J Math Chem (2013) 51:2718–2730
DOI 10.1007/s10910-013-0233-9

ORIGINAL PAPER

Information content of molecular graph and prediction of gas phase thermal entropy of organic compounds

Chandan Raychaudhury · Debnath Pal

Received: 12 April 2013 / Accepted: 16 July 2013 / Published online: 27 July 2013

Abstract Entropy is a fundamental thermodynamic property that has attracted a wide attention across domains, including chemistry. Inference of entropy of chemical compounds using various approaches has been a widely studied topic. However, many aspects of entropy in chemical compounds remain unexplained. In the present work, we propose two new information-theoretical molecular descriptors for the prediction of gas phase thermal entropy of organic compounds. The descriptors reflect the bulk and size of the compounds as well as the gross topological symmetry in their structures, all of which are believed to determine entropy. A high correlation ($r^2 = 0.92$) between the entropy values and our information-theoretical indices have been found and the predicted entropy values, obtained from the corresponding statistically significant regression model, have been found to be within acceptable approximation. We provide additional mathematical result in the form of a theorem and proof that might further help in assessing changes in gas phase thermal entropy values with the changes in molecular structures. The proposed information-theoretical molecular descriptors, regression model and the mathematical result are expected to augment predictions of gas phase thermal entropy for a large number of chemical compounds.

Keywords Thermal entropy · Molecular descriptor · Information content · Regression model

C. Raychaudhury · D. Pal (✉)
Bioinformatics Centre and Supercomputer Education and Research Centre,
Indian Institute of Science, Bangalore 560012, India
e-mail: dpal@serc.iisc.ernet.in

1 Introduction

The concept of measuring topological information content [1] or, information content of graph using Shannon's measure of information [2] was introduced in the 1950's along with the 'negentropy' principle [3]. However, attempts [4] to correlate entropy and information theory for chemical species were made much later in the next decade based on order-disorder considerations in living systems [5]. It took another decade before linear combination of graph-theoretical invariants (LCGI) scheme for the thermodynamics of alkanes was introduced [6] which included entropy measurements by considering graph like state of matter. Subsequently, several graph-theoretical approaches found important applications in chemistry and biology [7–13] leading to dramatic increase in interest for predicting physical / physicochemical properties as well as biological activities of small organic compounds considering graph-theoretical models of molecular structures [14–17] primarily using topological indices [18] derived from them. Notwithstanding, inference of entropy for large number of chemical compounds in thermodynamic context and computational point of view has remained a challenge.

Development of mathematical models to measure entropy of chemical compounds from their molecular structures using information theoretical formalism has been attempted only for particular classes of compounds [4]. Therefore, development of predictive models for more generalized set of compounds containing both cyclic and acyclic structures is desirable. From structure-activity relationship point of view, graph-theoretical indices encapsulating information content of molecular graphs have been used to explain molecular properties of different series of compounds and high degree of correlations were obtained in these studies [16]. Such indices were derived considering automorphism group of the vertices [19], topological distances between pairs of vertices [20–22], neighborhood of vertices [23–27] as well as from other graph-theoretical considerations [16]. The usefulness of such indices seems to stem from their additive-constitutive character, akin to physicochemical properties like partition coefficient [28] which is an important parameter in determining biological activities. An information-theoretical topological index (ITTI) accrue from a basic advantage that information-theoretical formalism closely resembles the equation of computing physical entropy [4]. Therefore, it is of interest to see how physical entropy can be related to information content values obtained from individual molecular structures [29]. Such predicted values may find applications in various important areas of research in chemistry and biology such as in studies on the thermodynamic properties of gas-phase hydrogen bonded complexes [30]. However, for developing useful predictive regression models from structure-activity correlation studies it is important that meaningful descriptors related to activity are used. For using topological indices, the indices may be derived in such a way that structural aspects which are believed to determine entropy are reflected in the indices and consideration of weighted graph models of chemical structures for deriving meaningful indices may be expected to serve the purpose suitably.

In the present study, we have developed a regression model using ITTIs as molecular descriptors for the prediction of gas phase thermal entropy (S^0) of organic compounds. We have created a training set of 100 compounds, having acyclic and cyclic structures,

for developing the model taking data from the literature [31]. Since size, volume (bulk) and symmetry are believed to be the major structural features that determine entropy [29,32], two measures of ITTIs namely 'information content' and 'total information content' have been defined keeping those aspects in mind. One set of these measures has been derived taking atomic numbers as vertex weights and have been categorized as 'TopoPhysical Molecular Descriptors' [12]. Another set has been derived on the basis of the shortest chemical path(s) emerging from each vertex and connecting all other vertices of a chemically labeled molecular graph. These indices are believed to reflect some kind of topological symmetry [33–35] present in a molecular structure and have been categorized as 'TopoChemical Molecular Descriptors' [10]. Two total information indices together have produced significantly high correlation ($r^2 = 0.92$) with S^0 and the predicted values of a set of 10 test compounds, obtained from the corresponding regression equation, have been found to be very close to experimentally observed values [31]. This seems to indicate the usefulness of the proposed ITTIs and the regression model for the prediction of gas phase thermal entropy of chemical compounds. In addition, we have also provided a mathematical result in the form of a theorem and proof for total information-content measure for use as an indicator of certain changes made in molecular structures. It appears that the regression model and the mathematical result may provide acceptable predicted gas phase thermal entropy values of a large number of chemical compounds.

2 Methods and theoretical results

2.1 Vertex weighted molecular graph

A vertex weighted molecular graph $V_W(G)$ is a connected graph [36] representing the structural formula of a chemical compound where some numerical values are assigned to the vertices, representing the atoms in the molecule, as vertex weights.

2.2 TopoPhysical Molecular Descriptor (TPMD)

If the weights assigned to the vertices and / or edges of a molecular graph correspond to certain physical properties of the atoms and chemical bonds, a molecular descriptor derived from such a graph model may be called TopoPhysical Molecular Descriptor [12].

2.3 Vertex labeled molecular graph

A vertex labeled molecular graph $V_L(G)$ is a connected graph representing the structural formula of a chemical compound where the vertices are labeled according to some physical and / or, chemical characteristic of the atoms they are representing. In this work, we will use the chemical symbols such as 'C' for carbon, 'O' for oxygen, 'H' for hydrogen etc. as vertex labels.

2.4 Chemical path

A chemical path p(u, v) between two vertices u and v in a chemically labeled molecular graph $V_C(G)$ may be defined as a sequence of chemically labeled vertices connecting u and v in $V_C(G)$. For example, some chemical path having the connectivity of the (Carbon–Oxygen–Hydrogen) may be represented by (C–O–H).

The shortest chemical path between two vertices u and v is the path which has the minimum number of chemically labeled vertices including u and v. It may be noted that a connected graph containing cycle(s) may have more than one shortest chemical path.

2.5 TopoChemical Molecular Descriptor (TCMD)

If the labels and / or, weights assigned to the vertices and / or, edges of a molecular graph correspond to certain chemical properties of the atoms and chemical bonds, molecular descriptors derived from such graph models may be regarded as a TopoChemical Molecular Descriptor [10].

2.6 Information content of graph

Shannon's measure of information content [2] of a system is obtained by partitioning the elements of the system into disjoint classes on the basis of an equivalence relation defined on the elements of the system. So, if there are n elements in a system S and they are partitioned into k disjoint classes having n_i elements in the i^{th} partitioned class, i = 1, 2, …, k, then information content of the system, I_s, may be obtained from the following equation:

$$\begin{aligned} I_s &= -\sum_i p_i \log_2 p_i \\ &= -\sum_i \frac{n_i}{n} \log_2 \frac{n_i}{n} \\ &= \sum_i \frac{n_i}{n} \log_2 \frac{n}{n_i} \end{aligned} \tag{1}$$

where p_i is the probability of finding an element in the i^{th} partitioned class, and $p_i = \frac{n_i}{n} \geq 0$, $\sum_i p_i = 1$ and the measure is expressed in bits. However, for convenience, this measure is often done taking '$\log_{10}$' and that has been followed in this work.

Using (1), a measure of 'total information (TI) content'[3] of a system S may be obtained from Eq. (2):

$$TI_s = n \times I_s \tag{2}$$

Following the same principle, information content of a graph may be computed considering a partition of graph elements, say vertices, into disjoint classes on the basis of an equivalence relation defined on the set of vertices. It is also possible that the weights assigned on the vertices of a graph be considered to form a discrete system and a partition scheme is defined on the sum of these weights.

2.7 Information content of vertex weighted graph

Let Z_i be the atomic number of the i^{th} atom in a molecule, i = 1, 2,. . ., n, where n is the number of atoms in the molecule. Considering Z_i as the weights on the respective vertices, one can get a vertex weighted molecular graph model $V_W(G)$ of the compound. In order to have the measure of information content of this vertex weighted graph, we proceed as follows:

Let,

$$Z = \sum_{i=1}^{n} Z_i \tag{3}$$

Now, considering, Z_i values as a partition of Z into n disjoint classes, we can use Shannon's information formula [2] to have a measure 'information content on atomic number' I^Z using Eq. (4):

$$I^Z[V_w(G)] = -\sum_{i=1}^{n} \frac{Z_i}{Z} \log_2 \frac{Z_i}{Z} \tag{4}$$

Having this measure of information content on atomic number, one can also have a measure of total information content [3] on atomic number, TI^Z and may be obtained from:

$$TI^Z[V_w(G)] = Z \times I^Z \tag{5}$$

2.8 Information content on shortest chemical path

Two vertices u and v in a vertex labeled molecular graph $V_C(G)$ are said to be equivalent if,

a) number of shortest chemical paths between u and all other vertices of $V_C(G)$ is the same as those between v and all other vertices in $V_C(G)$;
b) for each shortest chemical path of length L from u there is one shortest chemical path of length L from v;
c) chemical labels of u and other vertices in that shortest chemical path from u is the same as those of v and other vertices in the shortest chemical path from v.

Now, from the equivalence of two vertices of $V_C(G)$ satisfying above conditions, one can have a partition of the vertex set of $V_C(G)$ and a measure of information content [5] on shortest chemical path, I^{CP}, for $V_C(G)$. Thus, I^{CP} for $V_C(G)$ may be given by (6):

$$I^{CP}[V_C(G)] = \sum_{j} \frac{X_i}{X} \log_2 \frac{X}{X_i} \tag{6}$$

where X_j is the number of vertices in the j^{th} partitioned class and $X = \sum_j X_j$

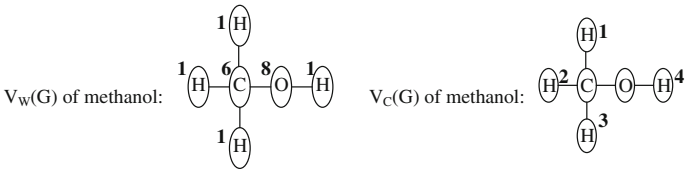

Fig. 1 Vertex weighted graph $V_W(G)$ and chemically labeled graph $V_C(G)$ of methanol. The atomic numbers are indicated in *bold* in the former, while the hydrogens are labeled for distinction in the latter

Subsequently, total information content [3] on shortest path, TI^{CP} may also be obtained from:

$$TI^{CP}[V_C(G)] = X \times I^{CP}[V_C(G)] \tag{7}$$

2.9 Illustration

Computation of the indices I^Z, TI^Z, I^{CP} and TI^{CP} is illustrated below (Fig. 1) taking methanol as an example:

In $V_W(G)$ of methanol, the respective atomic numbers have been put as weights in the vertices representing the atoms in the molecule. With one carbon, one oxygen and four hydrogen atoms in this molecule, the sum of the atomic numbers Z is:

$$Z[V_W(G)\text{methanol}] = 6 + 8 + (4 \times 1) = 18$$

Therefore, the I^Z and TI^Z for methanol may be obtained using (4) and (5):

$$\begin{aligned} I^Z[V_W(G)\text{methanol}] &= (8/18)\log_{10}(18/8) + (6/18)\log_{10}(18/6) \\ &\quad +4 \times (1/18)\log_{10}(18/1) \\ &= 1.9749 \\ TI^Z[V_W(G)\text{methanol}] &= 18 \times 1.9749 \\ &= 35.5490 \end{aligned}$$

Again, in $V_C(G)$ of methanol, chemical paths of different lengths (L) from its vertices are:

H1, H2 and H3	H4	C	O
H–C: L 1 × 3	H–O: L 1	C–H: L 1 × 3	O–H: L 1
H–C–H: L 2 × 2	H–O–C : L 2	C–O: L 1	O–C: L 1
H–C–O: L 2	H–O–C–H: L 3 × 3	C–O–H: L 2	O–C–H: L 2
H–C–O–H: L 3			

Looking into these chemical paths, one can find the equivalences and partition the vertex set of $W_C(G)$ of methanol into disjoint classes and can compute the information content indices using Eqs. (6) and (7). Clearly, the partition of the six vertices is (H1, H2, H3), (H4), (C) and (O) *i.e.*, (3,1,1,1).

Therefore,

$$\begin{aligned} I^{CP}[V_C(G)\text{methanol}] &= (3/6)\log_{10}(6/3) + 3 \times (1/6)\log_{10}(6/1) \\ &= 1.7920 \\ TI^{CP}[V_C(G)\text{methanol}] &= 6 \times 1.7920 \\ &= 10.7550 \end{aligned}$$

2.10 Theorem-1

If $P_1 = N(n_1, n_2, \ldots, n_i, \ldots, n_k)$ and $P_2 = N+1(n_1, n_2, \ldots, n_{i+1}, \ldots, n_k)$ be the partitions of two positive numbers N and N + 1 respectively, then

$$TI(P_2) > TI(P_1)$$

where TI stands for 'total information' as used in Eq. (2) and n_i, i = 1, 2, …, k, are positive integers, $k \geq 2$.

2.10.1 Proof

$$\begin{aligned} TI(P2) - TI(P1) &= (N+1)\log(N+1) - N\log N - (n_i+1)\log(n_i+1) + n_i \log n_i \\ &= N\log(N+1) + \log(N+1) - N\log N \\ &\quad - n_i\log(n_i+1) - \log(n_i+1) + n_i\log n_i \\ &= \log\left[\left\{(N+1)^{N+1} \times (n_i)^{ni}\right\} \Big/ \left\{N^N \times (n_i+1)^{ni+1}\right\}\right] \\ &= \log\Big[\left\{N(1+(1/N)\}^N/N^N\right\} \times \{(N+1)/n_i + 1\} \\ &\quad \times \{(n_i)^{ni}/n_i\{1+(1/n_i)\}^{ni}\Big] \\ &= \log\Big[\left\{(1+(1/N))^N/1 + (1/n_i)^{ni}\right\} \times \{(N+1)/(n_i+1)\}\Big] \\ &= \log[\{(N+1)/(n_i+1)\} \times \{(2 + (1/2)(1-(1/N)) \\ &\quad + (1/6)(1-(1/N)(1-(2/N) \\ &\quad + \cdots/2 + (1/2)(1-(1/n_i)) + (1/6)(1-(1/n_i))(1-(2/n_i))\}] > 0. \end{aligned}$$

Hence the theorem.

3 Results and discussion

The aim of the present study is to develop statistical regression model by using meaningful molecular descriptors as well as a mathematical result that could be used to predict gas phase thermal entropy (S^0) for a large number of organic compounds. For developing a useful regression model, we have defined two sets of information-theoretical topological indices (ITTIs) reflecting important structural features like size, bulk and topological symmetry of molecular structure which are believed to determine gas phase thermal entropy of organic compounds [29,32]. In carrying out this study, we have taken 100 compounds from the literature [31] comprising of both cyclic and acyclic structures. The idea is to use the entropy (S^0) values and the values of the ITTIs, defined in this paper, of these compounds as a training set for finding correlation between S^0 and ITTIs and develop a regression equation model which can be used to predict gas phase thermal entropy for a large number of organic compounds. The data comprising of S^0 values and those of two ITTIs, TI^Z and TI^{CP}, for 100 training set compounds are given in Table 1.

Since it is important to use meaningful molecular descriptors which can reflect molecular bulk, size as well as some kind of structural symmetry, we have found total information content indices more suitable since they include the sum of atomic numbers as well as the number of atoms in a molecule and not merely their partition into disjoint classes which are measured from I^Z and I^{CP}. By carrying out correlation studies, it has been found that both the total information content indices correlate highly ($r^2 \geq 0.86$) with S^0 values of the 100 training set compounds and therefore both these indices may be considered as suitable molecular descriptors for building up multiple regression models. By doing that, an improved correlation ($r^2 = 0.92$) is obtained taking two total information content indices, TI^Z and TI^{CP} together as variables and the predicted S^0 values are also close to the observed values for most of the compounds (Table 1). The trend is also apparent from the index values and S^0 (observed) values of the compounds. For example, the values of both TI^Z and S^0 for fluoro-, chloro-, bromo- and iodomethane have an increasing trend (Table 1: 2–5). From cycle containing compounds, the values of both TI^{CP} and S^0 for o-zylene and p-zylene have changed in the same direction. The corresponding multiple regression equation is given by Eq. (8).

$$\begin{aligned} S^0 &= 53.4 + 0.102(TI^Z) + 0.324(TI^{CP}) \\ N &= 100; \quad r^2 = 0.92; \quad s = 4.48; \quad F = 576.86 \end{aligned} \tag{8}$$

where, N is the number of data points (compounds) in the data set, r^2 is the square of correlation coefficient, s is standard deviation and F is the F statistic. The statistical results have been obtained using Minitab Statistical Software Minitab-16 [37].

In order to evaluate the predictive power of the regression model (8), a test set of 10 organic compounds comprising of both cyclic and acyclic structures, is created. A high predictive power of the regression model is evident from the closeness of the experimentally observed and predicted values for 10 test compounds (Table 2). It is apparent from this study that the two information indices TI^Z and TI^{CP} which reflect

Table 1 Observed and predicted gas phase thermal entropy (S^0) and the values of TI^Z and TI^{CP} indices for a training set of 100 compounds

	Compound	S^0(obs)[31]	S^0(pred) (Eq. 8)	TI^Z	TI^{CP}
1.	Methane	44.52	56.346	17.710	3.610
2.	Fluoromethane	53.25	58.761	31.019	6.855
3.	Chloromethane	55.97	59.396	37.214	6.855
4.	Bromomethane	58.76	60.212	45.179	6.855
5.	Iodomethane	60.64	60.714	50.071	6.855
6.	Ethane	54.76	59.977	44.039	6.490
7.	Fluoroethane	63.34	65.370	62.663	17.245
8.	Chloroethane	65.91	66.374	72.467	17.245
9.	Bromoethane	68.71	67.749	85.878	17.245
10.	Iodoethane	70.82	68.627	94.450	17.245
11.	Difluoromethane	58.94	60.914	49.640	7.610
12.	Dichloromethane	64.61	63.205	71.994	7.610
13.	Dibromomethane	70.10	67.684	115.702	7.610
14.	Diiodomethane	73.95	71.842	156.280	7.610
15.	Propane	64.58	67.125	75.682	18.544
16.	1-Fluoropropane	72.71	72.886	97.915	29.298
17.	1-Chloropropane	76.27	74.172	110.461	29.298
18.	1-Bromopropane	79.08	76.006	128.359	29.298
19.	1-Iodopropane	80.32	77.215	140.152	29.298
20.	2-Fluoropropane	69.82	70.051	97.915	20.544
21.	2-Chloropropane	72.70	71.336	110.461	20.544
22.	2-Bromopropane	75.53	73.171	128.359	20.544
23.	2-Iodopropane	77.55	74.379	140.152	20.544
24.	1,2-Dichloroethane	73.66	68.747	112.199	12.000
25.	1,2-Diiodoethane	83.30	78.500	207.371	12.000
26.	1-Bromobutane	88.30	84.810	172.383	42.549
27.	2-Bromobutane	88.50	84.575	172.383	41.793
28.	1-Chlorobutane	85.58	82.584	150.667	42.549
29.	2-Chlorobutane	85.94	82.340	150.667	41.793
30.	2-Mehylbutane	82.12	81.272	148.928	39.047
31.	Pentane	83.40	81.573	148.928	39.977
32.	2-Methylpentane	90.95	92.234	189.134	60.174
33.	3-Methylpentane	90.77	90.292	189.134	54.174
34.	Hexane	92.83	88.593	189.134	48.929
35.	2,2-Dimethylbutane	85.62	87.774	189.134	46.400
36.	2,3-Dimethylbutane	87.42	82.922	189.134	31.420
37.	Heptane	102.27	97.957	231.194	64.532
38.	2-Methylhexane	100.38	101.600	231.194	75.777
39.	3-Methylhexane	101.37	104.191	231.194	83.777
40.	2,2-Dimethylpentane	93.30	97.138	231.194	62.003

Table 1 continued

	Compound	S^0(obs)[31]	S^0(pred) (Eq. 8)	TI^Z	TI^{CP}
41.	2,3-Dimethylpentane	98.96	101.355	231.194	75.022
42.	2,4-Dimethylpentane	94.80	92.286	231.194	47.022
43.	3,3-Dimethylpentane	95.53	96.172	231.194	59.022
44.	2,2,3-Trimethylbutane	91.61	94.302	231.194	53.248
45.	2,2-Dimethylhexane	103.06	106.849	274.852	78.172
46.	2,3-Dimethylhexane	106.11	111.067	274.852	91.192
47.	2,4-Dimethylhexane	106.51	111.067	274.852	91.192
48.	2,5-Dimethylhexane	104.93	100.054	274.852	57.192
49.	3,3-Dimethylhexane	104.70	110.419	274.852	89.193
50.	3,4-Dimethylhexane	107.15	106.532	274.852	77.191
51.	1-Chloro-2-methylpropane	84.56	79.748	150.667	33.793
52.	2-Butanol	85.80	83.008	140.438	47.094
53.	Ethylamine	68.08	68.618	71.540	24.464
54.	Formic acid	59.45	61.891	46.529	11.610
55.	1-Heptanal	110.34	106.128	251.432	83.353
56.	1-Heptanethiol	117.89	112.683	284.006	93.284
57.	1-Hexanol	105.50	101.283	222.704	77.484
58.	Methanethiol	60.96	61.222	42.702	10.755
59.	Methanol	57.29	60.489	35.549	10.755
60.	2-Methyl-2-butanol	86.70	88.327	180.644	50.794
61.	2-Methyl-2-propanethiol	80.79	77.566	156.154	25.319
62.	2-Methyl-1-propanol	85.81	80.417	140.438	39.094
63.	1-Nitrobutane	94.28	89.998	195.521	51.245
64.	Nitroethane	75.39	72.990	114.206	24.464
65.	Nitromethane	65.73	65.414	76.839	12.897
66.	2-Nitropropane	83.10	78.396	153.897	28.596
67.	1-Octanethiol	127.20	123.022	333.241	109.627
68.	1-Pentanal	91.53	87.675	166.529	53.245
69.	1-Propanethiol	80.40	76.343	115.948	34.265
70.	1-Propanol	77.61	74.959	102.444	34.265
71.	2-Propanol	74.07	72.123	102.444	25.510
72.	Allylalcohol	73.51	71.995	89.471	29.219
73.	Propionaldehyde	72.83	71.103	89.471	26.464
74.	1-Propylamine	77.48	76.405	106.793	37.351
75.	n-Propylnitrate	92.10	86.311	187.031	42.548
76.	2-Thiabutane	79.62	75.451	115.948	31.510
77.	2-Thiaheptane	107.73	102.354	241.871	74.729
78.	3-Thiaheptane	108.27	102.354	241.871	74.729
79.	2-Thiahexane	98.43	92.963	198.214	59.549
80.	2-Thiapentane	88.84	83.971	156.154	45.094
81.	3-Thiapentane	87.96	79.436	156.154	31.094

Table 1 continued

	Compound	S^0(obs)[31]	S^0(pred) (Eq. 8)	TI^Z	TI^{CP}
82.	2-Thiahexane	68.32	64.920	77.954	11.020
83.	Triethylamine	96.90	91.063	227.053	44.559
84.	Cyclobutane	63.43	66.971	97.961	11.020
85.	Cyclohexane	71.28	76.652	175.020	16.529
86.	Cyclopropane	56.75	62.548	63.510	8.265
87.	m-Cresol	85.27	93.785	207.194	59.245
88.	p-Cresol	83.09	91.194	207.194	51.245
89.	1,2-Dichlorobenzene	81.61	86.720	227.467	31.020
90.	1,4-Dichlorobenzene	80.47	86.720	227.467	31.020
91.	1,2-Diethylbenzene	103.81	108.050	304.402	72.529
92.	1,3-Diethylbenzene	104.99	109.345	304.402	76.529
93.	1,4-Difluorobenzene	75.43	82.845	189.649	31.020
94.	Ethylbenzene	86.15	94.998	215.685	60.304
95.	Methylcyclohexane	82.06	96.774	216.643	65.484
96.	Thiacyclobutane	68.17	70.724	102.348	21.219
97.	Thiacycloheptane	86.50	93.045	226.941	50.711
98.	Thiacyclopentane	73.94	77.022	142.039	28.106
99.	o-zylene	84.31	90.867	215.685	47.549
100.	p-xylene	84.23	88.276	215.685	39.550

Table 2 Observed and predicted gas phase thermal entropy (S^0) and the values of TI^Z and TI^{CP} indices for a test set of 10 compounds

	Compound	S^0(obs)[31]	S^0(pred) (Eq. 8)	TI^Z	TI^{CP}
1.	2-Chloro-2-methylpropane	77.00	75.245	150.670	19.991
2.	2-Methyl-2-butanethiol	92.48	85.063	149.077	50.792
3.	1-Nitropropane	85.00	81.199	153.898	37.350
4.	1,1-Dichloroethane	72.91	72.247	112.200	22.849
5.	1,2-Dichloropropane	84.80	79.276	154.263	31.299
6.	3-Methylheptane	110.32	113.817	274.857	99.944
7.	2,2,3,3,-Tetramethylbutane	93.06	91.039	274.857	29.643
8.	m-xylene	85.49	92.102	215.685	51.548
9.	o-cresol	85.47	93.729	207.194	59.245
10.	1,3-Difluorobenzene	76.57	84.090	189.646	35.020

molecular bulk, size and some kind of topological symmetry, respectively, can be used as useful molecular descriptors for the prediction of gas phase thermal entropy (S^0) of organic compounds. It may be noted that for some compounds, TI^Z or, TI^{CP} may have redundant values although these indices together can discriminate one compound from another in most of the cases (Table 1).

In addition to that, Theorem-1 implies that other than mono-atomic species, the value of TI^Z will increase if an atom in a molecule is replaced by another atom of higher atomic number. This is reflected in the TI^Z values of the halogen substituted methane compounds (Table 1: 2–5). Again, symmetry in molecular structure is believed [29] to be another factor determining entropy. Although, information theoretical measures from symmetry of 3D structures of compounds have been proposed on the basis of point group of symmetry [29], the indices I^{CP} and TI^{CP} seem to reflect some kind of topological symmetry [33–35] in a molecule since it is based on the partition of its atoms (vertices) from the emergence of similar type of shortest chemical paths from them. This, in a way, puts topologically similar-positioned atoms in one partitioned class. Furthermore, total information index TI^{CP} includes the number of atoms in a molecule as a factor taking into account the size aspect too and therefore may be a more suitable index for the present purpose.

For these topochemical descriptors too, Theorem-1 may be used meaningfully. For example, the vertices of both cyclopropane (9 vertices) and cyclobutane (12 vertices) get partitioned into two disjoint classes as (class 1:3 carbons) & (class 2:6 hydrogens) for cyclopropane and (class 1:4 carbons) & (class 2:8 hydrogens) for cyclobutane. Therefore, the partitioned classes for cyclobutane may be obtained by adding vertices, one by one, to the partitioned classes for cyclopropane. For these two compounds, although the I^{CP} index gets the same value (0.918), the total information index TI^{CP} gets different values giving higher value for cyclobutane (11.020) than for cyclopropane (8.265) which corroborates with the result of Theorem-1. It may be noted that since we were interested in topological aspect of molecular symmetry, we did not consider bond types such as single bond, double bond etc. However, one can easily do that, if required, for a given problem.

4 Conclusions

It is evident from the present study that statistically significant regression model obtained using information theoretical topological indices, TI^Z and TI^{CP} defined in this work, along with Theorem-1 may be useful in predicting gas phase thermal entropy of a large number of organic compounds. The proposed approach is more general in that it is based on the structural information of both cyclic and acyclic compounds. While S^0 may be predicted from Eq. (8) having the two ITTIs computed, one can also use Theorem-1 to choose a molecule which can have a higher or, lower S^0 value as may be required. Since the coefficients of the dependent variables in Eq. (8) are positive, higher TI^Z and TI^{CP} values would give higher S^0 values. Our proposed theorem may be useful in better modeling gaseous state of matter, in general, including compressible gas flows, coefficient of expansion and contraction etc. Finally, we have also used some definitions as terminologies in this manuscript which may be useful in categorizing molecular descriptors in the studies where they are used.

Acknowledgments We sincerely acknowledge the financial assistance received from the Department of Biotechnology, New Delhi.

References

1. N. Rashevsky, Bull. Math. Biophys. **17**, 229 (1955)
2. C.E. Shannon, W. Weaver, *Mathematical Theory of Communication* (University of Illinois Press, Urbana, 1949)
3. L. Brillouin, *Science and Information Theory* (Academic Press, New York, 1956)
4. M. Valentinuzzi, M.E. Valentinuzi, Bull. Math. Biophys. **24**, 11 (1962)
5. H. Morowitz, Bull. Math. Biophys. **17**, 81 (1955)
6. M. Gordon, J.W. Kennedy, J. Chem. Soc. Faraday Trans. II **69**, 484 (1973)
7. N. Trinajstic, *Chemical Graph Theory* (CRC Press, Boca Raton, 1983)
8. A.T. Balaban (ed.), *Chemical Application of Graph Theory* (Academics Press, London, 1967)
9. A.J. Stuper, W.E. Brugger, P.C. Jurs, *Computer Assisted Studies of Chemical Structure and Biological Function* (Wiley-Interscience, New York, 1979)
10. G. Klopman, C. Raychaudhury, J. Comput. Chem. **9**, 232 (1988)
11. C. Raychaudhury, A. Banerjee, P. Bag, S. Roy, J. Chem. Inf. Comput. Sci. **39**, 248 (1999)
12. C. Raychaudhury, I. Ghosh, Internet Electron. J. Mol. Des. **3**, 350 (2004)
13. C. Raychaudhury, D. Pal, Curr. Comput.-Aided Drug Des. **8**, 128 (2012)
14. L.B. Kier, L.H. Hall, *Molecular Connectivity in Chemistry and Drug Research* (Academic Press, New York, 1976)
15. C. Raychaudhury, S.C. Basak, A.B. Roy, J.J. Ghosh, Indian Drugs **18**, 97 (1980)
16. D. Bonchev, *Information Theoretic Indices for Characterization of Chemical Structures* (Wiley-Research Studies Press, Chichester, 1983)
17. L.B. Kier, L.H. Hall, *Molecular Connectivity in Structure-Activity Analysis* (Wiley-Research Studies Press, Letchworth, 1986)
18. H. Hosoya, Bull. Chem. Soc. Jpn. **44**, 2332 (1971)
19. E. Trucco, Bull. Math. Biophys. **18**, 129 (1956)
20. D. Bonchev, N. Trinajstic, J. Chem. Phys. **67**, 4517 (1977)
21. C. Raychaudhury, S.K. Ray, J.J. Ghosh, A.B. Roy, S.C. Basak, J. Comput. Chem. **5**, 581 (1984)
22. G. Klopman, C. Raychaudhury, J. Chem. Inf. Comput. Sci. **30**, 12 (1990)
23. P. Sarkar, A.B. Roy, P.K. Sarkar, Math. Biosci. **39**, 299 (1978)
24. S.C. Basak, A.B. Roy, J.J. Ghosh, in *Second International Conference on Mathematical Modeling, vol. 2*, University of Missouri, Rolla (1979), p. 851
25. C. Raychaudhury, J.J. Ghosh, in *Third Annual Conference of the Indian Society for Theory of Probability and its Applications*, (Wiley Eastern Limited, New Delhi, 1981)
26. S.C. Basak, V.R. Magnuson, Arzneim.-Forsch./ Drug Res. **33**, 501 (1983)
27. A.B. Roy, C. Raychaudhury, J.J. Ghosh, S.K. Ray, S.C. Basak, in *Quantitative Approaches to Drug Design*, (Elsevier, Amsterdam, 1983), p. 75
28. C. Hansch, A. Leo, *Substituent Constant for Correlation Analysis in Chemistry and Biology* (Wiley, New York, 1979)
29. D. Bonchev, D. Kamenski, V. Kamenska, Bull. Math. Biol. **38**, 119 (1976)
30. L.A. Curtiss, M. Blander, Chem. Rev. **88**, 827 (1988)
31. J.E. Dean (ed.), *Langes' Hand Book of Chemistry. Table 9–2, 13 Edition* (McGraw-Hill, New York, 1985)
32. G.H. Wannier, *Statistical Physics* (Wiley, New York, 1966)
33. C.A. Shelley, M.E. Munk, J. Chem. Inf. Comput. Sci. **17**, 110 (1977)
34. C. Jochum, J. Gasteiger, J. Chem. Inf. Comput. Sci. **17**, 113 (1977)
35. R.E. Carhart, J. Chem. Inf. Comput. Sci. **18**, 108 (1978)
36. F. Harary, *Graph Theory* (Addison-Wesley, Reading, 1972)
37. Minitab-16: Minitab Statistical Software: PA, USA (2013)