Análises evolutivas baseadas em dados moleculares e morfológicos: exemplos da herpetofauna brasileira

Felipe Gobbi Grazziotin[1] & Ronaldo Fernandes[2]

Resumo

Abordagens moleculares e probabilísticas, enfocando a evolução de elementos da herpetofauna, têm ocupado cada vez mais as páginas de importantes periódicos científicos internacionais. Até o momento, poucos destes trabalhos envolvendo espécies brasileiras foram realizados no Brasil, sendo a maioria desenvolvida no exterior. Essas abordagens possibilitam inferências que não podem ser acessadas por outros métodos e dizem respeito às relações microevolutivas entre espécies ou populações, podendo indicar processos estocásticos e seletivos. O tamanho ancestral da população ou da espécie, a estruturação genética e magnitude do fluxo gênico entre áreas, o tempo de ancestralidade, entre outros parâmetros genéticos, podem ser estimados e contrastados com árvores filogenéticas e redes de haplótipos, permitindo um estudo mais completo das histórias evolutivas. A interação entre herpetólogos e laboratórios de biologia molecular seria um importante avanço para a formação de grupos de pesquisa que utilizem essas abordagens. As técnicas que possibilitam análises integradas de dados representam uma ferramenta importante para a realização dessa promissora interação.

Palavras-chave: Evolução molecular, Filogenia, Filogeografia, Caracteres contínuos.

Abstract

Molecular and probabilistic approaches focusing on the evolution of the herpetological fauna have been occupying an increasing space in important scientific journals. Until now, few of these studies addressing Brazilian taxa have been developed in Brazilian laboratories. These approaches allow inferences that cannot be accessed by other methods regarding the microevolutionary relationships among species or populations, and may reveal stochastic as well as selective processes. The size of the ancestral population or species, the genetic structure and magnitude of gene flow among areas, the time to the most recent common ancestor, and other genetic parameters can be estimated and contrasted to phylogenetic trees and haplotype networks, leading to a more comprehensive study of the group's evolutionary history. The interaction between herpetologists and molecular biology laboratories would be an important advance for the establishment of research groups that use these approaches. Methods that facilitate analyses of integrated data represent an important tool in this promising interaction.

Key-words: Molecular evolution, Phylogeny, Phylogeography, Continuous characters.

Aceito em agosto de 2006.

[1] Centro de Biologia Genômica e Molecular, PUCRS, Avenida Ipiranga 6681, Porto Alegre, RS, 90619-900. felipe.grazziotin@pucrs.br
[2] Departamento de Vertebrados, Museu Nacional, UFRJ, Quinta da Boa Vista s/n, Rio de Janeiro, RJ, 20940-040. ronnie@acd.ufrj.br

Introdução

O uso de uma abordagem molecular e probabilística no estudo da evolução da biodiversidade tem se tornado uma constante nas publicações atuais. Muitos herpetólogos já incorporam estes conhecimentos e discutem com propriedade os benefícios e as falhas dos mesmos, assim como existem outros que, apesar da falta de conhecimento básico, demonstram muito interesse sobre o assunto. Este capítulo é direcionado àqueles herpetólogos que buscam suprir as lacunas básicas neste tipo de abordagem e ao mesmo tempo mostrar a interação que pode ocorrer entre as abordagens morfológica e molecular, ressaltando que apesar de algumas diferenças teóricas e metodológicas a abordagem molecular apresenta-se como ferramenta auxiliar no estudo da evolução da herpetofauna.

Conceitos básicos em genética molecular

Optamos por iniciar esta revisão com um sucinto resumo sobre genética molecular básica, com o intuito de recordar conceitos básicos que auxiliam na compreensão de teorias mais complexas. Sugerimos a leitura de GRAUR & LI (2000) para uma revisão mais ampla sobre esse assunto. Já para aqueles que possuem esse conhecimento básico sugerimos começar a leitura pelo tópico "Dinâmica do processo evolutivo".

A grande maioria da variabilidade de formas (fenótipos) que encontramos entre os organismos possui uma base genética. As características que podem ser usadas em estudos evolutivos necessitam impreterivelmente desta base, pois somente os caracteres herdáveis podem reconstruir a história evolutiva. Apesar dessa base genética, o fenótipo de um organismo é sempre um produto da interação dos processos gênicos com o ambiente e as características com significado evolutivo são sempre herdáveis.

O genoma de um organismo representa todo o material hereditário que se encontra contido no núcleo e no citoplasma (mitocôndria e cloroplastos). O genoma pode ser dividido em uma porção que possui genes e uma porção que não os possui, sendo que nessa última encontram-se todos aqueles elementos genômicos cuja função biológica ainda não é bem conhecida (inserções *Alu*, microssatélites, espaçadores gênicos, introns, etc.). Esses elementos genômicos muitas vezes representam a maior porção do genoma (como é o caso dos vertebrados em geral).

Os genes ou grupo de genes são expressos em proteínas e interagem com as mais diversas substâncias presentes no organismo, gerando as estruturas morfológicas que por séculos foram e ainda são estudadas por aqueles que classificam a diversidade orgânica. Um gene é tradicionalmente definido como um segmento de DNA (ácido desoxirribonucléico) que é traduzido em uma cadeia polipeptídica ou é transcrito para uma molécula funcional de RNA (ácido ribonucléico). O DNA se estrutura através de duas fitas antiparalelas de nucleotídeos, os quais são compostos de uma base nitrogenada, uma pentose e um grupo fosfato. Essas bases são divididas em dois tipos: as purinas, adenina (A) e guanina (G); e as pirimidinas, citosina (C) e timina (T).

A dinâmica da expressão gênica inicia quando a informação genética do DNA é transcrita em RNA mensageiro (mRNA). Esse atua como molde para a síntese protéica e junto com o RNA

ribossômico (rRNA) e as proteínas ribossômicas constituem a maquinaria de tradução, a qual é capaz de ordenar a ligação de RNAs transportadores (tRNA) que conduzem os aminoácidos (AA) específicos e os polimeriza em polipeptídeos que formarão as proteínas do organismo. Essa tradução é baseada em um código onde cada três nucleotídeos do mRNA (códon) interagem com três nucleotídeos do tRNA (anticódon). Cada tRNA está associado a um único aminoácido, entretanto mais de um tRNA pode conduzir o mesmo aminoácido, como indica a Tabela 1, pois os quatro nucleotídeos formam 64 combinações três-a-três, enquanto os AAs mais comuns são apenas 20, existindo uma sobreposição de códons chamada degenerescência do código. O código genético é universal e quase a totalidade dos organismos compartilham esse mesmo código, contudo existem exceções como na mitocôndria, alguns protozoários e algumas bactérias.

Tabela 1: Código genético universal

Códon	Aminoácido	Códon	Aminoácido	Códon	Aminoácido	Códon	Aminoácido
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Ter	UGA	Ter
UUG	Leu	UCG	Ser	UAG	Ter	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

Aminoácidos representados por seu código de três letras. Ter códon de terminação.

Durante a replicação do DNA, nucleotídeos errados podem ser incorporados pelas enzimas de replicação. Da mesma forma, fatores ambientais como luz ultravioleta e substâncias químicas podem induzir mudanças na seqüência de DNA. Estas mudanças são chamadas de mutações e podem ser de dois tipos: troca de nucleotídeos e ganho ou perda de nucleotídeos.

Mesmo que uma mutação ocorra, devemos ter a noção que isso não significa que essa mutação terá importância evolutiva, pois essa base alterada pode ser reparada, ocorrer em células somáticas, o organismo que a possuir pode não ser viável (mutação deletéria) ou simplesmente não deixar descendentes. As mutações têm real importância evolutiva no momento que são passadas para a

próxima geração. Quando isso ocorre, as mutações de troca passam a ser chamadas de substituições e podem se apresentar de duas formas: transições, quando o tipo do nucleotídeo é mantido (troca de purina por purina ou pirimidina por pirimidina); e transversões quando a substituição ocorre entre tipos distintos (purina por pirimidina e vice-versa). Como podemos ver na Figura 1 as transversões, observando apenas as possibilidades de troca, deveriam ser mais freqüentes que as transições. Entretanto elas causam um impacto muito maior na expressão protéica (ver Tabela 1), por isso ocorrem geralmente com uma freqüência menor se comparadas com as transições. As substituições podem ser classificadas também quanto o seu resultado na expressão. Substituições que mudam os AAs são chamadas de não-sinônimas, enquanto as que mantêm os AAs originais são denominadas de sinônimas.

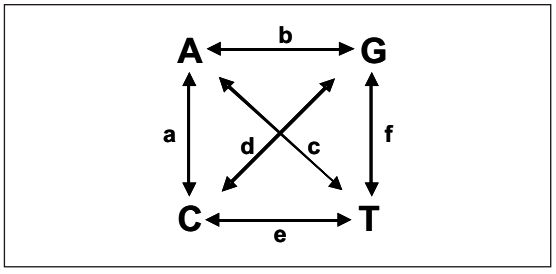

Figura 1: Representação gráfica das possibilidades de substituição nucleotídica. b, e representam transições; a, c, d, f representam transversões.

As mutações de ganho ou perda de nucleotídeos são denominadas de inserções e deleções (*indels*). Este tipo de mutação é freqüentemente causador de erros de expressão, pois podem modificar todo o quadro de leitura da proteína levando a um produto não funcional.

A mutação que não foi corrigida e passou para a próxima geração criou uma forma alternativa para o loco gênico original denominada de alelo. Um gene pode ter várias formas alélicas dependendo de quantos polimorfismos diferentes existem segregando na população. Entretanto, um indivíduo diplóide (dois conjuntos cromossômicos) pode ter apenas dois alelos. Se esses forem iguais, o organismo é definido como homozigoto para o gene em questão, enquanto que se ele apresentar dois alelos distintos para o mesmo gene será definido como heterozigoto. Quando várias mutações são herdadas em conjunto, com baixa ou nenhuma recombinação podemos denominá-las de haplótipos.

Existem regiões gênicas que estão sobre intensa pressão, onde qualquer mudança na estrutura do produto gênico ocasiona uma desvantagem para o organismo. O exemplo mais conhecido são as histonas, proteínas que se encontram associadas à fita de DNA e são envolvidas na sua compactação, formando os cromossomos. A variabilidade no gene que codifica as histonas é tão pequena que quase todos os organismos vivos possuem a mesma seqüência. Por outro lado, regiões de microssatélites, que se constituem de pequenas seqüências repetidas não codificantes ao longo do genoma, alteram o número de suas repetições com grande rapidez que podem ser detectadas em intervalos tão curtos quanto uma geração.

Buscando polimorfismos: as técnicas moleculares

Existem diversas técnicas moleculares para encontrar variabilidade genética. Estas técnicas podem ser separadas em três grandes grupos: análises químicas, análises de proteínas e

análises de DNA. Atualmente as análises de DNA são as técnicas mais utilizadas na qual ressaltamos: os polimorfismos de restrição (RFLP), os polimorfismos de segmentos amplificados (PCR, RAPD, AFLP, microssatélites) e o seqüenciamento de fragmentos específicos (ver HILLIS *et al*., 1996).

Cada uma destas metodologias é adequada ao estudo de determinados sistemas. Assim, os estudos filogenéticos baseiam-se principalmente no seqüenciamento, enquanto que os estudos de diversidade populacional requerem análises do tipo microssatélites ou seqüenciamento. Podemos ainda classificar as técnicas de acordo com o tipo de herança, esforço de interpretação, entre outras. Uma visualização das diferenças entre alguns marcadores encontra-se na Tabela 2.

Tabela 2: Comparação entre algumas técnicas moleculares.

Abordagem	Esforço	Custo	Herança	Uso	Interpretação	Maior Problema
AFLP	médio	médio	dominante	populacional	média	Falta de homologia, dominância
Aloenzimas	baixo	baixo	codominante	populacional	difícil	Necessidade de material fresco
Microssatélites	alto	médio	codominante	populacional	média	Construção dos iniciadores para PCR
PCR/RFLP	baixo	baixo	variável	populacional	média	Escolha do gene
RAPD	baixo	baixo	dominante	populacional	difícil	Falta de homologia, dominância e reprodutibilidade
Sequenciamento	baixo	alto	variável	populacional filogenético	fácil	Análise dos resultados

Esforço: esforço para a adequação da técnica ao projeto; **Custo**: investimento necessário para realizar a técnica; **Herança**: tipo de herança observável com o marcador, se codominante os heterozigotos podem ser identificados se dominante não; **Uso**: tipo de estudo mais adequado para utilização da técnica; **Interpretação**: facilidade de interpretar geneticamente os dados. Adaptada de SOLé-CAVA (2001).

A técnica de PCR, amplamente difundida, consiste na amplificação específica de uma região qualquer do genoma, para a qual são conhecidas pequenas porções das regiões que flanqueiam essa seqüência de interesse. Para essas regiões são sintetizadas pequenas seqüências complementares de poucos nucleotídeos (aproximadamente 20), chamadas *primers* ou iniciadores. Em um sistema *in vitro* são misturados: o DNA do organismo alvo, os iniciadores, os quatro nucleotídeos em igual proporção, uma enzima polimerizadora e co-fatores enzimáticos. Esse sistema é submetido a ciclos de elevação e redução de temperatura onde as fitas paralelas do DNA alvo são desnaturadas (separam-se), os iniciadores pareiam com suas regiões complementares nos flancos da seqüência desejada e a enzima sintetiza novas fitas de DNA amplificando a região de interesse. Esse processo é repetido por vários ciclos e gera uma grande quantidade de seqüências através de um aumento exponencial, pois para cada molécula da seqüência alvo na suspensão de DNA adicionada ao sistema, existirá 2^n amplicons ou produtos de PCR (onde n representa o número de ciclos).

Todas as técnicas da Tabela 2 visam em última análise obter informações da seqüência de DNA de uma dada região. O sequenciamento de fragmentos específicos do DNA é dentre elas a técnica que maximiza essa informação. Atualmente o sequenciamento é realizado em sofisticados aparelhos automáticos que minimizam o esforço humano. A técnica utilizada é baseada naquela

desenvolvida por SANGER *et al.* (1977), a qual adiciona, além dos nucleotídeos (desoxirnbonucleotídeos), uma certa proporção de nucleotídeos modificados (dideoxiribonucleotídeos) que cessam a polimerização das seqüências de DNA por não possuírem o grupo hidroxila na posição 3' da pentose. Devido à grande quantidade de fragmentos amplificados, todas as possibilidades de parada da polimerização em uma seqüência são encontradas depois da reação. Esses fragmentos são submetidos a uma eletroforese e, devido aos seus diferentes tamanhos, eles podem ser identificados e a seqüência pode ser completamente decifrada.

Dinâmica do processo evolutivo

O desafio para quem trabalha com a biodiversidade é tentar compreender os processos históricos que a moldaram, além de descrever as formas existentes na natureza, sua fisiologia ou seu relacionamento. A evolução por detrás da história dos organismos é o processo que pode ser interpretado e extrapolado. Os eventos históricos que moldaram a diversidade são quase sempre únicos e exclusivos de cada organismo, são idiossincráticos e só podem ser descritos. Infelizmente, os conceitos chave da evolução são muitas vezes mal compreendidos devido ao distanciamento atingido pelos diversos tipos de especialistas que trabalham com o tema. Entretanto, em última análise, tanto para aqueles que trabalham com macroevolução (sistemática) como para os que trabalham com microevolução (genética de populações), a evolução pode ser pensada como uma alteração na freqüência de formas gênicas (não estamos reduzindo a evolução apenas a essa alteração, mas sim indicando uma das diversas formas de pensá-la). Esses dois campos do estudo da evolução encontram-se, nos dias de hoje, muito mais ligados do que no passado. Sua junção possui, desde 1987, a denominação de filogeografia, nome dado por John Avise (AVISE *et al*., 1987). Os estudos filogeográficos têm crescido exponencialmente na literatura científica demonstrando que é sempre preferível uma abordagem que englobe os dois campos (micro e macroevolução) para um melhor entendimento de qualquer problema evolutivo (novas espécies, re-descrição de táxons, padrões de variabilidade genética, etc.). Os processos intraespecíficos que afetam os organismos não podem ser esquecidos quando trabalhamos com filogenias interespecíficas, sendo o contrário também válido. A formação da biodiversidade é complexa e a evolução não se processa de maneira particionada sem interação entre os níveis de variabilidade.

Atualmente, existem duas abordagens diferenciadas para inferir os processos evolutivos em espécies ou populações através da filogeografia. Uma, mais clássica, utiliza quase que exclusivamente árvores filogenéticas ou outras formas gráficas em conjunto com conhecimentos biogeográficos. Outra apresenta um arcabouço muito mais estatístico, utiliza estimativas de parâmetros microevolutivos fundamentalmente matemáticos possibilitando até mesmo a simulação computacional de distintos cenários evolutivos. Assim, como outros dilemas metodológicos, a utilização de árvores ou estatísticas em análises moleculares não representa uma real dicotomia de abordagem. Ambos os métodos podem e devem ser utilizados, pois para a compreensão do processo

evolutivo a variabilidade deve ser acessada de todas as formas possíveis. Portanto, esses métodos serão abordados aqui de forma integrada.

O conceito de variabilidade genética ou polimorfismo está intimamente ligado ao conceito moderno de evolução e essa variabilidade pode ser estimada de várias formas. A proporção relativa de cada um dos alelos em uma população é o que denominamos de freqüência alélica e é alterada principalmente por três forças evolutivas: seleção natural, deriva genética e migração.

Seleção natural

Todo organismo vivo é moldado pela seleção natural, que está agindo sobre o fenótipo do indivíduo e quase sempre selecionando características do DNA. São os genes que sofrem mutações, duplicações, transversões e desta forma geram novas estruturas e funções. O genoma de uma espécie é moldado pela seleção, um processo altamente direcionado que leva ao aumento da adaptabilidade (*fitness*). A seleção se processa pela reprodução diferencial de indivíduos com fenótipos distintos e pode ser dividida em: seleção positiva, quando um alelo particular possui maior adaptabilidade em um determinado ambiente; seleção negativa ou purificadora, quando os polimorfismos reduzem a adaptabilidade do organismo e são removidos da população; e seleção balanceadora que ocorre quando o alelo em heterozigose torna-se vantajoso e aumenta a adaptabilidade.

A deriva genética

Outro processo que altera as freqüências alélicas é a deriva genética. Este processo é totalmente aleatório e representa a chance dos alelos de uma geração estarem presentes na próxima, ou seja, é o processo estocástico de amostragem de alelos em uma população no momento da reprodução. Isso fica claro se pensarmos que em cada geração existe um número muito maior de alelos na população do que indivíduos que vão contribuir para a próxima geração. Sendo assim, muitos alelos serão perdidos, enquanto outros serão fixados, isto é, atingirão 100% de freqüência, pelo jogo de azar da reprodução (Fig.2).

O parâmetro que determina o impacto da deriva genética é o tamanho efetivo da população (N_e). Diferentemente do censo populacional, o N_e representa o número de indivíduos em uma população teórica ideal (com cruzamentos ao acaso, gerações discretas, tamanho populacional constante, igual proporção sexual e variação do tamanho da prole seguindo uma distribuição de Poisson), população essa que possui a mesma magnitude de mudança na média do coeficiente de endocruzamento causado pela deriva genética (perda de diversidade genética) atuante na população atual (WRIGHT, 1931). Isto quer dizer que o N_e representa o número de indivíduos que efetivamente se reproduzem em uma população real em relação a uma população ideal. O N_e pode ser histórico, ou seja, relativo à população ancestral, o qual comparado com o censo atual pode dizer muito sobre o que ocorreu com o organismo que está sendo estudado. As flutuações do N_e representam períodos de expansão ou redução populacional (Fig.2). Toda a vez que uma população reduz seu tamanho a deriva genética tem grandes efeitos sobre a fixação ou eliminação de alelos. Muitas vezes pensamos somente em explicações adaptativas (frutos de

seleção natural), enquanto muito da evolução do organismo pode ser explicada por eventos demográficos e de deriva (e.g. efeito fundador, onde poucos indivíduos, que representam uma parcela muito pequena da diversidade de uma população fundam outra, como nas colonizações de ilhas, e em poucos anos podem divergir enormemente das freqüências alélicas da população original).

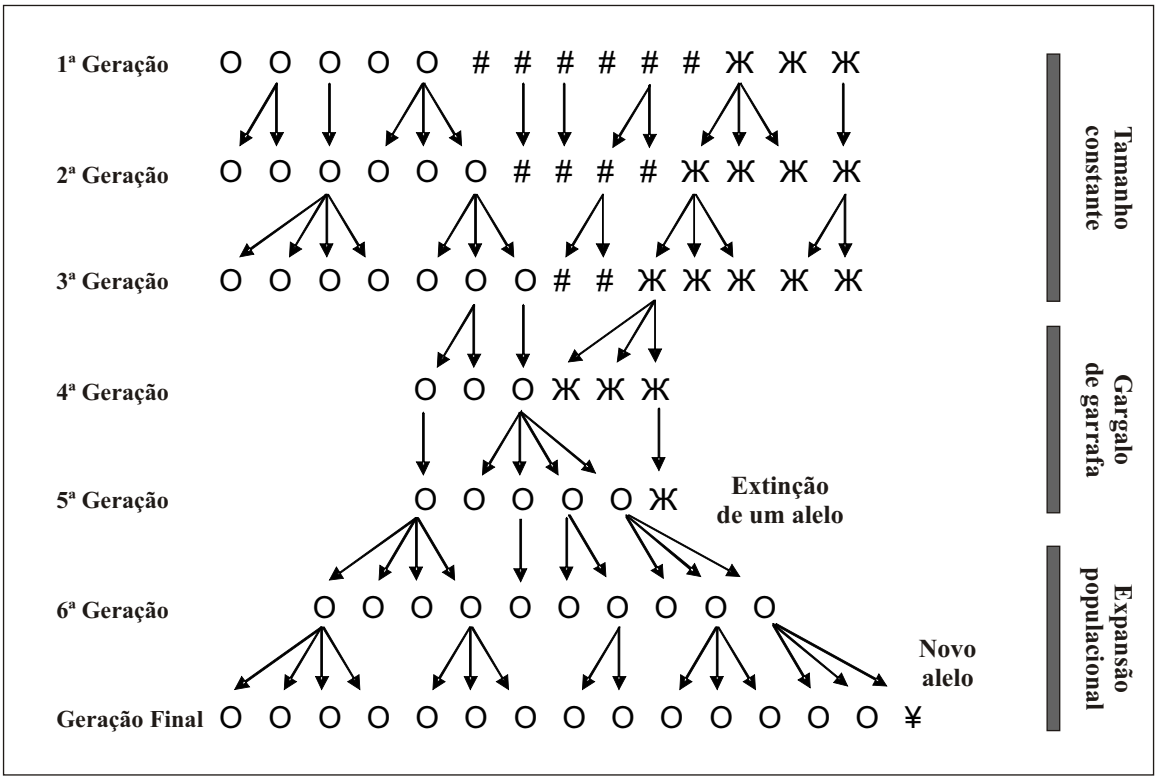

Figura 2: Representação gráfica do processo de deriva genética (baseado em VANDAMME, 2002)

Dentro desse contexto o parâmetro θ é freqüentemente estimado dos dados e utilizado em diversas análises genéticas (fluxo migratório, neutralidade de seqüências, definição do N_e, etc.). O θ representa a relação $4N_e\mu$ onde μ é a taxa de mutação por geração e $4N_e$ é, em um população de Wright-Fisher (cada geração da população é reconstituída pela amostragem com reposição baseada nos alelos na população, sendo essa amostragem binomial e a população constante), o tempo esperado para aleatoriamente duas formas gênicas, escolhida ao acaso, possuírem um ancestral comum, pois esse tempo é de $2N_e$ gerações. Sendo o θ então, o polimorfismo genético esperado para estes dois alelos multiplicados por 2, ou seja, duas vezes o número esperado de novos mutantes por geração. Existem várias formas de estimar o parâmetro θ e a mais utilizada, até a última década, foi a desenvolvida por WATTERSON em 1975. Ela utiliza o número de sítios segregantes (polimórficos), em uma amostra de seqüências, sob um Modelo de Sítios Infinitos, onde uma mutação nunca ocorre duas vezes no mesmo sítio. Essa forma de estimar o θ foi adaptada por TAVARÉ (1984) e HUDSON (1990) para um modelo de coalescência (ver abaixo) e atualmente as estimativas do θ

seguem uma abordagem genealógica baseado em coalescência, sendo testadas por máxima verossimilhança (KUNHER *et al*., 1998).

Migração

A última força evolutiva a ser abordada é a migração, ou seja, o movimento dos organismos entre sub-populações gerando um fluxo gênico entre esses agrupamentos. É o processo evolutivo que pode limitar nos mais variados graus a diferenciação, homogeneizando geneticamente as populações. No caso de populações geneticamente estruturadas, a migração também pode ocasionar a introdução de novos alelos. O nível de divergência entre populações pode ser expresso como Nm, ou seja, o produto do N_e e da taxa de migração (*m*). Dependendo do tamanho de uma população, *m* pode impedir ou não a diferenciação. Em um modelo neutro, para um mesmo *m* (e.g. 0,01 média de um indivíduo migrante a cada cem gerações), duas populações com um N_e alto (e.g. N_e 10000, Nm = 100) apresentarão, uma diferenciação relativamente menor se comparadas com duas populações com um N_e baixo (e.g. N_e 100, Nm = 1). Em populações com baixo N_e, *m* precisa ser alto para evitar a divergência causada unicamente pela deriva. Obviamente a seleção também estará atuando, o que pode direcionar a divergência mesmo com um alto *m*.

O índice mais utilizado para descrever indiretamente o fluxo gênico entre populações é o F_{ST} (Índice de fixação ou estatística-F) que foi descrito por Sewall Wright em 1951 (WRIGHT, 1950) e posteriormente adaptado por outros autores (SLATKIN, 1989; EXCOFFIER *et al*., 1992; HUDSON *et al*., 1992). Em sua descrição original Wright representa o coeficiente de endocruzamento de uma população, ou seja, a probabilidade que os alelos de uma população ancestral sejam combinados no zigoto. Seu cálculo nas definições atuais é a proporção da variância nas freqüências alélicas entre as populações (NEIGEL, 2002)

F_{ST} = Var p / mp (1 mp), onde mp é a média da proporção da freqüência alélica.

Ele pode ser compreendido como a estruturação genética de uma população, ou seja, a relação entre o quanto da diversidade genética está presente nos indivíduos, nas sub-populações ou na população como um todo (EXCOFFIER *et al*., 1992). O cálculo clássico leva em conta a heterozigozidade (número de indivíduos heterozigotos dividido pelo número total de indivíduos de cada sub-população) na fórmula:

F_{ST} = (HT HS) / HT, onde HS é heterozigozidade média da sub-população e HT é heterogozidade total da população.

O valor máximo de F_{ST} é 1 e significa que a população encontra-se completamente estruturada e cada sub-população apresenta um alelo exclusivo. Enquanto o valor mínimo é 0, onde a população não apresenta estruturação e as sub-populações apresentam freqüências gênicas iguais. Os valores intermediários representam diferentes níveis de isolamento das sub-populações.

Outra forma de interpretar o F_{ST} é como uma medida indireta do fluxo gênico (HUDSON *et al*., 1992).

F_{ST} = 1 / (1 + 4Nm), sendo possível estimar o próprio Nm do F_{ST} calculado para a população.

A estruturação gênica, o número de migrantes e o N_e são ferramentas importantes para a compreensão dos processos evolutivos sofridos pelos organismos e até mesmo para recomendações

de práticas conservacionistas.

A teoria neutralista

Para compreendermos muitos destes modelos matemáticos utilizados para interpretar a evolução molecular é necessário conhecer a base da Teoria Neutralista de Motoo Kimura (KIMURA, 1968). Deve ficar claro que esta teoria não tenta explicar a evolução no nível morfológico, mas exclusivamente no nível molecular. Ela admite que grande parte da evolução ocorre pela fixação de mutações neutras pela ação da deriva genética. Para Kimura a maioria dos polimorfismos encontrados é seletivamente neutra e por extensão só podem ser fixados por processos estocásticos. HARTL & CLARK (1997), em seu livro texto de genética de populações, se referem à Teoria Neutralista como uma das mais completas e elegantes de todo o campo da ciência biológica. Atualmente, depois de décadas discutindo o dilema neutralismo versus selecionismo, a maioria dos teóricos reconhece que as forças da deriva genética estão sempre presentes e que mesmo se os polimorfismos encontrados não forem estritamente neutros a Teoria Neutralista permanece útil gerando hipóteses nulas rigorosas.

Identificando neutralidade em seqüências

Atualmente, existem diversos índices que testam a neutralidade de um gene ou região genômica. Alguns destes testes buscam a relação entre as mutações sinônimas (dS) e não sinônimas (dN; ZHANG *et al.*, 1997) tentando encontrar sinais de seleção positiva, pois sob seleção a freqüência de dN é acelerada em relação a dS. Dessa forma, podemos calcular a proporção entre as taxas $\omega = dN/dS$, onde $\omega > 1$ representa um sinal de seleção positiva, $\omega < 1$ representa um sinal de seleção purificadora agindo sob o gene e $\omega = 1$ indica que a mudança de AAs é neutra (para uma revisão ver YANG & BIELAWSKI, 2000).

Outros testes de neutralidade avaliam não a relação entre mudanças na proporção dS/dN, mas estatísticas como por exemplo a relação entre sítios segregantes (S) e a diversidade média da variabilidade genética em uma população (k). TAJIMA (1989) desenvolveu seu teste D utilizando estas estatísticas, mais precisamente relacionando os θs calculados através dessas duas medidas (S e k).

Se existe algum valor adaptativo nos mutantes em uma população as duas estimativas θs devem ser diferentes. Se essa diferença não existir a região está evoluindo sem pressão de seleção. Por outro lado, um valor negativo representa uma seleção positiva, enquanto um valor positivo, uma seleção balanceadora favorecendo a diferenciação. Os parâmetros S e k também são afetados pelas mudanças demográficas e a diferença entre as estimativas de θ baseadas nos mesmos podem ser interpretadas como sinais genéticos dessas mudanças. Populações que apresentam um forte crescimento populacional aumentam primeiramente a diversidade genética através do aumento do número de alelos, ocorrendo assim, um aumento relativo de S em relação a k e o D de Tajima torna-se negativo. Por outro lado, se esta população reduzir significativamente seu tamanho o D ficará positivo, uma vez que os alelos raros serão os primeiros a se extinguirem levando a uma diminuição de S em relação a k. Já uma população estruturada com um tamanho populacional constante apresenta um D de Tajima próximo de

zero ou não significante, pois as diferenças relativas entre S e k não são marcantes.

O processo de coalescência

Atualmente, as análises populacionais baseadas em coalescência têm sido aplicadas nos mais diversos grupos taxonômicos (AVISE, 2000). O significado da coalescência pode ser visualizado se escolhermos dois alelos ao acaso em uma população e acompanharmos o processo genealógico até o ancestral comum mais recente (MRCA, *most recent common ancestor*) levando em conta o tempo em gerações. O modelo matemático para trabalhar com essa história genealógica foi desenvolvido por KINGMAN (1982). Nele a probabilidade de um alelo atual descender de um alelo ancestral qualquer é igual a freqüência original desse alelo ancestral. Podemos compreender intuitivamente o processo analisando o diagrama da Figura 3, onde as linhagens alélicas representadas na população atual estão ligadas em algum ponto do tempo por um MRCA e no processo genealógico alelos foram perdidos de geração em geração devido exclusivamente ao processo de deriva genética. Abordagens de coalescência tem possibilitado produzir distribuições nulas e probabilidades *a priori* que através de uma metodologia de máxima verossimilhança tem possibilitando o cálculo do tempo de alelos ancestrais e o cálculo de parâmetros populacionais como θ, fluxo gênico, crescimento populacional, etc.

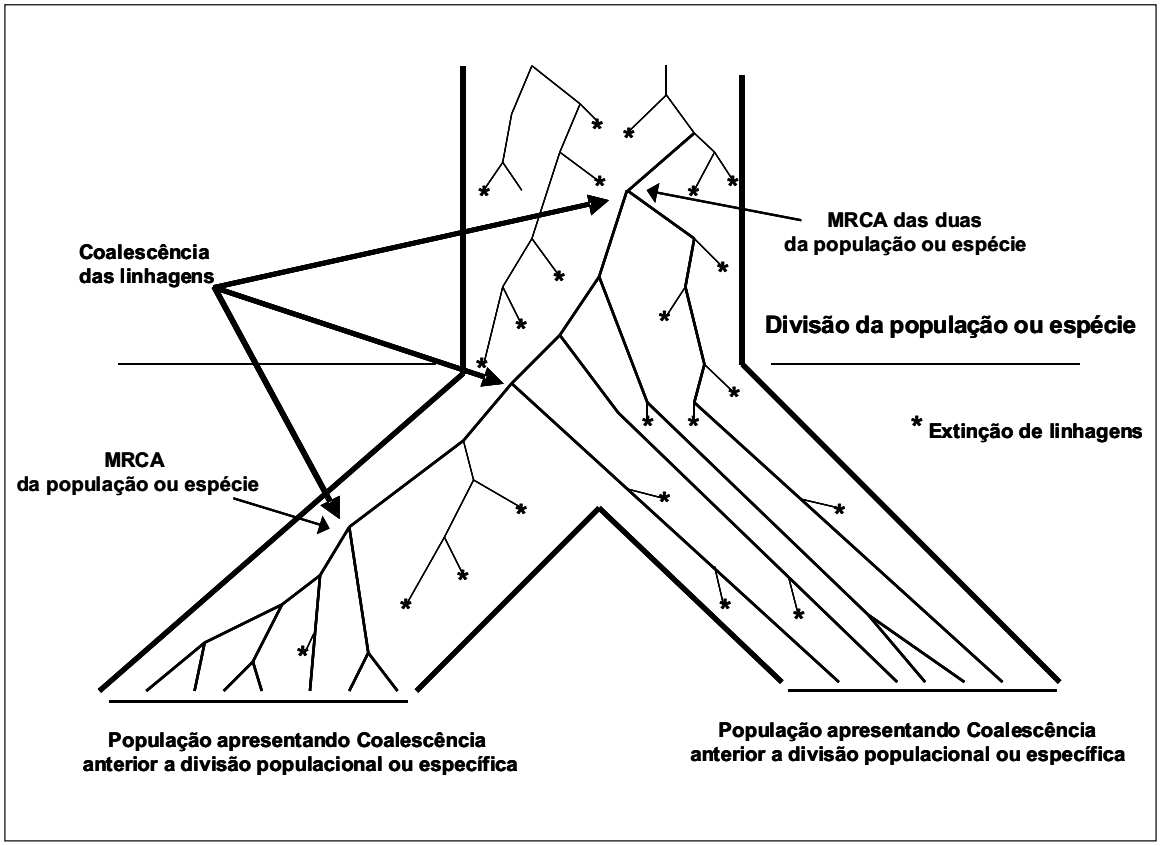

Figura 3: Representação do processo genealógico demonstrando a coalescência de linhagens gênicas, demonstrando também a diferença entre uma árvore de genes e uma árvore de espécies.

Aplicação na conservação

O conhecimento da estruturação de unidades infraespecíficas pode indicar prioridades para a conservação, assim como o conhecimento do fluxo migratório entre populações pode indicar a existência de unidades com relevante diferenciação genética dentro de uma população morfologicamente homogênea. A definição dessas unidades evolutivas tem sido discutida por diversos autores (MORITZ, 1994; CRANDAL *et al.*, 2000). Entretanto todos ressaltam a importância de compreender a variabilidade genética dentro da espécie, limitando suas discussões a aspectos metodológicos. A importância do conhecimento da estruturação genética recai sobre o conhecimento da variabilidade de unidades evolutivas regionais. Muitas vezes, espécies ou populações estruturadas possuem subdivisões que funcionam como reservatórios da maior parcela de suas diversidades genéticas, sendo indicadas para programas conservacionistas.

O valor da variabilidade encontra-se definido no teorema fundamental da seleção natural de FISCHER (1930), no qual a taxa de mudanças evolutivas em uma população é proporcional à soma da diversidade genética disponível. Sendo assim, a perda da diversidade genética reduz as futuras opções evolutivas para uma população. Da mesma forma que os cálculos de F_{ST} e Nm, as estimativas de N_e e outras que indiquem flutuação populacional (como θ e o D de Tajima) podem sinalizar aspectos relevantes para a conservação. Populações com N_e pequeno e que mantém esse tamanho pequeno por um certo tempo, podem, como vimos anteriormente (Fig.2), perder rapidamente a variabilidade genética. Esse tema é de relevante importância e a existência de revistas internacionais focadas freqüentemente ou unicamente nesse assunto atesta esse fato (e.g. *Conservation Genetics*, *Molecular Ecology*, etc.).

Filogenias com seqüências de dna

O debate entre qual abordagem filogenética - morfológica ou molecular - é preferível para responder uma determinada questão persistiu por décadas. Contudo, nos dias atuais, abordagens integradas vêm sendo utilizadas para a compreensão da evolução. A abordagem filogenética dentro da biologia molecular possui princípios iguais aos da filogenia morfológica. É necessário trabalhar com genes ou regiões do genoma que apresentem homologia, ou seja, caracteres que possuem uma história evolutiva comum. Os genes podem sofrer duplicações durante o processo de replicação, esses genes duplicados são chamados de parálogos e do momento de sua duplicação em diante apresentam história evolutivas independentes e não podem ser considerados simultaneamente em uma análise filogenética. Os genes que possuem a mesma história e não apresentaram duplicações são chamados de ortólogos e são os que possibilitam a inferência filogenética da mesma forma que os caracteres morfológicos homólogos.

Estes genes ou regiões genômicas, após serem seqüenciados, devem ser alinhados buscando-se a homologia entre os nucleotídeos para cada sítio. O alinhamento é uma parte crucial de uma análise molecular, pois erros cumulativos nos métodos matemáticos serão introduzidos se forem alinhados nucleotídeos não homólogos. A maioria dos alinhamentos é realizada em programas

computacionais (e.g. Clustal X desenvolvido por THOMPSON *et al.*, 1997) que empregam algoritmos que comparam a similaridade de cada par de seqüências formando grupos que possuem a maior similaridade, usando como guia uma árvore, para depois alinhar as outras seqüências em ordem decrescente de similaridade. As mutações do tipo *indel* que geram lacunas (*gaps*) nas seqüências são penalizadas por esses algoritmos em diferentes graus conforme o tipo de seqüência utilizada, pois como vimos anteriormente os indels podem modificar completamente a tradução de um polipeptídio. Todos os alinhamentos depois de realizados pelos programas devem ser analisados e editados manualmente, pois muitos padrões de homologia ainda não são encontrados pelos algoritmos utilizados. Existem outras formas de alinhamento, como por exemplo, a implementada no programa POY (WHEELER, 1996). Entretanto, a abordagem pormenorizada das formas de alinhamento não cabe neste capítulo (para maiores detalhes consultar VINGRON & WATERMAN, 1994). Contrariamente, vamos nos ater ao cálculo da distância genética, que é uma das chaves para a compreensão das técnicas de análise de evolução molecular. Apesar das diferenças teóricas entre cálculo de distância e modelos evolutivos ambos serão tratados em conjunto.

Modelos evolutivos

Uma distinção ainda existente entre os métodos de análise para dados moleculares e morfológicos se refere à possibilidade de se assumir ou não um modelo probabilístico que tente modelar a evolução do caractere utilizado. Essa diferença decorre da dificuldade em modelar mudanças evolutivas em caracteres morfológicos. Felizmente, a evolução dos dados moleculares é modelável, devido a sua maior simplicidade decorrente da ocorrência de apenas quatro estados (A, C, G e T), permitindo desenhar um arcabouço estatístico matematicamente testável.

Como vimos anteriormente, por décadas os pesquisadores de genética de populações vêm utilizando métodos probabilísticos para modelar os processos evolutivos dentro e entre populações. A utilização destes métodos matemáticos na filogenia é uma forma de mesclar micro e macroevolução, pois os processos evolutivos para os dados moleculares são os mesmos, não existindo uma delimitação entre o que só pode ser realizado em um nível de variabilidade ou em outro.

As possibilidades de substituição para os nucleotídeos estão representadas na Figura 1 e, para cada uma delas (representadas por a, b, c, d, e, f), podemos ter probabilidades distintas de substituição. Podemos também possuir diferentes composições nucleotídicas nas seqüências analisadas, com uma proporção maior de um ou outro nucleotídeo. Todas essas probabilidades podem ser incorporadas em um modelo evolutivo. Entretanto, podemos calcular apenas a distância *p*, que representa somente a proporção de diferenças entre duas seqüências, sendo:

$\boldsymbol{p} = \mathbf{n}p/\mathbf{n},$. onde n*p* representa o número de polimorfismos e n o número de nucleotídeos.

A partir de um certo tempo de diferenciação entre as seqüências, a distância *p* pode subestimar a real divergência entre elas, como podemos observar na Figura 4, pois as substituições podem ocorrer mais de uma vez em cada posição nucleotídica. Para acessar a diferença entre valor de

mudança evolutiva esperado e o observado, utilizamos métodos de correção de distâncias.

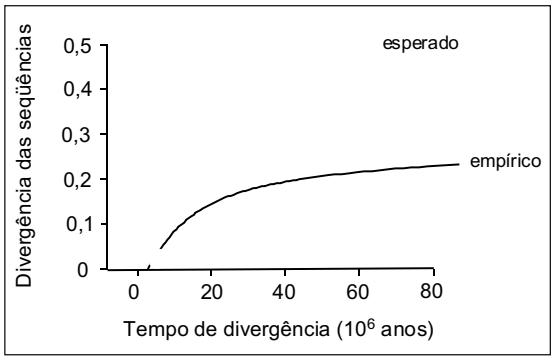

Figura 4: Comparação entre a divergência de seqüências de DNA empírica e esperada relacionadas tempo de divergência.

O modelo de evolução corrigido mais simples é o proposto por JUKES & CANTOR (1969), denominado JC, onde:

$d = -\frac{3}{4} \ln (1 - 4/3p)$

O modelo JC é por isso baseado em uma distribuição de Poisson levando em conta múltiplas substituições para um mesmo nucleotídeo onde *ln* é o logaritmo natural e *p* é a proporção de sítios diferentes. Apesar disso, esse modelo é muito simples e assume que as quatro bases possuem igual proporção (A = C = G = T) e todas as substituições possuem igual probabilidade (a = b = c = d = e = f).

O modelo proposto por KIMURA (1980) denominado K2P (Kimura dois parâmetros) leva em conta não apenas as substituições múltiplas e uma igual proporção de bases (A= C = G = T) como em JC, mas também a diferença entre transições e transversões (b = e, a = c = d = f), sendo a distância calculada por:

$d = \frac{1}{2} \ln [1/(1 - 2P - Q)] + \frac{1}{4} \ln[1/(1 - 2Q)]$, onde P é a proporção de diferenças contando somente as transições e Q é o mesmo para transversões.

Como anteriormente comentado, as transições são muito mais freqüentes que as tranversões nos mais diversos organismos já estudados e, em alguns casos (e.g. DNA mitocondrial) onde as diferenças são muito grandes, essa correção se faz necessária.

FELSENSTEIN (1981) desenvolveu um modelo evolutivo que leva em conta diferentes proporções dos nucleotídeos (A ? C ? G ? T). Esse modelo, F81 foi agrupado com K2P por HASEGAWA *et al*. (1985) no modelo HKY85, no qual cada nucleotídeo possui uma proporção diferente e são assumidas probabilidades diferentes para transições e transversões (b = e, a = c = d = f). Outro modelo muito próximo ao HKY85 é o TrN desenvolvido por TAMURA & NEI (1993) no qual, além da proporção diferencial de bases e a possibilidade de diferentes probabilidades de transição e transversão, os dois tipos de transição possuem probabilidades diferenciadas (b, e, a = c = d = f). Esse modelo se apresenta como um caso especial de HKY85, onde as diferentes transições são pesadas e apresenta uma fórmula para o *d* muito menos complexa que HKY85 (fórmulas não demonstradas).

O modelo evolutivo mais completo é o GTR (General Time Reversible Model) desenvolvido por RODRIGUEZ *et al*. (1990), no qual todos os parâmetros possíveis podem ter uma proporção individual. Portanto, corrige substituições múltiplas (distribuição binomial), proporção diferenciada de nucleotídeos (A ? C ? G ? T) e proporções distintas para as substituições (a, b, c, d, e, f).

Existem diversas variações dos modelos apresentados e até mesmo programas

computacionais (e.g. MODELTEST, POSADA & CRANDAL, 1998) testam quais modelos são mais verossímeis para explicar as diferenças encontradas em um conjunto de seqüências. Embora esses modelos possam incorporar diferenças na proporção dos nucleotídeos (e.g. maior proporção de AT do que CG), todas as seqüências necessariamente seguiram a mesma proporção. Um modelo que possibilita comparar seqüências com proporções de nucleotídeos diferentes é o LogDet, desenvolvido por LOCKHART *et al*. (1994), que através do determinante de uma matriz de substituição entre as seqüências analisadas permite esse tipo de comparação, o que possibilita uma certa correção para análises de seqüências de grupos taxonômicos muito distantes.

Outra forma de corrigir as diferenças entre a evolução dos sítios observada e esperada é assumir outros tipos de distribuição, como o caso da distribuição gama. Assumindo essa distribuição, podemos descrever seqüências onde cada sítio evolua com taxas diferenciadas. Para isso, adicionamos as fórmulas de modelos evolutivos o parâmetro alfa da distribuição gama, o qual, descreve essa variação (ver NEI & KUMAR, 2000 para uma definição matemática).

Reconstrução de árvores filogenéticas

As árvores filogenéticas são representações gráficas que tentam demonstrar as relações evolutivas entre organismos, genes ou regiões genômicas. As primeiras representações datam do século XIX e existem as mais diversas metodologias para a sua construção. Atualmente os métodos mais utilizados são de máxima parcimônia (MP) e os que utilizam a máxima verossimilhança (MV). Outros métodos são utilizados como agrupamentos de vizinhos (*Neighbor joining*), inferência Bayesiana, entre outros. Entretanto, abordaremos apenas o método de MV que é o mais utilizado para dados moleculares [para uma revisão de métodos ver HOLDER & LEWIS (2003) e para uma descrição detalhada sobre árvores filogenéticas ver SWOFFORD *et al.* (1996)]. Este método é baseado na teoria do matemático inglês R. A. Fisher e foi utilizado pela primeira vez para reconstruir uma árvore filogenética por CAVALLI-SFORZA & EDWARDS (1967). Contudo, foi FELSENSTEIN (1981) quem aprimorou o método e desenvolveu algoritmos para utilizar seqüências de nucleotídeos. A verossimilhança difere de uma probabilidade simples devido a sua característica condicional. Sinteticamente, verossimilhança é a probabilidade do dado (D), dada uma hipótese (H), ou seja:

P (D|H)

Isso significa que uma hipótese será testada em cima de quão bem ela prediz os dados observados. Nesse caso, o dado representa o alinhamento de uma seqüência de DNA, enquanto a hipótese se refere à árvore filogenética e o modelo evolutivo que modela as mudanças moleculares.

Para compreendermos melhor o processo do cálculo da verossimilhança, vamos utilizar a árvore e o alinhamento da Figura 5.

Para chegar ao valor da verossimilhança da hipótese filogenética, precisamos calcular a probabilidade de cada sítio em cada posição da árvore. Por exemplo, analisando o sítio I a seqüência 4 possui um T. Esse T é o nosso dado e devemos calcular a probabilidade da seqüência 4 possuir esse dado observado. Podemos assumir que o ancestral representado pelo nó 7 possuía um C e mudou para T

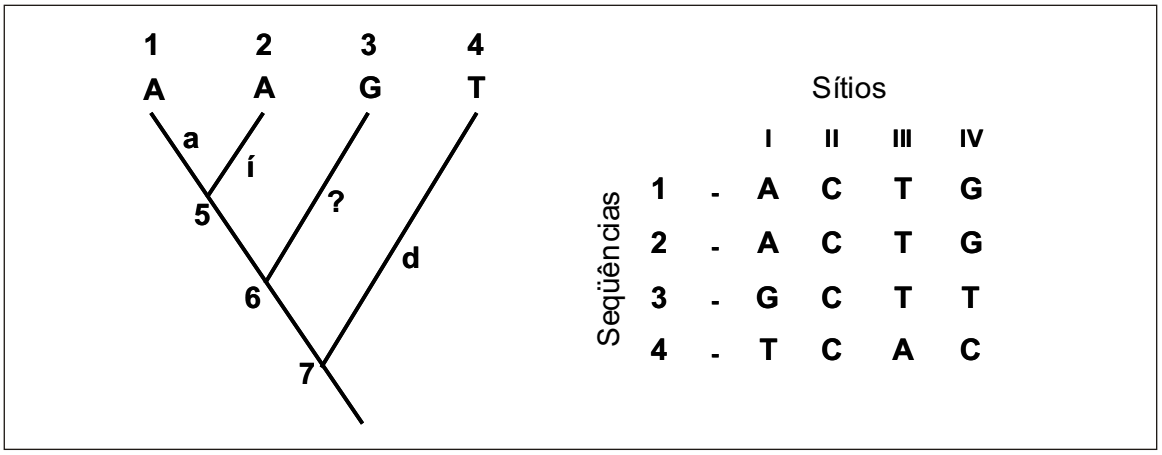

Figura 5: Árvore filogenética hipotética e alinhamento hipotético.

na seqüência 4. Dessa forma, podemos calcular a probabilidade usando um modelo evolutivo qualquer de que no nó 7 temos um C (usando a proporção de C) e que C mudou para T (substituição *e* na Figura 1). Entretanto não sabemos realmente qual era o nucleotídeo do sítio I no nó ancestral 7 e precisamos multiplicar todas as possibilidades de nucleotídeos para o nó 7 e todas as possibilidade de troca desses nucleotídeos para T na seqüência 4. Precisamos até mesmo levar em conta a possibilidade de que o nó 7 possuía um T e se manteve como T na seqüência 4. Desta forma, a verossimilhança de um nucleotídeo estar presente no nó 7 e mudar ou não para outro na seqüência 4 está relacionada com a proporção nucleotídica e a chance de cada substituição ocorrer como estão definidas no modelo evolutivo utilizado. Esse cálculo será repetido para todas as seqüências (1, 2, 3, 4) e todos os sítios (I, II, III, IV), sendo levado em conta ainda o tamanho dos ramos (α, β, γ, δ). O produto dessas probabilidades é muito pequeno e transformamos a verossimilhança em logaritmo natural para trabalharmos com números maiores. Dessa forma o produtório das probabilidades transforma-se no somatório do logaritmo das mesmas.

Até agora teríamos calculado a verossimilhança de apenas uma árvore, mas para o exemplo da Figura 5 existem 15 árvores enraizadas para serem testadas e diversos tamanhos de ramos. Desta forma, o método de MV é a comparação da verossimilhança das árvores possíveis, sendo que aquela que apresentar a maior verossimilhança será considerada a melhor hipótese filogenética. Entretanto, a quantidade de árvores possíveis para qualquer análise com mais de 20 táxons torna o processo praticamente inviável computacionalmente (para 20 táxons existem 8.200.794.532.637.891.559.000 árvores possíveis), pois o número de árvores enraizadas é dado por:

$(2n\ \ 3)!/\ 2^{n-2}(n-2))!$, onde n é o número de táxons analisados.

Para encontrarmos a árvore de MV somos forçados a empregar técnicas de busca heurística da mesma forma que ocorre com análises de MP, ou seja, realizamos o cálculo em um conjunto de árvores em vez de executarmos uma busca exaustiva ou uma busca somente eliminando soluções sub-ótimas (*Branch-and-baund*). Os algoritmos heurísticos mais utilizados são os do tipo *Branch-swapping* e foram inicialmente desenvolvidos para MP, onde uma árvore inicial é construída e o número

de passos é calculado. Após esta etapa, os táxons são rearranjados e a cada novo rearranjo o cálculo é refeito e a árvore com menor número de passos é escolhida. Essas abordagens reduzem enormemente o tempo de análise e a diferença entre MP e MV é a forma de escolher a melhor árvore que em MP se faz pelo menor número de passo, enquanto em MV se escolhe pela maior verossimilhança. Estas técnicas de reconstrução de árvores filogenéticas com dados moleculares são realizadas em programas como o PAUP* 4.0b10 (SWOFFORD, 2002) e Phylip (FESLSENSTEIN, 2004), entre outros.

Existem muitos artigos mostrando tanto falhas como vantagens encontradas nas análises com MV. Contudo o método é considerado pela maioria dos pesquisadores da área a mais completa ferramenta para o estudo de dados moleculares, pois possibilita testes de hipótese evolutiva e comparação de MVs. Atualmente, diferentes algoritmos visando à otimização do processo de busca de árvores tem sido desenvolvidos para MV (GUINDON & GASCUEL, 2003).

Networks

Muitos estudos moleculares representam a relação evolutiva entre as seqüências e não entre os organismos. Para isso não utilizam árvores bifurcadas (Fig.3), mas sim redes de haplótipos (*networks*). Essas redes são freqüentemente as melhores representações para estudos intraespecíficos ou de espécies próximas, pois assumem que seqüências ancestrais podem estar presentes na amostragem e demonstram a relação genealógica entre essas seqüências. Elas indicam também a freqüência de um haplótipo, sua relação com os demais e a quantidade de mutações que os

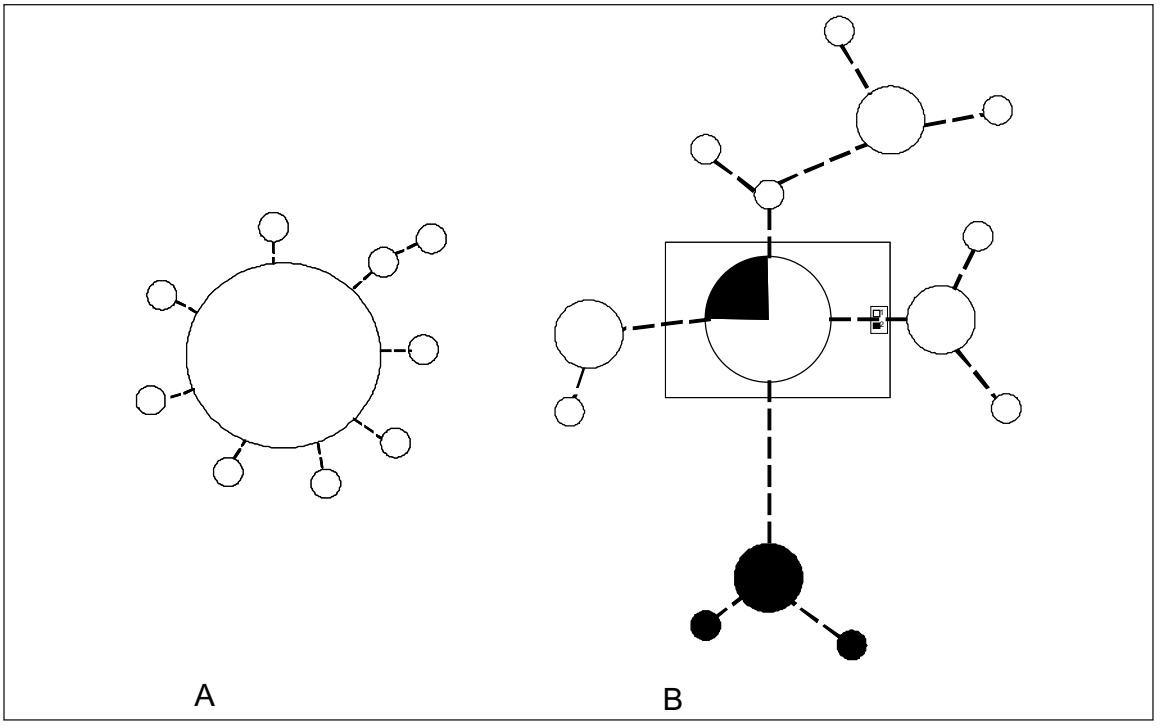

Figura 6: Network hipotético mostrando a relação evolutiva, freqüência e passos mutacionais entre haplótipos. A Clássica formação em estrela obtida de uma população em expansão demográfica, B Representação de uma população estruturada com compartilhamento de haplótipos. Círculos pretos população X, Círculos brancos população Y.

separam (Fig.6). Podemos visualizar através das redes, a forma gráfica de populações em expansão demográfica (Fig.6A), na qual um dos haplótipos possui uma elevada freqüência, sendo o ancestral de uma série de haplótipos raros (formação em estrela). Da mesma forma, podemos observar organismos com variabilidade genética estruturada (Fig.6B), possuindo vários haplótipos freqüentes com grande variabilidade entre eles, podemos até mesmo visualizar o compartilhamento de haplótipos entre populações diferentes. Esses dados gráficos podem ser comparados com as estatísticas anteriormente descritas (D de Tajima, F_{ST}, Nm, Crescimento populacional, etc.), buscando construir um cenário mais realista do processo histórico sofrido pelo organismo. Os programas mais utilizados para a construção de redes de haplótipos são: Network (BANDELT *et al*., 1999) e TCS (CLEMENTE *et al*., 2000), que utilizam algoritmos de *Median Joining* e parcimônia estatística, respectivamente.

Exemplos de estudos com dados moleculares

Atualmente diversos grupos estão trabalhando com ferramentas moleculares em problemas envolvendo a herpetofauna brasileira (Instituto Butantan, MZUSP, Instituto de Biociências da USP, UFAM, UNESP, UCG, entre outros). O Centro de Biologia Genômica e Molecular da PUC do Rio Grande do Sul (CBGM), no qual desenvolvemos estudos moleculares, é um dentre estes grupos que vêm realizando projetos que tentam responder perguntas sobre o padrão da variabilidade genética das espécies, bem como os processos que moldaram essa variabilidade. Em um destes estudos, GRAZZIOTIN *et al*. (no prelo), analisando através da variabilidade do DNA mitocondrial o padrão filogegráfico da espécie *Bothrops jararaca*, encontraram uma clara distinção entre as populações do sul e sudeste do país. Essa distinção se traduz em uma diferença genética de 3,2% e leva a conclusão de que estes grupos de populações possivelmente estavam separados geograficamente por alguma barreira que impediu o fluxo gênico a aproximadamente 3,8 milhões de anos. Neste estudo observa-se que a diversidade genética média de *B. jararaca* é de 2,1%, sendo esta comparativamente semelhante à diversidade de todo o grupo *atrox* (WÜSTER *et al*.,1999), que engloba no mínimo sete espécies. Esta informação de diversidade torna-se relevante para a compreensão dos processos que moldam os ecossistemas sul-americanos, uma vez que muito pouco se sabe sobre a variabilidade dos componentes da fauna neotropical. Em outro estudo em desenvolvimento no CBGM, estamos analisando a diversidade genética das espécies *Bothrops fonsecai* e *Bothrops cotiara*, buscando compreender a distribuição disjunta (Serra da Mantiqueira versus Planalto das Araucárias) desses dois táxons estritamente relacionados. Os resultados prévios deste estudo indicam que somente eventos muito antigos, como a orogênese das serras litorâneas (SIMPSON, 1979), podem explicar a grande divergência entre as espécies (seis milhões de anos) e que as espécies possuem uma diversidade baixa, o que pode influenciar as ações conservacionistas, uma vez que, ambas espécies encontram-se em listas estaduais de animais ameaçados de extinção.

Estes exemplos demonstram a importância da integração entre os dados moleculares e

conhecimentos sobre a biogeografia e diversidade da fauna neotropical. Contudo, a integração atualmente mais proeminente é a realizada entre dados moleculares e dados morfológicos.

A integração entre análises moleculares e morfológicas

As relações filogenéticas inferidas com dados moleculares vêm tornando-se freqüentes para todos os grupos de organismos. Entretanto, muitas vezes, não são observadas as diferenças básicas entre árvores de genes e árvores de espécies. Os nós de uma árvore filogenética molecular não necessariamente indicam um evento de especiação (Fig.5), assim como o compartilhamento de alelos (ou haplótipos) entre táxons não significa que um desses táxons em questão não seja válido. A definição de espécies ou outros níveis taxonômicos recai sobre outros fatores além de uma árvore filogenética molecular. Essa árvore não deve ser encarada como uma representação da real relação entre as espécies, mas sim como um dos sinais deixados pelo processo evolutivo que gerou essa relação. Esses sinais podem ou não estar diretamente relacionados com a relação entre as espécies, o que deve obrigatoriamente ser analisado caso a caso.

A melhor forma de testar hipóteses moleculares é contrapô-las às análises morfológicas visando integrar mais de um tipo de dado, comparando e buscando explicações para os resultados.

Em geral as análises moleculares não utilizam genes que possibilitem observar visualmente a variabilidade na aparência do organismo. Freqüentemente nas análises acima mencionadas não são utilizados genes de coloração, ou que produzem diferentes números de escamas, ou qualquer outra alteração morfológica. Tentamos trabalhar com genes que não evoluam estritamente por seleção natural, mais ainda, buscamos os genes mais evolutivamente neutros. Quando trabalhamos com genes moldados pela seleção, estamos contando a história da pressão de seleção que este organismo sofreu, o que muitas vezes pode não representar a história evolutiva geral do organismo.

Esta é uma das vantagens de trabalharmos com ambos tipos de dados morfológicos e moleculares de forma integrada. Uma vez que os eventos de especiação ocorrem constantemente, frutos de um processo de diferenciação entre unidades evolutivas, podemos encontrar atualmente organismos em diversos estágios de especiação, muitas vezes sem nenhuma característica visível de divergência morfológica, mas com marcantes divergências genéticas. Por outro lado, organismos claramente identificados por suas diferenças morfológicas, podem apresentar-se iguais para a maioria do genoma (excluindo os genes que apresentam essas diferenças).

Atualmente abordagens simultâneas, onde dados morfológicos e moleculares são utilizados, foram realizadas com espécies da nossa fauna (e.g. BURBRINK, 2005 e FERNANDES *et al*., 2004; entre outros). Nos próximos tópicos pretendemos demonstrar como estes trabalhos podem ser realizados abordando o uso de caracteres morfológicos com sobreposição em um trabalho que reinterpreta a taxonomia do gênero *Lachesis*, usando como base os dados de DNA mitocondrial.

Usando caracteres polimórficos com sobreposição em análises filogenéticas

Ao contrário dos dados provenientes de seqüências de DNA, dados morfológicos podem, em muitos casos, serem difíceis de codificar em análises filogenéticas. Com efeito, a descoberta e seleção de caracteres para uma análise cladística podem incluir diversas etapas subjetivas que, potencialmente, podem obscurecer o resultado da análise.

Enquanto as seqüências de DNA oferecem caracteres com distribuição discreta (limitada aos pares de bases), dados morfológicos podem apresentar distribuição contínua, muitas vezes com sobreposições [veja THIELE (1993) para uma análise detalhada da terminologia dos tipos de dados]. Mesmo caracteres tidos como qualitativos podem ser apenas uma interpretação simplória de caracteres com distribuição contínua na natureza. Por exemplo, estados de caracteres descritos como "longo" ou ainda "reduzido", nada mais são do que um comprimento, assim como "circular" e "oval" são uma razão. Assim sendo, termos como "qualitativo" e "quantitativo" referem-se a modos de expressão e não à qualidades intrínsecas dos dados (WILEY, 1981). Ainda assim, diversos autores (e.g. PIMENTEL & RIDGES, 1987; CHAPPILL, 1989; MURPHY & DOYLE, 1998) argumentam que caracteres com extensiva sobreposição, sejam eles de natureza polimórfica, merística ou morfométrica, não possuem informação filogenética apreciável e, assim sendo, devem ser excluídos de análises filogenéticas. Em oposição a essa visão, uma série de técnicas foram desenvolvidas para a codificação destes caracteres (THIELE, 1993; SMITH & GUTBERLET, 2001; WIENS, 2001) e vários estudos (GUTBERLET, 1998; POE, 1998; GUTBERLET & HARVEY, 2002; FERNANDES *et al.*, 2004) mostraram a eficácia de tais métodos aplicados à problemas de sistemática herpetológica.

Diversos métodos foram desenvolvidos para lidar com caracteres com sobreposição, tais como as contagens do número de escamas ventrais nas espécies de um determinado gênero de serpente. O mais simples, codificação simples de lacunas (MICKEVICH & JOHNSON, 1976), consiste na procura de lacunas na distribuição de freqüências dos caracteres para separar os diversos estados dos caracteres. Assim sendo, se considerarmos quatro espécies (A-D) com um caráter variando em cada espécie entre 120-130, 125-135, 150-160, 162-172, respectivamente, poderíamos usar as duas lacunas para definir três estágios (120-135, 150-160 e 162-172). Alternativamente, poderíamos utilizar um tamanho mínimo de lacuna para separar os estados. Ainda assim, a codificação simples de lacunas e métodos similares (em geral introduzindo testes estatísticos para determinar a significância das lacunas) não são capazes de aproveitar toda a informação disponível, como o fato da primeira lacuna ser bem maior que a segunda, ou ainda a sutil, porém real diferença entre as distribuições de freqüência nas espécies A e B, ou ainda a questão da realidade das próprias lacunas, uma vez que estas poderiam ser artefatos provocados pela amostragem insuficiente dos táxons.

O método da codificação ponderada de lacunas (THIELE, 1993) é baseado na pesagem diferencial das lacunas entre os estados codificados. Conforme a proposta original, os estados são atribuídos aos valores das médias em cada espécie dos dados padronizadas pelo alcance. O alcance é determinado pelo número máximo de estados permitidos pelos algoritmos de busca. Assim sendo, o táxon com a menor média é codificado com o mais

baixo valor de estado possível, enquanto a maior média recebe o maior valor. Valores intermediários são atribuídos proporcionalmente às médias. Voltando ao exemplo anterior, se as médias dos táxons A-D fossem, respectivamente, 125, 130, 155 e 167, e se fosse utilizado um algoritmo capaz de codificar 10 estados, os estados atribuídos seriam: 0, 1, 7 e 9. Como os estados são tratados como ordenados, o método captura a informação do tamanho das lacunas, associando-as a um número proporcional de passos (1, 6 e 2, no exemplo acima). Ainda que THIELE (1993) tenha utilizado média, ele não defendeu uma medida de tendência central em particular. Contudo, FERNANDES *et al.* (2004) preferiram utilizar as medianas, pois consideraram estas menos sujeitas a efeitos de *outliers*, que podem ser bastante influentes em casos onde a amostra é limitada.

O método de THIELE (1993) foi aperfeiçoado por WIENS (2001), que propôs a codificação ponderada de lacunas por matriz de passos. Usando o método de THIELE (1993), há um número limitado de estados que podem ser utilizados nas atuais implementações dos principais programas de busca de árvores mais parcimoniosas (10 no HENNIG86 e 32 no PAUP*). WIENS (2001) propôs que, alternativamente, construa-se uma matriz de passos para cada caráter, aproveitando-se do fato que, no PAUP*, o custo entre os estados de uma matriz de passos pode ser de até 1000 (999 no MacClade). Assim sendo, os valores das médias (ou medianas) poderiam ser codificados com uma precisão muito maior do que no método de THIELE (1993). Voltando ao exemplo anterior, as médias poderiam ser codificadas numa escala de 0 a 1000 (Tabela 5) e, posteriormente, introduzidas em uma matriz de passos para processamento (Tabela 4). A matriz de passos é tratada pelo PAUP* como um único caráter e os valores das células representam o número de passos adicionados à árvore quando ocorre a mudança de um estado para outro.

Tabela 3: Exemplo hipotético mostrando a conversão das médias em valores para a montagem da matriz de passos.

Táxon	Média	Valor	Código
Espécie A	125	0	0
Espécie B	130	119	1
Espécie C	155	714	2
Espécie D	167	1000	3

Tabela 4: Matriz de passos montada a partir dos dados da Tabela 1. Os códigos representam a diferença entre os valores obtidos para cada táxon.

	0	**1**	**2**	**3**
0	0	119	714	1000
1	119	0	595	881
2	714	595	0	286
3	1000	881	286	0

Tabela 5: Freqüência do caráter número de dentes nas espécies A-D

Táxons	Número de Dentes				
	8	9	10	11	12
Espécie A	1	0	0	0	0
Espécie B	0	0	0,2	0,5	0,3
Espécie C	0	0,6	0,4	0	0
Espécie D	0	1	0	0	0

Infelizmente, o uso de matrizes de passos é limitado pela implementação disponível no PAUP*. Não só o programa é capaz de trabalhar apenas com um número limitado de matrizes de passos, mas, também, a utilização desta ferramenta pode prolongar a busca das árvores mais parcimoniosas para uma escala de tempo proibitiva.

Tanto a codificação ponderada de lacunas quanto sua versão aplicada a uma matriz de passos utilizam-se apenas de uma medida de tendência central, desprezando qualquer outra informação da distribuição de freqüências do caráter. Assim sendo, SMITH & GUTBERLET (2001) desenvolveram o método de codificação geral de freqüências para aproveitar este tipo de informações. Este método combina idéias da codificação ponderada de lacunas com o método de codificação de freqüências, desenvolvido para aplicação em caracteres com estados polimórficos.

Polimorfismo é outro fenômeno difícil de ser codificado em análise filogenéticas, tanto que, freqüentemente, caracteres polimórficos são excluídos mesmo de estudos taxonômicos por serem considerados não informativos. Diversas técnicas foram desenvolvidas (ver a revisão de WIENS, 1999), em particular o método de codificação de freqüências (*frequency-bins*; WIENS, 1993, 1995), que permite a codificação de caracteres polimórficos binários, ou seja, caracteres com dois estados os quais ambos estão presentes em pelo menos um táxon. Neste método, os caracteres são codificados em um número de estados, em geral 25, baseados na freqüência observada da condição derivada em cada táxon. Letras são utilizadas como códigos para as freqüências, de forma que a=0-3%, b=4-7%, ..., y=96-100%. Assim sendo, um táxon no qual todos os espécimes apresentam o estado ancestral é codificado como a, enquanto um outro em que todos os espécimes apresentam o estado derivado é codificado como y. Contudo, um táxon que apresente 69% dos exemplares com o estado derivado seria codificado como r. Após a codificação, o caráter é tratado como ordenado, de forma que grandes mudanças na freqüência são representadas por um grande número de passos, em oposição a pequenas mudanças, as quais são representadas por poucos passos.

A codificação geral de freqüências (SMITH & GUTBERLET, 2001) transforma caracteres multiestado polimórficos em um número de subcaracteres, nos quais cada estado observado é tratado como um caráter individual. Codificação por freqüências é aplicada a cada um dos subcaracteres de forma a capturar a distribuição de freqüências encontrada em cada táxon. Finalmente, para evitar a supervalorização dos caracteres polimórficos, um esquema de pesagem é aplicado para que cada subconjunto de subcaracteres tenha o mesmo peso que um caráter binário. Considere que as espécies A-D possuam número variável de dentes, cuja distribuição é conhecida (Tabela 5). A mesma pode ser transformada em uma matriz de subcaracteres (Tabela 6), na qual cada elemento representa a freqüência acumulada dos estados a direita da coluna do elemento. Assim sendo, 30% dos exemplares da espécie B possuem mais de 11 dentes, enquanto 100% dos exemplares da espécie C possuem mais de oito dentes. É importante notar que, nesta transformação, elimina-se a última coluna já que seria não informativa, apresentando freqüência acumulada nula em todos os táxons. A matriz de subcaracteres (Tabela 6) pode ser agora

Tabela 6: Subcaracteres para o número de dentes das espécies A-D, apresentando a freqüência acumulada para cada estado.

Táxons	Subcaracteres para o Número de Dentes			
	8	9	10	11
Espécie A	0	0	0	0
Espécie B	1	1	0,8	0,3
Espécie C	1	0,4	0	0
Espécie D	0	0	0,5	0,5

codificada por freqüências resultando na matriz final (Tabela 7). Uma rápida análise da matriz final revela que o subcaracter para oito dentes é não informativo e, por isso, deve ser eliminado.

Este processo deve ser repetido para todos os caracteres polimórficos multiestado. Ao final, os diversos conjuntos de subcaracteres devem ser integrados aos caracteres convencionais em uma única matriz. Uma vez que os caracteres polimórficos foram divididos em inúmeros subcaracteres, é necessário um esquema de pesagem para evitar que estes tenham maior peso que os binários. O peso de cada subcaracter deverá ser igual ao de um caráter binário dividido pelo número de subcaracteres em seu conjunto e pelo número de passos entre o maior e menor código de freqüência em cada subcaracter. SMITH & GUTBERLET (2001) recomendam que se atribua o peso 32.767 para cada caráter binário, uma vez que este é o maior peso possível no PAUP*, de forma que as operações sejam realizadas com números inteiros, o que é muito mais rápido do que as computações com ponto flutuante. Assim sendo, o subcaracter 9 deverá ter peso 341 (32.767/4/24, 4 subcaracteres no conjunto e 24 passos [a-y]), enquanto o 10 deverá ter peso 410 (32.767/4/20, 4 subcaracteres no conjunto e 20 passos [a-u]). Assim como no método de codificação ponderada de lacunas (THIELE, 1993), os caracteres devem ser considerados ordenados para incorporar a informação das diferentes extensões das lacunas.

SMITH & GUTBERLET (2001) recomendam que se mantenha os subcaracteres não informativos na matriz, mesmo que sejam excluídos da análise. A razão para tanto é que o método não requer mudança nos estados dos subcaracteres com a adição de novos táxons, o que pode ser útil em caso de estudos continuados. Finalmente, há um programa disponível para a implementação da codificação geral de freqüências, CodeThis! (GUTBERLET *et al*., 2000).

Diversos estudos já foram realizados

Tabela 7: Subcaracteres para o número de dentes das espécies A-D, apresentando a codificação de freqüências.

Táxons	Subcaracteres para o Número de Dentes			
	8	9	10	11
Espécie A	a	a	a	a
Espécie B	y	y	u	h
Espécie C	y	k	a	a
Espécie D	y	y	n	n

usando métodos como estes descritos e, até mesmo, comparando os resultados obtidos sob técnicas de análises diferentes. As serpentes do gênero viperídeo *Lachesis* foram investigadas por FERNANDES *et al*. (2004), que reconstruíram a história filogenética de seis grupos geográficos, representando as duas populações centro americanas, das vertentes do Atlântico e do Pacífico, bem como quatro sul americanas, incluindo a da Mata Atlântica, a do Mato Grosso e duas amazônicas, ao norte e ao sul do rio Amazonas. As análises foram baseadas em uma matriz com 26 caracteres morfológicos, incluindo 18 caracteres merísticos com sobreposição e um polimórfico qualitativo. As análises incluíram o tratamento dos dados segundo os métodos de codificação ponderada de lacunas e codificação geral de freqüências. Com esta última técnica obteve-se uma matriz de 284 caracteres (sete convencionais e 277 subcaracteres), dos quais 95 eram não informativos, de forma que não foram considerados. Por último, realizou-se uma terceira análise combinando-se os dados morfológicos codificados através de ponderação de lacunas com as seqüências de DNA mitocondrial obtidas por ZAMUDIO & GREENE (1997), formando uma matriz de 812 caracteres, dos quais 254 caracteres eram informativos.

Todas as análises apresentaram resultados com topologias idênticas dentro do grupo de estudo (FERNANDES *et al.,* 2004), comprovando a eficácia dos caracteres polimórficos com sobreposição de recuperar o sinal filogenético (Fig.7). A análise por codificação geral de freqüências obteve uma única árvore com índices de consistência 0,744 e retenção 0,563. A análise por codificação ponderada de colunas obteve uma única árvore com índices 0,789 e 0,624, respectivamente. Por fim, a análise combinada obteve cinco árvores mais parcimoniosas com índices 0,817 e 0,612, cujo consenso estrito era similar as demais análises com as ambigüidades limitadas aos grupos externos. Apesar da congruência das árvores, é curioso notar que na avaliação dos clados internos por Bootstrap (Fig.7), a análise combinada apresentou melhores resultados que a codificação geral de freqüências que, por sua vez, se mostrou melhor que a codificação por ponderação de lacunas. Contudo, este padrão foi quebrado no clado formado pelas populações da Mata Atlântica e do Mato Grosso, no qual apenas a análise de codificação por ponderação de lacunas mostrou um resultado aceitável (acima de 70%, conforme sugerido por HILLIS & BULL, 1993).

Algumas das críticas mais contundentes a incorporação das freqüências de caracteres polimórficos (com ou sem sobreposição) em

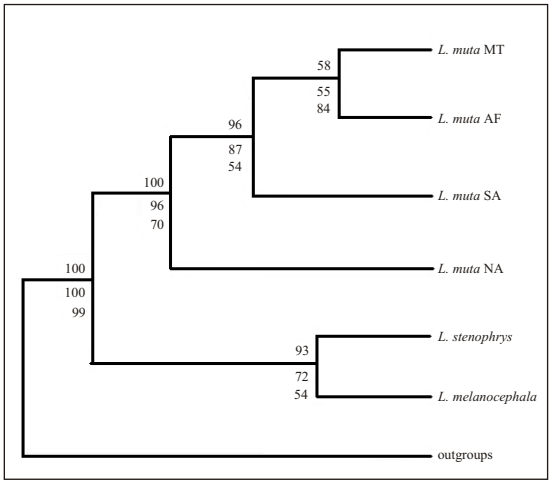

Figura 7: Árvore mais parcimoniosa recuperada para os grupos geográficos de *Lachesis*. Números nos ramos referem-se a valores de Bootstrap obtidos com 1.000 pseudoréplicas nas análises (de cima para baixo) combinadas, por codificação geral de frequências e codificação ponderada de lacunas. MT= Mato Grosso; SA= sul do rio Amazonas; AF= Floresta Atlântica; NA= norte do rio Amazonas (FERNANDES *et al*., 2004).

análises filogenéticas (e.g., MURPHY & DOYLE, 1998, SIDALL & KLUGE, 1997) são as afirmações de que freqüências variam rapidamente, tanto na escala temporal quanto espacial, e que freqüências não são características hereditárias. Em relação ao primeiro ponto, concordamos com a postura de SMITH & GUTBERLET (2001) que apontam que a variação temporal e espacial é um problema geral da reconstrução filogenética na forma de homoplasia. Caracteres baseados em freqüências não são especiais em relação à homoplasia e devem ser testados como qualquer outro caráter, através da congruência com os demais caracteres em uma análise filogenética. Em relação a heriditabilidade, WIENS (1999) lembra que freqüências podem ser herdadas caso as populações estejam em um equilíbrio de Hardy-Weiberg. Ainda que estas freqüências forçosamente sejam variáveis durante o processo evolutivo, aparentemente esta mudança não impede que métodos baseados em freqüência recuperem a informação filogenética.

Conclusões

As análises moleculares ou conjuntas (dados morfológicos e moleculares) permitem várias inferências que as abordagens clássicas sozinhas não permitem. Infelizmente o Brasil não possui muitos trabalhos publicados utilizando análises moleculares ou integrados com elementos da herpetofauna. Os principais trabalhos realizados com evolução molecular foram desenvolvidos fora do país, demonstrando a carência de pessoal e infra-estrutura para esses estudos no Brasil. Podemos citar os trabalhos coordenados por W. Wüster da Universidade de Bankor, Londres, que em associação com M. G. Salomão e G. Puorto, ambos do Instituto Butantan, realizaram diversos estudos moleculares com o gênero *Bothrops* (WÜSTER *et al*., 1999; PUORTO *et al.*, 2001; WÜSTER *et al*., 2002) Também podemos citar os trabalhos de J. Sites Jr. da Brigham Young University e N. J. da Silva da Universidade Católica de Goiás (SILVA e SITES, 2001), bem como o de K. Zamudio e H. Greene, ambos da Cornell University (ZAMUDIO & GREENE, 1997), que trabalharam com os gêneros *Micrurus* e *Lachesis*, respectivamente. Para outros grupos da herpetofauna neotropical, a realidade não é diferente. Existem trabalhos esparsos e quase que na sua totalidade realizados no exterior (e.g. FAIVOVICH *et al*., 2004). Cabe ressaltar os trabalhos realizados pelo grupo de I. P. Farias da Universidade Federal do Amazonas com quelônios e crocodilianos (FARIAS *et al*., 2004; FARIAS *et al*., no prelo) e os realizados na Universidade Católica de Goiás por K. C. M. Pellegrino e colaboradores com lagartos (PELLEGRINO *et al*., 2001, PELLEGRINO *et al*., 2005).

Essa realidade de carência de laboratórios brasileiros de biologia molecular trabalhando com herpetologia é em parte causada pelos próprios herpetólogos que, buscando realizar trabalhos com colaboração internacional, não procuram parceiros no país, dificultando assim o fortalecimento de grupos brasileiros voltados para essa área. Os problemas que podem ser abordados utilizando dados moleculares auxiliariam em muito o entendimento da biodiversidade brasileira. Muitas espécies da herpetofauna que possuem sistemática duvidosa ou questões acerca do *status* para a conservação podem e devem ser abordadas com diversas ferramentas.

As análises conjuntas entre caracteres morfológicos e moleculares indicam um caminho para a interação entre grupos de pesquisa que trabalham com a herpetofauna do Brasil. Esperamos que a discussão destas técnicas desperte a atenção dos herpetólogos para as aplicações das mesmas em seus problemas de estudo.

Referências Bibliográficas

AVISE, L.C. *Phylogeography: The history and formation of species*. Harvard University press, Cambridge. 2000.

AVISE, J.C.; ARNOLD, J.; BALL; R.M.; BERMINGHAM, E.; LAMB, T.; NEIGEL, J.E.; REEB, C.A. & SARNDERS, N.C. Intraspecific Phylogeography: The Mitochondrial DNA Bridge Between Population Genetics and Systematics. *Annual Review of Ecology and Systematic,* v.18, p.489-522. 1987.

BANDELT, H.J.; FORSTER P. & RÖHL, A. Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution,* v.16, n.1, p.37-48. 1999.

BURBRINK, F.T. Inferring the phylogenetic position of *Boa constrictor* among the Boinae. *Molecular Phylogenetics and Evolution,* v.34, n.1, p.167-180. 2005.

CAVALLI-SFORZA, L.L. & EDWARDS, A.W. Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics.* v.19, n.3, p.233-257. 1967.

CHAPPILL, J. Quantitative characters in phylogenetic analysis. *Cladistics*, v.5, p.217-234.1989.

CLEMENT, M.; POSADA, D. & CRANDALL, K.A. TCS: a computer program to estimate gene genealogies. *Molecular Ecology*, v.9, n.10, p.1687-1659. 2000.

CRANDALL, K.A.; BININDA-EMONDS, O.R.P.; MACE, G.M. & WAYNE, R.K. Considering evolutionary processes in conservation biology. *Trends in Ecolgy and Evolution*, v.15, n.7, p.290295. 2000.

EXCOFFIER, L.; SMOUSE, P.E. & QUATTRO, J.M. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics,* v.131, n.2, p.479-491. 1992.

FAIVOVICH, J.; GARCÍA, P.C.A; ANANIAS, F.; LANARI, L.; BASSO, N.G. & WHEELER, W.C. A molecular perspective on the phylogeny of the *Hyla pulchella* species group (Anura, Hylidae). *Molecular Phylogenetics and Evolution,* v.32, n.3, p.938-950. 2004.

FARIAS, I.P.; DA SILVEIRA, R.; DE THOISY, B.; MONJELÓ, L.A.S.; THORBJARNARSON, J. & HRBEK, T. Genetic diversity and population structure of Amazonian crocodilians. *Animal Conservation*, v.7, p.265-272. 2004.

FARIAS, I.P., JEROZOLIMSKI, A.; VIANA, M.N. & MONJELÓ, L.A.S. Population genetics of the Amazonian tortoises (*Geochelone denticulata*, *G. carbonaria* - Cryptodira: Testudinidae) from the Kayapó Indigenous Reserve. *Journal of Herpetology* (no prelo).

FELSENSTEIN, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, v.7, n.6, p.368376. 1981.

FELSENSTEIN, J. *PHYLIP 3.63*. (http://evolution.genetics.washington.edu/phylip.html). 2004.

FERNANDES, D.S.; FRANCO, F.L. & FERNANDES, R. Systematic revision of the genus *Lachesis* Daudin, 1803 (Serpentes, Viperidae). *Herpetologica*, v.60, n.2, p.245-260. 2004.

FISCHER, R.A. *The Genetical Theory of Natural Selection*. Oxford: Claredon Press, 1930.

GRAUR, D. & LI, W.H. *Fundamentals of Molecular Evolution*. Sunderland: Sinauer Associates. 2000.

GRAZZIOTIN, F.G.; MONZEL, M.;

ECHEVERRIGARAY, S. & BONATTO, S.L. Phylogeography of the Bothrops jararaca complex (Serpentes: Viperidae): past fragmentation and island colonization in the Brazilian Atlantic Forest. *Molecular Ecology*, (no prelo).

GUINDON, S. & GASCUEL, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, v.52, n.5, p.696-704. 2003.

GUTBERLET Jr., R.L. The phylogenetic position of the Mexican black-tailed pitviper (Squamata: Viperidae: Crotalinae). *Herpetologica*, v.54, n.2, p.184-206. 1998.

GUTBERLET Jr., R.L. & HARVEY, M.B. Phylogenetic relationships of New World pitvipers as inferred from anatomical evidence. In: SCHUETT, G.W.; HÖGGREN, M.; DOUGLAS, M.E. & GREENE, H.W. (eds.). *Biology of the Vipers*. Eagle Mountain: Eagle Mountain Publishing, 2002. p.51-68.

GUTBERLET, M.J.; GUTBERLET Jr.; R.L. & SMITH, E. N. CodeThis! Version 1.0. (ftp://ag.arizona.edu/dept/systbiol/issues/50_2/smith). 2000.

HARTL, D.L. & CLARK, A.G. *Principles of population genetics*. Sunderland: Sinauer Associates. 1997.

HASEGAWA, M.; KISHINO, H. & YANO, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal Molecular Evolution,* v.22, n.2, p.160 174. 1985.

HILLIS, D.M. & BULL, J.J. An empirical test of boot strapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*, v.42, p.182-192. 1993.

HILLIS D.M., MABLE B.K. & MORITZ C. Applications of molecular systematics. In: HILLIS, D.M.; MORITZ, C. & MABLE, B.K. (eds.). *Molecular Systematics*. Sunderland: Sinauer Associates, 1996. p.515-544.

HOLDER, M. & LEWIS, P.O. Phylogeny estimation traditional and Bayesian approaches. *Nature Reviews Genetics*, v.4, n.4: 4275-4284. 2003.

HUDSON, R.R. Gene genealogies and the coalescent process. In: FUTUYMA, D. & ANTONOVICS, J. (eds.). *Oxford Surveys in Evolutionary Biology*, v.7. New York: Oxford University Press, 1990. p.1-44.

HUDSON, R.R., SLATKIN, M. & MADDISON, W.P. Estimation of levels of gene flow from DNA sequence data. *Genetics*, v.132, n.2, p.583-589. 1992.

JUKES, T.H. & CANTOR, C.R. Evaluation of protein molecules. In: MUNRO, H.M. (ed.) *Mammalian protein metabolism*. New York: Academic Press, 1969. p. 21-132.

KIMURA, M. Evolutionary rate at the molecular level. *Nature*, v.217, n.129, p.624-626. 1968.

KIMURA, M. A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, v.16, p.111-120. 1980.

KINGMAN, J.F.C. The coalescent. *Stochastic Process and their Applications*, v.13, p.235-248. 1982.

KUHNER, M.K.; YAMATO, J. & FELSENSTEIN, J. Maximum likelihood estimation of population growth rates based on coalescent. *Genetics*, v.149, n.1, p.429-434. 1998.

LOCKHART, P.J.; STEEL, M.A.; HENDY, M.D. & PENNY, D. Recovering evolutionary trees under a more realistic model of sequence evolution. *Molecular Biology and Evolution*, v.11, n.4, p.605612. 1994.

MICKEVICH, M.F. & JOHNSON, M.F. Congruence between morphological and allozyme data in evolutionary inference and character evolution. *Systematic Zoology*, v.25, p.260-270. 1976.

MORITZ, C. Defining 'Evolutionarily Significant Units' for conservation. *Trends in Ecology and Evolution*, v.9, n.10, p.373-375. 1994

MURPHY, R.W. & DOYLE, K.D. Phylogenetics: frequencies and polymorphic characters in genealogical estimation. *Systematic Biology*, v.47, n.4, p.737-761. 1998.

NEI, M. & KUMAR, S. *Molecular Evolution and Phylogenetics*. New York: Oxford University Press. 2000.

NEIGEL, J.E. Is FST obsolete? *Conservation Genetics*, v.3, n.2, p.167173. 2002.

PELLEGRINO, K.C.M.; RODRIGUES, M.T.; YONENAGA-YASSUDA, Y. & SITES Jr., J.W. A molecular perspective on the evolution of microteiid lizards (Squamata, Gymnophthalmidae) and a new classification for the family. *Biological Journal of Linnean Society*, v.74, n.3, p.317-340. 2001.

PELLEGRINO, K.C.M.; RODRIGUES, M.T.; WAITE, A.; Morando, M.; YASSUDA, Y.Y.& SITES Jr., J.W. Phylogeography and species limits in the *Gymnodactylus darwinii* complex (Gekkonidae, Squamata): genetic structure coincides with river systems in the Brazilian Atlantic Forest. *Biological Journal of Linnean Society*, v.85, p.13-26. 2005

PIMENTEL, R.A. & RIGGINS, R. The nature of cladistic data. *Cladistics*, v.3, p.201-209. 1987.

POE, S. Skull characters and the cladistic relationships of the Hispaniolan dwarf twig *Anolis*. *Herpetological Monographs*, v.12, p.192-236. 1998.

POSADA, D. & CRANDALL, K.A. Modeltest: Testing the model of DNA substitution. *Bioinformatics*, v.14, n.9, p.817-818. 1998.

PUORTO, G.; SALOMÃO, M.G.; THEAKSTON, R.D.G.; THORPE, R.S.; WARRELL, D.A. & WÜSTER, W. Combining mitochondrial DNA sequences and morphological data to infer species boundaries: phylogeography of lanceheaded pitvipers in the Brazilian Atlantic forest, and the status of *Bothrops pradoi* (Squamata: Serpentes: Viperidae). *Journal of Evolutionary Biology*, v.14, n.4, p.527-538. 2001.

RODRÍGUEZ, F.; OLIVER, J.; MARÍN, A. & MEDINA, J. The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology*, v.142, n.4, p.485-501. 1990.

SANGER, F.; NICKELEN, S. & COULSON, A.R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the U.S.A.*, v.74, n.12, p.5463-5467. 1977.

SIDALL, M.E. & KLUGE, A.G. Probabilism and phylogenetic inference. *Cladistics*, v.13, n.4, p.313-336. 1997.

SIMPSON, B.B. Quaternary biogeography of the high montane regions of South America. In: DUELLMAN, W.E. (ed.). *The South American Herpetofauna: Its Origin, Evolution, and Dispersal*. Monograph of the Museum of Natural History, University of Kansas, 1979. p. 157188.

SLATKIN, M. Population structure and evolutionary progress. *Genome*, v.31, n.1, p.196-202. 1989.

SMITH, E.N. & GUTBERLET Jr., R.L. Generalized frequency coding: a method of preparing polymorphic multistate characters for phylogenetic analysis. *Systematic Biology*, v.50, n.2, p.156-169. 2001.

SILVA Jr., N.J. & SITES Jr., J.W. Phylogeny of the South America triad coral snakes (Elapidae: *Micrurus*) based on molecular characters. *Herpetologica*, v.57, p.1-22. 2001.

SWOFFORD, D.L. PAUP* Phylogenetic analysis using parsimony and other methods. http://paup.csit.fsu.edu/index.html. 2002.

SWOFFORD, D.L.; OLSEN, G.J.; WADDELL, P.T. & HILLIS, D.M. Phylogenetic inference. In: HILLIS, D.M.; CRAIG, M. & MABLE, B.K. (eds.) *Molecular systematics*. Sunderland: Sinauer Associates. 1996. p.441-501.

TAJIMA, F. Statistical method for testing the neutral mutation hypothesis. *Genetics*, v.123, n.3, p.585-595. 1989.

TAMURA, K. & NEI, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial-DNA in humans and chimpanzees. *Molecular Biology and Evolution*, v.10, n.3, p.512-526. 1993.

TAVARÉ, S. Line-of-descent and genealogical processes and their applications in population genetics models. *Theoretical Population Biology*, v.26, n.2, p.119-164. 1984.

THOMPSON, J.D.; GIBSON, T.J.; PLEWNIAK, F. & JEANMOUGIN, F. The Clustal X Windows interface: flexible strategies for multiple sequence

alignment aided by quality analysis tools. *Nucleic Acids Research*, v.25, n.24, p.4876-4882. 1997.

THIELE, K. The holy grail of the perfect character: the cladistic treatment of morphometric data. *Cladistics*, v.9, p.275-304. 1993.

VINGRON, M. & WATERMAN, M. Sequence alignment and penalty choice review of concepts, case studies and implications. *Journal of Molecular Biology*, v.235, p.1-12. 1994.

WHATTERSON, G.A. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, v.7, p.256-276. 1975.

WHEELER, W. Optimization Alignment: the end of Multiple alignment in Phylogenetics? *Cladistics*, v.12, p.1-9. 1996.

WIENS, J.J. Phylogenetic systematics of three lizards (genus *Urosaurus*). *Herpetologica*, v.44, p.399-420. 1993.

WIENS, J.J. Polymorphic characters in phylogenetic systematics. *Systematic Biology*, v.44, p.482-500. 1995.

WIENS, J.J. Polymorphism in systematics and comparative biology. *Annual Review of Ecology and Systematics*, v.30, p.363-395. 1999.

WIENS, J.J. Character analysis in morphological phylogenetics: problems and solutions. *Systematic Biology*, v.50, p.689-699. 2001.

WILEY, E.O. *Phylogenetics: the theory and practice of phylogenetic systematics*. New York: J. Wiley & Sons. 1981.

WRIGHT, S. Evolution in Mendelian populations. *Genetics*. v.16, n.2, p.97-159. 1931.

WRIGHT, S. The genetical structure of populations. *Annals of Eugenics*, v.15, p.323-354. 1950.

WÜSTER, W.; SALOMÃO, M.G.; DUCKETT, G.J.; THORPE, R.S. & B.B.B.S.P. Mitochondrial DNA phylogeny of *Bothrops atrox* species complex (Squamata: Serpentes: Viperidae). *Kaupia Darmstädter Beiträge zur Naturgeschichte*, v.8, p.135-144. 1999.

WÜSTER, W.; SALOMÃO, M.G.; QUIJADA-MASCAREÑAS, J.A.; THORPE, R.S. & B.B.B.S.P. Origin and evolution of the South American pitviper fauna: evidence from mitochondrial DNA sequence analysis. In: SCHUETT, G.W.; HÖGGREN, M.; DOUGLAS, M.E. & GREENE, H.W. (eds.). *Biology of the vipers*. Eagle Mountain: Eagle Mountain Publishing, 2002. p.111-128.

YANG, Z.; & BIELAWSKI, J.P. Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution*, v.15, n.12, p.496-503. 2000.

ZHANG, J.; KUMAR, S. & NEI, M. Small-sample tests of episodic adaptive evolution: A case study of primate lysozymes. *Molecular Biology and Evolution*, v.14, n.12, p.1335-1338. 1997.

ZAMUDIO, K.R. & GREENE, H.W. Phylogeography of the bushmaster (*Lachesis muta*: Viperidae): implications for Neotropical biogeography, systematics and conservation. *Biological Journal of the Linnean Society*, v.62, p.421-442. 1997.