Journal of Mathematical Chemistry Vol. 38, No. 3, October 2005 (© 2005)
DOI: 10.1007/s10910-005-6474-5

Asymptotically independent topological indices on random trees

Boris Hollas

Theoretische Informatik, Universität Ulm, D-89081 Ulm, Germany
E-mail: hollas@informatik.uni-ulm.de

Received 24 November 2003; revised 19 May 2005

Topological indices are graph invariants used in computational chemistry to encode molecules. A frequent problem when performing structure-activity studies is that topological indices are inter-correlated. We consider a simple topological index and show asymptotic independence for a random tree model. This continues previous work on the correlation among topological indices. These findings suggest that a size-dependence in a certain class of distance-based topological indices can be eliminated.

KEY WORDS: Topological indices, asymptotic independence, random graphs

AMS subject classification: 05C80, 60E10, 92E10

1. Introduction

An important class of molecular descriptors used by computational chemists are topological indices. *Topological indices* are graph invariants that are derived from the molecular graph, usually the hydrogen-depleted molecular graph [1–4]. Such a graph represents atoms and bonds in a molecule, regardless of distances between atoms, bond and torsion angles, and other parameters representing the threedimensional molecular geometry. Topological indices are either a function of the molecular graph only *(topostructural indices)* or also encode information on chemical properties of atoms *(topochemical indices)* [5]. For example, the most frequently used molecular descriptor, the *Randić index* [6]

$$\sum_{\text{adjacent } v,w} \deg(v)\deg(w)$$

is a topostructural index while *Moreau–Broto-autocorrelation* [7]

$$\sum_{\text{adjacent } v,w} p_v p_w,$$

whereby p_v, p_w are quantitative chemical properties of atoms v, w, is a topochemical index. The latter index is also used for pairs of atoms having distances (number of bonds between) $d > 1$.

Topological indices are used to characterize similarity of molecules and to predict physical, chemical or biological activities or properties [8]. Since topological indices can readily be computed using very little computation time they are especially suited to screen large virtual libraries, a common task in computer aided drug design.

The methods to relate the structure of a molecule to a specific activity or property are known as quantitative structure-activity relationship (QSAR) or quantitative structure-property relationship (QSPR) [9]. A frequent problem is that molecular descriptors are inter-correlated which makes QSAR/QSFR studies difficult or even impossible and raises doubt concerning the meaning of a large number of descriptors [10], In this paper we consider a simple topological index and show asymptotic independence for non-cyclic structures.

2. Preliminaries

Let $D_d = D_d(G) = \{(v, w)|v < w \wedge d(v, w) = d\}$ be the set of ordered pairs of vertices that have distance $d > 0$ in graph G and for all $v \in V$ let X_v be a variable associated with v. Many topological indices have the form

$$A_d(\mathbf{X}) = \sum_{(v,x)\in D_d} X_v X_w,$$

whereby $\mathbf{X} = (X_1, \ldots, X_n)$ is the vector of vertex-properties.

In [11–13], we used random graph models to investigate correlations among these indices. *A random graph model* is, in the most general case, a set of graphs together with a probability distribution defined on it.

For any random graph model, we proved the following [13]: let, for all vertices v, w, Y_v, Y_w be independent random variables that are independent of the graphical structure, and let $E(X)$ and $E(Y)$ be the common expectations of X_v and Y_v, respectively. For all distances $d > 0$ then holds

1. $A_d(\mathbf{X}), A_d(\mathbf{Y})$ are uncorrelated iff $E(X) = 0$ or $E(Y) = 0$;
2. $A_d(\mathbf{X}), A_d(\mathbf{Y})$ are linearly dependent for $E(X), E(Y) \to \pm\infty$.

However, uncorrelated random variables may still be dependent, even functionally dependent. Thus, the above result does not clarify if the mutual dependence, as expressed e.g., by the mutual information [14], is reduced for $E(X) = 0$ and to what extent. In this paper, we use characteristic functions to show asymptotic independence.

The *characteristic function* $\varphi\mathbf{X} : \mathbb{R}^k \to \mathbb{C}$ of a k-dimensional random vector $\mathbf{X}$ is defined as

$$\varphi_{\mathbf{X}}(\mathbf{X}) = E(e^{i\mathbf{x}\cdot\mathbf{X}})$$

whereby $\cdot$ denotes the scalar product. Note that $\varphi_{\mathbf{X}}$ is always finite since $|\varphi_{\mathbf{X}}(x)| \leqslant 1$. Characteristic functions have the following important properties [15,16]:

(1) φ_{X} is characteristic for X, i.e. $\varphi_X = \varphi_Y$ iff $X \sim Y$;

(2) $\varphi_{X_n} \to \varphi_X \Longleftrightarrow X_n \overset{\mathscr{L}}{\to} X$;

(3) for independent random variables X, Y holds $\varphi_{\mathrm{X+Y}} = \varphi_{\mathrm{X}} + \varphi_Y$(the converse is not true however);

(4) random variables X, Y are independent iff for all $x, y\varphi(X, Y)^{(x,y)} = \varphi_{\mathrm{X}}(x)\varphi_Y(y)$.

3. The random tree model

Let $(\mathcal{T}, P)$ be a probability space of trees whose number of vertices is distributed according to a random variable N, that is, for every tree $T \in \mathcal{T}$ holds $P(T$ has n vertices$) = P(N = n)$. We require that $N > 1$.

Furthermore, let $X_1, \ldots, X_N$ be random variables such that

(1) $X_1, \ldots, X_N$ are independent and uniformly distributed on $\{-1, 1\}$,

(2) $X_1, \ldots, X_N$ are independent of D_1.

Since we are going to use property (3) of characteristic functions, we need that

$$\sum_{(v,w)\in D_1(T)} X_v X_w$$

is a sum of independent random variables for every fixed $T \in \mathcal{T}$. While for any graph $G = (V, D_1)(X_v X_w)_{(v,w)\in D_1}$ are pairwise independent, these random variables are not independent for cyclic graphs: consider $G = K_3$ and let be $X_1X_2 = X_2X_3 = 1$. Then $X_1 = X_2 = X_3$, hence $X_1X_3 = 1$. However, independence holds for trees:

Lemma 1. For every tree T, $(X_v X_w)_{(v,w)\in D_1(T)}$ are independent.

Proof. Since T is a tree, let be w.l.o.g. 1 a leaf and $(1, 2) \in D_1$. For all $(v, w)D_1$, let be $x_{v,w} \in \{-1, 1\}$. Then

$$\begin{aligned}
&P(X_1X_2 = 1 \wedge \forall(v, w) \in D_1 X_v X_w = x_{v,w})\\
&\quad = P(X_1 = X_2 = 1 \wedge \forall(v, w) \in D_1 X_v X_w = x_{v,w})\\
&\qquad + P(X_1 = X_2 = -1 \wedge \forall(v, w) \in D_1 X_v X_w = x_{v,w})
\end{aligned}$$

since $P(X_1 = \pm 1) = 1/2 = P(X_1X_2 = 1)$, we get

$$\begin{aligned}
&= P(X_1X_2 = 1)(P(X_2) = 1 \wedge \forall(v, w) \in D_1 X_v X_w = x_{v,w})\\
&\quad + P(X_2 = -1 \wedge \forall(v, w) \in D_1 X_v X_w = x_{v,w}))\\
&= P(X_1X_2 = 1)P(\forall(v, w) \in D_1 X_v X_w = x_{v,w})\\
&= \cdots = \prod_{(v,w)\in D_1} P(X_v X_w = x_{v,w}).
\end{aligned}$$

For $P(X_1X_2 = -1)$, the result follows accordingly. Thus, for all $M \subset D_1$ follows

$$P\left(\bigcup_{(v,w)\in M} \{X_v X_w = x_{v,w}\}\right) = \prod_{(v,w)\in M} P(X_v X_w = x_{v,w})$$

since $M = D_1(F)$ for a forest F. □

As an unexpected consequence, we get:

Corollary 2. Let T_1, T_2 be trees on the same number of vertices. Then

$$\sum_{(v,w)\in D_1(T_1)} X_v X_w \quad \text{and} \quad \sum_{(v,w)\in D_1(T_1)} X_v X_w$$

have the same distribution.

Proof. The distribution of the sum depends on the number of summands only. □

4. Asymptotic independence

Lemma 3 is crucial to show asymptotic normality and independence in the main theorem. Note that $e^{-1/2x^2}$ is the characteristic function of the normal distribution.

Lemma 3. For all $x \in \mathbb{R}$, $\lim_{n\to\infty} \left(\cos \frac{x}{\sqrt{n}}\right)^n = e^{-1/2x^2}$.

Proof. By Taylor's theorem,

$$\cos(x) = 1 - \frac{1}{2}x^2 + r(x)$$

with $r(x) = \sin(\xi)$ for $\xi \in (0, x)$. For any $\varepsilon > 0$ there is an n such that

$$r\left(\frac{x}{\sqrt{n}}\right) = \left| r\left(\frac{x}{\sqrt{n}}\right)\right| \leqslant \frac{x^4}{6n^2} < \frac{\varepsilon}{n}.$$

Hence,

$$\left(1 - \frac{x^2}{2n}\right)^n \leqslant \left(\cos\frac{x}{\sqrt{n}}\right)^n \leqslant \left(1 - \frac{x^2}{2n} + \frac{\varepsilon}{n}\right)^n$$

For $n \to \infty$, we get

$$\mathrm{e}^{-\frac{1}{2}x^2} \leqslant \lim_{n\to\infty}\left(\cos\frac{x}{\sqrt{n}}\right)^n \leqslant \mathrm{e}^{-\frac{1}{2}x^2+\varepsilon}$$

The assertion follows for $\varepsilon \to 0$. □

Throughout this section, let $X_1, \ldots, X_N, Y_1, \ldots, Y_N$ be independent random variables with properties (1) and (2) from section 3. Then holds

Lemma 4. Indices $B(\mathbf{X})$, $B(\mathbf{Y})$ are uncorrelated.

Proof. Write

$$B(\mathbf{X}) = \sum_{v,w} X_v X_w \frac{1_{\{(v,w)\in D_1\}}}{\sqrt{N-1}},$$

whereby $1_{\{(v,w)\in D_1\}}$ is the indicator function for event $\{(v, w) \in D_1\}$. Then, by linearity and independence, $E(B(\mathbf{X})) = 0$ and $E(B(\mathbf{X})B(\mathbf{Y})) = 0$, hence $\rho(B(\mathbf{X}), B(\mathbf{Y})) = 0$. □

Thus, both $A(\mathbf{X})$, $A(\mathbf{Y})$ and $B(\mathbf{X})$, $B(\mathbf{Y})$ are uncorrelated, The following two theorems show what difference the factor $1/\sqrt{N-1}$ makes in terms of independence.

Theorem 5. Indices $B(\mathbf{X})$, $B(\mathbf{Y})$ are asymptotically normal and independent for $E(N) \to \infty$ and $\mathrm{Var}(N) \in O(E(N)^\alpha), \alpha < 1$.

Proof. We introduce a notation first. For a random vector X and an event A, let $\varphi_{(X|A)}(\mathrm{x})$ denote the conditional expectation, $E(\mathrm{e}^{i\mathrm{x}.X}|A)$.

$$\varphi_{(B(\mathbf{X}),B(\mathbf{Y}))}(x, y) = \sum_{n>1} \varphi_{(B(\mathbf{X}),B(\mathbf{Y})|N=n)}(x, y)P(N = n)$$

$$= \sum_{n>1} E\left(\exp \sum_{(v,w)\in D_1(T)} \left(i\frac{x}{\sqrt{n-1}}X_vX_w + i\frac{y}{\sqrt{n-1}}Y_vY_w\right)\right) P(N=n).$$

By lemma 1 and corollary 2, we get

$$= \sum_{n>1} E\left(\left(\exp\left(i\frac{x}{\sqrt{n-1}}X_1X_2 + i\frac{y}{\sqrt{n-1}}Y_1Y_2\right)\right)^{n-1}\right) P(N=n)$$

$$= \sum_{n>1} \left(\varphi_{X_1X_2}\left(\frac{x}{\sqrt{n-1}}\right)\right)^{n-1} \left(\varphi_{Y_1Y_2}\left(\frac{x}{\sqrt{n-1}}\right)\right)^{n-1} P(N=n).$$

Since $P(X_1X_2 = \pm 1) = 1/2$,

$$\varphi_{X_1X_2}(x) = \frac{1}{2}\mathrm{e}^{-ix} + \frac{1}{2}\mathrm{e}^{ix} = \cos(x)$$

hence, the sum above is

$$= \sum_{n>1} \underbrace{\left(\cos\left(\frac{x}{\sqrt{n-1}}\right)\right)^{n-1} \left(\cos\left(\frac{x}{\sqrt{n-1}}\right)\right)^{n-1}}_{:=f_n(x)} P(N=n).$$

Next, we show that this sum converges to $\mathrm{e}^{-1/2(x^2+y^2)}$. With the triangular inequality follows

$$\left|\varphi_{(B(\mathbf{X}).B(\mathbf{Y}))}(x,y) - \mathrm{e}^{-\frac{1}{2}(x^2+y^2)}\right|$$

$$\leqslant \sum_{n=2}^{\lfloor E(N)/2\rfloor} P(N=n) + \sum_{\lfloor E(N)/2\rfloor}^{\infty} |f_n(x) - \mathrm{e}^{-\frac{1}{2}(x^2+y^2)}| P(N=n)$$

$$\leqslant \lfloor E(N)/2\rfloor P(|N-E(N)| \geqslant E(N)/2) + \max_{n\geqslant E(N)/2} |f_n(x) - \mathrm{e}^{-\frac{1}{2}(x^2+y^2)}|$$

$$\leqslant \frac{2\mathrm{Var}(N)}{E(N)} + \max_{n\geqslant E(N)/2} |f_n(x) - \mathrm{e}^{-\frac{1}{2}(x^2+y^2)}|$$

by Chebyshev's inequality. By lemma 3, $f_n(x) \to \mathrm{e}^{-1/2(x^2+y^2)}$, thus

$$\left|\varphi_{(B(\mathbf{X}),B(\mathbf{Y}))}(x,y) - \mathrm{e}^{-1/2(x^2+y^2)}\right| \to 0$$

for $E(N) \to \infty$ and $Var(N) \in O(E(N)^\alpha), \alpha < 1$. Since $\mathrm{e}^{-\frac{1}{2}(x^2+y^2)}$ is the characteristic function of the bivariate normal distribution, we have shown that $B(\mathbf{X})$

(and $B(\mathbf{Y})$) are asymptotically normal (the marginal distributions of a bivariate normal distribution are normal). Further, random variables that are uncorrelated and whose joint distribution is the bivariate normal distribution are independent. Hence, the assertion follows by lemma 4. $\square$

Theorem 6. The assertions of theorem 5 do not hold for the untransformed indices $A(\mathbf{X}),A(\mathbf{Y})$.

Proof. From the proof of theorem 5 follows for $A(\mathbf{X}),A(\mathbf{Y})$

$$\varphi_{A(\mathbf{X})}(x) = \sum_{n>1} (\cos(x))^{n-1} P(N = n)$$

and

$$\varphi_{(A(\mathbf{X}),A(\mathbf{Y}))}(x, y) = \sum_{n>1} (\cos(x)\cos(y))^{n-1} P(N = n)$$

As distribution for $N = N_k$ we choose

$$P(N_k = n) = \left(\frac{1}{2}\right)^{n-1-k}, \quad n \geqslant k + 2.$$

Thus, $E(N) \to \infty$ is equivalent to $k \to \infty$ and $Var(N_k)$ is constant. We get

$$\begin{aligned}
\varphi_{A(\mathbf{X})}(x) &= \sum_{n=k+2}^{\infty} (\cos(x))^{n-1} \left(\frac{1}{2}\right)^{n-1-k} \\
&= 2^k \sum_{n=k+1}^{\infty} \left(\frac{1}{2}\cos(x)\right)^n \\
&= 2^k \frac{\left(\frac{1}{2}\cos(x)\right)^{k+1}}{1 - \frac{1}{2}\cos(x)} \\
&= \frac{(\cos(x))^{k+1}}{2 - \cos(x)}
\end{aligned}$$

by the formula for the sum of a geometric progression. Thus, $B(\mathbf{X})$ is not asymptotically normal. Accordingly,

$$\varphi_{(A(\mathbf{X}),A(\mathbf{Y}))}(x, y) = \frac{(\cos(x)\cos(y))^{k+1}}{2 - \cos(x)\cos(y)}.$$

Since

$$\frac{\varphi_{A(\mathbf{X})}\varphi_{A(\mathbf{Y})}}{\varphi_{(A(\mathbf{X}),A(\mathbf{Y}))}} = \frac{2 - \cos(x)\cos(y)}{(2 - \cos(x))(2 - \cos(y))} \not\equiv 1$$

indices $A(\mathbf{X})$, $A(\mathbf{Y})$ are not independent. $\square$

5. Discussion

Theorems 5 and 6 show that the factor $1/\sqrt{N-1}$ further reduces dependence among the already uncorrelated indices $B(\mathbf{X})$, $B(\mathbf{Y})$ in our random tree model, This makes this transform interesting for further research. However, our model has two shortcomings:

1. graphs must have no cycles, whereas chemical graphs may contain cycles;
2. the random variables X_v, Y_v assume only two distinct values.

An essential prerequisite for the proof of theorem 5 is that $\sum X_v X_w$ is a sum of independent random variables, as shown in lemma 1, This does not hold for arbitrary graphs: generalizing the example in section 3, it is easy to show that in any cycle $C_n X_1 X_n$ is a function of $\sum_{v=1}^{n-1} X_v X_{v+1}$. Also, if $X_1, \ldots, X_n$ are not uniformly distributed on $\{-1, 1\}$, then $(X_v X_w)_{\{(v,w)\in D_1\}}$ may not be independent even on trees: let $X_1, \ldots, X_n$ be i.i. d. with $P(X_1 = 0) = p > 0$. If $(1,2), (1,3) \in D_1$ then X_1X_2, X_1X_3 are not independent. However, these results suggest that the factor $1/\sqrt{|D_1|}$ eliminates an otherwise present *size-dependence* in general graphs (including cyclic graphs) for arbitrary random variables X_v that are symmetrically distributed with mean 0. The size of a graph can be coded as a separate descriptor.

References

[1] N. Trinajstić, *Chemical Graph Theory* (CRC Press, Boca, Raton, 1992).
[2] D. Bonchev and D. Rouvray, editors. *Chemical Graph Theory* (Abacus Press/Gordon & Breach, Amsterdam. 1991).
[3] D. Rouvray, ed. *Computational Chemical Graph Theory* (Nova Press, 1990).
[4] R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors* (Wiley, New York, 2000).
[5] S. Basak, B. Gute and G. Grunwald, Use of topostructural, topochemical, and geometric parameters in the prediction of vapor pressure. J. Chem. Inf. Comput. Sci. 39 (1999) 255.
[6] M. Randlć, On characterization of molecular branching. J. Am. Chem. Soc., 97 (1975) 6609.
[7] G. Moreau and P. Broto. Autocoxrelation of a topological structure: A new molecular descriptor. Nouv. J. Chim. 4 (1980) 359.
[8] J. Devillers and A. Balaban (ed.), *Topological Indices and Related Descriptors in QSAR and QSPR* (Gordon & Breach, Amsterdam, 1999).
[9] M. Karelson, *Molecular Descriptors in QSAR/QSPR* (Wiley, New York, 2000).
[10] I. Motoc, A. Balaban, O. Mekenyan and D. Bonchev, Topological indices: Inter-relations and composition, MATCH – Commun. Math. Comput. Chem. 13 (1982) 369.
[11] B. Hollas, Correlation properties of the autocorrelation descriptor for molecules. MATCH – Commun. Math. Comput. Chem. 45 (2002) 27.
[12] B. Hollas. An analysis of the autocorrelation descriptor for molecules. J. Math. Chem. 33(2) (2003) 91.
[13] B. Hollas. Correlations in distance-based descriptors, MATCH – Commun. Math. Comput. Chem. 47 (2003) 79.
[14] R. Roman, *Coding and Information Theory* (Springer, Berlin, 1992).

[15] H. Bauer, *Probability Theory* (Gruyter, 1996).
[16] W. Feller, *An Introduction to Probability Theory and its Applications* (Wiley, New York, 1970).