

The Signature Molecular Descriptor. 2. Enumerating Molecules from Their Extended Valence Sequences

Jean-Loup Faulon* and Carla J. Churchwell

Sandia National Laboratories, P.O. Box 969, MS 9951, Livermore, California 94551

Donald P. Visco, Jr.

Department of Chemical Engineering, Tennessee Technological University, Box 5013, Cookeville, Tennessee 38502

Received October 29, 2002

We present a new algorithm that enumerates molecular structures matching a predefined extended valence sequence or signature. The algorithm can construct molecular structures composed of about 50 non-hydrogen atoms in CPU seconds time scale. The algorithm is run to produce all molecular structures matching the binding affinities (IC_{50}) of some HIV-1 protease inhibitors. The algorithm is also used to compute the degeneracy, or the number of molecular structures, corresponding to a given signature. Signature degeneracy is systematically studied for varying signature heights on four molecular series, alkanes, alcohols, fullerene-type structures, and peptides. Signature degeneracy is compared with similar results obtained with popular topological indices (TIs). As a general rule, we find that signature degeneracy decreases as the signature height increases. We also find that alkanes, alcohols, and fullerene-type structures comprising n non-hydrogen atoms are uniquely characterized by signatures of height $n/4$, while peptides up to 4000 amino acids can be singled out with signatures of heights as small as 2 and 3.

INTRODUCTION

For motivation of this study, we recall the comments of A. Balaban that “Topological Indices (TIs) should show high correlation ability and at the same time should have low degeneracy. A TI which would be completely nondegenerate would be a molecular code, or molecular ID number, and, in principle, one should be able to reconstruct the structure from this code. So far, for a general problem of any graph, this is still an open problem.”¹ While we have already established that signatures such as other fragment-based descriptors have correlation abilities for biological activities and physical properties,² we present in this paper an algorithm that reconstructs molecular structures from their signatures. We then use this algorithm to study the degeneracy of signature.

Reconstructing molecules matching molecular descriptors or TIs is a long-standing problem. Surprisingly, there are not many reports in the literature providing answers to the question. Most of the proposed techniques are stochastic in nature and either use Genetic Algorithm or Monte Carlo methods to search and construct chemical structures matching predefined descriptor values. While Venkatasubramanian et al.³ and Sheridan et al.⁴ were the first to propose stochastic techniques based on Genetic Algorithms, methods based on Monte Carlo were reported later.^{5,6} Whereas other papers using stochastic techniques have appeared since then, there are still very few attempts to solve the reconstruction problem using a deterministic approach or, in other words, techniques that generate exhaustive lists of molecular structures matching predefined descriptor values. In a series of three papers

Kier, Hall, and co-workers^{7–9} reconstruct molecular structures from the count of paths, 1P , up to length $l = 3$. Their technique essentially computes all the possible degree sequences matching the count of paths up to length $l = 2$. Then for each degree sequence all the molecular structures are generated using an isomer generator, and the graphs that do not match the 3P count are rejected. Skvortsova et al.¹⁰ use a similar technique, but from the count of paths they derive an edge sequence, in addition to the degree sequence. An edge sequence counts the number of edges between each distinct pairs of atom degrees. The two sequences are then fed to an isomer generator that produces all the structures matching the sequences. Regrettably, the authors do not provide details on how the isomer generator deals with the edge sequence. In our previous paper,² we introduced the concept of signature which is a canonical representation of all atoms environments up to a predefined height h . Note that from any height-0 signature it is possible to derive a degree sequence, and from any height-1 signature an edge sequence can be compiled. Instead of computing all degree sequences and all edge sequences matching path counts, one can compute all height-1 signatures, or extended degree sequences, matching the same counts. Thus, the algorithm we propose here can be used to solve the path problem up to length 3. Furthermore, the algorithm can go beyond paths of length 3 as it enumerates all molecules matching signatures of arbitrary heights.

This paper is set up as follows. The first section describes an algorithm that enumerates all molecular structures corresponding to a given signature. We then prove this algorithm is correct, meaning the algorithm produces the set of all possible structures and there are no duplicates in the solution

* Corresponding author phone: (925)294-1279; fax: (925)294-3020; e-mail: jfaulon@sandia.gov.

set. The next section analyzes the computational complexity of the algorithm. A simple application of the algorithm is presented where we enumerate all compounds having the same signatures as some experimentally active HIV-1 protease inhibitors. The last section systematically studies the degeneracy of signature as we vary the signature's height. The study is carried out on four homogeneous series of compounds, alkanes, alcohols, fullerene-type structures, and peptides. We compare our results to the degeneracy of other popular TIs. Finally, we conclude by revisiting Balaban's comments from the Introduction.

ENUMERATION ALGORITHM

The algorithm that enumerates all molecular graphs corresponding to a target signature is based on an isomer enumeration algorithm published some time ago.¹¹ Starting with a molecular graph containing all atoms but no bonds, the algorithm belongs to the class of orderly algorithms,¹² where bonds are added in all possible ways in order to produce all nonisomorphic saturated graphs matching the target signature. The original idea of this algorithm is to saturate all equivalent atoms at once. Specifically, the atoms are first partitioned into equivalent classes, the classes being the orbits of the automorphism group of the graph, then an orbit is chosen, and all the possible graphs that can be generated by saturating all atoms of the orbit are generated. The initial atom partitioning is performed following the target signatures. The process is recursive until all orbits have been saturated. The algorithm is given next and illustrated in Figure 1.

```

enumerate-molecule-signature( $G, {}^h\tau, S$ )
Input:  $G$  a molecular graph
          ${}^h\tau$  the target molecular signature
Output:  $S$  the set of solution graphs

if  $G$  is connected and  ${}^h\sigma(G) = {}^h\tau$  then return  $S \cup \{G\}$ 
partition the atoms of  $G$  into orbits
let  $o$  be an orbit of  $G$  s.t.  $\text{valence}(o) - \text{degree}(o) \neq 0$ 
 $S_c = \text{enumerate-orbit-signature}(o, G, {}^h\tau, \emptyset)$ 
For all graphs  $G$  of  $S_c$  do
     $S = S \cup \text{enumerate-molecule-signature}(G, {}^h\tau, S)$ 
done
return  $S$ 

```

When called with an initial graph, G , composed of isolated atoms, a target signature, ${}^h\tau$, and a set $S = \emptyset$, the above algorithm returns all the molecular graphs that can be constructed from G and ${}^h\tau$. The two main steps are to compute the orbits of the automorphism group of G and to saturate all the atoms of a chosen orbit. Recall that the orbits of the automorphism group are the atoms equivalent classes. Two atoms are in the same class if they have the same target signature and the same height $D+1$ signature in G , D being the diameter of G . Indeed, according to Proposition 6 in the first paper of this series, two atoms having the same height $D+1$ signature are automorphic and belong to the same class.² Hence, the orbits are calculated by computing and comparing the signatures of all atoms of the graph being constructed, which can be performed efficiently according to Proposition 7 of our previous paper.² Note that since the initial graph has no edges, to each atom, one associates a target signature compatible with the atom type, and all atoms having the same target signature are merged into one orbit. Once the atoms have been partitioned, an orbit containing unsaturated vertices is chosen and saturated at once using

the scheme given next. This scheme is illustrated in Figure 2.

```

saturate-orbit-signature( $o, G, {}^h\tau, S$ )
Input:  $o$  an orbit of atoms
          $G$  a molecular graph
          ${}^h\tau$  the target molecular signature
Output:  $S$  the set of graphs saturating orbit  $o$ 

if  $o = \emptyset$  then return  $S \cup \{G\}$ 
let  $x$  be the first atom of  $o$ 
 $S_v = \text{saturate-atom-signature}(x, G, {}^h\tau, \emptyset)$ 
 $o = o - \{x\}$ 
For all graphs  $G$  in  $S_v$  do
     $S = S \cup \text{saturate-orbit-signature}(o, G, {}^h\tau, S)$ 
done
return  $S$ 

```

The above scheme is straightforward. Assuming the atoms of orbit o are sorted in increasing label order, one chooses and removes the first atom of o , and then computes all the graphs generated by saturating the chosen atom. The scheme is recursive for all graphs saturating the chosen atom, until o is empty. Below is an algorithm enumerating all graphs saturating an atom; the algorithm is illustrated in Figure 3.

```

saturate-atom-signature( $x, G, {}^h\tau, S$ )
Input:  $x$  a atom
          $G$  a molecular graph,  $E(G)$  is the set of bonds of  $G$ 
          ${}^h\tau$  the target molecular signature
Output:  $S$  the set of graphs saturating atom  $x$ 

 $d = \text{valence}(x) - \text{degree}(x)$ 
if  $d = 0$  then return  $S \cup \{G\}$ 
For all atoms  $y$  of  $G$  s.t.  $\text{valence}(y) - \text{degree}(y) \neq 0$  do
     $E(G) = E(G) \cup \{x, y\}$ 
    if  $\text{compatible-bond-signature}(x, y, G, {}^h\tau)$ 
       and  $\text{compatible-bond-signature}(y, x, G, {}^h\tau)$ 
       and  $G$  is canonical
       and  $G$  does not contain a saturated subgraph
    then
         $S = S \cup \text{saturate-atom-signature}(x, G, {}^h\tau, S)$ 
    fi
     $E(G) = E(G) - \{x, y\}$ 
done
return  $S$ 

```

The previous scheme repeats itself until the given atom x is fully saturated. Basically all bonds between atom x and all unsaturated atoms y of the graph are tested. Bond $[x, y]$ is accepted if it is compatible with the target signatures of x and y , provided the resulting graph is canonical and the bond does not create a saturated subgraph. Verifying that a saturated subgraph has been created consists of finding the monoconnected components of a graph and can be performed in linear time.¹³ Checking for canonicity is a common procedure of orderly enumeration algorithms,¹² the procedure guarantees that the graphs generated are nonisomorphic. Note that molecules being bounded valence graphs, canonization can be implemented in polynomial time.¹⁴ To verify that a graph is canonical, one labels the vertices of the graph in all possible ways. The graph is canonical if the initial labeling leads to a list of edges that is lexicographically smaller than the lists obtained with all other labelings. Note that one does not have to test every possible label for a given vertex but only test the labels of its orbit. Indeed, two atoms belonging to different orbits cannot exchange labels since they are not automorphic. Examples of verification for canonicity are given in Figure 3. In the present paper, we have implemented two algorithms to verify canonicity, Tarjan tree canonization algorithm¹⁵ if the tested graph is acycle and McKay's Nauty technique otherwise.¹⁶ With both algorithms it is possible to limit label exchange within atoms' orbits.

Aside from canonicity one also must verify that the added bond $[x, y]$ is compatible with the target signatures of x and y . Examples of compatible and incompatible bonds are given

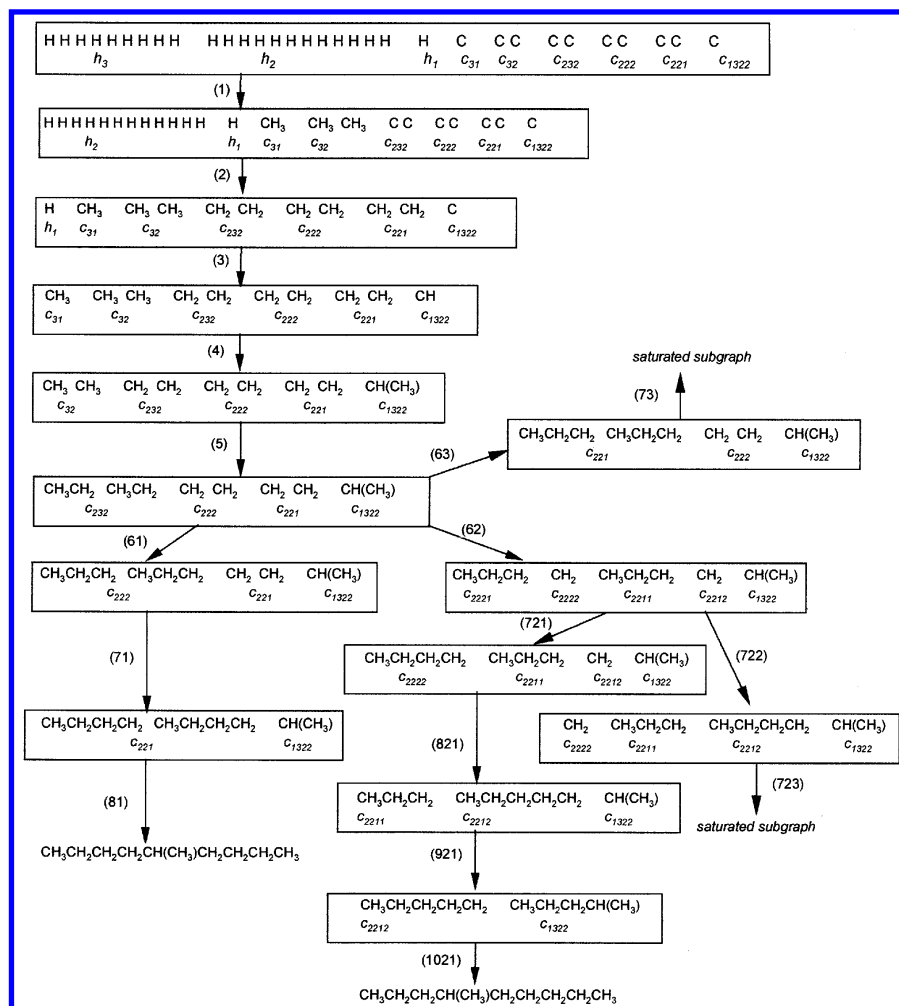


Figure 1. Graph signature enumeration algorithm. The figure depicts the enumeration of all $C_{10}H_{22}$ compounds matching the height-2 signature: $9h_3 + 12h_2 + h_1 + c_{31} + 2c_{32} + 2c_{232} + 2c_{222} + 2c_{221} + c_{1322}$, where $h_3 = H(C(HHC))$, $h_2 = H(C(HCC))$, $h_1 = H(C(CCC))$, $c_{31} = C(HHHC(HCC))$, $c_{32} = C(HHHC(HHC))$, $c_{232} = C(HHC(HHH)C(HHC))$, $c_{222} = C(HHC(HHC)C(HHC))$, $c_{221} = C(HHC(HHC)C(HCC))$, $c_{1322} = C(HC(HHH)C(HHC)C(HCC))$. As indicated in the text, the algorithm proceeds by saturating at once all atoms of a chosen orbit until there are no more orbits to be saturated. All saturation steps are indicated with an arrow labeled with a number. All graphs are represented by their list of atoms and edges, and the orbits are given in the line immediately following the graph. At each step of the algorithm, the orbit to be saturated is always the most left one. The initial graph is composed of 32 isolated atoms, 22 hydrogens, and 10 carbons. The atoms are partitioned according to their target signatures, i.e., 9 h_3 , 12 h_2 , 1 h_1 , 1 c_{31} , 2 c_{32} , 2 c_{232} , 2 c_{222} , 2 c_{221} , and 1 c_{1322} . The first orbit to be saturated is h_3 . According to the target signature, each of the nine atoms of the orbit must be linked to a carbon attached to a total of three hydrogens. There are three of such atoms belonging to two orbits, one in c_{31} and two in c_{32} . Obviously only one graph can be constructed, the three first atoms of h_3 are attached to the atom of orbit c_{31} , and the six remaining atoms are attached to the atoms of orbit c_{32} . Steps (2) and (3) are similar to step (1) and correspond to the saturation of the hydrogen orbits h_2 and h_1 . In step (4), orbit c_{31} is saturated, this orbit is composed of one atom, which according to its signature must be attached to a carbon linked to three other carbon, the only atom having this characteristic is the one belonging to c_{1322} . In step (5) the two atoms of orbit c_{32} are attached to the atoms of orbit c_{232} . Step (6) leads to three solutions graphs (61), (62), and (63). In this step, the two atoms to be saturated have the target signature c_{232} and thus must be linked to two carbons one carrying three hydrogens and one having two hydrogens. The two unsaturated atoms of c_{232} are already linked to carbons attached to three hydrogens, thus bonds only need to be created between c_{232} and carbons linked to two hydrogens. Such atoms are found in orbits c_{222} and c_{221} . The first solution (61) consists of linking c_{232} to the two atoms of c_{222} . Note that c_{232} cannot be linked to the same c_{222} atom since this would create the saturated subgraph $CH_3CH_2-CH_2-CH_2CH_3$. In the second solution (62), the first atom of c_{232} is linked to c_{222} and the second to c_{221} . In the resulting graph the orbits c_{222} and c_{221} are divided into c_{2221}, c_{2222} and c_{2211}, c_{2212} . In the last solutions (63) the two atoms of c_{232} are linked to c_{221} . The orbit to be saturated, c_{221} , in solution (63) is composed of two atoms, which according to their target signatures must be linked to a carbon atom attached to only one hydrogen, there is only one such atom in orbit c_{1322} . Attaching c_{221} to c_{1322} creates a saturated subgraph, the resulting graph and is therefore discarded. In solution (62) the orbit to be saturated is c_{2221} , the target signature is a carbon attached to two hydrogens and two carbons of its own type. Carbons carrying two hydrogens are in orbits c_{2222}, c_{2211} , and c_{2212} . Attaching c_{2221} to c_{2211} is discarded since it creates a saturated subgraph. Orbit c_{2221} can be attached to c_{2212} , but the following step (723) leads to a saturated subgraph. Linking c_{2221} to c_{2222} produces a valid graph leading to methyl-4-nonane. Similarly, solution (61) is a valid graph that leads to methyl-5-nonane.

in Figure 4. The algorithm given below checks if a bond can be created between two atoms x_1 and x_2 and be compatible with the signature of x_2 . This algorithm must be run twice, once to check compatibility with x_2 and once with x_1 . Let $h\tau(x_1)$ and $h\tau(x_2)$ be the target atomic h -signatures for x_1 and x_2 . One first computes n_{12} , the number of

permissible bonds between x_1 and x_2 according to their target signatures. Specifically, for each atom y_1 bonded to x_1 in $h\tau(x_1)$, one computes its atomic height $h-1$ signature, $h-1\sigma_{\tau(x_1)}([y_1])$. Note that aside from being a string of character, $h\tau(x_1)$ also represents a molecular subgraph as we have shown in our previous paper.² Hence, atomic signatures

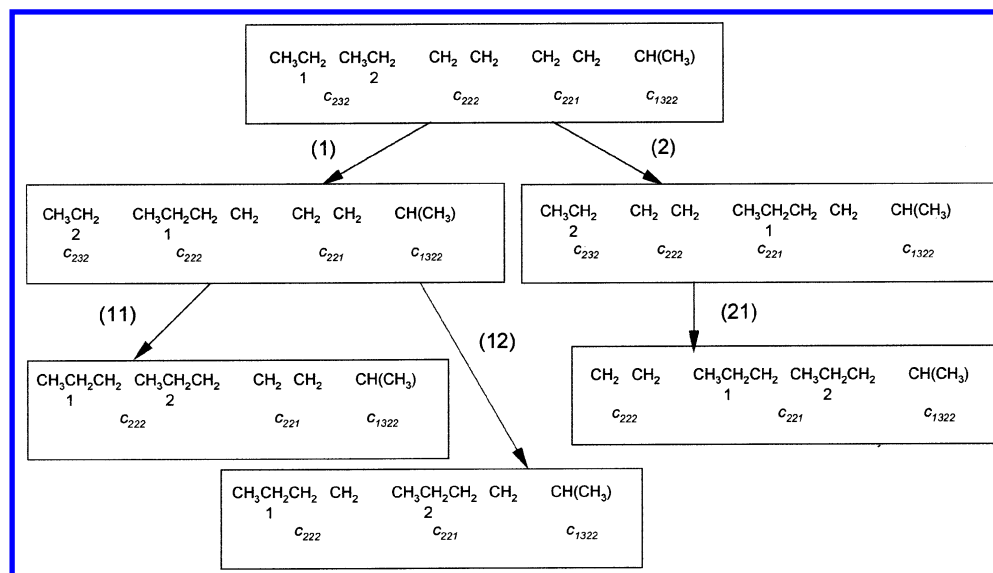


Figure 2. Orbit saturation algorithm. The initial graph is the graph produced by step (5) in Figure 1. The orbit to be saturated is c_{232} , the notation being defined in the caption of Figure 1. This orbit is composed of two atoms labeled 1 and 2. Atom 1 is first saturated producing two possible graphs, graph (1) where atom 1 is linked to orbit c_{222} and graph (2) where atom 1 is linked to orbit c_{221} . From graph (1), atom 2 is then saturated producing two solutions, graph (11) where atom 2 is attached to orbit c_{222} and graph (12), where atom 2 is attached to orbit c_{221} . From graph (2) only graph (21) is produced where atom 2 is connected to orbit c_{221} . The reason atom (2) when connected to orbit c_{222} in graph (2) does not produce a valid solution is detailed in Figure 3.

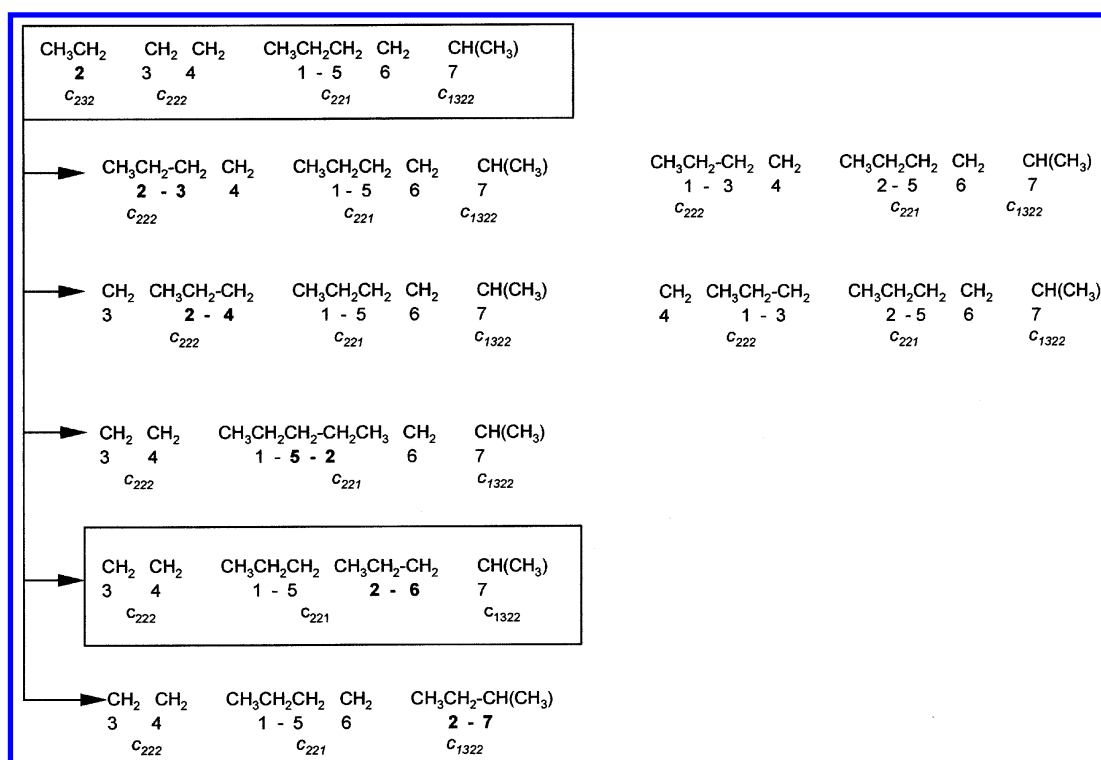


Figure 3. Vertex saturation algorithm. The initial graph is the graph produced by step (2) in Figure 2. The atom to be saturated is atom 2, since atom 1 has already been saturated and bond 1-5 has been created. The figure lists the five possible attachments for atom 2, namely, 2-3, 2-4, 2-5, 2-6, and 2-7. Attachments 2-3 and 2-4 do not produce canonical graphs, the corresponding canonical graphs being drawn on the right side of the figure. Attachment 2-5 is rejected because it creates a saturated subgraph. Attachment 2-6 produces a valid canonical graph. Attachment 2-7 produces a graph that violates the target signatures for atoms 2 (orbit c_{232}) and atom 7 (orbit c_{1322}), the reason for this is further detailed in Figure 4.

can be computed on ${}^h\tau(x1)$ the same way they are computed on molecular graphs. The bond $[x1,x2]$ is permissible if $y1$ can be found such that ${}^{h-1}\sigma_{\tau(x1)}([y1]) = {}^{h-1}\tau(x2)$, that is an atom can be found in ${}^h\tau(x1)$ having the same height $h-1$ signature than $x2$. We prove in the next section that this requirement is a necessary condition for the final graphs to match the target signature. The above test is performed for

all vertices $y1$ attached to $x1$; each time a match is found the number of permissible bonds, n_{12} , is augmented by one. Obviously, if $n_{12} = 0$, $x1$ and $x2$ cannot be connected. When the bond $[x1,x2]$ is permissible (e.g., $n_{12} \neq 0$), the second step checks that m_{12} , the number of bonds of type $[x1,x2]$ already created in G , does not exceed its limit (e.g., n_{12}). Precisely, one computes the number of atoms attached to $x1$

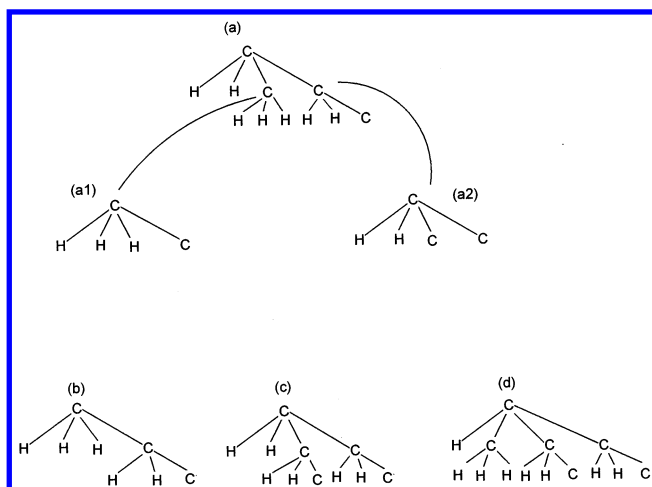


Figure 4. Example of signature compatible and signature incompatible bonds. The figure represents the height-2 signatures of four-atom (a), (b), (c), and (d). Using the notation given in the caption of Figure 1, atom (a) belongs to orbit c_{232} , atom (b) to orbit c_{32} , atom (c) to orbit c_{222} , and atom (d) to orbit c_{1322} . The signatures (a1) and (a2) are respectively the height-1 signatures of atom (a1) and (a2) taken in the graph representing the signature of (a). The 1-signature of atom (a1) is equal to the 1-signature of atom (b). Atom (a) and (b) or orbits c_{232} and c_{32} may thus be bounded according to their signatures. Similarly, the 1-signature of atom (a2) equals the 1-signature of atom (c), and atom (a) and (c) or orbits c_{232} and c_{222} can be bonded. Neither atom (a1) nor atom (a2) lead to a 1-signature equal of the 1-signature of atom (d). Atoms (a) and (d) or orbits c_{232} and c_{1322} cannot be bounded.

having the same height $h-1$ target signature than x_2 . When this number is equal to the limit, bond $[x_1, x_2]$ cannot be created, otherwise the above two steps are reapplied to verify that bond $[x_1, x_2]$ is compatible with the signature of x_1 . In short bond $[x_1, x_2]$ is signature compatible iff

$$n_{12} - m_{12} \geq 0 \text{ and } n_{21} - m_{12} \geq 0 \quad (1)$$

where $n_{12} = \{y_1 \text{ bonded to } x_1 \text{ in } {}^h\tau(x_1) \text{ such that } {}^{h-1}\sigma_{\tau(x_1)}(y_1) = {}^{h-1}\tau(x_2)\}$, $n_{21} = \{y_2 \text{ bonded to } x_2 \text{ in } {}^h\tau(x_2) \text{ such that } {}^{h-1}\sigma_{\tau(x_2)}(y_2) = {}^{h-1}\tau(x_1)\}$, $m_{12} = \{y_1 \text{ bonded to } x_1 \text{ in } G \text{ such that } {}^{h-1}\tau(y_1) = {}^{h-1}\tau(x_2)\}$, and $m_{21} = \{y_2 \text{ bonded to } x_2 \text{ in } G \text{ such that } {}^{h-1}\tau(y_2) = {}^{h-1}\tau(x_1)\}$.

```
compatible-bond-signature(x1,x2,g,h)
Input:  x1,x2 two atoms
        G a molecular graph
        h the target molecular signature
Output: TRUE or FALSE

Let n12 = 0 be the number of permissible bonds between x1 and x2
For all atoms y1 in hτ(x1) bounded to x1 do
  if h-1στ(x1)(y1) = h-1τ(x2) then n12 = n12+1
done
if n12 = 0 return FALSE
Let m12 = 0 be the number of bonds of type [x1, x2] already created
For all atoms y1 in G bounded to x1 do
  if h-1τ(y1) = h-1τ(x2) then m12 = m12+1
done
if n12 - m12 < 0 return FALSE
return TRUE
```

ALGORITHM CORRECTNESS

The intent of this section is to prove the algorithm we presented above is correct. Precisely, we show the algorithm produces the set of all possible graphs matching a given signature and that there are no duplicates in the solution set. In the following, G is a graph not necessarily fully saturated unless otherwise stated, and ${}^h\tau$ is a given target signature.

Proposition 1. Let x and y be two bonded atoms in a molecular graph G

$${}^{h-1}\sigma_G(y) = {}^{h-1}\sigma_{{}^h\sigma_G(x)}(y)$$

Proof. ${}^h\sigma_G(x)$ is a canonical representation of the subgraph composed of all atoms that are at most h distant from x . Because atom y is bonded to x , all atoms that are at most $h-1$ distant from y in G are contained in ${}^h\sigma_G(x)$. Now, ${}^{h-1}\sigma_{{}^h\sigma_G(x)}(y)$ is a subgraph of ${}^h\sigma_G(x)$ containing all atoms that are at most $h-1$ distant from y in ${}^h\sigma_G(x)$. Since all atoms that are at most $h-1$ distant from y in G are in ${}^h\sigma_G(x)$ and ${}^{h-1}\sigma_{{}^h\sigma_G(x)}(y)$, we have ${}^{h-1}\sigma_G(y) = {}^{h-1}\sigma_{{}^h\sigma_G(x)}(y)$.

Proposition 2. Let G be a graph having the signature ${}^h\tau$. All bonds in G are signature compatible.

Proof. Let x_1 and x_2 be two bonded vertices in G , we must prove eq 1 is valid. Let ${}^h\tau(x_1)$ and ${}^h\tau(x_2)$ be the target signatures of x_1 and x_2 . Since ${}^h\sigma(G) = {}^h\tau$, we have ${}^h\sigma_G(x_1) = {}^h\tau(x_1)$ and ${}^h\sigma_G(x_2) = {}^h\tau(x_2)$. Trivially, we also have ${}^{h-1}\sigma_G(x_1) = {}^{h-1}\tau(x_1)$ and ${}^{h-1}\sigma_G(x_2) = {}^{h-1}\tau(x_2)$. Now, according to eq 1, $m_{12} = \{y_1 \text{ bonded to } x_1 \text{ in } G \text{ such that } {}^{h-1}\tau(y_1) = {}^{h-1}\tau(x_2)\}$, and using the above equalities we have $m_{12} = \{y_1 \text{ bonded to } x_1 \text{ in } G \text{ such that } {}^{h-1}\sigma_G(y_1) = {}^{h-1}\sigma_G(x_2)\}$. All atoms bonded to x_1 in G are bonded to x_1 in ${}^h\sigma_G(x_1)$, thus, $m_{12} = \{y_1 \text{ bonded to } x_1 \text{ in } {}^h\sigma_G(x_1) \text{ such that } {}^{h-1}\sigma_G(y_1) = {}^{h-1}\sigma_G(x_2)\}$. Using Proposition 1, we obtain $m_{12} = \{y_1 \text{ bonded to } x_1 \text{ in } {}^h\sigma_G(x_1) \text{ such that } {}^{h-1}\sigma_{\tau(x_1)}(y_1) = {}^{h-1}\sigma_G(x_2)\}$. Because ${}^h\sigma_G(x_1) = {}^h\tau(x_1)$, ${}^{h-1}\sigma_G(x_1) = {}^{h-1}\tau(x_1)$, and ${}^{h-1}\sigma_G(x_2) = {}^{h-1}\tau(x_2)$, we have $m_{12} = \{y_1 \text{ bonded to } x_1 \text{ in } {}^h\tau(x_1) \text{ such that } {}^{h-1}\sigma_{\tau(x_1)}(y_1) = {}^{h-1}\tau(x_2)\}$. Thus, $m_{12} = n_{12}$, the number of permissible bonds between x_1 and x_2 . Using the same technique we prove $m_{21} = n_{21}$ and eq 1 is indeed valid.

Proposition 3. The enumerate-graph-signature algorithm constructs an exhaustive list of graphs matching the given target signature.

Proof. Let us recall that the enumerate-graph-signature algorithm generates the set of graphs that can be constructed by saturating all the orbits of an initial graph. In turn, generating all graphs saturating a given orbit is performed by constructing the set of graphs that saturate all the atoms of the orbit. Let us first assume that the tests for bond compatibility, graph canonicity, and saturated subgraph are not performed when saturating a given atom x . Clearly then, every bond between atom x and any other atom y of the graph produces a solution. Since all atoms x of the graph are considered one after another in saturate-orbit-signature, the algorithm generates all possible labeled graphs that can be constructed from the initial graph. Solutions are not missed when testing if the bonds are signature compatible, since according to Proposition 2, signature compatibility is a necessary condition for the final graphs to have the target signature.

Regarding canonicity, Read proved some time ago¹⁷ that any canonical graph containing $m+1$ edges is an augmentation of some canonical graph containing m edges. Since the initial graph contains no bonds it is therefore canonical, and because our algorithm tests all possible edge augmentations on graphs that are always canonical, checking for canonicity does not skip solutions. Finally, avoiding saturated subgraphs is a necessary condition for the final graphs to be connected, thus the saturated subgraph test does not miss solutions.

Table 1. CPU Running Times for the Series of Compounds $\text{CH}_3[\text{CH}_2]_x\text{CH}(\text{CH}_3)[\text{CH}_2]_x\text{CH}_3$, with $x = n/2 - 2^a$

n	signature height	CPU time (s)	nb calls to enumerate-molecule-signature
4	1	0.02	2
8	2	0.1	5
12	3	0.32	7
16	4	0.89	9
20	5	1.8	11
24	6	3.33	13
28	7	5.56	15
32	8	8.9	17
36	9	13.26	19
40	10	19.58	21
44	11	27.98	23
48	12	38	25
52	13	50	27

^a CPU times were recorded on an SGI O2 workstation.

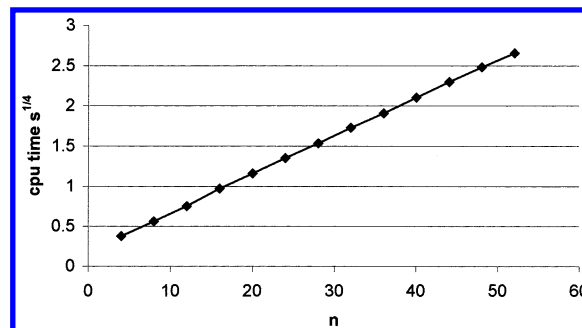
Proposition 4. The enumerate-graph-signature algorithm produces a list of nonisomorphic graphs.

Proof. We first notice that the set of graphs returned when saturating all the atoms of a given orbit does not contain isomorphic graphs, since each graph is canonical. To prove the final solutions are not isomorphic, we need to show that given two canonical graphs G_1 and G_2 both saturating the same orbit, no augmentation of G_1 is isomorphic to an augmentation of G_2 . The proof of this latter assertion can be found in an earlier paper.¹¹ For the reader's convenience we give the proof in the Appendix.

ALGORITHM EFFICIENCY

While each step of the algorithm presented earlier is polynomial in time, it does not mean the algorithm runs efficiently. In fact, enumeration algorithms can run for an exponential time, simply because the number of solutions may be exponential. It is customary with enumeration algorithms to evaluate their computational time per output. It has been shown that orderly enumeration algorithms run in theory in polynomial-time per output when the solution graphs are not constrained.¹⁸ In the present case, all solutions have to meet two constraints, connectivity and signature compatibility. Although we have implemented necessary conditions on the graphs being constructed in order to meet the constraints, these conditions are not sufficient, and intermediate graphs can still be produced that do not lead to valid final solutions. Examples of such unproductive intermediate graphs are shown in Figure 1, graphs (63) and (722). It is rather difficult to theoretically evaluate the number of unproductive intermediate graphs, but at best we can study the actual running CPU time per output on test examples.

We chose to generate all molecules matching the signatures of the series of compounds $\text{CH}_3[\text{CH}_2]_x\text{CH}(\text{CH}_3)[\text{CH}_2]_x\text{CH}_3$ with $x = n/4 - 1$, and the total number of non-hydrogen atoms, n , ranging from 4 to 52. The signature height chosen for each compound was the one leading to only one solution and the heights turned out to follow the expression $h = n/4$. This expression is further investigated in the latter section of the paper. The running CPU times of our algorithm are reported in Table 1 and Figure 5. The scaling appears to be $O(n^4)$. This is not surprising considering that for each molecule of the series, the n atoms have four possible bonds. Thus, there is a maximum of n^4 bonds that are created and

**Figure 5.** CPU running times for the series of compounds $\text{CH}_3[\text{CH}_2]_x\text{CH}(\text{CH}_3)[\text{CH}_2]_x\text{CH}_3$, with $x = n/2 - 2$. The time reported on the ordinate axis is $[\text{CPU time (s)}]^{1/4}$ derived from Table 1.

tested using our algorithm. Finally, it is interesting to note from Table 1 that the number of times our algorithm calls itself scales linearly with n . Thus, the number of unproductive intermediate graphs does not appear to worsen the overall computational complexity.

ENUMERATING COMPOUNDS HAVING SIMILAR ACTIVITIES OR PROPERTIES

A straightforward application of our algorithm is to enumerate compounds having similar activities or properties. We argued in our previous paper² why signature, like other fragment-based descriptors, performs well in QSAR and QSPR studies. A direct consequence of our arguments is that compounds having identical signatures should give similar activities or properties. The algorithm presented here can thus be used to generate all molecular compounds having the signature of a particular molecule found experimentally to be biologically active or having desired physical properties. To illustrate such an utilization of our algorithm we extracted from the HIV-1 protease inhibitor study we reported earlier² the seven most experimentally active compounds used in that study from Thompson et al.¹⁹ Signatures up to height 4 were computed for the seven compounds and the algorithm was used to produce all molecular graphs matching the calculated signatures. We found all seven compounds to have a unique height-4 signature, and, as reported in Table 2, none of them have a height-3 signature that is shared with more than one other molecule. The structures listed on the right side of Table 2 are new compounds that have not been tested by Thompson et al. They represent the exchange of a para-substituted group between two phenyls. According to other group exchanges on these phenyls reported by Thompson within this data set, it is likely that all of the new compounds reported in Table 2 have similar pIC_{50} values.

When we move to height-2 signatures, the number of compounds that we can form which have the same signatures as that of the seven compounds selected from the Thompson set increases, as expected. For example, in Figure 6 we present the eight unique compounds that have the same height-2 signature as that of the most active compound found by Thompson ($\text{pIC}_{50} = 10.155$). Note that two of those compounds (upper row) can also be found in Table 2 since they have the same height-3 signature. The other six compounds represent the extension of the aromatic ring off the cyclic pentane that has not been previously evaluated. Though, by inspection, sterics may eliminate some of these compounds from any experimental study, the structures in

Table 2. All Compounds Having the Same Height-3 Signature than the Seven Most Active HIV-1 Protease Inhibitors Found by Thompson et al.¹⁹

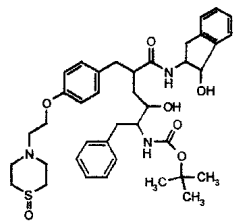
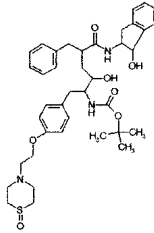
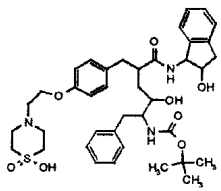
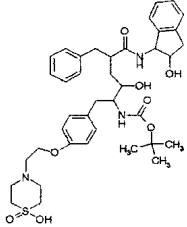
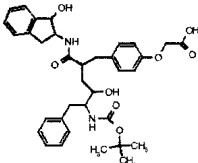
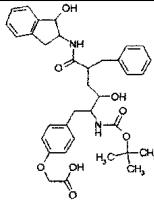
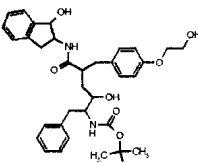
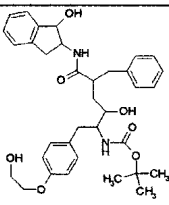
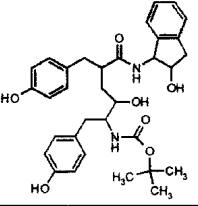
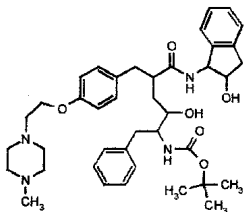
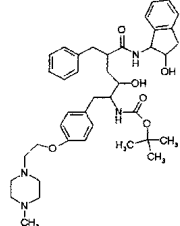
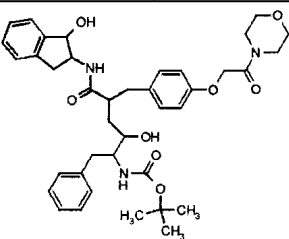
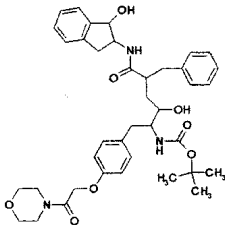
pIC ₅₀	Structure found by Thompson et al.	Structure having same signature
10.155		
10.097		
10.389		
10.046		
9.824		no isomer found
9.699		
9.699		

Figure 6 represent intermediates between the height-0 signature (same molecular formula – many compounds) and

the height-4 signature (unique compound). Probing for similar compounds in this manner by creating focused

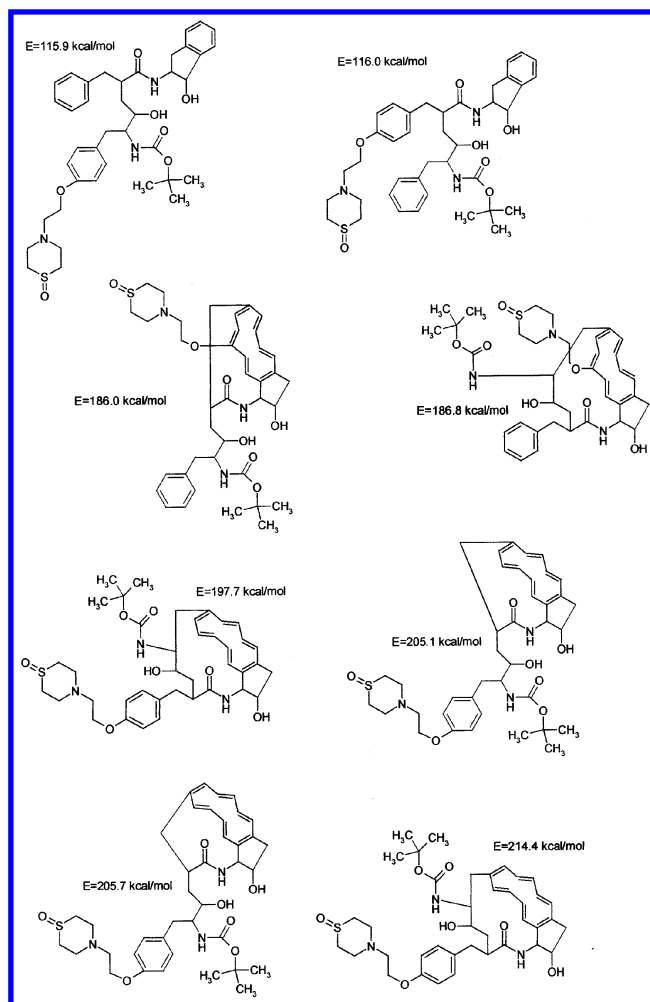


Figure 6. All compounds having the same height-2 signature as the most active HIV-1 protease inhibitor found by Thompson et al.¹⁹ Thompson et al. compound is on the right upper corner and has a pIC_{50} value of 10.155. Energy values were computed on minimized structures using the Dreiding force field.³⁶

libraries of many compounds, tuned by height of the signature used, may prove a fruitful methodology in drug development. Before pursuing further this direction, one first needs to be able to compute the number of compounds, or degeneracy, matching a given signature and to calculate the minimum height that uniquely characterize a compound. To this end, in the next section we investigate the degeneracy of signature for several families of hydrocarbons.

DEGENERACY OF SIGNATURE AND TOPOLOGICAL INDICES

We have established in our first paper that any molecular structure is uniquely characterized by its molecular signature of height $D+1$, where D is the diameter of the molecular graph.² We also have in hand an algorithm that can enumerate all molecules matching a given signature. The goal of this section is to apply this algorithm to study the actual degeneracy of signature for different heights and to compare this degeneracy with the degeneracy of popular TIs used in QSAR and QSPR studies. Since we have established that all TIs based on count of walks, paths, and distances can be computed from signature,² we expect signature to exhibit lower degeneracy than other TIs. To verify this assertion, we systematically computed the degeneracy of signature and

Table 3. Number of Isomeric Alkanes ($\text{C}_n\text{H}_{2n+2}$)

n	nb isomers	n	nb isomers
1	1	10	75
2	1	11	159
3	1	12	355
4	2	13	802
5	3	14	1858
6	5	15	4347
7	9	16	10359
8	18	total	18030
9	35		

selected TIs on four homogeneous series of compounds. To probe the effect of branching on descriptor degeneracy, we chose the alkane series up to 16 carbon atoms. To study the role of heteroelements, the alcohol series up to eight carbon atoms and four oxygen atoms was generated. To examine the effect of cyclicity, a fullerene-type structure series up to 16 carbon atoms was constructed. Finally, a diverse series of 450 peptide sequences ranging from 18 to 4260 amino acids was extracted from a protein databank to compute the degeneracy of signature with biochemical compounds.

Previous Studies on Degeneracy. Studies of and discussions on the degeneracy of molecular descriptors in the open literature are quite prevalent. Here we will review a few of those efforts. Gutman et al.²⁰ looked at the isomer degeneracy of the Wiener index for medium sized molecules and concluded that high-isomer degeneracy exists. Demirev et al.²¹ developed a new TI based on atomic charge distributions of organic molecules and demonstrated the low degeneracy of this descriptor. Randic et al.²² explored the degeneracy of novel bond-additive descriptors as well as the Wiener index, connectivity index, and others for some nonanes and hexanes. Estrada²³ reduced the degeneracy of the Randic index through a modification and showed the improved results for nearly two-dozen graphs. Ivanciuc et al.²⁴ explored the degeneracy of topological distance sum and distance degree sequences for small fullerenes. Finally, A. Balaban^{25,26} developed new methods to generate descriptors having the property of low degeneracy. While investigating the degeneracy of molecular descriptors is not new, to the best of our knowledge, the study reported below appears to be the first where degeneracy is systematically probed for homogeneous molecular series on a *wide-variety* of molecular descriptors.

Computational Protocol. For each of the compound series considered in the paper our computational protocol is composed of three steps, (1) compound construction, (2) descriptor computation, and (3) degeneracy calculation.

(1) *Compound Construction.* This first step was applied to the four series mentioned above. All alkanes up to 16 carbon atoms were generated in SMILES format using the enumeration algorithm presented earlier. To generate all isomers the algorithm was run for each height-0 signature of the molecular series, or in other words, for each molecular formula of the series. Note that the numbers of isomers we found in Table 3, totaling 18 030, are consistent with other literature sources.²⁷ Mono-, di-, tri-, and tetraalcohols were generated in SMILES format using the same procedure as for alkanes. A total of 5009 compounds were constructed for this series, the numbers of isomers vs carbon numbers listed in Table 4. Fullerene-type structures were generated up to 16 atoms using the Genreg generator²⁸ for regular graphs. More precisely, Genreg was used to compute the

Table 4. Number of Isomeric Alcohols ($C_nH_{2n+2}O_k$)

<i>n</i>	<i>k</i>					total
	0	1	2	3	4	
1	1	1	1	1	1	5
2	1	1	2	2	2	8
3	1	2	4	5	6	18
4	2	4	9	14	20	49
5	3	8	21	39	62	133
6	5	17	52	109	198	381
7	9	39	129	312	625	1114
8	18	89	332	890	1972	3301
total	40	161	550	1372	2886	5009

Table 5. Number of Fullerene Type (C_n) or Spheroalkane Type (CH_n) Structures

<i>n</i>	nb isomers	<i>n</i>	nb isomers
6	2	14	509
8	5	16	4060
10	19	total	4680
12	85		

adjacency matrices of all cubic graphs up to 16 vertices, and a postprocess was written and implemented to convert vertices into carbon atoms and adjacency matrices into SMILES. As depicted in Table 5, a total number of 4680 structures were built. Finally, a set of 450 peptide sequences was extracted from the FSSP database.²⁹ Note that all the structures in the FSSP database have been screened with the alignment algorithm DALI³⁰ to ensure the uniqueness of the sequences. The purpose of this procedure was to extract peptide sequences representing major distinct protein families and functions. Originally written in PDB, format all peptides were simply stored using their amino acid sequences. The sequence lengths of these peptides varied from 18 amino acids to 4260.

(2) *Descriptor Computation.* Signatures up to height 7 were systematically computed for all compounds generated in the first step. For alkanes, alcohols, and fullerenes, signatures were derived over the alphabet {C, H, O}, while for peptides signatures were computed over the one letter alphabet of 20 amino acids. For a complete description on how to compute signatures the reader is referred to our first paper.² For all compounds in the alkanes, alcohols, and fullerene series a total of 47 TIs were computed using the Molconnn-Z software package.³¹ We arbitrarily divided these TIs into five categories. (a) The connectivity indices composed of all connectivity ($^{0-10}\chi$) and valence connectivity indices ($^{0-10}\chi_v$) up to a path of length 10, the difference of chi-cluster-3 and path-cluster-4 ($^3\chi_c - ^4\chi_{pc}$), and the difference of the valence chi-cluster-3 and valence path-cluster-4 ($^3\chi_{vc} - ^4\chi_{vpc}$). (b) The shape indices comprising Kappa ($^{0-3}\kappa$) and Kappa alpha ($^{1-3}\kappa_\alpha$) indices up to path length 3 and the flexibility index (ϕ_α). (c) The distance indices including the Wiener index (*W*), the product Wiener index (*Wp*), the total Wiener index (*Wt*), and the Platt number (*F*). (d) The information indices composed of the Shannon information index (*I*) and the Bonchev and Trinajstić information indices I_D^W and I_D^E and their corresponding mean information indices \bar{I}_D^W and \bar{I}_D^E . (e) The hybrid or other indices which includes the total topological index (τ) and total topological index based on electrotopological states (τ_{ets}), the sum of the intrinsic state values ($sumI$), the sum of perturbation values of the intrinsic states ($sum\Delta I$), the number of atom equivalent classes (*nclass*), and the number of hydrogen donors (*HBd*)

Table 6. Signature Degeneracy

	signature height							
degeneracy	0	1	2	3	4	5	6	7
% Alkanes								
1	0.0	0.1	57.5	99.1	100.0	100.0	100.0	100.0
10	0.1	1.1	42.3	0.9	0.0	0.0	0.0	0.0
100	0.7	10.0	0.2	0.0	0.0	0.0	0.0	0.0
1000	7.3	64.4	0.0	0.0	0.0	0.0	0.0	0.0
10000	91.9	24.4	0.0	0.0	0.0	0.0	0.0	0.0
100000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
% Alcohols								
1	0.2	2.9	99.2	100.0	100.0	100.0	100.0	100.0
10	1.3	34.8	0.8	0.0	0.0	0.0	0.0	0.0
100	7.4	62.4	0.0	0.0	0.0	0.0	0.0	0.0
1000	51.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10000	39.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
100000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
% Fullerene-type Structures								
1	0.0	0.0	5.4	99.7	100.0	100.0	100.0	100.0
10	0.1	0.1	24.5	0.3	0.0	0.0	0.0	0.0
100	2.2	2.2	50.8	0.0	0.0	0.0	0.0	0.0
1000	10.9	10.9	19.3	0.0	0.0	0.0	0.0	0.0
10000	86.8	86.8	0.0	0.0	0.0	0.0	0.0	0.0
100000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
% Peptides								
1	0	7.9	95.1	98.9	99.6	99.8	99.8	100.0
10	0	10.2	2.7	0.4	0.4	0.2	0.2	0.0
100	0	11.5	0.4	0.0	0.0	0.0	0.0	0.0
1000	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
>1000	100	70.4	1.8	0.7	0.0	0.0	0.0	0.0

and acceptors (*HBa*). Most of the above TIs are defined in the first paper of the series.² For a complete set of definitions the reader is referred to the Molconnn-Z software package manual.³¹

(3) *Degeneracy Calculation.* With the exception of peptides, the degeneracy was simply computed for each series and each descriptor by counting the number of structures matching a given TI or signature value. For peptides it was found that no two sequences had the same signature even for height 0. This is not surprising considering that the peptides were selected to ensure uniqueness of the sequences. Consequently, all peptides have different 0-signatures, that is, different amino acid compositions. To probe the degeneracy of signature beyond our restricted data set of 450 sequences, we use the enumeration algorithm presented earlier with signatures written over the alphabet of amino acids. Note that the enumeration algorithm searches sequences matching the given signature in the entire space of all possible amino acid sequences. That is, our algorithm searches for sequences beyond naturally occurring or synthesized sequences that may be stored in protein databanks. Therefore, the degeneracy computed with the enumeration algorithm is an overestimate of the “real” degeneracy for signatures of existing proteins.

RESULTS AND DISCUSSION

Degeneracy results are presented in Tables 6–11. All tables must be read as follows. For the alkane, alcohol, and fullerene series, the degeneracy is compiled into 6 bins 1, 2–10, 11–100, 101–1000, 1001–10000, and 10001–100000. All numbers are percentages and sum up to 100 for each descriptor and each compound series. For example in Table 6, 57.5% appears for height-2 signature in the alkanes

Table 7. Connectivity Indices Degeneracy

degeneracy	$^0\chi_v$	$^1\chi_v$	$^2\chi_v$	$^3\chi_v$	$^4\chi_v$	$^5\chi_v$	$^6\chi_v$	$^7\chi_v$	$^8\chi_v$	$^9\chi_v$	$^{10}\chi_v$	$^3\chi_c-^4\chi_{pc}$	$^3\chi_{vc}-^4\chi_{vpc}$
% Alkanes													
1	0.1	2.1	26.6	42.8	37.2	25.1	13.1	5.8	2.4	1.0	0.4	19.9	19.9
10	1.2	32.9	69.3	57.1	62.5	72.5	67.0	37.8	16.5	6.9	2.3	53.1	53.1
100	9.9	61.2	4.1	0.1	0.3	2.4	17.7	33.8	26.4	14.1	7.3	27.1	27.1
1000	64.4	3.7	0.0	0.0	0.0	0.0	2.2	9.9	20.0	16.0	7.5	0.0	0.0
10000	24.4	0.0	0.0	0.0	0.0	0.0	0.0	12.7	34.7	0.0	0.0	0.0	0.0
100000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	62.0	82.5	0.0	0.0
% Alcohols													
1	0.9	11.5	63.5	61.7	44.4	22.3	7.4	1.7	0.3	0.1	0.0	1.6	39.6
10	7.0	72.8	36.5	38.0	54.5	58.1	27.0	7.8	2.8	0.4	0.0	44.6	56.7
100	39.1	15.7	0.0	0.3	1.1	9.9	27.7	15.4	2.3	0.0	0.0	53.8	3.7
1000	52.9	0.0	0.0	0.0	0.0	9.7	38.0	0.0	0.0	0.0	0.0	0.0	0.0
10000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	75.1	94.6	99.5	0.0	0.0	0.0
100000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
% Fullerene-type Structures													
1	0.0	0.0	0.0	0.1	0.4	0.6	1.3	2.3	4.5	8.8	13.9	0.0	0.0
10	0.1	0.1	0.1	0.4	2.5	8.0	17.9	53.1	84.3	88.5	83.5	0.1	0.1
100	2.2	2.2	2.2	5.0	38.2	74.3	80.8	44.6	11.3	2.6	2.5	2.2	2.2
1000	10.9	10.9	10.9	47.7	58.9	17.1	0.0	0.0	0.0	0.0	0.0	10.9	10.9
10000	86.8	86.8	86.8	46.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	86.8	86.8
100000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 8. Shape Indices Degeneracy

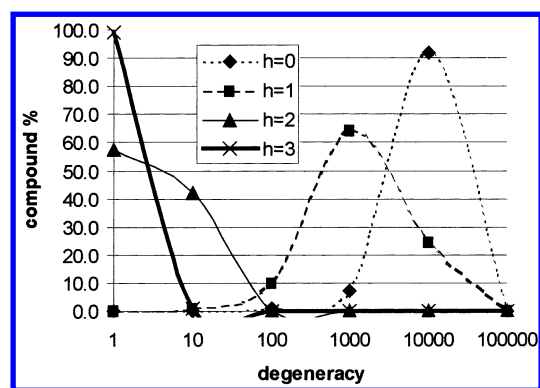
degeneracy	$^0\kappa$	$^1\kappa$	$^2\kappa$	$^3\kappa$	$^1\kappa_\alpha$	$^2\kappa_\alpha$	$^3\kappa_\alpha$	ϕ
% Alkanes								
1	0.4	0.0	0.1	0.0	0.0	0.1	0.0	0.1
10	2.4	0.1	0.8	1.1	0.1	0.8	1.1	0.8
100	13.4	0.7	5.9	9.7	0.7	5.9	9.7	5.9
1000	55.0	7.3	46.7	64.0	7.3	46.7	64.0	46.7
10000	28.8	34.4	46.6	25.1	34.4	46.6	25.1	46.6
100000	0.0	57.5	0.0	0.0	57.5	0.0	0.0	0.0
% Alcohols								
1	0.2	0.0	0.0	0.0	0.2	0.8	0.2	0.8
10	4.4	0.2	1.3	0.6	1.3	5.7	7.1	5.7
100	20.0	2.2	15.6	27.0	7.4	37.7	42.0	37.7
1000	75.4	27.9	83.1	72.4	51.8	55.8	50.7	55.8
10000	0.0	69.6	0.0	0.0	39.4	0.0	0.0	0.0
100000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
% Fullerene-type Structures								
1	0.8	0.0	0.0	0.1	0.0	0.0	0.1	0.0
10	3.9	0.1	0.1	0.9	0.1	0.1	0.9	0.1
100	15.5	2.2	2.2	5.1	2.2	2.2	5.1	2.2
1000	44.4	10.9	10.9	46.9	10.9	10.9	46.9	10.9
10000	35.4	86.8	86.8	46.8	86.8	86.8	46.8	86.8
100000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

series, thus 57.5% of all alkanes generated have a unique height-2 signature. Taking another example with alkanes in Table 7, we find that 69.3% compounds have a connectivity index, $^2\chi_v$, with a degeneracy in the bin 2–10, thus 69.3% of all alkanes have a $^2\chi_v$ index value shared with 2 to 10 other alkanes. As a final example, in Table 9, we find that 35.3% of all alkanes have a Wiener number W , that is shared with 101 to 1000 other alkanes. Comparing the distributions for $^2\chi_v$ and W , it is clear that high percentages occur at lower bins for $^2\chi_v$ than W , thus making $^2\chi_v$ less degenerate than W for the alkane series.

The peptide series appears only in Table 6 since signature was the only descriptor computed for these compounds. The bins for peptides are 1, 2–10, 101–1000, and above 1000. All numbers given for peptides in Table 6 are normalized percentages for the 450 tested sequences. Hence when 7.9% of sequences have a height-1 signature with a degeneracy of 1, $0.079 \times 450 = 36$ sequences have a signature that is unique. The results are discussed next for each compound series.

Table 9. Distance Indices Degeneracy

degeneracy	W	Wp	Wt	F
% Alkanes				
1	0.4	0.0	0.4	0.0
10	4.9	0.1	4.9	0.2
100	59.4	1.7	59.4	0.9
1000	35.3	28.6	35.3	13.8
10000	0.0	69.6	0.0	85.1
100000	0.0	0.0	0.0	0.0
% Alcohols				
1	0.1	0.0	0.1	0.0
10	5.8	0.2	5.8	0.5
100	89.6	7.1	89.6	6.2
1000	4.5	92.6	4.5	93.3
10000	0.0	0.0	0.0	0.0
100000	0.0	0.0	0.0	0.0
% Fullerene-type Structures				
1	0.9	0.1	96.3	0.0
10	11.8	0.4	3.6	0.1
100	66.5	5.0	0.1	2.2
1000	20.8	47.7	0.0	10.9
10000	0.0	46.8	0.0	86.8
100000	0.0	0.0	0.0	0.0

**Figure 7.** Signature degeneracy for alkanes. $h = 0, \dots, 4$ is the signature height.

(1) *Alkane Series.* As depicted in Table 6 and Figure 7, the degeneracy of signature decreases as the signature height increases. Full nondegeneracy is reached for height 4 and almost achieved for height 3. Such a trend is not observed with any other index even with the connectivity index of

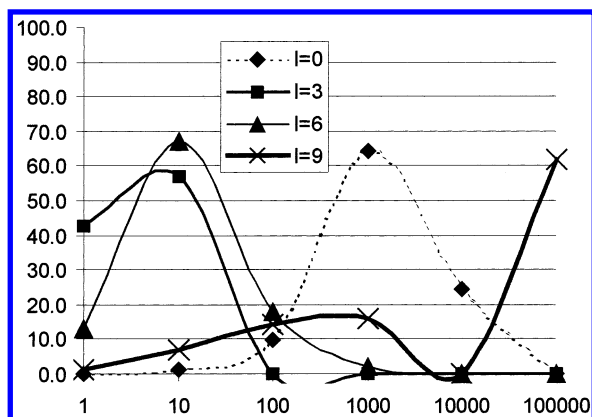


Figure 8. Connectivity indices ${}^0\chi_v$, ${}^1\chi_v$, ${}^3\chi_v$, ${}^6\chi_v$, and ${}^{10}\chi_v$ degeneracy for alkanes.

Table 10. Information Indices Degeneracy

degeneracy	I	I_D^W	I_D^W	I_D^E	I_D^E
% Alkanes					
1	0.3	81.1	1.2	59.7	1.0
10	2.4	18.9	10.2	40.2	13.2
100	13.1	0.0	59.2	0.1	85.7
1000	55.5	0.0	29.4	0.0	0.0
10000	28.8	0.0	0.0	0.0	0.0
100000	0.0	0.0	0.0	0.0	0.0
% Alcohols					
1	0.1	1.1	0.3	1.1	0.9
10	3.8	38.6	13.3	34.7	23.2
100	20.8	60.2	86.5	64.2	75.9
1000	75.4	0.0	0.0	0.0	0.0
10000	0.0	0.0	0.0	0.0	0.0
100000	0.0	0.0	0.0	0.0	0.0
% Fullerene-type Structures					
1	0.7	17.9	1.3	12.9	3.6
10	4.0	49.9	13.7	46.4	42.3
100	15.5	32.2	63.6	40.7	51.9
1000	44.4	0.0	21.5	0.0	2.2
10000	35.4	0.0	0.0	0.0	0.0
100000	0.0	0.0	0.0	0.0	0.0

increasing path lengths. In fact, as it can be seen in Figure 8 and Table 7, the degeneracy of the valence connectivity index decreases up to path length 3 but then increases for length 4 and above. The shape indices and the flexibility index (cf. Table 8) are less discriminating than the connectivity indices, and the distance-based indices (cf. Table 9) also exhibit high degeneracy. Interestingly, 81.1% of all alkanes have a unique total information index I_D^W (cf. Table 10). Mean information content indices I_D^W and I_D^E have a worse degeneracy than their corresponding total information indices I_D^W and I_D^E (cf. Table 10). Aside from signature, Kier and Hall total topological indices τ , τ_{ets} , and $sum \Delta I$ achieve the best degeneracy performances compared to all other indices (cf. Table 11). In particular, the topological index τ leads to a degeneracy close to the one obtained with height-3 signature. To further investigate the degeneracy of signature and the degeneracy of the total topological index, we extracted in the alkane series all compounds leading either to a degenerate signature or a degenerate total topological index. Examples of such compounds can be seen in Figure 9. While no specific structural patterns could be found for the total topological index, all isoalkanes $CH_3[CH_2]_xCH(CH_3)[CH_2]_yCH_3$ with $x+y+4 = n$ gave degenerate height- h signatures for $n > 4h$. This trend can clearly be observed in Table 12, where the degeneracy of signature was probed for

Table 11. Hybrid and Other Indices Degeneracy

degeneracy	τ	τ_{ets}	sumI	sum ΔI	nclass	HBd	HBa
% Alkanes							
1	98.5	94.5	0.1	90.8	0.0	n/a	n/a
10	1.5	5.5	0.9	9.2	0.0	n/a	n/a
100	0.0	0.0	8.0	0.0	0.7	n/a	n/a
1000	0.0	0.0	66.6	0.0	11.8	n/a	n/a
10000	0.0	0.0	24.4	0.0	87.4	n/a	n/a
100000	0.0	0.0	0.0	0.0	0.0	n/a	n/a
% Alcohols							
1	99.8	99.4	0.7	89.5	0.0	0.0	0.0
10	0.2	0.6	6.6	10.5	0.0	0.0	0.0
100	0.0	0.0	39.7	0.0	1.8	0.8	0.8
1000	0.0	0.0	53.0	0.0	31.1	14.2	14.2
10000	0.0	0.0	0.0	0.0	67.0	85.0	85.0
100000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
% Fullerene-type Structures							
1	99.9	99.96	0.0	n/a ^a	0.0	n/a	n/a
10	0.1	0.04	0.1	n/a	0.0	n/a	n/a
100	0.0	0.0	2.2	n/a	5.7	n/a	n/a
1000	0.0	0.0	10.9	n/a	58.9	n/a	n/a
10000	0.0	0.0	86.8	n/a	35.4	n/a	n/a
100000	0.0	0.0	0.0	n/a	0.0	n/a	n/a

^a Molconn-Z returned 0 values for all fullerene-type compounds.

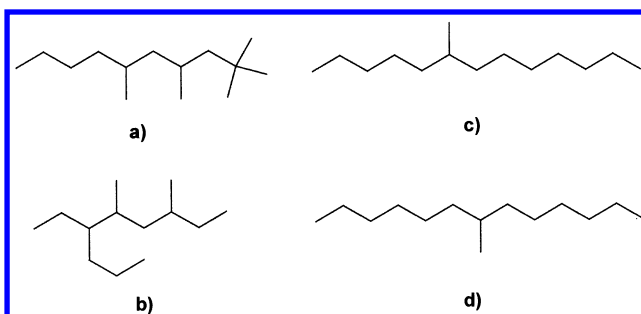


Figure 9. Examples of signature degeneracy and total topological index degeneracy for butadecane. Butadecane isomers (a) and (b) have the same total topological index value $\tau = 41.20524$. Butadecane isomers (c) and (d) have the same height-3 signature: $6H(C(HHC(HHC))) + 3H(C(HHC(HCC))) + 4H(C(HC(HHH)C(HHC))) + 12H(C(HC(HHC)C(HHC))) + 4H(C(HC(HHC)C(HCC))) + H(C(C(HHH)C(HHC)C(HHC))) + 2C(HHHC(HHC(HHC))) + C(HHHC(HC(HHC)C(HHC))) + 2C(HHC(HHH)C(HHC(HHC))) + 2C(HHC(HHC(HHH)C(HHC(HHC))) + 2C(HHC(HHC(HHC)C(HHC(HHC))) + 2C(HHC(HHC(HHC)C(HHC(HHC))) + 2C(HHC(HHC(HHC)C(HHC(HHC))) + C(HC(HHH)C(HHC(HHC)C(HHC(HHC)))$

the isoalkane series up to 40 carbon atoms. These results were obtained running our enumeration algorithm for each height-1 signatures in the series. For each solution the signatures up to height 10 were computed, and the degeneracy was calculated by counting the number of solutions having the same signature. Since isoalkanes belong to the class of compounds leading to the highest degeneracy in the alkane series, one may conjecture that all alkanes of n carbon atoms are uniquely characterized by signatures of height $\lceil n/4 \rceil$, where $\lceil \cdot \rceil$ is the ceiling function. This proposition appears to be true for any acyclic compound, as the introduction of heteroelements does not increase the degeneracy of signature, heteroelements being part of the signature notation (cf. alcohol series in Table 6 and signature definition in the Introduction). While we rigorously established that signature is nondegenerate for height $D+1$,² the $n/4$ threshold which may be smaller than the diameter D of the molecular graph remains a conjecture at the present time.

Table 12. Number of Isoalkanes Having Identical Signatures

		signature height									
n^a	0^b	1	2	3	4	5	6	7	8	9	10
4	2	1	1	1	1	1	1	1	1	1	1
6	5	2	1	1	1	1	1	1	1	1	1
8	18	4	1	1	1	1	1	1	1	1	1
10	75	7	2	1	1	1	1	1	1	1	1
12	355	10	3	1	1	1	1	1	1	1	1
14	1858	14	4	2	1	1	1	1	1	1	1
16	10359	19	5	3	1	1	1	1	1	1	1
18	60523	24	6	4	2	1	1	1	1	1	1
20	366319	30	7	5	3	1	1	1	1	1	1
22	2278658	37	8	6	4	2	1	1	1	1	1
24	1.4E+07	44	9	7	5	3	1	1	1	1	1
26	9.4E+07	52	10	8	6	4	2	1	1	1	1
28	6.2E+08	61	11	9	7	5	3	1	1	1	1
30	4.1E+09	70	12	10	8	6	4	2	1	1	1
32	2.8E+10	80	13	11	9	7	5	3	1	1	1
34	1.9E+11	91	14	12	10	8	6	4	2	1	1
36	1.3E+12	102	15	13	11	9	7	4	3	1	1
38	8.9E+12	114	16	14	12	10	8	6	4	2	1
40	6.2E+14	127	17	15	13	11	9	7	5	3	1

^a Number of carbon atoms. ^b Data taken from Trinajstić.²⁷

(2) *Alcohol Series.* Compound uniqueness for signature is reached for height 3, which is smaller than that for alkanes. The reason being that the alcohol series was generated up to eight carbon atoms and four oxygen atoms, thus giving a maximum number of 12 atoms. According to Table 12, uniqueness for alkanes composed of 12 atoms is also obtained with height-3 signature. In general, for signature and other indices, the trend for alcohols remains the same as for alkanes. Connectivity indices have a lower degeneracy than shape and distance indices (cf. Tables 7–9). Note that in Table 8, shape indices (κ) and valence shape indices (κ_v) have different degeneracy distributions due to the introduction of oxygen atoms. For the same reason, in Table 11, hydrogen bond donors (*Hbd*) and acceptors (*Hba*) have distributions for alcohols. These distributions appear to be highly degenerate. Also due to the presence of oxygen, the information indices (cf. Table 10) have greater degeneracy for alcohols than alkanes since these indices do not take account of atom types. Interestingly however, while the Molconn-Z distance indices of Table 9 do not take into account heteroelements, they do not exhibit higher degeneracy for alcohols than for alkanes. This may be due to the fact that compounds in the alcohol series contains less atoms than compounds in the alkane series. As with alkanes the lowest degeneracies are obtained with the total topological indices τ and τ_{ets} (cf. Table 11).

(3) *Fullerene-type Series.* Let us recall that the fullerene-type series is composed of all cubic graphs up to 16 vertices, where the vertices have been replaced by carbon atoms. Fullerenes, nanotubes, and spheroalkanes all belong to the class of cubic graphs. For signature, the fullerene-type degeneracy distribution is similar to the distribution of the alkane series, full nondegeneracy is achieved with height-4 signature (cf. Table 6). Furthermore, a thorough examination of the results revealed that height $n/4$ signatures uniquely characterize all fullerene-type structures up to $n = 16$ atoms. This is somewhat surprising as cyclic graphs are usually harder to characterize than acyclic ones. However, Jerrum and Skyum³² have shown that cubic graphs of n vertices have a diameter approximating $D = 1.47 \log_2(n)$. Since height

$D+1$ signatures uniquely characterized every molecular graph, cubic graphs should be uniquely identified with signature heights smaller than $n/4$ as n increases (for $n \geq 34$, $D+1 < n/4$). The connectivity indices have a higher degeneracy for fullerenes than for alkanes up to path lengths of 7, but the trend is reversed for higher path lengths (cf. Table 7). This may be due to the fact that in a cyclic structure, the number of high length paths is limited by the size of the structure. Indeed, there is at most one path of length n in a structure comprising n atoms. While the number of paths starting at any given atom in an acyclic structure is always bounded by the number of atoms of the structure, the number of paths may increase exponentially with the path length in a cyclic graph. Thus, high path length connectivity indices should do a better job discriminating for cyclic structures than for alkanes. For the same reason the total Wiener index (cf. *Wt* in Table 9), which is half of the sum of the length of all paths, is less degenerate for cyclic structures than alkanes. Note that the total Wiener index degeneracy distribution for alkanes and alcohols is identical to the Wiener index distribution (cf. *W* in Table 9). Indeed, for these two series the total Wiener index should be identical to the Wiener index because all paths are short paths in acyclic graphs. The above trend is not observed for the shape indices since their path length is limited to 3. Furthermore, the shape indices have a higher degeneracy for fullerenes than for alkanes and alcohols (cf. Table 8). In all, information indices are more degenerate for fullerenes than for alkanes (cf. Table 10). Finally, as with the alkane and alcohol series, the total topological indices τ and τ_{ets} give the overall lowest degeneracy excluding signature (cf. Table 11).

(4) *Peptide Series.* Let us recall that the signature degeneracy for peptides was computed using the enumeration algorithm outlined in the previous section on a data set of 450 peptide sequences. For each peptide tested, sequences having the same signature were searched in the entire space of all possible peptide sequences. This search space most certainly contains sequences that do not occur in living organisms or that have not yet been synthesized. It is then remarkable to notice that height-2 signatures uniquely identifies 95.1% of all tested peptide sequences (cf. Table 6). Full nondegeneracy is obtained for height 7, but for height 5 and above only one signature leads to several sequences. More precisely, four sequences correspond to the degenerate signature, each sequence is composed of 312 amino acids. Considering that the sizes of the tested sequences range between 18 and 4680 amino acids, signature degeneracy does not appear to increase with the size of the tested peptide. Borrowing Pevzner's arguments for DNA sequence reconstruction,³³ we provide next a simple explanation of the signature degeneracy behavior for peptides.

Let us recall that a height h peptide signature is a vector of occurrence numbers of height h atomic signatures. The atomic h -signature of a given amino acid a , $^h\sigma(a)$, is a string composed of the h amino acids preceding a in the sequence, the amino acid a , and the string of h amino acids following a in the sequence. Thus, the atomic h -signature of an amino acid can be presented as a string of $l = 2h+1$ characters over the amino acid alphabet. To simplify our calculations, we now assume that the amino acids of the sequence are independent and identically distributed with probability $p = 1/20$. According to Pevzner, the most likely cause of

nonunique reconstruction for a peptide or a DNA fragment of length n is the presence of interleaved pairs of repeated $l-1 = 2h$ character strings. The expected number of such repeats is $\binom{n}{2}(1-p)p^{2h}$. Solving $\binom{n}{2}(1-p)p^{2h} = 1$, leads to a rough estimates $h = \frac{1}{2} \log_{1/p} (n(n-1)(1-p)/2)$. For n ranging between 100 and 5000 we find h values between 1.9 and 3.2. Thus, signatures of heights 2, 3, and above should uniquely characterize any peptide comprising up to 5000 amino acids.

Signature appears to be a powerful tool to identify and reconstruct protein sequences. To use signature with a practical experimental setup one needs to develop analytical tools that could recover for each amino acid in the sequence its predecessors and successors. Such tools exist (NMR, Edman degradation)³⁴ but are not high-throughput; however, recent advances in peptide digest coupled with MS and MS/MS³⁴ may be able to generate the input necessary for signature usage.

CONCLUDING REMARKS

Is the problem of finding a TI, which would be nondegenerate an open problem? We do not believe so. In fact, molecular descriptors uniquely characterizing a molecular structure have existed in the open literature for nearly three decades. For instance, the Kudo and Sasaki connectivity stack³⁵ published in 1974 is a nondegenerate descriptor. The descriptor is an integer, in binary form, compiled from the upper triangle of the adjacency matrix of the canonical representation of the molecular graph. Similarly, the list of vertex numbers taken from the list of edges of any canonical molecular graph is also a nondegenerate descriptor. Finally, as we have shown earlier molecular signatures of height $D+1$ are nondegenerate representation of all molecular graphs of diameter D . As Balaban points out,¹ a TI should also show high correlation ability, and here is the open problem Balaban is talking about: can we find a TI that is both nondegenerate and suitable for QSAR and QSPR applications. The authors of this paper are not aware of any QSAR or QSPR studies involving connectivity stacks, and for realistic molecular sizes, performing a QSAR and QSPR using height $D+1$ signatures would be cumbersome and lead to an overfit of the data considering the sheer number of parameters (e.g., atomic signatures) involved. While we have not provided an answer to Balaban's question, we have proposed a descriptor with QSAR/QSPR correlation abilities, from which other descriptors can be computed and, most importantly, from which molecular structures can be enumerated.

Nondegeneracy may not be what is expected of a descriptor when used in QSAR/QSPR analyses. Compelling evidences of this are the usage of the highly degenerate connectivity indices. Typical QSARs and QSPRs often comprise descriptors with different degeneracy profiles. Thus, having control over the degeneracy of the descriptors used in a QSAR/QSPR analysis is certainly an advantage. As clearly depicted in Figure 7, by changing the signature height one can increase or decrease the degeneracy of the descriptor. None of the TIs tested in this paper exhibit such a behavior.

Programs to compute signatures and to generate molecular structures matching signatures for various i/o formats are available from the authors upon request.

APPENDIX

We provide here the proof of Proposition 4. Precisely, we show that considering two nonisomorphic graphs $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ both saturating the same orbit of graph $G = (V, E)$, no extension of G_1 can be found isomorphic to an extension of G_2 . To prove this assertion we define the notion of ideal graph and need to prove one additional lemma.

Ideal Graph. Let $G = (V, E)$ be a molecular graph. G is ideal if for any isomorphism, π , between two extensions of G , G is invariant: $\pi(E) = E$.

Note that if π is an isomorphism between two extensions of $G = (V, E)$, an ideal graph, then for any vertex x of V , x and $\pi(x)$ belong to the same orbit in G since $\pi(E) = E$.

Lemma. Let $G' = (V, E')$ be an extension of the ideal graph $G = (V, E)$ and let x be an unsaturated vertex in G . Assume G' saturates the orbit of x , $o(x)$, then G' is ideal.

Proof. Let G'_1 and G'_2 be two extensions of G' isomorphic by π . G is ideal therefore $\pi(E) = E$. Let $e = [x', x'']$ be an edge of $E' - E$. G' saturates $o(x)$, x' or x'' are in the orbit of x , arbitrarily we take x' . Now, x' and $\pi(x')$ belong to the same orbit in G since G is ideal and $\pi(e)$ has an extremity (e.g., x') in $o(x)$, thus $\pi(e)$ belongs to E' . Consequently $\pi(E' - E) \subseteq E'$. Since $\pi(E) = E$, we have $\pi(E') = E'$.

Proof of Proposition 4. Let $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ be two nonisomorphic extensions of $G = (V, E)$ saturating the orbit of vertex x , $o(x)$. Let $G'_1 = (V, E'_1)$ be an extension of G_1 , and $G'_2 = (V, E'_2)$ an extension of G_2 we prove the contradiction that if G'_1 and G'_2 are isomorphic by π , then G_1 and G_2 are also isomorphic by π .

We first notice that G , G_1 , and G_2 are ideal graphs. Indeed, the initial graph containing no edge is obviously ideal and any graph produced by saturating an orbit of the initial graph is also ideal. Now, according to the lemma, any graph obtained by saturating an orbit of an ideal graph is ideal, therefore, G , G_1 , and G_2 are ideal graphs.

Let $e = [x', x'']$ be an edge of $E_1 - E$. Because G_2 also saturates $o(x)$, $\pi(e)$ has an extremity in $o(x)$ and therefore belongs to E_2 , thus $\pi(E_1 - E) \subseteq E_2$. For the same reason every edge of $E_2 - E$ is an image of an edge of E_1 by π^{-1} . Therefore π is an isomorphism between G_1 and G_2 .

ACKNOWLEDGMENT

Funding for this work was provided by the U.S. Department of Energy and Sandia National Laboratories under Grant number DE-AC04-76DP00789. J.L.F. is also pleased to acknowledge funding provided by the Math Information and Computer Science program of the U.S. Department of Energy.

REFERENCES AND NOTES

- (1) Balaban, A. Chemical Graphs: Looking Back and Glimpsing Ahead. *J. Chem. Inf. Comput. Sci.* **1995**, 35, 339–350.
- (2) Faulon, J.-L.; Visco, D. P., Jr.; Pophale, R. S. The Signature Molecular Descriptor. 1. Extended Valence Sequences and Topological Indices. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 707–720.
- (3) Venkatasubramanian, V.; Chn, K.; Caruthers, J. M. Evolutionary Design of Molecules with Desired Properties. *J. Chem. Inf. Comput. Sci.* **1995**, 35, 188–195.
- (4) Sheridan, R. P.; Kearsley, S. K. Using the Genetic Algorithm To Suggest Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* **1995**, 35, 310–320.

- (5) Kvasnicka, V.; Pospichal, J. Simulated Annealing Construction of Molecular Graphs with Required Properties. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 516–526.
- (6) Faulon, J.-L. Stochastic generator of chemical structure. 2. Using simulated annealing to search the space of constitutional isomers. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 731–740.
- (7) Kier, L. B.; Hall, L. H.; Frazer, J. W. Design of Molecules from Quantitative Structure–Activity Relationship Models. 1. Information Transfert between Path and Vertex Degree Counts. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 143–147.
- (8) Hall, L. H.; Kier, L. B.; Frazer, J. W. Design of Molecules from Quantitative Structure–Activity Relationship Models. 2. Derivation and Proof of Information Transfert Relating Equations. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 148–152.
- (9) Hall, L. H.; Dailey, R. S.; Kier, L. B. Design of Molecules from Quantitative Structure–Activity Relationship Models. 3. Role of Higher Order Path Counts: Path 3. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 598–603.
- (10) Skvortsova, M. I.; Baskin, I. I.; Slovokhotova, O. L.; Palyulin, V. A.; Zefirov, N. S. Inverse problem in QSAR/QSPR studies for the case of topological indices characterizing molecular shape (Kier indices). *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 630–634.
- (11) Faulon, J.-L. On using graph-equivalent classes for the structure elucidation of large molecules. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 338–348.
- (12) Colbourn, C. J.; Read, R. C. Orderly Algorithms for the Generating Restricted Classes of Graphs. *J. Graph Theory* **1979**, *3*, 187–195.
- (13) Kucera, L. *Combinatorial algorithms*; Adam Hilger: Bristol, 1990.
- (14) Faulon, J. L. Isomorphism, Automorphism Partitioning, and Canonical Labeling Can Be Solved in Polynomial-Time for Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 432–444.
- (15) Hopcroft, J. E.; Tarjan, R. E. *Isomorphism of planar graphs*, In *Complexity of Computer Computation*; Miller, R., Thatcher, E., Eds.; Plenum: New York, 1972; pp 131–152.
- (16) McKay, B. D. *Nauty User's Guide Version 1.5*; 2002.
- (17) Read, R. C. Everyone a Winner or How to Avoid Isomorphism When Cataloguing Combinatorial Configurations. *Annals Discrete Mathematics* **1978**, *2*, 107–120.
- (18) Goldberg, L. A. Efficient Algorithms for Listing Unlabeled Graphs. *J. Algorithms* **1992**, *13*, 128–143.
- (19) Thompson, W. J.; Fitzgerald, P. M. D.; Holloway, M. K.; Emini, E. A.; Darke, P. L.; McKeever, B. M.; Schleif, W. A.; Quintero, J. C.; Zugay, J.; Tucker, T. J.; Schwering, J. E.; Homnick, C. F.; Nunberg, J.; Springer, J. P.; Huff, J. R. Synthesis and antiviral activity of a series of HIV-1 protease inhibitors with functionality tethered to P1 or P1' phenyl substituents: X-ray crystal structure assisted design. *J. Med. Chem.* **1992**, *35*, 1685–1701.
- (20) Gutman, I.; Luo, Y. L.; Lee, S. I. The Mean Isomer Degeneracy of the Wiener Index. *J. Chin. Chem. Soc.* **1993**, *40*, 195–198.
- (21) Demirev, P. A.; Dyulgerov, A. S.; Bangov, I. P. CTI: A Novel Charge-Related Topological Index with Low Degeneracy. *J. Math. Chem.* **1991**, *8*, 367–382.
- (22) Randic, M.; Mihalic, Z.; Nikolic, S.; Trinajstić, N. Graphical Bond Orders: Novel Structural Descriptors. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 403–409.
- (23) Estrada, E. Graph Theoretical Invariant of Randić Revisited. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1022–1025.
- (24) Ivanciuc, O.; Laidboeur, T.; Cabrol-Bass, D. Degeneracy of Topological Distance Descriptors for Cubic Molecular Graphs: Examples of Small Fullerenes. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 485–488.
- (25) Balaban, A. Local versus Global Numerical Modeling of Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 398–402.
- (26) Balaban, A. Real-number Local (Atomic) Invariants and Global (Molecular) Topological Indices. *Revue Roumaine Chim.* **1994**, *39*, 245–257.
- (27) Trinajstić, N. *Chemical Graph Theory*, 2nd ed. In *Mathematical Chemistry*; Klein, D. J., Randić, M., Eds.; CRC Press: Boca Raton, FL, 1992.
- (28) Meringer, M. Fast Generation of Regular Graphs and Construction of Cages. *J. Graph Theory* **1999**, *30*, 137–146.
- (29) Holm, L.; Sander, C. The FSSP database of structurally aligned protein fold families. *Nucl. Acid Res.* **1994**, *22*, 3600–3609.
- (30) Holm, L.; Ouzounis, C.; Sander, C.; Tuparev, G.; Vriend, G. A database of protein structure families with common motifs. *Protein Sci.* **1992**, *1*, 1691–1998.
- (31) Hall, L. H. *MOLCONN-Z*; Hall Associates Consulting: Quincy, MA, 1991.
- (32) Jerrum, M. R.; Shyrum, S. Families of fixed degree graphs for processor interconnection. *IEEE Trans. Comput.* **1984**, *33*, 190–194.
- (33) Pevzner, P. A. *Computational Molecular Biology. An Algorithmic Approach*; MIT Press: Cambridge, MA, 2000; pp 74–75.
- (34) Aebersold, R.; Goodlett, D. R. Mass Spectrometry in Proteomics. *Chem. Rev.* **2001**, *101*, 269–295.
- (35) Kudo, Y.; Sasaki, S.-I. The Connectivity Stack, a New Format for Representation of Organic Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1974**, *14*, 200–202.
- (36) Mayo, S. L.; Olafson, B. P.; Goddard, I. W. A. Dreiding: A generic force field. *J. Phys. Chem.* **1990**, *94*, 8897–8909.

CI0203460