Journal of Mathematical Chemistry Vol. 33, No. 2, February 2003 (© 2003)

An analysis of the autocorrelation descriptor for molecules

Boris Hollas

University of Ulm, Department of Theoretical Computer Science, D-89069 Ulm, Germany
E-mail: hollas@informatik.uni-ulm.de

Received 23 April 2002; revised 05 December 2002

We discuss properties of the autocorrelation descriptor, a topological descriptor encoding both molecular structure and physico-chemical properties of a molecule. We introduce two random graph models for molecules and show that this descriptor may exhibit unwanted correlation properties, making the generated data unusable for structure–activity relationship studies. This shortcoming can easily be eliminated by centering properties, facilitating subsequent statistical analysis.

KEY WORDS: molecular descriptor, random graphs, correlation

1. Introduction

For the needs of computational chemistry a variety of descriptors has been developed. A (numerical) descriptor is a function or an algorithm that accepts a representation of a molecule or an atom as input and outputs some numerical data. Descriptors are used in computer aided drug design for tasks such as similarity analysis, clustering, and quantitative structure–activity relationship (QSAR) studies [1], a method to relate the structure of a molecule to a specific biological property. Both descriptors for planar (2D) and for spatial (3D) molecule representations are utilized. 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. Topological descriptors, also called molecular connectivity indices [2], are 2D-descriptors computed from the molecular graph whereby hydrogen atoms and their bonds are usually omitted (for the numerous applications of graph theory on chemistry see [3,4]). Due to their minimal computational requirements, topological descriptors are frequently used to analyze virtual combinatorial libraries or large chemical databases with thousands or even millions of compounds. One of the first topological descriptors was proposed by Wiener [5] and successfully used to determine boiling points of paraffin. Several other topological descriptors have been proposed since, most of which do not account for physico-chemical properties located at atoms or bonds. Moreau and Broto [6] proposed a topological descriptor that not only encodes the structure of the molecule but also numerical properties assigned to atoms. This *autocorrelation descriptor* has been used to estimate log P-values [7], a number related

to membrane permeation, for pharmaceutical [8,10,11] and toxicological research [9]. A software package that uses 2D-autocorrelation is, e.g., DRAGON [12].

2. Preliminaries

For the autocorrelation descriptor, the molecular structure is represented as a graph G and physico-chemical properties of atoms (e.g., volume, electronegativity, hydrophobicity) as real values assigned to the vertices of G. To that, let $D_d = \{(u, v) \mid d(u, v) = d\}$ be 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 distances $d \leqslant d^*$ and the vector $(A_1, \ldots, A_{d^*})$ is then used to describe the molecule. Typical values are $d^* = 8$ or $d^* = 10$. Thus, all molecules in the dataset have a uniform description.

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 molecules as *random graphs* [13,14] that have a random variable associated with each vertex. We use two random graph models:

1. A general model with an arbitrary graphical structure. We do not make any assumptions on the edge distribution. Especially, this model is valid for all chemical structures.

2. A model in which the number N of vertices is a random variable and edges are selected independently with a probability that depends on the actual number of vertices. The expected number of edges equals the expected number of vertices $E(N)$. Any chemical graph can be regarded as a realization of such a random graph.
   This model is a generalization of a model we studied in [15].

To model physico-chemical properties of atoms, we associate with each vertex $v \in V = \{1, \ldots, N\}$ a random variable X_v. Hence, the function (1) becomes a random variable

$$A_d(\mathbf{X}) = \sum_{(u,v)\in D_d} X_u X_v, \qquad \mathbf{X} = (X_1, \ldots, X_N),$$

Figure 1. The graph of phenol and its edge graph.

and D_d is now a random set on the space of random graphs. In particular, $D_1 \subset V^2$ 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 $\mathbf{X} = \mathbf{1} = (1, \ldots, 1)$.

If we want to analyze properties of chemical bindings instead of atoms, the autocorrelation descriptor can be applied to the *edge graph* of the molecule: iff[1] e, f are adjacent edges (i.e., have a vertex in common) of G, its edge graph $\widehat{G}$ contains vertices v_e, v_f and an edge (v_e, v_f).

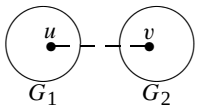
We assume that $E(X_u) = E(X_1)$, $u = 1, \ldots, N$, and $X_1, \ldots, X_N$ are independent and independent of D_1, i.e., independent of the graphical structure.

3. Basic properties

In this section, we examine some basic properties of the function $A_d = A_d(G)$, $d > 0$, for a graph G.

Note that in (1), every summand appears twice for $d > 0$ since $(u, v) \in D_d$ iff $(v, u) \in D_d$. In some of the derivations below we consider pairs (u, v) for which this symmetry is not given and the corresponding summands are multiplied by 2.

A_d has an additive property. Consider two graphs G_1, G_2 that are connected by at least one edge.



For this ensemble then holds

$$A_d(G_1, G_2) = A_d(G_1) + A_d(G_2) + 2 \sum_{\substack{u \in V(G_1), v \in V(G_2) \\ d(u,v)=d}} x_u x_v.$$

Especially, if G_2 is a single vertex v connected to G_1, we get

$$A_d(G_1, v) = A_d(G_1) + 2x_v \sum_{\substack{u \in V(G_1) \\ d(u,v)=d}} x_u. \tag{2}$$

Small changes of the values x_v or the graphical structure, therefore, result in small changes of the corresponding autocorrelation descriptors. If we assume that similar compounds will exhibit similar physico-chemical and biological properties (*similar property principle*) then we may expect that molecules with similar values A_d will be similar biologically.

[1] if and only if.

Figure 2. Different weighted graphs with identical values of A_d.

Figure 3. Different graphs with identical values of A_d.

For trees, from (2) it follows by induction

$$A_d(G) = \sum_{v \in V(G)} x_v C_d^{(v)}$$

with

$$C_d^{(v)} = \sum_{\{u | d(u,v)=d\}} x_u.$$

In this case, A_d is a sum of the *d-distance neighborhood* $C_d^{(v)}$ of v, weighted by x_v.

A_d is not characteristic for edge-weighted graphs. Though the graphs below have different real values at their edges, the values A_d, $d = 0, \ldots, 3$, are identical. This is also true if A_d is only to encode the structure of the graph, i.e., if $x_v = 1$ for all vertices v: for both the graphs below, $A_0, \ldots, A_6$ are identical.

Labeled graphs with equal structure are called *isomorphic*. Since it is widely believed (but not yet proved) that no polynomial-time algorithm exists for the decision problem whether two labeled graphs are isomorphic [16], results as those above are to be expected. However, molecular graphs that have identical values A_d have the same number of vertices $n = A_0(\mathbf{1})$, edges $m = A_1(\mathbf{1})$, and rings r: since molecular graphs are planar (can be drawn in a plane without crossing edges), Euler's polyhedron theorem $n - m + r = 1$ holds. This may in part explain successful clusterings [8,10] using the autocorrelation descriptor.

4. The autocorrelation descriptor for $E(X) = 0$

Consider an arbitrary random set D_1 and an arbitrary random vector $\mathbf{X}$ independent of D_1: this is our general model. D_1 represents the molecular structure and X_u a numerical property of atom u. Since we do not make any assumptions on D_1, this model is valid for arbitrary chemical structures.

As usual, we denote by $L_p = \{X \mid E(|X^p|) < \infty\}$ the set of p-times integrable random variables. Let $N > 0$, $N \in L_2$, be an integer-valued random variable and $\mathbf{X}$ a random vector having the property:

(iie) $\mathbf{X} = (X_1, \ldots, X_N)$ and all $X_u \in L_2$ are independent and independent of D_d with $E(X_u) = E(X_1)$ $(u = 1, \ldots, N)$.

Note, however, that N is not independent of D_d. For ease of notation, we write $E(X)$ instead of $E(X_1)$. Let

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

denote the indicator function of $\{(u,v) \in D_d\}$, $d \geqslant 0$. Then,

$$A_d(\mathbf{X}) = \sum_{u,v=1}^{N} X_u X_v \cdot 1_{\{(u,v)\in D_d\}}$$

and

$$\begin{aligned} E\big(A_d(\mathbf{X})\big) &= \sum_{n=1}^{\infty} E\big(A_d(\mathbf{X}|N=n)\big)P(N=n) \\ &= \sum_{n=1}^{\infty} E\left(\sum_{u,v=1}^{n} X_u X_v \cdot 1_{\{(u,v)\in D_d\}} \Big| N=n\right) P(N=n). \end{aligned}$$

If $d > 0$, this equals

$$E(X)^2 \sum_{n=1}^{\infty} E\left(\sum_{u,v=1}^{n} 1_{\{(u,v)\in D_d\}} \Big| N=n\right) P(N=n), \tag{3}$$

since X_u, X_v are independent for $u \neq v$.

Let $\mathbf{Y}$ be a random vector with property (iie) and X_k, Y_l be independent for $k \neq l$, but not necessarily independent of $\mathbf{X}$. This includes the case $\mathbf{X} = \mathbf{Y}$. Then

$$\begin{aligned} &E\big(A_{d_1}(\mathbf{X})A_{d_2}(\mathbf{Y})\big) \\ &= \sum_{n=1}^{\infty} E\big(A_{d_1}(\mathbf{X})A_{d_2}(\mathbf{Y})\big|N=n\big)P(N=n) \\ &= \sum_{n=1}^{\infty} E\left(\sum_{u,v,i,j=1}^{n} X_u X_v Y_i Y_j \cdot 1_{\{(u,v)\in D_{d_1}\}} \cdot 1_{\{(i,j)\in D_{d_2}\}} \Big| N=n\right) P(N=n). \end{aligned} \tag{4}$$

To determine $\mathrm{Cov}(A_{d_1}(\mathbf{X}), A_{d_2}(\mathbf{Y}))$ for distances $d_1, d_2 \geqslant 0$ and $E(X) = 0$ we consider the following cases:

1. $d_1 = d_2 = 0$ and X_k, Y_l are independent for all k, l. In this case,

$$E\big(A_0(\mathbf{X})|N=n\big) = E\left(\sum_{u=1}^{n} X_u^2\right) = nE\big(X^2\big)$$

and

$$E\big(A_0(\mathbf{X})A_0(\mathbf{Y})\big|N=n\big) = E\left(\sum_{u,v=1}^{n} X_u^2 Y_v^2\right) = n^2 E\big(X^2\big)E\big(Y^2\big).$$

Hence,

$$\begin{aligned}\operatorname{Cov}\big(A_0(\mathbf{X}),\, A_0(\mathbf{Y})\big) &= \sum_{n=1}^{\infty} n^2 E\big(X^2\big) E\big(Y^2\big) P(N=n) \\ &\quad - \sum_{n=1}^{\infty} n E\big(X^2\big) P(N=n) \cdot \sum_{n=1}^{\infty} n E\big(Y^2\big) P(N=n) \\ &= E\big(X^2\big) E\big(Y^2\big) E\big(N^2\big) - E\big(X^2\big) E(N) E\big(Y^2\big) E(N) \\ &= E\big(X^2\big) E\big(Y^2\big) \operatorname{Var}(N) > 0\end{aligned}$$

if $X, Y \not\equiv 0$, $N \not\equiv c$ for a constant c.
If, however, we define

$$\widetilde{A}_0(\mathbf{X}) = \sum_{u=1}^{N} \big(X_u^2 - 1\big) = A_0(\mathbf{X}) - N, \tag{5}$$

then

$$\operatorname{Cov}\big(\widetilde{A}_0(\mathbf{X}),\, A_0(\mathbf{Y})\big) = E\big(X^2 - 1\big) E\big(Y^2\big) \operatorname{Var}(N) = 0$$

if $E(X_u^2) = 1$ for all u. This condition is equivalent to $\operatorname{Var}(\mathbf{X}) = \mathbf{1}$ in the case $E(X) = 0$, i.e., if $\mathbf{X}$ is centered and normalized.

2. $d_1 > 0$, $d_2 \geqslant 0$, $d_1 \neq d_2$ and $XY \in L_2$. Then, without loss of generalization, X_u is independent of the other variables, hence,

$$\begin{aligned}&E\big(A_{d_1}(\mathbf{X}) A_{d_2}(\mathbf{Y}) \big| N = n\big) \\ &\quad = E(X) E\left(\sum_{u,v,i,j=1}^{n} X_v Y_i Y_j \cdot 1_{\{(u,v)\in D_{d_1}\}} \cdot 1_{\{(i,j)\in D_{d_2}\}} \Big| N = n \right) = 0.\end{aligned}$$

The second factor is always finite by Cauchy–Schwarz inequality and $XY \in L_2$. By (3), $E(A_{d_1}(\mathbf{X})) = 0$; by (4), $E(A_{d_1}(\mathbf{X}) A_{d_2}(\mathbf{Y})) = 0$, hence,

$$\operatorname{Cov}\big(A_{d_1}(\mathbf{X}),\, A_{d_2}(\mathbf{Y})\big) = 0.$$

3. $d_1 = 0$, $d_2 > 0$, $XY \in L_2$ and $E(X_u^2) = 1$ for all u. In this case, it follows as above that

$$\operatorname{Cov}\big(\widetilde{A}_0(\mathbf{X}),\, A_{d_1}(\mathbf{Y})\big) = 0.$$

4. $d_1 = d_2 > 0$ and X_k, Y_l are independent for all $k \neq l$. Then X_u is independent of the other variables and $\operatorname{Cov}(A_{d_1}(\mathbf{X}),\, A_{d_2}(\mathbf{Y})) = 0$.

Thus, we have shown: if $E(X) = 0$, then

1. $A_{d_1}(\mathbf{X})$ and $A_{d_2}(\mathbf{X})$ are uncorrelated for different distances d_1, d_2.
2. $A_{d_1}(\mathbf{X})$ and $A_{d_2}(\mathbf{Y})$ are uncorrelated for all distances $d_1, d_2 > 0$. If additionally $\operatorname{Var}(\mathbf{X}) = \mathbf{1}$ and modification (5) is applied for $d_1 = 0$ then this is true for all $d_1, d_2 \geqslant 0$.

In the next section, we will see that this is not the case in general.

Remember that the general model we used to derive these results is valid for arbitrary chemical structures.

5. The autocorrelation descriptor for $E(X) \neq 0$

We have not yet investigated the autocorrelation descriptor for $E(X) \neq 0$. In this section, we use our second random graph model to show that $E(A_1(\mathbf{X}))$ and $E(A_1(\mathbf{Y}))$ are correlated if $E(X), E(Y) \neq 0$.

Let N be an integer-valued random variable. We construct a random graph G as follows: for $N = n$, G is a graph on n vertices $\{1, \ldots, n\}$ whose edges are selected independently with probability $p_n = 2/(n-1)$. Note that isolated vertices may occur. For every fixed n the number of edges is binomially distributed with expectation $E(|D_1||N = n) = \binom{n}{2} p_n = n$. Thus, for a variable number of vertices N, the expected number of edges is $E(|D_1|) = \sum_{n=1}^{\infty} E(|D_1||N = n) P(N = n) = E(N)$. Hence, in the average, graphs have equally many edges and vertices.

Let $\mathbf{X}$, $\mathbf{Y}$ be random vectors having property (iie) and X_k, Y_l be independent for $k \neq l$. Since

$$E(1_{\{(u,v)\in D_1\}}|N = n) = \begin{cases} p_n, & \text{for } u \neq v, \\ 0, & \text{else,} \end{cases}$$

we get

$$E\big(A_1(\mathbf{X})\big) = E(X)^2 \sum_{n=1}^{\infty} 2\binom{n}{2} p_n P(N = n) = 2E(X)^2 E(N) \tag{6}$$

by (3). To determine $E(A_1(\mathbf{X})A_1(\mathbf{Y}))$, consider

$$\begin{aligned} &E\big(A_1(\mathbf{X})A_1(\mathbf{Y})\big|N = n\big) \\ &\quad = E\left(\sum_{u,v,i,j=1}^{n} X_u X_v Y_i Y_j \cdot 1_{\{(u,v)\in D_1\}} \cdot 1_{\{(i,j)\in D_1\}} \Big| N = n \right). \end{aligned} \tag{7}$$

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 $\binom{2}{0} + 2\big(\binom{2}{1} + \binom{2}{2}\big) = 7$ cases of which for symmetry reasons three have different expectations. By independence and linearity (7) thus becomes

$$\begin{aligned} &= \binom{2}{0} E\left(\sum_{\substack{u,v,i,j=1 \\ \text{all}\neq}}^{n} X_u X_v Y_i Y_j \cdot 1_{\{(u,v)\in D_1\}} \cdot 1_{\{(i,j)\in D_1\}} \right) \\ &\quad + 2\binom{2}{1} E\left(\sum_{\substack{u,v,i=1 \\ \text{all}\neq}}^{n} X_u X_v Y_i Y_v \cdot 1_{\{(u,v)\in D_1\}} \cdot 1_{\{(i,v)\in D_1\}} \right) \\ &\quad + 2\binom{2}{2} E\left(\sum_{\substack{u,v=1 \\ u\neq v}}^{n} X_u X_v Y_u Y_v \cdot 1_{\{(u,v)\in D_1\}} \right) \end{aligned}$$

$$
\begin{aligned}
&= E(X)^2 E(Y)^2 \sum_{\substack{u,v,i,j \\ \text{all}\neq}} E(1_{\{(u,v)\in D_1\}}) E(1_{\{(i,j)\in D_1\}}) \\
&\quad + 4E(XY)E(X)E(Y) \sum_{\substack{u,v,i \\ \text{all}\neq}} E(1_{\{(u,v)\in D_1\}}) E(1_{\{(i,v)\in D_1\}}) \\
&\quad + 2E(XY)^2 \sum_{u\neq v} E(1_{\{(u,v)\in D_1\}}) \\
&= E(X)^2 E(Y)^2 \cdot 4! \binom{n}{4} p_n^2 + 4E(XY)E(X)E(Y) \cdot 3! \binom{n}{3} p_n^2 \\
&\quad + 2E(XY)^2 \cdot 2! \binom{n}{2} p_n \\
&= 24E(X)^2 E(Y)^2 \binom{n}{4} p_n^2 + 24E(XY)E(X)E(Y) \binom{n}{3} p_n^2 \\
&\quad + 4E(XY)^2 \binom{n}{2} p_n. \qquad (8)
\end{aligned}
$$

By assumption, (8) holds for independent random vectors $\mathbf{X}$, $\mathbf{Y}$ as well as for $\mathbf{X} = \mathbf{Y}$.

In the following, let $\mathbf{X}$, $\mathbf{Y}$ be independent and $N > 2$. By (4), (8), and an elementary calculation

$$
\begin{aligned}
E\big(A_1(\mathbf{X})A_1(\mathbf{Y})\big) &= \cdots = E(X)^2 E(Y)^2 \sum_{n=3}^{\infty} \frac{4n(n^2-3)}{n-1} P(N=n) \\
&= 4E(X)^2 E(Y)^2 E\left(\frac{N(N^2-3)}{N-1}\right).
\end{aligned}
$$

For the variance, we get

$$
\begin{aligned}
\operatorname{Var}\big(A_1(\mathbf{X})\big) &= E\big(A_1(\mathbf{X})^2\big) - E\big(A_1(\mathbf{X})\big)^2 = \cdots \\
&= 4E(X)^4 E\left(\frac{N^3}{N-1}\right) + 4\big(6\operatorname{Var}(X)E(X)^2 + \operatorname{Var}(X)^2\big) E\left(\frac{N^2}{N-1}\right) \\
&\quad - 4\big(3E(X)^4 + 10\operatorname{Var}(X)E(X)^2 + \operatorname{Var}(X)^2\big) E\left(\frac{N}{N-1}\right) \\
&\quad - \big(2E(X)^2 E(N)\big)^2
\end{aligned}
$$

by (4), (8), (6), and an elementary calculation. Hence, the correlation is

$$
\begin{aligned}
\rho\big(A_1(\mathbf{X}), A_1(\mathbf{Y})\big) &= \frac{4E(X)^2 E(Y)^2 (E(N(N^2-3)/(N-1)) - E(N)^2)}{\sqrt{\operatorname{Var}(A_1(\mathbf{X}))\operatorname{Var}(A_1(\mathbf{Y}))}} \\
&= \frac{E(X)^2 E(Y)^2 \operatorname{Var}(A_1(\mathbf{1}))}{\sqrt{\operatorname{Var}(A_1(\mathbf{X}))\operatorname{Var}(A_1(\mathbf{Y}))}}.
\end{aligned}
$$

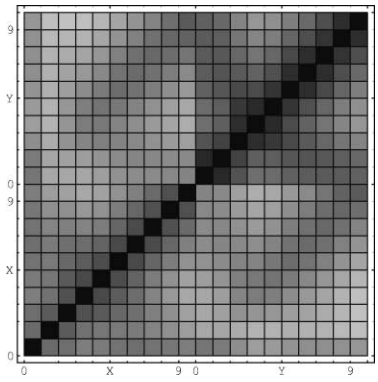


Figure 4. Correlation matrix for $X \sim \mathcal{N}(1, 1)$, $Y \equiv 1$.

Since

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

we get

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

and

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

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 versus infinity. The same is true for A_0; with (3) and (4), it can be shown that $\rho(A_0(\mathbf{X}), A_0(\mathbf{Y})) \to 1$ for $E(X), E(Y) \to \infty$ and $\rho(A_0(\mathbf{X}), A_0(\mathbf{1})) \to 1$ for $E(X) \to \infty$ in the general model.

6. Simulation with chemical structures

To validate our results, we carried out a simulation on 1128 randomly selected structures from the Available Chemicals Directory for a normally distributed property $\mathbf{X}$ and the identity $\mathbf{Y} = (1, 1, \ldots, 1)$. The figures below show the correlation matrices for $A_0(\mathbf{X}), \ldots, A_9(\mathbf{X})$ and $A_0(\mathbf{Y}), \ldots, A_9(\mathbf{Y})$ with $X \sim \mathcal{N}(1, 1)$ (figure 4) and $Y \equiv 1$, and $X \sim \mathcal{N}(0, 1)$ (figure 5) and $Y \equiv 1$, respectively. In figure 5 we also applied modification (5) for $d = 0$. Colors represent absolute values of the matrix entries, ranging from 0.0 (white) to 1.0 (black).

Figure 4 shows considerable correlation among $A_d(\mathbf{X})$ ($d = 0, \ldots, 9$) (lower left quadrant) and among $A_0(\mathbf{Y})$ ($d = 0, \ldots, 9$) (upper right quadrant) as well as between $A_{d_1}(\mathbf{X})$, $A_{d_2}(\mathbf{Y})$ ($d_1, d_2 = 0, \ldots, 9$) (upper left and lower right quadrants). As predicted,

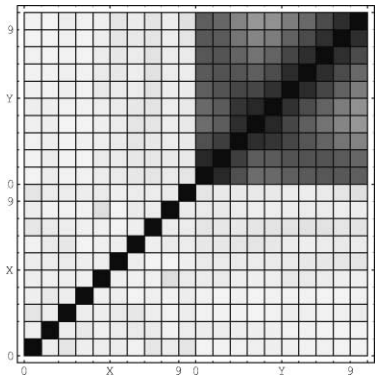


Figure 5. Correlation matrix for $X \sim \mathcal{N}(0, 1)$, $Y \equiv 1$.

no correlation is present in figure 5 among $A_d(\mathbf{X})$ and between $A_{d_1}(\mathbf{X})$, $A_{d_2}(\mathbf{Y})$ since $\mathbf{X}$ is centered ($E(\mathbf{X}) = 0$).

7. Conclusion

Not only does the strong correlation that is present for large values of $|E(X)|$ or $|E(Y)|$ complicate the statistical analysis of the generated data, it also makes meaningful QSAR studies very difficult as there is no clear distinction of contributions from property $\mathbf{X}$ and property $\mathbf{Y}$. This cannot be overcome by factor analysis since factors are linear combinations of all autocorrelation descriptors. Also, the functions $A_0(\mathbf{X})$ and $A_1(\mathbf{X})$ loose all physico-chemical information $\mathbf{X}$ on the vertices as $|E(X)|$ tends versus infinity; thus, $A_0(\mathbf{X})$ and $A_1(\mathbf{X})$ are merely structural descriptors in this case. These shortcomings are easily eliminated if property $\mathbf{X}$ is centered and modification (5) is applied for $d = 0$. The resulting data is uncorrelated and can be analyzed by multivariate statistics or used for QSAR studies. Properties should therefore always be centered and normalized before A_d is applied.

References

[1] H. Kubinyi, *QSAR: Hansch Analysis and Related Approaches* (VCH, 1993).
[2] L.B. Kier and L.H. Hall, *Molecular Connectivity in Structure Activity Analysis* (Wiley, 1986).
[3] N. Trinajstić, *Chemical Graph Theory* (CRC Press, 1992).
[4] D. Bonchev and D.H. Rouvray (eds.), *Chemical Graph Theory*, Vols. 1 and 2 (Gordon & Breach, 1991, 1992).
[5] H. Wiener, Structural determination of parrafin boiling points, J. Am. Chem. Soc. 69 (1947) 17–20.
[6] G. Moreau and P. Broto, Autocorrelation of a topological structure: A new molecular descriptor, Nouv. J. Chim. 4 (1980) 359–360.
[7] J. Devillers and D. Domine, Comparison of reliability of log P values calculated from a group contribution approach and from the autocorrelation method, SAR QSAR Environ. Res. 7 (1997) 195–232.

[8] M. Wagener, J. Sadowski and J. Gasteiger, Autocorrelation of molecular properties for modelling corticosteroid binding globulin and cytosolic Ah receptor activity by neural networks, J. Am. Chem. Soc. 117 (1995) 7769–7775.
[9] J. Devillers, Autocorrelation descriptors for modelling (eco)toxicological endpoints, in: *Topological Indices and Related Descriptors in QSAR and QSPR*, ed. J. Devillers (Gordon & Breach, 1999) pp. 595–612.
[10] H. Bauknecht, A. Zell, H. Bayer, P. Levi, M. Wagener, J. Sadowsky and J. Gasteiger, Locating biologically active compounds in medium-sized heterogeneous datasets by topological autocorrelation vectors: Dopamine and benzodiazepine agonists, J. Chem. Inf. Comput. Sci. 36 (1996) 1205–1213.
[11] J. Devillers, D. Domine and R.S. Boethling, Use of a backpropagation neural network and autocorrelation descriptors for predicting the biodegradability of organic chemicals, in: *Neural Networks in QSAR and Drug Design*, ed. J. Devillers (Academic Press, 1996) pp. 65–82.
[12] `http://www.disat.unimib.it/chm/Dragon.htm`
[13] B. Bollobas, *Random Graphs* (Academic Press, 1984).
[14] E.M. Palmer, *Graphical Evolution* (Wiley, 1985).
[15] B. Hollas, Correlation properties of the autocorrelation descriptor for molecules, Comm. Math. Chem. (MATCH) 45 (2002) 27–33.
[16] J. Köbler, U. Schöning and J. Toran, *The Graph Isomorphism Problem: Its Structural Complexity* (Birkhäuser Verlag, Boston, 1993).