Comparação de Genomas Completos de Espécies da Família *Vibrionacea* Empregando Rearranjo de Genomas

Este exemplar corresponde à redação final da Dissertação devidamente corrigida e defendida por Patrícia Pilisson Côgo e aprovada pela Banca Examinadora.

Campinas, 28 de março de 2008.

João Meidanis (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: Maria Júlia Milani Rodrigues – CRB8a / 2116

Côgo, Patrícia Pilisson
C655c Comparação de genomas completos de espécies da família Vibrionacea empregando rearranjo de genomas / Patrícia Pilisson Côgo -- Campinas, [S.P. :s.n.], 2008.

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

1. Biologia computacional. 2. Genômica. 3. Filogenia. I. Meidanis, João. II. Universidade Estadual de Campinas. Instituto de Computação. III. Título.

Título em inglês: A rearrangement-based approach to compare whole genomes of Vibronacea strains.

Palavras-chave em inglês (Keywords): 1. Computational biology. 2. Genomics. 3. Phylogeny.

Área de concentração: Biologia Computacional

Titulação: Mestre em Ciência da Computação

Banca examinadora:

Prof. Dr. João Meidanis (IC-UNICAMP)
Prof. Dr. Maria Emília Machado Telles Walter (DCC-UnB)
Prof. Dr. Jacques Wainer (IC-UNICAMP)
Prof. Dr. Orlando Lee (IC-UNICAMP)

Data da defesa: 28/03/2008

Programa de Pós-Graduação: Mestrado em Ciência da Computação

TERMO DE APROVAÇÃO

Dissertação Defendida e Aprovada em 28 de março de 2008, pela Banca examinadora composta pelos Professores Doutores:

Profª. Drª. Maria Emília Machado Telles Walter
DCC / UnB.

Prof. Dr. Jacques Wainer
IC - UNICAMP.

Prof. Dr. João Meidanis
IC - UNICAMP.

Instituto de Computação
Universidade Estadual de Campinas

Comparação de Genomas Completos de Espécies da Família *Vibrionacea* Empregando Rearranjo de Genomas

Patrícia Pilisson Côgo[1]

Maio de 2008

Banca Examinadora:

- João Meidanis (Orientador)
- Maria Emília M. T. Walter
  Departamento de Ciência da Computação – UnB
- Jacques Wainer
  Instituto de Computação – Unicamp
- Orlando Lee (Suplente)
  Instituto de Computação – Unicamp

[1]**Essa dissertação contou com auxílio financeiro das agências CNPq e FAPESP.**

Resumo

A evolução das técnicas de seqüenciamento tornou possível a obtenção de uma enorme quantidade de dados genômicos. O desafio atual é analisar esses dados e construir novos conhecimentos a partir deles.

Neste contexto, um problema importante e ainda em aberto é a criação de métodos de análise taxonômica de genomas completos. Especialmente para organismos procariontes, para os quais ainda não há um conceito claro de espécie, a comparação de genomas completos pode significar uma importante contribuição.

Neste trabalho propomos uma metodologia para comparação de genomas completos baseada na teoria de Rearranjo de Genomas, aplicando-a a organismos da família *Vibrionaceae* – uma família hetereogênea 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.

As distâncias genômicas obtidas quando analisamos separadamente cada um dos dois cromossomos que compõem o genoma desses organismos estão de acordo com as árvores filogenéticas construídas empregando outros métodos de comparação genômica. Esse resultado corrobora nosso método e a utilização da teoria de rearranjo de genomas como uma alternativa para análise de genomas completos. Além disso, pode indicar que os eventos modelados neste trabalho, como perda de genes, transferências horizontais, reversões entre outros, exercem um papel importante na evolução desses organismos. Compreender a dinâmica desses eventos e combiná-la a outros métodos de análise genômica pode significar um grande avanço para a construção de uma filogenia mais acurada para estes vibriões.

Abstract

The evolution in genomic sequencing techniques has resulted in a large amount of genomic data. The challenge that arises from this scenario is to analyze these data and to extract from them relevant biologic information.

In this context, taxonomic analysis of complete genome sequences is still an open problem. Futhermore, it is critical for procaryotes, which still lack a clear definition of species, and whose taxonomic classification is in continuous evolution, where complete genomes comparison may well play a significant role.

In this work, we propose a methodology to compare complete genomes based on genome rearrangement theory. We have applied our method to organisms of *Vibrionaceae* family – a heterogeneous family that comprehends organisms from five different genera, including the agent responsible for cholera, a severe disease in developing countries.

The genomic distances obtained when we analysed each chromosome individually are in agreement with phylogenic trees built using other genomic methods. This result validates our method and the genomic rearrangement theory as an alternative to analyze complete genomes. It also can indicate the importance played by rearrangement events in the vibrio genomic evolution. The understanding of these events, combined with other genomic methods, can play an important role in the construction of a robust vibrio phylogeny.

Agradecimentos

Estas linhas são a oportunidade de expressar minha gratidão e meu reconhecimento às pessoas que estiveram ao meu lado durante a realização deste trabalho.

Agradeço à minha família pelo amor, pela compreensão, pela confiança. Mamãe, Papai, Márcia: vocês foram minha fonte de alegria e meu porto seguro durante todos esses anos longe de casa. Sem o carinho de vocês nada seria possível. Eu os amo muito!

Agradeço ao Marcelo pela companhia em todas as horas, pelo carinho, pela dedicação. Seu auxílio e seu apoio, por vezes pessoais, por vezes técnicos, foram muito importantes para mim.

Agradeço ao João, por tudo que me ensinou durante esses anos de mestrado. Agradeço pela confiança depositada em mim e no meu trabalho e até mesmo pelos puxões-de-orelha, dados na devida medida e nas horas certas.

Finalmente, agradeço aos professores, amigos e funcionários do Instituto de Computação. Aqui aprendi muitas coisas, conheci pessoas muito boas e fiz muitos amigos, que levarei no coração para toda vida.

Conteúdo

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

A evolução na velocidade e capacidade de processamento ocorrida na computação entre o ábaco e os modernos computadores possibilitou o aprofundamento do saber humano em diversas áreas. Hoje, graças aos avanços nas técnicas de seqüenciamento genético, já podemos responder, em termos de milhões de bases A-C-T-G's, à filosofal pergunta "Do que somos feitos?" E o desafio atual é extrair novos conhecimentos, sobre o passado e também sobre o futuro, a partir dessa imensa quantidade de informação.

Atualmente, 1826 projetos de seqüenciamento de genomas completos estão concluídos ou em fase de conclusão [1]. Desses, 1420 (78%) são bacterianos — organismos que são objeto de estudo neste trabalho — e métodos eficientes de análise desses genomas são ainda um problema em aberto e temas de diversas pesquisas.

A análise taxonômica bacteriana teve um grande impulso com o surgimento de técnicas de análise genômica. Entre as técnicas existentes, atualmente, as mais importantes consistem na utilização de um gene ou um conjunto de genes como marcadores filogenéticos. Como exemplos, podemos citar o RNA ribossomal 16S [35] e a técnica conhecida como MLSA (*Multilocus sequence analysis*) [23, 29], que emprega não somente um gene marcador, mas a concatenação de um conjunto de genes que se mantiveram conservados no genoma de diferentes organismos ao longo de sua evolução.

Os métodos de análise genômica representaram uma revolução na taxonomia desses organismos, anteriormente baseada exclusivamente em informações fenotípicas. Entretanto, esses métodos apresentam algumas limitações e em alguns casos, como na comparação entre as espécies *Vibrio splendidus* e *Vibrio harveyi*, as técnicas tradicionais de análise genômica podem não ser suficientemente conclusivas [29].

Na nossa opinião, um problema grave nos métodos de análise taxonômica baseados em apenas um gene ou em um pequeno conjunto de genes marcadores é estender para todo genoma a evolução medida a partir de uma pequena fração deste. A princípio, não

[1] Fonte: NCBI, 24 de novembro de 2007.

há garantias de que a extrapolação da evolução medida em uma pequena porção de genes para todo o genoma seja verdadeira, especialmente para organismos procariontes — nos quais um grande conjunto de eventos evolutivos, além da substituição de nucleotídeos, podem ocorrer.

Em particular, os métodos tradicionais não consideram eventos que alteram conteúdo e ordem dos genes, como: perda de genes, transferências horizontais e rearranjos cromossomiais, descritos em maior detalhe na seção 1.3 — que ocorrem muito frequentemente em bactérias e são fortes responsáveis pela imensa diversidade verificada entre esses organismos [7, 26, 27].

Neste trabalho propomos uma metodologia alternativa para complementar os resultados obtidos a partir de marcadores moleculares e análise fenotípica, baseada na teoria de rearranjo de genomas para análise de genomas completos. Aplicamos nosso método em genomas da família *Vibrionaceae*, empregando variações no conteúdo gênico e rearranjo cromossomiais para estimar distâncias evolutivas.

As árvores filogenéticas obtidas quando aplicamos nossa metodologia de comparação a cada um dos dois cromossomos que compõem o genoma dos vibriões coincidem com aquelas construídas através de outros métodos de análise filogenética baseados em informações genômicas.

A análise de genomas da família *Vibrionaceae* constituiu um teste para validação do método. O próximo passo será empregá-lo na comparação entre outros grupos de organismos, para os quais não há consenso na análise filogenética baseada em métodos genômicos e fenotípicos tradicionais.

1.1 A família *Vibrionaceae*

Vibrionaceae é uma heterogênea família proteobacteriana que, atualmente, compreende organismos de cinco diferentes gêneros: *Vibrio*, *Photobacterium*, *Salinivibrio*, *Enterovibrio* e *Grimontia*. Vibriões, i.e., organismos da família *Vibrionacea*, são gram-negativos, apresentam respiração aeróbica ou anaeróbica (facultativa), possuindo um ou mais flagelos. Esses organismos habitam águas doces ou salgadas, sendo frequentemente encontrados em associação com outros organismos.

Nesse trabalho estudamos os genomas de seis vibriões, são eles:

- *Vibrio cholerae* — Agente causador da cólera. Foi o primeiro vibrião a ser estudado, pelo anatomista Filippo Pacini, ainda no século XIX;
- *Vibrio fisheri* — Habitante de águas marinhas temperadas, apresenta bioluminescência. Não é, em geral, patogênico, sendo frequentemente encontrado em associação simbiótica com outros animais marinhos;

- *Vibrio parahaemolyticus* — Habitante de águas marinhas, pode contaminar mariscos ou outros animais marinhos que, quando ingeridos pelo homem, causam grave gastroenterite;
- *Vibrio vulnificus* — Habitante de águas marinhas, pode contaminar animais marinhos, principalmente ostras, que, quando ingeridas cruas ou mal-preparadas, causam séria infecção;
- *Photobacterium profundum* — Habitante de águas marinhas, apresenta bioluminescência e é capaz de sobreviver em ambientes de baixas temperaturas e alta pressão.

Normalmente, vibriões apresentam dois cromossomos: um maior, que contém genes relacionados ao crescimento e desenvolvimento desses organismos, e um outro menor, que apresenta genes relacionados a adaptação ambiental e virulência [33]. A tabela 1.1 apresenta algumas informações sobre os genomas dos organismos analisados.

	Cromossomo 1		Cromossomo 2	
Organismo	Comprimento (Mb)	Genes	Comprimento (Mb)	Genes
P. profundum	4,08	2.603	2,24	2.030
V. choleare	2,96	2.889	1,07	1.119
V. fisheri	2,91	2.716	1,33	1.186
V. parahaemolyticus	3,29	3.223	1,88	1.769
V. vulnificus CMCP6	3,28	3.049	1,84	1.575
V. vulnificus YJ016	3,35	3.387	1,88	1.711

Tabela 1.1: Características dos seis genomas de vibriões analisados neste trabalho.

A figura 1.1 mostra a topologia considerada correta para estes seis vibriões. Esta árvore foi cosntruída a partir das seqüências do gene 16S rRNA, utilizando o método *neighbor-joining*, disponível na ferramenta MEGA 4 [31]. Como pode ser observado, as espécies mais próximas são *Vibrio vulnificus* e *Vibrio parahaemolyticus*, seguidos do *Vibrio cholerae*. A espécie *Vibrio fisheri* localiza-se entre o *Photobacterium profundum* e os demais vibriões. Outras árvores filogenéticas, que abrangem mais organismos da família *Vibrionacea*, baseadas em diversos genes marcadores podem ser encontradas nos trabalhos de Thompson e colegas [33, 29]. Em geral, as restrições dessas árovores para os organismos tratados neste trabalho coincidem com a topologia mostrada na figura 1.1.

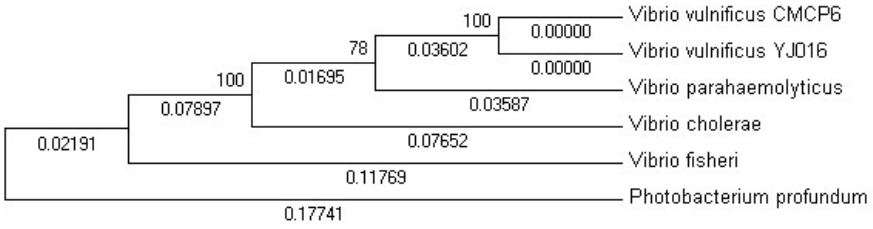

Figura 1.1: Árvore filogenética entre os vibriões analisadas baseada no método *neighbor-joining*, empregando o gene 16S rRNA. Em cada nó interno são mostrados os valores de *bootstrap*, ou seja, o percentual de vezes que o agrupamento foi reconstruído em 1.000 simulações distintas.

1.2 Evolução de Genomas Bacterianos

A partir da década de 70, métodos de hibridização de fitas de DNA têm complementado, ou substituído, análises fenotípicas para classificação de espécies bacterianas. De maneira prática, podemos definir uma espécie bacteriana como um conjunto de organismos que apresentam consistência fenotípica, grau relevante de hibridização de DNA (50-70%) e similaridade superior a 97% no RNA ribossomal 16S [11]. Entretanto, essa é uma regra empírica e novos métodos de análise, especialmente tratando de genomas completos, podem complementar e promover novas oportunidades na classificação taxonômica desses organismos.

Comparando diferentes genomas bacterianos, pode-se afirmar que não só substituição de nucleotídeos, mas também eventos mutacionais que envolvem grandes blocos do genoma ocorrem e direcionam a evolução desses organismos [11]. Dentre esses eventos, destacamos duplicações, transferências horizontais, perdas de genes e rearranjos cromossomiais – eventos que, individualmente ou em conjunto, motivaram diversas pesquisas [15, 26, 27, 17] na área de comparação genômica, os quais pretendemos abordar neste trabalho.

A ocorrência de *duplicações* facilita a adaptação dos organismos a ambiente hostis, como ocorre com o vibrião *Photobacterium profundum*, adaptado ao ambiente de alta pressão das profundidades oceânicas. Quando sucedidas por alterações na função gênica, duplicações são mecanismos muito importantes para adaptação às mudanças ambientais e para exploração de novos nichos ecológicos.

Outro mecanismo de adaptação a novos ambientes é chamado THG — *transferência horizontal de genes*. Este evento consiste na troca de material genético entre organismos que compartilham um mesmo *habitat*, mas que não estão na mesma linha de descendência.

Há muitas evidências de ocorrências de THG entre organismos procariontes. Entretanto, é muito complexo encontrar provas irrefutáveis da ocorrência de transferências horizontais, uma vez que não é trivial diferenciá-las de grandes mutações que alteram significativamente a funcionalidade gênica. Pesquisas relacionadas a este tema, usualmente, atribuem a THGs a aquisição de genes relacionados a adaptação ambiental e a virulência [11].

Além de duplicações e transferências horizontais, que aumentam o número de funções expressas em um genoma, podem ocorrer mutações que reduzem o número de genes. Chamamos de *perda de genes* o descarte de genes que perderam sua função ou cuja presença seja prejudicial a função de outro gene.

Além de eventos que alteram o conteúdo genômico, fenômenos que alteram a organização do genoma também ocorrem muito frequentemente em bactérias, como mostra a figura 1.2, onde são comparados cromossomos homólogos pertencentes ao *Vibrio cholerae* e ao *Vibrio parahaemolyticus*. Embora não haja indícios que a alteração na posição de um gene influencie sua função, o número de eventos de rearranjos cromossomiais pode fornecer uma forte estimativa do grau de divergência entre dois genomas ao longo de sua evolução.

Neste trabalho, propomos uma associação entre o número de eventos de perda e ganho de genes e as distâncias de rearranjo como uma maneira alternativa para estimar distâncias evolutivas entre genomas completos, aplicando-a aos genomas dos vibriões *Vibrio cholerae*, *Vibrio parahaemolyticus*, *Vibrio fisheri*, *Vibrio vulnificus* CMCP6 e YJ016 e *Photobacterium profundum*. Esses seis genomas foram escolhidos por estarem publicamente disponíveis no início deste trabalho.

1.3 Rearranjo de Genomas

Uma maneira de comparar grandes blocos genômicos é através da ordem relativa e da orientação de seus genes. A mudança na posição relativa ou na orientação de um gene entre os genomas sendo comparados é denominado *evento de rearranjo*. O mínimo número de eventos dessa natureza capazes de transformar um genoma em outro é chamado *distância de rearranjo*. A *Teoria de Rearranjo de Genomas* se preocupa em analisar problemas envolvendo eventos de rearranjo e definir algoritmos eficientes para estimar suas distâncias.

Dentre os eventos de rearranjo descritos na literatura, destacamos: reversões, translocações, *block-interchanges*, transposições, fissões, fusões, além de combinações dessas operações, que descreveremos nas próximas seções.

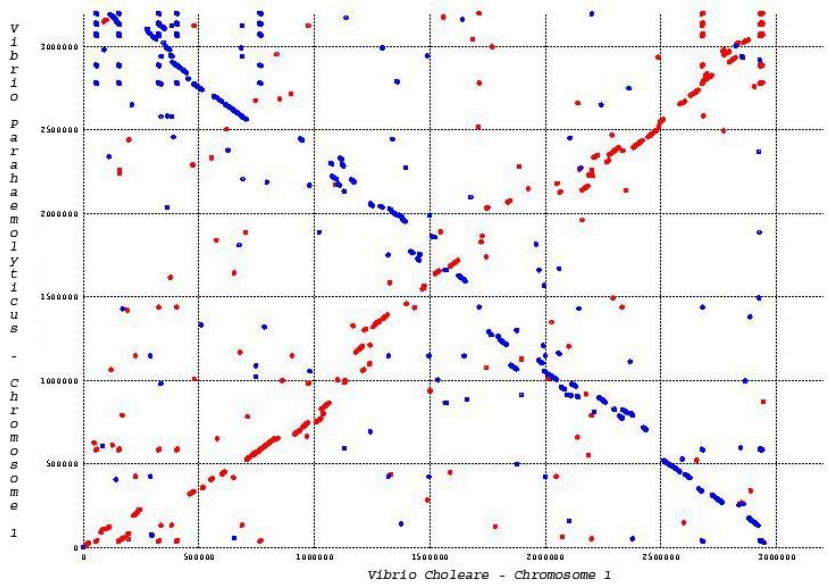

Figura 1.2: Comparação entre cromossomos homólogos do *Vibrio cholerae* (eixo horizontal) e do *Vibrio parahaemolyticus* (eixo vertical). Pontos representam regiões de similaridade. Observa-se a ocorrência de uma grande reversão, além de outros eventos menores, na ordem em que os genes se apresentam nos cromossomos desses dois organismos.

1.3.1 Reversão

O evento de reversão altera a ordem relativa de um gene ou um conjunto de genes consecutivos em um genoma, conforme mostra a figura 1.3. Quando a ordenação dos genes é conhecida, o problema de ordenação por reversões possui solução polinomial. Bader e colegas encontram uma solução linear para o cálculo da distância de reversão [3]. Quando se deseja conhecer a seqüência de reversões capaz de transformar um genoma em outro, não somente o número dessas operações, a melhor solução para o problema possui complexidade $O(n^2)$ [21].

1.3.2 Translocação

O evento de translocação troca de posição dois genes, ou dois conjuntos de genes consecutivos, pertencentes a cromossomos distintos, conforme figura 1.4.

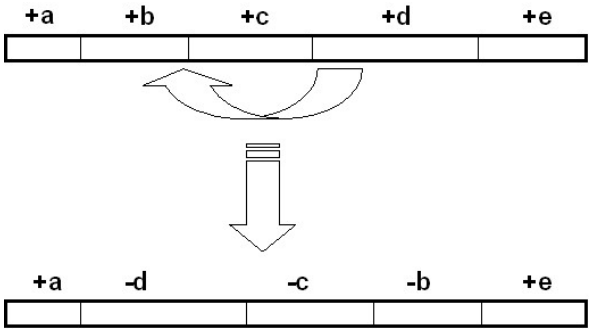

Figura 1.3: Exemplo de reversão.

1.3.3 Transposição

O evento de transposição consiste na troca de posição entre dois blocos genômicos adjacentes, como ilustrado na figura 1.5. Não existe algoritmo polinomial para solução do problema da distância de transposição. Algoritmos de aproximação para o problema foram apresentados por Bafna e Pevzner [4], Christie [10] e Walter *et. al* [34].

1.3.4 *Block-Interchange*

O evento de *block-interchange*, introduzido por Christie [9], consiste na troca de posição entre dois blocos genômicos, não necessariamente adjacentes, conforme figura 1.6. Portanto, essa operação pode ser considerada uma generalização da operação de transposição. Nesse mesmo trabalho, Christie mostrou que o problema de ordenação por *block-interchange* pode ser resolvido em $O(n^2)$.

1.3.5 Fissão e Fusão

A operação de fissão quebra um cromossomo em dois novos cromossomos e uma fusão atua sobre dois cromossomos, unindo-os, conforme mostrado na figura 1.7 (A) e (B), respectivamente. Dias e Meidanis [25], analisaram o problema de ordenação envolvendo fissões, fusões e transposições.

1.3.6 Combinando Diferentes Operações

Nos últimos anos, muitas pesquisas têm sido realizadas com o objetivo de obter distâncias de rearranjo mais realistas e com maior significado biológico. Nesse sentido, surgem

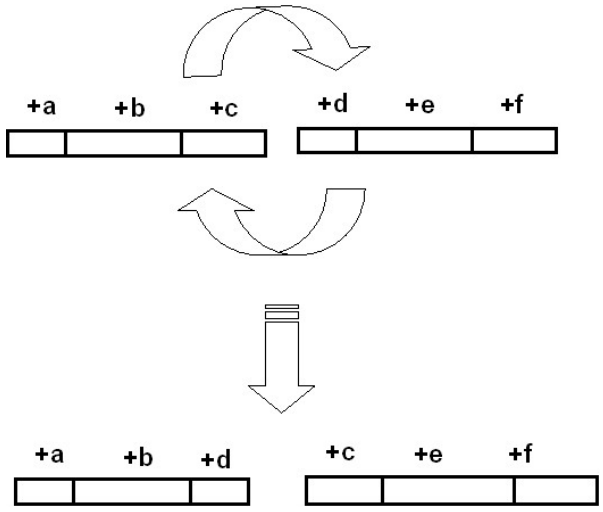

Figura 1.4: Exemplo de translocação.

trabalhos de abordam o problema considerando não somente uma única operação de rearranjo, mas um conjunto de operações, como os trabalhos de Dias e Meidanis envolvendo fissões, fusões e transposição [25] e o de Hannenhali e Pevzner, onde o problema da distância de rearranjo é estudado através de inversões, translocações, fissões e fusões [19].

Yancopoulos *et. al* introduziram uma operação chamada *double-cut-and-join* que, basicamente, consiste na quebra do genoma em dois pontos e na reconexão dos quatro terminais em uma maneira de diferente da original [36]. Dependendo da maneira como essa reconexão é feita, o resultado da operação pode ser uma reversão ou uma translocação (incluindo fissão e fusão como casos especiais). Duas operações sucessivas podem resultar em uma operação de *block-interchange*. Neste trabalho, utilizamos a solução descrita por Bergeron *et. al* [6] para estimar a distância de rearranjo entre dois genomas multicromossomiais utilizando a operação de *double-cut-and-join*.

1.3.7 Aplicações

Diferentes estudos foram realizados envolvendo aplicações práticas da teoria de rearranjo de genomas. Como exemplos, podemos citar Bafna e Pevzner [4], onde o cromossomo X de diferentes mamíferos é analisado, Hannenhalli *et al.* [17], onde são comparados vírus causadores da herpes, Belda *et al.* [5] que estudaram γ-proteobactérias utilizando diferentes técnicas, incluindo rearranjo de genomas, e Lin *et al.* [22], onde a operação de *block-interchange* foi empregada para comparação de espécies de vibriões.

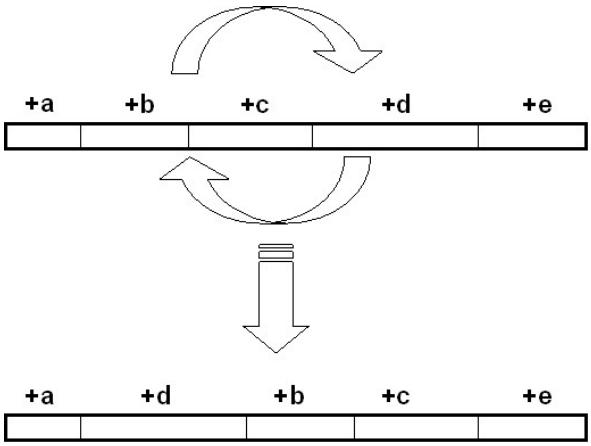

Figura 1.5: Exemplo de transposição.

A maioria das pesquisas envolvendo rearranjo de genomas consideram que o problema de determinação de homologia, *i.e.*, identificação dos genes que tiveram a mesma origem no genoma ancestral, já está previamente resolvido ou dedicam apenas um pequeno esforço em sua solução [28]. Entretanto, neste trabalho pretendemos utilizar uma abordagem diferenciada para solução desse problema, empregando as informações de famílias gênicas do projeto HAMAP [14] e desenvolvendo, para as proteínas de vibriões ainda não classificadas por esse projeto, nosso conjunto complementar de perfis [16] para descrever famílias de proteínas homólogas. Nosso método será apresentado em maior profundidade adiante.

Nos próximos capítulos descreveremos em maiores detalhes a metodologia empregada neste trabalho e os resultados obtidos durante sua realização. No capítulo 2 descreveremos a comparação de genomas completos de vibriões empregando uma metodologia muito simples e analisaremos os problemas encontrados nesses experimentos. No capítulo 3, apresentaremos nossa nova metodologia de comparação, incluindo soluções para os problemas de identificação de homologia, tratamento de duplicações e para o cálculo das distâncias de rearranjo. Os resultados obtidos quando aplicamos nosso método para comparar genomas completos de vibriões serão apresentados no capítulo 4. Finalmente, o capítulo 5 apresenta nossas conclusões e perspectivas de futuros trabalhos.

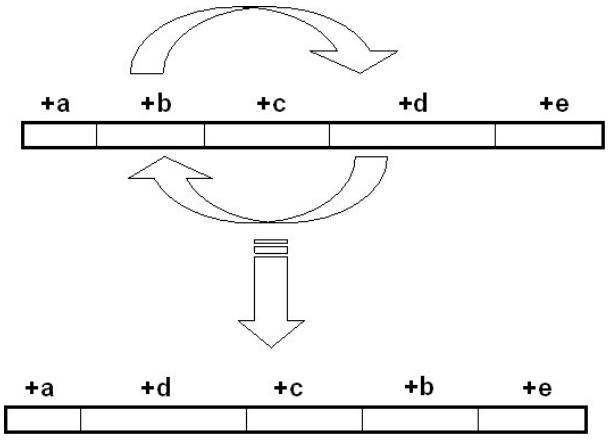

Figura 1.6: Exemplo de *block-interchange*.

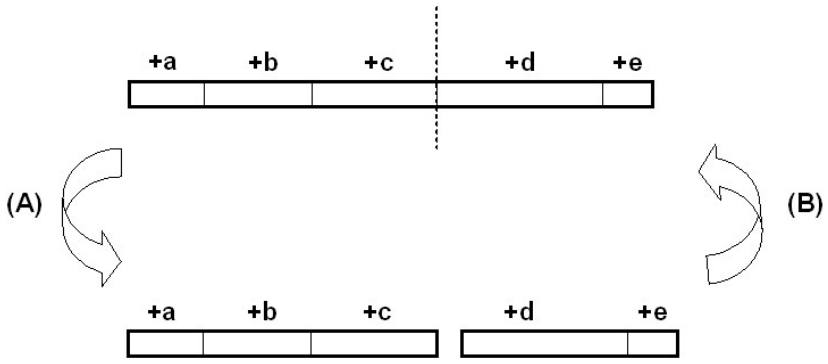

Figura 1.7: Exemplos de fissão e fusão.

Capítulo 2

Comparando Genomas

Neste capítulo descreveremos nossos experimentos iniciais de comparação de genomas completos de vibriões empregando rearranjo de genomas.

Nesses experimentos empregamos uma metodologia simples, que não nos permitiu reproduzir os resultados obtidos através da utilização de outros métodos de análise genômica. Contudo, a análise desses resultados nos levou a observar algumas deficiências nas técnicas que utilizamos e a identificar algumas propriedades importantes para a definição de uma nova metodologia de comparação.

2.1 Descrição da metodologia

Tradicionalmente, quando se deseja estimar evolução em bioinformática, existem 3 etapas fundamentais a serem seguidas:

1. Definir o modelo de comparação;
2. Identificar estruturas homólogas entre os organismos comparados;
3. Com base no modelo definido 1 e nas estruturas identificadas em 2, calcular o cenário mais parcimonioso capaz de transformar o genoma de uma espécie no de outra.

Nestes experimentos iniciais, utilizamos modelos descritos na literatura para identificação de estruturas homólogas e para o cálculo de distâncias de rearranjo.

Para identificar e agrupar genes homólogos, empregamos o método descrito por McLysaght *et al.* [24] em seu trabalho sobre a evolução de poxvírus. Em resumo, essa técnica de identificação de homologia consiste em comparar todo o conjunto de proteínas em análise, empregando o programa BLASTP (a versão utilizada nestes experimentos foi a 2.213) [2], classificando duas proteínas como homólogas se o resultado do alinhamento entre elas apresentar *e*-value menor ou igual a 10^{-5} e cobertura superior a 40% da maior proteína.

A figura 2.1 mostra uma representação gráfica desse procedimento. Nessa ilustração, cada vértice representa uma proteína. Proteínas que satisfazem o critério de similaridade definido são conectadas através de arestas.

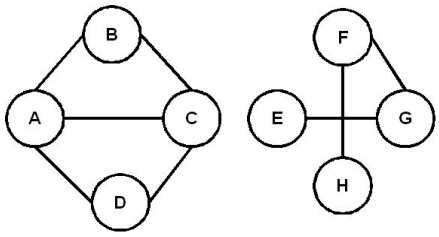

Figura 2.1: Exemplo de identificação de homologia baseada no grau de similaridade entre seqüências.

Após a comparação, uma variação do método de agrupamento *single-linkage* (ver seção 3.1.1) foi empregada para agrupar genes homólogos. Nesse método, um agrupamento é mantido se apresentar ao menos um membro que seja homólogo a todos os demais. Grupos de proteínas que não satisfaçam esse critério são desfeitos e seus membros tratados como *singletons*, ou seja, proteínas que não são homólogas a nenhuma outra. No exemplo da figura 2.1 o agrupamento {A,B,C,D} será mantido, visto que o vértice A está conectado a todos os demais. Por sua vez, o agrupamento {E,F,G,H} será desfeito.

Para tratamento de duplicações, mantivemos aleatoriamente uma única cópia de cada gene e descartamos todas as demais. De maneira semelhante, genes que não estavam presentes nos dois genomas comparados foram descartados. Após esse tratamento, ambos os genomas apresentavam o mesmo conteúdo e não havia duplicações.

As distâncias de rearranjo foram estimadas empregando a operação de *block-interchange*. A orientação apresentada por um mesmo gene em cada um dos genomas foi ignorada, uma vez que a operação de *block-interchange* não considera essa informação. Esse procedimento é similar ao que foi utilizado por Lin *et al.* na comparação entre três genomas completos de vibriões [22].

Os organismos analisados foram: *Vibrio cholerae O395*, *Vibrio fischeri ES114*, *Vibrio parahaemolyticus*, *Vibrio vulnificus CMCP6*, *Vibrio vulnificus YJ016* e *Photobacterium profundum SS9*. Esses genomas estão disponiveis no endereço:

ftp://ftp.ncbi.nih.gov/genomes/Bacteria[1].

Nas próximas seções descreveremos três experimentos empregando esse método, cada um deles resultando em uma ou mais matrizes de distâncias, dependendo dos parâmetros

[1]Último acesso em 08/04/20058

empregados e dos pesos atribuídos a cada operação (*block-interchange* ou perda de genes). Como veremos a seguir, somente quando consideramos a variação no conteúdo, e não somente a ordem relativa na qual os genes se apresentam em cada genoma, obtivemos distâncias biologicamente significativas.

2.2 Distâncias de *Block-Interchange*

As distâncias de *Block-Interchange* entre seis genomas de vibriões é mostrada na tabela 2.1. Como nenhum tipo de normalização foi empregada, as distâncias obtidas tendem a ser maiores entre organismos que possuem um maior número de genes em comum. Isso explica a grande distância entre os dois genomas da espécie *V. vulnificus*, para os quais esperávamos encontrar a menor distância.

	P. p.	*V. c.*	*V. f.*	*V. p.*	*V. v.* CMCP6	*V. v.* YJ016
P. profundum	0	780	808	877	1058	837
V. choleare		0	721	848	1089	881
V. fisheri			0	820	1011	802
V. parahaemolyticus				0	1276	868
V. vulnificus CMCP6					0	1291
V. vulnificus YJ016						0

Tabela 2.1: Matriz de distâncias empregando somente a operação de *block-interchange*.

2.3 Distâncias de *Block-Interchange* e *Perda de Genes*

Para solucionar o problema detectado no primeiro experimento, procuramos melhorar o modelo de estimação das distâncias evolutivas. Neste segundo experimento, calculamos as distâncias de rearranjo como a soma entre a distância de *block-interchange* mostrada na tabela 2.1 e o número de genes descartados por não possuírem um correspondente em outro genoma ou estarem duplicados. A matriz resultante é mostrada na tabela 2.2. Como podemos observar, os dois genomas do *V. vulnificus* apresentam nessa situação as menores distâncias.

Entretanto, quando modificamos os pesos atribuídos a cada operação (duplicando o peso para *block-interchange*, uma vez que essa operação é mais complexa), os problemas identificados no experimento anterior voltam a ocorrer, como mostra a tabela 2.3. Isso pode indicar que o conjunto de operações que utilizamos nesse experimento não é adequado para o problema que estamos tratando.

	P. p.	*V. c.*	*V. f.*	*V. p.*	*V. v.* CMCP6	*V. v.* YJ016
P. profundum	0	5531	6056	6740	6721	6674
V. choleare		0	4706	5302	5045	5254
V. fisheri			0	5293	5198	5436
V. parahaemolyticus				0	5490	5487
V. vulnificus CMCP6					0	4590
V. vulnificus YJ016						0

Tabela 2.2: Matriz de distâncias empregando as operações de *block-interchange* e *perda de genes*.

	P. p.	*V. c.*	*V. f.*	*V. p.*	*V. v.* CMCP6	*V. v.* YJ016
P. profundum	0	6249	6864	7617	7759	7511
V. choleare		0	5527	6150	6134	6135
V. fisheri			0	6113	6209	6238
V. parahaemolyticus				0	6766	6349
V. vulnificus CMCP6					0	6781
V. vulnificus YJ016						0

Tabela 2.3: Distâncias obtidas aplicando-se peso 2 à operação de *block-interchange* e 1 à *perda de genes*

2.4 Eliminando todas as duplicações

Este experimento é similar ao experimento anterior, exceto por descartamos todos os genes duplicados, ao invés de mantermos uma única cópia aleatoriamente. Novamente, dois conjuntos de pesos foram testados: aplicando peso unitário a todas as operações e duplicando o peso das operações de *block-interchange*. As distâncias obtidas para cada esquema de pesos são mostradas nas tabelas 2.4 e 2.5, respectivamente.

Apesar da simplicidade deste modelo, que não considera operações importantes como reversões e translocações, a árvore filogenética construída a partir da matriz 2.4 (aplicando peso unitário a todas as operações), mostrada na figura 2.2, apresenta algumas concordâncias com a árvore obtida através da análise de genes marcadores. Comparando-a com a árvore obtida empregando o gene 16S rRNA, ilustrada na figura 1.1, podemos observar que a relação de proximidade entre os dois genomas do *Vibrio vulnificus* e distância do *Photobacterium profundum* em relação aos demais organismos é mantida. Entretanto, esse

	P. p.	*V. c.*	*V. f.*	*V. p.*	*V. v.* CMCP6	*V. v.* YJ016
P. profundum	0	6221	6382	7065	7017	6246
V. choleare		0	4915	5548	5243	5411
V. fisheri			0	5570	5489	5716
V. parahaemolyticus				0	5750	5729
V. vulnificus CMCP6					0	4835
V. vulnificus YJ016						0

Tabela 2.4: Matriz de distâncias usando as operações de *block-interchange* e perda de genes, removendo todos os genes duplicados.

	P. p.	*V. c.*	*V. f.*	*V. p.*	*V. v.* CMCP6	*V. v.* YJ016
P. profundum	0	6939	7018	7853	7977	8002
V. choleare		0	5589	6258	6266	6247
V. fisheri			0	6323	6409	6506
V. parahaemolyticus				0	6938	6535
V. vulnificus CMCP6					0	6353
V. vulnificus YJ016						0

Tabela 2.5: Matriz de distâncias obtida com a remoção de genes duplicados e dobrando o peso da operação de *block-interchange*.

método não é capaz de classificar corretamente os genomas dos vibriões *Vibrio choleare*, *Vibrio parahaemolyticus* e *Vibrio fisheri*.

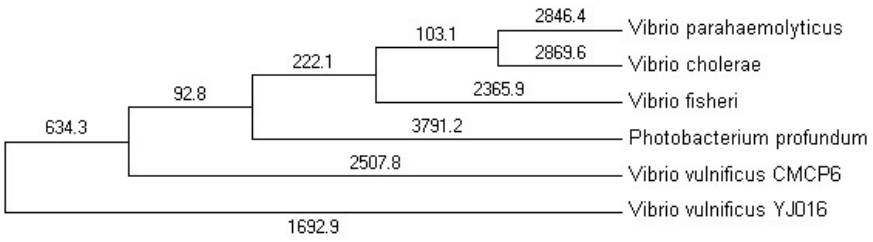

Figura 2.2: Árvore filogenética obtida a partir da distância de *block-interchange* e *perda de genes*, removendo todos os genes duplicados.

2.5 Problemas encontrados

Em nossos experimentos iniciais de comparação genômica de vibriões não conseguimos obter distâncias de rearranjo equivalentes às distâncias evolutivas obtidas utilizando outros métodos de comparação genômica. Isso nos levou a refletir sobre a metodologia que empregamos e a detectar algumas falhas.

Os primeiros problemas que identificamos foram relacionados ao método de classificação de homologia. Observamos que algumas famílias agrupavam proteínas bastante diferentes, pois bastava que apenas um membro fosse homólogo a todos os demais para que o agrupamento fosse mantido. Por outro lado, proteínas muito similares poderiam ser separadas e tratadas como *singletons*, pois todo o agrupamento seria desfeito se não houvesse ao menos um elemento homólogo universal.

Além disso, o conjunto de famílias construídas é dependente dos organismos analisados: se excluíssemos um dos vibriões, ou incluíssemos um novo organismo, as relações de homologia identificadas poderiam não ser mantidas.

Outros problemas estão relacionados ao modelo empregado para cálculo de distâncias de rearranjo. Em nosso primeiro experimento, descrito na seção 2.2, as distâncias de *block-interchange* divergem muito das obtidas quando são empregados outros métodos de comparação. Isso sugere que o número de genes que foram adquiridos, através de transferências horizontais, ou descartados ao longo da evolução desses organismos é significativo e estes genes devem ser considerados no cálculo de distâncias.

Um outro problema grave foi não considerarmos a orientação dos genes. Como mostra a figura 1.2, na qual são comparados cromossomos homólogos dos vibriões *V. cholerae* e *V. parahaemolyticus*, reversões são eventos comuns no processo de evolução desses organismos e não considerá-las pode ter sido a causa das distorções observadas em nossas estimativas.

Os questionamentos levantados durante a execução e análise desses experimentos levaram-nos a definir algumas propriedades que métodos mais robustos para identificação

de homologia, tratamento de duplicações e cálculo de distâncias devem satisfazer. A partir dessas considerações, começamos a construir uma metodologia mais robusta de comparação, que será apresentada ao longo do próximo capítulo.

Capítulo 3

Uma nova metodologia de comparação

Em nossos experimentos iniciais identificamos alguns problemas e definimos algumas propriedades que devem ser seguidas por uma metodologia robusta de comparação de genomas. Neste capítulo, apresentaremos nossas soluções para o problema de identificação de homologia e para o cálculo de distâncias de rearranjo.

3.1 Homologia

Segundo Fitch, em seu trabalho "*Homology: a personal view on some of the problems*" [13], podemos definir homologia entre dois ou mais genes como a origem em um mesmo gene ancestral. As relações de homologia podem ser classificadas em três diferentes sub-tipos:

- **Ortologia** – Os dois genes tiveram origem em uma mesma estrutura no genoma ancestral e a divergência entre eles seguiu um processo de *especiação*;
- **Paralogia** – Os dois genes homólogos tiveram origem em um mesmo genoma e a divergência entre eles seguiu um processo de *duplicação*;
- **Xenologia** – Os dois genes estão envolvidos em um processo de *transferência horizontal*;

Detecção de homologia em seus três sub-tipos é um problema crucial em muitas aplicações biológicas, incluindo genômica comparativa, análise taxonômica, inferência de estruturas de proteínas, entre outras. Por essa razão, diferentes métodos de identificação de genes homólogos foram desenvolvidos, variando entre técnicas simples de comparação de seqüências a complexos algoritmos de clusterização.

Apesar do grande número de pesquisas que se dedicam a esse tópico, não há concordância sobre qual o método mais adequado para identificação de homologia. Enquanto métodos simples, como alinhamento de seqüências, são suficientemente eficazes na análise de famílias bem conservadas [32], não há ainda uma metodologia eficiente para identificação de homólogos muito distantes.

Por essa razão, cada estudo envolvendo identificação de estruturas homólogas define metodologia e regras específicas [24, 12, 17], que podem não ser adequadas quando aplicadas a outro contexto.

Baseados em observações e em nossos experimentos preliminares, definimos algumas propriedades que um método de identificação de homologia eficiente e adequado ao nosso problema deve satisfazer, as quais não são necessariamente obedecidas por todos os métodos de detecção de homologia descritos na literatura.

3.1.1 Critérios para identificação de estruturas homólogas

A precisão das estimativas de evolução baseadas na comparação entre o conteúdo e a ordem na qual os genes aparecem em um genoma é, por natureza, fortemente dependente do método de identificação de estruturas homólogas empregado. Com o objetivo de definir uma estratégia robusta e acurada, estabelecemos duas propriedades que devem ser obedecidas.

A primeira delas é ***transitividade***, i.e., se o gene A é homólogo ao gene B e B é homólogo a C, então, A deve ser homólogo a C. Essa é uma regra natural das relações de homologia: se os pares {A,B} e {B,C} tiveram origem em um ancestral comum, deduze-se que também {A,C} a tiveram.

Um método largamente empregado para detecção de homologia é o grau de similaridade entre as seqüências sendo analisadas. Assim, diz-se que duas seqüências são homólogas se o resultado do alinhamento entre elas for superior a um limiar pré-estabelecido. Esse método pode violar a regra de transitividade, conforme exemplifica a figura 3.1.

Nesse exemplo, cada letra representa um gene e arestas conectam genes cuja similaridade está acima do limiar definido para identificação de homologia. Neste caso, {A,B} e {B,C} são considerados homólogos, mas a homologia entre {A,C} não é detectada, pois esses dois genes não são suficientemente similares.

Usualmente a falha desse método é solucionada através da utilização de métodos de agrupamento [15]. Neste exemplo, a aplicação de um método *single-linkage* (i.e, no qual cada membro de um grupo é similar ao menos a um dos demais) resulta em duas famílias gênicas {A, B, C, D, E} e {F, G, H}.

Apesar de eliminar a falha de transitividade, essa solução pode gerar novos problemas. No exemplo da figura 3.1, {A,B,C} são membros de uma mesma família. Entretanto, se B

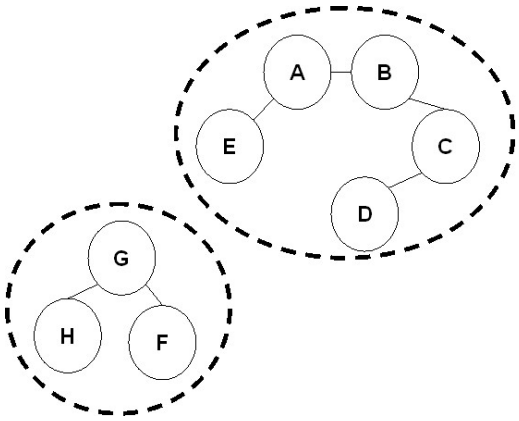

Figura 3.1: Identificação de homologia baseada em similaridade entre seqüências. As curvas identificam as famílias construídas a partir da aplicação de método *single-linkage*.

não fosse incluído na análise, A e C deixariam de ser considerados homólogos. Portanto, a classificação entre os genes A e C é dependente da presença do gene B, o que não é desejável.

Para evitar problemas como esse, a segunda regra que nosso método de detecção de homologia deve satisfazer é a ***independência do conjunto de genes analisados***. De acordo com essa regra, nossa identificação de homologia não pode ser baseada em comparação direta entre as seqüências analisadas, mas sim em um critério externo que associe um gene a uma família, independentemente do resultado da classificação dos demais genes analisados.

Em resumo, nosso método de identificação de estruturas homólogas deve satisfazer às seguintes regras:

1. Transitividade;

2. Independência do conjunto de genes analisados.

Seguindo esses critérios, adotamos um método de identificação de homologia baseado na informação *a priori* de famílias de proteínas disponibilizada pelo projeto HAMAP — *Automatic annotation of microbial proteomes in Swiss-Prot* [14].

O sistema HAMAP é baseado em descritores construídos manualmente para representar famílias de proteínas bem conservadas. Esses descritores são capazes de identificar, automaticamente e com alta confiabilidade, proteínas que pertençam à família representada e criar anotações semi-automáticas confiáveis. Atualmente, existem 1.445 famílias descritas pelo HAMAP. Esse conjunto de famílias é capaz de classificar 159.275 seqüências distintas [1].

Infelizmente, um grande número de genes contemplados neste trabalho ainda não são descritos pelo conjunto de famílias HAMAP. Com o objetivo de expandir essa classificação e cobrir totalmente os genomas analisados, criamos um conjunto próprio de perfis descritores de famílias de proteínas. Descrevemos nossa metodologia de construção desse conjunto de perfis nas próximas seções.

3.1.2 Perfis

Perfis são matrizes ponderadas que descrevem famílias de proteínas homólogas. Essas matrizes são construídas a partir do alinhamento múltiplo entre membros da família descrita e constituem um método eficiente de identificação de homologia, mesmo para seqüências distantes. Afirmamos que uma seqüência é representada por um perfil, ou seja, pertence à família descrita por esse perfil, quando o resultado do alinhamento entre eles for superior a um limiar pré-estabelecido [16, 30].

Esse método pode produzir falsos positivos ou falsos negativos, como pode ocorrer em quaisquer outros métodos de identificação de genes homólogos, mas satisfaz os critérios definidos na seção 3.1.1, pois:

- Todas as seqüências representadas pelo mesmo perfil irão pertencer à mesma família, satisfazendo o critério de transitividade;
- O resultado do alinhamento entre uma seqüência e um perfil é independente de qualquer outra seqüência analisada.

3.1.3 Identificando homologia entre os genomas de vibriões

O primeiro passo no nosso método de identificação de estruturas homólogas entre os genomas dos seis vibriões foi submetê-los, na forma de 27.282 proteínas, ao servidor HAMAP[1]. Essa consulta resultou na classificação de 3.217 seqüências, que correspondem a apenas 12% do total. A fim de identificar homologia entre as seqüências restantes, cobrindo completamente todos os genomas analisados, complementamos as informações do HAMAP construindo nosso próprio conjunto de perfis.

O processo de construção de um perfil se inicia com a seleção de seqüências pertencentes a família a qual se deseja descrever e geração de um alinhamento múltiplo entre elas. Nos projetos PROSITE e HAMAP [30, 14], essa escolha é feita por especialistas, através de uma seleção manual criteriosa. Em função do volume de famílias que precisamos representar, neste trabalho, uma análise tão cuidadosa não foi possível, sendo adotado um método automático de seleção de seqüências semelhantes.

[1]Data de submissão: 31/08/2006.

Nossa estratégia foi comparar, organismo a organismo, todas as proteínas analisadas às seqüências do banco de proteínas **NR**, um dos mais completos bancos de seqüências genéticas, mantido pelo NCBI (*U. S. National Center for Biology Information*). Para essa comparação, empregamos a ferramenta BLASTP [2], aplicando o seguinte critério para seleção das seqüências que serão empregadas para construção de nossos perfis:

- Se a proteína possuir 120 aminoácidos ou mais:
  - Selecione os 50 melhores alinhamentos com *e-value* $\leq 10^{-50}$ e cujo comprimento não difira em mais de 30% da proteína comparada;
- Se a proteína possuir menos de 120 aminoácidos:
  - Selecione os 50 melhores alinhamentos com *e-value* $\leq 10^{-3}$.

As seqüências selecionadas para cada proteína que não está representada pelas famílias HAMAP são alinhadas utilizando programa ClustalW [8]. Após a construção do alinhamentos múltiplo, a ferramenta PFMAKE, disponibilizada pelo projeto PROSITE [30], é empregada para construção de um novo perfil.

O perfil construído é chamado *não-escalado* (*"unscaled profile"*), *i.e.*, seus valores de alinhamento não possuem significado biológico e não podem ser comparados aos valores assinalados por outros perfis [30]. Para transformar esses valores em outros mais significativos, associamos a cada perfil uma função de normalização. Essa é uma função linear da forma:

$$f(x) = \frac{10x}{a}$$

onde x é o resultado não normalizado e a o resultado do alinhamento e entre a proteína original e seu próprio perfil. Dessa maneira, associamos uma seqüência à família descrita por perfil quando o resultado normalizado desse alinhamento é próximo a 10. Neste trabalho, assumimos que estes valores devem estar entre 8,5 e 11,75.

Naturalmente, antes de adicionarmos o perfil recém construído ao nosso conjunto, verificamos se este realmente descreve uma nova família ou se é equivalente a outro perfil, que foi construído anteriormente. Isso é feito alinhando a proteína original ao conjunto de perfis já construídos, a fim de verificar se essa proteína pertence a alguma das famílias representadas. Neste caso, o novo perfil é simplesmente descartado.

Esse procedimento é repetido para cada proteína que não foi classificada por alguma das famílias do projeto HAMAP, até que todos os genes analisados tenham sido associados a alguma família ou tenham sido classificados como *singletons*, ou seja, não homólogos a nenhum outro.

O processo de identificação de estruturas homólogas está representado graficamente no diagrama da figura 3.2, utilizando a notação de *diagrama de Jackson*. Segundo essa notação, retângulos simples são ações seqüenciais, retângulos marcados por I e $*$ representam iterações e retângulos marcados por S e $\circ$ representam ações condicionais [20].

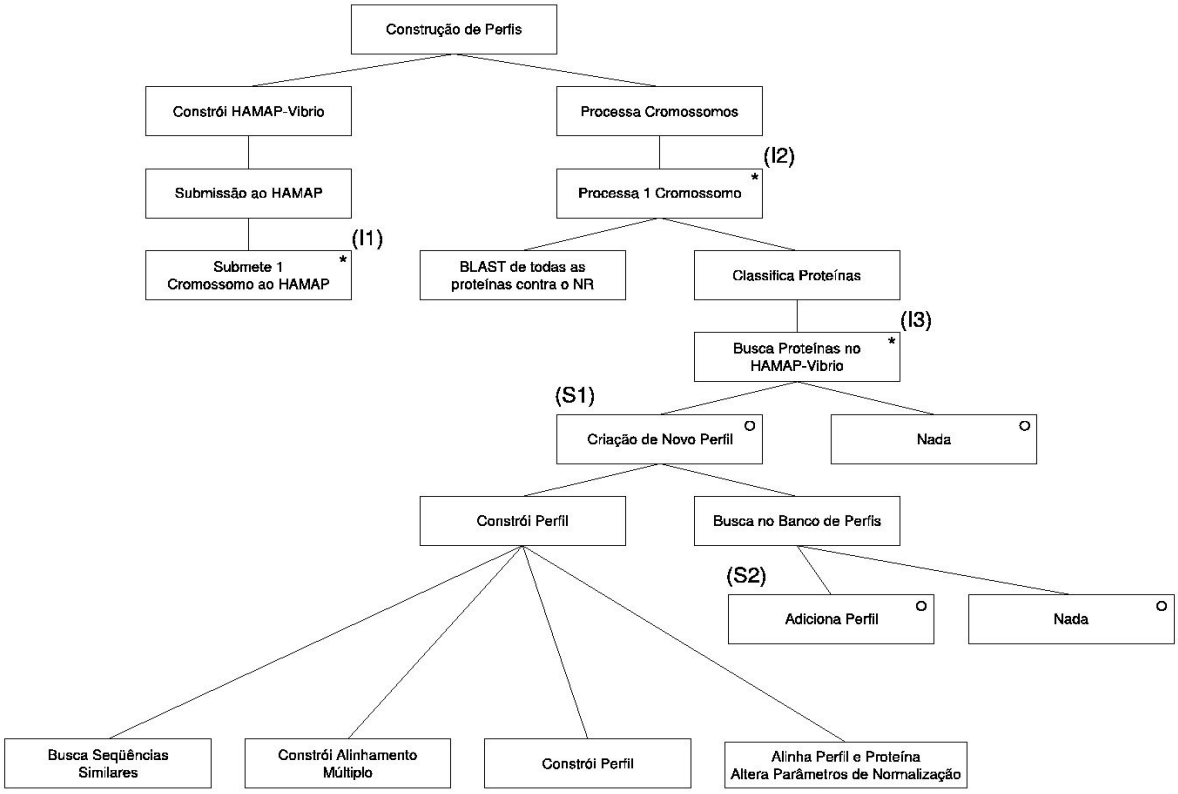

Figura 3.2: Diagrama de Jackson para o processo de construção de perfis. Segundo essa notação retângulos marcados I's e $*$ representam iterações, retângulos marcados com S's e $\circ$ representam ações condicionais

Como mostramos no diagrama da figura 3.2, o processo de construção de perfis foi realizado de maneira seqüencial, cromossomo a cromossomo, até que todas as proteínas de todos os organismos fossem classificadas como membro de uma família HAMAP, ou membro de uma família representada através de nossos perfis ou como *singleton*.

É importante esclarecermos que não há garantias de que a condição de independência do conjunto de genes analisados, definida na seção 3.1.1, seja verdadeira para todas as

proteínas classificadas através de nossos perfis. Em outras palavras, não podemos afirmar que se uma proteína Y é descrita pelo perfil construído para a proteína X, então X também é descrita pelo perfil construído para Y.

Todavia, nossos perfis constituem um primeiro passo para a adoção de um método mais robusto para identificação e catalogação de famílias de genes homólogos, de maneira semelhante, embora menos acurada, ao próprio projeto HAMAP. Acreditamos que esse banco de identificadores de famílias de proteínas homólogas possa ser empregados em outros trabalhos relacionados a este tema e, conseqüentemente, será alvo de incrementos e melhorias ao longo do tempo.

3.1.4 Aspectos práticos

Todos as etapas do processo de identificação de estruturas homólogas foram executadas através de *scripts* escritos em linguagem *Perl* e programas publicamente disponíveis, como BLASTP, ClustalW e PFMAKE, listados no apêndice A.

Duas etapas constituem grandes impactos no tempo de execução desse procedimento. A primeira é o alinhamento (BLAST) entre as proteínas do cromossomo analisado e o banco NR. A outra é a verificação se uma proteína já está representada no conjunto de perfis construídos, ou seja, se essa proteína é homóloga a outra tratada anteriormente, ou se o perfil construído a partir dela descreve uma nova família.

Em uma máquina com processador *Semprom*, *clock* de 1.8 gHz e 512 MB de memória RAM, o programa BLAST levava, em média, 4 dias para o conjunto de proteínas do maior cromossomo. Em uma máquina de mesmas características, o alinhamento entre o cromossomo e o conjunto de perfis já construídos levava entre 6 e 10 dias. O tempo de execução crescia a medida que mais cromossomos eram tratados pois, para cada um deles, novos perfis eram adicionados ao conjunto.

Ao final desse processo, representamos através do conjunto de perfis construídos 7.649 famílias de proteínas. O gráfico da figura 3.3 mostra como cada genoma analisado foi classificado, em termos do número de genes classificados pelas famílias do projeto HAMAP, classificados através do nosso conjunto de perfis e o número de genes *singletons*.

3.2 Estimando Distâncias Evolutivas

No capítulo 2 observamos que a operação de *block-interchange* é insufiente para modelar a evolução entre genomas completos de vibriões. Nossa solução para este problema é empregar um conjunto mais complexo de operações. O novo modelo para estimar distâncias de rearranjo é baseado na ocorrência de quatro operações fundamentais: fissões, fusões, reversões e perda de genes.

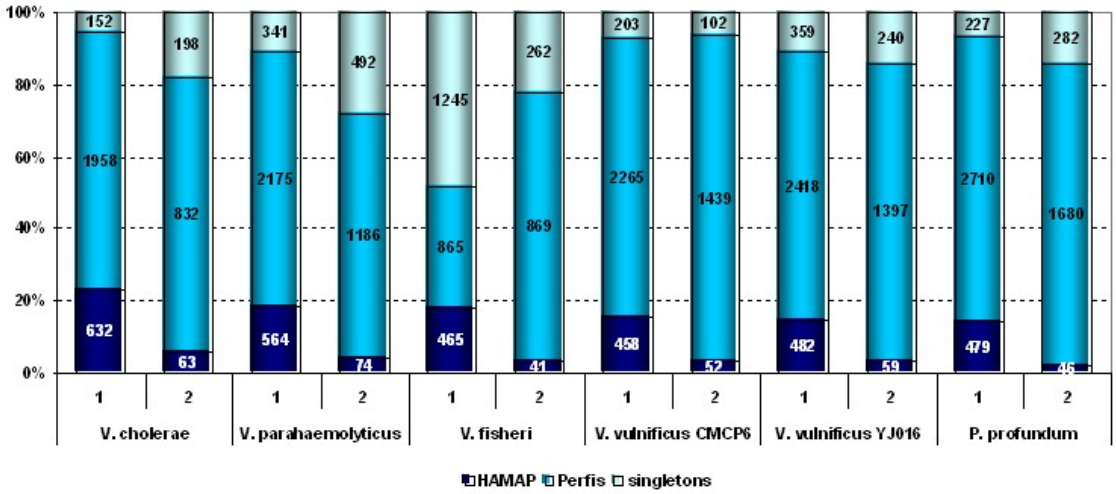

Figura 3.3: Distribuição das proteínas dos vibriões conforme sua classificação de homologia.

Translocações, *block-interchanges* e transposições não são consideradas operações fundamentais, mas ocorrências especiais de uma operação de fissão seguida por uma operação de fusão. De maneira semelhante, modelamos neste trabalho as operações de perda e ganho de genes e transferências horizontais como casos especiais de fissão/fusão, combinados com a perda ou ganho de um novo "cromossomo".

Nesta seção descreveremos a metodologia empregada para estimar distâncias evolutivas, i.e., o tratamento para perda de genes e duplicações e também o cálculo das distâncias de rearranjo.

3.2.1 Perda e Ganho de Genes

Algoritmos de cálculo de distâncias de rearranjo, usualmente, têm como premissa a igualdade entre o conteúdo dos genomas sendo comparados. Entretanto, especialmente para genomas procariontes, é extremamente difícil que essa hipótese seja verdadeira. Como discutimos na seção 1.2, eventos que alteram o conteúdo do genoma, como transferências horizontais e perda de genes, são comuns durante a evolução desses organismos, tanto para aquisição de novas características que aumentem sua capacidade de adaptação às mudanças ambientais quanto para eliminar pseudogenes ou genes cuja presença seja prejudicial a outro.

Para considerar em nossas distâncias evolutivas a ocorrência desses eventos, incluímos

uma nova operação:

- Perda e Ganho de genes.

Essa operação é tratada como um caso especial de fissão ou fusão e recebe peso unitário. Após sua aplicação a todos os genes que não estão presentes em ambos os genomas sendo comparados, a hipótese de igualdade de conteúdo será satisfeita, adequando o genoma as proposições assumidas pelos algoritmos de rearranjo.

É importante explicarmos que eventos de perda de genes e transferências horizontais são muito diferentes entre si e a complexidade associada a cada um deles é bastante distinta. Todavia, identificar a ocorrência desses eventos e diferenciá-los corretamente é uma tarefa difícil [26, 27]. Por essa razão, neste trabalho, adotamos a simplificação de representá-los como um só evento, que não necessariamente reflete a freqüência com a qual esses eventos ocorrem no processo de evolução.

3.2.2 Duplicações

A duplicação de um gene ou um conjunto de genes é um valioso mecanismo de adaptação às mudanças ambientais e exploração de novos nichos ecológicos. Gevers e Van der Peer, em seu estudo sobre duplicações em genomas de vibriões, enfatizam a importância desses eventos, representada pelas altas taxas de duplicação em genomas adaptados a ambientes hostis [15]. Como exemplo, podemos citar o vibrião *Photobacterium profundum*, adaptado a ambientes oceânicos de alta pressão, para o qual a taxa de duplicações é estimada em 36,7%.

Apesar da importância desses eventos, analisar uma ocorrência de duplicação e tratá-la corretamente não é uma tarefa simples, como ilustrado na figura 3.4. Neste exemplo, estamos comparando dois genomas e cada um deles contém duas cópias do gene X. Nesta situação, há duas possibilidades de tratamento para X: (i) manter uma única cópia em cada genoma, tratando cada parálogo como uma ocorrência de perda/ganho ou (ii) associar as cópias presentes nos dois genomas aos pares, de acordo com sua origem no genoma ancestral.

Infelizmente não é possível identificar cada um desses casos, uma vez que não possuímos nenhuma informação sobre a configuração do genoma ancestral. Dessa maneira, adotamos neste trabalho uma solução de compromisso: cada família que apresente mais de uma cópia entre os genomas sendo comparados será dividida em sub-famílias.

Cada sub-família será composta unicamente por genes pertencentes a organismos distintos, de maneira que não haja mais duplicações em cada uma delas.

Nosso método de sub-divisão das famílias com duplicações consiste em construir uma árvore filogenética entre todos os membros da família e encontrar o corte mais próximo à

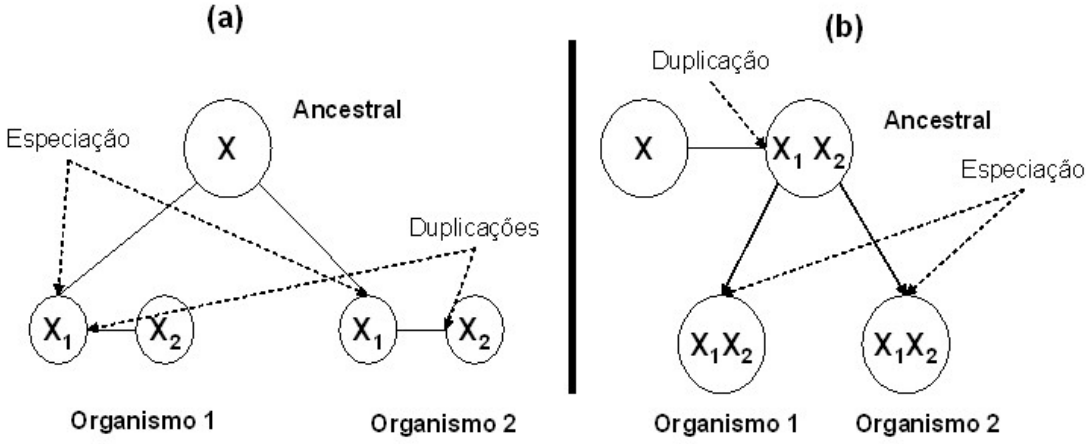

Figura 3.4: Comparação entre dois genomas contendo duas cópias do gene X. Em (a) os genes parálogos tiveram sua origem em uma duplicação; em (b) ambas as cópias foram herdadas do genoma ancestral.

raiz dessa árvore que crie somente sub-árvores que não contenham mais de um elemento pertencente a cada organismo. Cada sub-árvore constituirá uma nova família, na qual não há mais duplicações, como exemplificado na figura 3.5.

O tipo de árvore filogenética adotado foi a *árvore ultramétrica*. Uma árvore ultramétrica representa o tempo evolutivo através do comprimento de suas arestas, assumindo a hipótese de *relógio molecular*, ou seja, todos os elementos representados evoluíram a uma taxa constante. A topologia dessa árvore garante que sempre encontraremos um corte que satisfaça a condição mencionada, pois todas as folhas estão à mesma distância da raiz.

Podemos resumir o método de sub-divisão de famílias com duplicações nas seguintes etapas:

- Calcular a matriz de distância entre os elementos da família sendo tratada;
- Construir, a partir dessa matriz de distâncias, uma árvore ultramétrica;
- Encontrar o corte mais próximo à raiz da árvore que crie somente sub-árvores que não possuam mais de um elemento pertencente a um mesmo organismo;
- Identificar cada sub-árvore como uma nova família, na qual não ocorrem mais duplicações.

Neste trabalho, empregamos o programa *protdist* (ver apêndice A) para o cálculo das matrizes de distâncias. Para construção de árvores ultramétricas a partir das distâncias

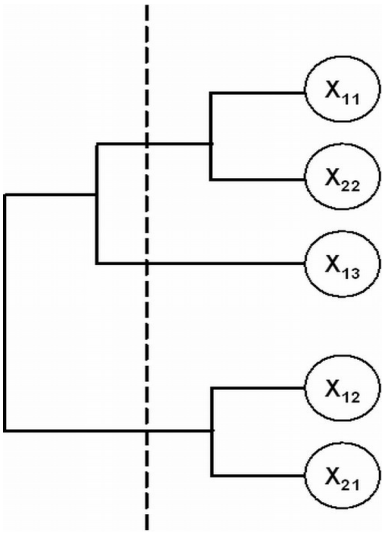

Figura 3.5: Neste exemplo, estamos analisando uma família composta por 5 genes:$\{X_{11}, X_{12}, X_{13}\}$, pertencentes ao organismo 1, e $\{X_{21}, X_{22}\}$, pertencentes ao organismo 2 (o primeiro sub-índice identifica o organismo ao qual cada gene pertence). Após o corte adequado, haverá 3 novas famílias: $\{X_{11}, X_{22}\}$, $\{X_{13}\}$, e $\{X_{12}, X_{21}\}$, as quais não apresentam nenhuma duplicação.

calculadas, utilizamos o método UPGMA (*Unweighted Pair Group Method with Arithmetic Mean*).

O algoritmo UPGMA parte de uma matriz de distâncias de dimensões $n \times n$ unindo os dois elementos mais próximos a fim de formar um novo agrupamento, reduzindo as dimensões da matriz para $(n-1) \times (n-1)$. A distância entre o agrumento formado e os demais elementos da matriz é calculado como a média aritmética entre as distâncias dos elementos que compuseram o agrupamento e os demais. Esse processo é repetido até que a matriz seja reduzida à dimensão 1×1.

O algoritmo 1 apresenta o pseudo-código para o cálculo da altura do corte em uma árvore ultramétrica, que divide uma família gênica em sub-famílias sem duplicações.

Algoritmo 1 Cálculo da altura do corte

Entrada: Árvore ultramétrica - T
Saída: Altura do corte - H

$V(T) \leftarrow$ conjunto de vértices
$H(v) \leftarrow$ altura de cada vértice em $V(T)$
$O(v) \leftarrow$ Organismo ao qual v pertence (somente se v é folha)
$G(v) \leftarrow$ Lista de organismos presentes na sub-árvore com raiz em v

para cada v em $V(T)$ **faça**
  se v é folha **então**
    $G(v) \leftarrow O(v)$
  senão
    $G(v) \leftarrow \{\}$
  fim-se
fim
$H \leftarrow H(Raiz\,\{T\}) + 1$
Explora$\{Raiz\,\{T\}\}$
retorne H

Procedimento Explora$\{v\}$
$U(v) \leftarrow$ Sub-árvore com raiz em v
para cada u em $U(v)$ **faça**
  Explora(u)
  se v é não folha **então**
    $G(v) \leftarrow$ G(filho a equerda$\{v\}$) + G(filho a direita$\{v\}$)
  fim-se
  se $G(v)$ possui duplicações e $H > H(v)$ **então**
    $H \leftarrow H(v)$
  fim-se
fim

Os métodos implementados para construção da árvore ultramétrica e para determinação da altura do corte estão listados no apêndice A.

Após a aplicação desse procedimento, não haverá mais famílias com elementos duplicados. Poderemos então submeter os genomas resultantes ao algoritmo de cálculo das distâncias de rearranjo, que será apresentado na seção 3.2.3.

Um exemplo real

A figura 3.6 ilustra um exemplo real de divisão de famílias com duplicações. Quando comparamos dois genomas de vibriões encontramos cinco cópias da proteína histidina quinase: duas pertencentes *V. cholerae* e três pertencentes ao *V. parahaemolyticus*. Cada gene está identificado por um código e pela sigla do organismo ao qual pertence. Essa família é dividida em três novas famílias: {15601492 vc, 28901075 vp}, {15641833 vc, 28897843 vp} e {28900955 vp}. A linha pontilhada mostra a altura do corte.

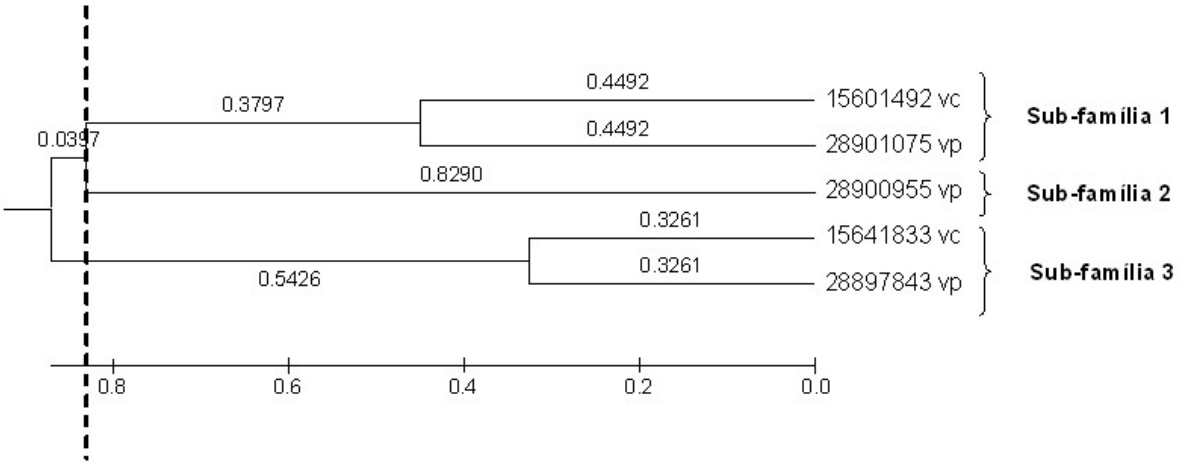

Figura 3.6: Exemplo real de subdivisão de uma família contendo cinco elementos: dois pertencentes *V. cholerae* e três pertencentes ao *V. parahaemolyticus*. O corte pontilhado divide esta família nas três sub-famílias mostradas.

3.2.3 Distâncias de Rearranjo

Após determinarmos as famílias de estruturas homólogas, tratado perda e ganho de genes e resolvido as ocorrências de duplicações, o problema final é calcular as distâncias de rearranjo entre os pares de genomas analisados, ou seja, encontrar o mínimo número de mutações, envolvendo genes ou conjunto de genes, que são capazes de transformar um genoma em outro.

Trabalhos clássicos de teoria de rearranjo de genomas usualmente analisam de maneira isolada uma operação e desenvolvem métodos e algoritmos eficientes para estimar distâncias empregando este único evento [4, 18, 21, 9]. Todavia, neste trabalho, nossa proposta é uma abordagem diferente, empregando um conjunto de operações básicas e suas combinações ao invés de uma única operação, a fim de obter distâncias de rearranjo mais relacionadas às distâncias biológicas.

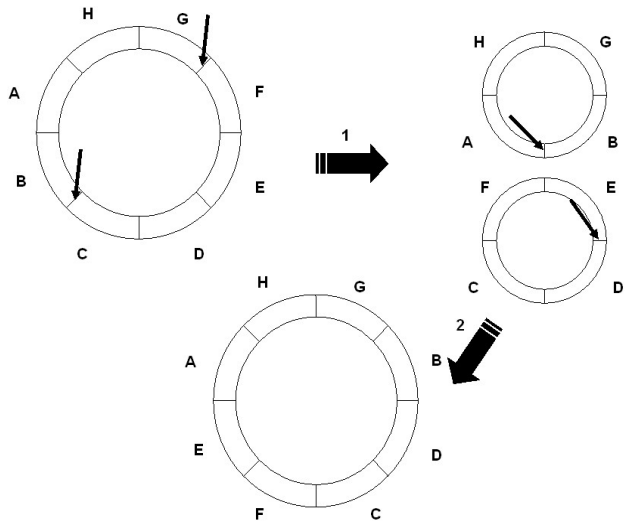

Figura 3.7: Exemplo de *block-interchange* entre {E,F} e {B} através de duas operações de *double-cut-and-join* consecutivas.

Ao encontro desse princípio, Yancopoulos *et al.* introduziram uma operação universal chamada *double-cut-and-join*, capaz de modelar um amplo conjunto de eventos: fissões, fusões, translocações, reversões e *block-interchanges* [36].

Uma operação de *double-cut-and-join* consiste em quebrar o genoma em 2 pontos, resultando em quatro terminais desconectados. Dependendo da maneira como a reconexão desses terminais é feita, a operação pode resultar em uma fissão, ou fusão, ou translocação ou reversão — cada uma delas recebendo peso 1. Uma operação de *block-interchange* pode ser obtida através de duas operações de *double-cut-and-join* sucessivas, por essa razão essa operação recebe peso 2, conforme mostra a figura 3.7.

Neste trabalho, empregamos o método descrito por Bergeron *et al.* [6] em seu artigo "*A unifying view of genome rearrangement*" para calcular distâncias de rearranjo usando a operação *double-cut-and-join*. Esse texto apresenta uma técnica simples para cálculo de distâncias em genomas multi-cromossomiais, com qualquer combinação de cromossomos lineares e circulares, que adaptamos nesse trabalho para o cálculo das distâncias de rearranjo entre os seis genomas de vibriões.

Os autores disponibilizam uma implementação *on-line* para cálculo das distâncias de *double-cut-and-join* chamada DCJ, disponível no endereço *http://bibiserv.techfak.uni-bielefeld.de/dcj/*[2]. Entretanto, essa ferramenta não suporta comparações entre genomas muito grandes, como os dos vibriões analisados neste trabalho. Por esse motivo, imple-

[2]Último acesso em: 03/02/2008.

mentamos em linguagem Java o algoritmo para cálculo das distâncias de rearranjo, listado no apêndice A, e empregamos a ferramenta DCJ somente para validação de pequenos exemplos.

Capítulo 4

Aplicando a nova metodologia

Neste capítulo, apresentaremos os resultados obtidos na comparação genômica de vibriões empregando a metodologia definida no capítulo 3 para identificação de estruturas homólogas e cálculo das distâncias evolutivas.

A tabela 4.1 apresenta a comparação entre essas técnicas e a metodologia que utilizamos em nossos experimentos iniciais, apresentados no capítulo 2.

	Metodologia Inicial	Nova metodologia
Identificação de estruturas homólogas	Similaridade	Informações do projeto HAMAP e Conjunto de Perfis
Tratamento de Duplicações	Escolha aleatória ou Exclusão	Subdivisão das famílias com genes duplicados
Operação de Rearranjo	*Block Interchange* *Perda de Genes*	*Double-cut-and-join* *Perda de Genes*

Tabela 4.1: Comparação entre as duas metodologias.

Os seis organismos analisados neste trabalho foram comparados aos pares, empregando nossa nova metodologia. As etapas necessárias para realizar essas comparações estão, em resumo, listadas a seguir.

- Classificar cada proteína como:
  - membro de uma família HAMAP, ou
  - membro de uma família descrita por um de nossos perfis adicionais, ou
  - singleton, *i.e.*, não homóloga a nenhuma outra proteína.

- Identificar famílias com múltiplos membros entre os genomas sendo comparados. Essas famílias serão subdivididas, conforme descrevemos na seção 3.2.2. Ao final deste procedimento, não haverá mais genes duplicados;

- Eliminar *singletons* e outros genes que não possuem um homólogo no outro genoma, conforme discutido na seção 3.2.1. Cada um desses genes será considerado uma operação de perda ou ganho;

- Aplicar os genomas resultantes das etapas anteriores ao algoritmo para cálculo da distância de rearranjo, apresentado na seção 3.2.3.

Ao final da execução dessas tarefas a todos os pares vibriões, teremos construído uma matriz de distâncias entre esses organismos. A matriz resultante pode então ser submetida a algoritmos de construção de árvores filogenéticas. Neste trabalho, utilizamos uma implementação do algoritmo *neighbor-joining*, disponível na ferramenta MEGA 4 [31].

A análise da árvore filogenética é muito importante, pois fornece uma representação gráfica das informações contidas na matriz de distâncias sobre a história desses organismos. Além disso, permite comparar nossos resultados com aqueles obtidos através de outros métodos de análise filogenética.

Construímos três matrizes de distâncias diferentes, empregando as seguintes abordagens:

1. Utilizando os dois cromossomos de cada genoma;

2. Analisando o cromossomo maior e o cromossomo menor separadamente.

No item 1, obtivemos uma árvore filogenética que não coincide com as tradicionalmente aceitas pela comunidade científica. Entretanto, está mais próxima as árvores adotadas para essas espécies que as obtidas em nossos experimentos iniciais, descritos no capítulo 2.

No item 2, obtivemos resultados equivalentes aos obtidos quando são empregados outros métodos de análise filogenética baseados em dados genômicos, tanto na análise do maior, quanto na análise do menor cromossomo.

Nas próximas seções apresentaremos as matrizes de distância e as árvores filogenéticas resultantes de cada uma dessas análises.

4.1 Analisando os dois cromomossomos

Nesta análise empregamos o par de cromossomos de cada vibrião para cálculo das distâncias evolutivas. Dessa maneira, estamos considerando a ocorrência de mutações entre pares de cromossomos não homólogos, como também mutações inter-cromossomiais para

estimar evolução entre esses organismos. A tabela 4.2 apresenta as distâncias evolutivas obtidas através dessa abordagem.

	V. c.	*V. p.*	*V. f.*	*V. v.* C.	*V. v.* Y.	*P. p.*
V. choleare	0	6292	6277	6296	5969	7351
V. parahaemolyticus		0	7154	6973	6614	7728
V. fisheri			0	7165	6745	8111
V. vulnificus CMCP				0	5835	7682
V. vulnificus YJ016					0	8114
P. profundum						0

Tabela 4.2: Distâncias evolutivas entre seis vibriões analisando os dois cromossomos de cada organismo.

A árvore filogenética construída a partir das distâncias evolutivas obtidas quando analisamos o par de cromossomos de cada genoma é mostrada na figura 4.1.

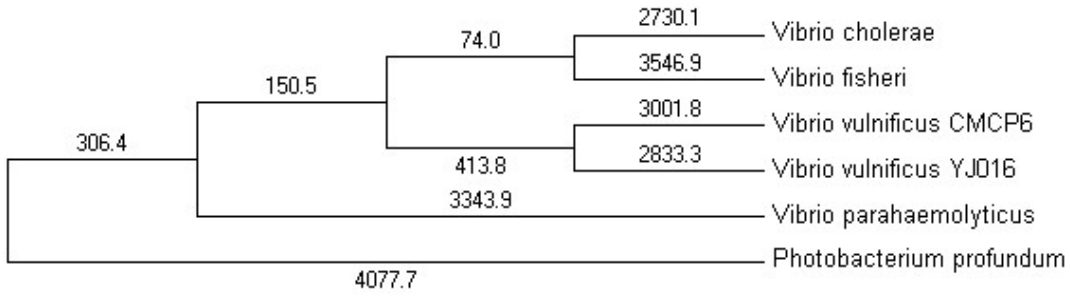

Figura 4.1: Árvore filogenética obtida empregando os dois cromossomos de cada vibrião.

Comparando a árvore obtida com as árvore da família *Vibrionacea* tradicionalmente aceitas pela comunidade científica, como a apresentada na figura 1.1, podemos dizer que os organismos *Vibrio vulnificus*, *Vibrio parahaemolyticus* e *Photobacterium profundum* estão corretamente classificados. Entretanto, ainda ocorrem problemas na classificação do *Vibrio choleare* e do *Vibrio fisheri*, que foram agrupados incorretamente.

4.2 Analisando o maior cromossomo

Nessa abordagem empregamos apenas o conteúdo do maior cromossomo de cada organismo para estimar distâncias filogenéticas. A matriz de distâncias resultante está representada na tabela 4.3.

	V. c.	*V. p.*	*V. f.*	*V. v.* C.	*V. v.* Y.	*P. p.*
V. choleare	0	3906	4403	3906	4150	4645
V. parahaemolyticus		0	4701	4031	4288	4929
V. fisheri			0	4548	4848	5062
V. vulnificus CMCP				0	3831	4757
V. vulnificus YJ016					0	5052
P. profundum						0

Tabela 4.3: Distâncias evolutivas entre seis vibriões analisando o maior cromossomo.

A árvore filogenética correspondente é equivalente às árvores filogenética tradicionalmente aceitas para esses organismos, conforme mostra a figura 4.2.

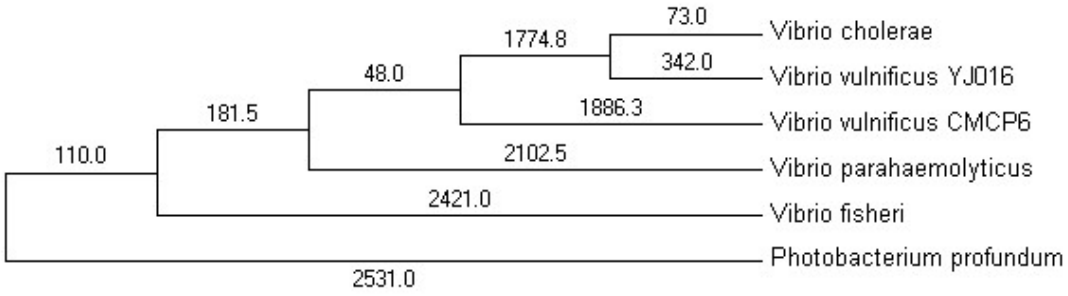

Figura 4.2: Árvore filogenética resultante da análise do maior cromossomo.

4.3 Analisando o menor cromossomo

Essa abordagem é semelhante a anterior, empregando o conteúdo do menor cromossomo de cada organismo, ao invés do maior, para cálculo das distâncias filogenéticas. A matriz de distâncias resultante está apresentada na tabela 4.4.

A figura 4.3 mostra a árvore filogenética construída a partir das distâncias obtidas nessa análise. Novamente, a árvore obtida mostra uma topologia similar àquela construída a partir de outros métodos genômicos.

	V. c.	*V. p.*	*V. f.*	*V. v.* C.	*V. v.* Y.	*P. p.*
V. choleare	0	2427	2061	2226	2336	2926
V. parahaemolyticus		0	2697	2667	2779	3512
V. fisheri			0	2416	2561	2942
V. vulnificus CMCP				0	2006	3266
V. vulnificus YJ016					0	3407
P. profundum						0

Tabela 4.4: Distâncias evolutivas entre seis vibriões analisando o menor cromossomo.

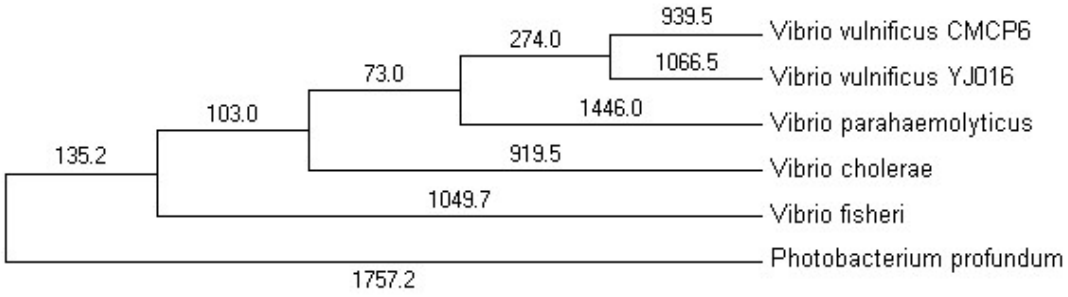

Figura 4.3: Árvore filogenética resultante da análise do menor cromossomo.

4.4 Discussão

Quando comparamos os resultados apresentados neste capítulo com aqueles obtidos em nossos experimentos iniciais, percebemos que houve ganhos na acurácia do método de comparação genômica quando empregamos técnicas mais robustas para identificação de estruturas homólogas e tratamento de duplicações e um conjunto mais complexo de operações de rearranjo.

Todavia, na análise em conjunto do par de cromossomos de cada genoma, ainda ocorrem problemas na classificação de alguns organismos. Essa falha é corrigida quando utilizamos nossa metodologia para analisar cada cromossomo isoladamente. Nesta última situação, tando na análise do maior, quanto na análise do menor cromossomo, as árvores obtidas são equivalentes às construídas através da aplicação de outros métodos de comparação genômica.

A análise de cada cromossomo separadamente também foi empregada por Lin *et al.* para o cálculo da distância de *block-interchange* entre os genomas dos organismos *Vibrio vulnificus*, *Vibrio parahaemolyticus* e *Vibrio cholerae*.

O resultado obtido é muito importante, pois corrobora nosso método e a teoria de

rearranjo de genomas como uma abordagem alternativa para análise de genomas completos. Além disso, pode indicar que alguns eventos de rearranjo modelados nesse trabalho exercem um papel importante na evolução de vibriões.

Entretanto, as diferenças encontradas entre a utilização dos dois cromossomos combinados e as análises individuais sugerem ser necessária uma melhor atribuição de pesos a cada operação de rearranjo. Neste caso, especificamente, há forte indicação de que o peso unitário atribuído à operação de translocação, que modela eventos entre cromossomos não-homólogos, foi sub-dimensionado e precisa ser reavaliado.

Capítulo 5

Conclusões

Neste trabalho apresentamos uma metodologia para análise de genomas completos utilizando a teoria de rearranjo de genomas e discutimos sua aplicação a seis organismos da família *Vibrionacea*.

Em nosso método, contemplamos todas as etapas de um processo de comparação genômica, incluindo: identificação de estruturas homólogas, tratamento para variações no conteúdo gênico e duplicações e cálculo de distâncias de rearranjo.

Para identificação de homologia, definimos algumas propriedades que um método robusto e eficiente deve satisfazer, são elas: *transitividade* e *independência do conjunto de genes analisados*.

Ao encontro desses princípios, definimos um método baseado nas informações do projeto HAMAP e em um conjunto complementar de perfis, para classificar proteínas que ainda não são contempladas pelas famílias HAMAP.

Devemos esclarecer que não há garantias de que a *independência do conjunto de genes analisados* seja verdadeira para todas as proteínas classificadas através do nosso conjunto de perfis. Todavia, acreditamos que, apesar das possíveis falhas, esse método apresenta vantagens em relação aos métodos de identificação de homologia tradicionais, baseados no alinhamento de seqüências. Além disso, os perfis construídos durante esse trabalho constituem uma base inicial de descritores de famílias de proteínas, que será alvo de incrementos e melhorias com novos trabalhos.

Neste trabalho introduzimos uma nova abordagem para eliminação de genes duplicados, uma vez que algoritmos de rearranjo, usualmente, não tratam duplicações.

Na comparação entre dois genomas, construímos uma árvore ultramétrica para cada família gênica que possuir mais de um gene em ao menos um desses genomas. A técnica de tratamento de duplicações consiste em encontrar o corte mais próximo à raiz da árvore, de maneira a criar somente sub-árvores que não possuam mais de um elemento pertencente ao mesmo organismo. Cada sub-árvore constituirá uma nova família na qual não há

duplicações.

Nosso modelo de cálculo de distâncias de rearranjo é baseado em duas operações:

- *Perda de Genes* – de maneira simplificada, modelamos através dessa operação perda de genes e transferência horizontais. A aplicação dessa operação a *singletons* e outros genes que não possuem um homólogo correspondente transforma genomas com conteúdos distintos em outros de mesmo conteúdo, que podem ser comparados através de algoritmos de cálculo de distâncias de rearranjo;

- *Double-cut-and-join* – uma operação universal capaz de modelar eventos de fissões, fusões, translocações, reversões e *block-interchanges*.

Aplicamos os modelos de identificação de homologia e cálculo de distâncias de rearranjo para comparar seis genomas de vibriões: *Vibrio cholerae*, *Vibrio parahaemolyticus*, *Vibrio fisheri*, *Vibrio vulnificus* CMCP6 e YJ016 e *Photobacterium profundum*.

Os resultados obtidos quando consideramos os dois cromossomos de cada organismo divergem das árvores filogenéticas construídas empregando outros métodos confiáveis de análise genômica. Especificamente, houve uma falha na classificação dos organismos *Vibrio cholerae* e *Vibrio fisheri*, que foram incorretamente agrupados.

Entretanto, quando aplicamos nosso método na análise em separado de cada cromossomo, obtivemos resultados equivalentes aos tradicionalmente aceitos pela comunidade científica.

Consideramos esse resultado positivo, pois corrobora nossa metodologia e a teoria de rearranjo de genomas como uma alternativa para a comparação de genomas completos. Além disso, fornece indicações de que alguns eventos de rearranjo modelados neste trabalho exerceram um papel importante durante a evolução desses vibriões.

As diferenças entre as árvores filogenéticas construídas a partir da combinação dos dois cromossomos e da análise individual de cada um deles assinalam que é necessária uma melhor regra para atribuição de pesos às operações de rearranjo. Neste caso em particular, há fortes indicações de que o peso da operação de *translocação*, que modela eventos entre cromossomos não homólogos, deveria ser superior ao valor unitário atribuído neste trabalho.

5.1 Trabalhos Futuros

Este trabalho é uma contribuição inicial para a definição de uma metodologia de comparação de genomas completos empregando a teoria de rearranjo de genomas. Estamos certos de que ainda há muito a ser feito.

Por exemplo, a definição de um esquema mais adequado de pesos às operações de rearranjo é uma forte melhoria que deve ser implementada em futuros trabalhos relacionados a esse tema.

Outro trabalho importante é a construção de uma ferramenta de edição de perfis. Sempre que fosse identificado um problema na classificação de uma família descrita por um dos perfis construídos automaticamente neste trabalho, essa ferramenta permitiria que esse perfil fosse editado ou substituído, melhorando a acurácia do método de identificação de estruturas homólogas.

Além dessas, outras contribuições importantes estão relacionados a extensão desse estudo, como:

- A inclusão de outros genomas de vibriões, como *Vibrio cholerae* O395 e *Vibrio harveyi* – que já estão publicamente disponíveis;
- A aplicação desta metodologia a outras famílias de bactérias – que validaria nosso método e traria um forte incremento ao banco de perfis descritores de proteínas que iniciamos durante a realização deste estudo.

Apêndice A

Listagem de programas utilizados:

- **Blast** *Basic search alignment tool*. Utilizado para identificação do grau de similaridade entre seqüências. Disponível para download no endereço:

  http://www.ncbi.nlm.nih.gov/BLAST/download.shtml

- **CLUSTALW** Ferramenta para construção de alinhamentos múltiplos entre seqüências de nucleotídeos ou proteínas. Disponível para download no endereço:

  ftp://ftp.ebi.ac.uk/pub/software/clustalw2/

- **PHYLIP** Pacote com diversos programas para análise filogenética. Entre eles o Protdist, utilizado para construção de matrizes de distâncias entre proteínas, empregado neste trabalho. Disponível para download no endereço:

  http://evolution.genetics.washington.edu/phylip/getme.html

- **pfmake** Ferramenta para construção de perfis a partir de um alinhamento múltiplo. Disponível no pacote **pftools**, no endereço:

  http://www.isrec.isb-sib.ch/ftp-server/pftools/

- **ps_scan** Ferramenta para alinhamento entre proteínas e perfis. Disponível no endereço:

  ftp://ftp.expasy.org/databases/prosite/tools/ps_scan

Os seguintes programas foram desenvolvidos ao longo da evolução deste trabalho. Estão disponíveis para *download* no endereço:

http://www.students.ic.unicamp.br/ ra041481

- **parseblast.pl** Script em linguagem *Perl* para parse do relatório gerado pela ferramenta BLAST, empregando os critérios de similaridade definidos na seção 3.1.3;

- **baixa_seq.pl** Script em linguagem *Perl* para copiar do site NCBI[1] o conjunto de proteínas identificadas como similares à cada proteína para a qual se deseja construir um novo perfil, ou seja, para todas aquelas não classificadas pelas famílias HAMAP;

- **roda_Clustalw.pl** Script em linguagem *Perl* para construção dos alinhamentos múltiplos que serão empregados para a construção de novos perfis. Isso é realizado através de chamadas ao software ClustalW;

- **criaPerfis.pl** Script em linguagem *Perl* para criação de perfis. Esse programa:

    1. Cria perfis a partir de alinhamentos múltiplos, executando chamadas ao programa *pfmake*;
    2. Altera a função de normalização desses perfis, conforme descrevemos na seção 3.1.3;
    3. Busca, na base de perfis já construídos, um perfil que descreva a proteína sendo analisada. Caso haja um perfil construído anteriormente que descreva a proteína em análise, o perfil recém construído será descartado.
    4. Cria uma nova base de perfis não redundantes, incluindo aqueles construídos para o cromossomo em análise.

- **lista_famílias.pl** Script em linguagem *Perl* que relaciona, para cada cromossomo dos organismos analisados, o sinal de suas proteínas e a família associada a cada uma delas;

- **famílias_repetidas.pl** Script em linguagem *Perl* que identifica, para cada par de genomas, as famílias que estão duplicadas em cada um deles ou em ambos. Para cada uma dessas famílias é construído um alinhamento múltiplo (utilizando o programa *CLUSTALW*) e, então, é calculada a matriz de distâncias (utilizando o programa *protdist*, disponível no pacote *PHYLIP*). Essa matriz será empregada para construção da árvore ultramétrica utilizada para subdivisão da família original em novas sub-famílias, as quais não apresentarão elementos duplicados;

- **representa.pl** Prepara cada par de genomas para serem aplicados ao algoritmo de cálculo da distância de rearranjo. Para isso, substitui cada família duplicada pelas sub-famílias correspondentes, elimina famílias que não estão presentes em ambos os genomas e calcula o número de operações de perda e ganho de genes;

[1] http://www.ncbi.nlm.nih.gov/

- **UPGMA** Pacote em Java para construção da árvore UPGMA, determinação da altura do corte e divisão das famílias com elementos repetidos em novas sub-famílias;
- **DCJ** Pacote em Java para cálculo das distâncias de rearranjo utilizando a operação de *Double-Cut-and-Join*.

Bibliografia

[1] HAMAP – High-quality Automated and Manual Annotation of microbial Proteomes. url: http://www.expasy.ch/sprot/hamap/index.html. Último acesso em 20 de Janeiro de 2008.

[2] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Computacional Biology*, 215:1403–1410, 1990.

[3] D. Bader, B. Moret, and M. Yan. A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *J. Comput. Biol.*, 5(8):483–491, 2001.

[4] B. Bafna and P. Pevzner. Sorting by transpositions. *SIAM J. Discrete Mathematics*, 11(2):224–240, 1998.

[5] E. Belda, A. Moya, and F. J. Silva. Genome rearrangement distances and gene order phylogeny in gamma-Proteobacteria. *Mol. Biol. Evol.*, 22(6):1456–1467, 2005.

[6] A. Bergeron, J. Mixtacki, and J. Stoye. A unifying view of genome rearrangements. In *proocedings of WABI 2006*, volume 4175 of *Lecture Notes in Computer Science*, pages 163–173, 2006.

[7] C. Chen, K. Wu, Chang Y., C. Chang, H. Tai, T. Liao, Y. Liu, H. Chen, A. Shen, J. Li, C. Shao, C. Lee, L. Hor, and S. Tsai. Comparative genome analysis of *Vibrio vulnificus*, a marine pathogen. *Genome Research*, 13:2577–2587, 2003.

[8] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson. Multiple sequence alignment with the CLUSTAL series of programs. *Nucleic Acids Res*, 31:3497–3500, 2003.

[9] D. A. Christie. Sorting permutation by block-interchanges. *Information Processing Letters*, 60(4):165–169, 1996.

[10] D. A. Christie. *Genome Rearrangement Problems*. PhD thesis, Institute of Computer Science - University of Glasgow, 1998.

[11] T. Coenye, D. Gevers, Y. Van der Peer, P. Vandamme, and J. Swings. Towards a prokaryotic genomic taxonomy. *FEMS Microbiology Reviews*, 29:147–167, 2005.

[12] A. C. Darling, B. Mau, F. R. Blattner, and N. T. Perna. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, 14(7):1394–1400, 2004.

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

[14] A. Gattiker, K. Michoud, C. Rivoire, A. H. Auchincloss, E. Coudert, T. Lima, P. Kersey, M. Pagni, C. J. Sigrist, C. Lachaize, A. L. Veuthey, E. Gasteiger, and A. Bairoch. Automatic annotation of microbial proteomes in Swiss-Prot. *Comp Biol Chem*, 27:49–58, 2003.

[15] D. Gevers and Y. Van de Peer. Gene duplicates in vibrio genomes. In F. L. Thompson, B. Austin, and J. Swings, editors, *Invited book chapter in The Biology of Vibrios*. ASM Press, 2006.

[16] M. Gribskov, A. D. Mclachlan, and D. Eisenberg. Profile analysis: detection of distantly related proteins. *Proc. Natl. Aca. Sci. USA*, 84:4355–4358, 1987.

[17] S. Hannenhalli, C. Chappey, E. Koonin, and P. A. Pevzner. Genome sequence comparison and scenarios for gene rearrangements: a test case. *Genomics*, 30:299–311, 1995.

[18] S. Hannenhalli and P. A. Pevzner. Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. In *Proceedings of the 27th Annual Symposium on the Theory of Computing (STOC 95)*, page 178–189, 1995.

[19] S. Hannenhalli and P. A. Pevzner. Transforming men into mice (polynomial algorithm for genomic distance problem). In *Annual IEEE Symposium on Foundations of Computer Science*, pages 581–592, 1995.

[20] M. A. Jackson. *Principles of Program Design*. Academic Press, 1975.

[21] H. Kaplan, R. Shamir, and R. E. Tarjan. Faster and simpler algorithm for sorting signed permutations by reversal. *SIAM Journal of Computing*, 29(3):880–892, 2000.

[22] C. Lin, C. Lu, H. Chang, and C. Tang. An efficient algorithm for sorting by block-interchanges and its application to the evolution of vibrio species. *Journal of Computacional Biology*, 12(1), 2005.

[23] M. C. J. Maiden, R. Urwin, et al. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *PNAS*, 95:3140–3145, 1998.

[24] A. McLysaght, P. F. Baldi, and B. S. Galt. Extensive gene gain associated with adaptive evolution of poxviruses. *Proceedings of the National Academy of Sciences of the USA*, 100(26):15655–15660, 2003.

[25] J. Meidanis and Z. Dias. Genome rearrangement distance by fusion, fission and transposition is easy. *Proceedings os the String Processing and Information Retrieval (SPIRE'2001)*, pages 250–253, 2001.

[26] H. Ochman, J. G. Lawrence, and E. A. Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405:299–304, 2000.

[27] H. Ochman and N. A. Moran. Genes lost and genes found: Evolution of bacterial pathogenesis and symbiosis. *Science*, 292(5519):1096–1098, 2001.

[28] D. Sankoff and N. El-Mabrouk. Genome rearrangement. *In Current Topics in Computational Biology*, 2001.

[29] T. Sawabe, K. Kita-Tsukamoto, and F. L. Thompson. Infering the Evolutionary History of Vibrios by Means of Multilocus Sequence Analysis. *Journal of Bacteriology*, 189(21), 2007.

[30] C. J. A. Sigrist, L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch, and P. Bucher. PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinformatics*, 3:265–274, 2002.

[31] K. Tamura, J. Dudley, M. Nei, and S. Kumar. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, 24:1596–1599, 2007.

[32] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, 28(1):33–36, 2000.

[33] F. L. Thompson, T. Iida, and J. Swings. Biodiversity of Vibrios. *Microbiology and Molecular Biology Reviews*, 68(3):403–431, 2004.

[34] M. E. M. T. Walter, Z. Dias, and J. Meidanis. A new approach for the transposition distance. In *proocedings of the String Processing and Information Retrieval (SPIRE'2000)*, 2000.

[35] C. Woese and G. Fox. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 74, pages 5088–5090, 1977.

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