ENTROPIES AND INFORMATION INDICES OF STAR FORESTS

Milan KUNZ

Chemopetrol, Research Institute of Macromolecular Chemistry,
656 49 Brno

Received February 8th, 1985

Entropy function H_m measures the difference of distances of m leaves from the root in a binary tree and in a forest of binary trees, when m leaves are divided by some information into n subgraphs. The enumerator of star forests with $(m + n)$ vertices and m edges is derived and using adjacency matrices of star forests their entropy functions and information indices are compared and their consistency discussed.

Graph theory is applied to the characterization of chemical structures and to correlations of their physical, chemical, and pharmacological properties by means of topological indices[1,2]. Information indices form a special class of topological indices. They use uncertainty function $\bar{H}$ to different graph characteristics (Table I).

The first one who used the uncertainty function $\bar{H}$

$$\bar{H} = -c \sum p_k \log p_k \,, \tag{1}$$

where c is a constant, k the index, and p_k are probabilities, was Boltzmann. But for him[3] it was a logarithmic measure of probability P

$$H = -c \log P \,, \tag{2}$$

where $H = n\bar{H}$ and P is given by a polynomial coefficient

$$P_n = n!/\prod_{k \geqq 0} n_k! \tag{3}$$

with a boundary condition

$$n = \sum_{k \geqq 0} n_k \,. \tag{4}$$

We get Eq. (*1*) from Eqs (*2*) and (*3*) using Stirling formula for log $n!$, thus the probability is then $p_k = n_k/n$.

Shannon used Eq. (*1*) as a measure of information content of a message. He introduced Eq. (*1*) as an axiom. But his measure $\bar{H}_m$ (the subscript m is added to distinguish it from Boltzmann function $\bar{H}$, which will be used with the subscript n, $\bar{H}_n$) can be derived from another polynomial coefficient P_m

$$P_m = m!/\prod_{j=1}^{n} m_j! = m!/\prod_{k \geq 0} m_k!^{n_k}\,, \tag{5}$$

where $m = \sum m_j = \sum n_k m_k$ and n is as in Eq. (*4*) and thus $p_k = m_k/m$.

The thermodynamical entropy is an abstract notion which is not understood completely till now and the same is valid for its information counterpart[4].

Function (*1*), applied to different graph characteristics, gives information indices which are considered as different from topological indices based on distances in a graph.

Recently Altenburg[5] has shown that information indices of Bonchev and Trinajstić are approximately quadratic functions of their basic topological indices.

The aim of this paper is to show that uncertainty function H_m is a measure of distance in a special class of graphs – decision trees, which can be formed to any set of m vertices, and to compare on a special class of graphs, star forests, different entropies, and topological indices.

Decision Trees

To any set of m vertices it is possible to construct binary decision trees with $(2m - 1)$ vertices and $(2m - 2)$ edges having m vertices of the m set as leaves. On the path

TABLE I

Entropies and topological indices[1,2]

Function	Author	Meaning
$\bar{H}_n$	Boltzmann	entropy
$\bar{H}_m$	Shannon	measure of information
I_{ORB}	Rashewsky	information content of graphs orbits
I_{CHR}	Mowshowitz	chromatic information index
$I_{\mathrm{D}}^{\mathrm{E}}$	Bonchev and Trinajstić	information index for the equality of distances
$I_{\mathrm{D}}^{\mathrm{W}}$	Bonchev and Trinajstić	information index for the magnitude of distances
I_{Z}	Bonchev and Trinajstić	information index on the Hosoya graph decompositions
I_{χ}^{E}	Bertz	information index for the edge connectivity

from the root to the leaf i there are l_i edges corresponding to binary decisions. Their sum $\sum l_i$ is the number of binary digits needed for indexing m set or, if we use graph terminology, it is the distance of the leaves from the root.

This sum has bounds

$$m \log_2 m \leqq \sum l_i \leqq \binom{m+1}{2} - 1 \,. \tag{6}$$

The lower bound we get when $m = 2^l$ and the decision tree has equal branches (Fig. 1*a*). Then $l_i = \log_2 m$ and $\sum l_i = m \log_2 m$. The upper bound corresponds to a comb (Fig. 1*b*) and it is a sum of indices i having 1 till $(m - 1)$ digits, the last one two times.

The sum of distances of the leaves from the root is a part of all distances in a decision tree and thus it is a topological distance index.

If there is some information dividing the m set into a forest of decision trees with n roots (Fig. 1*c*), then the sum of distances between the leaves and the roots is $\sum_j \sum_i l_i$. It has bounds

$$\sum m_j \log_2 m_j \leqq \sum_j \sum_i l_i \leqq \sum \binom{m_j - 1}{2} - n \,. \tag{7}$$

We can use the difference of decisions or digits needed for indexing the m set, spared by the given information dividing the m set into a forest of decision trees as a direct measure of information content. From Eqs (*6*) and (7) we get difference of the lower bounds as a possible estimation of this measure. Eq. (*1*) with the binary base of logarithm is thus a topological distance index.

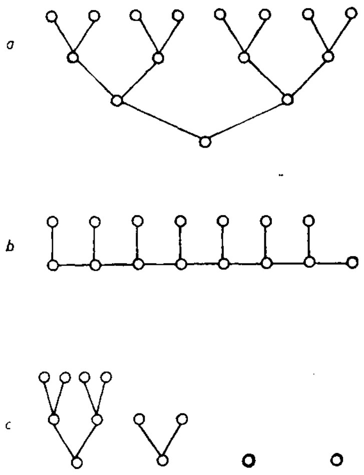

FIG. 1

Different decision trees. *a* Decision tree with equal branches. *b* Comb. *c* Decision tree divided into the forest of decision trees by information. Its information entropy: $H = 8\cdot3 - 4\cdot2 - 2\cdot1 - 2\cdot0 = 14$, $\bar{H} = 1.75$

Forests of Stars and their Adjacency Matrices

It has been shown[6,7] that a string of m symbols on an alphabet n in the transposed form $(\text{string})^{\mathrm{T}}$ can be written, using unit vectors $\mathbf{e}_j$ on place of symbols j of the alphabet n, as a naive matrix $\boldsymbol{A}(m, n)$. A naive matrix $\boldsymbol{A}(m, n)$ is a matrix with m rows and n columns, which elements a_{ij} are 0 or 1 and for which $\boldsymbol{Aj} = \boldsymbol{j}$ is valid, where $\boldsymbol{j}$ is the m dimensional unit vector-column. Naive matrices were interpreted as onedirectional Markov chains in n dimensional spaces or as stars S with indexed vertices and indexed multiedges. The enumerator of nonequivalent naive matrices $\boldsymbol{A}(m, n)$ is the sum of products of polynomial coefficients (*3*) and (*5*) over all partitions of m into n parts, 0 is allowed as a part too. Matrices $\boldsymbol{A}$ are not usually symmetrical and it is difficult to find characteristics needed for calculating all information indices.

A block matrix

$$\boldsymbol{W}(m + n, m + n) = \begin{Vmatrix} \mathbf{O} & \boldsymbol{A} \\ \boldsymbol{A}^{\mathrm{T}} & \mathbf{O} \end{Vmatrix}$$

is always a symmetrical one. It is the adjacency matrix of a bipartite graph with $(m + n)$ vertices and m edges from the set of m vertices to the set of n vertices. The edges are incident only in the n set and they form stars. $\boldsymbol{W}$ matrices correspond to a class of star forests, all star forests with m edges and $(m + n)$ vertices are obtained by symmetrical permutations $\boldsymbol{W} - \boldsymbol{P}^{\mathrm{T}}\boldsymbol{WP}$, where $\boldsymbol{P}^{\mathrm{T}}$ is the transposed unit permutation matrix $\boldsymbol{P}$.

The enumerator of star forests can be derived as follows.

In a forest of stars, there are n_0 stars S_1 with m_0 edges, n_1 stars S_2 with m_1 edges *etc.* till

$$(m + n) = \sum n_k(m_k + 1)\,, \quad m = \sum n_k m_k \quad \text{and} \quad n = \sum n_k\,.$$

We can choose stars S_1 from $(m + n)$ vertices using binominal coefficients.

$$\binom{m + n}{1}\binom{m + n - 1}{1}\cdots\binom{m + n - n_0 + 1}{1}.$$

As this stars (unconnected vertices) cannot be distinguished, we must divide the product of binomial coefficients by $n_0!$ and we get the first term

$$\binom{m + n}{n_0}\frac{1}{1!^{n_0}}\,.$$

Similarly we can choose stars S_2 from the rest of vertices

$$\binom{m + n - n_0}{2}\binom{m + n - n_0 - 2}{2}\dots\binom{m + n - n_0 - n_1 + 2}{2}$$

and we get the second term

$$\binom{m + n - n_0}{n_1}\frac{1}{2!^{n_1}}\,.$$

At stars with $k \geqq 2$ we can choose a center in the star from $(k + 1)$ vertices and thus we get the third term

$$\binom{m + n - n_0 - 2n_1}{n_2}\frac{3^{n_2}}{3!^{n_2}} = \binom{m + n - n_0 - 2n_1}{n_2}\frac{1}{2!^{n_2}}$$

and generally for $k \geqq 2$

$$\binom{m + n - \sum\limits_{k=0}^{k-1}(k + 1)\, n_k}{n_k}\frac{1}{m_k!^{n_k}}\,.$$

The final result for all possible stars we get by multiplying all terms and making sum over all partitions of number m into n parts

$$\sum \frac{(m + n)!}{2^{n_1}\prod\limits_{k\geqq 0} n_k!\, m_k!^{n_k}}\,. \tag{8}$$

Comparing the result with the enumerator of matrices **A**, equation (*8*) can be written as the product of three coefficients

$$\sum \left\{\left[2^{-n_1}\binom{m + n}{n}\right]\left[n!/\prod_{k\geqq 0} n_k!\right]\left[m!/\prod_{k\geqq 0} m_k!^{n_k}\right]\right\}. \tag{9}$$

The last two coefficients count block matrices **W** and thus naive matrices **A**, in the first term the binomial coefficient counts permutations between m and n sets, the divisor 2^{n_1} is due to stars S_1 forming permutation cycles of length 2.

Other topological characteristics of star forest follow. Paths of length of 1 and 2 are in a star forest. There are $\sum m_j$ paths of length 1 and $\sum \binom{m_j}{2}$ paths of length 2, the number of paths is $\sum\binom{m_j + 1}{2}$, their total length and thus Wiener number is $\sum m_j^2$ which is the sum of components in $\mathbf{AA}^{\mathrm{T}}$ and the square of Euclidean length of the corresponding vector $\mathbf{A}^{\mathrm{T}}\mathbf{A}$.

The characteristic polynom of an adjacency matrix of star forests is given by the expression

$$p(\mathbf{W}, k) = x^{|m-n|} \prod_{j=1}^{n} (x^2 - m_j)\,, \qquad (10)$$

where $|m - n|$ is the absolute term[8]. Polynom terms c_k count subgraphs with k isolated edges. $\mathbf{W}$ is a symmetrical matrix and thus its eigenvalues are square roots of eigenvalues of $\mathbf{W}^2$. Matrix $\mathbf{W}^2$ has two components, the diagonal matrix $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ with eigenvalues m_j and the matrix $\mathbf{AA}^{\mathrm{T}}$ with the same nonzero eigenvalues.

The star forest as a bipartite graph has the chromatic number always 2, the number of orbits O we get by counting $m_k > 0$. If $m_k = 0$ or 1, then $O_k = n_k^0$, if $m_k \geqq 2$, there are 2 orbits, 1 as leaves and 1 as the center and thus $O = 2[(\sum n_k^0) - 1]$.

To each star forest there are 6 different parameters function (*1*) can be applied for. As an example the partition (5, 5) parameters are given in Table II.

TABLE II
Characteristics of star forests W (5,5)

Parameters		Partition						
		50000	41000	32000	31100	22100	21110	11111
Term $1/2^{n_1} . 252$		252	126	252	63	126	63/2	63/8
Polynomial coefficient (*3*)		5	20	20	30	30	20	1
Polynomial coefficient (*5*)		1	5	10	20	30	60	120
Coefficients c_k of the polynom (*10*)	0	1	1	1	1	1	1	1
	1	5	5	5	5	5	5	5
	2	0	4	6	7	8	9	10
	3	0	0	0	3	4	7	10
	4	0	0	0	0	0	2	5
	5	0	0	0	0	0	0	1
Number of orbits		3	4	5	4	4	4	1
Orbit partition		541	4321	33211	4321	4222	6211	10
Path number		15	11	9	8	7	6	5
Wiener number		25	17	13	11	9	7	5

DISCUSSION

Practical importance of star forests in chemistry is rather low. They can be used as models of ansambles of molecules of hydrogen depleted lowest alkanes only. To find the analogous equation for all forests or at least for linear tree forests were more important.

But even so, star forests are interesting in many respects. Although the theory of partitions is a special field of the number theory with many theorems of the highest mathematical difficulty[9], it seems that nobody studied corresponding polynomial coefficients as systematically, as they are studied in graph theory[8].

Properties of star forests adjacency matrices allow to compare different definitions of entropy and information indices. In this field words of John von Neumann[10] to Shannon hold: "In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place and more important no one knows what entropy really is, so in a debate you will always have the advantage".

The statistical entropy is defined in statistical mechanics[11] as a logarithm of the number of possible states of the systems. Boltzmann and Shannon entropies are in accord, they are logarithms of corresponding polynomial coefficients. Polynomial coefficients of acyclic graphs are number of subgraphs with k isolated edges. Thus, if we wanted to calculate their polynomial entropy, we should calculate their product as a geometrical analog of the Hosoya index. But why to use function (*1*)? Its application should be based on a better theory than an analogy is. Gordon and Temple[12] proposed to define the entropy of alkanes as the logarithmic measure of symmetry of their graphs. There are some differences between their and this paper methods, but in general the approach is the same, it goes from the symmetry of molecule graphs.

Distance information indices I_D^E and I_D^W give at forests of stars trivial results similarly as chromatic information index I_{CHR}. In general case they should give some information about distance matrices as vectors in n^2 or $\binom{n}{2}$ dimensional spaces. Distances ij form a partition of $\mathbf{W}$, if we take them as leaves, we can measure their distance from the root in a decision tree. Thus formally there can not be any objections against such indices, but some cases of their practical application[13] are ahead of theory.

Complicated algorithmic relations between the adjacency and distance matrices of graphs exist and systematical study of distance matrices properties as n dimensional vectors just started[14]. If information indices should give more information then their original functions, they should be studied systematically as information content of graphs orbits and chromatic information index were by Mowshowitz[15].

REFERENCES

1. Sabljić A., Trinajstić N.: Acta Pharm. Jug. *31*, 189 (1981).
2. Balaban A. T., Motoc J., Bonchev D., Mekenyan O.: Topics Curr. Chem. *114*, 21 (1983).
3. Boltzmann L.: Wien Ber. *76*, 373 (1877).
4. Hamming R. W.: *Coding and Information Theory*. Prentice Hall, New York 1980.
5. Altenburg K.: Z. Phys. Chem. Leipzig *265*, 257 (1984).
6. Kunz M.: Chem. Prům. *33*, 542 (1983).
7. Kunz M.: Inform. Process. Management *20*, 519 (1984).
8. Cvetković D. M., Doob M., Sachs H.: *Spectra of Graphs*. Academic Press, Berlin 1983.
9. Andrews G. E.: *Theory of Partitions*. Addison-Wesley, London 1976.
10. Shaw D., Davis C. H.: J. Am. Soc. Inform. Sci. *34*, 67 (1983).
11. Kittel C.: *Thermal Physics*. Wiley, New York 1969.
12. Gordon M., Temple W. B.: J. Chem. Soc., A *1970*, 729.
13. Bonchev D., Mekenyan O.: J. Chem. Soc., Faraday Trans. *80*, 695 (1984).
14. Křivka P., Trinajstić N.: Aplik. Matem. *28*, 357 (1983).
15. Mowshowitz A.: Bull. Math. Biophys. *30*, 175, 225, 387, 415 (1968).