Available online at www.sciencedirect.com
SCIENCE DIRECT®

Journal of Molecular Graphics and Modelling 22 (2004) 263–273

Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

# The signature molecular descriptor 3. Inverse-quantitative structure–activity relationship of ICAM-1 inhibitory peptides

Carla J. Churchwell[a], Mark D. Rintoul[a], Shawn Martin[a], Donald P. Visco, Jr.[b], Archana Kotu[b], Richard S. Larson[c], Laurel O. Sillerud[c], David C. Brown[c], Jean-Loup Faulon[d,*]

[a] *Department of Computational Biology, Sandia National Laboratories, P.O. Box 5800, MS 0310, Albuquerque, NM, USA*
[b] *Department of Chemical Engineering, Tennessee Technological University, Box 5013, Cookeville, TN, USA*
[c] *Departments of Pathology, Biochemistry and Molecular Biology, School of Medicine, University of New Mexico, 915 Camino de Salud, CRF Rm 223, Albuquerque, NM, USA*
[d] *Department of Computational Biology, Sandia National Laboratories, 7011 East Avenue, P.O. Box 969, MS 9951, Livermore, CA 94550, USA*

Received 10 April 2003; received in revised form 6 October 2003; accepted 13 October 2003

**Abstract**

We present a methodology for solving the inverse-quantitative structure–activity relationship (QSAR) problem using the molecular descriptor called signature. This methodology is detailed in four parts. First, we create a QSAR equation that correlates the occurrence of a signature to the activity values using a stepwise multilinear regression technique. Second, we construct constraint equations, specifically the graphicality and consistency equations, which facilitate the reconstruction of the solution compounds directly from the signatures. Third, we solve the set of constraint equations, which are both linear and Diophantine in nature. Last, we reconstruct and enumerate the solution molecules and calculate their activity values from the QSAR equation. We apply this inverse-QSAR method to a small set of LFA-1/ICAM-1 peptide inhibitors to assist in the search and design of more-potent inhibitory compounds. Many novel inhibitors were predicted, a number of which are predicted to be more potent than the strongest inhibitor in the training set. Two of the more potent inhibitors were synthesized and tested in-vivo, confirming them to be the strongest inhibiting peptides to date. Some of these compounds can be recycled to train a new QSAR and develop a more focused library of lead compounds.

*Keywords:* QSAR; Inverse-QSAR; ICAM-1; LFA-1; Signature descriptor

## 1. Introduction

Current drug design is an iterative process, involving years of research, identification, synthesis, and subsequent testing of potential compounds that optimize a desired biological or chemical profile. Although databases containing millions of molecules exist, the most time-consuming step is the selection of high-quality lead compounds for possible synthesis. This inefficiency can be resolved through the use of computational tools to expedite the screening of molecular databases for a particular activity or property.

Virtual screening is a conventional computational technique that can be used in conjunction with high-throughput screening to refine the search for molecules matching a desirable property [1,2]. However, these tools are limited in that they can only provide solution molecules that are already in the database. Ideally, one would like to remove this constraint and identify compounds that are not currently in databases, but from which a high-quality lead compound can be produced. Here we present a novel and exciting technique to do just that; namely develop focused libraries of compounds that are not in the database but are predicted to have a desired value. This technique is rooted in the use of a powerful molecular descriptor that we have recently developed, called signature, which, in essence, involves the solution of the inverse-quantitative structure–activity relationship (QSAR) problem.

To demonstrate our inverse-QSAR approach, we applied it to a small set of inhibitory peptides directed against leukocyte trafficking and localization whose synthesis and testing in clinical trials is limited. A crucial event in leukocyte

[*] Corresponding author. Tel.: +1-925-294-1279; fax: +1-925-294-3020.
*E-mail address:* jfaulon@sandia.gov (J.-L. Faulon).

1093-3263/$ – see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2003.10.002

localization is binding to endothelium and subsequent migration from the blood into tissue. Recently, identification of cell surface adhesion molecules that mediate the adhesion of leukocytes to the endothelium, such as leukocyte functional antigen-1 (LFA-1) and its ligand intercellular adhesion molecule-1 (ICAM-1), have allowed investigation into leukocyte trafficking [3–5]. Collaborators at the University of New Mexico Health Sciences Center developed a novel antagonist of ICAM-1 that has in-vivo efficacy [6–9], and Kelly and coworkers [10,11] have developed a small molecule antagonist of LFA-1, although toxicity effects have limited its development as an in-vivo inhibitor. Using our technique, we can predict compounds that provide the proper inhibitory effect, some of which hopefully lack any adverse toxicity effects.

The paper is set up in the following manner. First, we will discuss QSARs in order to define the inverse problem. Next, we will present the signature molecular descriptor and review the various features of signature that have been previously investigated. Third, we will describe, in detail, the methodology used to generate solutions to the inverse-QSAR problem. Finally, we will provide an example demonstrating the use of signature in the solution of an inverse-QSAR problem; namely the construction of a focused library of compounds in rational peptide design.

## 2. Methodology

A quantitative structure–activity relationship is an empirical relationship between a molecule's structure and a specific biological activity or physical property possessed by that molecule. The independent variables in these equations are given in terms of molecular descriptors, which are operators on the molecular graph that strive to characterize the properties of the molecule [12–16]. QSARs have been used to quantify empirical data such as boiling points, electric moments, chromatography retention times, $IC_{50}$ values, lipophilicity, resonance, $\log P$, and polarity [17–20]. QSARs are generally trained against a large data set (called a training set) and validated through a subset not used in the parameterization (called a test set) in order to test the predictive ability of the QSAR. The quality and predictive capacity of the QSAR equation depends on several factors, including the size of the training set and the diversity of the molecules used in the construction of the equation. If the size of the training set is small, over-fitting of the QSAR can easily occur making for a relationship that provides poor predictions. If the molecules in the training set are of a certain type, then the QSAR developed is ill-equipped to predict properties of molecules of a type not included in the training set.

What is described above is called the forward-QSAR problem, which uses values for the independent variables of a particular compound in the QSAR to solve for the activity of that compound (the dependent variable). In contrast, the goal of the inverse-QSAR problem is to determine values for the independent variables given a desired activity. It is very important to note that the inverse-QSAR method discussed in this work applies to *any* property and not exclusively to activity.

The inverse-QSAR problem is quite challenging for a variety of reasons. First, one needs to be able to solve the QSAR for a given activity. This corresponds to generating the vectors of solutions (values for the independent variables, or descriptor values) that correspond to the given activity. If this first step can be completed, the generated solutions then need to be turned into actual compounds.

Reconstructing molecules that match molecular descriptor values is a long-standing problem. However, there are only a few reports in the literature providing solutions to this problem. Most of the proposed techniques are stochastic in nature and use either genetic algorithms or Monte Carlo methods to search for and construct chemical structures matching predefined descriptor values. Venkatasubramanian et al. [21] and Sheridan and Kearsley [22] were the first to propose stochastic techniques based on genetic algorithms, while methods based on a Monte Carlo approach were reported later [23,24]. Although other 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 and coworkers [25–27] reconstructed molecular structures from the count of paths, ${}^{l}P$, 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 ${}^{3}P$ count are rejected. Skvortsova et al. [28] used a similar technique, but from the count of paths they derived an edge sequence in addition to the degree sequence. An edge sequence counts the number of edges between each distinct pair 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.

Owing to the limited progress offered by the above approaches to the solution of the inverse problem, it is clear that the key to an effective solution methodology lies in the use of a molecular descriptor that facilitates reconstruction of the solutions into actual compounds. This descriptor needs to be information rich, have good correlation abilities in QSAR applications, and must also be computationally efficient. A computationally efficient descriptor should have a low degeneracy, that is, it should lead to a limited number of solutions when applied in inverse-QSAR. Next, we briefly present a descriptor that we believe matches the above criteria. The descriptor is called signature and is further detailed in two previous papers in this series [29,30].

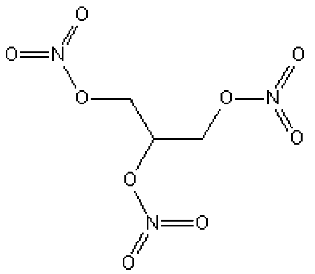
### 2.1. The signature descriptor

Signature is based on the molecular graph of a molecule, $G = (V_G, E_G)$, where the elements in $V_G$ denote the atoms in the molecule, and the edges of $E_G$ correspond to the bonds between those atoms. In this context, a molecule is characterized by a set of canonical subgraphs, each rooted on a different vertex with a predefined level of branching, which we refer to as the height $h$. The branching of a vertex is an extended degree sequence that describes the local neighborhood, up to the distance $h$ away from the root.

We define an atomic signature, ${}^h\sigma_G(x)$, as the canonical subgraph of $G$ consisting of all atoms a distance $h$ from the root $x$. A molecular signature, ${}^h\Sigma_G$, is then the set of all unique atomic signatures and the occurrence with which they appear in the molecular graph. Even though the atomic signatures are unique, they are, by construction, interrelated allowing information about the overall structure of the molecule to be conveyed at the end.

The atomic signatures make up the set of molecular descriptors for a molecule. These are expressed in terms of a string of characters that correspond to the canonized subgraph in a breath-first order. Branch levels are indicated by a set of parenthesis following the parent vertex. An example of the molecular signature for nitroglycerine is given in Fig. 1.

Signature is uniquely suited to address the issues related to the inverse-QSAR problem. First, signature produces QSARs on par with those obtained from conventional molecular descriptors. In fact, signature encapsulates information from which other molecular descriptors can be computed. Its usefulness in QSAR analysis was previously established by comparison to a QSAR developed from the commercial package, Molconn-Z, with similar results [29,31]. Second, signature is shown to be less degenerate than many other popular descriptors. The degeneracy of a molecular descriptor depends on how well it is able to map one property to one descriptor. A molecular descriptor with a low degeneracy is vital in limiting the number of solutions to the inverse-QSAR problem. Ideally, the descriptors should be orthogonal to one another, such that only one descriptor corresponds to the information for a single structural motif [32,33]. In a previous study, the degeneracy of signature was systematically probed and compared to a broad set of traditional molecular descriptors [30]. Signature proved to be less degenerate than the other descriptors, but more importantly, its degeneracy can be user controlled. Third, and foremost, signature provides a way to go from numerical solutions of the inverse-QSAR problem to actual structures that correspond to solutions. Indeed, the main advantage of signature versus other molecular descriptors is its readiness for inverse problems. An algorithm to both enumerate and sample chemical structures corresponding to solution vectors (i.e. molecular signatures), has already been developed and tested for a variety of compounds including alkanes, fullerenes, and HIV-1 protease inhibitors [30].



$^1\sigma$ (Nitroglycerin)

| | |
|---|---|
| 3 | o_(n_c_) |
| 6 | o=(n_) |
| 3 | n_(o_o=o=) |
| 5 | h_(c_) |
| 2 | c_(o_h_h_c_) |
| 1 | c_(o_h_c_c_) |

Fig. 1. Nitroglycerine and its corresponding height one molecular signature. The molecular signature is the sum of the atomic signatures. Note the different atom types between the single and double bonded oxygen atoms.

### 2.2. Inverse-QSAR scheme

The inverse-QSAR method can be broken into four steps. The first step is the QSAR analysis. Here, we generate every atomic signature of a desired height for the compounds in the training set. We then use those signatures to construct a QSAR equation relating compounds to their activities. The second step is to generate the set of constraint equations with integer coefficients (Diophantine equations) for the signatures. In the third step, we solve these equations for integer solutions using a Diophantine equation solver. The last step consists of building the molecular structures and predicting their activities using the QSAR equation.

#### 2.2.1. QSAR analysis

Here, we outline the use of signature in the QSAR analysis; for further details on the procedure, the reader is referred to our two previous papers [29,30]. Construction of the QSAR equation begins by expressing each compound in the training set in terms its molecular signature of height $h$. A list of the unique descriptors (atomic signatures) is compiled and provides a descriptor database for the QSAR. This set contains the minimum number of descriptors needed to span the activity/property space of compounds in the training set. Assuming, there are $m$ compounds in the training set and $n$ unique descriptors, an $n \times m$ “descriptor matrix” is constructed by screening each compound against the set of unique descriptors and storing that descriptor’s occurrence number in the matrix. Perfectly correlated rows, i.e. descriptors with the same predictive capabilities, are removed from the matrix to avoid redundancies, which can skew the results of the multiple linear regression analysis. This matrix and the corresponding property values determine the QSAR equation that will be developed.

In our studies, the QSAR equation is a linear equation of the form $\sum \alpha_i x_i - \alpha_0 = P$, where $\alpha_i$ represents the regression coefficients, $x_i$ represents the occurrence number of the molecular descriptor $i$, and $P$ is the property value of interest minus the regression constant. The number of molecular descriptors, unknown until the training set has been established, dictates the number of independent variables. To avoid the possibility of over-fitting the data, which occurs when the number of independent variables is equal to or exceeds the number of dependent variables, a forward selection procedure [34] was used in the multilinear regression. The forward selection adds variables one by one, according to those that have the most impact on the model, as determined by $r^2$ and $F$ statistics. This method is computationally efficient and can control the number of independent variables of the QSAR equation [34].

To test the accuracy of the model, a plot of the experimental values versus the QSAR correlated values is constructed. For a good, correlative model, the points on the graph should lie close to a 45° angle, which is quantitatively described by the $R^2$ value. If this is not the case, then adjustments can be made to modify the number of independent variables; the resultant equation is inhomogeneous with unknown occurrence values, $x_i$.

### 2.2.2. *Production of constraint equations*

In addition to the QSAR equation, constraints are needed to ensure the ability to reconstruct compounds from the solutions. There are two types of constraint equations, namely the graphicality equation and the consistency equations.

*2.2.2.1. Graphicality equation.* The graphicality equation ensures that at least one connected graph can be constructed from the molecular descriptors. This equation, taken directly from graph theory, uses only the degree of the vertices in the graph. In order to build a connected graph, we require that (1) the sum of all the vertex degrees must be even and (2) the number of vertices of odd degree must be even. The resulting equation can be expressed in terms of a degree sequence $N = \{n_1, n_2, \ldots, n_k\}$ where $n_i$ is the number of vertices of degree $i$. In this case, the degree sequence $N$ is graphical if and only if there exists an integer $z \geq 0$ such that

$$\sum_{i=2}^{k}(i-2)n_i - n_1 + 2 = 2z. \qquad (1)$$

The graphicality equation can be computed directly from the height zero molecular signature.

*2.2.2.2. Consistency equations.* The next set of equations is collectively referred to as the consistency equations. Recall that a molecular signature is a collection of interrelated atomic signatures, where each atomic signature describes a particular atom and its neighboring atoms to a predetermined height. In constructing the signature of a molecule, it is guaranteed that a bond in one atomic signature will match up with a bond in another atomic signature, albeit in reverse order. However, blind reconstruction of the molecule requires equations to enforce these conditions of interdependency among the atomic signatures. This is done by matching bonds between two atoms of one signature to the bonds involving the same atoms in all other signatures.

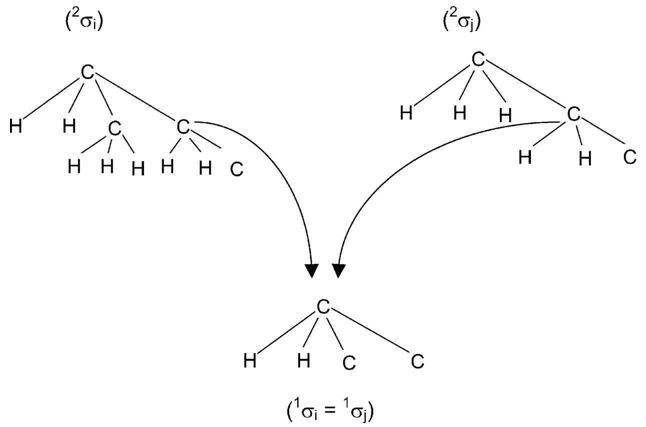


Fig. 2. Graphical depiction of the bond occurrence $({}^2\sigma_i \rightarrow {}^2\sigma_j)$. Here, ${}^2\sigma_i$ supplies the height one bond between the carbon root and the last carbon atom child. If a root swap is performed with this last carbon atom, then the resulting height one signature is C_(H_H_C_C_), which we designate as ${}^1\sigma_i$. Similarly, if a root swap is performed on ${}^2\sigma_j$ and its carbon atom child, then truncated to height one, its signature will match that of ${}^1\sigma_i$.

We will use the notation ${}^h\sigma_i$ to describe the atomic signature of height $h$ of an arbitrary atom $i$. Using ${}^h\sigma_i$ as a reference, any bond between the root and one of its children must be sought in all other atomic signatures in which the positions of the root and child are the transpose of ${}^h\sigma_i$. We use the notation $\#({}^{h-1}\sigma_i \rightarrow {}^{h-1}\sigma_j)$, to depict the number of bond types ${}^h\sigma_i$ has in common with ${}^h\sigma_j$. Clearly, then $\#({}^{h-1}\sigma_i \rightarrow {}^{h-1}\sigma_j) = \#({}^{h-1}\sigma_j \rightarrow {}^{h-1}\sigma_i)$. As depicted in Fig. 2, it is important to note that the signature of a bond is one height less than the height of the molecular signature. The reason is that when $\#({}^{h-1}\sigma_i \rightarrow {}^{h-1}\sigma_j)$ is computed, one has to transpose the root $i$ with a child $j$. While the neighborhood of $i$ was initially probed up to height $h$, the transposed signature with new root $j$ probes the neighborhood of $j$ only up to height $h{-}1$. In the case where $i = j$, then $\#({}^{h-1}\sigma_i \rightarrow {}^{h-1}\sigma_i)$ must be even.

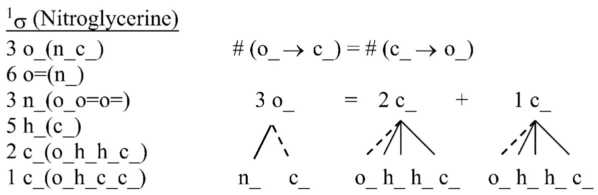


Fig. 3. Illustration of how a consistency equation for nitroglycerine is calculated. In particular, the number of oxygen–carbon single bonds must equal the number of carbon–oxygen single bonds. Only one atomic signature contains a single bonded oxygen root connected to a single bonded carbon with an occurrence of three. On the other hand, there are two atomic signatures with a single bonded carbon root. In each case, the carbon root is connected to a single bonded oxygen atom shown with a dashed line. There is only one instance of a carbon–oxygen single bond in both signatures, so the occurrence number is the sum of these, which is 3.

For example in Fig. 3 we calculate the number of oxygen–carbon single bonds and compare it to the number of carbon–oxygen single bonds in nitroglycerine. Here, the first signature supplies the bond type, in this case an oxygen atom single bonded to a carbon atom. To make sure bond types match up, we need to find all instances of a carbon atom single bonded to an oxygen atom in the other signatures. This is done by searching the list of unique signatures for all those that contain a single bonded carbon atom (c_) as the root. For each matching signature, an algorithm is called to re-root the tree, such that a child is now the root. This procedure is reiterated for each of the first children. Because the carbon in this example has four children, there will be four restructured graphs. If the signature contained a carbon–oxygen bond, then upon restructuring, that oxygen should appear as the root. A quick check is performed and if any oxygen atom exists as the roots, then the occurrence number of the original signature is returned. All matches are then summed to give a total for the number of carbon–oxygen bonds, which should equal the number of oxygen–carbon bonds, in this case three.

The consistency equations can be summarized as follows: a molecular signature of height $h$ ($^{h}\Sigma$) is consistent if and only if the two following conditions are verified:

(i) For all atomic signatures $^{h-1}\sigma_i$ and $^{h-1}\sigma_j$ in $^{h-1}\Sigma$, $\#(^{h-1}\sigma_i \rightarrow {}^{h-1}\sigma_j) = \#(^{h-1}\sigma_j \rightarrow {}^{h-1}\sigma_i)$.

(ii) For all $^{h-1}\sigma_i$ in $^{h-1}\Sigma$, $\#(^{h-1}\sigma_i \rightarrow {}^{h-1}\sigma_i)$ is even valued.

In the nitroglycerine example in Fig. 1, the occurrence numbers of the signatures were known quantities and should be consistent by construction. However, in the inverse-QSAR, the occurrence numbers are represented by unknown quantities, $x_i$. Construction of these equations will produce relationships that quantitatively convey the interdependencies of one signature on another in the set.

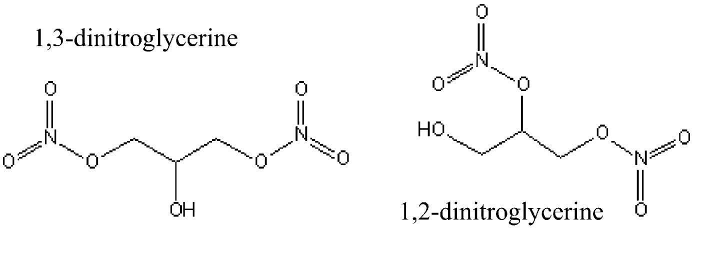
*2.2.2.3. Equation solver.* Together, the QSAR and constraint equations form a system of equations with unknown occurrence numbers $x_i$ corresponding to atomic signature $i$. These solutions must represent quantities meaningful to signature and, hence, the occurrence numbers should take on non-negative integer values. Equations in which only positive integer solutions are allowed are called Diophantine equations. Research on algorithms for solving linear integer equations has been widely investigated starting from the ancient Greeks [35]. Such systems arise in various areas of computer science and efficient algorithms are well-known for solving these systems over real numbers, rational numbers, and even integers. Unfortunately, restricting the domain to the natural numbers (positive integers) makes the problem much more difficult and the algorithms in the previous class are no longer suitable.

In the recent past, several techniques to directly solve linear Diophantine systems have been proposed. We have implemented such an algorithm adapted from Contejean and Devie [36]. This algorithm uses a geometric interpretation of Fortenbacher's algorithm [37], which efficiently solves homogeneous and inhomogeneous linear Diophantine equations. The output of the Diophantine solver is a set of basis vectors that spans the solution space of the system of equations.

Our system is comprised of three types of linear equations: inhomogeneous, homogeneous, and modulus equations. The modulus equations can be re-written in terms of a homogeneous equation by adding a dummy variable to enforce modularity. This equation can then be included into the system of equations that eventually feed into the Diophantine solver. Incorporation of the inhomogeneous QSAR equation has been found to slow the time it takes for the Diophantine solver to produce results. For this reason, we purposefully leave out the QSAR equation when determining the feasible solutions from the system of homogeneous equations and then use the QSAR equation to predict the activity values of the solutions.

*2.2.2.4. Structure generator.* Once a solution has been found, the corresponding molecule needs to be constructed from the molecular signature. As can be seen in Fig. 4, it is possible that more than one structure may exist that corresponds to the molecular signature. As a result, an enumeration routine was developed to find all possible structures with a given molecular signature. A brief overview of the enumeration routine is given here, but the reader is referred to a previous paper [30] 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



| | $^{1}\sigma$ |
|---|---|
| 2 | o_(n_c_) |
| 1 | o_(h_c_) |
| 4 | o=(n_) |
| 2 | n_(o_o=o=) |
| 1 | h_(o_) |
| 5 | h_(c_) |
| 2 | c_(o_h_h_c_) |
| 1 | c_(o_h_c_c_) |

Fig. 4. A molecular signature may correspond to multiple structural configurations. Both 1,3-dinitroglycerine and 1,2-dinitroglycerine have the same height one molecular signature given at the right. Increasing the height of the signatures will decrease the number of structural configurations that correspond to a molecular signature.

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 complete [30], meaning only the set of unique non-isomorphic graphs are produced.

## 3. Inverse-QSAR OF ICAM-1 inhibitory peptides

The inhibitory compound that UNM tested was a cyclic peptide composed of nine amino acids that bind to ICAM-1, thereby inhibiting LFA-1/ICAM-1 binding. Using alanine replacement and homologous amino acid substitutions, we identified residues strategic to the antagonist activity. A small set of such derived peptides was used as a training set for the inverse-QSAR. The activity or potency of the peptides is associated to an $IC_{50}$ value, which measures the concentration that leads to half-maximal inhibition of receptor to ligand. Table 1 lists the amino acid sequence of sixteen such derived peptides with their $IC_{50}$ values (given in μM), determined using a cellular aggregation blocking assay [7,8]. In brief, the aggregation of a cell line dependent on LFA-1 binding to ICAM-1 was used to measure the inhibitory capacity of each peptide. Inhibitory peptides and cells were seeded in flat-bottomed microtitier plates and allowed to aggregate. The number of aggregates and total number of free (single) cells were counted using inverted phase microscopy. The percent aggregation, *P*, was determined as

$$P = 100\left(1 - \frac{F}{I}\right) \quad (2)$$

where *F* is the final number of free cells and *I* is the initial number of free cells. This percent aggregation was then used to calculate the percent inhibition, $P_{\mathrm{i}}$ as follows

$$P_{\mathrm{i}} = 100\left(1 - \frac{P_{\mathrm{ip}}}{P_{\mathrm{c}}}\right) \quad (3)$$

where $P_{\mathrm{ip}}$ is the percent aggregation with inhibitory peptide and $P_{\mathrm{c}}$ is the percent aggregation in the control experiment. The $IC_{50}$ values were calculated from a line fit to the percent inhibition data as a function of inhibitory peptide concentration over the range from 10 μM to 1 mM. Each condition was performed in duplicate while each experiment was performed a minimum of three times.

The first and last amino acids in the sequence are connected to one another via a disulfide bridge, making the structures cyclic. In addition, the peptides are classified according to their inhibitory capabilities: peptides with $IC_{50}$ values less than or equal to 500 are considered strong inhibitors, peptides with $IC_{50}$ values between 500 and 1000 are considered weak inhibitors and peptides with $IC_{50}$ values greater than or equal to 1000 are said to be non-inhibitors.

Fourteen of these peptides were used as a training set for the inverse-QSAR process; peptides 4 and 13 were used as the test set. Here, the goal was to find any other compounds within the property space of the training set that similarly inhibit the binding of LFA-1/ICAM-1, but with greater efficacy, i.e. a lower $IC_{50}$ value.

Table 1
Training and test set for the ICAM-1/LFA-1 inhibitory peptides

| | Peptide sequence | Experimental $IC_{50}$ (μM)[a] | Inhibitor strength |
|---|---|---|---|
| 1 | CLLRMRSAC | 480 | Strong |
| 2 | CILRMRSAC | 190 | Strong |
| 3 | CVLRMRSAC | >1000 | Non |
| 4 | CLIRMRSAC* | 720 | Weak |
| 5 | CLVRMRSAC | >1000 | Non |
| 6 | CLLKMRSAC | 105 | Strong |
| 7 | CLLRMKSAC | 90 | Strong |
| 8 | CLLRMRSLC | >1000 | Non |
| 9 | CLLRMRSVC | 700 | Weak |
| 10 | CLLRMRSIC | 580 | Weak |
| 11 | CALRMRSIC | >1000 | Non |
| 12 | CLARMRSIC | >1000 | Non |
| 13 | CLLRARSIC* | >1000 | Non |
| 14 | CLLRMASIC | >1000 | Weak |
| 15 | CLLRMRAIC | 710 | Weak |
| 16 | CILKMKSAC | 40 | Strong |

Peptides indicated with an asterisk (peptides 4 and 13) indicate compounds in the test set. Listed are the amino acid sequences, the experimentally determined $IC_{50}$ value and the potency of the peptide in inhibiting the ICAM-1/LFA-1 complex.

[a] $IC_{50}$ values determined using cellular assay described in Sillerud et al. [8].

### 3.1. QSAR analysis

The training set contained 14 cyclic peptides composed of nine amino acids, which were expressed in terms of a linear, one letter amino acid sequence. Following the procedure previously outlined, 47 unique atomic signatures of height one were used, each of which was given an unknown occurrence number $x_i$ (See Table 2).

Once the compounds in the training set were expressed in terms of their signatures, a linear QSAR equation could be created. First, a 47 × 14 descriptor matrix was constructed, then screened for perfectly correlated rows, or rows containing identical entries. Recall that equivalent variables distort the multiple linear regression results if included in the analysis. In addition, rows containing a single or double entry were discarded in an attempt to generalize the signatures in the QSAR so that none would map to a specific activity. Second, the actual $IC_{50}$ value for non-binding peptides ($IC_{50}$ > 1000) was not experimental measured, thus we assigned them a value of 1000. Furthermore, to make all the values of the dependent variables the same order of

Table 2
Height one amino acid signatures for the ICAM-1/LFA-1 training set

| | | | | | |
|---|---|---|---|---|---|
| $x_1$ | A(CL) | $x_{17}$ | L(AC) | $x_{33}$ | R(AM) |
| $x_2$ | A(CS) | $x_{18}$ | L(AR) | $x_{34}$ | R(AS) |
| $x_3$ | A(IR) | $x_{19}$ | L(CI) | $x_{35}$ | R(IM) |
| $x_4$ | A(LR) | $x_{20}$ | L(CL) | $x_{36}$ | R(LM) |
| $x_5$ | A(MS) | $x_{21}$ | L(CS) | $x_{37}$ | R(MS) |
| $x_6$ | A(RR) | $x_{22}$ | L(CV) | $x_{38}$ | R(MV) |
| $x_7$ | C(AC) | $x_{23}$ | L(IK) | $x_{39}$ | S(AI) |
| $x_8$ | C(CI) | $x_{24}$ | L(IR) | $x_{40}$ | S(AK) |
| $x_9$ | C(CL) | $x_{25}$ | L(KL) | $x_{41}$ | S(AR) |
| $x_{10}$ | C(CV) | $x_{26}$ | L(LR) | $x_{42}$ | S(IR) |
| $x_{11}$ | I(AC) | $x_{27}$ | L(RV) | $x_{43}$ | S(LR) |
| $x_{12}$ | I(CL) | $x_{28}$ | M(AR) | $x_{44}$ | S(RV) |
| $x_{13}$ | I(CS) | $x_{29}$ | M(KK) | $x_{45}$ | V(CL) |
| $x_{14}$ | I(LR) | $x_{30}$ | M(KR) | $x_{46}$ | V(CS) |
| $x_{15}$ | K(LM) | $x_{31}$ | M(RR) | $x_{47}$ | V(LR) |
| $x_{16}$ | K(MS) | $x_{32}$ | R(AL) | | |

Table 3
Overall statistics for the QSAR equation with six signatures

| $F$ | $R^2$ | $s^2$ | $s^2$ (test set) |
|---|---|---|---|
| 16.9 | 0.935 | 0.015 | 0.011 |

Table 4
Individual descriptor statistics for the QSAR equation with six signatures

| Descriptor | $R^2$ | Variable inflation factor | $P$-value |
|---|---|---|---|
| $x_2$ | 0.3735 | 1.5962 | 0.0202 |
| $x_8$ | 9.55e−7 | 1.00000095 | 0.9974 |
| $x_{13}$ | 0.1692 | 1.2037 | 0.1439 |
| $x_{31}$ | 0.4726 | 1.8961 | 0.0066 |
| $x_{37}$ | 0.1609 | 1.1918 | 0.1551 |
| $x_{41}$ | 0.0057 | 1.0057 | 0.7976 |

magnitude, we used the base ten logarithm of each $IC_{50}$ value in the QSAR. Last, a forward stepping algorithm was applied to select the most statistically significant signatures, one at a time. We chose to use a QSAR equation with six variables, where the variables $x_i$ are the occurrence numbers of the signatures listed in Table 2.

$$\log_{10}(IC_{50}) = 2.81 - 0.739x_2 - 0.574x_8 + 0.662x_{13} + 0.728x_{31} + 0.727x_{41} - 0.644x_{37} \quad (4)$$

The training set contained biased activities; almost half of the compounds had activities equal to 1000, the other majority of compounds contained activities less than 500. This trend was inevitably captured in the QSAR equations, where the added signatures simply distinguished between strong and non-inhibitory compounds. Thus, the coefficients in the QSAR equation are not as stable as we would like; ideally, they should exhibit little to no variation when another descriptor is added. However, since our data set is small, the QSAR will be sensitive to perturbations, i.e. the addition of new signatures.

Fig. 5 illustrates the ability of Eq. (4) to correlate the $IC_{50}$ values of the training set as well as predict the values of the peptides in the test set. Although the statistics in Table 3 could be higher, the key is to choose a QSAR equation that not only correlates the signatures to the activities, but one that is also predictive. We chose our QSAR based on the statistics in Table 4 (which show our QSAR

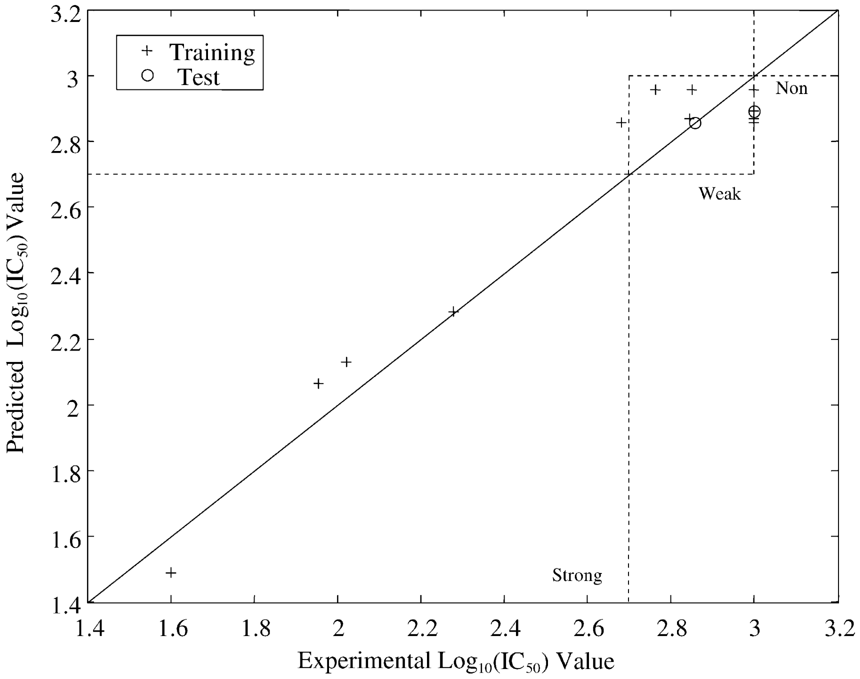


Fig. 5. Accuracy of the QSAR equation (Eq. (4)) for the LFA-1/ICAM-1 training set. The points lie near a 45° line and the points in the test set are accurately predicted, indicating that the QSAR equation is a good correlation of signatures to the $IC_{50}$ values.

Table 5
The predicted $IC_{50}$ values and their differences for peptides in the test set using Eq. (4).

| | Peptide sequence | Experimental $IC_{50}$ (μm) | Predicted $IC_{50}$ (μm) | Difference |
|---|---|---|---|---|
| 4 | CLIRMRSAC | 720 | 727.8 | 7.8 |
| 13 | CLLRARSIC | >1000 | 790.7 | 209.3 |

has not been overly affected by multicolinearity) as well the QSAR's ability to predict the $IC_{50}$ values for compounds in the test set. Table 5 lists the differences of the predicted and experimental $IC_{50}$ values for the compounds in the test set using Eq. (4).

### 3.2. Construction of constraint equations

The amino acids can be regarded as vertices of degree 2. Consequently, the graphicality equation will always be satisfied and need not be calculated for this particular training set.

The consistency equations were calculated from the unique signature set as described in Section 2. In addition, we wanted the resulting compounds to be cyclic structures composed of nine amino acids. To capture this requirement, we added a constraint that the number of amino acids in any solution was to total 9. These equations are listed in Table 6. Notice that the individual constraint equations do not contain the majority of the variables. The two modulus equations (Table 6, Eqs. (16) and (23)) were incorporated into the system of equations by adding dummy variables (one for each modulus equation) to make them homogeneous.

Table 6
Constraint equations for the height one amino acid signatures in the training set

(1) $-x_{44}+x_{46}=0$
(2) $-x_{38}+x_{47}=0$
(3) $-x_{22}-x_{27}+x_{45}+x_{47}=0$
(4) $-x_{10}+x_{45}+x_{46}=0$
(5) $-x_{34}-x_{37}+x_{41}+x_{42}+x_{43}+x_{44}=0$
(6) $-x_{21}+x_{43}=0$
(7) $-x_{16}+x_{40}=0$
(8) $-x_{13}+x_{39}+x_{42}=0$
(9) $-x_{2}-x_{5}+x_{39}+x_{40}+x_{41}=0$
(10) $-x_{28}-x_{30}-2x_{31}+x_{33}+x_{35}+x_{36}+x_{37}+x_{38}=0$
(11) $-x_{18}-x_{24}-x_{26}-x_{27}+x_{32}+x_{36}=0$
(12) $-x_{14}+x_{35}=0$
(13) $-x_{3}-x_{4}-2x_{6}+x_{32}+x_{33}+x_{34}=0$
(14) $-x_{15}-x_{16}+2x_{29}+x_{30}=0$
(15) $-x_{5}+x_{28}=0$
(16) $(x_{20}+x_{25}+x_{26})\%2=0$
(17) $-x_{15}+x_{23}+x_{25}=0$
(18) $-x_{12}-x_{14}+x_{19}+x_{23}+x_{24}=0$
(19) $-x_{9}+x_{17}+x_{19}+x_{20}+x_{21}+x_{22}=0$
(20) $-x_{1}-x_{4}+x_{17}+x_{18}=0$
(21) $-x_{8}+x_{11}+x_{12}+x_{13}=0$
(22) $-x_{3}+x_{11}=0$
(23) $(x_{7}+x_{8}+x_{9}+x_{10})\%2=0$
(24) $-x_{1}-x_{2}+x_{7}=0$

Eqs. (16) and (23) are modulus equations, which can be expressed as homogeneous equations by adding a dummy variable. For example Eq. (16) would read $x_{20}+x_{25}+x_{26}-2z_{1}=0$. The % sign indicates the modulus is to be used.

### 3.3. Equation solver

As mentioned previously, the inhomogeneous equations were intentionally excluded from the system in order to

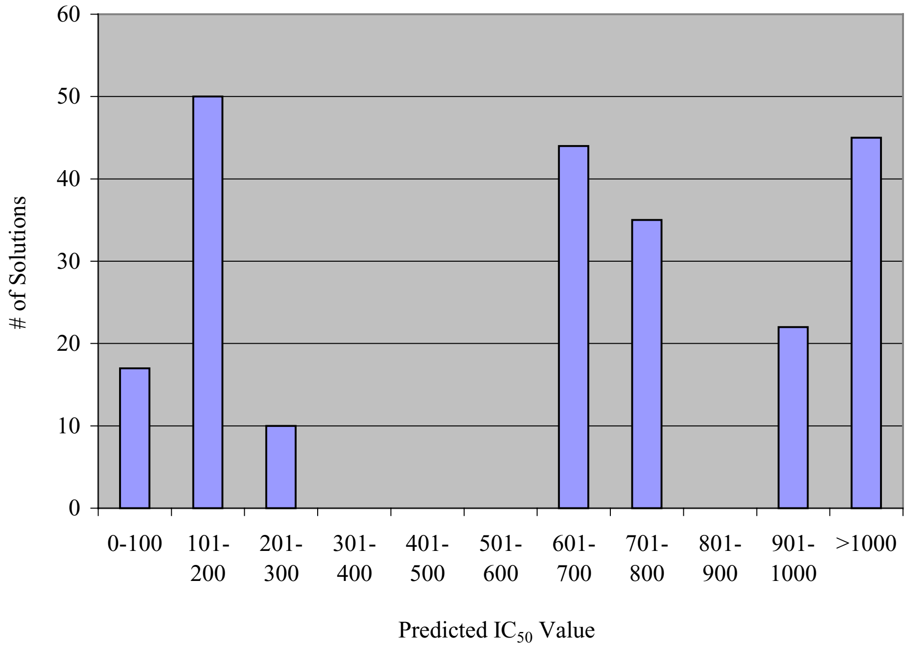


Fig. 6. Distribution of $IC_{50}$ values for the solutions of the inverse-QSAR using six signatures. Solutions are grouped according to their $IC_{50}$ values: 0–100, 101–200, 201–300, and so on, up to 1000.

obtain results in a reasonable amount of time. Thus, only the constraint equations were solved using the Diophantine solver. Due to the size constraint of the peptides, only those solutions containing nine or less amino acids were kept, the rest were discarded. Solutions with less than nine amino acids were used in making linear combinations, again, adhering to the size constraint of nine amino acids. By leaving out the QSAR equation, all solutions were obtained with activities spanning a wide range of $IC_{50}$ values. The distribution of predicted activities is given in Fig. 6, where the solutions were divided into bins of 100 ranging up to 1000.

### 3.4. Structure generator

The reconstruction of the peptides was straightforward in this case. By construction, each peptide only contained nine amino acids that formed a cyclic structure. Therefore, once the amino acid sequence of the peptide was known, the structure would also be known. From a solution, we start building the amino acid sequence by selecting a descriptor—it does not matter which one since the structure is cyclic. The children of each amino acid are used as guides to tell us what the previous and following amino acids are in the sequence. Fig. 7 illustrates how a sample solution is reconstructed from the amino acid signatures. Here we pick a signature, in this case C(AC), since we know that the first and last amino acids form a disulfide linkage. We know that it is connected to another signature with root A and a signature with root C, both of which should have C as their child. So, we choose the signature A(CS) as the next residue in the sequence. C is already connected to an A, so the next residue must be a signature with root S. This process is reiterated until no more amino acids are left and the last amino acid should be a child of the first one and vice versa.

Table 7 lists 20 sequences corresponding to compounds with the lowest predicted $IC_{50}$ values. Even though some of the peptides are predicted to be strong inhibitors, they may not be viable candidates for synthesis. For example peptide 12, which has the sequence CASICCLIC, contains two cysteine residues in the middle of the compound. These residues contain sulfur atoms which may form undesired disulfide bonds that potentially distort the three dimensional structure.

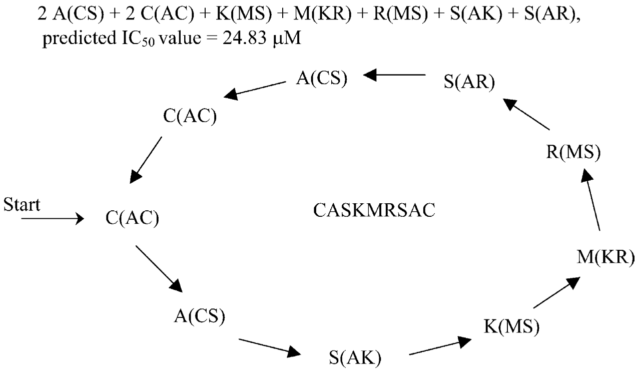

2 A(CS) + 2 C(AC) + K(MS) + M(KR) + R(MS) + S(AK) + S(AR),
predicted IC50 value = 24.83 μM
A(CS)
S(AR)
C(AC)
R(MS)
Start
C(AC)
CASKMRSAC
M(KR)
A(CS)
K(MS)
S(AK)

Fig. 7. Reconstruction of a solution peptide from the amino acid signature. Since the structure is cyclic, it does not matter which signature is used to start of the sequence. Here we choose a C(AC) to start. This is connected to both an A and a C. Selecting the signature A(CS) we know that it is already connected to a C, so the next signature must be S(AK). Continuing in this manner, the last signature should match up the first.

Table 7
Peptide sequences of the twenty lowest $IC_{50}$ values as predicted by the inverse-QSAR with six signatures

| | Peptide sequence | Predicted $IC_{50}$ value (μm) | Actual $IC_{50}$ value (μm) |
|---|---|---|---|
| 1 | CASKMKSAC | 21.48 | |
| 2 | CASKMRSAC | 24.83 | 23 |
| 3 | CASKMRSVC | 25.53 | |
| 4 | CASKMRSLC | 25.53 | |
| 5 | CASKMRSIC | 31.26 | |
| 6 | CASKMRLIC | 31.41 | |
| 7 | CASKMKLIC | 31.41 | |
| 8 | CASKMRAIC | 31.41 | |
| 9 | CASRMKLIC | 36.31 | |
| 10 | CILKMRSVC | 37.33 | 28 |
| 11 | CILKMRSLC | 37.33 | |
| 12 | CASICCLIC | 38.46 | |
| 13 | CILKMRSIC | 45.71 | |
| 14 | CILKMRLIC | 45.92 | |
| 15 | CILKMKLIC | 45.92 | |
| 16 | CILRARLIC | 45.92 | |
| 17 | CILKMRAIC | 45.92 | |
| 18 | CASKMKLLC | 117.8 | |
| 19 | CASKMRVLC | 117.8 | |
| 20 | CASKMRLVC | 117.8 | |

## 4. Discussion

From the inversion process, a total of 223 compounds were found, including the 14 original compounds in the training set and the two test set compounds. The trends found in the training set reappear in Fig. 6. Recall that activities in the training set are biased towards either the strong or non-inhibitory groups. A similar trend emerges in the predicted activities of the solution set, with a gap in $IC_{50}$ values ranging between 300 and 600. This can be controlled with a larger training set.

The goal of the inverse-QSAR method was to predict, if any, novel inhibitory compounds possessing a lower $IC_{50}$ value than those in the training set. There were 77 new peptides classified as strong inhibitors. Of these, 12 represent peptides with predicted $IC_{50}$ values less than 40—the $IC_{50}$ value of peptide 16, which was the strongest inhibitor in the training set. To provide feedback on these predictions, we synthesized two of these peptides, sequences 2 and 10, using cellular assays. Their experimental $IC_{50}$ values were very close to our predicted values (see Table 7), and to our knowledge are the strongest inhibiting peptides to date that work in-vivo as well.

Although our training set was relatively small, it was nevertheless a sufficient basis from which the inverse-QSAR produced promising results. If possible, the training set should be larger with more diverse activities, but this is a challenge for experimental researchers who need to work with what they already have. Some of the predicted compounds exhibiting the desired activity can be used to construct another more focused training set from which higher-quality lead compounds can be designed. In this manner, old and new data can be exploited to refine the design process of more potent compounds.

## Acknowledgements

Funding for this work was provided by the US Department of Energy and Sandia National Laboratories under grant number DE-AC04-76DP00789. MDR, SM and JLF are also pleased to acknowledge funding provided by the US Department of Energy's Genomes to life program (http://www.doegenomestolife.org) under project, "Carbon Sequestration in Synechococcus Sp.: From Molecular Machines to Hierarchical Modeling" (http://www.genomes-to-life.org). The University of New Mexico would like to acknowledge that support for this project was in part funded by grants from the American Cancer Society RPG-00-096-01-LBC and the American Heart Associate 9960318Z.

## References

[1] W.P. Walters, M.T. Stahl, M.A. Murcho, Virtual screening—an overview, Drug Discov. Today 3 (4) (1998) 160–178.

[2] J. Mestres, Virtual screening: a real screening complement to high-throughput screening, Biochem. Soc. Trans. 30 (4) (2002) 797–799.

[3] E.C. Butcher, L.J. Picker, Lymphocyte homing and homeostasis, Science 272 (1996) 60–66.

[4] T.A. Springer, Traffic signals for lymphocyte recirculation and leukocyte emigration: the multistep paradigm, Cell 76 (1996) 301–314.

[5] T.M. Carlos, J.M. Harlan, Leukocyte–endothelial adhesion molecules, Blood 84 (1994) 2068–2101.

[6] J. Shannon, M. Silva, D. Brown, R.S. Larson, Novel cyclic peptide inhibits ICAM-1 mediated cell aggression, J. Pept. Res. 58 (2001) 40–150.

[7] J. Shannon, D.C. Brown, M. Silva, R.S. Larson, A cyclic peptide inhibits LFA-1/ICAM-1 mediated cell aggregation, J. Pept. Res. 58 (2001) 1–14.

[8] L. Sillerud, E. Burks, D.C. Brown, R.S. Larson, NMR-derived solution model of potent ICAM-1 inhibitory peptide, J. Pept. Res. 62 (2003) 97–116.

[9] S.H. Merchant, D.M. Gurule, R.S. Larson, Amelioration of ischemia-reperfusion injury with cyclic peptide blockade of ICAM-1, Am. J. Phys-Heart Circ. 284 (2003) H1260–H1268.

[10] T.A. Kelly, D.D. Jeanfarve, D.W. McNeil, et al., Cutting edge: a small molecule antagonist of LFA-1 mediated cell adhesion, J. Immunol. 163 (1999) 5173–5177.

[11] K. Last-Barney, W. Davidson, M. Cardozo, et al., Binding site elucidation of hydantoin-based antagonists of LFA-1 against multi-disciplinary technologies: evidence for the allosteric inhibition of a protein–protein interaction, J. Am. Chem. Soc. 123 (2001) 5643–5650.

[12] N. Trinajstic, Chemical graph theory, in: D.J. Klein, M. Randic (Eds.), Mathematical Chemistry, second ed., CRC Press, Boca Raton, FL, 1992.

[13] L.B. Kier, Indexes of molecular shape from chemical graphs, Acta Pharm. Jugosl. 36 (1986) 171.

[14] L.H. Hall, L.B. Kier, The molecular and connectivity chi indexes and kappa shape indexes in structure–property modeling, In: K.B. Lipkowitz, D.B. Boyd (Eds.), Reviews in Computational Chemistry, VCH Publishers, New York, 1991, p. 367–422.

[15] M. Randic, Graph valence shells as molecular descriptors, J. Chem. Inf. Comput. Sci. 41 (2001) 627–630.

[16] A.T. Balaban, Topological index J for heteroatom-containing molecules taking into account periodicities of element properties, Math. Chem. (MATCH) 21 (1986) 115–122.

[17] R.R. Parakulam, M.L. Lesniewski, K.J. Taylor-McCabe, C. Tsai, QSAR studies of antiviral agents using molecular similarity analysis and structure activity maps, SAR QSAR Environ. Res. 10 (1999) 175–206.

[18] T.D. Le, J.G. Weers, QSPR and GCA models for predicting the normal boiling points of fluorocarbons, J. Phys. Chem. 99 (1995) 6739–6747.

[19] H. Weiner, Structural determination of paraffin boiling points, J. Am. Chem. Soc. 69 (1947) 17–20.

[20] P.V. Khadikar, S. Sharma, V. Sharma, S. Joshi, I. Lukovits, M.A. Kaveeshwar, QSAR study of the effect of benzohydroxamic acids on DNA synthesis, Bull. Soc. Chim. Belg. 106 (12) (1997) 167–172.

[21] V. Venkatasubramanian, K. Chen, J.M. Caruthers, Evolutionary design of molecules with desired properties, J. Chem. Inf. Comput. Sci. 35 (1995) 188–195.

[22] R.P. Sheridan, S.K. Kearsley, Using the genetic algorithm to suggest combinatorial libraries, J. Chem. Inf. Comput. Sci. 35 (1995) 310–320.

[23] V. Kvasnicka, J. Pospichal, Simulated annealing construction of molecular graphs with required properties, J. Chem. Inf. Comput. Sci. 36 (1996) 516–526.

[24] J.-L. Faulon, Stochastic generator of chemical structure. 2. Using simulated annealing to search the space of constitutional isomers, J. Chem. Inf. Comput. Sci. 36 (1996) 731–740.

[25] L.H. Hall, R.S. Dailey, L.B. Kier, Design of molecules from quantitative structure-activity relationship models. 3. Role of higher order path counts: path 3, J. Chem. Inf. Comput. Sci. 33 (1993) 598–603.

[26] L.B. Kier, L.H. Hall, J.W. Frazer, Design of molecules from quantitative structure–activity relationship models. 1. Information transfert between path and vertex degree counts, J. Chem. Inf. Comput. Sci. 33 (1993) 143–147.

[27] L.B. Kier, L.H. Hall, J.W. Frazer, Design of molecules from quantitative structure–activity relationship models. 2. Derivation and proof of information transfert relating equations, J. Chem. Inf. Comput. Sci. 33 (1993) 148–152.

[28] M.I. Skvortsova, I.I. Baskin, O.L. Slovokhotova, V.A. Palyulin, N.S. Zefirov, Inverse problem in QSAR/QSPR studies for the case of topological indices characterizing molecular shape (Kier indices), J. Chem. Inf. Comput. Sci. 33 (1993) 630–634.

[29] J.-L. Faulon, D.P. Visco Jr., R.S. Pophale, The signature molecular descriptor. 1. Extended valence sequences vs. topological indices in QSAR and QSPR studies, J. Chem. Inf. Comput. Sci. 43 (2003) 707–720.

[30] J.-L. Faulon, C.J. Churchwell, J.D.P. Visco, The signature molecular descriptor. 2. Enumerating molecules from their extended valence sequence, J. Chem. Inf. Comput. Sci. 43 (2003) 721–734.

[31] D.P. Visco, R.S. Pophale, M.D. Rintoul, J.-L. Faulon, Developing a methodology for an inverse quantitative structure–activity relationship using the signature molecular descriptor, J. Mol. Graphics Model. 20 (#6/SISI) (2002) 429–438.

[32] M. Randic, On computation of optimal parameters for multivariate analysis of structure–property relationship, J. Comput. Chem. 12 (1991) 970–980.

[33] M. Randic, Resolution of ambiguities in structure–property studies by use of orthogonal descriptors, J. Chem. Inf. Comput. Sci. 31 (1991) 311–320.

[34] N.R. Draper, H. Smith, Applied Regression Analysis, third ed., Wiley, New York, 1998.

[35] I.G. Bashmakova, Diophantus and Diophantine Equations, Math. Assoc. Amer., Washington, DC, 1997.

[36] E. Contejean, H. Devie, An efficient incremental algorithm for solving systems of linear Diophantine equations, J. Inf. Comput. 113 (1) (1994) 143–172.

[37] M. Clausen, A. Fortenbacher, Efficient solution of linear Diophantine equations, J. Symbolic Comput. 8 (1/2) (1989) 201–216.