UPOV

INTERNATIONALUNIONFORTHEPROTECTIONOFNEWVARIETIESOFPLANTS

GENEVA

TECHNICALWORKINGPA RTY
ON
AUTOMATIONANDCOMPU TERPROGRAMS

TwentiethSession
MexicoCity,June17to20,2002

BANDSCORING

DocumentpreparedbyexpertsfromFranceandtheUnitedKingdom

1. ScoringofMolecularData

1.   For a number of years protein electrophoresis has been recognized as a valuable tool in identifying and discriminating between crop varieties. Agr   eed methods are part of the UPOV TestGuidelinesforalimitednumberofcrops.Muchresearchhastakenplaceoverrecentyears demonstrating the potential of DNA approaches. It is the purpose of this section to outline issuesthatneedtobeaddressedw   hendealingwithDNAdata.

2.   TherearemanyDNAprofilingmethodsavailabletocropscientists.
- AFLP
- SSR
- (andvariantsoftheabove)
- Sequencedata

3.   Itisimportanttoselectscoringstrategiesthatreflectthegeneticsofthecrop   understudy (e.g. ploidy) and the mode of operation of the DNA profiling system (e.g. a co   -dominant system).Forexample,thenumberofsamplesthatareneededwilldependontheobserved(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 d   ealing 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 choic   e 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.Piep   ho 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 o   f 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. The data of two genotypes can be arranged as 2x2 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 x and 0 for y
n_{10} = number of band positions scored 1 for x and 0 for y
n_{01} = number of band positions scored 0 for x and 1 for y
$n_{11} = n_{xy}$ = number of band positions scored 1 for x and 1 for y
$n_x = n_{01} + n_{11}$ = number of bands present in x
$n_y = n_{10} + n_{11}$ = number of bands present in y
$n_{xy} = n_{00} + n_{11}$ = num ber 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 neg ative 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   1234

Genotype x 0011
             y 1111
             z 0111

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):

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

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.

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

22.   When allelic data from sever al 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)

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

23.   In the i mportant 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 R ogers 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 $j_{xy}=\ \textstyle\sum x_i y_i$, $j_x=\ \textstyle\sum x_i^2$, $j_y=\ \textstyle\sum y_i^2$ 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 $D_{xy} = -Ln(I_{xy})$ 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 popu lations 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 cu ltivars. 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 descri ptive distance measure.

26.   See examples in Appendix I.

3.   <u>Distinctness and Uniformity</u>

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 group s. 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 d iscriminant 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 a re 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 dat a 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 with in 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 w here 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 (for instance, set the principles of use, select which methods would be appropriate and reproducible and 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 b e 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 decisio n 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 a neasy 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 246 209/11/1996 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 documentsofinterestinpdfformat.

41. Theoptionchos enbyUPOVistheuseofSGML.SGMLisanabbreviationfor"Standard GeneralizedMarkupLanguage" whichwasfirstdefinedbyanInternationalStandardin1986 . In this system the information is provided in electronic text files and the recognition of the informationismadebytheuseoftags . Forinstance,intheUPOVROMthetag<600>isused 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 M arkup Language" is the referencefornewdevelopmentsforexchangeofinformation(descriptionoftheinformationand exchange), while HTML "Hyper Text Markup Language", is used to describe the display of information on the computer screens through the int ernet. 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. Todefin eadatabasesub -modelforthepurposeofinformationexchange,anotheroptionis possible. Insuchacase,theinformationcanbeprovidedwithouttheuseoftags,asthecontent andspecificationsareknownbythedefinitionofthedatabasemodel . Thisapproachisprobably betteriftheaimistoreachharmonizationandencouragecooperativework,asitallowseasier exchange,andthepossibilitytointerconnectdifferentdatabasesatrealtime .

43. Examplesofhowinformationcanbekeptin differentwaysisillustratedinAppendix2.

44. Inbothcases,(Markuplanguageordatabasesub -model)wheninformationisexchanged and grouped, the identification of the objects is essential in order to be able to search for informationcomin gfromdifferentsources,ortorelateobjects(varietiesforinstance)studiedin differentplaces. UPOVhasalreadydefinedsuchidentificationcodesforthecharacteristicsin the UPOV Test Guidelines . By contrast, the variety code is given by each co untry, and no databaselinkisprovidedintheUPOVROMtoidentifyifagivenvarietyhasbeendescribedby oneorseveralcountries . Theuseofspeciesnamecanhelptoidentifysuchcases,butthisisleft totheuseofthedatabase . Thecodingofthe speciesisnotyetharmonizedbyUPOValthough someworkhasbeendoneinthatdirection.

45. Itisprematuretoimaginehowbio -molecularresultswouldbeinputinanadministrative and technical database, as these results are not yet used by U POV members for DUS . Nevertheless,twopossibleusescanbeimaginedinthefuture. Thefirstonehasalreadybeen usedforelectrophorosis,whereUPOVdefined *ad hoc* characteristicsintheguidelinesforthe corresponding species. With this approach, t here 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 beforeafinal"note"i sgiven.Inthiscase,foreachvarietythereisasetofvalues(i.e . measure on60individualplantsfrom3differenttrials) . Wecanalreadyimagineinsuchacasethatat least the lot, the sample, the material used, the test, the locus, the allele, the frequency, will enableustocopewithmost,ifnotall,ofthesituationswherewewillhavetocomputecriteria fordistinctnessanduniformity.

APPENDIXI

46. Examplesof3similaritymeasures,eachonthesamedataset,andwiththeadd   itionof6 monomorphicbands,follow:

CorrelationBetweenOriginal'Raw'Data(10bands)&+Mono(10+6monomorphicbands)

NL	0.9444
J	0.9496
RT	0.9990

CorrelationBetweenSimilarityMethods'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

NL:-Nei&Li
J:-Jaccard's
RT:-RogerandTanimoto
NL-M(Nei&Libasedon'Raw'+6Monomorphicsbands)
J-M(Jaccard'sbasedon'Raw'+6Monomorphicsbands)
RT-M(RogersandTanimotobasedon'Raw'+6Monomo   rphicsbands)

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

	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	

JACCARDS   (MONO   BANDS   ADDED)

	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	

ROGERS and TANIMOTO   (MONO   BANDS   ADDED)

	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	

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	

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	

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	

APPENDIX2

STORAGEOFINFORMATI ONINDIFFERENTWAYS :

<u>EXTRACTFROMFILESENTTO UPOVFORCREATIONO FTHEUPOV -ROM</u>(SGML
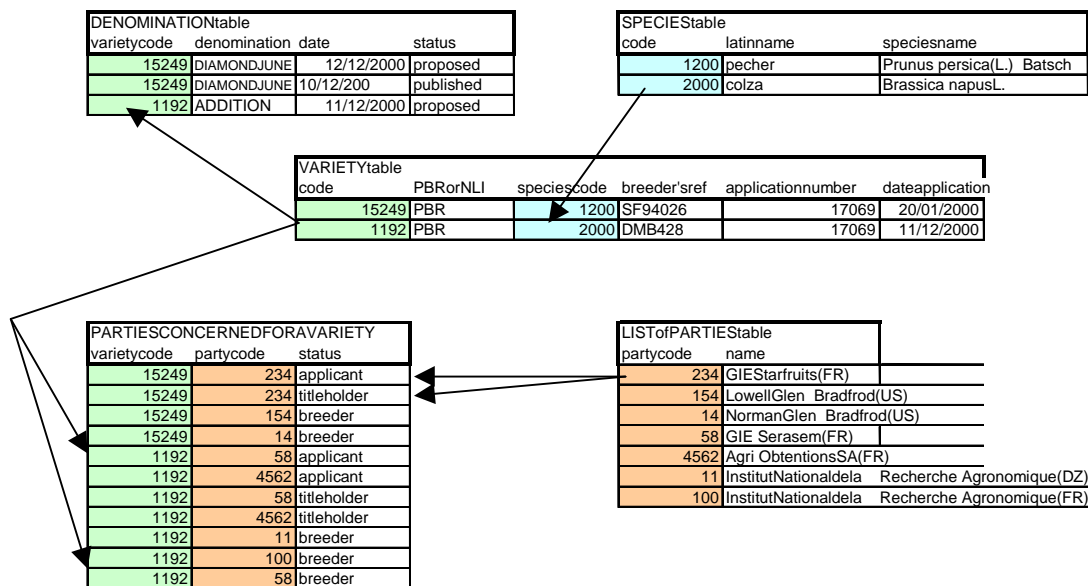FORMAT,INFORMATIONS AREIDENTIFIEDBYT AGSDESCRIBEDINCIR CULARU
2462)
ForinstanceTag190istoidentifyinformation"countrysendingtheinformation"
```
<000>0
<190>FR
<010>PBR 15249
<500>Prunus persica (L.) Batsch
<510>Pecher
<540>20001212 DIAMOND JUNE
<541>20001210 DIAMOND JUNE
<600>SF 94 026
<210>17069
<220>20000120
<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)
```

<u>Exampleofinformationkeptinspreadsheet</u> : Whenthereismorethanonesetofinformation
kept for a variety, the maintenance of the file can become a problem, e.g. there are two co -
breedersinvarietyDIAMONDJUNE;2co -applicants,3co -breeders,2co -titleholdersforthe
BrassicanapusL.variety.

PBRor NLI	Identification number	species name	Latinname	date denomination proposed	denominationproposed	date denominationnaccepted	denomination accepted	breeder's reference	applicationnumber	date application	publication ofdate …	Applicant's name	Breeder'sname	Titleholder's
PBR	15249	Pecher	Prunus persica(L.) Batsch	12/12/2000	DIAMOND JUNE	02/10/2000	DIAMOND JUNE	SF94026	17069	20/01/2000	02/10/2000	GIEStar fruits(FR)	LowellGlen Bradford(US) NormanGlen Bradford(US)	GIEStar fruits(FR)
"	"	"	"	"	"	"	"	"	"	"	"			"
PBR	1192	Colza	Brassica napusL.	11/12/2000	ADDITION		ADDITION	DMB428	17515	11/12/2000		GIE Serasem(FR)	InstitutNationaldela Recherche Agronomique(DZ)	GIESerasem (FR)
PBR	1192	Colza	Brassica napusL.	11/12/2000	ADDITION		ADDITION	DMB428	17515	11/12/2000		Agri Obtentions SA(FR)	InstitutNationaldela Recherche Agronomique(FR)	Agri obtentions SA(FR)
PBR	1192	Colza	Brassica napusL.	11/12/2000	ADDITION		ADDITION	DMB428	17515	11/12/2000			GIESerasem (FR)	

Example of information kept in a database : To avoid redundancy and to ease ma intenance, 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.

DENOMINATIONtable

varietycode	denomination	date	status
15249	DIAMONDJUNE	12/12/2000	proposed
15249	DIAMONDJUNE	10/12/200	published
1192	ADDITION	11/12/2000	proposed

SPECIEStable

code	latinname	speciesname
1200	pecher	Prunus persica(L.) Batsch
2000	colza	Brassica napusL.

VARIETYtable

code	PBRorNLI	speciescode	breeder'sref	applicationnumber	dateapplication
15249	PBR	1200	SF94026	17069	20/01/2000
1192	PBR	2000	DMB428	17069	11/12/2000

PARTIESCONCERNEDFORAVARIETY

varietycode	partycode	status
15249	234	applicant
15249	234	titleholder
15249	154	breeder
15249	14	breeder
1192	58	applicant
1192	4562	applicant
1192	58	titleholder
1192	4562	titleholder
1192	11	breeder
1192	100	breeder
1192	58	breeder

LISTofPARTIEStable

partycode	name
234	GIEStarfruits(FR)
154	LowellGlen Bradfrod(US)
14	NormanGlen Bradfrod(US)
58	GIE Serasem(FR)
4562	Agri ObtentionsSA(FR)
11	InstitutNationaldela Recherche Agronomique(DZ)
100	InstitutNationaldela Recherche Agronomique(FR)

References

Clifford, H.T. and Stephenson, W. (1975). An introduction to numerical classification. AcademicPress,NewYork.

Engqvist and Becker (1995). Genetic diversity for allozymes, RFLPs and RAPDs in resynthesizedrape.Proceedingsofthe9 [th]EUCAR PIAsectioninPlantBreeding,6 -8July1994, Wageningen.

Link,W.,Dixkens,C.,Singh,M.,Schwall,M.andMelchinger,A.E.(1995).Geneticdiversity inEuropean andMediterraneanfababean germplasmrevealedbyRAPDmarkers. TAG90: 27-32.

Loarce, Y., Gallego, R. and Ferrer, E. (1996). A comparative analysis of the genetic relationshipbetweenryecultivarsusingRFLPandRAPDmarkers.Euphytica.88:107 -115

Melchinger, A.E., Messmer, M.M., Lee, M., Woodman, W.L. and Lamkey, K.R. (1991). Diversity andrelationships amongU.S.maizeinbredsrevealedbyrestrictionfragmentlength polymorphisms.CropScience.31:669 -678.

Nei,M.(1972).Geneticdistancebetweenpopulations.AmericanNaturalist.106:283 -292.

Nei,M.(1987).Molecularevolu tionarygenetics.ColumbiaUniversityPress,NewYork.

Nei,M.andLi,W -H.(1979). Mathematicalmodelforstudyinggeneticvariationintermsof restrictionendonucleases.ProceedingsoftheNationalAcademyofSciences(USA).76:5269 - 5273.

O'Donoughue,L.S.,Souza,E.,Tanksley,S.D.andSorells,M.E.(1994).Relationshipsamong 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).Geneticdataanalysis.Sinauer,Sunderland.

[Endofdocument]