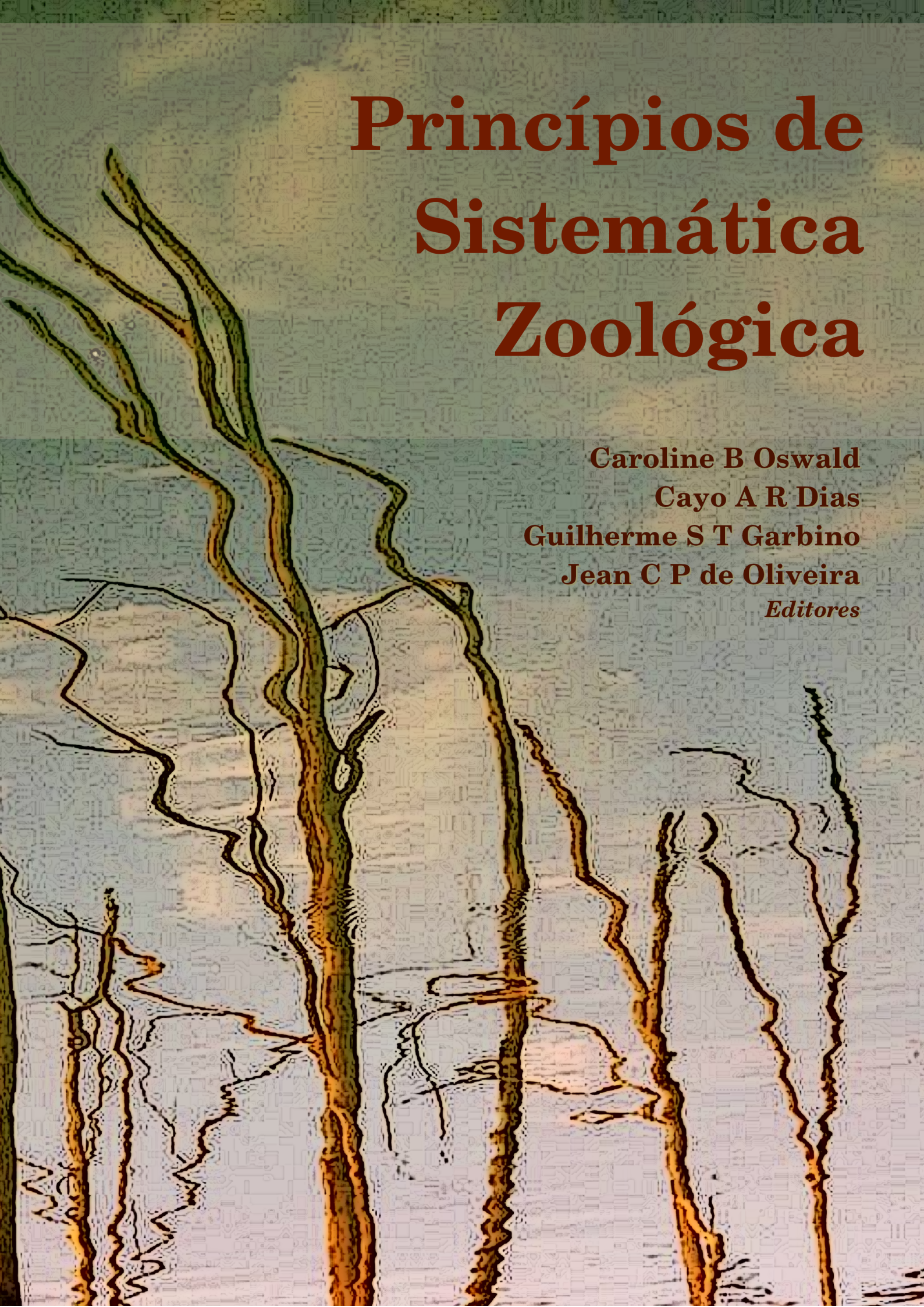
Princípios de Sistemática Zoológica

Caroline B Oswald
Cayo A R Dias
Guilherme S T Garbino
Jean C P de Oliveira
Editores



Caroline Batistim Oswald
Cayo Augusto Rocha Dias
Guilherme Siniciato Terra Garbino
Jean Carlo Pedroso de Oliveira
(Organizadores)

PRINCÍPIOS DE SISTEMÁTICA ZOOLÓGICA

Belo Horizonte, 2020

Organizadores:
Caroline Batistim Oswald
Universidade Federal de Minas Gerais, Programa de Pós-Graduação em Zoologia
e-mail: carolbatistim@gmail.com
Cayo Augusto Rocha Dias
Universidade Federal de Minas Gerais, Programa de Pós-Graduação em Zoologia
e-mail: cayodias@gmail.com
Guilherme Siniciato Terra Garbino
Universidade Federal de Minas Gerais, Programa de Pós-Graduação em Zoologia
e-mail: gstgarbino@hotmail.com
Jean Carlo Pedroso de Oliveira
Universidade Federal de Minas Gerais, Programa de Pós-Graduação em Genética
e-mail: pedrosojco@gmail.com

Capa: Ivan L. F. Magalhães (fotografia), Jean C. P. de Oliveira (edição)

Ficha catalográfica elaborada por Fabiane C. M. Reis – CRB: 6/2680

C977p

Curso de Verão em Sistemática Zoológica (CVSZ) do Programa de Pós-Graduação em Zoologia da UFMG (Belo Horizonte, MG : 2018)

Princípios de sistemática zoológica: material de apoio para o I CVSZ / Caroline Batistim Oswald (org.)... [et al.] . 1ª ed. – Belo Horizonte : PGZoo UFMG, 2020.

xiv, 77 p.

ISBN: 978-65-00-08035-3

Outros organizadores: Cayo Augusto Rocha Dias; Guilherme Siniciato Terra Garbino; Jean Carlo Pedroso de Oliveira.

1. Zoologia. I. Oswald, Caroline Batistim. II. Título. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas.

CDD:591
CDU:591

https://cursozooufmg.wordpress.com/

Apresentação

Para grande parte do público, não formalmente introduzido às Ciências Biológicas, a Zoologia remete à conservação da fauna ou o estudo sobre comportamento, especialmente de grandes vertebrados. No entanto, esse ramo da Biologia é muito mais complexo e diverso. Zoólogos estudam a conservação da fauna e o seu comportamento, tanto de vertebrados quanto de invertebrados, mas também a interação destes com outros seres vivos e o ambiente, bem como a sua evolução. Além disso, para uma análise holística, as pesquisas em Zoologia englobam aspectos que esbarram em outras disciplinas, como a Morfologia, Genética, Fisiologia, Embriologia, entre outras. A base de todos esses fascinantes ramos do conhecimento ligados ao estudo da fauna é a sistemática zoológica, que se ocupa em reconhecer e classificar as espécies animais, estimar as suas relações de parentesco e sua evolução no tempo e no espaço. Dessa forma, um conhecimento básico de sistemática é primordial para todos que se interessam por Zoologia.

A sistemática é um campo do conhecimento que não é amplamente abordado nos ciclos básicos da educação, e muitos cursos de graduação em Ciências Biológicas não possuem uma disciplina exclusivamente voltada para o tema. A ausência dessa disciplina tão importante gera uma grande lacuna na compreensão da Zoologia, já que apesar de ser um campo básico do conhecimento zoológico, a sistemática possui um extenso jargão técnico e utiliza de metodologias que necessitam de um amplo conhecimento de outras áreas, como por exemplo, bioquímica, estatística e bioinformática.

Para diminuir essa lacuna, algumas instituições de ensino passaram a oferecer cursos concentrados de treinamento e atualização na área, voltados especialmente para graduandos ou recém-formados do curso de Ciências Biológicas e áreas correlatas. Nesse contexto, o programa de Pós-Graduação em Zoologia da Universidade Federal de Minas Gerais (PGZoo UFMG) tomou a iniciativa de realizar o Curso de Verão em Sistemática Zoológica (CVSZ), que teve sua primeira edição em janeiro de 2018. Embora a ideia da criação do curso tenha partido dos docentes do programa, o planejamento e execução foram conduzidos pelos discentes. Desta forma, ainda que o objetivo primário do CVSZ fosse introduzir as ferramentas da pesquisa em sistemática zoológica aos interessados pela área, diversos outros benefícios vieram desta iniciativa: podemos salientar o importante envolvimento dos pós-graduandos na prática da docência, a experiência de organizar um evento de extensão, e o contato com estudantes de diferentes estados do Brasil. Com o sucesso do primeiro evento, a segunda edição do CVSZ foi realizada no início de 2020.

O escopo dessa experiência aumenta agora, com a edição deste guia, Princípios de Sistemática Zoológica. Os capítulos compilados aqui foram pensados para servir como material de apoio para o I CVSZ. Eles foram escritos pelos pós-graduandos e residentes de pós-doutorado do nosso programa, que trabalham na área de sistemática zoológica e que ministraram as aulas na primeira edição do curso. O resultado final mostrou um potencial maior, com textos ricamente ilustrados que servem tanto para uma introdução teórica, quanto para guiar nos primeiros passos nas análises utilizadas nos estudos de sistemática. Por essa razão, a organização do CVSZ tomou mais essa iniciativa, de democratização do conhecimento, tornando público o conteúdo do curso reunido nesta publicação. Os nove capítulos incluem temas básicos, começando com métodos para a coleta de material zoológico e coleções biológicas, passando pela taxonomia, classificação, nomenclatura, chaves de identificação interativas, evolução e processos biológicos. São também abordados métodos mais específicos da área, como análise filogenética com dados morfológicos e moleculares e métodos filogenéticos comparativos. Em cada capítulo, há uma bibliografia recomendada, permitindo ao leitor aprofundar-se no tema sem perder-se na imensidão da literatura disponível.

O CVSZ e a publicação desse volume, juntamente com a organização bianual do Simpósio em Zoologia Sistemática pela PGZoo UFMG têm ajudado a consolidar o nosso programa como um importante centro de ensino e pesquisa em sistemática zoológica no país. Apesar da PGZoo UFMG ser um programa novo, com o início de sua história em 2011, essas iniciativas são demonstrações da sua capacidade de atração de estudantes interessados em sistemática e da qualidade do sistema de pós-graduação pública e da ciência produzida no Brasil. A proposta da PGZoo UFMG é oferecer este curso periodicamente e esperamos que essa publicação contribua na compreensão dos vários aspectos da sistemática zoológica, auxiliando na formação de futuros zoólogos.

Kirstern Lica Follmann Haseyama
Gisele Yukimi Kawauchi

Prefácio

A sistemática é a área mais fundamental e a mais inclusiva das Ciências Biológicas. Mais fundamental porque é pré-requisito para uma comunicação clara e objetiva entre as demais áreas. Mais inclusiva porque o sistemata frequentemente compara e sintetiza conhecimentos de muitas áreas, de ecologia à genética, para produzir classificações. Apesar da óbvia importância, ainda estamos longe de um conhecimento adequado e sistematizado da biota terrestre. Esse déficit de conhecimento sobre a real diversidade das espécies do planeta, conhecido como déficit lineano, é particularmente pronunciado num país megadiverso como o Brasil. Em 2006, existiam 22 programas de pós-graduação em Zoologia no Brasil, a maioria em São Paulo, Rio de Janeiro e Paraná. Em 2011 inicia-se o programa de pós-graduação em Zoologia da Universidade Federal de Minas Gerais (UFMG), que ao longo de nove anos de existência vem contribuindo para a formação de zoólogos e consequentemente para a mitigação do déficit lineano.

Com o objetivo de apresentar os fundamentos teóricos e práticos da sistemática zoológica, organizamos o I Curso de Verão em Sistemática Zoológica da UFMG em 2018. Este Guia, **Princípios de Sistemática Zoológica** é oferecido como material suplementar aos alunos do curso e também será disponibilizada *online* para qualquer interessado no assunto. Nosso intuito com esse material é fornecer as bases teóricas da taxonomia e sistemática zoológica e também rudimentos em alguns programas analíticos. Os primeiros capítulos (1 a 4) apresentam princípios de sistemática filogenética (morfológica e molecular), taxonomia, e nomenclatura zoológica, assim como um breve histórico do desenvolvimento dessas áreas. Os capítulos 5 e 6 tratam das inúmeras aplicações de filogenias, tanto na sistemática como fora dela. Finalmente, os capítulos de 7 a 9 apresentam métodos de coleta, identificação e depósito de material zoológico.

Caroline Batistim Oswald
Cayo Augusto Rocha Dias
Guilherme Siniciato Terra Garbino
Jean Carlo Pedroso de Oliveira

Agradecimentos

A elaboração do **Princípios de Sistemática Zoológica** é fruto de um esforço conjunto de muitos pesquisadores, e por isso não poderíamos deixar de agradecê-los. A realização da primeira edição do Curso de Verão em Sistemática Zoológica da UFMG (CVSZ), evento que deu origem a este Guia, só foi possível graças ao incentivo da professora Dra. Kirsten Lica Haseyama e do trabalho conjunto de inúmero alunos e ex-alunos do programa de pós-graduação em Zoologia (PGZoo) da Universidade Federal de Minas Gerais (UFMG). Contamos ainda com a participação de alunos do programa de pós-graduação em Genética (PGGen), somando esforços para a realização do evento e material de apoio. Nesse sentido, somos gratos aos alunos e ex-alunos da PGZoo e da PGGen, que contribuíram para a escrita dos capítulos, ministraram aulas e auxiliaram na organização do curso.

Gostaríamos de agradecer também a professora Dra. Gisele Yukimi Kawauchi e o professor Dr. Almir Rogério Pepato, docentes do departamento de Zoologia da UFMG, que tornaram possível a realização da segunda edição do CVSZ. Estendemos nossos agradecimentos à Fundação de Desenvolvimento da Pesquisa (Fundep), pelo valoroso auxílio na organização desse evento.

Por último, agradecemos aos revisores, cujas sugestões e leitura crítica contribuíram significativamente para a melhora e consolidação dos capítulos: Marta Svartman, José Eduardo Serrano Villavicencio, Kirsten Lica Follmann Haseyama, Guilherme Henrique Fernandes de Azevedo, Marcelo Rodrigues Nogueira, Diogo Borges Provete, Sandra Ludwig, Livia Echternacht Andrade, Paulo Durães Pereira Pinheiro, Tiago Leite Pezzuti, Nancy França Lo Man Hung, Alice Fumi Kumagai e Priscila Guimarães Dias.

Caroline Batistim Oswald
Cayo Augusto Rocha Dias
Guilherme Siniciato Terra Garbino
Jean Carlo Pedroso de Oliveira

Colaboradores

Alessandro Rodrigues Lima
Universidade Federal de Minas Gerais, Centro de Coleções Taxonômicas
e-mail: alerolima@gmail.com

Bárbara Teixeira Faleiro
Universidade Federal de Minas Gerais, Programa de Pós-Graduação em Zoologia
e-mail: btf8@hotmail.com

Cayo Augusto Rocha Dias
Universidade Federal de Minas Gerais, Programa de Pós-Graduação em Zoologia
e-mail: cayodias@gmail.com

Daniel M. Casali
Universidade Federal de Minas Gerais, Programa de Pós-Graduação em Zoologia
e-mail: daniel_casali@yahoo.com.br

Daniela Lidia Nuñez Rodriguez
Universidade Federal de Minas Gerais, Programa de Pós-Graduação em Genética
e-mail: dnunezrodriguez@gmail.com

Guilherme S. T. Garbino
Universidade Federal de Minas Gerais, Programa de Pós-Graduação em Zoologia
e-mail: gstgarbino@hotmail.com

José Eustáquio dos Santos Júnior
Universidade Federal de Minas Gerais, Programa de Pós-Graduação em Genética
e-mail: jrsantos140782@yahoo.com.br

Larissa C. C. S Dumbá
Universidade Federal de Minas Gerais, Programa de Pós-Graduação em Zoologia
e-mail: larissa.dumba@gmail.com

Leonardo Sousa Carvalho
Universidade Federal do Piauí, *Campus* Amílcar Ferreira Sobral
e-mail: carvalho@ufpi.edu.br

Rafael Félix de Magalhães
Universidade Federal de São João del-Rei, *Campus* Dom Bosco
e-mail: rafaelfelixm@gmail.com

Rafaela V. Missagia
Universidade Federal de Minas Gerais, Programa de Pós-Graduação em Zoologia
e-mail: rafaelamissagia@gmail.com

Sumário

Capítulo 1
Evolução e Processos evolutivos

Rafael Félix de Magalhães

1.1 Introdução

A biologia evolutiva é a disciplina que sintetiza todo o conhecimento gerado na área das ciências biológicas. Evolução em biologia é sinônimo de mudança e frequentemente é categorizada de acordo com a escala temporal em que as mudanças orgânicas ocorrem nos seres vivos. O termo microevolução está historicamente associado aos processos evolutivos registrados em uma escala de tempo muito pequena, geralmente no âmbito de populações ou entre espécies proximamente aparentadas entre si. Alguns biólogos argumentam que todas as mudanças evolutivas têm como base os processos de microevolução. Dentre estes processos, podemos citar a seleção natural, a deriva genética e as migrações interpopulacionais como modificadores de frequências alélicas ao longo do tempo.

Por outro lado, investigações sobre os padrões evolutivos associados a longas escalas temporais (por exemplo: surgimento, evolução e extinção de clados ao longo de milhões de anos) e sistemáticas (isto é, acima do nível de espécie) estão classificados no universo dos estudos macroevolutivos. Os dois termos, ainda utilizados por muitos cientistas, foram propostos pelo zoólogo russo Yuri Filipchenko, um dos pioneiros na incorporação da genética Mendeliana à teoria evolutiva. Entretanto, muitos biólogos, incluindo o evolucionista Stephen J. Gould, começaram a criticar esta hierarquia reducionista. Estes pensadores defendiam uma visão plural, segundo a qual os chamados processos micro- e padrões macroevolutivos operam tanto a nível intraespecífico quanto a níveis supraespecíficos. Em resumo, a macroevolução pode ser interpretada como o acúmulo de processos microevolutivos.

Neste capítulo, abordaremos resumidamente os pontos da teoria evolutiva, essenciais para a aplicação e interpretação dos dados em sistemática filogenética. Utilizaremos os termos micro- e macroevolução apenas para fins didáticos.

1.2 Um breve histórico do desenvolvimento da teoria evolutiva

Diversos filósofos gregos, incluindo Aristóteles, Empédocles e Xenófanes, já conheciam o registro fóssil e o atribuíam a indícios da vida passada, supostamente extinta por catástrofes naturais. Entretanto, nenhum deles formalizou um conceito de evolução. Este quadro não mudou durante a Idade Média, já que as interpretações literais da bíblia foram instituídas como dogmas da fé cristã, desencorajando a investigação racional do mundo natural, o que poderia ser encarado como heresia naquela época.

Somente no início do século XIX, mais precisamente em 1809, o naturalista francês Jean-Baptiste Lamarck propôs em seu livro *Système des Animaux sans Vertèbres* a primeira explicação para a origem e evolução dos seres vivos. Sua teoria, conhecida como **herança dos caracteres adquiridos**, assumia que a evolução era um processo transformacional. Isto quer dizer que os organismos mudavam suas características fenotípicas ao longo do tempo, de acordo com as pressões ambientais, o uso e o desuso das partes do corpo. Lamarck não acreditava na ancestralidade comum dos organismos e postulava que os caracteres adquiridos seriam transmitidos para os descendentes de cada ancestral independente.

Mais tarde no mesmo século, entre os anos de 1830 e 1833, o geólogo britânico Charles Lyell propôs o princípio do uniformitarismo. Segundo ele, as leis da física e da química são imutáveis e os processos geológicos observados no presente devem ser réplicas dos eventos ocorridos ao longo da história da Terra. Suas ideias foram formalizadas nos três volumes do seu livro *Principles of Geology: being an attempt to explain the former changes of the Earth's surface, by reference to causes now in operation*. Nesta mesma época, um jovem naturalista conterrâneo de Lyell iniciava uma viagem de cinco anos ao redor do mundo a bordo de um navio da marinha britânica. Este navio se chamava *HMS Beagle* e tinha como objetivo realizar prospecções científicas marinhas e terrestres na América do

Sul e no Oceano Pacífico. O jovem cientista era Charles Darwin, que carregava consigo o primeiro volume do *Principles of Geology*.

A viagem no *Beagle*, somada à leitura do livro de Lyell, teve profundo impacto sobre as ideias de Darwin. Ele ficou convencido de que as forças naturais poderiam explicar a diversidade de formas dos organismos e as características geológicas da Terra. Darwin coletou e escreveu sobre a fauna e flora vivente e fóssil das regiões onde o *Beagle* aportou, mas foi a parada em setembro de 1835 nas Ilhas Galápagos, território do Equador, que teve grande influência na consolidação das suas ideias. O naturalista passou pelas ilhas de Cabo Verde, ao noroeste da África e, apesar deste arquipélago possuir clima e topografia similares aos das Galápagos, ambos possuíam biotas muito distintas. As espécies encontradas no arquipélago equatoriano eram semelhantes àquelas encontradas no continente sul-americano. Além disso, cada ilha possuía, por exemplo, espécies endêmicas de tentilhões e tartarugas gigantes, cada qual aparentada às espécies das outras ilhas. Com isso, Darwin concluiu que as espécies de Galápagos deveriam ser originárias de um ancestral oriundo do continente. Além disso, as formas distintas de cada ilha deveriam ter surgido através de modificações relacionadas às condições ambientais singulares de cada ilha. Em 1858, Darwin recebeu um manuscrito de um outro naturalista britânico. Este cientista, Alfred Russel Wallace, durante uma expedição ao arquipélago Malaio, chegou independentemente às mesmas conclusões que Darwin sobre a evolução dos organismos, conforme reportou nesta correspondência. Em julho daquele mesmo ano, Darwin preparou um manuscrito com suas perspectivas sobre evolução que foi publicado junto com o de Wallace no *Journal of the Linnean Society*. Finalmente, em 1859, Darwin publicou um livro com suas ideias expandidas, rico em exemplos, denominado *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle of Life*, ou simplesmente A Origem das Espécies.

Diferente das ideias transformacionais de Lamarck, a teoria de Darwin e Wallace é variacional. Isto quer dizer que as espécies evoluem ao longo do tempo a partir de variantes selecionadas nas populações. Entretanto, Darwin não conhecia os meios pelos quais as características eram herdadas. Por isso, ele se apropriou das ideias de Lamarck para explicar a herança dos caracteres, segundo as quais a hereditariedade seria resultado de um processo de mistura das caraterísticas parentais em suas proles, incluindo traços adquiridos ao longo da vida. Esta perspectiva mudou no final do século XIX, quando August Weismann, um biólogo do desenvolvimento, demonstrou através de experimentos que as características adquiridas pelos organismos ao longo da vida não eram herdadas pelos seus descendentes, rejeitando a explicação lamarckista de herança. Por fim, na década de 1930, um grupo de geneticistas e biólogos populacionais, inspirados pela redescoberta dos trabalhos de Gregor Mendel no início daquele século, revisou a teoria de Darwin sob uma perspectiva matemática. Gradualmente, foram incorporadas a genética mendeliana e os avanços teóricos da genética populacional à teoria, que ficou conhecida como **Teoria Sintética da Evolução**, ou **Nova Síntese Evolutiva**, que foi, nas décadas seguintes, ampliada incorporando conceitos de outros campos da Biologia, tais como: a Botânica, Paleontologia e Sistemática. Dentre estes cientistas, estavam Ronald Fisher, John B. S. Haldane, Sewall Wright, Theodosius Dobzhansky, Ernst Mayr, George G. Simpson, Julian Huxley, Bernhard Rensch e George Ledyard Stebbins.

Darwin acreditava que os organismos vivos descendiam de um ancestral comum. Esta ideia foi ilustrada em sua famosa árvore filogenética, rascunhada em seu caderno pessoal (notebook B; Figura 1.1), tida como um ícone da teoria evolutiva. Entretanto, apenas em 1950, o entomólogo alemão Willi Hennig publicou o livro *Grundzüge einer Theorie der Phylogenetischen Systematik*, no qual formalizava um método objetivo para a estimativa das relações filogenéticas entre os organismos. Somente em 1966 seu trabalho se tornou amplamente conhecido, uma vez que foi traduzido para o inglês sob o título *Phylogenetic Systematics*. Desde então, o estudo da evolução e da sistemática filogenética tem se tornado um dos campos mais frutíferos das ciências naturais.

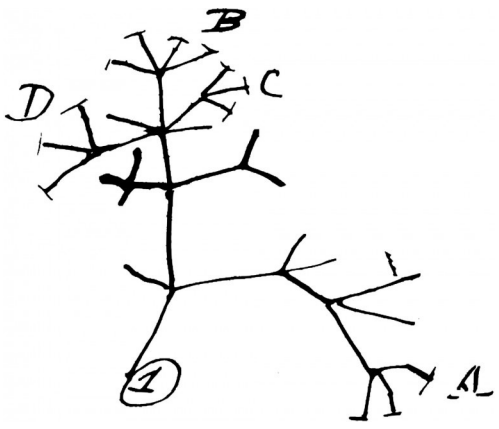


Figura 1.1 Esboço de Darwin sobre a inter-relação entre os seres vivos. Imagem de uso livre, acessível no Wikimedia Commons.

1.3 Evidências da evolução

Registro paleontológico

Uma importante premissa da teoria evolutiva é que os organismos são resultado do acúmulo de modificações hereditárias desde o passado até o presente. A paleontologia, uma ciência macroevolutiva, fornece importantes evidências que corroboram esta premissa. Os fósseis, objetos de estudo desta ciência, são remanescentes preservados da vida passada. A principal tarefa dos paleontólogos é datar e interpretar o registro fóssil sob a luz da teoria evolutiva.

A datação do registro fóssil pode ser feita de maneira relativa (isto é, através da interpretação estratigráfica das camadas de rochas ou de fósseis-guia) ou de maneira absoluta (isto é, através da radiometria aplicada ao decaimento radioativo de elementos naturais). Por exemplo, a presença de trilobitas fósseis em um estrato rochoso indica que a rocha data da era Paleozóica. Para determinar com exatidão a idade destas rochas, pode-se utilizar a datação por decaimento de Urânio em Chumbo ou de Potássio em Argônio, por exemplo.

O registro fóssil é enviesado. As partes mineralizadas do corpo de animais tais como ossos, conchas e carapaças são preservados com maior facilidade do que partes moles. Além disso, processos tafonômicos (isto é, relacionados ao modo de decomposição dos organismos) podem levar à separação das partes articuladas e à fragmentação, dentre outros. Em resumo, a maioria do registro fóssil é incompleto. Cabe aos paleontólogos identificar, interpretar e sistematizar seus objetos de estudos.

Apesar da incompletude, vários fósseis icônicos auxiliaram no entendimento da evolução da vida na Terra. Dentre eles, pode-se citar o *Tiktaalik roseae* (Figura 1.2 A), um peixe sarcopterígio do Período Devoniano (375Ma) que possuía características comuns entre peixes (por exemplo: escamas e barbatanas) e tetrápodes terrestres (por exemplo: cabeça achatada, pescoço e pulsos articulados), indicando que os primeiros tetrápodes devem ter surgido neste período. Outro fóssil importante é o *Archaeopteryx litographica* (Figura 1.2 B), do Período Jurássico superior (150Ma). Assim como o *T. roseae*, o arqueoptérix possuía características comuns entre dinossauros saurísquios (por exemplo: dentes e vértebras caudais não fusionadas) e aves modernas (por exemplo: membros anteriores modificados para planar e penas assimétricas). Entretanto, fósseis com tamanho significado evolutivo são raros e, quando são encontrados, provocam grande alvoroço tanto da comunidade científica quanto da sociedade, sendo popularmente chamados de **elos perdidos**.

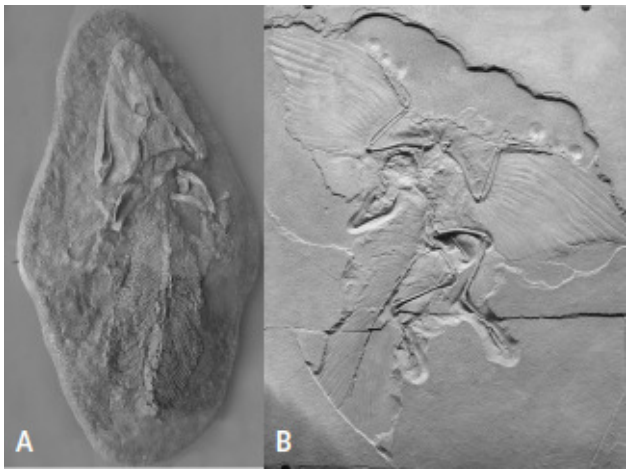


Figura 1.2 Fósseis de (A) *Tiktaalik roseae* e (B) *Archaeopteryx litographica*. Imagens de uso livre, acessíveis no Wikimedia Commons.

Homologias e ontogenia como evidências da ancestralidade comum

O termo homologia foi usado por Richard Owen, um naturalista conterrâneo e contemporâneo de Darwin, para descrever “o mesmo órgão, em organismos diferentes, sujeito a variações de forma e função”. Darwin reconhecia que as homologias eram a principal evidência para sua teoria. Dentre exemplos clássicos de órgãos homólogos com distintas formas e funções, podemos citar os membros dos vertebrados. Os braços de um macaco, as nadadeiras de um golfinho, as asas de um morcego e as patas de um cavalo são todas modificações de membros anteriores. Uma análise mais aprofundada revela que todos estes membros possuem os mesmos componentes ósseos: úmero, rádio, ulna, carpos, metacarpos e falanges, corroborando a hipótese de ancestralidade comum entre todos estes tetrápodes.

A ontogenia, ramo da biologia que estuda o desenvolvimento de um organismo ao longo da vida, também revela homologias quando se comparam estágios de desenvolvimento de distintos organismos. Por exemplo, embriões de

peixes, répteis, aves e mamíferos apresentam arcos branquiais morfologicamente comparáveis nos estágios iniciais do desenvolvimento. Em estágios mais avançados, estes arcos se desenvolvem em estruturas distintas nestes táxons. Nos peixes, o primeiro par dá origem à mandíbula, o segundo se desenvolve no complexo hiomandibular e os demais arcos formam as estruturas esqueléticas das brânquias. Por outro lado, nos mamíferos, o primeiro par de arcos branquiais também está associado à formação de dois dos três ossículos da orelha interna (martelo e bigorna). O terceiro ossículo, estribo, vem do segundo arco branquial, que também contribui para formar o complexo hióide. Apesar das diferenças no desenvolvimento, estas evidências apontam para um ancestral comum dotado de arcos branquiais. As evidências ontogenéticas podem ser corroboradas através das reconstruções filogenéticas, que serão tema dos Capítulos 2 e 3.

1.4 Seleção natural

A seleção natural é um dos principais processos geradores de evolução, sendo o agente natural que origina adaptações. A primeira premissa da seleção natural se baseia nas observações de que todos os organismos possuem um potencial reprodutivo elevado, que resultaria em crescimento exponencial das populações ao longo do tempo. Entretanto, nem todos os indivíduos de uma população se reproduzem e grande parte da prole daqueles que conseguem se acasalar não sobrevive até a idade reprodutiva. Isso ocorre porque há interações agonísticas dentro e entre espécies (competição, predação, parasitismo, entre outras), além de sobrevivência diferencial relacionada a fatores abióticos (frio, chuva, incêndios, entre outros). Além disso, os recursos naturais não suportariam o crescimento exponencial das espécies, aumentando severamente os níveis de competição. Sendo assim, apenas os indivíduos capazes de explorar recursos, sobreviver a predadores, manter seus territórios e atrair parceiros reprodutivos são capazes de deixar descendentes para a geração seguinte.

Isso só ocorre devido à segunda premissa da teoria da seleção natural: a presença de variações dentro das populações. Todos os indivíduos possuem diferenças, ainda que sutis, em características como tamanho corpóreo, fisiologia, comportamento e resistência a patógenos e parasitas. A seleção natural atua sobre essas diferenças, favorecendo indivíduos com combinações de traços ideais para a sobrevivência em um determinado ambiente. Entretanto, os fenótipos favoráveis só geram evolução caso sejam herdáveis. Nestes casos, a seleção natural promove a mudança das populações ao longo do tempo, gerando novas adaptações e, quiçá, novas espécies. Tudo isso gera um questionamento: qual é a fonte da variação herdável nas populações?

1.5 Processos geradores de diversidade genética e modificadores de frequências alélicas

Imagine uma população infinita cujos indivíduos possuam dois alelos de um determinado gene ε. Imagine também que os dois alelos, ε_1 e ε_2, possuem o mesmo valor adaptativo (isto é, não estão sob seleção diferencial). Neste caso, e na ausência de quaisquer outras forças externas, é completamente possível estimar a frequência de cada um dos genótipos na população. Esta estimativa foi proposta independentemente pelo inglês Godfrey Harold Hardy e pelo alemão Wilhelm Weinberg, através de um teorema matemático conhecido como equilíbrio de Hardy-Weinberg, dado por:

$$p^2 + 2pq + q^2 \,, \tag{1.1}$$

onde p^2 é a frequência do homozigoto $\varepsilon_1\varepsilon_1$, q^2 é a frequência do homozigoto $\varepsilon_2\varepsilon_2$ e $2pq$ é a frequência do heterozigoto $\varepsilon_1\varepsilon_2$. Este polinômio pode ser estendido para múltiplos alelos. As demais forças externas mencionadas acima incluem a mutação, a deriva genética e os acasalamentos não aleatórios, além da própria seleção natural.

As **mutações** são a fonte primária de geração de variação nas populações naturais. Elas ocorrem de forma aleatória, mas as mutações neutras são as mais comumente identificadas em amostras populacionais. Este é o caso das mutações silenciosas, que ocorrem devido à **degeneração do código genético**. A degeneração é uma propriedade universal do código e se refere ao fato de que um determinado aminoácido pode ser traduzido a partir de múltiplos códons do RNA. Tomemos como exemplo a leucina nos vertebrados. Este aminoácido pode ser traduzido a partir dos códons CUU, CUC, CUA, CUG, UUA e UUG. Sendo assim, a terceira base do códon CUU pode sofrer mutações para qualquer outro nucleotídeo sem resultar em mudanças fenotípicas. Entretanto, nem todas as mutações são silenciosas, e podem ocasionar pequenas mudanças na estrutura das proteínas. Caso estas mudanças resultem em organismos com valores adaptativos distintos, as frequências alélicas e/ou genotípicas podem ser alteradas ao longo do tempo como resultado da seleção natural. A seleção pode, inclusive, fixar um genótipo novo resultante de uma mutação aleatória caso os homozigotos para o novo alelo tenham valor adaptativo maior do que o dos indivíduos com outros genótipos.

Outro processo modificador de frequências genotípicas são os **acasalamentos não aleatórios**. Imagine uma população em equilíbrio de Hardy-Weinberg, cujas frequências genotípicas sejam p^2 = 25%, q^2 = 25% e $2pq$ = 50%. Caso os homozigotos se acasalem preferencialmente entre si, em poucas gerações a frequência de homozigotos na população aumentará. Quando as populações possuem tamanhos muito pequenos, a chance de acasalamento entre indivíduos aparentados também aumenta, levando a um processo denominado endogamia, cujas consequências genéticas incluem um excesso de homozigotos e a perda de diversidade genética nas populações. A intensificação da endogamia pode levar a um aumento de alelos recessivos raros, muitas vezes deletérios, que aumentam a chance de extinção populacional.

A expressão de alelos raros em endogamia é resultado de um processo evolutivo aleatório denominado **deriva genética**, que ocorre em todas as populações de tamanho finito. A deriva consiste na flutuação aleatória nas frequências alélicas entre gerações, incluindo perda de alelos. Este processo é tão mais intenso quanto menores forem o tamanho da população, o número de alelos e suas frequências. Na ausência da entrada de novos alelos na população, a deriva genética é um processo redutor de variabilidade. No final da década de 1960, o geneticista japonês Motoo Kimura demonstrou matematicamente que as populações podem evoluir por deriva genética na ausência de seleção natural. Esta proposta ficou conhecida como **Teoria Neutra da Evolução Molecular**. Além das mutações, uma população pode receber novos alelos através da **migração**. Enquanto populações isoladas tendem a se diferenciar por deriva e seleção natural, a migração é um processo metapopulacional homogeneizador de frequências alélicas, promovendo o intercâmbio de variantes interpopulacionais e diminuindo os efeitos deletérios da deriva genética.

1.6 Teoria da coalescência e estimativas filogenéticas

Duas cópias alélicas quaisquer podem ser idênticas por descendência ou por estado. O primeiro caso é observado, por exemplo, entre irmãos. Ambos podem compartilhar cópias alélicas vindas do pai ou da mãe. Isto quer dizer que, se observarmos a genealogia entre cópias idênticas de um alelo entre dois irmãos, chegaremos ao ancestral comum destes alelos em uma geração. Por outro lado, dois indivíduos sem nenhum parentesco também podem compartilhar cópias idênticas de um alelo. A filogenia destes alelos pode ser estimada, mas o ancestral comum entre as duas cópias alélicas é muito anterior ao dos alelos idênticos por descendência. Neste caso, diz-se que eles são idênticos por estado. A filogenia entre dois alelos distintos também pode ser estimada, mas é certo que o ancestral comum entre eles é ainda mais antigo do que o dos alelos idênticos por estado. O processo genealógico de busca por ancestrais alélicos a partir de cópias gênicas observadas é chamado de **coalescência**, e quando o ancestral comum entre dois alelos é estimado, diz-se que ambos coalescem neste ancestral (Figura 1.3).

Em 1982, o matemático britânico Sir John Frank Charles Kingman publicou o artigo *On the Genealogy of Large Populations*, no qual desenvolveu modelos probabilísticos de coalescência de alelos a partir de expectativas geradas pela teoria neutra de Kimura. Mais recentemente, estes modelos foram incorporados à sistemática filogenética, enriquecendo os campos da sistemática molecular e da filogeografia.

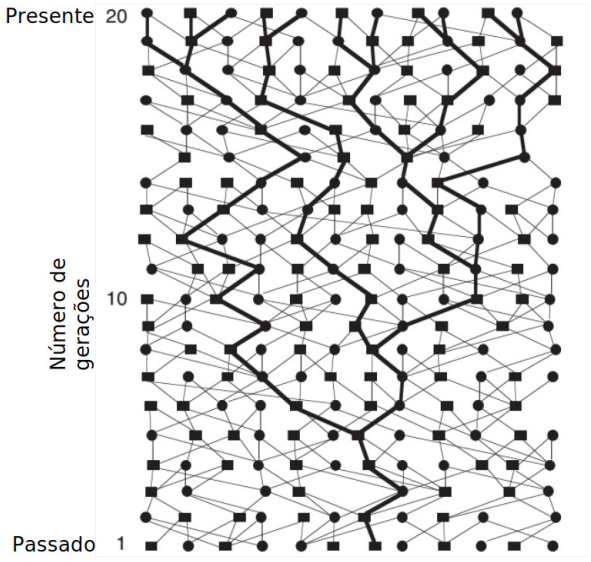


Figura 1.3 Genealogia gênica hipotética representado o pedigree de um gene durante 20 gerações. Quadrados representam machos e círculos representam fêmeas. As linhas conectam as cópias alélicas aos seus ancestrais comuns mais recentes. Modificado de Avise (2009).

1.7 Filogeografia

Como mencionado na introdução, apesar de Stephen J. Gould e outros contemporâneos defenderem uma abordagem pluralística e integrativa entre os estudos micro e macroevolutivos, esta discussão ficou por anos apenas no campo das ideias. Em 1987, o geneticista norte-americano John C. Avise propôs uma nova área de estudo, denominada **Filogeografia**, que integrava na prática as metodologias usualmente empregadas por micro- e macroevolucionistas. Esta disciplina pode ser definida, nas palavras de Avise, como o "campo de estudo que se ocupa com os princípios e processos que governam a distribuição geográfica de linhagens genealógicas, especialmente dentro e entre espécies estritamente relacionadas". Ela integra metodologias da genética de populações, taxonomia, história natural, ecologia, sistemática, biogeografia e paleontologia na elucidação dos princípios e processos acima mencionados.

Uma prática simples e interessante introduzida na filogeografia foi o uso de indivíduos, e não espécies, como unidades operacionais para a reconstrução de relações filogenéticas. Além disso, nas últimas duas décadas, dois grandes avanços metodológicos aprimoraram a acurácia das inferências filogeográficas. O primeiro deles foi a incorporação de algoritmos baseados na teoria da coalescência de Kingman na reconstrução de árvores gênicas e árvores de espécies, levando em consideração a incongruência entre ambas. Neste contexto, surgiram métodos de delimitação de espécies baseados no modelo *multispecies coalescent*, que aumentam a objetividade e replicabilidade do trabalho taxonômico. O segundo avanço foi a implementação de modelos estatísticos para o teste explícito de hipóteses, definindo a subárea da filogeografia estatística. Esta disciplina tem sido extensamente aplicada à Zoologia, com o objetivo de esclarecer questões tais como o tempo e o modo de especiação dos organismos, a origem das distribuições endêmicas e os mecanismos associados aos padrões de diversidade genética ao redor do planeta.

1.8 Bibliografia recomendada

AVISE, John C. Phylogeography: retrospect and prospect. Journal of biogeography, v. 36, n. 1, p. 3-15, 2009.

AVISE, John C.; ARNOLD, Jonathan; BALL, R. Martin; BERMINGHAM, Eldredge; LAMB, Trip; NEIGEL, Joseph E.; REEB, Carol A.; SAUNDERS, Nancy C. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Annual review of ecology and systematics, v. 18, n. 1, p. 489-522, 1987.

CALLAHAN, Hilary S. Microevolution and macroevolution: Introduction. eLS, 2002.

DARWIN, Charles; BYNUM, William F. The origin of species by means of natural selection: or, the preservation of favored races in the struggle for life. AL Burt, 2009.

ERWIN, Douglas H. Macroevolution is more than repeated rounds of microevolution. Evolution and development, v. 2, n. 2, p. 78-84, 2000.

KNOWLES, L. Lacey; MADDISON, Wayne P. Statistical phylogeography. Molecular Ecology, v. 11, n. 12, p. 2623-2635, 2002.

MATIOLI, Sérgio Russo; FERNANDES, Flora Maria de Campos. Biologia molecular e evolução. Holos/Sociedade Brasileira de Genética, 2012.

RIDLEY, Mark. Evolução. Artmed Editora, 2009.

Capítulo 2
Introdução à Sistemática Filogenética e Análise de Dados Morfológicos

Daniel M. Casali & Larissa C. C. S. Dumbá

2.1 O que é a sistemática e o que ela contempla?

A sistemática, em sentido amplo, é a disciplina das Ciências Biológicas que se ocupa de estudar a biodiversidade a partir de uma perspectiva organizacional evolutiva. A sistemática compreende as atividades de **taxonomia**, em que o objetivo é identificar, classificar e nomear as formas de vida, e também do estudo das relações de parentesco (**relações filogenéticas**) entre os organismos. A taxonomia, amparada pela filogenia, é a base teórica e prática para a produção das classificações biológicas. Tais classificações têm como objetivo estabelecer um sistema de referência que tenta ser o mais objetivo e replicável possível. É também parte da sistemática o desenvolvimento de metodologias que permitam estudar em uma perspectiva filogenética (considerando as relações de parentesco) a evolução de características biológicas e entender padrões gerais a partir dessas análises.

2.2 Por que ter um sistema de sistematização (organização/classificação) biológico?

A atividade de sistematização das relações entre as entidades biológicas, não diferente de outras classificações e organizações conceituais humanas, decorre da capacidade cognitiva primitiva (também compartilhada com outros animais) de categorizar elementos do seu ambiente de acordo com as suas necessidades. Essa capacidade é mais desenvolvida na nossa espécie que nas demais espécies que conhecemos, no entanto. Ao decorrer da nossa história, com o crescente aumento do conhecimento, bem como da população humana (tanto em quantidade como em grau de conectividade), sistemas ainda mais elaborados de sistematização foram sendo desenvolvidos e empregados, tornando-se indispensáveis às atividades cotidianas, das mais simples às mais complexas.

Além de servirem como repositório organizado do conhecimento por si só, as classificações quando aplicadas às ciências têm papel de permitir a realização de predições de acordo com as categorias reconhecidas, como é o caso da tabela periódica utilizada na Química, por exemplo – saber a coluna ou grupo em que um elemento se situa permite fazer algumas afirmações sobre suas propriedades. Na Biologia, a existência de um sistema formal de organização da biodiversidade se tornou necessário para a execução da própria ciência bem como de atividades humanas fundamentais que dependem destas informações, como a medicina, a agricultura, a pecuária, o manejo e conservação da biodiversidade, entre outras.

A organização das relações evolutivas entre os seres vivos nos permite desenvolver sistemas de referência que busquem refletir unidades naturais, ao contrário de apenas construir classificações arbitrárias ou artificiais. Essas últimas, embora tenham uma utilidade prática (como é o caso de um inventário de ervas medicinais, por exemplo), têm um escopo reduzido quando se trata de compreender propriedades históricas compartilhadas. Dessa forma, a classificação resultante de uma perspectiva filogenética tem também importante poder preditivo sobre as propriedades dos organismos.

2.3 Breve histórico do pensamento sistemático na Biologia

Muitos indivíduos contribuíram historicamente para a formação do nosso atual entendimento sobre como sistematizar as relações entre os seres vivos, estando fora do escopo deste texto uma relação completa desses e de suas contribuições. No entanto, algumas figuras centrais merecem alguma menção por suas contribuições terem sido desproporcionalmente importantes.

Um dos mais antigos pensadores envolvidos em entender como os seres vivos se relacionam, muito antes de uma visão evolutiva se fizesse presente, foi o filósofo Aristóteles (384-322 a.C.), que considerou características morfológicas dos indivíduos como indicativos de similaridade que permitiria agrupar estes. Muito tempo depois, no século XVIII, Carl Linnaeus, ainda se valendo dessa noção de similaridade e usando as melhores partes de várias outras classificações, propôs um sistema de classificação biológico que apresentava um padrão hierárquico e se tornou um padrão referencial, sendo em grande parte parecido ao que utilizamos atualmente ao classificar os seres vivos.

Durante o final do século XVIII e todo o século XIX, nascia e ganhava enfoque na Europa a disciplina da **anatomia comparada**, tendo especial destaque figuras como Georges Cuvier (1769-1832) na França e Richard Owen (1804-1892) na Inglaterra. Essa disciplina se propunha a estudar os organismos de forma comparativa e não apenas descritiva, e entender os padrões de similaridade e diferença de forma metódica. Esse enfoque teve importantes implicações para o entendimento dos padrões de relação entre os organismos.

Nesse contexto, surge um dos mais importantes conceitos para a anatomia comparada, e também para a sistemática, o conceito de **homologia**. Segundo Owen, que cunhou o termo em 1843, como visto no capítulo anterior, homologia seria: "o mesmo órgão, em diferentes animais, em toda a variedade de forma e função". Isso expande a noção de que duas estruturas reconhecidas e comparáveis entre si, em dois organismos distintos, não são apenas similares, mas **as mesmas**. Owen também reconheceu as estruturas análogas, que apesar de cumprir as mesmas funções, poderiam ou não apresentar as mesmas formas. No entanto foi Ray Lankester (1847-1908) que propôs pela primeira vez o termo **homoplasia**, para se referir às estruturas que eram morfologicamente similares, mas não eram verdadeiras homologias, já que não se tratavam da mesma estrutura. Exemplos de estruturas homoplásticas são asas dos morcegos, das aves e dos pterodátilos ou as nadadeiras dos cetáceos, dos "peixes" ou dos ictiossauros.

Em seu principal trabalho, A Origem das Espécies (1859), Charles Darwin (1809-1882), após uma extensa compilação de observações e experiências, contribuiu com duas principais ideias que mudariam a forma de entender a biologia, incluindo as relações entre os organismos. A primeira delas foi o mecanismo central pelo qual ele propunha que as formas de vida se modificavam ao decorrer do tempo, a **seleção natural**, descoberta paralelamente com Alfred Russel Wallace (1823-1913). A segunda ideia, que mais nos interessa no contexto da sistemática, é a ideia de que os seres vivos seriam todos aparentados entre si, e que toda a vida na terra poderia ser conectada através de relações de ancestralidade-descendência, formando um contínuo que refletiria o que hoje chamamos de árvore da vida.

O impacto dessas teorias foi imenso, passando por períodos de muita controvérsia acadêmica e social, mas hoje são consideradas, do ponto de vista científico, parte fundamental do nosso entendimento da biodiversidade e seus padrões e processos. A concepção dessa árvore da vida, na qual os organismos de fato se relacionavam com os outros através de parentesco, conseguia pela primeira vez explicar satisfatoriamente o porquê de as classificações biológicas serem tão naturalmente hierárquicas. Isso também explicou, através de um mecanismo, o que são as homologias ou estruturas herdadas de um mesmo ancestral comum. Desde então, se entende que as classificações serão tão melhores quanto se aproximarem de uma genealogia, refletindo o padrão natural evolutivo.

Diagramas de árvores começaram a se tornar comuns na literatura, sendo um dos mais famosos o utilizado por Ernst Haeckel (1834-1919), que foi também quem cunhou o termo **filogenia** (se referindo à árvore evolutiva da vida). Apesar do uso gráfico de diagramas de árvores, no início do século XX ainda não existiam critérios práticos e explícitos para definir as relações entre os organismos, e esses diagramas eram, inevitavelmente, especulativos.

2.4 Escolas de pensamento na sistemática moderna

Nas décadas iniciais do século XX (1918-1950) foram incorporados à teoria evolutiva os conhecimentos oriundos de várias disciplinas biológicas, como a genética, paleontologia, zoologia, botânica, embriologia e sistemática, dando origem a síntese moderna evolutiva. No que afeta a sistemática, teve origem a **taxonomia evolutiva**, a primeira escola moderna de sistemática. Nos anos subsequentes, em parte em resposta à falta de metodologias quantitativas na taxonomia evolutiva, surgiram duas outras escolas de pensamento: a **taxonomia numérica** e a **sistemática filogenética**.

Taxonomia evolutiva

A escola da taxonomia evolutiva ou escola gradista se originou como consequência imediata de se aplicar princípios evolutivos às classificações biológicas. Apesar do avanço conceitual, não havia uma metodologia explícita de como reconstruir as relações filogenéticas entre os organismos, que seria a base para as classificações produzidas. Os gradistas reconheciam a importância de distinguir as homologias das homoplasias na tentativa de reconhecer grupos naturais, e reconheciam grupos parafiléticos e monofiléticos. Eles se valiam não apenas dos padrões de parentesco, resultantes dos eventos **cladogenéticos** (eventos de especiação), mas também consideravam importante

a quantidade de mudanças resultantes de eventos **anagenéticos** (modificações em uma mesma linhagem). Essas últimas características muitas vezes eram utilizadas com um certo grau de subjetividade, atribuindo a algumas características maior ou menor grau de informação relacionado à complexidade da estrutura ou importância para a sobrevivência e reprodução dos organismos. Pesquisadores influentes dessa escola, bem como na formulação da síntese moderna evolutiva, foram o mastozoólogo George G. Simpson (1902-1984) e o ornitólogo Ernst Mayr (1904-2005).

Taxonomia numérica

Em busca de uma objetividade operacional, os taxonomistas numéricos (também chamados de feneticistas), entre os quais se destacam Robert R. Sokal (1926-2012) e Peter H. A. Sneath (1923-2011) advogavam que todas as características biológicas deveriam ser utilizadas para se agrupar e organizar a biodiversidade. Os dados deveriam ser, sempre que possível, quantitativos e analisáveis com metodologias estatísticas para mensurar distâncias e realizar agrupamentos. Esse requisito implicava que o mais importante para a fenética não eram os padrões evolutivos, mas apenas a similaridade geral entre os seres vivos e, portanto, não havia porque não utilizar também as homoplasias, além das homologias, no processo.

A fenética, desde o início, encontrou problemas tanto metodológicos como filosóficos. Metodologicamente, havia desentendimentos de qual seria a melhor forma de medir a similaridade ou agrupar os organismos, uma vez que as possibilidades matemáticas eram, e são, das mais variadas. Filosoficamente, era questionável excluir a informação evolutiva das classificações. Qual seria a validade de ter uma classificação com grupos que não eram resultantes do processo natural que gera a hierarquia biológica que sempre estimulou as classificações? Para os feneticistas, as classificações poderiam ser artificiais, bastando ser úteis em alguma finalidade.

Apesar da escola da taxonomia numérica não ter se tornado o padrão atual de classificação biológica, dois princípios que utilizamos hoje são em parte uma herança dessa proposta: a busca pela objetividade metodológica e o uso de métodos quantitativos (por exemplo: matrizes de dados codificados e algoritmos de escolha de árvores, respectivamente).

Sistemática filogenética

A terceira escola de pensamento é a sistemática filogenética (ou, como apelidada ironicamente por Ernst Mayr, cladística). A escola se desenvolveu a partir da publicação do livro *Phylogenetic Systematics* do alemão Willi Hennig (1913-1976). O livro de Hennig foi publicado inicialmente em alemão, em 1950, e apenas em 1966 foi traduzido para o inglês, quando recebeu a atenção de pesquisadores por todo mundo. A metodologia, diferentemente e em oposição a fenética, foi construída levando em conta elementos do processo evolutivo.

Além de diferenciar as homologias de homoplasias, Hennig propôs que as homologias deveriam ser distinguidas entre primitivas ou ancestrais (**plesiomorfias**) e derivadas (**apomorfias**), sendo unicamente estas últimas informativas para agrupar os organismos. Isso porque as apomorfias possuem características ancestrais que não guardariam informação sobre grupos mais derivados como, por exemplo, a presença de coração entre os vertebrados (que está presente antes da origem do grupo), não nos diz nada sobre as relações entre os subgrupos de vertebrados. No entanto, características derivadas compartilhadas por apenas alguns organismos, como a presença de crânio, nos informariam que existe um grupo dentro dos vertebrados (Craniata), diagnosticado por esta característica.

Quando essas características são compartilhadas por mais de um grupo, essas recebem o prefixo sin(m), do grego *syn-*, que significa em conjunto. Logo, temos **simplesiomorfias** (=plesiomorfias compartilhadas) e **sinapomorfias** (=apomorfias compartilhadas). Apomorfias de um único grupo ou espécie são chamadas de autapomorfias, e não são informativas para agrupar estes com outros organismos.

Hennig também propôs que haveria três tipos de agrupamento que poderiam ser propostos para um conjunto de seres vivos (Figura 2.1):

- Grupos monofiléticos: incluem todos os descendentes de um ancestral comum imediato (= ancestral comum exclusivo) – diagnosticados por sinapomorfias;
- Grupos parafiléticos: excluem ao menos um descendente de um ancestral comum imediato – diagnosticados por simplesiomorfias;
- Grupos polifiléticos: indivíduos agrupados não possuem ancestral comum imediato – diagnosticados por homoplasias.

Diante disso, Hennig defendeu que apenas os grupos monofiléticos são grupos naturais, resultantes do processo evolutivo, e que os demais grupos - para- e polifiléticos - são construções humanas arbitrárias ou baseadas em critérios que não refletiriam a genealogia, mas sim similaridade ou diferença fenotípica. Essa similaridade poderia ser um relicto de uma história evolutiva compartilhada (presença de crânio em um salmão e em um cavalo, por exemplo), ou poderia ser oriundo de origem independente da característica (por exemplo, cauda preênsil

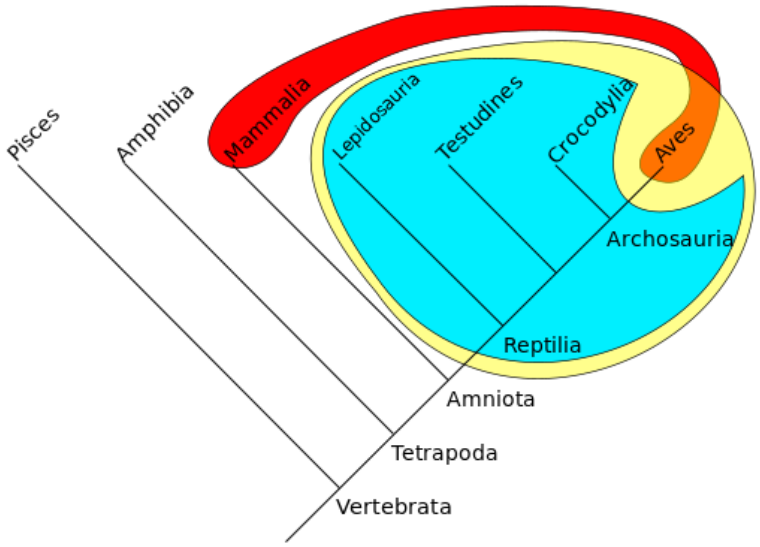


Figura 2.1 Cladograma representando as relações filogenéticas entre as linhagens de Tetrapoda (=vertebrados com quatro membros). Em amarelo, indicando um grupo monofilético (ancestral comum imediato + todos descendentes) - Sauropsida. Em azul, um grupo parafilético (ancestral comum imediato + descendentes, porém não todos) - "répteis" - que exclui as aves, e portanto não é um grupo natural. Em vermelho, um grupo polifilético - Mamíferos + Aves, um grupo artificial, pois não inclui o ancestral comum imediato dos dois grupos. Imagem modificada de Wikipédia, domínio comum.

em macacos e em roedores). A despeito da possível utilidade prática, esses grupos não refletiriam a realidade evolutiva e, portanto, constituiriam um sistema menos útil de classificação e organização da biodiversidade (é importante lembrar da função de predição sobre a biodiversidade, comentadas no início deste texto). Como grupos monofiléticos são diagnosticados através de sinapomorfias e apenas grupos monofiléticos seriam desejáveis, logo, apenas sinapomorfias deveriam ser utilizadas no processo inferência das relações filogenéticas e classificação biológica, concluiu Hennig.

Uma síntese dos grupos reconhecidos por cada uma das escolas, bem como as características biológicas utilizadas por cada uma delas pode ser vista na Tabela 2.1, abaixo.

O período histórico que se seguiu foi marcado por vários debates intensos entre os defensores de cada uma das três escolas, e ficou conhecido como o período das Guerras Sistemáticas. Eventualmente a sistemática filogenética se tornou o método filosófico e prático dominante, sendo o paradigma conceitual vigente atualmente, apesar das modificações que a sistemática passou e vem passando continuamente.

2.5 Cladogramas como representação das relações evolutivas hipotéticas

Os cladogramas, que vêm do nome clado (=grupo monofilético), são representações gráficas dos padrões de relacionamento filogenético hipotetizados para um conjunto de organismos (Figura 2.2). Os cladogramas exibem os grupos sendo estudados (geralmente espécies ou táxons superiores) como ramos externos, chamados de **terminais** ou táxons. Esses terminais se ligam a vértices, chamados de nós, que representam os ancestrais hipotéticos dos terminais estudados. Outros ramos internos, que ligam os nós uns aos outros, representam linhagens ancestrais hipotéticas. O nó mais basal da árvore é chamado de **raiz**, e representa o ancestral comum de todos os terminais sendo estudados (incluindo o grupo externo – mais detalhes logo abaixo). O padrão de relacionamento entre os terminais de um cladograma não representa o tempo evolutivo absoluto, mas apenas os padrões relativos dos eventos

Tabela 2.1 Grupos reconhecidos e caracteres utilizados por cada uma das escolas de classificação modernas (adaptado de Ridley, 2004).

Caracteristicas utilizadas						
	Grupos reconhecidos			Homoplasias	Homologias	
Escolas	Monofiléticos	Parafiléticos	Polifiléticos		Apomorfias	Plesiomorfias
Gradista	sim	sim	não	não	sim	sim
Fenética	sim	sim	sim	sim	sim	sim
Cladista	sim	não	não	não	sim	não

de cladogênese, definindo assim em que ordem relativa os organismos estudados são aparentados entre si. Esse padrão é chamado de topologia do cladograma.

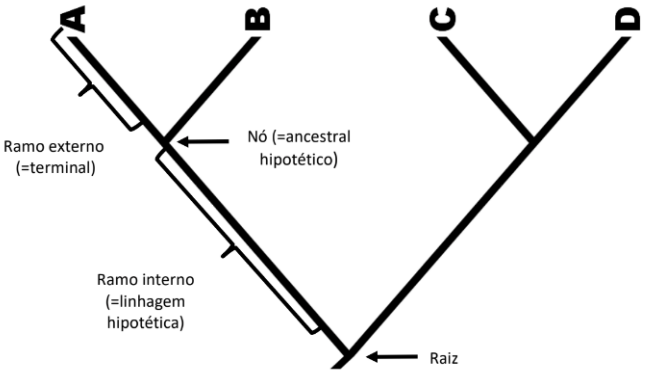


Figura 2.2 Um exemplo de cladograma e a terminologia utilizada para se referir aos seus elementos.

O lado em que os terminais aparecem não possui nenhuma implicação para os padrões evolutivos sendo exibidos, podendo-se rotacionar qualquer nó da árvore sem perda de informação. Isso ocorre porque é a ordem em que os organismos são conectados que define suas relações de parentesco. A figura 2.3 mostra como diferentes rotações dos nós não mudam a ordem em que os organismos se relacionam. Clados ou terminais ligados por um ancestral comum imediato são chamados de grupos irmãos.

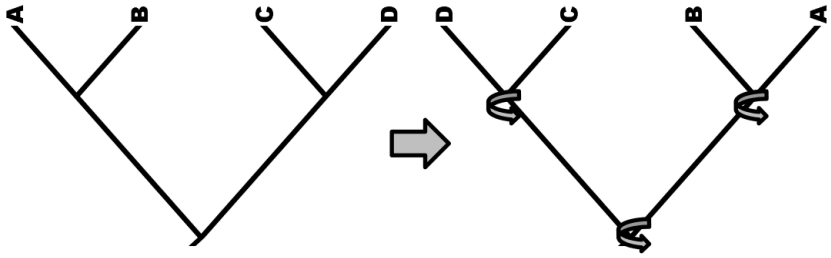


Figura 2.3 Esquema demonstrando que é possível rodar todos os nós de um cladograma sem alterar os padrões de relacionamento entre os táxons. Em ambos os cladogramas, A e B são grupos irmão mais próximos entre si, assim como C e D são entre si também. A relação entre os clados também não se altera ao rodar os nós, sendo AB irmão de CD.

2.6 Caracteres e estados de caráter

As características individualizadas que nos permitem comparar sua variação entre diferentes terminais são referidas como **caracteres** (plural de caráter), e as formas da variação observada em cada terminal são chamadas de **estados de caráter**. Uma definição mais formal, proposta por Wagner em 1999, seria que cada caráter é: "...uma propriedade de um organismo que é quase-independente de outras propriedades do organismo", e que os estados de caráter, segundo Wiley & Liebermann (2011) são "[a] interpretação do caráter de um organismo que é usado para comparar o caráter desse organismo com outro organismo".

Estabelecer um caráter e seus estados é o mesmo que propor uma hipótese de homologia entre estruturas. Para isso se deve utilizar critérios práticos como similaridade, posição, conectividade e desenvolvimento das estruturas, por exemplo. É uma atividade central do processo de inferência das relações filogenéticas e demanda um profundo estudo da morfologia dos organismos, bem como um bom arcabouço teórico sobre o processo em si.

No exemplo dos vertebrados citado anteriormente, o caráter seria **presença de crânio**, e os estados **ausente** e **presente**. Outro exemplo poderia ser a **cor da pelagem** em um mamífero, tendo como estados: **preto**, **marrom** e **cinza**. Os estados de caráter são codificados utilizando-se, mais frequentemente, números – exemplo: presença de crânio - ausente (0), presente (1); cor da pelagem - preto (0), marrom (1), cinza (2). Esses caracteres são chamados caracteres discretos, pois tem valores inteiros, e simbolizam variação em categorias bem definidas. Existem também caracteres de variação contínua, como de medidas, que podem ser discretizados em categorias (exemplo: 1-15 cm (0), de 16-30 cm (1)) ou utilizados de forma contínua, usando os valores mensurados diretamente. Esse último caso é menos comum na prática da inferência de filogenias e focaremos no restante do texto nos caracteres discretos (ou discretizados).

Do ponto de vista teórico, quando estamos estudando as relações filogenéticas de um determinado grupo, estamos interessados em caracteres que sejam ao menos em parte herdáveis e que não exibam extrema variação para cada terminal. No entanto, às vezes essa variação em um mesmo terminal existe, como entre indivíduos de uma mesma espécie apresentando diferentes estados de caráter. Nesse caso, é possível codificar um terminal com mais de um estado (0 & 1), definindo assim que o caráter é **polimórfico** para aquele terminal. Outro requisito teórico é que caracteres que sejam dependentes uns dos outros e que, portanto, forneçam informação filogenética redundante (pois evoluem em concerto) sejam evitados, utilizando apenas caracteres independentes entre si. Nem sempre é possível distinguir isso no momento de definir os caracteres, mas é sempre bom estar atento caso seja possível diagnosticar essa dependência.

Além dos organismos pertencentes ao grupo de interesse (chamados de **grupo interno**), para o qual se deseja entender as relações filogenéticas, é também necessário codificar os caracteres para um ou, preferencialmente, alguns **grupos externos**. Os grupos externos seriam organismos que estão fora do nosso grupo de interesse, e que servirão como comparativo para definirmos quais características do grupo interno são primitivas (plesiomorfias) e quais são derivadas (apomorfias). Usando novamente o exemplo dos vertebrados, imaginem que queremos entender se a presença de um crânio é uma sinapomorfia ou uma simplesiomorfia. Uma vez que temos alguns vertebrados que não possuem crânios (os Myxiniformes) e outros que possuem. Apenas comparando com a condição observada em um ou mais grupos de não-vertebrados (como os anfioxos, por exemplo) é que podemos concluir que a ausência de crânio é uma característica simplesiomórfica, ao passo que a presença dessa estrutura é uma característica sinapomórfica e, portanto, diagnóstica para um clado interno aos vertebrados (Craniata). Esse processo é chamado de polarização de caracteres. Existem outros métodos para a polarização de caracteres, como o uso de fósseis e estágios ontogenéticos (=do desenvolvimento), mas raramente são utilizados atualmente.

O processo de escolha dos grupos externos é de extrema importância para que a análise filogenética seja confiável, por isso se deve valer de outros estudos filogenéticos prévios ou outras informações biológicas para definir bem os grupos externos. É recomendável utilizar grupos proximamente aparentados, podendo incluir-se o grupo irmão, mas também alguns não tão próximos, na tentativa de que a amostra taxonômica tenha maior probabilidade de incluir espécies portando os estados ancestrais dos caracteres.

Com essas informações em mãos, deve-se montar uma matriz de dados, sendo as linhas os terminais, as colunas os caracteres, e as células os estados de caracteres codificados para cada terminal, como no exemplo abaixo. O uso de **0** para os estados presentes no grupo externo não é uma necessidade, mas pode ajudar na organização dos dados.

Existem alguns caracteres que têm uma relação mais complexa entre si no que se refere ao ato de codificação (não confundir com a dependência do tipo redundância, citada acima). Por exemplo, imaginando que dois caracteres estejam sendo usados em uma mesma matriz de dados: presença de asas e presença de listra nas asas. Apesar de serem duas informações independentes, a codificação do caráter referente a listra depende da presença da asa em si. Logo, indivíduos que não possuem a asa não podem nem mesmo ser avaliados sobre a presença da listra. Isso significa que o segundo caráter terá de ser codificado como não aplicável para os terminais que não possuem a asa no primeiro caráter, usando o símbolo (?) ou (–) (que também é como devemos simbolizar quando um dado está faltando, por qualquer razão que nos tenha levado a não possuir tal informação).

A codificação desses dois caracteres como acima exemplificados é chamada de **codificação redutiva**, em contraste a **codificação composta**, que consideraria o caráter da seguinte forma: Asa - ausente (0), presente e sem listra (1) e presente e com listra (2). Essa forma alternativa transforma os dois caracteres em um único. Apesar de não ser estritamente proibida, essa forma de codificação é geralmente desaconselhada, tanto porque comete um equívoco lógico de misturar duas propriedades em um único caráter (presença da asa e presença da faixa), mas também porque ela pode levar a perda de informação filogenética útil.

2.7 O princípio de parcimônia nas análises filogenéticas

Uma vez que todos os caracteres foram codificados levando em conta as hipóteses de homologia propostas (chamadas de **homologias primárias**), é necessário definir qual o critério para interpretar a evolução desses caracteres e assim obter o cladograma que melhor representará as possíveis relações filogenéticas entre os táxons

Tabela 2.2 Exemplo de uma matriz de dados padrão utilizada para as análises filogenéticas

Taxons	Caráter 1	Caráter 2	Caráter 3	Caráter 4	Caráter 5	Caráter 6
Gr. Exter.	0	0	0	0	0	0
Taxon A	1	1	?	?	1	1
Taxon B	1	0	1	1	1	1
Taxon C	0	1	1	1	1	1

estudados. Temos que considerar não apenas o critério para cada interpretar a evolução de cada caráter, mas para resolver também as possíveis incongruências que podem aparecer ao se juntar vários caracteres.

O critério que foi proposto por Hennig e posteriormente justificado filosoficamente por outros autores, especialmente por James S. Farris em 1983, foi o princípio chamado de parcimônia ou máxima parcimônia. O princípio da parcimônia (também conhecido como a navalha de Occam) tem a autoria atribuída ao frade franciscano Willian de Occam (1285-1347), que teria alegado que "pluralidade não deve ser postulada sem necessidade". Isso implica que se duas hipóteses explicam igualmente bem um fenômeno ou padrão, aquela que o fizer usando o menor número de alegações adicionais deve ser preferida, uma vez que essa não adiciona mais complexidade onde ela não é necessária. É importante ressaltar que isso não implica que a hipótese mais simples é sempre a melhor, uma vez que a mais simples pode não explicar algo tão bem como uma mais complexa. Logo, o princípio é útil apenas para escolher entre hipóteses igualmente explicativas, mas que variam no grau de complexidade. Aplicando o princípio à sistemática, Farris justifica que a parcimônia atuaria minimizando as **hipóteses ad hoc** (postulados extras) de homoplasias, uma vez que seria razoável assumir que por princípio os organismos se assemelhariam por ancestralidade comum e, portanto, as similaridades por homoplasias (convergências, paralelismos e reversões) seriam menos prováveis e não deveriam ser postuladas sem estrita necessidade.

Conforme o exemplo da Figura 2.4 e considerando os dados da tabela 2.2, podemos observar que para cada caráter, devemos assumir que a melhor explicação para a evolução desse é aquela em que o número de modificações evolutivas que este caráter passou é o menor possível. No entanto, alguns caracteres podem sugerir relações que são incongruentes com as sugeridas por outros e, portanto, é necessário também empregar o método da parcimônia para todos caracteres em conjunto. O cladograma final será aquele ou aqueles que minimizarem ao máximo o número de passos evolutivos necessários para explicar a evolução de todos eles em conjunto, maximizando a parcimônia.

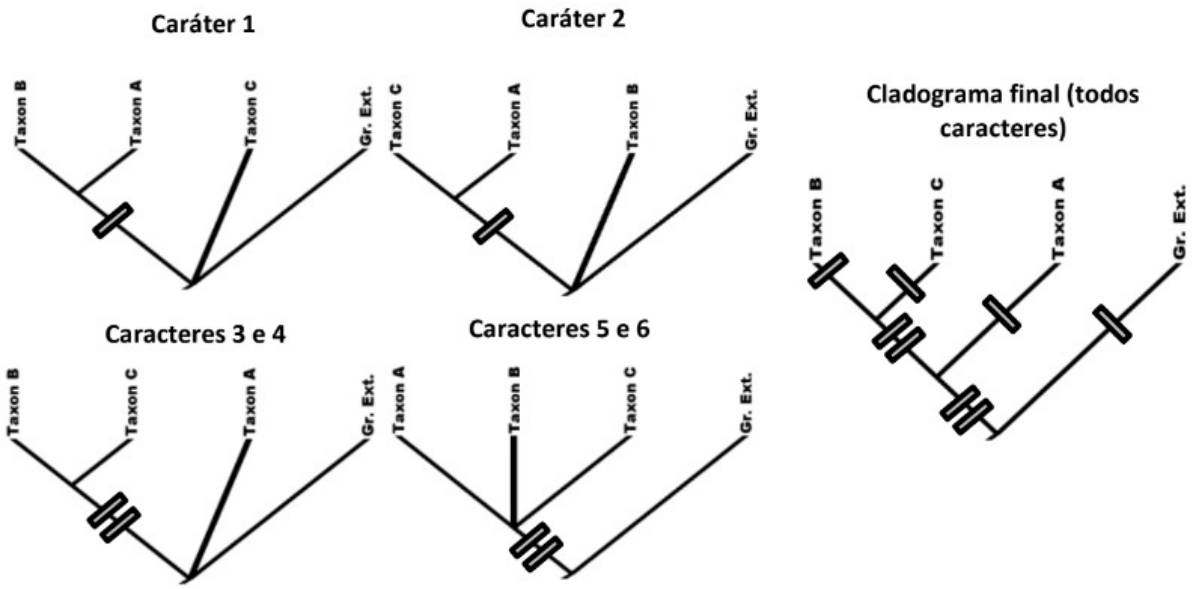


Figura 2.4 Critério da máxima parcimônia aplicado aos caracteres individualmente e o cladograma final, considerando a interação e conflito entre os caracteres. Os caracteres 3 e 4 são congruentes entre si, assim como os caracteres 5 e 6 também o são, já os caracteres 1 e 2 são incongruentes entre si, e incongruentes com os caracteres 3 e 4. Esses dois últimos, por sua vez, são congruentes com os caracteres 5 e 6, assim como os conjuntos de caracteres 1, 5 e 6 ou 2, 5 e 6 também são conjuntos que apresentam congruência. O cladograma final apresenta a hipótese mais parcimoniosa, onde os caracteres 1 e 2 são interpretados como homoplasias (podendo igualmente ser uma convergência ou uma reversão – a figura exibe o cenário de convergência) e os caracteres 3 a 6 são sinapomorfias. As barras cinzas simbolizam as transformações e os estados dos caracteres, de 0 para 1.

Nesse processo, algumas das homologias primárias serão confirmadas e passam a serem consideradas **homologias secundárias** (=sinapomorfias). Como vimos, as sinapomorfias definirão os padrões de agrupamento dos táxons. Outras hipóteses primárias não serão confirmadas, pois serão reconhecidas como plesiomorfias ou homoplasias no(s) cladograma(s) mais parcimonioso(s). Conforme o número de caracteres e terminais aumenta, mais difícil se torna de fazer tais análises manualmente, e por isso que atualmente todas as análises são conduzidas com o auxílio de programas de computador.

Algumas vezes um cladograma mais parcimonioso permite mais de uma interpretação da evolução dos caracteres, como seria o caso do cladograma final da figura 2.4. Os estados de caráter compartilhados por A e C são homoplasias, e podem ser igualmente explicados por um surgimento independente em cada terminal (dois passos) ou surgimento no clado ABC, e reversão em B (dois passos). O mesmo vale para a homoplasia de A e B, que poderia ter surgido

convergentemente nesses terminais, ou surgido em ABC, sendo perdida em C. Quando temos cenários igualmente parcimoniosos dessa forma, dizemos que a otimização dos caracteres é ambígua e não é possível escolher uma em detrimento da outra, a princípio. No entanto, é possível se ater apenas às otimizações não-ambíguas.

Em outros casos, temos como resultado da análise não apenas um, mas vários cladogramas mais parcimoniosos e como todos são igualmente ótimos perante o critério de parcimônia, não é possível de se escolher entre um ou outro, a não ser por outro critério adicional, o que pode ser controverso. Nesses casos, geralmente utilizam-se os métodos de consenso, fazendo-se então um sumário dos clados que são comuns em todos (**consenso estrito**) ou na maioria (**consenso de maioria**) dos cladogramas mais parcimoniosos (Figura 2.5). Essas são as formas de consenso mais utilizadas, apesar de existirem outras. O consenso de maioria precisa de um valor de corte, acima do qual se aceita que um determinado clado seja mantido. O exemplo da figura abaixo considera que se um clado aparecer em mais de 50% dos cladogramas, ele irá aparecer no consenso de maioria. O valor desse corte pode variar de 50% a 99%, sendo que o corte de 100% é o consenso estrito. É importante lembrar que o consenso nunca pode ser interpretado como uma árvore evolutiva, como os cladogramas individuais, mas apenas como um sumário da nossa incerteza diante dos dados e da análise realizada.

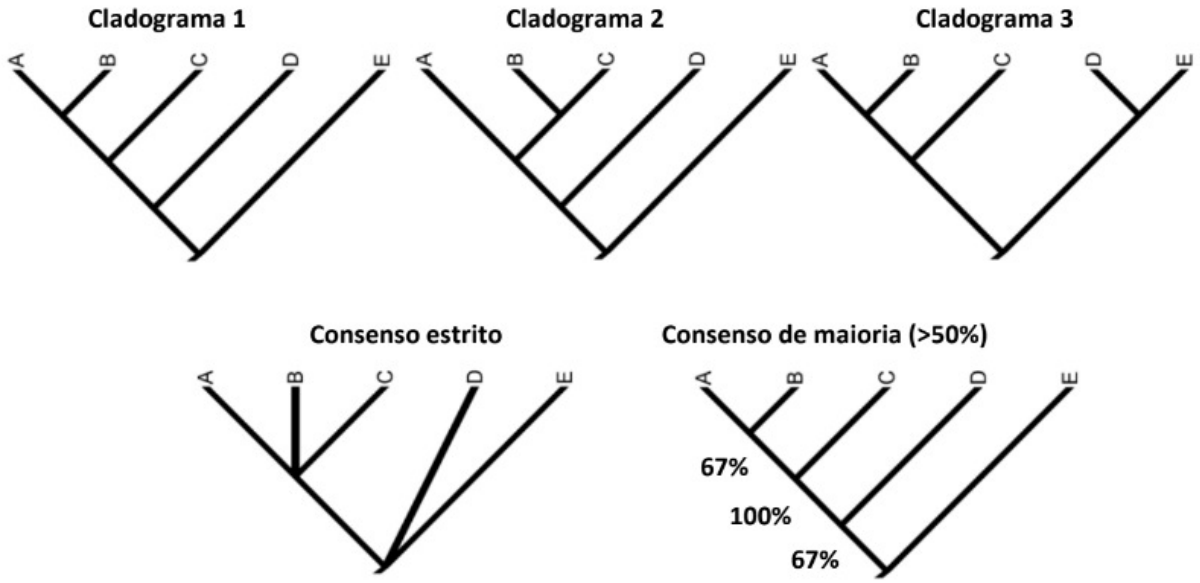


Figura 2.5 Três cladogramas igualmente parcimoniosos e seus consensos, estrito e de maioria (>50%). Os valores abaixo de cada clado do consenso de maioria representam a frequência daquele clado nos cladogramas mais parcimoniosos.

Esses valores não devem ser confundidos com valores que indicam suporte ou estabilidade dos clados em cladogramas individuais, que podem ser estimados através de várias metodologias diferentes. Tais medidas são um assunto importante na sistemática, mas por ser um assunto complexo por si só, não será abordado neste material introdutório. Mais informações sobre essas medidas poderão ser encontradas no material indicado para leituras complementares.

2.8 Ordenação e pesagem de caracteres

Tradicionalmente, cada mudança entre os estados de caráter é contabilizada igualmente, com peso igual a um. No entanto, existem formas de tratar os caracteres diferentemente de forma que nem sempre esses ou as transformações dos seus estados possuam igual contribuição para o comprimento (=número total de passos) que o(s) cladograma(s) final(is) terá(ão). Esses métodos são a **ordenação de caracteres** e a **pesagem diferencial de caracteres**.

Na ordenação dos caracteres, pode-se atribuir uma sequência necessária para as transformações entre os estados e contar como dois ou mais passos a mudança que não passar pelos estágios intermediários. Imaginando-se que a posição de uma estrutura possa ser definida em três categorias: anterior (0), média (1) e posterior (2). Se diferentes terminais possuírem a estrutura em diferentes posições, é possível definir que para passar da posição anterior (0) para a posição posterior (2), é necessário antes passar pela posição média (1). Isso implica que se for observada uma transformação dos estados 0 para 2 neste caráter, devemos contar como dois passos evolutivos e não apenas um, como seria se não estivéssemos ordenando o carácter.

Já a pesagem, atribui um peso diferencial para cada caráter como um todo ou para as mudanças dos estados. É possível definir os pesos anteriormente, durante ou posteriormente à análise. Diversos critérios foram propostos de forma a obter valores que definam esses pesos relativos entre os caracteres, incluindo medidas subjetivas de importância e complexidade dos caracteres ou algumas mais objetivas como grau de homoplasia daquele caráter em um determinado cladograma (caracteres mais homoplásticos teriam menor valor informativo sobre as relações filogenéticas, argumentam os que defendem essa abordagem). Todos esses métodos têm seus prós e contras e existe bastante discussão na literatura sobre o uso desses recursos. Mais detalhes sobre esses assuntos podem ser encontrados nos materiais para leitura complementar indicados no final deste texto.

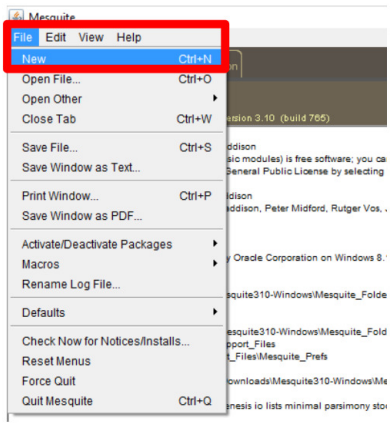
2.9 Comentários breves sobre dados moleculares e análises estatísticas

Apesar de ter sido utilizado originalmente (e ainda hoje, mais frequentemente) para analisar dados morfológicos, o critério da parcimônia pode também ser aplicado aos dados moleculares. Isso é menos frequente devido a algumas limitações metodológicas do critério da parcimônia, especialmente relevantes quando se utiliza os dados moleculares. A principal dessas limitações seria a chamada **atração de ramos longos**, onde dois ramos dos cladogramas são inferidos como sendo grupos irmãos, mesmo que esses não sejam proximamente aparentados e tenham apenas acumulado muitas mudanças convergentemente.

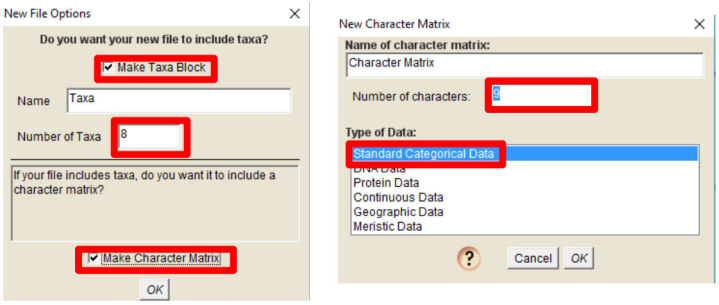
Na tentativa de buscar métodos que seriam menos susceptíveis a essas limitações, foram desenvolvidos procedimentos estatísticos de análise filogenética que utilizam modelos de evolução molecular e que são baseados nos critérios de otimização de máxima verossimilhança e inferência Bayesiana. Tais métodos estatísticos começaram, mais recentemente, a ser também empregados nas análises morfológicas. Apesar dos modelos atualmente disponíveis não serem de todo satisfatórios para lidar com a evolução de caracteres morfológicos, avanços têm acontecido nesse sentido e é possível que no futuro esses métodos baseados em modelos probabilísticos de evolução sejam tão relevantes quanto a parcimônia para inferências morfológicas ou nas análises combinando morfologia e moléculas.

2.10 Montando uma matriz de dados no programa Mesquite

O programa Mesquite é uma plataforma multi-funções, de autoria dos irmãos W. Maddison e D. Maddison e disponível desde 1997. O programa inclui montagem e organização de matrizes de caracteres, além de várias funcionalidades de edição dos dados e variadas análises. Não é objetivo aqui abordar todas as funcionalidades desse programa, mas apenas utilizar algumas funções bem simples para montagem de uma matriz de caracteres com dados morfológicos hipotéticos e exportação para o formato *.tnt*, que será utilizado por outro programa para as análises filogenéticas. Após abrir o programa, a primeira tela exibida será essa abaixo, e devemos clicar em [*File, New*] para criar um novo arquivo.

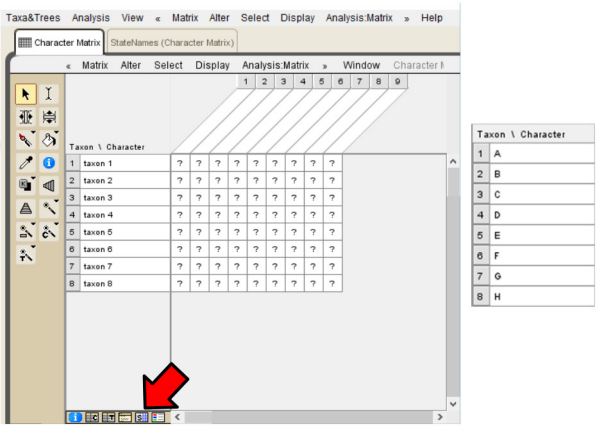


Após clicar em [*New*], o programa irá solicitar um destino no computador e um nome para o arquivo sendo gerado. Então a seguinte tela irá abrir, para definir algumas propriedades da matriz de dados.

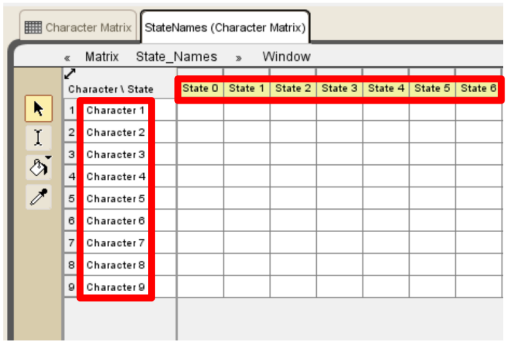


Devem-se marcar os boxes [*Make Taxa Block*] [*Make Character Matrix*] para que o programa crie os blocos de táxons e caracteres, respectivamente. Nessa tela é recomendável também definir o número de táxons que haverá no bloco sendo criado. Esse número pode ser alterado posteriormente. Também é dada a opção de nomear o bloco de táxons. Na próxima tela, bem similar, deve-se definir o número de caracteres (também editável posteriormente), podendo-se também nomear a matriz de dados. No último campo, deve ser definido o tipo de dado que será usado. No caso de caracteres morfológicos discretos, marcamos [*Standard*]. É possível observar que o programa suporta uma série de formatos, como caracteres contínuos, moleculares, entre outros.

Será exibida então a tela principal do programa, mostrando a matriz de dados, com os táxons nas linhas, os caracteres nas colunas e os estados a serem codificados (no momento, todos como **?**) sendo as células da matriz. Para editar os nomes dos táxons, basta dar um duplo-clique em cada um dos nomes pré-definidos e escrever o nome desejado. A próxima coisa que devemos fazer é informar os caracteres e seus estados para o programa. Para isso, primeiro deve-se clicar no símbolo de tabela com um **S**, canto inferior esquerdo.



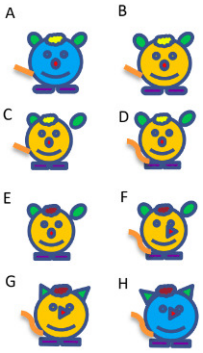
A tela abaixo será exibida. Nela, os caracteres estão exibidos nas linhas (para editar os nomes dos caracteres, basta dar um duplo-clique nos nomes pré-definidos) e os estados de caráter nas colunas. Na frente de cada caráter, deve-se escrever o nome do estado, que ficará então associado a um código (0, 1, 2...), conforme a coluna em que for colocado.



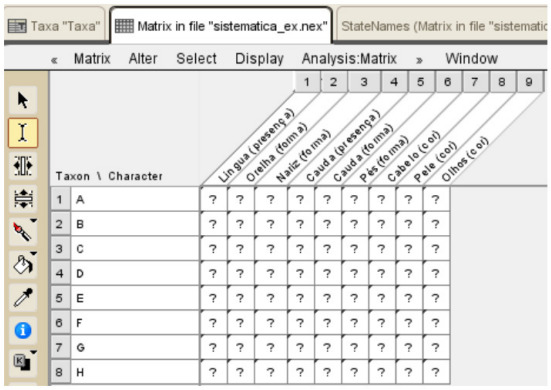
Após o preenchimento dos nomes e dos estados de caráter, essa tabela fica assim:

« Matrix State_Names » Window

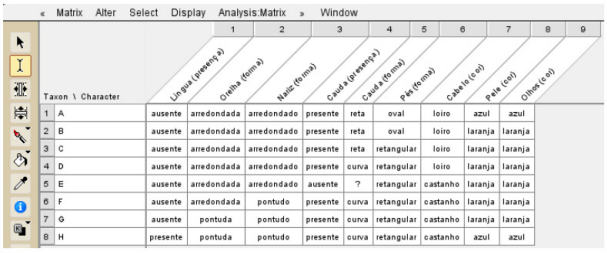
	Character \ State	State 0	State 1
1	Lingua (presença)	ausente	presente
2	Orelha (forma)	arredondada	pontuda
3	Nariz (forma)	arredondado	pontudo
4	Cauda (presença)	ausente	presente
5	Cauda (forma)	reta	curva
6	Pés (forma)	oval	retangular
7	Cabelo (cor)	loiro	castanho
8	Pele (cor)	azul	laranja
9	Olhos (cor)	azul	laranja



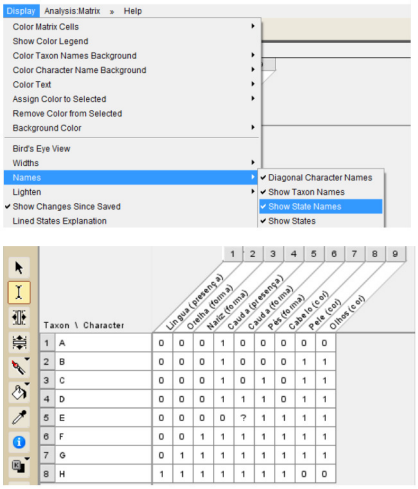
Voltando para a matriz principal (clicando na aba superior é possível alternar entre as matrizes e tabelas), teremos então já definidos os táxons e caracteres, bastando preencher os estados de caráter, utilizando os códigos (0, 1, 2...) atribuídos na planilha **S**.



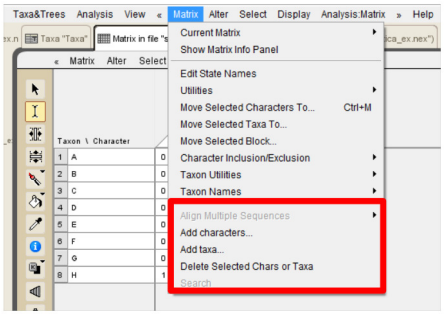
Após preenchida, a matriz ficará assim:



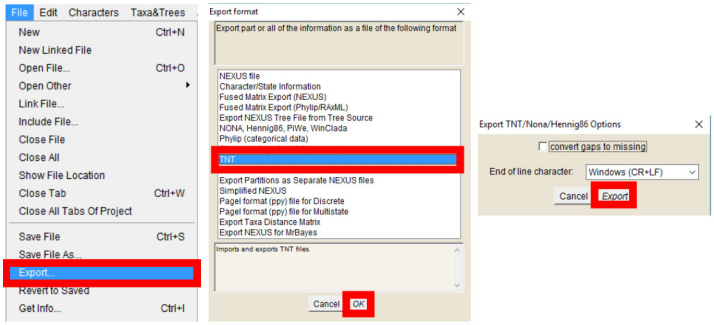
Note que estão exibidos os nomes dos estados de caráter para cada táxon e caráter. Para mudar para a exibição dos códigos numéricos, basta ir no menu [*Display, Names, Show State Names*] e desmarcar. Isso permite uma exibição mais compacta e fácil visualizar.



Caso exista a necessidade de se adicionar ou remover táxons ou caracteres, isso é feito no menu [*Matrix*], na parte inferior existem as funções [*add characters*], [*add taxa*] e [*delete selected characters or taxa*]. Para essa última, é necessário selecionar o caráter ou táxon antes de usar a função.



Por último, para exportar a matriz no formato *.tnt* e utilizar no programa de mesmo nome, deve-se ir no menu [*File, Export*], na próxima tela marcar [*TNT*, OK], e no último box [*Export*], salvando no local de preferência.

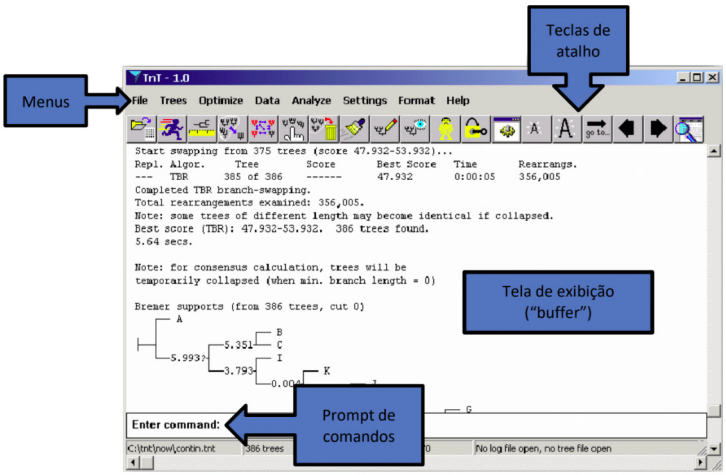


2.11 Realizando uma análise filogenética utilizando o programa TNT (Tree Analysis using New Technologies).

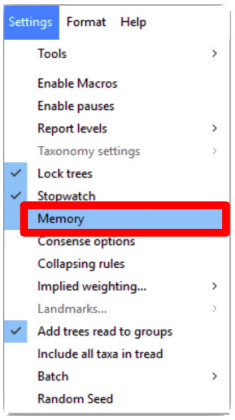
O programa **TNT**, desenvolvido por James S. Farris, Pablo Goloboff e Kevin C. Nixon foi disponibilizado pela primeira vez em 2000 e se tornou gratuito para download em 2008, pela Willi Hennig Society. É o programa mais utilizado atualmente para análises filogenéticas baseadas em parcimônia, dado a grande eficiência das buscas por árvores mais parcimoniosas, bem como uma versão com interface gráfica simples e intuitiva, ideal para usuários iniciantes ou pouco familiarizados com a utilização de linhas de comando. Assim como o Mesquite, o programa executa várias funcionalidades permitindo usar diferentes algoritmos de busca por árvores mais parcimoniosas

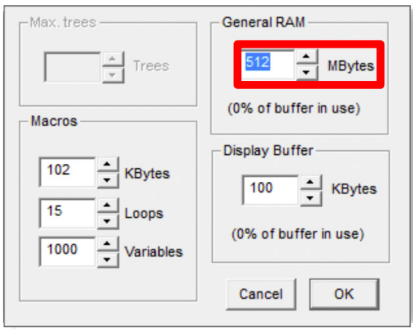
além de várias funções complementares como o uso de ordenação e pesagem de caracteres, consensos, medidas de suporte, otimização de caracteres, exibição de sinapomorfias, além de várias outras. Aqui, iremos nos ater ao uso da busca exata e otimização de caracteres e visualização de sinapomorfias, além de alguns aspectos gerais do programa.

A tela abaixo é a tela principal do TNT. Na parte superior, estão os menus onde encontramos as funcionalidades do programa e, logo abaixo, algumas dessas funções também podem ser acessadas através das teclas de atalhos. A tela principal ou *buffer* exibe as ações tomadas e os resultados dessas ações, podendo também armazenar as árvores resultantes das análises, caso o usuário deseje.

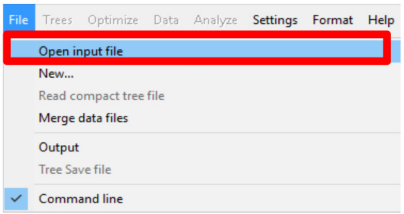


Antes de realizar qualquer procedimento, a primeira ação que se deve realizar ao usar o TNT pela primeira vez é ajustar a quantidade de memória RAM do seu computador que o programa poderá utilizar. Isso pode ser feito no menu [*Settings, Memory*] e na tela seguinte, em [*General RAM*] defina 512 MBytes. Essa quantidade geralmente é suficiente para a maioria das análises, mas caso tenha uma memória RAM considerável em seu computador, e queira, considere dobrar esse valor.



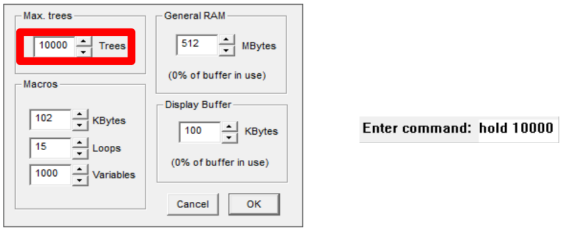


Agora devemos abrir a matriz, aquela montada no Mesquite. Para isso, deve-se ir em [*File, Open input file*] ou usando o atalho abaixo.

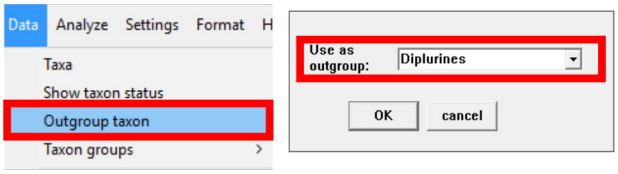


Caso deseje, é possível visualizar a matriz (para matrizes muito grandes, isso não é possível) clicando em [*Data, Show Matrix*]. Esse é um passo opcional, no entanto. Esse passo se refere apenas a visualizar a matriz, uma vez que ela já está aberta, o programa já pode acessá-la.

Agora, com a matriz aberta, deve-se voltar no menu onde foi configurada a memória RAM [*Settings, Memory*] e definir a memória de árvores. Isso é necessário para que, no caso do programa acessar mais árvores que o *default* definido (100 árvores) durante as buscas, o programa não interrompa as análises. Podemos definir nesse menu um limite até 99999 árvores. Usando o prompt de comandos, é possível usar valores maiores com o comando `hold N`, sendo N o número de árvores desejadas. Se quisermos então, definir 100.000 árvores deve-se usar `hold 100000`. Por hora, uma memória de 10.000 árvores basta. O limite do valor da memória de árvores tem relação com a alocação de memória RAM e as capacidades do próprio computador.



O TNT irá interpretar que o primeiro táxon (primeira linha) da sua matriz de dados é o seu grupo externo **mais externo**, ou seja, o que define a posição da raiz da árvore e a polarização dos caracteres nos demais táxons. Caso o grupo externo não esteja na primeira linha ou se deseje redefinir o grupo externo, basta ir em [*Data, Outgroup taxon*] e então selecionar o novo grupo externo na caixa de opções.



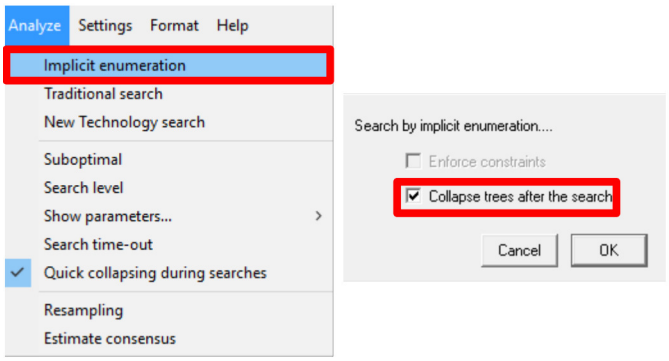
Agora podemos partir para a análise filogenética, que será realizada usando uma busca exata, nomeada no TNT de *Implicit enumeration*. Essa busca é chamada de exata pois garante que encontraremos a ou as árvores mais parcimoniosas com certeza, diferente das buscas heurísticas, também implementadas no TNT. Essas últimas não nos dão a certeza de ter encontrado a árvore com o menor número de passos, apesar de serem bastante eficientes e provavelmente também conseguirem encontrar a(s) árvore(s) mais parcimoniosa(s) nas buscas.

Para realizar uma busca exata, existem dois caminhos: um deles é testar como os caracteres evoluem em todos os cladogramas possíveis, para o número de táxons estudados. Esse método é muito trabalhoso computacionalmente e o número de árvores cresce exponencialmente conforme o número de táxons aumenta, tornando essa opção pouco viável. Uma alternativa a essa busca exaustiva, passando por todas as possibilidades, é o chamado algoritmo de *branch-and-bound*, que é o que o tipo de busca exata implementada no TNT, através da opção `Implicit enumeration`.

O algoritmo de *branch-and-bound* funciona da seguinte forma: primeiro o programa gera uma árvore inicial, adicionando aleatoriamente todos os táxons em uma posição qualquer, sendo a árvore resultante, quase certamente, uma que não é a mais parcimoniosa possível. Os caracteres são otimizados (isso é, é avaliada a evolução mais parcimoniosa do conjunto de caracteres naquela topologia) e o comprimento dessa árvore é tomado como um limite superior provisório (bound). A partir disso, o algoritmo começa a construir outras árvores por adição subsequente de terminais, usando uma busca exaustiva (tentando todas as possibilidades), e avaliando para cada terminal adicionado e otimização feita, se o valor em número de passos está abaixo, igual ou superior do valor do bound. A cada terminal adicionado, o número de "rotas" possíveis para se tentar adicionar os terminais aumenta exponencialmente. No entanto, caso uma ou mais dessas rotas tenha atingido o valor superior ao do bound, ela é abandonada, já que qualquer adição a ela não poderá produzir árvores mais curtas que o bound, e logo, não são as mais parcimoniosas. Sendo o valor menor ou igual que o do bound, o algoritmo continua fazendo essas adições e as verificações de tamanho até chegar a uma árvore final, com todos táxons adicionados, que se tiver o valor igual ao bound, é salva como a mais parcimoniosa (algo raro na primeira rodada), ou se for menor que o bound, esse novo comprimento (mais curto) é definido como novo limite superior, e se repete a análise até que o valor do comprimento do bound e da árvore final se igualem. A ideia é evitar testar árvores em que a otimização dos caracteres leve a valores em número de passos certamente maiores. Dessa forma, evita-se frequentemente parte do custo computacional das buscas exaustivas, e ainda assim se mantém a confiabilidade de ser uma busca exata.

Esse tipo de busca é viável até no máximo 30 táxons (esse número podendo ser menor, dependendo da complexidade dos dados). A partir disso, deve-se recorrer às buscas heurísticas. Para mais detalhes sobre essas, ver os materiais de leitura complementar.

Para realizar a busca exata, deve-se clicar em [*Search, Implicit enumeration*]. Em seguida, abrirá uma tela, onde deve-se marcar o item [*Colapse trees after the search*]. O que essa opção faz, simplificando um pouco, é garantir que nenhum cladograma que resulte dessa análise tenha clados (=grupos monofiléticos) sem alguma sinapomorfia que o suporte (chamados de ramos de comprimento 0). Esses ramos de comprimento 0 são artefatos matemáticos da busca, e não têm sentido dentro da sistemática filogenética (lembrando da associação necessária entre grupos monofiléticos e sinapomorfias).

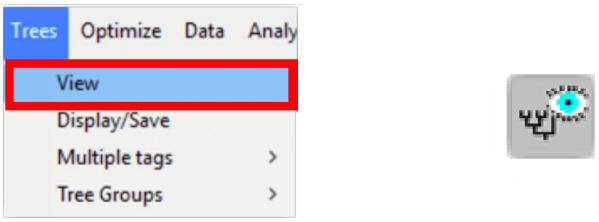


Depois de concluída a busca, será exibida na tela principal a seguinte mensagem, informando a busca realizada, o número de árvores encontradas, o número de passos para essas árvores e o tempo gasto na busca.

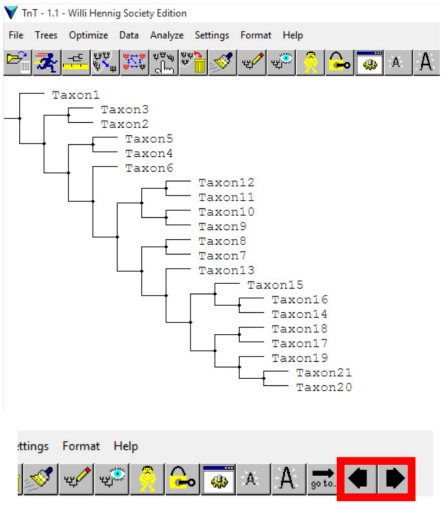
Nesse caso exemplificado, a busca encontrou sete árvores igualmente parcimoniosas, com 148 passos cada. Essas são as informações sobre a busca que geralmente nos interessam.

```
Implicit enumeration, 7 trees found, score 148.
Time 100.64 secs.
```

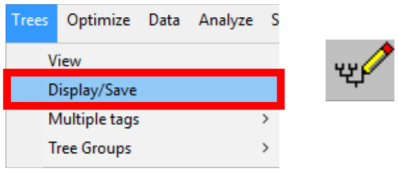
Para visualizar as árvores encontradas, deve-se clicar em [*Trees, View*] ou usar o atalho.



Será exibida a tela de visualização de árvores, e para alternar entre as árvores mais parcimoniosas, deve-se usar as setas no menu de atalho. Observe que o TNT rotula as árvores (e na verdade, tudo mais que ele conta) a partir do número 0. Ou seja, se foram encontradas sete árvores, elas serão as árvores de 0 a 6 e não de 1 a 7. Essa informação pode ser útil para algumas funções onde se precisa definir qual árvore será utilizada em um determinado procedimento.

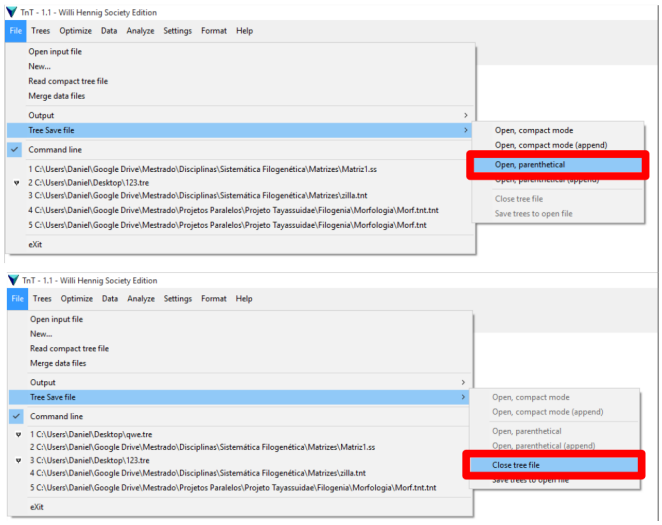


As árvores obtidas podem ser salvas de três formas diferentes. Para salvar as árvores na tela principal de exibição (*buffer*) ou um arquivo de imagem (*metafile*), acessamos a edição de árvores no menu [*Trees, Display/Save*] ou utilizando o atalho.

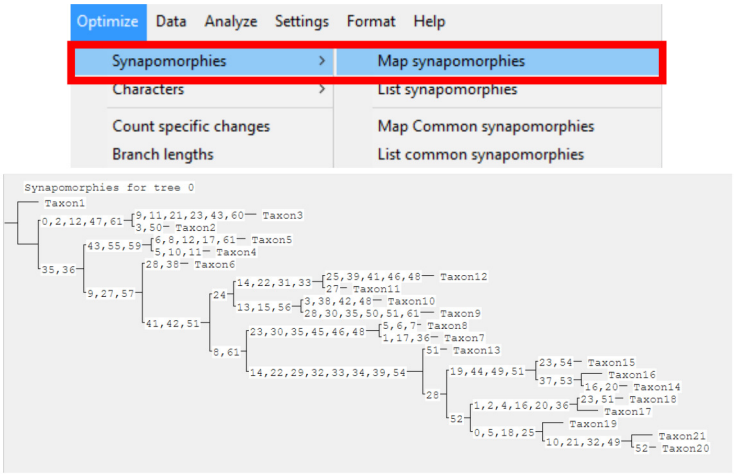


Em seguida, uma tela cinza abrirá exibindo uma das árvores. Nessa tela, apertando `s` a árvore exibida será salva no *buffer* e apertando `m` ela será salva como um arquivo de imagem. Na mesma tela, o comando `h` abre um menu de ajuda (*help*), com os comandos disponíveis para salvamento, edição e manipulação das árvores. Esses dois comandos de salvamento (`s` e `m`) apenas se aplicam para a árvore que está exibida na tela e, portanto, devem ser repetidos para todas as árvores caso queira todas salvas. Para alternar entre as árvores, use `enter` ou `F7` para ir para frente ou `backspace` ou F8 para voltar.

A terceira forma de salvar as árvores permite criar um arquivo para exportar até todas as árvores em um formato que permite ser aberto em outros programas de edição de árvores. Utilizamos mais frequentemente o chamado formato parentético, mas também existe o formato compacto. Para salvar todas as árvores ou alguma delas em formato parentético, vá ao menu [*File, Tree save file, Open parenthetical*] e então [*File, Tree save file, Close tree save file*]. O ato de abrir já salva as árvores que estão na memória no arquivo, mas é necessário fechar esse para que possa ser lido por outros programas.

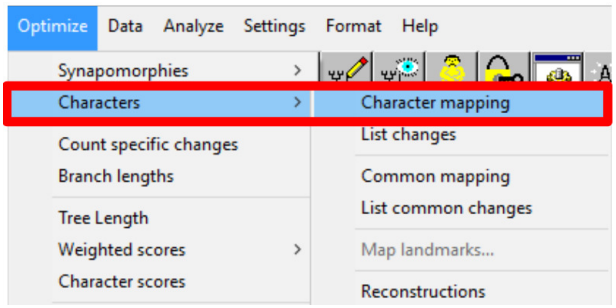


Além de obter as árvores mais parcimoniosas para um conjunto de táxons e caracteres, pode-se querer também entender como os caracteres utilizados para gerar este cladograma serão otimizados nele, ou seja, como eles evoluem da forma mais parcimoniosa em relação à topologia da árvore. Para isso podemos visualizar em conjunto todas as sinapomorfias que suportam os clados da árvore, no menu [*Optimize, Synapomorphies, Map synapormorphies*]. É possível escolher entre mapear as sinapomorfias em alguma árvore específica ou em todas (assumindo que existam mais de uma). Para alternar entre as árvores, use `enter` ou `F7` para ir para frente ou `backspace` ou `F8` para voltar.

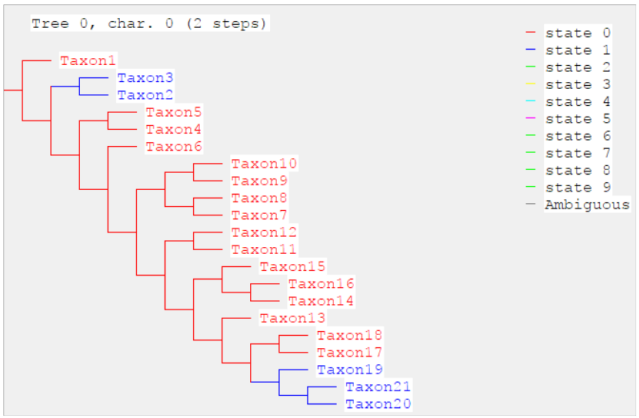


As sinapomorfias se referem aos números dos caracteres na matriz (lembre-se que o TNT começa a contar no 0). São sinapomorfias apenas os caracteres nos ramos que unem ao menos dois terminais, os números nos ramos terminais são suas autapomorfias.

Além de mapear as sinapomorfias em conjunto, é possível mapear a evolução de cada caráter individualmente, permitindo observar os padrões de homologias e homoplasias. Para isso, vá em [*Optimize, Characters, Character mapping*].



Assim como para as sinapomorfias, é possível escolher entre mapear os caracteres em alguma árvore específica ou em todas. Também é possível escolher apenas um ou alguns caracteres para ser mapeados. Para alternar entre as os caracteres em uma mesma árvore, use `enter` ou `F7` para ir para frente ou `backspace` ou `F8` para voltar. Após a exibição de todos os caracteres em uma árvore, o programa passa automaticamente para a próxima.



Os estados do caráter são associados a uma cor e a legenda do lado direito estabelece a referência entre os estados e as cores. Neste exemplo, o estado 0 (vermelho) é uma provável plesiomorfia pois está presente no grupo externo (Taxon1) e na maioria do grupo interno. Além disso, o estado 1 (azul) é uma homoplasia, que aparece duas vezes independentemente. No entanto, nesses dois clados azuis, ela é uma sinapomorfia para cada um. Esse é um ponto importante, dependendo da escala que se analisa (mais ou menos inclusiva), uma mesma característica pode ser simultaneamente uma sinapomorfia e uma homoplasia. Nesse caso, o estado 1 é sinapomofia para o clado (Taxon19(Taxon20+Taxon21)), sinapomorfia para (Taxon2+Taxon3) e homoplasia ao nível da árvore como um todo.

Existem dois comandos que podem ser úteis durante qualquer momento ao usar o TNT, um deles a **pá de lixo**, que limpa o *buffer* completamente, e o **cesto de lixo**, que apaga as árvores na memória. Eles podem ser acessados mais facilmente pelos atalhos abaixo.

Por fim é possível, caso o usuário deseje, salvar todo a informação do *buffer* em um arquivo de *output*. Esse arquivo pode ser aberto a qualquer momento em [*File, Output file, Open output file*]. Todo *buffer* até aquele momento, será salvo. Para atualizar qualquer adição ao *buffer*, basta clicar em [*File, Output, Save display buffer*]. Assim como para os arquivos de árvores, é necessário fechar o arquivo em [*File, Close output file*]. O *output* pode ser aberto em editores de texto comuns.

2.12 Bibliografia e leitura recomendada

FARRIS, James. The logical basis of phylogenetic analysis. 1983. In: Advances in Cladistics proceedings of the second meeting of the Willi Hennig Society (Platnick N, Funk VA, eds.). Columbia University Press, New York: 1-36.

GOLOBOFF, Pablo A. Estimating character weights during tree search. Cladistics, v. 9, n. 1, p. 83-91, 1993.

HENNIG, Willi; DAVIS, D. Dwight. Phylogenetic systematics. University of Illinois Press, 1999.

HULL, David L. Science as a process: an evolutionary account of the social and conceptual development of science. University of Chicago Press, 2010.

RIDLEY, Mark. Evolution. Malden. 2004.

SIDDALL, Mark E. Measures of support. Techniques in molecular systematics and evolution (R. DeSalle, G. Giribet, and W. Wheeler, eds.). Birkhäuser Verlag, Basel, p. 80-101, 2002.

WAGNER, Gunter P. A research programme for testing the biological homology concept. Homology, p. 125-134, 1999.

WILEY, Edward Orlando; LIEBERMAN, Bruce S. Phylogenetics: theory and practice of phylogenetic systematics. John Wiley & Sons, 2011.

Capítulo 3
Filogenética Molecular

Cayo Augusto Rocha Dias & José Eustáquio dos Santos Júnior

3.1 Introdução

Como apresentado no capítulo anterior, constitui parte primordial da sistemática, o estudo das relações históricas entre os seres vivos. Este, quando baseado em dados moleculares, por vezes, é denominado **Sistemática Molecular**. Note, contudo, que o termo **molecular** não se refere aos princípios que fundamentam a classificação biológica, mas ao conjunto de dados que serve de base para a proposição de hipóteses sobre as relações de parentesco entre os organismos. Assim, a Sistemática Molecular estaria inserida na Sistemática Filogenética e, nesse contexto, aquela seria melhor denominada como **Filogenética Molecular**.

O emprego de caracteres moleculares remonta ao início do século XX, quando a precipitação de proteínas do sangue foi utilizada por George Nuttall para inferir as relações evolutivas entre animais. Ao longo desse mesmo século, diferentes fontes de dados moleculares foram incorporadas às análises filogenéticas, sendo as precursoras as fontes indiretas, tais como eletroforese de proteínas e hibridização de DNA. Essas foram progressivamente substituídas pelo uso de sequências de aminoácidos (proteínas) e de nucleotídeos (DNA), sobretudo a partir dos anos 1960 e décadas seguintes, quando se iniciou o desenvolvimento dos primeiros métodos de sequenciamento desses polímeros. Ainda nesse período, foram também publicados os primeiros estudos a fazerem uso das sequências de aminoácidos em análises filogenéticas.

Bastaram algumas décadas, no entanto, para que o DNA se tornasse a principal fonte de informação em estudos filogenéticos. Particularmente por seu maior conteúdo informacional quando comparado às suas alternativas moleculares, as proteínas e as fontes indiretas. Isso ocorre porque nem sempre a mudança em uma das bases de um segmento de DNA irá refletir-se em mudanças nos aminoácidos da proteína por ele codificada. Além disso, vale ressaltar que grande parte dos genomas de eucariotos é composta por regiões não codificadoras de proteínas.

As análises filogenéticas baseadas nos caracteres moleculares seguem, essencialmente, as mesmas etapas apresentadas para os caracteres morfológicos no capítulo anterior. Elas têm início com a elaboração de uma matriz de dados que deverá conter os organismos de interesse (grupo interno) e os grupos externos. Essa etapa é crucial, pois dela dependem todas as demais. Procede-se então ao processo de inferência das relações filogenéticas por diferentes métodos, que podem ou não exigir etapas adicionais. O texto que será apresentado neste capítulo apresenta uma abordagem introdutória dessas etapas.

3.2 Conceitos básicos em evolução molecular

3.2.1 Tipos de mutação

A informação genética armazenada no DNA encontra-se na forma de sequências de nucleotídeos que são caracterizados pela base nitrogenada contida em cada um. As bases nitrogenadas podem ser classificadas em dois grupos: as **pirimidinas** e as **purinas**. Fazem parte do primeiro grupo citosina, timina e uracila (embora esta última só esteja presente nas moléculas de RNA), ao passo que o segundo é formado por adenina e guanina. Suas iniciais A, G, C e T (ou U, no caso do RNA) compõem o alfabeto usado para representar a informação contida nos ácidos nucleicos.

Parte da informação genética é transcrita nos diferentes tipos de RNA (mensageiro, transportador, entre outros) e alguns deles, especificamente os RNAs mensageiros (mRNA), são traduzidos em proteínas. A correspondência entre o alfabeto de quatro letras dos ácidos nucleicos e o de 20 letras das proteínas estabelece-se por meio de um **código genético** quase universal em que três bases correspondem a um aminoácido. Alguns aminoácidos, no

entanto, são codificados por mais de um **códon**, nome dado a cada uma das combinações possíveis das trincas de bases do mRNA. Por essa razão o código genético é dito **degenerado** (ou redundante).

Como são quatro nucleotídeos e cada códon é composto por três, existem 64 possibilidades ($4^3 = 64$). Além dos 20 aminoácidos primários, existem três códons de terminação (UAA, UAG, UGA). Vale ressaltar que existem variações sobre o "código padrão" descritas na literatura (por exemplo: UGA pode ser traduzido como Triptofano em algumas espécies), que devem ser levadas em consideração pelos usuários durante as análises. A redundância do Código genético reduz as chances de que mutações pontuais alterem a sequência de aminoácidos. Caso não houvesse redundância, existiriam 44 códons de terminação que levariam a interrupção da síntese proteica. Veja um exemplo da redundância do código genético na tabela 3.1.

O exemplo acima representa as mudanças em cada base do códon. Entre os aminoácidos existem aqueles que são representados por apenas um códon e os que são representados por até seis códons diferentes. Mudanças em cada uma das bases de um códon têm probabilidades diferentes de resultarem na mudança do aminoácido codificado, sendo a terceira e a segunda bases aquelas que apresentam o maior e menor probabilidades de mudança respectivamente. Ou seja, mudanças na segunda base do códon, resultarão em substituição do aminoácido codificado. Uma explicação para que não ocorra mudança na segunda base dos códons é que códons com pirimidinas na segunda base geralmente codificam aminoácidos hidrofóbicos, enquanto códons com purinas codificam aminoácidos carregados ou polares. Sendo assim uma mudança nos aminoácidos pode alterar a conformação da proteína afetando a sua função. Algo que não ocorre com a terceira base, onde mais da metade das mudanças possíveis não ocasionam modificação do aminoácido. Quanto a esse aspecto, as mutações, assim denominadas as mudanças na sequência nucleotídica de caráter permanente, podem ser classificadas como **sinônimas**, caso não haja mudança no aminoácido codificado, ou **não sinônimas**, se a mudança acontece.

As mutações podem ocorrer entre bases de uma mesma categoria, ou seja, entre bases púricas ($A \leftrightarrow G$) ou entre bases pirimídicas ($C \leftrightarrow T$), sendo chamadas de **transições**. De outro modo, se as mutações ocorrem entre bases de categorias distintas, ou seja, entre uma purina e uma pirimidina ou vice-versa ($A \leftrightarrow C$, $A \leftrightarrow T$, $G \leftrightarrow C$, $G \leftrightarrow T$), serão chamadas de **transversões**.

Mutações podem, além disso, caracterizar-se pela perda, **deleção**, ou ganho, **inserção**, de um ou mais nucleotídeos. Quando a adição de novos nucleotídeos ocorre em uma escala maior, ela pode dar origem a novas cópias de segmentos genômicos. A **duplicação gênica** é o exemplo mais relevante desse tipo de mutação no contexto filogenético, já que a existência de múltiplas cópias de um determinado gene pode levar a complicações durante a inferência filogenética.

Tabela 3.1 Aminoácidos e os códons a eles associados, ilustrando a redundância do código genético.

Aminoácidos	Metionina	Histidina	Isoleucina	Treonina	Serina
Códons	AUG	CAC	AUU	ACU	UCC
		CAU	AUC	ACC	UCU
			AUA	ACA	UCA
				ACG	UCG
					AGC
					AGU

3.2.2 Evolução molecular e teoria neutra

Como se pode depreender a partir do tópico anterior, as mutações são, em última análise, a fonte da variação em âmbito molecular. A existência desta, por sua vez, é condição necessária para que ocorra o processo evolutivo em âmbito nucleotídico. Assim como no caso dos traços fenotípicos, as variantes gênicas também estão sujeitas aos principais mecanismos evolutivos: seleção natural e deriva genética.

Seja por suas contribuições ao fitness de um organismo (positiva ou negativa), ou por razões estritamente aleatórias, as formas alternativas de um gene, que surgem após eventos de mutação, podem ter suas frequências aumentadas ou diminuídas ao longo das gerações. A frequência com que novas variantes surgem em uma população é chamada **taxa de mutação**. Uma variante pode ter sua frequência elevada de tal modo que se torna a única forma disponível, substituindo as demais. À frequência com que isso ocorre, dá-se o nome de **taxa de substituição**.

Sob uma perspectiva **neutralista** da evolução, que considera que a maior parte das mudanças evolutivas, em âmbito molecular, são explicadas pela deriva genética, as taxas de substituição são iguais às taxas de mutação. O que não acontece sob uma concepção quase neutra da evolução, que prediz que novas variantes, ainda que moderadamente vantajosas, poderão ser fixadas ou perdidas por deriva genética ou seleção natural, sendo maior o efeito da deriva quanto menor o tamanho populacional.

É importante destacar que as taxas de substituição podem variar ao longo do tempo, a despeito do que sugeria a primeira hipótese de **relógio molecular** proposta por Zuckerkandl e Pauling (1962), que pressupunha que as diferenças se acumulavam de maneira uniforme. Sabe-se, atualmente, que as taxas podem variar entre diferentes organismos, diferentes regiões do genoma e até mesmo entre as diferentes bases de um dado segmento de DNA. Diante disso, as hipóteses de relógio molecular foram **"relaxadas"** de modo que pudessem levar em conta a possibilidade de variação nessas taxas.

3.3 Dados genéticos: bancos de dados de sequências e principais formatos

3.3.1 GenBank

A informação genética utilizada como base para análises filogenéticas pode ser gerada em laboratório, a partir do sequenciamento de seguimentos específicos do DNA, aqui denominados **marcadores moleculares**, dos organismos de interesse. Contudo, diante do contínuo desenvolvimento de tecnologias de sequenciamento de DNA e a produção exponencial desses dados, fez-se necessário o desenvolvimento de bancos de dados que facilitassem o armazenamento e recuperação de sequências dos mais variados marcadores para diversos organismos.

GenBank é uma das principais bases de dados de acesso público que dispõe de sequências de nucleotídeos para mais de 260 mil espécies descritas. Os registros estão agrupados em divisões que refletem o grupo taxonômico ou a estratégia de sequenciamento empregada. Além das sequências propriamente ditas, cada registro contém **anotações** que trazem informações adicionais tais como o organismo, região do genoma, localidade, autores, entre outras (veja um exemplo de um registro anotado do *GenBank* em https://www.ncbi.nlm.nih.gov/genbank/samplerecord/#LocusNameB). A cada registro é atribuído um identificador (uma identidade) único chamado **número de acesso**, o qual deve ser utilizado para fazer referência a uma sequência em trabalhos científicos (Figura 3.1).

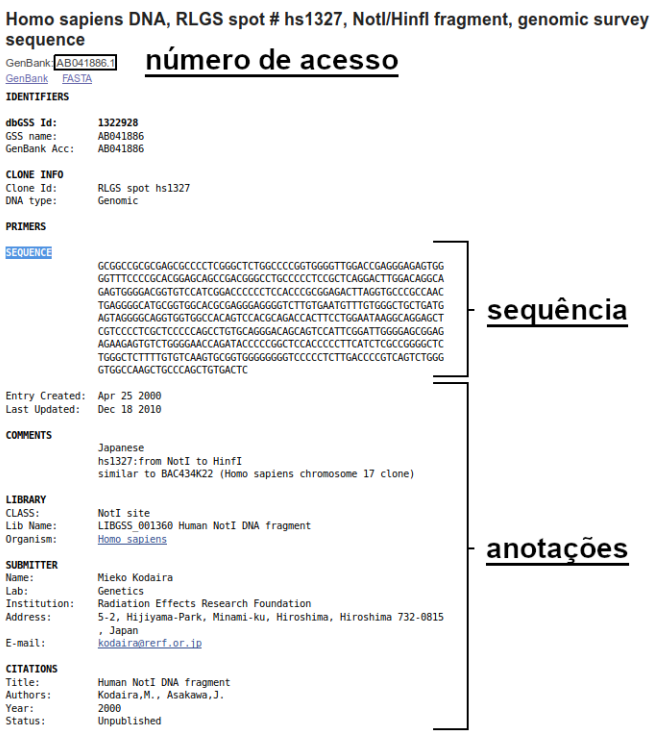


Figura 3.1 Exemplo de um registro do Genbank com o número de acesso e as anotações.

Os registros de sequências do *GenBank* podem ser acessados e recuperados por meio de um sistema chamado *Entrez*. Esse sistema permite a busca, em bancos de dados mais específicos, por palavras-chaves (ou combinações das mesmas) que incluem o número de acesso, os nomes das espécies, o marcador de interesse ou qualquer outro termo presente em um determinado registro (Figura 3.2).

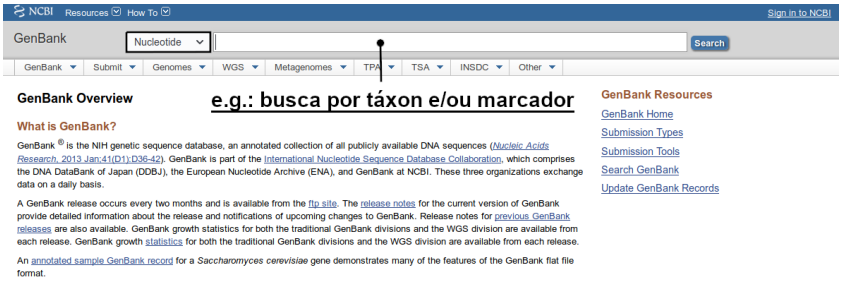


Figura 3.2 Exemplo da página do GenBank e sugestão de busca por sequências nucleotídicas com base no táxon e/ou marcador de interesse.

3.3.2 Sequenciamento de DNA

Como apontado anteriormente, os dados moleculares podem ser obtidos através de técnicas de sequenciamento de DNA, que podem ser divididas em dois conjuntos de métodos. O primeiro e mais comumente utilizado nas últimas décadas é o método de ***Sanger***, que consiste na amplificação do DNA em meio contendo nucleotídeos modificados, chamados didesoxirribonucleotídeos, e posterior sequenciamento dos fragmentos de DNA replicados (*amplicons*) pela DNA polimerase. As técnicas de **sequenciamento de nova geração**, ou **NGS** (da sigla em inglês *Next Generation Sequencing*), formam o segundo conjunto e são caracterizadas pelo alto rendimento resultante de sequenciamento massivo e paralelizado de fragmentos de DNA. Neste último grupo, destacam-se as plataformas da *Illumina*, *Life Technologies* e *Roche*.

Uma metodologia que está em ascensão em estudos filogenéticos é a seleção de genes ultra conservados **UCEs**, do inglês *Ultra Conserved Elements*. A técnica consiste no enriquecimento das regiões altamente conservadas dos genomas (*amplicons* gerados pela DNA polimerase), separação dos UCEs por *beads* magnéticas durante a montagem da biblioteca e posterior sequenciamento nas plataformas de NGS. Com isso há uma diminuição na representação, em pares de bases (pb), do genoma e uma maximização de leituras dos fragmentos de interesse (UCES) sequenciados.

Com a ampliação do uso das novas tecnologias de sequenciamento, o emprego de dados genômicos para inferir as relações entre as espécies está tornando-se cada vez mais frequente em estudos filogenéticos. Contudo, seu uso é limitado pela capacidade de processamento dos computadores e pela falta de métodos de inferência filogenética capazes de lidar de maneira eficiente com grande volume de dados.

3.3.3 Formatos de dados

Há diferentes tipos de formatos em que os dados podem ser recuperados no *GenBank*, aquele que é mais relevante para o contexto deste capítulo é o **FASTA**, que é utilizado para armazenar sequências de nucleotídeos (ou aminoácidos). Cada posição da sequência representada é ocupada por uma única letra indicando um nucleotídeo ou uma lacuna (normalmente referido pelo termo em inglês *gap* e representada por um (–)). As letras utilizadas devem respeitar o padrão IUPAC (*International Union of Pure and Applied Chemistry*), conforme Tabela 3.2, o que significa que, além das usuais iniciais das bases nitrogenadas (A, C, G, T e U), também há símbolos para retratar ambiguidade ou a incerteza relacionada ao nucleotídeo que ocupa determinada posição. A letra N, por exemplo, é usada para indicar que qualquer base pode ocupar uma dada posição, já a letra R é usada para indicar a presença de uma purina, e assim consequentemente. Observação: nas regiões diploides do genoma a presença de uma base ambígua em uma determinada posição indica um *locus* heterozigoto.

Sequências em formato *FASTA* seguem um padrão caracterizado por uma linha que contém a descrição da sequência, que é sempre iniciada pelo símbolo (>), seguida por linhas que compreendem a sequência propriamente dita (Figura 3.3).

Os arquivos de entrada para os programas desenvolvidos para análises de dados genômicos são os *FASTQ*. O arquivo FASTQ é um formato baseado em texto utilizado para armazenar tanto uma sequência de bases quanto seus respectivos valores de qualidade. Nos arquivos FASTQ os valores de qualidade são codificadas de forma compacta, através dos valores de qualidade. Os valores de qualidade são representados como caracteres, de acordo com os códigos ASCII (Figuras 3.4 e 3.5).

Outro tipo de arquivo muito comum no contexto filogenético é o formato *Nexus*, que diferentemente do formato FASTA, é capaz de armazenar diferentes tipos de dados além das sequências de nucleotídeos. Árvores filogenéticas, matriz de dados morfológicos e comandos para um programa são alguns exemplos. Este tipo de arquivo é sempre iniciado com a expressão *#nexus* e a informação é organizada em blocos delimitados por expressões específicas, *begin;* e *end;*, que iniciam e terminam o bloco respectivamente. As sequências e seus identificadores ocupam uma

Tabela 3.2 Codificação utilizada para representar os ácidos nucleicos (padrão IUPAC).

Código de Ácido Nucleico	Significado
A	Adenosina
C	Citosina
G	Guanina
T	Timidina
U	Uracila
R	G A (purina)
Y	T U C (Pirimidina)
K	G T U
M	A C
S	G C
W	A T U
B	G T U C
D	G A T U
H	A C T U
V	G C A
N	A G C T U
. ou -	lacuna de comprimento indeterminado

mesma linha (Figura 3.6). Para mais detalhes sobre o formato Nexus, leiam o artigo publicado por Maddison e colaboradores (1997).

Existem outros formatos de arquivos mais comumente associados a programas específicos e que têm suas próprias estruturas. Alguns desses formatos incluem *Clustal*, *MEGA* ou *PHYLIP* que, não por acaso, são também os nomes dos programas para os quais esses formatos foram desenvolvidos. Cabe ressaltar que há programas dedicados à conversão entre os diferentes formatos, dentre os quais PGDSpider destaca-se pela grande variedade de formatos suportados.

```
>Tarsius_syrichta
AAGTTTCATTGGAGCCACCACTCTTATAATTGCCCATGGCCTCACCTCCTCCCTATTATTTTGCCTAGCAAATACAAACTACGAACGAGTCCACAGTCGAAC
>Lemur_catta
AAGCTTCATAGGAGCAACCATTCTAATAATCGCACATGGCCTTACATCATCCATATTATTCTGTCTAGCCAACTCTAACTACGAACGAATCCATAGCCGTAC
>Homo_sapiens
AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGGCTTACATCCTCATTACTATTCTGCCTAGCAAACTCAAACTACGAACGCACTCACAGTCGCAT
>Pan
AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTTACATCCTCATTATTATTCTGCCTAGCAAACTCAAATTATGAACGCACCCACAGTCGCAT
>Gorilla
AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTTACATCATCATTATTATTCTGCCTAGCAAACTCAAACTACGAACGAACCCACAGCCGCAT
>Pongo
AAGCTTCACCGGCGCAACCACCCTCATGATTGCCCATGGACTCACATCCTCCCTACTGTTCTGCCTAGCAAACTCAAACTACGAACGAACCCACAGCCGCAT
>Hylobates
AAGCTTTACAGGTGCAACCGTCCTCATAATCGCCCACGGACTAACCTCTTCCCTGCTATTCTGCCTTGCAAACTCAAACTACGAACGAACTCACAGCCGCAT
>Macaca_fuscata
AAGCTTTTCCGGCGCAACCATCCTTATGATCGCTCACGGACTCACCTCTTCCATATATTTCTGCCTAGCCAATTCAAACTATGAACGCACTCACAACCGTAC
>M_mulatta
AAGCTTTTCTGGCGCAACCATCCTCATGATTGCTCACGGACTCACCTCTTCCATATATTTCTGCCTAGCCAATTCAAACTATGAACGCACTCACAACCGTAC
>M_fascicularis
AAGCTTCTCCGGCGCAACCACCCTTATAATCGCCCACGGGCTCACCTCTTCCATGTATTTCTGCTTGGCCAATTCAAACTATGAGCGCACTCATAACCGTAC
>M_sylvanus
AAGCTTCTCCGGTGCAACTATCCTTATAGTTGCCCATGGACTCACCTCTTCCATATACTTCTGCTTGGCCAACTCAAACTACGAACGCACCCACAGCCGCAT
>Saimiri_sciureus
AAGCTTCACCGGCGCAATGATCCTAATAATCGCTCACGGGTTTACTTCGTCTATGCTATTCTGCCTAGCAAACTCAAATTACGAACGAATTCACAGCCGAAC
```

Figura 3.3 Estrutura de um arquivo FASTA. As linhas iniciadas por (>) correspondem às descrições das sequências. Essas linhas precedem as linhas que contêm as sequências propriamente ditas.

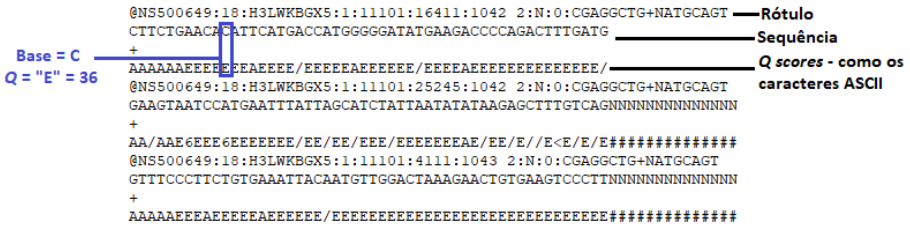


Figura 3.4 Estrutura de um arquivo FASTQ. As linhas iniciadas por "@"correspondem às descrições das leituras, seguida pela linha com a sequência, o caractere (+) e os Qscores como na tabela ASCII. Ver Figura 3.5.

Q	P	ASCII		Q	P	ASCII		Q	P	ASCII	
0	1.00000	33	!	15	0.03162	48	0	29	0.00126	62	>
1	0.79433	34	"	16	0.02512	49	1	30	0.00100	63	?
2	0.63096	35	#	17	0.01200	50	2	31	0.00079	64	@
3	0.50119	36	$	18	0.01585	51	3	32	0.00063	65	A
4	0.39811	37	%	19	0.01259	52	4	33	0.00050	66	B
5	0.31623	38	&	20	0.01000	53	5	34	0.00040	67	C
6	0.21623	39	'	21	0.00794	54	6	35	0.00032	68	D
7	0.19953	40	(	22	0.00631	55	7	36	0.00025	69	E
8	0.15849	41	)	23	0.00501	56	8	37	0.00020	70	F
9	0.12589	42	*	24	0.00398	57	9	38	0.00016	71	G
10	0.10000	43	+	25	0.00316	58	:	39	0.00013	72	H
11	0.07943	44	´	26	0.00251	59	;	40	0.00010	73	I
12	0.06310	45	-	27	0.00200	60	<	41	0.00008	74	J
13	0.05012	46	.	28	0.00158	61	=	42	0.00006	75	K
14	0.03981	47	/								

Figura 3.5 Tabela de caracteres base ASCII para Illumina, Ion Torrent, PacBio e Sanger e probabilidades de erro (P). Os Qscores são frequentemente representados como caracteres de ASCII.

```
#NEXUS

[ Arquivo de exemplo do MrBayes ]

begin data;
   dimensions ntax=12 nchar=898;
   format datatype=dna interleave=no gap=-;
   matrix
Tarsius_syrichta   AAGTTTCATTGGAGCCACCACTCTTATAATTGCCCATGGCCTCACCTCCTCCCTATTATTTTGCCTAGCAAATACAAACTACGAACGAGTCCACAGTCGAACAATAGCACTAGC
Lemur_catta        AAGCTTCATAGGAGCAACCATTCTAATAATCGCACATGGCCTTACATCATCCATATTATTCTGTCTAGCCAACTCTAACTACGAACGAATCCATAGCCGTACAATACTACTAGC
Homo_sapiens       AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGGCTTACATCCTCATTACTATTCTGCCTAGCAAACTCAAACTACGAACGCACTCACAGTCGCATCATAATCCTCTC
Pan                AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTTACATCCTCATTATTATTCTGCCTAGCAAACTCAAATTATGAACGCACCCACAGTCGCATCATAATTCTCTC
Gorilla            AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTTACATCATCATTATTATTCTGCCTAGCAAACTCAAACTACGAACGAACCCACAGCCGCATCATAATTCTCTC
Pongo              AAGCTTCACCGGCGCAACCACCCTCATGATTGCCCATGGACTCACATCCTCCCTACTGTTCTGCCTAGCAAACTCAAACTACGAACGAACCCACAGCCGCATCATAATCCTCTC
Hylobates          AAGCTTTACAGGTGCAACCGTCCTCATAATCGCCCACGGACTAACCTCTTCCCTGCTATTCTGCCTTGCAAACTCAAACTACGAACGAACTCACAGCCGCATCATAATCCTATC
Macaca_fuscata     AAGCTTTTCCGGCGCAACCATCCTTATGATCGCTCACGGACTCACCTCTTCCATATATTTCTGCCTAGCCAATTCAAACTATGAACGCACTCACAACCGTACCATACTACTGTC
M_mulatta          AAGCTTTTCTGGCGCAACCATCCTCATGATTGCTCACGGACTCACCTCTTCCATATATTTCTGCCTAGCCAATTCAAACTATGAACGCACTCACAACCGTACCATACTACTGTC
M_fascicularis     AAGCTTCTCCGGGCGCAACCACCCTTATAATCGCCCACGGGCTCACCTCTTCCATGTATTTCTGCTTGGCCAATTCAAACTATGAGCGCACTCATAACCGTACCATACTACTATC
M_sylvanus         AAGCTTCTCCGGTGCAACTATCCTTATAGTTGCCCATGGACTCACCTCTTCCATATACTTCTGCTTGGCCAACTCAAACTACGAACGCACCCACAGCCGCATCATACTACTATC
Saimiri_sciureus   AAGCTTCACCGGCGCAATGATCCTAATAATCGCTCACGGGTTTACTTCGTCTATGCTATTCTGCCTAGCAAACTCAAATTACGAACGAATTCACAGCCGAACAATAACATTTAC
   ;
end;
```

Figura 3.6 Estrutura de um arquivo Nexus. Note que o mesmo tem início com a expressão #NEXUS e que o bloco de dados está delimitado pelas expressões *begin;* e end;. O texto entre colchetes é um comentário, ou seja, ele é ignorado pelos programas de análise.

3.4 Alinhamento de sequências

Os dados moleculares, assim como os morfológicos, devem ser organizados em uma matriz de dados em que os táxons (organismos) estão dispostos ao longo das linhas e os caracteres (as bases ou aminoácidos) estão organizados ao longo das colunas. Diferentemente de uma matriz de dados morfológicos, em que números são comumente utilizados para representar os estados de caráter, no caso dos dados moleculares, a codificação deverá seguir o padrão de letras da IUPAC.

Alinhar duas ou mais sequências implica propor hipóteses de homologia primária entre as bases que ocupam uma dada coluna da matriz, o que vale dizer que se supõe que as bases ali representadas são homólogas, ou seja, foram herdadas de um ancestral comum. Isso é realizado buscando por regiões de **similaridade** entre pares de sequências. Nesse sentido, os alinhamentos não são apenas como fontes de dados para as análises filogenéticas, pois permitem a identificação de regiões conservadas, que, por sua vez, podem estar associadas à similaridade estrutural ou funcional (Figura 3.7).

Como se pode imaginar, alinhar longas sequências de vários indivíduos à mão não é uma tarefa trivial, além disso, o processo manual poderia introduzir problemas relacionados à subjetividade e replicabilidade. Por isso, o **alinhamento** de sequências é, atualmente, um processo automatizado realizado através de algoritmos computacionais.

Quanto maiores e mais numerosas as sequências, mais complexo, em termos computacionais, torna-se o processo de alinhá-las. Isso significa que o tempo para encontrar uma solução ótima aumenta muito rapidamente com o aumento do número e tamanho das sequências. Consequentemente, foram desenvolvidas várias classes de algoritmos que adotam estratégias distintas para abordar o problema. Algumas valem-se de métodos de otimização, buscando pelo melhor resultado, enquanto outros empregam métodos heurísticos que permitem que se chegue a um resultado satisfatório, ainda que este não represente a melhor solução. A **programação dinâmica**, por exemplo, é uma

Figura 3.7 Exemplo de um alinhamento múltiplo de sequências do gene mitocondrial *ND4* de primatas. Os asteriscos indicam regiões conservadas. A coluna em destaque ilustra a presença de *gaps*.

técnica de otimização que se baseia na atribuição de pontuações para as diferentes possibilidades de alinhamento dos nucleotídeos das sequências comparadas. A pontuação é definida a partir de um esquema composto por uma matriz de custos (de mudanças entre as bases) e penalização de *gaps*. O melhor alinhamento é então escolhido com base na melhor pontuação possível, calculada a partir de uma função matemática (função objetiva) que avalia a **qualidade** do alinhamento. O detalhamento dos cálculos está além do escopo desse texto, no entanto, é importante ter em mente que a solução ótima é dependente do esquema de pontuação adotado, ou seja, para uma dada matriz de custos, a escolha da penalidade atribuída às lacunas pode interferir no resultado. O algoritmo **Needleman-Wunsch** foi um dos precursores da aplicação de programação dinâmica ao alinhamento de sequências e ainda hoje é utilizado, com modificações, em casos específicos. Apesar de garantir que uma solução ótima seja encontrada, seu uso, assim como os outros métodos de programação dinâmica, está restrito a pequenos conjuntos de dados devido ao tempo e recursos demandados para ser executado.

Dentre os métodos de busca heurística, os **alinhamentos progressivos** (ou técnicas progressivas) estão presentes em alguns dos programas mais comumente utilizados como o Clustal, T-Coffee e algumas variantes do MAFFT. De modo geral, os métodos progressivos podem ser divididos em três etapas principais: (1) Alinhar todas as sequências aos pares (usando o método de Needleman-Wunsch) e estimar a similaridade entre elas (calculando a distância, por exemplo); (2) construir uma árvore que represente a relação entre as sequências (uma árvore guia) e (3) combinar os alinhamentos, partindo das sequências mais similares para as mais dissimilares. Apesar de ser um método rápido e de ser capaz de lidar com centenas de sequências, não há garantias de que a solução ótima será encontrada. Outras desvantagens incluem o fato de que erros que ocorrem durante as etapas iniciais são mantidos, sem correção, no alinhamento final e a influência da árvore guia, já que a ordem em que as sequências são adicionadas pode afetar o resultado.

Os **métodos iterativos** representam um aprimoramento sobre as técnicas progressivas, adotando uma abordagem que visa melhorar a qualidade dos alinhamentos iniciais, realinhando-os por várias iterações. Diferentes algoritmos fazem isso de maneiras diversas. **MAFFT**, por exemplo, faz uso de técnicas matemáticas para identificar mais rapidamente regiões homólogas. **Muscle**, por sua vez, envolve duas etapas de alinhamento progressivo, cada uma baseada em uma forma diferente de estimar as distâncias entre as sequências e uma etapa final de refinamento que compreende novos alinhamentos baseados em perfis de sequências.

Existem outros métodos como os **algoritmos genéticos**, que utilizam o conceito de populações de alinhamentos (sujeitas à mutação e recombinação) que evoluem ao longo das gerações por meio de um processo análogo à seleção natural para otimizar uma função objetiva. Métodos probabilísticos como *Hidden Markov Model*, que utiliza um conjunto de sequências não alinhadas de uma determinada família de sequências para identificar regiões conservadas para gerar um modelo que associa uma distribuição de probabilidade para cada posição em uma sequência indicando a verossimilhança de se observar um dado nucleotídeo (ou aminoácido) ou gap. Há ainda métodos que combinam a inferência filogenética e o alinhamento múltiplo de sequências que, apesar das vantagens com relação ao tratamento dos *gaps*, são muito lentos.

A presença de sequências repetitivas, *indels* (inserções e deleções) e posições saturadas por substituições múltiplas podem resultar em regiões de difícil alinhamento ou de alinhamento ambíguo. Nesses casos, mais comuns em sequências não codificadoras de proteínas, em que a determinação da homologia posicional fica comprometida, é recomendável que tais regiões sejam eliminadas do alinhamento final. Com o intuito de tornar esse processo mais objetivo e replicável, já que a supressão de segmentos ambíguos era feita manualmente, foi desenvolvido um método baseado na identificação de regiões conservadas que atendam a alguns parâmetros que podem ser ajustados pelo usuário. O método é implementado no programa Gblocks.

Novamente, vale destacar que o alinhamento é uma das etapas mais importantes do processo de inferência filogenética. Erros que por ventura tenham ocorrido durante este estágio serão propagados às etapas posteriores e as influenciarão.

3.5 Modelos de substituição e particionamento dos dados

3.5.1 Escolha dos modelos de substituição

Após obtido o alinhamento, tem-se em mãos uma matriz de dados moleculares pronta para ser utilizada como base para análises filogenéticas. Especialmente através de métodos que não adotam **modelos evolutivos** explícitos como a máxima parcimônia apresentada no capítulo anterior. Os métodos probabilísticos, no entanto, requerem a definição, a priori, de modelos estatísticos (**modelos de substituição de nucleotídeos**) que descrevem a evolução das sequências que compõem o alinhamento sem retratar os mecanismos subjacentes.

Os modelos adotam como principais premissas a neutralidade, independência e finitude dos sítios. Em outras palavras, assume-se que as substituições não estão sujeitas à seleção, que as mudanças ocorridas em um determinado sítio não interferem nas chances de mudanças nos outros sítios e que cada sítio poderá sofrer múltiplas mudanças.

Matematicamente, os modelos de substituição podem ser caracterizados como processos de Markov (ou markovianos), que, simplificadamente, são processos estocásticos em que a probabilidade de um dado evento futuro, não depende dos estados passados, mas apenas do estado presente. Em termos de biologia molecular, significa dizer que a probabilidade de mudança em um dado sítio, por exemplo, de A para T, em determinado tempo futuro, só depende do estado atual (A), não importando se aquele sítio, anteriormente, era ocupado por qualquer outra base. Probabilidades de mudança entre os diferentes nucleotídeos podem ser calculadas a partir de uma matriz de taxas instantâneas, denotada por **Q**, que retrata as taxas relativas de mudanças entre as bases (Figura 3.8).

$$Q = \begin{bmatrix} A\text{->}A & A\text{->}G & A\text{->}C & A\text{->}T \\ G\text{->}A & G\text{->}G & G\text{->}C & G\text{->}T \\ C\text{->}A & C\text{->}G & C\text{->}C & C\text{->}T \\ T\text{->}A & T\text{->}G & T\text{->}C & T\text{->}T \end{bmatrix}$$

Figura 3.8 Matriz de taxas instantâneas (Q), cada elemento fora da diagonal dessa matriz representa uma taxa relativa de substituição entre bases.

A principal diferença entre os modelos de substituição reside na forma de parametrização da matriz Q. Do mais simples para o mais complexo há uma variação crescente no número de parâmetros adotado por cada um. O modelo de **Jukes e Cantor (JC69)**, por exemplo, é definido por um único parâmetro que representa a taxa de mudança entre as bases, que, neste caso, é igual para todos os casos. **Kimura (K80)** propõe um modelo que leva em conta as taxas diferenciais com que ocorrem transições e transversões, resultando em um modelo com dois parâmetros. O aumento na complexidade dos modelos se dá pela inclusão de parâmetros que descrevem as diferenças nas frequências de cada uma das bases e as diferentes possibilidades de mudança. O modelo **GTR** (da sigla em inglês *General Time Reversible*), por exemplo, é aquele que conta com mais parâmetros, nove ao todo, assumindo taxas diferentes para cada tipo de substituição e frequências distintas para cada uma das bases.

Outro importante parâmetro que pode ser incorporado aos modelos de substituição diz respeito à heterogeneidade das taxas de substituição. É possível então que se leve em conta a variação das taxas de substituição entre os diferentes sítios (cada coluna do alinhamento), o que permite, por exemplo, admitir que alguns sítios tenham maiores taxas de substituição que outros. A variação das taxas de substituição entre os sítios é comumente modelada através de uma distribuição *gama*, que tem um único parâmetro usado para descrever como as taxas variam. É possível ainda combinar aos modelos um parâmetro que indica a proporção de sítios que são invariantes, isto é, que têm taxas de substituição nulas.

O uso de modelos muito simples, que não retratam com o detalhamento necessário a evolução das sequências, pode levar a conclusões equivocadas. Já os modelos mais complexos podem resultar em sobreajuste do modelo aos dados, levando a resultados com menor acurácia. O resultado é, portanto, dependente do modelo evolutivo utilizado no processo de inferência filogenética. Assim, a seleção rigorosa daquele que melhor se ajuste aos dados constitui etapa essencial do processo de inferência filogenética. Programas específicos para seleção de modelos podem tornar o processo de escolha mais objetivo e replicável.

3.5.2 Particionamento dos dados

O uso de vários marcadores concatenados em uma única matriz de dados tem se tornado prática cada vez mais comum. Os marcadores usados podem ter origens (organelas ou núcleo) ou funções (sequências codificadoras de

proteínas, *íntrons*, DNA ribossomal, regiões reguladoras, entre outras) distintas, fato que pode levar, em termos qualitativos, a padrões de substituição distintos entre os diferentes sítios do alinhamento e não apenas uma variação nas taxas relativas de substituição.

Análises que deixam de considerar a variação nos padrões de substituição podem resultar em hipóteses filogenéticas não acuradas, ainda que essas possam apresentar altos valores de suporte de ramos. Lidar com esse problema significa possibilitar que essa variação qualitativa nos padrões de substituição seja incorporada às análises filogenéticas. Uma forma de fazer isso é utilizando **modelos mistos**, que permitem avaliar, para cada sítio do alinhamento, diferentes modelos evolutivos em vez de um único definido a priori. Em outras palavras, para cada sítio poderá ser atribuída uma matriz Q (e seus respectivos parâmetros) específica em vez de uma única para todo o alinhamento. Essa saída, no entanto, é mais adequada a grandes volumes de dados, já que, em razão dos vários parâmetros a serem estimados, um pequeno conjunto de dados pode não conter informação suficiente para estimar todos os parâmetros com a acurácia necessária.

O particionamento (ou combinação) dos dados, que é outra opção, consiste em agrupar, a priori, diferentes sítios ou regiões do alinhamento que se supõe ter evoluído sob os mesmos processos. Dessa maneira, assume-se para cada **partição** um modelo evolutivo distinto que terá seus parâmetros estimados independentemente. O particionamento pode ser realizado "manualmente" ou de modo automatizado. No primeiro caso, as partições são definidas pelo pesquisador de acordo com alguma propriedade biológica compartilhada pelos sítios que serão agrupados. No segundo caso, programas como **PartitionFinder** podem ser utilizados para definir, concomitante e objetivamente, as partições e seus respectivos modelos evolutivos de melhor ajuste.

3.6 Inferências Filogenéticas

Uma vez determinados o alinhamento, o particionamento dos dados e os modelos evolutivos, é possível dar início à etapa de análises filogenéticas. Com relação à forma como os dados serão utilizados nas análises, pode-se classificar os métodos de inferência filogenética em dois grupos, os métodos baseados em distância e os métodos baseados em caracteres. No primeiro caso, a matriz de dados é utilizada para o cálculo de uma matriz de distâncias (genéticas) entre os pares de táxons. Essa matriz é então utilizada para a construção de uma árvore filogenética através de análises de agrupamento, que se utilizam de algoritmos como o *Neighbor-Joining*, que tenta minimizar a soma dos comprimentos de ramos. Já os métodos baseados em caracteres utilizam uma matriz de caracteres discretos para inferir uma ou mais árvores filogenéticas com base em um critério de otimização. Nesse grupo estão a máxima parcimônia (discutida no capítulo anterior) e os métodos probabilísticos, **máxima verossimilhança (ML)** e **inferência bayesiana (IB)**.

Máxima verossimilhança é uma técnica estatística comumente empregada na estimativa de um ou mais parâmetros de interesse. Os valores desses parâmetros, os quais compõem uma função de verossimilhança, são selecionados de modo que se obtenha o valor máximo para essa função, ou seja, os valores que maximizem a probabilidade de se obter os dados observados, dado um determinado modelo. Formalmente, a verossimilhança pode ser enunciada como a probabilidade dos dados dada uma hipótese (ou modelo), denotada por **P(D|H)**. No contexto filogenético, os dados (D) se referem ao alinhamento múltiplo de sequências e a hipótese (H) é representada pela topologia da árvore, os comprimentos dos ramos e os parâmetros do modelo de substituição.

Adotar o critério de máxima verossimilhança implica encontrar os valores dos parâmetros, sobretudo topologia e comprimentos dos ramos, para os quais é máxima a probabilidade de se observar o alinhamento dada a hipótese filogenética avaliada. Desse modo, idealmente, deve-se calcular a verossimilhança levando-se em conta cada possível árvore filogenética (caracterizada por sua topologia e comprimentos dos ramos) e então selecionar aquela que resultar no maior valor. Contudo, ao se considerar que para cada árvore, deve-se calcular a verossimilhança para cada uma das colunas do alinhamento e que o número de árvores possíveis cresce muito rapidamente com o aumento do número de táxons, logo fica claro que não é possível analisar todas as árvores e é necessário recorrer aos métodos heurísticos. De modo geral, parte-se de uma árvore inicial com base na qual se calcula a verossimilhança. Essa árvore é então submetida a sucessivos rearranjos (conduzidos por meio de algoritmos específicos), dando origem a novas árvores. Para cada nova árvore a verossimilhança é calculada e aquela que conferir a maior verossimilhança é selecionada para dar início a uma nova etapa de rearranjos. Esse processo se repetirá até que nenhuma melhoria na verossimilhança seja verificada. Por analogia, é possível imaginar, simplificadamente, o espaço multidimensional de verossimilhança como uma paisagem composta por morros e vales, com alguns morros mais altos do que outros. Quando se busca pela máxima verossimilhança, busca-se alcançar o morro mais alto (o máximo global), mas os métodos heurísticos não garantem que esse objetivo será alcançado, deixando muitas vezes em morros menos altos (máximos locais). Logo, é importante que a busca pela árvore de máxima verossimilhança seja baseada em mais de uma árvore inicial para que sejam reduzidas as chances de se ficar preso em máximos locais. Inferência Bayesiana é o nome dado a um conjunto de métodos estatísticos que utilizam o teorema de Bayes para estimar parâmetros e atualizar a probabilidade associada a uma hipótese (H) a partir da informação contida nos dados disponíveis. Essa "atualização" é expressa na forma de uma probabilidade condicional da hipótese (H) dado um conjunto de

dados observados (D), denotada por **P(D|H)** e denominada **probabilidade posterior**. Seu cálculo leva em conta a probabilidade a priori da hipótese (prior), isto é, a probabilidade da hipótese antes da observação dos dados, e a probabilidade dos dados dada a hipótese (a verossimilhança como discutido anteriormente). A relação entre a probabilidade posterior e a probabilidade a priori é dada por:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

onde P(D|H) corresponde à verossimilhança, P(H) à probabilidade a priori da hipótese e P(D) à probabilidade dos dados.

É possível propor o seguinte exemplo para ilustrar o conceito apresentado acima: Imagine uma caixa com 100 moedas, das quais 80 são moedas honestas e 20 são moedas viciadas. Jogando as moedas para o alto, constata-se que, no caso das moedas honestas, a probabilidade de se obter cara, P(k), é igual a probabilidade de se obter coroa, P(c), sendo ambas 0,50. Já para as moedas viciadas, P(k) = 0,75 e P(c) = 0,25. Suponha que desta caixa uma moeda é tomada ao acaso e lançada ao alto por duas vezes consecutivas, obtendo-se cara em ambas as ocasiões. Pode-se perguntar então qual a probabilidade de que a moeda selecionada seja uma moeda viciada dado que duas caras consecutivas foram observadas, e calcular utilizando o teorema de Bayes.

Como apresentado acima, o cálculo da probabilidade posterior depende da verossimilhança e dos *priors* e, neste caso, ficaria assim:

$$p(v|(k,k)) = \frac{p((k,k)|v) * p(v)}{p((k,k)|v) * p(v) + p((k,k)|h) * p(h)}$$

onde i. p((k,k)/v) é a probabilidade de se observar duas caras dado que a moeda é viciada, sendo igual a 0,75 x 0,75 = 0,5625; ii. p((k,k)/h) é a probabilidade de se observar duas caras dado que a moeda é honesta, sendo igual a 0,50 x 0,50 = 0,25; iii. p(v) é a probabilidade a priori de que a moeda seja viciada, sendo igual à fração de moedas viciadas na caixa, ou seja, 0,20; iv. p(h) é a probabilidade a priori de que a moeda seja honesta, sendo igual à fração de moedas honestas na caixa, ou seja, 0,80.

Dispondo dessas informações e utilizando a fórmula dada acima, é possível calcular a probabilidade posterior de que a moeda do exemplo seja viciada, como:

$$p(v|(k,k)) = \frac{0,5625 * 0.2}{0,5625 * 0.2 + 0.25 * 0.8}$$

Assim, como discutido anteriormente, o uso do teorema de Bayes permitiu que a crença sobre o vício da moeda fosse atualizada após a observação dos dados. Desse modo a probabilidade de se tomar, ao acaso, uma moeda viciada, que era igual 0,20, passou a ser 0,36.

Um aspecto interessante da estatística Bayesiana é que a incerteza dos parâmetros, dos dados e dos resultados (a probabilidade posterior) podem ser representados por distribuições de probabilidade. Uma consequência direta é que, diferentemente da ML, que tem como resultado uma única árvore, o método Bayesiano retorna um conjunto de árvores.

Considerando o contexto filogenético, conforme apresentado anteriormente na seção sobre a máxima verossimilhança, D se refere ao alinhamento múltiplo de sequências e H inclui a topologia da árvore, os comprimentos dos ramos e os parâmetros do modelo de substituição. Dispondo-se de D e H é possível estimar a verossimilhança (conforme discutido anteriormente), mas o cálculo da probabilidade posterior pelo teorema de Bayes ainda depende dos *priors*. Nas análises filogenéticas é comum utilizar *priors* pouco informativos para as topologias, ou seja, admite-se, a priori, que todas as topologias têm igual probabilidade de serem amostradas como uma forma de reduzir a introdução de subjetividade no processo de inferência.

O número de elementos envolvidos no cálculo da probabilidade posterior de uma filogenia faz com que a fórmula apresentada anteriormente se torne bastante complicada para que uma resposta seja obtida analiticamente. Métodos de simulação são empregados para aproximar resultado ao **ótimo**, em especial o método de simulação de Monte Carlo via Cadeias de Markov (MCMC; *Markov Chain Monte Carlo*), que amostra filogenias (aqui representada pela topologia, comprimentos de ramos e parâmetros do modelo de substituição) de acordo com suas probabilidades posteriores. O MCMC é conduzido por várias gerações e, em cada geração, são propostas modificações na topologia, nos valores dos parâmetros e nos comprimentos de ramos. Esses dados são usados para definir o estado a cada geração e então comparar estados de gerações sucessivas, permitindo assim calcular a probabilidade de se aceitar ou não o novo estado proposto. Nesse processo, repetido inúmeras vezes, o algoritmo promove então um passeio aleatório pelo espaço amostral e os estados visitados formam a cadeia de Markov. Uma amostra da distribuição posterior é então obtida a partir dos estados da cadeia de Markov que foram amostrados em intervalos definidos durante a análise. A probabilidade posterior de uma dada árvore filogenética é então aproximada a partir da proporção do tempo em que ela foi registrada nessas amostras.

Conforme discutido, o método Bayesiano resulta em uma distribuição de probabilidade posterior de H dado D. A partir dela é possível estimar a distribuição de probabilidade posterior das topologias, cuja informação sobre

cada clado pode ser sumarizada em um consenso de maioria. Esse aspecto constitui uma vantagem do método Bayesiano, que permite inferir, ao mesmo tempo, a topologia e o suporte dos clados (a probabilidade posterior), não demandando tempo e recursos adicionais para o cálculo do suporte como ocorre com a máxima parcimônia e máxima verossimilhança.

A discussão sobre os métodos de inferência filogenética foi apresentada de maneira sucinta em razão de seu caráter introdutório. Uma noção aprofundada sobre o tema exigiria a apresentação de conceitos matemáticos e métodos computacionais que vão além dos objetivos deste texto. Àqueles que buscam uma apresentação mais detalhada do tema, a bibliografia sugerida ao final do capítulo servirá como um ponto de partida.

3.7 Programas

3.7.1 Obtenção e curadoria das sequências

Como mencionado na seção 3.2 desse capítulo, existem dois grupos principais de métodos de sequenciamento, Sanger e NGS. Para a obtenção de sequências pelo método de Sanger, deve-se levar em consideração a produção de duas leituras para cada indivíduo, uma direta (*forward*) e outra reversa (*reverse*). A partir das leituras é formado um consenso, que é então utilizado nas análises posteriores. Mas é comum que alguns pesquisadores, com o intuito de aumentar a confiabilidade dos consensos, gerem mais leituras de diferentes reações de PCR para cada marcador. Esta estratégia objetiva averiguar inserções/deleções ou inclusões de bases erradas durante a PCR, pela Taq polimerase, uma vez que cada par de leituras é proveniente de uma reação diferente.

Após conferir os eletroferogramas (representação gráfica do resultado do sequenciamento de DNA), se houver erros, como exemplificado na Figura 3.9, leituras adicionais poderão ser sequenciadas. É recomendável que essa etapa seja repetida até a confirmação de todas as posições do fragmento de interesse ou o uso dos códigos da IUPAC. Entre os programas gratuitos para a conferência dos eletroferogramas e a obtenção das sequências consenso está o **BioEdit** (http://www.mbio.ncsu.edu/BioEdit/bioedit.html), disponível apenas para Windows. Para usuários das plataformas Linux e Mac os programas *Phred*, *Phrap* e *Consed* podem ser usados. Eles permitem ler os arquivos de saída do sequenciador e avaliar a qualidade de cada base, montar um consenso a partir de diferentes leituras (diretas e reversas), visualizar e editar a sequência consenso quando necessário (http://www.phrap.org/phredphrapconsed.html).

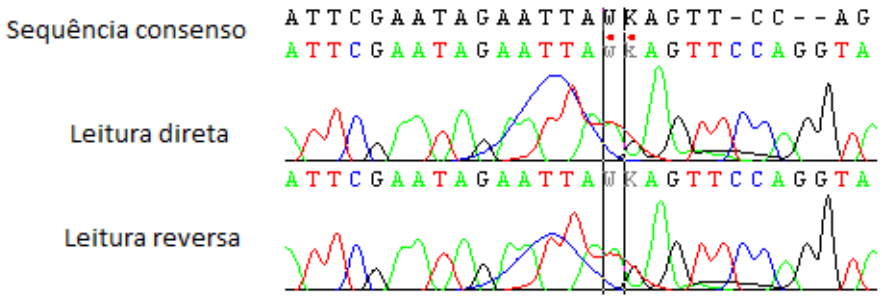


Figura 3.9 Exemplo de um eletroferograma. Pode ser observado, na coluna em destaque, que em uma das posições não foi possível determinar o nucleotídeo tanto na leitura direta quanto na reversa. Devido ao erro a sequência consenso recebeu a letra W (indicando a possibilidade de um T ou um A naquela posição). As cores verde, preta, azul e vermelha nos eletroferogramas correspondem os nucleotídeos cujas bases são Adenina, Guanina, Citosina e Timina respectivamente.

No caso de dados obtidos a partir de NGS (aqui, NGS será usado para fazer referência à plataforma Illumina), um número maior de leituras é necessário para que possa resolver os problemas relacionados aos erros de sequenciamento. Para estimar a cobertura adequada deve-se levar em consideração o tamanho do genoma e o conteúdo de DNA repetitivo. Na inexistência de bons genomas de referência estima-se que seja necessário de 20 a 100 vezes de cobertura. A cobertura mencionada acima é a cobertura média, já que esta não é homogênea ao longo do genoma, podendo variar conforme a representação do material genético (por exemplo: o genoma mitocondrial, em geral, tem uma cobertura mais alta quando comparado ao genoma nuclear devido ao número de cópias mitocondriais no genoma total).

A explicação para a necessidade de aumento da cobertura, no sequenciamento de NGS está no tamanho dos fragmentos gerados (por exemplo: 2 x 100 a 2 x 300pb na plataforma Illumina –NGS e 1000pb no ABI –Sanger) e na alta taxa de erros durante a chamada das bases (medidas pelos valores de qualidade *Phred*), quando comparado com o método de Sanger. Na literatura, o termo *Phred* é usado para se referir ao programa *Phred*, utilizado para calcular

a qualidade das leituras Sanger, e como uma medida de qualidade de nucleobases geradas pelo sequenciamento nas plataformas de NGS (valores de qualidade de qualidade de *Phred*). Nos dados oriundos do Sanger o programa Phred lê os dados do sequenciamento de DNA atribuindo valores de qualidade a cada base chamada (atribuída) em uma dada posição da sequência. Quando as bases são analisadas o *Phred* avalia o traço (*trace*) em torno de cada base chamada, usando quatro ou cinco parâmetros para quantificar a qualidade do traço. Os valores de qualidade de qualidade variam de quatro a cerca de 60, sendo os valores mais altos os de maior qualidade, tabela 3.3. Já em NGS, o formato FASTQ possuem as nucleobases codificas com pontuações *Phred*, como caracteres ASCII ao lado das sequências lidas (ver Figuras 3.4 e 3.4 da seção 3.3.3). Em NGS os valores de qualidade de qualidade de *Phred* são usados para caracterizar a qualidade das sequências, e como uma métrica para comparar a eficácia de diferentes métodos de sequenciamento.

Tabela 3.3 Valores de qualidade *Phred*. Os valores de qualidade estão logaritmicamente vinculados às probabilidades de erros.

Score de qualidade Phred	Probabilidade de erro da base	Precisão
10	1 em 10	90,00%
20	1 em 100	99,00%
30	1 em 1.000	99.9%
40	1 em 10.000	99.99%
50	1 em 100.000	99,999%
60	1 em 1.000.000	99.9999%

Os valores de qualidade Q são definidos como uma propriedade logaritmicamente vinculada às probabilidades de erros de chamada de base $(P)^2$. $Q = -10 \log_{10} P$.

Para os dados de Sanger (por exemplo: ABI 3130 *xl* Genetic Analizer) é recomendável selecionar apenas as bases com um valor de qualidade igual ou superior a 20, portanto, o valor *Q20*. O valor *Q20* corresponde uma precisão de 99% para a base chamada. Uma precisão de 99% (Q20) na chamada da base implicará uma probabilidade de chamar a base incorreta de 1 em 100, o que significa que cada leitura de sequenciamento de 100pb provavelmente conterá um erro (Tabela 3.3). Enquanto o *Q30* é considerado uma referência de qualidade do NGS. Baixos valores de Q podem aumentar os falsos positivos durante a detecção de variantes, o que pode resultar em conclusões imprecisas. A Figura 3.10 a qual demonstra a queda de qualidade em porcentagem (%>Q30), nos ciclos finais. Note que na segunda leitura, parte direita da figura a baixa qualidade no inicio é mais acentuada.

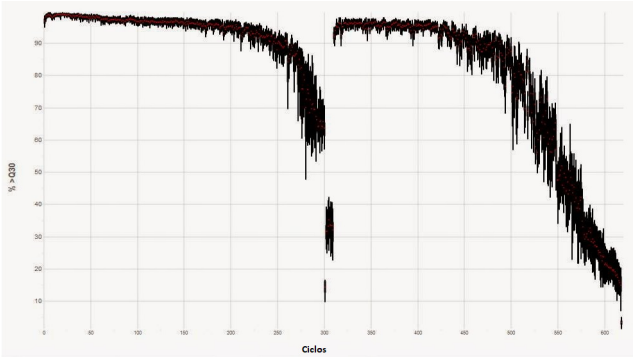


Figura 3.10 Relação entre o número de ciclos e à proporção de bases com Q ≥ 30 das leituras referentes a corridas do Kit v3 600 ciclos da Illumina. Note que há uma diminuição significativa na proporção de leituras com *Qscores* >Q30 em torno de 200 ciclos (primeira leitura a esquerda) e em torno do ciclo 450 da segunda leitura (direita).

Diversos programas e *scripts* foram desenvolvidos para o controle de qualidade e remoção (do inglês *trimming*) das bases com baixa qualidade de dados oriundos de sequenciadores NGS (FastQC, UrQt, Trimmomatic, NGS QC Toolkit, entre outros). Além das bases com baixa qualidade, as sequências utilizadas para identificar as amostras (*barcoding*) e os adaptadores (*index*) também devem ser removidos.

3.7.2 Programas de alinhamento de sequências

Existem programas que possuem algoritmos de alinhamento implementados, como o **MEGA** - *Molecular Evolutionary Genetics Analysis*, que traz os algoritmos de alinhamento **ClustalW** e **MUSCLE**. Há também o AliView, que permite ao usuário a personalização dos algoritmos de alinhamento que serão implementados.

O programa **MEGA** está disponível para *download* nos três principais sistemas operacionais em duas versões, linha de comando ou interface gráfica. Para fazer o *download* basta acessar o site http://www.megasoftware.net/. Já o **AliView** é um programa gratuito de código aberto que preza pela usabilidade e velocidade e dispõe de várias opções para visualização e edição de alinhamentos. Ele pode ser obtido acessando o site http://www.ormbunkar.se/aliview/.

O programa **MAFFT**, além de estar disponível para *download* nos três sistemas (Linux, Mac e Windows), possui uma versão online. Para usar a versão online basta enviar o arquivo com as sequências em formato fasta e escolher os parâmetros do alinhamento (https://mafft.cbrc.jp/alignment/server/). É importante salientar que o arquivo deve conter apenas fragmentos de um mesmo gene e que os parâmetros utilizados devem considerar as características do gene ou fragmento a ser analisado, como no caso de genes ribossomais ou RNA transportadores (RNAts). Nesses dois casos o parâmetro que utiliza a estrutura secundária no alinhamento deve estar selecionada.

No caso dos dados em escala genômica, provenientes do NGS, após as etapas iniciais de bioinformática, incluindo limpeza de leitura (eliminação das bases abaixo de Q30) e eliminação dos adaptadores (sequências *barcoding* utilizadas para a identificação das amostras e/ou adaptadores necessários para o sequenciamento), ocorre a montagem e alinhamento. Entre os softwares e/ou pacotes de softwares utilizados para a montagem e alinhamento estão: **Trinity** (https://github.com/trinityrnaseq/trinityrnaseq/wiki) –UCEs (ver seção 3.2.2) e transcriptomas; PHYLUCE (https://github.com/faircloth-lab/phyluce) –UCEs; **ABySS** (https://github.com/bcgsc/abyss) –genoma. Quando o objetivo é alinhar todos os *loci* individualmente o programa MAFFT, também pode ser utilizado. Após a tradução de *contigs* (sequências consenso obtidas de diferentes leituras) em sequências de aminoácidos, os genes devem ser avaliados com relação à ortologia (exemplo de pipeline OrthologID –http://nypg.bio.nyu.edu/orthologid/. O OrthologID avalia conjuntos completos de genes de todos os táxons e os atribui a conjuntos de genes específicos; em seguida, gera uma árvore pelo método de parcimônia para cada cluster de genes e extrai um ou mais conjuntos de genes ortólogos. O programa **PHYLUCE** (https://github.com/faircloth-lab/phyluce) deve ser utilizado após a montagem dos *contigns* de UCEs para identificar *loci* individuais de UCE a partir dos *contigs* montados. O PHYLUCE também remove os possíveis parálogos (cópias diferentes de um mesmo gene), durante a identificação dos UCEs. Ele também pode ser usado para combinar os contigs, como de táxons sequenciados para UCEs, com táxons sequenciados para genomas ou transcriptomas, e vise-versa, em um único arquivo fasta.

3.7.3 Programas utilizados para o tratamento dos alinhamentos

Tão importante quanto o alinhamento das sequências a serem analisadas é o tratamento dos dados pós-alinhamento. Entre os erros cometidos durante a fase de alinhamento estão a não observância da presença de paralógos e a não remoção das regiões de alinhamento ambíguo. No caso do primeiro problema, basta eliminar os indivíduos que foram sequenciados para as cópias paralógas. Sequências paralógas de um gene, geralmente, são reconhecidas quando três ou mais indivíduos de uma mesma espécie são sequenciados e alinhados. A presença de sequências muito divergentes e de códons de terminação em um gene codificante de proteína é um forte indício de genes com múltiplas cópias (Figura 3.11). Entre os programas utilizados para visualizar e conferir as sequências e os eletroferogramas estão o MEGA, o BioEdit e o Phred, Phrap e Consed.

No caso de genes com a presença de posições mal alinhadas e/ou regiões com mais de uma possibilidade de alinhamento de DNA ou proteína (regiões de alinhamento ambíguo), as posições podem não ser homólogas ou ter sido saturadas por múltiplas mutações. Não sendo possível estabelecer a homologia entre as bases é desejável a eliminação destas regiões antes das análises filogenéticas. Quando falamos de regiões saturadas por mutações, que não tiverem problemas no estabelecimento das homologias entre as bases, a correção das regiões saturadas ocorre através dos modelos mutacionais (ver seção 3.5 para a parte teórica e a seção 3.7.4 para os programas de seleção de modelos).

Um dos programas utilizados para remover as regiões de alinhamento ambíguo é o **Gblocks** que está disponível para Windows, Linux, Mac e online (http://phylogeny.lirmm.fr/phylo_cgi/one_task.cgi?task_type=gblocks). Ele seleciona blocos do alinhamento seguindo um conjunto de condições, como: a presença ou ausência de grandes fragmentos de posições não conservadas contíguas, falta e presença de *gaps* e a alta conservação de posições flanqueadoras. A remoção das regiões de alinhamento ambíguo torna o alinhamento final mais adequado para análises filogenéticas (ver seção 3.4 desse capítulo). O uso do Gblocks para essa finalidade reduz a necessidade de edição manual dos alinhamentos, além de facilitar a reprodução dos alinhamentos para análises posteriores

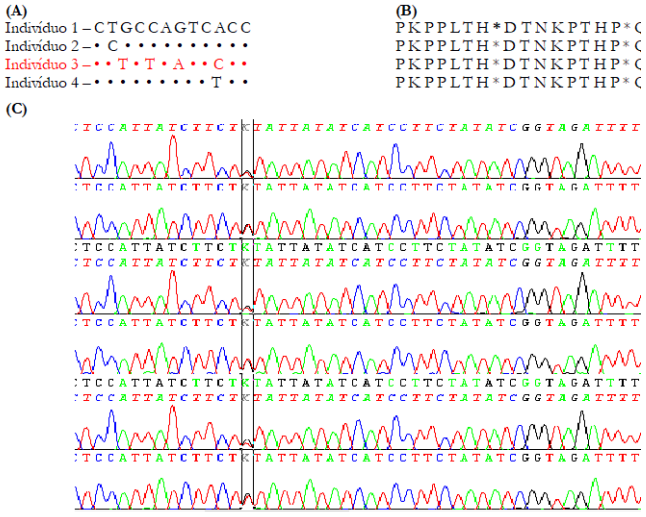


Figura 3.11 Exemplos de como reconhecer a presença de múltiplas cópias de um único gene. (A) sequências muito divergentes de uma mesma espécie; (B) presença de códons de terminação em um gene codificante de proteína, (*representação gráfica do códon de terminação).

realizadas por outros pesquisadores. O programa Gblocks também é utilizado para remover regiões de alinhamento ambíguo em dados de NGS, quando locus ou contigs são alinhados individualmente.

3.7.4 Programas para a seleção dos modelos evolutivos e particionamento dos dados

Vários programas podem ser utilizados para seleção de modelos evolutivos. O **MEGA**, o **jModelTest** e os próprios programas de inferência bayesiana (exemplo: BEAST2). Eles também estão disponíveis para *download* em Windows, Linux e Mac. Como forma de auxiliar na familiarização com esses programas, existem diversos tutoriais e modelos de entrada de dados na literatura. Um exemplo de um programa com tutorial está disponível na página: http://evomics.org/learning/phylogenetics/jmodeltest/. Programas específicos para seleção de modelos como o BEAST2 e jModeltest2 podem tornar o processo de seleção mais objetivo e replicável. Note que eles empregam métodos diferentes, enquanto o BEAST2 usa métodos bayesianos, o jModeltest2 usa o método de máxima verossimilhança, para determinação dos parâmetros dos modelos e critérios de informação para seleção.

O **PartitionFinder** permite inferir os modelos evolutivos de melhor ajuste aos dados ao mesmo tempo que infere o melhor esquema de particionamento. Sua principal vantagem consiste na rapidez com que é capaz de lidar com grandes volumes de dados. Um tutorial sobre o programa, apresentando os formatos de entrada e como interpretar os resultados pode ser encontrado em sua página (http://www.robertlanfear.com/partitionfinder/tutorial/).

3.7.5 Programas para inferências filogenéticas

Como mencionado no item 3.6 deste capítulo, aqui será apresentado apenas os programas que fazem inferências filogenéticas de ML e IB. O programa mais utilizado atualmente para as análises de ML é o **RAxML** (*Randomized Axelerated Maximum Likelihood*). RAxML é um programa para inferências de ML capaz de analisar grandes matrizes em um curto espaço de tempo, quando comparado com outros programas (por exemplo: **PAUP**). Ele também é usado após análises filogenéticas, para analisar conjuntos de árvores filogenéticas, alinhamentos e avaliar a interferência de terminais que possuem dados faltantes (*missing data*). Acesse a página do programa e utilizem os tutoriais disponíveis, https://sco.h-its.org/exelixis/web/software/raxml/index.html.

O programa mais utilizado para fazer as análises de IB é o **MrBayes**. Ele também é um programa grátis e disponível para *download* para os sistemas operacionais Windows, MAC e LINUX. Como já mencionado o programa MrBayes além de fazer as análises de IB também pode ser utilizado para a escolha de modelos filogenéticos e evolutivos. Outro programa também utilizado para inferir as relações entre as espécies é o BEAST (http://www.beast2.org/).

3.8 Bibliografia recomendada

ARENAS, Miguel. Trends in substitution models of molecular evolution. Frontiers in genetics, v. 6, 2015.

LARTILLOT, Nicolas; PHILIPPE, Hervé. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Molecular biology and evolution, v. 21, n. 6, p. 1095-1109, 2004.

BEVAN, Rachel B.; BRYANT, David; LANG, B. Franz. Accounting for gene rate heterogeneity in phylogenetic inference. Systematic biology, v. 56, n. 2, p. 194-205, 2007.

BUCKLEY, Thomas R. Model misspecification and probabilistic tests of topology: evidence from empirical data sets. Systematic Biology, v. 51, n. 3, p. 509-523, 2002.

CASTRESANA, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular biology and evolution, v. 17, n. 4, p. 540-552, 2000.

CHOUDHURI, Supratim. Phylogenetic Analysis** The opinions expressed in this chapter are the author's own and they do not necessarily reflect the opinions of the FDA, the DHHS, or the Federal Government – Chapter 9.

DARRIBA, Diego et al. jModelTest 2: more models, new heuristics and parallel computing. Nature methods, v. 9, n. 8, p. 772-772, 2012.

EDGAR, Robert C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research, v. 32, n. 5, p. 1792-1797, 2004.

FELENSTEIN, Joseph. Inferring phylogenies. Sunderland, MA: Sinauer associates, 2004.

FENG, Da-Fei; DOOLITTLE, Russell F. Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. Journal of molecular evolution, v. 25, n. 4, p. 351-360, 1987.

HOGEWEG, Paulien; HESPER, Ben. The alignment of sets of sequences and the construction of phyletic trees: an integrated method. Journal of molecular evolution, v. 20, n. 2, p. 175-186, 1984.

KAINER, David; LANFEAR, Robert. The effects of partitioning on phylogenetic inference. Molecular biology and evolution, v. 32, n. 6, p. 1611-1627, 2015.

KATOH, Kazutaka et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research, v. 30, n. 14, p. 3059-3066, 2002.

KROGH, Anders et al. Hidden Markov models in computational biology: Applications to protein modeling. Journal of molecular biology, v. 235, n. 5, p. 1501-1531, 1994.

KUHNER, Mary K.; FELSENSTEIN, Joseph. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Molecular biology and evolution, v. 11, n. 3, p. 459-468, 1994.

LANFEAR, Robert et al. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Molecular biology and evolution, v. 29, n. 6, p. 1695-1701, 2012.

LARTILLOT, Nicolas; PHILIPPE, Hervé. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Molecular biology and evolution, v. 21, n. 6, p. 1095-1109, 2004.

LE, Si Quang; LARTILLOT, Nicolas; GASCUEL, Olivier. Phylogenetic mixture models for proteins. Philosophical Transactions of the Royal Society of London B: Biological Sciences, v. 363, n. 1512, p. 3965-3976, 2008.

LIO, Pietro; GOLDMAN, Nick. Models of molecular evolution and phylogeny. Genome research, v. 8, n. 12, p. 1233-1244, 1998.

LISCHER, Heidi EL; EXCOFFIER, Laurent. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. Bioinformatics, v. 28, n. 2, p. 298-299, 2011.

MADDISON, David R.; SWOFFORD, David L.; MADDISON, Wayne P. NEXUS: an extensible file format for systematic information. Systematic biology, v. 46, n. 4, p. 590-621, 1997.

MAYROSE, Itay; FRIEDMAN, Nir; PUPKO, Tal. A Gamma mixture model better accounts for among site rate heterogeneity. Bioinformatics, v. 21, n. suppl_2, p. ii151-ii158, 2005.

MILLER, Kenneth G.; KOMINZ, Michelle A.; BROWNING, James V.; WRIGHT, James D.; MOUNTAIN, Gregory S.; KATZ, Miriam E.; SUGARMAN, Peter J.; CRAMER, Benjamin S.; CHRISTIE-BLICK, Nicholas; PEKAR, Stephen F. The Phanerozoic record of global sea-level change. Science, v. 310, n. 5752, p. 1293-1298, 2005.

MOUNT, David W. Using iterative methods for global multiple sequence alignment. Cold Spring Harbor Protocols, v. 2009, n. 7, p. pdb. top44, 2009.

NEEDLEMAN, Saul B.; WUNSCH, Christian D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology, v. 48, n. 3, p. 443-453, 1970.

NGUYEN, Lam-Tung; SCHMIDT, Heiko A.; VON HAESELER, Arndt; MINH, Bui Quang. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular biology and evolution, v. 32, n. 1, p. 268-274, 2014.

NOTREDAME, Cédric; HIGGINS, Desmond G. SAGA: sequence alignment by genetic algorithm. Nucleic acids research, v. 24, n. 8, p. 1515-1524, 1996.

PAGEL, Mark; MEADE, Andrew. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Systematic biology, v. 53, n. 4, p. 571-581, 2004.

POSADA, David. jModelTest: phylogenetic model averaging. Molecular biology and evolution, v. 25, n. 7, p. 1253-1256, 2008.

SALEMI, Marco; LEMEY, Philippe; VANDAMME, Anne-Mieke (Ed.). The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. Cambridge University Press, 2009.

SIEVERS, Fabian; WILM, Andreas; DINEEN, David; GIBSON, Toby J.; KARPLUS, Kevin; LI, Weizhong; LOPEZ, Rodrigo; McWILLIAN, Hamish; REMMERT, Michael; SÖDING, Johannes; THOMPSON, Julie D.; HIGGINS, Desmond G. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology, v. 7, n. 1, p. 539, 2011.

SMITH, Temple F.; WATERMAN, Michael S. Identification of common molecular subsequences. Journal of molecular biology, v. 147, n. 1, p. 195-197, 1981. WAGENMAKERS, Eric-Jan; FARRELL, Simon. AIC model selection using Akaike weights. Psychonomic bulletin & review, v. 11, n. 1, p. 192-196, 2004.

YANG, Ziheng. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. Journal of Molecular evolution, v. 39, n. 3, p. 306-314, 1994.

Capítulo 4
Taxonomia, Classificação e Nomenclatura

Guilherme S. T. Garbino & Alessandro Rodrigues Lima

4.1 Taxonomia e Classificação

Definição

A Taxonomia é uma ciência teórico-prática que lida, basicamente, com a organização do conhecimento biológico sobre os seres vivos. Conforme proposto por George G. Simpson, é definida como "o estudo teórico da classificação, incluindo suas bases, princípios, procedimentos e regras".

Aliada à taxonomia, Simpson define **classificação zoológica** como "a ordenação de animais em grupos (ou ordenações) com base em suas relações, ou seja, associações por contiguidade, similaridade ou ambos". Classificação zoológica, seguindo essa definição, é a parte operacional da taxonomia.

Nesta seção apresentamos um breve histórico sobre a taxonomia zoológica, relevante para entendermos a filosofia por trás do pensamento taxonômico atual. Em seguida apresentamos quais fundamentos teóricos são comumente seguidos na prática taxonômica vigente.

Breve histórico

A taxonomia zoológica mais antiga que conhecemos no mundo ocidental é a presente no livro *Historia Animalium*, escrita pelo filósofo grego Aristóteles (384-322 a.C.) no século IV a.C. A ideia inicial de Aristóteles era classificar os animais de maneira dicotômica – um exemplo do próprio autor é o agrupamento de animais com sangue *versus* animais sem sangue (equivalente aos vertebrados e "invertebrados"). Porém, a variedade de formas e de categorias para classificar os animais era tão vasta que Aristóteles reconheceu ser impossível agrupar todos os animais em uma única classificação dicotômica. O golfinho, por exemplo, ele classificou no grupo de animais aquáticos (*versus* terrestres) e também no grupo de animais vivíparos (*versus* ovíparos), duas categorias não mutuamente exclusivas. Embora considerada a primeira taxonomia zoológica, o *Historia Animalium* não contém uma classificação única e sistematizada, mas apenas uma série de divisões dicotômicas dos seres estudados por Aristóteles.

Além de uma tentativa de classificação dos animais, Aristóteles introduz quatro conceitos-chave em seu livro: **gênero** (γένος - *genos*), **espécie** (εἶδος - *eidos*), ***differentia*** (διαφορά - *diaphora*) e **essência** (*essentia*). Diferente do que usamos hoje, um **gênero** poderia referir-se a qualquer agrupamento, e uma **espécie** a membros desse agrupamento. Então, por exemplo, a ordem Chiroptera poderia ser um gênero, e as famílias Phyllostomidae e Vespertilionidae espécies dentro desse gênero, segundo a definição de Aristóteles.

O termo ***differentia*** (ou diferença, em português) ainda é usado na taxonomia, e refere-se às características que diferenciam membros de uma classe em relação à outra. Por exemplo, no Brasil existem duas espécies de porcos-do-mato, pertencentes à família Tayassuidae, que são caracterizadas por possuírem incisivos superior triangulares em secção transversal e com as pontas voltadas para baixo. As duas espécies são diferenciadas por uma possuir um colar branco na região do pescoço e a outra possuir uma mancha branca no queixo.

O termo **essência** é atribuído a Platão, e podemos defini-lo como o(s) atributo(s) necessário(s) para que uma classe de indivíduos seja o que ela é. Como um exemplo, para ser um cachorro é necessário ter quatro patas, latir, possuir um rinário, ter quatro pré-molares superiores e quatro inferiores e possuir pelos. Dos quatro conceitos utilizados por Aristóteles, gênero, espécie e *differentia* são importantes na taxonomia atual, enquanto que essência (como veremos a seguir) foi deixado de lado por ser incompatível com a biologia evolutiva.

O sueco Carl Linnaeus (Carlos Lineu em português) (1707-1778), também conhecido pelo seu nome de nobreza Carl von Linné, foi responsável por disseminar o sistema nomenclatural binomial, que utilizamos até hoje. Embora

não tenha sido o primeiro a utilizar a nomenclatura binomial –Garpard e Johann Bauhin, por exemplo, já utilizavam binômios 200 anos antes –Lineu foi o primeiro a utilizá-la de maneira consistente.

Um dos grandes méritos de Lineu, e provavelmente o principal motivo pelo qual seguimos seu sistema binomial até hoje, foi sua imensa produtividade em escrever livros e descrever espécies. Ao todo, o naturalista publicou mais de 70 livros e descreveu mais de 4.400 espécies de animais e 7.770 de plantas. O livro mais conhecido e importante de Lineu é o *Systema Naturae*, o qual teve a primeira edição publicada em 1735 e a 12ª (a última feita por Lineu) entre os anos de 1766 e 1768. Como veremos na seção de nomenclatura, a décima edição (1758) do *Systema Naturae* é a mais relevante para taxonomia e nomenclatura zoológica, pois essa edição marca o ponto de início da nomenclatura zoológica.

Embora tenhamos herdado o sistema nomenclatural de Lineu, os métodos de classificação dos seres vivos utilizados pelo autor sueco eram consideravelmente diferentes dos que utilizamos. No *Systema Naturae* e em outros livros, como o *Species Plantarum*, a grande ambição de Lineu, além de descrever todos os seres vivos do planeta, era desvendar a lógica divina por trás das espécies de plantas e animais. Para o naturalista, o agrupamento de espécies em gêneros, de ordens em classes, era parte do plano divino. Como ele mesmo gostava de dizer: *"Deus creavit, Linnaeus disposuit"* ("Deus criou, Lineu classificou").

Até o início do século XIX, a vasta maioria das classificações produzidas consideravam que as espécies eram entidades eternas e imutáveis. Decorrente desse conceito, os estudiosos acreditavam que cada espécie era definida por sua essência, uma herança do conceito de **tipo** formulado na Grécia Antiga por Platão. Segundo os autores essencialistas (alcunha forjada em tempos recentes e usada de maneira pejorativa), as variações encontradas seriam apenas imperfeições desse modelo ideal etéreo.

De maneira geral, as variações individuais que os essencialistas assumiram ser imperfeições de uma forma **ideal**, foram vistas, de maneira independente, por Charles Darwin (1809-1882) e Alfred Wallace (1823-1913), como evidência de que as espécies não constituíam entidades fixas, mas que mudavam ao longo do tempo. Portanto, num contexto evolutivo, a variação intraespecífica passa a ser vista como condição fundamental para a definição de uma espécie.

As ideias de Darwin e Wallace tiveram dois grandes impactos na taxonomia: o primeiro foi o entendimento de que as semelhanças encontradas entre as espécies eram devido à ancestralidade comum; o segundo grande impacto foi a compreensão de que variações intraespecíficas são essenciais para que ocorra evolução.

Infelizmente, a taxonomia tipológica perdurou por quase um século após a publicação de **A Origem das Espécies**. Entretanto, nas décadas de 1940 e 1950 surge o movimento conhecido como **Síntese Moderna**, cujos maiores expoentes foram Ernst Mayr, Theodosius Dobzhansky e George Simpson. A síntese moderna consolidou o pensamento populacional em biologia, além de unir a genética Mendeliana à teoria da seleção natural. Somente após esse período a ideia de que espécies são entidades dinâmicas e variáveis passa a ser amplamente disseminada em taxonomia.

A última grande influência teórica na taxonomia veio na década de 1960, principalmente através dos trabalhos de Willi Hennig (1913-1976). O entomólogo alemão defendia que grupamentos naturais de seres vivos deveriam ser definidos apenas com base em **características** derivadas (ou **apomórficas**). Esses grupos, denominados **monofiléticos**, isto é, que possuem um ancestral em comum **exclusivo**, seriam os únicos táxons admissíveis em classificações biológicas.

4.2 A prática taxonômica

Como mencionamos, durante a maior parte da história da taxonomia zoológica as espécies eram definidas com base em indivíduos (espécimes), o que só foi alterado a partir das proposições de Darwin e Wallace sobre evolução dos sistemas e do caráter populacional das espécies, que devem ser definidas por conjuntos de indivíduos. Se considerarmos, portanto, que na grande maioria das vezes as populações não podem ser observadas diretamente, concluímos que a taxonomia é uma ciência baseada em inferência. Isso significa que em taxonomia efetuamos generalizações sobre a(s) população(ões) estudadas com base na amostra que temos disponível.

A espécie é uma entidade definida com base em inferências sobre observações e, de maneira análoga às populações, é algo que não pode ser observado diretamente na natureza. Estritamente falando, e considerando que todos os seres vivos fazem parte de um contínuo evolutivo, a espécie, como a tratamos na prática, é um conceito e sua definição é, e sempre será, arbitrária. Isso não necessariamente implica que espécies não existem de fato, mas apenas quer dizer que nossos conceitos e critérios operacionais sobre o que seria uma espécie não necessariamente coincidem com o táxon real na natureza.

Embora a espécie seja a unidade fundamental em estudos evolutivos e, consequentemente, em taxonomia, sua definição ainda representa um dos temas mais discutidos em biologia. Atualmente existem mais de 25 conceitos de espécie, mas aqui iremos focar em três, que consideramos os mais comumente aplicados.

O **Conceito Biológico de Espécie**, com certeza o mais conhecido entre o público leigo, foi proposto por Ernst Mayr na década de 1940. Nele, uma espécie é "Um grupo de populações naturais potencialmente ou de fato intercruzantes que é reprodutivamente isolado de outros grupos".

De maneira mais ampla, o **Conceito Evolutivo de Espécie** proposto por George Simpson em 1951, e reelaborado em seu livro de 1961, diz que uma espécie é "uma linhagem evoluindo separadamente de outras e com seu próprio papel evolutivo e tendências".

Nos dois conceitos observamos a abordagem populacional, característica da síntese moderna. Um grande problema no conceito de Mayr é que ele automaticamente exclui espécies com reprodução assexuada, e na prática, também as espécies fósseis ou outras espécies que não coexistem temporalmente. O conceito de Simpson, por outro lado, falha em ser muito amplo, e por conter termos vagos como, por exemplo, o "papel evolutivo" de uma linhagem, fornecendo um conceito, mas não um critério para delimitar as espécies.

O conceito que aparentemente está sendo mais utilizado atualmente é o **Conceito Filogenético de Espécie** de Joel Cracraft (1983). Nele, uma espécie é "o menor agrupamento de organismos individuais nos quais existe um padrão parental de ancestralidade e descendência e que é diagnosticavelmente distinto de outros agrupamentos por uma combinação única de estados de caráter fixados". Outros autores propuseram variações nesse conceito, mas a ideia central é identificar grupos monofiléticos que sejam diagnosticáveis. A aplicação desse conceito e suas variantes tem sido útil no avanço da taxonomia zoológica, pois os dois requerimentos desse conceito, isto é, que uma espécie seja monofilética e que ela seja diagnosticável, são mais facilmente testáveis do que os requerimentos dos outros conceitos (como a presença ou não de isolamento reprodutivo do conceito biológico).

Um avanço importante a essa discussão foi trazido por Kevin de Queiroz, que argumenta que a maioria dos **conceitos** na verdade expõem **critérios** para definirmos espécies (diagnosticabilidade, papel evolutivo, isolamento reprodutivo). Embora os critérios possam variar entre si, existe um amplo consenso que, em teoria, espécies são linhagens de metapopulações evoluindo separadamente.

O biólogo e filósofo da ciência Massimo Pigliucci identifica que todos os conceitos de espécie buscam por uma **essência** na definição de espécie, quando é sabido que, devido à grande heterogeneidade dos seres vivos, é impossível "agradar a gregos e troianos" com um conceito baseado em um único elemento-chave. Para mitigar esse problema, Pigliucci adota a ideia de **semelhança de família** do filósofo Ludwig Wittgenstein para o problema do conceito de espécies. Segundo essa ideia, espécie é um "conceito de grupo, cuja sustentação é encontrada em uma série de características como relações filogenéticas, similaridade genética, compatibilidade reprodutiva e características ecológicas". Isso significa que não existe uma característica *sine qua non* para definirmos espécies, mas que mesmo assim podemos defini-la, de maneira objetiva, evocando um conjunto de qualidades comumente identificadas e consensualmente consideradas essenciais à definição de espécies, tais como monofiletismo, isolamento reprodutivo, nicho ecológico, entre outras.

Os táxons superiores

A delimitação de táxons superiores, isto é, as categorias mais inclusivas que espécie (gênero, família e acima), é ainda mais arbitrária que a delimitação de espécies. Uma importante diferença entre a espécie e as categorias superiores é que na primeira a sua coesão é mantida por processos evolutivos ocorrendo em **tempo real**, enquanto nas demais os processos evolutivos ocorreram no **passado**. O único pré-requisito para que uma categoria de nível superior seja considerada como válida é que ela represente um grupo **monofilético**.

Hennig propôs utilizar o **tempo de divergência** para definir os táxons superiores. Nessa proposta, um clado que surgiu entre seis e quatro milhões de anos seria classificado como gênero, por exemplo, e um que surgiu entre 23 e 22 milhões de anos seria uma família. Dessa maneira, segundo Hennig, seria possível comparar essas categorias entre táxons filogeneticamente distantes, como Myrtaceae, família de plantas, e Scarabeidae, família de besouros. Na prática, tal classificação "padronizada" ainda seria arbitrária, além de implicar mudanças radicais e desnecessárias em táxons já estabelecidos. Os gêneros *Drosophila* (moscas) e *Eucalyptus* (árvore), por exemplo, possuem mais de 50 milhões de anos, enquanto nosso gênero, *Homo*, possui entre dois e três milhões de anos.

4.3 Nomenclatura

Definição

Seguindo, como na seção anterior, a definição de George Simpson para **nomenclatura zoológica**, podemos defini-la como "a aplicação de nomes distintos para cada um dos grupos reconhecidos em uma dada classificação zoológica". O objetivo principal da nomenclatura zoológica é que cada entidade reconhecida pelos taxonomistas seja reconhecida apenas por um nome, e que esse nome seja exclusivo desse táxon. Em outras palavras, tenta-

se eliminar **sinônimos** (nomes diferentes para a mesma entidade) e **homônimos** (nomes iguais para entidades diferentes).

Os nomes científicos

O nome de uma espécie é composto por duas partes, o nome **genérico** e o **nome específico**, também conhecido como **epíteto específico**. No escorpião-amarelo *Tityus serrulatus*, por exemplo, o nome genérico é *Tityus*, o epíteto específico é *serrulatus* e o nome da espécie é *Tityus serrulatus*. A esse nome geralmente atribuímos a alcunha de **nome científico**, o que na verdade é uma designação bastante vaga, já que vários outros nomes usados em ciências também podem ser considerados nomes científicos (por exemplo: aldeído, polimerase, fototropismo). Uma definição mais precisa seria "nomes científicos latinizados de espécie biológicas". Por conveniência, no entanto, continuaremos a referir a tais nomes como **nomes científicos**.

Lineu foi o responsável por aperfeiçoar e promover o sistema binomial de nomenclatura que usamos hoje. A aceitação e uso geral desse sistema só foi possível porque ele estabeleceu uma base sólida ao aplicar a binômios de forma consistente nos livros *Philosophia Botanica* (1751), *Species Plantarum* (1753) e na décima edição do *Systema Naturae* (1758). Uma vantagem do sistema lineano, diferente por exemplo do sistema utilizado pelo contemporâneo Georges-Louis Leclerc, o Conde de Buffon, era permitir expansões, ou seja, que as novas formas descobertas fossem facilmente alocadas na sua grande classificação da vida.

De maneira mais ampla, o principal trunfo da utilização de nomes científicos ao invés de nomes populares é que os primeiros permitem uma comunicação não-ambígua entre pesquisadores. A onça-parda, uma espécie de felino que ocorre desde o Canadá até a Argentina, é conhecida por mais de 40 nomes populares, em línguas que vão desde o quéchua, taino e tupi-guarani até o espanhol, português e inglês. No entanto, qualquer pesquisador, seja ele anatomista, ecólogo, fisiologista, geneticista ou sistemata, reconhece a entidade taxonômica denotada pelo nome *Puma concolor*. Os nomes científicos ainda servem como um sistema de recuperação de dados. Ao buscarmos por um organismo-modelo, como *Drosophila melanogaster*, numa base de dados, encontraremos milhares de artigos científicos e outros tipos de trabalho que lidam com a anatomia, ecologia, genética, entre outros aspectos da biologia dessa espécie de mosca.

Nomes científicos são formados por palavras em latim ou latinizadas. O ratinho silvestre *Sooretamys angouya*, por exemplo, tem seu nome específico derivado do substantivo tupi-guarani *angudjá*, que significa rato. Nomes também podem ser formados a partir de combinações arbitrárias de letras, como o gênero de caracóis *Aaadonta*. Geralmente o epíteto específico é um adjetivo e o gênero é um substantivo. No entanto, também existem epítetos formados a partir de genitivos (por exemplo: *Gymnodactylus darwinii* e *Leopardus emiliae*), que são cunhados geralmente em homenagem a pessoas ou lugares.

É aconselhável escrever o autor e data (separados por vírgula) da publicação da espécie ao lado do binômio, ao menos na primeira vez que o nome aparece no trabalho. Portanto, quando encontramos a grafia *Rhinella marina* (Linnaeus, 1758), significa que essa espécie foi descrita por Linnaeus no ano de 1758, com uma combinação diferente da proposta atualmente (por isso o autor e data estão entre parênteses). No caso, a espécie foi originalmente descrita como *Bufo marinus*.

Embora o pai da taxonomia tenha difundido o sistema de nomes científicos de duas partes, Lineu não considerava o binômio como sendo o nome verdadeiro da espécie. O **nome completo** da onça-pintada, segundo ele, seria uma frase descritiva ao invés de um binômio: *Felis cauda elongata, corpore flavescente maculis nigris rotundato angulatis medio flavis*. Traduzindo, o nome significaria "gato com cauda longa, corpo amarelo com manchas negras e o meio amarelo."

Outro fato curioso é que nenhum trabalho de Lineu apresenta de forma clara os binômios. Voltando para o exemplo da onça-pintada, no *Systema Naturae* (1758) a espécie é apresentada da seguinte maneira dentro do gênero *Felis*:

> *Onca. 4. F. cauda elongata, corpore flavescente maculis nigris rotundato angulatis medio flavis*

O número 4 indica que essa é a quarta espécie de *Felis* descrita no *Systema Naturae*. "F." é uma abreviação de *Felis* e "Onca" subentende-se que seria o epíteto específico da espécie. Notem que em nenhum momento do livro existe a combinação *"Felis Onca"* claramente explicitada.

Portanto, para Lineu, o epíteto específico serviria como um *proxy* para se referir de forma prática à espécie de *Felis* que possui cauda longa, com corpo amarelo, entre outras características. Também seria uma forma mais fácil de memorizar a espécie, já que o estudante (taxonomista ou não) não precisaria decorar toda a frase de seu nome completo.

Com o grande aumento no entendimento sobre a distribuição e variação geográfica das espécies de animais, alguns zoólogos no século XIX começaram a utilizar **trinômios** ou **subespécies**, para se referir a variações, geralmente restritas geograficamente, encontradas dentro do que consideravam a mesma espécie. Para designar esses táxons, esses autores utilizavam nomes de três partes, como por exemplo *Boa constrictor amarali*. Subespécies estão caindo em desuso na taxonomia da maioria dos grupos zoológicos atuais, principalmente porque, segundo o

conceito filogenético de espécie, se um táxon é diagnosticável, então ele deve ser considerado uma espécie plena. No entanto, em alguns casos onde a variação geográfica é considerável e não existe uma clara separação filogenética ou fenotípica entre populações, embora existam formas geograficamente restritas, o trinômio ainda é passível de ser usado.

Diferentemente da subespécie, o subgênero não precisa aparecer obrigatoriamente no nome da espécie/subespécie e nem é considerado parte do nome. Dessa forma, *Saguinus (Leontocebus) weddelli melanoleucus* não é um quadrinômio, e o nome do mesmo táxon pode ser escrito da forma *Saguinus weddelli melanoleucus*. Subgêneros indicam agrupamentos monofiléticos dentro de gêneros, e são particularmente úteis quando a divisão de um gênero em vários pode resultar em mudanças nomenclaturais desnecessárias.

Tipos

Herdamos do essencialismo de Platão o nome **tipo**. Embora originalmente o tipo de uma espécie denotasse sua essência, hoje essa palavra é usada em um contexto totalmente diferente em taxonomia e nomenclatura.

O **espécime-tipo** de uma espécie é, simplesmente, o espécime físico no qual o nome dado é ancorado (Figura 4.1). Em outras palavras, o tipo é a ligação entre o conceito (o nome da espécie) e a realidade (a espécie na natureza). Como se pode imaginar, o espécime-tipo possui função extremamente importante na resolução de problemas taxonômicos. Quando houver dúvidas, por exemplo, se dois nomes científicos se referem à mesma entidade taxonômica, a resposta começa com a análise do espécime-tipo de cada um dos nomes em questão. Se o pesquisador entender que ambos espécimes se encaixam no conceito existente sobre a entidade estudada, ele pode propor que ambos pertencem à mesma espécie. Então os nomes são considerados sinônimos, e o nome mais antigo tem prioridade sobre o mais recente.

É obrigatório, ao descrever uma nova espécie, designar um espécime-tipo (ou **holótipo**) para ancorar o nome novo. Se o(s) autor(es) desejarem, também podem ser designados **parátipos**. Não necessariamente todos os espécimes utilizados na descrição, além do holótipo, são parátipos. No entanto, se na descrição original da espécie o holótipo não foi claramente identificado, todos os espécimes utilizados na descrição são denominados **síntipos**. Cabe a um autor subsequente escolher, dentre os síntipos, qual seria o espécime-tipo, que nesse caso passa a ser chamado **lectótipo**. Os outros síntipos passam a ser **paralectótipos**.

Nos casos em que o holótipo ou o lectótipo se perderam (como aconteceu em vários museus europeus bombardeados durante a 2ª guerra mundial), e **é necessário saber a identidade de um táxon**, é possível designar um **neótipo**. Como o próprio Código de Nomenclatura Zoológica sugere, a designação de neótipos não deve ser feita apenas por motivos curatoriais, mas somente se realmente for justificável.

Figura 4.1 Espécime-tipo (lectótipo) de *Chiroderma villosum*, Peters, 1860, depositado no *Museum für Naturkunde* em Berlim, sob o número ZMB 408. Notar a etiqueta de coloração vermelha, exclusiva para espécimes-tipo. Foto: Guilherme S. T. Garbino

Além dos espécimes-tipo de táxons do grupo da espécie, os táxons supraespecíficos do grupo do gênero e da família também possuem tipos. De maneira distinta da espécie, os tipos de gênero e família não consistem em espécimes físicos, mas sim em nomes. Um gênero (e táxons do grupo do gênero) possui uma espécie-tipo e uma família (e táxons do grupo da família) possui um gênero tipo.

A espécie-tipo do gênero de mutuns *Crax* é *Crax rubra*. Isso quer dizer que, das sete espécies hoje reconhecidas no gênero, o nome genérico *Crax* está permanentemente ancorado à espécie *Crax rubra*. No grupo da família, o nome do gênero-tipo serve para formar a base do nome da família. Então, a família Deltatheriidae possui como gênero-tipo *Deltatheridium*.

Assim como acontece na espécie, ao descrevermos um novo gênero ou família (ou tribo, subfamília, ou outros), é obrigatório designar uma espécie-tipo e gênero-tipo, respectivamente. Nos casos de trabalhos antigos quando essa obrigação não existia e tipos não eram designados, cabe a um autor subsequente fazer a designação dos tipos.

Código de nomenclatura

Atualmente existem seis principais códigos de nomenclatura biológica: o Código Internacional de Nomenclatura de algas, fungos e plantas, o Código Internacional de Nomenclatura de Bactérias, o Código Internacional de Nomenclatura de Plantas Cultivadas, o Código Internacional de Nomenclatura Fitossociológica, o Código Internacional de Nomenclatura Zoológica e o Comitê Internacional de Taxonomia de Vírus.

A função desses códigos é providenciar um conjunto de regras que governam a nomenclatura formal dos grupos de organismos compreendidos por eles. Além de possuir um conjunto de regras dizendo o que é e o que não é permitido na nomenclatura de cada um dos grupos citados acima, cada código possui um comitê internacional que pode julgar e decidir arbitrariamente casos ambíguos ou de difícil resolução. Os códigos de nomenclatura foram criados para evitar o caos nomenclatural que estava se instaurando no final século 19. Alguns autores, por exemplo, começaram a deliberadamente substituir todos os nomes "bárbaros" (nomes de espécimes dados com base em palavras indígenas, como *Mico*, *Pecari* e *Puma*) por nomes com raízes gregas e latinas. Outros, ignoravam nomes mais antigos e criavam novos que eram mais do seu agrado.

Existem algumas diferenças entre cada código de nomenclatura. Por exemplo, o código de algas, fungos e plantas não permite tautônimos, enquanto o código de nomenclatura zoológica não faz restrição ao seu uso, como acontece em *Gorilla gorilla* e *Troglodytes troglodytes*. Até recentemente, descrições de plantas tinham que ser apresentadas em latim, enquanto na nomenclatura zoológica isso nunca foi obrigatório.

O Código Internacional de Nomenclatura Zoológica (CINZ) foi criado em 1961, mas teve como precursor um conjunto de Regras Internacionais de Nomenclatura Zoológica, publicado em 1905. O CINZ está em sua quarta edição, do ano de 1999 e seus idiomas oficiais são inglês e francês. Uma versão em português da segunda edição do código pode ser encontrada no livro Fundamentos práticos de taxonomia zoológica editado pelo prof. Nelson Papavero.

Um dos preceitos fundamentais do CINZ, é o **princípio da prioridade**. Em linhas gerais, esse princípio diz que se existirem dois nomes diferentes para um mesmo táxon, o nome mais antigo possui prioridade. Dessa maneira, o nome mais antigo deve ser reconhecido como **sinônimo sênior**, ficando o mais recente como **sinônimo júnior**. Um caso recente onde essa regra foi corretamente aplicada foi o da redescoberta do macaco-prego-galego. Em 2006, um grupo de cientistas deu um nome novo ao que acreditaram ser uma espécie não descrita de macaco-prego, chamando-a de *Cebus queirozi*. Outros pesquisadores sugeriram que esse táxon de macaco prego já tinha sido descrito em 1774 pelo naturalista alemão Johann Schreber, que nomeou a espécie de *Cebus flavius*. Aplicando o princípio da prioridade, portanto, o nome *Cebus queirozi* Mendes-Pontes & Malta, 2006 é um sinônimo júnior de *Cebus flavius* Schreber, 1774. Uma característica importante do CINZ é que suas regras e princípios se aplicam somente a táxons do grupo de família, gênero e espécie. Portanto, é permitido chamar a ordem que inclui as baleias, hipopótamos, vacas, camelos e porcos de Cetartiodactyla Montgelard, Catzeflis & Douzery, 1997, embora o nome mais antigo para esse grupo seja Cetacea Brisson, 1762.

Para facilitar o trabalho de futuras revisões taxonômicas e evitar erros de redescrever um táxon quando já existe um nome disponível, como o que ocorreu com o macaco-prego-galego, é comum trabalhos de cunho taxonômico publicarem listas de sinônimos. Nelas, além dos sinônimos e homônimos, são listados também erros de grafia (por exrmplo: *Chiroderma trinitratum*: Linares & Moreno-Mosquera, 2010; grafia incorreta de *Chiroderma trinitatum* Goodwin, 1958) e combinações alternativas (por exemplo: o primeiro uso da combinação *Mico humeralifer* Rylands et al. 2000, ao invés de *Callithrix humeralifera* por de Vivo 1991).

É importante mencionar que o CINZ considera a décima edição do *Systema Naturae* como o ponto inicial da nomenclatura zoológica. Portanto, não podem existir binômios mais antigos que a data de publicação deste livro, que é 1758. A única publicação que tem prioridade sobre o *Systema Naturae* é o livro *Svenska Spindlar* (ou *Aranei Suecici*), de Carl Alexander Clerck (1757), mas os nomes presentes neste trabalho são tratados como sendo publicados em 1 de janeiro de 1758.

Validade dos nomes

Como vimos na seção sobre tipos, um dos requisitos para um nome científico ser válido é possuir um tipo. O código permite que o tipo seja um espécime coletado ou uma foto ou ilustração que **represente o espécime**. Deste modo, é possível descrever uma nova espécie apenas com base em uma foto, e designar como o holótipo o indivíduo **representado na foto** (note que a foto em si não é o tipo, mas sim o espécime retratado nela).

Atos nomenclaturais, como descrição de uma nova espécie, são considerados válidos pelo código apenas quando se enquadram nos critérios de publicação. Até recentemente, o código considerava publicado apenas trabalhos

impressos em papel, que possuíssem várias cópias iguais e fossem distribuídos amplamente. Em 2012, entretanto, foi permitida a publicação em meio digital, desde que o ato nomenclatural seja ser registrado na base de dados do ZooBank.

Outro requerimento para um nome ser válido é que ele seja acompanhado de uma descrição que indique quais caracteres diferenciam o táxon novo. O autor do nome novo também precisa dizer explicitamente que está propondo o nome como novo. Por esse motivo na frente dos nomes novos é comum encontrar escrito **sp. nov**, ***nomen novum***, **gen. nov.**, entre outros.

Se o nome não atende a esses critérios, como por exemplo, se ele aparece apenas em uma lista de espécies, sem qualquer descrição mais aprofundada, ele é considerado um *nomen nudum* não disponível para fins nomenclaturais. Qualquer autor subsequente pode utilizar o mesmo nome para um táxon novo e, caso a descrição atenda aos critérios do CINZ, o nome passa a ser disponível, mas com o autor e data da publicação mais recente.

4.4 Bibliografia recomendada

CAIN, Arthur J. Logic and memory in Linnaeus's system of taxonomy. In: Proceedings of the Linnean Society of London. Blackwell Publishing Ltd, 1958. p. 144-163.

GARBINO, Guilherme Siniciato Terra. Defining genera of New World monkeys: the need for a critical view in a necessarily arbitrary task. International Journal of Primatology, v. 36, n. 6, p. 1049-1064, 2015.

GROVES, Colin. The What, Why and How of Primate Taxonomy, International Journal of Primatology, v. 25, n. 5, p. 1105–1126, 2004.

HENNIG, Willi. Phylogenetic systematics. Annual review of entomology, v. 10, n. 1, p. 97-116, 1965.

ICZN (INTERNATIONAL COMMISSION ON ZOOLOGICAL NOMENCLATURE). International Code of Zoological Nomenclature, adopted by the International Union of Biological Sciences. 1999.

LINNAEUS, C. von. Systema naturae, 10th edn, vol. 1. Stockholm: L. Salvii, 1758.

MAYR, Ernst. Systematics and the origin of species, from the viewpoint of a zoologist. Harvard University Press, 1942.

PAPAVERO, Nelson. Fundamentos práticos de taxonomia zoológica. Unesp, 1994.

PATTON, James L.; CONROY, Christopher J. The conundrum of subspecies: morphological diversity among desert populations of the California vole (Microtus californicus, Cricetidae). Journal of Mammalogy, v. 98, n. 4, p. 1010-1026, 2017.

PIGLIUCCI, Massimo. Species as family resemblance concepts: The (dis-) solution of the species problem?. BioEssays, v. 25, n. 6, p. 596-602, 2003.

DE QUEIROZ, Kevin. Species concepts and species delimitation. Systematic biology, v. 56, n. 6, p. 879-886, 2007.

SIMPSON, George Gaylord. Types in modern taxonomy. American Journal of Science, v. 238, n. 6, p. 413-431, 1940.

SIMPSON, George Gaylord. The species concept. Evolution, v. 5, n. 4, p. 285-298, 1951.

SIMPSON, George Gaylord. Principles of animal taxonomy. 1961.

WINSTON, Judith E. Describing species: practical taxonomic procedure for biologists. Columbia University Press, 1999.

WRIGHT, John. The naming of the shrew: A curious history of Latin names. Bloomsbury Publishing, 2015.

Capítulo 5
Métodos Comparativos Filogenéticos

Rafaela V. Missagia & Daniel M. Casali

5.1 Introdução

Definição

São um conjunto de ferramentas analíticas que incorporam informações filogenéticas ao estudo dos padrões e processos evolutivos. Estes métodos têm como objetivo compreender estes padrões e processos a partir dos atributos (fenotípicos, moleculares, ecológicos, geográficos, entre outros) dos táxons. A maior parte desses estudos utiliza informações que podem ser observados em táxons atuais, como indicadores da sua história evolutiva, mas atributos de organismos fósseis também são frequentemente usados. A utilização dos Métodos comparativos filogenéticos (MCF) não deve ser confundida com a inferência filogenética. Esta tem como objetivo obter o relacionamento entre táxons a partir dos atributos dos organismos, já os MCF utilizam filogenias para estudar a evolução de atributos.

Breve histórico

Apesar de apenas recentemente as informações filogenéticas terem sido incorporadas às abordagens comparativas, a abordagem comparativa tem uma longa história na biologia, existindo desde o século XX, como uma forma de estabelecer interpretações sobre as relações de similaridade (e posteriormente, evolutivas) a partir da comparação entre táxons. Muitos pesquisadores contribuíram para este campo de estudo, mas alguns tiveram maior destaque e merecem ser mencionados.

Piérre Belon (1517-1564) foi um dos pioneiros da anatomia comparada moderna realizando estudos sobre a morfologia do esqueleto de pássaros e de seres humanos, diagnosticando padrões importantes de similaridade morfológica entre estes organismos. Edward Tyson (1651-1708) é muitas vezes considerado o **fundador da anatomia comparada**. Utilizando uma abordagem comparativa (mas não o método comparativo moderno), escreveu sobre a relação próxima entre chimpanzés e seres humanos. Jean-Baptiste Lamarck (1744-1829) foi um dos primeiros a considerar o ambiente e o tempo histórico como componentes da variação morfológica. Antes de Lamarck, a anatomia era primariamente descritiva, e não havia tanta preocupação na busca de mecanismos causais. Georges Cuvier (1769-1832), naturalista e zoólogo francês, avançou enormemente o campo da anatomia comparada, através de descrições e comparações entre organismos viventes e fósseis. O anatomista e paleontólogo inglês Richard Owen (1804-1892) cunhou o termo **homologia**, para se referir às estruturas comparáveis (a mesma, independente das variações de forma e função) entre dois ou mais organismos.

No início, o estudo comparativo era feito essencialmente por comparação anatômica entre espécies. A partir de Lamarck, os componentes ambiental e histórico passaram gradativamente a ser levados em consideração nas explicações das causas biológicas, culminando com a teoria evolutiva proposta independentemente por Charles Darwin (1809-1882) e Alfred Russel Wallace (1823-1913), dando início ao que se tornaria a biologia evolutiva. No início século XX, com o estabelecimento da síntese evolutiva moderna, ocorre a integração de diversas áreas da biologia à teoria evolutiva darwiniana, incluindo a sistemática. Alguns anos depois, na década de 1950, surge a escola sistemática filogenética, proposta por Willi Hennig (1913-1976), que em meio a um intenso debate entre divergentes escolas de pensamento, estabeleceu os princípios fundamentais para o processo de inferência das relações filogenéticas entre os organismos, sendo a base teórica de grande parte do método ainda hoje utilizado para obtenção das árvores filogenéticas e classificações biológicas.

Ao longo do século XX, programas de pesquisa em evolução emergiram baseados em duas linhas: o da genealogia e descendência comum, na filogenética; e o estudo dos fatores adaptativos relacionados ao ambiente, na biologia evolutiva, ecologia e etologia. O papel das filogenias em estudos ecológicos e adaptativos era bastante tímido. Mesmo com a revolução darwiniana e o fundamento teórico de que a diversidade biológica evoluiu através de uma

combinação de processos genealógicos e ambientais, os estudos filogenéticos e ecológicos seguiram sendo feitos de forma independente por muito tempo.

Até meados dos anos 70, os métodos analíticos da biologia comparativa se desenvolveram independente de aspectos filogenéticos e, em sua maioria, tentavam relacionar um determinado fenótipo a outro, ou a alguma variável ambiental, utilizando correlações como evidência. Dessa forma, os dados comparativos de atributos dos organismos eram analisados com os métodos estatísticos tradicionais, frequentemente assumindo a total independência das observações entre os táxons estudados, visando atender as premissas dos testes estatísticos. Pode-se destacar como exceção a ANOVA hierárquica (*nested analysis of variance*), que particiona a variação de um caráter contínuo entre espécies por diferentes níveis taxonômicos. O objetivo é identificar o nível taxonômico que abriga a maior parte da variação observada, que será então utilizado na análise. Esse método pode ser descrito como uma tentativa inicial de identificar a influência do parentesco na variância dos caracteres, podendo indicar até que ponto os dados podem ser tratados como filogeneticamente independentes.

Para testar se a evolução de um caráter era correlacionada a uma variável ambiental, ou a outro atributo, uma correlação significativa era entendida como uma evidência para a influência da variável X sobre a evolução da variável Y. O problema é que espécies são relacionadas entre si por meio de ancestrais comuns, formando uma filogenia hierarquicamente estruturada, e não podem ser consideradas como unidades amostrais independentes em testes estatísticos. Se espécies proximamente relacionadas apresentam fenótipos similares, isso pode ser um efeito de ancestralidade comum recente e não porque evoluíram independentemente este fenótipo. O problema da dependência filogenética ganhou maior popularidade após o artigo de Felsenstein (1985) propondo o método de contrastes independentes como uma forma de corrigir a não-independência das amostras decorrente da ancestralidade comum.

Com o crescimento das tecnologias e redução dos custos para sequenciamento molecular, surgimento de complexos modelos evolutivos, refinamentos das metodologias de análises probabilísticas e avanços nas capacidades computacionais, o número de filogenias disponíveis aumentou exponencialmente. Isso impulsionou o surgimento e desenvolvimento de diversas abordagens utilizando métodos comparativos filogenéticos.

Inicialmente, era possível analisar apenas a relação entre determinados caracteres utilizando uma hipótese filogenética, seja para a reconstrução de estados ancestrais dos caracteres, correlacionar variáveis (controlando o efeito da descendência comum para garantir a independência dos pontos) ou por associação entre o aparecimento de determinada característica no grupo e variáveis ambientais/eventos geológicos. Mais recentemente, métodos que permitem testar diferentes padrões evolutivos ao longo dos ramos da filogenia têm sido desenvolvidos. Isso se dá principalmente através de comparação e avaliação de ajuste entre diferentes modelos de evolução de caracteres e de modelos de diversificação, que constituem um enorme avanço no estudo da biologia evolutiva.

5.2 Aplicações

Os MCF são aplicados em estudos macroevolutivos, incluindo a inferência de estados ancestrais de caracteres, padrão de evolução de caracteres individuais, mudanças no tempo, e modo de evolução destes caracteres, padrões de correlação entre caracteres, teste de sinal filogenético, mudanças nas taxas de diversificação (especiação e extinção), influência de caracteres (ou outros atributos) nas taxas de diversificação, entre outros. Para esses estudos, é necessário uma hipótese filogenética, preferencialmente estimada de forma independente dos atributos utilizados nas inferências evolutivas dos MCF. Estas hipóteses podem ser uma simples topologia, contendo as relações de parentesco entre os táxons estudados (para casos mais simples como o uso de parcimônia), mas em sua maioria, requerem informações adicionais importantes como os comprimentos de ramos ou tempos de divergência entre os táxons. Cada método tem suas premissas, como aceitar somente uma árvore completamente resolvida, ou com comprimento de ramos, sendo necessário sempre considerar essas premissas ao aplicar determinado método à sua pergunta. Alguns métodos permitem a utilização de múltiplas árvores, considerando também a incerteza filogenética.

Ao se utilizar modelo, deve-se sempre testar a adequação destes aos dados. Para caracteres contínuos, o modelo mais utilizado, o de movimento Browniano, em sua versão mais simples descreve a evolução dos caracteres como um processo estocástico e assume que a variância de determinado atributo aumenta em função do tempo, sendo que esta variância pode assumir valores positivos ou zero, com uma taxa constante de evolução. Apesar de muito utilizado e ter diversas variantes (algumas que inclusive não assumem taxas constantes de evolução), o modelo Browniano pode ser insuficiente para descrever processos evolutivos mais complexos que originaram a variação morfológica. Desde então, alguns outros modelos foram propostos. Outros modelos mais complexos, tal como o Modelo de *Ornstein-Uhlenbeck* (OU), que incorpora um parâmetro (*alfa*) que representa forças seletivas que diminuem a variância dos atributos, se aproximando mais de processos de adaptação, apesar de permitir outras explicações não adaptativas também. O modelo OU também possui diversas variações, que podem ser testadas quanto ao ajuste aos dados. Já para caracteres discretos, necessita-se de uma matriz de transição que especifique as taxas de transição entre os estados do caráter.

5.2.1 Métodos para a inferência de estados ancestrais

É possível inferir os estados ancestrais de caráter para conjuntos de espécies a partir de informações dos terminais de uma filogenia e um modelo de evolução. O método de estimativa de estados ancestrais mais utilizado foi o da parcimônia, que, assim, como para a inferência de hipóteses filogenéticas, minimiza o número de mudanças ao longo da árvore, mas sem calcular a incerteza das estimativas.

Atualmente, a estimativa de estados ancestrais também tem sido feita por Máxima Verossimilhança ou métodos Bayesianos, onde os estados de caráter são estimados de acordo com a sua probabilidade, considerando um modelo evolutivo, uma hipótese filogenética e os comprimentos de ramo da filogenia. Nesse tipo de análise é possível calcular a incerteza das estimativas, que aumentam à medida que se avança dos terminais para a raiz da filogenia (Figura 5.1).

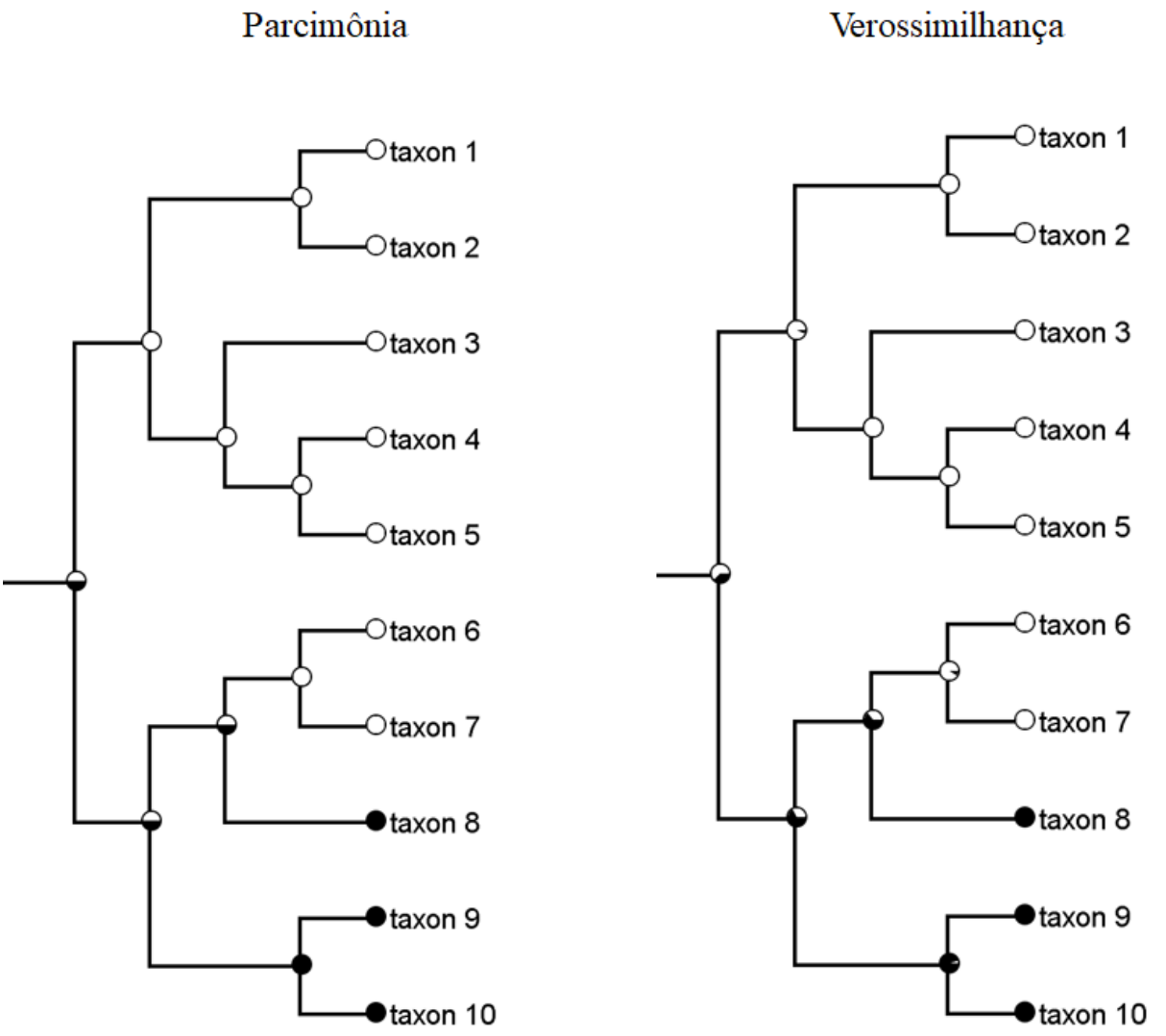


Figura 5.1 Reconstrução de estados ancestrais de um caráter discreto binário para dez táxons, utilizando parcimônia e máxima verossimilhança. Na reconstrução com parcimônia, ou os nós ancestrais são estimados como um dos estados (círculo branco ou preto nos nós) ou como uma ambiguidade ilustrada como 50% de probabilidade para cada estado (círculo metade branco, metade preto). Utilizando a verossimilhança, os círculos nos nós indicam o grau de suporte relativo para cada estado (dado o modelo), que pode variar de 0 a 100%. O suporte de cada estado é ilustrado por uma fatia do círculo nos nós na reconstrução por máxima verossimilhança (o mesmo se aplicaria para métodos Bayesianos).

5.2.2 Métodos para atributos individuais

Utilizando modelos de evolução de caracteres, tanto para atributos discretos como contínuos, é possível não apenas testar qual modelo melhor se ajusta a um caráter, mas também avaliar se as taxas da mudança de um estado são maiores ou menores que para a mudança reversa (exemplo: $A \rightarrow B > B \rightarrow A$), permitindo compreender padrões de mudanças das taxas e direcionalidade da evolução do caráter. Através destas análises também é possível testar padrões macroevolutivos relacionados ao modo que determinados atributos evoluem. Por exemplo, pode-se testar qual modelo se ajusta melhor aos dados, como um modelo gradual de evolução, com as mudanças igualmente distribuídas pelos ramos da topologia, ou um padrão pontuado, com poucas mudanças nos ramos, estando a maioria destas relacionadas aos eventos de cladogênese.

5.2.3 Métodos para correlação entre atributos

Métodos para caracteres discretos

A utilização de caracteres discretos nos MCF se dá principalmente através da correlação entre caracteres (Figura 5.2), com o objetivo de determinar se a evolução de um caráter está relacionada a evolução do outro. Assim como para dados quantitativos, apenas contar as incidências de correlação de dois ou mais caracteres entre os táxons não é suficiente, já que pelo efeito da ancestralidade comum, os estados de caráter podem não ser independentes. Ridley (1983) propôs um dos primeiros métodos, onde cada mudança de caráter é tratada como independente ao longo dos ramos. A correlação era então testada tratando um dos caráteres como dependente e o outro como independente, e através de uma tabela de contingência, se mudanças em um dos caracteres estariam relacionadas às mudanças no outro.

Mark Pagel (1994) foi o primeiro a propor um método específico para caracteres discretos, utilizando modelos probabilísticos para estimar as taxas de transição de um atributo em relação a outro. O Modelo de Pagel calcula a probabilidade de transição de um estado de caráter em dois cenários através de um modelo Markoviano: um onde a transição é independente da transição em outro atributo; e outro onde os atributos mudam de forma correlacionada, mas utilizando um modelo probabilístico, em vez de uma matriz de presença e ausência. O método foi estendido posteriormente para utilizar caracteres multiestado e para incorporar a incerteza quanto a topologia, em um arcabouço Bayesiano.

Métodos para caracteres contínuos

Felsenstein (1985) foi o primeiro a propor um método que incorpora a filogenia no estudo de caracteres contínuos, com os contrastes independentes. É necessário incorporar um modelo de evolução de caracteres, que vai estabelecer como determinado atributo evoluiu ao longo do tempo. O modelo mais utilizado nas análises com caracteres contínuos é o baseado no movimento Browniano (inclusive uma das premissas do método de contrastes independentes), mas outros modelos de evolução podem ser aplicados em outros testes, como no caso do método *Phylogenetic Generalized Least Squares* (PGLS), que permite modelos menos estritos que o movimento Browniano.

5.2.4 Métodos para inferência de taxas de diversificação

O número de espécies varia entre linhagens, regiões geográficas e períodos geológicos. Mudanças nas taxas de diversificação podem estar ligadas a eventos geológicos, tais como extinções em massa, que podem promover o aumento da diversificação dos táxons sobreviventes, oportunidades ecológicas relacionadas às radiações adaptativas, ou ao surgimento de inovações chave, entre outros.

A diversificação pode ser entendida como um balanço entre as taxas de especiação e extinção. As taxas de diversificação podem ser compreendidas de forma relativa ou absoluta, em relação a determinados intervalos temporais. Uma mudança nas taxas de diversificação pode ser identificada utilizando filogenias moleculares datadas, permitindo entender o período específico em que elas ocorreram.

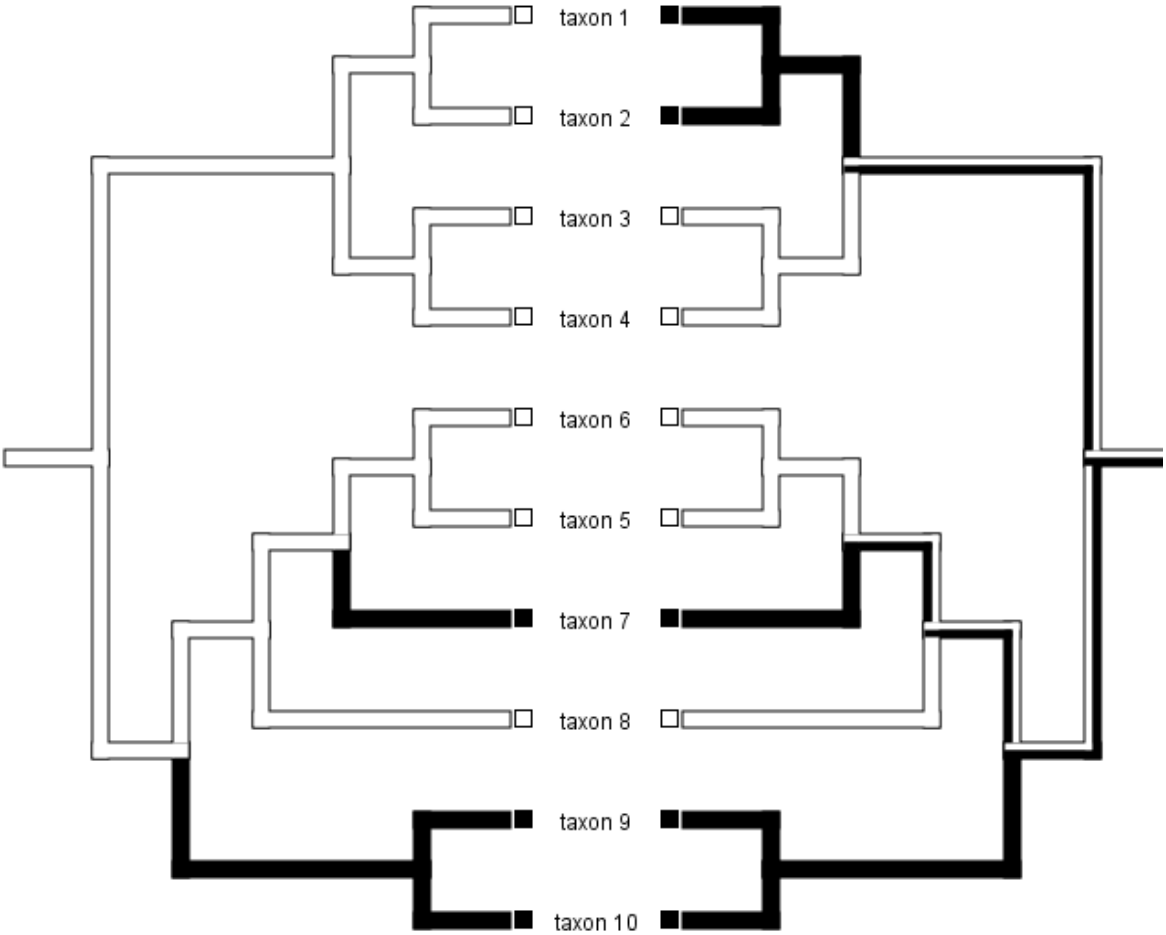


Figura 5.2 O padrão evolutivo de dois caracteres discretos independentes pode fornecer evidências de correlação entre eles. Na figura acima, cada árvore (esquerda e direita) mostra o padrão de evolução um caráter. Podemos observar que para 8 dos 10 táxons, os estados de caráter mudam de forma coincidente, sugerindo um padrão de correlação, que pode ser testado quanto ao seu suporte em relação a uma hipótese nula de não-correlação.

5.2.5 Métodos para inferência de taxas de diversificação dependentes de atributos

Há uma longa tradição na literatura paleontológica e macroevolutiva de questionar se atributos dos organismos poderiam afetar positiva ou negativamente as taxas de especiação e extinção, gerando a chamada "seleção de espécies". Algumas vezes, um determinado atributo pode promover maior estabilidade ou crescimento populacional, e consequentemente, suas taxas de especiação e extinção (Figura 5.3). Maddison e colaboradores (2007) propuseram um método (*BiSSE*) para estimar o efeito de transição de estados de um caráter binário nas taxas de extinção e especiação. Desde então, foram propostos novos métodos para lidar com a influência de caracteres com múltiplos estados (*MuSSE*; FitzJohn, 2012), caracteres contínuos (*QuaSSE*; FitzJohn 2010), caracteres geográficos (*GeoSSE*; Goldberg et al. 2011), entre outros. Todos esses métodos relacionam taxas de diversificação (especiação e extinção) às transições de caracteres.

5.3 Limitações e cuidados ao utilizar os métodos comparativos filogenéticos

Mesmo usando os modelos mais sofisticados e procedimentos para selecionar os melhores modelos, testes que correlacionam variáveis (caracteres entre si, diversificação x atributos) nos fornecem apenas padrões de correlação. Devemos lembrar que as interpretações de causalidade devem sempre ser feitas utilizando outros conhecimentos sobre os processos testados e do grupo de estudo, de forma a estabelecer interpretações mais fundamentadas.

Outra condição importante é ter uma amostragem ampla, contendo preferencialmente diversas réplicas do mesmo padrão estudado, a fim de que os testes de significância e suporte de hipóteses possam ser interpretados com maior segurança. Por exemplo, teríamos maior confiança se observássemos diversas vezes um determinado caráter correlacionado com altas taxas de extinção, do que apenas uma vez.

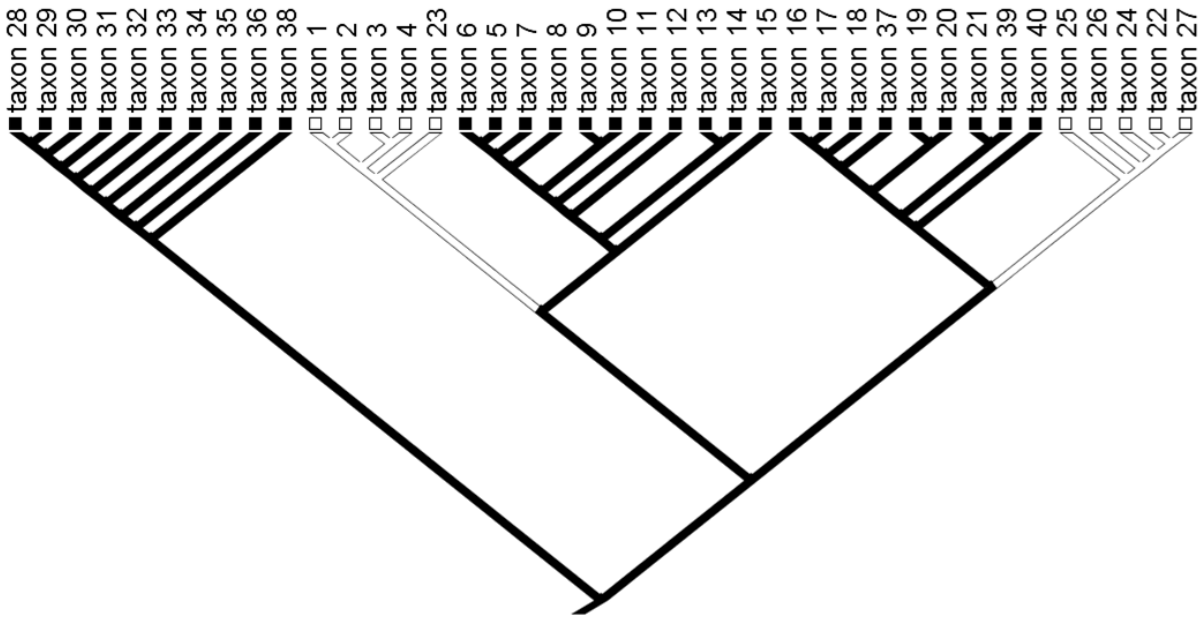


Figura 5.3 Linhagens com maior número de terminais podem ter maiores taxas de especiação ou menores taxas de extinção. É possível testar se modelos que atribuem mudanças nas taxas de diversificação em determinados clados estão associados à mudanças de estado de caracteres nestes clados. Na árvore acima, seria importante investigar se a presença do estado preto do caráter teria influenciado nos padrões de diversificação dos clados com mais espécies.

Ainda, é importante lembrar que assumimos que a árvore utilizada é correta ou a melhor estimativa que temos. Portanto, nossos resultados dependem de quão confiáveis são estas filogenias. Incorporar a incerteza filogenética é recomendável nestes casos.

5.4 Bibliografia recomendada

AUTUMN, Kellar; RYAN, Michael J.; WAKE, David B. Integrating historical and mechanistic biology enhances the study of adaptation. The Quarterly Review of Biology, v. 77, n. 4, p. 383-408, 2002.

BROOKS, Daniel R.; MCLENNAN, Deborah A. Phylogeny, ecology, and behavior: a research program in comparative biology. University of Chicago press, 1991.

CHEVERUD, James M.; DOW, Malcolm M.; LEUTENEGGER, Walter. The quantitative assessment of phylogenetic constraints in comparative analyses: sexual dimorphism in body weight among primates. Evolution, v. 39, n. 6, p. 1335-1351, 1985.

CLUTTON-BROCK, Timothy H.; HARVEY, Paul H. Primate ecology and social organization. Journal of Zoology (London), v. 183, p. 1-39, 1977.

CORNWELL, Will; NAKAGAWA, Shinichi. Phylogenetic comparative methods. Current Biology, v. 27, n. 9, p. R333-R336, 2017.

CROOK, John H. Sexual selection, dimorphism and social organization in the primates. In Sexual selection and the descent of man, 1871-1971. 231-281pp. Campbell, B. (ed.). Chicago: Aldine Atherton, 1972.

FELSENSTEIN, Joseph. Phylogenies and the comparative method. The American Naturalist, v. 125, n. 1, p. 1-15, 1985.

FITZJOHN, Richard G. Quantitative traits and diversification. Systematic biology, v. 59, n. 6, p. 619-633, 2010.

FITZJOHN, Richard G. Diversitree: comparative phylogenetic analyses of diversification in R. Methods in Ecology and Evolution, v. 3, n. 6, p. 1084-1092, 2012.

GARAMSZEGI, László Zsolt (Ed.). Modern phylogenetic comparative methods and their application in evolutionary biology: concepts and practice. Springer, 2014.

GOLDBERG, Emma E.; LANCASTER, Lesley T.; REE, Richard H. Phylogenetic inference of reciprocal effects between geographic range evolution and diversification. Systematic Biology, v. 60, n. 4, p. 451-465, 2011.

GOULD, Stephen Jay; LEWONTIN, Richard C. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Proceedings of the Royal Society of London B: Biological Sciences, v. 205, n. 1161, p. 581-598, 1979.

HANSEN, Thomas F.; PIENAAR, Jason; ORZACK, Steven Hecht. A comparative method for studying adaptation to a randomly evolving environment. Evolution, v. 62, n. 8, p. 1965-1977, 2008.

HARVEY, Paul H. et al. The comparative method in evolutionary biology. Oxford: Oxford university press, 1991.

HARVEY, Paul H.; MACE, Geordina M. Comparisons between taxa and adaptive trends: problems of methodology. 343-361pp. In: Current problems in sociobiology. King's College Sociobiology Group, ed. Cambridge University Press, Cambridge. 1982.

HARVEY, Paul H.; CLUTTON-BROCK, Timothy H. Life history variation in primates. Evolution, v. 39, n. 3, p. 559-581, 1985.

MADDISON, Wayne P. Confounding asymmetries in evolutionary diversification and character change. Evolution, v. 60, n. 8, p. 1743-1746, 2006.

MADDISON, Wayne P.; MIDFORD, Peter E.; OTTO, Sarah P. Estimating a binary character's effect on speciation and extinction. Systematic biology, v. 56, n. 5, p. 701-710, 2007.

O'MEARA, Brian C. Evolutionary inferences from phylogenies: a review of methods. Annual Review of Ecology, Evolution, and Systematics, v. 43, p. 267-285, 2012.

O'MEARA, Brian C.; BEAULIEU, Jeremy M. Past, future, and present of state-dependent models of diversification. American Journal of Botany, v. 103, n. 5, p. 792-795, 2016.

PAGEL, Mark. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proceedings of the Royal Society of London B: Biological Sciences, v. 255, n. 1342, p. 37-45, 1994.

PAGEL, Mark. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. Systematic biology, v. 48, n. 3, p. 612-622, 1999.

PAGEL, Mark. Inferring the historical patterns of biological evolution. Nature, v. 401, n. 6756, p. 877-884, 1999.

PENNELL, Matthew W.; HARMON, Luke J. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. Annals of the New York Academy of Sciences, v. 1289, n. 1, p. 90-105, 2013.

RONQUIST, Fredrik. Bayesian inference of character evolution. Trends in ecology & evolution, v. 19, n. 9, p. 475-481, 2004.

SWOFFORD, David L.; MADDISON, Wayne P. Reconstructing ancestral character states under Wagner parsimony. Mathematical Biosciences, v. 87, n. 2, p. 199-229, 1987.

Capítulo 6
Aplicabilidade da Sistemática Molecular

Daniela Nuñez

6.1 Introdução

A informação genética tem o potencial de auxiliar na conservação e na prática de manejo de populações silvestres e em cativeiro de duas principais formas. A primeira é que, baseado nela, é possível identificar espécies de forma precisa a partir de vários estágios do desenvolvimento (ovos, larvas, juvenis, adultos). Uma vez que as chaves taxonômicas são comumente elaboradas com base apenas em indivíduos adultos, os dados moleculares auxiliam na identificação dos demais estágios. A outra utilidade é o uso em análises de estrutura populacional, na reconstrução da história evolutiva e na definição de unidades de manejo (UMs), as quais são peças-chave na elaboração de estratégias de conservação.

O uso de técnicas da genética molecular é crucial para a estimativa da diversidade genética populacional. Desta forma, a gama de objetivos a serem atingidos exige o uso de vários marcadores, entre os quais se encontram disponíveis aqueles desenhados nos genomas mitocondrial (mtDNA) e nuclear (nDNA), não tendo nenhum deles uma definição suficiente para poder responder a todas as aplicações.

6.2 Marcadores moleculares para identificação de fauna

O genoma mitocondrial

O genoma mitocondrial é, na maioria dos casos, uma molécula circular, que nos animais possui um tamanho aproximado de 16 mil pares de base (pb). Sua estrutura é altamente conservada e inclui 13 genes codificadores de proteína (PCGs), dois genes ribossomais (rRNAs), 22 genes de RNAs de transferência e duas regiões não codificadoras que regulam a replicação e a transição do mtDNA: a região controle (*D-loop*) e a origem de replicação da cadeia de leve (OL).

O genoma mitocondrial é frequentemente usado para testar relações filogenéticas entre categorias taxonômicas menos inclusivas, devido à sua taxa de mutação ser, em média, maior do que a do genoma nuclear. Além disso, comparado com o genoma nuclear, o tamanho de sua molécula é menor e possui um maior número de cópias por célula, o que permite o uso de pequenas quantidades de tecido para obter uma alta qualidade nas análises.

Na maioria das espécies, o mtDNA provém apenas do óvulo e o mtDNA do esperma é degradado pela célula fecundada, ou em outros casos, o mtDNA do esperma é incapaz de ingressar no óvulo, fazendo com que o DNA mitocondrial seja herdado somente da mãe. Esta origem uniparental e a ausência de recombinação nos haplótipos, reduz a diversidade intraespecífica, fortalecendo assim a utilização destas regiões para identificar indivíduos de uma mesma espécie.

DNA barcode

A técnica do DNA *barcode* foi inicialmente proposta por Hebert e colaboradores em 2002, e consiste em comparar sequências de um fragmento de aproximadamente 650pb do gene Citocromo oxidase subunidade I (COI) entre espécimes a serem estudados.

O uso do DNA *barcode* tem sido amplamente utilizado na identificação de diversos metazoários, desde invertebrados a vertebrados. Entre os inúmeros exemplos que o gene COI permitiu diferenciar espécies filogeneticamente próximas, pode-se destacar o caso da borboleta *Astraptes fulgerator* (Walch, 1775), cujos resultados revelaram que o táxon, na verdade, representa um complexo de 10 espécies.

O método de análise do DNA *barcode* inclui a amplificação por reação em cadeia da polimerase (PCR) do gene COI e o sequenciamento através de eletroforese capilar dos *amplicons* obtidos. Logo, as sequências são comparadas nas bases de dados pré-existentes (BOLD, NCBI), e estima-se a distância e a similaridade genética entre a sequência alvo e as sequências depositadas nas bases de dados. O modelo e o algoritmo utilizados são Kimura-2 parâmetros (K2P) e *Neighbor Joinning* (NJ) propostos por Nei e Kumar (2000), respectivamente. O modelo K2P considera taxas diferentes para as transições e transversões, e NJ reconstrói árvores filogenéticas a partir de matrizes de distância geradas com base no modelo evolutivo.

Apesar de sua aplicabilidade, esta técnica também possui limitações. Foi apontado por Moritz e Cicero (2004), que sobreposições são frequentes quando táxons muito proximamente relacionados são incluídos no mesmo estudo. Um exemplo é o estudo de Smith e colaboradores (2008) realizado em anfíbios, onde houve uma sobreposição na divergência intra e interespecífica, consequência provavelmente da hibridação dos grupos, ou uma taxonomia ainda não totalmente resolvida.

Taxonomia integrativa

Desde pelo menos a metade do século XX, sistematas já sugeriam utilizar dados conjuntos de diferentes sistemas para elaborar uma hipótese taxonômica. O etólogo Nikko Tinbergen, por exemplo, advogava o uso de caracteres comportamentais para complementar hipóteses taxonômicas baseadas apenas em morfologia. Nos anos 2000, surge e populariza-se o termo **Taxonomia integrativa**. Da mesma maneira que o proposto no século anterior, a taxonomia integrativa, usualmente, visa integrar dados genotípicos e fenotípicos para a construção de uma hipótese filogenética mais robusta e com maior poder preditivo, embora também sejam utilizadas outras fontes de informação, como por exemplo em um estudo de poríferos, no qual a composição química das espículas pode ser usada como informação ecológica e taxonômica.

A escolha dos marcadores e sua especificidade

A identificação de espécies usando marcadores genéticos busca isolar regiões no genoma que sejam diferenciais e conservadas dentro do grupo taxonômico alvo. Em geral, a escolha entre as diferentes possibilidades de marcadores depende de alguns fatores, entre os quais está o custo e a eficácia para identificar o indivíduo, espécie ou população para que foi desenhado. A seguir, listamos brevemente algumas das principais características moleculares dos genes mitocondriais que têm demonstrado uma grande eficácia em estudos que envolvem identificação molecular de espécies:

Citocromo b (Cyt-*b*)

O gene mitocondrial Citocromo b é codificado na fita pesada que compreende uma região de tamanho aproximado de 1140pb, sendo sua posição na molécula dependente da espécie e relativa à origem de replicação. É a única proteína sintetizada pelo genoma mitocondrial que pertence ao conjunto das 10 proteínas do Complexo III do sistema mitocondrial de fosforilação oxidativa. Essa molécula apresenta uma grande variação interespecífica e pouca variação intraespecífica.

Citocromo Oxidase subunidade I (COI)

Pertence ao Complexo IV. É uma das regiões mitocondriais mais conservadas, com um tamanho de aproximadamente 1550pb. Possui reduzida taxa de evolução, devido ao fato deste gene codificar proteínas necessárias para a produção de energia celular. Essa característica faz com que esta região altamente conservada seja ideal para o desenho de *primers* específicos para um grupo taxonômico menos inclusivo.

Genes mitocondriais ribossomais: *12S*-RNA e *16S*-RNA

Estão classificadas pelo seu tamanho, sendo a subunidade *16S* maior do que a subunidade *12S*, com um tamanho de aproximadamente de 1559pb e 959pb, respectivamente. O gene *12S* é o primeiro gene estrutural após a região de controle e está precedido pelo tRNA-*Phe*. O gene *12S* está unicamente separado do *16S* pelo tRNA-*Val*, embora sua posição seja variável. Seu papel no metabolismo celular é realizar a tradução do RNA mensageiro em proteínas mitocondriais. Embora possuam muitas substituições, a estrutura de ambos genes compartilha funções e estruturas em diversos organismos.

Região controle (*D-loop*)

No genoma mitocondrial existem duas regiões não codificantes que abrangem menos de 7% do total da sequência, das quais, a maior e principal delas é o *loop* de deslocamento ou *D-loop*, contendo a característica tripla fita. Nesta região encontram-se os promotores de transcrição e a origem de replicação, servindo assim como região de controle da expressão do mtDNA. A região não codificante evolui de duas a cinco vezes mais rápido que o Cyt-*b* e o COI, que são os genes mitocondriais codificadores com as taxas evolutivas mais elevadas e, assim, apresentam maior variabilidade.

Devido a sua alta variabilidade, a região terminal da extremidade 5' do *D-loop* é amplamente utilizada para resolver questões evolutivas e populacionais no nível inter e intraespecífico em peixes, embora essa alta variabilidade não ocorra em todos os táxons. Quando usado como método de identificação, é esperado que os *primers* desenhados na região *D-loop*, tenham uma eficiência limitada para identificar espécies filogeneticamente mais próximas.

6.3 Problemática: Marcadores moleculares e Taxonomia

6.3.1 Importância dos espécimes-*voucher*

Todo tecido coletado deve estar associado a um espécime-*voucher* ou testemunho. Usualmente um espécime-testemunho (chamado em inglês de *voucher specimen*) é um indivíduo completo que serve de referência para o tecido coletado, e é mantido em uma coleção biológica (ver Capítulo 9). Tal espécime é imprescindível para verificação taxonômica e morfológica do táxon, assim como para a replicabilidade do estudo.

A informação associada ao espécime inclui dados sobre o coletor, dados biológicos (categoria taxonômica, características morfológicas que se perdem no processo de fixação), dados geográficos (localidade e coordenadas), dados ecológicos, entre outros. É muito importante associar uma etiqueta ao espécime que relacione a coleção zoológica onde está depositado o *voucher* com a coleção de tecidos, onde está depositada a amostra de músculo, escamas, penas, sangue, entre outras.

Erros na identificação taxonômica são frequentes inclusive quando os espécimes-testemunho são preservados. O maior problema ocorre quando se tem unicamente os tecidos preservados que torna mínima a possibilidade de corrigir o erro. Os tecidos coletados sem um espécime associado empobrecem as coleções de tecido, já que os erros de identificação podem ser perpetuados. Porém, existem algumas exceções como no caso de espécies raras ou ameaçadas e facilmente identificáveis no campo. Nestes casos, é possível coletar sangue ou pelos de maneira não-destrutiva, e realizar a identificação em campo.

6.3.2 Crimes ambientais: e quando não temos o espécime?

Atualmente, a fauna silvestre é um recurso de grande importância, que tem abastecido as necessidades alimentares mundialmente. Um grande número de espécies de peixes, por exemplo, é comercializado para o consumo e produção de alimento, representando 17% do total de proteína animal consumida no mundo. Segundo a Organização das Nações Unidas para a Alimentação e a Agricultura – FAO (2016), os peixes foram o maior incremento de consumo per-capita durante o ano 2016 nos países como o Brasil, Chile, China, México e Peru.

O controle da pesca e caça ilegal é dificultado pela identificação imprecisa de espécies, uma vez que espécies protegidas são usualmente comercializadas e referidas por nomes populares, que frequentemente não refletem as divisões taxonômicas. Erros de identificação e rotulagem incorreta de produtos derivados da pesca e caça ilegal ocorrem regularmente, e propositalmente com a finalidade de mascarar uma espécie de valor módico por outra de maior valor no mercado, ou para o comércio de espécies protegidas. Estas substituições ilegais são facilitadas pela dificuldade e/ou impossibilidade de reconhecimento das espécies usando caracteres morfológicos em produtos de músculo congelado ou industrializados.

Deste modo, a disponibilidade de marcadores moleculares específicos para as espécies envolvidas em crimes ambientais fortalece o controle e identificação de espécies. A efetividade de cada marcador irá depender da qualidade das amostras obtidas, sendo que quanto maior o grau de decomposição, menor será o tamanho do fragmento a ser obtido. Desta forma o DNA *barcode*, que amplifica fragmentos a partir de 600pb, pode não representar sempre a ferramenta mais efetiva em tecidos sometidos a altas temperaturas, processamento físico-químico, entre outros que provocam a fragmentação das fitas de DNA.

6.4 Bibliografia recomendada

FARRIS, James. The logical basis of phylogenetic analysis. 1983. In: Advances in Cladistics proceedings of the second meeting of the Willi Hennig Society (Platnick N, Funk VA, eds.). Columbia University Press, New York: 1-36.

AGOSTINHO, Ângelo A.; THOMAZ, SIDINEI M.; GOMES, LUIZ C. Conservação da biodiversidade em águas continentais do Brasil. Megadiversidade, v. 1, n. 1, p. 70-78, 2005.

ARDURA, Alba et al. DNA barcoding for conservation and management of Amazonian commercial fish. Biological Conservation, v. 143, n. 6, p. 1438-1443, 2010.

ARIF, Ibrahim A. et al. DNA marker technology for wildlife conservation. Saudi journal of biological sciences, v. 18, n. 3, p. 219-225, 2011.

ASAKAWA, Shuichi et al. Strand-specific nucleotide composition bias in echinoderm and vertebrate mitochondrial genomes. Journal of molecular evolution, v. 32, n. 6, p. 511-520, 1991.

BØRSTING, Claus; MORLING, Niels. Next generation sequencing and its applications in forensic genetics. Forensic Science International: Genetics, v. 18, p. 78-89, 2015.

DOOSTI, Abbas; DEHKORDI, Payam Ghasemi. Genetic polymorphisms of mitochondrial genome *D-loop* region in Bakhtiarian population by PCR-RFLP. International Journal of Biology, v. 3, n. 4, p. 41, 2011.

DAWNAY, Nick et al. Validation of the barcoding gene COI for use in forensic genetic species identification. Forensic Science International, v. 173, n. 1, p. 1-6, 2007.

DEAGLE, Bruce E. et al. DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. Biology letters, v. 10, n. 9, p. 20140562, 2014.

HEBERT, Paul DN; RATNASINGHAM, Sujeevan; DE WAARD, Jeremy R. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proceedings of the Royal Society of London B: Biological Sciences, v. 270, n. Suppl 1, p. S96-S99, 2003.

HEBERT, Paul DN et al. Biological identifications through DNA barcodes. Proceedings of the Royal Society of London B: Biological Sciences, v. 270, n. 1512, p. 313-321, 2003.

JAMANDRE, Brian Wade; DURAND, Jean-Dominique; TZENG, Wann-Nian. High sequence variations in mitochondrial DNA control region among worldwide populations of flathead mullet Mugil cephalus. International Journal of Zoology, v. 2014, 2014.

HUFFMAN, Jane E.; WALLACE, John R. Wildlife forensics: methods and applications. John Wiley & Sons, 2012.

VAN DIJK, Erwin L. et al. Ten years of next-generation sequencing technology. Trends in genetics, v. 30, n. 9, p. 418-426, 2014.

LARIZZA, Alessandra et al. Lineage specificity of the evolutionary dynamics of the mtDNA *D-loop* region in rodents. Journal of molecular evolution, v. 54, n. 2, p. 145-155, 2002.

LINACRE, Adrian; TOBE, Shanan. Wildlife DNA analysis: applications in forensic science. John Wiley & Sons, 2013.

MORITZ, Craig; CICERO, Carla. DNA barcoding: promise and pitfalls. PLoS biology, v. 2, n. 10, p. e354, 2004.

NEI, Masatoshi; KUMAR, Sudhir. Molecular evolution and phylogenetics. Oxford University Press. 2000.

OGDEN, Rob; DAWNAY, Nick; MCEWING, Ross. Wildlife DNA forensics—bridging the gap between conservation genetics and law enforcement. Endangered Species Research, v. 9, n. 3, p. 179-195, 2009.

PEREIRA, Luiz HG et al. Can DNA barcoding accurately discriminate megadiverse Neotropical freshwater fish fauna?. BMC genetics, v. 14, n. 1, p. 20, 2013.

ROJAS, María et al. Development of a real-time PCR assay to control the illegal trade of meat from protected capercaillie species (Tetrao urogallus). Forensic science international, v. 210, n. 1, p. 133-138, 2011.

SATOH, Takashi P. et al. Structure and variation of the mitochondrial genome of fishes. BMC genomics, v. 17, n. 1, p. 719, 2016.

VIEIRA, Fabio et al. Peixes. Em: DRUMMOND, Glaucia Moreira et al. Biota Minas: Diagnostico do conhecimento sobre a biodiversidade no Estado de Minas Gerais-subsidio ao Programa Biota Minas. In: Biota Minas: Diagnóstico do conhecimento sobre a biodiversidade no Estado de Minas Gerais-subsidio ao Programa Biota Minas. Fundação Biodiversitas, 81-121 p. 2009.

YANG, Dongya Y.; SPELLER, Camilla F. Co-amplification of cytochrome b and D-loop mtDNA fragments for the identification of degraded DNA samples. Molecular Ecology Resources, v. 6, n. 3, p. 605-608, 2006.

YE, Jian et al. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC bioinformatics, v. 13, n. 1, p. 134, 2012.

Capítulo 7
Chaves Taxonômicas

Bárbara Teixeira Faleiro

7.1 Chaves taxonômicas

O objetivo das chaves taxonômicas é auxiliar na identificação de um espécime até uma determinada categoria taxonômica. Elas podem ser construídas para qualquer nível taxonômico: para espécies, gêneros, famílias, entre outros. É válido destacar que as chaves de identificação não necessariamente refletem as relações filogenéticas entre os táxons, bem como os estados dos caracteres utilizados não necessariamente são homólogos entre si. São ferramentas que geralmente se restringem a um grupo taxonômico de uma dada região geográfica (por exemplo: Chave para as espécies Neotropicais do gênero *Lathrolestes* (Hymenoptera, Ichneumonidae)). Existem dois tipos principais de chaves taxonômicas: as chaves tradicionais ou dicotômicas, e as interativas ou de múltiplos-acessos.

7.2 Chaves taxonômicas tradicionais

A primeira chave taxonômica dicotômica foi desenvolvida por Richard Waller em seu trabalho *Tables of the English Herbs reduced to such an order, as to find the name of them by their external figures and shapes* (Figura 7.1). Hoje, as chaves tradicionais (Figura 7.2) são uma ferramenta amplamente utilizada para a identificação de táxons. Elas estão presentes em grande parte dos trabalhos de descrição e revisão de táxon. Essas chaves possuem um acesso único com duas (dicotômica) ou mais (politômica) alternativas mutuamente excludentes que conduzem ao próximo passo, e assim por diante até que se chegue na identificação do espécime.

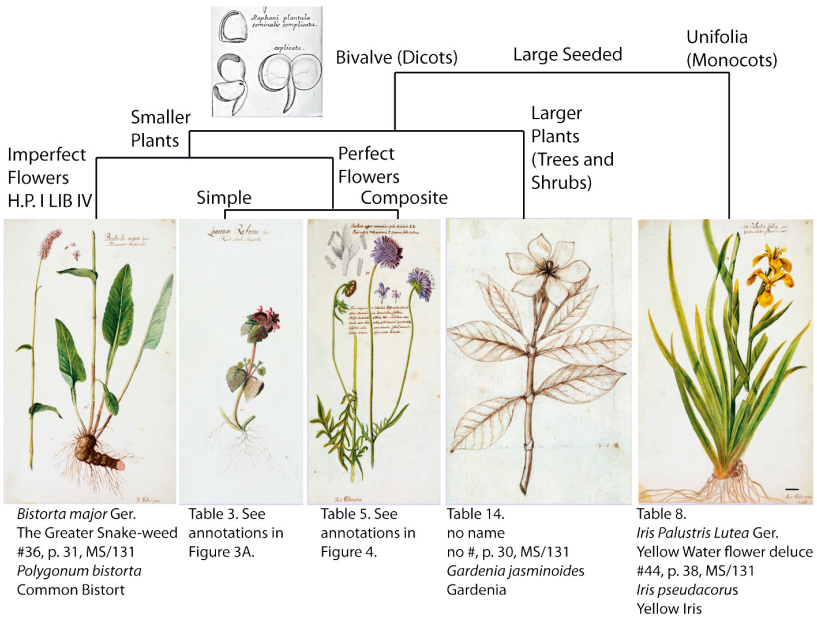


Figura 7.1 Representação das tabelas de Richard Waller, modificado de Griffing (2011). As tabelas de Waller apresentam os caracteres de identificação inseridos em um diagrama de árvore que se ramifica dicotomicamente, possuindo no final de seus ramos a pranchas das espécies de plantas, também conhecidas como chaves pictóricas.

1 Area superomedia of propodeum twice as wide as long; first metasomal tergite at most as long as wide apically; fore wing *3rs-m* stub like and almost evanescent, not forming a true areolet 2
1' Area superomedia of propodeum longer than wide; first metasomal tergite 1.25–2× as long as wide apically; fore wing *3rs-m* well pigmented, forming a conspicuous areolet 3
2 Metasomal tergite I with lateromedian longitudinal carina extending more than half its length; face, gena, hind legs and metasoma black (Ecuador) . . . . . *L. gauldi* Reshchikov, Veijalainen & Sääksjärv, 2012
2' Metasomal tergite I without lateromedian longitudinal carina; face, gena, hind coxa and femur, and metasomal tergites reddish; remaining parts of hind legs and the last three tergites black (Peru) . . . . . *L. fiedleri* Reshchikov, 2015
3 Mesoscutum with notaulus strongly impressed anteriorly; upper part of head and mesosoma granulate, mat; lateromedian longitudinal carinae of first metasomal tergite separated centrally by about the diameter of its spiracle. (Costa Rica) . . . . . *L. karenae* Gauld, 1997
3' Mesoscutum with notaulus vestigial or absent; upper part of head and mesosoma fairly smooth and polished; lateromedian longitudinal carinae of first metasomal tergite separated centrally by far more than the diameter of its spiracle 4
4 Tergites of metasoma reddish-yellow 5
4' Tergites of metasoma black and yellow or only some tergites reddish-yellow 7
5 Fore wing hyaline at apex; first tergite of metasoma 1.25× as long as wide apically (Mexico) . . . . . *L. tepeyollotlis* Reshchikov, 2011
5' Fore wing infuscate at apex; first tergite of metasoma 1.4–1.7× as long as wide apically 6
6 First tergite of metasoma 1.7× as long as wide apically; lower mandible tooth clearly longer than upper; clypeus reddish-yellow apically. Pronotum, mesoscutum and mesopleuron reddish-yellow. (Mexico) . . . . . *L. quetzalcoatlus* Reshchikov, 2011
6' First tergite of metasoma 1.4× as long as wide apically; lower mandible tooth about as long as upper; clypeus black apically. Pronotum black with anterior border yellow; mesoscutum black with yellow marks; mesoscutum reddish-yellow with apical black band and subalar prominence yellow. (Brazil) . . . . . *L. piranga* Lima & Kumagai **sp. n.**

Figura 7.2 Exemplo de uma chave taxonômica tradicional dicotômica. Chave para as espécies Neotropicais do gênero Lathrolestes (Hymenoptera, Ichneumonidae), retirado de Lima & Kumagai (2016).

7.3 Chaves taxonômicas interativas

As chaves taxonômicas interativas ou multiacesso são construídas em programas de computador (exemplo: Xper2) (Figura 7.3) e funcionam por eliminação da lista de possíveis resultados, ou seja, ao selecionar um estado de caráter ela mantêm todos os táxons que possuem aquele estado selecionado e exclui todos aqueles que não o possuem. Por isso, para se construir uma chave interativa é necessário construir uma matriz com todos os táxons e caracteres incluídos. As chaves interativas, embora possuam o mesmo objetivo das chaves tradicionais, diferem enormemente destas e apresentam algumas vantagens:

1. As chaves interativas são multiacesso, ou seja, o usuário pode escolher por qual caráter iniciar a identificação, assim como todos os seguintes. Pois, diferentemente das chaves tradicionais, as chaves interativas não possuem uma sequência hierárquica de passos. Assim, o usuário pode inserir as informações sobre os caracteres que ele dispõe do organismo que se pretende identificar.
2. Não há limites para a inclusão de informações nas chaves interativas. Elas permitem a inclusão de links, fotos, vídeos, sons, imagens, entre outras. Tornando-as uma ótima ferramenta para disseminar o conhecimento taxonômico. Além de torná-las mais atrativas e convidativas.
3. A atualização das chaves interativas é mais fácil e rápida. Para se incluir novos táxons e/ou caracteres na chave, basta adicioná-los à matriz. Dessa forma também é mais fácil alterar as informações associadas aos caracteres e/ou táxons.
4. As chaves interativas podem ser facilmente publicadas online (exemplo: Xper3), tornado o acesso a elas mais fácil e prático.

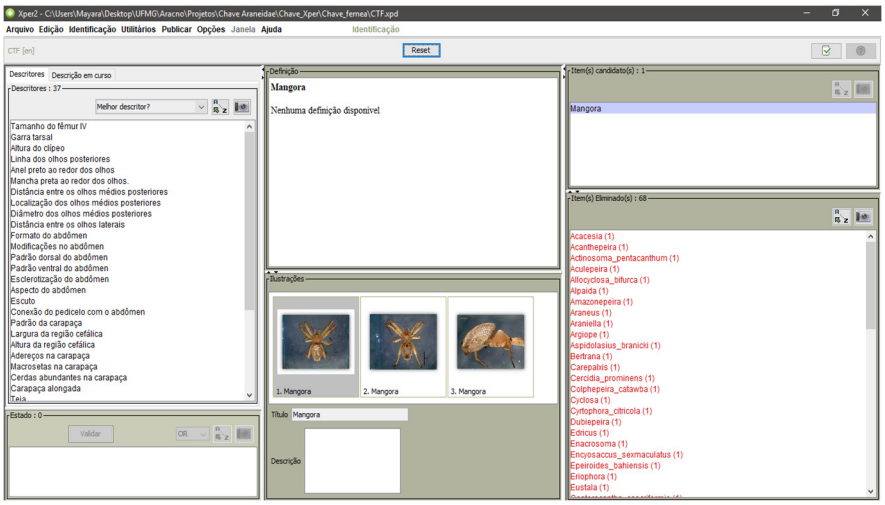


Figura 7.3 Exemplo de uma chave taxonômica interativa construída no programa Xper2. Chave taxonômica interativa para os gêneros da família Araneidae (Araneae).

7.4 Bibliografia recomendada

GRIFFING, Lawrence R. Who invented the dichotomous key? Richard Waller's watercolors of the herbs of Britain. American Journal of Botany, v. 98, n. 12, p. 1911-1923, 2011.

LIMA, Alessandro Rodrigues; KUMAGAI, Alice Fumi. Lathrolestes Förster, 1869 (Hymenoptera, Ichneumonidae) from Brazil, with description of two new species and a key to the Neotropical species. Zootaxa, v. 4170, n. 3, p. 587-593, 2016.

Xper2, Laboratoire Informatique & Systématique <http://www.infosyslab.fr/?q=en>

Capítulo 8
Métodos de coleta de material biológico, desenho experimental e vieses de amostragem

Leonardo Sousa Carvalho

8.1 Por que coletar material biológico?

Por que coletar material biológico? A coleta (conceituada aqui como captura seguida de eutanásia) de animais para estudos é uma atividade fundamental da pesquisa zoológica, praticada desde os primórdios da biologia moderna. Podemos, por exemplo, lembrar os clássicos estudos desenvolvidos por grandes cientistas como Carl Linnaeus (1707-1778), Charles Darwin (1809-1882) e Alfred Russel Wallace (1823-1913), dentre outros, que requereram estudos comparativos da anatomia, morfologia e história natural de animais de diversos grupos taxonômicos. Para o desenvolvimento destes estudos, a coleta de muitos exemplares foi requerida. É ainda mais surpreendente saber que alguns dos exemplares coletados e estudados por estes cientistas permanecem preservados em Museus pelo mundo.

No entanto, um leitor mais atento pode se perguntar se, hoje em dia, a coleta de animais ainda é um procedimento necessário para a pesquisa zoológica. Vivemos em uma época em que métodos modernos de estudo de animais permitem a coleta de um volume muito grande de informações. Câmeras, *datallogers*, e rádios-colares com GPS são rotineiramente utilizados em animais grandes como baleias e elefantes, até animais menores como caranguejeiras e abelhas. Esses equipamentos não provocam a morte do animal e ainda proveem informações sobre seus ritmos circadianos e sua área de vida. Paralelamente, vivemos em um período em que ainda há necessidade de ações de educação ambiental, a fim de conscientizar a população sobre a importância da preservação da biodiversidade, antes que a ação humana provoque a perda de um número ainda maior de espécies.

Ante as novas técnicas de estudo de animais e a atual crise da biodiversidade que passamos, cabe-nos fazer duas perguntas: seria a coleta para fins científicos um procedimento em desuso? Estaria a pesquisa zoológica afetando negativamente as populações animais, através da coleta de um número grande de exemplares? A resposta para ambas perguntas é, seguramente, não; e o principal motivo disto é que o conhecimento sobre a biodiversidade mundial é extremamente heterogêneo e urge por avanços ainda maiores. Estes problemas são denominados déficits científicos, sendo os mais conhecidos os déficits lineano, wallaceano, hutchinsoniano e prestoniano.

Ainda não conhecemos todas as espécies que existem na Terra (déficit lineano). Mesmo em regiões onde há reconhecidamente um menor número de espécies (por exemplo: regiões temperadas ou desertos) ou para táxons melhor conhecidos (como mamíferos e aves) ainda existem espécies que só foram reconhecidas e/ou descritas recentemente. Nos últimos cinco anos, por exemplo, foram reconhecidas novas espécies de baleias, grandes primatas, tamanduaís, antas, e diversas aves, entre muitos outros. Se estendermos esta análise a grupos megadiversos (como artrópodes e outros táxons de invertebrados), biomas tropicais (por exemplo: floresta amazônica ou mata atlântica) ou áreas de difícil acesso (por exemplo: regiões montanhosas ou localidades longe de vias de acesso), certamente o número de espécies ainda não descritas será extremamente elevado.

Além disso, mesmo para espécies já descritas, diversos estudos já apresentaram evidências que conhecemos menos da sua distribuição geográfica (déficit wallaceano) ou do nicho ambiental (características do ambiente que limitam a distribuição de uma espécie; déficit hutchinsoniano) ocupado por elas, que esperado. Ou seja, se, por exemplo, realizarmos a modelagem da distribuição geográfica esperada de uma espécie, para a maioria as espécies conhecidas, a área esperada será maior que as melhores estimativas da sua área de ocorrência previamente conhecida. Se pensarmos em uma escala mais refinada, sabemos menos ainda sobre os padrões de abundância das comunidades animais ou a dinâmica de populações da maioria das espécies de animais do planeta (déficit prestoniano). Especialmente, carecemos de séries temporais que permitam inferências sobre as respostas de populações animais diante das mudanças climáticas ou outras alterações estocásticas no ambiente. Em termos de espécie, raros são os casos em que a variação inter- e intraespecífica, seja ela comportamental, morfológica ou molecular, é conhecida.

Visto todos esses problemas no conhecimento acerca da biodiversidade mundial é notório que precisamos coletar mais. Entretanto, a segunda pergunta permanece não respondida. Estaria a coleta de espécimes para pesquisa

influenciando negativamente as populações de animais? Alguns cientistas acreditam que sim, especialmente para organismos com populações pequenas e isoladas. A argumentação desses cientistas cita o exemplo da alca-gigante, *Pinguinus impennis* Bonnaterre, 1791, uma espécie de ave endêmica da Islândia, que foi extinta em 1844. Acreditava-se que a caça excessiva para o consumo de carne e penas, associado a mudanças climáticas e a coleta demasiada de espécimes para museus tenha causado a sua extinção. Autores que propuseram estas causas para sua extinção, sugeriram a utilização de métodos alternativos, tais como fotografias, gravações de áudio e amostragem de DNA através de técnicas não-invasivas, como suficientes para identificar espécies ainda não descritas ou reconhecer espécies previamente consideradas extintas. Assim, seria desnecessária a realização de novas coletas.

Em contrapartida, essas ideias foram amplamente rebatidas. A alca-gigante, por exemplo, apresenta apenas 102 exemplares depositados em coleções científicas. A maioria desses espécimes são representados apenas por esqueletos coletados de carcaças e obtidos após a extinção dessa espécie, enquanto que milhões de indivíduos foram explorados para obtenção de óleo, carne e penas. Além disso, atualmente existem normas éticas e regulamentações que devem ser estritamente seguidas por pesquisadores, de modo que o número de exemplares coletados esteja abaixo daquilo que se considera necessário para a perpetuação de cada espécie. No contexto do pensamento evolutivo populacional, coletas são necessárias para estudar a diversidade intra e interespecífica – como mencionado anteriormente – e ainda entender a sua evolução. Adicionalmente, a existência de espécimes preservados permite o estabelecimento de marcos históricos importantes para o monitoramento da saúde dos indivíduos de uma espécie (exemplo: presença ou ausência de um determinado fungo), sua distribuição geográfica e variações no seu fenótipo.

Assim, a simples tarefa de coleta de material biológico permite a elucidação de muitas das questões citadas acima. Porém, para cada uma dessas questões e para cada grupo taxonômico, deve-se adequar a metodologia de coleta e o método de preservação aos organismos coletados. A coleta científica de um animal deve prover o máximo de informações possíveis e permitir o máximo de estudos futuros quanto possíveis. Não é raro, hoje em dia, coletarmos um animal e já armazenarmos uma amostra de seu tecido, para futuras análises utilizando métodos moleculares. Isto não era rotina, talvez apenas duas décadas atrás. Seguramente acreditamos que espécimes que coletamos hoje serão utilizados de maneiras inimagináveis pelos futuros cientistas.

Neste capítulo, não objetivo descrever detalhadamente cada método de coleta e preparação de vertebrados e invertebrados, pois a literatura sobre o assunto é abundante. Sugiro que o leitor interessado no assunto veja a bibliografia sugerida ao final deste capítulo, onde diversos livros e artigos sobre o assunto encontram-se referenciados. Focarei então em discutir questões mais amplas que influenciam a tomada de decisão sobre qual método escolher para cada pergunta.

8.2 Desenho experimental e vieses de amostragem

Um dos tópicos menos debatido durante a graduação em Ciências Biológicas, talvez seja o delineamento amostral. Paradoxalmente, essa disciplina é fundamental e muito difundida entre aqueles que realizam experimentos nas áreas de bioquímica, parasitologia, ciências agrárias, entre outras. Um bom delineamento amostral é importante para diminuir o viés de uma amostragem ou permitir a realização de testes robustos de hipóteses. Por exemplo, um pesquisador deseja testar o efeito de uma variável qualquer no crescimento de um determinado inseto, como por exemplo ingestão de cálcio. Então, ele decide criar todos os insetos com alguma dieta suplementada de cálcio. Dessa forma, mesmo se ele observar uma relação positiva entre o aumento da ingestão de cálcio e crescimento dos animais não será possível testar isoladamente o efeito desta variável, pela ausência de grupos-controle no experimento, ou seja, insetos criados sob dieta sem suplementação de cálcio e que sirvam de comparação com o grupo-teste.

Outro erro comum, é a utilização de variáveis não relacionadas diretamente ao problema. Por exemplo: medir a relação entre a fauna de artrópodes e a estrutura da vegetação de uma determinada localidade. Como variáveis de estrutura de vegetação, o pesquisador escolhe medir a densidade da vegetação, cobertura do dossel, número de árvores com diâmetro do tronco medido a altura do peito (em torno de 150cm acima do nível do solo) e profundidade da serapilheira. Os artrópodes serão coletados utilizando-se apenas guarda-chuva entomológico, um método ativo de coleta de animais que habitam os estratos herbáceo, subarbustivo e arbustivo da vegetação. Neste exemplo, correlações entre a fauna de artrópodes e a profundidade da serapilheira tendem a apresentar maior viés dado a efeitos indiretos ou estocásticos. A fauna de artrópodes que habita diretamente a serapilheira não foi amostrada, aumentando assim a possibilidade de encontrar correlações espúrias entre estas variáveis.

Os erros listados acima parecem um tanto fáceis de serem detectados. Vamos a outro exemplo: inventário da fauna de borboletas do Parque Nacional da Chapada Diamantina, na Bahia. Esta unidade de conservação ocupa uma extensa área de relevo bastante acidentado, havendo áreas de campos de altitude, enclaves florestais, matas de galeria ao longo de riachos e pequenos rios e ainda áreas de caatinga arbustiva-arbórea. Um bom desenho amostral das borboletas da região deveria incluir amostragens em diversos pontos em cada uma destas fitofisionomias, incluindo períodos de seca e de chuva. Se o trabalho ainda objetivar avaliar a sazonalidade das borboletas da região, uma amostragem mais duradoura (exemplo: três anos) é ainda recomendada, visto que a realização de coletas em anos atípicos (muito secos ou muito chuvosos) poderá enviesar significativamente a amostragem.

Poucos são os estudos que compararam explicitamente o efeito do desenho experimental sobre a amostragem de animais, especialmente em regiões tropicais. Para inventários de fauna, existem diversos protocolos propostos para grupos específicos. De maneira geral, eles diferem em duas propostas de organização da amostragem: design sistemático e design estratificado.

O design sistemático talvez seja a forma menos idiossincrática de amostragem, produzindo uma amostragem uniforme da área de estudo. Pode-se, por exemplo, dividir a área de estudo em regiões uniformes (exemplo: delimitação com grades com células de 1km² ou 10km²). Em cada evento amostral, uma ou mais áreas podem ser sorteadas ao acaso e a amostragem ser realizada nos locais sorteados. Alternativamente, pode-se escolher um número aleatório (exemplo: 6) e aplicar o protocolo de amostragem nas áreas múltiplas deste número (exemplo: 6, 12, 18, 24). Na figura 8.1 A, note que se propõe a amostragem a cada seis áreas a partir da primeira área no canto superior esquerdo, contando-se para a direita e para baixo. Este tipo de amostragem é indicado para ambientes pouco heterogêneos espacialmente, porém pode ser influenciado pela distância entre os pontos (ou tamanho da grade), forma de escolha dos pontos de coleta, número de pontos de coleta e ainda a relação destas variáveis e o táxon estudado. Ou seja, um delineamento amostral excelente para o exemplo do inventário de borboletas, pode não ser o ideal para a amostragem de aranhas ou de roedores. Deve-se avaliar criticamente um desenho experimental, considerando-se o objetivo do trabalho, a área de estudo e o táxon de estudo.

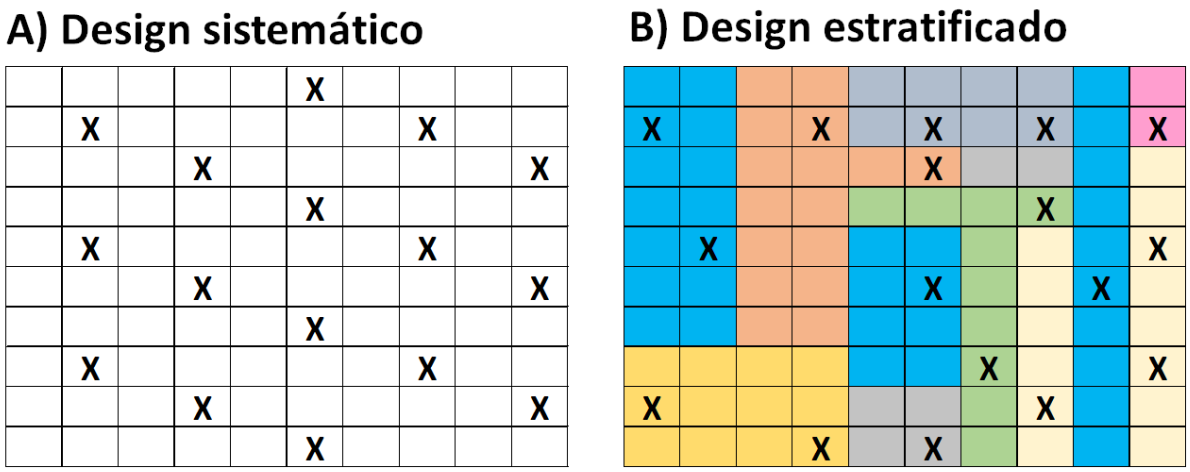


Figura 8.1 Desenho esquemático de um desenho amostral seguindo um design sistemático (A) ou estratificado (B). O retângulo delimita a totalidade da área de estudo do exemplo. Cada célula da grade representa uma parcela da área de estudo. Cores no design estratificado representam fitofisionomias diferentes da área de estudo. Áreas amostradas em cada design estão marcadas com um X.

O design estratificado, por sua vez, foi desenvolvido para minimizar a variação da amostragem entre estratos diferentes ou ainda quando as populações do táxon de estudo apresentam variações sistematicamente distribuídas pela área de estudo. Isto envolve um conhecimento prévio da região onde a amostragem será realizada, e/ou do táxon de estudo. Por exemplo: monitoramento da fauna de quelônios do Parque Nacional da Serra das Confusões, no Piauí. Na região, quelônios podem ser encontrados em lagoas, riachos e poças d'água temporárias. Logo, a aplicação de uma amostragem seguindo um design sistemático resultaria em eventos de amostragem em lugares onde quelônios podem não ocorrer. Assim, pode-se estratificar a área de estudo, reconhecendo os locais onde se espera que ocorram quelônios e então realizar a amostragem. Mas e se este mesmo estudo de monitoramento de quelônios fosse realizado em uma área na Amazônia ou no Pantanal, onde provavelmente existem extensas áreas propícias a sobrevivência de quelônios? Poderíamos ainda pensar em utilizar o design estratificado? Este tipo de reflexão deve sempre ser feita pelo pesquisador.

No design estruturado, a escolha dos locais de amostragem pode ser aleatória (como já exemplificado anteriormente) ou direcionada (opção que aumenta a subjetividade na amostragem). Pode-se ainda realizar sorteio dos pontos de coleta, considerando-se a disponibilidade de cada ambiente na região, garantindo assim que todos os tipos de ambientes ou fitofisionomias sejam amostrados (Figura 8.1 B). Este tipo de amostragem é indicado para ambientes muito heterogêneos espacialmente e pode ser influenciado pelos mesmos parâmetros que o design sistemático.

Estas conformações de design, embora exemplificados em uma perspectiva espacial, podem ser estendidos a uma perspectiva temporal. Em estudos de longa duração ou estudos que objetivem avaliar a sazonalidade de um determinado fenômeno, nem sempre é possível coletar dados continuamente. Assim, deve-se tomar decisões e espalhar temporalmente a amostragem. No exemplo do inventário de borboletas do Parque Nacional da Chapada Diamantina, pode-se optar por realizar amostragens mensais na área de estudo, coletando sempre em pontos fixos de cada fitofisionomia, permitindo então avaliar a dinâmica das comunidades de borboletas ao longo do tempo. Isto

segue, então, um design sistemático de amostragem. Este mesmo protocolo seria ineficaz para a amostragem da fauna de anfíbios da mesma região. Considerando que no semiárido brasileiro a sazonalidade é bem marcada, coletas de anfíbios em períodos secos seriam extremamente ineficazes. Assim, seria necessário estratificar temporalmente a amostragem, de modo a coletar mais durante o período chuvoso.

A organização espacial e/ou temporal da amostragem é um ponto extremamente importante do delineamento amostral. A realização de amostragens muito próximas espacialmente ou temporalmente pode resultar em amostras que não são independentes entre si. Estas amostras são então denominadas pseudoréplicas espaciais e/ou temporais. É impossível estabelecer regras gerais para estudos, visto que a interação das variáveis tempo, espaço, táxon de estudo, método de coleta e objetivo da pesquisa, resulta em combinações complexas. O pesquisador deve avaliar estas questões de maneira objetiva, considerando estes fatores. Ou seja, um protocolo de coleta de borboletas que resulte em amostras independentes, pode resultar em sérios problemas de dependência das amostras se for seguido para um inventário de aves, por exemplo.

Vejamos os modelos de disposição de armadilhas de queda para invertebrados (Figuras 8.2 A-B), ilustrados nas figuras 8.2 C-E. O primeiro modelo (Figura 8.2 C) apresenta quatro linhas de armadilhas, cada uma contendo 10 armadilhas, com distâncias de 10 metros entre armadilhas e 50 metros entre as linhas. O segundo modelo (Figura 8.2 D) apresenta armadilhas dispostas em um bloco, com espaçamento de 1 metro entre armadilhas. O terceiro modelo (Figura 8.2 E) apresenta armadilhas organizadas em estações de coleta. Neste modelo, um conjunto de cinco armadilhas são ligadas por pequenas lonas plásticas (Figura 8.2 B), formando uma estação de coleta; e as estações distam cerca de 100 metros entre si. Nestes três casos, se cada armadilha (ou estação de armadilhas) fosse tratado como uma amostra, há maior probabilidade de haver pseudoreplicação espacial entre amostras de invertebrados coletados pelo segundo modelo (Figura 8.2 D). Assim, a utilização de armadilhas de queda, dispostas em blocos, deveria considerar o conjunto de todos os invertebrados coletados por todas as armadilhas de um mesmo bloco em um evento amostral, como uma única amostra. Igualmente, pode-se considerar uma amostra o conjunto de todos os invertebrados coletados em cada armadilha, ou pelo conjunto de todas as armadilhas de cada linha, no primeiro modelo (Figura 8.2 C). O terceiro modelo (Figura 8.2 E), por sua vez, talvez apresente a definição mais clara do que constitui uma amostra: o conjunto de todos os invertebrados coletados por cada estação de armadilhas em um dado evento amostral.

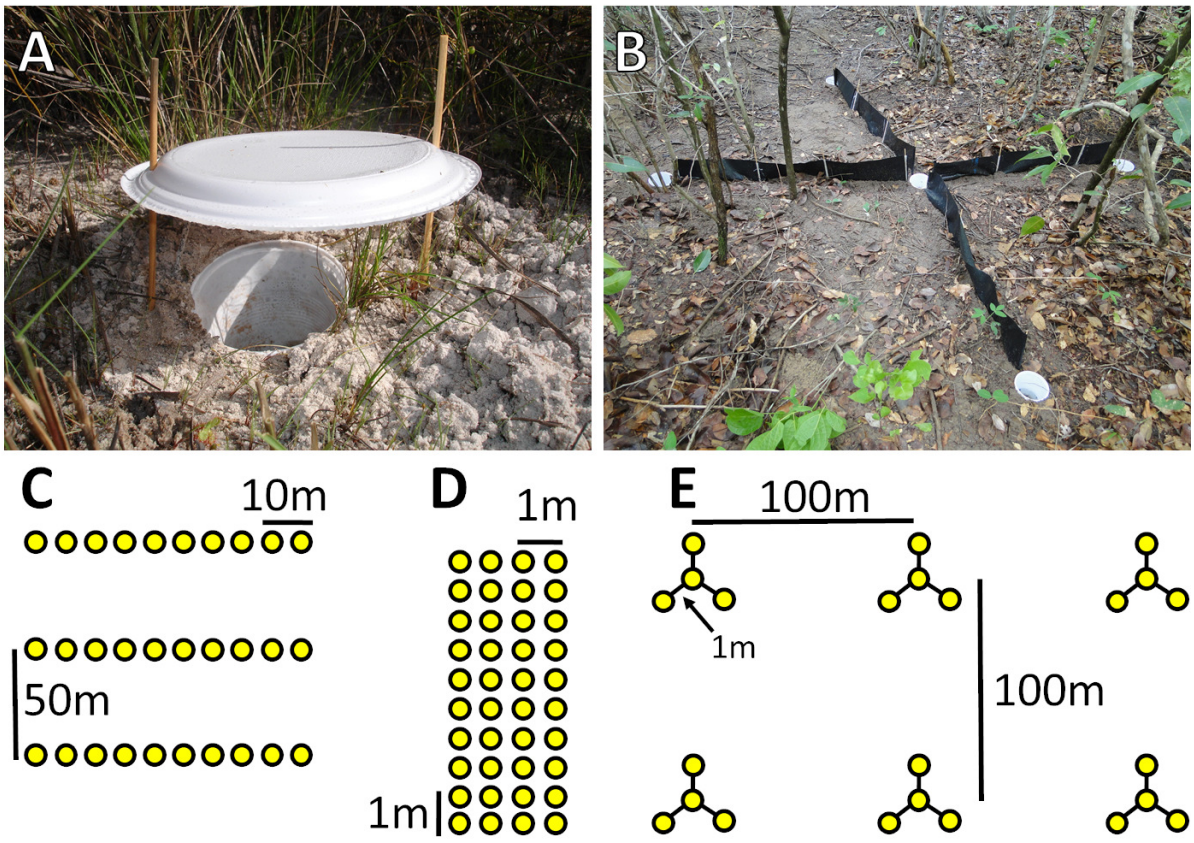


Figura 8.2 Armadilhas de queda para invertebrados (A-B) e variações de sua disposição (C-E), em linha (C), em blocos (D) ou em estações (E). Fotos: Leonardo Sousa Carvalho

8.3 Métodos de coleta de material biológico

A disposição e o tipo de armadilhas que serão utilizadas para a amostragem depende de três fatores: (1) objetivo da amostragem; (2) táxon de estudo; (3) área de estudo. Quando uma amostragem objetiva coletar o maior número de espécies de um táxon (exemplo: borboletas), sem se importar com padrões de abundância das espécies ou a comparação entre ambientes distintos na área de estudo, uma amostragem não padronizada poderá ser realizada. Neste caso, o conhecimento prévio do pesquisador será extremamente importante para a decisão de que tipo de método de coleta utilizar e em que ambientes os aplicar.

No inventário de borboletas, anteriormente exemplificado, poderíamos escolher utilizar armadilhas específicas para borboletas, que muitas vezes utilizam iscas atrativas compostas de misturas de frutas amassadas e caldo de cana. A decomposição dessa mistura libera odores que atraem os animais para o interior da armadilha, onde normalmente ficam presas. O pesquisador decidirá quantas armadilhas serão utilizadas, onde colocá-las, que tipo de isca será utilizada, a distância entre cada armadilha, entre outras. Igualmente, este mesmo pesquisador pode decidir realizar apenas coletas manuais, utilizando redes entomológicas que permitem a captura de insetos ativamente pelo pesquisador. A experiência do pesquisador é fundamental também neste caso, pois ele deverá decidir onde procurar por borboletas e deverá ter habilidade para manusear o equipamento e capturar os animais.

Neste exemplo, vemos a utilização de dois grupos de métodos de coleta que são complementares: um método passivo (armadilhas com iscas) e um método ativo (captura com rede entomológica). Cada uma desses grupos de métodos resulta em diferentes vieses de amostragem. De qualquer forma, a escolha do método de coleta a ser empregado deve sempre levar em consideração diversos fatores: (1) táxon de estudo, (2) relação custo-benefício, (3) fatores logísticos (transporte, necessidade e disponibilidade de energia elétrica, entre outros), (4) experiência e habilidade do coletor, e (5) os objetivos do trabalho.

Os métodos passivos são aqueles em que o pesquisador utiliza armadilhas para promover a captura ou coleta de animais. Dentre eles, podemos destacar as armadilhas de queda (*pit-fall traps*; Figura 8.2), armadilhas de queda com cercas-guia (*pit-fall traps with drift fences*), armadilhas fotográficas, amostragens indiretas (sons, pegadas, rastros, fezes, entre outras), termonebulizador de copa, gaiolas, ratoeiras, armadilhas de interceptação de voo (exemplo: malaise, redes de neblina) e qualquer tipo de armadilha que utilize iscas atrativas (exemplo: carne em putrefação, fezes, luz, cores, frutas, entre outras). Por outro lado, métodos ativos são aqueles em que o pesquisador captura ativamente os animais. Enquadram-se nessa categoria as coletas manuais, coletas crípticas, coletas com guarda-chuva entomológico, rede entomológica, rede de varredura, alguns tipos de redes de pesca, redes de plâncton, pesca com anzol, armas de fogo, entre outros.

8.4 Preparação de material biológico

Após a realização da coleta, é importante que o máximo de informações de cada espécime seja preservada, incluindo o próprio espécime. O depósito do material testemunho de uma pesquisa é parte fundamental do método científico, visto que permite a análise futura por outros pesquisadores. Para cada animal silvestre coletado e que será depositado em uma coleção científica, é imprescindível que informações sobre a procedência do animal seja armazenada na coleção. Outras informações dos animais coletados, tais como nome do coletor, data e horário de coleta, método de coleta, dados de história natural (exemplo: tamanho do grupo, ectoparasitas) e descrições de partes do corpo que perdem a cor ou são perdidas durante o preparo (exemplo: cor da íris de uma ave, ou a coloração in vivo de um anfíbio) podem ser igualmente armazenados e são sempre bem-vindos, porém não são obrigatórios. Da mesma forma, metadados associados a estes animais (tais como fotos, vídeos, gravações de vocalizações, entre outros) também podem ser incorporados aos acervos das coleções.

Cada grupo de animal apresenta formas apropriadas de preservação. Alguns, inclusive, apresentam mais de um método de preservação. A preservação de animais envolve duas etapas: a fixação e a conservação. Por exemplo, mamíferos podem ser mantidos em meio líquido, quando são fixados com solução de formaldeído e conservados em álcool etílico; ou em meio seco, sendo taxidermizados. Igualmente, um mamífero pode ser armazenado em meio líquido e, posteriormente, taxidermizado. Este não é um procedimento padrão e nem ideal, mas necessário em certos casos. Cabe ao pesquisador decidir o melhor método para realizar a preservação do material coletado. Normalmente, isso segue normas estabelecidas pelo local onde os animais serão depositados, bem como procedimentos de rotina para cada grupo de animal.

Como já mencionado anteriormente, não irei aqui detalhar métodos de preparação de material biológico, visto que há literatura abundante sobre o assunto. No entanto, é importante lembrar que a captura e eutanásia dos animais segue legislação específica e necessita de autorizações específicas, seja de algum Comitê de Ética em Pesquisa com Animais, do Instituto Chico Mendes de Conservação da Biodiversidade (ICMBio), ou ainda de órgãos ambientais estaduais e/ou municipais. Além disto, a Lei Federal nº 13.123, de 29 de maio de 2015 e o Decreto nº 8.772, de 11 de maio de 2016 criaram e regulamentaram as atividades sobre o acesso ao patrimônio genético (exemplo:

coleta, captura ou registro fotográfico de animais para fins de pesquisa, dentre outras), dentre outras atividades, devendo sua realização ser cadastrada no Sistema Nacional de Gestão do Patrimônio Genético e do Conhecimento Tradicional Associado – SisGen.

8.5 Bibliografia recomendada

ADIS, Joachim et al. Canopy fogging of an overstory tree-recommendations for standardization. Ecotropica, v. 4, p. 93-97, 1998.

ALTIG, Ronald. A primer for the morphology of anuran tadpoles. Herpetological conservation and biology, v. 2, n. 1, p. 71-74, 2007.

ARISTOPHANOUS, Marios. Does your preservative preserve? A comparison of the efficacy of some pitfall trap solutions in preserving the internal reproductive organs of dung beetles. ZooKeys, v. 34, p. 1-16, 2010.

BUCKLAND, Stephen T. et al. Point transect sampling with traps or lures. Journal of Applied Ecology, v. 43, n. 2, p. 377-384, 2006.

CALIXTO, Alejandro A.; HARRIS, Marvin K.; DEAN, Allen. Sampling ants with pitfall traps using either propylene glycol or water as a preservative. Southwestern Entomologist, v. 32, n. 2, p. 87-91, 2007.

CECHIN, Sônia Zanini; MARTINS, Marcio. Eficiência de armadilhas de queda (pitfall traps) em amostragens de anfíbios e répteis no Brasil. Revista brasileira de Zoologia, v. 17, n. 3, p. 729-740, 2000.

CODDINGTON, Jonathan A. et al. Designing and testing sampling protocols to estimate biodiversity in tropical ecosystems. In: The unity of evolutionary biology: Proceedings of the Fourth International Congress of Systematic and Evolutionary Biology. Portland: Dioscorides Press, 1991. p. 44-60.

COZZUOL, Mario A. et al. A new species of tapir from the Amazon. Journal of Mammalogy, v. 94, n. 6, p. 1331-1345, 2013.

FAO/DANIDA. Guidelines for the Routine Collection of Capture Fishery Data: Prepared at the FAO/DANIDA Expert Counsultation, Bangkok, Thailand, 18-30 May 1998. Food & Agriculture Org., 1999.

GASPER, André Luís de; EISENLOHR, Pedro V.; SALINO, Alexandre. Improving collection efforts to avoid loss of biodiversity: lessons from comprehensive sampling of lycophytes and ferns in the subtropical Atlantic Forest. Acta Botanica Brasilica, v. 30, n. 2, p. 166-175, 2016.

GIUPPONI, Alessandro PL; KURY, Adriano Brilhante. A new species of Metagovea Rosas Costa, 1950 from Napo Province, Ecuador (Opiliones, Cyphophthalmi, Neogoveidae). ZooKeys, n. 477, p. 1, 2015.

GRAHAM, Catherine H. et al. New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology & Evolution, v. 19, n. 9, p. 497-503, 2004.

JIMÉNEZ-VALVERDE, Alberto; LOBO, Jorge M. Establishing reliable spider (Araneae, Araneidae and Thomisidae) assemblage sampling protocols: estimation of species richness, seasonal coverage and contribution of juvenile data to species richness and composition. Acta Oecologica, v. 30, n. 1, p. 21-32, 2006.

KRELL, Frank-T.; WHEELER, Quentin D. Specimen collection: plan for the future. Science, v. 344, n. 6186, p. 815-816, 2014.

LOURENÇO, Marta C. Are university collections and museums still meaningful? Outline of a research project. International Committee for University Museums and Collections (UMAC) Proceedings, 2010.

LOVELL, S. J. et al. Assessment of sampling approaches for a multi-taxa invertebrate survey in a South African savanna-mosaic ecosystem. Austral Ecology, v. 35, n. 4, p. 357-370, 2010.

MCFALL, Waddy F. Taxidermy step by step. Winchester Press, 1975.

MESQUITA, Paulo CMD; PASSOS, Daniel C.; CECHIN, Sonia Z. Efficiency of snake sampling methods in the Brazilian semiarid region. Anais da Academia Brasileira de Ciências, v. 85, n. 3, p. 1127-1139, 2013.

MINTEER, Ben A. et al. Avoiding (re) extinction. science, v. 344, n. 6181, p. 260-261, 2014.

MIRANDA, Flávia R. et al. Taxonomic review of the genus Cyclopes Gray, 1821 (Xenarthra: Pilosa), with the revalidation and description of new species. Zoological Journal of the Linnean Society, p. zlx079, 2017.

MORIN, Phillip A. et al. Genetic structure of the beaked whale genus Berardius in the North Pacific, with genetic evidence for a new species. Marine Mammal Science, v. 33, n. 1, p. 96-111, 2017.

NATER, Alexander et al. Morphometric, behavioral, and genomic evidence for a new Orangutan species. Current Biology, v. 27, n. 22, p. 3487-3498. e10, 2017.

OLIVEIRA, Ubirajara et al. The strong influence of collection bias on biodiversity knowledge shortfalls of Brazilian terrestrial biodiversity. Diversity and Distributions, v. 22, n. 12, p. 1232-1244, 2016.

PARDO, Iker et al. A novel method to handle the effect of uneven sampling effort in biodiversity databases. PloS one, v. 8, n. 1, p. e52786, 2013.

RIBEIRO-JÚNIOR, Marco A.; GARDNER, Toby A.; ÁVILA-PIRES, Teresa CS. The effectiveness of glue traps to sample lizards in a tropical rainforest. South American Journal of Herpetology, v. 1, n. 2, p. 131-137, 2006.

RIBEIRO-JÚNIOR, Marco A. et al. Influence of pitfall trap size and design on herpetofauna and small mammal studies in a Neotropical Forest. Zoologia (Curitiba), v. 28, n. 1, p. 80-91, 2011.

RIBEIRO-JÚNIOR, Marco A.; GARDNER, Toby A.; ÁVILA-PIRES, Teresa CS. Evaluating the effectiveness of herpetofaunal sampling techniques across a gradient of habitat change in a tropical forest landscape. Journal of Herpetology, v. 42, n. 4, p. 733-749, 2008.

ROCHA, Luiz A. et al. Specimen collection: An essential tool. Science, v. 344, n. 6186, p. 814-815, 2014.

ROWE, Rebecca J. Elevational gradient analyses and the use of historical museum specimens: a cautionary tale. Journal of Biogeography, v. 32, n. 11, p. 1883-1897, 2005.

SEREDA, Elvira et al. Assessing spider diversity on the forest floor: expert knowledge beats systematic design. Journal of Arachnology, v. 42, n. 1, p. 44-51, 2014.

SNYDER, Bruce A.; DRANEY, M. L.; SIERWALD, P. Development of an optimal sampling protocol for millipedes (Diplopoda). Journal of Insect Conservation, v. 10, n. 3, p. 277, 2006.

SOUTHWOOD, Thomas Richard Edmund; HENDERSON, Peter A. Ecological methods. John Wiley & Sons, 2009.

SZINWELSKI, Neucir et al. Ethanol fuel improves arthropod capture in pitfall traps and preserves DNA. ZooKeys, n. 196, p. 11, 2012.

WARD, Darren F.; NEW, Tim R.; YEN, Alan L. Effects of pitfall trap spacing on the abundance, richness and composition of invertebrate catches. Journal of Insect Conservation, v. 5, n. 1, p. 47-53, 2001.

WORK, Timothy T. et al. Pitfall trap size and capture of three taxa of litter-dwelling arthropods: implications for biodiversity studies. Environmental Entomology, v. 31, n. 3, p. 438-448, 2002.

Capítulo 9
Coleções Biológicas Científicas

Alessandro Rodrigues Lima & Bárbara Teixeira Faleiro

9.1 Coleções biológicas

Introdução e histórico

O hábito de colecionar objetos pode ser identificado desde a Pré-história, quando o homem deixou de ser nômade e começou a estabelecer sociedades fixas. Mas, foi somente no século XV, com o surgimento dos Gabinetes de Curiosidade na Europa (Figura 9.1), que as coleções biológicas começaram a ganhar expressão e sistematização. Os Gabinetes reuniam todo tipo de objetos considerados estranhos, raros ou valiosos, incluindo pinturas, livros, minerais, artefatos, insetos, animais taxidermizados, fósseis, plantas, etc. Durante os séculos XVI e XVII, os Gabinetes tiveram sua era áurea, impulsionados pelas grandes explorações ao Novo Mundo e à Ásia que traziam uma infinidade de objetos exóticos. No decorrer dos séculos XVIII e XIX o material de muitos Gabinetes de Curiosidade foram incorporados ao acervo de grandes Museus de História Natural que estavam surgindo, como o *Muséum National d'Histoire Naturelle* em Paris, e o *Natural History Museum* em Londres, contribuindo assim para o surgimento das coleções biológicas modernas.

Definição

A expressão **coleções biológicas** tem um uso amplo e disseminado na literatura científica, contudo, o significado atribuído à expressão não está tão explícito em muitos dos textos onde ela aparece. Na sequência apresentamos três definições usadas por importantes entidades nacionais:

1. "Conjuntos de organismos, ou partes destes, organizados de modo a fornecer informações sobre a procedência, coleta e identificação de cada um de seus espécimes" (Fundação Instituto Oswaldo Cruz - FIOCRUZ);
2. "Um conjunto de organismos fósseis ou atuais, podendo ser exemplares completos ou somente parte deles, devidamente preservados e catalogados com a finalidade de estudos didático-científico" (Sistema de Informação Brasileiro sobre a Biodiversidade - SiBBr);
3. "Coleção Biológica Científica: coleção brasileira de material biológico devidamente tratado, conservado e documentado de acordo com normas e padrões que garantam segurança, acessibilidade, qualidade, longevidade, integridade e interoperabilidade dos dados da coleção, pertencente à instituição científica com objetivo de subsidiar pesquisa científica ou tecnológica e a conservação *ex situ*" (Instruções Normativas do IBAMA nº 160, 27/04/2007 e ICMBio nº 03, 01/09/2014).

Dada a diversidade de tipos de acervos existentes (organismos microscópicos mantidos vivos em cultura, espécimes preservados em via úmida ou seca, plantas prensadas ou vivas em jardins botânicos, fósseis armazenados em gavetas, amostras biológicas armazenadas em freezer) é difícil encontrar uma definição concisa que possa englobar todos os tipos de coleções existentes. Comumente, os conceitos se aplicam a níveis mais restritos. Os conceitos 1 e 3, por exemplo, não deixam claro se uma coleção de fósseis, algo na grande maioria das vezes composto por material inorgânico, seria considerada uma coleção biológica, apesar de nenhum pesquisador contestar sua validade como coleção biológica.

Na busca de uma definição robusta, que satisfaça às ideias de uma coleção e compreenda os distintos tipos de acervo, formulamos o conceito a seguir:

> As coleções biológicas são conjuntos organizados de registros espaço-temporais da biodiversidade.

Figura 9.1 *Ritratto del Museo di Ferrante Imperato*. Extraído da obra *Dell'Historia Naturale*, Nápoles, 1599. Essa foi a primeira ilustração de um Gabinete de Curiosidades.

A coleção pode ser organizada por procedência geográfica do material colecionado, por hierarquia taxonômica ou por ordem de entrada do material na coleção. O material colecionado deve ser um registro espaço temporal da biodiversidade. As unidades da coleção podem ser organismos (vivos ou preservados; inteiros ou apenas partes), impressões deixadas por organismos (exemplo: fósseis e icnofósseis) ou produtos resultantes de suas atividades (exemplo: ninhos e registros de vocalização), desde que configurem um registro da biodiversidade atrelado a um determinado período (ou data específica) e a um local, por isso espaço-temporal.

Objetivo das coleções

Pensando no objetivo das coleções biológicas, poderíamos classificá-las em três categorias:

1. *Coleções científicas*

   As coleções científicas localizam-se em sua maioria em institutos relacionados à pesquisa, geralmente com estreita ligação com programas de pós-graduação. O material testemunho (espécimes utilizados em algum trabalho científico) de descrições taxonômicas (espécimes Tipo) merece especial atenção. Esse material fixa o nome das espécies e, por isso, sua preservação e longevidade são fundamentais. Sistematas usarão como referências esses espécimes tipos para revisões e filogenias supra-específicas, sendo que esses espécimes também serão a referência para identificação da espécie.
   As coleções científicas também são o *background* dos cientistas (ecólogos, zoólogos, botânicos) para estudar a biodiversidade de uma determinada região. Com o acesso aos dados (taxonômicos e espaço-temporais) de uma coleção é possível estudar desde a variação da diversidade local em função do tempo e das modificações do espaço (exemplo: desmatamento e mudanças climáticas), bem como usar padrões identificados no material colecionado para gerar modelos preditivos.

2. *Coleções didáticas*

   As coleções didáticas localizam-se em sua maioria em institutos relacionados ao ensino, em diferentes níveis de instrução. Seu papel principal é ilustrar a teoria, complementando o aprendizado, de forma prática. A sensibilização para a realidade além da teoria é muito válida na formação pessoal. Expandindo os horizontes, nesse sentido, os zoológicos podem ser entendidos como grandes coleções didáticas vivas, servindo tanto ao papel de ilustrar a teoria quanto à sensibilização do ser humano para com as outras espécies animais.
   O acervo dessas coleções precisa ser constantemente renovado, uma vez que seus exemplares são danificados pelo frequente manuseio por pessoas com pouca experiência. Uma fonte comum de espécimes para coleções didáticas são as coleções científicas. Material que por algum motivo (exemplo: espécimes que não possuem dados mínimos de procedência; indivíduos jovens, quando é necessário que estejam adultos para serem identificados a um nível mais específico) não podem ser incorporados às coleções científicas, mas ainda sim podem ser utilizados na didática. Outras fontes de material são zoológicos e criadouros de animais domésticos, já que geralmente espécimes provenientes de tais instituições possuem pouco ou nenhum valor em estudos taxonômicos.

3. *Coleções de reposição*

   As coleções de reposição, também chamadas de Banco de Sementes ou Banco Ativo de Germoplasma (BAG), são verdadeiras coleções de *backup*, que conservam o material genético de uso imediato, geralmente espécies com interesse para agricultura, ou o material com potencial de uso futuro. A maior coleção é a *Svalbard Global Seed Vault*, fruto de uma parceria entre diferentes países e organizações. O Brasil possui a maior coleção de germoplasma das Américas, criada e mantida pela Empresa Brasileira de Pesquisa Agropecuária (Embrapa).
   Zoológicos (quando engajados em programas conservacionistas) e Jardins Botânicos atuam como repositórios da biodiversidade *ex situ*, e são inúmeros os casos de sucesso de reintrodução de espécies extintas ou quase extintas na natureza. Como por exemplo, o mico-leão-dourado (*Leontopithecus rosalia* Kleiman et al., 1982), o cavalo de Przewalski (*Equus ferus przewalskii* Poljakov, 1881), o bisão europeu (*Bison bonasus* Linnaeus, 1758), a raposa voadora de Rodriguez (*Pteropus rodricensis* Dobson, 1878) e o falcão das ilhas Maurício (*Falco punctatus* Jones et al., 1995).

9.2 Coleções científicas

Histórico

Embora fosse costume os antigos naturalistas possuírem suas coleções particulares, em sua maioria, as coleções científicas foram atreladas a instituições de ensino e pesquisa. Dessa forma podemos datar as primeiras coleções científicas pelo surgimento dos museus. O primeiro registro de um museu no mundo ocidental é o Museu de Alexandria (o mesmo da famosa Biblioteca de Alexandria) criado por Ptolomeu Sóter no século III a.C. Apesar de muito diferente do conceito de museu moderno e o fato de ser, em essência, uma academia de filosofia, já contava com um jardim zoológico e botânico.

No século XVIII, com o aumento do interesse científico provocado pelo Renascimento, surgiram os primeiros museus modernos (Museu Britânico em 1759 e Museu do Louvre em 1793) – embora, como ressaltamos no início do capítulo, os Gabinetes de curiosidades tenham precedido em séculos essas instituições. Neste século também surgiram os primeiros museus de história natural, entre eles o mais emblemático é o *Muséum National d'Histoire Naturelle* de Paris, fundado em 1793. Mas foi no século XIX que surgiu a maioria dos grandes museus de história natural do mundo: *Museum für Naturkunde* em Berlim (1810), *American Museum of Natural History* em Nova Iorque (1869) e *Natural History Museum* em Londres (1881). Esses museus são importantes pelo elevado número de espécimes depositados, por conservarem inúmeros espécimes-tipos, além da representatividade temporal enorme, uma vez que todos foram fundados há mais de 100 anos. Essas coleções são centros de referência para milhares de trabalhos ao redor do mundo.

No Brasil, o primeiro museu foi o **Museu Nacional**, fundado em 1818 por Dom João VI no Rio de Janeiro. Seu acervo era o maior de história natural e antropologia da América Latina. Infelizmente, na noite de 2 de setembro de 2018 o prédio do museu (edifício histórico que foi residência da família real portuguesa) foi atingindo por um grande incêndio. Quase a totalidade do seu acervo histórico e científico foi destruído no incêndio. Além do pioneirismo da instituição no país, o Museu criou o primeiro periódico brasileiro dedicado a ciências naturais, em 1876, os Arquivos do Museu Nacional.

Ainda no século XIX, duas outras importantes instituições de pesquisa foram criadas: o Museu Paraense Emílio Goeldi (1871), em Belém e o Museu de Zoologia da Universidade de São Paulo (1890), em São Paulo. O Museu Paraense Emílio Goeldi, inicialmente intitulado apenas Museu Paraense, chegou a fechar as portas em 1889 devido a questões políticas e falta de verba e pesquisadores, sendo reaberto apenas em 1891. Hoje é um dos mais importantes

centros de pesquisa da Amazônia, possuindo um acervo que contempla grande parte da biodiversidade conhecida da região. O Museu de Zoologia da Universidade de São Paulo (USP) foi formado a partir da transferência da coleção zoológica do Museu Paulista (também conhecido como Museu do Ipiranga) para o Departamento de Zoologia da Secretaria da Agricultura de São Paulo. Foi somente em 1969, que a coleção zoológica do departamento foi incorporada à Universidade de São Paulo e o museu recebeu seu nome atual. Um dos diretores do Museu de Zoologia da USP, Paulo Vanzolini, foi um dos responsáveis por disseminar a nova síntese na taxonomia zoológica brasileira, e crucial na formação de toda uma geração de zoólogos no país. Hoje, o Museu de Zoologia da USP e o Museu Paraense Emílio Goeldi são importantes centros de pesquisa zoológica e são as primeiras instituições brasileiras a serem reconhecidas como fiéis depositárias pelo Conselho de Gestão do Patrimônio Genético (Deliberação CGEN nº 002, 08/07/2002).

No último ano do século XIX foi criada uma importante instituição de pesquisa: a Fundação Oswaldo Cruz (Fiocruz), fundada em 1900 no Rio de Janeiro, como Instituto Soroterápico Federal e com o objetivo de fabricar soros e vacinas contra a peste bubônica. Sua história está intimamente relacionada com o desenvolvimento da saúde pública no país, encabeçando vários eventos marcantes como a Reforma Sanitária, a Revolta da Vacina e a criação do Departamento Nacional de Saúde Pública. A Fiocruz teve à sua frente alguns dos mais célebres e brilhantes pesquisadores brasileiros, como Oswaldo Cruz, homenageado com nome da instituição em 1918, e Carlos Chagas. Atualmente a Fiocruz apoia a manutenção e salvaguarda 31 coleções, divididas em quatro grandes categorias: microbiológica, zoológica, histopatológica e botânica. A Fundação é hoje a mais importante instituição de pesquisa e desenvolvimento em saúde da América Latina, contando com 20 unidades distribuídas por todo o Brasil e uma em Maputo, Moçambique.

No século XX outras importantes instituições foram criadas, entre elas o Instituto Butantan (1901), o Instituto Nacional de Pesquisas da Amazônia (INPA; 1952), e a Fundação Zoobotânica do Rio Grande do Sul (1972), reunindo assim os maiores e mais importantes centros de pesquisa do país.

Desde o final dos anos 90, a tendência é existir coleções regionais, geralmente associadas às Universidades, como por exemplo, a Coleção de Zoologia da Universidade Federal do Mato Grosso do Sul (ZUFMS), em Campo Grande, da Universidade Federal de Minas Gerais (UFMG), em Belo Horizonte, da Universidade Federal da Paraíba (UFPB), em João Pessoa, da Universidade Federal de Rondônia (UNIR), em Porto Velho, entre outras.

9.2.1 Acervo

Obtenção

O acervo das coleções científicas ativas está em constante expansão com a adição de exemplares. Existem quatro fontes principais de obtenção de novos exemplares para as coleções: expedições de coleta, permuta, retenção e doações.

As expedições de coleta, geralmente, são as responsáveis pela obtenção da maior parte do acervo de uma coleção. Essas expedições podem ser mais restritivas, nas quais coleta-se um ou poucos grupos específicos do interesse e/ou necessários ao pesquisador responsável; ou podem ser mais abrangentes, como as coletas de levantamento e inventários da fauna, flora ou microbiota.

Permuta é uma prática de troca de material comum entre coleções. O material que é muito comum no acervo de uma coleção (geralmente espécies que ocorrem próximas ao local da coleção) pode ser extremamente raro ou inexistente em várias outras coleções. Dessa forma, é possível que duplicatas sejam trocadas entre as instituições, enriquecendo o acervo de ambas.

Retenção é quando uma instituição fica com parte ou todo o material recebido para identificação. É comum que pesquisadores de outros locais enviem material para um especialista identificar. Como parte da compensação por esse trabalho firma-se um acordo no qual parte ou mesmo todo o material é incorporado à coleção da instituição onde está o especialista. Dessa forma é possível ampliar a representatividade biogeográfica do acervo a um baixo custo financeiro, quando comparado com a obtenção por coletas.

A doação de material para uma coleção científica pode vir de diferentes fontes. O mais comum é pessoas externas coletarem ocasionalmente algum espécime e o levarem para uma instituição, em geral também com o interesse de saber a espécie e se é perigosa para o ser humano. As doações também podem ser procedentes de coleções particulares, quando o próprio colecionador ou sua família leva o material para uma instituição. Outra forma de doação é o depósito de material testemunho utilizado em trabalhos de pesquisa de áreas relacionadas (exemplo: ecologia, genética, parasitologia) ou em Estudos de Impacto Ambiental (EIA).

Conservação

As coleções científicas são (ou deveriam ser) projetadas para serem **perenes**, com duração indefinida. Sendo assim, a conservação do acervo deve visar que o material colecionado esteja em condições adequadas de uso durante o maior período de tempo possível. A conservação envolve uma série bastante complexa e específica de parâmetros, como a preparação do material colecionado, a anotação adequada das informações, o espaço físico adequado (prevendo o crescimento do acervo), material de consumo usado para manter os itens colecionados, entre outros. Cada um dos parâmetros varia de acordo com as características intrínsecas do material biológico que se tem o interesse de preservar.

A maior parte do acervo zoológico e botânico consiste em material preservado, enquanto que coleções de microrganismos são coleções vivas, com o acervo sendo periodicamente replicado para garantir sua perenidade. O acervo pode ser conservado em via seca (exemplo: plantas prensadas em exsicatas, insetos alfinetados, insetos em envelopes entomológicos, aves e mamíferos taxidermizados), em via úmida (imersas em álcool hidratado, glicerina ou soluções à base de formol) ou em lâminas de microscopia (tecidos/órgãos de animais e vegetais, ou animais microscópicos).

Cada coleção precisa lidar com uma grande quantidade de peculiaridades associadas ao material colecionado, e todas essas características precisam ser consideradas para a correta conservação do acervo. Nesse contexto se insere a figura do **curador**. É função do curador, além de gerir a coleção, informar-se sobre as práticas corretas para a preparação e preservação do material colecionado. Mesmo tratando de grupos proximamente relacionados, podem existir diferenças profundas quando o assunto é curadoria. Por isso, idealmente, o curador deve ser uma pessoa que esteja bastante familiarizada com a taxonomia do material colecionado, para garantir que o material será catalogado e corretamente referenciado para estudos futuros. A atividade de curadoria também envolve a reposição do material de consumo usado para a manutenção do acervo (exemplo: reposição de álcool, naftalina ou cânfora). Idealmente, cada coleção conta com um curador. Porém, no cenário atual, o usual é que haja um único curador para mais de uma (às vezes várias) coleção, ainda dividindo seu tempo com outras atividades (pesquisa e/ou ensino). Quando a coleção tem um acervo pequeno, uma única pessoa geralmente é capaz de realizar a função de curadoria. Mas em uma coleção minimamente modesta já é imprescindível a presença de um **técnico**. A principal função do técnico da coleção é realizar as atividades de preparação e manutenção do acervo, sob orientação do curador.

Independente das particularidades associadas à natureza do material colecionado, existem agentes deterioradores que podem danificar irreversivelmente os acervos. Os principais são a luz, o calor, a umidade e as pragas. Dependendo do tipo de preservação e do material, a suscetibilidade pode ser maior a um que a outro desses agentes, mas idealmente a coleção deve ter mecanismos para controlar todos. O local onde a coleção está armazenada também influencia na preservação de seu material. A coleção do Museu Emílio Goeldi, por exemplo, é muito mais suscetível a contaminação por fungos devido a umidade, do que a coleção do Instituto Humboldt (Colômbia), localizada em área com baixíssima pluviosidade. Dessa forma, as instalações que abrigam as coleções devem levar em consideração as características físicas e climáticas do local no planejamento, construção e manutenção das mesmas.

A incidência da radiação da luz degrada os pigmentos, levando o material gradualmente a perder sua coloração original. Mesmo a luz produzida artificialmente pode ser o suficiente para que essa degradação ocorra. A estrutura física da coleção deve ser planejada de forma que o acervo permaneça sem iluminação durante todo o tempo em que não estiver em uso, e que durante o uso ele tenha apenas a iluminação necessária. Dependendo do tamanho e da distribuição do espaço físico, é possível planejar um esquema de iluminação em setores, acionados independentemente, de acordo com a necessidade.

Calor e umidade aceleram as reações químicas envolvidas na degradação do material colecionado, além de oferecerem um ambiente propício à proliferação de pragas. Nas coleções que preservam seu material em meio líquido, o calor pode ser especialmente perigoso, pois o aumento da temperatura acelera a evaporação do líquido preservante. Caso o recipiente não tenha uma vedação adequada, o vapor escapa, podendo em casos extremos deixar o material completamente seco, prejudicando-o de forma irreversível (Figura 9.2). Quando o líquido preservante é alguma solução inflamável, como geralmente é o caso, o risco é ainda maior, porque o vapor inflamável pode saturar o ambiente da coleção aumentando os riscos de incêndio.

Consideramos como pragas de coleções aqueles organismos que comumente vivem nas coleções utilizando-as como alimento e/ou substrato, e destruindo o acervo de alguma forma. As coleções em via seca costumam sofrer mais com as pragas, sendo as mais comuns os fungos e alguns grupos de insetos, em especial os piolho-de-livros (Psocoptera), os besouros-da-dispensa (família Dermestidae/Coleoptera), formigas (Hymenoptera), baratas (Blatodea) e outros. Porém, dependendo das condições do espaço da coleção, é possível também ter infestação de vertebrados, como ratos (Roedores) e pombos (Columbiformes). Tanto quanto possível, as salas da coleção devem ser hermeticamente isoladas do ambiente externo, com um controle rigoroso do fluxo de pessoas e do acervo. Deve haver um protocolo de limpeza e inspeção de todo material que chega para a coleção (sendo novo ou retorno de algum empréstimo), para minimizar o risco de que alguma praga presente nesse material contamine o resto do acervo.

Figura 9.2 À esquerda, frasco de vidro que armazenava insetos em via úmida. A vedação inadequada levou ao ressecamento irreversível do material (adultos de Ephemeroptera), mostrado em maior aumento à direita. Fotos: Alessandro R. Lima e Bárbara T. Faleiro

Abrangência

O acervo de uma coleção pode ser classificado em duas categorias em relação a sua abrangência: geral (global) ou local (regional).

Os acervos gerais possuem exemplares que representam a biodiversidade de grandes regiões do mundo, abrangendo vários biomas. Esses acervos são geralmente encontrados em grandes instituições de ensino e pesquisa: Museus Nacionais de história natural, Universidades federais, e grandes Herbários (exemplo: *Muséum National d'Histoire naturelle* em Paris e Museu Nacional no Rio de Janeiro). Em sua maioria, essas coleções são antigas, possuindo dessa forma uma grande representatividade histórica, tanto por reunir espécimes coletados há muitos anos como material estudado por grandes pesquisadores (naturalistas) do passado. Além disso, os acervos gerais tendem a possuir grande quantidade de material tipo.

Os acervos regionais apresentam uma representatividade geográfica menor. Geralmente abrangendo regiões e biomas ao redor da sede da coleção. São encontrados normalmente em Museus de história natural estaduais ou regionais, Universidades estaduais ou privadas, e centros de pesquisa (exemplo: Museu de Ciências e Tecnologia da Pontifícia Universidade Católica do Rio Grande do Sul). Esses acervos, geralmente, contam com uma representatividade da biodiversidade de uma determinada região, uma vez que possuem grandes séries coletadas ao longo de vários anos. Diante disso, é comum encontrar várias espécies endêmicas nessas coleções.

Essa classificação é baseada nas características da maior parte do acervo. Ou seja, não significa que em um acervo regional não existam exemplares de outras regiões. Mesmo as coleções globais, no geral sempre têm uma ou algumas regiões mais bem representadas que outras. Essa classificação não implica uma hierarquia de importância. As coleções regionais e globais se complementam para o conhecimento da biodiversidade. Dependendo do tipo de pergunta que uma pesquisa tenta responder, uma coleção pequena e regional pode oferecer melhor conjunto de dados que uma grande e antiga coleção mundial.

9.2.2 Metadados

Como apresentado acima, consideramos que cada unidade em uma coleção biológica representa um **registro espaço-temporal da biodiversidade**. Aqui é necessário entender outro conceito relacionado ao colecionismo, o **metadado**. De modo sumário, metadados são **dados sobre dados**. Para as coleções científicas, metadados podem ser entendidos como qualquer informação atrelada/associada ao material colecionado.

Se cada unidade de uma coleção representa um registro espaço-temporal, informações de procedência (local de origem e data) constituem o mínimo (essencial) de metadados que se espera encontrar. Esses metadados são obtidos no ato da coleta, e devem ser devidamente anotados pelo coletor para que sejam associados à unidade colecionada por meio de uma etiqueta de procedência. Há quem argumente que uma coleção biológica sem essas informações não tem valor para a ciência.

Os metadados associados à taxonomia geralmente são adicionados a posteriori, após exame do material pelos taxonomistas. Diferente dos metadados de procedência, as informações taxonômicas não são dados concretos obtidos na coleta. Tanto a identificação taxonômica quanto a classificação são hipóteses propostas por taxonomistas e sistematas. Dessa forma, caso haja uma mudança na proposta de classificação, ou caso um taxonomista tenha uma opinião diferente sobre a identificação, essas informações precisam ser atualizadas na base de dados. Porém, uma informação não pode simplesmente substituir a anterior, sendo fundamental que haja um histórico dessas diferentes propostas.

Tradicionalmente os metadados ficavam apenas junto das unidades da coleção, por meio das etiquetas. Nas coleções mais organizadas, as informações também eram compiladas em livros de tombo (Figura 9.3). No cenário

atual, com o acesso a computadores de baixo custo e boa capacidade de processamento, os livros de tombo tornaram-se obsoletos. Os bancos de dados são a melhor forma de armazenar e organizar os metadados da coleção. Em linhas gerais, bancos de dados são formados por planilhas eletrônicas relacionadas entre si e associadas às unidades da coleção. As planilhas são bem mais eficientes para acomodarem os metadados, praticamente sem restrições de quantidade ou formato das informações (nos livros de tombo e etiquetas é bastante comum encontrar abreviações e siglas devido à restrição de espaço. Outra grande vantagem das planilhas está relacionada à legibilidade das informações. Todo taxonomista que trabalhou em acervos antigos já teve problemas para interpretar as informações em etiquetas e/ou livros de tombo, os quais eram sempre preenchidos à mão. Isso sem falar dos problemas decorrentes do material usado para escrever: grafite que se apaga com tempo, tintas solúveis em água, etc. Usando um editor de textos é possível fazer um *script* para gerar etiquetas com formato padrão, contendo toda a informação básica de forma legível, impressa com *toner*, insolúvel em água, álcool ou outro líquido conservador (Figura 9.4). Uma outra característica das planilhas que modificou completamente o acesso às informações do banco de dados foram as ferramentas de busca e filtros. Imagina como seria ter que pesquisar nos livros de tombo ou procurar em todo o acervo todas os espécimes de um determinado gênero coletados em uma certa região do mundo. Em uma coleção de pequeno tamanho essa pesquisa já demandaria um tempo gigantesco, tornando inviável esse trabalho.

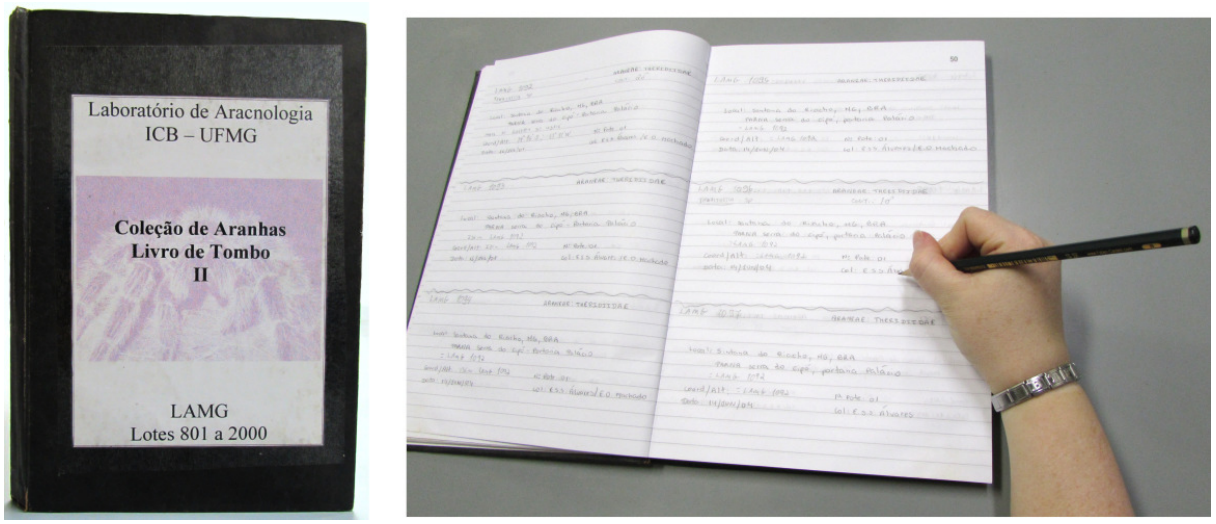


Figura 9.3 Exemplo de preenchimento (à direita) do livro tombo utilizado na coleção de aranhas da Universidade Federal de Minas Gerais (à esquerda). Fotos: Alessandro R. Lima e Bárbara T. Faleiro

Os bancos de dados podem ser facilmente armazenados *online*, permitindo acesso remoto e fácil intercâmbio desses metadados entre pesquisadores e instituições, por todo o mundo. Em 2001 uma organização internacional foi formada, a *Global Biodiversity Information Facility* (GBIF), com o objetivo de ser uma infraestrutura de pesquisa de dados sobre a biodiversidade mundial, financiada por diferentes governos, de acesso livre. Desde sua formação, cada vez mais instituições ao redor do globo têm aderido à iniciativa, alimentando o sistema com suas bases de dados. No Brasil, temos duas plataformas que alimentam o GBIF, o *speciesLink*, projeto desenvolvido pelo Centro de Referência em Informação Ambiental (CRIA), e o Sistema de Informação sobre a Biodiversidade Brasileira (SiBBr), iniciativa do governo brasileiro.

Na prática, cada coleção pode estruturar sua base de dados da forma mais adequada ao material colecionado. Mas é preciso levar em consideração que um banco de dados precisa atender a uma série de parâmetros lógicos para que seja uma base sólida e eficiente (não ambiguidade, chaves primárias, etc.). Idealmente, a construção da estrutura deveria contar com uma consultoria de alguém com conhecimento na área. Para facilitar a compreensão/comparação de informações sobre a biodiversidade em nível internacional, estabeleceu-se em 2009 um conjunto de normas, incluindo glossário de termos (propriedades, elementos, campos, colunas, atributos ou conceitos), que foi denominado *Darwin Core*. Esse é o padrão utilizado pelo GBIF, por exemplo. Esse padrão serve como uma base para construção de bases de dados de coleções científicas.

O banco de dados pode ser manipulado diretamente (edição direta das informações contidas nas células da planilha) ou indiretamente, usando um programa gerenciador. O programa oferece uma interface diferente da planilha, geralmente com caixas de diálogos. As informações são preenchidas nessa interface e o programa é quem se encarrega de gravar as informações na planilha. O uso de gerenciadores tende a reduzir os erros no banco de dados. Existem várias opções de gerenciadores de acesso gratuito, cada qual otimizado para um tipo de coleção. Uma grande vantagem do gerenciador é que ele permite automatizar facilmente tarefas relacionados ao gerenciamento do acervo, como por exemplo o controle e emissão de guias de remessa de empréstimos.

9.2.3 Coleções Particulares *x* Institucionais

Historicamente grande parte das coleções científicas brasileiras poderiam ser classificadas como coleções particulares. Isso vale para coleções hospedadas em instituições públicas também. Essas coleções são/foram formadas pelo esforço individual (algumas vezes, coletivo) de um pesquisador visando o seu grupo de estudo/interesse. Não são raros os casos de pesquisadores que investiram o próprio dinheiro para a criação e manutenção de robustas coleções taxonômicas, e que foram praticamente dizimadas pelo abandono assim que o pesquisador se afastou do seu comando. Existem inclusive casos de coleções que foram vendidas pelos familiares após a morte do pesquisador. Não basta que uma coleção esteja dentro de um instituto para que ela seja considerada institucionalizada. Todas essas situações ocorrem porque, mesmo que o material estivesse dentro de um instituto, a coleção não era institucionalizada.

O objetivo de institucionalizar uma coleção é garantir que ela seja reconhecida como parte do patrimônio da Instituição, que assume compromisso para a sua manutenção. A coleção institucional pode pleitear verba por meio de projetos e demandar profissionais técnicos, junto ao instituto, para que seja feita sua devida curadoria. Contar com suporte para possíveis questões burocráticas junto aos órgãos ambientais e para possíveis disputas geradas por movimentação do patrimônio (empréstimos) são importantes para as atividades de uma coleção. O status da coleção também pesa no reconhecimento das atividades relacionadas a ela. Se uma coleção é particular, o trabalho do curador pode ser considerado apenas como uma atividade de interesse próprio. Quando há institucionalização, o curador pode requerer que parte da sua jornada de trabalho esteja destinada às atividades da coleção.

Centro de Coleções Taxonômicas da UFMG (CCT-UFMG)

O Centro de Coleções Taxonômicas (CCT) é o resultado de um trabalho de quase 20 anos para a institucionalização das coleções taxonômicas depositadas no Instituto de Ciências Biológicas (ICB) da UFMG. Após uma reunião em 1997, pesquisadores dos departamentos de Biologia Geral, Botânica, Parasitologia e Zoologia, apresentaram à instituição um relatório das dimensões e da situação das diversas coleções existentes. Como resultado, o CCT foi reconhecido pela Universidade em junho de 2015 como um Órgão Complementar do ICB.

Conforme informações do site oficial https://www2.icb.ufmg.br/cct/, O CCT-UFMG têm hoje 25 coleções, com um leque taxonômico bastante abrangente, incluindo o herbário do Departamento de Botânica; a coleção de fungos do Departamento de Microbiologia; as coleções de moléculas e tecidos dos departamentos de Microbiologia, Biologia Geral e Zoologia; e as coleções zoológicas, dispersas nos departamentos de Biologia Geral, Parasitologia e Zoologia. De forma geral, as coleções do CCT têm amostragem mais extensiva e importante em

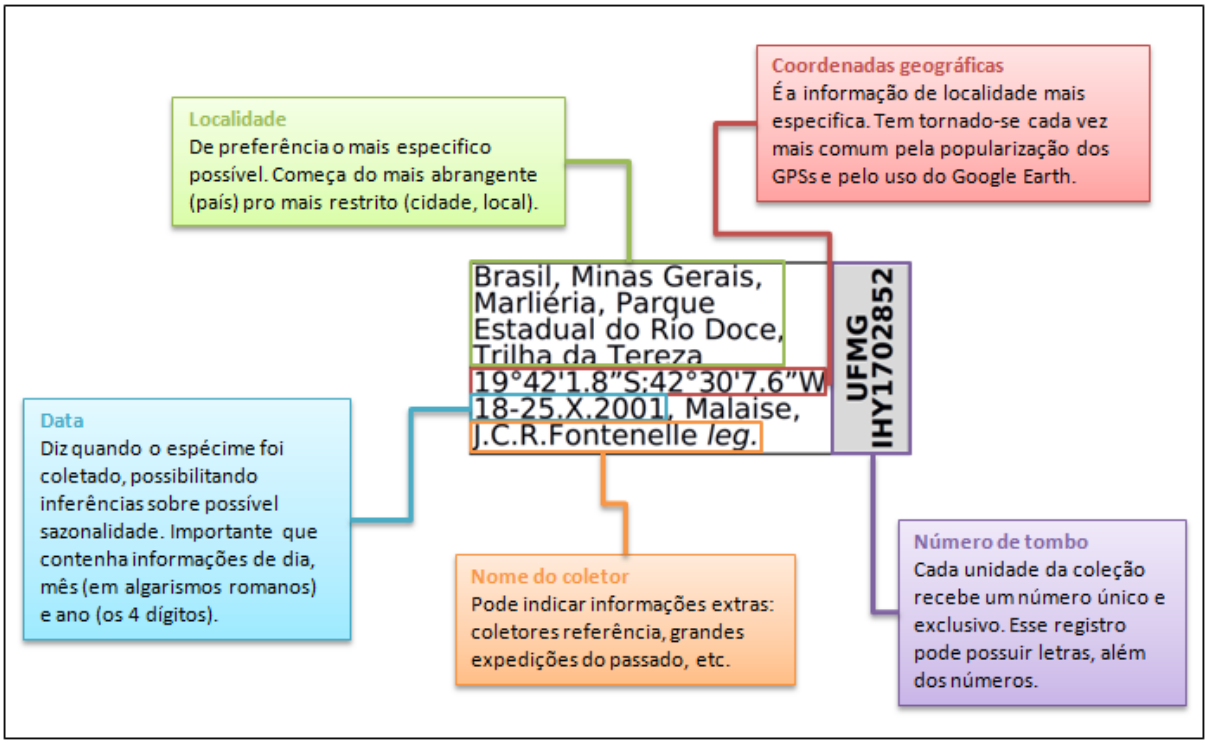


Figura 9.4 Infográfico com os detalhes das informações contidas na etiqueta de procedência da coleção de insetos CCT-UFMG. A etiqueta foi gerada automaticamente a partir da planilha de metadados. Fotos: Alessandro R. Lima e Bárbara T. Faleiro

Minas Gerais e região Sudeste, com exemplares de coletas que datam do início do século XX, em áreas sensíveis, que hoje estão muito degradadas, como o vale do Rio Doce.

O CCT é administrado por um Conselho Técnico-Científico e um Corpo de Curadores, com apoio de funcionários Técnico-Administrativos da Universidade. Ademais, o CCT possui um Laboratório de Ilustração Científica (LIC) e é responsável pela publicação do periódico científico *LUNDIANA: International Journal of Biodiversity*.

9.3 Bibliografia recomendada

Biodiversity Information Standards, Padrão Darwin Core http://rs.tdwg.org/dwc/

GBIF, Global Biodiversity Information Facility https://www.gbif.org/

PAPAVERO, Nelson. Fundamentos práticos de taxonomia zoológica. Unesp, 1994.

SILVEIRA, Fernando A.; ALVARENGA, Alessandra S. O acervo de abelhas da Coleção entomológica das coleções Taxonômicas da UFMG. MG BIOTA, v. 4, p. 5-24, 2012.

SIMMONS, John E & MUÑOZ-SABA, Yaneth (eds.). Cuidado, manejo y conservación de las colecciones biológicas.Universidad Nacional de Colombia, D.C. 286 pp., 2005.

Sistema de Informação sobre a Biodiversidade Brasileira http://www.sibbr.gov.br/

SpeciesLink, Centro de Referência em Informação Ambiental, CRIA http://splink.cria.org.br/

SUAREZ, Andrew V.; TSUTSUI, Neil D. The value of museum collections for research and society. AIBS Bulletin, v. 54, n. 1, p. 66-74, 2004.