Quantitative Structure-Selectivity Relationships (QSSR) in Enantioselective Catalysis: Past, Present and Future

Andrew F. Zahrt, Soumitra Athavale, and Scott E. Denmark*

Roger Adams Laboratory, Department of Chemistry, University of Illinois
Urbana, Illinois, 61801

SUPPLEMENTARY MATERIALS

Table of Contents *Page*

Detailed Description of Conformer Independent Chirality Codes (CICC)

Aires-de-Sousa and Gasteiger have developed chirality codes to represent chiral compounds.[1,2] Depending on the need to consider specific molecular conformations, these representations are termed conformer independent chirality codes (CICC) and conformer dependent chirality codes (CDCC). Both chirality codes are constructed by transforming the 3D molecular structure into a radial distribution function of the form:[3]

$$g(r) = \sum_j \sum_i a_j a_i e^{-b(r-r_{ji})^2}$$

In this general equation, a is a molecular property of atoms i and j, r_{ij} is the distance between atoms i and j, b is a smoothing parameter, and r is the radius scanned in the radial distribution function. However, the representation in this equation is limited because it is not able to distinguish between different enantiomers of a molecule. To address this limitation, Aires-de-Sousa and Gasteiger developed CICC by introducing two terms: E_{ijkl} and S_{ijkl}. The term E_{ijkl} considers four atoms belonging to four different "neighborhoods" of a molecules. A neighborhood is defined as different groups affixed to a stereogenic, tetrahedral center (Figure 43).

$$E_{ijkl} = \frac{a_i a_j}{r_{ij}} + \frac{a_i a_k}{r_{ik}} + \frac{a_i a_l}{r_{il}} + \frac{a_j a_k}{r_{jk}} + \frac{a_j a_l}{r_{jl}} + \frac{a_k a_l}{r_{kl}}$$

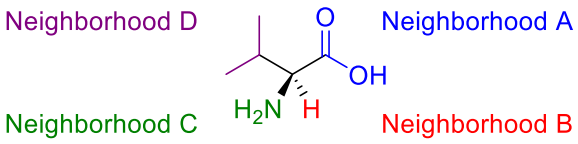

Figure S1. Depiction of the four neighborhoods of valine.

Here, a is an atomic property of atom i, j, k, or l, and r is the distance separating the two atoms of interest. In order to make the chirality code independent of molecular conformation, r is defined as the minimum sum of the bond lengths separating the two atoms of interest. For example, in the case of valine (Figure S1), the four atoms in neighborhoods A, B, C and D can

be the carboxylic acid carbon, hydrogen, nitrogen and methine carbon respectively, with the atomic property *a* selected as the charge on these atoms. Whereas the E_{ijkl} term accounts for the molecular properties at increasing radius from a stereocenter, the term S_{ijkl}, or chirality signal, accounts for the actual configuration of the stereocenter. This term only has a value of +1 or -1 and is calculated by first ranking the priority of atoms *i, j, k,* and *l* on the basis of their atomic property of interest, *a* (this value can be atomic number, atomic charge, etc). The coordinates of the three highest priority atoms are used to define a coordinate system in which all three points are in a plane. If the atoms are ordered clockwise by priority as defined by the atomic property *a* and the fourth atom is above the plane, S_{ijkl} obtains a value of +1. In this case, above refers to a positive value in the *z*-direction if the first 3 points are placed in the *xy* plane and the coordinate system is constructed in accordance to the left hand rule. Alternatively, if the fourth atom is in below the plane (negative in the *z*-direction), S_{ijkl} obtains a value of -1. Thus, when used together, E_{ijkl} and S_{ijkl} describe both the environment and configuration at incremental radii around a stereogenic center. These terms are used to calculate the CICC as follows:

$$f_{CICC}(x) = \sum_i \sum_j \sum_k \sum_l S_{ijkl} e^{-b(x-E_{ijkl})^2}$$

Detailed Description of Counterpropagation Network.

A depiction of this network architecture is given in Figure S1.

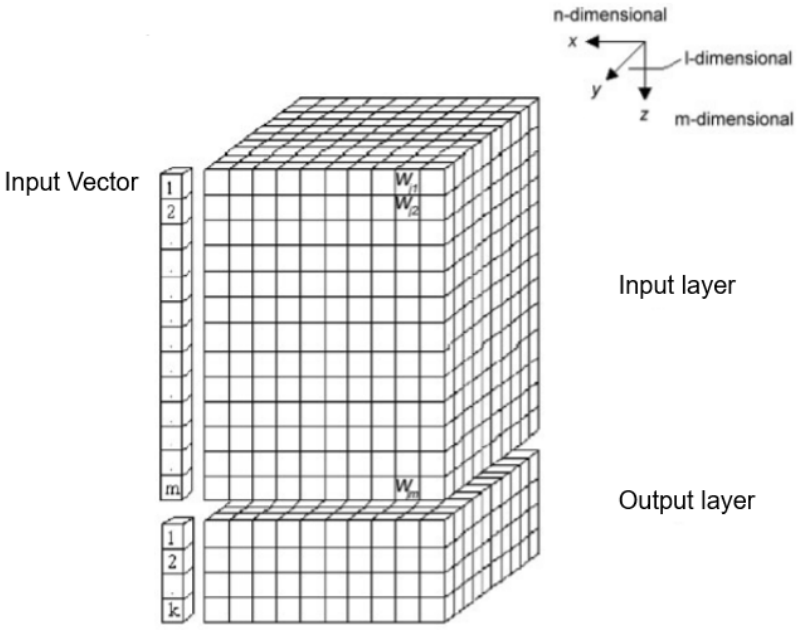

Figure S1. Depiction of a couterpropogation neural network. Reprinted (adapted) with permission from reference 135. Copyright 2001 American Chemical Society.

Counterpropagation neural networks are trained considering only the input vectors, making the training is unsupervised. The Kohonen layer consists of k by l nodes of m length. Each node can be thought of as a vector of m dimensions wherein the value in each dimension is a randomly assigned weight. Thus, a given node has numerous weights (w_1, w_2, … , w_m). An input vector of m dimensions is introduced to the network and the nearest node in m dimensional space is identified. An alternative way of describing this is that the node containing weights most similar to the input vector is identified. This node is termed the winning node, or best matching unit (BMU). The weight values of the winning node are then adjusted to more closely resemble the winning node (in counterpropagation neural networks, the output layer is accordingly modified). Similarly, the neighboring nodes are adjusted such that the extent of modification is

inversely proportional to their distance from the winning node. Compounds can then be classified on the basis of their proximity to nodes in the input layer. Because the input layer is linked to an output layer, the network can be used to make predictions on the basis of the input vector.

Conformer Dependent Chirality Codes (CDCC)

The form of CDCC is similar to that of CICC. The E_{ijkl} term is identical in form, with the only difference being r is the interatomic distance between the two atoms of interest. Instead of the S_{ijkl} term, CDCC contain a c_{ijkl} term, defined as:

$$c_{ijkl} = \frac{x_j y_k z_l}{x_j y_k + x_j|z_l| + y_k|z_l|}$$

In this equation, the atoms have been transformed into a Cartesian coordinate system in which atom i is at the origin, atom j is located along the x-axis, and atom k is on the xy-plane. Similarly to CICC, the atoms *i, k,* and *k* are used to define a plane and the sign of c_{ijkl} is dependent on whether atom l is in front of or behind that plane. However, the value of c_{ijkl} is now continuous with the magnitude proportional to the distance between atom l and the plane formed by atoms *i, j,* and *k.* Thus, CDCC is calculated with the equation:

$$f_{CDCC}(x) = \sum_i \sum_j \sum_k \sum_l c_{ijkl} e^{-b(x - E_{ijkl})^2}$$

Detailed Description of Select Topological Descriptors

The Randić index is derived from a connectivity matrix, such as the one constructed for 2-methylbutane (Figure 74). Three matrices are depicted, in which A represents atom pairs that are separated by 1 bond (1st order), B represents atom pairs separated by two bond (2nd order) and C represents atom pairs separated by 3 bonds (3rd order).

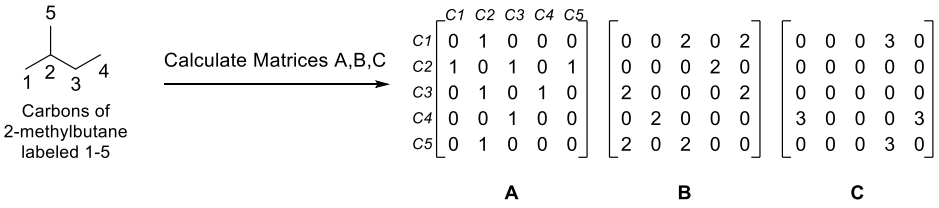

Figure S2. Connectivity matrices for 2-methylbutane.

The Randić index is then calculated for one connectivity matrix and is defined as the summation of the product of the square roots of the valences over each atom pair by the equation:

$$X^I = \sum (\delta_i \delta_j)^{-\frac{1}{2}}$$

Here, δ is the vertex degree of atom *i* and *j*. Thus, atoms 1 and 2 from Figure 74 contribute 0.7071 to the Randić index, and the total 1st order Randić index for 2-methylbutane is 2.2701.

The Kier and Hall index is a topological index defined as follows:

$$X^v = \sum \prod_{k=1}^{m+1} (\delta_k)^{-\frac{1}{2}}$$

In which δ_k is the is the valency for atom *k*, defined as follows:

$$\delta_k = \frac{(Z_k^v - H_k)}{(Z_k - Z_k^v - 1)}$$

Z_k^v is the number of valence electrons for atom *k*, H_k is the number of hydrogens bound to atom *k*, and Z_k is the total number of electrons in atom k. The index can be *n* order, in which *n* is 0, 1, 2, 3, etc. and defines the length of the bond path and *n* = 0 is simply the atomic valence connectivity index. The Kier shape index has multiple forms and attempts to grasp shape characteristics of a molecule. For example, 1st, 2nd, and 3rd order indices are given by

$$\kappa^1 = (N_{SA} + \alpha)(N_{SA} + \alpha - 1)^2({}^1P + \alpha)^2$$

$$\kappa^2 = (N_{SA} + \alpha - 1)(N_{SA} + \alpha - 2)^2({}^2P + \alpha)^2$$

$$\kappa^3 = (N_{SA} + \alpha - 1)(N_{SA} + \alpha - 3)^2({}^3P + \alpha)^2 \quad if\ N_{SA}\ is\ odd$$

$$\kappa^3 = (N_{SA} + \alpha - 3)(N_{SA} + \alpha - 2)^2({}^3P + \alpha)^2 \quad if\ N_{SA}\ is\ odd$$

Here, κ is the shape index, N_{SA} is the total number of non-hydrogen atoms, ${}^nP$ is the number of paths length *n*, and α is the van der Waal's radius of the atom normalized to the van der Waal's radius of a C_{sp3} atom, subtracted by one.

References

(1) Aires-de-Sousa, J.; Gasteiger, J. New Description of Molecular Chirality and Its Application to the Prediction of the Preferred Enantiomer in Stereoselective Reactions. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 369-375.

(2) Aieres-de-Sousa, J.; Gasteiger, J. Prediction of Enantiomeric Selectivity in Chromatography: Application of Conformation-Dependent and Conformation-Independed Descriptors of Molecular Chirality. *J. Mol. Graphics Model.* **2002**, *20*, 373-388.

(3) Hemmer, M.C.; Steinhauer, V.; Gasteiger, J. The Prediction of the 3D Structure of Organic Molecules from Their infrared Spectra. *Vibrat. Spectrosc.* **1999**, *19*, 151-164.