Classification of Some Active Compounds and Their Inactive Analogues Using Two Three-Dimensional Molecular Descriptors Derived from Computation of Three-Dimensional Convex Hulls for Structures Theoretically Generated for Them

Thy-Hou Lin,* Yih-Shiang Yu, and Hong-Jih Chen

Department of Life Science, National Tsing Hua University, Hsinchu, Taiwan, ROC

Received April 10, 2000

Two three-dimensional (3D) molecular descriptors are used to classify 73 protease inhibitors against the human immunodeficiency virus type 1 (HIV-1). X-ray structures of these HIV-1 protease bound inhibitors are used as templates to generate the most probable bioactive conformations of the inhibitors. A convex hull computation algorithm is applied to each structure generated. The frequency of atoms lying on the vertexes of each hull is counted. Vertexes of the same atomic charge state are then gathered together as a set of commonly exposed groups for all the structures generated. The first 3D descriptor is computed as the maximum molecular path length among any three distinct commonly exposed groups, while the second 3D one is computed as the maximum molecular path length among any three atoms of nonconvex hull vertexes. We find that the 73 HIV-1 protease inhibitors can be classified by the first 3D descriptor into two groups, which agrees with the result of visual classification using the activity data as a criterion for these compounds. The classification scheme is then used to classify a database of 427 active trypsin inhibitors and their inactive analogues. The structures of these compounds are generated theoretically from steps of energy minimization and molecular dynamics. Classification for all these compounds is performed using the SYBYL hierarchical clustering method on the first 3D descriptor and then the second 3D one computed. It is found that some inactive analogues are completely separated from the active inhibitors at the first stage of classification using the first 3D descriptor. Most of the highly active inhibitors are classified into a cluster at the second stage of classification using the second 3D descriptor. Finally, most of these highly active inhibitors are separated from all the accompanying inactive analogues in the cluster through a structural alignment process using a set of commonly exposed groups determined for them.

INTRODUCTION

There are several computational methods have been developed for exploring relationships between chemical structures and measured properties, especially biological effects of compounds.[1−4] Many of these methods have been reviewed recently. The basic problem is how to express an irregular object like a chemical structure in a regular form that allows the quantitative comparing and contrasting with other structures. One usually relies on using molecular descriptors,[5] which are numerical values representing selected features of the compounds. Each structure is represented as a list, or vector, of such numerical descriptors and thus may be thought of as a point in a high-dimensional space, with coordinates equal to the corresponding descriptor values. The problem of relating structure to biological activity becomes one of relating position (in the high-dimensional space) to activity. In general, molecular descriptors developed consist of two-dimensional (2D) or three-dimensional (3D) ones depending on the degree of complexity and dimensionality they represent.[6−8] The simplest molecular descriptors are counts of individual atoms, bonds, degrees of connectivity, etc. These can be extended to counts of rings, pharmacophore points, and any other feature that can be represented as a single node or arc in the graph or reduced graph representation of the molecule. Substantial molecular properties such as hydrophobicity, polarity, flexibility, shape, volume, hydrogen-bonding properties, and aromatic density have also been used as molecular descriptors.[9,10] In choosing a set of molecular descriptors for structure−activity studies or database classification or searching, one always needs to compromise efficiency, generality, ease of interpretation, and ease of automatic perception. Therefore, to select some highly effective descriptors, it is essential to perform some preliminary statistical analyses on the vast number of descriptors.[11]

Mathematical tools such as linear regression[12] and discriminant analysis[13] are often used to establish correlations between descriptor values and biological activity. To yield a mathematically well-behaved solution, the number of adjustable parameters must be kept rather small compared to the number of structures. Some sort of feature reduction processes are required if the number of descriptors that might be used greatly exceeds this number.[14−16] The selected descriptors are often those found to yield the best mathematical model after performing many trial-and-error runs. However, it has been shown[17] that this approach can introduce bias and can lead to artificially high “goodness of fit” parameters in linear regression analyses. For database classification and searching, it is found[18] that models based on complementary sets of descriptors would improve the reliability of predictions. Conventional descriptors such as

* Corresponding author. Telephone: (886) 03-574-2759. Fax: (886) 03-572-1746. E-mail: thlin@life.nthu.edu.tw.

10.1021/ci000328b CCC: $19.00
Published on Web 09/02/2000

molecular connectivity,[19] counts of self-avoiding paths,[20] and Moreaus's autocorrelation of topological molecular structure values[21] have the advantage that they can be computed easily from the connection table of a structure and can be applied to much more diverse sets of compounds. However, they are complex measures of the topology of the structure and, hence, may be hard to interpret even if they do correlate with activity. Hopfinger[22] has defined parameters from molecular shape analysis that relate to the space-filling characteristics and 3D electrostatic potential of molecules. While these can yield geometric models of biological activity, their computation requires a detailed conformational analysis of each structure in the series.

In this work, we describe two novel 3D descriptors that do not require feature selection and that deal uniformly with all the structures studied. Our 3D descriptors are based on the identification of structural convexity for 73 HIV-1 protease inhibitors[23−25] and then 427 structures in a data set consisting of 17 substance P antagonists[26] (designated as S structures), 32 melatonin agonists[27] (designated as M structures), 88 trypsin inhibitors[28] (designated as T structures), and 290 analogues of trypsin inhibitors searched from a MDL/ISIS database[29] (designated as Q structures). We use a 3D convex hull computation method published previously[30] to identify the exposed functional groups on the vertexes of a convex hull of each structure generated through molecular dynamics and energy minimization. A set of commonly exposed groups are determined for a series of structures from these vertexes by gathering together atoms of the same atomic charge state and of nonzero frequency identified as a convex hull vertex. By examining the set of commonly exposed groups determined for a series of structures, we define our first 3D descriptor as the maximum molecular path length computed among any three commonly exposed groups selected for each structure in the series. Our second 3D descriptor is defined as the maximum molecular path length computed among any three atoms of the nonconvex hull vertexes of each structure in the same series. The first 3D descriptor is used to perform a primary classification for all the structures. Clusters thus generated are classified further by the second 3D descriptor. We find that the 73 HIV-1 protease inhibitors can be classified into two groups by the first 3D descriptor, which agrees with the visual classification result using the activity data as a criterion.

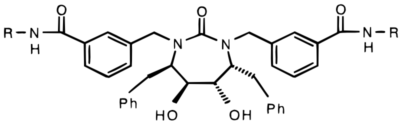
Classification for the set of 427 structures is performed by treating the 88 T structures as an active set. We find that at the first stage of classification all the T structures are separated from the sets of S and M structures. Further separation between some T and Q structures at the second stage of classification is evident since about half the total clusters generated at this stage contain no T structures. Most of the highly active T structures are also classified into a cluster at this stage. However, there are still some Q structures classified with the cluster. All the structures in the cluster are then aligned using some commonly exposed groups determined as a set of correspondence. Through computation of molecular path lengths among three commonly exposed groups selected for the aligned structures, most of the T structures in the cluster can be separated from the Q ones.

METHODS

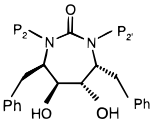
Construction of structures for 73 HIV-1 protease inhibitors was based on the two X-ray structures 1qbr[23] and 1dmp[24] which contain ligands XV638 ([(4*R*)−(4α,5α,6*β*,7*β*)-3,3′-[[tetrahydro-5,6-dihydroxy-2-oxo-4,7-bis(phenylmethyl)-1*H*-1,3-diazepine-1,3(2*H*)-diyl]bis(methylene)]bis[*N*-2-thiazolylbenzamide]) and DMP450 ((4*R*,5*S*,6*S*,7*R*)-hexahydro-5,6-dihydroxy-1,3-bis[(3-aminophenyl)methyl]-4,7-bis(phenylmethyl)-2*H*-1,3-diazepin-2-one), respectively, available in the protein data bank.[31] The XV638 structure was used as a template for constructing structures of the first 27 inhibitors while that of DMP450 was used for constructing structures of the remaining 46 inhibitors (Table 1). The construction of the molecules was done within the active site of the HIV-1 protease by replacing side chains of the template molecule as has been described previously by other group.[25] Structures generated were extracted from each protein−ligand complex using the SYBYL FlexiDock module.[32] Structures in each set were aligned against that of the first inhibitor of the set using the SYBYL FIT module[32] and a set of commonly exposed groups determined as the set of correspondence as described below. The goodness of each alignment was examined using the SYBYL CoMFA (comparative molecular field analysis) module[32] and the default settings within the program. Construction of 17 S[26] (molecules CP99994, MOLp, MOLh, MOLc, MOLo, CP96345, MOLe, MOLn, MOLl, MOLd, MOLf, MOLa, CGP47899, MOLm, L-73224, CIBA_45, MOLi), 32 M[27] (compounds 1, 2, 3, 5, 6, 8, 10, 12, 13, 17, 20, 21, 22, 27, 30, 31, 32, 36, 38, 39, 42, 45, 51, 55, 57, 58, 59, 60, 61, 62, 63, 64), and 88 T[28] (compounds 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72 of Table 1; and compounds 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88 of Table 5 of ref 28) structures were performed using the SYBYL 6.5 fragment database.[32]

All structures were subjected to 4000 steps of energy minimization and then 50 ps of molecular dynamics (MD) simulation using the TRIPOS force field.[32] A distance-dependent dielectric constant ($\epsilon = 32r$) where r was the distance was used, and the nonbonded cutoff was set at 8 Å. The 290 Q structures were searched from a MDL/ISIS database[29] using partial structures depicted in Tables 1 and 2 of ref 24 as queries. These searched structures were also constructed by the SYBYL 6.5 fragment library[32] and subjected to the same steps of energy minimization and molecular dynamics simulation. Computation of a convex hull[30] for each of the 427 structures generated was then performed. The Gasteiger−Huckel[33] option of the SYBYL 6.5 program[32] was used to calculate the atomic charge for all the structures. These charge values were sorted to find the maximum and the minimum ones of a range. A series of hypothetical charges were calculated by dividing the range by a number selected as the number of atomic charge state. The number of atomic charge state was set as 25, 17, or 5 for the sets of 27, 46, and 427 structures, respectively. A hypothetical charge state for each vertex atom was assigned by comparing the actual atomic charge with the hypothetical charges calculated. Each hypothetical charge state was

Table 1. Classification of Two Sets of HIV-1 Protease Inhibitors Using Values of the First 3D Descriptors Computed

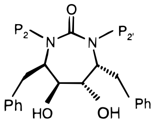


no.	R	K_i (nM)	commonly exposed groups	molecular path length[a]
86	2-(5-CH_3-pyridinyl)	0.011	11, 12, 21, 22, 27, 28, 29, 30, 35, 36, *44*, 45, 49, *50*, 53, 54, *55*, 56, 59	45(50^2)
89	2-(5-Cl-pyridinyl)	0.012	21, 22, 23, 28, 29, 30, *44*, *46*, 49, *50*, 53, 54	41(37)
102	2-imidazolyl	0.014	11, 12, 21, 22, 23, 27, 28, *29*, 30, 35, 36, 49, *50*, *53*	37(32)
88	2-(4,6-di-CH_3-pyridinyl)	0.016	11, 12, 21, *22*, 23, 29, 30, 35, 36, 44, *45*, *49*, 55	32(28)
87	2-(6-CH_3-pyridinyl)	0.020	11, 12, 21, *22*, 23, 27, 28, 29, 30, 35, *46*, 49, *50*, 53, 55, 58, 59	35(39)
69	OH	0.020	*11*, 12, 22, 23, 27, 28, *29*, 30, 35, 36, 47, *48*	29(33)
103	2-benzimidazolyl	0.024	11, 12, 21, 22, 23, 28, 29, 30, 35, 36, 49, *50*, 56, *57*, 58, 62, *63*, 64	45(46)
85	2-(4-CH_3-pyridinyl)	0.027	11, 12, 21, *22*, 23, 28, 29, 30, 35, 36, 45, *46*, 49, *50*, 54, 55, 58	35(38)
91	2-(5-Br-pyridinyl)	0.035	21, *22*, 23, 27, 28, 29, 30, *47*, 49, *50*, 54, 56, 58	34(36)
67	H	0.039	*11*, 12, 21, 22, 23, 24, 27, 28, 29, 30, *35*, 36, 46, *47*	29(34)
83	2-pyridinyl	0.043	11, 12, 22, 23, *29*, 30, 35, 36, *44*, 45, *49*, 55	35(39)
70	OCH_3	0.045	12, 21, *22*, 27, 28, 29, *35*, 36, 41, 47, *48*	29(32)
79	CH_2CN	0.063	11, 12, *17*, 21, 22, 23, 28, 29, 30, 36, *41*, 49, *50*	37(42)
71	CH_3	0.066	*11*, 12, 21, 22, 23, 28, 29, 30, *35*, 36, 45, *46*	29(35)
93	2-(5-CF_3-pyridinyl)	0.085	21, 22, 23, 28, 29, 30, *44*, *46*, 49, *50*	41(37)
92	2-(4-CH_3-pyrimidinyl)	0.115	21, 22, 23, 27, 28, 29, 30, *46*, 49, *50*, 53, *58*	45(42)
78	CH_2CF_3	0.210	10, 12, *13*, 19, 20, 21, 26, 27, *34*, 35, 41, *42*	29(28)
72	CH_2CH_3	0.210	12, 13, 14, *15*, 21, 22, 23, 28, 29, 30, 35, 36, 37, *38*, 43, *44*	33(31)
90	2-(3,5-di-Cl-pyridinyl)	0.245	11, 12, *17*, 21, 22, 23, 27, 28, *29*, 30, 49, *50*	33(38)
84	2-(3-CH_3-pyridinyl)	0.260	12, 21, *22*, 23, 28, 29, 30, 31, *35*, 36, 47, *48*	29(31)
82	3-pyridinyl	0.290	21, 22, 23, 28, *29*, 30, *44*, 49, *50*, 53	35(39)
74	$CH_2CH_2CH_3$	0.359	12, 21, 22, 23, 28, 29, 36, *44*, *46*, 47, 49, *50*, 54, 55	41(44)
81	4-pyridinyl	0.410	12, *13*, 20, 21, 22, 27, 28, 29, 35, *36*, 43, *44*	29(28)
75	$CH_2CH_2CH_2CH_3$	0.424	11, 12, 21, 22, 23, 27, 28, 29, 30, 35, 36, *47*, 49, *50*, *56*	45(34)
80	benzyl	0.430	11, 14, *15*, 21, 22, 27, 28, 29, 32, 34, 36, 37, *38*, 43, *44*	37(34)
73	$CH(CH_3)_2$	0.579	12, *22*, 23, 27, 28, 29, 30, *35*, 36, 49, *50*	29(32)
76	$C(CH_3)_3$	2.400	11, 12, 21, 22, 23, 28, 29, 30, 35, 36, 45, 47, 49, *50*, 54, 55, *56*, *57*, 58	45(43)



no.	$P_2/P_{2'}$	K_i (nM)	commonly exposed groups	molecular path length
53	*m*-hydroxybenzyl	0.12	*12*, 13, *17*, 21, 22, 23, 28, 29, 30, 33, 34, 39, *40*	29(29)
51	*m*-(hydroxymethyl)benzyl	0.14	*11*, 12, 21, 22, 23, 28, 29, 30, *35*, 36, 49, *50*	29(34)
54	*m*-aminobenzyl	0.28	11, 12, 21, 22, 23, 27, 28, 29, 30, 35, 36, 44, *45*, *46*, 47, 49, *50*, 51, 52, 53, 54, 55, 56, 57, 58	47(41)
50	*p*-(hydroxymethyl)benzyl	0.34	11, 12, 21, *22*, 23, *29*, 30, 35, 36, 49, *50*	28(36)
49	*m*-iodobenzyl	0.42	12, 13, 16, 21, *22*, 23, 28, *29*, 30, 33, 34, 39, *40*	28(35)
37	*m*-chlorobenzyl	0.89	16, *17*, 18, 23, *24*, 25, 35, *36*	28(36)
23	cyclobutylmethyl	1.3	11, 13, 21, *22*, 23, 28, *29*, 30, 33, 34, 39, *40*	28(36)
39	*m*-bromobenzyl	1.4	15, *16*, 17, 22, *23*, 24, 33, *34*	28(36)
4	*n*-butyl	1.4	10, 12, 13, *14*, 19, 20, 21, 26, 27, 28, *34*, 36, 41, *42*	29(30)
46	*m*-methoxybenzyl	1.6	16, *17*, 18, 23, *24*, 25, 35, *36*	28(36)
5	*n*-pentyl	1.6	*12*, 16, 17, 18, 23, 24, 25, *31*, 35, *36*	29(25)
19	isoprenyl	1.8	2, 8, 15, *16*, 17, 22, *23*, 24, 27, 33, *34*	28(36)
22	cyclopropylmethyl	2.1	*11*, 12, 21, 22, 23, 27, 28, 29, 30, *35*, 36, 49, *50*	29(35)
48	*m*-nitrobenzyl	2.8	21, *22*, 23, 28, *29*, 30, 35, 36, 43, *44*	28(36)
27	benzyl	3.0	11, 12, 13, 16, 17, 21, *22*, 23, 27, 28, *29*, 30, 33, 37, *38*	28(36)
34	*m*-fluorobenzyl	3.0	11, 12, 13, 18, *19*, 20, 21, 25, *26*, 27, 31, 33, 34, 39, *40*	28(35)
24	cyclopentylmethyl	4.3	11, 12, 13, 18, *19*, 20, 21, 25, *26*, 27, 31, 33, 34, 39, *40*	28(36)
6	*n*-hexyl	4.6	17, *18*, 19, 23, 24, *25*, 26, 37, *38*	28(35)
38	*p*-chlorobenzyl	5.2	10, 12, 15, 16, 21, *22*, 23, 28, *29*, 30, 39, *40*	28(36)
17	allyl	5.2	9, 10, 14, *15*, 16, 21, *22*, 23, 27, 31, *32*	28(36)
42	*p*-methylbenzyl	5.7	9, 10, 11, 12, *13*, 15, 16, 21, 22, 23, 28, *29*, 30, 39, *40*	29(30)
41	*m*-methylbenzyl	7.0	12, *13*, 21, 22, 23, 28, *29*, 30, 33, 34, 39, *40*	27(27)
13	isohexyl	7.0	21, *22*, 23, 28, *29*, 30, 37, *38*	28(36)
18	2-methylpropen-3-yl	7.3	2, 8, 9, 15, *16*, 17, 22, *23*, 24, 33, *34*	28(36)
3	*n*-propyl	8.0	14, *15*, 16, 21, *22*, 23, 31, *32*	28(36)
12	isopentyl	12	16, *17*, 18, 23, *24*, 25, 35, *36*	28(36)
43	*m*-(trifluoromethyl)benzyl	22	21, *22*, 23, 28, *29*, 30, *45*, 46	28(35)
40	*p*-bromobenzyl	27	11, 21, *22*, 23, 27, 28, *29*, 30, 34, 39, *40*	28(35)

Table 1. (Continued)



no.	$P_2/P_{2'}$	K_i (nM)	commonly exposed groups	molecular path length
14	isoheptyl	30	21, *22*, 23, 28, *29*, 30, 39, *40*	28(35)
33	*o*-fluorobenzyl	34	11, *12*, 13, 18, 19, 20, 24, 25, 26, 27, 32, *33*, 34, 39, *40*	28(29)
16	neohexyl	36	17, *18*, 19, 24, *25*, 26, 27, 37, *38*	28(36)
25	cyclohexylmethyl	37	18, *19*, 24, 25, *26*, 27, 39, *40*	28(35)
11	isobutyl	49	15, *16*, 17, 22, *23*, 24, 33, *34*	28(37)
44	*p*-(trifluoromethyl)benzyl	51	21, *22*, 23, 28, *29*, 30, 45, *46*	28(36)
20	$CH_2CH_2OCHCH_2$	60	10, *11*, 12, 16, 17, 18, 23, 24, 25, 29, 30, *31*, 35, *36*	30(28)
2	ethyl	100	13, *14*, 15, 20, *21*, 22, 29, *30*	28(36)
15	isooctyl	110	21, *22*, 23, 28, *29*, 30, 41, *42*	28(36)
47	*p*-methoxybenzyl	157	10, 11, 13, *14*, 19, 20, 21, 26, 27, 28, 33, *36*, 41, *42*	33(32)
36	*o*-chlorobenzyl	240	10, 11, *12*, 13, 16, *17*, 21, 22, 23, 28, 29, 30, 33, 39, *40*	29(29)
7	*n*-heptyl	260	18, *19*, 20, 25, *26*, 27, 39, *40*	28(36)
8	$CH_2CH_2OCH_3$	800	15, *16*, 17, 18, 22, *23*, 24, 33, *34*	28(36)
9	$CH_2CH_2OCH_2CH_3$	1100	16, *17*, 18, 19, 22, 23, *24*, 25, 29, 35, *36*	28(36)
45	*o*-methoxybenzyl	1870	*12*, 13, 19, 20, 21, 26, 27, 28, 32, 33, 34, *36*, 41, *42*	26(27)
26	*N*-morpholino-2-ethyl	4000	*11*, 12, 19, 20, 21, 25, 26, 27, 28, *34*, 35, 41, *42*	29(25)
1	methyl	5700	12, *13*, 14, 19, *20*, 21, 27, *28*	28(36)
10	$CH_2CH_2OCH_2CH_2OCH_3$	7700	*13*, 18, 19, 20, 25, 26, 27, *31*, 39, *40*	31(32)

[a] Values of the first 3D descriptor are computed as the maximum molecular path length among three distinct commonly exposed groups determined which are marked as italic digits in the table. [b] This is the descriptor value of an atom triangle formed by the three distinct commonly exposed groups described above. The descriptor value is computed as the sum of the three interatomic distances among the three commonly exposed groups.

represented as a character variable in FORTRAN. The commonly exposed groups were collected as vertex atoms on each structure in the set that had the same hypothetical charge states assigned. In other words, each of these commonly exposed groups identified would appear as a vertex at least once on each of the structures in the set. While the number of commonly exposed groups identified for each structure in the set might be varied, the actual number of hypothetical charge states determined for each structure in the set was the same and was smaller than the number of atomic charge state originally selected.

To compute the molecular path length, each structure was treated as a graph[34] in which atoms were nodes and chemical bonds were edges. Each pair of nodes were connected by an edge. A walk of a graph was an alternating sequence of nodes and edges that begins and ends with nodes. A walk became a path if all the nodes traversed were distinct. The length of a path was the number of all the edges in it. The interconnections or the topology of each graph was represented by a connection table.[35] The connection table was then used to compute the path length between each pair of nodes in each graph. This result was saved as a file for each graph. This file and that of the commonly exposed groups determined were read into a program in which the molecular path length among any three distinct commonly exposed groups (e.g., the molecular path length was computed as a sum of path lengths among groups 1−2, 1−3, and 2−3) selected for each structure was calculated. The first 3D descriptor was defined as the maximum path length identified among all such path lengths computed for each structure (Figure 1). The three interatomic distances among the three commonly exposed groups for which the first 3D descriptor was defined and computed were also calculated and summed for the 73 HIV-1 protease inhibitors. These were the descriptor values of so-called atom triangle.[9,38] To compare

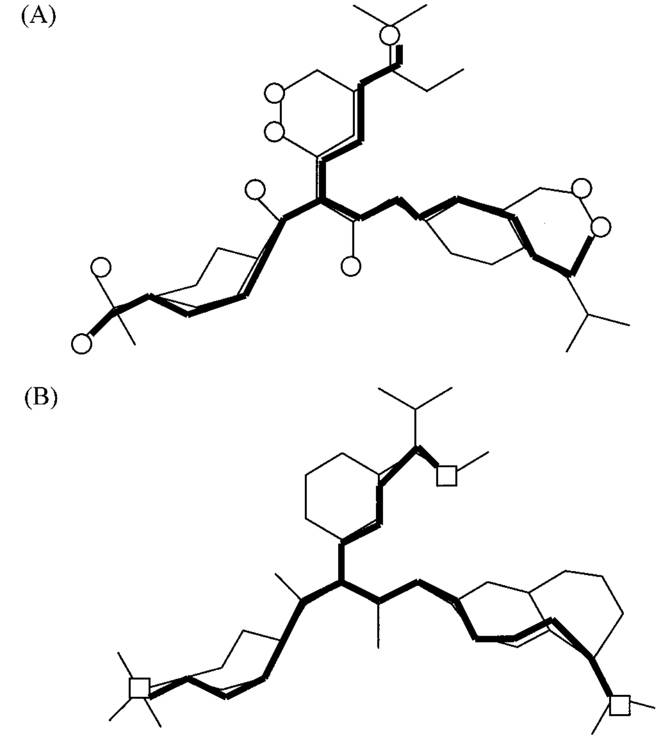


Figure 1. Schematic representation of (A) the first 3D and (B) the second 3D descriptor defined and computed for a hypothetical structure. A set of commonly exposed groups represented as open circles in (A) are determined first for the hypothetical structure. The first 3D descriptor represented by a thick line on the hypothetical structure is then computed as the maximum molecular path length among three distinct commonly exposed groups determined. The second 3D descriptor also represented by a thick line on the hypothetical structure in (B) is computed as the maximum molecular path length among any three distinct atoms of nonconvex hull vertexes (represented by open squares) determined for the structure.

with the pK_i values measured,[25] both the first 3D ones and the descriptor values of atom triangle[9,38] were scaled as follows:

$$f_i = \frac{D_i - D_{\min}}{D_{\max} - D_{\min}}$$

where f_i was the ith scaled descriptor value, D_i was the ith original descriptor value, and $D_{\max}$ and $D_{\min}$ were, respectively, the maximum and minimum descriptor values identified among all the original ones.

Values of the first 3D descriptor computed for the 427 structures were classified into several clusters using the SYBYL hierarchical clustering module.[32] The method for linking clusters chosen was Single and the input descriptor values were treated as Normal.[32] Each analysis result was generated as a table in which groupings of the descriptor values were listed as a function of the cluster levels given in the first column. The best grouping of the given cluster levels was chosen by inspecting a dendrogram automatically presented by the program after each analysis. The underlying principles on which the clustering method based are described in detail in the manual of SYBYL ligand-based design.[32] Each cluster thus generated was further classified using the second 3D descriptor. The second 3D descriptor was computed as the maximum molecular path length among any three distinct atoms of nonconvex hull vertexes on each structure (Figure 1). Most of the highly active T structures were classified into a cluster at this stage of classification. To further classify the cluster that containing most of the highly active T structures, a set of commonly exposed groups was determined by setting the number of charge state as 8 for all the structures in the cluster. Some of these commonly exposed groups were treated as a set of correspondence for aligning all the structures in the cluster using the FIT option of the SYBYL 6.5 program.[32] A T structure (T11, the one with the largest pK_i value measured) was treated as a common template, and then a pairwise matching was carried out for each other structures in the cluster against it. The aligned structures were then classified using molecular path lengths computed among three commonly exposed groups selected for each structure in the cluster.

RESULTS AND DISCUSSION

By treating a convex hull as a set of correspondence between matching points, it is possible to utilize it as a tool to preferentially screen conformational space for aligning structures generated for rather large and flexible molecules.[36] In principle, structures with exact similarity could give rise to exactly similar convex hulls computed. Therefore, the number of exposed groups that are common to several convex hulls computed for several structures can be used to examine the extent of structural similarity between them. The commonly exposed groups can be collected as vertex atoms of the same atomic type or the same atomic charge state. While the number of atomic type is fixed, the number of atomic charge state is a variable which varies with the number of structures studied. No commonly exposed groups will be determined if the number of atomic charge state selected is too large. Conversely, a small such number selected would result in too much generality in the commonly exposed groups determined. In general, a large number of atomic charge state should be used for structures of greater similarity and within the same series. This number is set as 17 and 25 for the sets of 46 and 27 HIV-1 protease inhibitors (Table 1) studied such that the number of commonly exposed groups determined varies from 7 to 16 and from 10 to 21 for the former and the latter sets, respectively.

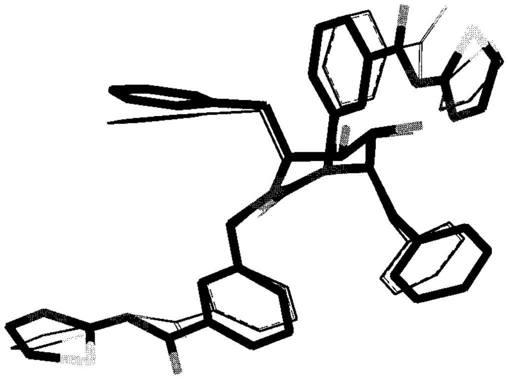


Figure 2. Superposition of the theoretically generated structure on the X-ray one of compound XV638.[23] The X-ray structure[23] is represented by a thick line while the theoretically generated one is represented by a thin line.

Figure 2 presents a comparison of two structures of compound XV638 in which one is determined by X-ray crystallography[23] and the other one is generated theoretically. The X-ray structure of compound DMP450[24] is used as a template for constructing the two side chains P_2 /$P_{2'}$ (*N*-2-thiazolylbenzamide) of the generated structure such as those depicted in Table 1 for the set of 46 inhibitors. The construction of the generated structure is done within the active site of the HIV-1 protease and is subjected to the same steps of energy minimization as those used for the set of 46 inhibitors. The generated structure is then aligned against the X-ray one using the SYBYL FIT module[32] and several commonly exposed groups determined for both as a set of correspondence. The root-mean-square value obtained after superposition of the two structures is 0.82. Apparently, the overall conformation of the generated structure agreed with the X-ray one although there are some differences in detail features of the two. A CoMFA[32] analysis on each set of structures generated and aligned using each set of commonly exposed groups determined as a set of correspondence gives values of cross-validated q^2 as 0.56 (46 inhibitor set) and 0.54 (27 inhibitor set), respectively. However, the number of atomic charge state is set as 8 for a combination of all 73 inhibitors which generates a range of commonly exposed groups from 8 to 25. A smaller number of atomic charge state is necessary to collect enough commonly exposed groups for each structure for classification since the diversity in structural feature is greatly increased if all 73 inhibitors are combined in one set.

The cross-validated q^2 computed for the aligned 73 structures by the set of commonly exposed groups against the structure of compound 86 (Table 1), which is the most active one in the series, is around 0.50. Values of the first 3D descriptor computed for all the aligned structures vary from 26 to 47 as shown in Table 1. For easy comparison with values of the first 3D descriptor, all the activity data (K_i values) in the table are arranged in descending order. This result shows that the activity data of all the inhibitors

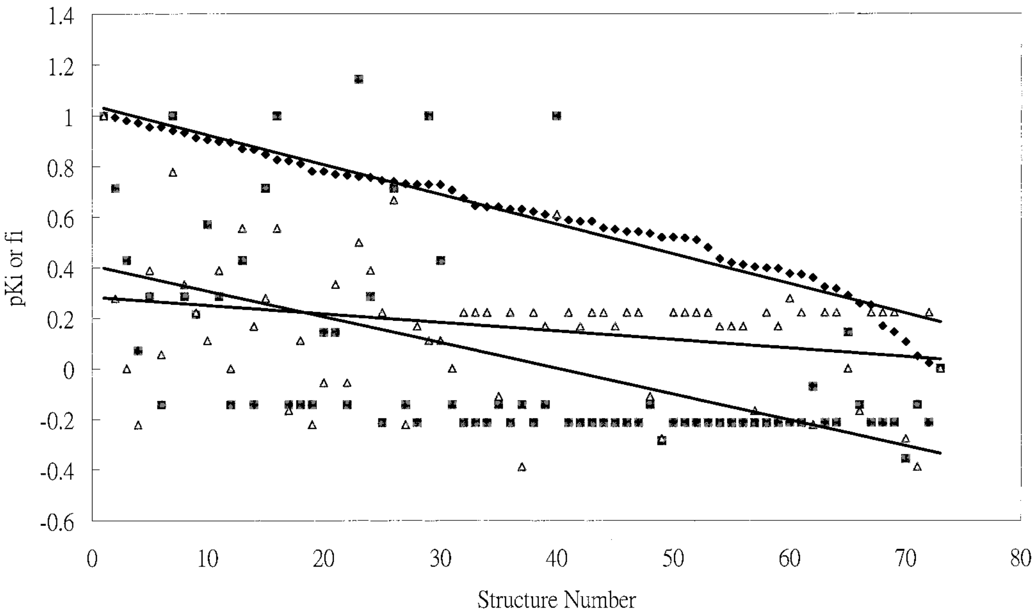


Figure 3. The measured[25] pK_i values (represented by solid diamonds), the scaled values of the first 3D descriptor (represented by solid squares), and the scaled values of the atom triangle descriptor (represented by triangles) plotted against the structure number of the 73 HIV-1 protease inhibitors. The trend of each data set is guided by a straight line computed through linear regression of the data set to the line.

Table 2. Commonly Exposed Groups Identified and Used in the Structural Alignment for Some T Structures and the Molecular Path Lengths Computed among Three Vertexes Selected for the Structures

structure	pK_i	alignment rule[a]	vertexes selected[b]	molecular path length
T11a	7.69	6, 10, 19, 20, 22, 23, 39	*32*, *20*, *41*	44
T9a	7.44	6(6), 19(19), 20(20), 22(22), 39(39)	*31*, *20*, *41*(0.90)[c]	46
T13a	6.85	6(6), 19(19), 10(20)	*32*, *10*, *40*(0.77)	44
T74a	6.58	6(6), 19(19), 10(20), 23(23)	*32*, *10*, *35*(1.38)	40
T78a	6.35	6(6), 19(19), 20(20), 22(22), 23(23)	*32*, *20*, *35*(1.55)	39
T22a	5.96	6(6), 19(19), 20(20), 23(22), 24(23), 40(39)	*33*, *20*, *42*(0.81)	38
T27a	5.66	6(6), 19(19), 20(20), 22(22), 23(23), 40(39)	*32*, *20*, *35*(1.29)	38
T59a	5.10	6(6), 19(19), 21(20), 23(22), 24(23)	*33*, *21*, *38*(1.85)	38
T88a	4.33	6(6), 19(19), 21(20), 23(22), 24(23)	*30*, *19*, *31*(1.76)	34
T68a	3.85	6(6), 19(19), 21(20), 22(22), 23(23)	*32*, *21*, *25*(2.04)	32

[a] Each structure is aligned against T11a, and the corresponding commonly exposed groups used for aligning the structure are parethesized. [b] Convex hull vertexes selected for computation of the molecular path length are represented with italic digits while those selected from the nonconvex hull ones are not. [c] The rms values of each aligned structure against T11a computed by the SYBYL Fit module[12] are parenthesized.

are correlated substantially with values of the first 3D descriptor computed. There is an apparent separation in the activity data[25] of both subsets (46 and 27) of inhibitors, although all of them can bind with the HIV-1 protease. Such a separation can be also seen in the values of the corresponding first 3D descriptor computed.

A comparison of values of pK_i and f_i computed from the first 3D descriptor and the atom triangle one for the 73 inhibitors is presented in Figure 3. A linear regression line computed and drawn through each data set is used to guide the trend of each data set. Values of the first 3D descriptor appear to be more similar in the trend than that of the atom triangle one to that of the pK_i, suggesting that the first 3D descriptor is superior in representing the data set to the conventional 3D one, the atom triangle descriptor.

Classification for 427 structures is performed next to test the classification ability of our 3D molecular descriptors for structural sets containing more diversified structures. The number of atomic charge state is set as 5 for all the 427 structures classified since this would guarantee that every structure has at least three commonly exposed groups determined. Since the largest pK_i measured is due to structure T11,[28] every other structure in the series is aligned against it using some commonly exposed groups selected as a set of correspondence (Table 2).

No further energy minimization is performed for each aligned structure. The goodness of each alignment is judged by the rms value computed (Table 2) for the matched atoms as given in the FIT option of the SYBYL 6.5 program.[32] By examining these aligned structures carefully, we have picked three vertex atoms on each structure so that the corresponding molecular path lengths computed among them can correlate with the pK_i values (Table 2). The pK_i values of 10 selected T structures are arranged in descending order and so are the molecular path lengths computed for each of them (Table 2). The number of atomic charge state is set as 10 for these 88 T structures. Apparently, the pK_i values and the molecular path lengths are highly correlated with each other since a correlation coefficient of 0.96 is computed for them (Table 2). Having classified these molecular path lengths using the SYBYL hierarchical clustering method,[32] we split the 88 T structures into several clusters as shown in Figure 4. The

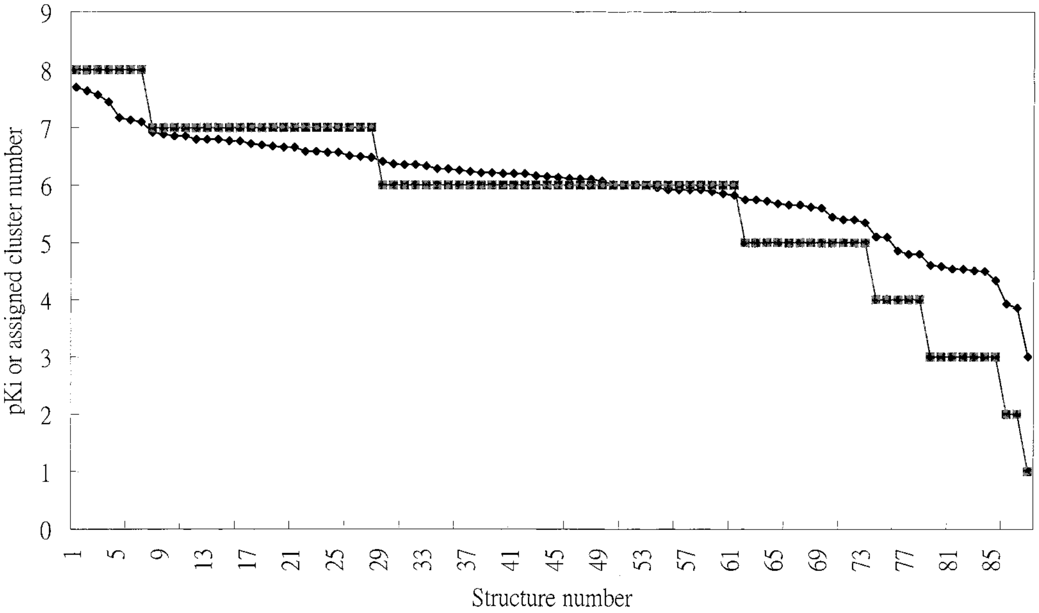


Figure 4. The sorted measured[28] pK_i values and the assigned cluster numbers for the 88 T structures classified using the SYBYL hierarchical clustering method[32] on the molecular path length computed among three commonly exposed groups determined plotted against the structure number. These structures have been aligned using a set of commonly exposed groups determined as a set of correspondences. The sorted pK_i values[28] are represented with the diamond points while the assigned cluster numbers are represented with the square points. The total number of clusters classified is eight.

Table 3. Classification of all 427 (S, M, T, and Q) Structures Using Values of the First 3D Descriptor Computed

structure series	clusters generated	S	M	T	Q
a	a1(186)[a]	2	0	58	126
	a2(37)	8	23	0	6
	a3(204)	7	9	29	159
c	c1(156)	2	0	52	102
	c2(58)	11	25	3	19
	c3(213)	4	7	32	170

[a] The number of structures classified in a cluster is parenthesized.

number of clusters classified is eight, and each member (structure) in a cluster is assigned a cluster number (Figure 4). The cluster number assigned for each structure is plotted against the structure number and so is the pK_i value of each structure as shown in Figure 4. These two plots are very similar to each other, which would imply that the pK_i value of each molecule within the set can be well represented using the molecular path lengths computed for them.

To classify a mixture of 427 structures of S, M, T, and Q series, the molecular path length among any three distinct commonly exposed groups selected on each structure is computed and sorted to find the maximum one. These maximum molecular path lengths are then treated as the first 3D descriptors and are classified by the SYBYL hierarchical clustering program[32] into several clusters (Table 3). All the structures are designated as the "a" series, and the number of clusters generated at this stage is three (Table 3). It appears that almost all the S and M structures at the stage are separated from the T and Q ones as indicated by the number of T structures classified in each cluster (Table 3). Since the 3D descriptor we describe here strongly depends on the conformation of a structure generated, we have also applied the same classification scheme to structures of the same series but generated with different conformations and designated as the "c" series in Table 3. The SYBYL Graphics toolbox[32] is used to directly modify the rotatable bonds on each "a" structure for generating a "c" starting structure. A "c" starting structure is also generated from an "a" one through 2000 steps of MD simulation on the latter if there are no rotatable bonds available. Each of these "c" starting structures is then subjected to the same steps of energy minimization and MD simulation as those employed for generating the "a" ones. Apparently, there are also three clusters generated for the series and most of the S and M structures are also separated from the T and Q ones at the stage of classification (Table 3).

To further classify those clusters that contain both T and Q structures of both series, namely clusters a1, a3, c1, and c3 (Table 3), the second 3D descriptor (Figure 1), which is computed as the maximum molecular path length among any three distinct atoms of nonconvex hull vertexes, is used. The classification creates 16, 20, 8, and 23 subclusters for clusters a1, a3, c1, and c3, respectively (Table 4). About half the total clusters generated at this stage of classification containing no T structures (Table 4). At this stage, most of the highly active T structures are also separated from those of the Q ones (Table 4). For example, T structures in subcluster 3, 4, and 5 of a1 or in subclusters 1, 2, and 3 of c1 are completely separated from Q ones (Table 4).

Separation of the first 10 T and Q structures by the first and then the second 3D descriptors is illustrated in Table 5. The maximum molecular path lengths computed among three distinct commonly exposed groups and three distinct atoms of nonconvex hull vertexes are compared for the selected structures in the table. Judging by the difference in maximum molecular path lengths computed, a separation for both sets of structures (e.g., structures T1a, T6a, T9a, T10a, T11a, Q10a, Q23a, Q24a, Q25a, and Q26a from structures T2a, T3a, T4a, T5a, T7a, Q1a, Q2a, Q3a, Q4a, and Q5a) using the first 3D descriptor is evident (Table 5). However, there is a less obvious separation of structures T1a, T6a, T9a, T10a,

Table 4. Classification of Clusters a1, a3, c1, and c3 of Table 3 Using Values of the Second 3D Descriptor Computed[a]

a1																							
	1	2	3*	4	5	6	7	8	9	10	11	12	13	14	15	16							
S	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0							
M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0							
T	4	3	14	15	8	2	2	7	3	0	0	0	0	0	0	0							
Q	14	12	15	12	5	3	2	18	14	14	7	6	1	1	1	1							
a3																							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20			
S	4	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
M	0	0	5	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
T	1	0	0	1	0	0	9	4	4	4	4	2	1	0	0	0	0	0	0	0			
Q	19	17	20	26	18	2	4	0	1	1	6	2	0	10	10	10	4	4	3	1			
c1																							
	1	2	3*	4	5	6	7	8															
S	2	0	0	0	0	0	0	0															
M	0	0	0	0	0	0	0	0															
T	11	20	15	5	2	0	0	0															
Q	25	29	22	4	11	7	2	1															
c3																							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
S	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	0	1	5	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	1	0	0	1	0	7	4	4	3	3	8	1	0	0	0	0	0	0	0	0	0	0	0
Q	14	19	18	22	6	0	8	3	1	0	2	3	21	17	12	7	7	3	2	2	1	1	1

[a] The two clusters marked with an asterisk (∗) contain the highly active T structures and are further classified through structural alignment using the commonly exposed groups determined. The results are presented in Table 6.

and T11a from structures Q10a, Q23a, Q24a, Q25a, and Q26a using the second 3D descriptor (Table 5). Therefore, some of these structures are classified into the same cluster at the second stage. On the contrary, a good separation between structures T2a, T3a, T4a, T5a and T7a, and structures Q1a, Q2a, Q3a, Q4a, and Q5a is obtained at the second stage (Table 5). These structures are thus classified into different clusters at the stage. Note that magnitude of the second 3D descriptors of structures T2a, T3a, T4a, T5a, and T7a is larger than that of the first 3D ones of these structures (Table 5). This is in contrast to that observed for structures T1a, T6a, T9a, T10a, and T11a, for which the magnitude of the first 3D descriptors is larger than that of the second 3D ones.

Our intuition would tell us that the magnitude of the first 3D descriptor should be larger than that of the second 3D one since the former are computed from points that are more extreme than those from the latter. However, this is not always true since all the first 3D descriptors are computed from commonly exposed groups selected rather than from the whole convex hull vertexes determined. This fact is highlighted by a plot shown in Figure 5, in which the first 3D descriptors are sorted in magnitude with the second 3D ones and plotted against the structure number of all the T structures. This plot shows that while the magnitude of the second 3D descriptors varies between 25 and 45, that of the first 3D ones gradually increase from 22 to 54. The plots also reveal that the magnitude of the second 3D descriptors of about 40% of T structures is larger than or equal to that of the first 3D ones.

Since the success of a classification method depends on its ability to separate the active structures from those of the inactive ones,[7,8] we also examine whether the T structures classified using the two 3D descriptors can correlate with the activities (the pK_i values) of the structures. This is illustrated by a plot of the percentage of active T structures counted in a cluster against the percentage of T structures identified for the cluster as shown in Figure 6. A T structure is considered active if its corresponding measured pK_i value is greater than 6.0. These plots show that, for both series of structures classified by the two 3D descriptors, the number of active T structures classified in a cluster is correlated substantially with the total number of T structures classified for the cluster. In other words, the two 3D descriptors can effectively extract the active T structures from the mixed ones. In fact, at this stage, there are 14 or 15 of the highly active T structures being classified into a cluster for both series of structures (Table 4).

Activities of these T structures and the accompanying Q ones classified into the same clusters are gathered together and listed in Table 6. Note that structures T1a, T6a, T9a, T11a, T16a, T34a, T51a, T53a, T84a, T86a, Q177a, Q206a, Q249a, Q253a, Q260a, Q267a, Q271a, Q275a, Q282a, Q284a, and Q289a are repeatedly classified into the cluster for both series of structures. This implies that these structures are somewhat less flexible than those others classified into the same cluster. Figure 7 is a presentation of all 26 accompanying Q structures classified with the set of highly active T ones. It appears that most of these Q structures contain nearly the same structural or topological features as those appearing on the set of highly active T ones. This is in accord with the fact that the two 3D descriptors used to classify them are indeed a mixed type of pure structural (commonly exposed groups) and topological (molecular path length) features. We have shown in a previous report[36] that the commonly exposed groups can be used to align structures

Table 5. Separation of the First 10 T and Q Structures Using the First 3D and Then the Second 3D Descriptors Computed

structures	first 3D descriptor[a]	second 3D descriptor[a]	structures	first 3D descriptor[a]	second 3D descriptor[a]
T1a	43, 17, 29(40)	18, 31, 41(38)	T2a	37, 27, 12(28)	17, 29, 35(38)
T6a	17, 29, 39(42)	18, 24, 40(38)	T3a	11, 31, 13(16)	17, 27, 37(36)
T9a	40, 31, 13(40)	17, 27, 38(36)	T4a	31, 6, 13(30)	17, 27, 35(34)
T10a	18, 40, 31(42)	17, 28, 37(34)	T5a	37, 46, 13(30)	17, 29, 43(40)
T11a	31, 18, 39(42)	17, 28, 40(38)	T7a	17, 26, 6(28)	18, 29, 36(32)
Q10a	26, 2, 27(34)	5, 22, 24(24)	Q1a	19, 20, 2(23)	6, 16, 18(15)
Q23a	33, 2, 6(44)	5, 25, 32(38)	Q2a	1, 3, 15(16)	4, 6, 17(16)
Q24a	16, 25, 3(32)	5, 15, 22(26)	Q3a	2, 17, 19(22)	4, 6, 15(14)
Q25a	16, 3, 23(34)	6, 9, 19(22)	Q4a	2, 13, 16(18)	4, 6, 14(12)
Q26a	28, 22, 3(38)	5, 21, 26(30)	Q5a	18, 21, 2(24)	6, 8, 18(18)

[a] All the descriptor values or the maximum molecular path lengths among three distinct commonly exposed groups computed are parenthesized.

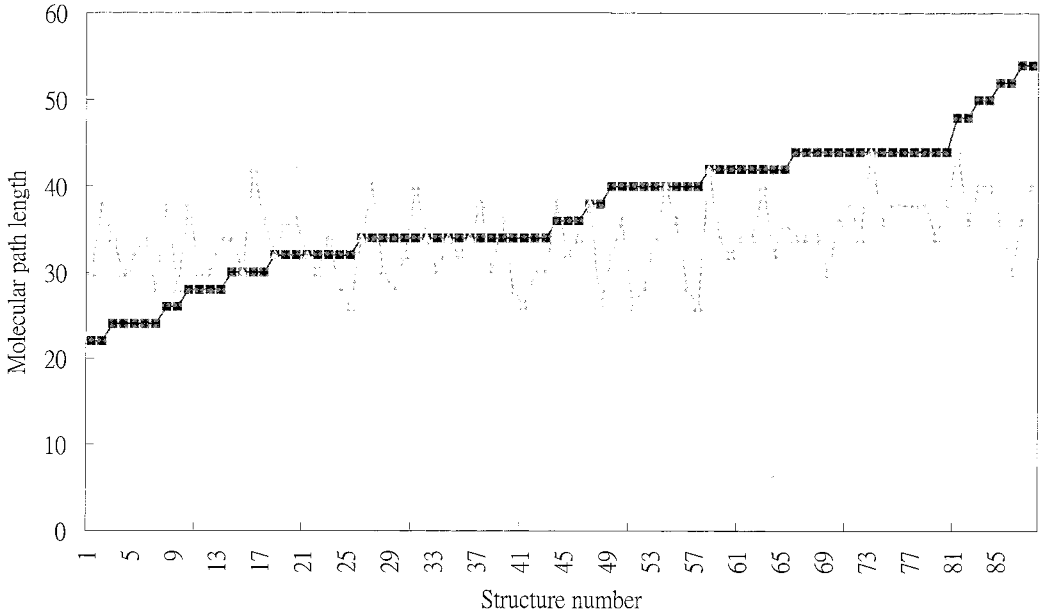


Figure 5. The first 3D descriptors computed as the maximum molecular path lengths among any three distinct commonly exposed groups selected for the 88 T structures plotted in increasing order against the structure number and represented by squares. The second 3D descriptors computed as the maximum molecular path lengths among any three atoms of nonconvex hull vertexes for the 88 T structures are plotted against the structure number and represented by triangles.

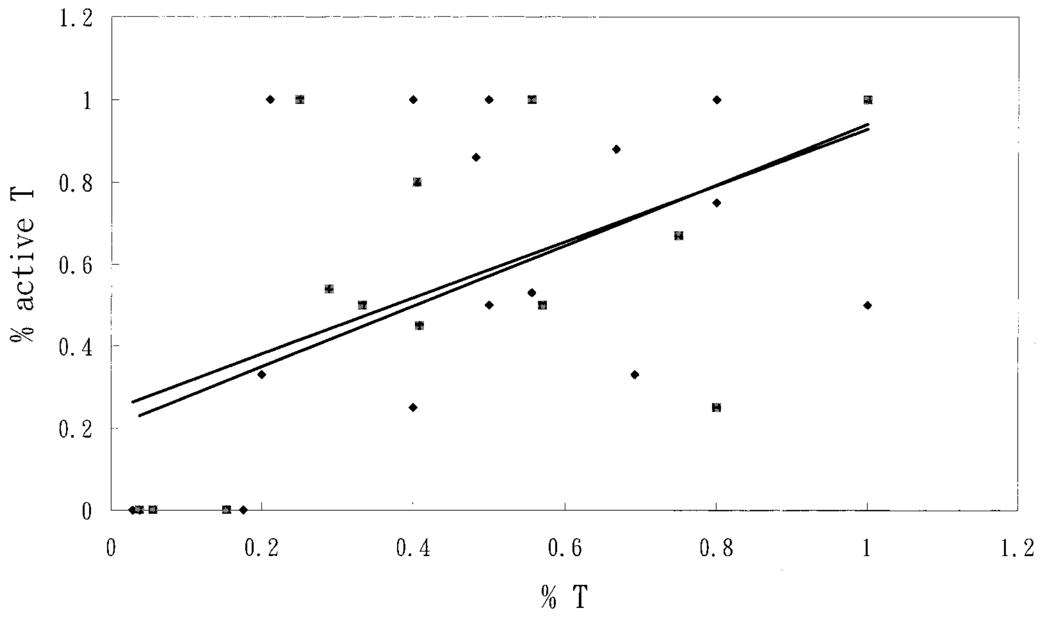


Figure 6. The percentage of active T structures in each cluster for clusters generated from classification of all 427 structures using the SYBYL hierarchical clustering method[32] on the first 3D and then the second 3D descriptors plotted against the percentage of T structures identified in each cluster. The results calculated for “a” series of structures are represented by solid diamonds while those calculated for “c” series of structures are represented by solid squares.

and to reflect the difference in biological activity caused by the difference in spatial locations of some functional groups. Therefore, it is possible to further separate the set of highly active T structures from the accompanying Q ones by aligning all these structures up using a set of commonly exposed groups determined for them. The three distinct commonly exposed groups selected for each structure for computation of a molecular path length among them are now also used for aligning each structure against that of the most active one. The three commonly exposed groups used for computation of molecular path lengths are selected from the most similar part of the two aligned structures. This is a rather easy task for the aligned T structures.

However, the commonly exposed groups determined for Q structures are more diversified as compared with those determined for T ones. Each of these Q structures is fitted to that of the most active T one several times using different combinations of groups of the commonly exposed groups determined for the structure as sets of correspondence. The three commonly exposed groups selected for computation of the molecular path length are then identified from the set that gives the best fit of the structure to that of the most active T one. The three commonly exposed groups selected and their corresponding molecular path lengths computed for all the aligned structures are listed in Table 6. These molecular path lengths are not further classified using a

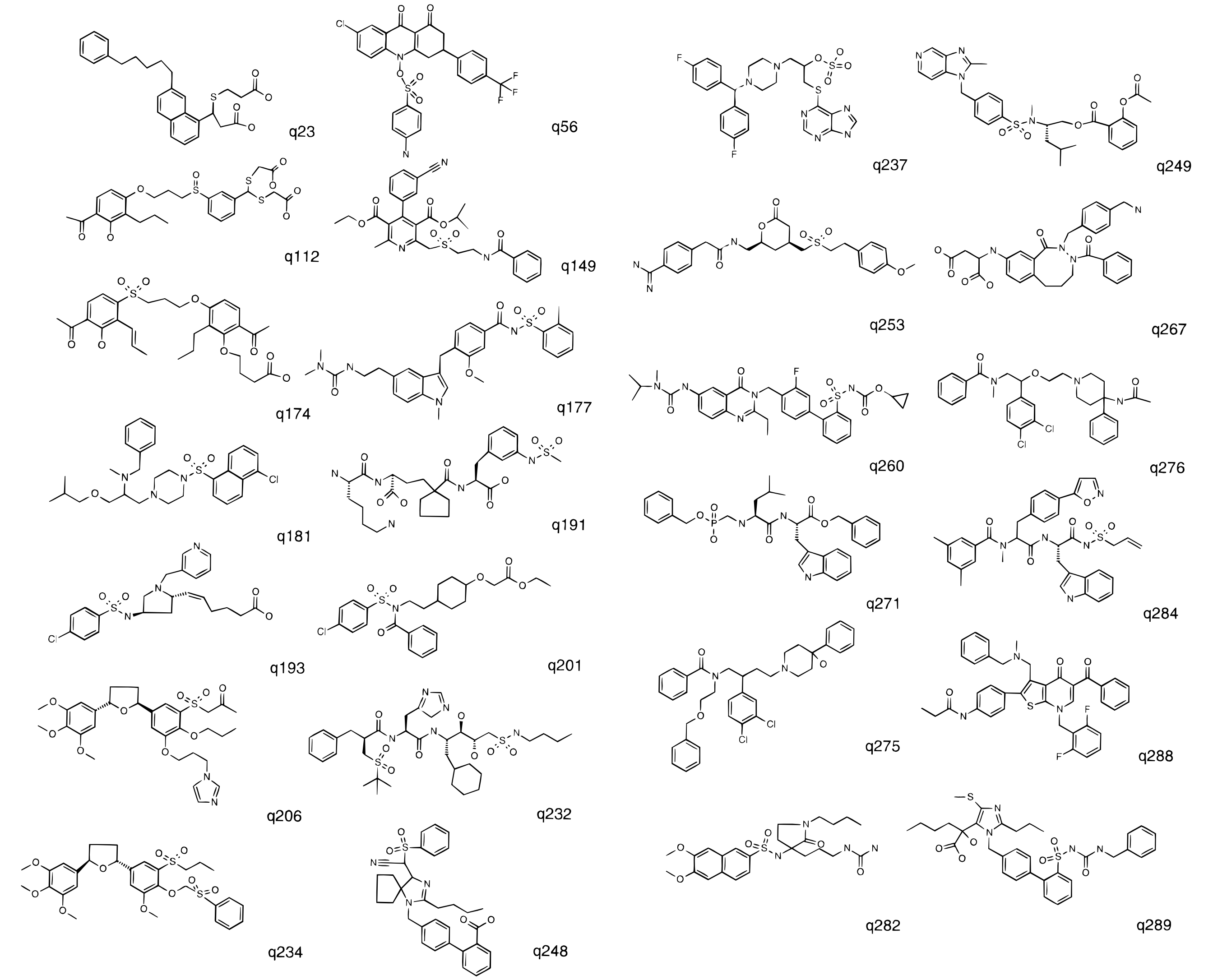


Figure 7. Twenty-six accompanying Q structures classified into the same clusters with some highly active T structures (see also Tables 4 and 6).

clustering method since it can be seen that the separation between T and Q structures is good after such a structural alignment. There are only 4 T and 1 Q structures found that are likely to disturb the separation and they are T84a, T32a, T57a, T53a, and Q56c, respectively (Table 6).

CONCLUSION

It is feasible to treat a set of commonly exposed groups identified as a set of correspondence for aligning structures and then derive a reasonable 3D QSAR (quantitative structure−activity relationship) using the method of CoMFA.[32] Since only commonly exposed groups are considered in comparing structures, the convex hull alignment rule is capable of aligning structures of greater structural dissimilarity.[36] Given its sensitivity to the fine structural changes, the commonly exposed groups identified can be used as a molecular descriptor to classify vast conformations of highly flexible molecules. The commonly exposed groups identified for a cluster of molecules can be envisaged as the most exterior parts of these molecules that are likely to similar to each other. Such a similarity also includes some basic properties of these atoms since they are selected as having the same atomic charge state. Except the geometrical ones,[10,37,38] some topological representations[34,35,39] have also been used to express the details of a molecule. It has been shown that the degree of similarity between two molecules can be studied by a topological representation such as the molecular path length.[34,35] Our 3D descriptors are a mixed type of geometrical and topological ones since they connect three convex hull vertex or nonvertex atoms with the molecular path length. Since our first descriptor is computed as the maximum molecular path length among any three distinct commonly exposed groups, it can select a local part of a molecule for comparison. Being computed as the maximum molecular path length among any three atoms of nonconvex hull vertexes, our second 3D descriptor in fact covers a more global part of the same molecule. Therefore, it is possible to classify a set of molecules of greater structural similarity through alternate use of these two 3D descriptors as we have presented here. We have found that there is no difference in the final classification result for all the structures of series "a" by reversing the usage of these two 3D descriptors.

A salient feature of our descriptors is that no encoding method or preliminary analysis for these descriptors is required. Classified structures can be further classified

Table 6. Separation of the Highly Active T Structures from the Accompanying Q Ones through Structural Alignment Using the Commonly Exposed Groups Determined for All the Structures in the Two Clusters Marked and Shown in Table 4

structures	pK_i	commonly exposed groups	structures	pK_i	commonly exposed groups
T11a	7.69	6, 12, 22(18)[a]	T6a	6.77	6, 12, 20(20)
T86a	7.63	6, 12, 31(20)	T81a	6.50	6, 12, 23(18)
T84a	7.56	10, 12, 23(22)	T32a	6.41	10, 12, 22(22)
T9a	7.44	6, 12, 22(18)	T57a	6.34	10, 19, 45(30)
T14a	7.13	6, 12, 22(18)	T5a	6.11	12, 17, 37(20)
T58c	7.09	6, 12, 22(18)	T51a	6.00	10, 19, 23(21)
T13a	6.85	6, 12, 27(18)	T22a	5.95	6, 12, 23(18)
T34a	6.79	6, 12, 22(18)	T53a	5.92	10, 12, 23(22)
T83c	6.79	6, 12, 23(18)	T16a	5.74	6, 12, 22(18)
T1a	6.77	6, 12, 20(20)			
Q23c		15, 19, 26(16)	Q56c		16, 21, 27(20)
Q112a		9, 21, 31(38)	Q149c		2, 33, 40(32)
Q149c		2, 33, 40(32)	Q174a		26, 27, 32(30)
Q177a		29, 32, 40(40)	Q181c		30, 32, 37(14)
Q191c		21, 34, 42(32)	Q193c		15, 20, 27(24)
Q201c		9, 18, 34(32)	Q206a		10, 31, 43(40)
Q232a		36, 37, 49(25)	Q234c		10, 26, 35(36)
Q237c		12, 21, 30(26)	Q248c		9, 35, 38(28)
Q249a		21, 32, 41(24)	Q253a		8, 26, 34(38)
Q260a		10, 19, 33(24)	Q271a		31, 32, 40(26)
Q275a		15, 17, 39(38)	Q276c		31, 30, 39(32)
Q282a		1, 2, 30(30)	Q284a		12, 19, 31(30)
Q288a		21, 40, 43(27)	Q289a		7, 30, 38(34)

[a] The molecular path lengths computed among the three commonly exposed groups selected are parenthesized.

through structural alignment using a new set of commonly exposed groups determined for them. Since there are structural alignments involved, our classification result would be more meaningful if the experimentally determined structures of some highly active molecules are available. Since the content of convex hull vertexes computed for flexible molecules can change more easily than that of less flexible ones, one can use our two 3D descriptors to separate molecules of small flexibility from those of highly flexible ones. A parameter that can affect the magnitude of our 3D descriptors computed is the number of atomic charge state selected in the identification of the commonly exposed groups. The number of atomic charge state cannot be set too large for a series of structures of great diversity. This number can be set as 1 for a huge database. In such a case, all the convex hull vertexes identified are treated as the commonly exposed groups. The classification would base on the comparison of two maximum molecular path lengths computed from the three most and then the three next geometrically extreme atoms of all the structures.

ACKNOWLEDGMENT

This work is supported in part by a grant from the National Science Council, ROC (NSC89-2314-B-007-002).

REFERENCES AND NOTES

(1) Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J., III. The collinearity problem in linear regression. The partial least squares approach to generalized inverses. *SIAM J. Sci. Stat. Comput.* **1984**, *5*, 735−743.

(2) Dammkoehler, R. A.; Karasek, S. F.; Shands, E. F. B.; Marshall, G. R. Constrained search of conformational hyperspace. *J. Comput.-Aided Mol. Des.* **1989**, *3*, 3−21.

(3) Vinter, J. G.; Davis, A.; Saunder, M. R. Strategic approaches to drug design. I. An integrated software framework for molecular modeling. *J. Comput.-Aided Mol. Des.* **1987**, *1*, 31−35.

(4) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic physicochemical parameters for three-dimensional structure directed quantitative structure−activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.

(5) Jurs, P. C.; Isenhour, T. L. *Chemical Applications of Pattern Recognition*; John Wiley: New York, 1975.

(6) Hall, L. H.; Kier, L. B. The molecular connectivity chi indexes and kappa shape indexes in structure−property modeling. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; John Wiley: New York, 1991; Vol. 2, p 367.

(7) Brown, R.; Martin, Y. Use of structure−activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(8) Matter, H. Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.

(9) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(10) Bath, P. A.; Poirrette, A. R.; Willett, P.; Allen, F. H. Similarity searching in fields of three-dimensional chemical structures: comparison of fragment-based measures of shape similarity. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 141−147.

(11) Ajay; Walters, W. P.; Murcko, M. A. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules. *J. Med. Chem.* **1998**, *41*, 3314−3324.

(12) Tabachnick, B. G.; Fidell, L. S. *Using Multivariate Statistics*; Harper and Row: Philadelphia, 1983.

(13) Dunn, W. J., III; Wold, S. A structure-carcinogenicity study of 4-nitroquinoline 1-oxides using the SIMCA method of pattern recognition. *J. Med. Chem.* **1978**, *21*, 1001−1007.

(14) Hodes, L.; Hazard, G. F.; Geran, R. I.; Richman, S. A statistical heuristic method for automated selection of drugs for screening. *J. Med. Chem.* **1977**, *20*, 469−475.

(15) Shemetulskis, N. E.; Dunbar, J. B., Jr.; Dunbar, B. W.; Moreland, D. Chemical database clustering and analysis. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 407−416.

(16) Menard, P. R.; Lewis, R. A.; Mason, J. R. Rational screening set design and compound selection: Cascaded clustering. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 497−505.

(17) Topliss, J. G.; Edwards, R. P. Chance factors in studies of quantitative structure−activity relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.

(18) Bemis, G. W.; Murcko, M. A. The properties of known drugs 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

(19) Kier, L. B.; Hall, L. H. Molecular connectivity in chemistry and drug research. *Med. Chem.* **1976**, 14.

(20) Randic, M.; Brissey, G. M.; Spencer, R. B.; Wilkins, C. L. Search for all self-avoiding paths for molecular graphs. *Comput. Chem.* **1979**, *3*, 5−13.

(21) Moreau, G.; Broto, P. The autocorrelation of a topological structure: a new molecular descriptor. *Nov. J. Chim.* **1980**, *4*, 359−360.

(22) Hopfinger, A. J. Theory and application of molecular potential energy fields in molecular shape analysis: a quantitative structure−activity relationship of 2,4-diamino-5-benzylpyrimidines as dihydrofolate reductase inhibitors. *J. Med. Chem.* **1983**, *26*, 990−996.

(23) Lam, P. Y.; Ru, Y.; Jadhav, P. K.; Alrich, P. E.; DeLucca, G. V.; Eyermann, C. J.; Chang, C. H.; Emmett, G.; Holler, E. R.; Daneker, W. F.; Li, L.; Confalone, P. N.; McHugh, R. J.; Han, Q.; Li, R.; Markwalder, J. A.; Seitz, S. P.; Sharpe, T. R.; Bacheler, L. T.; Rayner, M. M.; Klabe, R. M.; Shum, L.; Winslow, D. L.; Kornhauser, D. M.; Hodge, C. N. Cyclic HIV protease inhibitors: Syntheses, conformational analysis, P2/P2′ structure−activity relationship, and molecular recognition of cyclic ureas. *J. Med. Chem.* **1996**, *39*, 3514−3525.

(24) Jadhav, P. K.; Ala, P.; Woerner, F. J.; Chang, C. H.; Garber, S. S.; Anton, E. D.; Bacheler, L. T. Cyclic urea amides: HIV-1 protease inhibitors with low nanomolar potency against both wild type and protease inhibitor resistant mutant of HIV. *J. Med. Chem.* **1997**, *40*, 181−191.

(25) Debnath, A. K. Three-dimensional quantitative structure−activity relationship study on cyclic urea derivatives as HIV-1 protease inhibitors: Application of comparative molecular field analysis. *J. Med. Chem.* **1999**, *42*, 249−259.

(26) Takeuchi, Y.; Shands, E. F. B.; Beusen, D. D.; Marshall, G. R. Derivation of a three-dimensional pharmacophore model of substance P antagonists bound to the neurokinin-1 receptor. *J. Med. Chem.* **1998**, *41*, 3609−3623.

(27) Marot, C.; Chavatte, P.; Morin-Allory, L.; Viaud, M. C.; Guillaumet, G.; Renard, P.; Lesieur, D.; Michel, A. Pharmacophoric search and

3D-QSAR comparative molecular field analysis studies on agonists of melatonin sheep receptors. *J. Med. Chem.* **1998**, *41*, 4453−4465.

(28) Bohm, M.; Sturzebecher, J.; Klebe, G. Three-dimensional quantitative structure−activity relationship analysis using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa. *J. Med. Chem.* **1999**, *42*, 458−477.

(29) The MDL/ISIS database in installed at the National Center for High Performance Computing, Taiwan, ROC. The URL is http://saturn.nchc.gov.tw:9091/cds.

(30) Lin, T. H.; Peng, W. J.; Lu, Y. J. Identification of convexity as a common structure feature for structures generated for two short peptides. *Comput. Chem.* **1998**, *22*, 309−320.

(31) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. J. The protein databank: a computer-based archival file for macromolecular structure. *J. Mol. Biol.* **1977**, *112*, 535−542.

(32) *SYBYL 6.5*; The Tripos Associates, 1699 S. Hanley Rd., St. Louis, MO.

(33) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3228.

(34) Balaban, A. T. Applications of graph theory in chemistry. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 334−343.

(35) Randic, M.; Wilkins, C. L. Graph theoretical approach to recognition of structural similarity in molecules. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 31−37.

(36) Lin, T. H.; Lin, J. J.; Lu, Y. J. A comparative molecular field analysis study on several bioactive peptides using the alignment rules derived from identification of commonly exposed groups. *Biochim. Biophys. Acta* **1999**, *1429*, 476−485.

(37) Pepperrell, C. A.; Willett, P. Techniques for the calculation of three-dimensional structural similarity using interatomic distances. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 455−474.

(38) Fisanick, W.; Cross, K. P.; Rusinko, A. Similarity searching on CAS registry substance. 1. Global molecular property and generic atom triangle geometric searching. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 664−674.

(39) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical similarity using physicochemical property descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.

CI000328B