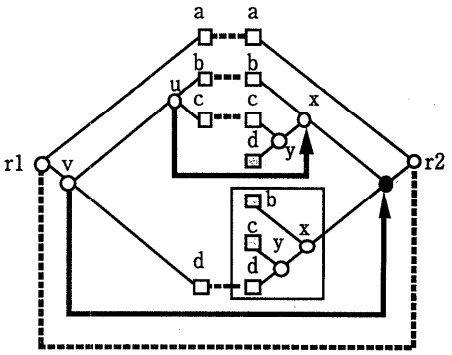


図12 部分木のコピーによる調停例

Fig. 12 An example of the results of reconciliation

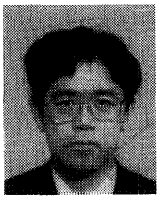
付録3 交換処理プログラム

```
procedure gen-swap( FixedLeves,Pivot, J, MatOL, NewJ, NewMatOL );
begin
    Search:="on";
    while( Search="on" and J≦Nmax ) do   /* Nmax は葉節点数 */
        if J=Pivot then do
            if Mat( Pivot, J)≒1 then do
                NewMat := Mat;   NewJ := J+1; Return(正常) ;
            end;
            J 列で 1 の存在する行 K（>Pivot）を見つける;
            /* K 行と Pivot 行を交換しても枝の交差ができなけ */
            /* れば NewMat に反映させる.                    */
            swap( "row", FxedLeves, MatOL, K, Pivot, NewMatOL );
            if swap-row が正常終了 then Search:="off";
        end;
        else   do   /* J > Pivot */
            if Mat( Pivot, J) = 1 then do
                /* J 列と Pivot 列を交しても枝の交差ができなけ */
                /*   れば NewMat に反映させる.                  */
                swap( "column". FixedLeves, MatOL,Pivot, J,NewMatOL );
                if swap-column が正常終了 then Search:="off";
            end;
        else do /* Mat( Pivot, J) = 0 */
            J 列で 1 の存在する行 K（>Pivot）を見つける;
            /* K 行と Pivot 行を交換しても枝の交差ができなけ */
            /* れば NewMat に反映させる.                    */
            swap( "row", FixedLeves, MatOL, K, Pivot, TempMatOL);
            if swap-row が正常終了 then do
                /* Pivot 列と J 列を交換しても枝の交差ができなけ */
                /* れば NewMat に反映させる.                    */
                swap("column",FixedLeves,TempMatOL,Pivot,J,
                        NewMatOL);
                if swap-column が正常終了 then Search:="off";
            end;
        end;
    end;
    J:= J+1;
end;
if search="on" and J>Nmax then return(異常);
NewJ := J;
return(正常);
end;
```

```
procedure swap( Mode, FixedLeves, MatOL, Pivot, J, NewMatOL);
begin
  Pivot 行と J 行に対応する 2 つの葉節点 SP1,SP2 を区別するための分岐点 B
  を見つける;
  Leaves1 := 分岐点 B から見て, SP1 と同じクラスターに所属する SP1 を含
  む葉節点集合;
  Leaves2 := 分岐点 B から見て, SP2 と同じクラスターに所属する SP2 を含
  む葉節点集合;
  if FixedLeaves∩Leaves1≠ φ then return(異常);
  if Mode="row" then do
        NewOL1=OL1;   /* 列間の順序の変化はなし */
        順序つき葉節点リスト OL2 に対して, Leaves1 と Leaves2 を交換し
        NewOL2 を作成する;（この交換に際しては, Pivot 位置に 1 がくる
        ようにする.）
        Mat に対して, Pivot 行と J 行を交換し NewMat を作成する;
  end;
  else do /* Mode="column" */
        順序つき葉節点リスト OL1 に対して, Leaves1 と Leaves2 を交換し
        NewOL1 を作成する;（この交換に際しては, Pivot 位置に 1 がくる
        ようにする.）
        NewOL2=OL2;   /* 行間の順序の変化はなし */
        Mat に対して, Pivot 列と J 列を交換し NewMat を作成する;
  end;
  MewMatOL := {NewMat, NewOL1, NewOl2};
  return(正常）;
end
```

（平成 10 年 12 月 20 日受付）
（平成 11 年 3 月 27 日採録）

（担当編集委員 中谷 多哉子）

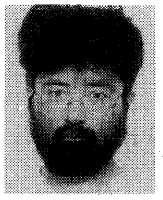


北上　始（正会員）

昭５１年 東北大学大学院修士課程修了．同年富士通株式会社入社．以後，富士通研究所，新世代コンピュータ技術開発機構，国立遺伝学研究所客員助教授を経て，平６年 広島市立大学情報科学部教授，現在に至る．科学データベース，分子生物学データベース，演繹データベース，知識ベース，画像データベース，知的情報検索，データマイニングなどの教育研究に従事．博士（工学）．情報処理学会２５周年記念論文．電子情報通信学会，情報処理学会，人工知能学会，日本遺伝学会，IEEE，ACM 各会員．

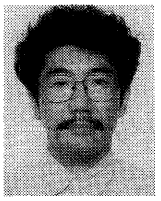
森　康真（正会員）

平６年 北陸先端科学技術大学院大学・情報科学研究科博士前期課程修了．同年広島市立大学情報科学部助手，現在に至る．情報フィルタリング，テキスト情報検索，発想支援，CSCW などの研究に従事．電子情報通信学会，情報処理学会，人工知能学会各会員．



太田　聡史

平成７年 北陸先端科学技術大学院大学 情報科学研究科博士課程前期修了．修士(情報科学)．平成１０年総合研究大学院大学生命科学研究科博士課程修了．博士（理学）．分子進化学の研究に従事．日本遺伝学会，日本分子生物学会各会員．



斎藤　成也

昭５６年 東大大学院修士課程終了．昭５７年～昭６１年にテキサス大学ヒューストン校にて Ph.D.取得．帰国後，日本学術振興会研究員，東大理学部助手を経て，平成３年より国立遺伝学研究所助教授，現在に至る．総合大学院大学生命科学研究科助教授を併任．遺伝子進化学，人類進化学．現在は血液型遺伝子の進化，遺伝子進化解析法の開発，器官・組織の進化，人間性を規定する遺伝子群の探索，人類集団間の遺伝的近縁関係などの研究に従事．日本遺伝学会，日本分子生物学会各会員．