

**TWC/21/3****ORIGINAL:** English**DATE:** May 15, 2003**INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS**

GENEVA

**TECHNICAL WORKING PARTY  
ON  
AUTOMATION AND COMPUTER PROGRAMS**

**Twenty-First Session  
Tjele, Denmark, June 10 to 13, 2003**

**PREDIP, A SOFTWARE PACKAGE TO PREDICT PHENOTYPIC DISTANCES WITH  
MOLECULAR DATA**

*Document prepared by experts from France*

## INTRODUCTION

1. Every year, hundreds of new varieties emerge from breeding activities. Among the conditions that have to be fulfilled to obtain a plant breeder's right, a variety must be distinct, uniform and stable (DUS).
2. The current method used to assess distinctness consists of the comparison between the candidate varieties and a set of reference varieties grown in one or two field locations in general, for two cycles. Experts check several phenotypic characteristics to assess if a candidate variety is distinct from known varieties. Many comparisons may be carried out and it would be useful to reduce this by eliminating reference varieties that are obviously distinct from the candidate.
3. As the collection of reference varieties and the number of candidate varieties to be tested increase, along with the experimental costs, it is crucial to find a new strategy to grow only the more relevant reference varieties even in the first year of trial; i.e. those that are similar to the candidate varieties.
4. Although distinctness is a phenotypic notion, GEVES (Groupe d'Etude et de contrôle des Variétés Et Semences) has studied the potential of molecular data for the rationalisation of field trials with two main goals: firstly to optimize the management of reference varieties and secondly to make examiners' work easier. This approach is in agreement with Option 2 defined by UPOV (see document TC/38/16 paragraphs 187 to 189).
5. Molecular data allow the determination of a genetic distance, which could give us some information on a possible proximity between varieties.
6. The studies of Bar-Hen *et al* [1 & 2] and, Burstin and Charcosset [3] on a maize line collection have demonstrated the existence of a triangular relationship between the Rogers' distance, estimated from the molecular data and the phenotypic distance computed from the phenotypic traits. Thus, the genetic distance cannot be used directly to estimate the phenotypic classification between varieties.
7. Nuel *et al* [5] have designed a pseudo-genetic distance, whose relationship with phenotypic distance is almost linear, using linear regression models. They have applied their method to a set of 144 maize lines using RFLP markers.
8. A software package, PREDIP, has been developed in SAS to implement this methodology for quantitative phenotypic data in self-pollinated and vegetatively-propagated varieties. PREDIP has recently been improved in collaboration with Mortier [4] who has generalised the methodology to treat quantitative and/or qualitative ordinal scaled phenotypic data.
9. We describe here the principles of this work, the parameters to be considered and some results obtained using data from rose.

## PREDIP

10. PREDIP is a software package developed in SAS (SAS institute Inc., version 8.01) using the modules SAS BASE, SAS STAT, SAS GRAPH and SAS IML. This software allows the prediction of phenotypic distances between self-pollinated or vegetatively-propagated varieties using molecular data. It works with several types of molecular data:

- Codominant molecular markers specific to a locus (e.g. RFLP, SSR). These are coded in terms of presence/absence of one allele at the locus: each marker produces as many bands as there are alleles revealed for the corresponding locus.
- Dominant molecular markers which are not specific to a locus (“fingerprinting”) and which are mass produced (e.g. RAPD, AFLP). These are coded in term of presence/absence: there is only one band for each molecular marker.

11. In this document, the term “banding data” refers to dominant molecular markers and “allelic” data refers to codominant molecular markers.

12. The calculations depend on the type of phenotypic data (quantitative and/or qualitative ordinal scaled data) and on the type of molecular data (allelic or banding data).

## Input files

13. Six files are required to use PREDIP:

- Three files deal with molecular data and consist of variety descriptions, marker descriptions and marker frequencies.
- Three files deal with phenotypic data and consist of variety descriptions, characteristic descriptions and phenotypic data.

14. Three files are optional:

- A file for locations if phenotypic data are measured in several places.
- A file for years if phenotypic data are measured in several years.
- A file for metric if users want to choose the weights of the phenotypic characteristics needed to calculate distances.

## 15. PREDIP Steps

- (a) Test of the existence and conformity of input files and directories.
- (b) Identification of:
  - Phenotypic characteristic type (quantitative and/or qualitative data).
  - Molecular marker type (allelic or banding data).
  - Total number of varieties and number of reference varieties.
  - Total number of molecular markers (number of bands for banding data, number of loci and alleles for allelic data).
- (c) Estimation of missing molecular data by the allelic frequency for all reference varieties.
- (d) Model selection:
  - For each phenotypic characteristic, PREDIP performs a selection of the molecular markers that best predict the trait.
  - If a model selection has already been performed, users can read data sets containing selected alleles lists (one list by phenotypic trait) in order to save computation time.
- (e) Estimation of the model parameters.
- (f) Estimation of phenotypic data predicted by the model for both candidate and reference varieties.
- (g) Computation of the predicted distance for each pair of varieties (Reference – Reference, Reference – Candidate, Candidate – Candidate) as the distance between predicted phenotypes. The Mahalanobis distance is computed by default, users can propose a specific metric to compute a distance with their own weights.
- (h) Estimation for each predicted distance of a one-sided or two-sided confidence interval at a given confidence level.
- (i) Creation of output files (text files and comma separated files) that contains predicted distances and confidence intervals for each pair of varieties.
- (j) Display of a classification tree based on the predicted distances.

## METHOD

16. This section describes the mathematical ideas incorporated in PREDIP. We refer to document TWC/19/10 for types of characteristics and their scale level definitions and document TWC/20/2 pages 2 to 6 for details on classical genetic distances.

### Main idea

17. The approach developed for phenotypic distance prediction lies in two main steps:

- (i) Prediction of phenotypic characteristics with molecular markers, based on a statistical model.
- (ii) Computation of an Euclidean distance between predicted phenotypes.

18. Quantitative characteristics are modeled on the basis of a classical linear model, whereas qualitative characteristics are modeled on the basis of a generalised linear model, using the notion of underlying gaussian variable.

19. The training data set used to fit the statistical model deals with reference varieties only. Phenotypic data are assumed to be unknown for candidate varieties. Once the model has been fitted with reference varieties, it is used to predict candidate characteristics according to their molecular profile.

### Model selection

20. In most cases, the number of reference varieties is not large enough to estimate all markers effects on phenotypic characteristics. Since not all the markers can be included in the model, a subset of relevant molecular markers must be determined for each phenotypic characteristic. These subsets may differ from one phenotypic characteristic to another.

21. For banding molecular data, the selection unit is the band, whereas for allelic data, it is the group of bands (alleles) associated with a locus. In order to select only the more relevant alleles, it is possible to perform a band by band selection on allelic data as if they were banding data.

22. For quantitative characteristics, selection is performed using a Forward/Backward Stepwise Selection algorithm based on a F statistic. The significance level for a marker to be included in the model is the same as the level to be deleted from the model.

23. For qualitative characteristics, selection is performed using a Forward Stepwise Selection algorithm based on a likelihood ratio test. The significance level can be different from the level used in the quantitative case.

Quantitative data: predicting characteristics with molecular markers

24. For each variety  $k$ ,  $k \in \{1, \dots, n\}$ , we assume that a set  $X = (X^1, \dots, X^p)$  of  $p$  quantitative characteristics and a set  $B$  of  $m$  markers ( $l \in \{1, \dots, m\}$  each of them with  $s_l$  bands) are observed. The details are provided in Nuel *et al.*

25. Let  $M$  be the matrix coding for genotypes  $B$  (with a dummy variable added for the intercept). For each characteristic  $X^i$ , a linear model using molecular markers as independent variables is fitted:

$$X^i = M\mathbf{b}_i + \mathbf{e}^i,$$

where  $\mathbf{b}_i$  is the vector of effects for the molecular markers and  $\mathbf{e}^i$  a gaussian random vector of errors. The effects corresponding to bands that were not chosen in the selection process are set to zero.

26. For each variety  $k$ , the random error  $\mathbf{e}_k^i$  and the observation  $X_k^i$  are normally distributed:

$$\mathbf{e}_k^i \approx N(0, \mathbf{s}_i^2), \quad X_k^i \approx N(M_k \mathbf{b}, \mathbf{s}_i^2),$$

where  $M_k$  is the marker profile for variety  $k$  ( $k^{\text{th}}$  row of matrix  $M$ ). The observations are assumed to be independent with the same variance  $\mathbf{s}_i^2$ .

27. A multivariate model is made by amalgamating the  $p$  univariate models:

$$(1) \quad X = M\mathbf{b} + E,$$

where  $\mathbf{b}$  is the matrix  $(\mathbf{b}_1, \dots, \mathbf{b}_p)$  of effects for all characteristics and  $E$  is a random matrix of errors. For each variety  $k$ , the vector of random errors  $E_k = (\mathbf{e}_k^1, \dots, \mathbf{e}_k^p)$  and the vector of observations  $X_k = (X_k^1, \dots, X_k^p)$  are normally distributed with covariance matrix  $\Sigma$ :

$$(2) \quad E_k \approx N(0, \Sigma), \quad X_k \approx N(M_k \mathbf{b}, \Sigma).$$

28. The individuals are assumed to be independent and characteristics are correlated with the same covariance matrix  $\Sigma$ .

Euclidean distance between varieties according to the predicted quantitative characteristics

29. The squared Euclidean distance between the varieties  $k$  and  $k'$  according to a given metric  $\Pi$  ( $p \times p$  matrix of weights) is:

$$d^2(L_k, L_{k'}) = (X_k - X_{k'})\Pi(X_k - X_{k'})^T.$$

30. The predicted distance is defined as the distance between the expected value of  $X_k$  and  $X_{k'}$  according to the multivariate linear model (*cf.* ( 1 ) and ( 2 ) ):

$$(3) \quad d^2(L_k, L_{k'}) = (M_k - M_{k'})\mathbf{b}\Pi\mathbf{b}^T(M_k - M_{k'})^T.$$

$(M_k - M_{k'})$  is the contrast between molecular profiles for varieties  $k$  and  $k'$ . The predicted distance can be considered as an Euclidean distance between molecular profiles in which the matrix of weights  $\mathbf{b}\Pi\mathbf{b}^T$  takes into account phenotypic information (phenotypic information lies in  $\mathbf{b}$ ). By default, PREDIP computes the predicted Mahalanobis distance which metric is  $\Pi = \Sigma^{-1}$ .

31. Figure 1 shows the improvement of the prediction for phenotypic distances using the pseudo-genetic distance designed by Nuel *et al* rather than the Rogers genetic distance in the case of maize inbred lines. Both graphs use the same vertical axis.

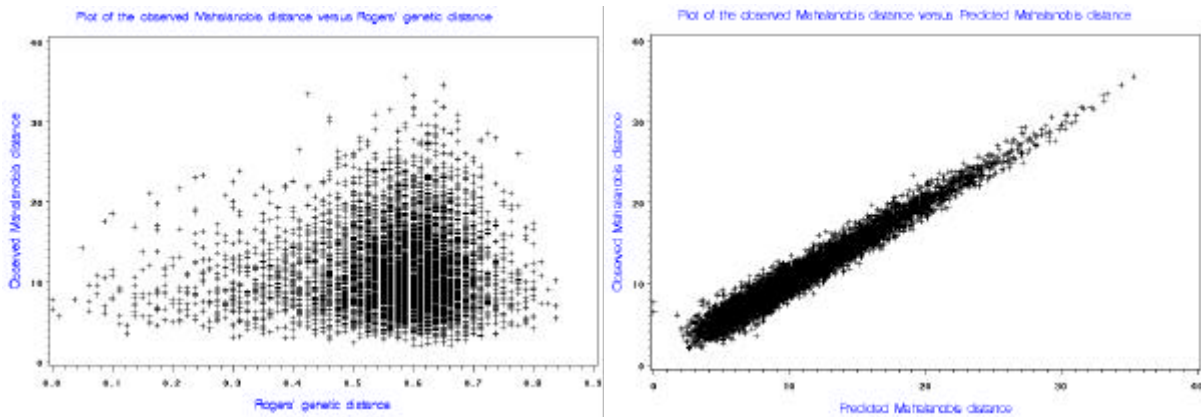


Figure 1: Observed Mahalanobis distance versus Rogers distance (left) and predicted Mahalanobis distance (right) for reference – reference couples, selection level for markers = 5%, 144 maize lines data set.

Qualitative ordinal scaled data: predicting underlying characteristics with molecular markers

32. Ordinal characteristics are neither suitable for computation of Euclidean distances nor for modelling with linear regression.

33. The approach developed by Mortier [4] is to simulate the quantitative case by assuming that underlying each ordinal characteristic is a normally distributed quantitative characteristic whose expected value is a linear combination of molecular markers and whose variance is arbitrarily set to one.

34. Each observed ordinal characteristic is considered as the expression of the corresponding quantitative underlying characteristic on an ordinal scale.

35. Although the underlying characteristic often has a physical basis, this is primarily a mathematical convenience and a physical basis is not required.

36. As in the quantitative case, the individuals (varieties) are assumed to be independent. For variety  $k$ , the underlying characteristic has an expected value  $M_k \mathbf{b}$  where  $M_k$  codes for the marker profile of variety  $k$  and  $\mathbf{b}$  is the vector of marker effects. Expected values for the underlying characteristic can differ from one variety to another.

37. The following example explains the principles of the model:

38. Measures for the characteristic “anthocyanin coloration of silk” ( $A$ ) are binary data, 0 for absence and 1 for presence. Intuitively, observation of anthocyanin coloration occurs when the anthocyanin rate ( $AR$ ) exceeds a given threshold  $\mathbf{a}$  :

$$\begin{cases} AR \approx N(M_k \mathbf{b}_{AR}, 1) \\ A = 0 \Leftrightarrow AR < \mathbf{a} \\ A = 1 \Leftrightarrow AR \geq \mathbf{a} \end{cases}$$

39. The characteristic “density of spikelets” ( $DS$ ) has three levels: 0 for lax, 1 for medium and 2 for dense. It can be modeled with a normal underlying characteristic  $U$  and 2 thresholds  $\mathbf{a}_1$  and  $\mathbf{a}_2$  :

$$\begin{cases} U \approx N(M_k \mathbf{b}_U, 1) \\ DS = 0 \Leftrightarrow U \leq \mathbf{a}_1 \\ DS = 1 \Leftrightarrow \mathbf{a}_1 < U \leq \mathbf{a}_2 \\ DS = 2 \Leftrightarrow U > \mathbf{a}_2 \end{cases}$$

40. Figure 2 illustrates the principle of this model for two varieties ( $k = 1, 2$ ): thresholds, markers effects and the distribution shape are the same for all varieties, however, the expected value changes with respect to their molecular profile. The areas under the gaussian curve and delimited by the thresholds correspond to the probabilities for the ordinal characteristic to take its values.

41. The correlation structure between the two ordinal characteristics,  $A$  and  $DS$ , is inherited from the correlation structure between their underlying variables  $AR$  and  $U$ .



42. The unknown parameters which need to be estimated in this model are the two sets of thresholds  $\mathbf{a}$  and  $\{\mathbf{a}_1, \mathbf{a}_2\}$ , the vectors of effects  $\mathbf{b}_{AR}$  and  $\mathbf{b}_U$  and the correlation between the two underlying variables  $\mathbf{r}$ . Estimates are calculated by maximizing the model likelihood function.

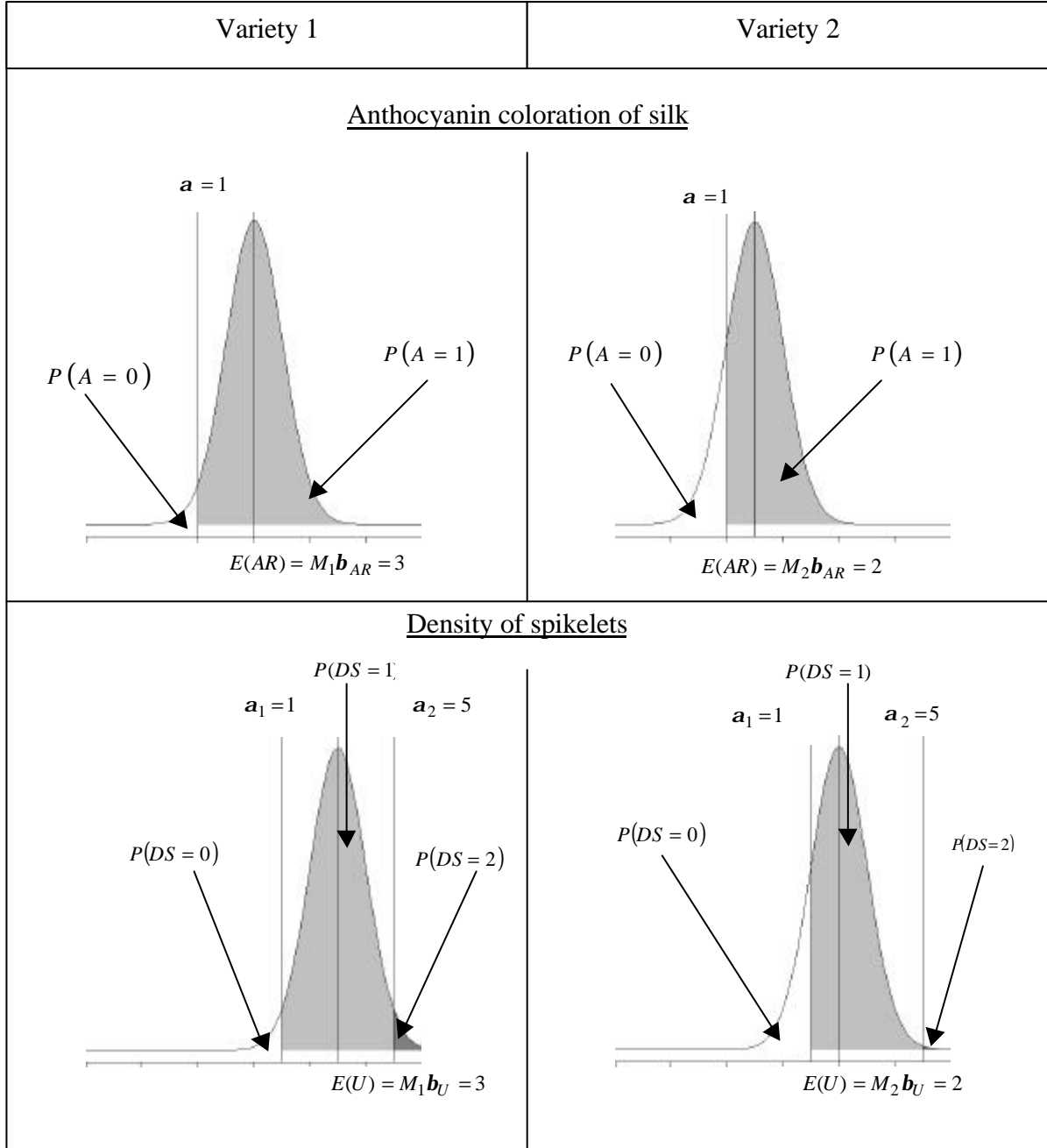


Figure 2: Relationship between ordinal traits and their underlying gaussian characteristic variable for varieties 1 and 2.

Euclidean distance between varieties according to the predicted underlying characteristics

43. Let us call  $Y$  the vector  $(AR, U)$  and  $\mathbf{b}$  the matrix of effects  $(\mathbf{b}_{AR}, \mathbf{b}_U)$ . The distance according to a metric  $\Pi$  between two varieties  $k$  and  $k'$  is based on their underlying phenotypes:

$$d^2(L_k, L_{k'}) = (Y_k - Y_{k'})\Pi(Y_k - Y_{k'})^T.$$

44. The expected value for the distance is the distance between the expected values of  $Y$  :

$$(4) \quad \mathbf{d}^2(L_k, L_{k'}) = (M_k - M_{k'})\mathbf{b}\Pi\mathbf{b}^T(M_k - M_{k'})^T.$$

45. Formula (4) is similar to formula (3) in the quantitative case; it is the expression of an Euclidean distance between molecular profiles where the metric is  $\mathbf{b}\Pi\mathbf{b}^T$ . The default metric  $\Pi$  used by PREDIP is  $R^{-1}$  the inverse of the correlation matrix  $\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$ .

46. Details about the general modelling of ordinal data and the modelling of a mixture of quantitative and ordinal data are available in Mortier's Thesis, pages 17-32. The expression for the corresponding Euclidean distances and their distributions can be found in pages 41-51.

## APPLICATION

47. This section deals with an application on rose data. 14 ordinal phenotypic traits and 393 AFLP<sup>TM</sup> (Amplified Fragment Length Polymorphism) markers were observed on 72 varieties.

48. Even where the phenotypic data were ordinal, they were considered to be quantitative because of sample size: at least 200 varieties are needed for the qualitative model, whereas 50 varieties are enough for the quantitative model. Furthermore, the ordinal characteristics have sufficient levels to provide a good variability.

49. The whole data set has been divided into three sets:

- A training set of 54 varieties, for which both phenotypic and marker data are known to fit the model. These varieties have been declared as reference varieties: observed and predicted distances can be compared.
- A validation set of 10 varieties (N° 85, 86, 87, 88, 89, 90, 94, v10, v6, v9). Both phenotypic and markers data are known but they have not been used to fit the model. These varieties have been declared as candidate varieties: observed and predicted distances can be compared.
- A set of 8 varieties (N° 3, 4, 42, 47, 70, 74, 75, 76) for which only marker data are known. Only predicted distances can be computed.

50. Markers have been selected for each phenotypic trait using Forward/Backward selection with a 2% level.

51. The distance computed is the usual Euclidean distance with each characteristic equally weighted one.

52. Figure 3 shows the relationship between the observed phenotypic distance, the classical Nei & Li molecular distance (left) and the distance predicted by PREDIP (right). The comparison is made for all types of variety pairs (Reference – Reference, Reference – Candidate, Candidate – Candidate).

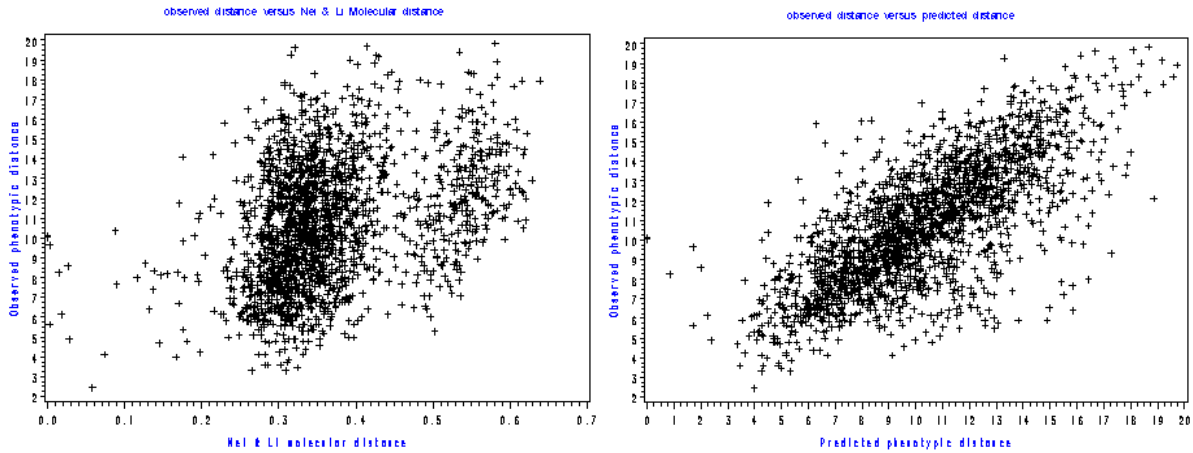


Figure 3: Observed phenotypic distance versus Nei & Li distance and predicted phenotypic distance.

53. Figure 3 and Table 1 show the improvement of the prediction when using the predicted distance rather than the classical Nei & Li distance: the observed phenotypic distance is more correlated with the predicted phenotypic distance than with the classical Nei & Li molecular distance.

<b>Pearson Correlation and Number of Observations</b>		
	<b>Nei &amp; Li molecular distance</b>	<b>Predicted phenotypic distance</b>
<b>Observed phenotypic distance</b>	0.37	0.72
	2016	2016

Table 1: Observed distance versus Nei & Li distance and predicted distance, correlation.

54. In Table 2, the observed phenotypic distance versus predicted phenotypic distance has been split according to the pair type: Reference – Reference (top), Reference – Candidate (middle), Candidate – Candidate (bottom). It shows the quality of the relationship between observed and predicted distances according to the three different cases. Dashed lines show where predicted distances equal observed distances.

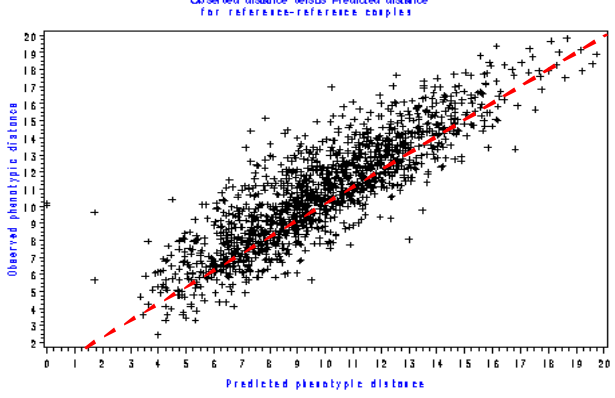
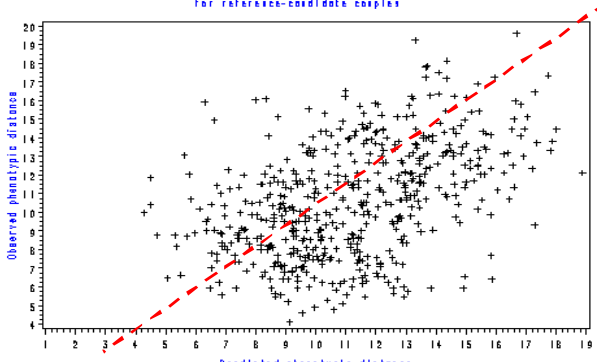
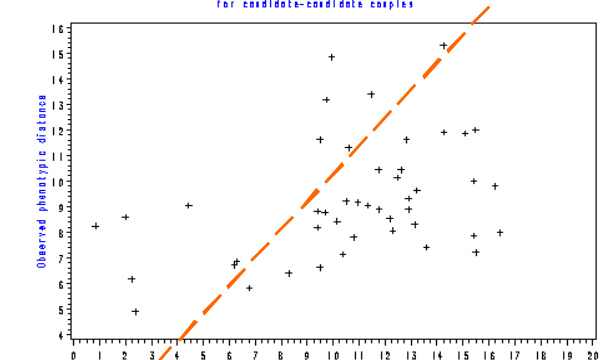
	Number of observations:  1431	<u>Predicted distance</u>
	<u>Observed distance</u>	Correlation:  0.86
Reference – Reference couples		
	Number of observations:  540	<u>Predicted distance</u>
	<u>Observed distance</u>	Correlation:  0.44
Reference – Candidate couples		
	Number of observations:  45	<u>Predicted distance</u>
	<u>Observed distance</u>	Correlation:  0.39
Candidate – Candidate couples		
Observed phenotypic distances versus predicted phenotypic distances.	Correlations between observed phenotypic distance and predicted phenotypic distance.	

Table 2: Relationship between observed phenotypic distance and predicted phenotypic distance according to couples type (pictures and correlations).

55. Table 2 shows that the relationship between the observed distance and the predicted distance is almost linear for pairs of reference varieties (training varieties). In the case of pairs of either one or two candidate varieties, distance prediction is less efficient but results can be improved using larger training sample sizes.

56. Figure 4, Figure 5 and Figure 6 in the annex show the classification dendrograms based on the observed distances, the predicted distances and the Nei & Li standard genetic distances. These three classifications have been carried out using the UPGMA (Unweighted Pair Groups Method of Aggregation) clustering method.

57. In Figure 4, classes have been highlighted for two levels of distance: 0.75 and 1.10.

58. The first classification (0.75 level) produces 9 groups whereas the second (1.10) produces 3 groups. Groups have been identified on the figure using colours in greyscale at the bottom for level 0.75 and at the top for level 1.10.

59. In Figure 5, classes have been made for the two equivalent distance levels (0.75 and 1.06). The former classification (0.75) produces 6 groups and the latter (1.06) produces 3 groups. Varieties with a square are those whose phenotype was not known and which are not shown in Figure 4. The colours added in greyscale represent the identification of the groups from the classification of Figure 4 in order to compare the two classifications (at the bottom for the level 0.75 and at the top for the level 1.10).

60. The comparison between the two classification trees shows that the predicted distance is not precise enough to recreate the classification from observed distances at the 0.75 level but some similarities are consistent from one classification to another (e.g. varieties 77, 55, 53 and 57).

61. For the level 1.10, the three main groups of the observed classification are almost the same as for the predicted classification.

62. Group I is analogous to group 1, Group II is analogous to group 2 and group III is analogous to group 3. Table 3 shows a cross-tabulation of observed groups against predicted groups at the 1.10 level. At the intersection of two groups is the number of varieties belonging to both of them.

		Groups from the observed classification			Sizes of predicted groups
		<b>1</b>	<b>2</b>	<b>3</b>	
Groups from the predicted classification	<b>I</b>	<b>8</b>	3	0	11
	<b>II</b>	3	<b>31</b>	0	34
	<b>III</b>	1	0	<b>18</b>	19
Sizes of observed groups		12	34	18	<b>64</b>

Table 3: Cross-tabulation of observed groups against predicted groups at the 1.10 level for the 64 varieties which observed phenotype is known.

63. Figure 6 in the annex shows the classification based on the Nei & Li molecular distance, colours in greyscale are the identification of groups 1,2 and 3 at the 1.10 level from the observed classification (Figure 4). Varieties that are not coloured are candidate varieties.

64. The comparison of Figure 5 and Figure 6 is another way to show that the predicted distance is better than the Nei & Li molecular distance for forming groups that are analogous to those built with the observed phenotypic distance.
65. Of ten candidate varieties, seven were predicted close to the same reference varieties as for the observed classification.
66. In particular, this example shows the quality of distance prediction for pairs of varieties in the training set (Reference – Reference). Simulations led by Mortier [4] pages 52-68 allow us to be optimistic about the prediction of distance for mixed pairs (Reference – Candidate) when a larger training sample size is available.

## CONCLUSION

67. The methodological approach initiated by Nuel *et al* and generalised by Mortier is now fully programmed. PREDIP is able to work with most of phenotypic data sets stemming from DUS trials. A validation step is now being developed.

68. Moreover, additive location and year effects can now be integrated in the model in order to take into account more phenotypic variability. This may also present a solution to the sample size problem, since it allows the use of multi-locational and over-years data sets. Interaction between location and year will be added in the next version of PREDIP.

69. It is planned to validate PREDIP on several species such as rapeseed, maize, rose, potato, sunflower, durum wheat, cherry and peach.

70. In the case of maize inbred lines, the reference collection management for the second year of trial is based on the phenotypic distances computed by the GAÏA software. These distances are computed from the phenotypic observations available for reference and candidate varieties after the first year of trial and some isoenzymatic data. Hence, it seemed interesting to treat the data that are currently used for GAÏA with PREDIP. Some varieties have been considered as candidates in order to compute actual predicted distances. Isoenzymatic data have been used to predict the phenotypic data and about 2000 varieties were available for training set. We have then compared PREDIP predicted distances to GAÏA observed distances. From this comparison, a 50 pairs experimental field trial has been designed to illustrate different cases: agreement or disagreement of PREDIP and GAÏA distances, according to different types of pairs (training, mixed or candidate pairs).

71. Moreover, it would be conceivable to take advantage of a DUS trial to check *a posteriori* the distance predictions from molecular data such as SSR markers.

72. In both situations, PREDIP distances could be compared to GAÏA distances and to experts' opinions.

73. In this way, we will have more information on the quality of phenotypic distance predictions according to several criteria, such as sample size and type of markers.

AUTHORS: M. FRANCK, S. LASSALVY

Acknowledgements for reviewing:

S. GREGOIRE, J. GUIARD, F. MORTIER, A. ROBERTS, S. ROBIN.

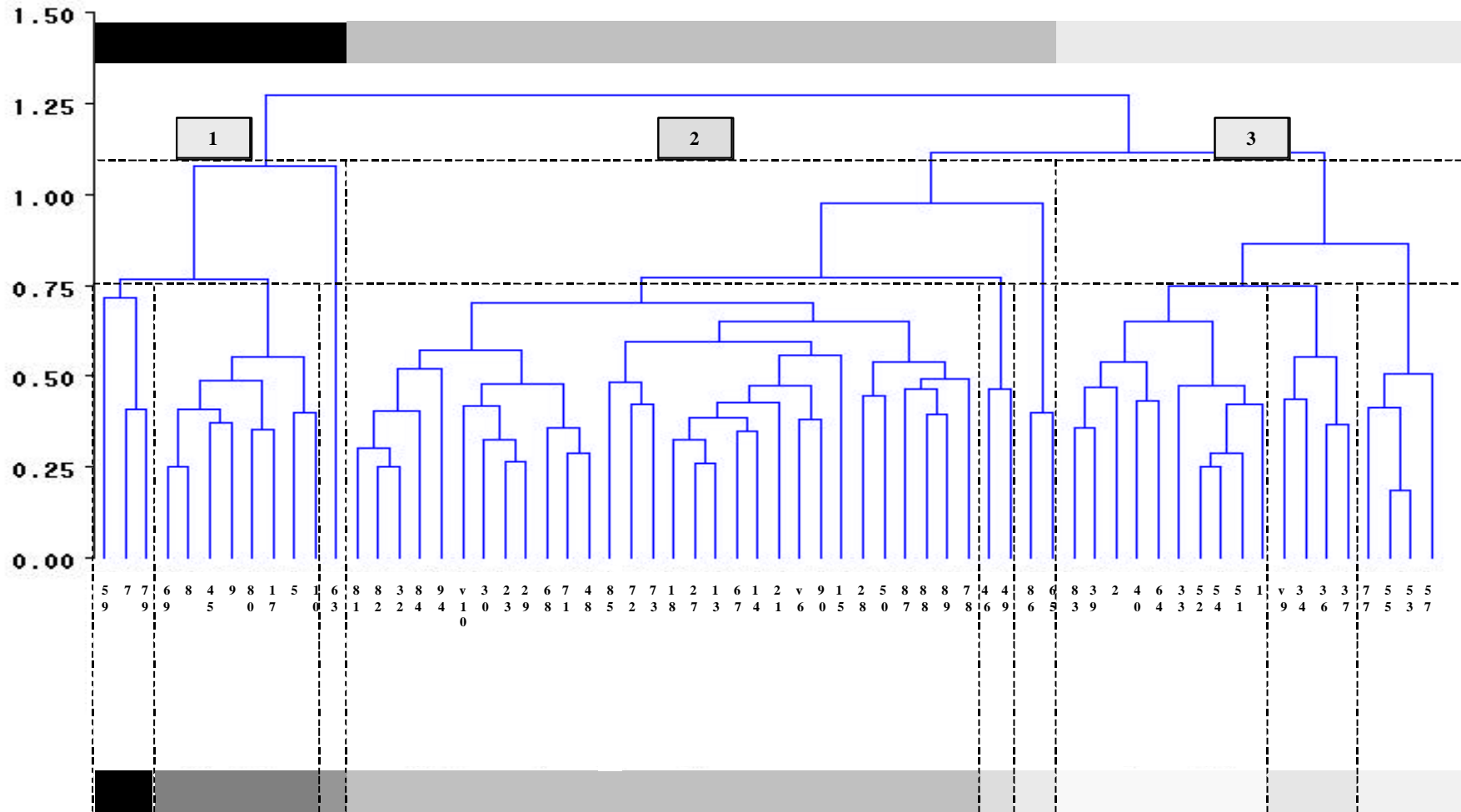


Figure 4: Classification tree based on the Observed phenotypic distances, clustering method: UPGMA.



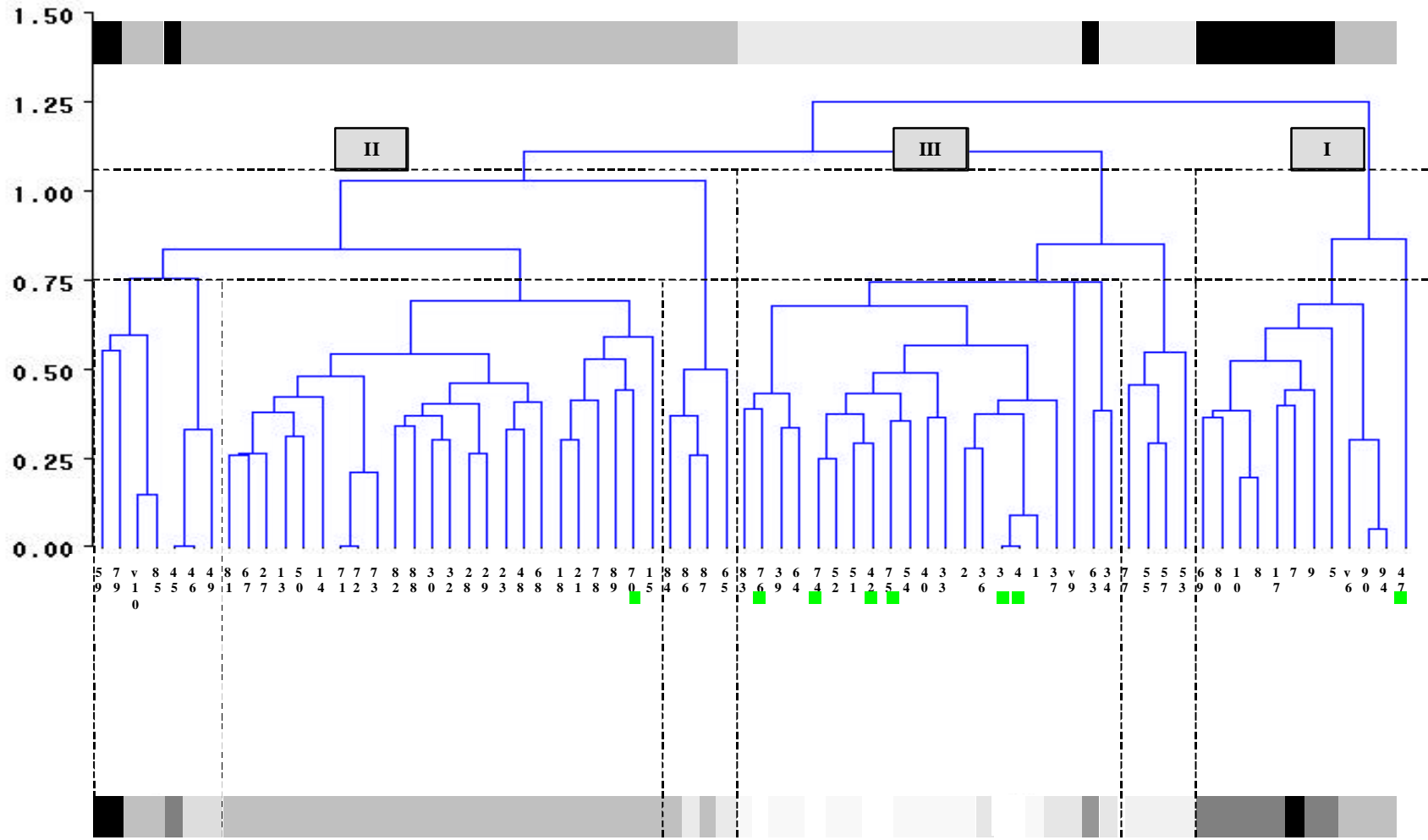


Figure 5: Classification tree based on the predicted phenotypic distances, clustering method: UPGMA.

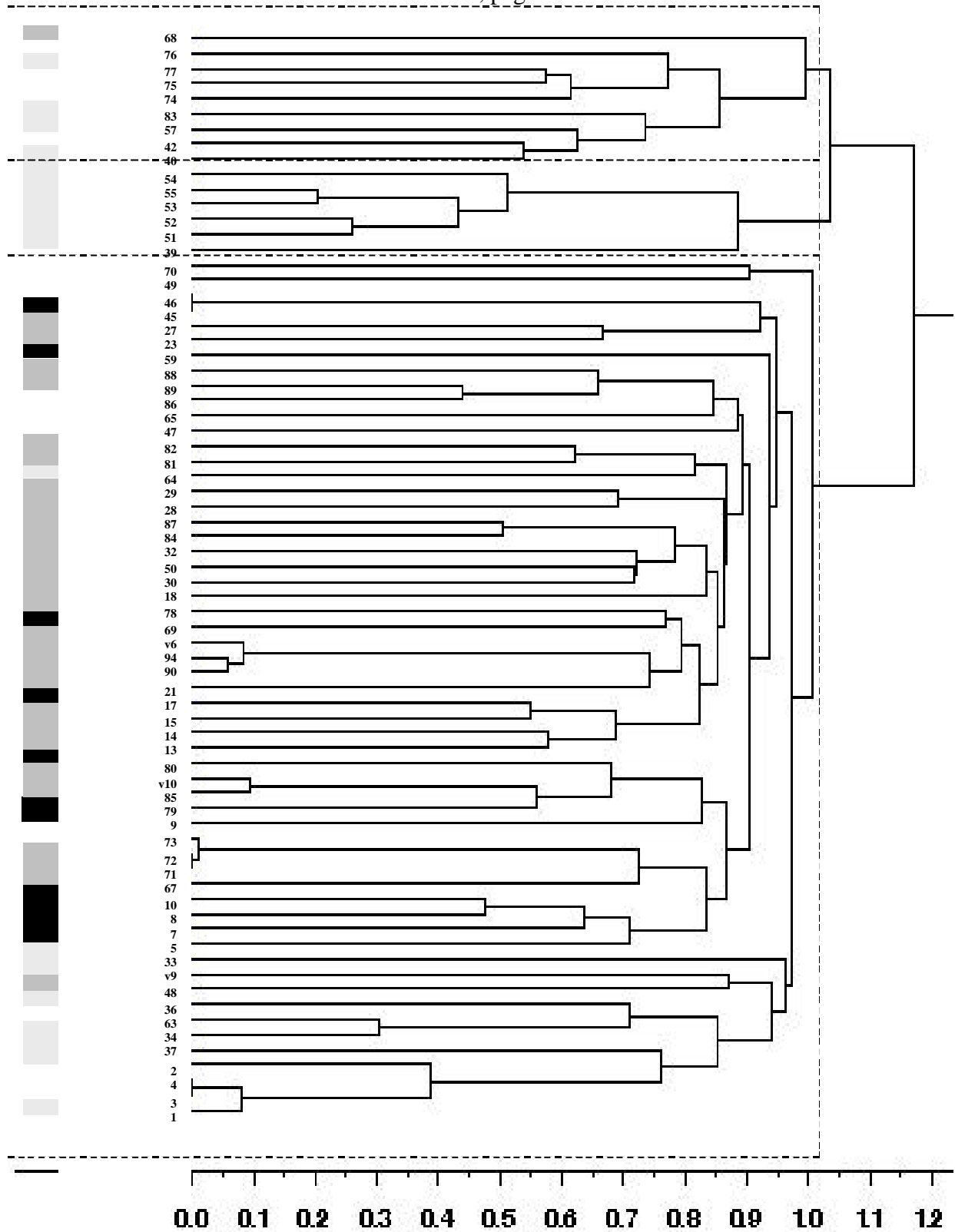


Figure 6: Classification tree based on the standard Nei & Li molecular distance, clustering method: UPGMA.

REFERENCES

- [1] Bar-Hen, A. & Charcosset, A. (1995) Relationship between molecular and morphological distances in a maize inbred line collection. Application for breeder's right protection. In: J.W. Van Oijen and J. Jansen (Eds) *Biometrics in Plant Breeding: Application of molecular markers* (Proceedings of the 9th meeting of the Eucarpia, Section Biometrics Plant Breeder) pp. 57-66 (CPRO-DLO, Wageningen).
- [2] Bar-Hen, A., Charcosset, A., Bourgoïn, M. & Guiard, J. (1995) Relationship between genetic markers and morphological traits in a maize inbred line collection, *Euphytica*, 84, pp. 145-154.
- [3] Burstin, J. & Charcosset, A. (1997) Relationship between phenotypic and marker distances: theoretical and experimental investigations, *Heredity*, 79(5), pp. 477-483. Available on internet at [www.nature.com/hdy](http://www.nature.com/hdy).
- [4] Mortier, F. (2002). Estimation de distances euclidiennes généralisées: Application à la distinction variétale, *Thèse de doctorat, Université des Sciences et Technologies de Lille*.
- [5] Nuel, G., Robin, S., & Baril, C.P. (2001) Predicting distances using a linear model: the case of varietal distinctness, *Journal of applied statistics*, 28(5), pp. 607-621.

[End of Annex and of document]