JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Stereo Signature Molecular Descriptor

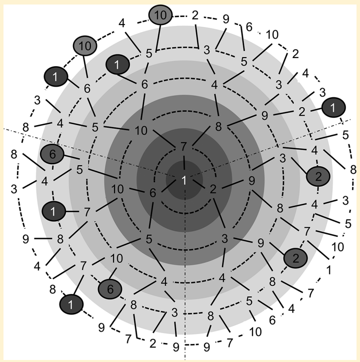
Pablo Carbonell,*[†] Lars Carlsson,[‡] and Jean-Loup Faulon[†]

[†]University of Evry, CNRS, Institute of Systems and Synthetic Biology, Évry, France

[‡]AstraZeneca Research and Development, SE-431 83 Mölndal, Sweden

Supporting Information

ABSTRACT: We present an algorithm to compute molecular graph descriptors considering the stereochemistry of the molecular structure based on our previously introduced signature molecular descriptor. The algorithm can generate two types of descriptors, one which is compliant with the Cahn–Ingold–Prelog priority rules, including complex stereochemistry structures such as fullerenes, and a computationally efficient one based on our previous definition of a directed acyclic graph that is augmented to a chiral molecular graph. The performance of the algorithm in terms of speed as a canonicalizer as well as in modeling and predicting bioactivity is evaluated, showing an overall better performance than other molecular descriptors, which is particularly relevant in modeling stereoselective biochemical reactions. The complete source code of the stereo signature molecular descriptor is available for download under an open-source license at http://molsig.sourceforge.net.



INTRODUCTION

It is well known that the ability to correctly distinguish enantiomers and diastereomers from their corresponding forms is important in various applications. One example is for drugs that can be a racemic mixture of two enantiomers where the biological activity may differ drastically between the two, especially when one of them may be toxic. Specific stereoisomers can be synthesized or separated from racemic mixtures so that a specific enantiomer or diastereomers can be run through different assays thus providing a growing amount of experimental data. In the early development of drugs it is common to use computational models to screen non-synthesized molecules for various structure–activity relationships to assess which molecules could be potential drug candidates. These models are often generated by relating entities describing the molecules, commonly denoted by descriptors, to experimental data. Hence, there is a need for descriptors that uniquely captures properties of chiral molecules.

Developing stereo descriptors is important because they allow for the extension of conventional quantitative structure–activity relationship (QSAR) models to data sets involving chiral molecules, i.e., where chemical or biological activity of molecules might be related to their chiral configuration. Stereochemistry information, thus, needs to be included in the descriptor representation in a way that allows discrimination of the observed activity. Because chirality is purely a topological feature, several structural descriptors that describe stereochemistry have been reported in the literature, see for example, Ortiz et al.[1] and Golbraikh et al.[2] A common approach for the inclusion of chiral information in two-dimensional topological descriptors is to correct atoms that represent chiral centers by assigning a value or an attribute depending on the underlying topological description. It is also possible to create three-dimensional conformations of the molecules to generate property-based descriptors.[3] Other approaches include incorporating physicochemical properties[4] that allow for more interpretable descriptors. A topological descriptor has been developed by one of the authors,[5] the signature descriptor, which canonically describes the connectivity of each atom in a molecule with its neighboring atoms in a tree-like fashion. The signature descriptor captures local compositions of atoms and their neighboring atoms, and this approach has been shown to be capable of covering the information that other non-topological descriptors contain.[6] Furthermore, the signature descriptor has also shown to be significantly relevant when interpreting QSAR predictions.[7] The signature descriptor was also the starting point for a chiral descriptor in Koichi et al.,[8] where the approach used by the authors was to compute first chirality prior to canonicalization. Assigning chirality to chiral centers, however, is not always a straightforwards process because it often involves traversing the neighboring atoms in order to detect mass differences. This is particularly noticeable for complex structures such as fullerenes, where several criteria might be possible to apply in order to determine chirality.[9]

Here, we take a different approach by including the determination of chirality as an integral component of the algorithm that computes our stereo signature molecular descriptor. More specifically, determination of chirality and of the signature descriptor are performed simultaneously through an iterative algorithm that can be made compliant with the Cahn–Ingold–Prelog or CIP priority rules.[10] The flowchart of this process is shown in Figure 1 and will be described in the

Received: December 8, 2012
Published: March 25, 2013

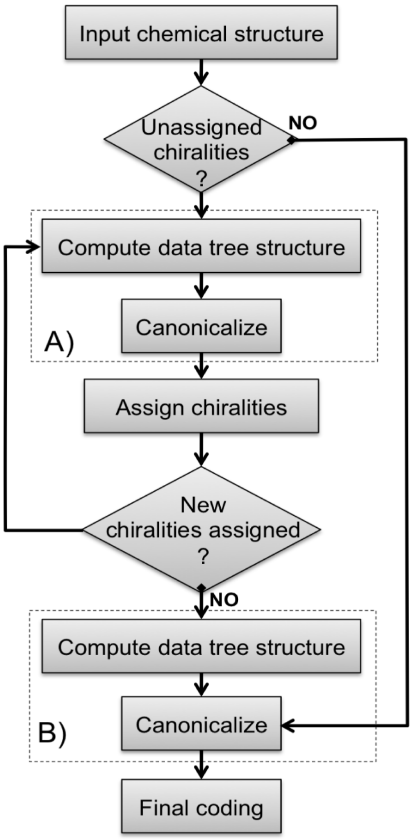


Figure 1. Flowchart of the signature canonicalization algorithm with stereochemistry. First, the algorithm reads the input structure and verifies its stereochemistry. If chirality needs to be assigned, the algorithm enters into a loop where at each iteration the tree data structure is computed and converted into its canonical form (procedure A) in order to assign parities. Once no more chirality can be assigned, the algorithm outputs the coding of the tree structure in its final canonical form (procedure B). Procedure A can be computed using the signature tree structure, following in that case the CIP rules, or using the signature DAG, which makes the algorithm computationally more efficient. Similarly, procedure B can use both data structures, although signature DAG is the default one because of its computational efficiency.

next sections, which is organized as follows. First, we define the tree data structures that are used in our algorithm. Next, the determination of chirality in the tree structure by the application of the CIP rules is explained, followed by its canonicalization. Results are then provided for execution times in comparison with the InChI code, assignment of chirality in fullerenes, and several examples are provided about the use of the stereo signature descriptor for structure−activity relationship modes and in the characterization of stereoselective biochemical reactions.

■ METHODS

Initial Data Structures. In this paper, we are considering two different data structures to represent atom neighborhoods. One data structure, the signature tree, is used to assign stereochemistry following the CIP rules. The second data structure, the signature DAG, is used to canonicalize the atom neighborhoods and can also be used to assign stereochemistry albeit not following necessarily the CIP rules. The definition of the signature tree and signature DAG given below can be found in the IUPAC recommendations[10] for signature trees and in Faulon et al.[5] for the signature DAG.

Let G be a molecular graph. and let x (b) be an atom (bond) of G. Initially, both tree data structures, i.e., signature trees and signature DAGs, rooted at atom x or bond b, respectively, are built from the molecular graph G. We use the terminology *atoms* and *bonds* when referring to the initial molecular graph, and the terms *vertices* and *edges* to describe the tree data structure, as several vertices in the structure may correspond to the same atom. Consider for instance the molecular graph for pseudotropine, an alkaloid stereoisomeric with tropine shown in Figure 2a. In Figure 2b, the tree for atom $x = 3$ is illustrated for both data structures. In both cases, the root of the tree is atom x itself. The first layer of the tree is composed of the neighbors of x; the second layer is composed of the neighbors of the vertices of the first layer except for atom x.

In the case of the **signature DAG**, the construction proceeds for the signature tree one layer at a time until no more layers can be added, that is, until all the bonds of G have been considered. Assuming that the tree has been constructed up to layer l, layer $l + 1$ is constructed considering each vertex y of layer l as follows: let z be a neighbor of y in G; vertex z and edge $[y,z]$ are added to layer $l + 1$ if the edges $[y,z]$ or $[z,y]$ are not already present in the previous layers of the tree. In the case of the **signature tree** vertex z and edge $[y,z]$ are added, in turn, to layer $l + 1$ if vertex y is present only one time in the path linking the root of z. In other words, in each branch of the signature tree atoms are allowed to be repeated no more than twice (CIP rule 1.b)[10]

Similarly, both tree data structures can be built for a bond b in a molecular graph G. The construction is identical to the one described above for atoms, although the root consists now of the two atoms forming the bond. In Figure 2c, the tree data structure rooted at bond $b = [3,4]$ of pseudotropine has been represented for signature DAG. Both data structures can be either defined up to the full depth spanned by the topology of the molecular graph or to a predefined diameter d related to the maximum layer of atom or bond neighborhood in the graph that is represented in the tree. By definition, even diameters $d = 2l_{\max}$ are used to describe data tree structures rooted at atoms containing a maximum layer $l_{\max} + 1$ of atom neighborhood, while odd values of diameters $d = 2l_{\max} + 1$ are used for trees rooted at bonds containing a maximum layer $l_{\max}$ of bond neighborhood.

We note that the size of the signature tree can grow exponentially because the number of branches is related to the number of paths in the molecular graph that are attached to the root. On the other hand, the size of the signature DAG, as shown in Faulon et al.[5] is no more than the number of bonds. Signature trees could, thus, lead to exponential complexity when computing stereochemistry. However, our results show that time difference when computing stereochemistry based on signature trees versus signature DAGs is rather small for molecules stored in cheminformatics databases.

Chirality Determination from Molecular Structure and Topology. We have implemented a procedure that will determine the stereochemistry of the molecular graph G based on the available information from the structure in a MOL file and the topology of the tree data structure. Two distinct classes of stereochemistry are represented, sp^2 (double bond) and sp^3 (tetrahedral). The sp^2 stereochemistry is determined for C, N, N+ atoms. Parity for stereo bonds is given by the sign of the dihedral angle between the planes determined by the double

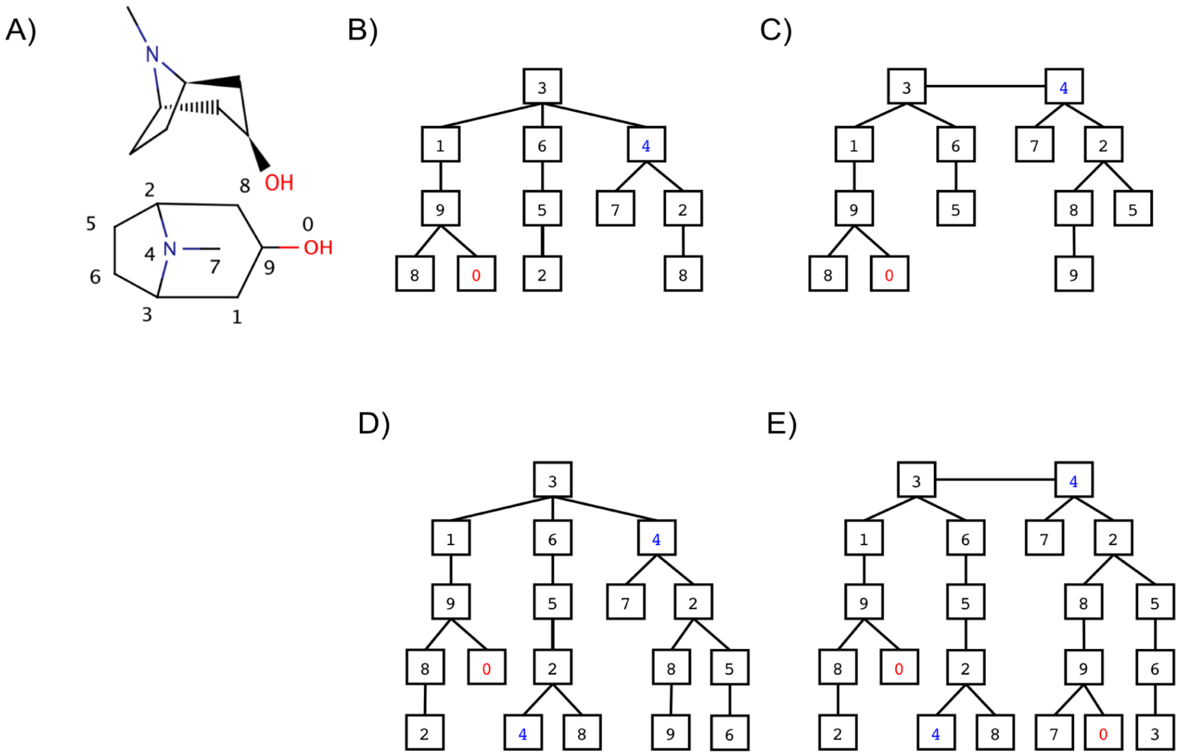


Figure 2. Tree data structures associated with pseudotropine. (A) Chemical structure of pseudotropine. (B) Signature DAG rooted at atom 3. (C) Signature DAG rooted at bond 3−4. (D) Signature tree rooted at atom 3. (E) Signature tree rooted at bond 3−4. In B and D, the signature DAG include all the bonds of the pseudotropine, and no layer needs to be added. In C and E, the signature tree is given up to layer 4. Note that up to layer 3 the signature tree and signature DAG are identical.

bond and the two heaviest substituents at both sides of the double bond. A positive angle corresponds to the *cis* configuration (Z), while a negative angle corresponds to the *trans* configuration (E). The sp^3 stereochemistry is considered for C, P, N+, S atoms with four different neighbors. The parity of a stereo center is given by the sign of the angle between the vector orthogonal to the plane that the stereo center forms with the two first heaviest substituents and the vector connecting the third substituent to the stereo center. A positive angle corresponds to the R configuration, while a negative angle to the S configuration. In case of coplanarity, the same operation is computed for the next combination as ordered by their weights. When no coordinates are supplied allowing the determination of stereochemistry, stereo information from the so-called 0D parity in the MOL file is used to assign parity to the atoms.

In order to univocally determine the parity of stereo centers and stereo bonds, priorities are to be assigned to each substituents from the heaviest to the lightest. Initially, stereo centers and stereo bonds are labeled as candidates. Next, through the iterative application of the well-known Cahn−Ingold−Prelog or CIP prioritization rules[10] to the signature tree data structure rooted at each center described in the previous section, parities are assigned. Such tree structures can either be defined up to the full depth spanned by the topology of the molecular graph or to a predefined diameter d where the maximum described layer of atom neighborhood is given by $d/2$ for even values or the maximum layer of bond neighborhood is given by $(d - 1)/2$ for odd values of d. Depending on the diameter d, chirality of a center might be lost if substituents become indistinguishable at such level of resolution. For instance, a minimum diameter of $d = 2$ is required for stereo centers and $d = 3$ for stereo bonds.

First, the geometry of the molecular structure is analyzed in order to flag all possible stereo center and stereo bond candidates. To assign priorities to each substituent of a stereo center or stereo bond candidate, the signature tree data structure described in previous section of the desired diameter d rooted either at the stereo center atom or at the sterobond corresponding to the candidate is computed. Priorities for each substituent are computed by comparing the branches that are children of the root in the signature tree data structure through the application of the CIP rules. The algorithm starts assigning an initial list of weights to each vertex consisting of the following terms in decreasing order of priority:

1. the atomic number Z;
2. the chirality of the atom, in case of stereo centers, where the highest priority is given to the R configuration, followed by the S configuration;
3. the maximum weight of the edge connecting the vertex to one of its parents in the tree data structure, where increasing precedence is assigned in the order of single, aromatic, double, and triple bonds, respectively;
4. the chirality of the edge, in case of stereo bonds, where the highest priority is given to the Z configuration, followed by the E configuration; and
5. the charge of the atom, in decreasing order of precedence.

Vertices at each layer are then compared based on their list of weights defined above. On the basis of this comparison, an initial invariant is assigned to each vertex so that the heaviest vertex receives the highest invariant number. Next, the algorithm enters into the iterative routine that applies the CIP rules in order to determine priorities of children. The

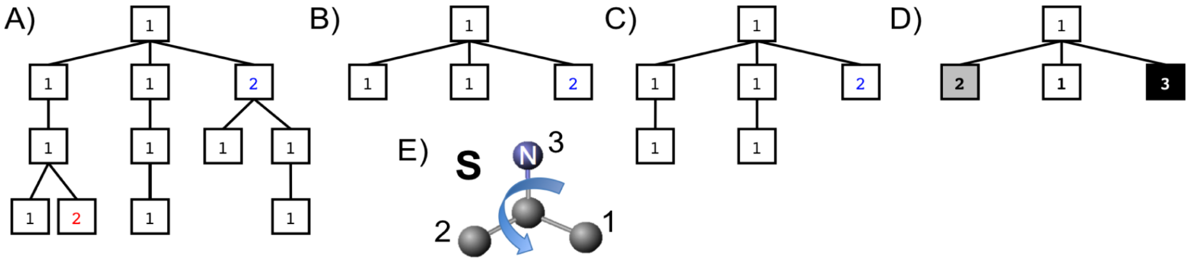


Figure 3. Determination of chirality for atom 3 in pseudotropine (Figure 2) by the application of the CIP rules. (A) Initial invariants assigned depending on the rules for atom type. (B) Prioritization of the substituents after looking at the first layer and the (C) the second layer. (D) Final assignment of priorities. (D) Resulting chiral configuration is *S*.

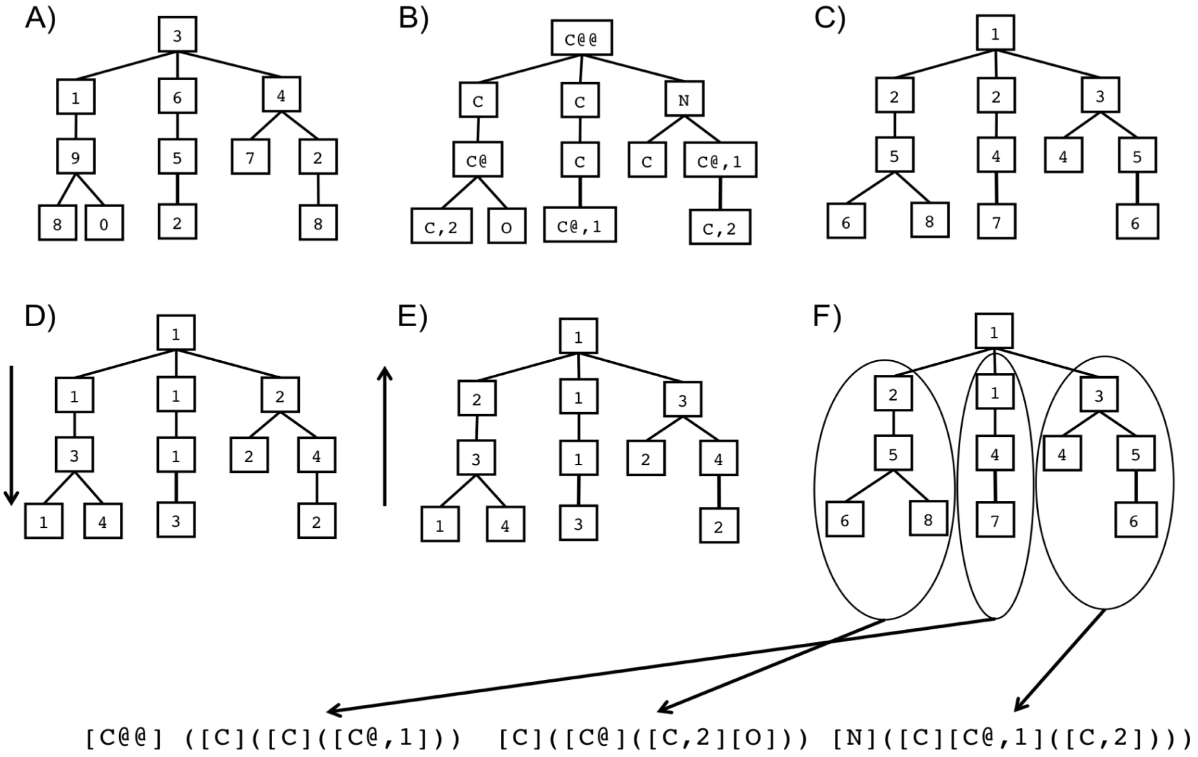

A)
B)
C)
D)
E)
F)
C@@
[C@@] ([C]([C]([C@,1])) [C]([C@]([C,2][O])) [N]([C][C@,1]([C,2])))

Figure 4. Signature canonicalization algorithm for atom 3 in pseudotropine (Figure 2). (A) Tree data structure rooted at atom 3. (B) Label substitution in the tree. (C) Assignment of initial invariants based on atom type, including chirality. (D) Vertex invariants after running the algorithm from root to leaves. (E) Vertex invariants after running the algorithm from leaves to root. (F) Final invariants after canonicalization sorted in order of precedence from right to left.

algorithm goes from parent to children, stopping when children have different configurations. Each pair of substituents is compared, and if they are equal, the algorithm descends to the next layer in the tree structure. The comparison between branches is based on the CIP rules (see detailed description in IUPAC recommendations[10]), i.e., substituents are sorted in decreasing order by the vertex invariants computed initially as described above. The branch containing the first vertex with the highest invariant is assigned a greater priority. If there is a tie, then if any of the vertices appeared before in a higher layer, then the one that appeared at the highest is greater. Otherwise, the algorithm proceeds to compare the children of the vertices at the next layer. At the end of this process, if new chirality was assigned but still some other chiral candidates remain to be assigned, the algorithm is repeated iteratively.

Consider for instance the determination of chirality for pseudotropine. In Figure 3 is shown the process followed in order to assign chirality to atom 3. Priorities are determined for the three heavy substituents in several steps. First, the signature tree data structure rooted at the stereo center candidate is developed. Next, initial priorities are assigned to each vertex. From here, the algorithm descends layer by layer while branches are equal. In the first layer, atom 4 already receives the highest priority, while priorities for the branches corresponding to atoms 1 and 6 are still undetermined. Two more iterations are actually necessary until the first difference between these two branches is found. Finally, priorities from heavier to lighter are 4, 1, 6, corresponding to a chirality *S* for atom 3. Similarly, the same procedure will determine that chirality for atom 2 is *R*. A different case is the stereo candidate atom 9. In this case, a first iteration will not assign chirality because the two branches of atoms 1 and 8 are symmetrical. The situation, however, changes at the second iteration because once the chirality has been assigned to atoms 2, 3, the symmetry does not hold anymore and chirality can be determined by the application of the CIP rules dealing with chiral centers. By these means, chirality is determined for atom 9 to be of *S* configuration.

Concerning stereo bonds, the procedure for chirality assignment is similar to the one described above, applied to the branches of the signature tree rooted at the stereo bond candidate. Once the priorities of the two substituents at each side of the bond have been determined, the *E* or *Z* configuration is assigned accordingly. As shown in Figure 1, through the application of the algorithm that assigns stereo centers and stereo bonds, chirality in the molecular graph are computed iteratively. At each iteration, the algorithm attempts to assign chirality to all the remaining unassigned stereo center and stereo bond candidates. If no new chiralities are possible to determine, the algorithm returns the molecular graph with the computed chiral annotations assigned to their corresponding atoms and bonds. As an illustration of how our algorithm can be used to assign stereochemistry, we provide as Supporting Information the assignments for the set of compounds in the DRUGBANK database.[11]

Canonicalization of Molecular Graphs. Once chirality has been fully determined for the molecular structure, the next step in order to compute the stereo signature descriptor of the molecule is to proceed to the determination of the canonical molecular graph associated with the structure. Our approach for canonicalization of the molecular graph is based on the classical Hopcroft and Tarjan's rooted-tree canonicalization algorithm.[12] The algorithm used here is an extension of the signature canonicalization algorithm proposed by one of us (Faulon et al.[5]). For any given molecule and any given diameter *d* greater than the one of the molecule, one can reconstruct in linear time the entire molecular graph from any of either its atomic or its bond signatures.[6] Thus, any algorithm canonicalizing a signature will also canonicalize a molecular graph as long as the signature's diameter is greater than the graph's diameter. When considering graph canonicalization, we should keep in mind that while graph canonicalization has not been shown in general to be intractable (NP-complete), no polynomial general solutions are known. Molecular graphs, however, belong to a restricted class of graphs to which the canonicalization is tractable and can be solved efficiently.[5] For instance, in addition to the signatures canonicalization algorithm here presented, the IUPAC project InChI provides also a canonicalizer of chemical structures.[13]

A detailed description of the process of signature canonicalization for a given atom in a molecular graph, which is illustrated in Figure 4 for a pseudotropine chiral atom, can be found in Faulon et al.[5] That algorithm is repeated for all atoms, and the resulting atomic signatures are compiled into a molecular signature. Because a molecular graph of maximum diameter *d* can be reconstructed from any of its atomic signatures, any atomic signature of at least diameter *d* suffices to represent the graph in a unique manner. Therefore, the algorithm offers the possibility of outputting a canonical code for the molecular graph, which is conventionally given for the signature that has fewer occurrences among the ones that are rooted at the highest atom in lexicographical order . For diameters lower than the maximum in the graph, the list of different atomic signatures and their occurrences are useful in order to model physical (QSPR) or biological activities (QSAR), as is illustrated through different examples in the Results and Discussion section.

We note that the canonicalization algorithm assigns invariants to all atoms of the signature DAG. These invariants can be also used to assign *R*/*S* and *E*/*Z* configurations, in a similar manner as it was done in the previous section. Indeed, invariants are computed based on CIP rules; however, the results might be different because the data structures are different. In particular, rule 1.b (which prevents any vertex to be repeated no more than twice in the branch linking to the root) is replaced in the signature DAG by a rule where bonds are not repeated if they already appeared at previous layers. We define computing chirality this way through the signature DAG as the *fast stereo signature mode* compared with the computation through the *stereo signature mode* using the signature tree, as described in previous section.

Output Notation. After canonicalization, the signature is written by reading the tree in a depth-first order sorted by the invariant of each subtree, parentheses are used to represent a parent−child '(' or a child−parent ')' relationship, and each vertex is represented by its atom type and a label if the vertex appears several times in the tree (see an example in Figure 4). Special symbols are used in order to denote stereochemistry. For double bonds (sp^2 stereochemistry), the /=\ symbol is used to indicate that the heavier atoms are on the same side (*E*) of the stereo center, while /=/ is employed when they appear at opposite sides (*Z*). For stereo center (sp^3 stereochemistry), the @@ label is used when the order of the neighbors seen from the lightest one is clockwise (*R*), whereas the @ label is used when they are in anticlockwise order (*S*).

Data Sets Used in the Performance Tests. *Execution Time Test.* To evaluate execution times, molecular structures of chemical compounds from the Fullerene Structure Library,[14] KEGG metabolic database[15] (total 14276 molecular structures), and the DRUGBANK database (total 6628 molecular structures)[11] were downloaded. We ran both the fast stereo and the stereo signature mode, as described in previous section, as well as the InChI chemical identifier[13] for the KEGG and DRUGBANK cases. Programs were run in 32 bits mode in a Linux box running at 2.93 GHz . Each program was called twice, and times were measured the second time in order to avoid cache differences when reading the same molecular structure file.

Prediction of Enantioselectivity. A series of 48 enantiomeric pairs of chiral amino alcohols that enantioselectively catalyze the addition of diethylzinc to benzaldehyde was taken from Zhang et al.[4] Each catalyst yields preferentially either *S* or *R*, while its opposite enantiomer produces the reverse. As in ref 4, the data set was split into a test set of 10 enantiomeric pairs and 38 pairs for training set. The second tested data set, also taken from Zhang et al.[4] was a series of 86 enantiomeric pairs of primary alcohols catalyzed by *Pseudomonas cepacia* lipase. Each enantiomer was identified as the preferred or non-preferred enantiomer of the reaction. The test set is formed by 14 enatiomeric pairs. Stereo signatures for each enantiomeric pair was computed for diameters 0, 2, 4, and 6 as descriptors. A support vector machine was trained using the kernlab R package[16] in order to build an enantioselectivity classifier.

Prediction of Ecdysteroids. Dinan et al.[17] present modeling results on Ecdysteroids. This is a set of 71 compounds that are active on the Ecdysone receptor, and Dinan et al. present experimental results for *Drosophila melanogaster* B_{II}. The measured activity is presented as the negative logarithm of ED_{50} values.

Here, a comparison is made between the non-stereo signatures and the stereo signatures by creating atom signature with diameters 0, 2, 4, and 6 as descriptors of the two types on the 71 Ecdysteroids used as a training set and the additional seven as a test set. The two different training sets were modeled

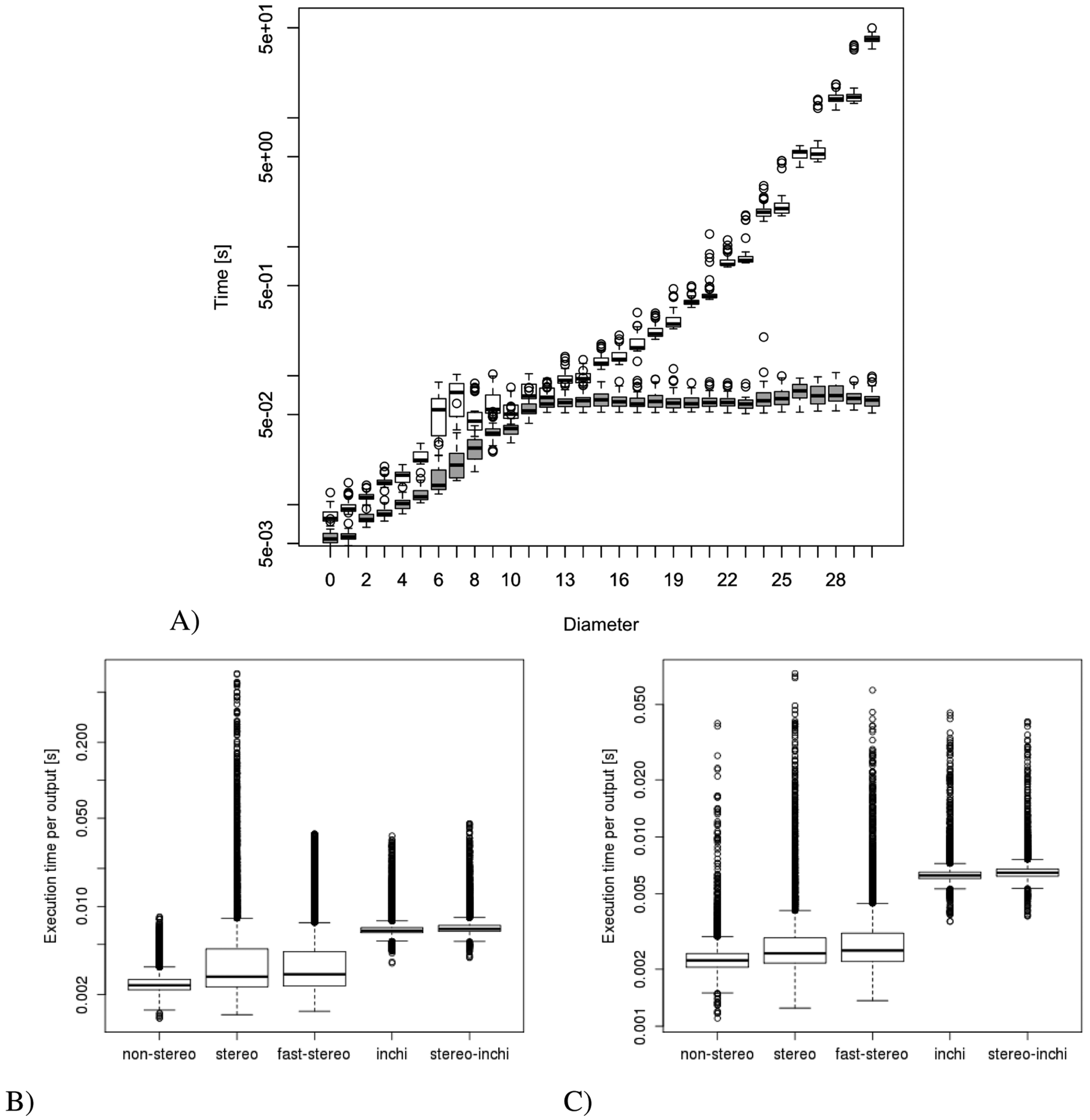


Figure 5. (A) Execution times of the stereo signature descriptor vs the fast stereo signature descriptor (solid boxes). (B) Molecular canonicalizer with non-stereo, stereo, and fast stereo signatures in comparison with the InChI canonicalizers with non-stereo and stereo descriptors for the set of metabolites in KEGG and (C) for the drug database DRUGBANK.

using ε-SVR and the RBF kernel in LIBSVM.[18] A leave-one-out cross validation was conducted to determine optimum values, with the objective to reduce mean square error, for the LIBSVM parameters C, γ, and p. These values were found using gridregression.py, within the LIBSVM package, which was used with its standard settings. The python script performs a grid search with the following ranges and step sizes: $C = 2^n, n = -1,...,6$, $\gamma = 2^n, n = -8,...,0$, and $p = 2^n, n = -8,...,-1$, where n is an integer. A model was built for each of the two training sets, and the test set was predicted by both corresponding models.

Prediction of Antimalarial Activity. Molecular structures and biological activity were downloaded from the authors of a QSAR model of bioactivity of artemisinin analogues.[19] Stereo signatures were computed using the stereo signature mode (see previous section). ECFPs were computed from Pipeline Pilot Academic License.[20] The QSAR model was developed by using package pls from the R statistical package.[21] The input feature vector to the support vector machine consisted of the set of all stereo signatures contained in the training set from diameter $d = 0$, up to predefined diameter (4 or 6).

Prediction of Enzymatic Reactions. The data set consisted of the list of unique metabolic reactions annotated in the KEGG database (release 50).[15] Reactions, with the general form $R{:}s_1S_1+...+s_nS_n \rightarrow p_1P_1+...+p_mP_m$, where s_i and p_j are the stoichiometric coefficients of substrates S_i and products P_j, were encoded by using the reaction signature descriptor,[22] as follows: (a) stereo and non-stereo signatures of substrates S_i and products P_i were computed using the stereo signature mode (see previous section), (b) a signature of the reaction is obtained as the net difference between the product and the substrate signatures multiplied by their respective stoichiometric coefficients, and (c) similarity between reactions was computed by the Tanimoto similarity coefficient between their signatures. For each enzyme family containing a significant amount of reactions (>20) with same non-chiral signatures in other families, we built and 10-fold cross-validated kernel-based predictors for enzymes sequences catalyzing specific reactions similar to the one proposed in Faulon et al.[23] Balanced test sets were formed by reactions in the selected enzyme family (positives) and reactions found in another enzyme families (negatives).

■ RESULTS AND DISCUSSION

Execution Time. Execution time for the stereo molecular signature descriptor was compared with those for the fast descriptor, as defined in the Methods section, for a set of structures with complex stereochemistry corresponding to 451 C_{100} fullerenes from the Fullerene Structure Library,[14] shown in Figure 5. For such structures, execution time increased significantly with the signature diameter for the stereo signature descriptor, while they remained essentially constant for the fast descriptor. This result shows the usefulness of the fast descriptor in cases of complex stereochemistry. However, as it is shown in our next test, in most of the common cases that might be found in applications such as metabolic or drug databases, the difference in execution time between both descriptors is not significant.

Here, we have compared the execution times between the signature canonicalizers described in the Methods section and the InChi chemical identifier[13] for two chemical databases: the set of metabolites in the KEGG database (total 14276 molecular structures) and the set of drugs in the DRUGBANK database (total 6628 molecular structures). The measured execution times in both databases are shown in Figure 5. Average times per output were of 2.79 ms in the stereo canonicalizer for KEGG compounds and of 2.43 ms in DRUGBANK, while the average execution time for the InChI program, including stereo, was of 6.63 ms per output for KEGG and 6.47 ms per output for DRUGBANK. Therefore the stereo signature canonicalizer was found to be almost 3 times faster than the InChI canonicalizer, a result that in part might be due to the fact that InChI performs a large amount of structure normalization, including tautomer/charge normalization.

Table 1. Average Time Per Output for InChI and Signature Canonicalizers

descriptor	KEGG	DRUGBANK
InChI	6.63 ms	6.47 ms
signature	2.79 ms	2.43 ms

CIP Assignment in Fullerenes. We considered the stereo assignment that performs our algorithm for the special cases of fullerenes and other carbon frameworks from the Fullerene Structure Database.[14] In these polyhedral carbon structures, chirality has to be often assigned based on rules other than the atomic mass precedence, such as the rule based on the presence of duplicated nodes in the graph or the rule for priorities assigned to chiral atoms. The examples given here compare the assignment of our algorithm with the ones performed in the work from Rassat et al.[9] In this study, authors proposed the application of a rule for priorities of branches of stereo centers based on the number of duplicated nodes. Branches are ranked as (L)arge, (M)edium, and (S)mall on the basis of duplicated nodes found at successive diameters in the graph rooted at the stereo center. The branch containing the lowest number of duplicated nodes is ranked Large, while the one containing the highest is ranked Small. This method differs notably from the IUPAC recommendations,[10] section P-91.1.1.2, rule 1.b, which is the one that we have adopted in our algorithm, where a duplicated atom, with its predecessor node having the same label closer to the root, ranks higher than a duplicated atom, with its predecessor node having the same label farther from the root, which ranks higher than any non-duplicated atom node.

For instance, we consider the C_2-symmetric $C_{10}H_{10}$ barettane shown in Figure 6. The assignment of chirality for node 1 is shown in Figure 7. The assignment is based on the presence of duplicated nodes. Root node 1 appears duplicated in both branches of vertex 5 and of vertex 6 at layer 4. According to the IUPAC rule, this duplicated node has the highest priority because its same predecessor corresponds to the root. Furthermore, branch 6 contains also the duplicated node 6 twice. Therefore, branch 6 gets the highest priority, while branch 7 is ranked second. Finally, branch 2, which contains node 2 duplicated twice at the same layer, gets the lowest priority. On the other hand, in the priority assignment method proposed by Rassat et al.[9] branch 7 receives the highest priority,

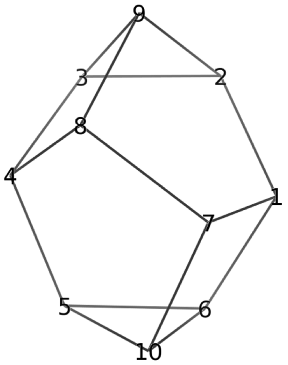


Figure 6. A C_2-symmetric spheroalkane.

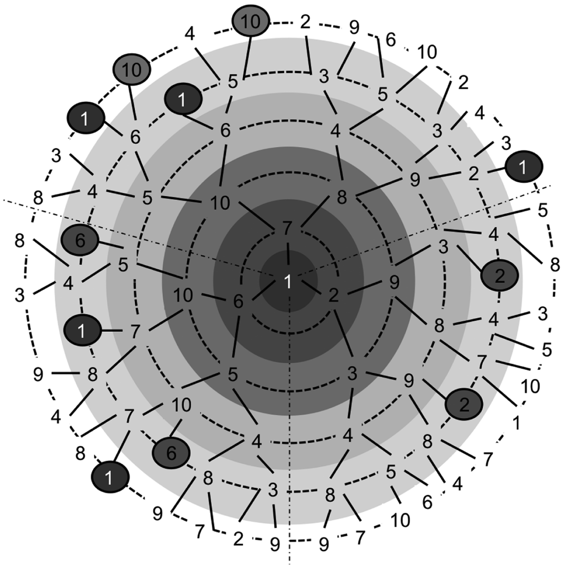


Figure 7. Rooted tree at vertex 1 of a C_2-symmetric spheroalkane (Figure 6). Lighter bands correspond to higher diameters, and duplicated vertices are highlighted.

Table 2. Duplicate Atoms Appearing in First Layers (up to 5) in Each Branch of Trees Rooted at Atoms of C_2-Symmetric $C_{10}H_{10}$ Spheroalkane of Figure 6[a]

atom	branch 1	branch 2	branch 3
1	2[4:2;5:1,9,3]	6[4:1,6;5:1,6,10,5]	7[4:1;5:1,7,10]
2	1[5:2,1,6]	3[3:2;4:3;5:2,3,9]	9[3:2;4:9;5:2,9,3]
3	2[3:3;4:2;5:3,9]	4[4:3;5:3,4,5]	9[3:3;4:3,9;5:9,2]
4	3[4:4,3;5:4,3,9,2]	5[4:5;5:4,10,6]	8[4:4;5:4,8,9]
5	10[3:5;4:10;5:5,10,6]	4[5:5,4,3]	6[3:5;4:6;5:5,6,10]
6	1[4:6;5:6,1,2]	10[3:6;4:6,10;5:10,5]	5[3:6;4:5;5:6,10]
7	1[4:7;5:7,1,2,6]	10[4:7,10;5:7,10,5,6]	8[5:7,8,9]
8	4[4:8;5:8,4,3,5]	7[5:8,7,10]	9[4:8,9;5:8,9,3,2]
9	2[3:9;4:2;5:9,3]	3[3:9;4:9,3;5:3,2]	8[4:9;5:9,8]
10	5[3:10;4:5;5:10,6]	6[3:10;4:10,6;5:6,5]	7[4:10;5:10,7]

[a] The first value in each branch column is the atom of this branch, whereas values between brackets provide duplicate atoms for each layer up to 5.

as it contains the lowest number of duplicated nodes in layer 4, branch 2 becomes second as it contains 2 duplicated nodes, while branch 6 gets the lowest priority, as it contains 3 duplicated node. According to this rule, thus, priorities are determined based on the number of duplicated nodes, independently of the distance of these duplicated nodes to their same predecessor node. The chiral descriptor, however, is the same in both cases (*S*) because priorities of the branches are related through a circular permutation.

In Table 3 the chiral assignment for the C_2-symmetric $C_{10}H_{10}$ barettane of Figure 6 is compared between the system based on the IUPAC rules and the system proposed in Rassat et al.[9] Generally, chiral descriptors are reversed in one case with respect to the other as we might expect because a branch containing a larger number of duplicate atoms has the highest priority according to the IUPAC rules but the lowest according to Rassat et al.[9] However, this is not the case for vertex 1, as well as for vertex 4, the two vertices of the hexagonal face that are *exo* to the triangular faces, because in that case the prevalent rule is the one that states that duplicate atoms being closer to the root have higher priority than the others. This rule, which has no counterpart in Rassat et al.[9] can create additional differences between both assignments. For instance, for atom 1 in Table 2, we see that atom 6 has the highest priority in the IUPAC method because it contains 2 duplicates at layer 4, while the rest only contain 1 duplicate at this layer. In the same fashion, the priority of this atom is the lowest in Rassat et al. (Table 3). However, the IUPAC rule will assign second priority to branch corresponding to atom 7 as its duplicate is the root. This aspect is not taken into account in the method from Rassat et al., which needs to go deeper into other layers in order to find the priorities between these two branches. This type of differences, which will appear only in cases where atoms belong to polygonal faces with low number of vertices, such as in the triangular faces, can easily create duplicates that are distant from the root and thus having lower priority than others in the IUPAC method.

As another example, we consider the chiral assignments for the D_3-symmetric C_{78} isolated-pentagon fullerene, whose Schlegel diagram is shown in Figure 8. Table 4 provides the comparison between the assignments given by our algorithm based on the IUPAC rules and the assignments based on the method in Rassat et al.[9] In this case, chirality assignments based on duplicated nodes according to the IUPAC rule are reversed by the rule based on the method proposed in Rassat et al.[9] In this case, as the topology of the structure is composed of either non-contiguous pentagons or hexagons, the rule in the IUPAC assignment based on the distance to the root for duplicates has no particular role. This is why the assignments for this molecular structure appear reversed in comparison with the ones in Rassat et al., as the prevalent rule here is based on the number of duplicates in each layer. This result, however, as we

Table 3. Assignment of Chiral Descriptors for C_2-Symmetric $C_{10}H_{10}$ Barettane According to IUPAC Rules (left) and Rules in Rassat et al.[9] (right)

	L	M	S		L	M	S	
1	6	7	2	S	7	2	6	S
2	9	3	1	S	1	3	9	R
3	9	2	4	R	4	2	9	S
4	3	8	5	S	8	5	3	S
5	10	6	4	S	4	6	10	R
6	10	5	1	R	1	5	10	S
7	10	1	8	R	8	1	10	S
8	9	4	7	R	7	4	9	S
9	3	2	8	S	8	2	3	R
10	6	5	7	S	7	5	6	R

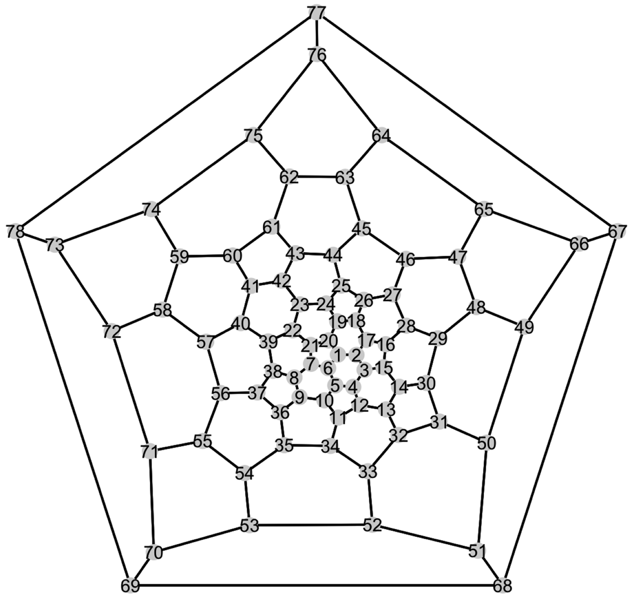


Figure 8. Schlegel diagram of a D_3-symmetric C_{78} isolated-pentagon fullerene.

Table 4. Assignment of Chiral Descriptors for D_3 C_{78} Fullerene According to IUPAC rules (left) and Rules in Rassat et al.[9] (right)[a]

	L	M	S		L	M	S	
1	2	10	6	S	6	10	2	R
2	1	12	3	R	3	12	1	S
7	6	21	8	R	8	21	6	S
9	8	26	10	R	10	26	8	S
10	1	11	9	S	9	11	1	R
11	12	10	28	R	28	10	12	S
22	21	23	39	S	39	23	21	R
23	24	42	22	S	22	42	24	R
24	8	25	23	S	23	25	8	R
25	24	26	44	S	44	26	24	R
26	9	25	27	R	27	25	9	S
27	26	28	46	S	46	28	26	R

[a] Chiral descriptors are given for one representative for each orbit at atomic sites.

have shown before, cannot be generalized because of the IUPAC rule of closeness of the atom node to the root.

Prediction of Enantioselectivity. Here, we investigated the ability of the stereo signature descriptor to classify chiral amino alcohols regarding their ability to enantioselectively catalyze the addition of diethylzinc to benzaldehyde. A similar predictor was previously developed by Zhang et al.[4] using physicochemical stereodescriptors and several machine-learning methods with a maximum accuracy of 96% in the training set and 90% in the test set (maximum overall accuracy of 96%). Using our stereo signature descriptor for diameter up to $d = 6$ and a support vector machine algorithm, we obtained 97% accuracy in the training set and 100% for the same test set as in Zhang et al.[4] (overall accuracy of 98%).

In yet another test, we tested the ability to predict the preferred enantiomer for primary alcohols involved in racemic resolutions of transesterifications or hydrolyses catalyzed by *Pseudomonas cepacia lipase*.[4] Using as before a classifier based on the stereo signature descriptor up to $d = 6$ and a support vector machine, we obtained an accuracy of 94% in the training set and of 86% in the test set (overall accuracy of 92%), which is higher than the maximum accuracy obtained by Zhang et al.[4] (83%, 93%, and 85%, respectively).

Prediction of Ecdysteroids. In this evaluation, the leave-one-out cross-validation procedure resulted in two models where the non-stereo signature model had a mean square error of 1.05 where the optimum parameters were $C = 2.0$, $\gamma = 2^{-8}$, and $p = 2^{-8}$. The corresponding values for the stereo signature model were 1.02 with $C = 32.0$, $\gamma = 2^{-7}$, and $p = 2^{-8}$. The mean square errors for the test set were 0.69 using the stereo signatures and 0.82 using the non-stereo signatures, whereas the mean square error obtained for the same set obtained in a model based on comparative molecular field analysis (CoMFA)[17] was of 1.78.

QSAR Model To Predict Biological Activity of Artemisinin Analogues. Development of accurate Quantitative Structure−Activity Relationship (QSAR) models based on the stereo signature descriptor of chemical compounds is another application of interest. It can be used to screen for the activity of interest for derivatives of a known drug. This application was used in one study to develop a QSAR model for analogues of artemisinin, a drug with significant antimalarial activity.[19] In this work, authors developed the model through PLS analysis through the ADPT (Automated Data Analysis and Pattern Recognition Toolkit) methodology by using a combination of geometric and topological descriptors. The maximum performance of the model was found to be $Q^2 = 0.687$.

Starting from the same data set and using the PLS methodology (see Methods), we obtained a performance of $Q^2 = 0.79$ using the extended connectivity fingerprints of both diameter $d = 4$ and 6 (Pipeline Pilot), whereas using stereo signatures we obtained a $Q^2 = 0.80$ and $Q^2 = 0.81$ for diameters $d = 4$ and $d = 6$, respectively (Table 5). These results, thus, illustrate how molecular signature descriptors can be used advantageously in order to develop QSAR models.

Table 5. Summary of Best Performances Obtained on QSAR Models Using ECFP and Stereo Signature Predictors of Diameters $d = 4$ and $d = 6$

descriptor	Q^2	R^2
ECFP ($d = 4$)	0.79	0.98
ECFP ($d = 6$)	0.79	0.99
Stereo signature ($d = 4$)	0.80	0.96
Stereo signature ($d = 6$)	0.81	0.98
ADPT	0.69	0.69

Prediction of Enzymatic Reactions. Chirality of substrates plays a crucial role in many enzymatic reactions as determining specificity. In order to characterize enzyme function, the use of molecular signature descriptors that contain chiral information, such as the ones here introduced, should facilitate the development of an accurate characterization of biochemical transformations, in particular for chiral specificity. To that end, we have already shown how reactions can be appropriately described using the molecular signature descriptor by the net difference between signatures of products and substrates in their respective stoichiometric proportions.[24] In order to illustrate how the use of chiral descriptors can increase the accuracy of the model, we screened metabolic network databases to look for enzyme families where chiral specificity is a distinctive feature. In particular, we were interested in considering those cases where a pair of reactions belonging to different enzyme families but with indistinguishable non-chiral signature description become significantly different once chiral descriptors are considered. In that case, chiral signatures are acting as discriminants, and thus, they might be an essential component enabling the development of accurate classifiers for enzymatic function.

In order to measure differences in chirality between two reactions, we have defined a similarity index for the signature descriptor, as described in Methods. After performing a screening of the entire enzyme families annotated in the KEGG metabolic database, we observed at least three families (at the third digit level of description of the EC classification) containing a significant amount of reactions showing a high level of similarity with other reactions outside of their own family (Table 6). Therefore, these enzymes might be potentially misclassified by the use of reaction descriptors that do not take into account chirality. To evaluate the general improvement in enzyme function classification obtained by the use of the chiral signatures, a support vector machine-based predictor was built for each of the enzyme families as given by the EC classification, as shown in Figure 9. The accuracy of the

Table 6. Average Accuracy Obtained by Cross-Validation in Enzyme Function Classifiers Trained with and without Chiral Descriptors

enzyme family	function	accuracy (non-chiral)	accuracy (chiral)
EC 1.1.1	oxidoreductases with NAD$^+$ or NADP$^+$ as acceptor	53.17%	93.17%
EC 2.4.1	glycosyltranferases acting on hexosyltransferases	48.88%	86.36%
EC 2.7.1	phosphotransferases with an alcohol group as acceptor	62.75%	92.85%

predictor based on descriptors with or without chirality was then tested through cross-validation for a diameter d = 6.

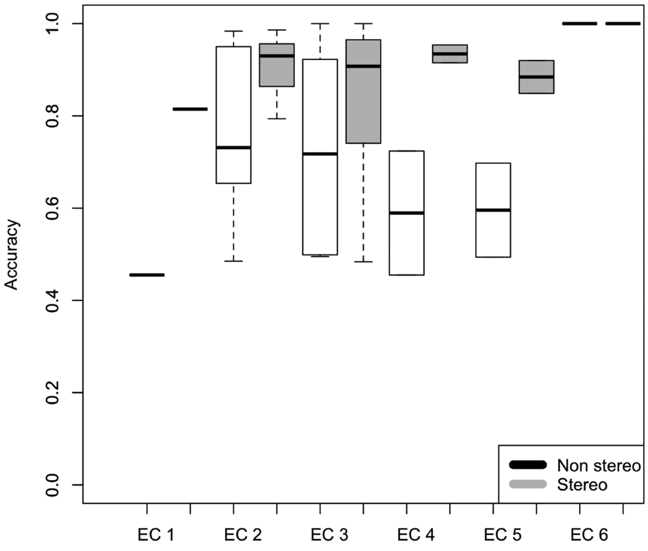


Figure 9. Accuracies obtained for enzymatic reaction classifiers using signatures without (left) and with (right) stereo descriptors.

CONCLUSIONS

We have introduced here a fast molecular graph-based algorithm to compute a topological descriptor that includes stereo chemical information, the stereo signature molecular descriptor, which is compliant with the CIP rules definitions that cover most of the important cases for practical use. To that end, two different structures from the molecular graph are defined, a signature tree in order to apply the CIP rules and the signature DAG, which is used to canonicalize the atom neighborhoods. In addition, signature DAGs can be used to obtain another canonical representation of the chiral molecular graph that allows a faster algorithm implementation for canonicalization. This representation, which does not necessarily follows the CIP rules, is in turn especially appropriate to use when dealing with structures with complex stereochemistry such as fullerenes.

We have shown that our descriptor has an excellent performance, being on average 3 times faster than the InChI canonicalizer. The descriptor showed high accuracy as a classifier in enantioselective catalysis, while prediction of bioactivity is at least at the same levels as the ones obtained by using Extended Connectivity Fingerprint Descriptors. In particular, the specificity of the descriptor as a classifier when using stereo information has been shown to improve significantly in the case of modeling enzymatic reactions, where enzyme stereoselectivity is often present. For the congeneric set of molecules represented by the ecdysteroids, the predictivity of models including stereo information was also improved compared to omitting this information. Such good performance can be explained because the assignment of chiralities is performed as an integral part of the graph-based algorithm that computes the stereo signature descriptor, a type of descriptor explicitly designed to capture molecular features relevant to activity.[23]

It could be difficult to set optimum diameters for the atom signatures; however, using the diameters 0, 2, 4, and 6, all combined as descriptors has been shown to give good results.[25] Also, internal model predictivity at AstraZeneca shows very good performance when signatures have been used with and without stereo information. Unfortunately, the data cannot be shown here, but this statement and the results presented here should be viewed as an indication that describing molecules with signatures, both with and without stereo information, can potentially be very useful for various modeling activities. The stereo signature predictor, thus, provides an improved version of the molecular signature predictor, whose usefulness has been highlighted for several applications. The complete source code of the stereo signature molecular descriptor is available for download under an open-source license at http://molsig.sourceforge.net.

ASSOCIATED CONTENT

Ⓢ Supporting Information

Assignment of stereochemistry by the algorithm for the DRUGBANK dataset; datasets used for prediction of enantioselectivity, ecdysteroids, antimalarial activity, and enzymatic reactions; predicted values for enantioselectivity, ecdysteroids, antimalarial activity; and obtained accuracies for enzymatic reactions. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: pablo.carbonell@issb.genopole.fr.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Funding: Genopole through an ATIGE grant; ANR through a Chair of Excellence.

REFERENCES

(1) de Julian-Ortiz, J.; Garcia-Domenech, R.; Gálvez Alvarez, J.; Soler Roca, R.; Garca-March, F.; Antón-Fos, G. Use of topological descriptors in chromatographic chiral separations. *J. Chromatogr., A* **1996**, *719*, 37−44.

(2) Golbraikh, A.; Tropsha, A. QSAR modeling using chirality descriptors derived from molecular topology. *J. Chem. Inf. Comput. Sci.* **2002**, *43*, 144−154.

(3) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(4) Zhang, Q.; Aires-de Sousa, J. Physicochemical stereodescriptors of atomic chiral centers. *J. Chem. Inf. Model.* **2006**, *46*, 2278−2287.

(5) Faulon, J.-L.; Collins, M. J.; Carr, R. D. The signature molecular descriptor. 4. Canonizing molecules using extended valence sequences. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 427−436.

(6) Faulon, J.-L.; Visco, D. P.; Pophale, R. S. The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 707−720.

(7) Carlsson, L.; Helgee, E. A.; Boyer, S. Interpretation of nonlinear QSAR models applied to Ames mutagenicity data. *J. Chem. Inf. Model.* **2009**, *49*, 2551−2558.

(8) Koichi, S.; Iwata, S.; Uno, T.; Koshino, H.; Satoh, H. Algorithm for advanced canonical coding of planar chemical structures that considers stereochemical and symmetric information. *J. Chem. Inf. Model.* **2007**, *47*, 1734−1746.

(9) Rassat, A.; Fowler, P. W.; de La Vaissière, B. Cahn−Ingold−Prelog descriptors of absolute configuration for carbon cages. *Chemistry* **2001**, *7*, 3985−3991.

(10) *Nomenclature of Organic Chemistry, 2005 ed. (Provisional Recommendations)*; International Union of Pure and Applied Chemistry (IUPAC): Research Triangle Park, NC, 2005.

(11) Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C. C.; Wishart, D. S. DrugBank 3.0: A comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* **2011**, *39*, D1035−D1041.

(12) Hopcroft, J. E.; Tarjan, R. E. Isomorphism of Planar Graphs. In *Complexity of Computer Computations*; Miller, R. E., Thatcher, J. W., Eds.; Plenum Press: New York, 1972; pp 131−152.

(13) Stein, S. E.; Heller, S. R.; Tchekhovskoi, D. An open standard for chemical structure representation: The IUPAC chemical identifier. *Proc. 2003 Int. Chem. Inf. Conf.* **2003**, 131−143.

(14) The Fullerene Structure Database. http://www.jcrystal.com/ (accessed December 1, 2012).

(15) Kanehisa, M.; Araki, M.; Goto, S.; Hattori, M.; Hirakawa, M.; Itoh, M.; Katayama, T.; Kawashima, S.; Okuda, S.; Tokimatsu, T.; Yamanishi, Y. KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **2008**, *36*, D480−484.

(16) Karatzoglou, A.; Smola, A.; Hornik, K.; Zeileis, A. kernlab - An S4 Package for Kernel Methods in R. *J. Stat. Soft.* **2004**, *11*, 1−20.

(17) Dinan, L.; Hormann, R.; Fujimoto, T. An extensive ecdysteroid CoMFA. *J.Comput.-Aided Mol. Des.* **1999**, *13*, 185−207.

(18) Chang, C.-C.; Lin, C.-J. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, **2011**, *2*, 27:1−27:27. http://www.csie.ntu.edu.tw/~cjlin/libsvm/ (accessed April 1, 2013).

(19) Guha, R.; Jurs, P. C. Development of QSAR models to predict and interpret the biological activity of artemisinin analogues. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1440−1449.

(20) Pipeline Pilot. http://accelrys.com/products/pipeline-pilot (accessed September 17, 2012).

(21) Mevik, B.-H.; Wehrens, R. The pls package: Principal component and partial least squares regression in R. *J. Stat. Soft.* **2007**, *18*, 1−24.

(22) Carbonell, P.; Faulon, J.-L. L. Molecular signatures-based prediction of enzyme promiscuity. *Bioinformatics* **2010**, *26*, 2012−2019.

(23) Faulon, J.-L.; Misra, M.; Martin, S.; Sale, K.; Sapra, R. Genome scale enzyme metabolite and drug target interaction predictions using the signature molecular descriptor. *Bioinformatics* **2008**, *24*, 225−233.

(24) Faulon, J.-L.; Carbonell, P. Reaction Network Generation. In *Handbook of Chemoinformatics Algorithms*, 1st ed.; Faulon, J.-L., Bender, A., Eds.; Chapman and Hall/CRC: Boca Raton, FL, 2010; pp 317−342.

(25) Norinder, U.; Ek, M. E. QSAR investigation of NaV1.7 active compounds using the SVM/Signature approach and the Bioclipse Modeling platform. *Bioorg. Med. Chem. Lett.* **2013**, *23*, 361−263.