

RAE-CEA-9521
RELATÓRIO DE ANÁLISE ESTATÍSTICA
SOBRE O PROJETO: AVALIAÇÃO ATRAVÉS DE TÉCNICAS
DE EPIDEMIOLOGIA MOLECULAR, DA BIODIVERSIDADE DE
***E.Coli* OBTIDAS DE AMOSTRAS DE ÁGUA PROVENIENTES**
DE AMBIENTES NÃO AFETADOS POR ATIVIDADE
ANTRÓPICA, FEZES DE ANIMAIS DOMÉSTICOS E
SELVAGENS, ESGOTO E DO HOMEM

Dalton Francisco de Andrade
Marcelo Leme de Arruda
Paula Andreucci da Silva Prado

- São Paulo, dezembro de 1995 -

CENTRO DE ESTATÍSTICA APLICADA**RELATÓRIO DE ANÁLISE ESTATÍSTICA - NÚMERO 9521
- CÓDIGO 95P21**

TÍTULO: Relatório de Análise Estatística sobre o projeto "Avaliação através de técnicas de epidemiologia molecular, da biodiversidade de *E.Coli* obtidas de amostras de água provenientes de ambientes não afetados por atividade antrópica, fezes de animais domésticos e selvagens, esgoto e do homem "

PESQUISADORA: Irma N. G. Rivera

INSTITUIÇÃO: Instituto de Ciências Biomédicas - Laboratório de Microbiologia Ambiental - USP

FINALIDADE: Publicação

RESPONSÁVEIS PELA ANÁLISE: Dalton Francisco de Andrade, Marcelo Leme de Arruda e Paula Andreucci da Silva Prado

REFERÊNCIA DESTE TRABALHO

ANDRADE, D.F., ARRUDA, M. L. e PRADO, P. A. S. **Relatório de Análise Estatística sobre o projeto "Avaliação através de técnicas de epidemiologia molecular, da biodiversidade de *E.Coli* obtidas de amostras de água provenientes de ambientes não afetados por atividade antrópica, fezes de animais domésticos e selvagens, esgoto e do homem "**. São Paulo, IME-USP, 1995 (RAE-CEA, 9521)

FICHA TÉCNICA**BIBLIOGRAFIA:**

BUSSAB, W.O., MIAZAKI, É. S. e ANDRADE, D. F. (1990). **Introdução à Análise de Agrupamentos**. São Paulo: Associação Brasileira de Estatística

GRIMONT, R. e GRIMONT, P. (1989). Ribotyping of enterobacteriaceae. **Journal of Applied Environmental Microbiology**, 57, 475-478

KRAMER, J. (1993). **Handbook of methods in aquatic microbiology ecology**. Florida, USA: Lewis Publishers, CRC. Press, Inc.

MURTEIRA, B.J. F e BLACK, G., H., J. (1983). **Estatística descritiva**. Lisboa: McGraw-Hill de Portugal

SPSS/PC+ for IBMPC/XT/AT version 2.0 (1988). USA: Microsoft Corporation

WARD, J. H. (1963). Hierarchical grouping to optimize an objective function, **Journal of the American Statistical Association**, 58, 236-244

YESTER, M. (1990). **Turbo Pascal: quick reference**. Rio de Janeiro: Campus

PROGRAMAS COMPUTACIONAIS UTILIZADOS:

Cluster versão 1990

SPSS/PC+ v4.0

Turbo Pascal versão 7.0

Word for Windows 6.0

TÉCNICAS ESTATÍSTICAS UTILIZADAS: (Entre parênteses a classificação ISI)

Análise de Agrupamentos (06:120)

Análise Descritiva (03:010)

ÁREA DE APLICAÇÃO:

Epidemiologia (14:040)

ÍNDICE

	Página
Resumo.....	5
1 Introdução	6
2 Descrição do estudo e das variáveis	6
3 Análise descritiva.....	7
3.1 Dados obtidos pelo método RAPD	7
3.2 Dados obtidos pelo método da Ribotipagem.....	12
4 Análise dos Dados	16
4.1 Procedimento Utilizado.....	16
4.2 O programa MATRIZES.EXE	18
4.3 Dendrogramas	20
4.4 Observações	20
4.5 O programa CONVERTE.EXE.....	21
4.6 Medida de distância usando a variabilidade dos dados.....	23
5 Conclusões e Sugestões	24
Apêndice A Tabelas.....	26
A.1 Dados obtidos pela técnica RAPD	27
A.2 Dados obtidos pela técnica de Ribotipagem.....	29
Apêndice B Dendrogramas	32
B.1 Dados obtidos para a subpopulação de animais pela técnica RAPD	33
B.1.1 Distância Euclideana	33
B.1.2 Distância de Jaccard.....	36
B.2 Dados obtidos para a subpopulação de animais pela técnica de Ribotipagem	39
B.2.1 Distância Euclideana	39
B.2.2 Distância de Jaccard.....	41
Apêndice C Uma alternativa para as medidas de distância	43
C.1 Apresentação da Medida	44
C.2 Exemplos da Medida.....	45

Resumo

Diversos exemplares da bactéria *Escherichia Coli* (*E.Coli*), utilizada como indicadora de qualidade da água, tiveram seu perfil genômico retratado através de duas técnicas: RAPD (mais barata) e Ribotipagem (mais precisa). Esses perfis deveriam ser comparados, via análise de agrupamentos (Bussab et al. (1990)) com o objetivo de identificar se as bactérias procedentes de ambientes diferentes (águas puras, animais e humanos) são ou não iguais a nível molecular.

Entretanto, o programa computacional utilizado pela pesquisadora não consegue processar toda a quantidade de dados coletados. Outro fato de interesse é que a medida utilizada por esse programa apresenta características que podem torná-la inadequada a este problema.

Foram construídos dendrogramas para todos os grupos de ambientes, através de três métodos diferentes de agrupamento e duas medidas de distância ou similaridade distintas.

Foram sugeridas medidas de distância ou similaridade alternativas àquela adotada pelo programa utilizado pela pesquisadora. Uma dessas medidas, já existente na literatura e nos pacotes estatísticos, foi utilizada na construção dos dendrogramas acima citados. A outra, introduzida neste trabalho, ainda é passível de discussão e carente de maiores análises quanto a propriedades e aplicações.

Foi, ainda, desenvolvido um programa que produz as matrizes envolvidas neste trabalho. Este programa soluciona, através das matrizes de distância ou similaridade, a eventual ausência de uma medida em algum pacote computacional. Este programa também produz a matriz de distância calculada a partir da medida alternativa introduzida neste trabalho.

1 - Introdução

A bactéria *Escherichia Coli* (*E. Coli*) é utilizada em saneamento básico como indicadora de qualidade de água e de riscos à saúde humana: quanto menor for sua concentração, melhor é a qualidade da água. Nos últimos anos, alguns trabalhos da área têm mostrado que essa bactéria também é encontrada naturalmente em águas supostamente não poluídas e não apenas nas fezes humanas e de animais. Assim, ela perderia seu caráter de indicadora de qualidade. O objetivo deste estudo é verificar se as bactérias *E. Coli* encontradas naturalmente no ambiente são, a nível molecular, diferentes das de origem humana ou animal. Se forem diferentes, pode-se continuar utilizando a bactéria *E. Coli* como indicadora de qualidade de águas; se forem iguais, ela perde essa sua característica.

2 - Descrição do Estudo e das Variáveis

Foram coletadas para o estudo, amostras (cepas) da bactéria *E. Coli* provenientes de três fontes (subpopulações): águas puras (de reservas florestais, não afetadas pela atividade antrópica, isto é atividade humana), humanos (enfermos e sadios) e animais (de áreas florestais e de zoológicos). Em seguida, essas cepas foram caracterizadas com relação à estrutura de seu DNA através das técnicas de RAPD e Ribotipagem. Foram utilizadas essas duas técnicas porque, embora a técnica de Ribotipagem produza resultados mais precisos, ela é mais demorada e mais cara do que a RAPD. Assim, foram analisadas 175 cepas pela técnica RAPD e apenas 51 pela Ribotipagem.

A técnica RAPD consiste em identificar as ocorrências de uma sequência específica de pares de bases no DNA. Como não é possível detectar visualmente a posição das ocorrências dessa sequência na molécula, utiliza-se um processo de eletroforese. A partir desse processo, obtém-se o perfil genômico da cepa, onde as ocorrências da sequência de pares de bases de interesse são identificáveis visualmente através de uma faixa com segmentos luminosos. A posição dessas

sequências na molécula é medida através da distância (em milímetros) de um extremo definido da faixa aos segmentos luminosos. Finalmente, essas distâncias em milímetros são convertidas, através do gráfico de uma função algébrica utilizada pelos cientistas da área, nos valores (em kilobases) das posições da sequência de pares de bases na molécula de DNA. A técnica de Ribotipagem também consiste em identificar as ocorrências de uma sequência específica de pares de bases no DNA. Nesse processo, como não é possível utilizar a molécula inteira de DNA, utiliza-se um processo enzimático de clivagem (corte) para dividi-la em diversos fragmentos. Esses fragmentos são, então, submetidos a um processo de eletroforese em gel. Nesse processo, aqueles fragmentos de menor peso molecular apresentam maior facilidade de movimentação no gel, afastando-se dos demais, de maior peso molecular e que têm maior dificuldade de se movimentar. Após isso, os fragmentos são transferidos para uma membrana de nylon (processo Southern Blott), onde sofrem uma hibridização por uma sonda de RNA. Essa sonda identifica e evidencia sequências de bases homólogas à sua própria. A partir desse processo de hibridização, obtém-se o perfil genômico dos fragmentos, onde as ocorrências da sequência de pares de bases identificadas e evidenciadas pela sonda são distinguíveis visualmente através de uma faixa com segmentos luminosos. Por fim, da mesma forma que no método RAPD, a posição dessas sequências na molécula é medida através da distância (em milímetros) de um extremo definido da faixa aos segmentos luminosos e convertida graficamente em kilobases.

Esses valores em kilobase são denominados bandas. Além disso, o número de bandas detectadas em cada bactéria também será estudado.

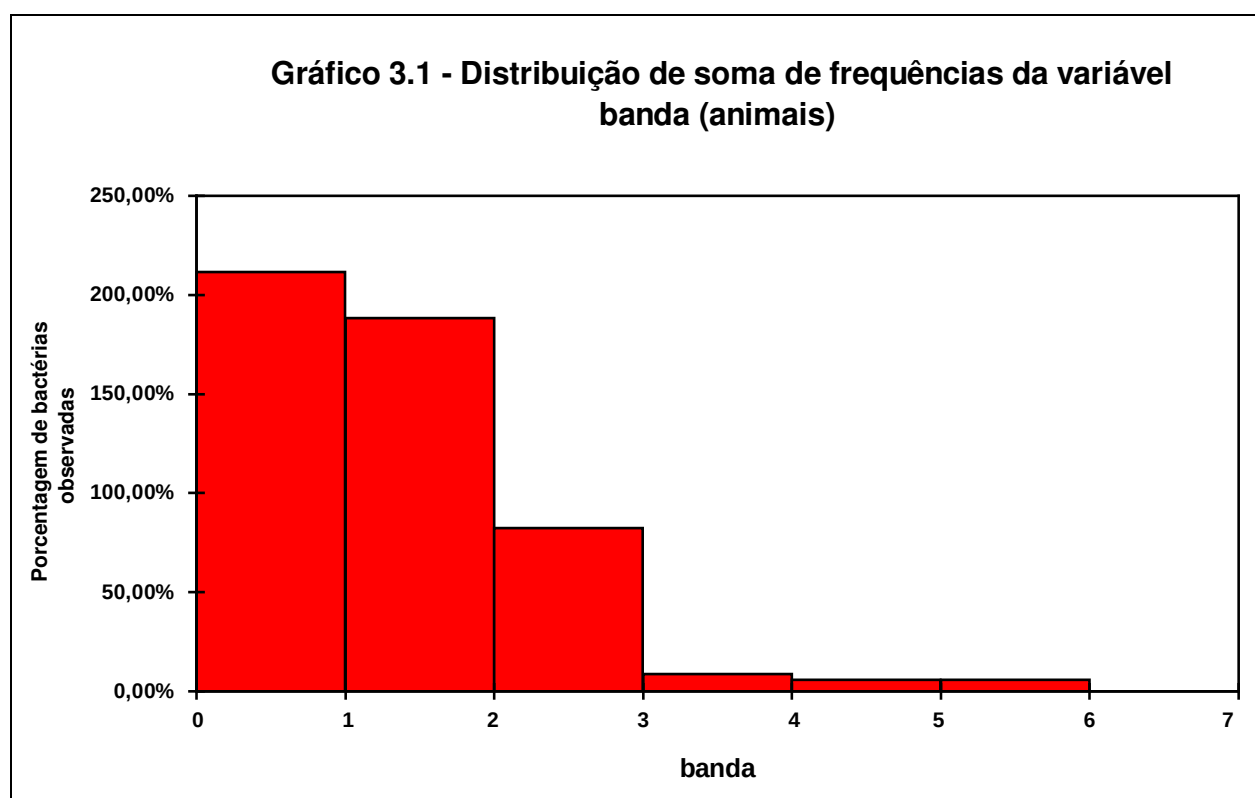
3 - Análise Descritiva

3.1 - Dados obtidos pelo método RAPD

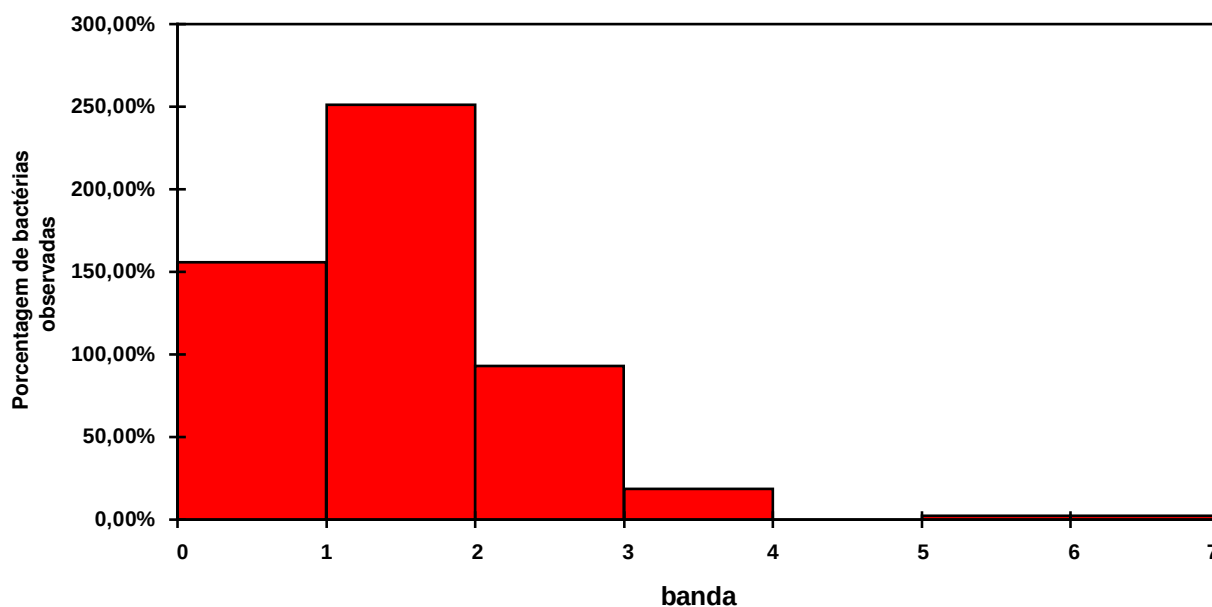
Para cada subpopulação (águas, animais ou humanos), foram produzidos tabelas e gráficos de frequência das variáveis bandas e número de bandas. Pode-se

ver, na Tabela A.1, do apêndice A e nos Gráficos 3.1 a 3.3, que, para as três subpopulações, há uma grande concentração de dados nos valores inferiores, com picos em torno da banda 1,0. Essa concentração, bastante assimétrica (de acordo com a Tabela 3.1), parece se estender até próximo da banda 3,0, a partir de onde há um escasseamento de bactérias que apresentem estes valores. A forma dessa grande concentração inicial varia entre as subpopulações.

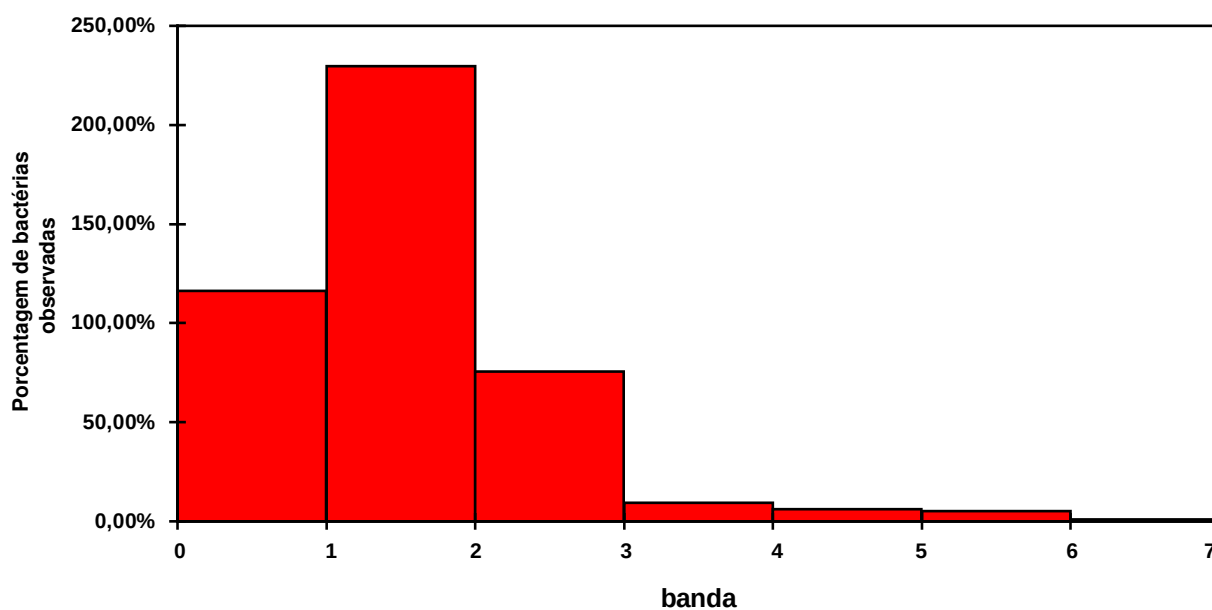
Para as bactérias coletadas em animais (Gráfico 3.1), essa grande concentração se mostra assimétrica com a frequência da faixa 0,1 a 1,0 maior que a da faixa 1,1 a 2,0, enquanto para as bactérias oriundas em humanos (Gráfico 3.2) e águas (Gráficos 3.3), a tendência assimétrica é mais leve, sendo a faixa 1,1 a 2,0 a de maior frequência.



**Gráfico 3.2 - Distribuição de soma de frequências da variável
banda (humanos)**



**Gráfico 3.3 - Distribuição de soma de frequências da variável
banda (águas)**

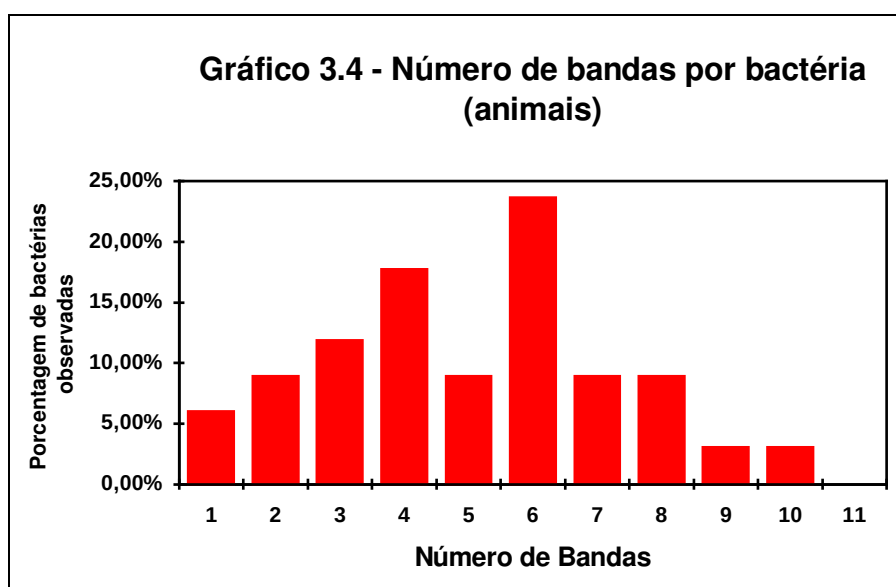


Pode-se notar, ainda, que foram coletadas mais bactérias das águas (98) do que das outras fontes (34 bactérias de animais e 43 de humanos), o que pode ter influenciado, para essa subpopulação, o aparecimento de uma “cauda” até a faixa 5 a 6 e do valor extremo de 7,0.

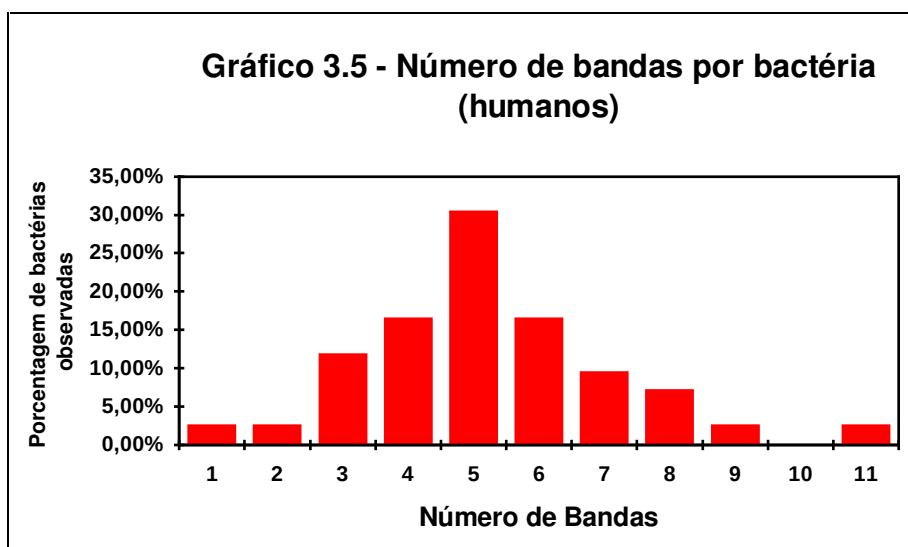
Tabela 3.1 - Coeficientes de assimetria de Bowley (Murteira e Black (1983))

Subpopulação	Bandas	Número de Bandas
Águas	0,250	0,000
Animais	0,333	-0,333
Humanos	0,333	0,000

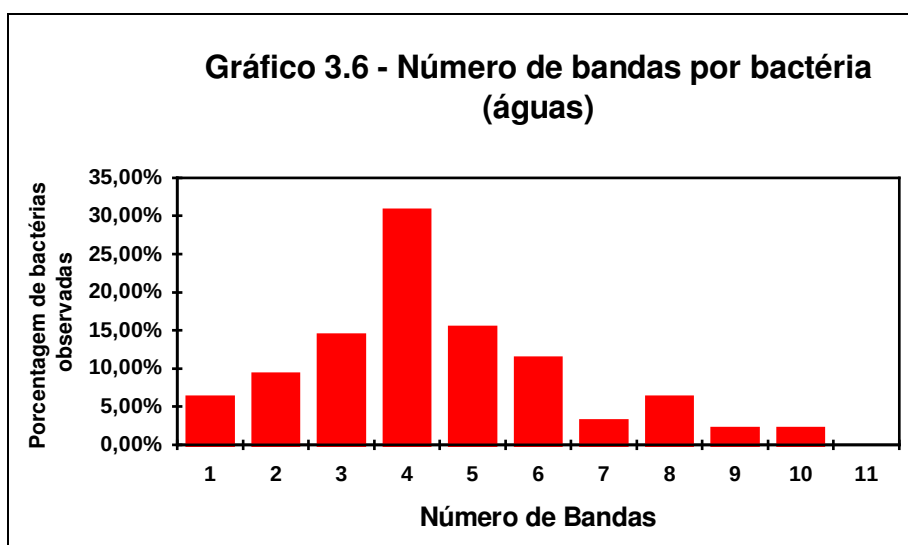
De acordo com os Gráficos 3.4 a 3.6 e Tabelas A.2 e A.3, pode-se notar que, para a maioria das bactérias, há uma concentração ligeiramente assimétrica (Tabela 3.1) de valores em torno da média observada de cada subpopulação. A tabela 3.1 apresenta os coeficientes de assimetria de Bowley, que apresentam o valor 0, quando a distribuição dos dados é simétrica. Para distribuições assimétricas, o coeficiente apresenta valores negativos (assimetria à esquerda) ou positivos (assimetria à direita).



Para as bactérias coletadas de animais (Gráfico 3.4), cujo valor médio observado foi 5,03 bandas, há picos de observação nos valores 4 e 6, com uma diminuição da frequência no valor 5.



Para as bactérias obtidas de humanos (Gráfico 3.5), cuja média observada foi de 5,23 bandas, há um comportamento aparentemente simétrico dos valores, que se concentram entre 3 e 8 bandas, com pico no valor 5.



Por fim, as bactérias encontradas em águas (Gráfico 3.6), que teve média do número de bandas igual a 4,43, há uma concentração assimétrica de valores na região de 2 a 6 bandas, com pico no valor 4.

Pode-se notar também que, segundo a Tabela A.3, as variâncias amostrais das três subpopulações são próximas entre si.

3.2 - Dados obtidos pelo método da Ribotipagem

Para cada subpopulação (águas, animais ou humanos), foram produzidos tabelas e gráficos de frequência das variáveis bandas e número de bandas. Pode-se ver, na Tabela A.4, do apêndice A, e nos Gráficos 3.7 a 3.9, que, para as três subpopulações, há uma grande concentração de dados nos valores inferiores, com picos em torno das bandas 4,0 a 5,0. Essa concentração assimétrica (Tabela 3.2) parece se estender desde próximo da banda 3, até as proximidades da faixa 8,1 a 9,0, fora de onde há um escasseamento de bactérias que apresentem estes valores. A forma e os limites dessa grande concentração variam entre as subpopulações.

Tabela 3.2 - Coeficientes de Assimetria de Bowley (Murteira e Black (1983))

Subpopulação	Bandas	Número de Bandas
Águas	0,179	0,000
Animais	0,280	-0,333
Humanos	0,533	-0,600

Para as bactérias oriundas de animais (Gráfico 3.7), a grande concentração parece abranger, assimetricamente, as faixas compreendidas entre as bandas 3,0 a 9,9. Há um grande pico em torno dos valores 4,5 a 5,0 e picos menores nas faixas 9,1 a 10,0 e 11,1 a 12,0.

Para as bactérias coletadas em humanos (Gráfico 3.8), a grande concentração parece abranger, decrescentemente, as faixas compreendidas entre as bandas 3,0 e 8,0, com o pico principal situado entre as faixas 3,1 a 4,0 e 4,1 a 5,0.

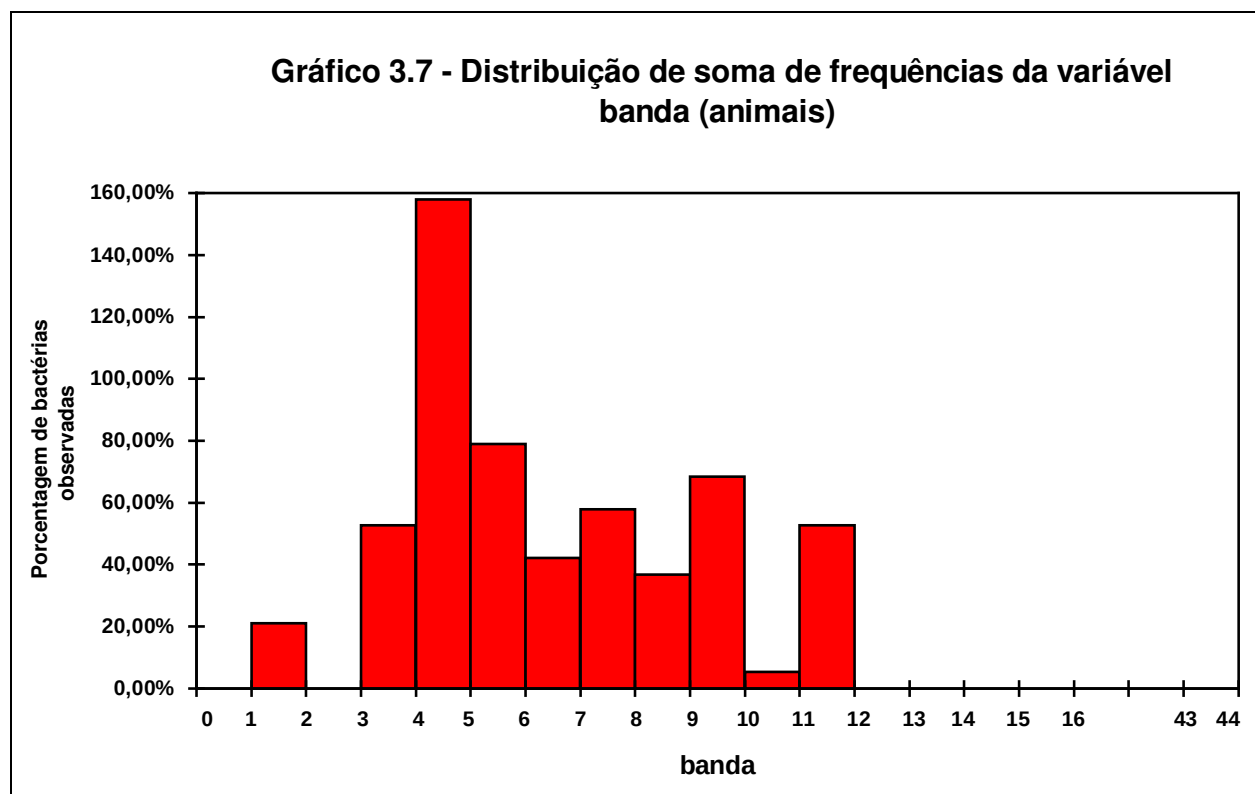


Gráfico 3.8 - Distribuição de soma de frequências da variável banda (humanos)

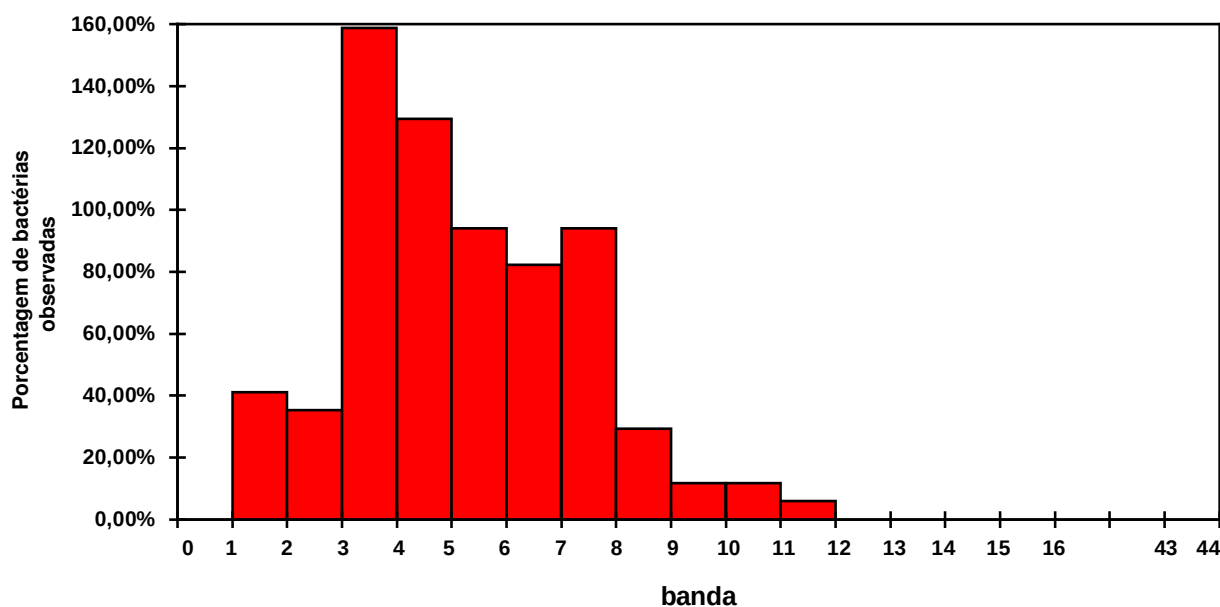
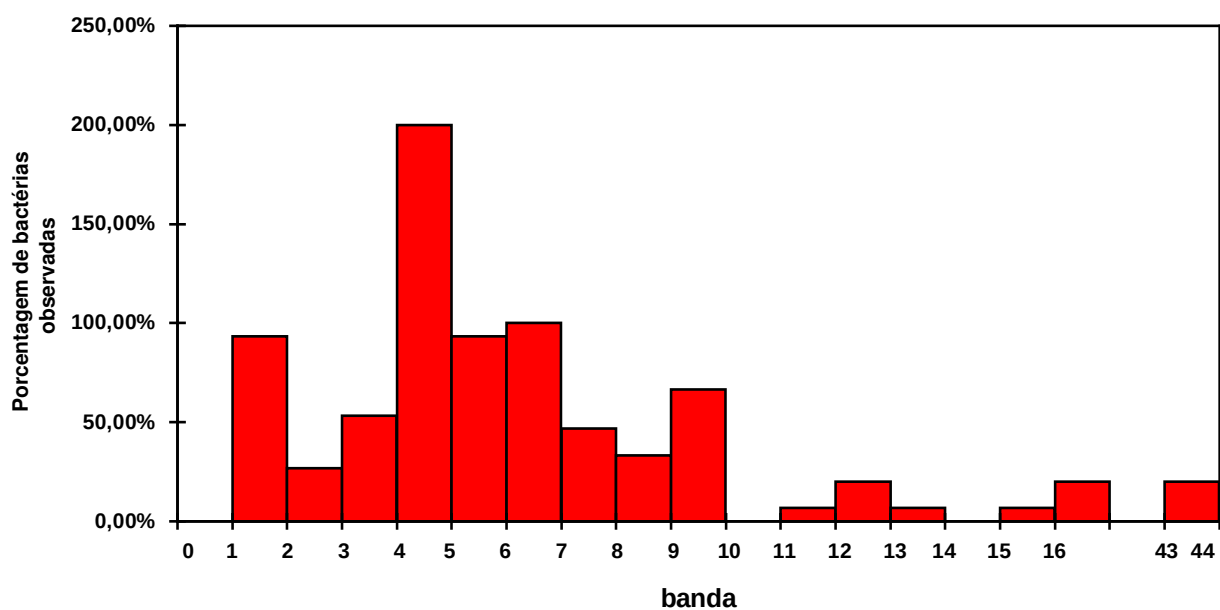


Gráfico 3.9 - Distribuição de soma de frequências da variável banda (águas)

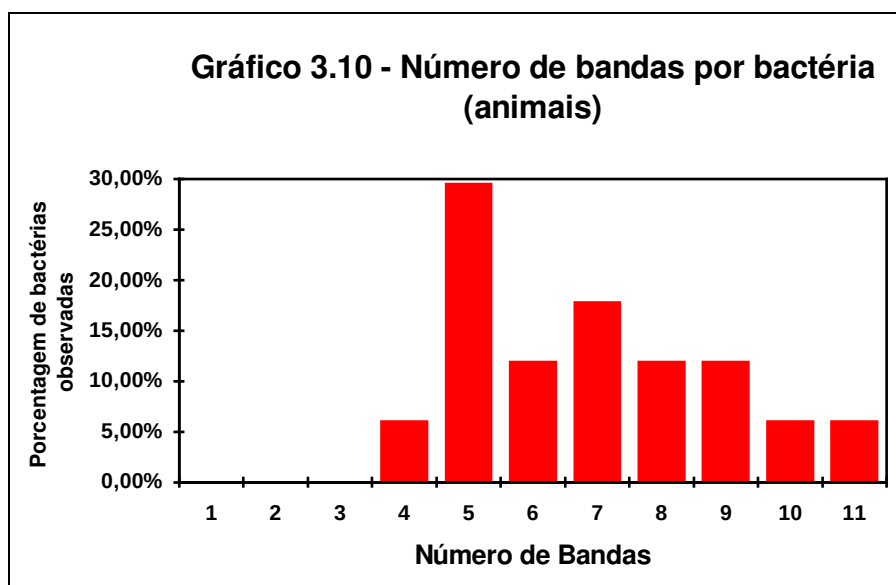


Por fim, para as bactérias obtidas de águas (Gráfico 3.9), a grande concentração parece abranger, assimétrica e irregularmente, as faixas compreendidas entre as bandas 3,0 a 8,0, com o pico principal situado entre as faixas 4,1 a 5,0 e 5,1 a 6,0.

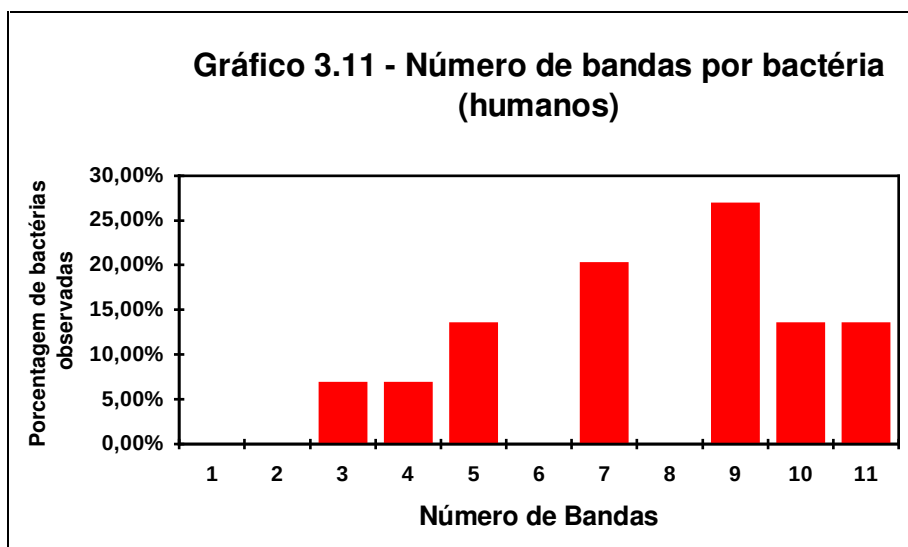
Pode-se notar, ainda, que nas bactérias das águas, houve ocorrências de bandas com o valor 43,8.

De acordo com os Gráficos 3.10 a 3.12 e Tabelas A.5 e A.6, pode-se notar que, para a maioria das bactérias, há uma concentração assimétrica (Tabela 3.2) de valores em torno da média observada de cada subpopulação. A exceção foi a subpopulação das bactérias observadas em humanos, que teve duas regiões de concentração, com a média se situando entre elas.

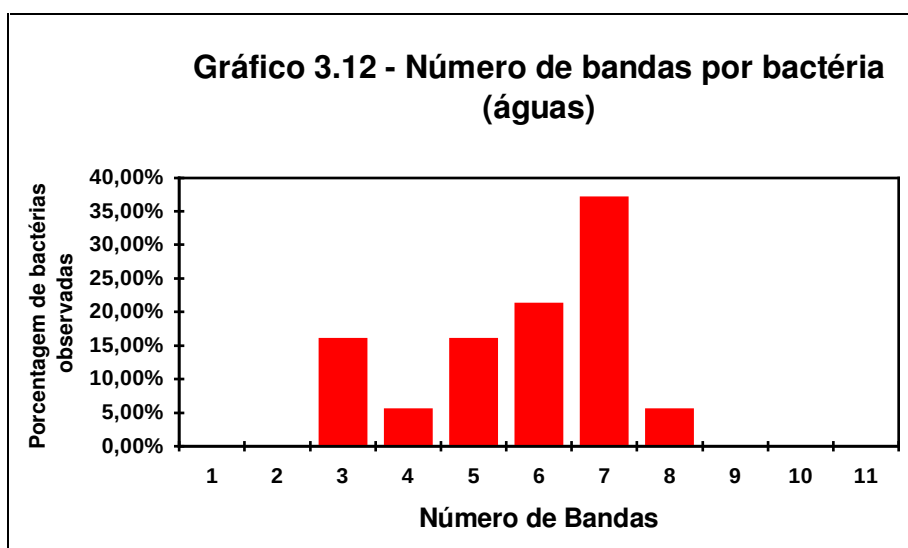
Para as bactérias coletadas de animais (Gráfico 3.10), cujo valor médio observado foi 6,88 bandas, há picos de observação nos valores 5 e 7, com uma diminuição da frequência no valor 6.



Para as bactérias obtidas de humanos (Gráfico 3.11), cuja média observada foi de 7,73 bandas, há um comportamento muito irregular e assimétrico dos valores, que se concentram entre 3 e 5 e entre 9 e 11 bandas, com picos nos valores 5, 7 e 9.



Por fim, as bactérias encontradas em águas (Gráfico 3.6), cuja média observada foi 5,74 bandas, há uma concentração assimétrica de valores na região de 5 a 7 bandas, com pico no valor 7.



Pode-se notar também que, segundo a Tabela A.6, as variâncias amostrais das três subpopulações são próximas entre si.

4. Análise dos Dados

4.1 Procedimento Utilizado

A análise de interesse para a pesquisadora é a análise de agrupamentos (Bussab et al. (1990)), com especial ênfase nos dendrogramas obtidos para cada um dos sete grupos de ambientes de estudo (águas, animais, humanos, águas+animais, águas+humanos, animais+humanos e águas+animais+humanos). Para tanto, é preciso definir uma medida de distância ou de similaridade entre cepas.

O procedimento adotado para medir as diferenças entre as moléculas de DNA das cepas será o seguinte: agrupar todas as bandas observadas de todos os indivíduos para cada ambiente em que os dados foram coletados, em um único vetor com todas as bandas ordenadas e não repetidas. A cada indivíduo será atribuído um novo vetor de dimensão igual à do vetor descrito acima. Os vetores dos indivíduos serão compostos apenas de uns e zeros. O valor *um* será atribuído quando a banda tiver sido observada e o valor *zero* quando ela não tiver sido observada na cepa em questão. Como exemplo, as tabelas 4.1 e 4.2 ilustram esse procedimento. A partir desses vetores, poderá ser calculada qualquer medida habitualmente utilizada em análise de agrupamentos. Particularmente, neste trabalho foram utilizadas as medidas euclidiana (de distância) e de Jaccard (de similaridade) (Bussab et al. (1990)).

Tabela 4.1 - Exemplo da planilha dos dados.

Cepa	Bandas Observadas				
56217	2,3	2,2	1,4	1,1	0,6
54111	7	1,8	0,9	0,6	
54211	2,8	1,1	0,9	0,6	
41616	1,7	1,3	1,2		
439I13	1,7	1,5	1,4	1,2	
500I14	4,2	2,1	1,3	0,9	

Tabela 4.2 - Exemplo da matriz de dados.

Cepa	7	4,2	2,8	2,3	2,2	2,1	1,8	1,7	1,5	1,4	1,3	1,2	1,1	0,9	0,6
56217	0	0	0	1	1	0	0	0	0	1	0	0	1	0	1
54111	1	0	0	0	0	0	1	0	0	0	0	0	0	1	1
54211	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1
41616	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0
439113	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0
500114	0	1	0	0	0	1	0	0	0	0	1	0	0	1	0

Esse procedimento será adotado para cada indivíduo criando, dessa forma, uma matriz com as linhas contendo os indivíduos e as colunas, as observações. Esse procedimento será adotado para cada ambiente de coleta de dados.

A partir dessa matriz, foram construídas matrizes de distâncias (pela medida de Jaccard e pela distância euclidiana) e, com base nessas matrizes, foram montados os dendrogramas, através dos métodos de Ward (Ward (1963)), da Ligação Simples e da Centróide, para se verificar o agrupamento das cepas.

4.2 O Programa MATRIZES.EXE

O programa MATRIZES.EXE, utilizado para montar as matrizes necessárias à análise estatística, foi desenvolvido e compilado na linguagem Turbo Pascal (a partir do arquivo fonte MATRIZES.PAS) e está disponível no disquete do projeto.

O pacote utilizado para a análise de agrupamentos (SPSS) não fornece, dentre as opções de medidas de distância, a medida de Jaccard. Por isso, além da matriz de dados (de zeros e uns), o programa MATRIZES.EXE produz também a matriz de distâncias de Jaccard. Além disso, esse programa também produz uma matriz de distâncias calculadas a partir de uma medida alternativa. Essa medida, cuja apresentação e explicação se encontram na Seção 4.6, é uma sugestão para resolver o problema de um par de cepas aparentemente mais distantes entre si, apresentarem o mesmo valor de distância de outro par de cepas aparentemente mais similares.

Figura 4.1 - Tela de execução do programa MATRIZES.EXE

```
Digite o nome do diretório de trabalho: b:\dados\ascii

Digite o nome do arquivo que contém a matriz de entrada: riboaah.prn

Este arquivo contém: (1) Dados Originais
                     (2) Matriz de 0 e 1
Digite a opção: 1

Digite o nome do arquivo que conterá a matriz de saída: riboaah.jac

Este arquivo conterá: (1) Matriz de 0 e 1
                     (2) Matriz de Distâncias de Jaccard
                     (3) Matriz de Distâncias Alternativas
Digite a opção: 2

Digite o número de elementos do arquivo B:\DADOS\ASCII\RIBOAAH.PRN: 51

Gravando arquivo B:\DADOS\ASCII\RIBOAAH.JAC: 4.59%
```

Para executar o programa a partir do prompt do DOS (C:\>) deve-se digitar **MATRIZES** e pressionar a tecla <ENTER>. Em seguida aparecerá a tela de execução do programa. A Figura 4.1 apresenta um exemplo dessa tela de execução.

A execução do programa **MATRIZES.EXE** consiste em fornecer, em cinco comandos, todos os parâmetros necessários ao trabalho. Após cada resposta, deve-se pressionar <ENTER> para se passar à próxima pergunta.

Inicialmente, deve-se indicar o diretório de trabalho (aquele em que está localizada a matriz de entrada e em que será gravada a matriz de saída). Caso o diretório de trabalho seja o mesmo em que se está executando o programa (diretório *default*), não é necessário digitar seu nome, sendo suficiente pressionar a tecla <ENTER>.

Em seguida, deve-se indicar o nome do arquivo que conterá a matriz de entrada (aquela a partir da qual será calculada a matriz que o usuário deseja obter). Depois, deve-se digitar 1 ou 2, conforme seja o tipo da matriz de entrada: 1 para matriz de dados originais (semelhante à tabela 4.1) e 2 para matriz de 0 e 1 (tal como a tabela 4.2).

Após isso, deve-se indicar o nome do arquivo que contém a matriz de saída (aquela que o usuário deseja obter). Depois, deve-se digitar 1, 2 ou 3, conforme seja o tipo da matriz de saída: 1 para matriz de 0 e 1, 2 para a matriz de distâncias de Jaccard e 3 para a matriz de distâncias alternativas (a medida explicada no Apêndice C).

Finalmente, deve-se indicar o tamanho (número de cepas) dos arquivos (matrizes) envolvidos, pois todos os procedimentos de cálculo do programa dependem deste valor.

Durante a execução do programa, um contador percentual, localizado na base da tela, indica o andamento do processo de cálculo e leitura/gravação de arquivos.

Cabe mencionar, ainda, que o programa é dotado de um procedimento interno de verificação de consistência, de modo que se for digitado o nome de um diretório ou arquivo de entrada inexistente, o programa é imediatamente abortado. Na especificação do tipo de matriz utilizada, se for digitado um caracter que não seja numérico, o programa será abortado. Se for digitado um número diferente de 1 ou 2 (para ambas as matrizes) ou 3 (para a matriz de saída), o programa ficará paralisado até que se digite um número válido. Finalmente, se a matriz de entrada for do tipo 2 (matriz de 0 e 1), o programa não aceitará a opção 1 (matriz de 0 e 1) para a matriz de saída, procedendo da mesma forma descrita acima.

4.3 Dendrogramas

Após a execução do programa MATRIZES.EXE, foram construídos dendrogramas através do pacote estatístico SPSS. Os arquivos de programa com os comandos para construir os dendrogramas estão disponíveis no disquete do projeto. Foram analisados dados de sete grupos de ambientes (águas, animais, humanos, águas+animais, águas+humanos, animais+humanos e águas+animais+humanos), coletados de duas formas (RAPD e Ribotipagem), utilizando-se duas medidas (euclídeana e Jaccard) e três métodos de agrupamento (Ligação Simples, Centróide e Ward), constituindo 84 análises ao todo. Por motivo de espaço, apenas os dendrogramas referentes à subpopulação animais para as técnicas de Ribotipagem e

RAPD, encontram-se neste trabalho (Apêndice B). O conjunto completo dos 84 dendrogramas, juntamente com as *Agglomeration Schedule* (quadros que mostram, nó a nó, os objetos agrupados e seus respectivos níveis de junção), está disponível no disquete do projeto.

4.4 Observações

Por causa das limitações dos programas (MATRIZES.EXE e SPSS), algumas providências devem ser tomadas para a correta manipulação dos arquivos:

- Para que possa utilizá-la como matriz de entrada em execuções futuras, o programa MATRIZES.EXE produz a matriz de 0 e 1 com um cabeçalho com o valor das bandas correspondentes a cada coluna. Para utilizar essa matriz como dado de entrada no SPSS, é necessário eliminar o cabeçalho, através de algum editor compatível com o padrão ASCII.
- No SPSS, quando se utiliza a matriz de distâncias de Jaccard, as cepas são consideradas como variáveis. Por adequação às regras de validação de nomes de variáveis do SPSS, algumas mudanças foram feitas nos nomes das cepas:
 - Só são aceitos letras e números nos nomes de variáveis. Assim, 255-H tornou-se 255H.
 - O primeiro caracter do nome deve ser uma letra. Assim, a todas as cepas, foi acrescentada a letra C antes de seu nome. Por exemplo, 255H passou a ser C255H.
 - O tamanho máximo permitido para um nome de variável é oito caracteres. Assim, C507jagll se tornou C507jagl.
- O SPSS, ao construir os dendrogramas, produz uma régua com graduação padronizada de 0 a 25 unidades. Para que se possa graduá-la em função dos níveis de agrupamentos, foi desenvolvido o programa CONVERTE, compilado na linguagem Turbo Pascal (a partir do arquivo fonte CONVERTE.PAS).

4.5 O Programa CONVERTE.EXE

O programa CONVERTE.EXE, utilizado para converter a graduação da régua do SPSS, está disponível no disquete do projeto.

Para executar o programa a partir do prompt do DOS (C:\>) deve-se digitar CONVERTE e pressionar a tecla <ENTER>. Em seguida aparecerá a tela de execução do programa. A Figura 4.2 apresenta um exemplo dessa tela de execução.

A execução do programa CONVERTE.EXE consiste em fornecer, em três comandos, todos os parâmetros necessários ao trabalho. Após cada resposta, deve-se pressionar <ENTER> para se passar à próxima pergunta.

Inicialmente, deve-se indicar o tipo de medida de distância utilizada (se de similaridade ou de dissimilaridade) para que se saiba o sentido da graduação: se a medida for de dissimilaridade, a graduação será decrescente, enquanto se for de similaridade, a graduação será crescente.

Figura 4.2 - Tela de execução do programa CONVERTE.EXE

A medida utilizada foi de: (1) Dissimilaridade (Por exemplo: Euclideana)			
(2) Similaridade (Por exemplo: Jaccard)			
Digite a opção: 2			
Nível mínimo ocorrido: 0.143			
Nível máximo ocorrido: 1.000			
Valor Padronizado:	0	Valor Calculado:	1.000
Valor Padronizado:	5	Valor Calculado:	0.829
Valor Padronizado:	10	Valor Calculado:	0.657
Valor Padronizado:	15	Valor Calculado:	0.486
Valor Padronizado:	20	Valor Calculado:	0.314
Valor Padronizado:	25	Valor Calculado:	0.143

Em seguida, deve-se indicar os valores mínimo e máximo dos níveis de junção listados na *Agglomeration Schedule* (quadro que mostra, nó a nó, os objetos agrupados e seus respectivos níveis de junção), pois os procedimentos de cálculo do programa dependem destes valores.

Após isso, o programa produz a listagem das seis graduações em valores padronizados (aqueles de 0 a 25, produzidos pelo SPSS) e calculados (em função dos níveis de junção).

Cabe mencionar, ainda, que o programa é dotado de um procedimento interno de verificação de consistência, de modo que se for digitado um número diferente de 1 ou 2, na especificação do tipo de medida, o programa ficará paralisado até que se digite um número válido.

4.6 Medida de distância usando a variabilidade dos dados

Esta seção tem como objetivo apresentar uma medida alternativa de distância (diferença) entre duas cepas, que seja mais fiel ao que foi observado.

As medidas tradicionalmente utilizadas (euclideana e Jaccard), têm como principal característica o fato de não considerarem a posição das bandas de *status* diferentes. Como *status*, define-se a ocorrência ou a não-ocorrência de uma banda em uma cepa.

Como exemplo, suponha-se que sejam observados os dados abaixo.

Tabela 4.3 - Exemplo 1

Cepa	Bandas Observadas		
C1	1.3	1.5	1.6
C2	1.3	1.5	1.8
C3	1.3	1.5	5.0

Tabela 4.4 - Matriz de Dados do Exemplo 1

Cepa	1.3	1.5	1.6	1.8	5.0
C1	1	1	1	0	0
C2	1	1	0	1	0
C3	1	1	0	0	1

Calculando-se as medidas tradicionais, estes dados levam às seguintes matrizes de distâncias:

Tabela 4.5 - Matriz de distâncias euclideanas para o Exemplo 1

	C1	C2	C3
C1	0.000		
C2	1.414	0.000	
C3	1.414	1.414	0.000

Tabela 4.6 - Matriz de distâncias de Jaccard para o Exemplo 1

	C1	C2	C3
C1	0.000		
C2	0.200	0.000	
C3	0.200	0.200	0.000

Embora as matrizes acima traduzam uma relação de equidistância, tem-se, observando os dados originais (Tabela 4.3), a noção intuitiva (através da proximidade entre as bandas observadas em cada cepa) de que as cepas C1 e C2 são mais semelhantes entre si do que com a cepa C3.

Pensando nesse fato, pode-se sugerir a utilização de uma medida de distância entre duas cepas, baseada nos valores de bandas que apresentem *status* diferentes (bandas que estejam presentes em uma cepa e ausentes em outra). O Apêndice C apresenta essa medida alternativa, acompanhada de alguns exemplos. A tabela 4.7 abaixo apresenta a matriz de distâncias segundo essa medida.

Tabela 4.7 - Matriz de distâncias alternativas para o Exemplo 1

	C1	C2	C3
C1	0.000		
C2	1,020	0.000	
C3	6,780	6,120	0.000

5 Conclusões e sugestões

Conforme se pode ver nos Gráficos C.1 a C.3 (do Apêndice C), a medida alternativa sugerida neste trabalho, realmente produz um dendrograma onde se nota o agrupamento de cepas que apresentem bandas observadas mais próximas entre si. Nesses gráficos, a cepa número 7 (C402) aparece isolada das demais, quando se utiliza as medidas euclideana e de Jaccard (medidas baseadas exclusivamente na

presença ou ausência de cada banda). Esse fato é justificado através das tabelas C.1 e C.2, onde se pode perceber que nenhuma outra cepa tem bandas em comum com a C402, resultando em total dissimilaridade.

Entretanto, quando se considera a medida alternativa (gráfico C.3), a cepa C402 aparece não mais isolada, mas unida às outras. Isso acontece porque, embora as bandas não sejam coincidentes com as de outras cepas, elas são próximas o suficiente, para a medida acusar a similaridade. Por exemplo, as bandas 9.4, 7.3 e 6.0 da cepa C402, embora tenham valores diferentes, são próximas das bandas 9.6, 7.8 e 6.3 das cepas, CH1, CH4 e C196.

Assim, a pesquisadora tem à sua disposição, as medidas (euclideana, Jaccard e alternativa) e programas (MATRIZES.EXE, SPSS e CONVERTE.EXE) citados na Seção 4, que poderão auxiliá-la na construção de quaisquer dendrogramas pelos métodos utilizados neste trabalho (ligação simples, centróide, Ward) ou por outros disponíveis em pacotes.

Cabe colocar que a interpretação e a análise dos dendrogramas pertinentes a este estudo devem ser realizadas pela pesquisadora, pois é ela quem possui os conhecimentos específicos necessários sobre as cepas agrupadas.

Finalmente, sugere-se à pesquisadora o estudo, e eventual adoção, das outras medidas apresentadas, como alternativas à euclideana, que tradicionalmente vem sendo utilizada nos trabalhos da área.

APÊNDICE A

Tabelas

A.1 - Dados obtidos pela técnica RAPD

Tabela A.1 - Distribuição de frequências do número de bactérias em que foi observada cada banda.

Banda	Animais		Humanos		Águas	
	Absoluto	Relativo	Absoluto	Relativo	Absoluto	Relativo
0,5	4	11,76%	1	2,33%	2	2,04%
0,6	11	32,35%	14	32,56%	26	26,53%
0,7	14	41,18%	11	25,58%	13	13,27%
0,8	6	17,65%	6	13,95%	11	11,22%
0,9	15	44,12%	12	27,91%	30	30,61%
1	22	64,71%	23	53,49%	32	32,65%
1,1	8	23,53%	24	55,81%	46	46,94%
1,2	14	41,18%	17	39,53%	34	34,69%
1,3	10	29,41%	14	32,56%	30	30,61%
1,4	9	26,47%	13	30,23%	25	25,51%
1,5	5	14,71%	8	18,60%	23	23,47%
1,6	8	23,53%	8	18,60%	12	12,24%
1,7	4	11,76%	8	18,60%	19	19,39%
1,8	2	5,88%	2	4,65%	9	9,18%
1,9	3	8,82%	13	30,23%	17	17,35%
2	1	2,94%	1	2,33%	10	10,20%
2,1	8	23,53%	10	23,26%	8	8,16%
2,2	3	8,82%	9	20,93%	8	8,16%
2,3	6	17,65%	1	2,33%	15	15,31%
2,4	4	11,76%	3	6,98%	10	10,20%
2,5	2	5,88%	9	20,93%	7	7,14%
2,6	2	5,88%	1	2,33%	11	11,22%
2,7	1	2,94%	4	9,30%	5	5,10%
2,8	0	0,00%	1	2,33%	3	3,06%
2,9	2	5,88%	1	2,33%	4	4,08%
3	0	0,00%	1	2,33%	3	3,06%
3,1	0	0,00%	0	0,00%	1	1,02%
3,3	0	0,00%	1	2,33%	0	0,00%
3,4	0	0,00%	1	2,33%	0	0,00%
3,5	2	5,88%	2	4,65%	2	2,04%
3,6	1	2,94%	0	0,00%	1	1,02%
3,7	0	0,00%	1	2,33%	3	3,06%
3,8	0	0,00%	3	6,98%	0	0,00%
3,9	0	0,00%	0	0,00%	1	1,02%
4	0	0,00%	0	0,00%	1	1,02%
4,2	0	0,00%	0	0,00%	3	3,06%
4,6	1	2,94%	0	0,00%	0	0,00%
4,7	0	0,00%	0	0,00%	1	1,02%
4,8	0	0,00%	0	0,00%	1	1,02%
4,9	1	2,94%	0	0,00%	0	0,00%
5	0	0,00%	0	0,00%	1	1,02%

5,2	1	2,94%	0	0,00%	2	2,04%
-----	---	-------	---	-------	---	-------

Tabela A.1 (continuação)- Distribuição de frequências do número de bactérias em que foi observada cada banda.

Banda	Animais		Humanos		Águas	
	Absoluto	Relativo	Absoluto	Relativo	Absoluto	Relativo
5,6	1	2,94%	0	0,00%	0	0,00%
5,8	0	0,00%	0	0,00%	3	3,06%
5,9	0	0,00%	1	2,33%	0	0,00%
6,8	0	0,00%	1	2,33%	0	0,00%
7	0	0,00%	0	0,00%	1	1,02%

Tabela A.2 - Distribuição de frequências do número de bandas em cada bactéria

Número de Bandas	Animais			Humanos			Águas		
	Absoluto	Relativo	Acumulado	Absoluto	Relativo	Acumulado	Absoluto	Relativo	Acumulado
1	2	5,88%	5,88%	1	2,33%	2,33%	6	6,12%	6,12%
2	3	8,82%	14,71%	1	2,33%	4,65%	9	9,18%	15,31%
3	4	11,76%	26,47%	5	11,63%	16,28%	14	14,29%	29,59%
4	6	17,65%	44,12%	7	16,28%	32,56%	30	30,61%	60,20%
5	3	8,82%	52,94%	13	30,23%	62,79%	15	15,31%	75,51%
6	8	23,53%	76,47%	7	16,28%	79,07%	11	11,22%	86,73%
7	3	8,82%	85,29%	4	9,30%	88,37%	3	3,06%	89,80%
8	3	8,82%	94,12%	3	6,98%	95,35%	6	6,12%	95,92%
9	1	2,94%	97,06%	1	2,33%	97,67%	2	2,04%	97,96%
10	1	2,94%	100,00%	0	0,00%	97,67%	2	2,04%	100,00%
11	0	0,00%	100,00%	1	2,33%	100,00%	0	0,00%	100,00%
Total	34	100,00%		43	100,00%		98	100,00%	

Tabela A.3 - Estatísticas Descritivas do número de bandas em cada bactéria

	Animais	Humanos	Águas
Média	5,03	5,23	4,43
Mediana	5	5	4
Mínimo	1	1	1
Máximo	10	11	10
Desvio-Padrão	2,23	1,88	1,99
Nº de Observações	34	43	98

A.2 - Dados obtidos pela técnica de Ribotipagem

Tabela A.4 - Distribuição de frequências do número de bactérias em que foi observada cada banda.

Banda	Águas		Animais		Humanos	
	Absoluto	Relativo	Absoluto	Relativo	Absoluto	Relativo
1,4	0	0,00%	1	5,88%	2	13,33%
1,5	0	0,00%	5	29,41%	5	33,33%
1,6	0	0,00%	0	0,00%	2	13,33%
1,7	4	21,05%	0	0,00%	0	0,00%
1,9	0	0,00%	1	5,88%	5	33,33%
2,1	0	0,00%	1	5,88%	1	6,67%
2,3	0	0,00%	1	5,88%	1	6,67%
2,7	0	0,00%	1	5,88%	0	0,00%
2,9	0	0,00%	0	0,00%	1	6,67%
3	0	0,00%	3	17,65%	1	6,67%
3,1	0	0,00%	5	29,41%	0	0,00%
3,2	0	0,00%	1	5,88%	0	0,00%
3,3	4	21,05%	1	5,88%	0	0,00%
3,4	0	0,00%	0	0,00%	1	6,67%
3,5	0	0,00%	1	5,88%	0	0,00%
3,6	0	0,00%	6	35,29%	5	33,33%
3,7	6	31,58%	0	0,00%	1	6,67%
3,9	0	0,00%	13	76,47%	1	6,67%
4,1	0	0,00%	0	0,00%	1	6,67%
4,2	6	31,58%	4	23,53%	3	20,00%
4,3	0	0,00%	0	0,00%	9	60,00%
4,5	5	26,32%	12	70,59%	1	6,67%
4,7	0	0,00%	0	0,00%	10	66,67%
4,8	6	31,58%	4	23,53%	2	13,33%
4,9	13	68,42%	1	5,88%	0	0,00%
5	0	0,00%	1	5,88%	4	26,67%
5,2	1	5,26%	10	58,82%	0	0,00%
5,5	3	15,79%	5	29,41%	2	13,33%
5,6	0	0,00%	1	5,88%	12	80,00%
6	11	57,89%	0	0,00%	0	0,00%
6,1	1	5,26%	5	29,41%	0	0,00%
6,3	0	0,00%	0	0,00%	10	66,67%
6,4	0	0,00%	6	35,29%	1	6,67%
6,5	0	0,00%	2	11,76%	0	0,00%
6,7	1	5,26%	0	0,00%	0	0,00%
6,8	5	26,32%	1	5,88%	2	13,33%

7	1	5,26%	0	0,00%	2	13,33%
---	---	-------	---	-------	---	--------

Tabela A.4 (continuação) - Distribuição de frequências do número de bactérias em que foi observada cada banda.

Banda	Águas		Animais		Humanos	
	Absoluto	Relativo	Absoluto	Relativo	Absoluto	Relativo
7,3	11	57,89%	13	76,47%	0	0,00%
7,6	0	0,00%	1	5,88%	0	0,00%
7,8	0	0,00%	0	0,00%	6	40,00%
7,9	0	0,00%	0	0,00%	1	6,67%
8	0	0,00%	2	11,76%	0	0,00%
8,3	6	31,58%	4	23,53%	1	6,67%
8,4	1	5,26%	0	0,00%	0	0,00%
8,5	0	0,00%	0	0,00%	3	20,00%
9	0	0,00%	1	5,88%	1	6,67%
9,3	0	0,00%	1	5,88%	0	0,00%
9,4	4	21,05%	0	0,00%	0	0,00%
9,6	0	0,00%	0	0,00%	10	66,67%
9,9	9	47,37%	1	5,88%	0	0,00%
11	1	5,26%	2	11,76%	0	0,00%
11,1	0	0,00%	0	0,00%	1	6,67%
11,3	10	52,63%	0	0,00%	0	0,00%
12	0	0,00%	1	5,88%	0	0,00%
13	0	0,00%	0	0,00%	3	20,00%
13,8	0	0,00%	0	0,00%	1	6,67%
15,8	0	0,00%	0	0,00%	1	6,67%
43,8	0	0,00%	0	0,00%	3	20,00%

Tabela A.5 - Distribuição de frequências do número de bandas em cada bactéria

Número de Bandas	Animais			Humanos			Águas		
	Absoluto	Relativo	Acumulado	Absoluto	Relativo	Acumulado	Absoluto	Relativo	Acumulado
1	0	0,00%	0,00%	0	0,00%	0,00%	0	0,00%	0,00%
2	0	0,00%	0,00%	0	0,00%	0,00%	0	0,00%	0,00%
3	0	0,00%	0,00%	1	6,67%	6,67%	3	15,79%	15,79%
4	1	5,88%	5,88%	1	6,67%	13,33%	1	5,26%	21,05%
5	5	29,41%	35,29%	2	13,33%	26,67%	3	15,79%	36,84%
6	2	11,76%	47,06%	0	0,00%	26,67%	4	21,05%	57,89%
7	3	17,65%	64,71%	3	20,00%	46,67%	7	36,84%	94,74%
8	2	11,76%	76,47%	0	0,00%	46,67%	1	5,26%	100,00%
9	2	11,76%	88,24%	4	26,67%	73,33%	0	0,00%	100,00%
10	1	5,88%	94,12%	2	13,33%	86,67%	0	0,00%	100,00%
11	1	5,88%	100,00%	2	13,33%	100,00%	0	0,00%	100,00%

Total	17	100,00%
--------------	----	---------

15	100,00%
----	---------

19	100,00%
----	---------

Tabela A.6 - Estatísticas Descritivas do número de bandas em cada bactéria

	Animais	Humanos	Águas
Média	6,88	7,73	5,74
Mediana	7	9	6
Mínimo	4	3	3
Máximo	11	11	8
Desvio-Padrão	1,97	2,46	1,52
Nº de Observações	17	15	19

APÊNDICE B

Dendrogramas

B.1 - Dados obtidos para a subpopulação de animais pela técnica RAPD

B.1.1 - Distância Euclideana

Gráfico B.1 - Dendrograma segundo o método da ligação simples

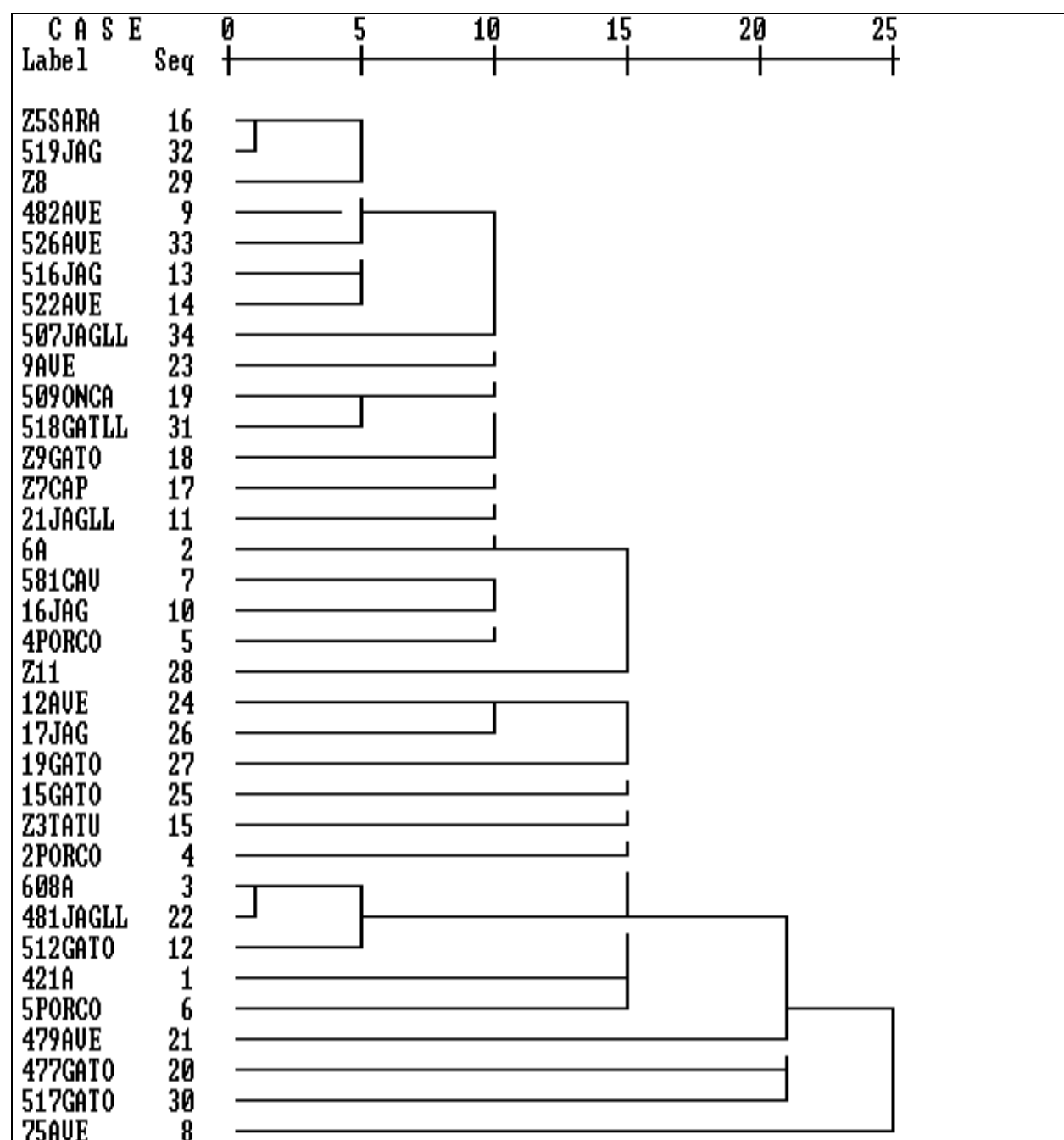


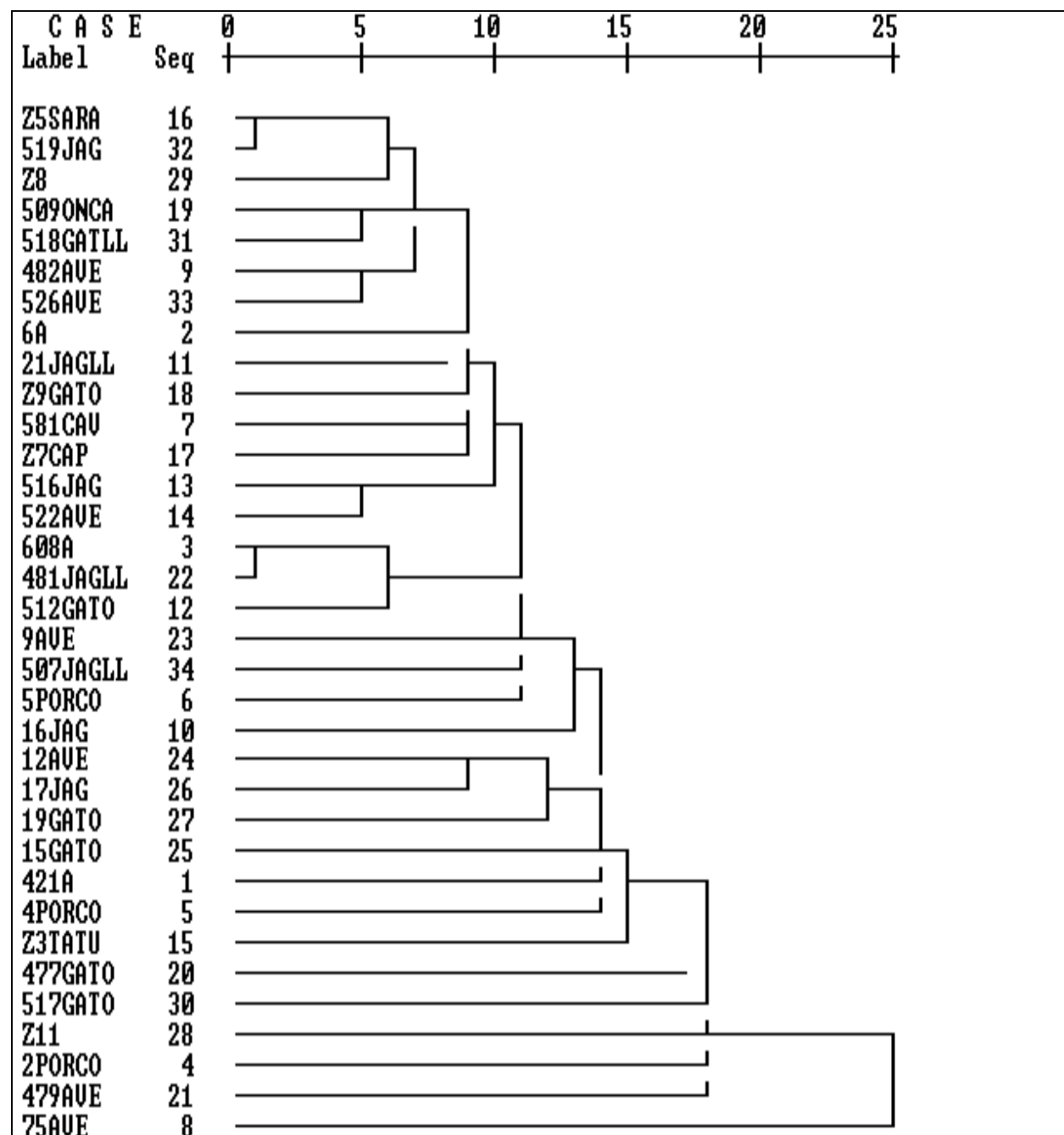
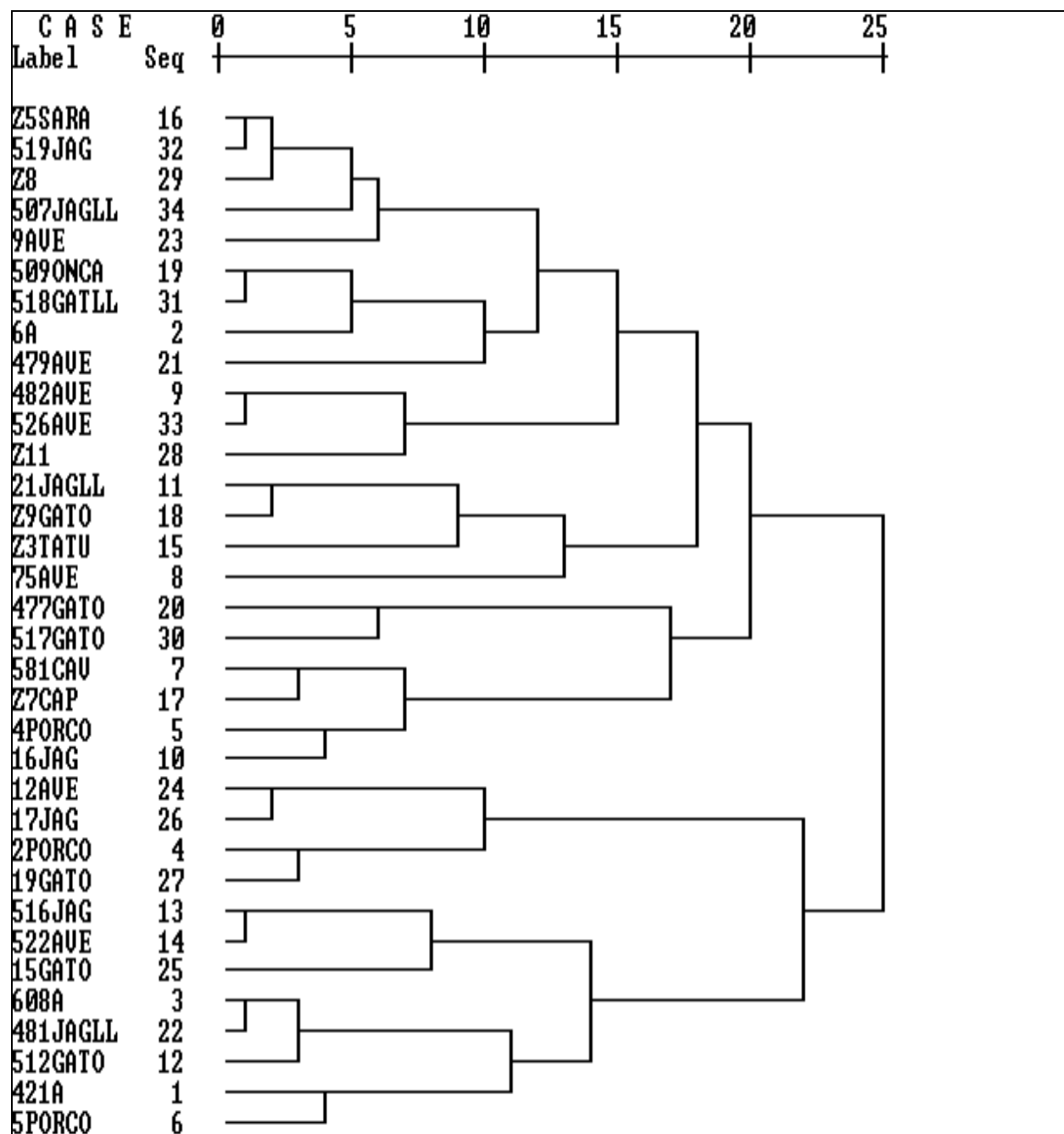
Gráfico B.2 - Dendrograma segundo o método da centróide

Gráfico B.3 - Dendrograma segundo o método de Ward

B.1.2 - Distância de Jaccard

Gráfico B.4 - Dendrograma segundo o método da Ligação Simples

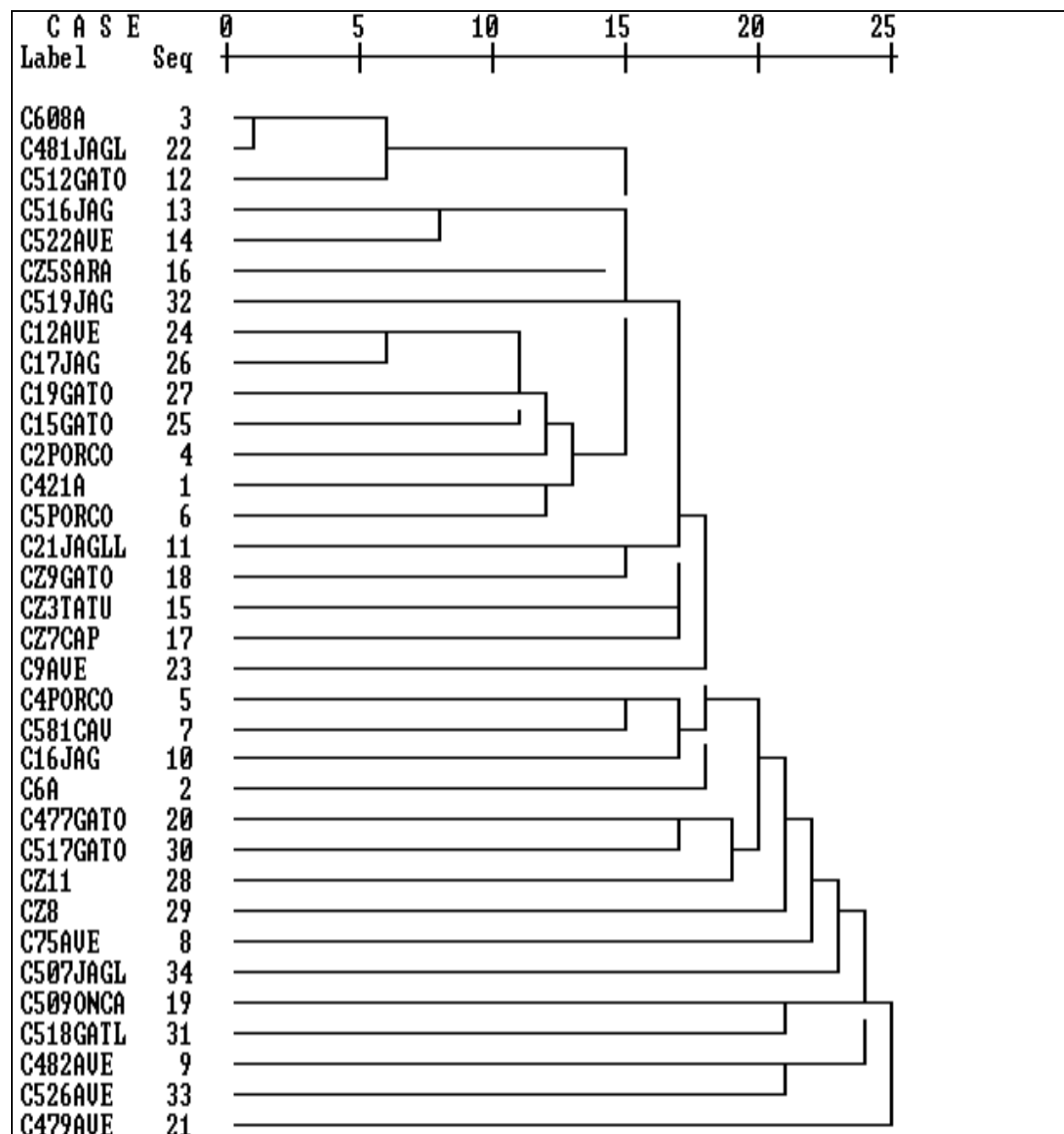


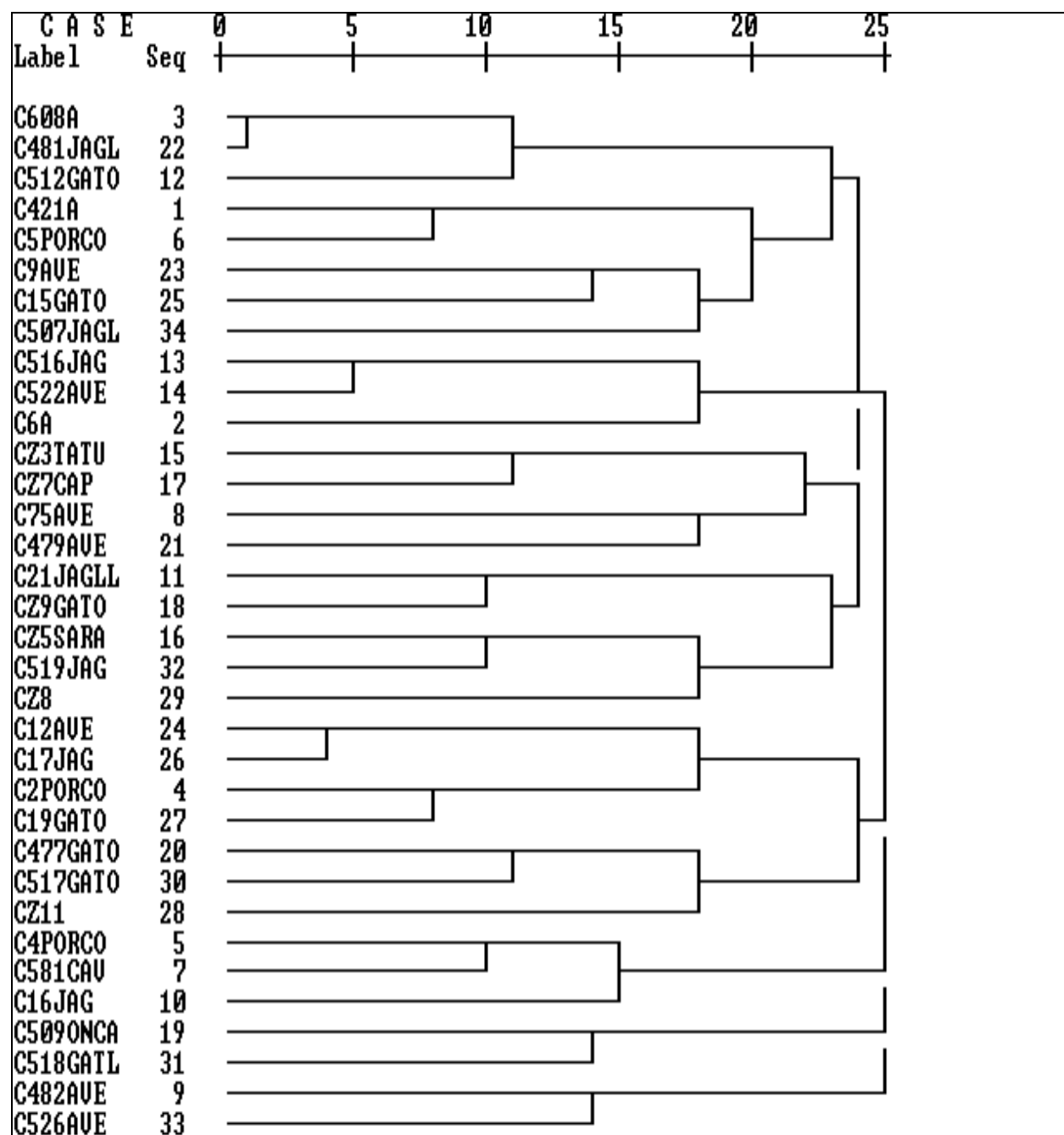
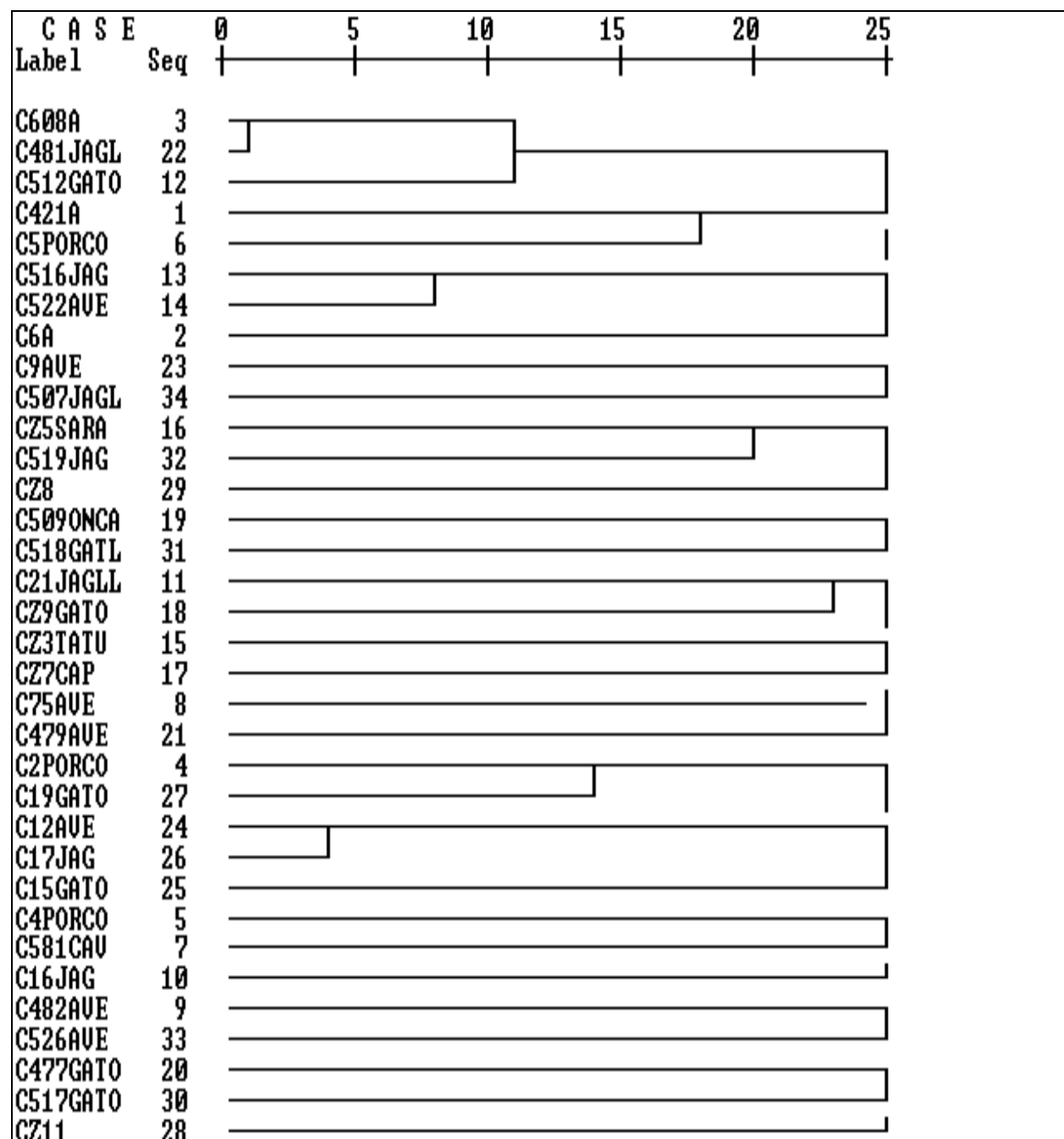
Gráfico B.5 - Dendrograma segundo o método da centróide

Gráfico B.6 - Dendrograma segundo o método de Ward

B.2 - Dados obtidos para a subpopulação de animais pela técnica de Ribotipagem

B.2.1 - Distância Euclideana

Gráfico B.7 - Dendrograma segundo o método da Ligação Simples

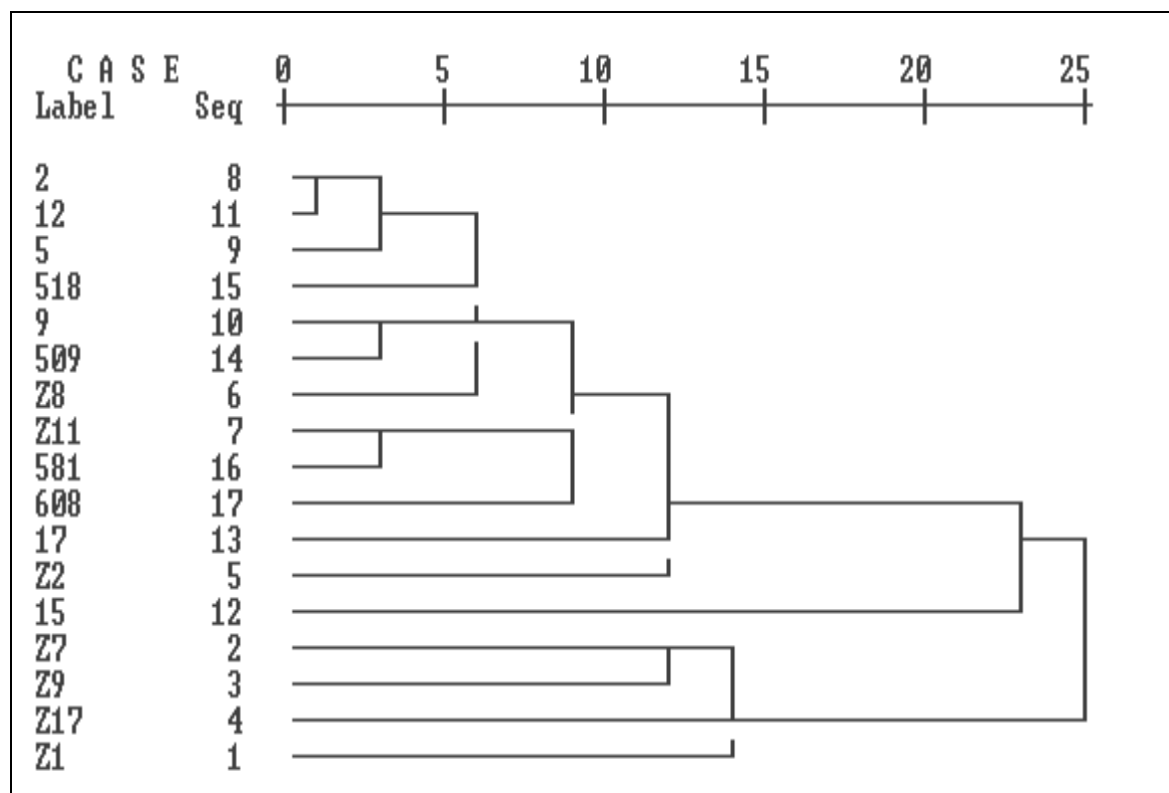
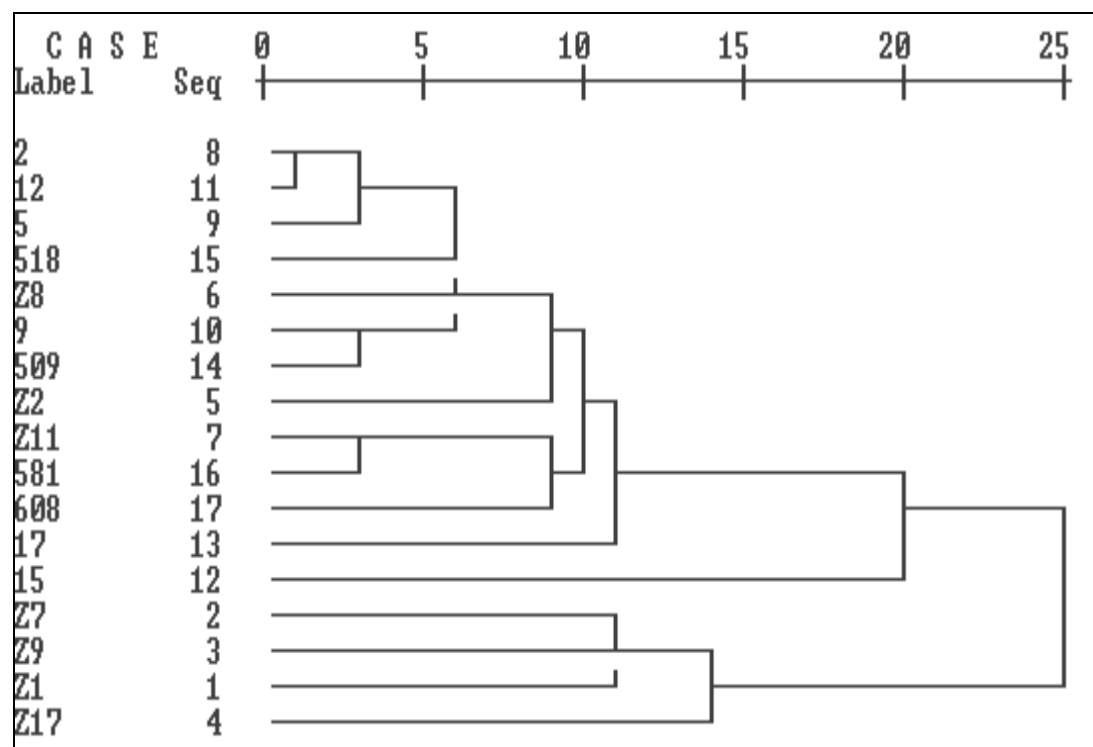
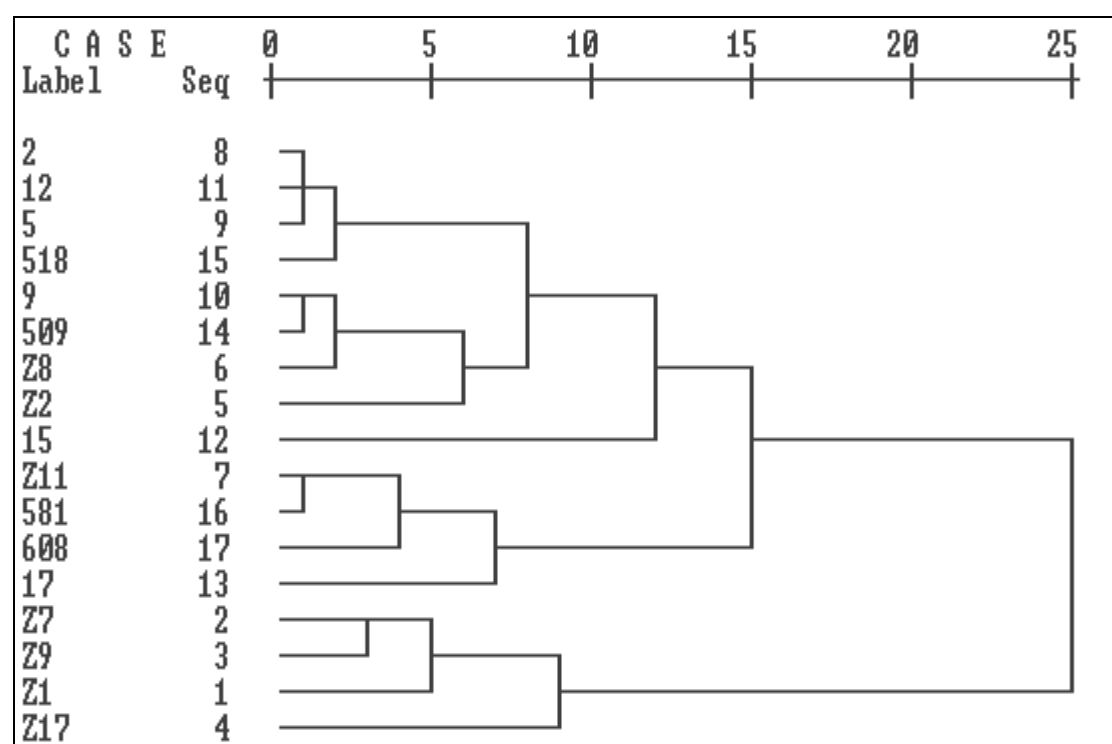


Gráfico B.8 - Dendrograma segundo o método da Centróide**Gráfico B.9** - Dendrograma segundo o método de Ward

B.2.2 - Distância de Jaccard

Gráfico B.10 - Dendrograma segundo o método da Ligação Simples

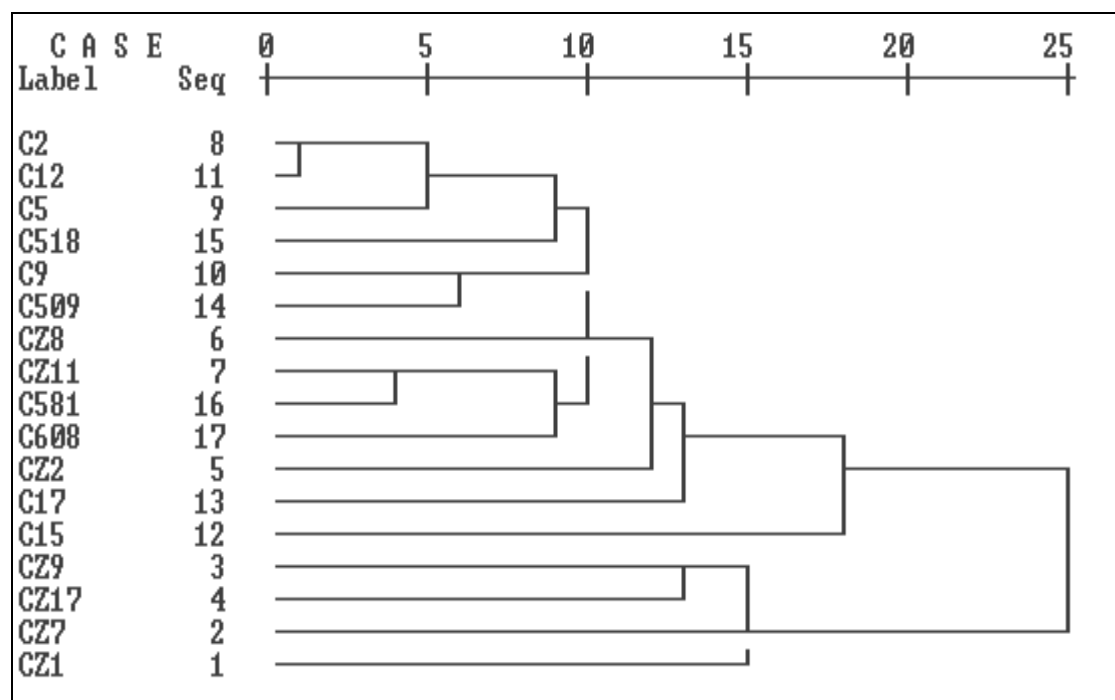


Gráfico B.11 - Dendrograma segundo o método da Centróide

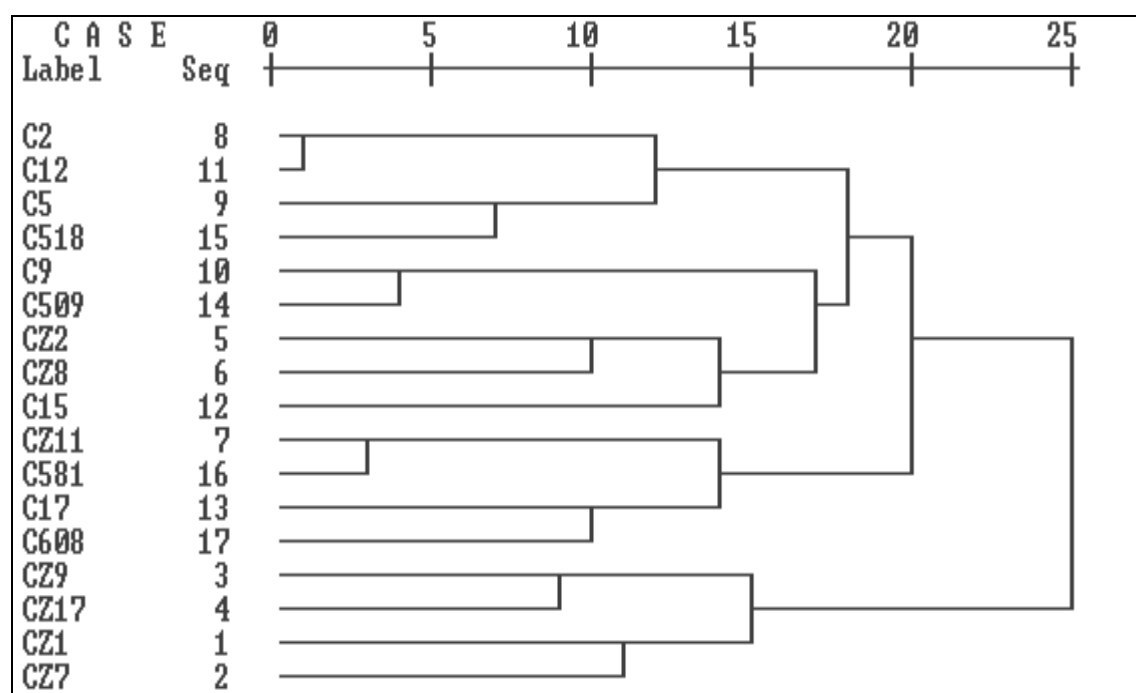
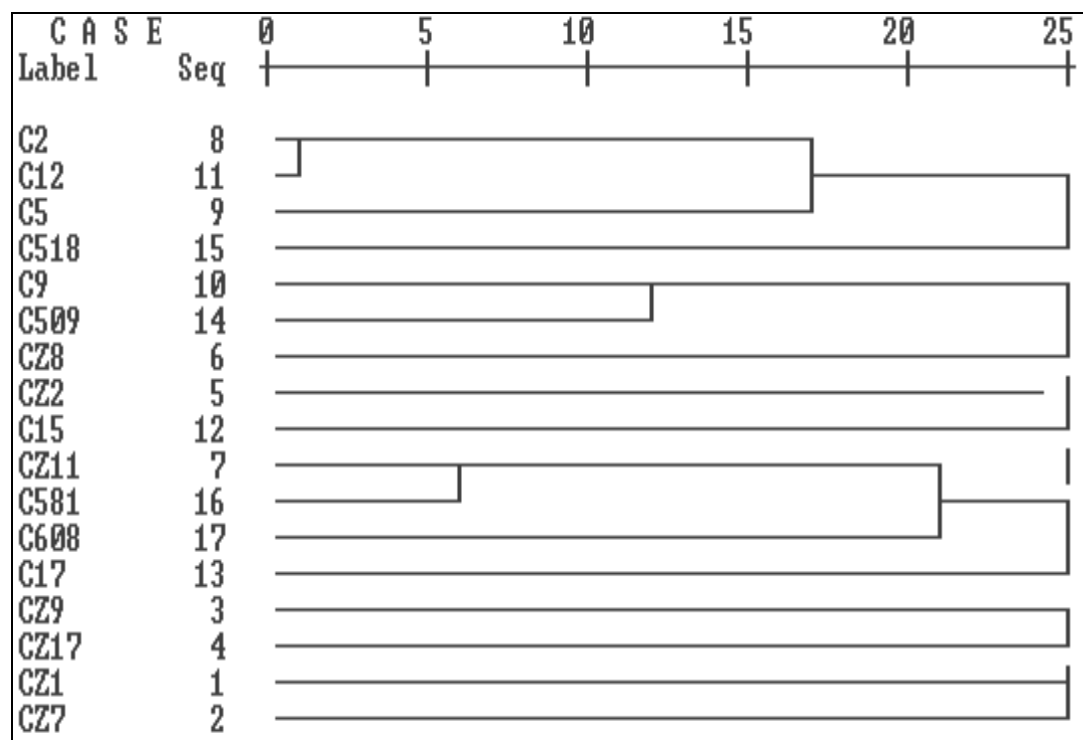


Gráfico B.12 - Dendrograma segundo o método de Ward

APÊNDICE C

Uma alternativa para as medidas de distância

C.1 - Apresentação da Medida

A primeira alternativa em que se poderia pensar é a de utilizar a variância desses valores. Assim, no Exemplo 1 (da Seção 4.6), as distâncias entre as cepas seriam dadas por:

$$d_{C1C2} = \text{Var}[\{16; 18\}] = 0.02;$$

$$d_{C1C3} = \text{Var}[\{16; 5.0\}] = 5.78 \text{ e}$$

$$d_{C2,C3} = \text{Var}[\{18; 5.0\}] = 5.12$$

Essa medida será referida, posteriormente, como distância D1.

Entretanto, essa medida apresenta um inconveniente, perceptível nos pares de cepas que possuam apenas uma banda com *status* diferente. A medida de distância, nesse caso, será dada pela variância de um conjunto de apenas um elemento, que é, por definição, igual a zero. Porém, se as duas cepas forem completamente idênticas (nenhuma cepa com *status* diferente), a medida de distância também será zero. Torna-se necessária, então, uma correção nessa medida para que esse fato não ocorra.

Seja n o número (quantidade) de bandas com *status* diferentes. O problema consiste, então, em encontrar uma medida que leve em consideração os valores das n bandas de *status* diferentes. A medida D1 (variância amostral) permite que, por exemplo, duas cepas com $n=3$ bandas de *status* diferentes sejam consideradas mais parecidas que outras duas cepas com $n=2$ bandas. Por exemplo, um par de cepas que sejam diferentes nas bandas 0.9, 1.1 e 1.2 é, algébrica (de acordo com a medida D1) e intuitivamente mais similar que outro par cujas bandas de *status* diferentes sejam 1.5 e 4.8, pois essas estão menos concentradas que as primeiras.

Além disso, um par de cepas que possua apenas 1 banda com *status* diferente ($n=1$), é necessariamente menos similar que outro par absolutamente idêntico ($n=0$), pois há uma banda de diferença. Esse par (com $n=1$ banda) também é necessariamente mais similar que outro com $n=2$ bandas, pois enquanto a concentração de suas bandas de *status* diferentes se limita a um único ponto, a do outro par está dispersa em dois pontos.

Assim, a correção para a medida de distâncias alternativas D1 que se propõe, considerando todos os fatos citados, é dada por:

$$D2 = \begin{cases} 0 & \text{se } n = 0 \\ 1 & \text{se } n = 1 \\ \sigma^2 + 1 & \text{se } n \geq 2 \end{cases}$$

Cabe ressaltar que qualquer número maior que zero poderia ser colocado no lugar do número 1. Assim, uma versão genérica da medida seria dada por

$$D2 = \begin{cases} 0 & \text{se } n = 0 \\ k & \text{se } n = 1 \\ \sigma^2 + k & \text{se } n \geq 2 \end{cases} \quad (k > 0),$$

onde o valor de k determina a intensidade da medida em um par de cepas diferentes. Quanto maior for o valor de k, maior será a diferença entre as medidas em um par de cepas diferentes ($D2 \geq k$) e em um par de cepas idênticas ($D2 = 0$). Da mesma forma, quanto menor for o valor de k, menor será esse contraste.

C.2 - Exemplos da medida

Foram escolhidas 10 cepas dentre as 226 observadas (175 pelo método RAPD e 51 pelo de ribotipagem) para exemplificar a medida alternativa. As tabelas C.1 e C.2 mostram as matrizes de dados originais e de 0 e 1 e as tabelas C.3 a C.5 ilustram as matrizes das três medidas de distâncias.

Tabela C.1 - Matriz de Dados Originais (Exemplo 2)

Cepa	Bandas Observadas								
244-H	1.0	0.9	0.8	0.7					
193-H	1.9	1.4	1.3	1.2	1.1	1.0	0.8	0.6	
252-H	2.5	1.4	1.1	0.9	0.7				
124-H	3.3	2.1	1.9	1.5	1.1	0.6			
239-H	3.8	1.7	1.3	1.1	0.6				
477GATO	5.6	2.2	1.4	1.3	1.2	1.1			
402	11.3	9.4	7.3	6.0	5.2	4.9	4.5		
H1	9.6	7.8	6.3	5.6	4.7	4.3	3.6	1.9	1.5
H4	9.6	7.8	6.3	5.6	4.7	4.3	3.6	1.9	1.5
196	9.6	7.8	6.3	5.6	5.0	4.7	3.9		

Tabela C.2 - Matriz de Dados (de 0 e 1) (Exemplo 2)

Cepa	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.7	1.9	2.1	2.2	2.5
244-H	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0
193-H	1	0	1	0	1	1	1	1	1	0	0	1	0	0	0
252-H	0	1	0	1	0	1	0	0	1	0	0	0	0	0	1
124-H	1	0	0	0	1	0	0	0	0	1	0	1	1	0	0
239-H	1	0	0	0	0	1	0	1	0	0	1	0	0	0	0
477GATO	0	0	0	0	0	1	1	1	1	0	0	0	0	1	0
402	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
H4	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
196	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tabela C.2 (continuação) - Matriz de Dados (de 0 e 1) (Exemplo 2)

Cepa	3.3	3.6	3.8	3.9	4.3	4.5	4.7	4.9	5.0	5.2	5.6	6.0	6.3	7.3	7.8
244-H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
193-H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
252-H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
124-H	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
239-H	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
477GATO	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
402	0	0	0	0	0	1	0	1	0	1	0	1	0	1	0
H1	0	1	0	0	1	0	1	0	0	0	1	0	1	0	1
H4	0	1	0	0	1	0	1	0	0	0	1	0	1	0	1
196	0	0	0	1	0	0	1	0	1	0	1	0	1	0	1

Tabela C.2 (continuação) - Matriz de Dados (de 0 e 1) (Exemplo 2)

Cepa	9.4	9.6	11.3
244-H	0	0	0
193-H	0	0	0
252-H	0	0	0
124-H	0	0	0
239-H	0	0	0
477GATO	0	0	0
402	1	0	1
H1	0	1	0
H4	0	1	0
196	0	1	0

Tabela C.3 - Matriz de Distâncias Euclidianas

	244-H	193-H	252-H	124-H	239-H	477GATO	402	H1	H4	196
244-H	0,00	2,83	2,23	2,83	3,00	3,16	3,32	3,61	3,61	3,32
193-H	2,83	0,00	3,00	2,83	2,65	2,45	3,87	3,87	3,87	3,87
252-H	2,23	3,00	0,00	3,32	2,83	2,65	3,46	3,74	3,74	3,46
124-H	2,83	2,83	3,32	0,00	3,00	3,46	3,61	3,32	3,32	3,61
239-H	3,00	2,65	2,83	3,00	0,00	2,65	3,46	3,74	3,74	3,46
477GATO	3,16	2,45	2,65	3,46	2,65	0,00	3,61	3,61	3,61	3,32
402	3,32	3,87	3,46	3,61	3,46	3,61	0,00	4,00	4,00	3,74
H1	3,61	3,87	3,74	3,32	3,74	3,61	4,00	0,00	0,00	2,45
H4	3,61	3,87	3,74	3,32	3,74	3,61	4,00	0,00	0,00	2,45
196	3,32	3,87	3,46	3,61	3,46	3,32	3,74	2,45	2,45	0,00

Tabela C.4 - Matriz de Distâncias de Jaccard

	244-H	193-H	252-H	124-H	239-H	477GATO	402	H1	H4	196
244-H	1,00	0,20	0,29	0,11	0,00	0,00	0,00	0,00	0,00	0,00
193-H	0,20	1,00	0,18	0,27	0,30	0,40	0,00	0,06	0,06	0,00
252-H	0,29	0,18	1,00	0,00	0,11	0,22	0,00	0,00	0,00	0,00
124-H	0,11	0,27	0,00	1,00	0,10	0,00	0,00	0,15	0,15	0,00
239-H	0,00	0,30	0,11	0,10	1,00	0,22	0,00	0,00	0,00	0,00
477GATO	0,00	0,40	0,22	0,00	0,22	1,00	0,00	0,07	0,07	0,08
402	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00
H1	0,00	0,06	0,00	0,15	0,00	0,07	0,00	1,00	1,00	0,46
H4	0,00	0,06	0,00	0,15	0,00	0,07	0,00	1,00	1,00	0,46
196	0,00	0,00	0,00	0,00	0,00	0,08	0,00	0,46	0,46	1,00

Tabela C.5 - Matriz de Distâncias Alternativas

	244-H	193-H	252-H	124-H	239-H	477GATO	402	H1	H4	196
244-H	0,00	1,17	1,45	1,87	1,97	3,13	14,38	9,66	9,66	10,44
193-H	1,17	0,00	1,39	1,62	2,02	4,49	12,79	9,29	9,29	9,33
252-H	1,45	1,39	0,00	1,70	2,16	3,88	13,13	8,83	8,83	9,44
124-H	1,87	1,62	1,70	0,00	1,95	2,84	11,95	8,69	8,69	8,53
239-H	1,97	2,02	2,16	1,95	0,00	4,07	12,41	8,48	8,48	8,89
477GATO	3,13	4,49	3,88	2,84	4,07	0,00	11,76	8,84	8,84	9,61
402	14,38	12,79	13,13	11,95	12,41	11,76	0,00	8,27	8,27	6,00
H1	9,66	9,29	8,83	8,69	8,48	8,84	8,27	0,00	0,00	2,90
H4	9,66	9,29	8,83	8,69	8,48	8,84	8,27	0,00	0,00	2,90
196	10,44	9,33	9,44	8,53	8,89	9,61	6,00	2,90	2,90	0,00

Assim, foram construídos dendrogramas a partir destas matrizes de distância, utilizando-se, para efeito de comparação de medidas, o método da Ligação Simples:

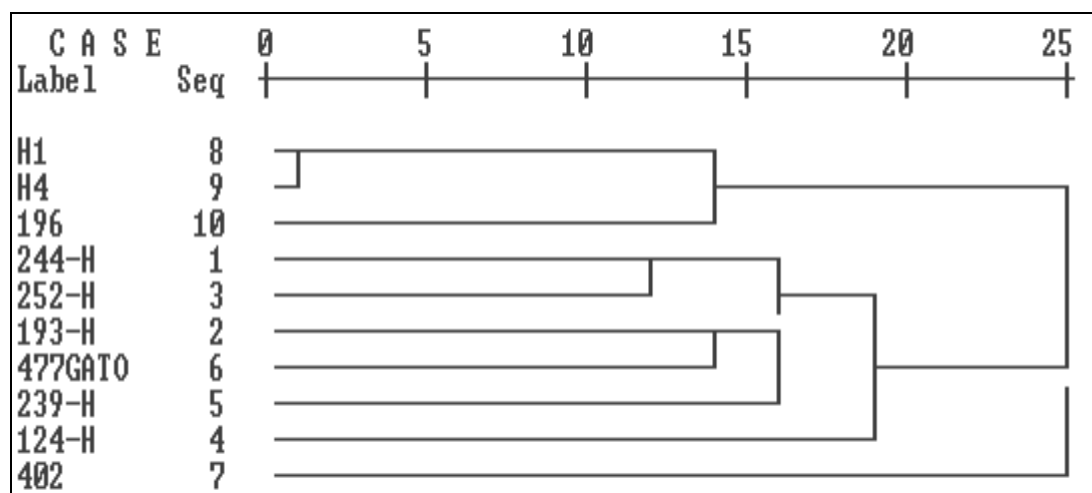
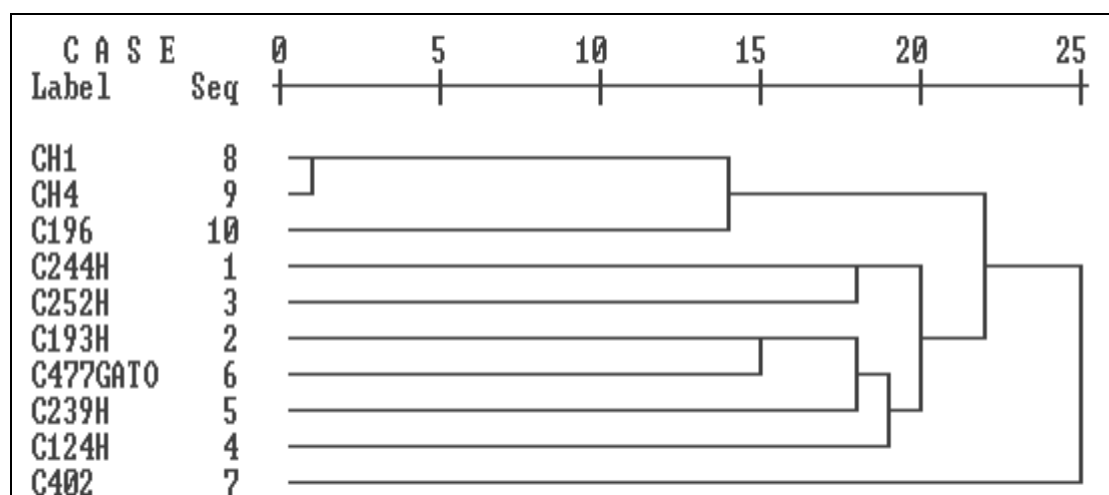
Gráfico C.1 - Dendrograma para a distância euclidiana**Gráfico C.2** - Dendrograma para a distância de Jaccard

Gráfico C.3 - Dendrograma para a distância alternativa