E

UPOV

BMT/6/8
ORIGINAL: English
DATE: February 7, 2000

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS

GENEVA

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

Sixth Session
Angers, France, March 1 to 3, 2000

RAPDS MATHEMATICAL ANALYSIS TO ESTABLISH RELIABILITY OF VARIETAL ASSIGNMENT IN VEGETATIVELY PROPAGATED SPECIES

prepared by experts from Spain

RAPDS MATHEMATICAL ANALYSIS TO ESTABLISH RELIABILITY OF VARIETAL ASSIGNMENT IN VEGETATIVELY PROPAGATED SPECIES

J. Ibañez
Instituto Madrileño de Investigacion Agraria y Alimentaria (IMIA). Apdo. 127. 28800 Alcala de Henares. Spain.

Abstract

An unknown grapevine sample (US) and a seedless grape variety (V) were compared by RAPDs analysis using 122 different dekamer oligonucleotides. No differences were found between them after studying 1741 bands. In order to determine the reliability of the assignment of US to V, two mathematical analysis have been developed. In the first one, individual bands are used as comparison unit. The probability of finding at least one non-shared band between V and US can be estimated as $P= 1- S^N$, being S a similarity coefficient, and N the number of bands. In the example, the calculated similarity between V and US is higher than 0.9974 with a P=0.9999, while the highest similarity found among a set of 45 seedless grape varieties was 0.9250. For the second approach, the complete RAPDs pattern obtained with each primer for V was compared with those obtained for the other 44 seedless grape varieties, and the number of non-differentiated pairs was established for each of the 10 primers used. Then, the probability of coincidence was calculated for these 10 primers ($2.3 \cdot 10^{-13}$) and for the 122 used between S and V ($7.7 \cdot 10^{-155}$). Both mathematical analyses clearly demonstrate that US belongs to the variety, and these analysis are applicable to other asexually propagated plants.

Keywords: *Vitis vinifera*, grapevine, similarity, probability-of-coincidence, legal-protection, patented-variety, breeders'-rights, intellectual-property.

1. Introduction

Assignment of a plant to a certain variety may be a natural, easy process in many species. In fact, many people are able to recognise simply with the naked eye the varieties they work with. The process is especially simpler in the case of vegetatively propagated species, since all the individuals belonging to a given variety have the same genotype. Nevertheless, when there are legal implications, like in the defence of patented varieties, accurate identification of cultivars is essential, and this applies especially to horticultural industry. In order to protect a newly developed variety through plant protection and patent laws, the owner of the patent must be able to clearly distinguish the protected variety from all other cultivars at the marketplace. For this purpose, morphological characters have been historically used, but molecular markers are currently being incorporated due to its superior power of discrimination. Assignment of a plant to a protected variety by an expert may result in legal disputes, where the producers will normally reject the assignment. It becomes necessary a mathematical analysis indicating the reliability of such assignment.

RAPDs technique (Random Amplified Polymorphic DNAs) is widely used for varietal identification, because it is technically simple and requires no previous genomic information (Williams et al., 1990). Nevertheless, the type of result obtained through this technique makes difficult its mathematical analysis. There is a huge amount of scientific literature concerning forensic uses of molecular markers and statistical approaches in humans. Nevertheless, no information could be found about the application of such mathematical analysis related to the protection of new varieties of plants through the RAPDs technique. As in humans, while

demonstrating that two DNA samples come from different individuals is relatively easy, difficulties arise when identity has to be proved. If an unknown sample (US) shows the same DNA profiles than a given variety (V), such coincidence could be casual, or could be due to the fact that US actually belongs to the variety in question. *A priori*, both alternatives are always possible and obtaining an estimation of their respective probabilities is convenient. In humans, the result is present in the form of a likelihood ratio (LR), which applied to the present case would be the ratio of the probabilities of the result obtained given that the US belongs (R/V) or not belongs to the variety (R/NV): $LR = \frac{P(R / V)}{P(R / NV)}$

In this study, a sample of grape is compared to a given variety and two mathematical analyses are used to test the reliability of the assignment of the sample to that variety. The analysis by primers allows the calculation of a likelihood ratio, while the analysis by bands gives the minimum similarity that both samples should share with a given confidence level.

2. Materials and methods

Plants of grapevine (*Vitis vinifera* L.) producing seedless table grapes have been used. An unknown sample (US) was compared with plants of the variety 'Sugraone' (also known as 'Superior seedless'), a legally protected variety. Also 45 seedless grape varieties, including 'Sugraone', were studied as a reference. The list of varieties is available on request.

Genomic DNAs were extracted from young leaves according essentially to Lodhi et al. (1994). RAPD reactions were performed in a volume of 25µl containing: 10mM Tris-HCl pH 8.3; 50mM KCl; 200µM of each dNTP; 5 mM $MgCl_2$; 0.4 µM primer; 2 u. Amplitaq DNA Polymerase, Stoffel Fragment (Perkin Elmer) and 25 ng. of genomic DNA. PCR program consisted of a previous denaturation step of 6 minutes at 94°C followed by 40 cycles of [94°C 60 seconds, 35°C 60 seconds, 35°C to 72°C 90 seconds and 72°C 6 minutes and 30 seconds]. The thermocycler used was a PTC-100 with hot bonnet, from MJ Research.

For the direct comparison between V and US 140 different oligonucleotide primers (10-mers) were used: A, D, E, H, I, O and S complete series from Operon Technologies (Alameda, CA). The group of 45 varieties was studied with 10 randomly chosen primers (Table 2). Amplification products were separated on 2% (P/V) agarose gels and stained with ethidium bromide. Gels were visualised under UV illumination, and the corresponding images were captured with an image analyser. Two different plants from each sample were independently analysed, and when existing doubts between V and US, another analysis was performed.

2.1. Mathematical analysis of coincidence

2.1.1 Analysis by bands

In the comparison by RAPDs analysis of two samples, I and J, one can find 'a' bands shared by the two samples (+/+), 'b' bands that appear only in the sample I (+/-) and 'c' bands that appear only in the sample J (-/+). The frequency of (+/+) bands in the whole number of bands present in I (a+b) plus J (a+c) is: $\frac{a + a}{(a + b) + (a + c)}$

This value is the Dice similarity coefficient (Dice, 1945, cited by Rohlf, 1990) and represents the probability that, choosing at random a band among all the possible ones (from I

plus from J), this band was of the type (+/+). Then, for N bands, the probability that all being of the type (+/+) is: $\left(\frac{2a}{2a+b+c}\right)^N = S^N$

This is the probability of finding no band distinguishing I and J after studying N bands. Thus, the probability of finding at least one different band between I and J studying N bands, if S is the actual similarity between both, is: $P = 1 - S^N$

Then, if no difference can be detected, it means that $S > \sqrt[N]{(1-P)}$

In other words: if, after studying N bands, it cannot be found any difference between I and J, then the similarity between them (measured with the Dice coefficient) should be at least S, with a probability P.

Pairwise Dice similarity coefficients were calculated for the 45 varieties using NTSYS-pc software (ver. 1.6) (Rohlf, 1990) from RAPDs data obtained with 10 primers. These similarity values were used as a reference for the theoretical values of S obtained.

2.1.2 Analysis by primers

This analysis allows the calculation of likelihood ratios. For that, it is necessary determining two alternative probabilities: the probability of the result if US belongs to the variety: P(R/V) and the probability of the result if US does not belong to the variety: P(R/NV). If US belongs to the variety, the probability of obtaining identical DNA profiles is: P(R/V)=1. Then, it is necessary calculate P(R/NV), i.e., to establish the probability that any other seedless grape variety show identical RAPDs pattern than V with a number M of primers. The individual probabilities of coincidence (C) between V and the other 44 varieties for each primer have been established simply by counting the number of varieties that show the same pattern than V and dividing by the total number of comparisons (44). These probabilities of coincidence have been also calculated for the whole group of varieties (990 different pairwise comparisons for each primer).

Assuming independence among RAPDs patterns obtained with different primers, the global probability of coincidence for the 10 primers is: $P_1(R/NV) = \prod_{m=1}^{10} C_m$

US was compared with V using a much higher number of primers. If the variability detected with the ten primers randomly chosen is representative of what it could be observed with the 140 primers, the results could be extrapolated. For that, the average of the probabilities of coincidence was calculated: $\overline{C} = \frac{\sum_{m=1}^{10} C_m}{10}$

Thus, the global probability of coincidence for M primers is: $P_2(R/NV) = \overline{C}^M$

3. Results

Consistent, reproducible DNA profiles were obtained with 122 out of the 140 primers used in the comparison between US and V. Only the results from these 122 primers were considered. ‘Sugraone’ and US showed identical patterns with every primer. A total of 3482 bands were considered: 1741 each sample, about 14 bands/primer. Only the bands that appeared in the two independent analysis carried out with each sample were considered.

3.1 Analysis by bands

Table 1A shows probability values obtained for different theoretical similarities after analysing 3482 bands. For instance, if the actual similarity between V and US is 0.9990, the probability of finding at least one different band between them will be 0.9693. In other words, if the similarity between US and V is equal or lower than S, we should find at least one different band between them with a probability P. Then, if we cannot find such different band is because the actual similarity is higher than S, with a probability P.

The average Dice similarity coefficient value found among the 45 seedless grape varieties was 0.817, being 0.925 the highest one (between 'Black seedless' and 'Sultanina'). 'Sugraone' presented the highest similarity with 'Blush seedless' (0.894) and the average value from its pairwise comparisons with the other 44 varieties was also 0.817. For these two maximum values of similarity (0.925 and 0.894) and studying 3482 bands, the probability of finding one or more different bands is 1 (with more than 30 significant decimals).

Table 1B shows another focus on the same question, determining the minimum similarity for different *a priori* significance levels (α=1-P). For instance, as no different band has been detected after studying 3482 bands, the similarity between V and US must be higher than 0.9974, with a confidence level of 99.99%.

3.2 Analysis by primers

V and US have the same DNA profiles. Since this is the expected result if US belongs to the variety, P(R/V)=1.

To calculate P(R/NV), the proportion of non-differentiated pairs in the comparisons of the RAPD profiles has been established for each primer, in the whole group of seedless varieties (Table 2, left) and with respect to 'Sugraone' (Table 2, right).

The global probability of coincidence for the whole group of varieties with the ten primers used is $P_1(R/NV)=1.3\cdot10^{-16}$ ($6.1\cdot10^{-15}$, using the upper 95% confidence limits). Then, the likelihood ratio is: $LR_1=7.7\cdot10^{15}$ ($LR_1^{95}=1.6\cdot10^{14}$). This means that, in general, if we found two seedless grape samples that show identical RAPDs profiles with 10 primers, it is $7.7\cdot10^{15}$ times more probable that both belongs to the same variety than they belongs to different varieties. In the case of the comparisons with 'Sugraone', $P_1(R/NV)=0$ because the profiles obtained for 'Sugraone' are unique among the varieties studied with three of the ten primers (D03, E02 and H13), and then three of the multiplicands are 0.

An alternative is to establish the average of the single probabilities of coincidence: $\overline{\mathbf{C}}$ (Table 2). The probability that other variety presents the same RAPDs profiles than 'Sugraone' with these ten primers is:

$$P_2(R/NV) = (0.0545)^{10} = 2.3\cdot10^{-13} \Rightarrow LR_2 = 4.3\cdot10^{12}$$

A conservative approach, using the upper 95% confidence limit for $\overline{\mathbf{C}}$, gives a value of $P_2^{95}(R/NV)=1.3\cdot10^{-10}$ for these ten primers, and $LR_2^{95}=7.8\cdot10^{9}$.

If the result obtained with ten primers is extrapolated to the number of primers used for the direct comparison between US and V, the probability of coincidence decreases to $7.7\cdot10^{-155}$, and consequently the likelihood ratio raises up to $1.3\cdot10^{154}$.

4. Discussion

RAPDs analysis shows that the unknown sample and the variety share the same DNA profiles with every primer used. Mathematical analyses, by bands and by primers, indicate that US can be assign to the variety beyond doubt.

The highest similarity found among the group of seedless varieties (0.925) is quite far from 0.997, the minimum value between V and US with a probability of 99.99%. The distance is significant since the collection studied include several groups of very close varieties: brothers (like 'Rodi', 'Sultana moscata', 'María Pirovano' and 'Basile Logothetis, or like 'Perlette', Delight' and 'Bruni 116') and parents and progeny (like 'Perlette' and 'Perlona' or 'Ruby seedless' and 'Marroo seedless' or 'Sultanina' and at least 12 of the varieties). The distance was greater with respect to the highest value found for the comparison of 'Sugraone' and the rest of varieties (0,894), even if a seeded variety named 'Cardinal', its only known parent, is included: 0,833 (data not shown).

Two aspects should be taken in consideration with respect to the analysis by bands: the independence of the bands and the possible existence of a little number of different bands between US and V. With respect to the first point, the probability determinations as well as the similarity coefficient calculations require band independence, which is normally ignored. It is very difficult to know how many bands are linked in all the DNA profiles studied, but the analysis is very robust: although only half of the bands considered (1740) were independent, similarity would be higher than 0.997, with a probability of 0.99; and using only 500 bands (250 each sample, about two bands/primer) S=0.991 with a P=0.99. (This also implies that, ignoring the question of bands independence, less than 20 primers would be needed to reach the last cited values of similarity and probability). The second point is very important, since a few different bands can sometimes appear between different plants of the same asexually propagated variety. These differences are caused by somatic mutations, which are the base of clonal selections within a variety, and also can be responsible for essentially derived varieties (EDV). In the previous analysis, we established the probability of finding one or more distinct bands (DB) between US and V (P(DB≥1)), but we can calculate the probability of finding more than one (or two, three, etc.): P(DB≥2) = 1 - P(DB=0) - P(DB=1) = P(DB≥1) - P(DB=1), and $P(DB=1)=\binom{3482}{1}\cdot S^{3481}\cdot(1-S)^{1}$

Thus, for a similarity value of 0.995, the probability of finding more than one different band after studying 3482 bands is 0.99999951. If only one different band is found, then S must be higher (with that confidence level) and, since this value is quite far from 0.925, it can still be concluded that both samples come from the same original embryo. Analogous calculations can be done for two, three or more different bands. Obviously the strength of the analysis diminishes, but it can still be powerful enough: for instance, for a similarity of 0.995, the probability of finding 4 or more distinct bands is 0.99997203.

The analysis of the RAPDs patterns by primers instead of by bands is not new: Tessier et al. (1999) established the frequencies of the different RAPDs profiles in a group of 224 varieties to calculate the confusion probability (C) for each primer, and then its discrimination power (D=1-C). This confusion probability is exactly the probability of coincidence defined here. And these probabilities of coincidence have showed that 10 randomly chosen primers are enough to assign US to the variety. This fact is due to the high variability found in grapevines and, very probably, more primers would be needed in other species to reach the very high likelihood ratios found here. In such a case, the extrapolation from 10 to 122 primers would be an important point. The base for that extrapolation was the choice of the ten primers: at random, instead of selecting the most discriminating ones. Probably, a number of primers higher than 10 would be more convenient in other cases, but again the use of 122 primers gives a huge robustness to the analysis: even if the average probability of coincidence with 'Sugraone' was very high, for instance 0.9 (more than 16 times higher than the probability found, 0.0545), the likelihood ratio of the analysis would still be very high: 382,308. Evett (1987) considered that a likelihood ratio above 1000 is a 'very strong evidence' in humans (and there are more human beings than grapevine varieties!).

Finally an important consideration about the group of varieties used as reference. It is clear that the selection of the varieties can greatly influence the result of the analysis: if the varieties are not related to V and among them, similarities and probabilities of coincidence will be lower and the mathematical analysis will be artificially stronger. Here, a very specialised group of varieties has been chosen: only those producing seedless grapes. The average similarities obtained for the whole group and for the group with respect to V are identical (0.817) while, for instance, that found in a variety of other species of the same genus (*Vitis rupestris* du Lot) was quite lower (0.561, data not shown); likewise, the average probabilities of coincidence are also very similar: 0.0534 and 0.0545 respectively. These data and the presence of close relatives in the group make of the selected varieties a completely suitable group for the pursued aim.

Acknowledgements

I wish to thanks Dr.J. Hornero for critically reviewing the manuscript.

References

Evett, I. 1987. Bayesian inference and forensic science: problems and perspectives. The Statisticians 36: 99-105

Lodhi M.A., Ye G.N., Weeden N.F. and Reisch B.I. 1994. A simple and efficient method for DNA extraction from grapevine cultivars and *Vitis* species. Plant Molec. Biol. Rep. 12: 6-13.

Rohlf F.J. 1990. NTSYS-pc, Numerical Taxonomy and Multivariate Analysis System. Version 1.6. Exeter Software. Setauket. New York

Tessier C., David J., This P., Boursiquot J.M. and Charrier A. 1999. Optimization of the choice of molecular markers for varietal identification in *Vitis vinifera* L. Theor. Appl. Genet. 98: 171-177.

Williams J.G.K., Kubelic A.R., Livak K.J., Rafalsky J.A. and Tingey S.V. 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res. 18: 6531-6535.

Table 1- A: Calculation of probability values for different theoretical similarities between V and US from 3.482 bands. B: Calculation of similarity values between V and US from 3.482 bands for different theoretical probabilities.

A		B	
S =0.9250	P=1.00000000	P=0.9500	S>0.99914002
S =0.9500	P=1.00000000	P=0.9900	S>0.99867831
S =0.9900	P=1.00000000	P=0.9950	S>0.99847953
S =0.9950	P=0.99999997	P=0.9990	S>0.99801812
S =0.9990	P=0.96930766	P=0.9995	S>0.99781947
S =0.9995	P=0.82473136	P=0.9999	S>0.99735837

Table 2 - Results of the pairwise comparisons of the RAPDs profiles among all the seedless grape varieties (Total) and between 'Sugraone' and each of the remain 44 varieties. C_m represents the proportion of identical patterns for each primer and $\overline{C}$ the average of C_m values. C_m^{95} columns show the upper 95% confidence limits for C_m and $\overline{C}$ values.

	Total		Sugraone	
Primer	C_m	C_m^{95}	C_m	C_m^{95}
D03	0.1121	0.1318	0.0000	0.0000
D04	0.0212	0.0302	0.0682	0.1427
D05	0.0030	0.0065	0.0227	0.0668
D19	0.0202	0.0290	0.0455	0.1070
E02	0.0010	0.0030	0.0000	0.0000
E06	0.1091	0.1285	0.1136	0.2074
E07	0.0747	0.0911	0.0227	0.0668
H13	0.0273	0.0374	0.0000	0.0000
H19	0.0293	0.0398	0.0227	0.0668
I11	0.1364	0.1577	0.2500	0.3779
Average	$\overline{C}$ = 0.0534	$\overline{C}^{95}$ =0.0845	$\overline{C}$ =0.0545	$\overline{C}^{95}$ =0.1025

[End of document]