Pedro Dias de Oliveira

Filogenética de Pilocarpinae (Rutaceae)

São Paulo

2007

Pedro Dias de Oliveira

Filogenética de Pilocarpinae (Rutaceae)

Tese apresentada ao Instituto de Biociências da Universidade de São Paulo, para a obtenção do Título de Doutor em Ciências, área de Botânica.

Orientador: Prof. Dr. José Rubens Pirani

São Paulo

2007

Dias, Pedro. Filogenética de Pilocarpinae (Rutaceae). 273p.

Tese (Doutorado) – Instituto de Biociências da Universidade de São Paulo. Departamento de Botânica.

1. Filogenia 2. Pilocarpinae. 3. Rutaceae. I. Universidade de São Paulo. Instituto de Biociências. Departamento de Botânica.

Comissão julgadora:

Prof(a). Dr(a).

Prof(a). Dr(a).

Prof(a). Dr(a).

Prof(a). Dr(a).

Prof. Dr. José Rubens Pirani

Orientador

Dedicatória

À *Renata*, pelo amor, paciência e colaboração infindáveis.

Epígrafe

While it is generally agreed that the reconstruction of evolutionary trees should ideally be regarded as a problem in statistical inference, few approaches to evolutionary taxonomy have taken into account the full implications of that premisse.

...

The only statistical optimality principle which would seem readily applicable to the case of estimated trees is that the selected tree should be the most probable on the basis of available data.

...

$$^{1}P\{E|D\} = \frac{P\{D|E\}P\{E\}}{P\{D\}}$$

J. S. Farris (p. 250) sobre inferência filogenética.

(*Syst. Biol.* 22: 250–256. 1973.)

[1] Teorema de Bayes.

Agradecimentos

- Ao Dr. José Rubens Pirani pela orientação, dedicação, ajuda e suporte imprescindíveis sempre que foram necessários.
- À Dra. Maria Elisabeth van den Berg pela iniciação na sistemática, pelas ajudas incondicionais e pelos "incentivos" para aprender inglês e alemão.
- À Dra. Jacquelyn Kallunki pela ajuda, colaboração e paciência durante minha estadia no Jardim Botânico de New York e pela bolsa concedida através de seu projeto, o que viabilizou minha ida a New York.
- Ao Dr. Kenneth Cameron e equipe pela utilização do Laboratório do International Plant Research Center do Jardim Botânico de New York e pelo fornecimento dos iniciadores de ITS.
- À Dra. Mariana Cabral de Oliveira e equipe pelo uso do Laboratório de Algas Marinhas, pelo auxílio sempre que necessário e pelas sugestões.
- À Dra. Roseli Aparecida Leandro pelas discussões e esclarecimentos sobre análise bayesiana.
- Ao Dr. Antonio Carlos Marques pelas discussões (e, apartir de agora, talvez, pelas discordâncias).
- Ao Dr. Sérgio Russo Matioli pelas discussões.
- À Comissão Avaliadora do exame de ingresso ao Curso de Pós-graduação (Drs. Eny I. S. Floh, Flávio A. S. Berchez, Maria Luiza F. Salatino e Nanuza L. Menezes) por ter recomendado a minha passagem para o Doutorado Direto. Espero ter correspondido à indicação.
- Aos Drs. Antonio Salatino e Maria Luiza Faria Salatino pelas sugestões iniciais.
- Aos funcionários da Secretaria do Departamento de Botânica (Carlos, Cesário & Norberto) e de Pós-graduação (Erika, Helder, Teresa & Vera) do IB-USP, sempre prestativos e bem dispostos.

- Às funcionárias da Biblioteca do IB-USP pela paciência e disposição em me atender incontáveis vezes às 9h30' da noite (minutos antes de fechar), especialmente quando os livros que eu pegava (uns que quase ninguém usava) precisavam ser cadastrados no novo sistema.

- Aos docentes e discentes do Laboratório de Sistemática Vegetal pela dinamicidade invejável.

- Aos vigilantes do prédio "Sobre-as-ondas" e "Genoma Humano" (Antonio, Naldo, Ulisses etc. – os "Cabras-sem-qualidade") pelos cafés que "filei" nas madrugadas.

- À Universidade de São Paulo e ao Instituto de Biociências pelo Curso de Pós-graduação.

- À Pró-reitoria de Pós-graduação da Universidade de São Paulo pelos auxílios concedidos para participação em eventos.

- Ao CNPq pelos 3 meses de bolsa concedida.

- À FAPESP pelos 45 meses de bolsa concedida (Processo 02/09762-6) e pelo Auxílio à Pesquisa concedido ao Dr. José Rubens Pirani (Processo 04/15141-0), o que viabilizou parte considerável desta tese.

- À assessoria da FAPESP por ter confiado em mim e no projeto proposto (embora um "pouco" diferente do que está sendo apresentado nesta tese). Espero que o resultado mereça a confiança depositada.

- À National Science Foundation pela bolsa concedida.

- Às Dras. Susan Pell e Allison Miller pela bolsa para apresentar palestra no Colóquio "Evolution and diversification in the Sapindales" (recursos oriundos da American Society of Botany - Genetics Section, do Brooklyn Botanical Garden e do Torrey Botanical Club), em Chicago.

- A todos os colaboradores que contribuíram para esta tese, entre eles:
  Abel Canguçú ("O Teimoso". Você precisa fazer o café igual ao da Botânica!)
  Oscar Iza e equipe - HBR & UNIVALI
  Osmar dos Santos Ribas e equipe - MBM
  Renato de Jesus e equipe - Companhia Vale do Rio Doce

- A todos os curadores de coleções, e respectivas equipes, que atenderam aos meus pedidos de empréstimo e/ou visitas: ALCB, CEPEC, HRCB, HUEFS, IAN, INPA, MBM, MG, MICH, NY, RB e US.

- Ao IBAMA pelas licenças de coleta (2005 e 2006), embora a renovação solicitada em maio ainda não tenha saído (e eu pensava que o pedido eletrônico seria mais rápido).

- Aos colegas da Bioinformática, Glauber Brito e Tarik el Jundi, pelos seminários sobre análise bayesiana, Cadeias de Markov, teoria de grafos, máxima verossimilhança, *microarrays*, p/z-value e PERL.

- Aos amigos de Santarém, pelos "remédios" e ajudas incondicionais[2] que tornaram possível (e posteriormente mais agradável) minha estadia nessa cidade:
  M.Sc. Chieno Suemitzu
  Dr. Reginaldo Rodrigues
  Dr. Ricardo Oliveira
  M.Sc. Rubens Yuki

- Às equipes do Departamento de Botânica, do Departamento de Zoologia e da Biblioteca (Santas Bibliotecárias!) do MPEG por terem tornado possível minha iniciação à sistemática (e principalmente pelos "remédios"). Entre essas pessoas estão: "Carlito", "Beleza", Cosme, Dr. Horácio Higuchi, Dr. Keid Nolan, M.Sc. Moisés Mourão, Luis Carlos, Jarilson ("Macaco"), Maria das Graças ("Do Céu"), D. Maria e D. Raimunda.

- À minha família por sempre estar na torcida - o que quer que signifique "filogenia" !

- À M.Sc. Renata Giassi Udulutsch ("Re") pela leitura crítica da versão final desta tese.

- A todos os outros colaboradores (que por algum motivo[3] não estou lembrando agora).

- Por último, mas não menos, ao povo brasileiro por pagar seus impostos, permitindo que pessoas como eu possam usufruir de bolsas de estudo, aqui vai o balanço do prejuízo durante o meu doutorado:

  Total de recursos recebidos (descontada a CPMF): R$ 113.307,63 (cento e treze mil, trezentos e sete reais e sessenta e três centavos).

Esta tese foi preparada com LaTeX 2_ε e GNU/Linux.
São gratuitos, use-os!

[2] "Remédios" também são incondicionais!
[3] Falta de remédio, talvez.

Sumário

Prefácio

Como disseram Meynk & Tweedie ([1]), nos prefácios podem ser colocadas diversas coisas: agradecimentos, desculpas, destaque da importância do que foi feito, as coisas novas (e as inacabadas) etc. Mas, em geral eles falam sobre o que vem a seguir. Assim, este (como o deles) não será exceção. Adicionalmente, usarei este prefácio como uma espécie[4] de resumo geral e meio informal (veja o Resumo para a versão formal).

Esta tese está dividida em três partes: I Filogenética Básica, II Filogenia de Pilocarpinae e III Novidades Taxonômicas em Esenbeckiinae.

Parte I Filogenética Básica - fornece uma rápida revisão de alguns dos métodos básicos atualmente em uso na filogenética, envolvendo aspectos teóricos e operacionais[5], assim como algumas de suas possíveis implicações. O **Capítulo 1** discute dois conceitos importantes (grupo e caráter) e interrelacionados, mostra como pode ser construído um modelo estocástico para o tratamento de caractres, enfatizando sua adequação e demonstranto como homologia e, por conseqüência grupos, podem ser adequadamente tratados sob ótica probabilística. O **Capítulo 2** apresenta os métodos de construção e otimização de árvores uzando máxima verossimilhnaça e análise bayesiana.

Parte II Filogenia de Pilocarpinae - apresenta a filogenia de Pilocarpinae baseada em dados moleculares (espaçadores ITS1, ITS2 e gene 5.8 S do DNA nuclear e espaçador *trn*G-S do DNA platidial) e a filogenia de *Pilocarpus* Vahl baseada em dados moleculares (mesmas regiões usadas anteriormente) e morfológicos. O **Capítulo 3** apresenta a filogenia em nível genérico da subtribo Pilocarpinae e de gêneros relacionados (*Helietta* e *Balfourodendron*), mostrando que, exceto *Esenbeckia*, os gêneros tradicionalmente reconhecidos (*Metrodorea*, *Pilocarpus* e *Raulinoa* - monoespecífico) emergem como monofiléticos (embora a subtribo não) e que *Balfourodendron* e *Helietta* (ambos da subtribo Pteleinae) possuem relações mais estreitas com parte dos gêneros de Pilocarpinae do que com o gênero-tipo de sua própria subtribo (Pteleinae) e reúnem-se em um clado caracterizado pela presença de inflorescências ramificadas, para o qual foi criada uma subtribo, ficando Pilocarpinae

[4] O que quer que seja uma.

[5] Principalmente, diriam alguns ...

monogenérica; além disso, este capítulo apresenta um protocolo[6] para detecção de *burn-in* em análises filogenéticas bayesianas usando métodos já bem estabelecidos em estudos de convergência de MCMC. Por sua vez, o **Capítulo 4** apresenta a filogenia das espécies de *Pilocarpus* baseada em dados morfológicos e moleculares; essa filogenia, associada a simulações computacionais, é utiliza como base para traçar hipóteses evolutivas sobre os padrões foliares e de estivação da corola no gênero, mostrando como os estados desses caracteres se comportam nas árvores obtidas e quão apropriado é utilizar os diferentes estados como sinapomorfias/homoplasias usando o método MCMC como base e contrastando com o mapeamento com parcimônia, deixando claro que sinapomorfia/homplasia é mais adequadamente tratada como uma questão de probabilidade.

Parte III Novidades Taxonômicas em Esenbeckiinae representa um reflexo das atividades de campo e de análise de material de herbário. No **Capítulo 5** é apresentada uma redescrição de *E. cowanii* Kaastra, espécie anteriormente conhecida apenas da Guiana Francesa e apenas pelo material tipo, cuja morfologia floral era desconhecida e foi encontrada nos Estados do Acre, Mato Grosso, Pará e Rondônia durante as expedições de campo que fiz para a Amazônia; além disso, é proposto um epítipo para o táxon. O **Capítulo 6** apresenta a descrição de uma nova espécie de *Esenbeckia* (embora ainda sem diagnose latina), coletada nos estados do Acre e Rondônia e caracterizada pela posse de brácteas persistentes.

Para os capítulos 3 e 4 existem Apêndices e Material Suplementar, os quais são apresentados no final do respectivo capítulo. Os apêndices apresentam informações essenciais geralmente exigidas quando da publicação do artigo, como matrizes, descrição dos caracteres morfológicos, informações sobre os *vouchers* utilizados nos estudos ou protocolos originais não encontrados em outro lugar na literatura (*i.e.*, desenvolvidos nesta tese). O Material Suplementar é formado por arquivos resultantes das análises realizadas e não são colocados como apêndices ou por serem muito volumosos ou por terem importância secundária, mas são apresentados no DVD-ROM que acompanha esta tese.

Por último, as Figuras 2.3, 2.11, 2.12, 2.13, 2.14 e 2.15, todas do Capítulo 2, foram modificadas de uma apresentação obtida na internet, mas não consegui achar o endereço novamente para citá-lo.

Para facilitar a leitura no computador[7], esta tese possui *hyperlinks* em toda a sua estrutura.

[6] Apenas o protocolo, não os métodos em si.
[7] Pensei em mim mesmo.

[1] Meynk, S. P. & Tweedie, R. L. 1993. *Markov chains and stochastic stability.* Springer-Verlag, London.

Pedro Dias
São Paulo
3 de Dezembro de 2007

Lista de Figuras

Lista de Tabelas

Parte I

Filogenética Básica

Capítulo 1

Grupos e Caracteres em Filogenética

Capítulo parcialmente publicado - Dias, P. *et al.* 2005. Monophyly vs. paraphyly in plant systematics. *Taxon* 54: 1039-1040.

1.1 Abstract

In this paper, I present a mini-review of two major concepts in phylogenetics, namely groups and characters, reinforce their importance and role, and also discuss some new interpretations to, and(or) appropriateness of, these terms. Among the concepts addressed here, monophyletic- and non-monophyletic groups (whose duality has recently grown up in the botanical literature) have a central position. Moreover, homology is discussed under the light of morphological and molecular data, taking into account both practical and operational issues, as well as its relevance to practicing phylogenetics. Then, I shall use the notion of stochasticity, demonstrate how to build an evolutionary model and emphasize the importance of models in phylogenetics. As an outcome, the meanings of character evolution and groups are reviewed and improved.

1.2 Resumo

Neste trabalho é apresentada uma rápida revisão sobre dois dos principais conjuntos de conceitos em filogenética (grupos e caracteres), destacando sua importância, reiterando seu papel ou discutindo novas formas de interpretação e/ou adequação dos mesmos. Dentre esses conceitos estão o de grupo monofilético e não-monofilético, cuja dualidade tem sido revitalizada recentemente na literatura botânica. Homologia é discutida em termos dos diferentes tipos de dados (morfológicos e moleculares), levando-se em conta sua adequabilidade tanto do ponto de vista prático como do operacional e suas respectivas relevâncias. Adicionalmente, é utilizada a noção de estocasticidade, é demonstrada a construção de um modelo probabilístico e enfatizada sua adequação e importante papel em filogenética. Conseqüentemente, a noção de evolução de caráter e de grupo são refinadas em termos probabilísticos.

1.3 Introdução

Desde as publicações iniciais de Hennig ([42], [43], [44]) a filogenética passou (e tem passado) por várias mudanças[1], as quais envolveram desde a maximização operacional do método (*e.g.*, Farris [18], Farris *et al.* [22]), a reinterpretação de aguns conceitos essenciais (*e.g.*, Nelson [72]), o desenvolvimento de novos termos (*e.g.*, Nelson [76]) ou até mesmo um completo divórcio (*e.g.*, Brower [3]) de algumas idéias consideradas fundamentais[2] quando da ampla divulgação das idéias henniguianas no ocidente.

Concomitante a essas mudanças, conseqüentemente, as discussões da área foram direcionadas basicamente a:

1) justificar o método dentro de um arcabouço filosófico (como pode ser visto, *e.g.*, em Heywood *et al.* [45], Kitts [55], Nelson [71], Platnick & Gaffney [86], Settle [103], Wiley [112]), atraindo atenção inclusive de filósofos (*e.g.*, Hull [47], [48], [49]);

2) (re)interpretar conceitos fundamentais (*e.g.*, Nelson [72], [74]; Platnick [87]); e

3) operacionalizar e automatizar o método (*e.g.*, Farris [18], Farris [17], Farris *et al.* [22], Kluge & Farris [58]).

Por outro lado, tão importante quanto os itens acima é vislumbrar a dimensão matemática e computacional do problema de inferir filogenias. Nesse sentido, alguns autores (*e.g.*, Cavalli-Sforza & Edwards [9], Edwards & Cavalli-Sforza [16], Felsenstein [25]) demonstraram que o problema de inferir as possíveis relações de um

[1] Para uma análise histórica mais completa sugere-se o livro de Hull [50].
[2] Por algumas pessoas, pelo menos (*e.g.*, Kluge [57]).

determinado grupo de organismos, dado o número de possibilidades (veja a Equação 2.1), era mais complicado do que poderia parecer inicialmente. O número total de árvores pode se tornar impossível de calcular analiticamente (NP completo) e soluções heurísticas tiveram que ser implementadas (veja o item 2.4.2 do Capítulo 2), o que levou a um desenvolvimento mais acelerado de algoritmos computacionais na área (*e.g.*, Felsenstein [23]). Adicionalmente, dado que as relações entre os elementos de um determinado grupo de organismos são desconhecidas (exceto se produzidos em laboratório, *e.g.*, Hillis *et al.* [46]), as análises filogenéticas passaram a ser encaradas como procedimentos de *inferência* sobre essas relações. Inferências sobre eventos desconhecidos (neste caso sobre os possíveis padrões dos eventos, *i.e*, as árvores) são adequadamente analisados sob ótica probabilística (*e.g.*, Farris [19], Felsenstein [23], Harper [38]) e é sob este enfoque que o restante deste trabalho será feito. Dessa forma, inicialmente neste trabalho será apresentada uma visão teórica de dois "grupos de conceitos" importantes na filogenética em geral (grupos e caracteres), posteriormente será introduzido o uso de modelos estocásticos e, por último, serão apresentadas reinterpretações[3] desses conceitos sob ótica probabilística.

[3]Ou opiniões prévias.

1.4 Grupos monofiléticos vs. não-monofiléticos

A noção de grupo monofilético é fundamental na filogenética. Desde Hennig [42], sua importância não foi questionada. Entretanto, associados ao conceito de grupo monofilético, existem outros conceitos de agrupamentos, *i.e.*, grupo parafilético e polifilético. Esses conceitos geraram certa confusão de interpretação, desde suas definições por Hennig [43], pois ele próprio usou critérios diferentes para definir, de um lado, grupo monofilético (topológico-dependente, p. 98) e, de outro, grupo parafilético e polifilético (caráter-dependentes, p. 103 e 104). Esses conceitos foram claramente redefinidos apenas em 1971 (Nelson [72]).

Confusão adicional é notada quando se tenta associar a noção de grupo monofilético à de hierarquia e ambas à de classificação. Nas últimas 3 décadas, tem havido uma forte tendência (não tanto entre os botânicos) de que as classificações sejam obrigatoriamente embasadas em filogenia(s) e, conseqüentemente, que apenas grupos mofiléticos sejam (in)formalmente nomeados. Quanto aos grupos polifiléticos não há discordância, mas em relação aos grupos parafiléticos tem havido uma mixórdia improcedente na literatura botânica, liderada especialmente por Brummitt (*e.g.* [5] [6], [7] [8]) e Cavallier-Smith (*e.g.* [10]). Essa discordância culminou, recentemente, em um "abaixo-assinado" de 150 autores de várias partes do mundo contrários à associação filogenia-classificação (Nordal & Stedje [80]). Assim, faz-se necessária uma rápida discussão sobre o assunto, dado o grande potencial de confusão.

Em geral, esses autores argumentam que:

1) Dividir uma árvore evolutiva [cladograma] em famílias, gêneros e espécies mutuamente excludentes, os quais sejam estritamente monofiléticos, é uma impossi-

bilidade lógica;

2) O surgimento do pensamento cladístico nos últimos 40 anos tem promovido uma obsessão por táxons monofiléticos, com classificação baseada somente em descendência ao custo da modificação;

3) A classificação linneana é a ferramenta ótima para catalogar a biodiversidade e requer o reconhecimento de grupos parafiléticos.

Esses argumentos podem ser facilmente subjugados, pois (modificado de Dias *et al.* [14]):

1) Primeiramente, não existem grupos "estritamente monofiléticos". Um grupo é ou não é monofilético. Em segundo lugar, não há qualquer impossibilidade lógica em se assumir apenas grupos monofiléticos como grupos válidos em uma determinada classificação. Os autores confundem *lógica* com a possibilidade de um grande número de nomes que poderia resultar de um esquema de subordinação aplicado à classificação filogenética (veja Nelson [73]), caso o autor desse esquema nomeie e atribua uma categoria para *todos* os grupos monofiléticos. Novamente, os autores negligenciam que nem *todo* e *qualquer* grupo monofilético tem que ter um nome formal (Nixon *et al.* [79]), o que já é uma prática comum (*e.g.*, Eudicotiledôneas e Asterídeas).

2) É notável a grande confusão feita pelos autores, pois fica clara falta de intimidade deles com o progresso que a filogenética teve desde Hennig [44]. Com essa afirmação, os autores tentam passar a idéia de que os filogeneticistas consideram apenas cladogênese (padrões de ramificação), mas não anagênese (modificação ao longo dos ramos). Essa é uma descrição simplista e equivocada da abordagem filogenética

à classificação, a qual foi freqüentemente usada nos anos 70 por defensores de outras escolas da sistemática (*e.g.*, Mayr [68]) e não possui qualquer valor para a questão sob análise (a aceitação de grupos parafiléticos). Como algum grupo poderia ser reconhecido como tal se nenhuma "modificação" (ou mais adequadamente, diferença em estados de caráter) pudesse ser detectada?

3) A "classificação linneana" não é *linneana*, é *aristotélica,* e não requer, em hipótese alguma, grupos parafiléticos (nem monofiléticos ou polifiléticos). Além disso, o reconhecimento de grupos parafiléticos não melhora em nada uma classificação e, como tal, não é necessário. Adicionalmente, grupos parafiléticos não possuem conteúdo informativo, característica chave em qualquer classificação. Por exemplo, que informação pode ser extraída (ou recuperada) de um grupo parafilético (*e.g.*, Farris [21], Platnick [88])? Qual, se algum, estado de caráter definiria tal tipo de grupo (*e.g.*, Dicotiledôneas)? Colocando de outra forma, que tipo de informação sobre estados de caráter poderia ser extraído de "Dicotiledôneas" que defina exclusivamente este táxon (veja Farris [20] [21], para discussão detalhada sobre conteúdo informativo)? O grupo parafilético "Dicotiledôneas" não descreve a distribuição de qualquer característica, qualquer que seja, e, portanto, não fornece qualquer informação que já não esteja disponível a partir de um outro grupo mais inclusivo. Além de tudo isso, como já discutido por diversos autores (*e.g.*, Keller *et al.* [53], Nixon *et al.* [79], Wheeler [110]), a taxonomia linneana pode ser facilmente integrada a um sistema no qual nomes sejam dados apenas a grupos monofiléticos.

1.5 Homologia e caracteres

A definição de homologia está no cerne da biologia comparada e, conseqüentemente, foi e ainda é um dos conceitos mais discutidos em estudos comparados. Após Owen [81], uma das obras mais influentes sobre homologia foi o trabalho de Remane [93], o qual elencou três critérios para se afirmar se "elementos" sob comparação seriam homólogos ou não:

1) critério da posição (*Kriterium der Lage*), a qual pode ser: a) topográfica (*Topographische Lageähnlichkeit*), b) geométrica (*Lageähnlichkeit bei geometrisch ähnlichen Figuren oder Körpern*) ou c) relativa às outras partes do corpo (*Lageähnlichkeit im Gefüge*);

2) critério das funções especiais das estruturas (*Kriterium der speziellen Qualität der Strukturen*); e

3) critério da conexão através de formas intermediárias (*Kriterium der Verknüpfung durch Zwischenformen: Stetigkeitskriterium*).

Desses, apenas a posição topográfica (critério 1) pode ser considerada válida para ser usada como *conjectura inicial de homologia*, dado que os outros dois critérios demandam que estudos adicionais de função e de busca por intermediários, respectivamente, sejam realizados, o que inviabiliza sua utilização em termos práticos.

Os critérios de Remane [93] deram espaço para que diferentes interpretações fossem feitas, o que acabou levando a uma confusão generalizada na literatura sobre o assunto. Essa situação caótica foi claramente resolvida por Pinna [85], que reiterou e justificou o uso do termo de forma objetiva. Pinna [85] endossou a visão

de alguns autores anteriores (*e.g.*, Cracraft [11], Nelson [75], Patterson [84], Wiley [112]), igualando homologia à sinapomorfia, demonstrando a utilidade prática de se definir homologia em termos operacionais. Homologia agora não é mais vista como um conceito, mas como um processo constituído por duas etapas distintas e inter-relacionadas: 1) estabelecimento de conjecturas iniciais (homologia primária[4]) e 2) a detecção destas conjecturas em um cladograma (homologia secundária).

Apesar do trabalho de Pinna [85], às vezes, ainda ocorrem interpretações equivocadas na literatura, como é o caso de, *e.g.*, Scotland [101] e Williams & Humphries [115], os quais reiteram idéias que já foram adequadamente refutadas por Pinna [85].

1.5.1 Caráter e estado de caráter

Em situação bastante semelhante à homologia está o conceito de caráter (*e.g.*, Hawkins [40]; Henning [44]; Kitching *et al.* [54]; Pogue & Mickevich [89]; Wiley [113]). Aqui, caráter será assumido como qualquer característica para a qual haja uma proposição de homologia primária e para a qual a detecção de uma homologia secundária seja possível[5], *i.e.*, caráter é uma *proposição* de que duas ou mais características são comparáveis, apesar de poderem ser diferentes. Dessa forma, caráter é uma abstração do pesquisador sobre as características dos organismos em estudo (veja Nelson [77], para uma discussão mais detalhada) e que pode ser usada *a posteriori* para propor possíveis relações de parentesco entre os mesmos. Estados de caráter constituem, conseqüentemente, as próprias características em si, as quais têm que

[4]Veja Brower & Schawaroch [4], para uma visão ligeiramente diferente.
[5]Embora não necessariamente detectável no grau de generalidade de uma determinada análise.

atender aos princípios de uma análise filogenética[6] (veja Grandcolas *et al.* [34]).

Levando isso em conta, vários caracteres usados na literatura não são adequados para uma análise filogenética. Por exemplo, é extremamente indefensável propor um caráter "distribuição geográfica" com os estados "pantropical", "América do Sul" e "America do Norte-Ásia", como feito por Pansarin ([83], p. 35).

Como pode ser notado, para o exemplo citado acima, é impossível fazer proposições de homologia primária; usar a distribuição geográfica conhecida dos organismos como caráter para inferir suas possíveis relações de parentesco, é tão absurdo quanto usar sua data de coleta (Grandcolas *et al.* [34]), ou, quem sabe, o nome de quem digitou os dados da planta.

1.5.2 Homologia em (e utilidade de) dados morfológicos e moleculares

Como já discutido anteriormente, homologia é uma proposição de relação entre características intrínsecas de organismos diferentes e divide-se em duas etapas: homologia primária e secundária (Pinna [85]). No caso de dados morfológicos, alguns autores têm argumentado que a proposição de homologia primária seria uma etapa extremamente subjetiva e que, dado um conjunto de organismos, diferentes autores poderiam "perceber" e codificar seus caracteres de maneira diferente (Hawkins [40]). Portanto, a proposição de homologia primária, constituiria "mais um fator de ambigüidade na análise filogenética" (Scotland *et al.* [102], p. 541 e Figura 1c, p. 540). Dessa forma, esses autores defendem que os caracteres morfológicos deveriam

[6] A "descendência com modificação" defendida pelos autores não está sendo endossada aqui.

ser apenas mapeados em filogenias inferidas a partir de dados moleculares e que somente estes seriam adequados para inferir as filogenias em si.

Entretanto, esses argumentos não se sustentam nem do ponto de vista operacional nem do prático. Operacionalmente, a proposição de homologia primária em dados moleculares pode ser tão (ou mais) problemática quanto nos morfológicos. Em dados moleculares, o espaço definido pelos estados é extremamente reduzido (4 ou 5[7]) se comparado ao possível espaço de estados em caracteres morfológicos. Esse espaço limitado de estados tem levado a uma falsa impressão de que a definição dos caracteres em dados moleculares é direta e objetiva (Jenner [51]), mas ao se analisar o espaço definido pelo número possível de alinhamentos (= proposição de homologia primária), será constatado que esta é uma tarefa matematicamente mais complicada do que a própria inferência das árvores filogenéticas em si (*e.g.*, Bonizzoni & Vedova [1], Felsenstein [30]).

Por outro lado, do ponto de vista prático, a utilidade de caracteres moleculares em grupos angiospérmicos ainda é nula. Sinapomorfias moleculares são inúteis no cotidiano e os táxons precisam ser definidos através de sinapomorfias morfológicas ou pelo menos caracterizados morfologicamente[8], senão não são passíveis de reconhecimento em atividade prática.

Adicionalmente, se os caracteres morfológicos podem ser mapeados em uma filogenia, então também são adequados para serem incluídos na própria matriz usada para inferir essa filogenia (*e.g.*, Jenner [51], Ronquist [96], Wiens [111]).

Ainda que com grande potencial, mesmo a técnica de código de barras de

[7]No caso de *gap* ser considerado um quinto estado

[8]Não implica que uma característica morfológica diagnóstica não incluída na análise filogenética seja chamada de sinapomorfia, como tem acontecido na literatura recente (*e.g.*, Ranker *et al.* [91]).

DNA (*e.g.*, Schindel & Miller [99]) apresenta várias limitações do ponto de vista prático e operacional e é dependente da existência de taxonomia morfológica prévia (*e.g.*, DeSalle *et al.* [13]), o que impõe que sua implementação cotidiana ainda esteja longe de ser viável na taxonomia da grande maioria dos grupos (*e.g.*, Ebach & Holdrege [15], Moritz & Cicero [70], Meyer & Paulay [69], Will *et al.* [114]).

Apesar da possível utilidade da morfologia, na literatura botânica recente tem havido uma tendência em desconsiderá-la. Mas, o mais grave talvez seja desconsiderar os caracteres morfológicos nas análises filogenéticas e depois mapeá-los em filogenias moleculares e chamá-los de "sinapomorfias" ou "homoplasias" (*e.g.*, Ranker *et al.* [91]). Isso mostra que ou o conceito de sinapomorfia não foi entendido por parte dos botânicos, ou foi redefinido sem ser publicado (ou ambos).

1.5.3 Modelos estocásticos

Considerando a história da sistemática após Hennig ([42] [43], [44]), o uso inicial de modelos estocásticos pode ser referido a Cavalli-Sforza & Edwards [9]. Posteriormente, Felsenstein (*e.g.*, [23] [24] [26]) foi um dos principais defensores do uso de "modelos evolutivos" (= modelos de substituição) na estimação de filogenias.

Atualmente, o uso de modelos está bastante difundido (*e.g.*, Posada & Crandall [90], Rodríguez *et al.* [94], Swofford *et al.* [107]), especialmente devido ao uso comum de máxima verossimilhança (*e.g.*, Felsenstein [23], [24] [26] [27] [28], [29] [30]; Swofford *et al.* [108]) e análise bayesiana (*e.g.*, Larget & Simon [60], Mau [65], Mau & Newton [66], Mau *et al.*, [67], Rannala & Yang [92], Ronquist & Huelsenbeck [97], Yang & Rannala [117]).

Modelos estocásticos desempenham um papel fundamental na filogenética atual (*e.g.*, Steel & Penny [104]) e são usados para estabelecer probabilidades de transição entre os estados de um determinado caráter (veja Kläre [56] para uma análise matemática). Usualmente, são representados na forma de uma matriz de substituição (Equação 1.13). Nessa matriz, os elementos de fora da diagonal representam a probabilidade de um estado ser substituído por um estado diferente (P_{ij}) após um dado intervalo (mínimo) de tempo t, enquanto os elementos da diagonal representam a probabilidade de um estado não ser substituído (P_{ii}) após um determinado intervalo (mínimo) de tempo (veja o item 1.5.3.1, abaixo, para detalhamento).

Adicionalmente, todos os modelos em filogenética representam um tipo particular de processo estocástico, a cadeia de Markov[9]. De forma simplificada, uma cadeia de Markov é uma seqüência de variáveis aleatórias $X = (X_0, X_1, X_2, \ldots, X_n)$ com a propriedade de que a probabilidade de estar em um determinado X_i no tempo t depende apenas de um número k de estados anteriores, sendo k a ordem da cadeia.

Esse tipo de cadeia está no cerne de qualquer abordagem probabilística à filogenética (*e.g.*, máxima verossimilhança, análise bayesiana, quebra-cabeça de quartetos) e é caracterizada por assumir que a probabilidade do estado atual depende apenas do estado imediatamente anterior (para uma visão geral dos diferentes tipos de cadeia e algumas aplicações, veja os trabalhos de, *e.g.*, Gilks *et al.* [33], Guttorp [35] e Häggström [36]).

[9] A não ser que expressamente declarado, será assumida uma cadeia de primeira ordem e homogênea.

1.5.3.1 Construção de modelo[10]

Para um determinado caráter binário $X = \{i, j\}$, $i \neq j$, que muda de forma idêntica e independentemente distribuída (i.i.d.[11]), é necessário atribuir probabilidades (P) para seus respectivos estados em um determinado intervalo de tempo t. Dessa forma, se considerarmos que $X = i$ no tempo t_0

$$P(X_{i_{t0}}) = 1 \tag{1.1}$$

logo,

$$P(X_{j_{t0}}) = 0 \tag{1.2}$$

então, precisamos saber duas probabilidades condicionais

$$P(X_{i_{t1}}|X_{i_{t0}}) \tag{1.3}$$

$$P(X_{j_{t1}}|X_{i_{t0}}) \tag{1.4}$$

Seja α a taxa de mudança[12] de i para j e vice-versa

$$P(X_{j_{t1}}|X_{i_{t0}}) = \alpha \tag{1.5}$$

então,

$$P(X_{i_{t1}}|X_{i_{t0}}) = 1 - \alpha \tag{1.6}$$

[10]Modificado de Li ([62]).
[11]Embora isso não seja necessário, facilita esta discussão inicial.
[12]Mudança na cadeia = iteração.

Dessa forma, em t_2 temos as seguintes situações

$$P(X_{i_{t2}}|X_{i_{t1}}) = P(X_{j_{t2}}|X_{j_{t1}}) \tag{1.7}$$

$$P(X_{i_{t2}}|X_{j_{t1}}) = P(X_{j_{t2}}|X_{i_{t1}}) \tag{1.8}$$

que levam a

$$\begin{aligned} P(X_{i_{t2}}) &= (1-\alpha)P(X_{i_{t1}}) + (\alpha - \alpha P(X_{i_{t1}})) \\ P(X_{i_{t2}}) &= (1-\alpha)P(X_{i_{t1}}) + \alpha(1 - P(X_{i_{t1}})) \\ P(X_{i_{t2}}) &= P(X_{i_{t1}}) - \alpha P(X_{i_{t1}}) + \alpha - \alpha P(X_{i_{t1}})) \\ P(X_{i_{t2}}) - P(X_{i_{t1}}) &= -2\alpha P(X_{i_{t1}}) + \alpha \end{aligned} \tag{1.9}$$

Considerando o tempo como contínuo,

$$\begin{aligned} \Delta P(X_{i_t}) &= -2\alpha P(X_{i_{t1}}) + \alpha \\ \frac{dP(X_{i_t})}{dt} &= -2\alpha P(X_{i_{t1}}) + \alpha \\ P(X_{i_t}) &= \frac{1}{2} + \left(P(X_{i_{t0}}) - \frac{1}{2}\right) e^{(-2\alpha t)} \end{aligned} \tag{1.10}$$

Se $P(X_{i_{t0}}) = 1$, como assumido anteriormente (veja a Equação 1.1), então

$$\begin{aligned} P(X_{i_{t1}}|X_{i_{t0}}) &= \frac{1}{2} + \left(1 - \frac{1}{2}\right) e^{(-2\alpha t)} \\ &= \frac{1}{2} + \frac{1}{2} e^{(-2\alpha t)} \end{aligned} \tag{1.11}$$

Por outro lado, se $P(X_{i_{t0}}) = 0$,

$$\begin{aligned} P(X_{i_{t1}}|X_{j_{t0}}) &= \frac{1}{2} + \left(0 - \frac{1}{2}\right) e^{(-2\alpha t)} \\ &= \frac{1}{2} - \frac{1}{2} e^{(-2\alpha t)} \end{aligned} \tag{1.12}$$

Então, a probabilidade de mudança entre estados $P(X_{ij})$ em um determinado intervalo de tempo t é dada pela Equação 1.12 e a probabilidade de não haver mudança é dada pela Equação 1.11.

O modelo aqui exemplificado é equivalente ao JC69 (Jukes & Cantor [52]).

1.5.3.2 Modelo e comprimentos de ramos[13]

Comprimentos de ramos representam o número esperado de mudanças num determinado intervalo de tempo. Apesar de claramente definido, é comum confundir comprimento de ramo com o tempo em si, como já destacado por vários autores (*e.g.*, Felsenstein [23], [30]). Dado que uma discussão abrangente sobre comprimentos de ramos foge ao escopo deste estudo, será apenas demonstrado como usar a informação obtida nas Equações 1.11 e 1.12, dado um intervalo de tempo arbitrário (para a estimação de comprimentos de ramos usando o método de mínimos-quadrados sugere-se o trabalho de Fitch & Margoliash ([31]) e usando máxima verossimilhança sugere-se o trabalho de Rogers & Swofford ([95])).

Sendo a matriz de transição entre os estados definida por

[13]Em análises que fazem uso explícito de modelos como o mostrado na seção anterior (item 1.5.3.1).

$$\varphi = \begin{bmatrix} ii_{t_n} & ij_{t_n} \\ ij_{t_n} & ii_{t_n} \end{bmatrix} \tag{1.13}$$

Substituindo as Equações 1.11 e 1.12 na 1.13 temos

$$\varphi = \begin{bmatrix} \frac{1}{2} + \frac{1}{2}e^{(-2t)} & \frac{1}{2} - \frac{1}{2}e^{(-2t)} \\ \frac{1}{2} - \frac{1}{2}e^{(-2t)} & \frac{1}{2} + \frac{1}{2}e^{(-2t)} \end{bmatrix} \tag{1.14}$$

Supondo que o tempo decorrido t desde o último evento cladogenético tenha sido de 0,5 unidade de tempo, a probabilidade de mudança de estado de caráter será

$$ij_{0,5} = 0.3160602794 \tag{1.15}$$

logo,

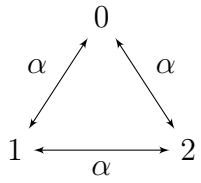
$$ii_{0,5} = 0.6839397206 \tag{1.16}$$

Substituindo as Equações 1.15 e 1.16 na 1.14, temos que o modelo de substituição para o caráter binário $X = \{i, j\}$, $i \neq j$, é representado por

$$\varphi = \begin{bmatrix} 0.6839397206 & 0.3160602794 \\ 0.3160602794 & 0.6839397206 \end{bmatrix} \tag{1.17}$$

Caso se alterasse o tempo para $t = 1$ e $t = 1,5$, teríamos, respectivamente

$$\varphi = \begin{bmatrix} 0.5676676416 & 0.4323323584 \\ 0.4323323584 & 0.5676676416 \end{bmatrix} \tag{1.18}$$



(a)

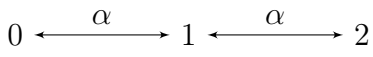
$$\varphi = \begin{bmatrix} 1-2\alpha & \alpha & \alpha \\ \alpha & 1-2\alpha & \alpha \\ \alpha & \alpha & 1-2\alpha \end{bmatrix}$$

(b)

Figura 1.1 Modelo para um caráter multi-estado com 3 estados e não-ordenado (α representa a probabilidade de substituição). (a) Representação esquemática de relações entre os estados. (b) Matriz de transição entre os estados.

$$\varphi = \begin{bmatrix} 0.5248935342 & 0.4751064658 \\ 0.4751064658 & 0.5248935342 \end{bmatrix} \tag{1.19}$$

Como mostram as Figuras 1.1 e 1.2, esses modelos podem ser aplicados tanto a caracteres não-ordenados como ordenados. Mesmo caracteres mais complexos, como o mostrado na Figura 1.3, podem tratados adequadamente com o uso de modelos.



(a)

$$\varphi = \begin{bmatrix} 1-\alpha & \alpha & 0 \\ \alpha & 1-\alpha & \alpha \\ 0 & \alpha & 1-\alpha \end{bmatrix}$$

(b)

Figura 1.2 Modelo para um caráter multi-estado com 3 estados e ordenado (α representa a probabilidade de substituição). (a) Representação esquemática de relações entre os estados. (b) Matriz de transição entre os estados.

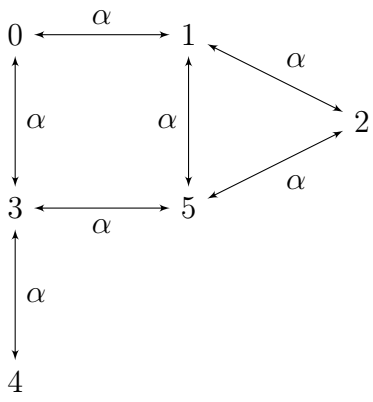


Figura 1.3 Representação esquemática de relações entre os estados para um caráter multi-estado parcialmente ordenado (α representa a probabilidade de substituição).

1.5.3.3 Dados moleculares[14]

Um dos trabalhos iniciais mais influentes que fez uso de modelos em análises de dados moleculares foi o de Jukes & Cantor [52], embora ainda não em contexto estritamente filogenético.

Atualmente, a utilização de modelos é bastante difundida e os que estão em uso variam desde os uniparamétricos, *e.g.*, o JC69 (Jukes & Cantor [52]), passando por modelos mais sofisticados, como o K2P (Kimura [109]), o F84 (Felsenstein [27]), o HKY85 (Hasegawa *et al.* [39]); até os mais complexos, como o GTR (Lanave *et al.* [59], Rodríguez *et al.* [94]). Todos esses modelos são hierarquizados, o que significa que é possível se chegar a um determinado modelo a partir de modificações feitas em um outro. Mais recentemente, esses modelos tiveram que ser modificados para que se tornassem biologicamente mais factíveis (e conseqüentemente mais complexos) e outros fatores (=parâmetros) tiveram que ser levados em conta, tais como a heterogeneidade de taxas entre sítios e a quantidade de sítios invariáveis (*e.g.*, Yang [116]), o que impulsionou fortemente seu próprio desenvolvimento e diversidade atual.

1.5.3.4 Dados morfológicos

Por outro lado, a utilização de modelos markovianos em análises de dados morfológicos tem tido relativamente menos atenção na literatura, apesar da abordagem não ser nova. Vários autores já propuseram seu uso (*e.g.*, Farris [19], Felsenstein [23], Losos [63], Martins [64], Neyman [78], Pagel [82], Schultz & Churchil [100]) e um modelo equivalente ao JC69 (Jukes & Cantor [52]) já havia sido usado por Haldane [37] em sua função de distância.

[14] A não ser que expressamente declarado, a referência será sempre a seqüências nucleotídicas.

Entretanto, apesar de existirem idéias anteriores, o uso de caracteres morfológicos discretos em análises probabilísticas teve como maior entrave à sua exploração o fato de não haver implementação computacional disponível. Por exemplo, até hoje, o PAUP* (Swofford [106]) só permite que se use máxima verossimilhança com dados moleculares (a não ser que se mascare os dados morfológicos com o uso de macros para enganar o programa, *e.g.*, Lewis [61]).

Talvez por essa dificuldade haja um ceticismo exacerbado de boa parte dos usuários, os quais, embora muitas vezes utilizem modelos em análises de dados moleculares, apresentam resistência quanto ao seu uso com dados morfológicos.

Esse ceticismo exagerado pode ser questionado tanto do ponto de vista biológico como do operacional. Se modelos são adequados para dados moleculares, porque não o seriam para dados morfológicos? Se é biologicamente admissível que, *e.g.*, transversões e transições podem ocorrer associadas a probabilidades diferentes, então também é biologicamente admissível que estados de caracteres morfológicos, *e.g.*, foliares *vs.* florais, podem mudar diferencialmente e, conseqüentemente, associados a probabilidades diferentes. Dessa forma, é biologicamente factível admitir que dados foliares possam mudar de forma diferenciada (*i.e.*, sob taxas diferentes) dos florais, sendo operacionalmente irrelevante qual apresenta maior ou menor taxa de mudança.

1.5.3.5 Pressuposições e uso de modelos

É comum ouvir de usuários adversos ao (ou menos avisados sobre o) uso de modelos e/ou análises probabilísticas que "não usam modelo algum" ou que "cadeia de Markov é coisa de matemático, não de biólogo". Essas duas afirmativas são infe-

lizes, pois demonstram a falta de intimidade que esses usuários têm com os próprios métodos que usam. Por exemplo, quando se usa parcimônia padrão como critério de otimização, implicitamente é assumido[15] *a priori* que:

1) as taxas de mudança são baixas e idênticas para todos os caracteres - o que não faz sentido ontológico nem biológico, dada a diversidade de sinal filogenético que os caracteres carreiam, *e.g.*, caracteres foliares *vs.* florais, regiões codificantes *vs.* não-codificantes;

2) no caso específico de dados moleculares, não existe viés entre transversões e transições - o que não tem suporte empírico; ou ainda

3) a chance de uma adenina parear com uma timina é igual à de parear com uma guanina – o que também não tem sustentação biológica, dado que a primeira e a segunda se unem por duas pontes de hidrogênio e a terceira possui três, tornando este último pareamento menos provável que o primeiro.

Ao sustentar que não usa modelo ou cadeia de Markov, o usuário de programas para análises filogenéticas desconhece o que, *e.g.*, o PAUP* (Swofford [106]) faz com os dados que são fornecidos a ele quando se faz uma análise de distância (*e.g.*, Neighbour-Joining, Saitou & Nei [98]) ou de máxima verossimilhança (*e.g.*, Felsenstein [23], [26]). Nessas análises, *sempre* é necessário o uso de uma cadeia de Markov (modelo de substituição, veja o item 1.5.3), a qual o PAUP* usa, querendo o usuário ou não. E se esse modelo for escolhido de forma equivocada ou, como é mais comum, seja usado o padrão do PAUP*, os resultados obtidos podem carrear nenhum valor, que de outra forma teriam (*e.g.*, Sullivan & Swofford [105], Swofford *et al.* [108]). Isso significa que, querendo ou não, sempre se usa modelo, seja ele

[15]Mesmo que por omissão.

simples ou complexo, explícito ou implícito.

1.5.4 Sinapomorfia enquanto probabilidade

Ao se assumir o uso de modelos markovianos, *i.e.*, a associação de probabilidades às mudanças entre estados de caráter, é possível vislumbrar uma forma diferenciada de se ver homologia. Se a substituição de um determinado estado de caráter por outro possui uma probabilidade associada (veja o item 1.5.3.1, acima), então o seu compartilhamento também possui uma probabilidade associada desconhecida. Essa probabilidade, que é condicional, pode ser adequadamente analisada sob o ponto de vista probabilístico (freqüentista ou bayesiano). Assim, a homologia pode ser *tratada* (não confundir com *quantificada*) como probabilidade (*e.g.*, Harper [38]) e, por conseqüência, também a monofilia.

Dada a confusão que essa afirmativa pode gerar, é genuíno ressaltar que não se está dizendo que homologia pode ser "quantificada", nem que é necessário construir um intervalo de confiança (no caso freqüentista) ou um intervalo de credibilidade (no caso bayesiano) para acessar o "grau de homologia", mas sim que é possível atribuir um valor de probabilidade ao caso de um determinado estado compartilhado representar uma homologia secundária ou não, dada(s) a(s) árvore(s) analisada(s) e os outros parâmetros de interesse. Dessa forma, o que se faz é estimar a probabilidade do evento ter acontecido de uma determinada maneira (padrão), levando-se em conta todas as outras possíveis maneiras analizadas (*i.e.*, a incerteza filogenética), não a probabilidade do evento em si ter acontecido (esta será sempre igual a 1, pois o evento já ocorreu). Portanto, a visão de que as "abordagens probabilísticas atribuem valores

a eventos pretéritos, os quais se sabe de antemão que aconteceram" (Helfenbein & DeSalle [41]) é equívoca.

O valor desse tipo de abordagem (sinapomorfia enquanto probabilidade) fica ainda mais patente quando se tem dados ausentes na matriz e/ou quando existe uma otimização ambígua em análise por parcimônia padrão, como por exemplo o caso mostrado na Figura 1.4.

1.5.5 Evolução de caracteres

Foi assumido anteriormente (veja o item 1.5.1, acima) que caráter é uma proposição e, como tal, *não* evolui (proposições não evoluem) (*e.g.*, Nelson [77]). Essa afirmativa, embora possa parecer extremamente simples (e consonante com o que foi escrito acima), já representou um problema considerável na história da sistemática (*e.g.*, Hull [50]) e foi até mesmo usada deturpadamente por criacionistas como suporte à sua tese (*e.g.*, http://emporium.turnpike.net/C/cs/evid4.htm [12]).

Quando se afirma que os caracteres, assim como os táxons, não podem ser considerados plesiomórficos/ancestrais ou apomórficos/descendentes (*e.g.*, Nelson [77]), não se está negando que a evolução ocorreu (e ocorre), mas dado que o caráter é uma contextualização/proposição, não há sentido ontológico em se dizer que ele em si evolui. Dessa forma, será assumido que o processo anagenético pode ser descrito como uma alternância entre os estados de caráter ao longo do tempo, o que pode ser adequadamente incorporado em um modelo, no qual sejam assumidas explicitamente as probabilidades associadas a cada tipo de alternância (veja o item 1.5.3, acima), mesmo que essas probabilidades sejam *a priori* e possuam distri-

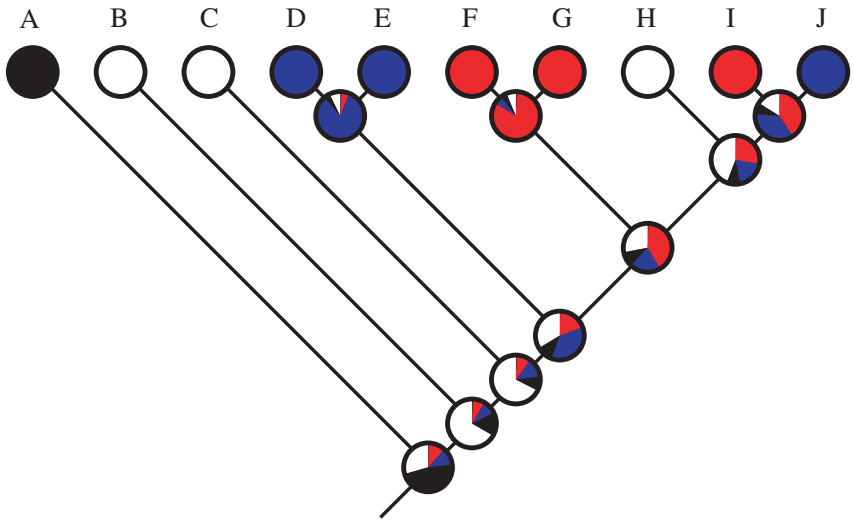


Figura 1.4 Otimização de caracteres usando máxima verossimilhança. Cores representam estados diferentes e probabilidades dos respectivos estados nos nós (A, B, e C representam grupos-externos).

buições não-informativas ou uniformes (veja, *e.g.*, Box & Tiao [2], Gelman *et al.* [32], para uma discussão mais detalhada sobre distribuições *a priori* não-informativas e distribuições uniformes).

1.6 Conclusões

Grupos e caracteres constituem dois dos principais conjuntos de conceitos em filogenética. Entretanto, apesar da existência de vasta literatura sobre o assunto, ainda é comum o uso equivocado desses termos, especialmente na literatura botânica. Adicionalmente, o tratamento de caracteres (especialmente morfológicos) com o uso de modelos estocásticos, apesar de sua adequação, ainda é uma área pouco explorada em filogenética. Nesse sentido, aqui foi apresentada uma rápida revisão

sobre esses conceitos, sua possível utilização com base em modelos explícitos, e vantagens associadas ao seu uso em filogenética. Dado que o conceito de caráter está estreitamente relacionado ao de grupo, o uso de modelos, conseqüentemente, leva a uma reinterpretação do conceito de sinapomorfia e, por sua vez, de monofilia.

1.7 Referências

[1] Bonizzoni, P. & Vedova, G. D. 2001. The complexity of multiple sequence alignment with SP-score that is a metric. *Theoret. Comp. Science* 259: 63–79.

[2] Box, G. E. P. & Tiao, G. C. 1992. *Bayesian inference in statistical analysis.* Wiley Classics Library, New York.

[3] Brower, A. V. Z. 2000. Evolution is not a necessary assumption of cladistics. *Cladistics* 16: 143–154.

[4] Brower, A. V. Z. & Schawaroch, V. 1996. Three steps of homology assessment. *Cladistics* 12: 265–272.

[5] Brummitt, R. K. 1997. Taxonomy versus cladonomy, a fundamental controversy in biological systematics. *Taxon* 46: 723–734.

[6] Brummitt, R. K. 2002. How to chop up a tree. *Taxon* 51: 31–41.

[7] Brummitt, R. K. 2003. Further dogged defense of paraphyletic taxa. *Taxon* 52: 803–804.

[8] Brummitt, R. K. 2006. Am I a bony fish? *Taxon* 55: 268–269.

[9] Cavalli-Sfforza, L. L. & Edwards, A. W. F. 1967. Phylogenetic analysis: models and estimation procedures. *Evolution* 21: 550–570.

[10] Cavallier-Smith, T. 1998. A revised sex-kingdom system of life. *Biol. Rev.* 73: 203–266.

[11] Cracraft, J. 1978. Science, philosophy and systematics. *Syst. Zool.* 27: 213–216.

[12] Creation Science. Disponível em http://emporium.turnpike.net/C/cs/evid4.htm. Acesso em: 27 maio 2007.

[13] DeSalle, R., Egan, M. G. & Siddall, M. 2005. The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360: 1905–1916.

[14] Dias, P., Assis, L. C. S. & Udulutsch, R. G. 2005. Monophyly vs. paraphyly in plant systematics. *Taxon* 54: 1039–1040.

[15] Ebach, M. C. & Holdrege, C. 2005. DNA barcoding is no substitute for taxonomy. *Nature* 437: 697.

[16] Edwards, A. W. F. & Cavalli-Sforza, L. L. 1964. Reconstruction of evolutionary trees. *In* Heywood, V. & McNeill, J. (eds.) *Phenetic and phylogenetic classification.* Systematics Association Publ. n. 6, London, 67–76.

[17] Farris, J. S. 1969. A successive approximations approach to character weighting. *Syst. Zool.* 18: 374–385.

[18] Farris, J. S. 1970. Methods for computing Wagner trees. *Syst. Zool.* 19: 83–92.

[19] Farris, J. S. 1973. On the use of the parsimony criterion for inferring evolutionary trees. *Syst. Zool.* 22: 250–256.

[20] Farris, J. S. 1977. On the phenetic approach to vertebrate classification. *In* Hecht, M., Goody, P. & Hecht, B. (eds.) *Major patterns in vertebrate evolution.* NATO Advanced Study Institute Series, no. 14. Plenum Press, New York, 823–850.

[21] Farris, J. S. 1979. The information content of the phylogenetic system. *Syst. Zool.* 28: 483–519.

[22] Farris, J. S., Kluge, A. G. & Eckardt, J. 1970. A numerical approach to phylogenetic systematics. *Syst. Zool.* 19: 172–191.

[23] Felsenstein, J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst. Zool.* 22: 240–249.

[24] Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. *Syst. Zool.* 27: 401–410.

[25] Felsenstein, J. 1978. The number of evolutionary trees. *Syst. Zool.* 27: 27–33.

[26] Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17: 368–376.

[27] Felsenstein, J. 1984. Distance methods for inferring phylogenies: a justification. *Evolution* 38: 16–24.

[28] Felsenstein, J. 1988. Phylogenies and quantitative characters. *Annual Rev. Ecol. Syst.* 19: 445–471.

[29] Felsenstein, J. 2001. The troubled growth of statistical phylogenetics. *Syst. Biol.* 50: 465–467.

[30] Felsenstein, J. 2004. *Inferring phylogenies.* Sinauer Associates, Sunderland.

[31] Fitch, W. M. & Margoliash, E. 1967. Construction of phylogenetic trees: a method based on mutation distances as estimated from cytochrome *c* sequences is of general applicability. *Science* 155: 279–284.

[32] Gelman, A., Carlin, J., Stern, H. & Rubin, D. 2003. *Bayesian data analysis.* Chapman and Hall, London.

[33] Gilks, W., Richardson, S. & Speigelhalter, D. 1996. *Markov chain Monte Carlo in practice.* Chapman & Hall, New York.

[34] Grandcolas, P., Deleporte, P., Desutter-Grandcolas, L. & Daugeron, C. 2001. Phylogenetics and ecology: as many characters as possible should be included in the cladistic analysis? *Cladistics* 17: 104–110.

[35] Guttorp, P. 1995. *Stochastic modelling of scientific data.* Chapman and Hall, London.

[36] Häggström, O. 2002. *Finite Markov chains and algorithmic applications.* Cambridge University Press, Cambridge.

[37] Haldane, J. B. S. 1919. The combination of linkage values and the calculation of distances between the loci of linked factors. *J. Genet.* 299-309: 8.

[38] Harper, C. W. 1979. A Bayesian probability view of phylogenetic systematics. *Syst. Zool.* 28: 547–553.

[39] Hasegawa, M., Kishino, H. & Yano, T. 1985. Dating the human-ape splitting a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22: 160–174.

[40] Hawkins, J. A. 2000. A survey of primary homology assessment: different botanists perceive and define characters in different ways. *In* Scotland, R. & Pennington, T. (eds.) *Homology and systematics.* The Systematics Association/Taylor and Francis, London, 22–53.

[41] Helfenbein, K. G. & DeSalle, R. 2005. Falsifications and corroborations: Karl Popper's influence on systematics. *Mol. Phylogen. Evol.* 35: 271–280.

[42] Hennig, W. 1950. *Grundzüge einer Theorie der phylogenetischen Systematik.* Deutscher Zentralverlag, Berlin.

[43] Hennig, W. 1965. Phylogenetic systematics. *Ann. Rev. Ent.* 10: 97–116.

[44] Hennig, W. 1966. *Phylogenetic systematics.* Translated by D. D. Davis & R. Zangerl. University of Illinois Press, Urbana.

[45] Heywood, V., Michener, C., Moss, W., Sokal, R. R., Koopman, C., Lenington, S., Farris, J. S., Griffiths, G. C., Crancraft, J., Nelson, G. J. & Johnson, L. 1973. Discussion of symposium papers on contemporary systematic philosophies. *Syst. Zool.* 22: 393–400.

[46] Hillis, D. M., Bull, J., White, M. E., Badgett, M. R. & Molineux, I. J. 1993. Experimental approaches to phylogenetic analysis. *Syst. Biol.* 42: 90–92.

[47] Hull, D. L. 1979. The limits of cladism. *Syst. Zool.* 28: 416–440.

[48] Hull, D. L. 1979. Philosophical issues in systematics: introduction. *Syst. Zool.* 28: 520.

[49] Hull, D. L. 1983. Karl Popper and Plato's Metaphor. *In* Platnick, N. & Funk, V. (eds.) *Advances in cladistics.* vol. 2, Columbia University Press, New York, 177–189.

[50] Hull, D. L. 1988. *Science as process: an evolutionary account of the social and conceptual development of science.* University of Chicago Press, Chicago.

[51] Jenner, R. A. 2004. Accepting partnership by submission? Morphological phylogenetics in a molecular millennium. *Syst. Biol.* 53: 333–342.

[52] Jukes, T. H. & Cantor, C. R. 1969. Evolution of protein molecules. *In* Munro, M. (ed.) *Mammalian protein metabolism.* vol. III, Academic Press, New York, 21–132.

[53] Keller, R. A., Boyd, R. N. & Wheeler, Q. D. 2003. The illogical basis of phylogenetic nomenclature. *Bot. Rev.* 69: 93–110.

[54] Kitching, I., Williams, D., Forey, P. L. & Humphries, C. J. 1998. *Cladistics: the theory and practice of parsimony analysis.* The Systematics Association, Oxford.

[55] Kitts, D. B. 1977. Karl Popper, verifiability, and systematic zoology. *Syst. Zool.* 26: 185–194.

[56] Kläre, S. 2005. *Stochastic models of molecular evolution: an algebraical and*

statistical analysis. Ph.D. dissertation, Ludwig-Maximilians-Universität, München.

[57] Kluge, A. G. 2001. Parsimony with and without scientific justification. *Cladistics* 17: 199–210.

[58] Kluge, A. G. & Farris, J. S. 1969. Quantitative phyletics and the evolution of anurans. *Syst. Zool.* 18: 1–32.

[59] Lanave, C., Preparata, G., Saccone, C. & Serio, G. 1984. A new method for calculating evolutionary substitution rates. *J. Mol. Evol.* 20: 86–93.

[60] Larget, B. & Simon, D. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* 16: 750–759.

[61] Lewis, P. O. 2001. A likelihood approach to estimating phylogeny from discrete morphological characters data. *Syst. Biol.* 50: 913–925.

[62] Li, W.-H. 1997. *Molecular evolution.* Sinauer Associates, Sunderland.

[63] Losos, J. B. 1994. An approach to the analysis of comparative data when a phylogeny is unavailable or incomplete. *Syst. Biol.* 47: 117–123.

[64] Martins, E. P. 1996. Conducting phylogenetic comparative studies when the phylogeny is not known. *Evolution* 50: 12–22.

[65] Mau, B. 1996. *Bayesian phylogenetic inference via Markov chain Monte Carlo methods.* Ph.D. dissertation, University of Wisconsin, Madison.

[66] Mau, B. & Newton, M. 1997. Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. *J. Comput. Graph. Stat.* 6: 122–131.

[67] Mau, B., Newton, M. & Larget, B. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics* 55: 1–12.

[68] Mayr, E. 1974. Cladistic analysis or cladistic classification? *Z. f. zool. Systematik u. Evolutionsforsch.* 12: 94–128.

[69] Meyer, C. P. & Paulay, G. 2005. DNA barcoding: error rates based on comprehensive sampling. *PLoS Biol.* 3: 2229–2238.

[70] Moritz, C. & Cicero, C. 2004. DNA barcoding: promise and pitfalls. *PLoS Biol.* 2: 1529–1531.

[71] Nelson, G. J. 1971. "Cladism" as a philosophy of classification. *Syst. Zool.* 20: 373–376.

[72] Nelson, G. J. 1971. Paraphyly and polyphyly: redefinitions. *Syst. Zool.* 20: 471–472.

[73] Nelson, G. J. 1973. Classification as an expression of phylogenetic relationships. *Syst. Zool.* 22: 344–359.

[74] Nelson, G. J. 1973. "Monophyly again?": a reply to P.D. Ashlock. *Syst. Zool.* 22: 310–312.

[75] Nelson, G. J. 1978. Ontogeny, phylogeny, paleontology, and the biogenetic law. *Syst. Zool.* 27: 324–345.

[76] Nelson, G. J. 1979. Cladistic analysis and synthesis: principles and definitions, with a historical note on Adanson's Familles des plantes (1763-1764). *Syst. Zool.* 28: 1–21.

[77] Nelson, G. J. 1994. Homology and systematics. *In* Hall, B. (ed.) *Homology: the hierarchical basis of comparative biology.* Academic Press, San Diego, 101–149.

[78] Neyman, J. 1971. Molecular studies of evolution: a source of novel statistical problems. *In* Gupta, S. & Yackel, J. (eds.) *Statistical decision theory and related topics.* Academic Press, New York, 1–27.

[79] Nixon, K. C., Carpenter, J. M. & Stevenson, D. W. 2003. The PhyloCode is fatally flawed, and the "Linnaean" system can easily be fixed. *Bot. Rev.* 69: 111–120.

[80] Nordal, I. & Stedje, B. 2005. Paraphyletic taxa should be accepted. *Taxon* 54: 5–6.

[81] Owen, R. 1843. *Lectures on comparative anatomy.* Longman, Brown, Green, and Longmans, London.

[82] Pagel, M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R. Soc. Lond. Ser. B* 255: 37–45.

[83] Pansarin, E. R. 2005. *Sistemática filogenética e biologia floral de Pogoniinae sul-americanas, e revisão taxonômica e análise das ceras epicuticulares do*

gênero ***Cleistes*** *Rich. ex Lindl. (Orchidaceae).* Tese de doutorado, Universidade Estadual de Campinas, Campinas.

[84] Patterson, C. 1982. Morphological characters and homology. *In* Joysey, K. A. & Friday, A. E. (eds.) *Problems of phylogenetic reconstruction.* Academic Press, London, 21–74.

[85] Pinna, M. C. C. 1991. Concepts and tests of homology in the cladistic paradigm. *Cladistics* 7: 317–338.

[86] Platnick, N. & Gaffney, E. 1978. Evolutionary biology: a Popperian perspective. *Syst. Zool.* 27: 138–141.

[87] Platnick, N. I. 1977. Paraphyletic and polyphyletic groups. *Syst. Zool.* 26: 195–200.

[88] Platnick, N. I. 1978. Gaps and prediction in classification. *Syst. Zool.* 27: 472–474.

[89] Pogue, M. G. & Mickevich, M. F. 1990. Character definition and character state delineation: the bête noire of phylogenetic inference. *Cladistics* 6: 319–361.

[90] Posada, D. & Crandall, K. A. 2001. Selecting the best-fit model of nucleotide substitution. *Syst. Biol.* 50: 580–601.

[91] Ranker, T. A., Smith, A. R., Parris, B. S., Geiger, J. M. O., Haufler, C. H., Straub, S. C. K. & Schneider, H. 2004. Phylogeny and evolution

of grammitid ferns (Grammitidaceae): a case of rampant morphological homoplasy. *Taxon* 53: 415–428.

[92] Rannala, B. & Yang, Z. 1996. Probability distribution of molecular evolutionary trees: a new method for phylogenetic inference. *J. Mol. Evol.* 43: 304–311.

[93] Remane, A. 1952. *Die Grundlagen des natürlichen Systems, der vergleichenden Anatomie und der Phylogenetik - theoretische Morphologie und Systematik.* Akademische Verlagsgesellschaft Geest und Portig, Leipzig.

[94] Rodríguez, F., Oliver, J. L., Marin, A. & Medina, J. R. 1990. The general stochastic model of nucleotide substitution. *J. Theor. Biol.* 142: 485–501.

[95] Rogers, J. S. & Swofford, D. L. 1998. A fast method for approximating maximum likelihoods of phylogenetic trees from nucleotide sequences. *Syst. Biol.* 47: 77–89.

[96] Ronquist, F. 2004. Bayesian inference of character evolution. *TREE* 19: 475–481.

[97] Ronquist, F. & Huelsenbeck, J. P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19: 1572–1574.

[98] Saitou, N. & Nei, N. 1987. The neighbour joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4: 406–425.

[99] Schindel, D. E. & Miller, S. E. 2005. DNA barcoding: a useful tool for taxonomists. *Nature* 435: 17.

[100] Schultz, T. R. & Churchil, G. A. 1999. The role of subjectivity in reconstructing ancestral character states: A Bayesian approach to unknown rates, states, and transformation asymmetries. *Syst. Biol.* 48: 651–664.

[101] Scotland, R. W. 2000. Taxic homology and three-taxon statement analysis. *Syst. Biol.* 49: 480–500.

[102] Scotland, R. W., Olmstead, R. G. & Bennett, J. R. 2003. Phylogeny reconstruction: the role of morphology. *Syst. Biol.* 52: 539–548.

[103] Settle, T. 1979. Popper on "When is a science not a science?" *Syst. Zool.* 28: 521–529.

[104] Steel, M. & Penny, D. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. *Mol. Biol. Evol.* 17: 839–850.

[105] Sullivan, J. & Swofford, D. L. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. *J. Mamm. Evol.* 4: 77–86.

[106] Swofford, D. L. 2002. *PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods), version 4.* Sinauer Associates, Sunderland.

[107] Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. 1996. Phylogenetic inference. *In* Hillis, D., Moritz, C. & Mable, B. (eds.) *Molecular systematics.* 2 ed. Sinauer Associates, Sunderland, 407–514.

[108] Swofford, D. L., Waddell, P. J., Huelsenbeck, J. P., Foster, P. G., Lewis, P. O. & Rogers, J. S. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst. Biol.* 50: 525–539.

[109] Tajima, F. & Nei, M. 1984. Estimation of evolutionary distance between nucleotide sequences. *Mol. Biol. Evol.* 1: 269–285.

[110] Wheeler, Q. D. 2004. Taxonomic triage and the poverty of phylogeny. *Phil. Trans. R. Soc. Lond. B* 359: 571–583.

[111] Wiens, J. J. 2004. The role of morphological data in phylogeny reconstruction. *Syst. Biol.* 53: 653–661.

[112] Wiley, E. O. 1975. Karl P. Popper, systematics, and classification: a reply to Walter Bock and other evolutionary taxonomists. *Syst. Zool.* 24: 233–242.

[113] Wiley, E. O. 1981. *Phylogenetics: the theory and pratice of phylogenetic systematics.* John Wiley and Sons, New York.

[114] Will, K. W., Mishler, B. D. & Wheeler, Q. D. 2005. The perils of DNA barcoding and the need for integrative taxonomy. *Syst. Biol.* 54: 844–851.

[115] Williams, D. & Humphries, J. 2003. Homology and character evolution. *In* Stuessy, T., Mayer, V. & Hörandl, E. (eds.) *Deep morphology.* A.R.G. Gantner Verlag, Liechenstein, 119–130.

[116] Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA se-

quences with variable rates over sites: approximate methods. *J. Mol. Evol.* 39: 306–314.

[117] Yang, Z. & Rannala, B. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Mol. Biol. Evol.* 14: 717–724.

Capítulo 2

Algoritmos Básicos em Filogenética

2.1 Abstract

Given the methodological improvements achieved in the last five decades, current phylogenetics could be viewed as a deeply modified discipline from the one that was initiated in the 50's. There are so many different protocols that it is truely impossible to group all of them at once. However, it is also true that any user should be able to handle and understand the basic concepts and tools available to him/her, otherwise his/her results could have no value due to a methodological misspecification. Thus, here I present an introduction to the very basic methods of tree construction and optimization using maximum likelihood and bayesian methods.

2.2 Resumo

A filogenética atual poderia ser considerada como uma área totalmente diferente do que poderia ter sido vislumbrado há cinco décadas atrás, dado o avanço metodológico que ocorreu no período. Atualmente, a quantidade de métodos disponível é tão alta que é impossível reuní-la em sua totalidade. Entretanto, uma visão geral dos principais recursos analílicos deveria ser uma preocupação de qualquer usuário da área; de outra forma, os resultados apresentados podem ser invalidados por uma opção metodológica equivocada. Nesse sentido, aqui é feita uma breve apresentação dos métodos básicos de construção e otimização de árvores usando máxima verossimilhança e análise bayesiana.

2.3 Introdução

Atualmente, a filogenética tornou-se, indubitavelmente, o método mais utilizado em estudos de biologia comparada, tanto em nível de organismo como de táxon. A filogenética já conseguiu, inclusive, evadir e/ou influenciar várias outras áreas[1] do conhecimento tão diversas quanto o desenvolvimento de vacinas (*e.g.*, Fitch *et al.* [22]), a epidemiologia (*e.g.*, Morgan *et al.* [50]), o estudo da diversificação de idiomas/dialetos (*e.g.*, Dunn [12]) e até mesmo estudos forenses (*e.g.*, Metzker *et al.* [49]). Todas essas aplicações, somadas à sua complexidade computacional, levou, conseqüentemente, ao desenvolvimento necessário e inevitável dos vários métodos que estão em uso corrente (veja Felsenstein [20] e Baxevanis *et al.* [2] para uma análise mais abrangente dos métodos disponíveis). Dada essa grande quantidade de algoritmos computacionais existentes (e em pleno uso) e a importância atual da filogenética, neste trabalho é apresentada uma visão geral e simplificada dos métodos mais utilizados na área, mas limitando-se apenas aos métodos de construção e otimização de árvores usados no PAUP* (Swofford [62]) e MrBayes (Ronquist & Huelsenbeck [60]), os dois programas mais usados atualmente.

Esses métodos não são, de maneira nenhuma, originais e a maioria deles já foi descrita na literaura há mais de 10 anos. Os métodos de busca descritos aqui já foram inclusive compilados por outros autores (*e.g.*, Marques [46]).

[1] Para uma lista de várias outras aplicações da filogenética, veja o trabalho de Fitch ([21]) e para aplicações específicas na saúde pública, veja o trabalho de Hillis ([34]).

2.4 Buscas por árvores ótimas

As buscas por árvores ótimas se dividem em duas categorias: 1) buscas exatas e 2) buscas heurísticas.

2.4.1 Buscas exatas

As buscas exatas se caracterizam por encontrar *todas as árvores ótimas* e se dividem em dois grupos: 1) buscas exaustivas e 2) buscas por *branch-and-bound*.

2.4.1.1 Buscas exaustivas

Durante uma busca exaustiva, *todas as árvores possíveis* são analisadas. O número possível de árvores enraizadas é dado por (Edwards & Cavalli-Sforza [13])

$$\frac{(2n-3)!}{2^{(n-2)}(n-2)!} \tag{2.1}$$

dado que a posição da raiz não influencia na otimização da árvore[2], os diferentes programas de análise filogenética fazem as buscas com árvores não-enraizadas (ou diagramas não-enraizados, DNE) e o número possível de DNE é dado por (Cavalli-Sfforza & Edwards [8], Felsenstein [18], Phipps [53])

$$\frac{(2n-5)!}{2^{(n-3)}(n-3)!} \tag{2.2}$$

Dada a Equação 2.2, fica claro que o número de árvores possíveis aumenta fatorialmente com a adição de terminais (Tabela 2.1) e que, portanto, nem sempre

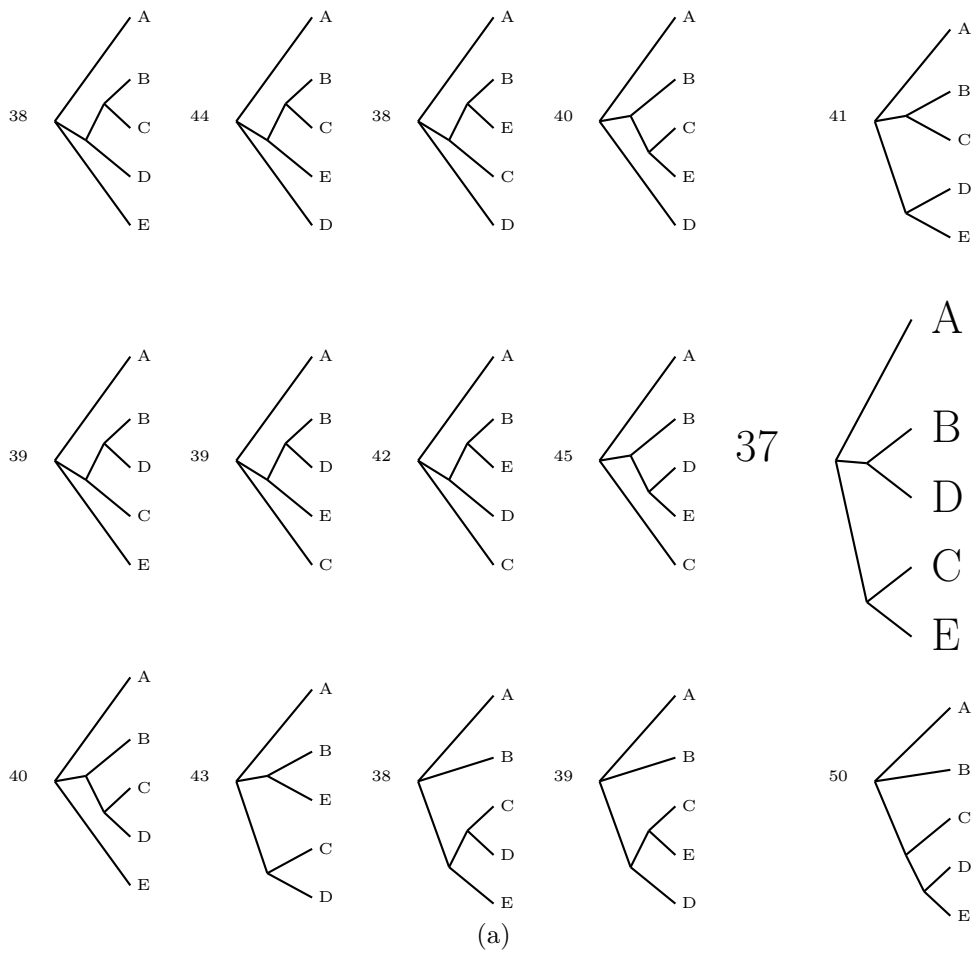
[2]No caso de análises que usam modelos evolutivos, estão sendo assumidos modelos reversíveis.

Tabela 2.1 Relação do número de terminais com o número de DNE.

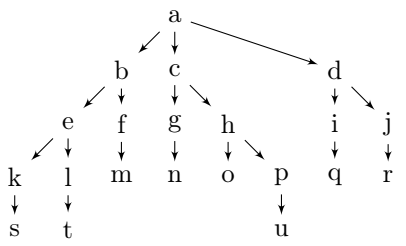
Número de terminais	Número de DNE
3	1
4	3
5	15
10	2027025
100	170045880929340613713932714496783633 x 10^{147}
1000	192672531682447655562275815720090839 x 10^{2825}

é possível realizar buscas exaustivas. Por exemplo, na versão de 32 bits do PAUP* (Swofford [62]), só é possível encontrar todas as árvores binárias possíveis com até 12 terminais, dado que um número maior que 12 exaure a capacidade do programa em buscar todas as árvores definidas pela Equação 2.2 num sistema de 32 bits (Swofford [62]).

Inicialmente, são escolhidos os 3 primeiros terminais da matriz e se constrói um DNE (o único possível, veja Equação 2.2 e Tabela 2.1). Então, um quarto terminal é adicionado ao DNE inicial e assim sucessivamente até que todas as possibilidades definidas pela Equação 2.2 sejam construídas, *i.e.*, até que todo o espaço de árvores seja varrido (a Figura 2.1(a) apresenta todos os DNE possíveis para cinco terminais). Então, os diferentes DNE são avaliados de acordo com o critério de otimização em uso (veja o item 2.5 abaixo) e o(s) melhor(es) é(são) escolhido(s). Dessa forma, dentre as classes de algoritmos de busca, uma busca exaustiva usa um algoritmo de largura (Figura 2.1(b)) na árvore de busca (Papadimitriou & Steiglitz [52]).



(a)



(b)

Figura 2.1 Representação esquemática de uma busca exaustiva (todas as árvores possíveis são construídas). (a) Ilustração de todas as árvores possíveis para cinco terminais (A, B, C, D, e E), a árvore com valor ótimo está destacada. (b) Grafo demonstrando a seqüência de análise do algoritmo de largura, neste caso a seqüência seria **a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u**.

2.4.1.2 *Branch-and-bound*

Achar a(s) árvore(s) ótima(s) representa um problema *NP-hard* (Foulds & Graham [23]), *i.e.*, a partir de um certo número de terminais torna-se inviável resolver o problema analiticamente, o que fica claro a partir dos resultados da Equação 2.2, apresentados na Tabela 2.1.

Adicionalmente, dada a limitação computacional imposta pelas buscas exaustivas, o uso de algoritmos *branch-and-bound* podem ser usados como uma solução "temporária" para se realizar análises exatas com um número de terminais superior ao usado em análises exaustivas. *Branch-and-bound* é uma classe de algoritmos bastante usada para resolver problemas de combinatória em diversas áreas (Papadimitriou & Steiglitz [52]). No caso de filogenética, o uso desse tipo do algoritmo foi introduzido por Hendy & Penny [31].

Uma busca por *branch-and-bound* (Figura 2.2(a)) se caracteriza por encontrar *todas as árvores ótimas*, mas no entanto, sem a necessidade de se analisar todas as árvores possíveis (só analisará todas as árvores possíveis no pior caso, o que a fará equivalente à busca exaustiva).

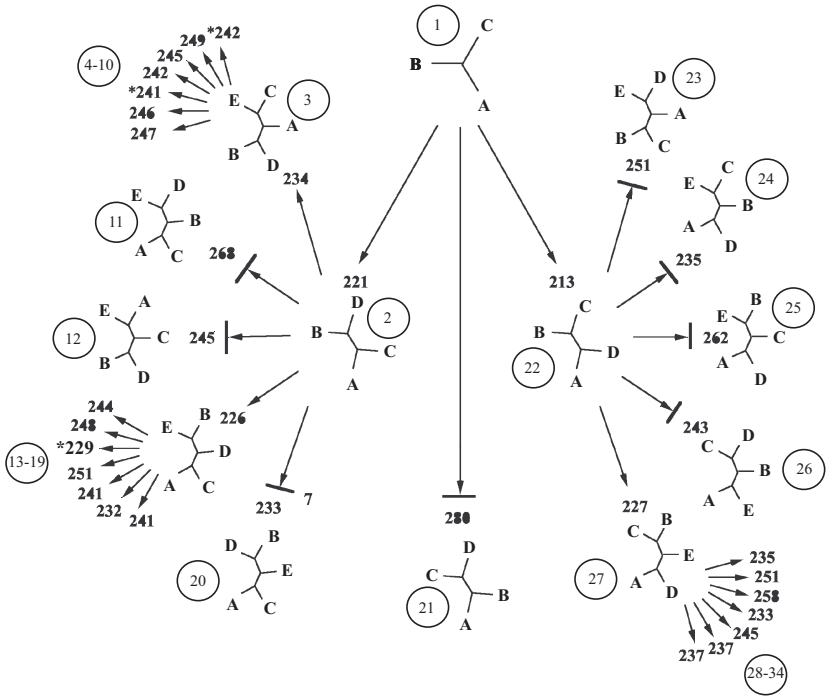
Como pode ser notado na Figura 2.2(a), a busca é muito parecida com a busca exaustiva: inicialmente, constrói-se um DNE com os três primeiros terminais da matriz. Então, adiciona-se o quarto terminal e constrói-se um dos três possíveis DNE e assim sucessivamente até que o último terminal seja adicionado à árvore de busca (=DNE). Entretanto, diferente da busca exaustiva, a cada passo o DNE construído é avaliado de acordo com o critério de otimização em uso (veja o item 2.5 abaixo) e cujo valor atribuído ao DNE é guardado como limite máximo (daí

o "bound") para aquele nível da busca (*i.e.*, para quando aqueles terminais estiverem agregados). Então, após incluir todos os terminais, o algoritmo retorna para um passo imediatamente anterior e continua em outra direção. Caso essa via apresente um *valor de otimização inferior* (não necessariamente menor, dado que em parcimônia o melhor valor é o menor) ao da via anterior naquela etapa, a via atual é descartada. O algoritmo, então, volta para o passo imediatamente anterior e segue em outra direção.

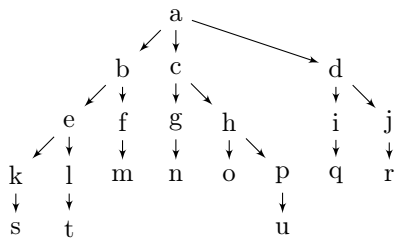
Como reportado por Hendy & Penny ([31], p. 277), às vezes o tempo de busca pode ser melhorado de 55 dias para menos de 5 minutos aplicando-se o algoritmo de *branch-and-bound* apresentado por eles. Apesar disso, o *branch-and-bound* também tem limites e a partir de um certo número de terminais também não é possível achar a(s) solução(ões) ótima(s). O PAUP* (Swofford [62]) não impõe limite para o número de terminais em buscas por *branch-and-bound*, pois não há uma relação direta entre o número de terminais e o número de DNE que pode ser analisado durante uma busca. Além disso, a capacidade do programa em achar soluções ótimas também será influenciada pela (in)congruência dos caracteres.

2.4.2 Buscas heurísticas

Soluções exatas são possíveis apenas para análises que envolvem um número pequeno ou mediano de terminais. No caso de grande número de terminais, buscas heurísticas (aproximadas) são necessárias. As buscas heurísticas *não garantem achar a (ou todas as) solução(ões) ótima(s)*, garantem apenas que a(s) solução(ões) encontrada(s) foi(foram) a(s) melhor(es) achadas pelo programa.



(a)



(b)

Figura 2.2 Representação esquemática de uma busca por *branch-and-bound*. (a) Ilustração do método de construção das árvores (números em círculos representam a ordem em que os DNE são avaliados e o número ao lado de cada um representa seu comprimento, modificado de Swofford, com. pess). (b) Grafo demonstrando a seqüência de análise do algoritmo de profundidade, neste caso a seqüência seria **a, b, e, k, s, l, t, f, m, c, g, n, h, o, p, u, d, i, q, j, r**.

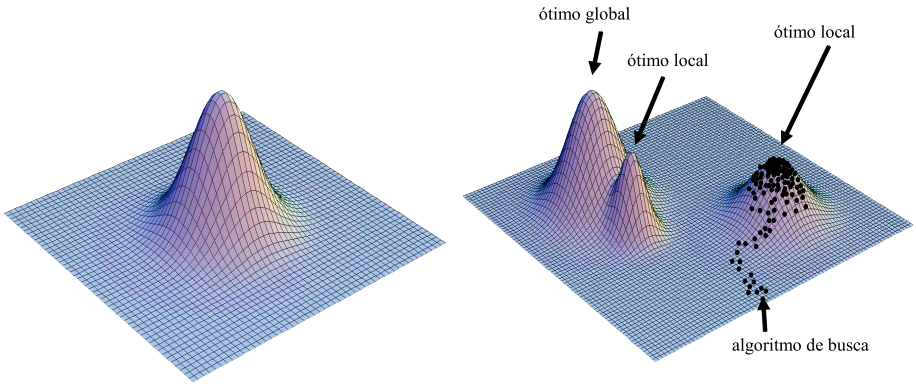
Os métodos heurísticos, em geral, são do tipo "escalada-de-morro", *i.e.*, o método tenta achar o ponto mais elevado (ou menos elevado, de acordo com o critério de otimização utilizado) na distribuição de árvores[3] (Figura 2.3(a)), mas não necessariamente o acha. Um DNE inicial é construído e tenta-se melhorá-lo buscando o ponto mais elevado da distribuição de valores do critério de otimização (Figura 2.3(a)). Entretanto, se essa distribuição for multimodal (com vários "morros") não há maneira de se saber se o método alcançou o pico do morro mais elevado da distribuição (ótimo global), o que seria o ideal, ou se subiu no morro que não é o mais alto da distribuição (ótimo local) (Figura 2.3(b)). Caso ocorra o último caso, o algoritmo chegará a uma solução que não é a ótima e não a abandonará, pois ele não conseguirá descer do morro em direção ao "vale", dado que no vale residem os DNE com valores de otimização piores que o atual. Esse é o principal problema dos métodos heurísticos em geral, dado que os algoritmos são "tolos" e *sempre* subirão no morro mais próximo e, uma vez no topo desse morro, não conseguem descer (mas veja o item 2.5.3.4 abaixo).

As buscas heurísticas normalmente são executadas em duas etapas: 1) construção de uma (ou mais) árvore inicial (seqüência de adição) e 2) otimização adicional dessa árvore (permutação de ramos).

2.4.2.1 Seqüência de adição

Seqüência de adição é a adição de terminais ao DNE em "desenvolvimento", até que todos os terminais sejam adicionados, e pode ser executada de diferentes formas.

[3] E outros "parâmetros", dependendo do critério de otimização em uso.



(a) (b)

Figura 2.3 Representação esquemática de duas possíveis distribuições de árvores definidas pela Equação 2.2. (a) Distribuição unimodal. (b) Distribuição multimodal.

Inicialmente, três terminais são escolhidos para compor o DNE inicial. Então, um quarto terminal é adicionado ao DNE inicial em todas as possibilidades, todos os DNE resultantes são avaliados de acordo com o critério de otimização em uso e o(s) que apresentar(em) o melhor valor é(são) mantido(s) e apenas este(s) será(ão) usado(s) na próxima rodada. Repete-se o passo anterior até que todos os terminais sejam adicionados. A representação esquemática de uma seqüência de adição é mostrada na Figura 2.4.

Entretanto, dado que o DNE inicial pode levar a resultados diferentes caso se altere os três terminais que o formam, são necessários métodos para definir quais serão os terminais que comporão o DNE inicial. Para tanto, existem basicamente 5 métodos (no PAUP*), descritos abaixo.

2.4.2.1.1 ***As is*** Os terminais são adicionados na ordem com que eles estão na matriz. Inicia-se com os primeiros três e adiciona-se o restante seqüencialmente.

2.4.2.1.2 ***Closest*** Inicialmente, são construídos todos os possíveis DNE de três terminais e seus valores de otimização são avaliados. O DNE que possuir o melhor valor, de acordo com o critério de otimização em uso, é escolhido como o DNE inicial. Posteriormente, cada um dos terminais restantes é adicionado ao DNE inicial em todas as posições possíveis e o DNE produzido que possuir o melhor valor é escolhido para a próxima rodada e assim sucessivamente até que todos os terminais sejam adicionados ao DNE. Dessa forma, fica claro que o *closest* é computacionalmente mais intensivo que o *as is*, dado que todas as combinações possíveis de 3 terminais, 4 terminais, 5 terminais etc. são avaliadas.

2.4.2.1.3 ***Furthest*** É o contrário do *closest* e o PAUP* permite seu uso apenas quando *todos* os caracteres são ordenados ou não-ordenados. Entretanto, no último caso o programa subsbtitui o *furthest* pelo *simple*.

2.4.2.1.4 ***Random*** Inicialmente, um número pseudo-aleatório é gerado para servir como semente no processo de permutação dos terminais a serem usados. Três terminais são escolhidos aleatoriamente para montar o DNE inicial. Então, o quarto terminal é escolhido aleatoriamente dentre os terminais restantes e assim sucessivamente.

2.4.2.1.5 ***Simple*** Os terminais são adicionados de acordo com o índice de avanço de Farris [14] que apresentam em relação a um terminal de referência. Inicialmente, calcula-se a distância entre cada um dos terminais e o terminal de referência. Então, os terminais são adicionados em ordem decrescente de distância (ou crescente de avanço). Dessa forma, o DNE inicial é formado pelo terminal de referência e os dois

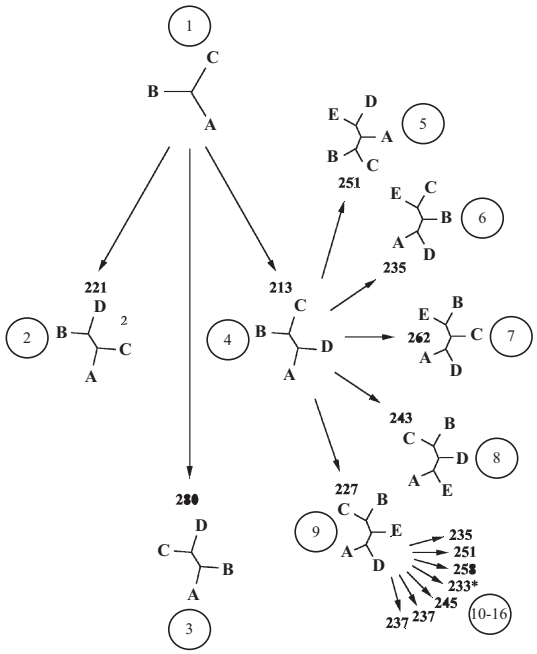


Figura 2.4 Representação esquemática do processo de seqüência de adição (números em círculos representam a ordem com que os DNE são avaliados e os números ao lado de cada DNE representam seus comprimentos, modificado de Swofford, com. pess).

terminais mais "próximos" (com menor distância) a ele. Os outros terminais são adicionados de acordo com seu ranqueamento quando se calculou as distâncias no primeiro passo.

2.4.2.2 Permutação de ramos

Como mostrado anteriormente, os métodos heurísticos são suscetíveis a ficarem presos em ótimos locais. Adicionalmente, os métodos de seqüência de adição são usados apenas para escolher a(s) árvore(s) de partida, *i.e.*, um ou mais pontos iniciais na distribuição de todas as árvores possíveis. Portanto, essas árvores iniciais precisam ser "melhoradas" para se tentar varrer o espaço de árvores. Esse "melhora-

mento" é feito por rearranjos pré-definidos na(s) árvore(s) construída(s) inicialmente. Em geral, esses rearranjos são perturbações executadas nas árvores obtidas pelos métodos de adição, como é o caso do NNI, SPR, TBR e derivações, ou em outras árvores oriundas de outras fontes (*e.g.*, árvores de consenso), como a fusão de árvores[4] (Goloboff [27]).

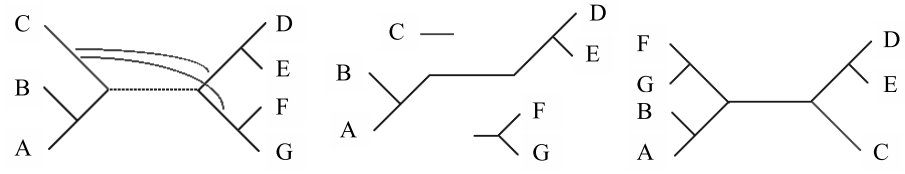
Dada a grande quantidade de derivações dos métodos básicos e suas respectivas aplicações, aqui serão apresentadas apenas as "classes" gerais dos métodos implementados no PAUP* (Swofford [62]) e no MrBayes ([60]). Por exemplo, apenas o MrBayes usa 52 métodos diferentes de perturbações[5], dentre as quais estão as diferentes versões do LOCAL[6] (descrito por Larget & Simon [41]), quatro tipos de NNI, nove de SPR, cinco de TBR, e uma nova "classe" o SS (Ronquist & Huelsenbeck [60]). Dos métodos descritos por Larget & Simon ([41]), apenas o LOCAL (com e sem relógio molecular) é usado no MrBayes, entretanto este procedimento é um tipo de NNI e não será apresentado aqui. Para a descrição de outras "classes" de métodos, veja, *e.g.*, Felsenstein ([20]).

2.4.2.2.1 NNI *Nearest Neighbor Interchanges* - trocas de vizinhos mais próximos. Cada ramo interno de um DNE define 4 sub-árvores (=vizinhos ou "sub-DNE"). Por exemplo, o ramo pontilhado na Figura 2.5(a) define as sub-árvores (C) e (A,B), do lado esquerdo, e (D,E) e (F,G), do lado direito (as linhas curvas mostram as possíveis posições onde (C) pode ser inserida). Na Figura 2.5(b) são destacadas as sub-árvores (C) do lado esquerdo e a (F,G) do lado direito, as candidatas à permuta-

[4] A fusão de árvores pode ser executada em quaisquer árvores, não apenas nas de consenso.

[5] Veja o código do programa, arquivo `mcmc.c`, linhas 282 a 333.

[6] Não confundir com o nome antigo do NNI usado nas versões anteriores à 3.1 do PAUP (Swofford & Begle [63]).



(a) (b) (c)

Figura 2.5 Representação esquemática do NNI (modificada de Siddall [61]). (a) DNE base. (b) Dois vizinhos isolados do DNE base. (c) Troca entre (C) e (F,G).

ção; e na Figura 2.5(c) duas sub-árvores separadas na Figura 2.5(b) são efetivamente permutadas.

A troca de uma sub-árvore de um lado por uma do outro lado do ramo caracteriza um procedimento/movimento NNI. Para cada ramo interno apenas dois NNI são possíveis e o NNI escolhido é o que maximiza o critério de otimização.

2.4.2.2.2 SPR *Subtree Pruning-Regrafting* - podagem e re-enxerto de sub-árvore. Neste procedimento, uma sub-árvore é podada do DNE através da dissolução de um nó interno (Figura 2.6(a), seta). Essa sub-árvore podada, que possui um ramo livre, é, então, re-enxertada em todas as posições possíveis do outro DNE (Figura 2.6(b)). A posição em que o re-enxerto maximiza o critério de otimização é escolhida como o ponto ótimo (Figura 2.6(c)).

2.4.2.2.3 TBR *Tree Bisection-Reconnection* - bissecção e reconexão de árvore. Neste método, o DNE-base é cortado em um determinado ramo interno (*i.e.*, um ramo é eliminado) levando à formação de duas sub-árvores (Figuras 2.7(a)-2.7(b)). Essas duas sub-árvores são, então, reconectadas por um determinado ramo de cada

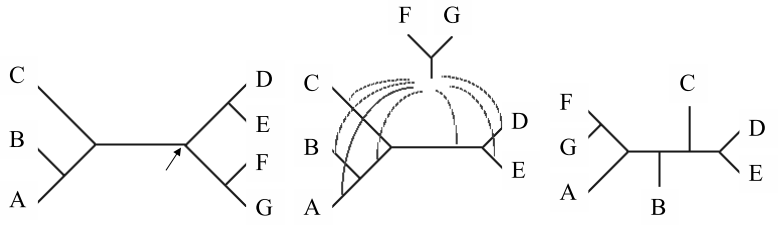


Figura 2.6 Representação esquemática do SPR (modificada de Siddall [61]). (a) DNE base (seta indica o nó alvo). (b) A sub-árvore (F,G) é podada e todas as posições possíveis de re-enxerto são indicadas pelas linhas curvas. (c) Re-enxerto da sub-árvore (F,G) no ramo que leva à sub-árvore (A).

uma delas (Figura 2.7(c)). Todas as possíveis bisecções e reconexões são tentadas e a reconexão que maximizar o critério de otimização é escolhida (Figura 2.7(d)). Melhoramentos ao TBR (e SPR) já foram propostos por vários autores, entre eles estão Goloboff ([26]) e Ronquist ([58]).

2.4.2.2.4 SS *Subtree Swapping* - permutação de subárvore. Como o SPR e o TBR, este procedimento é hierarquicamente mais abrangente que o NNI, mas, apesar de sua simplicidade, existem diferenças interessantes entre ele e o SPR e/ou TBR. O SS simplesmente seleciona duas subárvores aleatoriamente e troca suas posições (subárvores S_2 e S_5 na Figura 2.8). Note que com SPR ou TBR não é possível ir da árvore mostrada na Figura 2.8(a) para a 2.8(b) em apenas um movimento.

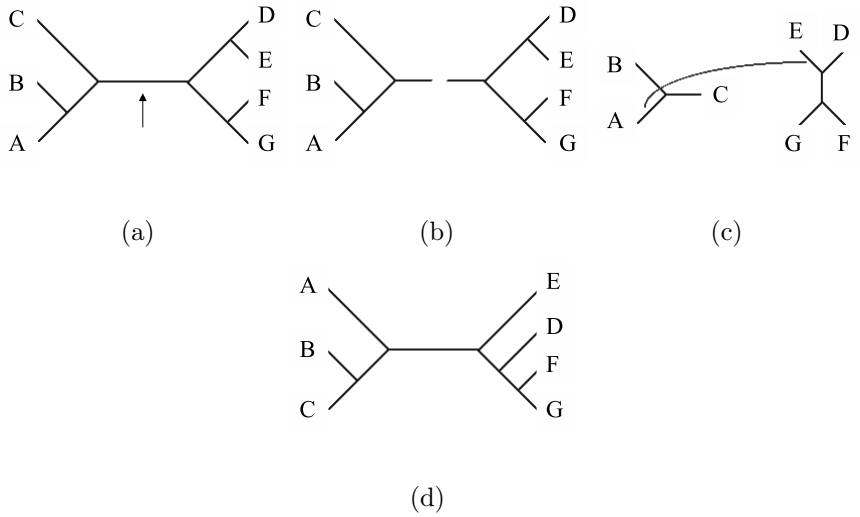


Figura 2.7 Representação esquemática do TBR (modificada de Siddall [61]). (a) DNE base (seta indica o ramo alvo a ser cortado). (b) Corte do ramo destacado em (a). (c) As duas sub-árvores resultantes do corte, (A,B,C) e ((D,E),F,G). (d) Reconexão das duas sub-árvores obtidas em (c) através de um ramo de cada.

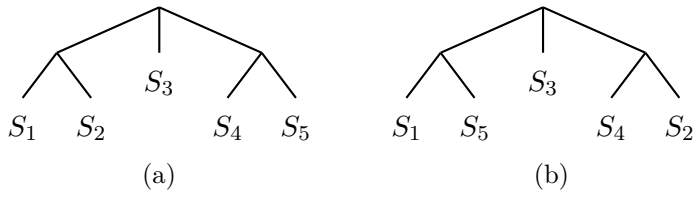


Figura 2.8 Representação esquemática do SS (S = subárvore). (a) Árvore inicial. (b) Árvore após o movimento SS, note que as subárvores S_2 e S_5 trocaram de posição.

2.5 Otimização

Embora seja feita de forma concomitante aos procedimentos descritos anteriormente (veja o item 2.4), a otimização é melhor entendida se descrita separadamente. Nesse sentido, aqui serão apresentados os principais critérios de otimização usados atualmente em filogenética, são eles: 1) parcimônia, 2) máxima verossimilhança e 3) probabilidade posterior (análise bayesiana).

2.5.1 Parcimônia

Atualmente, em filogenética, a parcimônia pode ser vista de duas maneiras diferentes: 1) como princípio filosófico e 2) como critério de otimização.

2.5.1.1 Princípio filosófico

Segundo Thorburn ([65]), desde a metade do século XIX, quase todo livro de lógica traz a frase "*Entia non sunti multiplicanda, prœter necessitatem*" citada como se fosse de William de Ockham (provavelmente um equívoco na história da filosofia repetido por diversas vezes e sacramentado por Hamilton (1852, *apud* Thorburn [65]) ao criar o termo "Occam's razor" e substituir "frugalidade", termo mais antigo, por "parcimônia", veja Burns [7] e Thorburn [65]).

Embora de origem controversa, a "lei da parcimônia [...] é um preceito puramente lógico" (Thorburn [65]), e, como tal, pervade qualquer atividade científica que envolva raciocínio lógico, como por exemplo as idéias newtonianas (Newton [51]). Isso significa que a parcimônia enquanto princípio filosófico é válida em qualquer abordagem à filogenética, seja ela qual for.

2.5.1.2 Critério de otimização

Apesar da validade da parcimônia enquanto princípio filosófico, o seu uso enquanto critério de otimização apresenta vários problemas associados. Assim, uma discussão mesmo que rápida se faz necessária por questões históricas.

O uso da parcimônia como critério de otimização pode ter sua origem (implícita) traçada ao trabalho de Hennig ([33], p. 104), em seu *princípio auxiliar*:

"*It must be recognized as a principle of inquiry for the practice of systematics that agreement in characters must be interpreted as synapomorphy as long as there are no grounds for suspecting its origin to be symplesiomorphy or convergence.*

Posteriormente, a parcimônia foi evocada de forma explícita por Wiley [68], em seus "axiomas popperianos".

A partir de então, vários autores têm defendido a utilização da parcimônia como critério de otimização, sendo o trabalho de Farris [15] um dos mais influentes.

Entretanto, uma das forma utilizadas pelos autores (*e.g.*, Farris [15]) para justificar a parcimônia como critério de otimização é a negação de outras possibilidades analíticas, tais como a máxima verossimilhança (*e.g.*, Felsenstein [16], [17]), especialmente no que se refere ao uso de modelos explícitos de substituição. Essa negação, entretanto, também incorporou uma certa quantidade de argumentos filosóficos (*e.g.*, Kluge [37], veja Hull [36] para uma abordagem histórica detalhada), na tentativa de justificá-la e encaixá-la no arcabouço filosófico-científico vigente na época, a versão popperiana (veja o item 2.5.1.3, abaixo).

Assim, uma das formas mais comuns de se defender a parcimônia é a suposição de que ao usá-la como critério de otimização, se está minimizando o número de

"hipóteses" *ad hoc* sobre o *explanandum* (Brower [5]). Entretanto, essa suposição de simplificação exacerbada (diminuição de "hipóteses" *ad hoc*) é infeliz, dado que ela normalmente não é realista nem tem o maior conteúdo empírico.

Infelizmente, a adoção da parcimônia como critério de otimização tem sido caracterizada por opiniões recheadas de imparcialidade (*e.g.*, Farris [15], Kluge [38], Kluge & Grant [39], Kluge & Wolf [40]), o que só tende a travar o desenvolvimento/melhoramento (*e.g.*, Wheeler [67]) dela própria.

2.5.1.3 Popperianismo e inferência filogenética

Como citado anteriormente, uma das formas encontradas para justificar a parcimônia e, conseqüentemente, negar outros critérios de otimização (*e.g.*, máxima verossimilhança), foi enquadrá-la em um arcabouço filosófico, o de Popper ([54], [55]). Por outro lado, alguns defensores da máxima verossimilhança (*e.g.*, DeQueiroz & Poe [10], [11]) também enquadram-na nas idéias popperianas e tentam justificá-la nos mesmos termos. Ambos os grupos analisam suas respectivas abordagens em relação ao popperianismo e chegam a conclusões opostas (Rieppel [57]).

Recentemente, Helfenbein & DeSalle [30] apresentaram um retrospecto e reavaliação da adequação dos principais critérios de otimização à filosofia popperiana. Nesse trabalho, os autores argumentam que a parcimônia se adequa aos critérios estabelecidos por Popper ([54]), dado que "a análise por parcimônia leva à descoberta de sinapomorfias e, conseqüentemente, de grupos monofiléticos". No entanto, por outro lado, a máxima verossimilhança e a inferência bayesiana não, justamente porque não levariam a esses resultados e, além disso, esses métodos "apresentam a característica ilógica de atribuírem probabilidade a um evento que foi historicamente

único" (Helfenbein & DeSalle [30], p. 278, mas veja o item 1.5.4 do Capítulo 1 desta tese).

Toda essa discussão, entretanto, precisa ser ponderada do ponto de vista biológico. Até que ponto a filosofia popperiana é importante para a biologia, se a própria filosofia de Popper não é, em si, popperiana (Rieppel [57])? Ou até que ponto a biologia, ou mais precisamente, a sistemática, em si se adequa ou precisa se adequar à filosofia de Popper?

Rieppel [57] tem feito questionamentos semelhantes e sugere que, talvez, seja mais interessante que a sistemática seja independente de qualquer filosofia, a não ser a dela própria.

2.5.2 Máxima verossimilhança

O uso da máxima verossimilhança foi introduzido por Edwards & Cavalli-Sforza [13] e posteriormente aprimorado principalmente por Felsenstein (*e.g.*, [16], [17], [19]). A máxima verossimilhança estima a probabilidade de se observar os dados considerando um determinado modelo e é representada da seguinte forma:

$$L = P(X|\theta) \tag{2.3}$$

onde, X representa o conjunto de dados da matriz e θ representa o modelo. Dessa forma podemos reescrever a Equação 2.3 como

$$L = P(X|\theta) = \prod_{i=1}^{n} P(X^{(i)}|\theta) \tag{2.4}$$

O modelo θ é constituído por uma árvore ("topologia" ou DNE e conjunto de comprimentos de ramos dessa árvore) e um modelo de substituição entre os estados de caráter (Equação 1.14). Assim, a Equação 2.4 pode ser reescrita como

$$P(X|\theta) = \prod_{i=1}^{n} P(X^{(i)}|\psi, \varphi) \tag{2.5}$$

onde, ψ representa a árvore e φ representa o modelo de substituição entre os estados de caráter (Equação 1.14). Note que ψ é constituído por uma "topologia" (padrão de ramificação) e comprimentos de ramos associados, da forma

$$\psi = (\tau, \nu) \tag{2.6}$$

onde, τ representa a topologia e ν representa o conjunto de comprimentos de ramos.

Dessa forma, a Equação 2.5 pode ser reescrita como

$$P(X|\theta) = \prod_{i=1}^{n} P(X^{(i)}|\tau, \nu, \varphi) \tag{2.7}$$

Uma vez que τ é definido pela Equação 2.2, para encontrar a(s) árvore(s) de máxima verossimilhança, o algoritmo teria que varrer o espaço definido pela Equação 2.2 usando a Equação 2.7 usando buscas exatas ou heurísticas (veja o item 2.4 acima).

Tabela 2.2 Matriz de caracteres.

	Caracteres		
Terminais	1	2	3
A	2	0	3
B	2	0	0
C	0	2	3
D	0	2	0

2.5.2.1 Exemplo[7]

Por facilitação, os caracteres serão considerados ordenados, do tipo $0 \rightarrow 1 \rightarrow 2 \rightarrow 3$ e irreversíveis e as árvores enraizadas[8].

A forma de aplicação seguirá a publicação original (mas veja Felsenstein [19] para uma versão melhorada do algoritmo). Considere a matriz apresentada na Tabela 2.2 e as árvores da Figura 2.9.

Seja α a probabilidade de mudança por unidade de tempo. Adicionalmente, seja αdt a probabilidade de mudança num intervalo (mínimo) de tempo dt. Se assumirmos que a probabilidade de mudança em qualquer intervalo de tempo for independente do número de intervalos e do número de mudanças que teriam ocorrido anteriormente[9], a probabilidade de ocorrerem k mudanças durante um intervalo de tempo t é dado pela distribuição de Poisson

[7] Modificado de Felsenstein 1973

[8] É importante deixar claro que essas restrições não são necessárias, mas permitem que o exemplo fique mais fácil de ser entendido.

[9] Cadeia de Markov.

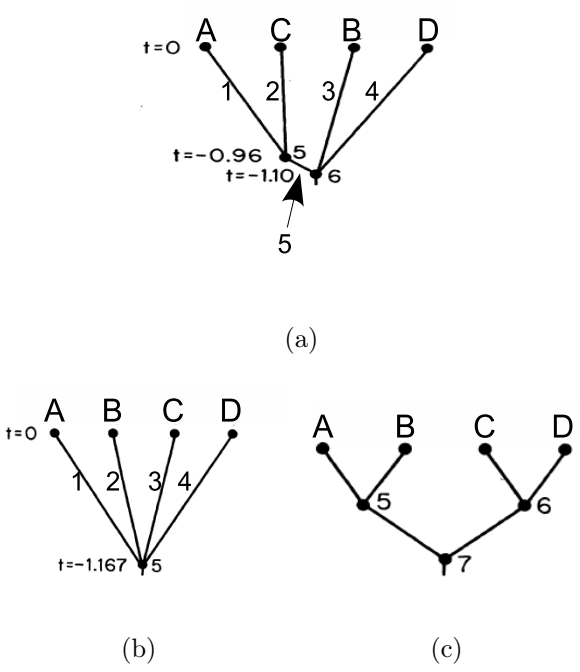


Figura 2.9 Árvores a serem utilizadas como exemplo para a estimação da máxima verossimilhança.

$$\frac{e^{(-\alpha t)}(\alpha t)^k}{k!} \tag{2.8}$$

Considerando que $\alpha = 1$ (num intervalo infinitamente pequeno, o número de mudanças tende a 1), α pode ser descartado e as unidades de tempo que serão estimadas representarão, na verdade, a quantidade de mudança esperada por caráter num determinado intervalo de tempo t. Dessa forma, a Equação 2.8 pode ser reescrita como[10]

$$\frac{e^{(-t)}t^k}{k!} \tag{2.9}$$

Agora, fica mais fácil calcular o valor do estimador de máxima verossimi-

[10]Compare com a Equação 1.12 do Capítulo 1 desta tese.

lhança para as diferentes árvores.

Considerando a árvore (a) e usando a Equação 2.8, temos

$$P_{C1_{r1}} = \frac{e^{(-0.96\alpha)}(0.96\alpha)^2}{2!} = 0.1764370419 \tag{2.10}$$

onde, C=caráter, r=ramo. Então,

$$P_{C1_{r2}} = \frac{e^{(-0.96\alpha)}(0.96\alpha)^0}{0!} = 0.3828928860 \tag{2.11}$$

$$P_{C1_{r3}} = \frac{e^{(-1.1\alpha)}(1.1\alpha)^2}{2!} = 0.2013870057 \tag{2.12}$$

$$P_{C1_{r4}} = \frac{e^{(-1.1\alpha)}(1.1\alpha)^0}{0!} = 0.3328710838 \tag{2.13}$$

Dessa forma, temos que a probabilidade de um determinado caráter (i) em uma determinada árvore ψ será dado por

$$P_{Ci_\psi} = \prod_{r=1}^{2T-2} P_{Ci_r} \tag{2.14}$$

onde, T é o número de terminais. Então,

$$P_{C1_a} = \prod_{r=1}^{3} P_{C1_r} = 0.003937071959 \tag{2.15}$$

Usando o mesmo procedimento para os outros caracteres, temos:

$$P_{C2_a} = \prod_{r=1}^{3} P_{C2_r} = 0.003937071959 \tag{2.16}$$

$$P_{C3_a} = \prod_{r=1}^{3} P_{C3_r} = 0.0008608872921 \tag{2.17}$$

Então, usando a Equação 2.7, temos a verossimilhança da árvore (a)

$$P(X|\psi_a, \varphi) = \prod_{i=1}^{n} P(X^{(i)}|\psi_a, \varphi) = 1.334421413 \text{ x } 10^{-8} \tag{2.18}$$

Usando o mesmo procedimento,

$$P(X|\psi_b, \varphi) = \prod_{i=1}^{n} P(X^{(i)}|\psi_b, \varphi) = 1.249423234 \text{ x } 10^{-8} \tag{2.19}$$

Assumindo que *t=0,5* para os nós 5 e 6 de ψ_c , temos

$$P(X|\psi_c, \varphi) = \prod_{i=1}^{n} P(X^{(i)}|\psi_c, \varphi) = 1.173976679 \text{ x } 10^{-9} \tag{2.20}$$

o que nos mostra que ψ_a é a árvore de máxima verossimilhança. Por outro lado, uma análise por parcimônia escolheria ψ_c como a árvore ótima, apesar de ser a de menor verossimilhança dentre as três analisadas.

Esse exemplo demonstra claramente que quando os comprimentos de ramos (quantidade de mudança esperada) variam de forma a termos ramos longos combinados com, e/ou separados por, um ramo curto, a parcimônia tenderá a escolher a árvore incorreta, levando ao que posteriormente foi chamado de atração de ramos longos ou "Zona de Felsenstein" (Hendy & Penny [32], Swofford *et. al.* [64]).

Apesar das vantagens apresentadas pela máxima verossimilhança, seu cálculo é computacionalmente muito intensivo, dado que seria necessário avaliar a Equação 2.7 para todas as árvores encontradas durante a busca. Por exemplo, dado que τ é definido pela Equação 2.2, uma busca exaustiva pela(s) árvore(s) de máxima verossimilhança precisaria avaliar a Equação 2.7 para todas as árvores possíveis, *i.e.*,

$$P(X|\theta) = \sum_{\tau=1}^{\frac{(2n-5)!}{2^{(n-3)}(n-3)!}} \prod_{i=1}^{n} P(X^{(i)}|\tau, \nu, \varphi) \tag{2.21}$$

daí o motivo das buscas serem tão lentas em máxima verossimilhança. Adicionalmente, no caso de estimativas de suporte, a situação fica ainda mais complicada e dependerá diretamente do número de réplicas usadas.

2.5.3 Análise bayesiana

2.5.3.1 O paradigma bayesiano

O método bayesiano (Bayes [3]) se caracteriza por estimar uma quantidade chamada "probabilidade posterior". Como foi visto no item 2.5.2, a máxima verossimilhança estima a probabilidade de se observar os dados, dada uma árvore com seus respectivos comprimentos de ramos e um modelo de substituição entre estados de caráter (Equação 2.5). A inferência bayesiana baseia-se na máxima verossimilhança (e, portanto, herda algumas de suas características já discutidas no item 2.5.2) para estimar a probabilidade do modelo com base nos dados. Entretanto, neste tipo de análise, os parâmetros são tratados como variáveis aleatórias e, além disso, também se faz uso de "probabilidades *a priori*" (ou simplesmente *priori*) sobre os parâmetros a serem estimados. As *priori* representam as distribuições de probabilidades dos

parâmetros (a serem estimados) independentemente dos dados.

Seja θ o modelo, então uma *priori* sobre θ pode ser representada como

$$P(\theta) \tag{2.22}$$

A *priori* é o "conhecimento" prévio que se tem sobre o modelo, o que pode ser incluído na análise. Por exemplo, se em todas as as análises filogenéticas feitas envolvendo os organismos A e B, eles sempre emergem formando um grupo monofilético, essa informação pode ser usada como uma priori sobre as árvores a serem estimadas em análises futuras (*i.e.*, podem ser usadas para ponderar diferencialmente as árvores). Ou o contrário, se não há nenhuma informação sobre o modelo, então assume-se uma priori não-informativa (note que não-informativa não necessariamente é uniforme, veja *e.g.*, Box & Tiao [4], para discussão sobre distribuição não-informativa e uniforme).

Por outro lado, as probabilidades posteriores possuem uma distribuição (distribuição *a posteriori*, ou simplesmente *posteriori*) geralmente desconhecida (dada sua complexidade) e são calculadas usando-se o teorema de Bayes [3], o qual é representado por

$$P(\theta|X) = \frac{P(X|\theta_i)P(\theta_i)}{P(X|\theta_j)P(\theta_j)} \tag{2.23}$$

onde, $P(X|\theta_i)$ representa a máxima verossimilhança de θ_i (veja a Equação 2.3), $P(\theta_i)$ representa a priori sobre θ_i (veja a Equação 2.22). $P(X|\theta_j)$ e $P(\theta_j)$ representam a máxima verossimilhança e a priori, respectivamente, de todos os outros valores de θ.

Dessa forma, a probabilidade posterior é o produto da máxima verossimilhança com a priori de θ_i, normalizado (para ter volume 1 em Θ) pelo produto da máxima verossimilhança com a priori de todos os outros valores de θ.

Considerando que $\theta = (\psi, \varphi)$ e que θ inclui mais de um ψ e pode incluir mais de um φ, o espaço definido por $\Theta = (\Psi, \Phi)$ inclui todos os ψ e todos os φ. Assim, uma análise bayesiana modela a incerteza da priori em θ com uma distribuição priori conjunta $P(\theta)$ para todos os parâmetros definidos em Θ (note que τ é discreto e particiona Θ). Então, pode-se reescrever a Equação 2.23 como

$$\begin{aligned} P(\theta|X) &= \frac{P(X|\theta)P(\theta)}{\int_\Theta P(X|\theta)P(\theta)d\theta} \\ &= \frac{P(X|\psi,\varphi)P(\psi,\varphi)}{\int_{(\Psi,\Phi)} P(X|\psi,\varphi)P(\psi,\varphi)d\psi d\varphi} \\ &= \frac{P(X|\tau,\nu,\varphi)P(\tau,\nu,\varphi)}{\int_{(T,N,\Phi)} P(X|\tau,\nu,\varphi)P(\tau,\nu,\varphi)d\tau d\nu d\varphi} \end{aligned} \tag{2.24}$$

onde T, N, e Φ representam as árvores, o conjunto dos comprimentos de ramos de cada uma das árvores e o conjunto dos modelos de substituição entre estados de caráter, respectivamente.

Como já visto no item 2.5.2, o numerador da Equação 2.24 pode ser avaliado para qualquer ponto θ_i, entretanto o cálculo do denominador da Equação 2.24 é extremamente complexo, mesmo para um número de terminais pequeno (*e.g.*, 6). Entretanto, uma vez que ele deve somar a (ou ao menos se aproximar muito de) 1, ele pode ser descartado. Conseqüentemente, a probabilidade posterior é proporcional ao produto da máxima verossimilhança pela probabilidade a priori.

Então, para se achar a probabilidade posterior de uma determinada topologia τ_i, é necessário encontrar o volume sob $P(\theta_i|X)$ na porção da partição de Θ definida por τ_i e isso é feito integrando-se todos os outros parâmetros. Então, pode-se separar um determinado ponto $\theta_i = (\tau_i, \nu_i, \varphi_i)$ nas partes que o compõem:

$$\begin{aligned} P(\tau_i|X) &= \frac{\int_N \int_\Phi P(X|\tau_i,\nu_i,\varphi_i)P(\tau_i,\nu_i,\varphi_i)d\nu_i d\varphi_i}{\sum_{\tau_j}\int_N \int_\Phi P(X|\tau_j,\nu_j,\varphi_j)P(\tau_j,\nu_j,\varphi_j)d\nu_j d\varphi_j} \\ &\propto \int_N \int_\Phi P(X|\tau_i,\nu_i,\varphi_i)P(\tau_i,\nu_i,\varphi_i)d\nu_i d\varphi_i \end{aligned} \tag{2.25}$$

2.5.3.2 Monte Carlo com Cadeia de Markov (MCMC)

Uma forma de encontrar uma solução aproximada para problemas computacionalmente complexos, para os quais (ainda) é impossível achar uma solução determinística, é usar o método Monte Carlo (MC). Aplicações do método Monte Carlo são conhecidas pelo menos desde o século XVIII (*e.g.*, Buffon [6]) e são utilizadas em diversas áreas, *e.g.*, econometria, meteorologia, física, psicologia, ecologia etc. (*e.g.*, Guttorp [28], Manly [45]). Basicamente, um método Monte Carlo é o uso (através de integração) de números (pseudo)aleatórios para examinar algum tipo de problema, o qual é geralmente muito complicado (multidimensional) para soluções analíticas e/ou representa um sistema estocástico (dinâmico ou estático).

Em filogenética, a proposição do método foi feita por três grupos independentes quase simultaneamente e dentro do paradigma bayesiano (Li [43], Mau [47], Rannala & Yang [56], para uma rápida revisão dos diferentes algoritmos usados inicialmente veja, *e.g.*, Larget & Simon [41]). Nesse caso, entretanto, foi proposta uma

variante do método MC, Monte Carlo com Cadeia de Markov (MCMC, Metropolis *et al.* [48], Hastings [29]). Uma cadeia de Markov é uma seqüência de variáveis aleatórias $X = (X_0, X_1, X_2, \ldots, X_n)$ com a propriedade de que a probabilidade de estar em um determinado X_i no tempo t depende apenas de um número k de estados anteriores, sendo k a ordem da cadeia.

O método proposto em filogenética faz uso da Equação 2.2 para estimar as árvores e, conseqüentemente, dado que os casos de um número grande, ou mesmo razoável, de terminais representam um problema insolúvel computacionalmente, devido ao grande número de possibilidades a serem analisadas (veja Tabela 2.1), o método MCMC é usado para "caminhar" no espaço definido pela Equação 2.24, amostrar pontos desse espaço e usar essas amostras como uma estimativa dos valores dos parâmetros (considerando a árvore como um dos parâmetros). Essas amostras são extraídas até que não haja variação significativa entre elas (Figura 2.10), *i.e.*, aumentar a amostragem não levaria a uma melhora nos valores já estimados, o que significa que a cadeia teria encontrado a distribuição de equilíbrio.

Essa abordagem traz várias vantagens, dentre elas:

1) o problema computacional é fortemente amenizado pelo uso do método Monte Carlo (embora uma solução exata não seja encontrada);

2) diferente da máxima verossimilhança, a qual faz estimativas de ponto no espaço definido pela Equação 2.2, os resultados de uma MCMC podem ser intuitivamente usados para se propor uma medida de suporte de ramos internos (usando-se a freqüência com que esses ramos aparecem nas diferentes árvores amostradas) e, conseqüentemente, para estimar a incerteza filogenética (*e.g.*, Ronquist [59], Larget

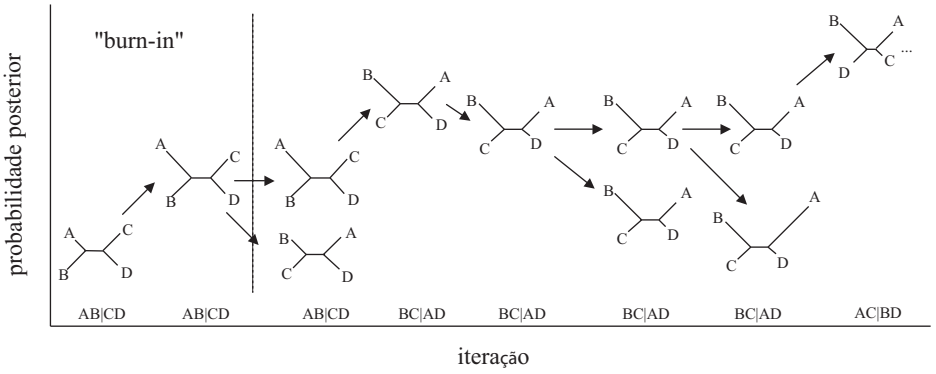


Figura 2.10 Representação esquemática de uma busca usando MCMC (modificado de Swofford, com. pess.). Após o período de "burn-in", a cadeia se aproxima da distribuição de equilíbrio.

& Simon [41]);

3) modelos mais complexos e biologicamente mais realistas podem ser usados.

Dentro do paradigma bayesiano, então, a MCMC é usada para extrair amostras de Θ. Assim, a MCMC caminha no espaço definido por Θ (Figura 2.11) e retira uma seqüência de amostras dependentes, $\theta_1, \theta_2, \cdots, \theta_n$, de forma que, após um determinado tempo, a distribuição das amostras realizadas se aproxima da distribuição posterior dos parâmetros alvos. Conseqüentemente, após uma caminhada suficientemente longa, a distribuição das amostras se aproxima da posteriori e as freqüências das topologias amostradas constituem uma representação válida de suas respectivas probabilidades posteriores (*e.g.*, Larget & Simon [41], Tierney [66]).

Uma vez que a MCMC retira amostras de Θ, cada uma dessas amostras também fornece informações sobre ψ e φ e, então, se obtém uma estimativa pontual de todos os parâmetros concomitantemente a cada movimento da MCMC.

2.5.3.3 MCMC em Θ

Suponha uma MCMC iniciando em um determinado θ_i de Θ (Figura 2.11). Para que essa MCMC se mova (*i.e.*, mude de estado), um movimento é proposto para θ_j de acordo com a função de densidade de probabilidade $Q(\theta_j|\theta_i)$. A MCMC, então, precisa aceitar ou rejeitar esse movimento e isso é feito com o algoritmo de Metropolis-Hastings (M-H, Hastings [29], Metropolis *et al.* [48]), o qual aceita o movimento para o novo estado proposto θ_j, a partir do estado atual θ_i usando a probabilidade r

$$\begin{aligned} r &= min\left(1, \frac{P(\theta_j|X)Q(\theta_i|\theta_j)}{P(\theta_i|X)Q(\theta_j|\theta_i)}\right) \\ &= min\left(1, \frac{P(X|\theta_j)}{P(X|\theta_i)}\frac{P(\theta_j)}{P(\theta_i)}\frac{Q(\theta_i|\theta_j)}{Q(\theta_j|\theta_i)}\right) \end{aligned} \tag{2.26}$$

Dado que Q é comumente (embora nem sempre) simétrica, a razão de Hastings ([29], último elemento do segundo termo da Equação 2.26) é igual a 1 (*i.e.*, não afeta a probabilidade de aceitação) e o segundo elemento do segundo termo da Equação 2.26 é anulado. Dessa forma, é necessário calcular apenas a razão de verossimilhança (primeiro elemento do segundo termo da Equação 2.26)

Após calcular r, gera-se uma variável aleatória U uniformemente distribuída no intervalo $(0, 1)$. Se U for menor que r, então aceita-se o estado proposto e $\theta_i = \theta_j$. Caso o movimento proposto seja rejeitado, o estado atual θ_i é repetido e uma nova proposição é feita.

A próxima etapa é fazer com que a MCMC percorra Θ e tente chegar à região

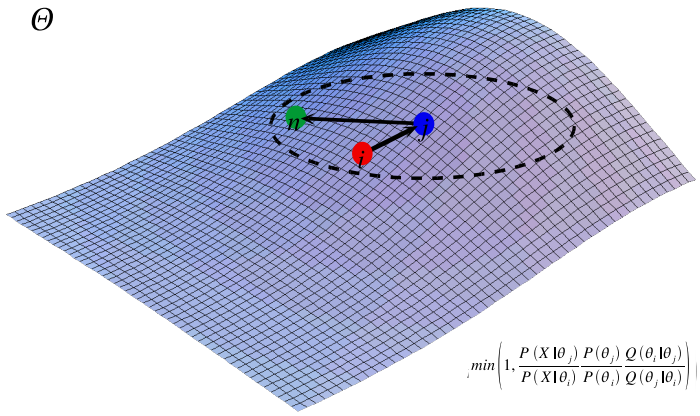


Figura 2.11 Representação esquemática do algoritmo M-H. Os círculos representam os diferentes θ que estão sendo amostrados pela MCMC, iniciando em θ_i (círculo vermelho). As setas representam o sentido da proposição do movimento, sendo que a aceitação ou não do movimento para θ_2 (círculo azul) depende da Equação 2.26.

de maior probabilidade posterior usando o método M-H (Figura 2.12). Para isso, inicialmente são escolhidos aleatoriamente, a partir de alguma distribuição priori, uma árvore qualquer e os valores dos parâmetros do modelo de substituição de estados de caráter, *i.e.*, $\theta_0 = (\psi_0, \varphi_0)$. Dado o estado atual de $\theta_i = (\psi_i, \varphi_i)$, um movimento (ciclo) em Θ envolve duas etapas (Altekar *et al.* [1]):

1) fixa ψ_i e propõe novos parâmetros para φ_j com uma MCMC Q_1 no espaço definido por Φ, os quais serão aceitos ($\varphi_{i+1} = \varphi_j$) ou rejeitados ($\varphi_{i+1} = \varphi_i$) levando em conta o resultado da Equação 2.26;

2) fixa φ_{i+1}, modifica ψ_i de acordo com algum critério (*e.g.*, NNI, TBR) e propõe uma nova ψ_{i+1} de acordo com uma MCMC Q_2 em Ψ, a qual é aceita ou rejeitada levando em conta o resultado da Equação 2.26.

Como visto na Figura 2.11, o algoritmo M-H garante que a MCMC irá para

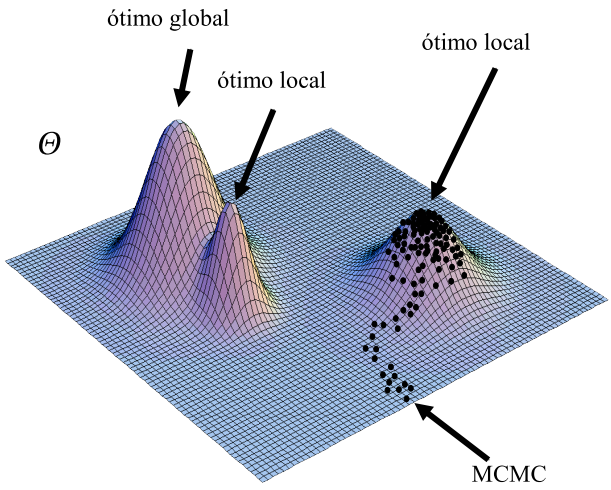


Figura 2.12 Representação esquemática de uma MCMC caminhando em um Θ multimodal. Note que a MCMC fica presa em um ótimo local, o qual fica separado do ótimo global por um grande vale intransponível para o algoritmo de M-H.

uma região de probabilidade posterior elevada. Entretanto, se a distribuição for multimodal (Figura 2.3(b) e 2.12), o método não garante que essa região encontrada é o ótimo global, podendo a MCMC ficar presa em um ótimo local ("ilha", veja Maddison [44]), como mostrado nas Figuras 2.12 e 2.13. Dessa forma, é necessária a utilização de métodos que permitam que a cadeia consiga explorar Θ de forma mais efetiva.

Uma forma de explorar Θ mais intensivamente e, conseqüentemente, aumentar a chance de visitar várias ilhas que podem ocorrer em Θ, é a execução de várias rodadas de MCMC (Figura 2.13), cada uma iniciando a partir de θ_i aleatórios em Θ, como já sugerido por vários autores (*e.g.*, Cowles & Carlin [9], Gelman [24], Larget & Simon [41], Li [43], Mau [47], Rannala & Yang [56]). Essa abordagem tem

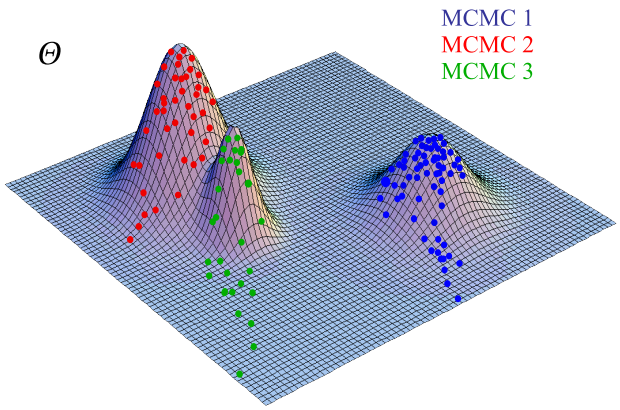


Figura 2.13 Representação esquemática de três MCMC (azul, vermelha e verde) caminhando em um Θ multimodal. Note que as MCMC azul e verde ficaram presas em diferentes ótimos locais e que somente a MCMC vermelha conseguiu chegar na região de probabilidade posterior máxima (ótimo global), embora nem sempre isso aconteça.

a vantagem não só de explorar Θ de forma mais adequada, mas também de fornecer um método de avaliação da mistura das cadeias. Entretanto, ela é extremamente custosa computacionalmente e torna-se rapidamente inviável (veja a Equação 2.2 e a Tabela 2.1), em especial se os dados apresentarem alta complexidade (incongruência). Assim, para uma distribuição posterior multidimensional, como as que ocorrem em problemas filogenéticos, essa abordagem pode ser completamente impraticável.

2.5.3.4 MCMCMC

Como visto anteriormente, a chance de uma MCMC ficar presa em uma determinada ilha em Θ é elevada, dada a natureza de M-H. Uma vez que a MCMC esteja em uma determinada ilha, o método M-H não garante que ela desça, atravesse o vale e tente explorar uma outra região em Θ. Essa limitação de M-H foi resolvida

por Geyer [25] ao propor uma modificação de M-H, a MCMCMC [ou $(MC)^3$].

$(MC)^3$ consiste, de forma simplificada, em executar n MCMC simultaneamente; dentre essas, n-1 são "aquecidas" através da potenciação da probabilidade posterior a um determinado valor β $(0 < \beta < 1)$. Por exemplo, se $P(\theta|X)$ representa a distribuição de probabilidade posterior dos parâmetros filogenéticos, então uma versão aquecida dessa distribuição seria $P(\theta|X)^{\beta}$. Aquecer uma cadeia significa aumentar sua probabilidade de aceitação de novas proposições.

Suponha que uma MCMC esteja em θ_i e que haja uma proposição para θ_j. Como já visto, a probabilidade de aceitação de θ_j em uma MCMC é dada pela Equação 2.26 (veja item 2.5.3.3, acima). Entretanto, para uma MCMC aquecida, a probabilidade de aceitação é dada por

$$r = min\left[1, \left(\frac{P(X|\theta_j)}{P(X|\theta_i)}\frac{P(\theta_j)}{P(\theta_i)}\right)^{\beta}\frac{Q(\theta_i|\theta_j)}{Q(\theta_j|\theta_i)}\right] \tag{2.27}$$

Se $P(\theta_i|X) > P(\theta_j|X)$, elevar cada uma a β faz com que r seja aumentado. Dessa forma, uma cadeia aquecida tende a aceitar mais estados do que uma não-aquecida (ou, simplesmente, "fria"), o que faz com que uma cadeia aquecida consiga cruzar os vales que separam os ótimos locais em Θ e, conseqüentemente, explorar várias ilhas.

Após calcular r, gera-se uma variável aleatória U uniformemente distribuída no intervalo $(0, 1)$. Se U for menor que r, então aceita-se o estado proposto e $\theta_i = \theta_j$. Caso o movimento proposto seja rejeitado, a MCMC continua em θ_i.

Depois que todas as n MCMC realizaram um determinado número de iterações (*e.g.*, um ciclo), duas (*e.g.*, k e l) são escolhidas aleatoriamente e é proposta uma

troca entre os estados atuais dessas cadeias. Essa troca ocorre com probabilidade R

$$R = min\left(1, \frac{P(\theta_l|X)^{\beta_k} P(\theta_k|X)^{\beta_l}}{P(\theta_k|X)^{\beta_k} P(\theta_l|X)^{\theta_l}}\right) \tag{2.28}$$

Após calcular R, gera-se uma variável aleatória U uniformemente distribuída no intervalo $(0, 1)$. Se U for menor que R, então a proposição é aceita e as cadeias trocam de estados entre si.

Uma troca de estados entre duas cadeias permite com que uma cadeia que esteja presa em uma determinada ilha consiga atravessar o vale e explorar outras ilhas e, conseqüentemente, consegue-se obter uma melhor aproximação da distribuição de probabilidade posterior. Usando como exemplo a Figura 2.14, a cadeia fria (representada em azul) está presa em um dos ótimos locais em Θ. Mas, uma troca de estado com a cadeia aquecida (representada em vermelho) permite com que ela saia dessa ilha e salte para outra, nesse caso o ótimo global, em uma única proposição.

Algo importante de ser destacado é que a cadeia aquecida "vê" Θ de forma ligeiramente plana (Figura 2.15), o que facilita com que ela passe pelos vales mais facilmente e eventualmente esteja visitando o ótimo global quando uma proposição com a cadeia fria seja feita.

Portanto, fica claro que o método $(MC)^3$ representa um avanço significativo em direção à aproximação correta da distribuição posterior dos parâmetros filogenéticos. Entretanto, seu custo computacional ainda é diretamente proporcional ao número de cadeias executadas.

Dado o elevado custo computacional de $(MC)^3$, Altekar *et al.* [1] propuseram uma nova versão de $(MC)^3$ baseada em computação paralela [p$(MC)^3$], que reduz

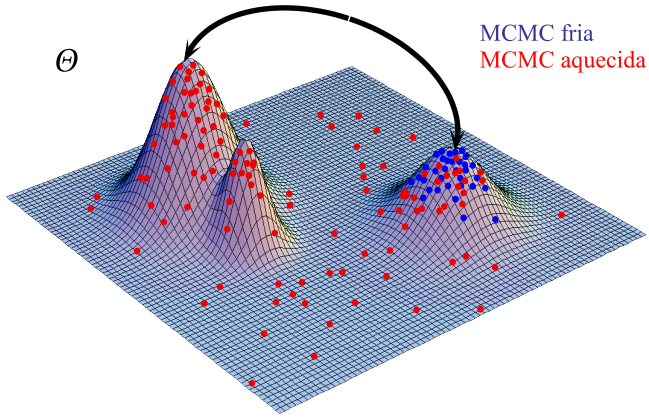


Figura 2.14 Representação esquemática do método de Geyer [25] usando duas MCMC, uma aquecida (vermelha) e uma fria (azul), caminhando em um Θ multimodal. Note que a MCMC azul está presa em um ótimo local quando é feita a proposição de troca de estados e, com isso, passará pelo vale e poderá chegar ao ótimo global.

consideravelmente o tempo de análise (Huelsenbeck *et al.* [35]) e é implementada no MrBayes a partir da versão 3 (Ronquist & Huelsenbeck [60]).

2.5.3.5 Exemplo

Considere a mesma matriz do exemplo descrito no item 2.5.2.1, acima. Apenas por facilitação, também serão consideradas as mesmas árvores.

2.5.3.5.1 Distribuições a priori

2.5.3.5.1.1 Caracteres Uma vez que não se tem informação alguma sobre φ, para quaisquer dois estados $\{i, j\}$, $i \neq j$, será assumido que

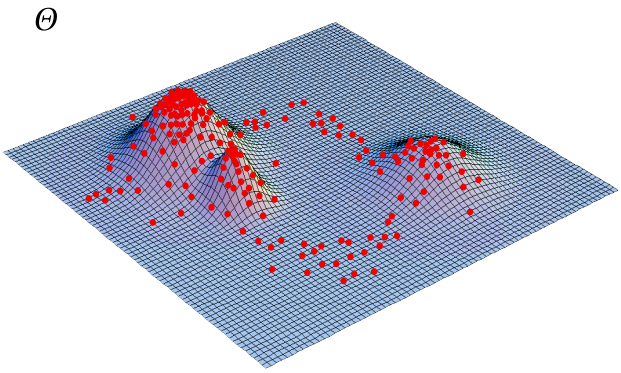


Figura 2.15 Representação esquemática de como uma cadeia aquecida "veria" o mesmo Θ representado na Figura 2.14.

$$\varphi = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix} \tag{2.29}$$

2.5.3.5.1.2 Árvores *Topologia* - a matriz que está sendo usada permite que sejam consideradas até 15 topologias (já que estão sendo consideradas árvores enraizadas, embora isso não seja necessário). Entretanto, para facilitar a comparação, serão analisadas as três topologias apresentadas no exemplo do item 2.5.2.1. Assim sendo, será assumido que só existem essas três topologias. Como também não se dispõe de informações sobre a distribuição a priori das probabilidades das diferentes topologias, será assumido que todas são, a priori, equiprováveis. Então,

$$P(\tau_A) = P(\tau_B) = P(\tau_C) = 0,33 \tag{2.30}$$

Comprimentos de ramos - para os ramos, caso não houvessem informações a priori sobre os seus comprimentos, poderiam ser usados, *e.g.*, uma distribuição

exponencial (como no MrBayes). Entretanto, para comparar os resultados com os obtidos no item 2.5.2.1 serão assumidos como priori os comprimentos de ramos utilizados no item 2.5.2.1 (*i.e.*, os comprimentos de ramos serão fixos). Dessa forma, usando a Equação 2.25 e considerando que as topologias e os comprimentos de ramos são fixos,

$$\begin{aligned} P(\tau_a|X) &\propto P(X|\tau_a,\nu,\varphi)P(\tau_a,\nu,\varphi) \\ &= (1,334421413 \text{ x } 10^{-8}) \text{ x } (0,33) \text{ x} \\ &\quad \{[(0,96)+(0,14)+(1,1)+(1,1)]/(4)\} \text{ x } (0,5) \\ &= (1,334421413 \text{ x } 10^{-8}) \text{ x } (0,33) \text{ x } (0.825) \text{ x } (0,5) \\ &= 0,181648115 \text{ x } 10^{-8} \end{aligned} \tag{2.31}$$

$$P(\tau_b|X) \propto 0,170077738 \text{ x } 10^{-8} \tag{2.32}$$

$$P(\tau_c|X) \propto 0,159807575 \text{ x } 10^{-9} \tag{2.33}$$

Como pode ser visto, a árvore de maior probabilidade posterior é a árvore de máxima verossimilhança. Isso tende a acontecer sempre que as priori forem uniformes e não-informativas, pois o segundo termo da Equação 2.25 será dominado pela máxima verossimilhança. Por outro lado, se fossem usadas diferentes priori de comprimentos de ramos para as diferentes árvores, seria possível a obtenção de resul-

tados totalmente diferentes (veja, *e.g.*, Yang & Rannala [69]), incluindo, até mesmo, uma politomia como a árvore de maior probabilidade posterior (*i.e.*, o paradoxo da politomia, Lewis *et al.* [42]).

É importante ressaltar que o exemplo dado acima é extremamente simples e não reflete fidedignamente uma análise bayesiana computacional, dado que no exemplo não foram usados os algoritmos de busca e de otimização discutidos anteriormente (*e.g.*, MCMCMC).

2.6 Considerações finais

Os diferentes métodos aqui apresentados representam uma parcela extremamente pequena das possibilidades analíticas disponíveis atualmente em filogenética. Mesmo Felsenstein [20] admite não conseguir reunir tudo que existe hoje na área. Como pôde ser notado, todos esses métodos são referentes à busca por árvores ótimas, ficando totalmente negligenciada uma parte importante relativa a dados moleculares, o alinhamento múltiplo, cuja complexidade talvez seja maior que a da própria busca por árvores ótimas.

Dessa forma, este trabalho deveria ser considerado apenas como um resumo simplificado dos métodos básicos em uso corrente.

2.7 Referências

[1] Altekar, G., Dwarkadas, S., Huelsenbeck, J. P. & Ronquist, F. 2004. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* 20: 407–415.

[2] Baxevanis, A. D., Davison, D. B., Page, R. D. M., Petsko, G. A., Stein, L. D. & Stormo, G. D. (eds.) 2003. *Current protocols in bioinformatics.* Wiley Intersciences, New York.

[3] Bayes, T. 1763. An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc. Lond.* 53: 370–418.

[4] Box, G. E. P. & Tiao, G. C. 1992. *Bayesian inference in statistical analysis.* Wiley Classics Library, New York.

[5] Brower, A. V. Z. 2000. Evolution is not a necessary assumption of cladistics. *Cladistics* 16: 143–154.

[6] Buffon, G. 1777. Essai d'arithmétique morale. *Histoire naturelle, générale et particulière,* Supplément 4: 46–123.

[7] Burns, C. D. 1915. Occam's razor. *Mind, New Ser.* 24: 592.

[8] Cavalli-Sfforza, L. L. & Edwards, A. W. F. 1967. Phylogenetic analysis: models and estimation procedures. *Evolution* 21: 550–570.

[9] Cowles, M. K. & Carlin, B. P. 1996. Markov chain Monte Carlo convergence diagnostics: a comparative review. *J. Amer. Statist. Soc.* 91: 883–904.

[10] DeQueiroz, K. & Poe, S. 2001. Philosophy and phylogenetic inference: a comparison of likelihood and parsimony methods in the context of Karl Popper's writings on corroboration. *Syst. Biol.* 50: 305–321.

[11] DeQueiroz, K. & Poe, S. 2003. Failed refutations: further comments on parsimony and likelihood methods and their relationship to Popper's degree of corroboration. *Syst. Biol.* 52: 352–367.

[12] Dunn, M., Terrill, A., Reesink, R. A., G. Foley & Levinson, S. C. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science* 309: 2072–2075.

[13] Edwards, A. W. F. & Cavalli-Sforza, L. L. 1964. Reconstruction of evolutionary trees. *In* Heywood, V. & McNeill, J. (eds.) *Phenetic and phylogenetic classification.* Systematics Association Publ. n. 6, London, 67–76.

[14] Farris, J. S. 1970. Methods for computing Wagner trees. *Syst. Zool.* 19: 83–92.

[15] Farris, J. S. 1983. The logical basis of phylogenetic analysis. *In* Platnick, N. & Funk, V. (eds.) *Advances in cladistics.* vol. 2, Columbia University Press, New York, 7–36.

[16] Felsenstein, J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst. Zool.* 22: 240–249.

[17] Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. *Syst. Zool.* 27: 401–410.

[18] Felsenstein, J. 1978. The number of evolutionary trees. *Syst. Zool.* 27: 27–33.

[19] Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17: 368–376.

[20] Felsenstein, J. 2004. *Inferring phylogenies.* Sinauer Associates, Sunderland.

[21] Fitch, W. M. 1995. Uses of evolutionary trees. *Phil. Trans. R. Soc. Lond. B* 349: 93–102.

[22] Fitch, W. M., Peterson, E. M. & de la Mazat, L. M. 1993. Phylogenetic analysis of the outer-membrane-protein genes of Chlamydiae, and its implications for vaccine development. *Mol. Biol. Evol.* 10: 892–913.

[23] Foulds, L. R. & Graham, R. L. 1982. The Steiner problem in phylogeny is NP-complete. *Advances Appl. Math.* 3: 43–49.

[24] Gelman, A., Carlin, J., Stern, H. & Rubin, D. 2003. *Bayesian data analysis.* Chapman and Hall, London.

[25] Geyer, C. J. 1991. Markov chain Monte Carlo maximum likelihood. *In* Keramidas, E. (ed.) *Computing Science and Statistics: Proceedings of the 23rd Symposium of the Interface.* Interface Foundation, Fairfax Station. 156–163.

[26] Goloboff, P. A. 1996. Methods for fast parsimony analysis. *Cladistics* 12: 199–220.

[27] Goloboff, P. A. 1999. Analysing large data set in reasonable times: solutions for composite optima. *Cladistics* 15: 415–428.

[28] Guttorp, P. 1995. *Stochastic modelling of scientific data.* Chapman and Hall, London.

[29] Hastings, W. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57: 97–109.

[30] Helfenbein, K. G. & DeSalle, R. 2005. Falsifications and corroborations: Karl Popper's influence on systematics. *Mol. Phylogen. Evol.* 35: 271–280.

[31] Hendy, M. D. & Penny, D. 1982. Branch and bound algorithms to determine minimal evolutionary trees. *Math. Biosc.* 60: 133–142.

[32] Hendy, M. D. & Penny, D. 1989. A framework for the study of evolutionary trees. *Syst. Zool.* 38: 297–309.

[33] Hennig, W. 1965. Phylogenetic systematics. *Ann. Rev. Ent.* 10: 97–116.

[34] Hillis, D. M. 2005. Health applications of the tree of life. *In* Cracraft, J. & Bybee, R. (eds.) *Evolutionary sciences and society: educating a new generation.* Biological Sciences Curriculum Study, Colorado Springs, 139–144.

[35] Huelsenbeck, J. P., F. Ronquist, R. N. & Bollback, J. P. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294: 2310–2314.

[36] Hull, D. L. 1988. *Science as process: an evolutionary account of the social and conceptual development of science.* University of Chicago Press, Chicago.

[37] Kluge, A. G. 1997. Testability and the refutation and corroboration of cladistic hypotheses. *Cladistics* 13: 81–96.

[38] Kluge, A. G. 2001. Parsimony with and without scientific justification. *Cladistics* 17: 199–210.

[39] Kluge, A. G. & Grant, T. 2006. From conviction to anti-superfluity: old and new justifications for parsimony in phylogenetic inference. *Cladistics* 22: 276–288.

[40] Kluge, A. G. & Wolf, J. 1993. Cladistics: What's in a word? *Cladistics* 9: 183–199.

[41] Larget, B. & Simon, D. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* 16: 750–759.

[42] Lewis, P. O., Holder, M. T. & Holsinger, K. E. 2005. Polytomies and bayesian phylogenetic inference. *Syst. Biol.* 54: 241–253.

[43] Li, S. 1996. *Phylogenetic tree construction using Markov chain Monte Carlo.* Ph.D. dissertation, Ohio State University, Columbus.

[44] Maddison, D. R. 1991. The discovery and importance of multiple islands of most-parsimonious trees. *Syst. Zool.* 40: 315–328.

[45] Manly, B. F. J. 1998. *Randomization, bootstratp and Monte Carlo methods in biology.* 2 ed. Chapmann and Hall, London.

[46] Marques, A. C. 1997. *Evolução basal nos Metazoa, com ênfase nas relações entre os Cnidaria.* Tese de doutorado, Universidade de São Paulo, São Paulo.

[47] Mau, B. 1996. *Bayesian phylogenetic inference via Markov chain Monte Carlo methods.* Ph.D. dissertation, University of Wisconsin, Madison.

[48] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. 1953. Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21: 1087–1092.

[49] Metzker, M. L., Mindell, D. P., Liu, X.-M., Ptak, R. G., Gibbs, R. A. & Hillis, D. M. 2002. Molecular evidence of HIV-1 transmission in a criminal case. *Proc. Nat. Acad. Science* 99: 14292–14297.

[50] Morgan, J. A. T., Dejong, R. J., Adeoye, G. O., Ansa, E. D. O., Barbosa, C. S., Brémond, P., Cesari, I. M., Charbonnel, N., Corrêa, L. R., Coulibaly, G., D'andrea, P. S., de Souza, C. P., Doenhoff, M. J., File, S., Idris, M. A., Incani, R. N., Jarne, P., Karanja, D. M. S., Kazibwe, F., Kpikpi, J., Lwambo, N. J. S., Mabaye, A., Magalhães, L. A., Makundi, A., Moné, H., Mouahid, G., Muchemi, G. M., Mungai, B. N., Maséne, M., Southgate, V., Tchuenté, L. A. T., Théron, A., Yousif, F., M.Zanotti-Magalhães, E., Mkoji, G. M. & Loker, E. S. 2005. Origin and diversification of the human parasite *Schistosoma mansonii*. *Mol. Ecol.* 14: 3889–3902.

[51] Newton, I. 1726. *Philosophiae naturais principia mathematica.* Guil. and Joh. Innys, London.

[52] Papadimitriou, C. & Steiglitz, K. 1982. *Combinatorial optimization: algorithms and complexity.* Prentice Hall, New York.

[53] Phipps, J. B. 1976. Dendrogram topology: capacity and retrieval. *Canad. J. Bot.* 54: 679–685.

[54] Popper, K. R. 1965. *Conjectures and refutations: the growth of scientific knowledge.* Harper Torchbooks, New York.

[55] Popper, K. R. 2002. *The logic of scientific discovery.* Routledge Classics, New York.

[56] Rannala, B. & Yang, Z. 1996. Probability distribution of molecular evolutionary trees: a new method for phylogenetic inference. *J. Mol. Evol.* 43: 304–311.

[57] Rieppel, O. 2003. Popper and systematics. *Sys. Biol.* 52: 271–280.

[58] Ronquist, F. 1998. Fast Fitch-parsimony algorithms for large data sets. *Cladistics* 14: 387–400.

[59] Ronquist, F. 2004. Bayesian inference of character evolution. *TREE* 19: 475–481.

[60] Ronquist, F. & Huelsenbeck, J. P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19: 1572–1574.

[61] Siddall, M. Disponível em http://research.amnh.org/~siddall/methods/day3.html. Acesso em: 27 maio 2007 .

[62] Swofford, D. L. 2002. *PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods), version 4.* Sinauer Associates, Sunderland.

[63] Swofford, D. L. & Begle, D. P. 1991. *PAUP. Phylogenetic Analysis Using Parsimony , version 3.1, user's manual.* Sinauer Associates, Illinois.

[64] Swofford, D. L., Waddell, P. J., Huelsenbeck, J. P., Foster, P. G., Lewis, P. O. & Rogers, J. S. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst. Biol.* 50: 525–539.

[65] Thorburn, W. M. 1918. The myth of Occam's razor. *Mind, New Ser.* 27: 345–353.

[66] Tierney, L. 1994. Markov chains for exploring posterior distributions. *Annals Stat.* 22: 1701–1762.

[67] Wheeler, W. C. 1995. Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data. *Syst. Biol.* 44: 321–331.

[68] Wiley, E. O. 1981. *Phylogenetics: the theory and pratice of phylogenetic systematics.* John Wiley and Sons, New York.

[69] Yang, Z. & Rannala, B. 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. *Syst. Biol.* 54: 455–470.

Parte II

Filogenia de Pilocarpinae

Capítulo 3

Filogenia de Pilocarpinae (Rutaceae) e Diagnose de MCMC em Estudos Filogenéticos

3.1 Abstract

We investigated generic-level phylogenetic relationships within the Pilocarpinae (Rutaceae), a neotropical subtribe with four genera (*Esenbeckia*, *Metrodorea*, *Pilocarpus* and *Raulinoa*), and among them and two closely related genera (*Balfourodendron* and *Helietta*) using the Bayesian MCMC method. Additionaly, we present a comparison among the methods currently used in Bayesian phylogenetics for burn-in detection and some of their implications. The relationships were analysed using nucleotide sequences of the spacer *trn*G-S of the cpDNA, the internal transcribed spacers (ITS1 and 2), and the 5.8S gene of the nrDNA of 30 species (4 outgroups). For the genera with more than one species included, the results support the monophyly of *Balfourodendron*, *Metrodorea* and *Pilocarpus*, although the subtribe itsef is not monophyletic. *Esenbeckia*, *Metrodorea* and *Raulinoa* are nested within a "Paniculate" clade (together with *Balfourodendron* and *Helietta*), which has maximmum posterior probability and *Pilocarpus* (also with maximmum support) as sister group. These two clades, "Paniculate" and *Pilocarpus*, can easily be recognized by their inflorescences, namely panicules and racemes, respectively. The "Paniculate" clade has a non-resolved base of the form (*E. grandiflora*, (other *Esenbeckia*), (*Metrodorea*, *Raulinoa*), *Helietta*, *Balfourodendron*), indicating that only *Esenbeckia* lacks support for its status as a (non-) monophyletic group. The segregation of *Pilocarpus* from the other genera gives stronger support for earlier suggestions on the need of recircunscription of Pilocarpinae as to include only *Pilocarpus*, what is assumed here. Moreover, a new subtribe is created to accomodate the "Paniculate" clade. With regard to the methods used to detect burn-in, the three approaches led to different results, and their implications may include topological changes and, accordingly, strong side effects on the conclusions based on the trees.

3.2 Resumo

Neste estudo são analisadas as relações filogenéticas dos gêneros da subtribo Pilocarpinae (Rutaceae), um grupo de plantas neotropicais formado por quatro gêneros (*Esenbeckia*, *Metrodorea*, *Pilocarpus* e *Raulinoa*), e gêneros próximos (*Balfourodendron* e *Helietta*) usando o método bayesiano com MCMC. Também é apresentada uma comparação entre os diferentes métodos de detecção de "burn-in" em análises filogenéticas e algumas de suas implicações. As relações filogenéticas foram estudadas usando seqüências nucleotídicas do espaçador *trn*G-S do DNA plastidial, dos espaçadores transcritos internos (ITS1 e 2) e do gene 5.8S do DNA nuclear de 30 espécies (4 grupos-externos). Os resultados sustentam que, dos gêneros com mais de uma espécie amostrada, *Balfourodendron*, *Metrodorea* e *Pilocarpus* são monofiléticos, mas a subtribo não. *Esenbeckia*, *Metrodorea* e *Raulinoa* emergem junto com *Balfourodendron* e *Helietta* em um clado com suporte máximo (clado "Paniculado"), o qual tem *Pilocarpus* como grupo-irmão (também com suporte máximo). Essa bifurcação é caracterizada morfologicamente pela posse de inflorescências paniculadas e racemos, respectivamente. Na base do clado "Paniculado" existe uma politomia na forma (*E. grandiflora*,(outras *Esenbeckia*),(*Metrodorea*,*Raulinoa*),*Helietta*,*Balfourodendron*), mostrando que dos gêneros de Pilocarpinae com mais de uma espécie, apenas *Esenbeckia* permanece com status duvidoso quanto à sua (não-) monofilia. A separação de *Pilocarpus* dos outros gêneros corrobora sugestões anteriores de reduzir a circunscrição da subtribo para *Pilocarpus*, o que é feito neste estudo. Adicionalmente, uma nova subtribo é criada para acomodar o clado "Paniculado". Por outro lado, os três métodos de detecção de "burn-in" apresentaram resultados diferentes, cujas implicações podem envolver alterações topológicas e, conseqüentemente, as inferências feitas sobre as árvores obtidas.

3.3 Introdução

Pilocarpinae é uma subtribo com distribuição neotropical, ocorrendo do México à Argentina e Paraguai (Kaastra [27], Figura 4, p. 23). Após a revisão apresentada por Kaastra ([27]), vários outros trabalhos já foram publicados sobre o grupo (especialmente com descrição de novos táxons, *e.g.*, Pirani [39], Skorupa [45], Skorupa & Pirani [46]) e, atualmente, a subtribo compreende 51 espécies: *Esenbeckia* com 28, *Metrodorea* com 5, *Pilocarpus* com 17 e *Raulinoa* com 1.

Esenbeckia e *Pilocarpus* têm ampla ocorrência nos neotrópicos, tendo sido encontrados representantes bem distribuídos em praticamente todas as formações vegetacionais, desde áreas abertas e de vegetação baixa, como campos, cerrados *s.s.*, campos rupestres, a áreas com vegetação mais exuberante e densa como a floresta amazônica e a mata atlântica. Nas formações florestais, também ocorrem de forma bastante diversificada, como próximo a igarapés e bem próximo ao nível do mar a topo de pequenas montanhas ou morros. Nesse último ambiente, geralmente encontram-se associados a córregos e cachoeiras. Apesar da ampla distribuição desses gêneros, algumas de suas espécies possuem distribuição bastante restrita, como é o caso de *E. decidua* Pirani (Norte de Minas Gerais), *E. scrotiformis* Kaastra (conhecida de duas localidades no Acre e uma em Rondônia), entre outras. Por exemplo, *P. jaborandi* Holm., outrora com distribuição ampla no nordeste do Brasil, hoje tem distribução conhecida restrita a apenas uma localidade, provavelmente devido à ação antrópica. *P. demerarae* Sandw. e *P. peruvianus* (Macbr.) Kaastra são conhecidos apenas da localidade-tipo e *P. trifoliolatus* Skorupa tem distribuiçao incerta[1].

[1] A espécie foi descrita com base em material doado, supostamente oriundo da região da Serra dos Carajás, Pará, Brasil.

Metrodorea, por outro lado, possui distribuição principalmente brasileira, com uma espécie (*M. flavida* Mart.) ocorrendo nos países com vegetação amazônica (ou ecótono amazônia/cerrado) que fazem fronteira com o Brasil.

Raulinoa é conhecida apenas da região de Itajaí/Ibirama (Santa Catarina, Brasil) e sua ocorrência conhecida está restrita à margem direita do rio Itajaí-açu. Adicionalmente, essa espécie é considerada símbolo da Prefeitura de Itajaí.

Em geral, os representantes de Pilocarpinae têm porte arbustivo, mas tamém existem espécies arbóreas (Figura 3.1), como *E. densiflora* e *M. flavida*, podendo alcaçar vários metros de altura e mais de 20cm de diâmetro.

Entre as espécies de Pilocarpinae há várias com importância econômica e com uso amplo na medicina popular, especialmente em *Esenbeckia* e *Pilocarpus*. Dentre as *Esenbeckia*, *E. leiocarpa* Engl. ("guarantã") se destaca por sua madeira de propriedades mecânicas singulares e utilização na fabricação de móveis. *E. febrifuga* (St.-Hil.) A. Juss. *ex* Mart., por sua vez, é bastante utilizada como febrífugo, principalmente no Brasil e Paraguai. Entre as espécies de *Pilocarpus*, o maior destaque é dado àquelas a partir das quais se extrai o alcalóide pilocarpina, bastante valorizado no mercado internacional e utilizado no tratamento de glaucoma (Kaastra [27], Pinheiro [36]). Esse alto valor associado à pilocarpina tem incentivado a atividade extrativista das folhas e, inclusive, sido responsabilizado pela inclusão de espécies de *Pilocarpus* em listas de plantas ameaçadas de extinção (Pinheiro [37]).

Pilocarpinae (Rutaceae) foi estabelecida por Engler em 1874 (como "Pilocarpeae", Engler [10]), incluindo os gêneros *Esenbeckia* Kunth, *Leptothyrsa* Hook. f., *Metrodorea* St.-Hil. e *Pilocarpus* Vahl (o gênero *Leptothyrsa* foi transferido por

Figura 3.1 Hábitos de representantes de Pilocarpinae. (a) *E. densiflora*, (b) *E.* sp. nov, (c) *M. flavida*, (d) *M. mollis*, (e) *P. sulcatus* e (f) *R. echinata*. (Fotos por R.G. Udulutsch)

ele mesmo para Cuspariinae em 1896), e junto com Cuspariinae Engl. formava a tribo Cusparieae DC. Além da freqüente simpetalia em Cuspariinae e dialipetalia em Pilocarpinae, uma das diferenças mais marcantes entre as duas subtribos é a forma do botão, arredondado em Pilocarpinae e alongado em Cuspariinae, o que é um resultado da forma, e principalmente do comprimento, das anteras e pétalas em cada subtribo (Kaastra [27]).

Mais recentemente, e após realizar detalhada análise histórica sobre o nome *Cusparia* (gênero-tipo de *Cuspariinae*), Elias ([9]) revelou que se tratava de um nome ilegítimo. Portanto, Cuspariinae e Cusparieae também são ilegítimos. Apesar disso, Elias ([9]) não sugeriu nenhuma alteração para os nomes de subtribo e tribo. Posteriormente, dado que *Cusparia* foi considerado um nome não validamente publicado, Kallunki & Pirani ([28]) sinonimizaram Cuspariinae em Galipeinae Kallunki e Cusparieae em Galipeeae Kallunki. Portanto, atualmente, as Pilocarpinae fazem parte das Galipeeae.

Questionamentos sobre o status de Pilocarpinae enquanto grupo e a inclusão ou exclusão de gêneros da subtribo não são novos. Como comentado anteriormente, o próprio Engler ([11]) excluiu um gênero (*Leptothyrsa*) da subtribo alguns anos após ele mesmo ter estabelecido o grupo. Por outro lado, outros autores também têm questionado o agrupamento de *Esenbeckia*, *Metrodorea*, *Pilocarpus* e *Raulinoa* numa mesma subtribo. Por exemplo, Kaastra ([27]) discute as diferenças em relação ao tipo de inflorescências nos quatro gêneros (Figura 3.2) e a ocorrência de imidazóis apenas em *Pilocarpus* e comenta uma possível redução de Pilocapinae compreendendo apenas seu gênero-tipo (*Pilocarpus*), excluindo os outros três gêneros. Entretanto,

apesar de favorável a essa redução, Kaastra ([27]) manteve os quatro gêneros, por achar a idéia ainda prematura.

Apesar da manutenção dos quatro gêneros em Pilocarpinae, diversos autores (inclusive o próprio Kaastra [27], p. 20) vêem como intrigante a semelhança morfológica (principalmente vegetativa e floral) de *Esenbeckia* com outros gêneros, como *Helietta* Tul., que por ter fruto alado era incluída na subtribo Pteleinae (tribo Toddalieae, subfamília Toddalioideae). Essa semelhança com outros gêneros, especialmente *Helietta* e *Balfourodendron* Mello *ex* Oliv., também de Pteleinae, causou (e ainda causa) vários problemas, tanto nomenclaturais (como pode ser visto na história taxonômica do grupo) quanto de identificação (como encontrado em herbários nacionais e internacionais). Exemplo dessa problemática causada pela semelhança morfológica entre *Esenbeckia* e *Balfourodendron*/*Helietta*, é o caso da espécie-tipo de *Balfourodendron*. *B. riedelianum* Engl. (Engl.) foi descrita, como *E. riedeliana* Engl. (Engler [10]) e, posteriormente, transferida para *Balfourodendron* pelo próprio autor (Engler [11]). Além disso, na mesma obra, uma espécie de *Helietta* (*H. multiflora* Engl.) foi descrita com base em materiais atualmente considerados como *B. riedelianum*. Para uma discussão mais detalhada da nomenclatura das espécies de *Balfourodendron* e *Helietta*, sugere-se o trabalho de Pirani ([38]).

Mais recentemente, o estudo de Groppo ([19]) sugeriu uma possível maior proximidade filogenética entre *Balfourodendron*, *Helietta*, *Esenbeckia* e *Metrodorea*, na forma (*Balfourodendron*,*Helietta*,(*Esenbeckia*,*Metrodorea*)), do que dos dois últimos gêneros com *Pilocarpus* (Groppo [19], Figura 3, p. 78).

Por outro lado, em estudos filogenéticos recentes, o uso da abordagem baye-

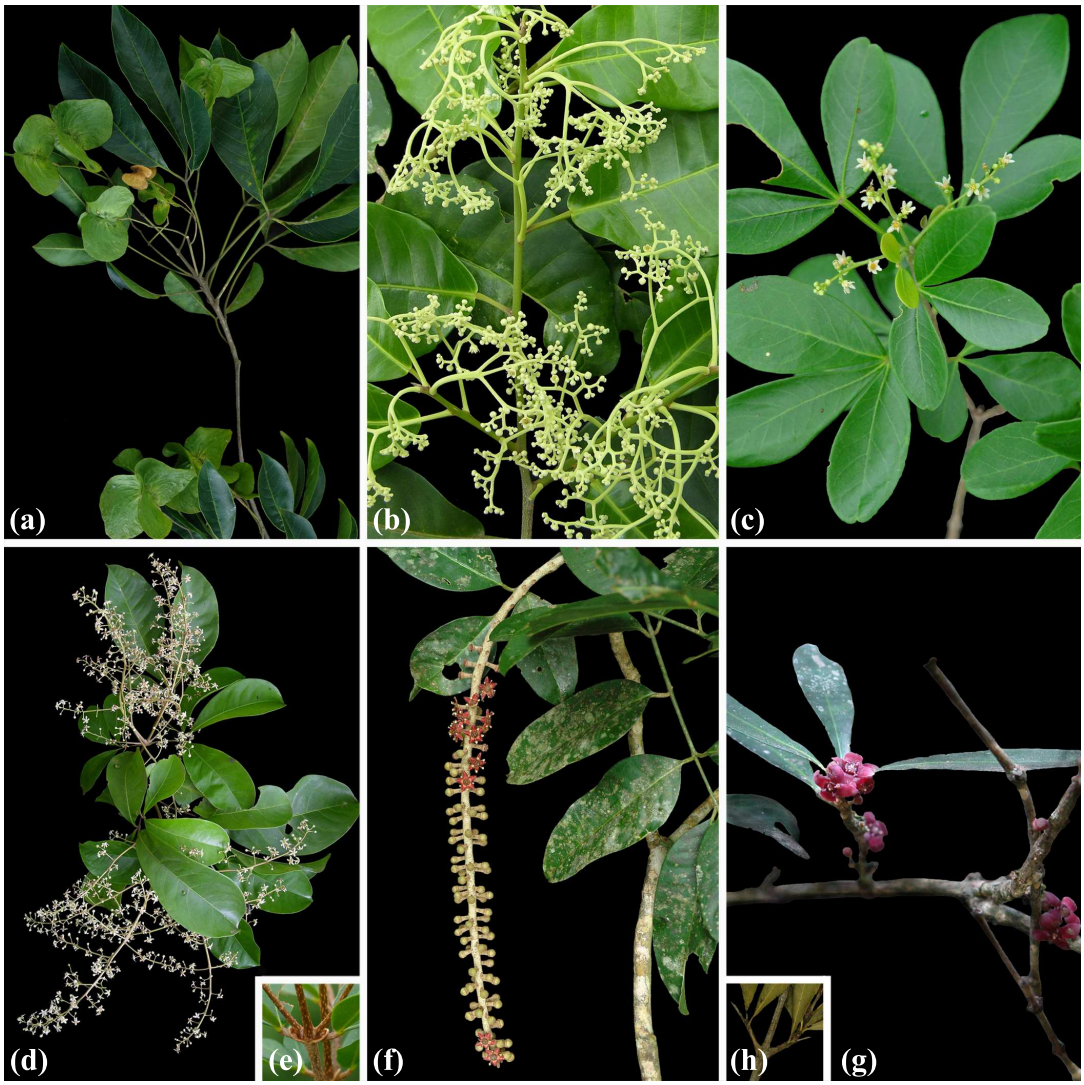


Figura 3.2 Padrões de de inflorescências em Pilocarpinae e gêneros próximos. (a) *B. riedelianum*, (b) *E. cowanii*, (c) *H. puberula*, (d) *M. flavida*, (f) *P. grandiflorus* e (g) *R. echinata*. (e) e (h) detalhes mostrando características diagnósticas de *Metrodorea* (bainha) e *Raulinoa* (espinhos), respectivamente. (Fotos por R.G. Udulutsch)

siana usando métodos MCMC (*e.g.*, Huelsenbeck *et al.* [23]), em especial o de Metropolis-Hastings (Metropolis *et al.* [33], Hastings [21]), tem sido bastante comum. Esses métodos têm sido utilizados não apenas na busca de árvores ótimas, mas também na construção de alinhamentos (*e.g.*, Löytynoja & Milinkovitch [30]), ou na inferência de ambos de forma concomitante (Redelings & Suchard [42], Suchard & Redelings [48]).

Dado que nesse tipo de análise se objetiva alcançar a distribuiçao de equilíbrio posterior, existem métodos que tentam avaliar se as MCMC chegaram a essa distribuição alvo (para revisões desses métodos, sugere-se os trabalhos de Brooks & Gelman [5], Brooks & Roberts [6] e Cowles & Carlin [7]). Entretanto, os métodos mais usados em estudos específicos sobre a convergência de MCMC ainda são ignorados por parte considerável dos sistematas, como pode ser observado na literatura atual.

Nos estudos publicados até o momento, em geral, a convergência das cadeias é verificada plotando-se as lnL contra as iterações (a partir daqui chamado de "método tradicional"). Se existir um platô em alguma região do gráfico, o(s) autor(es) do estudo assume(m) que as MCMC convergiram. Entretanto, como já discutido por vários autores (*e.g.*, Hillis *et al.* [22]), esse pode não ser o caso, dado que o platô não necessariamente significa convergência. Além disso, recentemente, Hillis *et al.* ([22]) propuseram o uso de um novo método para detectar a não-convergência das cadeias usando a distância de Robinson-Foulds ponderada (Robinson & Foulds [43]) e o escalonamento multidimensional (MDS). Esse método, segundo os autores, seria superior ao "método tradicional" por permitir que seja avaliada a distribuição posterior conjunta e não apenas a distribuição marginal dos parâmetros individuais.

Nesse sentido, os objetivos deste estudo vão em duas direções: de um lado, serão avaliados o status de Pilocarpinae enquanto grupo, o arranjo dos gêneros dessa subtribo e suas possíveis relações com os gêneros morfologicamente semelhantes (*Balfourodendron* e *Helietta*) e suas implicações taxonômicas no nível de subtribo; e do outro lado, uma vez que será necessário diagnosticar a convergência das MCMC, também será feita uma comparação entre as diferentes formas de avaliar a convergência das cadeias usadas na filogenética atual e discutidas algumas de suas implicações.

3.4 Material e métodos

3.4.1 Seleção de terminais

Os quatro gêneros de Pilocarpinae e dois dos quatro de Pteleinae foram incluídos neste estudo. O arranjo taxonômico dos gêneros em nível de subfamília, tribo e subtribo (de acordo com Engler [12], e modificações posteriores de Kallunki & Pirani [28]), assim como o número total de espécies (apenas para o grupo-interno) e número de espécies incluídas é mostrado na Tabela 3.1.

Como grupos-externos, foram selecionados representantes dos gêneros *Galipea*, *Neoraputia*, *Rauia* e *Zanthoxylum*. Esses grupos-externos foram selecionados com base em estudos anteriores de taxonomia "clássica" (*e.g.*, Engler [12]) e no de Groppo [19]. Para os grupos-externos, foi incluída apenas uma espécie de cada gênero.

A Tabela 3.2 apresenta informações sobre número de coletor, coleção e localidade de coleta dos "vouchers" para todos os terminais (informações completas sobre os pontos de coleta de todos os "vouchers" são apresentadas no Apêndice F).

Tabela 3.1 Arranjo taxonômico dos terminais de acordo com Engler ([12]). Galipeeae e Galipeinae estão de acordo com Kallunki & Pirani ([28])

Subfamília	Tribo	Subtribo	Gênero - #spp/spp. incluídas	ITS/*trn*G-S	Grupo
Rutoideae	Galipeeae	Galipeinae	*Galipea*	1/1	externo
			Neoraputia	1/1	externo
			Rauia	1/1	externo
		Pilocarpinae	*Esenbeckia* – 28/9	12/9	interno
			Metrodorea – 5/2	5/2	interno
			Pilocarpus – 17/11	11/11	interno
			Raulinoa – 1/1	1/1	interno
	Zanthoxyleae	Evodiinae	*Zanthoxylum*	1/1	externo
Toddalioideae	Toddalieae	Pteleinae	*Balfourodendron* – 2	2/2	interno
			Helietta – 8/1	2/1	interno

Tabela 3.2 Informações sobre os "vouchers" dos terminais.

Gênero	Espécie	Coletor e número	Coleção	Local de coleta
Balfoudendron	*B. molle* (Miq.) Pirani	P. Dias 213	SPF	Brasil, BA, Rio de Contas
	B. riedelianum (Engl.) Engl.	P. Dias 217	SPF	Brasil, SP, Piracicaba
Esenbeckia	*E. almawillia* Kaastra	P. Dias 233	SPF	Brasil, AC, Xapuri
	E. cowanii Kaastra	P. Dias 227	SPF	Brasil, MT, V. B. Santíssima Trindade
	E. decidua Pirani	P. Dias 202	SPF	Brasil, MG, Mato Verde
	E. grandiflora Engl.	P. Dias 273	SPF	Brasil, SC, Florianópolis
	E. hieronymi Engl.	P. Dias 271	SPF	Brasil, PR, Paranaguá
	E. oligantha Kaastra	P. Dias 310	SPF	Brasil, TO, Mateiros
	E. pumila Pohl	P. Dias 225	SPF	Brasil, MT, Chapada dos Guimarães
	E. scrotiformis Kaastra	P. Dias 298	SPF	Brasil, RO, Ouro Preto d'Oeste
	Esenbeckia sp. nv.	P. Dias 280	SPF	Brasil, MS, Ladário

...

(*Continua na próxima página*)

Tabela 3.2 Informações sobre os "vouchers" dos terminais. *(Continuada)*

Galipea	*G. trifoliata* Aubl.	P. Dias 230	SPF	Brasil, RO, Presidente Médice
Helietta	*H. puberula* R.E. Fr.	P. Dias 216	SPF	Brasil, MS, Corumbá
Metrodorea	*M. nigra* St.-Hil.	P. Dias 264	SPF	Brasil, SP, Rio Claro
	M. stipularis Mart.	P. Dias 263	SPF	Brasil, SP, Rio Claro
Neoraputia	*N. paraensis* (Ducke) Emmerich	P. Dias 245	SPF	Brasil, MA, Buriticupu
Pilocarpus	*P. alatus* Joseph ex Skorupa	P. Dias 247	SPF	Brasil, MA, Buriticupu
	P. giganteus Engl.	P. Dias 337	SPF	Brasil, ES, Linhares
	P. grandiflorus Engl.	P. Dias 339	SPF	Brasil, ES, Linhares
	P. jaborandi Holm.	P. Dias 252	SPF	Brasil[1]
	P. microphyllus Stapf ex Wardl.	P. Dias 235	SPF	Brasil[1]
	P. pauciflorus St.-Hil.	P. Dias 218	SPF	Brasil, SP, Piracicaba
	P. pennatifolius Holm.	P. Dias 215	SPF	Brasil, SP, Piracicaba

...

(*Continua na próxima página*)

[1]Informação retida por estar ameaçada de extinção.

Tabela 3.2 Informações sobre os "vouchers" dos terminais. *(Continuada)*

	P. peruvianus (Macbr.) Kaastra	P. Dias 291	SPF	Brasil, RO, Jaru
	P. spicatus St.-Hil.	P. Dias 325	SPF	Brasil, BA, Caetité
	P. sulcatus Skorupa	P. Dias 322	SPF	Brasil, BA, Maniaçu
	P. trachylophus Holm.	P. Dias 323	SPF	Brasil, BA, Maniaçu
Rauia	*R. resinosa* Nees & Mart.	P. Dias 243	SPF	Brasil, MA, Buriticupu
Raulinoa	*R. echinata* Cowan	P. Dias 257	SPF	Brasil, SC, Ibirama
Zanthoxylum	*Z. rhoifolium* Lam.	P. Dias 232	SPF	Brasil, RO, Ariquemes

3.4.2 Extração, amplificação e seqüenciamento de DNA

A obtenção de DNA foi feita a partir de amostras de folhas conservadas em sílica. O DNA total foi extraído com o uso do kit DNeasy (QiaGen), com pequenas modificações nas recomendações do fabricante (Sewell, com. pess.).

Foi feita amplificação de regiões do DNA nuclear (ITS1, ITS2 e gene 5.8S, daqui em diante referidas genericamente como ITS) e do DNA platidial (*trn*G-S). Como mostrado na Tabela 3.1, o número de seqüências obtidas para ITS foi maior que para *trn*G-S, portanto as análises ficaram restritas aos terminais com sequências para as duas regiões.

O ITS foi amplificado usando os iniciadores TATGCTTAAAYTCAGCGGGT e CCTTATCATTTAGAGGAAGGAG e de acordo com as condições descritas em Stanford *et al.* [47].

Para as seqüências de *trn*G-S, foram utilizados os iniciadores descritos por Hamilton ([20]) sob as mesmas condições de PCR, exceto que a cada ciclo o tempo de extensão foi aumentado em 3 segundos. Para todas as análises foram incluídos controles negativos.

O seqüenciamento foi realizado em seqënciador automático ABI 377 na Universidade de Washington em Seattle. Todas as seqüências foram obtidas usando os mesmo iniciadores usados nas amplificações.

3.4.3 Análise e qualidade das seqüências

Todas as seqüências (diretas e reversas) foram checadas no GenBank quanto à sua similaridade para verificar a possibilidade de seqüenciamento de material conta-

minante. Essa checagem foi feita de forma automatizada com o script "CheckGB.pl" (veja o Apêndice A), o qual utiliza o programa BLAST (Altschul *et al* [1]).

A qualidade de todas as seqüências foi verificada com o programa Phred (Ewing & Green [13], Ewing *et al.* [14]). O alinhamento das seqüências diretas e reversas foi feito com o programa Phrap (Green não-publicado), posteriormente visualizados com o programa Consed (Gordon *et al.* [18]) e somente a região Phred20 foi selecionada. Dado que é possível ocorrerem bases de qualidade inferior dentro da região Phred20, essas bases foram trocadas por "N" e as seqüências foram transformadas em formato FASTA com auxílio do script "Phred20.pl" (veja o Apêndice E). Após a obtenção das seqüências consenso para cada terminal, essas seqüências foram novamente checadas no GenBank quanto à sua identidade, como descrito acima.

3.4.4 Alinhamento e matriz

O alinhamento das seqüências foi feito com o programa ProAlign 0.5 (Löytynoja & Milinkovich [30]), usando 1000 réplicas e os parâmetros listados no Apêndice H. O alinhamento com maior probabilidade posterior foi selecionado e os gaps iniciais e finais ("trailing gaps") foram excluídos.

Apesar da recomendação de alguns autores (*e.g.*, Morrison [34]) e do uso comum de "ajustes manuais" ao alinhamento, nenhum "ajuste" foi feito. Dada a multidimensionalidade (geralmente maior que o da própria inferência filogenética, *e.g.*, Felsenstein [15]) e o número de parâmetros a serem considerados em um alinhamento múltiplo, "ajustes manuais" só tendem a enviesar o alinhamento para as presuposições pessoais do próprio autor.

Cada região foi considerada como uma partição distinta e, conseqüentemente, todos os parâmetros do modelo de substituição foram estimados de forma independente para cada uma delas (veja o Apêndice D para detalhes).

Como ponto de partida, foi utilizado um modelo de substituição equivalente ao GTR + I + Γ. Entretanto, todos os parâmetros foram tratados de maneira que ficassem livres e seus valores fossem determinados pelos próprios dados, ficando a possibilidade de o modelo ser literalmente substituído ao longo da análise, caso o inicial se mostrasse inadequado aos dados. Todos os caracteres foram tratados como não-ordenados.

3.4.5 Buscas por árvores e suporte de ramos

Todas as buscas foram feitas com o programa MrBayes 3.1.2 (Ronquist & Huelsenbeck [44]). Foram realizadas quatro rodadas simultâneas e independentes, cada uma com três cadeias aquecidas e uma fria, totalizando 16 MCMC simultâneas, durante 15 milhões de iterações, com espaçamento de 1000. A probabilidade posterior (PP) dos ramos foi assumida como suporte. As árvores de consenso do MrBayes foram transcritas para o formato NEXUS padrão com o script "ConToNex.pl" (veja o Apêndice B).

Adicionalmente, foi realizada uma análise com *Ptelea*, o gênero-tipo de Pteleinae, baseada no ITS1 e usando *Murraya* (subfamília Aurantioideae) e *Zanthoxylum* como grupos-externos (seqüências de *Ptelea trifoliata* L., *Murraya paniculata* (L.) Jacq. e *Murraya koenigii* (L.) Spreng. obtidas do GenBank, veja o Apêndice G). Neste caso, foram usadas duas seqüências de *Ptelea* devido o ITS1 apresentar varia-

ção[2] para a espécie utilizada e as análises foram rodadas por 10 milhões de iterações.

3.4.6 Diagnóstico de convergência e intervalos HPD

Os arquivos com os valores dos parâmetros estimados pelo MrBayes foram carregados no programa R (R Development Core Team [41]) e a convergência das cadeias das diferentes rodadas, para cada um dos parâmetros, foi diagnosticada pelo método de Gelman & Rubin ([16]) corrigido por Brooks & Gelman ([5], CSRF). A mistura para cada uma das cadeias foi verificada pela autocorrelação dentro da própria cadeia. Todas as medidas de diagnóstico foram executadas com o pacote CODA (Plumer *et al.* [40]) no programa R ([41]). Adicionalmente, foram feitas comparações entre os métodos usados e as abordagens mais comuns em filogenética para detecção de convergência[3], como a comparação das lnL por iteração ao longo da(s) cadeia(s), e o método proposto por Hillis *et al.* ([22]) para visualização das cadeias com escalonamento multidimensional (MDS).

Após a exclusão do "burn-in" e com base nos valores de lnL das iterações, foram construídos intervalos HPD de 95% para cada uma das rodadas. Então, somente as iterações desses intervalos HPD foram consideradas para as análises posteriores. Os intervalos HPD foram construídos com o pacote hdrcde (Hyndman & Einbeck [24]) no programa R e os quantis estimados foram usados pelo script "HPD-Trees.pl" (veja o Apêndice C) para filtrar as árvores de cada rodada com base em sua respectiva lnL. Todas as árvores dos intervalos HPD de todas as rodadas foram utilizadas para calcular o consenso de maioria com o próprio MrBayes.

[2]Diferença de um nucleotídeo entre as seqüências utilizadas.
[3]Detecção de não-convergência, na verdade.

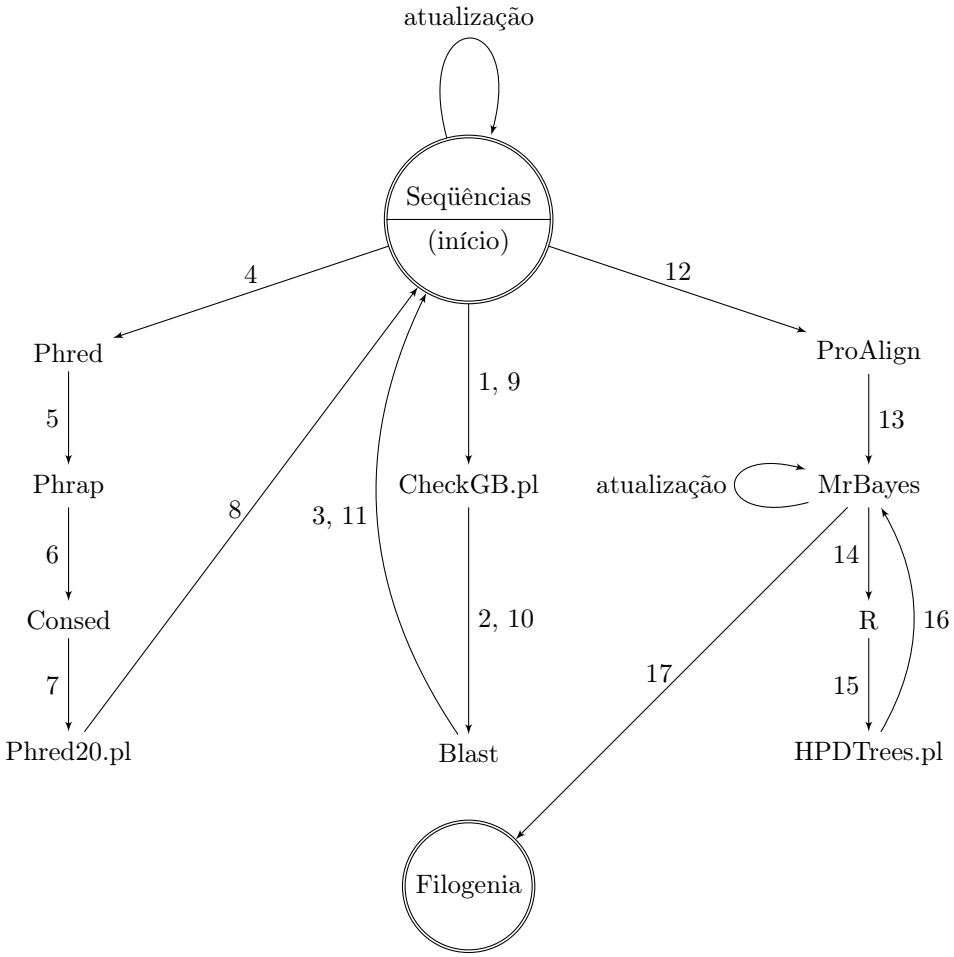


Figura 3.3 Fluxograma dos métodos utilizados neste estudo. Números representam ordem de execução.

A Figura 3.3 mostra de forma simplificada o fluxograma dos métodos e programas usados neste estudo.

3.5 Resultados e discussão

3.5.1 Amplificação, seqüenciamento e qualidade das seqüências

Para todos os terminais, tanto ITS como *trn*G-S foram facilmente amplificados. Apesar de o ITS ser comumente encontrado com bandas múltiplas em vários grupos de angiospermas (veja, *e.g.*, Alvarez & Wendel [2] para uma revisão), para os grupos em estudo, e usando os primers citados anteriormente, a checagem dos produtos amplificados em gel de agarose a 1% revelou banda única para todos os terminais. Por outro lado, o seqüenciamento do espaçador *trn*G-S de para alguns terminais não foi bem sucedido devido à baixa concetração do produto de PCR purificado.

Na primeira busca no GenBank, com exceção da seqüência reversa de *trn*G-S para *E. irwiniana*, a qual foi descartada, todas as seqüências tiveram valor de similaridade elevado (veja o Material Suplementar B para o resultado detalhado das buscas). Adicionalmente, o valor mais elevado para cada seqüência (direta e reversa) sempre foi relativo a algum representante de Sapindales, o que descarta a possibilidade de contaminação para todas as seqüências.

Em alguns casos, os valores atribuídos pelo Phred (Ewing & Green [13], Ewing *et al.* [14]) para as bases do espaçador *trn*G-S foram baixos e a região Phred20 não pôde ser identificada, caso de *E. densiflora*, *E. irwiniana*, *M. flavida* e *Metrodorea* sp. nov. Conseqüentemente, esses terminais foram excluídos. Adicionalmente, *M. maracasana* e *M. mollis* também foram excluídas devido sua região Phred20 possuir uma grande quantidade (cerca de 50%) de bases com valor baixo (veja o Material

Suplementar D).

3.5.2 Alinhamento e matriz de dados

Após as análises da qualidade das seqüências, e exclusão de terminais com dados apenas de ITS, as seqüências foram alinhadas, posteriormente combinadas e foi obtida uma matriz com 30 terminais e 1879 caracteres (1636 após a exclusão dos gaps das extremidades para ambas as partições). A Tabela 3.3 apresenta informações sumárias para cada uma das partições que compoêm a matriz obtida, a qual é apresentada no Apêndice D.

Apesar do alinhamento representar o cerne[4] de qualquer análise filogenética baseada em dados moleculares (*e.g.*, Giribet & Wheeler [17]), seu custo computacional (e temporal) pode limitar sua exploração de forma mais intensiva. Um exemplo disso é o alinhamento utilizado neste estudo. Certamente, as 1000 réplicas inferidas pelo programa representam um fragmento muito pequeno do espaço definido pelas possibilidades de alinhamento (*e.g.*, Bonizzoni & Vedova [4], Just & Vedova [25]). Entretanto, das poucas possibilidades computacionais que existem (considerando coerência metodológica entre construção de alinhamento e inferência de filogenia[5]), o uso de 1000 réplicas (veja o Material Suplementar E) pode ser considerado um avanço se comparado aos alinhamentos feitos pelo Clustal (Thompson *et al.* [49]) ou pelas "mãos" dos pesquisadores. Uma discussão mais abrangente específica sobre alinhamentos foge do escopo deste estudo e, para tanto, recomenda-se os trabalhos

[4]Pelo menos do ponto de vista semântico.

[5]Uma das alternativas ao ProAlign é o BAli-phy, mas seu custo computacional ainda é extremamente alto, o que impede seu uso de forma plena mesmo para matrizes de dados de tamanho pequeno a mediano, como é o caso deste estudo, em computadores "usuais".

de Edgar & Batzoglou [8], Iain *et al.* [50], Landan [29], Morrison [34], Notredame [35] e Thompson *et al.* [49].

3.5.3 Buscas por árvores

3.5.3.1 Diagnóstico das cadeias

Apesar de seus problemas (veja, *e.g.*, Ronquist & Huelsenbeck [44]), uma das formas mais comuns de se checar a não-convergência da(s) cadeia(s) em análises bayesianas de filogenias é plotar/monitorar as iterações por suas respectivas *ln*L (Figura 3.4) e supor que a existência de um platô no gráfico resultante seria indicação da convergência das cadeias.

A Figura 3.4 mostra os valores de cada uma das diferentes taxas de substituição do modelo (para cada uma das partições), para o comprimento das árvores (TL, comprimento conjunto de todos os ramos e para as duas partições) e para a *ln*L em relação às iterações. Como pode ser visto, em geral, todas as taxas de substituição variaram pouco do início ao fim das cadeias, tanto para a partição de ITS como para a de *trn*G-S, embora a variação tenha sido diferente entre as partições (*e.g.*, $(C \leftrightarrow G)_1$ e $(C \leftrightarrow G)_2$). Por outro lado, apesar de ter apresentado variação, o TL (e conseqüentemente a *ln*L) rapidamente se estabilizou. Aparentemente, houve uma rápida convergência e mistura das cadeias para todos os parâmetros.

Ainda na Figura 3.4, logo abaixo do TL e da *ln*L é apresentado o detalhe do início do platô (possível distribuição de equilíbrio), mostrando que provavelmente as cadeias alcançaram a distribuição alvo já próximo da iteração 20 mil.

Tabela 3.3 Sumário da composição das partições dos dados (gaps das extremidades das seqüências já excluídos). inv = sítios invariáveis, ? = dados "ausentes" e ambíguos.

Partição	Comprimento alinhado	% inv	% ?[1]	Freqüência (média)			
				A	C	G	T
ITS	656	58,69	33,69	0,17976	0,29807	0,32744	0,19474
*trn*G-S	980	70,92	60,41	0,37425	0,13242	0,14523	0,34810

[1]Gaps estão sendo considerados como "ausentes" e bases com valor baixo dentro da região Phred20 estão como ambíguas (embora os dois casos sejam tratados da mesma forma pelo MrBayes).

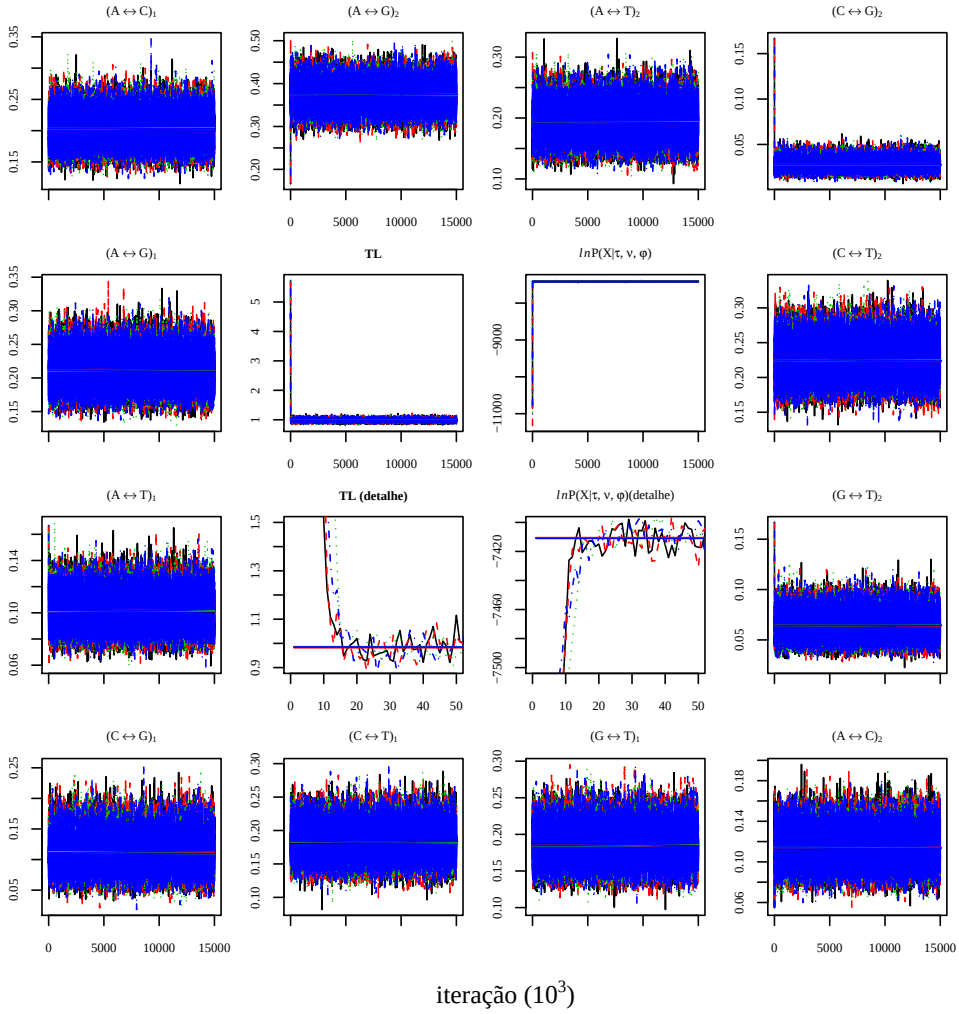


Figura 3.4 Variação dos valores dos parâmetros do modelo de substituição, comprimento de ramos (TL) e da *ln*L ao longo das cadeias. $(X \leftrightarrow Y)_i$ representa a taxa de transição entre X e Y para a partição i, TL = comprimento da árvore. (X e $Y \in$ {A, C, G, T}, $X \neq Y$ e $i \in$ {ITS, *trn*G-S}).

Um outro método gráfico proposto recentemente (Hillis *et al.* [22]) para monitorar a não convergência das cadeias é a visualização por escalonamento multidimensional (MDS) usando a distância de Robinson-Foulds ponderada (Robinson & Foulds [43]), como implementada no módulo Tree Set Visualization (Amenta *et al.* [3]) no programa Mesquite 1.01[6] ([32]). Segundo os autores do método (p. 477):

"this approach to visualizing the results of Bayesian analyses may prove to be a fruitful heuristic for assessing appropriate chain lengths and sampling strategies in MCMC Bayesian analyses of phylogeny. We have also used this approach with several empirical data sets (results not shown) and have found it to be a useful approach for assessing convergence among independent analyses."

Entretanto, uma primeira limitação dessa abordagem é a quantidade de memória necessária para estimar as distâncias. Por exemplo, para este estudo, não foi possível calcular a matriz de distâncias nem mesmo para as árvores dos intervalos HPD (9500 por rodada, veja o item 3.5.3.2 abaixo), muito menos para todas as árvores amostradas nas buscas (60000), pois a memória máxima alocável à JVM[7] é de 2048MB (independentemente da quantidade de memória física ou virtual que o computador possa ter) e a matriz de distâncias das 38000 árvores esgota esse limite. Uma solução para esse problema seria modificar o código do módulo (ou do programa) para que ele use o disco rígido para guardar informações durante certos cálculos e não a memória virtual. Isso tem que ser feito no nível de programação, *e.g.*, em vez de colocar informações grandes em uma escalar ou vetor, gravá-las em

[6] O TSV é incompatível com a versão 2 do Mesquite e somente até a 1.01 o TSV tem a opção da distância de Robinson-Foulds ponderada.

[7] Java Virtual Machine.

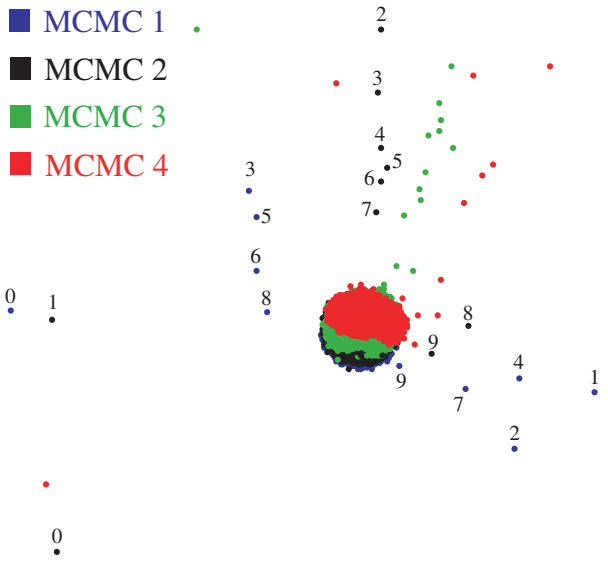


Figura 3.5 Visualização com escalonamento multidimensional usando a distância de Robinson-Foulds ponderada (Robinson & Foulds [43]) das primeiras 6000 iterações das 4 rodadas. Números representam iterações (mostrados apenas para as cadeias 1 e 2).

arquivo(s) temporário(s) para uso posterior pelo próprio programa poderia ser uma solução.

Além disso, como pode ser visto na Figura 3.5, o método por MDS não mostra nada que o método anterior não teria mostrado. De acordo com a Figura 3.5, as cadeias teriam convergido rapidamente, antes da iteração 20 mil, como foi "detectado" anteriormente (veja a Figura 3.4). Nesse caso, fica claro que a utilidade do método proposto por Hillis *et al.* ([22]) seria equivalente ao método "tradicional".

Uma outra forma de estimar a convergência das cadeias é usar os métodos já bem estabelecidos em estudos sobre MCMC. Exemplos disso são os métodos propostos por Brooks & Gelman ([5], para estimar a convergência entre cadeias) e a autocorrelação (para estimar a mistura de uma determinada cadeia). Uma discus-

são detalhada sobre os métodos de diagnose de MCMC foge ao escopo deste estudo, para tanto sugere-se os trabalhos de Brooks & Gelman [5], Brooks & Roberts [6] e Cowles & Carlin [7].

A Figura 3.6 mostra a estimativa de convergência das cadeias para os mesmos parâmetros da Figura 3.4 com base no CSRF ("corrected scale reduction factor", Brooks & Gelman [5]). Como pode ser notado, os resultados são discordantes dos obtidos com a abordagem "clássica" de detecção do "burn-in", assim como com o método proposto por Hillis *et al.* ([22]).

Para cada parâmetro analisado, o CSRF se aproxima de 1 em iterações diferentes. No caso do TL, ainda existe certa instabilidade um pouco além da iteração 4 milhões. Entretanto, a partir da iteração 5 milhões o CSRF já está bem próximo de 1 e estável para todos os parâmetros, podendo esse intervalo (5 milhões) ser usado como "burn-in". Portanto, a iteração na qual as cadeias supostamente convergiram é bastante diferente (cerca de 250 vezes maior) da que seria postulada pela simples detecção visual dos platôs apresentados na Figura 3.4 e dos resultados apresentados na Figura 3.5. A Figura 3.7 mostra um contraste mais marcante entre os resultados dos diferentes métodos. Na Figura 3.7(b) é mostrado que as cadeias teriam inicialmente convergido rápido, entretanto por volta da iteração 3000 elas foram para locais diferentes da distribuição, o que não é detectado pelos outros dois métodos[8].

O impacto preciso desses diferentes "burn-in" nos resultados filogenéticos (especialmente nos comprimentos de ramos), ainda precisa ser explorado de forma mais abrangente e foge ao escopo deste estudo (mas veja o item 3.5.5).

[8]Note que embora a *ln*L não seja um parâmetro, ela está sendo plotada na Figura 3.7(b) para possibilitar a comparação com os outros dois métodos.

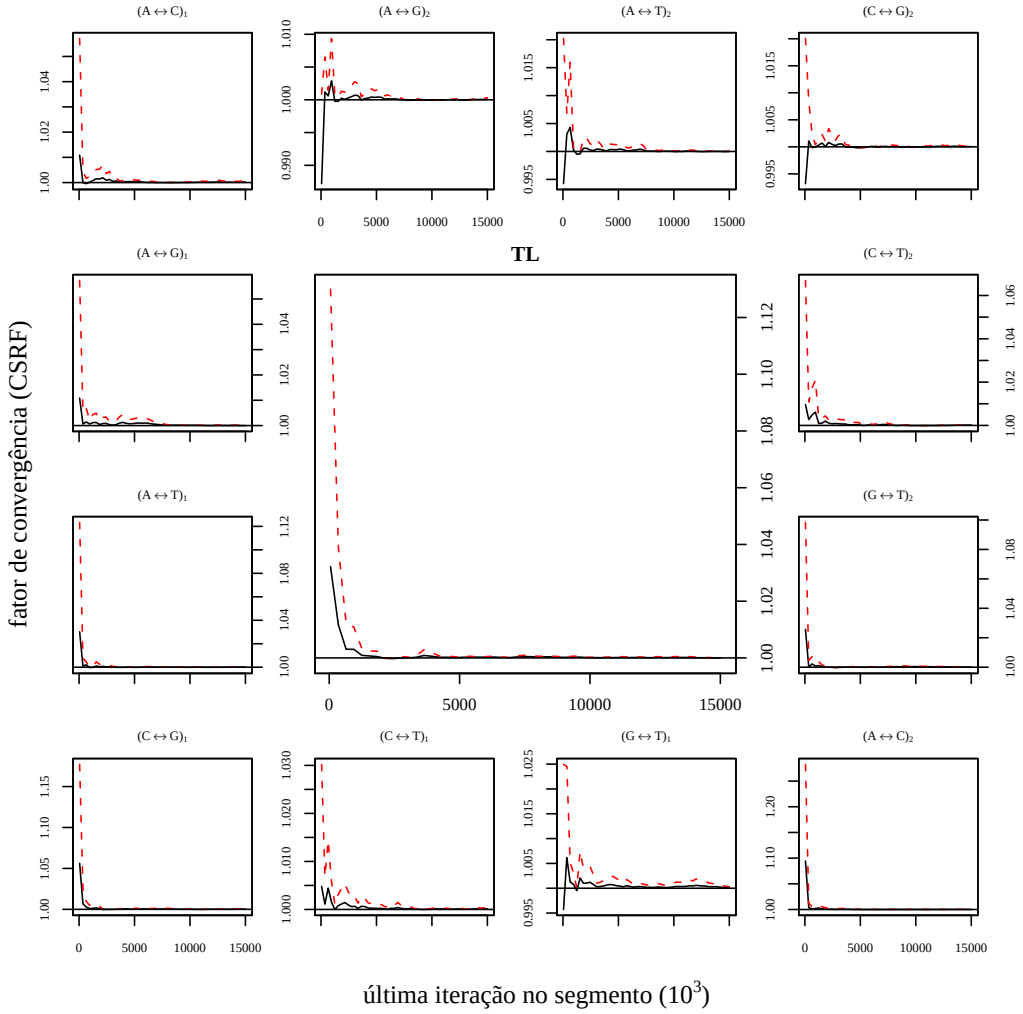
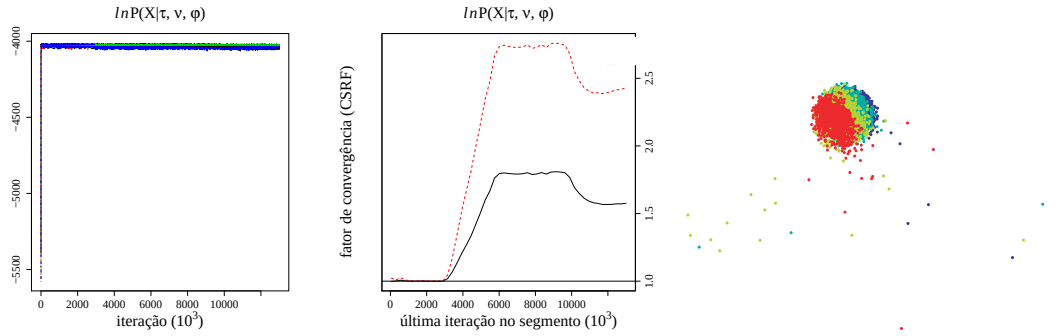


Figura 3.6 CSRF para cada um dos parâmetros do modelo de substituição e comprimento de ramos (TL). $(X \leftrightarrow Y)_i$ representa a taxa de transição entre X e Y para a partição i, TL = comprimento da árvore. Linha contínua = mediana, linha pontilhada = quantil 97,5% (X e $Y \in$ {A, C, G, T}, $X \neq Y$ e $i \in$ {ITS, *trn*G-S}).



(a) (b) (c)

Figura 3.7 Comparação entre os resultados dos três métodos. (a) "Método tradicional", mostra que as cadeias convergiram. (b) Método de Brooks & Gelman [5], mostra que as cadeias **não** convergiram e provavelmente estão em ótimos locais (ilhas, veja Maddison [31]) de alturas semelhantes (linhas como na Figura 3.6). (c) Método de Hillis *et al.* [22], mostra que as cadeias convergiram (cores representam cadeias diferentes).

Quanto às cadeias individuais em si, a Figura 3.8 mostra a autocorrelação da cadeia da Rodada 1. Como pode ser observado, os passos iniciais da cadeia foram esquecidos rapidamente em relação a todas as taxas de substituição do modelo e demorando um pouco mais para o TL. Conseqüentemente, pode-se assumir que houve mistura da cadeia. Dado que para as outras cadeias a autocorrelação apresentou resultados semelhantes, os gráficos referentes a essas cadeias estão sendo omitidos do texto, mas são apresentados no Material Suplementar A.

3.5.3.2 Intervalos HPD

Durante as buscas foram amostradas 60 mil árvores pelas quatro rodadas (15 mil por rodada) durante as 15 milhões de iterações. Como visto nas Figuras 3.4 e 3.6, existe uma variação na *ln*L das diferentes árvores, a qual representa uma distri-

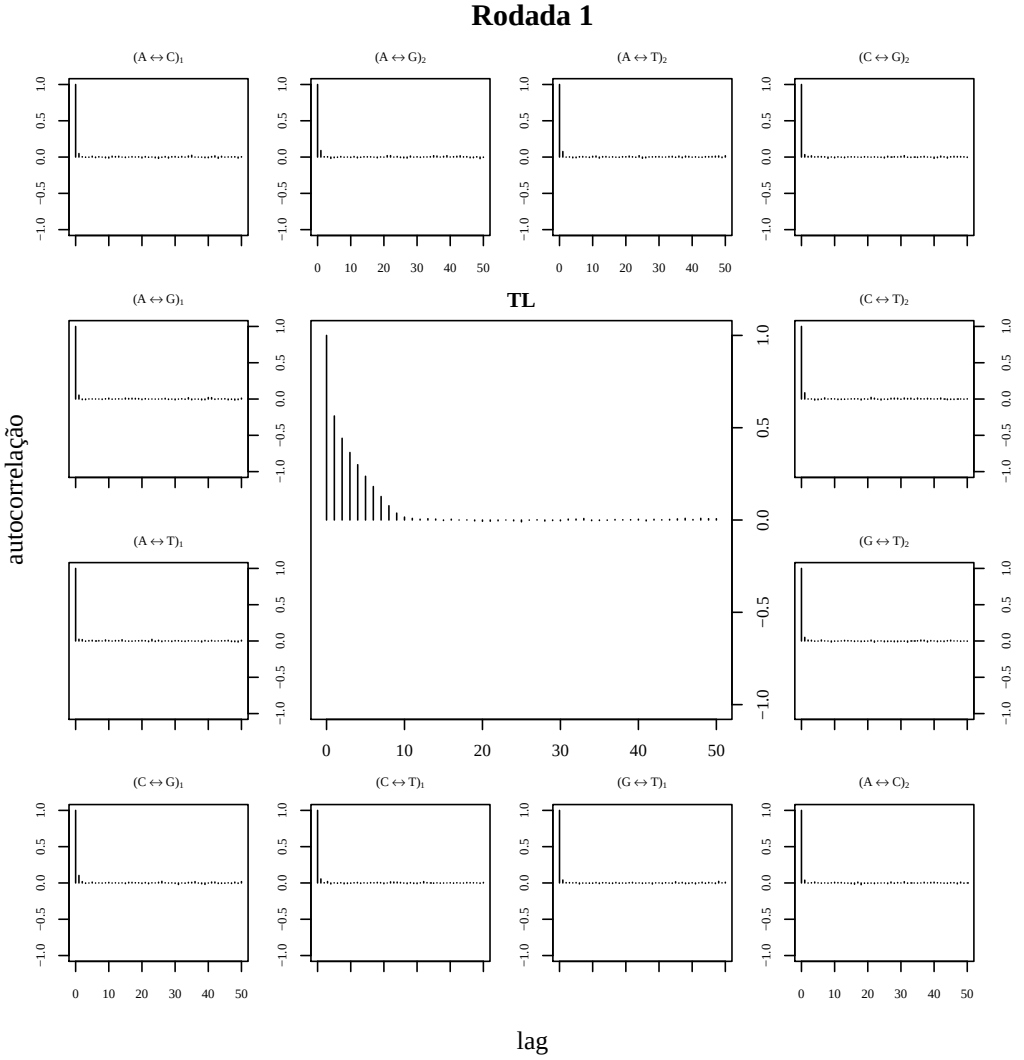


Figura 3.8 Autocorrelação para os parâmetros do modelo de substituição e comprimento de ramos (TL) na rodada 1. $(X \leftrightarrow Y)_i$ representa a taxa de transição entre X e Y para a partição i, TL = comprimento da árvore. (X e $Y \in$ {A, C, G, T}, $X \neq Y$ e $i \in$ {ITS, *trn*G-S}).

buição subjacente. Com base nessa distribuição de *ln*L, após a exclusão das iterações relativas ao "burn-in", como detectado pelo método de Brooks & Gelman ([5]), foram construídos intervalos HPD de 95% para cada uma das rodadas (Figura 3.9). Somadas, as iterações que compoẽm os intervalos HPD de todas as rodadas incluíram 38 mil árvores.

Para todas as rodadas, as iterações apresentaram *ln*L normalmente (ou quase) distribuída (Figura 3.9), o que está de acordo com o suposto alcançe da distribuição de equilíbrio.

3.5.4 Grupos monofiléticos, suporte e dados ambíguos

Pilocarpinae emerge como não-monofilética (provavelmente parafilética) e deve incluir grupos de Pteleinae (ao menos em relação aos gêneros amostrados) para que se configure como um grupo monofilético (Figura 3.11). O grupo formado pelos seis gêneros (os quatro de Pilocarpinae e dois de Pteleinae) possui PP máxima (PP=1) e separa-se em dois grandes clados (ambos também com PP=1): um composto apenas por *Pilocarpus* e o outro composto pelos outros cinco gêneros incluídos neste estudo (*Balfourodendron*, *Esenbeckia*, *Helietta*, *Metrodorea* e *Raulinoa*). Uma diferença morfológica marcante entre esses dois grupos é o tipo de de inflorescência quanto à ramificação (Figura 3.10): *Pilocarpus* apresenta racemo e os outros gêneros apresentam inflorescência ramificada (podendo se apresentar na forma de panícula - reduzida ou não, dicásio ou fasciculada - *Raulinoa*)[9]. Para facilitar a comunicação, o clado de inflorescência ramificada será informalmente chamado de "Paniculado".

Em *Pilocarpus*, dos nove ramos internos, apenas quatro possuem suporte >

[9] *Provavelmente* uma simplesiomorfia.

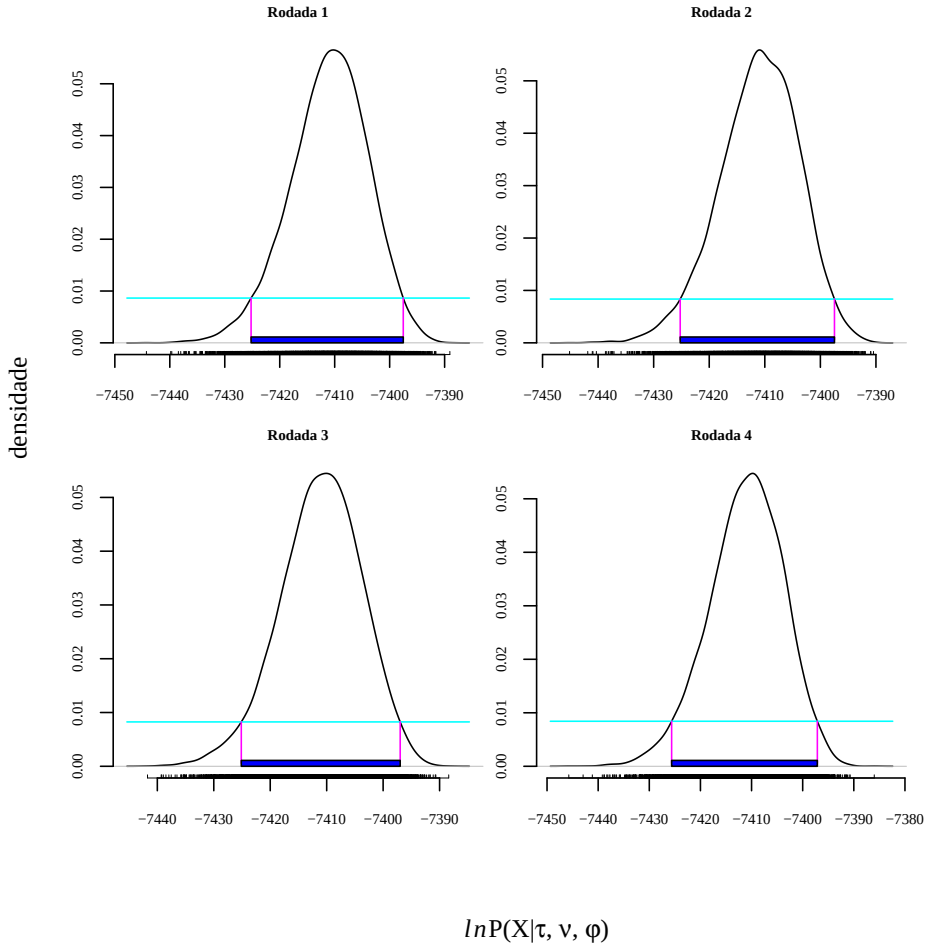


Figura 3.9 Intervalos HPD para as quatro rodadas. Apenas as árvore desses intervalos serão usadas nas análises posteriores.

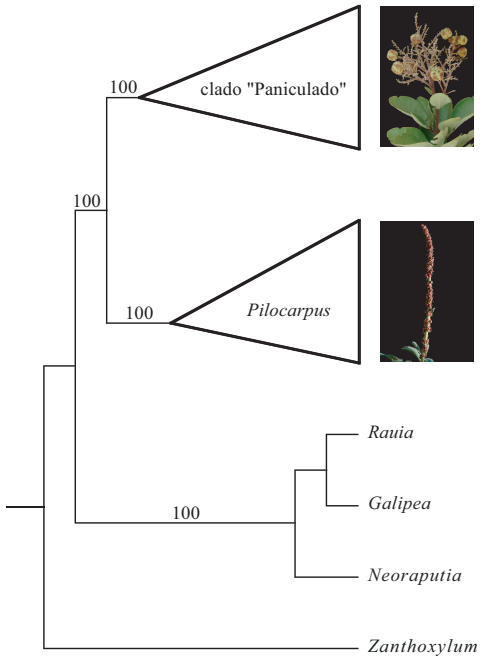


Figura 3.10 Filogenia de Pilocarpinae e gêneros próximos, principais clados, números sobre os ramos representam probabilidades posteriores. Imagens: *E. pumila* (superior), *P. trachylophus* (inferior). (Fotos por R.G. Udulutsch)

75% (Figura 3.11(a)), apesar do gênero emergir como monofilético e com suporte máximo. Como pode ser observado na Figura 3.12 essa falta de resolução dentro do gênero não tem relação direta nem com o número de dados ambíguos nem com o número de gaps (inverso do comprimento da seqüência), separadamente ou combinados. Dessa forma, fica claro que outras fontes de dados precisam ser avaliadas para que se obtenha uma filogenia mais resolvida e com suporte elevado para os ramos internos do grupo.

Por outro lado, no clado "Paniculado", as relações interespecíficas estão mais bem resolvidas e a maioria dos ramos internos possui suporte $> 75\%$ (apenas três dos 13 possuem suporte $< 75\%$, Figura 3.11(a)). Exceto *Esenbeckia*, cada um dos gêneros reconhecidos tradicionalmente emerge como monofilético (mas veja o item 3.5.4), embora a base do clado não seja resolvida (suporte $< 50\%$).

Comparando-se o nível de resolução nos dois clados (Figura 3.11(a)), fica claro que a informatividade das regiões utilizadas varia em diferentes pontos da filogenia. Na base do grupo-interno emergiram dois clados com suporte elevado e comprimento de ramos acima da média[10] (Figuras 3.11(a) e 3.11(b)). Entretanto, dentro de *Pilocarpus* os ramos dos grupos de espécies, na maioria, têm suporte baixo e seus comprimentos são bastante curtos (bem abaixo da média). Talvez, uma diversificação recente[11] (ou um processo anagenético[12] menos "ativo" após a separação da linhagem basal) poderia ser responsável por esse padrão. Por outro lado, dentro do clado "Paniculado" a maioria dos ramos (dos gêneros e dos grupos

[10]O comprimento do ramo que leva ao clado "Paniculado" é igual a 0,015794 e o ramo que leva a *Pilocarpus* tem comprimento igual a 0,024073, enquanto a média dos comprimentos de ramos (desconsiderando o ramo de *E. grandiflora*) é 0,0151320714285714.

[11]Para aceitar ou rejeitar essa hipótese teria que ser feita uma análise com relógio molecular.

[12]Taxa de substituição (ou probabilidde de transição entre os estados de caráter).

de espécies) tem suporte elevado e comprimento acima da média, mas os ramos que separam os grupos de gêneros tem comprimento bem abaixo da média e com suporte variado. Esse padrão poderia ser explicado por uma "diversificação de gêneros" em momentos "próximos" seguida por anagênese "rápida" dentro dos gêneros.

Como mostrado na Figura 3.11(a), no clado "Paniculado", os gêneros tradicionalmente reconhecidos emergem como monofiléticos, exceto *Esenbeckia*. Como pode ser visto na Figura 3.11(b), a linhagem supostamente basal desse gênero (*E. grandiflora*, ramo destacado em cinza) possui ramo cujo comprimento[13] ultrapassa em mais de cinco vezes a média, apesar de *E. grandiflora* ser a espécie com seqüência mais curta (maior número de gaps) e a quarta com maior quantidade de dados ambíguos em relação comprimento total da seqüência (Figura 3.12). Logo, qualquer relação entre esses fatores e o elevado comprimento do ramo estaria descartada, dado que o MrBayes ignora os dados ambíguos e gaps. Portanto, esse comprimento de ramo, provavelmente, é resultado de anagênese mais ativa nessa linhagem. Por outro lado, como resultado desse processo anagenético exacerbado, por chance, o comprimento desse ramo influenciará diretamente a estabilidade de *E. grandiflora* (Figura 3.13) entre as árvores amostradas e, conseqüentemente, no suporte dos ramos da região, como pode ser observado nas Figuras 3.11(a) e 3.11(c) (ramos destacados em cinza). Na Figura 3.11(c) pode ser visto que a exclusão de *E. grandiflora* leva a um aumento de cerca de 5-30% nos ramos destacados.

[13] 0,081551.

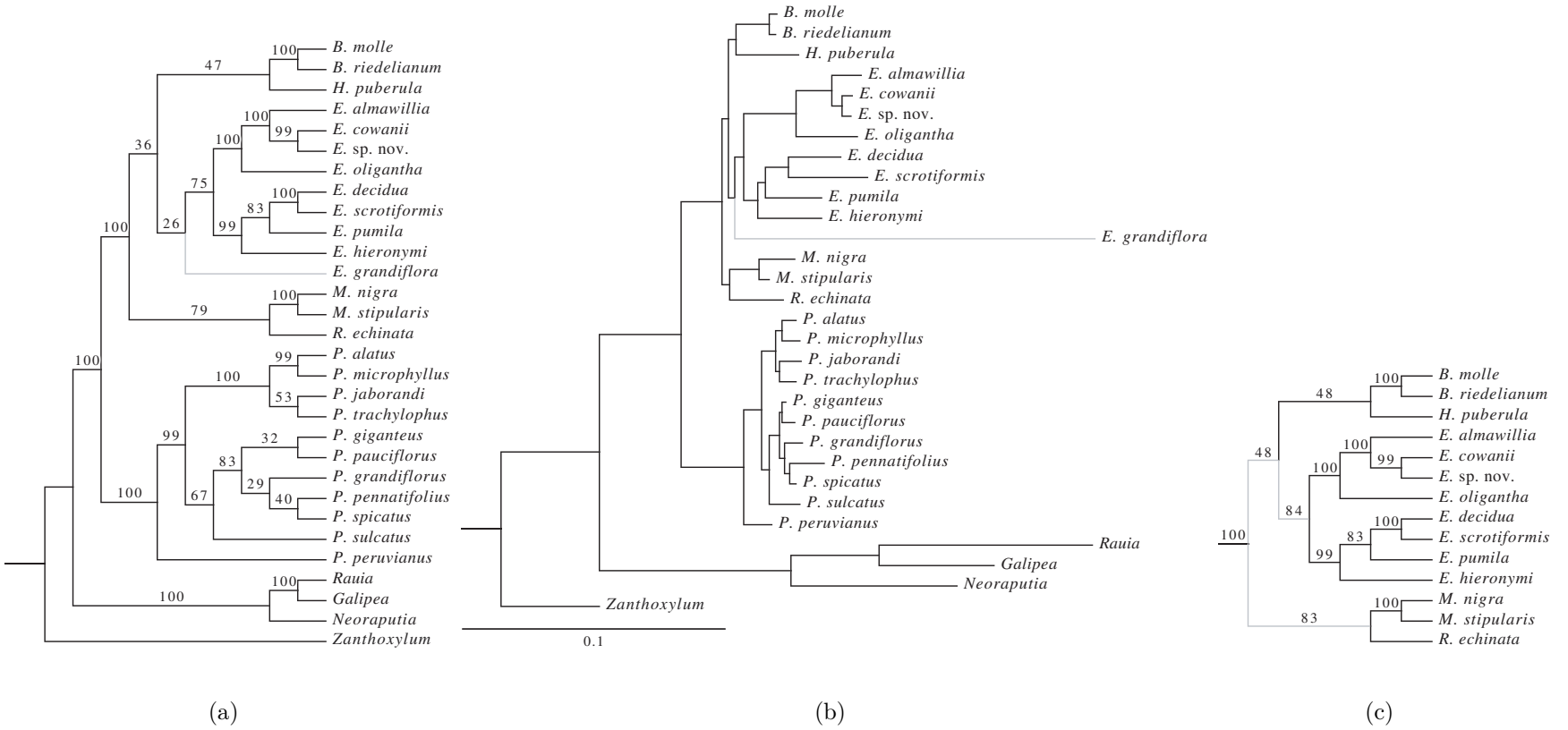


Figura 3.11 Filogenia de Pilocarpinae e gêneros próximos, consenso de maioria estendido das 38000 árvores incluídas nos intervalos HPD (veja a Figura 3.9). (a) Cladograma, números sobre os ramos representam probabilidades posteriores. (b) Filograma, note o comprimento do ramo de *E. grandiglora*. (c) Parte da filogenia enfatizando a influência de *E. grandiflora* no suporte dos ramos próximos (ramos destacados em cinza).

Ainda no clado "Paniculado", os outros gêneros representados por mais de uma espécie (*Balfourodendron* e *Metrodorea*) emergiram como monofiléticos e com suporte máximo (PP=1). Adicionalmente, *Metrodorea* e *Raulinoa*, gêneros morfologicamente bastante diferentes (Figura 3.2 com destaque para os detalhes vegetativos, (e) e (h)), também formam um clado com suporte $> 75\%$ (79% com *E. grandiflora* incluída na análise e 83% se excluída).

3.5.5 "Burn-in" e implicações filogenéticas

Como comentado anteriormente, os métodos usados para detecção do "burn-in" podem chegar a resultados diferentes. Então, uma pergunta que pode ser feita é: qual a possível influência dos diferentes métodos nos resultados filogenéticos? Uma resposta precisa a essa pergunta ainda está longe de ser alcançada, mas é possível observar algumas de suas implicações.

Considerando os resultados obtidos pelos métodos "tradicional" (Figura 3.4) e o de Hillis *et al.* ([22]) (Figura 3.5), as cadeias das diferentes rodadas teriam convergido antes da iteração 20 mil. Se isso tivesse acontecido, então todas as árvores amostradas após a iteração 20 mil estariam sendo amostradas da distribuição de equilíbrio. Assim, seria esperado que o consenso de maioria das árvores não mudasse[14], independentemente da banda amostrada. Entretanto, se analizarmos as primeiras 100 mil iterações e usarmos as primeiras 20 mil como "burn-in", veremos que existe diferença na topologia do consenso de maioria (Figura 3.14), o que deixa claro que o "burn-in" pode ter influência direta não apenas nos comprimentos de ramos (o que é esperado), mas também na própria topologia da árvore final obtida, embora a chance

[14]Em relação ao τ, não ao ν.

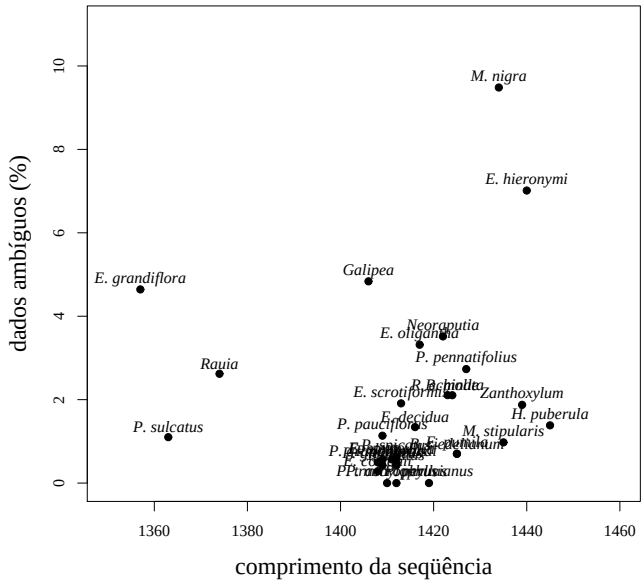


Figura 3.12 Relação entre o número de dados ambíguos e o comprimento da seqüência.

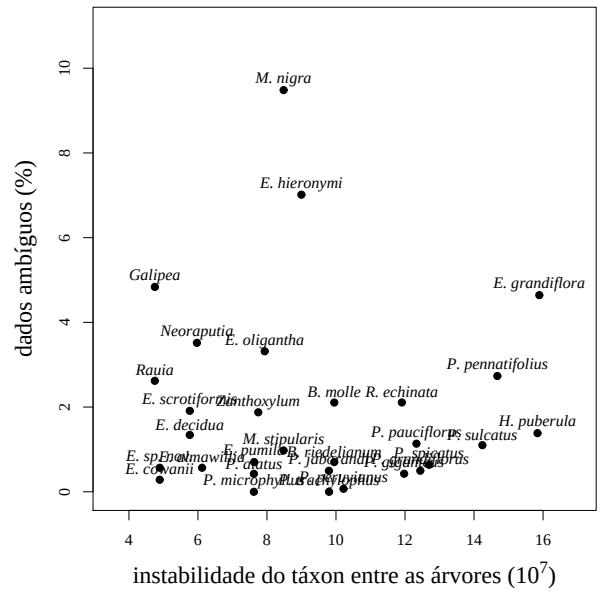


Figura 3.13 Relação entre o número de dados ambíguos e a instabilidade do terminal.

de ocorrerem tais diferenças sejam inversamente relacionadas ao suporte do ramo (o qual, em si, pode ser influenciado diretamente pelo "burn-in"). Nesse sentido, conseqüentemente, a influência do "burn-in" se extende a todas as conclusões que se tire com base na topologia encontrada (*e.g.*, diversificação de clados, otimização de caracteres etc.)

3.5.6 Relações filogenéticas e implicações taxonômicas

3.5.6.1 Relações e grupos

As relações mostradas na Figura 3.11 revelam que três dos quatro gêneros de Pilocarpinae (*Esenbeckia*, *Metrodorea* e *Raulinoa*) possuem parentesco maior com gêneros de outra subtribo (tribo e subfamília) do que com o gênero-tipo da própria subtribo, *Pilocarpus* (como sugerido por Groppo [19]). Adicionalmente, esses clados possuem suporte máximo e ramos com comprimentos acima da média. Embora o clado "Paniculado" tenha uma politomia na sua base, a maioria dos gêneros[15] amostrados com mais de uma espécie emergem como monofiléticos e possuem suporte $> 75\%$ (Figura 3.11).

Apesar de *Balfourodendron* possuir suporte máximo (PP=1), seu agrupamento com *Helietta* não possui sustentação (*i.e.*, PP $< 0{,}5$), o que deixa o status de Pteleinae incerto enquanto grupo, o que é concordante com os resultados obtidos por Groppo[16]. *Balfourodendron* e *Helietta* são gêneros facilmente reconhecidos quando férteis, em especial quando possuem frutos, mas para o não-taxonomista eles

[15]Exceção apenas para *Esenbeckia*.

[16]Apesar de Groppo ([19]) afirmar que Pteleinae é polifilética (p. 90), sua filogenia (Figura 3, p. 78) não corrobora essa afirmação.

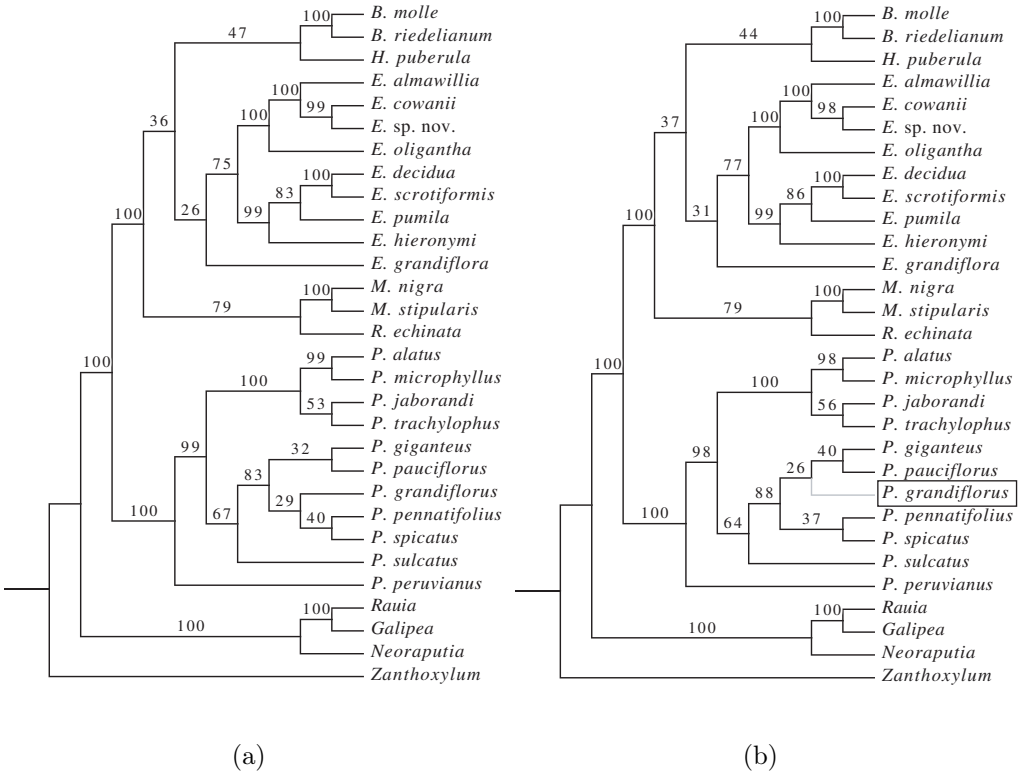


Figura 3.14 Filogenia de Pilocarpinae e gêneros próximos, consensos de maioria estendidos. Números sobre os ramos representam probabilidades posteriores. (a) Cladograma obtido com as árvores incluídas nos intervalos HPD (mesma árvore apresentada na Figura 3.11(a)). (b) Cladograma obtido com as primeiras 100 árvores (100 mil iterações com amostragem a cada 1000 e após a exclusão do "burn-in" de 20 mil). Diferença topológica destacada em cinza (*P. grandiflorus*).

são facilmente confundidos com espécies 3-folioladas de *Esenbeckia* quando estéreis ou mesmo quando floridos, como já mencionado anteriormente. Mas, se esses dois gêneros devem ou não ser combinados com *Esenbeckia* num único gênero, é algo que não encontra apoio nos resultados apresentados[17].

Por outro lado, para *Esenbeckia* não existe uma definição em relação ao seu status enquanto grupo (monofilético ou não), mas se *E. grandiflora* for excluída, o gênero se torna monofilético e seu suporte fica $> 75\%$. Adicionalmente, como comentado anteriormente, a exclusão de *E. grandiflora* leva a um aumento no suporte dos ramos próximos, dada a instabilidade desse terminal (Figura 3.13) e sua respectiva influência negativa na região (Figura 3.11(c)). Entretanto, sua simples exclusão do gênero não melhora em nada a taxonomia do grupo em termos práticos (*e.g.*, identificação). Dessa forma, fica claro que é necessária a utilização de outras fontes de dados na "base" de *Esenbeckia*. Apesar dessa indefinição em relação ao status de *Esenbeckia*, note que *Esenbeckia* é o nome com prioridade, logo qualquer alteração em termos de sinonimização dará prioridade a *Esenbeckia*. Conseqüentemente, o nome pode ser usado mesmo que o status do gênero não esteja definido enquanto grupo monofilético, pois na pior situação (sinonimização de todos os gêneros do clado "Paniculado"), *Esenbeckia* prevalecerá.

Por sua vez, *Metrodorea* surge com suporte máximo e sua relação com *Raulinoa* possui suporte $> 75\%$ (79% e 83%, com e sem *E. grandiflora* na filogenia, respectivamente). Esse agrupamento de *Metrodorea* com *Raulinoa* corrobora as suposições iniciais de Cowan (1960) sobre a provável maior proximidade de *Raulinoa* com *Metrodorea* do que com *Esenbeckia*, embora pareçam falsas suas proposições de

[17] *I.e.*, os dados não são "decisivos" nesse sentido.

equivalência entre a bainha de *Metrodorea* e os espinho de *Raulinoa* (Kaastra [26]).

3.5.6.2 Revisitando as Pteleinae: Pilocarpinae *s.s.* + clado "Paniculado"

De acordo com os resultados obtidos e as considerações anteriores, a solução para a situação taxonômica em Pilocarpinae envolve duas possibilidades:

1) reduzir a circunscrição de Pilocarpinae para incluir apenas o gênero-tipo (*Pilocarpus*) em uma subtribo "Pilocarpinae *s.s.*" e transferir os outros gêneros (*Esenbeckia*, *Metrodorea* e *Raulinoa*) para Pteleinae (subtribo dos gêneros *Balfourodendron* e *Helietta*); e

2) assumir "Pilocarpinae *s.s.*" (como antes), excluir *Balfourodendron* e *Helietta* de Pteleinae (esta ficando apenas com seu gênero-tipo, *Ptelea*) e eleger uma nova subtribo para o clado "Paniculado".

A redução de Pilocarpinae para "Pilocarpinae *s.s.*" (contemplada nas duas possibilidades) poderia[18] estar refletida no padrão filogenético obtido (Figura 3.10 e Figura 3.11) e estaria em consonância com a opinião de autores anteriores (*e.g.*, Kaastra [27]). Entretanto, a transferência de *Esenbeckia*, *Metrodorea*, e *Raulinoa* para Pteleinae não estaria corroborada nos resultados obtidos, dado que o gênero-tipo de Pteleinae (*Ptelea*) não está incluído na análise e, portanto, essa seria uma proposição duvidosa e sem embasamento em topologia alguma. Adicionalmente, *Ptelea* poderia não estar relacionada filogeneticamente com nenhum dos gêneros. Conseqüentemente, essa transferência só criaria confusão nomenclatural adicional.

Por outro lado, considerando a segunda alternativa, a exclusão de *Balfourodendron* e *Helietta* das Pteleinae encontra a mesma dificuldade relativa à inclusão

[18] O grupo está refletido, mas a categoria depende do autor.

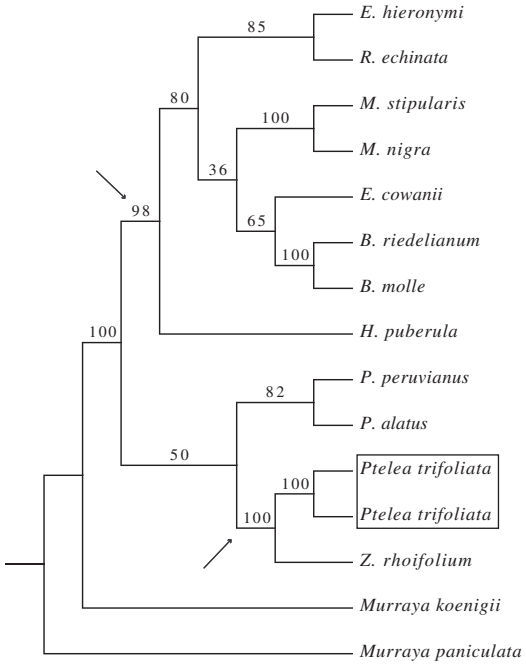


Figura 3.15 Consenso de maioria estendido baseado em 18000 árvores ("burn-in" de 5500 árvores para cada rodada) mostrando a posição filogenética de *Ptelea* em relação a *Pilocarpus* e ao clado "Paniculado". Números acima dos ramos representam probabilidades posteriores. Setas destacam os suportes dos ramos que levam ao clado "Paniculado" e ao clado (*Zanthoxylum*, *Ptelea*).

de *Esenbeckia*, *Metrodorea* e *Raulinoa* na subtribo.

Para resolver esse impasse, foi executada um análise específica para verificar o agrupamento de *Ptelea* com os outros gêneros baseada em dados do ITS1 (veja o Material Suplementar C para a matriz utilizada).

Como pode ser visto na Figura 3.15, *Ptelea* emerge fora do clado (*Pilocarpus*, clado "Paniculado") e, conseqüentemente, sua relação filogenética é mais estreita com outros grupos de Rutaceae do que com os gêneros *Balfourodendron* e *Helietta*. Dessa forma, fica claro que Pteleinae não é monofilética e que *Balfourodendron* e *Helietta*

devem ser realocados em outro grupo. Esse padrão mostrado na Figura 3.15 associado aos resultados apresentados nas Figura 3.10, a qual separa de um lado o clado "Paniculado" e de outro o gênero *Pilocarpus*, está de acordo com as suposições de Kaastra (1982) em separar *Pilocarpus* dos outros três gêneros (*Esenbeckia*, *Metrodorea* e *Raulinoa*), o que tem suporte tanto morfológico (*e.g.*, inflorescência em racemo - Figura 3.2, presença de gândulas póstero-dorsais nas anteras) como químicos (presença de imidazóis). Conseqüentemente, fica justificável uma nova circunscrição da subtribo Pilocarpinae, ficando monogenérica (apenas com *Pilocarpus*, veja o item 3.5.6.3 abaixo).

As Pteleinae, por sua vez, também mostraram-se não monofiléticas e parte delas (*Balfourodendron* e *Helietta*) formam o clado "Paniculado" junto com *Esenbeckia*, *Metrodorea* e *Raulinoa*. Essa relação, associada à maior proximidade filogenética de *Ptelea* com outros grupos de Rutaceae do que com *Balfourodendron* e *Helietta*, deixa o clado "Paniculado" "órfão" em relação à subtribo. Portanto, uma nova subtribo para, e topologicamente equivalente a, o clado "Paniculado" deve ser proposta (veja o item 3.5.6.3 abaixo).

3.5.6.3 Rearranjos genéricos em Pilocarpinae e uma nova subtribo[19]

Como já discutido anteriormente, parte dos gêneros de Pilocarpinae (*Esenbeckia*, *Metrodorea* e *Raulinoa*) são mais estreitamente relacionados com parte dos gêneros de Pteleinae (*Balfourodendron* e *Helietta*) do que cada uma dessas partes com os outros gêneros que compõem suas respectivas subtribos. Logo, existe a ne-

[19]Para efeitos de validade, esta tese não deve ser considerada uma publicação para qualquer alteração taxonômica, as quais serão publicadas de forma válida em outro lugar.

cessidade de se fazer rearranjos dos gêneros pertencentes às Pilocarpinae e parte dos gêneros de Pteleinae. Assim, a classificação para o grupo fica da seguinte forma:

Tribo **Galipeeae** Kallunki, *Kew Bulletin*, 53(2): 257. 1998 [= Cusparieae DC., *nom. illegit.*, *Mkm. Mus. Hist. Nat. Paris* 9: 141. 1822; tipo: *Galipea trifoliata* Aubl.]

Subtribo **Esenbeckiinae** P. Dias subtrib. nov. (tipo: *Esenbeckia* Engl.)

Subtribo **Galipeinae** Kallunki[20], *Kew Bulletin*, 53(2): 257. 1998 [= Cuspariinae Engl., *nom. illegit.*, Engler & Prantl, *Nat. Pflanzenfam.* 3(4): 111, 160 (1896)]

Subtribo **Pilocarpinae** Engl., *Fl. bras.* 12(2): 129-130, excl. gen. P. Dias

3.6 Conclusões

O método mais comun de diagnose da convergência das MCMC em estudos filogenéticos é o monitoramento da *ln*L em realação às iterações ao longo das cadeias. Entretanto, como já discutido por outros autores (*e.g.*, Ronquist & Huelsenbeck [44]) esse método pode levar a equívocos. O método proposto recentemente por Hillis *et al.* ([22]) usando a distância de Robinson-Foulds ponderada através de MDS (proposto como um método alternativo de diagnose) é equivalentes ao método anterior, portanto, deve sofre de problemas semelhantes. Nesse sentido, propõe-se que se use os métodos já bem estabelecidos em estudos de diagnose de MCMC, como o método de Brooks & Gelman ([5]). Embora estudos específicos tenham que ser conduzidos quanto à violação ou adequação de alguns dos princiípios desses métodos

[20]Embora não haja estudo definitivo sobre o status dessa subtribo enquanto grupo monofilético, Groppo [19] sugere que ela talvez não seja.

pelos dados filogenéticos, espera-se que a sugestão de seu uso incentive sua exploração em estudos futuros, dada a importância do método bayesiano com MCMC na filogenética atual.

Por outro lado, os resultados das análises filogenéticas sugerem que Pilocarpinae não é monofilética: os gêneros *Esenbeckia*, *Metrodorea* e *Raulinoa* (como tradicionalmente reconhecidos) são mais próximos de *Balfourodendron* e *Helietta* (subtribo Pteleinae) do que de *Pilocarpus* (gênero-tipo da subtribo). Os resultados mostram que os gêneros estudados se separam em dois grupos, de um lado *Pilocarpus* e de outro um clado caracterizado principalmente pela presença de inflorescência ramificada (clado "Paniculado") englobando todos os outros gêneros. Com exceção de *Esenbeckia*, todos os gêneros incluídos no estudo emergiram como monofiléticos e com suporte $> 75\%$, embora as relações entre os gêneros dentro do clado "Paniculado" não sejam totalmente conhecidas. Dentre as epécies de *Esenbeckia* (único gênero cujo status enquanto grupo está indefinido), *E. grandiflora* apresentou alta instabilidade nas árvores-fonte e, dado que sua posição parece ser basal no gênero, isso influenciou diretamente o suporte do grupo e dos ramos próximos. Uma vez que *Pilocarpus* ficou separado dos outros gêneros de Pilocarpinae e esses são mais aparentados com gêneros de outra subtribo, os resultados sustentam a recircunscrição da subtribo, a qual, como definida aqui, ficou monogenérica. Adicionalmente, uma subtribo **Esenbeckiinae** P. Dias foi criada e é equivalente ao clado "Paniculado".

3.7 Referências

[1] Altschul, S. F., Gish, W., Miller, W., Meyers, E. W. & Lipman, D. J. 1990. Basic Local Alignment Search Tool. *J. Mol. Biol.* 215: 403–410.

[2] Alvarez, I. & Wendel, J. F. 2003. Ribosomal ITS sequences and plant phylogenetic inference. *Mol. Phylogen. Evol.* 29: 417–434.

[3] Amenta, N., John, K. S., Klingner, J., Heath, T. A., Clarke, F., Edwards, D., Neris, S., Mahindru, R. & Postarnakevich, N. 2004. *Tree set visualization module for Mesquite: visualizing sets of phylogenetic trees.* http://comet.lehman.cuny.edu/treeviz/.

[4] Bonizzoni, P. & Vedova, G. D. 2001. The complexity of multiple sequence alignment with SP-score that is a metric. *Theoret. Comp. Science* 259: 63–79.

[5] Brooks, S. & Gelman, A. 1998. General methods for monitoring convergence of iterative simulations. *J. Comp. Graph. Stat.* 7: 434–455.

[6] Brooks, S. P. & Roberts, G. O. 1998. Convergence assessment techniques for Markov chain Monte Carlo *Stat. Computing* 8: 319–335.

[7] Cowles, M. K. & Carlin, B. P. 1996. Markov chain Monte Carlo convergence diagnostics: a comparative review. *J. Amer. Statist. Soc.* 91: 883–904.

[8] Edgar, R. C. & Batzoglou, S. 2006. Multiple sequence alignment. *Curr. Op. Struct. Biol.* 16: 368–373.

[9] Elias, T. S. 1970. The correct name for the genus *Cusparia* (Rutaceae). *Taxon* 19: 573–575.

[10] Engler, H. G. A. 1874. Rutaceae. *In* Martius, C. F. P. & Eichler, A. G. (eds.) *Flora brasiliensis.* vol. 12, Frid. Fleischer, Leipzig, 77–196.

[11] Engler, H. G. A. 1896. Rutaceae. *In* Engler, H. G. A. & Prantl, K. (eds.) *Die natürlichen Pflanzenfamilien.* 1 ed. Wilhelm Engelmann, Leipzig, 95–201.

[12] Engler, H. G. A. 1931. Rutaceae. *In* Engler, H. G. A. & Prantl, K. (eds.) *Die natürlichen Pflanzenfamilien.* 2 ed. Wilhelm Engelmann, Leipzig, 187–359.

[13] Ewing, B. & Green, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 8: 186–194.

[14] Ewing, B., Hillier, L., Wendl, M. C. & Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* 8: 175–185.

[15] Felsenstein, J. 2004. *Inferring phylogenies.* Sinauer Associates, Sunderland.

[16] Gelman, A. & Rubin, D. B. 1992. Inference from iterative simulation using multiple sequences. *Stat. Science* 7: 457–511.

[17] Giribet, G. & Wheeler, W. C. 1999. On Gaps. *Mol. Phylogen. Evol.* 13: 132–143.

[18] Gordon, D., Abajian, C. & Green, P. 1998. Consed: a graphical tool for sequence finishing. *Genome Res.* 8: 195–202.

[19] Groppo, M. 2004. *Filogenia de Rutaceae e revisão taxonômica de* ***Hortia*** *Vand. (Rutaceae).* Tese de doutorado, Universidade de São Paulo, São Paulo.

[20] Hamilton, M. B. 1999. Four primer pair for the amplification of chloroplast intergenic regions with interspecific variation. *Mol. Ecol.* 8: 513–525.

[21] Hastings, W. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57: 97–109.

[22] Hillis, D. M., Heath, T. A. & John, K. S. 2005. Analysis and visualization of tree space. *Syst. Biol.* 54: 471–482.

[23] Huelsenbeck, J. P., Larget, B., Miller, R. & Ronquist, F. 2002. Potential applications and pitfalls of bayesian inference of phylogeny. *Syst. Biol.* 51: 673–688.

[24] Hyndman, R. J. & Einbeck, J. 2007. *hdrcde: highest density regions and conditional density estimation.* R package version 2.07.

[25] Just, W. & Vedova, G. D. 2004. Multiple sequence alignment as facility location problem. *INFORMS J. Comput.* 16: 430–440.

[26] Kaastra, R. C. 1978. Leaf sheaths and obturators in Rutaceae-Pilocarpinae. *Beitr. Biol. Pflanzen* 53: 317–320.

[27] Kaastra, R. C. 1982. Pilocarpinae (Rutaceae). *Fl. Neotrop. Monogr.* 33: 1–198.

[28] Kallunki, J. A. & Pirani, J. R. 1998. Synopses of *Angostura* Roem. & Schult. and *Conchocarpus* J. C. Mikan (Rutaceae). *Kew Bull.* 53: 257–334.

[29] Landan, G. 2005. *Multiple sequence alignment errors and phylogenetic reconstruction.* Ph.D. dissertation, Tel-Aviv University, Tel-Aviv.

[30] Löytynoja, A. & Milinkovitch, M. C. 2003. A hidden Markov model for progressive multiple alignment. *Bioinformatics* 19: 1505–1513.

[31] Maddison, D. R. 1991. The discovery and importance of multiple islands of most-parsimonious trees. *Syst. Zool.* 40: 315–328.

[32] Maddison, W. P. & Maddison, D. R. 2004. *Mesquite: a modular system for evolutionary analysis. v. 1.01.* http://mesquiteproject.org.

[33] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. 1953. Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21: 1087–1092.

[34] Morrinson, D. A. 2006. Multiple sequence alignment for phylogenetic purposes. *Austral. Syst. Bot.* 19: 479–539.

[35] Notredame, C. 2007. Recent evolutions of multiple sequence alignment algorithms. *PLOS Comput. Biol.* 3: e123.

[36] Pinheiro, C. U. B. 1997. Jaborandi (*Pilocarpus* spp. and Rutaceae): a wild species and its rapid transformation into a crop. *J. Econ. Bot.* 51: 49–58.

[37] Pinheiro, C. U. B. 2002. Extrativismo, cultivo e privatização do jaborandi

(*Pilocarpus microphyllus* Stapf ex Holm.; Rutaceae) no Maranhão, Brasil. *Acta Bot. Bras.* 16: 141–150.

[38] Pirani, J. R. 1998. A revision of *Helietta* and *Balfourodendron* (Rutaceae-Pteleinae). *Brittonia* 50: 348–380.

[39] Pirani, J. R. 1999. Two new species of *Esenbeckia* (Rutaceae, Pilocarpinae) from Brazil and Bolivia. *Bot. J. Linn. Soc.* 129: 305–313.

[40] Plummer, M., Best, N., Cowles, K. & Vines, K. 2007. *coda: output analysis and diagnostics for MCMC.* R package version 0.12-1.

[41] R Development Core Team. 2007. *R: a language and environment for statistical computing.* Vienna, Austria. http://www.R-project.org.

[42] Redelings, B. D. & Suchard, M. A. 2005. Joint Bayesian estimation of alignment and phylogeny. *Syst. Biol.* 54: 401–418.

[43] Robinson, D. L. & Foulds, L. R. 1981. Comparison of phylogenetic trees. *Math. Biosc.* 53: 131–147.

[44] Ronquist, F. & Huelsenbeck, J. P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19: 1572–1574.

[45] Skorupa, L. A. 1998. Three new species of *Pilocarpus* Vahl (Rutaceae) from Brazil. *Novon* 8: 447–454.

[46] Skorupa, L. A. & Pirani, J. R. 2004. A new species of *Pilocarpus* (Rutaceae) from northern Brazil. *Brittonia* 56: 147–150.

[47] Stanford, A. M., Harden, R. & Parks, C. R. 2000. Phylogeny and biogeography of *Juglans* (Juglandaceae) based on *mat*K and ITS sequence data. *Amer. J. Bot.* 87: 872–882.

[48] Suchard, M. A. & Redelings, B. D. 2006. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics* 22: 2047–2048.

[49] Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucl. Acids Research* 24: 4876–4882.

[50] Wallace, I. M., O'Sullivan, O. & Higgins, D. G. 2005. Evaluation of iterative alignment algorithms for multiple alignment. *Bioinformatics* 21: 1408–1414.

Apêndices

A CheckGB.pl

```
#!/usr/bin/perl
# #
# # Works under Linux/Unix only!.
#
# To be used after phred/phrap/consed programs and 'Phred20.pl' script
#      (which can be found under 'consensuses' folders)
# #
# # Written for:
# # 1. concatenating all seq files into a big file
# # 2. performing Blast searches for every single sequence
# # 3. summarizing (top 10 hits) Blast information for every single sequence
# # 4. performing multiple alignments with all sequences
# # 5. building a nexus file with a PAUP block
# # 6. performing (dummy) phylogenetic searches
# #
# Warnings
# # 1. blastall, clustalw, and paup should be on your path
# # 2. this program must be in the same folder as *.phred20 files
# #
# # PDias, September 22, 2007
# #
# # Licence: GPL 3 is assumed
# # (visit: http://www.gnu.org/licenses/gpl.html)
#
# Part 1 - concatenating
#
unless (($lsr = `ls -l ./blast/*`) ne "") {`mkdir ./blast/`};
`rm ./blast/*`;
#
$lsresult = `ls -l ./`;
@lines = split (/\n/, $lsresult);
$taxon = "pilocarpinae";
foreach $line (@lines) {
  chomp($line);
  if ($line =~ /\.phred20/g) {
    @lineparts = split (/[\t\s]+/, $line);
    $file = $lineparts[-1];
    ($name, $gene, $ext) = split (/\./, $file);
    open (IN, $file)  || die "\n\n\nYou killed me ...IN\n\n";
    open (GENE, ">>./blast/$taxon.$gene.seq") || die "\n\n\nYou killed me again...GENE\n\n";
    while ($entry = <IN>) {
      chomp($entry);
      if ($entry =~ />/g) {
        $header = $entry;
      }else {
        $seq = $entry;
      }
    }
      print "$header\n$seq\n\n";
      print GENE "$header\n$seq\n\n";
    close (IN, GENE);
  }
}
#
# Part 2 - performing Blast searches
#
# Warning: local Blast is much faster, so it is used
#
`blastall -p blastn -d /usr/local/bioinf/blast/blast/data/nt -i ./blast/$taxon.$gene.seq -o
  ./blast/$taxon.$gene.blastted`;
#
# Part 3 - summarizing Blast results - looking on the top 10 hits
#
$blastted = "./blast/$taxon.$gene.blastted";
$/="BLASTN";
open (IN_SUM, $blastted) || die "\n\nNo '$blastted' file.\n\n";
open (OUT_SUM, ">$blastted.sum") || die "\n\nNo '$blastted.sum' file.\n\n";
```

```
while ($entry_sum = <IN_SUM>) {
  chomp ($entry_sum);
  @lines_sum = split (/\n/, $entry_sum);
  foreach $line_sum(@lines_sum) {
    $line_sum = "$line_sum\n";
  }
  @header_sum = @lines_sum[8 .. 9, 19 .. 20];
  $pdias_warning = "(Top 10 only - for detailed information see the '$taxon.$gene.blastted' file)";
  @top_ten = @lines_sum[21 .. 31];
  print "@header_sum", "$pdias_warning", "@top_ten\n\n\n";
  print OUT_SUM "@header_sum", "$pdias_warning", "@top_ten\n\n\n";
}
close (IN_SUM, OUT_SUM);
#
# Part 4 - (dummy) Phylogenetic analyses - e.g., to see the amount of
# parsimony-informative characters, measure distances, anything else your mind
# would allow you to do with PAUP (or any other program)
#
unless (($lsr = `ls -l ./paup/*`) ne "") {`mkdir ./paup/`};
`rm ./paup/*.*`;
`cp ./blast/$taxon.$gene.seq ./paup/$taxon.$gene.seq`;
# #
# A. Performing multiple alignment with ClustalW and saving a nexus file
#
`clustalw -INFILE=./paup/$taxon.$gene.seq -OUTPUT=NEXUS -OUTORDER=INPUT -OUTPUTTREE=nexus`;
#
# B. Building a PAUP block to automate PAUP searches
#
# Warning: you should be able to understand and modify anything between 'BEGIN PAUP' and 'END;'
#
`cp ./paup/*.nxs ./paup/$taxon.$gene.block.nxs`;
#
open (PAUP_BLOCK, ">>./paup/$taxon.$gene.block.nxs");
$block =
"BEGIN PAUP;
  CD ./paup/;
  SET CRIT=P INCREASE=A AUTOINC=200 TORDER=A AUTOCLOSE=Y TAXLABELS=F VISNOTIFY=NONE ERRORSTOP=Y
    WARNRESET=N STATUS=N CHECKEVTS=N ALLDIGLAB=NOW WARNREDEF=N;
  LOG FILE= pilocarpinae_its_block.log REPLACE;
  SHOWM;
  CSTATUS;
  TSTATUS;
  EXPORT FILE=pilocarpinae_its_block_matrix FORMAT=NEXUS REPLACE=YES;
  EXCLUDE GAPPED;
  SHOWM;
  CSTATUS;
  TSTATUS;
  EXPORT FILE=pilocarpinae_its_block_matrix_no_gap FORMAT=NEXUS REPLACE=YES;
  OUT Z_rhoifolium_ITS;
  TSTATUS;
  HS NR=1000 ADDS=R SWAP=T;
  [HS NR=10 ADDS=R SWAP=T;[!JUST A TEST]]
  ROOTTREES;
  SAVET FORMAT=N ROOT=Y BRLENS=Y FILE=hs_trees.tre REPLACE=Y;
  CONTREE / MAJ=Y TREEFILE=hs_contrees.tre;
  BOOT SEARCH=H NR=100 /NR=10 ADD=R SWAP=T;
  ROOTTREES;
  SAVET FORMAT=N ROOT=Y SAVEBOOTP=BOTH MAXDECIMALS=2 FROM=1 TO=1 FILE=boot_trees_both.tre REPLACE=Y;
  SAVET FORMAT=N ROOT=Y SAVEBOOTP=NODE MAXDECIMALS=2 FROM=1 TO=1 FILE=boot_trees_node.tre REPLACE=Y;
  SAVET FORMAT=N ROOT=Y SAVEBOOTP=BRL MAXDECIMALS=2 FROM=1 TO=1 FILE=boot_trees_brl.tre REPLACE=Y;
  LOG STOP;
END;
";
print PAUP_BLOCK $block;
close (PAUP_BLOCK);
#
# C. Calling PAUP to do the hard work
#
`paup -r ./paup/$taxon.$gene.block.nxs -l $taxon.$gene.block.nxs.log`;
exit;
```

B ConToNex.pl

```
#!/usr/bin/perl
# #
# # Works under Linux/Unix only!.
#
# Warnings:
# # 1. this script assumes a binary tree
# # 2. it is used to translate MrBayes' consensus tree files into standard NEXUS tree files
# #
# #
# # PDias, September, 2007
# #
# # Licence: GPL 3 is assumed
# # (visit: http://www.gnu.org/licenses/gpl.html)
#
$lsresult = `ls -l ./`;
@lines = split (/\n/, $lsresult);
foreach $line (@lines) {
  chomp($line);
  if ($line =~ /\.con$/g) {
    @lineparts = split (/[\t\s]+/, $line);
    $file = $lineparts[-1];
    $files .= "\n\t$file";
    open (IN, $file)  || die "\n\n\nYou killed me ...IN\n\n";
    open (TEMP, ">$file.temp") || die "\n\n\nYou killed me again...TEMP\n\n";
    while (<IN>){
      s/\r\n|\r/\n/g;
      print TEMP;
    }
    close(IN,TEMP);
    open(IN2, "$file.temp") || die "\n\n\nYou killed me again...TEMP2\n\n";
    open (T, ">$file.tre") || die "\n\n\nYou killed me again...TREE\n\n";
    $LineCounter = "";
    @TaxonNames = ();
    $TreeCounter = "";
    $AllTrees = "";
    $TranslateTable = "";
    while (<IN2>){
      if (/^[\s|\t]+tree[\s|\t]+/gi) {
        chomp;
        s/^[\s|\t]+//g;
        @Branches = ();
        $TreeCounter ++;
        $UpdtTreeDescription = "";
        ($Name,$TreeDescription) = split(/\=/,$_);
        ($TreeCmd, $TreeName) = split (/[\s|\t]+/,$Name);
        @Branches = split (/\:/, $TreeDescription);
        foreach $Branch(@Branches) {
          if ($Branch =~ /\)/g) {
            ($BranchLength,$BranchSupport) = split (/\)/,$Branch);
            if($BranchSupport !~ /\;/g && $BranchSupport ne "") {
              $UpdtBranchSupport = ($BranchSupport * 100);
              $Branch = "$BranchLength\)$UpdtBranchSupport";
            }
          }
          $UpdtTreeDescription .= "$Branch\:";
          if ($TreeCounter < 2) {
            $Branch =~ s/[\(|\|\s|\t)]+//g;
            if ($Branch =~ /\,/g) {
              ($BranchLength, $Terminal) = split (/\,/, $Branch);
              if ($Terminal =~ /[a-z]+/gi) {
                push(@TaxonNames,$Terminal);
              }
            }else {
              if ($Branch =~ /[a-z]+/gi) {
                push(@TaxonNames,$Branch);
              }
            }
```

```
          }
        }
        $AllTrees .= "\t$Name \= $UpdtTreeDescription\n";
      }
    }
    $TerminalCounter = "";
    @TaxonNames = sort(@TaxonNames);
    foreach $TaxonName(@TaxonNames) {
      $TerminalCounter ++;
      if ($TaxonName ne $TaxonNames[-1]) {
        $Translate = "\t\t$TerminalCounter '$TaxonName'\,\n";
      }else {
        $Translate = "\t\t$TerminalCounter '$TaxonName'\n\t\t\;\n";
      }
      $TranslateTable .= $Translate;
      if ($AllTrees =~ /$TaxonName/gi) {
        $AllTrees =~ s/$TaxonName/$TerminalCounter/g;
      }
    }
    $AllTrees =~ s/\;\:/\;/g;
    $Header = "\#NEXUS\n\nBegin trees\;\n\tTranslate\n";
    print T "$Header$TranslateTable$AllTrees";
    print T "end\;";
    close(IN2,T);
    unlink ("$file.temp");
  }
}
exit;
```

C HPDTrees.pl

```
#!/usr/bin/perl
#
# # Works under Linux/Unix only!.
#
# This program needs MrBayes's output files '.p' and '.t';
# It can be used to :
#   1. put the posterior probability of the tree (from '.p' file) into
#   the tree file ('.t' file) as a comment for each post-burnin tree
#   2. select trees according to specified quantiles based on their PP
#   3. select trees according to their PP and build credibility intervals;
#
# IMPORTANT NOTE: this code is a bit strange, but works!
#
# # PDias, September ??, 2007
# #
# # Licence: GPL 3 is assumed
# # (visit: http://www.gnu.org/licenses/gpl.html)
#
use Time::localtime;
use Statistics::Descriptive;
use Time::HiRes qw(gettimeofday);
$stat = Statistics::Descriptive::Full->new();
$Statistics::Descriptive::Tolerance = 1e-32;
#
$log = "";
#
$tm = localtime;
chomp($SystemDate = `date`);
print "\n\n$SystemDate\n";
$log .= "\n\n$SystemDate\n";

print "\n\nNumber of generations to be used as burnin: ";
$log .= "\n\nNumber of generations to be used as burnin: ";
chomp($Jump = <>);
$log .= $Jump;
#
print "\n\nSample frequency used in MrBayes: ";
$log .= "\n\nSample frequency used in MrBayes: ";
chomp($SampleFreq = <>);
$log .= $SampleFreq;
#
$Burnin = ($Jump * $SampleFreq);
#
$BurninLineFractionP = ($Jump + 3) ;
#
print "\n\nValue in \% for the CI (0.01/0.1/0.2/0.3/0.4/0.5/0.75/0.9/0.95/0.99): ";##
$log .= "\n\nValue in \% for the CI (0.01/0.1/0.2/0.3/0.4/0.5/0.75/0.9/0.95/0.99): ";##
#
chomp($IcValue = <>);
$log .= $IcValue;
#
$t1 = gettimeofday;
#
if ($IcValue eq /0.01|0.1|0.2|0.25|0.3|0.4|0.5|0.75|0.9|0.95|0.99/) {
  die "\nYour values don't make sense! Try again.\n";
  $log .= "\nYour values don't make sense! Try again.\n"; ;
}
if ($IcValue eq "0.01") {
  $Percentile = "49.5 49.5";
}if ($IcValue eq "0.1") {
  $Percentile = "45 55";
}if ($IcValue eq "0.2") {      #
  $Percentile = "40 60";  #
}if ($IcValue eq "0.25") {
  $Percentile = "37.5 62.5";
}if ($IcValue eq "0.3") {
  $Percentile = "35 65";
}if ($IcValue eq "0.4") {
```

```
  $Percentile = "30 70";
}if ($IcValue eq "0.5") {
  $Percentile = "25 75";
}if ($IcValue eq "0.75") {      #
  $Percentile = "12.5 87.5";  #
}if ($IcValue eq "0.9") {
  $Percentile = "5 95";
}if ($IcValue eq "0.95") {
  $Percentile = "2.5 97.5";
}if ($IcValue eq "0.99") {
  $Percentile = "0.5 99.5";
}
#
open (LOG, ">probintotree.b$Jump.ci$IcValue.log") || die "\n\n\nYou killed me again...LOG\n\n";
#
print "\n\nPlease wait, it may take some minutes...\n\t(maybe hours, if your runs have a lot of
  generations)";
$log .= "\n\nPlease wait, it may take some minutes...\n\t(maybe hours, if your runs have a lot of
  generations)";
#
@lsresult = `ls -l ./`;
#
foreach $line (@lsresult) {
  chomp($line);
  if ($line =~ /(run\d+)\.treewithprob$/gi) {
    $TreeWithProbRun = $1;
    $ConcatTFileCounter ++;
    @linepartsT = split (/[\t\s]+/, $line);
    $fileT = $linepartsT[-1];
    $linepartsT[-1] =~ s/\.treewithprob//gi;
    $FileTIN = $linepartsT[-1];
    push(@FilesTIN, $FileTIN);
  }
  if ($line =~ /(run\d+)\.p$/gi) {
    $PFileRun = $1;
    $ConcatPFileCounter ++;
    @linepartsP = split (/[\t\s]+/, $line);
    $fileP = $linepartsP[-1];
    $linepartsP[-1] =~ s/\.p//gi;
    $FilePIN = $linepartsP[-1];
    push(@FilesPIN,$FilePIN);
  }
}
#
if ($ConcatTFileCounter eq "") {
  die "\n\nThere is no '.treewithprob' file in this directory.\n\n"
}if ($ConcatPFileCounter eq "") {
  die "\n\nThere is no '.p' file in this directory.\n\n"
}
#
# **********************
# Credibility interval **
# **********************
#
foreach $FilesTIN(@FilesTIN) {
print "\n\nComputing quantiles ... Please wait...";
$log .= "\n\nComputing quantiles ... Please wait...";
#
open (PERCENTILE, $fileT)  || die "\n\n\nYou killed me again...PERCENTILE\n\n";
open (CIT, ">$FilesTIN.b$Jump.ci$IcValue.t")  || die "\n\n\nYou killed me again...CIT\n\n";
open (CIP, ">$FilesTIN.b$Jump.ci$IcValue.p")  || die "\n\n\nYou killed me again...CIP\n\n";
#
@TreeLnLAbsValue = "";
$entryPercCounter = "";
while ($entryPerc = <PERCENTILE>) {
  chomp($entryPerc);
  if ($entryPerc =~ /^tree run/gi) {
    ($Header, $Lnl, $TreeDescription) = split (/\=/, $entryPerc);
    ($SemiHeader,$RepNumber) = split (/\./, $Header);
    $RepNumber =~ s/\s//g;
    $RepNumber =~ s/\[TreeLnL//gi;
```

```
    if ($RepNumber > $Burnin) {
      $entryPercCounter ++;
      $entryPerc =~ /\[TreeLnL \= (\-\d+\.?\d+?)\]/gi;
      push(@TreeLnLAbsValue, $1);
      $Sum += $1;
    }
  }
}
$CiInferiorValue = "0";
$CiSuperiorValue = "0";
#
@OrderedTreeLnLAbsValue = sort { $a <=> $b } @TreeLnLAbsValue;
$lenghtTreeLnLAbsValue = @TreeLnLAbsValue;
$stat->add_data(@OrderedTreeLnLAbsValue);
$mean = $stat->mean();
$var  = $stat->variance();
$median = $stat->median();
$harmonic_mean = $stat->harmonic_mean();
$geometric_mean = $stat->geometric_mean();
$mode = $stat->mode();
$frequency_distributionP = $stat->frequency_distribution($partitions);
$frequency_distributionB = $stat->frequency_distribution(\@bins);
$frequency_distribution = $stat->frequency_distribution();
$least_squares_fit = $stat->least_squares_fit();
$least_squares_fitX = $stat->least_squares_fit(@x);
#
($QuantileInferior, $QuantileSuperior) = split (" ", $Percentile);
($x, $index) = $stat->percentile(($QuantileInferior));###
#
if ($x eq "") {
  $log .= "\nQuantile inferior '$QuantileInferior' for file '$FilesTIN' includes no tree!
    Your distribution has problems... Quitting :-(\n";
  print LOG $log;
  close (CIT, CIP);
  unlink ("$FilesTIN.$IcValue.ci.p", "$FilesTIN.$IcValue.ci.t");
  warn "\nQuantile inferior '$QuantileInferior' for file '$FilesTIN' includes no tree!
    Your distribution has problems... Quitting :-(\n";
}
$CiInferiorValue = $x;
#
($x, $index) = $stat->percentile(($QuantileSuperior));###
#
if ($index eq "") {
  $log .= "\nQuantile superior '$QuantileSuperior' for file '$FilesTIN' includes no tree!
    Your distribution has problems... Quitting :-(\n" ;
  print LOG $log;
  close (CIT, CIP);
  unlink ("$FilesTIN.$IcValue.ci.p", "$FilesTIN.$IcValue.ci.t");
  warn "\nQuantile superior '$QuantileSuperior' for file '$FilesTIN' includes no tree!
    Your distribution has problems... Quitting :-(\n" ;
}
$CiSuperiorValue = "$x";
#
close (PERCENTILE);
#
open (PERCENTILET, "$FilesTIN.treewithprob")  || die "\n\n\nYou killed me again...PERCENTILE2\n\n";
open (PERCENTILEP, "$FilesTIN.p")  || die "\n\n\nYou killed me again...PERCENTILE2P\n\n";
#
print "Done!";
$log .= "Done!";
#
print "\n\nSome important values about your data:";
print "\n\tNumber of trees: $entryPercCounter";
print "\n\tNumber of samples: $lenghtTreeLnLAbsValue ";
print "\n\tMean: $mean";
print "\n\tMedian: $median";
print "\n\tMode: $mode";
print "\n\tVariance: $var";
print "\n\tHarmonic mean: $harmonic_mean";
print "\n\tGeometric mean: $geometric_mean";
print "\n\tInferior quantile: $CiInferiorValue";
```

```
print "\n\tSuperior quantile: $CiSuperiorValue\n";
#
$log .= "\n\nSome important values about your data:\n\tNumber of trees:
  $entryPercCounter\n\tNumber of samples: $lenghtTreeLnLAbsValue\n\tMean: $mean\n\tMedian:
  $median\n\tMode: $mode\n\tVariance: $var\n\tHarmonic mean: $harmonic_meanp\n\tGeometric mean:
  $geometric_mean\n\tInferior quantile: $CiInferiorValue\n\tSuperior quantile:
  $CiSuperiorValue\n";
#
open (ORDERED_POST_BURNIN, ">$FilesTIN.b$Jump.OrderedPP.txt") || die "\n\nYou killed me...
  ORDERED_POST_BURNIN\n\n";
  foreach (@OrderedTreeLnLAbsValue) {
    if ($_ ne @OrderedTreeLnLAbsValue[-2] || $_ ne @OrderedTreeLnLAbsValue[-1]) {
      print ORDERED_POST_BURNIN "$_\, ";
    }else {
      print ORDERED_POST_BURNIN "$_";
    }
  }
#
if ($IcValue eq "0.5") {
  $CiInferiorValue = "-7414.804000000000";
  $CiSuperiorValue = "-7405.46332024622";
}if ($IcValue eq "0.75") {      #
  $CiInferiorValue = "-7418.452774139952";
  $CiSuperiorValue = "-7402.36900000000";
}if ($IcValue eq "0.9") {
  $CiInferiorValue = "-7422.402100428964";
  $CiSuperiorValue = "-7399.32780000000";
}if ($IcValue eq "0.95") {
  $CiInferiorValue = "-7425.211736056476";
  $CiSuperiorValue = "-7397.52262316598";
}if ($IcValue eq "0.99") {
  $CiInferiorValue = "-7430.191111664876";
  $CiSuperiorValue = "-7394.31696750681";
}
#
print "\n\nFiltering trees of the $IcValue credibility interval for $FilesTIN... ";
$log .= "\n\nFiltering trees of the $IcValue credibility interval for $FilesTIN... ";

print "\n\tTrees with $CiInferiorValue <= prob <= $CiSuperiorValue will be included.";
$log .= "\n\tTrees with $CiInferiorValue <= prob <= $CiSuperiorValue will be included.";
#
$entryPercCounterTDumb = "";
while ($entryPercT = <PERCENTILET>) {
  chomp($entryPercT);
  if ($entryPercT =~ /^tree/gi) {
    ($Header, $Lnl, $TreeDescription) = split (/\=/, $entryPercT);
    ($SemiHeader,$RepNumber) = split (/\./, $Header);
    $RepNumber =~ s/\s//g;
    $RepNumber =~ s/\[TreeLnL//gi;
    if ($RepNumber > $Burnin) {
      $TreeLnLAbsValue2 = "";
      $entryPercT =~ /\[TreeLnL \= (\-\d+\.?\d+?)\]/gi;
      $TreeLnLAbsValue2 = $1;
      if ($CiInferiorValue <= $TreeLnLAbsValue2 && $CiSuperiorValue >= $TreeLnLAbsValue2 ) {
        $entryPercCounterTDumb ++;
        print CIT "$entryPercT\n";
      }
    }
  }else {
    print CIT "$entryPercT\n";
  }
}
print "\n\t$entryPercCounterTDumb trees saved to '$FilesTIN.b$Jump.ci$IcValue.t'.\n\t...Done!";
$log .= "\n\t$entryPercCounterTDumb trees saved to '$FilesTIN.b$Jump.ci$IcValue.t'.\n\t...Done!";
#
print "\n\nFiltering tree probs of the $IcValue\% credibility interval for $FilesTIN...";
$log .= "\n\nFiltering tree probs of the $IcValue\% credibility interval for $FilesTIN...";
#
print "\n\t$CiInferiorValue <= prob <= $CiSuperiorValue will be included.";
$log .= "\n\t$CiInferiorValue <= prob <= $CiSuperiorValue will be included.";
#
```

```
$entryPercCounterPDumb = "";
while ($entryPercP = <PERCENTILEP>) {
  chomp($entryPercP);
  $FieldsPAbsValue = "";
  @entryPercPFields = split (/\t/, $entryPercP);
  if ($entryPercPFields[0] =~ /Gen/g) {
    print CIP "$entryPercP\n";
  }

  if ($entryPercPFields[0] =~ /^(\d+)/gi) {
    if ($entryPercPFields[0] > $Burnin) {
      if ($CiInferiorValue <= $entryPercPFields[1] && $CiSuperiorValue >= $entryPercPFields[1]) {
        $entryPercCounterPDumb ++;
        print CIP "$entryPercP\n";
      }
    }
  }
}
#
print "\n\t$entryPercCounterPDumb probs saved to '$FilesTIN.b$Jump.ci$IcValue.p'.\n\t...Done!";
$log .= "\n\t$entryPercCounterPDumb probs saved to '$FilesTIN.b$Jump.ci$IcValue.p'.\n\t...Done!";
#
close (CIT,CIP.PERCENTILET,PERCENTILEP);
#
}
#
$t2 = gettimeofday;
$elapsed = ($t2 - $t1);
#
print "\n\nTime used:  $elapsed seconds";
$log .= "\n\nTime used:  $elapsed seconds";
#
print "\n\n(<Enter> key to exit)";
$log .=  "\n\n(<Enter> key to exit)";
print LOG $log;
close(LOG);
#
exit unless (($quit = <>) =~ "");
```

D Matriz

```
#NEXUS

BEGIN DATA;
  DIMENSIONS  NTAX =30 NCHAR=1879;
  FORMAT DATATYPE = DNA GAP = - MISSING = ?;
  MATRIX

B_molle
----------------------GTCGCGATG-CGAGCGCCGCAG-ATGCGGAGCGTCAGGGTCCCTGAG-TCCCGAAACGGAGACTTCGGCACGCGACAT
GA-ACTCGAGAGGCTTGTTTT-CACCACCGATAGTCGCGGCCTCAGTCGTCGAGGACTCAAATTTGGGCCAACCGCGAGCG--GGAG-CGCACGGGAGGC
CATTCTCCGCCCGCACCGCCAGGCCC---CAATGG-TAAGGGGTGGGG-TGGGGCAACGATGCGTGACACCCAGGCAGACGTGCCCTCGGCCTAACGGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGATA---CAGTGAAAGAAGGCGTCGCGTCCCGAGGCACACCGTG-TCCGGGGCCCC-TGGAGCGTGCTCTCTCG
TTAC-AT-TTCCTTGGCGCATTCCGCGCCGGGGGTTCGTTGTCCGCCG---CAGGGAACG-GGAC----G--AGTCCCGCAC-CATTGGCGGCAGGGGGA
GTAGCACCCGTGG-GCGCGCCCCC--CCGGTGTTTT-AACATGTTCGCGGGTCGTTCTGCTAG-GCAGG--TTTCGACAAT-GATCCTTCCGCANNTTC-
ACCTACGNAA----------A----------TCGGATGTGATCAAGATAGAACAGAAATTCAAAAG
AATCATGAAATGCAAATTCGAAATTAGAACGTTGACGTCTTTGTCAGGAGTCCTA--ATGCAATTAGG------AAAGGAGGAGTTATTTCGTTTAAATA
ACTAAATTGAATAATTAACAATTTAATA---ATA--------------AAAAACAAACTCTTTTGGTGGACGAATTTTGACAGATATGGCTCGACAAAAC
AACTTAAGTCATAAGACAAAAAA--AACAAGTGGATTGTGAAAA-AATCCCTAGTTCCATTTTTCCTTTTTTTT--T-TCAGTATTTTTTTTTTCNNGTA
AAAGTCAAATAA--ATGGAAATACAAAAAA-A--------AGAA-ACGAAAGAATAATNAA---------------AAATNANNNTTTGATTNNAATTCT
ATATCNNNNNNNGANNNGNAATAGTCCTTCTTTTTCTTGTTCTTTGAATACAATA--ACAAGTATTTC----ATTTTTGGATTTTCAAATCAAATCCAAA
TAAATCATTTTTGTTTATGTTATGGTTCGCATTTTTTCTATGGTTTGTCGGTACGGTTCATTAGAACAAAAAA---AGGCCCGGCTGGGTACTGACCAGG
CCAGGCCGTCGAAGTGGAAATAAAAAAGGCCCCGTT----GAA---CGAAATTAAAGAGATATTCTTTTCTTTAGTTTTTT-------------CTATTT
TATCTCTTTCGAATGCTTTCTGTCTTTAAATTTTCTATAAATGANNNAAAAGGAA------T-----TG------CTGATAAAAGTTA-GATCTATTT--
-----ATATAA----------TAGAATCCAAAAGACAATAAAAAAA--A--------TATNNTCTAT----A---AGCTATATTTTNNTTGA----GCTA
TATAAT-------CAAATTACTTT---CTAGAT-------T--CTCTAN---------NNTATAGAG-AA--TCTAG-----AAAGTA--AAG-------
--ANNNAA------------A----------
B_riedelianum
--------------------GGGTCGCGATG-CGAGCGCCGCAG-ATGCGGAGCGTCAGGGTCCCTGAG-TCCCGAAACGGAGACTTCGGCACGCGACAT
GA-ACTCGAGAGGCTTGTTTT-CACCACCGATAGTCGCGGCCTCAGTCGTCGAGGACTCAAATTTGGGCCAACCGCGAGCG--GGAG-CGCACGGGAGGC
CATTCTCCGCCCGCACCGCCAGTCCC---CAATGG-TAAGGGGTGGGG-TGGGGCAACGATGCGTGACACCCAGGCAGACGTGCCCTCGGCCTAACGGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGATA---CAGTGAAAGAAGGCGTCGCGTCCCGAGGCACACCGTG-TCCGGGGCCCC-TGGAGCGTGCTCTCTCG
TTAC-AT-TTCCTTGGCGCATTCCGCGCCGGGGGTTCGTTGTCCGCCG---CAGGGAACG-GGAC----G--AGTCCCGCAC-CATTGGCGGCAGGGGGA
GGAGCACCCGTGG-GCGCGCCCCCCCGCCGGTGTTTT-AACATGTTCGCGGGTCGTTCTGCTAG-GCAGG--TTTCGACAAT-GATCCTTCCGCAGGTTC-
ACCTACGGAA----------A----------ATCGGATGTGATCAAGATAGAACAGAAATTCAAAAG
AATCATGAAATGCAAATTCGAAATTAGAACGTTGACGTCTTTGTCAGGAGTCCTA--ATGCAATTAGG------AAAGGAGGAGTTATTTCGTTTAAATA
ACTAAATTGAATAATTAACAATTTAATA---ATA--------------AAAAACAAACTCTTTTGGTGGACGAATTTTGACAGATATGGCTCGACAAAAC
AACTTAAGTCATAAGACAAAAAAA-AACAAGTGGATTGTGAAAA-AATCCCTAGTTCCATTTTTCCTTTTTTTT--T--CAGTATTTTTTT-TTCCAGTA
AAAGTCAAATAA--ATGGAAATACAAAAAAAA---------GAA-ACGAAAGAATAATANN---------------AAATGACACTTTGATTCGAATTCT
ATATCNNNATAGGAATGGAAATAGTCCTTCTTTTTCTTGTTCTTTGAATACAATA--ACAAGTATTTC----ATTTTTGGATTTGCAAATCAAATCCAAA
TAAATCATTTTTGTTTATGTTATGGTTCGCATTTTTTCTATGGTTCGTCGGTACGGTTCATTAGAACAAAAAA---AGGCCCGGCTGGGTACTGACCAGG
CCAGGCCGTCGAAGTGGAAATAAAAAAGGCCCCGTT----GAA---CGAAATTAAAGAGATATTCTTTTCTTTAGTTTTTT-------------CTATTT
TATCTCTTTCGAATGCTTTCTGTCTTTAAATTTTCTATAAATGATAGAAAAGGAA------T-----TG------CTGATAAAAGTTA-GATCTATTT--
-----ATATAA----------TAGAATCCAAAAGACAATAAAAAAA--AA--------TATTTNNTAT----A---AGCTATATTTTNTANGA----GCTA
TATAAT-------CAAATTACTTT---CTAGAT-------T--CTCTAN---------TTTATAGAG-AA--TCTAG-----AAAGTA--AAG-------
--ATCNAA--------A-ATA----------
E_almawillia
--------------------GGGTCGCAAAG-CGAGCACCGCAG-ATGCGGGGCGTCAGGGTCC-TGAG-TCCTAAAACGATGACTCCGGCACGCGACGT
GA-GCTCGAGAGGCTTGTTTT-CACCACCGATAGTCGCGGCCTCAGTCGATGAGGACTAGTATTTGGACCAACCGCGAGCG--GGAATCTCACGGGAGGC
CATTCTCCGCCCGCACCACCAGGGCC---CGATGGGTAAGGGGT------GGGGCAACGATGCGTGACACCCAGGCAGACGTGCCCTCGGCCTAACGGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGATA---CAGTGAAAGAAGGCACTGCGTCCTTAGGCACACCGTG-TCCGGGGCCCT-CGGAGCATGCTCTCTTG
TTGA-AT-TTCCTTGGCGCATTCCGCGCCGGGGGTTCGTTGTCCGCGG---CAGGGAACG-GGAC----A--AGTCCCGCAC-CTGAGGCGGCAGGTGGA
GGAGTGCCCGTGG-GCGCGCCCCC-ACCGATATTTT-AACATGTTCGCGGGTCTTTCTGCTAG-GCAGG--TTTCGACAAT-GATCCTTCCGCAGNTTC-
ACCT----------ATCAAGATAGAACAGAAATTCAAAAG
AATCATGAAATGCAAATTCGAAATTAGAACATTGACGTCTTTGTCAGGAGTCCTA--ATGCAATTAGG------AAAGGAGGAGTTTTTTCGTTTAAATA
ACTAAATTGAATAATTAACAATTTAATA---ATA--------------AAAAACAAA-----T-GGTGGACGAATTTTGACAGATATGGCTCGACAAAAC
AACTTAAGTCATAAGACAAAAAAAAAACAAGTGGATTGTGAAAA-AATCCCTAGTTCCATTTTTCCTTTTTTTT--T--CAGTATCTT-TT-TTCCAGTA
AAAGTCAAATAA--ATTGAAATACAAAAAAAA--------AGAA-ACGAAAGAATAATAAA---------------AAATGACACTTTGATTCGAATTCT
ATATCNTTATAGGANNGGAAATAGTCCTTCTTTTTCTTGTTCTTTGAATACAAT--------TATTTC----ATTTTTGGGTTTTCAAATCAAATCCAAA
AAAATCATTTTTGTTTATGTTATGGTTCGCTTTTTTTCTATGGTTCGGCGGTACGGTTCATTAGAACAAAAAA---AGGCCCGGCTGGGTACTGACCAGG
CCAGGCCGTCAAANNNGAAATAAAAAAGGCCCCGTT----GAA---CGAAATTAAAGAGATATTCTTTTCTTTAGTTTTTT-------------CTATTT
TATCTCTTTCTAATGCTTTCTGTCTTTAAATTTTCTATAAATGATAGAAAAGGAA------T-----TG------CTGATAAAAGTTA-GATCTATTT--
-----ATATAA----------TAGAATCCAAAAGCCAATAAAAAAAAAA--------TATNNTCTAT----A---AGCTATATTTTATATGA----GCTA
TATAAT-------CAAATTACTTT---CTAGAT-------T--CTCTAT---------ATTATAGAG-AA--TCTAG-----AAAGTA--AAG-------
----------
```

```
E_cowanii
--------------------GGGTCGCAAAG-CGAGCACCGCAN-NTGCGGGGCGTCAGGGTCC-TGAG-TCCCAAAACGATGACTCCGGCACGCGACGT
GA-GCTCGAGAGGCTTGTTTT-CACCACCGATAGTCGCGGCCTCAGTCGACGAGGACTAGAATTTGGGCCAACCGCGAAAG--GGATTCTCACGGGAGGC
CATTCTCCGCCCGCACCACCAGGGCC---CGATGGGTAAGGGGT------GGGGCAACGATGCGTGACACCCAGGCAGACGTGCCCTCGGCCTAACGGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGATA---CAGTGAAAGAAGGCACCGCGTCCTTAGGCACACCGTG-TCCGGGGCCCT-CGGAGCGTGCTCTCTCG
TTGC-AT-TTCCTTGGCGCATTCCGCGCCGGGGGTTCGTTATCCGCAA---CAGGGAACG-GGAC----A--AGTCCCGCAC-CCGAGGCGGCAGGCGGA
GGAGCGCCCGTGG-GCGCGCCCCC-ACCGAT-TTTT-AACATGTTCGCGGGTCTTTCTGCTAG-GCAGG--TTTCGACAAT-GATCCTTCCGCAGGTTC-
ACCTACGNNNA-----------------------------------------A----GA----GAATTGGATGTAATCAAGATAGAACAGAAATTCAAAAG
AATCATGAAATGCAAATTCGAAATTAGAACATTGACGTCTTTGTCAGGAGTCCTA--ATGCAATTAGG------AAAGGAGGAGTTTTTTCGTTTAAATA
ACTAAATTGAATAATTAACAATTTAATA---ATA--------------AAAAACAAA-----T-GGTGGACGAATTTTGACAGATATGGCTCGACAAAAC
AACTTAAGTCATAAGACAAAAAAA-AACAAGTGGATTGTGAAAA-AATCCCTAGTTCCATTTTTCCTTTTTTTT--T--CAGTATCTT-TT-TTCCAGTA
AAAGTCAAATAA--ATTGAAATACAAAAAAAA--------AGAA-ACGAAAGAATAATAAG--------------AAATGACACTTTGATTCGAATTCT
ATATCATCATAGGAATGGAAATAGTCCTTCTTTTTCTTGTTCTTTGAATACAAT--------TATTTC----ATTTTTGGGTTTTCAAATCAAATCCAAA
AAAATCATTTTTGTTTATGTTATGGTTCGCTTTTTTTCTATGGTTCGGCGGTACGGTTCATTAGAACAAAAAA---AGGCCCGGCTGGGTACTGACCAGG
CCAGGCCGTNNAAGTGGAAATAAAAAAGGCCCCGTT----GAA---CGAAATTAAAGAGATATTCTTTTCTTTAGTTTTTT-------------CTATTT
TATCTCTTTCTAATGCTTTCTGTCTTTAAATTTTCTATAAATGATAGAAATGGAA------T-----TG------CTGATAAAAGTTA-GATCTATTT--
-----ATATAA----------TAGAATCCAAAAGCCAATAAAAAAA-AA--------TATATTCTAT----A---AGCTATATTTTATATGA----GCTA
TATAAT-------CAAATTACTTT---CTAGAT-------T--CTCTAT---------ATTATAGAG-AA--TCTAG-----AAAGTA--AAG-------
--ATCTAA--------A-ATAAA-----------------------------------------------------------
E_decidua
--------------------GNG---CA----C---CNNNGG---ATGCGGAGCGTCAGGGTCCTTGAG-TCCCGAAACGACGACTCCGGCACGCGACGT
GA-GCTCGAGAGGCTTGTTTTTCACCACCGATGGTCGCGGCCTCGGTCGTCGAGGGCTCGAATTTGGGCCAACCGCGAGCG--GGAG-CGCACGGGAGGC
CATTCTCCGCCCGCACCGCCAGGCCC---CAATGG-AAAGGGGTGGGG-TGGGGCAATGATGCGTGACACCCAGGCAGACGTGCCCTCGGCCTAACGGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGGTA---AAGTGAAAGAAGACGCCGCGTCCCGAGGCGCACCGTG-TCCGGGGCCCC-TGGAGCGTGCTCTCTCG
TTAA-GT-TTCCTTGGCGCGTTCCGCGCCGGGGGTTCGTTGTCCGTCA---CGGGGAACG-AGAC----G--AGTCCCGCAC-CCGCGACGGTAGTGAGA
GGAGCGCCCGTGG-GCACACCCCCCGCTGGTGTTTT-AACATGTTCGCGGGTCGTTCTGCTGG-GCAGG--TTTCGACAAT-GATCCTTCCGCANNTTC-
ACCT-----------------------------------------------------------------A--------GAACAGAAATTCAAAAG
AATCATGAAATGCAAATTCGAAATTAGAACGTTGACGTCTTTGTCAGGAGTCCTA--ATGCAATTAGG------AAAGGAGG----------TTTAAATA
ACTAAATTGAATAATTAACAATTTAATA---ATA--------------AAAAAAAAACTCTTTTGGTGGACGAATTTTGACAGATGTGGCTCGACAAAAC
AACTTAAGTCATAAGACTTAAAAAAAACAAGTGGATTGTGAAAA-AATCCCTAGTTCCATTTTTCCTTTTTTTTTTTTTTCAGTATTTTTTT--TCCCGTA
AAAGTCAAATAA--ATGGAAATACAAAAAAAG--------AAA---CGAAAGAATAATAAA---------------AAATGANACTTTGATTCGAATTCT
ATATNNTCCTAGGANNNGAAATAGTCCTTCTTTTTCTTGTTCTTTGAATACAATA--ACAAGTATTTC----ATTTTTGGGTTTTCAAATCAAATCCAAA
TAAATCATTTTTGTTTATGTTATGGTTCGCATTTTTTCTATGGTTTGGCGGTACGGTTCATTAGAACAAAAAA---AGGCCCGGCTGGGTACTGACCAGG
CCAAACCGTCGAAGTGGAAATAAAAAAGGCCCCGTT----GAA---CGAAATTAAAGAGATATTCTTTTCTTTAGTTTTTT-------------CTATTT
TATCTCTTTCGAATGCTTTCTGTCTTTAAATTTTCTATAAATGATAGAAANNGAA------T-----TG------CTGATAAAAGTTA-GATCTATTT--
-----ATATAAT----------AGAATCCAAAAGANAATAAAAAAA--A-T-------ANNNTCTAT----A---AGCTATATTTTNNNNGA----GCTA
TATAAT-------CAAATTACTTT---CTAGAT-------T--CTCTAT---------ATTATAGAG-AA--TCTAG-----AAAGT----A--------
--------------------------------------------------------------------------------
>E_grandiflora
--------------------GGGTCGCAATG-CGAGCACTGTAAAATACAGAGCGTCAGGGTCCCTGAG-TCCTGAAAAGGAGACTAAGGCACGCGACGT
GA-GCTCGAGAGGCTTGTTTT-CACCAACGATAGTCGCGGCCTCGGTCGTCGAGGACTCGAATTTGGGCCAACCGCGAACA--GGAG-TGCACGGGAGGC
CATTCTCTGCCCGCACCTCCAGGCCC---TAATGG-TAAGGGGTGGGG-TGGGGCAACGATGCGTGACACCCAGGCAGACGTGCCCTCGGCCTAATGGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGATA---CATTGAAAGAAGGCACCGCGTCCTGAGGCGCACCGTG-TCCGGGGCCAT-CGGAGCGTGCTCTCTCG
TGAT-AT-TTCCTTGGCGCATTCCGCGCCGGGGGTTCGTTATCCACCA---TAGGGAACG-GGAC----G--AGTCCCGCAC-CTACGGCGGTAGGGGGA
GGAACGCCCGTGG-GCGTGCCCCCCGCCAGTGTTTT-AACATGTTCACGGGTCGTTCTGCTAG-GCAGG--TTTCGACAAT-GATCCTTCCGCANNTTC-
ACCTA-----------------------------------------------CTTTTGTCGAACAAGAGAATCGGATGTAATCAAGATAGAAATTCAAAAG
AATCATGAAATGCAAATTCGAAATTAGAACGTTGACGTCTTTGTCAGGAGTCCATTGATGAGAAGGGGCTAGGGAAAGGAGGAGTTATTTCGTTTAAATA
ACTAAATTGAATAATTAACAATTTAATA---ATA--------------AAAAACAAACTCTTTTGGTGGACGAATTTTGACAGATATGGCTCGACAAAAC
AACTTAAGTCATAAGACAAAAAAA--ACGAGTGGATTGTGAAAA-AATCCCTAGTTCCATTTTTCCTTTTTTTT---TTCAGTATTTTTTTTTTCCCGTA
AAAGTCAAATAA--ATGGAAATACAAAAAAA---------AGNA-ACANAAGAATAATAAA---------------AAATGACCCTTTGATTCCANTTCT
ATATCNNNANNNGANNNNNAATAGTCCTTCTTTTTCTTGTTCTTTGAATACAATA--ACAAGTATTTCTTTCATTTTTGGGTTTTCAAATCAAATCCAAA
TAAATCATTTTTTTTTTATGTTATGGTTCNNNTTTTTTCTATGGTTCNNNNGAATTGCTGATAAAAGTTAGA------------TCTATTTATATA--AAA
--AAA---TCCAAAAAAANAATAAAAA-A--------------------ATATTTNNNATTTNNTATAAGCTATATTTNN------------------NTN
GAGCTATATAATCAAATTACTTT--CNANA-T-TCTCTNNTTTATAGAGAA-------------------------TNNNNAAAGTAAAGATCTA-----
-----AAATA-----------AAGTAAAGTAAAGACGGG----CT-------------TTTTTCAAG---------CCTNNNNNTTTNNNNN-----NNN
NNNNNN--N---NGTAATNNNTTT-----------------------------------------------------TTT-------------------------
----TTT--------------------------------------------------------------------------
E_hieronymi
--------------------GGGGTCGCAATG-CGAGCACCGCAA-ATGCGGAGCGTCAGGGCCCTTGAG-TCCCAAAACGGTGACTCCGGCACGCGACGT
GA-GCTCGAGAGGCTTGTTTT-CACCACCGATAGTCGCGGCCTCAGTCGCCGAGGACTCGAATTTGGGCCAACCGCGAGCG--GGAG-CGCACGGGAGGC
CATTCTCCGCCCGCACCACCAGGCCC---CAACGG-TAAGGGGTGGGG-TGGGGCAATGATGCGTGACACCCAGGCAGACGTGCCCTCGGCCTAATGGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGATA---AAGTGAAAGAAGGCGCCGCGTCCCGAGGCGCACCGTG-TCCGGGGCCCC-TAGAGCGTGCTCTCTCG
TTAC-AT-TTCCTTGGCGCATTCCGCGCCGGGGGTTCGTTGTCCGCCA---CGGGGAACG-AGAC----G--AGTCCNNNNNCCCGTGGCGGTAGGGGGA
GGAGTGCCCGTGG-GCACGCCCCCCGCCGGTGTTTT-AACATGTTCGCGGGTCGTTCTGCTAG-GCAGG--TTTCGACAAT-GATCCTTCCGCAGGTTC-
ACCTACGGAAAC-C-------------------------------------------------------------------------CAGAAATTCAAAAG
AATCATGAAATGCAAATTCGAAATTAGAACGTTGACGTCTTTGTCAGGAGTCCTA--ATGCAATTAGG------AAAGGAGGAGTTATTTCGTTTAAATA
ACTAAATTGAATAATTAACAATTTAATA---ATA--------------AAAAACAAACTCTTTTAGTGGACGAATTTTGACAGATATGGCTCGACAAAAC
```

```
AACTTAAGTCATAAGACAAAAAAA-A-CAAGTGGATTGTGAAAA-AATCCCTAGTTCCATTTTTCCTTTTTTTT--TTTCAGNATTTTTTT-TTCCAGTA
AAAGTCAAATAA--ATGGAAATACAAAAAAAA--------ATA--CTNAAAAAAAAAAAAAACANANNNNNNNNAAAAAATNANCCTTTGANTNANATTCT
ANNNCNNNNNNNNNNAGNNAATAGTCCTTTTTTTTTTNNTNCNTTNAATACANNN--ACAAGTATTTC----ATTTTTGGGTTTTCAAATCAAATCCAAA
TAAATCATTTTTGTTNNNGTNNAGGTTNGCATTTTTTTTTNGGTTCGGNANNNCGGTTCATTAAANCAAAAAAAAAAGGCCCGGCTGGGNNCTGACCAGG
CC-----GNNNAANNNGAAATAAAAAAGGCCCCGTT----GAA---CNAAATTAAAGAGATATNNTTTTCTTTAGTTTTTN-------------NTATTT
TATCTCTTTCNAATGCTTTCTGTCTTTAAATTTTNTATAAATGANNGAAAAGGAA------T-----TG------TTGATAAAAGTTA-GATCTATTT--
-----ATATAA----------TANAATCNNAAANNNNNTAACAAAA--A--------TATTTNNNAN----N---NGCTATATTTNNNTNNA----GCTN
NNTAAT-------CAAATTACTTT---CTAGAT-------T--CTCNAN---------NNNNTAGAG-AA--TCT-------------------------
----ANAA--------------------------------------------------------------------------
E_oligantha
GGGTAATCCCGCCTGACCTGGGGTCGCAAAG-CGAGCACCGCAG-ATGCGGAGCGTCAGGGTCCGTGAG-TCCCAAAACGATGACTTTGGCACGCGACGT
GA-GCTCGAGAGGCTT-TTTT-GACCACCGATAGTCGCGGCCTCAGTCGACCAGGACTCGAATTTGGGCCAACCACGAGCA--AGAA-CTCACGGGAGGC
CATTCTCCGCCCGCACCACCAGGCCC---TGATGGGTAAGGGGT-----TGGGGCAACGATGCGTGACACCCAGGCAGACGTGCCCTCGGCCTAACGGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGATA---TAGTGAAAGAAGGCATTGCGTCCCTAGGCTCACCGTG-TCCGGGGCCCT-TGGAGCGTGCTCTCTCG
TTAC-AT-TTCCTTGGCGCATTCCGCGCCGGGGGTTCGTTGTCCGCAG---CGGGGAACG-GGAC----G--AGTCCCACAC-CCGCCGCGGTAGGTGAA
GGAGCGCCCGTGG-GCGCGCCCCC-ACCGATATTTT-AACATGTCCGCGGGTCTTTCTGCTAG-GCAGG--TTTCGACAAT-GATCCTTCCGCAGGTTC-
ACCTA----------------------------------TGTCGAA---CAA----GA----GAATCGGATGTAATCAAGATAGAACAGAAATTCAAAAG
AATCATGAAATGCAAATTCGAAATTAGAACATTGACGTCTTTGTCAGGAGTCCTA--ATGCAATTAGG------AAAGGAGGAGTTATTTCGTTTAAATA
ACTAAATTGAATAATTAACAATTTAATA---ATA--------------AAAAACAAA-----T-GGTGGACGAATTTTGACAGATATGGCTCGACAAAAT
AACTTAAGTCATAAGACAAAAAAA-AGCAAGTGGATTGTGAAAA-AATCCCTAGTTCCATTTTTCCTTTTTTTT--T-TCAGTATTTTTTT-TTCCCNTA
AAAGTCAAATAA--ATTGAAATACAAAAAAAA--------AAAA-ACGAAAGAATAATAAA---------------AAATGACACTTTGATTNNNATTCT
ATATCNNNANNNGANAGGAAATAGTCCTTCTTTTTCTTGTTCTTTGAATACAAT--------TATTTC----ATTTTTGGGTTTTCAAATCAAATCCAAA
TAAATCATTTTTGTTTATGTTATGGTTCGCATTTTTTCTATGGTTCGGCGGTACGGTTCATTAAAACAAAAAA---AGGCCCGGGTGGGTACTGACCAGG
CCAGGCCGTCAAANNNGAAATAAAAAAGGTCCCGTT----GAA---CGAAATTAAAGAGATNNTCTTTTCTTTAGTTTTTT-------------CTATTT
TATCTCTTTCGAATGCTTTCTGTCTTTAAATTTTCTATAAATGATANAAAAGGAA------T-----TN------CNGATAAAAGTTA-GATCTATTT--
-----ATATAA----------TAGAANNNNNNAAGCNNNTAAAAAA--AA--------TATTTNNNAN----N---NGCTATATTTTANNNNA----GCTA
TATAATATATAATCAAATTACTTT---CTAGAT-------T--CTCTAT---------NNTATANNN-NN--NNNAN-----AAAGTA--AAN-------
--NNNNNN--------A-ATAAAGNNACNNNGNNNN--TTTTTNNNGCNNNNNTTTTTGTGTAAT--------------
Rauia_sp
-------------------------GNNNAG-CGAG---CGCA---TG----GCGCTAGGGTCCTCGAG-TCCCGA-ACGG---C---G---CGCGACGC
GC-TCTGAGGAGGTCT-TTCGACACCACCGATCGTCGCGGCACCAATA-CCGGGGACTAGTATTTAGGCCAACCGCGCGCTCAG-AG-CGCACGGGAGGC
CAATATCCGCCC-CACCGCCACGGCCCGCACATGGATGCGGGGAGGGGGTGGGGCAAAGATGCGTGACACCCAGGCAGACGTGCCCTCGGCCTAGAGGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCGCGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTTTTGACATTATAGCGAAAGAAGACGTCCCCTCTCGGGGGGATCCGTG-TTTGGGTCCCCGAGCGGAGAGCTCTCTCG
TTAGAA-ATTCCTTGGCGCGTTCCGCGCCGGGGGTTCGTTACTCGCAG---CG-GGAGGAAGGACGTTAGTTAGTCCCACTC-CCGCCTCGAGACGCGGG
GG-AAGGGGCGGATGCCCCCACCCCGCGGGTGTTTTTAACAGGTTCGCGGGTCGTTC----TC---CGA--GTCCGACAAT-GATCCTTCCGCA------
-----------------------------------------------------------------GGATGGAATCAAGATCGAACAGAAATTCAAAAT
AATCATGAAATGCAAATTCGAAATTAGAACGTTGACGTCTTTGTCAGGAGTCCTA--ATGCAATTAGG------AAAGGAGGAGTTATTTCGTTTAAATA
ACTAAATTTCA--------AATTTAATA---ATT--------------TAAAACAAACTCTTTTGTTGGACGAATTTTGACAGATATGGCTCGACAAAAC
AACTTAAGTAATAAGATAAAAA-----CAAGTGGATTGTGAAAA-AATCCCAAGTTTAATTTTTCCTTTTTTTC--A-GTATTTTTTG----C--AGTA
AAAATCAAATAA--ATGGAAATAAAAAA-AAAAA----AAAAAA-ACGAAAGAATAATAAA---------------ATA---GGNNNNGA----------
------------------AATAGTCCTTCTT------GTTCTTTGAATACAATA--ACAA-----------ATTTTGAGGTTTTCAAATCAAATCAAAA
TAAATCATTTTTGTTNNTGTTATGGTTCG---------G---------CCGTATGGTTNNNNAAAACAAAAAA---AGGCCCGGCTGGGTACTGACCCGG
CCAGGCCGTCAAANNNGAAATAAAAAAGGCCCNNNTGAANAAAATTCAAAATTAAAGAGATATTCTTTTCTTTAGTTTTTT-------------CTATTT
TATCTCTTTTGAATGCTTTCTGTCTTTTAATTTTCTATAAATGNNNNAAAGGGAA------T-----TG------CTGATAAAAGTTA-GATTTTTTT--
-----ANNNAA----------TAGAATACAAAAGNNNNTAAAAAAA-GA--------TATTTTCTAT----T---AGCTATATTTTCGAAANNNNNTCTA
GAAGCG-------ATATTTTCNTNGAGCTATATAATAAATT--CTATATAATAAATTTACCNTCTATATT-CTCTAN-----TTTATA--NNN-------
-GATANNG-----A--AAANNNNNNNNNT-----AGAAA---------G--TAA--------------AGAAAAGA---
E_pumila
-----------------------------------AGC-----------------GTCAGGGTCCTTGAG-CCCCGAAACGGCAACTCCGGCACGCGACGT
GA-GCTCGAGAGGCTTGTTTT-CACCATCGNN-GTCGCGGCCTCAGTCGCCGAGGACTCATATTTGGGCCAACCGCGAGCG--GGAG-CGCACGGGAGGC
CATTCTCCGCCCGCACCACAAAGCCC---CAATGG-TAAGGGGAGGG--TGGGGCAATGATGCGTGACACCCAGGCAGACGTGCCCTCGGCCTAACGGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGATA---AAGTGAAAGAAGGCGCCGCGTCCCGAGGCGCACCGTGTTATGGGGCCCC-TGGAGCGTGCTCTCTCG
TTAC-AT-TTCCTTGGCGCATTCCGCGCCGGGGGTTCGTTT-CCACCA---CGGGGAACG-AGAC----G--AGTCCNNNNNCCCGTGGCGGTAGGGGGA
GGGGTGCCCGTGG-GCACGCCCCCCGCCGGTGTTTT-AACATGTTCGCGGGTCGCTCTGCTGG-GCAGG--TTTCGACAAT-GANNNTTCCGCA------
-------------------------------------------------------------------------------AACAGAAATTCAAAAG
AATCATGAAATGCAAATTCGAAATTAGAACGTTGACGTCTTTGTCAGGAGTCCTA--ATGCAATTAGG------AAAGGAGGAGTTATTTCGTTTAAATA
ACTAAATTGAATAATTAACAATTTAATA---ATA--------------AAAAACAAACTCTTTTGGTGGACGAATTTTGACAGATAGGGCTCGACAAAAC
AACTTAAGTCATAAGACAAAAAAA---CAAGTGGATTGTGAAAA-AATCCCTAGTTCCATTTTTCCTTTTTTTT--T-TCAGTATTTTTTT-TTCCAGTA
AAAGTCAAATAA--ATGGAAATACAAAAAAAA---------GAA-ACGAAAGAATAATAAG---------------AAATGACACTTTGATTCGAATTCT
ATATTATCATAGGAATGGAAATAGTCCTTCTTTTTCTTGTTCTTTGAACACAATA--ACAAGTATTTC----ATTTTTGGGTTTTCAAATCAAATCCAAA
TAAATCATTTTTGTTTATGTTATGGTTCGCATTTTTTCTATGGTTTGGCGGTACGGTTCATTAGAACAAAAAA---AGGCCCGGCTGGGTACTGACCAGG
CCAGGCCGTCGAAGTGGAAATAAAAAAGGCCCCGTT----GAA---CGAAATTAAAGAGATATTCTTTTCTTTAGTTTTTT-------------CTATTT
TATCTCTTTCGAATGCTTTCTGTCTTTAAATTTTCTATAAATGATAGAAATGGAAAAATGATAGAAATGGAATTGCTGATAAAAGTTA-GATCTATTT--
-----CTATAA----------TAGAATCCAAAAGGCAATAAAAAAA-AA--------TATATTCTAT----A---AGCTATATTTTCTATGA----GCTA
TATAAT-------CAAATTACTTT---CTAGAT-------T--CTCTAT---------ATTATAGAG-AA--TCTAG-----AAAGTA--AAGATCTA--
--AAATAA-----AGTAAAGACGGG-----------------------------------------------------------
E_scrotiformis
---------------------GGTCGCAATG-CGAGCACCGCAG-ATGCGGAGCGTCAGGGTCCCTGGG-TCCCGAAACGGTGACTCCGACACGCGACGT
```

```
GA-GCTCGAGAGGTTTTTTTT-CACCACCGATAGTCGCGGCCTCAGTCATCGAGGACTCGAATTTTGGCCAACCGCGAGCG--GGAG-CGCACGGGAGGC
CATTATCCGCCCGCACCGCGAAACCC---CGATGG-TAAGGGGTGGGG-TGGGGCAATGATGCGTGACACCCAGGCAGACGTGCCCTCGGCCTAACAGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGATA---AAGTGAAAGAAGGCGCCGCGTCCAGAGGCGCACCGTG-TACGGGGCCCC-TGGTGCGTGCTCTCTCG
TTAA-GT-TTCCTTGGCGCATTCCGCGCCGGGGGTTCGTTGTCCGCCA---CGGGGAACG-AGAC----G--AGTCCCGCAC-CCGCGGCGGCAGGGGGA
GGAGCGCCCGTAG-GCGCGCCACCC---GGTGTTTT-AACATGTTCGCGGGTCGTTCTGCTAG-GCAGG--TTTCGACAAT-GATCCTTCCGCAG-----
----------------------------------------------------------------------------GAA-----ATTCAAAAG
AATCATGAAATGCAAATTCGAAATTAGAACGTTGACGTCTTTGTCAGGAGTCCTA--ATGCAATTAGG------AAAGGAGG----------TTTAAATA
ACTAAATTGAATAATTAACAATTTAATA---ATA-------------AAAAACAAACTTTTTTGGTGGACGAATTTTGACAGATGTGTCTCGACAAAAC
AACTTAAGTCATAAGAC-AAAAAAAAACAAGTGGATTGTGAAAA-AATCCCTAGTTCCATTTTTCCTTTTTTTTT---------------T--TCCAGTA
AAAGTCAAATAA--ATGGAAATACAAAAAAAA--------AAAAAACGAAAGAATNNNNAA---------------AAATGANNNNNNNNTNCAAATTCT
ATNNNNNTNNNNGNN-TGAAATAGTCCTTCTTTTTCTTGTTCTTTGAATACAATA--ACAAGTATTTC----ATTTTTGGGTTTTCAAATCAAATCCAAA
TAAATCTTTTTTTTTTATGTTATGGTTCGCATTTTTTCTATGGTTTGGCGGTACGGTTCATTAGAACAAAAAG---AGGCCCGGCTGGGTACTGACCAGG
CCAAACCGTCGAAGTGGAAATAAAAAAGGCCCCGTT----GAA---CGAAATTAAAGAGATATTCTTTTCTTTAGTTTTTT-------------CTATTT
TATCTCTTTCGAATGCTTTCTGTCTTTAAATTTTCTATAAATGATAGAAATGGAA------T-----TG------CTGATAAAAGTTA-GATCTATTT--
-----ATATAATAATAA----TAGAATCCAAAAGACAAAAAAGAAT--A-TATATA-TANNNTCTAT----A---AGCTATATTTTCTATGA----GCTA
TATAAT-------CAAATTACTTT---CTAGAT-------T--CTCTAT---------AATATAGAG-AA--TCTAG-----AAAGT----A--------
--AA--GA----------TCT------------------------------------------------------------
E_spnov
------------------TGGGGTCGCAAAG-CGAGCACCGCAG-ATGCGGGGCGTCAGGGTCT-TGAG-TCCCAAAACGATGACTCCGGCACGCGACGT
GA-GCTCGAGAGGCTTGTTTT-CACCACCGATAGTCGCGGCCTCAGTCGACGAGGACTAGAATTTGGGCCAACCGCAAGCG--GGATTCTCACGGGAGGC
CATTCTCCGCCCGCACCACCAGGGCC---CGATGGGTAAGGGGT------GGGGCAACGATGCGTGACACCCAGGCAGACGTGCCCTCGGCCTAACGGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGATA---CAGTGAAAGAAGGCACCGCGTCCTTAGGCACACCGTG-TCCGGGGCCCT-CGGACCGTGCTCTCTCG
TTGC-AT-TTCCTTGGCGCATTCCGCGCCGGGGGTTCGTTGTCCGCAA---CAGGGAACG-GGAC----A--AGTCCCGCAC-CCGAGGCGGCAGGCGGA
GGAGCGCCCGTGG-GCGCGCCCCC-ACCGATATTTT-AACATGTTCGCGGGTCTTTCTGCTAG-GCAGG--TTTCGACAAT-GATCCTTCCGCAGGTTC-
ACCTACGGAAACC----------------------------------------------------------------------ACAGAAATTCAAAAG
AATCATGAAATGCAAATTCGAAATTAGAACATTGACGTCTTTGTCAGGAGTCCTA--ATGCAATTAGG------AAAGGAGGAGTTTTTTCGTTTAAATA
ACTAAATTGAATAATTAACAATTTAATA---ATA-------------AAAAACAAA-----T-GGTGGACGAATTTTGACAGATATGGCTCGACAAAAC
AACTTAAGTCATAAGACAAAAAAA-AACAAGTGGATTGTGAAAA-AATCCCTAGTTCCATTTTTCCTTTTTTTT--T--CAGTATCTT-TT-TTCCAGTA
AAAGTCAAATAA--ATTGAAATACAAAAAAAA--------AGAA-ACGAAAGAATAATAAG---------------AAATGACACTTTGATTCGAATTCT
ATATCATCATAGGAATGGAAATAGTCCTTCTTTTTCTTGTTCTTTGAATACAAT--------TATTTC----ATTTTTGGGTTTTCAAATCAAATCCAAA
AAAATCATTTTTGTTTATGTTATGGTTCGCTTTTTTTCTATGGTTCGGCGGTACGGTTCATTAGAACAAAAAA---AGGCCCGGCTGGGTACTGACCAGG
CCAGGCCGTCGAAGTGGAAATAAAAAAGGCCCCGTT----GAA---CGAAATTAAAGAGATATTCTTTTCTTTAGTTTTTT-------------CTATTT
TATCTCTTTCTAATGCTTTCTGTCTTTAAATTTTCTATAAATGATAGAAAAGGAA------T-----TG------CTGATAAAAGTTA-GATCTATTT--
-----ATATAA----------TAGAATCCAAAAGCCAATAAAAAAA-AA--------TANNNTCTAT----A---AGCTATATTTTANNNGA----GCTA
TATAAT-------CAAATTACTTT---CTAGAT-------T--CTCTAT---------NNTATAGAG-AA--TCTAG-----AAAGTA--AAG-------
--ATCTAA--------A-ATAAAG-----TA-------------------------------------------------
G_trifoliata
---------------------------------CGAG---CGCT---AA----GCGCTAGGGTCCGCAAG-TCCCGA-CGGG---C---GACGCGCGACGT
GT-TCTCGAGAGGTCT-TTCAACACCACCGATCGCCGCGGCGCCNNNCGCCGNNNACTGAAATTTGGGCCAACCGCGCNNA-AG-AG-CGCACGGGAGGC
CAATATCCGCCCTCACCGCCACGCCCCGNNNAAGACAACNAGGGAGGTGTGGGGCAACGATGCGTGACACCCAGGCAGACGTGCCCTCGGCNNNGCGGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGNGAGTCGTTTTAGACATTATAGCGAAAGAAGTCGTCGCCTCTCGGGGTTATCCGTG-TTCGGGTCCCCGAGGGGCGAGCTCTCTCG
TAAGTA--TTCCTTGGCGCGTTCCGCGCCGGGGGTTCGTTACTCGCAG---CGAGGAGGAAGGAC----GTTAGTCCGGCTC-CCGCTGCG--ACGCGGG
GG-GAGGTT-----GCCCCCTCCCCGCGGGTGATTTTAACATGTTCGCGGGTCGTTC----GT---AGG--GTCCGNNNNNNGATCCTTCC---------
----------------------------------------------------------------------------GAACAGAAATTCAAAAT
AATCATGAAATGCAAATTCGAAATTAGAGCGTTGACGTCTTTGTCAGGAGTCCTA--ATGCAATTAGC------AAAGGAGGAGTTATTTCGTTTAAATA
ACTAAATTTCA--------AATTTAATA---ATT--------------CAAAACAAACTCTTTTGTTGAACGAATTTTGACAGATATGGCTCGACAAAAC
AACTTAAGTAATAAGACAAAAA-----CAAGTGGATTGTGAAAA-AATCCCAAGTTTCATTTTTCCTTTTTTTT--T-TCAGTATTTTTT---TGCAGTA
AAAGTCAAATAA--ATGGAAATACAAAAAAAAAA------AAAA-ACGAAAAAATAATNAA---------------AAATGACCCTTTGATTCTAATTCT
ATATCATNNNNNNNNNNNNAATAGTCCTTCTTTTTCTTGTTCTTTGAATNNNAAA--ACAA-----------ATTTTGAGGTTTTCAAATCAAATCCAAA
TAAATCATTTTTGTTTATGTTATGGTTCGCCTTTTTTCTATGGTTCGGCCGNATGGTTCATTAGAACAAAAAA---AGGCCCGGCTGGGNACTGACNNGG
C-----CGTCGAAGNNGAAATAAAAAAGGCCCCGTT----GAA---CAAAATTAAAGAGATATTCTTTTCTTTAGTTTTTT-------------CTATTT
TATCTCTTTTGAATGCCTTCTGTCTTTTAATTTTCTATAAATGAGAGAAAGGGAA------T-----TGTTGATAATGATAAAAGTTA-GATCNNNNT--
-----ATCTAA----------TAGAATACAAAAGNNNNTAAAAAAA-GA--------TATCNTNNNT----A---AGCTAAATTTNNNNNGA----GCTA
GAAAAT-------AAATTTATTTTT--CTATAT-------T--CTCTANNNTATANNGAATATAGNNAAA--TATAG-----AGGATA--TAG-------
------AA-----AGTAAAGAAAAGATAT-----------------------------------------AAAA-----
H_puberula
---------------------GGTCGCAATG-NGAGCGCCGCGA-ATGCGGAGCGTCAGGGTCCCTGAG-TCCCGAAACGAAGACTCAAGCGCGCGACGT
GA-ACTCGAGAGGCTTTGTTT-CACCACCGATAGTCGCGGCCTCAGTCGTCGAGGACTCGAATTTGGGCCAACCGCGAGCG-AGG-G-CGCACGGGAGGC
CATTCTCCGCCCGCACCTCCAAGCCC---CGATGG-TAAGGGGTGGGG-TGGGGCAACGATGCGTGACACCCAGGCAGACGTGCCCTCGGCCTAACGGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGATA---CAGTGAAAGAAGGCGCCGCGTCCCGGGGCGCACCGTG-TCCGGGGCCCC-TGGAGCGTGCTC----G
TTAC-AT-TTCCTTGGCGCATTCCGCGCCCGGGGTTCGTTGTCCACCG---CGGGGAACG-GGAC----G--AGTCCCGCAC-CCGCGGCGGTGAGGGGA
GGAGCGCCCGTGG-GCGCGTCCCCCGTCGGTGTCTT-AACGTGTTCGCGGGTCGTTCTGCTAG-GCAGG--TTTCGACAAT-GATCCTTCCGCANNNNNN
ACCTACGNAAA----------------------------------------------------------------------------------AAAAG
AATCATGAAATGCAAATTCGAAATTAGAACGTTGACGTCTTTGTCAGGAGGCCTA--ATGCAATTAGG------AAAGGAGGAGTTATTTCGTTTAAATA
ACTAAATTGAATAATTAACAATTTAATATTAATAATCAAATTAATAATAAAAAAAAACTCTTTTGGTGGACGAATTTTGACAGATATGGCTCGACAAAAC
AACTTAAGTCATAAGACAAAAAAA--ACAAGTGGATTGTGAAAA-AATCCCTAGTTCCATTTTTCCTTTTTTT------CAGTATTTTTTT-TTCCAGTA
AAAGTCAAATAA--ATGGAAATACAAAAAAAA--------AGAA-ACGAAAGAATAATAAG---------------AAATGACACTTTGATTCGAATTCT
```

```
ATATCATCATAGGAATGGAAATAGTCCTTCTTTTTCTTGTTCTTTGAATACAATA--ACAAGTATTTC----ATTTTTGGATTTTCAAATCAAATCCAAA
TAAATCATTTTTGTTTATGTTATGGTTCGCATTTTTTCTATGGTTCGGCGGTACGGTTCATTAGAACAAAAAA---AGGCCCGGCTGGGTACTGACCAGG
CCAGGCCGTTGAAGTNGAAATAAAAAAGGCCCCGTT----GAA---CGAAATTAAAGAGATATTCTTTTCTTTAGTTTTTT------------CTATTT
TATCTCTTTCGAATGCTTTCTGTCTTTAAATTTTCTATAAATGATANAAATGGAA------T-----TG------CTGATAAAAGNNN-NATCTATTNNN
NGAATATATAA----------TAGAATCCCAAANACAATAAAAAAAAAA--------TATTTTCTAT----A---AGCTATATTTNNNTTGA----GCTA
GATAAT-------CAAATTACTTT---CTAGAT-------T--CTCTAN---------NNNNTAGAG-AA--TCNNN-----AAAGTA--AAG-------
--ANNNA-----------------------------------------------------------------------
M_nigra
--------------------GGGTCGCAATG-CGAGCGCCGCAA-ATGCGGAGCATCAGGGTCCCTGAG-TCCCGAAACGGAGACTGCGGCACGCGACGT
GA-GCTCGAGAGGCTTGTTTT-CACCACCGATAGTCACGGCCTCAGTCGCCGAGGACTCGAATTTCGGCCAACCGCGAGCG--GGAG-CGCACGGGAGGC
CATTCTCCGCCCACACCGCCAGGCCC---CAATGG-TAAGGGGTGGGG-TGGGGCAACGATGCGTGACACCCAGGCAGACGTGCCCTCGGCCTAAGGGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGATA---CAGTGAAAGAAGGCGTCGCGTCCGGAGGCGCACCGTG-TCCGGGGCCTC-TGGAGCGTGCTCTCTCG
TTAC-AT-TTCCTTGGCGCATTCCGCGCCGGGGGTTCGTTGTTCGCCG---CGGGGAACG-GGAC----G--AGTCCCGCAC-CCGTGGCGGTAGGGGGA
GGAGCGCCCGTGG-GCGCGCCCCCCGCCGGTGTTTT-AACATGTTCGCGGGTCGTTCTGCTAG-GCAGG--TTTCGACAAT-GATCCTTCCGC-------
----------------------A-----------------------------------------------------------ACAGAAATTCAAAAG
AATCATGAAATGTAAATTCGAAATTAGAACGTTGACGTCTTTGTCAGGAGTCCTA--ATGCAATTAGG------AAAGGAGGAGTTATTTCGTTTAAATA
ACTAAATTGAATAATTAACAATTTAATA---ATA--------------AAAAACAAACTCTTTTGGTGGACGAATTTTGATAGATATGGCTCGACAAAAC
AACTTAAGTCATAAGACAAAAAAA-AACAAGTGGATTGTGAAAA-AATCCCTAGTTCCATTTTTCCTTTTTTTT--TTTCNNNATTTTTTT-TC-CANTA
AAAGTCAAATAA--ATGGAAATACAAAAAAAA--------NNA--ANNNAANNATAATAAA---------------AAATGANCCTTTGATTNNNATTCT
ATATCNNTNNNNNNNNNGAAATAGTCCTTCTTTTTCTTGTTCTTTGAATNCANNA--ACAAGTATTTC----ATTTTTGGGTTTTCAAATCAAATCCAAA
TAAATCATTTTTGTTTTTGTTTTGGTTCGCATTTTTTCTATGGTTTGNNNGNACGGTTCATTAAANNAAAAAA---AGGCCCGGNNGGNTACTGACCAGG
CCAANNNNNCNNNGNNNNNNNAAAAAAGGCCCCGTT----GAN---NNNNATTAAAGANNNNNTTTTTNNTTTANNTTTTT------------TNNTTT
TANNNNTTTCNNNTNCTTTCNNTCTTTAAATTTNCTATAAATGATANAAAAGNNA------T-----TG------CNNNTAAAAGTTA-GANNNNNTN--
-----NNNTAA----------TAGAANNCCAAANNNNNTAAAAAAA--AA-------TATTTNNNNNTTNNNNTAAGCTATNNTTNNNTNNA----GCTA
TATAAT-------CAAATTNNNNN---NNAGAT-------T--NNNNNG---------AATANNNNN-NA--TCT-------------------------
--------------------------------------------------------------------------------
M_stipularis
-----------------------------------GCGCCGCAN-ATGCGGAGCATCAGGGTCCCTGAG-TCCCGAAACGGAGACTGCGGCACGCGACGT
GA-GCTCGAGAGGCTTGTTTT-CACCACCGATAGTCACGGCCTCAGTCGCCGAGGACTCGAATTTAGGCCAACCGCGAGCG--GGAG-CGCACGGGAGGC
CATTCTCCGCCCACACCGCCAGGCCC---CAATGG-TAAGGGGTGTGG-TGGGGCAACGATGCGTGACACCCAGGCAGACGTGCCCTCGGCCTAAGGGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGATA---CAGTGAAAGAAGGCGTCGCGTCCGGAGGCGCACCGTG-TCCGGGGCCCC-TGGAGCGTGCTCTCTCG
TTAC-AT-TTCCTTGGCGCATTCCGCGCCGGGGGTTCGTTGTTCGCCG---CGGGGAACG-GGAC----G--AGTCCCGCAC-CCGTGGCGGTAGGGGGA
GGAGCGCCCGTGG-GCGCGCCCCCCGCCGGTGTTTT-AACATGTTCGCGGGTCGTTCTGCTAG-GCAGG--TTTCGACAAT-GATCCTTCC---------
-------------------------------------------------------------------------ATCAAGATAGAACAGAAATTCAAAAG
AATCATGAAATGTAAATTCGAAATTAGAACGTTGACGTCTTTGTCAGGAGTCCTA--ATGCAATTAGG------AAAGGAGGAGTTATTTCGTTTAAATA
ACTAAATTGAATAATTAACAATTTAATA---ATA--------------AAAAACAAACTCTTTTGGTGGACGAATTTTGACAGATATGGCTCGACAAAAC
AACTTAAGTCATAAGAC-AAAAAAAAACAAGTGGATTGTGAAAA-AATCCCTAGTTCCATTTTTCCTTTTTTTTTTT-TCAGTATTTTTTT--TCNNGTA
AAAGTCAAATAA--ATGGAAATACAAAAAAAA--------GAAA--CGAAAGAATAATAAG---------------AAATGANACTTTGATTCGAATTCT
ATATCNCCANNGGANNNGNAATAGTCCTTCTTTTTCTTGTTCTTTGAATACAATA--ACAAGTATTTC----ATTTTTGGGTTTTCAAATCAAATCCAAA
TAAATCATTTTTGTTTATGTTATGGTTCGCATTTTTTCTATGGTTTGGCGGTACGGTTCATTAGAACAAAAAA---AGGCCCGGCTGGGTACTGACCAGG
CCAGACCGTCGAAGTGGAAATAAAAAAGGCCCCGTT----GAA---CGAAATTAAAGAGATATTCTTTTCTTTAGTTTTTT------------CTATTT
TATCTCTTTCGAATGCTTTCTGTCTTTAAATTTTCTATAAATGATAGAANNNGAA------T-----TG------CTGATAAAAGTTA-GATCTATTT--
-----ATATAAT----------AGAATCCAAAAGACAATAAAAAAA--AATATATTCTATATTCTAT----A---AGCTATATTTTCTATGA----GCTA
TATAAT-------CAAATTACTTT---CTAGAT-------T--CTCTAG---------AATATAGAG-AA--TCTAG-----AGAATCTAGA--------
--AAGTAA----------AGA-------------------------------------------------------
N_paraensis
----------------------TCGCNNNG-CGAG---CGCA---TA----GCGCTTGGGTACATGTG-TCCCAG-ACGA---T---G---CTCGACGC
GT-TCTCGAGAGGTAT-TATAACACTACCGATCGTCGCAGCACCATTAGCTGAGGACTNNAATTTAGGCCAACCGCGAGCT-AG-AG-CACACGGGAGGC
CAATATCCGCCCTCACCGCCTTCCTCCC-AAGAGAACGAGGGGGAGGGGTGGGGCAACAATGCGTGACACCCAGGCAGACGTGCCCTCGGCCTAAAGGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGATATTACAACGAAAGAAGGCGTTGCCTCCCGAGGAGATCCGTG-TCCGGGTCCNNNAGG-GCAAGCTCTCTCA
TTAGATTTTTCCTTGGCGCGGTCCGCGCCGGGGGTTTGTTGCTCGCAG---CAGAGAGCAAGGAT----GTTAGTCCAGCTC-CCGCTGCG--ATGCGAG
GG-GAGGAA-----GCCCTCTCCTCGCAGGTGTTTT-AACAAGTTCGCGGGTCGTTCTGCTTT-GCAGG--TTTCGACAAT-GATCCTTCCGCANNTTC-
ACCT-------------------------------------------------GTCGAACAAGAGAATCGGATGTAATCAAGATAGAACAGAAATTCAAAAT
AATCATGAAATGCAAATTCGAAATTAGAACGTTGACGTTTTTGTCAGGAGTCCTA--ATGCAATTAGG------AAAAGAGGAGTTATTTCGTTTAAATA
ACTAAATTTCAGAATTAAAAATTTAATA---ATT--------------CAAAACAAACTCTTTTGTTGGACGAATTTTGACAGATATGGCTTGACAAAAC
TACTTAAGTAATAAGACAAAAA-----CAAGTGGATTGTGAAAA-AATCCCTAGTTCCATTTTTCCTTTTTTTC--A-GTATTTTTTTGTTTTTGCAGTA
AAAGTCAAATAA--ATGGAAATACAAAAAAAAAAAAAAAAAGAA-ACGAAAGAATAATAAG---------------AAATNANNNTTTGATTCTAATTCT
ATATNNNNAAANGANNNGNAATAGTCCTTCTTTTTCTTGTTCTTTGAANACAAAA--ACAA-----------ATTTTTGGGTTTTCAAATCAAATCCAAA
TAAATCATTTTTGTTTNNGTTATGGTTTG---------A---------CGGTATGGTTCATTAGAACAAAAAA---AGGCCCGGCTGGGTACTGACCCGG
CCNNGCCGTCGAANNNGAAATAAAAAAGGCCCCNNT----GAA---CGAAATTAACGAGATATTCTTTTCTTTAGTTTTTT------------CTATTT
TATCTCTTTTGAATGCCTTCTGTCTTTTAATTTTCTATAAATGATAGANNNGGAA------T-----TG------CTGATAAAAGTTA-GATCTATTT--
-----ATCTAA----------TAGAATANAAAAGANAATAAAAAAA-GA--------TATATTCTAN----N---NGCTATATTTTCTAAATATTTTCTA
GAAGNT-------ATATTTTCNNNGAGCTATAGAATAAATTTACTTTCTCNNNTCTCNANNNTATAGAANNNTATAG-----AGAATA--GAGAAAGTAA
AGATATAAAATAAAGTAAAGACGGGNTTTTTTNNNNCCATTATTTTTTGTGTAATGTNNNNNNNGTAANNNNNNTTTTT
P_alatus
-------------------GGG-TCGCAGTG-CGAGCACCGNNN-NCGCGGAGCATAAGGGTCCATGAG-CCCCGAAAAGGAGACGAGGGCGCGCGGCGT
TTTGCTCGAGAGGCTTGATTT-CACCACCGATCGCAGCGGCCTCGGTCGCCGGGGACTCGAATTTGGGCCAACCGCGAGCG-GGGAG-CGCACGGGAGGC
CATTCTCCGCCCGCACCACCGGGCCCC-CCGATGT-CGAGGGGTGGGGGTGGGGCAACGATGCGTGACACCCAGGCAGACGTGCCCTCGGCCTAACGGCT
```

```
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGATA---CAGTGAAAGAAGGCGCCGCGTCCCGAGGCGCACCGTG-TCCGGGGCCCC-TGGAGCGTGCTCTCTCG
TTACATT-TTCCTTGGCGCATTCCGCGCCGGGGGGTTCGTTGTCCGCCG---CGGGGAACG-GGAC----G--AGTCCCGCAC-CCGCGGCGATAGGGGGA
GGAGCGCCCGAGGGGCGCGCCCCCCGCCGGTGTTGT-GACGGGTTCGCGGGTCGTTCTGCTGT-GCAGG--TTTTGACAAT-GATCCTTCCGCA------
-----------------------------------------------TTTGTCGAATAAGAGAATCGGATGTAATCAAGATAGAACAGAAATTCAAAAT
AATCATGAAATGCAAATTCGAAATTAGAACGTTGACGTCTTTGTCAGGAGTCCTA--ATGC----------------GGAGGAGTTATTTCGTTTAAATA
ACTAAATTGAATAATTAACAATTTAATA---ATT--------------AAAAACAAACTCCTTTGGTGGACGAATTTTGACAGATACGGCTCGACAAAAC
AACTTAAGTCATAAGACAAAAA-----CAAGTGGATTGTGAAAAAAATCCCTAGTTCCATTTTTCCTTTTTTTTT----CAGTATTTTTT---TCCAGTA
AAAGTCAAATAA--ATGGAAATACAAAAAAAAA------AAGAA-ACGAAAGAATACTAAG---------------AAATGACACTTTGATTCTAATTCT
ATATCATCATAGGAATGGAAATAGTCCTTCTTTTTCTTGTTCTTTGAATACAATA--ACAAGTATTTC----ATTTTTGGGTTTTCAAATCAAATCCAAA
TAAATCATTTTTGTTTATGTTATGGTTCGCATTTTTTCTATGGTTCGGCGGTACGCTTCATTAGAACAAAAAA---AGGCCCGGTCGGGTACTGACCAGG
CCAGGCCGTCGAAGTGGAAATAAAAAAGGCCCCGTT----GAA---CGAAATTAAAGAGATATTCTTTTCTTTAGTTTTTT-------------CTATTT
TATCTCTTTCGAATGCTTTCTGTCTTTAAATTTTCTATAAATGATAGAANTNGAA------T-----TG------CTGATAAAAGTTC-GATCTATTT--
-----ATATAA----------TAGAATCCAAAAGATAATAAAAAA---A--------TAGATTCTAG----A---AGCTATATTTTCTATGA----GCTA
TATAAT-------CAATTTACTTT---CTATAT-------T--CTCTAG---------AATATAG-----------------AGAATA--TAGA------
--AAGTAAA-----GTAA------------------------------------------------------------
P_giganteus
------------------GGGG-TCGCAGTG-CGAGCGCCGCTT-GCGCGGAGCGTCAGGGTCCCTGAG-CCCCGAAACGGAGACGAGGGCGCGCGGCGT
TTTGCTCGAGAGGCTTGATTT-CACCACCGATCGCAGCGGCCTCGGTCGCCGAGGACTCGAATTTGGGCCAACCGCGAGCG-GGGAG-CGCACGGGAGGC
CATTCTCCGCCCGCACCGCCGGGCCCC-CCGATGT-CGAGGGGTGGGGGTGGGGCAACGATGCGTGACACCCAGGCAGACGTGCCCTCGGCCTAACGGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGATA---CAGTGAAAGAAGGCGCCGCGTCCCGAGGCGCACCGTG-TCCGGGGCCCC-TGGAGCGTGCTCTCTCG
TTACTTT-TTCCTTGGCGCGTTCCGCGCCGGGGGTTCGTTGTCCGCCG---CGGGGAACG-GGAC----G--AGTCCCGCAC-CCGCGGCGATAGGGGGA
GGAGCGCCCGAGGGGCGCGCCCC-CGCCGGTGTTGT-GACAGGTTCGCGGGTCGTTCTGCTGT-GCAGG--TTTTGACAAT-GATCCTTCCGCAGGTNN-
-----------NCC----------------------------------TTTGTCGNNCAAGAGAATCGGATGTAATCAAGATAGAACAGAAATTCAAAAT
AATCATGAAATGCAAATTCGAAATTAGAACGTTGACGTCTTTGTCAGGAGTCCTA--ATGC----------------GGAGGAGTTATTTCGTTTAAATA
ACTAAATTGAATAATTAACAATTAAATA---ATT--------------AAAAACAAACTCTTTTGGTGGACGAATTTTGACAGATACGGCTCGACAAAAC
AACTTAAGTCATAAGACAAAAA-----CAAGTGGATTGTGAAAAAAATCCCTAGTTCCATTTTTCCTTTTTTTT-TT--CAGTATTTTTT---TCCAGTA
AAAGTCAAATAA--ATGGAAATACA-AAAAAAA------AAGAA-ACGAAAGAATACTAAG---------------AAATGACACTTTGATTCTAATTCT
ATATCANNNTAGGAATGGAAATAGTCCTTCTTTTTCTTGTTCTTTGAATACAATA--ACAAGTATTTC----ATTTTTGGGTTTTCAAATCAAATCCAAA
TAAATCATTTTTGTTTATGTTATGGTTCGCATTTTTTCTATGGTTTGGCGGTACGGTTCATTAGAACAAAAAA---AGGCCCGGTCGGGTACTGACCAGG
CCAGGCCGTCGAAGTGGAAATAAAAAAGGCCCCGTT----GAA---CGAAATTAAAGAGATATTCTTTTCTTTAGTTTTTT-------------CTATTT
TATCTCTTTCGAATGCTTTCTGTCTTTAAATTTTCTATAAATGATAGAANNNGAA------T-----TG------CTGATAAAAGTTA-GATCTATTT--
-----ATATAA----------TAGAATCCAAAAGACAATAAAAAA---A--------TAGATTCTAT----A---AGCTATATTTTCTATGA----GCTA
TATAAT-------CAATTTACTTT---CTATAT-------T--CTCTA------------TATT---------CTCT-----ATAATA--TAGA------
--AAGTAAA-----GTAAAGACAGGCTTTTTTCAANNNNNC-----TT---TTTTGTGTAA------------------
P_grandiflorus
-------------------GGGGTCGCAGTG-CGAGCGCCGCTT-GCGCGGAGCGTCAGGGTCCCTGAG-CCCCGAAACGGAGACGAGGGCGCGCGGCGT
TTTGCTCGAGAGGCTTGATTT-CACCACCGATCGCAGCGGCCTCGGTCGCCGAGGACTCGAATTTGGGCCAACCGCGAGCG-GGGAG-CGCACGGGAGGC
CATTCTCCGCCCGCACCGCCGGGCCCC-CCGATGT-CGAGGGGTGGGGGTGGGGCAACGATGCGTGACACCCAGGCAGACGTGCCCTCGGCCTAACGGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGATA---CAGTGAAAGAAGGCGCCGCGTCCCGAGGCGCACCGTG-TCCGGGGCCCC-TGGAGCGTGCTCTCTCG
TTACTTT-TTCCTTGGCGCATTCCGCGCCGGGGGTTCGTTGTCCGCCG---CGGGGAACG-GGAC----G--AGTCCCGCAC-CCGCGGCGATAGGGGGA
GGAGCGCCCGAGGGGCGCGCCTCCCGCCGGTGTTGT-CACAGGTTCGCGGGTCGTTCTGCTGT-GCAGG--TTTTGACAAT-GATCCTTCCGCAGNNNC-
ACCTAC-----------------------------------------CTTTTGTCGAACAAGAGAATCGGATGTAATCAAGATAGAACAGAAATTCAAAAT
AATCATGAAATGCAAATTCGAAATTAGAACGTTGACGTCTTTGTCAGGAGTCCTA--ATGC----------------GGAGGAGTTATTTCGTTTAAATA
ACTAAATTGAATAATTAACAATTAAATA---ATT--------------AAAAACAAACTCTTTTGGTGGACGAATTTTGACAGATACGGCTCGACAAAAC
AACTTAAGTCATAAGACAAAAA-----CAAGTGGATTGTGAAAAAAATCCCTAGTTCCATTTTTCCTTTTTTTT-TT--CAGTATTTTTT---TCCAGTA
AAAGTCAAATAA--ATGGAAATACAAAAAAAAAA----AAAGAA-AGGAAAGAATACTAAG---------------AAATGACACTTTGATTCTAATTCT
ATATCATCATAGGAATGGAAATAGTCCTTCTTTTTCTTGTTCTTTGAATACAATA--ACAAGTATTTC----ATTTTTGGGTTTTCAAATCAAATCCAAA
TAAATCATTTTTGTTTATGTTATGGTTCGCATTTTTTCTATGNNNNNNNGGTACGGTTCATTAGAACAAAAAA---AGGCCCGGTCGGGTACTGACCAGG
CC-----GTCGAAGTGGAAATAAAAAAGGCCCCGTT----GAA---CGAAATTAAAGAGATATTCTTTTCTTTAGTTTTTT-------------CTATTT
TATCTCTTTCGAATGCTTTCTGTCTTTAAATTTTCTATAAATGATAGAAATGGAA------T-----TG------CTGATAAAAGTTA-GATCTATTT--
-----ATATAA----------TAGAATCCAAAAGACAATAAAAAA---A--------TAGATTCTAT----A---AGCTATATTTTCTATGA----GCTA
TATAAT-------CAATTTACTTT---CTATAT-------T--CTAGA------------GAAT---------ATAG-----AGAATA--TAGA------
--AAGTAAA-----GTAAAGACAGGCTTTTTTCAAGCATTC-----TT---TTTTGTGTA-------------------
P_jaborandi
-------------------GGGGTCGCAGCG-CGAGCGCCGCTT-GCGCGGAGCGTCAGGGTCCATGAG-CCCCGAAAAGGAGACGAGGGCGCGCGGCGT
TTTGCTCGAGAGGCTTCATTT-CACCACCGATCGCAGCGGCCTCGGTCGCCGGGGACTCGAATTTGGGCCAACCGCGAGCG-GGGAG-CGCACGGGAGGC
CATTCTCCGCCCGCACCACCGGGCCCC-CCGATTTTCGAGGGGTGGGGGTGGGGCAACGATGCGTGACACCCAGGCAGACGTGCCCTCGGCCTAACGGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGATA---CAGTGAAAGAAGGCGCCGCGTCCCGAGGCGCACCGTG-TCCGGGGCCCC-TGGAGCGTGCTCTCTCG
TTACATT-TTCCTTGGCGCATTCCGCGCCGGGGGTTCGTTGTCCGCCG---CGGGGAACG-GGAC----G--AGTCCCGCAC-CCGCGGCGATAGGGGGA
GGAGCGCCCGAGGGGCGCGCCCCCCGCCGGTGTTGT-GACGGGTTCGCGGGTCGTTCTGCTGT-GCAGG--TTTTGACAAT-GATCCTTCCGCANNTTC-
ACCTACNGAAA-----------------------------------------------------GAATCGGATGTAATCAANNTNNAACAGAAATTCAAAAT
AATCATGAAATGCAAATTCGAAATTAGAACGTTGACGTCTTTGTCAGGAGTCCTA--ATGC----------------GGAGGAGTTATTTCGTTTAAATA
ACTAAATTGAATAATTAACAATTTAATA---ATG--------------AAAAACAAACTCTTTTGGTGGACGAATTTTGACAGATACGGCTCGACAAAAC
AACTTAAGTCATAAGACAAAAA-----CAAGTGGATTGTGAAAAAAATCCCTAGTTCCATTTTTCCTTTTTTTTTTT--CAGTATTTTTT---TCCCGTA
AAAGTCAAATAA--ATGGAAATAAA-AAAAAAA------AAGAA-ACGAAAGAATACTAAA---------------AAATGACACTTTGATTCTAATTCT
NNNTCATCATAGGANTGGAAATAGTCCTTCTTTTTCTTGTTCTTTGAATACAATA--ACAAGTATTTC----ATTTGTGGGTTTTCAAATCAAATCCAAA
TAAATCATTTTTGTTTATGTTATGGTTCGCATTTTTTCTATGGTTNNGCGGTACGGTTCATTAGAACAAAAAA---AGGCCCGGTCGGGTACTGACCAGG
```

```
CCAGGCCGTCGAAGTGGAAATAAAAAAGGCCCCGTT----GAA---CGAAATTAAAGAGATATTCTTTCCTTTAGTTTTTT-------------CTATTT
TATCTCTTTCGAATGCTTTCTGTCTTTAAATTTTCTATAAATGATAGAANTGGAA------T-----TG------CTGATAAAAGTTA-GATCTATTT--
-----ATATAA----------TAGAATCCAAAAGATACTAAAAAA---A--------TAGATTCTAT----A---AGCTATATTTTCTATGA----GCTA
TATAAT-------CAATTTACTTT---CTATAT-------T--CTCTA------------TATT---------CTAG-----AGAATA--TAGA------
--AAGNNNA----AGTAAAGACAGGCTNNNNNNA-GNNNNC-----TT---TTTTGTGTAA---TG-----------TA
P_microphyllus
------------------GGGGGTCGCAGTG-CGAGCGCCGCTT-GCGCGGAGCGTCAGGGTCTATGAG-CCCCGAAAAGGAGACGAGGGCGCGCGGCGT
TTTGCTCGAGAGGCTTGATTT-CACCACCGATCGCAGCGGCCTCGGTCGCCGGGGACTCGAATTTGGGCCAACCGCGAGCG-GGGAG-CGCACGGGAGGC
CATTCTCCGCCCGCACCACCGGGCCCC-CCGATGT-CGAGGGGTTGGGGTGGGGCAACGATGCGTGACACCCAGGCAGACGTGCCCTCGGCCTAACGGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGATA---CAGTGAAAGAAGGCGCCGCGTCCCGAGGCGCACCGTG-TCCGGGGCCCC-TGGAGCGTGCTCTCTCG
TTACATT-TTCCTTGGCGCATTCCGCGCCGGGGGTTAGTTGTCCGCCG---CGGGGAACG-GGAC----G--AGTCCCGCAC-CCGCTGCGATAGGGGGA
GGAGCGCCCGAGGGGCACGCCCCCCGCCGGTGTTGT-GACGGGTTCGCGGGTCGTTCTGCTGT-GCAGG--TTTTGACAAT-GATCCTTCCGCA------
-----------------------------------------------------------------------------------GAACAGAAATTCAAAAT
AATCATGAAATGCAAATTCGAAATTAGAACGTTGACGTCTTTGTCAGGAGTCCTA--ATGC----------------GGAGGAGTTATTTCGTTTAAATA
ACTAAATTGAATAATTAACAATTTAATA---ATT--------------AAAAACAAACTCCTTTGGTGGACGAATTTTGACAGATACGGCTCGACAAAAC
AACTTAAGTCATAAGACAAAAA-----CAAGTGTATTGTGAAAAAAATCCCTAGTTCCATTTTTCCTTTTTTTTT----CAGTATTTTTT---TCCAGTA
AAAGTCAAATAA--ATGGAAATACAAAAAAAAA------AAGAA-ACGAAAGAATACTAAG---------------AAATGACACTTTGATTCTAATTCT
ATATCATCATAGGAATGGAAATAGTCCTTCTTTTTCTTGTTCTTTGAATACAATA--ACAAGTATTTC----ATTTTTGGGTTTTCAAATCAAATCCAAA
TAAATCATTTTTGTTTATGTTATGGTTCGCATTTTTTCTATGGTTCGGCGGTACGGTTCATTAGAACAAAAAA---AGGCCCGGTCGGGTACTGACCAGG
CCAGGCCGTCGAAGTGGAAATAAAAAAGGCCCCGTT----GAA---CGAAATTAAAGAGATATTCTTTTCTTTAGTTTTTT-------------CTATTT
TATCTCTTTCGAGTGCTTTCTGTCTTTAAATTTTCTATAAATGATAGAAATGGAA------T-----TG------CTGATAAAAGTTA-GATCTATTT--
-----ATATAA----------TAGAATCCAAAAGATAATAAAAAA---A--------TAGATTCTAT----A---AGCTCTATTTTCTATGA----GCTA
TATAAT-------CAATTTACTTT---CTATAT-------T--CTCTAG---------AATATAG-----------------AGAATA--TAG-------
-----------------A-------------------------------------------------------------
P_pauciflorus
-----------------------CGCAGTG-CGAGCGCCGCTT-GCGCGGAGCGTCAGGGTCC-TGAG-CCCCGAAACGGAGANNNGGGCGCGCGGCGT
TTTGCTCGAGAGGCTTGATTT-CACCACCGATCGCAGCGGCCTCGGTCGCCGAGGACTCGAATTTGGGCCAACCGCGAGCG-GGGAG-CGCACGGGAGGC
CATTCTCCGCCCGCACCGCCGGGCCCC-CCGATGT-CGAGGGGTGGGGGTGGGGCAACGATGCGTGACACCCAGGCAGACGTGCCCTCGGCCTAACGGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGATA---CAGTGAAAGAAGGCGCCGCGTCCCGAGGCGCACCGTG-TCCGGGGCCCC-TGGAGCGTGCTCTCTCG
TTACTTT-TTCCTTGGCGCATTCCGCGCCGGGGGTTCGTTGTCCGCTG---CGGGGAACG-GGAC----G--AGTCCCGCAC-CCGCGGCGATAGGGGGA
GGAGCGCCCGAGGGCCGCGCCCCCCGCCGGTGTTGT-GACAGGTTCGCGGGTCGTTCTGCTGT-GCAGG--TTTTGACAAT-GATCCTTCCGCA------
-----------------------------------------------CTTTTGTCGAACAAGAGAATCGGATGTAATCAAGATAAAACAGAAATTCAAAAT
AATCATGAAATGCAAATTCGAAATTAGAACGTTGACGTCTTTGTCAGGAGTCCTA--ATGC----------------GGAGGAGTTATTTCGTTTAAATA
ACTAAATTGAATAATTAACAATTAAATA---ATT--------------AAAAACAAACTCTTTTGGTGGACGAATTTTGACAGATACGGCTCGACAAAAC
AACTTAAGTCATAAGACAAAAA-----CAAGTGGATTGTGAAAAAAATCCCTAGTTCCATTTTTCCTTTTTTTT-T---CAGTATTTTTT---TCCAGTA
AAAGTCAAATAA--ATGGAAATAAAAAAAAAAA------AAGAA-ACGAAAGAATACTAAA---------------AAATGANACTTTGATTCTAATTCT
ATATCNCCNNNGGAATGGAAATAGTCCTTCTTTTTCTTGTTCTTTGAATACAATA--ACAAGTATTTC----ATTTTTGGGTTTTCAAATCAAATCCAAA
TAAATCATTTTTGTTTATGTTATGGTTCGCATTTTTTCTATGGTTTGGCGGTACGGTTCATTAGAACAAAAAA---AGGCCCGGTCGGGTACTGACCAGG
CCAGGCCGTCGAANTGGAAATAAAAAAGGCCCCGTT----GAA---CGAAATTAAAGAGATATTCTTTTCTTTAGTTTTTT-------------CTATTT
TATCTCTTTCGAATGCTTTCTGTCTTTAAATTTTCTATAAATGATAGAANNNGAA------T-----TG------CTGATAAAAGTTA-GATCTATTT--
-----ATATAA----------TAGAATCCAAAAGACAATAAAAAA---A--------TAGATTCTAT----A---AGCTATATTTTCTATGA----GCTA
TATAAT-------CAATTTACTTT---CTATAT-------T--CTCTA------------TATT---------CTCT-----ATAATA--TAGA------
--AAGTAAA-----GTAAAGACAGGCTTTTTTCAAGCNNNN-----TT---TTTTGTGTAA---TC-----------TA
P_pennatifolius
-------------------GGN-TCNCNNNNNNNAGCGCCGCTT-NCGCGGAGCGTCAGGGT-CCTGAG-CCCCGAAAAGGAGNNNAGGGCGCGCGGCGT
TTTGCTCGAGAGGCTTGATTT-CACCACCGATCGCAGCGGCCTCNGNNNCCGAGGACTCGAATTTGGGCCAACCGCGAGCG-GGGAG-CGCACGGGAGGC
CCTNNTCCGCCCGCACCGCCGGGCCCC-CCGNNNT-CGAGGGGTGGGGGTGGGGCAACGATGCGTGACACCCAGGCAGACGTGCCCTTNGCNTAACGGNT
TGGGGCGCAACNNGCGTTCAAAGACTCGATGGTTCACGGGATTTTGCAATTCACACCAAGTATCGNNATTCGCTACGTTTTTTATCGATGCGNGAGCNGN
NNNNNCCGTTGCNCNGAGTCGTTATAGATA---NCGTGAAAGAAGGCGCCGCGTCCCGAGGCGCACCGTG-NCCGGGGCCCC-TGGAGCGTGCTCTCTCG
TTACTTT-TTCCTTGGCGNNNNNNGCGCCGGGGGTTCGTTGNCCGCCG---CGGGGAACG-GGAC----G--AGTCCCGCAC-CCGCGGCGATAGGGGGA
GGAGCGCCCGAGGGGCGCGCCCCCCGCCGGTGTTGT-GACAGGTTCGCGGGTCGTTCTGCTGT-GCAGG--TTTTGACAAT-GATCCTTCCGCAGNNNT-
CACCTACGGAAACC-------------------------------------GTCGAACAAGAGAATCGGATGTAATCAAGATAGAACAGAAATTCAAAAT
AATCATGAAATGCAAATTCGAAATTAGAACGTTGACGTCTTTGTCAGGAGTCCTA--ATGC----------------GGAGGAGTTATTTCGTTTAAATA
ACTAAATTGAATAATTAACAATTAAATA---ATT--------------AAAAACAAACTCTTTTGGTGGACGAATTTTGACAGATACGGCTCGACAAAAC
AACTTAAGTCATAAGACAAAAA-----CAAGTGGATTGTGAAAAAAATCCCTAGTTCCATTTTTCCTTTTTTTTT----CATTATTTTTT---TCCAGTA
AAAGTCAAATAA--ATGGAAATACAAAAAAAAA------A-GAA-ACGAAAGAATACTAAG---------------AAATGACACTTTGATTCTAATTCT
ATATCATCATAGGAATGGAAATAGTCCTTCTTTTTCTTGTTCTTTGAATACAATA--ACAAGTATTTC----ATTTTTGGGTTTTCAAATCAAATCCAAA
TAAATCATTTTTGTTTATGTTATGGTTCGCATTTTTTCTATGGTTCGGCGGTACGGTTCATTAGAACAAAAAA---AGGCCCGGTCGGGTACTGACCAGG
CCAGGCCGTCGAAGTGGAAATAAAAAAGGCCCCGTT----GAA---CGAAATTAAAGAGATATTCTTTTCTTTAGTTTTTT-------------CTATTT
TATCTCTTTCGAATGCTTTCTGTCTTTAAATTTTCTATAAATGATAGAAATGGAA------T-----TG------CTGATAAAAGTTA-GATCTATTT--
-----ATATAA----------TAGAATCCAAAAGACAATAAAAAA---A--------TAGATTCTAT----A---AGCTATATTTTCTATGA----GCTA
TATAAT-------CAATTTACTTT---CTATAT-------T--CTCTATATTCTCTATAATATAATTCTC--TATATTCTCTATAATA--TAGA------
--AAGTAAA-----GTAAAGACAGGCTT-----------------------T---------------------------
P_peruvianus
---------------------------------AGCGCCGCTT-GNGCGGAGCGTCAGGGTCCCTGAG-CCCCAAAACGGAGACGAGGGCGCGCTGCGT
TT-GCTCGAGAGGCTTGATTT-CACCACCGATCGCAGCGGCCTCGGTCGCCGAGGACTCGAATTTGGGCCAACCGCGAGCG-GGGAG-CGCACGGGAGGC
CATTCTCCGCCCACACCACCAGGCCCC-CCAATGACAAGGGGGTGGGGGTGGGGCAACGATGCGTGACACCCAGGCAGACGTGCCCTCGGCCTAACGGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGATA---CAGTGAAAGAAGGCGCCGCGTCCCGAGGCGCACCGTG-TCCGGGGCCCC-TGGAGCGTGCTCTCTCG
```

```
TTACTTT-TTCCTTGGCGCATTCCGCGCCGGGGGTTCGTTGTCCGCAG---CGGGGAACG-GGAC----G--AGTCCCGCAC-CCGACGCGATAGGGGGA
GGAGCGCCCGAGGGGCGCGCCCCC-GCCGGTGTTGT-GACGGGTTCGCGGGTCGTTCTGCTAT-GCAGG--TTTTGACAAT-GATCCTTCCGCANNNNNN
ANNTNNNNNNAC---------------------------------CTTTTGTCGAACAAGAGAATCGGATGTAATCAAGATAGAACAGAAATTCAAAAT
AATCATGAAATGCAAATTCGAAATTAGAACGTTGACGTCTTTGTCAGGAGTCCTA--ATGC----------------GGAGGAGTTATTTCGTTTAAATA
ACTAAATTGAATAATTAACAATTTAATA---ATT--------------AAAAACAAACTCTTTTGGTGGACGAATTTTGACAGATACGGCTCGACAAAAC
AACTTAAGTCATAAGACAAAAA-----CAAGTGGATTGTGAAAAAAATCCCTAGTTCCATTTTTCCTTTTTTTT-----CAGTATTTTTT---TCCAGTA
AAAGTCAAATAA--ATGGAAATACAAAAAAAAA-----AAAGAA-ACGAAAGAATACTAAG---------------AAATGACACTTTGATTCTAATTCT
ATATCATCATAGGAATGGAAATAGTCCTTCTTTTTCTTGTTCTTTGAATACAATA--ACGAGTATTCC----ATTTTTGGGTTTTCAAATCAAATCCAAA
TAAATCATTTTTGTTTATGTTATGGTTCGCATTTTTTCTATGGTTCGGCGGTACGGTTCATTAGAACAAAAAA---AGGCCCGGTCGGGTACTGACCAGG
CCAGGCCGTCGAAGTGGAAATAAAAAAGGCCCCGTT----GAA---CGAAATTAAAGAGATATTCTTTTCTTTAGTTTTTT-------------CTATTT
TATCTCTTTCGAATGCTTTCTGTCTTTAAATTTTCTATAAATGATAGAAAAGGAA------T-----TG------CTGATAAAAGTTA-GATCTATTT--
-----ATATAA----------TAGAATCCAAAAGACAATAAAAAA---A--------TAGATTCTAT----A---AGCTATATTTTCTATGA----GCTA
TATAAT-------CAATTTACTTT---CTATAT-------T--CTCTAG---------AATATAGAAGAA--TATAG-----AGAATA--TAGA------
--AAGTAAA-----GTAAAGACAGGCTTTTTTCAANCNNNC-----TT---TTT--------------------------
P_spicatus
------------------------CGCAGTG-CGAGCGCCGCTT-GCGCGGAGCGTCAGGGTCCCTGAG-CCCCGAAACGGAGACGAGGGCGCGCGGCGT
TTTGCTCGAGAGGCTTGATTT-CACCACCGATCGCAGCGGCCTCGGTCGCCGAGGACTCGAATTTGGGCCAACCGCGAGCG-GGGAG-CGCACGGGAGGC
CATTCTCCGCCCGCACCGCCGGGCCCC-CCGATGT-CGAGGGGTGGGGGTGGGGCAACGATGCGTGACACCCAGGCAGACGTGCCCTCGGCCTAACGGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGATA---CAGTGAAAGAAGGCGCCGCGTCCCGAGGCGCACCGTG-TCCGGGGCCCC-TGGAGCGTGCTCTCTCG
TTACTTT-TTCCTTGGCGCATTCCGCGCCGGGGGTTCGTTGTCCGCCG---GGGGGAACG-GGAC----G--AGTCCCGCAC-CCGCGGCGATAGGGGGA
GGAGCGCCCGAGGGGCGCGCCCCCCGCCGGTGTTGT-GACAGGTTCGCGGGTCGTTCTGCTGT-GCAGG--TTTTGACAAN-NNNNNT-CCGCAG-----
-------------------------------------------------------------------------------AGATAGAACAGAAATTCAAAAT
AATCATGAAATGCAAATTCGAAATTAGAACGTTGACGTCTTTGTCAGGAGTCCTA--ATGC----------------GGAGGAGTTATTTCGTTTAAATA
ACTAAATTGAATAATTAACAATTAAATA---ATT--------------AAAAACAAACTCTTTTGGTGGACGAATTTTGACAGATACAGCTCGACAAAAC
AACTTAAGTCATAAGACAAAAA-----CAAGTGGATTGTGAAAAAAATCCCTAGTTCCATTTTTCCTTTTTTTTTTT-TCAGTATTTTTT---TCCAGTA
AAAGTCAAATAA--ATGGAAATACAAAAAAAA-------AAGAA-ACGAAAGAATACTAAG---------------AAATGACACTTTGATTCTAATTCT
NNNTCATCATAGGAATGGAAATAGTCCTTCTTTTTCTTGTTCTTTGAATACAATA--ACAAGTATTTC----ATTTTTGGGTTTTCAAATCAAATCCAAA
TAAATCATTTTTGTTTATGTTATGGTTCGCATTTTTTCTATGGTTTGGCGGTACGGTTCATTAGAACAAAAAA---AGGCCCGGTCGGGTACTGACCAGG
CCAGGCCGTCGAAGTGGAAATAAAAAAGGCCCCGTT----GAA---CGAAATTAAAGAGATATTCTTTTCTTTAGTTTTTT-------------CTATTT
TATCTCTTTCGAATGCTTTCTGTCTTTAAATTTTCTATAAATGATAGAAATGGAA------T-----TG------CTGATAAAAGTTA-GATCTATTT--
-----ATATAA----------TAGAATCCAAAAGACAATAAAA-A---A--------TAGATTCTAT----A---AGCTATATTTTCTATGA----GCTA
TATAAT-------CAATTTACTTT---CTATAT-------T--CTCTAT---------AATATAG-----------------AAAGTA--AAG-------
---------------TAA-----------------------------------------------------------
P_sulcatus
--------------------------------CGAGCGCCGCTT-GCGCGGAGCGTCAGGGTCCCTGANCCNCCGNAACGGAGACGAGGGCGCGCGGCGT
TTTGCTCGAGAGGCTTGATTT-CACCACCGATCGCAGCGGCCTCGGTCGCCGAGGACTNNNATTTGGGCCAACCGCGAGCG-GGGAG-CGCACGGGAGGC
CATTATCCGCCCGCACCACCAGGCCCC-CCGGTGT-GGAGGGGTGGGGGTGGGGCAACGATGCGTGACACCCAGGCAGACGTGCCCTCGGCCTAACGGCT
NNNGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGATA---CAGTGAAAGAAGGCGCCGCGTCCCGAGGCGCACCGTG-TCCGGGGCCCC-TGGAGCGTGCTCTCTCG
TTACTTT-TTCCTTGGCGCATTCCGCGCCGGGGGTTCGTTGTCCGCCG---CGGGGAACG-GGAC----G--AGTCCCGCAC-CCGCGGCGATAGGGGGA
GGAGCGCCCGAGGGGCGCGCCCCCCGCCGGTGTTGT----------G-----------------------------------------------A------
-----------------------------------------------------TCGAACAAGAGAATCGGATGTAATCAAGATAGAACAGAAATTCAAAAT
AATCATGAAATGAAAATTAGAAATTAGAACGTTGACGTCTTTGTCAGGAGTCCTA--ATGC----------------GGAGGAGTTATTTCGTTTAACTA
ACTAAATTGAATAATTAACAATTAAATA---ATT--------------AAAAACAAACTCTTTTGGTGGACGAATTTTGACAGATACGGCTCGACAAAAG
AACTTAAGTCATAAGACAAAAA-----CAAGTGGATTGTGAAAAAAATCCCTAGTTCCATTTTTCCTTTTTTTT-T---CCGTATTTTTT---TCCAGTA
AAAGTCAAATAA--ATGGAAA-----AAAAAAA------AAGAA-ACGAAAGAATACTAAA---------------AAATGACACTTTGATTCTAATTCT
ATATCATCATAGGANNNGAAATAGTCCTTCTTTTTCTTGTTCTTTGAATACAATA--ACAAGTATTTC----ATTTTTGGGTTTTCA-------TCCAAA
TAAATCATTTTTGTTTATGTTATGGTTCGCATTTTTTCTATGGTTCGGCGGTACGGTTCATTAGAACAAAAAA---AGGCCCGGTCGGGTACTGACCAGG
CCAGGCCGTCGAAGTGGAAATAAAAAAGGCCCCGTT----GAA---CGAAATTAAAGAGATATTCTTTTCTTTAGTTTTTTTCTTTAGTTTTTCTATTT
TATCTCTTTCGAATTCTTTCTGTCTTTAAATTTTCTATAAATGATAGAANNNGAA------T-----TG------CTGATAAAAGTTA-GATCTATTT--
-----ATATAA----------TAGAATCAAAAAGACAATAAAAAA---A--------TAGATTCTAT----A---AGCTATATTTTCTATGA----GCTA
TATAAT-------CAATTTACTTT---CTATAT-------T--CTCTA------------TATT---------CTCT-----ATAATA--TAGA------
--AAGTAAA-----GTAAAGANNNGCTTTTTTCAAGNNNNN-----NT---TTTTGTGTAA---TG-----------T-
P_trachylophus
------------------------TCGCAGTG-CGAGCGCCGCTT-GCGCGGAGCGTCAGGGTCCATGAG-CCCCGAAAAGGAGACGAGGGCGCGCGGCGT
TTTGCTCGAGAGGCTTGATTT-CACCACCGATCGCAGCGGCCTCGGTCGCCGGGGACTCGAATTTGGGCCAACCGCGAGCG-GGGAG-CGCACGGGAGGC
CATTCTCCGCCCGCACCGCCGGGCCCC-CCGATGT-CGAGGGGTGGGGGTGGGGCAACGATGCGTGACACCCAGGCAGACGTGCCCTCGGCCTAACGGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGATA---CAGTGAAAGAAGGCGCCGCGTCCCGAGGCGCACCGTG-TCCGGGGCCCC-TGGAGCGTGCTCTCTCG
TTACATT-TTCCTTGGCGCTTTCCGCGCCGGGGGTTCGTTGTATGCCG---CGGGGAACG-GGAC----G--AGTCCCGCAC-CCGCGGCGATAGGGGGA
GGAGCGCCCGAGGGGCGCGCCCCCCGCCGGTGTTGT-GACGGGTTCGCGGGTCGTTCTGCTGT-GCAGG--TTTTGACAAT-GATCCTTCCGCA------
-----------------------------------------------CTTTTGTCGAACAAGAGAATCGGATGTAATCAAGATAGAACAGAAATTCAAAAT
AATCATGAAATGCAAATTCGAAATTAGAACGTTGACGTCTTTGTCAGGAGTCCTA--ATGC----------------GGAGGAGTTATTTCGTTTAAATA
ACTAAATTGAATAATTAACAATTTAATA---ATT--------------AAAAACAAACTCTTTTGGTGGACGAATTTTGACAGATACGGCTCGACAAAAC
AACTTAAGTCATAAGACAAAAA-----CAAGTGGATTGTGAAAAAAATCCCTAGTTCCATTTTTCCTTTTTTTT-----CAGTATTTTTT---TC-AGTA
AAAGTCAAATAA--ATGGAAATACAAAAAAAAAA----AAAGAA-ACGAAAGAATACTAAG---------------AAATGACACTTTGATTCTAATTCT
ATATCATCATAGGAATGGAAATAGTCCTTCTTTTTCTTGTTCTTTGAATACAATA--ACAAGTATTTC----ATTTTTGGGTTTTCAAATCAAATCCAAA
TAAATAATTTTTGTTTATGTTATGGTTCACATTTTTTCTATGGTTTGGCGGTACGGTTCATTAGAACAAAAAA---AGGCCCGGTCGGGTACTGACCAGG
CCAGGCCGTCGAAGTGGAAATAAAAAAGGCCCCGTT----GAA---CGAAATTAAAGAGATATTCTTTTCTTTAGTTTTTT-------------CTATTT
TATCTCTTTCGAATGCTTTCTGTCTTTAAATTTTCTATAAATGATAGAAATGGAA------T-----TG------CTGATAAAAGTTA-GATCTATTT--
```

```
-----ATATAA----------TAGAATCCAAAAGATAATAAAAAA---A--------TAGATTCTAT----A---AGCTATATTTTCTATGA----GCTA
TATAAT-------CAATTTACTTT---CTATAT-------T--CTCTA------------TATT---------CTAG-----AGAATA--TAGA------
--AAGTAAA-----GTAAAGACAGGCTTTTTTCAANNNNNC-----TT---TTTTG-----------------------
R_echinata
-------------------GGGGTCGCAATG-CGAGCACCGCAA-ATGCGGAGCGTCAGGGTCCCTTAG-TCCTGAA-CGGAGACTCCAGCACGCGACGT
GA-GCTCGAGAGGCTTGTTTT-CACCACCGATAGTCGCGGCCTCAGTCACCGAGGACTCGAATTTGGGCCAACCGCGAGCG--AGAG-CGCACGGGAGGC
CATTCTCCGCCCGCACCGCCAGGCCC---CAATGG-TAAGGGGTGGGG-TGGGGCAACGATTCGTGACACCCAGGCAGACGTGCCCTCGGCCTAATGGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGATA---TAGTGAAAGAAGGCGCCGCATCCCGAGGCGCACCGTG-TCCGGGGCCCC-TGGAGCGTGCTCTCTCG
TTAC-AT-TTCCTTGGCGCATTCCGCGCCGGGGGTTCGTTGTCCACCA---CGGGGAACG-AGAC----G--AGTCCCGCAC-CCGTGG-GGNNGGGGGA
GGAGCGCCCATGG-GCGCGCCCCC-ACCAGTGTTTT-AACATGTTCACGGGTCGTTCTGCTAG-GCAGG--TTTCGACAAT-GATCCTTCCGCAGGTTC-
ACCTACGG-----------------------------------------------------------------ATCAAGATAGAACAGAAATTCAAAAG
AATCATGAAATGAAAATTCGAAATTAGAACGTTGACGTCTTTGTCAGGAGTTCTA--ATGCAATTAGG------AAAGGAGGAGTTATTTCGTTTAAATA
ACTAAATTGAATAATTAACAATTTAATA---ATA-------------AAAAACAAACTCTTTTGGTGGACGAATTTTGACAGATATGGCTCGACAAAAC
AACTTAAGTAATAAGACAAAAAAAAAACAAGTGGATTGTGAAAA-AATCCCTAGTTCNNTTTTTCCTTTTTTTT----TCAGTATTTTTTT--TCNNGTA
AAAGTCAAATAA--AAGGAAATACAAAAAAAA--------AGAA-ACGAAAGAATAANAAA---------------AAATGACACTTTGATTNNAATTCT
ATATCNCCNNNNGANNNGAAATAGTCCTTATTTTTCTTGTTCTTTGAATACAATA--ACAAGTATTTC----ATTTTTGGGTTTTCAAATCAAATCCAAA
AAAATAATTTTTGTTTATGTTATGGTTCGCATTTTTTCTATGGTTTGGCGGTACGGTTCATTAGAACAAAAAA---AGGCCCGGCTGGGTACTGACCAGG
CCAGGCCGTCNAAGTGGAAATAAAAAAGGCCCCGTT----GAA---CGAAATTAAAGAGATATTCTTTTCTTTAGTTTTTT-------------CTATTT
TATCTCTTTCNAATGCTTTCTGTCTTTAAATTTTCTATAAATGATAGAANNNGAA------T-----TG------CTGATAAAAGTTA-GATCTATTT--
-----ATATAAT----------AGAATCCAAAAGACAATAAAAAAA--A--------TATNNNNTAT----A---AGCTATATTTTNNNNGA----GCTA
TATAAT-------CAAATTACTTT---CTAGAT-------T--CTCTAG---------AATATAGAG-AA--TCTAG-----AAAGT---AA--------
--AGATCN----------AAA-----------------------------------------------------------
Z_rhoifolium
--------------------GGGTCGCAATG-TGAGCACCGCTT-GCACGGAGCAAAAAGGTCCTTCAG-TCCCGTAATGGAGAGCCTGGCACACGACAT
GT-GCTCGAGAGGTTTGTTTA-CACCACCGATCGTCGCGGCTTTGGTCGCCGAGGACTCGAATTTGGGCCAACCGCGAGCT-AGAAG-CGCACGGGAGGC
CATTATCAGCCCGCACCACCAGGCCT--CCGAGGC---AGGGGTGGGG-TGGGGCAATGATGAGTGACACCCAAGCAGACGTGCCCTCGGCCTAAAGGCT
TGGGGCGCAACTTGCGTTCAAAGACTCGATGGTTCACGGGATTCTGCAATTCACACCAAGTATCGCATTTCGCTACGTTCTTCATCGATGCGAGAGCCGA
GATATCCGTTGCCGAGAGTCGTTATAGATA---GTGTGAAAGAAGGCATTATATCCCAGAGCACACCGTG-TCAGGGGCCCC-AAGAGCATGCTCTCTCG
TTAT-AT-TTCCTTGGCACAATCCGCGCCGGGGGTTTATTGTTCGCCCTCACGGGGAACG-GGAC----A--AGTCCCGCAC-CCACAAA--GACGTNAG
GGAGCGGCCCAAAAGCACGCCCCCCGAAAGGGTTAT-CACGAGTTCACGGGTCGTTCTGCTTTTGCAGAGGTTTCGACAAT-GATCCTTCCGCAGGTTC-
ACCTACGGAAAC-------------C---TTGTCTAAAAANNNNNNCTTTTGTCGAACAAGAGAATCGGATGTAATCAAGATAGAACAAAAATGAAAAAT
AATCATGAAATGCAAATTCGAAATTAGAACGTTGACGTCTTTGTCAGGAGTCCTA--ATGCAATTAGG------AAAGGAGGAGTTATTTCGTTTAAATA
ACTAAATTTAATAATTAAAAATTTCATA---ATA-------------AAAAA-AAACTCTTTTGTTGGACGAATTTTGACAGAGATGGCTCGACAAAAC
AACTTAAGTCATAAGACAAAAA-----CAAGTGGATTGTGAAAAAATTCC-TAGTTCCATTTTTCCTTT-TTT------CAGTATTTTTT---TCCAGTA
AAAGTAAAAAAAAAATGGAAATAAATAAAAAAA-----AAAGAA-ACGAAAGAATAATANN---------------AAATGACACTTTGATTCTAATTCT
ATATCNNNNNAGGAATGGAAATAGTCCTTCTTTTTCTTGTTCGTTGANNACAATAAAAGAAGTATTTC----ATTTTTGGGTTTTCAAATCAAATCCAAA
TAAATCATTTTTGTTTATGTTATGGTTCGCATTTTTTCTATGGTTCGACGGNNNGGTTCATTAGAACAAAAAA---AGGCCCGGCTGGGTACTGACCAGG
CCAGGCCGNNNAANTGGAAATAAAAAAGGCCCCGTT----GAA---CGAAATTAACGAGATATTCTTTTCTTTAGGTTTTT------------TCTATTT
TATCTCTTTCGAATGCTTTCTGTCTTTTAATTTTCTATAAATGATAGAAAAGGAA------T-----TG------CTAATAAAAGTTA-GATCTATTT--
-----ATANAATAGAATATAATAGAATACAAAAGACAATAAAAAAA--A--------TATTTTCTAT----A---AGCTATATTTTNNTNGA----GCTG
TATAAT-------CAATTTACTTT---CTATAT-------G--CNNNAG---------ANNNTCG-AGAA--TATAG-----AAAGTA--AAGATATA--
--AAATAAA-----GTAAAGACGGGCTTTTTTNNNNNNNNN-----TT---TTTTGTGTAA---TG-----------T-


;

END;

BEGIN MRBAYES;
  LOG START FILENAME=pilocarpinae_comb_bayes.log REPLACE;


          [********ROOT********]
  OUTGROUP Z_RHOIFOLIUM;


          [********DEFINE CHARACTER GROUPS********]
  CHARSET ITS = 1-729;
  CHARSET TRNG_S = 730-1879;
  CHARSET TRAILING_GAPS_ITS = 1-35 692-729;
  CHARSET TRAILING_GAPS_TRNGS = 730-795 1776-1879;
  EXCLUDE TRAILING_GAPS_ITS;
  EXCLUDE TRAILING_GAPS_TRNGS;


          [********DEFINE PARTITIONS********]
  PARTITION GENES = 2:TRNG_S, ITS;


          [********SET PARTITIONS********]
```

```
  SET PARTITION=GENES;


          [********SHOW TAXON INFO AND MATRIX********]
  TAXASTAT;
  SHOWMATRIX;


          [********MODELS********]
  LSET
    APPLYTO = (ALL)  NST=6 RATES=INVGAMMA;

  UNLINK STATEFREQ=(ALL) REVMAT=(ALL) SHAPE=(ALL) PINVAR=(ALL);



          [********PRIOR ON PRMS********]
  PRSET
    APPLYTO=(ALL)
    RATEPR=VARIABLE
    STATEFREQPR=DIRICHLET(1)
    TOPOLOGYPR=UNIFORM
    BRLENSPR=UNCONSTRAINED:EXPONENTIAL(10.0);


          [********SEARCH PRMS********]
  MCMC NRUNS=4 NGEN=1000000 SAMPLEFREQ=1000 PRINTFREQ=1000 NCHAINS=4 TEMP=0.2
    FILENAME= pilocarpinae_comb_bayes SAVEBRLENS=YES;


          [********SHOW MODEL********]
  SHOWMODEL;


          [********SUMARIZE RESULTS********]
  SUMP
    BURNIN=500
    NRUNS=4
    PRINTTOFILE=YES;
  SUMT
    BURNIN=500
    NRUNS=4
    SHOWTREEPROBS=YES;

  LOG STOP;
END;
```

E Phred20.pl

```
#!/usr/bin/perl
# #
# # Works under Linux/Unix only!.
#
# To be used after phred/phrap/consed programs
# #
# # Written for:
# # 1. making sure only Phred20 bases are being used
# # (i.e., bases with lower scores are assumed as 'N')
# #
# #
# # PDias, September 22, 2007
# #
# # Licence: GPL 3 is assumed
# # (visit: http://www.gnu.org/licenses/gpl.html)
#
$lsresult = `ls -l ./`;
@lines = split (/\n/, $lsresult);
#
@seq_files;
@qual_files;
#
foreach $line (@lines) {
  chomp($line);
  if ($line =~ /^-/) {
    @lineparts = split (/[\t\s]+/, $line);
    $file = @lineparts[@$lineparts-1];
    if ($file =~ /\.phd/i) {
      push(@phd_files, $file);
    }if ($file =~ /\.qual/i) {
      push(@qual_files, $file);
    }if ($file !~ /\./g) {
      push(@seq_files, $file);
    }
  }
}
foreach $phd_file(@phd_files) {
  chomp($phd_file);
  ($name, $ext) = split (/\./, $phd_file);
  open (PHD, "$phd_file") || die "\n\nI can't open '$phd_file'\n\n";
  open (SEQ_PHRED20, ">$name.trngs.phred20") || die "\n\nI can't open '$name.phred20'\n\n";
  print "\n>$name\_trnG\_S\n";
  print SEQ_PHRED20 ">$name\_trnG\_S\n";
  while ($phd_data = <PHD>) {
    chomp($phd_data);
    if ($phd_data =~ /[a|c|g|t][\s\t]{1}\d{1,}[\s\t]{1}\d{1,}/gi) {
      ($base, $quality, $last) = split (/[\s\t]/, $phd_data);
      if ($quality > 19) {
        print uc($base);
        print SEQ_PHRED20 uc($base);
      }else {
        print "N";
        print SEQ_PHRED20 "N";
      }
    }
  }
}
exit;
```

F Pontos de coleta dos "vouchers"

Balfourodendron

B. molle (Miq.) Pirani - P. Dias 213

Brasil. Bahia. Rio de Contas: Mato Grosso, Córrego da Fazendola, 12 km além da represa do rio Brumado em direção a Mato Grosso e 5,5 km aquém do arraial. Vegetação ripária degradada e campos. $13^0 28' 57''$ S - $41^0 51' 52''$ W.

B. riedelianum (Engl.) Engl. - P. Dias 217

Brasil. São Paulo. Piracicaba: ESALQ, às proximidades do Parque da ESALQ, borda direita da via que passa à esquerda do parque (indo do prédio central). Solo argiloso. Cultivado. $22^0 42' 49,9''$ S - $47^0 37' 50,2''$ W.

Esenbeckia

E. almawillia Kaastra - P. Dias 233

Brasil. Acre. Xapuri: Distrito de Porto Rico, Rodovia BR 317, sentido Rio Branco-Brasiléia, 20,05km após o trevo Xapuri-Brasiléia, estrada à esquerda que leva ao Distrito de Porto Rico (estrada antes da entrada para a vila Epitaciolândia), então 11,36km, trilha na margem esquerda, então 220m na trilha. Solo argiloso. Floresta ombrófila densa, dossel cerca de 30-35m, extração de madeira. $10^0 56' 34''$ S - $68^0 29' 11,7''$ W.

E. decidua Pirani - P. Dias 202

Brasil. Minas Gerais. Mato Verde: margem direita da rodovia Mato Verde - Monte Azul (BR 122), 8 km norte da cidade. Caatinga arbÃşreo-arbustiva, solo areno-argiloso. $15^0 20' 07''$ S - $42^0 53' 25''$ W.

E. grandiflora Engl. - P. Dias 273

Brasil. Santa Catarina. Florianópolis: Morro do Ribeirão, estrada Pântano do Sul - Ribeirão da Ilha, 7,86Km na SC-406, então 1,85Km pela estrada Rosália Paulina Ferreira, então 1,98Km na estrada de terra sentido Ribeirão da Ilha (saída à direita), então 330m ao longo da margem esquerda do córrego ("cachoeira", margem direita da estrada). Sobre rochas. Mata atlântica. $27^0 45' 09,1''$ S - $48^0 32' 33,3''$ W.

E. hieronymi Engl. - P. Dias 271

Brasil. Paraná. Paranaguá: APA Morro do Inglês, final da estrada principal (a que sai da BR-277), sítio Zé Bento (de propriedade de Maria do Carmo Santos Marcelino e D. Cleópatra), base do Morro, ca. de 160m do córrego (atrás da casa), margem direita da trilha. Solo argiloso com grandes afloramentos rochosos. Mata atlântica. $25^0 34' 18,7''$ S - $48^0 39' 19,2''$ W.

E. oligantha Kaastra - P. Dias 310

Brasil. Tocantins. Mateiros: Parque Estadual do Jalapão, 23,3Km na estrada para São Félix, então 7Km na estrada para Mumbuca (estrada à esquerda), então 1,4Km na estrada para Boa Esperança (estrada à esquerda), então 8,1Km na estrada à direita em direção ao menor morro, então ca. de 500m até a base do morro (lado esquerdo da estrada), então ca. de 20m subindo o morro. Cerrado campo sujo, entre rochas. $10^0 23' 35,4''$ S - $46^0 36' 45,2''$ W.

E. pumila Pohl - P. Dias 225

Brasil. Mato Grosso. Água Fria: Chapada dos Guimarães, estrada Guimarães-Água Fria, margem esquerda, 15,5 km do início da estrada. Solo argiloso avermelhado. Resto de cerrado na margem da estrada. $15^0 19' 47,8''$ S - $55^0 45' 16,4''$ W.

E. scrotiformis Kaastra - P. Dias 298

Brasil. Rondônia. Ouro Preto D'Oeste: Estação Ecológica do INPA, ca. de 160m na trilha principal, então ca. de 40m na trilha à direita, margem direita do córrego, sobre lageado, próximo a várias *Galipea* spp., *Metrodorea flavida* K. Krause e *Rauia* sp. Solo argiloso com afloramentos rochosos. Floresta ombrófila densa. $10^0 43' 00,6''$ S - $62^0 14' 36,3''$ W.

Esenbeckia sp. nv. - P. Dias 280

Brasil. Mato Grosso do Sul. Ladário: Estrada Parque, 7,7Km da BR-262, então 7,05Km na estrada da Fazenda Carandá ("estrada das Fazendas"), próximo ao morro do Rabicho (ou Rabichão), a noroeste do morro. Ecótono cerrado/pantanal. $19^0 06' 45,7''$ S - $57^0 31' 31,1''$ W.

Galipea

G. trifoliata Aubl.- P. Dias 230

Brasil. Rondônia. Presidente Médice: BR-364, sentido Presidente Médice - Cacoal/Vilhena, 6,5Km após a entrada para Alvorada d'Oeste, estrada para o morro da EMBRATEL (estrada à esquerda), então 2,6Km na estrada, margem direita do córrego (em frente à EMBRAPA), então ca. de 120m dentro da mata, ca. de 25m da margem do riacho. Floresta ombrófila densa. $11^0 15' 32,8''$ S - $61^0 62' 42,4''$ W.

Helietta

H. puberula R. E. Fr. - P. Dias 216

Brasil. Mato Grosso do Sul. Corumbá: Bairro Aeroporto, rua Alan Kardec, 30m após o final da rua subindo o morro pela trilha à direita, à 6m da trilha. Floresta estacional semidecídua. $19^0 01' 23,9''$ S - $57^0 39' 54,3''$ W.

Metrodorea

M. nigra St.-Hil. - P. Dias 264

Brasil. São Paulo. Rio Claro: Mata São José. Solo argiloso. Floresta estacional semidecídua.

M. stipularis Mart. - P. Dias 263

Brasil. Ronsônia. Alvorada D'Oeste: Estrada para Nova Brasilândia, 20km de Alvorada D'Oeste, margem esquerda, beira de córrego. Solo argiloso com pequenos afloramentos rochosos. Floresta ombrófila densa. $11^0 29' 21,9''$ S - $61^0 17' 24,8''$ W.

Neoraputia

N. paraensis (Ducke) Emmerich - P. Dias 245

Brasil. Maranhão. Buriticupu: Buriticupuzinho, cerca de 8km antes de chegar em Buriticupu pela Rodovia BR 222, sentido Bom Jesus das Selvas-Buriticupu, margem esquerda, 2^a Vila de Produção (condomínio) da Companhia Vale do Rio Doce, Reserva da CVRD, final do condomínio, trilha principal à direita, então 250m dentro da mata, margem direita da trilha. Solo argiloso. Resto de Floresta ombrófila densa, muito degradada (gado pastando e lixão). Várias Rutaceae, incluindo *Spiranthera*, *Metrodorea*, *Esenbeckia* e *Pilocarpus* (raro). $4^0 18' 32,1''$ S - $46^0 31' 34,9''$ W.

Pilocarpus

P. alatus Joseph *ex* Skorupa - P. Dias 247

Brasil. Maranhão. Buriticupu: Buriticupuzinho, cerca de 8km antes de chegar em Buriticupu pela Rodovia BR 222, sentido Bom Jesus das Selvas-Buriticupu, margem esquerda, 2^a Vila de Produção (condomínio) da Companhia Vale do Rio Doce, Reserva da CVRD, final do condomínio, trilha principal à direita, então 645m dentro da mata, margem direita da trilha. Solo argiloso. Resto de Floresta ombrófila

densa, muito degradada (gado pastando e lixão). Várias Rutaceae, incluindo *Spiranthera*, *Metrodorea*, *Esenbeckia* e *Pilocarpus* (raro). $4^0 16' 21,8''$ S - $46^0 31' 25,1''$ W.

P. giganteus Engl. - P. Dias 337

Brasil. Espírito Santo. Linhares: Reserva da Companhia Vale do Rio Doce, estrada da Bomba d'água, ca. de 670m antes do final da estrada. Solo arenoso com áreas parcialmente alagadas, margem esquerda da estrada. Mussununga (vegetação semelhante a uma restinga arbustivo-arbórea). $19^0 11' 13,2''$ S - $39^0 54' 50,5''$ W.

P. grandiflorus Engl. - P. Dias 339

Brasil. Espírito Santo. Linhares: Reserva da Companhia Vale do Rio Doce, estrada Farinha Seca, RFL-01/80 - Bloco E2. Solo argiloso. Mata atlântica (floresta de tabuleiro). $14^0 29' 06''$ S - $39^0 06' 07''$ W.

P. jaborandi Holm. - P. Dias 252

Brasil[21].

P. microphyllus Stapf *ex* Wardl. - P. Dias 235

Brasil[22].

P. pauciflorus St.-Hil. - P. Dias 218

Brasil. São Paulo. Piracicaba: Bairro Godinho, "Mata do Godinho", estrada ao lado direito da mata, cerca de 30m da estrada, Parcela 2, na borda de pequena clareira. Solo areno-argiloso. Fragmento de mata mesófila decídua cercada por canaviais. $22^0 42' 38,9''$ S - $47^0 37' 54,6''$ W.

P. pennatifolius Holm. - P. Dias 215

Brasil. São Paulo. Piracicaba: ESALQ, borda da "matinha" (área onde

[21]Informação retida por estar ameaçada de extinção.
[22]Idem.

foram plantadas vŕias espécies nativas), final da rua em frente ao ESA, próximo ao portão. Solo argiloso. Cultivado. $22^0 42' 33,9"$ S - $47^0 37' 40,2"$ W.

P. peruvianus (Macbr.) Kaastra - P. Dias 291

Brasil. Rondônia. Jaru: BR-364, 28,4Km de Ouro Preto d'Oeste, então 4,5Km na Linha 632, Fazenda do Sr. Zuza, fragmento de mata na margem direita da Linha. Solo argiloso com afloramentos rochosos. Floresta ombrófila densa. $10^0 32' 38,9"$ S - $62^0 25' 38,7"$ W.

P. spicatus St.-Hil. - P. Dias 325

Brasil. Bahia. Caetité: ca. 3km de Caetité, margem direita da estrada Caetité-Guanambi. Pequena mata de galeria com subosque.

P. sulcatus Skorupa - P. Dias 322

Brasil. Bahia. Maniaçu: estrada Caetité-Paramirim, a 29,2Km de Caetité e 1,8Km do cruzamento para Maniaçu, margem direita da estrada, beira da estrada. Solo areno-argiloso. Restos de caatinga arbustiva. $13^0 49' 27,9"$ S - $42^0 23' 24,2"$ W.

P. trachylophus Holm. - P. Dias 323

Brasil. Bahia. Maniaçu: estrada Caetité-Paramirim, a 32,7Km de Caetité e 5,3Km do cruzamento para Maniaçu, margem direita da estrada, beira da estrada. Solo areno-argiloso. Restos de caatinga arbustiva. $13^0 47' 46,7"$ S - $42^0 22' 45,9"$ W.

Rauia

R. resinosa Nees & Mart. - P. Dias 243

Brasil. Maranhão. Buriticupu: Buriticupuzinho, cerca de 8km antes de chegar em Buriticupu pela Rodovia BR 222, sentido Bom Jesus das Selvas-Buriticupu, margem esquerda, 2^a Vila de Produção (condomínio) da Companhia Vale do Rio

Doce, Reserva da CVRD, final do condomínio, trilha principal à direita, então 645m dentro da mata, margem direita da trilha. Solo argiloso. Resto de Floresta ombrófila densa, muito degradada (gado pastando e lixão). Várias Rutaceae, incluindo *Spiranthera*, *Metrodorea*, *Esenbeckia* e *Pilocarpus* (raro). $4^0 18' 32,1''$ S - $46^0 31' 34,9''$ W.

Raulinoa

R. echinata Cowan - P. Dias 257

Brasil. Santa Catarina. Ibirama: estrada à direita no portão de entrada da cidade, então estrada à direita após a ponte (estreita) reformada pela Usina, margem direita do Rio Itajaí-açu, entre afloramentos rochosos na beira da água. $27^0 04' 58,3''$ S - $49^0 29' 58,8''$ W.

Zanthoxylum

Z. rhoifolium Lam. - P. Dias 232

Brasil. Ariquemes: Rodovia BR 364, sentido Cuiabá-Porto Velho, km 495, cerca de 22km do perḿetro urbano de Ariquemes, estrada à direita, então 12km, margem direita, pequeno morro. Solo argiloso com afloramentos rochosos (0,5-1m). Florest ombrófila densa bastante degradada. $10^0 04' 13,6''$ S - $62^0 57' 41,2''$ W.

G Seqüências obtidas do GenBank

>gi|58737220|emb|AJ879084.1| Murraya koenigii ITS1 (partial), 5.8S rRNA gene and ITS2 (partial)
AGGGATATTGTCGAAACCTGCCAGCAGAACGACCCGCGAACCGGTTGAAATCACCGGCGGTGGGAGGGGG
GCGCGCCCCAGCTGCGGGCGCTCCCCCTCCCTTGCCTCGCCGGGGGGAGCGGAATTCGCCCCTTTCCCCT
GGGGCAAACCACCGAACCCCCGGGCGGAACCGCCCCAAGGAAATCAAACGAGAGAGGGGGATCCCCCGGC
CCCGGAAACGGGGCGCGGGGGGATGCGGGGCCTTCTTTCCCTCATTTCCAAAACAACTCTTGGCAACGGA
TATCTCGGCTCTCGCATCGATGAAGAACGTAGCGAAATGCGATACTTGGTGTGAATTGCAGAATCCCGTG
AACCATCGAGTCTTTGAACGCAAGTTGCGCCCCAAGCCGTTAGGCCGAGGGCACGTCTGCCTGGGTGTCA
CGCATCGTTGCCCCACCCCGCCCCCCTCGGGGGCCCGGCGGTGTGGGCGGAGATTGGCTTCCCGTGCGCT
CCCCGCTCGCGGTTGGCCCAAATCCGAGTCCCCGGCGACCGAAGCCGCGACGATCGGTGGTGAAACAAAA
AGCCTCTCGAGCTCCCGTCGCGTGCCTCGGTCTCCGCGAGGGGACCCGCAGACCCGACGATCCGCGCAGG
CGGACGCTCGCATCGCGACCCCAGGTCAGGCGGGATACCCCGCTATT

>gi|58737221|emb|AJ879085.1| Murraya paniculata ITS1 (partial), 5.8S rRNA gene and ITS2 (partial)
TGGGAACTGCGGAAGGATCATTGTCGAAAGCCTTCCCAGCAGAACGACCCGCGAACCAGTTGGAATCACC
GGCGGCGGGAGGGGGGACGCGCTCCGCTGCGGGCGCGCCTCCTCGCCCCCCCTTGCTGTCGGGAGTGGGA
CACGTCCCTATCCCGGCGGCGAAACAACGAACCCCCGGCGCGGACCGCGCCAAGGAAATCCAACGAGAGA
GCACGCTCCCGCGGCCCCGGAGACGGTGTGCGGCGGGACGCGGCGCCTTCTTTCACTTGTAATCCAAAAC
GACTCTCGGCAACGGATATCTCGGCTCTCGCATCGATGAAGAACATAGCGAAATGCGATACTTGGTGTGA
ATTGCAGAATCCCGTGCCATGAGTCTTTGAACGCAAGTTGCGCCCCAAGCCGTTAGGCCGAGGGCACGTC
TGCCTGGGTGTCACGCATCGTTGCCCCACCCCACCCCCCCGGGGGCCCGGCGGTGCGGGCGGATATTGGC
CTCCCGTGCGCTCCCCGCTCGCGGTTGGCCCAAATCTGAGTCCTCGGCGACCGAAGCCGCGGCGATCGGT
GGTGAATGAAAAGCCTCTCGAGCTCCCGCCGCGTGCCCGGTCTCCGCGAGGGGACTTCGCGACCCTGACG
CCCCGCGCAAGCGGCGCTCGCATCGCGACCCCAGGTCAGGCGGGATCACCCCG

>gi|78883190|gb|DQ225789.1| Ptelea trifoliata subsp. angustifolia isolate 1765 internal transcribed spacer 1, complete sequence
GGATCGCGGCGACGTGGGCGGTTCGCTGCCTGTGACGTCGCGAGAAGTCCACTGAACCTTATCATTTAGA
GGAAGGAGAAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTGTCGAAACTCGTAGAGC
AGAATGACCCGTGAACTCGTAGAAAAACAACATTGGCTGGAGGCACGCACTCTTTGTGGTGCTCCTCCCT
CTTTCACCGTCGGTGTGGGATTCTTCCTTCTCCCTGCGGTGAACAACGAACCCCCGGCACGGACTGTGCC
AAGGAAATATAACGAGAGAGAAGTATCTTGGGGCCCCGAAAACGGTGTGCCTTGGGATGTTGTGCCCTCT
TTCAATTTATCTTTAACGACTCTCGGCAACGGATATCTCGGCTCTCGCATCGATGAAGAACGTAGCGAAA
TGCGATACTTGGT

>gi|78883191|gb|DQ225790.1| Ptelea trifoliata isolate AA10 internal transcribed spacer 1, complete sequence
GGATCGCGGCGACGTGGGCGGTTCGCTGCCTGTGACGTCGCGAGAAGTCCACTGAACCTTATCATTTAGA
GGAAGGAGAAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTGTCGAAACTCGTAGAGC
AGAATGACCCGTGAACTCGTAGAAAAACAACATTGGCTGGAGGCACGCACTCTTTGTGGTGCTCCTCCCT
CTTTCACCGTCGGTGTGGGATTCTTCCTTCTCCCTGCGGTGAACAACGAACCCCCGGCACGGACTGTGCC
AAGGAAATATAACGAGAGAGAAGTATCTTGGGGCCCCGAAAACGGTGTGCCTTGGGATGTTGTGCCCCTC
TTTCAATTTATCTTTAACGACTCTCGGCAACGGATATCTCGGCTCTCGCATCGATGAAGAACGTAGCGAA
ATGCGATACTTGGT

H Valores dos parâmetros usados no ProAlign

```
Character frequencies
A-0.2
C-0.2
G-0.2
T-0.2
GAP-0.2

HMM model
Delta-0.1
Epsilon-0.75

Viterbi traceback-sample

Pairwise alignment
GOP-15
GEP-7

Terminal gaps-penalize
Trailing sequences-correct
Max allow-150
Distancec-correct
Scale-1

Bandwidth-505
```

Material suplementar

Por favor, veja o DVD-ROM.

A Autocorrelação

/Cap3/Autocorrelation/

B Buscas de similaridade no GenBank

/Cap3/GenBank/

C Matriz utilizada na análise com *Ptelea*

/Cap3/Ptelea/

D Regiões Phred20

/Cap3/ITS/Consensuses/

/Cap3/trnGS/Consensuses/

E Réplicas amostradas pelo ProAlign

/Cap3/ITS/Consensuses/ProAlign/

/Cap3/trnGS/Consensuses/ProAlign/

Capítulo 4

Phylogeny of *Pilocarpus* Vahl (Rutaceae) and Stochastic Mapping of Morphological Characters

4.1 Abstract

We used morphological and molecular characters to study morphological evolution within *Pilocarpus* (Rutaceae). *Pilocarpus* comprises 17 species and occur in a variety of vegetation types from Mexico to Argentina, and some of its species are well known as "jaborandi". For the phylogenetic analyses, we used nucleotide sequences of the internal trancribed spacers (ITS1 and 2), the 5.8S gene, and *trn*G-S spacer, plus 94 morphological characters, for 11 species. Leaf blade and corolla aestivation patterns were selected for further evolutionary investigation applying the Markov chain Monte Carlo method under the Bayesian paradigm. Our results support the monophyly of the genus, although some of the relationships within the genus are still uncertain. Our character histories study showed that 1) coumpound and unifoliolate leaves might have independently arisen from simple ones, and compound and simple leaves might be synapomorphies of two major clades in the genus (the last as a reversal); and 2) the quincuncial imbricate corolla might have been the pattern present at the most basal node of the genus, whereas proximal cochleate imbricate might be a synapomorphy of a major clade within the genus, and all other variations might have diversified (within *Pilocarpus*) from the proximal cochleate imbricate pattern. In addition, comparing parsimony and Bayesian character mappings reinforced the perspective of considering synapomorphy as a matter of probability.

4.2 Resumo

Neste trabalho foram utilizados dados morfológicos e moleculares de 11 espécies de *Pilocarpus* (Rutaceae) para estudar a evolução morfológica no gênero. *Pilocarpus*, popularmente conhecido como "jaborandi", possui 17 espécies, ocorre do México à Argentina e está bem representado nas principais formações vegetacionais. Para as análises filogenéticas foram utilizadas seqüências nucleotídicas dos espaçadores transcritos internos (ITS1 e 2), do gene 5.8S, do espaçador *trn*G-S e 94 caracteres morfológicos. O método MCMC foi utilizado para traçar hipóteses evolutivas sobre a diversificação dos padrões de lâmina foliar e de estivação da corola. Os resultados reiteram a monofilia do gênero, embora algumas das relações internas ainda sejam desconhecidas. Os estudos de evolução morfológica indicam que 1) as folhas compostas e unifolioladas teriam surgido independentemente de folhas simples e as folhas compostas e simples poderiam ser sinapomorfias de grupos dentro do gênero (embora as últimas através de reversão); e 2) o padrão imbricado quincuncial da corola provalvelmente é o padrão basal no gênero, a corola imbricada cocleada proximal representa uma sinapomorfia de um dos principais clados dentro do gênero e os outros tipos de corola se diversificaram a partir do padrão imbricado cocleado proximal. Adicionalmente, ao se comparar os mapeamentos com parcimômia e com o método bayesiano, fica claro que sinapomorfia pode ser considerada como uma questão de probabilidade.

4.3 Introduction

The genus *Pilocarpus* Vahl –"*jaborandi*" – (subtribe Pilocarpinae, tribe Galipeeae, subfamily Rutoideae) comprises 17 species and occurs in a variety of vegetation types from Mexico to Argentina.

Pilocarpus representatives are usually shrubs (Figure 4.1), but some species can also be trees (e.g., *P. grandiflorus* Engl., *P. pauciflorus* St.-Hil., and *P. pennatifolius* Engl.). The genus is generally characterized by having racemes (Figure 4.2) and by its anthers bearing a postero-dorsal gland.

Economically, *Pilocarpus* is noteworthy by the alkaloid pilocarpine, which is of great importance in the pharmaceutical industry (*e.g.*, MERCK [26]). However, the harvest of their leaves for more than a century has endangered some species, as such *P. jaborandi* Holm., *P. microphyllus* Stapf *ex* Wardl., and *P. alatus* Joseph *ex* Skorupa (Kaastra [19]). For example, *P. jaborandi* has been exploited since the 19$^{\text{th}}$ century, and its previously wide distribution throughout the humid submontane forests ("brejos de altitude" according to Daly & Mitchell [6]) of Northeastern Brazil, is now restricted to the "Serra do Ibiapaba" (northestern Brazil). In addition, more recent surveys (*e.g.*, Pinheiro [30], [31]) have shown that other species, *e.g.*, *P. microphyllus* and *P. alatus*, are also undergoing the same strong human pressure (as *P. jaborandi*), driven especially by the pharmaceutical industry.

Although taxonomic revisions of *Pilocarpus* have been presented by Kaastra ([19]) and Skorupa ([36]), to date, its representation in phylogenetic analyses of (and within) the Rutaceae is rather incipient. For example, in the analyses by Chase *et al.* ([5]) and by Groppo ([11]), *Pilocarpus* is represented by only one and the same

Figure 4.1 Examples of shrubby representatives of *Pilocarpus*. (a) *P. jaborandi*, (b) *P. spicatus*, and (c) *P. sulcatus*. (Photos by R.G. Udulutsch)

species, *P. spicatus*.

On the other hand, all phylogenetic analyses with Rutaceae representatives have been based only on molecular data; morphological characters being fully ignored in the phylogeny reconstruction, as one can note in, *e.g.*, Araújo *et al.* ([1]), Chase *et al.* ([5]), Duretto & Ladiges ([8]), Federici *et al.* ([9]), Groppo ([11]), Morton *et al.* ([27]), Samuel *et al.* ([34]) and Scott *et al.* ([35]). Further, only standard parsimony has been used as optimality criterion to find the optimal trees.

In this study, we use morphological and molecular characters to reconstruct species-level relationships within *Pilocarpus* and use this phylogeny as framework to investigate possible evolutionary scenarios for leaf blade and corolla aestivation patterns, the two most intriguing characters in the genus.

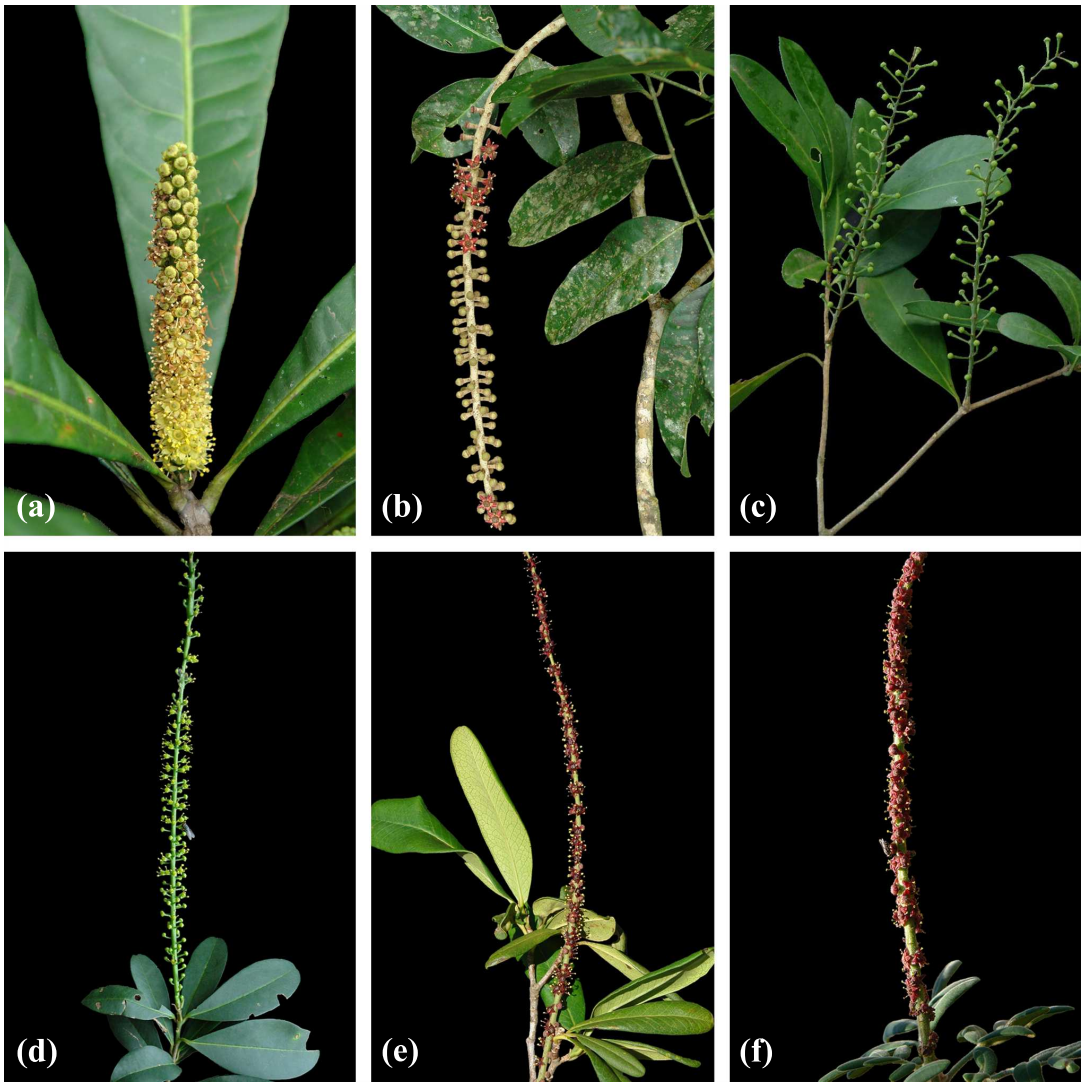


Figure 4.2 Examples of racemes of *Pilocarpus*. (a) *P. giganteus*, (b) *P. grandiflorus*, (c) *P. pauciflorus*, (d) *P. spicatus*, (e) *P. sulcatus*, and (f) *P. trachylophus*. (Photos by R.G. Udulutsch)

4.4 Material and methods

4.4.1 Taxon sampling

We included 11 of the 17 recognized species of *Pilocarpus* (Skorupa [36], Skorupa & Pirani [37]) in all the analyses. Although infra-specific taxa are accepted for several species (see Kaastra [19] and Skorupa [36]), our interest here is on major patterns of leaf and floral morphological evolution in the genus, and, therefore, those infra-specific ranks are irrelevant. However, we used representatives of those ranks whenever possible.

We used as outgroups (Nixon & Carpenter [28]) *Esenbeckia decidua* Pirani, *Raulinoa echinata* Cowan (Pilocarpinae), and *Balfourodendron molle* (Miq.) Pirani (all from Esenbeckiinae, the subtribe sister to Pilocarpinae, see Dias [7]). These outgroups were selected according to their affinities suggested by Dias ([7]).

For this study, we examined herbarium and fresh specimens, including type material whenever possible and our own collections (deposited at SPF[1]). For the morphological analyses, the number of collections studied per species was, in some degree, dependent on loans from other herbaria, and we had acess to all 17 species. However, because we obtained sequences for 11 species, our analyses were restricted to these especies. For the molecular study, all specimens (also included in the morphological analyses) are in the Table 4.1, and the especimens exclusively used for the morphological analyses are listed in the Appendix D).

[1]Some collections are also deposited at NY herbarium.

Table 4.1 Voucher information for the molecular analyses.

Genus	Species	Collector and number	Acronym	Locality
Balfoudendron	*B. molle* (Miq.) Pirani	P. Dias 213	SPF	Brasil, BA, Rio de Contas
Esenbeckia	*E. decidua* Pirani	P. Dias 202	SPF	Brasil, MG, Mato Verde
Pilocarpus	*P. alatus* Joseph *ex* Skorupa	P. Dias 247	SPF	Brasil, MA, Buriticupu
	P. giganteus Engl.	P. Dias 337	SPF	Brasil, ES, Linhares
	P. grandiflorus Engl.	P. Dias 339	SPF	Brasil, ES, Linhares
	P. jaborandi Holm.	P. Dias 252	SPF	Brasil[1]
	P. microphyllus Stapf *ex* Wardl.	P. Dias 235	SPF	Brasil[1]
	P. pauciflorus St.-Hil.	P. Dias 218	SPF	Brasil, SP, Piracicaba
	P. pennatifolius Holm.	P. Dias 215	SPF	Brasil, SP, Piracicaba
	P. peruvianus (Macbr.) Kaastra	P. Dias 291	SPF	Brasil, RO, Jaru
	P. spicatus St.-Hil.	P. Dias 325	SPF	Brasil, BA, Caetité

...

(*Continued on next page*)

[1]Not provided due to the species being endangered.

Table 4.1 Voucher information for the molecular analyses. (*Continued*)

	P. sulcatus Skorupa	P. Dias 322	SPF	Brasil, BA, Maniaçu
	P. trachylophus Holm.	P. Dias 323	SPF	Brasil, BA, Maniaçu
Raulinoa	*R. echinata* Cowan	P. Dias 257	SPF	Brasil, SC, Ibirama

4.4.2 Characters

4.4.2.1 Molecular data

Molecular data were obtained from Dias ([7]) and we are using the same alignment (trailing gaps excluded) used in that study. These data are concatenated nucleotide sequences from the internal transcribed spacers (ITS1 and ITS2), and the gene 5.8S (hereafter generically called "ITS"); and also from the plastidial *trn*G-S spacer.

4.4.2.2 Morphological data

94 morphological characters were used in our analyses. Unobserved states were scored as "?" and uncomparable as " -", although both are treated as missing by MrBayes 3.1.2 (Ronquist & Huelsenbeck [33]). The last category was rather common in leaf characters for the simple-leaved terminals. The list of characters and the coded data matrix are presented in the Appendices B and A, respectively.

Stem – the sole stem character used was related to whether the plant was armed or not.

Trichomes – unlike the taxonomic studies carried out by Kaastra [19] and Skorupa [36], in which different trichome types are considered usefull to identify taxa, we used presence or absence of trichomes, for this treatment is more defensible in terms of primary homology statements.

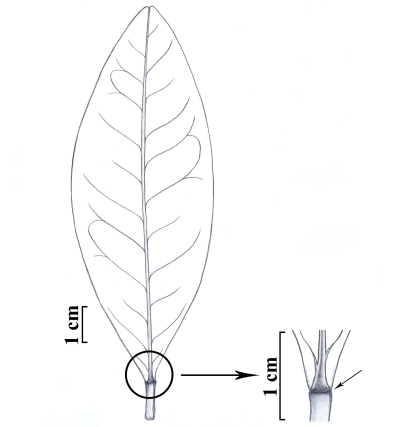
Leaf characters – these characters include presence/absence of some structures (wings and grooves of the petiole and petiolule), venation patterns, and blade division.

The rich variation found in *Pilocarpus* leaves is surely a challenge for appropriate discrete coding. However, some genetic and developmental studies with other simple and compound-leaved angiosperms (*e.g.*, Bharathan & Sinha [3], Hareven *et al.* [12], Kim *et al.* [20]) have provided fruitful insights on gene expression patterns behind these leaf blade types. One of the most interesting results is that class-1 *KNOX* genes (*e.g.*, Bharathan & Sinha [3], Bharathan *et al.* [2]) are extensively expressed in compound leaves (except some legumes, *e.g.*, pea) independently of the number of leaflets, whereas their expression is absent in simple leaves.

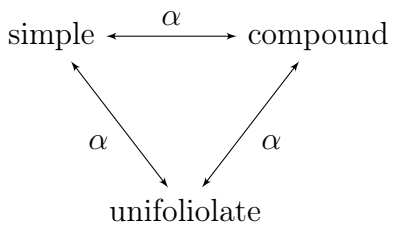
In addition, Bharathan *et al.* ([2]) have found that although a compound leaf may mimic a simple leaf's form, as *in vivo* in the Apiaceae genus *Lepidium*, there is no evidence for the reverse. Even when it was tried *in vitro* by Hareven *et al.* ([12]), the simple pattern was sufficiently stable genetically not to give rise to the compound pattern (although the simple pattern may give rise to irregularly lobed laminas, see Lenhard *et al.* [21]). Therefore, we assumed these two patterns (simple and compound) as states of the leaf blade character.

Pilocarpus, however, presents still another leaf pattern, the "unifoliolate" (Figure 4.3(a)), as yet genetically unstudied. Unifoliolate leaves in Rutaceae are remarkable by a conspicuous articulation located in the region between the petiole and the leaf blade (Figure 4.3(a)). That articulation has not been anatomically described, and we are unaware of its very nature in *Pilocarpus*.

The unifoliolate pattern was found also in other angiosperm families, *e.g.*, Leguminosae, and Hofer *et al.* ([16]) have provided evidence that unifoliolate leaves in pea are an outcome of a recessive mutation (*UNI*, exclusive of legumes) at the



(a)



(b)

Figure 4.3 Leaf blade patterns (character 2). (a) Section of a unifoliolate leaf, the articulation is indicated by the arrow. (b) Putative relationships among character states (α means probability of change).

locus *PEAFLO*. Nevertheless, *uni* pea mutants also show 3-foliolate leaves, and an increase in the *uni* expression gives rise to pinnately compound leaves, demonstrating the compound nature of these unifoliolate leaves. On the other side, however, our unifoliolate terminals (*P. pauciflorus* and *E. decidua*) show no compound leaves, suggesting a potentially different pattern. Thus, we thought it to be inappropriate to score the uniofoliolate terminals as either simple or compound, and we opt for a third state whose putative relationships to the other states are shown in the Figure 4.3(b).

Flower – Among flower characters, corolla aestivation is the most diverse,

but the diversity of patterns found in *Pilocarpus* has been masked in the literature by the use of few terms to describe all of them. For example, Kaastra ([19]) and Skorupa ([36]) have described all patterns as valvar, quincuncial, or imbricate. However, we found seven patterns[2] (Figure 4.4), which we used as states, and named, in general, according to Weberling's terminology (Weberling [39]).

Although recognition of several patterns allows us to better describe the morphological diversity of *Pilocarpus*, understanding their evolutionary relationships is a rather difficult task. Further complexity is provided by many species of *Pilocarpus* presenting floral dimorphism (4- and 5-merous flowers), *e.g.*, *P. alatus*, *P. carajaensis* Skorupa, *P. giganteus* Engl., *P. pauciflorus* St.-Hil., *P. riedelianus* Engl., *P. spicatus* St.-Hil., *P. trachylophus* Holm., and *P. trifoliolatus* Skorupa & Pirani. However, the corolla aestivation pattern between 4-and 5-merous flowers differ just by one piece and the basic pattern is the same. Therefore, we did not distinguish between 4- and 5-merous corollas.

Fruit – as the leaf blade, the fruit of *Pilocarpus* shows a number of variations for which conjectures about primary homology are sometimes difficult to propose. One such example is the number of carpels that develop in fruit. For example, only one carpel usually ripens in *P. alatus*, albeit it is four or five in *P. trachylophus*. Nonetheless, the fruits possess a number of other features phylogenetically usable, such as the appendages, ribs, etc.

Seeds – the seeds of *Pilocarpus* are rather similar and we restricted seed characters in our analyses to the form and presence/absence of appendages.

[2]Although we are using only five of them due to our taxon sampling strategy (see item 4.4.1 above), we thought it would be better to describe all the diversity found in *Pilocarpus*.

Finally, we partitioned the morphological data set into five partitions, according to the "structure" of the plant body (which each character was scored from), as: 1) vegetative (stem and leaf characters), 2) inflorescence, 3) flowers, 4) fruit, and 5) seeds. For each of these subsets, we set MrBayes to infer different, independent parameter values for the models, allowing different evolutionary rates across partitions.

4.4.3 Phylogenetic analyses

The nodes with PP ≥ 0.75 of the topology found by Dias ([7]) were used as constraints for all analyses performed here. Tree searches were run with MrBayes (Ronquist & Huelsenbeck [33]) using four independent MCMC for 15 million generations, sampling at each 1000. The diagnosis of the chains was implemented as suggested by Dias ([7]). All search commands used in MrBayes are provided in the Appendix C.

4.4.4 Stochastic mapping

In general, we used the method described by Huelsenbeck *et al.* [18] to map characters onto trees sampled by the MCMC, as applied by the SIMMAP program (Bollback, [4]). However, by the (mixed) nature of the data used to infer the very trees, we did not rescaled branch lengths and use them as rates.

We then simulated 200000 character histories for each character of interest (leaf blade and corolla aestivation patterns) to explore the data in order to build hypotheses of character evolution.

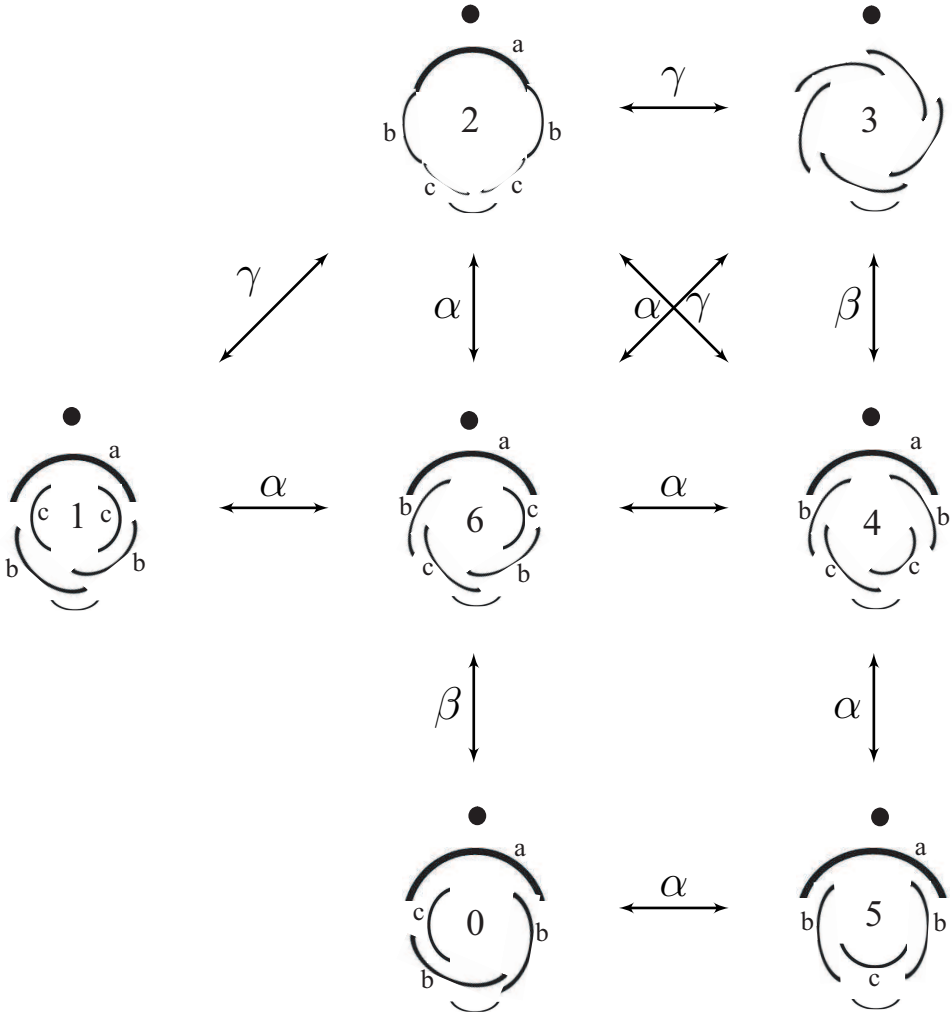


Figure 4.4 Relationships among corolla aestivation patterns (character 37). Greek letters represent (putative) different connections and denote both relative degrees of similarity and (putative) number of character state changes, and do not represent evolutionary pathways. (α indicates one character state change; β indicates two character state changes; γ indicates five character state changes). Roman letters represent homologous petals according to their relative topographical positions in the corolla. Numbers represent character states: 0 = proximal-cochleate imbricate, 1 = quincuncial imbricate, 2 = valvar, 3 = right-handed imbricate, 4 = descending distal-cochleate imbricate (5 petals), 5 = descending distal-cochleate imbricate (4 petals), 6 = proximal-cochleate imbricate.

4.5 Results and discussion

4.5.1 Tree searches

As the chains seemed to reach stationarity after 10 million generations (source files provided in the Supplemental Material C), our results are based on the last 5 millions generations.

4.5.2 Support, relationships, and synapomorphies

The monophyly of *Pilocarpus* was already found by Dias ([7]), and constrained or unconstrained searches did not change this result.

Some of the relationships within *Pilocarpus* are still unknown (low support, Figure 4.5(a)), however they have no influence on our character mapping procedure (although we are aware that some posterior probability values may be slightly biased, up or down, see, *e.g.*, Huelsenbeck *et al.* [17], Lewis *et al.* [23], Suzuki *et al.* [38]). As we can see on the Figure 4.5, the recovered tree was basically the same as the one found by Dias ([7]) with minor differences in support values for some nodes.

Morphologically, *Pilocarpus* can be distinguished easily from the other genera included in this study by its racemes (unbranched inflorescence, character 20, Figure 4.2). Furthermore, *Pilocarpus* has petals with inflexed apexes (character 42), erect stamens at anthesis (character 51), anthers bearing a postero-dorsal gland (character 58), an extra-staminal disc fully adnate to the ovary (character 64), and carpels with transversal ribs on the external surface (character 91).

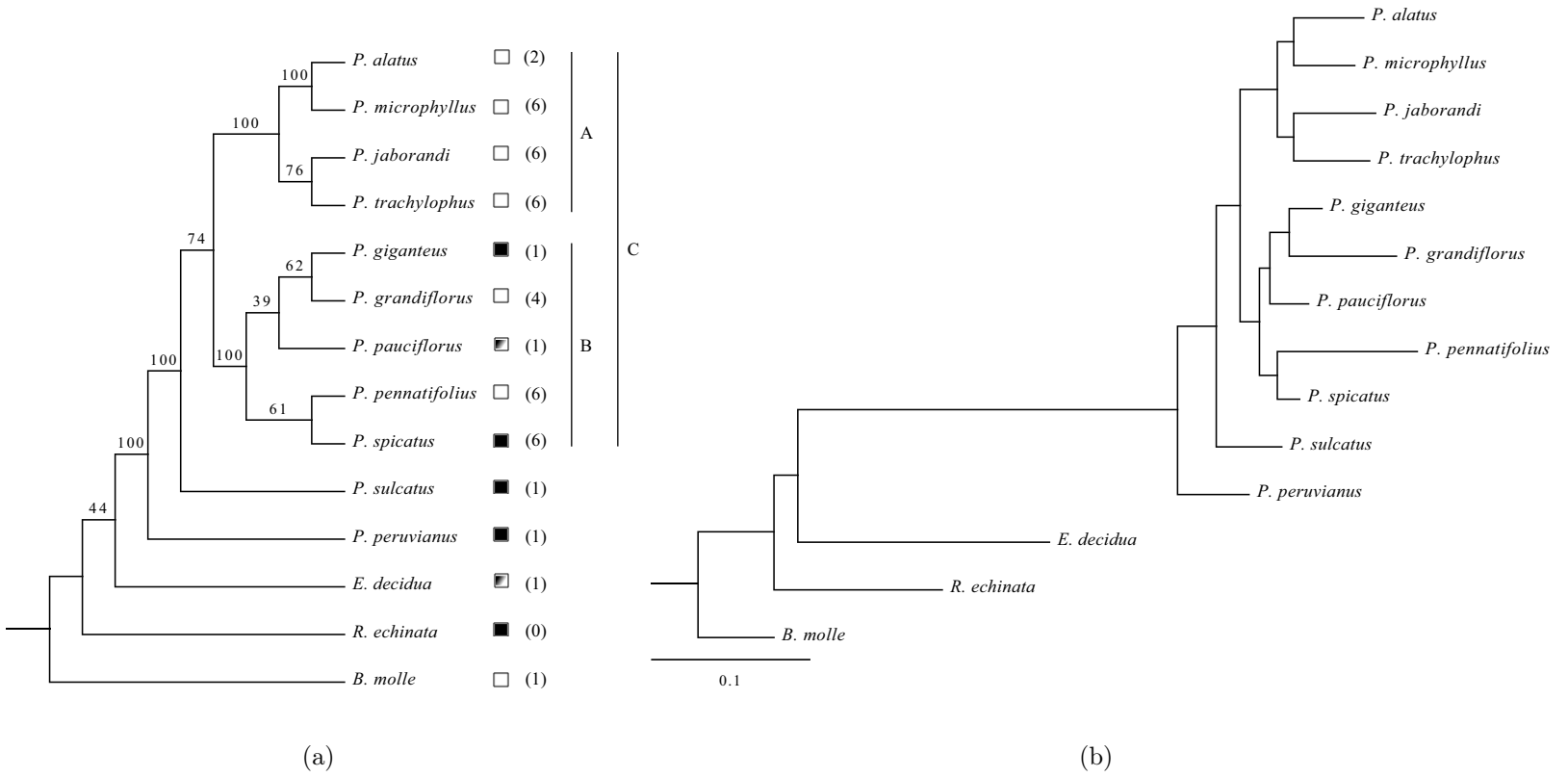


Figure 4.5 Extended majority consensus tree of the 20000 post-burn-in trees sampled by the Markov chains. (a) Cladogram. (b) Phylogram. Numbers above branches are posterior probabilities. White square = compound leaf, black square = simple leaf, black-and-white square = unifoliolate leaf. Numbers in parentheses represent the states of corolla aestivation (character 37): 0 = proximal-cochleate imbricate, 1 = quincuncial imbricate, 2 = valvar, 4 = descending distal-cochleate imbricate (5 petals), 6 = proximal-cochleate imbricate. A, B and C represent clades to be discussed under character mapping (see text).

4.5.3 Stochastic mapping and evolutionary hypotheses

Although using discrete morphological characters in probabilistic phylogenetic inference is not a new approach (*e.g.*, Felsenstein [10], Lewis [22], Losos [24], Martins [25], Pagel [29]), mapping characters onto trees using fully Bayesian methods has only recently been developed (Huelsenbeck *et al.* [18]). Such methodological advances have the benefit of being able to handle not only more realistic models of character change, but also the very uncertainty both in the data being mapped and in the phylogenetic trees being used to map them (Ronquist [32]).

As mentioned above, leaf blade patterns have posed some challenges for our procedure of character state delineation, and further exploration is needed. Additionally, corolla aestivation patterns represent a complex character, and revealing their relationships would help to understand what would be the basic pattern of *Pilocarpus* and, likewise, how this character have evolved within the genus. To accomplish these tasks, we applied the methods of Huelsenbeck *et al.* ([18]) to map the leaf and the corolla aestivation patterns (see Supplemental Material B, for detailed information), and hence to propose evolutionary hypotheses for these characters.

4.5.3.1 Leaf evolution

As we can note in the Figure 4.5, simple leaves would (as assumed) be a synapomorphy of the entire genus , whereas unifoliolate leaves would be an autapomorphy of *P. pauciflorus*. However, given that the basalmost relationship within the clade where *P. pauciflorus* emerged is unsolved (Figure 4.5(a), clade B), it is unclear what the condition at the basal node of that clade would be, whether sim-

ple or unifoliolate (or even compound). Therefore, one could ask why would simple/unifoliolate/compound be a synapomorphy or what is the probability of any state being a synapomorphy at that level of the phylogeny?

Thus, investigating character change at the basal node and on the branches immediately below and up that node, taking into account the associated tree uncertainty, is expected to clarify the appropriate status of either state as a synapomorphy at that node and demonstrate a major difference between parsimony and bayesian character mappings.

Because our interest is focused on the clade B, as a simplification for the mapping procedure, we will use the clade A as outgroup, and prune the other terminals. As all terminals of the clade A have compound leaves (state 1, white square), using a standard parsimony mapping would lead to independent changes[3] from compound to simple (Figure 4.6(a)). However, as the Figure 4.5(b) shows, the branches of *P. grandiflorus* and *P. pennatifolius* (both compound-leaved terminals) are more than twice as long as the branches of *P. giganteus* and *P. spicatus* (both simple-leaved terminals). Therefore, it is reasonable to assume that a change would occur more likely along these longer branches than along the shorter ones, and that is exactly what is suggested by the stochastic mapping procedure shown on Figure 4.6(b). While parsimony would fully ignore those branch lengths, they are taken into account when using the method of Huelsenbeck *et al.* ([18]), and, as we can see on Figure 4.6(b), the parsimonious reconstruction may not be the most likely to have happened on that tree.

Figure 4.7(a) shows a summary the results (for the entire trees, and all

[3]Note that the parsimony mapping is ambiguous if we consider the entire tree.

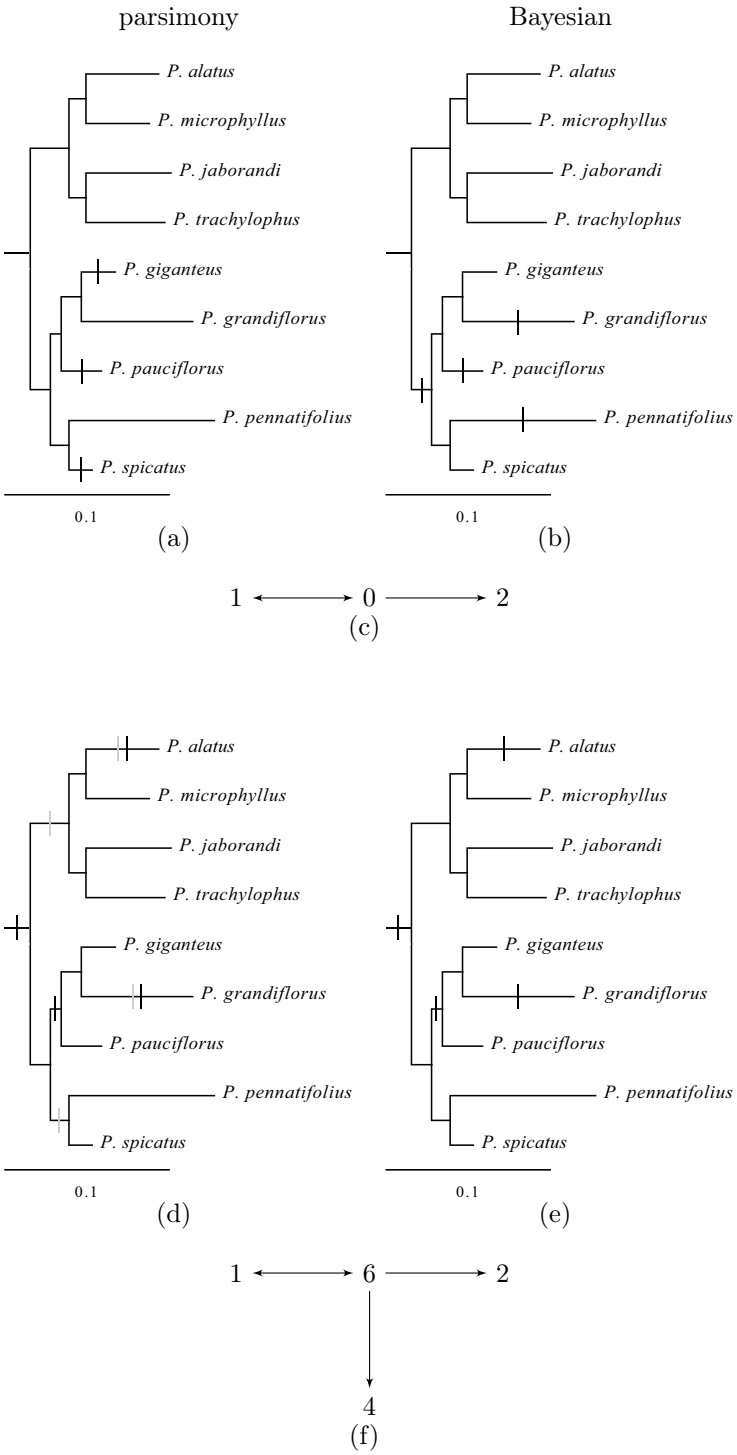


Figure 4.6 Optimization of the leaf blade (character 2) and corolla aestivation (character 37) patterns, vertical bars represent character state changes. (a) and (b) leaf blade. (d) and (e) corolla aestivation, ACCTRAN and DELTRAN optimizations are represented by black and gray bars, repectively. (c) and (f) character state tree as resulting from the optimization.

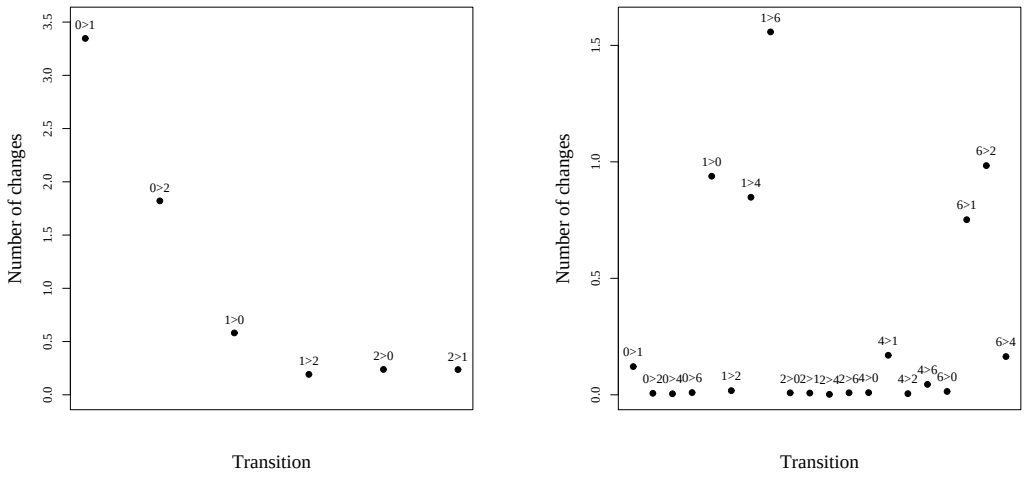
character state trasitions) of our 200000 character history simulations on all post-burn-in trees sampled by the Markov chains during the last 5 million generations of the tree searches. As we can see, the two most frequent transitions are (from a total of 6.4152 changes): 1) from simple to compound leaves ($0 \rightarrow 1$, number of changes = 3.3466) and 2) from simple to unifoliolate leaves ($0 \rightarrow 2$, number of changes = 1.8210), while a transition from compound to unifoliolate or vice versa (0.1918 and 0.2367, respectively) is far from the hilltop.

Our results suggest, therefore, that compound and unifoliolate leaves would have risen independently "from" simple ones in *Pilocarpus* (Figures 4.6(b) and 4.6(c)). They also suggest that any direct transition between compound and unifoliolate has a lower support, but that a transition from unifoliolate to compound has a higher posterior probability than the other way around. Accordingly, we assume that the state present at the basalmost node of the clade B is simple, and that compound and unifoliolate leaves originated on their own branches, as demonstrated in the Figure 4.6(b).

4.5.3.2 Corolla aestivation evolution

Unlike leaf blade, at the first glance corolla aestivation patterns (character 37) do not appear to represent a case of synapomorphy for any of the clades discussed before, and the importance of aestivation states may sometimes appear somewhat species-specific (*e.g.*, *P. alatus* and *P. grandiflorus*, Figure 4.5).

If we analyse the parsimony optimization for the clade C (*P. sulcatus* as outgroup), or even the entire tree and all terminals (including all outgroups), the state to be attributed to the basal node of clade C is ambiguous (both ACCTRAN and



(a) (b)

Figure 4.7 Number of all possible transitions among states of the leaf blade (character 2) and corolla aestivation (character 37) patterns. (a) Character 2, 0 = simple, 1 = compound, 2 = unifoliolate. (b) Character 37, 0 = proximal-cochleate imbricate, 1 = quincuncial imbricate, 2 = valvar, 4 = descending distal-cochleate imbricate (5 petals), 6 =proximal-cochleate imbricate.

DELTRAN optimizations have a four-step cost), either proximal-cochleate imbricate (state 6) or quincuncial imbricate (state 1) could represent the ancestral state at this node. Note that assuming either parsimony optimization, ACCTRAN or DELTRAN, at the basal node of C will directly influence on how homology (synapomorphies/homoplasies) may be interpreted at the basal nodes of each of the clades A and B.

However, if one takes into account some degree of uncertainty coming from the interaction among neighbour terminals, we can see that the posterior mapping postulates the proximal-cochleate imbricate (state 6) as synapomorphy of the clade C (equivalent to the ACCTRAN optimization[4]), which not only lead the quincuncial imbricate (state 1) to be considered as independent reversals (on the branches of *P. giganteus* and *P. pauciflorus*), but also maximizes the information content of the character at this node (note that the $1 \rightarrow 6$ change is on the highest position on the Figure 4.7(b)). All the results of ancestral state reconstructions for the clades A, B, and C are provided in the Supplemental Material A.

This way, it becomes clear that whether or not a given state would be considered a synapomorphy is dictated by the probability of the sequence/order of character state change along the trees. Consequently, as some authors have stated before (*e.g.*, Harper [13]), it is evident that synapomorphy can be seen as a matter of probability.

[4]This should not be considered an endorsement of the ACCTRAN procedure.

4.6 Conclusions

This study of *Pilocarpus* phylogeny and morphological evolution used molecular and morphological characters, and was performed using the MCMC algorithm within the Bayesian paradigm. Our results strongly support the monophyly of the genus, as already found by a previous study (Dias [7]).

Our character history simulations for the leaf blade patterns suggest that compound leaves have been derived from simple ones. The unifoliolate pattern, in turn, is likely to have occurred as an independent modification from simple leaves during the evolutionary history of *Pilocarpus* along the branch leading to *P. pauciflorus*. Also, according to the mapping assumed here, compound and simple leaves may be synapomorphies (the last as a reversal) of major clades in the genus (but see Supplemental Material A).

For the corolla aestivation evolutionary history, our simulations indicated that quincuncial imbricate is the pattern that may have occurred at the basalmost node of the genus, but it is suggested as a symplesiomorphy in *Pilocarpus*, while proximal-cochleate imbricate would be a synapomorphy of a major clade in the genus, and the other patterns exemplify later changes from this basic condition.

Furthermore, our stochastic mappings also show that the putative order of diversification is clearly dictated by the posterior probabilities associated with each transition between states and, therefore, whether or not a character may be considered a synapomorphy is clearly a matter of probability.

4.7 Acknowledgements

The authors are grateful to the Curators of F, HUEFS, IAN, INPA, MBM, MG, NY, RB, and TEGB for loans of herbarium specimens. We also thank Roseli .A. Leandro, from the Department of Statistics - ESALQ/USP, and Jacquelyn Kallunki, from The New York Botanical Garden, for their critical reading of an earlier draft of this manuscript and providing many helpful suggestions, although we are the only responsible for any mistakes. PD was supported by FAPESP (02/09762-6 & 04/15141-0), RGU by CNPq (140945/2004-0), and JRP by CNPq (304726/2003-6) & FAPESP (04/15141-0).

4.8 References

[1] Araújo, E. F., Queiroz, L. P. & Machado, M. A. 2003. What is *Citrus*? Taxonomic implications from a study of cp-DNA evolution in the tribe Citreae (Rutaceae subfamily Aurantioideae). *Org. Divers. Evol.* 3: 55–62.

[2] Bharathan, G., Goliber, T. E., Moore, C., Kessler, S., Pham, T. & Sinha, N. R. 2002. Homologies in leaf form inferred from *KNOXI* gene expression during development. *Science* 296: 1858–1860.

[3] Bharathan, G. & Sinha, N. 2001. The regulation of compound leaf development. *Plant Physiol.* 127: 1533–1538.

[4] Bollback, J. P. 2006. SIMMAP: stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics* 7: 88.

[5] Chase, M. W., Morton, C. M. & Kallunki, J. A. 1999. Phylogenetic relationships of Rutaceae: a cladistic analysis of the subfamilies using evidence from rbcL and atpB sequence variation. *Amer. J. Bot.* 86: 1191–1199.

[6] Daly, D. C. & Mitchell, J. D. 2000. Lowland vegetation of tropical South America - an overview. *In* Lentz, D. (ed.) *Imperfect balance: landscape transformations in the pre-Columbian Americas.* Columbia University Press, New York, 391–454.

[7] Dias, P. 2007. *Filogenética de Pilocarpinae (Rutaceae).* Tese de doutorado, Universidade de São Paulo, São Paulo.

[8] Duretto, M. F. & Ladiges, P. Y. 1999. A cladistic analysis of *Boronia* section *Valvatae* (Rutaceae). *Austral. Syst. Bot.* 11: 635–665.

[9] Federici, C. T., D. Q. Fang, R. W. S. & Roose, M. L. 1998. Phylogenetic relationships within the genus *Citrus* (Rutaceae) and related genera as revealed by RFLP and RAPD analysis. *Theoret. Appl. Genetics* 96: 812–822.

[10] Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. *Syst. Zool.* 27: 401–410.

[11] Groppo, M. 2004. *Filogenia de Rutaceae e revisão taxonômica de* ***Hortia*** *Vand. (Rutaceae).* Tese de doutorado, Universidade de São Paulo, São Paulo.

[12] Hareven, D., Gutfinger, T., Parnis, A., Eshed, Y. & Lifschitz, E. 1996. The making of a compound leaf: genetic manipulation of leaf architecture in tomato. *Cell* 84: 735–744.

[13] Harper, C. W. 1979. A Bayesian probability view of phylogenetic systematics. *Syst. Zool.* 28: 547–553.

[14] Harris, J. G. & Harris, M. W. 2001. *Plant identification terminology: an illustrated glossary.* Spring Lake Publishing, Spring Lake.

[15] Hickey, L. J. 1979. A revised classification of the architecture of dicotyledonous leaves. *In* Metcalfe, C. & Chalk, L. (eds.) *Anatomy of the dicotyledons.* vol. 1, 2 ed. Clarendon Press, Oxford, 25–39.

[16] Hofer, J., Turner, L., Hellens, R., Ambrose, M. & Matthews, P.

1997. UNIFOLIATA regulates leaf and flower morphogenesis in pea. *Curr. Biol.* 7: 581–587.

[17] Huelsenbeck, J. P., Larget, B., Miller, R. & Ronquist, F. 2002. Potential applications and pitfalls of bayesian inference of phylogeny. *Syst. Biol.* 51: 673–688.

[18] Huelsenbeck, J. P., Nielsen, R. & Bollback, J. 2003. Stochastic mapping of morphological characters. *Syst. Biol.* 52: 131–158.

[19] Kaastra, R. C. 1982. Pilocarpinae (Rutaceae). *Fl. Neotrop. Monogr.* 33: 1–198.

[20] Kim, M., McCormick, S., Timmermans, M. & Sinha, N. 2003. The expression domain of PHANTASTICA determines leaflet placement in compound leaves. *Nature* 424: 438–443.

[21] Lenhard, M., Jurgens, G. & Laux, T. 2002. The *WUSCHEL* and *SHOOTMERISTEMLESS* genes fulfil complementary roles in *Arabidopsis* shoot meristem regulation. *Development* 129: 3195–3206.

[22] Lewis, P. O. 2001. A likelihood approach to estimating phylogeny from discrete morphological characters data. *Syst. Biol.* 50: 913–925.

[23] Lewis, P. O., Holder, M. T. & Holsinger, K. E. 2005. Polytomies and bayesian phylogenetic inference. *Syst. Biol.* 54: 241–253.

[24] Losos, J. B. 1994. An approach to the analysis of comparative data when a phylogeny is unavailable or incomplete. *Syst. Biol.* 47: 117–123.

[25] Martins, E. P. 1996. Conducting phylogenetic comparative studies when the phylogeny is not known. *Evolution* 50: 12–22.

[26] Merck. 2001. *Merck index: an encyclopedia of chemical, drugs and biologicals.* 13 ed. Merck, Rahway.

[27] Morton, C., Grant, M. & Blackmore, S. 2003. Phylogenetic relationships of the Aurantioideae inferred from chloroplast DNA sequence data. *Amer. J. Bot.* 90: 1463–1469.

[28] Nixon, K. C. & Carpenter, J. M. 1993. On outgroups. *Cladistics* 9: 413–426.

[29] Pagel, M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R. Soc. Lond. Ser. B* 255: 37–45.

[30] Pinheiro, C. U. B. 1997. Jaborandi (*Pilocarpus* spp. and Rutaceae): a wild species and its rapid transformation into a crop. *J. Econ. Bot.* 51: 49–58.

[31] Pinheiro, C. U. B. 2002. Extrativismo, cultivo e privatização do jaborandi (*Pilocarpus microphyllus* Stapf ex Holm.; Rutaceae) no Maranhão, Brasil. *Acta Bot. Bras.* 16: 141–150.

[32] Ronquist, F. 2004. Bayesian inference of character evolution. *TREE* 19: 475–481.

[33] Ronquist, F. & Huelsenbeck, J. P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19: 1572–1574.

[34] Samuel, R., Ehrendorfer, F., Chase, M. C. & Greger, H. 2001. Phylogenetic analyses of Aurantioideae (Rutaceae) based on non-coding plastid DNA sequences and phytochemical features. *Plant Biol.* 3: 77–87.

[35] Scott, K. D., McIntyre, C. L. & Playford, J. 2000. Molecular analyses suggest a need for a significant rearrangement of Rutaceae subfamilies and a minor reassessment of species relationships within Flindersia. *Plant Syst. Evol.* 223: 15–27.

[36] Skorupa, L. A. 1996. *Revisão taxonômica de* ***Pilocarpus*** *Vahl (Rutaceae).* Tese de doutorado, Universidade de São Paulo, São Paulo.

[37] Skorupa, L. A. & Pirani, J. R. 2004. A new species of *Pilocarpus* (Rutaceae) from northern Brazil. *Brittonia* 56: 147–150.

[38] Suzuki, Y., Glazko, G. V. & Nei, M. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proc. Nat. Acad. Sciences* 99: 15138–16143.

[39] Weberling, F. 1989. *Morphology of flowers and inflorescences.* Cambridge University Press, Cambridge.

Appendices

A Data matrix

Table 4.2 Data matrix, polymorphisms are indicated as A={0,1} and B={1,2}.

B. molle	01100?0000010001112100000112220110101111013210010011100000011110000101112011111101111100011 10
E. decidua	02000??20010??0????2000001111101101010210212110100111000000010110100010120122232123012010021 1
P. alatus	0100011001110002121000000101211101102021000 10A000000100001000001 1A00100000000000100011010100 0
P. giganteus	00000??2201022033??0101101111101001040010000 0A000000100000001001 1A00100000100000100011010102 1
P. grandiflorus	0100000021010001212011001111111110001000101101000001100001100001111010012010000010001111010 20
P. jaborandi	0100111201110101212 0A1000111111000106001001101010000100000010001111001011110000010001001010 20
P. microphyllus	010001100110000????0000001111100101060010001010100011000000000001011001A000000000100011010102 0
P. pauciflorus	02000??10010??0????000000A1111101010100100000100000010000000000011100110000100000100011010120 0
P. pennatifolius	0100111201110101222001000A1111000010600100110101000010000001000101000101111000001000000111102 2
P. peruvianus	00000??0201002033??0010011111100011010010000001000000100001001001010011000010000010001101010 21
P. spicatus	00000??0001022033??0010001111100101060010000 0A000000100021000001 1A00110000100000100011010102 0
P. sulcatus	00000??2201022133??0000000111101001010010000010000 0A100000001001110011000010000010001101 01?21
P. trachylophus	0100000211010021?2B000000A1111010010600010100100000010000011100111001111120100000100001011102 0
R. echinata	10100??0100122133??0100001111201101000210210000100110000000110000001010120123223123112010001 0

B Description of the 94 characters

Description of the 94 characters used in this study. Unless otherwise stated, the terms used follow Hickey [15] for leaves, Weberling [39] for flowers and inflorescences, and Harris & Harris [14] for the remaining characters.

Stem

1. Thorns: (0) absent; (1) present.

Leaf

2. Blade pattern: (0) simple; (1) compound; (2) unifoliolate.

3. Attachment: (0) alternate; (1) opposite. The sub-opposite condition, which is often described in the botanical literature as another pattern, is rather the alternate pattern with shorter internodes. Thus, we treated it under the alternate state.

4. Surcurrent base: (0) absent; (1) present. Here we used succurrent base (Harris & Harris [14]) as an appropriate alternative to the term sheath, which is found in the literature (*e.g.*, Kaastra [19]) describing the leaf base in *Metrodorea*.

5. Venation: (0) brochidodromous; (1) eucamptodromous. Here we applied a topography-based primary homology conjecture between the whole blade of the simple or unifoliolate leaves and the terminal foliole of the compound leaves.

6. Rachis - wings: (0) absent; (1) present.

7. - appendage beyond the most apical pair of folioles: (0) absent; (1) present.

8. Primary vein - adaxial surface: (0) convex ; (1) plane; (2) concave. Here

we applied a topography-based primary homology conjecture between the central vein of the simple or unifoliolate leaves and the rachis of the compound leaves.

9. Secondary veins - adaxial surface: (0) convex; (1) plane; (2) concave. Here we applied a topography-based primary homology conjecture between the secondary veins of the simple or unifoliolate leaves and the primary vein of the lateral folioles of the compound leaves.

10. Petiole - cross-section - form: (0) semi-terete; (1) terete.

11. - wings: (0) absent; (1) present.

12. - canalicule: (0) absent; (1) present.

Lateral leaflets

13 Blade - surface: (0) plane; (1) bullate.

14 - margin: (0) smooth; (1) crenulate.

15. - Secondary veins - adaxial surface: (0) convex; (1) plane; (2) concave. Here we applied a topography-based primary homology conjecture between the secondary veins of the lateral folioles of compound leaves and the tertiary veins of the simple leaves.

16. Peciolule - wings: (0) absent; (1) present.

17. - canalicule: (0) absent; (1) present.

Terminal leaflet

18. - Petiolule - wings: (0) absent; (1) present;

19. - canalicule: (0) absent; (1) present;

Inflorescence

20. - branching: (0) non-branched; (1) opposite branching; (2) alternate branching. Here we opt for scoring the branching patterns as a multistate character to avoid character duplication.

21. - position: (0) terminal; (1) lateral.

22. - apex orientation at early development: (0) erect; (1) declined.

23. - cauliflory : (0) absent; (1) present.

Rachis

24. - canalicules: (0) absent; (1) present.

25. - longitudinal ribs : (0) absent; (1) present.

Flower

26. - development: (0) acropetal; (1) basipetal.

27. - pedicel: (0) absent (sessile); (1) present (pedicellate).

28. - basal bract: (0) absent (no bracts); (1) one; (2) two.

29. - bracteole - attachment: (0) alternate; (1) opposite.

30. - form: (0) ovate; (1) lanceolate;

Calyx

31. - sepal union : (0) free ; (1) connate.

32. - aestivation: (0) valvar; (1) quincuncial imbricate; (2) contort imbricate.

33. - symmetry: (0) zygomorphic; (1) actinomorphic.

34. - sepal apex: (0) triangular; (1) rounded.

35. - sepal - abaxial surface: (0) carinate; (1) smooth.

Corolla

36. petal union: (0) free; (1) connate.

37. aestivation: (0) proximal-cochleate imbricate; (1) quincuncial imbricate; (2) valvar; (3) right-handed imbricate; (4) descending distal-cochleate imbricate (5 petals); (5). descending distal-cochleate imbricate (4 petals); (6) proximal-cochleate imbricate

Petals

38. - curvature at anthesis: (0) reflexed; (1) patent.

39. - form: (0) ovate; (1) oblong; (2) lanceolate.

40. - abaxial surface: (0) carinate; (1) smooth.

41. - adaxial surface: (0) carinate; (1) smooth.

42. - apex - curvature at anthesis: (0) inflexed; (1) patent; (2) reflexed.

43. - venation: (0) acrodromous; (1) actinodromous; (2) cladodromous; (3) camptodromous.

44. primary vein(s) - prominence- adaxial surface : (0) flat (non-prominent); (1) prominent through a half of the blade length; (2) prominent throughout the blade length.

45. - glands: (0) absent; (1) present.

Androecium

46. number of stamens: (0) 4; (1) 5; (2) 7.

47. staminode: (0) absent; (1) present.

Fertile stamens

48. filament - form: (0) linear; (1) subulate.

49. - union: (0) free; (1) connate.

50. - adnation to petal: (0) free; (1) adnate.

51. - curvature at anthesis: (0) erect; (1) reflexed.

52. - apex - form: (0) acute; (1) rounded.

53. - base - adherence to disc: (0) free; (1) adherent.

54. - anthers - attachment: (0) dorsifixed; (1) basifixed.

55. - connective - basal appendage: (0) absent; (1) present.

56. - union: (0) free; (1) connate.

57. - form: (0) ovate; (1) oblong; (2) orbicular.

58 - dorso-apical gland: (0) absent; (1) present.

59. - apical lobes - curvature at anthesis: (0) erect; (1) declined.

60. - basal lobes - curvature at anthesis: (0) erect; (1) incurved.

61. - mobile anther: (0) absent; (1) present.

Disc

62. - lateral expansions (plicate disc): (0) absent; (1) present.

63. - form: (0) annular; (1) cupulate.

64. - union to carpels: (0) free; (1) adpressed; (2) adnate.

65. - glands: (0) absent; (1) present.

66. - elongated, unbranched, unicellular trichome: (0) absent; (1) present.

67. Floral merism: (0) 4; (1) 5.

Gynoecium

68 - gynophore: (0) absent; (1) present.

69. - carpel - union: (0) connate at base; (1) fully connate.

70. - elongated, unbranched, unicellular trichome: (0) absent; (1) present.

71. - lateral expansions: (0) absent; (1) present.

72. - apex extensions: (0) absent; (1) present.

73. - ovule - number per carpel: (0) 1; (1) 2.

74. - arrangement at anthesis: (0) 1 ovule; (1) 2 superposed; (2) 2 colateral.

75. - stigma - form: (0) capitate; (1) clavate.

76. - lateral expansions: (0) absent; (1) present.

Fruit

77. - type: (0) schizocarp; (1) samara; (2) capsule. Here we applied the often used term schizocarp to define *Pilocarpus* fruits. That term is commonplace in the literature (*e.g.*, Kaastra [19], Skorupa [36]), albeit the segments (carpels) open tardily at maturity.

78. - apex: (0) concave; (1) convex;

79. - lobes: (0) absent; (1) present;

80. - number: (0) 4; (1) 5;

81. - murication: (0) absent; (1) present.

82. - loculicidal dehiscence: (0) absent; (1) present.

83. - septicidal dehiscence: (0) septum absent; (1) absent; (2) present.

84. - dorsal apophysis: (0) absent; (1) present.

Carpel (at maturity)

85. - form: (0) obovate; (1) elliptic

86. - dorsal surface: (0) plane; (1) rounded.

87. - apex - form: (0) plane; (1) rounded; (2) angled.

88. - rostrum on the ventral surface: (0) absent; (1) present.

89. - epicarp - external surface - glands: (0) absent; (1) present.

90. - longitudinal ribs : (0) absent; (1) present.

91. - transversal ribs: (0) absent; (1) present.

Seed

92. - form - frontal view: (0) ovoid; (1) oblong; (2) ellipsoid.

93. - lateral view - median-ventral region: (0) plane; (1) convex; (2) concave.

94. - supero-ventral region - rostrum: (0) absent; (1) angled; (2) rounded.

C MrBayes block used

```
BEGIN MRBAYES;
 LOG START FILENAME=pilocarpus_comb_bayes_const.log REPLACE;

     [***ROOT***]
 OUTGROUP B_MOLLE;

     [***DEFINE CHARACTER GROUPS***]
 CHARSET ITS = 1-656;
 CHARSET TRNG_S = 657-1636;
 CHARSET MORPH_VEG =1637-1655;
 CHARSET MORPH_INFL = 1656-1661;
 CHARSET MORPH_FWR = 1662-1712 ;
 CHARSET MORPH_FR = 1713-1727;
 CHARSET MORPH_SEED = 1728-1730;

     [***DEFINE PARTITIONS***]
 PARTITION GENES_MORPH = 7:TRNG_S, ITS, MORPH_VEG, MORPH_INFL, MORPH_FWR, MORPH_FR, MORPH_SEED;

     [***SET PARTITIONS***]
 SET PARTITION=GENES_MORPH;

     [***SHOW TAXON INFO AND MATRIX***]
 TAXASTAT;
 SHOWMATRIX;

     [***DEFINE TAXON GROUPS (BASED ON THE SUBTRIBE ANALYSIS) TO BE USED AS CONSTRAINTS***]
     [***ONLY NODES WITH PP >= 75%***]
 [1] TAXSET PAM = P_alatus P_microphyllus;[HERE PAM IS NOT THE MATRIX OF GENETIC DISTANCES! :-)]
 [2] TAXSET PEARL_JAM = P_alatus P_jaborandi P_microphyllus P_trachylophus;[MAYBE IN HONOUR TO THE BAND]
 [3] TAXSET BIGGEST = P_giganteus P_grandiflorus P_pauciflorus P_pennatifolius P_spicatus;
     [THIS IS THE BIGGEST - IT HAS 'P_giganteus']
 [4] TAXSET NO_PERUVIAN_ALLOWED = P_alatus P_giganteus P_grandiflorus P_jaborandi P_microphyllus
     P_pauciflorus P_pennatifolius P_spicatus P_sulcatus P_trachylophus;[SELF-EXPLANATORY]
 [5] TAXSET PILOCARPUS = P_alatus P_giganteus P_grandiflorus P_jaborandi P_microphyllus P_pauciflorus
     P_pennatifolius P_peruvianus P_spicatus P_sulcatus P_trachylophus;[IDEM]

     [***SET CONSTRAINTS***]
     [***ONLY NODES WITH PP >= 75%***]
 [1] CONSTRAINT C_PAM 99 = PAM;
 [2] CONSTRAINT C_PEARL_JAM 100 = PEARL_JAM;
 [3] CONSTRAINT C_BIGGEST 83 = BIGGEST;
 [4] CONSTRAINT C_NO_PERUVIAN_ALLOWED 99 = NO_PERUVIAN_ALLOWED;
 [5] CONSTRAINT C_PILOCARPUS 100 = PILOCARPUS;


     [***MODELS***]
     [***DNA***]
 LSET
  APPLYTO = (1,2)  NST=6 RATES=INVGAMMA;

     [***MORPHOLOGY***]
 LSET
  APPLYTO = (3,4,5,6,7)  NBETACAT=7 CODING=VAR;

 UNLINK STATEFREQ=(ALL) REVMAT=(ALL) SHAPE=(ALL) PINVAR=(ALL);

     [***PRIOR ON PRMS***]
     [***MAKE SURE THE CONSTRAINTS ARE EFFECTIVE***]
 PRSET
  APPLYTO=(ALL)
  RATEPR=VARIABLE
  STATEFREQPR=DIRICHLET(1)
  BRLENSPR=UNCONSTRAINED:EXPONENTIAL(10.0)
  TOPOLOGYPR=CONSTRAINTS(C_PAM, C_PEARL_JAM, C_BIGGEST, C_NO_PERUVIAN_ALLOWED, C_PILOCARPUS);

     [***SEARCH PRMS***]
 MCMC NRUNS=4 NGEN=5000000 SAMPLEFREQ=1000 PRINTFREQ=1000 NCHAINS=4 TEMP=0.2
  FILENAME= pilocarpus_comb_bayes_const SAVEBRLENS=YES;
```

```
    [***SHOW MODEL***]
 SHOWMODEL;

    [***SUMMARIZE RESULTS***]
 SUMP
  BURNIN=500
  NRUNS=4
  PRINTTOFILE=YES;
 SUMT
  BURNIN=500
  NRUNS=4
  SHOWTREEPROBS=YES
  CONTYPE=ALLCOMPAT;

 LOG STOP;
END;
```

D Specimens used

Balfourodendron molle (Miq.) Pirani

D. Andrade-Lima 52-1025 (SPF), 70-5864 (SPF), 72-7189 (SPF); O.G. Araújo Filho 320 (SPF); A.M. Carvalho 3838 (SPF); L.V. Costa (SPF 152764); M.C. Ferreira 513 (SPF); M.R. Fonseca 1292 (SPF); A.M. Giulietti 1741 (SPF), CFCR6950 (SPF); A.P.S. Gomes 331 (SPF); R.M. Harley 16374 (SPF), 27142 (SPF); G. Hatschbach 56546 (SPF), 69925 (SPF); M.B. Horta (SPF 83096), (SPF 159211); J.A. Kallunki 406 (SPF); A.M. Miran 1488 (SPF); J.R. Pirani CFCR349 (SPF), 4265 (SPF); L.P. Queiroz 7301 (SPF); R.A. Silva 1438 (SPF); M. Sobral 7637 (SPF); J. Souza-Silva 699 (SPF); Teixeira (SPF 111109).

Esenbeckia decidua Pirani

R. Mello-Silva 770 (MBM, SPF), V.C. Souza 5454 (SPF); N.P. Taylor 1489 (SPF).

Pilocarpus alatus Joseph *ex* Skorupa

D.C. Daly D465 (SPF); P. Dias 247 (SPF); F.H. Muniz B1751 (SPF); L.A. Skorupa 1024 (SPF).

Pilocarpus giganteus Engl.

G.S. França 365 (SPF); J.A. Kallunki 698 (SPF); S.E. Martins 672 (SPF).

Pilocarpus grandiflorus Engl.

J.G. Jardim 2228 (SPF), 2623 (SPF), 3064 (SPF), 3084 (SPF); J.A. Kallunki 373 (SPF); J.R. Pirani 1113 (SPF), 4676 (SPF); T.S. Santos 4527 (SPF); W.W.

Thomas 6069 (SPF).

Pilocarpus jaborandi Holm.

Anonymous (RB 46768); P. Dias 252-256 (SPF).

Pilocarpus microphyllus Stapf *ex* Wardl.

P.B. Cavalcante 2678 (IAN); P. Dias 235 (SPF), 237 (SPF), 238 (SPF), 239 (SPF), 240 (SPF); A. Fernandes (INPA 188979); R.L. Fróes 34933 (IAN); J.G.S. Maia 19 (MG); E.F. Miranda 460 (INPA); R.S. Monteiro 487 (MG); F.H. Muniz B1253 (SPF); T. Plowman 9774 (INPA); B.G.S. Ribeiro 1338 (IAN); R.S. Secco 160 (MG), 479 (MG); E.G. Silva 1 (IAN); C.R. Sperling 6374 (MG); E.L. Taylor E1247 (SPF); R.F. Vieira 803 (SPF), 848 (SPF), 850 (SPF), 853 (SPF), 855 (SPF), 856 (SPF), 866 (SPF), 867 (SPF), 888 (SPF), 891 (SPF), 905 (SPF), 908 (SPF), 912 (SPF).

Pilocarpus pauciflorus St.-Hil.

O.T. Aguiar 147 (SPF); R.J. Almeida-Scabbia (SPF 114330), (SPF 114331); A.M. Amorim 1412 (SPF); P.F. Assis 228 (SPF); C.T. Assumpção (SPF 16356); K.D. Barreto 2220 (SPF); Cesar (SPF 32626); S.A.C. Chiea 673 (SPF); I. Cordeiro (SPF 46655); P. Dias 218 (SPF); H.Q.B. Fernandes 1333 (SPF); A. Gehrt (SPF 123623); G. Hatschbach 16324 (SPF); J.A. Kallunki 395 (SPF); M. Kuhlmann 872 (SPF), 3825 (SPF); M.F.R. Melo 680 (SPF); J.R. Pirani 2844 (SPF), 3233 (SPF); J.R. Stehman 1484 (SPF); J.Y. Tamashiro 1179 (SPF), 1229 (SPF); V.B. Zipparro (SPF 112896).

Pilocarpus pennatifolius Lem.

P.T. Alvim 1 (SPF); I. Andó 15 (SPF), 91 (SPF); Anonymous (SPF 143437),

(SPF 76982); M.M. Arbo 5999 (SPF); M.A. Assis (SPF 76983), (SPF 143439); A.E. Brina (SPF 122481); E.L.M. Catharino 321 (SPF); P. Dias 215 (SPF); H.H. Faria 108 (SPF); E.C. Fonseca (SPF 75962); F. França 2566 (SPF); G. Hatschbach 43200 (SPF), 49111 (SPF); A. Krapovickas 44142 (SPF); M.S. Pereira 15 (SPF); J.R. Pirani 612 (SPF); A.T. Shimoda (SPF 130644); A. Souza 1 (SPF), (SPF 143440); E. Tameirão Neto 1976 (SPF); M.E.B. Thomaz (SPF 130643); S.G. Tressens 3302 (SPF); R. Záchia 3140 (SPF).

Pilocarpus peruvianus (Macbr.) Kaastra

L. de Lima 597 (NY); A. Ducke (RB 23828); R.B. Foster 5804 (F), 11478 (NY); J.A. Kallunki 1983 (NY); E. Milgaki 73 (RB); M. Nee 34417 (NY), 49581 (NY); P. Nuñez 1804 (F); H.H. Rusby 2072 (NY); J.S. Vigo 3290 (F), 8180 (NY); L. Willians 4878 (F).

Pilocarpus spicatus St.-Hil.

A.M. Amorim 1054 (SPF); M.J.G. Andrade 181 (HUEFS); D. Araújo 8945 (SPF), 9761 (SPF), 9944 (SPF), 10230 (SPF), 10455 (SPF); F.S. Araújo 1333 (HUEFS), 1358 (HUEFS); J. Badini (SPF 133953); C.N. Fraga 652 (SPF); L.S. Funch FCD129 (HUEFS); A.M. Giulietti 2002 (SPF), 2083 (HUEFS); M.L. Guedes 8184 (HUEFS); G. Hatschbach 43962 (SPF), 65922 (SPF), 68298 (SPF); W. Hoehne 5951 (SPF), 5982 (SPF); J.G. Jardim 675 (SPF); J.A. Kallunki 334 (SPF), 577 (SPF), 635 (SPF); M. Kuhlmann (SPF 123626); W.P. Lopes 681 (SPF); L.A. Mattos Silva 2248 (SPF); P.H.A. Melo 28 (SPF); A. Oliveira 105 (HUEFS); C.A.L. Oliveira 2317 (SPF); O.J. Pereira 1033 (SPF); J.R. Pirani 2861 (SPF), 3392 (SPF), 4995 (SPF); L.P. Queiroz 4235 (SPF), 5255 (SPF), 9406 (SPF), (SPF 146365); T.S. Santos 4545 (SPF); B.

Stannard (SPF 91616); E. Tameirão Neto 3155 (SPF), 3244 (SPF); W.W. Thomas 11944 (SPF); R.F. Vieira 1201 (SPF).

Pilocarpus sulcatus Skorupa

M. Andrade (SPF 99185); D.S. Carneiro Torres 39179 (SPF); T.R.S. da Silva 126 (HUEFS); G. Hatschbach 65173 (SPF), 78478 (SPF); L.P. Queiroz 1593 (SPF); E. Saar (SPF 130166), (SPF 153929); V.C. Souza 26058 (MBM); N.P. Taylor 1488 (SPF).

Pilocarpus trachylophus Holm.

M. Andrade (SPF 99184), (SPF 99188); O.F. Carlos 32 (SPF); E.R. de Souza 291 (HUEFS); L. Emperaire (TEGB 2795); A. Furlan (SPF 22440); W. Ganev 800 (SPF); A.M. Giulietti (SPF 117242), (SPF 153930); R.M. Harley 21468 (SPF), 51555 (SPF), 51598 (SPF), 51881 (SPF), 51943 (SPF), 51998 (SPF), 54358 (HUEFS), 54360 (HUEFS); G. Hatschbach 65177 (SPF), 65875 (SPF), 67717 (SPF); J.A. Lombardi 1754 (SPF); E.B. Miranda Silva 214 (SPF); G. Pereira-Silva 8448 (SPF); L.P. Queiroz 5795 (HUEFS), 5973 (HUEFS); A. Salino 3334 (SPF); V.C. Souza 5396 (SPF); E. Tameirão Neto 617 (SPF).

Raulinoa echinata Cowan

Anonymous (SPF 134300); M.W. Biovatti 27 (SPF); P. Dias 258-263 (SPF).

Supplemental Material

Please, see the DVD-ROM.

A Ancestral state reconstructions

/Cap4/AncStates/

B Character history simulations

SIMMAP output files

/Cap4/Simulations/

C Tree searches

MrBayes output files

/Cap4/TreeSearches/

Part III

Novidades Taxonômicas em Esenbeckiinae

Capítulo 5

Re-description and Epitypification of *Esenbeckia cowanii* (Rutaceae)

Artigo submetido à *Novon* em Junho de 2007.

5.1 Abstract

A re-description of *Esenbeckia cowanii* Kaastra (Rutaceae) is presented and, as the type material of this taxon is ambiguous for identification purposes, an epitype is designated. In addition, we provide an updated picture of its geographic distribution, since this species was known only from the type locality.

5.2 Resumo

Neste trabalho é apresentada uma redescrição de *Esenbeckia cowanii* Kaastra (Rutaceae) e, dado que o material-tipo do táxon é ambíguo, um epítipo é designado. Adicionalmente, é fornecida uma ilustração detalhada da espécie e, uma vez que a espécie era conhecida apenas da localidade-tipo, um mapa com sua distribuição conhecida atualizada é apresentado.

5.3 Introduction

The genus *Esenbeckia* (subtribe Esenbeckiinae, tribe Galipeeae, subfamily Rutoideae, Rutaceae) comprises 28 species, and is widely distributed in the Neotropical region. A monograph to the genus was presented by Kaastra [3] for the Flora Neotropica Monograph series, and recent work by one of us (Pirani, [5]) has contributed with new species to the genus.

Esenbeckia cowanii was described by Kaastra [2] based on two collections from the same locality in French Guiana, Cowan 38833 (deposited at F, NY, and US) and 38757 (deposited at K, NY, and P). Nevertheless, Kastra [2] described his new taxon without knowing its flower morphology. By the fact that all of these specimens are non-flowering, they do not allow unambiguous recognition of the taxon. Thus, according to the International Code of Botanical Nomenclature (McNeill *et al.*, [4]), an epitype may be designated.

During recent field work in the Amazon basin and Central Brazil, we rediscovered *E. cowanii* Kaastra in Acre, Mato Grosso, and Rondônia States. Furthermore, while studying the collections of some herbaria, we found that several specimens of *E. cowanii* were deposited at the major Amazonian herbaria (IAN, INPA, and MG) as well as at NY herbarium. However, all of these specimens were misidentified as *E. almawillia* Kaastra, probably for its lateral infructescences, holding 1-2 mature fruits each.

In this paper, besides designating an epitype, we provide a re-description of *E. cowanii*, as to include its floral morphology and details of the fruit and seed, and update its geographic distribution.

5.4 Material and methods

For the morphological descriptions, we used our own collections (deposited at SPF) and additional herbarium specimens from F, IAN, INPA, MG, NY, and US. We used only fully developed structures. Flower and fruits were fixed with FAA in the field (for our own collections) or rehydrated (for herbarium specimens) before measurements and sketches were made. Unless otherwise stated, terms used follow Weberling [6] for flowers and inflorescences, and Hickey [1] for leaves.

5.5 Results and discussion

5.5.1 *Esenbeckia cowanii* Kaastra

Esenbeckia cowanii Kaastra, Acta Bot. Neerl. 26: 481, t. 7, 1977. - Holotype: French Guiana, Guyane, Montagne de Kaw, 250-270 m, 14 Dec 1954, fr., R. S. Cowan 38833 (US; isotypes, F, NY). - Epitype (designated here): Brazil, Mato Grosso, Vila Bela da Santíssima Trindade, estrada para Cachoeirinha (cascata), 17 km de Vila Bela da Santíssima Trindade, Parque Estadual "Serra de Ricardo Franco", cerca de 500 m, na margem direita do riacho, solo areno-argiloso, com afloramentos rochosos, mata de terra firme, 16 Nov 2005 (fl., fr.), P. Dias & R.G. Udulutsch 227, sheets A and B (SPF). Figure 5.1.

Tree 5–15 m tall with brown-hyaline pubescence of appressed hairs to 0,4 mm long; branchlets 3–6 mm in diam., greenish brown and puberulent when young, then glabrescent. with eliptic, pale lenticels. Leaves alternate, unifoliolate, with sessile leaflet; petiole subterete, slightly winged or not, 4–19 mm long, the base slightly

tumid, puberulent with appressed trichomes; leaf blade elliptic, (3.1-) 6.1–21.7 x 2.5–9.5 cm, acute to rounded at base, acuminate (acumen obtuse to retuse) at apex, the margin entire and flat (slightly subrevolute in silico), chartaceous, adaxial surface lustrous when fresh, puberulous to glabrous only on the midvein, abaxial surface paler and dull, puberulous to glabrous only on the midvein; venation brochidodromous, primary and secondary veins impressed to plane above and prominent beneath. Inflorescences lateral, not axillary (sometimes opposite to the leaf itself), along the branchlets, erect, paniculate, shorter than leaves, 5–15 x 3–10 cm, densely puberulent with small, appressed trichomes (to 0.4 mm long); side-branchlets (paraclades) alternate, 1–5 cm long, erect; bracts caducous, densely puberulent; pedicels 4.35–15 mm long, puberulent; bractlets (prophylls) caducous, 1 per flower, c. 1 mm long, triangular, puberulent. Flowers 4.2–5 mm in diam.; calyx lobes (4-)5(-6), persistent at young fruits, quincuncial, deltoid, 0.8–1.4 mm long, apex acute to atenuate, coriaceous, puberulent, venation hyphodromous (vein visible only on the adaxial surface); petals (4-)5, persistent at young fruits, quincuncial, widely spreading, ovate to slightly oblong, 2–2.7 x 0.9–1.2 mm, acute at apex, coriaceous, pale yellow (cream), puberulent on abaxial surface, glabrous on adaxial surface, venation hyphodromous (vein visible only on the adaxial surface); filaments (4-)5, sometimes persistent at young fruits, subulate, 1.6–2.05 mm long, glabrous, reflexed after anthesis; anthers heart-shaped, versatile, c. 0.8 mm long including the small mucronate tip, glabrous, early deciduous; disc annular, (4-)5-lobed, each lobe slightly 2-lobed, 1.8–2.1 mm in diam., fleshy, provided with some tiny glands, puberulent, pale yellow; carpels (4-)5, 0.9–1.2 mm high, adnate to the disc through the lower half and connate proximally,

free and ellipsoid distally, the free part protruding c. 0.1 mm beyond the disc, furnished with tiny glands, puberulent; ovules 2 per locule, collateral; style inserted proximally on the carpels but nearly completely free, 1.4–1.7 mm long, glabrous; stigma clavate. Fruit 1–2 per infructescence, a depressed, stellately 5-lobed capsule, 11.8–21.95 x 10.05–27 mm, puberulent, nerved externally when young, then becoming slightly smooth (except for the caducous dorsal apophysis), dehiscent septicidally along the dorsal commissures from apex down to 2–4 mm above the base or even reaching the very base, and loculicidally from the base along the ventral sutures up to an obtuse to rounded apophysis (2/3 to 1/2 from the base), slightly nerved to smooth on the inside of the mesocarp; endocarp smooth, ochre-yellow, hard, elastically (or explosively) dehiscent; seed 1–2 per locule (if 2, then superposed), ovoid, 8.5–13 x 4.8–5,7 mm (if 1 seed) or 5.5–7.3 x 5.1–6 mm (if 2 seeds), beaked at apex, testa brown, hilum narrow, elongated, running from apex down to the chalazal area, irregular in shape, dark brown to black; embryo not seen.

As stated by Kaastra [3], *E. cowanii*, together with *E. almawillia*, is included in *Esenbeckia* subg. *Lateriflorens* because of its lateral inflorescences.

Since Kaastra's description was based on so few specimens, some of his statements were found to be misleading. For example, the inflorescences are not axillary, but opposite to leaves; the fruits do not become fully free at maturity (Figure 5.1f), being the free carpels found by Kaastra most likely an outcome of the pressing process.

As observed in herbarium specimens, *E. cowanii* is usually misidentified as *E. almawillia*. However, the resemblance is only apparent from herbarium material,

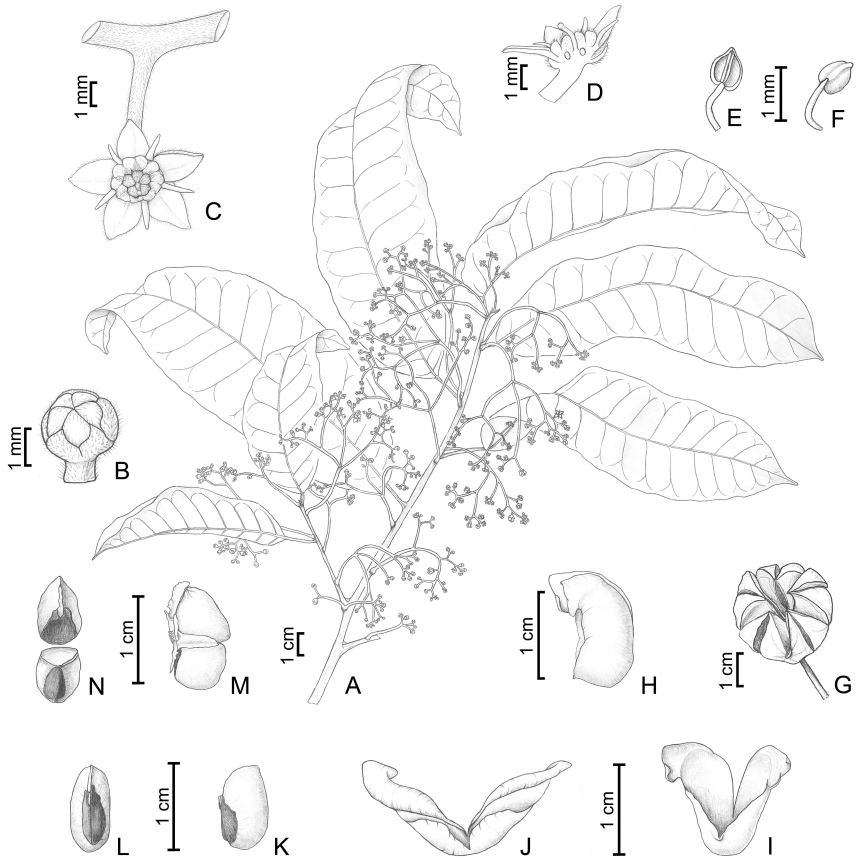


Figure 5.1 *E. cowanii*. A, flowering shoot; B, floral bud; C, flower at anthesis; D, flower in long section, note the ovary higher than the disc; E-F, stamen at anthesis, E, dorsal view, F, ventral view; G, dehisced capsule, only the dry exocarp and mesocarp remain; H-I, endocarp before elastic dehiscence and detached from the mericarp, H, lateral view, I, frontal view; J, endocarp elastically dehisced and detached from the mericarp, frontal view; K-N, seeds, K-L, when one seed per locule, K, lateral view, L, frontal view, M-N, when 2 seeds per locule, M, lateral view, N, frontal view. (A-G, P. Dias & R.G. Udulutsch 227 (SPF); H-N, M.F.F. da Silva *et al.* 1304 (INPA).

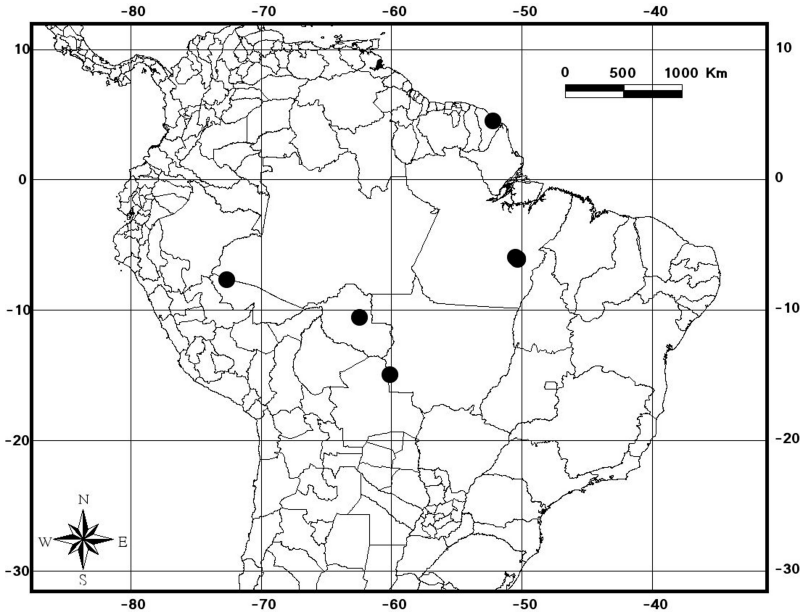


Figure 5.2 Map showing the known distribution of *E. cowanii*.

for *E. cowanii* has several distinguishing features not shared by *E. almawillia*, as summarized in Table 5.5.1.

E. cowanii, previously known only from French Guiana, has a wide distribution in South America (Figure 5.2), though it is poorly collected. In Brazil, it occurs in Acre, Pará, Rondônia (northern Brazil), and Mato Grosso (Central Brazil). In Acre and Rondônia, it was collected in terra firme forests with canopy of about 30 m tall; in Pará, it was found in terra firme forests and also in the forest-savanna edges of the Carajás Range (southern Pará); and in Mato Grosso it was collected at the base of the Parecis Range.

Table 5.1 Distinguishing features between *E. cowanii* and *E. almawillia*.

Feature	*E. cowanii*	*E. almawillia*
Habit	tree	shrub
Inflorescence type	panicule, Figure 5.1(a)	dichasium
Bracts and bractlets	caducous	persistent
Flowers	protandric, Figure 5.1(c)	non-protandric
Disc	lower than the ovary, Figure 5.1(d)	higher than the ovary
Capsule at maturity	rounded, Figure 5.1(g)	stellate
Dorsal apophysis	caducous, small, obtuse to rounded, Figure 5.1(g)	persistent, curved, thorn-like

5.5.2 Additional specimens examined

BRAZIL. Acre: Cruzeiro do Sul, BR-364, Km 42, ramal 4 do Projeto Santa Luzia - INCRA, mata de terra firme, solo argiloso, 11 Sep 1985 (fr.), A. Rosas Jr. *et al.* 259 (MG). Mato Grosso: Vila Bela da Santíssima Trindade, Caminho da cachoeira de Vila Bela, Km 23 de Vila Bela, plantas sobre pedras, mata de terra firme, solo argiloso, 5 May 1983 (fr.), L. Carreira *et al.* 729 (INPA, MG); estrada para Cachoeirinha (cascata), 17 Km de Vila Bela da Santíssima Trindade, Parque Estadual "Serra de Ricardo Franco", ca. de 500 m, na margem direita do riacho, solo areno-argiloso, com afloramentos rochosos, mata de terra firme, 16 Nov 2005 (fl., fr.), P. Dias & R.G. Udulutsch 226. Pará: Marabá, Carajás, Serra Norte, área de exploração de minÃľrio N-1, área de contato dos Campos Rupestres com mata, 2 June 1983 (fr.), M.F.F. da Silva *et al.* 1304 (INPA, MG); Marabá, Serra dos Carajás, N-1, transição da mata de terra firme para campo rupestre, 29 Mar 1984 (fr.), A.S.L. da Silva *et al.* 2005 (INPA, MG); Marabá, Serra dos Carajás, Mina de manganÃłs, mata de terra firme, solo argiloso, relevo ondulado, 16 Mar 1988 (fr.) J.G.S. Maia *et al.* 11 (MG); Parauapebas, Serra dos Carajás, Serra Norte, 6 Km on road southeast of Amza camp N-1, forest at edge of transition into cerrado vegetation, 19 May 1982 (fr.), C.R. Sperling *et al.* 5747 (INPA, MG, NY); Parauapebas, Serra dos Carajás, 1-4 Km along road from camp Azul toward AMZA camp N-1, forest on terra firme, 28 May 1982 (fr.), C.R. Sperling *et al.* 5852 (INPA, MG, NY); Parauapebas, Serra dos Carajás, 20-25 Km NW of Serra Norte mining camp, semi-deciduous forest and scrub, 6 Dec 1981 (fl., fr.), D.C. Daly *et al.* 1765 (INPA, MG, NY); Parauapebas, Serra dos Carajás, Mina de ferro do N1, próximo à mata, 8 June 1990 (fr.), N.A.

Rosa *et al.* 5140 (MG); Parauapebas, Serra dos Carajás, margem da estrada entre N-1 e Serraria, Km 12. Mata de terra firme, 29 Aug 1972 (fr.), N.T. Silva & B.S. Ribeiro 3644 (IAN); Parauapebas, Serra dos Carajás, margem da estrada do N-13. Capoeira, 24 Aug 1972 (fr.), N.T. Silva et B.S. Ribeiro 3589 (IAN). Rondônia: Jaru, BR-364, 28,4 Km de Ouro Preto d'Oeste, então 4,5Km na Linha 632, Fazenda do Sr. Zuza, fragmento de mata na margem direita da Linha, solo argiloso com afloramentos rochosos, floresta ombrófila densa, 06 Jan 2007 (fr.), P. Dias & R.G. udulutsch 284-286 (SPF).

5.6 Acknowledgments

The authors are grateful to the Curators of F, IAN, INPA, MG, NY, and US for loans of herbarium specimens, and to IBAMA for providing the collecting license N. 080/2005-COMON. PD was supported by FAPESP (Procs. 02/09762-6 & 04/15141-0), RGU by CNPq (Proc. 140945/2004-0) and JRP by CNPq (Proc. 304726/2003-6) & FAPESP (Proc. 04/15141-0).

5.7 Literature cited

[1] Hickey, L. J. 1979. A revised classification of the architecture of dicotyledonous leaves. *In* Metcalfe, C. & Chalk, L. (eds.) *Anatomy of the dicotyledons.* vol. 1, 2 ed. Clarendon Press, Oxford, 25–39.

[2] Kaastra, R. C. 1977. New taxa and combinations in Rutaceae. *Acta Bot. Neerl.* 26: 471–488.

[3] Kaastra, R. C. 1982. Pilocarpinae (Rutaceae). *Fl. Neotrop. Monogr.* 33: 1–198.

[4] McNeill, J., Barrie, F. R., Burdetand, H. M., Demoulin, V., Hawksworth, D. L., Marhold, K., Nicolson, D. H., Prado, J., Silva, P. C., Skog, J. E., Wiersema, J. H. & Turland, N. J. (eds.) 2006. *International Code of Botanical Nomenclature (Vienna Code) adopted by the Seventeenth International Botanical Congress Vienna, Austria, July 2005.* vol. 146 [Regnum Veg.], Koeltz Scientific Books, Königstein.

[5] Pirani, J. R. 1999. Two new species of *Esenbeckia* (Rutaceae, Pilocarpinae) from Brazil and Bolivia. *Bot. J. Linn. Soc.* 129: 305–313.

[6] Weberling, F. 1989. *Morphology of flowers and inflorescences.* Cambridge University Press, Cambridge.

Capítulo 6

A New Species of *Esenbeckia* Kunth (Rutaceae)

6.1 Abstract

A new species of *Esenbeckia* (Esenbeckiinae, Rutaceae) is described and illustrated. *Esenbeckia bracteata* P. Dias & Pirani sp. nov. is known from some collections from the Eastern Amazon (Acre and Rondônia States). It is a species with unifoliolate leaves, distinct from the other species of the genus with lateral inflorescences for its persistent bracts and a very distinctive area of the dorsal apophysis on its fruit.

6.2 Resumo

Um nova espécie de *Esenbeckia* (Esenbeckiinae, Rutaceae) é descrita e ilustrada. *Esenbeckia bracteata* P. Dias & Pirani sp. nov. está sendo descrita com base em coleções oriundas da Amazônia oriental (Estados do Acre e Rondônia). Trata-se de uma espécie com folhas unifolioladas que se distingue das outras espécies do gênero com inflorescências laterais por suas brácteas persistentes e uma área bem distinta ao redor da apófise dorsal do fruto.

6.3 Introduction

The genus *Esenbeckia* (subtribe Esenbeckiinae, tribe Galipeeae, subfamily Rutoideae, Rutaceae) comprises 28 species, and occurs in the Neotropical region. A monograph to the genus was presented by Kaastra [2] for the Flora Neotropica Monograph series, and recent work by us (Pirani [3], Dias *et al.* unpublished) has contributed with the taxonomy of the genus. During our herbarium work, we discover a flowering and some fruiting collections of a undescribed species of this genus. Additionally, in a recent field work in the Amazon basin, we recollected additional flowering and fruiting samples of that species. For the morphological descriptions, we used our own collections (deposited at SPF, MG, IAN, and INPA) and additional herbarium specimens from IAN, INPA, MG, and NY. Except for aestivation patterns, we used only fully developed structures. Flower and fruits were fixed with FAA in the field (for our own collections) or rehydrated (for fruiting herbarium specimens) before measurements and sketches were made. Unless otherwise stated, terms used follow Weberling [4] for flowers and inflorescences, and Hickey [1] for leaves.

6.4 Results and discussion

6.4.1 Description of the new taxon

Esenbeckia bracteata P. Dias & Pirani sp. nov.

Type. Brazil, Acre, Xapuri, Distrito de Porto Rico, Rodovia BR 317, sentido Rio Branco-Brasiléia, 20,05km após o trevo Xapuri-Brasiléia, estrada à esquerda que leva ao Distrito de Porto Rico (estrada antes da entrada para a vila Epitaciolândia), então 11.36km, trilha na margem esquerda, então 220m na trilha, 263m, 23.xi.2005, fl., fr., P. Dias 233 (holotype here designated, SPF; isotypes, CEPEC, HUEFS, IAN, INPA, MBM, MG,NY).

Treelet 1-3.2 m tall with yellowish to soft-greenish-brown pubescence of appressed hairs to 0.8 mm long; branchlets 2-4 mm in diam., pubescent when young, then glabrescent. with rare, eliptic, pale lenticels. Leaves alternate, unifoliolate, with sessile leaflet; petiole semiterete, canalicule with a concetration of appressed trichomes, 0.35–1.2 mm long, the base and the apex slightly tumid, pubescent with appressed trichomes; leaf blade elliptic, 3.2–16.1 x 1.2–5.7 cm, acute to obtuse or rounded at base, acuminate (acumen obtuse to retuse) at apex, the margin entire and waved, membranaceous, adaxial surface lustrous when fresh, glabrescent on the whole lamina (sparsed trichomes) and pubescent to puberulous on the midvein, abaxial surface paler and dull, glabrous, puberulous on the midvein; venation brochidodromous, primary and secondary veins prominent above and beneath. Inflorescences lateral, not axillary (sometimes opposite to the leaf itself) along the branches, erect, dichasial, with 7 flowers, shorter than leaves, 4.2–6.65 x 3.45–6.15

mm, densely pubescent with appressed trichomes (to 0.8 mm long); partial florescences, 1.2–2.2 mm long, erect; bracts, persistent, 4.1-5.95 x 0.45-0.65 mm, 1 per inflorescence, persistent, pubescent; pedicels 0.9–1.9 mm long, pubescent; bractlets (prophylls) persistent, 1 per flower, 1.55-2.4x0.6-0.75, triangular, pubescent. Flowers to 3.5-4.1 mm in diam.; calyx lobes 5(-6), persistent at mature fruits, ascending cochleate, ovate, 1.25-1.75 mm long, apex obtuse to slightly acute, coriaceous, pubescent, hyphodromous (vein visible only on the adaxial surface); petals 5(-6), persistent at mature fruits, descending cochleate, erect, lanceolate, 2.45–2.7 x 1.1–1.4 mm, acute at apex, coriaceous, pale yellow (cream), pubescent on abaxial surface, glabrous on adaxial surface, hyphodromous (vein visible only on the adaxial surface); filaments 5(-6), persistent at mature fruits, subulate, 1.9–2.2 mm long, glabrous, reflexed after anthesis; anthers heart-shaped, versatile, c. 0.8 mm long including the small mucronate tip, glabrous, early deciduous; disc annular, 5-lobed, each lobe slightly 2-lobed, 1.5-1.7 mm in diam., fleshy, with small rounded projections above c. 0.1 mm, higher than the ovary, glabrous, pale-yellow; carpels 5(-6), c. 0.8 mm high, adnate to the disc through the lower half and connate proximally, free and ellipsoid distally, furnished with tiny projections c. 0.1 mm, pubescent; ovules 2 per locule, collateral; style inserted proximally on the carpels but nearly completely free, c. 1.5 mm long, glabrous; stigma capitate, 5-lobed. Fruit 1-2 per infructescence, a slightly depressed, stellately 5-lobed capsule, 11.2–13.55 x 18.25–19 mm, glabrescent when young, then glabrous, muricate with numerous hooked prickles 0.9-2.9 mm long, dehiscent septicidally along the dorsal commissures from apex down to 2–3 mm above the base or even reaching the very base, and loculicidally from the base along

the ventral sutures up to the dorsal apophysis (1/2 from the base) which is thorn-like within a broad area not occupied by the prickles, nerved on the outside of the exocarp with nerves leaving from the dorsal apophysis; clearly transversally nerved on the inside of the mesocarp; endocarp smooth, ochre-yellow, hard, elastically (or explosively) dehiscent; Seed 1 per locule, ovoid, 8.6–11.14.4–5.4 mm, beaked at apex, testa brown to dark brown, hilum narrow, elongated, running from apex down to the chalazal area, irregular in shape, soft brown to brown toward the chalazal area; embryo not seen.

The small inflorescences resemble those of *E. almawillia* Kaastra, but this latter species has inflorescences with caducous bracts and bractlets. In addition, *E. bracteata* has many other noteworthy features, as the sepals, petals and often filaments persistent at the mature fruit (Figure 6.1K, vs. caducous in *E. almawillia* Kaastra); the fruit muricate (Figure 6.1J-L, vs. smooth, except for the dorsal apophysis), the erect dorsal apophysis encompassed by a distinguished area (Figure 6.1L, vs. inflexed dorsal apophysis only).

Esenbeckia bracteata is known from 7 collections from Acre (6) and Rondônia (1) States (Northeastern Brazil, Figure 6.2), where it occurs in 'terra firme' forests with canopy about 35 m tall, on clayish soils.

The inflorescence with a prominent, persistent bract has not been mentioned for any species in the genus and the epithet makes reference to this feature (Figure 6.1A, D, and K), as an important diagnostic feature which facilitates its recognition in the field, even under vegetative condition.

Illustration. Figure 6.1.

Paratypes. Brazil, Acre, Xapuri, Distrito de Porto Rico, Rodovia BR 317, sentido Rio Branco-Brasiléia, 20,05km após o trevo Xapuri-Brasiléia, estrada à esquerda que leva ao Distrito de Porto Rico (estrada antes da entrada para a vila Epitaciolândia), então 11.36 km, trilha na margem esquerda, então 220m na trilha, 263m, 23.xi.2005, fl., fr., P. Dias 234 (SPF); Brasiléia, Seringal Porangaba, Colocação São José, primary moist forest on gently undulating terrain, frequent in understory of primary forest, 23.v.1991, fr., D.C. Daly 6884 (INPA, NY); Brasiléia, Seringal Porvir, Colocação Tucumã (Zé Maria), n. 30, floresta densa, 05.ix.2000, fr., E.F. Orfanó 21 (IAN); Brasiléia, estrada para Assis Brasil, Km 20, a 3Km da margem da estrada, mata alterada de terra firme, solo argiloso, 03.xi.1980, fl., fr., C.A. Cid 3124 (INPA); Brasiléia, estrada Velha para Rio Branco, Km 31, ramal para a fronteira Brasil-Colômbia, mata de terra firme, solo argiloso, 02.vi.1991, fr., C.A. Cid 10229 (INPA). Rondônia, Nova Mamoré, Ramal 34, margem da linha D, 30.viii.1996, fr., L.C.B. Lobato 2297 (MG).

6.5 Acknowledgements

The authors are grateful to the Curators of IAN, INPA, MG, and NY for loans of herbarium specimens, and to IBAMA for providing the collecting license N. 080/2005-COMON. PD was supported by FAPESP (02/09762-6 & 04/15141-0), RGU by CNPq (Proc. 140945/2004-0), and JRP by CNPq (304726/2003-6) & FAPESP (04/15141-0).

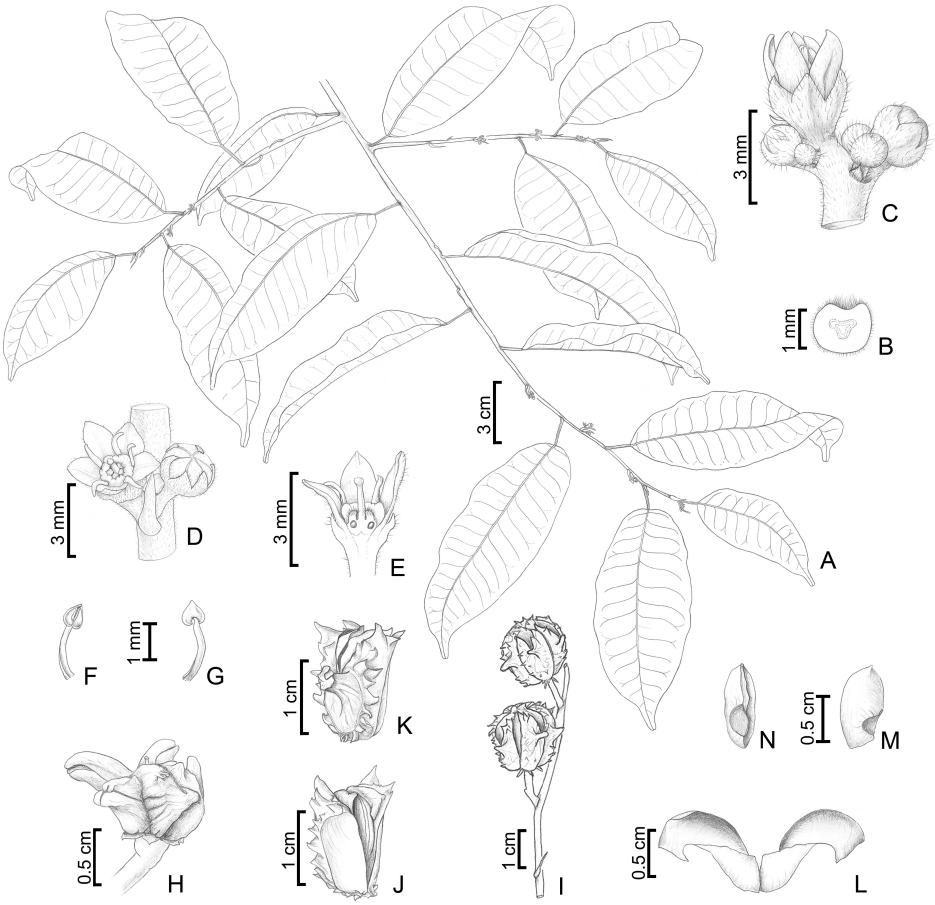


Figure 6.1 *Esenbeckia bracteata* P. Dias & Pirani. A, Flowering shoot. B, petiole, cross section. C, dichasium. D, flower at anthesis and 6-merous bud in the same inflorescence, note the persistent bract. E, flower in long section, note the disc higher than the ovary. F-G, stamens, F, stamen with dehisced anther, frontal view, G, stamen with dehisced anther, dorsal view. H-I, fruits, H, young fruit, note the persistent perianth, I, muricate, dehisced capsules, still with seeds. J-K, mature carpel detached from the capsule, J, carpel with endocarp and seed, latero-ventral view, K, carpel with endocarp and seed, latero-dorsal view, note the dorsal apophysis isolated within its own area. L, endocarp elastically dehisced and detached from the mericarp. M-N, seeds, M, lateral view, N, frontal view. A-H from P. Dias 233, I from C.A. Cid 10229, J-N from L.C.B. Lobato 2297.

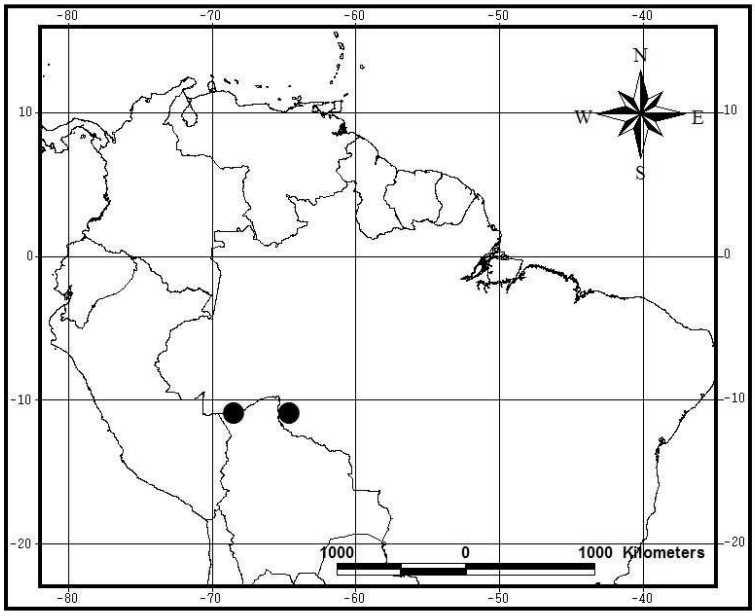


Figure 6.2 Map showing the knwon distribution of *Esenbeckia bracteata* in South America.

6.6 References

[1] Hickey, L. J. 1979. A revised classification of the architecture of dicotyledonous leaves. *In* Metcalfe, C. & Chalk, L. (eds.) *Anatomy of the dicotyledons.* vol. 1, 2 ed. Clarendon Press, Oxford, 25–39.

[2] Kaastra, R. C. 1982. Pilocarpinae (Rutaceae). *Fl. Neotrop. Monogr.* 33: 1–198.

[3] Pirani, J. R. 1999. Two new species of *Esenbeckia* (Rutaceae, Pilocarpinae) from Brazil and Bolivia. *Bot. J. Linn. Soc.* 129: 305–313.

[4] Weberling, F. 1989. *Morphology of flowers and inflorescences.* Cambridge University Press, Cambridge.

Considerações finais

Resumo[1]

Esta tese está dividida em três partes: I Filogenética Básica, II Filogenia de Pilocarpinae e III Novidades Taxonômicas em Esenbeckiinae.

Parte I Filogenética Básica - fornece uma rápida revisão de alguns dos métodos básicos atualmente em uso na filogenética, envolvendo aspectos teóricos e operacionais, assim como algumas de suas possíveis implicações. O **Capítulo 1** discute dois conceitos importantes (grupo e caráter) e interrelacionados, mostra como pode ser construído um modelo estocástico para o tratamento de caractres, enfatizando sua adequação e demonstranto como homologia e, por conseqüência grupos, podem ser adequadamente tratados sob ótica probabilística. O **Capítulo 2** apresenta os métodos de construção e otimização de árvores uzando máxima verossimilhnaça e análise bayesiana.

Parte II Filogenia de Pilocarpinae - apresenta a filogenia de Pilocarpinae baseada em dados moleculares (espaçadores ITS1, ITS2 e gene 5.8 S do DNA nuclear e espaçador *trn*G-S do DNA platidial) e a filogenia de *Pilocarpus* Vahl baseada em dados moleculares (mesmas regiões usadas anteriormente) e morfológicos. O **Capítulo 3** apresenta a filogenia em nível genérico da subtribo Pilocarpinae e de gêneros relacionados (*Helietta* e *Balfourodendron*), mostrando que, exceto *Esenbeckia*, os gêneros tradicionalmente reconhecidos (*Metrodorea*, *Pilocarpus* e *Raulinoa* - monoespecífico) emergem como monofiléticos (embora a subtribo não) e que *Balfourodendron* e *Helietta* (ambos da subtribo Pteleinae) possuem relações mais estreitas com parte dos gêneros de Pilocarpinae do que com o gênero-tipo de sua própria subtribo (Pteleinae) e reúnem-se (junto com *Esenbeckia*, *Metrodorea* e *Raulinoa*) em um clado caracterizado pela presença de inflorescências ramificadas, para o qual foi criada uma subtribo, ficando Pilocarpinae monogenérica; além disso, este capítulo apresenta um protocolo para detecção de *burn-in* em análises filogenéticas bayesia-

[1]Está sendo apresentado apenas por "norma", pois é uma cópia do prefácio.

nas usando métodos já bem estabelecidos em estudos de convergência de MCMC. Por sua vez, o **Capítulo 4** apresenta a filogenia das espécies de *Pilocarpus* baseada em dados morfológicos e moleculares; essa filogenia, associada a simulações computacionais, é utiliza como base para traçar hipóteses evolutivas sobre os padrões foliares e de estivação da corola no gênero, mostrando como os estados desses caracteres se comportam nas árvores obtidas e quão apropriado é utilizar os diferentes estados como sinapomorfias/homoplasias usando o método MCMC como base e contrastando com o mapeamento com parcimônia, deixando claro que sinapomorfia/homplasia é mais adequadamente tratada como uma questão de probabilidade.

Parte III Novidades Taxonômicas em Esenbeckiinae representa um reflexo das atividades de campo e de análise de material de herbário. No **Capítulo 5** é apresentada uma redescrição de *E. cowanii* Kaastra, espécie anteriormente conhecida apenas da Guiana Francesa e apenas pelo material tipo, cuja morfologia floral era desconhecida e foi encontrada nos Estados do Acre, Mato Grosso, Pará e Rondônia durante as expedições de campo que fiz para a Amazônia; além disso, é proposto um epítipo para o táxon. O **Capítulo 6** apresenta a descrição de uma nova espécie de *Esenbeckia* (embora ainda sem diagnose latina), coletada nos estados do Acre e Rondônia e caracterizada pela posse de brácteas persistentes.

Abstract

This dissertation is composed of three major parts: I Basic Phylogenetics, II Phylogeny of Pilocarpinae, and III Taxonomic Novelties in Esenbeckiinae.

Part I Basic Phylogenetics - provides a mini-review of some basic methods currently used in phylogenetics, covering theroretical and operational issues, and some of their implications as well. In the **Chapter** 1 I discuss two major concepts in phylogenetics, namely groups and characters, demonstrate how to build an evolutionary model and emphasize the importance of models in phylogenetics. As an outcome, the meanings of character evolution and groups are reviewed and improved under a probabilistic view. Chapter 2 presents an introduction to the very basic methods of tree construction and optimization using maximum likelihood and bayesian methods.

Part II Phylogenetics of Pilocarpinae - presents a phylogeny of Pilocarpinae based on molecular data (internal trancribed spacers ITS1, ITS2, gene 5.8 S - from the nuclear DNA , and spacer *trn*G-S - from the plastidial DNA), and a phylogeny of *Pilocarpus* based on morphological and molecular (same DNA regions used before) evidence. **Chapter 3** presents a generic-level phylogeny of the Pilocarpinae and allied genera (*Balfourodendron* and *Helietta*), which supports the monophyly of the traditionally recognized genera (*Metrodorea* and *Pilocarpus*, *Raulinoa* - monospecific), except *Esenbeckia* (which is included in a polytomy), whereas the subtribe itself is not monophyletic; it is also shown that *Balfourodendron* and *Helietta* (both from subtribe Pteleinae) are more closely related to some genera of Pilocarpinae than to the type genus of their own subtribe (Pteleinae), and emerge (together with *Esenbeckia*, *Metrodorea*, and *Raulinoa*) nested within a clade that has multi-axis inflorescences, for which I created a new subtribe, leaving the Pilocarpinae monogeneric;

moreover, this Chapter presents a new[2] protocol to be used in MCMC diagnosis in phylogenetic studies. In the **Chapter 4**, in turn, the phylogeny of *Pilocarpus* is investigated based on morphological and molecular data; that phylogeny, combined to computer simulations, is then used to propose evolutionary hypotheses of leaf blade and corolla aestivation patterns, and show how appropriate the use of character states as synapomorphy/homoplasy can be using the MCMC method; additionally the MCMC and parsimony character mapping procedures are compared, and it is shown that synapomorphy/homoplasy is just a matter of probability.

Part III Taxonomic Novelties in Esenbeckiinae is clearly a direct result of my field expeditions to the Amazon, and herbarium work. In the **Chapter 5** I present a re-description and epitypification of *E. cowanii* Kaastra, previously known only from the type locality (and by the type specimens) and whose floral morphology was unknown, which I collected in Acre, Mato Grosso, Pará, and Rondônia States during my collecting trips in the Amazon; further, given the poor type material, I propose an epitype for the species. In the **Chapter 6** I describe a new species of *Esenbeckia* (still without latin diagnose), which I collected in Acre and Rondonia states, whose diagnostic feature is the presence of persistent bracts.

[2]Just the protocol, not the analytical tools themselves.

Publicações

Artigos e capítulos de livro publicados ou submetidos

Dias, P., Assis, L. & Udulutsch, R. G. 2005. Monophyly vs. paraphyly in plant systematics. *Taxon* 54: 1039–1040.

Dias, P., Pirani, J. R. & Udulutsch, R. G. submetido (abril de 2007). Epitypification and re-description of *Esenbeckia cowanii* Kaastra (Rutaceae). *Novon.*

Furlan, A., Udulutsch, R. G., & **Dias, P.** aceito para 2008. Flora da Serra do Cipó, Minas Gerais: Amaranthaceae. *Boletim de Botânica da Universidade de São Paulo.*

Furlan, A., Udulutsch, R. G., & **Dias, P.** submetido (maio de 2007). Flora da Serra do Cipó, Minas Gerais: Nyctaginaceae. *Boletim de Botânica da Universidade de São Paulo.*

Lima, L.R., **Dias, P.** & Sampaio, P. 2004. Flora da Serra do Cipó: Flacourtiaceae. *Boletim de Botânica da Universidade de São Paulo* 22: 19–23.

Udulutsch, R. G., **Dias, P.**, Pinheiro, M. H. O & Furlan, A. 2007. Basellaceae. *In* Wanderley, M. G. L. *et al.* (eds.) *Flora fanerogâmica do Estado*

de São Paulo, vol. 5, São Paulo, Rima, 17–20.

Udulutsch, R. G., **Dias, P.**, Pinheiro, M. H. O, Tannus, J. & Furlan, A. 2007. Phytolaccaceae. *In* Wanderley, M. G. L. *et al.* (eds.) *Flora fanerogâmica do Estado de São Paulo*, vol. 5, São Paulo, Rima, 237–246.

Udulutsch, R. G., Souza, V. C., Rodrigues, R. R. & **Dias, P.** submetido (maio de 2007). Composição florística de lianas e suas formas de escalada na Estação Ecológica dos Caetetus, São Paulo, Brasil. *Revista Brasileira de Botânica.*

Tradução

Swofford, D. L. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods), version 4, Documentação Beta (Traduzido por **Pedro Dias**, 2004). Sinauer Associates, Sunderland.

Citações na *Web of Science*

Acesso em 28 de novembro de 2007: 5 (*Taxon* – 4, *Systematic Botany* – 1)

Citações totais

Checagem manual (*Taxon* e *Systematic Botany*, até novembro de 2007) – 7

Posfácio

Semelhante ao Prefácio, este Posfácio será utilizado mais ou menos informalmente (para uma versão formal sugiro o Resumo). Aproveitarei este espaço para um pouco de estória[3] e perspectivas futuras, dado que as conclusões já foram resumidas no Prefácio e estão em cada um dos capítulos.

Como a estória começou. Há 11 anos atrás, eu estava na biblioteca do Museu Goeldi e me deparei com uns livros incomuns, uma coleção de vários volumes que o Nelson Papavero havia publicado junto com o Llorente-Bousquets pela Universidade Nacional Autônoma do México, os *Principia taxonomica* ([4]). Um dos volumes mostrava a forte relação da sistemática com a matemática, mostrando que os sistemas taxonômicos tinham sua origem na teoria dos conjuntos[4]. Começei a ler os livros do Papavero e logo depois descobri a primeira edição do livro do Dalton Amorim ([1]). Mas não havia com quem discutir, pois os botânicos que eu conhecia não estavam muito atentos para essas coisas. Então fui conversar com uns zoólogos (principalmente o Horácio Higuchi) e me inscrevi como aluno-ouvinte em uma disciplina da Pós-graduação em zoologia. Começei a ler a *Cladistics* e a *Systematic Zoology* (recém mudada para *Systematic Biology*) e depois já não saía da biblioteca (tinha que ser "comunicado" pela bibliotecária que já era 17h30' e que o expediente estava encerrado). Vi um artigo do Felsenstein ([2]) discordando do que eu havia aprendido (100% parcimômia) e achei estranho, meio complicado, e ficou por isso.

Pulando uns anos, em 2002, fiz a disciplina do "Tim" (Antonio Carlos Marques) e fiquei realmente convencido sobre a parcimônia e achei interessante a idéia do 3TS. Pensei em usar 3TS na tese, mas o programa que havia, TAS (Nelson & Ladiges [3]), era/é tão ruim que talvez fosse mais fácil escrever um programa do que tentar usá-lo. Coincidindo a isso tem o fato de que vários problemas da sistemática podem ser resolvidos (ou pelo menos extremamente facilitados, veja os capítulos da Parte II desta tese) pela computação, então me matriculei numa disciplina da bioinformática. Resolvi[5] o problema do 3TS, mas já não estava convencido pelo método.

[3]Isso talvez devesse ter sido colocado no Prefácio, mas ...

[4]Talvez isso tenha influenciado minha visão "algoritmica" da sistemática.

[5]Só não sei onde está a versão final, só consigo encontrar uma versão inicial com alguns bugs.

Depois, 2003, fiz a disciplina do Matioli (Sérgio Russo Matioli) e, por coincidência, em uma das discussões tive que defender a máxima verossimilhança contra os outros métodos (foi sorteio). A partir daí tive que ler os artigos do Felsenstein, o proponente do método em filogenética. Coincidentemente, também, uns meses antes havia saído na *Systematic Biology* (vol. 51) um número quase completo com artigos de análise bayesiana. Li os artigos e descobri que precisava aprender algo de estatística. Me matriculei na disciplina da Roseli (Roseli Aparecida Leandro), na estatística, e descobri o quanto ainda preciso estudar de matemática e estatística para ter um conhecimento razoável sobre o assunto em filogenética. Mas, aprendi algumas coisas interessantes e o resultado disso foi utilizado nesta tese e tem possibilitado as seguintes perpectivas em curto prazo:

1. Inferir áreas de endemismo usando métodos bayesianos - os métodos propostos por Szumik *et al.* ([5]) e Szumik & Goloboff ([6]) podem ser mais adequadamente tratados do ponto de vista probabilístico;

2. Desenvolver um protocolo de análise (parcialmente feito no Capítulo 3) de convergência de MCMC em estudos filogenéticos utilizando os métodos já disponíveis na literatura. Apesar da existência de vários métodos em estudos específicos de MCMC, esses métodos não são utilizados em filogenética e não existe um protocolo para a área (o Capítulo 3 pode ser claramente desmembrado em dois artigos).

Em relação às Rutaceae estudadas, as questões mais intrigantes são (para médio prazo):

1. A distribuição de *Pilocarpus racemosus* Vahl no norte da América do Sul, América Central e México. Um estudo filogeográfico precisa ser feito;

2. A expressão dos genes responsáveis pelo padrões foliares.

[1] Amorim, D. S. 1994. *Elementos básicos de sistemática filogenética.* Holos Editora e Sociedade Brasileira de Entomologia, Ribeirão Preto.

[2] Felsenstein, J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst. Zool.* 22: 240–249.

[3] Nelson, G. J. & Ladiges, P. Y. 1991. *TAX: three taxon statements.* Published by the author.

[4] Papavero, N. & Llorente-Bousquets, J. (Orgs.). 1993-1996. *Principia taxonomica.* vol. 1-8, Universidade Nacional Autónoma de Mexico, Mexico.

[5] Szumik, C. A., Cuezzo, F., Goloboff, P. A. & Chalup, A. E. 2002. An optimality criterion to determine areas of endemism. *Syst. Biol.* 51: 806–816.

[6] Szumik, C. A. & Goloboff, P. A. 2004. Areas of endemism: an improved optimality criterion. *Syst. Biol.* 53: 968–977.

. . .