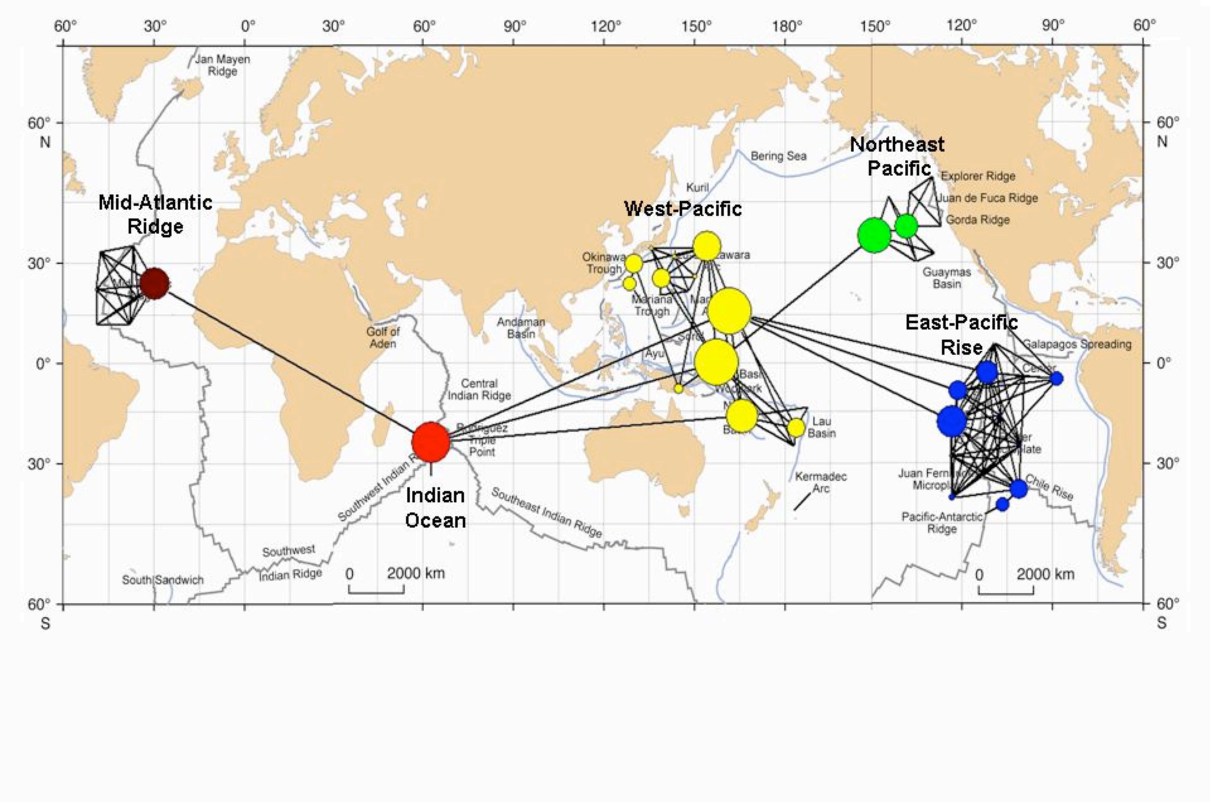
EDENetwork :
Ecological and Evolutionary Networks



Mikko Kivelä[1], Sophie Arnaund-Haond[2], Jari Saramäki[1]

[1] Complex Networks group ; Department of Biomedical Engineering and Computational Science Aalto, University School of Science and Technology, Helsinki, Finland
[2] Ifremer- Deep, Centre de Brest, France & Lab. of Marine Ecology and Evolution, Ccmar- Univ. Do Algarve, Faro, Portugal

EDENetwork manual 0.1

25/05/2010

Table of Contents

1. **Brief overview** 4

1.1. Glossary 4

1.2. Data types handled by EDENetwork 6

1.2.1 Genotype matrix for individual samples: 6
1.2.2 Distance matrix: 7
1.2.3 Network data: 7

1.3. System requirements 7

1.4. Installing and uninstalling EDENetwork 7

1.5. How to download and cite EDENetwork 7

1.6. Acknowledgements 7

2. **Getting started** 8

2.1.1 Workflow in a nutshell 8

2.2. Infiles 9

2.2.1 Individual-centred analysis 9
2.2.2 Population-level analysis 10
2.2.3 Distance matrix 11
2.2.4 Network data 11

2.3. Interface 13

2.3.1 Begin a new analysis project 13
2.3.2 Genotype matrix : Individual centred or location based 13
2.3.3 Distance matrix, network data 14

2.4. Data analysis 15

2.4.1 Distance distribution 15
2.4.2 Schematic overview of data input and analysis 16
2.4.3 Deriving, Analyzing and Visualizing Networks 17
2.4.4 Analyzing network structure 18
2.4.5 Visualizing networks 19
2.4.6 Randomizations 20
2.4.7 Outfiles 21

3. **Methodological outline: Details and references for the calculations** 22

3.1. Genetic distances implemented 22

3.1.1 Individual centred 22
3.1.2 Population centred 22

3.2. Network descriptors 23

3.2.1 Minimum Spanning Tree 23
3.2.2 Thresholded Network 23
3.2.3 Percolation 23
3.2.4 Clustering 23
3.2.5 Assortativity 24

3.3. Node descriptors 24

3.3.1 Degree 24

3.3.2 Clustering 24

3.3.3 Betweeness centrality 24

4. References 24

1. Brief overview

The goal of EDENetwork is to provide researchers with a set of methods to study population-genetic or ecological data sets in the form of networks built from genetic distance matrices. The built-in network analysis tools allow extracting information on the structure of a system of individuals, populations, or other genetic groups.

EDENetwork has been designed to visualize and analyze networks in order to study genetic relationships in an entire dataset, without *a priori* assumptions on the clustering of individuals, populations or genetic groups. The only underlying assumptions are linked to the genetic distance measure chosen by the user, and hence its careful and accurate selection is therefore crucial.

Network analysis is fairly new but very promising method in ecology and evolution (Bascompte, et al., 2003; Hernández-García, et al., 2007; Proulx, et al., 2005). We recommend a number of articles dealing with this methodology in order to understand both the usefulness and limitations of this holistic approach.
Examples of population genetics analysis based on networks exist at the level of individuals (Becheler, et al., 2010; Hernandez-Garcia, et al., 2006; Rozenfeld, et al., 2007) and populations (Fortuna, et al., 2009; Rozenfeld, et al., 2008). Most applications in population genetics are based on the methodology implemented in EDENetwork, where networks are derived from genetic distance matrices by thresholding out the largest distances. Note that there are alternatives, such as the method based on link correlations proposed by Fortuna et al., which is inspired by earlier population genetics graph analysis methods (Dyer and Nason, 2004). This method may be implemented in a future version.
Ecological networks based on the distance thresholding approach have also been recently proposed to illustrate and analyse relationships between communities (genetic groups) and define biogeographic provinces, based on ecological distances describing their taxonomic composition (Moalic, et al., soumis).

In addition to population-genetic data, EDENetwork allows for studies of any weighted network; however, its set of analysis methods has been selected for their usefulness in genetic analysis.

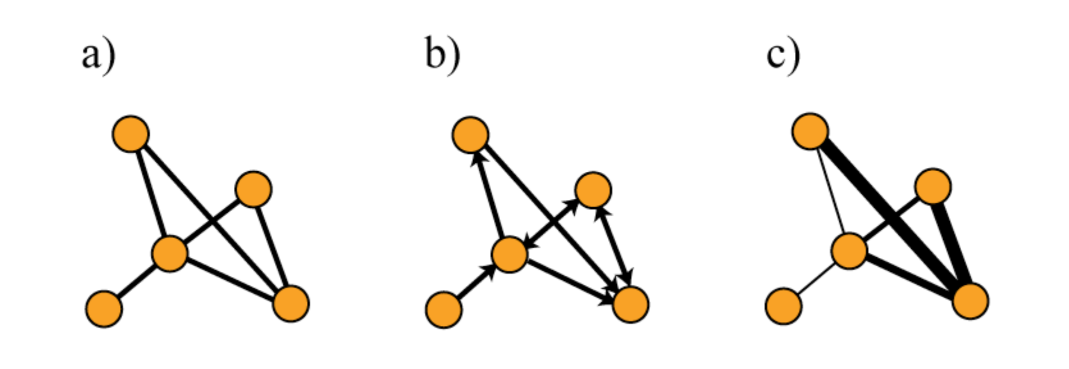
1.1. Glossary

In this section, we provide a brief overview of the terminology and concepts of network analysis. This section is by no means a substitute to complex networks literature, and we encourage the users of EDENetwork to get familiar with the fundamentals of network science. As an introduction we may recommend some reviews on network theory (Albert and Barabasi, 2002; Albert, et al., 2000; Newman, 2003; Watts, 2004; Watts and Strogatz, 1998).

The following glossary may however be useful in order to understand both the manual and analysis methods implemented in the software.

Introductory definitions

Networks consist of **nodes** (or **vertices**) linked by **links** (or **edges)**. The nodes represent the fundamental units of the system, such as individuals or populations, and links represent their interactions or relationships. The strength of such relationships can be taken into account in the form of edge weights; in the case of EDENetwork, a genetic distance is associated with every edge.



Examples of possible networks (a) simple and undirected, (b) directed (c) weighted. EDENetwork deals with networks of type (c).

Properties of individual nodes

Degree (also called **'connectivity degree')**: number of edges connected to a node.
Betweeness Centrality: number of shortest paths between other nodes passing through a node.
Average Nearest-Neighbour Degree: the average degree of the nodes to which the node is connected
Clustering Coefficient: the ratio of existing connections between a node's neighbours to the possible number of such connections.
Component : the component to which a node belongs is the set of nodes that can be reached from it by following the links of the network.

Descriptors of network topology

Degree distribution: probability density distribution of node degrees.
Average shortest path length (or **geodesic distance**): average number of links on the shortest paths between all pairs of nodes.
Clustering coefficient (or **transitivity**)**:** either the network average of the clustering coefficients of individual nodes, or the ratio of interconnected nodes triplets compared to the total of possible triplets in the network. As a high clustering coefficient value indicates non-randomness in the network structure, this index can be used as a measure of substructure (also understood as hierarchical structure in population genetics) in the network.

Diameter: The diameter of a network is the length (number of edges) of the longest path between any two nodes.
Small World: Small-world networks that have a low diameter/average path length and high clustering. Such a topology is typical for most natural of networks, including biological, social, and technological networks, as well as some population-genetic networks (Hernandez-Garcia et al., 2006; Rozenfeld et al., 2007).
Fully connected network: a network in which every node is directly connected to every other node.
Connected network: a network which consists of a single component, i.e. all nodes can reach every other via some path.
Percolation threshold: When links are removed from a connected network, it eventually fragments into small components. The point where this happens is called the percolation threshold. More accurately, this is the point where the so-called giant component (whose size is of the order of the network size) disappears and there is no long-range connectivity; even before the percolation threshold small disconnected fragments will appear, yet a substantial fraction of nodes belongs to the giant component.

1.2. Data types handled by EDENetwork

EDENetwork can handle a wide range of genetic data types. This data can be input at the level of individuals and used to construct genetic distance networks of these individuals or the populations they have been sampled from; in addition, EDENetwork can read pre-calculated genetic distance matrices or network files.

1.2.1 Genotype matrix for individual samples:

For individual samples, the genotypes can be input as

- Allozymes
- Microsatellites
- AFLP
- RFLP

Based on this data, the software can be used to construct and analyze either a network of individual samples, where the nodes represent samples and edges their genetic distances, or a population-level network, if the input data is augmented with population labels for each individual.

For individual centred analysis, the following distance measures can be chosen:

- ✓ the Allele Shared Distance or
- ✓ the Rozenfeld distance (only applicable to microsatellites under assumption of stepwise mutation model)
- ✓ Allele parsimony
- ✓ Hybrid

For population centred analysis, the available distance measure is:

- ✓ the Goldstein Distance

1.2.2 Distance matrix:

Instead of inputting genotypes of every individual, the user can pre-construct a genetic or ecological distance matrix outside EDENetwork, and input this matrix for network analysis. Such a matrix can represent either distances between individuals or distances between populations, such that rows/columns correspond to individuals or populations, respectively. The user is free to choose any distance measure, as long as it yields a pairwise distance matrix.

Examples include

- ✓ Genetic distance: Fst, Rst, Goldstein, Nei…
- ✓ Ecological distance: Jaccard, Bray Curtis, Manhattan…

1.2.3 Network data:

Genetic or ecological distance data may also be input directly in the form of a network, where nodes represent individuals or populations and edges their distances. This network may or may not be pre-thresholded; however, for fully connected networks where an edge links every node, we recommend inputting a distance matrix as above. Network data can either be stored as a GML file (Graph Markup Language, .gml), or an edge file (.edg) where one row denotes an edge (vertex1 vertex2 edge_weight).

1.3. System requirements

EDENetwork can be installed on any computer running Windows. At least 1 GB of memory is recommended for analyzing larger networks. Linux and OS X versions will be published in the near future.

1.4. Installing and uninstalling EDENetwork

Download the EDENetwork .zip package from the address below, and unzip the contents of the package. EDENetwork can now be launched by clicking on the EDEN_launcher_v2.exe icon in the EDENetwork folder. To uninstall, simply delete the entire folder.

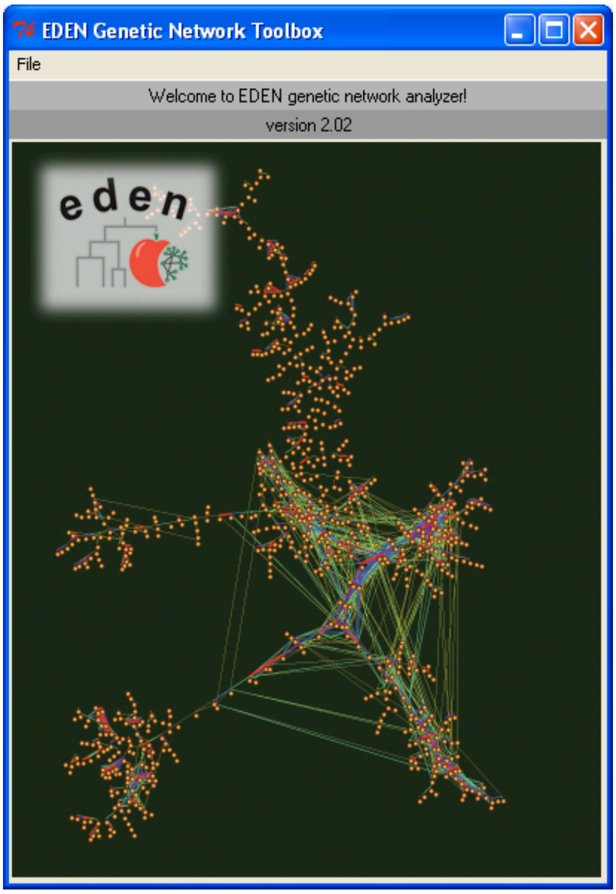
1.5. How to download and cite EDENetwork

The EDENetwork software can be downloaded from the website of project EDEN: http://ifisc.uib.es/EDEN/fichaOutput.php?idOutput=15. Instructions for citing will be published on this website in the near future.

1.6. Acknowledgements

This software had been made possible by the collaborations developed through the project EDEN: Ecological Diversity and Evolutionary Networks (http://ifisc.uib.es/EDEN/). This research project was supported by the New and Emerging Science and Technology programme (NEST) of the 6th Framework Programme of the European Commission, under the NEST-Pathfinder Complexity initiative. It started on 1st January 2007 and finished on December 2010.

2. Getting started



Welcome to EDEN genetic network analyzer!

version 2.02

As for any other software, we recommend to read this manual before using EDENetwork – it has been designed to provide you with all necessary information to perform your analysis. For assistance in interpreting the results, as well as recognizing the limitations of the analysis methods of EDENetwork, we recommend reading the references cited in this manual.

Examples of infiles are provided with the installation package and may be opened with a Text Editor such as Textpad (freely available on http://www.textpad.com) to check for the exact format (i.e. space and tabulation). Most infiles can be prepared either in a text editor or Excel, and then saved as text files (.txt) with tabulator separators.

One to two infiles should be prepared for each analysis. The first one is always required, containing the genetic data to be analyzed. Depending on what is being analyzed, the second optional file can contain auxiliary information either on sampled individuals or their populations.

2.1.1 Workflow in a nutshell

Typical use of EDENetwork consists of the following steps:

- Genetic data is imported
- Properties of the resulting distance matrix are inspected
- Distance matrix is thresholded, yielding a network
- The network is visually inspected for understanding structural features of the genetic system
- The network's properties are analyzed
- Key quantities such as betweenness centrality of nodes in population networks are inspected
- For population networks, the significance and robustness of key quantities is assessed
- Visualizations, graphs, data on nodes, and data on networks is saved.

Below, we will through these procedures step by step.

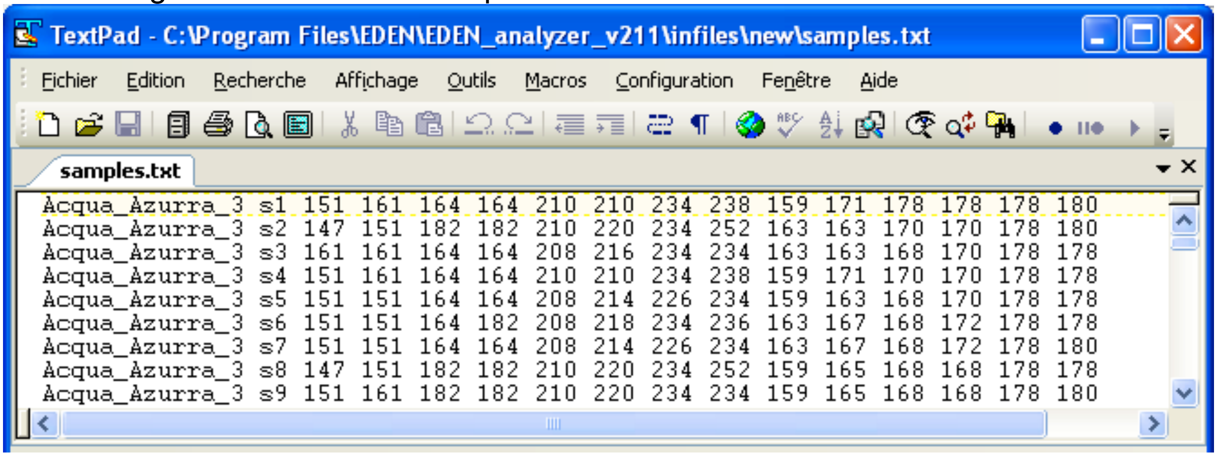
2.2. Infiles

2.2.1 Individual-centred analysis

For construction networks of genetic distances between individuals, the **first (main) infile** should contain the genotypes of each individual. The file should be a text file where each row corresponds to the genotype of an individual, such that

- the first column contains the sampling location identifier
- the second column contains the identifier of the individual sample
- the genotype is described starting from the third column, with each **allele encoded as three digits** and one column per allele

See the Figure below for an example.

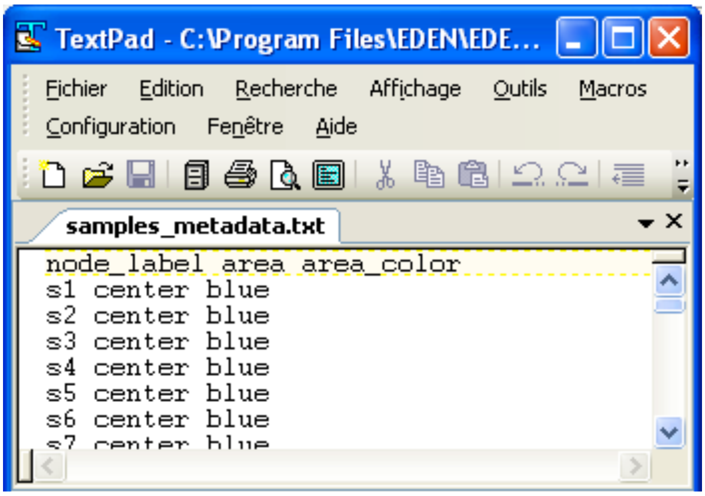


```
Acqua_Azurra_3 s1 151 161 164 164 210 210 234 238 159 171 178 178 178 180
Acqua_Azurra_3 s2 147 151 182 182 210 220 234 252 163 163 170 170 178 180
Acqua_Azurra_3 s3 161 161 164 164 208 216 234 234 163 163 168 170 178 178
Acqua_Azurra_3 s4 151 161 164 164 210 210 234 238 159 171 170 170 178 178
Acqua_Azurra_3 s5 151 151 164 164 208 214 226 234 159 163 168 170 178 178
Acqua_Azurra_3 s6 151 151 164 182 208 218 234 236 163 167 168 172 178 178
Acqua_Azurra_3 s7 151 151 164 164 208 214 226 234 163 167 168 172 178 180
Acqua_Azurra_3 s8 147 151 182 182 210 220 234 252 159 165 168 168 178 178
Acqua_Azurra_3 s9 151 161 182 182 210 220 234 234 159 165 168 168 178 180
```

In addition, the user has the choice for adding auxiliary information for each sampled individual (**second infile**). This auxiliary information may contain anything from pre-assigned sample classes to be used later in the analysis to colours for individual nodes to be used in visualization. As in the above file, one row corresponds to one sampled individual, however, an additional header row is required. Each column corresponds to one auxiliary variable; however, the first column must contain identifiers of individual samples (as in the second column of the first infile). See the Figure below for an example.

This file is constructed as follows:

- The first row contains the header
- The header contains labels for all auxiliary variables
- The first column of the header must be node_label
- The rest are user-specified
- The rest of the rows should begin with sample identifier and then contain values/labels/colors for each auxiliary variable (see below)



```
node_label area area_color
s1 center blue
s2 center blue
s3 center blue
s4 center blue
s5 center blue
s6 center blue
s7 center blue
```

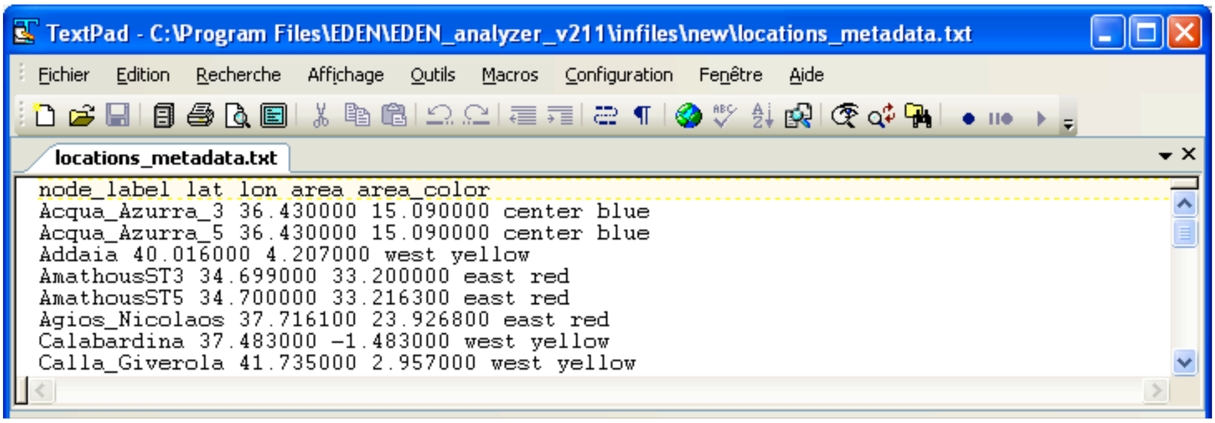
2.2.2 Population-level analysis

For population-level analysis, the **first infile** has the same format as for individual-level analysis (2.1.1. above).

The (optional) **second infile** should contain auxiliary data for sampling locations, which were input in the first column of the first infile. Each row corresponds to one sampling location and a header row is required. Each column corresponds to an auxiliary variable; however, the first column must contain identifiers of sampling locations (as in the first column of the first infile). See the Figure below for an example.

This file is constructed as follows:

- The first row contains the header
- The header contains labels for all auxiliary variables
- The first column of the header must be node_label
- The rest are user-specified
- The rest of the rows should begin with sampling location identifier and then contain values/labels/colors for each auxiliary variable (see below)



```
node_label lat lon area area_color
Acqua_Azurra_3 36.430000 15.090000 center blue
Acqua_Azurra_5 36.430000 15.090000 center blue
Addaia 40.016000 4.207000 west yellow
AmathousST3 34.699000 33.200000 east red
AmathousST5 34.700000 33.216300 east red
Agios_Nicolaos 37.716100 23.926800 east red
Calabardina 37.483000 -1.483000 west yellow
Calla_Giverola 41.735000 2.957000 west yellow
```

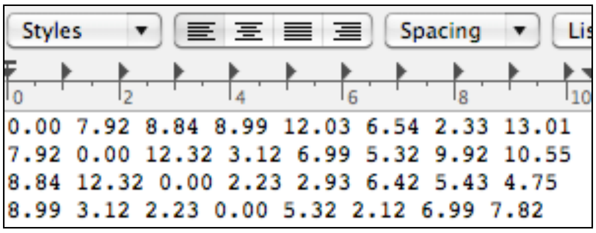
Depending on user choice, the auxiliary variables may include information such as

- the GPS coordinates
- (x,y) coordinates for user-defined layouts for network visualization

- the cluster of sampling locations, if known a priori, e.g. for testing purposes
- the colour code to be used in network visualization

2.2.3 Distance matrix

If the user wishes to work with pre-computed distance matrices instead of individual genotypes, a text file containing the distance matrix is required (first infile). This file should only contain the entries of the matrix, separated by tabs or white spaces. The distance matrix can represent population-level distances or distances between individuals.



```
0.00 7.92 8.84 8.99 12.03 6.54 2.33 13.01
7.92 0.00 12.32 3.12 6.99 5.32 9.92 10.55
8.84 12.32 0.00 2.23 2.93 6.42 5.43 4.75
8.99 3.12 2.23 0.00 5.32 2.12 6.99 7.82
```

Additionally, auxiliary information on the elements of the matrix can optionally be input (second infile). As with the auxiliary files for individual-centered or location-based analysis, the first row should contain header information, such that its first entry is node_label. The rest of the columns should contain labels for the additional information (as in 2.1.1 and 2.1.2). In order to correctly assign the entries of the auxiliary file to the nodes, the rows of this file should correspond to the order in which the matrix rows are given, i.e. the first row corresponds to the node whose distances to others are given in the first row of the matrix, and so on.

If this second infile is not given by the user, the nodes of the matrix are labelled with integer numbers (1..N).

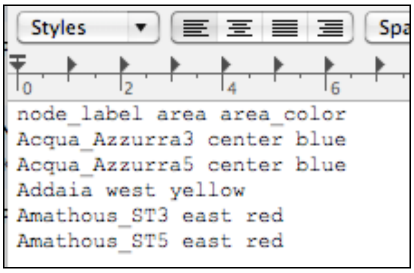
2.2.4 Network data

The user may also choose to work with pre-computed network data, either prepared outside EDENetwork analyzer, or calculated from genetic distance data with EDENetwork and saved in its network window. In this case, there are two options for the type of the **first infile**:

- Graph Markup Language (GML, *.gml) – see http://www.infosun.fim.uni-passau.de/Graphlet/GML/ for a full description
- Edge files (*.edg) which list all edges in the network.

For the latter, the format is as follows: every row in the text infile lists one edge with three columns: first vertex, second vertex, and edge weight:
`vertex1 vertex2 weight12.`

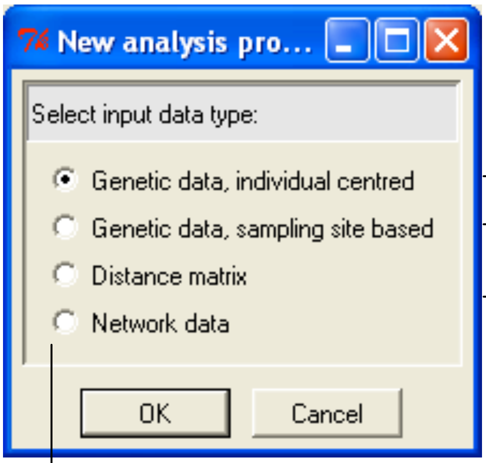
The second infile (optional) should be formatted similarly to the above; the first header row lists all auxiliary variables beginning with node_label and the rest of the rows contain node names (the vertex labels of the network file) and the values of their auxiliary variables.



```
node_label area area_color
Acqua_Azzurra3 center blue
Acqua_Azzurra5 center blue
Addaia west yellow
Amathous_ST3 east red
Amathous_ST5 east red
```

2.3. Interface

2.3.1 Begin a new analysis project



Select input data type:

● Genetic data, individual centred

○ Genetic data, sampling site based

○ Distance matrix

○ Network data

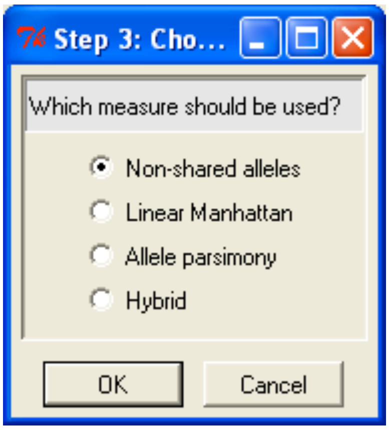
Genetic data
For analysis based on genotype matrix (see section 1.1.1), to build either individual centred (vertices/nodes are individuals) or population based (vertices/nodes correspond to sampling sites) networks.

Distance matrix
Allows entering an infile with a pre-computed pairwise distance matrix with any chosen distance (ecological, genetic, or any other pairwise distance) and at any level (individual, population, or location centred)

Network data
Allows entering an infile with a pre-computed network generated with EDENetwork or other software (individual, population, or location centred)

2.3.2 Genotype matrix : Individual centred or location based

For individual centred choose the distance measure to be used (for location based analysis the Goldstein distance will be used.)



Which measure should be used?

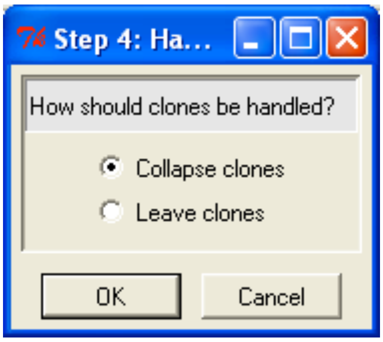
● Non-shared alleles

○ Linear Manhattan

○ Allele parsimony

○ Hybrid

When opening a genotype matrix for individual centred analysis, the program will ask the following


Step 4: Ha...

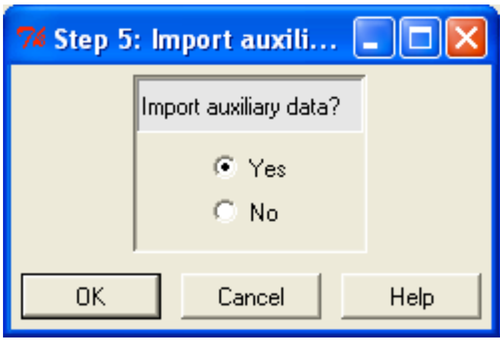
How should clones be handled?

● Collapse clones

○ Leave clones

Datasets on non clonal organisms should be analyzed ‘leaving clonal replicates’. As for clonal organisms it is recommended that you ascertain clonal membership (Arnaud-Haond et al., 2007) before choosing to collapse clones – when clones are collapsed, a single node will represent all samples with exactly similar genotypes. Note that if there are no similar genotypes, collapsing clones will have no effect on the data.

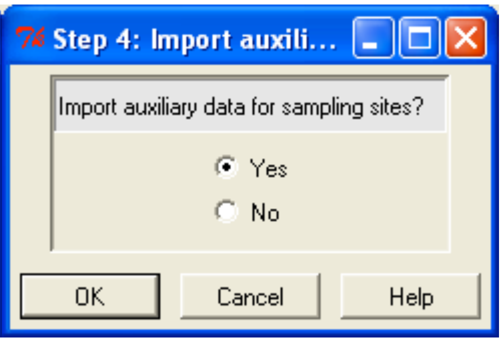
Then for either individual centred or location based analysis the program will propose the option to import auxiliary data (see § I).


Step 5: Import auxili...

Import auxiliary data?

● Yes

○ No


Step 4: Import auxili...

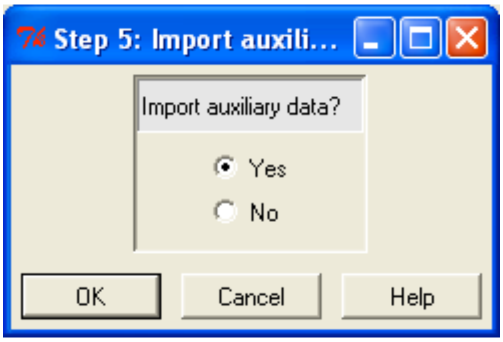
Import auxiliary data for sampling sites?

● Yes

○ No

2.3.3 Distance matrix, network data

When opening a distance matrix or a network file, the program will propose the option to import auxiliary data as above.


Step 5: Import auxili...

Import auxiliary data?

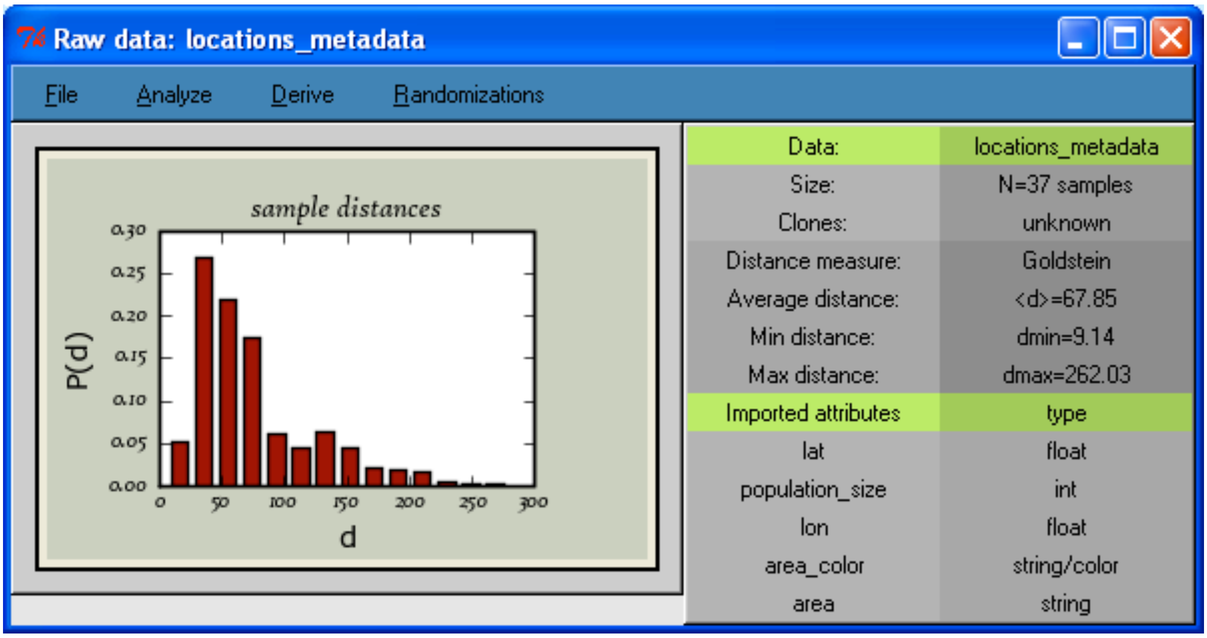
● Yes

○ No

2.4. *Data analysis*

2.4.1 Distance distribution

After entering a genotype or distance matrix, the data will be first analyzed in terms of distance distribution.



Data:	locations_metadata
Size:	N=37 samples
Clones:	unknown
Distance measure:	Goldstein
Average distance:	<d>=67.85
Min distance:	dmin=9.14
Max distance:	dmax=262.03
Imported attributes	type
lat	float
population_size	int
lon	float
area_color	string/color
area	string

All analysis will then start from this distance distribution.

The distance distribution can be viewed separately and investigated (from the Analyze menu), using linear, logarithmic, or semilogarithmic axes, or presented as a cumulative distribution. From this window, one can save the plot or the distribution as numbers (File -> Save graph, File->Export data).