Relatório da Disciplina
MAC 5701 – Tópicos em Ciência da Computação

Fábio Henrique Viduani Martinez
Orientador: Prof. José Augusto Ramos Soares

02 de julho de 2001

Sumário

1 Introdução **1**

2 Método da Verossimilhança Máxima para Seqüências de DNA **2**

2.1 Visão Geral 3

2.2 Modelos de Evolução de Seqüências 4

2.3 Cálculo da Verossimilhança de uma Árvore 5

2.4 Heterogeneidade de Taxas entre *Sites* 5

2.5 Estimativa de Parâmetros do Modelo 5

2.6 Busca por Árvores Ótimas 6

3 Considerações Finais **6**

1 Introdução

Uma das principais questões da Biologia é tentar explicar o processo evolutivo das espécies existentes e, ainda, de que forma essas espécies se relacionam em termos de ancestrais comuns. A essa tentativa de detectar, descrever e explicar a diversidade biológica existente dá-se o nome de **sistemática**. Essa questão central é respondida geralmente através da construção de árvores — chamadas **árvores filogenéticas** — cujas folhas contêm espécies e cujos nós internos representam ancestrais hipotéticos.

Como freqüentemente não existem informações suficientes acerca dos ancestrais de espécies atuais, há a necessidade de inferir a história evolucionária desses organismos através da reconstrução da sua árvore filogenética, utilizando algum tipo de critério de comparação entre essas espécies.

Árvores filogenéticas têm sido construídas desde o século XIX. Os métodos utilizados até aproximadamente 1960 eram baseados nas observações das características morfológicas das espécies envolvidas ou no exame de registros fósseis coletados. Desde então, com a introdução gradual de técnicas biomoleculares e de formalismos matemáticos na área, diversos métodos para reconstrução de árvores filogenéticas têm sido propostos.

Uma árvore filogenética pode ser construída a partir de espécies, populações, gêneros, seqüências moleculares ou quaisquer outras unidades operacionais taxonômicas (*operational taxonomic units – OTU's*)[1]. Nas folhas de uma árvore filogenética estão representadas OTU's e seus nós internos representam as relações de ancestralidade ou especialização entre essas unidades. Se uma árvore filogenética possui uma raiz, isso significa que todas as OTU's possuem um ancestral comum conhecido ou hipotético.

Um dos objetivos do estudo de filogenias é a recuperação da ordem de especialização entre as espécies de um conjunto, representada pela **topologia**[2] da árvore filogenética correta ou verdadeira. Outro aspecto importante é a **distância** entre um par de nós da árvore, determinada por alguma medida, representando a quantidade relativa de divergência evolucionária entre as espécies sob algum modelo de evolução.

Como mencionado, centenas de métodos diferentes têm sido desenvolvidos para reconstrução de árvores filogenéticas e podem ser basicamente agrupados em duas classes principais: (*i*) métodos baseados em distâncias, que trabalham sobre dados numéricos comparativos entre as OTU's, e (*ii*) métodos baseados em características discretas, que trabalham sobre os estados de cada característica das OTU's envolvidas. Quando, por exemplo, as OTU's representam seqüências de DNA, uma característica pode ser entendida como uma coluna ou *site* das seqüências alinhadas e o conjunto de estados a que cada característica pode pertencer é representado pelo conjunto de nucleotídeos. Um dos métodos baseados em características mais estudados na atualidade é o método da verossimilhança máxima (*maximum likelihood – ML*). Este método para construção de árvores filogenéticas foi nossa principal fonte de estudo nesta disciplina.

2 Método da Verossimilhança Máxima para Seqüências de DNA

O método da verossimilhança máxima foi proposto inicialmente por Fisher [10]. Esse método depende da especificação completa dos dados e do modelo de probabilidades que descreve os dados. A probabilidade de observar os dados tendo assumido um modelo depende também dos parâmetros do modelo. O método da verossimilhança máxima escolhe o valor de um parâmetro que maximiza a probabilidade dos dados observados.

A estimativa de verossimilhança máxima foi primeiro utilizada por Cavalli-Sforza e Edwards [3]. Felsenstein [7, 8] foi o primeiro a utilizar o princípio da verossimilhança máxima para inferência de árvores filogenéticas baseadas em seqüências de nucleotídeos. Depois, a verossimilhança máxima foi também aplicada em seqüências de aminoácidos por Kishino et al. [14] e Adachi e Hasegawa [1].

[1]**Taxonomia** é a ciência da classificação; é um sistema para nomear e organizar espécies em agrupamentos que compartilham características similares. Sistemática e taxonomia são termos equivalentes na literatura.

[2]Alguns autores também usam o termo **padrão de ramificação** da árvore.

2.1 Visão Geral

O método de verossimilhança máxima para construção de árvores filogenéticas procura inferir a história que melhor descreve os dados observados, ou seja, as seqüências de nucleotídeos de entrada. As incógnitas são a topologia e os tamanhos das arestas da árvore. Para aplicar a técnica da verossimilhança máxima, um modelo do processo evolucionário que descreve a conversão de uma seqüência em outra deve ser especificado. Este modelo pode ser completamente definido ou pode conter parâmetros que são estimados das seqüências de entrada. A técnica da verossimilhança máxima para inferência filogenética computa a probabilidade que o modelo evolucionário escolhido produzirá as seqüências de entrada. Filogenias são então inferidas pela escolha das árvores que possuem maiores valores de verossimilhanças.

Os princípios básicos envolvidos no cálculo da verossimilhança de uma árvore são apresentados a seguir. Considere um conjunto de seqüências de nucleotídeos alinhadas de quatro táxons:

	1							j							N
(1)	C	...	G	G	A	C	A	$\boxed{\text{C}}$	G	T	T	T	A	...	C
(2)	C	...	A	G	A	C	A	$\boxed{\text{C}}$	C	T	C	T	A	...	C
(3)	C	...	G	G	A	T	A	$\boxed{\text{A}}$	G	T	T	A	A	...	C
(4)	C	...	G	G	A	T	A	$\boxed{\text{G}}$	C	C	T	A	G	...	C

Para computar a verossimilhança de uma árvore sem raiz apresentada na Figura 1(a), considerando o modelo evolucionário um modelo **reversível**[3], a verossimilhança da árvore é em geral independente da localização da raiz. É então conveniente posicionar a raiz da árvore em um nó interno arbitrário, como mostra a Figura 1(b). Sob a suposição que os *sites*[4] evoluem independentemente, pode-se calcular a verossimilhança para cada *site* separadamente e em seguida combinar esses resultados.

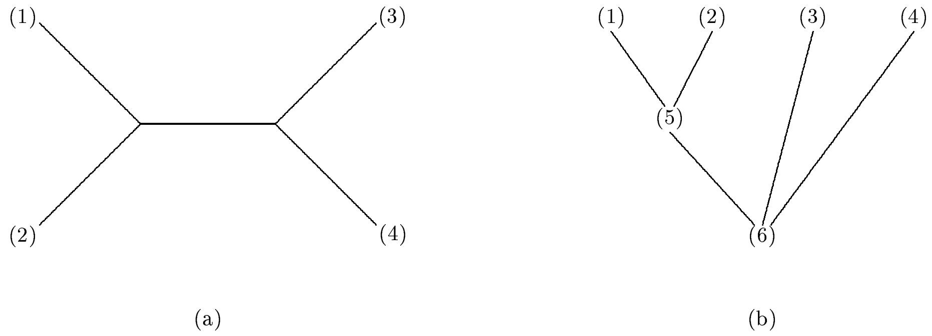
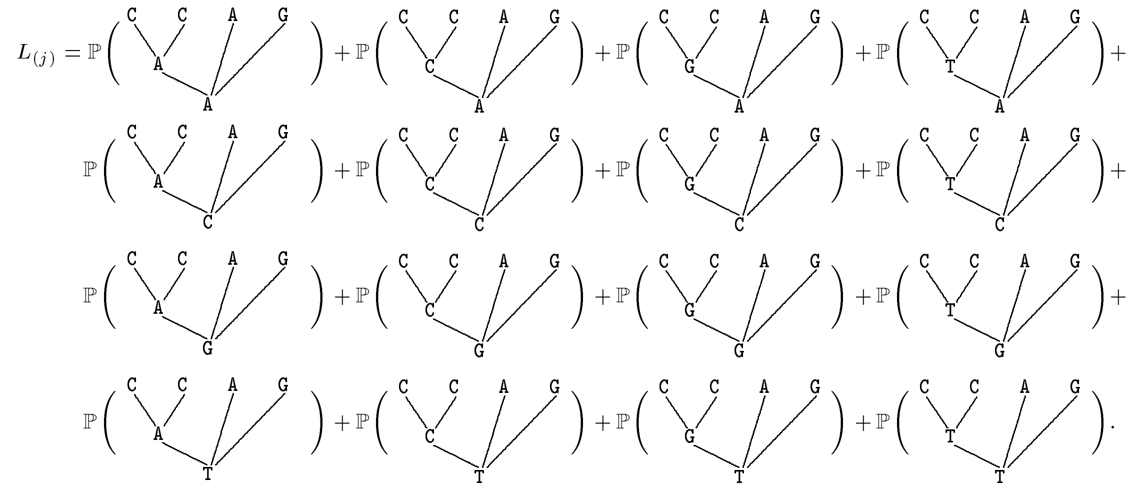


Figura 1: **(a)** Uma árvore binária sem raiz para as quatro táxons cujas seqüências são dadas anteriormente. **(b)** Árvore após o posicionamento da raiz em um nó interno arbitrário.

[3] *time-reversible*.

[4] Um *site* é uma coluna da matriz das seqüências alinhadas, como por exemplo a j-ésima coluna ou o *site* j acima.

Para calcular a verossimilhança para algum *site* j, considera-se todas as possíveis configurações evolutivas:



Considerando que o modelo de evolução é um **modelo de Markov**, as mudanças ao longo das diferentes arestas são independentes.

Calculadas as verossimilhanças para cada *site*, a probabilidade das seqüências descritas por todos os *sites*, dada essa árvore e o modelo específico, é computada como o produto das verossimilhanças nos *sites* individuais:

$$L = L_{(1)} \cdot L_{(2)} \cdot \ldots \cdot L_{(N)} = \prod_{j=1}^{N} L_{(j)}.$$

Em geral, computa-se o logaritmo da verossimilhança, já que as probabilidades envolvidas são números extremamente pequenos:

$$\ln L = \ln L_{(1)} + \ln L_{(2)} + \ldots + \ln L_{(N)} = \sum_{j=1}^{N} \ln L_{(j)}.$$

2.2 Modelos de Evolução de Seqüências

As probabilidades das substituições de bases dependem de várias suposições sobre o processo evolutivo e estas definem o modelo de substituição de bases. O estudo realizado restringiu-se ao modelo de Markov em que a probabilidade de uma substituição de um estado i para um estado j em um dado *site* não depende da história do *site*. Supõe-se também que as probabilidades de substituição não se modificam em diferentes partes da árvore.

A expressão matemática de um modelo de substituição é uma tabela de **taxas**, ou seja, substituições por *site* por unidade de distância evolucionária, em que cada nucleotídeo é trocado por um outro nucleotídeo. Essas taxas podem então ser expressas como uma matriz Q de taxas instantâneas, de dimensão 4×4, em que cada elemento q_{ij} representa a taxa de substituição da base i para base j durante algum período de tempo infinitesimal dt.

Para computar as verossimilhanças são necessárias as probabilidades de susbstituição de qualquer estado para qualquer outro estado ao longo de uma aresta de tamanho t. A **matriz de probabilidade de substituição** P é calculada como

$$P(t) = e^{Qt}.$$

Todos os modelos de substituição de nucleotídeos baseados no modelo de Markov apresentam esta estrutura geral, como, por exemplo, o modelo de de Jukes e Cantor [12] e o modelo de Felsenstein [7], que foram estudados durante a disciplina.

2.3 Cálculo da Verossimilhança de uma Árvore

Para calcular a verossimilhança de uma árvore é necessário considerar as verossimilhanças da ocorrência de cada nucleotídeo em cada nó da árvore como uma função de sua topologia e dos tamanhos de suas arestas. Já que este método define a árvore ótima em termos de um critério de otimalidade, pode-se supor que uma árvore é dada e que deseja-se saber quão boa ela é.

Assim, o método para avaliar a verossimilhança de uma dada árvore inicia a partir de um nó raiz hipotético em uma posição conveniente e combina as verossimilhanças de cada uma de suas sub-árvores esquerda e direita, recursivamente. Como trata-se de um modelo reversível, a escolha da localização da raiz não deve alterar a verossimilhança da árvore.

Nessa discussão, supõe-se implicitamente que os tamanhos das arestas são conhecidos, mas estes em geral não são conhecidos e devem ser estimados como parte do proceso de computar uma verossimilhança.

2.4 Heterogeneidade de Taxas entre *Sites*

Heterogeneidade de taxas de evolução podem ser incorporadas no cálculo da verossimilhança de uma árvore pela inclusão de taxas relativas no modelo. Diversos modelos de heterogeneidade de taxas entre *sites* têm sido desenvolvidos, considerando que os *sites* são log-normalmente distribuídos, considerando uma classe invariante de taxas, estimando as taxas em diferentes porções das seqüências separadamente ou utilizando uma combinação de diferentes distribuições de taxas. Provavelmente, o modelo mais amplamente usado é aquele que supõe que as taxas são gama-distribuídas [30, 31].

Uma vantagem da técnica da verossimilhança é que os modelos podem se tornar mais complexos para incorporar outros processos biologicamente importantes. Por exemplo, os modelos de substituição podem ser modificados para explicar eventos de inserção e remoção em alinhamentos, estrutura secundária de proteínas e taxas evolucionárias correlatas entre *sites* [9, 18, 32].

2.5 Estimativa de Parâmetros do Modelo

O modelo geral descrito acima contém uma variedade de parâmetros que devem ser estimados dos dados de entrada ou fornecidos com base em evidências. Esses parâmetros consistem da topologia da árvore, estimativas dos tamanhos das arestas, parâmetros de taxas relativas do modelo de substituição, parâmetros de freqüências de bases e parâmetros utilizados na modelagem da heterogeneidade de taxas.

Idealmente, uma busca por valores globalmente ótimos deste parâmetros é realizada em um espaço n-dimensional de parâmetros. Isto é, toda possível árvore é considerada e então otimizam-se todos os parâmetros do modelo simultaneamente para cada árvore, escolhendo a árvore resultante com maior valor de verossimilhança. Para uma dada árvore, pode-se executar uma otimização multidimensional usando o método de Newton-Raphson [6]. Esta técnica é de difícil implementação já que necessita do conhecimento da primeira e segunda derivadas parciais da função de verossimilhança com respeito a cada um dos parâmetros. O programa `DNAML` do pacote `PHYLIP`, distribuído por Felsenstein [8] contém a implementação deste método. Métodos alternativos são também encontrados na literatura, como aquele implementado no programa `PAUP`* [28].

2.6 Busca por Árvores Ótimas

Métodos para reconstrução de árvores filogenéticas que possuem um critério de otimalidade explícito, como o método da verossimilhança máxima, separam o problema de avaliar uma árvore particular sob o critério de otimalidade, do problema de encontrar árvores ótimas. Até aqui, a descrição apresentada trata do primeiro problema.

Observe que uma árvore binária sem raiz com N nós folhas tem $N-3$ nós internos e $2N-3$ arestas. O número total de árvores binárias sem raiz com N nós folhas é

$$B(T) = \prod_{i=3}^{N} (2i-5).$$

Dessa forma, a busca exaustiva de possíveis árvores é inviável computacionalmente.

Algumas heurísticas têm sido desenvolvidas com o objetivo de superar essa dificuldade. Dentre elas, pode-se citar **adição passo a passo** contida nos programas `PHYLIP` [8] e `PAUP`* [28], **decomposição em estrela** utilizada no programa `NucML` do pacote `MOLPHY` [1], **troca entre vizinhos mais próximos** (ou *Nearest Neighbor Interchange*) também contida no `PAUP`* [28] e ***quartet puzzling*** contida no `TREE-PUZZLE` [26, 25] .

3 Considerações Finais

Durante a disciplina, estudamos o método da verossimilhança máxima baseado em seqüências de nucleotídeos proposto por Felsenstein [7, 22]. Para obter uma visão geral da área e também do método, utilizamos alguns livros e revisões [5, 11, 13, 16, 23, 25, 29]. O programa `DNAML` do pacote `PHYLIP` [8] foi estudado com alguma profundidade, mas sua compreensão global ainda está aquém de nossas expectativas. Apenas uma visão superficial dos programas `NucML` do pacote `MOLPHY` [1] e `TREE-PUZZLE` [26, 25] foi obtida.

A compreensão do programa `DNAML` deverá ser completada pela melhor compreensão e estudo dos trabalhos relacionados à heterogeneidade e correlação das taxas evolucionárias entre os *sites*

[9, 18, 32]. O estudo de outros aspectos importantes, tais como modelos de substituição de bases, busca por árvores ótimas e teste de confiabilidade devem completar o estudo para compreensão total do método de verossimilhança máxima.

A paralelização dos algoritmos existentes, tal como em [17], nos parece um próximo passo natural a ser seguido.

Referências

[1] Adachi, J., and Hasegawa, M. MOLPHY: Programs for molecular phylogenetics I — PROTML: maximum likelihood inference of protein phylogeny. Computer Science Monographs 27, 1992. Institute of Statistical Mathematics, Tokyo.

[2] Becker, Y., and Darai, G. *Molecular Evolution of Viruses – Past and Present – Evolution of Viruses by Acquisition of Cellular RNA and DNA*. Kluwer Academic Publishers, December 2000.

[3] Cavalli, L. L., and Edwards, A. W. F. Phylogenetic analysis: models and estimation procedures. *Evolution 32* (1967), 550–570.

[4] Chang, J. T. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Mathematical Biosciences 137* (1996), 51–73.

[5] Clote, P., and Backofen, R. *Computational Molecular Biology: an Introduction.* John Wiley & Sons, Ltd, August 2000, ch. All About EVE.

[6] Edwards, A. W. F. *Likelihood.* Cambridge University Press, Cambridge, 1972.

[7] Felsenstein, J. Evolutionary tress from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution 17* (1981), 386–376.

[8] Felsenstein, J. PHYLIP: Phylogeny inference package (version 3.5c). Distributed by the author, 1993. Department of Genetics, University of Washington, Seattle.

[9] Felsenstein, J., and Churchill, G. A. A hidden Markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution 13* (1996), 93–104.

[10] Fisher, R. A. On the mathematical foundations of theoretical statistics. *Phylosophical Transactions of the Royal Society of London – Series A 68* (1922), 222–309.

[11] Huelsenbeck, J. P., and Crandall, K. A. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Reviews on Ecology and Systematics 28* (1997), 437–466.

[12] Jukes, T. H., and Cantor, C. R. Evolution of protein molecules. In *Mammalian Protein Metabolism*, H. M. Munro, Ed. Academic, New York, 1969, pp. 21–132.

[13] Kim, J., and Warnow, T. Tutorial on phylogenetic tree estimation. In *The Seventh International Conference on Intelligent Systems for Molecular Biology* (Heidelberg, Germany, August 1999), T. Lengauer, R. Schneider, P. Bork, D. Brutlag, J. Glasgow, H.-W. Mewes, and R. Zimmer, Eds., The American Association for Artificial Intelligence Press.

[14] Kishino, H., Miyata, T., and Hasegawa, M. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution 31* (1990), 151–160.

[15] Leitner, T., Escanilla, D., Franzén, C., Uhlén, M., and Albert, J. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Evolution 93* (October 1996), 10864–10869.

[16] Li, W.-H. *Molecular Evolution.* Sinauer Associates, Inc., 1997.

[17] Matsuda, H., Yamashita, H., and Kaneda, Y. Molecular phylogenetic analysis using both DNA and amino acid sequence data and its parallelization. Department of Computer and Systems Engineering, Faculty of Engineering, Kobe University, Japan.

[18] Olsen, G. J., Matsuda, H., Hagstrom, R., and Overneek, R. fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Computer Applications in the Biosciences – CABIOS 10*, 1 (1994), 41–48.

[19] Page, R. D. M., and Holmes, E. C. *Molecular Evolution: A Phylogenetic Approach.* Blackwell Science Inc, October 1998.

[20] Penny, D., Hendy, M. D., and Steel, M. A. Progress with methods for constructing evolutionary trees. *Trends in Ecology and Evolution – TREE 7*, 3 (March 1992), 73–78.

[21] Rannala, B., and Yang, Z. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of Molecular Evolution 43* (April 1996), 304–311.

[22] Saitou, N. Property and efficiency of the maximum likelihood method for molecular phylogeny. *Journal of Molecular Evolution 27* (1988), 261–273.

[23] Setúbal, J., and Meidanis, J. *Introduction to Computational Molecular Biology.* PWS Publishing Company, 1997.

[24] Siddall, M. E., and Kluge, A. G. Probabilism and phylogenetic inference. *Cladistics 13* (July 1997), 313–336.

[25] Strimmer, K. S. *Maximum Likelihood Methods in Molecular Phylognetics.* PhD thesis, Fakultät für Biologie der Ludwig-Maximilians – Universität Müchen, October 1997.

[26] Strimmer, K. S., and von Haeseler, A. Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution 13*, 7 (1996), 964–969.

[27] Strimmer, K. S., and von Haeseler, A. *The Phylogenetic Handbook.* Cambridge University Press, Cambridge, UK, 2001, ch. Probabilistic Models of Nucleotide Substitution.

[28] Swofford, D. L. *PAUP*: Phylogenetic Analysis Using Parsimony (and Others Methods), Version 4.0.* Sinauer Associates, Inc., Sunderland, MA, 1996.

[29] Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. *Molecular Systematics*, 2 ed. Sinauer Associates, Inc., Cambridge, UK, 1996, ch. Phylogenetic Inference, pp. 407–514.

[30] Yang, Z. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution 10* (1993), 1396–1401.

[31] Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution 39* (March 1994), 306–314.

[32] Yang, Z. A space-time process model for the evolution of DNA sequences. *Genetics 139* (1995), 993–1005.

[33] Yang, Z. Phylogenetic analysis using parsimony and likelihood methods. *Journal of Molecular Evolution 42* (1996), 294–307.

[34] Yang, Z. Maximum likelihood analysis of adaptive evolution in HIV-1 GP120 ENV gene. In *Pacific Symposium on Biocomputing* (2001), pp. 226–237. (pre-print).