投稿先：日本動物分類学会和文誌タクサ

直接最適化法 direct optimization の問題点

Problems in direct optimization

（ランニングタイトル：直接最適化法の問題点）

吉澤和徳[1)]

Kazunori Yoshizawa[1)]

[1)]北海道大学農学部昆虫体系学教室

〒060-8589　札幌市北区北９条西９丁目

Systematic Entomology, Graduate School of Agriculture, Hokkaido University, Sapporo, 060-8589 Japan

E-mail: psocid@res.agr.hokudai.ac.jp

ABSTRACT

Problems in direct optimization method as implemented in the software POY were reviewed. Alignment of highly variable regions using POY is problematic because it likely provides artificial maximization of pseudo-homology within the alignment. If the variable regions are aligned with conserved regions, then the bias could be emerged in two ways. When the conserved regions contain only weak phylogenetic signal, then stochastic similarities within the variable regions strongly affect to the final alignment. This type of bias can be emerged even if the variable regions are aligned independently. As a result, aligned variable sequences may provide superior amount of pseudo-signal which could even mask true weak signal contained in the conserved regions. When the conserved regions contain significant phylogenetic signal, then the variable regions will play as adherents of the conserved data even if the variable regions contain no or even contradicting phylogenetic signal. The adherent bias is especially problematic when data set contains heterogeneous sequences by gene introgression, recombination, or lineage sorting. Structure- or similarity-based approaches should be used in aligning DNA sequences.

Key Words: direct optimization, POY, ILD, DNA sequence, alignment, phylogeny, homology

■ はじめに

「...照見五蘊皆空度一切苦厄...」（般若心経の一節）

直接最適化法 （direct optimization method） は，塩基配列間の多重配列アライメント（以下単にアライメントとする）と，系統樹の推定を同一の認識論的問題として扱うアルゴリズムである．この方法のもとでは，アライメントの最適化と樹形探索は，1ステップ解析によって，同一の最適化基準に基づき同時に行われる（図１）．これは通常の系統推定の手順である２ステップ解析（アライメントをした後，そのアライメント結果に基づいて系統推定する方法）とは大きく異なる．アライメントの最適化と樹形探索を複数回繰り返す直接最適化法（図１）は，多大なマシンパワーを要求することもあり （Giribet *et al.*, 2001），この方法を実装したソフトウェア POY （Wheeler *et al.*, 2006）が，系統解析を行う研究者の間で多くの利用者を得るにはいたっていない．その一方，POY の開発者である Ward Wheeler や彼の共同研究者達は，スーパーコンピュータや並列コンピュータを用いて大規模データの解析を行い，活発に POY を用いた論文を発表している．それらの中には，節足動物 （Giribet *et al.*, 2001） やナナフシ （Whiting *et al.*, 2003） の系統の論文など， *Nature* 誌に載った影響力の極めて大きい論文もある．その他にも，多くの節足動物 （Giribet and Edgecombe, 2006; Wheeler and Hayashi, 1998; Wheeler *et al.*, 2001; Terry and Whiting, 2005 など），脊椎動物 （Bertelli and Giannini, 2005; Frost *et al.*, 2001, 2006; Faivovidh *et al.*, 2010; Giannini and Simmons, 2003 など），軟体動物（Giribet *et al.*, 2006 など），環形動物 （Worsaae *et al.*, 2005） などの高次レベルの系統関係に関する重要な論文が，POY を用いて作り出されている．

直接最適化法および POY の特徴は，その独特のアライメントの最適化法にある．アライメントソフトとしての POY のパフォーマンスに関するテストはこれまでいくつか行われてきた．例えば，Ogden and Rosenberg (2007) および Simmons *et al.* (2008a) は，シミュレーションデータを用いることにより，POY のパフォーマンスを他のアライメントソフトと比較している．さらに直接最適化法の方法論に対しては，「独特なアルゴリズムで無理矢理アライメントを行って，系統解析に利用する」（上島, 2008），"Problematic data are analyzed with questionable methods"（Wiens, 2007） など，問題視する意見も多い．しかし，直接最適化法の方法論的問題の具体的な検討は Simmons (2004) を除き，ほとんどなされてこなかった．本稿では，この直接最適化法の問題点を，Yoshizawa (2010) にそって議論する．解析の方法について，本稿では大まかな流れを述べるにとどめるので，実験の再現にあたっては Yoshizawa (2010) を参照いただきたい．本稿ではそれに加えて，Morgan and Kelchner (2010) によってなされた直接最適化法に関する哲学的な問題提起についても紹介する．

なお筆者が知る限り，これまでに日本語で書かれた直接最適化法への批判として，上島（2008）と吉澤（2008）がある．このうち上島は，直接最適化法の方法論のみならず，全証拠解析 （total

evidence analysis）および missing data を含むデータセットの解析も批判している．混同を避けるために念のため付け加えておくが，筆者が本稿および吉澤 （2008）, Yoshizawa (2010) で批判したのは，直接最適化法それ自身のみであって，分子データと形態データを統合した系統解析（広い意味での全証拠解析：以下で述べるように直接最適化法の一部としての全証拠解析には重大な問題がある）や，missing data を含むデータの解析を批判している訳ではない．

■ アライメント

塩基配列のアライメントは，異なる配列間の塩基座位の相同性の推定にほかならい．これは系統推定を行うにあたり，最適化基準の選択（最節約，最尤，ベイズ）や塩基置換モデルの選択以上に重要な過程と言える．なぜなら，非相同な形質同士の比較は全く意味をなさず，そのような比較は，系統推定に当たってはノイズしかもたらさないからである （De Pinna, 1991）．村上（1998）は「分子情報の方は実験を失敗しない限り，何時，誰がどこでやっても結果は同じで，厳密さは要求されない」と述べている．確かに分子データの場合，同一のサンプルの同一の領域を，同じプライマーを用いて PCR で増幅しシーケンスすれば，誰が実験しても常に同一の配列が得られることが期待される．しかし残念ながら，アライメントに関しては，何時，誰がどこでやっても同じ結果が得られるとは期待できない．

アライメントが問題となるのは，塩基の挿入欠損が頻繁におこっている場合である．なぜならこのような領域では，配列同士の塩基座位間のアライメントが一義的に決定できないからである．一般に，塩基配列同士のアライメントは，コンピュータソフトウエアを用いて行われる．多くのソフトウェアでは，塩基の置換パターン（トランジッション／トランスバージョン）や挿入欠損に異なるコストを与え，そうして与えられたパラメータのもとで，アライメントの良否を判定する．パラメータ設定はアライメント結果に大きな影響を及ぼすが，それを決定する客観的な基準は存在しない．さらに，ソフトウェアに組み込まれたアルゴリズムや最適化基準の違いによってもアライメント結果は変わってくる （Wong *et al.*, 2008）．また，アライメントの信頼性の低い領域の取り扱い（解析に用いる/除外する）も問題となる．高次レベルの系統解析で頻繁に用いられるリボソーム RNA コード領域は，挿入欠損をほとんど含まない保存性の高い領域がある一方，挿入欠損を多数含むアライメント困難な領域も存在する．例えば，挿入欠損を多数含む領域のうち，E23 と呼ばれる領域は hyperbariable region と呼ばれ，比較的近縁なグループ間でもアライメントが困難となる．そして，こういった領域の取り扱い方によって，系統解析の結果は大きく変わってくる．直接最適化法では，こういった領域もアライメントし，解析に用いる．

■ 直接最適化法とその問題点

まず直接最適化法の手順を大まかに紹介する（図１）．直接最適化法の第一段階として，入力された生データに対して通常のアライメントが行われる．そうやって生成された初期のアライメント結果に基づいて樹形探索が行われ，そして得られた樹状図の上で，塩基の置換や挿入欠損が最小化されるようなアライメントが再度生成される．この樹形探索とアライメントの最適化サイクルを，両者が固定されるまで繰り返すことで，塩基配列のアライメントの最適化と，最適樹の選択が同時に行われる．

複数の遺伝子領域や形態データが解析される場合，すべてのデータが単一のデータセットにまとめられ，同時に最適化される．この場合，データセット中の全てのデータが同一の樹形に基づいて最適化されることになるため，データ全体をとおして系統情報が均質化されるようなアライメントが生成される （Simmons, 2004）．分子データの場合，配列間の個々の塩基座位の相同性が未確定のまま塩基配列が得られ，アライメントによってそれらの相同性が推定されるのに対し，形態データの場合，相同性を推定した後でなければ，形質の比較もコーディングも意味をなさない．つまり形態データは，マトリクスを作成した段階で相同性の判断が固定される．このようなデータが結合され，直接最適化法にかけられると，形態データに含まれる系統情報（直接最適化法を通して固定）が，分子データのアライメント（直接最適化法を通して生成される）に対して強い影響を及ぼし，その結果，分子データは形態に含まれる系統情報に合わせるようにアライメントされることになる．データセット中に形態データが含まれない場合でも，蛋白質コード領域のように挿入欠損をほとんど含まない配列と，リボソーム RNA コード領域のように挿入欠損を多数含んだ配列を直接最適化法にかけた場合には，同様の問題が生じる．POY を用いた解析では，形態形質や蛋白質コード領域，RNA の保存性の高い領域は prealigned data として解析に先立ってアライメントされ，POY での解析を通して相同性の判断が固定される．一方，挿入欠損を含んだ領域の配列間の相同性は POY での解析を通して決定されるため，それらの領域のアライメント結果は，挿入欠損を含まない配列の影響を強く受けることになる．なお直接最適化法に基づく解析でも，prealigned 領域の決定には，アミノ酸配列やRNAの二次構造の情報が用いられる．しかしこれらの情報はあくまでも計算時間を短縮するために用いられているもので （Giribet *et al.*, 2000; Giribet, 2001），直接最適化法の支持者らが，これらの情報の有用性を最適化基準に上回るものとして用いているわけではない．

以上は複数データを統合した解析で生じうる問題点だが，個別のデータセットに対して直接最適化法を適用することも可能である．しかし，Wheeler and Hayashi (1998) は，複数のデータに基づく解析では，全てのデータを統合したうえで最適化してこそ，全証拠解析の相補性が最大化されるとして，直接最適化法は統合したデータに適用されるべきだと述べている．さらに，もし個別にデータを解析したとしても，そのデータセット中の配列間で組み替えなどがおこっている場合，複数データの解析

と同様の問題が生じうる （Ohshima and Yoshizawa, 投稿中）．組み替え等を含まない単独のデータであっても，挿入欠損を多数含む不安定な領域への直接最適化法の適用には問題が生じるが，これについてはランダムデータを用いた検討の中で明らかにする．

なお直接最適化法は，分子データのみではなく，コオロギの音響データ （Robillard *et al.*, 2006）やトビムシの刺毛配列の相同性決定 （Agolin and D'Haese, 2009） など，相同性の決定が困難な行動データや形態データへの適用例もある．例えば後者の場合，通常の形態のコーディングとは異なり，相同性を固定しないまま，刺毛の形質状態を 0,1.. の配列としてコードし，POY を使って刺毛の相同性を決定している．また，POY は最節約基準に基づいてアライメントの最適化と最短樹の探索を行うが，尤度を基準とした直接最適化法の拡張版も開発されている （Liu *et al.*, 2009）．

■ ランダムデータを用いた検討

本節では，前節で指摘したような問題をランダムデータの解析を通して検討する．まず実際の生物からとられた，強い系統情報を含む配列（実データ）に，ランダムに生成された系統情報を全く含まない配列を加え，POY を用いて全体を最適化した．なお解析に当たっては，ギャップ：トランジッション：トランスバージョンのコストを 1:1:1 に設定した．あとで述べるが，ここで与えているギャップのコストは，通常のアライメントソフトの初期設定値と比べ，極めて低い値となっている．生成されたランダムデータのアライメント結果のみを用いて解析を行った結果，実データから推定される系統樹（図２上）と完全に整合性があり，また全ての枝がブーツストラップ確率 100% で支持される樹状図が得られた（図２下右）．さらに，このデータから計算される一致指数，保持指数（CI=0.70, RI=0.64） の値も高く，アライメントされたランダムデータに含まれる情報のこの樹形への整合性が非常に高いことが示される．つまり，POY をとおすことで，もともと系統情報を全く含まない配列から，非常に強い「偽」の系統情報が生成されたことになる．そして，こうして得られた偽の系統情報は，実データから得られた樹形との整合性の高さからも分かる通り，実データに含まれる系統情報の影響が明らかである．上で指摘したデータの非独立性の問題を顕著に表した結果となっている．なお，ギャップコストを低く設定した場合，Clustal, MAFFT, MUSCLE, T-Coffee 等の累進法を用いたアライメントソフトでもデータの非独立性の影響が生じうる．なぜなら，これらのソフトではアライメントの精度を上げるために，配列の対比較によって生成された guide tree と呼ばれる樹状図に基づいてアライメントが行われ，また再検討されるが，この guide tree の樹形が最終的なアライメント結果にも影響を及ぼしうるからである．しかし，これらのソフトウェアを用いた場合，POY のように極めて低いギャップコスト（ギャップコストの影響については後述する）が適用されることはまずあり得ないことや，個々の遺伝子ごとにアライメントを行うことが一般的であることなどか

ら，データの非独立性の問題が強く影響するケースは少ないと考えられる （Yoshizawa, 2010）．

データの非独立性が問題であるならば，各データを単独で解析するとどうなるであろうか？ランダムデータを単独で解析した場合でも先ほどと同様，全ての枝がブーツストラップ確率 100% で支持される非常に強い偽の系統情報が生成された（図３右）．一致指数，保持指数も高い （CI=0.72, RI=0.68）．ランダムデータの単独解析からも強い偽の系統情報が生成された理由は，以下のように考えられる．直接最適化法は，アライメントと樹形を相互に参照し合うサイクルをとおして両者が最適化される（図１）．つまり，ランダムな配列単独で解析した場合でも，そこに含まれる偶然の類似に基づき初期樹が作成され，そしてひとたび樹が作成されると，その樹形を支持するようなアライメントが最適と判断され，結果として非常に強い偽の系統情報が生成されることになる．

以上の結果はギャップ：トランジッション：トランスバージョンのコストを 1:1:1 に設定して得られたものである．上でも述べたが，これは他のアライメントソフト初期設定値と比較し，ギャップのコストが極めて低い設定となっている．一方，ギャップのコストを塩基置換の15倍に設定して POY を用いてランダムデータを解析した場合（これは，Clustal の初期設定値にほぼ一致する），結合解析，単独解析いずれからも上で得られたような強い偽の系統情報は生成されなかった（図２, ３下左）．しかし，直接最適化法では，アライメントのコストを決定する際，データパーティション間の不一致を定量化した ILD （Mickevich and Farris, 1981）という数値を用い，データパーティション間の不一致を最小化するコストパラメータを最良なものとして選択する （Wheeler and Hayashi, 1998）．さらに Grant & Kluge (2003) は，パラメータ設定を客観的に擁護できる唯一の方法は，全ての変化に対して同じ重み付けをすることであるとも述べ，1:1:1 を推奨している．そのため，この 1:1:1 コストパラメータは POY の実質的なデフォルト値としてほぼ全ての研究で採用されている (Ogden & Rosenberg, 2007)．しかし，1:1:1 は言わばワイルドカードのような設定値で，ある塩基座位に好みの塩基の状態やギャップを，同じコストによって与えることができるパラメータとなっている．したがって，このパラメータのもとでは，都合の良いアライメントが作成されやすくなるのは当然で，つまり，直接最適化法はそのアルゴリズムのみならず，パラメータ選択法の点でも，人為的な偽情報を生成しやすい条件を備えていることになる．

このように，直接最適化法には，(1) 異なるデータ同士が影響を及ぼし合い，データ間の矛盾を最小化するアライメントが生成されること，(2) どんな配列からでも，特定の樹形を強く支持する偽の系統情報が生成されること，といった問題があり，さらに (3) ギャップの挿入に対して非常に低いコストが与えられることにより，(1), (2) の問題がより強化されることになる．

■　実データに基づく検討

ここでは，Whiting *et al.* (2003) によるナナフシの高次系統と Terry and Whiting (2005) による昆虫の高次系統の研究を通して，実際の生物から得られたデータの解析結果への直接最適化法の影響を検討する．多様な分類群を専門とする会員を含む動物分類学会の特性も考慮し，昆虫以外の解析可能なデータも探したが，アライメント結果を公開している例が見つからなかった．例えば，Giribet *et al.* (2001) による節足動物の高次系統の解析の例では，アライメントされていない生の配列データと POY のバッチファイルのみが公開されている．これらを用いれば彼らのアライメントは再現できるかもしれないが，一方で彼らの解析は256台の並列コンピュータを用いて行われており，これらの解析を一台のコンピュータで行った場合，解析を終えるには42年かかると述べられている （Giribet *et al.*, 2001）．さらに公開されている生データには，バッチファイルでは解析対象として示されている一部のデータが含まれておらず，たとえ42年（もちろんパソコンの性能は著しく向上しており，もっと短い時間で済むとは思われるが）の時間をかけたとしても，彼らの結果は再現できない．直接最適化法および POY の支持者たちは，解析にあたっての客観性および再現可能性を非常に強く訴えているにも関わらず （Ogden *et al.*, 2005），研究結果の再現性が全く期待できないこの状態は，彼らの主張と大きく矛盾しており，また科学研究論文としての要件すら満たしていないお粗末なものと言える．

ここで用いる Whiting *et al.* (2003) のデータに関しても，アライメント結果は公開されているものの，事前にアライメントされ，相同性が固定された領域の指定が行われていないため，実際に POY の解析対象となった範囲の特定が出来ない．そこでここでは，Whiting らによってアライメントされたデータを，挿入欠損を含まない形質と含む形質に分け，解析を行った．なお，オリジナル論文で含められていた外群は今回の解析では除外したが，内群のみの解析で得られた樹状図の樹形は，オリジナルの論文で示されたものと矛盾しない．一方，Terry and Whiting (2005) はアライメント結果を公開しておらず，また論文で示されたサンプルリストには，GenBank Accession Number に間違いが多数含まれており，データを再構成することが出来ない （Yoshizawa, 2010）．そこで，一部のデータの再解析結果と，論文で示された結果とを比較することによって，直接最適化法の問題点を検討する．

図４は，Whiting *et al.* (2003) が用いたアライメントの一部（リボソームRNAコード領域）を示している．白く抜けている部分がギャップまたは unknown data を示しているが，一見して分かるように，ギャップをほとんど含まない，アライメントの信頼性の高い領域（図の両端の領域）と，多数のギャップを含む，アライメントの信頼性が極めて低い領域（図の中央部分）から成っている．この多数のギャップを含む，アライメントの信頼性が極めて低い領域が，POY によってアライメントされた領域に相当する．このデータ全体を最節約法を用いて解析した場合，図５の中央に示した樹が得られる．なお，個々のギャップは第５の形質として解析している（ギャップを missing data とせず，形質として扱う点も，直接最適化法の特徴である）．*印で示している通り，多くの枝がブーツスト

ラップ確率 90–100% で支持される，一見非常に解像度の高い樹が得られることが分かる．一方，ギャップを含む形質を全て解析から取り除いた場合，図５右に示した系統樹が得られる．対応する末端分類群を結ぶ線を見て分かる通り，右の系統樹は全データから得られた樹とは樹形が全く異なっている．さらに，アライメントの安定した領域のみからは，多くの枝のブーツストラップ確率が50%以下という，解像度の低い系統樹しか得られない．

逆に，ギャップを含まない，アライメントの信頼性の高い形質を全て取り除き，ギャップを含む，アライメントの信頼性の低い形質のみで解析を行うと，図５左に示した樹が得られる．グレーの囲みで示した通り，アライメントの信頼性の低い形質から得られた樹は，全データから得られた樹と完全に樹形の整合性があり，さらに多くの枝がブーツストラップ確率 90–100% で支持される，非常に解像度の高いものとなっている．さらに，これらの樹形から計算された一致指数，保持指数を比較すると，ギャップを含むアライメントが不確かな形質から計算された値 （CI=0.64, RI=0.64） の方が，ギャップを含まない形質から得られた値 （CI=0.47, RI=0.44） より高く，つまり，後者の形質には，前者より多くのホモプラシーが含まれていることを示している．そもそも相同性の判断が不明確になるほど挿入欠損が頻繁に起こっているデータに含まれるホモプラシーが，ほぼ一義的にアライメントできるデータに含まれるそれより少ないという結果は通常では考え難い．さらに，ランダムデータを POY で解析したアライメント結果に含まれるギャップの割合（58.5–60.0%） や一致指数，保持指数を比較すると，ギャップを含むアライメントが不確かな形質のそれらの値（ギャップの割合は 47.2–65.2%）と近似の値を示しており，ギャップを含む領域のアライメント結果は，全く系統情報を含まない配列から得られたアライメント結果と大差ない．

上記の結果から，全データから推定された樹形は，ギャップを含むアライメントが不確かな領域に強く依存したものであること，そしてギャップを含む領域に含まれる情報は，本来の系統情報ではなく，直接最適化法によって人為的に生成された偽の系統情報である可能性が高いことが分かる．つまり Whiting *et al.* (2003) に示された樹状図は，偽の情報によって作り出された，全く意味のないものである．全く同様の傾向が別のデータセットからも見いだされていることから（Yoshizawa, 2010），ここで検討したデータは特殊なケースではなく，直接最適化法の普遍的な傾向を示していると考えられる．このような偽の系統情報が作り出された原因は，次のように推定できる．挿入欠損の無い形質は，解析を通してその相同性が固定されていたと考えられる（先に述べたように，Whiting らが公開したデータには prealigned data の指定が無いため，挿入欠損を含まない形質の中に，一部 POY による解析対象が含まれている可能性は否定できないが）．そしてこれらの形質に含まれている系統情報は，挿入欠損を含む形質のアライメント結果に影響を及ぼす．実際，挿入欠損の無い形質からブーツストラップ確率90%以上の支持を得た枝は，挿入欠損を含む形質の解析でも全てブーツストラップ確率100%で支持されており，ランダムデータの節で示したデータの非独立性の影響が強く認められ

る．一方，挿入欠損の無い形質のみの解析では，特に深い枝の多くのブーツストラップ確率が50%以下で，これらの形質には深い系統関係に関する明瞭な情報が含まれていないことが分かる．一方で，挿入欠損を含む形質の独立解析では，深い枝もその多くがブーツストラップ確率90%以上の支持を得ており，さらにそれらの枝は，挿入欠損無しの形質から推定された枝とは一致しない．これは，初期アライメントの偶然の類似が強化された結果と考えられる．

一方，挿入欠損を多数含んだ領域に真の系統情報が含まれていた可能性は否定できない．しかし，次に示す Terry and Whiting (2005) の例は，他のデータとは矛盾するような系統情報を含んだ配列からでさえ，データの非独立性の影響により，他のデータに含まれる系統情報と調和するアライメントが生成されることを示す．上述した通り，彼らのアライメント結果を再構築することは不可能なので，ここでは昆虫綱絶翅目（ジュズヒゲムシ）の単系統性に問題をしぼって，彼らによって示された結果を，GenBank に登録されている 18S のデータを二次構造に基づいてアライメントして得られた結果（Yoshizawa and Johnson, 2005) と比較することで，データの非独立性の問題を検討する．

そもそも筆者が POY の決定的な問題点に気づいたのも Terry らの論文がきっかけであった．この論文では，18S, 28S, Histone 3 の塩基配列および形態データに基づいて，昆虫の高次系統関係が推定されている．しかし，彼らの利用したジュズヒゲムシの18S塩基配列を再解析した結果，この中に革翅目（ハサミムシ）の18Sと判断できる配列が混入していることが明らかとなった．図６には，ジュズヒゲムシとハサミムシの18S塩基配列に基づき構築した近隣結合樹を示している．問題の配列は，Zorotypus_hubbardi_BYU_ACZO001 で，これは同じ種から決定された他の２本の配列（BYU_ACZORAP および DQ013288）とはわずか 73-76% の類似性しか示さないのみならず（図６の矢印），他のジュズヒゲムシから得られた７本の配列との類似性も低く，むしろハサミムシの18S塩基配列と類似していることを示している．二次構造等の詳細な解析結果からも，BYU_ACZO001 は他のジュズヒゲムシの配列とは大きく異なり，ハサミムシのそれと一致することが示された（Yoshizawa, 2010）．一方，BYU_ACZO001 と同一標本から得られたとされる28Sの配列は，BYU_ACZORAP, DQ013228 と同一の個体から得られた28S塩基配列と99%以上の類似性を示している．以上の結果は，BYU_ACZO001 の18S塩基配列が，ハサミムシのそれの混入であることを示す十分な証拠と言える．

このように明らかに別の生物に由来する配列が混入した場合，混入した配列に含まれる系統情報は，目的の生物に由来する他の遺伝子の配列に含まれる系統情報と矛盾するはずである．Terry らは，POY によって得られた結果と同時に，初期設定状態の Clustal で得られたアライメントを PAUP* で解析した結果も示している．そして，Clustal でアライメントしたデータの分割崩壊指数（partitioned Bremer support value）は，ジュズヒゲムシの単系統性に対して28Sが49，Histone 3 が17，形態が33といずれも強い支持を与えているのに対し，18S は-25と強い負の値（つまりジュズヒゲム

シの単系統性を支持しない値）を示している．この結果は，ハサミムシに由来する18S配列の混入を考えれば極めて妥当な結果である．一方，POY を用いてアライメントしたデータでは，系統的に遠く離れた生物からの配列の混入があるにも関わらず，ジュズヒゲムシの単系統性に対する18Sの分割崩壊指数は３と，ジュズヒゲムシの単系統性を支持する値となっている．Terry らの解析では，Histone 3 および28S, 18S の保存性の高い領域は事前にアライメントされ，解析を通して相同性の判断は固定されている．したがって，Histone 3 および 28S遺伝子が適切に増幅，シーケンスされていれば，これらの塩基配列の事前にアライメントされた領域は，ジュズヒゲムシの単系統性を支持する強い系統情報を含んでいると考えられる（実際，上で示した分割崩壊指数はそれを支持する）．さらに，形態データは目（order）を単位にコードされている．つまり，形態データは目を通して一定であるかのようにコードされており，したがって目の単系統性に対して非常に強いシグナルを含んでいる．このような状況のもと，全てのデータが統合され，POY による解析にかけられれば，ハサミムシからの混入である18Sの配列のうち，POY の解析対象となる挿入欠損を多数含む領域は，あたかもそれがジュズヒゲムシの18Sと近縁な配列であるかのようにアライメントされる．またナナフシの例で示した通り，こうしてアライメントされた挿入欠損を多数含んだ領域から生成される偽の系統情報は，事前にアライメントされた領域に含まれる真の系統情報を圧倒しうる．結果として，ハサミムシの配列からでさえ，ジュズヒゲムシの単系統性を支持する偽の系統情報が生成されたものと考えられる．

ここでは人為的に混入した配列の場合を検討したが，自然現象としても遺伝子ごとにその系統的背景が異なる場合は生じうる（遺伝子浸透，組み替え，lineage sorting など）．これらの現象は，species-tree を推定する上ではノイズと見なされることもあるが，特に集団レベルでの研究では非常に重要な情報と見なされる（例えば Ohshima and Yoshizawa, 2010 など）．直接最適化法による最適化は，このような重要な系統情報をかき消すことにもつながる．

■ Morgan and Kelchner (2010) による哲学的問題の指摘

以上，直接最適化法によってもたらされるバイアスの問題を Yoshizawa (2010) にそって指摘してきたが，この論文の出版とほぼ時を同じくして，Morgan and Kelchner (2010) によって直接最適化法の哲学的問題点も指摘されている．Morgan らが問題としているのは，樹形のみに基づいて相同性の決定を行う直接最適化法の方法が正当化できるか否かという点である．筆者は Morgan らの主張に強く同意し，またこれは直接最適化法に対する批判の決定打の一つになりうると考えているので，紹介したい．

直接最適化法では，樹形への一致性「のみ」に基づいて，塩基置換や挿入欠損を最小化するアライ

メントが生成される．このように樹形への一致性に基づいて判定される相同性は，secondary homology と呼ばれ，系統推定に先立って類似性等に基づいて推定される相同性（primary homology）とは区別される （De Pinna, 1991）．そして，直接最適化法では初期アライメントに基づき最短樹が推定され，その樹の上で塩基の置換や挿入欠損を最小化するように，相同性の判断が変更される．この相同性の判断過程は，dynamic homology approach と呼ばれる （Wheeler *et al.*, 2006）．また，データマトリクス上のそれぞれの列は，相同と判断された「形質」を表しており，そして同一の列に入力された値 （0,1 とか ACGT-）はその相同な形質の「形質状態」を表している．

Morgan らが問題とするのは，このような樹形への一致性のみに基づく，「形質」の相同性の判断の変更である．彼らの示した例にそって，形態形質を例に考えてみる．鳥とコウモリの翼は，四足動物の前脚という点で相同である．したがって，データマトリクスの作成に際して翼は，前脚という「形質」の「形質状態」としてコードされる．そして，系統解析により鳥とコウモリが遠縁の生物と判断され，翼がこれらの生物で独立に起源したことが明らかとなったとしても，この系統樹に基づいて独立起源（secondary homology としての非相同性）が明らかにされるのは，あくまでも翼という「形質状態」であり，それらが前脚として「相同」であることには変わりはない．系統樹との不一致だけをもとに，「形質」の相同性の判断が変更されることはありえない．これは分子の場合でも同様である．例えばあるアライメントされた塩基配列データの第11番目の塩基の状態が，系統推定の結果，独立に２回 A から T に変わったことが示されたとする．しかし，独立に生じた T 同士は，依然としてこの配列の第11番目の塩基座位と言う「形質」として相同であり，系統樹に基づいて独立起源が示されるのは，あくまでも11番目の塩基座位の「形質状態」である．このように系統樹への一致基準は，「形質状態」の相同性は検証できる一方，「形質」の相同性に対し，何の手がかりももたらさない．「形質」の相同性の推定は，系統樹への一致性「以外」の基準，つまり primary homology を推定するための基準を用いて決定されなければならない．

Primary homology としての相同性をどのように定義し，推定し，検証するかは非常に議論のある所であるが（この問題は深く追求しないが，興味のある方は，倉谷，1999 が参考となる．一方，secondary homology としての相同性の客観的検証基準は，上記の通り確立されている），比較生物学や体系学を実践する上で，類似性が最も重要な相同性の推定基準として用いられてきた．直接最適化法でも，塩基の個々の状態 （ACGT） の類似性（compositional similarity）は考慮される．しかし，ACGT が表すのはあくまでも「形質状態」であり，異なる配列間の同一の塩基状態が相同かどうかを判断するためには，それらの塩基座位 （topographical similarity or topological correspondence） としての相同性が検討されなければならない．逆に，塩基座位として相同であると判断されれば，配列間で異なる状態を示す A と T が相同形質となる場合も当然ある．つまり，単なる塩基の状態の一致は，相同性の推定基準としては適切ではなく（ことに塩基配列の場合，1/4 の確率で同じ形質状態を示

す），塩基座位としての相同性の判断には，配列のブロック単位での類似や，RNA の二次構造等を参照する必要がある．個々の「形質状態」の類似と，それらの樹形への一致度のみで相同性の判断を決定する直接最適化法および dynamic homology という概念は正当化できない．

では，hypervairable region のように，形質の類似に基づくアライメントが困難な領域はどのように扱うべきであろうか？Morgan らは，これらの領域は解析から除外すべきであると主張する．それは，それらの領域が系統情報を含まないからでも，ノイズしか含まないからでもなく（Lee, 2001），それらの領域の配列間の相同性を判断する根拠が示せないからである．単に数学的な最適化のみに基づいてアライメントされた領域からは，生物学的に意味のある情報は得られない．

以上が，Morgan らの指摘の概要である．一方で Wheeler and Giribet (2009) は"Alignments are not an attribute of nature...their role can only be as a heuristic tool in the solution of phylogenetic problem"と述べ，そもそも真のアライメントは存在しないとの立場を取っている．しかし，変化を伴う由来が現実の自然現象として存在する以上，相同性も現実の自然現象として存在し，そしてアライメントは配列間の塩基座位の相同性に関する言明に他ならない．直接最適化法によって生成されたアライメントが，単に樹長を最小化するという「だけ」の理由で最良と判断されるのであれば，その中には生物学的に正当化できるような相同性に関する言明は存在しえない．相同形質が持つ系統情報に基づいて推定されたものではない「樹」は，見た目は似ていたとしても系統樹とは似て非なるものである．相同性が存在し，そしてそれは発見，認識できるという考えは，比較生物学，進化生物学の中心的概念である（Morgan and Kelchner, 2010）．Wheeler らの言明は，その最も重要な概念すら否定するものである．

■ おわりに

"...the safest bet is to simply ignore this study..." (Wiens, 2007)

以上，Yoshizawa (2010) および Morgan and Kelchner (2010) のレビューを通して，POY によって生成された系統樹のようなものは，文字通り「ポイ」と捨て去るべきであるとする筆者の考えを示した．「はじめに」で般若心経の一節を引用したが，POY によって生成された５つの形質（ACGT-）のアライメントも，まさに皆空である．本稿が，POY 樹による苦厄から解放される一助になれば幸いである．もちろん以上示してきた筆者の考えには誤りがあるかもしれない．その場合，ぜひ問題点の指摘や反論等お聞かせ頂きたい．

参考までに，挿入欠損を多数含む解析を行う際，筆者が用いている方法を紹介したい．まず，リボソーム RNA コード領域などの解析にあたっては，分子のより高次の構造である二次構造の相同性を利用すべきと考えている．二次構造に基づくアライメント法は，Kjer *et al.* (2009) によって詳しく紹介されている．またイントロンなど，分子の高次の構造を利用できない領域の解析にあたっては，Dialign-TX (http://dialign-tx.gobics.de/) が非常に良好なアライメント結果を出す（Simmons *et al.*,

2008a）．Dialign は他の多くのアライメントソフトと次の２点で大きく異なる: (1) 配列全体を通してアライメントを最適化するのではなく，類似したローカルな配列断片を集積することによりアライメントを行う（断片集積法）．そのため，組み替えが生じているような配列のアライメントにも非常に高いパフォーマンスを示す （Ohshima and Yoshizawa, 投稿中）; (2) 塩基置換や挿入欠損に対するコスト，つまりアライメントのパラメータを設定する必要がない．言うまでもないことだが，いずれの方法を用いても，二次構造の相同性が確定できた領域や，配列のブロック単位での類似性が相同性に由来すると推測できる領域のみを解析に用いるべきで，hypervariable region のような領域を解析に含めるべきでない．

最後にあらためて強調したいのは，「アライメントと系統推定を同一の認識論的問題として見なす」とする直接最適化法の大目標が，すでに崩されているということである（Simmons, 2004; Morgan and Kelchner, 2010）．この点は，最適化基準として尤度や事後確率を採用するといった微調整 （Liu *et al.*, 2009）では修正できない．RNA の二次構造なども情報として組み込み，系統樹と相互に参照しながらアライメントと樹形を「同時推定」するための自動化アルゴリズムは可能だろうし，そのワークフローは図 1 と類似したものとなるだろう．しかしこうした「同時推定法」が開発されたとしても，大目標が崩されている以上，それは直接最適化法の発展型と見なすべきではなく，Hennig (1966) の相互参照法（reciprocal illumination method: checking, correcting, and rechecking: p. 21）を共通祖先とする，別クレードのアルゴリズムと見なすべきである．

■ 謝辞

Karl Kjer，David Morrisson 両博士には，本稿で述べた問題点について有益な議論をいただいた．松村洋子氏には，投稿前の原稿を通読していただき，様々な指摘をいただいた．三中信宏博士には，論文の改訂にあたって議論いただいた．お礼申し上げる．もちろんこれは，上記の方々が筆者の主張に同意していることを意味するものではない．

■ 文献

Agolin, M. and D'Haese, C. A. 2009. An application of dynamic homology to morphological characters: direct optimization of setae sequences and phylogeny of the family Odontellidae (Podouromorpha, Collembola). *Cladistics*, 25: 353–385.

Bertelli, S. and Giannini, N. P. 2005. A phylogeny of extant penguins (Aves: Sphenisciformes) combining morphology and mitochondrial sequences. *Cladistics*, 21: 209–239.

De Pinna, M. C. C. 1991. Concepts and tests of homology in the cladistic paradigm. *Cladistics*, 7: 367–394.

Faivovich, J., Haddad, C. F. B., Baêta, D., Jungfer, K. H., Álvares, F. R., Brandão, R. A., Sheil, C.,

Barrientos, L. S., Barrio-Amorós, C. L., Cruz, C. A. G. and Wheeler, W. C. 2010. The phylogenetic relationships of the charismatic poster frogs, Phyllomedusinae (Anura, Hylidae). *Cladistics*, 26: 227–261.

Frost, D. R., Etheridge, R., Janies, D. and Titus, T. A. 2001. Total evidence, sequence alingment, evolution of polychrotid lizards, and a reclassification of the Iguania (Squamata: Iguania). *American Museum Novitates*, 3348: 1–38.

Frost, D. R., Grant, T., Faivovich, J. *et al*. 2006. The amphibian tree of life. *Bulletin of the American Museum of Natural History*, 297: 1–370.

Giannini, N. P. and Simmons, N. B. 2003. A phylogeny of megachiropteran bats (Mammalia: Chiroptera: Pteropodidae) based on direct optimization analysis of one nuclear and four mitochondrial genes. *Cladistics*, 19: 496–511.

Giribet, G. 2001. Exploring the behavior of POY, a program for direct optimization of molecular data. *Cladistics*, 17: S60–70.

Giribet, G., Distel, D. L., Polz, M., Sterrer, W. and Wheeler, W. C. 2000. Triploblastic relationships with emphasis on the acoelomates and the position of Gnathostomulida, Cyclophora, Plathelminthes, and Chaetognatha: a combined approach of 18S rDNA sequences and morphology. *Systematic Biology*, 49: 539–562.

Giribet, G. and Edgecombe, G. D. 2006. Conflict between datasets and phylogeny of centipedes: an analysis based on seven genes and morphology. *Proceedings of the Royal Society (B)*, 273: 531–538.

Giribet, G., Edgecombe, G. D. and Wheeler, W. C. 2001. Arthropod phylogeny based on eight molecular loci and morphology. *Nature*, 413: 157–161.

Giribet, G., Okusu, A., Lindgren, A. R., Huff, S. W., Schrödl, M. and Nishiguchi, M. K. 2006. Evidence for a clade composed of molluscs with serially repeated structures: Monoplacophorans are related to chitons. *Proceedings of the National Academy of Science of the United States of America*, 103: 7723–7728.

Grant, T. and Kluge, A. G. 2003. Data exploration in phylogenetic inference: scientific, heuristic or neither. *Cladistics*, 19: 379–418.

Hennig, W. 1966. *Phylogenetic Sytematics*, 263 pp. University of Illinois Press, Urbana.

Kjer, K. M., Roshan, U. and Gillespie, J. J. 2009. Structural and evolutionary consideration for multiple sequence alignment of RNA, and the challenges for algorithms that ignore them. *In* Rosenberg, M. S. (ed.), *Sequence Alignment: Methods, Models, Concepts, and Strategies*, pp. 105-149, University of California Press, Berkley.

倉谷　滋　1999. 相同性とは何か–発生と進化とを結びつける形態学的認識–. 棚部一成，森　啓（編），古生物の科学２ 古生物の形態と解析，pp. 1–33, 朝倉書店，東京.

Lee, M. S. Y. 2001. Unalignable sequences and molecular evolution. *Trends in Ecology and Evolution*, 16: 681–685.

Liu, K., Raghavan, S., Nelesen, S., Linder, C. R. and Warnow, T. 2009. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science*, 324: 1561–1564.

Mickevich, M. F. and Farris, J. S. 1981. The implications of congruence in *Menidia*. *Systematic Zoology*, 30: 351–370.

Morgan, M. J. and Kelchner, S. A. 2010. Inference of molecular homology and sequence alignment by direct optimization. *Molecular Phylogenetics and Evolution*, 56: 305–311.

村上哲明　1998. [evolve:3375] Molecular vs morphology at Kyushu U（ポストシンポ –メーリングリスト EVOLVE より–). *Panmixia*, 12: 24–26.

Ogden, T. H. and Rosenberg, M. S. 2007. Alignment and topological accuracy of the direct optimization approach via POY and traditional phylogenetics via ClustalW + PAUP*. *Systematic Biology*, 56: 182–193.

Ogden, T. H., Whiting, M. F. and Wheeler, W. C. 2005. Poor taxon sampling, poor character sampling, and non-repeatable analyses of a contrived data set do not provide a more credible estimate of insect phylogeny: a reply to Kjer. *Cladistics*, 21: 295–302.

Ohshima, I. and Yoshizawa, K. 2010. Differential introgression causes genealogical discordance in host races of Acrocercops transecta (Insecta: Lepidoptera). *Molecular Ecology*, 19: 2106–2119.

Ohshima, I. and Yoshizawa, K. 投稿中. The utility of indels in population genetics: the *Tpi* intron for host race genealogy of *Aceocercops transecta* (Insecta: Lepidoptera).

Robillard, T., Legendre, F., Desutter-Grandcolas, L. and Grandcolas, P. 2006. Phylogenetic analysis and alignment of behavioral sequences by direct optimization. *Cladistics*, 22: 602–622.

Simmons, M. P. 2004. Independence of alignment and tree search. *Molecular Phylogenetics and Evolution*, 31: 874–879.

Simmons, M. P., Richardson, D. and Reddy, S. N. 2008a. Incorporation of gap characters and lineage-specific regions into phylogenetic analysis of gene families from divergent clades: and example from the kinesin superfamily across eukaryotes. *Cladistics*, 24: 372–384.

Simmons, M. P., Müller, K. F. and Webb, C. T. 2008b. The relative sensitivity of different alignment methods and character coding in sensitivity analysis. *Cladistics*, 24: 1039–1050.

Terry, M. D. and Whiting, M. F. 2005. Mantophasmatodea and phylogeny of the lower neopterous insects. *Cladistics*, 21: 240–257.

上島　励　2008. 節足動物の分子系統学，最近の展開．石川良輔（編），バイオディバーシティ・シリーズ６ 節足動物の多様性と系統，pp. 28–48, 裳華房，東京.

Wheeler, W. C., Aagesen, L., Arango, C. P. *et al.* 2006. *Dynamic Homology and Phylogenetic Systematics: A unified Approach using POY*. 365 pp. American Museum of Natural History, New York.

Wheeler, W. C. and Giribet, G. 2009. Phylogenetic hypotheses and the utility of multiple sequence alignment. In Rosenberg, M. S. (ed.), *Sequence Alignment: Methods, Models, Concepts, and Strategies*, pp. 95–104, University of California Press, Berkley.

Wheeler, W. C. and Hayashi, C. Y. 1998. The phylogeny of the extant chelicerate orders. *Cladistics*, 14: 173-192.

Wheeler, W. C., Whiting, M., Wheeler, Q. D. and Carpenter, J. 2001. The phylogeny of the extant hexapod orders. *Cladistics*, 17: 113–169.

Whiting, M. F., Bradler, S. and Maxwell, T. 2003. Loss and recovery of wings in stick insects. *Nature*, 421: 264–267.

Wiens, J. J. 2007. The Amphibian Tree of Life. Bulletin of the American Museum of Natural History, Number 297. By Darrel R Frost, Taran Grant, Julián Faivovich, Raoul H Bain, Alexander Haas, Célio F B Haddad, Rafael O De Sá, Alan Channing, Mark Wilkinson, Stephen C Donnellan, Christopher J Raxworthy, Jonathan A Campbell, Boris L Blotto, Paul Moler, Robert C Drewes, Ronald A Nussbaum, John D Lynch, David M Green, and Ward C Wheeler. *Quarterly Review of Biology*, 82: 55–56.

Wong, K. M., Suchard, M. A. and Huelsenbeck, J. P. 2008. Alignment uncertainty and genomic analysis. *Science*, 319: 473–476.

Worsaae, K., Nygren, A., Rouse, G. W., Giribet, G., Persson, J., Sundberg, P and Pleijel, F. 2005. Phylogenetic position of Nerillidae and Aberranta (Polychaeta, Annelida), analysed by direct optimization of combined molecular and morphological data. *Zoologica Scripta*, 34: 313–328.

吉澤和徳　2008. 六脚類の高次分類体系と進化．石川良輔（編），バイオディバーシティ・シリーズ 6 節足動物の多様性と系統，pp. 297–329, 裳華房，東京．

Yoshizawa, K. 2010. Direct optimization overly optimizes data. *Systematic Entomology*, 35: 199–206.

Yoshizawa, K. and Johnson, K. P. 2005. Aligned 18S for Zoraptera (Insecta): Phylogenetic position and molecular evolution. *Molecular Phylogenetics and evolution*, 37: 572–580.

図の説明

図１．直接最適化法のフローシート．

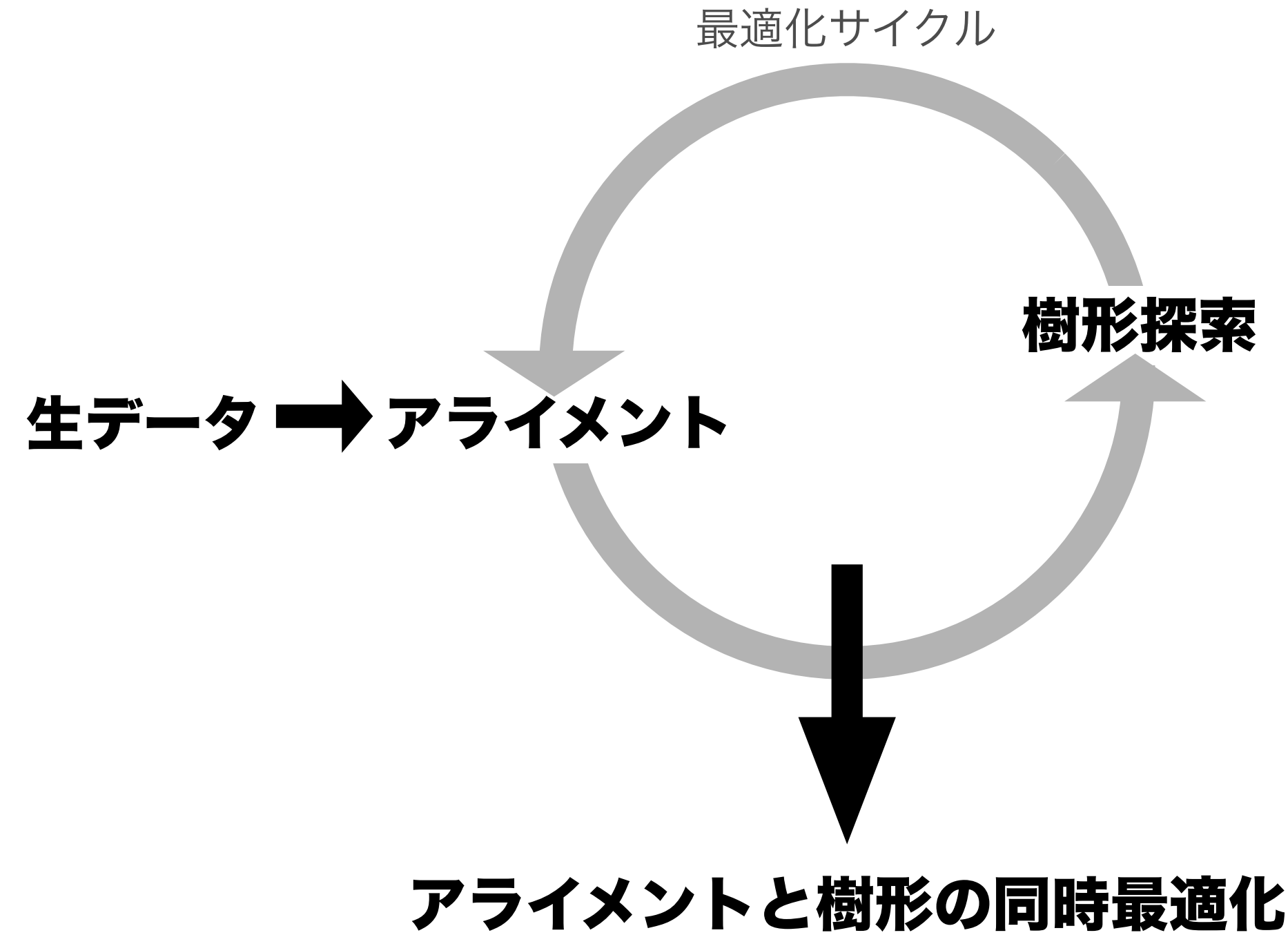
図２．実データ（上）と，実データと統合した上で POY でアライメントしたランダムデータ（下）から推定したブーツストラップ合意樹．ランダムデータの解析には，２つのパラメータ設定を適用した（ギャップ：トランジッション：トランスバージョン = 15:1:1 および 1:1:1）．破線は，実データの系統樹と矛盾した枝を，*印は枝のブーツストラップ確率を示す（***=100%, **≥90%, *≥70%, 無印≥50%）．

図３．POY を用いて単独でアライメントしたランダムデータから推定したブーツストラップ合意樹．詳細は図２の説明を参照．

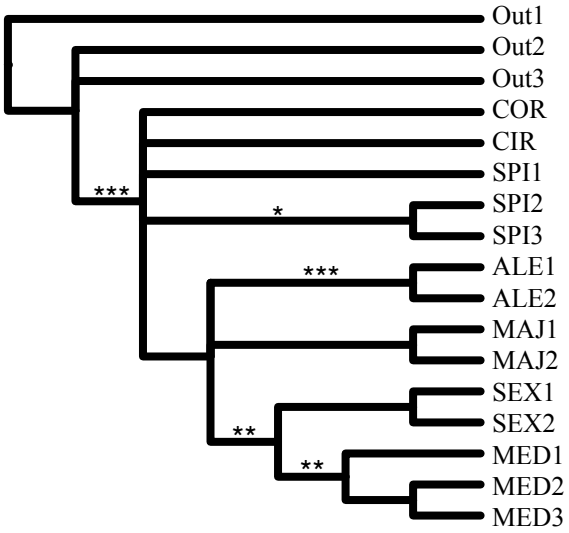
図４．Whiting *et al.* (2003) で解析されたデータの一部の鳥瞰図．濃度の違うグレーは異なる塩基の状態を，白抜きはギャップもしくは missing data を示す．

図５．Whiting *et al.* (2003) で解析されたデータの全データ（中央），挿入欠損の無い形質（右），および挿入欠損を含む形質（左）から推定した再節約樹．*印は図２の説明参照．Xはブーツストラップ確率50%以下の枝を示す．

図６．ジュズヒゲムシとハサミムシの18S塩基配列から構築した近隣結合樹．矢印は，同一種（*Zorotypus hubbardi*）としてラベリングされた配列を示す．


最適化サイクル
樹形探索
生データ
アライメント
アライメントと樹形の同時最適化

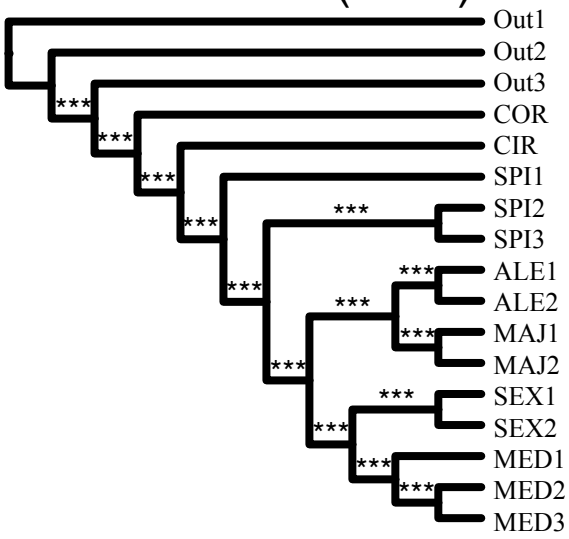
real data



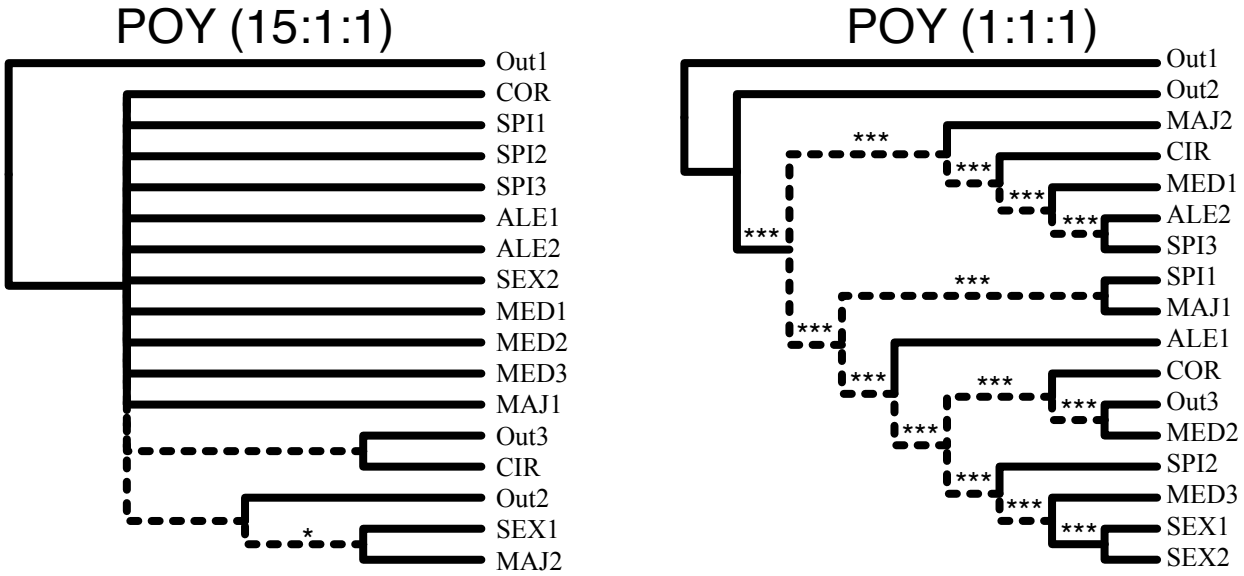
random data: aligned with real data

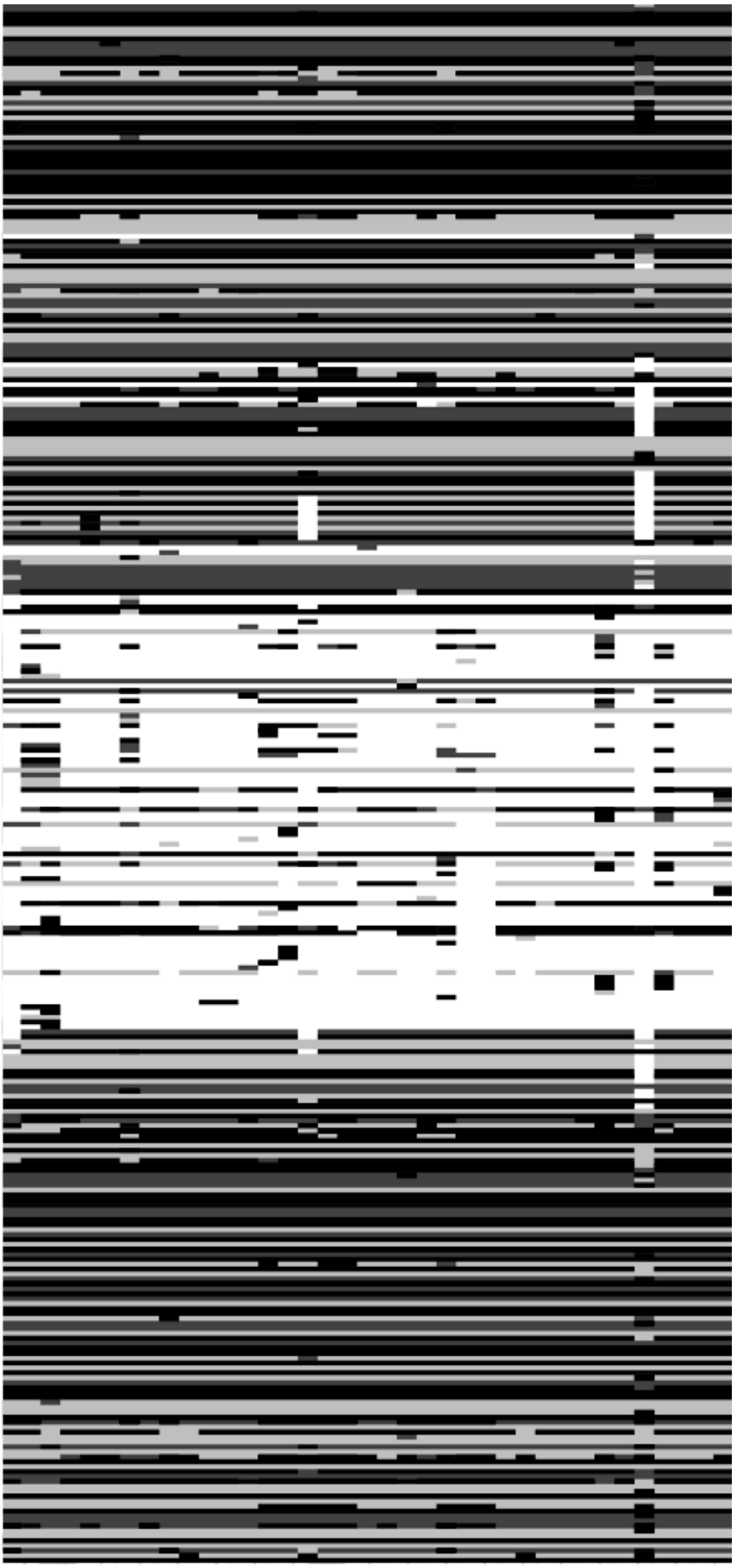
POY (15:1:1)

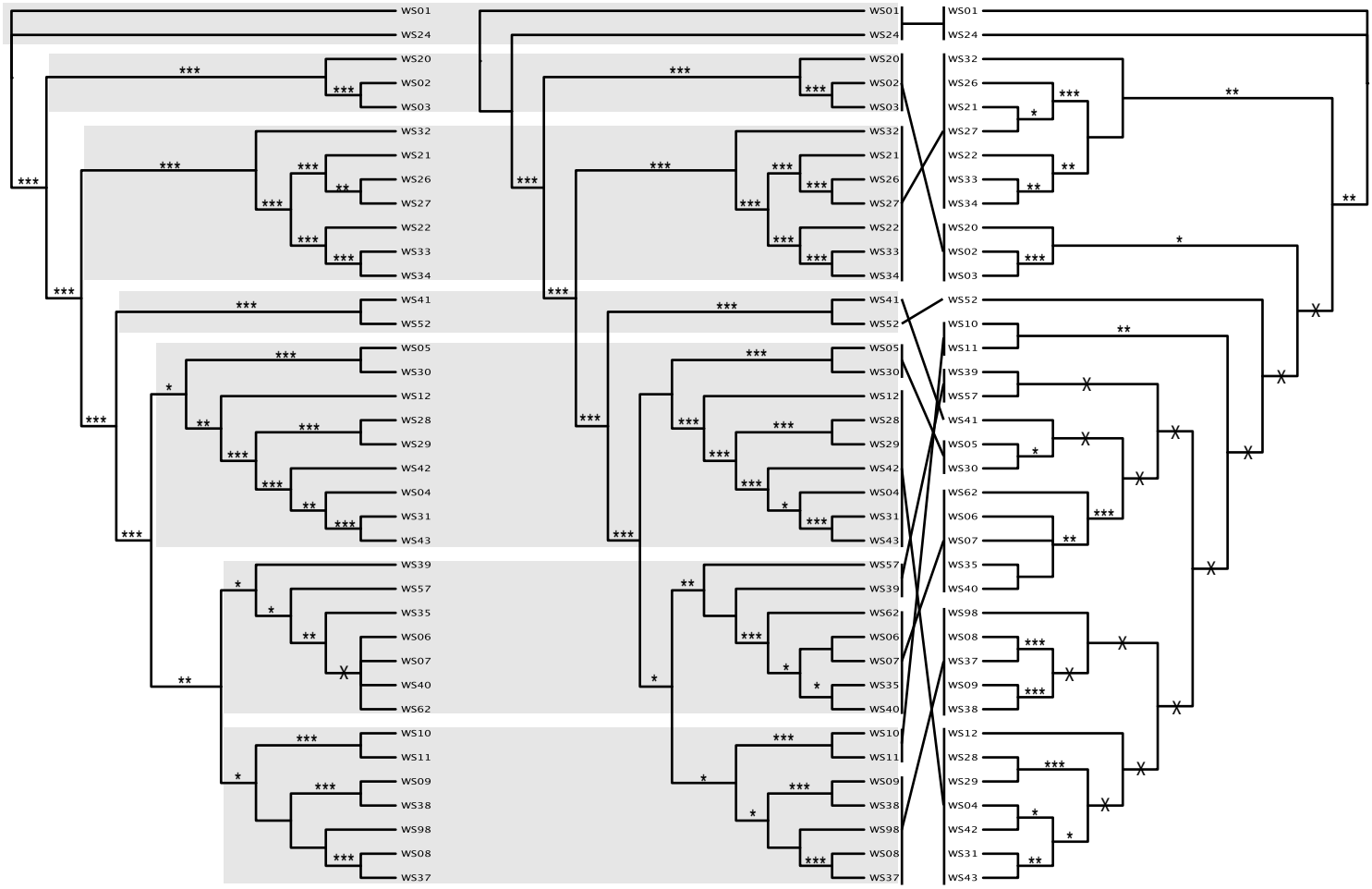
POY (1:1:1)

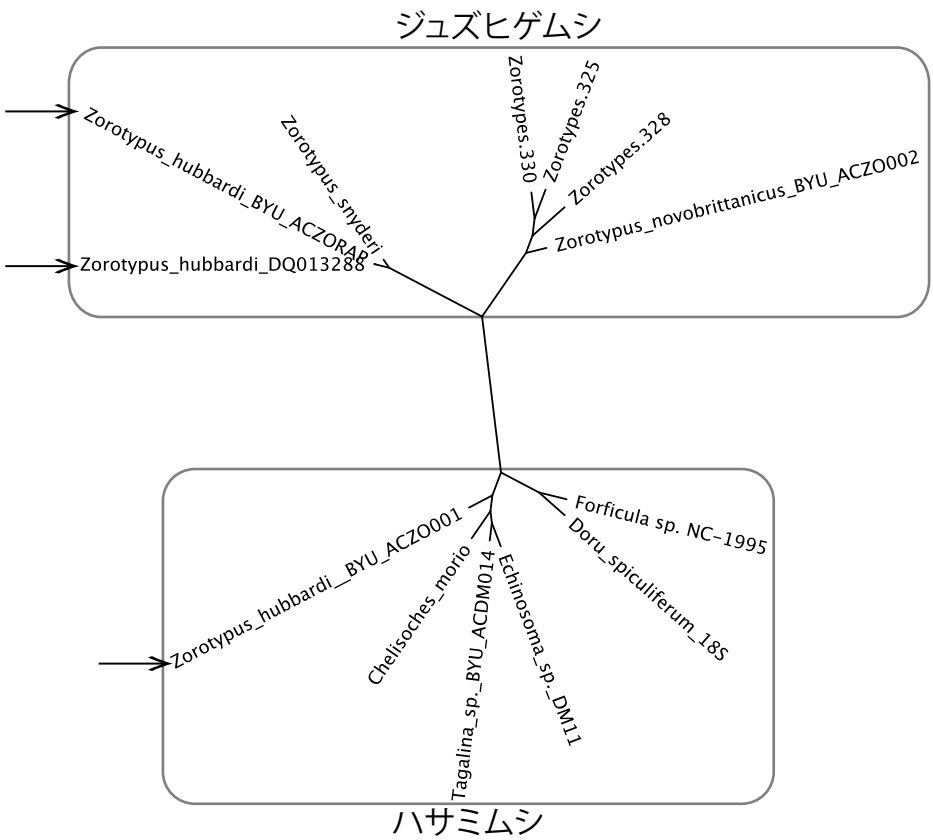


random data: aligned independently






WS01
WS24
WS20
WS02
WS03
WS32
WS21
WS26
WS27
WS22
WS33
WS34
WS41
WS52
WS05
WS30
WS12
WS28
WS29
WS42
WS04
WS31
WS43
WS39
WS57
WS35
WS06
WS07
WS40
WS62
WS10
WS11
WS09
WS38
WS98
WS08
WS37


ジュズヒゲムシ
Zorotypus_hubbardi_BYU_ACZORAB
Zorotypus_snyderi
Zorotypes.330
Zorotypes.325
Zorotypes.328
Zorotypus_novobrittanicus_BYU_ACZO002
Zorotypus_hubbardi_DQ013288
Zorotypus_hubbardi__BYU_ACZO001
Chelisoches_morio
Tagalina_sp._BYU_ACDM014
Echinosoma_sp._DM11
Forficula sp. NC-1995
Doru_spiculiferum_18S
ハサミムシ