Cláudio Luís Marques Sampaio

Busca de um tamanho ótimo de gene e proteína para maximização da qualidade da filogenia resultante

Dissertação apresentada ao Instituto de Biologia da Universidade Estadual de Campinas para obtenção do Grau de Mestre em Genética e Biologia Molecular.

Orientador:
Gonçalo Guimarães Pereira

LABORATÓRIO DE GENÔMICA E EXPRESSÃO (LGE)
INSTITUTO DE BIOLOGIA
UNIVERSIDADE ESTADUAL DE CAMPINAS

Campinas - SP, Brasil

19 de fevereiro de 2010

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DO INSTITUTO DE BIOLOGIA – UNICAMP

Sa47b	Sampaio, Cláudio Luís Marques Busca de um tamanho ótimo de gene e proteína para maximização da qualidade da filogenia resultante / Cláudio Luís Marques Sampaio. – Campinas, SP: [s.n.], 2010. Orientadores: Gonçalo Amarante Guimarães Pereira. Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Biologia. 1. Filogenia. 2. Entropia. 3. DNA. I. Pereira, Gonçalo Amarante Guimarães. II. Universidade Estadual de Campinas. Instituto de Biologia. III. Título.

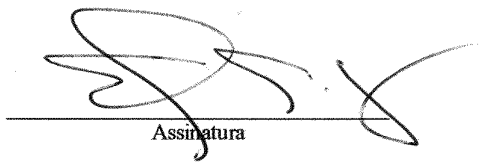
Título em inglês: The search for an optimal size of gene and protein for maximum philogeny quality.
Palavras-chave em inglês: Phylogeny; Entropy; DNA.
Área de concentração: Bioinformática.
Titulação: Mestre em Genética e Biologia Molecular.
Banca examinadora: Gonçalo Amarante Guimarães Pereira, Nilce Maria Martinez-Rossi, Marcelo Brocchi.
Data da defesa: 19/02/2010.
Programa de Pós-Graduação: Genética e Biologia Molecular.

Dissertação de mestrado sob o título *"Busca de um tamanho ótimo de gene e proteína para maximização da qualidade da filogenia resultante"*, defendida por Cláudio Luís Marques Sampaio e aprovada em 19 de fevereiro de 2010, em Campinas, Estado de São Paulo, pela banca examinadora constituída pelos professores:

Campinas, 19 de Fevereiro de 2010

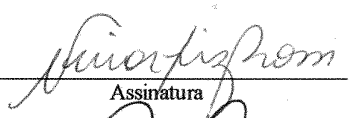
BANCA EXAMINADORA

Prof. Dr. Gonçalo Amarante Guimarães Pereira



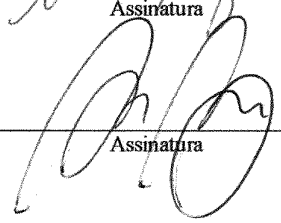
Assinatura

Profa. Dra. Nilce Maria Martinez-Rossi



Assinatura

Prof. Dr. Marcelo Brocchi



Assinatura

Prof. Dr. Johana Rincones

Assinatura

Prof. Dr. Fernando von Zuben

Assinatura

Resumo

Um problema recorrente em Filogenética é saber de antemão que melhores tamanhos de genes ou proteínas se deve ter para a construção de dendrogramas mais precisos. Neste trabalho, examinamos quais os efeitos de variados tamanhos de um alinhamento conhecido na qualidade da inferência de filogenia, em particular a filogenia dos fungos, utilizando 198 táxons fungais e 16 de grupo externo. Adicionalmente, calculamos a entropia de Shannon de cada ponto do alinhamento e fizemos iterações semelhantes por seus limiares. Para isto construímos um programa open-source baseado no toolkit *bioperl* que calcula estes dados. Concluímos que tanto para as iterações por tamanho quanto os para entropia, os limiares ideais são aquém do tamanho total do gene, podendo justificar uso de drafts de seqüenciamentos em inferências filogenéticas usando um pequeno número de regras.

Abstract

A recurring issue in phylogenetics is knowing beforehand which best sizes for genes or proteins one should have for building more accurate cladograms. Herein we examine the effects or varying sizes of a known aligment on the quality of its inferred phylogeny, specifically considering the fungi phylogeny by using 198 fungal taxa plus 16 outgroup taxa. Additionally, we calculate the Shannon entropy of each point of the alignment and iterate similarly by its thresholds. To that end, we developed an open-source software based on the *bioperl* toolkit to calculate this data. Finally, we concluded that either for the size iterations or for the entropy iterations, the ideal thresholds are below the gene full size, justifying the use of sequencing drafts in phylogenetic inferrences using a handful of rules.

Dedicatória

Dedico este trabalho aos meus pais, sempre orgulhosos de minhas realizações e sempre fazendo o possível para ajudá-las a se tornarem realidade.

À toda a minha família, formada também de irmão e irmã, tios e tias, primos e primas, avô e avó, que, morando longe na cidade de Juiz de Fora, sempre me apoiaram nesta empreitada.

À minha *team leader* Rosana Elias e meu gerente Delmar Demarchi, assim como meu gerente anterior Claudio Yamazaki que, ao concederem inúmeras exceções e folgas da IBM para o trabalho na dissertação e no programa, permitiram que esse trabalho chegasse à sua conclusão.

Agradecimentos

Agradeço principalmente ao professor Gonçalo que com sua sabedoria cirúrgica, sempre soube dizer com absoluta precisão o que eu deveria preservar ou remover do meu trabalho, assim como o melhor jeito de apresentá-lo.

Indispensável também foi o apoio, ajuda e acompanhamento de meu colega de laboratório Ricardo Tibúrcio, que com seu extenso conhecimento de Filogenética foi uma grande força impulsionadora de meu trabalho, sempre me dizendo as direções e dizendo como e quando fazer a coisa certa.

Agradeço também ao time de bioinformática do Laboratório de Genômica e Expressão, em especial ao Marcelo Carazzolle que me ajudou com os recursos necessários para eu processar meus dados, assim como dicas pontuais.

O pessoal de lista de discussão da API bioperl, que utilizei em meu programa, sempre foi paciente e prestativo com minhas dúvidas, em particular Mark Jensen com sua ajuda na parte de entropia.

E por fim, ao utilizar várias partes diferentes para fazer meu processamento, topei com amigos que, apesar de morarem em outros países, não deixaram de atender meus pedidos de orientação, como o Dr. Glenn Hickey e o Dr. Pablo Goloboff.

Sumário

Referências Bibliográficas **85**

Lista de Figuras

1 Introdução

"*Na visão tradicional em que as espécies*
foram independentemente criadas,
não obtemos uma explicação científica."
Charles Darwin

1.1 Estrutura da dissertação

Neste capítulo 1 são apresentados os conceitos úteis a este trabalho, começando pela caracterização do problema seguido de uma introdução à filogenética com ênfase aos tópicos utilizados no trabalho. Em seguida, a Teoria de Informação, um ramo da matemática orientado para ciência de computação, é brevemente explicada para fazer referência à entropia, grandeza escolhida para aprimorar os resultados além de aumentar a abrangência do escopo do trabalho.

Esclarecidos o problema e os fundamentos, se segue a explicação dos objetivos do trabalho no capítulo 3.

Logo após, no capítulo 4, são apresentadas as ferramentas computacionais utilizadas, como foram montadas e o que foi levado em conta no desenvolvimento do *software* open-source que viria a produzir os dados pretendidos, assim como uma introdução ao seu funcionamento, modo de uso e recursos.

Estes dados são analisados no capítulo 5, iniciando pelos critérios de limitação e filtragem utilizados e procedendo com as conclusões e possíveis aplicações dos dados no capítulo 6.

Após as conclusões, é apresentado um pequeno glossário de termos comuns tanto da filogenética quanto de Teoria da Informação na seção 6.

1.2 Caracterização do problema

Apesar do número gradativamente maior de espécies com genoma completamente seqüenciados hoje em dia, ainda se enfrenta um dilema. Todo processo tem um início, e é comum relativamente cedo neste processo de seqüenciamento ter-se rascunhos do material que, se considerados suficientemente informativos, podem gerar dados que adiantariam seu estudo e anotação, possibilitando inferir uma série de propriedades sobre o organismo que são cruciais para seu adequado tratamento biológico.

Destas propriedades,a mais importante é a classificação filogenética, seja do organismo, seja de outro táxon como um de seus genes. O posicionamento filogenético é importante porque contextualiza o táxon - estabelece suas relações evolutivas e esclarece sua história, levando à inferência de propriedades químicas e bioquímicas, físicas, biológicas, anatômicas, ecológicas e geográficas que não estariam inicialmente aparentes. Na medida em que estas propriedades contribuem para esta contextualização, sua retroalimentação aos dados tem um papel importante

de diminuir o tempo necessário para o seqüenciamento final do táxon, cortando ambigüidades e aumentando a precisão, como no caso típico da classificação de um gene como parálogo ou ortólogo.

Para esta classificação prévia ser calculada é necessário descobrir um critério que possa dizer, para nossos rascunhos, um tamanho ideal de filamento abaixo do total cuja filogenia possa ser inferida com qualidade; cabe aqui notar que esta inferência é um cálculo que por si só toma um tempo considerável.

Procurar descobrir este critério apenas pelo tamanho já seria importante; no entanto, é também do interesse deste estudo encontrar alguma outra medida que possa aumentar nossas chances de obter uma análise filogenética precisa, que usaremos em conjunção com o tamanho.

O escopo deste trabalho é exploratório: ao invés de usar fórmulas e caminhos matemáticos conhecidos para se chegar a uma conclusão, procura-se trabalhar com o sinal biológico e dados conhecidos para, com suas análises e comparações, serem descobertos os padrões desejados. Neste contexto, a conhecida especificidade dos dados para diferentes grupos taxonômicos nos impele a restringir o estudo a um grupo específico de organismos. Fatores como diferentes taxas de evolução, saturação, ocorrência de seleção natural e outros, que introduzem artefatos nos cálculos, serão assim minimizados.

O laboratório de Genômica e Expressão da Unicamp vem há anos realizando seqüenciamentos e estudos em cima de uma praga a um dos anteriormente principais produtos de exportação da agricultura brasileira, o cacau. A praga, conhecida informalmente como Vassoura-de-Bruxa e tendo como nome científico *Moniliophtora perniciosa* (recentemente reclassificada filogeneticamente; seu nome anterior era *Crinipellis perniciosa*), é um fungo basidiomiceto de grande resistência e por isso imensa importância comercial. Torna-se assim um excelente caso de estudo, evidenciado até pela sua recente classificação filogenética. Tomamos como parâmetro deste trabalho, então, o grupo maior a que ele pertence, o reino Fungi.

Não há ferramentas prontas, no entanto, que nos permitam fazer este tipo de análise comparativa com iterações. Como objetivo secundário deste estudo, desenvolveremos uma ferramenta open-source, com modos gráfico e de linha de comando, que nos permitirá, de forma flexível, gerar estes dados. A ferramenta será disponibilizada para o público.

2 Conceitos utilizados

"*Coisas complexas e estatisticamente improváveis são por sua própria natureza mais difíceis de explicar do que coisas simples, estatisticamente prováveis.*"
Richard Dawkins

2.1 Filogenia

Uma filogenia, árvore filogenética ou dendrograma é uma exibição em forma de árvore das relações evolutivas entre várias espécies, genes, seqüências de DNA ou características fenotípicas.

Em uma árvore filogenética, cada nodo com descendentes representa o mais recente antepassado comum destes e os comprimentos dos ramos podem representar estimativas do tempo evolutivo ou da quantidade de mudança evolutiva.

Cada nó terminal em uma árvore filogenética é chamado de unidade taxonômica ou táxon. Nós internos geralmente são chamados de Unidades Taxonômicas Hipotéticas. As árvores filogenéticas são confeccionadas a partir de uma matriz contendo os dados disponíveis (morfológicos, químicos ou genéticos) sobre os táxons estudados. Estes dados são comparados, e os táxons agrupados pelas semelhanças e diferenças entre si em clados. Por questão de simplicidade e facilidade de cálculos, usamos sempre grupos de dois elementos - isto é, as árvores terão apenas bifurcações, e uma árvore multifurcada pode geralmente sem prejuízo ser transformada em uma árvore equivalente bifurcada.

Os métodos de agrupamento dependem do modelo de evolução adotado (hipótese) e envolvem diferentes métodos estatísticos, que podem considerar ou não diferentes parâmetros como variabilidade dos dados, diferentes taxas evolutivas entre eles e transmissão horizontal. Diferentes métodos e modelos apresentarão diferentes confianças na inferência a partir dos dados; essa confiança, ou probabilidade de acerto, é geralmente estimada com métodos estatísticos como o bootstrap, que gera uma grande quantidade de inferências e compara a proporção de similaridade entre elas.

Uma árvore filogenética pode ser *enraizada* ou *não-enraizada*. Em uma árvore enraizada, temos um referencial - a raiz - que indica o ancestral comum mais recente entre os táxons (Fig. 2.1).

Para nossos propósitos, é interessante notar que o número de árvores bifurcadas possíveis que podemos obter é função do número da táxons de que dispomos.

Especificamente, temos que para árvores enraizadas, o número de árvore possíveis segue a fórmula

$$\frac{(2n-3)!}{2^{n-2}(n-2)!} \tag{2.1}$$

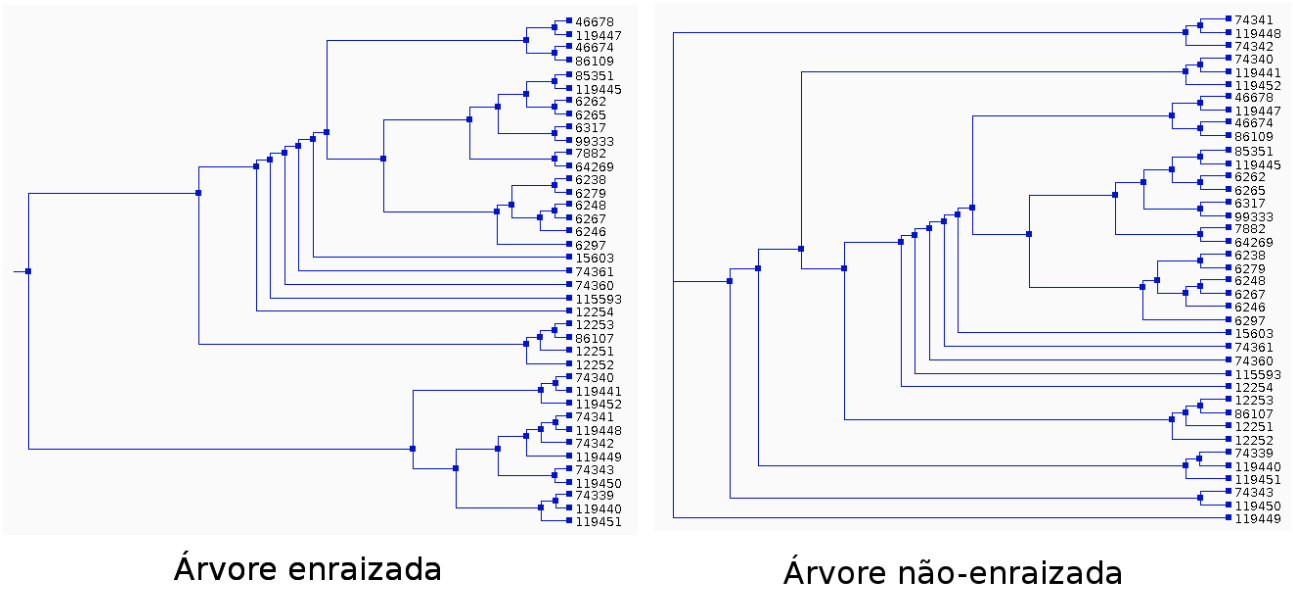


Figura 2.1: Tipos de árvore filogenética em relação à presença de raiz.

e para árvores não-enraizadas, o número de árvores possíveis é

$$\frac{(2n-5)!}{2^{n-3}(n-3)!} \tag{2.2}$$

É fácil ver que ambos são números que crescem muito rapidamente - a primeira fórmula dá 105 árvores para 5 táxons e 34.459.425 para 10 táxons, enquanto a segunda dá 15 árvores para 5 táxons e 2.027.025 árvores para 10 táxons. Isto é importante porque alguns métodos procuram por todo o espaço de árvores para achar a mais adequada!

Os métodos de inferência sempre resultam na relação entre os táxons - sem determinar exatamente onde fica o extremo mais antigo da árvore, ou seja, sem determinar a raiz. Isto é vantajoso principalmente nos métodos exaustivos (onde se examinam todos os arranjos combinatórios possíveis), pois o número de árvore não-enraizadas é menor que o de enraizadas.

Para determinar a raiz de uma árvore não-enraizada, a solução é simples. Adiciona-se ao alinhamento um táxon (ou um pequeno grupo de táxons, para aumentar a verificação de consistência) que se sabe de antemão ser bem mais distante de todos aqueles avaliados. Este táxon, sendo o mais distante de todos, obrigatoriamente deverá aparecer na árvore inferida como tendo uma ligação (um caminho direto) com a raiz verdadeira (Fig. 2.2).

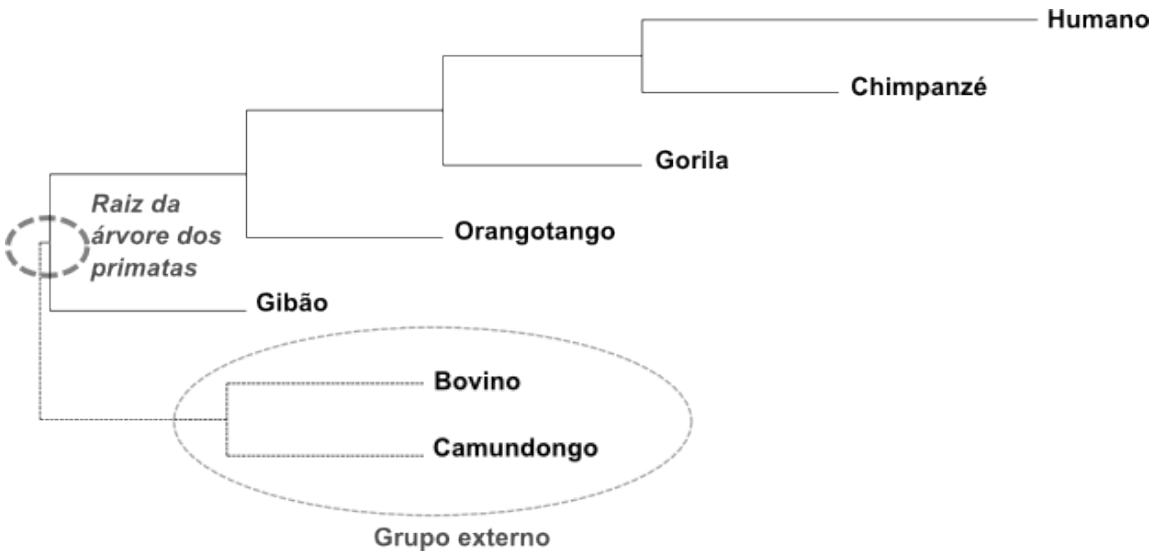


Figura 2.2: Um *outgroup* ou *grupo externo* (Bovino e camundongo) de uma árvore filogenética, usado para localizar a raiz da árvore dos primatas (mostrada na figura). Escolhe-se uma ou mais espécies propositadamente distantes evolutivamente das espécies de interesse pois se sabe que terão de ser colocadas em um galho separado. Caso isso não aconteça, é sinal claro que o método escolhido falhou.

2.1.1 Métodos de Inferência

A inferência filogenética atualmente, ao invés de usar somente dados anatômicos e geográficos, usa principalmente da análise molecular, ou seja, o processamento digital de seqüências de aminoácidos (no caso de proteínas) ou nucleotídeos (no caso de DNA/RNA) para fazer os agrupamentos citados. São desses métodos que nos valeremos, visto que são os que podem ser quantificados com exatidão e automatizados.

Os métodos de inferência filogenética podem ser agrupados em três grupos principais:

- *Métodos baseados em Matriz de Distância* (Neighbor Joining e UPGMA): São métodos que se utilizam do conceito de *distância* entre táxons do alinhamento (gene ou proteína) que serve de dado para a inferência: calcula-se um número, baseado nas diferenças de cada táxon para cada táxon, que denotaria a distância evolutiva entre eles. Baseando-se nesses números, constrói-se uma matriz diagonal e os táxons são agrupados sucessivamente em duplas, dando origem à árvore (Fig. 2.3).

  Os métodos baseados em distância são computacionalmente baratos, crescem pouco em função do número de táxons e comprimento do alinhamento e são adequados para filogenias pequenas, mas têm pouca precisão. O método de Neighbor Joining não tem como pressuposto

uma taxa de evolução constante, mas, por ser um algoritmo "ganancioso" e eliminar caminhos desde o princípio, não percorre todo o espaço de árvores: ele obtém uma árvore curta mas é possível que haja uma árvore mais curta que não tenha sido encontrada.

- *Método da (máxima) parcimônia*: trazido da filogenética tradicional em que se usavam traços anatômicos/morfológicos, o método da Máxima Parcimônia (por vezes chamado somente de Parcimônia) traça uma matriz com os caracteres de interesse na horizontal e os táxons na vertical (Fig. 2.3), e percorre o espaço de árvores procurando as que atendem o critério de otimalidade que respeite os agrupamentos de táxons com as maiores coincidências de caracteres, recursivamente. Considera-se nesse caso cada nucleotídeo ou aminoácido como um caráter particular para fazer o cálculo. É um método que costuma dar melhores resultados que Neighbor Joining mas por outro lado é mais propenso a artefatos (por questões como seleção convergente, inversões, evolução lateral e outros fatores) e, diferente de neighbor joining, não é *estatisticamente consistente*, ou seja, não tem garantias de encontrar a árvore correta com suficientes dados. Por percorrer todo o espaço de árvores, é um método mais caro do que Neighbor-Joining e também cresce em tempo computacional mais depressa do que esse.

- *Método da Máxima Verossimilhança*: Como o método da parcimônia, também é um método exaustivo, percorrendo todo o espaço de árvores. Diferentemente da Máxima Parcimônia, entretanto, o critério de otimalidade da Máxima Verossimilhança é mais consistente: ele percorre todo o espaço de árvores e calcula a probabilidade, para cada árvore, de que, com aquele conjunto de dados, ela tenha resultado, de acordo com o modelo escolhido (que pode ter, por exemplo, diferentes taxas de transição e transversão entre nucleotídeos). Guarda então a árvore com a maior probabilidade calculada e a apresenta como resultado.
  É um método mais caro do que o de Máxima Parcimônia - o cálculo da probabilidade de cada árvore já é, por si só, caro. No entanto, também tem muito menos pressupostos e pode ser relativo ao modelo de evolução que se sabe sobre aqueles táxons. Tende a ter muito menos artefatos do que o da Máxima Verossimilhança e chegar mais facilmente à árvore certa. É também estatisticamente consistente e dispõe de boas heurísticas para chegar mais rapidamente à árvore mais provável (Russo et al. 1996).

Não é obviamente aparente pela descrição dos métodos qual é o melhor tipo de entrada para seu processamento, isto é, percebem-se as diferenças para o cálculo de filogenias para o mesmo tipo

	C1	C2	C3	C4
OTU 1	1	1	0	0
OTU 2	0	0	1	0
OTU 3	1	1	0	0
OTU 4	0	1	1	1
OTU 5	1	1	1	0

Matriz de Caracteres (Ausentes ou Presentes)

	OTU 1	OTU 2	OTU 3	OTU 4	OTU 5
OTU 1	0				
OTU 2	3	0			
OTU 3	2	1	0		
OTU 4	3	3	2	0	
OTU 5	1	3	3	2	0

Matriz de Distâncias

Figura 2.3: Matriz de distância (abaixo) e Matriz de caracteres (acima). Na Matriz de caracteres, as colunas são cada caracter (asas, cauda, corpo segmentado, etc.) e as linhas são cada *OTU*, Unidade Taxonômica Operacional, que denota gene ou espécie de que tratamos. As células terão "0" (ausente) ou "1" (presente) de acordo com a existência da característica na OTU de interesse. No caso de dados biológicos digitais como DNA ou proteína, a presença ou ausência será em relação a um nucleotídeo ou aminoácido em determinado sítio. No caso da matriz de distâncias, temos um "placar" calculado para representar a distância filogenética ou evolutiva entre os táxons considerados (tanto linhas quanto colunas são as OTUs, portanto temos uma matriz triangular).

de dado; mas a pergunta deste trabalho consiste em saber com quais dados essenciais mínimos podemos ter maior qualidade, considerando os métodos descritos. Há muitos parâmetros que poderiam hipoteticamente melhor a qualidade, entre eles aumento do número de amostras, qualidade do seqüenciamento e tamanho do trecho alinhado.

O tamanho do trecho tem importância especial para a praticidade do estudo e pode ser considerado relativamente independente de outros fatores. Fatores também influentes são a variabilidade das seqüências e a forma como essa variabilidade se distribui ao longo do filamento. É necessário classificar de acordo com o método de construção da árvore, visto que os métodos conhecidos variam em eficiência de acordo com o número de OTUs (*Operational Taxonomic Units*) (Servedio 1998). Ocasionalmente uma filogenia de determinados organismos tem que ser construída a partir de genes incompletos (Hannula e Hanninen 2007).

A construção dessas árvores é feita por programas de computador especializados, por causa da alta quantidade de contas e tomadas de decisão baseadas em regras.

Um fator muito importante é que é possível automatizar o processo, ou seja, dado um conjunto de seqüências, é possível realizar o alinhamento, a escolha de um modelo e a construção da árvore de forma automática, pois os dados podem ser avaliados previamente à construção e ajudarem na decisão de que seqüências ou partes de seqüências devem ser realmente usadas, bem como que partes podem ser consideradas confiáveis. De modo semelhante, a possibilidade de automatização também possibilita testar vários modelos e trechos e compará-los, verificando a fidedignidade da análise prévia.

2.1.2 Homologia, parálogos e ortólogos

Em termos evolucionários, homologia se refere a quaisquer similaridades entre características de organismos causadas por sua ancestralidade compartilhada. Na genética, homologia se refere a similaridades de seqüências de RNA ou DNA (ou as correspondentes seqüências protéicas): seqüências de DNA são geralmente similares e não idênticas e essa similaridade, presume-se, é oriunda de um ancestral comum.

O conceito de homologia se contrasta com o de *analogia*, que se refere a quando duas estruturas cumprem funções idênticas ou semelhantes por mecanismos similares mas evoluíram separadamente, em um processo conhecido como *evolução convergente*.

Temos dois tipos principais de homologia: *ortologia* e *paralogia*.

- Seqüências homólogas são chamadas de *ortólogas* se foram separadas por um evento de especiação: quando uma espécie diverge em duas espécies diferentes, as cópias (geralmente ligeiramente divergentes) de um gene particular nas espécies resultantes são denominados genes ortólogos.
  A mais robusta evidência da ortologia de genes semelhantes em espécies diferentes é a análise filogenética da linhagem do gene. Genes que pertencem a um mesmo clado são ortólogos, descendentes de um ancestral comum. Genes ortólogos comumente - mas nem sempre - têm a mesma função.
  Seqüências ortólogas disponibilizam informação bastante útil para a classificação taxonômica e estudos filogenéticos de um organismo. O padrão de divergência genética pode ser usado para traçar o grau de parentesco dos organismos. Dois organismos cuja ancestralidade comum seja recente são propensos a apresentar seqüências muito semelhantes nos genes ortólogos; organismos mais distantes tenderão a apresentar maior grau de divergências nessas seqüências.

- Seqüências homólogas são chamadas de *parálogas* se foram separadas por um evento de *duplicação de gene*, ou seja, um gene do organismo foi duplicado de forma a ocupar duas posições diferentes no mesmo genoma.
  Um conjunto de seqüências que sejam parálogas algumas vezes têm a mesma função ou similar, mas outras vezes não têm: devido à falta de pressão seletiva sob uma cópia do gene duplicado - por causa da redundância de função -, ela fica livre para sofrer mutações e adquirir novas funções no organismo.
  Seqüências parálogas mostram informações valiosas de como os genomas evoluem. Os genes de mioglobina codificadora e hemoglobina são considerados parálogos antigos. As quatro classes de hemoglobina (A, A2, B e F) são parálogas entre si.
  Genes parálogos podem pertencer à mesma espécie, mas isso não é necessário. Por exemplo, o gene de hemoglobina dos humanos e o gene de mioglobina dos chimpanzés são parálogos. Isso é um problema comum na bioinformática: Quando genomas de diferentes espécies são seqüenciados e genes homólogos são encontrados, não é possível concluir de pronto se esses genes têm função similar, pois podem ser parálogos cuja função divergiu.

Uma vez que tenham sido escolhidos somente genes ortólogos para fazer as análises deste trabalho (Matioli e Russo 2001), as filogenias dos genes deverão coincidir com as filogenias das espécies.

2.1.3 Bootstrapping e medidas de qualidade

Um dos mais usados modos de se medir a confiabilidade de uma árvore inferida a partir de quaisquer dos métodos citados é o teste chamado *bootstrap* ou *bootstrapping* proposto por Felsenstein(Felsenstein 1985). Se há *m* seqüências, cada uma com *n* nucleotídeos (ou códons ou aminoácidos), uma árvore filogenética pode ser reconstruída usando um método de inferência. Trata-se o alinhamento como uma matriz *nxm* e colunas da matriz são aleatoriamente escolhidas e duplicadas em outras colunas da matriz, criando um novo alinhamento de mesmo tamanho. Uma nova árvore é inferida a partir destas novas seqüências usando o mesmo método e parâmetros.

A seguir a topologia desta nova árvore é comparada com a da árvore original. A cada galho interior da árvore de bootstrap que for diferente da árvore original é dado uma pontuação de 0; todos os outros ganham a pontuação 1. Este procedimento de reamostragem dos sítios e reconstrução da árvore é repetido várias vezes, tradicionalmente centenas ou milhares de vezes por árvore inferida (caracterizando o processo como computacionalmente caro), e a porcentagem de vezes que cada galho interior ganha o valor de 1 é anotado. Este é o "valor de bootstrap".

Tem sido comum no estudo de filogenia usar o bootstrap como medida de qualidade de uma árvore, com o pressuposto de que se um galho tem um valor de bootstrap de 95% ou maior, ele está biologicamente correto. No entanto este uso tradicional tem muitos problemas e mesmo a literatura existente sobre o assunto já mostra que o bootstrap, embora seja uma boa medida de consistência interna dos dados, tem muito pouco a dizer sobre a corretude biológica da árvore inferida (Sanderson 1995).

A meta originalmente proposta para o bootstrap é que o processo de amostragem com substituição simulasse a obtenção de dados adicionais na natureza para os táxons em questão. Dessa forma os 100 ou mais conjuntos de dados obtidos por essa técnica simulam os diferentes conjuntos de dados que poderiam ter sido amostrados se mais e mais dados tivessem sido repetidamente coletados. Em suma, o bootstrap parte da distribuição de caracteres observada em uma amostra para inferir a distribuição apresentada pelos dados reais; dessa forma busca quantificar até que ponto os grupos inferidos a partir da matriz de dados são conseqüências de artefatos amostrais.

É importante notar que altos valores de bootstrap podem ser obtidos para inferências incorretas: isto acontece em situações nas quais os dados disponíveis e o método empregado levam todas as amostragens a inferirem árvores erradas (os sucessivos eventos de amostragem geram matrizes que também endossam a árvore incorreta).

As principais situações que levam um bootstrap a estar incorreto são: taxas de mudança muito desiguais; taxas de mudança são muito altas tornando a informação aleatória e aumentando a saturação; e viés sistemático nos dados, conseqüência, por exemplo, da ocorrência de seleção (Hillis e Bull 1993).

Um fator a notar sobre bootstraps: como são medidas *internas* relativas aos próprios dados, é consenso na literatura que não são comparáveis entre filogenias diferentes. Então, ainda que se espere que ao se ter uma filogenia com galhos com mais de 95% de concordância de bootstraps indique boa qualidade, não se espera que uma árvore com galhos de mais de 95% de concordância de bootstraps para certas espécies tenha melhor qualidade que, por exemplo, uma árvore com galhos de mais de 90% de concordância para espécies diferentes.

2.1.4 Distâncias entre árvores filogenéticas

É comum que a filogenia de organismos, populações e espécies não seja determinada somente pelos dados moleculares, mas também por informações morfológicas, bioquímicas e geográficas. Não raro, temos a inferência digital de uma árvore filogenética complementada ou mesmo corrigida com essas informações adicionais. A isso se chama de dar "significado biológico" a uma filogenia, pois, para *provar* que os cálculos da inferência estão certos, não se podem usar os mesmos dados: a verificação independente por outras fontes é necessária (Bininda-Emonds 2000).

Se os dados incluírem uma árvore modelo já bastante corroborada pela literatura, podemos inferir uma árvore usando nossos métodos e procurar sabermos o quão próxima a inferência resultante se posicionou dessa árvore ideal, caso a inferência não seja idêntica (o que será raro para um número grande de táxons). Essa proximidade, ou medida de "semelhança" com a árvore original, é dada por um número chamado de *distância* entre filogenias.

A grandeza mais comumente usada para expressar a distância entre árvores é a chamada distância "SPR" (Subtree Prune and Regraft)(Hickey et al. 2008), que consiste no número mínimo de operações de "recorte" e "colagem" de galhos que precisamos realizar em uma árvore para que ela se transforme na árvore de referência. Uma operação atômica (singular) feita entre duas árvores pode ser demonstrada na figura 2.4.

É interessante notar que a determinação da distância entre árvores é um problema NP-difícil, ou seja, o único jeito de testar deterministicamente a solução é tentar todas as possíveis e compará-las. Isso torna o problema computacionalmente muito caro, embora haja heurísticas disponíveis para

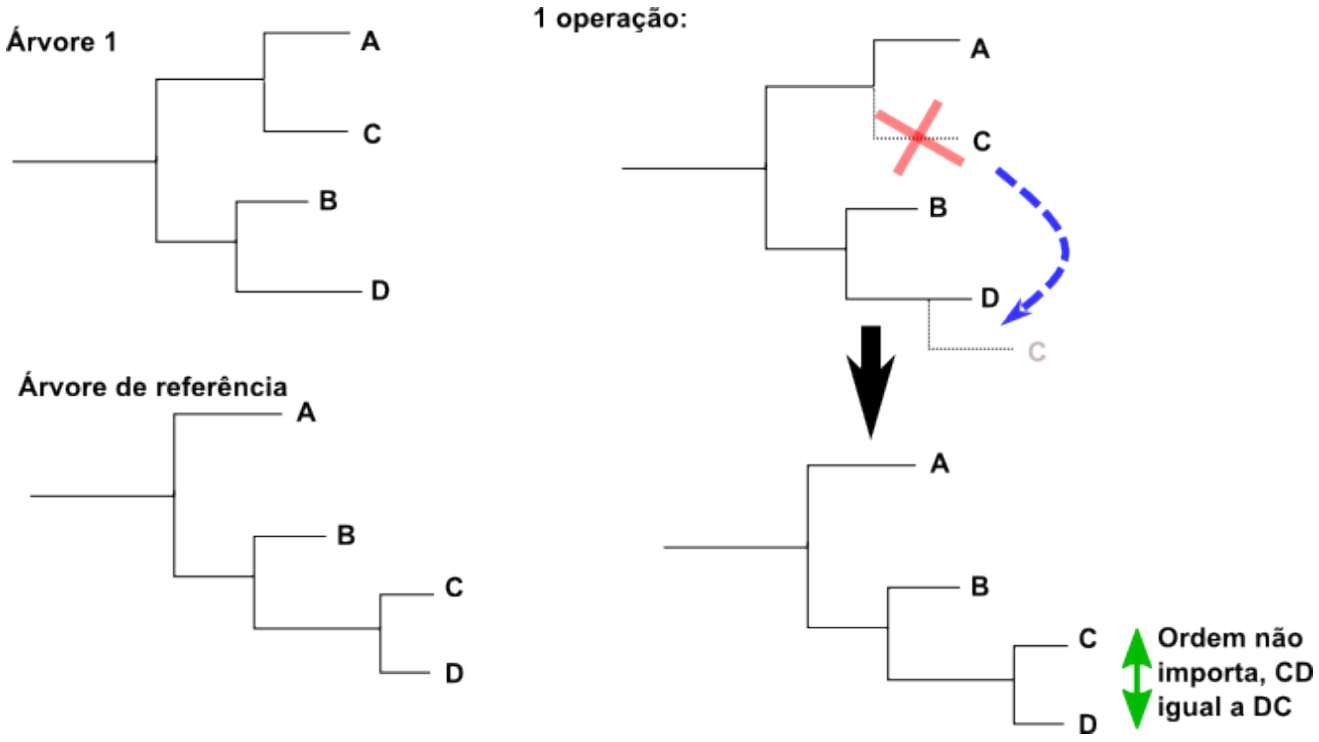


Figura 2.4: Operação de recorte e colagem feita para transformar a Árvore 1 na Árvore de referência. Na árvore 1, o galho "C" é removido de sua localização adjacente a "A" e enxertado a um ponto de ligação com "D", gerando uma árvore equivalente à de referência. Como foi necessária apenas uma operação de recorte e colagem, diz-se que as árvore são distantes em "1" unidade SPR.

melhorar seu tempo de execução.

2.2 Introdução - Teoria da Informação e Entropia

A Teoria da informação ou Teoria matemática da comunicação é um ramo da teoria da probabilidade e da matemática estatística que lida com sistemas de comunicação, transmissão de dados, criptografia, codificação, teoria do ruído, correção de erros, compressão de dados, etc. Ela não deve ser confundida com tecnologia da informação e biblioteconomia.

Claude E. Shannon (1916-2001) é conhecido como "o pai da teoria da informação". Sua teoria foi a primeira a considerar comunicação como um problema matemático rigorosamente embasado na estatística e deu aos engenheiros da comunicação um modo de determinar a capacidade de um canal de comunicação em termos de ocorrência de bits. A teoria não se preocupa com a semântica dos dados, mas pode envolver aspectos relacionados com a perda de informação na compressão e na transmissão de mensagens com ruído no canal.

É geralmente aceito que a moderna disciplina da teoria da informação começou com duas publicações: a do artigo científico de Shannon intitulado *Teoria Matemática da Comunicação* ("*A Mathematical Theory of Communication*"), no Bell System Technical Journal, em julho e outubro de 1948 (Shannon 2001); e do livro de Shannon em co-autoria com o também engenheiro estadunidense Warren Weaver (1894-1978), intitulado *Teoria Matemática da Comunicação* (*The Mathematical Theory of Communication*), e contendo reimpressões do artigo científico anterior de forma acessível também a não-especialistas - popularizando deste modo os conceitos.

2.2.1 Entropia de Shannon

No processo de desenvolvimento de uma teoria da comunicação que pudesse ser aplicada por engenheiros eletricistas para projetar sistemas de telecomunicação melhores, Shannon definiu uma medida chamada de entropia, definida como:

$$H(X) = -\sum_{i=1}^{n} p(x_i) log_b p(x_i) \tag{2.3}$$

onde log é o logaritmo - cujas bases mais freqüentes são 2 e número de Neper (logaritmo natural, o usado em nosso software) - que determina o grau de caoticidade da distribuição de probabilidade $p(x_i)$ e pode ser usada para determinar a capacidade do canal necessária para transmitir a informação. $p(x_i)$ é a probabilidade do símbolo x na posição i.

A medida de entropia de Shannon passou a ser considerada como uma medida da informação contida numa mensagem, em oposição à parte da mensagem que é estritamente determinada (portanto prevísivel) por estruturas inerentes, como por exemplo a redundância da estrutura das linguagens ou das propriedades estatísticas de uma linguagem, relacionadas às frequências de ocorrência de diferentes letras (monemas) ou de pares, trios, (fonemas) etc., de palavras.

A entropia como definida por Shannon está intimamente relacionada à entropia definida por físicos. Boltzmann e Gibbs fizeram um trabalho considerável sobre termodinâmica estatística. Este trabalho foi a inspiração para se adotar o termo entropia em teoria da informação. Há uma profunda relação entre entropia nos sentidos termodinâmico e informacional. Por exemplo, o "*demônio de Maxwell*" necessita de informações para reverter a entropia termodinâmica e a obtenção dessas informações equilibra exatamente o ganho termodinâmico que o demônio alcançaria de outro modo.

A teoria da informação de Shannon é apropriada para medir incerteza sobre um espaço desordenado. Uma medida alternativa de informação foi criada por Fisher para medir incerteza sobre um espaço ordenado. Por exemplo, a informação de Shannon é usada sobre um espaço de letras do alfabeto, já que letras não tem 'distâncias' entre elas. Para informação sobre valores de parâmetros contínuos, como as alturas de pessoas, a informação de Fisher é usada, já que tamanhos estimados tem uma distância bem definida.

2.2.2 Entropia de Shannon para alinhamentos

Sendo a entropia a medida mais sensível para medir a diversidade de um sistema, é natural que se aplicasse esse conceito para a bioinformática. Uma maneira de entendê-la em termos de seqüências é que se uma amostra é tirada de uma população maior, a entropia de Shannon pode ser considerada uma medida indicativa de sua habilidade de inferir que aminoácidos ou nucleotídeos estariam na próxima seqüência que se tirasse da população, baseada em sua amostra anterior.

Imagine-se, por exemplo, que tivéssemos interesse em uma determinada posição onde mutações possam conferir resistências a drogas. O conhecimento das freqüências dos aminoácidos nesta posição tirada de populações resistentes e susceptíveis permitiria que se calculassem as entropias de Shannon, uma reflexão da qualidade da estimativa de quais aminoácidos seriam os próximos em uma amostra desconhecida extraída da população. É possível estreitar ou definir sítios de resistência a droga em genomas complexos ao definir posições em proteínas que são invariáveis em populações susceptíveis a drogas (baixa entropia) mas altamente variáveis em populações resistentes (entropia mais alta). Mesmo que o aminoácido de consenso seja o mesmo em ambos os conjuntos, sítios que variam mais podem ser identificados quando lidamos com vírus resistentes.

Quando esta medida de incerteza é usada como estratégia para quantificar a variabilidade de seqüência em uma coluna em um alinhamento de seqüências, ela incorporas ambas a freqüência (por exemplo, uma coluna que tenha 50% A e 50% T tem maior entropia do que uma coluna que é 90% A e 10% T) e o número de possibilidades (uma coluna que é 90% A, 5% T e 5% G tem maior entropia do que uma com 90% A e 10% T). Uma coluna invariante tem uma entropia zero.

A entropia máxima é dependente do número de variáveis discretas no seu conjunto; por exemplo, se você está considerando DNA, você pode ter A, C, G e T, e sua máxima entropia ocorreria se todas estivessem presentes em igual freqüência, 25% cada. Esta entropia máxima é o logaritmo do número de variantes na base e – a entropia máxima em um ponto do alinhamento é 1,39 no caso

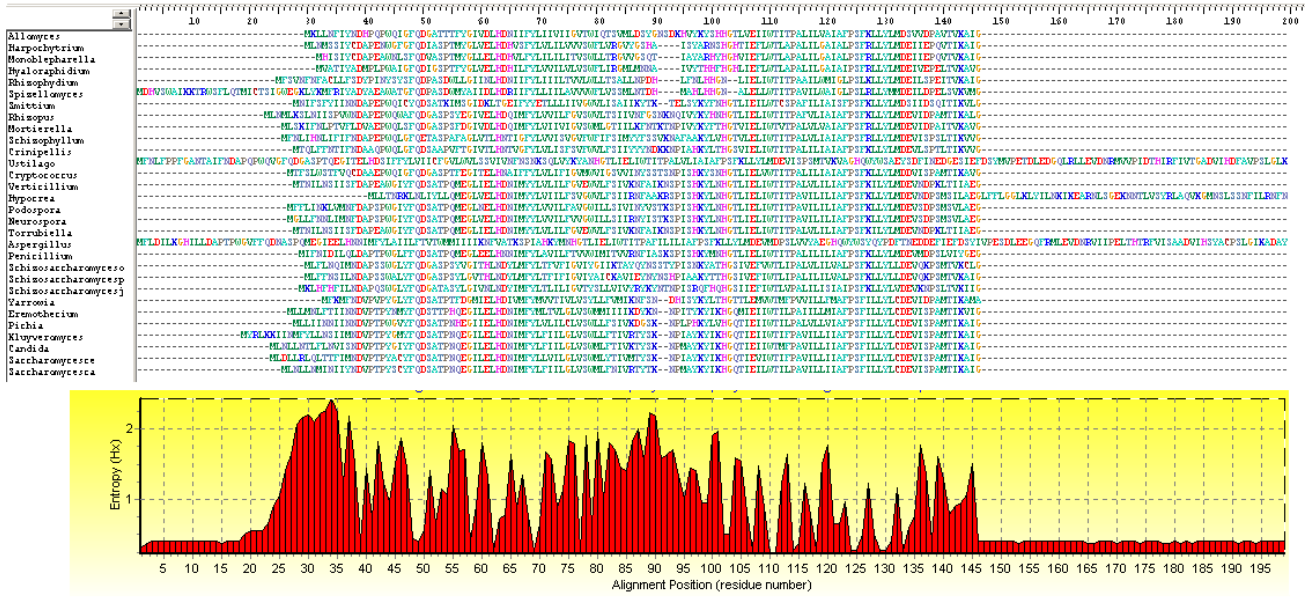


Figura 2.5: Representação de um alinhamento da proteína do gene mitocondrial *cox2* de espécies de fungos e a entropia correspondente a cada sítio no programa *bioedit*. Note que apesar da alta variabilidade de alguns pontos e a máxima entropia possível de 3,05, temos entropias bem menores do que esse máximo.

dos nucleotídeos ($log4$), 1,61 ($log5$) se considerarmos um quinto estado para representar as "*gaps*" do alinhamento e 1,79 ($log6$) caso tenhamos os 4 nucleotídeos, um estado para GAP e um para nucleotídeo desconhecido – este último caso usado no programa que construímos. No caso de 20 aminoácidos, temos entropia máxima de 3 ou, se considerarmos os "*gaps*" e desconhecidos como estados à parte, 3,09 (Fig. 2.5). A entropia não tem unidade.

3 Objetivos

"*Toda ação deve ter um objetivo.*

Devemos sofrer, devemos trabalhar, devemos pagar nosso lugar no espetáculo,

mas é para ver; ou ao menos para que um dia outros vejam."
Jules Poincaré

O objetivo deste projeto é a identificação de tamanhos de alinhamentos de material genético (proteína ou DNA) inferiores ao tamanho do alinhamento final com os genes completos para determinar uma faixa de tamanhos em que se poderá, mesmo com dados incompletos, obter uma filogenia adequada (isto é, a mais próxima possível da "real") dos táxons em estudo.

Adicionalmente, será usada a entropia de Shannon, uma medida de informação de cada posição do alinhamento, para aprimorar a determinação dos trechos que poderão ser usados para se obter esta filogenia ótima.

Fatores que diminuem a confiança da estimativa, como GAPs e artefatos posicionais, serão eliminados ou minimizados.

As iterações entre os tamanhos serão feitas por um programa "open source" construído para este trabalho que avalia a inferência de filogenias resultante para cada tamanho e produz os gráficos e dados que nos permitirão chegar à conclusão.

4 Metodologia e ferramentas

"*Um objetivo sem um plano é apenas um desejo.*"
Antoine de Saint-Exupery

4.1 Ferramentas

O campo da bioinformática, que é a computação aplicada à biologia molecular e genética, é bastante vasto. Por isso mesmo, a disponibilidade de ferramentas e soluções é gigantesca e variada. No entanto, existem muitas soluções bastante semelhantes para alguns problemas pontuais e poucas soluções realmente flexíveis, que satisfaça às nossas necessidades para este trabalho.

A estratégia deste projeto consistiu em aproveitar da altíssima disponibilidade de ferramentas pontuais e combiná-las para formar o necessário para conseguir nossos dados. Deste modo, foi possível montar e desenvolver uma solução própria - mais especificamente, criar um software que, alimentado com certos dados, pudesse processá-los e nos entregar os gráficos e medidas de que necessitamos para chegar à conclusão.

4.1.1 Equipamento

Utilizamos um computador Intel Xeon 3GHz com 4 núcleos e 16 GB de RAM. Para segurança e continuidade das execuções, as executamos dentro de uma máquina virtual "XEN" que utilizava apenas 2 núcleos e 2 GB de RAM para o processamento. O algoritmo do programa mostrou-se bastante linear, utilizando apenas um processador por vez, e quantidade de RAM não foi um fator importante (a qualquer momento, o programa utilizava bem menos RAM do que o disponível, não passando de 300 MB). Como disco, foi utilizado um disco virtual de 12 GB (10 GB livres para dados) por cima de um sistema de arquivos compartilhado em RAID-5. Velocidade de escrita em disco também não é um fator importante para o cálculo dos dados.

O método usado para melhor velocidade foi ter apenas duas execuções por vez, de modo que cada uma usasse um processador inteiro, em diretórios diferentes do sistema de arquivos virtualizado.

4.1.2 Plataforma

Tradicionalmente a plataforma UNIX (incluindo nisso o sistema Linux) já é utilizada para a maioria das tarefas de bioinformática que exigem poder de computação alto e disponibilidade contínua. Tem, portanto, boa disponibilidade de software disponíveis, especialmente softwares livres e grátis, que nos interessam pela economia que representam.

Entre várias distribuições, foi utilizado um Linux Fedora 10 para a execução dos trabalhos "produtivos" em cima dos dados e um Ubuntu 8.10 como plataforma de desenvolvimento e testes do software. Funcionando nessas duas plataformas diferentes, garante-se que o programa rode em outros tipos de Linux e até outros Unixes diferentes (como FreeBSD).

4.1.3 Ferramenta interativa de filogenia

Antes do desenvolvimento da ferramenta e mesmo durante seus testes, utilizamos um programa de filogenia livremente disponível para Unix e Windows denominado HYPHY (Pond e Muse 2005) (Fig. 4.1). Este programa é interativo e guiado por menus e apresenta um visualizador/editor de alinhamentos razoável. O programa HYPHY serviu para a escolha dos dados e conferência das árvores finais do software desenvolvido.

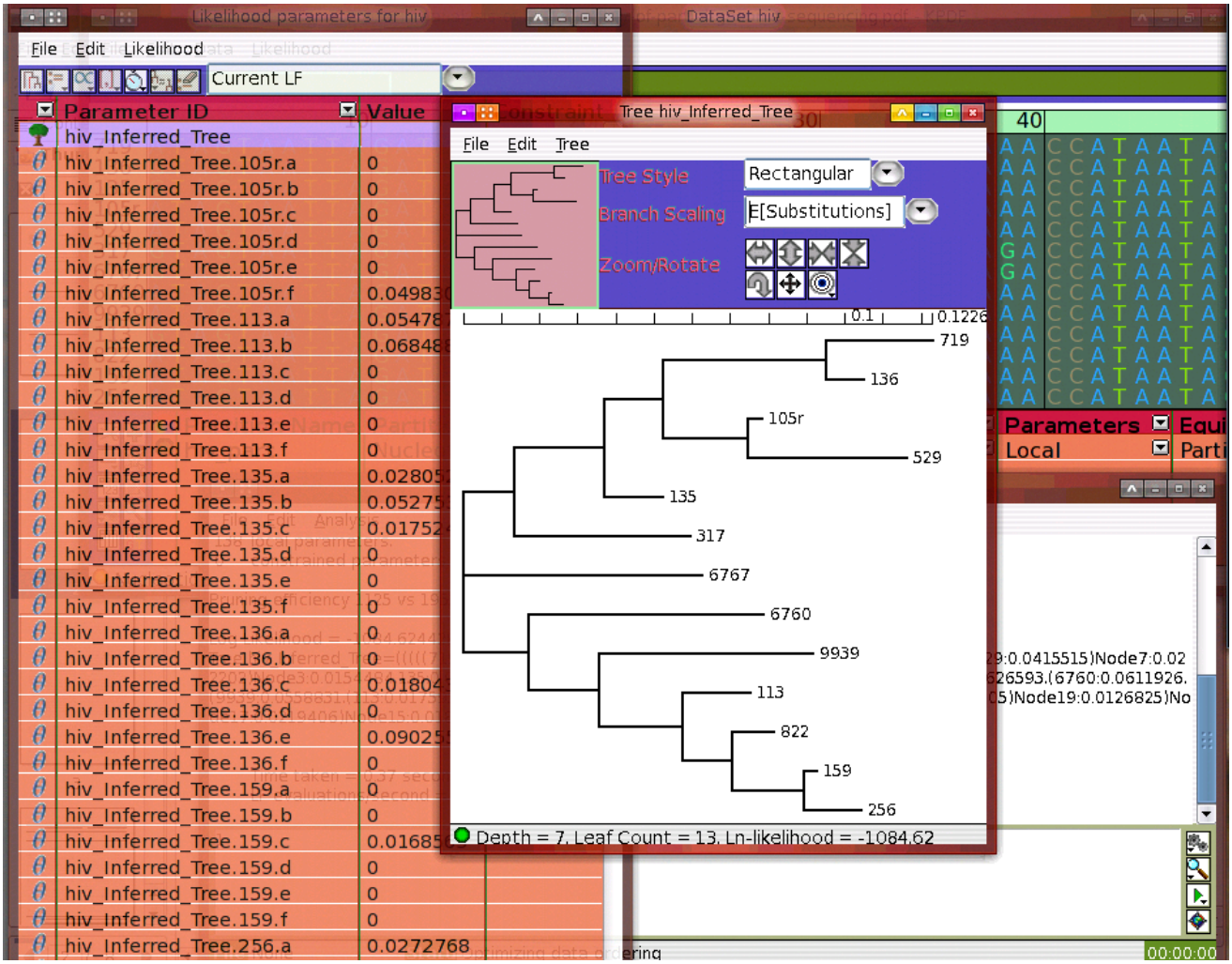


Figura 4.1: Exemplo de tela do programa interativo HyPhy, mostrando tanto uma árvore de exemplo com táxons numerados inferida pelo método de Neighbor-Joining quanto a edição de alinhamentos ao fundo. Outros diálogos que aparecem são o "terminal", onde se vêem as saídas detalhadas das operações executadas (direita, abaixo) e o resultado detalhado da distância calculada de cada nó para a inferência da árvore (esquerda).

4.1.4 Ferramenta de edição de alinhamentos

Como a matéria-prima do estudo são alinhamentos - de DNA, RNA ou proteína -, havia a necessidade de um programa que fosse bastante completo, fizesse uma série de análises e filtragens automatizadas e nos permitisse editar precisamente o alinhamento para passarmos para o nosso software. O editor incluso na ferramenta HyPhy não era suficiente.

Nota-se que na disciplina "filogenética" é raro ter-se a captação de dados biológicos sendo alimentada crua em um *software* de computador. É prática comum, dada a especificidade dos dados, tratá-los de formas que se julguem adequadas para torná-lo processável. Devido às incompletudes e irregularidades dos dados escolhidos, houve a necessidade de seguir este expediente, ainda que o *software* desenvolvido tenha, ele mesmo, uma série de filtros para selecionar os dados de entrada e ter menos ruído.

Foi escolhido o software BioEdit (Hall 1999), um software para Windows que rodou sem problemas emulado no Linux escolhido como plataforma de teste e desenvolvimento. Este software permite não somente permite edição fina de todos os aspectos de um alinhamento de DNA, RNA ou proteína, como também levantar uma série de estatísticas automaticamente sobre o alinhamento e passar uma série de filtros (de retiradas de *gaps*, por exemplo) de que necessitamos. Uma das estatísticas interessantes do BioEdit é a entropia que mencionamos no capítulo 2, a qual usamos para comparar com a saída do nosso software.

Foram utilizados também, em conjunto com o bioedit, pequenos scripts em perl para ajustar e converter pequenas massas de dados dos alinhamentos, além do tradicional software ClustalW (Aiyar 1999).

4.1.5 Linguagem e programas auxiliares

Os seguinte critérios foram utilizados na escolha de uma linguagem adequada para construir o *software*:

- Ser livremente disponível;
- Ter disponível, também livremente, uma API de bioinformática com módulos razoavelmente completos de processamento de alinhamentos e filogenias;
- Ter a capacidade de produzir gráficos científicos dos dados;

- Ter facilidade de integração com programas auxiliares que inevitavelmente são necessários usar para processar alguns dados;
- Ser relativamente bem conhecida pelos profissionais de bioinformática e tivesse casos de uso comprovados.

Para tal, foi escolhida a linguagem perl, uma linguagem *script* multiplataforma bem estabelecida que roda com grande desempenho em todos os Unixes e tem disponível para si uma API bem completa de rotinas de bioinformática, tendo entre estas partes bem avançadas e completas de tratamento de alinhamentos e filogenias. Esta API tem nome *bioperl* (Stajich 2002) e sítio web próprio em http://www.bioperl.org.

Apesar do avançado e maduro estado desta API (que foi o principal componente de software do projeto Genoma Humano), já se sabia previamente que certos processamentos precisariam ser feitos com ferramentas externas, tanto por motivos de desempenho (uma linguagem script, mesmo uma pré-compilada, sempre é menos eficiente que uma linguagem compilada) quanto de funcionalidades (certos algoritmos heurísticos necessários para a inferência de filogenias com Máxima Verossimilhança, por exemplo, não fazem parte da *bioperl*). Usamos os seguintes softwares - todos livremente disponíveis pela Internet:

PhyLIP

PhyLIP é um agregado de pequenos programas que fazem processamento de seqüências e árvores de forma eficiente para geração de filogenias (Felsenstein 2005, Retief 2000). Bastante acoplável, os vários programas são usados para gerar árvores de consenso, calcular bootstraps e inferir árvores pelos métodos de neighbor joining, parcimônia e máxima verossimilhança.

PhyML

PhyML(S e O. 2003) é um software isolado, porém intercambiável com a peça do PhyLIP que constrói filogenias com o método de máxima verossimilhança, no entanto com heurística bastante superior e com excelente descoberta automática de parâmetros. Durante os testes, ficou claro que os algoritmos de máxima verossimilhança do PhyLIP são lentos e insuficientes para este trabalho.

EEEP

EEEP (Efficient Evaluation of Edit Paths) é o programa que faz cálculos da distância entre duas árvores (Beiko e Hamilton 2006). A API bioperl também tem este recurso, mas sua heurística de funcionamento é pobre e não traz bons resultados. Note-se que o EEEP é utilizado pelo nosso software com heurística mínima, caso contrário ele facilmente estouraria os limites de tempo e memória de que dispomos, no entanto mesmo assim consegue um resultado bastante satisfatório.

gnuplot

GNUPLOT é um software gratuito tradicional para plotagem de gráficos científicos e financeiros (Williams et al. 2004). Ele é utilizado pelo calcphyl para produzir todos os seus gráficos, com exceção da plotagem da árvore filogenética.

4.1.6 O software desenvolvido - CalcPhyl

O software é open-source e está hospedado no sítio web do Laboratório de Genômica e Expressão no endereço http://lge.ibi.unicamp.br/calcphyl.

Ainda que o software use três conjuntos de auxiliares binários externos, isto não faz dele um software simples. Muito do trabalho que o programa desempenha consiste em converter e reprocessar entradas e saídas desses auxiliares, tratar erros, levantar estatísticas e escolher opções e fluxos.

O programa tem como entrada um alinhamento de DNA ou proteínas (o tipo é autodetectado) e opcionalmente uma árvore de referência - que, se dada, será usada como gabarito para calcular a distância SPR de cada árvore calculada.

Apesar de ser do interesse da estratégia exploratória do estudo usar um dado simples de entrada - o trecho seqüenciado de rascunho que precisaríamos avaliar -, esta abordagem seria bem menos viável do que a de utilizar o alinhamento já feito porque:

- Este é um problema bem diferente de escolher a melhor filogenia. Um alinhamento tem seus próprios parâmetros de qualidade, sua política de *gaps* e suas idiossincrasias. Fazer um alinhamento bom é um problema em si e se o estudo almejasse resolver dois problemas conjuntamente, poderia não resolver nenhum.

- Existem programas que fazem esta parte de, dado um rascunho, alinhar automaticamente (ou semi-automaticamente) com genes extraídos do genebank ou outros repositórios, gerando um alinhamento em formato compreensível pelo software do nosso estudo (como o Hal do já mencionado *aftol*). Portanto, podem ser acoplados a ele para esta automatização, caso interesse ao usuário.

- Com a parte de criação do alinhamento automatizada, uma das práticas mais ubíquas de estudos de filogenias – a edição manual do alinhamento, importante para gerar uma boa filogenia – seria impedida ou dificultada.

- Mesmo se fosse gerado um alinhamento automaticamente, seria impossível obter também automaticamente uma árvore de referência para ele. E este é justamente o problema que o projeto se propõe a medir.

Existem várias opções que se podem passar ao programa - número de pontos dos gráficos (o que determinará o número de iterações), método de inferência de filogenia, que tipo de método de distância usar, se é necessário usar bootstrapping e neste caso, quantas permutações serão feitas e muitos outros. Caso não sejam explicitamente indicadas, cada opção tem um *default* razoável.

O software então entrará em seu laço principal, fazendo iterações com o alinhamento dado - por tamanho ou, caso tenha sido escolhido assim, por limite inferior de entropia - e gerando, no final da execução, vários arquivos de saída:

- Um arquivo-texto contendo vários dados e estatísticas sobre a execução:
  - Qual o arquivo de entrada utilizado;
  - O arquivo com a árvore de referência utilizada;
  - O método do cálculo de distância SPR (pelo software EEEP ou interno ao bioperl);
  - Qual o nome identificador da execução;
  - Qual formato de saída gráfica (bitmap ou vetorial);
  - Qual o formato da planilha gerada;
  - Se os gráficos apresentam curva bezier *auxiliar;*
  - O diretório de trabalho;
  - O método de iteração (posição ou entropia);

  - O diretório-base dos executáveis do PhyLIP;
  - Se o modo de depuração foi selecionado;
  - O número de permutações para cálculo do bootstrap;
  - O número de pontos do gráfico (e iterações do código);
  - Se a execução preservou ou não arquivos temporários;
  - Se os dados foram embaralhados ou não;
  - Se os terceiros nucleotídeos foram removidos ou não;
  - O valor do filtro de entropia mínima;
  - O método de inferência de filogenia;
  - O raio de transições para transversões;
  - O parâmetro "jumble" utilizado em máxima verossimilhança;
  - O número de resíduos do alinhamento;
  - O comprimento do alinhamento;
  - O número de seqüências (táxons);
  - A porcentagem de identidade do alinhamento;
  - Se o alinhamento tem todas as seqüências de mesmo tamanho ou não;
  - O alfabeto do alinhamento (proteína, DNA, RNA);
  - O tamanho com o maior bootstrap obtido para o alinhamento, incluindo o valor obtido para o bootstrap;
  - Data e hora de início da execução;
  - Data e hora de término da execução;
  - Duração da execução.

- Uma planilha XLS (formato Excel) com todos os dados calculados, consistindo de uma tabela com:

  - O tamanho do alinhamento avaliado em cada iteração;
  - O bootstrap calculado neste tamanho;

  - O número de bootstraps calculados para este tamanho (que pode ser menor que o número de galhos, dado que em alguns casos a informação do alinhamento é insuficiente para inferir uma árvore);
  - O bootstrap "relativo", isto é, o bootstrap dividido pelo número de bootstraps calculados;
  - A distância SPR entre a árvore inferida para este tamanho e a árvore-gabarito;
  - O menor valor de entropia obtido para qualquer ponto do trecho obtido com este tamanho.

- Se foi passado o parâmetro de bootstrap, é traçado o gráfico da árvore filogenética com melhor bootstrap que encontrou.
- Se foi passado o parâmetro de bootstrap, é renderizado o gráfico de medida do bootstrap das árvores por tamanho encontrado.
- Se foi passado o parâmetro de árvore-gabarito, é renderizado o gráfico com a distância SPR em cada tamanho de alinhamento.
- Se foi passado o parâmetro de árvore-gabarito, os arquivos de dados com as melhores distâncias SPR são preservados.
- O gráfico de entropia em cada ponto do alinhamento. É calculado também um segundo gráfico - a entropia ordenada em ordem decrescente em cada sítio, de forma a podermos visualizar claramente a concentração de diversidade que um determinado alinhamento apresenta.

O software tem dois modos de uso:

- Uma linha de comando flexível e poderosa, adequada para trabalhos em *batch* iterados, ou seja, execução em lote tipicamente levando vários dias para completar. (Fig. 4.2)
- Um modo gráfico, mais fácil de usar, porém menos flexível (Fig. 4.3) e adequado apenas para execuções individuais. Serve para a familiarização com as opções do programa e saber como usá-lo: quando se chama a execução, a interface mostra a linha de comando que será usada e termina, deixando a execução rodando em segundo plano.

Para o modo gráfico, é possível usar a interface gráfica remotamente, isto é, vê-la em uma máquina diferente da máquina em que a execução acontece, utilizando do mecanismo de exportação de janelas padrão do Unix. O progresso da execução pode ser acompanhado pelo seu arquivo de log, gravado enquanto os dados são processados.

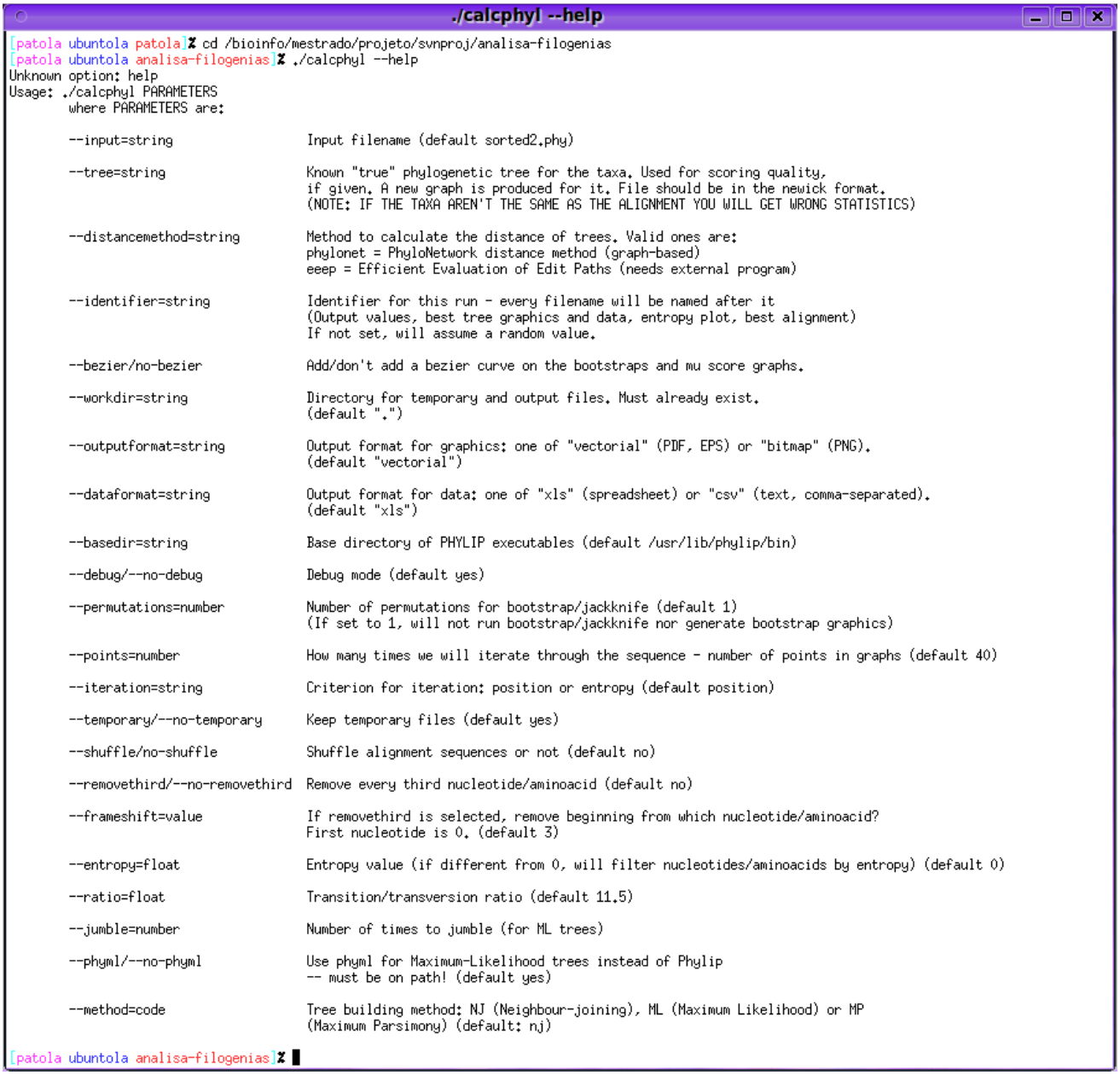


Figura 4.2: Captura de tela das opções de linha de comando do CalcPhyl. Como o programa estará livremente disponível para a comunidade, optou-se por usar o inglês na linha de comando. Algumas opções mais básicas são: “input”, o arquivo de alinhamento de entrada; “workdir”, o diretório de trabalho; “method”, o método de inferência de árvores; “points”, o número de pontos no gráfico; “iteration”, que tipo de iteração (por entropia ou por posição); e “shuffle” para embaralhar o alinhamento.

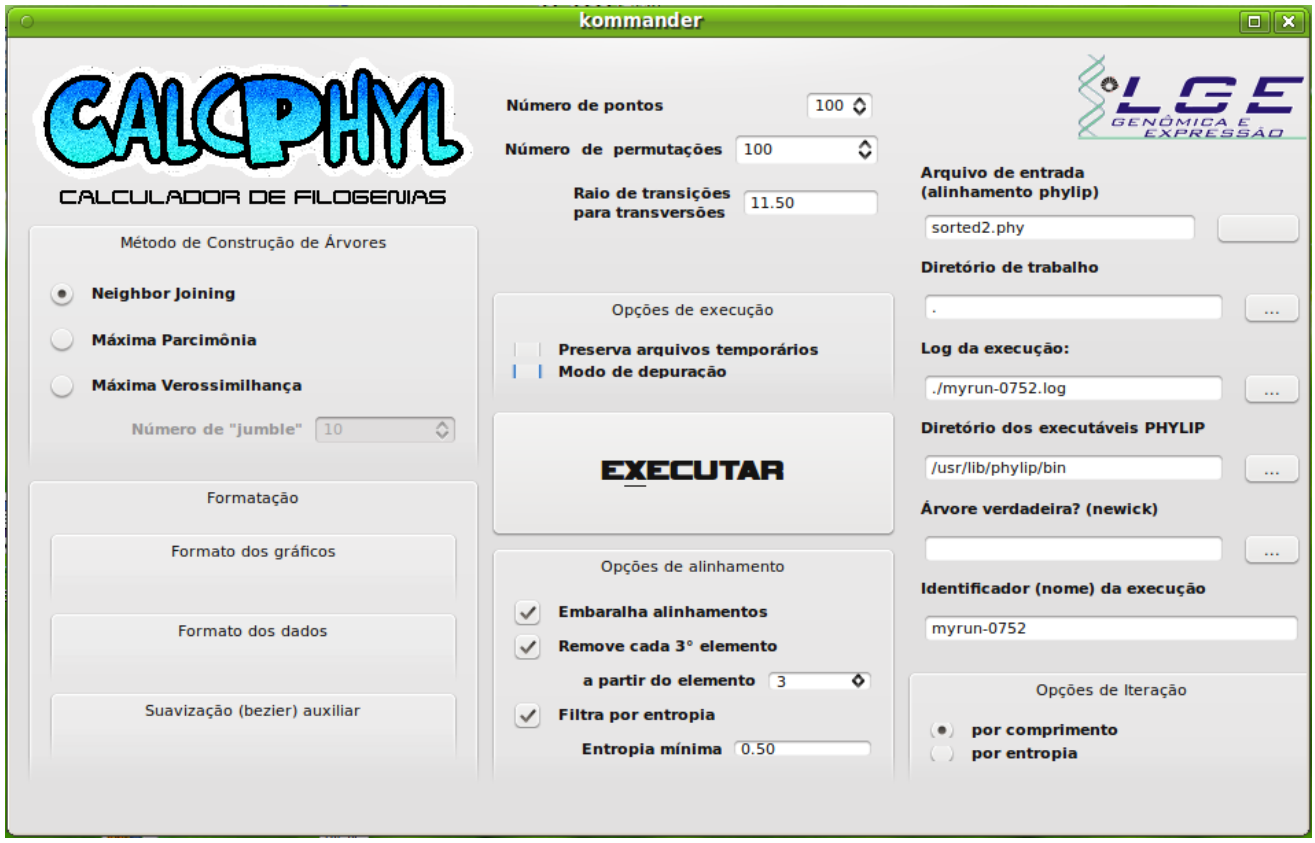


Figura 4.3: Modo gráfico do CalcPhyl. Diferentemente da linha de comando, modos gráficos, por objetivarem maior facilidade de uso, são "internacionalizados", isto é, preparados para funcionar em diferentes linguagens. Neste caso, a interface aparece em português. Podem-se ver as mesmas opções que se vêem na linha de comando, como "método de construção de árvores", "embaralha alinhamentos", "iteração por comprimento (posição) ou por entropia", "Arquivo de entrada" e "diretório de trabalho".

Fluxograma

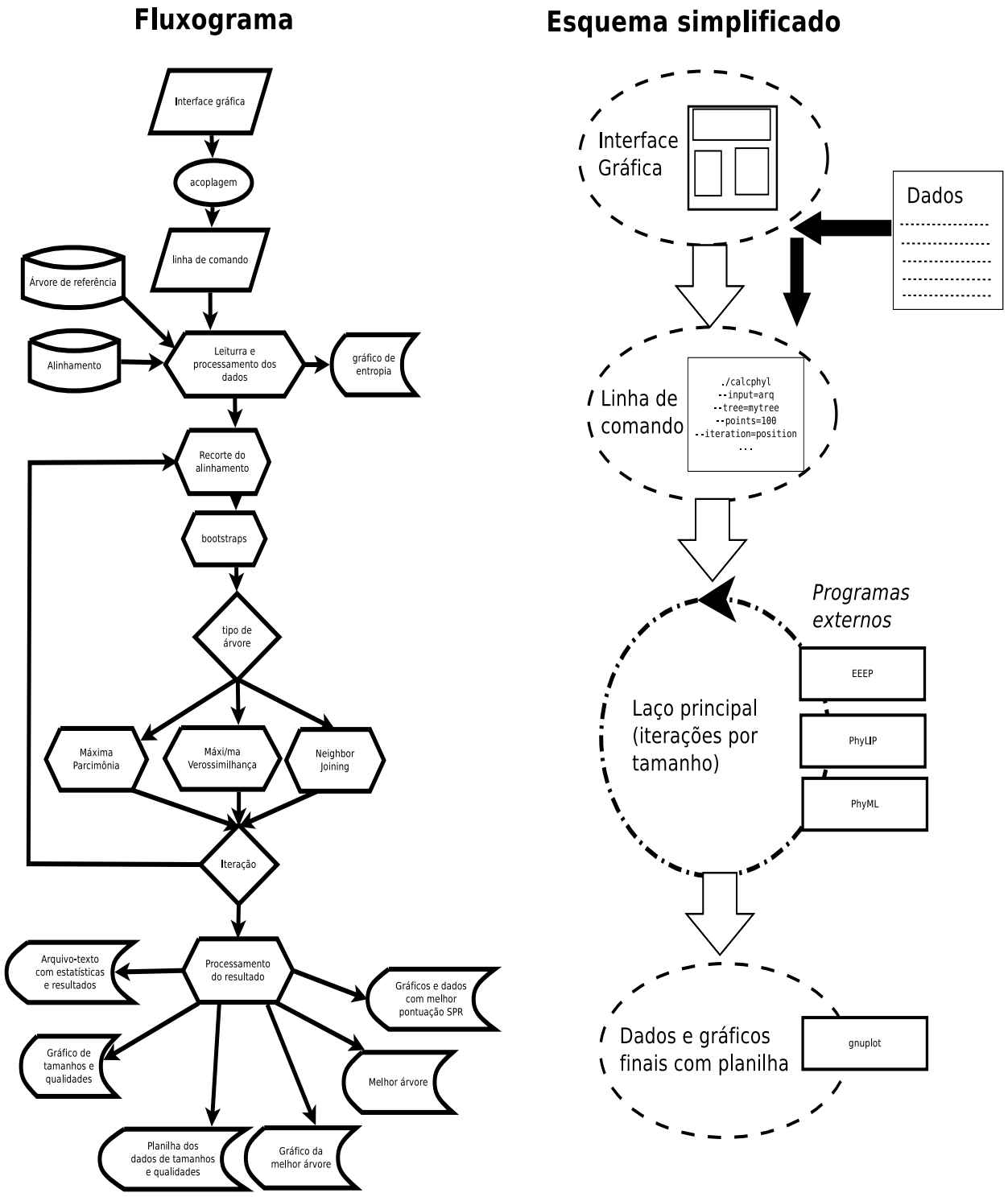


Figura 4.4: Fluxograma de execução do CalcPhyl (à esquerda) e diagrama simplificado do fluxo do programa (à direita). No fluxograma vemos um detalhamento das decisões tomadas pelo software durante a execução assim como o processamento e geração de dados, enquanto no diagrama simplificado temos um desenho intuitivo do processo de execução e dos programas exteriores usados.

Saída do programa (exemplos)

Alguns exemplos de saídas produzidas:

- Valor da entropia em cada ponto do alinhamento (Figura 4.5);
- distância SPR da árvore inferida para a árvore de referência (Figura 4.6);
- a planilha de dados para cada ponto (tamanho) calculado (Figura 4.7);
- A árvore renderizada para o tamanho com melhor bootstrap (Figura 4.8);
- O arquivo final de estatísticas da execução (Figura 4.9);
- O valor da entropia de cada ponto ordenada em ordem decrescente (Figura 4.10).

As saídas gráficas são renderizadas pelo próprio software e suas APIs não necessitando de softwares auxiliares fora os já mencionados.

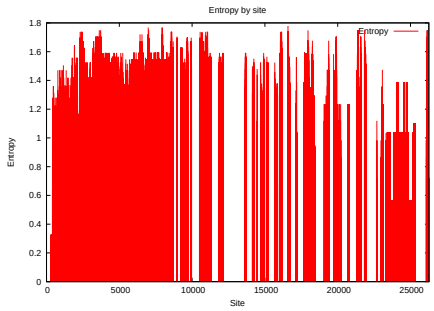


Figure 4.5: Saída do gráfico de entropia dos 6 genes nucleares. Este é o primeiro gráfico produzido pelo *calcphyl* em sua execução.

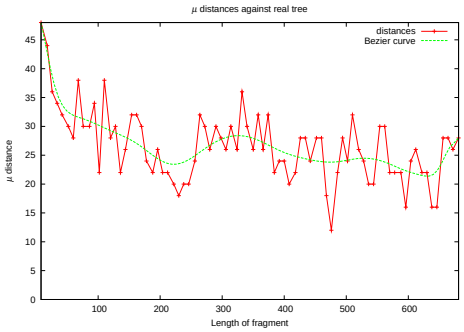


Figure 4.6: Gráfico do cálculo de distâncias SPR (ou distâncias μ) para o gene *cox1* usando o método Neighbor-Joining com curva "bezier" habilitada.

	A	B	C	D	E
1	**Size (nucleotides)**	**bootstraps sum**	**number of bootstraps**	**relative bootstrap**	**SPR distance**
2	8	3317,1	52	63,7903846153846	45
3	18	3685,04	56	65,8042857142857	44
4	26	3767,91	56	67,2841071428571	32
5	34	4140,4	56	73,9357142857143	26
6	42	4248,2	56	75,8607142857143	24
7	52	4392,6	56	78,4392857142857	18
8	60	4359,9	56	77,8553571428571	24
9	68	4340,38	56	77,5067857142857	32
10	76	4416,5	56	78,8660714285714	20
11	86	4438,5	56	79,2589285714286	20
12	94	4479,2	56	79,9857142857143	22
13	102	4428,4	56	79,0785714285714	24
14	110	4501	56	80,375	18
15	120	4532,7	56	80,9410714285714	16
16	128	4633,1	56	82,733928571428	18
17	136	4648,2	56	83,0035714285714	24
18	144	4627,7	56	82,6375	22
19	154	4740,1	56	84,6446428571428	20
20	162	4729,3	56	84,4517857142857	24
21	170	4667	56	83,3392857142857	22
22	178	4690,8	56	83,7642857142857	24
23	188	4693,5	56	83,8125	22
24	196	4684,3	56	83,6482142857143	20
25	204	4690,4	56	83,7571428571428	16
26	212	4712,2	56	84,1464285714286	18

Figure 4.7: Trecho da planilha resultante dos cálculos para os 6 genes nucleares com dados de distância SPR e bootstraps para o método de neighbor-joining.

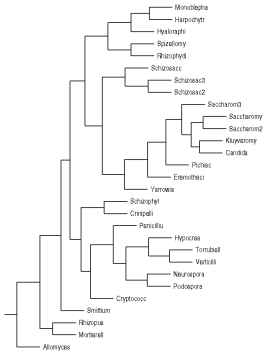


Figure 4.8: Gráfico da árvore com melhores bootstraps para as inferências feitas utilizando Neighbor-Joining para o gene *cox2* de 29 espécies.

```
*** Variable values:
        input file=cox1.phy
        True tree: ./cox1-run-ml-position-to-compare-tree/cox1-run-ml-tree.nwk
        identifier=cox1-cmp-nj
        output format:  vectorial
        data format: xls
        bezier curve: yes
        work directory: ./comparison-2-mp
        iteration method: position
        basedir=/home/lge/patola/bioinfo/phylip-3.68/exe
        debug=0
        permutations=100
        points=80
        keep temporary files=0
        shuffle=1
        remove every third nucleotide=0
        entropy threshold=0.0000000000
        tree building method=mp
        Transition/transversion ratio=11.5000000000
        Number of times to jumble=3
*** Number of residues in alignment: 15432
*** Length: 681
*** Number of sequences: 29
*** Percentage of identity: 64.879137
*** Alignment is flush: yes
*** Alphabet of alignment: protein

Best size obtained for this alignment: 681 sites.
*** Started execution at Tue Mar 24 22:15:21 2009
*** Ended execution at Wed Mar 25 02:10:35 2009
*** Execution took 14114 seconds. (0 days, 3 hours, 55 minutes and 14 seconds)
```

Figure 4.9: Arquivo-texto contendo dados da execução do *calcphyl* para o gene cox1 usando Máxima Verossimilhança.

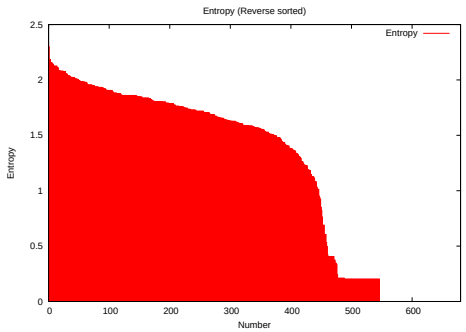


Figure 4.10: Gráfico de entropia ordenada (maiores entropias primeiro) para uma execução usando o gene *cox1* com Máxima Verossimilhança.

4.2 Dados escolhidos

"*O problema com fatos é que há tantos*!"
Samuel McChord Crothers

É de interesse deste trabalho ter como dado um caso prático associado ao fungo *Moniliophthora perniciosa* e por isso o reino *Fungi* foi escolhido como grupo a ser estudado. As duas fontes principais para este propósito foram:

4.2.1 Artigo de filogenia dos fungos

Um dos artigos mais influentes na determinação da filogenia dos fungos, de caráter acadêmico bastante respeitado, é o de Timothy James e seu time de colaboradores (James et al. 2006). Alguns dos fatores que levaram à escolha destes dados:

- Os dados moleculares (genes, alinhamentos) estão disponíveis para download;
- Temos a árvore-gabarito (filogenia) resultante;
- Citado por 72 artigos no *Pubmed*;
- Riqueza de dados: no total são analisados 214 táxons (neste caso, espécies) diferentes, 15 delas de grupo externo. Temos 6 genes nucleares de DNA: Ribossomal 18S, ribossomal 28S, região de gene ITS, EF1α, RPB1, RPB2.

4.2.2 Base filogenética dos fungos: aftol.org

O sítio http://www.aftol.org (Assembling the Fungal Tree of Life) é uma grande base de dados agregando os resultados de diversos artigos, análises e resultados de computações sobre a filogenia dos fungos, com a possibilidade de se baixar os alinhamentos pertinentes. Com esta base:

- Foram usados pelo estudo dados concordantes com a filogenia de (James et al. 2006);
- Houve a possibilidade de escolher critérios diferentes para desempate - para isso foram utilizados os genes da citocromo oxidase 1, 2 e 3 (cox1, cox2, cox3), que são mitocondriais e de proteína.

4.2.3 Refinamento dos dados de entrada

Para ambos os casos, foi necessário diminuir os alinhamentos, extirpar os mais incompletos e procurar obter um número razoável de táxons "modelo", em cujo conjunto foi inserida a Moniliophthora perniciosa. Apesar de no escopo deste trabalho haver variação no número desses táxons por execução, procuramos manter consistentes as espécies usadas. Na execução usando menos espécies (as de genes mitocondriais, tanto por haver menos disponibilidade quanto por no caso estudado estes serem de proteína e exigirem mais processamento), usamos 13 táxons e o total de táxons escolhidos foi 29 (Figura 4.11).

Esse refinamento é necessário e na verdade aprimora o cálculo da filogenia, removendo partes dos dados que potencialmente gerariam artefatos e aumentando o grau de certeza dos cálculos finais (Bininda-Emonds et al. 2001). Além disso, o tempo de inferência de uma filogenia, especialmente para o método de Máxima Verossimilhança, aumenta vertiginosamente em função do número de táxons, mas não tanto em função do número de nucleotídeos e DNA (Servedio 1998).

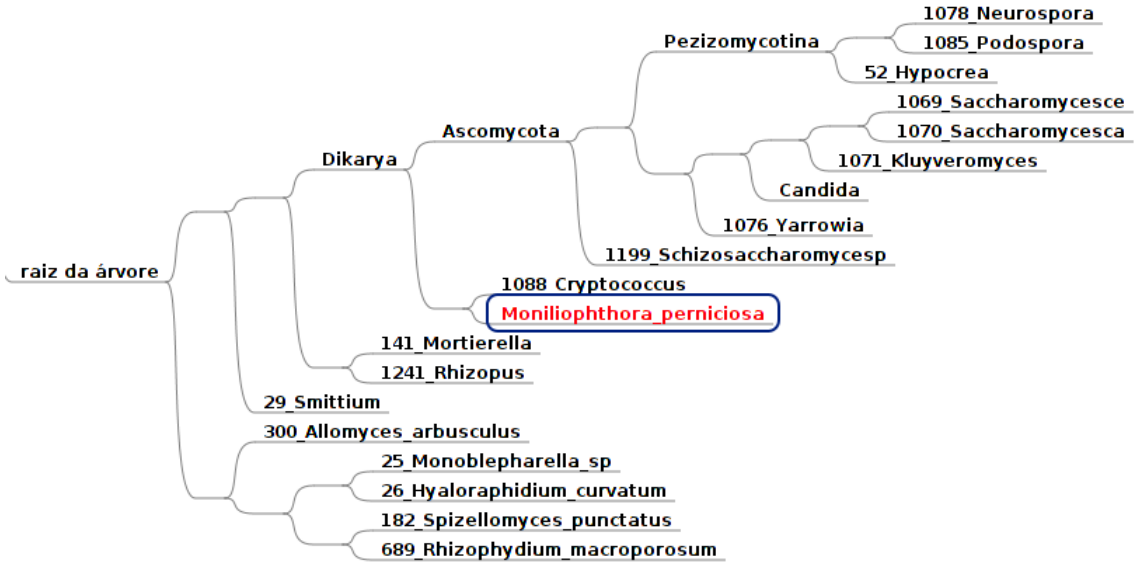


Figura 4.11: Representação gráfica das espécies-modelo que utilizamos para o cálculo das filogenias de interesse. Este é um subconjunto da árvore de *James et al* com a Moniliophthora perniciosa incluída. Foi usado apenas um representante de cada gênero ou classe (no caso de classes pequenas), para assegurar alguma distância entre os táxons e portanto consistência da filogenia.

4.2.4 Variação de parâmetros

Como a abordagem é exploratória, seguintes variações foram exploradas nas execuções do software desenvolvido:

Primeira escolha: método de inferência filogenética

Sabe-se que existe um balanço entre tempo de processamento e qualidade da inferência. O *calcphyl* foi construído de modo que os três métodos principais de inferências pudessem ser utilizados e assim se pudesse escolher o mais adequado:

- Neighbor-Joining;
- Parcimônia;
- Máxima Verossimilhança (com heurística).

Ao utilizar cada um destes três métodos nas inferências iteradas do estudo, as perguntas que se deseja responder são:

- A penalidade de qualidade dos métodos mais rápidos é consistente, isto é, ocorre sempre?
- A penalidade de velocidade dos métodos mais precisos é tolerável?

Segunda escolha: bootstrap *vs* árvore-gabarito

A literatura atual descreve o bootstrapping (e métodos estatísticos semelhantes, como o "jack-knife") como insuficientes para auferir a qualidade (no sentido de "probabilidade de ser a história evolutiva real") de uma filogenia (Hillis e Bull 1993). Além disto, este trabalho necessita da mensuração da variação dos índices de bootstrap de acordo com o tamanho, visto que o intuito é conhecer um ponto de máximo que se possa usar com genes incompletos. Caso o bootstrap acompanhe o gráfico de qualidade comparada da filogenia inferida com a árvore-gabarito, ter-se-á um índice independente de dados externos para ser usado.

O software *calcphyl* foi projetado com o recurso de escolha de método de comprovação de qualidade – boostrapping ou árvore-gabarito – de modo não-exclusivo, isto é, é possível usar ambos simultaneamente e assim comparar os resultados de um método em relação ao outro.

Terceira escolha: iteração por entropia ou iteração por posição

Ao processar um alinhamento, o programa por *default* itera por posição, ou seja, começando com um tamanho n, retira uma amostra do alinhamento inicial começando de uma posição arbitrária (o centro), calcula a filogenia pelo método configurado, aufere a qualidade e vai fazendo o mesmo para amostras progressivamente maiores dos dados, até chegar ao tamanho original do alinhamento. O número de vezes que isso é feito é configurado por um parâmetro da linha de comando - o número de pontos no gráfico, cujo default é 40.

Passando o parâmetro de iteração por entropia, o programa ao invés de amostrar o alinhamento original por um trecho crescendo a partir do centro, utiliza o *limiar mínimo de entropia*. A entropia de cada posição do alinhamento é anotada; as n posições com as maiores entropias são coletadas para fabricar um alinhamento que então é inferido e sua qualidade é auferida, passando para a próxima iteração com n maior.

Note-se que a entropia, por sua natureza, assume um número discreto (limitado) de valores, sendo comum a ocasião em que teremos várias posições do alinhamento com a mesma entropia. Neste caso, ao escolher entre posições com a mesma entropia, o programa volta para o critério original de posição, crescendo a partir do centro.

Note-se que o início da curva de qualidade de acordo com o tamanho será tipicamente bastante ruim, isto é, com qualidade muito baixa, visto que entropia alta significa grande variação – e teremos um alinhamento pequeno com muitas diferenças, algo notavelmente ruim para o cálculo da inferência. Isto é proposital e é um dos motivos de usarmos este dado: uma curva mais acentuada mostra de forma mais clara as tendências de mudança.

A literatura existente confirma nossas impressões, achando um valor de 0,5 como mínimo de entropia ideal para o cálculo de uma árvore filogenética a partir de um alinhamento de DNA (Miyazaki et al. 1996). Neste estudo no entanto se objetiva estender o conhecimento para uma curva em relação ao tamanho disponível e também para alinhamentos com proteínas.

Quarta escolha: embaralhar alinhamento ou seguir a ordem.

A abordagem deste trabalho é exploratória, portanto inicialmente tomamos uma decisão arbitrária - crescer a amostra de alinhamento a partir do centro, moderadamente justificável quando temos um gene (cuja qualidade de seqüenciamento costuma ser mais concentrada longe das bordas) mas sem boa justificativa quando o alinhamento envolve genes concatenados ou segmentos de comprimentos maiores.

Neste caso procurou-se a anulação do efeito posicional permitindo ao usuário do software a opção de *embaralhar o alinhamento*, isto é, considerando o alinhamento como uma matriz de linhas (táxons) e colunas (DNA ou aminoácidos), aleatoriza-se a ordem das colunas sem mudar a das linhas. Fazendo suficientes inferências com diferentes embaralhamentos nos mesmos dados, estas concordando entre si, pode-se ter segurança de que o efeito posicional não tem muito impacto nas medidas.

Note-se que isso pode gerar uma distorção no caso de usarmos DNA, pois o *quadro (frame) de alinhamento* tem sua quantidade de informação variável – o terceiro nucleotídeo de um *códon* varia mais do que o primeiro e segundo, pois tem mais aminoácidos sinônimos associados. Se ainda assim tivermos consistência nos dados, podemos intuir que este efeito é mitigado nos casos que tratamos.

Quinta escolha: efeito do terceiro nucleotídeo (DNA somente)

Como o terceiro nucleotídeo de um códon é associado a mais aminoácidos sinônimos que os outros, é comum, para diminuir o tempo de cálculo de uma filogenia de DNA e aumentar a quantidade de informação, remover este quadro do alinhamento. A opção fui incluída no software desenvolvido, levando em conta esta prática.

Sexta escolha: proteínas ou DNA

Foram escolhidos tanto dados protéicos quanto de DNA para este estudo, buscando entre estes dados concordâncias e divergências.

Sétima escolha: genes nucleares e mitocondriais

Foram escolhidos tantos genes mitocondriais quanto nucleares para este estudo, buscando entre estes dados concordâncias e divergências.

4.2.5 Quantidade de variáveis x tempo disponível

É simples ver, pela enumeração de fatores variáveis, que há uma quantidade de variáveis muito grande pra tratar. Se houvesse o tratamento de cada uma em particular variando todas as outras, teríamos

- 3 métodos de inferência;
- 2 medidas de qualidade;
- 2 tipos de iteração;
- 5 ordens (uma a partir do centro e pelo menos quatro embaralhadas para consistência);
- 2 opções - incluir ou excluir o terceiro nucleotídeo;
- 2 tipos de dados, proteína ou DNA;
- 2 tipos de genes, nucleares e mitocondriais;
- 2 conjuntos de dados no mínimo (o de fungos e um auxiliar).

Um cálculo rápido mostraria que seria necessário avaliar $3 \times 2 \times 2 \times 5 \times 2 \times 2 \times 2 \times 2 = 960$ casos diferentes, sem contar os diferentes algoritmos possíveis (phylip vs. phyml, além de dois métodos de cálculo de distância filogenética). Levando-se em conta que o cálculo de filogenias iterado já é extremamente demorado (como será visto mais tarde, uma única execução do software demorou aproximadamente 40 dias para completar), avaliar todas as permutações possíveis de opções é inviável.

Por questões de tempo e disponibilidade dos dados, portanto, procurou-se trabalhar uma ou duas variações por vez, eliminando as alternativas que se descobrissem menos relevantes para o objetivo do estudo.

5 *Apresentação dos Resultados*

"*Contentamento não é o cumprimento do que você quer, mas a percepção de quanto você já tem.*"
Anônimo

5.1 Dados iniciais e filtragem

Quando do software ainda em construção, os primeiros dados utilizados e obtidos foram apenas exploratórios em busca dos melhores modelos e parâmetros para estudar, não constando dos resultados. Como havia muitos modelos diferentes, muitos métodos de construção de árvore e até uma boa quantidade de programas acessórios para escolher acoplados ao *CalcPhyl*, estes dados serviram para orientar as decisões de *como* processá-los para chegar à conclusão do estudo.

A seguir, os objetivos que cumprimos para filtrar progressivamente o tratamento dos dados.

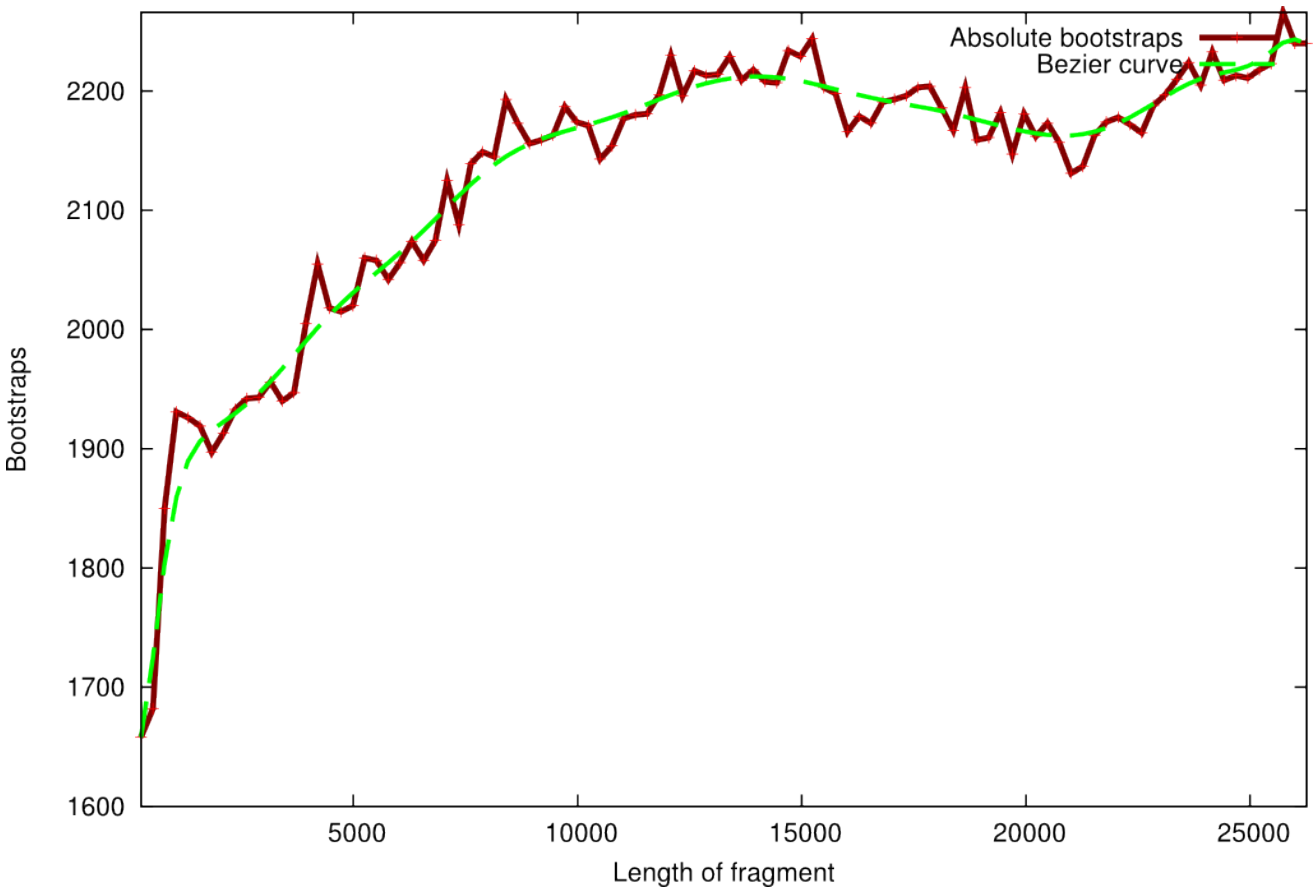


Figura 5.1: Medida de soma dos bootstraps da filogenia resultante do alinhamento de 6 genes nucleares concatenados de fungos de acordo com o tamanho. Uma curva bezier aparece interposta aos pontos para mostrar a direção da curva. (Eixos não têm unidade).

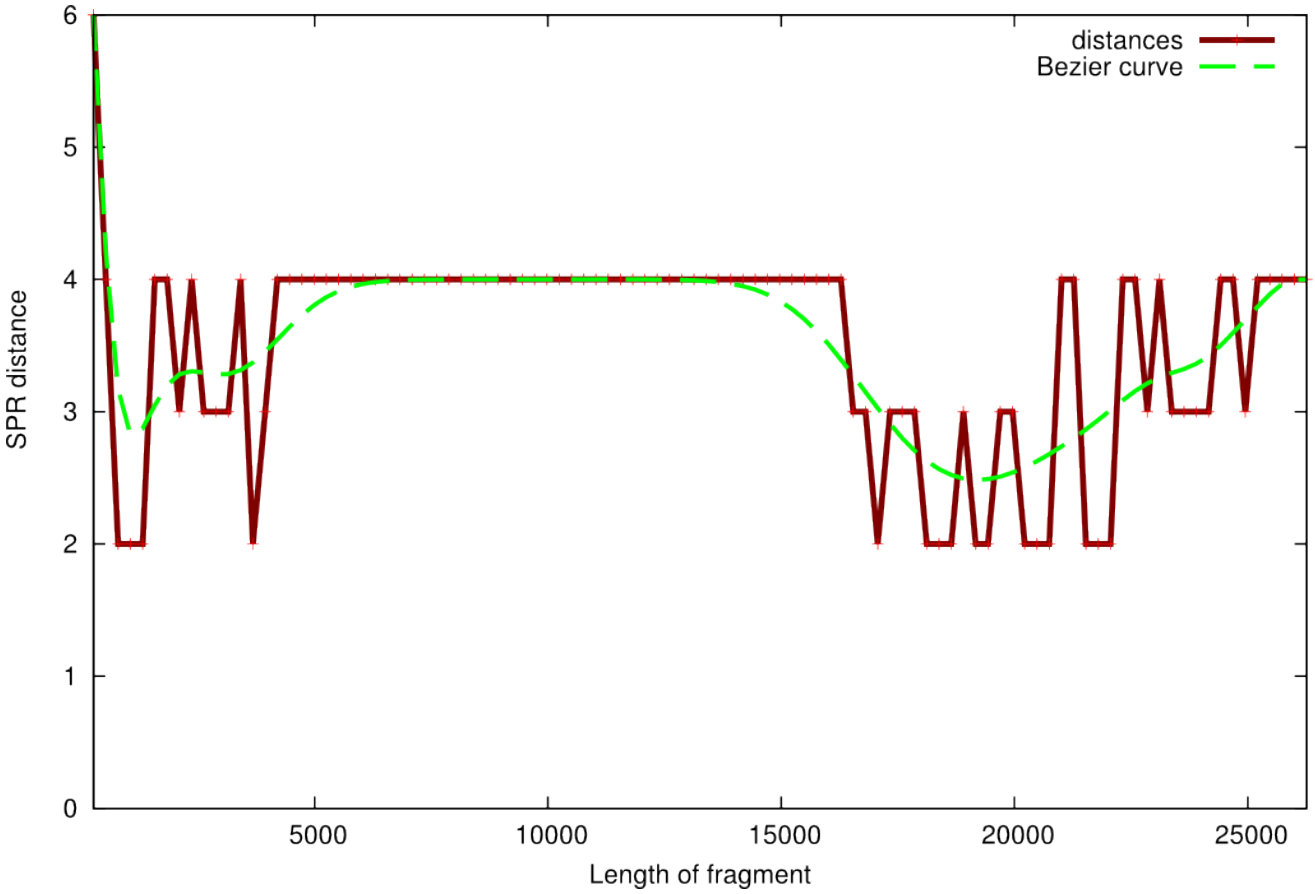


Figura 5.2: Medida da distância de árvore ideal da filogenia resultante do alinhamento de 6 genes nucleares concatenados de fungos de acordo com o tamanho. Uma curva bezier aparece interposta aos pontos para mostrar a direção da curva.

5.1.1 Melhor estimador de qualidade: bootstrap ou árvore-gabarito

Como o objetivo principal é auferir a qualidade das árvores de acordo com o tamanho (ou limiar de entropia de um dado tamanho) de gene, o indicador mais crítico é o método para medir a qualidade de uma filogenia. Outros critérios foram usados na etapa prévia dos dados, contudo dado o impaco deste critério na quantidade de processamento envolvida, tornou-se este o mais elementar, visto que nesta escolha residiria a diferença de duas ordens de grandeza de tempo computacional.

Nesta primeira execução, usamos como representativos aqui os dados dos 6 genes nucleares (DNA) da amostra de *James et. al* concatenados, totalizando um alinhamento de 26261 sítios (porcentagem de identidade das seqüências de 76,57%), usando 13 táxons com embaralhamento de seqüências, 100 permutações de bootstrap (Figura 5.1) e a árvore-gabarito dos fungos (Figura 5.2). A execução demorou 13 dias e 3 horas.

Se o bootstrap fosse uma medida real da qualidade da filogenia inferida - ao invés de somente uma medida da consistência interna de partições dos dados - esperaríamos que ele seguisse as inflexões da curva da distância da árvore ideal de maneira inversa: quanto menor a distância (melhor qualidade), maiores os bootstraps calculados. E enquanto a curva de bootstraps vai aumentando de forma semelhante a uma parábola, alcançando um máximo por volta da quantidade de 15000 nucleotídeos, a de distância estabiliza bem antes disso, tendo um aprimoramento da qualidade exatamente nos pontos em que a de bootstrap tem uma decréscimo.

Note-se que as duas medidas foram calculadas na mesma execução, desaventando a hipótese de um artefato de diferentes embaralhamentos da seqüência.

Como seria de esperar, procurou-se buscar mais uma confirmação disto na forma das seqüências auxiliares de *cox*. Foi preciso ser criterioso nos parâmetros a usar pois o tempo de cálculo de filogenias a partir de proteínas é mais longo do que o tempo a partir de DNA. Mesmo assim, tentou-se usar um número maior de bootstraps (200) para se ter maior confiança nos dados.

Rodou-se o programa para o alinhamento envolvendo o gene *cox3*, que tem 313 sítios e porcentagem de identidade de 49,69%, com embaralhamento da seqüência, 100 pontos de gráfico e 19 táxons. O resultado de bootstraps pode ser visto na Figura 5.3 (note que os valores de bootstraps são bem diferentes daqueles do gráfico de genes nucleares; para evitar confusões, ao invés de colocar esses valores como uma porcentagem, foi colocada a soma dos bootstraps dos galhos com seu valor absoluto, visto que bootstraps *não podem ser comparados entre si*) e o resultado com distâncias na

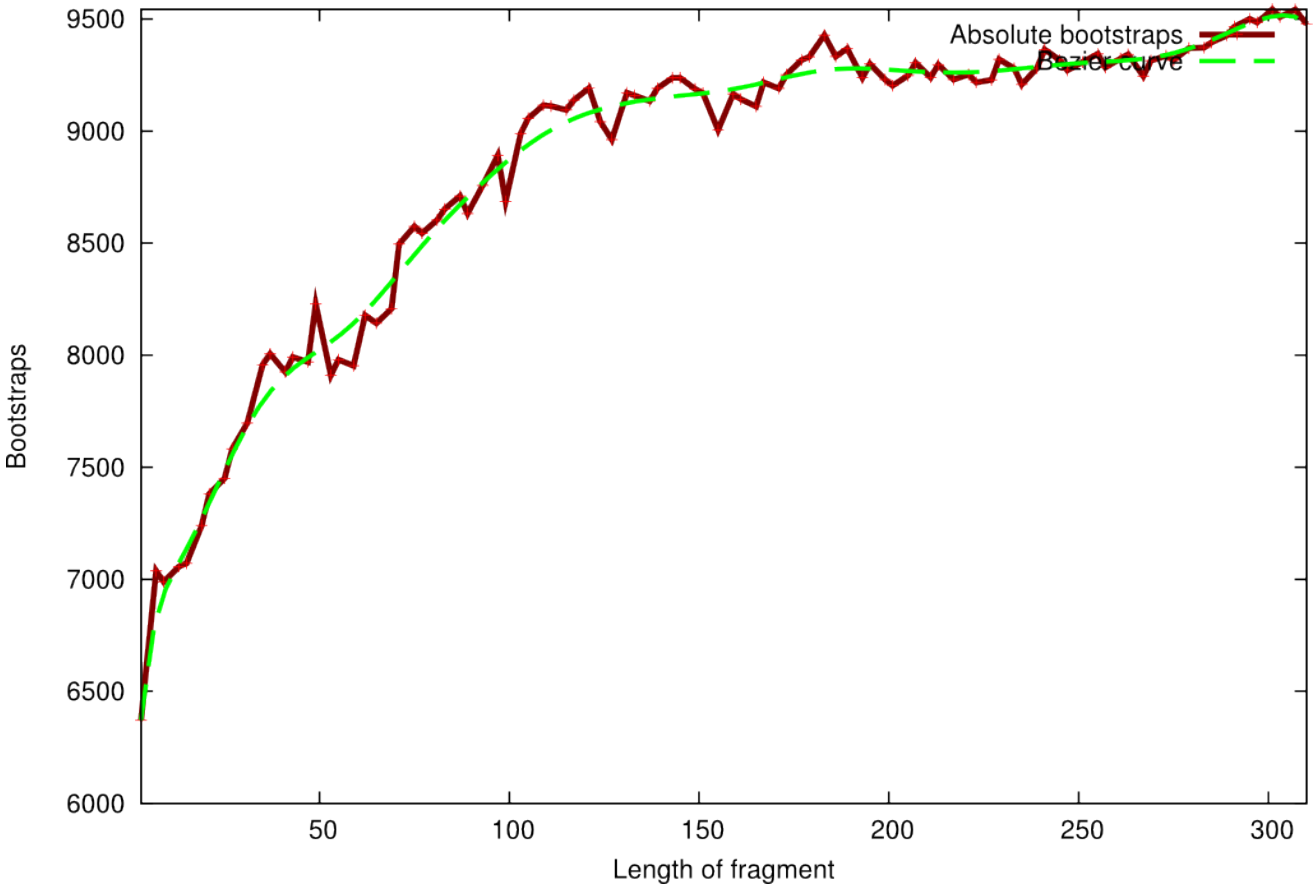


Figura 5.3: Bootstraps calculados para o alinhamento do gene de proteína *cox3*, com curva bezier auxiliar.

Figura 5.4. A execução demorou 26 dias e 11 horas.

Novamente nota-se que as curvas são bastante diferentes em suas inflexões e não são opostas como se esperaria. A curva de bootstraps aumenta proporcionalmente ao comprimento enquanto a de distância atinge melhor qualidade (menor valor) por volta de 2/3 do alinhamento para então tender para valores maiores, piorando de qualidade.

Pode-se ver, por essa análise, que o bootstrap **não satisfaz os critérios** de medida de adequação de uma filogenia, concordando com a literatura sobre o assunto, e é, a partir desse momento, abandonada em favor da comparação com uma árvore-gabarito. Se isso por um lado traz algumas limitações - para um estudo como o proposto, fica-se limitado àqueles com uma filogenia bem determinada e conhecida -, por outro deixa de exigir trilhar o caminho computacionalmente caro de realizar centenas de iterações adicionais para o cálculo do índice de bootstrap.

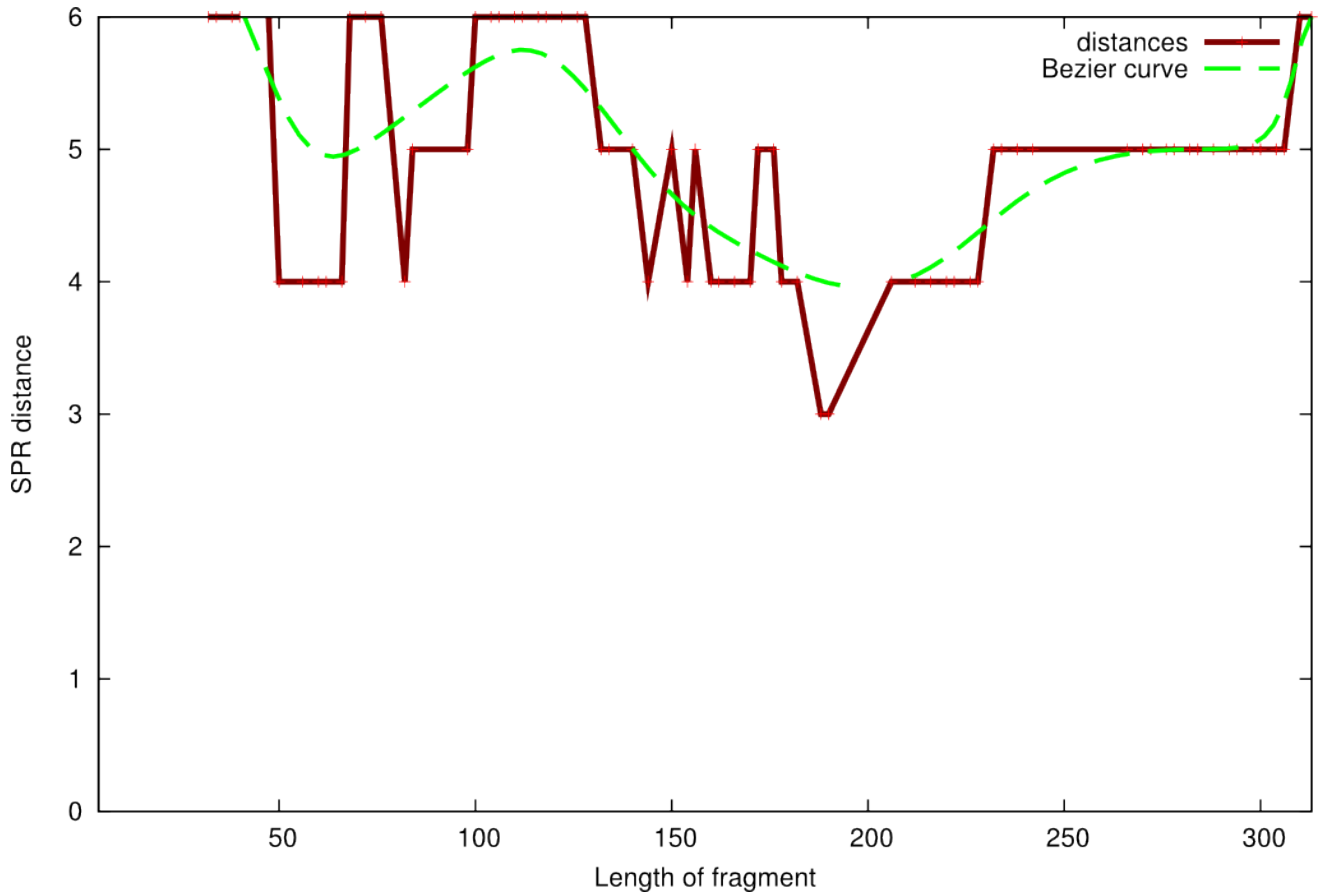


Figura 5.4: Distâncias da árvore ideal para o gene *cox3*, com curva bezier auxiliar.

5.1.2 Método de Inferência: Máxima Verossimilhança versus Máxima Parcimônia e Neighbor-Joining

Sabe-se pela literatura que, dos três tipos principais de métodos de inferência, a Máxima Verossimilhança tem mais chances de chegar ao resultado correto que as outras, razão pela qual a usamos para atestar a adequação do bootstrap. Cabe a dúvida, entretanto, se ela é estritamente necessária. A pergunta se torna relevante somente se a diferença em qualidade for estatisticamente desprezível e consistente, afinal almeja-se neste estudo a melhor qualidade possível tendo somente resultados preliminares de um seqüenciamento.

Consistentemente, as filogenias de Máxima Parcimônia alcançavam uma qualidade (medida pela distância) 10% a 20% pior do que a de Máxima Verossimilhança, ilustrado, por exemplo, nestes gráficos em cima do alinhamento de 6 genes ordenado (Figura 5.5 e Figura 5.6). Para a Máxima Verossimilhança a execução levou 3 horas e 22 minutos enquanto que para a Parcimônia foram necessários apenas 44 minutos, usando os mesmos parâmetros.

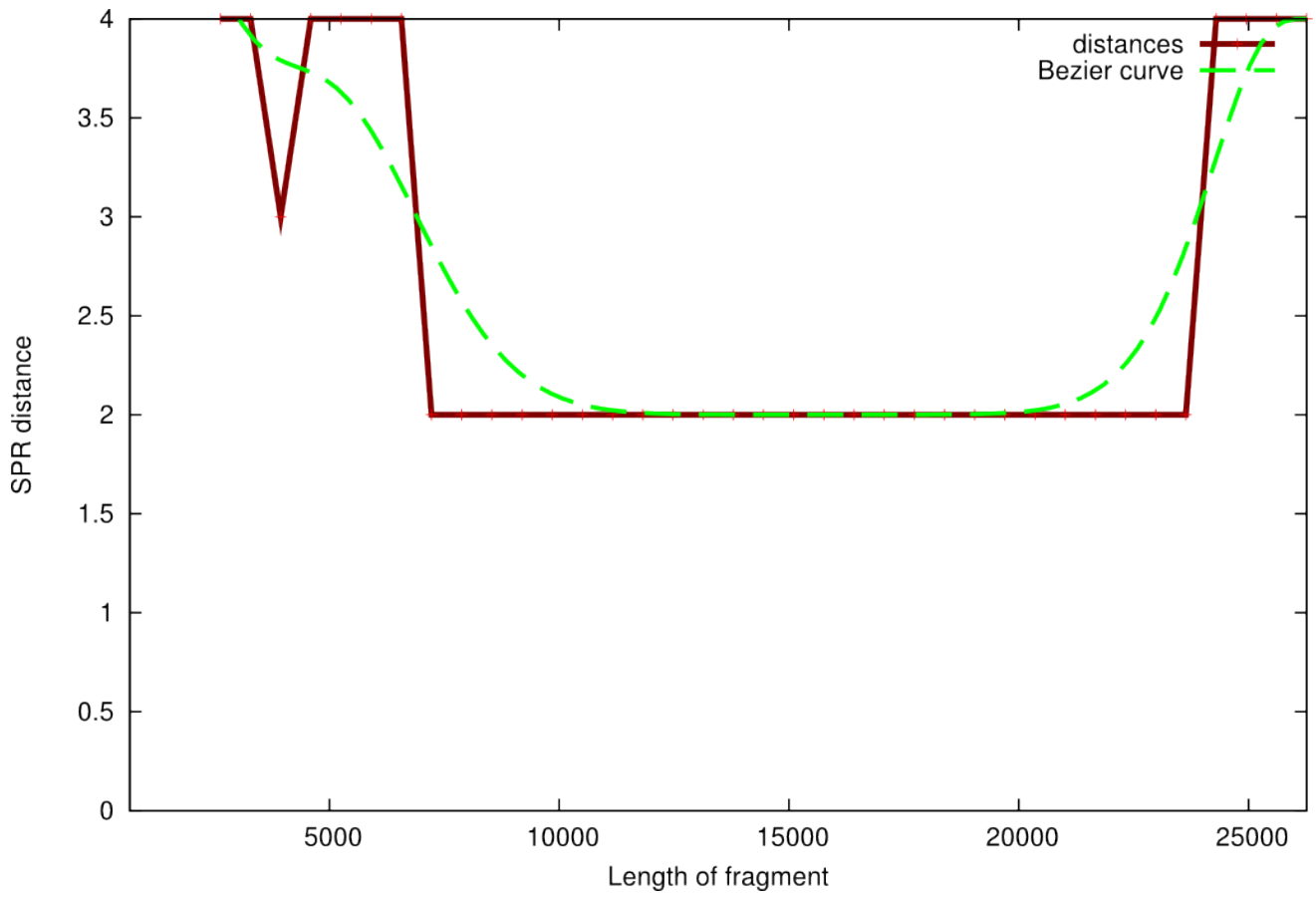


Figura 5.5: Gráfico da distância da Verossimilhança Máxima para a seqüência de 6 genes nucleares. Note que temos um máximo de qualidade mais para o centro e não no comprimento total do alinhamento.

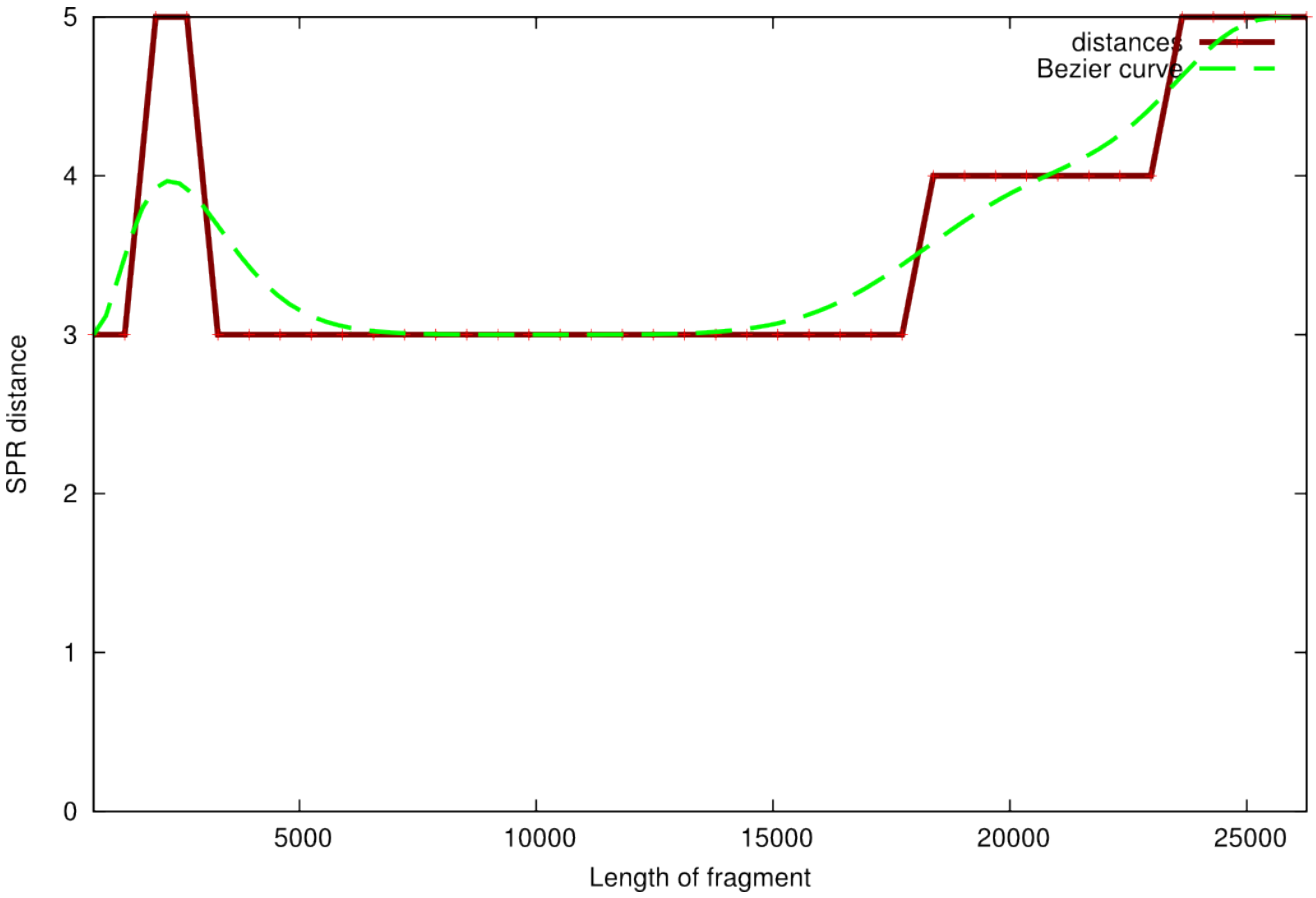


Figura 5.6: Gráfico da distância da Parcimônia Máxima para a seqüência de 6 genes nucleares. Note que temos um máximo de qualidade mais para o centro e não no comprimento total do alinhamento, como no caso da Máxima Verossimilhança.

O gráfico das duas figuras é bastante semelhante, apresentando uma melhoria de qualidade perto do centro e decrescendo ao se aproximar do comprimento total, com a Máxima Verossimilhança nunca obtendo uma distância menor que 2 unidades da árvore ideal e Parcimônia nunca tendo menor que 3, embora fiquem estabilizadas nessas distâncias. Essa tendência se manteve para outros casos (e.g., dois dos genes *cox*) e com outros parâmetros. Como a curva da Máxima Parcimônia acaba se assemelhando com a da Máxima Verossimilhança mas demora bem menos tempo, podemos recomendá-la apenas quando se tiver parcos recursos computacionais ou tempo bastante curto pra produzir a filogenia. A Máxima Parcimônia, entretanto, tem maior probabilidade de apresentar inconsistências que a de Máxima Verossimilhança.

Foi este o caso com o gene *cox1*, de 681 aminoácidos e com porcentagem de identidade de 65%. Iteramos por posição com quarenta pontos e Máxima Parcimônia não retornou resultados

bons ou consistentes. Com este método, a distância permaneceu excessiva na filogenia resultante para praticamente todos os comprimentos (Figura 5.7)

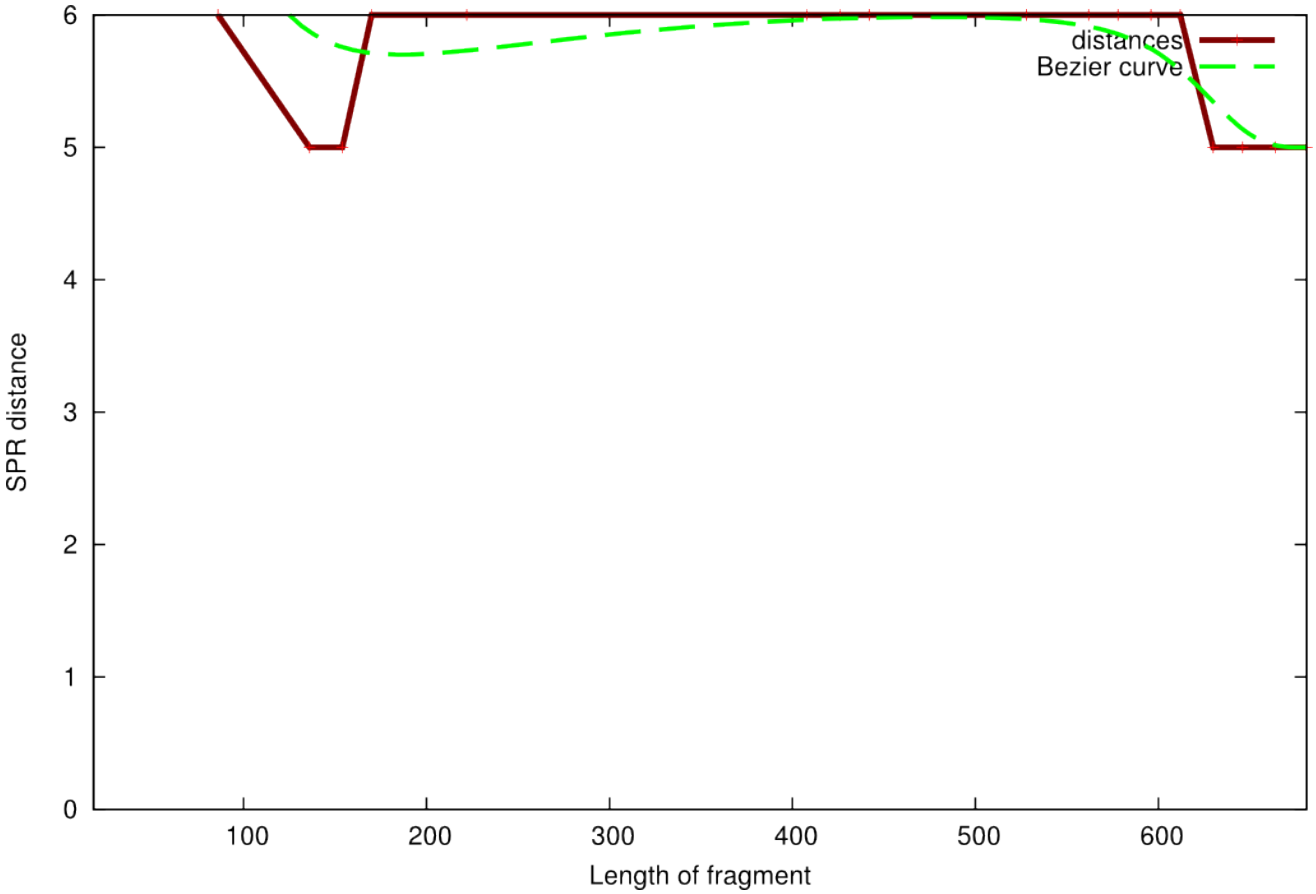


Figura 5.7: A distância filogenética de acordo com o tamanho do alinhamento usando o método de Máxima Parcimônia para o gene *cox1*. Note que com 19 táxons, 6 é uma distância considerável.

No entanto, o mesmo gene *cox1* demonstra uma tendência comum aos outros alinhamentos de a análise com Máxima Verossimilhança (Figura 5.8) ser mais precisa do que a de Neighbor-Joining (Figura 5.9), e esta última apresentar artefatos (picos e vales) nas bordas:

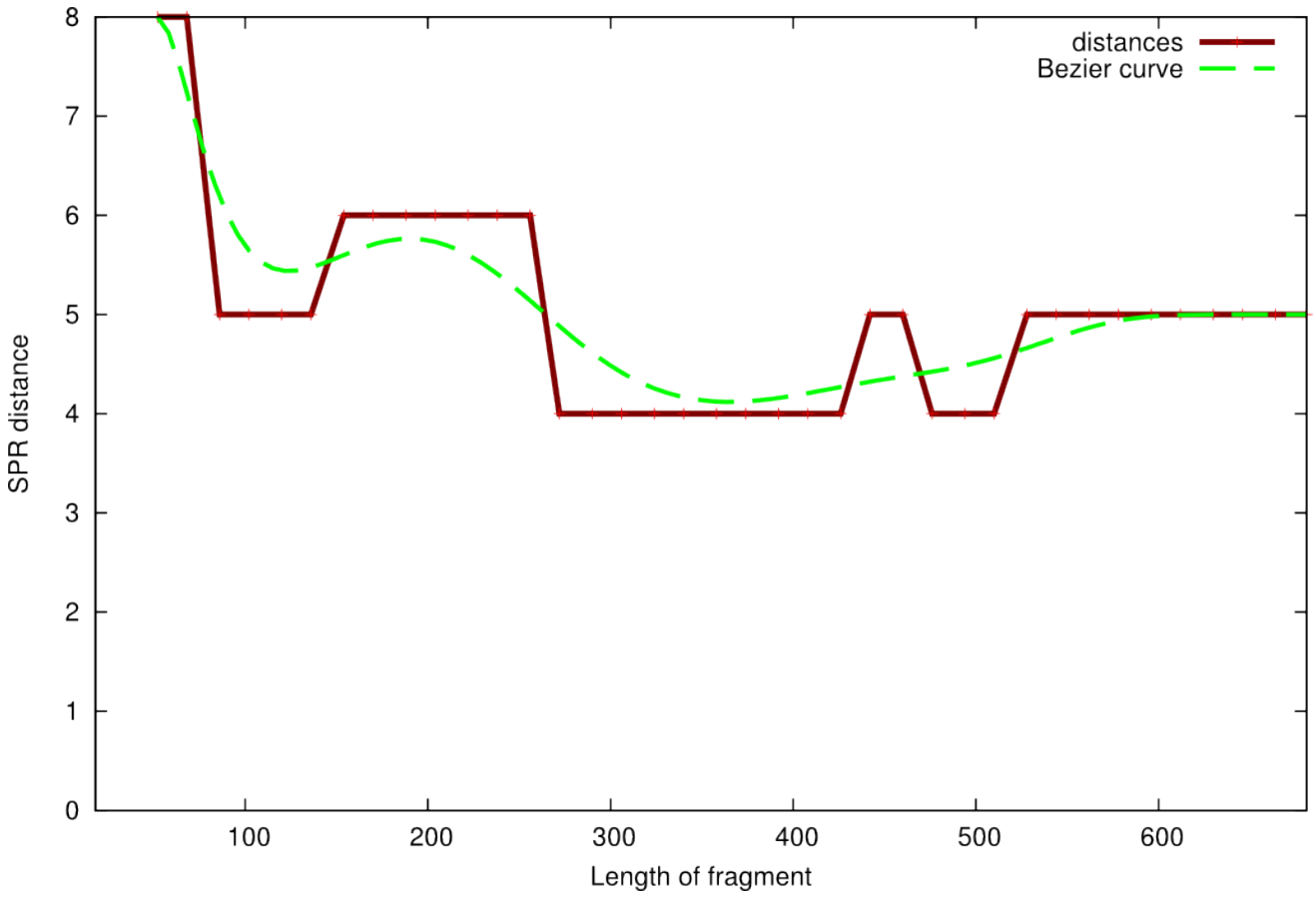


Figura 5.8: A distância filogenética de acordo com o tamanho do alinhamento usando o método de Máxima Verossimilhança para o gene *cox1*. Note que novamente os extremos têm menos qualidade, um tema recorrente.

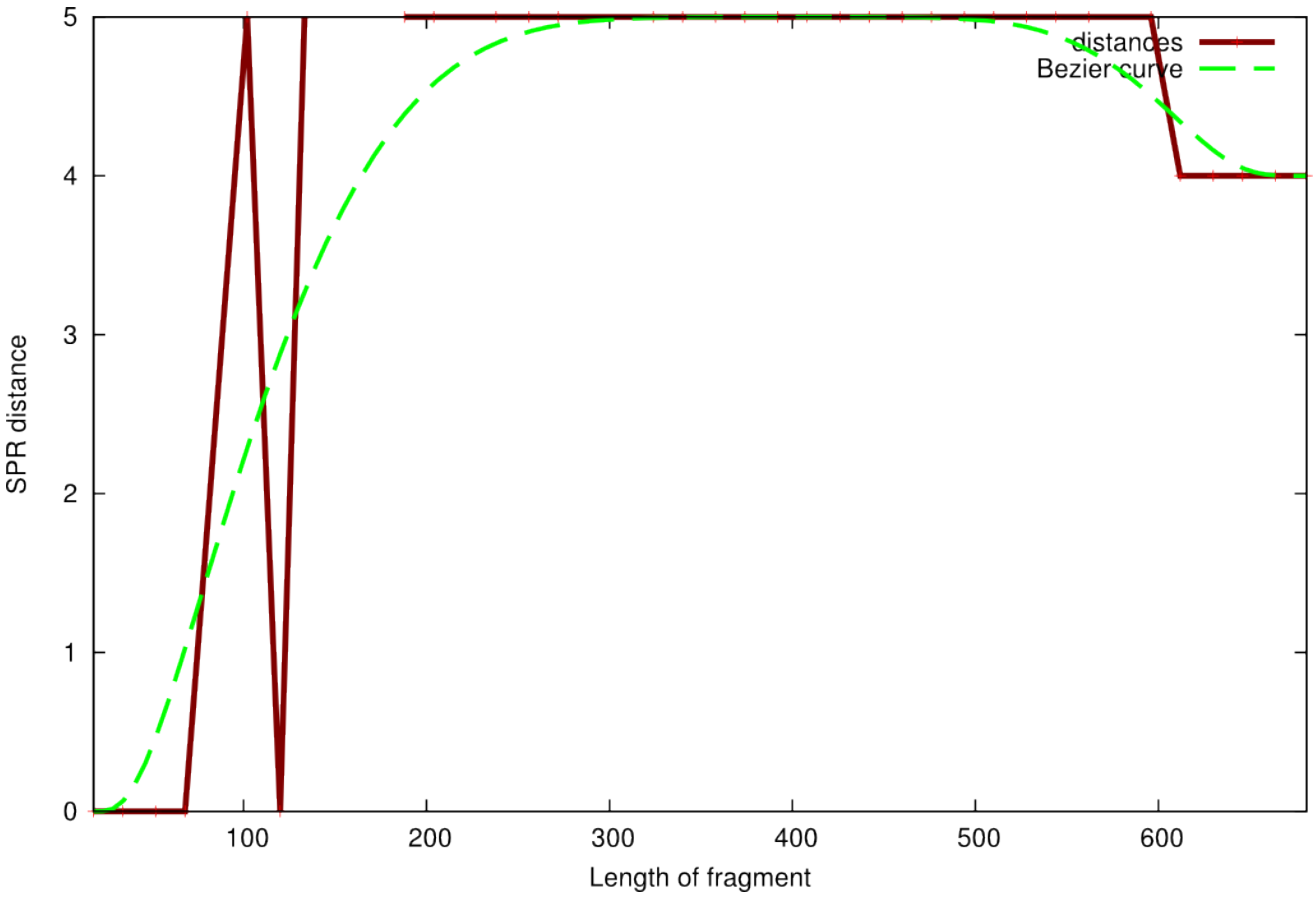


Figura 5.9: A distância filogenética de acordo com o tamanho do alinhamento usando o método de Neighbor-Joining para o gene *cox1*. Note o artefato que aparece no extremo esquerdo do gráfico (quando o alinhamento é pequeno), dando a impressão de que a qualidade é boa. Note também que com a curva estável, a qualidade do Neighbor-Joining fica entre o método de Máxima Verossimilhança e Parcimônia.

Novamente, os tempos ficaram bem diferentes: meros um minuto e 14 segundos para Neighbor-Joining, 4 minutos e 16 segundos para Máxima Parcimônia e uma hora e 35 minutos para Máxima Verossimilhança. No entanto, Neighbor-Joining, mais rápido que Parcimônia, obteve maior qualidade, o que levanta dúvidas sobre a segurança de usá-la na hipótese de não poder arcar com os custos de Máxima Verossimilhança.

A partir desta seção, a não ser que se diga o contrário, pressupõe-se que é usada Máxima Verossimilhança.

5.1.3 Efeito do embaralhamento

Importanta para a análise destes dados é também medir o efeito do embaralhamento dos dados, um recurso incluído no software. Possíveis conseqüências como degenerar o formato da curva de qualidade precisam ser medidos. Visto que procurou-se distribuir as análises entre alinhamentos embaralhados e normalmente ordenados em várias das execuções, é importante avaliar se o efeito positivo de evitar viés de posicionamento não é atrapalhado pelos efeitos negativos de se perder a ordem.

Usaram-se os genes *cox1* (681 aminoácidos, porcentagem de identidade no alinhamento 65%) e *cox3* (313 aminoácidos, porcentagem de identidade no alinhamento 50%), com os mesmos parâmetros pra ambos: uma execução iterando por posição, ordenadamente (Figura 5.10 para *cox1* e Figura 5.12 para *cox3*) e quatro execuções iterando por posição com o alinhamento embaralhado(Figura 5.11 para *cox1* e Figura 5.13 para *cox3*).

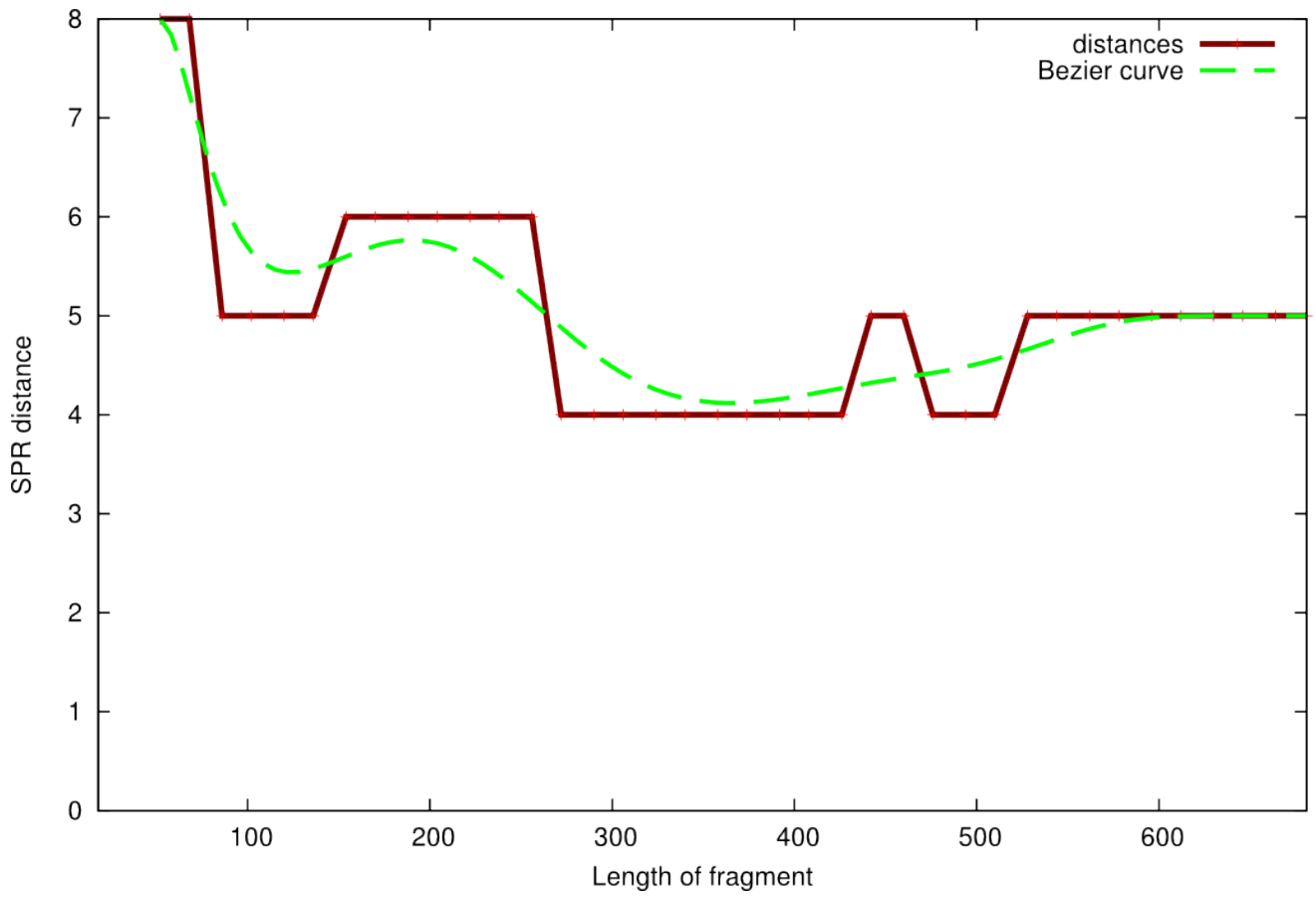


Figura 5.10: Distância da árvore ideal da filogenia do gene protéico *cox1* iterando por tamanho, ordenado. A curva bezier (em verde) sobreposta ao gráfico mostra de forma suave a variação da qualidade, melhor no centro do gráfico onde o gene de proteína tem de 272 aminoácidos até 442 aminoácidos. O tamanho total do gene é 681 aminoácidos.

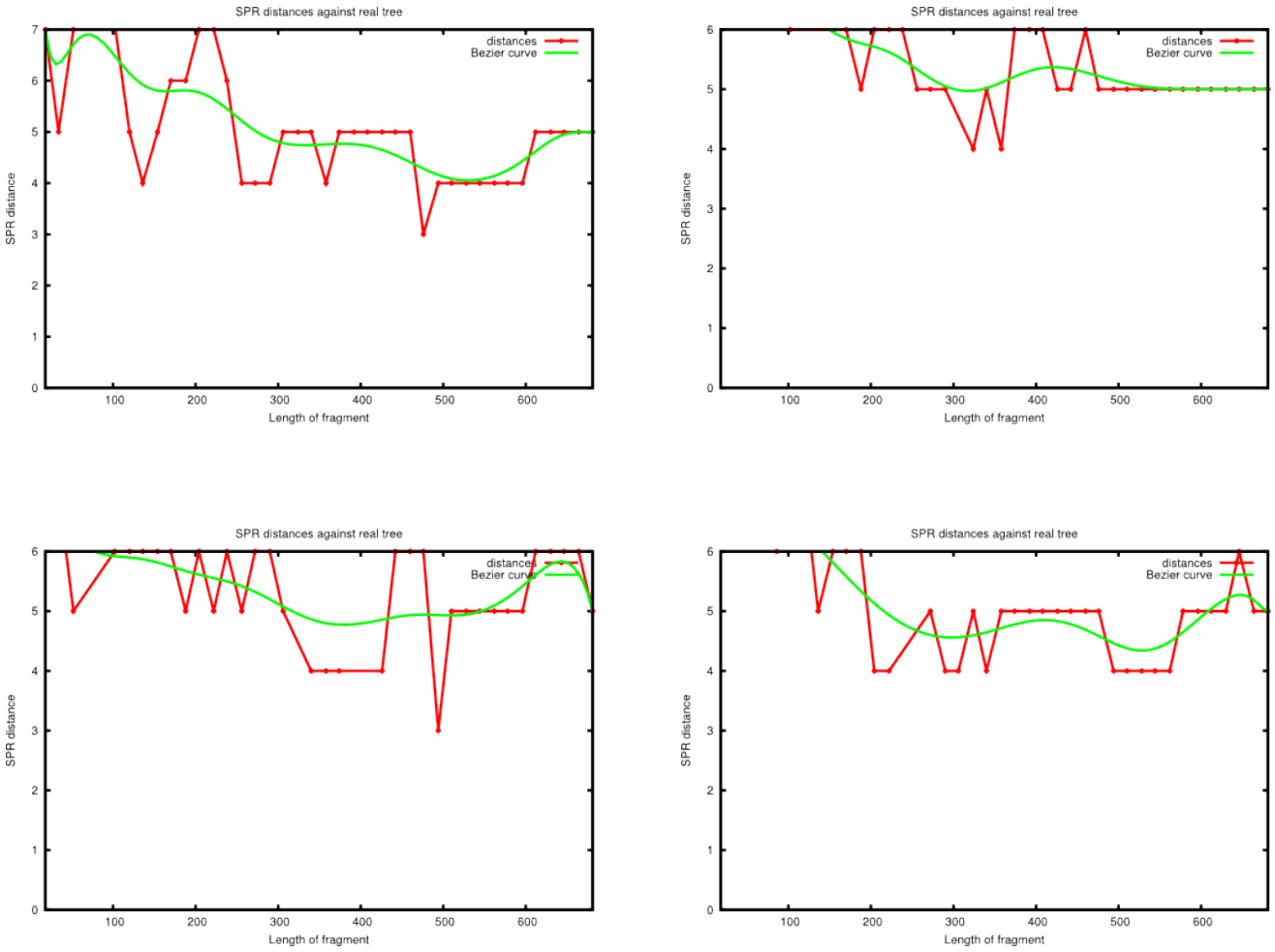


Figura 5.11: Quatro iterações com os mesmos parâmetros: gene protéico *cox1* iterando por tamanho com a seqüência embaralhada. Percebe-se que embora os gráficos variem bastante de acordo com o tamanho, eles mantêm a tendência de qualidades boas mais para o centro (menor que o tamanho total do gene, 681 aminoácidos). Valores notáveis estão nos gráficos da esquerda - o gráfico superior à esquerda tem a melhor qualidade (3 unidades SPR) com 476 aminoácidos e o gráfico inferior à esquerda tem a mesma qualidade com 494 aminoácidos. Os gráficos à direita não apresentaram estes mínimos notáveis.

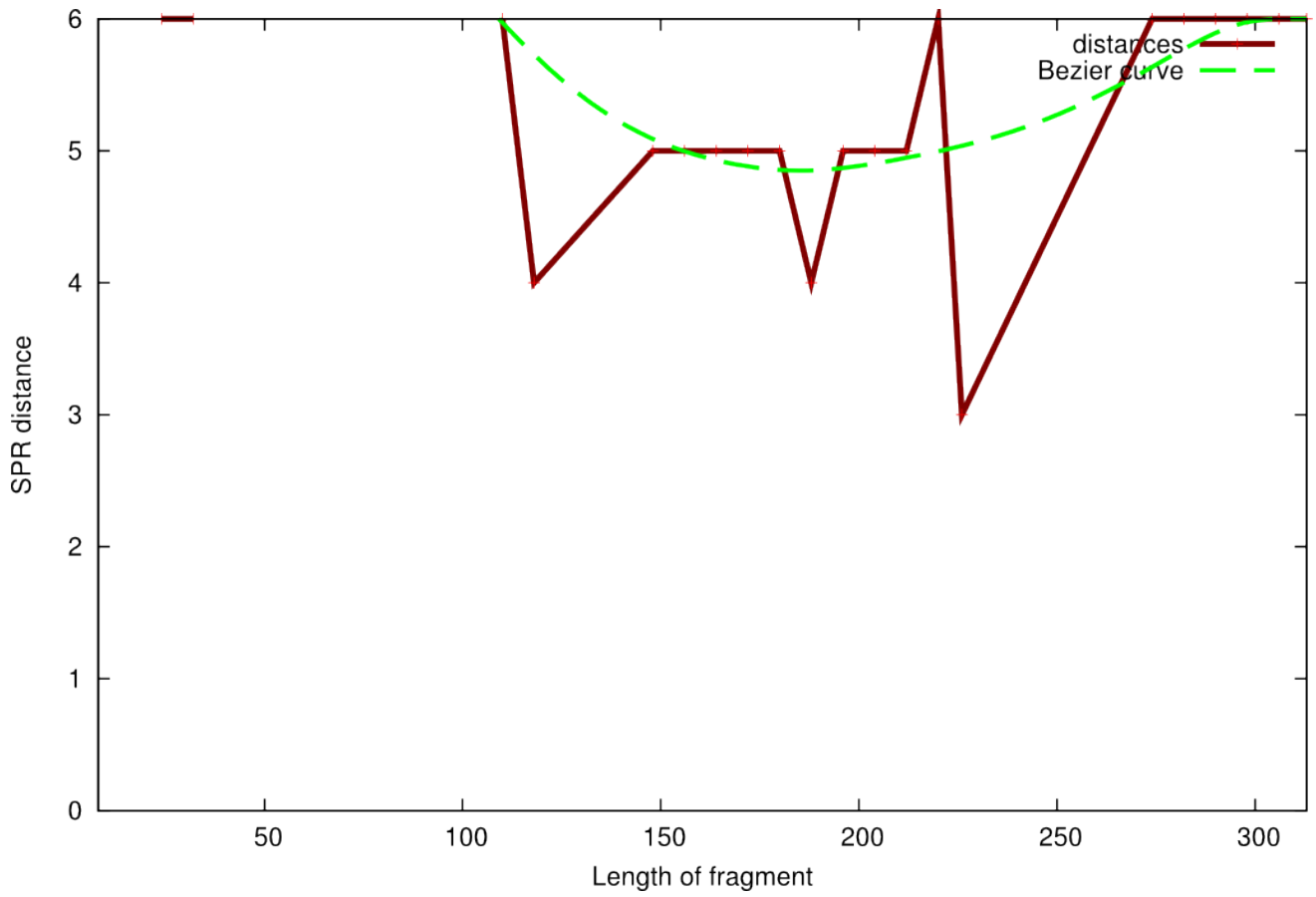


Figura 5.12: Distância da árvore ideal da filogenia do gene protéico *cox3* iterando por tamanho, ordenado. Nota-se melhor qualidade (menores valores de distância da árvore ideal) mais próximas ao centro do gráfico, isto é, fora do seu tamanho total. A menor distância é de 3 unidades na posição 226 (gene protéico de 226 aminoácidos). O tamanho total do gene é de 313 aminoácidos.

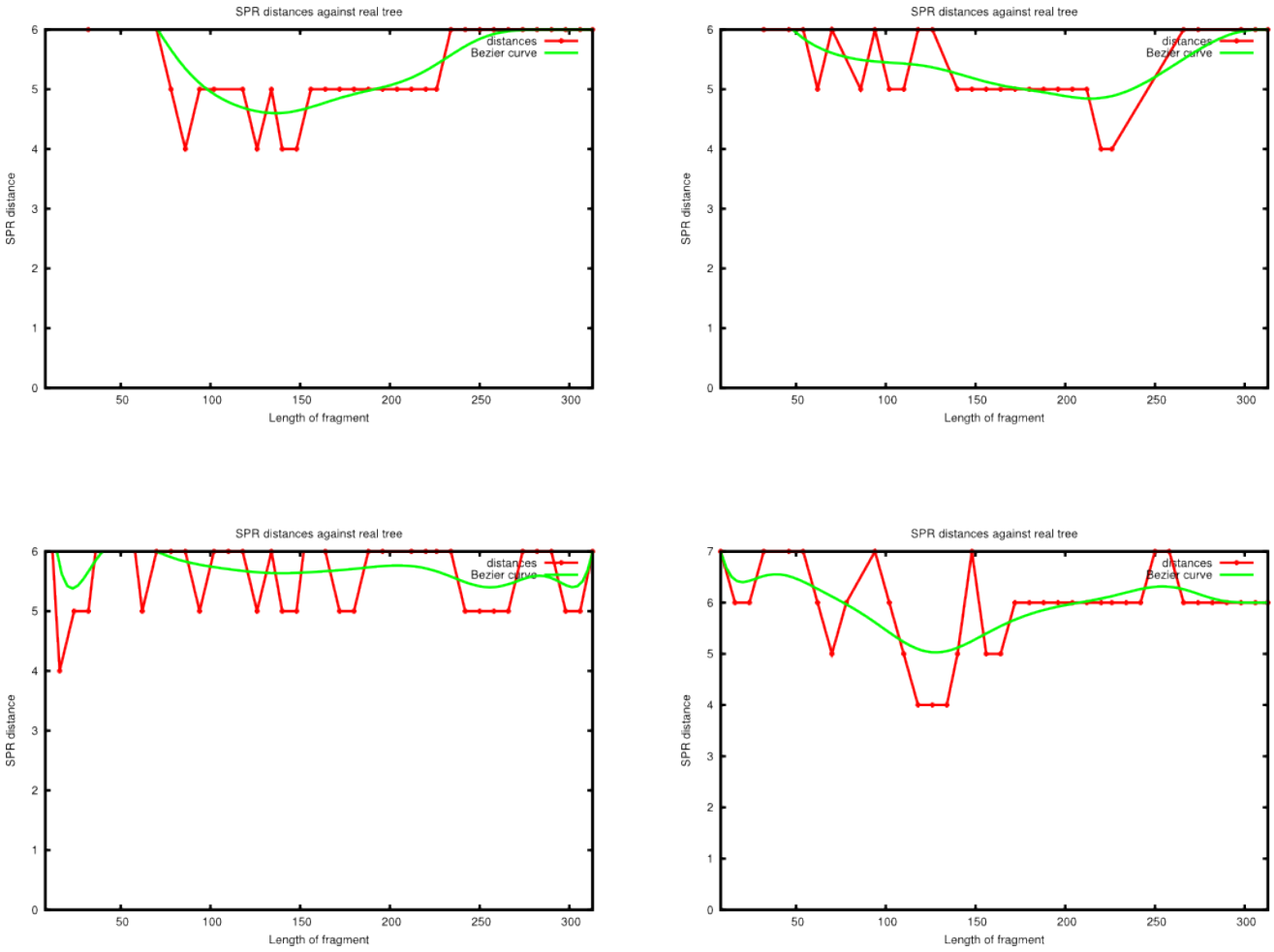


Figura 5.13: Quatro iterações com os mesmos parâmetros: gene protéico *cox3* iterando por tamanho com a seqüência embaralhada. Excetuando o gráfico inferior à esquerda, que parece ter distribuído os vales e picos igualmente e apresentado um artefato (distância mínima de 4 unidades SPR) na posição inicial de 16 aminoácidos, nota-se uma tendência a melhores (menores) valores de qualidade mais para o centro do gráfico, isto é, quando o gene está com menos de seu comprimento total. O tamanho do gene *cox3* é 313 aminoácidos.

Percebe-se que há uma deformação na curva de qualidade quando do embaralhamento no sentido de diminuir aclividades e declividades, em algumas execuções chegando quase a aplainar a curva, mas em geral a tendência de a menor distância encontrar-se no centro do gráfico (onde o gene está em metade do seu tamanho) se mantém.

5.1.4 Remoção do terceiro nucleotídeo

Surpreendentemente, esta técnica que se usa para diminuir o tamanho do alinhamento (e portanto o tempo para processá-lo) preservando parte da informação revelou-se ser menos eficiente do que fora antecipado. Sem remover o terceiro nucleotídeo, temos um gráfico bastante regular (Figura 5.14).

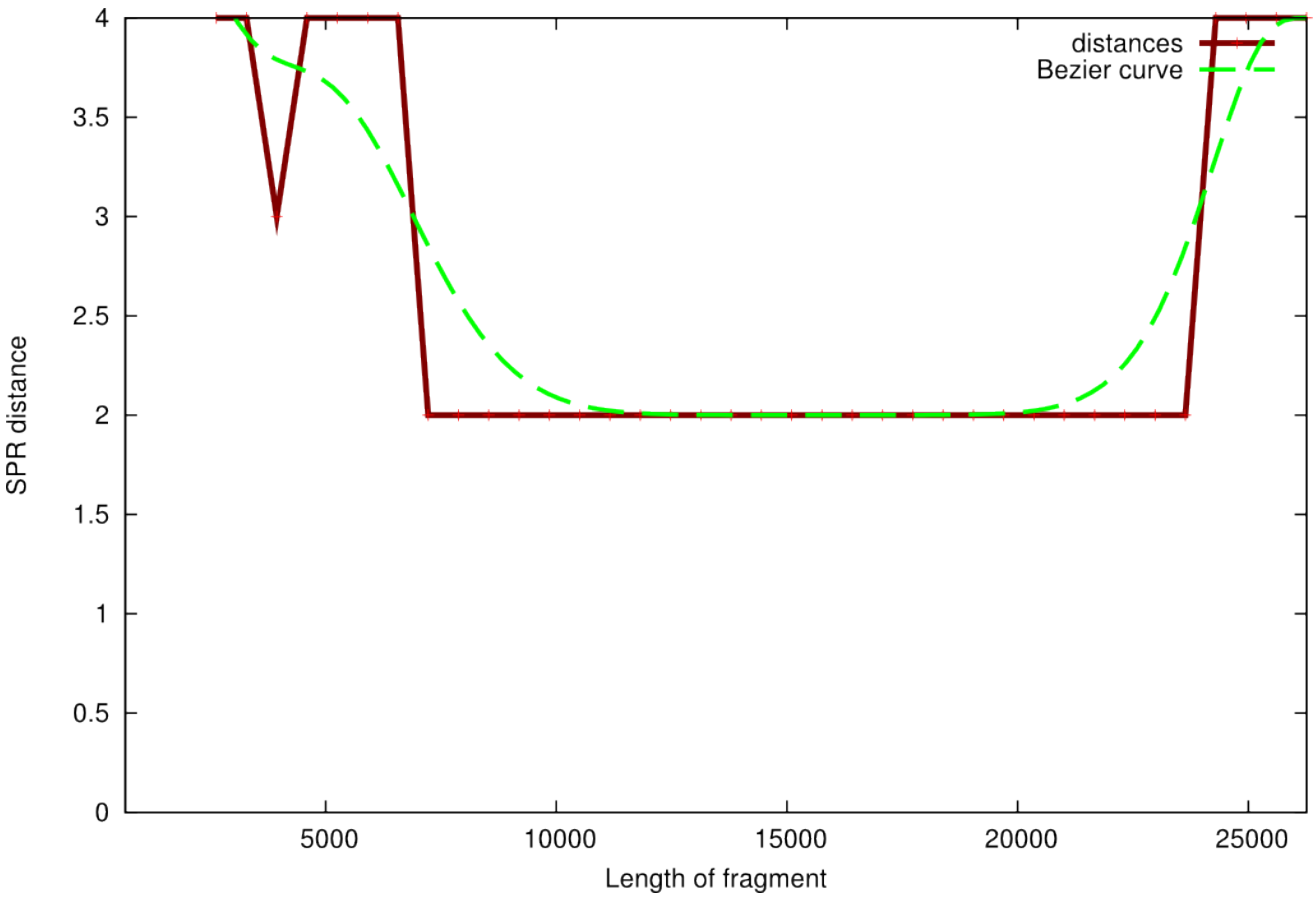


Figura 5.14: Distância da árvore ideal da filogenia do 6 genes nucleares iterando por tamanho, ordenado.

Ao se remover o terceiro nucleotídeo, no entanto, a iteração por tamanho surpreendentemente retorna para quase todos os pontos a mesma distância: os tamanhos centrais conseguem ter sua distância da árvore-gabarito auferida (igual a aproximadamente o máximo do gráfico com os nucleotídeos), mas os do extremo acabam recebendo um símbolo "Not a Number" (NaN), significando que a árvore ficou consideravelmente diferente a ponto de não ser possível ter sua distância medida. Como o software utiliza de uma biblioteca gráfica que não espera tantas incongruências, o gráfico

aparece em branco, mesmo que devesse ter uma reta central.

Não foi achado bug no software e a remoção do terceiro nucleotídeo, além de somente ter sentido para DNA, é um tópico menor. Considerou-se que é uma situação que perturba o suficiente na consistência da filogenia para não ser usada neste estudo – além de diminuir o rascunho seqüenciado para 2/3 do tamanho, algo não desejado.

5.1.5 Entropia

Por fim, procurou-se achar um parâmetro que, dado um tamanho já limitado, pudesse escolher os melhores trechos do alinhamento para a inferência da filogenia. Este parâmetro é a entropia, que preenche perfeitamente esta necessidade. Nos itens prévios, não foram apresentados os resultados com entropia, deixando-se esta análise para os últimos dados.

O que todos os algoritmos de inferência filogenética têm em comum é buscar um padrão entre as diferenças e semelhanças no conteúdo dos diferentes táxons alinhados para, a partir deste padrão, encontrar os prováveis agrupamentos (galhos) que perfazem a árvore filogenética. Os algoritmos de distância como Neighbor-Joining contam a diferença de nucleotídeos ou aminoácidos - informação digital - para com esta diferença chegar a uma distância evolutiva. O algoritmo de parcimônia busca pelas mudanças que podem ter produzido essas diferenças e a partir disso procura encontrar um caminho de menor resistência. E por fim, Máxima Verossimilhança busca, nessas mesmas diferenças, um "placar" para classificar, entre o espaço de árvores existentes, a mais provável segundo um modelo de evolução dado.

Se todos estes métodos têm em comum se concentrar nas diferenças usando as semelhanças apenas como substrato – mesmo porque as semelhanças são imensa maioria – algo que, em um rascunho, com informações faltando, atue como uma "lupa" nas diferenças, nos facilitará a visão necessária para colocar cada táxon em seu devido galho da árvore inferida.

O que se procura é a dose certa de informação a ser usada para se obter a melhor filogenia. Como uma lupa, existe uma distância ideal de seu alvo para a melhor visão: mais perto, e os detalhes se perdem. Mais longe, a visão fica embaçada.

Utiliza-se então um dado que o CalcPhyl calcula em todas as execuções incondicionalmente: o valor de entropia cada sítio do alinhamento. O tempo de processamento desta medida é quase instantâneo (computacionalmente "barato"), e obtém-se algo nos moldes da Figura 5.15, que pode ser ilustrado para cada dado que usamos:

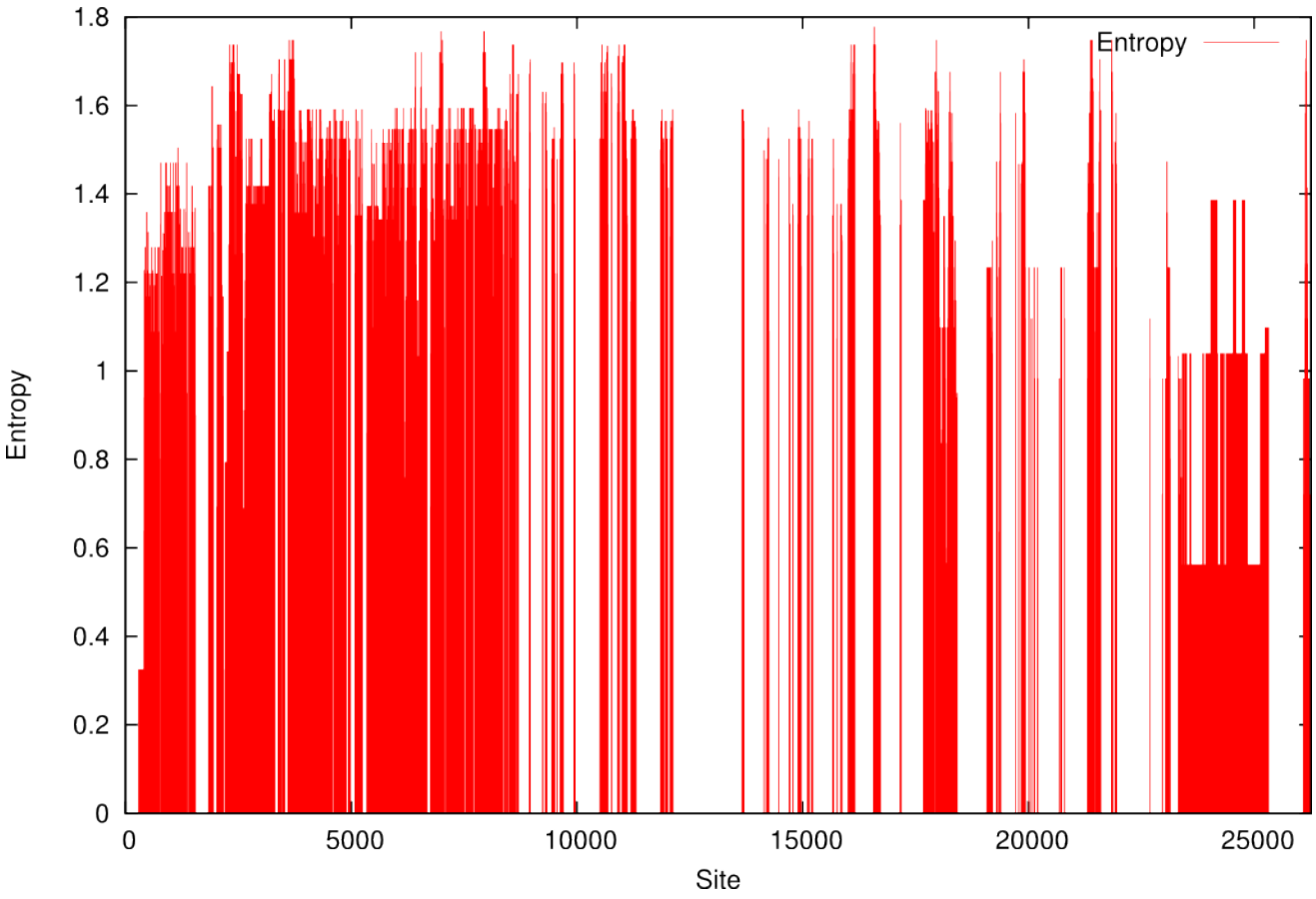


Figura 5.15: Entropia do alinhamento por sítio de 6 genes nucleares (DNA). Percebe-se no gráfico que há alguns sítios com entropia zero (significando que todos os táxons têm o mesmo nucleotídio neste sítio) ou com entropia perto do máximo de 1,79 (significando que a distribuição de estados entre os táxons no sítio determinado é uniforme e máxima).

É claramente perceptível que o alinhamento tem sítios com muito mais variabilidade do que outros. São eles que dão as diferenças de informação – a quantidade de variação e o aleatoriedade desta variação – de interesse do trabalho. Se é construído, a partir deste, um alinhamento com os sítios de maiores entropias - digamos, as **1314** posições que contenham as maiores entropias, mantendo a ordem - e é inferida a árvore usando o método de Máxima Verossimilhança, esta árvore já não estará tão distante da árvore ideal, necessitando de apenas **5** operações de recortar e colar para se transformar nesta referência.

Neste alinhamento de 1314 posições construído, o sítio com a entropia original de menor valor terá entropia de **1,524707** e alguns outros sítios terão este mesmo valor, visto que a entropia tem um número discreto de valores possíveis.

A ver o que acontece quando é construído um alinhamento ainda maior, seguindo a regra das maiores entropias, obtém-se um alinhamento de **2626** posições (aproximadamente o dobro do anterior), para então ser inferida a distância da árvore ideal. É percebido que a qualidade vai se aprimorando: neste alinhamento, são necessárias somente **4** operações de recortar e colar para obter a árvore de referência. O sítio de menor entropia original agora terá **1,424130**.

Aumenta-se a quantidade de nucleotídeos, diminuindo-se o limiar de entropia mais uma vez, até ser obtido o tamanho original de **26261** nucleotídeos. Antes disso, próximo do tamanho de **12000** nucleotídeos, já se terá amostrado pelo menos um nucleotídeo cujo valor de entropia é **zero** e nossa árvore inferida já está bastante similar à original, distante apenas **2** operações de recortar e colar. Os próximos alinhamentos a partir de 12000, portanto, só acrescentarão substrato (semelhanças), já tendo os anteriores fixado todas as diferenças. Em algum momento, aumentar o substrato aumentará a distância da árvore inferida, diminuindo a qualidade novamente.

A Figura 5.16 ilustra este conceito formando um “vale” claramente definido.

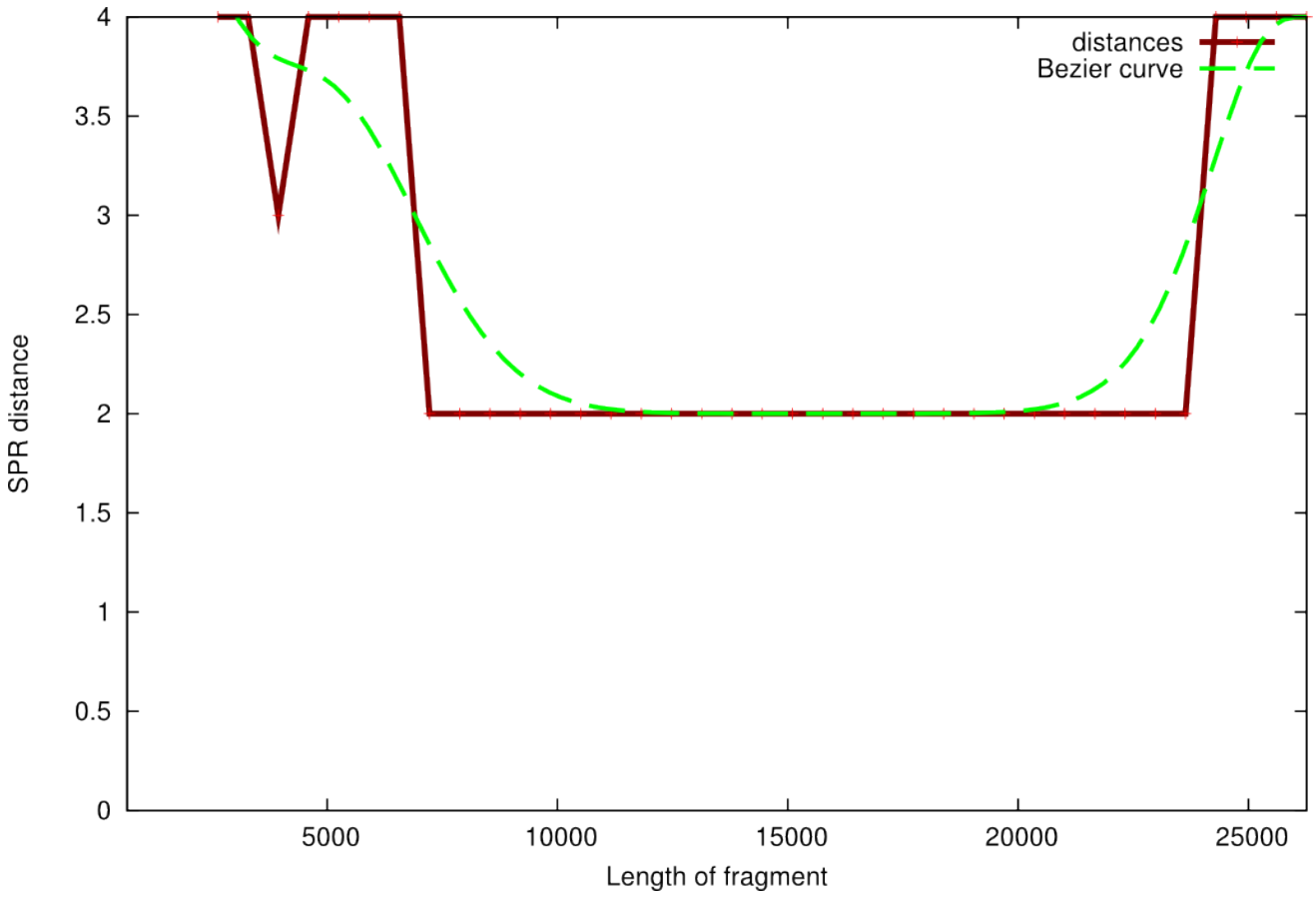


Figura 5.16: Distância da árvore ideal do alinhamento de 6 genes nucleares iterando por entropia. Note-se que é bastante claro o intervalo de melhor qualidade (menor distância da árvore ideal). Mesmo com um tamanho menor de filamento temos uma qualidade melhor do que teríamos com o gene completo.

Tem-se, ao que parece, uma boa receita para achar uma filogenia adequada, mesmo com apenas fragmentos de um trecho seqüenciado. O próximo passo é conferir se a mesma situação acontece com os genes de proteína. Vê-se o resultado na Figura 5.17 (*cox1*), Figura 5.18 (*cox2*) e Figura 5.19 (*cox3*).

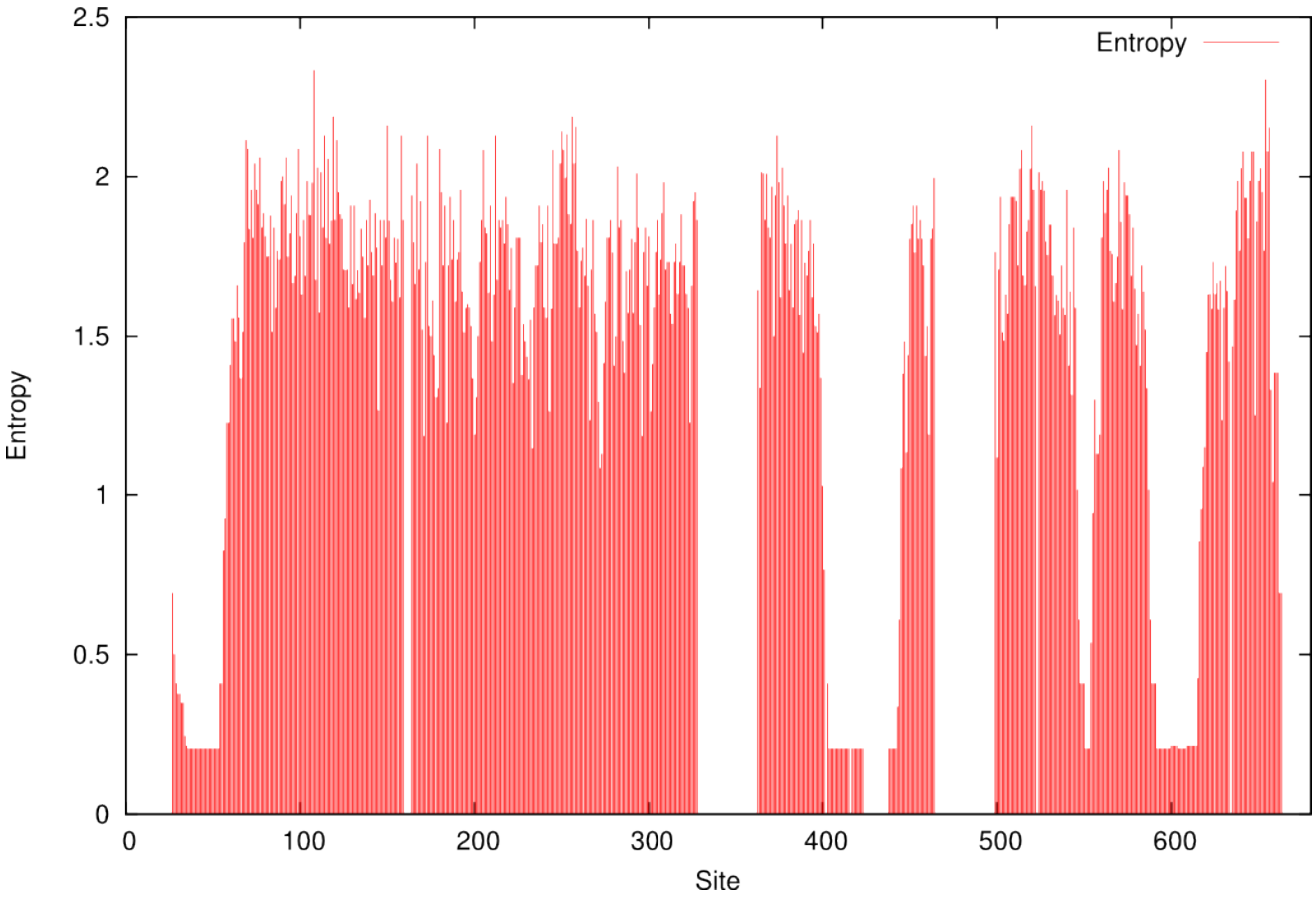


Figura 5.17: Entropia por sítio do alinhamento de gene protéico *cox1*. O alinhamento possui algumas partes de muita variabilidade (pouca correspondência entre os sítios) e pequenos trechos de concordância (entropia pequena ou zero).

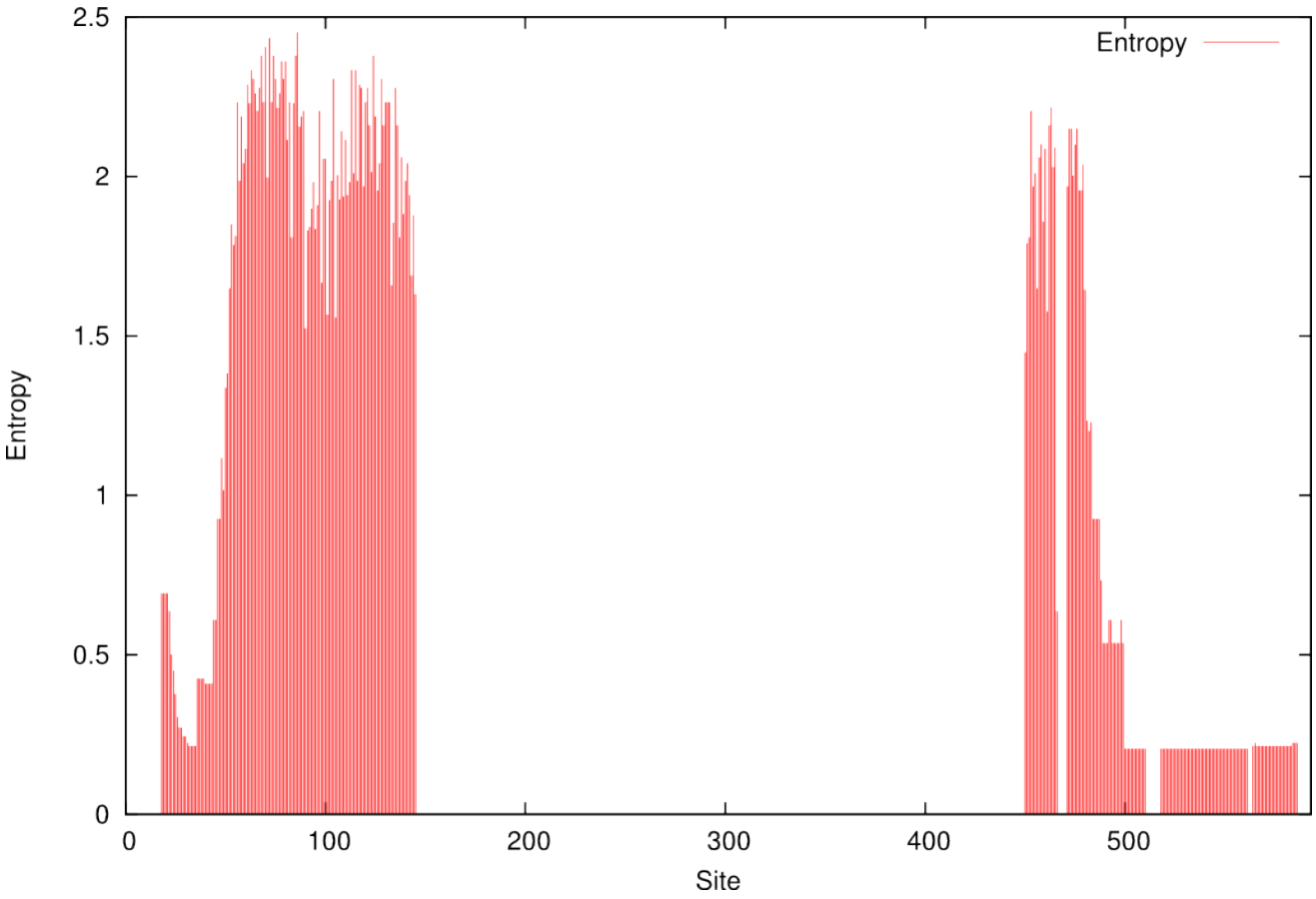


Figura 5.18: Entropia por sítio do alinhamento de gene protéico *cox2*. O alinhamento possui um trecho central de grande concordância e trechos periféricas de grande variabilidade.

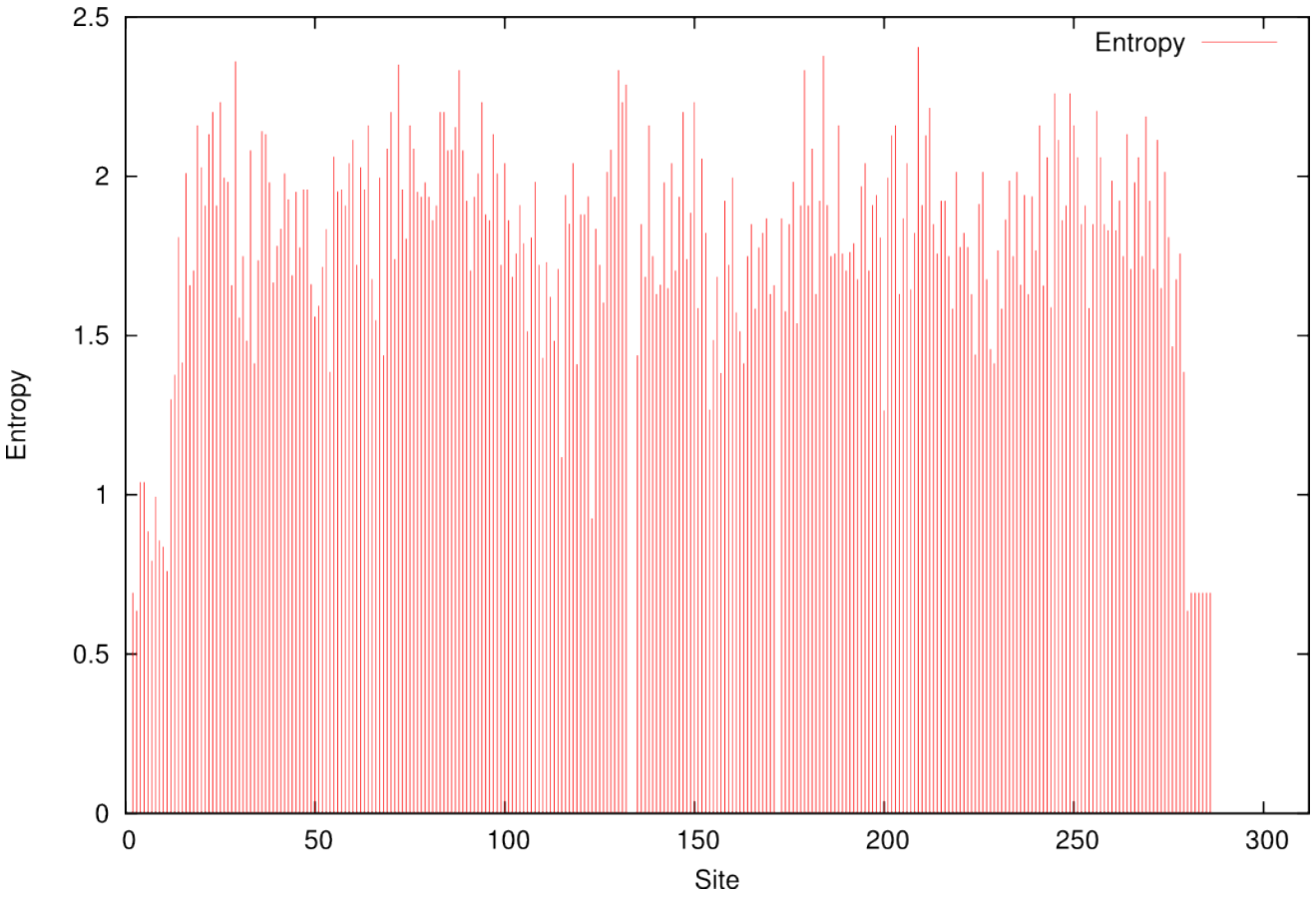


Figura 5.19: Entropia por sítio do alinhamento de gene protéico *cox3*. O alinhamento possui grande variabilidade (entropia) em praticamente toda a sua extensão.

3 genes de citocromo oxidase com tanta diferença em variação são uma boa "prova de fogo" para o critério determinado: se se conseguir, como no caso dos genes nucleares, uma boa qualidade de filogenia (pequena distância) em determinada região do gráfico (um vale central), ter-se-á logrado o objetivo do trabalho.

O resultado é obtido para os genes de proteína *cox1* (Figura 5.17), *cox2* (Figura 5.18) e *cox3* (Figura 5.19).

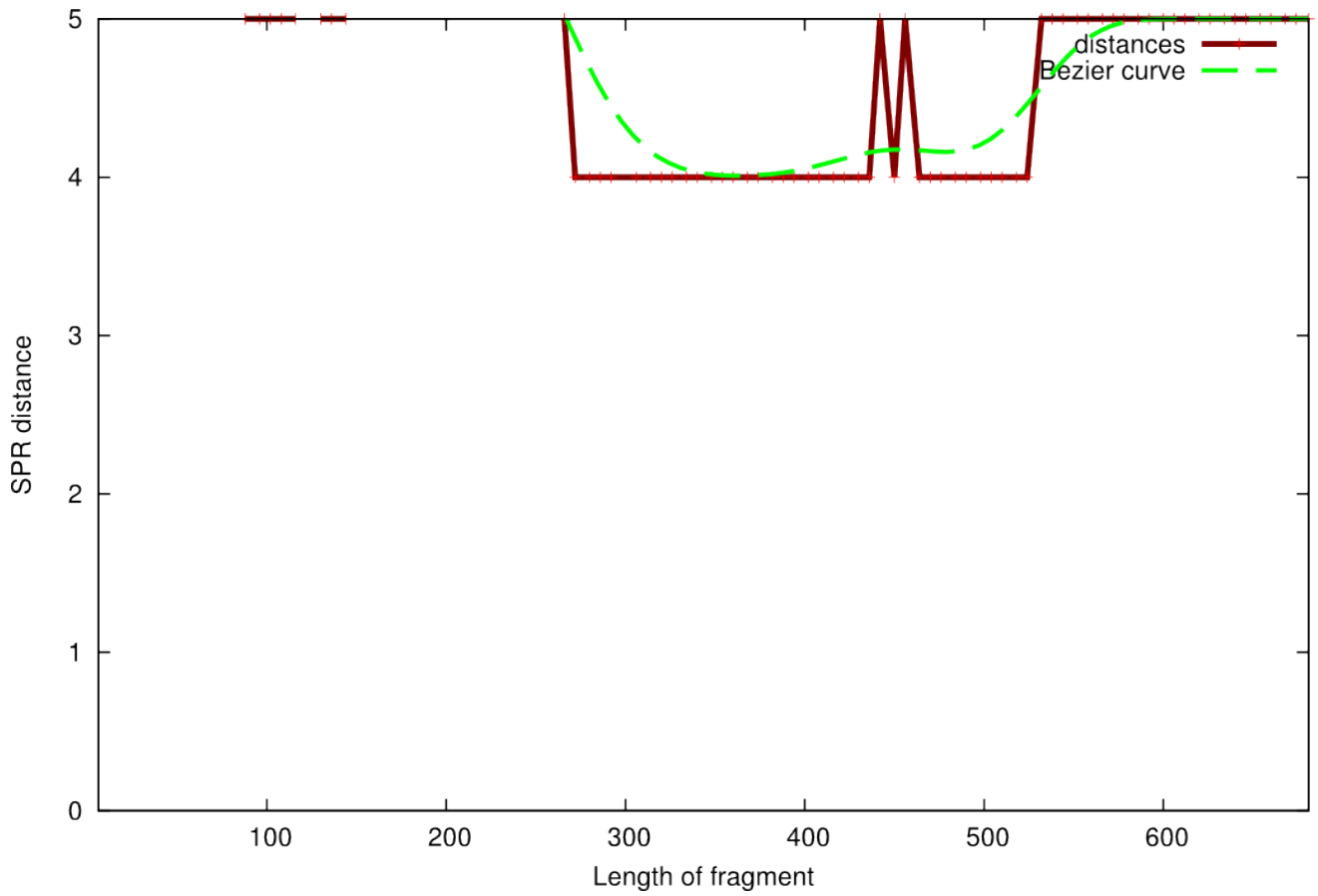


Figura 5.20: Distância à àrvore ideal do gene *cox1*, iterando por entropia.

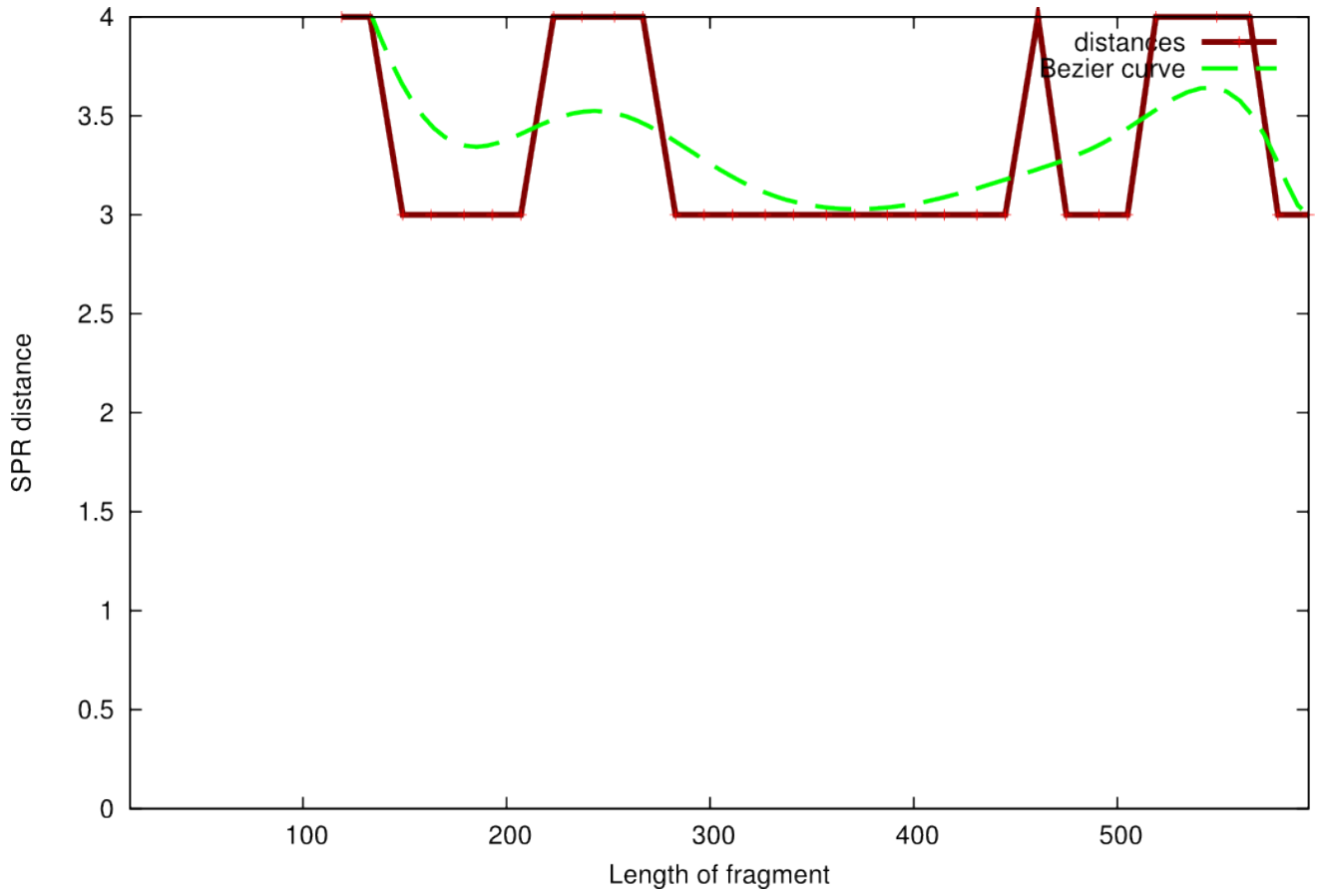


Figura 5.21: Distância à àrvore ideal do gene *cox2*, iterando por entropia.

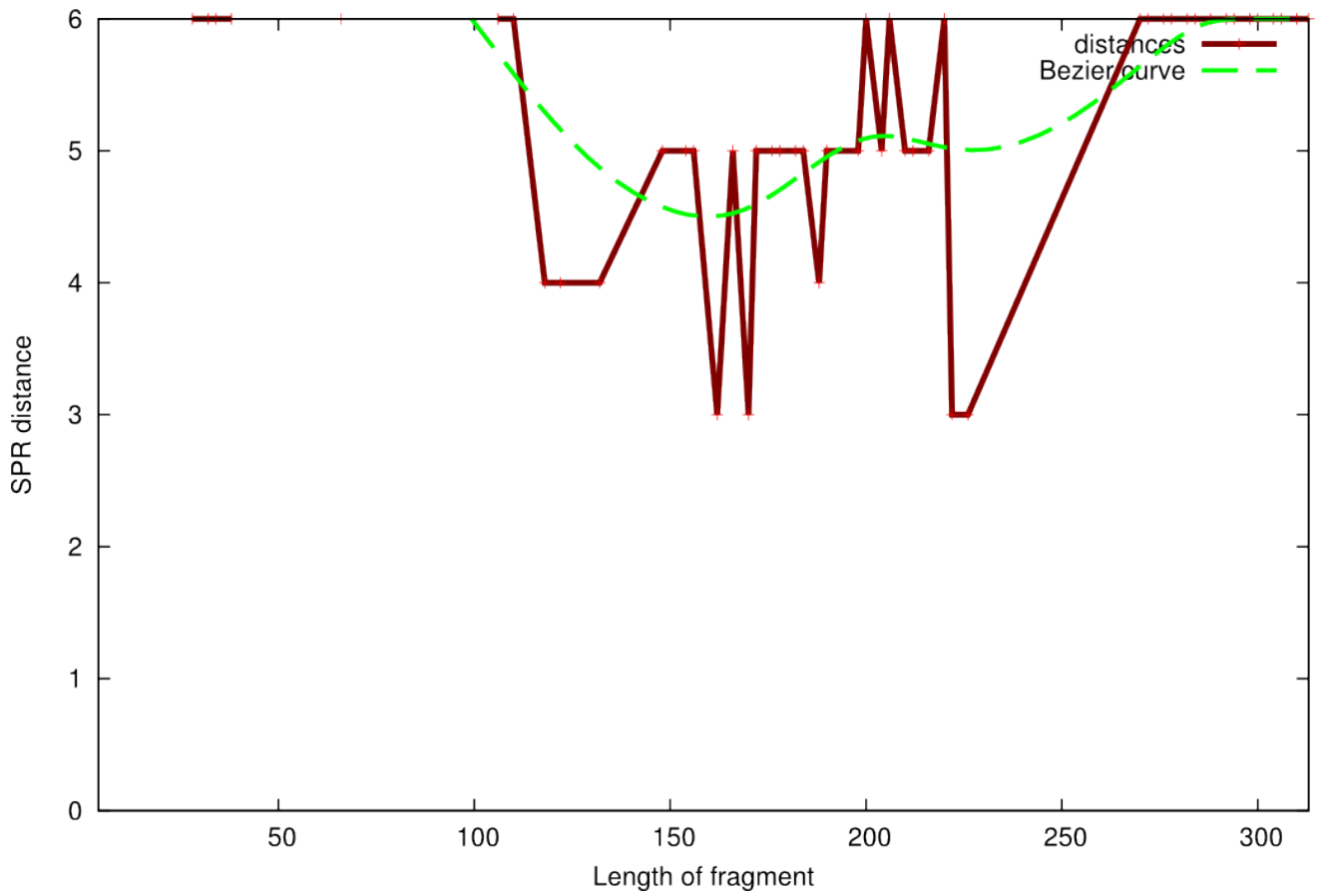


Figura 5.22: Distância à àrvore ideal do gene *cox3*, iterando por entropia.

Note-se a presença de vales de melhor qualidade - ainda que com pequena diferença, pois os dados da amostra, contendo poucos elementos, são menos ideais para uma inferência filogenética - em todos os gráficos. Com genes com distribuições de entropias por sítio tão díspares dando resultados tão consistentes, chega-se a um resultado bastante usável.

6 Conclusões

- Logrou-se o objetivo almejado, de não somente auferir a qualidade de um gene incompleto para a produção de uma filogenia quanto de arrumar pelo menos um critério útil para selecionar os trechos mais informativos de um gene incompleto com o qual nos deparemos.

- Este critério é o da entropia de Shannon para um alinhamento. Selecione-se uma fração com maior entropia deste alinhamento para inferir a filogenia, e ela terá uma qualidade bem próxima à da ideal.

- Mesmo não usando a entropia, percebe-se que usar um trecho incompleto do gene (de 60% a 80% de tamanho) freqüentemente não traz diminuições na qualidade da inferência, quando se usa um método preciso (Máxima Verossimilhança).

- Embaralhar o alinhamento para diminuir a influência de posição, no entanto, parece ter o efeito de gerar um pequeno decréscimo na influência do tamanho para a filogenia, tornando-o mais consistentemente distante da ideal.

- Usar um rascunho de boa qualidade logo no início de um processo de seqüenciamento para gerar a filogenia melhora e acelera o próprio processo de seqüenciamento e anotação, por permitir contextualizar o segmento genético cedo (em paralogia ou ortologia, por exemplo), o que cumulativamente torna o resto mais rápido.

- Como vimos, a literatura da filogenia mostra que filtrar alinhamentos pela entropia mínima dá bons resultados, e chegamos a essa conclusão de forma independente, sem de antemão conhecer o artigo que o dizia. Seria interessante, no entanto, conhecer o efeito de outras filtragens de sítios por entropia - filtrar um intervalo, ou filtrar pela máxima entropia. Como seria o efeito na qualidade?

Nomenclatura

Apomorfia: Um estado de caráter derivado.

Caráter: Um recurso variável que pode assumir um entre diferentes estados.

Caracteres ordenados: Caracteres onde a transição entre estados segue (ou se presume que siga) um padrão particular, limitando as transições permitidas entre estados.

Caracteres reversíveis: Caracteres onde o padrão de mudança é reversível.

Clado: Um grupo (monofilético) de organismos relacionados por descendência a um ancestral comum.

Cladograma: A representação cladística de uma filogenia, onde somente a ordem de ramificação é mostrada.

Curva bezier: Uma representação suavizada da curva que liga todos os pontos em um gráfico de distribuição de medidas.

Dendrograma: A representação gráfica de uma filogenia.

Distância: A medida de diferença entre dois objetos.

Entropia: Medida física da quantidade de desordem de um sistema.

Entropia de Shannon: Medida da Teoria da Informação relacionada com a Entropia física, denotando a desordem do sistema como sua quantidade de variação.

Exaustivo: Diz-se do método que examina todos os arranjos ou combinações no espaço de possibilidades.

Filogenia: Hipótese para as relações entre os organismos.

Filogenias enraizadas: Filogenias onde pelo uso de um *outgroup* o último ancestral comum do clado da OTU sob consideração pode ser colocado.

Filogenias não-enraizadas: Filogenias onde nenhum *outgroup* é especificado.

Filograma: A representação de uma filogenia onde as relações evolucionárias são mostradas tanto pela ordem de ramificação quanto por uma medida de distância.

Galho: Um segmento ligando um nó a outro, ou um nó com uma OTU terminal.

Genes conservados: Genes que durante sua história evolutiva sofreram poucas mudanças.

Genes COX: Genes mitocondriais da proteína Citocromo Oxidase, altamente conservados.

Grupo externo: Um grupo de OTUs para o qual se assume, *a priori*, que estejam fora da monofilia da OTU sob análise. Usado para dar direção à filogenia resolvida onde todas as relações são representadas como bifurcações.

Grupo interno: A OTU sob análise.

Homologia: Ancestralidade comum entre dois genes, caracteres ou posições.

Homoplasia: Derivação independente do estado de um caráter em duas linhagens.

Ingroup: O mesmo que grupo interno.

Monofilético: Um clado onde todas as OTUs descendem de um único ancestral comum mais recente, e que inclui todos os descendentes deste ancestral.

Não-Exaustivo: Diz-se do método que analisa somente um subconjunto dos elementos do espaço de possibilidades.

Nó ou ***nodo***: Um ponto de bifurcação de uma árvore (um OTU ancestral presumido).

Paralogia: Homologia surgida a partir de duplicação de gene.

Plesiomorfia: O estado do caráter ancestral.

Ortologia: Homologia "verdadeira", surgida a partir de especiação.

Outgroup: O mesmo que grupo externo.

OTU: Operational Taxonomic Unit ou Unidade Taxonômica Operacional, um termo não-comprometedor usado para designar os objetos de estudo filogenéticos, sejam genes, populações, espécies ou indivíduos.

Parafilético: Um grupo taxonômico que não inclui todos os descendentes de um táxon ancestral.

Polifilético: Um grupo taxonômico que depende de mais de um táxon ancestral.

Saturação: O fenômeno de múltiplas substituições no mesmo sítio (especialmente de mutações do tipo transição no terceiro nucleotídeo de cada códon), que apaga o sinal filogenético .

Similaridades: A medida da semelhança entre dois objetos.

Simplesiomorfia: Um estado de caráter ancestral compartilhado.

Sinal filogenético: Quantidade de informação que se pode inferir de um alinhamento para gerar sua filogenia.

Sinapomorfia: Um estado de caráter derivado compartilhado (em referência a uma hipótese filogenética)

Referências Bibliográficas

Aiyar, A. 1999. The use of clustal w and clustal x for multiple sequence alignment. Bioinformatics Methods and Protocols **132**(132): 221.

Beiko, R. G. e Hamilton, N. 2006. Phylogenetic identification of lateral genetic transfer events. BMC Evolutionary Biology **6**(1): 15.

Bininda-Emonds, O. R. 2000. Factors influencing phylogenetic inference: A case study using the mammalian carnivores. Molecular Phylogenetics and Evolution **16**(1): 113.

Bininda-Emonds, O. R. et al. 2001. Scaling of accuracy in extremely large phylogenetic trees. Pac Symp Biocomput 547–558.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution **39**(4): 783.

Felsenstein, J. 2005. Phylip (phylogeny inference package) version 3.6. Department of Genome Sciences, University of Washington, Seattle .

Hall, T. A. 1999. Bioedit: a user-friendly biological sequence alignment editor and analysis. Nucleic acids symposium series .

Hannula, M. e Hanninen, M.-L. 2007. Phylogenetic analysis of helicobacter species based on partial gyrb gene sequences. International Journal of Systematic and Evolutionary Microbiology **57**(3): 444.

Hickey, G. et al. 2008. Spr distance computation for unrooted trees. Evol Bioinform Online **4**: 17–27.

Hillis, D. M. e Bull, J. J. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Systematic Biology **42**(2): 182–192.

James, T. Y. et al. 2006. Reconstructing the early evolution of fungi using a six-gene phylogeny. Nature **443**(7113): 818.

Maria, J. W. e R. Servedio, J. 1998. Phylogenetic analysis and intraspecific variation: Performance of parsimony, likelihood, and distance methods. Systematic Biology **47**(2): 228.

Matioli, S. R. e Russo, C. A. M. 2001. Biologia Molecular e Evolucao, Holos Editora, chap. 12, 130–136.

Miyazaki, S. et al. 1996. The efficiency of entropy evolution rate for construction of phylogenetic trees. Genes and Genetic Systems **71**(5): 323.

Pond, S. L. K. e Muse, S. V. 2005. Hyphy: Hypothesis testing using phylogenies. Statistical Methods in Molecular Evolution 125.

Retief, J. D. 2000. Phylogenetic analysis using phylip. Methods Mol Biol **132**: 243–258.

Russo, C. A. M. et al. 1996. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. Molecular Biology and Evolution **13**(3): 525–536.

S, G. e O., G. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology **52**(5): 696–704.

Sanderson, M. J. 1995. Objections to bootstrapping phylogenies: A critique. Systematic Biology **44**(3): 299.

Servedio, J. W. M. R. 1998. Phylogenetic analysis and intraspecific variation: Performance of parsimony, likelihood, and distance methods. Systematic Biology **47**(2): 228.

Shannon, C. E. 2001. A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review **5**(1): 3.

Stajich, J. E. 2002. The bioperl toolkit: Perl modules for the life sciences. Genome Research **12**(10): 1611.

Williams, T. et al. 2004. gnuplot 4.4: An Interactive Plotting Program. URL `http://www.gnuplot.info/docs/gnuplot.html`.