ON TOPOLOGICAL INDICES INDICATING BRANCHING PART I. THE PRINCIPAL COMPONENT ANALYSIS OF ALKANE PROPERTIES AND INDICES

A. Perdih*, M. Perdih
Mala vas 12, SI-1000 Ljubljana, Slovenia

Received 9.2.2000

Abstract

The suitability of topological indices J, W, Z, D, MTI, Xu, ID, χ, $\lambda\lambda_1$, EA_{max}, and λ_1 as branching indices, as well as of physicochemical properties MON, BP, d, Vi, Vm, Vc, Tc, Pc, dc, Zc, αc, ΔHv, A, B, C, n_D, MR, a_0, b_0, ΔHf°g, ΔGf°g, S, R^2, and ω, as reference properties for the branching of alkanes is tested by means of the Principal Component Analysis (PCA). On the PCA plots, alkanes are separated by several criteria in the following descending order of importance: carbon number, number of branches, whether the carbons are tertiary or quaternary, the position of branches, the shape and symmetry of molecules. Most properties and indices correlate highly with the carbon number of alkanes and the influence of branching on them is much lower. MON (motor octane number) depends first of all on the number of CH_2 groups. The properties are divided into intrinsic and interaction-dependent ones. It is explained why the latter ones are less suitable as primary references for branching. Two definitions of branching are presented, the *Methane-based* definition as a general definition and the *n-Alkane-based* definition as a special definition more familiar to chemists.

Introduction

Several hundred topological indices have been developed and tested for their performance as branching indices or indices of substances' properties. They have been correlated with several physical, chemical, and biological properties of molecules and the interest in this has grown remarkably during the past years. Therefore, the study of branching indices remains important. In his recent paper, Randić gave an overview of efforts to present the measures of branching in molecules [1]. He stressed the statement of Rouvray [2] that "…ultimately any definition of branching must rest on an intuitive basis. Because of this circumstance, the use of sophistry in defining the concept of branching appears unlikely to lead to a more viable definition". Randić compared several

approaches to this topic and proposed a novel branching index $\lambda\lambda_1$ [1]. He found the support for his novel index in the parallelism between $\lambda\lambda_1$ and λ_1, the largest eigenvalue of the adjacency matrix, as well as in regression with the Motor Octane Number (MON).

If a physicochemical property should be a good reference for the quality of an index, it must depend mostly on branching. In this respect, the primary question is how to find a property that can be used as a measure of branching. One aim of this paper is to find out which alkanes' properties are the most suitable as references for branching. The second aim is to find out which topological indices are the most appropriate as descriptors of branching. As mentioned before, there are several hundred topological indices and also a lot of physicochemical properties. Ideally, one should test all indices as well as all physicochemical properties. Recent studies, however, indicate that a limited number of indices may suffice for this purpose.

Mendiratta and Madan [3] reported that besides the Wiener index (W) [4] the most useful indices are the Hosoya index (Z) [5], the Randić index (χ) [6], and the Balaban index (J) [7]. The most popular branching index, the Wiener index, W, was used even to define molecular branching [8], although it was developed to determine the paraffin boiling points [4]. Another important index is λ_1, the largest eigenvalue of the adjacency matrix [9]. At most a dozen indices emerge as the best single characterisation of diverse physicochemical properties of octanes [10].

On the basis of these findings we decided to study only the most frequently used and some recently presented topological indices (later on: indices). Our decision is based on the assumption that with this selection of indices no relevant information about the molecular structure contained in the information space of all indices is lost.

Since the number of indices and properties is rather large for binary comparisons we first decided to find out which properties and indices contain similar information. For this purpose one of the methods for grouping can be used. We decided to use the Principal Component Analysis (PCA) method. PCA is one of the methods applicable for the analysis of data sets, where features of several objects (alkanes in our case) are presented with several variables (physicochemical properties and/or indices in our case). As the result of the method, objects and variables are grouped according to their

similarity in the space of the principal components. From the grouping patterns it is possible to extract which properties and/or indices are related with branching, or at least to identify those, which are not. A more detailed description of the method is given below.

Results

The data sets used in the Principal Component Analysis are presented in Table 1; the details are given in the text accompanying each Figure. Five sets of data were analysed:

1. A set of properties and indices of alkanes for which the MON value was available;
2. A set of properties available for all alkanes from methane through octanes;
3. A set of indices for all alkanes from ethane through octanes;
4. A set of properties available for all octanes;
5. A set of indices for all octanes;

Table 1. The data sets used to derive the figures.

Fig. No.	Alkanes	Topological indices	Physicochemical properties	Markers
1	36	11	21	7
2	40	0	18	6
3	39	11	0	6
4	18	0	22	4
5	18	11	0	4

The first set was chosen to test the suitability of MON as a reference property and to get the first impression of the relations between the properties and indices.

The second set was chosen to study the relations between those properties for which the data for all 40 alkanes from methane through octanes were available as well as to see how the alkanes were spread in the space of principal components.

The third set was chosen to see how the indices disperse alkanes in the space of principal components to be compared with the dispersion under the influence of properties obtained with the second data set.

The fourth set was chosen to study the relations between the properties of the largest group of alkanes of equal carbon number for which a large set of properties was available, the octanes. The dispersion of octanes under the influence of their properties was also sought to compare it with the dispersion due to the influence of indices.

The fifth set was chosen to study the relations between the indices of the octanes. The dispersion of octanes under the influence of their indices was compared with the dispersion due to the influence of properties.

PCA of properties and indices of alkanes with known MON value

To see whether an index is an appropriate measure of branching it was usually correlated with one or several physical properties of lower alkanes. The correlations were usually good when alkanes of different carbon numbers were taken into account but became in most instances worse or even bad when only data of isomers with equal carbon number were taken into test [22]. For this reason we studied, by means of PCA loading plots, the relations between the most useful indices, the alkanes' properties, and some markers that might be of help in understanding the relations. In this step the data for 36 alkanes out of 40 were studied, from propane to the octanes for which MON data were found:

- 21 properties: MON, BP, d, Vm, V_i, Tc, Pc, Vc, Zc, αc, dc, ΔHf°g, ΔHv, a_0, b_0, ω, A, B, C, n_D, and MR,
- 11 indices: W, Z, J, MTI, ID, χ, $\lambda\lambda_1$, EA, Xu, D, and λ_1,
- 7 markers: Mw, N_C, *Np*, *Ns*, *Nss*, *Nt*, and *Nq*,

For higher alkanes the data available to us were far less complete and therefore they were not included. The results are presented as the loading plot in Fig. 1a. The axes PC1 and PC2 in Fig. 1a explain 74% and 15% of data variance. The following two axes, PC3 and PC4 explain 5% and 2% of variance. They are not shown. In Fig. 1a a dense cluster of properties and indices can be seen on the right side. This cluster is presented enlarged in Fig. 1b. Pc and C correlate highly and negatively with the properties and indices in that cluster. Other properties, i.e., dc, A, MON, Zc, and indices J, EA, and λ_1 are dispersed across Fig. 1a. The markers in the cluster indicate that the axis PC1 is strongly influenced

by the molar mass (carbon number) of alkanes (Mw, N_C, see Fig. 1b). This axis does not separate the number of primary (*Np*) and secondary (*Ns*) carbons in alkanes from one another (see Fig. 1a); it separates a little the number of tertiary carbons (*Nt*), and more the number of quaternary carbons (*Nq*). The axis PC2 separates *Np* well from *Ns* and *Nq* from *Nss*, whereas the axis PC3 separates *Nt* from *Nq*. The separation by the axis PC4 is small.

Fig. 1b presents the dense cluster of Fig. 1a. Since Pc and C highly negatively correlate with the features in this cluster, they were projected through the coordinate origin in order to study the relations between Pc, C, and other features in this cluster more easily. The position of projected variables is indicated in parenthesis.

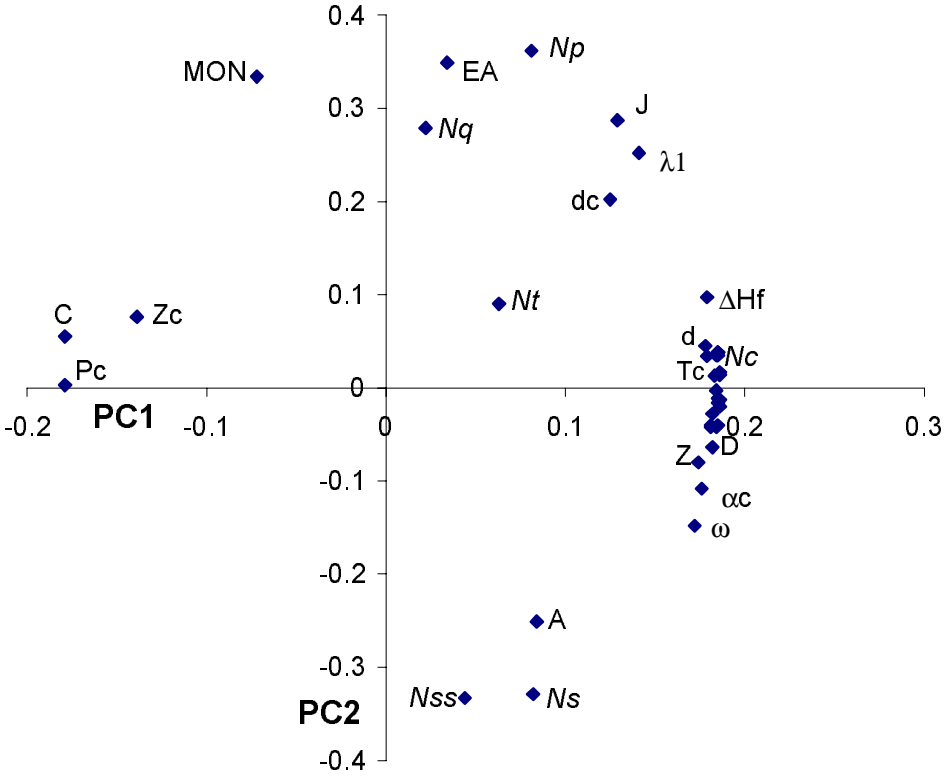

Fig. 1a. The loading plot for the first two principal components of properties and indices of alkanes for which MON is known. The dense cluster on the right side is presented enlarged in Fig. 1b.

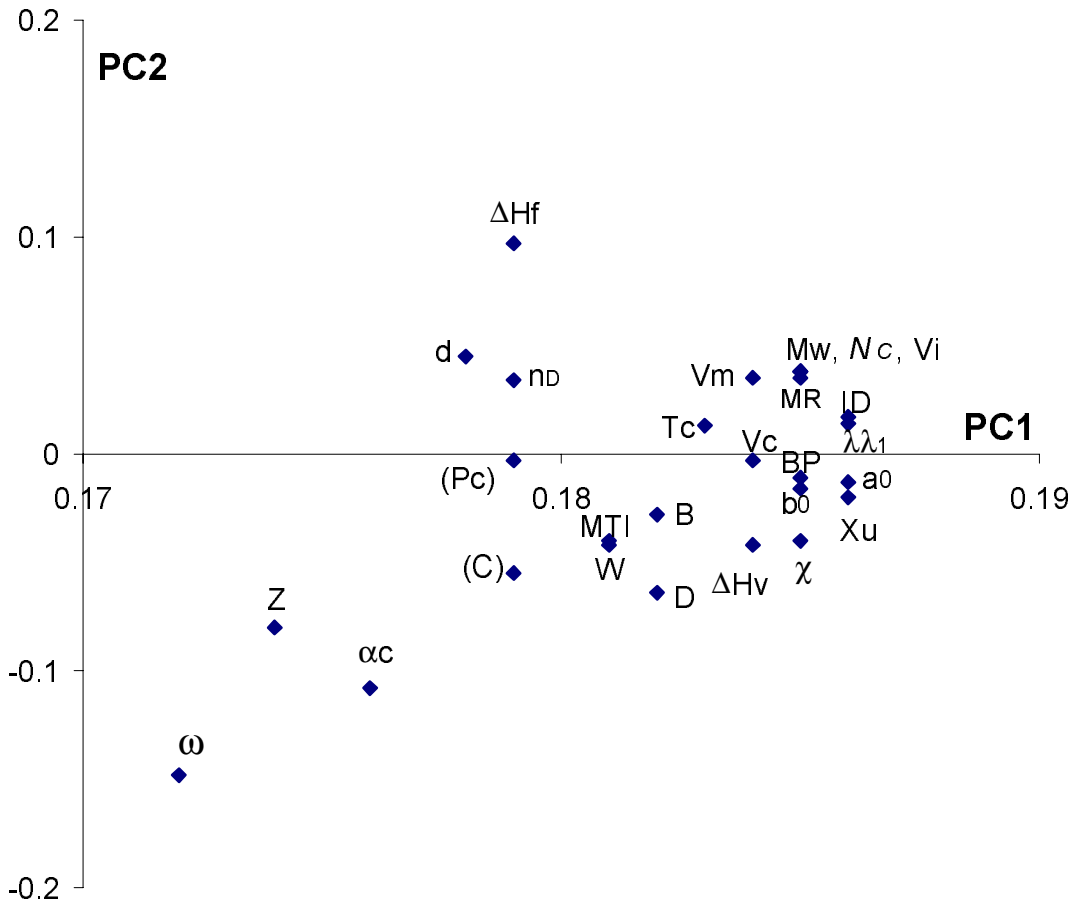

Fig. 1b. The dense cluster in the loading plot for the first two principal components of properties and indices of alkanes for which MON is known.

The molar weight, Mw, the carbon number, N_C, the following properties: Vi, Vm, Vc, BP, Tc, αc, d, ΔHv, B, n_D, MR, a_0, and b_0, as well as the indices $\lambda\lambda_1$, ID, χ, D, MTI, W, Xu and Z form this dense cluster. ΔHf°g is placed above this cluster and ω below it. The properties contained in this dense cluster correlate highly with the carbon number and molar mass and this correlation is presented by r_{NC}. The series of correlation coefficients with the carbon number, r_{NC}, is as follows: 1 = N_C ~ Mw ~ Vi > MR > Vm > a_0 > 0.99 > b_0 > Vc ~ BP > Tc ~ ΔHf°g > ΔHv > B > n_D > d > 0.95 > αc > 0.90 >>> C > -0.95 > Pc. The same holds true for indices: 1 = N_C > ID ~ $\lambda\lambda_1$ > 0.99 > Xu > χ > D ~ MTI ~ W > 0.95 > Z > 0.90. These indices correlate to a high gegree (r = 0.927 to over 0.999) with one another as well as with the properties in the dense cluster.

Mw and N_C have the same, whereas Vi has almost the same value among isomers. They measure only the size of molecules and do not measure their branching. Regarding the branching they can be considered as references for influences other than branching. The high correlation coefficients with them indicate that properties and indices contained in this dense cluster depend first of all on carbon number (and molar weight) of alkanes and much less on other things like branching.

These indices are thus good measures of molar weight and carbon number of alkanes as well as of their properties that are dependent first of all on molar weight. They are not so good measures of their branching. The indices EA, J, and λ_1 correlate less well with molar weight or carbon number (r_{NC} = 0.261, 0.752, and 0.808, respectively). Their correlation reflects their separation from the cluster by the axis PC1: 0 < EA < J < λ_1 < cluster (in the cluster: Z < MTI ~ W < D < $\chi \leq$ **Mw** < Xu < $\lambda\lambda_1$ ~ ID). Also interesting is the position of indices EA, J, and λ_1 regarding the axis PC2. They are placed above the zero line, whereas all other indices are placed lower, below the position of Mw, N_C, and Vi: EA > J > λ_1 > **Mw** > ID ≥ $\lambda\lambda_1$ > Xu > MTI ~ W ~ χ > D > Z. This series corresponds to the fact that the values of indices EA, J, and λ_1 increase with branching, whereas those of indices ID, $\lambda\lambda_1$, Xu, MTI, W, χ, D, and Z decrease with branching.

Motor octane number is often used as a prototype property of alkanes for testing branching indices [1]. According to Fig. 1a, it poorly and negatively correlates with molar weight and carbon number of alkanes. Its highest (and negative) correlation is with the number of CH_2 groups in alkanes ($r_{MON,Ns}$ = -0.863) and slightly less with the number of adjacent CH_2 groups ($r_{MON,Nss}$ = -0.856). Motor octane number is thus in a way related to a property of CH_2 groups and consequently depends on branching indirectly, i.e., only as much as branching diminishes the number of CH_2 groups.

PCA of properties of all alkanes from methane to octanes

To see how different alkanes disperse in the space of principal components we analysed only complete properties' data for all 40 alkanes from methane to octanes. Due to lacking data among other properties, only the following ones were included:

- 18 properties: BP, d, Vm, Tc, Pc, dc, Vc, ΔHf°g, ΔHv, A, B, C, Zc, ω, αc, Vi, a_0, b_0,

- 6 markers: Mw, N_C, *Np*, *Ns*, *Nt*, and *Nq*.

The influence of alkanes' properties was analysed separately from the influence of indices for two reasons. The first one is to see whether the properties and the indices give the same pattern in the spread of alkanes across the figure. The second one is the fact that the Xu index of methane is -∞ and cannot be included into analysis.

The results are presented as score plots in Figs. 2a,b. In Fig. 2a the axes PC1 and PC2 explain 78% and 10% of variance, while in Fig. 2b the axes PC3 and PC4 explain 6% and 3% of variance. The loading plot, corresponding to the score plot in Fig. 2a, is very similar to that in Fig. 1a, forming a dense cluster of properties that correlate highly with carbon number. Therefore it is not shown. The score plot, Fig. 2a, indicates that the axis PC1 separates alkanes first of all by carbon number with some influence of the

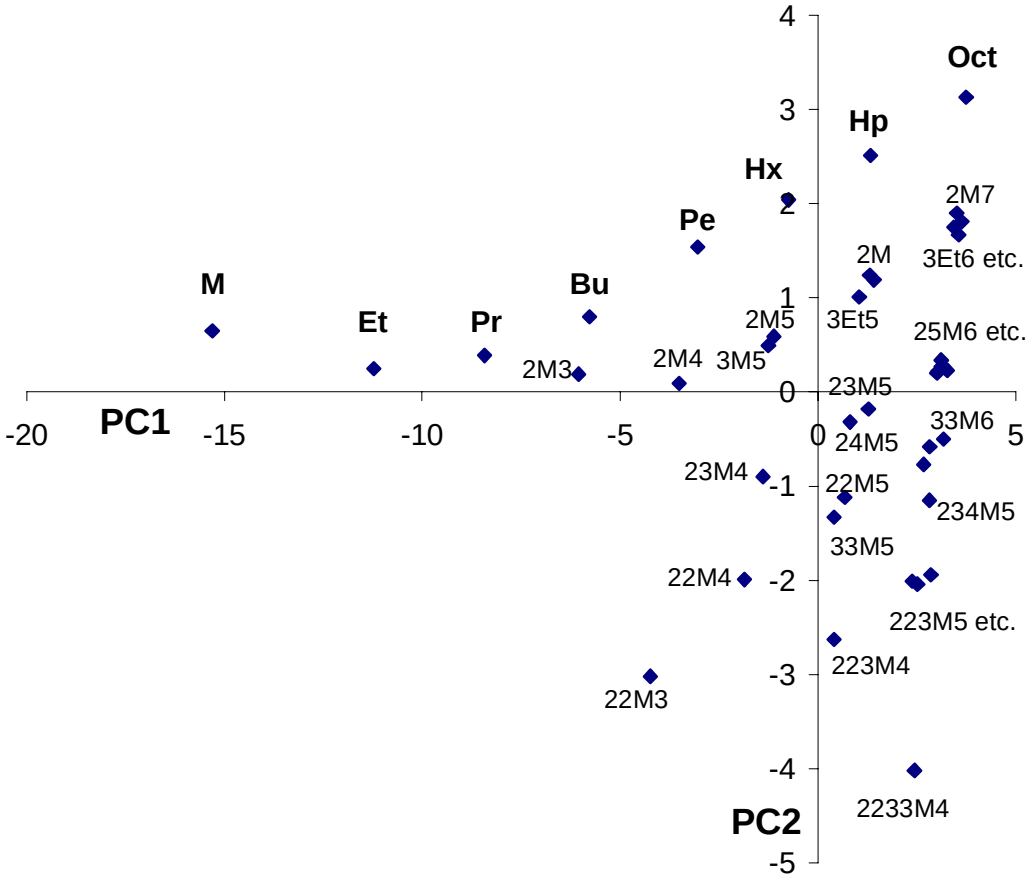

Fig. 2a. The score plot of alkanes under the influence of their properties in the plane of the first two principal components.

number of branches. The axis PC2 separates them by the number and type of branches. The axis PC3, Fig. 2b, emphasises the distinction between the presence of tertiary carbons (right side) and that of secondary and quaternary ones (left side). The axis PC4 separates first of all the elongated and flat molecules from the spherical ones and, among the former, the symmetric from the asymmetric ones. On the other hand, on the axis PC4 is indicated also some separation of the structures having peripheral branches from those having central branches, as well as those having adjacent branches from those having distant branches. Several properties correlate well with one another: N_C, Mw, Vi (r = 1); BP, TC, B; N_C, Vc, a_0, b_0; ΔHv, BP, B (r > 0.99); d, Tc; BP, Vm, Pc, Vc, ΔHf°g, C; BP, αc; Vm, ΔHv, B (|r| > 0.95).

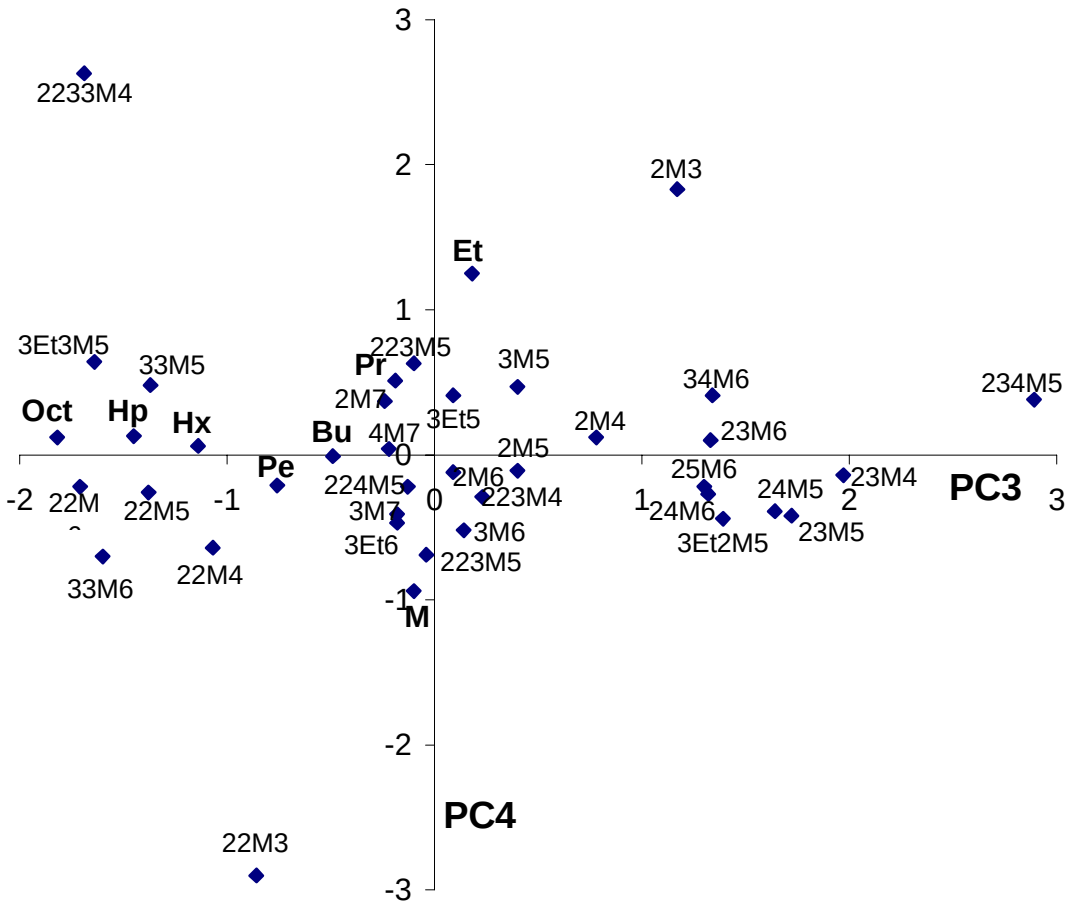

Fig. 2b. The score plot of alkanes under the influence of their properties in the plane of the third and the fourth principal component.

PCA of tested indices of all alkanes from ethane to octanes

Figs. 3a and 3c represent the score plots, i.e. the grouping of alkanes due to the influence of indices and markers, while Fig. 3b presents a detail from Fig. 3a. The features of 39 alkanes, contained in the data set, are presented using 11 tested indices and 6 markers.

The axis PC1 explains 68% of information, i.e. less than in Fig. 2. The axis PC2 explains 20% of information, i.e. twice that in Fig. 2. The axis PC3 explains 9% of information and the axis PC4 only 2%. As in Fig. 2a, also in Fig. 3a the axis PC1 separates alkanes by their molar weight (carbon number).

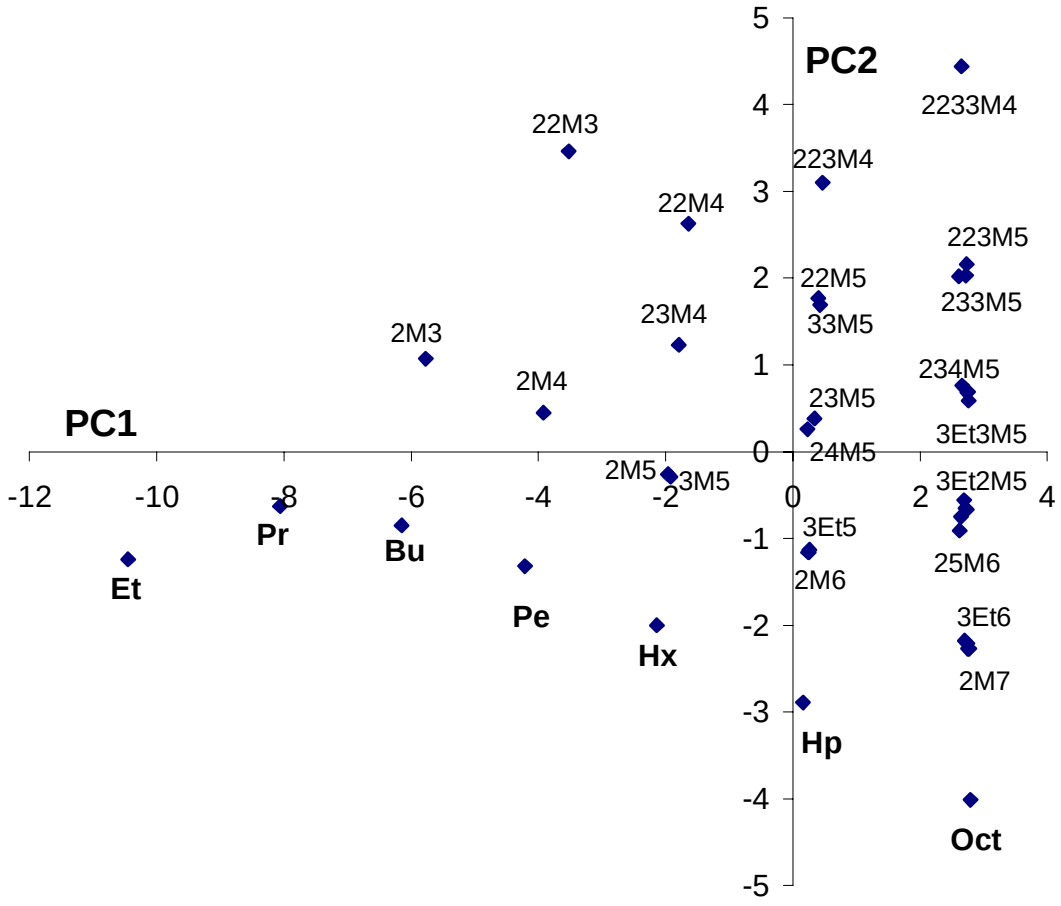

Fig. 3a. The score plot of alkanes under the influence of indices in the plane of the first two principal components.

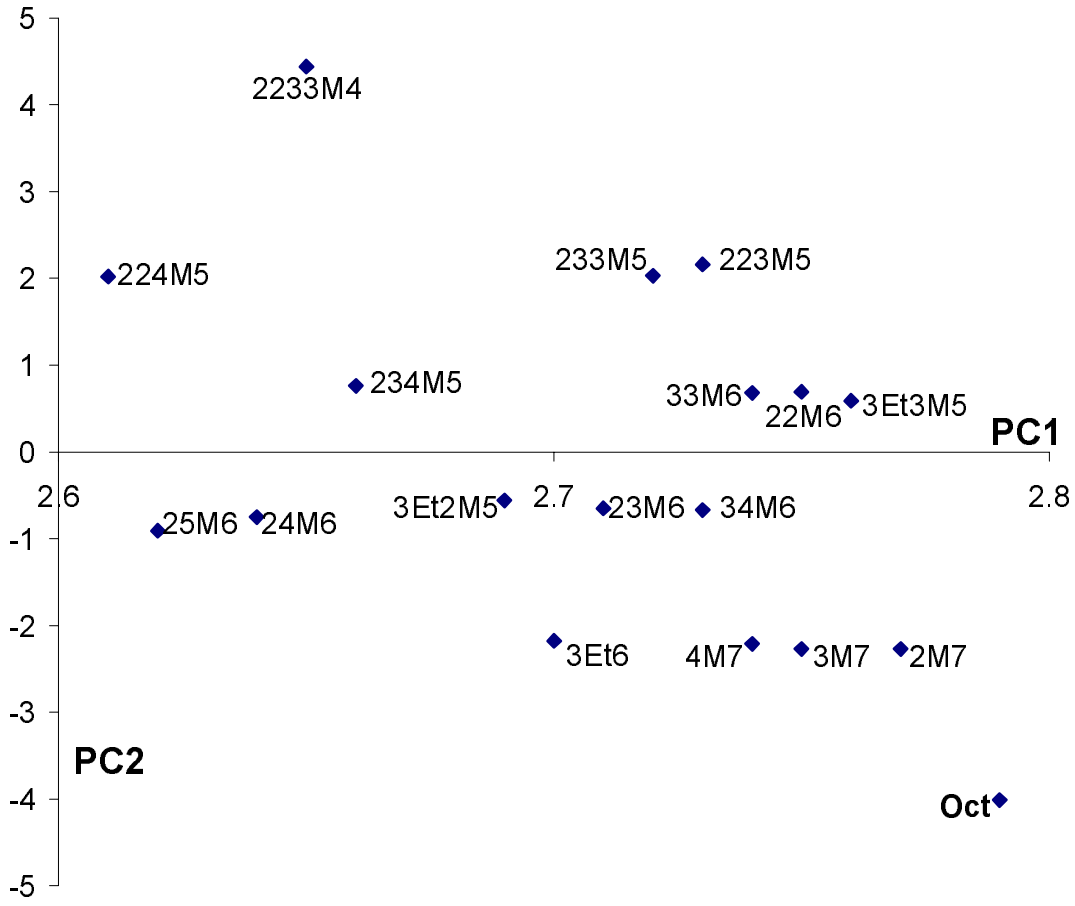

Fig. 3b. The octanes of Fig. 3a.

Alkanes of the same carbon number are grouped into several clusters along the axis PC2. The enlarged view of the grouping of octanes in Fig. 3a is presented in Fig. 3b. These clusters, composed of horizontally placed data points in Fig. 3b, are characterised by the number of branches and they also contain a small amount of information of the type and position of branches.

Fig. 3c presents the distribution of alkanes in the plane of the axes PC3 and PC4. *n*-Alkanes (in bold) are distributed in a parabolic shape. The axis PC3 separates the alkanes with regard to the presence and number of tertiary vs. secondary or quaternary carbons. The axis PC4 shows no dependence on shape and symmetry, but some separation by the type and position of branches, their peripheral vs. central position, and whether they are adjacent or distant. The indices in the loading plots are grouped as in Fig. 1 and for this

reason are not presented. Their correlation with carbon number is $\lambda\lambda_1$ > ID > Xu > χ > D ~ MTI ~ W > Z > λ_1 > J > EA.

The separation pattern in Fig. 2a and Fig. 3a is similar if the inversion of the axis PC2 is disregarded. Different is only the separation by the axis PC4 in Fig. 2b and Fig. 3c. In Fig. 2b the axis PC4 separates first of all the elongated and flat molecules from the spherical ones and among the former the symmetric from the asymmetric ones. Also indicated is some separation of the structures having peripheral branches from those having central branches, as well as of those having adjacent branches from those having distant branches. In Fig. 3c, the axis PC4 shows no dependence on shape and symmetry, but some separation by the type and position of branches, their peripheral vs. central position, and whether they are adjacent or distant.

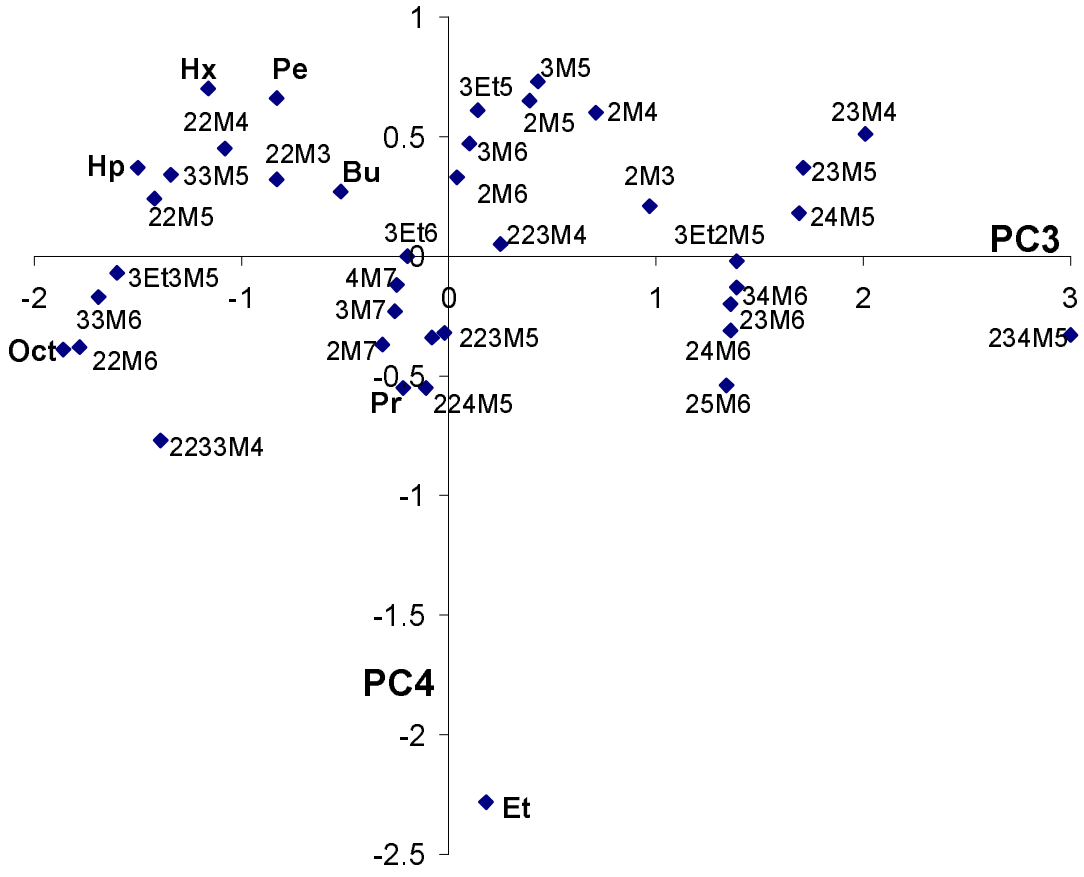

Fig. 3c. The score plot of alkanes under the influence of indices in the plane of the third and the fourth principal component.

PCA of properties and indices in octanes

Since the carbon number has a major influence on the alkanes' properties as well as on indices, we tried to exclude this influence. Like Kirby [22], we tested in this step the properties only for octanes. Octanes were chosen because this is the largest group of isomeric alkanes for which a number of data of their physicochemical properties are known. The resulting plots are shown in Figs. 4a, 4b and 5. In Figs. 4a and 4b the results obtained using the properties known for all 18 octanes are presented:

- 22 properties: BP, d, Vm, Vi, Tc, Pc, Vc, Zc, dc, αc, ΔHf°g, ΔHv, A, B, C, n_D, MR, ω, a_0, b_0, S, and R^2,
- 4 markers: *Np*, *Ns*, *Nt*, and *Nq*.

In Fig. 5 the results obtained using the 11 indices for all octanes are presented.

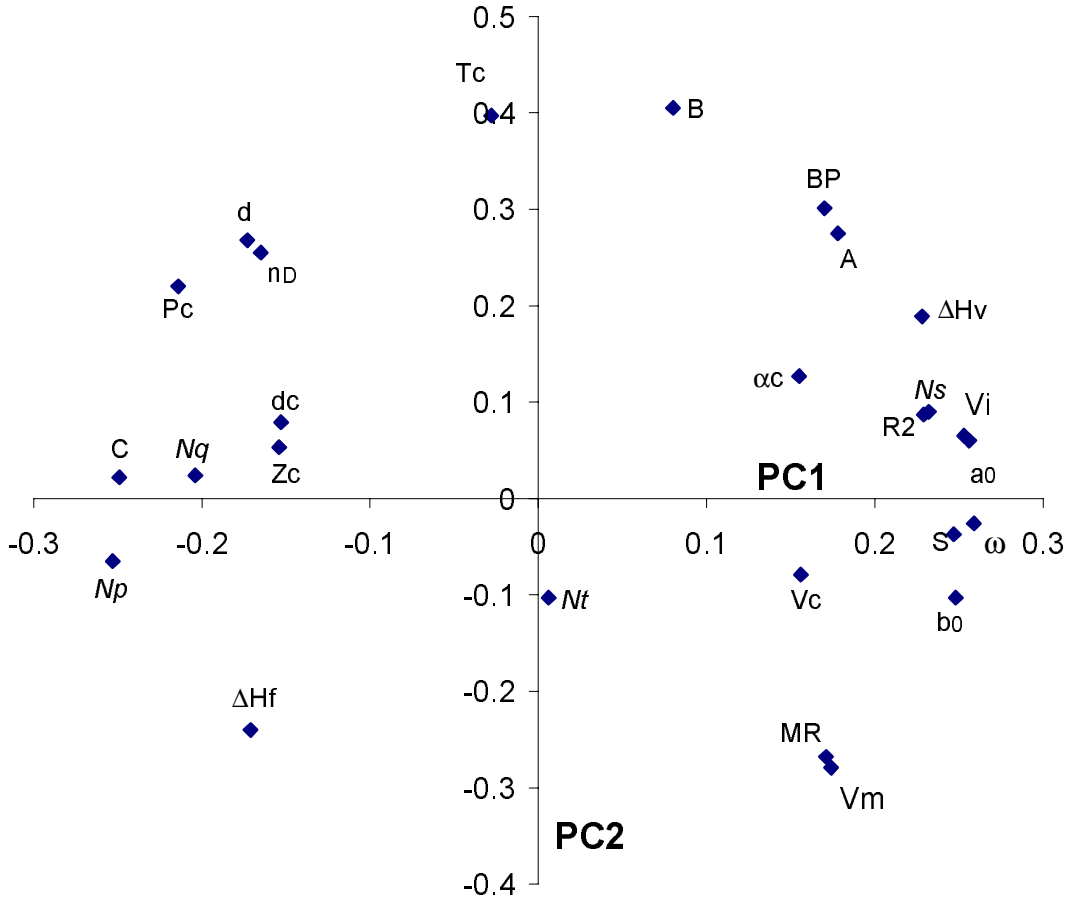

Fig. 4a. The loading plot of the octanes' properties for the first two principal components.

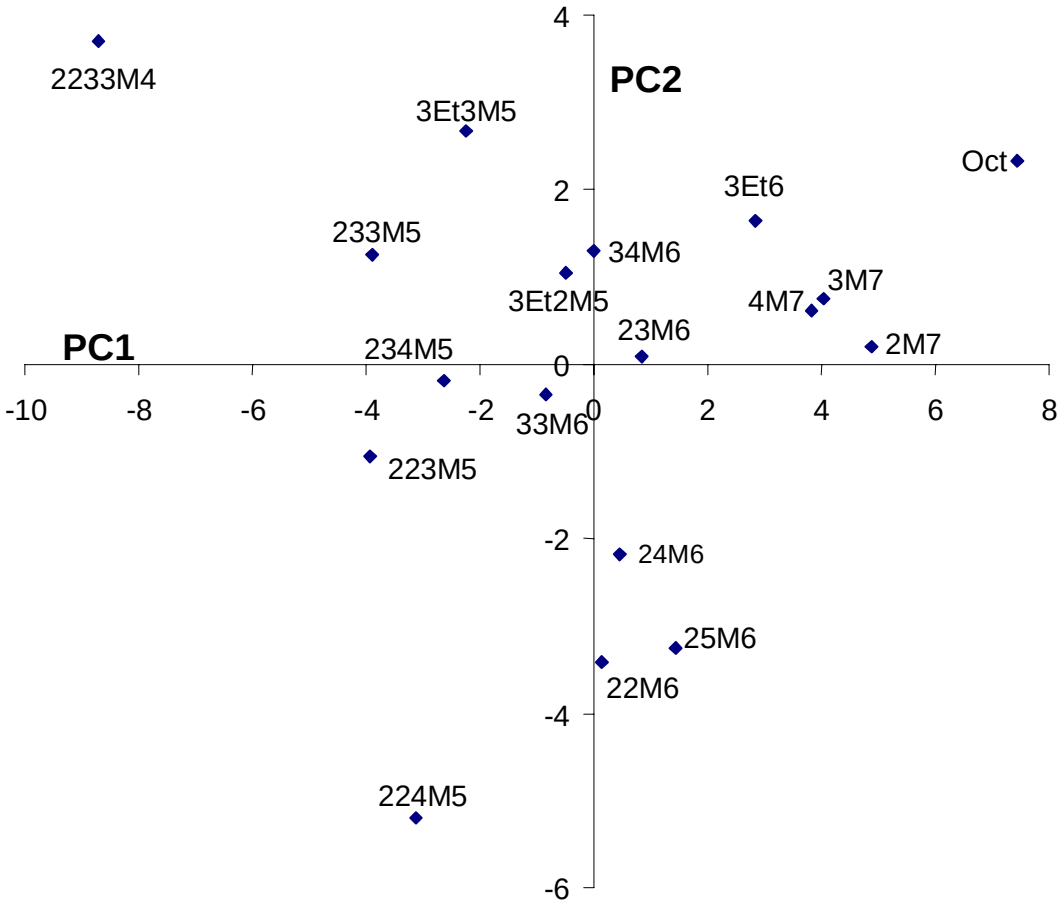

Fig. 4b. The score plot of the octanes in the plane of the first two principal components.

If only the properties known for all octanes are taken into account, then in the PCA plot the axis PC1 explains 57% of variance, i.e. less than in previous cases. The other three axes explain more information than in previous cases: PC2, PC3, and PC4 explain 20%, 12%, and 7% of variance. One of the reasons for the lower information content on the axis PC1 is the fact that the most important difference between the alkanes in Figs. 1-3 was the carbon number, while the alkanes in Figs. 4 have the same carbon number. In the loading plot, Fig. 4a, the properties are spread in a circular manner around the centre. The axis PC1 separates them regarding the changes of their values with increasing branching. The reasons for the separation by the axis PC2, e.g. ΔHf°g from d, or Vm from BP are not clear from this plot. With regard to the position on the axis PC3 (not shown), the properties dc, Vm, Tc, Pc, BP, and S seem to be more dependent on tertiary

than quaternary carbons, whereas the properties d, Vc, Zc, αc, ΔHf°g, and n_D more on quaternary than tertiary ones; the other properties seem to be more or less indifferent.

On the score plot, Fig. 4b, the axis PC1 separates octanes first of all by the number of branches (4 < 3 < 2 < 1 < 0). Much lower is the contribution of the type of branched structure (quaternary < tertiary), the position of branches (central < peripheral), and type of branches (ethyl < methyl). The axis PC2 separates them first of all by the adjacency of branches. The separation criteria on the axis PC3, Fig. 4c, do not seem clear-cut. Separation according to the position of branches (central vs. peripheral, adjacent vs. distant) and symmetry of molecules seems to be indicated. The axis PC4 separates mainly quaternary carbons containing structures from those containing tertiary ones.

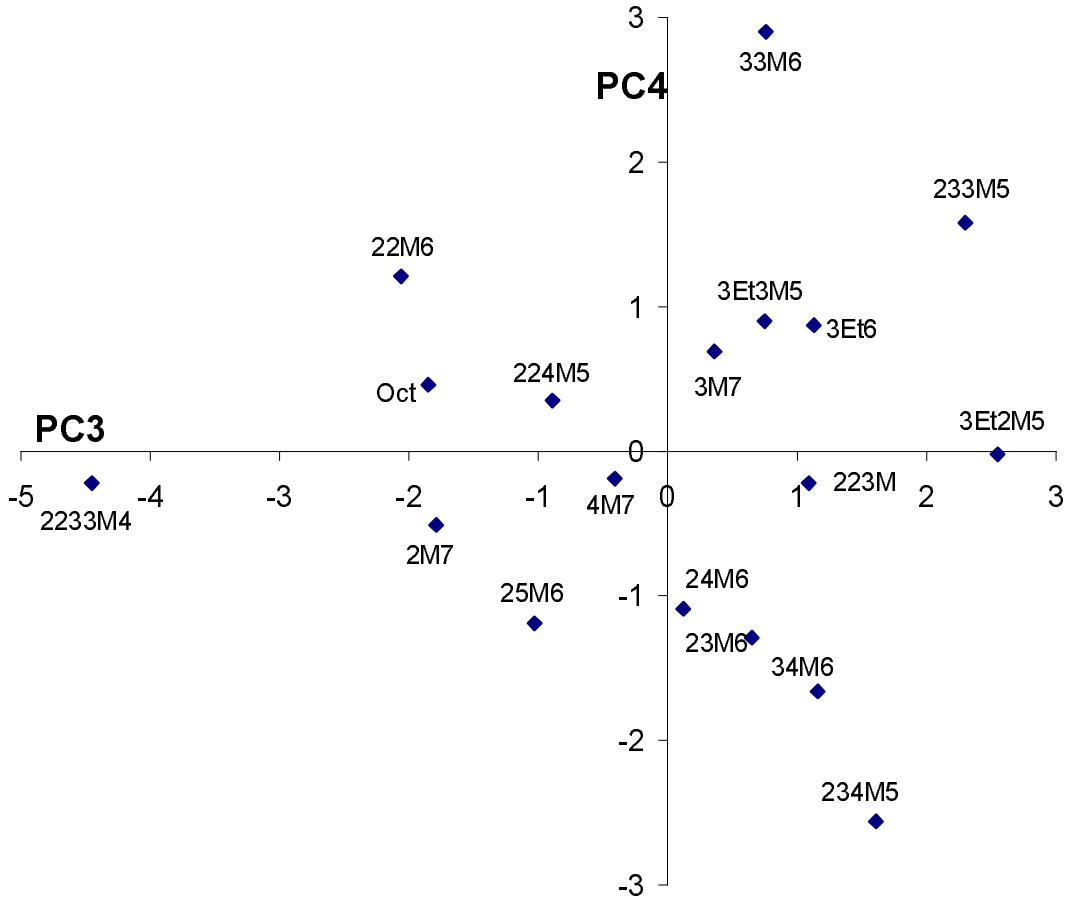

Fig. 4c. The score plot of the octanes in the plane of the third and the fourth principal component.

If the properties known for all octanes but 2,2,3,3-tetramethylbutane are tested (because the data for MON and ΔGfºg were not available for 2,2,3,3-tetramethyl butane; results are not shown because of similarity with Fig. 4), the axes PC1, PC2, PC3, and PC4 explain 54%, 30%, 7%, and 5% of variance, respectively. This group of data is important to test MON and ΔGf°g. The value of MON increases with branching, whereas the value of ΔGf°g either increases or decreases. The influence of the axes PC1 and PC2 remains largely unchanged. The axes PC3 and PC4 do not separate well most of the properties except dc, Vc, and Zc. This means that most of information about the properties except dc, Vc, and Zc is contained on the first and second axis.

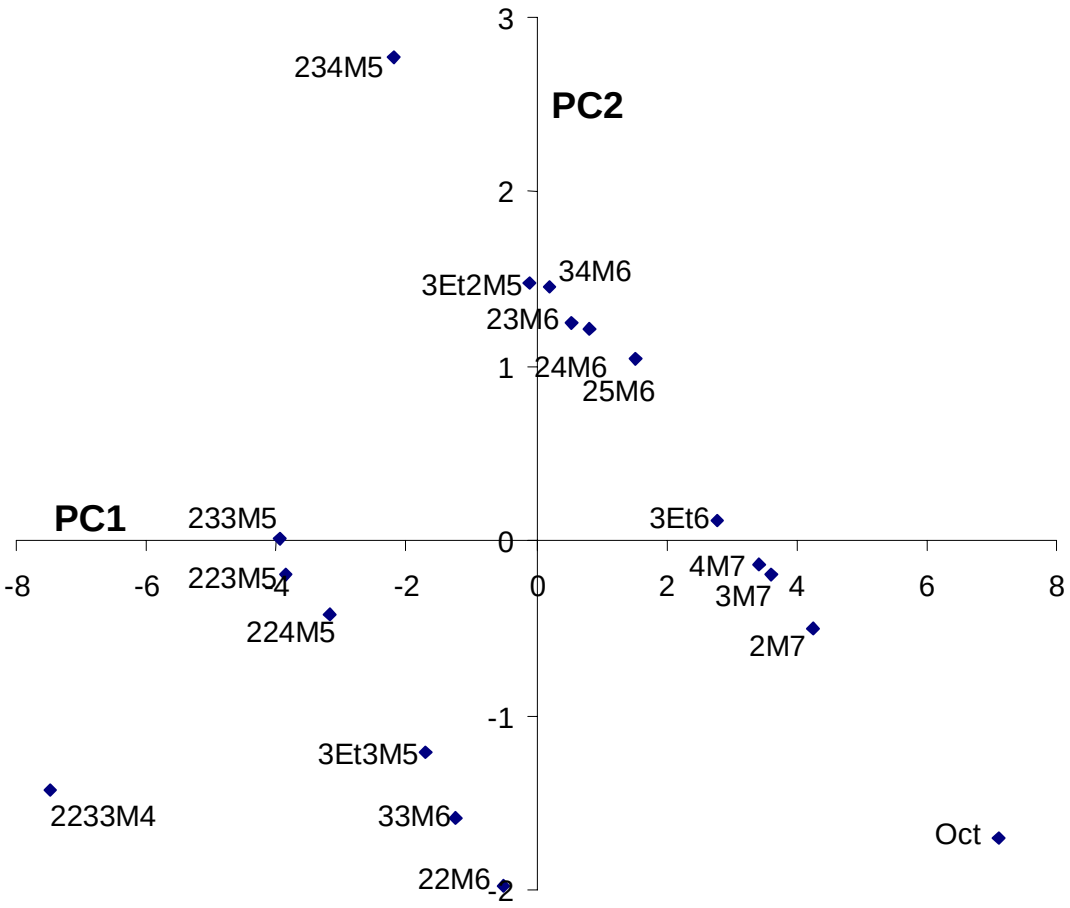

Fig. 5. The score plot of the octanes in the plane of the first two principal components.

On the score plot of Fig. 5 the octanes are grouped under the influence of indices. This score plot is quite different from that in Fig. 4b, where the influence of tested properties is presented. In Fig. 5, octanes form three groups. Two of them are shaped as nearly straight lines. One of them contains octanes possessing no quaternary carbon and it is divided into four subgroups consisting of *n*-octane, all mono-substituted heptanes and hexanes, all i,j-disubstituted hexanes and pentanes, and the only i,j,k-trisubstituted pentane. The second group contains octanes possessing one quaternary carbon. It is divided into two subgroups consisting of the i,i-substituted hexanes and pentanes on the one hand and the i,i,j-substituted pentanes on the other hand. The third cluster contains the only octane containing two quaternary carbons. The separation rules are as follows. The axis PC1 separates octanes first of all by the number of branches. Some separation by the type of branched structure (i.e. whether the carbons are tertiary or quaternary) and the adjacency of branches can be seen, too. The axis PC2 separates octanes containing tertiary carbons from those containing quaternary ones. The axis PC3 separates them by the position of branches: peripheral < central as well as distant < adjacent. The axis PC4 separates the symmetric from the asymmetric ones.

On the loading plot (not shown) corresponding to the score plot, Fig. 5, the axis PC1 groups the indices into two clusters according to their dependence on branching. One cluster is formed by those that increase with branching, i.e. EA, J, and λ_1. The other cluster contains the indices that decrease with branching, i.e. Z, χ, MTI, W, D, Xu, $\lambda\lambda_1$, and ID. The axis PC2 separates in the former cluster the index EA from λ_1 and J, and the latter cluster into three subclusters. The first subcluster contains Z and χ, the second one contains $\lambda\lambda_1$ and ID, and the third one contains W, D, MTI, and Xu. The axes PC3 and PC4 contribute additional separation of EA, λ_1, and J, as well as Z from χ and both from the others, ID from the others, whereas MTI, W, D on the one hand, and Xu and $\lambda\lambda_1$ on the other hand, form two distinct subclusters.

Discussion

Table 2 presents the summary of the composition of the data sets as well as of the percentage of variance explained by particular PCi axes.

Table 2. The information content (% of variance) of PCi axes on analysis of the data sets used to derive the figures.

Fig.	No. of data in the set				Part of variance explained by the PCi axis (%)			
No.	Alkanes	Indices	Propert.	Markers	PC1	PC2	PC3	PC4
1	36	11	21	7	74	15	5	2
2	40	0	18	6	78	10	6	3
3	39	11	0	6	68	20	9	2
4	18	0	22	4	57	20	12	7
5	18	11	0	4	83	12	5	

When all alkanes taken into consideration are studied regarding their physicochemical properties, Fig. 2, then about three quarters of variance is explained by the axis PC1, which separates alkanes mainly by their carbon number, i.e. the number of vertices in their graphs. A similar result is obtained regarding the considered topological indices, Fig. 3. In Fig. 2, studying the influence of properties, the axis PC2 explains 10% of variance. This axis separates the alkanes by the number of branches, with some influence of the type of branches. Studying the influence of indices, Fig. 3, the axis PC2 explains 20% of variance, i.e. twice more than if the properties are considered. The indices seem to give rise to less information about the type of branches than the properties do. This difference, as well as the difference in the amount of the explained variance, indicates that the properties and the indices, considered as two groups of data, are not entirely equivalent.

The axis PC3 explains 6% of variance in Fig. 2 and 9% in Fig. 3. In both cases the alkanes are separated by the type and frequency of occurrence of the branched structure, i.e. those having the structure containing tertiary carbons are separated from those containing the secondary or quaternary carbons. The information content of the axis PC4 is small. In Fig. 2 it explains 3% of variance and in Fig. 3 about 2%. This difference is reflected also in the type of information. Whereas the properties bear information about the shape and symmetry of molecules, as well as of the adjacency of branches, the indices present hardly any information on shape and symmetry.

If only data of octanes are considered, i.e. if the influence of carbon number (and molar mass) presenting the main part of information on the axis PC1 of Figs. 1 - 3 is excluded, then the main part of information that could be derived from Fig. 4b and 4c (influence of properties) is:

- On the axis PC1 (57% of variance) about the number of branches,
- On the axis PC2 (20% of variance) whether the branches are adjacent or distant,
- On the axis PC3 (12% of variance) it could not be clearly recognised,
- The axis PC4 (7% of variance) separates mainly molecules containing quaternary carbons from those containing tertiary ones.

The information derived from Fig. 5 (influence of indices) is much more straightforward:

- On the axis PC1 (83% of variance) about the number of branches > tertiary vs. quaternary structure > adjacency,
- On the axis PC2 (12% of variance) mainly whether the branched structure is tertiary or quaternary,
- On the axis PC3 (5% of variance) whether the branches are central or peripheral,
- On the axis PC4 (<0.5% of variance!) whether the molecule is symmetric or not.

Thus, regarding the branching of alkanes, the information content of the properties and the indices, considered as two groups of data, is not entirely equivalent. In spite of that, several conclusions can be drawn from the above analysis when all alkanes from methane through octanes are considered, Figs 1 - 3:

- The branching of alkanes is not directly and unequivocally reflected in their properties.
- The major influence (around 75% of variance) on tested properties has the molar weight (carbon number, number of vertices) and much less the branching, as indicated earlier by Kirby [22].
- Next to carbon number, the number of branches is important (10 - 20% of variance).
- Next to the number of branches, the structures having tertiary and quaternary carbons are distinguished (6 - 9% of variance).

– The least influence has the position of branches; the properties separate alkanes also regarding their shape and symmetry (3% of variance), whereas the indices do it to a lesser extent (2% of variance).

Figs. 4 and 5 indicate that the tested indices disperse octanes in a different way than the properties. The contribution of structure details does not seem to be as clear as when all alkanes are considered. These facts stimulate consideration of the reasons for the observed differences.

Motor octane number

When all available data of MON among alkanes up to octanes are considered, the highest (and negative) correlation of MON is observed with the number of CH_2 groups ($r_{MON,Ns,all} = -0.86$). Since other correlation coefficients are $r_{MON,Np,all}$ (0.55) > $r_{MON,Nq,all}$ (0.39) > $r_{MON,Nt,all}$ (0.18), the tertiary carbons seem to be of low importance, but the primary carbons attached to quaternary ones may have some additional influence. If only octanes are considered, the correlation of MON with the number of CH_2 groups is slightly higher, $r_{MON,Ns,octanes} = -0.88$, but that with the number of CH_3 groups is distinctly higher than previously: $r_{MON,Np,all} = 0.55$ and $r_{MON,Np,octanes} = 0.93$. The correlation with the number of tertiary carbons, $r_{MON,Nt,all} = 0.18$ and $r_{MON,Nt,octanes} = 0.30$ as well as with that of quaternary carbons, $r_{MON,Nq,all} = 0.39$ and $r_{MON,Nq,octanes} = 0.54$ also indicates some increase, but these types of functional groups have obviously less influence on MON, especially the tertiary ones. Thus, both groups of data indicate that MON is influenced first of all by conversion of secondary carbons into primary ones. Conversion of one secondary carbon into a primary one in an alkane causes a decrease of the number of secondary carbons by one (on formation of a quaternary structure) or two (on formation of a tertiary structure). In this respect, the influence of tertiary carbons on MON would be expected to be higher than it is observed.

The reason for the low influence of tertiary carbons might be in the reactivity of alkanes. During the alkanes' combustion, the alkanes containing tertiary carbons convert to tertiary carbon-centred radicals. These radicals are quite stable. Their structure causes some steric hindrance, which decreases the reactivity and consequently increases the

MON. Therefore, the increase of MON with branching has to be ascribed not only to the decrease of the number of secondary carbons, but also to other phenomena, e.g. the steric hindrance of the remaining secondary carbons by primary carbons attached to the quaternary ones and not to branching as such. Accordingly, motor octane number is not the best possible measure of branching.

Intrinsic and interaction-dependent properties of alkanes

The values of some physicochemical properties of alkanes increase, but the majority of them decrease with branching. This fact, as well as the fact that physicochemical properties of alkanes correlate well with one or another topological index but only few of the known topological indices correlate best with more than one property [10] as well as the fact that most of them correlate very well with carbon number, raises several questions. The first question is whether the properties are directly dependent on branching or not. To answer this question, one has to consider whether a physicochemical property is an intrinsic property of a molecule itself or a consequence of interaction between molecules. Intrinsic properties are e.g. Mw, V_i, ω, and $\Delta Hf°g$, whereas properties dependent on interactions between molecules are e.g. BP, d, Vm, Tc, Pc, dc, Vc, Zc, αc, a0, b0, ΔHv, A, B, and C. Hosoya et al. [23] divide these properties into dynamic (ΔS, BP, $\Delta Hf°g$), static (d, nD, Vm, MR, Pc), and dynamic + static (Tc, dc, Vc) ones. We consider their division to be less fundamental.

Among the intrinsic properties, Mw does not change between isomers; V_i varies by $\leq 0.05\%$, and both of them vary with carbon number. $\Delta Hf°g$ and ω, on the other hand, vary with carbon number, as well as between isomers. Thus, since Mw is not dependent on branching and V_i can also be considered independent, they cannot be used to indicate branching. Of the intrinsic properties considered here, $\Delta Hf°g$ and ω may be useful.

To understand the influence of branching on properties dependent on interactions, let us look at the consequences of branching. Branching influences the ability of molecules for interaction in several ways. The type and number of functional groups that are exposed to intimate interaction with functional groups of other molecules is changed on branching. Different functional groups have different contributions to intermolecular

attraction. In alkanes, the contribution to intermolecular attraction at the equilibrium distance is CH_3 < CH_2 < CH << C [24]. On branching, the number of CH_2 groups is decreased and the number of CH_3 groups is increased. The latter are placed at the surface of the molecule. Therefore, the contribution to intermolecular attraction is decreased. The direct consequence of branching is thus decrease in intermolecular attraction, causing lower BP, Tc, the need for higher Pc, etc. On the other hand, the tertiary carbons and especially the quaternary carbons are buried in the interior of the molecules, so their interaction distances are greater than those of groups at the surface of molecules. Consequently, in spite of their greater possible contribution to intermolecular attraction at the equilibrium distance, their effective contribution to attraction can be at most 10% of that at equilibrium distance.

Simultaneously with the change in type, number and position of functional groups involved in intermolecular attraction, branching influences also the shape of molecules. The shape influences the packing. Thus, the change in packing is an indirect consequence of branching. The packing influences the distances between the functional groups of adjacent molecules. Because of the high short-range repulsion, these distances are usually not shorter than the equilibrium ones, but may be (and usually at least some of them are) longer. When the intermolecular distance (d_i) is greater than the equilibrium distance (d_e), the intermolecular attraction depends on it by $(d_i/d_e)^{-6}$. Consequently, a small increase in the intermolecular distance caused by looser packing of molecules due to the change in their shape can appreciably decrease the effective intermolecular attraction. Better packing, on the other hand, decreases the intermolecular distances and as a result, the intermolecular attraction increases, and vice versa. To sum up, on increasing branching fewer functional groups are involved in intimate intermolecular attraction, they individually contribute less to the intermolecular attraction, and their effective distances vary with the ability of molecules to pack effectively, as well as with the effective attraction forces. Consequently, branching influences the physicochemical properties that are dependent on intermolecular interactions in an indirect, complex way that makes them less suitable as criteria to assess the branching indices.

Among the properties dependent on interaction between molecules, the influence of branching on density is the easiest to comprehend. Looking at the molecular level, the density is the ratio of the mass that is contained in the molecule and of the volume that is the sum of the intrinsic volume of the molecule and of the corresponding part of the "free" space between the molecules. Since molar mass does not depend on branching and intrinsic volume can be considered to be independent of branching, too, it is predominantly the "free" space between molecules that depends on branching. On the one hand, on increasing branching the mutual attraction between the molecules becomes lower and the "free" space increases to some extent, giving rise to lower density. On the other hand, branching influences the shape of molecules, the shape influences their packing and, due to worse or better packing, the density decreases or increases. In the case of density, this latter influence seems to be more important than the decrease of intermolecular attraction and therefore the density in some cases decreases and in others increases with branching, the latter cases prevailing. Also interesting is the series of increasing "free" space among octanes which is the same as that for Vm and the reverse of that of density: 2233M4 < 3Et3M5 < 233M5 < 34M6 < 3Et2M5 < 234M5 < 223M5 < 3Et6 < 23M6 < 33M6 < 3M7 < 4M7 < 8 < 24M6 < 2M7 < 22M6 < 25M6 < 224M5. This series is quite different from that of known melting points, although some segments of the series retain their order: 2233M4 >> 8 > 3Et3M5 ~ 25M6 > 233M5 > 224M5 > 2M7 ~ 234M5 > 223M5 > 3Et2M5 > 3M7 ~ 4M7 ~ 22M6 > 33M6. It seems as if the criteria for good packing were different for the liquid and the solid state because of differences in the mobility of molecules.

A similar case is the boiling point, the property known for the greatest number of alkanes and very often used to assess the suitability of topological indices. The boiling point is the temperature at which the molecules have their thermal energy equal to the sum of energy due to external pressure and that of intermolecular attraction. If intermolecular attraction decreases due to increased branching, a lower temperature is needed to satisfy the condition presented above. If due to branching the packing of molecules becomes looser, greater intermolecular distances cause an additional decrease of attraction and again a lower temperature is needed to satisfy the condition presented

above. As a rule, however, the decrease of attraction due to changes of the type of interacting groups does not go parallel to the decrease of interaction due to increased intermolecular distances. If due to an increase of branching the packing becomes denser, then on the one hand, the intermolecular attraction decreases due to lower ability for interaction contributed by the greater number of CH_3 groups; on the other hand, it increases due to lower distances caused by closer packing and the needed decrease in temperature is lower. If we compare, for example, *n*-octane, 2,2-dimethylhexane, and 2,2,3,3-tetramethylbutane, the number of less interacting functional groups at the surface of the molecule increases in this series. The number of functional groups that can come into close contact decreases in this series of increasing branching. Both of these consequences give rise to lower and lower attraction in this series and hence to lower BP. But the packing, as deduced from Vm, is the best in the case of 2,2,3,3-tetramethylbutane and the worst in the case of 2,2-dimethylhexane. It does not follow the former series and it introduces some disorder in the extent of attraction causing BP to be not 8 > 22M6 > 2233M4 but 8 > 22M6 ~ 2233M4. This fact is reflected much more clearly in their melting points than in their boiling points, since the melting points are much more affected by the packing than the boiling points. Thus, the boiling point and other physicochemical properties dependent on intermolecular interaction are influenced by branching in a too complex way to depend on branching in a simple and straightforward manner that would be desired if they should serve as reference properties. In spite of that, the decrease of intermolecular attraction on branching explains the decrease of the values of these properties on branching. Consequently, only some of the intrinsic properties, such as $\Delta Hf°g$ might be used as reference properties, whereas the properties dependent on intermolecular interaction such as the boiling point, critical data, etc., can only be of secondary use.

Definition of branching

Intuitive branching rules have been known for decades [8]. When we use them and consider the data of topological indices, one similarity seems striking. All of indices considered here, except the Hosoya index Z, which has the value of 1 for methane by definition, and Xu that is log 0 by definition, set the value 0 to methane. The fact that the value set to methane is 0, indicates that methane should be considered nonbranched. If so, then the definition of branching is straightforward: "*Methane is nonbranched. Each departure from its structure is branching. Each type of departure has its own contribution to the extent of branching.*" Or: "*Methane is not branched. Replacement of any H atom by another atom is considered as an increase of branching. … .*" (Replacement by isotopes is not considered in this paper). According to this *Methane-based* definition of branching, the value of a branching index must increase with increasing carbon number at the same type of branching. The analysis of lumped data of physicochemical properties presented above enables us to make a rough estimation of the information contribution to the values of indices obeying the Methane-based definition. The number of vertices should contribute around 75% of information content of a branching index, the number of branches should contribute around 10%, the type of branches around 5%, while the contribution of the information about the shape and symmetry of molecules, as far as it is possible, as well as of the adjacency of branches should contribute a small percentage of information.

The Methane-based definition of branching is mathematically correct. But it is not in line with the general chemical sense about branching. According to Abraham et al. [25], the results of application of equations based on some chemical model must make and maintain general chemical sense. According to the general chemical sense the *n*-alkanes are not branched, whereas by the Methane-based definition of branching the higher the *n*-alkane the more it is branched. Therefore, an additional definition of branching not insulting the general chemical sense has been sought and found.

The *n-Alkane-based* definition of branching is: "*n-Alkanes are not branched. Replacement in them of any H atom except those placed on the peripheral carbon by another atom or group causes branching.*" or, "*n-Alkanes are not branched. Any departure from the n-alkane structure is defined as branching. Each type of departure*

has its own contribution to the extent of branching." This definition maintains the general chemical sense.

The relation between these definitions of branching is as follows. The *Methane-based* definition of branching is an absolute, general definition, whereas the *n-Alkane-based* definition of branching is a special definition that must be based on the absolute one, i.e. on the Methane-based definition of branching, and they should be used accordingly.

No one of the tested properties and indices is consistent with the *n*-Alkane-based definition of branching. For properties this is natural since they depend first of all on the number and type of atoms they are composed of, as well as on the molar mass and the size of molecules. The tested indices, on the other hand, were developed to fit the properties, not this definition. It could reasonably be expected that the properties as well as the indices are consistent with the Methane-based definition. In fact, only the intrinsic property $\Delta Hf°g$ and the indices λ_1, J, and EA seem to be consistent with this definition. On the other hand, the properties and indices increasing with carbon number and decreasing with branching do not follow the Methane-based definition for the reasons explained in the previous section. These indices cannot be considered as branching indices but as indices of interaction-dependent properties.

Data and method

The indices

We decided to take into account the group of the most frequently used indices and some novel indices. Altogether eleven indices are used. The data for Wiener index, W, the Hosoya index, Z, the Randić index χ, the Balaban index, J, the Yang-Xu-Hu index EA_{max} (denoted in present paper as EA) were taken from Yang et al. [11]. The data for the Randić ID number were taken from [12] and the data for the Schulz MTI number were taken from [13]. The data for the Xu index were taken from Ren [14]. The following indices, the Randić index $\lambda\lambda_1$ [1], the largest eigenvalue of the distance matrix (D), and the largest eigenvalue of the adjacency matrix (λ_1) were calculated from the corresponding matrices.

The alkanes' properties

In this work 24 properties are taken into account. The data for Motor Octane Number (MON) of alkanes were taken from Ren [14] and Pogliani [15], those of boiling point (BP), entropy (S), and quadratic mean radius (R^2) were taken from Ren [14]. The data of melting point (MP), density (d), the critical data Tc, Pc, Vc, Zc, αc, and dc, as well as the standard enthalpy of formation for the ideal gas (ΔHf°g), the standard Gibbs energy of formation for the ideal gas (ΔGf°g), the enthalpy of vaporisation (ΔHv), the Antoine constants A, B, and C as well as the Pitzer's acentric factor (ω) and the refractive index (n_D) were taken from CRC Handbook [16] or from Lange's Handbook [17]. The data for the liquid molar volume (Vm), the intrinsic molar volume (V_i), the van der Waals parameters a_0 and b_0, and molar refraction (MR) were calculated from data presented in those handbooks.

Other variables

For an easier identification of the features, which have the largest impact on indices and properties, some additional variables were included in the data sets. These variables are the molar weight (Mw), the carbon number (N_C), the number of primary carbons (*Np*), the number of secondary carbons (*Ns*), the number of adjacent secondary carbons (*Nss*), the number of tertiary carbons (*Nt*), and the number of quaternary carbons (*Nq*) in the structure of alkanes. They contain the information about molecular topology, which is already contained in properties and indices. For this reason, no additional information is introduced in the data sets with these variables. Indices, for example, which mostly depend on the size of the molecules, will group with Mw and N_C. These variables therefore only serve as a kind of markers and will be referred to as markers in the following text.

The Principal Component Analysis

The Principal Component Analysis (PCA) was performed as described in [18-21]. Each principal component PC_i is a new co-ordinate expressed as a linear combination of the old features x_j: $PC_i = \sum_j b_{ij} x_j$. The old features, x_j, are in our case indices, properties and other variables (markers), mentioned above. The new co-ordinates PC_is are called scores, while coefficients b_{ij} are called loadings. The scores (new co-ordinates or PC_is) are ordered according to their information content (variance content) with respect to the total variance among all objects. The score-score plots show the positions of compounds (in our case the alkanes) in the new co-ordinate system, while loading-loading plots show the position of features that represent compounds (in our case the indices, the properties, and the markers) in the new co-ordinate system.

Typically, most of the information contained in the data set is explained by the first few principal components. This is the case when the variables correlate with each other, as it is the situation in the present work. The concentration of information in the space of only the first few principal components leads to the reduction of the information space, i.e., with principal components the relevant part of the information is presented with a smaller number of variables than before. In the reduced information space it is easier to interpret the information contained in the data set.

The identity of alkanes is presented in shorthand. M, Et, Pr, Bu, Pe, Hx, Hp, and Oct are *n*-alkanes from methane (M) to *n*-octane (Oct). For other alkanes the following system is used, illustrated here with 2-methylheptane (2M7), 3-ethyl hexane (3Et6), 2,2,3-trimethylpentane (223M5), and 3-ethyl-2-methylpentane (3Et2M5) as examples.

References

1. M. Randić, *Acta Chim. Slov*. **1997**, *44*, 57-77.
2. D.H. Rouvray, *J. Comput. Chem.* **1987**, *8*, 470-480.
3. S. Mendiratta, A.K. Madan, *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 867-871.
4. H. Wiener, *J. Am. Chem. Soc*. **1947**, *69*, 17-20.
5. H. Hosoya, *Bull. Chem. Soc. Japan* **1971**, *44*, 2332-2339.
6. M. Randić, *J. Am. Chem. Soc*. **1975**, *97*, 6609-6615.
7. A.T. Balaban, *Chem. Phys. Lett*. **1982**, *89*, 399-404.
8. D. Bonchev, N. Trinajstić, *J. Chem. Phys*. **1977**, *67*, 4517-4533.
9. L. Lovasz, J. Pelikan, *Period. Math. Hung*. **1973**, *3*, 175-182.
10. M. Randić, S.C. Basak, *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 261-266.

11. Y.-Q. Yang, L. Xu, C.-Y. Hu, *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 1140-1145.
12. M. Randić, *J. Chem. Inf. Comput. Sci*. **1984**, *24*, 164-175.
13. Z. Mihalić, S. Nikolić, N. Trinajstić, *J. Chem. Inf. Comput. Sci*. **1992**, *32*, 28-37.
14. B. Ren, *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 139-143.
15. L. Pogliani, *J. Phys. Chem*. **1995**, *99*, 925-937.
16. D.R. Lide, CRC Handbook of Chemistry and Physics, 76th Ed., CRC Press, Boca Raton 1995-1996.
17. J.A. Dean, Lange's Handbook of Chemistry. McGraw-Hill, New York 1985.
18. D.L.Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte, L. Kaufman, *Chemometrics: a textbook*. Elsevier, Amsterdam, 1988.
19. R.C. Graham, *Data Analysis for the Chemical Sciences*. VCH, Weinheim,1993, pp. 329-343.
20. R. G. Brereton, *Chemometrics: Applications of Mathematics and Statistics to Laboratory Systems*. Ellis Horwood, New York, 1990.
21. S. Wold, K. Esbensen, P. Geladi, *Chemometr. Intell. Lab. Syst*. **1987**, *2*, 37-52.
22. E.C. Kirby, *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 1030-1035.
23. H. Hosoya, M. Gotoh, M. Murakami, S. Ikeda, *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 192-196.
24. F.M. Fowkes, in R.L. Patrick, *Treatise on Adhesion and Adhesives. Vol. 1: Theory*, M. Dekker, New York, 1967, pp 325-449.
25. M.H. Abraham, P.L. Grellier, J.-L.M. Abboud, R.M. Doherty, R.W. Taft, *Can. J. Chem*.**1988**, *66*, 2673-2686.

Povzetek

Ustreznost topoloških indeksov J, W, Z, D, MTI, Xu, ID, χ, $\lambda\lambda_1$, EA_{max} in λ_1 kot indeksov razvejanosti ter fizikokemijskih lastnosti MON, BP, d, Vi, Vm, Vc, Tc, Pc, dc, Zc, αc, ΔHv, A, B, C, n_D, MR, a_0, b_0, ΔHf°g, ΔGf°g, S, R^2 in ω kot referenčnih lastnosti zanje sva ugotavljala pri alkanih s pomočjo metode glavnih osi (PCA). Na PCA diagramih so alkani lo čeni po številnih kriterijih. Najpo membnejše je število o gljiko v o z. mo lska masa, sledijo ji število vej, razlo čevanje terciarnih o gljiko v o d kvarternih, po lo žaj in medsebo jna lega vej, o blika in simetrija mo lekul. Večina lastno sti in indekso v viso ko korelira s številom ogljikov, vpliv razvejanja pa je manjši. Oktansko število ogljikovodikov je odvisno predvsem od števila CH_2 skupin. Lastnosti deliva na notranje, to je lastnosti mo lekul samih, in na tiste, ki so o dvisne o d medsebojnega vpliva mo lekul. Razlo ženo je, zakaj slednje niso primerne ko t o sno vne referenčne vredno sti. Po dajava dve definiciji razvejanosti, metansko kot splošno ter *n*-alkansko kot posebno, bolj sprejemljivo za kemike.