

Chapter 2

DNA Marker Analysis

Genetic markers are polymorphic genetic sequences, like RFLPs or microsatellites, that differ within chromosomal alleles. Rather than analysing the sequence directly, this gene is inferred through analysis of a genetic marker. Marker analysis approach is quite helpful in population biology and ecology studies that can be trace patterns in populations like plants, animals, humans, etc.

2.1 Genetic Analysis Using NTSYSpc (Numerical Taxonomy System)

Introduction

Numerical taxonomy deals with grouping taxonomic units by numerical methods, a concept first developed by Sokal and Sneath in 1963, divided the field into **phenetics** and **cladistics** in which the former was based on the patterns of overall similarities and in the latter classifications were based on the branching patterns of the estimated evolutionary history of the taxa.

NTSYSpc generate phenogram based on characters by performing various types of agglomerative cluster analysis with similarity or dissimilarity matrix like standardizing data matrix and computing distance coefficients among the columns of the matrix, using UPGMA to cluster the distance matrix, computing cophenetic-value (ultrametric) matrix and also computing the cophenetic correlation as a measure of goodness of fit, and then finally plotting the results in the form of a phenogram and generating the distance matrix also as an output. Matrix which is the actual input can be generated based on the presence (1) or absence (0) of characters of the variables.

Exploration Activity

- 1. Based on the gel picture (Fig. 2.1), identify the distinct bands and prepare a table (Table 2.1). Enter 1 for the presence of band.

Using NTedit

- 2. Click and activate the NTedit program in your computer.
- 3. You can create a table of desired rows and columns by entering appropriate values in ‘No. Rows’ and ‘No. Cols’ (Fig. 2.2).
- 4. You can edit the row label and column label by clicking ‘Row Lab’ and ‘Col Lab’ button and editing accordingly.
- 5. Once the matrix has been prepared, enter the values as you have tabulated earlier. Enter 1 for the presence of band and 0 for the absence of band.

Fig. 2.1 The RAPD gel picture of six different samples

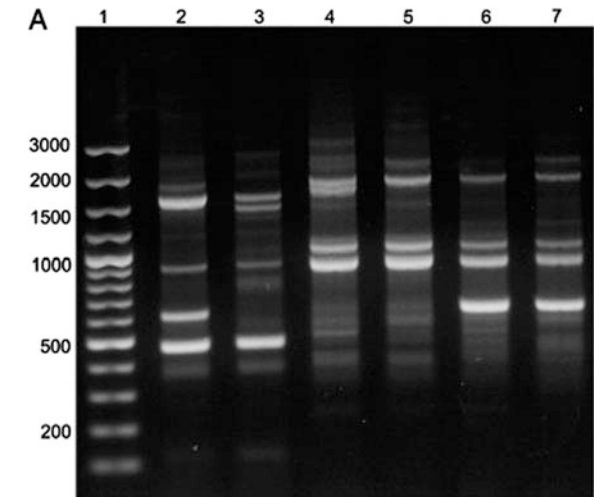


Table 2.1 Table prepared with according to the gel picture where one indicates presence of bands

Marker (MW)	Sp1	Sp2	Sp3	Sp4	Sp5	Sp6
3000			1	1		
2500			1			
2400			1	1		
2000			1	1	1	1
1600	1	1	1			
1500		1	1	1		
1000			1	1	1	1
700					1	1
600	1					
500	1	1				
400	1	1				
120		1				

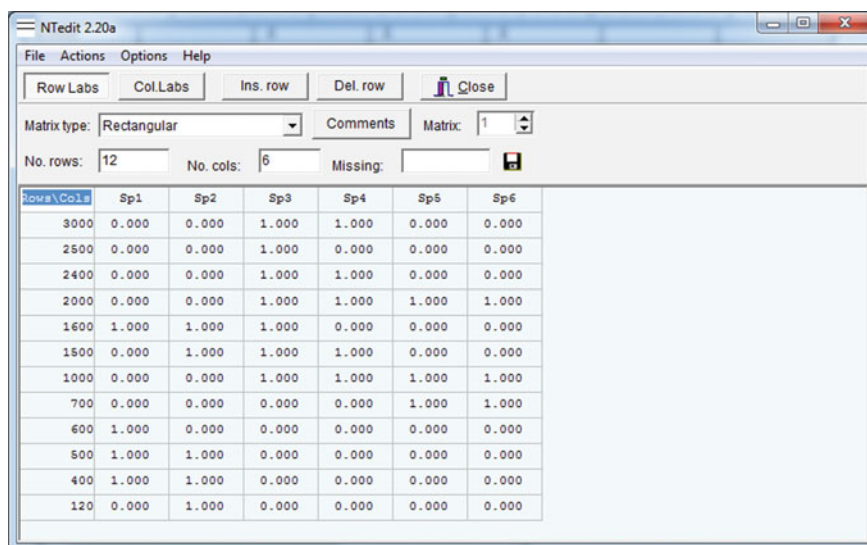


Fig. 2.2 The NTedit program for preparing data

6. Save the file by choosing **'Save'** option from **'File'** in the menu bar or by clicking the floppy icon (data.NTS).

Using NTSYSpc

7. Now click and activate the NTSYSpc program from your computer.
8. The options and the menus of the program have been listed on the left side of the program window (Fig. 2.3).
9. Click on the **'Transformation'** option (Fig. 2.4) and select **'Standardization'** (Stand).
10. Specify the file which you prepared and saved earlier using NTedit as input file (Data.NTS). And, specify a name for the output file (Stand.NTS).
11. Click **'Compute'** button.
12. The program generates a **'Report listing'** window (Fig. 2.5) whose contents can be copied and saved for future requirements, if necessary.
13. You can open the output file to visualize the distance matrix (Fig. 2.6).

Constructing Dissimilarity Tree

14. Click on the **'Dis/similarity'** menu and select **'Genetic Distance'** (Simgend) option. Specify the output file you generated earlier in Standardization step as input file (Stand.NTS). Then, specify a new output file (simgend_out.NTS) (Fig. 2.7).
15. Click **'Compute'** button.

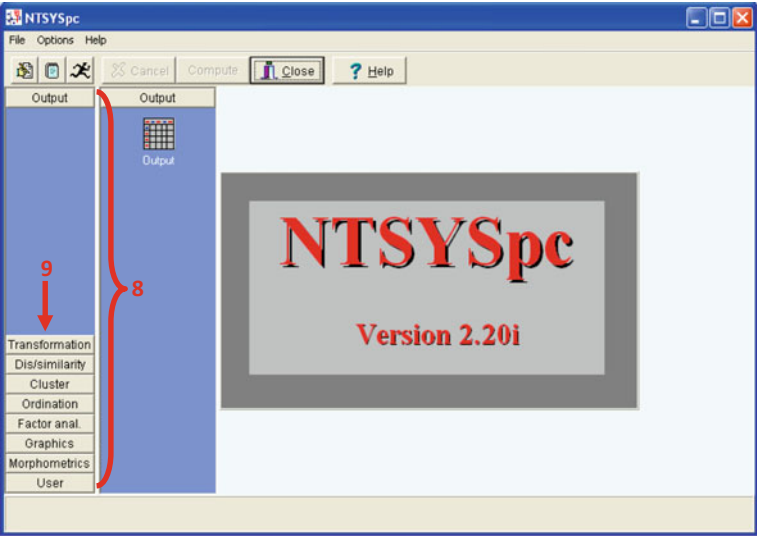


Fig. 2.3 The NTSYSpc program interface

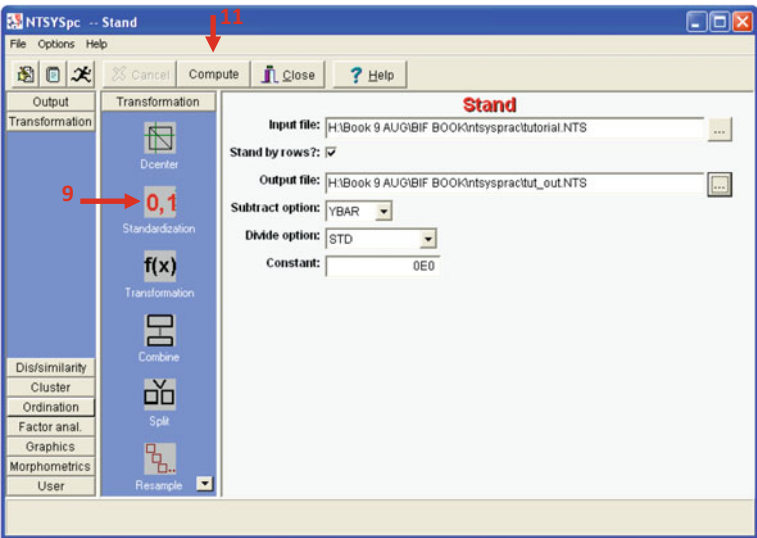


Fig. 2.4 Performing standardization under transformation

- 16. The program generates a **‘Report listing’** tab whose contents can be copied and saved for future requirements, if necessary.
- 17. You can open the output file to visualize the distance matrix (Fig. 2.8).

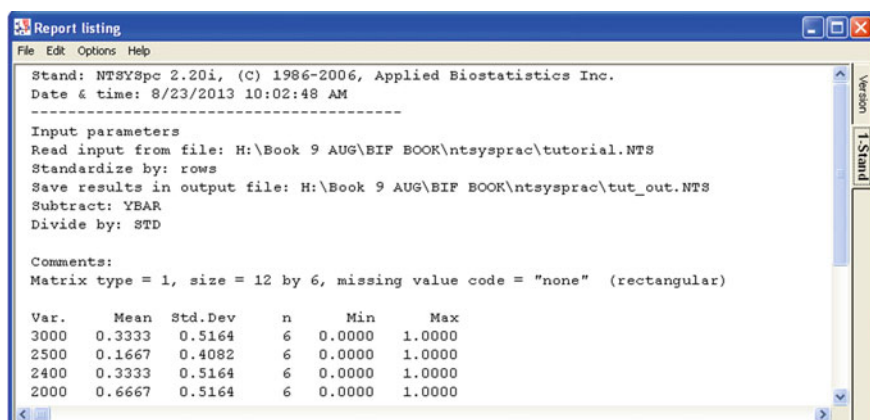


Fig. 2.5 The report listing window

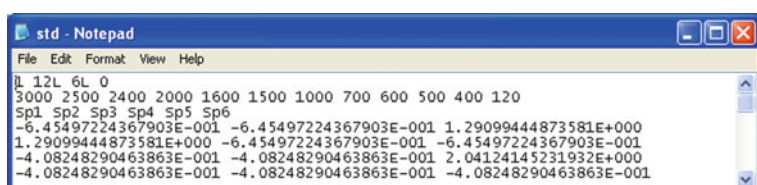


Fig. 2.6 The content of output file displaying the distance matrix

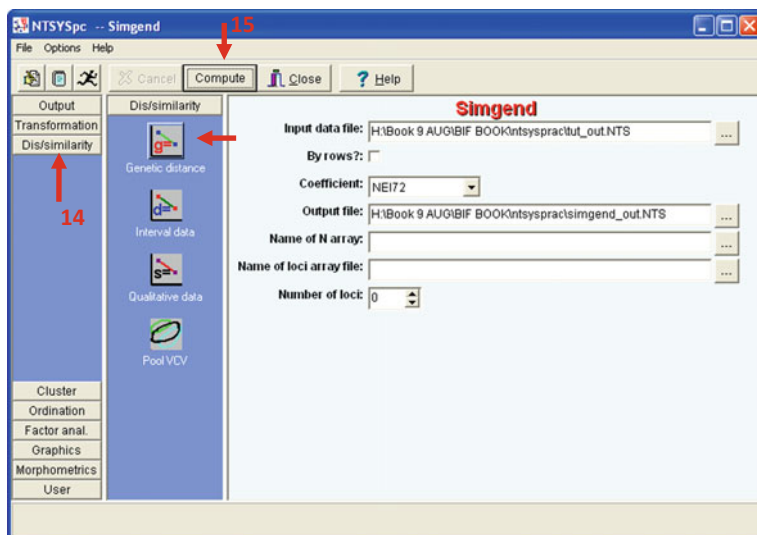
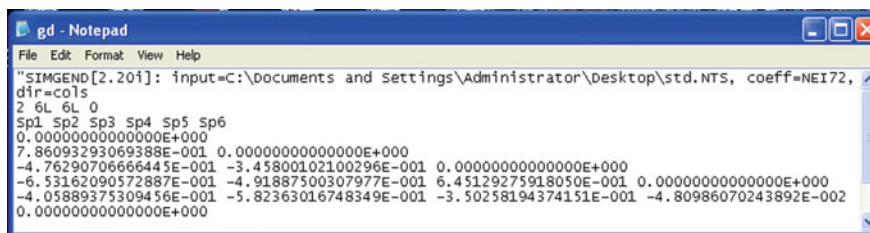


Fig. 2.7 Calculating genetic distance from standardized data



```

"SIMGEND[2.201]: input=C:\Documents and Settings\Administrator\Desktop\std.NTS, coeff=NEI72,
dir=cols
2 6L 6L 0
Sp1 Sp2 Sp3 Sp4 Sp5 Sp6
0.000000000000000E+000
7.86093293069388E-001 0.000000000000000E+000
-4.76290706666445E-001 -3.45800102100296E-001 0.000000000000000E+000
-6.53162090572887E-001 -4.91887500307977E-001 6.45129275918050E-001 0.000000000000000E+000
-4.05889375309456E-001 -5.82363016748349E-001 -3.50258194374151E-001 -4.80986070243892E-002
0.000000000000000E+000

```

Fig. 2.8 Simgend report

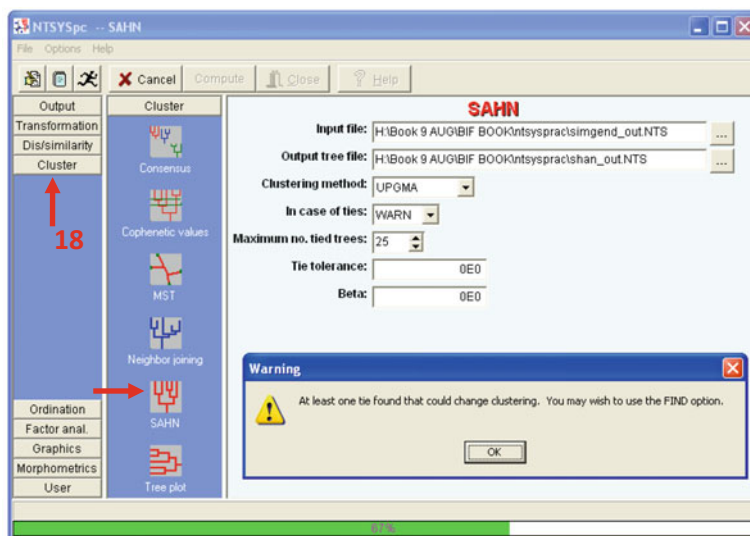


Fig. 2.9 Preparing SAHN tree

18. Click '**Cluster**' from the menu and select '**SAHN**' from the option (Fig. 2.9). Specify the output file you generated earlier in Genetic distance step as input file (simgend_out.NTS). Then, specify a new output file (sahn_out.NTS).
19. Click '**OK**' on the warning popup generated to proceed processing. You can see an icon at left corner which can be clicked to view the generated tree (Fig. 2.10).
20. The tree is displayed on a new window '**Tree plot**' (Fig. 2.11).
21. The display options like flipping the tree, changing the font or thickness of the line can be changed by clicking the '**Options**' from the menu bar and selecting '**Plot options...**' (Fig. 2.12).
22. You can view the UPGMA file content (sahn_out.NTS) by opening the output file (Fig. 2.13).

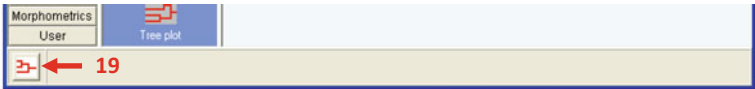


Fig. 2.10 The button to display the generated tree

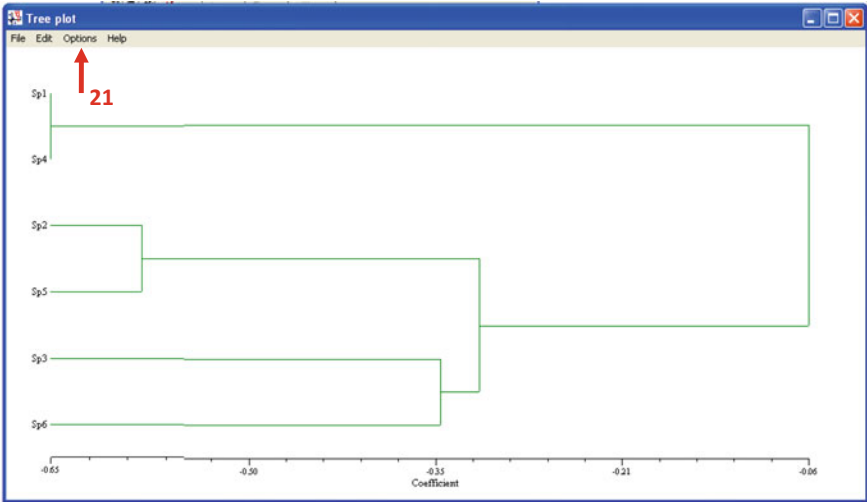
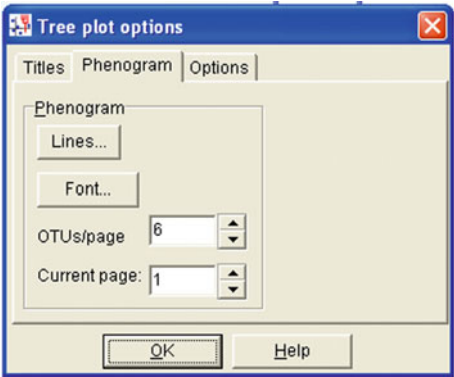


Fig. 2.11 The dissimilarity tree

Fig. 2.12 The tree plot options



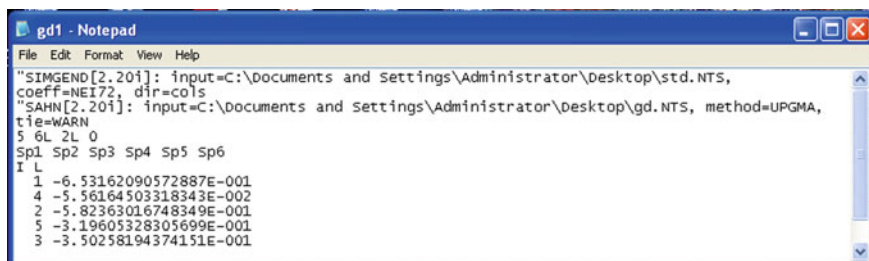


Fig. 2.13 UPGMA tree generated by SAHN

Constructing Similarity Tree

23. Click on the '**Dis/similarity**' menu and select '**Interval data**' (Simint) or '**Qualitative data**' (Simqual) option. Specify the output file you generated earlier in Standardization step as input file (stand.NTS). Then, specify a new output file (simint_out.NTS or simqual_out.NTS).
24. Click '**Compute**' button.
25. The program generates a '**Report listing**' tab and an output file whose contents can be copied and saved for future requirements, if necessary.
26. Click '**Cluster**' from the menu and select '**SHAN**' from the option to generate tree.

Exercise

1. What are the applications of NTSYSpc?
2. Define the following terms:
 1. Ordination
 2. Factorial analysis
 3. Cluster analysis
 4. SAHN
 5. UPGMA
 6. Neighbour Joining

2.2 Principal Coordinate Analysis (PCOORDA) Using NTSYSpc

Principal component analysis (PCA) is a mathematical procedure invented in 1901 by Karl Pearson and later developed by Harold Hotelling in the 1930s. This method used orthogonal transformation for converting a set of possibly correlated variables into a set of linearly uncorrelated variables named principal components.

PCA can be done by Eigen value decomposition of a data matrix, usually after mean centring (and normalizing or using Z-scores) the data matrix for each attribute. The results of a PCA are usually discussed in terms of Eigen value. Largest Eigen values usually correspond to the principal components that are associated with most of the co-variability among a number of observed data.

PCOORDA is an alternative to PCA which will give the same results as PCA. When there are fewer points than variables, the computation time will be much less than for the PCA. The data matrix is standardized by variables (rows), a matrix of distances between the objects is computed, the double-centred distance matrix is then factored and a plot is drawn showing the objects in a three-dimensional space.

1. Now click and activate the NTSYSpc program from your computer.
2. The options and the menus of the program have been listed on the left side of the program window (Fig. 2.14).

Preparation of Similarity Matrix Using Qualitative Data

3. Click on the '**Dis/similarity**' option and select '**Qualitative data**' (Simqual).
4. Specify the file which you prepared and saved earlier using NTedit as input file (**Data**). And for the **output file** specify a name (**DJ**) (Fig. 2.15).
5. Choose '**J**' under '**Coefficient**'.
6. Leave other options as default (Positive code 1 and Negative code 0).
7. Click '**Compute**' button.

Preparation of Double-Centred Matrix

8. Click '**Transformation**' and select '**Dcenter**' (Fig. 2.16).

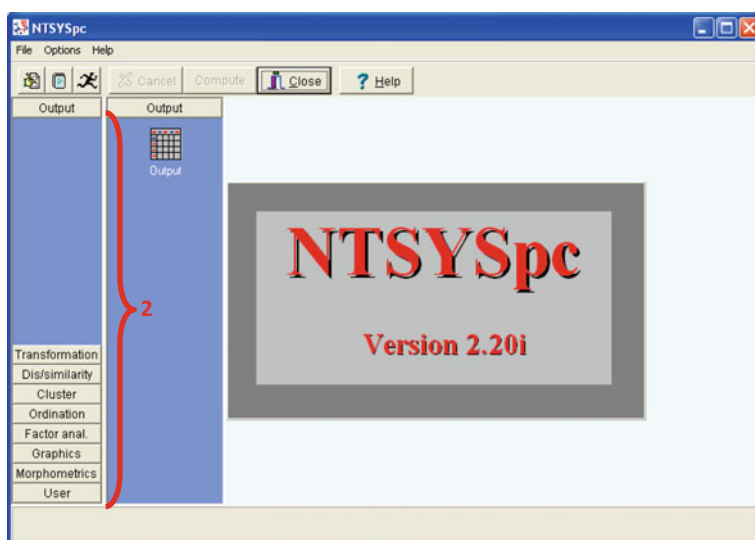


Fig. 2.14 The NTSYSpc program interface

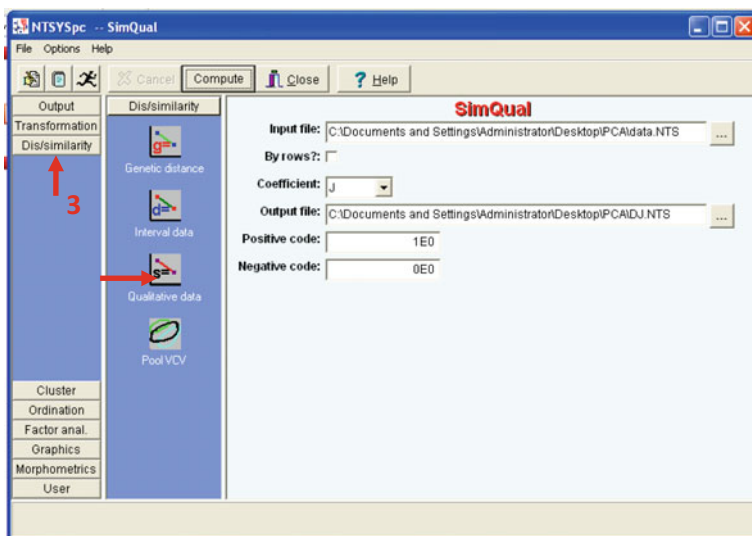


Fig. 2.15 Calculating similarity matrix using qualitative data

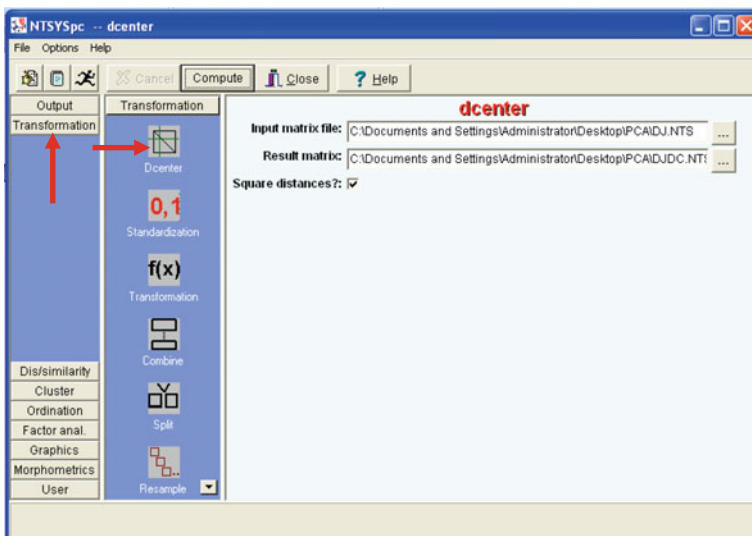


Fig. 2.16 Calculating Double-centred matrix

9. Specify the file which you prepared and saved earlier (**DJ**) as Input matrix file. And for the **Result matrix** (output file) specify a name (**DJDC**).
10. Click '**Compute**' button.

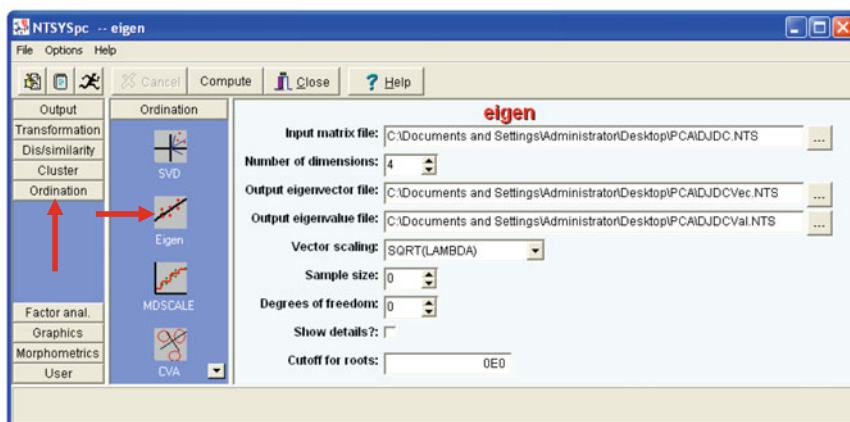


Fig. 2.17 Calculating Eigen values

Calculation of Eigen Value

11. Click '**Ordination**' and select '**Eigen**' (Fig. 2.17).
12. Specify the file which you prepared and saved earlier (**DJDC**) as input matrix file.
13. Leave **no. of dimensions** as default (4).
14. For the **output Eigenvector file** specify a name (**DJDCVec**) and for **output Eigenvalue file** a name (**DJDCVal**).
15. Leave other options as default (sample size and degrees of freedom = 0).
16. Click '**Compute**' button.
17. The Eigen values are displayed on the '**Report listing**' window (Fig. 2.18).

Plotting Coordinate Graph

18. Click '**Graphics**' and select '**Mod 3D Plot**' (Fig. 2.19).
19. Specify the file which you prepared and saved earlier (DJDCVec) as input file. Deselect '**Plot by rows**'.

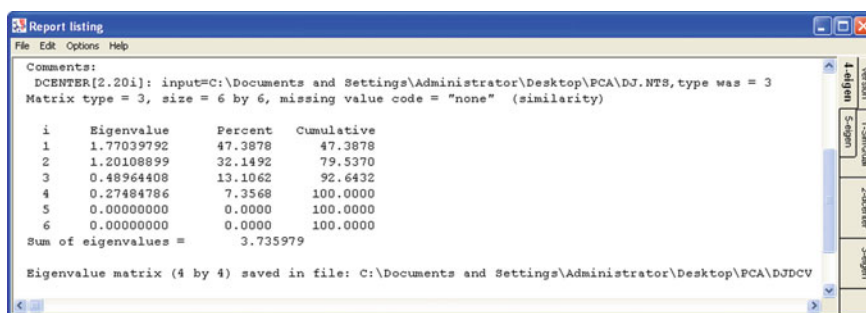


Fig. 2.18 Eigen values displayed on the report listing window

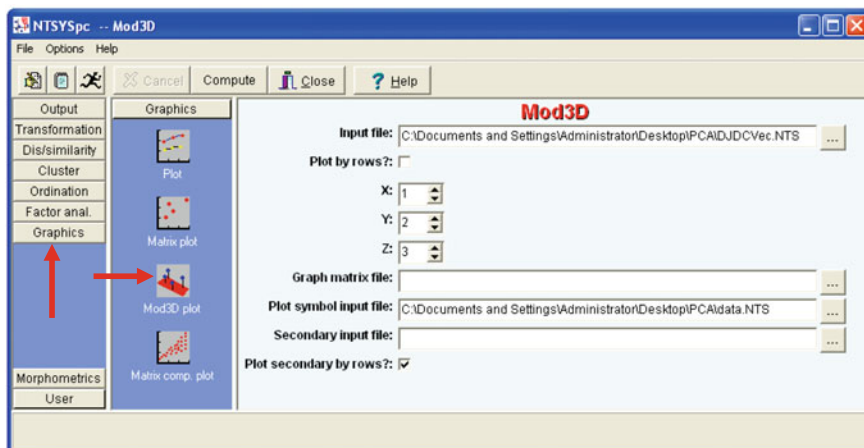


Fig. 2.19 Preparing plot

20. For **Plot symbol input file** (dominant symbol), specify the matrix file you created with NTedit (Data).
21. Leave '**Graph matrix file**' as blank.
22. Click '**Compute**' button.
23. The graph is displayed by which can be rotated by mouse. The display options can be changed by changing the options in '**Plot options**' (Fig. 2.20).

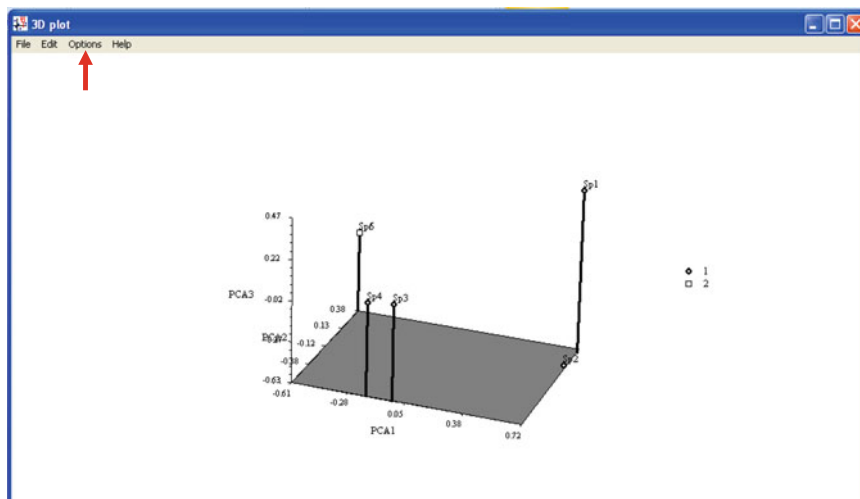


Fig. 2.20 The plot in 3D. *Diamond* symbol indicates one species (1, 2, 3 and 4) and the *square* symbol indicates 2 species (5 and 6)

Exercise

1. What do you mean by multivariate analysis and PCA?
2. How PCA is different from PCOORDA?
3. Define the following
 - A. Eigen value
 - B. PCA 1 and PCA 2
 - C. SVD.

2.3 Population Genetic Analysis Using PowerMarker

Introduction

PowerMarker is a comprehensive collection of both new and traditional statistical methods that can be employed for population genetic analysis which is basically designed for SSR/SNP data analysis (Liu and Muse 2005).

Preparation of Matrix Data Based on Banding Pattern

1. Open a new Excel (xls) file and create a data file containing **1 for the presence of band, 0 for the absence of band** and **9 for missing data** (Fig. 2.21).
2. Find and replace all 0 to **Z**, 1 to **Y** and 9 to **Q** (Fig. 2.22).
3. Copy the whole table and paste in a Notepad (Fig. 2.23). Now replace all the Z to **0/0**, Y to **1/1** and Q to **9/9** and save the file (Data.txt).

	A	B	C	D	E	F	G	H	
1	popn	accession	p1000	p800	p600	p400	p200	p100	
2	p1	A1	1	0	1	0	9	1	
3	p1	A2	1	1	0	0	0	0	
4	p1	A3	1	0	0	0	1	0	
5	p1	A4	1	1	0	1	1	1	
6	p1	A5	1	0	1	1	1	1	
7	p2	A6	0	0	1	1	0	1	
8	p2	A7	9	1	1	1	0	9	
9	p2	A8	1	0	0	1	1	1	
10	p2	A9	0	1	0	0	1	0	
11	p2	A10	0	0	1	0	0	0	
12	p3	A11	0	1	0	9	0	0	
13	p3	A12	1	0	9	0	1	1	
14	p3	A13	0	0	9	1	1	1	
15	p3	A14	1	0	0	1	0	1	
16	p3	A15	1	9	1	1	9	1	
17									

Fig. 2.21 Preparing data on excel sheet

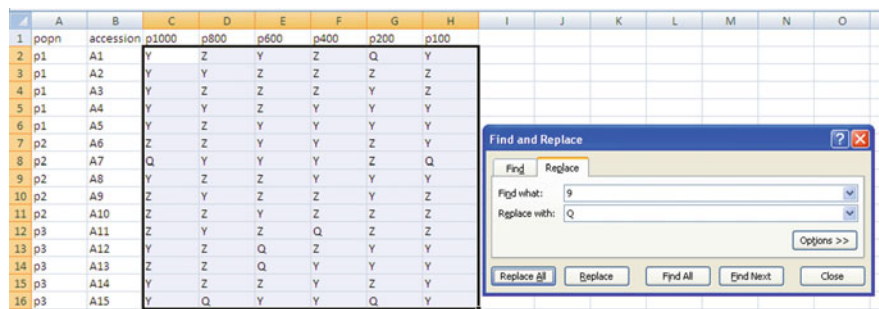


Fig. 2.22 Finding and replacing 0's and 1's

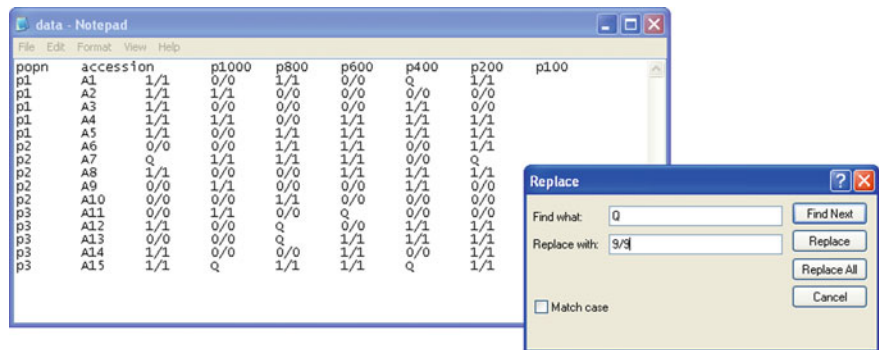


Fig. 2.23 Modifying data on a notepad

Using Power Marker

1. Click and initiate PowerMarker program from your computer.
2. Choose 'File' and select 'Close all Projects' from the menu to close all the default projects.
3. Create a new project by choosing 'File' and selecting 'New Project' by giving a name (PMTut).
4. Click 'File' and select Import and Dataset. A dataset import wizard window now opens. Click on the browse and select the 'Data.txt' you saved earlier. Click 'Next' button (Fig. 2.24).
5. Change the 'popn' and 'accession' to categorical type by clicking the column item and clicking the 'Categorical' link at the bottom of the table (Fig. 2.25).
6. Then under 'Hierarchy', change the Level-1 to 'accession' and level-2 to 'popn' by clicking on the dropdown menu and selecting. Click 'Next' button (Fig. 2.25).
7. Change the 'Missing allele' and 'Missing category' values to 9. Click 'Next' button (Fig. 2.26).

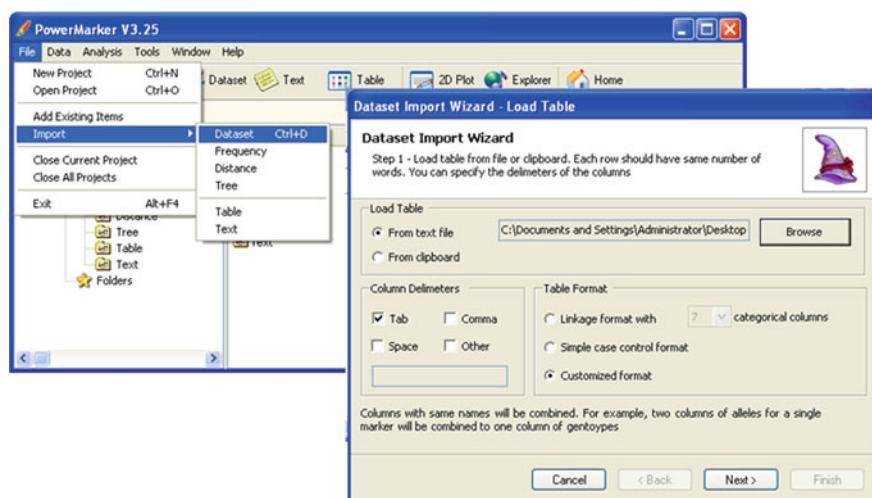


Fig. 2.24 Importing dataset into PowerMarker

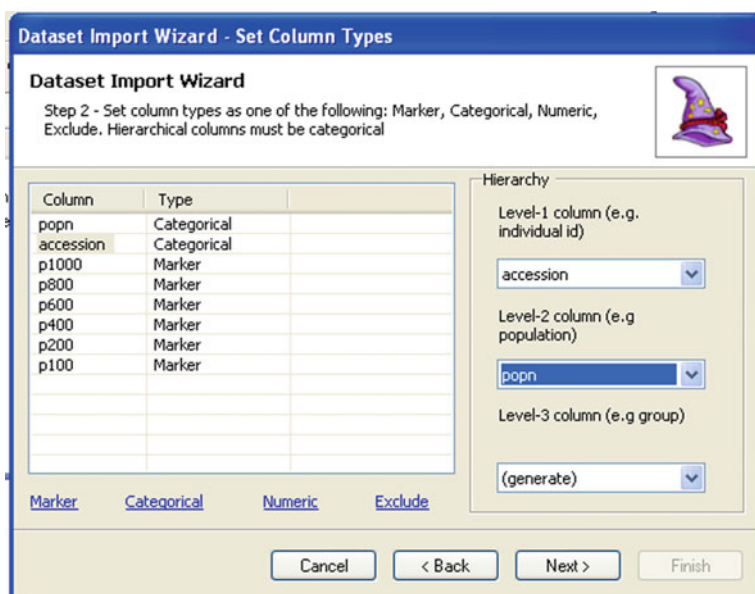


Fig. 2.25 Choosing options in the dataset import wizard

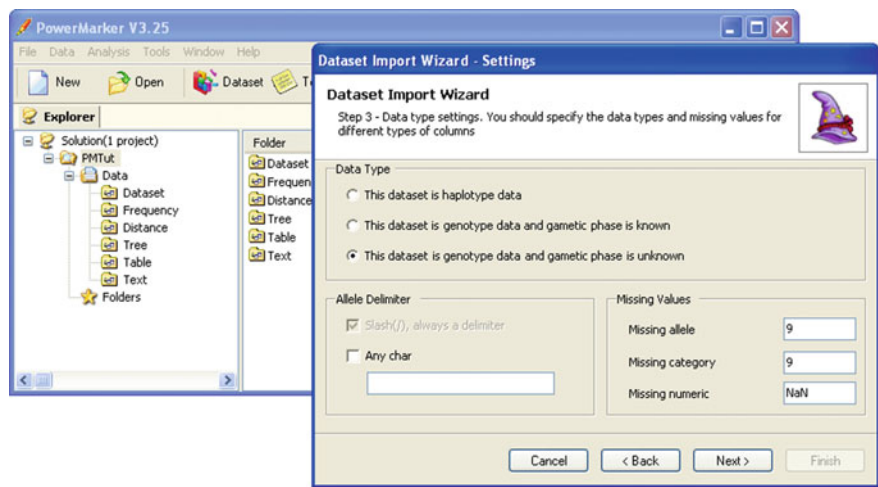


Fig. 2.26 Specifying the data types and missing values

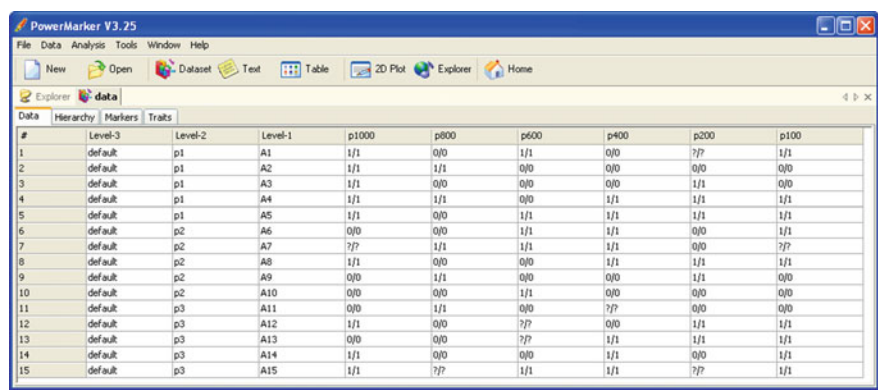


Fig. 2.27 The final matrix

8. It will display a final matrix that is ready to be imported to the project. Click ‘Finish’ button (Fig. 2.27).

Statistical Analysis

9. Click on the ‘Analysis’ from the main menu to see the list of analyses that can be performed for the provided data (Fig. 2.28). The type of analyses that can be performed is listed in Table 2.2.
10. Important features to look in ‘Summary’ (Table 2.2).
11. Important features (Fig. 2.29) to look in ‘Structure’ (Table 2.3).
12. The list of analyses that can be done with PowerMarker are provided in the developer’s website.

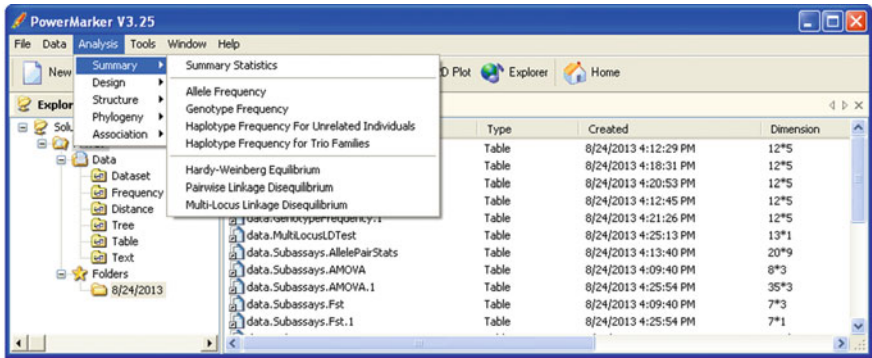


Fig. 2.28 The Analysis options in PowerMarker

Table 2.2 Options under ‘summary statistics’

Summary statistics	Major allele, Gene diversity, Heterozygosity, PIC
Allele frequency	Marker, variance
Genotype frequency	Covariance, marker
Hardy-Weinberg equilibrium	Chi square, p value
Pairwise linkage disequilibrium	Multi allelic D and N allelic D

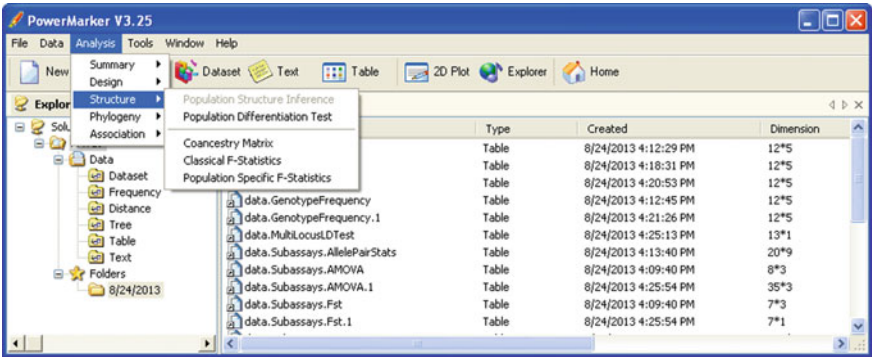


Fig. 2.29 The structure options under analysis

Table 2.3 Classical F statistics

Classical F-statistics	Select ‘ANOVA table’
Population specific F-statistics	Mean SD values of beta, alpha and covariance

Exercise

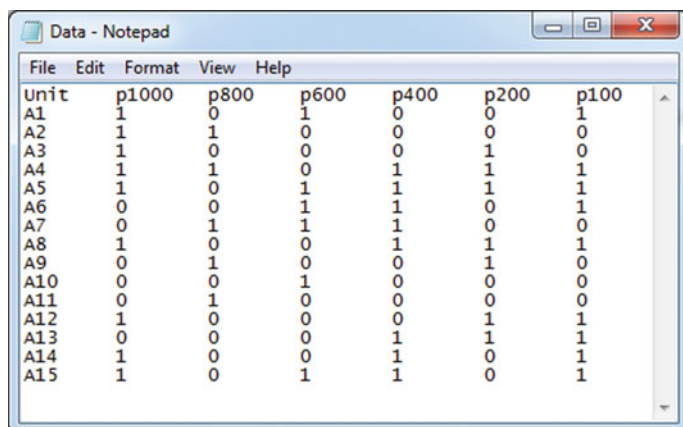
1. Define the following
 1. Covariance
 2. Heterozygosity
 3. PIC
 4. Hardy Weinberg equilibrium
 5. Exact Test
 6. Mantel Test
2. Explain the basic steps involved in AMOVA and state their applications.

2.4 Dissimilarity Analysis—DARwin5 (Dissimilarity Analysis and Representation for Windows)

DARwin offers both traditional and more original approaches and focuses mainly on diversity structure description which is based on the distance methods.

Exploration Activity

1. DARwin program can be downloaded from <http://darwin.cirad.fr/Download.php> for free after registration.
2. Prepare the data in excel sheet and save it in a notepad (data.txt) (Fig. 2.30). Type Unit in column A. Row may be accession, species or variety. Each column represents the marker/bands.
3. Activate the Darwin5 program from your computer (Fig. 2.31).



unit	p1000	p800	p600	p400	p200	p100
A1	1	0	1	0	0	1
A2	1	1	0	0	0	0
A3	1	0	0	0	1	0
A4	1	1	0	1	1	1
A5	1	0	1	1	1	1
A6	0	0	1	1	0	1
A7	0	1	1	1	0	0
A8	1	0	0	1	1	1
A9	0	1	0	0	1	0
A10	0	0	1	0	0	0
A11	0	1	0	0	0	0
A12	1	0	0	0	1	1
A13	0	0	0	1	1	1
A14	1	0	0	1	0	1
A15	1	0	1	1	0	1

Fig. 2.30 Preparation of data



Fig. 2.31 The DARwin5 interface

4. Click on '**File**' and select '**Import Data Matrix**'. Locate the data file you prepared earlier in notepad (data.txt). You can now see the name on the imported file under the button '**Txt File to Import**'.
5. Then click on the '**Save Data as...**' button and save the file (data.var). Once done you can see the file name under the button '**Save Data as...**'.
6. Click the check box next to 'View file when done' to see the file.
7. Click the '**OK**' button.
8. Click on the '**Dissimilarity**' option on the main menu and select '**Calculate from Single Data**'. Now locate the var file you saved earlier (Data.var) (Fig. 2.32).
9. In case there are missing data in your data file you can click the '**Options**' button in the '**Missing Data**' column, tick the '**Some data are missing**' and under '**Pairwise variable deletion**' choose 50%. Click '**OK**' (Fig. 2.33).
10. If you don't have any missing data, you can skip the earlier step.
11. Under '**Dissimilarity index**' choose '**Jaccard**'.
12. Click '**Save dissimilarity as...**' button on the top to save (Data.dis).
13. Click the '**OK**' at bottom of the program window to execute the program.

Factorial Analysis

Principal Coordinate analysis produces Euclidean plans which preserve at best the distances between units through graphical representations.

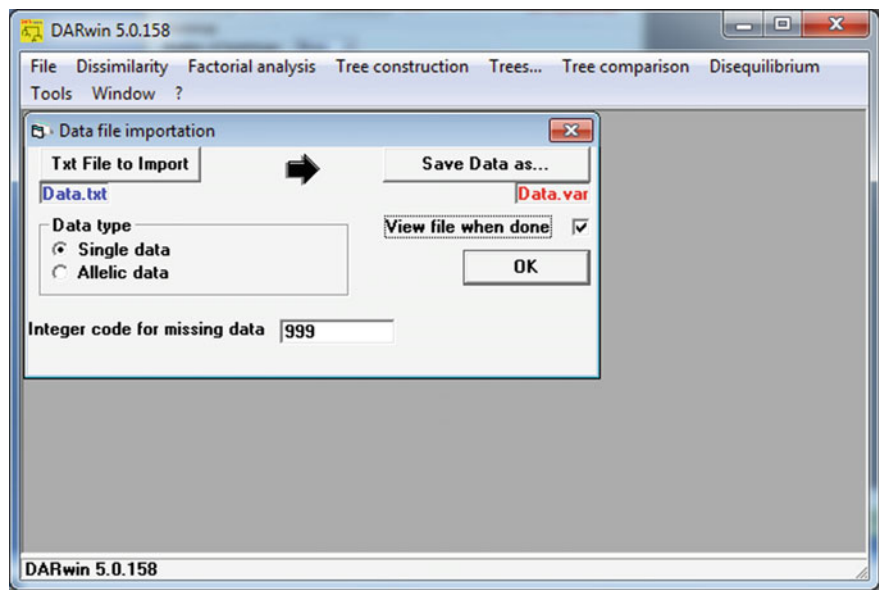


Fig. 2.32 Importing the data file and converting it to var file

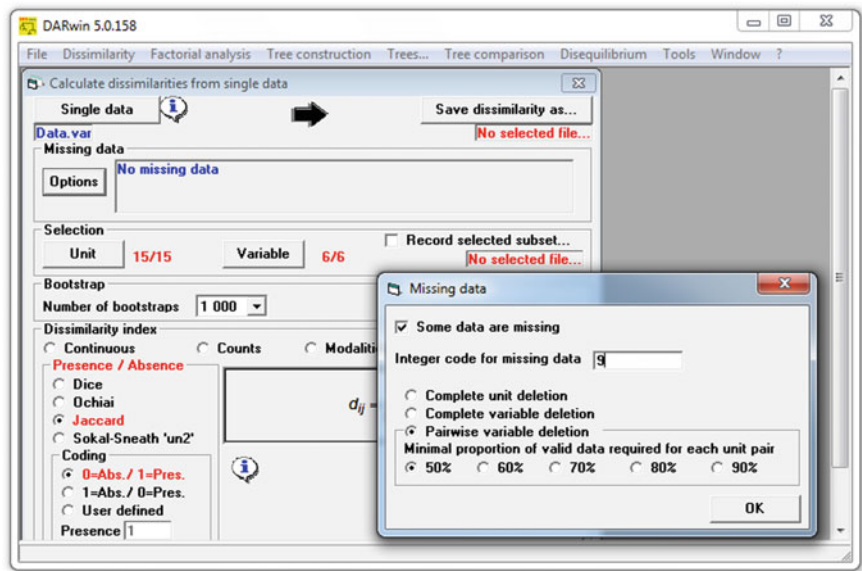


Fig. 2.33 Preparing dissimilarity file

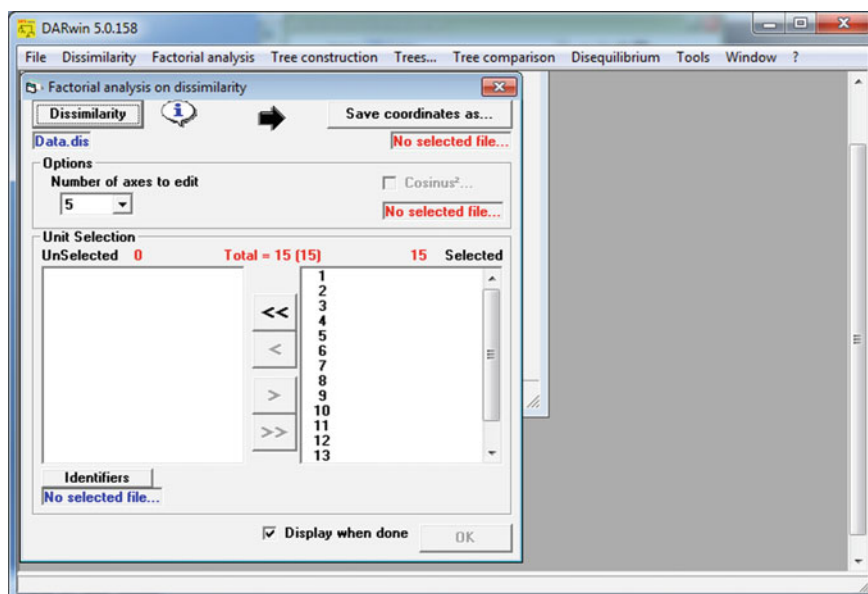


Fig. 2.34 Performing factorial analysis

14. Click on the '**Factorial analysis**' option on the main menu and select '**Analysis**'. Now locate the dis file you saved earlier (Data.dis).
15. Once loaded, now click on the '**Save coordinates as...**' button on the top of the program interface and save the file Factorial coordinate file (Data.AFT). Click '**OK**' button at the bottom to execute the procedure (Fig. 2.34).
16. The program now displays the plot and a report on the Factorial analysis done (Fig. 2.35).
17. To save the report as text file (FA.txt) click on the floppy disk icon on the '**AFTD sur**' window.

Tree Construction by Neighbour-Joining Method

Tree construction can be done by methods like aggregation criteria, Neighbour Joining or by Scores. The first two methods can be either weighted or unweighted. Methods like NJTree and Scores produce more reliable trees by attempting to reduce sensibility to data error. More over bootstrapping, a statistical measure, in NJTree can be employed to estimate how well the tree is supported by the data.

18. Click on the '**Tree construction**' option on the main menu and select '**Neighbour joining**'. Now locate the dis file you saved earlier (Data.dis). Check the '**Bootstrap analysis**' and '**Display when done**' options at the bottom (Fig. 2.36).

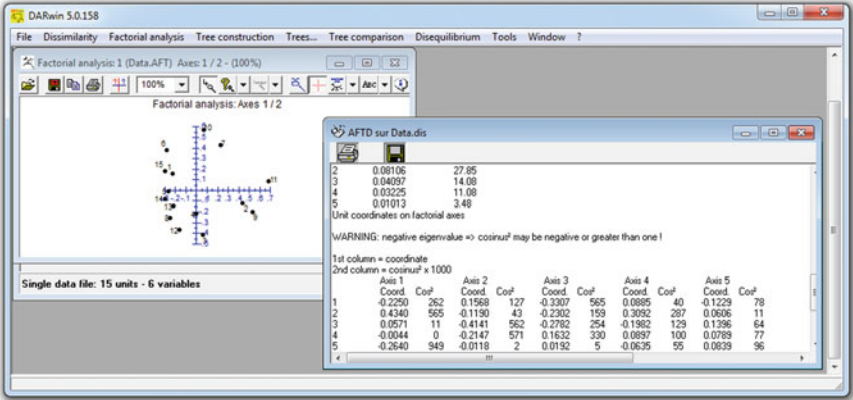


Fig. 2.35 The output from factorial analysis

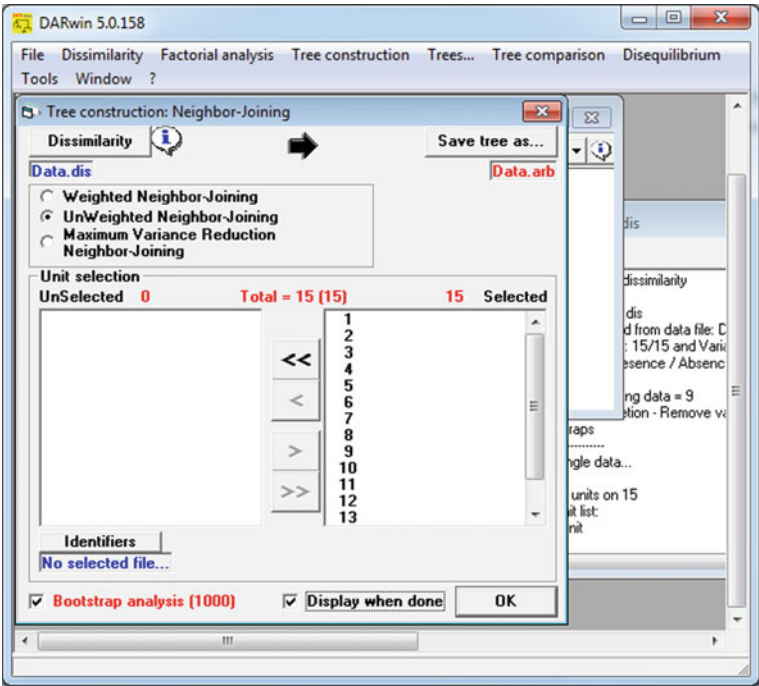


Fig. 2.36 Tree construction using neighbour-joining method

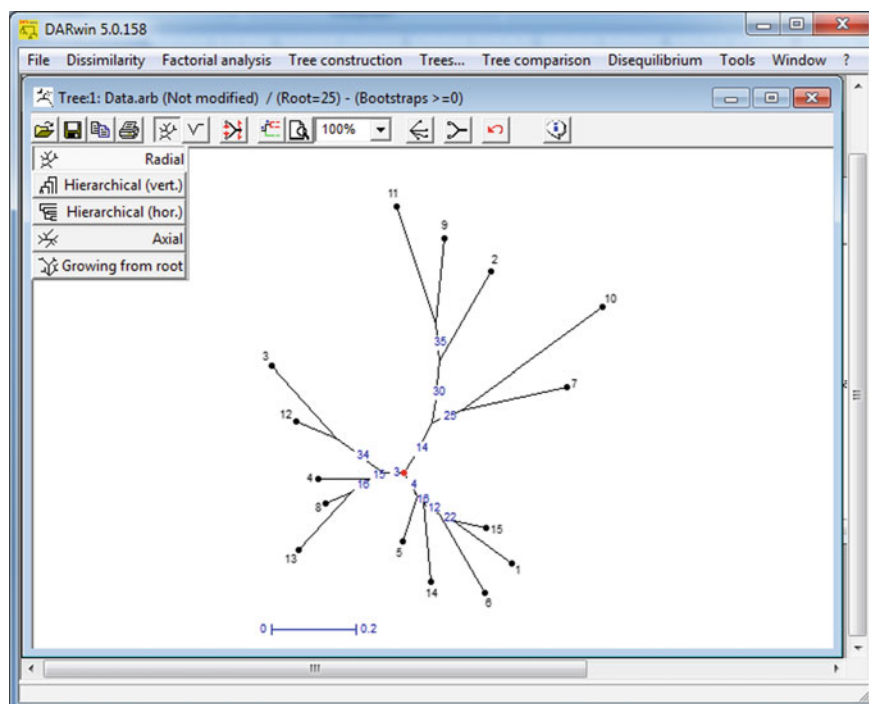


Fig. 2.37 The neighbour-joining tree

19. Select '**Unweighted Neighbour joining**' from the options and click '**Save tree as...**' button on the top of the program interface to save the tree file (Data.arb).
20. By default radial tree is displayed. The tree can be displayed in other forms by clicking '**Tree representation**' button from the menu bar (Fig. 2.37).

Tree Representation

21. The menu bar provides options to change the display of the tree. For example to specify a root, click on the '**Root selection tool**' and specify or to edit the tree, click on the '**Edition**' button.
22. To save the final tree, click on the floppy disk icon on the menu.

Exercise

1. What do you mean by dissimilarity analysis?
2. What do you mean by weighted and unweighted Neighbour-Joining method?
3. Define Jaccard Index of dissimilarity.

Bioinformatics - A Student's Companion

Syed Ibrahim, K.; Gurusubramanian, G.; Zothansanga;
Yadav, R.P.; Senthil Kumar, N.; Pandian, S.K.; Borah, P.;
Mohan, S.

2017, XV, 283 p. 333 illus., Hardcover

ISBN: 978-981-10-1856-5