B

- **backward Fukui function** → quantum-chemical descriptors (⊙ Fukui functions)
- **Balaban centric indices** → centric indices
- **Balaban DJ index** → Balaban distance connectivity index

Balaban distance connectivity index

The Balaban distance connectivity index (also called **distance connectivity index** or **average distance sum connectivity**), denoted as J, is one of the most known graph invariant. It is a very discriminating → *molecular descriptor* and its values do not increase substantially with molecule size or number of rings; it is defined in terms of the → *vertex distance degrees* σ_i, which are the row sums of the → *distance matrix* **D** [Balaban, 1982, 1983a]:

$$J = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} a_{ij} \cdot (\sigma_i \cdot \sigma_j)^{-1/2} = \frac{1}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} a_{ij} \cdot (\bar{\sigma}_i \cdot \bar{\sigma}_j)^{-1/2}$$

where σ_i and σ_j are the vertex distance degrees of the vertices v_i and v_j, a_{ij} the elements of the → *adjacency matrix* equal to one for pairs of adjacent vertices and zero otherwise, A the number of graph vertices, B the number of graph edges, and C the → *cyclomatic number*, that is, the number of rings. The denominator $C + 1$ is a normalization factor against the number of rings in the molecule. $\bar{\sigma}_i = \sigma_i / B$ is the **average vertex distance degree**; it was observed that within an isomeric series the average distance degrees are low in the more branched isomers.

To better discriminate among graph size, cyclicity, and branching, two modifications of the original Balaban distance connectivity index were later proposed [Balaban, Mills *et al.*, 2006]. The resulting indices, denoted as F and G, are defined as

$$F = B \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} a_{ij} \cdot (\sigma_i \cdot \sigma_j)^{-1/2} = (C+1) \cdot J$$

$$G = \frac{A^2 \cdot F}{A+C+1} = \frac{A^2 \cdot (C+1)}{A+C+1} \cdot J = \frac{A^2 \cdot B}{A+C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} a_{ij} \cdot (\sigma_i \cdot \sigma_j)^{-1/2}$$

where the summation goes over all pairs of graph vertices but only pairs of adjacent vertices are accounted for by means of the elements a_{ij} of the adjacency matrix. A, B, and C are the number of vertices, edges, and rings, respectively. The index G is defined in terms of the index F and seems to be able to account for → *molecular complexity*.

Molecular Descriptors for Chemoinformatics, Volume I: Alphabetical Listing
Roberto Todeschini and Viviana Consonni

ISBN: 978-3-527-31852-0

Balaban-like indices are molecular descriptors calculated applying the same mathematical formula as the distance connectivity index J, but substituting the vertex distance degrees σ_i by row sums VS_i of → *graph-theoretical matrices* other than the distance matrix **D** or other → *local vertex invariants* $\mathcal{L}_i$. They are usually derived from → *weighted matrices* computed from vertex- and edge-weighted graphs, which properly represent molecules containing heteroatoms and/or multiple bonds:

$$J(\mathbf{M}; w) = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} a_{ij} \cdot (VS_i(\mathbf{M}; w) \cdot VS_j(\mathbf{M}; w))^{-1/2}$$

where **M** is a graph-theoretical matrix, a_{ij} the elements of the adjacency matrix **A** equal to one for pairs of adjacent vertices and zero otherwise, A the number of graph vertices, w the → *weighting scheme*, and VS the → *vertex sum operator* applied to the matrix **M**.

This formula for the calculation of the Balaban-like indices was called **Ivanciuc–Balaban operator** by Ivanciuc and denoted as *IB* [Ivanciuc, Ivanciuc *et al.*, 1997; Ivanciuc, 2001c; Nikolić, Plavšić *et al.*, 2001].

The most general formula for computing Balaban-like indices in terms of any local vertex invariant is

$$J(\mathcal{L}) = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} a_{ij} \cdot (\mathcal{L}_i \cdot \mathcal{L}_j)^{-1/2}$$

where $\mathcal{L}_i$ and $\mathcal{L}_j$ are the local invariants of vertices v_i and v_j.

The **extended Ivanciuc–Balaban operator** was also defined as [Ivanciuc, Ivanciuc *et al.*, 2002e]

$$IB(\mathbf{M}; w, \lambda) = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} a_{ij} \cdot (VS_i(\mathbf{M}; w) \cdot VS_j(\mathbf{M}; w))^{\lambda}$$

where λ is a variable exponent.

The $\mathbf{J}_t$ **index** is a Balaban-like index defined as [Balaban, 1994a]

$$J_t = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} a_{ij} \cdot (t_i \cdot t_j)^{-1/2}$$

where t_i an t_j are local invariants for vertices v_i and v_j defined as a combination of vertex distance degree σ and → *vertex degree* δ to obtain a greater discriminant power among isomers:

$$t_i = \frac{\sigma_i}{\delta_i}$$

where δ_i is the ith vertex degree of the vertex v_i. The idea behind these LOVIs is that usually the vertices with the highest distance sums have the lowest vertex degrees, thus enhancing the intramolecular differences.

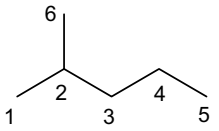
The J index for multigraphs is calculated by the distance sums of the → *multigraph distance matrix* $^*\mathbf{D}$ where the distances are obtained by weighting each edge with the reciprocal of its → *conventional bond order* (→ *relative topological distance*):

$$J({}^{*}\mathbf{D}) = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} a_{ij} \cdot ({}^{*}\sigma_i \cdot {}^{*}\sigma_j)^{-1/2}$$

where ${}^{*}\sigma_i$ and ${}^{*}\sigma_j$ are the → *multigraph distance degrees* of vertices v_i and v_j.

Example B1

Calculation of the Balaban distance connectivity index J and J_t index for 2-methylpentane. **D** is the topological distance matrix; σ_i and δ_i are the vertex distance sums and the vertex degrees. *B* equals 5 and *C* is zero.



$\mathbf{D} =$

Atom	1	2	3	4	5	6	σ_i	δ_i	t_i
1	0	1	2	3	4	2	12	1	12
2	1	0	1	2	3	1	8	3	2.667
3	2	1	0	1	2	2	8	2	4
4	3	2	1	0	1	3	10	2	5
5	4	3	2	1	0	4	14	1	14
6	2	1	2	3	4	0	12	1	12

$$J = \frac{B}{C+1} \times \left[(\sigma_1 \times \sigma_2)^{-1/2} + (\sigma_6 \times \sigma_2)^{-1/2} + (\sigma_2 \times \sigma_3)^{-1/2} + (\sigma_3 \times \sigma_4)^{-1/2} + (\sigma_4 \times \sigma_5)^{-1/2} \right] =$$
$$= 5 \times \left[(12 \times 8)^{-1/2} + (12 \times 8)^{-1/2} + (8 \times 8)^{-1/2} + (8 \times 10)^{-1/2} + (10 \times 14)^{-1/2} \right] = 2.6272$$

$$J_t = \frac{B}{C+1} \times \left[(t_1 \times t_2)^{-1/2} + (t_6 \times t_2)^{-1/2} + (t_2 \times t_3)^{-1/2} + (t_3 \times t_4)^{-1/2} + (t_4 \times t_5)^{-1/2} \right] =$$
$$= 5 \times \left[(12 \times 2.667)^{-1/2} + (12 \times 2.667)^{-1/2} + (2.667 \times 4)^{-1/2} + (4 \times 5)^{-1/2} + (5 \times 14)^{-1/2} \right] = 5.0141$$

To account for both bond multiplicity and heteroatoms, **Balaban modified distance connectivity indices** J^X and J^Y were proposed [Balaban, 1986a; Balaban, Catana *et al.*, 1990]. They are derived from the → *multigraph distance matrix* ${}^{*}\mathbf{D}$ as

$$J^X = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} a_{ij} \cdot ({}^{*}\sigma_i^X \cdot {}^{*}\sigma_j^X)^{-1/2}$$

$$J^Y = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} a_{ij} \cdot ({}^{*}\sigma_i^Y \cdot {}^{*}\sigma_j^Y)^{-1/2}$$

where *B* is the number of graph edges, *C* the number of graph rings, a_{ij} the elements of the adjacency matrix equal to one for pairs of adjacent vertices and zero otherwise, and *A* the number of graph vertices. Each edge is weighted by the inverse square root of the product of modified → *multigraph distance degrees* of the incident vertices according to the → *X weighting scheme* and → *Y weighting scheme*, respectively, as

$${}^{*}\sigma_i^X = X_i \cdot {}^{*}\sigma_i = X_i \cdot \sum_{j=1}^{A} [{}^{*}\mathbf{D}]_{ij} \quad \text{and} \quad X_i = 0.4196 - 0.0078 \cdot Z_i + 0.1567 \cdot G_i$$

$$^{*}\sigma_i^{Y} = Y_i \cdot {}^{*}\sigma_i = Y_i \cdot \sum_{j=1}^{A} [{}^{*}\mathbf{D}]_{ij} \quad \text{and} \quad Y_i = 1.1191 + 0.0160 \cdot Z_i - 0.0537 \cdot G_i$$

The quantities X and Y are recalculated atomic Sanderson electronegativities and covalent radii relative to carbon atom, respectively, obtained as a function of the atomic number Z_i and the group number of the Periodic System short form G_i of the atom; for atoms different from B, C, N, O, F, Si, P, S, Cl, As, Se, Br, Te, and I the X and Y values are set at one. X_i and Y_i are local indices that account for the presence of heteroatoms in the molecule.

The **3D Balaban index**, denoted as ^{3D}J, is a Balaban-like index derived from the → *geometry matrix* **G** as [Mihalić, Nikolić *et al.*, 1992]

$$^{3D}J(\mathbf{G}) = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} a_{ij} \cdot ({}^{G}\sigma_i \cdot {}^{G}\sigma_j)^{-1/2}$$

where $^{G}\sigma_i$ and $^{G}\sigma_j$ are the → *geometric distance degrees* of the vertices v_i and v_j, which are the row sums of the geometry matrix.

The **E-state topological parameter**, denoted as TI^E, is derived by applying the → *Ivanciuc–Balaban operator* to the → *E-state index* values used to characterize molecule atoms [Voelkel, 1994]:

$$TI^E = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} a_{ij} \cdot (S_i \cdot S_j)^{-1/2}$$

where S_i and S_j are the E-state values for the vertices v_i and v_j.

It has to be pointed out that the proposed formula for the E-state topological parameter cannot be used for every molecule because it presents two drawbacks: (1) TI^E cannot be calculated when there exists one atom in the molecule with negative E-state value *S*; (2) TI^E assumes very large values even when one *S* value tends to zero.

To overcome these drawbacks of the original formula, an alternative formula [Authors, This Book], adopted in the → *DRAGON descriptors*, is the following:

$$TI^E = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} a_{ij} \cdot (1 + e^{S_i} \cdot e^{S_j})^{-1/2}$$

In this case, descriptor values can be obtained also for molecules with negative *S* values; moreover, they are in a suitable range for any molecule.

Other Balaban-like indices are → *Balaban-like information indices*, → *Barysz index*, → *reversed Balaban index*, → *Harary–Balaban index*, → *Balaban-like resistance index*, → *variable Balaban index*, → L_z *index*, → *quotient Balaban index of the first kind*, and → *quotient Balaban index of the second kind*.

The **Balaban DJ index** was still defined in terms of modified vertex distance degrees σ_i but using the formula of the → *matrix sum indices* as [Balaban and Diudea, 1993]

$$DJ = \sum_{i=1}^{A} dj_i = \sum_{i=1}^{A} \sum_{j=1}^{A} a_{ij} \cdot \left(\frac{\sigma_i}{w_i(1+f_i)} \cdot \frac{\sigma_j}{w_j(1+f_j)} \right)^{-1/2}$$

where A is the number of graph vertices, f the → *multigraph factor*, w a weighting factor accounting for heteroatoms, a_{ij} the elements of the adjacency matrix equal to one for pairs of adjacent vertices and zero otherwise, and dj_i → *local vertex invariants* accounting for heteroatoms and bond multiplicity. When the factor w is equal to one and the multigraph factor is equal to zero then the index DJ is related to the Balaban index J by the following:

$$DJ = 2 \cdot \frac{C+1}{B} \cdot \mathrm{J} = 2 \cdot \frac{C+1}{B} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} a_{ij} \cdot (\sigma_i \cdot \sigma_j)^{-1/2}$$

🕮 [Balaban and Quintas, 1983; Barysz, Jashari *et al.*, 1983; Balaban and Filip, 1984; Balaban, Ionescu-Pallas *et al.*, 1985; Sabljić, 1985; Mekenyan, Bonchev *et al.*, 1987; Balaban and Ivanciuc, 1989; Balaban, Ciubotariu *et al.*, 1990; Balaban, Kier *et al.*, 1992a; Nikolić, Medicsaric *et al.*, 1993; Guo and Randić, 1999; Montanari, Cass *et al.*, 2000; Estrada and Gutierrez, 2001; Nikolić, Plavšić *et al.*, 2001; Balaban, Mills *et al.*, 2002; Ivanciuc, 2002a; Ivanciuc, Ivanciuc *et al.*, 2002e]

- **Balaban ID number** → ID numbers
- **Balaban-like information indices** → topological information indices
- **Balaban-like indices** → Balaban distance connectivity index
- **Balaban-like resistance index** → resistance matrix
- **Balaban modified distance connectivity indices** → Balaban distance connectivity index
- **Barnard keys** ≡ *BCI keys* → substructure descriptors (⊙ structural keys)
- **Baroni–Urbani similarity coefficient** → similarity/diversity (Table S9)
- **Bartell resonance energy** → delocalization degree indices
- **barycenter** ≡ *center of mass* → center of a molecule
- **Barysz index** → weighted matrices (⊙ weighted distance matrices)
- **Barysz distance matrix** → weighted matrices (⊙ weighted distance matrices)
- **basic graph** ≡ *Sachs graph* → graph
- **basis of descriptors** → vectorial descriptors
- **Bate–Smith–Westall retention index** → chromatographic descriptors
- **BC(DEF) coordinates** ≡ *BC(DEF) parameters*

■ **BC(DEF) parameters** (≡ *BC(DEF) coordinates*)
Proposed by Cramer III in 1980, they are five → *principal properties* (i.e., significant components calculated by → *Principal Component Analysis*) of a data matrix comprised of the values of six physico-chemical properties collected for 114 diverse liquid-state compounds [Cramer III, 1980a, 1983b].

The → *physico-chemical properties* used to derive BC(DEF) descriptors are activity coefficient in water, → *octanol–water partition coefficient*, boiling point, → *molar refractivity*, liquid state → *molar volume*, and heat of vaporization. The eigenvalues and corresponding cumulative explained variances of the five principal properties (denoted by B, C, D, E, and F) are reported in Table B1. It can be noted that the first two principal properties B and C already explain 95.7% of the original variance of the six physico-chemical properties; further analysis using different compounds and properties showed B and C to be independent of the data set used in their derivation, identifying them as measures of molecular bulk and cohesiveness, respectively. The other three parameters, D, E, and F, are of minor importance, however they were

retained due to their significance in the correlations involving some physico-chemical properties.

In general BC(DEF) parameters describe molecular properties related to nonspecific intermolecular interactions in the liquid state and could therefore be useful in predicting biological activity or physico-chemical properties depending on such nonspecific interactions; 29 linear models were calculated by multivariate regression analysis that correlate BC(DEF) parameters to 29 different physico-chemical properties.

Table B1 Eigenvalues and cumulative variances of BC(DEF) principal properties.

Principal property	Eigenvalue	Cumulative variance (%)
B	3.870	64.4
C	1.870	95.7
D	0.168	98.5
E	0.045	99.2
F	0.029	99.7

Calculation of BC(DEF) parameters for new compounds different from the original 114 compounds can be accomplished either by their physico-chemical properties or their structure.

The property-derived BC(DEF) values are calculated from a set of known property values and the corresponding property models previously derived from the original 114×6 data set. A property model has the general form:

$$y = b_0 + b_1 \cdot \mathrm{B} + b_2 \cdot \mathrm{C} + b_3 \cdot \mathrm{D} + b_4 \cdot \mathrm{E} + b_5 \cdot \mathrm{F}$$

where y is the known experimental property value and b the known regression coefficients taken from the specific property model. Using a set of at least six property models, all the BC(DEF) values together with their confidence intervals can be obtained as solutions of the linear equation system [Cramer III, 1983a]. In this case, the physico-chemical properties should be considered as independent variables and the BC(DEF) values as dependent variables.

Alternatively, the BC(DEF) values can be obtained by → *additive-constitutive models* based on the contributions of individual fragments and some correction factors to each parameter [Cramer III, 1980b]. A hierarchical additive-constitutive model was derived by multivariate regression of the BC(DEF) values of 112 original compounds (water and methane were excluded from the model) and occurrence frequencies of 35 molecular fragments. Moreover, in the same way a linear additive-constitutive model was also proposed; the fragment contributions to BC(DEF) parameters are reported in Table B2.

Table B2 Fragment contributions to BC(DEF) parameters.

Fragment	B	C	D	E	F
Intercept	−0.506	−0.056	0.007	0.031	0.028
−H	0.066	0.018	−0.027	−0.019	−0.019
$-CH_3$	0.142	−0.020	−0.016	−0.023	−0.015
$-CH_2-$	0.076	−0.038	0.011	−0.004	0.003
>CH−	0.003	−0.058	0.053	0.018	0.015
>C<	−0.075	−0.076	0.091	0.043	0.034

(*Continued*)

Table B2 (*Continued*)

Fragment	B	C	D	E	F
−CH=CH−	0.147	−0.043	0.028	0.010	0.003
−CH=CH_2	0.212	−0.025	0.000	−0.009	−0.015
>CH=CH_2	0.147	−0.043	0.028	0.010	0.003
−C≡CH	0.171	0.074	0.027	0.002	−0.012
−C_6H_5	0.467	−0.007	0.012	0.007	−0.017
≈CH−(aromatic)	0.088	0.002	−0.007	0.001	−0.003
−Naphthyl	0.766	0.018	−0.026	0.024	−0.028
−Cyclohexyl	0.489	−0.148	0.004	−0.029	−0.009
−F[a]	0.078	0.088	0.009	−0.019	−0.020
−Cl	0.165	0.087	−0.024	−0.012	−0.021
−Br[a]	0.213	0.095	−0.033	−0.008	−0.020
−I[a]	0.302	0.103	−0.056	−0.010	−0.031
−CF_3	0.150	0.017	0.035	−0.037	−0.013
−CCl_3	0.410	0.015	−0.009	−0.017	−0.017
−OH[a]	0.202	0.324	−0.012	−0.015	0.003
−O−[a]	0.044	0.155	0.061	0.019	−0.022
−C=O−[a]	0.135	0.246	0.061	0.023	−0.021
−CH=O[a]	0.219	0.244	0.010	−0.014	−0.027
−COO−[a]	0.167	0.170	0.062	0.015	−0.027
−COOH[a]	0.323	0.342	−0.011	−0.017	0.008
−NH_2[a]	0.167	0.269	0.037	0.027	−0.014
−NH−[a]	0.082	0.251	0.095	0.056	−0.010
−N−[a]	−0.006	0.189	0.125	0.069	0.014
−CN	0.241	0.269	−0.007	−0.023	−0.041
−N=[a] (pyridine)	0.102	0.183	0.031	−0.011	−0.020
−NO_2[a]	0.238	0.241	−0.012	−0.027	−0.037
−$CONH_2$[a]	0.444	0.499	−0.019	−0.039	−0.012
−S−[a]	0.136	0.130	0.028	0.032	−0.020
−SH[a]	0.231	0.155	−0.026	−0.011	−0.013

[a]Value when attached to an aliphatic system.

- **BCF** ≡ *bioconcentration factor* → environmental descriptors
- **BCI keys** → substructure descriptors (⊙ structural keys)
- **BCUT descriptors** → spectral indices (⊙ Burden eigenvalues)
- **benzene-likeliness index** → delocalization degree indices
- **Bertz branching index** → molecular complexity (⊙ molecular branching)
- **Bertz complexity index** → molecular complexity
- **Bertz–Herndon relative complexity index** → molecular complexity
- **best hydrophilic volumes** → grid-based QSAR techniques (⊙ VolSurf descriptors)
- **best hydrophobic volumes** → grid-based QSAR techniques (⊙ VolSurf descriptors)
- **Beteringhe–Filip–Tarko descriptor** → MPR approach
- **Betti numbers** → Mezey 3D shape analysis
- **betweenness centrality** → center of a graph
- **Bhattacharyya distance** → similarity/diversity (⊙ Table S7)
- **bilinear indices** → TOMOCOMD descriptors

- **binary descriptors** → indicator variables
- **binary distance measures** → similarity/diversity
- **binary QSAR analysis** → scoring functions
- **binary sparse matrix** → algebraic operators (⊙ sparse matrices)
- **binding affinity** → drug design
- **binding property pairs** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **binding property torsions** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **binding site cavity** → drug design
- **binormalized centric index** → centric indices (⊙ Balaban centric index)
- **binormalized quadratic index** → Zagreb indices
- **bioaccumulation** → environmental indices
- **bioconcentration factor** → environmental indices

biodescriptors

These are numerical quantities encoding information about biochemical systems, addressing the problem of numerical characterization of macromolecules like proteins and nucleic acids and complex systems like proteomics maps.

The term *biodescriptors* was introduced by analogy with the common molecular descriptors, which are *chemodescriptors* since they are derived from the molecular structure of chemicals.

Biodescriptors cover a large field of mathematical strategies, including graph-theoretical approaches [Randić and Basak, 2002] and general theoretical links between topological structure of molecules and molecular biology networks were also investigated [Bonchev and Buck, 2007].

Most of biodescriptors were proposed to characterize sequences of peptides and nucleic acids trying to account for the sequential disposition of the constitutive elements of the considered biological system.

If each element of the sequence (e.g., an amino acid) is considered as an "atomic" unit, the physico-chemical properties of each atomic unit can be evaluated and used as "local bioinvariant" for the calculation of several descriptors defined for classical organic molecules.

The term **Quantitative Sequence-Activity Models** (QSAMs) is used instead of quantitative structure-activity relationships when referred to the research on relationships between structure and activities of molecules of biological interest [Jonsson, Norberg *et al.*, 1993].

Below, some common strategies for description of peptide and DNA sequences are briefly reviewed starting from amino acid descriptors, which are of fundamental importance for deriving most of the protein descriptors. The last section deals with some approaches for proteomics map characterization.

• amino acid descriptors

Due to the relevance and the complexity of proteins, some descriptors were defined to represent amino acid side chains, these being responsible for the packing of the regular elements of secondary structure and then for the tertiary structure of a protein. As a consequence, the structure of a protein can be expressed quantitatively by means of side chain amino acid properties. Starting from the pioneering work of Sneath, who described peptide sequences by semiquantitative experimental parameters of the 20 coded amino acids [Sneath, 1966], several amino acid descriptors have been proposed that contain information about properties of side chains of amino acids.

Ten principal properties were calculated by → *Principal Component Analysis* on 188 physico-chemical properties for the 20 coded amino acids [Kidera, Konisci *et al.*, 1985a, 1985b]. These 10 properties were called **KOKOS descriptors** by Pogliani on the basis of the Authors' names [Pogliani, 1994a]; they describe most of the conformational, bulk, hydrophobicity, α-helix, and β-structure properties of amino acids.

To calculate → *ACC transforms* of peptide sequences, each amino acid in the peptide sequence was described by three orthogonal → *z-scores* (Table B3), derived from a → *Principal Component Analysis* on 29 → *physico-chemical properties* of the 20 coded amino acids [Hellberg, Sjöström *et al.*, 1986, 1987a, 1987b; Wold, Eriksson *et al.*, 1987; Jonsson, Eriksson *et al.*, 1989]. These → *principal properties* were later extended to 87 amino acids including natural amino acids [Sandberg, Eriksson *et al.*, 1998]. Moreover, amino acid 3D principal properties (Table B3) were also derived from → *molecular interaction fields* [Norinder, 1991; Cocchi and Johansson, 1993; Cruciani, Baroni *et al.*, 2004].

Table B3 Z-scores (columns 4–6) and principal properties (PP) from molecular interaction fields, in the original form (columns 7–9) and reoriented and scaled between –1 and + 1 (columns 10–12) [Cruciani, Baroni *et al.*, 2004].

ID	Code	Code	z_1	z_2	z_3	PP_1 Polarity	PP_2 Hydroph.	PP_3 H-bond	PP^S_1 Polarity	PP^S_2 Hydroph.	PP^S_3 H-bond
1	Ala	A	10.07	–1.73	0.09	3.19	–2.21	–0.82	–0.96	–0.76	0.31
2	Arg	R	2.88	2.52	–3.44	–2.94	3.44	–2.56	0.80	0.63	0.99
3	Asn	N	3.22	1.45	0.84	–3.03	–1.45	–0.04	0.82	–0.57	0.02
4	Asp	D	3.64	1.13	2.36	–3.66	–2.74	2.60	1.00	–0.89	–1.00
5	Cys	C	0.71	–0.97	4.13	1.77	–1.02	–0.49	–0.55	–0.47	0.19
6	Glu	E	3.08	0.39	–0.07	–3.45	–1.34	2.58	0.94	–0.54	–0.99
7	Gln	Q	2.18	0.53	–1.14	–2.89	–0.34	0.99	0.78	–0.30	–0.38
8	Gly	G	2.23	–5.36	0.30	2.91	–3.20	–1.26	–0.88	–1.00	0.49
9	His	H	2.41	1.74	1.11	–2.51	0.43	–0.95	0.67	–0.11	0.37
10	Ile	I	–4.44	–1.68	–1.03	3.11	0.65	0.47	–0.94	–0.05	–0.18
11	Leu	L	–4.19	–1.03	–0.98	2.99	0.99	0.61	–0.90	0.03	–0.24
12	Lys	K	2.84	1.41	–3.14	–2.25	1.27	–2.60	0.60	0.10	1.00
13	Met	M	–2.49	–0.27	–0.41	2.69	1.01	0.22	–0.82	0.03	–0.08
14	Phe	F	–4.92	1.30	0.45	2.80	2.81	1.52	–0.85	0.48	–0.58
15	Pro	P	–1.22	0.88	2.23	2.65	–0.76	0.17	–0.81	–0.40	–0.07
16	Ser	S	1.96	–1.63	0.57	–1.58	–2.46	–1.48	0.41	–0.82	0.57
17	Thr	T	0.92	–2.09	–1.40	–1.55	–1.72	–0.95	0.40	–0.64	0.37
18	Trp	W	–4.75	3.65	0.85	–0.38	4.94	1.11	0.06	1.00	–0.47
19	Tyr	Y	–1.39	2.32	0.01	–1.23	2.59	0.51	0.31	0.42	–0.20
20	Val	V	–2.69	–2.53	–1.29	3.33	–0.87	0.36	–1.00	–0.43	–0.14

VHSE descriptor (*principal component score Vector of Hydrophobic, Steric, and Electronic properties*) is a → *vectorial descriptor*, containing eight principal properties, derived from Principal Component Analysis on 50 physico-chemical properties of the 20 coded amino acids [Mei, Liao *et al.*, 2005]. $VHSE_1$ and $VHSE_2$ are related to hydrophobic properties of amino acids, $VHSE_3$ and $VHSE_4$ to steric properties, and $VHSE_5$–$VHSE_8$ to electronic properties (Table B4).

Table B4 VHSE descriptor for the 20 coded amino acids [Mei, Liao *et al.*, 2005].

ID	Code	Code	$VHSE_1$	$VHSE_2$	$VHSE_3$	$VHSE_4$	$VHSE_5$	$VHSE_6$	$VHSE_7$	$VHSE_8$
1	Ala	A	0.15	–1.11	–1.35	–0.92	0.02	–0.91	0.36	–0.48
2	Arg	R	–1.47	1.45	1.24	1.27	1.55	1.47	1.30	0.83
3	Asn	N	–0.99	0.00	–0.37	0.69	–0.55	0.85	0.73	–0.80
4	Asp	D	–1.15	0.67	–0.41	–0.01	–2.68	1.31	0.03	0.56
5	Cys	C	0.18	–1.67	–0.46	–0.21	0.00	1.20	–1.61	–0.19
6	Gln	Q	–0.96	0.12	0.18	0.16	0.09	0.42	–0.20	–0.41
7	Glu	E	–1.18	0.40	0.10	0.36	–2.16	–0.17	0.91	0.02
8	Gly	G	–0.20	–1.53	–2.63	2.28	–0.53	–1.18	2.01	–1.34
9	His	H	–0.43	–0.25	0.37	0.19	0.51	1.28	0.93	0.65
10	Ile	I	1.27	–0.14	0.30	–1.80	0.30	–1.61	–0.16	–0.13
11	Leu	L	1.36	0.07	0.26	–0.80	0.22	–1.37	0.08	–0.62
12	Lys	K	–1.17	0.70	0.70	0.80	1.64	0.67	1.63	0.13
13	Met	M	1.01	–0.53	0.43	0.00	0.23	0.10	–0.86	–0.68
14	Phe	F	1.52	0.61	0.96	–0.16	0.25	0.28	–1.33	–0.20
15	Pro	P	0.22	–0.17	–0.50	0.05	–0.01	–1.34	–0.19	3.56
16	Ser	S	–0.67	–0.86	–1.07	–0.41	–0.32	0.27	–0.64	0.11
17	Thr	T	–0.34	–0.51	–0.55	–1.06	–0.06	–0.01	–0.79	0.39
18	Trp	W	1.50	2.06	1.79	0.75	0.75	–0.13	–1.01	–0.85
19	Tyr	Y	0.61	1.60	1.17	0.73	0.53	0.25	–0.96	–0.52
20	Val	V	0.76	–0.92	–0.17	–1.91	0.22	–1.40	–0.24	–0.03

SSIA descriptors (*Scores of Structural Information for Amino acids)* are *z*-scores derived from Principal Component Analysis on → *3D VAIF descriptors* for the 20 coded amino acids [Zhou, Zhou *et al.*, 2006].

T-scale is a five-dimensional vectorial descriptor (Table B5) derived from Principal Component Analysis on 67 → *topological indices* of 135 amino acids [Tian, Zhou *et al.*, 2007].

Table B5 T-scale for the 20 coded amino acids [Tian ian, Zhou *et al.*, 2007].

ID	Code	T_1	T_2	T_3	T_4	T_5	ID	Code	T_1	T_2	T_3	T_4	T_5
1	Ala	–9.11	–1.63	0.63	1.04	2.26	11	Leu	–4.38	0.28	–0.49	1.45	0.02
2	Arg	0.23	3.89	–1.16	–0.39	–0.06	12	Lys	–2.59	2.34	–1.69	0.41	–0.21
3	Asn	–4.62	0.66	1.16	–0.22	0.93	13	Met	–4.08	0.98	–2.34	1.64	–0.79
4	Asp	–4.65	0.75	1.39	–0.40	1.05	14	Phe	0.49	–0.94	–0.63	–1.27	–0.44
5	Cys	–7.35	–0.86	–0.33	0.80	0.98	15	Pro	–5.11	–3.54	–0.53	–0.36	–0.29
6	Gln	–3.00	1.72	0.28	–0.39	0.33	16	Ser	–7.44	–0.65	0.68	–0.17	1.58
7	Glu	–3.03	1.82	0.51	–0.58	0.43	17	Thr	–5.97	–0.62	1.11	0.31	0.95
8	Gly	–10.61	–1.21	–0.12	0.75	3.25	18	Trp	5.73	–2.67	–0.07	–1.96	–0.54
9	His	–1.01	–1.31	0.01	–1.81	–0.21	19	Tyr	2.08	–0.47	0.07	–1.67	–0.35
10	Ile	–4.25	–0.28	–0.15	1.40	–0.21	20	Val	–5.87	–0.94	0.28	1.10	0.48

VSW descriptor (*Vector of principal component Scores for WHIMs*) is a vectorial descriptor derived from Principal Component Analysis on the 99 → *WHIM descriptors* calculated for the

20 coded amino acids [Tong, Liu *et al.*, 2008]. The VSW descriptor contains nine principal properties for each amino acid (Table B6).

Table B6 VSW descriptor for the 20 coded amino acids [Tong, Liu *et al.*, 2008].

ID	Code	Code	VSW_1	VSW_2	VSW_3	VSW_4	VSW_5	VSW_6	VSW_7	VSW_8	VSW_9
1	Ala	A	–11.634	–1.897	1.978	–2.606	–1.715	–2.031	–0.818	0.640	1.080
2	Arg	R	11.871	–2.870	2.748	1.257	1.143	–0.477	–2.722	1.769	1.440
3	Asn	N	–5.350	7.683	4.117	4.174	4.249	–0.189	–1.065	–0.128	–0.839
4	Asp	D	–4.027	2.993	–3.359	–3.770	1.923	0.672	1.557	1.210	–0.301
5	Cys	C	–5.650	–2.879	–2.990	2.344	0.878	–1.945	1.069	–1.562	2.619
6	Gln	Q	2.176	–2.400	0.845	3.572	–1.201	–1.092	–0.114	0.052	–0.882
7	Glu	E	2.367	0.152	–4.048	0.804	2.037	0.990	1.087	2.345	0.166
8	Gly	G	–11.782	–13.698	3.470	0.201	0.965	3.074	0.440	–0.282	–0.702
9	His	H	2.339	0.361	–1.565	–1.076	2.002	–1.041	–1.300	–2.067	–1.570
10	Ile	I	0.412	6.404	–1.244	–1.622	–1.234	1.424	0.041	–0.179	–0.419
11	Leu	L	0.269	8.116	2.897	0.982	–1.934	3.156	0.058	–1.196	1.530
12	Lys	K	9.006	–2.097	–3.355	2.392	0.378	1.327	–0.462	–0.186	0.332
13	Met	M	4.363	–1.665	–3.977	–1.023	0.130	0.817	–0.540	–1.703	0.084
14	Phe	F	7.264	–4.366	–1.091	1.621	–3.196	–0.093	0.408	–0.110	–0.667
15	Pro	P	–5.307	3.184	0.595	4.277	–1.525	–1.512	3.067	0.320	–0.683
16	Ser	S	–9.155	2.320	–0.499	–2.269	–0.129	0.372	–0.853	0.997	0.546
17	Thr	T	–4.220	–0.272	–1.391	–2.538	1.070	–1.557	–1.338	–0.608	–0.558
18	Trp	W	11.702	0.162	5.620	–4.919	0.564	–0.732	2.216	–0.973	0.320
19	Tyr	Y	8.540	–1.526	1.741	–1.285	0.109	–0.953	1.142	1.100	–0.575
20	Val	V	–3.184	2.294	–0.492	–0.516	–4.515	–0.210	–1.874	0.561	–0.920

→ *Isotropic surface area* and → *electronic charge index* were proposed as the descriptors of steric character and local dipole of amino acid side chains [Collantes and Dunn III, 1995]. Moreover, amino acids were described, for example, by → *substituent descriptors* [Charton and Charton, 1982; Charton, 1990], → *connectivity indices* [Gardner, 1980; Pogliani, 1992a, 1992b, 1993a, 1993b, 1994a, 1994c, 1995a, 1996a, 1997a, 1997c, 1999a; Lučić, Nikolić *et al.*, 1995b], → *G-WHIM descriptors* [Zaliani and Gancia, 1999], → *side chain topological index* [Raychaudhury, Banerjee *et al.*, 1999], → *Molecular Holographic Distance Vector* [Liu, Yin *et al.*, 2001a], and → *WHIM descriptors* (Table B7) [Mauri, Ballabio *et al.*, 2008].

Table B7 Global *WHIM* descriptors for the 20 coded amino acids [Mauri, Ballabio *et al.*, 2008].

ID	Code	Code	Am	Km	Dm	ID	Code	Code	Am	Km	Dm
1	Ala	A	0.3634	0.4430	0.2330	11	Leu	L	1.1486	0.5210	0.3370
2	Arg	R	1.9266	0.7980	0.3130	12	Lys	K	1.5369	0.8120	0.3340
3	Asn	N	0.9274	0.4970	0.2960	13	Met	M	1.0385	0.4610	0.2940
4	Asp	D	0.8575	0.4290	0.3700	14	Phe	F	1.3731	0.5790	0.2710
5	Cys	C	0.6683	0.4990	0.2530	15	Pro	P	0.5536	0.4870	0.2910
6	Gln	Q	0.9970	0.5840	0.3810	16	Ser	S	0.4656	0.3910	0.2810
7	Glu	E	1.1128	0.4040	0.3260	17	Thr	T	0.6918	0.4850	0.3070
8	Gly	G	0.2343	0.5420	0.3220	18	Trp	W	2.3415	0.6410	0.2970

(*Continued*)

Table B7 (*Continued*)

ID	Code	Code	Am	Km	Dm	ID	Code	Code	Am	Km	Dm
9	His	H	1.0631	0.7590	0.2740	19	Tyr	Y	1.6385	0.6620	0.2840
10	Ile	I	0.9845	0.5820	0.2660	20	Val	V	0.7066	0.5100	0.2980

Am, Km, and Dm are the size, shape, and atom density global WHIM descriptors, respectively, weighted by the atomic masses.

One of the most comprehensive resources of amino acid properties freely available on line is the amino acid index database (*AAindex*), which includes numerical indices representing various physico-chemical, biochemical, and statistical properties of amino acids and pairs of amino acids. AAindex database has been made publicly available by the Japanese GenomeNet database service (http://www.genome.jp/aaindex/).

🕮 [Sneath, 1966; Wolfenden, Andersson *et al.*, 1981; Fauchère and Pliška, 1983; Sjöström and Wold, 1985; Abraham and Leo, 1987; Skagerberg, Sjöström *et al.*, 1987; Nakayama, Shigezumi *et al.*, 1988; Tsai, Testa *et al.*, 1991; El Tayar, Tsai *et al.*, 1992; El Tayar and Testa, 1993; Naray-Szabo and Balogh, 1993; Eriksson, Hermens *et al.*, 1995; Šoškić, Klaić *et al.*, 1995; Vallat, Gaillard *et al.*, 1995; Chapman, 1996; Pogliani, 1997a, 2000b; Randić and Krilov, 1997a; Sotomatsu-Niwa and Ogino, 1997; Pérez and Contreras, 1998; Grgas, Nikolić *et al.*, 1999; Raychaudhury and Nandy, 1999; Stein, Gordon *et al.*, 1999; Tao, Wang *et al.*, 1999; Testa, Raynaud *et al.*, 1999; Alifrangis, Christensen *et al.*, 2000; Nyström, Andersson *et al.*, 2000; Randić, Mills *et al.*, 2000; Nikolić and Raos, 2001; Oprea and Gottfries, 2001b; Pacios, 2001; Wold, Sjöström *et al.*, 2001; Shen, LeTiran *et al.*, 2002; Estrada, 2004b; Marrero-Ponce, Marrero *et al.*, 2004; Boon, Van Alsenoy *et al.*, 2005; Restrepo and Villaveces, 2005; Liang, Zhou *et al.*, 2006; Zhang, Ding *et al.*, 2008]

• **peptide sequences** (≡ *amino acid sequences*)
A peptide sequence is the ordered sequence of amino acid residues, connected by peptide bonds, which compose a peptide or protein. The sequence is generally reported from the N-terminal end containing free amino group to the C-terminal end containing free carboxyl group. Peptide sequences are often called **protein sequences** if they represent the protein primary structure. The primary structure of a peptide (or protein) is just the sequence of amino acids along its backbone. The secondary structure of proteins is defined by patterns of hydrogen bonds between backbone amide and carboxyl groups, while the tertiary structure is the three-dimensional structure, as defined by the atomic coordinates.

Side chains of amino acids are responsible for the packing of the regular elements of secondary structure and then for the tertiary structure of a protein. As a consequence, the structure of a protein can be expressed quantitatively by means of side chain amino acid properties. Several → *amino acid descriptors* have been proposed, which contain information about properties of side chains of amino acids.

A general index based on the → *total information content* of biological compounds, such as peptide sequences, is the **information index on amino acid composition**, defined as [Bonchev, 1983]

$$I_{\mathrm{AAC}} = k \cdot \left(\ln N! - \sum_{g=1}^{G} \ln n_g! \right)$$

where k is the Boltzmann constant, N the total number of amino acid residues, G the number of → *equivalence classes*, that is, the number of different amino acid residues, and n_g the number of

amino acid residues of type *g*, that is, belonging to the *g*th equivalence class. Unlike other information indices, factorials are used in the expression to take into account combinations of amino acid residues.

A simple approach to protein description consists of representing a protein by a sequence of properties of its constituent amino acids. Each amino acid is described by one ore more properties and therefore the total number of protein descriptors is given by the product of the number of amino acids in the protein and the number of selected amino acid properties. As this number of descriptors increases very fast with the size of proteins, this approach is usually applied to small- and medium-size peptides. Moreover, in QSAR studies that require → *uniform-length descriptors*, it can be used only to describe a series of peptide analogues, which are peptide sequences with the same length. To enable QSAR studies of peptide sequences with different length, some method is required that is able to translate the peptide sequences into → *vectorial descriptors* with the same number of variables. For example, → *ACC transforms* were applied to compress information about → *principal properties* of amino acids into peptide sequences with different length.

To characterize size and shape of side chains in amino acids, a topological descriptor was proposed [Raychaudhury, Banerjee *et al.*, 1999] based on a graph-theoretical approach applied to rooted weighted molecular graphs (hydrogen included) representing the side chains. Each vertex of the chain other than the link vertex (C_α carbon atom) is weighted and all shortest weighted paths between the link vertex C_α (assumed at zero position) and terminal vertices are taken into account; the weight of each path is given by the sum of the atomic weights of all involved atoms. Moreover, if there are more than one shortest path between two vertices, then the selected path is that with the minimum sum of the weights of its vertices.

A probability value p_i is assigned to the directed path connecting the link vertex to each *i*th terminal vertex (Figure B1), calculated as the following:

$$p_i = (\delta'_1 \cdot \ldots \cdot \delta'_{i-1})^{-1}$$

where δ' is the number of incident bonds of each atom involved in the path without considering those already counted at the previous step. Bonds in the rings not involved in any path should be deleted to get probability values.

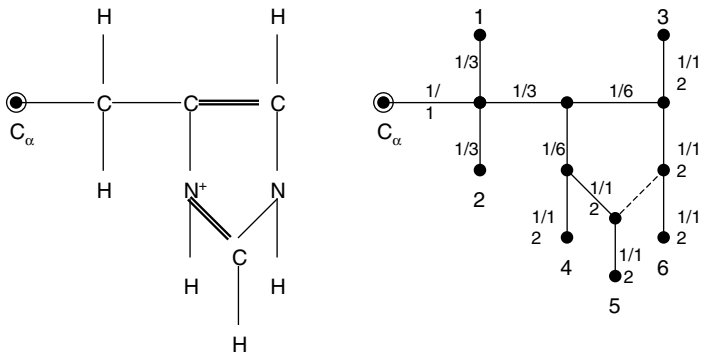


Figure B1 Histidine molecule and corresponding molecular graph. The number associated to each bond is the probability corresponding to the incident vertex, calculated starting from C_α.

By using the calculated probability values, the path value $P_{i,\mathrm{p}}$ is calculated as

$$P_{i,\mathrm{p}} = \mathrm{p}_i \cdot \sum\nolimits_k w_k$$

where the sum runs over all the vertices between the link atom and the ith vertex; w represents the weights of the atoms involved in the path. A molecular shape and size related index $M_{\mathrm{S}}^{\mathrm{S}}$ (Figure B2), here called **side chain topological index**, is calculated for the link vertex C_α as the sum of all the path values:

$$M_{\mathrm{S}}^{\mathrm{S}} = \sum_{i=1}^{\mathrm{N_T}} P_{i,\mathrm{p}}$$

where $\mathrm{N_T}$ is the number of terminal vertices in the side chain.

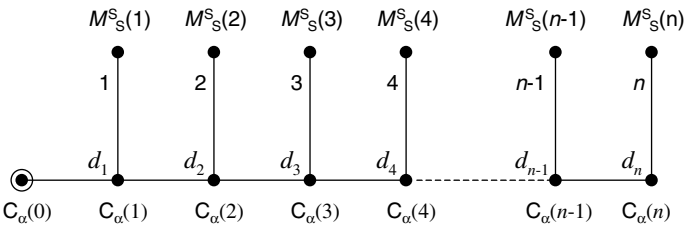


Figure B2 Amino acid sequence starting from $C_\alpha(0)$.

From this index, a descriptor for a sequence of amino acids, called **distance exponent index** D^x, was also proposed, defined as

$$D^x = \sum\nolimits_k (M_{\mathrm{S}}^{\mathrm{S}})_k \cdot d_k^x$$

where the sum runs over the considered sequence, being each term at topological distance d_k from the C_α representing the origin of the sequence. Each side chain topological index $M_{\mathrm{S}}^{\mathrm{S}}$ is calculated independently of its link atom C_α. The exponent x may take any real values; values $x = -3$ and -4 were usefully proved in modeling side chain properties [Raychaudhury and Klopman, 1990].

A general approach to derive protein descriptors is based on representing a protein by a **macromolecular graph** in which vertices represent the α-carbon of the amino acid residues and edges represent the covalent peptidic bonds. Loops on vertices can be added to account for noncovalent interactions within a chain or between chains.

Then, → *amino acid descriptors* are used to weight vertices in the macromolecular graph in the same way as the atomic properties are used to weight vertices in a common molecular graph. At this stage, all the classic → *graph invariants* can be calculated from the weighted macromolecular graph and used as the protein descriptors in QSAR studies. Examples of these descriptors are linear, bilinear, and quadratic → *TOMOCOMD descriptors*.

By means of the macromolecular graph, the peptide description is simplified, considering that (a) the physico-chemical properties of the amino acids are responsible for the 3D structure and the functionality of the peptide and (b) all amino acids share common structural features, including an α-carbon to which an amino group, a carboxyl group, and a variable side chain are

bonded. The macromolecular graph allows reducing the complexity of the structures, since the number of amino acids in a peptide is significantly lower than the number of atoms.

To be able to calculate 3D descriptors, amino acids have to be characterized by (x, y, z) Cartesian coordinates. Being the α-carbon present in all coded amino acids, the Cartesian coordinates of that atom are selected as the coordinates of the whole amino acid. From the peptide topological representation and/or the corresponding geometrical representation (using only α-carbon spatial coordinates), several constitutional, topological and geometrical descriptors can be calculated.

For instance, several protein descriptors both topological and geometric were calculated by weighting amino acids with the → *WHIM descriptors* related to size (Am), shape (Km), and atom distribution density (Dm) of the single amino acids [Mauri, Ballabio *et al.*, 2008]. These amino acid properties were calculated on the isolated 3D structure of amino acids and are collected in Table B7.

An important characteristic of the 3D structure of proteins is the degree of folding of the protein chain. The degree is a quantitative measure of how folded a protein backbone is [Estrada and Uriarte, 2005]. Protein fold to optimize the conformational preferences of amino acids subject to local and global constraints. The → *folding degree index* obtained by diagonalization of the → *distance/distance matrix* [Randić, Kleiner *et al.*, 1994; Randić and Krilov, 1999] is an example of quantitative measure of folding degree, together with → *molecular profiles* [Randić and Krilov, 1997a]. Other size and/or shape descriptors of proteins and, in general, macromolecules, are the → *characteristic ratio*, → *span*, → *Kuhn length*, → *end-to-end distance*, → *persistence length*, and → *radius of gyration*.

Moreover, the **protein folding degree index**, denoted as I_3, is based on the torsion angles of the protein backbone chain (ϕ, ψ, and ω) [Estrada, 2000, 2002a, 2004a; Estrada, Uriarte *et al.*, 2006]. The torsion angle ϕ_i describes the rotation about N_i–$C_{\alpha i}$ peptide bond, ψ_i the rotation about the $C_{\alpha i}$–C_i peptidic bond, and ω_i describes the rotation around the C_i–N_{i+1} bond. A graph is defined whose vertices represent ϕ, ψ, and ω torsion angles and two vertices are connected if, and only if, the corresponding angles are contiguous in the backbone chain of the protein. Then, a matrix **B** is defined to represent the protein backbone as

$$\mathbf{B} = \mathbf{A} + \mathbf{T}$$

where **A** is the → *adjacency matrix* of the graph torsion angles and **T** is a diagonal matrix of the cosine of ϕ, ψ, and ω angles. Finally, the protein folding degree index I_3 is defined as

$$I_3 = \frac{1}{N-3} \cdot \sum_{j=1}^{N-3} e^{\lambda_j}$$

where N is the number of atoms in the protein backbone and λ_j are the eigenvalues of the **B** matrix. Note that this index is strictly related to the → *Estrada index* derived from the adjacency matrix of a molecular graph [Estrada and Hatano, 2007], then it can be expressed as the infinite sum of → *spectral moments* of **B** divided by $k!$ [Estrada, 2004b]. Consequently, the protein folding degree index can be interpreted as the sum of contributions from the sequences of torsion angles of different lengths, in such a way that large sequences of contiguous angles receive lower weights than shorter ones. It was shown that this index well describes the degree of folding of protein chains and that it takes larger values when the folded regions are close to the center of the chain. Moreover, it was demonstrated that local contribution of the ith amino acid

to the global protein folding is expressed as follows [Estrada, 2004b; Estrada and Uriarte, 2005; Estrada and Rodríguez-Velásquez, 2005b]:

$$I_3(i) = \sum_{j=1}^{N} e^{\lambda_j} \cdot \{[\ell_j(\psi_i)]^2 + [\ell_j(\phi_i)]^2\}$$

where $\ell_1, \ell_2, \ldots, \ell_N$ are the eigenvectors of **B** associated to the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_N$; $\ell_j(\psi_i)$ and $\ell_j(\phi_i)$ are the components of the eigenvector ℓ_j corresponding to the torsion angles ψ_i and ϕ_i of the *i*th amino acid (the angle ω_i is not considered because it corresponds to the peptidic bond shared by two contiguous amino acids).

🕮 [Lee and Richards, 1971; Richards, 1977; Connolly, 1983b; Wagner, Colvin *et al.*, 1985; Eisenberg and McLachlan, 1986; Åqvist and Tapia, 1987; Arteca and Mezey, 1990; Wang, Shi *et al.*, 1990; Wold, Jonsson *et al.*, 1993; Leicester, Finney *et al.*, 1994b; Liang and Mislow, 1994; Kuz'min, Trigub *et al.*, 1995; Arteca, 1996; Poirrette, Artymiuk *et al.*, 1997; Andersson, Sjöström *et al.*, 1998; Štambuk, 1999; Lin and Lin, 2001; Liu, Yin *et al.*, 2001a, 2001b; Tusnády and Simon, 2001; Gironés, Amat *et al.*, 2002; Ivanciuc, Schein *et al.*, 2002; Torrens, 2002; Allen, Grant *et al.*, 2003; Ivanciuc, 2003d; Rost, Liu *et al.*, 2003; Ivanciuc, Oezguen *et al.*, 2004; Randić, Zupan *et al.*, 2004; Bai and Wang, 2005; Estrada, 2006b; Pissurlenkar, Malde *et al.*, 2007; Župerl, Pristovšek *et al.*, 2007]

• **DNA sequences**

The genome sequencing research projects are among the most challenging enterprises of these last decades. Elucidation of complete DNA sequences or protein sequences constitutes only a first step, while the further step lies in the interpretation of this huge number of data by automatic procedures.

A DNA sequence is a sequence of four letters A, T, G, and C that, respectively, denote four nucleic acid bases: adenine, thymine, guanine, and cytosine. RNA sequences contain the base uracil U in place of thymine T.

Graphical representations of DNA sequences were proposed by Hamori [Hamori, 1983, 1985, 1989], Gates [Gates, 1985], Nandy [Nandy, 1994, 1996a, 1996b; Nandy and Nandy 1995; Ray, Raychaudhury *et al.*, 1998; Nandy and Basak, 2000], and Leong and Mogenthaler [Leong and Mogenthaler, 1995].

The methods proposed by Gates and Nandy are based on choosing the four cardinal directions in (*x*, *y*) coordinate two-dimensional Cartesian system to represent the four bases in DNA sequences (Figure B3). The method essentially consists of plotting a point corresponding to a base by moving one unit in the positive or negative direction *x*- or *y*-axis depending on the defined association of a base with a cardinal direction. The cumulative plot of such points produces a graph that corresponds to the sequence.

In the Gates axis system (**TCAG-axis system**), one would move one unit in the positive *x*-direction for a cytosine (C), along the positive *y*-direction for a thymine (T), the negative *x*-direction for a guanine (G), the negative *y*-direction for an adenosine (A), implying a cumulative plot of the count of instantaneous C–G against T–A. The Nandy axis system (**CGTA-axis system**) associates G with positive *x*-direction, C with positive *y*-direction, A with negative *x*-direction, and T with negative *y*-direction (Figure B4). In the Leong and Morgenthaler axis system (**TAGC-axis system**), A is associated with positive *x*-direction, T with positive *y*-direction, C with negative *x*-direction, and G with negative *y*-direction.

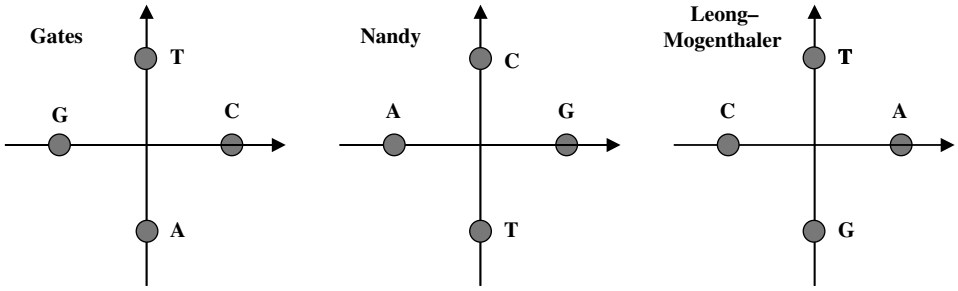


Figure B3 Graphical representation of a DNA sequence by two-dimensional Cartesian systems, as proposed by Gates, Nandy, and Leong–Mogenthaler.

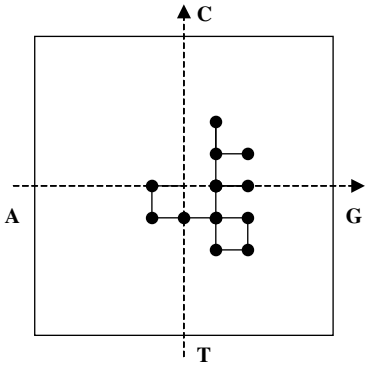


Figure B4 Graphical representations of a DNA sequence by the Nandy coordinate system.

A set of moments relative to the distribution of the graph points around the origin was proposed as the descriptor of the DNA sequence depicted by the Nandy graphical representation. These moments are derived as the weighted average of both *x*- and *y*- coordinates as [Raychaudhury and Nandy, 1999]

$$\mu_x = \frac{\sum_{i=1}^{N} x_i}{N} \qquad \mu_y = \frac{\sum_{i=1}^{N} y_i}{N}$$

where N is the total length of the DNA sequence.

Then, the **graph radius**, denoted as g_R, was proposed as a further sequence descriptor, defined as

$$g_R = \sqrt{\mu_x^2 + \mu_y^2}$$

and the corresponding → *Euclidean distance* between two DNA sequences *s* and *t* was proposed as the measure of sequence dissimilarity:

$$d_{st} = \sqrt{\left(\mu_x^2(s) - \mu_x^2(t)\right)^2 + \left(\mu_y^2(s) - \mu_y^2(t)\right)^2}$$

To reduce the degeneracy of Nandy's graphical representation, an approach based on the idea to deviate from the original cardinal axes directions more than two of the four unit vectors that represent the corresponding bases [Liu, Guo *et al.*, 2002].

Starting from the Nandy's representation of DNA sequences, the eigenvalues obtained from the → *geometric distance/topological distance quotient matrix* and its increasing powers were used to perform similarity/diversity analysis of DNA sequences [Randić, 2000a].

Following the same philosophy of the previous graphical approaches, a representation into the 3D spaces was proposed assigning the four nucleic acid bases, the four directions associated with the regular tetrahedron [Randić, Vračko *et al.*, 2000]. To specify directions, the origin of the Cartesian (x, y, z) coordinate system was assigned in the center of a cube so that the four corners of the cube, which define the tetrahedral directions, constitute the main axes. Then, each basis is moved along the directions as arbitrarily defined in Table B8.

Table B8 The directions of the four nucleic bases in the tetrahedron space as proposed by [Randić, Vračko *et al.*, 2000].

Base	x	y	z
A	+1	−1	−1
G	−1	+1	−1
C	−1	−1	+1
T	+1	+1	+1

The → *leading eigenvalues* obtained from the → *geometric distance/topological distance quotient matrix* and its higher order matrices were proposed to describe the sequence. In any case, the degeneracy of this approach still remains large.

Still trying to remove degeneracy of the DNA sequence representations, another 2D representation was proposed moving the nucleic bases into the 2D plane, with coordinates that depend on an integer parameter d [Guo, Randić *et al.*, 2001] (Table B9). Values $d = 4$ and $d = 8$ were found to generate DNA graphical representations with lower degeneracy.

Table B9 The directions of the four nucleic bases in the tetrahedron space, as proposed by [Guo, Randić *et al.*, 2001].

Base	x	y	Base	x	y
A	−1	$+\frac{1}{d}$	C	$+\frac{1}{d}$	+1
G	+1	$+\frac{1}{d}$	T	$+\frac{1}{d}$	−1

Also in this approach, the → *leading eigenvalues* obtained from the → *geometric distance/topological distance quotient matrix* and its higher order matrices were proposed as the descriptors of DNA sequences.

In another approach, all the possible 64 combinations of three out of four nucleic bases, called triplets, were considered. A cubic matrix $4 \times 4 \times 4$ was constructed whose entries denote the frequencies of occurrence of all the 64 triplets in a DNA sequence [Randić, Guo *et al.*, 2001]. However, in practice, the cubic matrix is not used directly, but three groups of four bidimen-

sional matrices of size 4×4 are derived, each group of which contains all entries of the cubic matrix; thus a total of 12 matrices of size 4×4 are obtained.

From these 12 matrices, the → *leading eigenvalues* were calculated and used for similarity/diversity analysis.

Another 2D representation of the four nucleic bases of DNA sequences was proposed through a scatter plot where the *x*-axis is defined by the actual sequence of nucleic bases and the *y*-axis by the four ordered labels C, G, T, A for nucleic bases [Randić, Vračko *et al.*, 2003] (Figure B5).

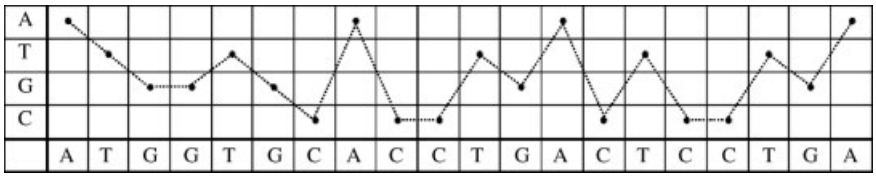


Figure B5 Scatter plot of a DNA sequence as proposed by [Randić, Vračko *et al.*, 2003].

Each point of the scatter plot indicates which basis is present in each position of the sequence. By joining two consecutive points by a line, a zigzag curve is obtained, which is a graphical representation of the sequence.

The numerical characterization of this zigzag curve is performed by the → *Euclidean-distance matrix*, where geometrical distances between every pair of vertices of the zigzag curve are collected, and two other → *graph-theoretical matrices*, called *M/M quotient matrix* and *L/L quotient matrix*. The **M/M quotient matrix** is a symmetric matrix whose off-diagonal elements are given as the ratio of the Euclidean distance between two vertices of the curve over the number of edges between the two vertices, that is, their → *topological distance*; diagonal elements are equal to zero. The M/M quotient matrix is the analogue of the → *geometric distance/topological distance quotient matrix* defined for molecular structures.

The **L/L quotient matrix** is a symmetric matrix whose off-diagonal elements are defined as the ratio of the Euclidean distance between two vertices of the curve over the sum of the geometrical lengths of the edges along the path connecting the two vertices. Note that this matrix is called → *quotient map matrix*, denoted as **Q**, in the framework of proteomics maps [Golbraikh, Bonchev *et al.*, 2001b; Randić, 2001e].

From these Euclidean-distance matrix, *M/M* quotient matrix and *L/L* quotient matrix, → *leading eigenvalues* were calculated to perform similarity/diversity analysis.

The **average distance between pairs of bases** (X, Y), being X, Y = A, C, G, T, is another descriptor of DNA sequences [Randić and Basak, 2001b]. To calculate this descriptor, the 16 pairs of DNA bases are arranged into a square matrix as

AA	AC	AG	AT
CA	CC	CG	CT
GA	GC	GG	GT
TA	TC	TG	TT

For each element of type X in the sequence, the → *topological distance* in the sequence between this element and the next one of type Y is recorded into a matrix, which has size $(N_X \times N_Y)$, where N_X and N_Y are the number of occurrences of the basis type X and Y.

Note that calculating the distance between pairs, their order is not considered, that is, the pair XY is considered the same as the pair YX. For example, given a DNA sequence as

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	T	G	G	T	G	C	A	C	C	T	G	A	C	T	C	C	T	G	A

the pair AA is characterized by the following distance matrix:

A/A	1	8	13	20
1	0	7	12	19
8	7	0	5	12
13	12	5	0	7
20	19	12	7	0

and the pair CA by the following unsymmetrical distance matrix:

A/C	7	9	10	14	16	17
1	6	8	9	13	15	16
8	1	1	2	6	8	9
13	6	4	3	1	3	4
20	13	11	10	6	4	3

From each matrix **XY**, the average distance between bases X and Y is calculated as

$$\overline{\mathrm{XY}} = \frac{\sum_i \sum_j [\mathbf{XY}]_{ij}}{N_\mathrm{X} \cdot N_\mathrm{Y}}$$

where $[\mathbf{XY}]_{ij}$ are the elements of the matrix **XY** and N_X and N_Y are the number of occurrences of the bases of type X and Y, respectively.

Then, for the example given above, the average sum of distances between pairs AA in the 4×4 matrix is $124/(4 \times 4) = 7.75$, while for the AC pair is $162/(6 \times 4) = 6.75$.

Characteristic sequences of DNA were defined in terms of three different classification criteria [He and Wang, 2002]. All the possible combinations of two out four nucleic acid bases were assigned to six different classes:

$R = \{\mathrm{A}, \mathrm{G}\}$ $Y = \{\mathrm{C}, \mathrm{T}\}$ classification based on chemical structure
$M = \{\mathrm{A}, \mathrm{C}\}$ $K = \{\mathrm{G}, \mathrm{T}\}$ classification based on distinguishing amino/keto groups
$W = \{\mathrm{A}, \mathrm{T}\}$ $S = \{\mathrm{G}, \mathrm{C}\}$ classification based on hydrogen bond strength

The first classification criterion consists in assigning each basis in a DNA sequence a value 1 if the basis belongs to the class *R*, that is, is A or G, and zero if the basis belongs to the class *Y*, that is, the basis is C or T. Similar operations are performed considering the other class partitions, thus obtaining three binary vectors of characteristic sequences (*R, Y*), (*M, K*), and (*W, S*), each having length equal to the length of the DNA sequence.

Exploiting this binary representation of sequences, three $2 \times 2 \times 2$ cubic matrices are generated, accounting for the eight possible triplets in each characteristic sequence: 000, 001, 010, 011, 100, 101, 110, and 111. The entries of these matrices are defined as

$$f_{ijk}^{X} = \frac{100 \cdot m_{ijk}^{X}}{N-2}$$

where m_{ijk} is the number of occurrences of the triplet *i–j–k* in *X* and *N* is the length of the string (i.e., the length of the DNA sequence). *X* stands for one of the three classes, (*R, Y*), (*M, K*), or (*W, S*).

The $2 \times 2 \times 2$ cubic matrices are splitted into six 2×2 matrices, considering separately the four entries with triplets beginning with zero (F_0^X) and the four entries with triplets beginning with one (F_1^X), as shown in Figure B6.

F_0^R	0	1
0	10	10
1	9	14

F_1^R	0	1
0	10	13
1	13	20

Figure B6 2×2 matrices for class (*R, Y*), according to the He–Wang approach.

The → *leading eigenvalues* of the six matrices were proposed to describe the whole DNA sequence.

Another approach to the description of DNA sequences is based on the partial-ordering given by the → *Hasse diagrams*. The order relationships between the C, T, A, G bases of a DNA sequence are recorded into the → *Hasse matrix* [Todeschini, Consonni *et al.*, 2006]. The variables used in building the Hasse matrix are the position of the bases in the sequence and a physico-chemical property of the bases, such as the mass. For example, the same small DNA sequence given above:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	T	G	G	T	G	C	A	C	C	T	G	A	C	T	C	C	T	G	A

can be described as shown in Table B10. From these data, the Hasse matrix (20×20) is calculated simply comparing, for each pair of bases in the sequence, the values of the two

Table B10 Data relative to the sequence in the text.

Basis	ID	MW
A	1	135.13
T	2	126.00
G	3	151.13
G	4	151.13
T	5	126.00
...	...	...
...	...	...
T	18	111.10
G	19	135.13
A	20	151.13

ID is the basis position in the sequence and MW the corresponding molecular weight.

variables, that is, position in the sequence (ID) and molecular weight (MW). The obtained corresponding Hasse diagram is shown in Figure B7.

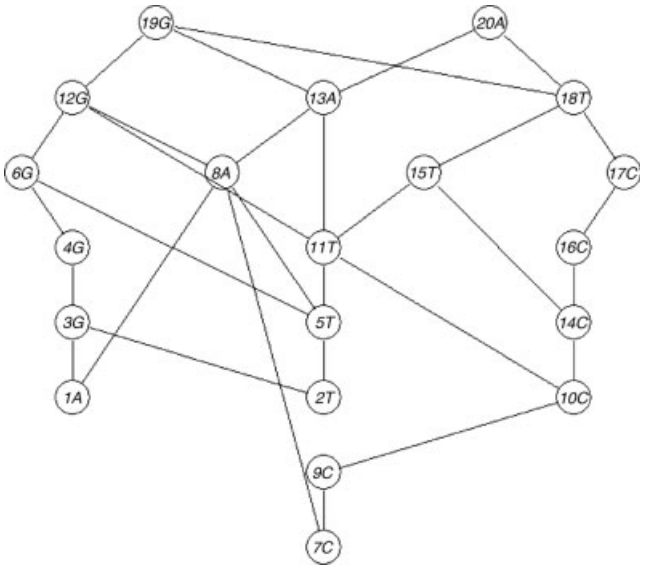


Figure B7 The Hasse diagram obtained by the sequence in the text. For each element, the number corresponds to its absolute position in the sequence.

From each property that ranks the four bases in a different way, a different Hasse diagram is obtained. Descriptors of the DNA sequence are finally obtained from the absolute values of the Hasse matrix elements and the largest eigenvalue was proposed to analyze similarity/diversity of DNA sequences.

🕮 [Le, Nussinov *et al.*, 1989; Shapiro and Zhang, 1990; Wold, Jonsson *et al.*, 1993; Norinder, 1994; Bucher, Karplus *et al.*, 1996; Ray, Raychaudhury *et al.*, 1998; Randić, 2000b; Randić and Vračko, 2000; Štambuk, 2000; Nandy, Nandy *et al.*, 2002; Gan, Pasquali *et al.*, 2003; Guo and Nandy, 2003; Nandy and Nandy, 2003; Randić, Vračko *et al.*, 2003; Randić and Balaban, 2003; Yan, Wang *et al.*, 2003; Yuan, Liao *et al.*, 2003; Dobeš, Kmunicek *et al.*, 2004; Liao and Wang, 2004a, 2004b, 2004c, 2004d, 2005; Liao and Ding, 2005; Liao, Zhang *et al.*, 2005; Liao, Tan *et al.*, 2005a, 2005b; Liao, 2005; Liao, Ding *et al.*, 2005; Liao, Wang *et al.*, 2005; Nandy and Basak, 2005; Randić, Lerš *et al.*, 2005a; Zhang, Liao *et al.*, 2005; Dai, Liu *et al.*, 2006; Gao and Zhang, 2006a, 2006b; Luo, Liao *et al.*, 2006; Nandy, Harle *et al.*, 2006; Randić, Novič *et al.*, 2006; Wang and Wang, 2006; Zhang, Liao *et al.*, 2006; Zhang and Chen, 2006; Liao, Zhu *et al.*, 2007; Nandy, Basak *et al.*, 2007; Zhang, Luo *et al.*, 2007; Zhang, 2007]

- **proteomics maps**

Proteomics maps, together with NMR spectral maps, graphical representation of DNA, and protein sequences, belong to the general class of graphical and visual data represented by a 2D

map [Jeffrey, 1990; Blackstock and Weir, 1999; Bradfield, 2004]. A map is intended as a region of a plane where N discrete points are described by their Cartesian coordinates and relative intensities. **Map invariants** are numerical quantities that characterize the map and are independent of the orientation of coordinate axes and labeling of points [Randić, Lerš *et al.*, 2004b]. Advantages of numerical characterization of 2D maps are the possibility of visual data storage in digital format, significant data compression, quantitative evaluation of → *similarity/diversity* between maps and modeling of relationships between the structure of foreign agents and, for instance, the effects they have on a proteome.

A proteomics map is the result of horizontal separation of proteins by electrophoresis and vertical separation by chromatography, so that proteins on the left of the map have greater charge and proteins at the top have greater mass [Bajzer, Randić *et al.*, 2003]. Proteomics data can be reported as tables of x, y Cartesian coordinates (i.e., charge and mass) of protein spots and their abundance, or formatted into the *bubble diagram* (Figure B8), in which a point, whose coordinates represent charge and mass of a protein, is the center of a circle with radius proportional to the abundance of that protein [Randić, Witzmann *et al.*, 2001].

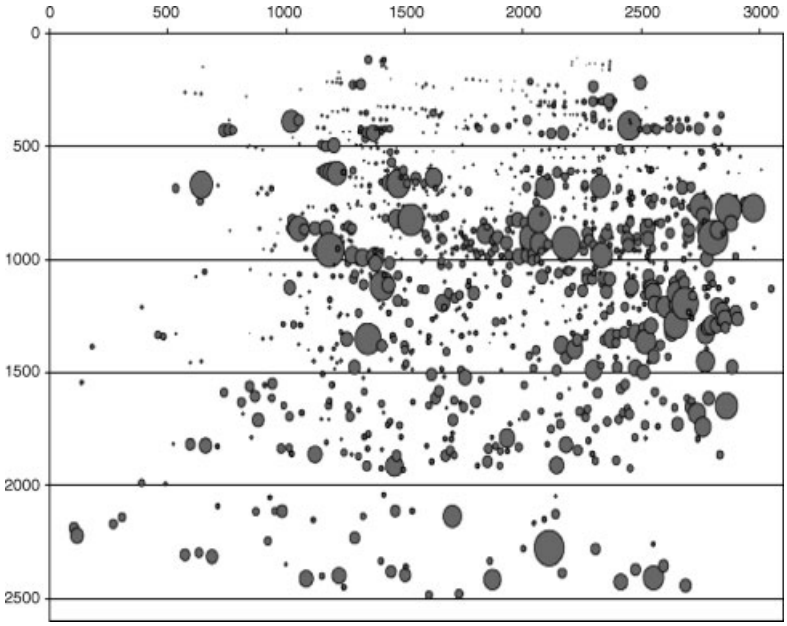


Figure B8 Example of a proteomics map, after a preliminary data pretreatment [Randić, Witzmann *et al*., 2001].

Alternatively, proteins can be represented by points in three-dimensional space where x, y, and z coordinates are proportional to charge, mass, and abundance, respectively. Since proteomics data represent different physical quantities, they are usually scaled, for instance, to the interval (–1, 1) or in such a way as the average charge, the average mass, and the average abundance are all equal to one or to a selected reference value [Bajzer, Randić *et al.*, 2003].

To generate map invariants, the following procedure is used [Randić, 2001e, 2002a; Randić, Witzmann *et al.*, 2001; Randić, Zupan *et al.*, 2001; Randić and Basak, 2002; Randić, Novič *et al.*,

2002]. The first step consists of associating a suitable mathematical object of fixed geometry with a map; then, for the selected mathematical object a numerical representation is constructed in the form of a matrix; once a matrix representing the map has been derived, local invariants and matrix invariants can be calculated in a similar way to → *local vertex invariants* and → *graph invariants* which encode information about a molecular graph.

Examples of mathematical objects used to generate a matrix representation of a map are (1) an *embedded zigzag curve* (or embedded path graph), (2) an *embedded graph of partial ordering*, (3) an *embedded cluster graph*, and (4) an *embedded neighborhood graph*. These will be briefly explained below.

To construct an **embedded zigzag curve**, first points in the map are ordered by assigning them with labels that rank points relative to their abundance giving the most abundant protein point label 1 [Randić, 2001e; Randić, Zupan *et al.*, 2001; Randić, Witzmann *et al.*, 2001; Randić, Novič *et al.*, 2002]. Then, points with adjacent numerical labels are connected by an edge thus resulting into a complicated path, which overlaps itself many times; this path is called zigzag curve. The zigzag curve is then the result of a total ordering of protein spots relative to their abundance (Figure B9).

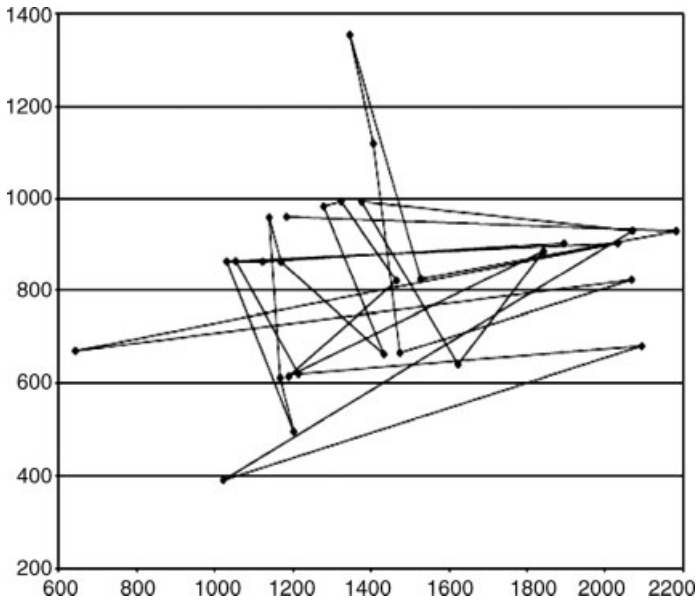


Figure B9 Zigzag curve of a proteomics map [Randić, Witzmann *et al.*, 2001].

The **embedded graph of partial order** is based on the partial order of proteins obtained by ordering them relatively to their charge and mass, respectively [Randić and Basak, 2002; Randić, 2002a; Randić, Zupan *et al.*, 2002]. In this graph, only those protein spots are connected that either dominate or are dominated in both the mass and the charge by the neighboring spots. If a direction from left to right is associated with each connection line, then a directed graph is obtained, which leads to an → *adjacency matrix* with positive and negative values depending on the direction of the edge connecting two vertices.

In the embedded graph of partial order the vertices, representing protein spots, are at fixed geometrical location and all the edges have positive slopes; this is a consequence of the partial order in which vertices at the top and right location dominate vertices which are at lower height and shifted towards the left.

The **embedded cluster graph** is obtained by making connections between the protein spots that are separated by Euclidean geometrical distances shorter than or equal to a selected critical distance [Randić and Basak, 2002; Bajzer, Randić *et al.*, 2003].

The **embedded neighborhood graph** is constructed by using the following procedure [Randić, Lerš *et al.*, 2004a, 2004b; Randić, Novič *et al.*, 2005]. First, x and y coordinates are scaled dividing them by the maximal Euclidean distance between two spots in the map. Relative abundances are calculated by dividing each protein abundance by the abundance of the protein corresponding to the spot with label 1, which is the maximally abundant protein. Then, the clustering method KNN is applied as follows: Euclidean distances between pairs of protein spots are calculated, for each spot a short list of the nearest neighbors is constructed, and, finally, the considered spot is connected by lines with its nearest neighbors. Different graphs are obtained by varying the number of nearest neighbors. The nearest neighbors can be 2D if Euclidean distance between spots are calculated in the space defined by the coordinates (x, y) or 3D if distances are calculated using (x, y, z) coordinates, where z refers to protein abundance. In any case, when dealing with 2D neighborhood graphs, information about relative abundances of spots can be accounted for by assigning each spot a two-component vector containing a local invariant derived from a map matrix **M** (e.g., the matrix row sum) and relative abundance z. Then, the length of this vector is

$$|\mathbf{v}_i| = \sqrt{\left(\sum_{j=1}^{N} [\mathbf{M}]_{ij}\right)^2 + z_i^2}$$

where **M** is any matrix describing relationships among protein spots and N is the number of spots. The vector length can be used as the descriptor of each spot and the average length of the vectors of all the spots as a map descriptor.

From the selected graph representation of a proteomics map, different *map matrices* can be derived which encode information about distances and adjacency between protein spots. Examples of these matrices are reported below.

The **Euclidean-distance map matrix**, denoted as **ED**, is the analogue of the → *geometry matrix* **G** derived from a molecular graph. In this case, vertices of the map graph are assigned (x, y) or (x, y, z) coordinates, z being intended as the → *weighting scheme* for vertices; it is defined as [Bajzer, Randić *et al.*, 2003]

$$[\mathbf{ED}]_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$

where x, y, and z are the spatial coordinates of a protein spot (x and y) in the map and its abundance (z), respectively. The Euclidean-distance map matrix can also be calculated by considering only x and y coordinates. To describe the embedded zigzag curve, Euclidean distances through space were measured directly (e.g., in millimeters) from the map for all the pairs of vertices [Randić, 2001e].

The **path-distance map matrix**, denoted as **PD**, resembling the → *bond length-weighted distance matrix* of a molecular graph, is defined as [Bajzer, Randić *et al.*, 2003]

$$[\mathbf{PD}]_{ij} = \min_{p_{ij}} \left(\sum\nolimits_{kq} [\mathbf{ED}]_{kq} \right)_{ij}$$

where $[\mathbf{ED}]_{kq}$ denotes entries of the Euclidian-distance map matrix, p_{ij} a path connecting vertices i and j, and the summation goes over all the pairs of adjacent vertices along the considered path. Then, each entry of the path-distance map matrix is the shortest distance between two vertices measured along the path by summing the geometrical length of the edges connecting adjacent vertices along the path.

The **quotient map matrix**, denoted as **Q**, is defined in a similar way to the → *geometric distance/topological distance quotient matrix*, whose entries are defined in terms of the ratio of distances between a pair of vertices measured through the space and along the bonds. The elements of the quotient matrix **Q** are formally defined as [Randić, 2001e; Bajzer, Randić *et al.*, 2003]

$$[\mathbf{Q}]_{ij} = \begin{cases} \dfrac{[\mathbf{ED}]_{ij}}{[\mathbf{PD}]_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where $[\mathbf{ED}]_{ij}$ and $[\mathbf{PD}]_{ij}$ are the elements of the Euclidean-distance and path-distance map matrix, respectively. The elements of the quotient matrix are equal to 1 for all the pairs of adjacent protein spots and smaller than 1 for pairs of nonadjacent spots.

The **neighborhood-distance map matrix**, denoted as **ND**, encodes information about vertex proximities; its elements are different from zero only for those pairs of protein spots that are within a certain neighborhood. This matrix is the analogue of the → *neighborhood geometry matrix* derived from a molecular graph; then it is defined as [Bajzer, Randić *et al.*, 2003]

$$[\mathbf{ND}]_{ij} = \begin{cases} [\mathbf{ED}]_{ij} & \text{if } [\mathbf{ED}]_{ij} \leq D_C \\ 0 & \text{if } [\mathbf{ED}]_{ij} > D_C \text{ or } i,j \text{ are not connected} \end{cases}$$

where $[\mathbf{ED}]_{ij}$ are the elements of the Euclidean-distance map matrix and D_C is a critical distance.

The **Euclidean-adjacency map matrix**, denoted as **EA**, is the analogue of the → *bond length-weighted adjacency matrix* defined for molecular graphs. It is defined by replacing elements equal to one, corresponding to pairs of adjacent spots in the → *adjacency matrix* with the corresponding elements in the Euclidean-distance map matrix **ED** as [Randić and Basak, 2002; Randić, Lerš *et al.*, 2004b]

$$[\mathbf{EA}]_{ij} = \begin{cases} [\mathbf{ED}]_{ij} & \text{if } i,j \text{ are connected} \\ 0 & \text{if } i = j \text{ or } i,j \text{ are not connected} \end{cases}$$

The **map connectivity matrices** are another set of map matrices based on partitioning of the → *Randić connectivity index* and the higher order → *connectivity indices* into contributions arising from paths of length k [Randić and Basak, 2002]. They are defined as

$$\begin{aligned} \left[\mathbf{D}_{1\chi}\right]_{ij} &= (\delta_i \cdot \delta_j)^{-1/2} \cdot \delta(d_{ij}, 1) \\ \left[\mathbf{D}_{2\chi}\right]_{ij} &= (\delta_i \cdot \delta_l \cdot \delta_j)^{-1/2} \cdot \delta(d_{ij}, 2) \\ &\cdots\cdots\cdots \\ \left[\mathbf{D}_{k\chi}\right]_{ij} &= (\delta_i \cdot \ldots \cdot \delta_j)^{-1/2} \cdot \delta(d_{ij}, k) \end{aligned}$$

where δ_i indicates the → *vertex degrees* and $\delta(d_{ij}, k)$ indicates the Kronecker delta function that is equal to one for pairs of vertices v_i and v_j at a topological distance of k, and zero otherwise. The term $(\delta_i \cdot \ldots \cdot \delta_j)$ indicates the product of the degrees of the vertices along the path connecting the vertices v_i and v_j.

Higher order map matrices are → *power matrices* derived by a map matrix **M** by either using the standard matrix multiplication of linear algebra or by using the → *Hadamard matrix product*, which leads to matrices whose elements are defined as $[\mathbf{M}]_{ij}^{k}$, where k is an integer exponent.

Map invariants usually calculated from map matrices are the → *leading eigenvalue* λ_1 of a map matrix and the normalized leading eigenvalues of its higher order matrices (e.g., $\lambda_1/k!$, k being the matrix order) [Bajzer, Randić *et al.*, 2003]. Other map invariants are the average row sum of a map matrix and the average of those row sums corresponding to protein spots lying in a selected region of the proteomics map [Randić, Lerš *et al.*, 2004b]. Moreover, → *Wiener-type indices* of map matrices were also investigated.

📖 [Marengo, Leardi *et al.*, 2003; Randić and Basak, 2004; Vračko and Basak, 2004; Randić, Lerš *et al.*, 2005b; Marengo, Robotti *et al.*, 2006; Randić, Witzmann *et al.*, 2006; Marengo, Robotti *et al.*, 2008]

➢ **bioisosterism** → drug design

■ biological activity indices

These are molecular properties related to the effect of a substance produces on an organism or any biological target. Biological activity depends on peculiarities of compounds (molecular structure and → *physico-chemical properties*), biological entity (species, gender, age, etc.), and mode of treatment (dose, route, exposure, etc.).

The *dose* is the amount of a substance that is administrated to an organism (animal, human) trough food or other administration routes (topic, gavage, injection, etc.).

The *dose–response curve* is a sigmoidal curve (Figure B10) that highlights the relation between the amount of a drug or chemical administered to an organism and the degree of response it produces. This response is measured by the percentage of the exposed population that shows the defined effect.

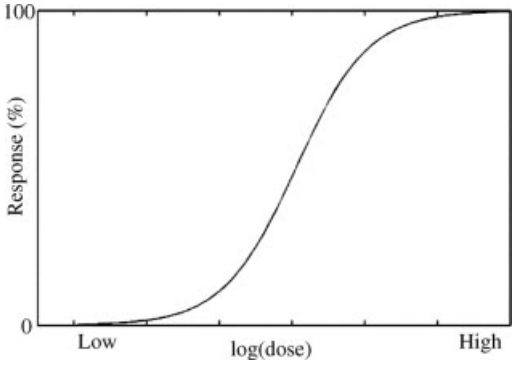


Figure B10 Dose–response curve.

Dose–response experiments typically use 10–20 doses, approximately spaced on a logarithmic scale. For example,

Dose (nM)	1	3	10	30	100	300	1000	3000	10000
Dose (log)	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4

- **pharmacological indices**

In pharmacology, the **Effective Dose** (ED) is the minimal dose that produces the desired effect of a drug. The effective dose is often determined based on analyzing the dose–response relationship specific to the drug. The dosage that produces a desired effect in half the test population is referred to as the **median effective dose** ED_{50}, that is, the amount of drug that produces a therapeutic response in 50% of the people taking it.

The **therapeutic index** (or **therapeutic ratio**) is a comparison of the amount of a therapeutic agent that causes the therapeutic effect to the amount that causes toxic effects. Quantitatively, it is the ratio of the dose required to produce the toxic effect over the therapeutic dose. A commonly used measure of therapeutic index is the lethal dose of a drug for 50% of the population (LD_{50}) divided by the effective dose for 50% of the population (ED_{50}):

$$\textit{therapeutic index} = \frac{LD_{50}}{ED_{50}}$$

The therapeutic index of a drug indicates the selectivity of the drug and consequently its usability. It should be noted that a single drug can have many therapeutic indices; while some are for each of its undesirable effects relative to a desired drug action, the others for each of its desired effects if the drug has more than one action.

- **toxicological indices**

Toxicity is a relative property of a chemical that refers to its potential to have a harmful effect on a living organism. It is experimentally determined trough toxicity tests in which organisms are exposed trough food (*oral toxicity*) or are exposed at a concentration of the chemical in a given environmental compartment, such as water, air, or soil (*environmental toxicity*)

In acute oral tests, organisms are subject to a single dose of the chemical and the toxicity is not a function of exposure time. Several → *structural alerts* were proposed for identify toxicological effects, particularly for carcinogenicity and mutagenicity of the chemicals.

The **lethal dose** (LD) is the dose of a chemical or biological preparation that is likely to cause death, giving an indication of the lethality of a given substance. Because resistance varies from one individual to another, the "lethal dose" represents a dose (usually recorded as weight of the dose per kilogram of subject body weight, e.g., mg/kg b.w.) at which a given percentage of subjects will die. The most commonly used lethality indicator is the **median lethal dose** LD_{50}, a dose at which 50% of subjects will die.

In long-term oral toxicity tests, the organism is fed for several days (in some cases months or years) with food contaminated by the tested chemical. In this case, results are expressed as concentration in food and are a function of exposure time. The total dose may also be calculated from the concentration and the total amount of food ingested.

In environmental toxicity tests, the **Lethal Concentration** (LC) is a measure (as weight/weight or weight/volume) of the concentration of the toxic chemical producing death in a given percentage of organisms. The concentration at which 50% of subjects will die is denoted as LC_{50} and is called **median lethal concentration**. It is always a function of exposure time (for example, 96 h LC_{50} in short-term toxicity tests on fish).

The **Lethal Time** (LT_{50}) is the time needed for 50% of the subjects to die after the exposure at a determined concentration of a substance.

Median Inhibitory Concentration (IC_{50}) is a measure of the concentration required for producing 50% inhibition of a biological activity (i.e., an enzyme reaction, cell growth, reproduction, etc.). In simpler terms, it measures how much of a particular substance/molecule is needed to inhibit some biological process by 50%. IC_{50} is commonly used as a measure of drug-receptor binding affinity.

No-Observed-Effect Level (NOEL) is the greatest concentration or amount of a substance, found by experiment or observation, that causes no alterations of morphology, functional capacity, growth, development, or life span of target organisms distinguishable from those observed in normal (control) organisms of the same species under the same defined conditions of exposure.

The **No-Observed-Adverse-Effect Level** (NOAEL) is used for those chemicals that at low levels may be beneficial or necessary (e.g., natural micronutrients, such as some heavy metals).

The **Lowest-Observed-Effect-Level** (LOEL) is the lowest level to which a studied effect is observed.

The **Acceptable Daily Intake** (ADI) is the daily intake of a chemical that, during an entire lifetime, appears to be without appreciable risk. It is expressed as in milligrams of the chemical per kilogram of body weight (mg/kg b.w.). It is usually estimated as

$$\mathrm{ADI} = \frac{\mathrm{NOAEL}}{\mathrm{SF}}$$

where NOAEL is the No-Observed-Adverse-Effect Level and SF is a safety factor. Safety factors are a function of the level of uncertainty of the NOAEL and the toxicological mode of action and may range from 10 to 10000.

The **Iball index** is defined as the percentage of skin cancer or papilloma-developing mice (skin painting experiments) divided by the average latent period in days for the affected animals multiplied by 100 [Daudel and Daudel, 1966; Herndon and Szentpály, 1986; Barone, Camilo Jr. *et al.*, 1996; Braga, Barone *et al.*, 1999; Barone, Braga *et al.*, 2000].

📖 [Wang and Milne, 1993; Benigni, 2003; Öberg, 2004a]

- **biological activity profile score** → scoring functions
- **Bird aromaticity indices** → delocalization degree indices
- **BLOGP** → lipophilicity descriptors

Blurock spectral descriptors

They are atomic or bond descriptors derived from a spectral representation of molecules, like the following: a property is associated with each atom (or bond) in such a way as to also represent the atom and its environment, the range of the property values in the whole data set of molecules

is divided into equal-sized intervals and the number of times the values of the molecule atoms fall within each interval is counted. The spectrum of the molecule is the distribution of these property values [Blurock, 1998].

The atomic properties considered are partial charges, electron densities, and polarizabilities, calculated by → *computational chemistry* methods; moreover, bond properties have been proposed as the difference between the property values of the atoms forming the bond. The range of each property is determined by the maximum and minimum values for all the atoms in all the molecules, thus obtaining uniform spectrum length for all the molecules in the data set.

Inductive learning was suggested for the prediction of the molecular property values.

- **Bocek–Kopecky analysis** → Free–Wilson analysis
- **Bocek–Kopecky model** → Free–Wilson analysis
- **Bodor hydrophobic model** ≡ *BLOGP* → lipophilicity descriptors
- **Bodor LOG*P*** ≡ *BLOGP* → lipophilicity descriptors
- **boiling point** → physico-chemical properties
- **Bonchev centric information indices** → centric indices
- **Bonchev complexity information index** → molecular complexity
- **Bonchev topological complexity indices** → molecular complexity
- **bond alternation coefficient** → delocalization degree indices
- **bond angles** → molecular geometry
- **bond connectivity index** ≡ *edge connectivity index* → edge adjacency matrix
- **bond connectivity indices** ≡ *extended edge connectivity indices* → edge adjacency matrix
- **bond count** ≡ *bond number*
- **bond dipole moment** → bond ionicity indices
- **bond distances** → molecular geometry
- **bond distance-weighted edge adjacency matrix** → edge adjacency matrix
- **bond eccentricity** → edge distance matrix
- **bonded pair descriptors** → substructure descriptors
- **bond *E*-state index** → electrotopological state indices
- **bond flexibility** → flexibility indices
- **bond flexibility index** → flexibility indices
- **bond index** → quantum-chemical descriptors
- **bonding information content** → indices of neighborhood symmetry
- **bonding orbital information index** → information theoretic topological index

bond ionicity indices

Such indices encode information about the bond character, being the importance of bond character to the physical and chemical behavior of compounds well known. Bond character is closely related to the capacity of bonded atoms to exchange electrons and such capacity is commonly well represented by the → *electronegativity* χ of the bonded atoms.

The difference in electronegativity between two bonded atoms was called **bond dipole moment** [Malone, 1933], defined as

$$\mu_{ij} = \left|\chi_i - \chi_j\right|$$

but there was poor correlation between this index and bond ionicity. Therefore, starting from bond dipole moment, several empirical relationships have been proposed to define bond ionicity indices f_b. The most popular are [Barbe, 1983]:

1. $f_b = 1 - \exp\left[-0.25 \cdot (\chi_i - \chi_j)^2\right]$
2. $f_b = 1 - \exp\left[-0.21 \cdot (\chi_i - \chi_j)^2\right]$
3. $f_b = 0.160 \cdot (\chi_i - \chi_j) + 0.035 \cdot (\chi_i - \chi_j)^2$
4. $f_b = \dfrac{\chi_i - \chi_j}{\chi_i + \chi_j}$
5. $f_b = \dfrac{\chi_i - \chi_j}{2}$
6. $f_b = \dfrac{\chi_i - \chi_j}{\chi_i}$ with $\chi_i > \chi_j$

- **Bondi volume** → volume descriptors (⊙ van der Waals volume)
- **bond length-corrected connectivity index** → connectivity indices (⊙ Kupchik modified connectivity indices)
- **bond length-weighted adjacency matrix** → molecular geometry
- **bond length-weighted distance matrix** → weighted matrices (⊙ weighted distance matrices)
- **bond length-weighted Wiener index** → weighted matrices (⊙ weighted distance matrices)
- **bond matrix** ≡ *edge adjacency matrix*

bond number (B) (≡ *edge counting; bond count*)

This is the simplest graph invariant defined as the number of edges in the simple → *molecular graph* G where multiple bonds are considered as single edges. The bond number is calculated from the → *adjacency matrix* **A** as half the → *total adjacency index* A_V:

$$B = \frac{1}{2} \cdot A_V = \frac{1}{2} \cdot \sum_{i=1}^{A} \sum_{j=1}^{A} a_{ij}$$

where a_{ij} are the elements of the adjacency matrix and A is the total number of graph vertices.

The bond number is related to molecular size and gives equal weight to chemically nonequivalent groups, such as CH_2-CH_2, $CH_2=CH$, CH_2-NH_2, and CH_2-Cl.

When bond multiplicity in the molecule must be considered several → *multiple bond descriptors* can be used instead of the bond number.

The number of bonds is considered in the → *cyclomatic number* and appears in several → *molecular descriptors* such as the → *Balaban distance connectivity index*, the → *mean Randić branching index*, the → *information bond index*, and several → *topological information indices*.

- **bond order** → quantum-chemical descriptors
- **bond order–bond length relationships** → bond order indices

bond order indices

These are descriptors for molecule bonds proposed with the aim of estimating the → *bond order* defined in quantum-chemical theory or of generally defining bond weights so as to distinguish the bonds in a → *molecular graph*.

The most common definitions of bond order indices are reported below. Moreover, the term **fractional bond order** was suggested to refer to the inverse of any bond order index. Fractional bond order permits individual treatment of σ and π molecular systems; σ bonds give simple graphs, while π bonds introduce a weighted molecular framework with weights smaller than one [Randić, Brissey *et al.*, 1980].

• **conventional bond order** (π^*)

Within the framework of the graph theory, the conventional bond order π^* is defined as being equal to 1, 2, 3, and 1.5, for single, double, triple, and aromatic bonds, respectively. The → *bond vertex degree* of an atom is an important local invariant defined as the sum of the conventional bond orders of the edges incident to a vertex.

To consider chemical information relative to multiple bonds in terms of topological bond lengths, the inverse powers of the conventional bond order were proposed [Balaban, 1993c; Balaban, Bonchev *et al.*, 1993]. The **relative topological distance** is defined as

$$RTD_{ij} = \left(\pi^*_{ij}\right)^{-1}$$

where i and j refer to adjacent vertices in the graph. To obtain values more related to the standardized experimental interatomic average distances (as reference is taken the distance of single bonds, Table B11), the **chemical distance** was defined as

$$CD_{ij} = \left(\pi^*_{ij}\right)^{-1/4}$$

Using relative topological distance or chemical distance as well as conventional bond order to weight each edge in the graph several → *weighted matrices* were proposed which account for information about bond multiplicity.

Table B11 Average experimental bond length *r*, carbon relative value r^*, conventional bond order π^*, relative topological distance *RTD*, and chemical distance *CD*.

Bond type	*r* (Å)	r^*	π^*	*RTD*	*CD*
C−C	1.54	1.00	1	1.00	1.00
C≈C	1.40	0.91	1.5	0.67	0.90
C=C	1.33	0.86	2	0.50	0.84
C≡C	1.20	0.78	3	0.33	0.76
C−N	1.47	1.00	1	1.00	1.00
C=N	1.29	0.88	2	0.50	0.84
C≡N	1.16	0.79	3	0.33	0.76
N−N	1.45	1.00	1	1.00	1.00
N=N	1.26	0.87	2	0.50	0.90
C−O	1.41	1.00	1	1.00	1.00
C=O	1.21	0.86	2	0.50	0.84
O−O	1.45	1.00	1	1.00	1.00
N−O	1.47	1.00	1	1.00	1.00
N=O	1.15	0.78	2	0.50	0.84
C−S	1.81	1.00	1	1.00	1.00
C=S	1.61	0.89	2	0.50	0.84

The symbol ≈ stands for aromatic bonds.

From the conventional bond order, the **atomic multigraph factor** (or **multigraph factor**) is a → *local vertex invariant*, denoted as f_i, and defined as [Balaban and Diudea, 1993]

$$f_i = \sum_{j=1}^{A} a_{ij} \cdot (\pi_{ij}^{*} - 1) \qquad \pi_{ij}^{*} = 0 \quad \text{if } (i,j) \notin \mathcal{E}(\mathcal{G})$$

where the summation goes over all graph vertices, but the only nonvanishing elements are those corresponding to pairs of adjacent vertices (a_{ij} are the elements of the adjacency matrix); π_{ij}^{*} is the conventional bond order associated to the edge connecting vertices v_i and v_j for pairs of adjacent vertices, and zero otherwise. The multigraph factor is zero for atoms without multiple bonds and it is used, for instance, to derive local invariants form → *layer matrices* and the → *Balaban DJ index*. Moreover, the atomic multigraph factor is closely related to the → *bond vertex degree* δ_i^b as

$$\delta_i^b = \sum_{j=1}^{A} a_{ij} \cdot \pi_{ij}^{*} = \delta_i + f_i \qquad \pi_{ij}^{*} = 0 \quad \text{if } (i,j) \notin \mathcal{E}(\mathcal{G})$$

where δ is the simple → *vertex degree*, that is, the number of adjacent vertices.

• **graphical bond order**

The graphical bond order of the bth bond, denoted as $(TI'/TI)_b$, is derived from the → *H-depleted molecular graph* of the molecule by calculating a → *graph invariant* TI' for the subgraph $\mathcal{G}'$ obtained by erasing an edge b from the graph and then dividing it by the corresponding graph invariant TI calculated on the whole molecular graph $\mathcal{G}$ [Randić, Mihalić *et al.*, 1994]. If more than one subgraph is obtained by the erasure of each edge, the single contributions are summed up to give the graphical bond order or, alternatively, they can be multiplied [Mekenyan, Bonchev *et al.*, 1988a]. The ratio $(TI'/TI)_b$ was interpreted as a measure of the relative importance of the edge in the graph. The first proposed graphical bond order was that calculated from the → *Hosoya Z index* and was originally called **topological bond order**; it was shown to represent the weight of a bond in distributing π-electrons over the molecular graph [Hosoya, Hosoi *et al.*, 1975; Hosoya and Murakami, 1975]. Moreover, graphical bond orders can be considered special cases of → *normalized fragment topological indices*.

Molecular descriptors are derived by the additive contributions of the graphical bond orders of all bonds in the molecule as

$$TI'/TI = \sum_{b=1}^{B} \left(\frac{TI'}{TI} \right)_b$$

They are usually called **graphical bond order descriptors**.

The graphical bond order calculated using the → *total path count* P as molecular invariant was called **path graphical bond order** and denoted by the ratio $(P'/P)_{ij}$, where i and j refer to the vertices incident to the bth edge erased from the graph [Randić, 1991b; Randić and Trinajstić, 1993a; Plavšić, Šoškić *et al.*, 1996b].

From the path graphical bond orders, a square symmetric → *weighted adjacency matrix* of dimension $A \times A$, called **P-matrix** (or **path matrix**) and denoted as **P**, was derived [Plavšić and Graovac, 2001]; its elements are defined by the following:

$$[\mathbf{P}]_{ij} = \begin{cases} (P'/P)_{ij} & (i,j) \in \mathcal{E}(\mathcal{G}) \\ 0 & \text{otherwise} \end{cases}$$

where $\mathcal{E}(\mathcal{G})$ is the set of graph edges.

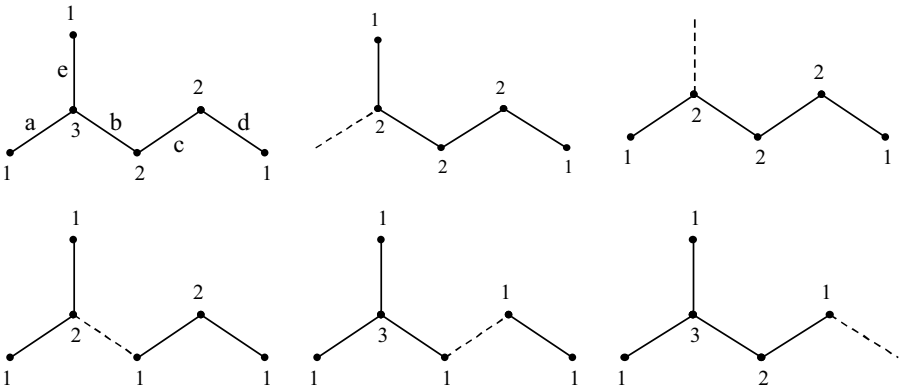
From the P-matrix, a → *Wiener-type index*, called ***P′/P* index**, is calculated applying the → *Wiener operator Wi* as

$$\frac{P'}{P} \equiv Wi(\mathbf{P}) = \frac{1}{2} \cdot \sum_{i=1}^{A} \sum_{j=1}^{A} [\mathbf{P}]_{ij}$$

Other encountered graphical bond order descriptors are **χ′/χ index**, ***W′/W* index**, ***WW′/WW* index**, ***J′/J* index**, ***CID′/CID* index**, and → *Z′/Z index* derived, respectively, from → *Randić connectivity index*, → *Wiener index*, → *hyper-Wiener index*, → *Balaban distance connectivity index*, → *Randić connectivity ID number*, and → *Hosoya Z index*.

Example B2

χ'/χ graphical bond order index for 2-methylpentane.



$${}^1\chi(G) = 2(1\cdot 3)^{-1/2} + (3\cdot 2)^{-1/2} + (2\cdot 2)^{-1/2} + (2\cdot 1)^{-1/2} = 2.7701$$

$${}^1\chi(G{-}\mathrm{a}) = {}^1\chi(G{-}\mathrm{e}) = 2\cdot(2\cdot 1)^{-1/2} + 2\cdot(2\cdot 2)^{-1/2} = 2.4142$$

$$\left(\frac{\chi'}{\chi}\right)_{\mathrm{a}} = \left(\frac{\chi'}{\chi}\right)_{\mathrm{e}} = \frac{{}^1\chi(G{-}\mathrm{a})}{{}^1\chi(G)} = \frac{2.4142}{2.7701} = 0.8715$$

$${}^1\chi(G{-}\mathrm{b}) = 4\cdot(2\cdot 1)^{-1/2} = 2.8284$$

$$\left(\frac{\chi'}{\chi}\right)_{\mathrm{b}} = \frac{{}^1\chi(G{-}\mathrm{b})}{{}^1\chi(G)} = \frac{2.8284}{2.7701} = 1.0210$$

$${}^1\chi(G{-}\mathrm{c}) = (1\cdot 1)^{-1/2} + 3\cdot(1\cdot 3)^{-1/2} = 2.7321$$

$$\left(\frac{\chi'}{\chi}\right)_{\mathrm{c}} = \frac{{}^1\chi(G{-}\mathrm{c})}{{}^1\chi(G)} = \frac{2.7321}{2.7701} = 0.9863$$

$$^1\chi(G{-}\mathrm{d}) = 2\cdot(1\cdot 3)^{-1/2} + (1\cdot 2)^{-1/2} + (2\cdot 3)^{-1/2} = 2.2701$$

$$\left(\frac{\chi'}{\chi}\right)_{\mathrm{d}} = \frac{^1\chi(G{-}\mathrm{d})}{^1\chi(G)} = \frac{2.2701}{2.7701} = 0.8195$$

$$\frac{\chi'}{\chi} = \left(\frac{\chi'}{\chi}\right)_{\mathrm{a}} + \left(\frac{\chi'}{\chi}\right)_{\mathrm{e}} + \left(\frac{\chi'}{\chi}\right)_{\mathrm{b}} + \left(\frac{\chi'}{\chi}\right)_{\mathrm{c}} + \left(\frac{\chi'}{\chi}\right)_{\mathrm{d}} = 2\cdot 0.8715 + 1.0210 + 0.9863 + 0.8195 = 4.5698$$

An explicit formula for the direct calculation of the χ'/χ index from the molecular graph was derived as

$$\frac{\chi'}{\chi} = \frac{1}{\chi(G)}\cdot\sum_{i=1}^{A-1}\sum_{j=i+1}^{A} a_{ij}\cdot\frac{(A+C-\delta_i-\delta_j)\cdot(\delta_i\cdot\delta_j)^{1/2} + \delta_i\cdot\left[(\delta_i-1)\cdot\delta_j\right]^{1/2} + \delta_j\cdot\left[(\delta_j-1)\cdot\delta_i\right]^{1/2}}{\delta_i\cdot\delta_j}$$

where $\chi(G)$ is the → *Randić connectivity index* for the whole molecular graph, the summation goes over all pairs of vertices, but the only nonvanishing terms are those corresponding to pairs of adjacent vertices, a_{ij} being the elements of the adjacency matrix; δ_i and δ_j are the → *vertex degrees* of the vertices v_i and v_j, A the number of graph vertices, and C the → *cyclomatic number*. This formula holds for every connected graph with $A > 1$ vertices [Plavšić, Šoškić *et al.*, 1998].

In the same way, a general formula valid for any graph based on the number A of graph vertices and the distances between pairs of vertices was derived for the calculation of the WW'/WW index as [Plavšić, 1999]

$$\frac{WW'}{WW} = A + 1 - \frac{\sum_{i=1}^{A-1}\sum_{j=i+1}^{A} d_{ij}\cdot(d_{ij}+1)\cdot(d_{ij}+2)}{\sum_{i=1}^{A-1}\sum_{j=i+1}^{A} d_{ij}\cdot(d_{ij}+1)},$$

where d_{ij} is the topological distance between vertices v_i and v_j.

Only for acyclic graphs, after special rearrangement, does the formula take the form

$$\frac{WW'}{WW} = A + 1 - \frac{\sum_{m=1}^{D}\frac{(m+2)!}{(m-1)!}\cdot{}^mP}{\sum_{m=1}^{D}\frac{(m+1)!}{(m-1)!}\cdot{}^mP}$$

where m is the length of the considered paths, ${}^mP$ the → *path count* of mth order, and D the → *topological diameter*, that is, the maximum topological distance in the graph.

- **bond order–bond length relationships**

Several bond order–bond length relationships were proposed in literature [Paolini, 1990; Alkorta, Rozas *et al.*, 1998]. The most known relationships are collected in Table B12.

Table B12 Relationships between bond length and bond order.

Equation	$r = f(\pi)$	$\pi = f(r)$	Reference
Pauling (1947)	$r_{ij} = \hat{r}_{ij} - 0.71 \cdot \log(\pi_{ij})$	$\pi_{ij} = \exp\left[-\frac{(r_{ij} - \hat{r}_{ij})}{0.71}\right]$	[Pauling, 1947]
Pauling (1986)	$r_{ij} = \hat{r}_{ij} - 0.700 \cdot \log\{\pi_{ij} \cdot [1 + 0.064 \cdot (\nu - 1)]\}$	$\pi_{ij} = \frac{\exp\left[-\frac{(r_{ij} - \hat{r}_{ij})}{0.700}\right]}{1 + 0.064(\nu - 1)}$	[Pauling and Kamb, 1986]
Paolini (1990)	$r_{ij} = \hat{r}_{ij} - 0.78 \cdot (\pi_{ij}^{1/3} - 1)$	$\pi_{ij} = \left[1 - \frac{(r_{ij} - \hat{r}_{ij})}{0.78}\right]^3$	[Paolini, 1990]
Gordy (1947)	$r_{ij} = \sqrt{a/(b + \pi_{ij})}$	$\pi_{ij} = a \cdot r_{ij}^{-2} - b$	[Gordy, 1947]
Lendvay (2000)	$r_{ij} = \hat{r}_{ij} - 0.25 \cdot \ln\pi_{ij}$	$\pi_{ij} = \exp\left[-\frac{(r_{ij} - \hat{r}_{ij})}{0.25}\right]$	[Lendvay, 2000]

$\hat{r}$ is the equilibrium bond length of a single bond. The parameters a and b of the Gordy equation are given in Table B13. For Pauling (1947) formula, some equilibrium distances $\hat{r}$ are C–C = 1.542 Å, C=C = 1.330 Å, and C≡C = 1.204 Å; for Lendvay formula, C–C = 1.54 Å, C=O = 1.43 Å, and C–H = 1.08 Å.

Gordy's bond order is used in the calculation of the → *Bird aromaticity indices* and the empirical constants a and b are given in Table B13 [Gordy, 1947; Krygowski and Cyranski, 2001].

Table B13 Values of a and b constants used in the calculation of Gordy's bond order.

Bond	a	b	Bond	a	b
C≈C	6.80	1.71	N≈N	5.28	1.41
C≈N	6.48	2.00	N≈O	4.98	1.45
C≈O	5.75	1.85	N≈S	10.53	2.50
C≈S	11.9	2.59	O≈O	4.73	1.22
C≈P	13.54	3.02	O≈S	17.05	5.58
B≈B	9.12	1.94	S≈S	19.30	3.46
B≈C	8.05	2.11	C≈Se	15.24	3.09
B≈N	7.15	2.10	C≈Te	21.41	3.81
B≈O	6.75	2.14	N≈Se	13.31	2.86

The symbol ≈ stands for aromatic bonds.

🕮 [Pauling, Brockway *et al.*, 1935; Bernstein, 1947; Gutman, Bosanac *et al.*, 1978; Randić, 1991g; Randić, 1993c; Hansen and Zheng, 1994; Randić, 1994b; Oláh, Blockhuys *et al.*, 2006; Sedlar, Andelic *et al.*, 2006]

- **bond order-weighted edge adjacency matrix** → edge adjacency matrix
- **bond order-weighted edge connectivity index** → edge adjacency matrix
- **bond order-weighted vertex connectivity indices** → connectivity indices
- **bond order-weighted Wiener index** → weighted matrices (⊙ weighted distance matrices)
- **bond profiles** → molecular profiles
- **bond rigidity** → flexibility indices
- **bond rigidity index** → flexibility indices (⊙ bond flexibility index)

- **bond spectral moments** → edge adjacency matrix
- **bond type *E*-state indices** → electrotopological state indices
- **bond vertex degree** → vertex degree
- **bootstrap** → validation techniques
- **Bowden–Wooldridge steric constant** → steric descriptors (⊙ number of atoms in substituent specific positions)
- **Bowden–Young steric constant** → steric descriptors (⊙ Charton steric constant)
- **branching ETA index** → ETA indices
- **branching index** ≡ *Randić connectivity index* → connectivity indices
- **branching indices** → molecular complexity (⊙ molecular branching)
- **branching layer matrix** → layer matrices
- **Braun–Blanque similarity coefficient** → similarity/diversity (⊙ Table S9)
- **Bray–Curtis distance** ≡ *Lance–Williams distance* → similarity/diversity (⊙ Table S7)
- **Brillouin redundancy index** → information content
- **Broto–Moreau–Vandicke hydrophobic atomic constants** → lipophilicity descriptors
- **Buckingham potential function** → molecular interaction fields (⊙ steric interaction fields)
- **bulk descriptors** → steric descriptors
- **bulkiness of an atom** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **bulk representation** → molecular descriptors
- **Burden eigenvalues** → spectral indices
- **Burden matrix** → weighted matrices (⊙ weighted adjacency matrices)
- **Burden modified eigenvalues** → spectral indices (⊙ Burden eigenvalues)
- **Buser distance** ≡ *Baroni–Urbani distance* → similarity/diversity (⊙ Table S7)