Theor Chem Acc (2003) 110: 371–376
DOI 10.1007/s00214-003-0490-y

Theoretical
Chemistry Accounts

Regular article

New structural parameters of fullerenes for principal component analysis

Francisco Torrens

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

Received: 2 September 2002 / Accepted: 12 May 2003 / Published online: 1 December 2003

Abstract. The Kekulé structure count and the permanent of the adjacency matrix of fullerenes are related to structural parameters involving the presence of contiguous pentagons p, q, r, q/p and r/p, where p is the number of edges common to two pentagons, q is the number of vertices common to three pentagons and r is the number of pairs of nonadjacent pentagons adjacent to another common pentagon. The cluster analysis of the structural parameters allows classification these parameters. Principal component analysis (PCA) of the structural parameters and the cluster analyses of the fullerenes permit their classification. PCA clearly distinguishes five classes of fullerenes. The cluster analysis of fullerenes is in agreement with PCA classification. Cluster analysis shows greatest similarity for the q–q/p and r–r/p pairs. PCA provides five orthogonal factors F_1–F_5. The use of F_1 gives an error of 28%. The inclusion of F_2 decreases the error to 2%.

Keywords: Cluster analysis – Dendrogram – Split decomposition – Principal component analysis – Similarity matrix

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, for example, a structure—activity study. If the vectors are only two-dimensional, they can be plotted in a plane. This allows the 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 2, 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 two-dimensional ones. 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 two-dimensional 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, PCA) or nonlinear (e.g., nonlinear mapping, self-organizing map). Comparisons of the quality of projection methods were described elsewhere [1, 2, 3, 4, 5, 6].

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, or factors) of the covariance matrix as a new basis [8]. Principal components corresponding to the two largest eigenvalues (variance) are used to produce two-dimensional plots [9]. The quality of the projection is commonly expressed by the retained variance of the first two principal components. 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.

From the Proceedings of the 28th Congreso de Químicos Teóricos de Expresión Latina (QUITEL 2002)

e-mail: francisco.torres@uv.es

In order 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 five structural parameters. For this example, the 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, 11, 12].

Balaban et al. [13] reported extensive computations of a number of graph invariants for many fullerenes. Diudea et al. [14] devised a novel way to construct toroidal fullerenes from square tiled tori. Aihara and Hirama [15] concluded that antiaromatic species are scarcely formed in interstellar space. Aihara [16] studied the spherical aromaticity in charged fullerenes and the $2(N+1)^2$ rule. Ivanciuc et al. [17] presented a qualitative resonance-theoretic view for the description of a variety of conjugated π-network species identified with subgraphs of the graphite network [17].

In a previous paper, the calculation of the Kekulé structure count and the permanent of adjacency matrices [18] was applied to fullerenes with different structural parameters involving the presence of contiguous pentagons [19]. PCA of the structural parameters was carried out [20]. In this work, two new structural parameters have been introduced and PCA has been performed. The aim of this paper is to analyze the interdependence between the structural parameters, to classify them, and to classify the fullerenes. The computational methods are presented in Sect. 2. The calculation results for fullerenes are discussed in Sect. 3. The conclusions are summarized in Sect. 4.

Computational methods

PCA is used to transform a number of potentially correlated variables into the same number of independent variables, which can then be ranked on the basis of their contributions for explaining the whole data set. The transformed variables that can explain all the information in the data are called principal components or factors. The first principal component, 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. Principal components having minor contribution to the data set may be discarded without losing too much information. If the number of principal components is less than 4 then the multidimensional data can be graphed in two-dimensional or three-dimensional space, i.e., PCA can be used to reduce dimensionality. The main purpose of employing PCA is to reduce the number of variables (principal components) used in the analysis. PCA creates new variables as linear combinations of all the initial variables so that the first principal component contains the largest variance, the second principal component contains the second largest variance, and so on, until the last principal component 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 jth variable are compared with the mean of this jth variable. In fact, the transformed value $x'_{ij} = \left(x_{ij} - \bar{x}_j\right)/\sigma_j$ where σ_j is the standard deviation of the jth variable. PCA, which consists in finding the eigenvalues and eigenvectors of the covariance matrix, produces standardized variables to diagonalize the correlation matrix of the initial variables. In effect, principal components 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 principal components F_1 and F_2. The profile of a principal component F_i is the vector of the squared C_{ik} coefficients $\left(C_{i1}^2, C_{i2}^2, \ldots, C_{iP}^2\right)$. Each C_{ik}^2 represents the weight of variable k in principal component F_i. It gives the fraction of each variable in principal component F_i.

On the other hand, one approach to the diversity problem is to cluster a structural database or virtual library on the basis of some kind of structural criteria. Standard approaches for clustering can be broken into hierarchical and nonhierarchical. Hierarchical approaches can be further categorized as agglomerative or divisive. In a nonhierarchical approach, a nearest-neighbor list is created and used to assemble members into related clusters [21]. There are many reasons to cluster a database of molecular structures [22, 23, 24, 25].

A program based on the IMSL [26] subroutine CLINK has been written to carry out hierarchical cluster analysis from a correlation or similarity matrix. Initially, each data point is considered to be a cluster, numbered 1 to $n = N_{\mathrm{pt}}$, where N_{pt} is the number of data points to be clustered. Clustering proceeds in five steps. Step 0. Set the counter $k = 1$. 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. 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 (Fig. 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.

Calculation results and discussion

The structural features involving adjacent pentagons are encoded by the p, q and r parameters as illustrated in Fig. 2. The p and q parameters enumerate, respectively, the number of edges common to two pentagons and the number of vertices common to three

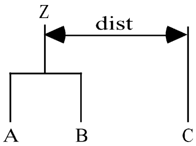

Fig. 1. Distance between clusters Z and C

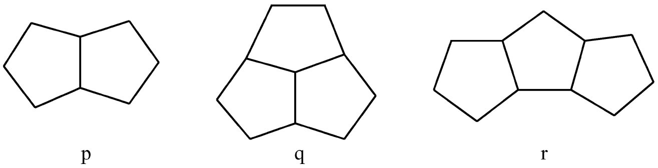

Fig. 2. Substructures that contribute to the p, q and r counts

pentagons [27]. The r count enumerates the number of pairs of nonadjacent pentagon edges shared with two other pentagons [28]. Thus, q and r complement each other by counting both possible arrangements of three contiguous pentagons. However, there is a close relationship between p, q and r, for example, the minimum structure with $q=1$ (Fig. 1, q) needs $p=3$, and the minimum structure with $r=1$ (Fig. 1, r) requires $p=2$. The interdependence between p, q and r suggests expanding the set of parameters with the q/p and r/p quotients.

The values for the structural parameters involving the presence of contiguous pentagons are given in Table 1. Much chemical graph-theory work revolved around the adjacency matrices **A**. 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. A motivation for the consideration of K is that K is never zero for fullerenes [29]. 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)]/ln$K$ with p, q and r. Despite his good results, three important remarks were made: (1) parameters p, q and r include some redundant information; (2) the error of some parameters is large; (3) nonlinear effects of p, q and r can affect ln[per(**A**)]/ln$K$, ln$K$ or ln[per(A)]/ln$K$ [18]. Therefore, a different strategy was used: (1) smaller superpositions of $p–q$ and $p–r$ were sought; (2) not all the three parameters were necessarily retained; (3) nonlinear correlations were allowed.

The best linear correlation of ln[per(**A**)]/ln$K$ with {p}, {q}, {r}, the three pairs and {p,q,r} for the fullerenes is

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

The mean absolute percentage error (MAPE) is 1.21% and the approximation error variance (AEV) is 0.4803. There are 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 of p, q and r, yet with rather widely varying values of per(**A**) and K. For instance, 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 application of the present fits is restricted to smaller fullerenes ($N<70$). As there are several fullerenes with the same set of p, q and r parameters, Eq. (1) explains 95% of the correlation coefficient of the mean ($n=24$, $R=0.757$). On the other hand, the best nonlinear correlation of ln[per(**A**)]/ln$K$ with {p}, {q}, {r}, {p,q},....,{p,q,r} is

Table 1. Values of p, q and r counts for fullerenes

Fullerene	K	per(**A**)	ln[per(**A**)]/ln$K$	p	q	r	q/p	r/p
C_{20} (I_h)	36	1,392	2.0199	30	20	30	0.6667	1.0000
C_{24} (D_{6d})	54	4,692	2.1192	24	12	36	0.5000	1.5000
C_{26} (D_{3h})	63	8,553	2.1853	21	8	30	0.3810	1.4286
C_{28} (T_d)	75	15,705	2.2378	18	4	24	0.2222	1.3333
C_{28} (D_2)	90	16,196	2.1540	20	8	24	0.4000	1.2000
C_{30} (C_{2v}) I	107	29,621	2.2034	17	4	20	0.2353	1.1765
C_{30} (C_{2v}) II	117	30,053	2.1651	18	6	20	0.3333	1.1111
C_{30} (D_{5h})	151	31,945	2.0672	20	10	20	0.5000	1.0000
C_{32} (D_3)	144	55,140	2.1968	15	2	18	0.1333	1.2000
C_{32} (C_2) I	151	55,705	2.1780	16	4	16	0.2500	1.0000
C_{32} (C_2) II	168	57,092	2.1375	17	6	16	0.3529	0.9412
C_{32} (D_2)	184	58,384	2.1045	18	8	15	0.4444	0.8333
C_{34} (C_{3v})	195	103,665	2.1902	15	3	15	0.2000	1.0000
C_{34} (C_s)	196	104,484	2.1896	15	3	16	0.2000	1.0667
C_{34} (C_2) I	204	103,544	2.1714	14	2	14	0.1429	1.0000
C_{34} (C_2) II	212	107,720	2.1632	17	6	16	0.3529	0.9412
C_{36} (D_{6h})	272	192,528	2.1706	12	0	12	0.0000	1.0000
C_{36} (D_{2d})	288	192,720	2.1489	12	0	12	0.0000	1.0000
C_{36} (C_{2v})	312	197,340	2.1231	13	2	10	0.1538	0.7692
C_{36} (D_{3h})	364	207,924	2.0764	15	6	6	0.4000	0.4000
C_{38} (C_{2v})	360	366,820	2.1768	14	2	14	0.1429	1.0000
C_{38} (C_{3v})	378	363,300	2.1572	12	1	9	0.0833	0.7500
C_{38} (D_{3h})	456	411,768	2.1116	18	8	18	0.4444	1.0000
C_{40} (D_{5d}) I	562	515,781	2.0775	10	0	10	0.0000	1.0000
C_{40} (T_d)	576	704,640	2.1185	12	4	0	0.3333	0.0000
C_{40} (D_{5d}) II	701	803,177	2.0750	20	10	20	0.5000	1.0000
C_{44} (T)	864	2,478,744	2.1775	12	4	0	0.3333	0.0000
C_{44} (D_{3h})	960	2,436,480	2.1416	9	2	0	0.2222	0.0000
C_{60} (I_h)	12,500	395,974,320	2.0986	0	0	0	–	–
C_{70} (D_{5h})	52,168	–	–	0	0	0	–	–
C_{82} (C_s)	–	–	–	0	0	0	–	–

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

and AEV decreases by 49%. If q/p and r/p are included in the model the best linear fit is

$$
\begin{aligned}
\ln[\mathrm{per}(A)]/\ln K = 1.88 &+ 0.0361p - 0.0490q\\
&+0.00953r + 0.0497q/p\\
&-0.253r/p
\end{aligned}
\tag{3}
$$
$$
n = 28,\ R = 0.941,\ s = 0.019,\ F = 34.2
$$
$$
\mathrm{MAPE} = 0.66\,\%,\quad \mathrm{AEV} = 0.1558
$$

and AEV decreases by 68%. Equation (3) explains 98% of the correlation coefficient of the mean ($n=23$, $R=0.956$). The best nonlinear model does not improve the results.

There are already powerful exact computational approaches for K. For arbitrary chemical graphs enumeration via Heilbronner recursion is feasible up to about 90 atoms. Better efficiency occurs with Kasteleyn's method as applies for all planar graphs (including all fullerenes). This simply involves the evaluation of the determinant of a signed adjacency matrix $\mathbf{A}'$ [30], where extension has been made to deal with conjugated circuits counts, using the inverse of $\mathbf{A}'$. For InK alone, the best linear correlation for the first 30 fullerenes in Table 1 is

$$
\begin{aligned}
&\ln K = 10.1 - 0.376p + 0.255q\\
&n = 30,\ R = 0.965,\ s = 0.401,\ F = 181.6,\\
&\mathrm{MAPE} = 4.21\%,\quad \mathrm{AEV} = 0.0692
\end{aligned}
\tag{4}
$$

Equation (4) explains 98% of the correlation coefficient of the mean ($n=24$, $R=0.982$). The use of nonlinear models or the inclusion of q/p and r/p does not improve the results.

For In[per(**A**)] alone, the best linear correlation for the fullerenes in Table 1 is

$$
\begin{aligned}
&\ln[per(A)] = 20.2 - 0.660p + 0.383q\\
&n = 29,\ R = 0.949,\ s = 0.757,\ F = 118.5\\
&\mathrm{MAPE} = 4.05\%,\quad \mathrm{AEV} = 0.0988
\end{aligned}
\tag{5}
$$

Equation (5) explains 97% of the correlation coefficient of the mean ($n=24$, $R=0.977$). On the other hand, the best nonlinear correlation is

$$
\begin{aligned}
&\ln[per(A)] = 20.0 - 0.666p + 0.616q - 0.00850pq\\
&\mathrm{MAPE} = 3.91\%,\quad \mathrm{AEV} = 0.0871
\end{aligned}
\tag{6}
$$

and AEV decreases by 12% with respect to the linear fit. The inclusion of q/p and r/p does not improve the results. Small p–q and p–r superpositions are observed in Equations (1), (2), (4), (5) and (6). This diminishes the risk of collinearity [31] in the fits given the close relationship among p, q and r. The correlation coefficient between cross-validated representatives and the property values R_{cv} has been calculated with the leave-n-out procedure [32]. The ln[per(**A**)]/ln$K$ versus $\{p,q,r,q/p,r/p\}$ method gives greater R_{cv} than the ln[per(**A**)]/ln$K$ versus $\{q,r\}$ and ln[per(**A**)]/ln$K$ versus $\{p,q,r\}$ methods. Both ln$K$ and ln[per(**A**)] versus $\{p,q\}$ methods give greater R_{cv} than the ln$K$ and ln[per(**A**)] versus $\{p,q,r\}$ methods. The corresponding interpretation is that the $\{p,q,r,q/p,r/p\}$ set is more predictive than $\{q,r\}$ or $\{p,q,r\}$ for ln[per(**A**)]/ln$K$, and that $\{p,q\}$ is more predictive than $\{p,q,r\}$ for both ln$K$ and ln[per(**A**)].

On the other hand, the upper triangle of the correlation matrix **R** for $\{p,q,r,q/p,r/p\}$ is

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

High correlation is obtained between q–q/p, p–q, r–r/p and p–r. The correlation between the derived q/p and r/p parameters is 20 times smaller than that between the primary q and r.

Both single-linkage and complete-linkage hierarchical cluster analyses allow building the dendrogram for p, q, r, q/p and r/p of fullerenes [33]. The cluster analysis performs a binary taxonomy of the structural parameters that separates first r–r/p from p–q–q/p. Further, the p count is set apart. Finally, q is disconnected from q/p, and r from r/p. The earliest separation of r (with r/p) is in agreement with the high value of R_{pq}. From the cluster analysis, the radial tree is built for p, q, r, q/p and r/p of the fullerenes. The radial tree is in agreement with the dendrogram. On the other hand, the method of split decomposition takes a distance matrix or a set of clustering data and produces a graph that represents the relationships between the taxa [34]. For ideal data, this graph is a tree, whereas less ideal data will give rise to a treelike 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 treelike given data are. The splits graph for p, q, r, q/p and r/p reveals that a conflicting relationship exists between p and parameters q–q/p and r–r/p. This is due to the interdependence between p, q and r. Therefore, the splits graph indicates a spurious relationship resulting from base composition effects. The r–p–q portion of the splits

Table 2. Coefficients for the principal component analysis factors $F_i = ap + bq + cr + dq/p + er/p$

Factor	*a*	*b*	*c*	*d*	*e*
F_1	0.523	0.480	0.470	0.421	0.314
F_2	−0.045	−0.342	0.385	−0.501	0.694
F_3	0.389	0.439	−0.010	−0.753	−0.297
F_4	−0.141	−0.240	0.779	−0.002	−0.561
F_5	0.744	−0.634	−0.154	0.068	−0.130

graph is in qualitative agreement with a previous study of the $\{p,q,r\}$ set, which also indicated a spurious relationship between p, q and r resulting from base composition effects.

PCA for the structural parameters p, q, r, q/p and r/p results in five factors, which are linear combinations of p, q, r, q/p and r/p. The coefficients for the factors are listed in Table 2. 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 explains 72% of the variance and gives a relative error of 28%. Moreover, the use of the first two factors explains 98% of the variance, reducing the relative error to 2%. Furthermore, the use of the first three factors explains 99.4% of the variance, reducing the relative error to only 0.6%.

Table 3. Importance of the principal component analysis factors

Factor	Eigenvalue	Percentage	Accumulated percentage
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

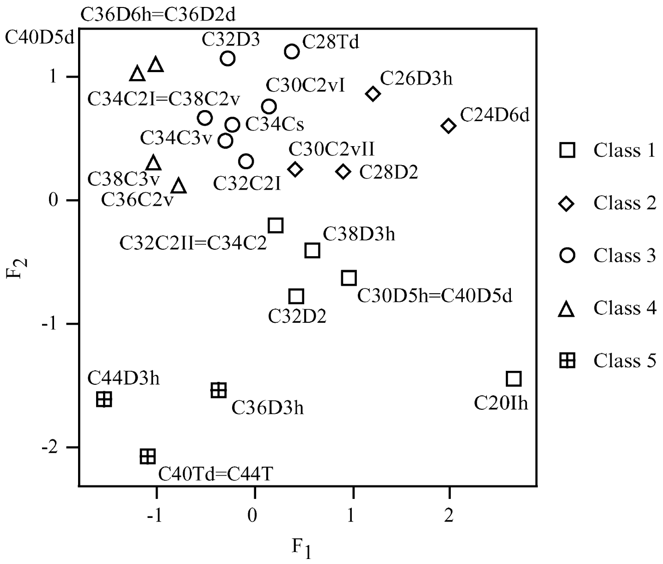

Fig. 3. Principal component analysis F_2 versus F_1 plot for the fullerenes

The PCA F_2 versus F_1 plot for the fullerenes is illustrated in Fig. 3. Fullerenes in classes 1, 3, 4 and 5 with the same set of p, q, r, q/p and r/p values in Table 1 appear superposed in Fig. 3. Five classes are clearly distinguished: class 1 with seven members (below the bisector, $F_1 \gg F_2$, middle right of Fig. 3), class 2 with four members (under the bisector, $F_1 > F_2 > 0$, top right of Fig. 3), class 3 with eight members (over the bisector, $F_1 < F_2$, top of Fig. 3), class 4 with five members (above the bisector, $F_1 \ll F_2$, top left of Fig. 3) and class 5 with four members (under the bisector, $0 > F_1 > F_2$, bottom left of Fig. 3). In general, fullerenes with the same number of atoms belong to the same class. The exceptions are the isomers of C_{28}, C_{30}, C_{32}, C_{34}, C_{36}, C_{38} and C_{40}, which are members of two or three classes. However, no fullerene has isomers belonging to four or five classes. With the purpose of classifying C_{60}, C_{70} and C_{82}, PCA analysis was repeated with the $\{p,q,r\}$ set. This PCA F_2 versus F_1 plot grouped C_{60}—C_{82} in class 5, close to C_{44} (D_{3h}).

On the other hand, instead of N fullerenes 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 was built and the similarity of the fullerenes was compared. The dendrogram for the fullerenes matching to p, q, r, q/p and r/p was calculated. The tree provides a binary taxonomy of the fullerenes in Table 1, which separates the fullerenes in the same classes as PCA (Fig. 3). With the purpose of classifying C_{60}—C_{82}, the dendrogram was repeated for $\{p,q,r\}$. The result was the inclusion of C_{60}–C_{82} in a new branch connected to C_{44} (D_{3h}). The radial tree for the fullerenes relating to p, q, r, q/p and r/p was calculated. It separates first the seven fullerenes in class 1, then the four fullerenes in class 2, the eight fullerenes in class 3, the five fullerenes in class 4 and the four fullerenes in class 5. These classes correspond to those obtained by PCA (Fig. 3) and the dendrogram. With the purpose of classifying C_{60}–C_{82}, the radial tree was repeated for $\{p,q,r\}$. The result was the inclusion of C_{60}–C_{82} in a new branch connected to C_{44} (D_{3h}), as shown in Fig. 4.

Conclusions

From the preceding results the following conclusions can be drawn.

1. The results for the Kekulé structure count and the permanent of the adjacency matrix of fullerenes are given for a series of structures up to C_{60}. A great deal of work remains to be done to characterize the relationship of the permanent to chemical structure and properties.
2. Linear and nonlinear 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 nonlinear regression for ln[per(**A**)]/ln$K$ has been improved. The variance decreased by 68%. It has

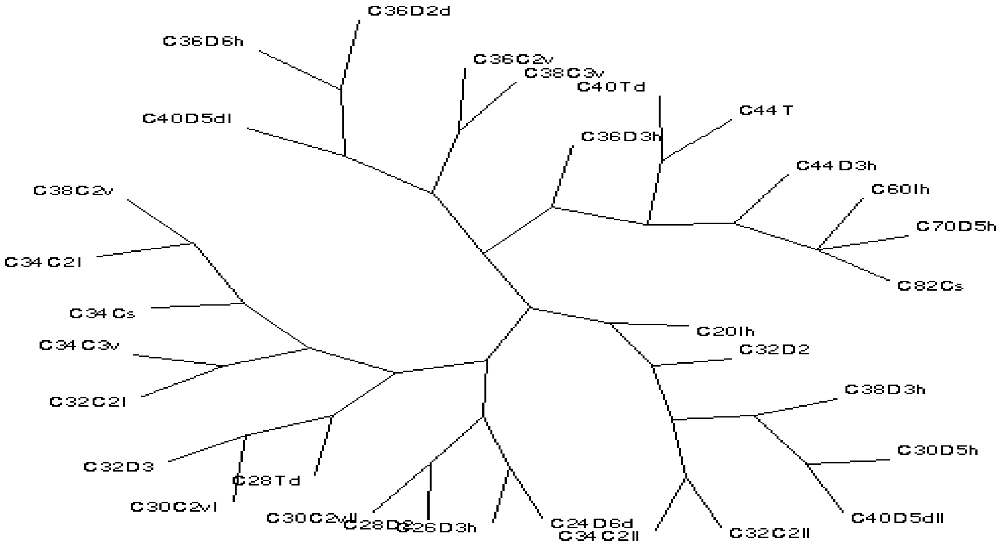

Fig. 4. Radial tree graph for the fullerenes

diminished the risk of collinearity. The most predictive set is $\{p,q,r,q/p,r/p\}$ for $\ln[\mathrm{per}(\mathbf{A})]/\ln K$, and $\{p,q\}$ for both $\ln K$ and $\ln[\mathrm{per}(\mathbf{A})]$.

3. The cluster analysis shows greater similarity for the p–q parameters than with r. Split decomposition indicates a spurious relationship resulting from base composition effects.
4. PCA provides three orthogonal factors. The use of only F_1 gives a relative error of 28%. The use of F_1 and F_2 decreases the error to 2% and groups the fullerenes in five 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.

Acknowledgements. I thank E. Besalú for providing me with 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).

References

1. Kraaiveld MA, Mao J (1995) IEEE Trans Neural Networks 6:548
2. Biswas G, Jain AK, Dubes RC (1981) IEEE Trans Pattern Anal Machine Intell 3:701
3. Sammon JW Jr (1969) IEEE Trans Comput C 18:401
4. Kowalski BR, Bender CF (1972) J Am Chem Soc 94:5632
5. Domine D, Devillers J, Chastrette M, Karcher W (1993) J Chemom 7:227
6. Bienfait B, Gasteiger J (1997) J Mol Graphics Modell 15:203
7. Hotelling H (1933) J Educ Psychol 24:417
8. Hotelling H (1933) J Educ Psychol 24:489
9. Wold S, Esbensen K, Geladi P (1987) Chemom Intell Lab Syst 2:37
10. Brown RD, Martin YC (1996) J Chem Inf Comput Sci 36:572
11. Brown RD, Martin YC (1997) J Chem Inf Comput Sci 37:1
12. Matter H (1997) J Med Chem 40:1219
13. Balaban AT, Liu X, Klein DJ, Babic D, Schmalz TG, Seitz WA, Randić M (1995) J Chem Inf Comput Sci 35:396
14. Diudea MV, Silaghi-Dumitrescu I, Pârv B (2002) Internet Electron J Mol Des 1:10
15. Aihara J, Hirama M (2002) Internet Electron J Mol Des 1:52
16. Aihara J (2002) Internet Electron J Mol Des 1:236
17. Ivanciuc O, Bytautas L, Klein DJ (2002) J Chem Phys 116:4735
18. Torrens F (2002) Int J Quantum Chem 88:392
19. Torrens F (2002) Internet Electron J Mol Des 1:351
20. Torrens F (2003) Internet Electron J Mol Des 2:96
21. Jarvis RA, Patrick EA (1973) IEEE Trans Comput C 22:1025
22. McGregor MJ, Pallai PV (1997) J Chem Inf Comput Sci 37:443
23. Doman TN, Cibulskis JM, Cibulskis MJ, McCray PD, Spangler DP (1996) J Chem Inf Comput Sci 36:1195
24. Turner DB, Tyrrell SM, Willett P (1997) J Chem Inf Comput Sci 37:18
25. Reynolds CH, Druker R, Pfahler LB (1998) J Chem Inf Comput Sci 38:305
26. Integrated Mathematical Statistical Library (IMSL) (1989) IMSL, Houston
27. Liu X, Klein DJ, Schmalz TG, Seitz WA (1991) J Comput Chem 12:1252
28. Cash GG (1997) Polycyclic Aromat Compd 12:61
29. Klein DJ, Liu X (1992) J Math Chem 11:199
30. Klein DJ, Liu X (1991) J Comput Chem 12:1260
31. Hocking RR (1976) Biometrics 32:1
32. Besalú E (2001) J Math Chem 29:191
33. Page RDM (2000) Program TreeView. Universiy of Glasgow, Glasgow
34. Huson DH (1998) Bioinformatics 14:68