Inter*net* Electronic Journal of Molecular Design

August 2003, Volume 2, Number 8, Pages 546–563

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Nenad Trinajstić on the occasion of the 65th birthday
Part 2

Guest Editors: Douglas J. Klein and Sonja Nikolić

Principal Component Analysis of New Structural Parameters for Fullerenes

Francisco Torrens

Institut Universitari de Ciència Molecular, Universitat de València, Dr. Moliner 50, E–46100 Burjassot (València), Spain

Received: March 25, 2003; Revised: May 13, 2003; Accepted: May 17, 2003; Published: August 31, 2003

Citation of the article:
F. Torrens, Principal Component Analysis of New Structural Parameters for Fullerenes, *Internet Electron. J. Mol. Des.* **2003**, *2*, 546–563, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Principal Component Analysis of New Structural Parameters for Fullerenes[#]

Francisco Torrens*

Institut Universitari de Ciència Molecular, Universitat de València, Dr. Moliner 50, E–46100 Burjassot (València), Spain

Received: March 25, 2003; Revised: May 13, 2003; Accepted: May 17, 2003; Published: August 31, 2003

Internet Electron. J. Mol. Des. **2003**, *2* (8), 546–563

Abstract

Motivation. Novel carbon allotropes, with finite molecular structure, including spherical fullerenes are nowadays currently produced and investigated. The Kekulé structure count and permanent of the adjacency matrix of these molecules are related to structural parameters involving the presence of contiguous pentagons. The close relationship between these parameters suggests considering also their quotients.

Method. Both single– and complete–linkage cluster analyses of the structural parameters allow classifying these parameters. PCA (principal component analysis) of the structural parameters and the cluster analyses of the fullerenes permits classifying these molecules.

Results. Cluster analysis provides a binary taxonomy of the structural parameters that separates first the *r*–*r*/*p* from the *p*–*q*–*q*/*p* parameters. PCA clearly distinguishes five classes of fullerenes. The cluster analysis of fullerenes is in agreement with the PCA classification.

Conclusions. Cluster analysis shows that the greatest similarity is between the *q*–*q*/*p* and *r*–*r*/*p* pairs of parameters. Split decomposition indicates a spurious relationship resulting from base composition effects. PCA provides five orthogonal factors F_1–F_5. The use of F_1 gives an error of 28%. The use F_1 and F_2 decreases the error to 2%. PCA groups the fullerenes in five classes. Some fullerenes with different numbers of atoms belong to the same class, while some fullerene isomers are members of different classes.

Keywords. Cluster analysis; dendrogram; split decomposition; principal component analysis; PCA; similarity matrix; fullerene.

1 INTRODUCTION

Multivariate data often consist of sets of high–dimensional vectors. In chemical applications, a vector could be a series of physical measurements or calculated properties made on a molecule. A dataset of compounds may be a series of related molecules collected for, *e.g.*, a structure–activity study. If the vectors are only two–dimensional (2D), they can be plotted on a plane. This allows the visual inspection of the structure of the dataset to identify clusters and particular objects, *i.e.*, to

[#] Dedicated to Professor Nenad Trinajstić on the occasion of the 65th birthday.

* Correspondence author; phone: 34–963–543–182; fax: 34–963–543–156; E–mail: Francisco.Torrens@uv.es.

perform an exploratory data analysis.

When dealing with vectors whose dimensions are larger than two, it is not possible to represent them graphically on a plane. One way to overcome this problem is to transform the *N*–dimensional vectors into 2D. Many projection methods were developed for this task. A good projection method preserves as faithfully as possible the original structure of the high–dimensional data. Unfortunately, the true distances between the vectors in the original high–dimensional space cannot be preserved exactly in the projected 2D display. The 2D plot thus obtained must distort in some way the original picture. Such distortions can cause misleading plots. Among the many papers concerned with the projection of multivariate data, the checking of the projections remains mostly an exception. Projection algorithms can be either supervised or unsupervised. Because this article deals with exploratory data structure analysis, only unsupervised methods are used. Unsupervised algorithms can be either linear (*e.g.*, principal component analysis) or non–linear (*e.g.*, non–linear mapping, self–organizing map). Comparisons of the quality of projection methods were described elsewhere [1–6].

Principal component analysis (PCA) is probably one of the most popular projection methods [7]. Its principal feature is to rotate the vector space using the eigenvectors (principal components, PCs or factors) of the covariance matrix as a new basis [8]. PCs corresponding to the two largest eigenvalues (variance) are used to produce 2D plots [9]. The quality of the projection is commonly expressed by the retained variance of the first two PCs. In addition, plots of other PCs, such as the first against the third, *etc*., might be useful. PCA facilitates the statistical analysis, but the interpretation is obscured as each new variable results from the combination of others. In order to illustrate the usefulness of this method, PCA is applied to a dataset of 31 fullerenes represented by five structural parameters. For this example, PCA projection method is applied. On the other hand, a method is described for cluster analysis (CA) data. The relative efficiency of CA algorithms and similarity descriptors was the subject of several articles [10–12].

In previous works, the calculation of the Kekulé structure count and the permanent of adjacency matrices [13] were applied to fullerenes with different structural parameters involving the presence of contiguous pentagons [14]. PCA of the structural parameters was carried out [15,16]. In this report, two new parameters have been introduced. The aim of this report is to analyse the interdependence between the structural parameters, to classify them and to classify the fullerenes. Section 2 presents the computational methods. Section 3 discusses the calculation results for fullerenes. Section 4 summarizes the conclusions.

2 COMPUTATIONAL METHODS

2.1 Principal Component Analysis

PCA is used to transform a number of potentially correlated variables into the same number of independent variables, which can then be ranked based upon their contributions for explaining the whole data set. The transformed variables that can explain all the information in the data are called principal components (PCs) or factors. The first PC, F_1, accounts for as much of the variability in the data as possible and each succeeding component, F_i, accounts for as munch of the remaining variability as possible. PCs having minor contribution to the data set may be discarded without losing too much information. If the number of PCs is less than four then the multidimensional data can be graphed in 2D or 3D space, *i.e.*, PCA can be used to reduce dimensionality. The main purpose of employing PCA is to reduce the number of variables (PCs) used in the analysis. PCA creates new variables as linear combinations of all the initial variables so that the first PC contains the largest variance, the second PC contains the second largest variance, and so on, until the last PC can be truncated. PCA also allows diminishing the number of total variables in a data set.

The comparison of the measures of two different variables has no sense. However, the initial measures can be transformed: the N values of the *j–th* variable are transformed using the mean $\bar{x}_j$ and standard deviation σ_j of this *j–th* variable. In fact, the converted value is:

$$x'_{ij} = \left(x_{ij} - \bar{x}_j\right) \Big/ \sigma_j \tag{1}$$

PCA, which consists in finding the eigenvalues and eigenvectors of the covariance matrix, proceeds the standardized variables to diagonalize the correlation matrix of the initial variables. In effect, PCs have the form:

$$F_i = \sum_{k=1}^{P} C_{ik} x'_k \tag{2}$$

On the (F_1, F_2) plane, each point (variable) k has as coordinates some numbers proportional to the C_{1k} and C_{2k} coefficients of the PCs F_1 and F_2. The profile of a PC F_i is the vector of the squared C_{ik} coefficients (C_{i1}^2, C_{i2}^2,…, C_{iP}^2). Each C_{ik}^2 represents the weight of variable k in PC F_i. It gives the fraction of each variable in PC F_i.

2.2 Cluster Analysis

The term cluster analysis (CA) was first used by Tryon, in 1939 [17]. Actually, CA encompasses a number of different classification algorithms. A general question in many areas of an inquiry is how to organize the observed data into meaningful structures, *i.e.*, how to develop taxonomies. Conceptually, the approach used by CA to address this problem can well be described by the saying birds of a feather flock together. Since its initial introduction, many CA algorithms have been invented. They belong to two categories: hierarchical cluster analysis (HCA) and non–hierarchical

(partitional) cluster analysis (NHCA) [18]. HCA rearranges objects in a tree–structure. In HCA, either the database is divided successively until a predetermined number of clusters have been created, or members are successively grouped together until the predetermined number of clusters has been assembled. In either case, a dendrogram (binary tree) is created that maps *N* members in one cluster to *N* members in *N* clusters. In NHCA, a nearest–neighbour list is created and used to assemble members into related clusters. An example of this is the Jarvis–Patrick NHCA algorithm, which has been widely used to cluster chemical structures and structural databases. In HCA, each object (*e.g.*, the 31 studied fullerenes) is initially assumed to be a lone cluster. A distance matrix is built, generally calculating the Euclidean distance between all the objects and scanned for the minor values. The corresponding objects are clustered together and treated as a single cluster. Successive iterations lead to the total clustering of all objects, generating a dendrogram with the objects clustered together according to their similarity level.

Correct CA results rely on: (*a*) proper structure representation (bioactivity–related descriptors), (*b*) suitable data normalization, and (*c*) carefully selected CA algorithms and proper parameter settings. Data normalization is the basis for comparing experiments with large series when experimental conditions may not be identical. Normalization ensures that the experimental quality of the data is comparable and sound mathematical algorithms were employed. Normalization includes various options to standardize data, and to adjust background levels and correct gradients. The commonly used normalization functions follow.

1. Linear normalization:

$$x'_i = X'_{\min} + \frac{(X'_{\max} - X'_{\min})(x_i - X_{\min})}{X_{\max} - X_{\min}} \tag{3}$$

2. Ratio normalization:

$$x'_i = \frac{x_i}{\sum_{i=1}^{n} |x_i|} \tag{4}$$

3. Z–score normalization:

$$x'_i = \frac{x_i - \bar{x}}{\delta} \tag{5}$$

where δ is the standard deviation. Generally, linear normalization is recommended [if $X'_{\max} = 1$ and $X'_{\min} = 0$, x'_i is normalized in percentage by Eq. (3)]. Z–score assumes x_i obeys Gaussian distribution. If x_i has a different distribution, then the normalization will twist the pattern (variance will be far away from the standard deviation) and leads to incorrect pattern recognition. One of the puzzling problems of CA algorithms is that they require a user in some ways to guess the number of clusters before carrying out the CA computation. In addition, CA cannot tolerate the heterogeneity

of the data.

There are many reasons why one might want to cluster a database of molecular structures [19]. Two of the most practical reasons are to identify representative compounds from a structural database or virtual compound library for screening or synthesis [20]. It is sometimes useful just to be able to determine if a database offering is rather diverse, or if most of the structures fall into a small number of homologous structural classes [21]. Three objectives must be in mind when designing a CA algorithm [22]. (*a*) A method would divide a database into an appropriate number of clusters based on the structures and their relative similarity, rather than some predefined number. Having to specify the number of clusters is a significant shortcoming of most CA algorithms, which create a defined number of clusters, without regard to the fact that this sometimes requires grouping very unlike structures together. (*b*) A method would allow clustering additional structures without starting from scratch. This objective requires an algorithm that can begin with a set of clusters and add future structures to existing clusters, or create new clusters as their structural topology dictates. (*c*) Any method has to be computationally tenable for very large structural databases. Speed is one of the most significant problems with HCA, but even the more efficient NHCA scale formally as N^2.

A program has been written using the IMSL [23] subroutine CLINK to carry out HCA based upon either a distance or a similarity (*e.g.*, correlation) matrix. Initially, each data point is considered to be a cluster, numbered 1 to $n = N_{pt}$, where N_{pt} is the number of data points to be clustered. HCA proceeds in four steps.

Step 1. If the data matrix contains similarities they are converted to distances.

Step 2. A search is made of the distance matrix to find the two closest clusters. These clusters are merged to form a new cluster, numbered $n + k$.

Step 3. Based upon the method of CA, updating of the distance measure corresponding to the new cluster is performed.

Step 4. Set $k = k + 1$. If $k < n$, go to step 2.

The procedure allows two methods of computing the distances between clusters. The single–(SLHCA) and complete–linkage hierarchical cluster analyses (CLHCA) differ primarily in how the distance matrix is updated, after two clusters have been joined. To understand these measures, suppose in the following discussion that clusters *A* and *B* have just been joined to form cluster *Z*, and interest is in computing the distance of *Z* with another cluster called *C*. In SLHCA, the distance from *Z* to *C* is the minimum of the distances (*A* to *C*, *B* to *C*). In CLHCA, the distance from *Z* to *C* is the maximum of the distances (*A* to *C*, *B* to *C*). In general, SLHCA will yield long thin clusters, while CLHCA will yield clusters that are more spherical.

2.3 Leave–*n*–Out

In modelling, it is essential to determine the complexity of the model to avoid overfitting. The predictive capability of the resulting model depends on the quality of the data (the more and better the data available, the more accurate prediction is possible), and on the number k of significant latent necessary variables. Cross–validation is a practical and reliable method for testing this significance. The leave–one–out approach consists in developing a number of models with one sample omitted at the time. After developing each model, the omitted data are predicted and the differences between actual and predicted y (*e.g.*, ln[per(**A**)]/ln K) values are calculated. Leave–one–out can be generalized [leave–many(n)–out] for the cross–validated properties that will be obtained when n fullerenes are being separated from the original group. Leave–n–out protocols are more adequate to obtain significant and optimal results. A result obtained by cross–validation possesses some intrinsic robustness, and even more if a leave–n–out protocol has been considered [24].

3 CALCULATION RESULTS AND DISCUSSION

The structural features involving adjacent pentagons are encoded by the parameters p, q and r. The counts p and q enumerate, respectively, the number of edges common to two pentagons and the number of vertices common to three pentagons [25]. The count r enumerates the number of pairs of non–adjacent pentagon edges shared with two other pentagons [26]. Thus, q and r complement each other by counting both possible arrangements of three contiguous pentagons. However, there are close relationships between p and q, and between p and r. For instance, the minimum structure with $q = 1$ needs $p = 3$, and the minimum structure with $r = 1$ requires $p = 2$. The interdependences p–q and p–r suggest expanding the count set of a previous work [15] with the quotients q/p and r/p.

Table 1. Values of p, q and r Counts for Fullerenes

Fullerene	K	per(**A**)	ln[per(**A**)]/ln K	q/p	r/p
C_{20} (I_h)	36	1392	2.0199	0.6667	1.0000
C_{24} (D_{6d})	54	4692	2.1192	0.5000	1.5000
C_{26} (D_{3h})	63	8553	2.1853	0.3810	1.4286
C_{28} (T_d)	75	15705	2.2378	0.2222	1.3333
C_{28} (D_2)	90	16196	2.1540	0.4000	1.2000
C_{30} (C_{2v}) I	107	29621	2.2034	0.2353	1.1765
C_{30} (C_{2v}) II	117	30053	2.1651	0.3333	1.1111
C_{30} (D_{5h})	151	31945	2.0672	0.5000	1.0000
C_{32} (D_3)	144	55140	2.1968	0.1333	1.2000
C_{32} (C_2) I	151	55705	2.1780	0.2500	1.0000
C_{32} (C_2) II	168	57092	2.1375	0.3529	0.9412
C_{32} (D_2)	184	58384	2.1045	0.4444	0.8333
C_{34} (C_{3v})	195	103665	2.1902	0.2000	1.0000
C_{34} (C_s)	196	104484	2.1896	0.2000	1.0667
C_{34} (C_2) I	204	103544	2.1714	0.1429	1.0000
C_{34} (C_2) II	212	107720	2.1632	0.3529	0.9412
C_{36} (D_{6h})	272	192528	2.1706	0.0000	1.0000

Table 1. (Continued)

Fullerene	*K*	per(**A**)	ln[per(**A**)]/ln *K*	q/p	r/p
C_{36} (D_{2d})	288	192720	2.1489	0.0000	1.0000
C_{36} (C_{2v})	312	197340	2.1231	0.1538	0.7692
C_{36} (D_{3h})	364	207924	2.0764	0.4000	0.4000
C_{38} (C_{2v})	360	366820	2.1768	0.1429	1.0000
C_{38} (C_{3v})	378	363300	2.1572	0.0833	0.7500
C_{38} (D_{3h})	456	411768	2.1116	0.4444	1.0000
C_{40} (D_{5d}) I	562	515781	2.0775	0.0000	1.0000
C_{40} (T_d)	576	704640	2.1185	0.3333	0.0000
C_{40} (D_{5d}) II	701	803177	2.0750	0.5000	1.0000
C_{44} (T)	864	2478744	2.1775	0.3333	0.0000
C_{44} (D_{3h})	960	2436480	2.1416	0.2222	0.0000
C_{60} (I_h)	12500	395974320	2.0986	–	–
C_{70} (D_{5h})	52168	–	–	–	–
C_{82} (C_s)	–	–	–	–	–

The values for the new structural parameters involving the presence of contiguous pentagons are listed in Table 1. Much chemical graph–theory work revolved around the adjacency matrices **A** of the compounds under investigation. The determinant of the 3×3 matrix [*a b c*, *d e f*, *g h i*] is *aei* – *ahf* – *dbi* + *dhc* + *gbf* – *gec*. The permanent of this matrix, per(**A**), is the *sum* of the same six terms. *K* denotes the Kekulé structure count. A motivation for the consideration of *K* is that *K* is never zero for fullerenes [27]. Per(**A**) is bounded below by K^2. As per(**A**) and *K* increase exponentially with system size, several authors used their logarithms. Cash selected a group of 27 fullerenes (included in Table 1) to correlate ln[per(**A**)]/ln *K*, ln *K* and ln[per(**A**)] with the structural parameters *p*, *q* and *r*. Despite the good results obtained by Cash, three important remarks were made. (1) Counts *p*, *q* and *r* include some redundant information. (2) The error of some fitted parameters is large. (3) Non–linear effects of *p*, *q* and *r* can affect ln[per(**A**)]/ln *K*, ln *K* or ln[per(**A**)]. Therefore, a different strategy was used. (1) Smaller superpositions of the pairs *p–q* and *p–r* were sought. (2) Not all the three counts *p–q–r* were necessarily retained in the fits. (3) Non–linear correlations were allowed.

The best linear correlation of ln[per(**A**)]/ln *K* with *p*, *q* and *r* for the first 29 fullerenes in Table 1 results:

$$\ln[\mathrm{per}(\mathbf{A})]/\ln K = 2.14 - 0.0108q + 0.00364r \qquad (6)$$

$$n = 29 \quad R = 0.721 \quad s = 0.036 \quad F = 14.1 \quad \mathrm{MAPE} = 1.21\% \quad \mathrm{AEV} = 0.4803$$

The mean absolute percentage error (MAPE) is 1.21% and the approximation error variance (AEV) is 0.4803. There are general *degeneracy* problems with trying to fit per(**A**) and *K* with the structural invariants *p*, *q* and *r*. Even with restriction to fullerenes there are numerous cases of whole families of fullerenes with exactly the same values for *p*, *q* and *r*, yet with rather widely varying values of per(**A**) and *K*. For instance, bucky–tubes with fixed fullerenic caps but of varying length are fullerenes all with the same values of *p*, *q* and *r*, while the values of per(**A**) and *K* increase without bound as the length of the tubes are increased. In addition, fairly large fullerenes surely almost all have $p = q = r = 0$, although the values for per(**A**) and *K* increase exponentially

with N (the number of sites of the fullerene). As N has not been included in the correlations, the applicability of the present fits is restricted to smaller fullerenes ($N < 70$). All other models with greater MAPE and AEV have been discarded. As there are several fullerenes with the same set of counts p, q and r, Equation (6) explains 95% of the correlation coefficient of the means ($n = 24$, $R = 0.757$). On the other hand, the best quadratic correlation of ln[per(**A**)]/ln K with quadratic functions of p, q and r gives:

$$\begin{aligned} \ln[\mathrm{per}(\mathbf{A})]/\ln K &= 2.13 + 0.0515 z_{41} \\ z_{41} &= 0.225 z_{31} + 1.20 z_{32} \\ z_{31} &= -1.16 + 0.232 q \\ z_{32} &= 1.05 z_{22} - 0.875 z_{21} z_{22} \\ z_{21} &= 1.22 - 0.0983 r + 0.00277 qr \\ z_{22} &= -0.726 z_{11} - 0.921 z_{12} \\ z_{11} &= -1.16 + 0.232 q \\ z_{12} &= 1.22 - 0.0983 r + 0.00277 qr \\ \mathrm{MAPE} &= 0.87\% \quad \mathrm{AEV} = 0.2432 \end{aligned} \tag{7}$$

and AEV decreases 49%.

If q/p and r/p are included in the model the best linear fit for the first 28 fullerenes in Table 1 results:

$$\begin{gathered} \ln[\mathrm{per}(\mathbf{A})]/\ln K = 1.88 + 0.0361 p - 0.0490 q + 0.00953 r + 0.0497 q/p - 0.253 r/p \\ n = 28 \quad R = 0.941 \quad s = 0.019 \quad F = 34.2 \quad \mathrm{MAPE} = 0.66\% \quad \mathrm{AEV} = 0.1558 \end{gathered} \tag{8}$$

and AEV decreases 68%. Eq. (8) explains 98% of the correlation coefficient of the means ($n = 23$, $R = 0.956$). The best non–linear model does not improve the results.

There are already powerful exact computational approaches for K, which are fairly reasonable or general. For arbitrary chemical graphs enumeration *via* Heilbronner recursion is feasible up to ~90 atoms. Better efficiency is found with Kasteleyn's method, which is generally applicable for all planar graphs (including all fullerenes). This simply involves the evaluation of the determinant of a signed adjacency matrix **A'** [28], where the method is neatly extended to deal with conjugated circuit counts, simply using the inverse of **A'**. Indeed this has been applied for fullerenes of up to 980 atoms [29], and even fullerenes of up to at least 2000 atoms could presumably be similarly treated if desired. The method has also been applied for infinite translationally symmetric networks (*i.e.*, with a finite number of sites per unit cell) [30]. For ln K alone, the best linear correlation for the first 30 fullerenes in Table 1 results:

$$\begin{gathered} \ln K = 10.1 - 0.376 p + 0.255 q \\ n = 30 \quad R = 0.965 \quad s = 0.401 \quad F = 181.6 \quad \mathrm{MAPE} = 4.21\% \quad \mathrm{AEV} = 0.0692 \end{gathered} \tag{9}$$

Eq. (9) explains 98% of the correlation coefficient of the means ($n = 24$, $R = 0.982$). The use of non–linear models or the inclusion of q/p and r/p does not improve the results.

For ln[per(**A**)] alone, the best linear correlation for the first 29 fullerenes in Table 1 results:

$$\ln[\mathrm{per}(\mathbf{A})] = 20.2 - 0.660\,p + 0.383\,q$$
$$n = 29 \quad R = 0.949 \quad s = 0.757 \quad F = 118.5 \quad \mathrm{MAPE} = 4.05\% \quad \mathrm{AEV} = 0.0988 \tag{10}$$

Eq. (10) explains 97% of the correlation coefficient of the means ($n = 24$, $R = 0.977$). On the other hand, the best quadratic correlation gives:

$$\ln[\mathrm{per}(\mathbf{A})] = 20.0 - 0.666\,p + 0.616\,q - 0.00850\,pq$$
$$\mathrm{MAPE} = 3.91\% \quad \mathrm{AEV} = 0.0871 \tag{11}$$

and AEV decreases 12% with respect to the linear fit. The inclusion of the q/p and r/p indices does not improve the results.

No superposition of the variables p–q or p–r is observed in Eqs. (6) and (7). This diminishes the risk of co–linearity in the fit given the close relationships p–q and p–r [31]. The signs and magnitudes of the coefficients in Eqs. (6)–(11) are of some interest. One would intuitively expect that, for some property determined in part by the presence of abutting pentagons p, an arrangement such as q would make less of a contribution than would three isolated p–type pairs of pentagons. If this is true, then the sign of the q coefficient would be opposite that of the p coefficient, as is the case in Eqs. (8)–(11).

By the same sort of argument, one would expect the magnitude of the r coefficient to be smaller than that of the q coefficient on the assumption that an r–type cluster is intermediate in properties between a q–type cluster and two isolated p–type pairs. In Eq. (6) and (8), the expected situation obtains. These findings indicate that ln[per(**A**)]/ln K may be the quantitative structure–property relationship variable of choice for some properties.

Table 2. Cross–Validation Correlation Coefficient in a Leave–*n*–Out Procedure for Fullerenes.

n	ln[per(**A**)]/ln K *vs.* p, q, r	ln[per(**A**)]/ln K *vs.* q, r	ln[per(**A**)]/ln K *vs.* p, q, r, q/p, r/p	ln K *vs.* p, q, r	ln K *vs.* p, q	ln[per(**A**)] *vs.* p, q, r	ln[per(**A**)] *vs.* p, q
1	0.551	0.623	0.818	0.935	0.943	0.930	0.932
2	0.550	0.623	0.819	0.935	0.943	0.930	0.932
3	0.548	0.622	0.820	0.936	0.944	0.930	0.932
4	0.546	0.622	0.821	0.937	0.944	0.930	0.932
5	0.544	0.622	0.822	0.938	0.944	0.929	0.932
6	0.542	0.621	0.824	0.939	0.945	0.929	0.932
7	0.540	0.621	0.824	0.939	0.945	0.929	0.932
8	0.538	0.620	0.825	0.940	0.946	0.928	0.932
9	0.536	0.619	0.826	0.941	0.946	0.928	0.932
10	0.534	0.619	0.827	0.942	0.946	0.927	0.932

Table 2. (Continued)

n	ln[per(**A**)]/ln *K vs*. *p*, *q*, *r*, *q*/*p*, *r*/*p* (means)	ln *K vs*. *p*, *q*, *r* (means)	ln *K vs*. *p*, *q* (means)	ln[per(**A**)]/ln *K vs*. *p*, *q*, *r*, *q*/*p*, *r*/*p* (27 points)
1	0.821	0.974	0.975	0.899
2	0.823	0.974	0.975	0.899
3	0.825	0.973	0.975	0.898
4	0.827	0.973	0.974	0.897
5	0.829	0.973	0.974	0.896
6	0.830	0.972	0.974	0.895
7	0.832	0.972	0.974	0.895
8	0.833	0.972	0.974	0.894
9	0.835	0.971	0.974	0.893
10	0.836	0.971	0.974	0.892

The correlation coefficient found between cross–validated representatives and the property values R_{cv} has been calculated with the leave–*n*–out procedure [32]. Leave–*n*–out furnishes a new method for selecting the best set of descriptors according to the criterion of maximization of the value of R_{cv}. The R_{cv} calculated for fullerenes are given in Table 2 for $1 \le n \le 10$. In general, R_{cv} decreases with *n*. However, for ln[per(**A**)]/ln *K vs*. {*p*, *q*, *r*, *q*/*p*, *r*/*p*} and for both ln *K* methods, R_{cv} increases with *n*. In general, the effect is corrected when the set of points is substituted by the set of their means (*cf*. the three columns labelled *means* in Table 2). Nevertheless, for the method ln[per(**A**)]/ln *K vs*. {*p*, *q*, *r*, *q*/*p*, *r*/*p*}, R_{cv} increases again with *n*. The effect is finally corrected when the first point in Table 1 is eliminated (*cf*. the last column of Table 2). In particular, for ln[per(**A**)]/ln *K*, the set of descriptors {*p*, *q*, *r*, *q*/*p*, *r*/*p*} gives the greatest R_{cv} for the whole range of *n* given in Table 2. However, for both ln *K* and ln[per(**A**)], {*p*, *q*} gives the greatest R_{cv}. The corresponding interpretation is that the set {*p*, *q*, *r*, *q*/*p*, *r*/*p*} is more predictive than {*q*, *r*} or {*p*, *q*, *r*} for modelling ln[per(**A**)]/ln *K*, and that {*p*, *q*} is more predictive than {*p*, *q*, *r*} for representing both ln *K* and ln[per(**A**)]. The upper triangle of the 3x3 symmetrical correlation matrix **R** calculated for the structural parameters *p*, *q* and *r* results:

$$\mathbf{R} = \begin{pmatrix} 1.000 & 0.836 & 0.864 \\ & 1.000 & 0.691 \\ & & 1.000 \end{pmatrix}$$

High correlation is observed for the pairs *p*–*r* and *p*–*q*. The correlation increases in the order $R_{qr} << R_{pq} < R_{pr}$. The upper triangle of the 5×5 correlation matrix **R** calculated for the structural parameters *p*, *q*, *r*, *q*/*p* and *r*/*p* results:

$$\mathbf{R} = \begin{pmatrix} 1.000 & 0.929 & 0.857 & 0.805 & 0.542 \\ & 1.000 & 0.635 & 0.934 & 0.225 \\ & & 1.000 & 0.457 & 0.875 \\ & & & 1.000 & 0.029 \\ & & & & 1.000 \end{pmatrix}$$

High correlation is obtained for the pairs *q*–*q*/*p*, *p*–*q*, *r*–*r*/*p* and *p*–*r*. Notice that the correlation

between the derived q/p and r/p parameters (0.029) is 20 times smaller than that between the primary q and r parameters (0.635). The correlation increases in the order $R_{qr} << R_{pr} < R_{pq}$. The difference with $\mathbf{R}_{3\times3}$ is due to the smaller number of points in the calculation of $\mathbf{R}_{5\times5}$, because the parameters q/p and r/p are undefined for the last three fullerenes in Table 1. From both HCAs, the radial tree [33] is built for the parameters p, q, r, q/p and r/p of the fullerenes (*cf*. Figure 1).

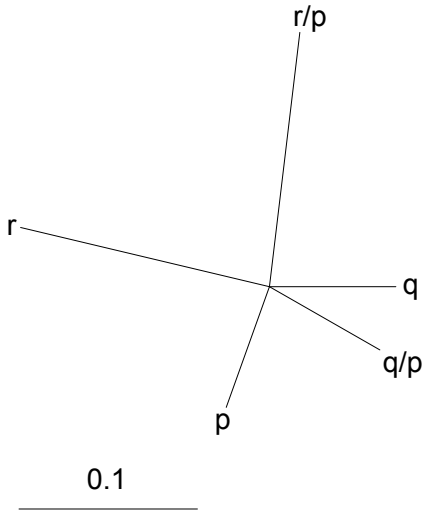

Figure 1. Radial tree graph for the parameters p, q, r, q/p and r/p of fullerenes.

SplitsTree is an interactive program for analysing and visualizing CA data [34]. Based on the method of split decomposition, it takes as input a distance matrix or a set of CA data and produces as output a graph that represents the relationships between the taxa. For ideal data, this graph is a tree, whereas less ideal data will give rise to a tree–like network that can be interpreted as possible evidence for different and conflicting data. Further, as split decomposition does not attempt to force data onto a tree, it can provide a good indication of how tree–like given data are. The splits graph for the structural parameters p, q, r, q/p and r/p of the fullerenes is displayed in Figure 2. The splits graph in Figure 2 reveals that a conflicting relationship exists between p, and parameters q–q/p and r–r/p. This is due to the interdependences p–q and p–r. Therefore, the splits graph indicates a spurious relationship resulting from base composition effects. The portion r–p–q of the splits graph is in qualitative agreement with a previous study of the set $\{p, q, r\}$, which also indicated a spurious relationship between p, q and r resulting from base composition effects.

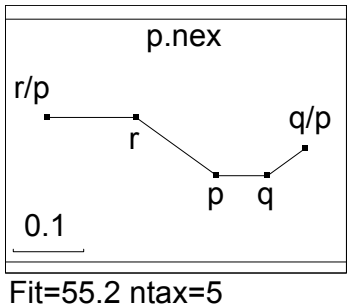

Figure 2. The splits graph for the parameters p, q, r, q/p and r/p of fullerenes.

Table 3. Importance of the Principal Component Analysis Factors

Factor	Eigenvalue	Percentage of variance	Cumulative percentage of variance
F_1	3.59677680	71.94	71.94
F_2	1.31894604	26.37	98.31
F_3	0.05633613	1.13	99.44
F_4	0.01775178	0.36	99.80
F_5	0.01018925	0.20	100.00

The importance of PCA factors F_1–F_5 for the structural parameters of the fullerenes is collected in Table 3. In particular, the use of only the first factor F_1 explains 72% of the variance and gives a relative error of 28%. Moreover, the combined use of the first two factors, F_1 and F_2, explains 98% of the variance, reducing the relative error to 2%. Furthermore, the use of the first three factors, F_1–F_3, explains 99.4% of the variance, reducing the relative error to only 0.6%.

Table 4. Principal Component Analysis Loadings for the Structural Parameters [a]

	PCA factor loadings				
Property	F_1	F_2	F_3	F_4	F_5
p	0.523	–0.045	0.389	–0.141	**0.744**
q	0.480	–0.342	0.439	–0.240	–0.634
r	0.470	0.385	–0.010	**0.779**	–0.154
q/p	0.421	–0.501	**–0.753**	–0.002	0.068
r/p	0.314	0.694	–0.297	–0.561	–0.130

[a] Loadings greater than 0.7 are boldfaced

Table 5. Profile of the Principal Component Analysis Factors.[a]

Factor	Percentage of p	Percentage of q	Percentage of r	Percentage of q/p	Percentage of r/p
F_1	27.33	23.07	22.06	17.71	9.83
F_2	0.20	11.69	14.83	25.09	48.19
F_3	15.16	19.30	0.01	**56.73**	8.80
F_4	2.00	5.76	**60.75**	0.00	31.49
F_5	**55.31**	40.18	2.36	0.46	1.69

[a] Percentages greater than 50% are boldfaced.

The factor loadings of PCA are shown in Table 4.The profile of PCA factors F_1–F_5 for the structural parameters of the fullerenes is resumed in Table 5. In particular, for both F_1 and F_5 factors, variable *p* has the greatest weight in the profile. However, factor F_1 cannot be reduced to three variables (*p*, *q* and *r*) without making a relative error of 28% (the sum of both *q*/*p* and *r*/*p* percentages). On the other hand, for factor F_2 the most important variable is *r*/*p*. For F_3, the variable with greatest weight is *q*/*p*. For F_4, the variable with greatest weight is *r*. In some way, factors F_1 and F_5 could be considered as linear combinations of *p*, *q* and *r* (with relative errors of 28% and 2%, respectively). Nevertheless, factor F_2 can be expressed as a linear combination of *r*, *q*/*p* and *r*/*p* with a relative error of 12%.

PCA F_2 *vs*. F_1 plot for the fullerenes is illustrated in Figure 3. The fullerenes with the same set of *p*, *q*, *r*, *q*/*p* and *r*/*p* values in Table 1, belonging to classes 1, 3, 4 and 5, appear superposed in Figure 3. Five classes of fullerenes are clearly distinguished: class 1 with 7 members (below the bisector,

$F_1 >> F_2$, middle right of Figure 3), class 2 with 4 members (under the bisector, $F_1 > F_2 > 0$, top right of Figure 3), class 3 with 8 members (over the bisector, $F_1 < F_2$, top of Figure 3), class 4 with 5 members (above the bisector, $F_1 << F_2$, top left of Figure 3) and class 5 with 4 members (under the bisector, $0 > F_1 > F_2$, bottom left of Figure 3). In general, those fullerenes with the same number of atoms belong to the same class. The exceptions are the isomers of the fullerenes C_{28}, C_{30}, C_{32}, C_{34}, C_{36}, C_{38} and C_{40}, which fit in two or three classes. However, no fullerene has isomers going to four or five classes.

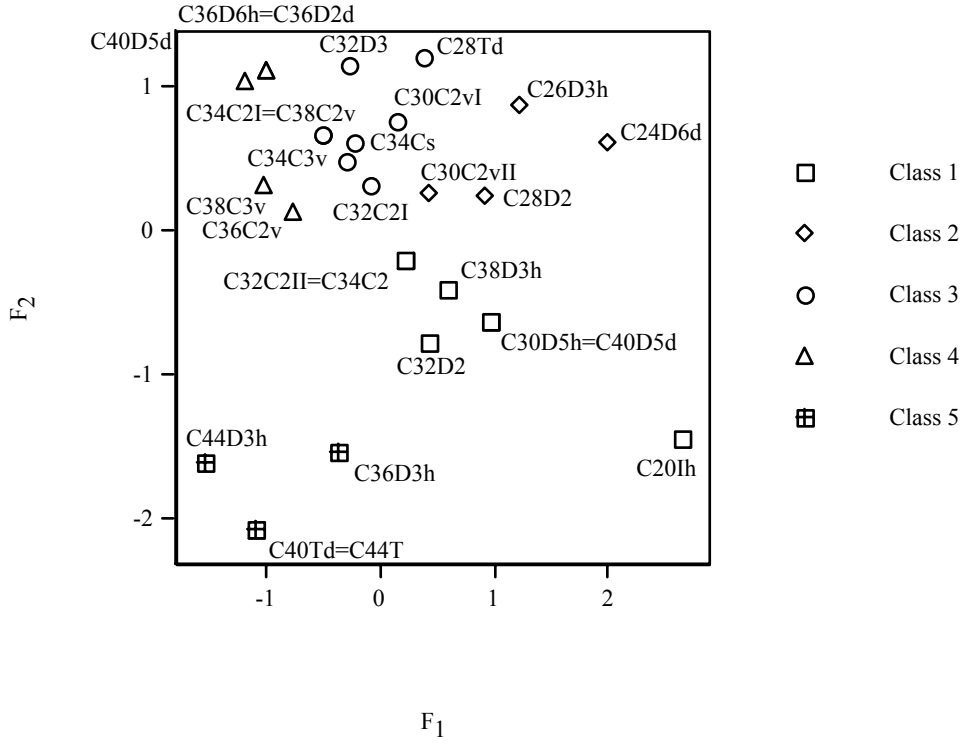

Figure 3. PCA F_2 *vs*. F_1 plot for the fullerenes.

With the purpose of classifying the C_{60} (I_h), C_{70} (D_{5h}) and C_{82} (C_s) fullerenes, PCA was repeated with the set {*p*, *q*, *r*}. PCA F_2 *vs*. F_1 plot grouped these fullerenes in class 5, close to C_{44} (D_{3h}). Therefore, the final consideration is the inclusion of C_{60}, C_{70} and C_{82} in class 5. The patterns in Figure 3 are rather similar to those in Figure 6 of the previous study [15]. The number of reported classes is different by the following reasons. (*a*) The previous study uses all the 31 fullerenes in Table 1. (*b*) It is limited to only 3 counts (*p*, *q* and *r*). However, as it can be seen from the profiles in the present Table 5, the new F_1 depends 18% on *q*/*p* and 10% on *r*/*p*, and the new F_2 depends 25% on *q*/*p* and 48% on *r*/*p*. (*c*) In the previous study, the distance is non–metric. Triangle inequalities are not satisfied. Worst violating triplets are 29–2–20, 30–2–20 and 31–2–20. This limitation also affects the dendrogram and radial tree. (*d*) The number of classes has been maximized because merging classes always gives a loss of information. Using only PCA, alternate classifications could

be proposed. For instance, classes 2 and 3 are somewhat close and could be candidates to merge. (*e*). The reported classification has been selected because it maximizes the number of classes and it is compatible with the five classes suggested by the dendrogram (*cf*. Figure 4) and with the five classes suggested by the radial tree (Figure 5). Unfortunately, in the previous study the dendrogram and radial tree provide no privileged cutting point. They are compatible with any number of classes. Therefore, they cannot be used to support any particular number of classes, which had to be determined exclusively by PCA.

Instead of N fullerenes (points) in the $\Re^P$ space of P parameters, let us consider P structural parameters in the $\Re^N$ space of N fullerenes. A table with P rows and N columns has been built and the similarity of the fullerenes is compared. The dendrogram for the fullerenes matching to the structural parameters p, q, r, q/p and r/p is shown in Figure 4. The tree provides a binary taxonomy of the fullerenes in Table 1, which separates first the 7 fullerenes in class 1 [from C_{20} (I_h) to C_{34} (C_2) II, Figure 4 top], then the 4 fullerenes in class 2 [from C_{24} (D_{6d}) to C_{30} (C_{2v}) II, Figure 4 top middle], the 8 fullerenes in class 3 [from C_{28} (T_d) to C_{38} (C_{2v}), Figure 4 middle], the 5 fullerenes in class 4 [from C_{40} (D_{5d}) I to C_{38} (C_{3v}), Figure 4 bottom middle] and the 4 fullerenes in class 5 [from C_{36} (D_{3h}) to C_{44} (T), Figure 4 bottom]. The classes correspond to those obtained by PCA (Figure 3). With the purpose of classifying the last three fullerenes in Table 1, the dendrogram was repeated for the set $\{p, q, r\}$. The result was the inclusion of C_{60}, C_{70} and C_{82} in a new branch connected to C_{44} (D_{3h}).

The radial tree for the fullerenes relating to the parameters p, q, r, q/p and r/p is displayed in Figure 5. It separates first the 7 fullerenes in class 1 [C_{20} (I_h)–C_{34} (C_2) II, Figure 5 bottom right], then the 4 fullerenes in class 2 [C_{24} (D_{6d})–C_{30} (C_{2v}) II, Figure 5 bottom], the 8 fullerenes in class 3 [C_{28} (T_d)–C_{38} (C_{2v}), Figure 5 left], the 5 fullerenes in class 4 [C_{40} (D_{5d}) I–C_{38} (C_{3v}), Figure 5 top] and the 4 fullerenes in class 5 [C_{36} (D_{3h})–C_{44} (T), Figure 5 right]. The classes correspond to those obtained by PCA (Figure 3) and dendrogram (Figure 4). With the purpose of classifying the last three fullerenes in Table 1, the radial tree was repeated for the set $\{p, q, r\}$. The result was the inclusion of C_{60}, C_{70} and C_{82} in a new branch connected to C_{44} (D_{3h}).

The present report allows the classification of fullerenes and fullerene isomers. The application of the work to different fullerene or hydrocarbon datasets would give a classification of these molecules. For instance, the fullerenes C_{60} (I_h), C_{70} (D_{5h}) and C_{82} (C_s) classified together. The application of the method to C_{80} (I_h), C_{120} (I_h), C_{140} (I_h), C_{180} (I_h) and C_{240} (I_h) all with $p = q = r = 0$ is expected to group these five fullerenes with C_{60}, C_{70} and C_{82}. On the other hand, the correlation of structural parameters against properties gives a classification of the properties. This is an added value of the method. The elimination of close properties would diminish the risk of co–linearity in the fits.

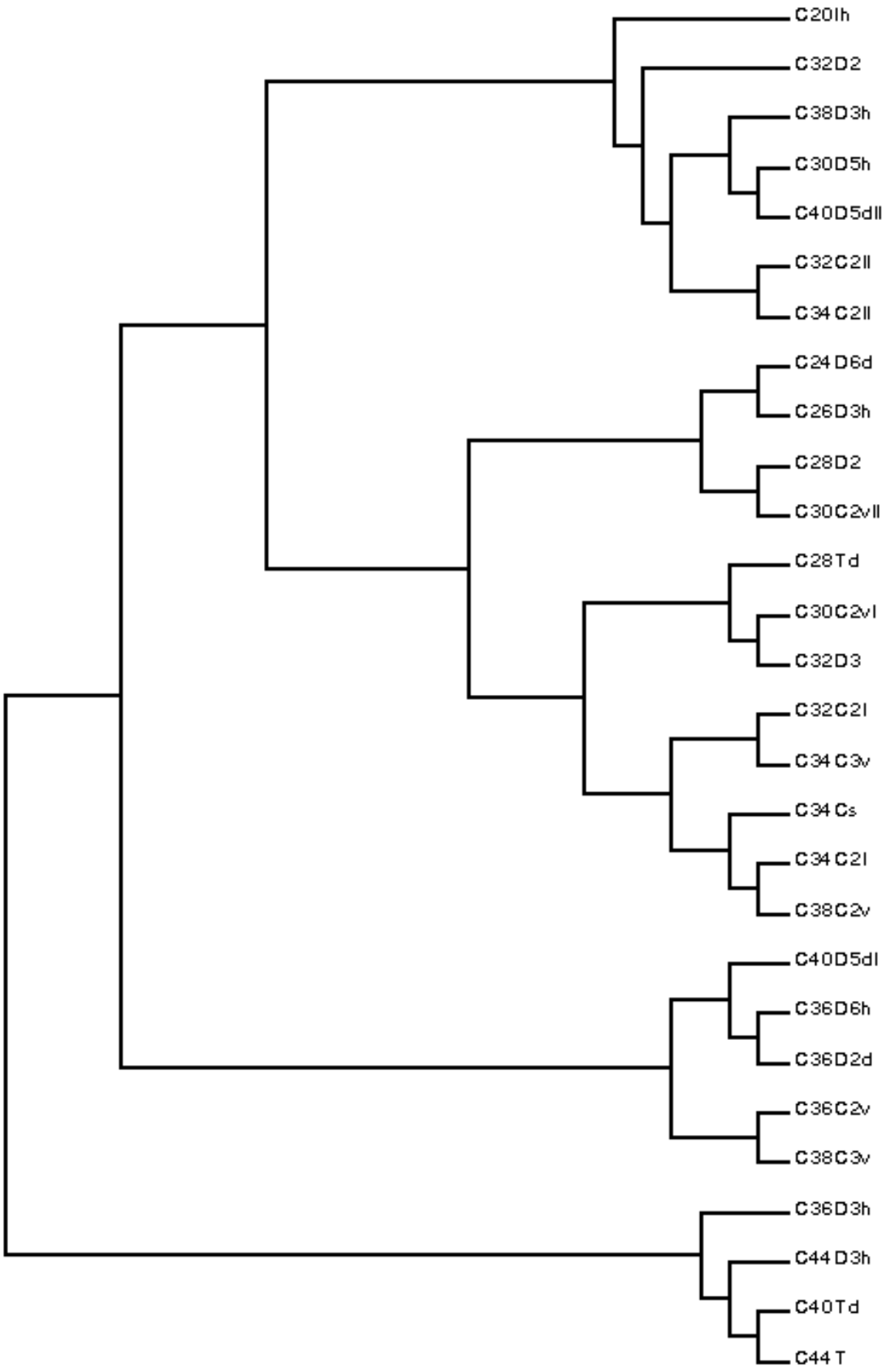

Figure 4. Dendrogram for the fullerenes.

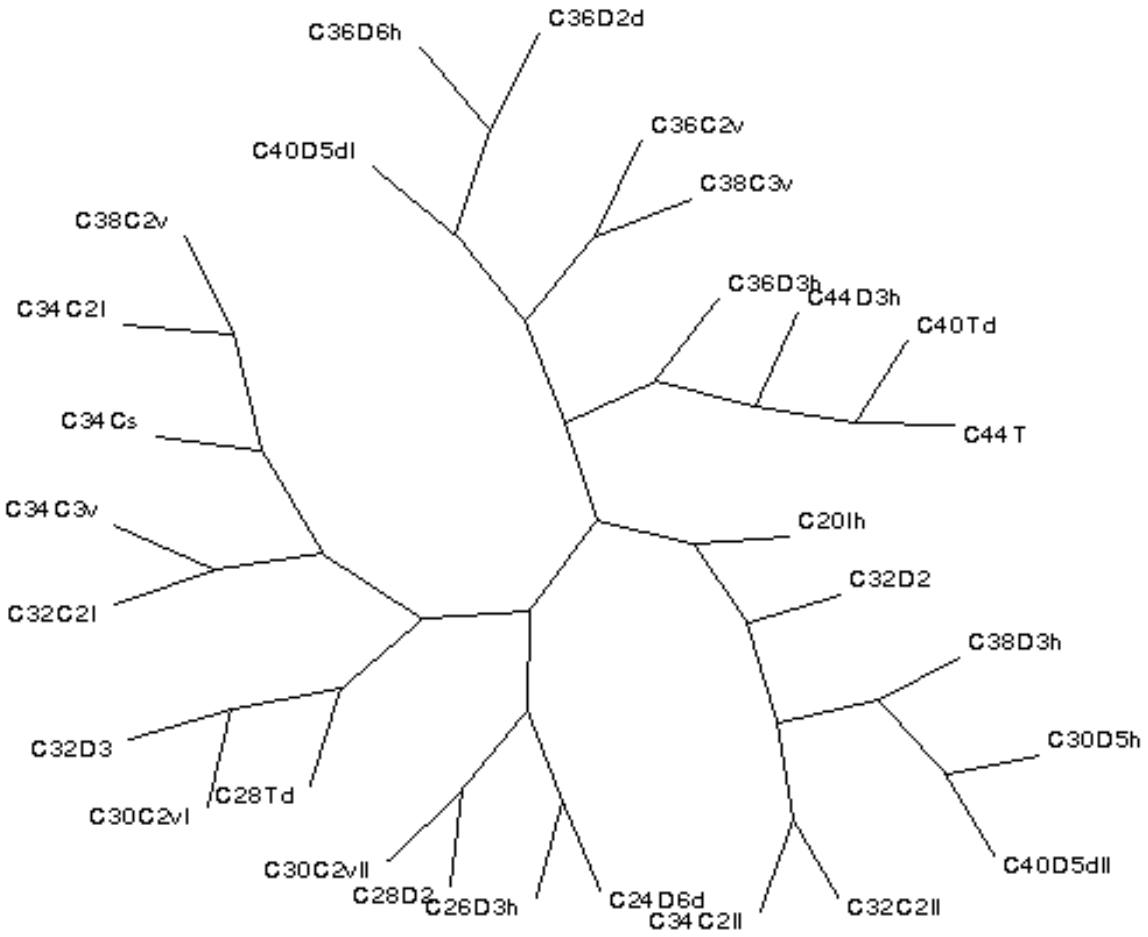

Figure 5. Radial tree graph for the fullerenes.

Cluster analyses are correct provided the dataset is complete. In the present case, a small subset of the fullerene structures has been examined for larger fullerenes. Therefore, the comparison of distances between C_{82} (C_s) and other fullerenes is limited. However, other points are part of wider subsets and no limitation is expected, as for the three C_{34} isomers in Class 3. Any clustering based on p, q, r or combinations thereof will necessarily group all isolated–pentagon isomers together. Similarly, any regression equation will yield the same value of the dependent value for all isolated–pentagon isomers. For instance, C_{60} (I_h), C_{70} (D_{5h}), C_{80} (I_h), C_{82} (C_s), C_{120} (I_h), C_{140} (I_h), C_{180} (I_h) and C_{240} (I_h) all with $p = q = r = 0$, which are represented by one only point in the space of parameters, must be used as one only point in the fits.

4 CONCLUSIONS

From the preceding results the following conclusions can be drawn.

1. Linear and non–linear correlation models have been obtained for ln[per(**A**)]/ln K, ln K and ln[per(**A**)] of fullerenes as functions of structural parameters involving the presence of contiguous

pentagons. The non–linear regression equation for ln[per(**A**)]/ln *K* has been improved. The variance of the fit has decreased 68%. It has also diminished the risk of co–linearity in the fit. The cross–validation leave–*n*–out procedure shows that the most predictive sets of descriptors according to the criteria of maximization of R_{cv} are {*p*, *q*, *r*, *q*/*p*, *r*/*p*} for ln[per(**A**)]/ln *K*, and {*p*, *q*} for both ln *K* and ln[per(**A**)]. Leave–*n*–out has been successfully used to identify outliers.

2. CA shows greater similarity for the parameters *p*–*q* than when comparing with the count *r*. Split decomposition indicates a spurious relationship resulting from base composition effects.

3. PCA provides five orthogonal factors F_1–F_5. The use of F_1 gives a relative error of 28%. The use of F_1 and F_2 decreases the relative error to 2%. The fullerenes have been grouped in five classes. Some fullerenes with different numbers of atoms belong to the same class. However, some fullerene isomers are members of different classes. Nevertheless, no fullerene belongs to four classes.

4. The similarity between fullerenes has been compared with CA of these molecules. CA is in agreement with PCA classification.

Acknowledgment

I wish to thank Dr. E. Besalú for providing me several versions of his full linear leave–many–out program prior to publication. The author acknowledges financial support of the Spanish MCT (Plan Nacional I+D+I, Project No. BQU2001–2935–C02–01).

5 REFERENCES

[1] M. A. Kraaiveld and J. Mao, A Nonlinear Projection Method Based on Kohonen's Topology Preserving Maps, *IEEE Trans*. *Neural Networks* **1995**, *6*, 548–559.
[2] G. Biswas, A. K. Jain and R. C. Dubes, Evaluation of Projection Algorithms, *IEEE Trans*. *Pattern Anal*. *Machine Intell*. **1981**, *PAMI–3*, 701–708.
[3] J. W. Sammon, Jr., A Nonlinear Mapping for Data Structure Analysis, *IEEE Trans*. *Comput*. **1969**, *C–18*, 401–409.
[4] B. R. Kowalski and C. F. Bender, Pattern Recognition. A Powerful Approach to Interpreting Chemical Data, *J. Am*. *Chem*. *Soc*. **1972**, *94*, 5632–5639.
[5] D. Domine, J. Devillers, M. Chastrette and W. Karcher, Non–linear Mapping for Structure–Activity and Structure–Property Modelling, *J*. *Chemom*. **1993**, *7*, 227–242.
[6] B. Bienfait and J. Gasteiger, Checking the Projection Display of Multivariate Data with Colored Graphs, *J*. *Mol*. *Graphics Mod*. **1997**, *15*, 203–215.
[7] H. Hotelling, Analysis of a Complex of Statistical Variables into Principal Components, *J*. *Educ*. *Psychol*. **1933**, *24*, 417–441.
[8] H. Hotelling, Analysis of a Complex of Statistical Variables into Principal Components, *J*. *Educ*. *Psychol*. **1933**, *24*, 498–520.
[9] S. Wold, K. Esbensen and P. Geladi, Principal Component Analysis, *Chemom*. *Intell*. *Lab*. *Syst*. **1987**, *2*, 37–52.
[10] R. D. Brown and Y. C. Martin, Use of Structure–Activity Data To Compare Structure–Based Clustering Methods and Descriptors for Use in Compound Selection, *J*. *Chem*. *Inf*. *Comput*. *Sci*. **1996**, *36*, 572–584.
[11] R. D. Brown and Y. C. Martin, The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand–Receptor Binding, *J*. *Chem*. *Inf*. *Comput*. *Sci*. **1997**, *37*, 1–9.
[12] H. Matter, Selecting Optimally Diverse Compounds from Structural Databases: A Validation Study of Two–dimensional and Three–dimensional Molecular Descriptors, *J*. *Med*. *Chem*. **1997**, *40*, 1219–1229.
[13] F. Torrens, Computing the Kekulé Structure Count for Alternant Hydrocarbons, *Int*. *J*. *Quantum Chem*. **2002**, *88*, 392–397.

[14] F. Torrens, Computing the Permanent of the Adjacency Matrix for Fullerenes, *Internet Electron. J. Mol. Des.* **2002**, *1*, 351–359, http://www.biochempress.com.
[15] F. Torrens, Principal Component Analysis of Structural Parameters for Fullerenes, *Internet Electron. J. Mol. Des.* **2003**, *2*, 96–111, http://www.biochempress.com.
[16] F. Torrens, New Structural Parameters of Fullerenes for Principal Component Analysis, *Theor. Chem. Acc.*, in press.
[17] R. C. Tryon, *J. Chronic Dis.* **1939**, *20*, 511–524.
[18] R. A. Jarvis and E. A. Patrick, Clustering Using a Similarity Measure Based on Shared Nearest Neighbors, *IEEE Trans. Comput.* **1973**, *C22*, 1025–1034.
[19] M. J. McGregor and P. V. Pallai, Clustering of Large Databases of Compounds: Using the MDL "Keys" as Structural Descriptors, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443–448.
[20] T. N. Doman, J. M. Cibulskis, M. J. Cibulskis, P. D. McCray and D. P. Spangler, Algorithm 5: A Technique for Fuzzy Similarity Clustering of Chemical Inventories, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1195–1204.
[21] D. B. Turner, S. M. Tyrrell and P. Willett, Rapid Quantification of Molecular Diversity for Selective Database Acquisition, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 18–22.
[22] C. H. Reynolds, R. Druker and L. B. Pfahler, Lead Discovering Using Stochastic Cluster Analysis (SCA): A New Method for Clustering Structurally Similar Compounds, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 305–312.
[23] *Integrated Mathematical Statistical Library* (*IMSL*), IMSL, Houston, 1989.
[24] P. Constans and J. D. Hirst, Non–Parametric Regressors Applied to Quantitative Structure–Activity Relationships, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 452–459.
[25] X. Liu, D. J. Klein, T. G. Schmalz and W. A. Seitz, Generation of Carbon–Cage Polyhedra, *J. Comput. Chem.* **1991**, *12*, 1252–1259.
[26] G. G. Cash, Permanents of Adjacency Matrices of Fullerenes, *Polycycl. Arom. Compounds* **1997**, *12*, 61–69.
[27] D. J. Klein and X. Liu, Theorems for Carbon Cages, *J. Math. Chem.* **1992**, *11*, 199–205.
[28] D. J. Klein and X. Liu, Many–Body Conjugated–Circuit Computations, *J. Comput. Chem.* **1991**, *12*, 1260–1264.
[29] D. J. Klein, H. Zhu, R. Valenti and M. A. Garcia–Bach, Many–Body Valence–Bond Theory, *Int. J. Quantum Chem.* **1997**, *65*, 421–438.
[30] H. Zhu, A. T. Balaban, D. J. Klein and T. P. Živković, Conjugated–Circuit Computations on Two–Dimensional Carbon Networks, *J. Chem. Phys.* **1994**, *101*, 5281–5292.
[31] R. R. Hocking, The Analysis and Selection of Variables in Linear Regression, *Biometrics* **1976**, *32*, 1–49.
[32] E. Besalú, Fast Computation of Cross–Validated Properties in Full Linear Leave–Many–Out Procedures, *J. Math. Chem.* **2001**, *29*, 191–203.
[33] R. D. M. Page, Program TreeView, Universiy of Glasgow, 2000.
[34] D. H. Huson, SplitsTree: Analizing and Visualizing Evolutionary Data, *Bioinformatics* **1998**, *14*, 68–73.

Biographies

Francisco Torrens is lecturer of physical chemistry at the Universitat de València. After obtaining a Ph.D. degree in molecular associations in azines and macrocycles from the Universitat de València, Dr. Torrens undertook postdoctoral research with Professor Rivail at the Université de Nancy I. More recently, Dr. Torrens has collaborated on projects with Professor Tomás–Vert. Major research projects include characterization of the electronic structure of electrically conductive organic materials, theoretical study of new electrically conductive organic materials, modellization of proteins, electronic correlation, development and applications of high–precision mono and multireferential electronic correlation methods, and development and application of high–precision quantum methods. Scientific accomplishments include the first implementation in a computer at the Universitat de Valencia of a program for the elucidation of crystallographic structures, and the construction of the first computational–chemistry program adapted to a vector–facility supercomputer at a Spanish university.