J Math Chem (2008) 43:1560–1572
DOI 10.1007/s10910-007-9284-0

ORIGINAL PAPER

Statistical spectroscopy as a tool for the study of molecular similarity

Dorota Bielińska-Wąż · Wiesław Nowak · Łukasz Pepłowski · Piotr Wąż · Subhash C. Basak · Ramanathan Natarajan

Received: 15 March 2007 / Accepted: 24 May 2007 / Published online: 3 September 2007

Abstract This paper aims at an application of the statistical theory of spectra to the classification of chemical compounds. It has been shown that the moments of the intensity distributions may be used as molecular descriptors. The new descriptors have been tested using spectra of nitriles and amides. The dependence of the accuracy of the classification on the number of moments (up to the twelveth order) is discussed using model spectra.

Keywords Statistical spectroscopy · Method of moments · Molecular similarity · genetic algorithm · DFT methods

1 Introduction

The degree of similarity of two objects depends on the kind of features considered in the comparison of these objects. Two molecules that are very similar with respect to some structural aspects need not be similar with respect to their fitting to a biological receptor. In the case of enantiomers, the two structures are very similar in several aspects but are entirely opposite with respect to their interaction with the plane polarized light.

D. Bielińska-Wąż (✉) · W. Nowak · L. Pepłowski
Instytut Fizyki, Uniwersytet Mikołaja Kopernika, Grudziądzka 5, Toruń 87-100, Poland
e-mail: dsnake@phys.uni.torun.pl

P. Wąż
Centrum Astronomii, Uniwersytet Mikołaja Kopernika, Gagarina 11, Toruń 87-100, Poland

S. C. Basak · R. Natarajan
Natural Resources Research Institute, 5013 Miller Trunk Highway, Duluth, MN 55811-1442, USA

The concepts of molecular similarity, derived from the structural features of the molecules, proved to be useful in the selection of the structural analogs for the drug design and for the prediction of toxicity [1,2]. The estimates of the chemical activity or toxicity based on the molecular similarity has been derived from the fundamental principle that chemicals with similar structures have similar properties. Calculated molecular descriptors, the presence of specific pairs of atoms and even physicochemical properties can be used for quantitative molecular similarity analysis [3,4]. The degree of molecular similarity depends on the choice of the set of the structural descriptors, and on the selection of a particular mathematical function used to quantify similarity from the chosen set of descriptors [5,6]. Basak et al. [7] preferred to use a tailored similarity space. The construction of the space depends on the property under consideration. The molecular wave function (or the molecular density function) contains a complete information about the molecular structure and properties. Therefore, it may be used as a basis for defining quantum similarity measures [8–10]. Another way to determine the similarity of molecules is by the comparison of their spectra. Molecular spectra encode more information than simple descriptors derived from the molecular topology because the position and the shape of the spectral lines are influenced by intra and inter molecular interactions. Therefore we advocate using molecular spectra to derive similarity among molecules. In the earlier papers [11] we used the spectra to generate a new type of descriptors to characterize the electronic states of molecules. In particular, using these descriptors examples of dissimilarity maps have been obtained [12]. It is worth to notice that the method is very general—the calculation of the new descriptors can be performed for all kinds of spectra.

The basic idea of using the new kind of descriptors (statistical moments of intensity distribution) is to treat the spectra as statistical distributions. This idea comes from the statistical theory of spectra and has already been applied in many areas of physics. The basic concepts of this theory originated in the thirties [13]. For many years statistical spectroscopy was mainly used in the nuclear physics, where not exactly known character of the interparticle interactions motivated using the language of statistics [14]. The motivation for the introduction of the statistical description to atomic spectra was completely different. The first statistical studies of atomic spectra were performed by Rosenzweig and Porter [15]. The authors tried to create a global description of the detailed features of the spectra. They studied the "repulsion of the energy levels" in complex atomic spectra. Since then, the methods of statistical spectroscopy were applied in atomic and molecular physics in order to avoid detailed calculations, to reduce the computing time, or in order to notice some new global features of the systems. Let us just mention methods of determining envelopes of the molecular electronic bands [16], or statistical studies on properties of spectra of the Heisenberg Hamiltonian [17]. In all these considerations the basic quantities which have been calculated are moments of different types of distributions. Therefore the application of the statistical moments in the theory of similarity seems to be very attractive.

The new descriptors are related to the shapes of the spectra. It is assumed that the degree of similarity of molecules is determined by the degree of similarity of the shapes of their spectra. The aim of this paper is to give another example showing that this assumption is correct. We collect a'priori two groups of molecules (nitriles and amides). The next step is the calculation of their IR spectra and then the calculation

of the moments of these distributions. The separation of the moments calculated for the nitriles from these for the amides gives us a possibility for clustering other groups of compounds without knowing their properties a'priori. The presented treatment is a test of the reliability of the classification based on the statistical moments of the intensity distributions.

In general, the process of comparison is very complex and its final result strongly depends on the number of aspects considered. In this paper we try to extract all the properties hidden in the shapes of the spectra that should be considered in the comparison. On the other hand, we have to exclude from the consideration these properties which are correlated. For this purpose a minimal basis (linearly independent) of statistical moments of intensity distributions has been extracted. This basis appears to be sufficient for the identification of the spectra and, as a result, of the corresponding chemical compounds. The studies on the selection of the minimal basis of the moments have been performed using model spectra taken as a sum of two Gaussian functions.

2 Theory

In this section, the properties of the new descriptors are studied using model spectra.

The new descriptors (statistical moments) correspond to the shapes of the intensity distributions. The basic idea stems from the observation that similar spectra have similar distribution moments. Similar systems correspond to intensity distributions and consequently, to similar moments. However, sometimes it may happen that different systems have similar particular moments. In these cases we expect that a larger set of moments should be taken into account in order to distinguish spectra. We study this problem using an infinite set of model spectra taken (as in Refs. [12], [18]) as linear combinations of two Gaussian distributions centered at ϵ_i with dispersions σ_i, defined by the parameters $c_i = 1/(2\sigma_i^2)$, $i = 1, 2$:

$$I^{\beta}(E) = N^{\beta}\left[a_1 \exp\left[-c_1(E-\epsilon_1)^2\right] + a_2 \exp\left[-c_2(E-\epsilon_2)^2\right]\right], \qquad (1)$$

where $\beta = \{c_1, a_1, \epsilon_1, c_2, a_2, \epsilon_2\}$ and E is the energy. The particular parameters characterize the width (c_i), the amplitude (a_i) and the locations of the maxima (ϵ_i) of the i-th Gaussian component $a_i \exp\left[-c_i(E-\epsilon_i)^2\right]$ of $I^{\beta}(E)$, where $i = 1, 2$. The normalization constant N^{β} is determined so that the zeroth moment of the distribution $I^{\beta}(E)$ is equal to 1. The analytical expressions for the moments of $I^{\beta}(E)$ as functions of parameters c_i, a_i, ϵ_i and also the definitions of the n-th moment M_n, of the centered n-th moment M'_n and of the scaled one M''_n, both for continous and for discrete spectra are presented in our previous paper [19]. Usually, in the statistical theory of spectra, only four lowest moments that have direct connection to the shapes of the spectra, are used. The first moment M_1 describes the mean value of the distribution, the second centered moment M'_2 is the variance, the third scaled moment M''_3 describes the asymmetry, and the fourth scaled moment M''_4 is the excess of the distribution. In this paper we consider twelve moments: M_1, M'_2, M''_3, M''_4 and M''_5, M''_6, ..., M''_{12}. The aim of this work is to find the minimal number of moments that carry the maximal information about the corresponding spectra.

Let us consider the spectra $I^{\beta}(E)$, where

$$\beta = \{5.0, 1.0, 1.2, 5.0 + \delta c, 1.0 + \delta a, 2.7 - \delta\epsilon\}. \tag{2}$$

The parameters δc, δa, $\delta\epsilon$ are related to the second Gaussian component $(1 + \delta a)$ $\exp[-(5+\delta c)(E-2.7+\delta\epsilon)^2]$ of $I^{\beta}(E)$ distributions. In this paper only this Gaussian component is subjected to variations and the changes are restricted by the parameter ranges:

$$\delta c \in \langle 0, 20\rangle, \delta a \in \langle 0, 10\rangle, \delta\epsilon \in \langle 0, 1\rangle. \tag{3}$$

Let us extract for the consideration the distributions with the same M_4'' excluding in this way this moment as a good descriptor. As an example we take $M_4'' = 2.390760$.

In order to get $I^{\beta}(E)$ distributions, genetic algorithms [20] have been used in this paper. These methods may be classified as computational techniques used for solving problems of optimization. Many classes of problems in physical sciences can be treated as optimization problems. The most common mathematical optimization tasks are minimization and maximization. In this work, searching for $I^{\beta}(E)$ with a condition of constant M_4'' is treated as the optimization problem. In practical calculations we are looking for the maximum of the function $1/[M_4''(\delta c, \delta a, \delta\epsilon) - 2.390760]$. Using genetic algorithm Pikaia [21], the search for parameters $\{\delta c, \delta a, \delta\epsilon\}$ has been performed within the restricted space defined in (3). With the termination condition 500 generations, and the accuracy of 10^{-8}, sets of parameters $\{\delta c, \delta a, \delta\epsilon\}$ have been obtained. The number of sets has been restricted to 24. The results are collected in Table 1.

Some representative cases are presented in Fig. 1. Labels in the Figure denote the numbers in Table 1. Their order depends on the value of δc. The first distribution, labeled by 10, characterizes the largest δc. It corresponds to the smallest width of the second Gaussian component. The last distribution, labeled by 18, characterizes the smallest δc. It corresponds to the largest width of this component. We observe that distributions with different δc, δa, $\delta\epsilon$ have the same M_4''. This phenomenon is referred to as *degeneracy*. As expected, the accuracy of distinguishing spectra strongly depends on the number of moments taken into account. In order to reduce the degeneracy, several different moments should be taken into account simultaneously in the process of comparison.

In order to exclude the moments that are correlated, Pearson's correlations coefficients between a pair of distributions x and y

$$P(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4}$$

have been calculated, where $P(x, y) \in \langle -1, 1\rangle$ and $n = 24$. In the present work our earlier studies [18] have been generalized by the enlargement of the number of moments taken into account from four to twelve. All the coefficients are collected in

Table 1 Values of the parameters corresponding to the distributions with $M_4'' = 2.390760$

No.	δc	δa	$\delta\epsilon$
1	7.450801	1.988400	0.528170
2	12.421802	2.435700	0.529050
3	13.673402	3.532001	0.117660
4	4.131801	1.719800	0.499220
5	18.205603	2.885500	0.532230
6	4.266201	1.791800	0.473660
7	17.337403	2.837900	0.527380
8	17.991003	2.842600	0.539230
9	15.145802	3.527601	0.206310
10	19.122603	3.788201	0.262540
11	15.398002	3.901101	0.002900
12	13.390002	0.557400	0.912120
13	18.726203	4.165201	0.060140
14	8.470401	2.805000	0.193390
15	18.848203	2.298700	0.676040
16	0.859800	0.249600	0.807730
17	3.008000	1.153900	0.664020
18	0.127400	0.465600	0.756890
19	0.816200	1.212100	0.560520
20	13.626002	2.759500	0.459420
21	6.981601	2.622600	0.190300
22	6.613001	0.270200	0.895760
23	10.935802	0.833000	0.856170
24	4.138201	1.288700	0.655300

the matrix P. The rows and the columns correspond to the orders of the moments, for example $P_{1\,3} \equiv P(M_1, M_3'')$.

$$P = \frac{1}{100}\begin{pmatrix}
100 & 98 & -84 & -18 & -70 & -45 & -37 & -67 & 12 & -83 & 67 & -83 \\
 & 100 & -76 & -18 & -60 & -55 & -25 & -73 & 23 & -80 & 65 & -77 \\
 & & 100 & 20 & 97 & -9 & 80 & 16 & 42 & 51 & -31 & 59 \\
 & & & 100 & 19 & 2 & 14 & 7 & 6 & 7 & -2 & 9 \\
 & & & & 100 & -30 & 92 & -6 & 62 & 33 & -12 & 44 \\
 & & & & & 100 & -61 & 96 & -87 & 67 & -68 & 54 \\
 & & & & & & 100 & -42 & 87 & -2 & 22 & 11 \\
 & & & & & & & 100 & -79 & 82 & -81 & 72 \\
 & & & & & & & & 100 & -47 & 61 & -34 \\
 & & & & & & & & & 100 & -96 & 98 \\
 & & & & & & & & & & 100 & -94 \\
 & & & & & & & & & & & 100
\end{pmatrix}$$

This matrix is symmetric and therefore only the upper triangle is here presented. The diagonal elements correspond to the correlation between the same moments and are equal to 1.

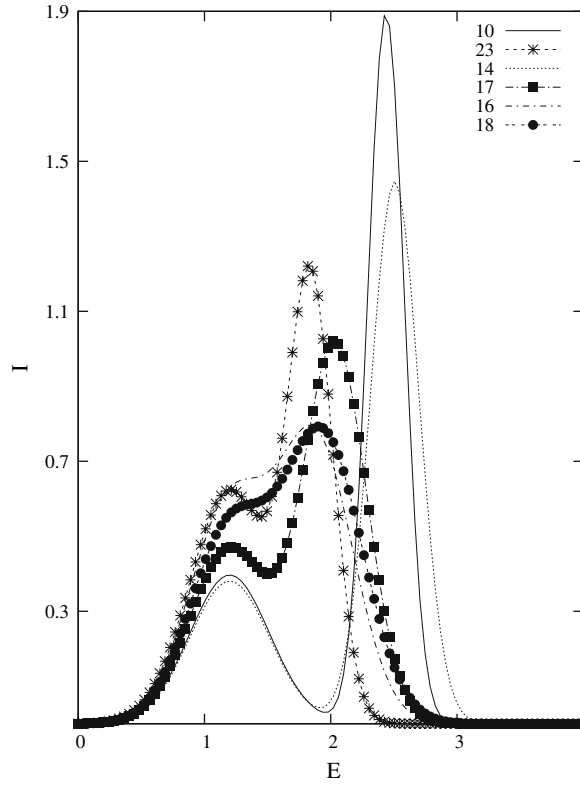

Fig. 1 Intensity distributions corresponding to $M_4'' = 2.390760$

We take into account only strong linear correlations, i.e. the cases when either $P_{xy} \in \langle 0.9, 1\rangle$ or $P_{xy} \in \langle -1, -0.9\rangle$. In the latter case we have the so called strong negative correlation, i.e. one of the quantities increases while the second one decreases.

As one can see, strong linear correlations appear between the following moments:

$M_1 \sim M_2'$ ($P_{1\,2} = 0.98$),
$M_3'' \sim M_5''$ ($P_{3\,5} = 0.97$),
$M_5'' \sim M_7''$ ($P_{5\,7} = 0.92$),
$M_6'' \sim M_8''$ ($P_{6\,8} = 0.96$),
$M_{10}'' \sim -M_{11}'' \sim M_{12}''$ ($P_{10\,11} = -0.96$, $P_{10\,12} = 0.98$, $P_{11\,12} = -0.94$),
$M_6'' \sim -M_9''$ ($P_{6\,9} = -0.87$),
$M_7'' \sim M_9''$ ($P_{7\,9} = 0.87$).

We conclude that there are four linearly independent moments, that can be chosen as M_1, M_3'', M_6'', and M_{10}''. Generally, we expect that the four lowest moments are sufficient for a complete description of the spectral similarity problems. The probability of a specific behavior of the lower order moments (as for example constant M_4'', as it is in the present case), though rather low, can not be excluded. Therefore, in some cases, the higher order moments should be considered in order to get a good classification of spectra.

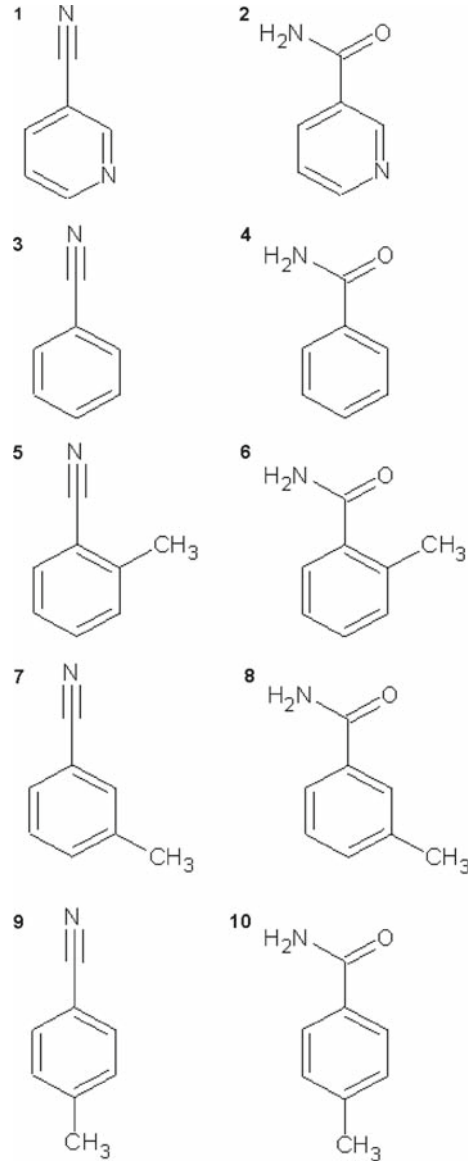

Fig. 2 Chemical compounds considered in this work (nitriles—left column, amides—right column)

3 Results and discussion

The new descriptors are tested using spectra of nitriles and amides. In Fig. 2 the chemical compounds considered in this work are collected.

The nitriles are displayed in the left column and the amides in the right one. The numbers presented in the Figure correspond to:

1. nicotinonitrile,
2. nicotinamide,
3. benzonitrile,
4. benzamide,
5. *o*-methylbenzonitrile,
6. *o*-methylbenzamide,
7. *m*-methylbenzonitrile,
8. *m*-methylbenzamide,
9. *p*-methylbenzonitrile,
10. *p*-methylbenzamide.

The spectra of all compounds have been calculated using *The Density Functional Theory method* (DFT) implemented in Gaussian 98 code [22]. The hybrid-type exchange-correlation potential by Becke, Lee, Yang and Paar (B3LYP) [23,24] was used. The B3LYP formula includes the Slater and Hartree-Fock exchange potential, Becke's gradient correction to exchange potential [25], Lee–Yang–Paar correlation potential [26] and Vosko–Wilk–Nusair correlation potential [27]. The 6-31G(d,p) basis set was used to optimize the structures and to predict the energies and IR spectra. The results have been presented in our previous paper [28]. This method has also been successfully used in many quantum-chemical studies of IR spectra [29]. The calculated spectra are presented in Fig. 3.

The numbers labeling the spectra correspond to those in Fig. 2.

The compounds studied in this work were carefully selected. They are important substrates (nitriles) and products (amides) of biotechnological enzyme nitrile hydratase (NHase). This protein is used for the kiloton-scale "green" production of amides [30]. The analysis of the IR spectra may help to elucidate the mechanism of the enzymatic activity of various variants of NHases [31,32].

Figures 4 and 5 present the first twelve moments calculated for the nitriles (crosses) and for the amides (squares).

All the moments have been calculated for all (10) chemical compounds. The numbers in the horizontal axes correspond to the numbering in Figs. 2 and 3.

We observe a clear separation of the moments for the nitriles from the ones for the amides. As one can see, the first four moments contain sufficient information necessary for the classification. For the higher order moments (from the fifth to the twelveth) we observe the same qualitative relations. Similarly as for the model case discussed in the previous section, four moments create the minimal and sufficient set of descriptors (within the 12 considered). As it is expected, there are no correlations between the lower order moments. Therefore, the minimal set of moments include M_1, M_2', M_3'', M_4''. The higher order moments do not contain any new information relevant for the classification of the considered molecules.

The first moment (smaller for the amides than for the nitriles, Fig. 4) gives the information that in the spectra of the amides more intensity is located in lower energy region than it is in the spectra of the nitriles. This global observation can be seen in Fig. 3. The observed behavior of M_1 has its origin in the substantial shift of the C−N stretching mode which is located close to $2400\,\mathrm{cm}^{-1}$ in the nitriles (triple C−N bond)

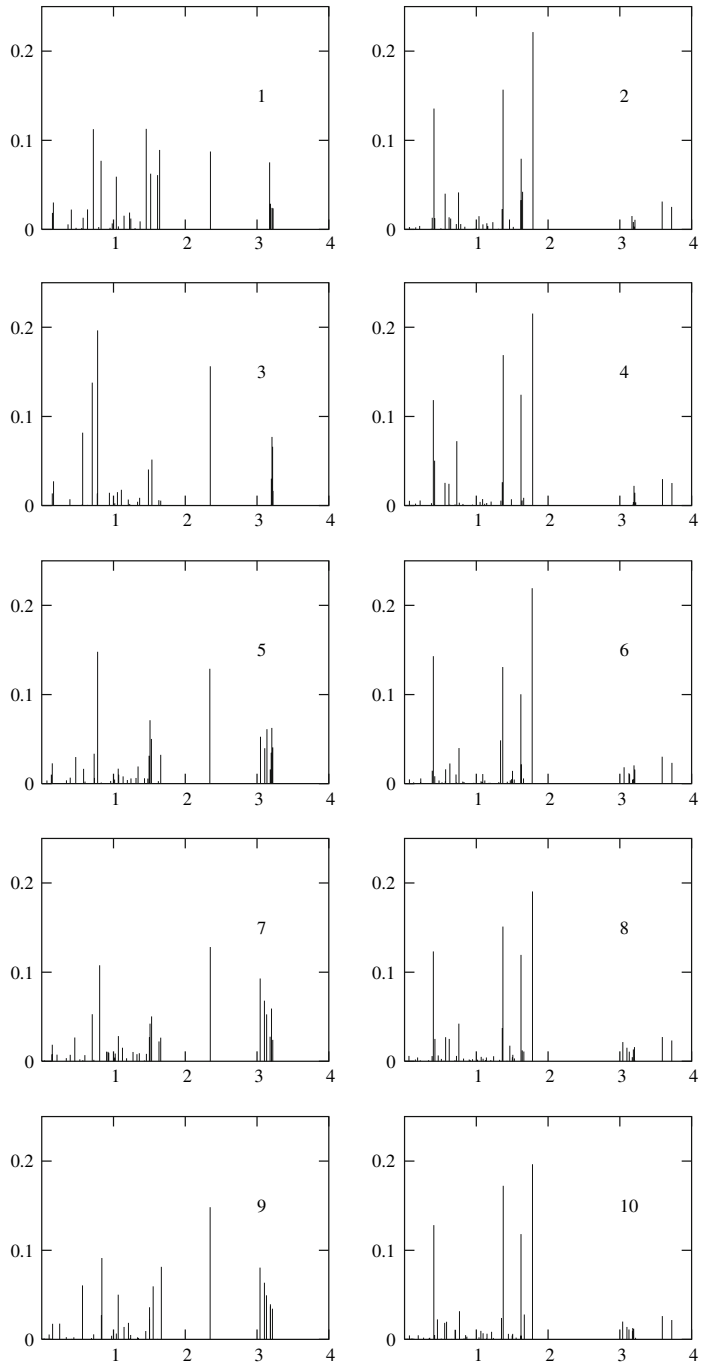

Fig. 3 Normalized intensities versus frequencies divided by 1000 [cm^{-1}] for nitriles (left column) and amides (right column)

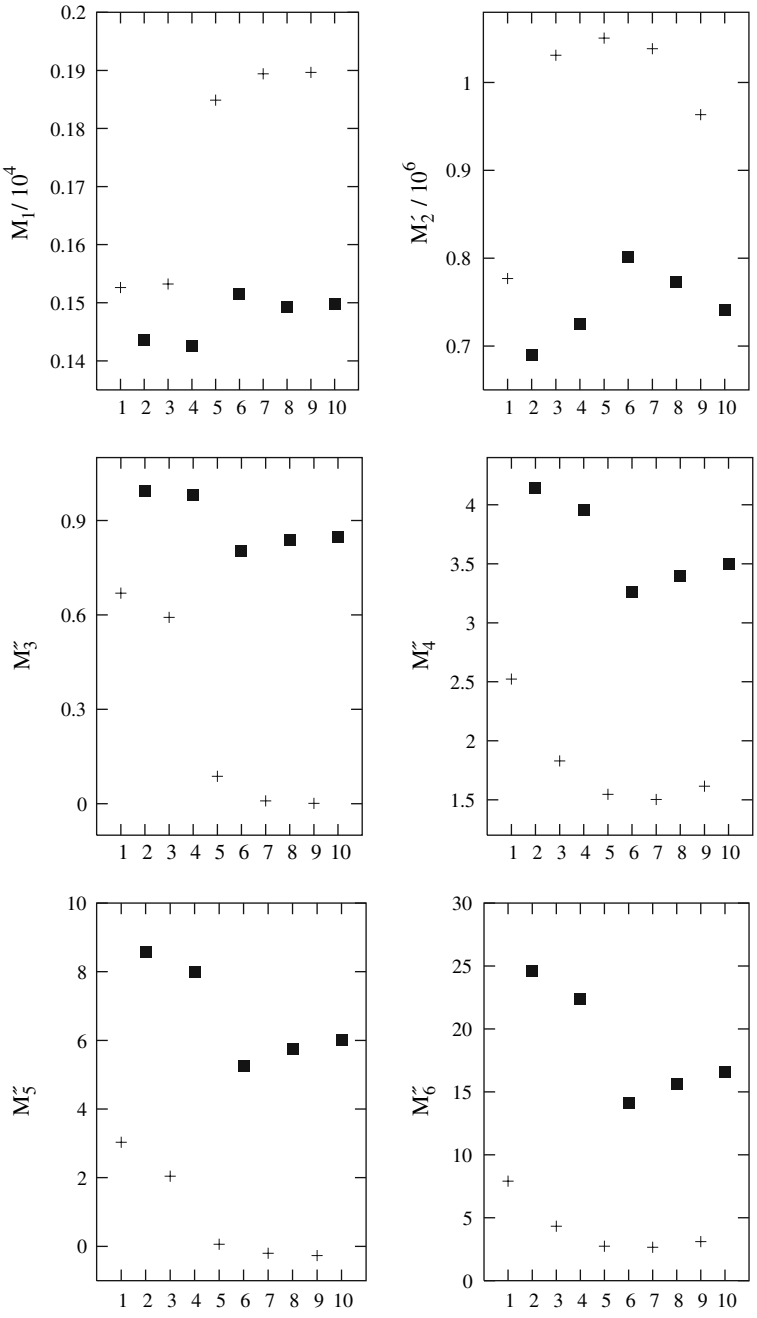

Fig. 4 Moments of the intensity distributions for nitriles (crosses) and for amides (squares)

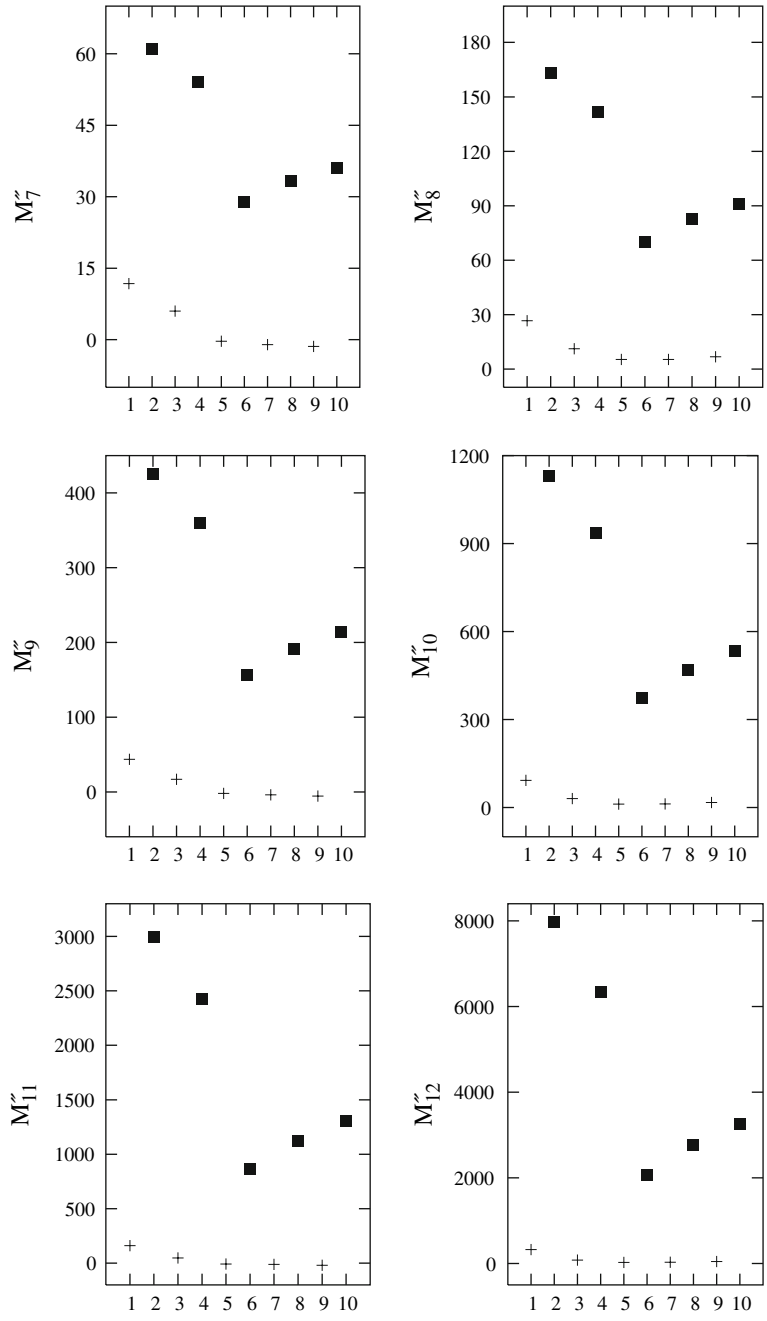

Fig. 5 Moments of the intensity distributions for nitriles (crosses) and for amides (squares)

and moves towards the lower energies (about 1400 cm^{-1}) in the amides (single $C-NH_2$ bond). The next factor differentiating these moments is the lower $C_{ring}-H$ stretching intensity in the amides than that in the corresponding nitriles. The last factor which influences the location of M_1 is a much lower relative intensity of the vibrations of the methyl group atoms observed for the amides (region of 3000 cm^{-1}) in comparison with the same modes in the nitriles [29]. The molecules labeled by 1, 2, 3 and 4 differ from the others — they do not contain CH_3 group. The spectral lines corresponding to this group are located in the high energy region. Therefore, as one can see in Fig. 3, the number of the spectral lines in this region for these compounds is smaller than for the other ones (labeled by 5–10). Consequently, M_1 for the compounds 1–4 is smaller than for the other compounds.

The second centered moments (also smaller for the amides than for the nitriles, Fig. 4) indicate that the variance of the intensity distributions for the amides is smaller than the one for the nitriles. This observation is also clearly seen in Fig. 3. The third and the fourth moments for the amides are bigger than the corresponding moments for the nitriles (Fig. 4). The third moment equal to zero corresponds to a symmetric distribution of the intensity in the spectrum. The intensities for the amides corresponding to the higher frequencies are smaller than the ones corresponding to lower frequencies. As a result the asymmetry (M_3'') of the spectra of the amides is larger.

Summarizing, moments of the intensity distributions can be used as molecular descriptors. The presented kind of classification gives a chance to find new aspects of similarities between molecules. This statistical method will be particularly efficient when we aim at the classification of a large number of molecules. In such a case other methods may be too time consuming.

Acknowledgment Supported by Ministry of Education and Science, Grant No. 2P04A 07229.

References

1. M.A. Johnson, G.M. Maggiora, *Concepts and Applications of Molecular Similarity* (Wiley, New York, 1990)
2. M.A. Johnson, S.C. Basak, G.A. Maggiora, Mathl. Comput. Model. **11**, 630 (1998)
3. S.C. Basak, V.R. Magnuson, G.J. Niemi, R.R. Regal, Discrete Appl. Math. **17**, 17 (1998)
4. S.C. Basak, S. Bertelsen, G.D. Grunwald, Chem. Inf: Compur. Sci. **34**, 270 (1994)
5. P. Willet, V.A. Winterman, Quant. Struct-Act. Relar. **5**, 18 (1986)
6. R. Carbo-Dorca, P.G. Mezey (eds.), *Advances in Molecular Similarity*, vol. 2, (JAI Press, Stamford, CN, 1998), p. 297
7. S.C. Basak, B.D. Gute, D. Mills, D.M. Hawkins, J. Mol. Struct. (Theochem) **622**, 127 (2003)
8. S.E. O'Brien, P.L.A. Popelier, Can. J. Chem. **77**, 28 (1999)
9. S.E. O'Brien, P.L.A. Popelier, J. Chem. Inf. Comp. Sci. **41**, 764 (2001)
10. S.E. O'Brien, P.L.A. Popelier, J. Chem. Soc. Perkins Trans. **2**, 478 (2002)
11. D. Bielińska-Wąż, P. Wąż, S.C. Basak, R. Natarajan, in *Symmetry, Spectroscopy and SCHUR*, eds. by R.C. King et al. (Nicolaus Copernicus University Press, Toruń, 2006), pp. 27–32
12. D. Bielińska-Wąż, P. Wąż, S.C. Basak, J. Math. Chem. (2007) DOI: 10.1007/s10910-006-9155-0
13. H.A. Bethe, Phys. Rev. **50**, 332 (1936)
14. C.E. Porter, *Statistical Theory of Spectra: Fluctuations* (Academic, New York, 1965)
15. W. Rosenzweig, C.E. Porter, Phys. Rev. **120**, 1698 (1960)
16. D. Bielińska-Wąż, in *Symmetry and Structural Properties of Condensed Matter*, eds. by T. Lulek et al. (World Scientific, Singapore, 1999), pp. 212–221
17. D. Bielińska-Wąż, N. Flocke, J. Karwowski, Phys. Rev. B **59**, 2676 (1999)

18. D. Bielińska-Wąż, P. Wąż, J. Math. Chem. (2007) DOI: 10.1007/s10910-007-9241-y
19. D. Bielińska-Wąż, P. Wąż, S.C. Basak, Eur. Phys. J. B **50**, 333 (2006)
20. D. E. Goldberg, *Genetic Algoritm in Search, Optimization & Machine Learning* (Addison-Wesley, 1989)
21. P. Charbonneau, Astr. J. Sup. Ser. **101**, 309 (1995)
22. M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, V.G. Zakrzewski, J.A. Montgomery Jr., R.E. Stratmann, J.C. Burant, S. Dapprich, J.M. Millam, A.D. Daniels, K.N. Kudin, M.C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G.A. Petersson, P.Y. Ayala, Q. Cui, K. Morokuma, D.K. Malick, A.D. Rabuck, K. Raghavachari, J.B. Foresman, J. Cioslowski, J.V. Ortiz, A.G. Baboul, B.B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R.L. Martin, D.J. Fox, T. Keith, M.A. Al-Laham, C.Y. Peng, A. Nanayakkara, M. Challacombe, P.M.W. Gill, B. Johnson, W. Chen, M.W. Wong, J.L. Andres, C. Gonzalez, M. Head-Gordon, E.S. Replogle, J.A. Pople, *Gaussian 98, Revision A.9* Gaussian, Inc., Pittsburgh, (1998)
23. A.D. Becke, J. Chem. Phys. **98**, 5648 (1993)
24. P.J. Stevens, F.J. Devlin, C.F. Chablowski, M.J. Frish, J. Phys. Chem. **98**, 11623 (1994)
25. A.D. Becke, Phys. Rev. A **38**, 3098 (1988)
26. C. Lee, W. Yang, R.G. Paar, Phys. Rev. B **37**, 785 (1988)
27. S.H. Vosko, L. Wilk, M. Nusair, Can. J. Chem. **58**, 1200 (1980)
28. Ł. Pepłowski, K. Kubiak, S. Zelek, W. Nowak, Int. J. Quant. Chem. DOI: 10.1002/qua.21357
29. M.D. Halls, H.B. Schlegel, J. Chem. Phys. **109**, 10587 (1998)
30. H. Yamada, S. Shimizu, M. Kobayashi, Chem. Rec. **1**, 152 (2001)
31. V. Mylerova, L. Martinkova, Curr. Org. Chem. **7**, 1 (2003)
32. W. Nowak, Y. Ohtsuka, J. Hasegawa, H. Nakatsuji, Int. J. Quantum Chem. **90**, 1174 (2002)