Análise da Divergência entre Matrizes Baseadas nos Modelos Evolutivos de Distância P, Jukes-Cantor e Kimura dois Parâmetros

Álvaro Eduardo Mascarenhas Ribas[1,2], Ryan Ribeiro de Azevedo[2,3]

[1]Universidade do Estado da Bahia, UNEB CAMPUS VIII

[2]Faculdade Sete de Setembro,

48600-000, Paulo Afonso, BA

alvaro.ribas@gmail.com

[3]Centro de Informática - Universidade Federal de Pernambuco

50732-970, Recife, PE

rra2@cin.ufpe.br

Marcelo José Siqueira Coutinho Almeida[4]

[4]Coordenação de Informática – Centro Federal de Educação Tecnológica da Paraíba

58015-430, João Pessoa, PB

marcelo@cefetpb.edu.br

Resumo: *Reconstruções filogenéticas são representações hipotéticas das relações evolutivas entre várias espécies que podem ter um antepassado comum. Os métodos de reconstrução filogenética podem ser classificados em dois grupos principais, que são: os métodos quantitativos: métodos de distância (pairwise distance) e métodos qualitativos: métodos de parcimônia e máxima verossimilhança. Esse artigo tem como objetivo avaliar o coeficiente de variação (CV) da média das distâncias em uma matriz contendo dados quantitativos baseados nos modelos evolutivos de: Distância P, Jukes Cantor e Kimura dois parâmetros. Essa avaliação é dada nos seguintes parâmetros: Mudança de percentagem de cada base A, C, T, G (Adenina, Citosina, Timina e Guanina) (parametro1), Mudança na quantidade de OTU's (Operational Taxonomic Unit) (parametro2) e Mudança na quantidade de sítios (parametro3). Para o trabalho foram geradas seqüências de nucleotídeos sintéticos usando software Random DNA sequence generator estas, foram geradas nos parâmetros citados.*

Palavras-chave: Bioinformática, Reconstrução filogenética, Modelos evolutivos, Estatística.

1. Introdução

Árvores filogenéticas são representações hipotéticas das relações evolutivas entre várias espécies que podem ter um antepassado comum. Nestas árvores, cada nodo com descendentes, representa o mais recente antepassado comum e os comprimentos dos ramos podem representar estimativas do tempo evolutivo. Cada nodo terminal em uma arvore é chamado de unidade taxonômica, *táxon* ou OTU's que podem ser famílias, gêneros, espécies, populações ou qualquer nível taxonômico. Já os nodos internos são chamados de unidades hipotéticas. As árvores são inferidas a partir de uma matriz contendo os dados disponíveis que podem ser morfológicos, químicos ou genéticos. Esses dados são comparados e os táxons, agrupados pelas semelhanças e diferenças entre si em clados.

Segundo [5] os primeiros critérios objetivos para a reconstrução filogenética baseavam em dados morfológicos. Com o acesso recente à estrutura de macromoléculas (DNA, RNA e proteínas) a análise filogenética passou a ter um avanço vertiginoso.

Nas metodologias envolvendo filogenia molecular, cada posição do nucleotídeo ou aminoácido ocupado na seqüência é considerado como um caráter do tipo multiestado. A variação destes caracteres em seus estados fornecerá informações filogenéticas.

O artigo está estruturado da seguinte forma: a Seção 2 trata do critério de classificação de uma árvore filogenética, a Seção 3 foca nos métodos de reconstrução de árvores filogenéticas, na seção 4 são apresentados os materiais e métodos utilizados nos experimentos. Na Seção 5, os resultados com os experimentos são exibidos e, por fim, a seção 6, conclui e apresenta os trabalhos futuros.

2. Critério de Classificação

O principal critério de classificação de uma árvore filogenética se baseia na presença ou ausência de raiz. Uma árvore com raiz transmite melhor a idéia da ancestralidade por haver hierarquia entre os nós. Já a árvore sem raiz possui uma distinção completa entre ancestrais e entre os nós, além disso, permite determinar quais espécies são mais próximas das

outras. É apresentado na Figura 1 exemplos de árvores com raiz e sem raiz respectivamente.

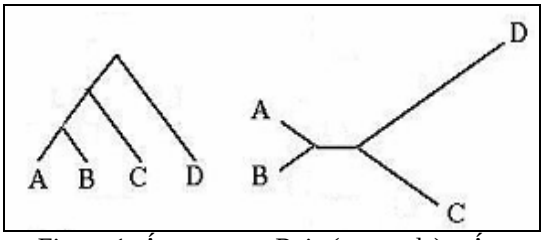

Figura 1: Árvore com Raiz (esquerda) e Árvore sem raiz (direita).

Em todas as metodologias envolvendo filogenia molecular, cada posição do nucleotídeo ou aminoácido ocupada na seqüência é considerada como um caráter do tipo multiestado, que pode ser um dos quatro nucleotídeos ou um dos vinte aminoácidos. Cada caractere é considerado independente dos demais. A variação destes caracteres em seus estados fornecerá informações filogenéticas.

A filogenia ou sistemática filogenética é a ciência que tem como objetivo estudar as relações evolutivas entre grupos de organismos (relações filogenéticas), ou seja, a determinação das relações ancestrais entre as espécies conhecidas. Essa ciência foi proposta por *Willi Hennig* e tem como finalidade testar a validade de grupos e classificações taxonômicas. A sistemática filogenética é uma das bases para a cládistica que é um método de análise das relações evolutivas entre grupos de seres vivos de modo a conseguir a sua genealogia. As árvores filogenéticas ou árvores da vida são os principais métodos utilizados por essa ciência.

3. Métodos de Reconstrução de Árvores Filogenéticas

Segundo [2] existe uma variedade de métodos que propõem a reconstrução de árvores filogenéticas. Ainda segundo [2], os métodos de reconstrução filogenética podem ser classificados em dois grupos principais, que são:

- Métodos quantitativos: *pairwise distance* e;
- Métodos qualitativos: métodos de parcimônia e máxima verossimilhança.

De acordo com [6] o processo de substituições nucleotidicas nas populações poderá ser observado no decorrer do tempo, através das gerações que herdam estas substituições. Uma vez que a taxa de substituições nucleotidicas e a história evolutiva são inferidas a partir deste processo, modelá-lo permite compreender melhor como ocorrem as substituições ao longo do tempo.

Muitos modelos de Substituição de nucleotídeo foram propostos. Segundo [6], Uma característica comum a estes modelos é que eles podem ser representados por uma matriz onde a probabilidade de substituição permanece constante ao longo do tempo e a freqüência das bases se encontra em equilíbrio. É apresentado na Figura 2 uma matriz de substituição.

$$Pt = \begin{bmatrix} pAA & pAC & pAG & pAT \\ pCA & pCC & pCG & pCT \\ pGA & pGC & pGG & pGT \\ pTA & pTC & pTG & pTT \end{bmatrix}$$

Figura 2: Matriz de probabilidade de substituição.

De acordo com [6] na matriz, *pAC* é a probabilidade de que o nucleotídeo *A* seja substituído pelo nucleotídeo *C* ao final do intervalo de tempo *t*, e assim por diante. Os elementos da diagonal *pAA*, *pCC*, *pGG* e *pTT* representam a probabilidade de não ser observada uma substituição de nucleotídeo ao final do tempo *t*.

Os modelos mais utilizados pelos biólogos na reconstrução filogenética são: Distância P, *Jukes-Cantor* e *Kimura* dois parâmetros, os dois métodos citados serão abordados ao longo deste artigo.

3.1 Distância P

O modelo conhecido e mais simples, é o chamado Distância P. Onde o valor é calculado com a observação do número de diferenças (mutações) entre um par de seqüências. Sendo assim, o número de mutações é contado se possui um cálculo estimativo. O método tem como formula:

$$P = (nd/n)$$

Onde *P* é a distância que se quer encontrar, *nd* o numero de diferenças (mutações) e *n* o número de sítios.

3.2 Jukes-Cantor

Jukes Cantor, *T.H.Jukes* e *C. Cantor* propuseram no ano de 1969 que a probabilidade α de um nucleotídeo mudar para outro diferente num intervalo ΔT é proporcional ao tempo transcorrido. Neste método o intervalo ΔT varia de acordo com a necessidade, ou seja, é desconsiderado. Com o tamanho de ΔT desconsiderado, se supõe que a probabilidade de uma determinada posição de uma seqüência sofrer mutação é maior em um intervalo de tempo maior, ou seja, α é diretamente proporcional a ΔT. Segundo [9], o modelo de *Jukes Cantor* tem como idéia principal a suposição que a probabilidade da

mudança de um estado ao outro é sempre igual. O cálculo da distância através desse modelo é feito pela fórmula:

$$ut = -3/4 \ln (1- 4/3p)$$

Onde *ut* é a distância que se quer encontrar e *p* é dado pela divisão da quantidade de posições que sofreram mutações pela quantidade de caracteres. Certos tipos de mutações são mais freqüentes que outras.

3.3 Kimura Dois Parâmetros

As mutações nos nucleotídeos podem ser por Transversão que é o tipo de mutação por substituição de uma purina por uma pirimidina e vice-versa ou Transição que é a substituição no DNA ou RNA de uma purina (Adenina e Guanina) por outra purina, ou de uma pirimidina (Citosina, Timina e Uracila) por outra pirimidina.

Em 1980, *Kimura* propôs um modelo probabilístico para substituição em seqüências de nucleotídeos que leva em conta as diferenças nas taxas de transições e transversões por posição na seqüência de bases no DNA.

O modelo propõe que num intervalo fixo de tempo ΔT em uma determinada posição de uma seqüência de nucleotídeo é α para transição e β para transversão. Assim como em *Jukes-Cantor* α é diretamente proporcional a ΔT. A matriz com as relações αdt e βdt é apresentada na Figura 2.

$$Pij(dt) = \begin{cases} \alpha dt & \text{para transição} \\ \beta dt & \text{para transversão} \end{cases} \quad \text{e sua matriz } Pt = \begin{bmatrix} . & \beta & \alpha & \beta \\ \beta & . & \beta & \alpha \\ \alpha & \beta & . & \beta \\ \beta & \alpha & \beta & . \end{bmatrix}$$

Figura 2: Matriz com as relações αdt e βdt.

O cálculo da distância segundo *Kimura* é dado por:

$$Dt= 1\backslash 2\ln(1/(1-2\,P-Q))+1/4\ln(1/(1-2Q))$$

Onde *Dt* é a distância que se quer encontrar, *P* é igual ao número de transições e *Q* é igual ao número de transversões.

4. Materiais e Métodos

Esse trabalho tem como objetivo avaliar o coeficiente de variação (CV) da média das distâncias em uma matriz triangular contendo dados quantitativos baseados nos modelos evolutivos de: Distância P, *Jukes Cantor* e *Kimura* Dois Parâmetros.

Avaliação do coeficiente de variação foi dada nos seguintes parâmetros: Mudança de percentagem de cada base A, C, T, G (Adenina, Citosina, Timina e Guanina) (parametro1), Mudança na quantidade de *OTU's* (parametro2) e Mudança na quantidade de sítios (parametro3).

O trabalho foi iniciado com o uso do *software Random DNA sequence generator* (Disponível em: http://www.birc.au.dk/fabox/random_sequence_generator.php) que foi usado para gerar seqüências aleatórias que foram agrupadas nos parâmetros citados. Para cada um deles foram geradas quatro amostras com os valores diferentes escolhidos aleatoriamente. Tais valores são:

- No parâmetro Um, cada amostra continha 30 OTU's, totalizando 120 espécies sintéticas. Para a amostra um foi distribuído 25% para cada base. A segunda teve 30% para A e C, 25% para G e 15% para T. A terceira recebeu 25% para A e T, 30% para C e 10% para G. A quarta recebeu 10% para A, 15% para C e 55% para G e 20% para T.

- No parâmetro Dois, foram gerados 20 OTU's para cada amostra no total de 80 espécies sintéticas. A porcentagem de bases não variou ficando 25% para cada. Na amostra um desse parâmetro foram gerados 40 sítios, na dois 50, na três 60 e na quarta 70.

- O parâmetro Três, não sofreu variedade em percentagem, tendo 25% para cada base. Houve variação apenas na quantidade de espécies, na primeira amostra possuía 30, na segunda 40, na terceira 50 e na quarta 60.

Após a coleta das seqüências, deve ser feito um alinhamento com o uso de alguma ferramenta computacional. Esse alinhamento é um pré- requisito antes do uso das seqüências e consiste em alinhar os nucleotídeos homólogos em uma mesma coluna, que é chamada de sitio, tal ação constitui uma matriz de nucleotídeo. Essa ação agrupa os trechos onde há maior similaridade e acabam por destacar os trechos onde há divergências, que seriam as mutações [4].

Com a matriz de nucleotídeo já arrumada em regiões homologas, ou seja, os sítios já organizados vêm à escolha do modelo e método a ser utilizado. Cada amostra foi submetida ao alinhamento em homologia usando o *Clustaw* (Disponível em: http://www.ebi.ac.uk/clustalw/), os parâmetros utilizados são apresentados na Tabela 1.

Tabela 1: Parâmetros usados no *Clustaw*.

Parâmetro	**Valor**
Gap openning penalty	15
Gap extension penalty	6.66
DNA Weight Matrix	*Clustalw* (1.6)
Trasition weight	0.5
Delay divergent cutoff	30%

Após o alinhamento, foi calculada a média das distâncias da matriz de cada seqüência para os modelos de Distância P, *Jukes Cantor* e *Kimura* Dois Parâmetros. Também foram calculados o número de mutações em todas as seqüências, para esse procedimento foi utilizado o pacote *MEGA* (*Molecular evolutionary genetics analysis.* Disponível em: http://www.megasoftware.net/).

Com as médias calculadas, foi gerado um arquivo, contendo as médias das amostras para cada parâmetro cujos valores são apresentados na Tabela 2. Foi utilizado o software *Netbook* (Disponível em: http://www.cin.ufpe.br/~autosim/netbook/) para os cálculos estatísticos.

Tabela2: Médias amostrais calculadas nos modelos estudados.

Amostra	**Média Distância P**	**Média *Jukes-Cantor***	**Média *Kimura* 2**
1	0.704	1.490	1.373
2	0.734	2.087	1.930
3	0.716	1.965	1.834
4	0.624	1.419	1.306

Alem das médias amostrais também foram computados os coeficientes de variação.

5. Resultados

O modelo de distância P, em todos os parâmetros estudados atingiu as menores médias. Os modelos de *Jukes Cantor* e *Kimura* dois parâmetros tiveram valores de média próximos, divergindo principalmente no parâmetro dois, em que as médias foram 2,11 e 1,95 respectivamente. Apesar da aparente singularidade entre os dois, as médias do modelo de *Jukes Cantor* em todos os parâmetros foi a maior. Isso pode ser verificado no Gráfico 1. Esses valores mostram o número médio da distância nos parâmetros escolhidos para os modelos estudados.

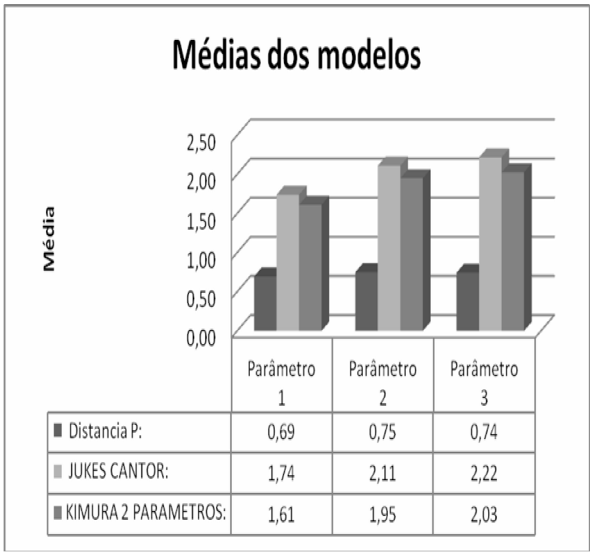

Gráfico 1:Valores médios das Distâncias nos parâmetros escolhidos nos modelos estudados.

Na avaliação do coeficiente de variação, o modelo de distância P, atingiu os menores valores, atingindo um valor próximo de zero (0,002872) no parâmetro 3.

Os valores de CV para *Jukes Cantor* ficaram bem próximos, chegando a ficar com valor iguais no parâmetro 2, tal resultado demonstra que a consistência dos dois métodos é praticamente equivalente nos parâmetros estudados. Os valores são apresentados no Gráfico 2.

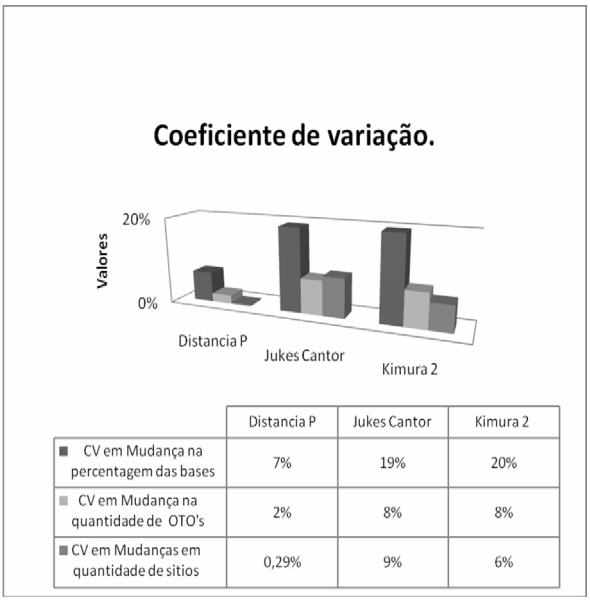

Gráfico 2:Valores de CV dos modelos estudados nos parâmetros escolhidos.

Segundo [1] deve haver valores razoáveis para a medida do CV para a área de atuação da pesquisa, pois ele varia de acordo com a espécie e com a variável resposta estudada. Para [8], é importante que se

conheça na literatura os valores mais freqüentes do CV para a resposta que está sendo estudada. Mas, de acordo com [7] em iguais condições, o CV da uma idéia da precisão do experimento, considerando que o experimento mais preciso é o que possui menor coeficiente de variação.

6. Conclusões e Trabalhos Futuros

Diante dos valores aferidos, pode-se afirmar que nos parâmetros: Mudança de percentagem de cada base A, C, T, G (Adenina, Citosina, Timina e Guanina), mudança na quantidade de OTU's e Mudança na quantidade de sítios, há pouca dispersão entre os modelos de Jukes Cantor e Kimura dois parâmetros.

O estudo mostrou que diante dos parâmetros escolhidos a mudança na percentagem das bases é a característica que dá maior dispersão aos modelos.

Foi concluído também que mudanças na quantidade de sítios provocam dispersão maior em *Jukes Cantor* do que em *Kimura* dois parâmetros. Este, sofre maior dispersão quando a quantidade de sítios é alterada na amostra.

A mudança na percentagem de bases não infere nenhuma dispersão quando o modelo usado for o de Distância P, pois essa, depois da mudança na percentagem de bases sofre maior dispersão quando a quantidade de sítios é alterada.

O estudo comprova que diante dos parâmetros usados, a distância P sofre a menor dispersão em relação a média dos três modelos estudados.

A ausência em referência de CV para análise de modelos de reconstrução filogenética com parâmetros pré-estabelecidos dificulta a classificação deste como baixo ou alto quando comparados um ao outro. Mas, pode-se concluir com o estudo que o modelo de Distância P possui a maior precisão dos três métodos avaliados por possuir os menores CV's.

Agradecimentos

Os autores agradecem à professora Dra. Marcília Andrade Campos do CIN-UFPE – Centro de Informática da Universidade Federal de Pernambuco pela revisão e sugestões de melhoria deste trabalho.

Referências

[1] FEDERER, WT. Experimental design. New York: Ed J wile, 1955.

[2] FELSENSTEIN, J. Phylip Home Page. Phylogeny Programs. Disponivel em : http://evolution.genetics.washington.edu/phylip.html, 2006. Acesso:28/05/2007.

[3] FREUND, E, JOHN. SIMON, A, GARY. Estatística aplicada. Porto Alegre: Ed bookman, 2000

[4] GONÇALVES, GLAUBER DIAS. TORRES, MARTHA. Analise de desempenho de métodos para reconstrução de arvores filogenéticas. Anais ERBASE, 2007.

[5] MATIOLI, FLORA MARIA. Noções de Filogenética molecular. São Paulo 2001.

[6] PEREIRA, LUIZ, SERGIO. Filogenia e evolução molecular em Cracidae (Aves). São Paulo, Tese (Doutorado) - Instituto de Biociências da Universidade de São Paulo, Departamento de Biologia, 2000.

[7] PIMENTEL-GOMES, F. Curso de estatística experimental. 12. ed. São Paulo, Nobel,1987.

[8] SAMPAIO,I.B.M. Estatística aplicada a experimentação animal. Belo Horizonte:FEPMVZ-UFMG,1998.

[9] SANKOFF, DAVID. BLANCHETTE, Mathieu. Comparative genomics via phylogenetic invariants for Jukes-Cantor semigroups. Canadian Mathematical Society,2006.