MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **71** (2014) 645-656

ISSN 0340 - 6253

Use of CTI Index for Perception of Duplicated Chemical Structures in Large Chemical Databases

Emil Petrov[1], Borislav Stoyanov[1], Nikolay Kochev[2], Ivan Bangov[3*]

[1] *Department of Computer Informatics*
Faculty of Mathematics and Informatics, Konstantin Preslavski University of Shumen,
115 Universitetska Str., Shumen, Bulgaria, epetrov1990@gmail.com, borislav.stoyanov@shu-bg.net

[2] *Department of Analytical Chemistry and Computer Chemistry,*
Faculty of Chemistry, University of Plovdiv,
24 Tsar Asen Str., Plovdiv, Bulgaria, nick@uni-plovdiv.net

[3] *Faculty of Natural Sciences, Konstantin Preslavski University of Shumen,*
115 Universitetska Str., Shumen, Bulgaria, ivan.bangov@gmail.com

(Received June 21, 2013)

Abstract: The employment of *C*harge-related *T*opological *I*ndex (*CTI*) devised by one of the authors (IB) for perception of duplicated structures in large structure collections has been studied. It is shown on a structural database of 249 000 chemical structures that the *CTI* values with precision more than 7 digits after the decimal point can produce safe discrimination between equivalent (isomorphic) and non-equivalent structures. Also the tests show that the *CTI* index does not give degenerate values for all alkane isomers of 17 carbon atoms.

Introduction

Duplicated structures frequently emerge in large chemical databases. Mathematically they are represented by isomorphic molecular graphs. Their perception and recognition by computers is a serious problem. The task of structure identification is particularly important in the context of modern chemical databases where multiple information sources (both free and commercial) are used and merged in order to provide large chemical collections. There are several approaches to the solution of this problem: use of hash codes, one-to-one comparison by pair wise mapping of the chemical structures, creation of a unique linear notation form or

[*] Corresponding author

electronic nomenclature, as well as the employment of topological indices. The use of hash code does not produce satisfactory results in all cases (especially when it is of 32 bits). Mapping of chemical structures and/or forming unique linear notation forms are rather slow procedures.

Molecular connectivity indices reduce the chemical structure representation to a number. However, it should be stated here that there is no mathematical proof for their discriminating efficiency. For example, the most popular topological indices such of Wiener [1] and Randić [2], frequently produce degenerated values, i.e., constitutionally different structures producing the same index values, as it is shown in Figure 1.

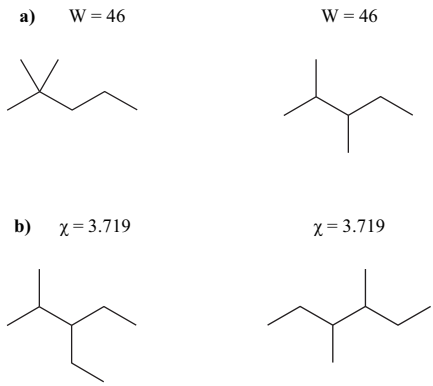


Figure 1: Non-equivalent structures producing the same Wiener (a) and Randić (b) index.

A *C*harge-related *T*opological *I*ndex (*CTI*) was introduced in 80's by one of the authors (IB) for solving the problem of 2D structure isomorphism within the computer-assisted structure generation from a gross formula [3-5]. This index was further extended toward 3D structures and it was named *C*harge-related *G*eometrical *I*ndex (*CGI*) [6]. The applicability of the two indices towards structure branching and to the quantitative structure/property relationships (QSPR) was investigated [6-8]. The purpose of this paper is to examine the usability of the Charge-related Topological Index more deeply to the problem of perception of duplicated (isomorphic) structures in large 2D collections of structures.

CTI index

The Charge-related Topological Index has the following form:

$$CTI = \sum_i \sum_j \frac{L_i L_j}{D_{ij}} \tag{1}$$

Here D_{ij} is the inter-atomic topological distance, i.e. the minimal number of bonds between atoms i and j (D_{ij} is an element of the topological distance matrix) and L_i and L_j are local indices characterizing the individual atoms i and j. L_i is defined as follows:

$$L_i = L_{0,i} - N_{H,i} + q_i\,, \tag{2}$$

where $L_{0,i}$ is the atom valence, $N_{H,i}$ is the number of hydrogen atoms attached to atom i (if atom i is a hydrogen then $N_{H,i} = 0$), and q_i is the corresponding charge density computed by either the topological empirical method of Gasteiger-Marsili [9] or by any of the more sophisticated quantum chemistry methods on semi-empirical or non-empirical level. In case we consider only the 2D topology of the molecule, the Gasteiger-Marsili method for calculation of the atomic charges and the topological distance matrix with inter-atomic distances are employed. In this case we have a Charge-related Topological Index. Vice versa, in case of using 3D molecular models, the distances and charge densities are calculated from the atom coordinates employing any of the quantum chemistry methods – in this case the index is no more topological. It could be used for 3D structure and conformer perception. Apparently, the generation of the *CGI* index is computationally very intensive procedure. Hence some quantum chemistry programs on a semi-empirical level could be used to this end. Additionally, the L_i values could be employed to the perception of the structure symmetry (solving the atom equivalence within the chemical structure) as reported in [10].

As mentioned above, like the other indices, the efficiency of the *CTI* cannot be strictly mathematically proved. Hence, the only way to examine the *CTI* efficiency is to apply the index for large collections of structures. It is expected that equivalent (isomorphic) structures produce the same *CTI* values (within a given precision), and different (non equivalent) structures – different values.

As seen from the relation (1) *CTI* consists of two parts – the numerator and denominator. Whereas the denominator accounts for the branching of the chemical structure in a way similar to that of the Wiener index, the numerator features the atomic type differences and their polarity. Furthermore, the charges (especially these produced from the

Gasteiger - Marsili method using iteratively practically all atom environments) experience the influence of the electron density of the whole molecule on each separate atom.

The Gasteiger-Marsili Iterative Partial Equalization of the Orbital Electronegativities (IPEOE) method is based on the following procedure [9]: The orbital electronegativities are defined according to Mulliken on the basis of ionization potentials I_p and electron affinities E_a:

$$\chi_{iv} = \frac{1}{2}(I_p + E_a) \tag{3}$$

Further, the electronegativity χ is considered to be dependent on the charge Q_i of the considered atom *i,* and for each atom it is calculated according to the following relation:

$$\chi_{iv} = a_i + b_i Q_i + c_i Q_i^2 . \tag{4}$$

Table 1: Coefficients a_i, b_i and c_i used by the Gasteiger-Marsili method.

Element	a_i	b_i	c_i
H	7.17	13.17	-0.56
Csp_3	7.98	9.18	1.88
Nsp_3	11.54	10.82	1.36
Osp_3	14.20	12.92	1.39
Fluor	16.96	13.85	2.31
Clor	11.85	9.69	1.35
Brom	10.08	8.47	1.16
Iod	9.90	7.96	0.96
Ssp_3	12	11	1.20
C_ar	8.79	9.32	1.51
N_ar	12.87	11.15	0.85
O_ar	17.07	13.79	0.47
Csp_2	8.79	9.32	1.51
Nsp_2	12.87	11.15	0.85
Osp_2	17.07	13.79	0.41
Ssp_2	16	13	0.3
Csp	10.39	9.45	0.73
Nsp	15.68	11.70	-0.27
Ph	8.79	9.32	1.51

An iterative procedure is applied as χ_{iv} for each iteration (except the first one) is calculated according to the equation (4), and χ_{iv} for the first iteration is given by the equation:

$$\chi_{iv} = a_i + b_i + c_i \tag{5}$$

The coefficients a_i, b_i and c_i are provided in Table 1. Further, for each iteration n, a contribution to each atom i charge density from adjacent atom j is calculated as follows:

$$\Delta q_{ij}^{(n)} = \left(\chi_{ijv}^{+}\right)^{-1}\left(\chi_i^{(n)} - \chi_j^{(n)}\right)\left(\frac{1}{2}\right)^n \tag{6}$$

where $\chi_{ijv}^{+} = max(\chi_i, \chi_j)$ is the electronegativity of the positive state (comparing the atoms i and j) used to scale the electronegativity values. Accordingly, the contributions are added to the previously calculated charge density:

$$Q_i^{(n)} = Q_i^{(n-1)} + \sum_j \Delta q_{ij}^{(n)} \tag{7}$$

The iterative procedure is carried out by calculating initially electronegativities χ_{iv} from eq. (5), then calculating charge contributions (6), and adding to the charge densities (7) and calculating again the electronegativities, χ_{iv}, for the next iteration (eq. (4)) and so on. Although the original method assumes a convergence, we use by default a fixed number of 6 iterations in our program. In this respect this approach algorithmically resembles some of the procedures for involvement of different environments around each atom within the structure (such as the HOSE code of Bremser [11], or the generation of the Daylight fingerprints [12]). Here, it should be mentioned that a modification of *CTI* was developed to include fragments [13]. Apparently, the charge density produced by this procedure experiences the influence of all surroundings of the whole structure. Accordingly, this combination of the two parts the nominator and the denominator makes the index very discriminative.

A serious problem arises from this approach. We do not have the orbital electronegativities and the coefficients a_i, b_i and c_i for all elements of the Periodic table. However, we must emphasize here that we do not need physically correct charge values for the solution of this particular task – the perception of isomorphic (constitutionally equivalent) structures and discrimination of non-equivalent structures. Basically we need charge densities which play a role similar to hash codes. As far as, for those elements we lack data we have used the atomic electronegativity to this end. Thus, we accepted the following approach:

1. For elements which have no orbital electronegativities we have exploited the atomic electronegativities.
2. For the a_i, b_i and c_i coefficients we have assumed a linear relationship between the atomic numbers and the coefficients in Table 1. Thus, the following relations have been obtained:

a_μ = element electronegativity of atom μ;

b_μ = 1.352 * En_μ;

c_μ = (1.2341194 * En_μ) – 3.7549954E-4,

where En_μ is the element number from the Periodic table.

Results and Discussions

A software program was developed in C++ for the calculation of *CTI* index and it was tested on the National Cancer Institute (NCI*)* database [14] consisting of about 249 000 chemical structures. An inner validation was carried out by comparing the *CTI* of each one of the structures against the other structures by using different precisions (significant digits after the decimal point).

The structures in the database were coded by their *SMILES* representations. The program transforms these representations into connectivity tables, filling the free valences with H atoms, calculating the distance matrices and charge densities and the *CTIs* for each structure subsequently computed. Thus, a file of the *CTIs* and the corresponding structure *SMILES* codes of all database structures is formed.

The NCI database was used in its original form working with the raw SMILES file. We employed Release 1 of the NCI file where one can download the latest release [14] (Release 4 - includes additionally about 17 000 molecules and some other improvements of the database). NCI file contains diverse molecules with aromatic and non-aromatic cycles, hetero cycles, hetero atoms etc. This set is quite popular and used for multiple chemoinformatics tasks and QSAR modeling. The basic characteristics of the used NCI database in accordance to the structure sizes are summarized by the histogram in Figure 2. In our opinion, the NCI database is a diverse structure collection and the tests with it are indicative enough to show the real characteristic of the studied *CTI* index. Within the performed *CTI* tests we discovered different types of duplications in the NCI database. There are entries which are represented by practically equivalent SMILES codes utilizing two different ways to describe the implicit H atoms. For example, the molecule of *hexahydro-1-benzofuran-2-one* was found represented by two different (but equivalent) SMILES notations: O=C1C[CH]2CCCC[CH]2O1 and O=C1CC2CCCCC2O1, both structures producing *CTI*=274.29004386765. In this example “[CH]” and the corresponding “C” from the second notation are equivalent according to the SMILES syntax. Both locations of “[CH]” describe a

tertiary carbon with one H atom attached to it described as an attribute within brackets "[ ]". On the other hand, according to the standard syntax of SMILES, "C" (without brackets) is interpreted as a carbon which is automatically "filled" with needed number of implicit H atoms in order to comply the normal valence – in our example it is a tertiary carbon and thus it is filled with one H atom.

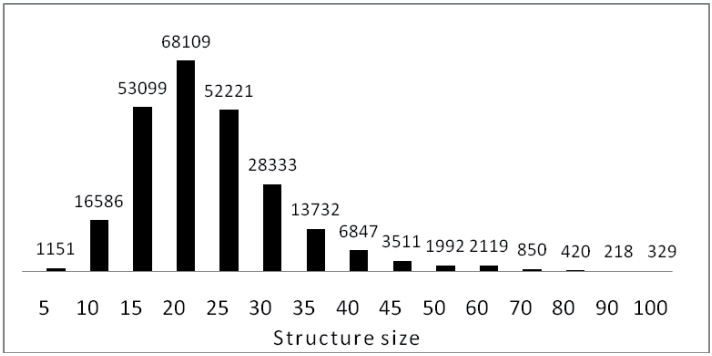
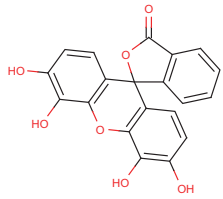
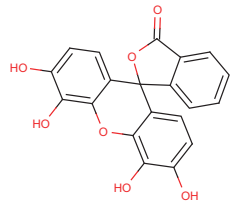


Figure 2: Structure size frequency distribution for the NCI database.

Furthermore, searching for duplicated *CTI* values, we found aromatic structures in the structure collection that have different Kekule presentations of their aromatic parts, thus their SMILES codes differ in the mutual double-single bond positions as shown in Figure 3.



CTI = 1406.95614249472
OC1=C(O)C2=C(C=C1)C3(OC(=O)C4=C3C=CC=C4)C5=C(O2)C(=C(O)C=C5)O



CTI = 1406.95614249472
OC1=CC=C2C(=C1O)OC3=C(O)C(=CC=C3C24OC(=O)C5=CC=CC=C45)O

Figure 3: Two equivalent aromatic structures producing the same *CTI* having different Kekule representation in the NCI file (corresponding SMILES codes are shown).

Additionally, some single atoms have been found in the database which were omitted from our consideration. The single atoms could be seen as separate records (structures) or as additional fragments to the basic structures. For the latter cases as well for the cases of fragmented structures we considered only the main (largest) fragment for those records. For example there are structure records which differ only by the presence of a disconnected ion:

Cl . NC1=CC(=CC=C1O)[As]=O

NC1=CC(=CC=C1O)[As]=O

As we mentioned, in principle we have worked with the basic fragment for each record but, if *CTI* was calculated for both example structures from the above SMILES codes, the same *CTI* would be obtained since the Chlorine is disconnected from the other fragment.

In order to study the discrimination power of *CTI* index we carried out several tests varying the number of used digits after the decimal point, the usage of *CTI* expression (1) with and without H atoms (but the H atoms being estimated within the charge calculation procedure) and varying the number of charge calculation iterations. The basic test procedure with the NCI database was performed as follows:

(i) CTI values for all NCI structures (for particular calculation parameters: digits, H atoms, charge iterations) are calculated and stored.

(ii) The result file with the *CTI* indices is analyzed and all duplication pairs of structures having same *CTI* values are stored.

(iii) The duplication pairs are analyzed in order to determine which duplications are real i.e. the cases where two chemically different structures have equal *CTI* values.

In order to perceive such structures, a mapping procedure was developed in our group and applied to the structures producing the same *CTI*, but having different SMILES linear notations. Additionally, the aromaticity information was processed in order to recognize equivalent structures in the NCI database which have different Kekule representations (see the example from Figure 3).

The results presented in Table 2 describe the number of pairs of structures having the same *CTI* values at different precisions (considered significant digits after the decimal point). One can see that there are no erroneous duplications when more than 7 digits are used for the case of included H atoms in the *CTI* calculations.

Table 2: Number of pairs of non-isomorphic structures producing the same *CTI* values (the tests were performed with and without H atoms included in the *CTI* calculation).

Number of used *CTI* digits after decimal point	**Number of pairs that produce same *CTI* values**	
	H atoms are included	***H atoms are not included***
1	1975982	2830004
2	197262	287559
3	19885	28951
4	2062	2876
5	209	302
6	21	32
7	3	5
8	0	1
9	0	0
10	0	0

The tests were carried out for two principle cases regarding the usage of H atoms. Table 2 shows that if H atoms are omitted in the calculation of *CTI*, then the discriminating power of *CTI* is slightly decreased (one additional digit is required). Also one can notice the strict exponential dependence of the number of duplications in respect to the number used digits. The logarithm of the number of duplications correlates linearly with the number of used digits with quite high correlation coefficient: R^2=0.9996 and R^2=0.9972 respectively for the cases with and without H atoms included in the calculations. The number of duplications decreases 10 times when additional digit is used. The latter results are logical since the number of possible values of the *CTI* index increases exactly 10 times when an additional digit is used and hence the chance for duplication is reduced 10 times. Without speculation we could extrapolate the results from Table 2 in the following manner:

- it could be expected that using *CTI* with 9 digits would produce save discrimination within a collection of at least 2 million structures (since 8 digits are enough to discriminate a collection of 250 000 and one additional digit gives a space for 10 time more structures);

- analogously it could be expected that using *CTI* with 10 digits would produce safe discrimination within a collection of 20 million structures. The latter result based on the

current tests is considered by us as a quite lower estimation of the limit of the capacity of *CTI* index since the potential different values of 10 digit *CTI* index are tens of billions.

Since the dawn of the mathematical chemistry development, the discrimination power of the newly presented indices is classically benchmarked against all alkane isomers. In this tradition we present the *CTI* performance results for various alkane isomers of different orders. Table 3 summarizes the results for the alkane isomers with 10 to 17 carbon atoms.

Table 3: Number of *CTI* duplications for various isomer sets of alkanes (*CTI* was calculated with 10 digits).

Number of carbons	Formula	Number of isomers	Number of *CTI* duplications
10	$C_{10}H_{22}$	75	0
11	$C_{11}H_{24}$	159	0
12	$C_{12}H_{26}$	355	0
13	$C_{13}H_{28}$	802	0
14	$C_{14}H_{30}$	1858	0
15	$C_{15}H_{32}$	4347	0
16	$C_{16}H_{34}$	10359	0
17	$C_{17}H_{36}$	24894	0

As it can be seen no duplications were obtained for any of the alkanes up to 17 carbon atoms. We also preformed tests with the smaller alkanes from 1 to 9 atoms (not presented in Table 3), where also no duplication was found. Accordingly, by extrapolating these results we could expect that *CTI* with 10 significant digits would not degenerate even for alkanes with 28 atoms or more considering the exponential rising of the number of isomers (C_{18}→60523, C_{19}→ 148284, C_{20}→ 366319,…,C_{28}→ 617105614) and the possible number of 10 billion variations of the 10 digit *CTI* index.

These results can be attributed to the fact we have found in early papers [6, 7] that *CTI* reproduces very well the branching of the chemical structures thus discriminating between different isomers.

The results from Table 4 show that number of charge iterations within the 10^{th} sign after decimal point precision (see eq. (4-7)) does not influence the *CTI* discrimination power. Thus, if needed, *CTI* can be used with smaller number of charge iterations when the issue of computation speed is addressed.

Table 4: Number of *CTI* duplications for the NCI database for different charge iterations.

Usage of H atoms	Charge Iterations	Number of *CTI* duplications
yes	3	0
yes	4	0
yes	5	0
yes	6	0
no	3	0
no	4	0
no	5	0
no	6	0

Conclusions

The results show that the *CTI* index can be safely used for a quick perception of the isomorphic (duplicated) structures in large databases as well as for very fast identity (full structure) search. *CTI* index having value up to the 10^{th} precision of its real number can be used in databases with millions of compounds. We also showed that *CTI* does not degenerate for alkanes containing 17 carbons.

Acknowledgements

This work has been supported by the National Fund for Scientific Research to the Bulgarian Ministry of Education and Science – Grant NFSR-IO1/7, which support is gratefully acknowledged.

References

[1] H. Wiener, Structural determination of paraffin boiling points, *J. Am. Chem. Soc*. **69** (1947) 17-20.

[2] M. Randić, Characterization of molecular branching, *J. Am. Chem. Soc.* **97** (1975) 6609-6615.

[3] I. Bangov, Computer-assisted structure generation from a gross formula. 3. Alleviation of the combinatorial problem, *J. Chem. Inf. Comput. Sci.* **30** (1990) 277-289.

[4] I. Bangov, Topological structure generators, in: J. Gasteiger (Ed.), *Handbook of Chemoinformatics*, Vol.1, Wiley, New York, 2003, pp. 178-194.

[5] N. Kochev, V. Monev, I. Bangov, Searching chemical structures, in: J. Gasteiger, T. Engel (Eds.), *Chemoinformatics: A Textbook*, Wiley, 2004, pp. 291-318.

[6] I. Bangov, M. Moskovkina, A. Patleeva, Charge-related molecular index (CMI), a novel descriptor for quantitative structure/property relationship (QSPR) models. I. General considerations, *Bulg. Chem. Commun.* **42** (2010) 338-342.

[7] P. Demirev, A. Dyulgerov, I. Bangov, CTI: A novel charge-related topological index with low degeneracy, *J. Math. Chem.* **8** (1991) 367-382.

[8] M. Moskovkina, I. Bangov, Employment of the charge-related molecular index (CMI) in chromatographic quantitative structure retention time relationships, *C. R. Acad. Bulg. Sci.* **65** (2012) 1199-1202.

[9] J. Gasteiger, M. Marsili, Iterative partial equalization of orbital electronegativity – a rapid access to atomic charges, *Tetrahedron* **36** (1980) 3219-3228.

[10] I. Bangov, Structure generation from a gross formula. 7. Graph isomorphism: A consequence of the vertex equivalence, *J. Chem. Inf. Comput. Sci*. **34** (1994) 318-324.

[11] W. Bremser, Hose - a novel substructure code, *Anal. Chim. Acta* **103** (1978) 355-365.

[12] http://www.daylight.com/dayhtml/doc/theory/theory.finger.html – Daylight fingerprint theory manual (Accessed August 3, 2013).

[13] I. Bangov, Use of the charge-related topological index for 13C NMR chemical shift predictions, *Ann. Univ. Sofia* **91** (2001) 103-113.

[14] NCI database files, Release 1 (October 1999), http://cactus.nci.nih.gov/download/nci/ (Accessed August 3, 2013).