Nereis. Revista Iberoamericana Interdisciplinar de Métodos, Modelización y Simulación	5	41-51	Universidad Católica de Valencia "San Vicente Mártir"	Valencia (España)	ISSN 1888-8550

Molecular Categorization of Yams by Principal Component and Cluster Analyses

Fecha de recepción y aceptación: 15 de febrero de 2013, 1 de marzo de 2013

Francisco Torrens Zaragozá*

Instituto Universitario de Ciencia Molecular, Universitat de València, Valencia, España
* Correspondencia: Instituto Universitario de Ciencia Molecular, Edificio del Instituto de Paterna, Universitat de València. P. O. Box 22085. 46071 Valencia. España. *E-mail*: francisco.torrens@uv.es

ABSTRACT

Wild yam species tubers of Ivory Coast are categorized by principal component analyses (PCAs) of nutrients and antinutritional factors and yams cluster analyses (CAs), which agree. Species group into three classes. Compositional PCA and yams CA allow classifying them and agree. The first PCA axis explains 90% of variance. Meta-analysis allows increasing samples numbers and data variety. Different yam behaviour depends on *energy*. Most antinutritional factors are grouped into the same class.

KEYWORDS: Principal component analysis, Meta-analysis, Distribution, Class, Tuber, Wild yam species, Antinutritional factor, Toxin, Nutrient, Individual, Variable.

RESUMEN

Especies de tubérculos de batatas silvestres de Costa de Marfil se clasifican por análisis de componentes principales (ACPs) de nutrientes y factores antinutricionales y análisis de agregados (AAs) de batatas, los cuales están de acuerdo. Las especies se agrupan en tres clases. El ACP composicional y AA de batatas permiten clasificarlos y están de acuerdo. El primer eje de ACP explica el 90% de la varianza. El metaanálisis permite aumentar el número de muestras y variedad de datos. El comportamiento de diferentes batatas depende de la *energía*. La mayoría de los factores antinutricionales se agrupan en la misma clase.

PALABRAS CLAVE: Análisis de componentes principales, Metaanálisis, Distribución, Clase, Tubérculo, Especie de batata silvestre, Factor antinutricional, Toxina, Nutriente, Individuo, Variable.

INTRODUCTION AND NOTATION

While some yam components give only a bitter taste to tubers after cooking, some others, *e.g*. alkaloids, are toxic [1]. These toxins, when they are ingested provoke grave/mortal symptoms [2,3]. In West Africa toxins of a number of wild yams were at all times exploited by hunters, fishermen and farmers [4]. Alexis and Georges determined relations between composition (14 nutrients/antinutritional factors) of nine wild yam species tubers (*cf*. Table 1) [5].

Figure 1 shows a dendrogram of eight wild yam species tubers according to nine nutrients after Alexis and Georges' classification. Tuber energy value was inversely proportional to its moisture content [6]. Individuals *D. minutiflora*, *D. hirtiflora* and *D. bulbifera* bulbil presented the highest levels of *moisture*. Species *D. burkilliana*, *D. dumetorum*, *D. bulbifera* tuber, *D. praehensilis* and *D. mangenotiana* showed high levels of *lipid*, *starch*, *proteins* and *energy*; they had the best nutritional potential. Yams *D. burkilliana*, *D. bulbifera* tuber, *D. dumetorum* and *D. praehensilis* indicated high levels of *soluble carbohydrates*. Tubers *D. minutiflora*, *D. bulbifera* bulbil, *D. mangenotiana* and *D. hirtiflora* denoted *th highest* levels of *ash*.

Table 1. Composition (antinutritional factors) of wild yam species tubers

Species	Oxalic acid (mg/100g d.m.)[a,b]	Tannins (mg/100g d.m.)	Hydrocyanic acid $\times 10^{-2}$ (mg/100g d.m.)	Alkaloid (mg/100g d.m.)	Sapogenins (%d.m.)
1. *D. minutiflora*	10.03	385.73	2.00	1.03	0.90
2. *D. hirtiflora*	6.50	15.03	7.00	107.63	1.34
3. *D. bulbifera* bulbil	9.33	470.03	1.97	165.63	0.22
4. *D. burkilliana*	5.43	489.23	4.03	187.03	0.20
5. *D. bulbifera* tuber	6.83	421.00	10.03	248.23	0.08
6. *D. dumetorum*	12.93	560.10	33.03	167.70	0.78
7. *D. praehensilis*	9.03	570.30	4.00	150.63	0.05
8. *D. mangenotiana*	8.33	410.73	20.03	175.70	0.26
9. *D. togoensis*	12.63	456.03	2.00	214.80	1.49

[a] Components: nutrients (i_1, moisture; i_2, proteins; i_3, lipid; i_4, soluble carbohydrates; i_5, starch; i_6, total carbohydrates; i_7, cellulose; i_8, ash; i_9, energy) and antinutritional factors (i_{10}, oxalic acid; i_{11}, tannins; i_{12}, hydrocyanic acid; i_{13}, alkaloid; i_{14}, sapogenins).
[b] d.m.: dry mass.

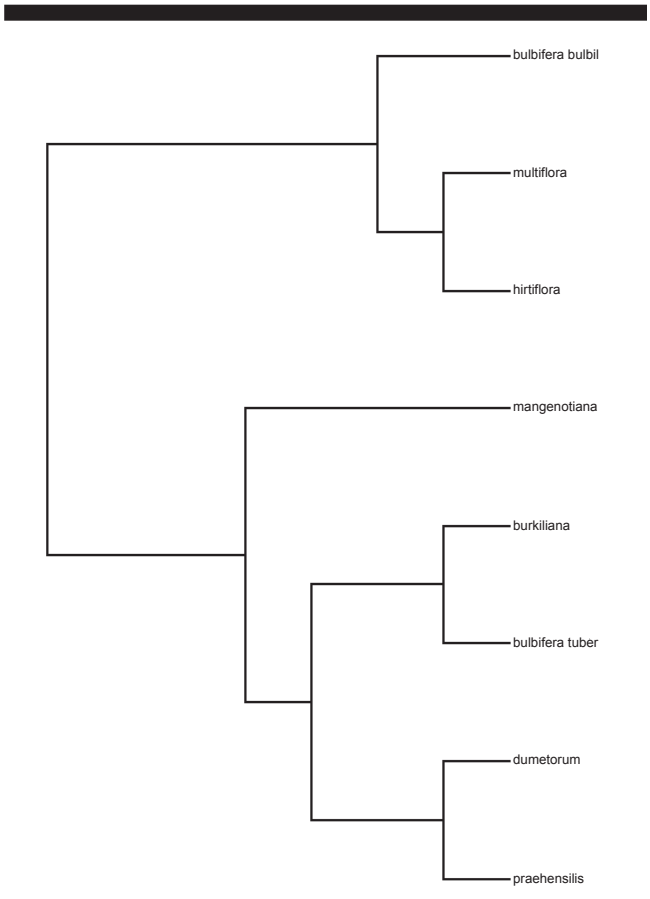

Fig. 1. Dendrogram of wild yam species tubers according to their nutrients

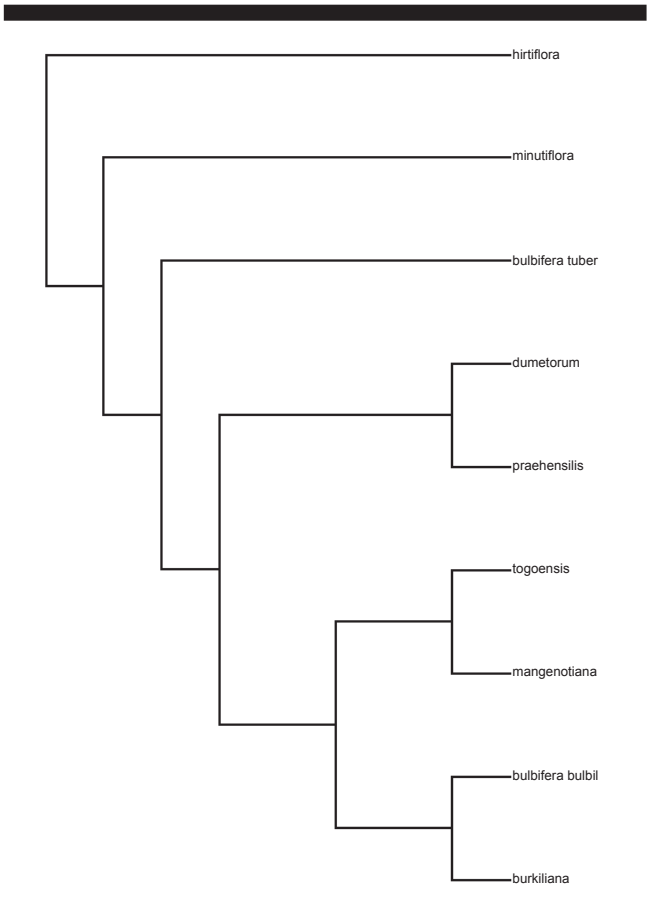

Fig. 2. Dendrogram of wild yam species tubers according to their antinutritional factors

On the other hand, Fig. 2 illustrates a dendrogram of nine wild yam species tubers according to five antinutritional factors (Alexis and Georges). Individuals *D. togoensis, D. bulbifera* bulbil, *D. mangenotiana*, *D. bulbifera* tuber, *D. burkilliana*, *D. praehensilis* and *D. dumetorum* presented the highest contents of *tannins/alkaloids* and *hydrocyanic acid* (HC≡N). Most toxins in the wild yam tubers are soluble alkaloids; during digestion they give severe symptoms. Hydrocyanic acid is a recognized toxin; it causes disorders of the thyroid preventing iodine from settling in it; it entrained goitres/stupidity incidences. Tannins are phenolic polymers; they developed, according to their concentration in food product, a positive/negative organoleptic note when their astringency/bitterness became excessive. Species *D. dumetorum*, *D. minutiflora*, *D. bulbifera* bulbil, *D. togoensis*, *D. mangenotiana* and *D. praehensilis* presented the highest levels of *oxalic acid* HO–C(=O)–C(=O)–OH, which, with some metals, forms insoluble salts; during digestion, if it remains in the digestive tract it is as poorly soluble alkaline oxalates; the ingested portion is toxic. Yams *D. dumetorum*, *D. togoensis*, *D. minutiflora* and *D. hirtiflora* showed the highest *sapogenins* contents, which present a steroidal structure and are aglycon portions of *saponins*; sapogenins haemolyze erythrocytes, which explains toxicity and makes them inedible.

The main aim of the present report is to develop code learning potentialities and, since molecules are more naturally described *via* varying size structured representation, the study of general approaches to structured-information processing. The objective was to categorize yams with principal component analysis (PCA) and cluster analysis (CA), which distinguished yams. Additional interest relied on yam domestication. The next section shows method. Following that, two sections present and discuss results. Finally, the last section summarizes our conclusions. Matrices will be denoted with capital letters. Assume that data matrix X has dimensions ($n \times p$), where n stands for number of observations and p, number of variables. A vector is always indicated in bold, *e.g.* $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ stands for the i-th observation. Classical estimates are denoted by means of a tilde.

COMPUTATIONAL METHOD

Principal components analysis (PCA) is a dimension reduction technique [7–12]. From the original variables set X_j, PCA constructs a new uncorrelated/orthogonal-variables set $\tilde{P}_j$, which are linear combinations of mean-centred variables $\tilde{X}_j = X_j - \overline{X}_j$, and called loadings/principal components (PCs), which correspond with the eigenvectors of the sample co-variance matrix $S = 1/(n-1)\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})$ of the data. For each loading vector $\tilde{P}_j$, corresponding eigenvalue $\tilde{l}_j$ of S tells how much data variability is explained by $\tilde{P}_j$ *via*: $\tilde{l}_j = \mathrm{Var}(\tilde{P}_j)$. Loading vectors are sorted in eigenvalues decaying order. First k PCs explain most data variability. After selecting k, one projects p-dimensional data points onto the subspace spanned by the k loading vectors and computes their co-ordinates with respect to $\tilde{P}_j$, which yields the scores:

$$\tilde{\mathbf{t}}_i = \tilde{P}'(\mathbf{x}_i - \overline{\mathbf{x}}) \tag{1}$$

for each $i = 1, \ldots, n$, which have trivially zero mean. With respect to the original co-ordinate system, the projected data point is computed as fitted value:

$$\hat{\mathbf{x}}_i = \overline{\mathbf{x}} + \tilde{P}\tilde{\mathbf{t}}_i \tag{2}$$

The $p \times k$ loading matrix $\tilde{P}$ contains the loadings column-wise. The ($k \times k$) diagonal matrix $\tilde{L} = (\tilde{l}_j)_j$ denotes eigenvalues. In order to choose appropriate number of loadings k, criteria exist. A graphical one is based on the scree plot which exposes eigenvalues in decaying order; the last-component index before the plot flattens is selected. Formal criterion considers variation explained by the first k loadings requiring:

$$\left(\sum_{j=1}^{k}\tilde{l}_j\right) \Big/ \left(\sum_{j=1}^{p}\tilde{l}_j\right) \geq 80\% \tag{3}$$

Cluster analysis (CA) encompasses different classification algorithms [13,14]. The general question is how to organize data into meaningful structures. The approach of CA is described by saying *birds of a feather flock together*. The starting point of the CA method is the $n \times p$ data matrix $\mathbf{X}$, which contains p features, *e.g.*, p components measured in n samples. One assumes that data were

pre-processed to remove artefacts and that missing values were imputed. The CA organizes samples into a small number of groups/clusters, such that samples within the same group tend to be similar, while samples from different groups tend to be dissimilar to each other. Similarity notion or equivalently dissimilarity/distance between samples is required to start CA. Popular distances between samples $x,x' \in \Re^p$ are l_q distances:

$$\|x - x'\|_q = \left(\sum_{i=1}^{p} |x_i - x'_i|^q \right)^{1/q} \tag{4}$$

in particular l_2 Euclidean/l_1 Manhattan distances. When comparing samples *via* relative values within samples, it is advantageous a similarity based on *Pearson's correlation coefficient* (PCC):

$$r(x - x') = \frac{\sum_{i=1}^{p} (x_i - \bar{x})(x'_i - \bar{x}')}{\left[\sum_{i=1}^{p} (x_i - \bar{x})^2 \sum_{i=1}^{p} (x'_i - \bar{x}')^2 \right]^{1/2}} \tag{5}$$

where $\bar{x} = \left(\sum_{i=1}^{p} x_i \right) / p$ is the mean value of the measures for sample x. The PCC ranges between 1 for identical and –1 for anticorrelated samples; it is transformed into a dissimilarity measure for CA: $1 - r(x,x')$ or $1 - |r(x,x')|$. When a similarity up to a nonlinear transform of the data is expected, other measures are used: *Spearman's rank correlation coefficient*, which is similar to the PCC when the measures exact values are replaced by their rank in the list of p measures sorted by decaying value for a sample, or *mutual information*, which captures relations between two-sample measures [15]. Samples similarity is a research topic: measure affects CA and guidelines exist to direct it [16,17]; it is driven by knowledge about data or is optimized to reach criterion. Strategy to optimize a distance, to fulfil dis/similarity constraints between particular sample pairs, is amenable to automatization with *metric learning* algorithms [18]. Two CA categories exist: hierarchical (HCA) and non-hierarchical (NHCA) [19,20]. The HCA rearranges objects in tree structure [21]. Members are grouped until predetermined number of clusters are assembled. A dendrogram is created that maps N members in one cluster to N members in N clusters. In NHCA a nearest-neighbour list assembles members into related clusters. In HCA each object is assumed to be a lone cluster. Distance matrix is scanned for the minor values. Objects are clustered. Iterations lead to objects total CA generating a dendrogram with objects clustered according to similarity. Results rely on: (1) structure representation, (2) data normalization and (3) algorithms/parameter settings. Data normalization is basis for comparing experiments with series when experimental conditions are not identical; it ensures that the experiments quality is comparable. Normalization functions follow: (1) linear $x'_i = X'_{\min} + (X'_{\max} - X'_{\min})(x_i - X_{\min}) / (X_{\max} - X_{\min})$, (2) ratio $x'_i = x_i / \sum_{i=1}^{n} |x_i|$ and (3) Z-score $x'_i = (x_i - \bar{x}) / s$, where s is standard deviation. Once distance measure is chosen CA is used to organize data into groups according to space. The HCA provides not only CA into K groups for fixed K if one cuts the dendrogram at a particular depth, but also hierarchical organization of the data into nested clusters with individual samples at the dendrogram bottom leaves and increasing-size clusters when one goes up the tree toward the root. Tree branch length is related to how strong the separation at the branch upper part is. Cutting the tree at a given depth defines data CA into groups finite number. The HCA is *agglomerative* when groups are formed by a *bottom-up* strategy iteratively joining most similar groups into larger groups, or *divisive* when groups are split into *top-down* strategy starting from a single group with all instances and iteratively splitting groups into two subgroups as separated as possible; it depends on a linkage function, which defines how the distance between two groups is computed from gaps between the samples they contain; it presents advantage of visually appealing organization of data, providing a multi-resolution view of groups within data and suggesting biointerpretations. Drawback is: it outputs dendrogram when samples have no reason to be organized into tree. Way to assess CA statistical significance is to value stability, problem related to choosing number of clusters, which are sensitive to errors in tree construction.

CALCULATION RESULTS

Nine wild yam species tubers reported by Alexis and Georges were used as the model dataset. The PCC matrix **R** was calculated between the pairs of nine yams; the upper triangle turned out to be:

$$\mathbf{R} = \begin{pmatrix} 1.000 & 0.629 & 0.949 & 0.936 & 0.892 & 0.941 & 0.949 & 0.940 & 0.940 \\ & 1.000 & 0.592 & 0.573 & 0.645 & 0.512 & 0.510 & 0.647 & 0.647 \\ & & 1.000 & 0.999 & 0.984 & 0.994 & 0.994 & 0.996 & 0.996 \\ & & & 1.000 & 0.987 & 0.995 & 0.994 & 0.994 & 0.994 \\ & & & & 1.000 & 0.969 & 0.964 & 0.991 & 0.991 \\ & & & & & 1.000 & 0.999 & 0.984 & 0.984 \\ & & & & & & 1.000 & 0.982 & 0.982 \\ & & & & & & & 1.000 & 1.000 \\ & & & & & & & & 1.000 \end{pmatrix}$$

The $R_{i,9}$ are taken as $R_{i,8}$. Some PPC correlations are high, *e.g.*, *D. bulbifera* bulbil–*burkilliana* or *D. dumetorum–praehensilis* $R_{3,4} = R_{6,7} = 0.999$. Correlations of PCC are illustrated in partial correlation diagram, which contains high ($r \geq 0.75$), medium ($0.5 \leq r < 0.75$), low ($0.25 \leq r < 0.50$) and *zero* ($r \leq 0.25$) partial correlations. Pairs of yams with high partial correlations show similar nutrients and antinutritional factors. Partial correlation diagram contains 28 high (*cf.* Fig. 3, *red*) and 8 medium (*orange*) partial correlations. It is in qualitative agreement with previous results (Figs. 1 and 2).

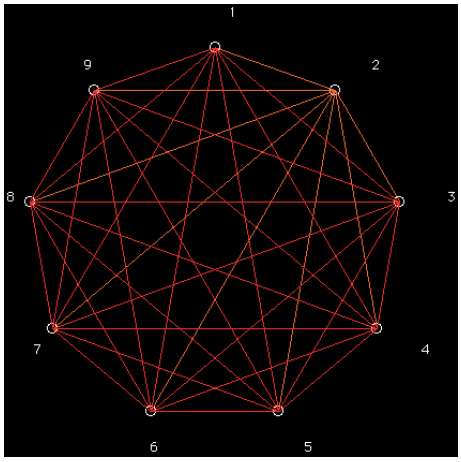

Fig. 3. Partial correlations diagram: high (*red*) and medium (*orange*) partial correlations

The dendrogram of nine wild yam species tubers according to 14 nutrients/antinutritional factors, *cf.* Fig. 4, shows different behaviour depending on *energy*. Three classes are clearly recognized:

(1,2)(3–5,8,9)(6,7)

Individuals *D. minutiflora* and *hirtiflora* present the highest levels of *moisture*, the lowest levels of *energy* and are grouped into class 1. Species with high levels of *alkaloids* are clustered into grouping 2. Yams *D. dumetorum* and *D. praehensilis* showed the highest levels of *energy*, *soluble carbohydrates* and *tannins*, high levels of *lipids*, *starch* and *proteins*, low levels of *ash* and are grouped into class 3. The yams belonging to the same class appear highly correlated in partial correlation diagram, in qualitative agreement with previous results (Figs. 1-3).

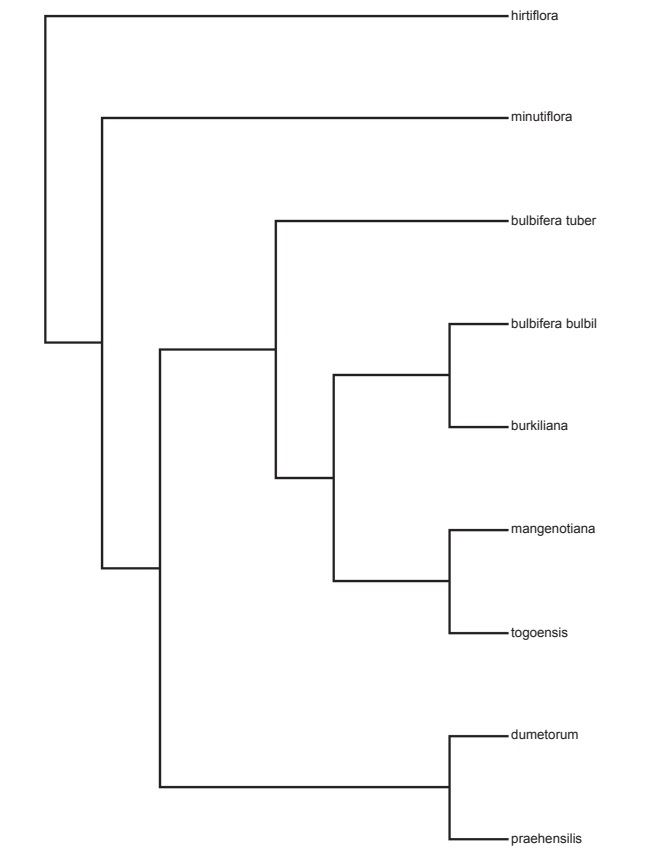

Fig. 4. Dendrogram of wild yam species tubers according to nutrients/antinutritional factors

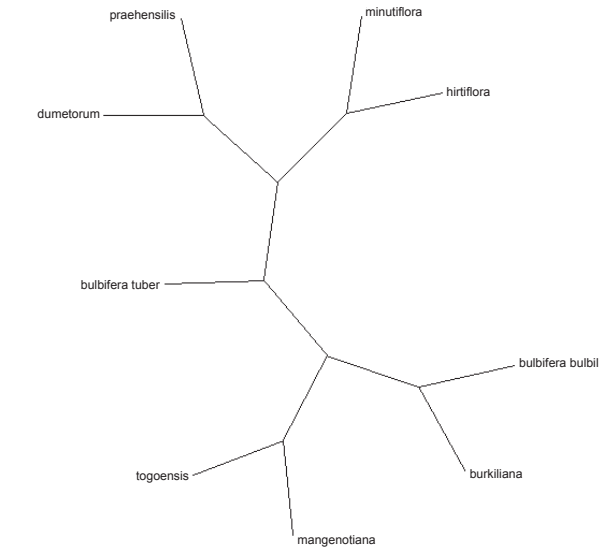

Fig. 5. Radial tree of wild yam species tubers according to nutrients/antinutritional factors

Radial tree classification (*cf.* Fig. 5) shows yams different behaviour depending on *energy*. The same classes are recognized in agreement with partial correlation diagram, dendrogram and previous results (Figs. 1–4). Again individuals with the highest levels of *moisture* are grouped into class 1, etc.

The splits graph for nine yams in Table 1 (*cf.* Fig. 6) shows that species 4 and 6-9 appear superimposed on 3. It reveals conflicting relationships between classes because of interdependences [22]. Therefore, it indicates spurious relationships resulting from base-composition effects. It illustrates the different behaviour of yams depending on *energy*. It is in qualitative agreement with partial correlation diagram, binary/radial trees and previous results (Figs. 1–5).

Usually in quantitative structure–property relationships (QSPRs), the data file contains less than one hundred objects and several thousands of *X*-variables. In fact, there are so many *X*-variables that no one can discover

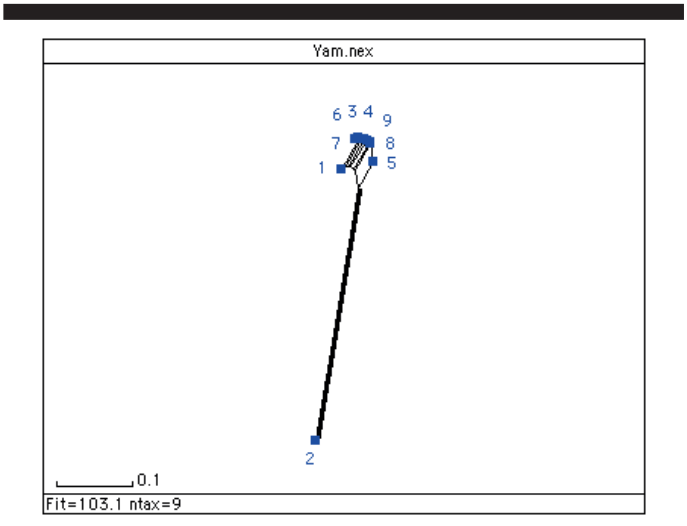

Fig. 6. Splits graph of wild yam species tubers according to nutrients/antinutritional factors

by *inspection* patterns, trends, clusters, *etc*. in the objects. *Principal components analysis* (PCA) is a technique, extremely useful to *summarize* all the information contained in the **X**-matrix and put it in a form understandable by humans. The PCA works by decomposing the **X**-matrix as the product of two smaller matrices **P** and **T**. The *loading matrix* **P** with information about the variables contains a few vectors, the so-called principal components (PCs), which are obtained as linear combinations of original *X*-variables. The *score matrix* **T**, with information about the objects, is such that every object is described in terms of the projections onto PCs, instead of the original variables: **X** = **TP**' + **E** where ' denotes transpose matrix. The information not contained in the matrices remains as *unexplained* X-*variance* in a *residual matrix* **E**. Every PC_i is a new co-ordinate expressed as a linear combination of the old features x_j: $PC_i = \Sigma_j b_{ij} x_j$. The new co-ordinates PC_i are called *scores* or *factors* while coefficients b_{ij} are called *loadings*. The scores are ordered according to the information content with regard to total variance among all objects. The *score-score plots* show the positions of compounds in the new co-ordinate system, while *loading-loading plots* show the position of features that represent compounds in the new co-ordination. The PCs present two interesting properties. (1) They are extracted in decaying order of importance. The first PC F_1 always contains more information than the second F_2, F_2 more than the third F_3, *etc*. (2) Every PC is orthogonal to one another. There is no correlation between the information contained in different PCs. A PCA was performed for yams. The importance of PCA factors $F_{1–14}$ for properties is collected in Table 2. In particular the use of only the first factor F_1 explains 36% of the variance (64% of the error), the combined application of the first two factors $F_{1/2}$ accounts for 64% of variance (36% of error), the utilization of the first three factors $F_{1–3}$ rationalizes 80% of variance (20% of error), *etc*.

Table 2. Importance of principal component analysis factors for composition of wild yam species

Factor	Eigenvalue	Percentage of variance	Cumulative percentage of variance
F_1	4.97682031	35.55	35.55
F_2	4.00570650	28.61	64.16
F_3	2.18458722	15.61	79.77
F_4	1.42318844	10.16	89.93
F_5	0.82188252	5.87	95.80
F_6	0.51457327	3.68	99.48
F_7	0.07324171	0.52	100.00
F_8	0.00000002	0.00	100.00
F_9	0.00000000	0.00	100.00
F_{10}	0.00000000	0.00	100.00
F_{11}	0.00000000	0.00	100.00
F_{12}	0.00000000	0.00	100.00
F_{13}	0.00000000	0.00	100.00
F_{14}	0.00000000	0.00	100.00

The PCA factors loadings of the first seven factors were calculated.

The PCA $F_{1–14}$ profiles for the properties were computed. In particular for F_1 variable i_8 shows the greatest weight in the profile; however, F_1 cannot be reduced to two variables $\{i_6,i_8\}$ without a 71% error. For F_2 variable i_3 presents the greatest weight in the profile; notwithstanding, F_2 cannot be reduced to two variables $\{i_1,i_3\}$ without a 72% error. For F_3 variable i_{10} assigns the greatest weight in the profile; nevertheless, F_3 cannot be reduced to two variables $\{i_{10},i_{13}\}$ without a 45% error. For F_4 variable i_{12} consigns the greatest weight in the profile; however, F_4 cannot be reduced to two variables $\{i_1,i_{12}\}$ without a 49% error. For F_5 variable i_{11} represents the greatest weight in the profile; notwithstanding, F_5 cannot be reduced to two variables $\{i_7,i_{11}\}$ without a 66% error. For F_6 variable i_9 displays the greatest weight in the profile; nevertheless, F_6 cannot be reduced to two variables $\{i_9,i_{11}\}$ without a 53% error. For F_7 variable i_4 exhibits the greatest weight in the profile; however, F_7 cannot be reduced to two variables $\{i_4,i_{10}\}$ without a 41% error, *etc*. Factors $F_{1–7}$ can be considered as linear combinations of $\{i_6,i_8\}$, $\{i_1,i_3\}$, $\{i_{10},i_{13}\}$, $\{i_1,i_{12}\}$, $\{i_7,i_{11}\}$, $\{i_9,i_{11}\}$ and $\{i_4,i_{10}\}$, respectively, with 71%, 72%, 45%, 49%, 66%, 53% and 41% errors.

The PCA F_2-F_1 *scores plot* of yams (*cf.* Fig. 7) shows that yam 9 appears superimposed on 8. It illustrates yams different behaviour depending on *energy*. Three classes are clearly distinguished: class 1 with 2 compounds ($0 > F_1 > F_2$, *bottom*), grouping 2 with 5 substances ($F_1 < F_2$, *middle left*) and class 3 with 2 molecules ($F_1 > F_2 > 0$, *right*). Plot is in qualitative agreement with partial correlation diagram, binary/radial trees, splits graph and previous results (Figs. 1-6).

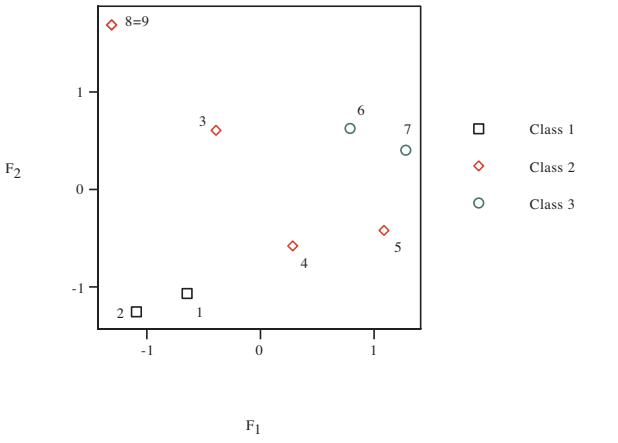

Fig. 7. PCA scores plot of wild yam species tubers according to nutrients/antinutritional factors

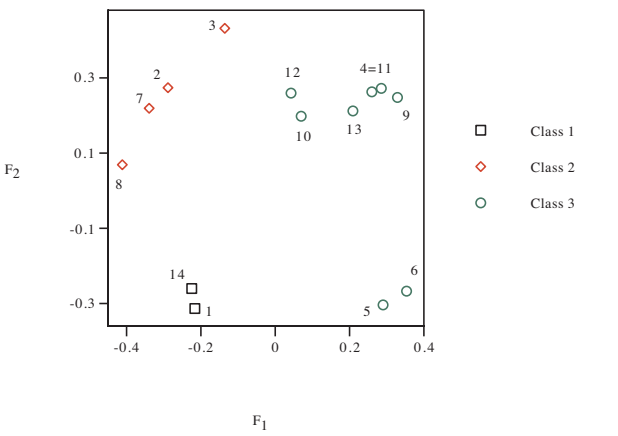

Fig. 8. PCA loadings plot of wild yam species tubers according to nutrients/antinutritional factors

From PCA factors loadings of yams, F_2–F_1 *loadings plot* (*cf.* Fig. 8) depicts 14 nutrients/antinutritional factors. Component 11 appears superimposed on 4. Three groupings are clearly distinguished: class 1 with 2 components {1,14} ($0 > F_1 > F_2$, *bottom*), grouping 2 with 4 nutrients {2,3,7,8} ($F_1 << F_2$, *middle left*) and class 3 with 8 components {4–6,9–13} ($F_1 > F_2 \approx 0$, *right*). In general, nutrients {1–9} appear more separated than antinutritional factors {10–14}. In particular, most antinutritional factors (toxic oxalic/hydrocyanic acids, tannins and alkaloids) appear grouped into class 3. In addition as a complement to scores plot (Fig. 7) for loadings (Fig. 8), it is confirmed that yams in class 1 located at the bottom present a more pronounced contribution from components in grouping 1, situated in the same position in Fig. 7. Yams in class 2 in the middle left side show a contribution from components in grouping 2, positioned in the same side. Yams in class 3, at the right side, indicate a more pronounced contribution from components in grouping 3 placed in the same side; in particular, toxic yam 6 (*D. dumetorum*) displays a contribution from antinutritional factors 10 and 12 (oxalic/hydrocyanic acids) placed in the same side.

Instead of nine yams in $\Re^{14}$ space of 14 components, consider 14 components in $\Re^{9}$ space of nine yams. Upper triangle of PCC matrix **R** between pairs of 14 components for yams resulted:

$$\mathbf{R} = \begin{pmatrix} 1.000 & -0.129 & -0.469 & -0.273 & -0.181 & 0.062 & 0.262 & 0.259 & -0.612 & 0.003 & -0.534 & -0.534 & -0.714 & 0.613 \\ & 1.000 & 0.544 & -0.069 & -0.610 & -0.947 & 0.515 & 0.608 & -0.149 & 0.302 & -0.059 & 0.353 & -0.358 & 0.191 \\ & & 1.000 & 0.131 & -0.722 & -0.649 & 0.711 & 0.416 & 0.258 & 0.009 & 0.140 & 0.283 & 0.458 & -0.445 \\ & & & 1.000 & -0.160 & 0.226 & -0.120 & -0.572 & 0.739 & 0.685 & 0.808 & 0.221 & 0.162 & -0.437 \\ & & & & 1.000 & 0.732 & -0.930 & -0.641 & 0.149 & -0.219 & 0.015 & -0.036 & 0.095 & 0.053 \\ & & & & & 1.000 & -0.690 & -0.811 & 0.327 & -0.102 & 0.166 & -0.258 & 0.243 & -0.137 \\ & & & & & & 1.000 & 0.707 & -0.263 & 0.003 & -0.334 & -0.011 & -0.036 & 0.097 \\ & & & & & & & 1.000 & -0.759 & -0.276 & -0.423 & -0.042 & -0.219 & 0.223 \\ & & & & & & & & 1.000 & 0.372 & 0.584 & 0.259 & 0.450 & -0.522 \\ & & & & & & & & & 1.000 & 0.445 & 0.565 & -0.280 & 0.177 \\ & & & & & & & & & & 1.000 & 0.217 & 0.318 & -0.723 \\ & & & & & & & & & & & 1.000 & 0.297 & 0.141 \\ & & & & & & & & & & & & 1.000 & -0.648 \\ & & & & & & & & & & & & & 1.000 \end{pmatrix}$$

High PPC correlations result not only between nutrients and between antinutritional factors but also combining both types, *e.g.* soluble carbohydrates and tannins $R_{4,11}$ = 0.808. A dendrogram for nutrients/antinutritional factors (*cf.* Fig. 9) separates the same three classes, in agreement with PCA loadings plot and previous results (Fig. 8). Again most toxins group into class 3.

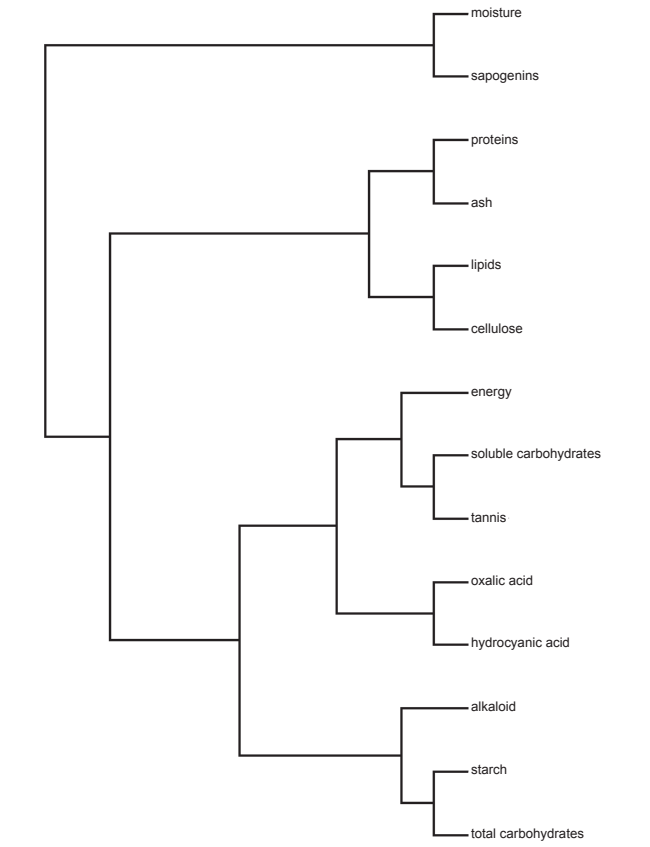

Fig. 9. Dendrogram of nutrients/antinutritional factors for wild yam species tubers

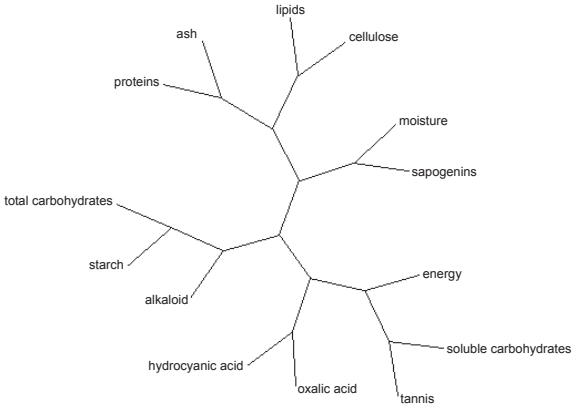

Fig. 10. Radial tree of nutrients/antinutritional factors for wild yam species tubers

The radial tree of 14 nutrients/antinutritional factors for wild yam species tubers, *cf.* Fig. 10, separates the same classes in agreement with PCA loadings plot, dendrogram and previous results (Figs. 8 and 9). One more time most toxins group into class 3.

Splits graph of nutrients/antinutritional factors for wild yam species tubers, *cf.* Fig. 11, indicates conflicting relations between all classes because of interdependences *via* base-composition effects. It separates the same groupings in agreement with PCA loadings plot, binary/radial trees and previous results (Fig. 8–10). Once more most toxins group into class 3.

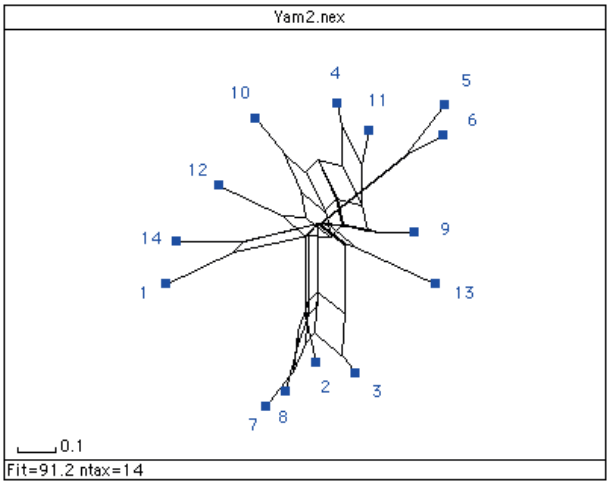

Fig. 11. Splits graph of nutrients/antinutritional factors for wild yam species tubers

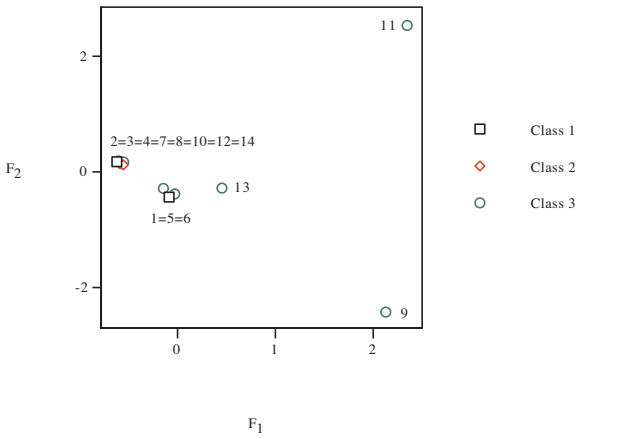

Fig. 12. PCA scores plot of components corresponding to wild yam species tubers

A PCA was performed for the components. Notice that factor F_1 explains 90% of variance (10% error). Factors $F_{1/2}$ account for 98% of variance (2% error), factors $F_{1–3}$ rationalize 99.95% of variance (0.05% error), *etc*. The PCA F_2–F_1 *scores plot* of components corresponding to wild yam species, *cf*. Figure 12, shows that components 5 and 6 appear superimposed on 1, and 3, 4, 7, 8, 10, 12 and 14, on 2. Three classes are clearly distinguished: class 1 with 2 components {1,14} ($F_1 < F_2 \approx 0$, *bottom*), grouping 2 with 4 nutrients {2,3,7,8} ($F_1 << F_2$, *left*) and class 3 with 8 components {4-6,9-13} ($F_1 > F_2$, *right*). In general, antinutritional factors {10–14} appear more separated than nutrients {1–9} and most toxins group into class 3. Plot separates the same classes in agreement with PCA loadings plot, binary/radial trees, splits graph and previous results (Fig. 8-11). The F_3-F_1 and F_3-F_2 scores plots result similar improving only the separation of component 1 from 5-6.

DISCUSSION

It is tempting to speculate that the molecular heterogeneity could explain the diversity of morphologies, properties and functions of yams. Having draft the molecular profile of tubers nutrients and antinutritional factors should give the basic elements, which cause and accompany composition, and provide the necessary information for defining yam and toxin subtypes in a rational way. What was so far mainly defined by morphological observation could also be approached hopefully better with full molecular characterization, which information could maybe not replace but it certainly complements the morphology. The question is much more than the intellectual exercise of ordering observations: conceptually it provides a map of the yams composition and facilitates reason and drawing hypotheses, which will lead to understanding the nature of the nutrients, toxins and bioprinciples that govern it; in practice it presents of course huge implications for the toxins and the atlas of yams and component types. The success of these attempts, to define new molecular classification of yams, should not hide the fact that clustering data remains a challenging task from a methodological viewpoint. In particular many parameters influence the classification obtained by clustering methods, *e.g*. the features/metric used to compare samples, the clustering algorithm itself and the procedure to select the number of clusters. Notwithstanding these limitations it is fair to say that the new molecular classifications of yams, obtained by automatic clustering of data, started to revolutionize the way one apprehends yam heterogeneity. As larger collections of samples are analyzed it is likely that finer classifications, into well-specified and robust subtypes, will emerge from clustering methods and allow a more precise stratification of nutrients/antinutritional factors into subcategories, which would not be captured by only morphological parameters. As different subgroups can present different uses or toxins, a more precise and robust classification of nutrients and antinutritional factors can improve nutritional use.

CONCLUSIONS

From the discussion of the present results the following conclusions can be drawn.

1. Several criteria were selected to reduce analysis to manageable quantity, from enormous set of yam-tubers components; they refer to nutrients/antinutritional factors. Integrating data analysis, *e.g*. meta-analysis, was useful to increase numbers of samples and variety of analyzed data. Different behaviour of yams depends on *energy*. With regard to components, most antinutritional factors grouped into the same class.

2. Principal components analyses of components, and cluster analyses of yams, allowed classifying them and agreed. Clustering is difficult; *e.g*. although oranges/apples present differences they are both fruit. Is a pomegranate more like an apple or an orange? When a clustering problem is poorly specified, or variation within each cluster is greater than that between different clusters, usually meaningful clustering becomes almost impossible. Progression in methods development is hampered by lack of *gold standards*, against which to judge quality of any clustering exercise. Chemistry/computational-methods understanding is essential for tackling associated *data mining* tasks, without being distracted by abundant fool's gold. If a small number of data clusters are easy to fit, model predictive ability could be guaranteed only if deviations inside clusters do not diverge.

ACKNOWLEDGEMENT

The author acknowledges financial support from the Spanish Ministerio de Ciencia e Innovación (Project No. BFU2010–19118).

REFERENCES

[1] COURSEY, D. G.; RUSSEL, J. C. 1969. A note on endogenous and biodeterioration factors in the respiration of dormant yam tubers, Int. Biodeterior. Bull., 5, 27–30.
[2] HLADIK, A., BAHUCHET, S., DUCATILLION, C., HLADIK, C. M. 1984. Les plantes à tubercules de la forêt dense d'Afrique Centrale. Rev. Ecol. (Terre Vie), 39, 248–290.
[3] WEBSTER, J., BECK, W., TERNAI B. 1984. Toxicity and bitterness in Australian *Dioscorea bulbifera L*. and *Dioscorea hispida Dennst*. from Thailand., J. Agric. Food Chem., 32, 1087–1090.
[4] MIÈGE, J. 1977. Stratégies végétales. Musée de Genève, 172, 13–17.
[5] ALEXIS, S. D., GEORGES, A. N'G. 2012. Classification of some wild yam species tubers of Ivory Coast forest zone, *Int. J. Biochem. Res. Rev*., 2, 137–151.
[6] BRADBURY, J. H. 1988. The chemical composition of tropical root crops, ASEAN Food J., 4, 3–136.
[7] HOTELLING, H. Analysis of a complex of statistical variables into principal components, J. Educ. Psychol., 1933, 24, 417–441.
[8] KRAMER, R. Chemometric Techniques for Quantitative Analysis; Marcel Dekker: New York, 1998.
[9] PATRA, S. K.; MANDAL, A. K.; PAL, M. K. J. 1999. Photochem. Photobiol., Sect. A, 122, 23.
[10] JOLLIFFE, I. T. 2002. Principal Component Analysis. Springer, New York.
[11] XU, J.; HAGLER, A. 2002. Molecules, 7, 566.
[12] SHAW, P. J. A. 2003. Multivariate Statistics for the Environmental Sciences, Hodder-Arnold: New York.
[13] IMSL, Integrated Mathematical Statistical Library (IMSL); IMSL: Houston, 1989.
[14] TRYON, R. C. J. 1939. Chronic Dis., 20, 511–524.
[15] PRINESS, I., MAIMON, O. AND BEN-GAL, I. 2007. Evaluation of gene-expression clustering via mutual information distance measure, BMC Bioinformatics, 8, 111.
[16] STEUER, R., KURTHS, J., DAUB, C. O., WEISE, J. AND SELBIG, J. 2002. The mutual information: Detecting and evaluating dependencies between variables, *Bioinformatics*, 18(Suppl. 2) S231–S240.
[17] D'HAESELEER, P., LIANG, S., AND SOMOGYI, R. 2000. Genetic network inference: From co-expression clustering to reverse engineering, *Bioinformatics*, 16, 707–726.
[18] PEROU, C. M., SØRLIE, T., EISEN, M. B., VAN DE RIJN, M., JEFFREY, S. S., REES, C. A., POLLACK, J. R., ROSS, D. T., JOHNSEN, H., AKSLEN, L. A., FLUGE, O., PERGAMENSCHIKOV, A., WILLIAMS, C., ZHU, S. X., LØNNING, P. E., BØRRESEN-DALE, A. L., BROWN, P. O., BOTSTEIN, D. 2000. Molecular portraits of human breast tumours, Nature (London), 406,747–752.
[19] JARVIS, R. A.; PATRICK, E. A. 1973. Clustering using a similarity measure based on shared nearest neighbors, IEEE Trans. Comput., C22, 1025–1034.
[20] PAGE, R. D. M. 2000. Program TreeView; Universiy of Glasgow, UK.
[21] EISEN, M. B., SPELLMAN, P. T., BROWN, P. O., AND BOTSTEIN, D. 1998. Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. U.S.A., 95, 14863–14868.
[22] HUSON, D. H. 1998. Bioinformatics, 14, 68.