E

TWC/21/4
ORIGINAL: English
DATE: May 14, 2003

UPOV

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

GAIA SOFTWARE: CROP EXPERT PHENOTYPIC DISTANCES BETWEEN VARIETIES

Document prepared by experts from France

GAIA SOFTWARE

CROP EXPERT PHENOTYPIC DISTANCES BETWEEN VARIETIES

1. The principle is to compute a phenotypic distance between two varieties, which is a sum of distances for individual characteristics.

2. For the difference observed between two varieties, in a given characteristic, a distance/weighting is derived from the absolute value of the difference and a metric defined for the characteristic.

3. The global distance is a sum of the distances on each characteristic:

$$dist(i, j) = \sum_{k=1,nchar} W_k(i, j)$$

where:

$dist(i, j)$ is the computed distance between variety i and variety j.

k is the k^{th} characteristic, from the *nchar* characteristics selected for computation.

$W_k(i,j)$ is a function of the difference observed between variety i and variety j for characteristic k.

OV_{ki} is the observed value on characteristic k for variety i.

$$W_k(i, j) = f\left(\left|OV_{ki} - OV_{kj}\right|\right)$$

4. For a given characteristic, a weighting is attributed to the absolute difference between two varieties. The weightings have been previously defined by the crop expert and stored in the GAIA database. The same weighting is attributed to any pair of varieties whose absolute differences between observed values are the same. If i, j, n and m are varieties.

$$\left|OV_{ki} - OV_{kj}\right| = \left|OV_{kn} - OV_{km}\right| \Rightarrow W_k(i, j) = W_k(n, m)$$

If you wish to look at a practical example first, please read Annex I.

5. The weighting is equivalent to a distance contribution. Crop experts prefer to use the word "weighting" when they consider the distance contribution on a given characteristic, and "distance" for the global distance on all characteristics.

6. The word "weighting" is not correct, but nevertheless we will use it for the *distance contribution, made by each characteristic*, in order to simplify communication and exchange between experts.

7. Weighting depends on the size of the difference and on the individual characteristic.

8. The weightings are defined by crop experts on the basis of their expertise in the crop and on a "try-and-check" learning process. The values for the weightings defined by the experts are stored in GAIA database.

9. Experts can give zero weighting to small differences. Thus, even if two varieties have different observed values in many characteristics, the resulting distance might be zero.

10. Varieties are compared in pairs. The crop expert can compare different combinations of pair-wise comparisons, for instance:

- compare two varieties,
- compare a given variety to all available varieties,
- compare all candidate varieties to all [candidate + reference] varieties,
- compare all possible combinations.

11. The crop expert can also select all the available characteristics, or different subsets of the characteristics.

12. The crop expert obtains a comprehensive report for each pairwise comparison. The software computes a global expert distance, but also provides all the individual absolute values and the distance contribution of each characteristic (see Annex III for an example).

13. The use of the results may differ from expert to expert. The most frequent use of the software in France is at present to fix and apply a threshold for the distance which enables the crop expert:

- to eliminate from subsequent growing cycles all pairs of varieties reaching or surpassing the GAIA distance threshold;
- to focus on close varieties, having a GAIA distance lower than the threshold, for the next growing cycle(s).

14. The threshold determined by the crop expert is at a level which ensures that all pairs of varieties having a GAIA distance equal or greater than the threshold are clearly distinct in the field or in the greenhouse. Therefore, they do not require further comparison in the field or in the greenhouse.

15. The threshold has to be based on experience gained with known varieties and must minimize the risk of taking a wrong decision. It would be a wrong decision to eliminate a pair of varieties which should be further compared in the field.

16. In France, greater weighting values are chosen for characteristics which are known to have polygenic control and are little influenced by environmental conditions. Monogenic controlled characteristics, or characteristics for which the level of expression is dependent on environmental conditions, are considered with care and lower values, or even a zero value is given for the weighting.

17. GAIA software computes information for the crop expert. The crop expert can use this information according to his own needs. Six cases are given below as a list of possible uses.

Case 1

18. After one growing cycle in the examination of an ornamental crop, the absolute data and distance computations are an objective way to confirm the opinion or the decision of the expert. There might be cases where pairs of varieties have a small distance, but nevertheless the expert has clear evidence of distinctness. If more growing cycles are necessary, before a decision is taken, the software helps to identify on which cases the expert will need to focus.

Case 2

19. In a "small" agricultural crop, there are relatively few candidate and reference varieties, which enables the crop expert to sow all candidates, and the appropriate reference varieties, in two or three successive growing cycles. The same varieties are sown in growing cycles 1, 2 and 3, and the layout is randomized. The software will help to identify the pairs with a small distance, to enable the expert to focus his attention on these particular cases when visiting the field.

Case 3

20. In a vegetable crop, there are many candidate and reference varieties. There is wide variability in the species, so on the one hand there are already obvious differences after only one cycle, but on the other hand some varieties are very similar. In order to be more efficient in their checks, the crop experts wish to grow "similar" varieties close to each other. The raw results and distances will help to select the "similar" varieties and decide on the layout of the trial for the next growing cycle.

Case 4

21. In a difficult crop, there are varieties which are so similar that it is common practice to make side-by-side comparisons for such varieties, identified after the first cycle. If the number of varieties in the crop is not too large, the crop expert will easily detect the cases which should be checked. However, when the number of varieties in a trial increases, it becomes less easy to identify all the problem situations. The software can help to "not miss" the less obvious cases.

Case 5

22. In vegetatively-propagated ornamentals, the examination lasts for one or two growing cycles. After the first growing cycle, some reference varieties in the trial are obviously different from all candidates, and their inclusion in the second growing cycle is not necessary. When the number of varieties is large, the raw data and distance(s) can help the expert to detect reference varieties for which the second growing cycle is unnecessary.

Case 6

23. A crop has a large number of reference varieties, and uses only qualitative characteristics on a 1 to 9 scale.

24. For characteristics where a difference of 1 note is considered to be clear evidence of distinctness, a weighting of 1 is entered in the matrices. For characteristics where a difference of 2 notes (1 and 3, 3 and 5) is considered to be clear evidence for distinctness, a weighting of 0 is entered when the notes differ by only 1 note, and a weighting of 1 is entered when there is a difference of at least 2 notes.

25. In this case the distance computed will be the number of characteristics where there is clear evidence of distinctness.

26. Each reference variety will be classified as zero distance from all other varieties of the same kind. This allows the selection of reference varieties of any kind that is needed if it is not possible to put all the reference varieties in the trial. This information is already available beforehand, and can be used to plan the first growing cycle trials as well as the subsequent growing cycles.

27. At present the software can use qualitative, quantitative and/or electrophoretic data. These types of data can be used alone or in combination, as shown in Diagram 1.

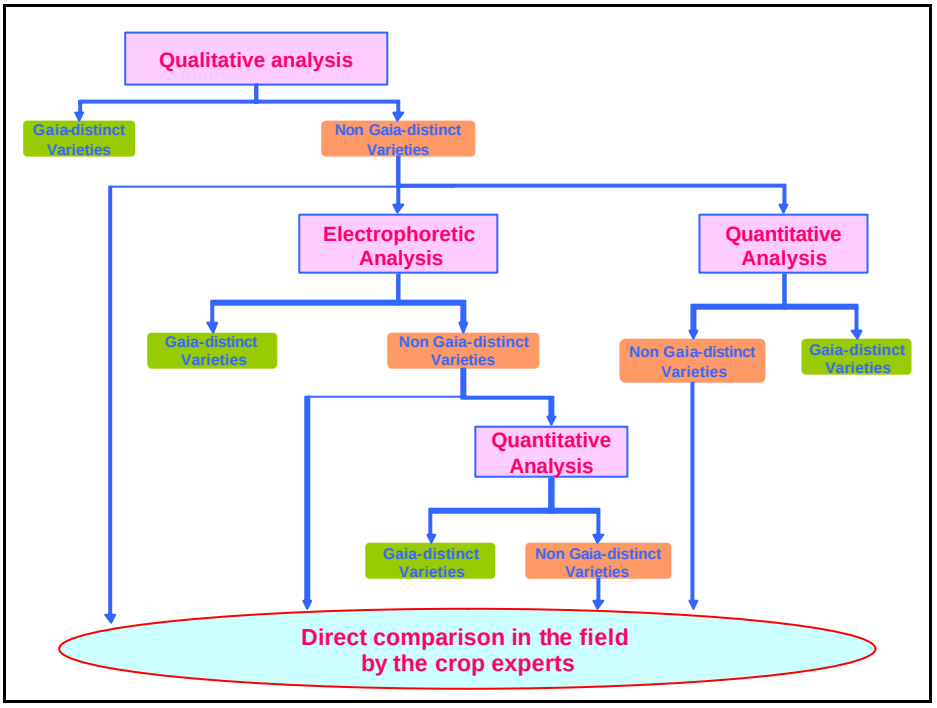


Diagram 1: Use of different types of characteristics

28. Software options change according to qualitative, electrophoretic or quantitative characteristics.

29. The user decides not only the type of data in the computation, but also the set of characteristics to use from those characteristics which are available.

30. For practical reasons, a distance threshold is used. This enables the crop expert to identify similar varieties which have a small distance (below the threshold) between them. This threshold can be used in different ways. The crop expert can use:

- a low threshold, which helps to find the more difficult cases (similar varieties);

- intermediate thresholds (different levels according to the needs);

- a large threshold when there is a need to have a comparison which uses all the available characteristics.

31. In order to minimize computation time, as soon as the threshold is achieved for a comparison between two given varieties, the software proceeds to the next pair of varieties. Remaining characteristics and their raw values will not be shown in the summary output, and will not contribute to the distance.

32. Often the crop expert looks for varieties which are similar. A low threshold is then appropriate.

33. If the crop expert wishes to see all available raw data and the different weighting for each characteristic, he must choose a threshold which is greater than the maximum distance possible on all characteristics.

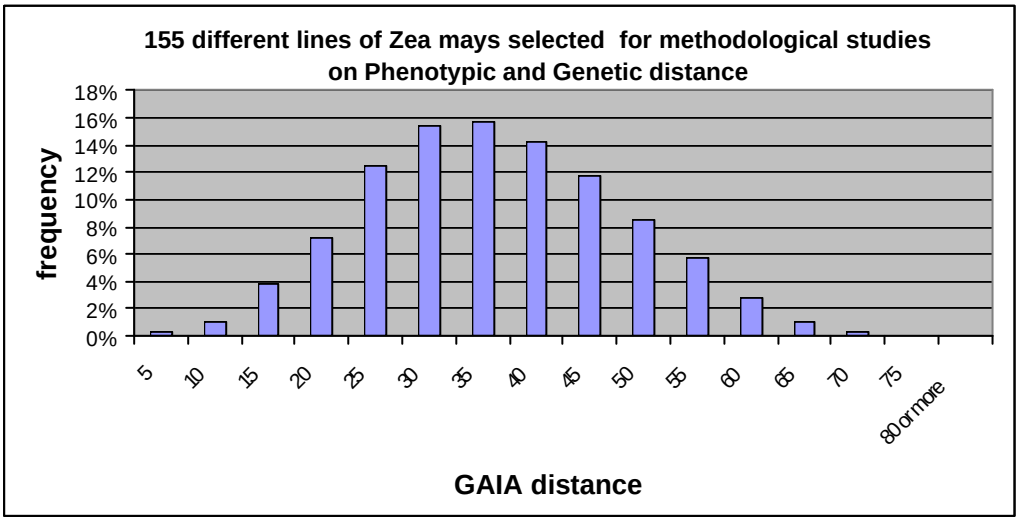
34. There is no absolute rule to decide whether a distance is "small" or "big". The crop experts themselves define the distance values.

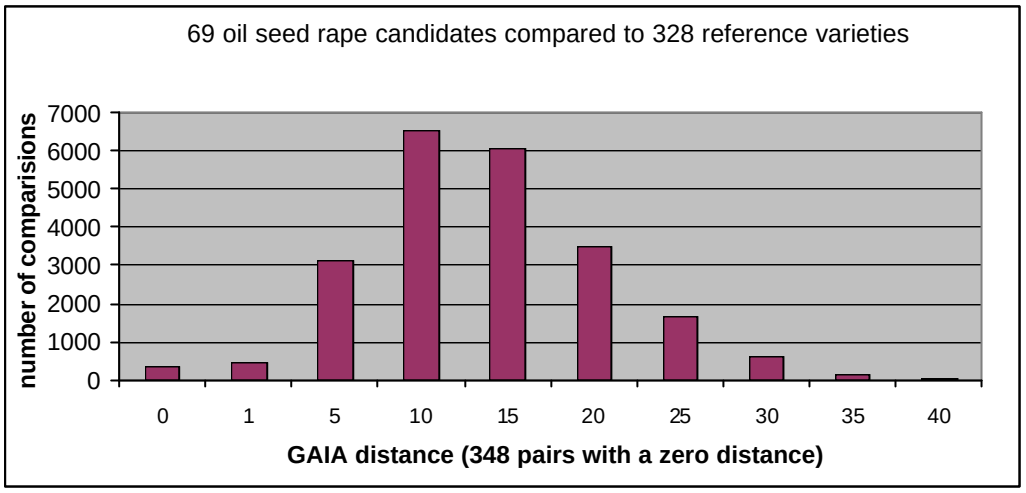
- Experts can choose different values as the weighting/distance for a characteristic (1, 2, 5, etc.).

- Some crops have more characteristics than others.

- The crop expert can use all available information, or only a subset of characteristics.

35. For these reasons the absolute values of distances vary. The same applies for the threshold.





36. The definition of the weighting "by characteristic" is necessary prior to use the software, and is important. There is no unique way to define these values; some practical considerations are described below.

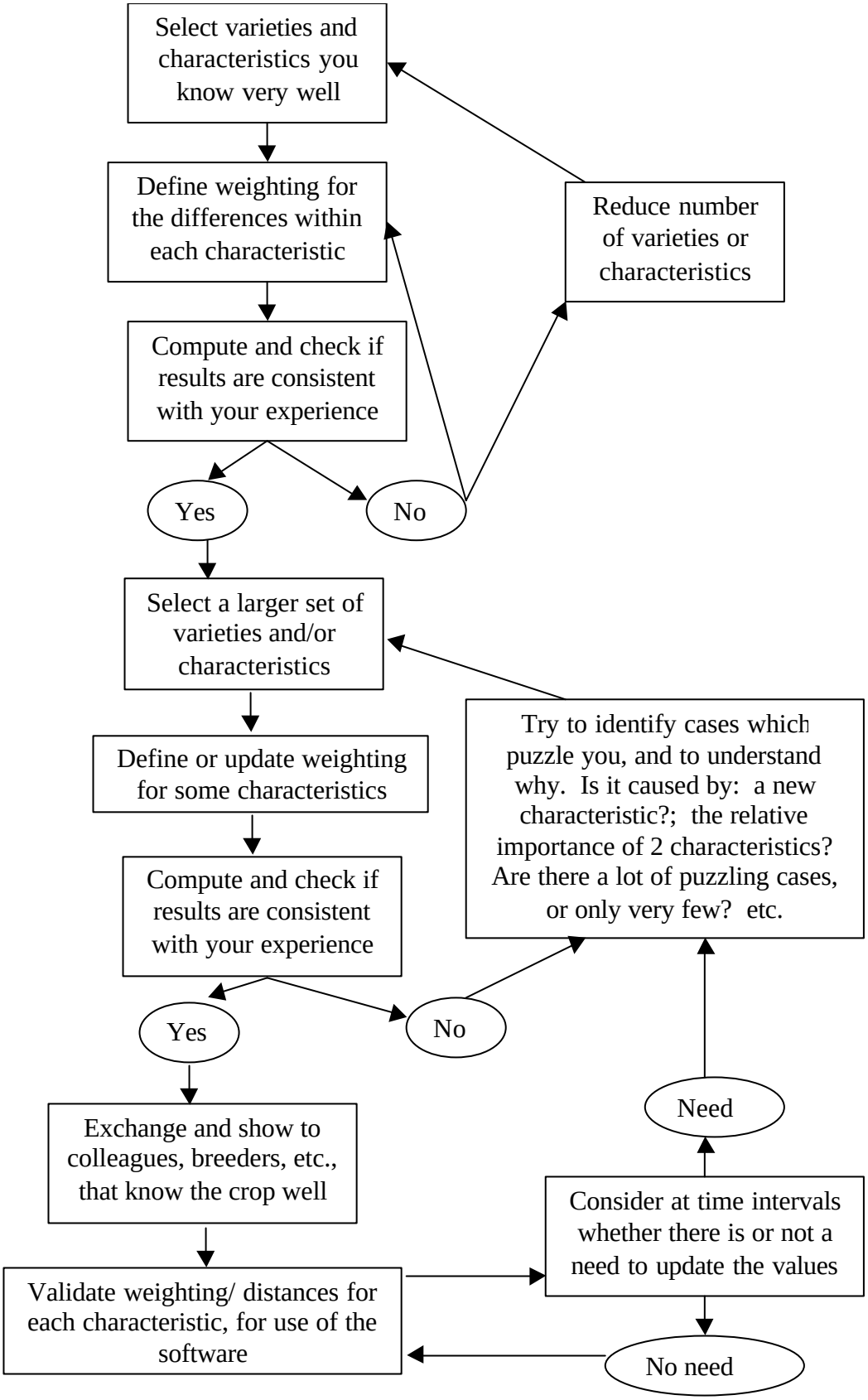
37. The two key aspects are simplicity and consistency; three simple "rules of thumb" are given here:

- the distances by characteristic should be integer values, for instance 0, 1, 2, 3, etc. where 3 is a distance or a weighting which is considered to be about 3 times greater than 1;

- if for a characteristic a given difference "expressed as an absolute value" is considered as a double distance for character *a* compared to character *b*, the distance value for this difference should be double that in character *a* than it is in character *b*;

- define the values by "try-and-check" iterations as shown in Diagram 2.

Diagram 2: "Try-and-check" process to define and revise the weightings for a crop



38. Annex I describes, a simple example of the computation of the distance between two varieties on the basis of 5 qualitative characteristics.

39. Annex II provides, in more detail, an example where successively qualitative, electrophoretic, and quantitative characteristics are used to compare two varieties.

40. Annex III provides a screen copy of a display tree which shows how the expert can navigate and visualise the results of computations.

41. A user manual and a description of the software are also available with the software (e-mail to christophe.chevalier@geves.fr).

42. The software is freely available for members of the International Union for the Protection of New Varieties of Plants (UPOV), but it is forbidden to distribute the software to other parties.

43. GAIA software has been developed with WINDEV-7.5. The general information (species, characteristics, weighting, etc.), the data collected on the varieties and the results of computations are stored in an integrated database. Import and Export facilities allow the use of your own information system in connection with GAIA software. ODBC allows access to the GAIA database and to other databases simultaneously.

44. For qualitative characteristics, 1 or 2 notes per variety can be used. In general, two notes are present when there are two trial locations. For electrophoresis data, only one description can be entered per variety. For quantitative characteristics at least 2 values (different trials, repeats, etc.) are necessary and the user selects which to use in the computation.

45. GAIA is mainly used for self-pollinated and vegetatively-propagated crops, but GAIA does not have special restrictions according to the crop.

46. A computer-based demonstration will be made during the TWC presentation of this paper. Experts interested in having more explanations, demonstrations, or asking specific questions are invited to contact Christophe Chevalier any time during the TWC meeting.

Author: Sylvain GRÉGOIRE

Thanks for review to: Joël GUIARD, Françoise BLOUET, Stéphane LASSALVY, Christelle GUITOUNI, Christophe CHEVALIER, Sally WATSON

ANNEX I

A SIMPLE EXAMPLE OF DISTANCE COMPUTATION ON 5 QUALITATIVE CHARACTERISTICS

1. The software examines differences for each characteristic and attributes the appropriate weighting. The weighting (stored in matrices in the database) is defined by the crop experts for each characteristic before the computation.

2. Weighting matrices are established by the crop experts on the basis of their expertise.

3. For a given difference in absolute values, the weighting can change according to the characteristic.

	Ear shape	Husk length	Type of grain	Number of rows of grain	Ear diameter	
Notes for variety A (1 to 9 scale)	1	1	4	6	5	
Notes for variety B (1 to 9 scale)	3	3	4	4	6	
Difference observed	2	2	0	2	1	
Weighting, according to the crop expert	*6*	*0*	*0*	*2*	*0*	*8*

Sum of weighting = Estimation of the phenotypic distance between A and B

4. In this crop, a difference of 2 notes in the absolute value is attributed:

- a weighting/distance of 6 for the characteristic Ear shape,
- a weighting/distance of 0 for the characteristic Husk length,
- a weighting/distance of 2 for the characteristic Number of rows of grain,

5. The crop experts, therefore, consider that the difference of 2 notes on “Ear shape” indicates a greater distance between two varieties than it does on “Number of rows of grain.”

6. The crop experts also consider that, for characteristic “Husk length”, note 1 for one variety and note 3 for another variety is not sufficient to indicate a distance between two varieties.

ANNEX II

EXAMPLE WITH QUALITATIVE, ELECTROPHORETIC AND QUANTITATIVE CHARACTERISTICS (*ZEA MAYS* DATA)

1. Qualitative characteristics are observed on a 1 to 9 scale. For each characteristic, weighting according to differences between levels of expression are pre-defined in a matrix of distances.

Example

2. For the characteristic “Shape of ear”, observed on a 1 to 3 scale, the crop experts have attributed weighting to differences which they consider significant:

1 = conical
2 = conico-cylindrical
3 = cylindrical

		Variety i		
		1	2	3
Variety j	1	0	2	6
	2		0	2
	3			0

3. When the crop experts compare a variety i with conical ear (noted 1) to a variety j with cylindrical ear (noted 3), they attribute a weighting of 6.

4. For the characteristic “Length of husks”, observed on a 1 to 9 scale, the crop experts have defined the weighting matrix:

1 = very short
2 = very short to short
3 = short
4 = short to medium
5 = medium
6 = medium to long
7 = long
8 = long to very long
9 = very long

		Variety i								
		1	2	3	4	5	6	7	8	9
Variety j	1	0	0	0	2	2	2	2	2	2
	2		0	0	0	2	2	2	2	2
	3			0	0	0	2	2	2	2
	4				0	0	0	2	2	2
	5					0	0	0	2	2
	6						0	0	0	2
	7							0	0	0
	8								0	0
	9									0

5. For this characteristic, the weighting between a variety i with very short husks (noted 1) and a variety j with short husks (noted 3) is 0.

6. Experts consider a difference of 3 notes is necessary in order to recognise a non-zero distance between two varieties.

7. Even if the difference in notes is bigger than 3, the experts do not increase the distance more than 2.

8. The reason for using a lower weighting for some characteristics compared to others can be that they are less "reliable" or "consistent" (e.g. more subject to the effect of the environment); and/or they are considered to indicate a lower distance between varieties.

9. A weighting matrix must be defined for each qualitative characteristic.

10. In this example, we will assume the crop expert has decided to use a distance threshold $\mathbf{S_{dist}}$ of 10 as an indicator of whether two varieties are close or not.

11. Let us take the first example with A and B observed for 5 qualitative characteristics:

	Ear shape	Husk length	Type of grain	Number of rows of grain	Ear diameter	
Notes for variety A (1 to 9 scale)	1	1	4	6	5	
Notes for variety B (1 to 9 scale)	3	3	4	4	6	
Difference observed	2	2	0	2	1	
Weighting according to the crop expert	*6*	*0*	*0*	*2*	*0*	$D_{qual} = 8$

12. In our example $\mathbf{D_{qual} = 8 < S_{dist}}$ so varieties A and B are declared "GAIA NON-distinct" and can be passed on to electrophoretic analysis.

Electrophoretic analysis

13. The electrophoretic characteristic is a homozygous allele in the UPOV Test Guidelines (see Diagram 3). The software does not allow the use of heterozygous alleles.

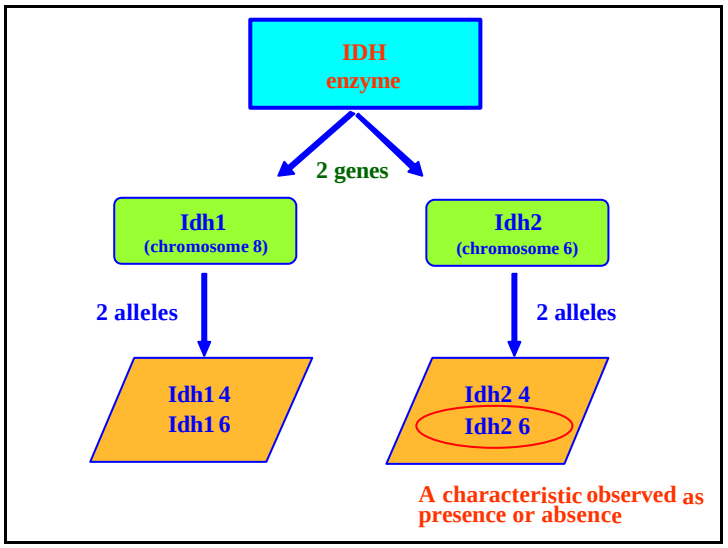
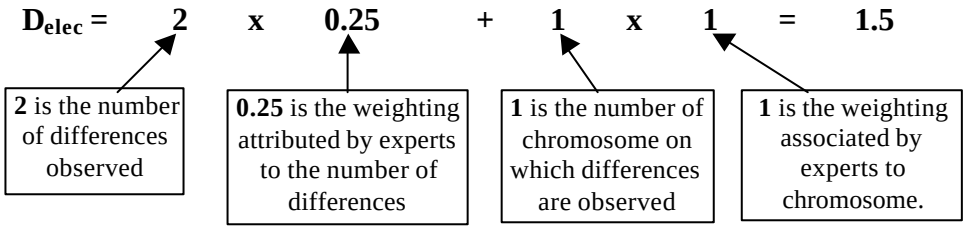


Diagram 3: The Isocitrate Deshydrogenase (IDH) enzyme has two genes (Idh1 and Idh2) located on two different chromosomes. Each of them has two alleles which are observed as 1 (presence) or 0 (absence).

14. Electrophoretic characteristics are noted 0 or 1 as absence or presence. The decision rule, used to give a weighting to two varieties, is the addition of the weighting number of differences observed and the weighting number of chromosomes related to these differences (see example below).

	Chromosome 8		Chromosome 6	
	Idh1 4	Idh1 6	Idh2 4	Idh2 6
Variety A	0	1	1	0
Variety B	0	1	0	1
Difference	0	0	1	1

15. In this example, varieties A and B are described for 4 electrophoretic characteristics: Idh1 4, Idh1 6, Idh2 4 and Idh2 6. The software looks at differences and gives the phenotypic distance using the following computation:



16. This formula, which might be difficult to understand, was established by the crop experts in collaboration with biochemical experts. Both the *number of differences* and the *number of chromosomes on which differences are observed* are used. Thus, less importance is attached to differences when these occur on the same chromosome, than when they occur on different chromosomes.

17. After qualitative and electrophoretic analysis, the phenotypic distance between varieties A and B is equal to:

$$\mathbf{D = D_{qual} + D_{elec} = 8 + 1.5 = 9.5}$$

18. The phenotypic distance is *lower than* $\mathbf{S_{dist}}$, *therefore varieties A and B are considered "GAIA NON-distinct"*.

Note: It is not possible to establish distinctness solely on the basis of electrophoretic analysis. It is necessary to have a minimal phenotypic distance in qualitative analysis in order to take into account the electrophoresis results. This minimal phenotypic distance must also be defined by crop experts. (For example, in France this value is 3 for rapeseed and 1 for maize with a distinction threshold equal to 6.)

Quantitative Analysis

19. For each quantitative characteristic, the comparison of two varieties is made by looking for consistent differences in at least two different experimental units. Experimental units are defined by the user depending on data present in the database.

20. It can, for example, be the data from two geographic locations of the first growing cycle, or 2 or 3 replications in the case of a single geographical location.

21. For a comparison to be made, the two varieties must be present in the same experimental units.

22. Differences observed must be greater than one of the two threshold values (or minimal distances), fixed by the crop experts.

- **$D_{min-inf}$** is the lower value from which a weighting is attributed,

- **$D_{min-sup}$** is the higher minimal distance. These values could be chosen arbitrarily or calculated (15% and 20% of the mean for the trial, or LSD at 1% and 5%, etc.)

23. For each minimal distance a weighting is attributed:

- **$D_{min-inf}$** a weighting P_{min} is attributed;

- **$D_{min-sup}$** a weighting P_{max} is attributed;

- the observed difference is lower than **$D_{min-inf}$** a zero weighting is associated.

24. Varieties A and B have been measured for characteristics “Width of blade” and “Length of plant” in two trials.

25. For each trial, and each characteristic, the crop experts have decided to define **$D_{min-inf}$** and **$D_{min-sup}$** by calculating respectively the 15% and 20% of the mean for the trial:

	Width of blade		Length of plant	
	Trial 1	Trial 2	Trial 1	Trial 2
$D_{min-inf}$ = 15% of the mean	1.2 cm	1.4 cm	28 cm	24 cm
$D_{min-sup}$ = 20% of the mean	1.6 cm	1.9 cm	37 cm	32 cm

26. For each characteristic: the crop experts have attributed the following weighting:

A weighting **P_{min}** = 3 is attributed when the difference is greater than **$D_{min-inf}$**.

A weighting **P_{max}** = 6 is attributed when the difference is greater than **$D_{min-sup}$**

	Width of blade		Length of plant		
	Trial 1	Trial 2	Trial 1	Trial 2	
Variety A	9.9 cm	9.8 cm	176 cm	190 cm	
Variety B	9.6 cm	8.7cm	140 cm	152 cm	
Difference	0.3 cm	1.1 cm	36 cm	38 cm	
Weighting according to the crop expert	0	0	3	6	D_{quan} =?

27. In our example, for the characteristic "Width of blade", the differences observed are lower than $\mathbf{D_{min\text{-}inf}}$, so no weighting is associated.

28. On the other hand, for the characteristic "Length of plant" one difference is greater than the $\mathbf{D_{min\text{-}inf}}$ value and the other is greater than the $\mathbf{D_{min\text{-}sup}}$ value. These two differences are attributed different weightings.

29. The user must, therefore, decide which weighting will be used for the analysis:

- minimalist option: the weighting chosen is that attributed to the lowest difference;
- maximalist option: the weighting chosen is that attributed to the highest difference;
- mean option: the weighting chosen is the mean of the others.

30. In this example, the crop experts have decided to choose the lowest of the two weightings, so the phenotypic distance based on quantitative characteristics is $\mathbf{D_{quan} = 3}$.

31. In summary, at the end of all analysis, the phenotypic distance between varieties A and B is:

$$\mathbf{D = D_{qual} + D_{elec} + D_{quan} = 8 + 1.5 + 3 = 12.5 > S_{dist}}$$

32. The phenotypic distance is greater than the distinction threshold $\mathbf{S_{dist}}$, fixed by the crop experts at 10, so varieties A and B are declared "GAIA-distinct".

33. In this example, the use of electrophoresis data "confirms" a distance between the two varieties; but on the basis of qualitative and quantitative data alone, the threshold is exceeded (8 + 3 = 11 is greater than 10).

34. If the threshold had been set at 6, the difference on the characteristic ear shape would have been sufficient, as variety A is conical and variety B is cylindrical, which is already a clear difference.

1 = conical
2 = conico-cylindrical
3 = cylindrical

Variety i			
	1	2	3
1	0	2	6
2		0	2
3			0

Quantitative and qualitative analysis on the same characteristics

35. For some crops, it is common practice to produce notes on a 1 to 9 scale for quantitative characteristics. Sometimes the transformation process is very simple, sometimes it is a complex process where all available data are used, but with a special manipulation of example varieties to adjust the raw values to the notes on the scale.

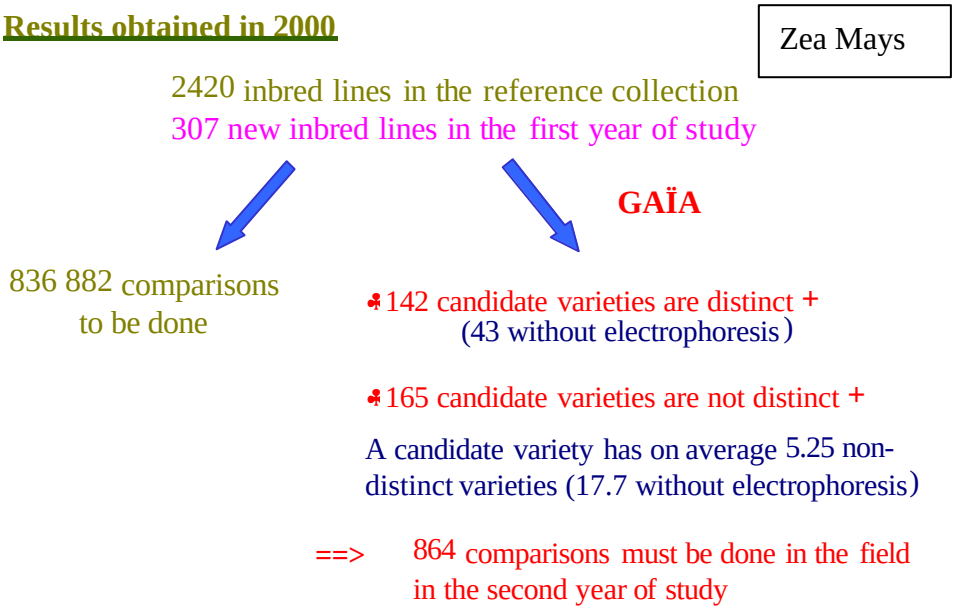
36. GAIA can include both as two separate characteristics: the original quantitative scale; and the "transformed into qualitative notes" scale. They are associated in the description of the characteristics.

37. Using the knowledge of this association, when quantitative and qualitative characteristics are both present, only one characteristic is kept, in order to avoid the information being used twice.

Conclusion of Annex II

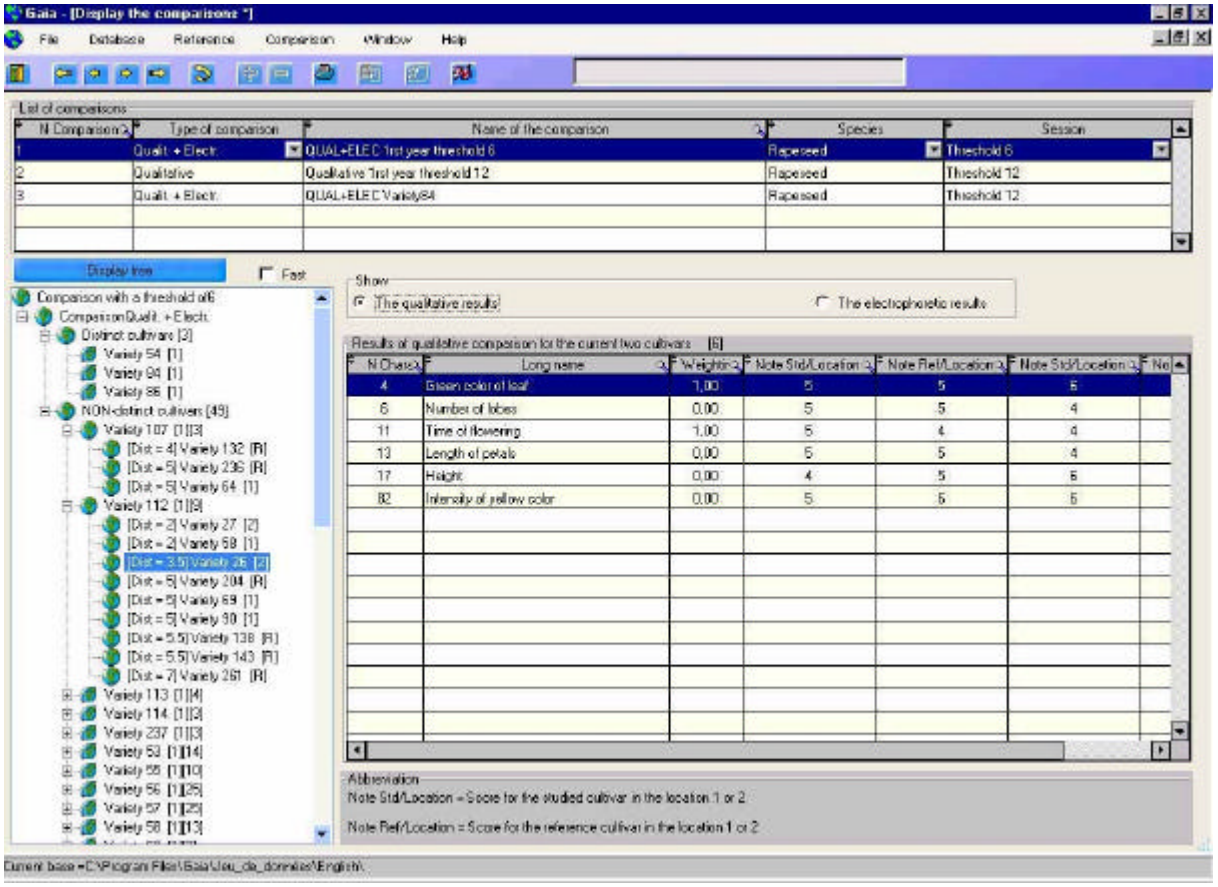
38. The above example was described in order to explain how GAIA uses different types of characteristics in a practical case.

39. The efficiency of the use of GAIA depends on the species. The following extract from the Powerpoint presentation shown at the TWC in Mexico in 2002 illustrates the potential in a crop where many years of experience are available.



ANNEX III

ANNEX III: SCREEN COPY



40. The upper part shows 3 different computations which have been kept in the database.

41. The display tree on the left shows results for a [qualitative + electrophoresis at threshold of 6] computation.

42. *Distinct cultivars [3]* demonstrates that 3 varieties were found distinct from all others. There was a total of 52 (49 + 3) cultivars in the computation.

43. The display tree is used to navigate through all possible pairs.

44. The user can expand or reduce the branches of the tree according to his needs.

45. *NON-distinct cultivars [49]*. Forty-nine cultivars were found “not distinct from all others” with a threshold of 6.

46. The first variety, *Variety 107*, has only 3 close varieties, whereas the second, *Variety 112*, has 9 close varieties, the third, *Variety 113*, 4 close varieties, etc.

47. The raw data for *Variety 112* and *Variety 26* are visible for the 6 qualitative characteristics observed on both varieties.

48. *Variety 112 [1][9]* indicates variety 112 is in the first year of examination *[1]*; and has 9 close varieties according to the threshold of 6 *[9]*.

49. *[dist=3.5]Variety 26 [2]* indicates variety 26 has a GAIA distance of 3.5 from variety 112, which is in second year of examination.

50. The third column is the weighting according to the pre-defined matrices. The notes for both varieties are displayed for the two available locations (Std stands for “studied” which are the candidate varieties).

51. In this screen copy the varieties have been numbered for sake of confidentiality, the crop experts can name the varieties according to their need (lot or application number, name, etc.).

[End of Annex III and of document]