communications in mathematical
match
and in computer chemistry

no. 45, March 2002
MATCDY (45) 27 - 33 (2002)
ISSN 0304-6253

Correlation Properties of the Autocorrelation Descriptor for Molecules

Boris Hollas *
University of Ulm, Department of Theoretical Computer Science

May 16, 2002

Abstract

The autocorrelation descriptor is a molecular descriptor encoding both molecular structure and physico-chemical properties attributed to atoms as a vector. Applications include QSAR studies and screening of large databases. Using random graphs, we show that the autocorrelation descriptor may contain highly redundant information even if the encoded properties are independent. We show that this shortcoming can easily be eliminated by centering properties, facilitating subsequent statistical analysis of the generated data.

1 Introduction

To computationally analyze large chemical databases with millions of compounds, a variety of numerical descriptors has been developed. A numerical descriptor is a function that, given a molecule or an atom as input, outputs numerical data such as molecular weight, number of atoms, surface area, or atomic charge. A major application are quantitative structure-activity relationship (QSAR) studies [1], a method to relate the structure of a molecule to a specific biological property. For QSAR, both descriptors for planar (2D) and for spatial (3D) molecule representations are used. While a 3D-descriptor usually changes its values if the molecule shifts to a different spatial conformation, a 2D-descriptor does not do so, which can be an advantage if the final conformation is not known in advance. Since a graph can be derived from any molecule, numerous applications of graph theory on chemistry were published [2, 3]. A topological descriptor is a numerical descriptor

*e-mail: hollas@informatik.uni-ulm.de

that is computed from the molecular graph whereby hydrogen atoms and their bonds are usually omitted. Thus, topological descriptors are 2D-descriptors. One of the first topological descriptors was proposed by Wiener [4] and successfully used to determine boiling points of parrafin. Several other topological descriptors have been proposed since, most of which do not account for physico-chemical properties located at atoms or bonds. The autocorrelation descriptor, first proposed by Moreau and Broto [5], is a topological descriptor that not only encodes the structure of the molecule but also numerical properties assigned to atoms. Apart from QSAR, this descriptor has been used to estimate logP-values [6]-[8], a number related to membrane permeation, for pharmaceutical [9, 10] and toxicological research [11].

For the autocorrelation descriptor, the molecular structure is represented as a graph G and physico-chemical properties of atoms as real values assigned to the vertices of G. To that, let be $D_d = \{(u,v) \mid d(u,v) = d\}$ the set of pairs of vertices (u,v) having distance d (length of shortest path from u to v) and x_u a real-value assigned to vertex u. Then

$$A_d = \sum_{(u,v)\in D_d} x_u x_v \tag{1}$$

is the *d-distance autocorrelation descriptor* of G. As a distance-based function, A_d is invariant for different labellings of G, hence, (1) can be defined as the autocorrelation descriptor of the molecule corresponding to G.

In practice, since not all molecular graphs in a chemical dataset have the same maximum distance, A_d is calculated for a distance $d \leq d^*$ with $d^* = 5$ or $d^* = 10$. For various physico-chemical properties the vector $(A_1, \ldots, A_{d^*})$ or a set of respective vectors is then used to describe the molecule.

The name 'autocorrelation descriptor' is a misnomer, (1) is actually a convolution. Still, we use the former name to be consistent with the literature.

To analyze mathematical properties of the autocorrelation descriptor, we model molecular structures as *random graphs* [13, 14]. These are graphs $G_{n,p}$ on a fixed set of vertices $V = \{1, \ldots, n\}$ whose edges are selected independently with probability $p \in (0,1)$. Thus, the number of edges is binomially distributed with expectation $\binom{n}{2}p$. To model molecular structures, we set $p = \frac{2}{n-1}$ in section 3 so that the expectation of the number of edges equals the number of vertices n. To model physico-chemical properties, we associate with each vertex $v \in V$ a random variable X_v. Hence, the function (1) becomes a random variable

$$A_d(\mathbf{X}) = A_d(\mathbf{X}, G_{n,p}) = \sum_{(u,v)\in D_d} X_u X_v\,, \quad \mathbf{X} = (X_1, \ldots, X_n)$$

and D_d is now a random set on the space of random graphs. In particular, D_1 is the random set of edges. $\mathbf{X}$ is the vector of properties X_u attributed to atom u $(u = 1, \ldots, n)$. To represent the molecular structure only, we set $\mathbf{X} = \mathbf{1} = (1, \ldots, 1)$. We assume that $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) and independent of D_d, i.e. independent of the graphical structure.

2 Moments of A_1

The moments we deduce for A_1 will be needed in section 3. Let

$$1_{\{(u,v)\in D_d\}} = \begin{cases} 1 & \text{if} \quad (u,v) \in D_d \\ 0 & \text{else} \end{cases}$$

be the indicator function of $\{(u,v) \in D_d\}$. Then,

$$E\left(1_{\{(u,v)\in D_1\}}\right) = \begin{cases} p & \text{for} \quad u \neq v \\ 0 & \text{else} \end{cases}$$

for random graphs.
The formulae for $E\left(1_{\{(u,v)\in D_d\}}\right)$ become quickly complicated for $d > 1$ and yet, a general formula for all $d > 1$ is not known [12]. We therefore restrict our analysis to $d = 1$ for the rather tedious case $E(X) \neq 0$; the simpler case $E(X) = 0$ can be handled for all $d > 0$.
Let $X_1, \ldots, X_n$ be i.i.d. random variables independent of $1_{\{(u,v)\in D_1\}}$. The expectation of $A_1(\mathbf{X})$ is

$$E(A_1(\mathbf{X})) = E\left(\sum_{u,v} X_u X_v \cdot 1_{\{(u,v)\in D_1\}}\right) =$$

$$E(X_1)^2 \sum_{u\neq v} E\left(\cdot 1_{\{(u,v)\in D_1\}}\right) = 2E(X_1)^2 \binom{n}{2} p \tag{2}$$

Let $Y_1, \ldots, Y_n$ be random variables such that

1. $Y_1, \ldots, Y_n$ are i.i.d. and independent of $1_{\{(u,v)\in D_1\}}$
2. X_u, Y_v are independent for $u \neq v$

Note that X_v, Y_v are not necessarily independent. Then

$$E(A_1(\mathbf{X})A_1(\mathbf{Y})) = E\left(\sum_{u,v,i,j} X_u X_v Y_i Y_j \cdot 1_{\{(u,v)\in D_1\}} \cdot 1_{\{(i,j)\in D_1\}}\right) \tag{3}$$

Since $(u,v),(i,j) \in D_1$, equality between variables in $\{u,v\}$ and $\{i,j\}$ can only occur for $u = k_1$ or $v = k_2$ with $\{k_1,k_2\} = \{i,j\}$. Also, all variables can be unequal, hence we have to consider $\binom{2}{0} + 2\left(\binom{2}{1} + \binom{2}{2}\right) = 7$ cases of which for symmetry reasons 3 have different expectations. By independence and linearity (3) thus becomes

$$= \binom{2}{0} E\left(\sum_{\substack{u,v,i,j=1 \\ all\neq}}^{n} X_u X_v Y_i Y_j \cdot 1_{\{(u,v)\in D_1\}} \cdot 1_{\{(i,j)\in D_1\}}\right)$$

$$+2\binom{2}{1} E\left(\sum_{\substack{u,v,i=1 \\ all\neq}}^{n} X_u X_v Y_i Y_v \cdot 1_{\{(u,v)\in D_1\}} \cdot 1_{\{(i,v)\in D_1\}}\right)$$

$$+2\binom{2}{2} E\left(\sum_{\substack{u,v=1 \\ u\neq v}}^{n} X_u X_v Y_u Y_v \cdot 1_{\{(u,v)\in D_1\}}\right)$$

$$= E(X)^2\,E(Y)^2 \sum_{\substack{u,v,i,j\\ all\neq}} E\Big(1_{\{(u,v)\in D_1\}}\Big)\,E\Big(1_{\{(i,j)\in D_1\}}\Big)$$

$$+4E(XY)\,E(X)\,E(Y) \sum_{\substack{u,v,i\\ all\neq}} E\Big(1_{\{(u,v)\in D_1\}}\Big)\,E\Big(1_{\{(i,v)\in D_1\}}\Big)$$

$$+2E(XY)^2 \sum_{u\neq v} E\Big(1_{\{(u,v)\in D_1\}}\Big)$$

$$= E(X)^2\,E(Y)^2 \cdot 4!\binom{n}{4}p_n^2 + 4E(XY)\,E(X)\,E(Y)\cdot 3!\binom{n}{3}p_n^2$$

$$+2E(XY)^2 \cdot 2!\binom{n}{2}p_n$$

$$= 24E(X)^2\,E(Y)^2\binom{n}{4}p_n^2 + 24E(XY)\,E(X)\,E(Y)\binom{n}{3}p_n^2$$

$$+4E(XY)^2\binom{n}{2}p_n \tag{4}$$

Note that this result is valid for independent random variables X_i, Y_j as well as for $X_i = Y_i$.

3 Correlation Analysis of A_1

We set $p = \frac{2}{n-1}$ so that by (2), $E(A_1(\mathbf{X})) = 2E(X)^2 n$. By (4) and an elementary calculation, we get

$$E(A_1(\mathbf{X})A_1(\mathbf{Y})) = \ldots = 4E(X)^2E(Y)^2\frac{n^3}{n-1}\left(1-\frac{3}{n^2}\right) \tag{5}$$

and for the variance

$$Var(A_1(\mathbf{X})) = \ldots = 4E(X)^4\,\frac{n^3}{n-1} + 4\frac{n^2}{n-1}\left(6\,Var(X)\,E(X)^2 + Var(X)^2\right)$$

$$-4\frac{n}{n-1}\left(3E(X)^4 + 10\,Var(X)\,E(X)^2 + Var(X)^2\right) - \left(2E(X)^2\,n\right)^2$$

In particular,

$$Var(A_1(\mathbf{1})) = 4\frac{n^2}{n-1}\left(1-\frac{3}{n}\right)$$

Finally, the correlation of interest is

$$\rho(A_1(\mathbf{X}), A_1(\mathbf{Y})) = \frac{E(X)^2E(Y)^2\,Var(A_1(\mathbf{1}))}{\sqrt{Var(A_1(\mathbf{X}))\;Var(A_1(\mathbf{Y}))}}$$

by (5). Since

$$\lim_{E(X)\to\pm\infty}\frac{Var(A_1(\mathbf{X}))}{E(X)^4} = Var(A_1(\mathbf{1}))$$

for $Var(X)$ constant, we get

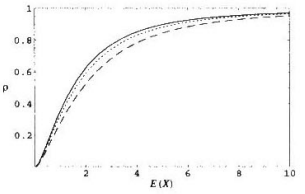
$$\lim_{E(X)\to\pm\infty}\rho\left(A_1(\mathbf{X}), A_1(\mathbf{1})\right) = 1$$

and

$$\lim_{E(X),E(Y)\to\pm\infty} \rho\left(A_1(\mathbf{X}), A_1(\mathbf{Y})\right) = 1$$

This means that A_1 contains highly redundant information for large values of $|E(X)|$ and $|E(Y)|$ even if properties $\mathbf{X}$ and $\mathbf{Y}$ are independent. Also, A_1 contains almost only structural information in this case, all physico-chemical information on the vertices is lost as $|E(X)|$ tends to infinity.

To show the rate of growth, we have plotted $\rho(A_1(\mathbf{X}), A_1(\mathbf{1}))$ as a function of $|E(X)|$ for $n = 5$ (dashed line), $n = 10$ (dotted line) and $n \to \infty$ (solid line) and $Var(\mathbf{X}) = 1$. From the chart, we see that for all reasonable molecular sizes the autocorrelation descriptors of property $\mathbf{X}$ and of the molecular structure are strongly correlated for $E(X) > 3$.



Centered properties

As we have seen for $d = 1$, the correlation is 0 for $E(X) = 0$ or $E(Y) = 0$. This result is easily verified for all positive distances.
Without loss of generality, assume that $E(X) = 0$ and $d_1, d_2 > 0$. For $G_{n,p}$, by (2)

$$Cov(A_{d_1}(\mathbf{X}), A_{d_2}(\mathbf{Y})) = \sum_{u,v}\sum_{i,j} Cov\Big(X_u X_v 1_{\{(u,v)\in D_{d_1}\}}, Y_i Y_j 1_{\{(i,j)\in D_{d_2}\}}\Big)$$

$$= \sum_{u,v}\sum_{i,j} E(X_u X_v Y_i Y_j)\, E\Big(1_{\{(u,v)\in D_{d_1}\}} 1_{\{(i,j)\in D_{d_2}\}}\Big) \tag{6}$$

We consider two cases:

1. If $X_i, Y_j\ (i, j = 1, \ldots, n)$ are independent then $Cov(A_{d_1}(\mathbf{X}), A_{d_2}(\mathbf{Y})) = 0$
2. Assume $X_i = Y_i$ for $i = 1, \ldots, n$ and $d_1 \neq d_2$. This implies that one variable out of u, v, i, j is unequal to the others. Without loss of generality, let be $u \neq v, i, j$. Then $covA_{d_1}(\mathbf{X})A_{d_2}(\mathbf{Y}) = 0$ by (6).

Thus, we have shown that

1. $\rho(A_{d_1}(\mathbf{X}), A_{d_2}(\mathbf{Y})) = 0$ for all $d_1, d_2 > 0$ if $X_i, Y_j\ (i, j = 1, \ldots, n)$ are independent and $E(X) = 0$. Note that $E(Y)$ may be different from zero. Of course, this result is also valid for $\mathbf{Y} = \mathbf{1}$.
2. $\rho(A_{d_1}(\mathbf{X}), A_{d_2}(\mathbf{X})) = 0$ for all $d_1 \neq d_2$ and $E(X) = 0$.
3. $\rho(A_1(\mathbf{X}), A_1(\mathbf{Y})) \to 1$ for $n \to \infty$ and $E(X), E(Y) \to \pm\infty$ or $E(X) \to \pm\infty, \mathbf{Y} = \mathbf{1}$.

4 Discussion

For a simple probabilistic model of molecules we studied correlations of the autocorrelation descriptor. It turned out that $A_1(\mathbf{X})$ and $A_1(\mathbf{Y})$ are strongly correlated even for independent properties $\mathbf{X}$ and $\mathbf{Y}$ if $|E(X)|$ and $|E(Y)|$ are large. In this case, $A_1(\mathbf{X})$ also correlates with the molecular structure as expressed by $A_1(\mathbf{1})$. $A_1(\mathbf{X})$, $A_1(\mathbf{Y})$, and $A_1(\mathbf{1})$ become linearly dependent as n tends to infinity. However, QSAR studies with highly correlated descriptors are impractical since most physico-chemical information is lost and the influence of different properties can hardly be separated. This cannot be overcome by factor analysis since factors are linear combinations of all descriptors. If however properties are centered (we call a property $\mathbf{X}$ centered if $E(\mathbf{X}) = \mathbf{0}$) all autocorrelation descriptors are uncorrelated and we can separate the influence of different properties, including the topology of the molecule. Properties should therefore always be centered before the autocorrelation descriptor is applied to facilitate subsequent statistical analysis.

As a drawback, the random graph model we used does not properly reflect the structure of molecules in a typical data set. For instance, the number of vertices in $G_{n,p}$ is constant, also $G_{n,p}$ is not necessarily planar or even connected. On the other hand, any chemical graph can be conceived as a realization of some random graph and numerical simulations we carried out on a chemical data set confirmed our results. Nevertheless, a more general model should be developed for a more precise analysis.

References

[1] Kubinyi, H.: *QSAR: Hansch analysis and related approaches.* VCH, 1993

[2] Trinajstić, N.: *Chemical Graph Theory.* CRC Press, 1992

[3] Bonchev, D., Rouvray, D. H., Editors: *Chemical Graph Theory.* Gordon and Breach Science Publishers, Vol. 1 1991, Vol. 2 1992.

[4] Wiener, H.: *Structural determination of parrafin boiling points.* J. Am. Chem. Soc. 69, 17-20 (1947)

[5] Moreau, G. and Broto, P.: *Autocorrelation of a topological structure: A new molecular descriptor.* Nouv. J. Chim. 4, 359-360 (1980).

[6] Devillers, J., Domine, D., and Karcher, W.: *Estimating n-octanol/water partition coefficients from the autocorrelation method.* SAR QSAR Environ. Res. 3, 301-306 (1995)

[7] Devillers, J., Domine, D., Guillon, C., Bintein, S., and Karcher, W.: *Prediction of partition coefficients using autocorrelation descriptors.* SAR QSAR Environ. Res. 7, 151-172 (1997)

[8] Devillers, J. and Domine, D.: *Comparison of reliability of log P values calculated from a group contribution approach and from the autocorrelation method.* SAR QSAR Environ. Res. 7, 195-232 (1997).

[9] Wagener, M., Sadowski, J., Gasteiger, J.: *Autocorrelation of molecular properties for modelling corticosteroid binding globulin and cytosolic Ah receptor activity by neural networks.* J. Am. Chem. Soc. 117, 7769-7775 (1995).

[10] Bauknecht, H., Zell, A., Bayer, H., Levi, P., Wagener, M., Sadowsky, J., Gasteiger, J.: *Locating biologically active compounds in medium-sized heterogeneous datasets by topological autocorrelation vectors: Dopamine and benzodiazepine agonists.* J. Chem. Inf. Comput. Sc. 36, 1205-1213 (1996).

[11] Devillers, J: *Autocorrelation descriptors for modelling (eco)toxicological endpoints.* In: Devillers, J., Balaban, A. T., Editors: Topological Indices and Related Descriptors in QSAR and QSPR, pp. 595-612. Gordon and Breach Science Publishers, 1999.

[12] Simon, K., Personal Communication

[13] Bollobas, B.: *Random Graphs.* Academic Press, 1984.

[14] Palmer, E. M.: *Graphical evolution.* Wiley, 1985.