

Tutorial 7

Dados Moleculares - Introdução

BIZ0433 - INFERÊNCIA FILOGENÉTICA: FILOSOFIA, MÉTODO E APLICAÇÕES.

Conteúdo

Objetivo	108
7.1 GenBank	109
7.1.1 Formatos de arquivos	109
7.1.2 Obtendo sequências do GenBank	111
7.2 Manipulação de arquivos de sequência	114
7.2.1 Compondo arquivos FASTA	114
7.2.2 Verificação e edição FASTA em AliView	117
7.3 Configuração de matrizes de dados moleculares	122
7.3.1 SequenceMatrix	123
7.4 Referências	128

Objetivo

O objetivo deste tutorial é apresentar ao aluno algumas ferramentas de manipulação de dados moleculares desde a obtenção de dados no GenBank/NCBI à configurações de matrizes de dados passíveis de serem analisadas em programas de análise filogenética. Ao concluir esse tutorial, o aluno será capaz de reanalisar dados já publicados utilizando TNT. Os arquivos associados a este tutorial estão disponíveis no [GitHub](#). Você baixar todos os tutoriais com o seguinte comando:

```
svn checkout https://github.com/fplmarques/cladistica/trunk/tutorials/
```

7.1 GenBank

7.1.1 FORMATOS DE ARQUIVOS

Antes de entrarmos na parte analítica de dados moleculares é necessário entender alguns formatos de arquivos que contenham dados moleculares. Há vários deles, mas em caráter introdutório iremos apresentar apenas dois. Se você está familiarizado com artigos que usam dados moleculares para inferência filogenética, já deve ter notado que grande parte deles deposita sequências nucleotídicas ou peptídicas em um repositório chamado GenBank. De fato, todos os periódicos obrigam autores a depositarem esses dados em repositórios como o **GenBank** e informar aos leitores os números de tombo. Observe a tabela no APPENDIX I de Dias *et al.* [1], publicação anexada a este tutorial. As colunas sob o termo "GenBank accession number" contém os números de tombo das sequências nucleotídicas ou peptídicas depositadas pelos autores para cada uma das regiões genômicas consideradas nesse estudo (*i.e.*, cyt b, 16S, Rag-1 e 12S). Estes números de tombo permitem que qualquer pessoa verifique e use esses dados.

O acesso ao banco de sequências nucleotídicas do GenBank é feito pelo endereço: <http://www.ncbi.nlm.nih.gov/nucleotide>. Se você possui um determinado número de tombo, por exemplo FJ685663 – que de acordo com a tabela apresentada por Dias *et al.* [1] refere-se ao fragmento de Citocromo B de *Cycloramphus acangatan*–, você poderá obter os dados deste número de tombo simplesmente usando-o na busca disponível nessa página.

Por *default*, quando você executa uma busca no GenBank com um único número de tombo, o GenBank retorna o resultado em seu formato mais detalhado. Neste caso em particular isso seria:

```
Cycloramphus acangatan voucher AF1605 cytochrome b (cytb) gene, partial cds; mitochondrial
GenBank: FJ685663.1
FASTA Graphics PopSet
Go to:
LOCUS      FJ685663                611 bp    DNA        linear    VRT 15-APR-2009
DEFINITION Cycloramphus acangatan voucher AF1605 cytochrome b (cytb) gene,
           partial cds; mitochondrial.
ACCESSION  FJ685663
VERSION    FJ685663.1  GI:226510745
KEYWORDS   .
SOURCE     mitochondrion Cycloramphus acangatan
  ORGANISM Cycloramphus acangatan
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Amphibia; Batrachia; Anura; Neobatrachia; Hyloidea; Cycloramphidae;
            Cycloramphus.
REFERENCE  1  (bases 1 to 611)
  AUTHORS  Amaro,R.C., Pavan,D. and Rodrigues,M.T.
  TITLE    On the generic identity of Odontophrynus moratoi Jim & Caramaschi,
            1980 (Anura, Cycloramphidae)
  JOURNAL  Zootaxa 2071, 61-68 (2009)
REFERENCE  2  (bases 1 to 611)
```

```

AUTHORS    Amaro,R.C., Pavan,D. and Rodrigues,M.T.
TITLE      Direct Submission
JOURNAL     Submitted (30-JAN-2009) Departamento de Zoologia, Instituto de
            Biociencias, Universidade de Sao Paulo, Rua do Matao, ravessa 14,
            101, Sao Paulo 05508-090, Brazil
FEATURES    Location/Qualifiers
            source             1..611
                                /organism="Cycloramphus acangatan"
                                /organelle="mitochondrion"
                                /mol_type="genomic DNA"
                                /specimen_voucher="AF1605"
                                /db_xref="taxon:635146"
            gene               <1..>611
                                /gene="cytb"
            CDS                <1..>611
                                /gene="cytb"
                                /codon_start=3
                                /transl_table=2
                                /product="cytochrome b"
                                /protein_id="AC059916.1"
                                /db_xref="GI:226510746"
                                /translation="LIMQIAPGLFLAMHYTADTSLAFSSIAHICRDVNNGWLLRSLHA
                                NGASFFFCICIYLIHIGRGLYYGSYLFKETWNIGVILLFMTMATAFVGIVLPWQMSFWG
                                ATVITNLLSAAPYIGTDLVQWIWGGFSVDNATLTRFFTFHFILPFIVTGLILLHLIFL
                                HETGSSNPTGLNPNPKVPFHTYFSYKDILGFAIMLSLLASLS"
ORIGIN
            1  gcttaattat acaaattgca ccaggactat tcttagccat acattacacc gcggatacct
            61  cattagcatt ttcatctatt gcccatatct gccgagatgt aaacaacggg tgactttctt
            121  gaagcctcca cgcaaatggg gcctcattct tctttatctg tatctacctt catattggcc
            181  gaggattata ctacggctca tacttattta aagaaacatg aaacattgga gtgattctcc
            241  tatttataac catagccaca gcctttgtcg gatacgtagt accatgagga caaatatcat
            301  tctggggggc cacagtcac accaacttat tatctgcagc cccctatatt ggcacagact
            361  tagtgcaatg aatctgaggc ggattttcag tagataacgc caccctcaca cgcttcttca
            421  catttcactt tctcctccca ttcatgttta caggattaat cctcctacac ctaatctttc
            481  ttcatgaaac aggatcttca aacccacag gcctaaaccc taaccagat aaagtcccat
            541  tccacaccta cttctctat aaagacatcc taggatttgc catcatactc tcccttcttg
            601  cctcactatc a

```

Observe o resultado da busca acima no qual o GenBank lhe fornece uma série de informações sobre essa sequência de nucleotídeos, tais como: autores, região genômica, posicionamento taxonômico do organismo sequenciado, tradução para aminoácidos (para regiões codificantes) e a sequência nucleotídica propriamente dita. Desta forma, este é o formato que contém o maior número de informações sobre a sequência nucleotídica ou peptídica associada ao número de tombo. Minha sugestão é que sempre que possível, obtenha as sequências no Genbank neste formato, pois isso evita ter que revisitar este banco de dados para obter informações adicionais no futuro caso você esteja usando dados disponíveis nesse repositório.

Há um outro formato que é muito utilizado na manipulação e análise de sequências nucleotídicas ou peptídicas: o formato **FASTA**. Na sua forma mais simples, ele contém terminais precedidos

de ">" e em outra linha a sequência nucleotídica ou peptídica. Por exemplo:

```
>sequencia_1
ACGTACGTACGT
>sequencia_2
AAGTAAGTAAGT
>sequencia_3
AAATAASTAAAT
```

Esse formato é bem simples, útil e frequentemente utilizado para transformar sequências nucleotídicas ou peptídicas em matrizes de dados. No entanto, este formato permite inserir comentários e outras informações no cabeçalho da sequência. Vejamos por exemplo como o GenBank retornaria a mesma sequência acima (*i.e.*, FJ685663) no formato **FASTA**:

```
Cycloramphus acangatan voucher AF1605 cytochrome b (cytb) gene, partial cds; mitochondrial
GenBank: FJ685663.1
GenBank Graphics PopSet
>gi|226510745|gb|FJ685663.1| Cycloramphus acangatan voucher AF1605 cytochrome b (cytb) ...
GCTTAATTATACAAATTGCACCAGGACTATTCTTAGCCATACATTACACCGCGGATACCTCATTAGCATT
TTCATCTATTGCCATATCTGCCGAGATGTAAACAACGGGTGACTTCTTGAAGCCTCCACGCAAATGGT
GCCTCATTCTTTATCTGTATCTACCTTCATATTGGCCGAGGATTATACTACGGCTCATACTTATTTA
AAGAAACATGAAACATTGGAGTGATTCTCTATTATAACCATAGCCACAGCCTTTGTCGGATACGTACT
ACCATGAGGACAAATATCATTCTGGGGGGCCACAGTCATCACCAACTTATTATCTGCAGCCCCCTATATT
GGCAGAGACTTAGTGCAATGAATCTGAGGCGGATTTTCAGTAGATAACGCCACCCCTCACACGCTTCTTCA
CATTTCACTTTATCTCCCATTCATTGTTACAGGATTAATCCTCCTACACCTAATCTTTCTTCATGAAAC
AGGATCTTCAAACCCACAGGCCTAAACCCTAACCCAGATAAAGTCCATTCCACACCTACTTCTCTCTAT
AAAGACATCCTAGGATTTGCCATCATACTCTCCCTTCTTGCTCACTATCA
```

7.1.2 OBTENDO SEQUÊNCIAS DO GENBANK

Todos os arquivos de sequências são arquivos texto. Portanto, nada impede que você copie e cole os dados diretamente da página no Genbank no formato que desejar em um editor de texto. Isso porém pode se tornar inviável e extremamente tedioso – dependendo do número de sequências que você precisa retirar do repositório. O GenBank possui uma ferramenta para retirar sequências nucleotídicas de forma bem eficiente. Essa ferramenta é chamada *Batch Entrez* e pode ser acessada pelo endereço <http://www.ncbi.nlm.nih.gov/sites/batchentrez>. A única coisa que você precisa para acessar essa ferramenta é possuir um arquivo texto contendo todos os números de tombo que deseja.

Vejamos o exemplo abaixo. Suponha que você possua um arquivo com o seguinte conteúdo¹:

```
FJ685663
FJ685664
FJ685665
...
KF214177
```

¹ Estes são alguns números de tombo para as sequências de Citocromo B de Dias *et al.* [1]

KF214178

FJ685662

Basta acessar a página do *Batch Entrez*/Genbank e fazer o *upload* do seu arquivo selecionando a opção *Choose File* (veja Figura 7.1) e em seguida pressionar *Retrieve*.

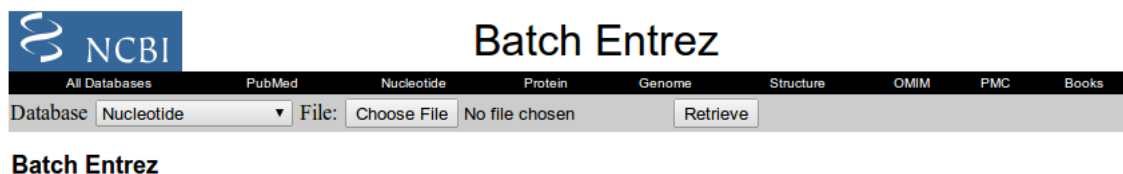


Figura 7.1: Página do *Batch Entrez* do GenBank.

Você deverá obter:

```
Received lines: 40
Rejected lines: 0
Removed duplicates: 0
Passed to Entrez: 40
Retrieve records for 40 UID(s)
```

Neste exemplo em particular, meu arquivo possuía 40 números de tombo e todos eles foram recuperados sem erro.

O próximo passo é obter os registros das sequências pressionando "Retrieve records for 40 UID(s)". Você deverá obter o resultado ilustrado na Figura 7.2.

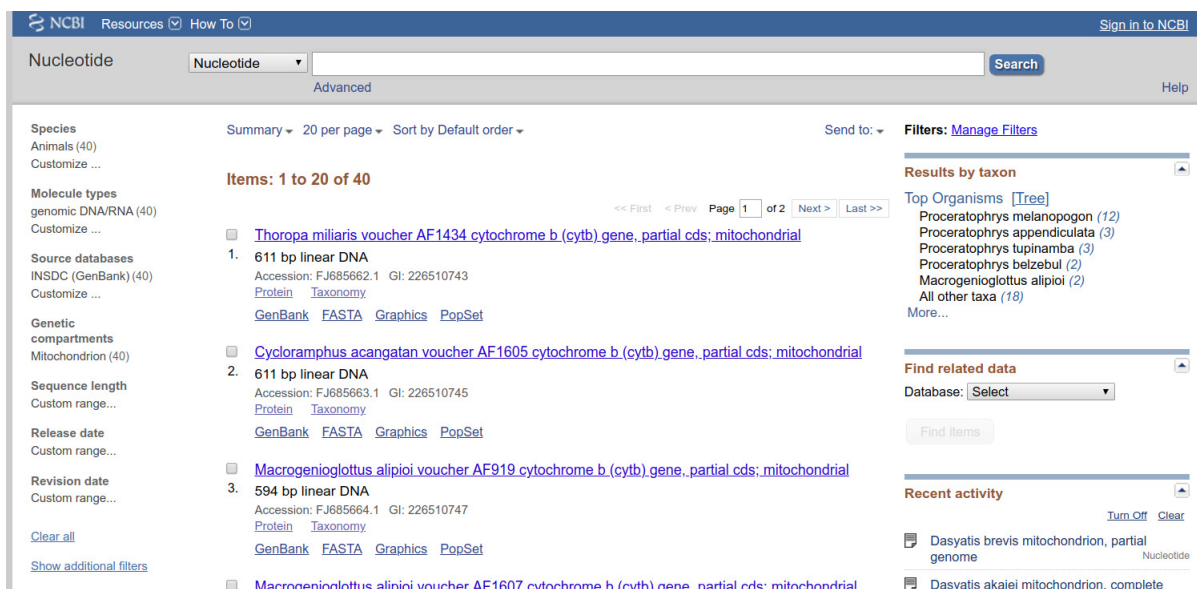


Figura 7.2: Página do *Batch Entrez* do GenBank exibindo resultado de uma busca.

Por *default* o GenBank irá apresentar as sequências solicitadas de forma sumarizada (veja Figura 7.2). Isto lhe permite verificar se os números de acesso de fato referem-se às sequências que estava interessado e/ou selecionar parte delas para exame mais detalhado e/ou para baixá-las. Para baixar

essas sequências, você precisa definir o formato que deseja utilizar. A definição do formato deve ser selecionada dentre as opções disponíveis nas opções de *Summary* (veja Figura 7.3).

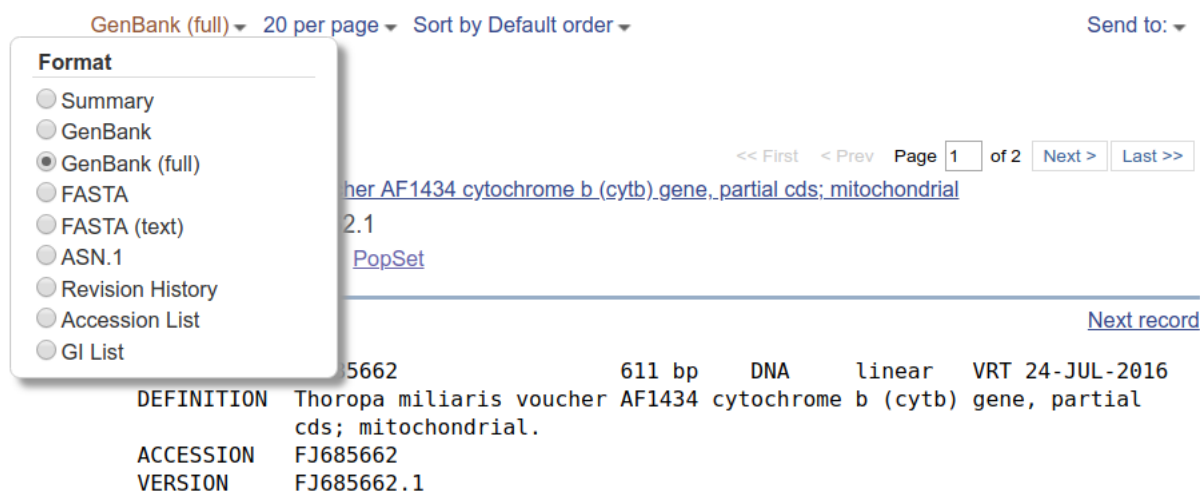


Figura 7.3: Página do *Batch Entrez* do GenBank exibindo as opções para exibição dos resultados de busca.

Ao selecionar, por exemplo, o formato *Genbank (full)*, você deverá obter a exibição das sequências no formato referido.

No entanto, para baixar as sequências em um determinado formato, não é necessário exibí-las da mesma forma. O *download* de sequências é executado pela seleção das opções em *Send to* (veja Figura 7.3). Ao pressionar *Send to*, você deverá obter uma janela com as opções exibidas na Figura 7.4.

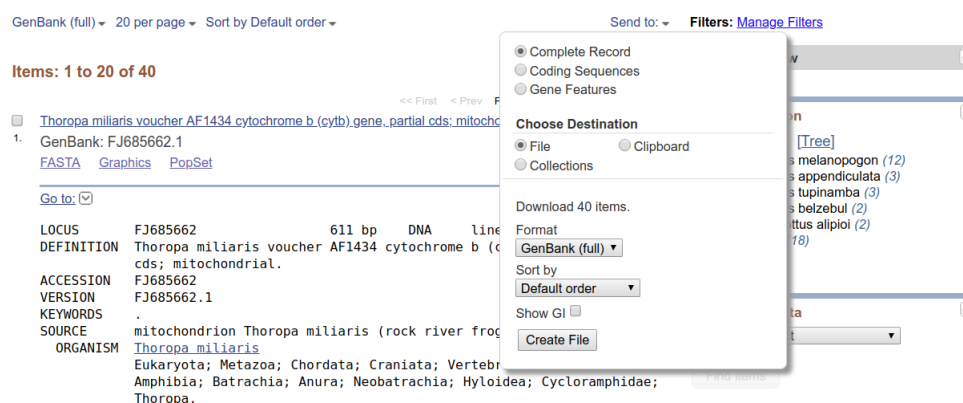


Figura 7.4: Página do *Batch Entrez* do GenBank exibindo as opções para baixar os resultados de busca.

Uma vez selecionadas as opções que deseja nesta janela, basta pressionar *Create File*. O arquivo será salvo com o nome de "sequence.gb" no diretório de *default* para *download* de seu sistema. Minha recomendação é que você modifique o nome deste arquivo de forma que seu nome indique seu conteúdo logo após baixar as sequências – ao fazê-lo, mantenha a extensão ".gb" para indicar o formato do arquivo.

Exercício 7.1

O arquivo "dias_et_al_2013_accession_numbers.csv" contém todos os números de tombo para as sequências utilizadas por Dias *et al.* [1] em um arquivo texto no formato CSV (*i.e.*, *Comma Separated Values*). Arquivos neste formato podem ser abertos em programas que manipulam planilhas, tais como EXCEL da Microsoft e OpenOffice Calculator. Este arquivo em particular possui 5 colunas, cada uma referente a uma região genômica utilizada por Dias *et al.* [1]. Para cada uma destas colunas você deverá

- i. Gerar um arquivo texto que contenha apenas os números de tombo referentes a respectiva região genômica. Para cada um dos arquivos, utilize a extensão ".acc" – de *accession* (*e.g.*, 12s.acc). Esta não é uma extensão particular, somente uma sugestão para que você possa mater seus arquivos organizados.

Dica: Inspeção o arquivo em CSV, você poderia usar a seguinte linha de comando para extrair os números de acesso do cyt-b, que estão na segunda coluna, com a seguinte linha de comando: `tail -n +2 dias_et_al_2013_accession_numbers.csv | cut -d',' -f5 | sed 's/"//g' > 12s.acc`

- ii. Utilizar a ferramenta do *Batch Entrez* para recuperar os registros referentes a cada um dos arquivos.

Atenção: Ao fazer o download do GenBank/NCBI o repositório criará um documento chamado `sequence.gb` que será salvo no diretório de Download de seu sistema.

- iii. Salvar as sequências de cada região genômica em seus respectivos arquivos no formato *GenBank (full)*. Para cada um dos arquivos, utilize a extensão ".gb" – de *GenBank* (*e.g.*, 12s.gb).

7.2 Manipulação de arquivos de sequência

Nosso objetivo a partir deste momento é extrair as informações dos arquivos gerados pelo GenBank para transformar essas sequências em um arquivo de dados que possa ser analisado por programas de inferência filogenética. No entanto, programas distintos são capazes de ler arquivos em formatos diferentes. No que se segue, manipularemos os arquivos contendo as sequências do Genbank com o objetivo de gerar uma matriz de dados que possa ser analisada em TNT [2]. Os passos apresentados a seguir poderão ser modificados à medida que você se familiarize com as inúmeras ferramentas disponíveis para manipulação de arquivos de sequência. Seja qual for a ferramenta que você venha a escolher no futuro, há dois componentes principais que eu considero importante na preparação de arquivos de sequências para análises filogenéticas. O primeiro deles é o nome dos terminais. O segundo é verificar se as sequências estão alinhadas e no mesmo sentido.

7.2.1 COMENDO ARQUIVOS FASTA

Os arquivos no formato *GenBank (full)* possuem um grande número de informações. No entanto, muitas delas não podem ser inseridas em um arquivo que se destine a análise filogenética. Para este último propósito, os nomes dos terminais devem ser curtos e bem definidos e as sequências devem estar alinhadas². O grande número de informações contido nos arquivos no formato *GenBank*

² Alinhamentos serão discutidos com detalhes nos próximos tutoriais.

(full), por outro lado, permite que nós possamos extrair o que queremos.

A ferramenta que iremos utilizar nesta primeira etapa de manipulação de arquivos de sequência é o *script* chamado *phyloconvert*. Esse *script* foi concebido primariamente para gerar arquivos no formato FASTA para serem analisados diretamente ou serem subsequentemente transformados em arquivos passíveis de serem analisados em TNT e/ou PAUP [2, 3].

O *script* chamado *phyloconvert* faz parte dos aplicativos disponíveis na imagem utilizada no curso, mas caso você não esteja utilizando a imagem, ele está disponível no diretório "tutorial_7". Este *script* requer que BioPython esteja instalado no seu sistema.

A execução do *script* chamado *phyloconvert* é feita do terminal:

```
$ phyloconvert3
```

Ao executá-lo você obterá as seguintes opções:

```
PhyloConvert 0.0.4 - May 2022 by Fernando Marques
```

```
#####
#           THIS PROGRAM CONVERTS FILES FOR PHYLOGENETIC ANALYSES           #
#                                                                                   #
# YOUR OPTIONS ARE:                                                                #
#                                                                                   #
# 1. For GenBank (*.gb) format to FASTA/POY using accession numbers.             #
# 2. For GenBank (*.gb) format to FASTA/POY using terminal names.                 #
# 3. For GenBank (*.gb) format to FASTA/POY using numbers and names.             #
# 4. For clean FASTA files to XREAD format (for TNT).                             #
# 5. For clean FASTA files to NEXUS format (for PAUP).                           #
# 6. Generate accession numbers list from GenBank (*.gb) file.                   #
# 7. Generate taxon name list from GenBank (*.gb) file.                           #
# 8. Generate a file containing translation rules for SED.                         #
# x. Exit program.                                                                #
#                                                                                   #
#####
```

```
Select the option desired:
```

As três primeiras opções manipulam arquivos no formato *GenBank (full)* transformando-os em arquivos FASTA. A diferença entre estas opções reside na informação que será inserida nos terminais. Por exemplo:

```
Select the option desired: 1
```

³Se você quer executar o script que está no diretório você deverá digitar “./phyloconvert”.

Resulta em:

```
>FJ685662
GCCTAATTACACAAATTATTACAGGACTTTTTTT...
>FJ685663
GCTTAATTATACAAATTGCACCAGGACTATTCTT...
>FJ685664
GTCACAGGACTCTTCCTTGCAATACACTATACTG...
...
```

Select the option desired: 2

Resulta em:

```
>Thoropa_miliaris
GCCTAATTACACAAATTATTACAGGACTTTTTTT...
>Cycloramphus_acangatan
GCTTAATTATACAAATTGCACCAGGACTATTCTT...
>Macrogenioglottus_alipioi
GTCACAGGACTCTTCCTTGCAATACACTATACTG...
...
```

Select the option desired: 3

Resulta em:

```
>FJ685662_Thoropa_miliaris
GCCTAATTACACAAATTATTACAGGACTTTTTTT...
>FJ685663_Cycloramphus_acangatan
GCTTAATTATACAAATTGCACCAGGACTATTCTT...
>FJ685664_Macrogenioglottus_alipioi
GTCACAGGACTCTTCCTTGCAATACACTATACTG...
...
```

A escolha destas opções depende de uma série de fatores. Por exemplo, se sua intenção é concatenar diferentes bases de dados, então você deve escolher a opção que considera apenas o nome dos terminais. Neste caso, assume-se que todas as partições (*i.e.*, distintas regiões genômicas ou diferentes fonte de dados) terão o mesmo nome para cada terminal. No entanto há um problema a considerar. Pode existir no seu banco de dados terminais com o mesmo nome. Neste caso, você terá problemas quando for analisar os dados ou mesmo quando for concatenar as partições – isso é evidente nesse conjunto de dados. A inserção do número de tombo não permite a concatenação da base de dados, pois cada número é único para determinado táxon e determinada região genômica. No entanto há como contornar esse problema, como veremos adiante. Seja qual for sua opção, pense bem em sua estratégia de análise, pois você poderá encontrar problemas à medida em que executa inúmeras tarefas em seu procedimento analítico.

Exercício 7.2

Neste exercício você deverá gerar um arquivo no formato FASTA para cada um dos arquivos que você obteve no Exercício 7.1. Estes arquivos deverão ser criados com a opção **1** de *phyloconvert*, ou seja, os terminais deverão conter o número de tombo. A razão pela qual não iremos incluir os nomes é porque isso geraria terminais com o mesmo nome e isso seria um problema quando você for concatenar os dados. Por outro lado, a adoção dos números de tombo faria com que cada terminal, em cada partição, fosse considerado como um terminal distinto – quando, na realidade, você dispõe de 5 regiões genômicas distintas para a grande maioria dos terminais. Portanto esse exercício deverá ser completado por dois procedimentos distintos. No primeiro deles, você irá usar o *phyloconvert* para fazer com que os nomes das sequências recebam o número de tombo do GenBank (opção 1). Na segunda etapa, você deverá substituir os números de tombo pelos nomes dos táxons na primeira coluna do arquivo `dias_et_al_2013_accession_numbers.csv`. Isso será feito com o auxílio do programa *sed* – que foi apresentado no Tutorial 2. No diretório `tutorial_7` há 5 arquivos de substituição (`*_sub.sed`). Após usar o *phyloconvert* você deverá executar o *sed* utilizando a seguinte linha de comando, por exemplo:

```
$ sed -f 12s_sub.sed 12s.fas > 12s_renamed.fas
```

A opção `-f` diz ao *sed* que as instruções de substituição estão no arquivo `12s_sub.fas` que serão aplicadas ao arquivo `12s.fas` e o resultado será redirecionado (`>`) para o arquivo `12s_renamed.fas`.

7.2.2 VERIFICAÇÃO E EDIÇÃO FASTA EM ALIVIEW

Uma vez editado os nomes dos terminais é necessário verificar as sequências nucleotídicas e, via de regra, editar o arquivo. Para isso usaremos o aplicativo chamado [AliView](#) [4]. AliView é uma aplicativo muito útil e intuitivo de usar. Há outras ferramentas disponíveis para esse propósito, tais como [SeaView](#) e [Geneious](#) – entre outros, caso ele não atenda suas necessidades. Para os propósitos de nosso curso, ele será suficiente.

O AliView é iniciado em um terminal. Abra um terminal e execute o comando `aliview`. Após a execução, você deverá obter a janela do programa tal qual na Figura 7.5.

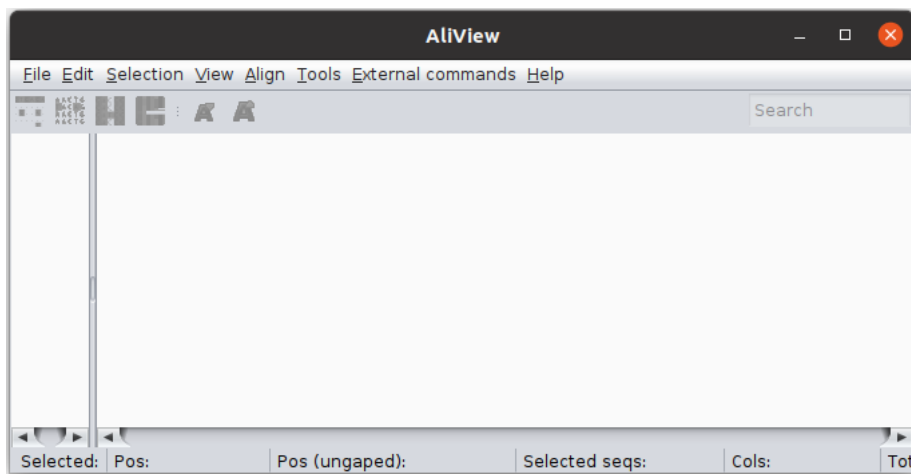


Figura 7.5: Janela inicial do AliView.

Para demonstrar alguns atributos de AliView, eu irei apresentar uma série de comandos no programa utilizando o arquivo `exemplo_aliview_cyt-b.fas` que encontra-se disponível no diretório `tutorial_07`. Esse arquivo foi gerado a partir das sequências de Citocromo B de Dias *et al.* [1] após eu ter criado um arquivo FASTA (usando a opção 1) utilizando o *script* `phyloconvert`, como vocês fizeram no Exercício 7.1.

No AliView, vá em **File > Open File** e abra o arquivo `exemplo_aliview_cyt-b.fas` que está no diretório `tutorial_07`. Uma vez aberto, corra a barra de rolagem horizontal até o final das sequências nucleotídicas como ilustrado na Figura 7.6.

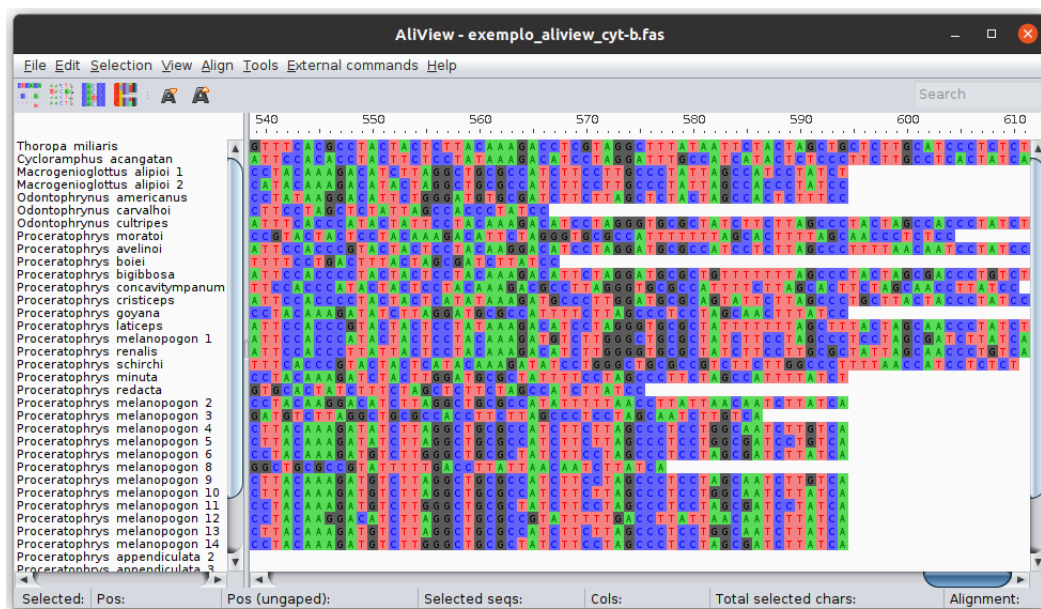


Figura 7.6: Arquivo `exemplo_aliview_cyt-b.fas` aberto no AliView.

Você deverá observar que no arquivo `exemplo_aliview_cyt-b.fas` as sequências nucleotídicas diferem de tamanho (Figura 7.6). Estas diferenças se dão por sequenciamento diferencial das amostras. Adicionalmente, o Citocromo B é um gene codificante, onde

raramente ocorrem inserções e deleções e quando ocorrem tendem a atuar sobre códon – portanto três nucleotídeos. Para que possamos proceder com análises filogenéticas desses dados é necessário alinhar essas sequências. Alinhamento é o processo pelo qual sequências de tamanhos diferentes são igualadas quanto ao tamanho com a inserção de *gaps* (“-”). Nós iremos discutir detalhadamente algoritmos de alinhamento e a relação entre alinhamento e hipóteses filogenéticas no próximo tutorial. No momento, por se tratar de uma região codificadora, a expectativa é que os *gaps* sejam inseridos no começo e no final de cada sequência.

O alinhamento de sequências em AliView requer inicialmente que todas as sequências sejam selecionadas. Para fazer isso clique em **Selection > Select all**. Uma vez selecionadas, clique em **Align > Realign everything** e clique em OK para iniciar o alinhamento. Na configuração padrão de AliView, o programa de alinhamento é o **Muscle** – que será discutido no próximo tutorial. Quando o processo de alinhamento terminar – indicado pela mensagem “Done”, feche a janela Align with Muscle. O resultado final deverá ser semelhante ao que é apresentado na Figura 7.7.

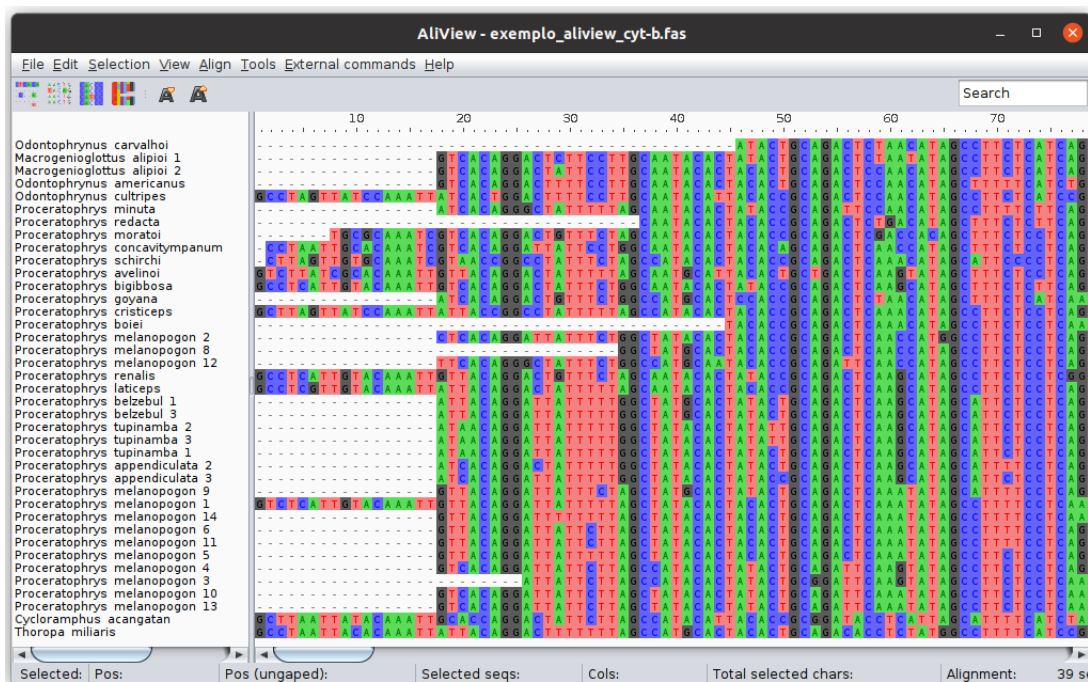


Figura 7.7: Alinhamento via Muscle em AliView das sequências em exemplo_aliview_cyt-b.fas.

Observe que após esse procedimento todas as sequências possuem o mesmo comprimento (612 pares de base [bp]). O Muscle inseriu a maioria dos *gaps* na porção inicial das sequências, chamados *leading gaps*, nenhum *gap* no meio das sequências, e algumas sequências com *gaps* no final, estes chamados de *trailing gaps*. Os *gaps* iniciais e finais resultam do processo de sequenciamento diferencial para cada terminal. Se as sequências são mantidas como estão, estes *gaps* podem ser considerados como um quinto estado de caráter (*i.e.*, A, C, G, T e -) quando não deveriam. Desta forma, é necessário editar as regiões iniciais destas sequências. Isso pode ser feito de duas formas. Uma delas é removendo as regiões iniciais das sequências, a outra forma é

substituindo os *gaps* por “Ns”. Considere que a substituição de *leading* e *trailing gaps* por “N” insere ambiguidade nos dados, pois “N” é considerado como “?” durante análises filogenéticas. Por outro lado, a remoção destas regiões pode resultar na perda de dados que pode ser importante para resolver a topologia. Por exemplo, há 7 sequências em que há 300 pares de base faltantes no final. Isso levou a inserção de muitos *gaps*. Se você optar pela exclusão desta região, você perde a metade dos seus dados aproximadamente. A decisão entre remover ou substituir é arbitrária.

Neste exemplo iremos fazer os dois procedimentos. O primeiro deles é remover os 17 pares de base da região inicial do alinhamento. Para remover esta região no AliView basta você selecionar esta região com o mouse e pressionar Ctrl+Del. Você deverá obter o que é ilustrado na Figura 7.8.



Figura 7.8: Arquivo exemplo_aliview_cyt-b.fas com a remoção dos primeiros 17 pares de base.

O segundo passo será transformar os *gaps* iniciais e finais em “Ns”. Isso pode ser feito de duas formas. A primeira delas é manualmente, mas isso pode ser indesejável quando o número de sequências é relativamente grande, a outra é usando um *script* que faz isso por você. Vamos usar as duas estratégias. Para fazer isso manualmente na região inicial, basta selecionar a região de *gaps* de uma determinada sequência e pressionar a tecla “n”. Você deverá obter o que está ilustrado na Figura 7.9. A segunda forma é utilizar o *script* subedgesgaps.py para executar a mesma tarefa, mas agora na região final das sequências nucleotídicas. Para isso, eu vou salvar este alinhamento parcialmente editado em **File > Save as Fasta** sob o nome de exemplo_aliview_cyt-b_aln_trim.fas e vou fechar o AliView.



Figura 7.9: Inserção manual de “Ns” na região inicial do alinhamento.

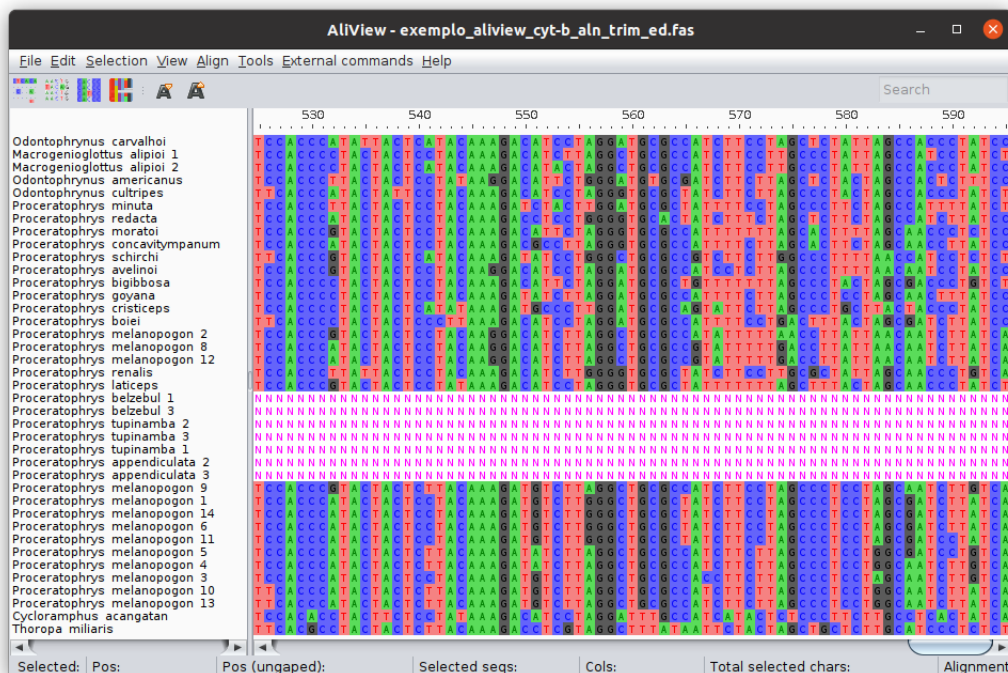


Figura 7.10: Inserção de “Ns” na região final do alinhamento utilizando o *script* subedgesgaps.py.

Agora vamos editar a parte final deste alinhamento utilizando o *script* subedgesgaps.py. Essa operação é relativamente simples. Em um terminal execute o seguinte comando:

```
$ ./subedgesgaps.pl -i exemplo_aliview_cyt-b_aln_trim.fas > exemplo_aliview_cyt-b_aln_trim_ed.fas
```

Se abrirmos o arquivo `exemplo_aliview_cyt-b_aln_trim_ed.fas` no AliView e observarmos o final do alinhamento iremos ver que todos aqueles *gaps* terminais foram preenchidos com “Ns” como ilustrado na Figura 7.10. Essa estratégia poderia ter sido adotada logo no início, após a deleção dos primeiros 17 pares de base. A operação anterior foi simplesmente para demonstrar que o AliView permite a edição das sequências manualmente.

Recomendo que em cada etapa de edição parcial você salve o arquivo editado sob outro nome. Geralmente eu uso a notação “_aln” para arquivos editados, “_trim” para arquivos em que eu removi as regiões iniciais e/ou finais (de *trimmed*) e “_ed” quando eu implemento edições. Isso possibilita que você volte a alguma etapa anterior caso detecte algum erro. Seu objetivo é chegar a uma edição final no qual o arquivo está no formato compatível com o programa de inferência filogenética que você deseja utilizar, incluindo como as regiões de inserção e deleção devam ser tratadas. Uma última ponderação sobre o exemplo acima; em alguns casos, o alinhamento apresentará *gaps* internos. Para genes ribossomais, estes devem ser tratados como quinto estado de caráter. Em regiões codificantes, a existência desses *gaps* não é usual e deve ser acompanhada de *triplets* – três *gaps* via de regra.

Exercício 7.3

Para que você se familiarize com o AliView, você deverá editar e salvar as edições de todos os arquivos que você gerou no Exercício 7.1.

7.3 Configuração de matrizes de dados moleculares

Transformar um arquivo FASTA em um arquivo passível de análises filogenéticas é relativamente simples – embora em alguns casos isso seja desnecessário como veremos ao longo desses tutoriais. Como vocês estão familiarizados com o formato de arquivos de entrada para TNT (veja Seção 4.3 do Tutorial 4), iremos explorar como transformar um arquivo FASTA em formato `xread`.

Considere as seguintes sequências em formato FASTA:

```
>taxon_1
GCCTAATTACACAAATTATT
>taxon_2
GCTTAATTATACAAATTGCA
>taxon_3
GTCACAGGACTCTTCCTTGC
>taxon_4
GTCACAGGACTCATACACTA
```

Neste exemplo temos 4 terminais com sequências de 20 pares de base. A edição deste arquivo para que seja passível de análise filogenética em TNT é simples e pode ser feita manualmente. Veja como seria:


```

nstates dna;

xread

20 4

taxon_1      GCCTAATTACACAAATTATT
taxon_2      GCTTAATTATACAAATTGCA
taxon_3      GTCACAGGACTCTTCCTTGC
taxon_4      GTCACAGGACTCATACTACTA;

proc/;

```

ou ainda:

```

xread

20 4

taxon_1      21130033010100033033
taxon_2      21330033030100033210
taxon_3      23101022013133113321
taxon_4      23101022013103010130;

proc/;

```

No primeiro caso, as sequências estão inseridas como vieram do arquivo FASTA e isso requer que a primeira linha do arquivo contenha "nstates dna;". No segundo caso, o arquivo é idêntico aos que vocês já viram anteriormente. Neste caso, as bases A, C, G e T foram substituídas por 0, 1, 2 e 3, respectivamente. Para arquivos simples como esse, a edição manual é possível, mas pode se tornar inviável à medida em que o número de terminais e/ou de pares de base aumentam. Neste caso, o melhor a fazer é usar *scripts* ou aplicativos concebidos para esse propósito.

Como vocês viram na seção 7.2.1, o *script phyloconvert* possui duas opções que possibilitam a configuração de arquivos para TNT e PAUP a partir de arquivos FASTA (opções 4 e 5, veja acima). Acredito que não terá dificuldades em usar *phyloconvert* para transformar arquivos no formato FASTA em arquivos para TNT e PAUP. Portanto, vamos explorar outra ferramenta.

7.3.1 SEQUENCEMATRIX

[SequenceMatrix](#) [5] é um aplicativo relativamente simples, mas muito útil para transformar arquivos FASTA em outros formatos. O aplicativo também permite concatenar bancos de dados distintos de forma muito amigável. [SequenceMatrix](#) tem uma série de funções que não serão exploradas em detalhe neste tutorial, consulte a página do aplicativo e o trabalho de Vaidya *et al.* [5] para maiores informações. Vejamos como ele funciona considerando as três partições ilustradas na Figura 7.11.

Partição 1	Partição 2	Partição 3
>taxon_1	>taxon_1	xread
AAAACCCC	GGGGTTTT	8 4
>taxon_2	>taxon_2	taxon_1
AAA-CCCC	GGG-TTTT	00000010
>taxon_3	>taxon_3	taxon_2
AAAA-CCC	GGGG-TTT	11000001
>taxon_4	>taxon_4	taxon_3
AAAACCC-	GGGGTTT-	11110000
		taxon_4
		11111100
		taxon_5
		11111110;
		proc/;

Figura 7.11: Partições de dados utilizadas no exemplo de SequenceMatrix. As duas primeiras partições estão em formato FASTA. A terceira partição é uma arquivo para dados morfológicos no formato xread de TNT. Observe que a terceira partição contém um terminal a mais que as demais.

A execução do SequenceMatrix é feita do terminal da seguinte forma:

```
$ sequencematrix
```

Ao executá-lo você deverá obter a seguinte janela:

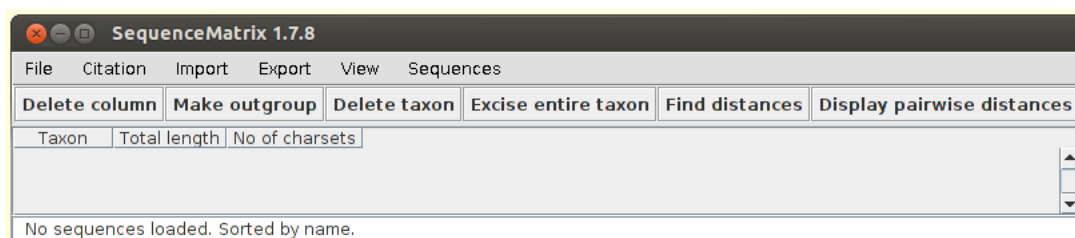


Figura 7.12: Janela de abertura de SequenceMatrix.

Para importar partições de dados basta selecionar `Import/Add sequence` no menu principal. No exemplo abaixo (Figura 7.13), eu adicionei a primeira partição que estava em um arquivo chamado `sequencematrix_1.fas`.

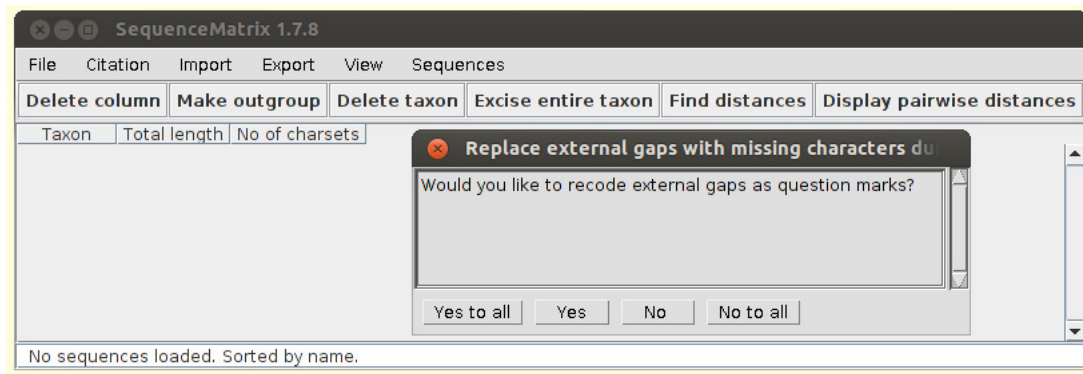


Figura 7.13: Janela de importação de SequenceMatrix.

SequenceMatrix irá lhe perguntar se você deseja substituir os *gaps* por ?. Você deverá sempre selecionar *No to all*, a não ser que você realmente queira tratar eventos de INDELs como *missing data*. Em breve iremos discutir esse aspecto da análise filogenética da dados moleculares.

Em alguns casos, o programa poderá lhe perguntar se você deseja usar o nome do táxon ou da sequência (Figura 7.14). Você deverá optar pelo nome da sequência, caso contrário, o programa gerará – neste caso em particular – terminais com o mesmo nome. Isso lhe traria proplemas mais adiante. Minha sugestão é que você configure os nomes dos terminais (ou sequências) da forma mais simples e informativa possível nesse estágio de sua análise – antes de concatená-los. Lembre-se que você poderá substituir esses nomes no final, quando for gerar figuras e ou avaliar resultados. Quanto mais simples os nomes são, mais fácil se torna a tarefa de gerenciar arquivos e evitar erros, principalmente se você está lidando com várias bases de dados para os mesmos terminais

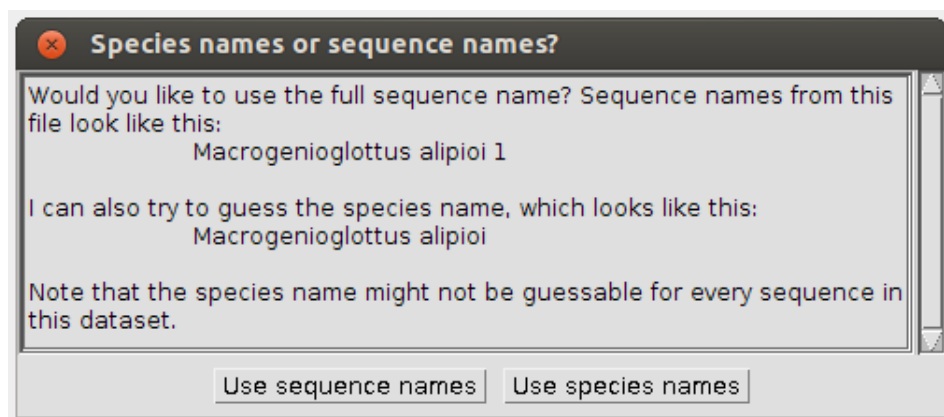
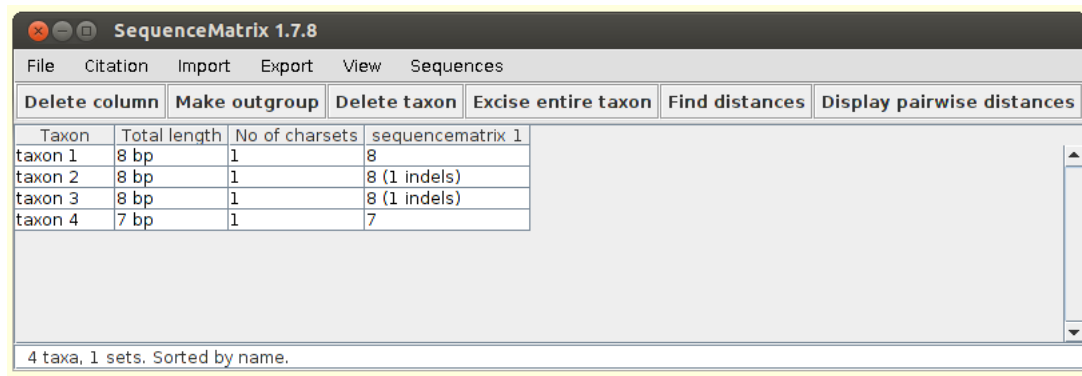


Figura 7.14: Janela de de seleção de nomes em SequenceMatrix.

Ao importar as sequências, você deverá obter resultado semelhante ao ilustrado na Figura 7.15. SequenceMatrix informa alguns detalhes sobre os dados importados, tais como número de terminais, tamanho das sequências e presença de INDELs.



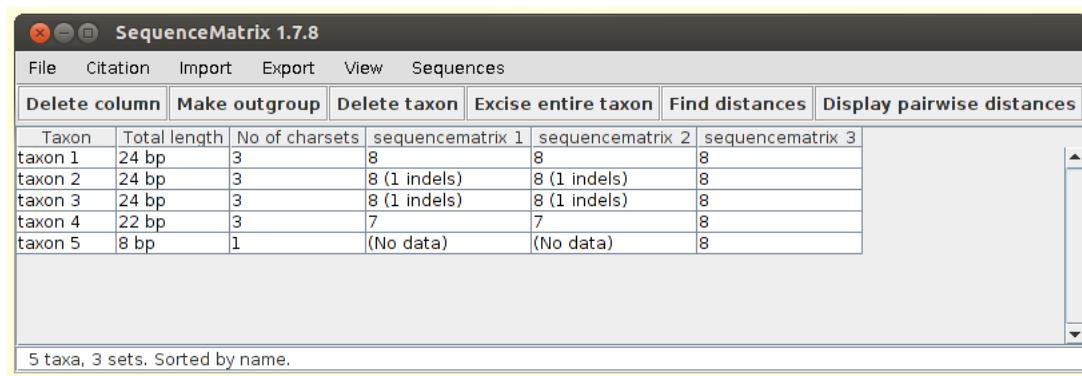
The screenshot shows the SequenceMatrix 1.7.8 interface. The menu bar includes File, Citation, Import, Export, View, and Sequences. Below the menu bar are buttons: Delete column, Make outgroup, Delete taxon, Excise entire taxon, Find distances, and Display pairwise distances. The main table displays the following data:

Taxon	Total length	No of charsets	sequencematrix 1
taxon 1	8 bp	1	8
taxon 2	8 bp	1	8 (1 indels)
taxon 3	8 bp	1	8 (1 indels)
taxon 4	7 bp	1	7

At the bottom, it states: 4 taxa, 1 sets. Sorted by name.

Figura 7.15: Sequências importadas em SequenceMatrix.

Para importar todas as partições da Figura 7.11 basta repetir a tarefa. SequenceMatrix importa matrizes de TNT da mesma forma que importa sequências em formato FASTA. A importação de todas as partições é ilustrada na figura seguinte (Figura 7.16):



The screenshot shows the SequenceMatrix 1.7.8 interface with 5 taxa imported across 3 matrices. The main table displays the following data:

Taxon	Total length	No of charsets	sequencematrix 1	sequencematrix 2	sequencematrix 3
taxon 1	24 bp	3	8	8	8
taxon 2	24 bp	3	8 (1 indels)	8 (1 indels)	8
taxon 3	24 bp	3	8 (1 indels)	8 (1 indels)	8
taxon 4	22 bp	3	7	7	8
taxon 5	8 bp	1	(No data)	(No data)	8

At the bottom, it states: 5 taxa, 3 sets. Sorted by name.

Figura 7.16: Partições importadas em SequenceMatrix.

Para exportar esses dados, basta selecionar uma das opções disponíveis em Export do menu principal. No exemplo abaixo eu selecionei a opção Export/Export sequences as TNT. O arquivo de exportação terá a seguinte configuração:

```
nstates dna;
xread
'Exported by SequenceMatrix 1.7.8 on Tue Apr 22 09:39:44 BRT 2014.'
24 5
taxon_1 AAAACCCCGGGTTTT00000001
taxon_2 AAA-CCCGGG-TTTT11000000
taxon_3 AAAA-CCCGGG-TTT11110000
taxon_4 AAAACCC-GGGGTTT-11111100
taxon_5 ?????????????11111110;

xgroup
=0 (sequencematrix_1) 0 1 2 3 4 5 6 7
=1 (sequencematrix_2) 8 9 10 11 12 13 14 15
=2 (sequencematrix_3) 16 17 18 19 20 21 22 23
;
```

```

agroup
=0 (CHARSETS_ATLEAST_3) 0 1 2 3
=1 (CHARSETS_ATLEAST_2) 0 1 2 3
=2 (TAXA_HAVING_sequencematrix_1) 0 1 2 3
=3 (TAXA_HAVING_sequencematrix_2) 0 1 2 3
=4 (TAXA_HAVING_sequencematrix_3) 0 1 2 3 4
;

```

No exemplo acima, a matrix de dados contempla todas as partições e insere pontos de interrogação para os caracteres das duas primeira partições nas quais o `taxon_5` não está representado. Há dois blocos adicionais neste arquivo. O primeiro deles é o `xgroup` que define grupos de caracteres. O segundo é o `agroup` que define grupos de táxons em TNT. Esses blocos são utilizados em TNT para manipular grupos como um todo – para maiores detalhes veja documentação de TNT.

Exercício 7.4

Neste exercício você deverá transformar os arquivos que gerou no Exercício 7.3 em arquivos no formato de TNT. As tarefas específicas a serem realizadas são as seguintes:

- i. Gerar arquivos no formato TNT para cada partição de Dias *et al.* [1].
- ii. Gerar um arquivo no formato TNT no qual todas as partições estão representadas.
- iii. Analisar estes arquivos em TNT e fazer um resumo abaixo no qual você compara os resultados que você obteve com a topologia apresentada na Figura 11 (pg. 294) de Dias *et al.* [1].

7.4 Referências

1. Dias, P. H. S.; Amaro, R. C.; de Carvalho-e Silva, A. M.P. T. & Rodrigues, M. T. 2013. Two new species of *Proceratophrys* Miranda-Ribeiro, 1920 (Anura; Odontophrynidae) from the Atlantic forest, with taxonomic remarks on the genus. *Zootaxa* **3682**: 277–304.
2. Goloboff, P.; Farris, J. S. & Nixon, K. 2008. TNT a free program for phylogenetic analysis. *Cladistics* **24**: 1–14.
3. Swofford, D. 2003–2016. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods, Version 4.0a131). Sunderland, Massachusetts: Sinauer Associates, 2003–2016.
4. Larsson, A. 2014. AliView: a fast and lightweight alignment viewer and editor for large data sets. *Bioinformatics* **30**(22): 3276–3278.
5. Vaidya, G.; Lohman, D. J. & Meier, R. 2010. SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information. *Cladistics* **27**: 171–180.