Journal of Mathematical Chemistry, Vol. 42, No. 4, November 2007 (© 2006)
DOI: 10.1007/s10910-006-9155-0

Similarity studies using statistical and genetical methods

Dorota Bielińska-Wąż*
Instytut Fizyki, Uniwersytet Mikołaja Kopernika, Grudziądzka 5, 87-100 Toruń, Poland
E-mail: dsnake@phys.uni.torun.pl

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

Subhash C. Basak
Natural Resources Research Institute, 5013 Miller Trunk Highway, MN 55811-1442, USA

Received 27 April 2006; revised 22 May 2006

This paper aims at demonstrating the applicability of statistical spectroscopy and genetic algorithms to the similarity studies. Statistical moments of the intensity distributions are used as a basis for defining similarity distances between pairs of model spectra. Model spectrum is taken as a sum of two Gaussian distributions characterized by different parameters. As a result, dissimilarity maps are presented.

KEY WORDS: data mining, statistical theory of spectra, molecular similarity, genetic algorithms

AMS subject classification: 76M25

1. Introduction

First studies on molecular similarity have been started about 25 years ago [1]. Since then, many indices of molecular similarity have been defined and successfully used in establishing criteria of molecular similarity [2]. In the computation of molecular similarity a large number of mathematical functions can be used to derive measures of similarity for a pair of molecules starting from the same set of structural descriptors [3,4]. Quantitative molecular similarity analysis uses molecular descriptors such as topological indices [5] to identify molecular similarity for property prediction and for environmental risk assessment. One should also mention other methods of similarity measures, as the well known 3D QSAR method, comparative molecular field analysis [6].

*Corresponding author.

Another class of methods may be derived from the quantum-mechanical description of a molecule and its interaction with the environment. Theoretically it is possible to calculate the interaction of a molecule with the environment by solving the appropriate Schrödinger equation. In particular, the electrostatic potential energy function calculated using Born–Oppenheimer approximation encodes all important features of a molecule. In order to extract this information one can use the molecular spectra. Then, quantum similarity among molecules may be derived from quantum chemical calculations [7–9]. One may also use infra-red spectra to calculate molecular descriptors derived from the normal coordinate eigenvalues [10].

In papers [11] and [12], we proposed a new set of similarity indices. These indices relate shapes of molecular spectra. It is assumed that the degree of similarity of molecules is correlated with the degree of similarity of their spectra. According to the so called *principle of moments* [13], we expect that if, we identify the lower moments of two distributions, we bring these distributions to approximate identity. Similarity of distributions in two- and three-moment approximations, in the context of the construction of envelopes of electronic bands, has been analyzed in [14,15].

In this paper, the principle of moments is applied to the theory of molecular similarity. We assume that molecules have similar properties if their intensity distributions and, consequently, the corresponding statistical moments, are approximately the same. A very clear meaning has the first moment, M_1, which describes the mean value of the distribution. The second centered moment, M_2', is the variance, which gives the width of the distribution. M_3'' is the skewness coefficient which describes the asymmetry of the spectrum. The kurtosis coefficient M_4'' is connected to the excess of the distribution. By the classification of the molecules according to the spectral density distribution moments we can discover new characterictics in the field of molecular similarity.

In this paper, dissimilarity maps are presented. The correlation of particular descriptors (based on statistical moments), evident in similarity measures, is shown. The problems restict to finding global maxima of multidimensional functions and are solved using genetic algorithms.

2. Statistical theory of spectra and similarity

Moments of the intensity distribution, $\mathcal{I}^{\gamma}(E)$, belong to a set of fundamental concepts of statistical theory of spectra. The i_kth moment of the continous intensity distribution is defined as:

$$M_{i_k}^{\gamma} = \frac{\int_{C(E)} \mathcal{I}^{\gamma}(E) E^{i_k} \mathrm{d}E}{\int_{C(E)} \mathcal{I}^{\gamma}(E)\ \mathrm{d}E}, \tag{1}$$

where $C(E)$ is the range of the energy for which the integrand does not vanish. It is convenient to consider normalized spectra $I^{\gamma}(E) = N^{\gamma}\mathcal{I}^{\gamma}(E)$ for which the area below the distribution function is equal to 1. Convenient characteristics of the distributions may be derived from the properly scaled distribution moments. Moments normalized to the mean value equal to zero ($M_1^{\gamma'} = 0$) are referred to as the *centered moments*. The i_kth centered moment reads:

$$M_{i_k}^{\gamma'} = \int_{C'(E)} I^{\gamma}(E)(E - M_1^{\gamma})^{i_k}\ \mathrm{d}E. \tag{2}$$

The moments, for which additionally the variance is equal to $1(M_2^{\gamma''} = 1)$ are defined as

$$M_{i_k}^{\gamma''} = \int_{C''(E)} I^{\gamma}(E)\left[\frac{(E - M_1^{\gamma})}{\sqrt{M_2^{\gamma} - \left(M_1^{\gamma}\right)^2}}\right]^{i_k}\ \mathrm{d}E. \tag{3}$$

In this work the model spectrum is approximated by a continous function taken as a linear combination of two unnormalized Gaussian distributions centered at ϵ_i with dispersions σ_i, defined by the parameters $c_i = 1/(2\sigma_i^2), i = 1, 2$:

$$I^{\gamma}(E) = N^{\gamma}\sum_{i=1}^{2} a_i \exp\left[-c_i(E - \epsilon_i)^2\right]. \tag{4}$$

The normalization constant N^{γ} is determined so that the zeroth moment of the distribution $I^{\gamma}(E)$ is equal to 1.

The i_kth moment of the distribution is equal to:

$$M_{i_k}^{\gamma} = N^{\gamma}\sum_{i=1}^{2}\int_{C(E)} a_i \exp\left[-c_i(E - \epsilon_i)^2\right] E^{i_k}\ \mathrm{d}E. \tag{5}$$

The analytical expressions for these moments as functions of parameters c_i, a_i, ϵ_i are presented in [11].

Using moments as descriptors, we can define the similarity distances in the sense of the i_kth, property as [11]:

$$D_{i_k} = 1 - \exp\left[-\left(P_{(i_k)}^{\alpha} - P_{(i_k)}^{\beta}\right)^2\right], \tag{6}$$

where $i_k = 1, 2, \ldots, n\ \ (k = 1, 2, \ldots, n)$, correspond to a specific property and n is the total number of properties taken into account in the comparison of a pair of spectra. The first property is taken as the mean value, the second as the width, the third as the asymmetry, and the fourth as the excess of the compared distributions, i.e. $P_{(1)}^{\gamma} = M_1^{\gamma}$, $P_{(2)}^{\gamma} = M_2^{\gamma\prime}$, $P_{(3)}^{\gamma} = M_3^{\gamma\prime\prime}$, $P_{(4)}^{\gamma} = M_4^{\gamma\prime\prime}$, where $\gamma = \alpha, \beta$. The values of all the descriptors may vary from 0 (identical properties) to 1.

Using D_{i_k}, similarity measures $\mathcal{S}_k^{i_1 i_2,\ldots,i_k}$ (k is the number of properties taken into account in the process of comparison) can be defined as a normalized information derived from a comparison of a pair of distributions, referred to as α and β [11]:

$$\mathcal{S}_k^{i_1 i_2,\ldots,i_k} = \sqrt{\frac{1}{k}\left(D_{i_1}^2 + D_{i_2}^2 + \cdots + D_{i_k}^2\right)}, \tag{7}$$

where $i_1 < i_2 <, \ldots, i_k$.

3. Genetical algorithm and dissimilarity of spectra

The development of the genetic algorithms [16–18] has been inspired by the evolutionary biology, in particular by inheritance, mutation, natural selection, recombination or crossover. They may be classified as computational techniques used for solving problems of optimization. The most common mathematical optimization tasks are minimization and maximization and many classes of problems in physical sciences can be treated as optimization problems. Data fitting using least-squares or root finding problems for non-linear, coupled systems of equations can be treated as minimization problems. Any algebraic system of equations can be solved as a residual minimization problem. Generally, genetic algorithm is a search for solution which starts by generating a set of trial solutions, called population, usually by choosing random values for all model parameters. For each member of the population the goodness of the fit (fitness) is evaluated. The next step is to generate the second generation (population) of solutions. Pairs of solutions (parents) are selected and using genetic operators: crossover (or recombination), and mutation, new solutions are obtained. The child population replaces the old one, the goodness of fit is evaluated, and the process of selection of parents repeates again. Generally, the average fitness increases by this procedure, since only the best organisms are selected for breeding. The generational process is repeated until a terminational condition is reached which can be a minimum criterium, reaching fixed number of generations, computation time, or combination of these condistions.

In this paper, the problem of searching for the most disimilar spectra is treated as optimization. The maximization of functions defined in equations (6) and (7) is performed.

4. Results and discussion

We have considered an infinite number of spectra of the type

$$I^{\gamma}(E) = N^{\gamma}\left[a_1 \exp\left[-c_1(E-\epsilon_1)^2\right] + a_2 \exp\left[-c_2(E-\epsilon_2)^2\right]\right], \tag{8}$$

where $\gamma = \{c_1, a_1, \epsilon_1, c_2, a_2, \epsilon_2\}$. As the reference spectrum is taken $I^{\alpha}(E)$, where

$$\alpha = \{5.0, 1.0, 1.2, 5.0, 1.0, 2.7\}. \tag{9}$$

The particular parameters characterize the width (c_i), the amplitude (a_i) and the locations of the maxima (ϵ_i) of the ith Gaussian component $a_i \exp[-c_i (E-\epsilon_i)^2]$ of $I^{\gamma}(E)$, where $i = 1, 2$.

The aim of this paper is to find spectra $I^{\beta}(E)$ which are the most *dissimilar* to the reference spectrum in the sense of arbitrary properties (one or several). The set of parameters which define the space in which the spectra $I^{\beta}(E)$ are searched, has been restricted to

$$\beta = \{5.0, 1.0, 1.2, 5.0 + \delta c, 1.0 + \delta a, 2.7 - \delta\epsilon\}. \tag{10}$$

Figures 1–4 present pairs of spectra with maximum dissimilarity defined by specific dissimilarity conditions, referred to as *dissimilarity maps*. Solid lines correspond to the reference spectrum $I^{\alpha}(E)$ and the dashed ones represent $I^{\beta}(E)$. In particular, Figure 1 corresponds to the conditions $D_{i_k} = \max$ for $i_k = 1$ (upper left), $i_k = 2$ (upper right), $i_k = 3$ (lower left), and $i_k = 4$ (lower right). Figure 2 corresponds to the conditions $S_2^{i_1 i_2} = \max$ for i_1, i_2, as indicated in the figure. Figure 3 corresponds to the conditions $S_3^{i_1 i_2 i_3} = \max$ for i_1, i_2, i_3, as indicated in the figure. Figure 4 corresponds to the conditions $S_4^{1234} = \max$. The maximum value of D or S means that the respective pair of distributions corresponds to the smallest similarity in the sense of the considered properties within the assumed range of parameters δc, δa, $\delta\epsilon$. The minimum value of D and S is zero. This value is reached when a pair of identical spectra is compared and $\delta c = \delta a = \delta\epsilon = 0$. 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 $I^{\beta}(E)$. In this paper, only the amplitude, the width and the location of this Gaussian component is subjected to variations and the ranges of the changes are restricted by the parameter ranges:

$$\delta c \in \langle 0; 20\rangle, \delta a \in \langle 0; 10\rangle, \delta\epsilon \in \langle 0; 1\rangle. \tag{11}$$

The problem of finding global maxima of functions $D(\delta c, \delta a, \delta\epsilon)$ and of $S(\delta c, \delta a, \delta\epsilon)$ within the assumed ranges of parameters (11) has been solved using the genetical algorithm Pikaia [19] and the results are presented in table 1. The initial population has been created by giving as the input the ranges (11) of the parameters. The search for the parameters $\delta c, \delta a, \delta\epsilon$ for which maxima of D or

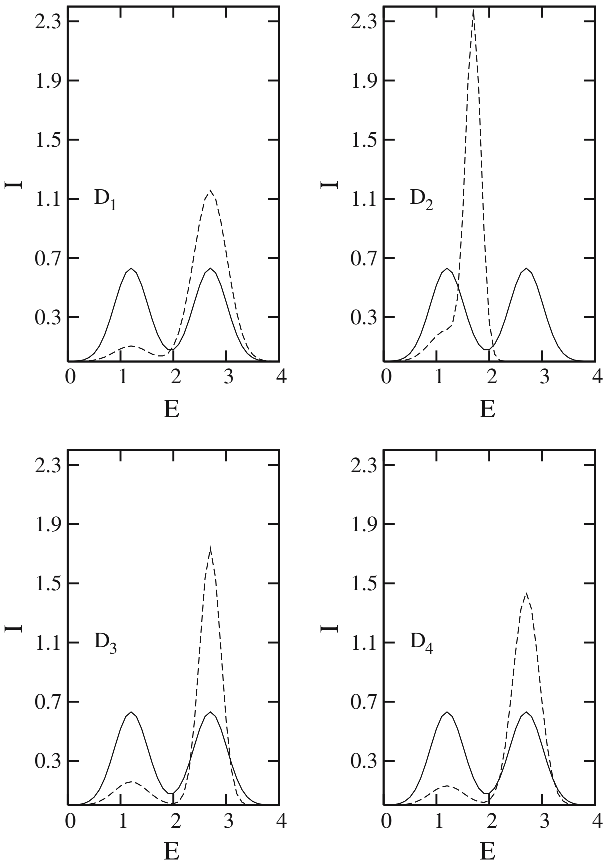

Figure 1. Dissimilarity maps corresponding to the maximum values of D_{i_k} indicated in the figures. Pairs of distributions I^{α} and I^{β} correspond to solid and dashed lines, respectively.

S are obtained has been terminated at the 500th generation. The second column presents the maximum values of the parameters displayed in the first column. The maximum values of D_1 and D_2 are rather small (0.323365 for D_1 and 0.297217 for D_2). This results in a small value of S_2^{12} (0.274429). The maximum values of all the remaining similarity parameters are larger than 0.5. The maximum values of D_{i_k} in Figure 1 appear for maximal amplitude of the second Gaussian component of I^{β} (maximal δa). D_2 has its maximum for very close location of the second Gaussian component of I^{β} to the first one ($\delta\epsilon$ is maximal, the last column in table 1). The remaining maxima in figure 1 appear for the maximal shift of the component Gaussians of I^{β} ($\delta\epsilon = \min = 0.0$). As one should expect, the most narrow distribution I^{β} is the one which is the most dissimilar to I^{α} in the sense of the width. Then, D_2 has its maximum for a large value of δc. A big dissimilarity in the sense of the mean values of I^{β} and I^{α}

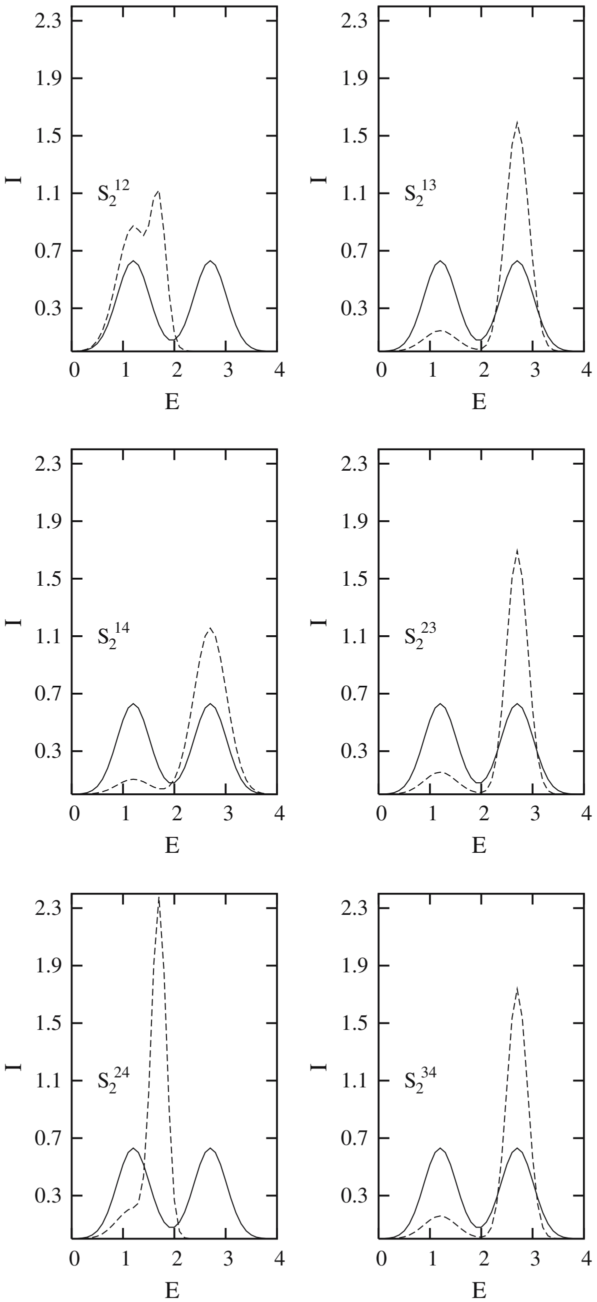

Figure 2. Dissimilarity maps corresponding to the maximum values of $S_2^{i_1 i_2}$ indicated in the figures. Pairs of distributions I^{α} and I^{β} correspond to solid and dashed lines, respectively.

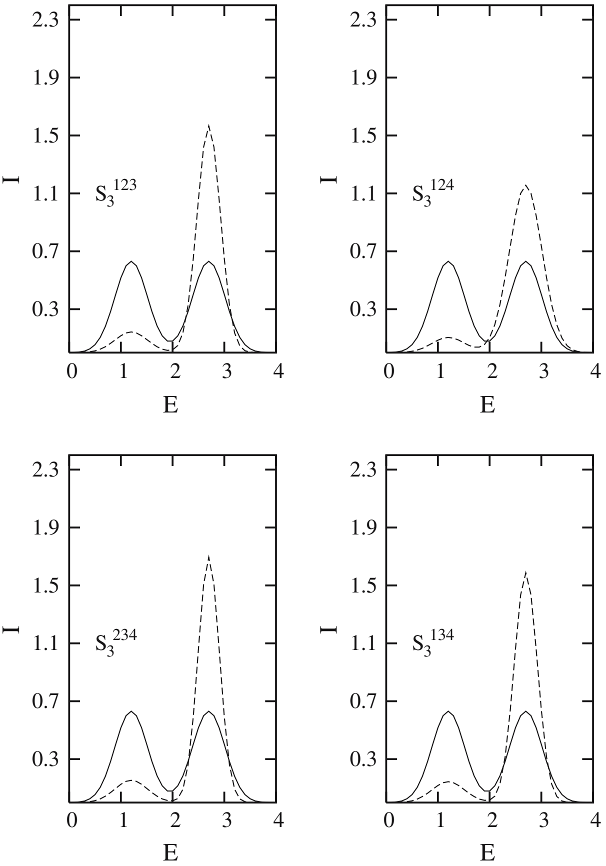

Figure 3. Dissimilarity maps corresponding to the maximum values of $S_3^{i_1 i_2 i_3}$ indicated in the figures. Pairs of distributions I^{α} and I^{β} correspond to solid and dashed lines, respectively.

is for the minimal value of δc, i.e. the intensity of I^{β} is located in its second Gaussian component with a very small first component. The largest dissimilarity in asymmetry results in a medium value of δc.

In figures 2–4 the pairs of distributions which characterize the smallest similarity in sense of several properties are presented. These pairs of spectra are combinations of two (figure 2) of three (figure 3) of four (figure 4) properties which are independently shown in figure 1. The conditions for the largest dissimilarity in sense of all S are realized by the maximal amplitude of the second Gaussian component of $I^{\beta}(E)$ (maximal δa). The exception is S_2^{12} case where the intensity of $I^{\beta}(E)$ is mainly located around the first component ($\delta a = 0$). In this way the smallest similarity of $I^{\alpha}(E)$ and $I^{\beta}(E)$, in sense of the combination of the mean values and of the widths, has been obtained. However δa is big both for D_1 and for D_2 and this feature is not observed for S_2^{12} (there is no correlation between

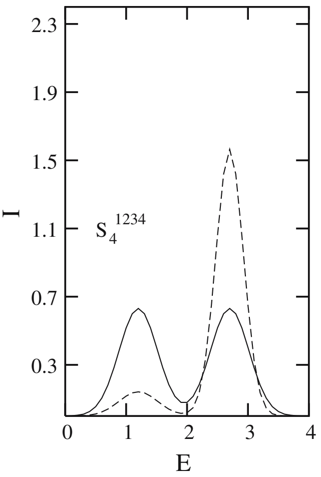

Figure 4. Disimilarity map corresponding to the maximum value of S_4^{1234}. A pair of distributions I^α and I^β correspond to solid and dashed lines, respectively.

Table 1
Values of parameters.

	Max	δc	δa	$\delta \epsilon$
D_1	0.323365	0.000000	9.999900	0.000000
D_2	0.297217	19.39400	9.999900	0.999990
D_3	0.974955	7.377800	9.999900	0.000000
D_4	1.000000	3.091800	9.999900	0.000010
S_2^{12}	0.274429	19.99940	0.000000	0.999990
S_2^{13}	0.716882	5.134200	9.999800	0.000000
S_2^{14}	0.743155	0.000200	9.999800	0.000010
S_2^{23}	0.695311	6.678400	9.999900	0.000000
S_2^{24}	0.737678	19.34680	9.999900	0.999990
S_2^{34}	0.987557	7.390200	9.999900	0.000000
S_3^{123}	0.590453	4.786400	9.999900	0.000000
S_3^{124}	0.612259	0.000000	9.999800	0.000020
S_3^{234}	0.809715	6.690400	9.999900	0.000000
S_3^{134}	0.822160	5.094400	9.999900	0.000000
S_4^{1234}	0.715175	4.769400	9.999900	0.000000

δa for S_2^{12} and δa for D_1 and D_2). A small shift of the second component (maximal $\delta\epsilon$) of $I^\beta(E)$ in the case of S_2^{12} is the second factor which results in the compensation of intensity of $I^\beta(E)$ around the low energy limit of spectrum. Contrary to S_2^{12} case, the parameters δc, δa, $\delta\epsilon$ for other $S_2^{i_1 i_2}$ similarity measures are rather well correlated with their counterpounts of the component D_{i_1} and

D_{i_2} similarity distances. The dissimilarity map for S_2^{13} is intermediate between maps for D_1 and for D_3. The parameters for S_2^{13} seen in table 1 are also intermediate between those for D_1 and D_3. The maps for S_2^{14}, S_2^{23}, S_2^{24}, S_2^{34} contain mainly information about one component. This is also evident in the parameters in table 1. Similarity measure S_2^{14} mainly resembles D_1, S_2^{23} – D_3, S_2^{24} – D_2, andS_2^{34} – D_3.

Taking into account three properties (figure 3) results in dissimilarity maps which are very similar to each other for all considered i_1, i_2, i_3. Considering more properties we converge to one dissimilarity picture which is presented in figure 4. The dissimilarity map for four properties is not very different from the maps for three properties. The parameters δc, δa, $\delta\epsilon$ for S_3 and S_4 are in between for the component D distances.

Summarizing, the genetical algorithms are an efficient tool for study of molecular similarity. Using these methods, similarity, and in particular dissimilarity, maps for arbitrary given similarity conditions, with a high precision and saving computing time, can be created. The application of the presented method to real cases is straighforward and is prepared for the publication.

Acknowledgments

This work has been supported by Ministry of Education and Science, grant no 2 PO3B 033 25.

References

[1] R. Carbo, L. Leyda and M. Arnau, Int. J. Quantum Chem. 17 (1980) 1185.

[2] R. Carbo and B. Calabuig, in: *Molecular Similarity*, ed. M. A. Johnson et al. (Wiley, New York, 1990).

[3] R. Carbo-Dorca and P.G. Mezey (eds.), *Advances in Molecular Similarity*, Vol. 2 (JAI Press, Stamford, CN, 1998) p. 297.

[4] M. Johnson and G.M. Maggiora, *Concepts and Applications of Molecular Similarity* (Wiley, New York, 1990), p. 393.

[5] J. Devillers and A.T. Balaban, (eds.), *Topological Indices and Related Descriptors in QSAR and QSPR*, (Gordon and Breach Science Publishers, The Netherlands 1999), p. 811.

[6] R.D. Cramer III, D.E. Patterson and J.D. Bunce, J. Am. Chem. Soc. 110 (1988) 5959.

[7] P.J. Smith and P.L.A. Popelier, J. Comput.-Aided Mol. Des. 18(2) (2004) 135.

[8] S.E. O'Brien and P.L.A. Popelier, J. Chem. Inf. Comp. Sci. 41 (2001) 764.

[9] S.E. O'Brien and P.L.A. Popelier, J. Chem. Soc. Perkin Trans. 2 (2002) 478.

[10] A.M. Ferguson, T. Heritage, P. Jonathan, S.E. Pack, L. Philips, J. Rogan and P.J. Smith, J. Comp-Aided Mol. Des. 11 (1997) 143.

[11] D. Bielińska-Wąż, P. Wąż and S.C. Basak, Eur. Phys. J. B 50 (2006) 333.

[12] D. Bielińska-Wąż, P. Wąż, S.C. Basak and R. Natarajan in *Symmetry, Spectroscopy and SCHUR*, (eds.), R.C. King et al. (Nicolaus Copernicus University Press, Toruń, 2006), pp. 27–32.

[13] V.S. Ivanov, and V.B. Sovkov, Opt. Spectrosc. 74 (1993) 30.

[14] D. Bielińska-Wąż and J. Karwowski, Phys. Rev. A 52 (1995) 1067.

[15] D. Bielińska-Wąż in: *Symmetry and Structural Properties of Condensed Matter*, eds. T. Lulek et al. (World Scientific, Singapore, 1999), pp. 212–221.

[16] D.E. Goldberg, *Genetic Algoritm in Search, Optimization & Machine Learning* (Addison-Wesley, Reading, MA, 1989).

[17] J.R. Koza, *Genetic Programming: on the Programming of Computers by Means of Natural Selection* (MIT Press, Cambridge, 1992).

[18] Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs* (Springer, New York, 1994).

[19] P. Charbonneau, Astr. J. Sup. Ser. 101 (1995) 309.