Construção de Filogenias Baseadas em Genomas Completos

Este exemplar corresponde à redação final da Dissertação devidamente corrigida e defendida por Karina Zupo de Oliveira e aprovada pela Banca Examinadora.

Campinas, 05 de Março de 2010.

João Meidanis
Instituto de Computação - UNICAMP
(Orientador)

Dissertação apresentada ao Instituto de Computação, UNICAMP, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO IMECC DA UNICAMP
Bibliotecária: Crisllene Queiroz Custódio – CRB8 / 7966

Oliveira, Karina Zupo de
OL4c Construção de filogenias baseadas em genomas completos/ Karina Zupo de Oliveira -- Campinas, [S.P. : s.n.], 2010.

Orientador : João Meidanis
Dissertação (Mestrado) - Universidade Estadual de Campinas, Instituto de Computação.

1. Biologia Computacional. 2. Filogenia - processamento de dados. 3. Genomas. 4. Homologia (Biologia). 5. Vibrio. 6. Vibrionaceae. I. Meidanis, João. II. Universidade Estadual de Campinas. Instituto de Computação. III. Título.

Título em inglês: Phylogenies construction based on whole genomes

Palavras-chave em inglês (Keywords): 1. Computational biology. 2. Phylogeny - data processing. 3. Genomes. 4. Homology (Biology). 5. Vibrio. 6. Vibrionaceae.

Área de concentração: Biologia Computacional

Titulação: Mestre em Ciência da Computação

Banca examinadora: Prof. Dr. João Meidanis (IC-UNICAMP)
Prof. Dr. Fabiano L. Thompson (IB-UFRJ)
Prof. Dr. Zanoni Dias (IC-UNICAMP)

Data da defesa: 05/03/2010

Programa de Pós-Graduação: Mestrado em Ciência da Computação

TERMO DE APROVAÇÃO

Dissertação Defendida e Aprovada em 05 de março de 2010, pela Banca examinadora composta pelos Professores Doutores:

Prof. Dr. Fabiano Lopes Thompson
Instituto de Biologia / UFRJ

Prof. Dr. Zanoni Dias
IC / UNICAMP

Prof. Dr. João Meidanis
IC / UNICAMP

Instituto de Computação
Universidade Estadual de Campinas

Construção de Filogenias Baseadas em Genomas Completos

Karina Zupo de Oliveira

Março de 2010

Banca Examinadora:

- João Meidanis (Orientador)
- Fabiano L. Thompson
  Instituto de Biologia - UFRJ
- Zanoni Dias
  Instituto de Computação - Unicamp
- Marcelo Menossi (Suplente)
  Instituto de Biologia - Unicamp
- Arnaldo Moura (Suplente)
  Instituto de Computação - Unicamp

Resumo

Contexto: A classificação de espécies começou sendo determinada pelas características fenotípicas dos organismos. Logo que o DNA foi descoberto, o sistema de classificação passou também a utilizar-se das características genotípicas. Ao longo dos últimos anos, avanços científicos permitiram que fossem sequenciados genomas completos. A cada ano, o número de genomas completamente sequenciados aumenta, e, com isso, é cada vez maior o número de trabalhos que tentam utilizar-se do maior número possível de genes para comparar dois ou mais organismos com o objetivo de melhor entender o relacionamento entre as diversas espécies.
Experimento: Este trabalho executa comparações de pares de cromossomos de um grupo de 10 genomas completos da família *Vibrionaceae* e um genoma completo da bactéria *Escherichia coli* como externo ao grupo. As homologias entre as proteínas são determinadas através da base de famílias *Protein Clusters* (NCBI). A seguir, árvores ultramétricas e a classificação COG das proteínas são utilizadas para resolver as paralogias correspondentes. Após isto, as proteínas únicas, que representam os eventos de perda e ganho de genes, são eliminadas, de forma a igualar o conteúdo dos cromossomos. Tipicamente, 50% das proteínas originais do pares de organismos de mesma família "sobrevivem" para serem utilizadas no cálculo da distância de rearranjo. Menos proteínas sobrevivem nas comparações com a bactéria externa ao grupo. A distância total é calculada pela soma do número de proteínas eliminadas e da distância de ordenação, medida através da distância de rearranjo dos cromossomos.
Resultados: As comparações produziram matrizes de distâncias utilizadas para inferir árvores filogenéticas através do algoritmo *Neighbor-Joining* (NJ). As árvores filogenéticas encontradas mostraram-se congruentes em topologia com a árvore produzida pelo gene 16S rRNA. Isto mostra que a comparação de genomas completos é uma proposta sensata. Os desafios agora são aperfeiçoar os detalhes. O material suplementar (Apêndice A) contém uma implementação computacional dos experimentos.

Abstract

Context: Species classification was originally determined by phenotypic characteristics. With the advent of DNA sequencing, the classification system started using genotypes as well. Over the last decades, scientific progress allowed complete sequencing of genomes. Each year, the number of genomes completely sequenced increases, and with it, the number of works trying to use as much genes as possible to compare two or more organisms, in order to get a better understand of the relationship between several species.
Experiment: This work executes a pairwise chromosome comparison from a set of 10 complete genomes from the *Vibrionaceae* family and one complete *Escherichia coli* genome as an outgroup. In our experiment, the homologies between proteins are assessed using the *Protein Clusters* (NCBI) database. In the next step, paralogies are resolved using ultrametric trees and COG classification. In the sequel, the loss and gain events are treated, thus, proteins present in only one chromosome from the pair are eliminated, in order to equalize the set of families in both chromosomes. Typically, 50% of the original proteins survive in comparisons between organisms of the same family (comparisons with the outgroup yield less survivors). The total distance is calculated by adding the number of eliminated proteins with the order distance, which is measured by the rearrangement distance beetween the chromosomes.
Results: Genome comparison produces distance matrices used to infer the phylogenetic trees through the *Neighbor-Joining* (NJ) algorithm. The phylogenetic trees generated are congruent regarding the topology with the tree inferred using the 16S rRNA gene. Also, in order to run a deeper investigation, the experiment was executed with some variations such as not resolving the paralogies using ultrametric trees or only classifying proteins using COG database. Supplemental material (Appendix A) contains the experiment computational implementation.

Agradecimentos

Ao meu orientador, Professor João Meidanis, pela competência com que orientou esta minha tese e por todos os ensinamentos transmitidos a mim. Também agradeço pela sua grande paciência, pela sua compreensão às minhas falhas e limitações, pelo seu incentivo constante, e por acreditar em mim.

À colega Patrícia Pilisson Côgo por sempre estar disposta a responder meus questionamentos e por me enviar todas as informações necessárias para compreender seu trabalho.

Aos meu pais, Marco e Rosangela, por me proverem a melhor educação possível, mesmo nos momentos mais difíceis, nunca falhando.

Ao meu esposo Claudio por seu companheirismo e por sua compreensão.

Aos meus chefes que, permitindo minha ausência do trabalho, demonstraram compreensão e incentivo a este mestrado.

Aos colegas de trabalho que cursaram matérias comigo pelas valiosas horas de estudo em conjunto.

Por fim, agradeço ao Instituto de Computação da UNICAMP e a seus excelentes funcionários e professores.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Taxonomia é a ciência que lida com a classificação (= criação de novos taxa), identificação (= alocação de linhagens dentro de espécies conhecidas) e nomenclatura [64]. O sistema de classificação dos organismos de espécies até reinos e domínios, criado por Lineu, é regido pelas regras de nomenclatura desta ciência. No passado, a taxonomia de bactérias era definida através de testes fenotípicos e características morfológicas. O fenótipo é o conjunto de características físicas possuídas por um organismo, em parte influenciadas pelo genótipo, e o genótipo é o conjunto de genes e regiões intergênicas de um organismo. Nos anos 70, a técnica de hibridização de DNA-DNA [11], produziu um refinamento dos grupos taxonômicos. Com o sequenciamentos de genes, a utilização das sequências do gene 16S RNA passaram a ser utilizadas como base da estrutura de taxonomia dos procariontes [42].

Cohan [10] num estudo da validade do sistema atual de classificação de espécies, tendo focado-se nos organismos procariontes, nos diz que vem crescendo o consenso entre os cientistas de que conceito de espécie de bactérias não exibe as mesmas propriedades dinâmicas especiais apresentadas pelo conceito de espécies biológico clássico, aquele aplicado aos organismos eucariontes. Em uma espécie de eucariontes, além dos indivíduos possuírem semelhança fenotípica e genotípica, estes indivíduos são capazes de procriar, produzindo descendentes férteis. Já nos indivíduos procariontes, a reprodução nem sempre é sexuada. Décadas de estudos utilizando hibridização de DNA-DNA mostraram que existe grande diversidades entre as bactérias classificados numa mesma espécie. Por fim, Cohan [10] ainda discute três visões contemporâneas da natureza da diversidade biológica entre as bactérias: o conceito biológico de espécies aplicado às bactérias [13], o conceito de ecótipos [9], e o conceito de sem espécie [28].

Para Gevers e colegas [26] as espécies dos procariontes são definidas com base em um caráter operacional e centradas em humanos e doenças, sendo importante em divesas áreas (indústrias farmacêutica e alimentícia, por exemplo). Acreditam que já é tempo de considerar um casamento entre este caráter operacional e o conceito teórico de espécies, o que traria melhorias e avanços no sistema taxonômico atual, principalmente para as espécies que ainda estão pouco caracterizadas ou que ainda nem foram descobertas. Mesmo aplicando as diversas técnicas para classificar organismos em espécies hoje existentes, estes autores encontraram casos em que organismos agrupados numa mesma espécie têm características fenotípicas diferentes, ou seja, organismos não tão próximos podem ser agrupados numa mesma espécie. Isso porque importantes características fenotípicas podem estar em genes que não são muito estáveis ou bem conservados, e, portanto, desconsiderados por algumas destas técnicas de classificação.

Também, com o sequenciamento dos genomas completos, iniciaram-se investigações mais detalhadas sobre outros tipos de eventos evolutivos. Por eventos evolutivos entendem-se, além das mutações em genes, eventos de perda e ganho de genes, duplicações de genes, transferência horizontal de genes entre espécies, reversões de grandes trechos do genoma, transposições, translocações, fissões e fusões de cromossomos, entre outros. Técnicas baseadas em comparações de genomas completos surgiram na tentativa de suprir deficiências dos métodos anteriores. Uma das vantagens destas técnicas é que elas utilizam todo o conjunto de genes, ao invés de basear-se em apenas um gene (16S rRNA) ou um grupo de genes (MLST), e, por utilizarem mais genes, espera-se que as relações de distâncias entre os ramos da árvore filogenética sejam mais acuradas. Além disso, importantes eventos evolutivos tais como reversões, transposições, duplicações, perdas de genes e transferências horizontais acontecem em níveis da ordem de genes em genomas e não da ordem de nucleotídeos em genes.

Visando aproveitar o aumento da disponibilidade de genomas completos e utilizar o máximo de genes possíveis para inferir relações entre espécies, neste trabalho, executamos um experimento que compara onze genomas completos. Destes, dez genomas, disponíveis no NCBI em Junho de 2009, são da família dos *Vibrionaceae* - uma família que compreende organismos de cinco diferentes gêneros, incluindo o vibrião causador da cólera, uma doença grave e que ainda causa anualmente milhares de mortes em países em desenvolvimento. Também utilizamos um genoma completo da bactéria *Escherichia coli*, que pertence à família das *Enterobacteriaceae*, como grupo externo (*outgroup*, em inglês). O modelo de comparação é dividido em três fases. A primeira fase tratará de classificar os genes de um genoma em famílias universais de genes homólogos. A segunda fase tem por objetivo restringir cada par de genomas a um conjunto comum de genes, dando trata-

mento adequado aos eventos de duplicações de genes (paralogias) e aos eventos de perda e ganho de genes. Por fim, a distância de rearranjo é calculada para o par de genomas. A árvore filogenética é inferida com base na matriz das distâncias. Neste trabalho foi desenvolvida uma ferramenta de comparação de genomas que poderá apoiar biólogos no estudo, entendimento e melhoraria do sistema de taxonomia dos organismos procariontes.

Capítulo 2

Conceitos

2.1 Homologia e Famílias

A palavra homologia (do grego *homo*, igualmente e *logia*, raciocínio) significa "concordância", Fitch esclarece que, no âmbito biológico, homologia é o estudo da ancestralidade comum de estruturas funcionais e genômicas contidas em organismos diferentes. Genes são homólogos se estes genes têm origem em um ancestral comum [18]. Sendo em muitos casos impossível obter os ancestrais comuns, homologia entre proteínas e DNA é frequentemente avaliada com base na similaridade de sequências, e, desta forma, se duas sequências de nucleotídeos tem alto grau de similaridade então provavelmente estas são homólogas. Porém, mesmo sendo muito similares, tais sequências poderiam ter surgido de ancestrais diferentes. Portanto, a noção de ancestralidade é parte chave da definição. Além disso, não se pode confundir estruturas homólogas com estruturas análogas. As asas de uma ave e de um inseto são análogas, pois ambas permitem voar, porém, estas asas não são estruturas homólogas, pois não tem origem numa estrutura ancestral comum. Já as asas do morcego e dos pássaros, além de serem análogas, possuem estrutura óssea homóloga, estrutura esta originada da nadadeira de peixe (Figura 2.1). Note que, apesar de possuírem estrutura óssea homóloga, as asas de ambas espécies se desenvolveram independentemente, apresentando uma estrutura diferente.

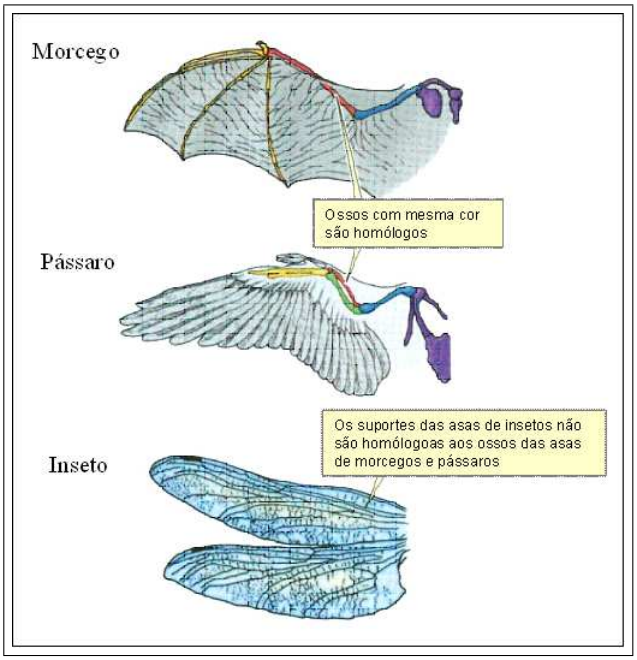

Figura 2.1: Exemplo de estrutura homóloga: estrutura óssea das asas de alguns pássaros e dos morcegos tem como ancestral comum a nadadeira do peixe.

Existem três tipos principais de homologia: ortologia, paralogia e xenologia. Dois genes, pertencentes a espécies diferentes, são **ortólogos** se tem origem num gene ancestral comum e estes genes forem adquiridos via transferência vertical, ou seja, por hereditariedade. A Figura 2.2 apresenta o gene 1.A da espécie A que é ortólogo ao gene 1.B da espécie B. Comparações de genomas mostram que relações de ortologia entre genes de espécies distantes podem ser estabelecidas para a grande maioria dos genes [37].

Genes **parálogos** são genes originados de eventos de duplicação de genes numa mesma espécie. Eventualmente, durante a evolução, estes genes duplicados podem assumir funções diferentes das executadas pelo gene original. Na Figura 2.3 acontece um evento de duplicação criando-se um novo gene 1.1.A que é parálogo ao gene 1.A. Dois genes são ditos **xenólogos**, se são homólogos, e, além disso, um deles for adquirido por evento de transferência horizontal de genes (THG). A dificuldade de reconhecerem-se genes xenólogos introduz desvios na criação de filogenias.

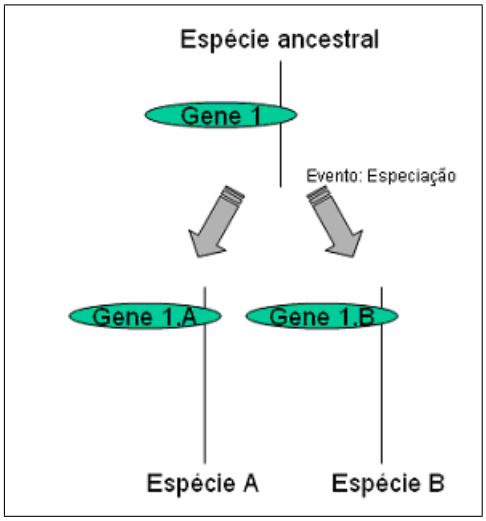

Figura 2.2: Exemplo de ortologia.

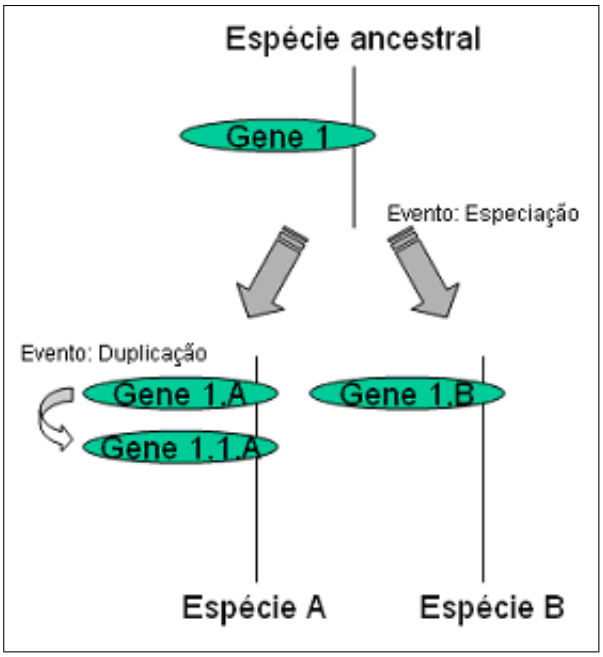

Figura 2.3: Exemplo de paralogia.

A Figura 2.4 mostra um exemplo de xenologia. O gene 1.C da espécie C é transferido horizontalmente para o genoma da espécie B, que adquiri assim o gene 1.C.B. Assim, os genes 1.A e 1.C.B são xenólogos.

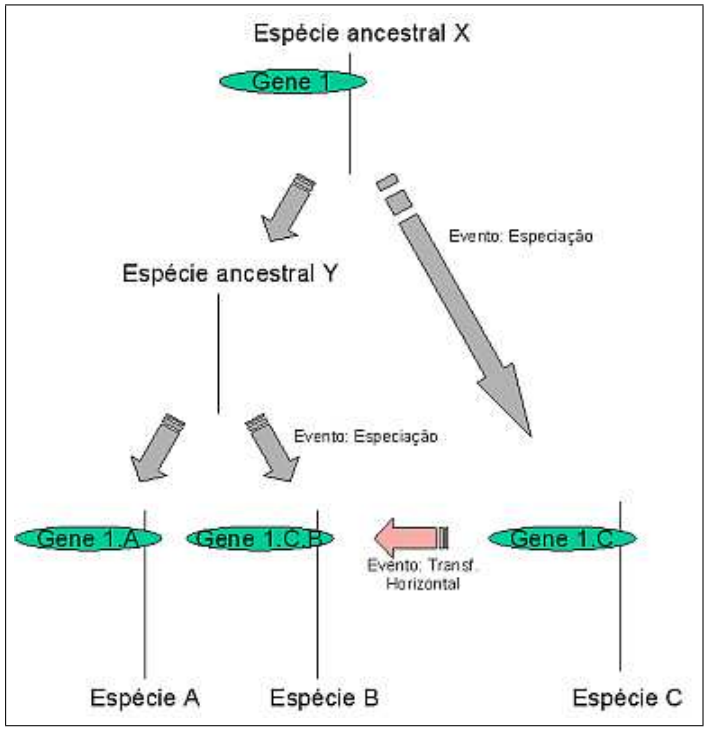

Figura 2.4: Exemplo de xenologia.

Um exemplo clássico de homologia é a encontrada no gene da hemoglobina. Um gene duplicado de hemoglobina dos mamíferos evoluiu permitindo que o feto realize a extração de oxigênio do sangue da mãe. Este gene é parálogo ao gene de hemoglobina que permite a um indivíduo adulto transportar o oxigênio. Estes dois genes são ortólogos ao gene da hemoglobina que possui a função de transporte de oxigênio nos pássaros.

A determinação de homologia não é um problema simples. Analise a Figura 2.5. Nesta Figura, após um evento de duplicação é produzido um novo gene Y a partir do gene X. A seguir acontece um evento de especiação, originando as espécies A, B e C. Os genes X.A, X.B e X.C são ortólogos. Os genes Y.A, Y.B e Y.C são ortólogos. Os genes X e Y de cada espécie são parálogos entre si. Após esta especiação, na espécie A, o gene X.A perde-se, e, na espécie C, o gene Y.C perde-se. O relacionamento entre os genes Y.A e X.C restantes é o de paralogia e não de ortologia. Em algumas ocasiões pode ser muito difícil determinar-se corretamente este relacionamento, principalmente quando na análise não se possui o genoma da espécie ancestral ou o genoma da espécie B, que manteve os dois genes ancestrais X e Y. Este problema foi denominado de problema da perda de gene por Fitch [20]. Koonin [37] chama estes genes de pseudoortólogos.

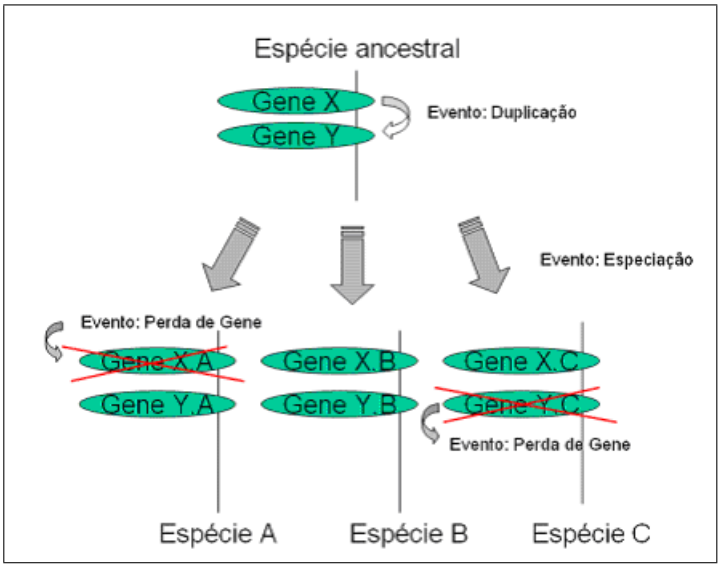

Figura 2.5: Exemplo de genes pseudoortólogos.

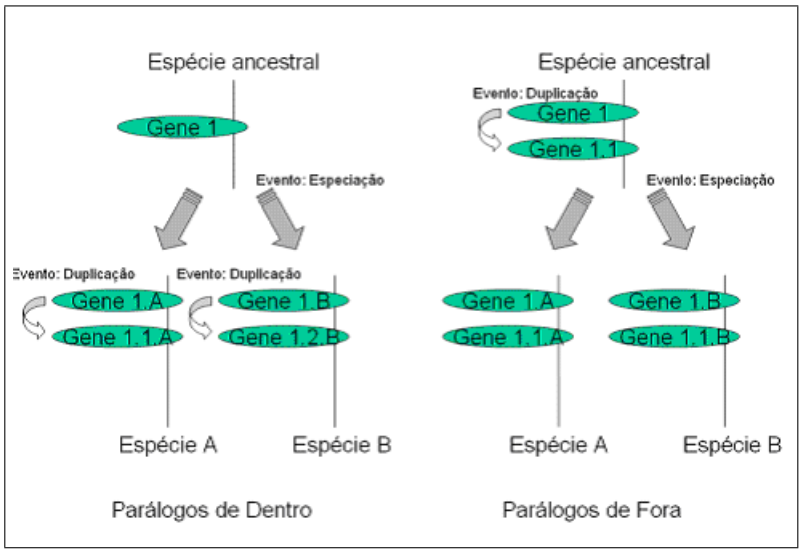

Figura 2.6: Genes parálogos de dentro e de fora.

Genes parálogos podem ser divididos em dois casos: parálogos de dentro (*inparalogs*, em inglês) e parálogos de fora (*outparalogs*, em inglês) [57]. Parálogos de dentro são os genes parálogos cuja duplicação ocorreu depois da especiação. Parálogos de fora são os genes parálogos cuja duplicação ocorreu antes da especiação. A Figura 2.6 esquematiza estes casos. Além disso, existem ainda os genes pseudoparálogos [37], que, numa análise contendo apenas um genoma, podem ser classificados falsamente como parálogos, porém,

são genes que resultam da combinação de herança vertical e transferência horizontal. Na Figura 2.7 os genes 1.B e 1.C.B, encontrados na espécie B, são pseudoparálogos.

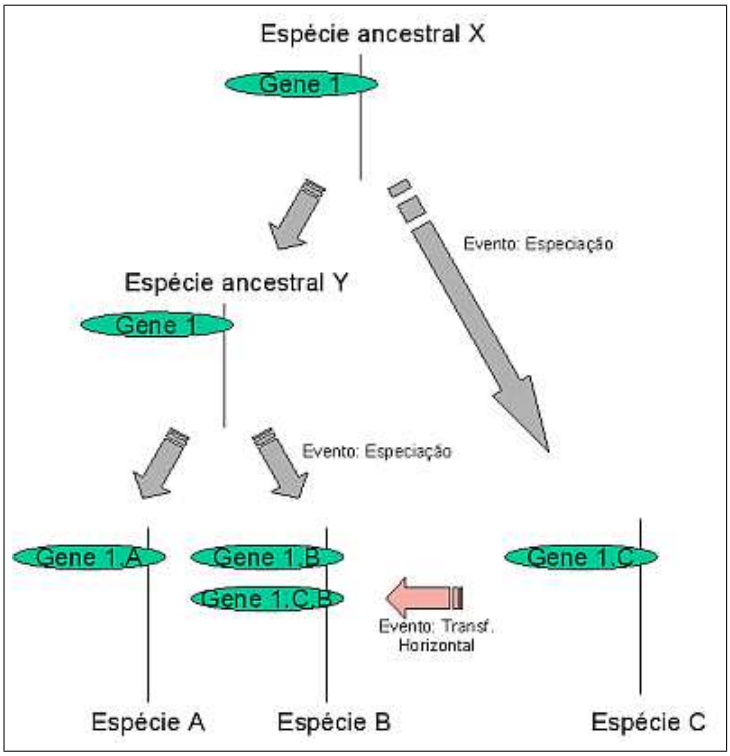

Figura 2.7: Exemplo de genes pseudoparálogos.

2.2 Rearranjo de Genomas

Considere dois cromossomos distintos, contendo ambos um conjunto igual e conhecido de genes, isto é, possuem o mesmo conteúdo. Estes cromossomos possuem uma ordenação conhecida para seus conjuntos de genes. Rearranjo de genomas é o processo pelo qual são realizados sucessivas operações de rearranjo de genes num dos cromossomos com o objetivo de ordenar seus genes na mesma ordenação do outro cromossomo.

Na maior parte dos casos, o que se deseja encontrar é a sequência mínima de operações necessárias para que um cromossomo tenha seus genes rearranjados na mesma sequência dos genes do outro cromossomo. Supõe-se que os rearranjos de genomas com número mínimo de operações sejam os mais prováveis de ter acontecido na Natureza, ou seja, são os mais parcimoniosos. No final desta ordenação, um valor de distância entre os dois cromossomos é calculado de acordo com o número de operações realizado e seus respectivos pesos. Este rearranjo é estudado com o objetivo de entender melhor as relações

de parentesco e evolução das espécies.

As principais operações de rearranjo são reversão, transposição, fusão, fissão e translocação. Algumas vezes, na literatura, o termo evento de rearranjo é utilizado no mesmo sentido que atribuímos aqui o termo operação de rearranjo de gene. A seguir serão explicadas cada uma destas operações.

Para explicar a operação de reversão precisamos antes explicar sobre **orientação** de genes. No cromossomo, os nucleotídeos que formam os genes estão organizados no que chamamos de estrutura de dupla hélice [55], que é composta por dois filamentos de nucleotídeos em formato espiral. De acordo com a composição química dos nucleotídeos, estes filamentos têm o que é entendido por uma direção. Uma ponta destes filamentos expõe uma molécula de hidroxila do grupo da dioxoribose. Esta ponta é conhecida por *terminal 3*. A outra ponta expõe uma molécula de fosfato. Esta outra ponta é conhecida por *terminal 5*. No DNA, estes dois filamentos estão dispostos no cromossomo de tal forma que, no início da hélice, um dos filamentos começa com as moléculas *terminal 3* e o outro filamento começa com o *terminal 5*. No final da hélice, o filamento que inicia com o *terminal 3* termina com o *terminal 5*, e o outro filamento, que começou com o *terminal 5*, termina com o *terminal 3*. Em cada cromossomo adota-se um dos filamentos como referência e os genes neste filamento são os que possuem orientação positiva. Os genes no outro filamento possuem orientação negativa.

A **reversão** é uma operação onde um ou mais genes tem sua ordenação invertida, bem como sua orientação. A Figura 2.8 mostra um cromossomo com cinco genes, numerados de 1 até 5. Após a operação de reversão do conjunto de genes $\{2, 3, 4\}$, estes genes tem sua ordem bem como sua orientação invertidas. As operações de reversões também podem ser chamadas de inversões.

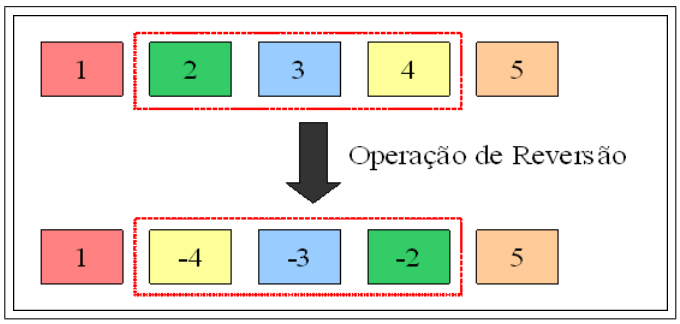

Figura 2.8: Operação de reversão de genes.

A **transposição** é uma operação onde dois grupos contíguos de genes são trocam de

posição entre si. Uma generalização desta operação é chamada de inter-troca de blocos. Na **inter-troca de blocos**, dois conjuntos de genes, contíguos ou não, trocam de posição entre si num mesmo cromossomo. A Figura 2.9 apresenta um evento de transposição onde o grupo de genes $\{2, 3\}$ é inserido depois do grupo de gene $\{4, 5\}$.

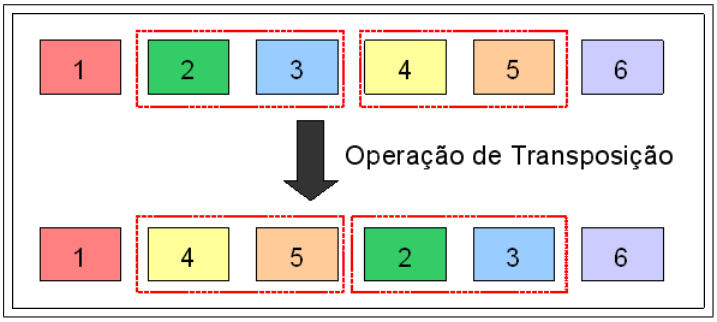

Figura 2.9: Operação de transposição de genes.

Enquanto as operações de reversão e transposição ocorrem num mesmo cromossomo, a **translocação** é uma operação de movimentação de genes entre dois cromossomos de um mesmo organismo. A operação de translocação pode ser do tipo simples e ou do tipo recíproca. Na translocação simples, um ou mais genes são removidos de um cromossomo e inseridos em outro cromossomo. Na translocação recíproca, dois blocos de genes contíguos, pertencentes a cromossomos diferentes de um mesmo organismo, trocam de posição entre si. O primeiro conjunto do primeiro cromossomo assume a posição do segundo conjunto no segundo cromossomo e vice-versa. A Figura 2.10 mostra uma operação de translocação recíproca. Nesta operação, o conjunto de genes $\{1.1, 1.2\}$ do primeiro cromossomo são translocados com o conjunto de genes $\{2.10, 2.11, 2.12\}$ do segundo cromossomo.

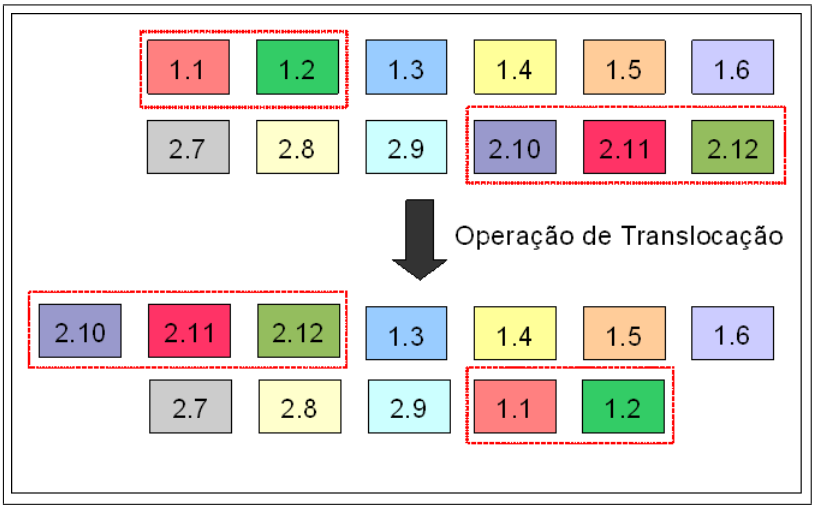

Figura 2.10: Operação de translocação recíproca de genes.

As operações de fissão e fusão são operações opostas. Na **fissão**, um cromossomo é dividido em dois ou mais cromossomos. Na **fusão**, dois ou mais cromossomos são unidos em um único cromossomo. A Figura 2.11 apresenta a operação de fissão de um cromossomo com cinco genes, numerados de 1 até 5. A fissão ocorre entre os genes 3 e 4 resultando em dois cromossomos, um com o conjunto de genes $\{1.1, 1.2, 1.3\}$ e outro com o conjunto de genes $\{2.4, 2.5\}$.

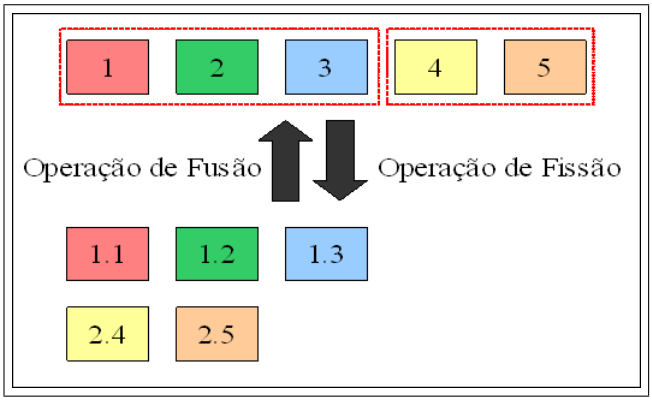

Figura 2.11: Operações de fissão e fusão de genes.

Outro conceito importante relacionado a rearranjo de genomas é o de circularidade do cromossomo. Normalmente, nos organismos eucariontes, os cromossomos são lineares, ou seja, o gene de uma extremidade não é conectado (quimicamente) ao gene da outra extremidade do cromossomo. Já em alguns organismos procariontes, os cromossomos são circulares, isto é, o cromossomo não tem extremidades abertas, e todos os seus genes são

conectados formando uma estrutura circular. A maioria dos problemas de comparação de genomas não é mais difícil de ser resolvido para os genomas circulares do que para os lineares [54]. Todos os organismos analisados neste trabalho possuem genomas circulares. O modelo de rearranjo de genomas utilizado neste trabalho contempla soluções tanto para genomas com cromossomos lineares quanto para genomas com cromossomos circulares.

Diversos pesquisadores estudam problemas de rearranjo de genomas utilizando uma ou mais das operações acima mencionadas. Para informações mais detalhadas sobre os avanços recentes, remetemos o leitor ao trabalho de Feijão e Meidanis [14].

A seguir explicaremos, em linhas gerais, como funciona o algoritmo para rearranjo de genomas *Double-Cut-And-Join* (DCJ). Considere as permutações iniciais de dois genomas A e B, ambos contendo o mesmo conjunto de genes:

$$\textit{Permutação } A = \{a, c, -d, b, e, f, g\} \text{ e } \textit{Permutação } B = \{a, b, c, -d, e, f, g\}$$

A primeira etapa é a construção do conjunto das adjacências de genes das permutações dos genomas A e B. Por exemplo, na *Permutação A* os genes a e c são adjacentes, formando assim uma adjacência. Abaixo, os conjuntos das adjacências da *Permutação A* e da *Permutação B*. As letras t e h representam o início ("tail") e o final ("head") dos genes, e servem para codificar sua orientação. Um gene em orientação positiva terá t antes de h; ao contrário, um gene em orientação negativa terá h antes de t.

$$\textit{Adjacências } A = \{\{NULL, at\}, \{ah, ct\}, \{ch, dh\}, \{dt, bt\}, \{bh, et\}, \{eh, ft\}, \{fh, gt\}, \{gh, NULL\}\}$$

$$\textit{Adjacências } B = \{\{NULL, at\}, \{ah, bt\}, \{bh, ct\}, \{ch, dh\}, \{dt, et\}, \{eh, ft\}, \{fh, gt\}, \{gh, NULL\}\}$$

A segunda etapa é ordenação dos genes, transformando a *Permutação A* na *Permutação B*. Para ordenar os genes, tomamos o conjunto das adjacência de genes $\{p, q\}$ da *Permutação B*, e, para cada uma destas adjacências, deve-se encontrar na *Permutação A* as adjacências que contenham os genes p e q e realizar a operação de DCJ. Para exemplificar a operação, tomemos da *Permutação B* a adjacência $\{ah, bt\}$, onde $p{=}ah$ e $q{=}bt$. Na *Permutação A*, p e q estão nas adjacências $\{ah, ct\}$ e $\{dt, bt\}$. Realizamos um corte na adjacência $\{ah, ct\}$ e um corte na adjacência $\{dt, bt\}$ para depois juntar e formar as novas adjacências $\{ah, bt\}$ e $\{dt, ct\}$. A seguir descrevemos o algoritmo do DCJ em pseudo-código. A implementação utilizada neste trabalho está baseada numa implementação simplificada do algoritmo DCJ feita por Bergeron e colegas [4]. Esta implementação não é multicromossomal.

Algoritmo 1 Calcule a Distância DCJ

```
1.
2.  adjA ← conjunto das adjacências de genes do cromossomo A
3.  adjB ← conjunto das adjacências de genes do cromossomo B
4.  distancia ← 0
5.
6.  for all {p, q} ∈ adjB do
7.     {p, x} ← adjacência de adjA contendo p
8.     {q, y} ← adjacência de adjA contendo q
9.     if x ≠ q then
10.       adjA ← adjA - {p, x} + {p, q}
11.       adjA ← adjA - {q, y} + {x, y}
12.       distancia ← distancia + 1
13.    end if
14. end for
15.
16. return distancia
```

2.3 Árvores Filogenéticas

Árvores filogenéticas são utilizadas para representar as relações evolutivas entre as espécies. Do ponto de vista computacional, estas árvores podem ser representadas por grafos. Cada nó do grafo representa uma unidade taxonômica. Uma unidade taxonômica é uma unidade do sistema de classificação de espécies, podendo ser a própria espécie, ou outros agrupamentos como gêneros e até mesmo reinos.

As arestas representam as relações de herança genética ou parentesco entre as unidades taxonômicas. Se as arestas forem orientadas ou se o grafo ou subgrafo possuir uma raiz, é possível determinar quem é o ancestral e quem é o descendente numa relação. Esta relação de descendência é também denominada de transferência vertical. O padrão de ramificação das arestas é chamado de topologia. Foi observado que os genes evoluem a taxas de mudança constante em função do tempo, se não existirem fatores externos modificadores destas taxas. Isso é chamado de relógio molecular. Estas taxas de evolução entre duas unidades taxonômicas também podem ser representadas neste grafo através da atribuição de pesos às arestas.

Para construir tais árvores a partir de um conjunto de sequências de DNA, existem dois tipos de métodos computacionais: os baseados em características discretas e os baseados

em matrizes de distância.

Características discretas são características significativas dos organismos para os quais se deseja construir uma árvore. Geralmente estas características são morfológicas ou biomoleculares. Por exemplo, uma característica poderia ser o número de dedos ou poderia ser a presença ou ausência de rabo. Cada uma destas características deve possuir um número finito de estados. Sendo assim, para número de dedos poderíamos ter os estados: três dedos, quatro dedos e cinco dedos, e, para presença ou ausência de rabo poderíamos ter os estados: com rabo e sem rabo. Em outro exemplo, cada posição de uma sequência de nucleotídeos de DNA poderia ser uma característica. Neste caso, cada posição possui quatro estados, representados pelos próprios nucleotídeos. A partir disso, a árvore é construída usando-se uma matriz de objetos, neste caso, unidades taxonômicas, versus suas características. A matriz é preenchida com os estados assumidos por estas características em cada uma das unidades taxonômicas em questão. Os métodos de máxima parcimônia tais como Fitch [19] e Sankoff [52] são baseados em características.

Matrizes de distância são matrizes preenchidas com o valor das distâncias resultantes da comparação entre cada dois objetos. Estes objetos podem ser, por exemplo, unidades taxonômicas ou espécies. Estas distâncias são calculadas através de diversos modelos de evolução tais como, por exemplo, os modelos Jukes-Cantor, Kimura, Dayhoff e *Jones-Taylor-Thornton* (JTT). Após o cálculo das distâncias e preenchimento da matriz, algoritmos tais como UPGMA e *Neighbor-Joining* (NJ) são utilizados para inferir as árvores filogenéticas.

A seguir explicaremos, em linhas gerais, como funcionam os algoritmos de UPGMA e *Neighbor-Joining* (NJ).

A sigla UPGMA é do termo em inglês *Unweighted Pair Group Method with Arithmetic mean* e é um método para construção de árvores filogenéticas introduzido por Sokal e Michener [56]. O algoritmo UPGMA produz árvores filogenéticas do tipo ultramétricas. Uma árvore ultramétrica é uma árvore binária, na qual a distância (peso das arestas) da raiz, passando pelos nós internos, até qualquer nó folha é sempre a mesma. Numa árvore ultramétrica, para toda tripla de unidades taxonômicas x, y e z, as distâncias entre (x, y), (x, z) e (y, z) ou são iguais ou duas delas são iguais e a distância restante é menor. Em linhas gerais, o algoritmo UPGMA agrupa sucessivamente os nós mais próximos, isto é, os dois nós com a menor distância. Para cada novo agrupamento é criado um nó interno. Após cada novo agrupamento, é calculada a distância entre o novo nó interno, pai do agrupamento, e seus nós filho. Também são recalculadas todas as distâncias entre este novo nó e todos os outros nós da árvore. A Figura 2.12 contém a árvore gerada a partir

da execução do algoritmo UPGMA para a matriz de distâncias contida na Tabela 2.1.

	A	B	C	D	E
A	0	20	60	100	90
B	20	0	50	90	80
C	60	50	0	40	50
D	100	90	40	0	30
E	90	80	50	30	0

Tabela 2.1: Matriz de distâncias exemplo.

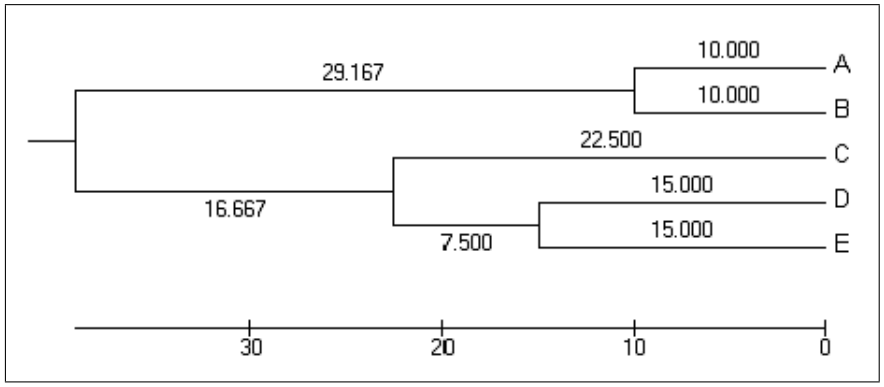

Figura 2.12: Árvore filogenética produzida pelo algoritmo UPGMA.

Neighbor-Joining (NJ) é um método para construção de árvores filogenéticas introduzido por Saitou e Nei [51]. Neste método, a matriz de distâncias utilizada não necessita possuir apenas distâncias ultramétricas e as linhagens da árvore não precisam evoluir na mesma taxa. Produz uma árvore binária e sem raíz. A construção de uma árvore utilizando-se o algoritmo NJ é similar à construção da árvore utilizando o algoritmo UPGMA mudando-se apenas o cálculo das distâncias feito a cada novo agrupamento da árvore. A Figura 2.13 contém a árvore gerada a partir da execução do algoritmo NJ para a matriz de distâncias contida na Tabela 2.1.

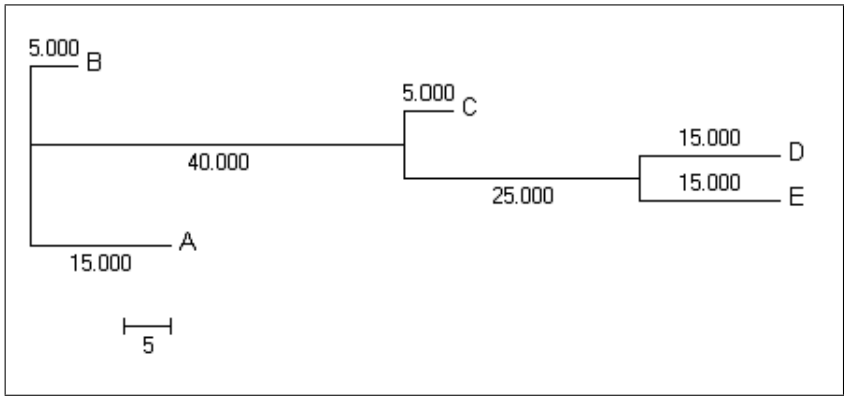

Figura 2.13: Árvore filogenética produzida pelo algoritmo NJ.

Informações mais detalhadas sobre os métodos de Fitch, Sankoff e Dollo, e sobre os algoritmos UPGMA e *Neighbor-Joining* (NJ) podem ser encontradas no livro *Inferring phylogenies* [16].

2.4 Sistema de Taxonomia

Neste capítulo falaremos um pouco sobre o sistema de taxonomia, e, mais especificamente, de como os organismos procariontes são classificados em espécies e quais os desafios enfrentados nesta classificação.

Uma das técnicas usadas na definição de espécies é a hibridização entre DNAs (*DNA-DNA Hybridization* ou DDH). Esta técnica mede o grau de similaridade entre sequências de DNA. O critério de 70% de similaridade é usado para agrupar sequências de DNA num mesmo grupo. Os problemas da DDH são que esta é demorada, é feita por poucos laboratórios e não pode ser utilizada em organismos não cultiváveis. Outro agravante é que uma análise de DDH não pode ser realizada contra uma base de dados de genomas.

Outra técnica utiliza-se da comparação de sequências de rRNA. O gene 16S rRNA geralmente é utilizado nesta técnica por ser um gene bem conservado, isto é, que evolui mais lentamente ou que sofre poucos eventos evolutivos. Entre as vantagens desta técnica destacamos que pode utilizar base de dados de genomas e que pode ser aplicada a organismos não cultiváveis. Porém, esta técnica não é suficiente para determinar uma espécie, fato também observado para a técnica DDH. Rohwer e colegas [50], realizaram um estudo com bactérias encontradas em corais no Panamá e Bermuda, sequenciando mais de 1000 16S rRNA genes. Metade destas sequencias apresentam menos de 93% de similaridade com sequências 16S rRNA previamente publicadas, e, dessa forma, provavelmente representam

novas espécies e gênero de bactérias.

Ainda, De acordo com a comunidade científica, organismos podem ser agrupados numa mesma espécie de procariontes se: *(i)* possuem certas similaridade de fenótipo, *(ii)* possuem sequências de DNA com 70% de similaridade medida através de DDH e *(iii)* suas sequências do gene 16S rRNA são 97% idênticas [26].

MLST (*Multilocus Sequence Typing*) é outra técnica para agrupar organismos procariontes em espécies. Esta técnica classifica organismos em espécies baseando-se na similaridade de um conjunto determinado de genes bem conservados. No entanto, não é possível usar o mesmo conjunto de genes para analisar todos os organismos. Geralmente, para cada família ou gênero é necessário escolher um conjunto de genes mais apropriado. Depois de escolhido o conjunto, estas sequências de genes são então concatenadas, comparadas entre si e usadas para construir árvores filogenéticas através de métodos de distâncias. Esta técnica foi utilizada por Thompson e colegas [63] para realizar a análise da posição taxonômica de seis novas cepas de bactérias da família dos *Vibrionaceae*, isto é, vibriões, obtidas em corais de ilhas da Austrália em 2002. Os vibriões estão relacionados a muitas doenças humanas, entre elas a cólera, causada pela bactéria *Vibrio cholerae*. Os genes utilizados como marcadores foram o 16S rRNA, o *recA* e o *rpoA*. O gene *rpoA* evolui tão lentamente quanto o gene 16S rRNA, é resistente a transferência horizontal e é também um excelente parâmetro para estudos com vibriões. Os autores também se apoiaram em análises de propriedades fenotípicas para determinar a que espécies as cepas pertencem, e finalizam o estudo propondo a criação de duas novas espécies de bactérias.

Cohan [10] num estudo da validade do sistema atual de classificação de espécies, tendo focado-se nos organismos procariontes, nos diz que vem crescendo o consenso entre os cientistas de que conceito de espécie de bactérias não exibe as mesmas propriedades dinâmicas especiais apresentadas pelo conceito de espécies biológico clássico, aquele aplicado aos organismos eucariontes. Em uma espécie de eucariontes, além dos indivíduos possuírem semelhança fenotípica e genotípica, estes indivíduos são capazes de procriar, produzindo descendentes férteis. Já nos indivíduos procariontes, a reprodução nem sempre é sexuada. Décadas de estudos utilizando hibridização de DNA-DNA mostraram que existe grande diversidades entre as bactérias classificados numa mesma espécie. Por fim, Cohan [10] ainda discute três visões contemporâneas da natureza da diversidade biológica entre as bactérias: o conceito biológico de espécies aplicado às bactérias [13], o conceito de ecótipos [9], e o conceito de sem espécie [28]. O autor tem como hipótese que as espécies de procariontes atualmente existentes estão mais para gêneros do que para espécies, e aprofunda o ramo taxonômico subdividindo uma espécie em ecótipos. Ecótipos são populações de or-

ganismos ocupando o mesmo nicho ecológico, nos quais espera-se que a seleção periódica seja uma poderosa forma de coesão e que sua recorrência altere a diversidade genética para próximo de zero. Afirma ainda que a técnica MLST determina os ecótipos ao invés das espécies de procariontes.

Para Gevers e colegas [26] a divisão de espécies dos procariontes são definidas com base em um caráter operacional e centradas em humanos e doenças, sendo importante em diversas áreas (indústrias farmacêutica e alimentícia, por exemplo). Discordam que o critério de 70% de similaridade entre organismos usado em técnicas de DDH seja suficiente para formar uma espécie. Os autores citam casos onde dois organismos têm 99% de similaridade de rRNA, porém possuem apenas 47% de similaridade quando utilizada técnica de DDH. Para resolver tais ambigüidades, estes autores propõem o uso da técnica de comparação de sequência de rRNA para agrupar os organismos em gêneros e famílias, seguido do uso da técnica MLST para agrupá-los em espécies. Propõem que este método seja chamado de *Multilocus Sequence Analysis* (MLSA). Mesmo assim, os resultados dependem da escolha do conjunto de genes e dos valores de similaridade utilizados. Por fim, esta técnica ainda não seria suficiente, pois existem casos onde mesmo agrupados numa mesma espécie, organismos podem ter características fenotípicas diferentes. Isso porque importantes características fenotípicas podem estar em genes que não são muito estáveis ou bem conservados, o que impede que estes genes sejam usados pela técnica MLST.

Capítulo 3

Trabalhos Anteriores

Neste capítulo será apresentada uma breve revisão de trabalhos de pesquisa que estudam métodos para comparação de genomas completos com o objetivo de construir árvores filogenéticas. Fazendo um paralelo com o nosso trabalho, apontamos algumas das soluções utilizadas para determinar homologias bem como tratar as paralogias.

Existem abordagens onde os genomas completos podem ser comparados sem determinação de homologias, apenas realizando o alinhamento de dois ou mais genomas. Por exemplo, Henz e colegas [29] em seu método comparam os nucleotídeos de todo o genoma, utilizando uma variação de algoritmos de similaridade local para determinar o que chamam de segmentos de pares de nucleotídeos com alta similaridade, ou HSP (do inglês, *High-scoring Segment Pairs*), e, com bases nestes HSPs calculam a distância. Darling e colegas [12] utilizam-se dos algoritmos tradicionais de alinhamento, modificados para alinhar regiões conservadas, encontrando assim o que chamam de regiões de colinearidade local entre os genomas, alinhando múltiplos genomas completos simultaneamente.

Outros métodos necessitam determinar as regiões homólogas entre dois genomas, sejam estas regiões formadas pelas próprias proteínas ou genes, pelos domínios estruturais de proteínas, por grupos de genes homólogos, e até por regiões com sobreposição de genes. Em grande parte dos casos, as regiões homólogas são determinadas através da utilização de ferramentas que implementam algum algoritmo de similaridade local, tais como o programa *blast*. Neste caso, os grupos ou famílias de proteínas homólogas são construídos agrupando as proteínas com valores de similaridade maiores que um valor definido. A construção destes grupos ou famílias também depende do universo de proteínas sendo utilizado. Por exemplo, Fitz-Gibbon e colegas [21], utilizam-se apenas do universo de proteínas dos genomas sendo analisados para construir as famílias de proteínas homólogas. Fukami-Kobayashi e colegas [23] se baseiam na organização dos domínios estruturais de

uma proteína, onde duas proteínas com organização de domínios estruturais similares são consideradas homólogas. Araújo e Almeida [3] utilizam-se tanto de genes homólogos quanto de homologias encontradas na forma de grupos contíguos de genes. Jiang [32] utiliza pares de genes com sobreposição da sua sequência de nucleotídeos, existentes nos dois genomas, alegando que estes genes são bons marcadores filogenéticos uma vez que tendem a sofrem menos mutações, pois uma mutação afetaria os dois genes. Note que, a abordagem de Jiang só pode ser aplicada a procariontes, pois genes sobrepostos são muitíssimos mais raros em eucariontes. Jiang também remove os genes desconhecidos, hipotéticos ou putativos não os classificando em homologias.

Alguns pesquisadores optam por construir suas bases de homologias, porém, atualmente, existem algumas bases [35, 47, 65] de homologias disponíveis, e este número vem crescendo. Utilizar-se de bases existentes pode ser uma boa opção, pois estas bases são curadas por especialistas e geralmente contém um grande universo de proteínas. As bases de homologias podem ser criadas manualmente, onde cada proteína tem vários aspectos estudados antes de ser inserida num grupo de homologias. Também podem ser criadas automaticamente, através do uso de ferramentas. Por exemplo, existem bases criadas automaticamente a partir da similaridade de sequências, utilizando a ferramenta *blast* ou similares. Exemplos destas bases são o *Protein Clusters* [35], utilizado neste trabalho, e o SYSTERS [47]. Outras bases, tais como o *Ensembl* [65], utilizam árvores filogenéticas para determinar as homologias.

Homologias podem conter, além dos genes ortógolos, um ou mais casos de genes parálogos, bem como genes xenólogos. Existem métodos de comparação de genomas que não necessitam determinar as ortologias e paralogias, e, dentre estes, estão aqueles que se baseiam apenas na distância do conteúdo de genes, ou seja, somente importa a presença ou a ausência da estrutura homóloga, e não sua localização, orientação ou multiplicidade. Outros métodos de comparação, tais como os que se baseiam na ordem dos genes ou os que se baseiam na análise de congruências entre as árvores de espécie e de genes, necessitam selecionar inequivocamente os genes ortólogos dentro de um grupo de genes homólogos. Para resolver as ortologias e paralogias, também chamadas de duplicações, existem várias abordagens. Almeida e Araújo [3] tentam construir suas famílias de proteínas homólogas de forma a minimizar os possíveis casos de paralogias. Fitzpatrick e colegas [22] resolvem os casos de paralogia com o auxílio da ferramenta YGOB (*Yeast Gene Order Browser*) [7]. Sankoff [53] adota uma estratégia chamada de método do exemplar, onde os genomas comparados são iniciados com o conjunto de proteínas classificadas em famílias sem parálogos, para, a seguir, família a família de homologias, escolher-se a proteína a ser inserida no conjunto de forma a minimizar a distância de rearranjo. Tang e colegas [59] também

procuram manter mais de uma duplicata com uma estratégia chamada casamento máximo. Nesta estratégia, os genes ortólogos de uma determinada família são renomeados em novas famílias, de forma combinada, e é calculada a distância de inversão para cada uma destas combinações.

Para inferir árvores filogenéticas são utilizados vários métodos. Como já foi dito, os genomas completos podem ser alinhados. Uma vez alinhados, basta aplicar um algoritmo de pontuação para gerar a matriz de distâncias. Outro método, chamado super árvores, executa a conciliação de árvores de espécies a partir de árvores de genes [22]. Outros métodos se baseiam no conteúdo dos genomas [3, 21, 22, 23, 32], ou seja, na presença ou ausência de estruturas homólogas para determinar a matriz de distâncias. Temos ainda os métodos que se baseiam na ordenação do genoma, isto é, que se baseiam na ordem das estruturas homólogas. Aqui se incluem os métodos de comparação por distância de *breakpoint* [59] ou por outros eventos de rearranjo [32, 34] tais como reversão e transposição. Além disso, existem os métodos que, ao analisarem um conjunto de genomas, selecionam apenas os ortólogos presentes em todos os genomas sendo analisados [22], enquanto outros, selecionam o máximo número de ortólogos encontrados [59]. Nota-se ainda que alguns autores [3, 32, 34] utilizam-se de combinações destes métodos para calcular as distâncias entre os genomas.

Neste trabalho estamos tratando os eventos de transferência horizontal de genes [5, 36, 38, 39] de maneira simplificada, isto é, como um evento de duplicação ou como um evento de perda de genes. Jiang [32] permite que os genes transferidos horizontalmente possam ser removidos da análise, utilizando a base HGT-DB [24]. Acreditamos que, para analisar eventos de transferência horizontal seria necessária uma análise que realize a inferência de redes filogenéticas [6, 31, 44] e não somente de árvores, bem como, deveria ser utilizado um conjunto de genomas mais abrangente e não somente dentro de uma família taxonômica tal como é feito neste estudo.

Capítulo 4

Comparação de Genomas Completos

A Figura 4.1 apresenta os passos do experimento realizado neste trabalho. Esta figura contém a visão macroscópica e auxiliará o leitor a acompanhar a sequência de atividades realizadas e também a leitura do resto deste documento. Para este experimento foram obtidas informações sobre o conteúdo de onze genomas completos da base do *Entrez* [43] no NCBI. Para cada genoma obtivemos seus cromossomos e suas proteínas, e, para cada proteína, seu nome, sua identificação, suas posições de início e término no cromossomo, seu tamanho e sua orientação. Através da base do *Entrez* também foram obtidos os grupos ortólogos (COG) [60] bem como as sequências de aminoácidos que descrevem cada uma destas proteínas.

O modelo de comparação é dividido em três fases. A primeira fase tratará de classificar os genes de um genoma em famílias universais de genes homólogos. A segunda fase tem por objetivo restringir cada par de genomas a um conjunto comum de genes, dando tratamento adequado aos eventos de duplicações de genes (paralogias) e aos eventos de perda e ganho de genes. Por fim, a distância de rearranjo é calculada para cada par de genomas. A árvore filogenética é inferida com base na matriz das distâncias. Este trabalho é uma evolução do trabalho de Côgo [8]. A seguir, neste capítulo, descreveremos brevemente cada uma destas fases, comparando-as com o trabalho de Côgo.

Na primeira fase, as proteínas homólogas são determinadas utilizando a base de famílias *Protein Clusters* [35] e sua ferramenta *rpsblast*. Este procedimento será detalhado no Capítulo 6. Ao longo deste trabalho, chamaremos as famílias do *Protein Clusters* de famílias PRK, seguindo convenção do próprio *Protein Clusters*. As proteínas que não podem ser classificadas em famílias PRK são desconsideradas. Os cromossomos dos genomas são então comparados dois a dois, através das famílias, que representam as estruturas homólogas entre estes dois cromossomos. Numa comparação entre dois cromossomos, cada família

pode estar presente ou em um ou em outro cromossomo, ou em ambos. Se uma família está em apenas um dos cromossomos e esta família classifica apenas uma proteína do par de cromossomos, temos o que denominamos de família unária. Geralmente, famílias unárias representam os eventos de perda e ganho de genes entre as espécies. Se uma família classifica duas ou mais proteínas de um mesmo cromossomo, esta família apresenta casos de paralogia. Neste trabalho, este tipo de família é chamado de família com duplicações. O último tipo de família são as que classificam exatamente duas proteínas, cada uma oriunda de um cromossomo do par, e estas são aqui denominadas de famílias binárias. No trabalho de Côgo foi construída uma base de famílias utilizando a ferramenta *blast*, tendo como estrutura e conteúdo inicial a base do HAMAP [41]. Com relação a este ponto, resolvemos executar o experimento com duas variantes:

1. Classificando em famílias todas as proteínas sem aplicar qualquer restrição ao *e-value* retornado pela ferramenta *rpsblast*

2. Classificando em famílias apenas as proteínas para as quais a ferramenta *rpsblast* retorna um *e-value* inferior ou igual a 10^{-5}

Além disso, uma variação do experimento foi executada onde, ao invés de classificarmos inicialmente as proteínas em famílias da base *Protein Clusters*, as proteínas foram inicialmente agrupadas por seus grupos ortólogos (COG).

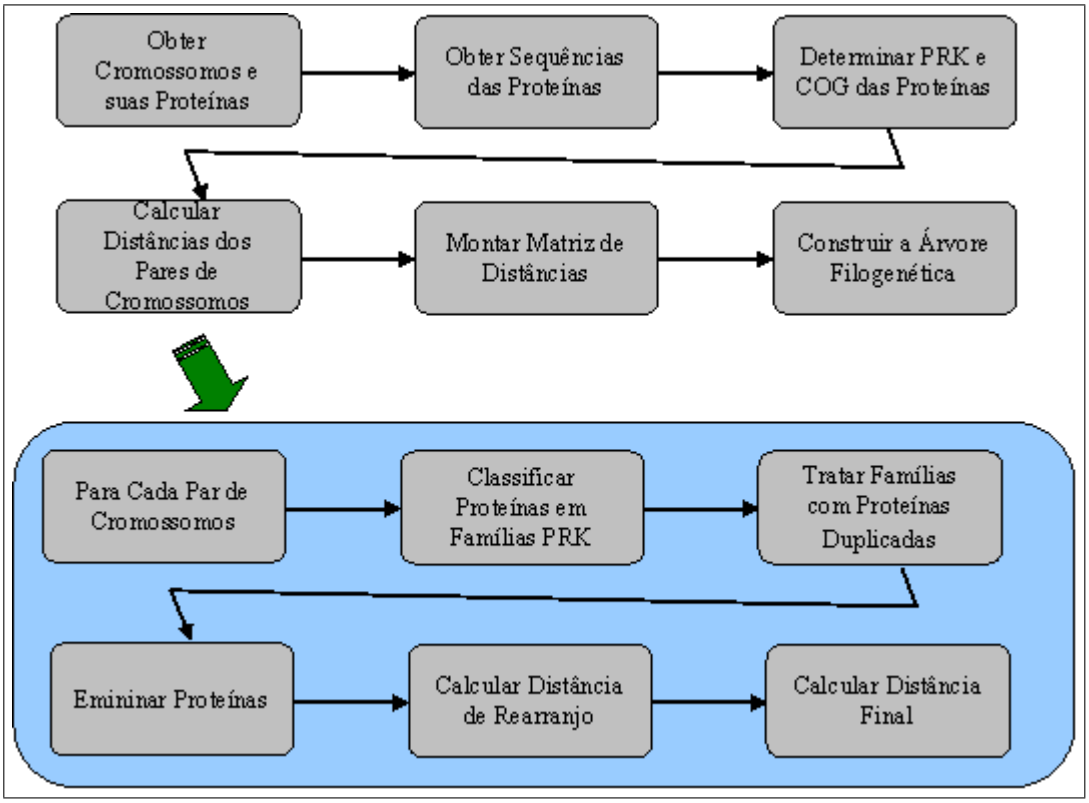

Figura 4.1: Passos realizados no experimento. PRK são as famílias do *Protein Clusters* (NCBI). COG são os grupos do *Clusters of Orthologous Groups* (NCBI).

Para inferir a árvore filogenética dos cromossomos analisados, utilizamos o método *Neighbor-Joining* (NJ). Este método recebe como entrada uma matriz de distâncias contendo todas as distâncias entre cada par de cromossomos. O valor da distância entre dois cromossomos é composto pela distância de eliminação e pela distância de ordenação, também chamada de distância de rearranjo. Para calcular a distância de rearranjo foram utilizados modelos que necessitam que o par de cromossomos possua o mesmo conteúdo, isto é, ambos devem ser reduzidos a um mesmo conjunto de famílias. Com isto, na segunda fase do modelo de comparação, são realizados dois passos. O primeiro passo é tratar as famílias com duplicações com o objetivo de determinar as ortologias e as paralogias correspondentes. Este procedimento será detalhado no Capítulo 7. Para tratar as paralogias encontradas em cada família com duplicações, Côgo utiliza árvores ultramétricas de forma a redistribuir as proteínas em novas subfamílias unárias ou binárias. Este trabalho adiciona aqui uma melhoria, propondo que as famílias com duplicações sejam redistribuídas em novas famílias de acordo com seus grupos ortólogos. Para melhor entender o desempenho de cada método, o experimento foi executado em quatro combinações:

1. Sem tratar duplicações

2. Tratando duplicações apenas utilizando árvores ultramétricas
3. Tratando duplicações apenas utilizando os grupos ortólogos
4. Tratando duplicações com árvores ultramétricas, e a seguir, melhorando este tratamento utilizando os grupos ortólogos

Após o tratamento das famílias com duplicações, as proteínas que foram classificadas em famílias unárias, e as paralogias restantes são eliminadas, compondo a distância de eliminação. Este procedimento será detalhado no Capítulo 8. Côgo eliminou estas proteínas uma a uma, adicionando à distância de eliminação 1 ponto para cada proteína eliminada. Este trabalho implementa aqui uma segunda opção, onde as proteínas serão eliminadas em blocos contíguos, e será adicionado à distância de eliminação 1 ponto para cada bloco de proteínas eliminado. O experimento foi executado com ambos os métodos.

Após a eliminação das proteínas, ambos os cromossomos contém apenas famílias binárias, isto é, famílias que classificam exatamente duas proteínas, cada uma oriunda de um dos cromossomos do par sendo comparado. Dizemos que estes cromossomos foram reduzidos a um mesmo conteúdo, contendo o que denominamos de famílias finais. Estas proteínas finais são colocadas na ordem e orientação em que aparecem no cromossomo, antes de ser calculada a distância de rearranjo. O procedimento utilizado para este cálculo é explicado no Capítulo 9. Côgo utilizou o modelo *Double-Cut-And-Join* (DCJ) para calcular a distância de rearranjo. Este trabalho utiliza o mesmo modelo.

O experimento de comparação foi executado num total de 18 variantes: 16 variantes quando as famílias são inicialmente classificadas pela base *Protein Clusters* e 2 variantes quando famílias são inicialmente classificadas por grupos ortólogos (COG). Note que só foi possível realizar todas estas variações no experimento devido a implementação e automatização total do procedimento de comparação. A automatização facilitou a coleta de dados numéricos que foram utilizados para comparar os resultados obtidos em cada uma das variantes do experimento.

A distância total para cada par de cromossomos é calculada, e a matriz de distâncias é utilizada para inferir a árvore filogenética dos cromossomos sendo comparados. No Capítulo 10 apresentamos a árvore filogenética dos genomas analisados inferida com base no gene 16S rRNA, e algumas árvores filogenéticas inferidas a partir de comparações entre os cromossomos número 1. Com relação ao trabalho de Côgo, existe ainda mais um ponto a ser mencionado. Côgo, em seu trabalho, além de executar as comparações entre os cromossomos número 1 e depois entre os cromossomos número 2, separadamente,

também executou as comparações utilizando ambos os cromossomos número 1 e número 2. A árvore filogenética produzida a partir dos valores de distância da comparação que utiliza ambos os cromossomos números 1 e 2 juntos é muito diferente em topologia da árvore inferida a partir do gene 16S rRNA. Neste trabalho executaremos o experimento comparando separadamente cada um dos cromossomos número 1 e número 2 dos vibriões.

A Figura 4.2 exemplifica os passos realizados para comparar dois cromossomos, até a fase anterior ao cálculo da distância de rearranjo. Começamos com o conjunto de proteínas originais do par de cromossomos *Alfa* e *Beta* a ser comparado. A seguir, estas proteínas são classificadas em famílias PRK. Note que, após esta classificação inicial, temos as famílias *PRK-G* e *PRK-M* com duplicações, e as famílias *PRK-E*, *PRK-X*, *PRK-Z* e *PRK-K* com apenas uma proteína. Note também que, a proteína *263* do cromossomo *Alfa* não pôde ser classificada em nenhuma família PRK. Esta proteína *263*, representada pelo símbolo *(*)* na figura, está excluída da comparação a partir deste momento. O próximo passo é tratar as famílias com duplicações. Neste ponto, a família *PRK-G* é desmembrada nas subfamílias *PRK-G.1*, *PRK-G.2* e *PRK-G.3* e a família *PRK-M* nas subfamílias *PRK-M.1* e *PRK-M.2*. O próximo passo a ser realizado é o de agrupar as proteínas ainda não classificadas em família binárias utilizando os grupos ortólogos. Neste ponto, as proteínas antes classificadas nas famílias unárias *PRK-N* e *PRK-M1* foram agrupadas na família binária *COG-A*. Por fim, as famílias não binárias são eliminadas. Neste passo são eliminadas as famílias (e as subfamílias) *PRK-E*, *PRK-X*, *PRK-Z*, *PRK-K*, *PRK-G.2* e *PRK-M.2*. Finalmente, o par de cromossomos é reduzido a um mesmo conjunto de proteínas finais representado pelas famílias: *PRK-A*, *PRK-B*, *PRK-C*, *PRK-D*, *PRK-F*, *PRK-G.1*, *PRK-G.3*, *PRK-H*, *PRK-I*, *PRK-L* e *COG-A*.

Proteínas Originais															
Cr. α	251	252	253	254	255	256	257	258	259	260	261	262	263	264	
Cr. β	351	352	353	354	355	356	357	358	359	360	361	362	363	364	365
Classificação Inicial em Famílias PRK															
Cr. α	PRK-A	PRK-B	PRK-D	PRK-C	PRK-E	PRK-F	PRK-G	PRK-G	PRK-G	PRK-H	PRK-I	PRK-L	(*)	PRK-N	
Cr. β	PRK-A	PRK-B	PRK-C	PRK-D	PRK-F	PRK-G	PRK-X	PRK-Z	PRK-G	PRK-K	PRK-L	PRK-I	PRK-H	PRK-M	PRK-M
Tratamento de Famílias com Duplicação - Utilizando Árvores Ultramétricas															
Cr. α	PRK-A	PRK-B	PRK-D	PRK-C	PRK-E	PRK-F	PRK-G.1	PRK-G.2	PRK-G.3	PRK-H	PRK-I	PRK-L	(*)	PRK-N	
Cr. β	PRK-A	PRK-B	PRK-C	PRK-D	PRK-F	PRK-G.1	PRK-X	PRK-Z	PRK-G.3	PRK-K	PRK-L	PRK-I	PRK-H	PRK-M.1	PRK-M.2
Tratamento de Famílias com Duplicação - Refinando por Grupos Ortólogos															
Cr. α	PRK-A	PRK-B	PRK-D	PRK-C	PRK-E	PRK-F	PRK-G.1	PRK-G.2	PRK-G.3	PRK-H	PRK-I	PRK-L	(*)	COG-A	
Cr. β	PRK-A	PRK-B	PRK-C	PRK-D	PRK-F	PRK-G.1	PRK-X	PRK-Z	PRK-G.3	PRK-K	PRK-L	PRK-I	PRK-H	COG-A	PRK-M.2
Eliminação de Famílias - Proteínas Finais															
Cr. α	PRK-A	PRK-B	PRK-D	PRK-C	-	PRK-F	PRK-G.1	-	PRK-G.3	PRK-H	PRK-I	PRK-L	(*)	COG-A	
Cr. β	PRK-A	PRK-B	PRK-C	PRK-D	PRK-F	PRK-G.1	-	-	PRK-G.3	-	PRK-L	PRK-I	PRK-H	COG-A	-

Figura 4.2: Exemplo de comparação entre dois cromossomos de acordo com descrição do experimento. O símbolo *(*)* representa as proteínas não classificadas. O símbolo - representa as proteínas classificadas mas eliminadas antes de efetuar-se o rearranjo do genoma.

O material suplementar (Apêndice A) fornece a implementação computacional deste experimento, bem como os resultados numéricos das comparações. De forma geral, nos próximos capítulos, apresentamos os resultados para as comparações entre os cromossomos número 1. Os resultados das comparações entre os cromossomos número 2 podem ser encontrados no material suplementar.

Capítulo 5

Apresentação dos Genomas Analisados

Como entrada para nosso experimento utilizamos dez genomas completos de organismos da família *Vibrionaceae*. Esta família muito estudada pela comunidade científica. São organismos encontrados em água doce ou salgada, e muitos dos organismos que fazem parte desta família são causadores de doenças. Entre estes organismos temos o *Vibrio cholerae* que é o agente causador da cólera. Já os organismos *Vibrio parahaemolyticus* e *Vibrio vulnificus* são causadores de gastroenterites. Muitas bactérias desta família são bioluminescentes e tipicamente vivem em relação de mutualismo com organimos de águas profundas. De acordo com o NCBI, esta família é dividida nos gêneros: *Aliivibrio*, *Allomonas*, *Catenococcus*, *Enterovibrio*, *Ferrania*, *Grimontia*, *Listonella*, *Photobacterium*, *Photococcus*, *Salinivibrio* e *Vibrio*. Os organismos neste trabalho analisados englobam os gêneros: *Aliivibrio*, *Photobacterium* e *Vibrio*. Para que o experimento possuísse na sua entrada um organismo que não fizeste parte da família dos *Vibrionaceae*, o genoma completo do organismo *Escherichia coli* foi utilizado, permitindo assim o posicionamento da raiz da árvore. A Tabela 5.1 descreve brevemente a importância de cada espécie. Mais informações sobre as bactérias da famílias *Vibrionaceae* podem ser encontradas no livro *The Biology of Vibrios* [61].

Organismo	Descrição
Photobacterium profundum	Encontrado no habitat marinho, é um agente catalizador da produção do ácido eicosapentaenóico (EPA), um Omega-3, que pode ser obtido do óleo (azeite) de pescados. EPA vem sendo utilizado no tratamento de doenças tais como esquizofrenia e pesquisas indicam que melhora da resposta nos pacientes em tratamento de quimioterapia. [1]
Vibrio cholerae	É o agente causador da cólera, e pode ser encontrado principalmente nas superfícies de plantas, algas, zooplâncton, crustáceos e insetos. [61]
Vibrio fischeri	É uma bactéria com propriedades bioluminescentes, encontrada em habitat marinho, principalmente em águas com temperaturas subtropicais. Organismos marinhos tais como as lulas Sepiolidas dependem desta bactéria para gerar luz, vivendo em colónias dentro do corpo do hospedeiro, numa relação de mutualismo. [61]
Vibrio harveyi	São encontradas livremente em águas tropicais, é uma patogenia qua ataca a florta intestinal de animais marinhos tais como corais, ostras elagostas. É também responsável por uma patogenia que ataca os camarões cultivados em cativeiro para fins comerciais. Também acredita-se que é responsável pelo fenômeno de água fluorescente em grandes massas de oceano, chamado de "mar de leite". [48]
Vibrio parahaemolyticus	São encontradas livremente em águas marinhas ou em peixes e moluscos. Essa bactérias é causadora de gastroenterite, sendo, na maioria dos casos, uma doença leve ou moderada. A doença é causada quando a bactéria fixa-se no intestino delgado e excreta uma toxina. [45]
Vibrio splendidus	São suspeitos de causar patologias com grande número de mortalidade em animais marinhos, tais como ostras. [27]
Vibrio vulnificus	São encontradas em habitat marinho, e são agentes causadores de infecções, principalmente ocasionadas por ingestão de alimentos crus ou mal cozidos, principalmente ostras, ou contaminação de lesões na pele com estas bactérias. Causam gastroenterites e infecções mais graves geralmente ocorrem em pessoas imunodeprimidas, e em alguns casos podendo levar a morte. [61]
Escherichia coli	Encontrada no lúmen intestinal dos humanos e de outros animais de sangue quente. A presença desta bactéria em água ou nos alimentos indica contaminação por fezes humanas, e a sua quantidade por mililitro de águas, o índice coliforme da água, é uma das principais medidas utilizadas no controle da qualidade da água potável. É agente causador de gastroenterites e infecções do tracto urinário.

Tabela 5.1: Informações sobre as espécies dos genomas analisados.

A Tabela 5.2 apresenta informações sobre o conteúdo dos genomas analisados. Todas estas informações form obtidas do NCBI. A coluna 'Organismo' refere-se ao nome da espécie e da cepa do organismo, enquanto a coluna 'Cr.' refere-se ao número do cromossomo. Os organismos da família *Vibrionaceae* possuem dois cromossomos, enquanto o organismo

Escherichia coli possui apenas um. A coluna 'RefSeq' contém o código de identificação do NCBI de cada cromossomo. A coluna 'Comp (nt)' refere-se ao número de bases de cada cromossomo e a coluna 'No. Prot' refere-se o número de proteínas contidas naquele cromossomo. Por fim, a coluna 'Dt. Criação' informa quando este cromossomo foi incluso na base de dados do NCBI.

Organismo	Cr.	RefSeq	Comp (nt)	No. Prot.	Dt. Criação
Photobacterium profundum SS9	1	NC_006370	4.085.304	3416	30-abr-04
Photobacterium profundum SS9	2	NC_006371	2.237.943	2006	30-abr-04
Vibrio cholerae O1 biovar El Tor str. N16961	1	NC_002505	2.961.149	2742	10-set-04
Vibrio cholerae O1 biovar El Tor str. N16961	2	NC_002506	1.072.315	1093	10-set-04
Vibrio cholerae O395	1	NC_009456	1.108.250	1133	18-mai-07
Vibrio cholerae O395	2	NC_009457	3.024.069	2742	18-mai-07
Vibrio fischeri ES114	1	NC_006840	2.897.536	2586	14-fev-05
Vibrio fischeri ES114	2	NC_006841	1.330.333	1175	14-fev-05
Vibrio fischeri MJ11	1	NC_011184	1.418.848	2590	9-mar-09
Vibrio fischeri MJ11	2	NC_011186	2.905.029	1254	9-mar-09
Vibrio harveyi ATCC BAA-1116	1	NC_009783	3.765.351	3561	6-set-07
Vibrio harveyi ATCC BAA-1116	2	NC_009784	2.204.018	2374	6-set-07
Vibrio parahaemolyticus RIMD 2210633	1	NC_004603	3.288.558	3080	10-mar-03
Vibrio parahaemolyticus RIMD 2210633	2	NC_004605	1.877.212	1752	10-mar-03
Vibrio splendidus LGP32	1	NC_011753	3.299.302	2946	7-mai-09
Vibrio splendidus LGP32	2	NC_011744	1.675.519	1485	7-mai-09
Vibrio vulnificus CMCP6	1	NC_004459	3.281.944	2927	23-dez-02
Vibrio vulnificus CMCP6	2	NC_004460	1.844.853	1557	23-dez-02
Vibrio vulnificus YJ016	1	NC_005139	3.354.505	3259	15-out-03
Vibrio vulnificus YJ016	2	NC_005140	1.857.073	1696	15-out-03
Escherichia coli str. K-12 substr. MG1655	1	NC_000913	4.639.675	4132	15-out-01

Tabela 5.2: Informações sobre o conteúdo dos genomas analisados.

Cabe ressaltar aqui algumas observações interessantes sobre as informações contidas na Tabela 5.2. Primeiramente, nota-se que todos os cromossomos número 1 de cada organismo são maiores em números de bases e proteínas do que seus respectivos cromossomos número 2, exceto no caso do organismo *Vibrio cholerae O395*. Após obtermos um resultado não esperado para este organismo no nosso experimento, executamos novamente o experimento trocando os cromossomos deste organismo. Os resultados obtidos com esta troca corresponderam aos resultados esperados. Tudo leva a crer que houve uma troca dos nomes destes cromossomos no NCBI. Deste ponto em diante, as comparações serão feitas entre o cromossomo número 2 do organismo *Vibrio cholerae O395* e os cromossomos número 1 dos demais organismos, e vice-versa.

Outra informação a ser observada na Tabela 5.2, é o número de proteínas contidas em cada cromossomo. Note que, apesar de serem os organismos de uma mesma família, existem grandes diferenças entre o número de proteínas contidos em cromossomos de mesmo número, isto é, entre os cromossomos de número 1 ou entre os cromossomos de

número 2. Ignorando por um momento o organismo *Vibrio cholerae O395*, e o organismo *Escherichia coli* que é de outra família, note que o cromossomo de número 1 com maior número de proteínas é o *Vibrio harveyi ATCC BAA-1116*, com 3561 proteínas, enquanto o cromossomo de número 1 com menor número de proteínas é o *Vibrio fischeri ES114*, com 2586 proteínas, uma diferença de 975 proteínas. Para calcular a distância de rearranjo entre os dois cromossomos citados, no mínimo e melhor dos casos, onde nenhuma proteína do cromossomo de número 1 do organismo *Vibrio fischeri ES114* é removida, é necessário que se eliminem 975 proteínas do cromossomo de número 1 do organismo *Vibrio harveyi ATCC BAA-1116*, ou seja, uma porção de 27% das proteínas deste cromossomo.

Capítulo 6

Determinação das Famílias de Proteínas

Em nosso procedimento de comparação de genomas completos, desejamos comparar, proteína a proteína, os pares de cromossomos. Existem genes que expressam proteínas como seu produto final, porém, existem genes que expressam como seus produtos finais ácidos nucléicos tais como RNA transportadores ou RNA ribossômicos. Doravante no texto, ao encontrar o termo gene, poderá considerar que estamos nos referindo somente a aqueles genes que produzem proteínas. Para comparar dois cromossomos, sem compará-los nucleotídeo a nucleotídeo (que seria outra metodologia válida), precisamos determinar, para cada proteína de um dos cromossomos, qual é a proteína homóloga no outro cromossomo. Este é um problema difícil, e nossa abordagem consiste em classificar todas as proteínas em famílias de homologia, de forma que, proteínas de uma mesma família serão tidas como correspondentes. É claro que esta abordagem traz desafios: o que ocorre se há duas proteínas que possam corresponder a uma dada? Ou se não houver proteína correspondente? Abordaremos estas questões mais adiante no texto.

O primeiro passo nesta classificação é determinar a base de famílias que será usada. Existem bases de famílias de proteínas públicas tais como o *Protein Clusters* [35], o SYSTERS [47], o PFam [17], o InterPro [30] e o HAMAP [41], para citar algumas. A Tabela 6.1 contém informações destas bases de famílias, as quais fazem uso da similaridade de sequências, entre outros critérios, para definir as famílias.

Base de Famílias	Versão	Número de Famílias
HAMAP	092308	1501
InterPro	18	11128
Pfam	23	10340
Protein Clusters	May_2008	6524
SYSTERS	4	158153

Tabela 6.1: Bases de famílias de proteínas.

Estas bases de famílias diferem principalmente em tamanho, metodologia, definição e cobertura. Por tamanho deve-se entender o número de famílias que esta base contém. A metodologia refere-se a como estas famílias são agrupadas, e se estes agrupamentos são feitos por processo automático ou manual, bem como se estes agrupamentos são ou não verificados por especialistas. A definição da família refere-se aos objetivos da classificação. A cobertura é o número de sequências de proteínas utilizadas para a criação da base. Dadas as diferenças, fica claro que a escolha da base afeta o resultado final do experimento.

Neste experimento, escolhemos utilizar a base de famílias *Protein Clusters*. Esta base é mantida pelo NCBI, e suas famílias são agrupadas por função e similaridade de sequência. Estas famílias são criadas automaticamente, agrupando-se proteínas similares através da utilização da ferramenta *Blast* [2]. A seguir, estas famílias são nomeadas e tem suas funções atribuídas, sendo anotadas manualmente. Esta base contém famílias constituídas por proteínas de organismos procariontes, plasmídeos, fagos e organelas e as sequências de proteínas advém da base de proteínas do NCBI.

Como o leitor poderá notar, o número e o conteúdo das famílias variam entre as bases. Isto se deve a vários fatores, como já mencionamos no parágrafo anterior. Por exemplo, a base de famílias SYSTERS, na versão 4 [47], contém 158.153 famílias. Porém, dentre estas famílias, existem 110.332 famílias geradas por somente a sequência de uma proteína, ou seja, classificam apenas uma proteína. Também, nesta mesma base, somente 35.345 de todas as famílias são perfeitas, isto é, são famílias compostas por sequências que fazem parte apenas de uma família. Dependendo da base, existem casos de proteínas que poderão ser classificadas em mais de uma família.

Além da base de famílias *Protein Clusters*, também obtivemos a base de famílias criada por Côgo [8]. Esta base foi construída tendo como conjunto inicial as famílias da base do HAMAP. A base do HAMAP, na versão *23-Sep-08*, possui somente 1.501 famílias, as quais somente classificam uma pequena parte das proteínas dos genomas completos. Assim, para seu experimento, Côgo desenvolveu uma metodologia de descrição e criação

de famílias, baseada em similaridade de sequências. Partindo inicialmente da base de famílias do HAMAP, novas famílias foram adicionadas automaticamente, totalizando um número de 8.820 famílias. Porém, estas novas famílias foram criadas utilizando-se somente do universo das proteínas dos seis genomas completos analisados, ou seja, o conjunto de famílias não é independente do conjunto de genomas analisados. Usando a base de famílias *Protein Clusters*, que é construída utilizando-se de todas as proteínas da base do NCBI, podemos então garantir a independência entre as famílias e o conjunto de genomas analisados. A Tabela 6.2 mostra o número de famílias de cada base e sua cobertura, isto é, o número de proteínas utilizadas para construí-la.

Base de Famílias	Versão	Número de Famílias	Cobertura (Proteínas Usadas)
Protein Clusters	May_2008	6524	2248112
Côgo	1	8220	27289

Tabela 6.2: Número de famílias e cobertura.

Após escolhermos a base de famílias para nosso experimento, segue-se a classificação propriamente dita. Para classificar uma proteína numa família da base *Protein Clusters* deve ser utilizada a ferramenta *rpsblast* [46]. Para este trabalho, a ferramenta *rpsblast*, versão 2.2.18 para sistema operacional *MS Windows XP*, foi executada utilizando como entrada arquivos *fasta* com as sequências das proteínas de um determinado cromossomo. Também, esta ferramenta foi executada com as opções *default* contra a base *Protein Clusters* versão *May_2008*. Todas as proteínas de todos os cromossomos foram classificadas em famílias PRK em menos de um dia. Tentamos realizar o procedimento de classificação utilizando a base de famílias de Côgo e a ferramenta *ps_scan* [25] do HAMAP. Após 25 horas de execução ininterrupta conseguimos classificar 2.474 proteínas do cromossomo número 1 do organismo *Vibrio vulnificus YJ016*, que possui um total de 3.259 proteínas. Baseado nisso, suponha que, em média, classifica-se uma proteína a cada 36,38 segundos. Sendo assim, todas as 49.490 proteínas dos onze genomas analisados levariam 20,83 dias para serem classificadas pela base de famílias de Côgo. A máquina utilizada neste experimento foi um Pentium 4, 2.8Ghz com 1.5Gb de memória RAM.

A ferramenta *rpsblast* classifica as proteínas em mais de uma família atribuindo pontuações e valores de *e-value* para cada classificação. Este sistema é similar ao utilizado pela ferramenta *Blast*. Dado que uma sequência de tamanho t, ao ser buscada numa base de tamanho T resultou num alinhamento de pontuação p, o *e-value* é o número esperado de vezes que um alinhamento de pontuação igual, ou melhor, aconteça em buscas de sequências de tamanho t em bases de tamanho T.

Tomemos como exemplo a busca feita com a ferramenta *rpsblast* para a proteína transcriptional regulator *MalT*, com identificação *gi* (*GenInfo Identifier* - GI) *15600782*, do organismo *Vibrio cholerae O1*, cromossomo número 2. A Tabela 6.3 mostra a saída da ferramenta *rpsblast* para esta busca. Os termos desta tabela não foram traduzidos, pois são exatamente os termos impressos pela saída da ferramenta *rpsblast*. A coluna '*e-value*' desta tabela fornece o valor esperado que outra sequência, isto é, outra proteína, tenha um alinhamento com a família dada pela coluna 'Subject id' que possua melhor ou igual pontuação. A ferramenta *rpsblast* poderá classificar uma proteína em mais de uma família PRK, porém, selecionaremos apenas a classificação em família com melhor pontuação. Poderá, também, retornar uma lista vazia, ou seja, significando que a entrada não foi nem um pouco similar a qualquer das famílias e portanto não pôde ser classificada.

Subject id	% identity	align. length	mis.	gap open.	q. start	q. end	s. start	s. end	e-value	bit score
prk04841	57,21	902	386	0	20	921	1	902	0	1469
prk10403	49,06	53	27	0	856	908	152	204	1,00E-07	53,7
prk10651	39,62	53	32	0	856	908	154	206	1,00E-07	53,6
prk10100	31,76	85	52	1	839	917	131	215	2,00E-07	52,6
prk09935	36,84	57	36	0	857	913	149	205	4,00E-05	44,9
prk10360	38,89	54	33	0	854	907	134	187	4,00E-04	41,6
prk09390	35,59	59	38	0	854	912	138	196	4,00E-04	41,6
prk13719	21,43	70	51	1	839	908	131	196	5,00E-04	41,3
prk12526	27,54	69	44	2	832	900	136	198	0,16	32,9
prk05084	26,79	56	41	0	88	143	77	132	0,18	32,6
prk09053	36,21	58	29	2	488	537	196	253	0,3	32
prk09958	22,95	61	47	0	854	914	140	200	0,34	31,8
prk07106	45,65	46	21	2	183	224	97	142	0,48	31,4
prk11281	26,32	76	56	0	434	509	157	232	1	30,3
prk03815	40,82	49	20	3	52	98	2	43	1,2	30
prk12370	26,23	61	38	2	393	450	371	427	1,6	29,6
prk00080	41,18	34	11	2	35	68	45	69	1,6	29,7
prk11475	28,17	71	51	0	848	918	125	195	1,7	29,3
prk09483	28	50	36	0	858	907	149	198	1,8	29,3
prk06930	26,79	56	38	2	856	910	114	167	2	29,3
prk09646	26,47	68	47	2	829	894	112	178	2	29,3
prk10840	31,25	48	33	0	857	904	150	197	2,3	29,2
prk12519	41,03	39	20	2	857	894	143	179	2,5	28,9
prk05294	41,94	31	14	2	872	902	491	517	4	28,1
prk13501	28,05	82	45	3	382	449	168	249	4,3	28,1
prk00300	34,38	32	20	1	52	83	8	38	4,6	28,2
prk04019	32,69	52	30	2	447	493	234	285	4,7	28
prk00090	24,64	69	45	2	347	415	51	112	5,5	27,9
prk00440	64,29	14	5	0	55	68	41	54	5,8	27,6
prk08990	66,67	15	5	0	750	764	214	228	6,6	27,7
prk11447	26,56	64	47	0	714	777	675	738	6,9	27,4
prk03381	53,85	26	11	1	361	386	202	226	7,2	27,3
prk01683	36,67	30	19	0	222	251	200	229	8,2	27,4

Tabela 6.3: Exemplo de saída de busca utilizando a ferramenta *rpsblast*.

Quanto menor o *e-value*, mais significativa é a pontuação e a classificação. Analisando-se os valores de *e-value*, como saber se manteremos a classificação de uma proteína em uma determinada família, ou, se devemos escolher deixá-la sem classificação? Para avaliar isto, realizamos a classificação de duas formas distintas: uma delas sem restrição de *e-value* e na outra as proteínas foram classificadas em famílias do *Protein Clusters* somente se o valor de *e-value*, atribuído pela ferramenta *rpsblast*, fosse menor ou igual a 10^{-5}.

A Tabela 6.4 mostra, para os onze genomas analisados, quantas proteínas foram classificadas se não impusermos nenhuma restrição no valor do *e-value*, e quantas proteínas foram classificadas se o valor do *e-value* for menor ou igual a 10^{-5}. Analisando-se as informações da Tabela 6.4, notamos que a base de famílias *Protein Clusters* classifica 98,78% das proteínas dos onze genomas analisados quando não há restrições para o valor do *e-value*. No entanto, somente 58,15% das proteínas são classificadas pela base do *Protein Clusters* quando restringimos esta classificação para somente aquelas classificações cuja ferramenta *rpsblast* atribuiu valores de *e-value* menores ou iguais 10^{-5}.

Base de Famílias	Proteínas Classificadas	Percentual (%)
Protein Clusters	48888	98,78%
Protein Clusters (*e-value* $\leq 10^{-5}$)	28776	58,15%

Tabela 6.4: Número de proteínas classificadas pela base *Protein Clusters*, para os onze genomas analisados, com e sem restrição de *e-value*.

Ora, se somente podemos classificar 58,15%, então, de fato, o resultado do experimento, neste caso, será baseado em somente um pouco mais da metade das proteínas de um cromossomo e não em todas elas. Nos próximos capítulos, o leitor poderá acompanhar o resultado da execução do procedimento de comparação de genomas completos para ambos os casos. Como veremos adiante, faz pouca diferença no resultado final impor ou não este limiar. A imposição do limiar, no entanto, significa que menos proteínas são classificadas em famílias, o que implicará num tempo computacional menor para comparar dois cromossomos.

A Tabela 6.5 apresenta, cromossomo a cromossomo, o número de proteínas classificadas em famílias da base *Protein Clusters* (famílias PRK), com e sem a restrição de valores de *e-value*. A Figura 6.1 apresenta o gráfico em barra dos valores da Tabela 6.5.

Organismo	Cr.	Total de Proteínas	Protein Clusters (sem restrição (de *e-value*)	Protein Clusters (sem restrição de *e-value*) %	Protein Clusters (*e-value* $\leq 10^{-5}$)	Protein Clusters (*e-value* $\leq 10^{-5}$) %
Escherichia coli str. K-12 substr. MG1655	1	4132	4117	99,64%	3146	76,14%
Photobacterium profundum SS9	1	3416	3413	99,91%	2121	62,09%
Photobacterium profundum SS9	2	2006	2004	99,90%	826	41,18%
Vibrio cholerae O1 biovar El Tor str. N16961	1	2742	2681	97,78%	1760	64,19%
Vibrio cholerae O1 biovar El Tor str. N16961	2	1093	1058	96,80%	478	43,73%
Vibrio cholerae O395	2	2742	2711	98,87%	1764	64,33%
Vibrio cholerae O395	1	1133	1107	97,71%	484	42,72%
Vibrio fischeri ES114	1	2586	2570	99,38%	1787	69,10%
Vibrio fischeri ES114	2	1175	1163	98,98%	590	50,21%
Vibrio fischeri MJ11	1	2590	2577	99,50%	1780	68,73%
Vibrio fischeri MJ11	2	1254	1237	98,64%	609	48,56%
Vibrio harveyi ATCC BAA-1116	1	3546	3499	98,67%	1864	52,57%
Vibrio harveyi ATCC BAA-1116	2	2374	2292	96,55%	818	34,46%
Vibrio parahaemolyticus RIMD 2210633	1	3080	3006	97,60%	1881	61,07%
Vibrio parahaemolyticus RIMD 2210633	2	1752	1708	97,49%	874	49,89%
Vibrio splendidus LGP32	1	2945	2926	99,35%	1888	64,11%
Vibrio splendidus LGP32	2	1485	1472	99,12%	730	49,16%
Vibrio vulnificus CMCP6	1	2927	2915	99,59%	1842	62,93%
Vibrio vulnificus CMCP6	2	1557	1554	99,81%	838	53,82%
Vibrio vulnificus YJ016	1	3259	3210	98,50%	1872	57,44%
Vibrio vulnificus YJ016	2	1696	1668	98,35%	824	48,58%

Tabela 6.5: Número de proteínas classificadas em famílias PRK, por cromossomo, com e sem restrição de *e-value*.

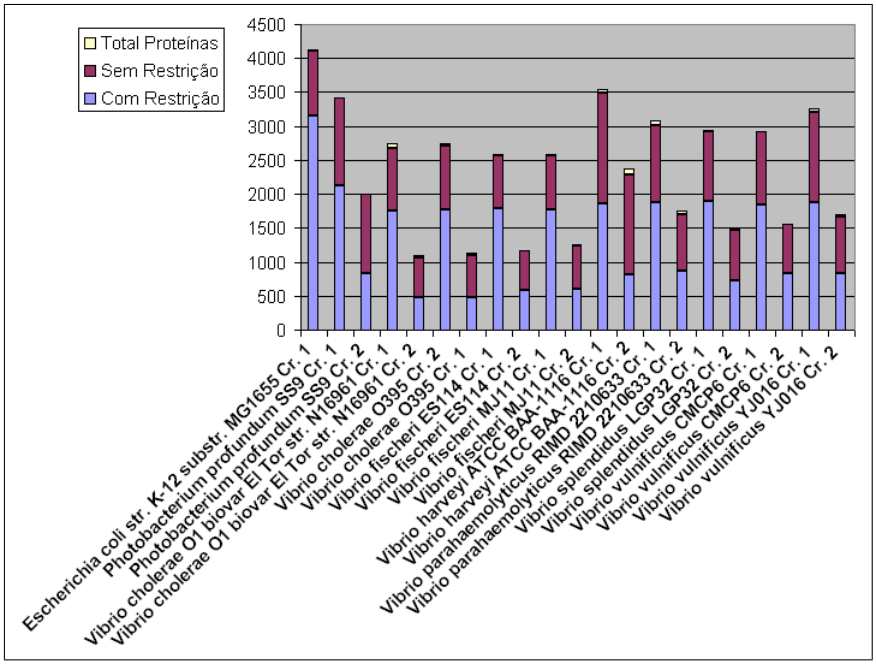

Figura 6.1: Gráfico em barra das proteínas classificadas em famílias PRK, por cromossomo, com e sem restrição de *e-value*.

Capítulo 7

Tratamento de Famílias com Duplicações

Após atribuir as proteínas de cada cromossomo a famílias PRK, nota-se que várias proteínas de um mesmo cromossomo são classificadas em uma mesma família. Para calcular a distância de rearranjo, precisamos que cada cromossomo contenha não mais que um elemento de uma família. E mais ainda, precisamos que estes cromossomos possuam o mesmo número de famílias e que cada uma destas famílias esteja presente nos dois cromossomos, ou seja, em pares.

Neste capítulo, explicaremos o tratamento dado a estas famílias PRK que ocorrem mais de uma vez num mesmo cromossomo, ou seja, que classificam duas ou mais proteínas num mesmo cromossomo. Estas famílias são chamadas de famílias com duplicações. No experimento foram aplicados dois métodos para tratamento de famílias com duplicações, um que reagrupa as proteínas de uma família utilizando árvores ultramétricas e outro que reagrupa as proteínas por grupos ortólogos. Mais adiante neste capítulo explicaremos cada um destes métodos.

A Tabela 7.1 mostra ao leitor, acima da diagonal, o número total de famílias PRK encontradas para cada par de cromossomos número 1 dos genomas analisados. Abaixo da diagonal, tem-se o número de famílias PRK com duplicações para cada par de cromossomos. Estas informações são referentes à classificação feita pela base do *Protein Clusters* na qual não foi aplicada nenhuma restrição de *e-value*. A Tabela 7.2 apresenta conteúdo similar ao da Tabela 7.1 no qual foi aplicada restrição de *e-value* na classificação inicial de famílias. As Tabelas 7.1 e 7.2 contém o número de famílias encontradas para cada par, e, cada família contém uma, duas ou mais proteínas.

	Escherichia coli str. K-12 substr. MG1655	*Photobacterium profundum SS9*	*Vibrio cholerae O1 biovar El Tor str. N16961*	*Vibrio cholerae O395*	*Vibrio fischeri ES114*	*Vibrio fischeri MJ11*	*Vibrio harveyi ATCC BAA-1116*	*Vibrio parahaemolyticus RIMD 2210633*	*Vibrio splendidus LGP32*	*Vibrio vulnificus CMCP6*	*Vibrio vulnificus YJ016*
Escherichia coli str. K-12 substr. MG1655		3579	3474	3448	3433	3436	3597	3537	3516	3495	3588
Photobacterium profundum SS9	1107		2846	2817	2778	2788	2986	2918	2908	2850	2980
Vibrio cholerae O1 biovar El Tor str. N16961	944	850		2187	2606	2595	2788	2710	2688	2625	2767
Vibrio cholerae O395	976	881	489		2583	2579	2757	2687	2661	2600	2741
Vibrio fischeri ES114	934	833	639	671		2161	2776	2698	2661	2633	2770
Vibrio fischeri MJ11	941	837	640	671	500		2794	2701	2659	2617	2768
Vibrio harveyi ATCC BAA-1116	1091	996	802	836	808	805		2766	2852	2768	2888
Vibrio parahaemolyticus RIMD 2210633	1028	928	738	767	733	740	841		2774	2700	2811
Vibrio splendidus LGP32	1004	900	714	740	696	695	846	782		2702	2825
Vibrio vulnificus CMCP6	1022	912	704	732	718	719	841	785	749		2533
Vibrio vulnificus YJ016	1056	958	753	785	767	768	885	829	797	708	

Tabela 7.1: Número de famílias PRK encontradas, para cada par de cromossomos **número 1**, acima da diagonal, versus o número de famílias com duplicações, abaixo da diagonal, **sem restrição** de *e-value*.

	Escherichia coli str. K-12 substr. MG1655	*Photobacterium profundum SS9*	*Vibrio cholerae O1 biovar El Tor str. N16961*	*Vibrio cholerae O395*	*Vibrio fischeri ES114*	*Vibrio fischeri MJ11*	*Vibrio harveyi ATCC BAA-1116*	*Vibrio parahaemolyticus RIMD 2210633*	*Vibrio splendidus LGP32*	*Vibrio vulnificus CMCP6*	*Vibrio vulnificus YJ016*
Escherichia coli str. K-12 substr. MG1655		2791	2748	2746	2757	2759	2763	2778	2765	2750	2761
Photobacterium profundum SS9	524		1852	1854	1851	1844	1861	1872	1871	1835	1866
Vibrio cholerae O1 biovar El Tor str. N16961	410	331		1533	1731	1731	1690	1703	1712	1658	1690
Vibrio cholerae O395	414	334	148		1738	1738	1696	1708	1718	1661	1694
Vibrio fischeri ES114	450	357	242	247		1536	1757	1757	1751	1715	1747
Vibrio fischeri MJ11	448	353	238	243	208		1759	1757	1749	1711	1742
Vibrio harveyi ATCC BAA-1116	456	361	235	240	277	271		1661	1727	1653	1686
Vibrio parahaemolyticus RIMD 2210633	447	360	233	237	276	268	248		1730	1658	1682
Vibrio splendidus LGP32	463	371	245	249	280	279	279	277		1689	1722
Vibrio vulnificus CMCP6	462	373	242	246	291	283	270	260	284		1583
Vibrio vulnificus YJ016	442	360	228	232	273	269	257	249	267	224	

Tabela 7.2: Número de famílias PRK encontradas, para cada par de cromossomos **número 1**, acima da diagonal, versus o número de famílias com duplicações, abaixo da diagonal, **com restrição** de *e-value*.

Na Tabela 7.3, acima da diagonal, é possível visualizar qual o percentual das proteínas que foram classificadas em famílias PRK com duplicações em relação ao total de proteínas

classificadas. Abaixo da diagonal, temos o percentual de famílias PRK com duplicações em relação ao número total de famílias PRK encontradas para cada par de cromossomos. Estas informações são referentes à classificação feita pela base do *Protein Clusters* na qual não foi aplicada nenhuma restrição de *e-value*. Tabela 7.4 apresenta conteúdo similar ao da Tabela 7.3 na qual foi aplicada restrição de *e-value* na classificação inicial de famílias.

	Escherichia coli str. K-12 substr. MG1655	*Photobacterium profundum SS9*	*Vibrio cholerae O1 biovar El Tor str. N16961*	*Vibrio cholerae O395*	*Vibrio fischeri ES114*	*Vibrio fischeri MJ11*	*Vibrio harveyi ATCC BAA-1116*	*Vibrio parahaemolyticus RIMD 2210633*	*Vibrio splendidus LGP32*	*Vibrio vulnificus CMCP6*	*Vibrio vulnificus YJ016*
Escherichia coli str. K-12 substr. MG1655		54,13%	48,06%	49,53%	47,90%	48,31%	54,35%	51,00%	50,06%	51,07%	51,78%
Photobacterium profundum SS9	30,93%		52,30%	53,63%	52,13%	52,52%	58,35%	54,84%	54,09%	55,23%	55,65%
Vibrio cholerae O1 biovar El Tor str. N16961	27,17%	29,87%		41,23%	44,12%	44,56%	52,39%	48,22%	47,53%	47,89%	48,74%
Vibrio cholerae O395	28,31%	31,27%	22,36%		45,81%	46,10%	53,85%	49,59%	48,84%	49,36%	50,28%
Vibrio fischeri ES114	27,21%	29,99%	24,52%	25,98%		41,50%	52,55%	48,17%	46,78%	48,08%	48,88%
Vibrio fischeri MJ11	27,39%	30,02%	24,66%	26,02%	23,14%		52,52%	48,43%	46,96%	48,34%	48,92%
Vibrio harveyi ATCC BAA-1116	30,33%	33,36%	28,77%	30,32%	29,11%	28,81%		54,34%	53,95%	54,63%	55,13%
Vibrio parahaemolyticus RIMD 2210633	29,06%	31,80%	27,23%	28,54%	27,17%	27,40%	30,40%		50,03%	50,92%	51,59%
Vibrio splendidus LGP32	28,56%	30,95%	26,56%	27,81%	26,16%	26,14%	29,66%	28,19%		49,60%	50,24%
Vibrio vulnificus CMCP6	29,24%	32,00%	26,82%	28,15%	27,27%	27,47%	30,38%	29,07%	27,72%		49,73%
Vibrio vulnificus YJ016	29,43%	32,15%	27,21%	28,64%	27,69%	27,75%	30,64%	29,49%	28,21%	27,95%	

Tabela 7.3: Percentual de proteínas classificadas em famílias com duplicações em relação ao total de proteínas classificadas, para cada par de cromossomos **número 1**, acima da diagonal, versus o percentual de famílias com duplicações em relação ao total de famílias encontradas para cada par, abaixo da diagonal, **sem restrição** de *e-value*.

Ainda com relação a Tabela 7.3, note que o percentual de famílias PRK com duplicações em relação ao total de famílias PRK encontradas varia pouco, e, calculando a média dos valores de todos os pares, temos que, em média, 28,41% das famílias apresentam duplicações. Calculando-se a média dos valores acima da diagonal, temos que, em média, 50,27% do total de proteínas classificadas são classificadas em famílias PRK com duplicações. Para a Tabela 7.4, as médias são 15,82% e 30,37%, respectivamente. A Tabela 7.5 sumariza estas informações. Note que, com a restrição do *e-value* menos proteínas são classificadas (vide Tabela 6.4), apenas 58,15% do total de proteínas originais, e destas, apenas 30,37% (aproximadamente 17,66% do total das proteínas originais) estão classificadas em famílias PRK com duplicações.

Pela Tabela 6.4 sabemos que são classificadas 20.112 (40,63%) mais proteínas quando não é aplicada nenhuma restrição ao *e-value* na classificação inicial em famílias, considerando as proteínas de ambos os cromossomos número 1 e 2. Destas 20.112 proteínas, 9.439 são classificadas em famílias com duplicação, isto é, 19,31% do total de proteínas.

	Escherichia coli str. K-12 substr. MG1655	*Photobacterium profundum SS9*	*Vibrio cholerae O1 biovar El Tor str. N16961*	*Vibrio cholerae O395*	*Vibrio fischeri ES114*	*Vibrio fischeri MJ11*	*Vibrio harveyi ATCC BAA-1116*	*Vibrio parahaemolyticus RIMD 2210633*	*Vibrio splendidus LGP32*	*Vibrio vulnificus CMCP6*	*Vibrio vulnificus YJ016*
Escherichia coli str. K-12 substr. MG1655		35,54%	29,29%	29,55%	31,54%	31,40%	31,74%	31,57%	32,32%	32,42%	31,49%
Photobacterium profundum SS9	18,77%		32,85%	33,05%	34,85%	34,63%	34,98%	34,86%	36,02%	36,26%	35,01%
Vibrio cholerae O1 biovar El Tor str. N16961	14,92%	17,87%		22,19%	27,21%	26,92%	27,26%	27,14%	28,34%	28,54%	27,12%
Vibrio cholerae O395	15,08%	18,02%	9,65%		27,43%	27,14%	27,56%	27,41%	28,61%	28,76%	27,34%
Vibrio fischeri ES114	16,32%	19,29%	13,98%	14,21%		26,86%	30,05%	30,10%	30,64%	31,55%	30,12%
Vibrio fischeri MJ11	16,24%	19,14%	13,75%	13,98%	13,54%		29,72%	29,58%	30,51%	31,09%	29,71%
Vibrio harveyi ATCC BAA-1116	16,50%	19,40%	13,91%	14,15%	15,77%	15,41%		29,11%	30,76%	30,81%	29,52%
Vibrio parahaemolyticus RIMD 2210633	16,09%	19,23%	13,68%	13,88%	15,71%	15,25%	14,93%		30,83%	30,33%	29,31%
Vibrio splendidus LGP32	16,75%	19,83%	14,31%	14,49%	15,99%	15,95%	16,16%	16,01%		32,04%	30,48%
Vibrio vulnificus CMCP6	16,80%	20,33%	14,60%	14,81%	16,97%	16,54%	16,33%	15,68%	16,81%		28,86%
Vibrio vulnificus YJ016	16,01%	19,29%	13,49%	13,70%	15,63%	15,44%	15,24%	14,80%	15,51%	14,15%	

Tabela 7.4: Percentual de proteínas classificadas em famílias com duplicações em relação ao total de proteínas classificadas, para cada par de cromossomos **número 1**, acima da diagonal, versus o percentual de famílias com duplicações em relação ao total de famílias encontradas para cada par, abaixo da diagonal, **com restrição** de *e-value*.

Base de Famílias	Média Fam. Dup. (%)	Média Prot. em Fam. Dup. (%)
Protein Clusters	28,41%	50,27%
Protein Clusters ($e\text{-}value \leq 10^{-5}$)	15,82%	30,37%

Tabela 7.5: Médias percentual das famílias com duplicações, **com e sem restrição** de *e-value*, para o cromossomo **número 1**.

Como já foi dito, o procedimento para cálculo da distância de rearranjo requer que cada cromossomo seja composto apenas por famílias binárias. A primeira solução seria, para cada par de cromossomos, eliminar todas as proteínas classificadas em famílias com duplicações, eliminando assim, em média, 50,27% ou 30,37% das proteínas classificadas, considerando a classificação inicial em famílias sem e com restrição de *e-value*. Uma segunda solução seria escolher, para cada família com duplicação, as proteínas homólogas, uma de cada cromossomo do par sendo analisado, que melhor representem aquela família com duplicações, descartando as demais proteínas. A seguir explicaremos os dois métodos implementados neste trabalho. O objetivo é determinar o maior número de paralogias, mantendo assim o maior número possível de proteínas na comparação do par.

7.1 Utilizando Árvores Ultramétricas

Um dos métodos para tratamento de famílias com duplicações implementado neste trabalho foi desenvolvido por Côgo [8] e o denominamos de tratamento de famílias com duplicações utilizando árvores ultramétricas. A Figura 7.1 mostra o fluxograma dos passos que devem ser realizados para cada família com duplicações.

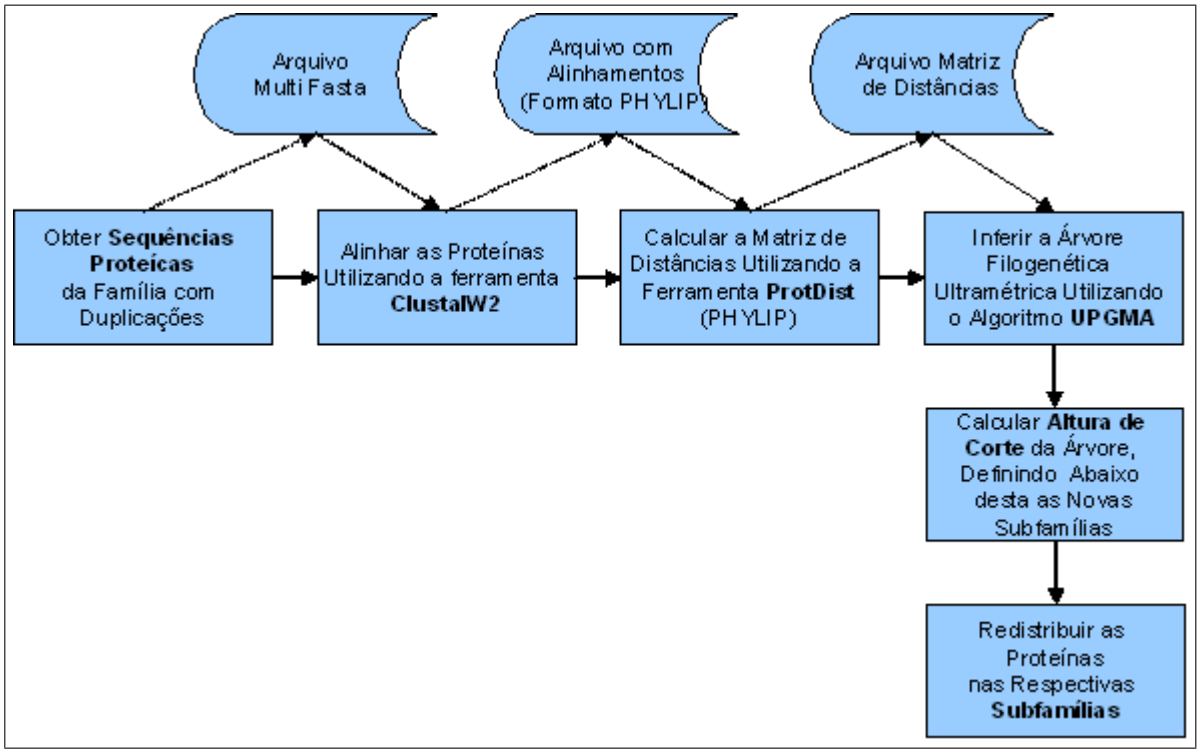

Figura 7.1: Passos para tratamento de famílias com duplicações.

A seguir apresentamos um exemplo da execução deste procedimento. Selecionamos a família *PRK11308* encontrada na comparação do par de cromossomos número 1 dos organismos *Photobacterium profundum SS9* e *Vibrio cholerae O1 biovar El Tor str. N16961*. A Tabela 7.6 contém todas as proteínas classificadas nesta família para a comparação realizada sem restrição de *e-value*.

GI	Organismo	Nome Proteína	Orien.	Posição Início	Comp. Comp.	Sub-Fam. Associada	COG Associado
54308328	*Photobacterium profundum SS9*	putative oligopeptide ABC transporter,ATP-binding protein	+	1258466	329	1	COG4608E
15641108	*Vibrio cholerae O1 biovar El Tor str. N16961*	oligopeptide ABC transporter, ATP-binding protein	+	1163262	336	1	COG4608E
54309610	*Photobacterium profundum SS9*	putative ABC-type antimicrobial peptide transport system, ATPase component	+	2830394	274	2	COG4167V
15641688	*Vibrio cholerae O1 biovar El Tor str. N16961*	peptide ABC transporter, ATP-binding protein	+	1818269	262	2	COG4167V
54307726	*Photobacterium profundum SS9*	putative ABC-type oligopeptide transport system, ATPase component	-	550194	331	3	COG4608E
15640636	*Vibrio cholerae O1 biovar El Tor str. N16961*	peptide ABC transporter, ATP-binding protein	-	651445	331	3	COG4608E
54310050	*Photobacterium profundum SS9*	putative ABC-type oligopeptide transportsystem, ATPase component	-	3363212	340	4	COG4608E
15641801	*Vibrio cholerae O1 biovar El Tor str. N16961*	eha protein	-	1943306	383	5	COG3267U

Tabela 7.6: Proteínas da família com duplicações PRK11308.

Após obter as sequências destas proteínas, o alinhamento é realizado com a ferramenta *ClustalW2* [40], versão 2.0.6 para sistema operacional *MS Windows XP*. Para este trabalho, a ferramenta ClustalW2 foi executada utilizando como entrada arquivos *fasta* com as sequências das proteínas da família. O arquivo com o alinhamento produzido como saída desta ferramenta deve estar no formato *PHYLIP* [15].

A seguir, é calculada a matriz de distâncias, utilizando-se a ferramenta *Protdist* contida no pacote *PHYLIP* [15], versão 3.67 para sistema operacional *MS Windows XP*. Foram utilizadas as opções *default* desta ferramenta, e, para este caso, o modelo de substituição utilizado é o *Jones-Taylor-Thornton* (JTT). A matriz de distâncias produzida como saída da execução da ferramenta *Protdist* é usada para inferir uma árvore filogenética ultramétrica, apresentada na Figura 7.2, através do método UPGMA [56]. Os nós folha desta árvore contêm o número de identificação *gi* (*GenInfo Identifier* - GI) destas proteínas, seguidos pelas letras '*p*' ou '*v*', significando que esta proteína pertence ou ao organismo *Photobacterium profundum SS9* ou ao organismo *Vibrio cholerae O1 biovar El Tor str. N16961*, respectivamente.

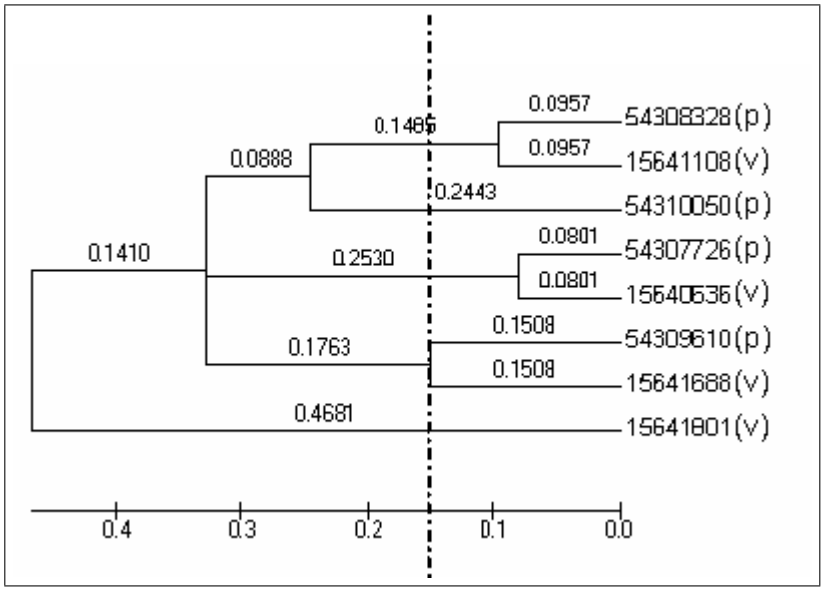

Figura 7.2: Árvore da família PRK11308.

Note como é possível visualizar o agrupamento de subfamílias. Por exemplo, é possível visualizar que a proteína com *gi 54310050* do organismo *Photobacterium profundum SS9* compõe uma subfamília com uma única proteína, sem sua correspondente no organismo *Vibrio cholerae O1 biovar El Tor str. N16961*. A linha tracejada da Figura 7.2 representa a altura de corte da árvore. A altura de corte representa o ponto onde a árvore deverá ser cortada de forma que os novos grupos não contenham proteínas do mesmo organismo. Escolhe-se o corte mais distante das folhas com esta propriedade.

O passo final do procedimento é o de redistribuir as proteínas da família *PRK11308*, nas suas respectivas novas subfamílias *(PRK11308.1, PRK11308.2, PRK11308.3, PRK11308.4, PRK11308.5)*. Três destas famílias possuem proteínas oriundas de cromossomos diferentes, e duas destas famílias possuem uma única proteína. Após este procedimento, cada nova subfamília ou é uma família unária, contendo somente uma proteína, ou é uma família binária, contendo duas proteínas oriundas de cromossomos diferentes.

Note por fim que as novas subfamílias PRK da família *PRK11308* agrupam proteínas que estão associadas a um mesmo grupo ortólogo (COG). Esta relação entre as subfamílias PRK e os COG foi observada em grande número de casos conferidos manualmente.

Na Tabela 7.7 são apresentados alguns resultados numéricos obtidos por este método para o conjunto de cromossomos número 1 analisados. A primeira coluna, 'Média de Famílias PRK Encontradas - Início', contém a média das famílias classificadas inicialmente por

par de cromossomos. As colunas 'Média de Famílias PRK Binárias - Início' e 'Média de Famílias PRK com Duplicações - Início' informam os números médios de famílias PRK binárias e com duplicações antes do tratamento. A seguir, a coluna 'Média Famílias PRK Binárias - Após Tratamento', contém o número médio de famílias PRK binárias, após o tratamento de famílias com duplicações utilizando árvores ultramétricas. A última coluna, 'Percentual de Aumento de Famílias PRK Binárias - Após Tratamento', mostra o aumento percentual das famílias PRK binárias após o tratamento.

Ainda de acordo com a Tabela 7.7, este procedimento redistribuiu as famílias aumentando em 77,37% o número de famílias binárias, quando não é aplicada a restrição de *e-value*. Quando é aplicação a restrição do *e-value*, este aumento é de 30,80%. Portanto, com o método descrito neste capítulo, conseguimos salvar um grande número de homologias que de outra forma seriam consideradas perda ou ganho de genes.

Base de Famílias	Média Famílias PRK Encontradas - Início	Média Famílias PRK Binárias - Início	Média Famílias PRK com Duplicações - Início	Média Famílias PRK Binárias - Após Tratamento	Percentual de Aumento Famílias PRK Binárias - Após Tratamento
Protein Clusters	2861,91	980,75	816,2	1737,22	77,37%
Protein Clusters (*e-value* $\leq 10^{-5}$)	1918,13	1129,18	305,65	1476,84	30,80%

Tabela 7.7: Comparativo do número de famílias PRK binárias antes e depois tratamento de duplicações, utilizando árvores ultramétricas, para o cromossomo **número 1**.

7.2 Utilizando Grupos Ortólogos

Nesta seção explicaremos como é feito o tratamento de famílias com duplicações utilizando grupos ortólogos (COG). Este método seleciona todas as proteínas que foram classificadas em famílias PRK unárias ou em famílias PRK com duplicações e reclassifica estas proteínas por seus grupos ortólogos. Neste ponto, lembramos ao leitor que nem todas as proteínas possuem informação de grupo ortólogos. As proteínas classificadas em famílias PRK binárias permanecem inalteradas. Esta reclassificação apresentará os mesmos três tipos de grupos: unários, binários e com duplicações. As proteínas classificadas em grupos ortólogos binários serão mantidas. Todas as demais proteínas, classificadas em grupos ortólogos unários ou com duplicações, são descartadas.

A seguir apresentaremos os resultados obtidos pela execução de ambos os métodos para tratamento de famílias com duplicações. Assim, primeiramente, as famílias PRK com duplicações são subdivididas em novas subfamílias utilizando árvores ultramétricas. A seguir, as proteínas que não estiverem classificadas em famílias (ou subfamílias) PRK binárias serão reagrupadas por grupos ortólogos (COG). A Tabela 7.8 apresenta, acima

da diagonal, para cada par de cromossomos número 1, o número total de famílias binárias após a reclassificação utilizando grupos ortólogos. Abaixo da diagonal, tem-se o número de famílias binárias após o tratamento utilizando árvores ultramétricas. Estas informações são referentes à classificação feita pela base do *Protein Clusters* na qual não foi aplicada nenhuma restrição de *e-value*. A Tabela 7.9 apresenta conteúdo similar ao da Tabela 7.8 na qual foi aplicada restrição de *e-value* na classificação inicial de famílias. Note que, neste caso, as proteínas não classificadas inicialmente em famílias PRK, também não serão incluídas na execução do procedimento descrito nesta seção, mesmo se estas proteínas possuírem informações de grupo ortólogos.

	Escherichia coli str. K-12 substr. MG1655	*Photobacterium profundum SS9*	*Vibrio cholerae O1 biovar El Tor str. N16961*	*Vibrio cholerae O395*	*Vibrio fischeri ES114*	*Vibrio fischeri MJ11*	*Vibrio harveyi ATCC BAA-1116*	*Vibrio parahaemolyticus RIMD 2210633*	*Vibrio splendidus LGP32*	*Vibrio vulnificus CMCP6*	*Vibrio vulnificus YJ016*
Escherichia coli str. K-12 substr. MG1655		1821	1692	1671	1640	1655	1798	1781	1749	1726	1780
Photobacterium profundum SS9	1707		1793	1777	1794	1789	1908	1910	1883	1869	1892
Vibrio cholerae O1 biovar El Tor str. N16961	1592	1662		2432	1683	1693	1886	1882	1836	1868	1912
Vibrio cholerae O395	1565	1648	2416		1675	1708	1901	1867	1852	1862	1895
Vibrio fischeri ES114	1540	1660	1569	1560		2298	1781	1788	1756	1751	1763
Vibrio fischeri MJ11	1543	1658	1570	1546	2248		1791	1788	1792	1757	1770
Vibrio harveyi ATCC BAA-1116	1667	1770	1740	1725	1653	1632		2181	1989	2012	2045
Vibrio parahaemolyticus RIMD 2210633	1664	1766	1730	1714	1644	1643	2058		1953	2047	2082
Vibrio splendidus LGP32	1643	1752	1698	1689	1642	1644	1810	1808		1951	1975
Vibrio vulnificus CMCP6	1615	1737	1739	1725	1624	1624	1862	1878	1812		2438
Vibrio vulnificus YJ016	1678	1769	1778	1758	1644	1652	1907	1938	1839	2392	

Tabela 7.8: Número de famílias binárias obtido após a reclassificação utilizando grupos COG, para cada par de cromossomos **número 1**, acima da diagonal, versus o número de famílias binárias obtido após o tratamento de famílias PRK com duplicações utilizando árvores ultramétricas, abaixo da diagonal, **sem restrição** de *e-value*.

	Escherichia coli str. K-12 substr. MG1655	*Photobacterium profundum SS9*	*Vibrio cholerae O1 biovar El Tor str. N16961*	*Vibrio cholerae O395*	*Vibrio fischeri ES114*	*Vibrio fischeri MJ11*	*Vibrio harveyi ATCC BAA-1116*	*Vibrio parahaemolyticus RIMD 2210633*	*Vibrio splendidus LGP32*	*Vibrio vulnificus CMCP6*	*Vibrio vulnificus YJ016*
Escherichia coli str. K-12 substr. MG1655		1523	1420	1425	1401	1404	1463	1465	1445	1435	1464
Photobacterium profundum SS9	1459		1481	1479	1486	1493	1539	1547	1528	1521	1530
Vibrio cholerae O1 biovar El Tor str. N16961	1366	1450		1737	1417	1410	1533	1535	1506	1543	1560
Vibrio cholerae O395	1366	1449	1733		1410	1411	1532	1535	1500	1540	1552
Vibrio fischeri ES114	1336	1447	1385	1380		1724	1453	1479	1453	1445	1447
Vibrio fischeri MJ11	1336	1455	1377	1373	1714		1462	1483	1459	1457	1455
Vibrio harveyi ATCC BAA-1116	1390	1495	1498	1495	1421	1421		1694	1560	1605	1615
Vibrio parahaemolyticus RIMD 2210633	1390	1506	1502	1501	1436	1436	1668		1580	1639	1647
Vibrio splendidus LGP32	1382	1494	1482	1477	1421	1423	1533	1549		1564	1566
Vibrio vulnificus CMCP6	1363	1484	1508	1508	1410	1418	1579	1598	1536		1764
Vibrio vulnificus YJ016	1398	1494	1523	1517	1414	1425	1592	1616	1539	1758	

Tabela 7.9: Número de famílias binárias obtido após a reclassificação utilizando grupos COG, para cada par de cromossomos **número 1**, acima da diagonal, versus o número de famílias binárias obtido após o tratamento de famílias PRK com duplicações utilizando árvores ultramétricas, abaixo da diagonal, **com restrição** de *e-value*.

A Tabela 7.10 apresenta os resultados numéricos obtidos pelo método de tratamento de famílias com duplicações utilizando grupos ortólogos, para o conjunto de cromossomos número 1. Por esta tabela é possível verificar uma pequena melhoria no número de famílias binárias após este tratamento. As colunas 'Início Atribuindo-se Famílias PRK', 'Após Tratamento por Árvores Ultramétricas' e 'Após Tratamento por Grupos Ortólogos' contém a média das famílias binárias existentes, por par de cromossomos, após cada uma destas fases. A última coluna, 'Percentual de Aumento de Famílias Binárias - Após Tratamento por Grupos Ortólogos', mostra o aumento percentual das famílias binárias após a reclassificação por grupos ortólogos (COG).

De acordo com a Tabela 7.10, e comparando com os resultados apresentados na Tabela 7.7, o procedimento que reagrupa as proteínas por seus grupos ortólogos produz um aumento de, em média, 7,49% no número de famílias binárias, para o experimento onde não foi aplicada a restrição de *e-value*. Quando é aplicada a restrição do *e-value*, este aumento é de, em média, 2,64%.

Base de Famílias	Início Atribuíndo-se Famílias PRK	Após Tratamento por Árvores Ultramétricas	Após Tratamento por Grupos Ortólogos	Percentual de Aumento de Famílias Binárias - Após Tratamento por Grupos Ortólogos
Protein Clusters	980,75	1737,22	1865,24	7,49%
Protein Clusters (*e-value* $\leq 10^{-5}$)	1129,18	1476,84	1514,93	2,65%

Tabela 7.10: Evolução do número de famílias binárias após o tratamento de famílias PRK com duplicações e após o agrupamento de famílias PRK unárias por grupos COG, para o cromossomo **número 1**.

Por fim, executamos uma variante do experimento. Nesta variante, as proteínas são inicialmente classificadas em famílias PRK, porém, não é realizado o tratamento das famílias PRK com duplicações utilizando árvores ultramétricas. No lugar deste tratamento, resolvem-se as paralogias apenas através da reagrupamento das proteínas por grupos ortólogos (COG), conforme descrito nesta seção. A Tabela 7.11 apresenta os resultados do experimento acima mencionado, e pode ser comparada com a Tabela 7.7. Na Tabela 7.7, após o tratamento das duplicações, temos 1.737 e 1.476 famílias binárias por par, em média, quando as proteínas são classificadas sem e com restrição de *e-value*, respectivamente (Veja coluna 'Média Famílias PRK Binárias - Após Tratamento'). Voltando à Tabela 7.11, após as proteínas serem classificadas inicialmente em famílias PRK, e realizando diretamente a resolução das paralogias apenas utilizando os grupos ortólogos, temos, 1.380 e 1.209 famílias binárias, quando as proteínas são classificadas sem e com restrição de *e-value*, respectivamente (Veja coluna 'Média Famílias PRK Binárias - Após Tratamento'). Sem realizar o cálculo da distância de rearranjo de proteínas e inferir as árvores filogenéticas, ainda não é possível afirmar qual dos dois métodos para tratamento de famílias com duplicações é melhor e identifica, corretamente, mais homologias, porém, os resultados obtidos por ambos os métodos se mostram promissores. Na verdade, este resultado era esperado, pois os grupos ortólogos definem uma classificação de mais alto nível — basta verificar o número de grupos COG existentes e o número de famílias PRK existentes. Inferimos a árvore filogenética produzida a partir das distâncias obtidas pela execução desta variante do experimento e esta árvore apresenta topologia idêntica à árvore inferida quando ambos os métodos de tratamento de família com duplicações são aplicados.

A seguir descrevemos todas as colunas da Tabela 7.11. A primeira coluna, 'Média de Famílias PRK Encontradas - Início', contém a média das famílias classificadas inicialmente por par de cromossomos, antes do tratamento das famílias com duplicações. As colunas 'Média de Famílias PRK Binárias - Início' e 'Média de Famílias PRK com Duplicações - Início' informam os números médios de famílias PRK binárias e com duplicações antes do tratamento. A seguir, a coluna 'Média Famílias PRK e COG Binárias - Após Trata-

mento', contém o número médio de famílias PRK e COG binárias após o procedimento de reclassificação das proteínas por grupos ortólogos. A última coluna, 'Percentual de Aumento de Famílias Binárias - Após Tratamento', mostra o aumento percentual das famílias binárias após o tratamento. Estes valores foram obtidos a partir do conjunto de cromossomos número 1.

Base de Famílias	Média Famílias PRK Encontradas - Início	Média Famílias PRK Binárias - Início	Média Famílias PRK com Duplicações - Início	Média Famílias PRK e COGs Binárias - Após Tratamento	Percentual de Aumento Famílias Binárias - Após Tratamento
Protein Clusters	2861,91	980,75	816,20	1380,89	41,10%
Protein Clusters (*e-value* $\leq 10^{-5}$)	1918,13	1129,18	305,65	1209,71	7,17%

Tabela 7.11: Comparativo do número de famílias binárias, realizado o tratamento das famílias PRK com duplicações apenas utilizando grupos ortólogos, para o cromossomo **número 1**.

Finalmente ressaltamos que este método possui tempo de execução menor do que o método que utiliza árvores ultramétricas. A Tabela 7.12 mostra os resultados da comparação entre os cromossomos número 1 dos organismos *Photobacterium profundum SS9* e *Vibrio cholerae O1 biovar El Tor str. N16961*, quando não é aplicada a restrição de *e-value* na classificação inicial de famílias PRK. Logo após a classificação inicial, temos 911 famílias binárias. Após o tratamento de duplicações utilizando árvores ultramétricas, que leva um tempo de 632 segundos para ser executado, o número final de famílias binárias é 1.662. Após o tratamento de duplicações utilizando grupos ortólogos, que leva apenas um segundo para ser executado, o número final de famílias binárias é 1.290. Na implementação computacional - encontrada no material suplementar - do método que utiliza árvores ultramétricas, os programas *ClustalW2* e *Protdist* são executados externamente ao programa do experimento, o que é um fator que aumenta o tempo de execução. Além disso, é necessário escrever em disco o arquivo com as sequências das proteínas para executar o programa *ClustalW2*. Então, o programa *ClustalW2* escreve, também em disco, um arquivo com os alinhamentos, que é entrada para a execução o programa *Protdist*. Por fim, o programa *Protdist* escreve em disco um arquivo com a matriz de distâncias que a seguir é lido pelo programa. Acredita-se que uma melhoria no tempo de execução pode ser alcançada, através de implementações internas dos programas *ClustalW2* e *Protdist* e sem a utilização do disco. Mesmo assim, o método que utiliza árvores ultramétricas provavelmente não seria mais rápido que o método que utiliza grupos ortólogos, pois precisa executar o alinhamento das proteínas de cada família com duplicação, o que é uma operação que consome muito tempo.

Base de Famílias - Início	Tratamento por Árvores Ultramétricas	Tratamento por Grupos Ortólogos	Núm. Famílias Binárias Antes do Tratamento	Núm. Famílias Binárias Após o Tratamento	Tempo de Execução (Segundos)
PRK	Sim	Não	911	1662	632
PRK	Não	Sim	911	1290	1

Tabela 7.12: Comparação entre os tempo de execução dos métodos para tratamento de duplicações.

Capítulo 8

Eliminação de Proteínas

Para se calcular a distância de rearranjo entre os cromossomos, é necessário reduzí-los a um mesmo conjunto de famílias. Neste ponto, são eliminadas as proteínas que não tem homólogo no outro cromossomo. Esta operação representa os eventos de perda ou ganho de genes durante a evolução destes genomas. Podemos considerar que serão eliminadas da comparação todas as proteínas que não estão classificadas nem em famílias PRK binárias nem em grupos COG binários. No experimento foram aplicados dois métodos para eliminação das proteínas, um que elimina as proteínas uma a uma e outro que elimina blocos de proteínas contiguas (vizinhas). No método em que as proteínas são eliminadas uma a uma, cada proteína eliminada é computada ao valor da distância de eliminação. No método em que as proteínas contíguas são eliminadas aos blocos, cada bloco eliminado é computado ao valor da distância de eliminação.

A Tabela 8.1 apresenta, acima da diagonal, o número de proteínas eliminadas, uma a uma, de cada par de cromossomos número 1. Abaixo da diagonal é apresentado o percentual de proteínas que foi eliminada em relação ao total de proteínas classificadas inicialmente em famílias PRK. Note que estes percentuais de eliminação variam entre 31,63% e 45,72%, aproximadamente, para organismos de espécies diferentes da família *Vibrionaceae*. Estes números aumentam, variando de 50,22% a 52,78% quando se compara um organismo da família *Vibrionaceae* com o organismo *Escherichia coli*. Este percentual de eliminação é menor nas comparações entre os pares de cromossomos de cepas da mesma espécie, tais como *Vibrio cholerae O395* e *Vibrio cholerae O1 biovar El Tor str. N16961*, ou como *Vibrio vulnificus CMCP6* e *Vibrio vulnificus YJ016*. Os dados apresentados na Tabela 8.1 foram computados realizando-se o experimento sem restrição de *e-value*. A Tabela 8.2 apresenta conteúdo similar ao da Tabela 8.1 quando o experimento é realizado aplicando-se a restrição de *e-value* na classificação inicial das proteínas em famílias. Estas tabelas apresentam resultados obtidos pela variante do experimento que executa os dois métodos

para tratamento de duplicações: primeiro tratando duplicações com árvores ultramétricas, e a seguir, melhorando este tratamento utilizando os grupos ortólogos.

	Escherichia coli str. K-12 substr. MG1655	*Photobacterium profundum SS9*	*Vibrio cholerae O1 biovar El Tor str. N16961*	*Vibrio cholerae O395*	*Vibrio fischeri ES114*	*Vibrio fischeri MJ11*	*Vibrio harveyi ATCC BAA-1116*	*Vibrio parahaemolyticus RIMD 2210633*	*Vibrio splendidus LGP32*	*Vibrio vulnificus CMCP6*	*Vibrio vulnificus YJ016*
Escherichia coli str. K-12 substr. MG1655		3888	3414	3486	3407	3384	4020	3561	3545	3580	3767
Photobacterium profundum SS9	51,63%		2508	2570	2395	2412	3096	2599	2573	2590	2839
Vibrio cholerae O1 biovar El Tor str. N16961	50,22%	41,16%		528	1885	1872	2408	1923	1935	1860	2067
Vibrio cholerae O395	51,05%	41,97%	9,79%		1931	1872	2408	1983	1933	1902	2131
Vibrio fischeri ES114	50,95%	40,03%	35,90%	36,57%		551	2507	2000	1984	1983	2254
Vibrio fischeri MJ11	50,55%	40,27%	35,60%	35,40%	10,71%		2494	2007	1919	1978	2247
Vibrio harveyi ATCC BAA-1116	52,78%	44,79%	38,96%	38,78%	41,31%	41,05%		2143	2447	2390	2619
Vibrio parahaemolyticus RIMD 2210633	49,99%	40,49%	33,81%	34,69%	35,87%	35,95%	32,94%		2026	1827	2052
Vibrio splendidus LGP32	50,33%	40,59%	34,51%	34,29%	36,10%	34,87%	38,09%	34,15%		1939	2186
Vibrio vulnificus CMCP6	50,91%	40,93%	33,24%	33,81%	36,15%	36,02%	37,26%	30,86%	33,20%		1249
Vibrio vulnificus YJ016	51,41%	42,87%	35,09%	35,99%	39,00%	38,83%	39,04%	33,01%	35,63%	20,39%	

Tabela 8.1: Número de proteínas **eliminadas uma a uma**, para o par de cromossomos **número 1**, acima da diagonal, versus o percentual de proteínas eliminadas em relação ao total de proteínas classificadas em famílias, abaixo da diagonal, **sem restrição** de *e-value*.

	Escherichia coli str. K-12 substr. MG1655	*Photobacterium profundum SS9*	*Vibrio cholerae O1 biovar El Tor str. N16961*	*Vibrio cholerae O395*	*Vibrio fischeri ES114*	*Vibrio fischeri MJ11*	*Vibrio harveyi ATCC BAA-1116*	*Vibrio parahaemolyticus RIMD 2210633*	*Vibrio splendidus LGP32*	*Vibrio vulnificus CMCP6*	*Vibrio vulnificus YJ016*
Escherichia coli str. K-12 substr. MG1655		2221	2066	2060	2131	2118	2084	2097	2144	2118	2090
Photobacterium profundum SS9	42,17%		919	927	936	915	907	908	953	921	933
Vibrio cholerae O1 biovar El Tor str. N16961	42,11%	23,68%		50	713	720	558	571	636	516	512
Vibrio cholerae O395	41,96%	23,86%	1,42%		731	722	564	575	652	526	532
Vibrio fischeri ES114	43,20%	23,95%	20,10%	20,59%		119	745	710	769	739	765
Vibrio fischeri MJ11	43,00%	23,46%	20,34%	20,37%	3,34%		720	695	750	708	742
Vibrio harveyi ATCC BAA-1116	41,60%	22,76%	15,40%	15,55%	20,41%	19,76%		357	632	496	506
Vibrio parahaemolyticus RIMD 2210633	41,71%	22,69%	15,68%	15,78%	19,36%	18,98%	9,53%		609	445	459
Vibrio splendidus LGP32	42,59%	23,77%	17,43%	17,85%	20,93%	20,45%	16,84%	16,16%		602	628
Vibrio vulnificus CMCP6	42,46%	23,24%	14,33%	14,59%	20,36%	19,55%	13,38%	11,95%	16,14%		186
Vibrio vulnificus YJ016	41,65%	23,37%	14,10%	14,63%	20,91%	20,32%	13,54%	12,23%	16,70%	5,01%	

Tabela 8.2: Número de proteínas **eliminadas uma a uma**, para o par de cromossomos **número 1**, acima da diagonal, versus o percentual de proteínas eliminadas em relação ao total de proteínas classificadas em famílias, abaixo da diagonal, **com restrição** de *e-value*.

A Tabela 8.3 mostra quantas proteínas são eliminadas uma a uma, em média, nas comparações entre os pares de cromossomos número 1 analisados. Pelo fato do organismo *Escherichia coli* não fazer parte da família *Vibrionaceae*, estamos desconsiderando-o neste momento. A coluna 'Média Proteínas Eliminadas em Relação a Classificadas' informa que, em média, para cada par de cromossomos número 1, 35,55% e 17,44% das proteínas classificadas inicialmente em famílias PRK, respectivamente sem e com restrição de *e-value*, são eliminadas. A coluna 'Média Proteínas Eliminadas em Relação a Totais' informa o percentual médio das proteínas classificadas que foram eliminadas em relação ao número total de proteínas originais contidas no par de cromossomos sendo comparado. A última coluna 'Média Proteínas Eliminadas e Não Classificadas em Relação a Totais' informa o percentual médio das proteínas classificadas que foram eliminadas somadas as proteínas que não foram inicialmente classificadas por famílias PRK ao número total de proteínas originais. Ainda na Tabela 8.3, verifique na última coluna que, quando não é aplicada a restrição de *e-value*, a soma das proteínas não classificadas e das proteínas eliminadas, é menor do que quando aplica-se a restrição de *e-value*.

Base de Famílias	Média Total de Proteínas	Média Proteínas Classificadas em Famílias PRK	Média Proteínas Eliminadas	Média Proteínas Eliminadas em Relação a Prot. Classificadas	Média Proteínas Eliminadas em Relação a Prot. Totais	Média Proteínas Eliminadas e Não Classificadas em Relação a Totais
Protein Clusters	5966,60	5901,60	2111,60	35,55%	35,17%	36,26%
Protein Clusters ($e\text{-}value \leq 10^{-5}$)	5966,60	3711,80	650,64	17,44%	10,91%	48,50%

Tabela 8.3: Média de proteínas **eliminadas uma a uma** em relação as as proteínas classificadas e as proteínas totais (originais), para o par de cromossomos **número 1**.

A Tabela 8.4 apresenta conteúdo similar ao da Tabela 8.3 para o método que elimina os blocos de proteínas. O resultado numérico de interesse aqui é o valor da distância de eliminação para cada um dos métodos de eliminação, que pode ser visualizado através da comparação entre a coluna 'Média Proteínas Eliminadas' da Tabela 8.3 e a coluna 'Média Blocos Eliminados' da Tabela 8.4.

Base de Famílias	Média Total de Proteínas	Média Proteínas Classificadas em Famílias PRK	Média Blocos Eliminados	Média Blocos Eliminados em Relação a Prot. Classificadas
Protein Clusters	5966,60	5901,60	850,91	14,38%
Protein Clusters ($e\text{-}value \leq 10^{-5}$)	5966,60	3711,80	303,09	8,13%

Tabela 8.4: Média de **blocos eliminados** em relação as proteínas classificadas, para o par de cromossomos **número 1**.

Por fim, apresentamos os resultados da execução das variantes do experimento:

1. Classificando inicialmente com famílias PRK, tratando duplicações com árvores ultramétricas, e a seguir, melhorando este tratamento utilizando os grupos ortólogos
2. Classificando inicialmente com famílias PRK, tratando duplicações apenas utilizando com árvores ultramétricas
3. Classificando inicialmente com famílias PRK, tratando duplicações apenas utilizando os grupos ortólogos
4. Classificando inicialmente com famílias PRK, sem tratar duplicações
5. Classificando inicialmente com famílias COG

Considere primeiramente os resultados obtidos pelo método que elimina as proteínas uma a uma. A Tabela 8.5 mostra os resultados deste experimento, quando não é aplicada a restrição de *e-value*, e desconsiderando o organismo *Escherichia coli*. As colunas numéricas apresentadas nesta tabela são similares às colunas existentes na Tabela 8.3. A Tabela 8.6 apresenta conteúdo similar ao da Tabela 8.5 na qual foi aplicada a restrição de *e-value* na classificação inicial de famílias. Os resultados obtidos pelo método que elimina os blocos de proteínas são apresentados nas Tabelas 8.7 e 8.8.

Base de Famílias - Início	Tratamento por Árvores Ultramétricas	Tratamento por Grupos Ortólogos	Média Proteínas Eliminadas	Média Proteínas Eliminadas em Relação a Prot. Classificadas	Média Proteínas Eliminadas em Relação a Prot. Totais	Média Proteínas Eliminadas e Não Classificadas em Relação a Totais
PRK	Sim	Sim	2111,60	35,55%	35,17%	36,26%
PRK	Sim	Não	2375,69	40,03%	39,60%	40,69%
PRK	Não	Sim	3111,87	52,48%	51,92%	53,01%
PRK	Não	Não	3941,24	66,54%	65,82%	66,91%
COG	Não	Não	4263,64	72,11%	71,33%	72,42%

Tabela 8.5: Variações do Experimento - Média de proteínas **eliminadas uma a uma** em relação ao total de proteínas, para o par de cromossomos **número 1**, **sem restrição** de *e-value*.

Base de Famílias - Início	Tratamento por Árvores Ultramétricas	Tratamento por Grupos Ortólogos	Média Proteínas Eliminadas	Média Proteínas Eliminadas em Relação a Prot. Classificadas	Média Proteínas Eliminadas em Relação a Prot. Totais	Média Proteínas Eliminadas e Não Classificadas em Relação a Totais
PRK	Sim	Sim	650,64	17,44%	10,91%	48,50%
PRK	Sim	Não	714,47	19,16%	11,99%	49,57%
PRK	Não	Sim	1292,73	34,71%	21,68%	59,27%
PRK	Não	Não	1442,82	38,75%	24,20%	61,79%
COG	Não	Não	2351,49	63,25%	39,49%	77,08%

Tabela 8.6: Variações do Experimento - Média de proteínas **eliminadas uma a uma** em relação ao total de proteínas, para o par de cromossomos **número 1**, **com restrição** de *e-value*.

Base de Famílias - Início	Tratamento por Árvores Ultramétricas	Tratamento por Grupos Ortólogos	Média Blocos Eliminados	Média Blocos Eliminados em Relação a Prot. Classificadas	Média Blocos Eliminados em Relação a Prot. Totais
PRK	Sim	Sim	850,91	14,38%	14,23%
PRK	Sim	Não	908,96	15,36%	15,20%
PRK	Não	Sim	1053,04	17,84%	17,65%
PRK	Não	Não	1036,91	17,61%	17,42%
COG	Não	Não	929,18	15,79%	5,62%

Tabela 8.7: Variações do Experimento - Média de **blocos eliminados** em relação ao total de proteínas, para o par de cromossomos **número 1**, **sem restrição** de *e-value*.

Base de Famílias - Início	Tratamento por Árvores Ultramétricas	Tratamento por Grupos Ortólogos	Média Blocos Eliminados	Média Blocos Eliminados em Relação a Prot. Classificadas	Média Blocos Eliminados em Relação a Prot. Totais
PRK	Sim	Sim	303,09	8,13%	5,08%
PRK	Sim	Não	319,07	8,56%	5,35%
PRK	Não	Sim	558,24	15,02%	9,38%
PRK	Não	Não	592,38	15,95%	9,96%
COG	Não	Não	709,02	19,12%	11,93%

Tabela 8.8: Variações do Experimento - Média de **blocos eliminados** em relação ao total de proteínas, para o par de cromossomos **número 1**, **com restrição** de *e-value*.

Capítulo 9

Cálculo de Distância de Rearranjo

Após eliminarmos as proteínas, o conjunto de proteínas homólogas de ambos os cromossomos é o mesmo, contendo apenas as proteínas classificadas em famílias binárias. Este conjunto de proteínas é chamado de proteínas finais. Neste ponto, para calcular a distância de ordenação entre dois cromossomos, executaremos um método de rearranjo de genomas. Os métodos de rearranjo de genomas aplicam uma sequência sucessiva de operações ao conjunto de proteínas de um dos cromossomos com o objetivo de deixar estas proteínas na mesma ordenação do outro cromossomo. O valor da distância reflete o número operações de rearranjo realizadas. Os métodos para cálculo da distância de rearranjo de genomas podem se utilizar de vários tipos de operações de rearranjo, e, algumas vezes, também atribuem diferentes valores de peso para cada tipo de evento.

Um dos métodos utilizado para calcular a distância de rearranjo foi o modelo *Double-Cut-And-Join* (DCJ), desenvolvido por Yancopoulos e colegas [66]. O método *Double-Cut-And-Join* (DCJ), quando comparado a outros métodos, apresenta melhor desempenho em termos de tempo de execução, é de fácil implementação e utiliza várias operações de rearranjo: translocação (incluindo fissões e fusões) - com peso 1, inversões - com peso 1, e inter-troca de blocos (incluíndo transposições) - com peso 2. Para a execução do experimento foi utilizada uma implementação simplificada do algoritmo DCJ, desenvolvida por Bergeron e colegas [4].

	Escherichia coli str. K-12 substr. MG1655	*Photobacterium profundum SS9*	*Vibrio cholerae O1 biovar El Tor str. N16961*	*Vibrio cholerae O395*	*Vibrio fischeri ES114*	*Vibrio fischeri MJ11*	*Vibrio harveyi ATCC BAA-1116*	*Vibrio parahaemolyticus RIMD 2210633*	*Vibrio splendidus LGP32*	*Vibrio vulnificus CMCP6*	*Vibrio vulnificus YJ016*
Escherichia coli str. K-12 substr. MG1655		940	889	854	860	866	985	965	933	907	971
Photobacterium profundum SS9	1821		635	608	585	581	674	672	653	632	659
Vibrio cholerae O1 biovar El Tor str. N16961	1692	1793		18	528	527	451	423	441	419	455
Vibrio cholerae O395	1671	1777	2432		516	529	432	407	444	394	412
Vibrio fischeri ES114	1640	1794	1683	1675		32	548	552	538	510	543
Vibrio fischeri MJ11	1655	1789	1693	1708	2298		533	557	552	508	540
Vibrio harveyi ATCC BAA-1116	1798	1908	1886	1901	1781	1791		262	429	361	399
Vibrio parahaemolyticus RIMD 2210633	1781	1910	1882	1867	1788	1788	2181		410	342	392
Vibrio splendidus LGP32	1749	1883	1836	1852	1756	1792	1989	1953		433	484
Vibrio vulnificus CMCP6	1726	1869	1868	1862	1751	1757	2012	2047	1951		102
Vibrio vulnificus YJ016	1780	1892	1912	1895	1763	1770	2045	2082	1975	2438	

Tabela 9.1: Distâncias **DCJ**, para o par de cromossomos **número 1**, acima da diagonal, versus o número de famílias finais, abaixo da diagonal, **sem restrição** de *e-value*.

A Tabela 9.1 apresenta, acima da diagonal, os valores das distâncias DCJ de cada par de cromossomos número 1. Abaixo da diagonal, apresenta o número de famílias finais, ou seja, de famílias PRK ou COG binárias. Note que, como esperado, os organismos da mesma espécie apresentam menores distâncias, bem como os organismos de famílias diferentes apresentam maiores distâncias. Por exemplo, verifique que, para os cromossomos número 1 das cepas da mesma espécie, *Vibrio vulnificus CMCP6* e *Vibrio vulnificus YJ016*, o número de proteínas finais é 2.438, o maior dentre os pares, e a distância DCJ é apenas 102. Note ainda que, os valores não necessariamente satisfazem à desigualdade triangular, pois os conjuntos de proteínas finais não são os mesmos para todos os pares. Por exemplo, observe os resultados na Tabela 9.1 acima da diagonal e verifique que a desigaldade triangular não vale para o trio de cromossomos dos organismos *Vibrio cholerae O1 biovar El Tor str. N16961*, *Vibrio cholerae O395* e *Vibrio fischeri MJ11*. Os valores apresentados na Tabela 9.1 foram computados utilizando as proteínas classificadas sem restrição de *e-value*. A Tabela 9.2 apresenta conteúdo similar ao da Tabela 9.1 na qual foi aplicada restrição de valor de *e-value* na classificação inicial de famílias. Estas tabelas apresentam resultados obtidos pela variante do experimento que executa os dois métodos para tratamento de duplicações: primeiro tratando duplicações com árvores ultramétricas, e a seguir, melhorando este tratamento utilizando os grupos ortólogos.

	Escherichia coli str. K-12 substr. MG1655	*Photobacterium profundum SS9*	*Vibrio cholerae O1 biovar El Tor str. N16961*	*Vibrio cholerae O395*	*Vibrio fischeri ES114*	*Vibrio fischeri MJ11*	*Vibrio harveyi ATCC BAA-1116*	*Vibrio parahaemolyticus RIMD 2210633*	*Vibrio splendidus LGP32*	*Vibrio vulnificus CMCP6*	*Vibrio vulnificus YJ016*
Escherichia coli str. K-12 substr. MG1655		622	588	597	606	601	620	624	613	593	610
Photobacterium profundum SS9	1523		359	356	343	340	341	353	336	331	335
Vibrio cholerae O1 biovar El Tor str. N16961	1420	1481		4	328	321	195	175	200	194	204
Vibrio cholerae O395	1425	1479	1737		324	326	196	179	196	190	194
Vibrio fischeri ES114	1401	1486	1417	1410		6	295	326	292	279	274
Vibrio fischeri MJ11	1404	1493	1410	1411	1724		300	327	303	283	275
Vibrio harveyi ATCC BAA-1116	1463	1539	1533	1532	1453	1462		64	142	114	116
Vibrio parahaemolyticus RIMD 2210633	1465	1547	1535	1535	1479	1483	1694		148	127	122
Vibrio splendidus LGP32	1445	1528	1506	1500	1453	1459	1560	1580		152	145
Vibrio vulnificus CMCP6	1435	1521	1543	1540	1445	1457	1605	1639	1564		29
Vibrio vulnificus YJ016	1464	1530	1560	1552	1447	1455	1615	1647	1566	1764	

Tabela 9.2: Distâncias **DCJ**, para o par de cromossomos **número 1**, acima da diagonal, versus o número de famílias finais, abaixo da diagonal, **com restrição** de *e-value*.

A Tabela 9.3 apresenta o número médio de proteínas finais e os valores médios das distâncias DCJ, nas comparações entre os pares de cromossomos número 1. Pelo fato do organismo *Escherichia coli* não fazer parte da família dos *Vibrionaceae*, estamos desconsiderando-o neste momento. As colunas 'Média Proteínas Finais em Relação a Classificadas' e 'Média Proteínas Finais em Relação a Totais' contém o percentual médio, por par de cromossomos, de proteínas finais em relação as proteínas inicialmente classificadas em famílias PRK e em relação as proteínas originais, respectivamente. Note que as proteínas finais, as quais foram utilizadas no cálculo da distância de rearranjo pelo algoritmo DCJ, são, em média, 63,74% do total das proteínas originais do par, quando não é aplicada restrição de valor de *e-value* na classificação inicial em famílias, e são, em média, 51,50% quando a restrição é aplicada.

Base de Famílias	Média Total de Proteínas	Média Proteínas Classificadas em Famílias PRK	Média Proteínas Finais	Média Proteínas Finais em Relação a Classificadas	Média Proteínas Finais em Relação a Totais	Média Distâncias DCJ	Percentual Médio Distâncias DCJ em Relação às Proteínas Finais
Protein Clusters	5966,60	5901,60	3790,00	64,45%	63,74%	469,38	12,38%
Protein Clusters (*e-value* $\leq 10^{-5}$)	5966,60	3711,80	3061,16	82,56%	51,50%	231,98	7,58%

Tabela 9.3: Média de proteínas finais e médias de distâncias **DCJ**, para o par de cromossomos **número 1**.

Ainda na Tabela 9.3, a coluna 'Percentual Médio das Distâncias DCJ em Relação às Pro-

teínas Finais' contém o percentual médio entre a distância DCJ e o número de proteínas finais. Note pela coluna 'Média Proteínas Finais' que o número de proteínas finais é maior quando não é aplicada a restrição de valor de *e-value*. Note também que esta diferença de 729 (=3.790-3.061) proteínas produz um aumento de 238 (=469-231) operações na distância DCJ, o que implica em praticamente dobrar o valor da distância DCJ.

Por fim, apresentamos os resultados da execução das variantes do experimento:

1. Classificando inicialmente com famílias PRK, tratando duplicações com árvores ultramétricas, e a seguir, melhorando este tratamento utilizando os grupos ortólogos
2. Classificando inicialmente com famílias PRK, tratando duplicações apenas utilizando com árvores ultramétricas
3. Classificando inicialmente com famílias PRK, tratando duplicações apenas utilizando os grupos ortólogos
4. Classificando inicialmente com famílias PRK, sem tratar duplicações
5. Classificando inicialmente com famílias COG

A Tabela 9.4 apresenta os resultados deste experimento, desconsiderando o organismo *Escherichia coli*. As colunas numéricas apresentadas nesta tabelas são similares as colunas existentes na Tabela 9.3. A Tabela 9.5 apresenta conteúdo similar a Tabela 9.4 na qual foi aplicada a restrição de *e-value* na classificação inicial de famílias.

Base de Famílias - Início	Tratamento de Famílias PRK com Duplicações	Agrupamento de Proteínas por COG	Média Proteínas Finais	Média Prot. Finais em Relação a Classificadas	Média Prot. Finais em Relação a Totais	Média Distâncias DCJ	Perc. Médio das Dist. DCJ em Relação às Prot. Finais
PRK	Sim	Sim	3790,00	64,45%	63,74%	469,38	12,38%
PRK	Sim	Não	3525,91	59,97%	59,31%	423,62	12,01%
PRK	Não	Sim	2789,73	47,52%	46,99%	352,47	12,63%
PRK	Não	Não	1960,36	33,46%	33,09%	267,71	13,66%
COG	Não	Não	1637,96	27,89%	27,58%	137,13	8,37%

Tabela 9.4: Variações do Experimento - Média de proteínas finais e média de distâncias **DCJ**, para o par de cromossomos **número 1**, **sem restrição** de *e-value*.

Base de Famílias - Início	Tratamento de Famílias PRK com Duplicações	Agrupamento de Proteínas por COG	Média Proteínas Finais	Média Prot. Finais em Relação a Classificadas	Média Prot. Finais em Relação a Totais	Média Distâncias DCJ	Perc. Médio das Dist. DCJ em Relação às Prot. Finais
PRK	Sim	Sim	3061,16	82,56%	51,50%	231,98	7,58%
PRK	Sim	Não	2997,33	80,84%	50,43%	207,36	6,92%
PRK	Não	Sim	2419,07	65,29%	40,73%	169,00	6,99%
PRK	Não	Não	2268,98	61,25%	38,21%	150,38	6,63%
COG	Não	Não	1360,31	36,75%	22,92%	107,82	7,93%

Tabela 9.5: Variações do Experimento - Média de proteínas finais e média de distâncias **DCJ**, para o par de cromossomos **número 1**, **com restrição** de *e-value*.

Por fim, na Tabela 9.6 são apresentados os valores médios de proteínas finais e de distâncias DCJ calculados somente a partir das comparações, sem restrições de *e-value*, entre os cromossomos número 1 de cepas da mesma espécie, ou seja, das comparações entre os pares *Vibrio cholerae O1 biovar El Tor str. N16961* e *Vibrio cholerae O395*, *Vibrio fischeri ES114* e *Vibrio fischeri MJ11* e *Vibrio vulnificus CMCP6* e *Vibrio vulnificus YJ016*.

Base de Famílias - Início	Tratamento de Famílias PRK com Duplicações	Agrupamento de Proteínas por COG	Média Proteínas Finais	Média Prot. Finais em Relação a Classificadas	Média Prot. Finais em Relação a Totais	Média Distâncias DCJ	Perc. Médio das Dist. DCJ em Relação às Prot. Finais
PRK	Sim	Sim	4778,67	86,37%	85,44%	50,67	1,06%
PRK	Sim	Não	4704,00	85,02%	84,10%	45,00	0,96%
PRK	Não	Sim	3537,33	64,09%	63,39%	41,00	1,16%
PRK	Não	Não	2716,67	49,32%	48,78%	26,00	0,96%
COG	Não	Não	1812,67	32,79%	32,44%	8,33	0,46%

Tabela 9.6: Variações do Experimento - Média de proteínas finais e média de distâncias **DCJ**, para o par de cromossomos **número 1** de **cepas de mesma espécie**, **sem restrição** de *e-value*.

Capítulo 10

Construção e Análise de Filogenias

Neste capítulo serão analisados os resultados do experimento, através dos valores das matrizes de distâncias bem como das árvores filogenéticas produzidas por estas matrizes. As árvores obtidas serão comparadas com a árvore filogenética inferida a partir do gene 16S rRNA dos genomas analisados.

O valor da distância total entre dois cromossomos é composto por dois termos: o valor da distância de eliminação e o valor da distância de ordenação. A distância de eliminação é o número das proteínas eliminadas de ambos cromossomos do par que estavam classificadas em famílias não binárias. A distância de eliminação também pode ser o número de blocos de proteínas contíguas eliminadas, dependendo do método de eliminação selecionado. A este valor de distância de eliminação não está sendo computado o número de proteínas que não puderam ser inicialmente classificadas em famílias PRK. O objetivo aqui é o de não penalizar o valor da distância total por falhas no método de classificação de proteínas em famílias. O valor da distância de ordenação é o valor da distância de rearranjo calculada pelo modelo *Double-Cut-And-Join* (DCJ).

O valor da distância total é dado pela fórmula é:

$$DistTotal = DistElim + DistOrdem$$

Os valores que compõem esta fórmula são:

- $DistElim$ = Distância de eliminação, que é o número de proteínas eliminadas uma a uma ou o número de blocos de proteínas eliminadas
- $DistOrdem$ = Distância de ordenação calculada pela distância de DCJ

Para este trabalho, os valores de distâncias de ordenação e de eliminação possuem mesmo peso na composição da fórmula. Não existe consenso entre os pesquisadores sobre quais

seriam os pesos ideais. O peso relativo destas parcelas na equação final é controverso e objeto atual de pesquisas tais como a realizada por Mirkin e colegas [49], onde, em seu trabalho sobre ancestral comum, experimentam variações de pesos nos eventos de ganho e perda de genes.

A Tabela 10.1 apresenta a matriz de distâncias totais calculada selecionando a variação do experimento em que as proteínas são eliminadas uma a uma e que a distância de rearranjo é calculada pelo modelo DCJ, para cada par de cromossomos número 1. Abaixo da diagonal, os valores foram obtidos a partir da classificação de famílias sem qualquer restrição de valor de *e-value*. Acima da diagonal, os valores foram obtidos a partir da classificação de famílias com a restrição do valor de *e-value*. A Tabela 10.2 apresenta a matriz de distâncias totais calculada selecionando a variação do experimento em que as proteínas são eliminadas em blocos e que a distância de rearranjo é calculada pelo modelo DCJ, para cada par de cromossomos número 1. Ambas as Tabelas 10.1 e 10.2 apresentam resultados obtidos pela variação do experimento que executa os dois métodos para tratamento de duplicações: primeiro tratando duplicações com árvores ultramétricas, e a seguir, melhorando este tratamento utilizando os grupos ortólogos.

	Escherichia coli str. K-12 substr. MG1655	*Photobacterium profundum SS9*	*Vibrio cholerae O1 biovar El Tor str. N16961*	*Vibrio cholerae O395*	*Vibrio fischeri ES114*	*Vibrio fischeri MJ11*	*Vibrio harveyi ATCC BAA-1116*	*Vibrio parahaemolyticus RIMD 2210633*	*Vibrio splendidus LGP32*	*Vibrio vulnificus CMCP6*	*Vibrio vulnificus YJ016*
Escherichia coli str. K-12 substr. MG1655		2843	2654	2657	2737	2719	2704	2721	2757	2711	2700
Photobacterium profundum SS9	4828		1278	1283	1279	1255	1248	1261	1289	1252	1268
Vibrio cholerae O1 biovar El Tor str. N16961	4303	3143		54	1041	1041	753	746	836	710	716
Vibrio cholerae O395	4340	3178	546		1055	1048	760	754	848	716	726
Vibrio fischeri ES114	4267	2980	2413	2447		125	1040	1036	1061	1018	1039
Vibrio fischeri MJ11	4250	2993	2399	2401	583		1020	1022	1053	991	1017
Vibrio harveyi ATCC BAA-1116	5005	3770	2859	2840	3055	3027		421	774	610	622
Vibrio parahaemolyticus RIMD 2210633	4526	3271	2346	2390	2552	2564	2405		757	572	581
Vibrio splendidus LGP32	4478	3226	2376	2377	2522	2471	2876	2436		754	773
Vibrio vulnificus CMCP6	4487	3222	2279	2296	2493	2486	2751	2169	2372		215
Vibrio vulnificus YJ016	4738	3498	2522	2543	2797	2787	3018	2444	2670	1351	

Tabela 10.1: Valores de distâncias totais, calculada pela **eliminação uma a uma** das proteínas e pela distância **DCJ**, para os cromossomos **número 1**, classificados **sem restrição** de *e-value*, abaixo da diagonal, versus, valores de distâncias totais, para cromossomos **número 1**, classificados **com restrição** de *e-value*, acima da diagonal.

	Escherichia coli str. K-12 substr. MG1655	*Photobacterium profundum SS9*	*Vibrio cholerae O1 biovar El Tor str. N16961*	*Vibrio cholerae O395*	*Vibrio fischeri ES114*	*Vibrio fischeri MJ11*	*Vibrio harveyi ATCC BAA-1116*	*Vibrio parahaemolyticus RIMD 2210633*	*Vibrio splendidus LGP32*	*Vibrio vulnificus CMCP6*	*Vibrio vulnificus YJ016*
Escherichia coli str. K-12 substr. MG1655		1371	1241	1243	1264	1237	1283	1288	1296	1259	1280
Photobacterium profundum SS9	2178		753	756	745	737	750	757	752	767	746
Vibrio cholerae O1 biovar El Tor str. N16961	1990	1569		44	654	643	465	431	486	455	450
Vibrio cholerae O395	1912	1512	318		652	657	474	445	492	455	449
Vibrio fischeri ES114	1917	1465	1358	1283		73	633	652	644	645	619
Vibrio fischeri MJ11	1907	1472	1351	1290	327		632	654	658	636	608
Vibrio harveyi ATCC BAA-1116	2281	1742	1364	1284	1446	1413		267	429	382	370
Vibrio parahaemolyticus RIMD 2210633	2165	1688	1293	1233	1419	1418	1156		411	387	370
Vibrio splendidus LGP32	2108	1667	1292	1229	1359	1367	1330	1299		452	414
Vibrio vulnificus CMCP6	2047	1615	1268	1199	1352	1346	1291	1210	1297		127
Vibrio vulnificus YJ016	2182	1695	1345	1264	1415	1431	1383	1330	1386	642	

Tabela 10.2: Valores de distâncias totais, calculada pela **eliminação em blocos** das proteínas e pela distância **DCJ**, para os cromossomos **número 1**, classificados **sem restrição** de *e-value*, abaixo da diagonal, versus, valores de distâncias totais, para cromossomos **número 1**, classificados **com restrição** de *e-value*, acima da diagonal.

Para analisar os resultados, na Figura 10.1 apresentamos a árvore filogenética inferida a partir do conjunto de genes 16S rRNA dos organismos analisados neste trabalho. Através do programa *MEGA 4* [58], realizamos o alinhamento dos genes 16S rRNA, e, a seguir, a árvore filogenética foi construída utilizando-se o método *Neighbor-Joining* (NJ) e com modelo de substituição *Jones-Taylor-Thornton* (JTT). A Figura 10.2 contém a árvore filogenética inferida a partir do conjunto de genes 16S rRNA construída utilizando-se o método *Neighbor-Joining* (NJ) e com modelo de substituição *PAM Matrix (Dayhoff)*. Compare estas duas árvores e verifique que existe uma troca de ordem entre os organismos *Vibrio cholerae* e *Vibrio splendidus*.

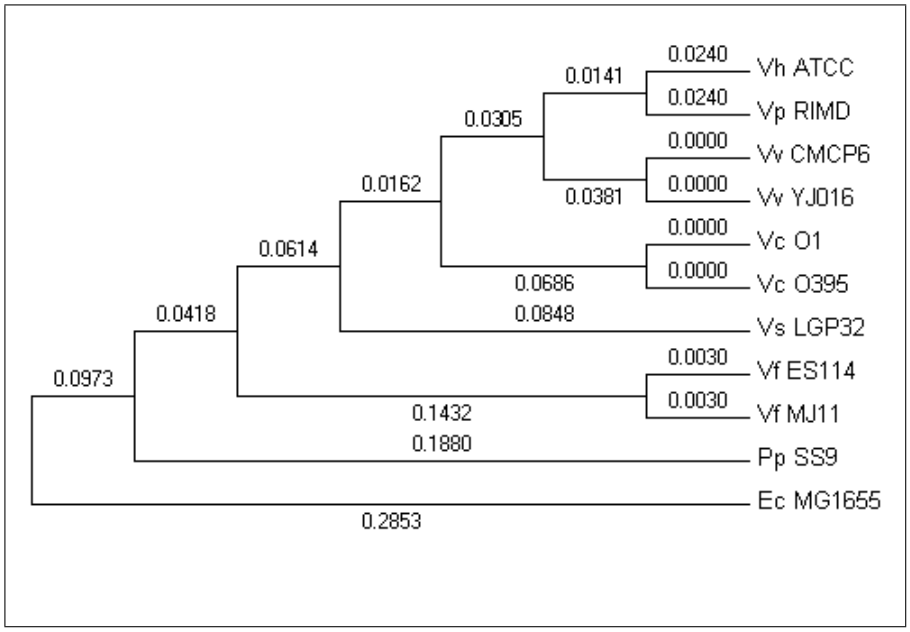

Figura 10.1: Árvore filogenética dos genes **16S rRNA** dos organismos analisados, calculada com modelo de substituição **JTT** e inferida pelo método *Neighbor-Joining*.

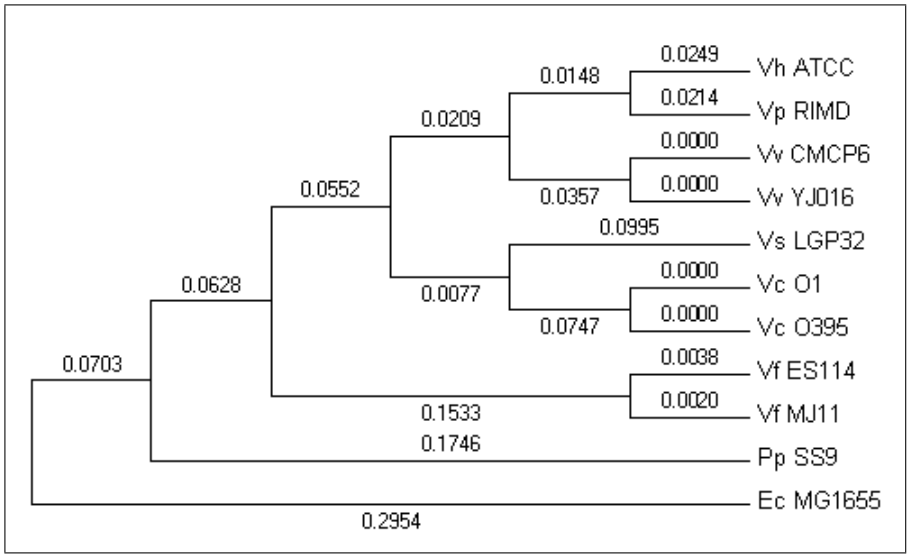

Figura 10.2: Árvore filogenética dos genes **16S rRNA** dos organismos analisados, calculada com modelo de substituição **PAM Matrix** e inferida pelo método *Neighbor-Joining*.

A Figura 10.3 apresenta a árvore filogenética produzida a partir dos valores das distâncias totais apresentados na Tabela 10.1, para os cromossomos número 1, sem restrição de *e-value* na classificação inicial de proteínas em famílias. A Figura 10.4 apresenta a árvore filogenética produzida a partir dos valores de distâncias totais apresentados na

Tabela 10.1, com restrição de *e-value*. A matriz com as distâncias totais é entrada para o programa *neighbor.exe* do pacote de ferramentas *PHYLIP* [15], que constrói a árvore filogenética utilizando o método *Neighbor-Joining* (NJ). As árvores aqui apresentadas foram visualizadas através da ferramenta *Tree Explorer* do programa *MEGA 4*, com a opção para organizar os genomas por balanceamento.

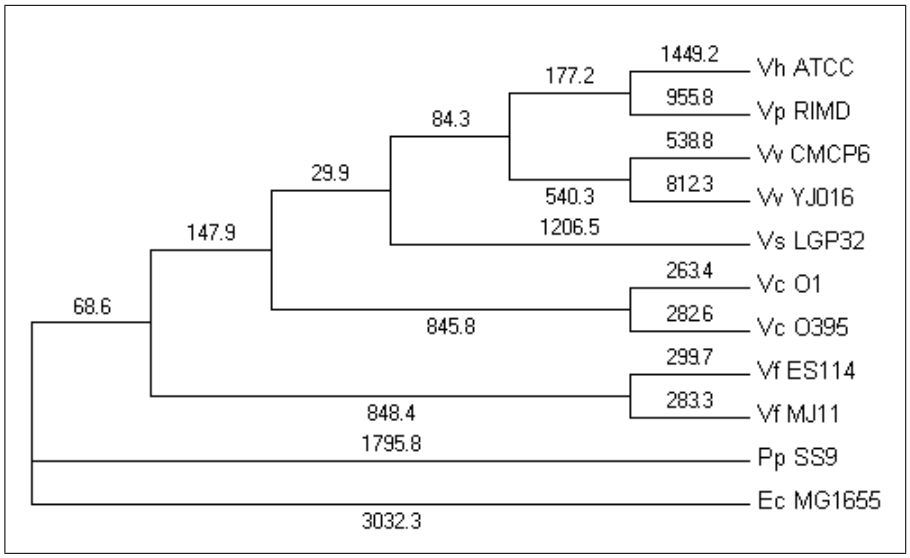

Figura 10.3: Árvore filogenética dos cromossomos **número 1**, com distância de proteínas **eliminadas uma a uma** somada a distância **DCJ sem restrição** de *e-value*.

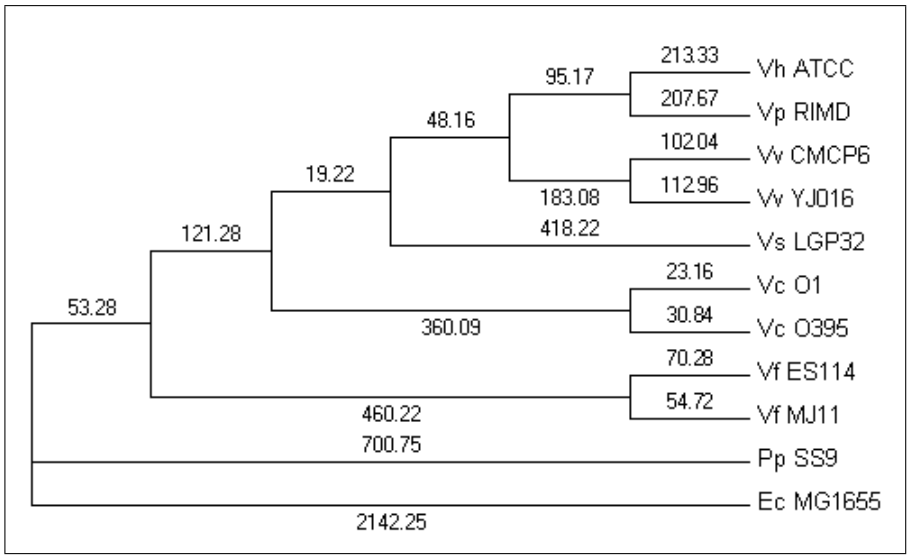

Figura 10.4: Árvore filogenética dos cromossomos **número 1**, com distância de proteínas **eliminadas uma a uma** somada a distância **DCJ com restrição** de *e-value*.

As árvores filogenéticas das Figura 10.3 e Figura 10.4 são congruentes em topologia com árvore filogenética inferida a partir dos genes 16S rRNA utilizando o modelo de substituição *PAM Matrix*. As cepas de mesma espécie, *Vibrio cholerae*, *Vibrio fischeri* e *Vibrio vulnificus* permaneceram agrupadas.

As Figuras 10.5 e 10.6 apresentam as árvores filogenéticas produzidas a partir dos valores das distâncias totais obtidos nas comparações entre os cromossomos número 2, sem e com a restrição de *e-value* aplicada a classificação inicial de proteínas em famílias, respectivamente. Estas árvores são congruentes em topologia com as árvores inferidas a partir das comparações entre os cromossomos número 1.

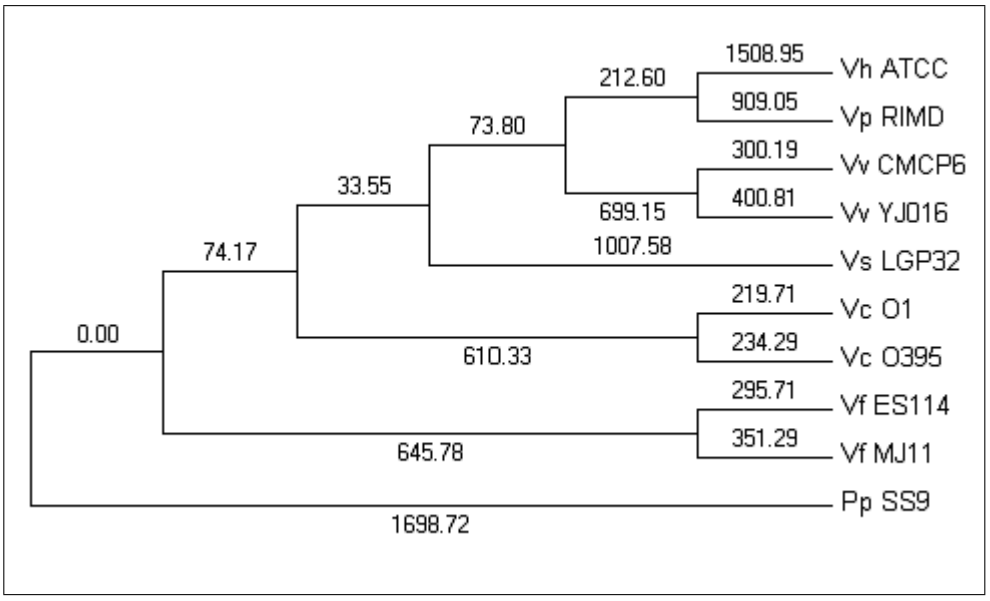

Figura 10.5: Árvore filogenética dos cromossomos **número 2**, com distância de proteínas **eliminadas uma a uma** somada a distância **DCJ sem restrição** de *e-value*.

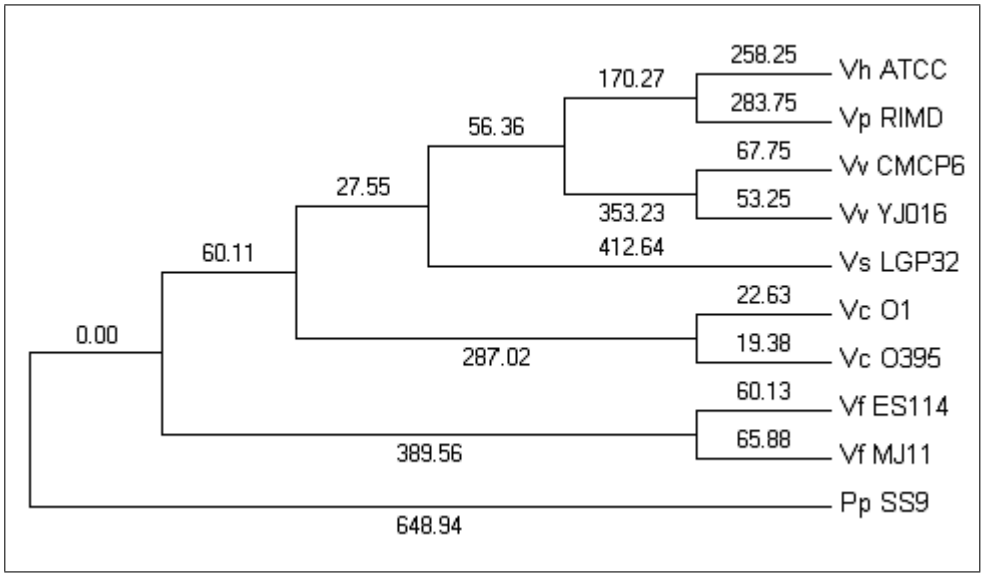

Figura 10.6: Árvore filogenética dos cromossomos **número 2**, com distância de proteínas **eliminadas uma a uma** somada a distância **DCJ com restrição** de *e-value*.

As Figuras 10.7 e 10.8 apresentam as árvores filogenéticas produzidas por Côgo. Em relação às árvores apresentadas nas Figuras 10.3, 10.4, 10.5 e 10.6 note que, retirando-se os organismos não utilizados por Côgo, resulta a mesma topologia.

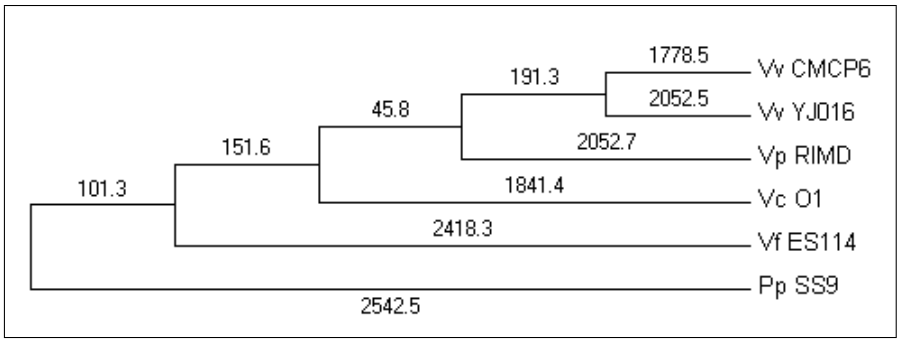

Figura 10.7: Árvore filogenética dos cromossomos **número 1** produzida por **Côgo**.

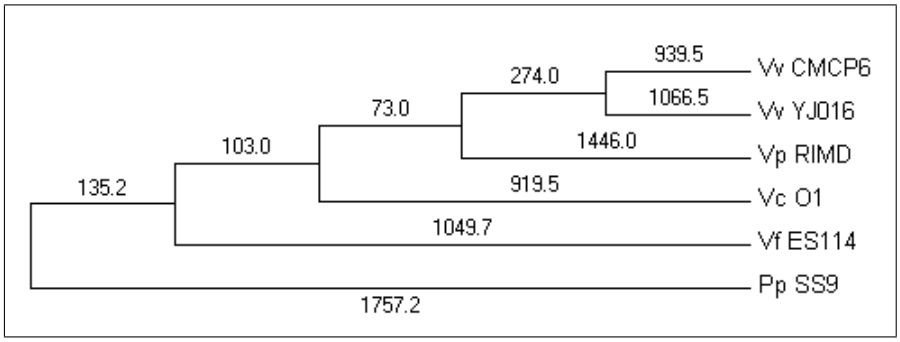

Figura 10.8: Árvore filogenética dos cromossomos **número 2** produzida por **Côgo**.

A Figura 10.9 apresenta a árvore filogenética produzida a partir dos valores das distâncias totais apresentados na Tabela 10.2, para os cromossomos número 1, sem restrição de *e-value* na classificação inicial de proteínas em famílias. A Figura 10.10 apresenta a árvore filogenética produzida a partir dos valores de distâncias totais apresentados na Tabela 10.2, com restrição de *e-value*. Estas árvores não são congruentes em topologia com as árvores apresentadas nas Figuras 10.3 e 10.4 havendo uma pequena diferença no posicionamento do organismo *Photobacterium profundum*.

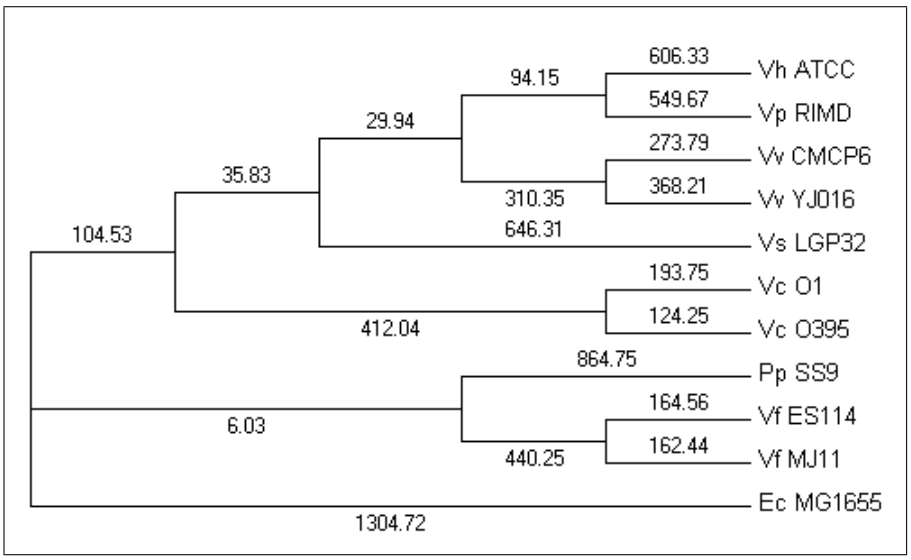

Figura 10.9: Árvore filogenética dos cromossomos **número 1**, com distância de proteínas **eliminadas em blocos** somada a distância **DCJ sem restrição** de *e-value*.

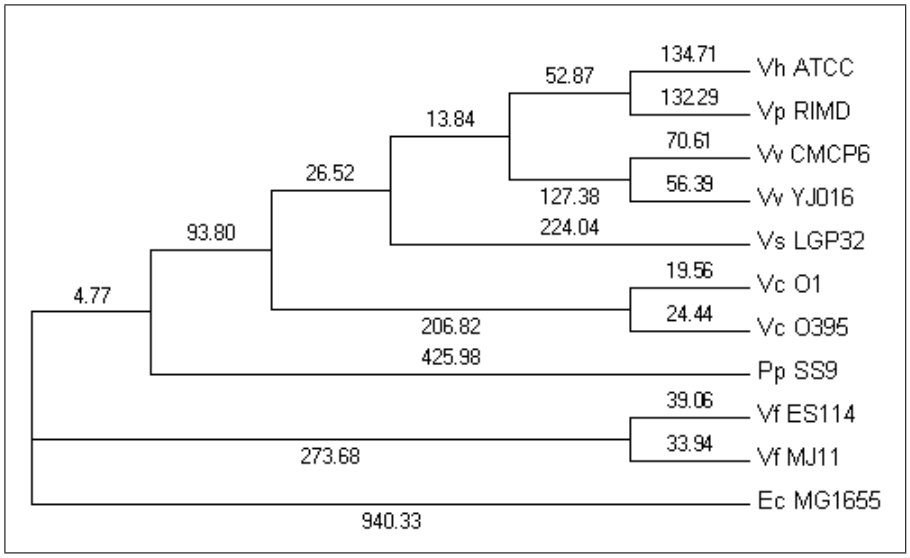

Figura 10.10: Árvore filogenética dos cromossomos **número 1**, com distância de proteínas **eliminadas em blocos** somada a distância **DCJ com restrição** de *e-value*.

A seguir, a Figura 10.11 apresenta a árvore inferida a partir das distâncias de eliminação de proteínas uma a uma para os cromossomos número 1, sem restrição de *e-value*. Esta árvore se mostra congruente em topologia com árvore filogenética inferida a partir dos genes 16S rRNA e utilizando o modelo de substituição *JTT*, apresentando assim uma troca no posicionamento dos organismos *Vibrio cholerae* e *Vibrio splendidus* em relação as árvores apresentadas nas Figuras 10.3 e 10.4. A seguir, a Figura 10.12 apresenta a árvore inferida a partir das distâncias de eliminação de proteínas em blocos para os cromossomos número 1, sem restrição de *e-value*. Em relação às árvores apresentadas nas Figuras 10.9 e 10.10 esta árvore apresenta uma difença no posicionamento do organismo *Photobacterium profundum* e também no posicionamento do organismo *Vibrio splendidus*. Por fim, a Figura 10.13 apresenta a árvore inferida a partir das distâncias de rearranjo DCJ para os cromossomos número 1, sem restrição de *e-value*. Esta árvore é congruente em topologia com as árvores apresentadas nas Figuras 10.3 e 10.4 e com com árvore filogenética inferida a partir dos genes 16S rRNA e utilizando o modelo de substituição *PAM Matrix*. Todas estas árvores apresentam resultados obtidos pela variação do experimento que executa os dois métodos para tratamento de duplicações: primeiro tratando duplicações com árvores ultramétricas, e a seguir, melhorando este tratamento utilizando os grupos ortólogos.

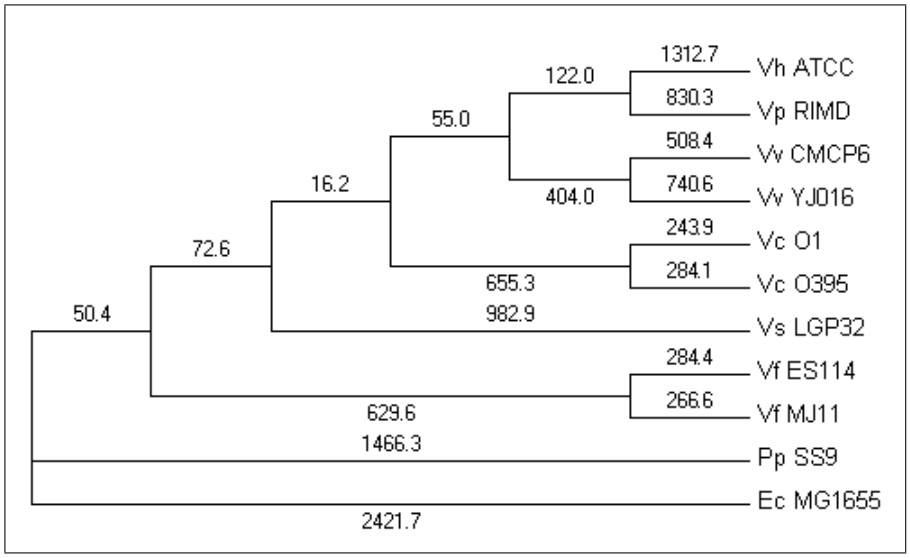

Figura 10.11: Árvore filogenética das distância de proteínas **eliminadas uma a uma** dos cromossomos **número 1**, **sem restrição** de *e-value*.

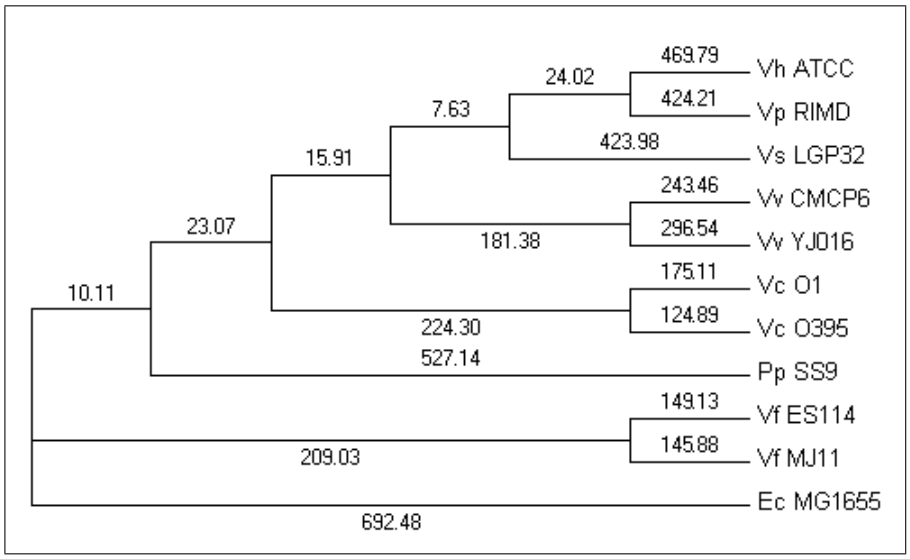

Figura 10.12: Árvore filogenética das distância de proteínas **eliminadas em blocos** dos cromossomos **número 1**, **sem restrição** de *e-value*.

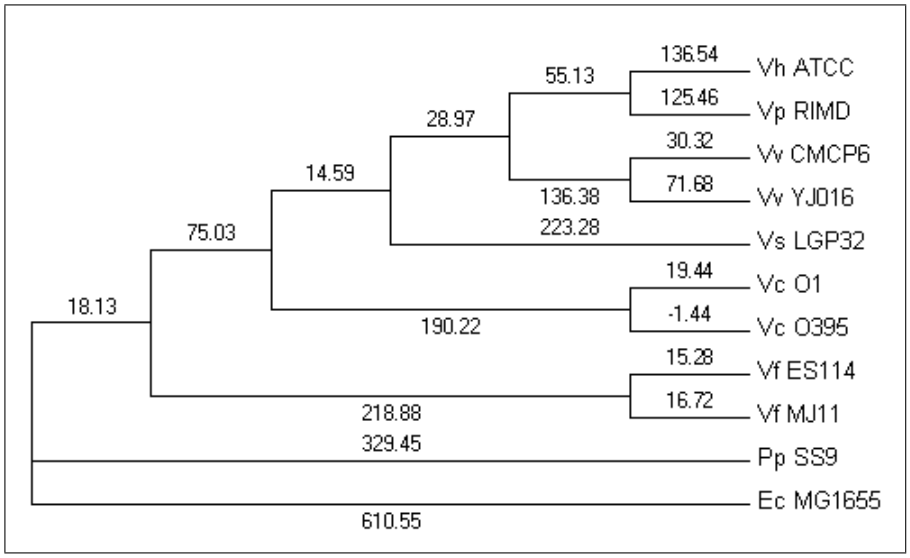

Figura 10.13: Árvore filogenética das distância de rearranjo **DCJ** dos cromossomos **número 1**, **sem restrição** de *e-value*.

Analisemos agora algumas variantes do experimento. Dado que as proteínas foram eliminadas uma a uma e a distância de rearranjo foi calculada pelo modelo DCJ, considere as seguintes comparações entre os cromossomos número 1, sem a aplicação da restrição ao *e-value* na classificação inicial de proteínas em famílias:

1. Classificando inicialmente com famílias PRK, tratando duplicações com árvores ultramétricas, e a seguir, melhorando este tratamento utilizando os grupos ortólogos
2. Classificando inicialmente com famílias PRK, tratando duplicações apenas utilizando com árvores ultramétricas
3. Classificando inicialmente com famílias PRK, tratando duplicações apenas utilizando os grupos ortólogos
4. Classificando inicialmente com famílias PRK, sem tratar duplicações
5. Classificando inicialmente com famílias COG

A árvore inferida a partir da variante número 1 foi apresentada na Figura 10.3. As árvores inferidas a partir dos valores de distâncias totais calculados a pelas variantes número 2, número 3 e número 4 apresentam topologia idêntica à árvore produzida pelo variante número 1. A árvore inferida a partir dos valores de distâncias totais calculados pela variante número 5 do experimento, que se utiliza apenas da classificação COG, não

apresenta topologia similar à árvore produzida pela variante número 1. Porém, ainda para a variante número 5, a árvore inferida a partir somente dos valores de distâncias de ordenação (DCJ) é congruente em topologia à árvore produzida pela variante número 1.

Por fim, no material suplementar está incluído uma implementação em Java do procedimento de comparação descrito neste trabalho. A Tabela 10.3 apresenta os tempos médios de execução das comparações entre os cromossomos número 1 quando é aplicada e quando não é aplicada a restrição de *e-value* na classificação inicial de famílias. A execução foi realizada em um computador Pentium 4, 2.8Ghz com 1.5Gb de memória RAM e sistema operacional *MS Windows XP*.

Base de Famílias	Tempo de Execução (Minutos)	Número de Comparações	Tempo Médio (Minutos)
Protein Clusters	318	55	5,78
Protein Clusters (*e-value* $\leq 10^{-5}$)	197	55	3,58

Tabela 10.3: Tempos de execução da comparação entre os pares de cromossomos número 1.

Capítulo 11

Conclusão

Este trabalho executou um experimento de comparação entre dez genomas completos da família *Vibrionaceae* e o genoma completo do organismo *Escherichia coli*. O modelo de comparação é dividido em três fases. A primeira fase classifica os genes de um genoma em famílias universais de genes homólogos. A segunda fase tem por objetivo restringir cada par de genomas a um conjunto comum de genes, dando tratamento adequado aos eventos de duplicações de genes (paralogias) e as eventos de perda e ganho de genes. Por fim, a distância de rearranjo é calculada para cada par de genomas. A árvore filogenética é inferida com base na matriz das distâncias.

Esta tese é, de certa forma, uma continuação natural da tese de mestrado da Patrícia Pilisson Côgo [8]. Continuando seu trabalho, adicionamos à análise os genomas completos de mais quatro vibriões: *Vibrio cholerae O395*, *Vibrio fischeri MJ11*, *Vibrio harveyi ATCC BAA-1116* e *Vibrio splendidus LGP32*. Também adicionamos um genoma completo da bactéria *Escherichia coli*, que pertence à família das *Enterobacteriaceae*, representando grupo externo.

Para identificar as homologias, Côgo construiu sua base de famílias utilizando inicialmente a base HAMAP. Porém, a base do HAMAP classifica apenas uma pequena parcela das proteínas. Assim, para seu experimento, Côgo desenvolveu uma metodologia de descrição e criação de famílias, baseada em similaridade de sequências. Partindo inicialmente da base de famílias do HAMAP, novas famílias foram adicionadas automaticamente. Porém, estas novas famílias foram criadas utilizando-se somente do universo das proteínas dos seis genomas completos analisados, ou seja, o conjunto de famílias não era independente do conjunto de genomas analisados. Usando a base de famílias *Protein Clusters*, que é construída utilizando-se de todas as proteínas da base do NCBI, podemos então aumentar significativamente a independência entre as famílias e o conjunto dos genomas analisados.

Além disso, esta base é curada continuamente por especialistas. A utilização da base do *Protein Clusters* permitiu também que o tempo para inclusão de um novo genoma na análise fosse reduzido de uma semana para algumas horas.

Para a identificação das ortologias e paralogias das famílias com duplicações, utilizamos a abordagem de Côgo, que se utiliza de árvores ultramétricas, e adicionamos um refinamento ao tratamento de duplicações realizando a reclassificação das proteínas por grupos ortólogos (COG). Na solução proposta, após ser realizado o tratamento utilizando árvores ultramétricas, as proteínas que ainda estiverem classificadas em famílias unárias ou em famílias com duplicações são reclassificadas por grupos ortólogos, formando três tipos de grupos: unários, binários e com duplicações. As proteínas classificadas em grupos ortólogos binários serão mantidas. Todas as demais proteínas, classificadas em grupos ortólogos unários ou com duplicações, são descartadas. A aplicação do tratamento de duplicações por reclassificação em grupos ortólogos, quando aplicado após o tratamento por árvores ultramétricas, produz pequenas melhorias, com um aumento no número de proteínas finais em média de 7,49% quando não é aplicada a restrição de *e-value* na classificação inicial de famílias e de 2,64% quando é aplicada a restrição. A aplicação isolada do procedimento de tratamento de duplicações por reclassificação em grupos ortólogos também apresenta bons resultados, com um aumento no número de proteínas finais em média de 41,10% quando não é aplicada a restrição de *e-value* na classificação inicial de famílias e de 7,17% quando é aplicada a restrição. Este método não apresenta melhores resultados que o tratamento por árvores ultramétricas, porém, tem a vantagem de ser executado em muito menos tempo.

Após os procedimentos de identificação das ortologias e paralogias, se aplicarmos ambos os tratamentos de duplicações (utilizando árvores ultramétricas seguido pela reclassificação por grupos ortólogos), aproximadamente 63,74% das proteínas originais são mantidas nas comparações entre os cromossomos número 1 da família dos *Vibrionaceos*, quando não é aplicada restrição de *e-value* na classificação inicial de famílias. Quando é aplicada a restrição, aproximadamente 51,50% das proteínas originais são mantidas. Se selecionarmos somente as comparações entre cepas de mesma espécie, aproximadamente 85,44% das proteínas originais são mantidas quando não é aplicada a restrição de *e-value* na classificação inicial de famílias.

Ao eliminar as proteínas, Côgo adicionou à distância de eliminação 1 ponto para cada proteína eliminada. Este trabalho implementa aqui uma segunda opção, onde as proteínas são eliminadas em blocos contíguos, e é adicionado à distância de eliminação 1 ponto para cada bloco de proteínas eliminado. O experimento foi executado para ambos os métodos.

Quando as proteínas são eliminadas uma a uma, a árvore filogenética produzida se mostra congruente com a árvore inferida a partir dos genes 16S rRNA e utilizando o modelo de substituição *PAM Matrix*. Quando as proteínas são eliminadas em blocos, a árvore produzida é similar a anterior havendo apenas uma pequena diferença no posicionamento do organismo *Photobacterium produndum*.

Após a eliminação das proteínas, ambos os cromossomos foram reduzidos a um mesmo conteúdo, contendo o que denominamos de famílias finais. É sobre este conjunto de proteínas classificadas nas famílias finais que é calculada a distância de rearranjo. Côgo utilizou o modelo *Double-Cut-And-Join* (DCJ), modelo também utilizado neste trabalho para calcular a distância de rearranjo.

O experimento de comparação foi executado num total de 18 variantes: 16 variantes quando as famílias são inicialmente classificadas pela base *Protein Clusters* e 2 variantes quando famílias são inicialmente classificadas por grupos ortólogos (COG). Note que só foi possível executar todas estas variantes do experimento devido à implementação computacional e automatização total do procedimento de comparação. A automatização facilitou a coleta de dados numéricos que foram utilizados para comparar os resultados obtidos em cada uma das variantes do experimento.

A Tabela 11.1 compara e mostra as principais diferenças entre este trabalho e o trabalho de Côgo.

Item	Côgo	Neste Trabalho
Número de Genomas Analisados	Seis	Onze
Genomas Fora do Grupo	Não tem	*Escherichia coli*
Determinação de Homologias	Base criada a partir do HAMAP	Base *Protein Clusters*
Tratamento de Duplicações	Utilização de árvores ultramétricas	Utilização de árvores ultramétricas e reclassificação por grupos ortólogos
Eliminação de Proteínas	Eliminação simples, uma a uma	Eliminação simples, uma a uma ou eliminação por blocos
Cálculo da Distância de Rearranjo	Pelo modelo DCJ	Pelo modelo DCJ
Automatização do Procedimento de Comparação	Praticamente inexistente	Total
Coleta de Resultados Númericos	Não	Sim, em cada etapa

Tabela 11.1: Sumário comparativo do experimento realizado por Côgo com o experimento realizado neste trabalho.

11.1 Trabalhos Futuros

Como trabalhos futuros cremos que as regiões intergênicas deveriam ser incluídas na análise. Com relação a homologias, seria interessante implementar um mecanismo que permitisse incluir ou não na comparação as proteínas desconhecidas, hipotéticas ou putativas. Como foi dito, tratamos os eventos de transferência horizontal de genes (THG) como eventos de perda de genes. Uma melhoria neste ponto seria remover da comparação as proteínas oriundas de eventos THG consultando a base do HGT-DB [24], e somando estes eventos ao valor da distância total com um peso apropriado.

Também seria interessante incluir mais genomas de vibriões no experimento, e utilizar como *benchmark* árvores inferidas em trabalhos de biólogos, tais como as encontradas no trabalho de Thompson e colegas [62]. Thompson inferiu árvores filogenéticas para um conjunto de trinta genomas de vibriões utilizando o método MLSA (*Multilocus Sequence*

Analysis). Também, os resultados poderiam ser comparados com árvores produzidas a partir de distâncias calculadas pelo índice de Karlin [33].

Por fim, otimizações no programa incluído no material suplementar, que implementa o procedimento de comparação descrito neste trabalho, permitiriam comparar um número maior de genomas completos num tempo menor. A etapa que mais consome tempo é o tratamento de famílias com duplicações por árvores ultramétricas. Para realizar este tratamento, os programas *ClustalW2* e *Protdist* são executados externamente ao programa do experimento, o que é um fator que aumenta o tempo de execução.

Apêndice A

Material Suplementar

Este capítulo descreve o conteúdo e organização do material suplementar, que basicamente é uma implementação computacional do experimento descrito nesta tese. Este material pode ser obtido na localização `http://www.ic.unicamp.br/~meidanis/PUB/Mestrado/2006-Zupo/material_suplementar.zip`. O arquivo *material_suplementar.zip* é composto por:

- *diretório raíz*: contém os arquivos de projeto da *IDE Eclipse*.
- *bin*: diretório que contém os programas binários *clustalw2.exe*, *protdist.exe* e *rpsblast.exe*.
- *classes*: diretório que contém os arquivos objeto do código Java
- *input*: diretório que contém os arquivos de entrada, contendo informações sobre os cromossomos e suas proteínas
- *output*: diretório que contém os arquivos de saída, contendo logs e as tabelas com os resultados das comparações
- *src*: diretório que contém os arquivos fonte do código Java
- *temp*: diretório utilizado para gravação de arquivos temporários

As seguintes ferramentas, compatíveis com sistema operacional *MS Windows XP*, foram utilizadas na implementação:

- Java SE 6.0 (`http://java.sun.com/javase/6/`)
- ClustalW2 2.0.6 [40]

- Protdist, pacote PHYLIP 3.67 [15]
- RpsBlast 2.2.18 [46]
- Eclipse 3.3.2 (`http://www.eclipse.org/downloads/moreinfo/classic.php`)

Para cada cromossomo utilizado neste experimento, existem 3 tipos de arquivos de entrada:

- *<NCBI RefSeq>-protein.txt*: tabela com informações de suas proteínas
- *<NCBI RefSeq>-protein-FASTA.txt*: sequências, em formato *fasta*, de suas proteínas
- *<NCBI RefSeq>-protein-FASTA.rpsblast*: classificação das proteínas em famílias *Protein Cluster*

Por exemplo, para obter os arquivo de entrada do cromossomo número 1 da bactéria *Escherichia coli str. K-12 substr. MG1655*, *RefSeq NC_000913*, acesse sua página no NCBI `http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome&cmd=Retrieve&list_uids=115`. Navegue para a página de detalhes da bactéria *Escherichia coli str. K-12 substr. MG1655* clicando em *'NC_000913'*. Para obter a tabela com informações de suas proteínas, selecione a opção *'Protein Table'* da lista *'Display'*, e a opção *'Text'* da lista *'Show´*. Para obter as sequências das proteínas, selecione a opção *'Protein FASTA'* da lista *'Display'*, e a opção *'Text'* da lista *'Show´*. Para obter a classificação das proteínas em famílias *Protein Cluster*, consulte o apêndice B.

O arquivo de saída *cromo1-results.csv* contém os resultados da execução da comparação entre os cromossomos número 1 e o arquivo *cromo2-results.csv*, da comparação entre os cromossomos número 2. O arquivo *results-from-excel.csv* contém os mesmos resultados numa melhor organização e pode ser diretamente visualizado em editores de planilhas de cálculo tais como *MS Excel* ou *OpenOffice.org Calc*.

O experimento pode ser executado a partir da execução do método *main* contido na classe *genomecomparison.main.Main*.

Apêndice B

Base de famílias *Protein Clusters*

Para criar a base do *Protein Clusters* e classificar as proteínas em famílias deve-se baixar o arquivo `PRK_Clusters.pssm.tgz`, com os perfis das famílias, do NCBI. Por exemplo, a localização `ftp://ftp.ncbi.nih.gov/genomes/Bacteria/CLUSTERS/Sep_2009/PRK` fornece o arquivo com os perfis de famílias para bactérias, na versão de Setembro de 2009. Lembrando que neste trabalho foi utilizada a versão *May_2008* dos perfis de famílias para bactérias.

Depois de descompactar o arquivo anterior, deve-se executar o programa *formatrpsdb* para criar a base:

```
formatrpsdb -i Prk.pn -o T -f 9.82 -n Prk -S 100.0
```

Este comando criará a base que contém arquivos tais como: *Prk.rps* e *Prk.loo*. Atenção, este programa depende do arquivo *blosum62*, com a matriz de substituição usada no alinhamento das sequências de proteínas com limiar definido em 62%. Para a descrição detalhada das opções de linha de comando do programa *formatrpsdb* acesse `http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/formatrpsdb.html`. A opção *-n* define o nome da base criada.

Para classificar as proteínas em famílias deve-se executar o programa *rpsblast* fornecendo como entrada o arquivo *fasta* com as sequências de proteínas que se deseja classificar:

```
rpsblast -i <arquivo fasta> -p T -d Prk -o <arquivo de saída> -m 9
```

Execute o programa *rpsblast* no diretório da base criada, fornecendo o nome da base através da opção *-d*.

Referências Bibliográficas

[1] E. E. Allen and D. H. Bartlett. Structure and regulation of the omega-3 polyunsaturated fatty acid synthase genes from the deep-sea bacterium photobacterium profundum strain ss9. *Microbiology*, 148(Pt 6):1903–1913, Jun 2002.

[2] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990.

[3] G. S. Araújo and N. F. de Almeida Jr. Phylogeny from whole genome comparison. In A. L. C. Bazzan, editor, *1st Brazilian Workshop on Bioinformatics, WOB 2002, pp 9-15, Gramado RS, Brazil*, pages 9–15, 2002.

[4] A. Bergeron, J. Mixtacki, and J. Stoye. A unifying view of genome rearrangements. In *Algorithms in Bioinformatics, 6th International Workshop, WABI 2006, Zurich, Switzerland*, pages 163–173, 2006.

[5] Y. Boucher, C. J. Douady, R. T. Papke, D. A. Walsh, M. E. R. Boudreau, C. L. Nesbo, R. J. Case, and W. F. Doolittle. Lateral gene transfer and the origins of prokaryotic groups. *Annu Rev Genet*, 37:283–328, 2003.

[6] D. Bryant and V. Moulton. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol*, 21(2):255–265, Feb 2004.

[7] K. P. Byrne and K. H. Wolfe. Visualizing syntenic relationships among the hemiascomycetes with the Yeast Gene Order Browser. *Nucleic Acids Res*, 34(Database issue):D452–D455, Jan 2006.

[8] P. P. Côgo. Comparação de genomas completos de espécies da família vibrionacea empregando rearranjo de genomas. Master's thesis, IC-UNICAMP, 2008.

[9] F. Cohan. What are bacterial species? *Annu Rev Microbiol*, 56:457–487, 2002.

[10] F. Cohan. *Concepts of bacterial biodiversity for the age of genomics.*, chapter 11, pages 175–194. Springer-Verlag New York, LLC, 2004.

[11] R. R. Colwell. Polyphasic taxonomy of the genus vibrio: numerical taxonomy of vibrio cholerae, vibrio parahaemolyticus, and related vibrio species. *J Bacteriol*, 104(1):410–433, Oct 1970.

[12] A. C. E. Darling, B. Mau, F. R. Blattner, and N. T. Perna. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res*, 14(7):1394–1403, Jul 2004.

[13] D. E. Dykhuizen and L. Green. Recombination in escherichia coli and the definition of biological species. *J Bacteriol*, 173(22):7257–7268, Nov 1991.

[14] P. C. Feijão and J. Meidanis. A survey on genome rearrangement problems and gene order based phylogenies. Technical report, IC-UNICAMP, 2008.

[15] J. Felsenstein. *PHYLIP (PHYLogeny Inference Package) version 3.6a2.* Distributed by the author, Department of Genetics, University of Washington, Seattle, 1993.

[16] J. Felsenstein. *Inferring phylogenies.* Sinauer Associates, 2003.

[17] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H.-R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. L. Sonnhammer, and A. Bateman. The Pfam protein families database. *Nucleic Acids Res*, 36(Database issue):D281–D288, Jan 2008.

[18] W. M. Fitch. Distinguishing homologous from analogous proteins. *Syst Zool*, 19(2):99–113, Jun 1970.

[19] W. M. Fitch. Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology*, 20:406–416, 1971.

[20] W. M. Fitch. Homology a personal view on some of the problems. *Trends Genet*, 16(5):227–231, May 2000.

[21] S. T. Fitz-Gibbon and C. H. House. Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res*, 27(21):4218–4222, Nov 1999.

[22] D. A. Fitzpatrick, M. E. Logue, J. E. Stajich, and G. Butler. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evol Biol*, 6:99, 2006.

[23] K. Fukami-Kobayashi, Y. Minezaki, Y. Tateno, and K. Nishikawa. A tree of life based on protein domain organizations. *Mol Biol Evol*, 24(5):1181–1189, May 2007.

[24] S. Garcia-Vallve, E. Guzman, M. A. Montero, and A. Romeu. HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res*, 31(1):187–189, Jan 2003.

[25] A. Gattiker, E. Gasteiger, and A. Bairoch. ScanProsite: a reference implementation of a PROSITE scanning tool. *Appl Bioinformatics*, 1(2):107–108, 2002.

[26] D. Gevers, F. M. Cohan, J. G. Lawrence, B. G. Spratt, T. Coenye, E. J. Feil, E. Stackebrandt, Y. V. de Peer, P. Vandamme, F. L. Thompson, and J. Swings. Opinion: Re-evaluating prokaryotic species. *Nat Rev Microbiol*, 3(9):733–739, Sep 2005.

[27] J. Gómez-León, L. Villamil, M. L. Lemos, B. Novoa, and A. Figueras. Isolation of vibrio alginolyticus and vibrio splendidus from aquacultured carpet shell clam (ruditapes decussatus) larvae associated with mass mortalities. *Appl Environ Microbiol*, 71(1):98–104, Jan 2005.

[28] J. P. Gogarten, W. F. Doolittle, and J. G. Lawrence. Prokaryotic evolution in light of gene transfer. *Mol Biol Evol*, 19(12):2226–2238, Dec 2002.

[29] S. R. Henz, D. H. Huson, A. F. Auch, K. Nieselt-Struwe, and S. C. Schuster. Whole-genome prokaryotic phylogeny. *Bioinformatics*, 21(10):2329–2335, May 2005.

[30] S. Hunter, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, A. Laugraud, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, N. Mulder, D. Natale, C. Orengo, A. F. Quinn, J. D. Selengut, C. J. A. Sigrist, M. Thimma, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu, and C. Yeats. InterPro: the integrative protein signature database. *Nucleic Acids Res*, 37(Database issue):D211–D215, Jan 2009.

[31] D. H. Huson and D. Bryant. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol*, 23(2):254–267, Feb 2006.

[32] L.-W. Jiang, K.-L. Lin, and C. L. Lu. OGtree: a tool for creating genome trees of prokaryotes based on overlapping genes. *Nucleic Acids Res*, 36(Web Server issue):W475–W480, Jul 2008.

[33] S. Karlin and C. Burge. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet*, 11(7):283–290, Jul 1995.

[34] N. Khiripet. Bacterial whole genome phylogeny using proteome comparison and optimal reversal distance. In *Fourth International IEEE Computer Society Computational Systems Bioinformatics Conference Workshops Poster Abstracts (CSB 2005 Workshops), Stanford, CA, USA*, pages 63–64, 2005.

[35] W. Klimke, R. Agarwala, A. Badretdin, S. Chetvernin, S. Ciufo, B. Fedorov, B. Kiryutin, K. O'Neill, W. Resch, S. Resenchuk, S. Schafer, I. Tolstoy, and T. Tatusova. The National Center for Biotechnology Information's Protein Clusters Database. *Nucleic Acids Res*, 37(Database issue):D216–D223, Jan 2009.

[36] E. V. Koonin. Horizontal gene transfer: the path to maturity. *Mol Microbiol*, 50(3):725–727, Nov 2003.

[37] E. V. Koonin. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*, 39:309–338, 2005.

[38] E. V. Koonin, K. S. Makarova, and L. Aravind. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol*, 55:709–742, 2001.

[39] C. G. Kurland, B. Canback, and O. G. Berg. Horizontal gene transfer: a critical view. *Proc Natl Acad Sci U S A*, 100(17):9658–9662, Aug 2003.

[40] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947–2948, Nov 2007.

[41] T. Lima, A. H. Auchincloss, E. Coudert, G. Keller, K. Michoud, C. Rivoire, V. Bulliard, E. de Castro, C. Lachaize, D. Baratin, I. Phan, L. Bougueleret, and A. Bairoch. HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic Acids Res*, 37(Database issue):D471–D478, Jan 2009.

[42] W. Ludwig and H. Klenk. Overview: a phylogenetic backbone and taxonomic framework for procaryotic systematics. In *In Bergey's Manual of Systematics Bacteriology. Second Edition.*, pages 49–65. Springer-Verlag. Berlin., 2001.

[43] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*, 35(Database issue):D26–D31, Jan 2007.

[44] V. Makarenkov. T-REX: reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics*, 17(7):664–668, Jul 2001.

[45] K. Makino, K. Oshima, K. Kurokawa, K. Yokoyama, T. Uda, K. Tagomori, Y. Iijima, M. Najima, M. Nakano, A. Yamashita, Y. Kubota, S. Kimura, T. Yasunaga, T. Honda, H. Shinagawa, M. Hattori, and T. Iida. Genome sequence of vibrio parahaemolyticus: a pathogenic mechanism distinct from that of v cholerae. *Lancet*, 361(9359):743–749, Mar 2003.

[46] A. Marchler-Bauer, A. R. Panchenko, B. A. Shoemaker, P. A. Thiessen, L. Y. Geer, and S. H. Bryant. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res*, 30(1):281–283, Jan 2002.

[47] T. Meinel, A. Krause, H. Luz, M. Vingron, and E. Staub. The SYSTERS Protein Family Database in 2005. *Nucleic Acids Res*, 33(Database issue):D226–D229, Jan 2005.

[48] S. D. Miller, S. H. D. Haddock, C. D. Elvidge, and T. F. Lee. Detection of a bioluminescent milky sea from space. *Proc Natl Acad Sci U S A*, 102(40):14181–14184, Oct 2005.

[49] B. G. Mirkin, T. I. Fenner, M. Y. Galperin, and E. V. Koonin. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol*, 3:2, Jan 2003.

[50] F. Rohwer, V. Seguritan, F. Azam, and N. Knowlton. Diversity and distribution of coral-associated bacteria. *Marine ecology progress series*, 243:1–10, 2002.

[51] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425, Jul 1987.

[52] D. Sankoff. Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, 28(1):35–42, 1975.

[53] D. Sankoff. Genome rearrangement with gene families. *Bioinformatics*, 15(11):909–917, Nov 1999.

[54] D. Sankoff and N. El-Mabrouk. Genome rearrangement. In T. Jiang, T. Smith, Y. Xu, and M. Zhang, editors, *Current Topics in Computational Biology*, pages 135–155. MIT Press, 2002.

[55] J. Setubal and J. Meidanis. *Introduction to computational molecular biology.* PWS Publishing Company, 1997.

[56] R. R. Sokal and C. D. Michener. A quantitative approach to a problem of classification. *Evolution*, 11:130–162, 1957.

[57] E. L. L. Sonnhammer and E. V. Koonin. Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet*, 18(12):619–620, Dec 2002.

[58] K. Tamura, J. Dudley, M. Nei, and S. Kumar. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol*, 24(8):1596–1599, Aug 2007.

[59] J. Tang and B. M. E. Moret. Phylogenetic reconstruction from gene-rearrangement data with unequal gene content. In *Algorithms and Data Structures, 8th International Workshop, WADS 2003, Ottawa, Ontario, Canada*, pages 37–46, 2003.

[60] R. L. Tatusov, E. V. Koonin, and D. J. Lipman. A genomic perspective on protein families. *Science*, 278(5338):631–637, Oct 1997.

[61] B. J. S. Thompson, Fabiano; Austin, editor. *The Biology of Vibrios (1st Edition)*. American Society for Microbiology, 2006.

[62] C. C. Thompson, A. C. P. Vicente, R. C. Souza, A. T. R. Vasconcelos, T. Vesth, N. Alves, D. W. Ussery, T. Iida, and F. L. Thompson. Genomic taxonomy of vibrios. *BMC Evol Biol*, 9:258, 2009.

[63] F. L. Thompson, C. C. Thompson, S. Naser, B. Hoste, K. Vandemeulebroecke, C. Munn, D. Bourne, and J. Swings. Photobacterium rosenbergii sp. nov. and Enterovibrio coralii sp. nov., vibrios associated with coral bleaching. *Int J Syst Evol Microbiol*, 55(Pt 2):913–917, Mar 2005.

[64] P. Vandamme, B. Pot, M. Gillis, P. de Vos, K. Kersters, and J. Swings. Polyphasic taxonomy, a consensus approach to bacterial systematics. *Microbiol Rev*, 60(2):407–438, Jun 1996.

[65] A. J. Vilella, J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, and E. Birney. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research*, 19(2):327–335, February 2009.

[66] S. Yancopoulos, O. Attie, and R. Friedberg. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, 21(16):3340–3346, Aug 2005.