Journal of Computer-Aided Molecular Design (2005) 19: 637–650
DOI 10.1007/s10822-005-9007-1

Reverse engineering chemical structures from molecular descriptors: how many solutions?

Jean-Loup Faulon[a,*], W. Michael Brown[b] & Shawn Martin[b]
[a]*Computational Bioscience, Sandia National Laboratories, 969, Livermore, CA, 94551-9292, USA;*
[b]*Computational Biology, Sandia National Laboratories, 5800, Albuquerque, NM, 87185-0310, USA*

Received 12 May 2005; accepted 28 July 2005

Key words: enumeration, molecular fragments, molecular design, structure–properties relationships, topological indices

Summary

Physical, chemical and biological properties are the ultimate information of interest for chemical compounds. Molecular descriptors that map structural information to activities and properties are obvious candidates for information sharing. In this paper, we consider the feasibility of using molecular descriptors to safely exchange chemical information in such a way that the original chemical structures cannot be reverse engineered. To investigate the safety of sharing such descriptors, we compute the degeneracy (the number of structure matching a descriptor value) of several 2D descriptors, and use various methods to search for and reverse engineer structures. We examine degeneracy in the entire chemical space taking descriptors values from the alkane isomer series and the PubChem database. We further use a stochastic search to retrieve structures matching specific topological index values. Finally, we investigate the safety of exchanging of fragmental descriptors using deterministic enumeration.

Introduction

Drug companies and other organizations often want to exchange information about the physical, chemical, and biological properties of different compounds without actually providing the structure of the compounds under consideration. One obvious way to exchange such information is via molecular descriptors that map structural information to activities and other properties. Is the exchange of such descriptors safe? In other words, can we exchange these descriptors without inadvertently providing enough information to reverse engineer the structures? In this paper, we investigate this question in the context of some popular 2D descriptors.

For this investigation, we limit ourselves to topological indices and molecular fragments. These descriptors can be computed from the 2D molecular graph of a chemical structure. Topological indices are based largely on connectivity, distance, and information theory. In this paper, we cover these three categories as we make use of shape [1] and connectivity [2] indices, the Wiener [3] and Balaban J and J_t distance indices [4], and the Bonchev-Trinajstic information theoretic index [5]. Molecular fragments are simply lists of molecular subgraphs and have been used to predict properties and activities since the development of group contribution. Group contribution uses a precompiled list of fragments defined by an expert

*To whom correspondence should be addressed. Tel: +925-294-1279; E-mail: jfaulon@sandia.gov

for the property or activity being investigated. Recently, techniques have been proposed to systematically compile such lists without expert knowledge. Some of these systematic techniques enumerate lists of subgraphs up to a predefined size. In this category are the hologram method of Tong et al. [6], used to predict chemical binding to estrogen receptors; and the fragments proposed by Zefirov and Palyulin [7], compiled to predict retention indices and normal boiling points. Other techniques are based on atomic neighborhoods. These methods include atom environments [8], used in similarity studies; multilevel neighborhood atoms [9], used to predict biological activities in large databases [10]; and atomic and molecular signatures [11], used in QSAR and QSPR analyses [12]. Figure 1 illustrates the differences between these three descriptors. Atom environments are composed of lists of neighbors for each heavy atom, with neighbors compiled up to a predefined distance. Multilevel neighborhood atoms are similar to the feature trees used in atom environments, but the connectivity between the neighbors is stored. Signatures are similar to multilevel neighborhood atoms, but ring closure and bond order are taken into account. Furthermore, signatures are canonized and are thus canonical representations of molecular fragments [13].

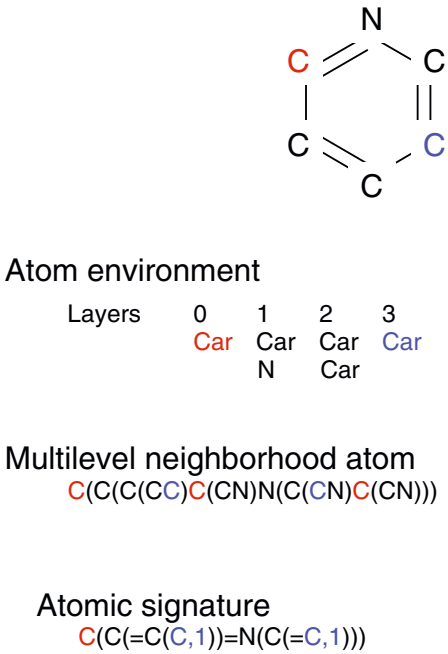

Figure 1. Differences between atom environment, multilevel neighborhood of atom, and atomic signature. In all cases the environment of the carbon atom marked in red is probed up to three bonds away (layers). For atom environment, atoms are replaced by Sybyl mol2 atom types. For multilevel neighborhoods, all the bonds between neighborhood atoms are taken into account including bonds returning to atoms already visited at a previous layer. For signature, atoms already encountered at previous layers are not repeated, bond order is taken into account, and ring closure is indicated by marking the atom where the closure occurs with a number (number 1 in this example).

From a computational point of view, it is relatively easy to check if the above descriptors are safe to exchange in the context of a database. In this case, one computes the descriptors for all compounds in the database, and then compiles the descriptor degeneracy, where the degeneracy of a descriptor is the number of structures having the same descriptor value. Clearly, descriptors having high degeneracy are safe to exchange, as many structures can be reverse engineered from the descriptor values. However, it is not necessarily true that a descriptor is unsafe to exchange if it has a low degeneracy, especially if the degeneracy was computed on a database. There are two reasons for this phenomenon. First, even if a descriptor has a low degeneracy, there is no general algorithm to reverse engineer structures from descriptor values, although such algorithms are available for few descriptors (cf. Section Reverse engineering methods). Second, a degeneracy calculation performed on a database at best underestimates the degeneracy computed in the entire chemical space. We note that a degeneracy calculation should in fact be performed in the entire chemical space, as it is unlikely that anyone would share information about a structure that is already present in a database.

All degeneracy calculations in the present paper are carried out in the entire chemical space. We start by computing the degeneracy of various topological indices for an alkane isomer series. We next address the problem of retrieving structures in the chemical space corresponding to topological indices that cannot be reverse engineered using a deterministic algorithm. For these indices, we use a stochastic process based on simulated annealing. Finally, we deterministically reverse engineer structures matching signature fragments of various sizes. Our results are presented in Section Results and discussion and the methods used to search and reverse engineer structures are given in Section Methods.

Methods

In order to investigate the security of chemical information exchange, we employ several tools developed previously for the purpose of designing molecules matching specified chemical properties [14–16]. These tools are discussed in Section Reverse engineering methods and include a stochastic structure generator as well as a method for deterministic enumeration of molecules. The second method uses the signature molecular descriptor, which has been presented in the Introduction. Since signatures of different sizes can be exchanged, we also examine the effect of signature size on the usefulness of signature in quantitative structure-activity relationships (QSARs) and quantitative structure-property relationships (QSPRs). QSARs and QSPRs are discussed in Section QSARs.

Reverse engineering methods

Reconstructing molecules that match chemical property values is a long-standing problem. Surprisingly, there are not many reports in the literature providing solutions to this problem. Most of the proposed techniques are stochastic in nature and use genetic algorithms, evolutionary computing, or Monte Carlo methods to search for and construct chemical structures matching predefined descriptor values. The first stochastic methods proposed were based on genetic algorithms [17, 18], while methods based on Monte Carlo were reported later [15, 19]. Although many papers using stochastic techniques have appeared since then, there are still very few attempts to solve the reconstruction problem using a deterministic approach, i.e. using techniques that generate exhaustive lists of molecular structures matching predefined descriptor values. In a series of three papers Kier, Hall, and co-workers reconstructed molecular structures from the count of paths, ${}^{l}P$, up to length $l=3$ [20–22]. 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 ${}^{3}P$ count are rejected. A similar technique was used in [23], but from the count of paths an edge sequence was derived in addition to the degree sequence. The two sequences were then fed to an isomer generator that produces all the structures matching the sequences. More recently [16], a deterministic technique was proposed to enumerate structures matching a predefined list of signature molecular fragments.

In the following two subsections, we describe the techniques that we used to explore the degeneracy of various molecular descriptors. We first describe a stochastic method that searches for chemical structures matching topological indices, and we then describe the deterministic technique mentioned above that enumerates compounds matching signatures.

Stochastic search

The stochastic structure generator used here for investigation of the security of chemical information exchange is designed around a bond-switch algorithm [15]. This algorithm is based on the conservation of bonds and bond order in a molecular graph. Connectivity can be changed by deleting bonds, creating bonds, or modifying bond order. If we follow the convention that a bond is deleted when its order is set to zero, and a bond is created when its order is switched from zero to a positive value, then all changes in connectivity can be performed by modifying bond orders. The fact that all structural isomers have the same number of bonds implies that when a bond order is increased (or decreased), another bond order must be decreased (or increased). Hence changing the connectivity implies the selection of at least two bonds (four atoms) in the molecular graph. Suppose x_1, x_2, x_3, and x_4 are the four selected atoms; a_{11}, a_{12}, a_{21}, and a_{22} are the orders of the bonds $[x_1, y_1]$, $[x_1, y_2]$, $[x_2, y_1]$, and $[x_2, y_2]$ in the initial molecular graph; and b_{11}, b_{12}, b_{21}, b_{22} are the orders of the same bonds after a random displacement. Because the valences of the four selected atoms must remain constant, the following equations hold

$$\begin{aligned} b_{11} + b_{12} &= a_{11} + a_{12} \\ b_{11} + b_{21} &= a_{11} + a_{21} \\ b_{21} + b_{22} &= a_{21} + a_{22} \\ b_{12} + b_{22} &= a_{12} + a_{22} \\ 0 \le a_{ij} \le 3, 0 \le b_{ij} &\le 3 \end{aligned} \tag{1}$$

where it is assumed that the maximum bond order is 3 (triple bond). Examples of bond switching are illustrated in Figure 2. In [15] it is shown that for any given molecular formula every possible

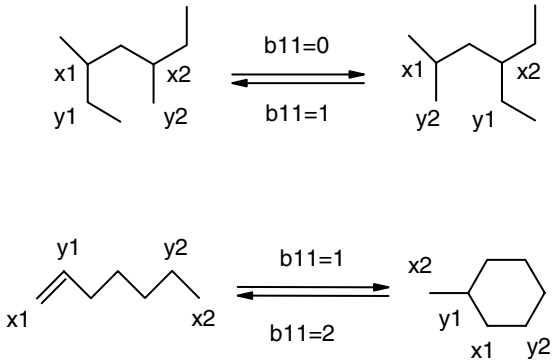

Figure 2. Bond switch examples. The two examples illustrate the technique described in Section Reverse engineering methods. In this figure, $b_{11} = \text{order}[x_1, y_1]$, $b_{12} = \text{order}[x_1, y_2]$, $b_{21} = \text{order}[x_2, y_1]$, and $b_{22} = \text{order}[x_2, y_2]$.

structural isomers can be reach by a series of bond switches. To optimize activities or properties, the bond switch algorithm is embedded in a simulated annealing algorithm. At each stage of the algorithm a random displacement is obtained by choosing a (random) value for b_{11} different from a_{11} and solving from the remaining variable subject to the inequality constraint.

Deterministic enumeration

In contrast to stochastic search, deterministic enumeration proceeds in an orderly and exhaustive manner. We have investigated two types of enumeration in previous studies. In [14] we derive a linear system of Diophantine equations whose solutions correspond to all possible molecular signatures that correspond to molecules. In [16] we describe an algorithm, which enumerates all possible molecules that can generate a given molecular signature. In the present study, we use only the second method, as we are interested in the possibility of reconstructing a molecule when a molecular signature is exchanged. A brief overview of the enumeration routine is given here, but the reader is referred to the previous paper [16] for a more detailed description of the algorithm.

Starting with a molecular graph, G, composed of isolated vertices and no edges, the edges are added in every possible combination to produce all non-isomorphic saturated graphs matching the molecular signature. There are two primary steps: (1) determine the orbits or atoms with equivalent atomic signatures of G, and (2) saturate each atom of a chosen orbit. Once the orbits have been defined, then one is selected that contains unsaturated vertices and is saturated in an orderly manner. This process is repeated until all the vertices have been saturated and the resulting bonds are compatible with the target signatures and it does not create a saturated subgraph of G. This algorithm was proven to be exhaustive and irredundant, meaning all solutions are produced and no two structures are identical (i.e., isomorphic).

QSARs

When reverse engineering molecules from QSAR and QSPR descriptors, it is important to decide which descriptors are generally useful for molecular property prediction and therefore warrant analysis. In the case of signature, we must choose the appropriate choice of signature height. While it is not straightforward to decide which signature heights will be optimum without consideration of a particular molecular property, there are two general qualities of molecular descriptors which are desirable in developing QSARs – descriptor variability and descriptor correlation. That is, for a property that varies with molecular structure, the descriptors for this structure should also vary in order to make prediction feasible. Additionally, the correlation between descriptors should be low such that there is little information overlap between descriptors, which can be selected for a given QSAR.

Shannon entropy has been suggested as a metric for descriptor variability [24] and is particularly convenient in this case due to the discrete nature of signature. The Shannon entropy of a discrete random variable X with alphabet χ and probability mass function $p(x) = \Pr\{X = x\}, x \in \chi$ is defined by [25]

$$H(X) = -\sum_{x \in \chi} p(x) \log_2 p(x). \tag{2}$$

In the case of signature, X represents a given atomic signature, $p(x)$ is the probability of the atomic signature occurring x times within each molecule in the dataset, and χ is the set of non-negative integers. For any dataset, the Shannon entropy for a descriptor will be zero when there is no variability (the atomic signature occurs in each molecule with the same count). Maximum variability, when the count of an atomic signature is different within each molecule in the dataset, will

yield a maximum Shannon entropy of $-\log_2(n^{-1})$ for a dataset of size n.

Our QSARs were trained using multiple linear regression (MLR) and support vector regression (SVR). For MLR, QSAR equations were obtained using forward stepping feature selection as implemented in the Matlab 7 [26] routine stepwisefit.m which was modified for efficiency and to prevent the addition of descriptors which would result in a QR factorization matrix that was close to singular. The predictive accuracy of QSAR equations was evaluated using cross-validation in order to prevent over fitting [27]. For the HIV and T_g datasets, leave-one out cross-validation was performed and for the log P dataset, 10-fold cross-validation was performed. For optimization of the signature height, the cross-validation squared correlation coefficient was used as the objective function. For MLR, all possible combinations of signature heights ranging from 0 to 7 and QSAR equation sizes (in terms of number of descriptors selected through forward selection) were evaluated and the signature height resulting in the highest q^2 was reported.

For SVR, the source code for SVM^{light} [28] was modified to perform leave-one-out cross-validation. The SVR was parameterized by the regularization parameter C (the trade-off between the training error and margin) and ε (the tube width for regression). In order to automate parameterization, evolutionary pattern search (as implemented in SGOPT [29]) was performed using a population size of 25 and a cross-over probability of 0.8 using a domain [0,1000] for C and [0.01,2] for ε. Five trials of evolutionary pattern search were performed using the cross-validation q^2 as the objective function. The best signature height from the five trials was taken as the optimum.

Results and discussion

We investigated the security of descriptor exchange using a variety of topological indices and descriptors on a number of datasets. Descriptors examined included the connectivity indices ${}^{0-3}\chi$; the shape indices ${}^{0-3}\kappa$; the Wiener number W; the Platt number P_f; the total topological index τ (also called TOTOP); the Shannon information SI; and the Bonchev-Trinajstic index I_D^W; the Balaban J and J_t indices; and the signature molecular descriptor. All descriptors, except Balaban indices and signature, were computed using the Molconn-Z software package (http://www.eslc.vabiotech.com/molconn). Datasets examined included alkane isomer series; compounds taken from the PubChem database; HIV-1 protease inhibitors; organics with log P measurements available; and polymer glass transition temperatures. Each of these datasets will be described in greater detail when they are introduced.

Degeneracy of topological indices for isomer series

We first investigated the degeneracy of the topological indices computed using Molconn-Z on the alkane isomer series. All alkanes up to 16 carbon atoms were generated using an isomer generator based on graph equivalence classes [30], resulting in 18,030 compounds. For each compound in this isomer series, we computed the degeneracy of that compound with respect to each topological index computed using Molconn-Z. The degeneracy of a given compound was computed by counting the number of compounds with the same value of the topological index under consideration. From the point of view chemical information exchange, a descriptor or topological index with low degeneracy is unsafe to exchange, since a small number of compounds will correspond to a given value of the descriptor. We rather arbitrarily defined a descriptor to be unsafe if more than 25% of compounds had degeneracy up to 10, that is, at most 10 structures corresponded to the descriptor value.

Our degeneracy calculations are given in Tables 1, 2, 3. These tables reveal that many popular descriptors are safe to exchange, including ${}^0\chi$, ${}^{0-3}\kappa$, W, P_f, and SI, but that some descriptors such as Kier and Hall total topological index τ, and the Bonchev-Trinajstic index I_D^W are not degenerate and are therefore unsafe to exchange. These results are not new, although they have not been interpreted in terms of security of information exchange. The reader interested by degeneracy of other specific indices is reported to our earlier paper [12], where the degeneracy of about 50 topological indices are examined for various isomer series including alkanes, alcohols, and fullerenes.

We next considered combinations of descriptors. While an individual descriptor may be degenerate, we find in Tables 1–3 that combinations of

Table 1. Degeneracy of connectivity indices for alkanes up to C_{16} (18,030 compounds).

Degeneracy	${}^0\chi$	${}^1\chi$	${}^2\chi$	${}^3\chi$	${}^0\chi+\ldots+{}^3\chi$
1	0.1	2.1	26.6	42.8	77.0
10	1.2	32.9	69.3	57.1	13.0
100	9.9	61.2	4.1	0.1	0.0
1000	64.4	3.7	0.0	0.0	0.0
10,000	24.4	0.0	0.0	0.0	0.0
100,000	0.0	0.0	0.0	0.0	0.0

In the last column the degeneracy is given for the combination of all previous connectivity indices.

highly degenerate descriptors may again have low degeneracy. This implies that while certain descriptors may be safe to exchange by themselves, they are unsafe to exchange in combinations. This phenomenon is best exemplified by W and SI in Table 3. In this case we see that 5% of alkanes have a degeneracy lower than 10 for each W and SI but that more than 35% of alkanes have a degeneracy lower than 10 for the combination $W+SI$.

Stochastic search for chemical structures matching molecular descriptors

Our initial study on isomer series reveals that there are descriptors or combinations of descriptors with low degeneracy, and that these descriptors or combinations may be unsafe to exchange. Even if a low degeneracy descriptor is exchanged, is it really possible to reverse engineer the corresponding structure? As mentioned in Section Reverse engineering methods, there is no general algorithm that can enumerate all possible structures corresponding to any descriptor. However, in most cases stochastic algorithms may still be used to search the chemical space. In this section, we investigate the use of such an algorithm in combination with the PubChem database.

Precisely, we used PubChem (http://pubchem.ncbi.nlm.nih.gov/) [31] to investigate the ability of the bond-switch search algorithm introduced in Section Reverse engineering methods to uncover structures matching molecular formulas, Wiener indices, and Balaban J and J_t indices [4]. These two last indices where selected because there have been successfully used for property activity predictions and are known to have low degeneracy. We recall the expression of J and J_t:

$$J = \frac{m}{\mu+1}\sum_{[x,y]} s_x s_y, \text{ where } s_x = \sum_y d(x,y), \quad (3)$$

where m is the number of bonds, μ the number of independent cycles, and $d(x,y)$ is the shortest path (number of bonds) between atoms x and y in the molecular graph.

$$J_t = \frac{m}{\mu+1}\sum_{[x,y]} t_x t_y, \text{ where } t_x = s_x/\deg(x), \quad (4)$$

where $\deg(x)$ is the number of bonds of atom x.

As of January 2005, PubChem contained 741,361 substances, with 529,542 unique two-dimensional structures in canonical SMILES format. From these structures we selected 1000 at random, each having less than 200 atoms. Molecular formulas and indices were computed on carbon skeleton structures, excluding heteroatoms and multiple bonds. The bond switch algorithm was then executed with the goal of matching the provided molecular formula, Wiener W, and Balaban J and J_t indices. The algorithm was run for 20,000 and 50,000 steps. In the case of 20,000 steps, the algorithm used 4 simulated annealing schedules with an initial temperature of 1000, a final temperature of 0, and a decrease in temperature of 10 at each step. In the case of 50,000 steps, the algorithm used 10 simulated annealing schedules. Our results are shown in Figure 3. In 1000 trials of the algorithm running for 20,000 steps, a structure was found 99.3% of the time matching W, 40.9% of the time matching J, and 9.9% of the time matching J_t. The differences between our results using W, J and J_t is explained by looking at the number of times the initial PubChem structures were retrieved by the algorithm. In the case of W the initial structure was retrieved only 0.8% of the time, which indicates that while it is easy to find a structure matching a given W value, the chance of finding the correct structure is small, this confirm the results given in Table 3 where W is found to be an highly degenerated index. In the case of J, the initial structures were retrieved 8.9% of the rime and in the case of J_t, the structures were recovered 8.6% of the time. Clearly, there are more structures matching J than J_t, so that J_t is

Table 2. Degeneracy of Kappa shape indices for alkanes up to C_{16} (18,030 compounds).

Degeneracy	${}^0\kappa$	${}^1\kappa$	${}^2\kappa$	${}^3\kappa$	${}^0\kappa+\ldots+{}^3\kappa$
1	0.4	0.0	0.1	0.0	9.3
10	2.4	0.1	0.8	1.1	31.9
100	13.4	0.7	5.9	9.7	54.9
1000	55.0	7.3	46.7	64.0	3.9
10,000	28.8	34.4	46.6	25.1	0.0
100,000	0.0	57.5	0.0	0.0	0.0

In the last column the degeneracy is given for the combination of all previous shape indices.

Table 3. Degeneracy of distance and information theory indices for alkanes up to C_{16} (18,030 compounds).

Degeneracy	W	P_f	τ (TOPOP)	SI	I_d^W	$W+SI$
1	0.4	0.0	98.5	0.3	81.1	5.9
10	4.9	0.2	1.5	2.4	18.9	29.1
100	59.4	0.9	0.0	13.1	0.0	63.0
1000	35.3	13.8	0.0	55.5	0.0	1.9
100,00	0.0	85.1	0.0	28.8	0.0	0.0
100,000	0.0	0.0	0.0	0.0	0.0	0.0

W is the Wiener index, P_f is the Platt number, τ is the Kier and Hall total topological index, SI is the Shannon entropy index and I_d^W is the Bonchev-Trinajtic information index. In the last column the degeneracy is given for the combination of Wiener and Shannon indices.

less degenerate than J. Having a low degenerated index was in fact Balaban's intention when developing J_t. Despite the differences between J and J_t, almost 9% of the time the initial PubChem structures were retrieved, and this number rose to 12.2% when the algorithm was run for 50,000 steps.

The bond switch algorithm uses as input a molecular formula. In the absence of such information, the algorithm can be run on a series of possible formulas. In the case of shape indices, for example, it is possible to derive formulas from the index values [20–22]. For others indices, series of potential formulas can be compiled from large databases, such as PubChem. As discussed in the next section and depicted in Figure 4, about 24% of compounds in PubChem have a unique height-0 signature, and consequently a unique molecular formula. Once removed from PubChem, these compounds cannot be retrieved from their index values using the molecular formulas remaining in PubChem. Conversely, all other compounds (76% of PubChem) can potentially be reverse-engineered. According to the results we obtained with the bond switch algorithm using 50,000 steps, there is a 12.2% chance of finding the correct structure from J_t values for 76% of PubChem compounds. This accuracy yields an

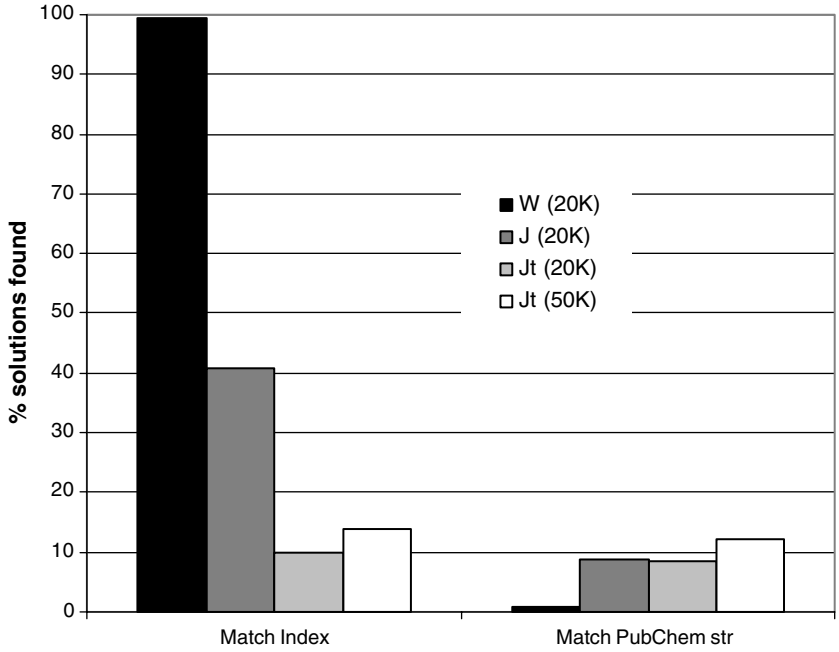

Figure 3. Searching structures matching 1000 molecular formulas Wiener W, and Balaban's J and J_t indices. Structures were searched using simulated annealing for 20,000 steps and 50,000 steps. The plots show the percentages of correct solutions found matching the given index values (left), and matching the PubChem structures for which the indices were computed (right).

overall rate of success just above 9% for correctly retrieving any PubChem compound from its J_t value without a molecular formula. However, to achieve the above rate of success, computing power is required. The bond switch algorithm takes about 4 min of CPU time to process 50,000 steps on a SGI O2 R10000 workstation, and PubChem is currently composed of about 200,000 different molecular formulas. Therefore, retrieving a structure from its J_t value requires 639 days of CPU time on a single processor. Note, however, that each potential molecular formula can be sent to a different processor without communicating the results to other processors, so that a maximum parallel speedup can be achieved, and structures can be retrieved using (for instance) 100 processors in less than a week of computing time.

Deterministic enumeration of chemical structures matching molecular descriptors

In addition to stochastic search, a deterministic algorithm can be used with the signature descriptor to reverse engineer structures (see Section 2.1). In this section, we investigate the security of exchanging the signature descriptor in terms of reverse engineering the original structures. We find that small fragments are safe to exchange and further investigate the utility of these fragments in terms of correlating the signature descriptor with chemical properties using QSARs.

Degeneracy of signature in PubChem

The degeneracy of the signature molecular descriptor within known compounds was evaluated using the January 2005 release of the PubChem database [31]. As mentioned in Section Stochastic search for chemical structures matching molecular descriptors, canonization of the database using signature resulted in 529,544 unique molecules. Signatures were calculated for each molecule at heights 0–3. Height 0 calculations produced 291 atomic signatures with each signature occurring in an average of 16,755 molecules. Height 1 calculations produced 5783 atomic signatures (average of 12,187 molecules per signature), height 2 produced 98,936 atomic signatures (average 5017 molecules per signature), and height 3 calculations produced 528,384 atomic signatures.

The number of atomic signatures calculated increases asymptotically to the limit reached when the signature height is equal to the largest molecule in the dataset. The increase in the number of signatures is consistent with a decrease in the degeneracy of the atomic signatures, and likewise, a decrease in the degeneracy of the molecular signatures generated within PubChem. Figure 4 illustrates the molecular signature degeneracy for the first four signature heights. At height 2, the percentage of molecules with unique signatures was found to be above 98% and at height 3 it was found to be 99.62%. A very small percentage of molecules were found to be highly degenerate in terms of signature. In fact, two molecules were degenerate up to a height of 18 (PubChem CIDs 174042 and 174046).

In general, however, for height 2 and 3 QSPRs, the reverse-engineering problem within the known molecules in the PubChem database reduces to a simple database search to find the only compound matching a given signature. Figure 4 clearly demonstrates that within the current PubChem space it is unsafe to exchange molecular signature of any heights, as even with height 0 more than 25% of PubChem compounds have a signature degeneracy lower than 10, and above height 0 most compounds have a unique signature.

Degeneracy of signature in chemical space

While within a given database the size of PubChem we observe that most compounds have a unique signature, does the observation hold for the entire chemical space? The question is relevant because one wishes to exchange information about new structures; not structures already present in a database. In other words, an individual trying to reverse engineer chemical structures from a safe information exchange system will have to search the structures in the entire chemical space, not just existing databases.

To answer this question, we used the same 1000 randomly selected PubChem structures used in Section Stochastic search for chemical structures matching molecular descriptors. For each structure, molecular signatures up to height 5 where computed. For height 0, the signature degeneracy was calculated using the isomer generator based on graph equivalence classes [30]. For all other heights we ran the algorithm outlined in the section Reverse engineering methods (see also

[16]). The results shown in Figure 5 are clearly different than those obtained in Figure 4. According to our safety criteria, we find that signature of heights 0 and 1 are safe to exchange, since only 0.7% of the compounds have a height 0 molecular signature degeneracy lower than 10. This number increases to 15.9% for height 1. Furthermore, according to Table 4, safety increases with compound size as we find that signatures of height 2 are safe to exchange for compounds with more than 70 atoms.

Signature height

While signature of height 1 and possibly signature of height 2 for larger compounds may be safe to exchange is such information useful for QSAR analysis? To answer that question three datasets were used for QSAR development with signature. The first two, a dataset of 130 HIV protease inhibitors with IC_{50} values and a dataset of 12,865 molecules with octanol/water partition coefficient (log *P*) data, have been previously reported in our work [12]. For this analysis, however, a random subset consisting of 1000 molecules out of the 12,865 was used. The third dataset consisted of 262 linear homopolymers with glass transition temperature (*T*g) data compiled by Bicerano [32]. In the case of the polymer dataset, a single monomer residue from the polymer chain was used for QSAR analysis. For each molecule, signature descriptors were calculated at heights 0–7 using publicly available software (http://www.cs.sandia.gov/~jfaulon/QSAR/index.html). Any descriptors which occurred in less than three molecules or that were found to be perfectly correlated with another descriptor were deleted to give a set of descriptors available for feature selection as shown in Table 5. QSARs were trained using multiple linear regression (MLR) and support vector regression (SVR) as described in section QSARs.

To analyze descriptor correlation and variability in terms of signature height, we first plotted the mean Shannon entropy of the calculated signature descriptors as a function of signature height is shown in Figure 6a for each of the data sets. The results show a decrease in the mean entropy with increasing signature height – the lower the height, the better the variability. While the mean entropy can be misleading (only a few descriptors with high variability might be necessary for a QSAR), we also observed a consistent decrease in the maximum descriptor entropy above height 1. An opposite trend, in terms of appropriate signature height, is observed when the mean pairwise correlation coefficient is calculated for the dataset (Figure 6b). As the descriptor height increases, the occurrence of a given descriptor becomes rare within a dataset and is of little use in developing QSARs. The results show that the selection of

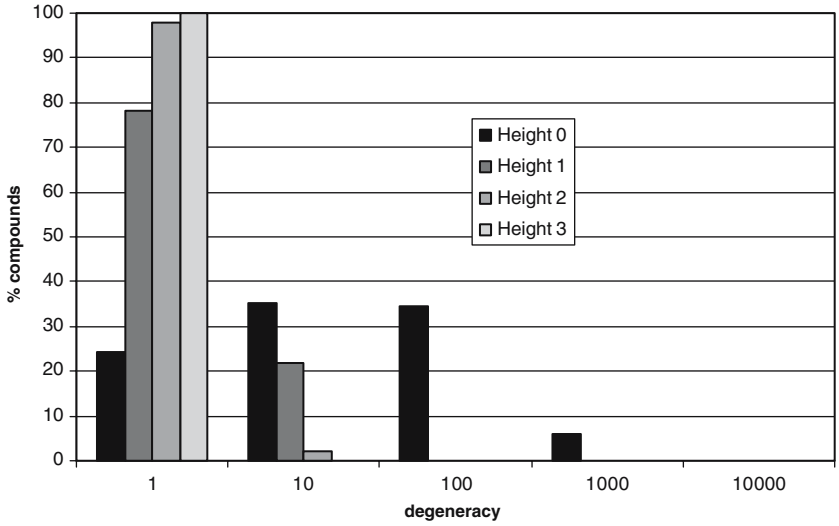

Figure 4. Molecular Signature Degeneracy within the PubChem Database. Degeneracy was calculated for 529,544 structurally unique compounds.

general signature heights for developing QSARs will result from competing effects of a high variation at low signature heights and a low correlation at high signature heights.

When specific properties are concerned, we can analyze the signature heights that balance these effects through optimization of signature height in QSAR regressions. Analyzing the three datasets mentioned above, we can obtain the optimum signature height as the height which produces the highest predictive accuracy for a QSAR as measured using a cross-validation squared correlation coefficient (q^2). For multiple linear regression (MLR), this is performed by simultaneous optimization of both the signature height and the number of descriptors selected by forward stepping – through a brute force evaluation of all possible pairs up to some limit (see Section QSARs). A characteristic plot of this optimization is shown in Figure 7 for the HIV dataset. The results from all three datasets are shown in Table 5. In all cases, the optimum height is found to be within 1–3. We also tested this effect using support vector regression (SVR), which has received recent attention in the QSAR literature [33], on the polymer dataset. In this case, the optimum height is found through simultaneous optimization of the signature height, the support vector regularization parameter (C), and the support vector tube width (ε) for regression (for details on these parameters and support vector machines see [28]). Here, we used evolutionary

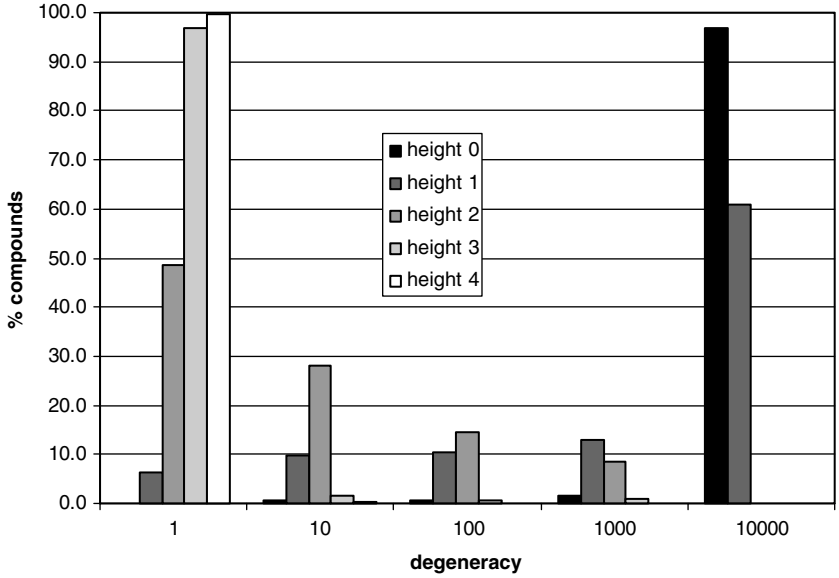

Figure 5. Molecular Signature Degeneracy within the Chemical Space. Degeneracy was calculated for 1000 compounds.

Table 4. Signature degeneracy versus compound size.

Degeneracy	Height 1				Height 2			
	Size > 0	Size > 30	Size > 50	Size > 70	Size > 0	Size > 30	Size > 50	Size > 70
1	6.2	0.9	1.0	0.0	48.6	31.6	13.8	9.3
10	9.7	2.8	0.5	0.0	28.0	33.9	28.2	13.3
100	10.3	3.3	1.0	0.0	14.6	17.1	23.1	20.0
1000	13.0	10.2	7.7	6.7	8.6	11.4	20.0	28.0
10,000	60.9	82.9	89.7	93.3	0.1	6.0	14.9	29.3

Degeneracy was calculated for height 1 and 2 signature computed on 1000 random structures extracted from PubChem. Among the set of 1000 selected structures, 579 compounds had more than 30 atoms, 195 more than 50, and 75 more than 70.

pattern search [34, 35] for optimization, again using q^2 as the objective function. In this case, the optimum signature height is found to be consistent with the results obtained from MLR (Table 5).

In addition to the datasets presented here, our overall experience with the signature descriptor has resulted in optimum heights lying in the range 1–3. This is consistent with reports describing the use of other fragmental descriptors such as atom environments [8, 36], where optimal heights have been reported as 2–3, and multilevel neighborhoods of atoms [9, 10], with reported optimal heights in the range 1–2. While there is some tolerance for a trade-off between reverse engineering security and signature height, moving far outside the optimal height range will result in poor predictive accuracy for QSARs in most cases.

Conclusion

Sharing useful information about chemical compounds without revealing their structures is a challenging problem. On one hand, we would like to share enough information such that meaningful properties and activities can be predicted. On the other hand, as we have seen in this paper, too much information may reveal the structure.

We have used degeneracy to examine the safety of information exchange. Degeneracy is the number of compounds matching the shared information. Highly degenerate information is safe to exchange, while slightly degenerate information is not. The degeneracy must be computed in the chemical space instead of an existing database, as it is unlikely that anyone would want to provide information about a compound already stored in a database. Furthermore, as demonstrated in Figures 4 and 5, degeneracy computed from databases can be greatly underestimated.

According to our somewhat arbitrary safety criteria (less than 25% of compounds have a degeneracy $\leq$10), we found that many individual topological indices are safe to exchange. However, when several topological indices are shared for the same structure, we observed that the combination of these indices may be unsafe. Being able to share several topological indices is important because QSAR developers use their own favorite descriptors, and also because QSARs often make use of multiple descriptors to obtain good predictive ability.

While individual indices and combination of indices may be unsafe to exchange, in most instances we do not have a systematic technique to reverse engineer structures. This good news is mitigated by the fact that stochastic techniques can always be utilized to search chemical structures matching descriptor values. While the convergence of these search techniques is not guaranteed, we nonetheless found random structures from PubChem were correctly reverse engineered 10% of

Table 5. Optimum QSAR predictive accuracy as a function of signature height using multiple linear regression for all three datasets and support vector regression for the polymer dataset.

HIV-1 Protease Inhibitors IC_{50}, 130 compounds in training set								
Height	0	1	2	3	4	5	6	7
Total Descriptors	8	74	385	1009	1663	2184	2666	3143
Descriptors Used for Forward Selection	6	48	142	186	158	142	119	108
q^2 (MLR)	0.75	0.80	0.77	0.81	0.77	0.69	0.65	0.66
Log P, 1,000 compounds in training set								
Total Descriptors	9		3424	9652	15516	18832	20626	21629
Descriptors Used for Forward Selection	9	192	935	1130	768	434	229	122
q^2 (MLR)	0.61	0.83	0.78	0.58	0.31	0.23	0.16	0.11
Glass Transition (T_g), 262 polymers in training set								
Total Descriptors	8	83	496	1325	2228	2926	3447	3813
Descriptors Used for Forward Selection	8	50	164	233	212	158	107	71
Q^2 (MLR)	0.65	0.78	0.75	0.81	0.65	0.60	0.44	0.32
q^2 (Linear SVR)	0.64	0.80	0.81	0.81	0.73	0.22	0.29	0

In all cases the optimum height is found to be within 1–3.

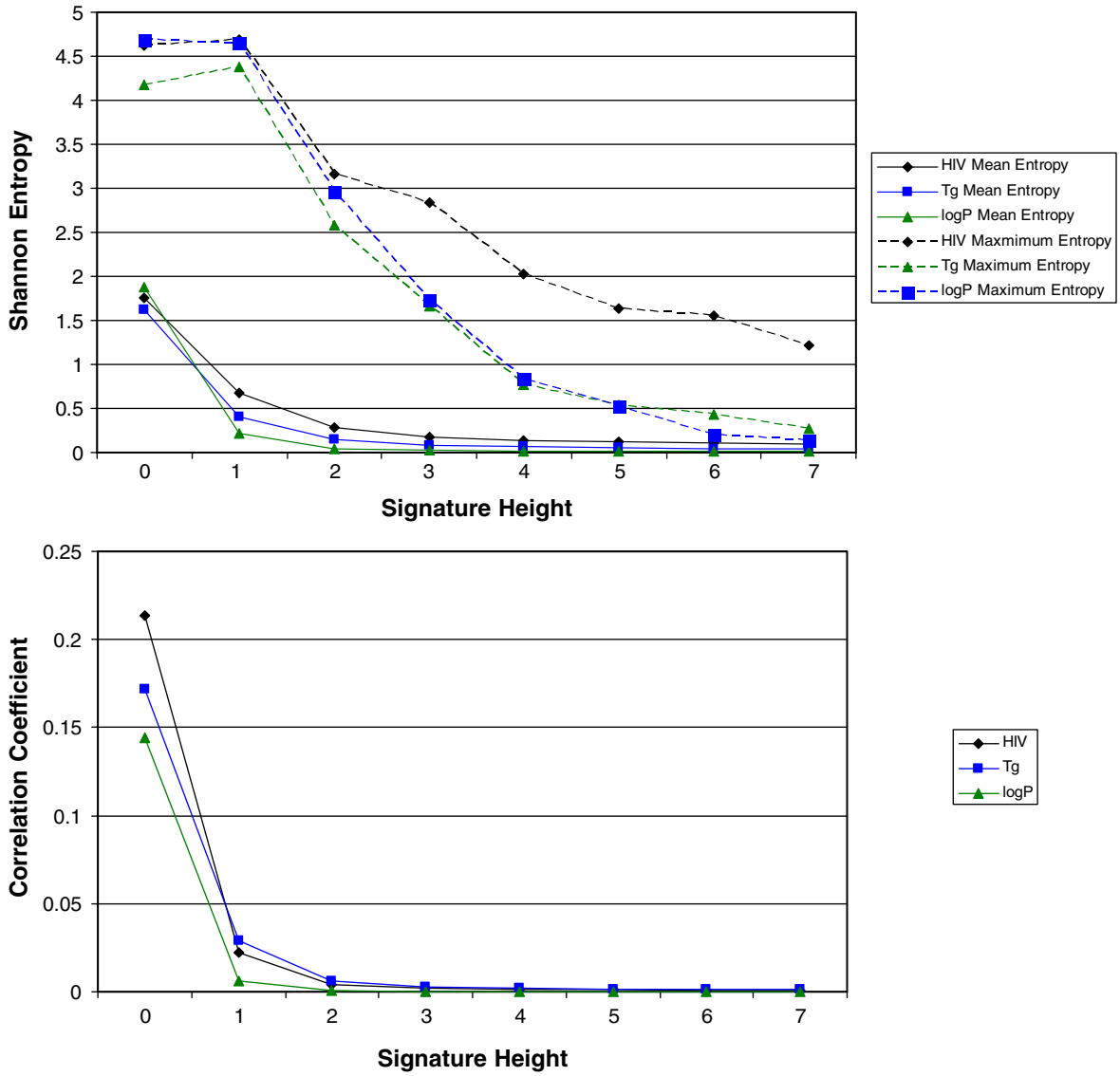

Figure 6. Top (a) mean and maximum Shannon entropy calculated for all atomic signatures at a given signature height. The results show that decreasing signature height results in an average increase in the descriptor variability. Bottom (b) mean pairwise correlation coefficients for every pair of atomic signatures at a given signature height.

the time from low degeneracy indices and molecular formulas using a simple simulated annealing algorithm. Consequently, if one wishes to build an information sharing system based on topological indices, one must find the appropriate combination of indices that are both highly degenerate and have predictive power in QSARs.

An alternative to topological indices is molecular fragments. While topological indices are rather abstract and difficult to interpret, molecular fragments provide direct chemical information. Furthermore, several studies have demonstrated that molecular fragments perform as well as topological indices in QSAR analyses [6, 7, 12], and that all topological indices can be computed from molecular fragments. Thus every QSAR involving topological indices can in principle be replaced with a similar QSAR based on molecular fragments [12, 37, 38]. Our results obtained with the signature molecular fragment (cf. Figure 5, Table 4), indicate that fragments of small sizes can be shared, and that reliable QSARs can be

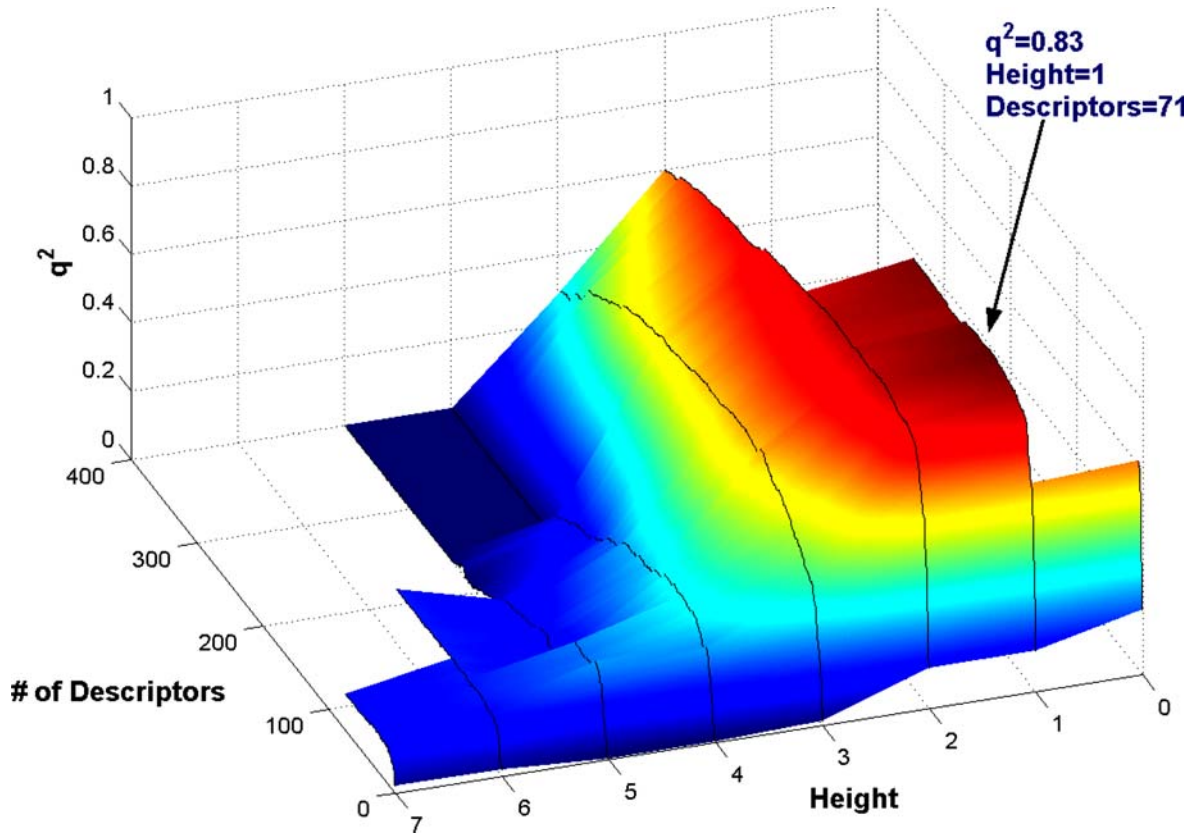

Figure 7. Optimization of the signature height for a set of 1000 molecules with log P data. The optimum signature height is selected as the one which produces the highest cross-validation squared correlation coefficient in order to preserve predictive accuracy.

obtained with these fragment sizes (cf. Table 5). Furthermore, according to Table 4, when the compound size increases so does the fragment size one can share. Nonetheless, also shown in Table 4, sharing molecular fragments is not 100% safe, and before sharing fragments, we recommend tests with a reverse engineering algorithm such the one that we described in Section Reverse engineering methods.

Acknowledgements

This work was funded in part by the U.S. Department of Energy's Genomics: GTL program (www.doegenomestolife.org) under project, "Carbon Sequestration in Synechococcus Sp.: From Molecular Machines to Hierarchical Modeling" (www.genomes-to-life.org). This work was also funded by Sandia National Laboratories Computer Science Research Fund. Sandia is a multiprogram laboratory operated by Sandia Corporation, a LockheedMartin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

1. Kier, L.B., Quant. Struct.-Act. Relat., 4 (1985) 109.
2. Randic, M., J. Am. Chem. Soc., 97 (1975) 6609.
3. Wiener, H., J. Am. Chem. Soc., 69 (1947) 17.
4. Balaban, A.T., J. Chem. Inf. Comput. Sci., 34 (1994) 398.
5. Bonchev, D. and Trinajstic, N., Int. J. Quantum Chem., 16 (1982) 463.
6. Tong, W., Lowis, D.R., Perkins, R., Chen, Y., Welsh, W.J., Goddette, D.W., Heritage, T. and Sheehan, D.M., J. Chem. Inf. Comput. Sci., 38 (1998) 669.
7. Zefirov, N.S. and Palyulin, V.A., J. Chem. Inf. Comput. Sci., 45 (2002) 1112.
8. Bender, A., Mussa, H.Y., Glen, R.C. and Reiling, S., J. Chem. Inf. Comput. Sci., 44 (2004) 170.
9. Filimonov, D.A., Poroikov, V., Borodina, Y. and Gloriozova, T., J. Chem. Inf. Comput. Sci., 39 (1999) 666.

10. Poroikov, V.V., Filimonov, D.A., Ihlenfeldt, W.-D., Gloriozova, T.A., Lagunin, A.A., Borodina, Y.V., Stepanchikova, A.V. and Nicklaus, M.C., J. Chem. Inf. Comput. Sci., 43 (2003) 228.
11. Faulon, J.-L., J. Chem. Inf. Comput. Sci., 34 (1994) 1204.
12. Faulon, J.L., Visco, D.P. Jr. and Pophale, R.S., J. Chem. Inf. Comput. Sci., 43 (2003) 707.
13. Faulon, J.-L., Collins, M.J. and Carr, R.D., J. Chem. Inf. Comput. Sci., 44 (2004) 427.
14. Churchwell, C.J., Rintoul, M.D., Martin, S., Visco, D.P., Kotu, A., Larson, R.S., Sillerud, L.O., Brown, D.C. and Faulon, J.-L., J. Mol. Graph. Model, 22 (2004) 263.
15. Faulon, J.-L., J. Chem. Inf. Comput. Sci., 36 (1996) 731.
16. Faulon, J.-L., Churchwell, C.J. and , J.D.P.V. Jr., J. Chem. Inf. Comput. Sci., 43 (2003) 721.
17. Sheridan, R.P. and Kearsley, S.K., J. Chem. Inf. Comput. Sci., 35 (1995) 310.
18. Venkatasubramanian, V., Chen, K. and Caruthers, J.M., J. Chem. Inf. Comput. Sci., 35 (1995) 188.
19. Kvasnicka, V. and Pospichal, J., J. Chem. Inf. Comput. Sci., 36 (1996) 516.
20. Hall, L.H., Dailey, R.S. and Kier, L.B., J. Chem. Inf. Comput. Sci., 33 (1993) 598.
21. Kier, L.B., Hall, L.H. and Frazer, J.W., J. Chem. Inf. Comput. Sci., 33 (1993) 143.
22. Kier, L.B., Hall, L.H. and Frazer, J.W., J. Chem. Inf. Comput. Sci., 33 (1993) 148.
23. Skvortsova, M.I., Baskin, I.I., Slovokhotova, O.L., Palyulin, V.A. and Zefirov, N.S., J. Chem. Inf. Comput. Sci., 33 (1993) 630.
24. Godden, J.W., Stahura, F.L. and Bajorath, J., J. Chem. Inf. Comput. Sci., 40 (2000) 796.
25. Cover, T.M. and Thomas, J.A., Elements of Information Theory. Wiley Series in Telecommunications, ed. Wiley. John Wiley & Sons, Inc., New York, 1991, 542 pp.
26. Matlab 7. MathWorks, (2005).
27. Hawkins, D.M., J. Chem. Inf. Comput. Sci., 44 (2004) 1.
28. Joachims , T., In Scholkopf, B., Burges, C.J.C., Smola, A.J., (Eds.), Advances in Kernel Methods-Support Vector Learning MIT Press Cambridge, MA 169, 1999.
29. Hart, W.E., SGOPT: A C++ Library of Global Optimization Methods. in IMSL. 1997.
30. Faulon, J.-L., J. Chem. Inf. Comput. Sci., 32 (1992) 338.
31. PubChem. National Library of Medicine, (2005).
32. Bicerno, J., Prediction of Polymer Properties. 3rd Edtion. Marcel Dekker, New York, 2002.
33. Zernov, V.V., Balakin, K.V., Ivaschenko, A.A., Savchuk, N.P. and Pletnev, I.V., J. Chem. Inf. Comput. Sci., 43 (2003) 2048.
34. Momma, M. and Bennett, K.P., In SIAM Proceedings Series, Arlington, 2002.
35. Quang, A.T., Zhang, Q.-L. and Xing, L., In Proceedings of the First International Conference on Machine Learning and Cybernetics, Beijing, 2002.
36. Bender, A., Mussa, H.Y., Glen, R.C. and Reiling, S., J. Chem. Inf. Comput. Sci., 44 (2004) 1708.
37. Baskin, I.I., Skvortsova, M.I., Stankevich, I.V. and Zefirov, N.S., J. Chem. Inf. Comput. Sci., 35 (1995) 527.
38. Skvortsova, M.I., Baskin, I.I., Skvortsova, L.A., Palyulin, V.A., Stankevich, I.V. and Zefirov, N.S., Theochem: J. Mol. Struct., 466 (1999) 211–217.