IPS Instituto Politécnico de Setúbal
Escola Superior de Tecnologia do Barreiro

JOÃO PEDRO
COELHO RAIMUNDO

Estudo das metodologias *BEAST* e *StarBEAST,* aplicadas à inferência filogenética molecular – relógios moleculares

Relatório de Estágio
Licenciatura em Bioinformática

ORIENTADORES

Professor Francisco Pina-Martins

Professor Octávio S. Paulo

dezembro 2019

Agradecimentos

A elaboração deste projeto não teria sido possível sem a colaboração direta ou indireta, estímulo e apoio de diversas pessoas e instituições às quais estou profundamente grato.

É imprescindível começar por agradecer ao orientador professor Francisco Pina-Martins pelo contínuo apoio prestado ao longo do projeto. À coordenação disponibilizada em todas as etapas do projeto, a todo o conhecimento que me foi partilhado, a todas as sugestões essenciais ao meu crescimento na área, sem nunca esquecer o estímulo diário que me deram forças para continuar a cada dia o projeto. A paciência que teve é notória e estou completamente agradecido.

Agradeço ao orientador professor Octávio S. Paulo por todas as bases prestadas que foram fulcrais para o começo da realização do projeto, à coordenação inicial que me ajudaram a definir uma linha de orientação coesa e a todo o apoio prestado. Um sincero obrigado.

Um sincero agradecimento aos investigadores Eduardo Marabuto e Miguel Simões Nunes que me integraram de mãos abertas no seu projeto de inferência filogenética a borboletas. Onde me foram partilhadas grandes noções de interpretação biológica e onde desenvolvi verdadeiramente a capacidade de inferir resultados com base na filogenia molecular.

Agradeço à Equipa CoBiG² (Computacional Biology and Population Genomics) da cE3c (Centro de Ecologia, Evolução e Alterações Ambientais) por me ter acolhido nesta importante etapa da minha vida. Onde me foi transmitido um espírito de equipa e amizade, prestada ajuda essencial no decorrer do projeto e sempre demonstraram interesse em partilhar o seu conhecimento, sugestões e experiências que me ajudaram a crescer profissionalmente e como pessoa. À Ana Sofia Rodrigues, ao Duarte Balata, ao Gonçalo Costa, ao Paulo Sousa, ao Pedro David, à Sara Ema Silva e à Sofia Seabra, um honesto obrigado.

Não podia deixar de agradecer à Escola Superior de Tecnologia do Barreiro do Instituto Politécnico de Setúbal, onde tive o prazer de conhecer grandes profissionais como o professor Ricardo Leite que agradeço grande parte do conhecimento que tenho hoje em Bioinformática. Agradeço também à Faculdade de Ciências da Universidade de Lisboa que me acolheu para a realização do presente estágio e me permitiu conhecer excelentes pessoas e profissionais.

É essencial agradecer à minha família pela força, carinho e compreensão que sempre me foi prestado ao longo de toda a minha académica. Bem como à minha namorada

que sempre esteve ao meu lado, aconselhando-me, apoiando-me e incentivando-me a cada dia. Agradeço-lhe toda a paciência e força notória prestada que me levaram a conseguir concluir esta etapa. Estou-vos eternamente grato.

Finalmente agradeço aos meus amigos e colegas mais próximos que indiretamente me ajudaram durante esta jornada, ao interesse e apoio prestado.

Índice

Índice de Figuras

Índice de Tabelas

Lista de Siglas e Abreviaturas

BEAST - *Bayesian Evolutionary Analysis Sampling Trees*

OTU - *Operacional Taxonomic Unit*

MRCA - *Most Recent Common Ancestral*

UPGMA - *Unweighted Pair Group Method with Arithmetic Mean*

NJ - *Neighbour-Joining*

LS - *Least Square*

ME - *Minimum Evolution*

MP - *Maximum Parsimony*

ML - *Maximum Likelihood*

MAS - *Multiple sequence alignment*

MCMC - *Markov chain Monte Carlo*

pp - posterior probability

Ma - Milhão de anos

MRK - Grupo constituído pelas cigarras do género Maoricicada, Rhodopsalta e Kikihia

NCBI - *National Center for Biotechnology Information*

API - *Application programming interface*

Entrez - Global Query Cross-Database Search System

ESS - *Effective Sample Size*

Cyt b - Cytochrome b

COI - Cytochrome c Oxidase Subunidade I

EF-1α - Elongation Factor 1-alpha

Wg - Wingless

β Fibrinogen - Fibrinogen beta chain

C-mos - Oocyte maturation factor

Resumo

O desenvolvimento de novos métodos, que possibilitam a implementação de relógios moleculares na inferência filogenética, tem ao longo do tempo permitido à comunidade científica obter resultados mais robustos, suportados pelo uso de registos fósseis e geológicos. Neste projeto foram comparados os métodos de inferência filogenética molecular implementados no *software BEAST2* e *StarBEAST2,* para que se tenha uma ideia de em que cenário um é mais vantajoso aplicar um ou o outro.

Deste modo, foram realizadas inferências moleculares independentes a espécies de cigarras dos géneros *Maoricicada*, *Rhodopsalta* e *Kikihia*, a espécies de lagartos ocelados do género *Timon* e por último a espécies de borboletas dos géneros *Lycaena* e *Melanargia.* As análises tiveram por base a aplicação da metodologia *StruggleBeast,* um *workflow* desenvolvido durante o projeto que permite a realização de inferências filogenéticas moleculares inter e intraespecíficas, com base nos recursos disponibilizados pelo *package* do *software BEAST2.*

Os resultados obtidos demonstram que é fundamental o uso de um *outgroup* adequado na calibração da árvore filogenética, que garanta suporte de confiança à inferência realizada. Não obstante, a proporcionalidade entre o número de genes e os *taxa* envolvidos na inferência são fulcrais à resolução dos nodos. Relógios inapropriados ou calibrações incorretas conduzem a resultados ilusórios de escalas de tempo. Por isso, de modo a garantir a idade mínima de existência dos *taxa*, é importante o uso de registos fósseis.

O *software StarBEAST2* demonstra obter resultados com melhor resolução, intervalos de divergência menores e um tempo de processamento no mínimo duas vezes mais rápido comparativamente ao *software BEAST2*, quando aplicado a *datasets* com pouca variabilidade interespecífica. Deve-se ter sempre em atenção o contexto biológico quando os resultados são inferidos, pois mesmo que sejam bem suportados pela análise dos valores de confiança, o ponto de vista informático nunca deve ser interpretado isoladamente.

Palavras Chave

Inferência filogenética; Relógios Moleculares, *BEAST2*; *StarBEAST2;*

Abstract

The development of new methods for implement molecular clocks in the phylogenetic inference, has over time allowed the scientific community to achieve more prudent results, supported by the use of fossil and geological records. In this project, the *BEAST2* and *StarBEAST2* molecular phylogenetic inference methods were compared, in order to conclude their diferences, and in which context one is more relevant to be applied than the.

Thus, independent molecular inferences were performed to cicadas' species of the genera *Maoricicada*, *Rhodopsalta* e *Kikihia*, to ocellated lizards' species of the genus *Timon* and lastly to butterflies' species of the genera *Lycaena* and *Melanargia*. The study were based on the implementation of the *StruggleBeast* methodology, a workflow developed during the project that allows inter and intraspecific molecular phylogenetic inferences, based on the resources provided by the *BEAST2* software package.

The results show that it is essential to use an adequate outgroup in the phylogenetic tree calibration, for the purpose of granting support value for the carried out inference. Furthermore, proportionality between the number of genes and *taxa* involved in the inference are fundamental to ensure node resolution. Inappropriate pacemakers or incorrect calibrations lead to illusory timescale results. Therefore, in order to ensure the minimum age of *taxa'* existence, the use of fossil records is very important.

StarBEAST2 software demonstrates better node' resolution results, shorter divergence times and at least twice fast processing times compared with *BEAST2 software* results, when applied to datasets with low interspecific variability. The biological context must always be taken into account, because even when well supported by confidence analysis, the IT point of view should never be interpreted in isolation

Keywords

Phylogenetic inference; Molecular Clocks, *BEAST2*; *StarBEAST2*;

1 Introdução

1.1 Objetivos do estágio

O recurso a relógios moleculares é cada vez mais frequente nos estudos filogenéticos. Isto levou ao surgimento de múltiplas implementações destes métodos que podem apresentar resultados diferentes.

O objetivo do projeto é reanalisar o resultado de publicações com diferentes software/ algoritmos e compará-los, permitindo assim inferir como diferentes fatores podem influenciar as conclusões biológicas.

1.2 Caracterização da Instituição

O presente estágio foi desenvolvido no "Center for Ecology, Evolution and Environmental Changes" (CE3c) da Universidade de Lisboa, coordenado pela professora Cristina Máguas. Mais propriamente, na equipa de "Computational Biology and Population Genomics" (CoBiG^2), liderada pelo professor Octávio S. Paulo.

A missão do CoBiG^2 é estudar a diversificação evolutiva e ecológica de espécies em ambientes naturais, bem como o processo genómico de adaptação dos organismos aos seus habitats, tendo como abordagem principal a genómica populacional.

1.3 Atividades desenvolvidas

Tarefa A – Aprendizagem dos *software BEAST2* e **BEAST2*: leitura e escolha das publicações adequadas ao projeto, leitura da documentação do *software*; reprodução de inferências a partir de exemplos;

Tarefa B – Inferência filogenética molecular em espécies de cigarras dos géneros *Maoricicada, Rhadopsalta* e *Kikihia*, com os métodos *BEAST2* e **BEAST2* baseada no gene *COI;*

Tarefa C – Inferência filogenética molecular em lagartos ocelados (género *Timon*), com os métodos *BEAST2* e **BEAST2* a partir do gene *Cyt b*;

Tarefa D – Inferência filogenética molecular em borboletas do género *Lycaena*, com o *software BEAST2* a partir dos genes *EF-1α e COI* (*dataset* concatenado)*;*

Tarefa E – Inferência filogenética molecular em borboletas do género *Melanargia* (subgénero *Argeformia*), com o *software BEAST2* a partir dos genes *16S, COI, EF-1α e Wg* (*dataset* concatenado)*;*

Tarefa F – Inferência filogenética molecular em lagartos ocelados (género *Timon*), recorrendo aos métodos *BEAST2* e **BEAST2* a partir dos genes *12S, 16S, Cyt b, B Fibrinogénio* e *C-mos* (*dataset* concatenado)*;*

1.4 Cronograma do estágio

	Semanas																		
Atividades	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Tarefa A																				
Tarefa B																				
Tarefa C																				
Tarefa D																				
Tarefa E																				
Tarefa F																				

Figura 1 - Cronograma do estágio realizado no grupo de investigação CoBiG², de 11 de março a 26 de julho de 2019.

2 Fundamentos teóricos

A comunidade científica ao longo do tempo têm-se focado em dar resposta a questões direcionadas à evolução dos organismos. Questões como: "A evolução é progressiva?" e "A seleção natural leva necessariamente aos organismos evoluírem para seres mais complexos?", têm ao longo do tempo vindo a ser fundamentadas a partir do desenvolvimento de novos métodos [1, 2].

Os métodos de inferência filogenética visam recriar uma história evolutiva da divergência das espécies a partir da sua informação genética, tendo por base datações de acontecimentos geológicos, calibrações fósseis entre outros [1, 2].

2.1 Inferência Filogenética

A inferência filogenética tem como objetivo retratar, a partir de métodos estimativos, a história evolutiva de um grupo de organismos, denominados como *taxa*, ou famílias de genes a partir da relação entre *OTUs* (*Operacional Taxonomic Units*) [3].

Hoje em dia, a inferência filogenética é fundamental em várias áreas de investigação, tais como o estudo da sistemática biológica e da biodiversidade [4], epidemiologia molecular [5], identificação de funções genéticas [6], estudos do microbioma [7], análises forenses [8], descoberta de novos fármacos [9], e até mesmo no desenvolvimento de vacinas [10]. [3]

2.1.1 Análise filogenética com base em dados de sequências

As análises filogenéticas são efetuadas com base na informação genética dos organismos, sequências de nucleótidos adquiridas a partir de métodos de sequenciação na qual se dá o nome de alinhamento. A criação de um ficheiro que possua os alinhamentos de todos os organismos permite dar resposta aos *software* de filogenética. Este ficheiro pode ter vários tipos de formatação, porém a mais comum é o formato *FASTA.* A partir do ficheiro *FASTA* é possível obter um *input* matricial que pode ser modelado noutros formatos como o *Nexus*. [3]

2.1.2 Árvores Filogenéticas

Os resultados de inferência filogenética são representados a partir de gráficos em esquema de árvore. A partir da sua interpretação é possível inferir as relações evolutivas num grupo de *OTUs*. Geralmente, o *OTU* representa uma espécie, mas também pode representar individualmente organismos de uma população, sequências de genes e proteínas ou um *taxon*, independentemente do *rank* taxonómico (família, ordem, classe, filo) [3].

Tendo por via uma abordagem *top-down,* aos nodos situados no extremo da árvore dá se o nome de *external nodes*, que representam cada uma das *OTUs* envolvidas na análise enquanto o ancestral hipotético mais recente entre duas *OTUs* é designado de *internal node*. Assim, a ligação que se dá entre estes dois tipos de nodos é denominada por *branch* (ou ramo), sobre a qual se demonstra a relação evolutiva entre os *taxa*. O *branch* que liga dois *internal nodes* é classificado como *internal branch*, sendo caracterizado como uma relação antiga. Da mesma forma que, o *branch* que liga um *internal node* a um *external node* é denominado por *external branch*. A *root* (ou raiz) representa a primeira divergência da árvore, o mais recentre ancestral em comum (*Most Recent Common Ancestral* - MRCA) de todos os *taxa* / *OTU' s* ou *external nodes;* [3]. A *Figura 2* exemplifica a composição de uma árvore filogenética, bem como os principais termos acima referidos.

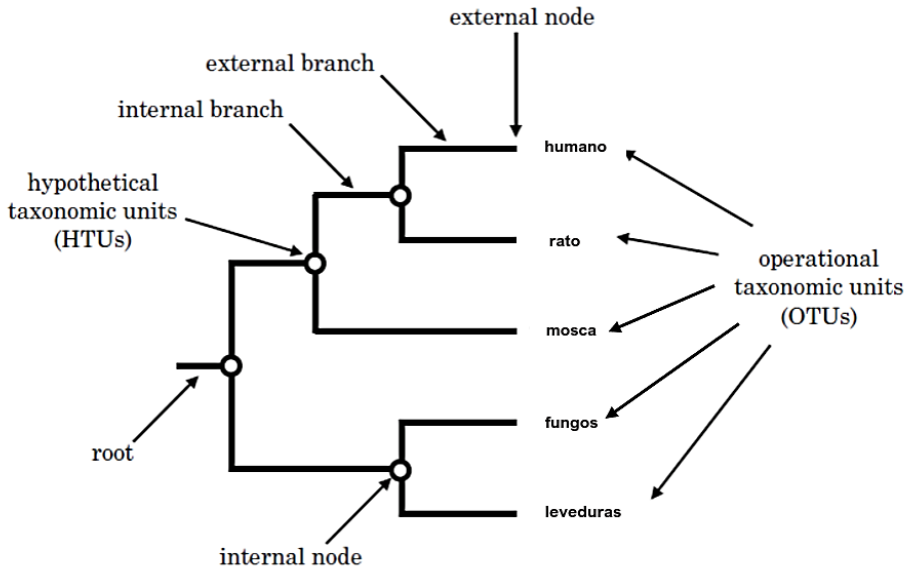


Figura 2 - Composição de uma árvore filogenética. Terminologia frequentemente usada; Fonte [3].

Geralmente, o *software* para obtenção de uma filogenia apenas consegue reconstruir uma *unrooted tree*. De forma a conferir um maior valor ao diagrama num contexto evolutivo, é desejável inferir uma *rooted tree* a partir da identificação da origem de todos os *taxa*. O melhor método é adicionar um *outgroup* ao *dataset* [3]. O *outgroup* é geralmente um *OTU*, *taxon* ou grupo de organismos mais distante das espécies a serem estudas, que serve de referência na determinação das relações evolutivas do *ingroup* (que é composto pelos *taxa* sob estudo) de modo a calibrar as relações da árvore. De forma que, teoricamente a *root* da árvore esteja localizada entre o *outgroup* e os restantes *taxa*. Assim, um bom *outgroup* será um *OTU* que tenha divergido recentemente dos restantes *taxa*, mas suficientemente distinto dos mesmos a ponto de não inferir com o *ingroup.* [3]

Todos os *taxa* que descendam do mesmo ancestral em comum são definidos como grupo monofilético ou *clade*. Já a um grupo de *OTU* que partilham o mesmo ancestral, mas que não possuem todos os membros descendentes, dá se o nome de grupo parafilético ou *glade* [3].

2.1.3 Métodos de Inferência Filogenética Convencionais

Os métodos de inferência filogenética têm como componentes básicos os modelos evolutivos, os critérios de otimização e a avaliação de árvores. Os modelos evolutivos são desenvolvidos matematicamente tendo por base testes de hipóteses e são usados para modelar o processo evolutivo. Já os critérios de otimização maximizam ou minimizam o valor dos parâmetros que fornecem uma base de comparação entre as árvores, como por exemplo a parcimónia, método dos mínimos quadrados e máxima verosimilhança (*parsimony*, a *least-squares distances* e *maximum likelihood)*, a serem abordados mais à

frente. Finalmente as árvores são avaliadas no decorrer de um processo de *search* como o *stepwise addition with branch swapping* e o *branch-and-bound*, de acordo com os critérios de otimização de modo a selecionar a melhor árvore [3].

Os métodos de inferência filogenética podem ser classificados a partir de duas abordagens distintas: métodos de distância e métodos de caracteres [3].

2.1.3.1 Métodos de Distâncias

Os métodos de distâncias têm por base a formulação da matriz de alinhamentos criada com a informação genética dos taxa. O método cria uma nova matriz convertendo o emparelhamento das sequências em valores de distância (dissimilaridades). Esta é designada como matriz de distâncias [3].

Uma vez a matriz de distâncias criada, a matriz constituída pelos alinhamentos já não será mais usada na inferência filogenética. Existem vários métodos que possibilitam inferir qual a melhor árvore como a UPGMA (*Unweighted Pair Group Method with Arithmetic Mean*) [11], *Neighbour-Joining* (NJ, consultar *Tabela 1*) [12], *Least Square* (LS) [13] ou *Minimum Evolution* (ME) [14]; [3].

A maior vantagem dos métodos de distâncias é a grande velocidade de processamento, devido à redução drástica da quantidade de informação a ser processada pela conversão matricial. Além disso, este método pode obter resultados fiáveis se a homoplasia [15] (evolução independente dos mesmos caracteres) for rara e se estiver distribuída aleatoriamente pela árvore. No entanto, a redução dos dados leva a uma perda de informação genética e pode eventualmente gerar *branch lengths* negativos, que carece de significado biológico. Contudo, métodos baseados em distâncias, como o NJ, são recomendados para grandes bases de dados (>1000 sequências) com um elevado nível de semelhança. [3]

2.1.3.2 Métodos de Caracteres

Ao longo do tempo, vários métodos foram desenvolvidos para lidar com caracteres (sequências de DNA e proteínas alinhadas), tais como a *Maximum Parsimony* (MP) [16], a *Maximum Likelihood* (ML) bem como métodos de inferência Bayesiana. Estas abordagens visam fazer uma reconstrução filogenética usando diretamente os alinhamentos sem qualquer tipo de transformação. Comparativamente aos métodos de distâncias o processo é mais lento, porém a inferência da árvore final é substancialmente mais precisa biologicamente. Sucintamente, o algoritmo começa por fazer o *score* a todas as possibilidades filogenéticas a partir de *n taxa*. A árvore mais adequada será a aquela que apresenta o melhor *score,* considerando os dados das sequências. [3]

Os algoritmos de pesquisa, implementados na maioria dos *software* de filogenia são o *branch-and-bound* [17] e o método heurístico [18], já referidos anteriormente (*Tabela 1*). O

primeiro método referido, numa primeira instância, começa por criar uma árvore *core* a partir de três *taxa* do *dataset*. De seguida, aleatoriamente e de forma incremental são adicionados os restantes *taxa* da matriz, selecionando sempre a árvore com o melhor *score*. Por outro lado, o método heurístico é semelhante ao *branch-and-bound*, mas em vez de ser adicionado um novo *taxon* à árvore que possui o melhor *score*, relativamente à anterior, o método heurístico usa apenas a árvore com o melhor *score* em cada etapa incremental. [3]

Tabela 1 - Métodos de construção e de search de árvores filogenéticas; Fonte [35].

Método	Descrição	Vantagens	Desvantagens
Métodos de construção de árvores filogenéticas			
"Stepwise addiction"	Contrói uma árvore completa, começando com três sequências e adicionando incrementlmente novas sequências, uma de cada vez, à branch que possui a árvore mais adquada.	Método de rápida implementação; Etapas posteriores podem reverter processos anteriores.	Processa uma árvore, que muitas vezes não é a mais indicada; sequências alternativas adicionais podem contruir árvores com uma topologia diferente; não é tão rápido como o "neighbour-joining"
"Star decomposition"	Constrói uma "resolved tree", começando por todas as sequências estarem conectadas a um único 'hub node'. A cada passo, duas linhagens são adicionadas ao 'hub node', tornando-se "neighbours", escolhidos de forma a que a árvore seja mais adequada à inferência.	Método de rápida implementação; sequencias adicionais são irrelevantes	Processa uma árvore, que muitas vezes não é a mais indicada; os "neighbours" não podem ser separados em etapas posteriores; não é adquado a alguns métodos.
"Neigbour joining"	Um método de "Star decomposition" que usa a aproximação a um minimo evolutivo (critério de optimização)	Um dos métodos mais rápidos na construção de árvores evolutivas	Processa uma árvore, que muitas vezes não é a mais indicada; os "neighbours" não podem ser separados em etapas posteriores; não é adquado a alguns métodos.
Métodos de "search" a árvores filogenéticas			
"Heuristic search"	Fornece uma árvore i exaustiva :ontem todas as sequências, executa o processo de "branch swapping" para produzir árvores alternativas, de modo a encontrar uma árvore mais optimizada	Máis rápido que as "Exact searches"	Pode não encontrar a árvore mais adquada
"Exact search"	Processa uma search "exautiva" de modo a examinar todas as árvores possiveis, garantindo a melho como output. As técnicas de "branch-and-bound" podem eliminar as piores árvores continuando a garantir o melhor output.	O único método que garante inferir a melhor árvore	Tempo de processamento: apenas praticável para processar algumas sequencias (<20)

2.1.3.2.1 Método *Maximum Parsimony*

O método *Maximum Parsimony* (MP) [16, 19] foi um dos primeiros modelos de substituição a surgir. É o mais usado para a inferência de árvores que têm por base um *dataset* de carácter morfológico, na qual é difícil de calcular a taxa (*rate*) de mutação evolutiva. Quando o método MP é aplicado, cada coluna da MSA (*multiple sequence alignment*) é processada como um caracter individual. Porém, nem todas as posições dos alinhamentos são propícias a esta metodologia, como é o caso dos sítios ou locais invariáveis *(invariable sites*). Caracteres (colunas da MSA) que possuem mais do que uma diferença entre os seus nucleótidos ou aminoácidos são designados de *parsimony informative sites*. Apenas estes são usados na inferência pelo método MP, a qual se baseia na pesquisa pela árvore mais parcimoniosa, isto é, aquela que menos passos necessitou para ser construída. Deste modo, a árvore mais curta possível e que consegue inferir relações entre taxa é considerada a mais adequada. Quanto menor for a homoplasia nas sequências, mais preciso será o

resultado da inferência (a MP visa minimizar a homoplasia dos alinhamentos). A MP é um critério de otimização simples e intuitivo, podendo ser facilmente aplicado a qualquer tipo de dados, como os *indels.* Porém, é pouco eficiente ao lidar com alinhamentos com elevado nível de variação e matrizes concatenadas com múltiplos genes. [3]

2.1.3.2.2 Método *Maximum Likelihood*

A *Maximum Likelihood* (ML) é um método estatístico usado para estimar parâmetros de um modelo a partir dos dados (sequências de DNA, proteínas entre outros), implementado pela primeira vez, em inferência filogenética, por Felsenstein [20]. O modelo ML visa estimar as *branch lenghts* e a topologia da árvore com base em modelos de substituição (anteriormente referidos) e alinhamentos. O *output* da análise é a probabilidade do *fit* entre a topologia, bem como o modelo, com a informação genética das matrizes. O processo é repetido para todas as possibilidades topológicas a partir de n *taxa*. A topologia da árvore com o maior score de *likelihood* é reportada como sendo a melhor árvore inferida, a *Maximum Likelihood tree*.

Para inferir a evolução das sequências, são usados modelos de substituição que calculam os valores das distâncias. O método mais simples, que pode inferir quer a distância de sequência de nucleótidos quer de proteínas é designado de *p-distance* que designa a percentagem de diferença entre as sequências observadas. Existem modelos de substituição como o *Jukes-Cantor* (JC64) considerado o modelo mais simples, que assume que todas as substituições dos nucleótidos ocorrem com o mesmo *rate* [21], enquanto que o modelo *Kimura* de dois parâmetros (K80) trata a ocorrência de transições e transversões como eventos diferentes, com probabilidades diferentes de ocorrer [22]. Ambos assumem que a substituição nucleotídica se aproxima de um equilíbrio, caso contrário é necessário implementar outros modelos que se adaptem às variações observadas como é o caso do *F81* [23], *HKY85* [24], *TN93* [25], *GTR* [26] entre outros (para mais detalhes consultar [27–29]). Usar o modelo mais apropriado para a inferência da árvore filogenética é fulcral para evitar erros de *clustering.* Existem vários software que visam testar qual o modelo de substituição mais adequado, tomamos por exemplo o *ModelTest* [30] e o *jModelTest 2* [31].

O melhor aspeto do método ML é a sua precisão, devido ao facto de depender dos modelos de substituição, e usar toda a informação genética disponibilizada nas matrizes para inferir os resultados, ao contrário do método MP. Contudo, o tempo de processamento é longo tornando-se um ponto fraco do modelo, tornando-se impraticável a análise a *datasets* de grande escala. Isto porque necessita de recursos computacionais relativamente avultados [3], [32].

2.1.3.2.3 Método Bayesiano

A inferência Bayesiana é um método recente, implementado pela primeira vez na filogenia há cerca de duas décadas atrás [33]. O algoritmo do método Bayesiano infere a árvore com a maior *posterior probability* (*pp*) [34], tendo em conta um grande número de possibilidades, através do algoritmo *Markov chain Monte Carlo* (MCMC) [35], aprofundado mais à frente. Existem alguns célebres programas filogenéticos que implementam o algoritmo de inferência Bayesiana, como é o caso do *RAxML* [36], *MrBayes* [37, 38] e *BEAST2* [39].

2.1.4 Avaliação dos valores de suporte das árvores filogenéticas - confiança

Uma fraqueza há muito discutida, no que diz respeito a métodos filogenéticos, é o facto dos resultados serem estimativas pontuais da filogenia. De modo que, é necessário dar resposta às perguntas “Quão suportadas são as árvores resultantes do processo de inferência? Como garantimos a sua robustez?”. Tradicionalmente estas questões são encaradas pelo algoritmo estatístico designado de *bootstrap*. A ideia deste método é a matriz original ser aleatoriamente re-amostrada, com substituição dos *sites*, para produzir data sets pseudo-replicados. Quando usados métodos que têm por base critérios de otimização, é iniciada uma *tree search* (*Figura 3-a*, caixa verde) para cada pseudo *dataset*, com base em algoritmos desenvolvidos para tal efeito: *Heuristic search* ou *Exact search* (*Tabela 1*). Uma árvore inicial é escolhida de forma aleatória ou a partir do resultado de um algoritmo, como *Neighbour joining*, *Stepwise addiction* ou *Star decomposition* (*Tabela 1*). A nova árvore é classificada e, se aceite, adicionada à coleção de árvores final. O *bootstrap* é um processo cíclico com termino de acordo com o número de iterações definidas (*Figura 3-a*). O número de vezes que um grupo de sequências ocorre na árvore, durante o processo de amostragem, pode ser usado como medida de quão fortemente o grupo é suportado pelos dados. De forma que, a árvore com melhor *score* de *likelihood* será a que terá a que terá maior nível de confiança [35].

Em análises Bayesianas, a avaliação dos valores de suporte não tem por base o uso do algoritmo *bootstrap*. Uma vez que, como abordado anteriormente, a estatística Bayesiana assenta na especificação de modelos e *priors* (parâmetros definidos durante o processo de modelação, normalmente assentam sobre valores de distribuição num dado tipo de função) para determinar a *posterior probability* (*pp*) de cada árvore, tendo em conta a integração dos valores dos parâmetros. De modo que, os algoritmos de *likelihood* se tornam complexos para serem integrados analiticamente em modelos filogenéticos. Assim, os métodos Bayesianos dependem da MCMC para avaliar a confiança das árvores inferidas. A MCMC é um algoritmo notável usado na aproximação de distribuições probabilísticas, numa ampla variedade de contextos. Esta tem por via um processo cíclico (*Figura 3-b*) na qual se cria uma *chain* com uma serie de passos independentes. A cada passo, uma nova localização para os parâmetros é proposta, de modo a criar uma nova ligação da *chain*. A localização proposta é similar à usada anteriormente, devido a ser criada aleatoriamente a partir do

reajuste dos parâmetros. A densidade da *pp* na nova localização é calculada; caso esta tenha um *score* superior ao reajuste anterior, é criada uma nova localização e a chain é movida [35].

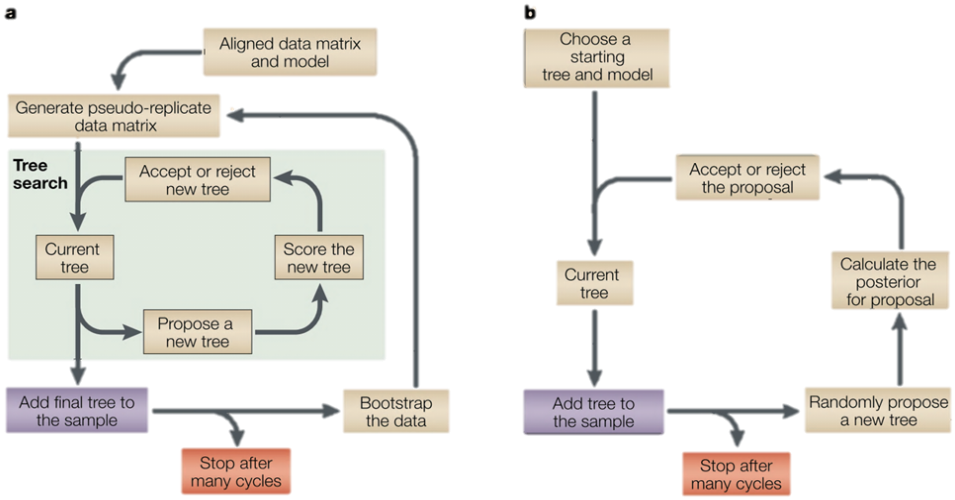


Figura 3 - Workflow dos algoritmos bootstrap (a) e MCMC (b); Fonte [35].

2.2 Relógios Moleculares

Durante os anos 60, grupos de cientistas, incluindo Emile Zuckerkandl e Linus Pauling [40], inferiram através de experiências com proteínas, que estas sofriam substituições entre os aminoácidos a uma taxa (*rate*) surpreendentemente consistente ao longo do tempo, em várias espécies distintas. De modo que, esta *rate* evolutiva, única e uniforme, foi descrita como relógio molecular. Rapidamente os biólogos aperceberam-se da forma como este método poderia ser útil para inferir uma escala de tempo para a evolução das espécies: estimar o *rate* de substituições dos aminoácidos, por unidade de tempo, e aplicá-la às diferenças proteicas num grupos de organismos, permitiria inferir os tempos de divergências dessas respetivas linhagens. [41]

50 anos depois, os avanços das tecnologias de sequenciação e o aparecimento de novas ferramentas, revelaram que as *rates* de variação genética variavam ao longo das linhagens. O termo relógio molecular refere-se atualmente a um conjunto de métodos e modelos, que inferem a forma de como as *rates* de evolução genética variam. Estes métodos possibilitam estimar uma escala de tempo relativa, de modo a permitir obter uma perspetiva cronológica dos grandes acontecimentos que levaram à divergência das espécies, como por exemplo a formação das ilhas Havaianas representada no *Anexo 1*. [41, 42].

2.2.1 Taxa de Variação Evolutiva

Hoje em dia, os relógios moleculares conseguem lidar com vários tipos de *rates* de variação evolutiva. As *rates* variam de acordo com as partes do genoma (*site effects*), ao nível dos taxa (*lineage effects*), e ao longo do tempo (*epoch effects*). Os *site effects* ocorrem quando diferentes partes do genoma evoluem a *rates* distintos (*Figura 4-A*). Um grande exemplo é o facto de os genes de codificação das proteínas possuírem uma elevada *rate* de variação evolutiva na terceira posição dos codões. Isto devido à variação da primeira e segunda posição terem maior probabilidade de alterar o aminoácido codificado, causando grandes consequências na função das proteínas. Os *site effects* foram as primeiras informações a serem conhecidas e caracterizadas, durante a investigação genética, sobre *taxa* heterogéneos. [41]

As *lineage effects* ocorrem quando diferentes *taxa* apresentam diferentes *rates* de evolução molecular (*Figura 4-B*). Por exemplo, os roedores apresentam elevadas taxas de variação genética em comparação com outros mamíferos, em parte devido ao curto tempo de geração. O estudo dos *lineage effects* deu origem à metodologia dos *relaxed-clocks*, que visam estatisticamente modelar as *rates* de variação ao longo dos *branches* da árvore evolutiva. Com o uso de relógios biológicos, este método permite inferir uma escala de tempo evolutivo quando as *rates* variam ao longo das linhagens. [41]

Os *epoch effects* ocorrem quando as *rates* evolutivas diferem em períodos de tempo (*Figura 4-C*). Por exemplo, verificou-se que os rates evolutivos da gripe aumentaram acentuadamente por volta de 1990. De modo que, esta heterogenia temporal torna-se mais difícil de ser detetada e estudada do que os *site effects* ou as *lineage effects*. Isto porque ocorre a criação de padrões de divergência genética ao longo dos *taxa* que são muito semelhantes aos previstos quando os *taxa* permanecem constantes ao longo do tempo. [41]

Novas abordagens surgem quando dois ou mais tipos de heterogenias interagem entre si. Os *site effects* e as *lineage effects* interagem quando genes diferentes possuem padrões no *rate* de variabilidade distintos ao longo da *taxa* (*Figura 4*-D). Para lidarmos com estes padrões complexos de variação, podemos socorrer-nos dos *partitioned clock models*, na qual diferentes partes do genoma evoluem de acordo com diferentes tipos de relógios (*pacemakers)*. [41]

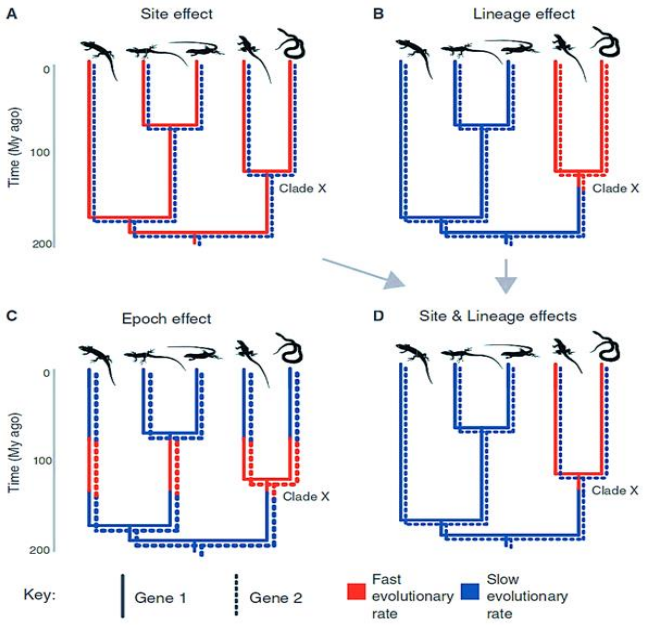


Figura 4 - Tipos de rates de variação genética ao longo da árvore evolutiva. | (A) Rate de variação ao longo dos sites. (B) Rate de variação ao longo das linhagens. (C) Rate de variação ao longo dos períodos de tempo, epochs. (D) Interação dos site effects e lineage effects. Fonte [41]

2.2.2 Calibração dos Relógios Moleculares

Mesmo quando analisado com os modelos mais sofisticados, o estudo das divergências genéticas apenas conseguem providenciar uma escala de tempo relativa. Desta forma, os relógios moleculares necessitam de ser calibrados de forma a inferir uma escala temporal absoluta, para que nos possamos referir ao intervalo de tempo em que uma divergência ocorre [42].

Tipicamente, as calibrações são efetuadas através de registos fósseis, como datas de acontecimentos geológicos (formação de ilhas, rifts continentais entre outros, exemplo anteriormente dado no *Anexo 1*), presumindo que a divergência das espécie possa ter sido afetada por estes eventos. Com precaução, poderão ser também usadas *rates* evolutivos estimados anteriormente, bem como tempos de divergência e dados biogeográficos [42].

2.2.3 Escalas de Tempo Evolutivas

Os relógios moleculares são essenciais para a reconstrução detalhada da escala de tempo, bem como a modelação dos *branches* da árvore evolutiva. Por sua vez, podem demonstrar o quão importante foi a história da Terra para a evolução dos seres vivos. Porém, o uso de relógios inapropriados ou calibrações incorretas podem conduzir a resultados ilusórios nos resultados estimados das escalas de tempo. Este problema tem criado debates sobre o tempo e fatores, dos momentos evolutivos mais importantes, incluindo a origem do Reino animal, a divergência entre as aves e os mamíferos, ou a origem das plantas com flor. Algumas das primeiras análises realizadas com relógios moleculares na divergência entre

o filo animal, concluíram que os metazoários terão divergido há cerca de mil milhões de anos atrás - quase o dobro da eclosão de animais fossilizados em rochas Cambrianas (compreendidas aproximadamente entre 542 milhões e 488 milhões de anos atrás) da era Paleozoica. A origem destes resultados deve-se à falha de se ter considerado os *lineage effects*: a variação genética geralmente ocorre mais lentamente em vertebrados do que em invertebrados, mas as primeiras análises moleculares inferiram uma baixa *rate* evolutiva para os vertebrados ao longo do tempo, levando os tempos de divergência animal a estarem associados ao Pré-Câmbrico. Posteriormente, foram efetuadas analises com modelos que possuem *rates* evolutivas melhoradas e aplicadas calibrações de forma minuciosa, que associavam os animais ao início do período Ediacariano, onde ocorreram grandes eventos de glaciação. No entanto, estes estudos continuam a proceder os fósseis dos metazoários por dezenas de milhões de anos. Em grupos com outros organismos, a melhoria dos métodos melhorou também a congruência entre as escalas de tempo inferidas a partir dados moleculares e registos fósseis. No entanto, a datação inferida é ainda substancialmente mais antiga do que a sugerida pelos registos fósseis. Estas discrepâncias entre as datas moleculares e a datação fóssil ainda necessitam de ser explicadas. [42]

2.2.4 Aplicação dos Relógios Moleculares através da Estatística Bayesiana

O uso da estatística Bayesiana em filogenética é relativamente recente, mas abre já horizontes no mundo científico, devido ao facto de permitir inferir árvores de incerteza para diferentes grupos taxonómicos presentes na mesma árvore evolutiva. O método Bayesiano está fortemente associado à ML. De modo que, a hipótese ideal é aquela que maximiza a *pp*, que é proporcional à *likelihood* multiplicada pela *prior probability*, para cada hipótese. *Prior probabilities* de diferentes hipóteses possibilitam inferir resultados antes dos dados serem analisados. Em vários métodos, os investigadores definem a distribuição da *prior probability* que acreditam ser a mais abrangente, para que a maioria das diferenças na *pp* seja atribuível a diferenças da *likelihood*. Uma das maneiras é aplicar um *prior* uniforme, com a mesma probabilidade a todos os parâmetros possíveis. Para permitir abordagens mais rápidas que a ML *bootstraping*, a inferência Bayesiana permite implementar modelos de sequenciação evolutiva complexos, como a estimativa de tempos de divergência, a deteção de resíduos importantes na seleção natural, bem como pontos de recombinação genética [35].

Como já abordado anteriormente, a ML não consegue lidar com modelos que possuem parâmetros muito complexos. Quando a relação entre os parâmetros e os dados é baixa, a inferência pode tornar-se incerta. Na inferência Bayesiana o resultado final é dado tendo em conta todos os parâmetros, ao contrário da ML. Isto porque existem grandes diferenças na forma de como o método Bayesiano e a ML abordam os parâmetros dos modelos evolutivos. Normalmente, a ML tem por base a metodologia *join estimation* que visa encontrar o ponto mais alto do *parameter landscape*. Enquanto que a estatística Bayesiana estima o volume entre um conjunto de *posterior-probabilities*; os parâmetros são integrados

entre si, de forma a obter a *marginal posterior probability* da árvore evolutiva. Além do uso desta metodologia o método Bayesiano usa o algoritmo MCMC (abordado anteriormente no ponto *2.1.3*) para lidar com modelos complexos [35].

2.2.4.1 A Estatística Bayesiana no cálculo dos Tempos de Divergência

Através do uso da metodologia MCMC, a estatística Bayesiana permite estimar os tempos de divergência entre as espécies através da calibração da *rate* de substituição, bem como os parâmetros dos modelos evolutivos, anteriormente falados: JC, HKY, GTR. Hoje em dia o *software* filogenético mais celebre que permite, através da modelação da estatística Bayesiana, inferir os tempos de divergência entre as espécies é o *BEAST*, *Bayesian Analysis Sampling Trees* [39].

O *BEAST* é um *software* de inferência Bayesiana, que utiliza o MCMC. É totalmente orientado para a inferência dos tempos de divergência de filogenias que têm por base uma *rooted tree* a partir do uso de relógios moleculares *strict* ou *relaxed*. Pode ser usado como um método de reconstrução filogenética, mas também como uma *framework* que visa testar hipóteses evolutivas. O *BEAST* possibilita ao utilizador fazer vários tipos de análises, tais como [43, 44]:

- Inferência de árvores evolutivas tendo por uso modelos que variam o *rate* de substituição (relógios moleculares com *rate* constante, *uncorrelated relaxed clocks*, *random local molecular clocks*);

- Estimar o tempo de divergência das espécies e calibrações fósseis, através de modelos de *branch-time* e métodos de calibração;

- Análise de sequências não-contemporâneas;

- Aplicação de modelos de substituição heterogéneos ao longo das partições;

- Análises populacionais, como a modelação de parâmetros demográficos (tamanho da população, crescimento/ declínio, migração), criação de *Bayesian skyline plots* e filogeografia;

- Inferência a árvores de genes e espécies tendo por base o uso da vertente *Star BEAST* (**BEAST*) [45];

Ambos as metodologias, *BEAST* e **BEAST*, estimam a topologia de árvores de espécies, tempos de divergência, tamanho da população entre uma amostra de genes sobre modelos de *multispecies coalescent*, porém existem várias diferenças na modelação. O *BEAST* requere um *outgroup*, o tamanho da população é assumido constante ao longo da *branch* e o *prior* para as árvores de espécies é uniforme. Computacionalmente infere cada gene da árvore individualmente, em duas etapas distintas. Em contraste, o **BEAST* infere a árvore

de espécies (*multi-individual*, *multi-locus method)*, bem como todos os genes num único processo com o uso do algoritmo MCMC e não necessita de um outgroup [42].

Atualmente existem duas versões disponíveis do software: o *BEAST v1.8* (*BEAST*) e *BEAST v2.5.2* (*BEAST2*) [46-49]. O *BEAST2* é uma versão completamente rescrita do *BEAST* que permite implementar novos modelos de distribuição e métodos através do uso de *plugins*. Como por exemplo a SNAPP (*phylogenetic analysis using SNP and AFLP data*) e BDSSM (*birth-death skyline model for serially-sampled data*). A preparação dos ficheiros de *input*, a modelação dos ficheiros de *output*, bem como a visualização é realizada por uma serie de programas disponibilizados pelo *package do software BEAST*, como o *BEAUti (Bayesian Evolutionary Analysis Utility)* [49]: um programa com uma *graphical user interface* (GUI) que visa modelar os ficheiros de *input* para o *BEAST* e **BEAST* em formato *eXtensible Markup Language* (XML); *LogCombiner* [49] que permite combinar os ficheiros de *log* e de *trees* a partir de múltiplas análises independentes; *TreeAnnotator* [49] *que sumariza a informação de um conjunto de árvores numa só*; *Tracer* [50] *que permite analisar e visualizar a MCMC* descrita pelo ficheiro de *log,* e *Figtree* [51] que permite visualizar, sumarizar e anotar árvores filogenéticas*.* [45]

Vários fatores podem influenciar a *rate* de substituição numa população, tais como a *rate* de mutação, o tamanho da população, o tempo de geração e a seleção. Como resultado, vários modelos foram desenvolvidos para dar resposta de como o *rate* de substituição varia ao longo da árvore da vida. Muitos destes modelos são aplicados através de *priors* que têm por base os métodos de inferência Bayesiana. As implementações dos métodos de datação fornecem uma forma flexível de modelar o *rate* de variação e obter tempos de divergência com confiança, conferindo se os modelos são adequados. Quando usados com métodos numéricos como a MCMC, para a aproximação da distribuição de parâmetros da *pp*, os métodos Bayesianos demonstram ser extremamente poderosos na inferência dos parâmetros dos modelos estatísticos [47]; o *Anexo 2* demonstra um *workflow* de um processo MCMC.

Várias componentes são aplicadas durante o cálculo dos tempos de divergência com uso dos métodos Bayesianos. Uma delas é o *prior* que define a datação dos nodos, também chamado de *tree prior*. Este descreve como os eventos que deram origem à evolução das espécies estão distribuídos ao longo do tempo. Quando este modelo é combinado com o modelo que calcula a *rate* das *branches*, a inferência Bayesiana permite estimar tempos de divergência relativos [47].

No *BEAST*, os *priors* disponíveis para este tipo de abordagem, cálculo dos tempos de divergência inter-espécies, são variantes do *birth-death prior*, que incluem o *calibrated Yule model*, o modelo *birth-death* com amostras incompletas das espécies, bem como *serially-sampled birth-death processes* [47].

3 Metodologia

As atividades desenvolvidas ao longo do projeto tiveram todas como ponto de partida a seleção das sequências a serem analisadas, a criação de datasets com as amostras, bem como a análise da qualidade das sequências, manipulação e estruturação nucleotídica após o processo de alinhamento.

A aquisição das amostras necessárias foi realizada através de várias vias. O *dataset* das espécies de cigarras dos géneros *Maoricicada*, *Rhodopsalta* e *Kikihia* (MRK) foi facultado pelos autores da publicação; As sequências dos lagartos ocelados foram adquiridas a partir do portal *National Center for Biotechnology Information* (NCBI), através da criação de um *script (Apêndice1,* repositório *online*: https://github.com/ray2g/StarBeast/blob/master/efetch.sh) com base na API (application programming interface*) -eutils* (que simplifica a pesquisa, aquisição, e análise de registos do NCBI) do sistema *Entrez* (*Global Query Cross-Database Search System*) desenvolvido pela plataforma da NCBI: o comando *efetch* (usado para fazer o *download* de amostras a partir da base de dados NCBI) com base nos *accession numbers* facultados pelas respetivas publicações. O *script* foi executado na *zsh* (*v5.7.1*) do sistema operativo *Ubuntu v18.04*, criando um ficheiro em formato *FASTA* na diretoria desejada com todas as sequências selecionadas; A informação genética usada para a criação dos *datasets* das borboletas do género *Lycaena* e *Melanargia* foi facultada pelos investigadores envolvidos no projeto.

As sequências em formato *FASTA* são em primeiro lugar alinhadas com uso do *software MAFFT v7.419* [52], com os parâmetros *standard (--auto).* Após o alinhamento, o *dataset* é analisado com o programa *AliView V1.26* [53] de forma a inferir a qualidade dos alinhamentos. Em alguns casos foi necessário cortar (*trimming*) as extremidade das sequências para que tivessem igual número de pares de bases.

Por último o ficheiro *FASTA* é convertido em formato *Nexus* para ser usado como *input* nos seguintes passos. O primeiro passo será com recurso ao *software jModelTest2*, estimar para cada *dataset,* qual o modelo de substituição que melhor se adequa a cada *dataset*.

3.1 Estudo do enquadramento do *software BEAST* na inferência filogenética molecular

Como ponto de partida do projeto, foram efetuadas as leituras dos documentos de apoio disponibilizadas *online* pelo software *BEAST*, quer da versão 1 e 2 para inferência *inter* e *intraespecífica*, bem como a realização de tutoriais. Os materiais estudados de maior relevância, estão apresentados no *Anexo 3*.

Esta primeira atividade foi fulcral para entender o enquadramento do programa na inferência filogenética, tendo por via os relógios moleculares, e interiorizar todos os passos

envolvidos na análise. Foram ainda pesquisadas e estudadas as publicações mais adequadas ao projeto, apresentadas no *Anexo 4* [54-59].

3.2 Reprodução do *paper* referente à inferência molecular das espécies da tribo *Cicadettini*, e descrição da *pipeline* comum do *software BEAST*

Neste ponto, foi reproduzida a análise "E" da publicação "*Inflation of Molecular Clock Rates and Dates: Molecular Phylogenetics, Biogeography, and Diversification of a Global Cicada Radiation from Australasia (Hemiptera: Cicadidae: Cicadettini)*" [60]. O *dataset* foi disponibilizado pelo autor a partir de um repositório, composto por amostras do gene *Cytochrome c Oxidase Subunidade I (COI)* das espécies de cigarras dos géneros *Maoricicada*, *Rhodopsalta* e *Kikihia* (MRK) habitantes da Nova Zelândia.

O artigo retrata a inferência intraespecífica com recurso à metodologia **BEAST2,* contudo foi realizada uma análise interespecífica adicional, de modo a comparar os resultados das duas abordagens. Para a inferência filogenética molecular com o software *BEAST2*, as sequências em formato *Nexus* foram carregadas no *software BEAUti* de modo a calibrar os parâmetros da MCMC e criar o ficheiro de controlo (ficheiro em formato XML que irá ser interpretado pelo *software*). Os parâmetros foram aplicados com base nas informações dadas pelos autores (*Tabela 2)*. Após a criação do ficheiro XML, foi executado o *software BEAST2* e carregado o ficheiro. Quando concluído o processo da MCMC, obteve-se um ficheiro de *log* e um ficheiro em formato *trees*.

Tabela 2 - Parâmetros usados na calibração do gene mitocondrial COI das espécies de cigarras do grupo MRK, formatados para BEAST2.

Parâmetros / Genes	*Site Models*	*Clock Model*	*Priors*	MCMC
COI	Modelo de Substituição: GTR+I+G; • *Rate* de Substituição = 1.0 (não estimado) • *Gamma Category Count* = 4 • *Shape* = 0.5 (estimado) • *Proportion Invariante* = 0.5	*Uncorrelated relaxed clock lognormal,* com *rate* de 0.0112 (não estimado): • M= 0.01172 • S = 0.288	*Birth Death Tree Model*: • *Mean grow rate* = 0.01 • *Relative death rate* = 0.5 • *Population size* = 100	10M de iterações guardadas a cada 1000; Os ficheiros de *log* e de *trees* são registados a cada 700 iterações.

Para a inferência intraespecífica, **BEAST2,* foi criado um novo ficheiro XML com a calibração apresentada na *Tabela 3, no software BEAUti*. Apesar de uma diferente formatação o ficheiro de controlo da é igualmente processado pela metodologia *BEAST2.*

*Tabela 3 - Parâmetros usados na calibração do gene mitocondrial COI das espécies de cigarras do grupo MRK, formatados para *BEAST2.*

Parâmetros / Genes	*Site Models*	*Clock Model*	*Priors*	MCMC
COI	Modelo de Substituição GTR+I+G: • *Rate* de Substituição = 1.0 (não estimado) • *Gamma Category Count* = 4 • *Shape* = 0.5 (estimado com distribuição Dirichlet) • *Proportion Invariante* (distribuição Beta) = 0.5	*Uncorrelated relaxed clock lognormal*, com *rate* de 0.0112 (não estimado): • M = 0.01172 • S = 0.288, com distribuição exponencial	*Birth Death Tree Model*: • *Mean grow rate* = 0.01 • *Relative death rate*= 0.5 (com distribuição uniforme) • *Population size* = 100	10M de iterações guardadas a cada 1000; Os ficheiros de *log* e de *trees* são registados a cada 700 iterações.

De modo a avaliar a qualidade da inferência, foi carregado o ficheiro de *log* no programa *Tracer* e verificados os valores de ESS (*Effective Sample Size*) dos parâmetros (exemplo apresentado nos *Anexos 5 e 6*), que testam as tentativas independentes da distribuição da *posterior,* na qual a cadeia de *Markov* é equivalente. Caso os parâmetros tivessem convergido corretamente (valores acima de 200), eram efetuadas mais uma série de *runs* independentes (com *seeds* diferentes), e consequentemente os ficheiros de *log* eram combinados a partir do programa LogCombiner, bem como os ficheiros em formato trees; caso contrário, os parâmetros seriam novamente calibrados *no BEAUti*. Devido à inferência ser suportada por métodos estocásticos, por vezes a distribuição acaba por não convergir corretamente. A combinação de *logs* permite garantir um valor de suporte às retas de distribuição dos *priors* ao longo das gerações do MCMC, levando a completarem-se entre si. Neste caso, foram combinados 8 ficheiros de *log* e de *trees*, a partir de 8 *runs* independentes, após várias calibrações. Seguidamente, os ficheiros de *trees* foram combinados com 10% de *burn-in* e anotados com o programa *TreeAnnotator*, de modo a reportar a *Maximum Clade Credibility Tree*, com os nodos calibrados pela *Mean Height*. De modo a não se perder informação, não foi aplicado novo *burn-in* à anotação devido ao facto de este já ter sido aplicado anteriormente na combinação. Quando terminado o processo de anotação, com uma grande necessidade de tempo de processamento, o *output* (a árvore filogenética final) foi analisado e editado a partir do software *FigTree*.

Esta metodologia retrata o *workflow* comum à inferência filogenética molecular com as metodologias *BEAST2* e **BEAST2*, denominado como *StruggleBeast* (*Figura 5*). As seguintes análises partilharam do mesmo método retratado. A análise de resultados desta inferência não faz parte dos objetivos do projeto, uma vez que este apenas serve como *dataset* de treino por abordar a vertente **BEAST2*.

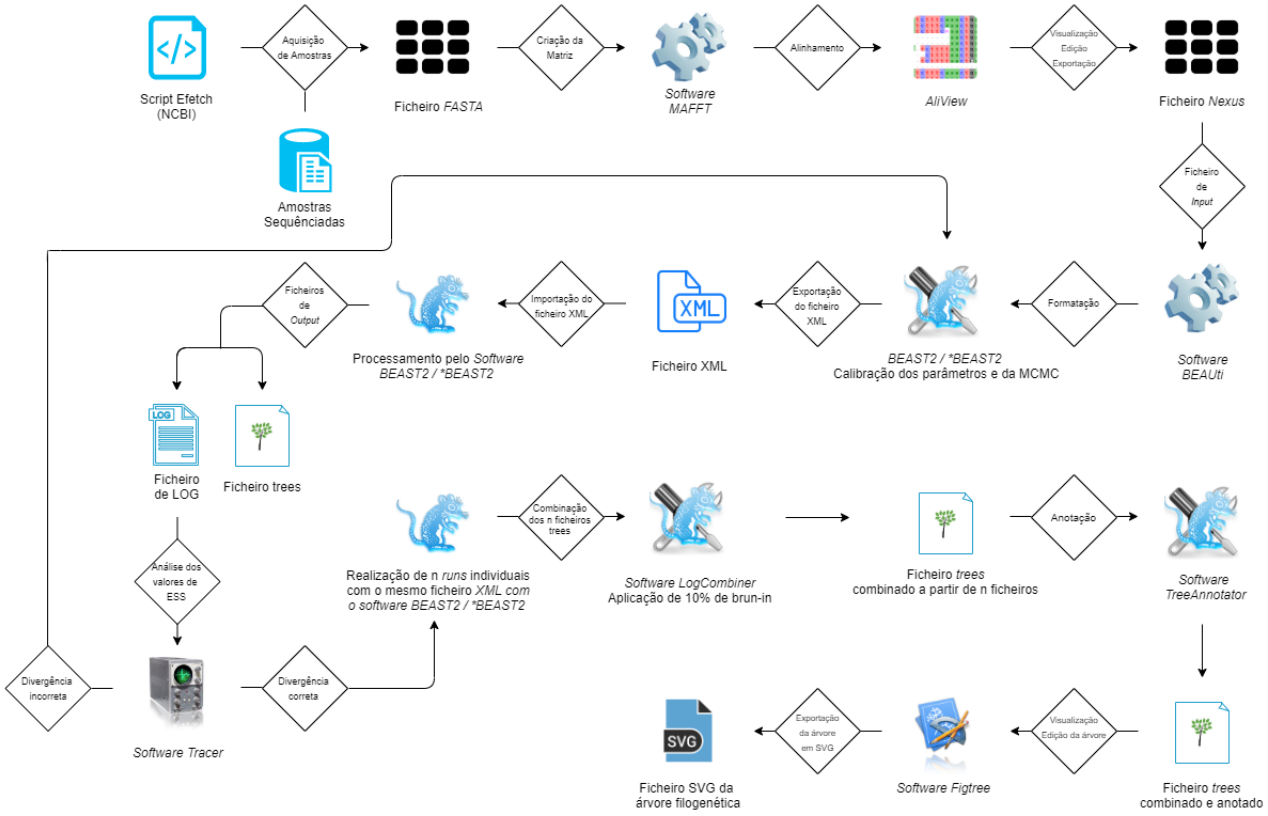


Figura 5 - Workflow da metodologia aplicada, StruggleBeast

3.3 Reprodução do *paper* relativo à inferência filogenética molecular das espécies de lagartos ocelados

Neste terceiro ponto foram replicados os resultados da publicação “The role of vicariance vs. dispersal in shaping genetic patterns in ocellated lizard species in the western Mediterranean” [54]. Com os *accession numbers* do *GenBank* disponibilizados pelo autor, foram descarregadas as sequências do gene mitocondrial *Cytochrome b* (*Cyt b)* em formato *FASTA* pelo método *efetch*, anteriormente descrito, e manipuladas com o *software AliView* após o alinhamento com o software *MAFFT*, seguidamente exportadas em formato *Nexus*.

Procedeu-se à aplicação da metodologia *StruggleBeast*. O dataset em formato *Nexus* do *Cytochrome b* foi carregado no software *BEAUti* de modo a calibrar os parâmetros da MCMC, e criar o ficheiro XML para ser processado pelo *BEAST2*. Os parâmetros usados na calibração da inferência interespecífica estão designados na *Tabela 4*.

Tabela 4 - Parâmetros usados na calibração do gene mitocondrial Cytochrome b das espécies de lagartos ocelados, para inferência interespecífica com o método BEAST2.

Parâmetros / Genes	*Site Models*	*Clock Model*	*Priors*	MCMC
Cyt b	Modelo de Substituição TPM2uf+I+G: • *Rate* de Substituição = 1.0 (não estimado) • *Gamma Category Count* = 4 • *Shape* = 1.0 (estimado) • *Proportion Invariante* = 0.5	*Strict clock* com *rate* de 0.0125 (não estimado).	*Coalescent Exponential Population Tree Model*: • *Population size* = 6.9: lower = 0.0, upper = 6900 • *Grow rate* = 0.0: lower = -12500, upper = 12500	10M de iterações com guardadas a cada 100; Os ficheiros de *log* e de *trees* são registados a cada 1000 iterações.

Quando finalizado o processamento, o ficheiro de *log* resultante foi analisado no *software* Tracer, até encontrarmos os *parâmetros* adequados, de acordo com os valores da distribuição de ESS. Após várias tentativas de calibração, procedeu-se à execução de sete novas análises independentes (com diferentes *seeds*) a partir do mesmo ficheiro XML para serem combinadas num único ficheiro a partir do *software LogCombiner*, bem como o ficheiro de *trees* combinado com a informação das 8 runs com 10% de *burn-in*. Procedeu-se à anotação com o *software TreeAnnotater* sem *burn-in e* os nodos calibrados pela *Mean Height*, prosseguindo-se a análise com a visualização da árvore com a maior credibilidade no *Figtree*.

Foi seguidamente feita a análise intraespecífica do mesmo *dataset*, tendo como metodologia a *StruggleBeast*, mas formatado no *BEAUti* para **BEAST2*, e calibrado com os parâmetros apresentados na *Tabela 5*.

*Tabela 5 - Parâmetros usados na calibração do gene mitocondrial Cytochrome b das espécies de lagartos ocelados, para inferência intraespecífica com o método *BEAST2.*

Parâmetros / Genes	*Site Models*	*Clock Model*	*Priors*	MCMC
Cyt b	Modelo de Substituição HKY+I: • *Kappa* = 7.52 (estimado) • *Rate* de Substituição = 1.0 (não estimado) • *Gamma Category Count* = 0 • *Proportion Invariante* = 0.572 (não estimado)	*Strict clock* com *rate* de 0.0125 (não estimado).	*Species Tree Population Size:* • *Pop Function = constant* • *Population Mean* = 8.7 (estimado) • *Ploidy = Y or mitocondrial* *Birth Death Tree Model* com os parâmetros *default*	10M de iterações guardadas a cada 100; Os ficheiros de *log* e de *trees* são registados a cada 1000 iterações.

3.4 Inferência filogenética molecular de borboletas dos géneros *Lycaena* e *Melanargia*

Esta tarefa, foi enquadrada num projeto de filogenia a ser desenvolvido conjuntamente com dois investigadores do grupo CoBiG[2]: Eduardo Marabuto e Miguel Nunes. O projeto consistiu em calcular os tempos de divergência para as borboletas dos géneros *Lycaena* e *Melanargia* com apenas a vertente interespecifica do software *BEAST2,* tendo em conta a sistemática biológica. A informação genética foi em parte disponibilizada pelos investigadores, através de amostragem e sequenciação, e outra parte obtida através do portal NCBI. Cada gene foi calibrado individualmente, e só depois foi realizada a calibração concatenada.

3.4.1 Inferência filogenética molecular de borboletas do género *Lycaena*

Primeiramente, procedeu-se à criação dos *datasets*. Neste procedimento foram utilizados dois genes; o gene mitocondrial *Cytochrome c Oxidase subunidade I* (*COI*) e o gene nuclear *Elongation Factor 1-alpha* (*EF-1α)*. Após as matrizes devidamente criadas e convertidas no formato *Nexus*, procedeu-se à metodologia *StruggleBeast*. A calibração foi realizada de acordo com os parâmetros apresentados na *Tabela 6*. As árvores com a calibração individual dos genes encontram-se presentes nos *Anexos 12* (*COI*) e *13* (*EF-1α)*.

Tabela 6 - Parâmetros usados na calibração do dataset concatenado composto pelos gene mitocondrial COI e nuclear EF-1α das borboletas do género Lycaena, formatados para BEAST2.

Parâmetros / Genes	*Site Models*	*Clock Model*	*Priors*	MCMC
COI	Modelo de Substituição JC69+I+G: • *Rate* de Substituição = 1.0 (não estimado) • *Gamma Category Count* = 4 • *Shape* = 0.76 (estimado) • *Proportion Invariante* = 0.54	Modelo *Relaxed Clock Log Normal*, com *rate* de 0.0115, (não estimado).	Partições dos genes *linked* para a inferência da árvore, de modo a partilharem a mesma topologia e *branch times*; *Birth Death Tree Model* com os parâmetros *default*; *Outgroup* calibrado com uma distribuição uniforme: [56.0, 63.0 Ma], com monofilia.	20M de iterações guardadas a cada 1000; Os ficheiros de *log* e de *trees* são registados a cada 100 iterações.
EF-1α	Modelo de Substituição TIM2+I+G: • *Rate* de Substituição = 1.0 (não estimado) • *Gamma Category Count* = 4 • *Shape* = 1.21 (estimado) • *Proportion Invariante* = 0.57	Modelo *Relaxed Clock Log Normal* com *rate* de 0.001277 (não estimado).		

3.4.2 Inferência filogenética molecular de borboletas do género *Melanargia*

Para a concretização desta tarefa, foi também usada a metodologia *StruggleBeast*. Foram criados quatro *datasets* para cada um dos genes usados na inferência ao género *Melanargia* (subgénero *Argeformia*): dois genes mitocondriais, o *COI* e a componente *16S* do RNA ribossómico, e dois genes nucleares: *Elongation Factor 1-alpha (EF-1α)* e *Wingless* (*Wg*). Os *datasets* dos genes *16S* e *Wg* são complementares à análise uma vez que não se encontram completos para todos os *taxa*, mas que ao serem concatenados garantem um nível de suporte adicional. Os parâmetros usados na calibração dos genes estão descritos na *Tabela 7*. As árvores com a calibração individual dos genes encontram-se presentes nos *Anexos 14* a *17.*

Tabela 7 - Parâmetros usados na calibração do dataset concatenado composto pelos gene mitocondrial COI e nuclear EF-1α das borboletas do género Lycaena, formatados para BEAST2.

Parâmetros / Genes	*Site Models*	*Clock Model*	*Priors*	MCMC
16S	Modelo de Substituição TPM2uf: • *Rate* de Substituição = 1.0 (não estimado)	Modelo *Relaxed Clock Log Normal* com *rate* de 0.0086 (não estimado).	Partições dos genes *linked* para a inferência da árvore, de modo a partilharem a mesma topologia e *branch times*; *Yule Tree Model* com os parâmetros *default*; *Outgroup* calibrado com uma distribuição uniforme: [32, 36 Ma], com monofilia.	20M de iterações guardadas a cada 1000; Os ficheiros de *log* e de *trees* são registados a cada 100 iterações.
COI	• Modelo de Substituição JC69+G; • *Rate* de Substituição = 1.0 (não estimado) • *Gamma Category Count* = 4 • *Shape* = 0.16 (estimado)	Modelo *Relaxed Clock Log Normal* com *rate* de 0.0115 (não estimado).		
EF-1α	• Modelo de Substituição TN93+G: • *Kappa* 1 e 2 = 2.0 (estimado) • *Rate* de Substituição = 1.0 (não estimado) • *Gamma Category Count* = 4 • *Shape* = 0.3 (estimado)	Modelo *Relaxed Clock Log Normal* com *rate* de 0.001277 (não estimado).		
Wg	Modelo de Substituição HKY+G: • *Kappa* = 2.0 (estimado) • *Rate* de Substituição = 1.0 (não estimado) • *Gamma Category Count* = 4 • *Shape* = 0,19 (estimado)	Modelo *Relaxed Clock Log Normal* com *rate* de 0.007044 (não estimado).		

3.5 Inferência de lagartos ocelados – dataset concatenado

De modo a complementar a inferência realizada no *ponto 3.3*, realizou-se uma nova inferência com quatro genes adicionais de lagartos ocelados. Foi efetuada a análise a três genes mitocondriais, duas componentes do RNA ribóssomico (*12S* e *16S)* e *Cytochrome b (Cyt b)*, e a dois genes nucleares: *Fibrinogen beta chain* (*ß Fibrinogen*) e *Oocyte maturation factor* (*C-mos)*. À semelhança da abordagem realizada só com o *Cytochrome b*, as amostras foram adquiridas com base no *script efetch (Apêndice 1)*, pelos *accession numbers* do *GenBank* providenciados pelo autor. Após analisada a qualidade das sequências com recurso ao *software AliView* e convertidas para formato *Nexus*, recorreu-se ao *StruggleBeast* de modo a inferir a *Maximum clade credibility tree* com base nos parâmetros apresentados na *Tabela 8* para a análise interespecífica com o software *BEAST2*, e a *Tabela 9* para a análise intraespecífica com recurso à dependência **BEAST2*, de modo a estimar os tempos de divergência relativos das espécies de lagartos ocelados. Os genes foram calibrados individualmente e de forma concatenada. Os *Anexos de 7 a 11* apresentam as árvores com calibração individual dos genes.

Foi ainda criado um *script* em *Python 3 (Apêndice 2,* repositório *online:* https://github.com/ray2g/StarBeast/blob/master/namefastaseq_changer_2gencode.py) de modo a substituir o nome das amostras por um código incremental de acordo com o número de sequências do *dataset*, como por exemplo "*S001")* de modo a facilitar o manuseamento das sequências. Pois para diferentes genes, a associação entre amostras é feita pelo nome, de modo que estes têm de ser iguais entre as diferentes matrizes. Após a concatenação com 10% de *burn-in*, o ficheiro em formato *trees* anotado é editado, de modo a substituir os códigos pelos nomes de origem dos *taxa*. A árvore com a maior credibilidade é posteriormente analisada e editada com recurso ao *software FigTree*.

Tabela 8 - Parâmetros usados na calibração do dataset concatenado dos lagartos ocelados composto pelos genes mitocondriais 12S, 16S e Cyt b, e nucleares β Fibrinogen e C-mos; formatados para BEAST2.

Parâmetros / Genes	*Site Models*	*Clock Model*	*Priors*	MCMC
12S	Modelo de Substituição TIM2+I: • *Kappa1* e *Kappa2* =2.0 • *Rate* de Substituição = 1.0 (não estimado) • *Proportion Invariante* = 0.77	Modelo *Strict Clock* com *rate* de 0.002837 (não estimado).	Partições dos genes *linked* para a inferência da árvore, de modo a partilharem a mesma topologia e *branch times*; *Coalescent Exponential Population Tree Model*, com os parâmetros *default.*	20M de iterações guardadas a cada 1000. Os ficheiros de log e de trees são registados a cada 200 iterações.
16S	• Modelo de Substituição TIM2+I: • Kappa1 e Kappa2 =2.0 • *Rate* de Substituição = 1.0 (não estimado) • *Gamma Category Count*= 4 • *Shape* = 0.14 (estimado)	Modelo *Strict Clock* com *rate* de 0.003518 (não estimado).		
Cyt b	Modelo de Substituição HKY+I: • Kappa1 = 9.97 • *Rate* de Substituição = 1.0 (não estimado) • *Proportion Invariante* =0.61 (estimado)	Modelo *Strict Clock* com *rate* de 0.0125 (não estimado).		
β Fibrinogen	Modelo de Substituição HKY+I+G: • Kappa1 =4.43 • *Rate* de Substituição = 1.0 (não estimado) • *Gamma Category Count* = 4 • *Shape* = 0.42 (estimado)	Modelo *Strict Clock* com *rate* de 0.001156 (não estimado).		
C-mos	Modelo de Substituição HKY+I: • Kappa1 = 19.28 • *Rate* de Substituição = 1.0 (não estimado) • *Proportion Invariante* =0.93 (estimado);	Modelo *Strict Clock* com *rate* de 0.0004678 (não estimado).		

*Tabela 9 - Parâmetros usados na calibração do dataset concatenado dos lagartos ocelados composto pelos genes mitocondriais 12S, 16S e Cyt b, e nucleares: β Fibrinogen e C-mos; formatados para *BEAST2.*

Parâmetros / Genes	*Site Models*	*Clock Model*	*Priors*	MCMC
12S	Modelo de Substituição TIM2+I: • *Kappa1* e *Kappa2* = 2.0 • *Rate* de Substituição = 1.0 (não estimado) • *Proportion Invariante* = 0.77	Modelo *Strict Clock* com *rate* de 0.002837 (não estimado).	Partições dos genes *linked* para a inferência da árvore, de modo a partilharem a mesma topologia e *branch times*; *Species Tree Population Size:* • *Pop Function Constante* • *Population Mean* = 2.892; • *Ploidy = Y or mitochondrial*; Yule Tree Model, com os parâmetros *default*.	20M de iterações guardadas a cada 1000. Os ficheiros de *log* e de *trees* são registados a cada 200 iterações.
16S	Modelo de Substituição TIM2+I: • *Kappa1* e *Kappa2* =2.0 • *Rate* de Substituição = 1.0 (não estimado) • *Gamma Category Count* = 4 • *Shape* = 0.14 (estimado)	Modelo *Strict Clock* com *rate* de 0.003518 (não estimado).		
Cyt b	Modelo de Substituição HKY+I: • *Kappa1* = 9.97 • *Rate* de Substituição = 1.0 (não estimado) • *Proportion Invariante* = 0.61 (estimado)	Modelo *Strict Clock* com *rate* de 0.0125 (não estimado).		
β Fibrinogen	Modelo de Substituição HKY+G: • *Kappa1* = 4.43 • *Rate* de Substituição = 1.0 (não estimado) • *Gamma Category Count* = 4 • *Shape* = 0.42 (estimado)	Modelo *Strict Clock* com *rate* de 0.001156 (não estimado).		
C-mos	Modelo de Substituição HKY+I: • *Kappa1* = 19.28 • *Rate* de Substituição = 1.0 (não estimado) • *Proportion Invariante* = 0.93 (estimado)	Modelo *Strict Clock* com *rate* de 0.0004678 (não estimado).		

4 Descrição de resultados

4.1 Espécies de cigarras dos géneros *Maoricicada, Rhodopsalta, Kikihia*

Nas *Figuras 6* e *7* estão apresentadas as árvores filogenéticas das cigarras do género *Maoricicada, Rhadopsalta* e *Kikihia* provenientes da Nova Zelândia, inferidas através das abordagens *BEAST* e **BEAST* respetivamente, resultante da metodologia apresentada no *ponto 3.2* e calibradas com base nos parâmetros da *Tabela 2* e *3*.

A análise topológica de ambas as figuras, demonstra que as inferências possuem tempos de divergência muito recentes contrariamente ao esperado (tendo por referência os tempos de divergência apresentados na publicação do autor), onde não se recorre ao uso de um *outgroup* para calibrar a *root* da árvore filogenética (uma vez que o *dataset* foi disponibilizado pelo autor e não incluí *taxa outgroup)*. Deste modo, os resultados corretos são os apresentados no artigo do autor, uma vez que a inferência realizada neste projeto é irreal.

As árvores filogenéticas resultantes de ambas as inferências realizadas neste projeto são compostas por três grandes grupos. Os géneros *Maoricicada* (a primeira separação que se presume ter surgido, correspondente à *clade I*), *Rhadopsalta* (*clade II*) e *Kikihia* (*clade III)*. Os resultados obtidos pelo método *BEAST2*, sugerem que as espécies do género *Kikihia* terão divergido dos restante *taxa* no intervalo [0.372, 0.116 Ma], com um bom suporte de agrupamento no valor de *pp*=0,973. A separação dos géneros *Maoricicada* e *Rhadopsalta* terá ocorrido no intervalo [0.311, 0.094 Ma], suportada pela *pp*=0,998 e *pp*=1 respetivamente. Já o método **BEAST2* supõe a divergência do género *Kikihia* da restante taxa no intervalo [0.311, 0.094 Ma] com *pp*=0,087, e a separação dos géneros *Maoricicada* e *Rhadopsalta* no intervalo [0.21, 0.056 Ma], com *pp*=0,995 e *pp*=0,997 respetivamente.

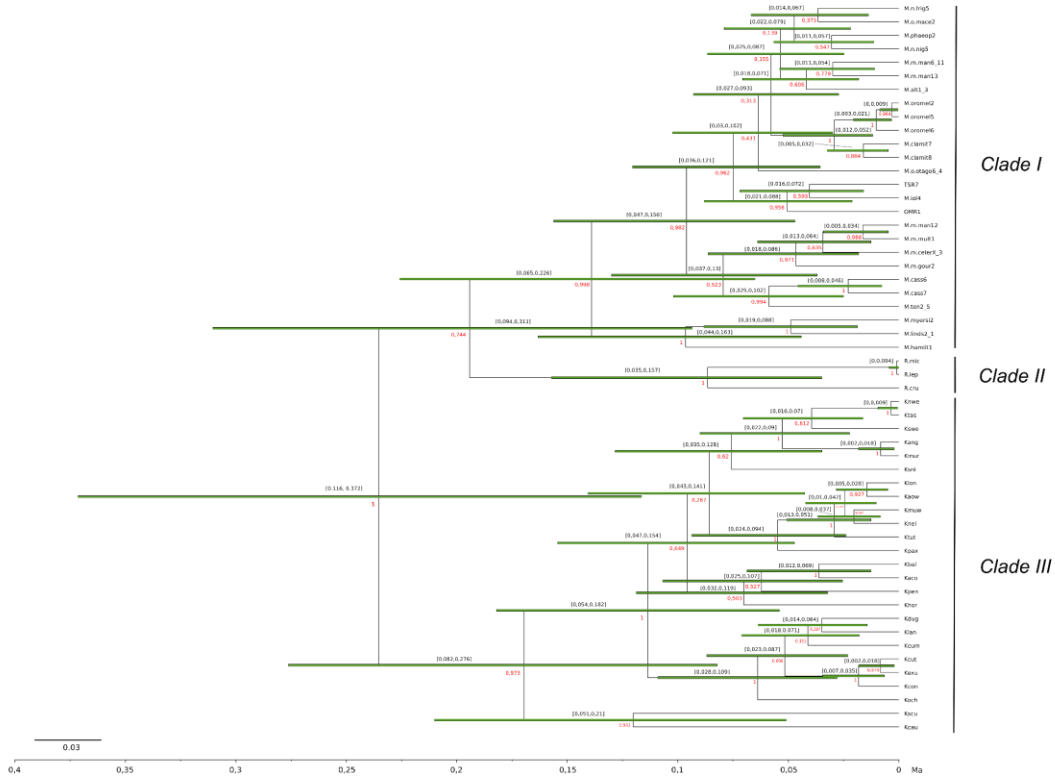


Figura 6 - Árvore filogenética resultante da inferência interespecífica (BEAST2) das espécies de cigarras do grupo MRK, calibrada com os parâmetros da Tabela 2; dataset composto pelo gene mitocondrial COI, inferido pela combinação de 8 runs independentes. Imagem com melhor resolução: https://raw.githubusercontent.com/ray2g/BEAST2-StarBEAST2/master/final_phylogenetic_trees/MRK_beast.png.

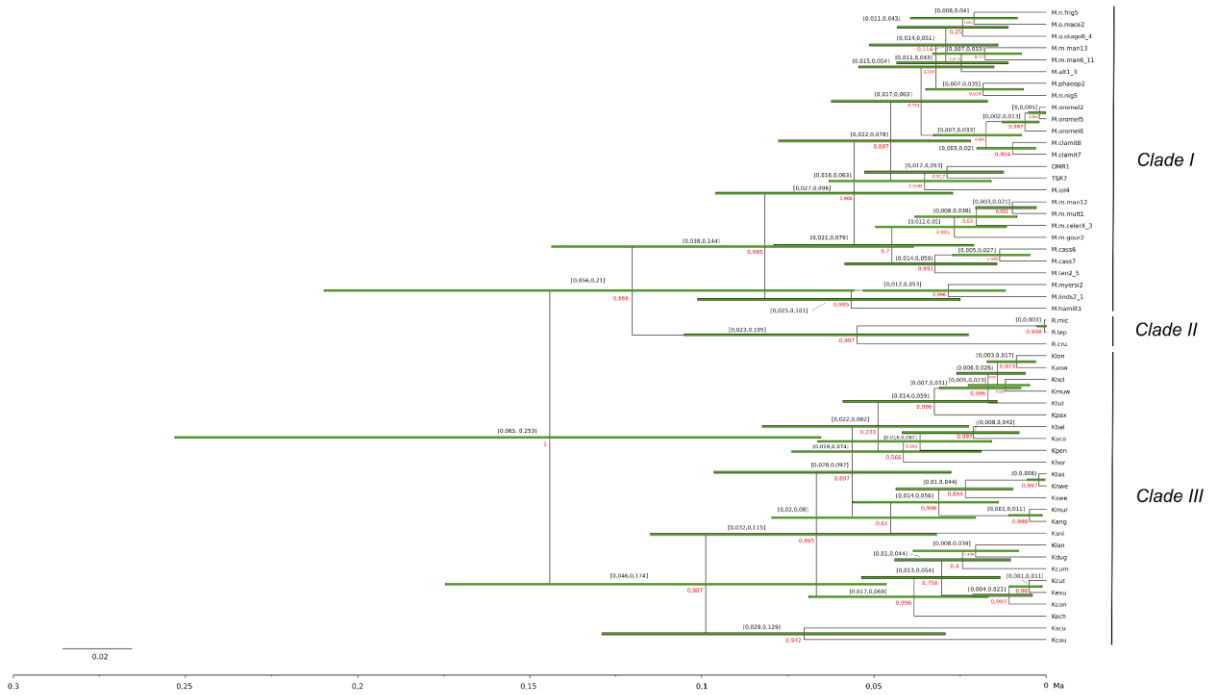


*Figura 7 - Árvore filogenética resultante da inferência intraespecífica (*BEAST2) das espécies de cigarras do grupo MRK, calibrada com os parâmetros da Tabela 3; dataset composto pelo gene mitocondrial COI, inferido pela combinação de 8 runs independentes. Imagem com melhor resolução: https://raw.githubusercontent.com/ray2g/BEAST2-StarBEAST2/master/final_phylogenetic_trees/MRK_starbeast.png.*

4.2 Lagartos ocelados – *dataset* constituído pelo gene *Cyt b*

4.2.1 Inferência interespecífica, *BEAST2*

Na *Figura 8* está apresentada a árvore filogenética resultante da inferência interespecífica ao gene *Cyt b* das espécies de lagartos ocelados, resultante da metodologia descrita no ponto 3.3. Esta foi calibrada de acordo com os parâmetros apresentados na *Tabela 4* e obtida a partir da combinação de 8 *runs* independentes.

Topologicamente, o *outgroup* é constituído pelas espécies *Lacerta agilis* e *Lacerta schreiberi*, e o *ingroup* pelo género *Timon* propriamente dito. O *ingroup* revelou ser constituído por dois grandes *clades*: o *clade I* constituído pelas espécies *Timon lepidus* e *Timon nevadensis* provenientes da Península Ibérica, e o *clade II* constituído pelas espécies *Timon tangitanus* e *Timon pater*, do Norte de África. A separação entre o *outgroup* e o *ingroup* poderá ter ocorrido no intervalo [26.384, 15.536 Ma], durante o período terciário da era cenozoica. Os *clades I* e *II* poderão ter divergido entre si no intervalo [11.132, 7.339 Ma], bem suportados (*pp*=1). No *clade I*, as espécies *Timon lepidus* (Portugal e Espanha) e *Timon nevadensis* (Serra Nevada) terão sofrido durante o intervalo [8.782, 5.034 Ma] segregação geográfica, separando-se em dois *clades* distintos bem suportados (*pp*=1 para ambas os *clades*). Já na *clade II*, poderá ter ocorrido também agrupamento das espécies *Timon tangitanus* e *Timon pater* em dois *clades* distintos no intervalo [9.644, 5.934 Ma], bem suportado (*pp*=1 para ambos os *clades*). Porém, com um baixo suporte no nodo ancião com *pp*=0,512.

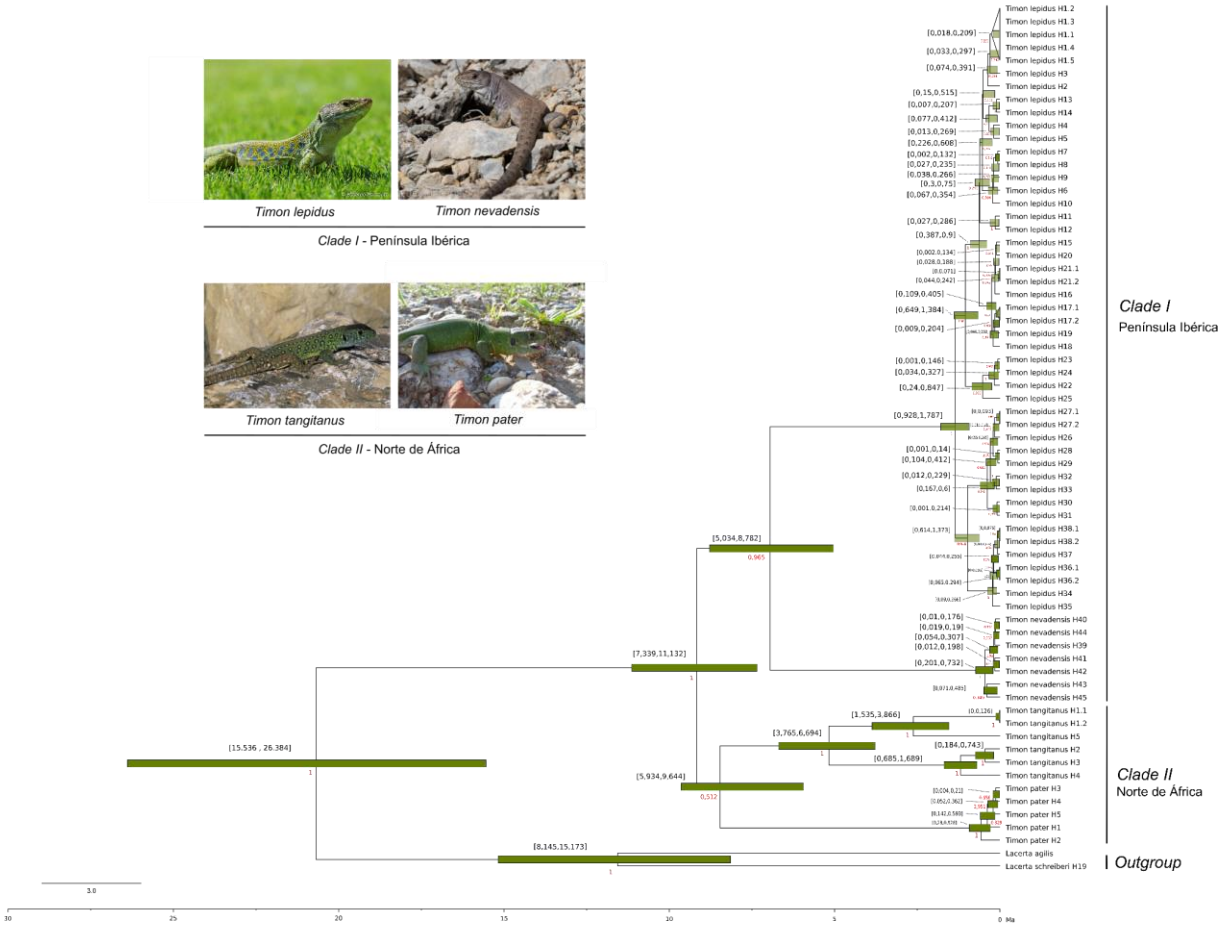


Figura 6 - Árvore filogenética resultante da inferência interespecífica (BEAST2) ao dataset de lagartos ocelados composto pelo gene mitocondrial COI, calibrado com os parâmetros da Tabela 4 a partir da combinação de 8 runs independentes. Fonte das fotografias: Timon lepidus © Eduardo Marabuto, Timon nevadensis © EUROLIZARDS, íNaturalist (Timon tangitanus © Roberto Sindaco, Timon pater © Karim Chouchane). Imagem com melhor resolução: https://raw.githubusercontent.com/ray2g/BEAST2-StarBEAST2/master/final_phylogenetic_trees/lizards_cyt_beast.png.

4.2.2 Inferência intraespecífica, **BEAST2*

Na *Figura 9*, está representada a árvore de inferência filogenética intraespecífica às espécies de lagartos ocelados (gene *Cyt b*), a partir da metodologia descrita no *ponto 3.3*, calibrada com os parâmetros da *Tabela 5 e obtida* a partir da combinação de 8 *runs* independentes.

A análise topológica evidencia o *outgroup* ser constituído pelas espécies *Lacerta agilis* e *Lacerta schreiberi*, e o *ingroup* pelas espécies do género *Timon*. O *ingroup* revela ser constituído por três grandes separações: a *clade I* que agrupa a espécie *Timon lepidus* + *Timon nevadensis* provenientes da Península Ibérica, *clade II* constituída pela espécie *Timon pater* (Marrocos) e a *clade III* da *espécie Timon tangitanus* (Tunísia). Porém os *clades II* e *III* aparentam estar mal resolvidos devido ao baixo suporte da divergência entre a espécie *Timon pater* e a *clade I* (*pp*=0,694), que poderá ter ocorrido no intervalo [8.654, 5.821 Ma]. A separação entre o *outgroup* e o *ingroup* poderá ter ocorrido no intervalo

[14.106, 10.792 Ma], durante o período terciário da era cenozoica. Seguidamente, a espécie *Timon tangitanus* poderá ter divergido do restante *ingroup* no intervalo [8.816, 6.661 Ma], com um bom suporte de agrupamento (*pp*=1). As espécies do *clade I* no intervalo [6.887, 4.575 Ma], terão divergido entre si dando origem a dois novos agrupamentos, o clade dos *Timon lepidus* provenientes de Portugal e Espanha, e o clade da espécie *Timon nevadensis* da Serra Nevada; suportadas favoravelmente pelos valores *pp*=0,8894 e *pp*=0,999, respetivamente.

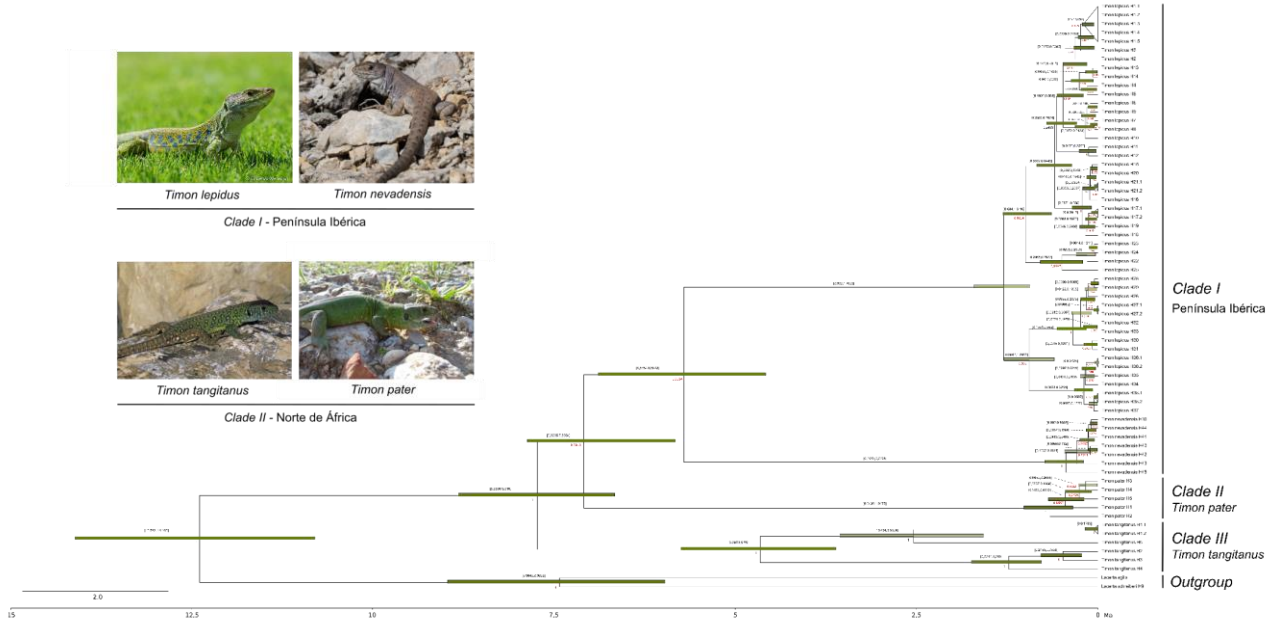


*Figura 7 - Árvore filogenética resultante da inferência intraespecífica (*BEAST2) ao dataset de lagartos ocelados composto pelo gene mitocondrial Cyt b, calibrado com os parâmetros da Tabela 5 a partir da combinação de 8 runs independentes. Fonte das fotografias: Timon lepidus © Eduardo Marabuto, Timon nevadensis © EUROLIZARDS, íNaturalist (Timon tangitanus © Roberto Sindaco, Timon pater © Karim Chouchane). Imagem com melhor resolução: https://raw.githubusercontent.com/ray2g/BEAST2-StarBEAST2/master/final_phylogenetic_trees/lizards_cyt_starbeast.png.*

4.3 Lagartos ocelados – *dataset* concatenado

4.3.1 Inferência interespecífica, *BEAST2*

A *Figura 10* apresenta a árvore filogenética das espécies de lagartos ocelados (*dataset* constituído pelos genes *12S*, *16S*, *Cyt b*, *β Fibrinogen* e *C-mos* concatenados) resultante da inferência interespecífica com o software *BEAST2*, a partir da metodologia apresentada no *ponto 3.5*. Esta foi calibrada com os parâmetros presentes na *Tabela 8 e* obtida a partir da combinação de 8 *runs* independentes.

Topologicamente, o *outgroup* é constituído pelas espécies *Lacerta agilis* e *Lacerta schreiberi*, e *ingroup* pelas espécies do género *Timon*. O *ingroup* revela ser constituído por dois grandes *clades*: o *clade I* constituído pelas espécies *Timon lepidus* e *Timon nevadensis*

(ambos provenientes da Península Ibérica) que poderão ter divergido entre si no intervalo [5.001, 7.268 Ma] (fortemente suportados, *pp*=1 para ambos os agrupamentos), e o *clade II* pelas espécies *Timon tangitanus* e *Timon Pater* do Norte de África. A separação entre o *outgroup* e o *ingroup* poderá ter ocorrido no intervalo [21.147, 15.592 Ma]. Seguidamente, os *clade I* e *II* poderão ter divergido entre si no intervalo [9.299, 6.992 Ma], com um forte suporte de agrupamento (*pp*=1 e *pp*=0,9693, respetivamente). As espécies da *clade II* também poderão ter divergido entre si, agrupando-se em dois *clades* distintos no intervalo [8.005, 5.738 Ma], fortemente suportados pela *pp*=1.

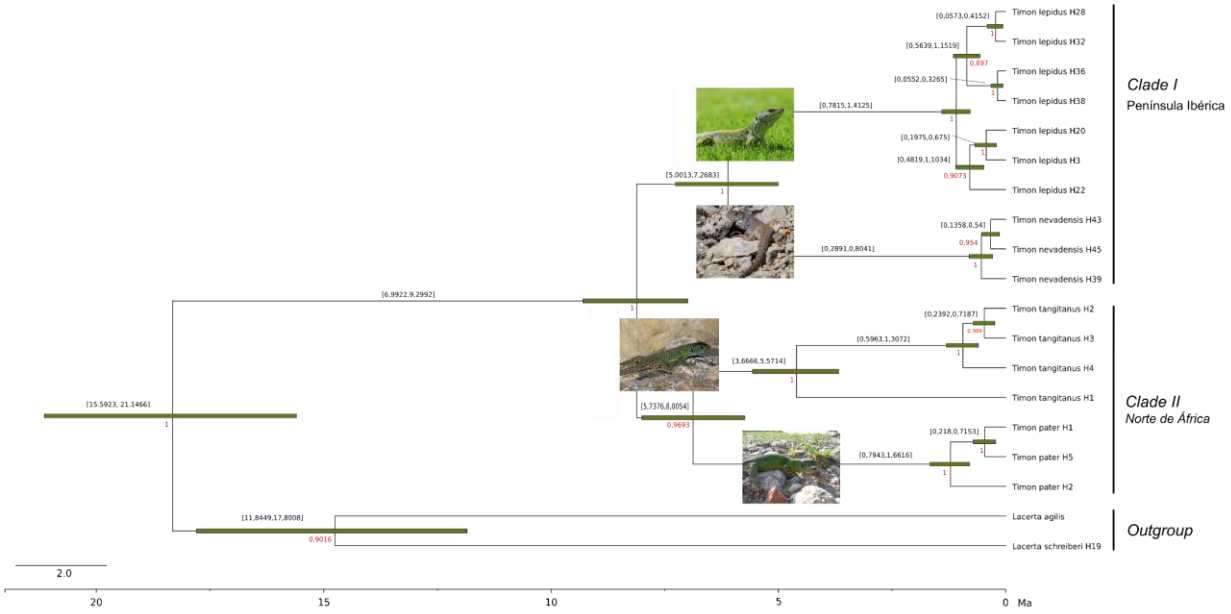


Figura 8 - Árvore filogenética resultante da inferência interespecífica ao dataset concatenado dos lagartos ocelados composto pelos genes mitocôndrias 12S, 16S e Cyt b, e nucleares: β Fibrinogen e C-mos; calibrado com os parâmetros da Tabela 8 a partir da combinação de 8 runs independentes. Fonte das fotografias: Timon lepidus © Eduardo Marabuto, Timon nevadensis © EUROLIZARDS, íNaturalist (Timon tangitanus © Roberto Sindaco, Timon pater © Karim Chouchane). Imagem com melhor resolução: https://raw.githubusercontent.com/ray2g/BEAST2-StarBEAST2/master/final_phylogenetic_trees/lizards_concatenated_beast.png.

4.3.2 Inferência Intraespecífica, **BEAST 2*

A *Figura 11* apresenta a árvore de inferência intraespecífica ao dataset concatenado com genes de lagartos ocelados, resultante da metodologia apresentada no *ponto 3.5*, calibrada com base nos parâmetros da *Tabela 9*, com combinação de 8 *runs* independentes.

A análise topológica realizada no ponto anterior, *4.3.1*, aplica-se igualmente a este ponto uma vez que as duas árvores partilham a mesma topologia apesar das abordagens serem diferentes. Porém, o valor dos tempos de divergência e a *posterior probability* do *nodo* de agrupamento, apresentados na *Tabela 10,* são diferentes.

Tabela 10 - Tempos de divergência relativos e valores de pp dos nodos de agrupamento correspondentes, das principais separações ocorridas na Figura 11.

Separações	Tempos de Divergência Relativos	*PP* do *nodo* de agrupamento
Outgroup* / *Ingroup	[21.002, 14.728 Ma]	1
***Clade* Ibérica / *Clade* de Magrebe**	[9.2029, 6.3021 Ma]	1
Timon lepidus* / *Timon nevadensis	[7.071, 4.473 Ma]	1
Timon tangitanus* / *Timon pater	[8.009, 5.290 Ma]	0,956

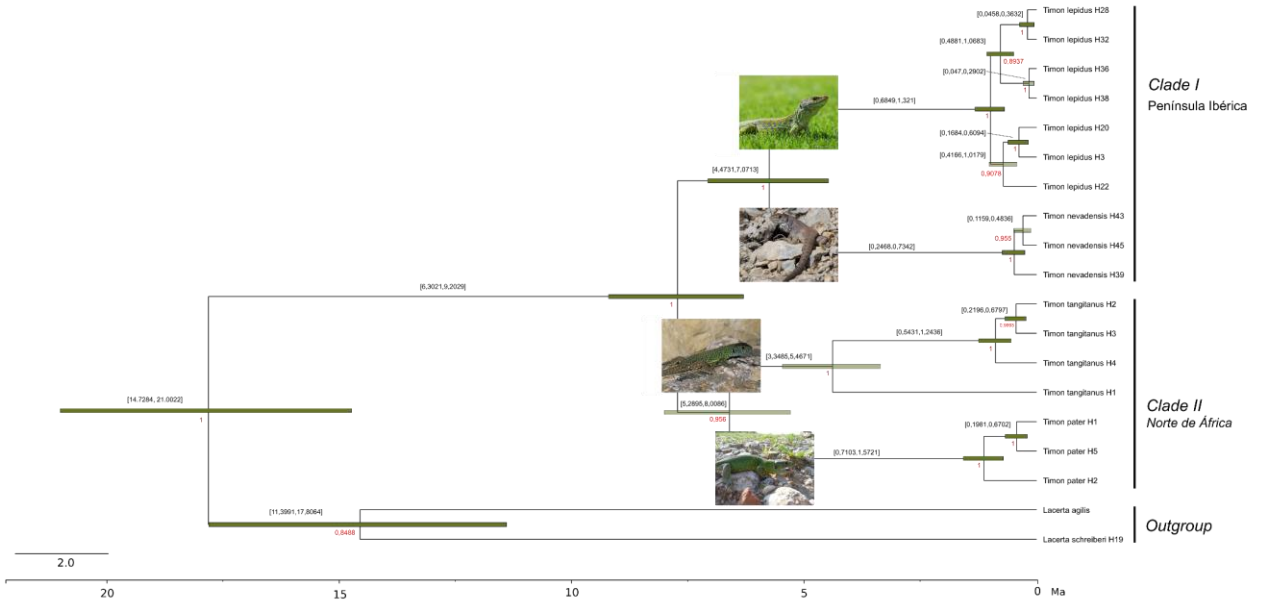


Figura 9 - Árvore filogenética resultante da inferência intraespecífica ao dataset concatenado dos lagartos ocelados composto pelos genes mitocondriais 12S, 16S e Cyt b, e nucleares: β Fibrinogen e C-mos; calibrado com os parâmetros da Tabela 9 e obtido a partir da combinação de 8 runs independentes. Fonte das fotografias: EUROLIZARDS (Timon lepidus + nevadensis), íNaturalist (Timon tangitanus © Roberto Sindaco, Timon pater © Karim Chouchane). Imagem com melhor resolução: https://raw.githubusercontent.com/ray2g/BEAST2-StarBEAST2/master/final_phylogenetic_trees/lizards_concatenated_starbeast.png.

4.4 Borboletas do género *Lycaena*

Na *Figura 12*, encontra-se apresentada a árvore filogenética de inferência interespecífica às espécies de borboletas do género *Lycaena* a partir dos genes COI e EF-1α, resultante da metodologia apresentada no *ponto 3.4.1*. Esta foi calibrada de acordo com os parâmetros da *Tabela 6* e obtida a partir da combinação de 6 *runs* independentes.

Topologicamente, o *outgroup* é constituído pela espécie *Lampides boeticus* (LBO_Lampi) e o *ingroup*, o género *Lycaena* propriamente dito. Este último revelou ser constituído por três *clades*, sendo que apenas dois deles estão bem suportados. A separação do *outgroup* em relação ao *ingroup* foi calibrada num tempo de divergência ancorado com base na publicação [61] (*Tabela 6)*; o intervalo obtido para este *nodo* foi [62.395, 56 Ma], ocorrido durante o paleocénico do período terciário. O *ingroup* revelou dois grandes *clades* bem suportados (*pp*=1 para a *clade I* e *pp*=0,976 para a *clade II*), consultar a *Figura 12*. Porém um conjunto de espécies na base da filogenia (*Lycaena lhelle* (LHELLE_Lhelle) que aparece como *clade* individualizado, tendo divergido no intervalo [31.649, 16.505 Ma], e *Lycaena li* (LLI_Li) + *pang* (LPA_Lpang) que aparecem associadas ao *clade II* mas com pouco suporte) não teve uma posição bem resolvida. O *clade I*, constituído maioritariamente por espécies de distribuição Euroasiática (exceto a *Lycaena cupreus* (LCU_Lcupr) proveniente da América do Norte, que divergiu do restante *clade* no intervalo [14.778, 4.910 Ma]), terá divergido das restantes espécies no intervalo [28.851, 15.641 Ma], a par das espécies orientais *Lycaena li* e *Lycaena pang* que terão divergido das restantes espécies do *clade II* no intervalo [26.913, 14.336 Ma] com baixo suporte (*pp*=0,401). A diferenciação na base da árvore está, portanto, pouco suportada e tanto as relações filogenéticas entre os grandes grupos como os seus tempos de divergência devem ser interpretados com cautela. Um dos objetivos desta filogenia era averiguar a posição do par de espécies *Lycaena bleusei* (LBL) e *Lycaena tityrus* (LTI). A árvore filogenética obtida, demonstra que estas duas espécies são irmãs, com elevado suporte (*pp*=0,998), tendo divergido entre si no intervalo [10.3195, 3.140 Ma] e incluídas perfeitamente no *clade I*. Após a análise, é possível evidenciar que populações da mesma espécie podem ter divergências muito antigas, como por exemplo a *Lycaena alciphron* cujas duas amostras terão divergido no intervalo [6.936, 0.846 Ma], ou em contraste divergências muito recentes como a registada entre populações de *Lycaena dorcas* durante o intervalo [1.136, 0.002 Ma]. Por outro lado, espécies diferentes podem ter divergências muito recentes, como por exemplo *Lycaena candens* Vs. *Lycaena hippothoe* que poderão ter divergido entre si no intervalo [3.042, 0.0501 Ma], *Lycaena solskyii* Vs. L*ycaena alpherakii* no intervalo [2.53, 0.07 Ma], e *Lycanea dione* Vs*. Lycaena editha durante* o intervalo [2.604, 0.132 Ma].

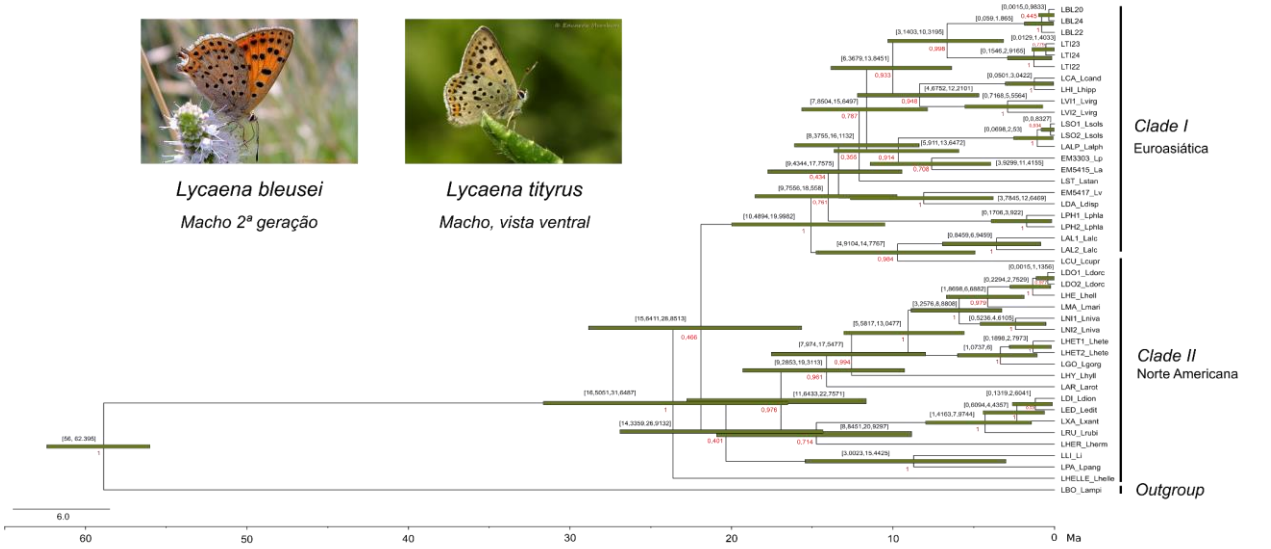


Figura 10 - Árvore filogenética resultante da inferência interespecífica do dataset das borboletas do género Lycaena composto pelo gene mitocondrial COI e nuclear EF-1α, calibrado com os parâmetros da Tabela 6. Fonte das fotografias: © Eduardo Marabuto. Imagem com melhor resolução: https://raw.githubusercontent.com/ray2g/BEAST2-StarBEAST2/master/final_phylogenetic_trees/lycaena_beast.png.

4.5 Borboletas do género *Melanargia*

Na *Figura 13*, encontra-se apresentada a árvore filogenética de inferência interespecífica às espécies de borboletas do género *Melanargia (genes 16S, COI, EF-1α* e *Wg*), obtida a partir da metodologia apresentada no *ponto 3.4.3* com combinação de 6 *runs* independentes e calibrada com os parâmetros da *Tabela 7*.

A análise topológica evidência o *outgroup* ser constituído pela espécie *Maniola jurtina* (*M.jurtina*), e o *ingroup* composto por dois grandes *clades*. O *clade I* composto pela espécie *Melanargia occitanica* (*M.occi*) com a sua espécie irmã, *Melanargia arge* (M.arge), e o *clade II composto pela espécie Melanargia ines (M.ines).* O tempo de divergência entre o *outgroup* e o *ingroup* foi calibrado com base na publicação [61, 62], (*Tabela 7) e estabelecido* neste nodo entre [35.705, 32 Ma], durante a época Eocénica do período Terciário. A divergência entre o *clade I* e o *clade II* terá ocorrido no intervalo [20.905, 8.519 Ma]. Seguidamente, ter-se-á dado a separação entre as espécies irmãs presentes no *clade I*, *Melanargia arge* e *Melanargia occitanica* no intervalo [8.763, 2.92 Ma], bem suportada (*pp*=1 e *pp*=0,996 respetivamente). No decorrer do tempo, as espécies *Melanargia occitanica* e *Melanargia ines* terão sofrido ambas segregação geográfica. De modo que, no intervalo [5.534, 1.636 Ma] a espécie *Melanargia occitanica* poderá ter-se separado em dois *clades* distintos, agrupando-se num *clade* Europeu (*M.occi: SS, WP, NS, France* e *NI*) e num *clade* Norte Africano *(M.occi: HA, MA* e *Sicily*), ambos bem suportados com *pp*=1. A árvore revela ainda que a espécie *Melanargia ines* se terá segregado no intervalo [6.796, 1.733 Ma] dando origem aos *clades* Norte Africano (*M.ines: AA, HA e MA*) e Ibérico (*M.ines: CS, SS* e *NS*); igualmente com um bom suporte (*pp*=1).

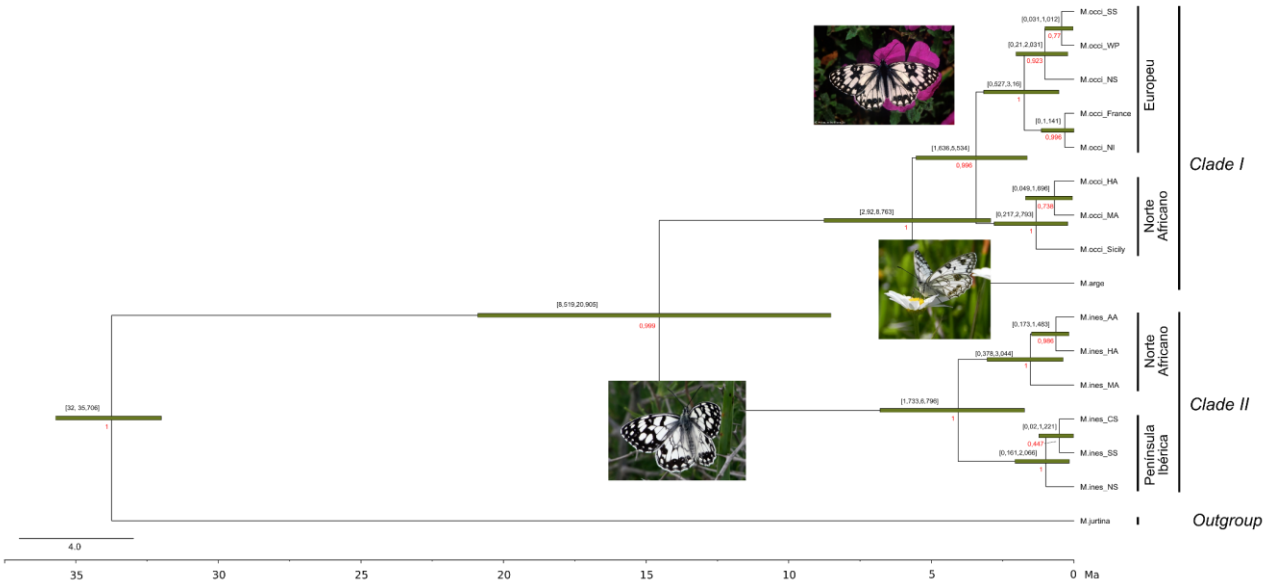


Figura 11 - Árvore filogenética resultante da inferência interespecífica ao dataset das borboletas do género Melanargia composto pelos genes mitocondriais 16S e COI e nucleares: EF-1α e Wg; calibrado com os parâmetros da Tabela 7 e obtido a partir da combinação de 6 runs independentes. Fonte das fotografias: Naturdata (Melanargia occitanica © Fernando Romão, Melanargia ines © Luís Nunes Alberto) e íNaturalist (Melanargia arge © Giuseppe Cagnetta). Imagem com melhor resolução: https://raw.githubusercontent.com/ray2g/BEAST2-StarBEAST2/master/final_phylogenetic_trees/melanargia_beast.png.

5 Discussão de resultados

Os valores dos tempos de divergência, a *posterior probability* do nodo de agrupamento e o tempo de processamento são fatores essenciais que dão resposta a qual dos métodos, *BEAST* ou **BEAST*, que atualmente é o mais viável de ser aplicado tendo conta o suporte e a resolução da árvore filogenética.

5.1 Espécies de cigarras dos géneros *Maoricicada, Rhodopsalta e Kikihia*

Como anteriormente mencionado, a inferência filogenética molecular das cigarras do grupo MRK foi realizada sem a calibração de um *outgroup*. De acordo com o autor, a idade de referência esperada para a *clade* MRK situa-se no intervalo [39, 38 Ma] em contraste com o intervalo obtido na inferência intraespecífica realizada, correspondente a [0.253, 0.065 Ma]. A árvore inferida pelo autor é composta pelas espécies da tribo *Cicadettini* da família *Cicadidae* provenientes da Australásia, sendo um dos clades o grupo MRK (aqui analisado), calibrada com um *outgroup* constituído pelas espécies *Lembeja paradoxa*, *Lembeja vitticollis* e *Cystosoma saundersii*. Deste modo, os resultados obtidos com ambas as abordagens (*BEAST2* e **BEAST2*) apresentam má resolução e suporte dos *clades*, apesar da análise dos valores de ESS no software *Tracer* mostrar boa convergência dos parâmetros; os *Anexos 5* e *6* apresentam o resultado das distribuições dos parâmetros de

ambas as abordagens. É importante realçar que o *dataset* é constituído por um amplo grupo de espécies MRK suportado apenas pela inferência a partir de um único gene (*COI*), o que faz com que a sua resolução possa não ser suficiente para inferir com confiança todas as relações filogenéticas entre os *taxa* presentes.

Tabela 11 - Resultados dos tempos de divergência relativos, valores de pp no nodo de agrupamento e tempo de processamento das principais separações referentes à inferência filogenética molecular inter e intraespecífica, às cigarras dos géneros Maoricicada, Rhodopsalta e Kikihia a partir do gene COI. Especificações técnicas do computador onde foram executados os processos: Intel® Core™ i5-7200U, Intel® HD Graphics 620, 8 GB DDR4.

Dataset* de cigarras dos géneros *Maoricicada, Rhodopsalta* e *Kikihia* - *COI						
Principais Separações	Tempos de Divergência Relativos		*PP* no nodo de agrupamento		Tempo de Processamento (min)	
	BEAST	**BEAST*	*BEAST*	**BEAST*	*BEAST*	**BEAST*
Kikihia / Maoricicada + Rhadopsalta	[0.372, 0.116 Ma]	[0.253, 0.065 Ma]	1	1	31	16.6
Maoricicada / Rhadopsalta	[0.311, 0.094 Ma]	[0.21, 0.056 Ma]	0.744	0.666		

Os resultados dos tempos de divergência com ambas as abordagens apresentados na *Tabela 11* demonstram que, apesar da diferença ser mínima, o **BEAST* apresenta intervalos de tempo de divergência menores. Quanto aos valores de *pp,* estes baixam à medida que a divergência dos *clades* se torna mais recente em ambas as abordagens; consultar as *Figuras 6* e *7*. Ainda assim, o método de inferência filogenética molecular **BEAST2* apresenta ser ~1,86 vezes mais rápido que o método *BEAST2* no que diz respeito ao tempo de processamento.

5.2 Lagartos Ocelados

5.2.1 *Dataset* composto pelo gene *Cytochrome b*

Os resultados da inferência filogenética molecular realizado em espécies de lagartos ocelados, apenas com o gene *Cytochrome b,* evidenciam uma má resolução da árvore filogenética pelo método *BEAST2* relativamente aos obtidos através do **BEAST2*. Uma vez que a espécies *Timon tangitanus* e *Timon pater* são irmãs suportadas com *pp*=0,512, e *pp*=1 para ambas os *clades* após a divergência. O método **BEAST* sugere uma melhor conformação topológica garantindo uma boa resolução entre os *clades* norte africanos. Primeiramente, o *clade* da espécie *Timon tangitanus* terá divergido do restante *ingroup* suportado pela *pp*=1 e só depois poderá ter ocorrido a divergência da espécie *Timon pater com pp*=0,694. O *clade I* encontra-se bem suportado pela pp=0,998, bem como a

divergência das espécies ibéricas (*pp*=1). De modo, que é possível inferir que o *clade* da espécie *Timon pater* se encontra mal suportado em ambas as abordagens e mal resolvida pelo método *BEAST2*. A implementação de mais genes na inferência, possibilitaria a resolução do *clade* da espécie *Timon pater,* e melhorar o suporte de agrupamento dos clades (valores de *pp*) na globalidade das duas abordagens.

Tabela 12 - Resultados dos tempos de divergência relativos, valores de pp no nodo de agrupamento e tempo de processamento das principais separações referentes à inferência filogenética molecular inter e intraespecífica, ao dataset dos lagartos ocelados a partir do gene COI. Especificações técnicas do computador onde foram executados os processos: Intel® Core™ i5-7200U, Intel® HD Graphics 620, 8 GB DDR4.

Dataset de lagartos ocelados – *Cyt b*						
Principais Separações	Tempos de Divergência Relativos		*PP* no nodo de agrupamento		Tempo de Duração de Processamento (min)	
	BEAST	**BEAST*	*BEAST*	**BEAST*	*BEAST*	**BEAST*
Outgroup / Ingroup	[26.384, 15.536 Ma]	[14.106, 10.792 Ma]	1	1	26,8	5
Clade I / Clade II	[11.132, 7.339 Ma]	----------	1	----------		
Timon pater / Clade I	----------	[8.654, 5.821 Ma]	----------	0.694		
Timon lepidus / Timon nevadensis	[8.782, 5.084 Ma]	[6.887, 4.575 Ma]	0.965	0.998		
Timon tangitanus / Clade I + T. pater	----------	[8.816, 6.661 Ma]	----------	1		
Timon tangitanus / Timon pater	[9.644, 5.934 Ma]	----------	0.512	----------		

De acordo com a *Tabela 12*, o método **BEAST2* evidência intervalos de tempos de divergência mais pequenos, bem como os valores *pp* em separações comuns aos dois métodos. O método **BEAST2* destaca-se ainda por ter um tempo de processamento ~5.36x mais rápido.

5.2.2 *Dataset* concatenado

A inferência filogenética molecular ao *dataset* concatenado com cinco genes das espécies de lagartos ocelados, evidencia que as árvores filogenéticas resultantes de ambos os métodos, *BEAST2* e **BEAST2*, possuem a mesma topologia. Ambas as árvores aparentam estar bem resolvidas e suportadas pelos valores de *pp* de acordo com a *Tabela 13* e com as *Figuras 10* e *11*. Isto devido à remoção de haplótipos (pois nem todas as amostras tinham a informação genética comum aos cinco genes) e adicionados mais genes neste *dataset* comparativamente ao *dataset* da inferência com apenas o *Cytochrome b.* Na *Tabela*

13 podemos observar que os métodos possuem intervalos de divergência semelhantes, porém os do **BEAST2* são relativamente curtos. O método **BEAST2* destaca-se quando analisados os tempos de processamento. Este foi ~2,7x mais rápido que o método *BEAST2*.

Tabela 13 - Resultados dos tempos de divergência relativos, valores de pp no nodo de agrupamento e tempo de processamento das principais separações referentes à inferência filogenética molecular inter e intraespecífica, ao dataset concatenado dos lagartos ocelados. Especificações técnicas do computador onde foram executados os processos: Intel® Core™ i5-7200U, Intel® HD Graphics 620, 8 GB DDR4.

Dataset de lagartos ocelados - concatenado						
Principais Separações	Tempos de Divergência Relativos		Posterior do nodo referente		Tempo de Duração de Processamento (min)	
	BEAST	**BEAST*	*BEAST*	**BEAST*	*BEAST*	**BEAST*
Outgroup / Ingroup	[21.147,15.592 Ma]	[21.002, 14,728 Ma]	1	1	42.82	16.1
Clade I / Clade II	[9.299, 6.992 Ma]	[9.2029, 6.302 Ma]	1	1		
Timon lepidus / Timon nevadensis	[7.727, 5.001 Ma]	[7.071, 4.473 Ma]	1	1		
Timon tangitanus / Timon pater	[8.005, 5.738 Ma]	[8.009, 5.290 Ma]	0.9693	0.956		

5.2.3 Comparação dos *datasets*: 1 gene (*Cytochrome b*) *Vs.* 5 genes (concatenado)

Ao se comparar os resultados dos tempos de divergência obtidos a partir da inferência aos *datasets* com o gene *Cyt b* (*Tabela 12*) e ao concatenado com os genes *12S*, *16S*, *Cyt b*, *β Fibrinogen* e *C-mos* (*Tabela 13*), verifica-se que com ambos os *software* obteve-se intervalos de divergência muito discrepantes, principalmente na separação entre o *outgroup* e o *ingroup.* Resultado o qual não foi apresentado no artigo de referência [54] pelo autor.

Comparativamente aos resultados do artigo onde foi apenas inferido o gene *Cyt b com o software BEAST v1,* para a separação entre o *clades I* (Península Ibérica) e *clade II* (Norte de África) obteve-se o intervalo de divergência [12.56, 6.59 Ma]. O estudo realizado neste projeto demonstra que para esta separação, obteve-se com o *software BEAST2* a partir do *dataset* apenas com o gene *Cyt b* o intervalo [11.132, 7.339 Ma] e no *dataset* concatenado o intervalo [9.299, 6.992 Ma]. Já o **BEAST2* no dataset do *Cyt b* obteve uma má resolução deste intervalo, mas para o dataset concatenado inferiu o intervalo [9.2029, 6.302 Ma], muito semelhante ao obtido com o *BEAST2.*

Para a separação entre as espécies *Timon lepidus* e *Timon nevadensis* o artigo apresenta o intervalo de divergência [11.66, 4.03 Ma] como referência. Na inferência realizada com o

software BEAST2 obteve-se o intervalo [11.132, 7.339 Ma] para o *dataset* do *Cyt b e* [9.299, 6.992 Ma] para o concatenado. Nesta separação o **BEAST* apresentou novamente uma má resolução no *dataset* do *Cyt b*, e o intervalo [9.2029, 6.302 Ma] para o concatenado (mais uma vez muito semelhante ao inferido pelo *BEAST2*).

No que diz respeito à separação das espécies *Timon tangitanus* e *Timon pater* o artigo de referência apresentou com resultado o intervalo [10.53, 4.57 Ma]. A inferência realizada neste projeto demonstrou que para esta divergência, o *BEAST2* obteve o intervalo [9.644, 5.934 Ma] para o *dataset* do *Cytochrome b* e o intervalo [8.005, 5.738 Ma] para o *dataset* concatenado.

Deste modo, verifica-se que os resultados obtidos que estão mais em linha com os do artigo de referência são os do *dataset* do *Cyt b* inferidos com o *BEAST2,* uma vez que se trata do mesmo gene e recorre-se à mesma abordagem (*BEAST2*), porém numa versão diferente e melhorada. Porém o uso da nova abordagem **BEAST* com o *dataset* concatenado permitiu obter intervalos de divergência mais curtos comparativamente aos apresentados no artigo.

5.3 Borboletas dos géneros *Lycaena* e *Melanargia*

A inferência filogenética, com o *software BEAST2*, ao género *Lycaena* demonstrou que a árvore resultante possui *clades* com posições mal resolvidas e com baixo suporte estatístico, como anteriormente referido no ponto *4.4*. Este resultado deve-se à grande quantidade de espécies envolvidas na inferência face ao número de genes usados, neste caso o *COI* e o *EF-1α*.

Devido à grande proximidade fenotípica entre os *taxa Lycaena bleusei* e *Lycaena tityrus*, poderia ser esperada uma estimativa de tempo de divergência mais recente entre estas do que aquele que foi de facto inferida. No entanto, o resultado obtido reflete a diferenciação dos genes *COI* e *EF-1α* entre as duas espécies, que é também superior à esperada.

No caso da inferência filogenética realizada ao género *Melanargia* foram usadas muito menos espécies em comparação com *dataset* do género *Lycaena*, tendo sido por outro lado analisados o dobro dos genes (*16S, COI, EF-1α* e *Wg*) o que conferiu à árvore final um bom suporte dos clades, que são apresentados como bem resolvidos. Os tempos de divergência estão em linha com o esperado de acordo com os fenómenos biogeográficos envolvidos na separação das linhagens e a evolução do Mediterrâneo, desde a crise salínica do Messiniano (6.8 – 5.3 Ma) às glaciações do Pleistoceno que isolaram diversas populações.

6 Conclusões

A implementação de relógios moleculares em análises filogenéticas permite relacionar o tempo de divergência entre espécies com o número de diferenças moleculares calculadas com base numa taxa (*rate*) de mutação constante. Mesmo não fornecendo uma informação exata, os relógios moleculares permitem estimar uma escala de tempo dos eventos da história evolutiva e inferir as diferentes hipóteses biológicas durante a evolução das espécies.

Os resultados das inferências filogenéticas moleculares realizadas ao longo do projeto com os métodos *BEAST2* e **BEAST2*, demonstram que o uso de um *outgroup* adequado na calibração da árvore filogenética é essencial de forma a garantir suporte de confiança à inferência realizada. Bem como a proporcionalidade entre o número de genes e *taxa* envolvidos na inferência, que garantem a sua resolução. O uso de relógios inapropriados ou calibrações incorretas podem conduzir a resultados ilusórios das escalas de tempo. De modo que, é relevante o uso de registos fósseis que garantem a idade mínima de existência dos *taxa*.

Ao compararmos os dois métodos, *BEAST2* e **BEAST2*, quando aplicados a *datasets* com pouca variabilidade, o **BEAST2* demonstra obter resultados com melhor resolução, intervalos de divergência menores e um tempo de processamento pelo menos duas vezes mais rápido. Esta conclusão é fundamentada com os resultados obtidos na inferência ao género *Lycaena* apenas com o *software BEAST2* (inferência interespecífica), que apresentou *nodos* mal resolvidos na árvore filogenética resultante.

É ainda importante referir que, quando inferidos os resultados deve ter-se sempre em consideração o contexto biológico, uma vez que o ponto de vista informático, não deve de ser interpretado isoladamente, mesmo quando bem suportado pela análise de confiança.

Conclusions

The implementation of molecular clocks in phylogenetic analyses allows relating divergence time between species with the number of molecular differences calculated based on a constant mutation rate. Even without providing accurate information, molecular clocks allow estimating a timescale of evolutionary history events and infer different biological hypotheses during the species' evolution.

Results of molecular phylogenetic inferences carried throughout the project with *BEAST2* and **BEAST2* methods, show that the use of an adequate outgroup in the phylogenetic tree calibration is essential in order to ensure confidence support to the inference. Futhermore, proportionality between the number of genes and the taxa involved in inference are fundamental to ensure node resolution. Use of innapropriate pacemakers or incorrect calibrations may lead to illusory timescale results. Therefore, the use of fossil records is relevant as they ensure the minimum age of taxa existence.

By comparing the two methods, *BEAST2* e **BEAST2*, when applied to datasets with low variability, **BEAST2* demonstrates better node resolution, lower divergence times and at least twice as fast processing times. This conclusion is supported by the results obtained in the inference with butterflies' species of the genus *Lycaena*, only with *BEAST2* software (interspecific inference), which showed poorly node resolution in the resultant phylogenetic tree.

It is also important to note that, when inferring the results, the biological context must always be taken into account, since the computer point of view should not be interpreted in isolation, even when well supported by confidence values' analyses.

7 Referências Bibliográficas

[1] D. E. Soltis e P. S. Soltis, «The Role of Phylogenetics in Comparative Genetics», *Plant Physiology*, vol. 132, n. 4, pp. 1790–1800, Ago. 2003.

[2] M. R. Dietrich, «Paradox and Persuasion: Negotiating the Place of Molecular Evolution within Evolutionary Biology», *Journal of the History of Biology*, vol. 31, n. 1, pp. 85–111, Mar. 1998.

[3] P. Ajawatanawong, «Molecular Phylogenetics: Concepts for a Newcomer», *Adv. Biochem. Eng. Biotechnol.*, vol. 160, pp. 185–196, 2017.

[4] C. Senés-Guerrero, G. Torres-Cortés, S. Pfeiffer, M. Rojas, e A. Schüßler, «Potato-associated arbuscular mycorrhizal fungal communities in the Peruvian Andes», *Mycorrhiza*, vol. 24, n. 6, pp. 405–417, Ago. 2014.

[5] E. Kenah, T. Britton, M. E. Halloran, e I. M. L. Jr, «Molecular Infectious Disease Epidemiology: Survival Analysis and Algorithms Linking Phylogenies to Transmission Trees», *PLOS Computational Biology*, vol. 12, n. 4, p. e1004869, Abr. 2016.

[6] A. B. Chang, R. Lin, W. Keith Studley, C. V. Tran, e M. H. Saier, «Phylogeny as a guide to structure and function of membrane transport proteins», *Mol. Membr. Biol.*, vol. 21, n. 3, pp. 171–181, Jun. 2004.

[7] J. B. H. Martiny, S. E. Jones, J. T. Lennon, e A. C. Martiny, «Microbiomes in light of traits: A phylogenetic perspective», *Science*, vol. 350, n. 6261, p. aac9323, Nov. 2015.

[8] M. Siljic *et al.*, «Forensic application of phylogenetic analyses - Exploration of suspected HIV-1 transmission case», *Forensic Sci Int Genet*, vol. 27, pp. 100–105, 2017.

[9] K. A. Jacobson, S. Costanzi, e S. Paoletta, «Computational studies to predict or explain GPCR polypharmacology», *Trends Pharmacol Sci*, vol. 35, n. 12, pp. 658–663, Dez. 2014.

[10] S. Ojosnegros e N. Beerenwinkel, «Models of RNA virus evolution and their roles in vaccine design», *Immunome Res*, vol. 6, n. Suppl 2, p. S5, Nov. 2010.

[11] H. A. Khan, I. A. Arif, A. H. Bahkali, A. H. Al Farhan, e A. A. Al Homaidan, «Bayesian, maximum parsimony and UPGMA models for inferring the phylogenies of antelopes using mitochondrial markers», *Evol. Bioinform. Online*, vol. 4, pp. 263–270, Out. 2008.

[12] N. Saitou e M. Nei, «The neighbor-joining method: a new method for reconstructing phylogenetic trees», *Mol. Biol. Evol.*, vol. 4, n. 4, pp. 406–425, Jul. 1987.

[13] J. Wen, Y. Xu, Z. Li, Z. Ma, e Y. Xu, «Inter-class sparsity based discriminative least square regression», *Neural Netw*, vol. 102, pp. 36–47, Jun. 2018.

[14] S. Bastkowski, V. Moulton, A. Spillner, e T. Wu, «The minimum evolution problem is hard: a link between tree inference and graph clustering problems», *Bioinformatics*, vol. 32, n. 4, pp. 518–522, Fev. 2016.

[15] D. Ortega-Del Vecchyo, D. Piñero, L. Jardón-Barbolla, e J. van Heerwaarden, «Appropriate homoplasy metrics in linked SSRs to predict an underestimation of demographic expansion times», *BMC Evolutionary Biology*, vol. 17, n. 1, p. 213, Set. 2017.

[16] C.-B. Stewart, «The powers and pitfalls of parsimony», *Nature*, vol. 361, n. 6413, pp. 603–607, Fev. 1993.

[17] E. L. Lawler e D. E. Wood, «Branch-and-Bound Methods: A Survey», *Operations Research*, vol. 14, n. 4, pp. 699–719, Ago. 1966.

[18] A. Goëffon, J.-M. Richer, e J.-K. Hao, «Heuristic Methods for Phylogenetic Reconstruction with Maximum Parsimony», em *Algorithms in Computational Molecular Biology*, John Wiley & Sons, Ltd, 2010, pp. 579–597.

[19] L. Kannan e W. C. Wheeler, «Maximum Parsimony on Phylogenetic networks», *Algorithms Mol Biol*, vol. 7, n. 1, p. 9, Mai. 2012.

[20] J. Felsenstein, «Evolutionary trees from DNA sequences: A maximum likelihood approach», *J Mol Evol*, vol. 17, n. 6, pp. 368–376, Nov. 1981.

[21] A. Som, «Theoretical foundation to estimate the relative efficiencies of the Jukes-Cantor+gamma model and the Jukes-Cantor model in obtaining the correct phylogenetic tree», *Gene*, vol. 385, pp. 103–110, Dez. 2006.

[22] M. Kimura, «A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences», *J Mol Evol*, vol. 16, n. 2, pp. 111–120, Jun. 1980.

[23] J. Felsenstein, «Evolutionary trees from DNA sequences: A maximum likelihood approach», *J Mol Evol*, vol. 17, n. 6, pp. 368–376, Nov. 1981.

[24] M. Hasegawa, H. Kishino, e T. Yano, «Dating of the human-ape splitting by a molecular clock of mitochondrial DNA», *J Mol Evol*, vol. 22, n. 2, pp. 160–174, Out. 1985.

[25] K. Tamura e M. Nei, «Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees», *Mol. Biol. Evol.*, vol. 10, n. 3, pp. 512–526, Mai. 1993.

[26] L. Gatto, D. Catanzaro, e M. C. Milinkovitch, «Assessing the Applicability of the GTR Nucleotide Substitution Model Through Simulations», *Evol Bioinform Online*, vol. 2, pp. 145–155, Fev. 2007.

[27] P. Liò e N. Goldman, «Models of molecular evolution and phylogeny», *Genome Res.*, vol. 8, n. 12, pp. 1233–1244, Dez. 1998.

[28] J. Sullivan e P. Joyce, «Model Selection in Phylogenetics», *Annual Review of Ecology, Evolution, and Systematics*, vol. 36, n. 1, pp. 445–466, 2005.

[29] M. Arenas, «Trends in substitution models of molecular evolution», *Front Genet*, vol. 6, Out. 2015.

[30] D. Posada e K. A. Crandall, «MODELTEST: testing the model of DNA substitution.», *Bioinformatics*, vol. 14, n. 9, pp. 817–818, Jan. 1998.

[31] D. Darriba, G. L. Taboada, R. Doallo, e D. Posada, «jModelTest 2: more models, new heuristics and parallel computing», *Nature Methods*, vol. 9, n. 8, pp. 772–772, Ago. 2012.

[32] J. Bertl, G. Ewing, C. Kosiol, e A. Futschik, «Approximate maximum likelihood estimation for population genetic inference», *Stat Appl Genet Mol Biol*, vol. 16, n. 5–6, pp. 387–405, 27 2017.

[33] Z. Yang e B. Rannala, «Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds», *Mol. Biol. Evol.*, vol. 23, n. 1, pp. 212–226, Jan. 2006.

[34] M. E. Alfaro e M. T. Holder, «The Posterior and the Prior in Bayesian Phylogenetics», *Annual Review of Ecology, Evolution, and Systematics*, vol. 37, n. 1, pp. 19–42, 2006.

[35] M. Holder e P. O. Lewis, «Phylogeny estimation: traditional and Bayesian approaches», *Nat. Rev. Genet.*, vol. 4, n. 4, pp. 275–284, Abr. 2003.

[36] A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, e A. Stamatakis, «RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference», *Bioinformatics*, vol. 35, n. 21, pp. 4453–4455, Nov. 2019.

[37] J. P. Huelsenbeck e F. Ronquist, «MRBAYES: Bayesian inference of phylogenetic trees», *Bioinformatics*, vol. 17, n. 8, pp. 754–755, Ago. 2001.

[38] F. Ronquist *et al.*, «MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space», *Syst. Biol.*, vol. 61, n. 3, pp. 539–542, Mai. 2012.

[39] R. Bouckaert *et al.*, «BEAST 2: A Software Platform for Bayesian Evolutionary Analysis», *PLOS Computational Biology*, vol. 10, n. 4, p. e1003537, Abr. 2014.

[40] G. J. Morgan, «Emile Zuckerkandl, Linus Pauling, and the Molecular Evolutionary Clock, 1959-1965», *Journal of the History of Biology*, vol. 31, n. 2, pp. 155–178, 1998.

[41] L. Bromham e D. Penny, «The modern molecular clock», *Nature Reviews Genetics*, vol. 4, n. 3, pp. 216–224, Mar. 2003.

[42] M. S. Y. Lee e S. Y. W. Ho, «Molecular clocks», *Current Biology*, vol. 26, n. 10, pp. R399–R402, Mai. 2016.

[43] M. A. Suchard, P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, e A. Rambaut, «Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10», *Virus Evol*, vol. 4, n. 1, p. vey016, Jan. 2018.

[44] J. Barido-Sottani *et al.*, «Taming the BEAST-A Community Teaching Material Resource for BEAST 2», *Syst. Biol.*, vol. 67, n. 1, pp. 170–174, 01 2018.

[45] J. Heled e A. J. Drummond, «Bayesian inference of species trees from multilocus data», *Mol. Biol. Evol.*, vol. 27, n. 3, pp. 570–580, Mar. 2010.

[46] T. A. Heath, D. S. Divergences, e F. B. Process, «Divergence Time Estimation using BEAST v2.2.0 Dating Species Divergences with the Fossilized Birth-Death Process», n. Mcmc, pp. 1–44, 2016.

[47] R. Bouckaert *et al.*, «BEAST 2: A Software Platform for Bayesian Evolutionary Analysis», *PLOS Computational Biology*, vol. 10, n. 4, p. e1003537, Abr. 2014.

[48] Z. Yang e B. Rannala, «Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds», *Mol. Biol. Evol.*, vol. 23, n. 1, pp. 212–226, Jan. 2006.

[49] R. Bouckaert *et al.*, «BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis», *PLOS Computational Biology*, vol. 15, n. 4, p. e1006650, Abr. 2019.

[50] A. Rambaut, A. J. Drummond, D. Xie, G. Baele, e M. A. Suchard, «Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7», *Syst. Biol.*, vol. 67, n. 5, pp. 901–904, 01 2018

[51] A. Rambaut, «FigTree, version 1.4.3», 2009.

[52] K. Katoh e D. M. Standley, «MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability», *Mol Biol Evol*, vol. 30, n. 4, pp. 772–780, Abr. 2013.

[53] A. Larsson, «AliView: a fast and lightweight alignment viewer and editor for large datasets», *Bioinformatics*, vol. 30, n. 22, pp. 3276–3278, Nov. 2014.

[54] O. S. Paulo, J. Pinheiro, A. Miraldo, M. W. Bruford, W. C. Jordan, e R. A. Nichols, «The role of vicariance vs. dispersal in shaping genetic patterns in ocellated lizard species in the western Mediterranean», *Molecular Ecology*, vol. 17, n. 6, pp. 1535–1551, 2008.

[55] I. Ribera *et al.*, «Ancient origin of a Western Mediterranean radiation of subterranean beetles», *BMC Evolutionary Biology*, vol. 10, n. 1, p. 29, Jan. 2010.

[56] E. Sérusiaux, J. C. V. A, T. Wheeler, e B. Goffinet, «Recent origin, active speciation and dispersal for the lichen genus Nephroma (Peltigerales) in Macaronesia», *Journal of Biogeography*, vol. 38, n. 6, pp. 1138–1151, 2011.

[57] I. R. Amorim, B. C. Emerson, P. A. V. Borges, e R. K. Wayne, «Phylogeography and molecular phylogeny of Macaronesian island Tarphius (Coleoptera: Zopheridae): why are there so few species in the Azores?», *Journal of Biogeography*, vol. 39, n. 9, pp. 1583–1595, 2012.

[58] A. Faille, C. Andújar, F. Fadrique, e I. Ribera, «Late Miocene origin of an Ibero-Maghrebian clade of ground beetles with multiple colonizations of the subterranean environment», *Journal of Biogeography*, vol. 41, n. 10, pp. 1979–1990, 2014.

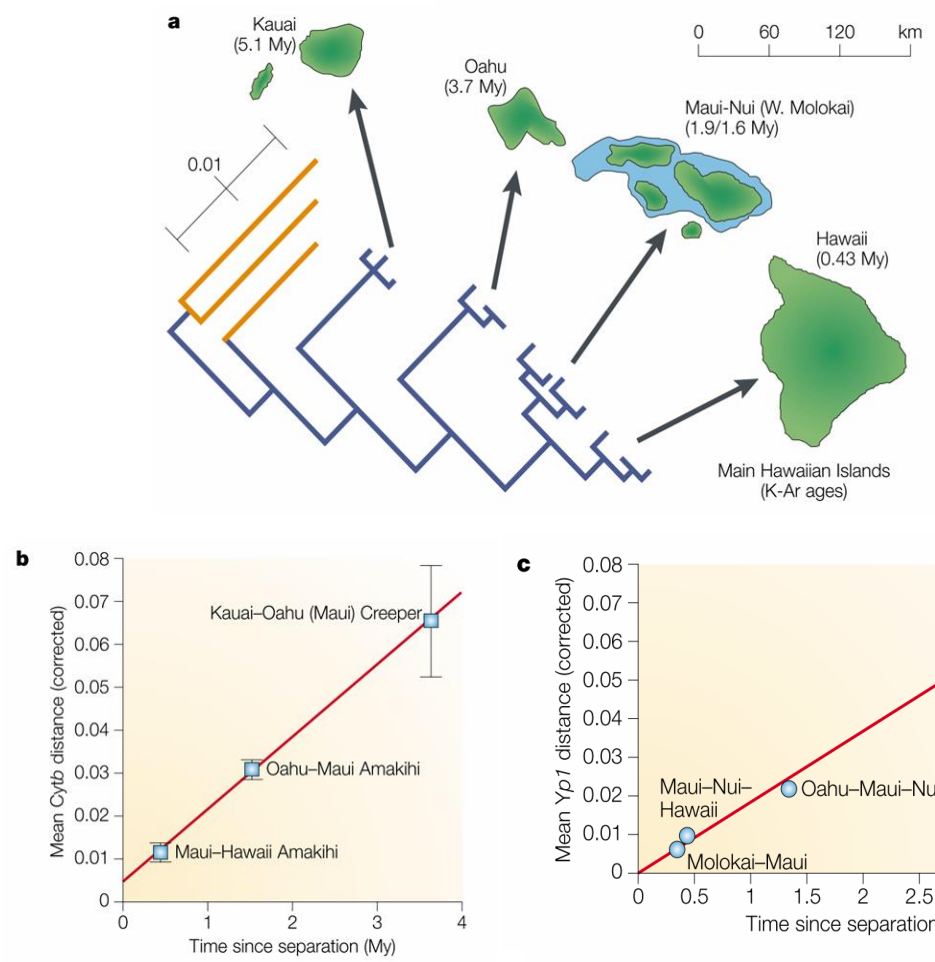
[59] T. Menezes, M. M. Romeiras, M. M. de Sequeira, e M. Moura, «Phylogenetic relationships and phylogeography of relevant lineages within the complex Campanulaceae family in Macaronesia», *Ecology and Evolution*, vol. 8, n. 1, pp. 88–108, 2018.

[60] D. C. Marshall *et al.*, «Inflation of Molecular Clock Rates and Dates: Molecular Phylogenetics, Biogeography, and Diversification of a Global Cicada Radiation from Australasia (Hemiptera: Cicadidae: Cicadettini)», *Syst. Biol.*, vol. 65, n. 1, pp. 16–34, Jan. 2016.

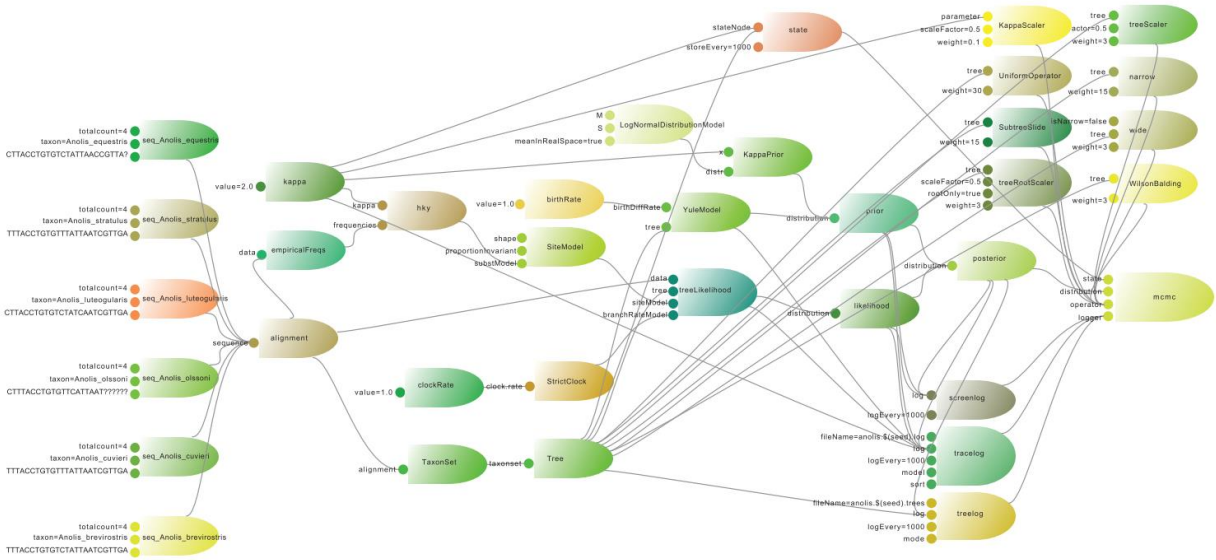
[61] M. Espeland *et al.*, «A Comprehensive and Dated Phylogenomic Analysis of Butterflies», *Current Biology*, vol. 28, n. 5, pp. 770-778.e5, Mar. 2018.

[62] C. Peña, S. Nylin, e N. Wahlberg, «The radiation of Satyrini butterflies (Nymphalidae: Satyrinae): a challenge for phylogenetic methods», *Zoological Journal of the Linnean Society*, vol. 161, n. 1, pp. 64–87, 2011.

8 Anexos



*Anexo 1 - Relógios moleculares das ilhas Havaianas. a: A origem vulcânica das ilhas Havaianas ocorreu com a formação de uma cadeia de ilhas com idades geológicas crescestes. As relações filogenéticas de aves insulares endémicas (como por exemplo espécies de drepananina: amakihi, Hemignathus virens e akiapolaau Hemignathus wilsoni; e moscas da fruta (Drosophila spp.)) refletem a "conveyer belt", com as espécies das primeiras ilhas mais antigas formarem os ramos mais internos da árvore evolutiva, e as ilhas mais recentes nas pontas. As "branches" as laranjas representam os outgroups. b,c : Datas moleculares para a espécie Hemignathus (**b**) e Drosophila (**c**) confirmam a ordem de colonização, e infere uma relação linear entre as distâncias da divergência genética e o tempo de separação em relação à idade da ilha. My, "million years" (milhões de anos); Fonte [41] reproduzidas com a permissão da REF. 10 © (1998) Blackwell Publishing.*



Anexo 2 - Workflow com seis sequências, tendo por uso o modelo de substituição HKY, strict clock, Yule tree prior, e ficheiros de logs que produzem o output para o Tracer, a árvore e o "screen output" através do processo da MCMC. doi:10.1371/journal.pcbi.1003537.g001. Fonte [45]

Materiais de estudo

http://beast.community/tree_priors

http://beast.community/rates_and_dates

http://beast.community/tip_dates

http://beast.community/treeannotator

http://beast.community/analysing_beast_output

https://github.com/Taming-the-BEAST/Substitution-model-averaging

http://beast.community/ess_tutorial

http://beast.community/logcombiner

https://academic.oup.com/ve/article/3/2/vex025/4100592

https://journals.plos.org/ploscompbiol/article/file?id=info%3Adoi/10.1371/journal.pcbi.1003537.s004&type=supplementary

https://beast.community/second_tutorial

https://beast.community/tempest_tutorial

https://beast.community/phylodynamics_of_epidemic_influenza

https://beast.community/phylodynamics_of_seasonal_influenza

https://beast.community/continuous_traits

https://beast.community/taxon_sets

https://beast.community/tip_dates

https://beast.community/tip_date_sampling

https://beast.community/constructing_models

https://beast.community/analysing_beast_output

https://beast.community/tracer_convergence

https://beast.community/summarizing_trees

https://beast.community/custom_substitution_models

https://beast.community/hierarchical_models

https://beast.community/markov_jumps_rewards

https://beast.community/adaptive_mcmc

https://beast.community/model_averaging_clocks

https://beast.community/time_dependent_rate_model

https://taming-the-beast.org/tutorials/StarBeast-Tutorial/

https://taming-the-beast.org/tutorials/StarBeast-Tutorial/StarBeast-Tutorial.pdf

Anexo 3 – Lista de links dos materiais de estudo principais usados.

Título da Publicação	Ref.	Ano de Publicação	Espécies de Estudo	Localização	Escala de Tempo
“The role of vicariance vs. dispersal in shaping genetic patterns in ocellated lizard species in the western Mediterranean”	[54]	2008	*Lagartos ocelados*	Ocidente do Mediterrâneo	Milhões de anos (Ma)
“Ancient origin of a Western Mediterranean radiation of subterranean beetles”	[55]	2010	Escaravelhos subterraneos da tribo *Leptodirini (Coleoptera, Leiodidae, Cholevinae)*	Cordilheiras Ocidente do Mediterrâneo	Milhões de anos (Ma)
“Recent origin, active speciation and dispersal for the lichen genus Nephroma (Peltigerales) in Macaronesia”	[56]	2011	*Gene Nephroma (Peltigerales) do Lichen*	Cosmopolitano com foco nas ilhas da Macaronésia: Açores, Madeira, Ilhas Canárias	Milhões de anos (Ma)
“Phylogeography and molecular phylogeny of Macaronesian island Tarphius (Coleoptera: Zopheridae): why are there so few species in the Azores?”	[57]	2012	*Tarphius (Coleoptera:Zopheridae)*	Macaronésia: Açores, Madeira, Ilhas Canárias	Milhões de anos (Ma)
“Late Miocene origin of an Ibero-Maghrebian clade of ground beetles with multiple colonizations of the subterranean environment”	[58]	2014	*Escaravelhos da espécie Trechus fulvus*	Paleártico Ocidental, com foco na área entre o sudeste da Península Ibérica e o Norte de Marrocos.	Milhões de anos (Ma)
“Phylogenetic relationships and phylogeography of relevant lineages within the complex Campanulaceae family in Macaronesia”	[59]	2017	Família das *Campanulaceae*	Macaronésia: Açores, Madeira, Ilhas Canárias	Milhões de anos atrás (Ma)

Anexo 4 - Lista de publicações estudadas com adequação ao projeto.

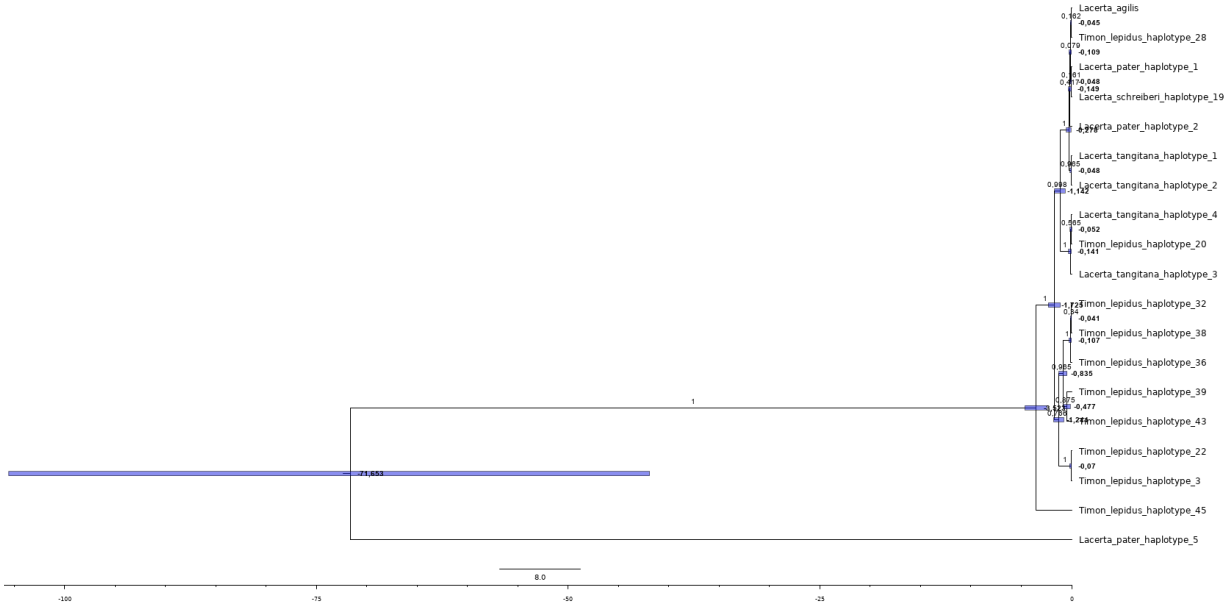
Traces:

Statistic	Mean	ESS	Type
posterior	-5890.725	2129	R
prior	83.81	2886	R
likelihood	-5974.536	686	R
treeModel.rootHeight	0.242	4701	R
tmrca(untitled1)	0.242	4701	R
birthDeath.meanGrowthRate	8.657	7853	R
birthDeath.relativeDeathRate	0.455	3652	R
ac	8.31E-2	5151	R
ag	0.759	1502	R
at	2.442E-2	1916	R
cg	3.803E-2	987	R
gt	7.588E-3	592	R
frequencies1	0.379	1541	R
frequencies2	8.21E-2	1891	R
frequencies3	9.365E-2	1965	R
frequencies4	0.445	1520	R
alpha	1.391	3703	R
pInv	0.606	2632	R
ucld.mean	0.665	5033	R
ucld.stdev	0.251	388	R
meanRate	0.647	4530	R
coefficientOfVariation	0.253	385	R
covariance	9.831E-3	5886	R
treeLikelihood	-5480.211	686	R
branchRates	-494.325	-	*
speciation	90.425	4235	R

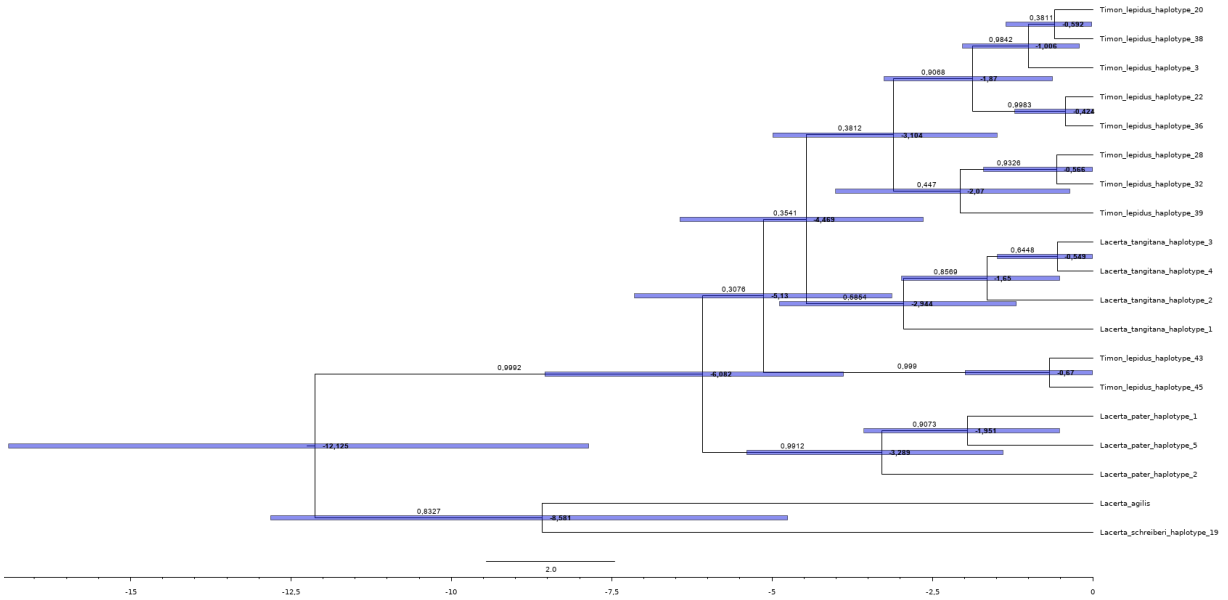
Anexo 5 - Resultado dos valores de convergência dos parâmetros usados na calibração da MCMC do dataset MRK, analisados pelo software Tracer; executado com BEAST.

Statistic	Mean	ESS	Type
species.popMean	6.386E-2	420	R
speciesTree.splitPopSize1	0.376	402	R
speciesTree.splitPopSize2	0.504	419	R
speciesTree.splitPopSize3	0.252	987	R
speciesTree.splitPopSize4	0.191	1146	R
speciesTree.splitPopSize5	0.253	994	R
speciesTree.splitPopSize6	8.011E-2	1831	R
speciesTree.splitPopSize7	7.931E-2	1795	R
speciesTree.splitPopSize8	8.825E-2	1938	R
speciesTree.splitPopSize9	8.112E-2	2056	R
speciesTree.splitPopSize10	8.182E-2	1730	R
speciesTree.splitPopSize11	8.905E-2	1621	R
speciesTree.splitPopSize12	7.973E-2	2084	R
speciesTree.splitPopSize13	7.876E-2	1860	R
species.birthDeath.meanGrowth...	8.241	1012	R
species.birthDeath.relativeDeath...	0.471	1580	R
speciesTree.rootHeight	0.107	354	R
treeModel.rootHeight	0.153	266	R
ac	8.221E-2	2553	R
ag	0.74	601	R
at	2.45E-2	701	R
cg	3.823E-2	866	R
gt	6.404E-3	203	R
frequencies1	0.377	922	R
frequencies2	8.174E-2	569	R
frequencies3	9.544E-2	528	R
frequencies4	0.445	828	R
alpha	1.244	1106	R
pInv	0.606	1207	R
ucld.mean	1.049	249	R
ucld.stdev	0.174	262	R
meanRate	1.038	249	R
coefficientOfVariation	0.175	260	R
covariance	2.056E-3	4273	R
treeLikelihood	-5484.396	493	R
branchRates	-494.325	-	*

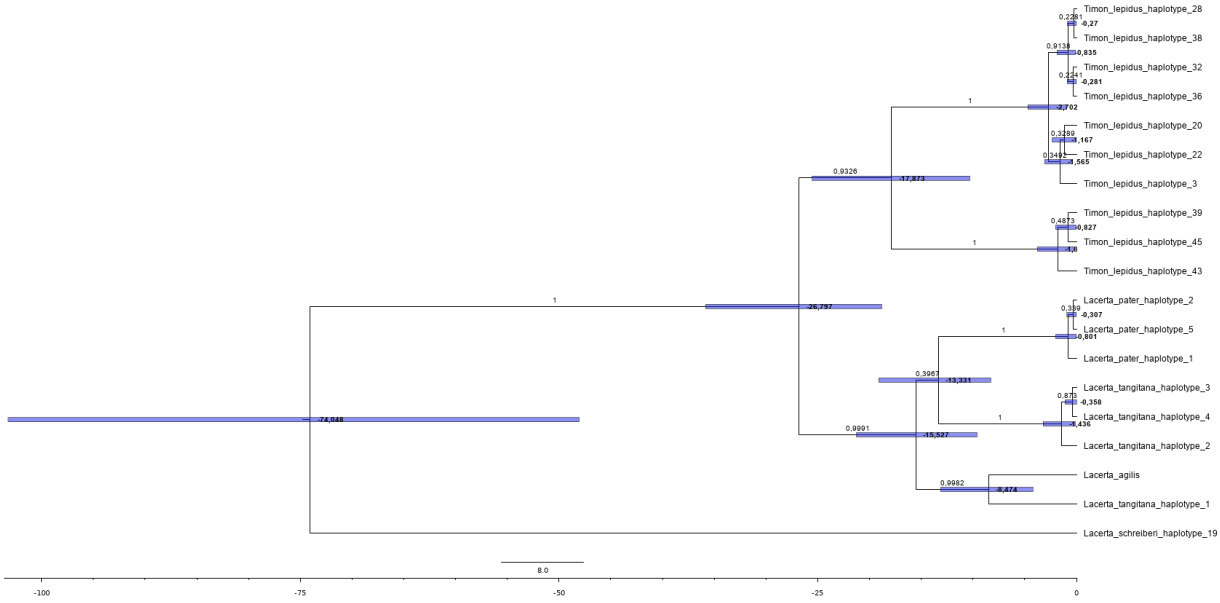
*Anexo 6 – Resultado dos valores de convergência dos parâmetros usados na calibração da MCMC do dataset MRK, analisados pelo software Tracer; executado com *BEAST.*



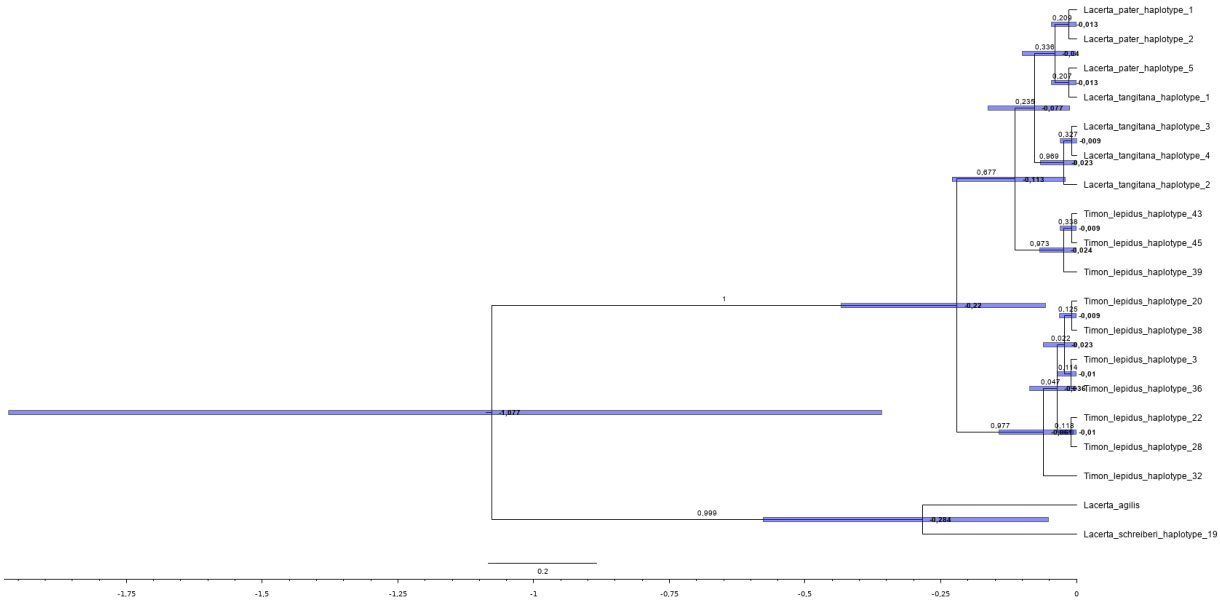
Anexo 7 – Árvore filogenética resultante da calibração individual do gene mitocondrial 12S, de acordo com os parâmetros apresentados na Tabela 8, do dataset concatenado dos lagartos ocelados.



Anexo 8 - Árvore filogenética resultante da calibração individual do gene mitocondrial 16S, de acordo com os parâmetros apresentados na Tabela 8, do dataset concatenado dos lagartos ocelados.

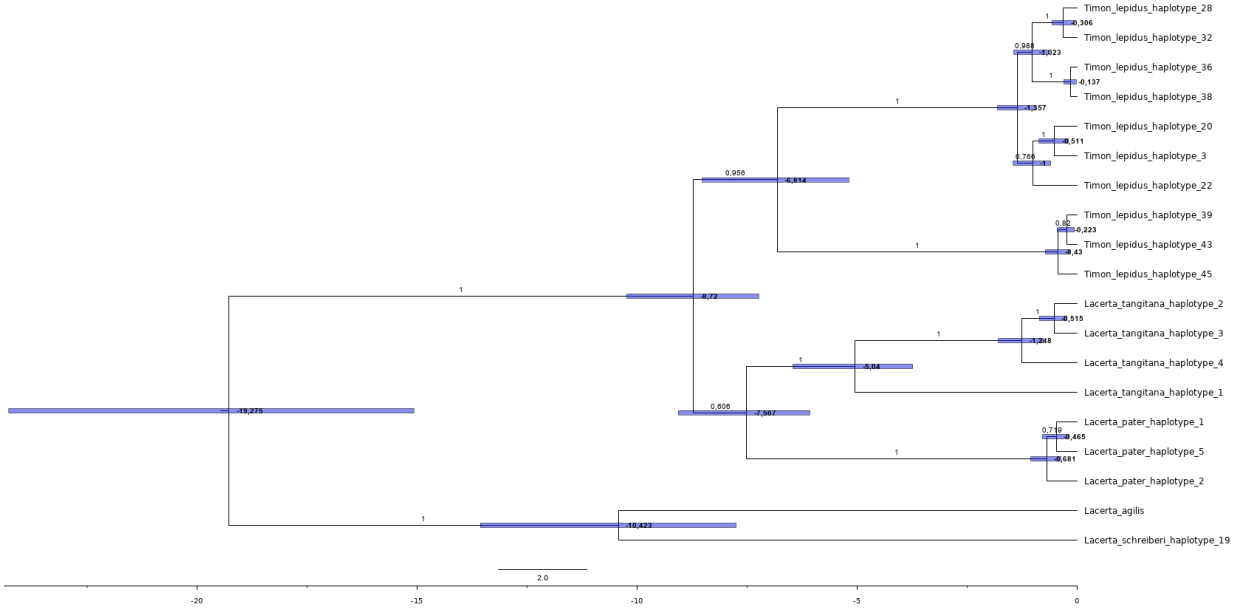


Anexo 9 - Árvore filogenética resultante da calibração individual do gene nuclear β Fibrinogen, de acordo com os parâmetros apresentados na Tabela 8, do dataset concatenado dos lagartos ocelados.

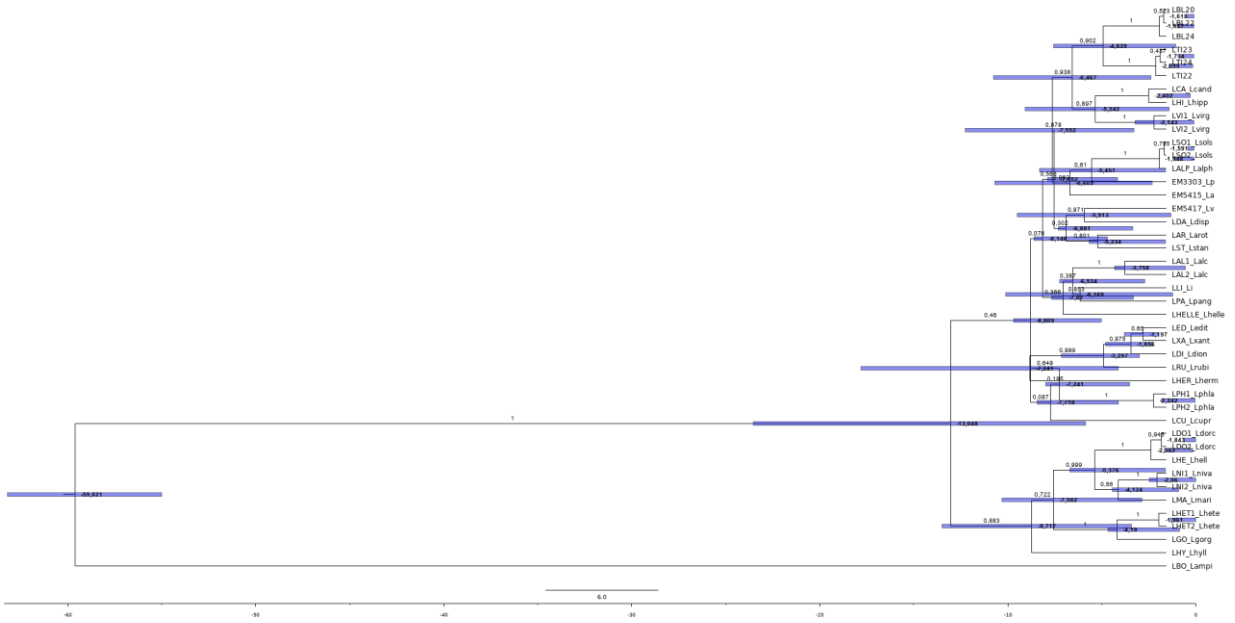


Anexo 10 - Árvore filogenética resultante da calibração individual do gene nuclear C-mos, de acordo com os parâmetros apresentados na Tabela 8, do dataset concatenado dos lagartos ocelados.

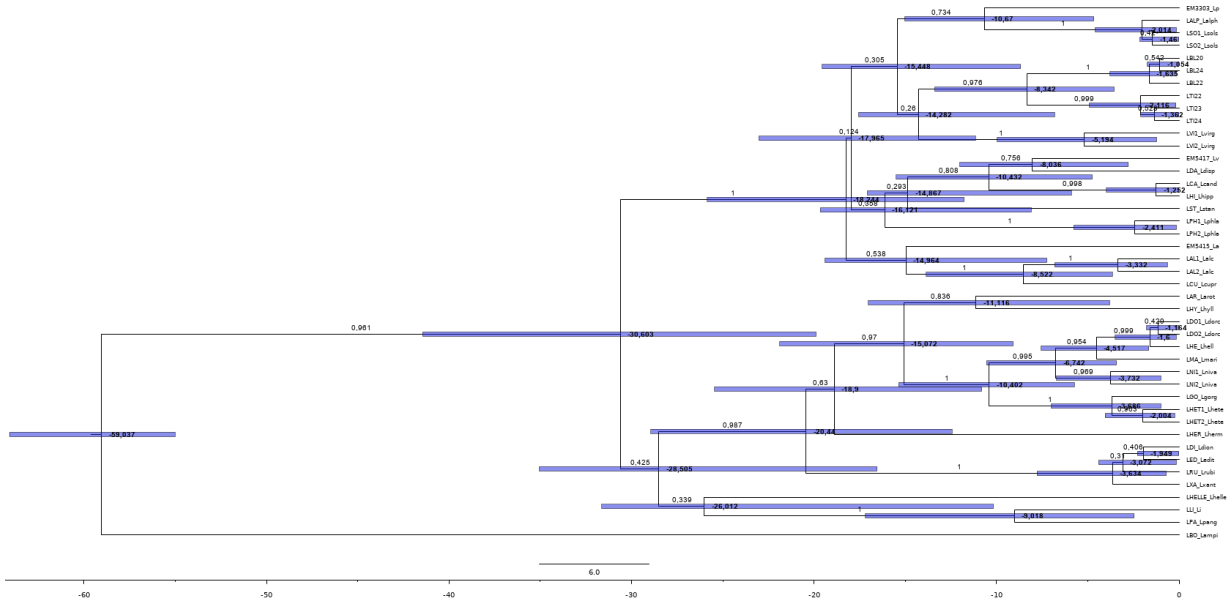
.



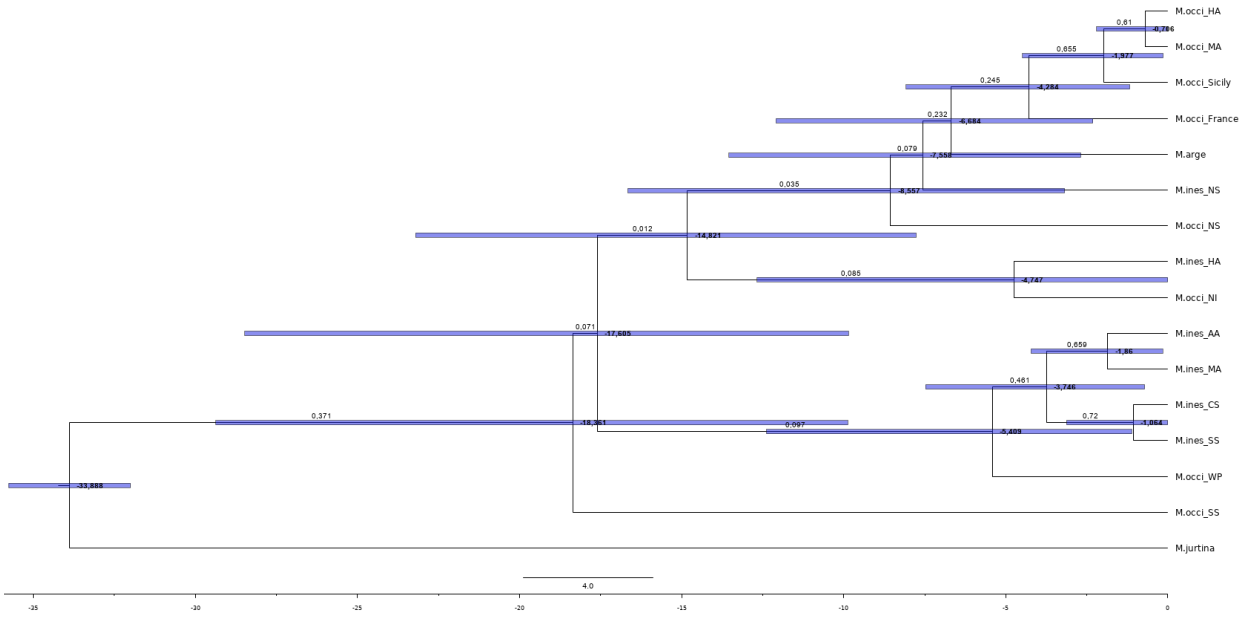
Anexo 11- Árvore filogenética resultante da calibração individual do gene mitocondrial Cytochrome b, de acordo com os parâmetros apresentados na Tabela 8, do dataset concatenado dos lagartos ocelados.



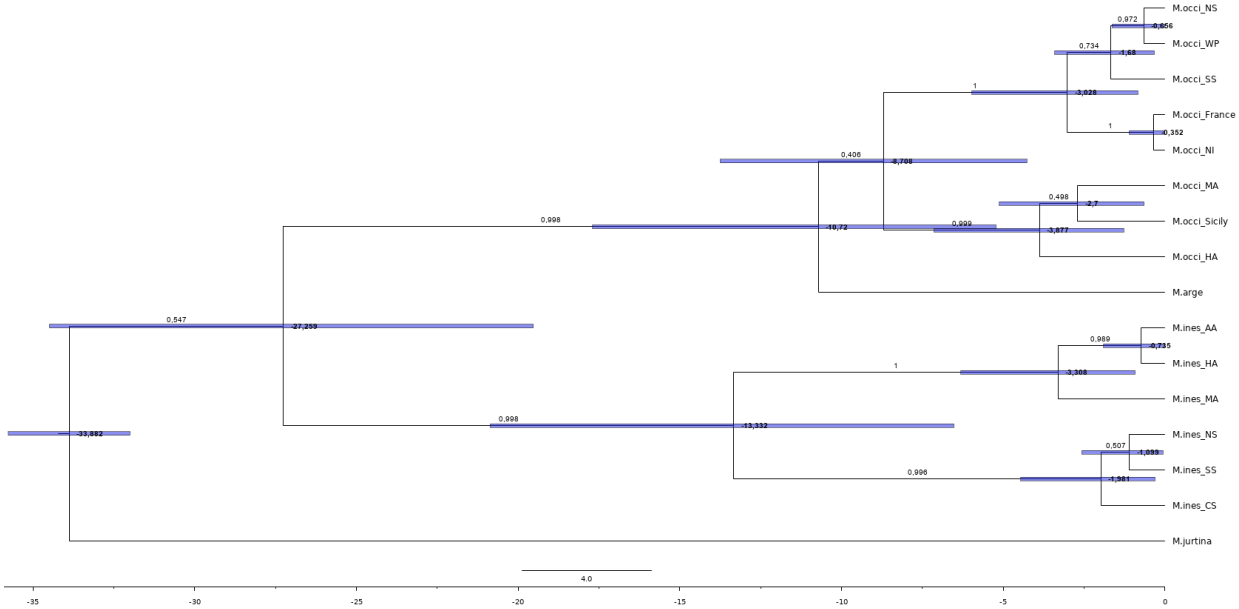
Anexo 12 - Árvore filogenética resultante da calibração individual do gene mitocondrial COI, de acordo com os parâmetros apresentados na Tabela 6, do género Lycaena.



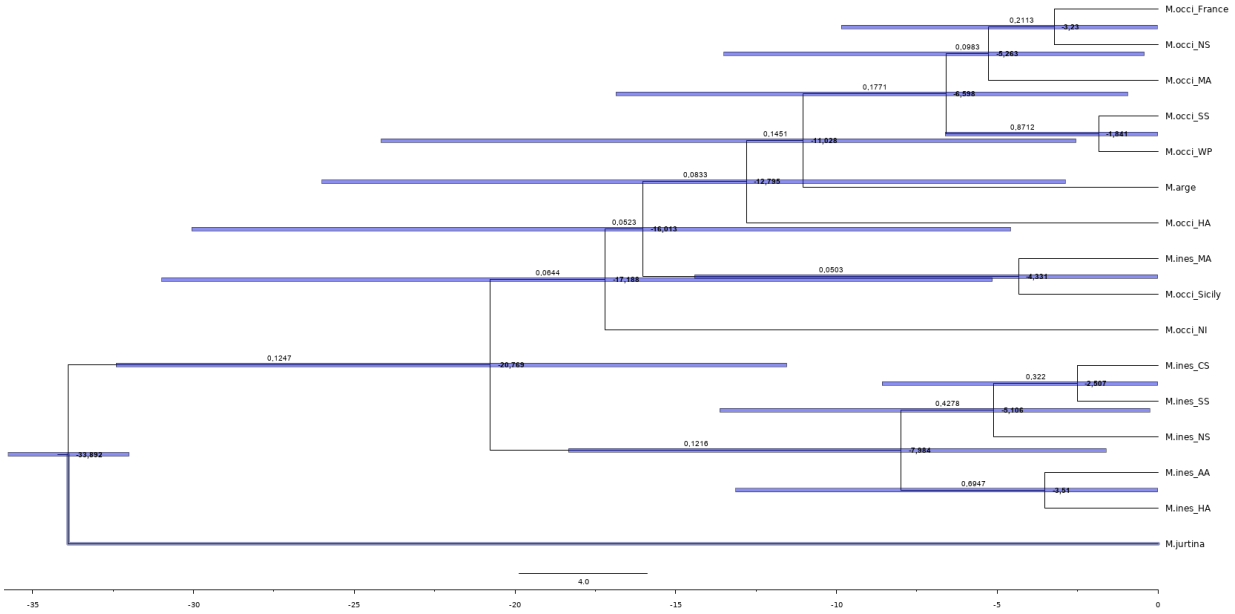
Anexo 13 - Árvore filogenética resultante da calibração individual do gene nuclear EF-1α, de acordo com os parâmetros apresentados na Tabela 6, do género Lycaena.



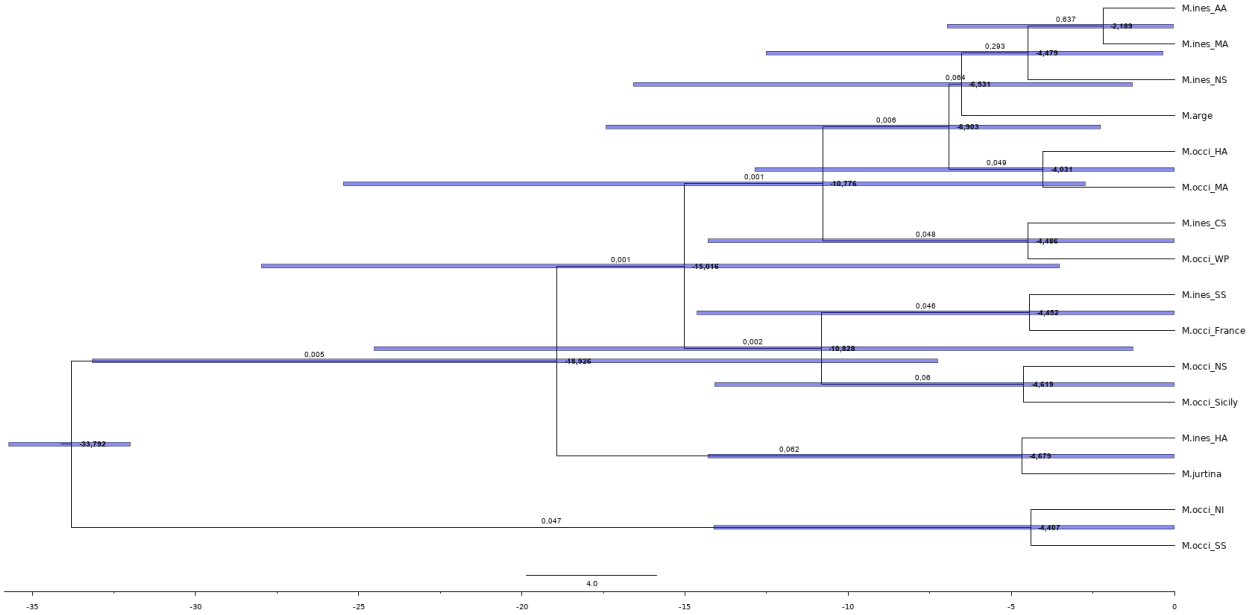
Anexo 14 - Árvore filogenética resultante da calibração individual do gene mitocondrial 16S, de acordo com os parâmetros apresentados na Tabela 7, do género Melanargia.



Anexo 15 - Árvore filogenética resultante da calibração individual do gene mitocondrial COI, de acordo com os parâmetros apresentados na Tabela 7, do género Melanargia.



Anexo 16 - Árvore filogenética resultante da calibração individual do gene nuclear EF-1α, de acordo com os parâmetros apresentados na Tabela 7, do género Melanargia.



Anexo 17 - Árvore filogenética resultante da calibração individual do gene nuclear Wg, de acordo com os parâmetros apresentados na Tabela 7, do género Melanargia.

9 Apêndices

```
# script efetch
wget https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi/\?db\=nucleotide\&id\= "ACCESSION NUMBERS DO NCBI SEPARADOS POR VIGULAS"\&rettype\=fasta -O ~/"PATH PARA A DIRETORIA DO SAVE".fasta
```

Apêndice 1 - Script efetch que faz o download dos alinhamentos das samples em formato fasta: Guardado num repositório online do GitHub com o endereço: https://github.com/ray2g/StarBeast/blob/master/efetch.sh conforme commit 29c3291.

```
# -*- coding: utf-8 -*-

import argparse

parser = argparse.ArgumentParser(prog='python3')


parser.add_argument("input_fasta", metavar="input.fasta", type= str,
                    help= "Fasta file to replace sequence names/description for a generate code, e.g. S001")

                     #input file parser


parser.add_argument("output_fasta", metavar="output.fasta", type= str,
                    help= "Output fasta file where the sequence description/name was changed to a generated code")

                     #output file parser


arguments = parser.parse_args()


def seq_number_gen(count):
  """
  generate the code for the sequence description assume
  """
    seq_number= str(count)
    while len(seq_number) < 3:
        seq_number = "0" + seq_number
    return seq_number


def changer(input_file, output_file):
  #change de sequence description/name to the generated code


    file_handle = open(input_file, 'r') #open the input file

    output= open(output_file, 'w') #open the output file

    dictio={} #dicionary creation, description | code

    count= 0 #count variable

    for line in file:

        if line.startswith('>'): #or line[0]=='>'

            count+= 1

            new_seqname = '>S' + seq_number_gen(count)+'\n'

            output.write(new_seqname) # write on the new output file the trade of description | generated code


        else:                        # if the line dont start with ">" write the the sequence normaly
            output.write(line)

            """
            aggregate the sequence description with the correnpondent generated code
            dictio[new_seqname]=line[1:]
            """

changer(arguments.input_fasta,arguments.output_fasta)
```

Apêndice 2 – Script que troca o nome das samples de um ficheiro em formato Fasta para um código incremental. Guardado num repositório online do GitHub com o endereço: https://github.com/ray2g/StarBeast/blob/master/namefastaseq_changer_2gencode.py conforme commit 7593295.