João Lídio da Silva Gonçalves Vianez Júnior

AVALIAÇÃO CRITERIOSA DAS SEQÜÊNCIAS DOS GENES *rrn*, *rpoB* E *gyrB* COMO FERRAMENTAS EM TAXONOMIA MICROBIANA

Monografia apresentada ao Instituto de Microbiologia Prof. Paulo de Góes, como pré-requisito para a obtenção do grau de Bacharel em Microbiologia e Imunologia

INSTITUTO DE MICROBIOLOGIA PROFESSOR PAULO DE GÓES
UNIVERSIDADE FEDERAL DO RIO DE JANEIRO – UFRJ
RIO DE JANEIRO
2005

VIANEZ JÙNIOR, João Lídio da Silva Gonçalves;

Avaliação criteriosa das seqüências dos genes *rrn*, *rpoB* e *gyrB* como ferramentas em taxonomia microbiana, Rio de Janeiro, Instituto de Microbiologia Professor Paulo de Góes / UFRJ, 2005;

IX 69p
Monografia: Bacharel em Microbiologia e Imunologia
1. Bioinformática
2. Taxonomia
3. Enterobactérias
4. Evolução
4. *gyrB*
5. *rpoB*
6. 16S rRNA
I. Universidade Federal do Rio de Janeiro
II. Avaliação criteriosa das seqüências dos genes *rrn*, *rpoB* e *gyrB* como ferramentas em taxonomia microbiana

Trabalho realizado no Departamento de Microbiologia Geral, do Instituto de Microbiologia Prof. Paulo de Góes, UFRJ, sob a orientação do Professor Dr. Andrew Macrae.

AGRADECIMENTOS

Agradeço primeiramente a benção da vida, e da condição humana, que me permitiram ter prazer em realizar meus deveres e obrigações. Agradeço a toda minha família, pelos exemplos e pelo suporte material que me forneceram. Agradeço a todos os meus amigos pelos momentos que me proporcionaram, os quais foram fundamentais em minha vida.

Agradeço especialmente ao meu amigo Felipe Assunção por me apoiar em todos os momentos e respeitar minhas decisões.

Agradeço a minha namorada, Fabíola Leoni, por estar ao meu lado incondicionalmente, pensar sempre em meu bem estar, me trazer paz e alegria e por tudo o que me fez aprender. Agradeço a ela por me amar e por me fazer ter a certeza de que minha felicidade segue a dela.

Agradeço ao meu orientador, Andrew Macrae pela inspiração, pela boa instrução e apoio, fundamentais ao meu desenvolvimento. Agradeço aos membros de minha banca, Prof. Fernando Portela e Fábio Mota por seu apoio e disponibilidade.

Agradeço especialmente a Profa. Bernadete Carvalho por sua dedicação como coordenadora e por sua sensibilidade comigo.

Agradeço de coração a minha amiga Fernanda Machado Pereira, por estar ao meu lado e iluminar minha vida em um momento onde isto se fazia extremamente necessário. Agradeço pelos momentos e reflexões proporcionados. Foram coisas imprescindíveis para chegar onde estou.

Agradeço ao Rio de Janeiro por sua beleza tranqüilizante e por suas noites alegres e relaxantes, sem as quais tudo ficaria mais difícil.

Meus sinceros agradecimentos a todos.

ÍNDICE

RESUMO viii
SUMMARY ix

1. **INTRODUÇÃO** 1
1.1 Introdução geral 1
1.1.1 Primeiras classificações da vida 1
1.1.2 Primeiras tentativas de se determinar relações filogenéticas entre microrganismos 2
1.1.3 Classificações posteriores 2
1.1.4 Procariotos X eucariotos 3
1.1.5 Seqüenciamento de ácidos nucléicos e a nova classificação da vida 4
1.2 Taxonomia bacteriana 4
1.2.1 Sorologia e quimiotaxonomia 5
1.2.2 Taxonomia numérica 6
1.2.3 Taxonomia molecular 6
1.3 Bioinformática 6
1.4 Filogenia molecular 8
1.4.1 Cronômetros moleculares 9
1.4.2 Alinhamento múltiplo 10
1.4.2.1 Alinhamento múltiplo em genes codificadores de proteína 11
1.4.3 Modelos de distância 12
1.4.3.1 Distância p 12
1.4.3.2 Jukes & Cantor 13
1.4.3.3 Kimura 2 parâmetros 13
1.4.3.4 Outros modelos 14
1.4.4 Conceito de árvores filogenéticas 14

1.4.5 Reconstrução de árvores filogenéticas 16
1.4.5.1 Reconstrução das árvores por
métodos de distância 16
1.4.5.1.1 Fitch e Margoliash (1967) …..... 18
1.4.5.1.2 Neighbour – Joining 21
1.4.5.2 Reconstrução de árvores filogenéticas
por métodos de caráter 22
1.4.5.2.1 Maximum Parsimony 22
1.4.5.2.2 Maximum Likelihood 23
1.4.6 Avaliação das árvores 24
1.4.7 Considerações Filogenéticas 24
1.4.8 Diferenças entre árvores de genes
e árvores de espécies 25
1.5 Genes escolhidos 25
1.5.1 *rpoB* 26
1.5.2 *rrn* 27
1.5.3 *gyrB* 29
1.6 Objetivo geral 31
1.6.1 Teoria da evolução endosimbiótica 31
1.6.2 Enterobactérias 32
1.7 Objetivos específicos e hipóteses 33

2. **MATERIAL E MÉTODOS** 34
2.1 Seqüências 34
2.2 Alinhamentos múltiplos 36
2.3 Correção das distâncias e reconstrução
das árvores filogenéticas 37
2.4 Avaliação das árvores 38
2.5 Matrizes de distância 38

3. **RESULTADOS** 39
3.1 Avaliação das técnicas de filogenética para testar a hipótese da evolução endosimbiótica utilizando seqüências dos genes *rrn* e *rpoB* 39
3.2 Genes *rpoB* e *gyrB* como ferramentas taxonômicas alternativas ao *rrn* no grupo das proteobactérias 42
3.3 Genes *rpoB* e *gyrB* como ferramentas taxonômicas alternativas ao *rrn* na família das *Enterobacteriacea* 46
3.3.1 Matrizes de distância e gráficos de entropia 46
3.3.2 Análises Filogenéticas 51

4. **DISCUSSÃO** 57
4.1 Gene *rrn* 57
4.2 Gene *rpoB* 57
4.3 Gene *gyrB* 58
4.4 Considerações finais 58
5. REFERÊNCIAS BIBLIOGRÁFICAS 61

RESUMO

Espécies bacterianas são definidas com base em diversos caracteres fenotípicos, relações DNA-DNA e mais recentemente com base no seqüênciamento do 16S rRNA. Porém é sabido que o 16S rRNA não é polimórfico o suficiente para diferenciação e inferências filogenéticas confiáveis em espécies muito relacionadas. Seqüências completas dos genes 16S rRNA, *rpoB* e *gyrB* de 52 espécies bacterianas e cloroplastos foram retiradas do Genbank e analisadas por três métodos; Neighbor Joining, Maximum Parsimony e Maximum Likelihood, para acessar sua utilidade como ferramentas em taxonomia e filogenia microbiana. Os resultados mostram que os genes *rpoB* e *gyrB* são mais polimórficos que o 16S rRNA. Em todos os casos, a porcentagem de similaridade entre diversas espécies foi menor utilizando os genes *rpoB* e *gyrB*, mesmo para espécies com alto grau de similaridade. Seqüências de 16S rRNA e *rpoB* também suportaram a hipótese de evolução dos cloroplastos a partir das cianobactérias. Testes *in silico* foram realizados com os três genes em membros do grupo das proteobacterias, e mais especificamente, com membros da familia das *Enterobacteriacea*. Vários clusters suportados pelos três métodos analíticos com altos valores de bootstrap foram observados utilizando-se os genes *gyrB* e *rpoB*. Seqüências dos genes *rpoB* e *gyrB* foram consideradas melhores que o 16S rRNA para diferenciação entre membros da família das *Enterobacteriacea*, a qual possui importância clínica.

SUMMARY

Bacterial species are defined on the basis of several phenotypic characters, DNA-DNA relatedness and, more recently, 16S rRNA gene sequencing. It is known, however, that 16S rRNA is not polymorphic enough for reliable phylogenetic inferences and differentiation of closely related species. Complete sequences of 16S rRNA, *rpoB* and *gyrB* of 52 bacterial species and chloroplasts were retrieved from Genbank and analyzed by three methods; Neighbor Joining, Maximum Parsimony and Maximum Likelihood, to assess their usefulness in bacterial taxonomy and phylogeny. The results show that *rpoB* and *gyrB* genes are more polymorphic than 16S rRNA. In all cases the percent similarities between different species were lower than those observed in 16S rRNA, even for species with a high degree of similarity. 16S rRNA and *rpoB* sequences strongly support a hypothesis of the evolution of chloroplasts from cyanobacteria. Further *in silico* tests were carried out with the three genes on members of the proteobacteria group, and more specifically, in the *Enterobacteriacea* family. Several phylogenetic clusters were strongly supported by high bootstrap values for all analytical methods. We show that sequences of *rpoB* and *gyrB* are better choices than 16S rRNA for differentiating between members of the *Enterobacteriaceae* family which is of clinical importance.

1. INTRODUÇÃO

1.1 Introdução Geral

1.1.1 Primeiras classificações da vida

Desde a antiguidade o homem tenta estabelecer relações entre os seres que o rodeiam. Uma das primeiras distinções que o homem fez ao observar a natureza foi a classificação dos seres vivos em plantas e animais. Por isso, toda a biologia foi dividida nas ciências da zoologia e da botânica, que estudavam os reinos *Animalia* e *Plantae* respectivamente. Com estes critérios, os fungos e bactérias foram incluídos no reino *Plantae*, pois apesar de não terem clorofila apresentavam parede celular. Já os protozoários foram considerados como pertencentes ao reino *Animalia* por se alimentarem por ingestão e serem móveis.

A inclusão de todos os organismos conhecidos nestes dois reinos teve seus primeiros problemas com a descoberta de organismos fotossintéticos móveis, como a alga unicelular *Euglena*, e de seres fixos heterotróficos, como as anêmonas e os corais. Assim sendo, o homem sempre tentou estabelecer relações entre os seres vivos, classificando-os em grupos e tentando determinar suas origens. Isto foi feito com certa facilidade pelos biólogos que estudavam os seres macroscópicos, como as aves, os mamíferos e as plantas. Estes apresentam uma grande riqueza de detalhes morfológicos, alguns dos quais podiam ser considerados como complexos demais para ocorrerem em dois organismos por acaso (Woese, 1987). Tais detalhes acabaram por servir como base para sua classificação filogenética (Woese, 1998).

Com a invenção dos primeiros microscópios e a descoberta dos microrganismos, a classificação baseada em características morfológicas tornou-se insegura.

1.1.2 Primeiras tentativas de determinar relações filogenéticas entre microrganismos

Os biólogos dos últimos séculos estavam tão preocupados em estabelecer relações evolutivas entre microrganismos quanto entre organismos macroscópicos. O estudo destas relações foi uma das principais forças na microbiologia na primeira metade do século passado (Woese, 1987). Este contexto serviu como pano de fundo para o surgimento de obras como o Bergey´s Manual (Holt *et al*., 1984), uma tentativa de estabelecer estas relações, adotando para as bactérias o mesmo sistema usado para agrupar os animais e plantas filogenéticamente.

Porém, a busca de uma filogenia bacteriana deste modo não foi bem sucedida, pois os microrganismos possuem uma morfologia muito simples, em contraste com os animais e as plantas (Mayr, 1998). Foram desenvolvidos esquemas complicados e falhos numa tentativa de classificar estes organismos (Woese, 1998), o que acabava complicando ainda mais o problema. O clássico exemplo resultante deste sistema foi a criação do grupo pseudomonas, composto de pelo menos de cinco grupos distintos de microrganismos (Palleroni, 2003). A própria origem da palavra pseudomonas reflete esta confusão (do grego *pseudes* e *monas*, que significa falsa unidade). Isto acabou diminuindo a importância do estudo destas relações perante os microbiologistas como um todo.

1.1.3 Divisões posteriores

Em 1866, Haeckel propôs um terceiro reino (*Protista*), onde foram incluídos todos os organismos de posição incerta, como as eubactérias, cianobactérias e protozoários. No entanto, apenas no início do século XX o reino *Protista* foi seriamente considerado, originando a classificação dos seres vivos em 3 grupos primários: plantas, animais e protistas.

No século XIX sabia-se da existência de organismos sem núcleo visível, mas apenas na segunda metade do século XX estes foram separados dos restantes em um único reino contendo todos os procariontes, o reino *Monera* (Copeland, 1938). A meio do século XX ter-se-ia, então, os reinos *Monera*, *Protista*, *Plantae* e *Animalia*.

Esta classificação foi modernizada criando-se um reino somente para os fungos (*Fungi*) originando a clássica divisão nos cinco reinos (Whittaker, 1969).

1.1.4 Procariotos x Eucariotos

A divisão da vida em procariotos e eucariotos (Chatton, 1937) foi baseada em uma característica celular, de se possuir ou não um núcleo verdadeiro, ou seja, circundado por uma membrana. Esta característica pode ser observada por microscopia eletrônica e pode também ser definida por estudos moleculares. Isto levou a consolidação desta visão da divisão primária da vida.

Porém, esta divisão em procariotos e eucariotos define um grupo pela ausência de características inerentes a outro (tudo o que não tivesse um núcleo bem definido era considerado procarioto). Implicando que todos os procariotos teriam que ser de um só tipo, embora a diversidade fisiológica e morfológica entre as bactérias fosse bem clara (Woese, 1987).

Além disso, o próprio conceito de bactéria, indiscutivelmente imprescindível para se determinar uma filogenia bacteriana, nunca havia sido bem definido. Uma definição baseada em características próprias, e não na ausência delas, era necessária.

Mas para isso são precisos estudos comparativos, os quais sempre foram escassos. Além disso, só é conhecida uma pequena fração da real quantidade de espécies de microrganismos existentes (Torsvik, Josten & Frida, 1990), e das poucas conhecidas, somente algumas foram bem estudadas. Posteriormente, se admitiu que certas propriedades de alguns microrganismos razoavelmente bem estudados (como a *E. coli*) eram propriedades dos procariotos em geral, resultando em mais incertezas.

Então, até pouco tempo atrás, era considerado ser impossível determinar relações filogenéticas microbianas, (Woese, 1998), e foi dado como fato os procariotos serem todos relacionados e pertencentes a um mesmo grupo filogenético.

1.1.5 Tecnologia de seqüenciamento de ácidos nucléicos e a nova divisão da vida

Uma forma encontrada para se determinar relações evolutivas entre os organismos, foi deixar de lado o modelo fenotípico (morfológico e fisiológico), por um modelo genotípico unidimensional. No modelo unidimensional das seqüências, conceitos como similar ou parecido (que são subjetivos) podem ser substituídos. Os elementos de uma seqüência podem ser avaliados de forma mais simples e direta, pois são restritos em número e bem definidos.

Através da comparação de seqüências de RNA ribossomal (rRNA) foi provado ser possível determinar uma filogenia microbiana (Woese, 1987). Com esta abordagem, todos os taxa dos procariotos, assim como as relações entre os mesmos e com os eucariotos se revelaram. A filogenia universal, resultante destas análises, demonstrou que o grupo dos procariotos na verdade engloba dois grupos distintos de organismos. O novo grupo identificado (arqueobactérias) é suportado por evidencias fenotípicas, além de estar mais próximo dos eucariotos (Keeling & Doolitte, 1995).

1.2 Taxonomia Bacteriana

A taxonomia engloba classificação, nomenclatura e identificação. A classificação é o arranjo dos organismos em grupos taxonômicos (taxa) baseando-se em suas similaridades ou relações. Nomenclatura é a designação de nomes a grupos taxonômicos de acordo com as normas internacionais. Identificação é o processo pelo qual se determina que um novo isolado pertence a um dos taxa estabelecidos. Já a sistemática procura organizar os diferentes tipos de bactérias em um sistema útil e coerente (Krieg, 1989).

Historicamente, a classificação das bactérias é baseada em características fenotípicas como morfologia, cultivo, nutrição, bioquímica, metabolismo, patogenicidade, sorologia e ecologia. Vários métodos baseados nessas propriedades foram desenvolvidos para classificar bactérias.

1.2.1 Sorologia e quimiotaxonomia

A sorologia e a quimiotaxonomia são métodos utilizados para investigar a arquitetura molecular da célula bacteriana. Técnicas sorológicas dependem da habilidade dos constituintes das células bacterianas se comportarem como antígenos, levando a produção de anticorpos em animais (Jones & Krieg, 1984). Tais técnicas incluem aglutinação, precipitação, fixação de complemento e imunofluorescência.

Basicamente, a taxonomia por sorologia pode ser dividida em duas classes: (a) aquelas que detectam diferenças ou similaridades entre as bactérias com base de sua superfície celular e de seus complementos antigênicos (flagelo, pili, parede celular, cápsula, etc.) e (b) o uso de anticorpos específicos para enzimas purificadas, permitindo verificar similaridades estruturais entre proteínas homologas de diferentes bactérias (Jones & Krieg, 1984).

A quimiotaxonomia engloba técnicas que permitem elucidar a composição química da célula bacteriana ou de partes dela. Tais estudos incluem análises da composição da parede celular, composição dos lipídios, análise de quinonas, composição do citocromo, caracterizações enzimáticas e análises dos produtos de fermentação (Jones & Krieg, 1984).

1.2.2 Taxonomia numérica

A grande quantidade de dados e tabelas resultantes da analises de propriedades fisiológicas e bioquímicas, entre outras, não podem ser prontamente analisadas a olho nu. A taxonomia numérica (ou taxometria) foi desenvolvida na década de 50 (Sneath, 1989), como forma de analisar grandes quantidades de dados com auxilio de computadores, tendo como função principal derivar esquemas objetivos para a classificação das bactérias a partir de dados oriundos de análises taxonômicas, filogenéticas e sorológicas entre outras. Classicamente, a taxonomia numérica primariamente avalia relações fenotípicas, e tem como básico a definição de taxoespécies (um grupo de estirpes com alta similaridade fenotípica).

1.2.3 Taxonomia molecular

As técnicas da biologia molecular possibilitaram uma caracterização taxonômica mais apurada dos microrganismos. Como complemento da descrição dos caracteres fenotípicos convencionais (e.g., coloração de Gram), a análise molecular pode permitir a caracterização mais objetiva e fundamentada dos microrganismos.

A taxonomia molecular envolve o estudo dos ácidos nucléicos microbianos, principalmente DNA cromossômico e RNA ribossômico, para a obtenção de informações taxonômicas (Goodfellow & O´Donnell, 1993). As informações derivadas de ácidos nucléicos podem ser empregadas na classificação de linhagens microbianas em diversos níveis taxonômicos hierárquicos, desde o estabelecimento de relações intra-específicas entre linhagens até relações entre espécies, gêneros e níveis taxonômicos supra genéricos.

1.3 Bioinformática

Em face do crescimento exponencial do número de seqüências disponíveis, era inevitável o recurso às bases de dados informatizadas, como forma de guardar, organizar e indexar os dados, bem como a utilização e desenvolvimento de ferramentas especializadas para a visualização e análise. A bioinformática é uma área interdisciplinar da ciência aplicada que engloba a biologia, a ciência da computação, a química, a física e a matemática, tendo como objetivo principal a capitalização de tecnologias emergentes, aplicadas à investigação biológica (Lesk, 2002).

Reconhecida há pelo menos 15 anos, é uma ciência que fornece ferramentas que visam aproveitar ao máximo estas informações com o auxílio dos computadores, seja desenvolvendo e aplicando novos algoritmos computacionais à problemas existentes na biologia, seja identificando novos genes, ou ainda procurando por padrões existentes que facilitem estas tarefas. Outros exemplos do uso da bioinformática incluem os alinhamentos de seqüências (tanto de nucleotídeos quanto de aminoácidos), predição filogenética, procura por similaridade das seqüências em bancos de dados e a classificação e predição de estruturas tridimensionais de proteínas. Estas ferramentas são apresentadas na forma dos diversos bancos de dados

e softwares existentes (muitos dos quais são gratuitos) capazes de realizar as mais diversas tarefas.

Vários fatores têm contribuído para o avanço da bioinformática. O advento da World Wide Web (WWW) possibilitou a partilha das informações em bancos de dados ou páginas da web que podem ser acessados instantaneamente de qualquer lugar do mundo. A potencialização dos computadores, principalmente nas últimas décadas, também tem contribuído para pesquisas na área da bioinformática. A própria globalização da informação como um todo permitiu o aparecimento de biólogos capazes de programar e de matemáticos interessados em aplicar sua ciência para auxiliar a resolução dos problemas existentes na biologia.

Além disso, várias técnicas de biologia molecular tem sido aplicadas no estudo da diversidade e ecologia dos microrganismos que não podiam ser cultivados em seu habitat natural. Tais métodos levaram ao crescente número de seqüências de nucleotídeos e aminoácidos disponíveis nos inúmeros bancos de dados como o RDP - Ribosomal Database Project - (http://rdp.cme.msu.edu/html/), GenBank (http://www.ncbi.nlm.nih.gov), EMBL – European Molecular Biology Laboratory - (www.embl-heidelberg.de) e o DDBJ – DNA Data Bank of Japan - (http://www.ddbj.nig.ac.jp) entre outros. Estes por sua vez têm se tornado uma fonte fundamental de informação para os pesquisadores das mais diversas áreas, principalmente da biologia molecular. Uma das aplicações da bioinformática na microbiologia é a construção de árvores filogenéticas.

1.4 Filogenia molecular

A natureza do DNA permite que este seja usado como um "documento" da história evolutiva. Comparando-se seqüências de DNA de diversos genes entre diferentes organismos, pode-se inferir relações entre estes que não poderiam ser determinadas somente pela observação morfológica (NCBI – http://www.ncbi.nlm.nig.gov/About/primer/).

Quando seqüências de ácidos nucléicos ou proteínas encontradas em organismos diferentes são similares, é provável que estas tenham sido originadas de uma seqüência ancestral comum. Um alinhamento de seqüências revela quais posições foram conservadas e quais divergem entre os descendentes de um mesmo ancestral. Quando duas seqüências possuem uma relação evolutiva, elas podem ser denominadas seqüências homólogas.

Como os genes evoluem através do acumulo de mutações, o número de diferenças entre eles pode indicar a quanto tempo estes divergiram de um ancestral comum. Portanto, comparando-se diferentes genomas, é possível derivar relações evolutivas entre eles. A este tipo de abordagem dá-se o nome de análise filogenética. Uma análise filogenética direta consiste de quatro passos:

1 – Definição do cronômetro molecular – ver 1.4.1
2 – Alinhamento das seqüências – ver 1.4.2
3 – Determinação de um modelo de substituição (consideração
da variação das seqüências) – ver 1.4.3
4 – Construção da árvore – ver 1.4.4
5 – Avaliação da árvore – ver 1.4.5

1.4.1 Cronômetros moleculares

Certas macromoléculas podem funcionar como cronômetros evolutivos, ou seja, podem ser usadas para medir distâncias evolutivas. Uma molécula cuja seqüência varie aleatoriamente no tempo pode ser considerada um cronômetro. É denominada distância a quantidade de mudanças ocorridas em uma seqüência no decorrer do tempo. Quanto mais mudanças forem computadas entre duas seqüências de uma determinada molécula, maior será a distância entre elas (distância = taxa de mutação x tempo). Porém, nem todas as seqüências são de grande valor ao se tentar estabelecer relações filogenéticas. As principais características necessárias para uma seqüência poder ser empregada em filogenia são:

1. Distribuição universal

Por razões obvias o cronômetro deve ser universalmente distribuído no grupo sendo estudado.

2. Função homóloga conservada

Uma função conservada garante a presença na seqüência de regiões conservadas que poderão ser alinhadas, favorecendo análise filogenética confiável.

3. Espectro filogenético

A taxa de mutação do cronômetro deve ser condizente com o espectro evolucionário das distâncias sendo avaliadas. Quanto maior a distância evolutiva sendo avaliada, menor deve ser a taxa de evolução do cronômetro. Um cronômetro que possua uma taxa de mutação baixa tem pouca utilidade para avaliar distâncias evolutivas pequenas, pois ele seria demasiadamente conservado para discernir entre os organismos.

4. Tamanho

O tamanho da seqüência deve ser grande o suficiente para que esta produza dados estatísticos satisfatórios. Além disso, para um melhor resultado, a seqüência deve conter partes cujas taxas de evolução sejam relativamente independentes umas das outras. Isto permite que o cronômetro mantenha sua sensibilidade mesmo que em alguma de suas partes não ocorram mutações aleatórias (ex. caso alguma parte esteja sujeita a fortes pressões seletivas).

1.4.2 Alinhamento múltiplo

Por alinhamento múltiplo de seqüências entende-se o procedimento de comparar duas ou mais seqüências procurando-se uma série de caracteres individuais ou padrões de caracteres que estão na mesma ordem nas seqüências (Mount, 2001).

Um alinhamento múltiplo é uma matriz de dados onde as colunas representam caracteres homólogos. Existem dois tipos de alinhamento:

1. Alinhamento global: Alinhamento de seqüências nucleotídicas ou de aminoácidos em toda a sua extensão.
2. Alinhamento local: É o processo de busca de alinhamento de regiões altamente similares de duas seqüências de DNA ou proteína.

O alinhamento de seqüências pode ser útil para descobrir informações funcionais, estruturais e evolutivas. É importante obter o alinhamento ótimo para ter acesso a estas informações. Os métodos de análises filogenéticas assumem que os alinhamentos múltiplos respeitam as seguintes afirmações (Mount, 2001):

a) As seqüências são homólogas – todas descendem de um ancestral comum.

b) Cada posição do alinhamento é homóloga a todas as outras daquela mesma coluna.

c) Cada uma das seqüências incluídas tem uma história filogenética em comum.

d) A amostragem dos taxa é adequada para resolver o problema em estudo.

e) A variação na seqüência das amostras é representativa dos grupos aos quais elas pertencem.

f) As variações nas seqüências das amostras possuem um sinal filogenético adequado para resolver o problema em estudo.

g) As seqüências estão corretas e são oriundas de suas respectivas fontes.

1.4.2.1 Alinhamento múltiplo em genes codificadores de proteína

Quando se utiliza genes codificadores de proteínas, têm-se dois níveis de informação: o da seqüência de aminoácidos e o da seqüência de DNA. Portanto, deve-se valer destas informações para realizar um alinhamento que seja mais próximo da realidade biológica.

Um dos problemas existentes ao se realizar alinhamentos múltiplos de seqüências é o de dar valor (penalidades) aos gaps, que representam eventos de deleção ou inserção ("indels"). Estes valores deveriam ser altos onde a probabilidade de acontecer um indel for baixa. Devido a natureza altamente deletéria de uma mudança na janela de leitura de uma proteína, a probabilidade de acontecer uma inserção ou uma deleção dentro de um dos códons é baixíssima, sendo portanto necessário uma penalidade alta para a inserção de um gap nesta situação.

Por isso não é muito sensível alinhar seqüências codificadoras de proteína como DNA, sendo recomendado traduzir a seqüência de DNA para aminoácidos, realizar o alinhamento e depois efetuar uma tradução reversa. Deve-se no final realizar a tradução reversa para usar outra vantagem dos genes codificadores de proteína. O código genético é degenerado e existe uma variabilidade maior na terceira posição dos códons. Este fato pode ser explorado, quando se utiliza genes codificadores de proteína como o *rpoB* e o *gyrB*, para tentar discriminar espécies muito relacionadas.

1.4.3 Modelos de distância

Quando o tempo de divergência das seqüências é pequeno, a distância observada é mais próxima da distância real. Conforme o tempo de divergência aumenta, também aumenta a probabilidade de uma ou mais substituições ocorrerem no mesmo sítio. Portanto para todas as seqüências analisadas, exceto as com baixíssimo tempo de divergência, as distâncias observadas subestimam a distância real (Jukes & Cantor, 1969).

Em ordem para construir uma árvore válida, as distâncias observadas devem ser corrigidas de modo que levem em consideração substituições múltiplas no mesmo sítio. Vários métodos foram desenvolvidos para realizar esta correção, e cada um assume diferentes hipóteses a respeito do processo de substituição.

1.4.3.1 Distância p

Esta é a distância não corrigida, ou distância observada. Corresponde a proporção (p) de sítios de nucleotídeo onde duas seqüências diferem. Este valor é obtido dividindo-se o número de diferenças entre duas seqüências pelo número de nucleotídeos sendo comparados.

1.4.3.2 Jukes & Cantor

Este método (Jukes & Cantor, 1969) foi desenvolvido assumindo-se que a taxa de substituição de nucleotídeos é a mesma para os quatro tipos (A, T, C e G). Também assume que as taxas de substituição em todos os sítios são iguais e que a freqüência dos quatro tipos de nucleotídeos são as mesmas. O modelo não diferencia transições de transversões.

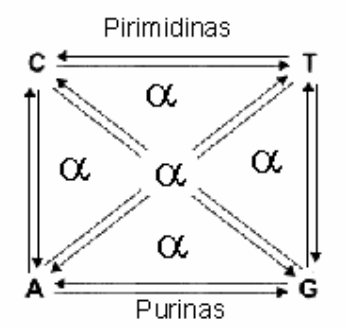


Figura 1: Esquema do modelo Jukes & Cantor, onde α é a taxa de substituição

1.4.3.3 Kimura 2 parâmetros

Este modelo (Kimura, 1980) corrige substituições múltiplas assumindo as mesmas hipóteses do modelo de Jukes & Cantor, e adicionalmente diferencia transições (substituições purina ↔ purina ou pirimidina ↔ pirimidina) de transversões (substituições purina ↔ pirimidina).

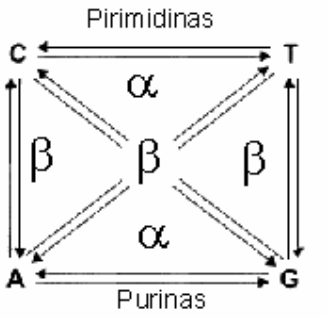


Figura 2: Esquema demonstrando o modelo de kimura, onde α é a taxa de transições β a taxa de transversões.

1.4.3.4 Outros modelos

Existem ainda outros modelos, que levam em consideração a diferentes porcentagens de GC (Tamura, 1992) ou ainda taxas de evolução diferentes na mesma molécula (Galtier & Gouy, 1995).

1.4.4 Conceito de árvores filogenéticas

As relações evolutivas entre grupos de organismos podem ser ilustradas em gráficos chamados de árvores filogenéticas. Uma árvore filogenética é composta de nós e ramos, sendo que cada ramo conecta nós adjacentes. Os nós representam unidades taxonômicas e os ramos definem as relações entre essas unidades em termos de descendência e ancestralidade. O padrão de ramificação de uma árvore é chamado de topologia. O tamanho dos ramos (em um filograma) representam o número de mudanças que ocorrem em relação ao último nó. As unidades taxonômicas representadas pelos nós podem ser espécies, populações, indivíduos, proteínas ou genes.

Quando trabalhamos com árvores filogenéticas devemos distinguir entre nós internos e externos. Os nós externos representam as unidades taxonômicas que estão sendo comparadas e podem ser chamadas de OTUs (Operational Taxonomic Units). Os nós internos representam unidades ancestrais. A raiz de uma árvore pode ser escolhida incluindo-se uma OTU que especula-se ter sido originada anteriormente às outras, e ao mesmo tempo deve ter alguma relação com as restantes. Também é

possível predizer uma raiz assumido-se a hipótese dos cronômetros moleculares, que diz que as taxas de mutação nos ramos da árvore são uniformes.

A figura 3 mostra uma árvore filogenética na qual os ramos estão em escala (seus tamanhos são proporcionais ao número de mudanças a partir do último ancestral). Nesta figura, os nós em preto (C, D, E, G e H) são externos e representam OTUs, enquanto os nós azuis A, B, F e J são internos e representam linhagens ancestrais.

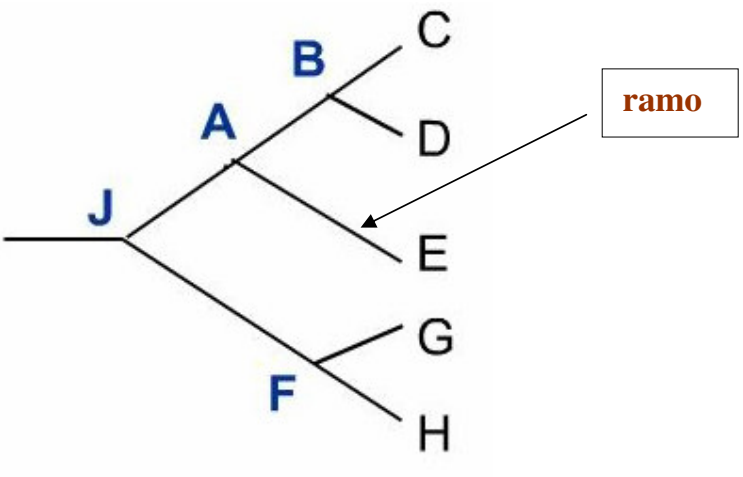


Figura 3: Exemplo de árvore com raiz. As linhagens C/D e G/H são derivadas de ancestrais comuns, representados pelos nós B e F respectivamente. Os ancestrais em destas linhagens também possuem um ancestral em comum (nó J). Note que os nós internos representam uma separação do caminho evolutivo que o gene ou organismo percorreu. Depois deste ponto, quaisquer mudanças evolutivas nos ramos formados ocorrem independentemente.

O período evolutivo transcorrido desde a separação de C e D geralmente não é conhecido. O que se procura estimar pela análise filogenética é a quantidade de mudanças ocorridas entre C e o nó interno B e também entre este e D. Observando-se o comprimento dos ramos pode-se deduzir que C e D tiveram a mesma quantidade de mudanças em suas seqüências. Porém, algumas vezes, por alguma razão biológica ou ambiental, uma OTU pode apresentar mais mudanças do que outra, o que é representado por comprimento de ramos diferenciados (ex. OTU´s A e B da figura abaixo).

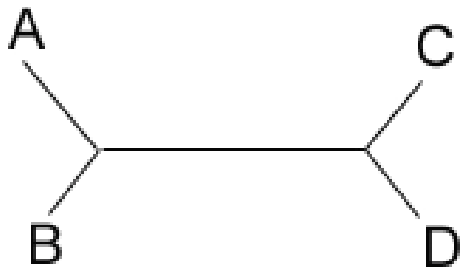


Figura 4: Exemplo de árvore sem raiz. A OTU A sofreu mais mudanças do que a B (Mount, 2001).

1.4.5 Reconstrução de árvores filogenéticas

Na filogenia molecular, os métodos de reconstrução de árvores podem ser distinguidos entre abordagens de distância e de estado de caráter. Os métodos pertencentes à primeira abordagem são baseados em medidas de distância, como o número de substituições de nucleotídeos ou de aminoácidos. Já os métodos relativos à segunda abordagem estão relacionados ao estado de um caráter, como a presença ou não de um nucleotídeo/aminoácido em num local particular, ou a presença/ausência de uma deleção/inserção em uma certa porção do DNA.

1.4.5.1 Reconstrução de árvores filogenéticas pelo método de distância

Nos alinhamentos múltiplos de seqüências, normalmente se calcula um índice de similaridade, o qual é definido pela soma das posições idênticas mais o número de substituições conservadas entre duas seqüências. Para uma analise filogenética, o que é utilizado é o índice de distância. Este valor é definido pelo número de posições não idênticas ou o número de posições que precisam ser alteradas para gerar a outra seqüência. Métodos de distância são métodos quantitativos.

Conjunto de seqüências idealizadas para as quais os comprimentos dos ramos de uma árvore assumida são aditivas (Mount, 2001)

Seqüências

seqüência A A C G C G T T G G G C G A T G G C A A C
seqüência B A C G C G T T G G G C G A C G G T A A T
seqüência C A C G C A T T G A A T G A T G A T A A T
seqüência D A C A C A T T G A G T G A T A A T A A T

Distâncias entre as seqüências (número de passos requeridos para uma seqüência se igualar a outra)

n_{AB} 3
n_{AC} 7
n_{AD} 8
n_{BC} 6
n_{BD} 7
n_{CD} 3

Matriz de distância (tabela 1)

	A	B	C	D
A	-	3	7	8
B	-	-	6	7
C	-	-	-	3
D	-	-	-	-

Árvore filogenética resultante para das seqüências A-D (figura 5), demonstrando os comprimentos dos ramos. A soma dos comprimentos entre duas seqüências quaisquer tem o mesmo valor das distâncias entre as seqüências

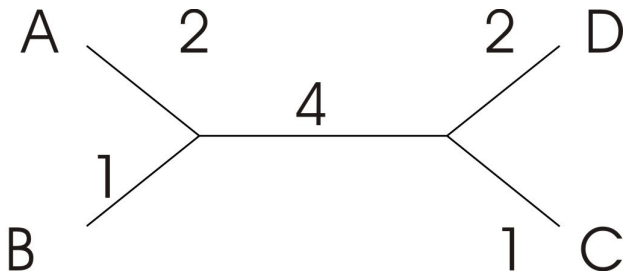


Figura 5

Os métodos de distância têm como objetivo construir uma árvore filogenética tomando como base uma matriz de distância. Esta matriz demonstra as distâncias evolutivas entre cada uma das seqüências alinhadas. Os pares de seqüências que possuírem as menores distâncias entre eles são denominados vizinhos. Em uma árvore, tais seqüências compartilham um mesmo nó interno e são ligadas ao mesmo por um ramo.

O objetivo destes métodos é identificar uma árvore que posicione os vizinhos corretamente e que também determine o tamanho dos ramos de modo que reproduzam os dados originais da melhor forma possível. Existem alguns algoritmos que utilizam métodos de distância para construir árvores filogenéticas.

1.4.5.1.1 Fitch e Margoliash (1967)

Este método utiliza tabelas de distância como a da tabela 7 acima. As seqüências são combinadas em trios para definir a arvore em questão e o comprimento dos seus ramos.

Passos no algoritmo (Mount, 2001):

Desenha-se uma árvore sem raiz com um trio de seqüências emanando do mesmo nó, como na figura 6.

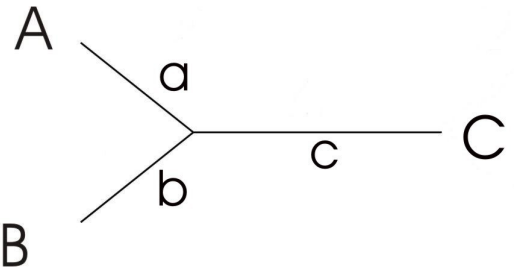


Figura 6

Calcula-se o tamanho dos ramos algebricamente

O tamanho dos ramos podem ser calculados observando-se a árvore:

Distância de A para B = a + b = 22 (1)

Distância de A para C = a + c = 39 (2)

Distância de B para C = b + c = 41 (3) Obs: Ver tabela 2

Subtraindo-se (3) de (2), a – b = -2 (4)

Somando-se (1) e (4), 2a = 20, a = 10 (5)

Então de (1) e (2) tem-se; b = 12, c = 29

Deve-se observar que nestes cálculos a distância de A e B para seu ancestral comum são diferentes, ou seja, A e B tem taxas de mutação diferentes neste modelo.

Algoritmo de Fitch -Margoliash (1967) para cinco seqüências (tabela 2):

	A	B	C	D	E
A	-	22	39	39	41
B	-	-	41	41	43
C	-	-	-	18	20
D	-	-	-	-	10
E	-	-	-	-	-

As seqüências mais próximas na tabela 2 são D e E.

As distâncias médias entre D para A, B e C e de E para A, B e C são calculadas (tabela 3):

	D	E	$D_{média}$ ABC
D	-	10	32.7
E	-	-	34.7
$D_{média}$ ABC	-	-	-

D e E agora são tratadas como uma seqüência única composta (DE), e uma nova tabela de distância é feita. A distância de A para (DE) e a distância média de A para D e de A para E. As outras distâncias são calculadas de mesmo modo (tabela 4):

	A	B	C	(DE)
A	-	22	39	40
B	-	-	42	43
C	-	-	-	19
(DE)	-	-	-	-

O próximo par de seqüências mais similares é identificado, neste caso C com (DE) (tabela 5):

	DE	C	$D_{média}$ AB
DE	-	19	41
C	-	-	40
$D_{média}$ AB	-	-	-

Calculando-se algebricamente como anteriormente, c = 9

Algoritmo de Fitch -Margoliash (1967) para mais de cinco seqüências:

1. Encontrar o par de seqüências mais próximos, por exemplo, A e B.
2. Tratar as seqüências restantes (ex. C, D, E e F) como sendo únicas e calcular uma nova matriz de distância entre as demais seqüências e esta nova seqüência composta (distância média).
3. Usar os dados para calcular os valores de A e B algebricamente (como no exemplo com 3 seqüências).
4. Depois, deve-se tratar A e B como sendo uma única seqüência, e calcular as distâncias entre AB e as demais seqüências. Fazer uma nova matriz de distância.
5. Identifica-se o próximo par de seqüências mais próximas e repete-se passo 1 para calcular os próximos valores dos tamanhos dos ramos.

6. Deve-se repetir o procedimento inteiro para todos os possíveis pares de seqüências (A e B, A e C, A e D, etc). Por fim deve-se calcular as distâncias entre cada par de seqüências para cada árvore e escolher aquela que melhor refletir os dados originais.

1.4.5.1.2 Neighbor – Joining

Neighbor – Joining (Saitou & Nei, 1987) é especialmente útil quando as taxas de evolução das linhagens consideradas variam. Este método é muito parecido com o anterior, sendo a que a diferença básica esta presente no algoritmo que determina quais seqüências irão ser escolhidas para serem consideradas compostas.

1.4.5.2 Reconstrução das árvores por métodos de caráter

1.4.5.2.1 Maximum Parsimony

O princípio da parcimônia máxima ou mínima evolução está relacionado à identificação da árvore filogenética que apresente o menor número de mudanças evolutivas para explicar as diferenças entre as OTUs estudadas (Mount, 2001). Esse método cladístico resulta em uma árvore chamada de árvore de parcimônia máxima.

Pode acontecer de ser obtida mais de uma árvore contendo o mesmo número mínimo de mudanças encontradas entre as OTUs. Assim, podemos obter duas árvores distintas a partir dos mesmos dados de entrada (Swofford *et al*., 1996).

Tabela 6. Sítios considerados informativos pelo método de Maximum Parsimony estão em verde. Não informativos em vermelho.

Espécie / Sítio	**1**	**2**	**3**	**4**
I	A	T	A	T
II	A	T	C	T
III	G	C	A	T
IV	G	C	C	T

Para ser montada a topologia da árvore mais parcimoniosa, não são observadas todas as seqüências de nucleotídeos, apenas aquelas que apresentam sítios informativos (Mount, 2001). Na tabela 5, os sítios 1, 2 e 3 são informativos. Já o sítio 4 não é.

Para calcularmos uma árvore de parcimônia máxima devemos, primeiramente, identificar todos os sítios informativos presentes nas seqüências de DNA das OTUs. A seguir, para cada árvore filogenética possível, devemos calcular o número mínimo de substituições em cada sítio informativo. Finalmente, somamos o número de mudanças de todos os sítios informativos para cada árvore e escolhemos a árvore associada com o menor número de substituições.

1.4.5.2.2 Maximum Likelihood

O princípio deste método é avaliar a probabilidade de que um determinado modelo de mudanças evolutivas possa explicar a origem dos dados observados. Estimativas de Maximum Likelihood foram utilizadas pela primeira vez por Cavalli-Sforza & Edwards em 1967, embora não utilizando dados de seqüência de macromoléculas (feito por Felsenstein em 1981).

O objetivo do método é inferir a história ou o conjunto de histórias evolutivas que sejam os mais consistentes em relação aos dados estudados. Para a aplicação do Maximum Likelihood, é necessário que um modelo concreto de mudanças evolutivas que leve à conversão de uma seqüência em outra seja especificado. O modelo pede ser:

a) completamente definido

b) conter uma série de parâmetros os quais serão estimados a partir dos dados.

Sumarizando, nesta metodologia, os modelos de mudanças evolutivas são avaliados quanto à sua probabilidade de explicar um conjunto de dados de forma que reflita a história evolutiva mais próxima da realidade, ou seja, a história mais verossímil (likelihood). Nessa avaliação, os modelos recebem valores de verossimilhança e aquele que apresentar o melhor valor é o que será utilizado para se inferir a árvore filogenética.

1.4.6 Avaliação das árvores

É de fundamental importância reconhecer que as árvores filogenéticas obtidas pelos métodos descritos acima são hipóteses. Deve-se portanto testar a estabilidade e a confiança dos dados.

Existem alguns métodos que permitem saber se os dados são bem estruturados ou se contém um bom sinal filogenético. Um dos métodos mais utilizados é o do bootstrap.

1. Índices de "bootstrap"

Os índices de bootstrap (Felsenstein, 1985) são verificados através da análise de reamostragem de caracteres da matriz de dados, gerando novas matrizes. Se, na reamostragem de caracteres, determinados ramos permanecem sempre juntos nas novas topologias simuladas, aquele nó que os une receberá um valor de 100% na árvore original.

1.4.7 Considerações Filogenéticas

É importante ressaltar que não existe uma única metodologia que se aplique a todos os estudos filogenéticos realizados com dados moleculares. Embora existam inúmeros algoritmos, procedimentos e softwares desenvolvidos para este tipo de análise, sua confiabilidade e praticidade são em todos os casos dependentes do tamanho e da estrutura dos dados analisados. As vantagens e desvantagens destes métodos estão sujeitas a debates científicos, devido ao fato de que o perigo de gerar resultados incorretos é maior na filogenia molecular computacional do que em muitos outros campos da ciência. Por vezes, o fator limitante neste tipo de análise não é o poder computacional, mas sim o entendimento que o pesquisador tem do que método computacional utilizado está fazendo com os seus dados.

1.4.8 Diferenças entre árvores de genes e árvores de espécies

É assumido que uma árvore filogenética constituída a partir de dados moleculares, poderá ser uma representação da árvore de espécie menos ambígua do que se obtida por comparações morfológicas. Porém, isso não significa que este tipo de árvore filogenética será a condizente com a historia evolutiva das espécies sendo analisadas. Para isto ser verdade, os nós internos em ambas as árvores (da espécie e do gene) devem ser equivalentes, o que nem sempre é o caso. Um nó interno em uma árvore de gene indica a divergência de um gene ancestral em dois genes com seqüências de DNA diferentes, um evento de paralogia (Jensen, 2001). Já em uma árvore de espécie, o nó interno representa um evento de especiação (ortologia). Nem sempre os dois ocorrem ao mesmo tempo.

1.5 Genes escolhidos

Diversos genes podem ser utilizados em estudos filogenéticos, porém neste estudo foram escolhidos três genes: *rrn* (16S RNA), *rpoB* (subunidade β da RNA polimerase de procariotos) e *gyrB* (subunidade β da DNA girase).

Existem pelo menos dois bancos de dados especializados em RNA ribossomal, o Ribosomal Database Project (RDP - http://rdp.cme.msu.edu/) e o European Ribosomal Database (http://www.psb.ugent.be/rRNA/), sendo que o RDP conta com 136,335 seqüências (junho de 2005). Existe também um banco de dados público especializado em *gyrB*, o Identification and Classification of Bactéria (ICB - http://seasquirt.mbio.co.jp/icb/index.php), cujo objetivo é prover a comunidade científica com recursos genéticos para a identificação de bactérias. Atualmente ele conta com 2789 seqüências (junho de 2005). A existência destes bancos de dados foram fatores influenciaram a escolha dos genes.

O Genbank, um dos maiores bancos de dados existentes (40,604,319 seqüências em junho de 2005) possui também uma considerável quantidade de seqüências destes genes - 16S rRNA, *rpoB* e *gyrB* contam com 32.947, 2.877 e 3.203 seqüências respectivamente (junho de2005).

1.5.1 *rpoB*

A RNA polimerase catalisa a síntese de RNA a partir de uma fita molde de DNA. Em bactérias uma única RNA polimerase (RNAP) é responsável pela síntese de mRNA´s, rRNA´s e tRNA´s. A enzima central é composta por quatro subunidades peptídicas: alfa (α), beta (β), beta` (β`)e omega (ω), assumindo a forma α2ββ`ω.

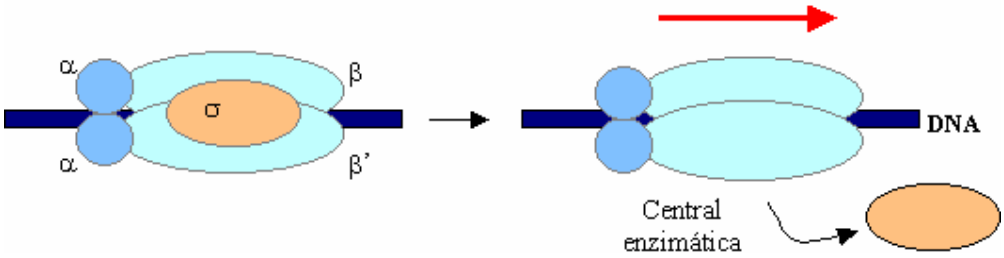


Figura 7: Subunidades da RNA polimerase de E. coli: duas alfa (40.000 kDa), uma beta (150.000 kDa), uma b' (160.000 kDa) e uma sigma (70.000 kDa). A especificidade para o sítio correto da iniciação da transcrição é determinada pela subunidade sigma.

Na forma α2ββ`ω a RNA polimerase pode se ligar ao DNA e catalisar a síntese de RNA não especificamente. Uma outra subunidade sigma (σ) completa a enzima conferindo especificidade. Ela é responsável pelo reconhecimento do promotor, estando presente no momento que a enzima começa a interagir com o DNA, se dissociando logo após. A *E. coli* sintetiza ao menos seis tipos de fatores sigma, cada um conferindo a RNA polimerase afinidade por promotores distintos.

As subunidades alfa, beta, e beta` são produtos dos genes *rpoA*, *rpoB* e *rpoC* respectivamente. O fator sigma é produto do gene *rpoD*.

Seqüências do gene *rpoB* têm sido usadas como uma ferramenta alternativa para determinação da filogenia ou identificação de bactérias entéricas (Mollet & raoult, 1997), *Mycobacterium* (kim *et al*., 1999), e espiroquetas (Lee *et al.*, 2000; Renesto, Drancourt & Raoult, 2000). O gene *rpoB* aparentemente possui uma só copia nos genomas bacterianos e também aparenta possuir uma taxa de mutação mais elevada que o 16S rRNA (Dahllöf, Harriet & Kjelleberg, 2000).

1.5.2 *rrn*

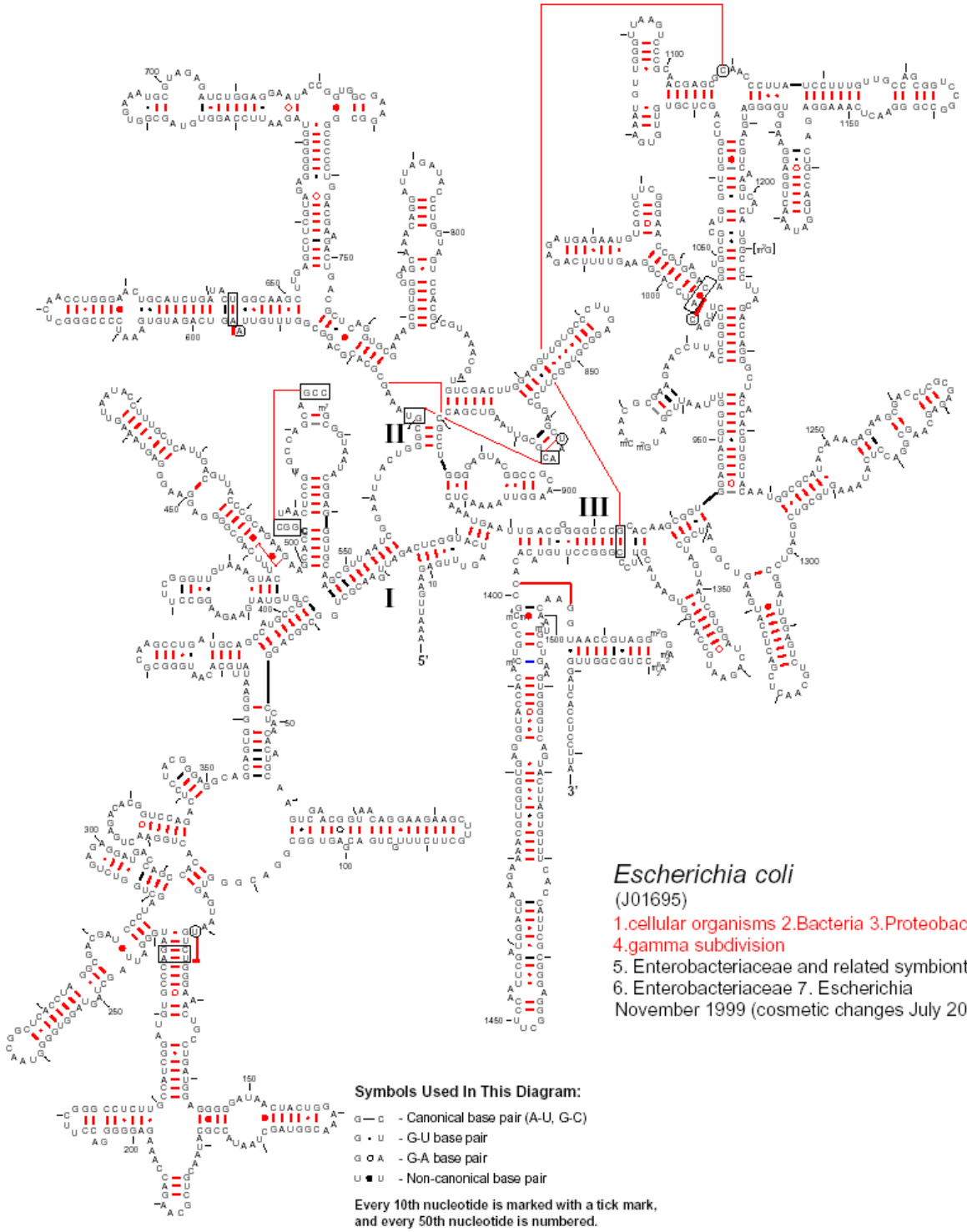
rRNA´s são moléculas antigas, funcionalmente constantes, universalmente distribuídas e relativamente bem conservadas mesmo entre organismos com grandes distâncias filogenéticas (Woese, 1987). O gene *rrn* codifica o 16S rRNA.

Existem três tipos de rRNA´s, que nos procariotos são, o 5S , o 16S e 23S com aproximadamente 120 (5S), 1500 (16S) e 2900 (23S) nucleotídios. Os maiores (16S e 23S) contém diversas regiões conservadas úteis para obter um alinhamento adequado, e ainda assim apresentam variabilidade suficiente em outras regiões da molécula para servir como cronômetros filogenéticos.

O rRNA 5S também já foi utilizado para análises filogenéticas (Kumazaki, Hori & Osawa, 1983; Erdmann *et al*., 1987), mas seu pequeno tamanho limita a informação possível de ser obtida desta molécula.

O rRNA 16S consiste de aproximadamente 1500 nucleotídeos com pareamento intramolecular, determinando uma estrutura secundária complexa com 4 domínios conservados. A estrutura da molécula de RNA determina sua interação com as proteínas ribossômicas e a própria conformação da subunidade ribossômica pequena.

Secondary Structure: small subunit ribosomal RNA



Escherichia coli
(J01695)
1.cellular organisms 2.Bacteria 3.Proteobacteria
4.gamma subdivision
5. Enterobacteriaceae and related symbionts
6. Enterobacteriaceae 7. Escherichia
November 1999 (cosmetic changes July 2001)

Symbols Used In This Diagram:

- G — C - Canonical base pair (A-U, G-C)
- G • U - G-U base pair
- G o A - G-A base pair
- U ■ U - Non-canonical base pair

Every 10th nucleotide is marked with a tick mark,
and every 50th nucleotide is numbered.

Tertiary interactions with strong comparative data are connected by solid lines.

Citation and related information available at http://www.rna.icmb.utexas.edu

Moléculas de DNA de todos os organismos estão sujeitas a forças topológicas originadas pelo tamanho e pela estrutura bi-helicoidal do DNA. Esta estrutura impede a rotação de qualquer uma das fitas do DNA ao redor da outra, a não ser por uma quebra na cadeia fosfodiester. As DNA´s topoisomerases são as enzimas responsáveis por aliviar as tensões torcionais no DNA, participando em todas as transações que requerem o desespiralamento de duas fitas de DNA (ex. replicação, transcrição, recombinação e remodelamento de cromatina).

As topoisomerases podem ser divididas em tipo I e II de acordo com suas propriedades enzimáticas, sendo que a DNA girase bacteriana é uma topoisomerase do tipo II (Danie´le *et al*, 2003). Esta enzima é um tetrâmero composto de duas proteínas (A2B2). A proteína A (*gyrA*) tem o peso molecular de aproximadamente 100 kDa enquanto a B (g*yrB*) tem 90kDa (em proteobactérias) ou 70kDa. Comparações entre as duas classes revelaram que as de 90 kDa possuem uma inserção de aproximadamente 170 aminoácidos na região 560 da seqüência da enzima de 70 kDa.

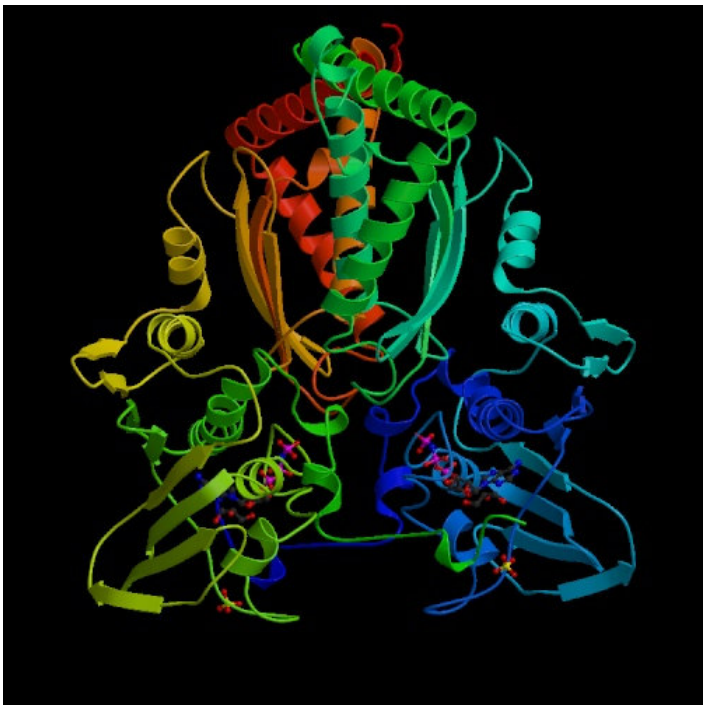


Figura 9: Estrutura da DNA girase de *E. coli* retirada do PDB (Protein Databank - http://www.rcsb.org/pdb/).

O gene *gyrB* possui todos os requisitos necessários para ser utilizado em análises filogenéticas. Aparenta estar presente em todas as espécies bacterianas e não ser transferido horizontalmente. Sua taxa de mutação aparenta ser maior do que a de genes ribossomais e do que outros genes que codificam de proteínas, o que o torna mais útil para a discriminação de estirpes.

Além destes fatores, este gene apresenta duas vantagens adicionais. A primeira é a disponibilidade de primers universais conhecidos, que são demonstradamente capazes de amplificar o gene por um amplo espectro de classes bacterianas (Yamamoto, S. & Harayama, S., 1996). A amplificação resulta em fragmentos de tamanho entre 1.2 e 1.4kb, que além de ter tamanho suficiente para ser usado em análises filogenéticas, apresenta regiões conservadas e de variabilidade.

1.6 Objetivos Gerais

A taxonomia de microrganismos é um processo complexo que leva em consideração a análise de características fenotípicas, químicas e moleculares (taxonomia polifásica). Metodologias moleculares, mais comumente o seqüenciamento do rDNA 16S, permitem a colocação direta de organismos em uma classificação filogenética, a nível de família e gênero. Pesquisas indicam que o rDNA 16S possui limitações como ferramenta taxonômica entre alguns grupos bacterianos devido fatores como a presença de copias múltiplas com seqüências variadas deste gene em um mesmo microrganismo. Neste trabalho foi avaliada a utilidade dos genes *rpoB* e *gyrB* como ferramentas taxonômicas e filogenéticas alternativas.

1.6.1 Teoria da evolução endosimbiótica

Evidências microscópicas e moleculares indicam que a célula eucariótica moderna é uma quimera genética, que evoluiu incorporando células de endosimbiontes quimiorganotróficos e fototróficos. Esta teoria postula que uma bactéria aeróbica estabeleceu residência no citoplasma de um eucarioto primitivo, fornecendo energia em troca de um ambiente estável, protegido e com nutrientes (Margulis & Stolz, 1984). Esta bactéria seria o ancestral da mitocôndria moderna.

De maneira similar, a incorporação de um organismo endosimbiótico fototrófico teria dado origem aos cloroplastos modernos.

Alguns eucariotos não chegaram a estabelecer relações endosimbióticas com bactérias, ou se estabeleceram as perderam posteriormente. Existem eucariotos que embora contenham um núcleo circundado por membrana, não possuem organelas. Interessantemente, estes organismos se localizam nas ramificações mais antigas na árvore filogenética universal (Madingan, Martinko & Parker, 2000).

Mitocôndrias e cloroplastos possuem ribossomos do tipo procarioto (sendo inclusive inibidos por antibióticos que inibem a função dos ribossomos bacterianos), contém pequenas quantidades de DNA circular covalentemente fechados (característica típica nos procariotos) e mostram seqüências de RNA ribossomal típica de certas bactérias (Margulis, 1984).

1.6.2 Enterobactérias

Devido a sua importância médica, as enterobactérias são um dos mais bem definidos e estudados grupos de bactérias Gram-negativas. Diversos isolados foram bem estudados e caracterizados

As enterobactérias compõem um grupo relativamente homogêneo filogenéticamente, dentro do grupo das Gamaproteobactérias (Madingan, Martinko & Parker, 2000), e são caracterizadas fenotipicamente como bacilos gram-negativos, não esporulantes, não móveis ou móveis por meio de flagelo peritríqueo, aeróbios facultativos, oxidase negativos e capazes de fermentar açúcares a diversos tipos de produtos finais.

Recombinação genética e homologia DNA-DNA demonstraram que as bactérias entéricas são intimamente relacionadas geneticamente. Porém, devido a sua importância médica os gêneros distintos são mantidos para auxiliar sua identificação.

1.7 Objetivos específicos e hipóteses

1. Avaliar as técnicas de filogenética utilizando seqüências dos genes 16S rRNA e *rpoB* para testar a hipótese da evolução endosimbiótica dos eucariotos, verificando as relações entre cianobactérias e cloroplastos.

Hipóteses:

- Os genes *rrn* e *rpoB* de cloroplastos e cianobactérias formariam um grupo distinto em relação a outros microrganismos.

- Utilizando-se o gene *rpoB*, o comprimento dos ramos da árvore seriam maiores do que com o *rrn*, demonstrando maior capacidade de discriminação do *rpoB*.

2. Comparar os genes *rpoB* e *gyrB* como ferramentas taxonômicas alternativas ao 16S rRNA no grupo das proteobactérias.

Hipótese:

- Os genes *rpoB* e *gyrB* podem ser mais eficientes para diferenciação do que o *rrn* no grupo das proteobactérias

3. Comparar os genes *rpoB* e *gyrB* como ferramentas taxonômicas alternativas ao 16S rRNA na família das *Enterobacteriacea*.

Hipótese:

- Os genes *rpoB* e *gyrB* podem ser mais eficientes para diferenciação do que o *rrn* na familia das *Enterobacteriaceae*

2. MATERIAL E MÉTODOS

2.1 Seqüências e analises filogenéticas

A comparação entre os genes *rrn*, *gyrB* e *rpoB* foi feita entre organismos pouco relacionados (organismos de diferentes classes) sendo que o espectro filogenético avaliado foi gradativamente afunilado até atingir organismos intimamente relacionados (espécies ou estirpes diferentes) tendo como referência o Bergey´s Outline (http://www.cme.msu.edu/bergeys/). Foi dada prioridade a seqüências de espécies tipo, quando estas eram disponíveis.

As seqüências foram obtidas nos bancos de dados Genbank - http://www.ncbi.nlm.nih.gov/ - (database do NCBI) e RDP - http://rdp.cme.msu.edu/ - (Cole *et al*., 2003) no caso do gene *rrn*. O número de acesso de todas as seqüências utilizadas neste trabalho são mostrados nas tabelas 6 e 7. Todas as seqüências foram retiradas diretamente no formato FASTA (figura 10), o qual é reconhecido pela maioria dos softwares utilizados em bioinformática. Posteriormente as seqüências foram armazenadas no computador como arquivos de texto simples no formato FASTA com o uso do Open Office.

```
>seqüência 1
ATGCTAGTCAT.....
...TTG

>seqüência 2
ATGCCG...
...GTG
```

```
>16s RNA E.coli parcial
ATCCTGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATG
CAAGTCGAACGGTAACAGGAAGAAGCTTGCTTCTTTGCTGAC
GAGTGGCGGACGGGTGAGTAATGTCTGGGAAACTGCCTGAT
GGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATA
ACGTCGCAAGACCAAAGAGGGGGACCTTCGGGCCTCTTGCCA
TCGGATGTGCCCAGATGGGATTAGCTAGTAGGTGGGGTAACG
GCTCACCTAGGCGACGATCCCTAGCTGGTCTGAGAGGATGAC
CAGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGG
AGGCAGCAGTGGGGAATATTGCACAATGGGCGCAAGCCTGA
TGCAGCCATGCCGCGTGTATGAAGAAGGCCTTCGGGTTGTAA
AGTACTTTCAGCGGGGAGGAAGGGAGTAAAGTTAATACCTTT
GCTCATTGACGTTACCCGCAGAAGAAGCACCGGCTAACTCCG
TGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATC
GGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTTGTTAAGT
CAGATGTGAAATCCCCGGGCTCAACCTGGGAACTGCATCTGA
TACTGGCAAGCTTGAGTCTCGTAGAGGGGGGTAGAATTCCAG
GTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGT
GGCGAAGGCGGCCCCCTGGACGAAGACTGACGCTCAGGTGC
GAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGGTAGTC
CACGCCGTAAACGATGTCGACTTGGAGGTTGTGCCCTTGAGG
```

Figura 10: estrutura do arquivo FASTA (acima) e exemplo de um arquivo (direita).

Tabela 7: Número de acesso das seqüências utilizadas.

Seqüências utilizadas – parte 1	
Seqüências	**Número de acesso (Genbank)**
Amborella trichopoda (c)	**AJ506156**
Agrobacterium tumefaciens str. C58	**AE007869**
Anaplasma marginale str. St. Maries	**CP000030 / M60313**
Bacillus subtilis subsp. subtilis str. 168	**AL009126**
Bradyrhizobium japonicum USDA 110	**BA000040**
Bordetella pertussis Tohama I	**BX640411**
Chlamydia trachomatis D/UW-3/CX	**AE001273**
Chlorella vulgaris C-27 (c)	**AB001684**
Chromobacterium violaceum ATCC 12472	**AE016910**
Cyanidium caldarium strain RK1 (c)	**AF022186**
Clostridium acetobutylicum ATCC 824	**NC_003030**
Corynebacterium diphtheriae NCTC 13129	**BX248353**
Deinococcus radiodurans R1	**AE000513**
Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough	**AE017309**
Escherichia coli (Migula) Castellani and Chalmers K-12 MG1655	**U00096**
Escherichia coli O157:H7 EDL933	**AE005174**
Escherichia coli O157:H7	**BA000007**
Escherichia coli CFT073	**AE014075**
Erwinia carotovora subsp. atroseptica SCRI1043	**BX950851**
Geobacter sulfurreducens PCA	**AE017180**
Helicobacter pylori J99	**AE001439**
Neisseria gonorrhoeae FA 1090	**AE004969**
Nitrosomonas europaea ATCC 19718	**BX321856**
Nostoc	**BA000019**
Nymphaea Alba (c)	**AJ627251**
Oryza sativa (c)	**X15901**
Panax ginseng (c)	**AY582139**
***Pasteurella multocida* subsp. multocida str. Pm70**	**AE006184**

2.2 Alinhamentos múltiplos

Alinhamentos múltiplos das seqüências foram realizados com o auxílio do software ClustalW 1.8 (Thompson *et al*., 1997), que utiliza a técnica do alinhamento progressivo. Esta técnica produz um alinhamento global tendo como guia uma árvore filogenética calculada pelo método de Neighbor–Joining (Saitou & Nei, 1987). Esta árvore é criada por comparações par a par pela técnica da programação dinâmica, que garante como resultado o melhor alinhamento possível.

No caso dos genes codificadores de proteínas (*rpoB* e *gyrB*) as janelas de leitura corretas dos genes foram encontradas, sendo que toda a parte das seqüências antes do códon iniciador foi apagada. As seqüências foram então traduzidas com o auxilio do Bioedit e alinhadas de modo a não permitir a introdução de gaps dentro de códons. Após o alinhamento as seqüências foram traduzidas de aminoácidos para nucleotídeos para dar prosseguimento as análises.

Posteriormente os alinhamentos foram verificados e editados a olho nu usando-se o Bioedit 7.0.4.1 (Hall, 1999). Os alinhamentos produzidos foram salvos no formato PHYLIP (arquivo .phy) com o auxilio do clustalx. Somente foram utilizadas as regiões das seqüências consideradas passiveis de comparação após a verificação com o Bioedit.

Tabela 8: Número de acesso das seqüências utilizadas (parte 2).

Seqüências utilizadas – parte 2	
Seqüências	**Número de acesso (Genbank)**
Photorhabdus luminescens subsp. laumondii TTO1	**BX571875**
Porphyra purpúrea (c)	**U38804**
Pseudomonas aeruginosa PAO1	**AE004440**
Psilotum nudum (c)	**AP004638**
Prochlorococcus marinus str. MIT 9313	**BX548175**
Rhodopseudomonas palustris CGA009	**BX572603**
Synechococcus	**BX548020**
Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67	**AE017220**
Salmonella enterica subsp. enterica serovar Typhi Ty2	**AE016846**
Salmonella enterica subsp. enterica serovar Paratyphi A str. ATCC 9150	**CP000026**
Salmonella enterica subsp. enterica serovar Typhi str. CT18	**AL627279**
Salmonella typhimurium LT2	**AE008894**
Shigella flexneri 2a str. 2457T	**AE016991**
Shigella flexneri 2a str. 301	**AE005674**
Synechocystis	**BA000022**
Thermosynechococcus elongatus BP-1	**BA000039**
Thermotoga maritima strain MSB8	**AE000512**
Thermus thermophilus HB27	**AE017221**
Yersinia pestis biovar Medievalis str. 91001	**AE017142**
Yersinia pestis KIM	**AE014012**
Yersinia pestis CO92	**AJ414160**
Yersinia pseudotuberculosis IP 32953	**BX936398**
Wolinella succinogenes DSM 1740	**BX571657**
***Zea mays* (c)**	**X86563**

2.3 Correção das distâncias e construção das árvores filogenéticas

Com o auxílio dos softwares TREECONW 1.3b (Van de Peer & Wachter, 1993) e Phylo_Win (Galtier, Gouy & Gautier, 1995) as distâncias evolucionarias entre as seqüências foram corrigidas pelo modelo de Galtier & Gouy (1995). Os métodos adotados para inferi-las foram Neighbor-Joining (Saitou & Nei, 1987; Studier & Keppler, 1988), Maximum Parsimony (Swofford & Olsen, 1996) e Maximum Likelihood (Felsenstein 1981).

2.4 Avaliação das árvores

Para se avaliar a estabilidade das árvores geradas foi usada a análise de "bootstrap" (Efron & Gong, 1983; Felsenstein, 1985; Swofford & Olsen, 1996) com 1000 repetições para os métodos Neighbor-Joining e Maximum Parsimony e com 100 repetições para o método de Maximum Likelihood.

2.5 Matrizes de distância e gráficos de entropia

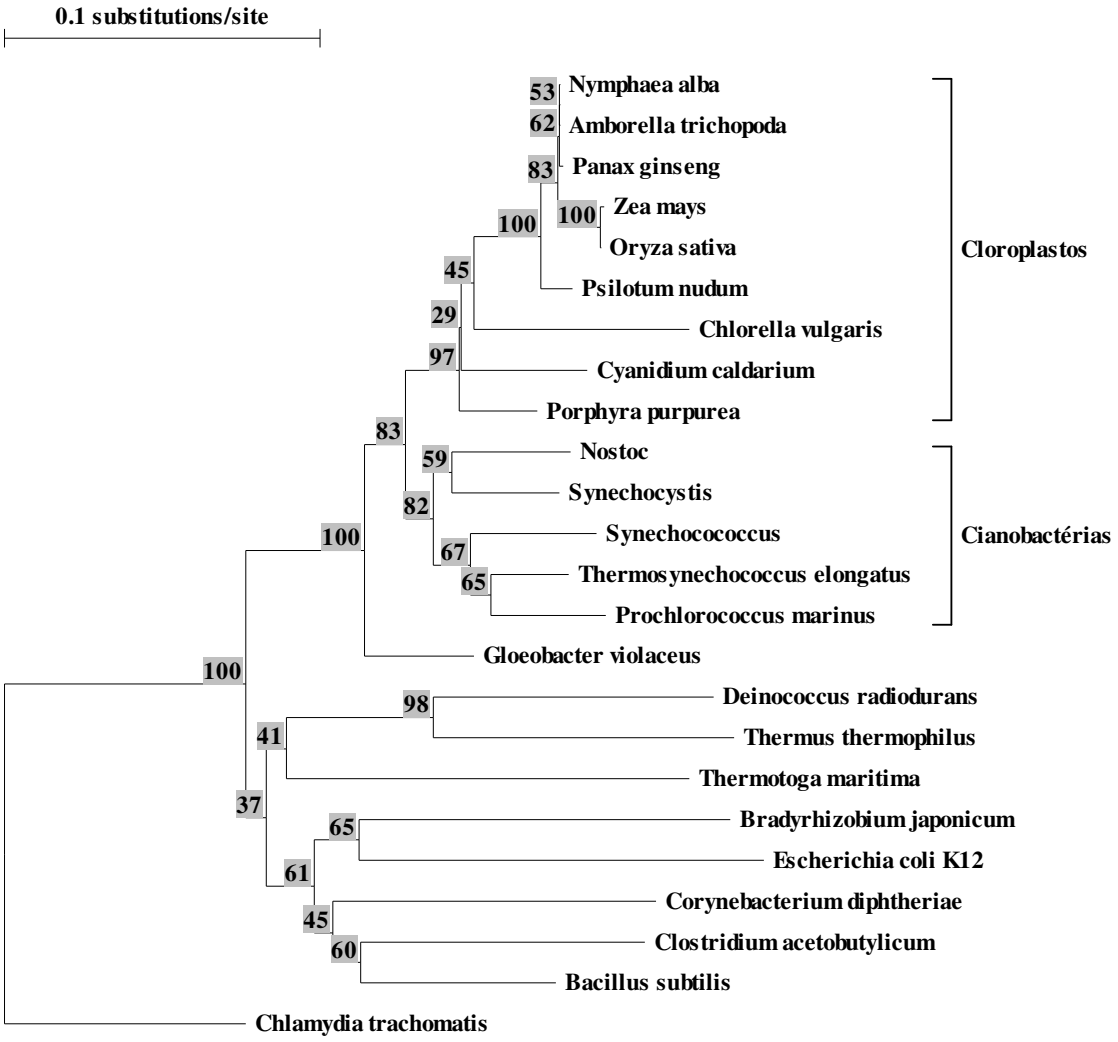
As matrizes de distância foram computadas com o auxilio do software MEGA3 (Kumar, Tamura & Nei, 2004), analisando-se só as posições do alinhamento onde foi possível comparar todas as seqüências (sem gaps). As distâncias evolutivas não foram corrigidas, utilizando-se a distância p, que é a proporção de sítios onde duas seqüências que estão sendo comparadas diferem (Kumar, Tamura & Nei, 2004). Após computadas as matrizes, foi calculada a média da distância entre todas as seqüências para cada gene avaliado. Os gráficos de entropia foram computados com o auxilio do software Bioedit 7.0.4.1 (Hall, 1999), tendo como base os alinhamentos das seqüências dos genes *rrn*, *rpoB* e *gyrB* de membros da família das *Enterobacteriaceae* produzidos pelo software clustalx.

3. RESULTADOS

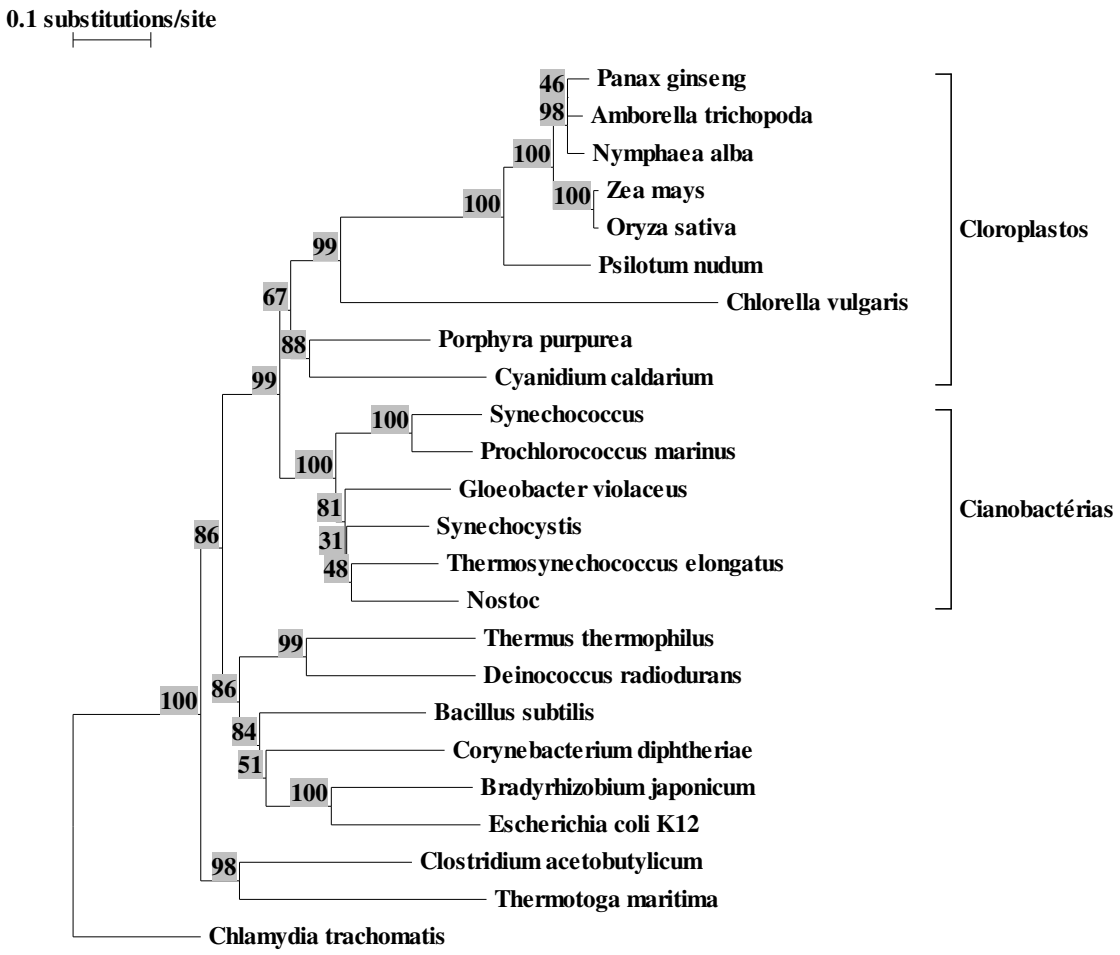
3.1 Avaliação das técnicas de filogenética para testar a hipótese da evolução endosimbiótica utilizando seqüências dos genes *rrn* e *rpoB*.

Análises filogenéticas de Neighbor – Joining, Maximum Likelihood e Maximum Parsimony baseadas nas seqüências dos genes 16S rRNA e *rpoB* resultaram em uma organização similar e confiável dos dois clusters evidenciados nas árvores 1 e 2, os quais foram suportados por altos valores de bootstrap.

Utilizando-se o método de Neighbor – Joining, valores de bootstrap maiores ou iguais a 95% foram observados em 6 dos 22 nós para o 16S rRNA, enquanto que valores maiores ou iguais a 95% foram observados em 12 dos 22 nós para o gene *rpoB*. Valores bootstrap para o gene *rpoB* também foram maiores nas análises com Maximum Likelihood e Maximum Parsimony. A cianobactéria *Gloeobacter violaceus* se mostra mais intimamente relacionada com as demais cianobactérias nas análises baseadas em *rpoB* do que no 16S rRNA.



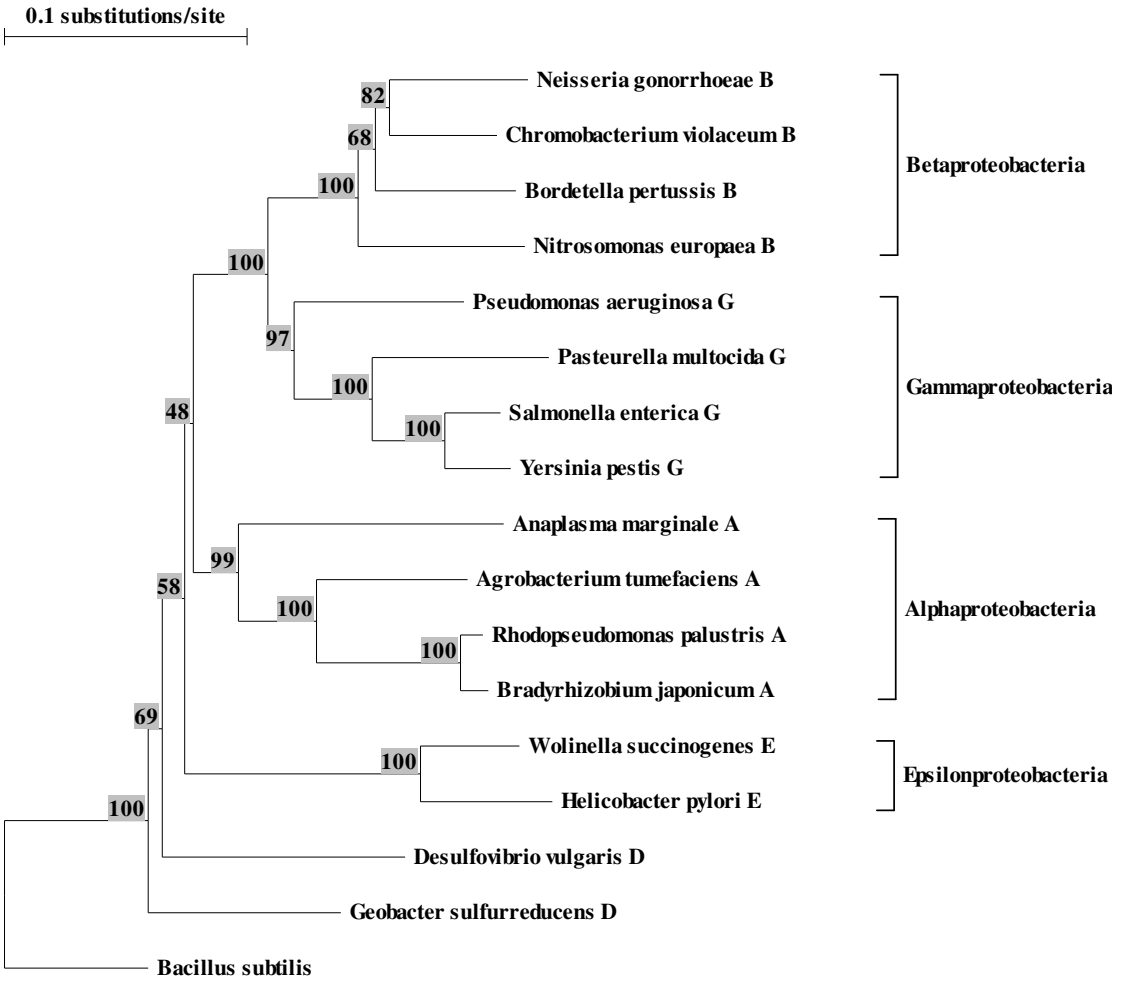
Árvore 1 - Relações filogenéticas entre cloroplastos, cianobactérias e outros microrganismos baseadas em 16S rRNA. Método: Neighbor Joining. 1264 posições analisadas. 1000 bootstraps (valores percentuais). Clusters destacados com colchetes foram suportados pelos três métodos utilizados (Neighbor – Joining, Maximum Parsimony e Maximum likelihood).



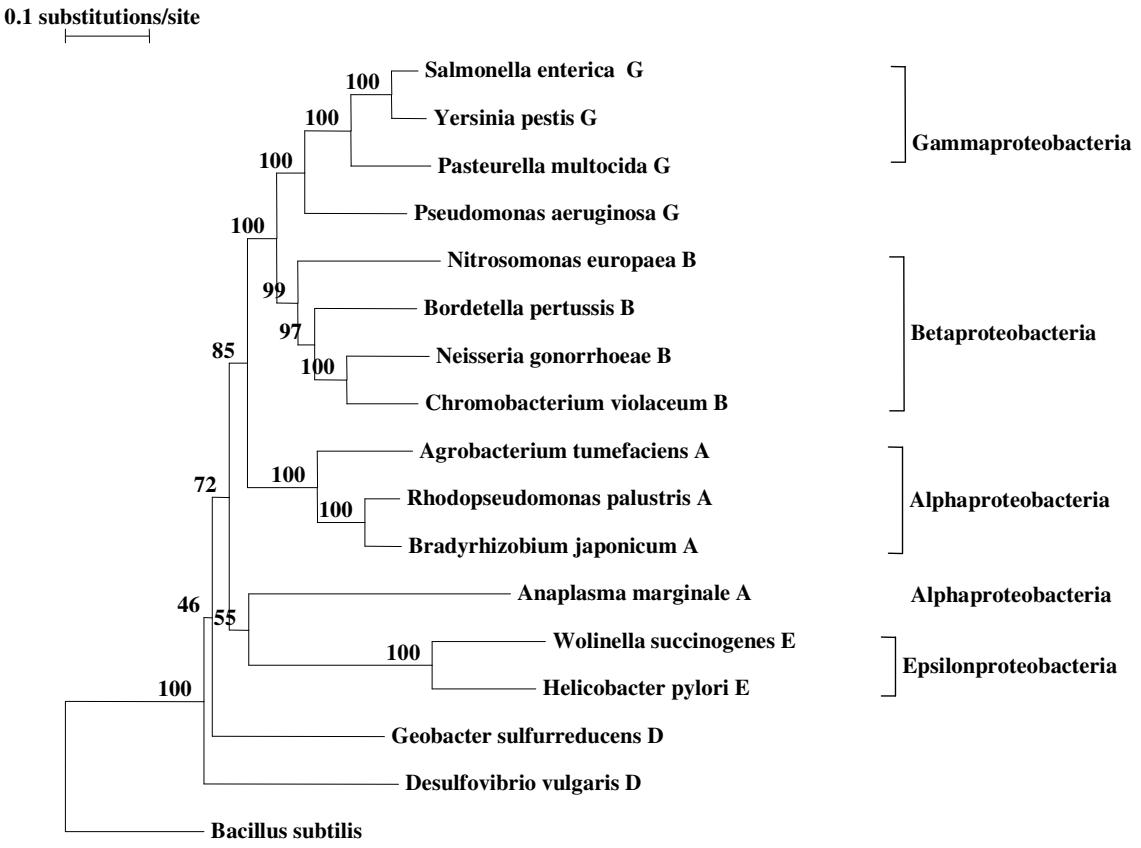
Árvore 2 - Relações filogenéticas entre cloroplastos, cianobactérias e outros microrganismos baseadas em *rpoB*. Método: Neighbor Joining. 2757 posições analisadas. 1000 bootstraps (valores percentuais). Clusters destacados com colchetes foram suportados pelos três métodos utilizados (Neighbor – Joining, Maximum Parsimony e Maximum likelihood).

3.2 Comparação dos genes *rpoB* e *gyrB* como ferramentas taxonômicas alternativas ao *rrn* no grupo das proteobactérias

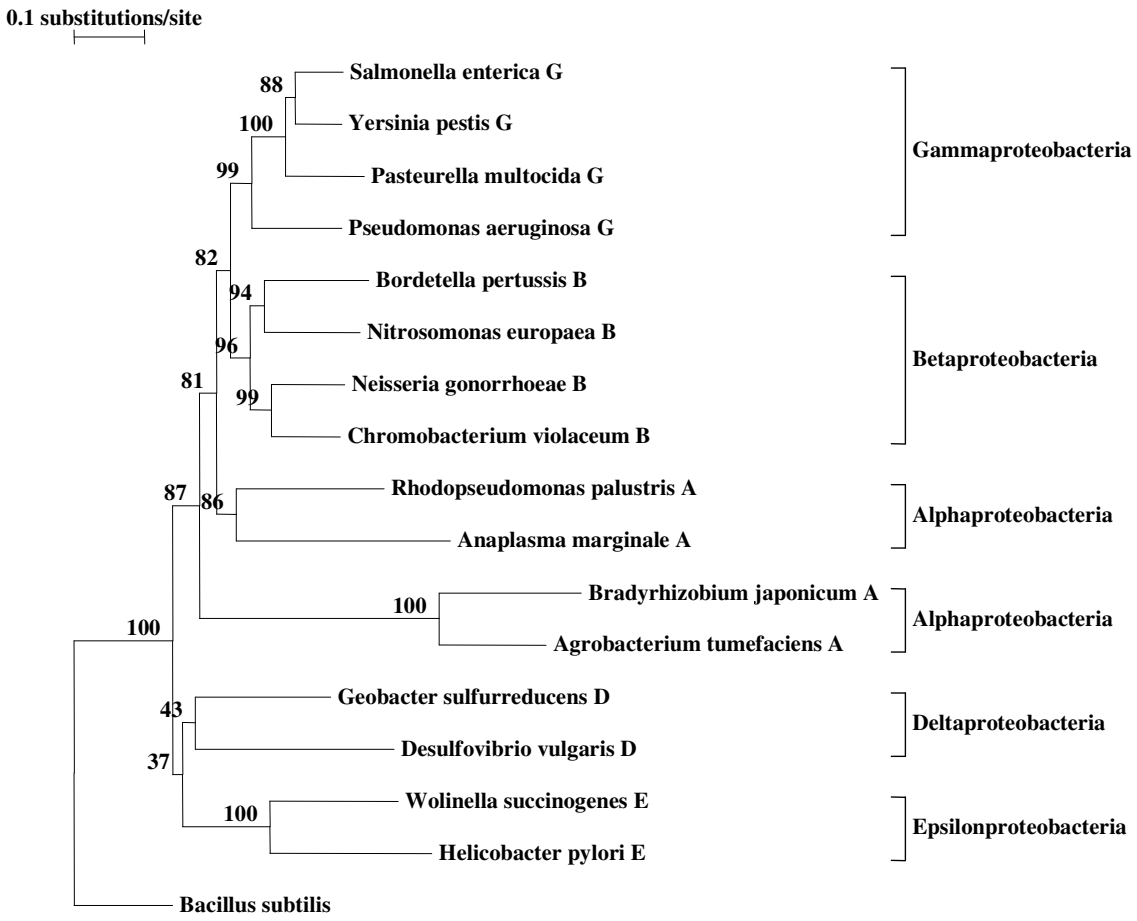
Análises filogenéticas de Neighbor – Joining, Maximum Likelihood e Maximum Parsimony baseadas nas seqüências dos genes 16S rRNA, *rpoB* e *gyrB* resultaram em uma organização similar e confiável nos clusters evidenciados nas árvores 3, 4 e 5. Valores bootstrap maiores ou iguais a 95% obtidos pelo método de Neighbor – Joining foram observados em 10 de 15 nós para o gene 16S rRNA, em 7 de 15 para o gene *gyrB* e em 11 de 15 para o *rpoB*. Proporções similares foram observadas nos demais métodos.



Árvore 3 - Relações filogenéticas proteobactérias baseadas em 16S rRNA. Método: Neighbor - Joining. 1352 posições analisadas. 1000 bootstraps (valores percentuais). Clusters destacados com colchetes foram suportados pelos três métodos utilizados (Neighbor – Joining, Maximum Parsimony e Maximum likelihood).



Árvores 4 - Relações filogenéticas proteobactérias baseadas em *rpoB*. Método: Neighbor - Joining. 3234 posições analisadas. 1000 bootstraps (valores percentuais). Clusters destacados com colchetes foram suportados pelos três métodos utilizados (Neighbor – Joining, Maximum Parsimony e Maximum likelihood).



Árvore 5 - Relações filogenéticas proteobactérias baseadas em *gyrB*. Método: Neighbor – Joining. 1737 posições analisadas. 1000 bootstraps (valores percentuais). Clusters destacados com colchetes foram suportados pelos três métodos utilizados (Neighbor – Joining, Maximum Parsimony e Maximum likelihood).

3.3 Comparar os genes *rpoB* e *gyrB* como ferramentas taxonômicas alternativas ao *rrn* na família das *Enterobacteriacea*.

3.3.1 Matrizes de distância e gráficos de entropia

As matrizes de distância dos genes 16S rRNA, *rpoB* e *gyrB* demonstraram que os genes *gyrB* e *rpoB* apresentam um maior grau de polimorfismo.

Gráficos representando a entropia de cada posição do alinhamento das seqüências dos genes 16S rRNA (figura 11)

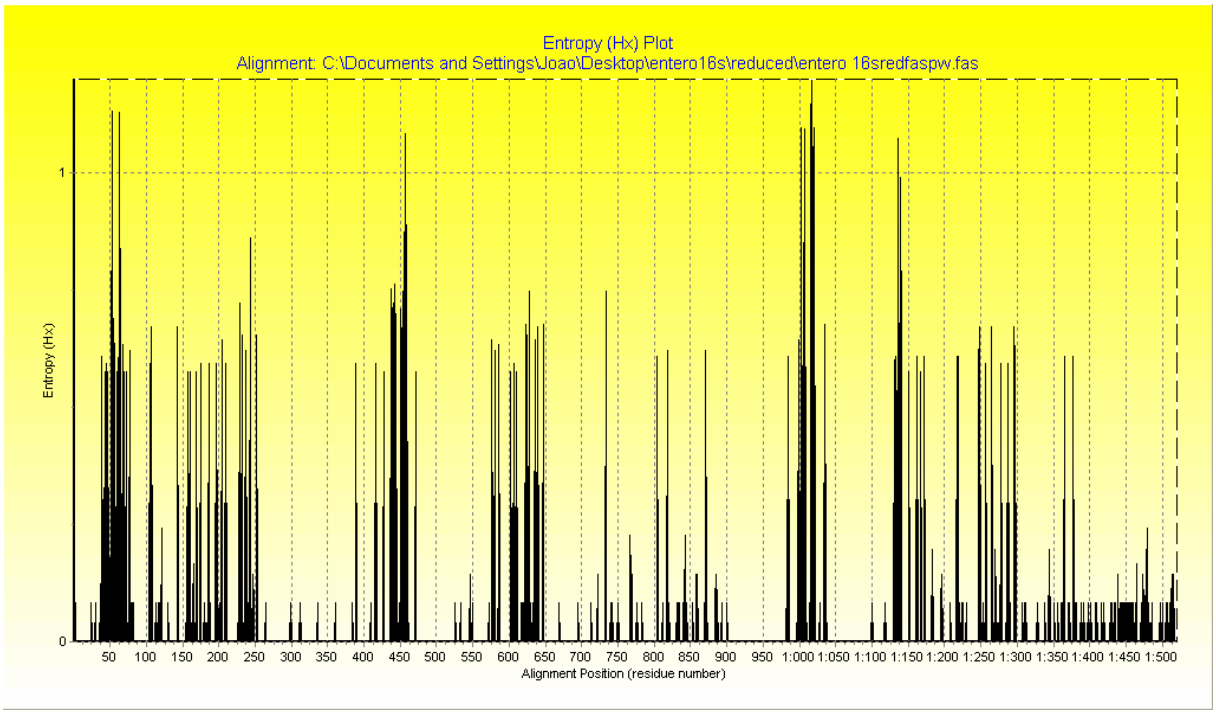


Figura 11

Gráficos representando a entropia de cada posição do alinhamento das seqüências dos genes *rpoB* (figura 12) e *gyrB* (figura 13)

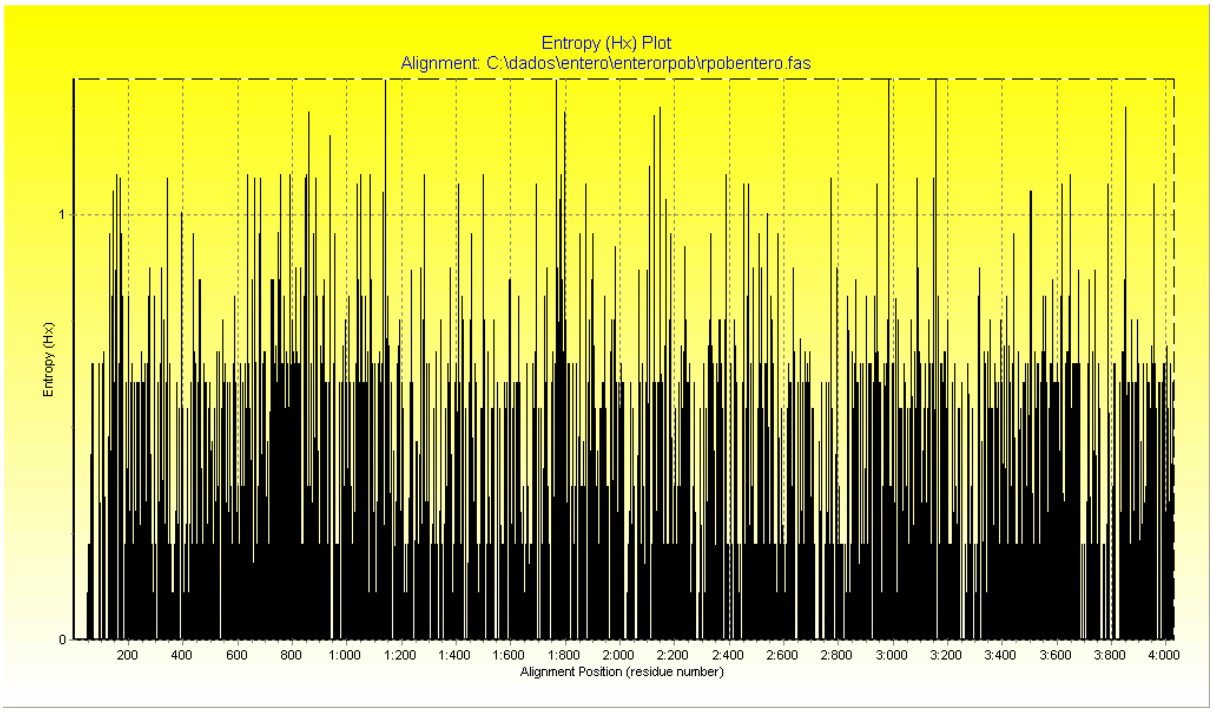


Figura 12

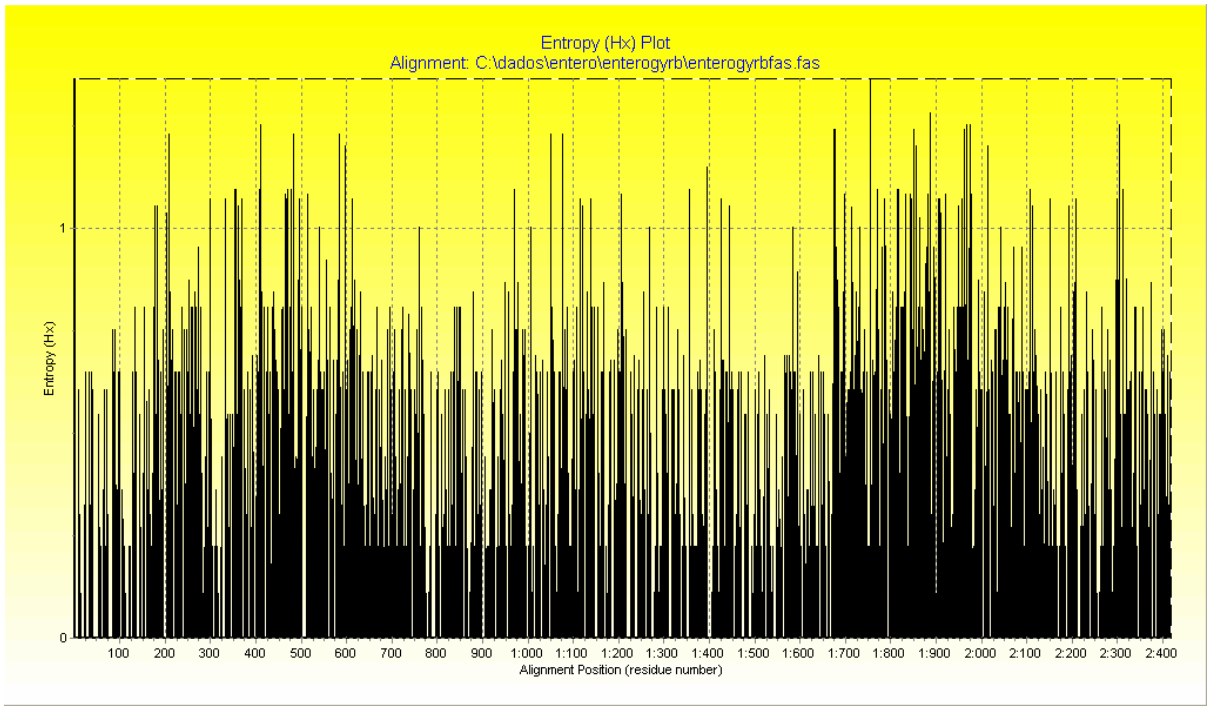
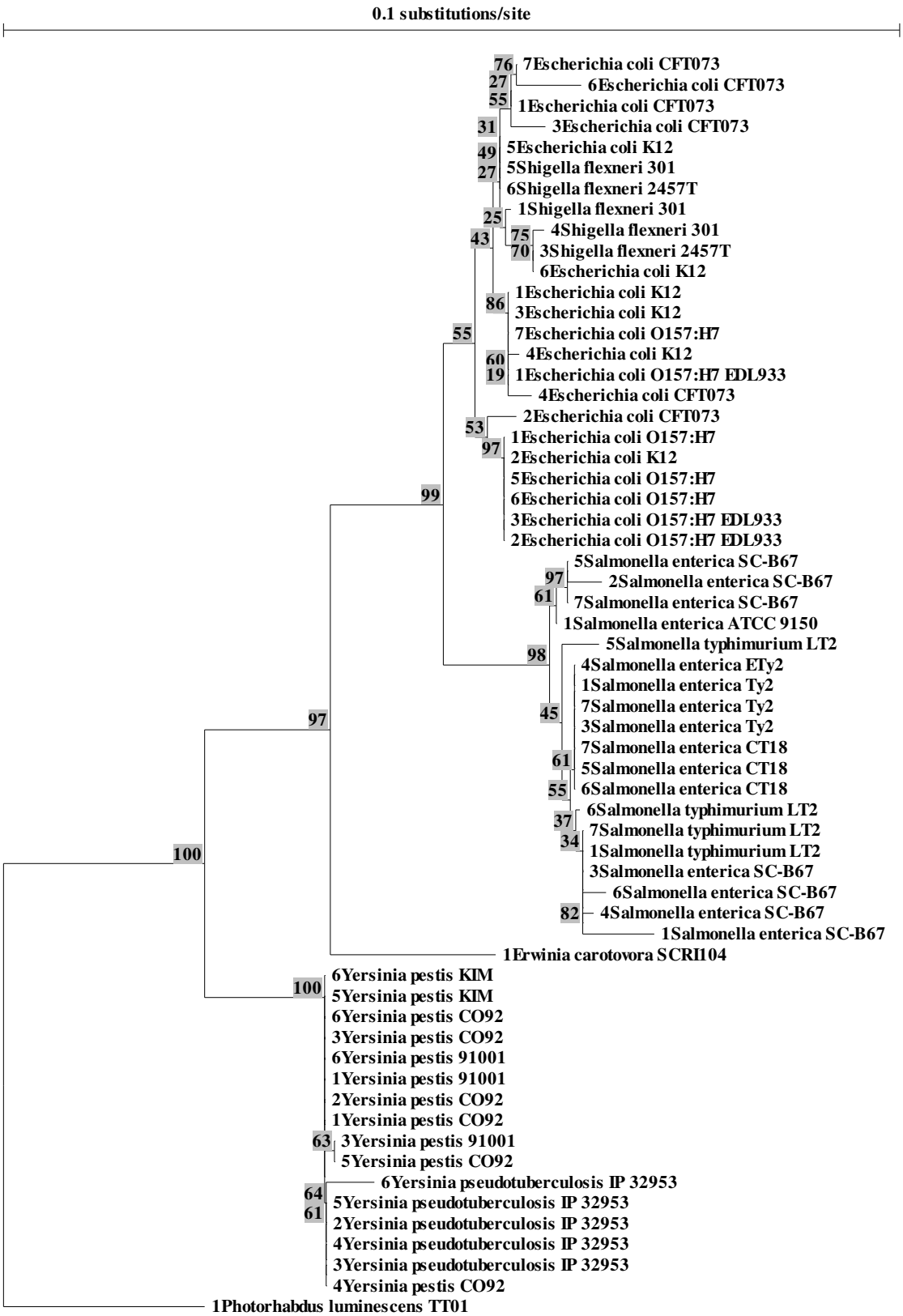


Figura 13

3.3.2 Análises Filogenéticas

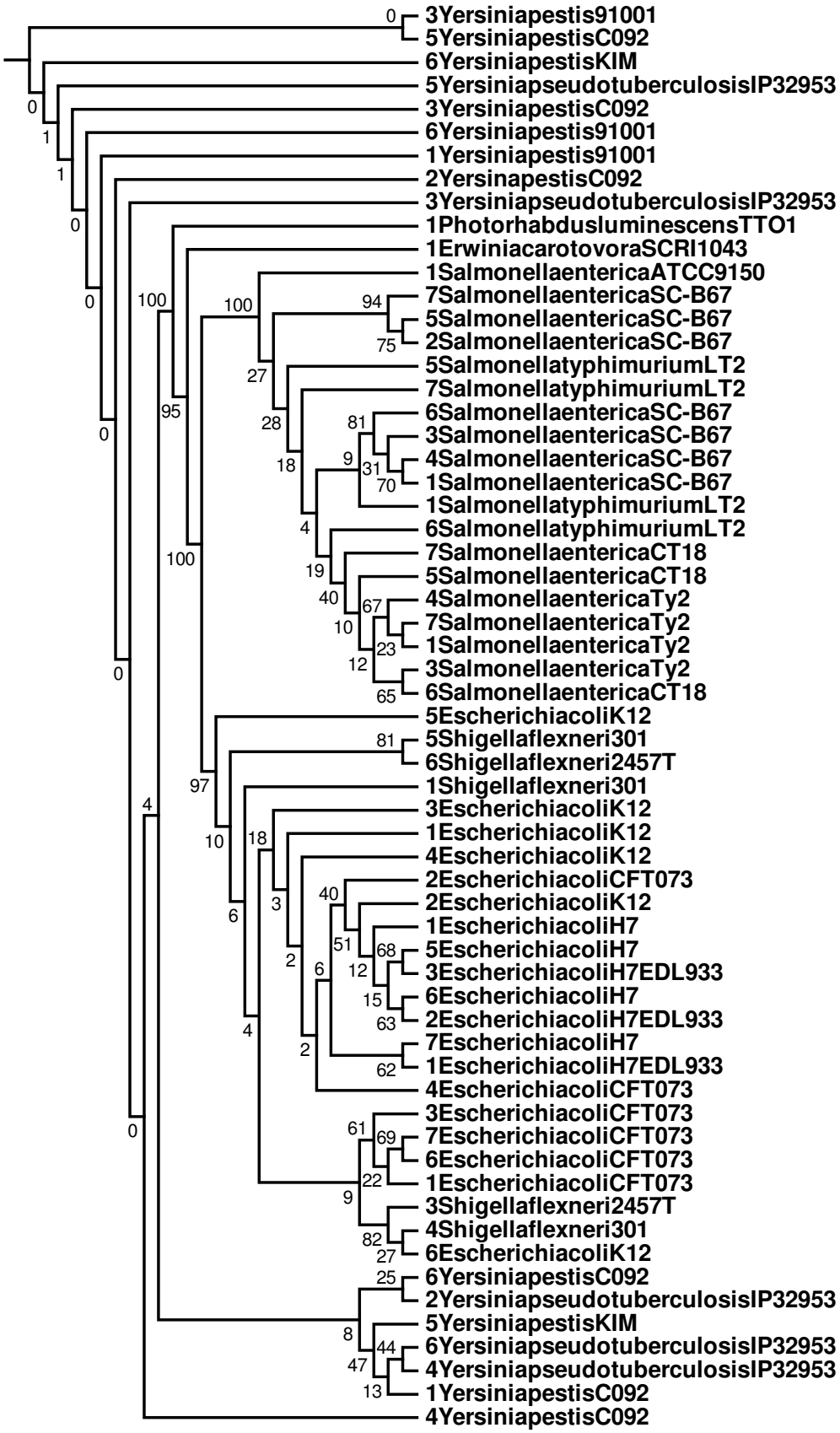
Utilizando-se o 16S rRNA foram obtidas 5 árvores mais parcimoniosas (mesmo número de passos necessários), 1 árvore pelo método de Neighbor – Joining e 1 árvore pelo método de Maximum Likelihood.. Análises filogenéticas entre membros da família *Enterobacteriaceae* não demonstraram clusters condizentes com a taxonomia apresentada no Bergey´s Outline (http://www.cme.msu.edu/bergeys/) em nenhuma das análises (árvores 6, 7 e 8).

Análises filogenéticas de Neighbor – Joining, Maximum Likelihood e Maximum Parsimony baseadas nas seqüências dos genes, *rpoB* e *gyrB* resultaram em uma organização similar e confiável nos clusters evidenciados nas árvores 9 e 10 suportadas por altos valores bootstrap.

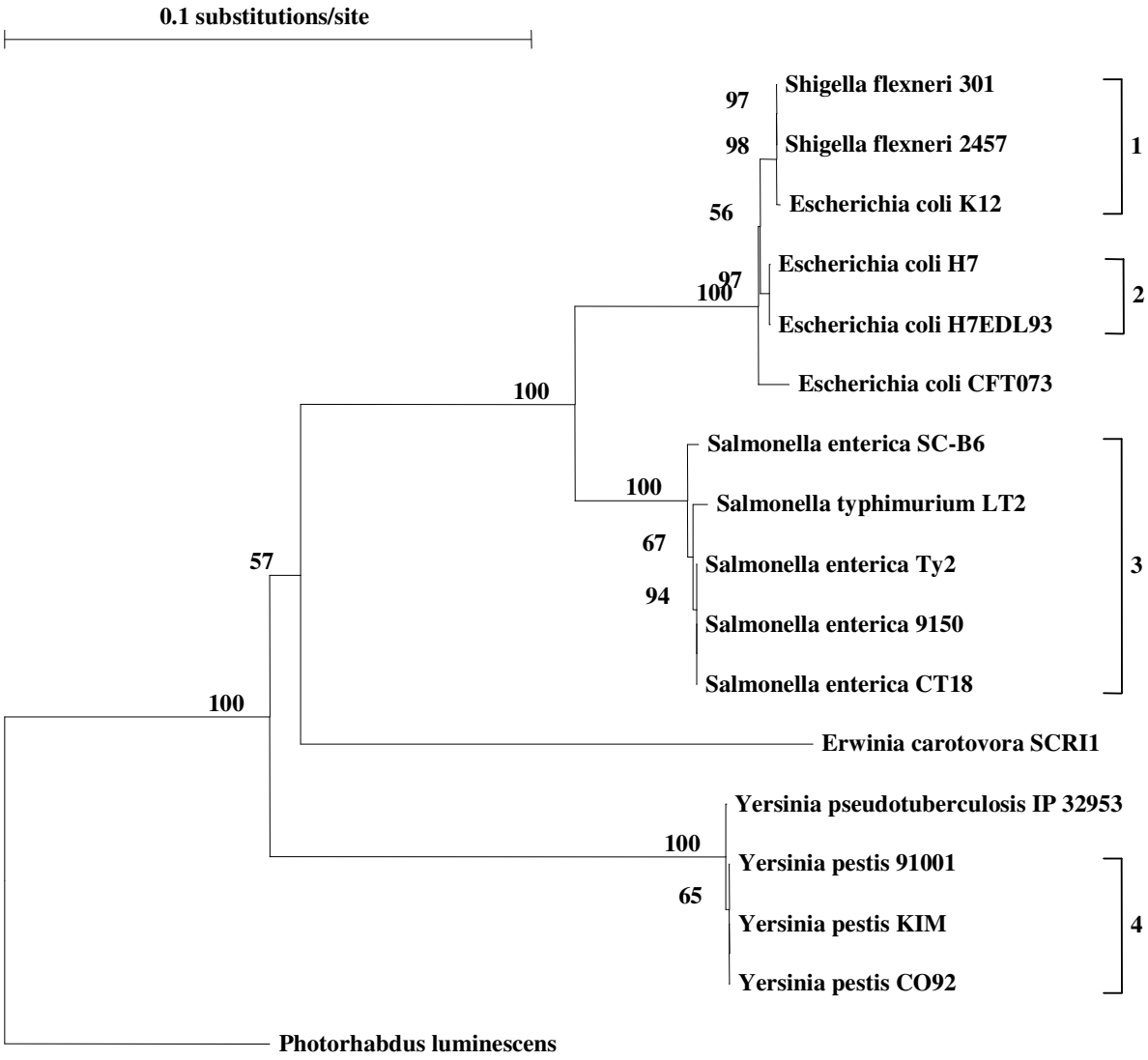


Árvore 6 - Relações filogenéticas enterobactérias baseadas em 16S rRNA contendo diversas cópias do gene. Método: Neighbor – Joining. 1440 posições analisadas. 1000 bootstraps (valores percentuais).

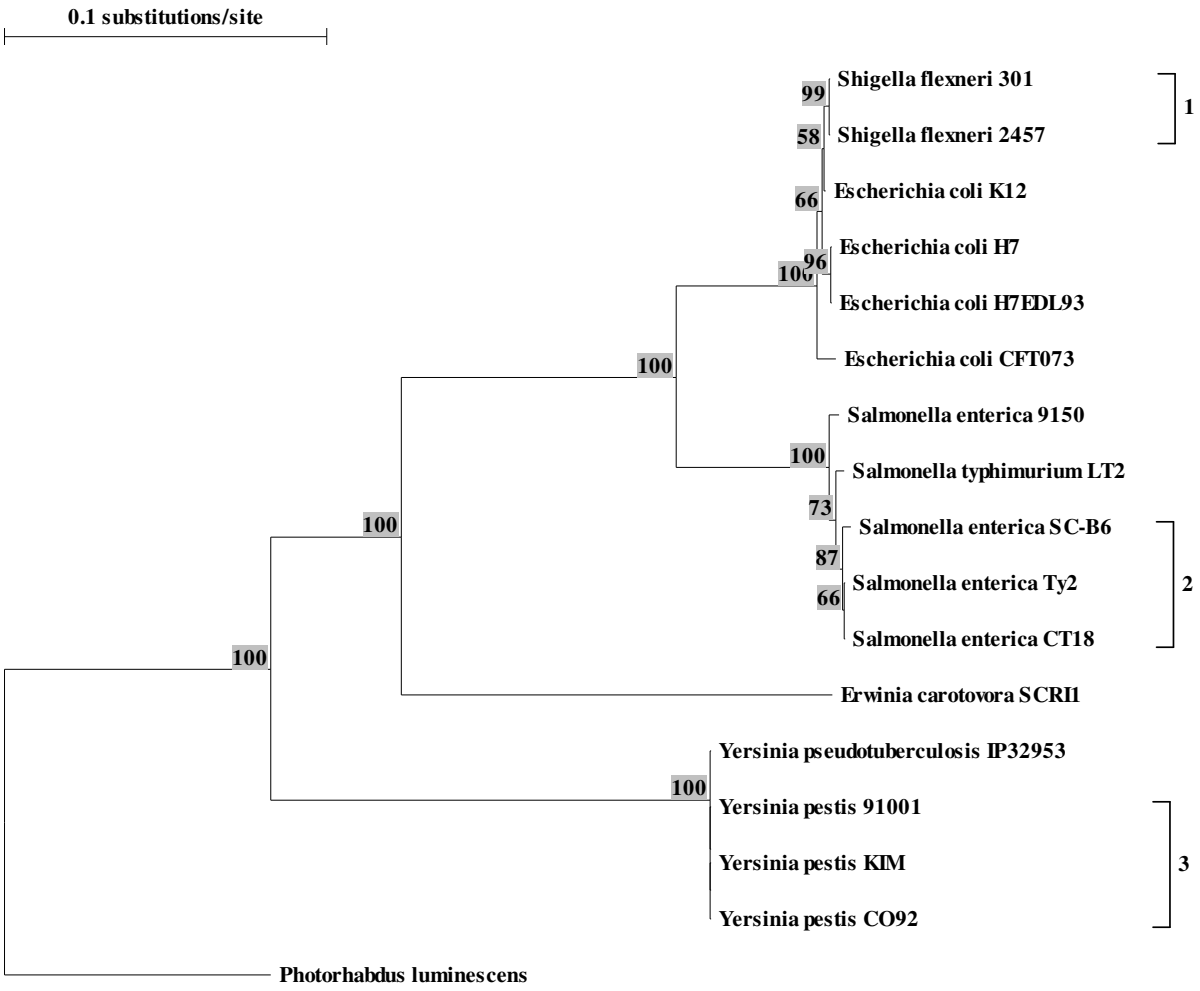
Árvore 7 - Relações filogenéticas enterobactérias baseadas em 16S rRNA contendo diversas cópias do gene. Método: Maximum Likelihood. 1440 posições analisadas. 100 bootstraps (valores percentuais).



Árvore 8 - Relações filogenéticas enterobactérias baseadas em 16S rRNA contendo diversas cópias do gene. Método: Maximum Parsimony. 1440 posições analisadas (118 informativas). 324 passos necessários. 1000 bootstraps (valores percentuais).



Árvore 9 - Relações filogenéticas enterobactérias baseadas em *rpoB*. Método: Neighbor - Joining. 4028 posições analisadas. 1000 bootstraps (valores percentuais). Clusters destacados com colchetes foram suportados pelos três métodos utilizados (Neighbor – Joining, Maximum Parsimony e Maximum likelihood).



Árvore 10 - Relações filogenéticas enterobactérias baseadas em *gyrB*. Método: Neighbor - Joining. 2412 posições analisadas. 1000 bootstraps (valores percentuais). Clusters destacados com colchetes foram suportados pelos três métodos utilizados (Neighbor – Joining, Maximum Parsimony e Maximum likelihood).

4. DISCUSSÃO

4.1 Gene *rrn*

O 16S rRNA é ferramenta mais utilizada em taxonomia e filogenia microbiana, mas deve-se ter em mente que este gene está presente em cópias múltiplas e variáveis na maioria dos genomas bacterianos seqüenciados (tabela 13) e que todas as analises filogenéticas e taxonômicas baseadas no sistema 16S rRNA estão sujeitas a este fato. Além disso, freqüentemente a variação existente nas seqüências de rRNA não são suficientes para discriminar espécies muito relacionadas (árvores 6,7 e 8). Por ser um gene muito conservado, o 16S rRNA é útil para estabelecer relações entre organismos distantes filogenéticamente. O gene demonstra que cloroplastos e cianobactérias formam um grupo distinto de outras bactérias, além de agrupar taxonomicamente as proteobactérias de acordo com o Bergey´s manual. Porém, a sua utilidade decai gradativamente conforme o espectro filogenético analisado diminui, ficando muito difícil diferenciar espécies muito relacionadas e virtualmente impossível diferenciar estirpes.

4.2 Gene *rpoB*

O gene *rpoB* já foi utilizado por outros autores *Bartonella* spp. (Renesto, 2001), *Staphylococcus* (Drancourt & Raoult, 2002), *Bosea* spp. e *Afipia* spp. (Khamis *et al*., 2003), *Mycobacterium* spp. (Kim *et al*., 1999) e *Legionella* spp. (KO *et al*., 2002).

Analises baseadas no gene *rpoB* foram capazes de agrupar cloroplastos e cianobactérias em um cluster separado de outros microrganismos. O gene também foi capaz de agrupar membros do grupo das proteobactérias de acordo com a taxonomia apresentada no Bergey´s Manual, sendo que *Pseudomonas aeruginosa* e *Anaplasma marginale* foram consideradas de difícil acesso, embora sua posição na árvore filogenética tenha sido suportada pelos 3 métodos analíticos empregados com altos valores bootstrap. Na família das *Enterobacteriaceae* o gene foi capaz de diferenciar não só espécies como estirpes intimamente relacionadas.

4.3 Gene *gyrB*

Analises baseadas no gene *gyrB* foram capazes de agrupar membros do grupo das proteobactérias de acordo com a taxonomia apresentada no Bergey´s Manual sem exceções pelos 3 métodos analíticos empregados com altos valores bootstrap. Na família das *Enterobacteriaceae* o gene foi capaz de diferenciar não só espécies como estirpes intimamente relacionadas.

4.4 Considerações finais

Desde o trabalho de Carl Woese em 1987 o 16S rRNA é a molécula mais utilizada para estabelecer relações evolutivas entre os microorganismos, apresentando diversas vantagens, como ser universalmente distribuída e apresentar regiões conservadas alternadas com regiões variáveis, o que teoricamente permitiria uma boa resolução filogenética. Porém, alguns autores destacaram as limitações deste gene como ferramenta em análises filogenéticas e taxonômicas (figura 9). Estas limitações incluem o fato do gene estar presente em várias cópias no genoma de diversos microorganismos (ver tabela 13). Além disto, a resolução filogenética intra específica em alguns casos não é boa o suficiente (figura 9). Neste trabalho foi avaliada a aplicabilidade dos genes *rpoB* e *gyrB* como ferramentas taxonômicas e filogenéticas.

Tabela 12. Variabilidade intra-genomica do 16S rRNA de bactérias e arqueobactérias cujos genomas foram completamente seqüenciados. (J. Hashimoto & J. Klappenbach rrndb: the Ribosomal RNA Operon Copy Number Database, http://rrndb.cme.msu.edu/rrndb)

Organismo	***Número de copias 16S rRNA***	***16S Diferenças (NT)***
Aquifex aeolicus VF5	2	-
Bacillus subtilis ATCC 23857	10	01-15
Campylobacter jejuni ATCC 700819	3	-
Deinococcus radiodurans ATCC 13939	3	0 - 2
Escherichia coli ATCC 10798	7	0-19
Haemophilus influenzae ATCC 51907	6	-
Helicobacter pylori 26695	2	-
Methanococcus jannaschii DSMZ 2661	2	3
M.thermoautotrophicum ATCC 29096	2	2
Neisseria menigitidis MC 58	4	-
Treponema pallidum ATCC 25870	2	-
Ureaplasma urealyticum serovar 3	2	1
Vibrio cholerae ATCC 39315	8	0-14
Xyella fastidosa 9a5c	2	-

Os resultados deste trabalho indicam que os genes *rpoB* e *gyrB* tem uma maior taxa de evolução do que o *rrn* (figuras 9, 10 e 11). O poder de discriminação destes genes pode ser acessado rapidamente com a média dos valores de divergência demonstrados nas matrizes de distância (62, 389 e 350 para os genes 16S rRNA, *rpoB* e *gyrB* respectivamente – ver tabelas 10, 11 e 12).

Embora existam membros da família das *Enterobacteriaceae* que sejam genotipicamente muito semelhantes, estes necessitam de métodos de discriminação devido a sua importância clínica. Os resultados deste trabalho indicam que tanto o gente *rpoB* quanto o *gyrB* são bons candidatos para estudos nos campos da microbiologia médica e ambiental.

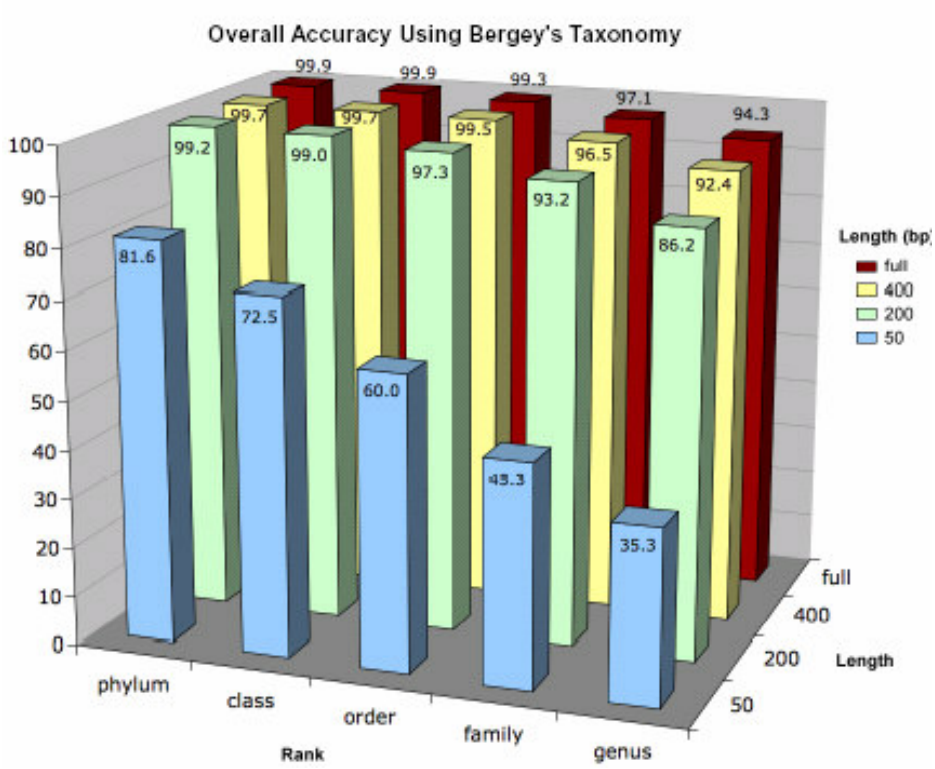


Figura 14. Acuidade do 16S rRNA para identificação de microrganismos tendo como base a taxonomia apresentada no Bergey´s Manual (fonte: Ribosomal Database Project http://rdp.cme.msu.edu/)

Tabela 9: Matriz de distância. Gene 16S rRNA 1467 posições analisadas (sítios sem dados ou com gaps foram removidos antes da análise).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1																	
2	0.000																
3	0.001	0.001															
4	0.001	0.001	0.000														
5	0.055	0.055	0.055	0.055													
6	0.055	0.055	0.056	0.056	0.002												
7	0.056	0.056	0.057	0.057	0.003	0.003											
8	0.053	0.053	0.054	0.054	0.004	0.003	0.005										
9	0.054	0.054	0.055	0.055	0.005	0.004	0.005	0.001									
10	0.053	0.053	0.054	0.054	0.010	0.010	0.011	0.006	0.007								
11	0.059	0.059	0.060	0.060	0.031	0.030	0.031	0.029	0.030	0.023							
12	0.059	0.059	0.060	0.060	0.031	0.030	0.031	0.029	0.030	0.023	0.001						
13	0.059	0.059	0.060	0.060	0.032	0.031	0.033	0.032	0.033	0.026	0.004	0.004					
14	0.070	0.070	0.070	0.070	0.042	0.042	0.043	0.041	0.042	0.037	0.015	0.015	0.012				
15	0.059	0.059	0.059	0.059	0.024	0.023	0.025	0.022	0.023	0.029	0.007	0.007	0.010	0.018			
16	0.048	0.048	0.049	0.049	0.047	0.048	0.047	0.048	0.048	0.045	0.049	0.049	0.048	0.060	0.051		
17	0.077	0.077	0.078	0.078	0.083	0.085	0.084	0.084	0.085	0.082	0.088	0.089	0.089	0.100	0.091	0.081	

[1] #*Yersinia pseudotuberculosis IP 32953*

[2] # *Yersinia pestis 91001*

[3] # *Yersinia pestis CO92*

[4] # *Yersinia pestis KIM*

[5] # *Shigella flexneri 2457T*

[6] # *Shigella flexneri 301*

[7] # *Escherichia coli CFT073*

[8] # *Escherichia coli K12*

[9] # *Escherichia coli H7EDL933*

[10] # *Escherichia coli H7*

[11] # *Salmonella enterica CT18*

[12] # *Salmonella enterica Ty2*

[13] # *Salmonella typhimurium LT2*

[14] # *Salmonella enterica SC-B67*

[15] # *Salmonella enterica 9150*

[16] # *Erwinia carotovora SCRI1043*

[17] # *Photorhabdus luminescens*

Média: 0.042

Média com as 7 cópias do gene 16S rRNA: **0.038**

Tabela 10: Matriz de distância. Gene *rpoB* 4028 posições analisadas (sítios sem dados ou com gaps foram removidos antes da análise).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1																	
2	0.148																
3	0.139	0.144															
4	0.148	0.001	0.144														
5	0.148	0.000	0.144	0.001													
6	0.140	0.145	0.008	0.144	0.145												
7	0.138	0.148	0.064	0.148	0.148	0.066											
8	0.139	0.146	0.062	0.146	0.146	0.063	0.011										
9	0.139	0.146	0.062	0.146	0.146	0.063	0.011	0.000									
10	0.139	0.144	0.001	0.144	0.144	0.007	0.064	0.062	0.062								
11	0.139	0.144	0.000	0.144	0.144	0.007	0.064	0.061	0.061	0.000							
12	0.166	0.154	0.163	0.154	0.154	0.165	0.160	0.161	0.161	0.162	0.162						
13	0.139	0.145	0.008	0.145	0.145	0.008	0.067	0.063	0.063	0.008	0.008	0.166					
14	0.136	0.147	0.062	0.147	0.147	0.063	0.011	0.008	0.008	0.062	0.061	0.161	0.063				
15	0.137	0.148	0.062	0.148	0.148	0.063	0.011	0.008	0.008	0.062	0.062	0.162	0.063	0.002			
16	0.137	0.148	0.062	0.148	0.148	0.064	0.011	0.008	0.008	0.062	0.062	0.161	0.064	0.002	0.000		
17	0.148	0.000	0.144	0.001	0.000	0.145	0.148	0.146	0.146	0.144	0.144	0.154	0.145	0.147	0.148	0.148	

[1] # *Erwinia carotovora SCRI1043*

[2] # *Yersinia pestis CO92*

[3] # *Salmonella enterica CT18*

[4] # *Yersinia pseudotuberculosis IP 32953*

[5] # *Yersinia pestis KIM*

[6] # *Salmonella typhimurium LT2*

[7] # *Escherichia coli FT073*

[8] # *Escherichia coli H7EDL933*

[9] # *Escherichia coli H7*

[10] # *Salmonella enterica 9150*

[11] # *Salmonella enterica Ty2*

[12] # *Photorhabdus luminescens*

[13] # *Salmonella enterica SC-B67*

[14] # *Escherichia coli K12*

[15] # *Shigella flexneri 2457T*

[16] # *Shigella flexneri 301*

[17] # *Yersinia pestis 91001*

Média: 0.097

Tabela 11: Matriz de distância. Gene *gyrB* 2412 posições analisadas (sítios sem dados ou com gaps foram removidos antes da análise).

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
1																	
2	0.227																
3	0.200	0.189															
4	0.001	0.228	0.199														
5	0.000	0.227	0.200	0.001													
6	0.223	0.011	0.188	0.223	0.223												
7	0.211	0.100	0.193	0.211	0.211	0.097											
8	0.210	0.104	0.194	0.209	0.210	0.101	0.019										
9	0.225	0.016	0.192	0.226	0.225	0.013	0.097	0.101									
10	0.227	0.000	0.189	0.228	0.227	0.011	0.100	0.104	0.016								
11	0.218	0.052	0.240	0.218	0.218	0.252	0.246	0.247	0.252	0.253							
12	0.225	0.015	0.193	0.225	0.225	0.016	0.101	0.104	0.019	0.015	0.253						
13	0.214	0.101	0.192	0.212	0.214	0.098	0.018	0.015	0.097	0.101	0.248	0.100					
14	0.213	0.103	0.194	0.212	0.213	0.101	0.022	0.008	0.101	0.103	0.247	0.104	0.012				
15	0.213	0.103	0.194	0.212	0.213	0.101	0.022	0.008	0.101	0.103	0.247	0.104	0.012	0.000			
16	0.210	0.104	0.194	0.209	0.210	0.101	0.019	0.000	0.101	0.104	0.247	0.104	0.015	0.008	0.008		
17	0.000	0.227	0.200	0.001	0.000	0.233	0.211	0.210	0.225	0.227	0.218	0.225	0.214	0.213	0.213	0.210	

[1] # *Yersinia pestis CO92*

[2] # *Salmonella enterica CT18*

[3] # *Erwinia carotovora SCRI1043*

[4] # *Yersinia pseudotuberculosis IP 32953*

[5] # *Yersinia pestis KIM*

[6] # *Salmonella typhimurium LT2*

[7] # *Escherichia coli CFT073*

[8] # *Escherichia coli H7EDL933*

[9] # *Salmonella enterica 9150*

[10] # *Salmonella enterica Ty2*

[11] # *Photorhabdus luminescens*

[12] # *Salmonella enterica SC-B67*

[13] # *Escherichia coli K12*

[14] # *Shigella flexneri 2457T*

[15] # *Shigella flexneri 301*

[16] # *Escherichia coli H7*

[17] # *Yersinia pestis 91001*

Média: 0.145

5. REFERÊNCIAS BIBLIOGRÁFICAS

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W., LIPMAN, D. L. 1990. Basic local alignment search tool. J. Mol. Biol. 215: 403-410

BROWN, R. & DOOLITTLE, W. 1997. Archaea and the Prokaryote-to-Eukaryote Transition. Microbiology and Molecular Biology Reviews. Dec:456–502

CANNONE, J.J., SUBRAMANIAN, S., SCHNARE, M.N., COLLET, J.R., D'SOUZA, L.M., DU, Y., FENG, B., LIN, N., MADABUSI, L.V., MULLER, K.M., PANDE, N., SHANG, Z., YU, N. & GUTELL, R.R. 2002. The Comparative RNA Web (CRW) Site: An Online Database of Comparative Sequence and Structure Information for Ribosomal, Intron, and other RNAs. BioMed Central Bioinformatics, 3:2. [Correction: BioMed Central Bioinformatics. 3:15.

CAVALLI – SFORZA, L. L., EDWARDS, A. W. 1967. Phylogenetic analisys. Models and estimation procedures. Am. J. Hum. Genet. May 19 (3): Suppl 19:233.

CHATTON, E. 1937. Titres et travoux scientifiques (1906-1937) de Edouard Chatton, Seton, France: Sottano.

COLE, J. R., CHAI, B., MARSH, T. L., FARRIS, R. J., WANG, Q., KULAM, S. A., CHANDRA, S., MCGARRELL, D. M., SCHMIDT, T. M., GARRITY, G. M. & TIEDJE, J. M. 2003. The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. Nucleic Acids Res. Jan 1,31(1):442-3

COPELAND, H.F. 1938. The kingdoms of organisms. Quarterly Review of Biology 13:384-420.

COULOMBE, B. & BURTON, Z. F. 1999. DNA Bending and Wrapping around RNA Polymerase: a "Revolutionary" Model Describing Transcriptional Mechanisms. Microbiology and Molecular Reviews. June p. 457–478

DAHLLÖF, I., HARRIET, B. & KJELLEBERG, S. 2000. *rpoB*-Based Microbial Community Analysis Avoids Limitations Inherent in 16S rRNA Gene Intraspecies Heterogeneity. Applied and Environmental Microbiology. Aug. 2000, p. 3376–3380

DANIE` LE GADELLE, JONATHAN FILE´, BUHLER, C. & FORTERRE, P. 2003. Phylogenomics of type II DNA topoisomerases. BioEssays. 25:232–242, Wiley Periodicals, Inc.

DRANCOURT, M. & RAOULT, D**.** 2002. *rpoB* gene sequence-based identification of *Staphylococcus* species. J. Clin. Microbiol. **40:**1333–1338.

EFRON, B. & G. GONG, G. 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. The American Statistician, 37(1):36-48.

EFRON, B. & HALLORAN, E. 1996. Bootstrap confidence levels for phylogenetic trees. Proc. Natl. Acad. Sci. USA, 93:7085-7090.

ERDMANN, V. A., WOLTERS, J., PIELER, T., DIGWEED, M., SPECHT, T., ULBRICH, N. 1987. Evolution of organisms and organelles as studied by comparative computer and biochemical analises of ribosomal 5S RNA structure. Ann N Y Acad Sci. 1987;503:103-24.

FELSENSTEIN, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783-791

FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981;17(6):368-76.

FITCH, W. M. & MARGOLIASH, E. 1967. Constrution of phylogenetic trees: A method based on mutation distances as estimated from cytocrome c sequences is of general applicability. Science. 155:279-284.

GALTIER, N. & GOUY, M. 1995. Inferring phylogenies from sequences of unequal base composition. Proceedings of the National Academy of Science USA. 92: 11317-11321.

GALTIER, N., GOUY, M. and GAUTIER, C. 1996. SeaView and Phylo_win, two graphic tools for sequence alignment and molecular phylogeny. Comput. Applic. Biosci., **12**, 543-548.

GOODFELLOW, M. & O'DONNELL, A.G. 1993. Roots of Bacterial Systematics. In: HandBook of New Bacterial Systematics, pp 3-54 (eds). Academic Press, London.

HALL, T. A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids. Symp. Ser. 41:95-98.

HOLT, J. G.; KRIEG, N. R.; SNEATH, P. H. A.; STALEY, J. T. & WILLIANS, S. T. 1994. Bergey's Manual of Determinative Bacteriology. 9ª ed., Baltimore, Williams & Wilkins Co.

JENSEN, R. A. 2001. Orthologs and paralogs – We need to get it right. Genome Biol. 2(8): interactions 1002.1-1002.3.

JIN, L. & NEI, M. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. 7(1):82-102

JONES & KRIEG. 1984. Serology and Chemotaxonomy. In: Bergey's Manual of Sytematic Bacteriology. S. T. Williams, M. E. Sharpe & J. G. Holt (eds.) vol.4, Baltimore, Willians & Wilkins, pp. 2303 – 2305.

JUKES, T. H. & CANTOR, C.R. 1969. Evolution of protein molecules. In HN Munro, ed., Mammalian Protein Metabolism, pp. 21-132, Academic Press, New York

KASAI, H., WATANABE, K., GASTEIGER, E., BAIROCH, A., ISONO, K., YAMAMOTO, S. & HARAYAMA, S. 1998. Construction of the *gyrB* database for the identification and classification of bacteria. pp13-21 in Genome Informatics, Miyano,S. and Takagi,T.(eds), Universal Academic Press Inc., Tokyo.

KHAMIS, A., COLSON, P., RAOULT, D. & LA SCOLA, B. 2003. Usefulness of *rpoB* gene sequencing for identification of *Afipia* and *Bosea* species, including a strategy for the choice of discriminative partial sequences. Appl. Environ. Microbiol. 69:6740–6749.

KEELING, J. & DOOLITTE, W. 1995. Archaea: Narrowing the gap between prokaryotes and eukaryotes. Proc. Natl. Acad. Sci. 92:5761-64.

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

KIM, B. J., LEE, S. H., LYU, A.M., KIM, S.J., BAI. G.H., KIM, S.S., CHAE, G.T., KIM, E.C., CHA, C.Y., KOOK, Y.H. 1999. Identification of mycobacterial species by comparative sequence analysis of the RNA polymerase gene (*rpoB*). J. Clin. Microbiol. **37:**1714-1720

KO, K. S., LEE, H. K. PARK, M. Y., LEE, K. H., YUM, Y. J., YUN, S. Y., WOO, H. Miyamoto & KOOK, Y. H. 2002. Application of RNA polymerase betasubunit gene (*rpoB*) sequences for the molecular differentiation of *Legionella* species. J. Clin. Microbiol. 40:2653–2658.

KOONING, E. V. 2001. An apology for orthologs – or brave new nemes. Genome biol. 2(4): comment1005.1–comment1005.2.

KRIEG, N. 1989. Introduction to Systematics. In: Bergey's Manual of Sytematic Bacteriology. S. T. Williams, M. E. Sharpe & J. G. Holt (eds.) vol.4, Baltimore, Willians & Wilkins, pp. 2303 – 2305.

KUMAR, S., Tamura K., Nei, M. 2004. **M**EGA3: Integrated Software for Molecular Evolutionary Genetics Analysis and Sequence Alignment Briefings in Bioinformatics 5:150-163.

KUMAZAKI, T., HORI, H. OSAWA, S. 1983. Phylogeny of protozoa deduced from 5S rRNA sequences. J Mol Evol. 19(6):411-9.

LEE, S.H., KIM, B.J., KIM, K.H., PARK, K.H., KIM, S.J., KOOK, Y.Y. 2000. Differentiation of *Borrelia burgdorferi* sensu lato on the basis of RNA polymerase gene (*rpoB*) sequences. *J. Clin. Microbiol.* 38:2557-2562.)

LESK, A. M. 2002. Introduction to Bioinformatics Oxford: Oxford University Press.

MACRAE, A. 2000. The use of 16S rDNA methods in soil microbial ecology. Braz. J Microbiol 31, 77-82.

MACRAE, A., RIMMER D. L. & O´Donnell A. G. 2000. Novel Bacterial diversity recovered from the rhizosphere of oilseed rape (*Brassica napus*) determined by the analysis of 16S ribosomal DNA. Antonie Leeuwenhoek Int. J. Gen. Microbiol. 78 (1). 13-21.

MADIGAN, M. T., MARTINKO, J. M. & PARKER, J. 2000. Brock Biology of microorganisms.

MARGULIS, L. & SCHWARTZ, K.V. 1998. Five Kingdoms: an Illustrated Guide to the Phyla of Life on Earth. 3rd Ed. WH Freeman & Co, San Francisco.

MARGULIS, L & STOLZ, J. F. 1984. Cell symbiosis [correction of symbioisis] theory: status and implications for the fossil record. Adv Space Res. 1984;4(12):195-201.

MAYR, E. 1998. Two empires or three?. Proc. Natl. Acad. Sci. USA. Vol. 95, pp. 9720–9723, August.

MOLLET, C. M. & RAOULT, D. 1997. *rpoB* sequence analysis as a novel basis for bacterial identification. Mol. Microbiol. 26:1005-1011

MOUNT, D.W. 2001. Bioinformatics. Sequence and Genome Analysis. 1 ed., New York, Cold Spring Harbor Laboratory Press.

PALLERONI, N.J. 2003. Prokaryote taxonomy of the 20th century and the impact of studies on the genus Pseudomonas: a personal view. Microbiology. 149:1-7.

RENESTO, P., DRANCOURT, M., RAOULT, D. 2000. *rpoB* gene analysis as a novel strategy for identification of spirochetes from the genera *Borrelia*, *Treponema*, and *Leptospira*. *J. Clin. Microbiol.* 38:2200-2203)

RENESTO, P., J. GOUVERNET, M. DRANCOURT, V. Roux, and D. Raoult. 2001. Use of *rpoB* gene analysis for detection and identification of *Bartonella* species. J. Clin. Microbiol. 39:430–437.

SAITOU, N. & NEI, M. 1987. The neighbour joining method: a new method for constructing phylogenetic trees. Mol. Biol. Evol. 6: 514-525.

SNEATH, P. H. A. 1989. Numerical taxonomy. In: Bergey's Manual of Sytematic Bacteriology. S. T. Williams, M. E. Sharpe & J. G. Holt (eds.) vol.4, Baltimore, Willians & Wilkins, pp. 2303 – 2305.

STUDIER, J. A. & KEPPLER, K. J. 1988. A note on the neighbor-joining algorithm of Saitou and Nei. Mol Biol Evol. 1988 Nov;5(6):729-31.

SWOFFORD, D.J. and G.J. Olsen, G.J. 1996. Phylogenetic Inference. In Molecular Systematics. D.M. Hillis & C. Moritz, 2nd ed., Massachusetts, Sinauer Associates, p 407 - 514

TAJIMA, F. & NEI, M. 1984. Estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. 1(3):269-285.

TAMURA, K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol Biol Evol. 1992 Jul;9(4):678-87.

THOMSON, J.D., HIGGINS, G.D. & GIBSON , T. J. 1994. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting positions-specific gap penalities and weight matrix choice. Nucleic Acids Res. 22: 4673-4680

THOMPSON, J.D., GIBSON, T.J., PLEWNIAK, F., JEANMOUGIN, F. and HIGGINS, D.G. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research, 24:4876-4882.

TORSVIK, V., JOSTEN, G. & FRIDA, L. 1990. High diversity in DNA of soil bacteria. Applied and environmental microbiology. P. 782-787

VAN DE PEER, Y. & DE WACHTER R. 1993. TREECON: a software package for the construction and drawing of evolutionary trees. J. Mol. Evol. 30: 463-476.

WEN-HSUING LI. 1981. Simple method for constructing phylogenetic trees from distance matrices. Proc. Natl. Acad. Sci. Vol 78, No 2, pp. 1085-1089

WHITTAKER, R. H. 1969. New concepts of kingdoms or organisms. Evolutionary relations are better represented by new classifications than by the traditional two kingdoms. Science. Jan 10,163(863):150-60.

WOESE, C. 1987. Bacterial Evolution. Microbiol Rev. 1987 Jun,51(2):221-71.

WOESE, C. R. 1998. Default taxonomy: Erns Mayr´s view if the microbial world. Proc. Natl. Acad. Sci. 95: 11043-11046.

WOESE, C. R. 1998. The universal ancestor. Proc. Nat. Acad. Sci. 95: 6854–6859.

WOESE, C. 2002. On the evolution of cells. Proc Natl Acad Sci U S A. 2002 Jun 25,99(13):8742-7.

YAMAMOTO, S. & Harayama, S. 1996. Phylogenetic analysis of Acinetobacter strains based on the nucleotide sequences of gyrB genes and on the amino acid sequences of their products.
Int J Syst Bacteriol. 1996 Apr;46(2):506-11.