E

BMT/8/14
ORIGINAL: English
DATE: August 13, 2003

UPOV

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS

GENEVA

WORKING GROUP ON BIOCHEMICAL AND MOLECULAR TECHNIQUES AND DNA-PROFILING IN PARTICULAR

Eighth Session
Tsukuba, Japan, September 3 to 5, 2003

BAND SCORING, DISTANCES, USE FOR DISTINCTNESS AND UNIFORMITY, DATA STORAGE

Document prepared by experts from France and the United Kingdom

1. Scoring of Molecular Data

1. For a number of years protein electrophoresis has been recognized as a valuable tool in identifying and discriminating between crop varieties. Agreed methods are part of the UPOV Test Guidelines for a limited number of crops. Much research has taken place over recent years demonstrating the potential of DNA approaches. It is the purpose of this section to outline issues that need to be addressed when dealing with DNA data.

2. There are many DNA profiling methods available to crop scientists.
- AFLP
- Micro-satellites (SSR and variants)
- Sequence data (SNP)

3. It is important to select scoring strategies that reflect the genetics of the crop under study (e.g. ploidy) and the mode of operation of the DNA profiling system (e.g. a co-dominant system). For example, the number of samples that are needed will depend on the observed (or assumed) level of within variety uniformity. For self-pollinated varieties, clonal material and inbred lines, a limited number of samples may be required to have a sample representative of such varieties. For allogamous crops large samples sizes are required with individuals making up the population generating band frequencies.

4. Before starting to score DNA profiling systems a number of issues need to be addressed. Some of these are given below and will depend on the end-use of the results.

- Are monomorphic bands to be scored or only polymorphic ones?
- Include “faint” bands?
- How are rare alleles to be handled?
- Are the data to be added to a database?

5. Once these points have been agreed then standard protocols need to be drawn up to ensure consistency of scoring in the long-term irrespective of the personnel involved.

6. When scoring is done manually, it may be necessary to have two independent scorings and include only those identified by both scorers. This will become more important when assessments are made across-gels.

7. Where ‘automated scoring’ is applied, the band intensity threshold, or other user tunable parameters, needs to be researched and fixed, as far as possible, to increase the robustness of the scoring system (re-scoring the same gels on another occasion or in different laboratories using the same ‘system’). With high-intensity multiplexed systems, manual scoring may be difficult.

8. While in some circumstances missing data can be accommodated in the calculation of the distance estimator, there will be occasions where a consistent treatment of missing data will be required. Removing an individual from the subsequent analysis just because it has a small fraction of missing data, is a drastic step. There are a number of proposed strategies for dealing with modest amounts of missing data. (See paper to be published by Law and Van Eeuwijk).

9. For data required for databasing, the choice of markers is as important as the quality assurance applied to any generated data. Markers of choice may not always be those with optimal PIC values, but will depend much more on the ease of scoring, "robustness", freely available, sufficiently polymorphic and other factors such as multiplexibility.

2. Distance Measures (based on a paper by H. P. Piepho and F. Laidig)

10. Based on either banding data or allelic data, distances and similarity measures can be computed. Such measures may be viewed as convenient means of data reduction, and they need not involve any genetic concept (Weir, 1990: 163). Some of the measures for allelic data were designed based on genetic models specifying the processes underlying the divergence of populations. It should be checked whether or not these assumptions are met in practice.

2.1 Binary banding data

11. Sneath and Sokal (1973) list four classes of similarity measures: (i) distance coefficients, (ii) association coefficients, (iii) correlation coefficients, and (iv) probabilistic similarity coefficients. Most measures relevant for the analysis of binary banding data fall within the class of association measures, which are based on qualitative data (multi-state or two-state). Occasionally association measures turn out to be special cases of distance coefficients or correlation coefficients. For a comprehensive overview, see Sneath and Sokal (1973) and Clifford and Stephenson (1975). In the following, we will give a few measures which we found to be frequently used in genetical studies. For each genotype, when a set of markers has been checked, for each marker the presence/absence is recorded as 0 or 1, the data of two genotypes can be summarized as 2 x 2 frequency table

		x	
		0	1
y	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

12. From this basic table the following frequencies can be computed (Armstrong *et al.*, undated)

n_{00}	= number of band positions scored 0 for *y* and 0 for *x*
n_{10}	= number of band positions scored 1 for *y* and 0 for *x*
n_{01}	= number of band positions scored 0 for *y* and 1 for *x*
$n_{11} = n_{xy}$	= number of band positions scored 1 for *y* and 1 for *x*
$n_x = n_{01} + n_{11}$	= number of bands present in *x*
$n_y = n_{10} + n_{11}$	= number of bands present in *y*
$m_{xy} = n_{00} + n_{11}$	= number of matches
$n = n_{00} + n_{10} + n_{01} + n_{11}$	= number of band positions

13. The most important distinction is between measures that ignore negative matches (0, 0 comparisons) and measures that do not. It is debatable whether or not exclusion of negative matches is useful in the context of DNA profiles. Take the following simple example with three genotypes *x*, *y* and *z*:

Table 3: Example of scores for banding data of two genotypes *x*, *y* and *z* (4 band positions):

Band position		1 2 3 4
Genotype	*x*	0 0 1 1
	y	1 1 1 1
	z	0 1 1 1

14. In a way, *z* is as similar to *x* as it is to *y*, because in both comparisons, three of four comparisons are concordant. The only difference is that in the *z*-*x* comparison, only two of the three concordant observations are positive matches, while in the *z*-*y* comparisons all are positive matches. On the other hand, for a negative match to be observable, the corresponding band must be observed for at least one of the other genotypes. Thus, any similarity measure, which takes into account negative matches, will depend on the particular set of genotypes included in the study, which is a point in favor of measures ignoring negative matches. Moreover, there are several ways in which a genotype may lose a band/DNA fragment, so it may be argued that basing similarity on the mutual absence of a character is inappropriate (Vierling and Nguyen, 1992).

15. The similarity measures (*s*) given here take values in the range from zero to unity. For identical genotypes $s = 1$, while for completely distinct measures $s = 0$. The distance measure corresponding to these similarity measures may be computed as $1 - s$.

2.1.1 Measures that ignore negative matches

(1) Nei and Li (1979):

$$NL_{xy} = 2n_{xy}/(n_x + n_y) = 2n_{11}/(2n_{11} + n_{01} + n_{10})$$

16. This is probably the most popular similarity measure in genetic analyses. It is equivalent to the Dice coefficient (Sneath and Sokal, 1973: 131) and assesses the proportion of bands shared by two genotypes *x* and *y*. Under certain statistical assumptions, NL_{xy} may be employed to derive an estimate of the mean number of nucleotide substitutions per nucleotide site (Nei and Li, 1979), which is a useful parameter in evolutionary studies. The underlying assumptions are probably realistic in natural populations, but doubtful in plant breeding. If the computations are exclusively based on single-banded RFLP patterns, then NL_{xy} is equal to Rogers distance (see below) (Melchinger, 1993).

(2) Jaccard (Sneath and Sokal, 1973: 131):

$$J_{xy} = n_{11}/(n - n_{00}) = n_{11}/(n_{01} + n_{10} + n_{11})$$

17. NL_{xy} is the same as J_{xy}, except that positive matches carry double weight. It has been suggested (Link *et al*., 1995) that NL_{xy} is more appropriate for RFLP data, while J_{xy} should be used with RAPD data. The reasoning is as follows (Engqvist and Becker, 1995): RAPD markers either produce a band in a certain position or the band is absent. Thus, one band

position usually corresponds to one marker locus. By contrast, RFLPs produce fragments of varying lengths for different alleles. For two cultivars differing at a marker locus, fragments are produced for both alleles, but they differ in their position on the gel. Hence a locus is represented by two band positions. When the cultivars are identical, however, the locus is manifest in only one band position for the pairwise comparison. Thus, matches should receive double weight compared to mismatches, as in NL_{xy}. This reasoning implies that RAPDs show no length polymorphisms and that each RFLP allele produces only one band on the gel. Both of these assumptions are idealizations, but may be reasonable approximation in practice.

2.1.2 Measures, which treat positive and negative matches alike

18. These measures are symmetric in n_{00} and n_{11}, i.e. the formula stays the same when n_{00} and n_{11} are exchanged. Only the most popular measure is given here. For other measures, see Sneath and Sokal (1973) and Clifford and Stephenson (1975).

(3) Simple Matching (Sneath and Sokal (1973: 132)

$SM_{xy} = m_{xy}/n = (n_{11} + n_{00})/n$

19. The simple matching coefficient measures the proportion of positive and negative matches. In order to compare SM_{xy} with measures that ignore negative matches, we computed some similarities for the example in Table 3. SM_{xy} yields the same similarity/distance for the pairs *x*-*z* and *y*-*z*, while measures ignoring negative matches such as J_{xy} and NL_{xy} indicate a larger similarity between *y* and *z*.

$J_{xz} = 0.67$ $NL_{xz} = 0.67$ $SM_{xz} = 0.75$
$J_{yz} = 1$ $NL_{yz} = 0.86$ $SM_{yz} = 0.75$

2.2 Allelic frequency data and band frequency data

20. In the following x_i and y_i will denote the frequencies of allele *i* at a given locus for genotypes *x* and *y*, respectively. Alternatively, x_i and y_i may denote the band frequency at band position *i*, when banding data are used.

(1) Euclidean distance

21. The frequencies x_i and y_i can be viewed as coordinates of points in a multidimensional space. The geometric distance may be interpreted as distance between populations *x* and *y*.

$E_{xy} = [\sum_i (x_i - y_i)^2]^{0.5}$

This distance is mentioned here as it is always given as the first possible distance, but it is not useful in the scope of UPOV work.

22. When allelic data from several loci are available, the distances for individual loci may be averaged. E_{xy} takes a value between 0 and $\sqrt{2}$. A standardization to values between 0 and 1 leads to

(2) Rogers' distance (Nei, 1987: 211)

$RD_{xy} = [0.5 \sum_i (x_i - y_i)^2]^{0.5}$

23. In the important case that x and y are inbred lines and allelic data are used, Rogers' distance (RD_{xy}) equals the percentage of loci which differ between lines *x* and *y*. Its expectation is related to the coefficient of coancestry (Melchinger *et al*., 1991). The Rogers distance has the following deficiency: When the two populations are both polymorphic but share no common alleles, this measure can become much smaller than unity even if the populations have entirely different sets of alleles (Nei, 1987: 209).

(3) Nei's standard genetic distance

24. Nei's measure is intended for allelic data. When no allelic information is available, it may be computed from band frequency data. If this is done, however, the measure does not have the genetic interpretation as if computed from allelic data. The *normalized identity of genes* or simply *genetic identity* is given by

$I_{XY} = J_{XY}/(J_X J_Y)^{1/2}$

25. Where $j_{xy} = \sum x_i y_i$, $j_x = \sum x_i^2$, $j_y = \sum y_i^2$ and J_x, J_y and J_{xy} are the averages of j_x, j_y and j_{xy} over all scored loci. I_{xy} is 1 when the two populations have identical gene frequencies over all loci and is 0 when they share no alleles. Because of this property, I_{xy} has been used for measuring the extent of genetic similarity between populations. The quantity $D_{xy} = -\text{Ln}(I_{xy})$ is the *standard genetic distance*. Under the assumption that the rate of gene substitution per locus is uniform across both loci and lineages and some other assumptions, it is an estimator for the number of codon differences per locus between two populations *x* and *y* (Nei, 1987: 218; Nei, 1972). While I_{xy} ranges from zero to unity, D_{xy} varies between zero and infinity.

Loarce *et al*. (1996) computed the genetic identity based on band frequencies of RAPD fragments from bulked DNA samples of two rye cultivars. O'Donoughue *et al*. (1994) computed D_{xy} for band frequencies from RFLPs in oats. When computed from band frequency data, D_{xy} probably does not allow the interpretation as a measure of the number of codon differences, though it is a valid descriptive distance measure.

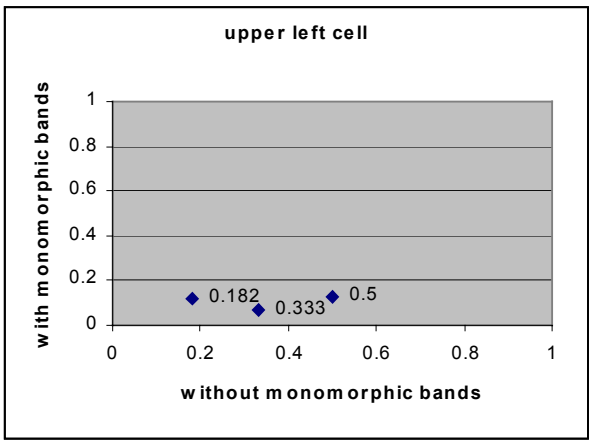
26. See examples in Appendix I.

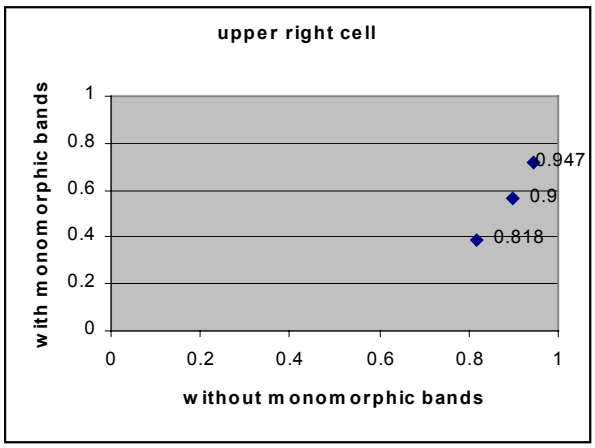
The 3 tables of distances, show that each distance measure will give a different result on a given set of data, for instance in the upper right cell respectively 0.818 0.900 and 0.947 in table A B and C, and in the upper left cell 0.333 0.500 0.182

The addition of 6 monomorphic bands illustrate the influence of "non informative data", as all genotypes have the same profile for these bands.

The values for the upper right cells become 0.391 0.563 0.720 in table D E and F, and in the upper left cell 0.067 0.125 0.118.

In the case of the upper right cell the ranking order of the values is kept when monomorphic bands are added, while the ranking order changes for upper left cells. So not only the distance values are different but also the ranking.





3. Distinctness and Uniformity

27. UPOV has not yet established how biomolecular data could be used in the process of DUS testing. Discussions are ongoing in BMT and other groups. During the last sessions of BMT, a number of papers on different crops were presented with the aim of exploring how biomolecular data could help to describe and distinguish varieties, and also to check how uniform the varieties are.

28. The discriminant power of biomolecular data is great. On the one hand, there has been some concern about this discriminating power being too big, or more appropriate to assess essential derivation than to assess DUS. On the other hand, the phenotypic traits are often susceptible to the effect of environment, while it should not be the case for biomolecular data if the method is carefully selected and technically well controlled.

29. Some concern about varieties not being uniform for biomolecular data has also been expressed. A very large majority of the papers submitted to the BMT indicated that varieties do not have the same level of uniformity according to their reproduction system, as is the case for phenotypic traits; but within plant, and within variety uniformity can be compared to what is found using phenotypic traits.

30. The uncontrolled use of biomolecular techniques could open the gate to the use of hundreds of new “characteristics”. This is not acceptable in the UPOV context where the strength of the protection right is essential. UPOV is considering how the use of biomolecular techniques might be used in a way which does not undermine the value of protection. The BMT Review Group had considered the proposals set out in document TC/38/14-CAJ/45/5 and gave some indications that have been endorsed by the technical committee (TC/38/15). Different elements shall be defined for instance set the principles of use, select which methods would be appropriate and reproducible, describe how the data should be used in the scope of DUS testing.

31. The data already provided through the various sessions of the BMT have shown that biomolecular data have a potential interest and might be taken into account when the

principles and modalities of their use can be established. There are different types of data available, different ways to compare or summarize them, and UPOV will have to establish how the data will be used in the DUS decision process to ensure harmonization between countries.

4. Databases for variety information

32. The storage of the information related to the varieties examined is very often kept in computer files, although some information and notes are kept on different papers as well.

33. In common language, the existence of information in a computer file is incorrectly referred to as "the data is kept in a database". The storage of information in a spreadsheet, for instance Excel©, where lines are different varieties (or lots) and columns are information in order to identify the variety (i.e. variety code) and information obtained (i.e. allele found) is such an example. This is very convenient when the number of lines and columns is small as it allows a lot of functions (sort, average,…) on the data and nice outputs in order to provide descriptions or reports.

34. When the amount of information is increasing, and when it is necessary to provide simultaneous access to many users, a database is needed. Access to the data is then controlled by software which allows different levels of access to the data (right to read, create, update, delete).

35. When a database is needed the data has to be analyzed in order to provide models to describe them and their relationships. A model can comprise several tables, in the most simple cases, to more than a hundred "tables". Each table is the equivalent of a row x column spreadsheet, where the rows are the different elements of information, and the columns the different types of information needed to identify the object, and describe it. Database design and implementation is usually performed by computer experts in cooperation with the users of the information. Database design is not such an easy task, and usually a design has a cost and so an economical value, which is why designs are sometimes protected, and rarely "given on request".

36. Administrative and technical description of the varieties already known, and under study, is commonly stored in databases by the offices in charge of DUS testing and the issue of titles of protection.

37. The notes which are used to provide the description are sometimes in a database and sometimes directly in a separate document per variety.

38. The results from biomolecular studies are sometimes in databases, sometimes in special files. For the time being (except for electrophoresis), as biomolecular results are not used in routine DUS tests, it is not necessary to include the results in the official administrative and technical databases.

39. International organizations such as UPOV or the European Union, for instance, developed databases in which information on known varieties and varieties under study are available. In this paper, only the UPOV approach is described.

40. In 1996, UPOV *ad hoc* subgroups proposed a format description for providing variety information. Since then, a UPOV ROM has been produced by UPOV and sent to members. The UPOV ROM is also available on subscription. The format is described in the Circular U 2462 09/11/1996 in which 30-40 types of information are described and identified by TAGs. The description of the varieties is not yet included, but this is not due to computer restrictions. Plant variety denominations was one of the reasons to develop this database. UPOV provides on the CD-ROM the database itself, the data, a program to search the database, and other documents of interest in pdf format.

41. The option chosen by UPOV is the use of SGML. SGML is an abbreviation for "Standard Generalized Markup Language" which was first defined by an International Standard in 1986. In this system the information is provided in electronic text files and the recognition of the information is made by the use of tags. For instance, in the UPOV ROM the tag <600> is used to indicate that the following information in the file is the breeder's reference of the variety under treatment. In the general scope of SGML, XML "Extensible Markup Language" is the reference for new developments for exchange of information (description of the information and exchange), while HTML "Hyper Text Markup Language", is used to describe the display of information on the computer screens through the internet. Many software packages provide basic tools to create or input XML files, whilst, with the present SGML UPOV format, each organization must write and maintain its own specific software to cope with the required specifications.

42. To define a database sub-model for the purpose of information exchange, another option is possible. In such a case, the information can be provided without the use of tags, as the content and specifications are known by the definition of the database model. This approach is probably better if the aim is to reach harmonization and encourage cooperative work, as it allows easier exchange, and the possibility to interconnect different databases at real time.

43. Examples of how information can be kept in different ways is illustrated in Appendix 2.

44. In both cases, (Markup language or database sub-model) when information is exchanged and grouped, the identification of the objects is essential in order to be able to search for information coming from different sources, or to relate objects (varieties for instance) studied in different places. UPOV has already defined such identification codes for the characteristics in the UPOV Test Guidelines. By contrast, the variety code is given by each country, and no database link is provided in the UPOV ROM to identify if a given variety has been described by one or several countries. The use of species name can help to identify such cases, but this is left to the use of the database. The coding of the species is not yet harmonized by UPOV although some work has been done in that direction.

45. It is premature to imagine how bio-molecular results would be input in an administrative and technical database, as these results are not yet used by UPOV members for DUS. Nevertheless, two possible uses can be imagined in the future. The first one has already been used for electrophorosis, where UPOV defined *ad hoc* characteristics in the guidelines for the corresponding species. With this approach, there is no difference in essence with the way agronomic characteristics are used. A variety has a value or note for a characteristic. The second one is already used for agronomic characteristics where different values are obtained before a final "note" is given. In this case, for each variety there is a set of values (i.e. measure on 60 individual plants from 3 different trials). We can already imagine in such a case that at least the lot, the sample, the material used, the test, the locus,

the allele, the frequency, will enable us to cope with most, if not all, of the situations where we will have to compute criteria for distinctness and uniformity.

APPENDIX I

46. Examples of 3 similarity measures, each on the same data set, and with the addition of 6 monomorphic bands, follow:

Correlation Between Original 'Raw' Data (10 bands) & + Mono (10 + 6 monomorphic bands)

NL	0.9444
J	0.9496
RT	0.9990

Correlation Between Similarity Methods 'Raw' Data

	NL	J
J	0.9843	
RT	0.8349	0.8609

	NL-M	J-M
J-M	0.9968	
RT-M	0.9662	0.9733

Table A NL:- Nei & Li
Table B J:- Jaccard's
Table C RT:- Roger and Tanimoto
Table D NL-M (Nei & Li based on 'Raw' + 6 Monomorphics bands)
Table E J-M (Jaccard's based on 'Raw' + 6 Monomorphics bands)
Table F RT-M (Rogers and Tanimoto based on 'Raw' + 6 Monomorphics bands)

Table A NEI and LI (=DICE)

	0.333	0.500	0.600	1.000	1.000	0.750	0.778	0.800	0.818
0.333		0.200	0.333	0.714	0.750	0.556	0.600	0.636	0.667
0.500	0.200		0.143	0.500	0.556	0.400	0.455	0.500	0.538
0.600	0.333	0.143		0.333	0.400	0.273	0.333	0.385	0.429
1.000	0.714	0.500	0.333		0.091	0.167	0.231	0.286	0.333
1.000	0.750	0.556	0.400	0.091		0.077	0.143	0.200	0.250
0.750	0.556	0.400	0.273	0.167	0.077		0.067	0.125	0.176
0.778	0.600	0.455	0.333	0.231	0.143	0.067		0.059	0.111
0.800	0.636	0.500	0.385	0.286	0.200	0.125	0.059		0.053
0.818	0.667	0.538	0.429	0.333	0.250	0.176	0.111	0.053	

Tables B JACCARDS

	0.500	0.667	0.750	1.000	1.000	0.857	0.875	0.889	0.900
0.500		0.333	0.500	0.833	0.857	0.714	0.750	0.778	0.800
0.667	0.333		0.250	0.667	0.714	0.571	0.625	0.667	0.700
0.750	0.500	0.250		0.500	0.571	0.429	0.500	0.556	0.600
1.000	0.833	0.667	0.500		0.167	0.286	0.375	0.444	0.500
1.000	0.857	0.714	0.571	0.167		0.143	0.250	0.333	0.400
0.857	0.714	0.571	0.429	0.286	0.143		0.125	0.222	0.300
0.875	0.750	0.625	0.500	0.375	0.250	0.125		0.111	0.200
0.889	0.778	0.667	0.556	0.444	0.333	0.222	0.111		0.100
0.900	0.800	0.700	0.600	0.500	0.400	0.300	0.200	0.100	

Table C ROGERS and TANIMOTO

	0.182	0.333	0.462	0.750	0.824	0.750	0.824	0.889	0.947
0.182		0.182	0.333	0.667	0.750	0.667	0.750	0.824	0.889
0.333	0.182		0.182	0.571	0.667	0.571	0.667	0.750	0.824
0.462	0.333	0.182		0.462	0.571	0.462	0.571	0.667	0.750
0.750	0.667	0.571	0.462		0.182	0.333	0.462	0.571	0.667
0.824	0.750	0.667	0.571	0.182		0.182	0.333	0.462	0.571
0.750	0.667	0.571	0.462	0.333	0.182		0.182	0.333	0.462
0.824	0.750	0.667	0.571	0.462	0.333	0.182		0.182	0.333
0.889	0.824	0.750	0.667	0.571	0.462	0.333	0.182		0.182
0.947	0.889	0.824	0.750	0.667	0.571	0.462	0.333	0.182	

Table D (MONO BANDS ADDED)
NEI and LI (=DICE)

	0.067	0.125	0.176	0.333	0.368	0.300	0.333	0.364	0.391
0.067		0.059	0.111	0.263	0.300	0.238	0.273	0.304	0.333
0.125	0.059		0.053	0.200	0.238	0.182	0.217	0.250	0.280
0.176	0.111	0.053		0.143	0.182	0.130	0.167	0.200	0.231
0.333	0.263	0.200	0.143		0.043	0.083	0.120	0.154	0.185
0.368	0.300	0.238	0.182	0.043		0.040	0.077	0.111	0.143
0.300	0.238	0.182	0.130	0.083	0.040		0.037	0.071	0.103
0.333	0.273	0.217	0.167	0.120	0.077	0.037		0.034	0.067
0.364	0.304	0.250	0.200	0.154	0.111	0.071	0.034		0.032
0.391	0.333	0.280	0.231	0.185	0.143	0.103	0.067	0.032	

Table E (MONO BANDS ADDED)
JACCARDS

	0.125	0.222	0.300	0.500	0.538	0.462	0.500	0.533	0.563
0.125		0.111	0.200	0.417	0.462	0.385	0.429	0.467	0.500
0.222	0.111		0.100	0.333	0.385	0.308	0.357	0.400	0.438
0.300	0.200	0.100		0.250	0.308	0.231	0.286	0.333	0.375
0.500	0.417	0.333	0.250		0.083	0.154	0.214	0.267	0.313
0.538	0.462	0.385	0.308	0.083		0.077	0.143	0.200	0.250
0.462	0.385	0.308	0.231	0.154	0.077		0.071	0.133	0.188
0.500	0.429	0.357	0.286	0.214	0.143	0.071		0.067	0.125
0.533	0.467	0.400	0.333	0.267	0.200	0.133	0.067		0.063
0.563	0.500	0.438	0.375	0.313	0.250	0.188	0.125	0.063	

Table F (MONO BANDS ADDED)
ROGERS and TANIMOTO

	0.118	0.222	0.316	0.545	0.609	0.545	0.609	0.667	0.720
0.118		0.118	0.222	0.476	0.545	0.476	0.545	0.609	0.667
0.222	0.118		0.118	0.400	0.476	0.400	0.476	0.545	0.609
0.316	0.222	0.118		0.316	0.400	0.316	0.400	0.476	0.545
0.545	0.476	0.400	0.316		0.118	0.222	0.316	0.400	0.476
0.609	0.545	0.476	0.400	0.118		0.118	0.222	0.316	0.400
0.545	0.476	0.400	0.316	0.222	0.118		0.118	0.222	0.316
0.609	0.545	0.476	0.400	0.316	0.222	0.118		0.118	0.222
0.667	0.609	0.545	0.476	0.400	0.316	0.222	0.118		0.118
0.720	0.667	0.609	0.545	0.476	0.400	0.316	0.222	0.118	

APPENDIX 2

STORAGE OF INFORMATION IN DIFFERENT WAYS:

EXTRACT FROM FILE SENT TO UPOV FOR CREATION OF THE UPOV-ROM (SGML FORMAT, INFORMATIONS ARE IDENTIFIED BY TAGS DESCRIBED IN CIRCULAR U 2462)

For instance Tag 190 is to identify information "country sending the information"

```
<000>0
<190>FR
<010>PBR 15249
<500>Prunus persica (L.) Batsch
<510>Pecher
<540>20001212 DIAMOND JUNE
<541>20001210 DIAMOND JUNE
<600>SF 94 026
<210>17069
<220>20000120
<400>20000210
<730>GIE Star Fruits (FR)
<731>Lowell Glen Bradford (US)
<731>Norman Glen Bradford (US)
<733>GIE Star Fruits (FR)
<000>0
<190>FR
<010>PBR 1192
<500>Brassica napus L.
<510>Colza
<540>20001211 ADDITION
<541>         ADDITION
<600>DMB 428
<210>17515
<220>20001211
<730>GIE Serasem (FR)
<730>Agri Obtentions SA (FR)
<731>Institut National de la Recherche Agronomique (DZ)
<731>Institut National de la Recherche Agronomique (FR)
<731>Serasem (FR)
<733>GIE Serasem (FR)
<733>Agri Obtentions SA (FR)
```

Example of information kept in spreadsheet: When there is more than one set of information kept for a variety, the maintenance of the file can become a problem, e.g. there are two co-breeders in variety DIAMOND JUNE; 2 co-applicants, 3 co-breeders, 2 co-title holders for the Brassica napus L. variety.

PBR or NLI	Identification number	species name	Latin name	date denominati on proposed	denominatio n proposed	date denominatio n accepted	denominati on accepted	breeder's reference	applicatio n number	date application	publication of date…	Applicant's name	Breeder's name	Title holder's
PBR	15249	Pecher	Prunus persica (L.) Batsch	12/12/2000	DIAMOND JUNE	02/10/2000	DIAMOND JUNE	SF 94 026	17069	20/01/2000	02/10/2000	GIE Star fruits (FR)	Lowell Glen Bradford (US) Norman Glen Bradford (US)	GIE Star fruits (FR)
"	"	"	"	"	"	"	"	"	"	"	"	"		"
PBR	1192	Colza	Brassica napus L.	11/12/2000	ADDITION		ADDITION	DMB 428	17515	11/12/2000		GIE Serasem(FR) Agri Obtentions SA (FR)	Institut National de la Recherche Agronomique (DZ) Institut National de la Recherche Agronomique (FR)	GIE Serasem (FR) Agri obtentions SA (FR)
PBR	1192	Colza	Brassica napus L.	11/12/2000	ADDITION		ADDITION	DMB 428	17515	11/12/2000			GIE Serasem (FR)	

Example of information kept in a database: To avoid redundancy and to ease maintenance, information is kept in a set of tables, for instance the name and address of a party is unique in the table LIST of PARTIES even if this party is used for dozens of varieties. Identification of appropriate party is done via a party code.

DENOMINATION table

variety code	denomination	date	status
15249	DIAMOND JUNE	12/12/2000	proposed
15249	DIAMOND JUNE	10/12/200	published
1192	ADDITION	11/12/2000	proposed

SPECIES table

code	latin name	species name
1200	pecher	Prunus persica (L.) Batsch
2000	colza	Brassica napus L.

VARIETY table

code	PBR or NLI	species code	breeder's ref	application number	date application
15249	PBR	1200	SF 94 026	17069	20/01/2000
1192	PBR	2000	DMB 428	17069	11/12/2000

PARTIES CONCERNED FOR A VARIETY

variety code	party code	status
15249	234	applicant
15249	234	title holder
15249	154	breeder
15249	14	breeder
1192	58	applicant
1192	4562	applicant
1192	58	title holder
1192	4562	title holder
1192	11	breeder
1192	100	breeder
1192	58	breeder

LIST of PARTIES table

party code	name
234	GIE Star fruits (FR)
154	Lowell Glen Bradfrod (US)
14	Norman Glen Bradfrod (US)
58	GIE Serasem (FR)
4562	Agri Obtentions SA (FR)
11	Institut National de la Recherche Agronomique (DZ)
100	Institut National de la Recherche Agronomique (FR)

References

Clifford, H.T. and Stephenson, W. (1975). An introduction to numerical classification. Academic Press, New York.

Engqvist and Becker (1995). Genetic diversity for allozymes, RFLPs and RAPDs in resynthesized rape. Proceedings of the 9th EUCARPIA section in Plant Breeding, 6-8 July 1994, Wageningen.

Link, W., Dixkens, C., Singh, M., Schwall, M. and Melchinger, A.E. (1995). Genetic diversity in European and Mediterranean faba bean germ plasm revealed by RAPD markers. TAG 90: 27-32.

Loarce, Y., Gallego, R. and Ferrer, E. (1996). A comparative analysis of the genetic relationship between rye cultivars using RFLP and RAPD markers. Euphytica. 88: 107-115

Melchinger, A.E., Messmer, M.M., Lee, M., Woodman, W.L. and Lamkey, K.R. (1991). Diversity and relationships among U.S. maize inbreds revealed by restriction fragment length polymorphisms. Crop Science. 31: 669-678.

Nei, M. (1972). Genetic distance between populations. American Naturalist. 106: 283-292.

Nei, M. (1987). Molecular evolutionary genetics. Columbia University Press, New York.

Nei, M. and Li, W-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences (USA). 76: 5269-5273.

O'Donoughue, L.S., Souza, E., Tanksley, S.D. and Sorells, M.E. (1994). Relationships among North American oat cultivars based on restriction fragment length polymorphisms. Crop Science. 34:1251-1258.

Sneath, P.H.A. and Sokal, R.R. (1973). Numerical taxonomy. Freeman and Company, San Francisco.

Weir, B.S. (1990). Genetic data analysis. Sinauer, Sunderland.

[End of document]