

<i>Nereis. Revista Iberoamericana Interdisciplinar de Métodos, Modelización y Simulación</i>	7	39-50	Universidad Católica de Valencia San Vicente Mártir	Valencia (España)	ISSN 1888-8550
--	---	-------	---	-------------------	----------------

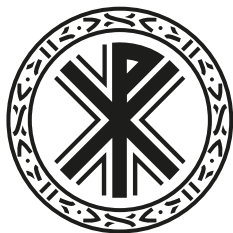
Classification of Fruits Proximate and Mineral Content: Principal Component, Cluster, Meta-Analyses

Fecha de recepción y aceptación: 3 de diciembre de 2014, 8 de enero de 2015

Francisco Torrens Zaragoza¹

¹ Institut Universitari de Ciència Molecular, Universitat de València.

Correspondencia: Edifici d'Instituts de Paterna, P. O. Box 22085. 46071 Valencia. España. E-mail: francisco.torrens@uv.es



ABSTRACT

Fruits from Nigeria are classified by principal component analyses (PCAs) of proximates and minerals, and plants cluster analyses (CAs), which agree. Samples group into three classes. Compositional PCA and fruit CA allow classification and concur. The first axis explains 39%, the first two, 59%, the first three, 73% variance. *Moisture* and K contents are high; *ash* and *carbohydrate*, low. Fruit behaviour depends on *ash*, *fibra* and K. Most nutritional constituents are grouped into the same class.

KEYWORDS: *natural product, phytochemical, toxic, metal, phytopharmacy, phytomedicine, chemotaxonomy, cytochemical, medicinal plant, ethnomedicine, functional food.*

RESUMEN

Se clasifica frutas de Nigeria por análisis de componentes principales (ACPs) de constituyentes principales y minerales, y análisis de agregados (AAGs) de plantas, los cuales están de acuerdo. Las muestras se agrupan en tres clases. El ACP de composición y AAG de frutas permiten la clasificación y asienten. El primer eje explica el 39% de la varianza, los dos primeros, el 59%, los tres primeros, el 73%. Los contenidos de *humedad* y K son altos; *ceniza* y *carbohidrato*, bajos. El comportamiento de las frutas depende de la *ceniza*, *fibra* y K. Se agrupa en la misma clase la mayoría de constituyentes nutricionales.

PALABRAS CLAVE: *producto natural, fitoquímico, tóxico, metal, fitofarmacia, fitomedicina, quimiotaxonomía, citoquímico, planta medicinal, etnomedicina, alimento funcional.*

INTRODUCTION

In Nigeria, food shortage resulted of population growth, competition for fertile land, poverty and lack of agricultural inputs [1]. Fruits contain water, carbohydrates, vitamins (Vit-A, B_{1/2}, C, D, E), minerals (Ca, Mg, Zn, Fe, K) and needed compounds [2–5]. Water is essential as solvent and in metabolism. Lipids are necessary because they provide maximum energy and facilitate intestinal absorption and transportation of fat-soluble Vits (A, D, E, K) and carotenoids. Fruit carbohydrates could supplement scarce cereal. Their deficiency causes depletion of body tissue. Iron is needed for energy and endurance because it delivers O₂ *via* the body. It is important in the diet of pregnant women, nursing mothers, infants, convulsing patients and elderly to prevent anaemia and related diseases [6]. Its recommended daily allowance is 7mg·day⁻¹ (men) and 12–18mg·day⁻¹ (pregnant women) [7]. Magnesium prevents muscle cramping, enhances nerve functioning, relieves tight sore muscles, improves bone density and relaxes muscles *via* the airway to the lung allowing asthma patients to breathe easier. Its deficiency causes severe diarrhoea, migraines, hypertension, cardiomyopathy, arteriosclerosis and stroke. Calcium regulates muscle contractions, transmission of nerve impulses and bone formation. Zinc deficiency because of high phytic-acid content of diets leads to poor growth, impairs immunity, increases morbidity from common infectious diseases and rises mortality; it is essential in cell growth, testosterone production, protein and nucleic



acid syntheses, development, growth and recovery from illness [8]. Too much Zn is toxic. Ekpete *et al.* reported 12 proximates and mineral compositions of nine fruits (*cf.* Table 1) [9,10].

Table 1. The results of the mineral composition of the analyzed fruits (mg·g⁻¹)

Species	K ^a	Fe	Ca	Mg	Zn
1. Banana <i>Musa paradisiaca</i>	380.05	0.86	7.24	45.71	0.45
2. Guava <i>Psidium guajava</i>	166.18	2.40	16.46	15.67	0.06
3. Pawpaw <i>Carica papaya</i> L.	358.37	0.46	84.90	44.04	0.24
4. Orange <i>Citrus sinensis</i>	126.36	1.19	33.35	20.10	0.35
5. Apple <i>Malus domestica</i>	280.10	0.43	18.65	27.12	0.51
6. Watermelon <i>Citrullus lanatus</i>	125.00	0.40	7.00	30.21	0.75
7. Soursop <i>Annona muricata</i>	26.15	1.40	17.43	32.50	0.57
8. Bush mango <i>Irvingia gabonensis</i>	77.95	3.21	14.35	14.67	1.26
9. Pineapple <i>Ananas comosus</i>	261.03	3.28	16.45	25.20	0.15

^a Compositions: proximates (i_1 , moisture; i_2 , ash; i_3 , dry matter; i_4 , carbohydrate; i_5 , protein; i_6 , lipid; i_7 , crude fibre) and minerals (i_8 , K; i_9 , Fe; i_{10} , Ca; i_{11} , Mg; i_{12} , Zn).

Figure 1 shows the *dendrogram* (binary tree) of nine fruits according to seven proximates after Ekpete *et al.* Species show different behaviour depending on *moisture* and *dry matter*. Individuals *Carica papaya* L., *Citrullus lanatus* and *Ananas comosus* presented highest *moisture*. Low moisture guarantees a good keeping period. Fruit energy value is inversely proportional to moisture content. Species *Citrus sinensis*, *C. lanatus* and *Irvingia gabonensis* showed highest *ash*. Samples with high percentages of ash contents present high concentrations of minerals that catalyze metabolic processes, and improve growth and development. Fruits *Musa paradisiaca*, *C. sinensis* and *I. gabonensis* indicated highest *dry matter*. Individuals *M. paradisiaca*, *C. papaya* and *Annona muricata* denoted highest *carbohydrate*, *protein* and *lipid*; they had best nutritional potential. Samples with low carbohydrate content are ideal for diabetic and hypertensive patients requiring low-sugar diets. Proteins are essential component of diet needed for survival of animals and men. Their basic function is to supply amino acids (AAs) required for nutrition. Their deficiency causes growth retardation, muscle wasting, oedema, belly swelling and collection of fluids in the body. All the fruits showed low lipid content and can be recommended in weight-reducing diets. Fruits *C. sinensis*, *C. lanatus* and *I. gabonensis* signified highest *crude fibre*. All the fruits contained low fibre levels and may be included in weaning diets.

On the other hand, Fig. 2 illustrates the dendrogram of nine fruits according to five minerals (Ekpete *et al.*). Individuals *M. paradisiaca*, *C. papaya* and *Malus domestica* presented highest K. Samples *Psidium guajava*, *I. gabonensis* and *A. comosus* contained highest Fe. Species *C. lanatus*, *A. muricata* and *I. gabonensis* took in highest Zn. Iron and Zn were high in *I. gabonensis*. Fruits *C. papaya*, *C. sinensis* and *M. domestica* included highest Ca contents. Individuals *M. paradisiaca*, *C. papaya* and *A. muricata* comprised highest Mg. Calcium and Mg were high in *C. papaya*.

Earlier publications in *Nereis* classified yams [11] and lactic acid bacteria [12] by principal component (PCA), cluster (CA) and meta-analyses. The main aim of the present report is to develop code learning potentialities and, since fruit components are more naturally described *via* varying size-structured representation, study of general approaches to processing of structured information. In view of the nutritional and health benefits of fruits, the objective was to categorize them with PCA and CA, which distinguished their phytochemical constituents: proximate and mineral content. The next section presents the computational method. Following that, two sections illustrate and discuss the calculation results. Finally, the last section summarizes my conclusions.



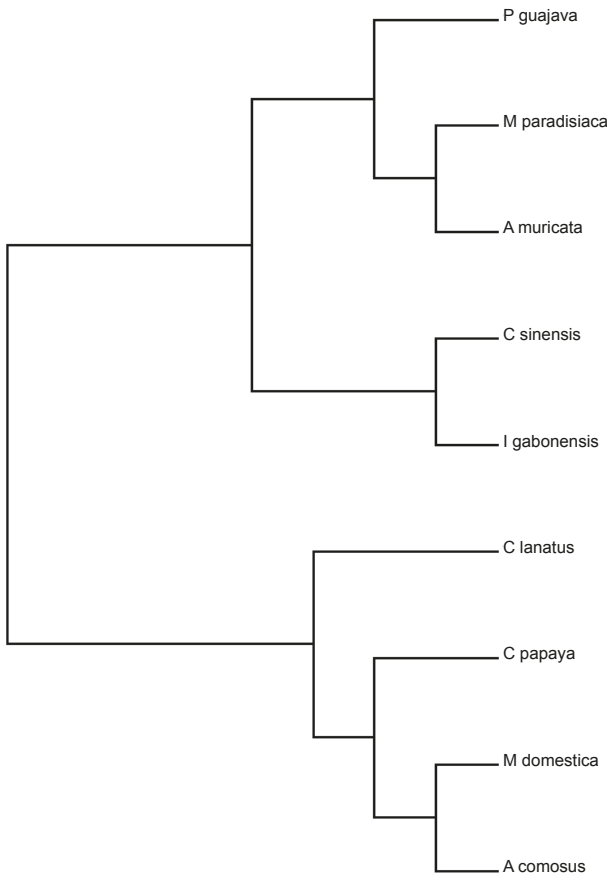


Figure 1. Dendrogram of fruits according to proximate content.

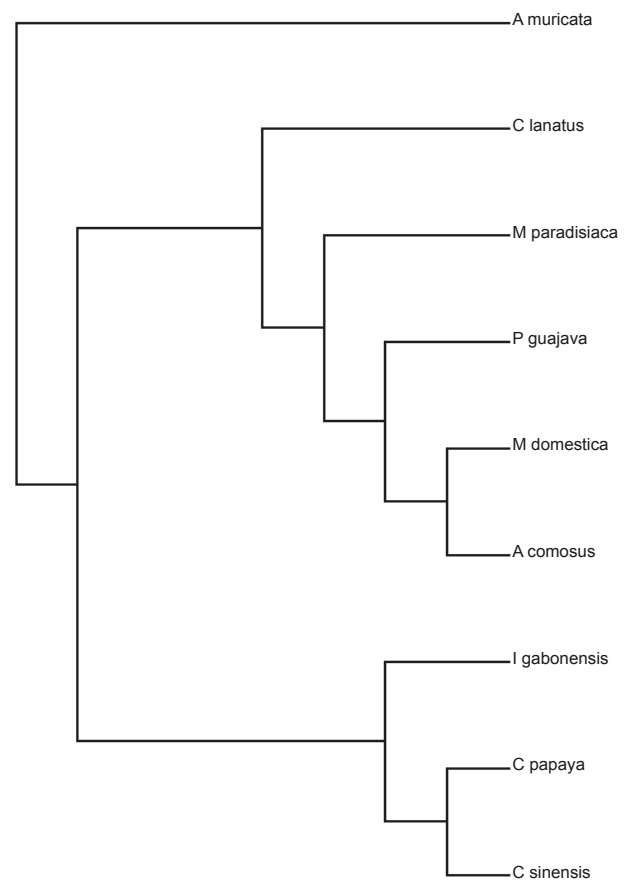


Figure 2. Dendrogram of fruits according to their mineral content.

COMPUTATIONAL METHOD

Principal components analysis (PCA) is a dimension reduction technique [13–18]. From original variables X_j , PCA builds orthogonal variables \tilde{P}_j linear combinations of mean-centred ones $\tilde{X}_j = X_j - \bar{X}_j$ corresponding to eigenvectors of sample co-variance matrix $S = 1/(n-1) \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})$. For every loading vector \tilde{P}_j , matching eigenvalue \tilde{l}_j of S tells how much data variability is explained: $\tilde{l}_j = \text{Var}(\tilde{P}_j)$. Loading vectors are sorted in decaying eigenvalues. First k PCs explain most variability. After selecting k , one projects p -dimensional data on to subspace spanned by k loading vectors and computes co-ordinates *vs.* \tilde{P}_j , yielding scores:

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

for every $i = 1, \dots, n$ having trivially zero mean. With respect to original co-ordinate system, projected data point is computed fitting:

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



Loading matrix \tilde{P} ($p \times k$) contains loadings column-wise and diagonal one $\tilde{L} = (\tilde{l}_j)$ ($k \times k$), eigenvalues. Loadings k explain variation:

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

Cluster analysis (CA) encompasses different classification algorithms [19,20]. Starting point is $n \times p$ data matrix X containing p components measured in n samples. One assumes data were pre-processed to remove artifacts, and missing values, imputed. The CA organizes samples into small number of clusters such that samples within cluster are similar. Distances l_q between samples $x, x' \in \mathfrak{R}^p$ are:

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

(e.g., Euclidean l_2 , Manhattan l_1 distances). Comparing samples, *Pearson's correlation coefficient* (PCC) is advantageous:

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

where $\bar{x} = (\sum_{i=1}^p x_i) / p$ is measure mean value for sample x [21–27].

CALCULATION RESULTS

Nine fruits juices taken from Ekpete *et al.* were used as data. The PCC matrix R was computed between plants; upper triangle is:

$$R = \begin{pmatrix} 1.000 & 0.962 & 0.975 & 0.916 & 0.994 & 0.899 & 0.346 & 0.793 & 0.992 \\ & 1.000 & 0.956 & 0.984 & 0.985 & 0.971 & 0.557 & 0.925 & 0.988 \\ & & 1.000 & 0.941 & 0.982 & 0.887 & 0.378 & 0.799 & 0.979 \\ & & & 1.000 & 0.951 & 0.965 & 0.645 & 0.952 & 0.954 \\ & & & & 1.000 & 0.931 & 0.421 & 0.846 & 1.000 \\ & & & & & 1.000 & 0.705 & 0.963 & 0.938 \\ & & & & & & 1.000 & 0.815 & 0.437 \\ & & & & & & & 1.000 & 0.856 \\ & & & & & & & & 1.000 \end{pmatrix}$$

Some correlations are high, e.g., *M. domestica*–*A. comosus* $R_{5,9} = 0.9997$; however, those including *A. muricata* are low, e.g., $R_{1,7} = 0.346$, and this fruit can be an outlier. They are illustrated in partial correlation diagram (PCD) that 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. The PCD contains 29 high (cf. Fig. 3, red), three medium (orange) and four low (yellow) partial correlations. Entry 7 (*A. muricata*) presents an only high partial correlation (with Entry 8) and can be an outlier. Pairs of fruits with high partial correlations show similar proximate and mineral compositions. The PCD is in qualitative agreement with previous results (Figs. 1 and 2).

The dendrogram of nine fruits according to 12 proximate and mineral compositions (cf. Fig. 4) shows different behaviour depending on ash, crude fibre and K. Three classes are clearly recognized:

(1,3,5,9)(2,4,6,8)(7)



Individuals *M. paradisiaca*, *C. papaya* and *M. domestica* present highest K and are grouped into class 1. Species with high *ash* and *crude fibre* are clustered into grouping 2. Fruit *A. muricata* shows high *carbohydrate*, *lipid*, Mg and Zn, and forms class 3, which should be taken with care because it involves an only fruit and can be an outlier. The fruits in the same class appear highly correlated in PCD in qualitative agreement with previous results (Figs. 1–3). However, the results should be taken with care because the branching with a unique individual (*A. muricata*) is a possible outlier.

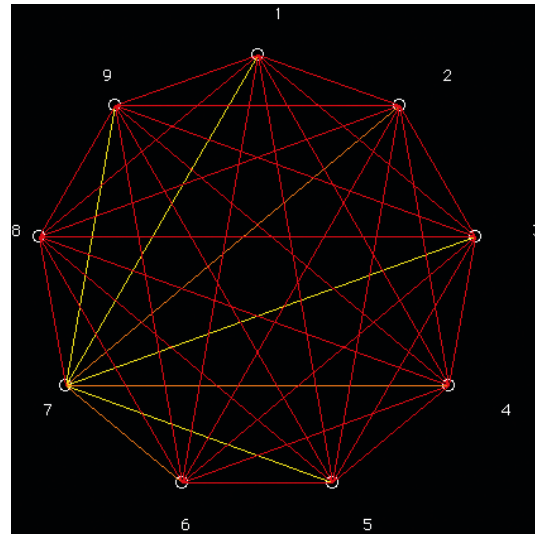


Figure 3. Partial correlation diagram: 29 high (red), three medium (orange) and four low (yellow) partial correlations.

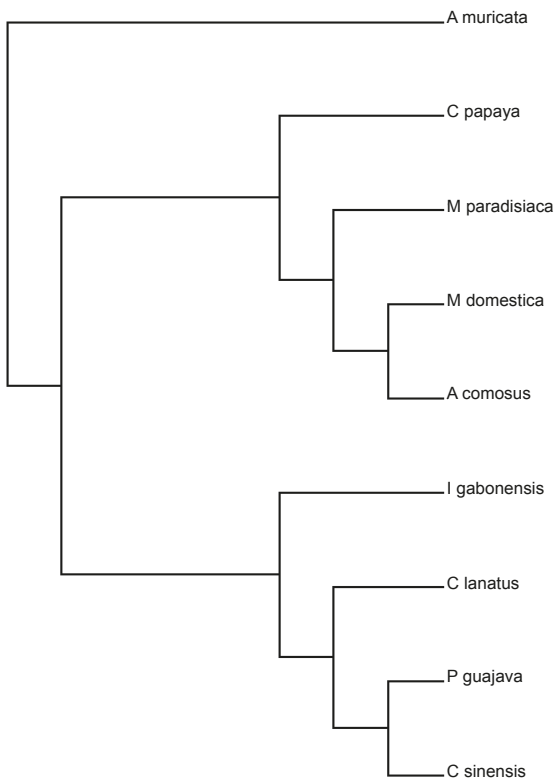


Figure 4. Dendrogram of fruits according to proximates and minerals.

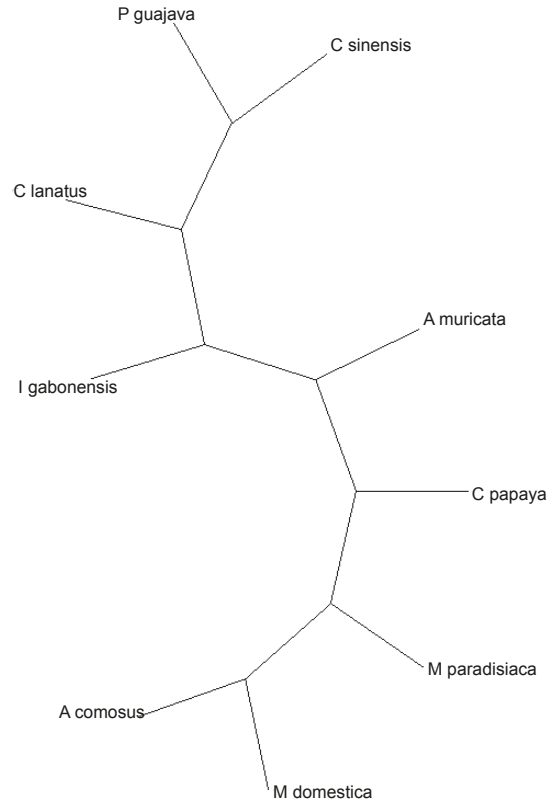


Figure 5. Radial tree of fruits according to proximates and minerals.



Radial tree (cf. Fig. 5) shows fruits different behaviour depending on *ash*, *crude fibre* and K. The same classes above are recognized in agreement with PCD, dendrogram and previous results (Figs. 1–4). Again individuals with highest K are grouped into class 1, *etc.*

The splits graph for nine fruits in Table 1 (cf. Fig. 6) reveals conflicting relations between classes 1 and 2 because of interdependences [28]. It indicates spurious relations between groupings resulting from base-composition effects. It illustrates different behaviour of fruits depending on *ash*, *crude fibre* and K. It is in qualitative agreement with PCD, binary/radial trees and previous results (Figs. 1–5).

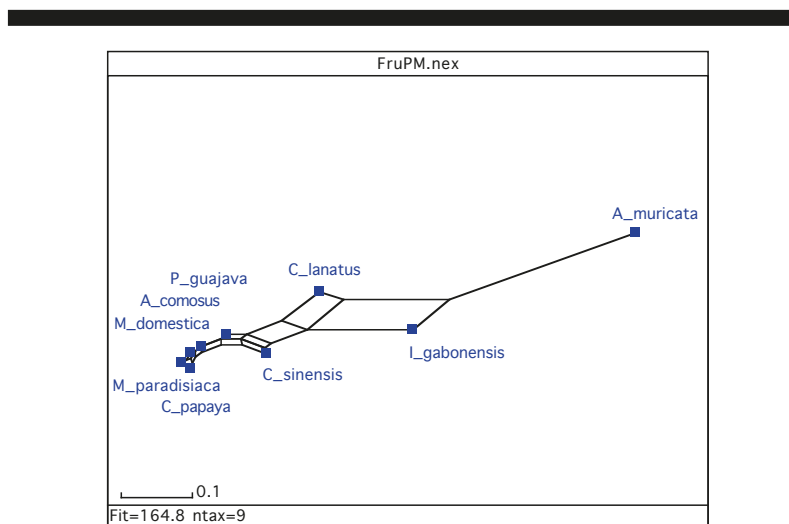


Figure 6. Splits graph of fruits according to proximate and mineral compositions.

Usually in quantitative structure–property relations (QSPRs), data contain <100 objects and >1000 X -variables. *Principal components analysis* (PCA) allows *summarizing* information contained in X -matrix. It decomposes X -matrix as product of two matrices P and T . *Loading matrix* P with information about variables contains a few vectors: PCs that are obtained as linear combinations of original X -variables. In *score matrix* T with information about objects, every object is described in terms of projections on to PCs instead of the original variables: $X = TP' + E$, where ' denotes transpose matrix. Information not contained in matrices remains *unexplained X-variance* in *residual matrix* E . Every PC_i is a new co-ordinate expressed as linear combination of the old x_j ; $PC_i = \sum_j b_{ij} x_j$. New co-ordinates PC_i are *scores (factors)* while coefficients b_{ij} are *loadings*. Scores are ordered according to information content *vs.* total variance among objects. *Score–score plots* show positions of compounds in new co-ordinate system, while *loading–loading plots* show location of features that represent compounds in new co-ordination. The PCs present properties: (1) they are extracted by decaying importance; (2) every PC is orthogonal to each other. A PCA was performed for fruits. Importance of PCA factors F_{1-12} for compositions (cf. Table 2) shows that first factor F_1 explains 39% variance (61% error), first two factors $F_{1/2}$, 59% variance (41% error), first three factors F_{1-3} , 73% variance (27% error), *etc.* For F_1 variable i_{11} shows greatest weight; however, F_1 cannot be reduced to two variables $\{i_2, i_{11}\}$ without 71% error. For F_2 variable i_4 presents greatest weight; notwithstanding, F_2 cannot be reduced to two variables $\{i_1, i_4\}$ without 45% error. For F_3 variable i_5 assigns greatest weight; nevertheless, F_3 cannot be reduced to two variables $\{i_5, i_6\}$ without 42% error. For F_4 variable i_7 consigns greatest weight; however, F_4 cannot be reduced to two variables $\{i_7, i_{10}\}$ without 54% error. For F_5 variable i_5 represents greatest weight; notwithstanding, F_5 cannot be reduced to two variables $\{i_5, i_{12}\}$ without 39% error. For F_6 variable i_8 displays greatest weight; nevertheless, F_6 cannot be reduced to two variables $\{i_8, i_9\}$ without 26% error, *etc.*

Table 2. Importance of principal component analysis factors for proximate and mineral compositions of the studied fruits.

Factor	Eigenvalue	Percentage of variance	Cumulative percentage of variance
F_1	4.62750696	38.56	38.56
F_2	2.50799682	20.90	59.46
F_3	1.56872055	13.08	72.54
F_4	1.28283270	10.69	83.23
F_5	0.93450887	7.78	91.01
F_6	0.51713185	4.31	95.32
F_7	0.38486869	3.21	98.53
F_8	0.17643357	1.47	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

In PCA F_2-F_1 scores plot of fruits (cf. Fig. 7) individuals 2 and 7 collapse. It illustrates different behaviour depending on *ash*, *crude fibre* and *K*. Three clusters are clearly distinguished: class 1 with four fruits ($F_1 < F_2$, left), grouping 2 with four plants ($F_1 > F_2$, right) and cluster 3 with one species ($F_1 \approx F_2$, middle). The diagram is in qualitative agreement with PCD, binary/radial trees, splits graph and previous results (Figs. 1–6).

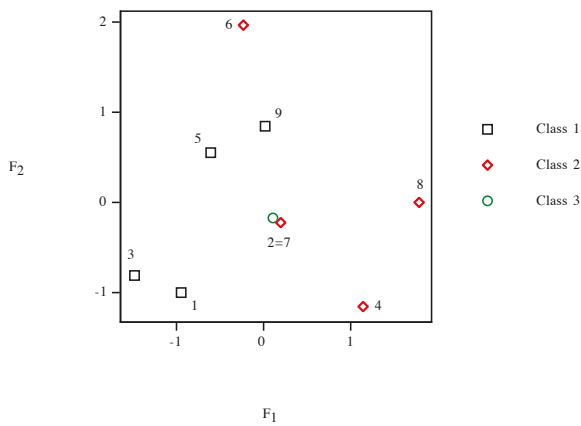


Figure 7. PCA scores plot of fruits according to components.

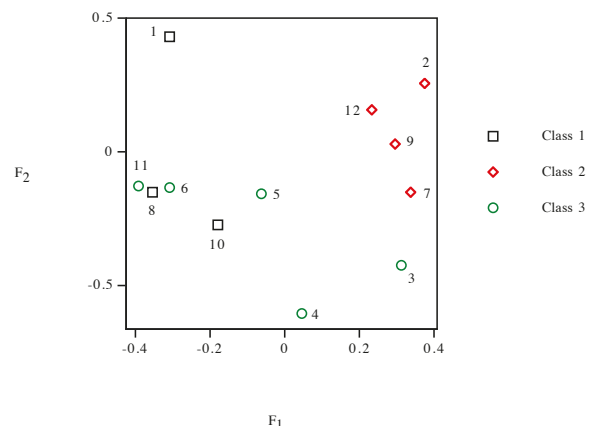


Figure 8. PCA loadings plot of fruits according to components.

From PCA factors loadings of fruits, F_2-F_1 loadings plot (cf. Fig. 8) depicts 12 proximate and mineral compositions (footnote a in Table 1). Three clusters are clearly distinguished: class 1 with three components {1,8,10} ($F_1 < F_2$, left), grouping 2 with four constituents {2,7,9,12} ($F_1 > F_2 \approx 0$, right) and cluster 3 with five substances {3,4,5,6,11} ($0 \approx F_1 > F_2$, bottom). Proximates {1–7} appear more separated than mineral components {8–12}. Most nutritional proximates (macronutrients: *carbohydrate*, *protein*, *lipid*) result grouped into class 3 and lipid is closer to protein in agreement with matrix **R**. Macronutrients are relatively separated. Mineral composition Zn appears close to Fe and Mg results closer to K than to Ca in concordance with matrix **R**. In addition, as a complement to scores diagram for loadings, it is confirmed that fruits in class 1, located at the left side, present a more pronounced



contribution from components in grouping 1 situated on the same side in Fig. 7. Fruits in cluster 2 in the right side show a contribution from components in class 2 positioned on the same side.

Instead of nine fruits in space \mathfrak{R}^{12} of 12 components, consider 12 constituents in space \mathfrak{R}^9 of nine fruits. Matrix \mathbf{R} upper triangle is:

$$\mathbf{R} = \begin{pmatrix} 1.000 & -0.216 & -1.000 & -0.658 & -0.119 & 0.195 & -0.536 & 0.354 & -0.382 & 0.212 & 0.407 & -0.169 \\ & 1.000 & 0.216 & -0.346 & 0.110 & -0.551 & 0.661 & -0.750 & 0.385 & -0.387 & -0.718 & 0.513 \\ & & 1.000 & 0.658 & 0.119 & -0.195 & 0.536 & -0.354 & 0.382 & -0.212 & -0.407 & 0.169 \\ & & & 1.000 & 0.051 & 0.053 & 0.297 & 0.113 & 0.046 & 0.539 & 0.072 & -0.171 \\ & & & & 1.000 & 0.507 & 0.032 & 0.143 & -0.193 & -0.057 & 0.162 & -0.205 \\ & & & & & 1.000 & -0.482 & 0.358 & -0.541 & 0.023 & 0.755 & -0.092 \\ & & & & & & 1.000 & -0.438 & 0.204 & 0.099 & -0.371 & 0.552 \\ & & & & & & & 1.000 & -0.321 & 0.369 & 0.642 & -0.465 \\ & & & & & & & & 1.000 & -0.299 & -0.639 & 0.100 \\ & & & & & & & & & 1.000 & 0.366 & -0.323 \\ & & & & & & & & & & 1.000 & -0.207 \\ & & & & & & & & & & & 1.000 \end{pmatrix}$$

High correlation appears between proximates *moisture–dry matter* $|R_{1,3}| = 1.0$ but low between pairs of nutritional proximates $R_{4,5}$ and $R_{4,6}$ except *protein–lipid* $R_{5,6} = 0.507$. For mineral compositions, high correlation results between $K–Mg$ $R_{8,11} = 0.642$ but low between $Fe–Zn$ $R_{9,12}$ and $Ca–Mg$ $R_{10,11}$. High correlations show not only between proximates and between mineral compositions, but also combining both types, *e.g.*, *lipid–Mg* $R_{6,11} = 0.755$. The dendrogram for 12 proximate and mineral compositions of fruits (*cf.* Fig. 9) separates the same three classes above, in agreement with PCA loadings plot and previous results (Fig. 8). Again most nutrients group into class 3 and Zn appears close to Fe.

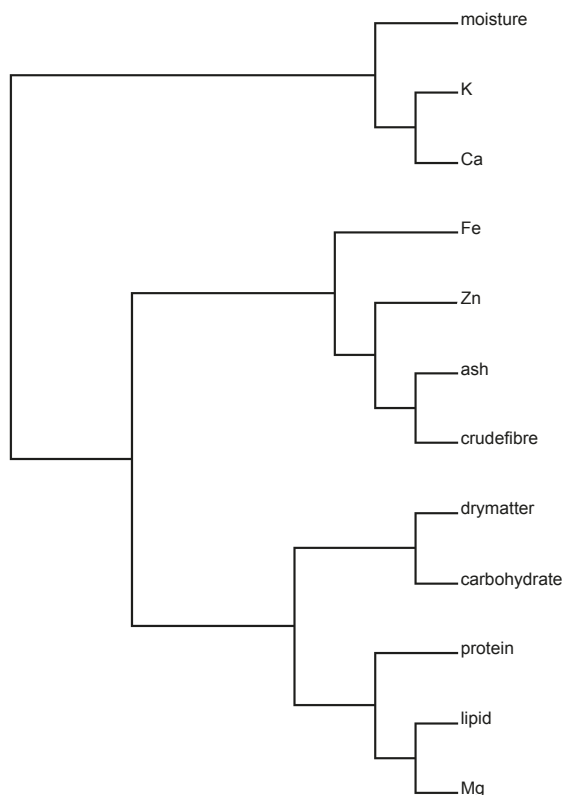


Figure 9. Dendrogram of proximate and mineral content for fruits.

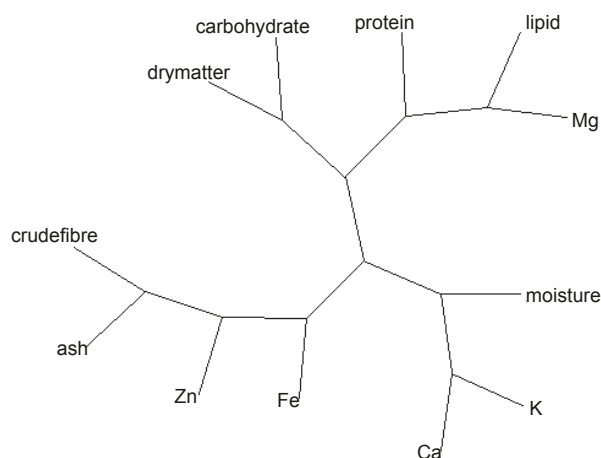


Figure 10. Radial tree of proximate and mineral content for fruits.



The radial tree of 12 proximate and mineral compositions for fruits (*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 nutrients group into class 3 and Zn is close to Fe.

Splits graph of proximate and mineral compositions for fruits (*cf.* Fig. 11) indicates conflicting relations between all classes because of interdependences. It separates the same groupings above in agreement with PCA loadings plot, binary/radial trees and previous results (Fig. 8–10). Once more most nutrients group into class 3 and Zn appears close to Fe.

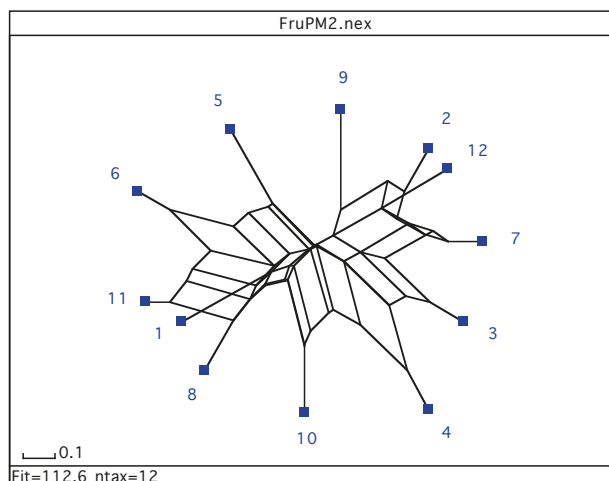


Figure 11. Splits graph of proximate and mineral content for fruits.

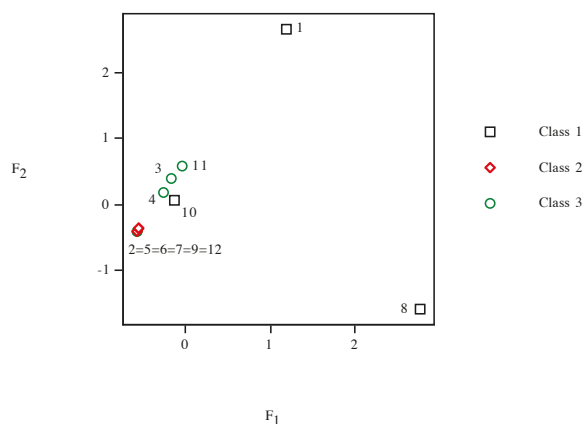


Figure 12. PCA scores plot of proximate/mineral content for fruits.

A PCA was performed for proximate and mineral compositions. Factor F_1 explains 88% variance (12% error), $F_{1/2}$ 99% variance (1% error), F_{1-3} 99.5% variance (0.5% error), *etc.* Scores plot of PCA F_2-F_1 for compositions corresponding to fruits (*cf.* Fig. 12) shows that constituents 2, 5, 6, 7, 9 and 12 collapse. Three clusters are clearly distinguished: class 1 with three components {1,8,10} ($F_1 > F_2$, *right*), grouping 2 with four constituents {2,7,9,12} ($F_1 < F_2 < 0$, *left*) and cluster 3 with five components {3,4,5,6,11} ($F_1 \ll F_2$, *middle*). Minerals {8–12} appear more separated than proximates {1–7}, most nutritional proximates group into class 3 and Zn appears close to Fe. Again macronutrients are relatively separated. The diagram separates the same classes above in qualitative agreement with PCA loadings plot, binary/radial trees, splits graph and previous results (Fig. 8–11).

DISCUSSION

Technologies development in different fields of applied science allowed the diffusion of tools based on computer-controlled instrumentation generating *big data* in a short time *via* one-step analytical procedures. Data analysis techniques exist allowing users to extract the maximum useful information from complex datasets, simplifying interpretation and resolution of the problem. The ultimate purpose of a molecular classification experiment is the extraction of physicochemical parameters (*e.g.*, proximates, minerals, carbohydrates, antinutrients, phytochemicals, antioxidants, bioprinciples contents) providing a process condensed description and, when combined with parameters obtained from other categorization experiments, are used to predict a species behaviour. In the analysis of experimental clustering data, it is necessary to introduce a model, which is difficult to decide. While some simple models will overlook particular classification details, some others, more complex with many adjustable parameters indiscriminately, fit any empirically observed behaviours. The quality of the experiential results (signal-to-noise ratio, observed errors) is the limiting factor in determining how complex the categorization model is. The primary objective of a phytochemical clustering algorithm is not only accuracy but also to understand which substances appear in plants to help scientists in cytochemistry, phytopharmacy, phytomedicine, *etc.* The molecular heterogeneity of the phytochemical compositions explains fruits morphologies, properties and functions diversity. Drafting fruits molecular profile, proximates and minerals give the basic elements accompanying constitution



and providing information for defining species/nutrient subtypes in a rational way. What was so far defined by morphology is better approached *via* complete chemical characterization, which could not replace but complement morphotypes. The question is much more than the intellectual exercise of ordering observations: conceptually it provides a fruits-composition map and facilitates reason/drawing hypotheses that will lead to understanding the nature of the nutrients, minerals and bioprinciples governing it; in practice, it presents implications for the nutrients and an atlas of fruit compositional types. The success to define fruits molecular classification should not hide that clustering data remains challenging. Many parameters influence the categorization obtained by clustering (*e.g.*, features/metric used to compare samples, clustering algorithm, procedure to select the number of groupings). Notwithstanding limitations, fruits molecular classifications obtained by automatic clustering of data revolutionized the way one apprehends fruits heterogeneity. As larger samples collections are analyzed, it is likely that finer categorizations into well-specified and robust subtypes will emerge from clustering methods and allow a more precise constituents stratification into subcategories, which would not be captured *via* only morphology. As different subgroups present dissimilar nutrients and uses, a precise and robust classification of proximates and minerals improves nutritional use.

The diet of Nigerians is high in carbohydrates and deficient in proteins. Eating five portions of fruits and vegetables a day helps people to maintain health, protecting them from heart disease, cancer, type-2 diabetes and kidney stones. The studied fruits with low antinutritional level, and high elemental composition, macronutrients (*proteins, lipids, carbohydrates*) and Vit-C serve as supplementary sources of essential nutrient to man and livestock being indicated in this work and technical literature. The essential elements are needed for metabolism. The healthy biofunctions are influenced by necessary elements: too little lead to deficiency disease and too much are toxic. Food effects result of major components and minor-constituents synergism/antagonism. (1) Vitamin-C in fruits is a strong Fe promoter. It enhances non-haeme-Fe absorption. (2) Phytic acid in foods binds essential mineral nutrients in the digestive tract causing deficiencies. It binds P converting it to phytate while metals (Ca, Zn, Mn, Fe, Mg) are changed to phytic complexes being indigestible substance decreasing elements bioavailability for absorption. It presents a negative effect on AA digestibility, posing problem to non-ruminants because of insufficient amount of intrinsic phytase necessary to hydrolyze the phytic acid complex but its presence results beneficiary because of its positive nutritional role as an antioxidant and anticancer agent. (3) Oxalate has negative effect on mineral availability. (4) Polyphenol *tannin* acts as antioxidant preventing oxidative stress causing illnesses (*e.g.*, coronary heart disease, cancer, inflammation, ageing). Plants synthesize a wide variety of chemical compounds being sorted by chemical class, biosynthetic origin and functional groups into primary and secondary metabolites. Chemical composition was employed in devising taxonomic classification. Chemotaxonomic studies showed distinguishing phytoplankton communities in aquatic systems being an economical approach bearing advantage for a large-ecosystem study. Bamboos evolution was investigated and results derived from leaf wax *n*-alkanes were consistent with morphology. Citrus and related-genera categorization was studied based on chemotaxonomic analysis from leaf, peel and flower oils, and the results (pomelo, citron, mandarin) were consistent with morphology and genetics. The studies and present work suggested that chemotaxonomic analysis was reliable, revealing a research tool for organizational investigations. The chemical signatures were analyzed. Obvious candidates, *e.g.*, nutritional proximates (macronutrients, *etc.*), made the list but the technique highlighted a swath of lesser-known compounds [*e.g.*, minerals (non/toxic metals, *etc.*), carbohydrates (total/reducing sugars, starch, Vit-C, *etc.*), antinutrients (tannin, trypsin inhibitor, phytate, oxalate, *etc.*)]. High correlations are obtained because only fruit juices were considered. However, the results changed if different fruit parts (seed, mesocarp, pulp, peel, *etc.*) or whole fruits/plants were considered (*e.g.*, in seeds, protein is greater and Vit-C is lesser than in mesocarp, mango peel contains 2–6-fold Vit-C than pulp).

A worldwide trend exists to the use of the natural phytochemicals present in berry crops, teas, herbs, oil seeds, beans, fruits and vegetables. Fruits chemical compositions vary with region because of geography. The studied fruits contain proximates, minerals and phytochemicals being linked to beneficial effects on health. The reported data show that the fruits result rich in nutrients. Low level of antinutrients and high quantity of Vit-C, indicated in this work and the bibliography, enhance mineral availability in composite meals. Too much Zn is toxic, risk rises in contaminated fruits grown in polluted industrial/residential areas. Equipment type used to produce and its quality to store fruits, after harvesting, are possible sources of contamination with trace metals; *e.g.*, storing fruits in galvanized containers is Zn-pollution source. Further study is needed to determine the digestibility and bioavailability of these plant foods. The work can be extended to additional mineral elements (Na, Mn, Cu), carbohydrates and antinutrients. An expansion of this research to other fruits and plants is desired. If a fruit juice could not serve as nutrient, the investigation should determine if some different part of the fruit (seed, mesocarp, pulp, juice, peel) or plant (root, stem, stem bark, leaf, flower) would work. Further study should be carried out to isolate, purify and characterize the functional constituents responsible for the fruits activity. Need exists to search for medicinal plants with the aim of validating compounds ethnomedicinal use and characterization, which will be added to the pharmacopoeia. Phytochemical analysis reveals flavonoids, anthraquinones, saponins, phenols, tannin, alkaloids, cardiac glycosides, terpenoids, *etc.* Work is encouraged to elucidate action mechanism of fruits constituents. The research would assess the generality of this discussion.



CONCLUSIONS

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

1. Never blindly trust what you get (do your results have chemical sense?).
2. Mathematical and statistical models are not the panacea (always ask yourself three questions: Why am I doing it? What the results might be? Will I be successful?).
3. Use this knowledge in data analysis to guide your investigation or experimentation, not as an end in itself.
4. Criteria reduced analysis to a manageable quantity from enormous set of fruits compositions: they refer to proximates and minerals. Meta-analysis was useful to rise numbers of samples and variety of analyzed data. Different behaviour of fruits depends on *ash*, *crude fibre* and *K*. With regard to components, most nutritional proximates grouped into the same class and *Zn* was close to *Fe*.
5. Principal components analyses of compositions and fruits cluster analyses allowed classifying them and agreed. A cluster with one only individual, *Annona muricata*, is a possible outlier. Phytochemistry and understanding of computational methods are essential for tackling associated *data mining* tasks.

REFERENCES

- [1] M. O. Bello, O. S. Falade, S. R. A. Adewusi, N. O. Olawore, Studies on the chemical compositions and antinutrients of some lesser known Nigeria fruits, *Afr. J. Biotechnol.* **7** (2008) 3972-3979.
- [2] D. E. Okwu, I. N. Emelike, Evaluation of the phytonutrients and vitamin content of citrus fruits, *Int. J. Mol. Med. Adv. Sci.* **2** (2006) 1-6.
- [3] V. O. Onibon, F. O. Abulude, L. O. Lawal, Nutritional and anti-nutritional composition of some Nigerian fruits, *J. Food Technol.* **5** (2007) 120-122.
- [4] O. O. Dosumu, O. O. Oluwaniyi, G. V. Awolola, M. V. Okunola, Stability studies and mineral concentration of some Nigerian packed fruits juices, concentrate and local beverages, *Afr. J. Food Sci.* **3** (2009) 82-85.
- [5] G. A. Dimari, S. S. Hati, Vitamin C composition and mineral content of some Nigerian packaged juices drinks, *J. Life Phys. Sci.* **3** (2010) 129-134.
- [6] E. A. Oluyemi, A. A. Akinlua, A. A. Adenuga, M. B. Adebayo, Mineral content of some commonly consumed Nigerian foods, *Sci. Focus* **11** (2006) 153-157.
- [7] M. F. Waller, E. M. Haymes, The effects of heat and exercise on sweat iron loss, *Med. Sci. Sports Exercise* **28** (1996) 197-203.
- [8] U. Melaku, E. W. Clive, F. Habtamon, Content of zinc, iron, calcium and their absorption inhibitors in foods commonly consumed in Ethiopia, *J. Food Compos. Anal.* **18** (2005) 803-817.
- [9] O. A. Ekpete, O. S. Edori, Proximate and mineral composition of some Nigerian fruits, *Int. J. Res. Rev. Pharm. Appl. Sci.* **3** (2013) 287-295.
- [10] O. A. Ekpete, O. S. Edori, E. P. Fubara, Proximate and mineral composition of some Nigerian fruits, *Br. J. Appl. Sci. Technol.* **3** (2013) 1447-1454.
- [11] F. Torrens-Zaragozá, Molecular categorization of yams by principal component and cluster analyses, *Nereis* (5) (2013) 41-51.
- [12] F. Torrens-Zaragozá, Classification of lactic acid bacteria against cytokine immune modulation, *Nereis* (6) (2014) 27-37.
- [13] H. Hotelling, Analysis of a complex of statistical variables into principal components, *J. Educ. Psychol.* **24** (1933) 417-441.
- [14] R. Kramer, *Chemometric techniques for quantitative analysis*, Marcel Dekker, New York, 1998.
- [15] S. K. Patra, A. K. Mandal, M. K. J. Pal, *Photochem. Photobiol., Sect. A* **122** (1999) 23-31.
- [16] I. T. Jolliffe, *Principal component analysis*, Springer, New York, 2002.
- [17] J. Xu, A. Hagler, *Molecules* **7** (2002) 566-600.
- [18] P. J. A. Shaw, *Multivariate statistics for the environmental sciences*, Hodder-Arnold, New York, 2003.
- [19] IMSL, *Integrated Mathematical Statistical Library (IMSL)*, IMSL, Houston, 1989.
- [20] R. C. J. Tryon, *Chronic Dis.* **20** (1939) 511-524.
- [21] I. Priness, O. Maimon, I. Ben-Gal, Evaluation of gene-expression clustering via mutual information distance measure, *BMC Bioinformatics* **8** (2007) 111-1-12.
- [22] R. Steuer, J. Kurths, C. O. Daub, J. Weise, J. Selbig, The mutual information: Detecting and evaluating dependencies between variables, *Bioinformatics* **18** (Suppl. 2) (2002) S231-S240.
- [23] P. D'Haeseleer, S. Liang, R. Somogyi, Genetic network inference: From co-expression clustering to reverse engineering, *Bioinformatics* **16** (2000) 707-726.
- [24] C. M. Perou, T. Sørlie, M. B. van Eisen, M. de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lønning, A. L. Børresen-Dale, P. O. Brown, D. Botstein, Molecular portraits of human breast tumours, *Nature (London)* **406** (2000) 747-752.
- [25] R. A. Jarvis, E. A. Patrick, Clustering using a similarity measure based on shared nearest neighbors, *IEEE Trans. Comput.* **C22** (1973) 1025-1034.



- [26] R. D. M. Page, *Program TreeView*, University of Glasgow, UK, 2000.
- [27] M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. U. S. A.* **95** (1998) 14863-14868.
- [28] D. H. Huson, SplitsTree: Analyzing and visualizing evolutionary data, *Bioinformatics* **14** (1998) 68-73.

ACKNOWLEDGEMENT

The author thanks support from the Spanish Ministerio de Economía y Competitividad (Project No. BFU2013-41648-P) and EU ERDF.

