J. theor. Biol. (1999) **197,** 193–205
Article No. jtbi.1998.0866, available online at http://www.idealibrary.com on IDEAL

Characterization and Comparison of *Escherichia coli* Transfer RNAs by Graph Theory based on Secondary Structure

Clara I. Bermúdez,* Edgar E. Daza† and Eugenio Andrade*‡

**Grupo de Biología Molecular y Teórica, Departamento de Biología and †Grupo de Química Teórica, Departamento de Química, Facultad de Ciencias, Universidad Nacional de Colombia, Santa Fe de Bogotá, D.C., Colombia*

(*Received on* 4 *February* 1998, *Accepted in revised form on* 9 *November* 1998)

We have developed a model to characterize the tRNA structures of *Escherichia coli* using graph theory. First of all, tRNAs were represented as graphs, whose vertices correspond to nucleotides and the edges to phosphodiester and hydrogen bond linkages. Vertices and edges were weighted using the results of a preliminary quantum study of the nucleotides and the possible coupling between pair bases using the semiempirical method AM1. For each vertex, we defined a nucleotide valence that measures the capability of forming hydrogen bonds. Edges were differentiated by using bond orders.

We have proposed weighted structural descriptors—closely related to molecular Randic connectivity and Balaban distance indices—as a distinctive characteristic of each structure. Molecules were characterized by a set of weighted structural descriptors and classified by a clustering method and discriminant function analysis. Two main groups of tRNAs that correspond to the biosynthetic amino acid pathways, in agreement with Wong's coevolution theory of the genetic code, were obtained.

© 1999 Academic Press

Introduction

Transfer RNAs (tRNAs) are essential molecules in protein biosynthesis that couple specific amino acids with corresponding codons. This process is achieved by molecular interaction between tRNA and aminoacyl-tRNA synthetases (aaRSs) that assures translation fidelity by accurate discrimination between tRNAs and cognate or non-cognate aaRSs. Structural features, required for tRNA aminoacylation by aaRSs, act as recognition motifs and they must play a major role in maintaining the tRNA amino acid specificity (Sherman & Söll, 1996).

This work is aimed at exploring a new approach to the concept of structure, that can be applied to tRNA structure and elucidates tRNA structure–function relationships. The characterization and modeling of tRNA are highly dependent on the concept of molecular structure. It has been commonplace to identify molecular structure with the position and arrangement of atoms in ordinary three-dimensional space. This point of view has prevailed in a variety of fields: from the very well known results of Watson & Crick (1953)—about the molecular geometry of DNA—to the successful determination of mol-

‡Author to whom correspondence should be addressed. E-mail: eandrade@ciencias.ciencias.unal.edu.co.

0022–5193/99/006193 + 13 $30.00/0 © 1999 Academic Press

ecular structures for a multitude of biologically important compounds just to mention one of the most popular cases.

Transfer RNAs are highly differentiated nucleic acids comprised of 74–95 nucleotides that are folded into a compact three-dimensional pattern which is believed to accommodate most of the known tRNA sequences (Schimmel, 1989). tRNA structure has also been modeled in terms of its geometry, i.e. based upon the information of the three-dimensional structure obtained by X-ray diffraction (Sussman & Kim, 1976; Clark, 1978; Woo *et al.*, 1980; Westhof *et al.*, 1985). These conceptions of tRNA structure share the advantages and disadvantages of the more general molecular structure concept in use.

Molecular structures have been usually treated as geometric models, in which molecules are depicted as spatial arrays of nuclei according to a mechanical approach that depends upon distance and angle measurements that vary according to the modeling hypothesis and experimental conditions. Therefore, the same results are not always expected, and specific corrections have to be made for each particular method (Villaveces & Daza, 1997). We believe that the concept of *molecular structure* is richer than the simple geometrical relationships of distances between atoms. The mechanistic view that proposes a one to one correlation between molecular structure and molecular geometry, has prevented the development of alternative definitions of molecular structure. Hence, it would be more adequate to regard geometry at most as an approximation to what the concept of molecular structure could be.

New theoretical approaches in mathematical chemistry, such as graph theory and topology might be helpful to uncover information underlying in a set of "structures" (Villaveces & Daza, 1990, 1997; Mezey, 1993; Bonchev & Seitz, 1997; Rouvray, 1997). These new tools are being explored in order to provide a more rigorous approach to the concept of molecular structure; so one might hope that it would be possible in future to construct a robust theory of molecular structure surpassing the limitations of the naïve semiclassical theory founded on mechanical analogies, by aiming to free it from the strait jacket imposed by the paradigm of a unique geometry.

A key idea, developed during the second half of the nineteenth century, is that atoms have a valence, making it possible to derive a structural formula. Thus, atomic composition is not sufficient to define a molecular structure but it is necessary to know how the component atoms are related to, that is which atoms are bonded to which. A structural formula in the classical chemical sense is a collection of symbols with lines joining them; sometimes single, sometimes double or even triple; and it is the existence of these links between atoms that carries the main information about the chemical properties of a compound. A chemical bond is a particular relationship between a pair of atoms. The collection of atoms and bonds constituting a molecule may be thought of as a mathematical object whose elements are the atoms and pairs of bonded atoms; together they form the kind of structure that mathematicians call a **graph** (Harary, 1969). Its pictorial representation has been traditionally called a chemical structural formula.

The graphical formula, much more than the theory of atomic arrangements, became an especially concise method of representing a great variety of chemical knowledge. That is relationships of chemical neighborhood—understood in the sense of which atoms are linked to which—are more fundamental than the physical or geometrical questions of interatomic distances and angles. This consideration leaves us with a notion of molecular structure nearer to that of Butlerov (1861), and with a mathematical graph as a better tool to represent structures than an ensemble of geometrical points.

A complete chemical graph is then a structure composed of a set of atoms, $\{v_i\}$, and a set of pairs of atoms $\{(v_i, v_j)\}$ denoting which atoms are bonded and denoted by **G**: $(\{v_i\}, \{(v_i, v_j)\})$. The elements of the set $\{v_i\}$ have been called vertices and the elements of $\{(v_i, v_j)\}$ are known as edges. This approach posits the existence of graph-theoretical invariants, called the descriptors or indices allowing a characterization of a molecular structure (Trinajstic, 1983; Kier & Hall, 1986; Randic *et al.*, 1988). These indices in the first instance are based on the connectivity

between the vertices and some other properties of the graph (Randic, 1975, 1984; Balaban, 1985; Mihalic & Trinajstic, 1992). Chemical graphs can be used to represent a variety of chemical objects such as molecules, reactions, crystals, polymers and clusters (Mihalic & Trinajstic, 1992); we propose that this approach can be extended to RNA, DNA and proteins, where vertices map to monomeric units.

Two of the more relevant indices used in molecular characterization are those of Randic (Randic, 1975; Randic & Trinajstic, 1994) and Balaban (1985), that exhibit high discriminating capacity. The Randic indices χ are based on a connectivity matrix and were first proposed as the sum of bond contributions. The index is defined as:

$$\chi^1 = \sum_{\text{bonds}} (m_i \times n_j)^{-1/2},$$

where m_i and n_j are the valences or the number of incident edges to each vertex. The index was generalized later by introducing "higher" connectivities which can be viewed as subgraphs assigned to molecular fragments of different length. These subgraphs are called paths in which no vertex occurs more than once. Furthermore:

$$\chi^L = \sum_{\text{paths}} (m_i \times n_j \ldots \times z_k)^{-1/2},$$

where L is the path length, i.e. the number of bond occurrences between those vertices that defined the path.

On the other hand, Balaban index J (Balaban, 1985) was constructed from the distance matrix D and is calculated as follows:

$$J = \frac{K}{u(G)+1} \sum_{\text{bonds}} (d_i \times d_j)^{-1/2}$$

and

$$d_i = \sum_{i=1}^{n} (D)_{ij},$$

where d_i stands for the sum of distances from every vertex i to the others. Each element of D is the length of the shortest path between the vertices i and j in G (Harary, 1969). The cyclomatic number $u(G)$ is the number of independent cycles in G and it is related to K (number of edges) and V (number of vertices) by Euler's formula $u(G) = K - V + 1$. The index encodes both the distance among all vertices and the existence of cycles, so recuperating in an abstract way the concept of molecular entanglement.

However, aiming to find a rigorous model of structure, we suggest that it is important to build a model which embraces, besides molecular connectivity, other concepts like interaction potentials between atoms or fragments, and some interactions with the molecular environment that can be captured by thermodynamic properties.

One way to introduce interaction potentials into the model is by using quantum theory. Thence, a molecular system is represented by its wave function which can be found by solving the time-independent Schrödinger equation (McWeeny, 1989; Stewart, 1990; Zerner, 1991; Szabo & Ostlund, 1996). Wave functions allow us to evaluate some chemical properties in order to characterize the system. Nonetheless, for molecules like tRNA, it is almost impossible to solve Schrödinger equations with enough precision to be able to account for biological behavior. Therefore, we propose that quantum chemistry may be used to identify univocally some fragments; in our case we have considered that nucleotides and some base pairs exhibit some properties that enable us to distinguish among them.

Finally, we consider two ways in which thermodynamic properties can be incorporated into the model. Direct thermodynamic measurements, if available, may be used as descriptors. Otherwise, thermodynamics appear in the model in an indirect way, as happens when we use semiempirical quantum programs which employ heat of formation as a parameter (Stewart, 1990; Zerner, 1991) or when we handle secondary structure prediction programs that use free energy minimization inferred from previous

empirical measurements (Jaeger *et al.*, 1990; Lybrand, 1990; McCaskill, 1990).

Materials and Methods

We have used the program RNADRAW 1.01 (Matzura, 1995) to predict and test alignments reported by the European Laboratory of Molecular Biology, 42 sequences were used in this analysis. However, it was assumed that modified nucleotides cannot make hydrogen bonds, rendering a base pairing unattainable (except for pseudouridine which is able to couple with Guanine and Adenine). tRNA sequence numbers are RA1660, RA1661, RA1662, RC1660, RD1660, RE1660, RE1661, RE1662, RF1660, RG1660, RG1661, RG1662, RH1660, RI1660, RI1661, RI1662, RK1660, RL1660, RL1661, RL1662, RM1660, RN1660, RQ1660, RQ1661, RR1660, RR1661, RR1662, RR1663, RR1664, RS1660, RS1661, RS1662, RS1663, RS1664, RT1660, RT1661, RV1660, RV1661, RV1662, RW1660, RY1660, RY1661.

All quantum mechanical calculations were run with GAUSSIAN 94 (Frisch *et al.*, 1995) to obtain AM1 semiempirical wave function. Mulliken atomic charges were calculated for each isolated nucleotide and bond orders were estimated from calculations for all possible coplanar base pairs with an intermolecular distance of 2.9 ± 0.04 Å (Westhof *et al.*, 1985).

Modified Randic and Balaban indices were calculated using NUCLEIA, a Pascal code based on PCDM2.0 (Oliveros *et al.*, 1995), developed by the authors. All calculations were performed with double precision.

To quantify structural similarity we have used clustering analysis. Groups of tRNA molecules were then established by UPGMA (unweighted pair group method using arithmetic average) and based on Euclidean distance. Discriminant function analysis is employed to test the statistical significance of the grouping.

Results

A MODEL FOR THE TRNA

Two shapes have been accepted as standard for tRNA molecules: L-shape and cloverleaf. It is recognized that when the three-dimensional L-shape is projected over a plane it gives rise to a cloverleaf the standard secondary structure (Clark, 1978; Westhof *et al.*, 1985). This secondary structure provides a pattern of bonds between the basic units of the tRNA, the nucleotides. Therefore, in order to study tRNA structure, one starts with the primary structure that has been classically considered as fundamental, because it contains all the information required to predict the secondary structure by using thermodynamic algorithms that converge to the cloverleaf shape, as the more stable secondary structure (Jaeger *et al.*, 1990; McCaskill, 1990).

In this paper instead of studying secondary structure by focusing on those motifs that are responsible for three-dimensional folding, we propose an approach where base stacking and tertiary base pairs are not included. tRNA folding is presumed to be the outcome of thermodynamically stable base pair couplings that can be projected over a plane as a graph corresponding to the cloverleaf. Free energy calculations allow to estimate the most likely secondary structure that can be described by connectivity patterns. Therefore, tRNA connectivity patterns are of the utmost importance in grasping structure–function correlations. This justifies the choice of secondary structure as a starting point for our study.

We propose to represent each tRNA molecule as a graph. The set of vertices corresponding to the nucleotides and the set of edges corresponding to pairs of nucleotides either joined by covalent or hydrogen bonds. Since different functional tRNAs may collapse into the same graph, it is required to develop a more precise treatment necessary to make a distinction among the different kinds of vertices and edges. Instead of giving a simple name to each one of them, we weighted edges and vertices based on quantum properties.

WEIGHTING VERTICES AND EDGES

Vertices

In order to characterize nucleotides and possible differences arising from their position in the cloverleaf, we have defined for each

nucleotide a valence b_i, as an analogue of atomic valence, that discriminates among the four main types of nucleotides as well as the 23 modified structures present in the tRNA (Table 1). This valence depends on the incident edges, i.e. bonds on the nucleotide, which are of two classes: the phosphodiester linkage and the binding by hydrogen bonds (regular and non-standard pairs). Thus:

$$b_i = I + (1 + H)|q|,$$

where I is the number of phosphodiester bonds over each nucleotide. H (1 or 0) indicates if the nucleotide is forming hydrogen bonds or not. Finally, q is the sum of the Mulliken atomic charges of the atoms that would be involved in the hydrogen bonds; this is a discriminating factor among all different types of nucleotides. Atomic charges were calculated from an AM1 semiempirical wave function (Dewar *et al.*, 1985).

This valence was used to calculate indices following the formula proposed by Randic for their connectivity indices (Randic, 1975; Randic & Trinajstic, 1994). We calculated indices from order one to four, using the following expression:

$$\chi^1 = \sum_{\text{bonds}} (b_i \times b_j)^{-1/2},$$

$$\chi^2 = \sum_{\text{paths}} (b_i \times b_j \times b_k)^{-1/2},$$

$$\chi^3 = \sum_{\text{paths}} (b_i \times b_j \times b_k \times b_l)^{-1/2},$$

$$\chi^4 = \sum_{\text{paths}} (b_i \times b_j \times b_k \times b_l \times b_m)^{-1/2},$$

where the power of each index indicates the length path, i.e. the number of edges between those vertices that defined the path. b is the valence for each vertex. An example of the procedure is depicted in Fig. 1. The results for each tRNA are presented in Table 2.

We have orthogonalized these descriptors in order to avoid the mutual correlations (Randic, 1991a,b, 1996). The four orthogonalized descriptors were taken as the components of a four-dimensional vector characterizing each tRNA.

TABLE 1

Characterization of nucleotides by valence (b_i). *Nucleotide valences were estimated in order to weight differences arising from their location in the cloverleaf by applying the following formula*: $b_i = I + (1 + H)|q|$ *where* I *is the number of phosphodiester bonds over each nucleotide*. H (1 *or* 0) *indicates if the nucleotide is forming hydrogen bonds or not and* q *is the sum of the Mulliken atomic charges of the atoms that would be involved in the hydrogen bonds*; *this is a discriminating factor among all different types of nucleotides*. *Modified nucleotides are abbreviated as in Sprinzl* et al. (1989)

Nucleotides	I	$\|q\|$	H	Valence b_i
U	2	0.4225	0	2.423
U	1	0.4225	1	1.846
U	2	0.4225	1	2.846
S4U	2	0.1501	0	2.150
D	2	0.4270	0	2.427
Ψ	2	0.4539	0	2.454
Ψ	2	0.4539	1	2.908
acp3U	2	0.6443	0	2.644
mnm5s2U	2	0.3587	0	2.359
mnm5U	2	0.4277	0	2.428
m5U	2	0.4181	0	2.418
cmo5U	2	0.3979	0	2.398
Um	2	0.4302	0	2.430
A	1	0.3753	0	1.375
A	1	0.3753	1	1.750
A	2	0.3753	0	2.375
A	2	0.3753	1	2.751
ms2i6A	2	0.3239	0	2.324
m6A	2	0.3104	0	2.310
m2A	2	0.3836	0	2.384
t6A	2	0.2155	0	2.216
m6t6A	2	0.3032	0	2.303
I	2	0.3976	0	2.398
C	2	0.7870	0	2.787
C	2	0.7870	1	3.574
ac4C	2	0.7324	0	2.732
acp2c	2	0.7137	0	2.714
s2C	2	0.6300	0	2.630
Cm	2	0.7957	0	2.796
G	2	0.5277	0	2.528
G	1	0.5277	1	2.056
G	2	0.5277	1	3.056
Q	2	0.5650	0	2.565
m7G	2	0.8276	0	2.828
m1G	2	0.7239	0	2.724
Gm	2	0.5329	0	2.533

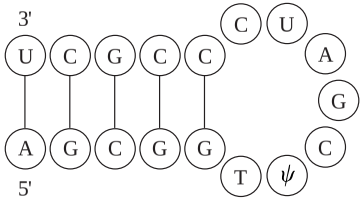

FIG. 1. Example of molecular descriptor calculations for the hairpin TΨC of $tRNA^{ALA}_{GGC}$.
χ^1 is defined as the sum of bond contributions and **b** is the valence for each nucleotide (Table 1):
$\chi^1 = \{[\mathbf{b}(U)\mathbf{b}(C)]^{-1/2} + [\mathbf{b}(U)\mathbf{b}(A)]^{-1/2} + [\mathbf{b}(C)\mathbf{b}(G)]^{-1/2} + [\mathbf{b}(C)\mathbf{b}(G)]^{-1/2} + \cdots\}$.
$\chi^1 = \{[(1.846)(3.574)]^{-1/2} + [(1.846)(1.750)]^{-1/2} + [(3.574)(3.056)]^{-1/2} + [(3.574)(3.056)]^{-1/2} + \cdots\} = 7.470$
χ^2 involves combinations of two consecutive bonds:
$\chi^2 = \{[\mathbf{b}(U)\mathbf{b}(C)\mathbf{b}(G)]^{-1/2} + [\mathbf{b}(U)\mathbf{b}(A)\mathbf{b}(G)]^{-1/2} + [\mathbf{b}(U)\mathbf{b}(C)\mathbf{b}(G)]^{-1/2} + [\mathbf{b}(C)\mathbf{b}(G)\mathbf{b}(C)]^{-1/2} + \cdots\}$.
$\chi^2 = \{[(1.846)(3.574)(3.056)]^{-1/2} + [(1.846)(1.750)(3.056)]^{-1/2} + [(1.846)(3.574)(3.056)]^{-1/2} + \cdots\} = 6.615$
and so forth.
The Balaban Index uses the expression:

$$J = \frac{K}{u(G)+1} \sum_{\text{bonds}} (d_i \times d_j)^{-1/2} \quad \text{and} \quad d_i = \sum_{i=1}^{n} (D)_{ij},$$

D is the distance matrix where each element D_{ij} is the length of the shortest weighted path and d_i sums for all entries in the corresponding row or column.

3′	U	C	G	C	C	C	U	A	G	C	Ψ	T	G	G	C	G	A5′	d_i
U	0.0000	1.0000	2.0000	3.0000	4.0000	5.0000	6.0000	7.0000	8.0000	7.0307	6.0307	5.0307	4.0307	3.0307	2.0307	1.0307	0.0307	64.2456
C	1.0000	0.0000	1.0000	2.0000	3.0000	4.0000	5.0000	6.0000	7.0000	6.0376	5.0376	4.0376	3.0376	2.0376	1.0376	0.0376	1.0307	51.2939
G	2.0000	1.0000	0.0000	1.0000	2.0000	3.0000	4.0000	5.0000	6.0000	5.0376	4.0376	3.0376	2.0376	1.0376	0.0376	1.0376	2.0307	42.2939
C	3.0000	2.0000	1.0000	0.0000	1.0000	2.0000	3.0000	4.0000	5.0000	4.0376	3.0376	2.0376	1.0376	0.0376	1.0376	2.0376	3.0307	37.2939
C	4.0000	3.0000	2.0000	1.0000	0.0000	1.0000	2.0000	3.0000	4.0000	3.0376	2.0376	1.0376	0.0376	1.0376	2.0376	3.0376	4.0307	36.2939
C	5.0000	4.0000	3.0000	2.0000	1.0000	0.0000	1.0000	2.0000	3.0000	4.0000	3.0376	2.0376	1.0376	2.0376	3.0376	4.0376	5.0307	45,2563
U	6.0000	5.0000	4.0000	3.0000	2.0000	1.0000	0.0000	1.0000	2.0000	3.0000	4.0000	3.0376	2.0376	3.0376	4.0376	5.0376	6.0307	54,2187
A	7.0000	6.0000	5.0000	4.0000	3.0000	2.0000	1.0000	0.0000	1.0000	2.0000	3.0000	4.0000	3.0376	4.0376	5.0376	6.0376	7.0307	63.1811
G	8.0000	7.0000	6.0000	5.0000	4.0000	3.0000	2.0000	1.0000	0.0000	1.0000	2.0000	3.0000	4.0000	5.0000	6.0000	7.0000	8.0000	72.0000
C	7.0307	6.0376	5.0376	4.0376	3.0376	4.0000	3.0000	2.0000	1.0000	0.0000	1.0000	2.0000	3.0000	4.0000	5.0000	6.0000	7.0000	63.1811
Ψ	6.0307	5.0376	4.0376	3.0376	2.0376	3.0376	4.0000	3.0000	2.0000	1.0000	0.0000	1.0000	2.0000	3.0000	4.0000	5.0000	6.0000	54.2187
T	5.0307	4.0376	3.0376	2.0376	1.0376	2.0376	3.0376	4.0000	3.0000	2.0000	1.0000	0.0000	1.0000	2.0000	3.0000	4.0000	5.0000	45.2563
G	4.0307	3.0376	2.0376	1.0376	0.0376	1.0376	2.0376	3.0376	4.0000	3.0000	2.0000	1.0000	0.0000	1.0000	2.0000	3.0000	4.0000	36.2939
G	3.0307	2.0376	1.0376	0.0376	1.0376	2.0376	3.0376	4.0376	5.0000	4.0000	3.0000	2.0000	1.0000	0.0000	1.0000	2.0000	3.0000	37.2939
C	2.0307	1.0376	0.0376	1.0376	2.0376	3.0376	4.0376	5.0376	6.0000	5.0000	4.0000	3.0000	2.0000	1.0000	0.0000	1.0000	2.0000	42.2939
G	1.0307	0.0376	1.0376	2.0376	3.0376	4.0376	5.0376	6.0376	7.0000	6.0000	5.0000	4.0000	3.0000	2.0000	1.0000	0.0000	1.0000	51.2939
A	0.0307	1.0307	2.0307	3.0307	4.0307	5.0307	6.0307	7.0307	8.0000	7.0000	6.0000	5.0000	4.0000	3.0000	2.0000	1.0000	0.0000	64.2456
5′																		

$u(G)$ is the cyclomatic number that represents the number of independent cycles in G and is obtained from the Euler's formula $u(G) = K - V + 1$, where K and V stand, respectively, for the number of edges and the number of vertices. So, $u(G) = 21 - 17 + 1 = 5$, and

$$\sum_{\text{bonds}} (d_i \times d_j)^{1/2} = 0.449120$$

then:

$$J = \frac{21}{5+1}(0.449120) \quad J = 1.57192$$

TABLE 2

Modified Randic Indices from order one to four calculated for each tRNA. Modified Randic Indices (χ^{1-4}) *were estimated by using the following expressions*:

$$\chi^1 = \sum_{\text{bonds}} (b_i \times b_j)^{-1/2}, \qquad \chi^2 = \sum_{\text{paths}} (b_i \times b_j \times b_k)^{-1/2},$$

$$\chi^3 = \sum_{\text{paths}} (b_i \times b_j \times b_k \times b_l)^{-1/2}, \quad \chi^4 = \sum_{\text{paths}} (b_i \times b_j \times b_k \times b_l \times b_m)^{-1/2},$$

where the power of each index indicate the length path, i.e the number of edges between those vertices that defined the path. b *is the valence for each vertex.* Ω^{1-4} *correspond to orthogonal modified Randic Indices*

Graph	χ^1	χ^2	χ^3	χ^4	Ω^1	Ω^2	Ω^3	Ω^4
LeuAAA	38.443958	35.918808	29.310738	26.466193	3.730982	−0.025059	0.022931	0.076021
ArgACG	33.372490	30.972211	25.123474	22.611826	−1.340486	0.078158	−0.042951	−0.008223
ArgACG1	33.772004	31.220366	25.271917	22.699588	−0.940972	−0.071495	−0.036250	0.053160
LeuCAG	38.354356	35.923130	29.326866	26.632717	3.641380	0.068482	−0.006676	0.158410
IluCAU	33.356019	30.791184	24.943193	22.326843	−1.356957	−0.086468	0.003378	−0.005695
MetCAU	34.170270	31.765616	25.908511	23.419073	−0.542706	0.077189	0.074027	0.080970
TrpCCA	33.478598	30.985549	25.085746	22.453475	−1.234378	−0.014159	−0.050374	−0.064296
GlyCCC	32.709010	30.602214	24.970490	22.588854	−2.003966	0.368808	−0.014981	−0.103403
ArgCCG	33.145282	30.001253	23.930942	20.948224	−1.567694	−0.666562	−0.082261	−0.030227
SerCGA	39.684433	37.306331	30.466239	27.545986	4.971457	0.127285	−0.059202	−0.019507
GlnCUG	32.857832	30.069581	24.255331	21.616070	−1.855144	−0.312011	0.025008	0.081602
PheGAA	33.358162	30.853298	25.022647	22.446770	−1.354814	−0.026488	0.003505	−0.002803
ValGAC	33.702829	31.055630	25.104737	22.500238	−1.010147	−0.167352	−0.021494	0.078110
ValGAC1	34.199365	31.905839	26.105988	23.748351	−0.513611	0.188442	0.103293	0.147762
LeuGAG	38.392014	35.587384	28.950040	25.865875	3.679038	−0.304761	0.067346	0.013425
IluGAU	34.051660	31.524052	25.546224	22.959760	−0.661316	−0.046271	−0.029169	0.042929
IluGAU1	34.052326	31.524681	25.546758	22.960172	−0.660650	−0.046305	−0.029150	0.042876
CysGCA	32.615522	30.047843	24.471123	21.937337	−2.097454	−0.092474	0.160324	0.046362
GlyGCC	33.384269	31.004186	25.204584	22.702590	−1.328707	0.098404	0.002107	−0.005870
SerGCU	41.079760	38.596117	31.433441	28.410032	6.366784	0.027702	−0.133916	0.027344
SerGGA	38.964654	36.580261	29.900749	27.038844	4.251678	0.117921	−0.008717	0.000904
SerGGA1	39.066681	36.685749	30.012774	27.156682	4.353705	0.121818	0.012674	0.013416
AlaGGC	33.543761	31.252786	25.502758	23.057046	−1.169215	0.188193	0.050414	0.009750
ThrGGU	33.738912	31.498653	25.690284	23.226510	−0.974064	0.239743	0.007584	−0.031863
ThrGGU1	33.290020	31.225236	25.533575	23.141594	−1.422956	0.413301	0.003145	−0.100306
TyrGUA	37.286329	34.395250	27.958178	24.660104	2.573353	−0.395931	0.120961	−0.212910
TyrGUA1	37.290556	34.417366	27.984125	24.676584	2.577580	−0.378024	0.120216	−0.233433
AspGUC	33.768641	31.238833	25.340847	22.790938	−0.944335	−0.049680	0.007303	0.066105
HisGUG	34.339475	32.349270	26.489526	24.132505	−0.373501	0.492361	−0.023553	−0.046360
AsnGUU	33.964696	31.805609	26.032637	23.626030	−0.748280	0.321879	0.054345	−0.006188
ValUAC	33.384805	30.973011	25.157155	22.683770	−1.328171	0.066696	−0.004786	0.042018
GlyUCC	32.773216	30.002411	24.183792	21.503522	−1.939760	−0.294927	0.002376	0.021752
ArgUCU	33.758379	31.097946	25.127813	22.497843	−0.954597	−0.180348	−0.028223	0.063908
ArgUCU1	33.588381	31.512383	25.811104	23.413943	−1.124595	0.403361	0.043207	−0.074788
SerUGA	39.026555	36.612195	29.899925	27.017699	4.313579	0.088218	−0.023081	0.002011
AlaUGC	33.014026	30.357981	24.452629	21.828709	−1.698950	−0.179139	−0.080483	0.013122
AlaUGC1	33.291887	30.866310	25.063896	22.547722	−1.421089	0.052516	−0.001761	0.002134
GluUUC	33.052429	30.453075	24.530099	21.866030	−1.660547	−0.122284	−0.108720	−0.063091
GluUUC1	33.106698	30.546467	24.648531	22.009418	−1.606278	−0.082929	−0.086687	−0.056934
GluUUC2	32.747889	30.331538	24.520323	21.932769	−1.965087	0.059419	−0.097860	−0.120530
GlnUUG	32.925816	30.140241	24.327904	21.694887	−1.787160	−0.309045	0.036711	0.091815
LysUUU	33.841026	31.612404	25.833412	23.408358	−0.871950	0.251816	0.049437	0.010522

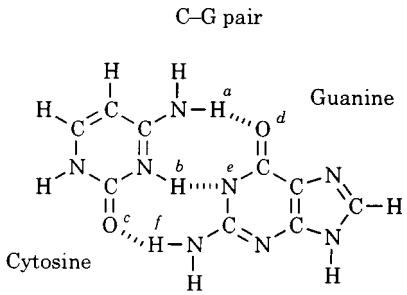

FIG. 2. Atoms that participate in hydrogen bonds. Intermolecular distance between coplanar base-pairs were fixed at 2.9 ± 0.04 Å (Westhof *et al.*, 1985). *a*, *b*, *c*, *d*, *e* and *f* indicate the atoms involved in the hydrogen bond.

Edges

Bond orders were used to weight edges in order to build a Distance Index like that proposed by Balaban (Balaban, 1985). Phosphodiester linkages were assumed to have the typical order of a covalent bond: one. Hydrogen bonds were considered as one edge, their weights were calculated as the sum of bond orders between atoms that yield or accept the hydrogen to make a bridge. (As an example see the C:G pair in Fig. 2).

We calculated bond orders for six base pairs occurring at the secondary structure. The mismatch C:A was included because this has been recognized as a tautomer base pair present in some nucleic acids (Drake & Baltz, 1976; Asahara *et al.*, 1993; Meinnel *et al.*, 1993) located in the stem position for tRNAs of MetCAU, LeuCAG and AlaUGC. We can see that Watson–Crick pairs have bond orders that agree with their stability (Table 3). The weighted distance Balaban indices were calculated as follows

$$J = \frac{K}{u(G) + 1} \sum_{\text{bonds}} (d_i \times d_j)^{-1/2}$$

and

$$d_i = \sum_{i=1}^{n} (D)_{ij},$$

Thus, each element of D is the length of the shortest weighted (i.e. using bond orders) path between the vertices i and j in G. An example of this procedure is also depicted in Fig. 1. The results for each tRNA are presented in Table 4.

Discussion

Summarizing, we calculated the four weighted connectivity indices (Randic indices), that were later transformed into orthogonal ones (Randic & Trinajstic, 1994). In order to quantify structural similarity and to establish relationships with biological function, we have employed clustering analysis considering each tRNA as a vector with four components. This allowed us to group tRNAs according to the Euclidean distance used as a similarity coefficient.

The cluster for tRNAs viewed as a four-dimensional vector is presented in Fig. 3. There are

TABLE 3

Bond orders between atoms that participated in hydrogen bonding. Bond orders were used to weight edges in order to build a Balaban Index. Phosphodiester linkages were assumed to have the order of a covalent bond: 1. *Hydrogen bonds were considered as one edge whose weight was calculated as the sum of bond orders between atoms that establish the bridge by either giving or accepting the hydrogen. It is shown that Watson and Crick pairs have the higher bond orders in agreement with their stability*

Base pairs	Atoms	Bond orders	Total
C:G	aH . . . dO	0.0110	
	bN . . . eH	0.0163	
	cO . . . fH	0.0103	0.0376
	aO . . . cH	0.0080	
A:U	bN . . . dH	0.0227	0.0307
	aO . . . cH	0.0076	
G:U	bH . . . dO	0.0122	0.0198
	aH . . . cN	0.0059	
A:C	bH . . . dN	0.0040	0.0099
	aO . . . cH	0.0049	
A:Ψ	bH . . . dN	0.0031	0.0080
	aO . . . bH	0.0005	
G:Ψ	aO . . . cH	0.0046	0.0051

TABLE 4

Modified Balaban Index for each tRNA. The weighted Balaban index was calculated according to:

$$J = \frac{K}{u(G) + 1} \sum_{\text{bonds}} (d_i \times d_j)^{-1/2} \quad d_i = \sum_{i=1}^{n} (D)_{ij},$$

D *is the distance matrix where each element* D_{ij} *is the length of the shortest weighted path and* d *sums for all entries in the corresponding row or column.* u(G) *is the cyclomatic number that represents the number of independent cycles in* G *and obtained from the Euler's formula* $u(G) = K - V + 1$, *where* K *and* V *stand, respectively, for the number of edges and the number of vertices*

tRNA graph	V	K	$u(G)$	$\Sigma(d_i.d_j)^{-1/2}$	J
LeuAAA	87	111	25	0.165236	0.705431
ArgACG	76	96	21	0.157794	0.688556
ArgACG1	77	97	21	0.155746	0.686698
LeuCAG	87	111	25	0.165320	0.705789
IluCAU	76	96	21	0.157792	0.688547
MetCAU	77	97	21	0.155752	0.686725
TrpCCA	76	96	21	0.157749	0.688359
GlyCCC	74	94	21	0.162343	0.693647
ArgCCG	77	96	20	0.151641	0.693216
SerCGA	90	115	26	0.157449	0.670616
GlnCUG	75	94	20	0.157786	0.706280
PheGAA	76	96	21	0.157778	0.688486
ValGAC	77	97	21	0.155706	0.686522
ValGAC1	77	97	21	0.155736	0.686654
LeuGAG	87	110	24	0.156851	0.690144
IluGAU	77	97	21	0.155814	0.686998
IluGAU1	77	97	21	0.155814	0.686998
CysGCA	74	93	20	0.160638	0.711397
GlyGCC	76	96	21	0.157781	0.688499
SerGCU	93	119	27	0.153401	0.651954
SerGGA	88	112	25	0.159415	0.686711
SerGGA1	88	112	25	0.159415	0.686711
AlaGGC	76	96	21	0.157760	0.688407
ThrGGU	76	96	21	0.157720	0.688233
ThrGGU1	75	95	21	0.160377	0.692537
TyrGUA	85	107	23	0.155050	0.691265
TyrGUA1	85	107	23	0.155050	0.691265
AspGUC	77	97	21	0.155889	0.687329
HisGUG	77	98	22	0.157594	0.671487
AsnGUU	76	96	21	0.157791	0.688543
ValUAC	76	96	21	0.157756	0.688390
GlyUCC	75	94	20	0.158572	0.709798
ArgUCU	77	97	21	0.155728	0.686619
ArgUCU1	75	95	21	0.160427	0.692753
SerUGA	88	112	25	0.159445	0.686840
AlaUGC	76	96	21	0.157927	0.689136
AlaUGC1	76	96	21	0.157756	0.688390
GluUUC	76	96	21	0.158335	0.690916
GluUUC1	76	96	21	0.158260	0.690589
GluUUC2	75	95	21	0.160431	0.692770
GlnUUG	75	94	20	0.157750	0.706119
LysUUU	76	96	21	0.157794	0.688556

two main groups: the first one (1) corresponds to tRNAs that have an extra arm (tRNAs for Ser, Leu and Tyr) and the second one (2) the tRNAs without an extra arm. The second group is divided into two subgroups (2a) and (2b); in the subgroup (2a) the tRNA corresponding to His has a tendency to become isolated. The remaining tRNAs in this subgroup have in common that their cognate aminoacids belong to the same biosynthetic pathway: aspartate family; i.e. the tRNAs corresponding to Asp, Asn, Lys, Thr, Met and Ile (Wood *et al.*, 1981). The tRNAs for Gly and Ala also are present in this group and are related with the metabolism of pyruvate, glycerate and glioxylate (Greenberg, 1969). The tRNAs for Trp may be related with Ala via Ser, so to say that Trp being a biosynthetic product of Ser, can be considered by extension as a member of the same biosynthetic pathway. In the subgroup (2a) there are some tRNAs that can be considered as exceptions: the tRNAs for Phe, Val and Arg. Phe and Val are related by a common biosynthetic pathway with Tyr and Leu, respectively (Wong, 1975), but tRNAs for these last two amino acids are in the first group (1). tRNAs for Arg will be discussed afterwards.

The subgroup (2b) gathers tRNAs that recognize amino acids belonging to the glutamate family: Gln, Glu, and Arg (Wood *et al.*, 1981). However, most of the isoacceptors for Arg have a tendency to cluster within group (2a). The biosynthetic pathway can be extended to include the tRNAs for Gly and Cys. The former can be related with the tRNAs of Ala via the mentioned metabolic pathway and the latter with a product of the Ser family. Although, it must be remembered that Ser belongs to the first group (1). Glu, Asp and Ala are known to be interconvertible via the tricarboxylate cycle (Wong, 1975), a fact that justifies our results, since some structural ambiguity was detected for the tRNAs isoacceptors of Ala, Gly and Arg which showed a tendency to be either in the subgroup (2a) or subgroup (2b). This ambiguity is justified for Ala and Gly since their isoacceptors play the role of bridges extending the biosynthetic pathways of the aspartate and glutamate families.

In Fig. 4, we present the cluster for tRNAs represented by the Balaban Index. In this case, the descriptor employed is less sensitive to the number of nucleotides by definition, so that the group (1), obtained using Randic indices, is spread out into the rest of the groups, although most of the previous results still hold. Thus, for this cluster, the group (2b2a) corresponds to the group (2a) in Fig. 3, but now some of the tRNAs for Ser fall within this group as an indication that it also belongs to the biosynthetic pathway between Ala and Trp. Likewise, the glutamate family was divided between the groups (2a) and (2b2b) according to different levels of similarity whose biochemical relations we have not been able to elucidate.

In general, one can observe the formation of some small clusters having only a few types of isoacceptors; in particular Gln and Leu independently seem to have the least variation between

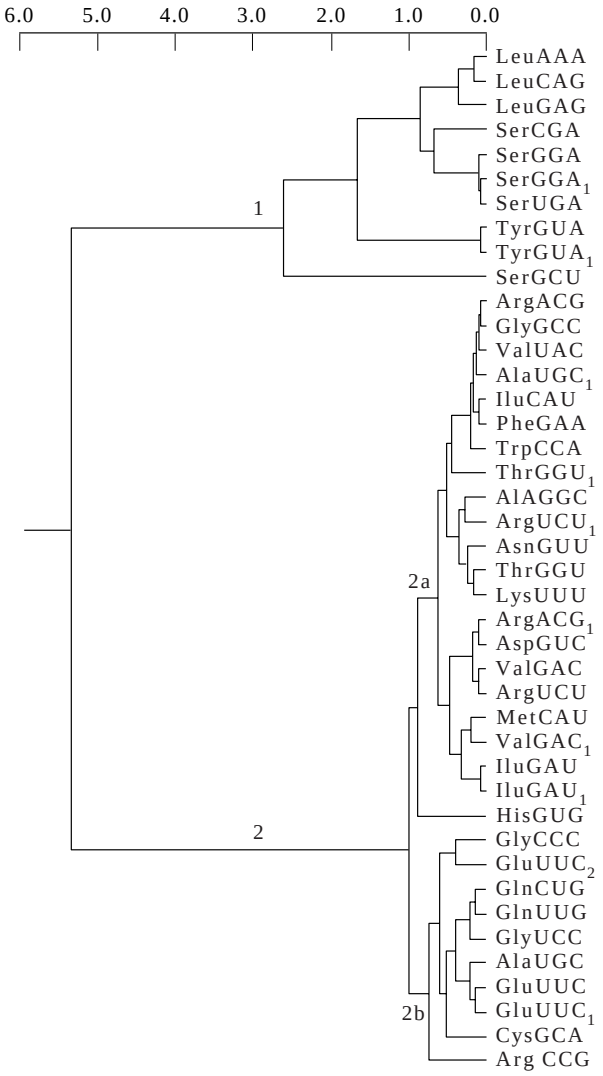

FIG. 3. Cluster obtained from orthogonal modified Randic Indices. Two main groups are shown. Group 1 includes all tRNAs having an extra arm and group 2 is split into two subgroups. Group 2a clusters tRNAs whose cognate aminoacids belong to the same biosynthetic pathway (aspartate family) and group 2b embraces those related with the glutamate biosynthetic pathway.

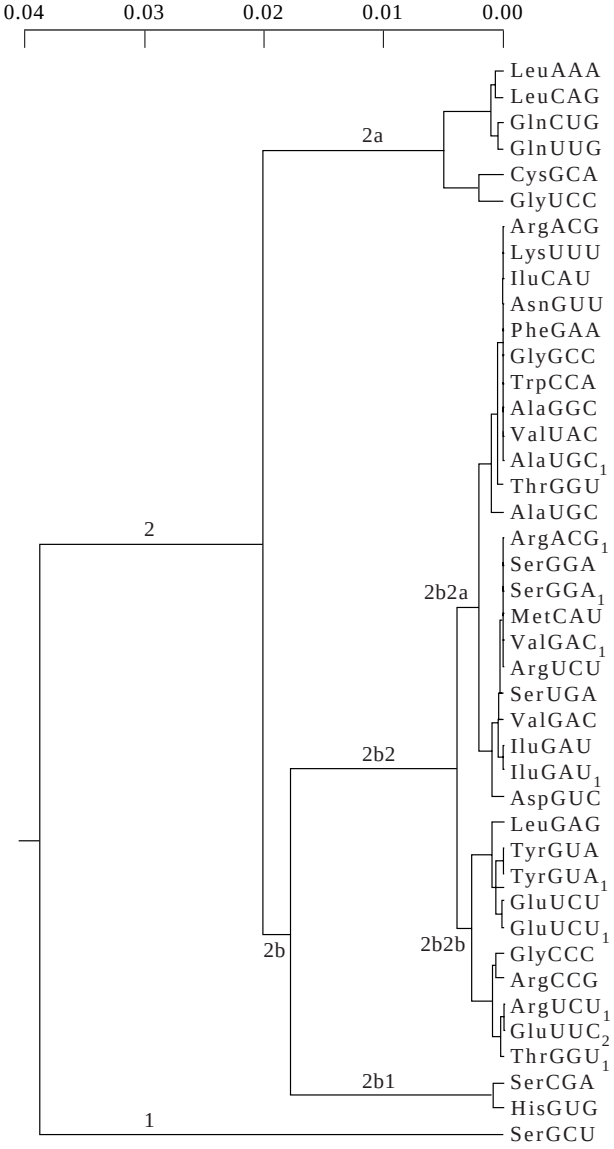

FIG. 4. Cluster obtained from modified Balaban distance index. Three main groups are shown. Groups 2a and 2b2b cluster all tRNAs that recognize aminoacids belonging to the glutamate biosynthetic pathway, and group 2b2a those tRNAs related to the aspartate biosynthetic pathway.

TABLE 5

Discriminant Function Analysis for Randic and Balaban Indices Clustering. Discriminant Function Analysis is employed to test the statistical significance of the grouping obtained by using linear combinations of the variables that permit group separations. (5*a*) *Randic Indices Clustering and* (5*b*), *Balaban Index Clustering*

(5a)

Group	Predicted (%)	Discriminant F	Sig. level	Eigenvalue	D.F.	Chi-square
1	100	1	0.000	37.125	8	148.241
2a	95	2	0.008	0.3664	3	11.708
2b	90					

(5b)

Group	Predicted (%)	Discriminant F	Sig. level	Eigenvalue	D.F.	Chi-square
1	100	1	0.000	66.480	4	160.04
2a	100					
2b2a	100					
2b2b	100					
2b1	100					

their isoacceptors, so that both of them have a high molecular similarity. Glu and Ser are also very similar. These results agree with the noise effect reported by Cedergren *et al*. (1980): the distribution of nucleotide differences between tRNAs specific for the same amino acid and tRNAs of different specificities overlaps extensively.

The clusters obtained are statistically significant. Matrix correlation between the cophenetic value matrix and the Euclidean distance matrix of Cluster for Randic indices is 0.95076, using normalized Mantel statistic Z, and using approximate Mantel t-test: $t = 11.305$ and probability random $Z <$ Observed Z: $p = 1.000$. Matrix correlation of Cluster for Balaban index is 0.95220 (using normalized Mantel statistic Z), and approximate Mantel t-test: $t = 7.483$ and Probability random $Z <$ Observed Z: $p = 1.000$. Discriminant function analysis is employed to test the statistical significance of the grouping by using linear combinations of the variables as functions that permit group separation; the results are shown in Table 5.

Since tRNA-aaRS complementarity results in specific tRNA aminoacylation, these recognition motifs are recorded in tRNA structure that to some extent are captured by the weighted graph. Furthemore, if aminoacylation of tRNA by aaRS recapitulates the genetic code by dictating the association between amino acids and tRNA anticodons (Saks & Sampson, 1995), it is likely that tRNA bears structural features permitting to trace the evolution of the genetic code. Since there are certain indications that tRNA possesses detectable phylogenetic information (Hopfield, 1978; Fitch & Upper, 1987; Di Giulio, 1994), we propose that the origin of the genetic code can be traced by studying molecular structure as it is characterized in our model. If tRNAs still maintain structural trails of the process that led to genetic code organization, then tRNA similarities must be explained in the context of Wong's coevolution theory (Wong, 1975). We therefore suggest that the similarity observed among tRNAs relates to functional restrictions that became incorporated into tRNA structure as a whole.

Conclusion

Following the above considerations, tRNAs were grouped in clusters that include the three basic elements proposed for molecular structure description. Our results indicate that the grouping shows a striking correlation between tRNA isoacceptors that share structural features with tRNAs that recognize amino acids belonging to similar biosynthetic pathways. Although tRNAs are considered to be molecules that diverged in primary structure, our results show

that this evolution took place under structural restrictions. The highly conserved secondary structure appears to reveal a high functional restriction that involves the integrity of the molecule as shown by our model. The obtained grouping favors the hypothesis that tRNAs bear structural traces that would help to comprehend how the genetic code evolved. It has been stated that the genetic code could have been established by means of its versatility in codon usage, that codons originally assigned to precursor amino acids were used to define derivative amino acids. Thence, tRNAs must have been modified to bridge the recognition among amino acids that belong to the same biosynthetic route. Along these lines, we propose to search for correspondence between this model and Wong's coevolution theory in order to achieve a better understanding of the organization of the genetic code.

REFERENCES

Asahara, H., Himeno, H., Tamura, K., Hasegawa, T., Watanabe, K. & Shimizu, M. (1993). Recognition nucleotides of *Escherichia coli* $tRNA^{Leu}$ and its elements facilitating discrimination from $tRNA^{Ser}$ and $tRNA^{Tyr}$. *J. Mol. Biol.* **231,** 219–229.

Balaban, A. T. (1985). Applications of graph theory in chemistry. *J. Chem. Inf. Comput. Sci.* **25,** 334–343.

Bonchev, D. & Seitz, W. A. (1997). The concept of complexity in chemistry. In: *Concepts in Chemistry: a Contemporary Challenge* (Rouvray, D. H., ed.) pp. 353–378. England: RSP Ltd.

Butlerov, A. Z. (1861). *Chem. Pharm.* **4,** 546. In: Villaveces, J. L. & Daza. E. E. (1997). The concept of chemical structure. In: *Concepts in Chemistry: a Contemporary Challenge* (Rouvray, D. H., ed.) pp. 101–131. England: RSP Ltd.

Cedergren, R. J., LaRue, B., Sankoff, D., Lapalme, G. & Grosjean, H. (1980). Convergence and minimal mutation criteria for evaluating early events in tRNA evolution. *Proc. Nat. Acad. Sci. U.S.A.* **77**(5), 2791–2795.

Clark, B. F. C. (1978). General features and implications of primary, secondary and tertiary structure. In: *Transfer RNA* (Altman, R., ed.) pp. 14–47. Cambridge, MA: MIT Press.

Dewar, M. J. S., Zoebisch, E. G., Healy, E. F. & Stewart, J. J. P. (1985). A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **107,** 3902–3909.

Di Giulio, M. (1994) The phylogeny of tRNA molecules and the origin of the genetic code. *Origins Life Evol. Biosphere* **24,** 425–434.

Drake, J. W. & Baltz, R. H. (1976). The biochemistry of mutagenesis. *Ann. Rev. Biochem.* **45,** 11–38.

Fitch, W. M. & Upper, K. (1987). The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. *Cold Spring. Harbor Symp. Quant. Biol.* **52,** 759–767.

Frisch, M. J., Trucks, G. W., Schlegel, H. B., Gill, W., Jhoson, B. G., Robb, M. A., Cheeseman, J. R., Keith, T. A, Petersson, G. A., Montgomery, J. A., Raghavachari, K., Lahan, A., Zakrzewski, A., Ortiz, J. V, Foresman, J. B., Cioslowsky, J., Stefanov, B. B., Nanyakkara, A., Challacombe, M., Peng, C. Y, Ayala, P. Y, Chen, W., Wong, M. W., Andres, J. L., Replogle, E. S., Gomperts, R., Martin, R. L, Fox, D. J, Binkley, J. S., Defree. D. J, Bakes, J., Stewar, J. J. P., Head, M., Gordon, Gonzales, C. & Pople, J. A. (1995). *GAUSSIAN* 94. Pittsburgh, PA: Gaussian 94 Inc.

Greenberg, D. M. (1969). In: *Metabolic Pathways* (Greenberg, D. M., ed.), Vol III, 3rd Edn., pp. 237–315. London: Academic Press. In: Wong, J. T.-F. (1975). A coevolution theory of the genetic code. *Proc. Nat. Acad. Sci. U.S.A.* **72**(5), 1909–1912.

Harary, F. (1969). *Graph Theory*, pp. 1–42, Reading, MA: Addison-Wesley.

Hopfield, J. J. (1978). Origin of the genetic code: a testable hypothesis based on tRNA structure, sequence, and kinetic proof-reading. *Proc. Nat. Acad. Sci. U.S.A.* **75**(9), 4334–4338.

Jaeger, J. A., Turner, D. H. & Zucker, M. (1990). Predicting optimal and suboptimal secondary structure for RNA. *Methods Enzymology* **183,** 281–330.

Kier, L. M. & Hall, D. D. (1986). *Molecular Connectivity in Structure–Activity Analysis*, pp. 1–27. England: Research Studies Press.

Lybrand, T. P. (1990). Computer simulation of biomolecular systems using molecular dynamics and free energy perturbation methods. In: *Reviews in Computational Chemistry* (Lipkowitz, K. B. & Boyd, D. B., eds), Vol. I, pp. 295–315. New York: VCH Publishers.

Matzura, O. (1995). RNADRAW 1.01. http://mango.mef.ki.se/~ole/rnadraw/rnadraw.html

McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29,** 1105–1119.

McWeeny, R. (1989). *Methods of Molecular Quantum Mechanics*. New York: Academic Press.

Meinnel, T., Mechulam, Y., Lazennec, C., Blanquet, S. & Fayat, G. (1993). Critical role of the acceptor stem of $tRNAs^{Met}$ in their aminoacylation by *Escherichia coli* methionyl-tRNA synthetase. *J. Mol. Biol.* **229,** 26–36.

Mezey, P. G. (1993). *Shape in Chemistry: an Introduction to Molecular Shape and Topology*. New York: VCH Publisher.

Mihalic, Z. & Trinajstic, N. (1992). A graph-theoretical approach to structure–property relationships. *J. Chem. Educ.* **69**(9), 701–712.

Oliveros, J., Payares. P., Diaz, D. & Mercado, J. (1995) PCDM2.0 Colombia: Universidad de Cartagena.

Randic, M. (1975). On characterization of molecular branching. *J. Am. Chem. Soc.* **97**(23), 6609–6615.

Randic, M. (1984). Non-empirical approach to structure–activity studies. *Int. J. Quantum Chem. Quantum Biol. Symp.* **11,** 137–153.

Randic, M. (1991a). Orthogonal molecular descriptor. *New. J. Chem.* **15,** 517–525.

Randic, M. (1991b). Resolution of ambiguities in structure–property studies by use of orthogonal descriptors. *New. J. Chem. Inf. Comput. Sci.* **31,** 311–320.

RANDIC, M. (1995). Molecular profiles novel geometry-dependent molecular descriptors. *New. J. Chem.* **19,** 781–791.

RANDIC, M. (1996). Orthosimilarity. *J. Chem. Inf. Comput. Sci.* **36**, 1092–1097.

RANDIC, M. & TRINAJSTIC, N. (1994). Isomeric variations in alkanes: boiling points of nonanes. *New. J. Chem.* **18,** 179–189.

RANDIC, M., SABLJIC, A., NIKOLIC, S. & TRINAJSTIC, N. (1988). A rational selection of graph-theoretical indices in the QSAR. *Int. J. Quant. Chem: Quant. Biol. Symp.* **15,** 267–285.

ROUVRAY, D. H. (1997). Are the concepts of chemistry all fuzzy?. In: *Concepts in Chemistry: a Contemporary Challenge* (Rouvray, D. H., ed.) pp. 1–13. England: RSP Ltd.

SAKS, M. E. & SAMPSON, J. R. (1995). Evolution of tRNA recognition systems and tRNA gene sequences. *J. Mol. Evol.* **40,** 509–518.

SCHIMMEL, P. (1989). Parameters for the molecular recognition of transfer RNAs. *Biochemistry* **28**(7), 2747–2759.

SHERMAN, J. M. & SÖLL, D. (1996). Aminoacyl-tRNA synthetases optimize both cognate tRNA recognition and discrimination against noncognate tRNAs. *Biochemistry* **35,** 601–607.

SPRINZL, M., HARTMANN, T., WEBER, J., BLANK, J. & ZEIDLER, R. (1989). Compilation of tRNA sequences and sequences of tRNA genes. *Nuc. Acids Res.* **17**(supplement).

STEWART, J. J. P. (1990). Semiempirical molecular orbital methods. In: *Reviews in Computational Chemistry* (Lipkowitz, K. B. & Boyd, D. B., eds), Vol. I, pp. 45–81, New York: VCH Publishers.

SUSSMAN, J. L. & KIM, S. H. (1976). Three-dimensional structure of a transfer RNA in two crystal forms. *Science* **192,** 853–858.

SZABO, A. & OSTLUND, N. (1996). *Modern Quantum Chemistry*. New York: Dover Publications.

TRINAJSTIC, N. (1983). *Chemical Graph Theory*, Vol. II, pp. 105–144. Cleveland OH: CRC Press.

VILLAVECES, J. L. & DAZA. E. E. (1990). On the topological approach to the concept of chemical structure. *Int. J. Quant. Chem.: Quant. Chem. Symp.* **24,** 97–106.

VILLAVECES, J. L. & DAZA. E. E. (1997). The concept of chemical structure. In: *Concepts in Chemistry: a Contemporary Challenge* (Rouvray, D. H., ed.) pp. 101–131.England: RSP Ltd.

WATSON, J. D. & CRICK, F. H. C. (1953). Molecular structure of nucleic acid. A structure for deoxirribose nucleic acid. *Nature* **171,** 737–738.

WESTHOF, E., DUMAS, P. & MORAS, D. (1985). Crystallographic refinement of yeast aspartic acid transfer RNA. *J. Mol. Biol.* **184,** 119–145.

WONG, J. T.-F. (1975). A coevolution theory of the genetic code. *Proc. Nat. Acad. Sci. U.S.A.* **72**(5), 1909–1912.

WOO, N. H., ROE, B. A. & RICH, A. (1980). Three-dimensional structure of *Escherichia coli* initiator $tRNA^{Met}$. *Nature* (*London*) **286,** 346–351.

WOOD, W. B., WILSON, J. H., BENBOW, R. M. & HOOD, L. E. (1981). *Biochemistry. A Problems Approach*, pp. 290–307. Reading MA: Benjamin/Cummings.

ZERNER, M. C. (1991). Semiempirical molecular orbital methods. In: *Reviews in Computational Chemistry* (Lipkowitz, K. B. & Boyd, D. B., eds.), Vol II, pp. 313–365. New York: VCH Publishers.

APPENDIX

Abbreviations

U:	uridine 5′P
s4U:	4-thiouridine 5′P
D:	dihydrouridine 5′P
Ψ:	pseudouridine 5′P
acp3U:	3-(3-amino-3-carboxypropyl)uridine 5′P
mnm5s2U:	5-methylaminomethyl-2-thiouridine 5′P
mnm5U:	5-methylaminomethyluridine 5′P
m5U:	ribosylthymine 5′P
cmo5U:	uridine 5-oxyacetic acid 5′P
Um:	2′-O-methyluridine 5′P
A:	Adenine 5′P
ms2i6A:	2-methylthio-N6-isopentenyladenosine 5′P
m6A:	N6-methyladenosine 5′P
m2A:	2-methyladenosine 5′P
t6A:	N6-threonylcarbamoyladenosine 5′P
m6t6A:	N6-methyl-N6-threonylcarbamoyladenosine 5′P
I:	inosine 5′P
C:	Cytidine 5′P
Cm:	2′-O-methylcytidine 5′P
ac4C:	N4-acetylcytidine 5′P
acp2c:	N2-(5-amino-5carboxypentil)cytidine 5′P
s2C:	2-thiocytidine 5′P
G:	Guanosine 5′P
Q:	queuosine 5′P
m7G:	7-methylguanosine 5′P
m1G:	1-methylguanosine 5′P
Gm:	2′-O-methylguanosine 5′P