Relatório de TCC1

Henrique Luiz Rieger[1]

[1]Departamento de Informática – Universidade Federal do Paraná (UFPR)
Curitiba – PR – Brasil

Resumo. *O objetivo deste trabalho é propor a aplicação de algoritmos genéticos como mecanismos de busca de árvores filogenéticas, baseando-se nos trabalhos de [Moilanen 1999], [Cotta and Moscato 2002] e [Zwickl 2006], para aplicações em matrizes de dados morfológicos. Uma breve introdução sobre o problema é apresentada e são citadas algumas pesquisas e* softwares *já disponibilizados. Também são realizados alguns experimentos práticos com dados reais usando o programa TNT.*

1. Introdução

Árvores filogenéticas são fundamentais para compreender a evolução da vida no planeta Terra e apresentam usos em diversas áreas do conhecimento, como na dinâmica de populações e no estudo de proliferação de doenças [Villalobos-Cid et al. 2023]. Em especial, pesquisas em taxonomia dependem de filogenias para traçar a história evolutiva e poder explicar determinados fenômenos observados nos grupos de foco [Azouri et al. 2021].

Fósseis são materiais usados para melhorar estudos filogenéticos, abrindo oportunidades para perceber o passado, permitindo que não sejam feitos apenas a partir de táxons viventes [Mongiardino Koch et al. 2021]. Trabalhar com táxons fósseis, porém, é desafiador, uma vez que é um recurso incompleto e dependente de caracteres morfológicos, de difícil coleta. Além disso, em comparação a sequências moleculares, os resultados de análises filogenéticas usando apenas morfologia tendem a ser menos precisos [Berger and Stamatakis 2010]. No entanto, a grande maioria dos trabalhos em Paleontologia depende exclusivamente de fósseis para as análises.

Com essas questões em vista, *softwares* focados em filogenia foram desenvolvidos. Mesmo assim, toda filogenia é uma busca por recuperar um passado que não pode ser observado diretamente, bem como é um problema muitas vezes intratável de forma exata, dependendo de métodos heurísticos.

Este relatório apresenta uma proposta de aplicação de algoritmos genéticos, um tipo especial de heurística de busca, na análise filogenética com caracteres morfológicos. Inicialmente, são contextualizados os temas abordados neste trabalho. Depois, é realizada uma revisão de pesquisas similares realizadas nessas áreas, bem como uma amostra dos programas existentes para análises filogenéticas. Experimentos utilizando um desses *softwares* (TNT) foram realizados para fins de demonstração dos métodos atuais. Por fim, a proposta para um experimento de investigação do uso de algoritmos genéticos em filogenia é apresentada.

1.1. Análises filogenéticas

Uma filogenia é uma representação da história evolutiva de um grupo de seres vivos. Essa representação é feita por meio de uma árvore, em que cada nó corresponde a um

táxon, uma unidade básica que pode representar uma família, um gênero, uma espécie ou mesmo um indivíduo. Um grupo de táxons que inclui um ancestral em comum e todos os seus descendentes é denominado um **clado**. Um exemplo de árvore filogenética pode ser observado na Figura 1.

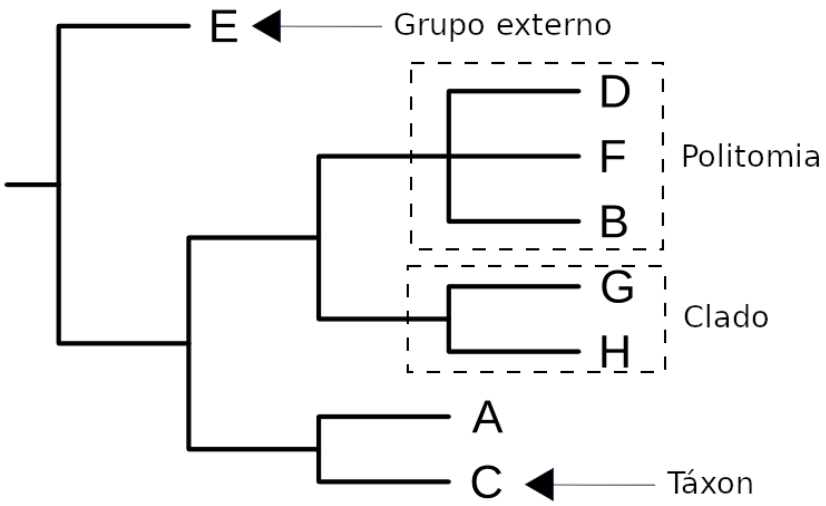


Figura 1. Exemplo de árvore filogenética

Em uma análise filogenética, os táxons são representados por sequências de caracteres, *strings* dentro de um alfabeto geralmente discreto em que cada símbolo representa uma característica do táxon que pode ser passada hereditariamente. Essas sequências podem ser formadas por trechos de DNA, aminoácidos de proteínas ou mesmo características fenotípicas, como aspectos morfológicos. Para que a análise funcione, as sequências de todos os táxons devem ter o mesmo tamanho, e cada posição deve representar a mesma característica. O conjunto de todas as sequências analisadas forma a matriz de caracteres [Felsenstein 1985].

O modelo assume que os táxons se diversificam ao longo do tempo em dois (ou mais) táxons descendentes pelo processo de evolução. Na árvore, essa relação fica representada pelos descendentes apresentados como nós filhos do táxon ancestral. Assume-se que o mais provável é que um táxon evolua apenas para dois descendentes, embora possam ser aceitos mais numa relação conhecida como **politomia**.

Uma filogenia pode ser enraizada ou não-enraizada. Em uma árvore filogenética enraizada, o tempo "flui"da raiz para as folhas, ou seja, quanto mais próximos das folhas, mais recentes são os táxons. Em uma árvore não-enraizada, não é possível determinar quais nós internos representam os táxons mais antigos. Uma árvore não-enraizada pode ter uma raiz atribuída com a inclusão de um **grupo externo** à análise, isto é, um conjunto de táxons que sabe-se ter divergido antes do grupo analisado, chamado de **grupo interno** [Felsenstein 2004].

O objetivo de uma análise filogenética é encontrar a melhor filogenia para um determinado grupo de táxons, isto é, dada uma matriz $n \times m$ com n táxons e m caracteres,

a análise retorna uma árvore com exatamente n folhas que melhor representa a história evolutiva do grupo a partir de um determinado critério.

1.2. Etapas de uma análise filogenética

Os métodos para análises filogenéticas podem ser divididos em dois grupos principais: os baseados em distância e os baseados em sequências [Cotta and Moscato 2002]. Métodos por distância agreagam iterativamente os clados mais próximos entre si baseados em alguma métrica absoluta, criando uma matriz de distâncias entre todos e resumindo cada táxon a um vetor de números. Fazem parte desse grupo os métodos UPGMA e Neighbor-Joining, por exemplo. Métodos baseados em sequências dependem da análise individual de cada caracter das sequências associadas aos táxons, com referência a algum critério. Análises por sequência podem ser divididas grosseiramente em quatro etapas: busca, avaliação, consenso e suporte. Essas etapas não necessariamente representam um processo linear e não são totalmente dissociadas, porém tal separação facilita a compreensão do método.

A etapa de busca consiste em determinar um método para percorrer o *espaço de busca* de árvores filogenéticas, isto é, o espaço que contém todas as possíveis filogenias válidas para um determinado conjunto de táxons de entrada. Assume-se para esta etapa que apenas árvores binárias completas (sem politomias) são válidas. [Cavalli-Sforza and Edwards 1967] calcularam o tamanho do espaço das árvores não-enraizadas U_n para um conjunto de n táxons como

$$U_n = (2n - 5)!!, \tag{1}$$

onde !! representa a operação fatorial duplo. Para as árvores enraizadas o tamanho do espaço corresponde a

$$R_n = U_{n+1} = (2n - 3)!!. \tag{2}$$

Como o espaço para árvores não-enraizadas é menor, a maioria dos métodos foca nesse tipo, enraizando as respostas com a inclusão de grupos externos, se necessário.

Determinar a filogenia ótima, procurando por todos os elementos do espaço de busca, consiste em um problema NP-Difícil [Chor and Tuller 2005], tornando uma busca ingênua impraticável para mais do que alguns poucos táxons. Mesmo sem percorrer todo o espaço, usando métodos como *branch and bound* [Hendy and Penny 1982], realizar uma busca exata (garantindo o resultado ótimo) ainda é muito custoso. Por essa razão, a maioria das análises utiliza métodos heurísticos, que consistem em trocar de lugar iterativamente ramos na melhor árvore encontrada até o momento, na esperança de achar um resultado melhor (algoritmos de *hill climbing*). Alguns operadores comuns para essa busca são o TBR (*tree bissection-reconnection*), o SPR (*subtree prune-regraft*) e o NNI (*nearest-neighbor interchange*) [Zwickl 2006].

A fase de avaliação ocorre intercaladamente com a fase de busca e permite analisar a qualidade de cada árvore gerada. Para isso, um critério de avaliação precisa ser definido. Os mais comuns são *máxima parcimônia* (MP) [Edwards and Cavalli-Sforza 1963], que determina que as árvores que necessitam do menor número de passos evolutivos são as mais prováveis, e *máxima verossimilhança* (MV) [Cavalli-Sforza and Edwards 1967, Felsenstein 1973], que se baseia em modelos estatísticos de substituição dos caracteres

para determinar a filogenia mais provável. Ambos os modelos apresentam benefícios e fraquezas para determinadas situações, mas todos são modelos simplificados de processos evolutivos e, portanto, não são capazes de refletir com fidelidade a realidade [Cavalli-Sforza and Edwards 1967].

A criação de uma árvore de consenso se faz necessária uma vez que a busca pode retornar mais de uma filogenia com a mesma avaliação, ou valores muito próximos dentro de um intervalo determinado. Surge então o desafio de quais clados serão mantidos para a resposta final. O método mais comum é o *consenso estrito*, que determina que apenas os clados presentes em todas as filogenias obtidas pela busca devem ser mantidos na resposta final [Felsenstein 2004]. Note que "colapsar" clados exige quebrar a propriedade binária das árvores obtidas nas fases anteriores, gerando politomias. Outros métodos para consenso incluem a regra da maioria [Margush and McMorris 1981] e o consenso de Adams [Adams III 1972].

Por fim, é preciso avaliar a qualidade da solução obtida. Para essa finalidade, o método mais comum é o *bootstrapping*, adaptado do campo da estatística por [Felsenstein 1985]. Esse se baseia em replicar a análise diversas vezes, reamostrando os caracteres, utilizando sorteios com repetição, e medindo a porcentagem de análises nas quais os clados são obtidos. Clados com boas porcentagens representam chances altas de serem representativos no mundo real. Como alternativa, *jackkniffing* realiza réplicas da análise descartando um caracter por vez.

No caso de buscas utilizando o critério de máxima parcimônia, o suporte de Bremer [Bremer 1988] pode ser utilizado como complemento ao *bootstrap*. Esse analisa quantas **homoplasias** são necessárias para que um clado seja eliminado do consenso, isto é, quantos passos evolutivos extras precisam ser considerados para que a filogenia seja desmanchada. Valores maiores representam maior robustez da resposta.

1.3. Algoritmos genéticos

Algortimos genéticos (AGs) pertencem a uma classe de algoritmos denominados *evolutivos*, que se baseiam em princípios de evolução darwiniana para obter respostas para problemas computacionais. AGs funcionam codificando possíveis soluções para um problema de otimização em *cromossomos* ou indivíduos, cujo conjunto forma uma população. Esses cromossomos são inicializados aleatoriamente e então recombinados entre si a cada iteração do algoritmo, gerando novas respostas. Cada novo cromossomo também está sujeito a mutações, pequenas alterações estocásticas que podem ocorrer em cada pedaço da resposta. Dois cromossomos são escolhidos para recombinação baseados em um critério de *aptidão*, que determina a qualidade da solução. Cromossomos mais aptos devem escolhidos mais vezes, mas dependem da escolha eventual de respostas não-ótimas para manter uma certa "variabilidade genética" na população [Zwickl 2006]. Os operadores de seleção, recombinação e mutação precisam ser definidos para cada problema.

Em comparação a métodos de *hill climbing*, algoritmos genéticos são menos propensos a ficar presos em ótimos locais [Zwickl 2006]. Isso é proporcionado tanto pela inclusão de cromossomos sub-ótimos na recombinação, quanto pela inclusão de pequenas variações individuais por meio das mutações. AGs também são particularmente úteis em problemas multiobjetivo [Zambrano-Vega et al. 2016]. No entanto, por depen-

derem de fatores aleatórios, não são facilmente reprodutíveis e podem necessitar de várias execuções independentes para obter respostas consistentes. Também dependem de vários hiperparâmetros (tamanho das populações, taxas de mutação e recombinação, número de gerações etc.), e não garantem a resposta ótima, podendo igualmente cair em ótimos locais, bem como não atingir convergência.

No contexto de filogenia, algoritmos genéticos podem ser usados para a etapa de busca de métodos baseados em sequência, servindo como alternativa às buscas utilizando *hill climbing*. Como cromossomos, podem ser utilizadas as várias topologias possíveis de árvores, ou mesmo codificações em vetores lineares das mesmas [Cotta and Moscato 2002]. Já os clássicos SPR, TBR e NNI podem ser adaptados como operadores de mutação, enquanto o operador PDR (*prune-delete-regraft*) pode ser utilizado para recombinação.

2. Trabalhos relacionados

[Moilanen 1999] apresenta o *software* PARSIGAL, que realiza buscas utilizando algoritmos genéticos para encontrar topologias candidatas e heurísticas de *hill climbing* para alterar os candidatos, avaliando a melhor filogenia pelo critério de máxima parcimônia. O trabalho também apresenta diversas otimizações feitas na linguagem C para o cálculo da parcimônia de uma árvore. Também é apresentado o algoritmo PDR como operador de recombinação.

[Cotta and Moscato 2002] comparam diversos métodos para obter filogenias baseadas em distâncias usando algoritmos evolutivos. Os autores compararam o uso da estrutura da árvore e codificações em *strings* numéricas como cromossomos para o algoritmo, chegando à conclusão de que, para a maioria dos casos, a representação em árvore obtém maior acurácia.

Em [Zambrano-Vega et al. 2016], os autores utilizam o algoritmo de otimização multiobjetivo NSGA-II para implementar o *software* MO-Phylogenetics, capaz de realizar análises filogenéticas baseadas em máxima parcimônia e máxima verossimilhança simultaneamente, retornando árvores ótimas via fronteira de Pareto. Já [Villalobos-Cid et al. 2023] estendem a aplicação do método para redes filogenéticas, modelos evolutivos que aceitam a transmissão de informações entre ramos que já divergiram, ampliando as possibilidades de usos da inferência multiobjetivo através do *software* MO-PhyNet.

A partir da análise baseada em máxima verossimilhança, a tese de [Zwickl 2006] explica em detalhes um algoritmo genético utilizando apenas mutações (sem recombinações), compilado no *software* GARLI. Também são apresentados diversos *benchmarks* do método em comparação a *softwares* tradicionais, medindo, entre outros aspectos, a acurácia da busca e o tempo de execução. No mesmo trabalho, ainda é apresentada a variante p-GARLI, utilizando diversas populações com intercâmbio de indivíduos rodando em paralelo, a fim de permitir buscas simultâneas na mesma ordem de tempo da versão sequencial. [Brauer et al. 2002] também apresentam uma versão paralela do *software* GAML, porém focando na paralelização do cálculo de aptidão de cada população. [Noutahi and El-Mabrouk 2018] utilizam um algoritmo genético para prever e corrigir filogenias por MV em conjunto com o modelo "Duplication-Transfer-Loss".

Com outros algoritmos de busca, [Azouri et al. 2021] desenvolveram um método

para inferência filogenética otimizando a escolha de passos em uma busca heurística usando *random forest*. Já [Berger and Stamatakis 2010] demonstram um método de inclusão de táxons fósseis, codificados por morfologia, em uma árvore obtida previamente por máxima verossimilhança com sequências moleculares de táxons viventes.

3. *Softwares* atuais

Entre os *softwares* de análise filogenética, a maior parte dos artigos consultados usa PAUP* [Wilgenbusch and Swofford 2003], podendo realizar buscas pelos critérios de máxima parcimônia, verossimilhança ou por métodos baseados em distância. Já o TNT [Goloboff and Morales 2023] é focado em análises por parcimônia e contém uma implementação do algoritmo de avaliação de Máxima Parcimônia de Pesos Implícitos [Goloboff 1993], ou MPPI, que atribui pesos aos caracteres da análise proporcionalmente à quantidade de homoplasias que esses apresentam em uma determinada árvore.

Para análises sob o critério de máxima verossimilhança, além do PAUP*, o programa RAxML [Stamatakis 2014] tem ampla aplicação, devido ao uso de algoritmos de rápida execução, em conjuntos grandes de dados. O *software* PhyML [Guindon et al. 2010] também é bastante utilizado, originalmente aplicando o operador NNI para as buscas heurísticas. O mesmo também faz uso de cálculos de parcimônia para eliminar árvores pouco promissoras.

Além dos métodos mais tradicionais já elaborados, uma técnica que vem ganhando espaço é o uso de Inferência Bayesiana (IB)[Tschopp and Upchurch 2018], partindo de técnicas de MCMC, como Metropolis-Hastings. Alguns dos *softwares* que permitem o uso de caracteres morfológicos por essa técnica são o MrBayes [Ronquist et al. 2012] e o BEAST [Suchard et al. 2018]. Embora relativamente recentes, [Mongiardino Koch et al. 2021] demonstraram que o modelo "Fossil Birth-Death rate", cominado com IB, pode produzir bons resultados para matrizes com dados paleontológicos, enquanto [Goloboff et al. 2018] comprovaram que o mesmo apresenta acurácia bastante próxima de análises por parcimônia em *datasets* sintéticos. Já [Cau 2017] realizou com sucesso uma análise utilizando BEAST 2 [Bouckaert et al. 2014] para filogenia de peixes Dipnoi a nível de espécime.

4. Experimentos preliminares

Para fins de demonstração dos métodos atuais usados em análises filogenéticas, os dados de três trabalhos de filogenia foram testados. As matrizes de cada publicação apresentam características distintas, como tipo e quantidade de caracteres, bem como quantidade de táxons e de dados faltantes, de forma a demonstrar os desafios de se realizar inferência filogenética, revelando como os dados demandam processos diferentes.

A análise de [Vega-Dias et al. 2004] apresenta uma matriz de caracteres morfológicos cranianos e pós-cranianos de dicinodontes, um grupo extinto de répteis parentes dos mamíferos atuais. O objetivo era encontrar o posicionamento da espécie brasileira *Jachaleria candelariensis* na árvore deste grupo. A busca original foi feita por *branch and bound*, pelo critério de máxima parcimônia. A matriz de caracteres consiste em 14 táxons, sendo dois pertencentes ao grupo externo, com 44 caracteres. Como resultado original, foi obtido um consenso estrito a partir de seis árvores igualmente parcimoniosas, com 111 passos cada.

O artigo de [Bremer 1988] apresenta a ideia por trás do suporte de Bremer. Neste trabalho, o autor exemplifica a ideia de um estudo de caso com dois alinhamentos de sequências de RNA de plantas angiospermas: um com nove sequências de 64 nucleotídeos e outro com seis sequências de 82 nucleotídeos. Em ambos os casos, um táxon (Poacea) foi reservado como grupo externo. As buscas foram feitas por método exato e pelo critério MP. A análise da primeira matriz retornou duas árvores de 151 passos, enquanto a segunda retornou uma filogenia de 161 passos.

Por fim, o trabalho de [Smith et al. 2007] faz uma descrição osteológica completa dos fósseis do dinossauro *Cryolophosaurus ellioti*. A partir dos dados obtidos por esse processo, uma análise foi realizada para determinar sua posição filogenética dentro do grupo Theropoda. A matriz usada apresenta 347 caracteres, com 56 táxons amostrados (seis no grupo externo). Essa enorme quantidade de dados só pode ser processada por métodos heurísticos, de forma que os autores optaram por realizar a busca usando o algoritmo TBR no programa PAUP* com 25000 réplicas de sequência de adição aleatória. O consenso resultante da avaliação por MP consiste em 108 árvores de 833 passos, agrupadas por consenso estrito e de Adams, com o mesmo resultado. Suportes por *bootstrapping* e Bremer também foram calculados.

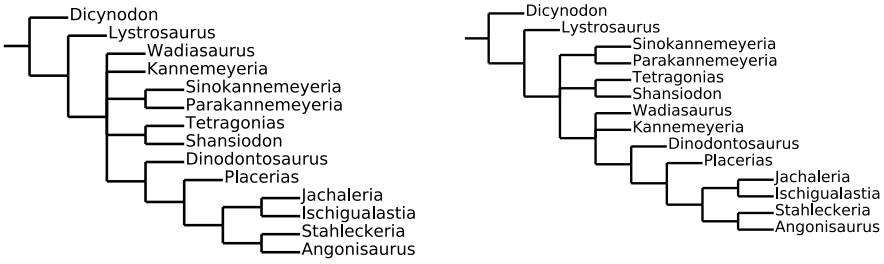
Para os testes, foi utilizado o *software* TNT na versão 1.6. As análises de terópodes foram feitas utilizando a opção "Traditional search", que corresponde à análise por *hill climbing*, com os resultados das matrizes menores (dicinodontes e angiospermas) sendo calculados com a opção "Implicit enumeration" (*branch and bound*, busca exata). Em todos os casos, os caracteres foram configurados como discretos e não-ordenados. Foram testados tanto o critério de máxima parcimônia padrão quanto parcimônia de pesos implícitos (função de concavidade padrão, $k = 3$). Todas as árvores foram agregadas utilizando consenso estrito. Por simplicidade, não foram calculados valores de suporte nem índices de convergência e retenção.

4.1. Resultados

A filogenia obtida pelos dados de dicinodontes usando máxima parcimônia padrão (MPP) está presente na Figura 2a. O resultado é exatamente o mesmo descrito no artigo original, à exceção de que a busca forçando a monofilia do grupo interno (obrigando-o a ser um clado exclusivo, sem nenhum táxon do grupo externo) retornou apenas quatro árvores, em vez das seis descritas. O algoritmo retorna seis árvores caso seja permitido que o gênero *Lystrosaurus* esteja presente dentro do grupo interno, formando uma politomia na base do mesmo.

Já a busca utilizando pesos implícitos, presente na Figura 2b, obteve um consenso um pouco mais resolvido de três árvores com pontuação 11,55, no qual os gêneros *Wadiasaurs* e *Kannemeyeria* deixaram de fazer parte da politomia na base do grupo interno e passam a enraizar um novo clado seguido pelo mesmo clado enraizado por *Dinodontosaurus* na análise anterior. Por não haver cálculo de suportes para esse exemplo, e por falta de uma explicação com fundamentação biológica, fica difícil precisar a coerência desse resultado.

As árvores obtidas pelos alinhamentos de sequência de angiospermas avaliadas por MPP, tanto de nove (Figura 3a) quanto seis táxons (Figura 3c), resultaram exatamente nas filogenias descritas originalmente. Como ambos os trabalhos foram realizados por

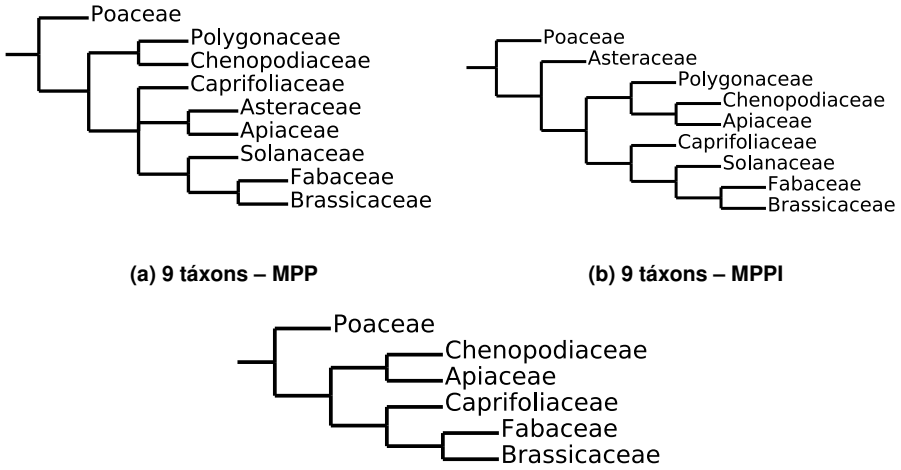


(a) Máxima parcimônia padrão

(b) Máxima parcimônia de pesos implícitos

Figura 2. Árvore de consenso do *dataset* de dicinodontes.

buscas exatas, é de se esperar que os valores de suporte também sejam os mesmos. Já a análise usando MPPI no alinhamento de nove táxons, apresentado na Figura 3b, mostra algumas diferenças significativas em relação ao resultado original, com a família Asteracea na base do grupo interno e Apiacea como grupo irmão de Chenopodiacea. No entanto, como o suporte de Bremer do resultado original já era baixo (≤ 3), também pode se esperar que o suporte nesse caso seja baixo. A aplicação de MPPI no *dataset* de seis táxons não obteve nenhuma diferença do original.



(a) 9 táxons – MPP

(b) 9 táxons – MPPI

(c) 6 táxons – MPP e MPPI

Figura 3. Árvore de consenso dos *datasets* de angiospermas.

Por fim, a análise da matriz geral de terópodes sob o critério de máxima parcimônia padrão (Figura 4) obteve um resultado bastante similar ao reportado originalmente, embora os números obtidos tenham sido significativamente diferentes. Em vez das 108 árvores originais de 833 passos, foram retornadas apenas oito árvores de 853 passos. O número reduzido de árvores, similar ao ocorrido no caso dos dicinodontes, é provavelmente um artefato da monofilia forçada no relatório oferecido pelo TNT, enquanto o valor das árvores mais parcimoniosas possivelmente se deve a limitações de memória do *software*. Em relação ao consenso obtido, a única diferença em relação ao original foi o posicionamento do gênero *Coelurus*, que passou a integrar a posição mais ancestral em (*Coelurus* + "Paraves"). O resultado da análise com o critério MPPI (Figura 5) retornou

três árvores com pontuação $\approx 80,76$, cuja maior diferença em relação à analise por MPP se dá nos posicionamentos dentro do clado Spinosauroidea.

Figura 4. Árvore de consenso do *dataset* de terópodes sob o critério de máxima parcimônia padrão

5. Proposta de trabalho

Para este trabalho, a proposta consiste na elaboração de um algoritmo genético para busca de árvores filogenéticas, utilizando como entrada matrizes de caracteres morfológicos sob o critério de máxima parcimônia. O algoritmo será testado utilizando tanto bases de dados sintéticas, cuja resposta correta pode ser determinada, mas que apresentam baixa correspondência com problemas reais; quanto com bases de dados reais retiradas de artigos da área de Paleontologia, cuja acurácia apenas pode ser obtida em relação aos resultads publicados e aplicação de outros métodos semelhantes.

O algoritmo será baseado principalmente em três dos trabalhos citados: no *software* GARLI de [Zwickl 2006], utilizando-se majoritariamente de operadores de

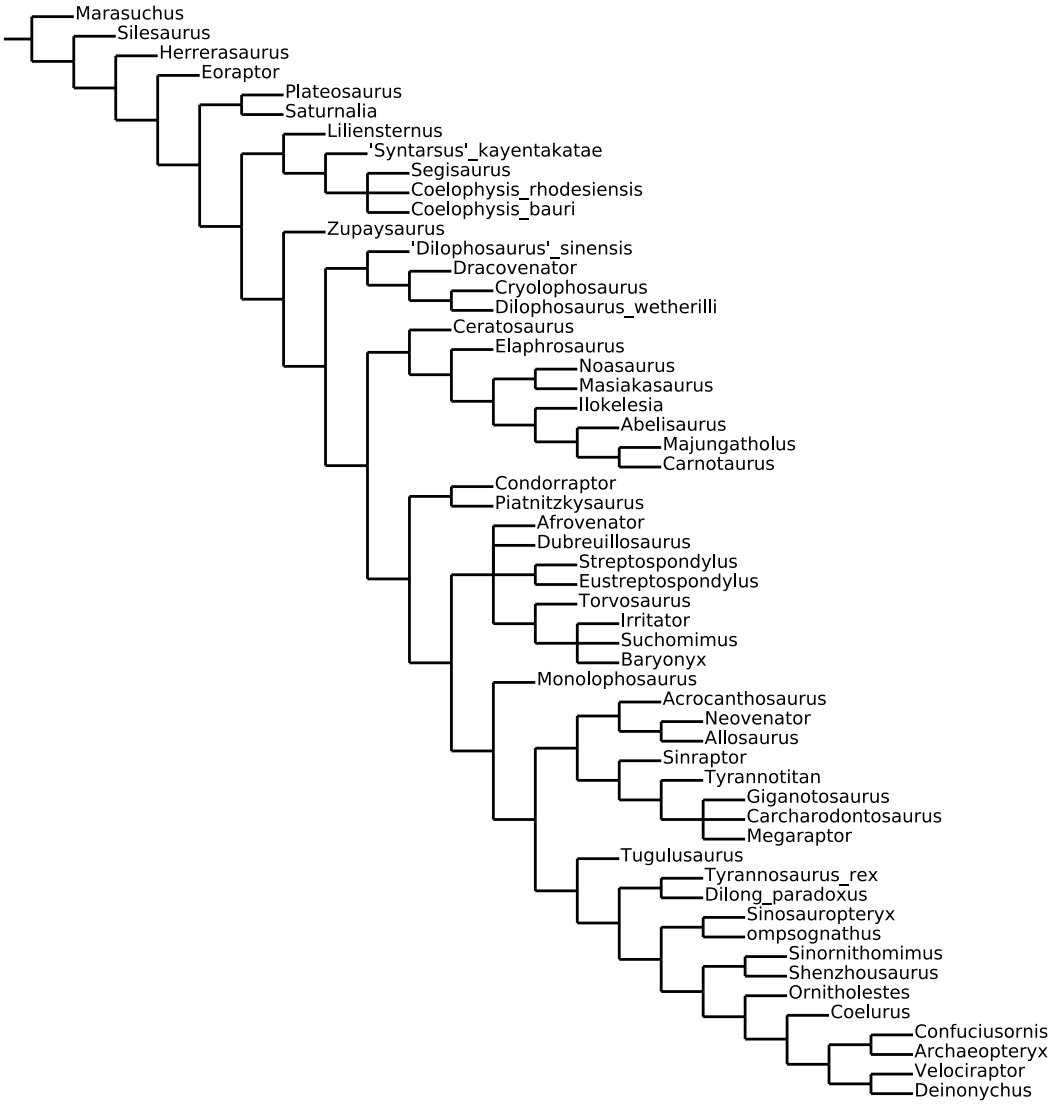


Figura 5. Árvore de consenso do *dataset* de terópodes sob o critério de máxima parcimônia de pesos implícitos

mutação, hiperparâmetros definidos para o *software* e método de avaliação dos resultados obtidos; e nos trabalhos de [Cotta and Moscato 2002] e [Moilanen 1999], adaptando o operador de recombinação PDR e possivelmente, utilizando as codificações de cromossomos e otimizações de código desenvolvidas para ambos os trabalhos. Para as análises, serão utilizadas avaliações por MP tanto convencional quanto de pesos implícitos, aplicadas como funções de aptidão para o AG. O consenso será calculado de forma estrita, e os suportes deverão se basear em *bootstrapping* e pelo método de Bremer. O algoritmo será comparado com buscas *hill climbing* tradicionais sob o mesmo *framework*, bem como com o algoritmo de *branch and bound*, quando cabível.

Além da acurácia da resposta, também serão coletadas outras métricas de desempenho do algoritmo, como distância para a resposta correta, tempo de execução, e possivelmente uso de memória e CPU. O objetivo é avaliar o custo do algoritmo em relação

aos métodos tradicionais caso tenha acurácia superior aos já usados, ou mesmo identificar se gasta mais recursos com desempenho similiar. Para isso, é importante que todos os algoritmos sejam executados em uma mesma base de código, para eliminar ao máximo artefatos de implementação nesses resultados. Em último caso, o resultado pode ser comparado a *softwares* existentes, como TNT e PAUP*.

Por fim, vale ressaltar que o objetivo deste trabalho não será desenvolver um *software* completo de análise filogenética, mas sim testar o desempenho de algoritmos genéticos como mecanismos de busca para filogenias baseadas em dados paleontológicos. O desenvolvimento de uma ferramenta pronta para uso requer a elaboração de vários detalhes que não serão levados em conta para este trabalho, como otimizações de código, criação de interface de usuário intuitiva, habilitação de configurações para o usuário, entre outras. O objetivo é demonstrar (ou não) que há benefício na implementação futura de *softwares* com esse conceito, e pavimentar o caminho para as próximas pesquisas na área de algoritmos genéticos para análises filogenéticas.

Reconhecimentos

Este trabalho foi feito usando a versão 1.6 do software TNT, disponibilizado gratuitamente para usuários finais pela Willi Hennig Society.

Referências

Adams III, E. N. (1972). Consensus techniques and the comparison of taxonomic trees. *Systematic Biology*, 21(4):390–397.

Azouri, D., Abadi, S., Mansour, Y., Mayrose, I., and Pupko, T. (2021). Harnessing machine learning to guide phylogenetic-tree search algorithms. *Nature communications*, 12(1):1983.

Berger, S. A. and Stamatakis, A. (2010). Accuracy of morphology-based phylogenetic fossil placement under maximum likelihood. In *ACS/IEEE International Conference on Computer Systems and Applications-AICCSA 2010*, pages 1–9. IEEE.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., and Drummond, A. J. (2014). BEAST 2: a software platform for bayesian evolutionary analysis. *PLoS computational biology*, 10(4):e1003537.

Brauer, M. J., Holder, M. T., Dries, L. A., Zwickl, D. J., Lewis, P. O., and Hillis, D. M. (2002). Genetic algorithms and parallel processing in maximum-likelihood phylogeny inference. *Molecular Biology and Evolution*, 19(10):1717–1726.

Bremer, K. (1988). The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution*, 42(4):795–803.

Cau, A. (2017). Specimen-level phylogenetics in paleontology using the fossilized birth-death model with sampled ancestors. *PeerJ*, 5:e3055.

Cavalli-Sforza, L. L. and Edwards, A. W. (1967). Phylogenetic analysis. Models and estimation procedures. *American journal of human genetics*, 19(3 Pt 1):233.

Chor, B. and Tuller, T. (2005). Maximum likelihood of evolutionary trees: hardness and approximation. *Bioinformatics*, 21(suppl_1):i97–i106.

Cotta, C. and Moscato, P. (2002). Inferring phylogenetic trees using evolutionary algorithms. In *International Conference on Parallel Problem Solving from Nature*, pages 720–729. Springer.

Edwards, A. W. and Cavalli-Sforza, L. L. (1963). The reconstruction of evolution. In *Annals of Human Genetics*, volume 27, pages 105–106.

Felsenstein, J. (1973). Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Biology*, 22(3):240–249.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39(4):783–791.

Felsenstein, J. (2004). *Inferring phylogenies*. Sinauer Associates, Inc.

Goloboff, P. A. (1993). Estimating character weights during tree search. *Cladistics*, 9(1):83–91.

Goloboff, P. A. and Morales, M. E. (2023). TNT version 1.6, with a graphical interface for macos and linux, including new routines in parallel. *Cladistics*, 39(2):144–153.

Goloboff, P. A., Torres, A., and Arias, J. S. (2018). Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology. *Cladistics*, 34(4):407–437.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology*, 59(3):307–321.

Hendy, M. D. and Penny, D. (1982). Branch and bound algorithms to determine minimal evolutionary trees. *Mathematical Biosciences*, 59(2):277–290.

Margush, T. and McMorris, F. R. (1981). Consensus$n$-trees. *Bulletin of Mathematical Biology*, 43(2):239–244.

Moilanen, A. (1999). Searching for most parsimonious trees with simulated evolutionary optimization. *Cladistics*, 15(1):39–50.

Mongiardino Koch, N., Garwood, R. J., and Parry, L. A. (2021). Fossils improve phylogenetic analyses of morphological characters. *Proceedings of the Royal Society B*, 288(1950):20210044.

Noutahi, E. and El-Mabrouk, N. (2018). GATC: a genetic algorithm for gene tree construction under the duplication-transfer-loss model of evolution. *BMC genomics*, 19:97–107.

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, 61(3):539–542.

Smith, N. D., Makovicky, P. J., Hammer, W. R., and Currie, P. J. (2007). Osteology of cryolophosaurus ellioti (dinosauria: Theropoda) from the early jurassic of antarctica and implications for early theropod evolution. *Zoological Journal of the Linnean Society*, 151(2):377–421.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, 4(1):vey016.

Tschopp, E. and Upchurch, P. (2018). The challenges and potential utility of phenotypic specimen-level phylogeny based on maximum parsimony. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 109(1-2):301–323.

Vega-Dias, C., Maisch, M. W., and Schultz, C. L. (2004). A new phylogenetic analysis of Triassic dicynodonts (Therapsida) and the systematic position of *Jachaleria candelariensis* from the Upper Triassic of Brazil.(with 8 figures and 1 table). *Neues Jahrbuch für Geologie und Paläontologie Abhandlungen*, 231(2):145–166.

Villalobos-Cid, M., Dorn, M., Contreras, Á., and Inostroza-Ponta, M. (2023). An evolutionary algorithm based on parsimony for the multiobjective phylogenetic network inference problem. *Applied Soft Computing*, 139:110270.

Wilgenbusch, J. C. and Swofford, D. (2003). Inferring evolutionary trees with PAUP. *Current protocols in bioinformatics*, (1):6–4.

Zambrano-Vega, C., Nebro, A. J., and Aldana-Montes, J. F. (2016). MO-phylogenetics: a phylogenetic inference software tool with multi-objective evolutionary metaheuristics. *Methods in Ecology and Evolution*, 7(7):800–805.

Zwickl, D. J. (2006). *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion*. The University of Texas at Austin.