**Inter*net* Electronic Journal of Molecular Design**

February 2003, Volume 2, Number 2, Pages 96–111

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Haruo Hosoya on the occasion of the 65th birthday
Part 6

Guest Editor: Jun–ichi Aihara

# Principal Component Analysis of Structural Parameters for Fullerenes

Francisco Torrens

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

Received: July 18, 2002; Revised: October 11, 2002; Accepted: November 19, 2002; Published: February 28, 2003

**Citation of the article:**
F. Torrens, Principal Component Analysis of Structural Parameters for Fullerenes, *Internet Electron. J. Mol. Des.* **2003**, *2*, 96–111, http://www.biochempress.com.

**Inter*net* Electronic Journal of Molecular Design**

**BioChem Press**
http://www.biochempress.com

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

Francisco Torrens*

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

Received: July 18, 2002; Revised: October 11, 2002; Accepted: November 19, 2002; Published: February 28, 2003

*Internet Electron. J. Mol. Des*. **2003**, *2* (2), 96–111

**Abstract**

**Motivation.** Novel carbon allotropes with finite molecular structure, including spherical fullerenes, are nowadays currently produced and investigated. The Kekulé structures count and permanent of the adjacency matrix of these molecules are related to structural parameters involving the presence of contiguous pentagons.

**Method.** Both single– and complete–linkage cluster analyses of the structural parameters allow classifying these parameters. Principal component analysis (PCA) 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 $q$ parameter (the number of edges common to two pentagons). PCA clearly distinguishes three classes of fullerenes. The cluster analysis of fullerenes is in agreement with PCA classification.

**Conclusions.** Cluster analysis shows the greatest similarity for the $p$ and $r$ (the number of vertices common to three pentagons) parameters. Split decomposition indicates a spurious relationship resulting from base composition effects. PCA provides three orthogonal factors $F_1$–$F_3$. The use of only $F_1$ gives an error of 13%. The use $F_1$ and $F_2$ decreases the error to 3%. PCA groups the fullerenes in three classes. Some fullerenes with different numbers of atoms belong to the same class, while some fullerene isomers are members of different classes.

**Availability.** The software programs are available on request from the author (Francisco.Torrens@uv.es) and are free for academics.

**Keywords.** Cluster analysis; dendrogram; split decomposition; principal component analysis; 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 in a plane. This allows the

[#] Dedicated to Professor Haruo Hosoya 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.

visual inspection of the structure of the dataset to identify clusters and particular objects, *i.e*., to perform an exploratory data analysis.

When dealing with vectors whose dimensions are larger than two, it is not possible to represent them graphically in a plane. One way to overcome this problem is to transform the *N*–dimensional vectors into 2D. Many projection methods have been 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 two–dimensional 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 components, 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.

To illustrate the usefulness of this method, a projection method and a dataset of molecules are studied. The dataset deals with a series of 31 fullerenes represented by three structural parameters. For this example, PCA projection method is applied. On the other hand, a method is described for clustering data. The relative efficiency of clustering algorithms and similarity descriptors has been the subject of several recent articles [10–12].

In a previous paper, the calculation of the Kekulé structures count and permanent of adjacency matrices [13] was applied to fullerenes with different structural parameters involving the presence of contiguous pentagons [14]. In this work, PCA of the structural parameters has been carried out. The aim of this paper is to analyse the interdependence between the structural parameters, to classify them, and to classify the fullerenes. Section 2 presents the computational method. Section 3 discusses the calculation results for fullerenes. Section 4 summarizes the conclusions.

# 2 COMPUTATIONAL METHOD

## 2.1 Principal Component Analysis

Data may be viewed as $N$ (number of points) vectors in $P$ (number of calculated parameters) dimensions. The data for each set can be represented by a matrix $\mathbf{X}$ which has $N$ rows and $P$ columns. Each pattern is therefore represented by a point in $\Re^P$, where $\Re$ is the field of real numbers. If each pattern $s$ was represented in $\Re^2$, then one could plot and investigate the extent of relationship between individual parameters. In $\Re^P$ such a simple analysis is not possible. However, if many of the data are highly intercorrelated, the points in $\Re^P$ can likely be represented by a subspace of fewer dimensions. The method of PCA or the Karhunen–Loeve transformation is a standard method for reduction of dimensionality. The first PC, $F_1$, is the line that comes closest to the points in the sense of minimizing the sum of the squared Euclidean distances from the points to the line. The second PC, $F_2$, is given by projections onto the basis vector orthogonal to $F_1$. For points in $\Re^P$, the first $r$ PCs give the subspace that comes closest to approximating the $N$ points. $F_1$ is the first axis of the points. Successive axes are major directions orthogonal to previous axes. PCs are the closest approximating hyperplane, and because they are calculated from eigenvectors of a $P \times P$ matrix, the computations are relatively accessible. However, there are important scaling choices, because PCs are scale dependent. To control this dependence, the most commonly used convention is to rescale the variables so that each variable have a mean of zero and a standard deviation of one. The co–variance matrix for these rescaled variables is the correlation matrix.

For each one of the $N$ fullerenes, one has $P$ values for the structural parameters. Therefore, one can build a table with $N$ rows and $P$ columns. Let us consider an $\Re^P$ space with an orthonormal basis set. Each axis of this basis set is a direction of one of the $P$ variables. For each fullerene in the table, a point is associated in $\Re^P$ where the coordinates on the $P$ axes are the values of its $P$ parameters. For the $N$ data in the table there are associated $N$ points in $\Re^P$, making up a cloud in this space. The objective of the analysis is to represent this cloud in a space with dimension lower than $P$, with the minimal loss of information. For accomplishing this, PCs of the cloud and the correlations of these axes with the $P$ variables are determined. A new system of uncorrelated factors is thus obtained. Each variable can be expressed by means of these $P$ factors. Anyway, certain factors contribute stronger than others to the variation of the variables. In general, the importance of a factor is represented by its percentage of variance. Therefore, by projecting the cloud over the plane containing the two most important factors, $F_1$ and $F_2$, one obtains a representation that contains the greatest part of the information. To call a third factor can be necessary if the two firsts are insufficient.

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 compared with the mean of this

*j*–th variable. In fact, the transformed value results

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

where $\sigma_j$ is the standard deviation of the *j*–th variable. The factorial analysis method, 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, the factors have the form:

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

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 $F_1$ and $F_2$ factors. The profile of a factor $F_i$ is the vector of the squared $C_{ik}$ coefficients $\left(C_{i1}^2, C_{i2}^2, ..., C_{iP}^2\right)$. Each $C_{ik}^2$ represents the weight of variable *k* in factor $F_i$. It gives the fraction of each variable in factor $F_i$.

## 2.2 Cluster Analysis

One approach to the diversity problem is to cluster a structural database or virtual library based on some kind of structural criteria. Standard approaches for clustering can be broken into two broad categories: hierarchical and non–hierarchical. Hierarchical approaches can be further categorized as agglomerative or divisive. In these approaches, 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 a non–hierarchical approach, a nearest–neighbour list is created and used to assemble members into related clusters. An example of this is the Jarvis–Patrick clustering algorithm, which has been widely used to cluster structural databases [15].

There are many reasons why one might want to cluster a database of molecular structures [16]. Two of the most practical reasons are to identify representative compounds from a structural database or virtual compound library for screening or synthesis [17]. In addition, one can be interested in using a clustering algorithm to validate similarity methods and descriptors. If it is possible to cluster databases where one has some biological data in a way that groups compounds with like activity, that will serve to validate the methods used to assign similarity. Further, 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 [18].

Three objectives must be in mind when designing a clustering algorithm [19]. First, 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 clustering algorithms that create a defined number of clusters without regard to the fact that this sometimes requires grouping very unlike structures together. Second, 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. Third, of course, any method has to be computationally tenable for very large structural databases. Speed is one of the most significant problems with hierarchical methods, but even the more efficient non–hierarchical approaches scale formally as $N^2$.

Using the IMSL [20] subroutine CLINK, a program has been written to carry out the cluster analysis from a correlation or similarity matrix. The algorithm performs hierarchical cluster analysis based upon a distance matrix or upon a similarity matrix. Hierarchical clustering proceeds in four steps. Initially, each pattern 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.

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 clustering, 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 and complete methods 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 (cf.* Figure 1). In the single linkage method, the distance from *Z* to *C* is the minimum of the distances *(A* to *C*, *B* to *C)*. In the complete linkage method, the distance from *Z* to *C* is the maximum of the distances *(A* to *C*, *B* to *C)*. In general, single linkage will yield long thin clusters, while complete linkage will yield clusters that are more spherical.

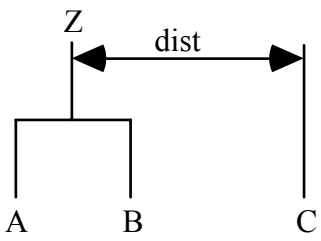
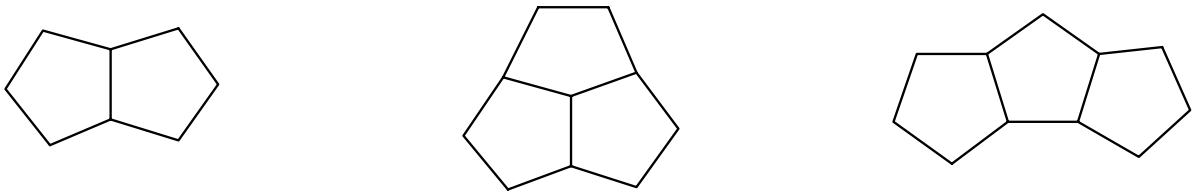



**Figure 1**. Distance between clusters *Z* and *C*.

# 3 RESULTS AND DISCUSSION

The structural features involving adjacent pentagons are encoded by the *p*, *q* and *r* parameters as illustrated in Figure 2. The *p* and *q* parameters enumerate, respectively, the number of edges common to two pentagons and the number of vertices common to three pentagons [21]. The *r* parameter enumerates the number of pairs of non–adjacent pentagon edges shared with two other pentagons [22]. Thus, *q* and *r* complement each other by counting both possible arrangements of three contiguous pentagons.



p q r

**Figure 2.** Substructures that contribute to the *p*, *q* and *r* counts.

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

| Fullerene | ***K*** | per(**A**) | ln[per(**A**)]/ln*K* | *p* | *q* | *r* |
|---|---|---|---|---|---|---|
| $C_{20}$ ($I_h$) | 36 | 1392 | 2.0199 | 30 | 20 | 30 |
| $C_{24}$ ($D_{6d}$) | 54 | 4692 | 2.1192 | 24 | 12 | 36 |
| $C_{26}$ ($D_{3h}$) | 63 | 8553 | 2.1853 | 21 | 8 | 30 |
| $C_{28}$ ($T_d$) | 75 | 15705 | 2.2378 | 18 | 4 | 24 |
| $C_{28}$ ($D_2$) | 90 | 16196 | 2.1540 | 20 | 8 | 24 |
| $C_{30}$ ($C_{2v}$) I | 107 | 29621 | 2.2034 | 17 | 4 | 20 |
| $C_{30}$ ($C_{2v}$) II | 117 | 30053 | 2.1651 | 18 | 6 | 20 |
| $C_{30}$ ($D_{5h}$) | 151 | 31945 | 2.0672 | 20 | 10 | 20 |
| $C_{32}$ ($D_3$) | 144 | 55140 | 2.1968 | 15 | 2 | 18 |
| $C_{32}$ ($C_2$) I | 151 | 55705 | 2.1780 | 16 | 4 | 16 |
| $C_{32}$ ($C_2$) II | 168 | 57092 | 2.1375 | 17 | 6 | 16 |
| $C_{32}$ ($D_2$) | 184 | 58384 | 2.1045 | 18 | 8 | 15 |
| $C_{34}$ ($C_{3v}$) | 195 | 103665 | 2.1902 | 15 | 3 | 15 |
| $C_{34}$ ($C_s$) | 196 | 104484 | 2.1896 | 15 | 3 | 16 |
| $C_{34}$ ($C_2$) I | 204 | 103544 | 2.1714 | 14 | 2 | 14 |
| $C_{34}$ ($C_2$) II | 212 | 107720 | 2.1632 | 17 | 6 | 16 |
| $C_{36}$ ($D_{6h}$) | 272 | 192528 | 2.1706 | 12 | 0 | 12 |
| $C_{36}$ ($D_{2d}$) | 288 | 192720 | 2.1489 | 12 | 0 | 12 |
| $C_{36}$ ($C_{2v}$) | 312 | 197340 | 2.1231 | 13 | 2 | 10 |
| $C_{36}$ ($D_{3h}$) | 364 | 207924 | 2.0764 | 15 | 6 | 6 |
| $C_{38}$ ($C_{2v}$) | 360 | 366820 | 2.1768 | 14 | 2 | 14 |
| $C_{38}$ ($C_{3v}$) | 378 | 363300 | 2.1572 | 12 | 1 | 9 |
| $C_{38}$ ($D_{3h}$) | 456 | 411768 | 2.1116 | 18 | 8 | 18 |
| $C_{40}$ ($D_{5d}$) I | 562 | 515781 | 2.0775 | 10 | 0 | 10 |
| $C_{40}$ ($T_d$) | 576 | 704640 | 2.1185 | 12 | 4 | 0 |
| $C_{40}$ ($D_{5d}$) II | 701 | 803177 | 2.0750 | 20 | 10 | 20 |
| $C_{44}$ (T) | 864 | 2478744 | 2.1775 | 12 | 4 | 0 |
| $C_{44}$ ($D_{3h}$) | 960 | 2436480 | 2.1416 | 9 | 2 | 0 |
| $C_{60}$ ($I_h$) | 12500 | 395974320 | 2.0986 | 0 | 0 | 0 |
| $C_{70}$ ($D_{5h}$) | 52168 | – | – | 0 | 0 | 0 |
| $C_{82}$ ($C_s$) | – | – | – | 0 | 0 | 0 |

The values for the 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* is the Kekulé structure count. 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: (*a*) parameters *p*, *q* and *r* include some redundant information, (*b*) the error of some parameters is large, and (*c*) non–linear effects of *p*, *q* and *r* can affect ln[per(**A**)]/ln *K*, ln *K* or ln[per(**A**)] [14]. In this work, a different strategy has been used: (*a*) smaller superpositions of the *p*–*q* and *p*–*r* pairs were sought, (*b*) not all the three structural parameters were necessarily retained in the fits, and (*c*) non–linear correlations were allowed. The best linear correlation of ln[per(**A**)]/ln *K* for the first 29 fullerenes in Table 1 is:

$$\ln[\mathrm{per}(\mathbf{A})]/\ln K = 2.14 - 0.0108q + 0.00364r$$
$$n = 29 \quad R = 0.721 \quad s = 0.036 \quad F = 14.1 \quad \mathrm{MAPE} = 1.21\% \quad \mathrm{AEV} = 0.4803 \tag{1}$$

The mean absolute percentage error (MAPE) is 1.21% and the approximation error variance (AEV) is 0.4803. All other models with greater MAPE and AEV have been discarded. As there were several fullerenes with the same set of *p*, *q* and *r* parameters, Eq. (1) explains 95% of the correlation coefficient of the means ($n = 24$, $R = 0.757$). On the other hand, the best non–linear correlation of ln[per(**A**)]/ln *K* with the structural parameters results:

$$\begin{aligned}
\ln[\mathrm{per}(\mathbf{A})]/\ln K &= 2.13 + 0.0515 z_{41} \\
z_{41} &= 0.225 z_{31} + 1.20 z_{32} \\
z_{31} &= -1.16 + 0.232q \\
z_{32} &= 1.05 z_{22} - 0.875 z_{21} z_{22} \\
z_{21} &= 1.22 - 0.0983r + 0.00277qr \\
z_{22} &= -0.726 z_{11} - 0.921 z_{12} \\
z_{11} &= -1.16 + 0.232q \\
z_{12} &= 1.22 - 0.0983r + 0.00277qr \\
\mathrm{MAPE} &= 0.87\% \qquad\qquad \mathrm{AEV} = 0.2432
\end{aligned} \tag{2}$$

and AEV decreases 49%. For ln *K* alone, the best linear correlation for the first 30 fullerenes in Table 1 is:

$$\ln K = 10.1 - 0.376p + 0.255q$$
$$n = 30 \quad R = 0.965 \quad s = 0.401 \quad F = 181.6 \quad \mathrm{MAPE} = 4.21\% \quad \mathrm{AEV} = 0.0692 \tag{3}$$

Equation (3) explains 98% of the correlation coefficient of the means ($n = 24$, $R = 0.982$). The best non–linear model does not improve the results. For ln[per(**A**)] alone, the best linear correlation for the first 29 fullerenes in Table 1 is:

$$\ln[\mathrm{per}(\mathbf{A})] = 20.2 - 0.660\,p + 0.383q$$
$$n = 29 \quad R = 0.949 \quad s = 0.757 \quad F = 118.5 \quad \mathrm{MAPE} = 4.05\% \quad \mathrm{AEV} = 0.0988 \tag{4}$$

Equation (4) explains 97% of the correlation coefficient of the means ($n$ = 24, $R$ = 0.977). On the other hand, the best non–linear correlation results are:

$$\ln[\mathrm{per}(\mathbf{A})] = 20.0 - 0.666\,p + 0.616\,q - 0.00850\,pq$$
$$\mathrm{MAPE} = 3.91\% \quad \mathrm{AEV} = 0.0871 \tag{5}$$

and AEV decreases 12% with respect to the linear fit. Small superpositions of the $p$–$q$ and $p$–$r$ pairs are observed in Eqs. (1)–(5). This diminishes the risk of co–linearity in the fits given the close relationship between each pair $p$–$q$ and $p$–$r$ [23].

**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*K* vs *p*,*q*,*r* | ln*K* vs *p*,*q* | ln[per(**A**)] vs *p*,*q*,*r* | ln[per(**A**)] vs *p*,*q* | ln*K* vs *p*,*q*,*r* (means) | ln*K* vs *p*,*q* (means) |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.551 | 0.623 | 0.935 | 0.943 | 0.930 | 0.932 | 0.974 | 0.975 |
| 2 | 0.550 | 0.623 | 0.935 | 0.943 | 0.930 | 0.932 | 0.974 | 0.975 |
| 3 | 0.548 | 0.622 | 0.936 | 0.944 | 0.930 | 0.932 | 0.973 | 0.975 |
| 4 | 0.546 | 0.622 | 0.937 | 0.944 | 0.930 | 0.932 | 0.973 | 0.974 |
| 5 | 0.544 | 0.622 | 0.938 | 0.944 | 0.929 | 0.932 | 0.973 | 0.974 |
| 6 | 0.542 | 0.621 | 0.939 | 0.945 | 0.929 | 0.932 | 0.972 | 0.974 |
| 7 | 0.540 | 0.621 | 0.939 | 0.945 | 0.929 | 0.932 | 0.972 | 0.974 |
| 8 | 0.538 | 0.620 | 0.940 | 0.946 | 0.928 | 0.932 | 0.972 | 0.974 |
| 9 | 0.536 | 0.619 | 0.941 | 0.946 | 0.928 | 0.932 | 0.971 | 0.974 |
| 10 | 0.534 | 0.619 | 0.942 | 0.946 | 0.927 | 0.932 | 0.971 | 0.974 |

The correlation coefficient found between cross–validated representatives and the property values $R_{cv}$ has been calculated with the leave–*n*–out procedure [24]. The procedure furnishes a new method for selecting the best set of descriptors according to the criterion of maximization of the value of $R_{cv}$. The $R_{cv}$ calculations for fullerenes are given in Table 2 for $1 \le n \le 10$. In general, $R_{cv}$ decreases with $n$. However, for both ln $K$ methods $R_{cv}$ increases with $n$. The effect is corrected when the set of points is substituted by the means *(cf.* the last two columns in Table 2). In particular, the method ln[per(**A**)]/ln $K$ *vs.* {$q$,$r$} gives greater $R_{cv}$ than *vs.* {$p$,$q$,$r$} for the whole range of $n$ given in Table 2. The same happens for both ln $K$ and ln[per(**A**)] *vs.* {$p$,$q$}, which give greater $R_{cv}$ than *vs.* {$p$,$q$,$r$}. The corresponding interpretation is that the {$q$,$r$} set of descriptors is more predictive than the {$p$,$q$,$r$} set 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**)].

On the other hand, the upper triangle of the 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 between *p*–*r* and *p*–*q*. Both single– and complete–linkage hierarchical cluster analyses allow building the dendrogram for the structural parameters *p*, *q* and *r* of fullerenes (Figure 3) [25]. The cluster analysis performs a binary taxonomy of the structural parameters that separates first the *q* parameter. Further, the *p* and *r* counts are set apart.

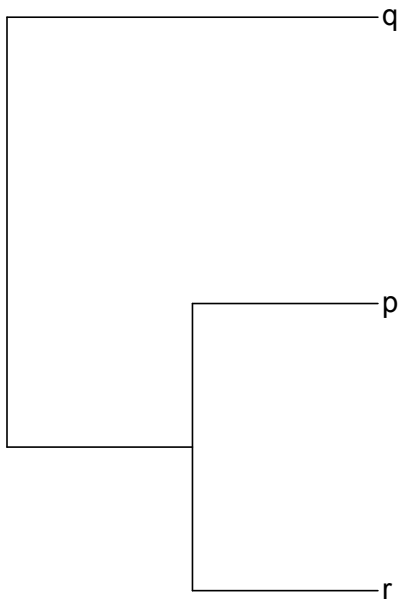



**Figure 3.** Dendrogram for the *p*, *q* and *r* counts of fullerenes.

From both cluster analyses, the radial tree is built for the structural parameters *p*, *q* and *r* of fullerenes *(cf*. Figure 4). The radial tree in Figure 4 is in agreement with the dendrogram (Figure 3).

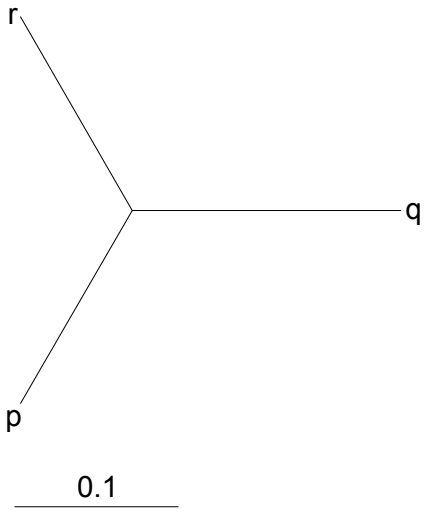



**Figure 4.** Radial tree graph for the *p*, *q* and *r* counts of fullerenes.

SplitsTree is an interactive program for analyzing and visualizing clustering data [26]. Based on the method of split decomposition, it takes as input a distance matrix or a set of clustering 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* and *r* of the fullerenes is displayed in Figure 5. The splits graph in Figure 5 reveals that a conflicting relationship exists between *p*, and parameters *q* and *r*. This is due to the interdependence between *p*, *q* and *r*. Hence, the splits graph indicates a spurious relationship resulting from base composition effects.

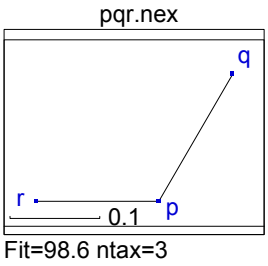



**Figure 5.** The splits graph for the *p*, *q* and *r* counts of fullerenes.

The PCA for the structural parameters *p*, *q* and *r* results in three factors $F_1$– $F_3$, which are linear combinations of *p*, *q* and *r*. The coefficients for factor $F_1$ are:

$$F_1 = 0.602p + 0.561q + 0.569r \quad (6)$$

The coefficients for factor $F_2$ are:

$$F_2 = -0.055p + 0.739q - 0.671r \quad (7)$$

The coefficients for factor $F_3$ are:

$$F_3 = 0.797p - 0.372q - 0.476r \quad (8)$$

**Table 3.** Importance of the Principal Component Analysis Factors.

| Factor | Eigenvalue | Percentage | Accumulated percentage |
|---|---|---|---|
| $F_1$ | 2.59619787 | 86.54 | 86.54 |
| $F_2$ | 0.31005411 | 10.34 | 96.88 |
| $F_3$ | 0.09374803 | 3.12 | 100.00 |

The importance of PCA factors $F_1$–$F_3$ for the structural parameters of the fullerenes is collected in Table 3. In particular, the use of only the first factor $F_1$ explains 87% of the variance and gives a relative error of 13%. However, the use of the two first factors, $F_1$ and $F_2$, explains 97% of the variance, reducing the relative error to 3%.

**Table 4.** Profile of the Principal Component Analysis Factors.

| Factor | Percentage of p | Percentage of q | Percentage of r |
|---|---|---|---|
| $F_1$ | 36.19 | 31.49 | 32.32 |
| $F_2$ | 0.31 | 54.64 | 45.05 |
| $F_3$ | 63.51 | 13.86 | 22.63 |

The profile of PCA factors $F_1$–$F_3$ for the structural parameters of the fullerenes is resumed in Table 4. In particular, factor $F_1$ cannot be reduced to two variables ($p$ and $r$) without making a relative error of 31% (the percentage of $q$). For both $F_1$ and $F_3$ factors, variable $p$ has the greatest weight in the profile. On the other hand, for factor $F_2$ the most important variable is $q$. In some way, factors $F_1$ and $F_3$ could be considered as linear combinations of $p$ and $r$ (with relative errors of 31% and 14%, respectively). However, factor $F_2$ can be expressed as a linear combination of $q$ and $r$ within a relative error of only 0.3%.

PCA $F_2$ *vs*. $F_1$ plot for the fullerenes is illustrated in Figure 6. Fullerenes in classes 2 and 3 with the same set of $p$, $q$ and $r$ values in Table 1 appear superposed in Figure 6. Three classes are clearly distinguished: class 1 with 7 members (below the bisector, $F_1 > F_2$, bottom of Figure 6), class 2 with 17 members (near the bisector, $F_1 \approx F_2$, middle of Figure 6) and class 3 with 7 members (above the bisector, $F_1 < F_2$, top of Figure 6). In general, fullerenes with the same number of atoms belong to the same class. The exceptions are the isomers of $C_{30}$, $C_{32}$, $C_{36}$ and $C_{40}$ fullerenes, which are members of two classes. However, no fullerene has isomers belonging to the three classes.

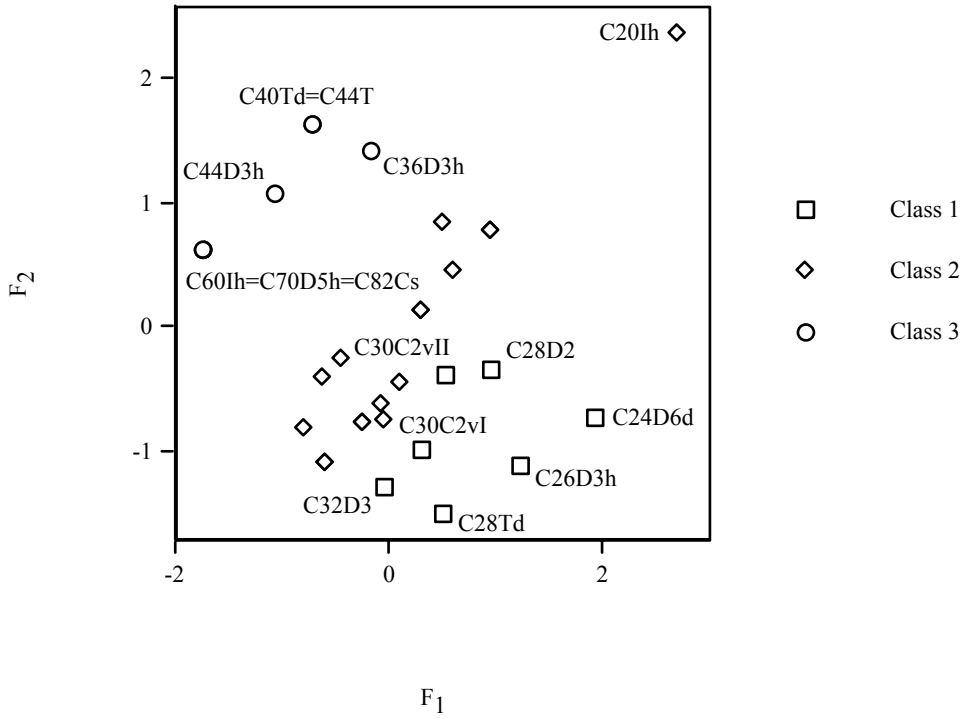



**Figure 6.** PCA $F_2$ *vs*. $F_1$ plot for the fullerenes.

On the other hand, 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 $p$, $q$ and $r$ structural parameters is shown in Figure 7. The tree provides a

binary taxonomy of the fullerenes in Table 1, which separates first the 7 fullerenes in class 1 [from $C_{28}$ ($T_d$) to $C_{26}$ ($D_{3h}$), *top* of Figure 7], then the 17 fullerenes in class 2 [from $C_{34}$ ($C_s$) to $C_{38}$ ($C_{3v}$), *middle* of Figure 7] and finally the 7 fullerenes in class 3 [from $C_{36}$ ($D_{3h}$) to $C_{82}$ ($C_s$), *bottom* of Figure 7]. These classes correspond to those obtained by PCA (Figure 6).

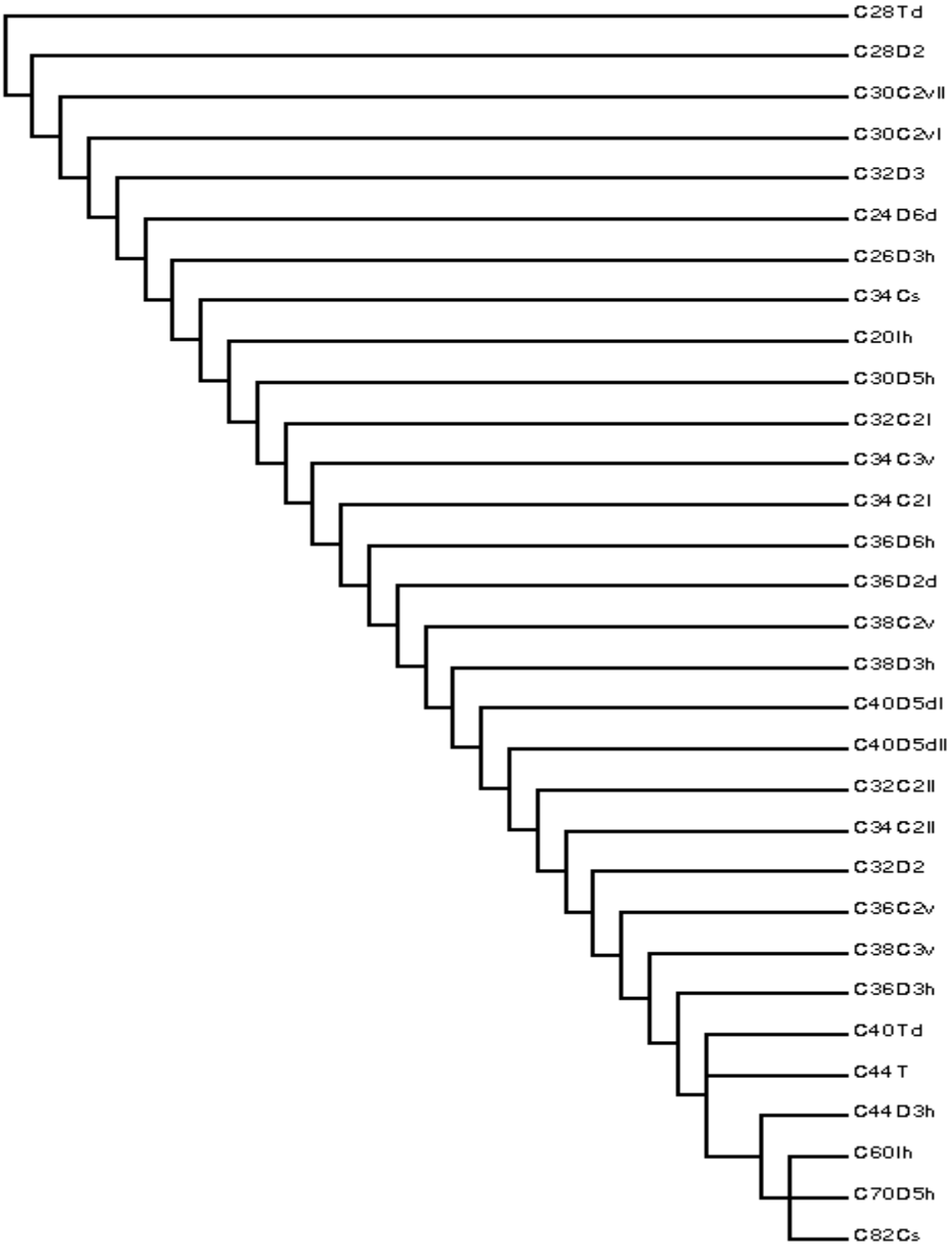



**Figure 7.** Dendrogram for the fullerenes.

The radial tree for the fullerenes relating to the *p*, *q* and *r* structural parameters is displayed in Figure 8. It separates first the 7 fullerenes in class 1 [$C_{28}$ ($T_d$)–$C_{26}$ ($D_{3h}$), right of Figure 8], then the 17 fullerenes in class 2 [$C_{34}$ ($C_s$)–$C_{38}$ ($C_{3v}$), middle of Figure 8] and finally the 7 fullerenes in class 3 [$C_{36}$ ($D_{3h}$)–$C_{82}$ ($C_s$), left of Figure 8]. These classes correspond to those obtained by PCA (Figure 6) and dendrogram (Figure 7).

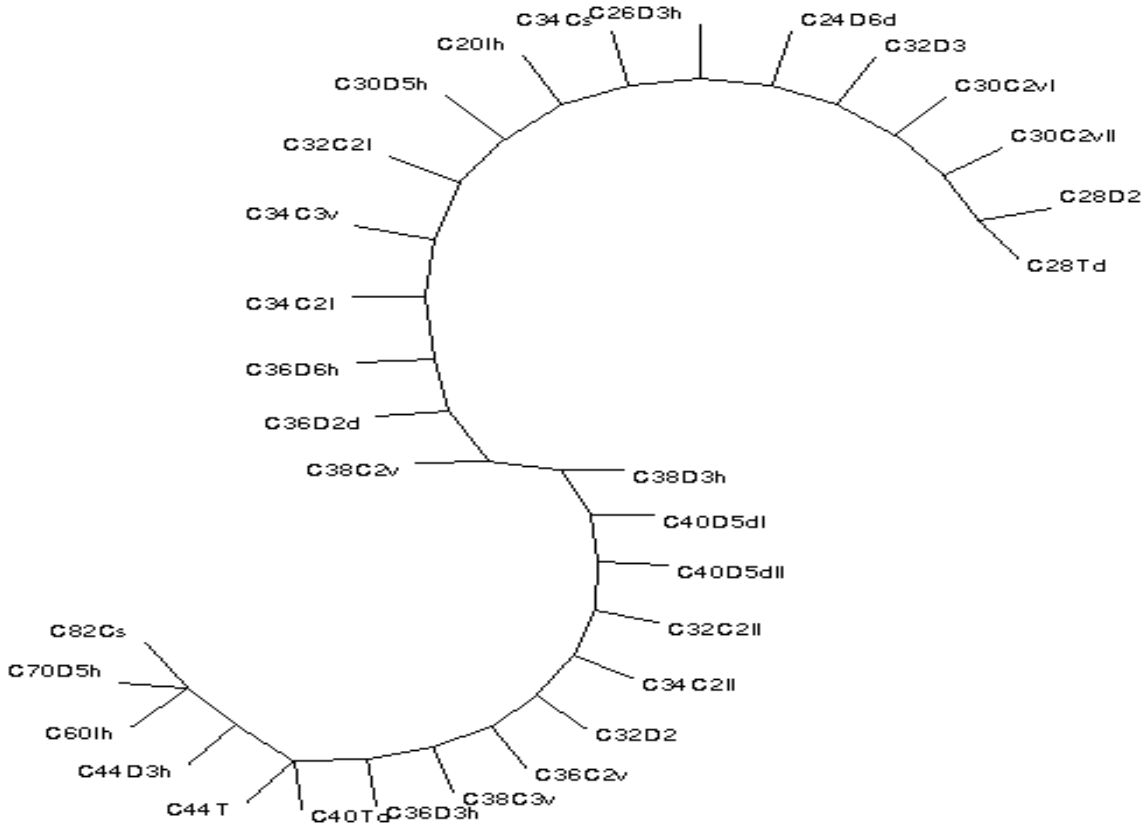



**Figure 8.** Radial tree graph for the fullerenes.

Class 1 shows rather high values of the *p*, *q* and *r* structural parameters for its small number of C atoms. These arrangements (especially *q* and *r)* decrease the stability of the fullerenes. This low stability is corroborated by the greatest ln[per(**A**)]/ln *K* values (2.18 on average) in Table 1. This result is far from that expected for alternant hydrocarbons (2.00), which are more stable. The corresponding interpretation is that high *p*, *q* and *r* indices drop the Kekulé structure count *K*, with a subsequent rise in ln[per(**A**)]/ln *K*. Class 2 presents relatively high values of *p*, *q* and *r* for its moderate size. These reasonable *q* and *r* counts cause an intermediate stability of the fullerenes. This intermediary stability is in agreement with lower ln[per(**A**)]/ln *K* values (2.13 on average). The interpretation is that relatively high *p*, *q* and *r* parameters decrease *K*, causing moderate values of ln[per(**A**)]/ln *K*. Class 3 exhibits low values of *p*, *q* and *r* for its great largeness. These low *q* and *r* counts increase the stability of the fullerenes. This high stability corresponds to the lowest ln[per(**A**)]/ln *K* values (2.12 on average). The interpretation is that low *p*, *q* and *r* sums rise *K*, with

a resultant drop in ln[per(**A**)]/ln *K*.

**Table 5.** Heats of Formation and Related Data for Carbon Clusters $C_n$.

| Class | Fullerene | Conjugated–Circuit Per–Site Ratios to Graphite [a] | Corrected Hückel Delocalization Energy Per–Site Ratios to Graphite [a] | Heat of formation ($\Delta H_f$) per atom [b] |
|---|---|---|---|---|
| 1 | $C_{28}$ ($T_d$) | –0.008 | –0.394 | 31.11 |
| 2 | $C_{40}$ ($D_{5d}$) | 0.322 | 0.313 | 28.35 |
| 3 | $C_{60}$ ($I_h$) | 0.712 | 0.676 | 14.49 |

[a] Taken from Ref. [29]
[b] In kcal·mol$^{-1}$, taken from Ref. [30]

Referring to the mass spectra for fullerenes, Rohlfing *et al*. [27] showed that some class–3 species have the highest relative abundances and therefore are observed to have the greatest relative stability of the fullerenes. Campbell and Hertel [28] concluded that the presence of such a large number of fullerenes in the mass spectra supports the stability of fullerenes generally. The large intensities for certain fullerene masses indicate a specially high stability for certain fullerenes, such as $C_{60}$ and $C_{70}$, which belong to class 3. Schmalz *et al*. [29] carried out conjugated–circuit counts and Hückel molecular orbital (HMO) calculations *(cf*., *e.g*. Table 5), and concluded that both conjugated–circuit values per atom and HMO delocalization energies per atom corrected for $\pi$–strain predict a greater stability for larger (class 3) clusters. Bakowies and Thiel [30] calculated with modified neglect of diatomic overlap (MNDO) the heats of formation *(e.g*. Table 5) and concluded that the heats of formation per atom predict a greater stability for larger (class 3) clusters.

## 4 CONCLUSIONS

From the preceding results the following conclusions can be drawn.

1. The results for the Kekulé structure count and permanent of the adjacency matrix of fullerenes are given for a series of structures up to $C_{70}$ and $C_{60}$, respectively. With the permanent now open to computation, a great deal of work remains to be done to characterize the relationship of the permanent to chemical structure and properties. Much future work remains to be done in elucidating the extent to which the permanent encodes structural features in a quantitative way as well as in exploring the relationship of the permanent to structure in fullerenes.

2. 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 49%. It has also diminished the risk of co–linearity in the fit. The cross–validation leave–*n*–out procedure shows that the most predictive set of descriptors according to the criteria of maximization of $R_{cv}$ are {*q*,*r*} for ln[per(**A**)]/ln *K*, and {*p*,*q*} for both ln *K* and ln[per(**A**)].

3. The cluster analysis shows the greatest similarity for the *p* and *r* parameters. Split

decomposition indicates a spurious relationship resulting from base composition effects.

4. PCA provides three orthogonal factors $F_1$–$F_3$. The use of only $F_1$ gives a relative error of 13%. The use of $F_1$ and $F_2$ decreases the relative error to 3%. The fullerenes have been grouped in three 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 the three classes.

5. The similarity between fullerenes has been compared with the cluster analysis of these molecules. The cluster analysis 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. Naural 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. Chemometrics* **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, *J. Educ. Psychol.* **1933**, *24*, 417.
[8] H. Hotelling, *J. Educ. Psychol.* **1933**, *24*, 489.
[9] S. Wold, K. Esbensen and P. Geladi, *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37.
[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] R. A. Jarvis and E. A. Patrick, Clustering Using a Similarity Measure Based on Shared Nearest Neighbours, *IEEE Trans. Comput.* **1973**, *C–22*, 1025–1034.
[16] 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.
[17] 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.
[18] 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.
[19] 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.

[20] Integrated Mathematical Statistical Library (IMSL), IMSL, Houston, 1989.
[21] X. Liu, D. J. Klein, T. G. Schmalz and W. A. Seitz, Generation of Carbon–Cage Polyhedra, *J. Comput. Chem.* **1991**, *12*, 1252–1259.
[22] G. G. Cash, Permanents of Adjacency Matrices of Fullerenes, *Polycycl. Arom. Compounds* **1997**, *12*, 61–69.
[23] R. R. Hocking, The Analysis and Selection of Variables in Linear Regression, *Biometrics* **1976**, *32*, 1–49.
[24] E. Besalú, Fast Computation of Cross–Validated Properties in Full Linear Leave–Many–Out Procedures, *J. Math. Chem.* **2001**, *29*, 191–203.
[25] R. D. M. Page, Program TreeView, Universiy of Glasgow, 2000.
[26] D. H. Huson, SplitsTree: Analizing and Visualizing Evolutionary Data, *Bioinformatics* **1998**, *14*, 68–73.
[27] E. A. Rohlfing, D. M. Cox and A. Kaldor, Production and Characterization of Supersonic Carbon Cluster Beams, *J. Chem. Phys.* **1984**, *81*, 3322–3330.
[28] E. E. B. Campbell and I. V. Hertel, Molecular Beam Studies of Fullerenes, *Carbon* **1992**, *30*, 1157–1165.
[29] T. G. Schmalz, W. A. Seitz, D. J. Klein, and G. E. Hite, Elemental Carbon Cages, *J. Am. Chem. Soc.* **1988**, *110*, 1113–1127.
[30] D. Bakowies and W. Thiel, MNDO Study of Large Carbon Clusters, *J. Am. Chem. Soc.* **1991**, *113*, 3704–3714.

## 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, development and application of high–precision quantum methods, and theoretical and computational chemistry. Scientific accomplishments include the first implementation in a computer at the Universitat de València 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.