communications in mathematical
match
and in computer chemistry

no. 47, January 2003
MATCDY (47) 79- 86 (2003)

ISSN 0340 - 6253

Correlations in distance-based descriptors

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

(Received August 7, 2002)

Abstract

We introduce random graphs as a concept to model molecules. For any random graph model we show that a class of distance-based topological descriptors (or indices) may be strongly correlated, which makes structure-activity relationship studies more difficult. We show that centering results in uncorrelated descriptors if vertex-properties are independent of the graph. A simulation with chemical structures confirms our findings.

1 Introduction

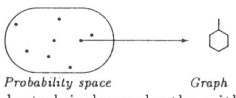
For the needs of computational chemistry a variety of descriptors has been developed. A *descriptor* is a function or an algorithm that accepts a representation of a molecule or an atom as input and outputs some data (real numbers, bitstrings, vectors). Descriptors are used in computational chemistry 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* or *indices* [2] are 2D-descriptors computed from the molecular graph, usually without hydrogen atoms (hydrogen-suppressed-graph). Though these descriptors were the first to be used by chemists, they are still useful in QSAR and, owing to their minimal computational requirements, to analyze virtual combinatorial libraries or large chemical databases.

In this paper, we represent the molecular structure as a graph G and properties of vertices or atoms as real values assigned to the vertices of G. In applications, these may be physicochemical properties of atoms or graph-theoretical properties of vertices. The descriptors we analyze here are of the form

$$A_d = \frac{1}{2} \sum_{(u,v) \in D_d} x_u x_v \tag{1}$$

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

whereby $D_d = \{(u,v) \mid d(u,v) = d\}$ is 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.
As a distance-based function, A_d is invariant for different labellings of G, hence, (1) can be defined as the descriptor of the molecule corresponding to G.
Many common descriptors are of the form (1). Let $x_u = \deg(u)^\alpha$, whereby $\deg(u)$ is the *degree* (valence) of vertex u. For $\alpha = 1$, (1) is called *Zagreb group index* [3]. Randic [4] considered the cases $\alpha = -1$ and $\alpha = -1/2$; especially, in the latter case, (1) is called *Randic connectivity index*, which is one of the most often applied topological indices in QSAR-studies. The Randic index measures the branching of the carbon skeleton of a molecule. If x_u is some physico-chemical property of atom u, (1) is called *autocorrelation descriptor* [5]. Devillers [6, 7] and Gasteiger [8, 9] have used this descriptor for toxicological and pharmaceutical research, and for clustering and similarity analysis of chemical data.
To analyze correlation properties of descriptor (1), we model molecules as *random graphs* [10, 11] that have a random variable associated with each vertex. While randomly generated were already applied to computational chemistry [12, 13], the concept of random graphs has for long only been used in mathematics and theoretical physics.
In the most general case, a random graph can be defined as a probability distribution on a set of graphs, or, equivalently, as a random element mapping to a set of graphs.



In [14], we used graphs on n vertices whose edges are selected independently with a probability p_n so that the expected number of edges equals n. For this random graph model G_{n,p_n} and distance $d = 1$ we provided explicit formulas for the correlation of descriptor (1).
In this paper, we use a distribution-free model that does not make any assumptions on the random graphs. That is, the results we derive are true for any probability distribution on any set of graphs. Especially, this model is valid for all chemical structures.
To model 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

$$A_d(\mathbf{X}) = \frac{1}{2} \sum_{(u,v) \in D_d} X_u X_v \,, \quad \mathbf{X} = (X_1, \ldots, X_N)$$

and D_d is now a random set describing the random graph. 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, $u = 1, \ldots, N$. The number of vertices N is an arbitrarily distributed random variable.
For the vertex properties, we assume that for every fixed N, $X_1, \ldots, X_N$ are independent with same expectation $E(X)$ and independent of D_d, i.e. independent of the graphical structure.
In chemistry, vertex properties are more or less dependent on each other and on the graphical structure. However, we show that even with these idealistic assumptions, descriptors may be strongly correlated.

2 Analysis with random graphs

Let N be any integer-valued random variable and $X_1, \ldots, X_N$, $Y_1, \ldots, Y_N$ be independent and independent of D_d for every fixed N. Also, let $E(X) = E(X_i)$, $E(Y) = E(Y_i)$ be the respective expectations. By $\mathbf{1}$ we denote the constant vector $(1, \ldots, 1)$.

In section 4 we prove for all $d \geq 0$

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

and

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

Remember that a correlation $\rho = 1$ means that the two random variables are linearly dependent. These results show that $A_d(\mathbf{X})$, $A_d(\mathbf{Y})$, and $|D_d| = 2A_d(\mathbf{1})$ are strongly correlated for large values of $|E(X)|$ and $|E(Y)|$ even if properties $\mathbf{X}$ and $\mathbf{Y}$ are independent. This correlation converges rapidly. In [14], we show for random graph model G_{n,p_n} that $\rho\,(A_1(\mathbf{X}), A_1(\mathbf{1})) > 0.8$ for $|E(X)| > 4$ if the edge probability p_n is chosen in a way such that the expected number of edges equals the number of vertices. If $\mathbf{X}$ is a physico-chemical property, then $A_d(\mathbf{X})$ contains almost only structural information, all physico-chemical information on the vertices is lost as $|E(X)|$ tends versus infinite.
In contrast, descriptors are uncorrelated if vertex properties are centered, i.e. if $E(X) = 0$.
In section 3 we prove that for $E(X) = 0$ holds:

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 $Var(\mathbf{X}) = \mathbf{1}$ and

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

   is used instead of $A_0(\mathbf{X})$ then this is true for all $d_1, d_2 \geq 0$.

The non-mathematical reader may skip the following sections 3 and 4.

3 The case $E(X) = 0$

For simplicity, we omit factor $\frac{1}{2}$ in (1) in the following calculations.
As usual, we denote by $L_p = \{X \mid E(|X^p|) < \infty\}$ the set of p-times integrable random variables. Let D_1 be a random set of edges, N the number of vertices of the graph defined by D_1 (without isolated vertices), 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_1 with $E(X_u) = E(X_1) \quad (u = 1, \ldots, N)$.

For ease of notation, we write $E(X)$ instead of $E(X_1)$. Let

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

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

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

and

$$E(A_d(\mathbf{X})) = \sum_{n=1}^{\infty} E(A_d(\mathbf{X}|N=n))\, P(N=n)$$

$$= \sum_{n=1}^{\infty} E\left(\sum_{u,v=1}^{n} X_u X_v \cdot 1_{\{(u,v)\in D_d\}} \mid N=n \right) P(N=n)$$

(Remember that the conditional expectation $E(X|A)$ for a random variable X and an event A is defined as $\frac{1}{P(A)} \int_A X dP$, which requires to multiply the sum by $P(N=n)$.)
If $d > 0$, this equals

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

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

$$E(A_{d_1}(\mathbf{X}) A_{d_2}(\mathbf{Y})) = \sum_{n=1}^{\infty} E(A_{d_1}(\mathbf{X}) A_{d_2}(\mathbf{Y}) \mid N=n)\, P(N=n)$$

$$= \sum_{n=1}^{\infty} E\left(\sum_{u,v,i,j=1}^{n} X_u X_v Y_i Y_j \cdot 1_{\{(u,v)\in D_{d_1}\}} \cdot 1_{\{(i,j)\in D_{d_2}\}} \mid N=n \right) P(N=n) \tag{4}$$

To determine $Cov(A_{d_1}(\mathbf{X}), A_{d_2}(\mathbf{Y}))$ for distances $d_1, d_2 \geq 0$ 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,

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

and

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

Hence,

$$Cov(A_0(\mathbf{X}), A_0(\mathbf{Y})) = \sum_{n=1}^{\infty} n^2 E(X^2) E(Y^2) P(N=n)$$

$$- \sum_{n=1}^{\infty} nE(X^2) P(N=n) \cdot \sum_{n=1}^{\infty} nE(Y^2) P(N=n)$$

$$= E(X^2) E(Y^2) E(N^2) - E(X^2) E(N) E(Y^2) E(N)$$

$$= E(X^2) E(Y^2)\, Var(N) > 0$$

if $X, Y \neq 0$, $N \not\equiv c$ for a constant c.

If however we apply modification (2) then

$$Cov\left(\tilde{A}_0(\mathbf{X}), A_0(\mathbf{Y})\right) = E(X^2 - 1) E(Y^2)\, Var(N) = 0$$

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

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

$$E(A_{d_1}(\mathbf{X})A_{d_2}(\mathbf{Y}) \mid N = n) =$$

$$E(X)\,E\left(\sum_{u,v,i,j=1}^{n} X_v Y_i Y_j \cdot 1_{\{(u,v)\in D_{d_1}\}} \cdot 1_{\{(i,j)\in D_{d_2}\}} \mid N = n\right) = 0$$

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

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

$$Cov\Big(\tilde{A}_0(\mathbf{X}), A_{d_1}(\mathbf{Y})\Big) = 0$$

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

4 The case $E(X), E(Y) \neq 0$

Let $N > 0$, $d > 0$, $N \in L_2$ and let $\mathbf{X}, \mathbf{Y}$ be independent random vectors having property (iie).
By (3),

$$E(A_d(\mathbf{X})) = E(X)^2\,E(A_d(\mathbf{1})) \tag{5}$$

By (4),

$$E(A_d(\mathbf{X})A_d(\mathbf{Y})) = E(X)^2\,E(Y)^2\,E\Big(A_d(\mathbf{1})^2\Big)$$

Hence, by (5),

$$\begin{aligned} Cov(A_d(\mathbf{X}), A_d(\mathbf{Y})) &= E(X)^2\,E(Y)^2\,Cov(A_d(\mathbf{1}), A_d(\mathbf{1})) \\ &= E(X)^2\,E(Y)^2\,Var(A_d(\mathbf{1})) \end{aligned} \tag{6}$$

To determine $E(A_d(\mathbf{X})^2)$, consider

$$E\Big(A_d(\mathbf{X})^2 | N = n\Big)$$

$$= E\left(\sum_{u,v,i,j=1}^{n} X_u X_v X_i X_j \cdot 1_{\{(u,v)\in D_d\}} \cdot 1_{\{(i,j)\in D_d\}} \mid N = n\right) \tag{7}$$

We dissect the sum according to $|\{u,v\} \cap \{i,j\}| = k$ for $k = 0, 1, 2$. For symmetry reasons, these cases each have same expectation. By independence and linearity (7) thus becomes

$$= \binom{2}{0} E\left(\sum_{\substack{u,v,i,j=1 \\ \{u,v\}\cap\{i,j\}=\emptyset}}^{n} X_u X_v X_i X_j \cdot 1_{\{(u,v)\in D_d\}} \cdot 1_{\{(i,j)\in D_d\}} \mid N = n\right)$$

$$+2\binom{2}{1} E\left(\sum_{\substack{u,v,i=1 \\ u\neq i}}^{n} X_u X_v^2 X_i \cdot 1_{\{(u,v)\in D_d\}} \cdot 1_{\{(i,v)\in D_d\}} \mid N = n\right)$$

$$+2\binom{2}{2}E\left(\sum_{u,v=1}^{n} X_u^2 X_v^2 \cdot 1_{\{(u,v)\in D_d\}} \mid N=n\right)$$

$$= E(X)^4 \sum_{\substack{u,v,i,j \\ \{u,v\}\cap\{i,j\}=\emptyset}}^{n} E\left(1_{\{(u,v)\in D_d\}} \cdot 1_{\{(i,j)\in D_d\}} \mid N=n\right)$$

$$+4E\left(X^2\right)E(X)^2 \sum_{\substack{u,v,i \\ u\neq i}}^{n} E\left(1_{\{(u,v)\in D_d\}} \cdot 1_{\{(i,v)\in D_d\}} \mid N=n\right)$$

$$+2E\left(X^2\right)^2 \sum_{u,v=1}^{n} E\left(1_{\{(u,v)\in D_d\}} \mid N=n\right)$$

It follows that

$$\frac{E(A_1(\mathbf{X})^2|N=n)}{E(X)^4} \to E\left(A_1(\mathbf{1})^2|N=n\right) \quad (E(X)\to\pm\infty)$$

converges monotonously and we may apply the monotonous convergence theorem, giving

$$\frac{E(A_d(\mathbf{X})^2)}{E(X)^4} \to E\left(A_d(\mathbf{1})^2\right) \quad (E(X)\to\pm\infty)$$

By (5),

$$\frac{Var(A_d(\mathbf{X}))}{E(X)^4} \to Var(A_d(\mathbf{1})) \quad (E(X)\to\pm\infty)$$

Finally, for the correlation follows by (6)

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

and

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

With a very similar calculation it can be seen that these results also hold for $d = 0$.

5 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 constant $\mathbf{Y} = (1, 1, \ldots, 1)$. The figures below show the correlation matrices $\rho((A_0(\mathbf{X}), \ldots A_9(\mathbf{X})), (A_0(\mathbf{Y}), \ldots, A_9(\mathbf{Y})))$ with $X \sim \mathcal{N}(1,1)$ [1] and $Y \equiv 1$ (figure 1), and $X \sim \mathcal{N}(0.1)$ and $Y \equiv 1$ (figure 2), respectively. In figure 2 we also applied modification (2) for $d = 0$. Shades represent absolute values of the matrix entries, ranging from 0.0 (white) to 1.0 (black). Note that in both charts matrix position (1,1) is in the lower left corner.

Figure 1 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, no correlation is present in figure 2 among $A_d(\mathbf{X})$ and between $A_{d_1}(\mathbf{X}), A_{d_2}(\mathbf{Y})$ since $\mathbf{X}$ is centered $(E(\mathbf{X}) = 0)$.

[1] normal distribution with mean 1 and variance 1

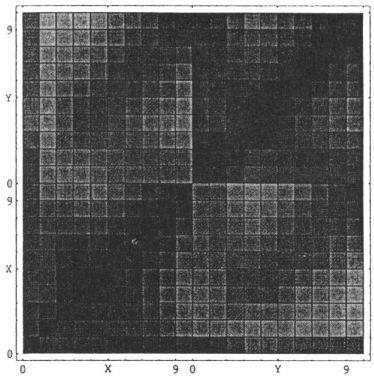


Figure 1: Correlation matrix for $X \sim \mathcal{N}(1,1), Y \equiv 1$

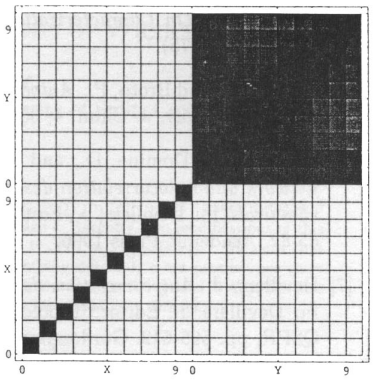


Figure 2: Correlation matrix for $X \sim \mathcal{N}(0,1), Y \equiv 1$

6 Discussion

We proved that descriptors $A_d(\mathbf{X})$ and $A_d(\mathbf{Y})$ or $A_d(\mathbf{1})$ are strongly correlated for large values of $|E(X)|$ and $|E(Y)|$. Correlated descriptors complicate the statistical analysis of the generated data and make QSAR studies very difficult as there is no clear distinction of contributions from property $\mathbf{X}$ and property $\mathbf{Y}$.

Principal component analysis (PCA) is a method from multivariate statistics to obtain uncorrelated principal components from correlated data and to reduce the number of dimensions. However, principal components are linear combinations of all components. Thus, applying PCA to strongly correlated descriptors results in essentially one linear combination (the first principal component) of these descriptors, which is of little use for QSAR.

We proved that descriptors $A_d(\mathbf{X})$ and $A_d(\mathbf{Y})$ or $A_d(\mathbf{1})$ are uncorrelated for $E(X) = 0$. This does not mean that centering properties (the step $X \to X - \bar{X}$ whereby $\bar{X}$ is the mean of X) always yields uncorrelated descriptors: since the assumption that $X_1, \ldots, X_N$ are independent and independent of the graphical structure does not hold for chemical descriptors, we cannot expect them to become uncorrelated. However, it is reasonable to assume that centering decreases correlations.

References

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

[2] Kier, L. B. and Hall, L. H.: Molecular Connectivity in Structure Activity Analysis. Wiley, 1986.

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

[4] Randic, M.: *On Characterization of Molecular Branching.* J. Am. Chem. Soc. 97, 6609-6615 (1975),

[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 Boethling, R. S.: *Use of a backpropagation neural network and autocorrelation descriptors for predicting the biodegradability of organic chemicals.* In: Devillers, J., Editor: Neural Networks in QSAR and Drug Design, pp. 65-82. Academic Press 1996.

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

[8] 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).

[9] 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).

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

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

[12] Gutman, I., Soldatović, T., Vidović, D.: *The energy of a molecular graph and its size dependence. A Monte Carlo approach.* Chem. Phys. Letters 297, 428-432 (1998)

[13] Gutman, I., T., Vidović, D.: *Quest for molecular graphs with maximal energy: a computer experiment.* J. Chem. Inf. Comput. Sci. 41, 1002-1005 (2001)

[14] Hollas, B.: *Correlation Properties of the Autocorrelation Descriptor for Molecules.* Comm. in Math. and Comp. Chem. (MATCH) 45, 27-33 (2002)