QCODES − Fast Topological Descriptors for Macromolecules

Edgardo Garcia*

Laboratório de Química Computacional, Instituto de Química, Universidade de Brasilía, Brasília DF 70910-900, Brasil

Received May 24, 2002

A fast iterative algorithm to obtain topological atomic and molecular descriptors, called Qcodes, is presented in this work. Its linear time dependence with the size of the structures and low memory requirement allows practical use for small molecules as well as macromolecular systems. The descriptors are unique and are able to sense small topological changes in atom's neighborhood. These properties make them suitable as atom and molecular identifiers to be used in database searching, force field atom typing, isomorphism, and automorphism algorithms. Examples of descriptors, their convergence behavior, sensitivity, ambiguity, and algorithm's performance are shown in this paper.

INTRODUCTION

Atomic and molecular descriptors are numbers, or sets of numbers, that are invariant with molecular structure labeling. They are of enormous utility in any application where atom neighborhood or molecular similarity needs to be accessed quickly such as database structure manipulation and screening. Experimental data and quantum mechanical properties can be used together or separately to build atomic and molecular descriptors. Another approach is to obtain descriptors from molecular graph invariants, generally called topological descriptors. The idea is as old as the application of graph theory in chemistry and a broad diversity of such descriptors have been described in the literature.[1−5] They can be classified, according to the nature of algorithm employed, in three general classes: based on detection and count of known graph patterns such as rings and functional groups, based on self-avoiding path enumeration and based on iterative procedures. Iterative algorithms have the advantage of not requiring preestablished patterns, which could limit the representation power in the first class of descriptors. Also iterative procedures do not have the strong memory and computational time dependence with the size of the system that path enumeration algorithm do. The Qcode algorithm presented in this paper is of iterative nature, and codes are formed by sets of real numbers that encode changes in atom's chemical topological radial neighborhoods. QCodes are finding practical use in new force field development strategies, isomorphism-automorphism checking, automatic isomer and structure generators, similarity indices, and other applications which require local topological discrimination. In force field development they are being used to verify atomic topological equivalence, to find stereocenters as a replacement of CIP rules, in automatic parameter generators and as a new consistent way to name atom types according to their local topology. Qcodes are currently implemented in MATCHEM,[6,7] a PROLOG language extension for representation and manipulation of chemical structures and reactions. Some important characteristics of Qcodes are the simplicity and high performance of its algorithm, good convergence, transferability, and sensitivity to subtle changes in atomic neighborhood, achieving a good representation power with very low ambiguity. Although exact zero ambiguity cannot be formally proved, descriptor's ambiguity can be tested for practical purposes in molecular sets containing a broad range of structural diversity. Qcodes were tested from small to big systems including organic, inorganic, and organometallic molecules as well as zeolites, bucky-balls, bucky-tubes, polymers, and biomolecules containing hundreds of atoms. There was not observed a loss of representation power with increase in molecular size or complexity. The algorithm can work with explicit and implicit hydrogen structures as well as hybrid representations. Any element of the periodic table can be present. It has low memory requirements and linear time dependence with the size of the molecule, since no path enumeration is needed.

QCODE ALGORITHM FOR ATOMIC DESCRIPTORS

An extension of Smith's[8] electronegativity equilibration concept was employed in an iterative way to obtain the Qcodes, which could be considered as topological atomic partial charges although no linear correlation with quantum mechanical charges is assumed in this work. They are not intended to substitute atomic charges but to be used only as a descriptor of atoms in molecules, i.e., just a number that varies in a consistent and unique way with changes in atom's neighborhood. Possible correlation with other electro-topological descriptors, like Gasteiger-Marsili[9] charges and Hall-Mohney-Kier[10] E-State indices, which are implemented in commercial software, were not accessed at the present time. Qcodes are unique, depending only on the atom's local radial neighborhood, and are more sensible to topological changes than Gasteiger charges which need a damping factor to reduce oscillations and guarantee fast convergence for its self-consistent procedure. Also, and especially important in applications involving macromolecular systems, Qcodes are computationally less demanding than E-state indices which are based on path enumeration. To obtain QCode atomic descriptors (AQcodes) only the molecule's topological structure and one parameter per atomic element are needed. Pauling atomic electronegativities were chosen as parameters

10.1021/ci025542+ CCC: $22.00 © 2002 American Chemical Society
Published on Web 10/08/2002

because they are unique numbers tabulated for all atoms of the periodic table. A modification to the original Qcode algorithm[7] is described in this paper as its currently implemented in MATCHEM.

Given the molecular graph (molecular constitution and connectivity table) the atomic Qcode values are obtained in the following way:

1. Each atomic element (i) of the structure receives a starting numeric value (X_i). Pauling atomic electronegativities, as tabulated in the literature[11] for neutral atoms, are used as the starting values (X_i).

2. Before beginning with the iterative process a new value (X_i^o) is calculated for each atom (i) to include the effect of the atom's coordination number where α_i = the number of bonds originating from atom i:

$$X_i^o = X_i/(\alpha_i + 1)^{1/2} \quad (1)$$

3. In each iteration (k) the arithmetic average $\langle X_\alpha^{k-1}\rangle_i$ of the atom's (X_α^{k-1}) values directly connected to atom (i) are calculated and new atomic values for each atom are given by the average:

$$X_i^k = (X_i^o + \langle X_\alpha^{k-1}\rangle_i)/2 \quad (2)$$

4. After each iteration (k) AQcode values (Q_i^k) are calculated from the change in electronegativity relative to the unperturbed atomic center (X_i^o):

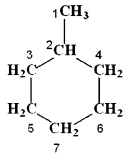
$$Q_i^k = (X_i^k - X_i^o)/X_i^o \quad (3)$$

5. Return to step 3 until reaching the desired number of iterations.

Note that no self-consistency in the "total charge" is required, and the sum of all (Q_i^k) for the molecule does not need to be zero. In fact this constraint in the sum is not desired at all, avoiding oscillatory behavior and also leading to a "residual total charge" for the molecule which turns out to be a useful molecular descriptor (MQcode).

Steps 1 and 2 are performed just once, before the iterative cycle begins, and (X_i^o) values for all atoms are saved. At each iteration the Q_i^k values for all atoms are also saved as well as (X_i^k) values for the current iteration that would be needed in the next. Steps 3 and 4 can be repeated at will, in practice until convergence is reached within the numerical floating point precision used. The limit of iterations will depend on the purpose and decimal places used to represent the values. In this paper convergence behavior is analyzed using double numerical precision with 15 decimal places. Care must be taken when working with floating numbers specially in comparing two numbers; the same rounding or truncating procedure and number of decimal places must be used as an internal representation and file storage for both. For applications in which Qcodes have been used, namely isomorphism and automorphism prescreening, force-fields atom typing names and automatic parameter generators, calculations were done at double precision with up to 12 iterations, but all codes were rounded to 10 decimal places for internal use and storage. In this paper 15 decimal places internal representations and storage were employed only to test code's convergence and sensitivity for up to 26 iterations (see Figure 7).

Methyl-Cyclohexane AQcodes[4]



Atoms	Iterations 1	2	3	4
C1	0.136545,	0.073962,	0.101698,	0.088452
C2	0.045515,	0.062583,	0.050493,	0.057694
C3 & C4	0.091030,	0.062583,	0.073251,	0.068806
C5 & C6	0.091030,	0.068272,	0.073251,	0.072451
C7	0.091030,	0.068272,	0.073962,	0.072362
H(C1)	-0.133463,	-0.083415,	-0.106354,	-0.096187
H(C2)	-0.133463,	-0.116780,	-0.110524,	-0.114956
H(C3,C4)	-0.133463,	-0.100097,	-0.110524,	-0.106614
H(C5,C6)	-0.133463,	-0.100097,	-0.108439,	-0.106614
H(C7)	-0.133463,	-0.100097,	-0.108439,	-0.106354
MQcode[4]	**-1.231276,**	**-0.901470,**	**-1.003160,**	**-0.968118**

Figure 1. Explicit hydrogen representations of methylcyclohexane with their respective 4 iterations atomic and molecular Qcode descriptors, rounded to 6 decimal places.

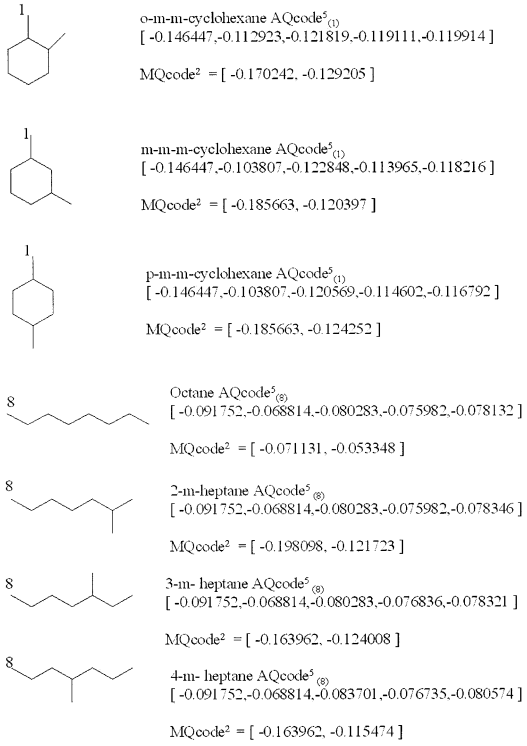


Figure 2. Qcode sensitivity test to variations in topology for C_8 isomer structures with their 5 iterations atomic and 2 iterations molecular Qcode descriptors, rounded to 6 decimal places.

The descriptor *AQcode* for each atom (i) in the molecule is built as the vector of (Q_i) values in all (k) iterations:

$$\mathrm{AQcode}_i^k = \{Q_i^1, Q_i^2, Q_i^3, ..., Q_i^k\}$$

A consequence of this algorithm is that the Q_i^k values at each iteration (k) respond to the effect of atoms from the (k) radial shell in the molecule. Note that these effects are included indirectly through a perturbation in the atom's first neighbor shell, giving good accuracy, stability, and convergence to the algorithm. Code values change through a perturbation wave that extends with the number of iterations. This characteristic leads to low memory requirements, fast execution, and linear dependence with number of iterations and system's size. For some applications only the (Q_i^k) values for the last iteration can be employed as an atomic index *AQID*, no longer needing to keep all previous values:

$$\mathrm{AQID}_i^k = Q_i^k$$

Atomic indices are actually being used in MATCHEM[7] system to improve the efficiency of exact isomorphism and automorphism algorithms. Their advantage over full vector descriptors are their short size with no loss of performance for this task, where the purpose is to prepartition the molecular graph atomic classes before the application of an exact atom−atom matching procedure. This implementation will be further explored in a future paper.

In Figure 1 Qcodes with 4 iterations are presented for methylcyclohexane molecule. Note that for clarity all values were rounded to 6 decimal places, but all internal calculations used 15 decimal places by default. As a matter of fact no more than 5 decimal places are necessary to differentiate among the atom's topological environment in the example. An interesting point is that, by looking only at individual AQID values, the C3's and C5's hydrogen atoms in methylcyclohexane were considered different at iteration 3, but in the next iteration they were incorrectly considered equivalent. Moreover, C2's and C3's hydrogen atoms have different AQID values at iteration 2 but the same ones at iteration 3. A similar case is observed between carbon atoms C3 and C5. This behavior is also present when using all 15 decimal places, its therefore not due to rounding errors, illustrating the fact that AQID with a sufficient number of iterations should be used for a given molecule. In the cases just sited AQID indices can distinguish correctly among all atoms with at least 4 iterations for carbons and at least 5 for hydrogens. Even when satisfying this point there is no way to guarantee that the AQID values could not be ambiguous at higher iterations. For the last reason it is always recommended to choose the complete AQcode if a better representation power is desired. Note that with the full AQcode, and keeping 5 decimal places values, just 4 iterations can correctly partition all atomic topological classes for the molecule in Figure 1. If a one number index is necessary for a given application, a less ambiguous ID like the sum of AQID iteration values for each atom could be employed as a good compromise between representation power and descriptor size.

To give an idea of how atomic Qcodes respond to changes in structural topology some cyclic and acyclic C_8 isomers had their Qcodes compared in Figure 2. One atomic element structures were selected to serve as a sensitivity test because code's variations will result solely from changes in topology. In the example the AQcodes for a terminal carbon atom, marked with a label, were compared to show how each iteration includes the effects from each respective neighbor radial shell. All code values were calculated with 5 iterations and rounded to 6 decimal places in this example, which is enough to classify all atoms within and across the molecules is the set. The first two acyclic molecules have the same 4 atom radial shell neighborhood for terminal atom 8; therefore, their $AQCode_{(8)}$ is exactly the same if just the first 4 iterations are considered; the change happens in the 5th iteration at the 4th decimal place. An interesting fact is that there is no need to use 5 iterations if the only goal is to distinguish among the molecules, just 2 iteration molecular codes (MQCodes[2]) are able to do so as can be seen in Figure 2.

MOLECULAR DESCRIPTORS

A molecular vector descriptor *MQcode* is built by summing all the atomic (Q_i) values at each iteration:

$$\mathrm{MQcode} = \{\Sigma Q_i^1, \Sigma Q_i^2, \Sigma Q_i^3, ..., \Sigma Q_i^k\}$$

Analogously a molecular index *MQID* that considers only the last iteration values for all atoms in the molecule is defined as

$$\mathrm{MQID}^k = \Sigma Q_i^k$$

Examples of molecular descriptors, with 4 iterations (MQcode[4]) and 6 decimal places, are presented in Figure 3 for some carbon isomer structure pairs. Most of the graphs shown in this figure have no chemical reality but are usually employed to check the ability of molecular codes to distinguish among similar isomers of diverse complexity. The examples include fused ring structures and other acyclic graphs as well as dodecane polymers many of which are well-known for having the same path count codes. For the graphs in the example no more than 4 iteration MQcodes and 6 decimal places were needed to check their isomorphism.

Although with the summations involved in calculation of MQcodes some representation power loss would be expected, such molecular descriptor and indices are still highly sensitive to changes in topology even in macromolecular structures using small number of iterations and few decimal places code representation. In this study MQID values with less than 5 iterations and 6 decimal places, for medium size molecules, and up to 9 iterations with 7 decimal places in proteins and DNA/RNA were enough to discriminate among isomeric structures with subtle topological changes. Comparisons of $MQID^1$ (at iteration 1), rounded to 7 decimal places, are shown in Figure 4 for macromolecular isomers created by slight modifications in the original structures of carbon bucky-tube, NaCl cubic crystal, trypsin protein, and RNA strand. Arrows indicate the molecule's regions that were altered. Care was taken to keep the same number and type of bonds and atoms in the original and modified molecules. For high cyclic complexity molecules a defect in the structure of the material was created, as in the bucky-tube case, or moved from one corner to another, as done in the NaCl crystal. In the biomolecules the position of one terminal oxygen atom was moved to an adjacent carbon in amino acid tyrosine (TYR10) and in a ribose sugar unit, for the modified trypsin (A) and modified RNA strand, respectively. In all cases changes introduced by the modifications propagate through the entire structures as reflected by changes in code values. The effects are readily sensed by the Qcodes giving different MQID values already at the first iteration and needing just 2 decimal places. Only in modified trypsin (B) 9 iterations were required, with at least 7 decimal places, to verify that the structures are not isomorphic. Modification (B) consisted of a protein primary sequence change, exchanging positions of two similar amino acids, tyrosine and phenylalanine, in similar surroundings making it more difficult to detect the topological changes. An interesting property of the MQID index is that it absolute value does not show strong dependence on the size of the molecule, which is usually the case of some indices based on path counts. MQcodes could be used in conjunction with the later to build QSPR/QSAR models. MQID indices seem to encode more strongly a measure of topological electronegativity diversity within the molecules, as expected from their

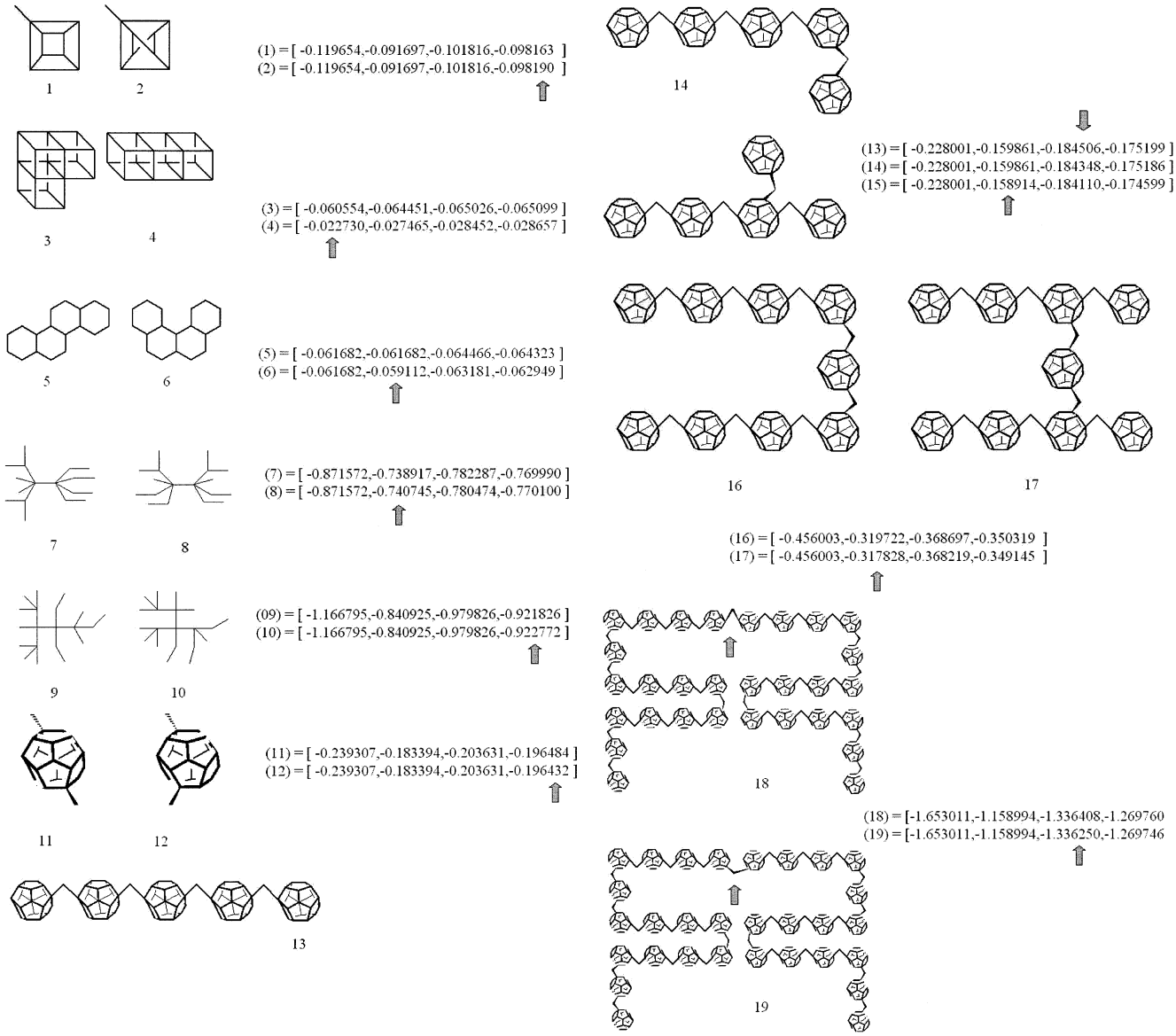


Figure 3. Nonisomorphic isomer graphs with their respective 4 iteration molecular codes (MQcode[4]) rounded to 6 decimal places. Arrows indicate where changes in MQcode values occur.

definition. Nevertheless a careful study of interdependence with other indices and correlation with molecular properties would be needed to reach serious and useful conclusions on this regard. At the present time Qcodes are being used only as topological atomic and molecular identifiers.

BOND ORDER INCLUSION

Purposely no explicit consideration of bond order was included in the algorithm's formulas, the descriptor values represent indirectly the valence hybridization state trough the atom's first shell when hydrogen atoms are included. This choice avoids problems such as equivalent atoms in different resonance structure representations having different descriptors. Bond order is an arbitrary concept, a myriad of commercial molecular file formats as well as force fields treat them in a different way. For example structures with moieties such as $-NO_2$ can be represented in distinct forms such as O−N═O (single and double bond), O−N−O (two single bonds), and O∼N∼O (1.4 or 1.5 bond order values). All forms will have the same values of atomic and molecular Qcodes with the present algorithm. On the other hand, if we allow bond orders to be considered the atomic and molecular descriptors will depend on the bond order representation chosen and would be different for each of the above cases. If such descriptor bond order dependence is desired, a simple substitution of eq 3 by eq 4, in step 2 of the original algorithm, can account for it

$$X_i^o = X_i/[\Sigma(\beta_i)^{1/2} + 1)]^{1/2} \qquad (4)$$

where $\Sigma(\beta_i)^{1/2}$ is the sum of bond orders square roots originating from atom i.

The modified descriptors and indices obtained in this way are called *Qβcodes*. Both Qcodes and Qβcodes are identical for a molecule containing only single bonds (like methylcyclohexane), since α and β are the same for each atom. Atomic and molecular Qβcodes will be different for all $-NO_2$ representations described above. Another important application of Qβcodes is to distinguish among structure representations without explicit hydrogen atoms but contain-

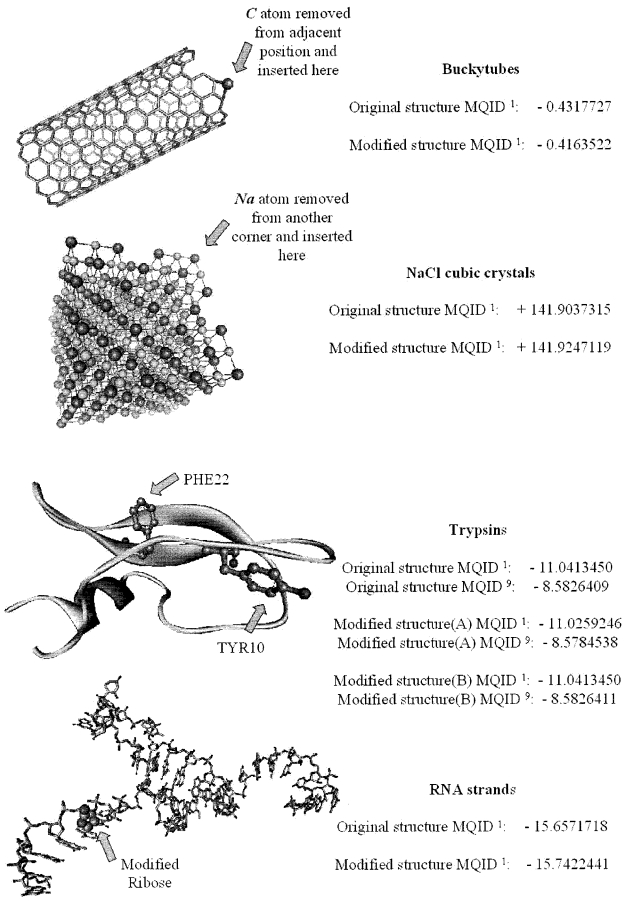
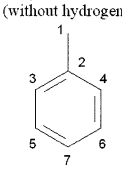


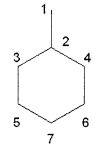
Figure 4. MQID discrimination power for macromolecular isomers. In all structural modifications the number and type of atoms and bonds are not altered. Places where modifications to the original molecules were done are marked with arrows. Code's values are rounded to 7 decimal places and calculated with 1 iteration ($MQID^1$), except in one structure that required 9 iterations ($MQID^9$).

Methyl-Benzene AQβcodes^4 (without hydrogen atoms)



Atom	Iterations 1	2	3	4
C1	-0.164780,	-0.122149,	-0.139673,	-0.132568
C2	0.127172,	0.074897,	0.096093,	0.086673
C3 & C4	-0.029879,	-0.001885,	-0.01526,	-0.008844
C5 & C6	0,	-0.007470,	-0.000471,	-0.004749
C7	0,	0,	-0.003735,	-0.000236
MQβcode^4	**-0.097365,**	**-0.065962,**	**-0.078777,**	**-0.073316**

Methyl-Cyclohexane AQβcodes^4 = AQcodes^4 (without hydrogen atoms)



Atom	Iterations 1	2	3	4
C1	-0.146447,	-0.103807,	-0.120569,	-0.113462
C2	0.120602,	0.073193,	0.093293,	0.084586
C3 & C4	-0.033494,	-0.007382,	-0.019740,	-0.013756
C5 & C6	0,	-0.008373,	-0.001846,	-0.005982
C7	0,	0,	-0.004187,	-0.000923
MQβcode^4	**-0.092831,**	**-0.062126,**	**-0.074634,**	**-0.069276**

Figure 5. m-Benzene and m-cyclohexane without hydrogen atoms and their 4 iteration Qβcodes rounded to 6 decimal places.

ing bond order information. One example is when comparing descriptors for toluene and methylcyclohexane, if we remove hydrogen atoms both structures will have the same Qcodes but different Qβcodes. The advantage of using explicit hydrogen together with Qcodes is a greater discrimination power due to the increase in neighborhood diversity. However, sometimes it is preferred to store molecules without explicit H's in order to reduce memory requirements in databases or simply because they are not available from the experimental source, as in macromolecular X-ray data.

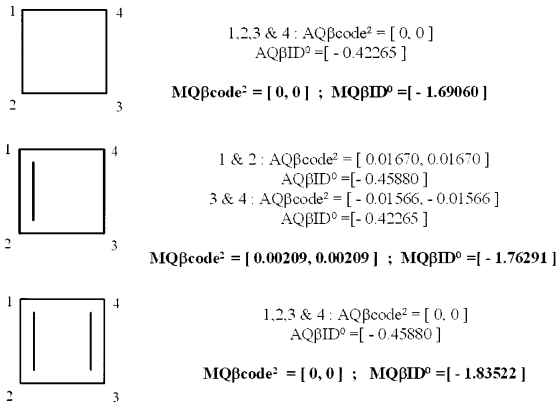


Figure 6. Cyclobutane, cyclobutene, and cyclobutadiene 2 iterations Qβcodes and QβID° values rounded to 5 decimal places.

In Figure 5 we can see that some atoms, like toluene's C5 to C7, have null Qβcode values. This happens when atom's neighbors are equivalent in the current iteration. It has to be pointed out that the full vector code will be null if all of the atoms in the structure are of the same type and topologically equivalent, i.e., one equivalence class. This will happen in some structures such as one atom class buckyballs as well as in no explicit hydrogen representations of cyclic molecules such as cyclobutane and cyclobutadiene shown in Figure 6. In the last case both molecules will have null molecular codes, even when using Qβcodes; therefore, the codes have no use in discriminating between the molecules neither among their atoms. One simple way to solve this problem is to always use explicit hydrogen representations for calculating the descriptors, only removing hydrogen atoms for storage purposes. As a second alternative the full Qβcodes vector descriptors can be augmented by adding a zero order iteration value (Q_i^o) for the atomic codes and the sum of such values (ΣQ_i^o) for the molecular codes. This would be calculated as the difference in electronegativity relative to the original atom's electronegativity tabulated value (X_i), by inserting eq 5 in step 2 of the algorithm before the iterative process begins. For each atom (i)

$$Q_i^o = (X_i^o - X_i)/X_i \qquad (5)$$

The Q^o_i values can be applied for both Qcodes and Qβcodes; however, it is in the last that they could play a more important role. AQβID^o_i and MQβID°, calculated with eqs (4) and (5), and Qβcodes with up to 2 iterations and rounded to 5 decimal places are shown in Figure 6. Now all atoms and molecules can be discriminated by the augmented descriptor. Even for explicit hydrogen molecular representations the inclusion of zero iteration values in atomic full vectors can be used as a first numeric identifier to build consistent names for atom types in force field development. If full vector descriptors are not desired, single numbers made by summing the zero and other iteration values can be adopted as atom and molecular indices.

CONVERGENCE AND AMBIGUITY

Since Qcodes are floating point numbers, the atomic codes variation dependence on the number of iterations has to be analyzed. The goal is to know the required number of decimal places in the representation and practical limit of

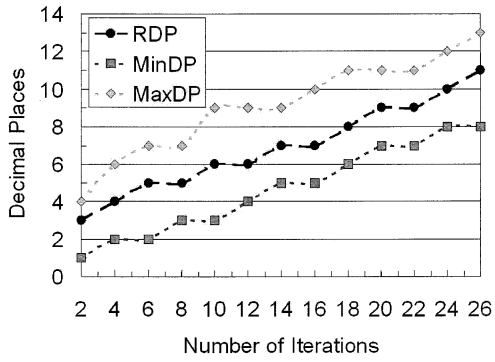


Figure 7. AQID values variation range for up to 26 iterations in 5 molecular sets containing 288 molecules and 14024 atoms with up to 10039 unique AQID.[26]

Table 1. Data Sets Used for Qcode Descriptors Convergence and Ambiguity Study[a]

set	molecules	atoms	unique AQID26	molecule's size (atoms) min.	max.	av	RMS
dekanes (H)	8	266	150	32	35	33	1
organics (H)	77	2381	1642	5	106	31	21
FLC (H)	31	2354	1356	38	93	76	14
FLC high P	167	6635	5250	28	52	40	5
biomolecules	5	2388	1641	243	783	493	222
total	288	14024	10039				

[a] Sets formed by molecules with explicit hydrogens are marked with (H), the rest contain no hydrogen atoms.

iterations imposed by the lowering of sensitivity and convergence behavior. This study involved a great number of atoms in molecular sets containing a broad diversity of functional groups, topologies, and sizes. Some molecules with very low diversity were included, like systems with only one atomic element and few topological equivalence classes. A summary of the data sets properties is presented in Table 1. The five test sets employed in the study contained a total of 288 molecules with sizes ranging from 5 to 783 atoms, some sets with explicit hydrogen representations, marked with (H) in Table 1, and others with no hydrogen atoms. The *Biomolecules* set is formed by the 3 trypsin isomeric forms and original RNA strand from Figure 3 plus a DNA strand. Ferroelectric liquid crystals with high polarization, taken from LiqCryst Database[11] and others designed in Prof. David Walba's group,[12] form the *FLC high P* and *FLC* sets, respectively. The *Organics* set includes organic and some inorganic molecules, chosen from a commercial software structures database[13] to get a good diversity of shapes and topologies. Finally monosubstituted alkanes of the form $C_{10}H_{21}-R$, with R: NH_2, F, Cl, OH, CN, CH_3, H, $CH(CH_3)_2$ at a terminal position, constitute the *Dekanes* set.

A total of 14024 atoms had their AQID indices calculated and compared up to 26 iterations, generating a total 10039 unique AQID[26] indices, an average of about 70% diversity in atomic topological environments for all sets. Absolute variations in AQID^k values at each iteration were calculated as $\Delta Q^k = |Q^k - Q^{k-1}|$ for all data set atoms using double precision 15 decimal places representation, on a PII-400 MHz processor running Arity-Prolog32 interpreter[14] v1.1. Atomic indices values (AQID) present approximate linear convergence behavior with a number of iterations as can be seen in Figure 7 by the convergence limiting curve marked with squares (MinDP). This curve points are the number of decimal places correspondent to the highest observed variations (ΔQ^k_{MAX}) at each iteration k for all atoms across all sets. They are an indication of the minimum number of decimal places needed to grasp changes in code values for a given iteration. The curve at the top (MaxDP), marked with diamonds, shows the decimal places correspondent to the lowest observed variations (ΔQ^k_{MIN}) at each iteration k. The middle curve, marked with balls, is the recommended minimum number of decimal places (RDP) required to capture most of the observed index value variations. RDP were obtained by subtracting ΔQ^k_{RMS} values from average values (ΔQ^k_{AVRG} - ΔQ^k_{RMS}) at each iteration k. Let us suppose as an example that codes with up to 10 iterations are desired. The maximum observed variation occurred at the third decimal place (MinDP); therefore, all indexes for the sets converged in the second decimal place ($\Delta Q^k_{MAX} < 1 \times 10^{-2}$). It is recommended that at least 6 decimal places (RDP) are used to represent index values, but if maximum code sensitivity is wanted then 9 decimal places (MaxDP) will suffice. Calculation of averages, RMS, minimum, and maximum ΔQ^k values included all atoms in all sets of molecules employed for the present study. Repeated and null AQID values were excluded from the statistics to avoid biasing the results due to differences in atom environment diversity within each particular set. The resulting data do not constitute a formal proof of convergence and code behavior but a practical way to evaluate expected code variation ranges as a function of increasing radial neighbor shells.

The ambiguity of the indices was measured by comparing the number of unique atomic and molecular indices, at each one of the 26 iterations, relative to the number of full vector atomic and molecular codes, respectively. A maximum ambiguity of 0.5% was observed in the atomic indices; 8 out of 2381 atoms had ambiguous AQID^k but different AQcodes^k in the *Organics* set. Regarding the molecular indices (MQID^k), the maximum observed relative ambiguity was 3%, 1 molecule in 31 for the *FLC* set. These results are only valuable to make decisions as whether to use full codes or single number indices. Full codes are less ambiguous than indices relative to results of an exact isomorphism algorithm;[6,7] no ambiguity at all was observed for MQcodes with up to 10 iterations in the set of 288 molecules used in this study, neither among the 26 structures in Figures 2 and 3. Of course this is a small set of molecules to guarantee a high discrimination power for MQcodes in any molecular environment. Exact topological ambiguity for the atomic descriptors can only be verified against an exact atom matching algorithm, that was not done at the present moment and will be presented in a future paper for a much bigger set of molecules with high atomic topological diversity.

COMPUTATIONAL PERFORMANCE

Computational time and memory dependence relative to the size of the system are of practical importance. For some applications such characteristics can determine the type of descriptors to be used, the number of iterations, and therefore decimal places in representations. In the present study Qcodes were obtained for all molecules of test sets discussed previously. Both memory and time increased linearly with the size of the system and number of iterations. Results for

Table 2. Time Performance for Calculation of AQcodes with 10 Iterations for Structures Shown in Figures 2 and 3 on a P2-400 MHz CPU Running Arity-PROLOG32 Interpreter

structure	atoms	bonds	CPU$^{(10)}$ time (s)
1, 2	9	13	0.02
3, 4	16	28	0.04
5, 6	18	21	0.04
7, 8	18	17	0.04
9, 10	20	19	0.05
11, 12	22	32	0.06
13, 14, 15	104	158	0.50
16, 17	188	286	1.26
buckytubes	336	490	3.90
NaCl crystals	336	855	6.00
trypsins	454	468	5.80
18, 19	629	958	9.90
RNA strands	783	876	12.50

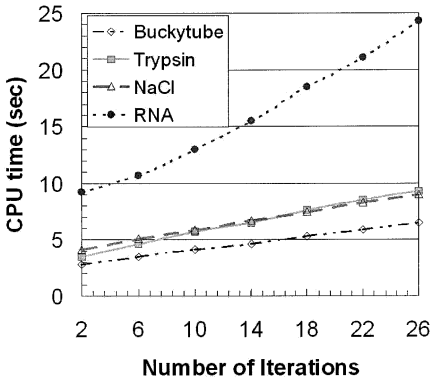


Figure 8. Linear time dependence on the number of iterations. PII-400 MHz CPU running Arity-Prolog32 interpreter.

graphs and molecules of Figures 2 and 3 are summarized in Table 2. Examples of time performance versus iterations for Buckytube, NaCl, trypsin, and RNA strand are presented in Figure 8.

Although CPU time is dominated by the number of atoms, there is also a time dependence on the number of bonds. Trypsin has almost the same number of bonds than the buckytube but only 35% more atoms showing a 49% increase in CPU time. Buckytube and NaCl structures have an equal number of atoms, but the latter has 75% more bonds leading to a less expressive 54% increase in time. A slight nonlinearity with system size is present. This is due to the way the atom X_i^k values are retrieved from the internal database to calculate the averages in step 3 of the algorithm. The current implementation saves the values in the order they are generated, a Hash-table database structure or a list of neighbors could improve performance considerably. The present algorithm's implementation does some redundant computations; also all parts of the program are interpreted so its speed could be increased at least 4 times by simply optimizing and compiling. Memory requirements to generate and store the full Qcodes are of the order of 16 bytes per atom per iteration. To generate and keep in a text file all full AQcodes for a 1000 atom structure, with 10 iterations and 15 decimal places, approximately 160Kbytes of memory will be used. The 288 molecules data set used in the convergence study, containing full AQcodes with 26 iterations for 14026 atoms, required almost 6 MB of storage in a regular ASCII text file. If necessary this can be reduced by a factor of 5 when compacted with appropriate software like WinZip[16] or similar.

CONCLUSIONS

The descriptors presented in this work have excellent discriminating power, as seen in high atomic environment diversity test sets, not showing any degradation of sensitivity even in macromolecules. Their sensitivity is limited by convergence, and it decreases in magnitude with an increase in the number of shells because codes are damped by averaging. However it is not a practical limitation if the descriptors are intended for local atom property correlation (QSPR), similarity checks, and automorphism prepartition. Properties such as quantum mechanics atomic partial charges, RMN shield factors, bond lengths and bond orders, and harmonic force constants, among others, are also damped and seem to converge before the 10th topological neighbor radial shell limit. In test cases for molecular descriptors (MQcodes), as the ones shown in Figures 2 and 3, it could be verified that a small number of iterations suffices for detecting subtle changes within molecular structures containing from 9 atoms to more than 700 atoms. The atomic descriptor *AQcode* is being used to name atom types in the development of new force fields, parametrized on demand for specific systems from quantum mechanical databases.[17,18] Such development strategies require an automatic and consistent form of nomenclature, imposed by the large amounts of available data and great number of generated force field atom type parameters, avoiding the confusion caused by commonly used arbitrary naming rules. The element name followed by the AQID° augmented full vector atomic descriptor, up to a given number of iterations, constitutes a text string that represents in a unique and extensible way the atom and its topological radial shell. There is no size limitation to this naming protocol leading to the possibility of a virtually infinite number of atom types, certainly more than enough to supply the needs of classical force field development.

Note Added after ASAP Posting. This article was released ASAP on 10/8/2002 before additional changes were made. The correct version was posted on 10/9/2002.

REFERENCES AND NOTES

(1) Pogliani, L. From molecular connectivity indices to semiempirical connectivity terms. Recent trends in graph theoretical descriptors. *Chem. Rev.* **2000**, *100*, 3827−3858.
(2) Mihalic, Z. Comparative study of molecular descriptors derived from distance matrix. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 28−37.
(3) Trinajstic, N. Chemical graph theory, 2nd ed.; CRC Press: Boca Raton, 1992.
(4) Balaban, A. T., Ed. *Chemical applications of graph theory*; Academic Press: London, 1986.
(5) Balaban, A. T., et al. *Topological indices for structure−activity correlations. Topics in current chemistry*; Springer: Berlin, Heidelberg, New York, Tokyo, 1983; Vol. 114, pp 21−55.
(6) Garcia, E.; Reyes, L. M. PROLOG representation of molecular structures and pattern recognition. *J. Mol. Struct. (TEOCHEM)* **1993**, *282*, 175−185.
(7) Garcia, E. MATCHEM: A symbolic model for the computer representation and manipulation of chemical structures and reactions, Ph.D. Thesis, Instituto de Química, Universidade de Brasilia, Brasilia DF, Brasil, 1994. MATCHEM is currently implemented in Arity-Prolog 32 for MS Windows operating systems.
(8) Smith, D. W. A new method of estimating atomic charges by electronegativity equilibration. *J. Chem. Educ.* **1990**, *67*, 7, 559−562.
(9) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity − a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3288.

(10) Hall L. H.; Mohney B.; Kier, L. B. The electrotopological state: structure information at the atomic level for molecular graphs. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76−82.
(11) Isaacs, N. S. *Physical Organic Chemistry*; Longman Group UK Limited: 1987; p 31.
(12) WebLab ViewerLite 3.5; Molecular Simulations Inc., Now Accelrys DS ViewerLite, 1999 (www.accelrys.com).
(13) Volkmar Vill, LiqCryst − Database of Liquid Crystalline Compounds, (http://liqcryst.chemie.uni-hamburg.de), LCI Publisher GmbH, Eichenstr. 3, D-20259, Hamburg, Germany.
(14) Glaser, M. A.; García, E.; Ginzburg, V. V.; Malzbender, R.; Clark, N. A.; Walba, D. M. Computer-aided design of ferroelectric liquid crystals. *Mol. Phys. Reports* **1995**, *10*, 26-47.
(15) Arity-Prolog32, A freely available Prolog interpreter and compiler, Arity Corporation (www. Arity.com), Damonmill Square, Concord, MA.
(16) WinZip, WinZip Computing Inc. (www.winzip.com), P.O. Box 540, Mansfield, CT.
(17) García, E.; Glaser, M. A.; Clark, N. A.; Walba, D. M. HFF: A force field for liquid crystal molecules. *J. Mol. Struct. (TEOCHEM)* **1999**, *464*, 39−48.
(18) Glaser, M. A.; García, E.; Clark, N. A.; Walba, D. M. Quantum chemistry based force fields for soft matter. *Spectrochim. Acta Part A* **1997**, *53*, 1325−1340.

CI025542+