Development of Three-Dimensional Descriptors Represented by Tensors: Free Energy of Hydration Density Tensor

Su Hwan Son,[†] Cheol Kyu Han,[†,⊥] Soon Kil Ahn,[⊥] Jeong Hyeok Yoon,[§] and Kyoung Tai No*[,†,§,‡]

Department of Chemistry and CAMD Research Center, Soong Sil University, Seoul 156-743, Korea, and
Chong Kun Dang Research Institute, 410 Shindorim-dong, Seoul 152-070, Korea

Received October 28, 1998

In order to describe the degree of interaction of a molecule with its environments by descriptors, several three-dimensional descriptors have been proposed. With the physical properties calculated around a molecule, scalar, vector, and tensor (zeroth, first, and second moments) of the physical properties were calculated and were used as descriptors for calculating the similarity index between the molecules. The tensors contain the information on the spatial distribution of those physical properties around the molecule. Hydration Free Energy Density (HFED) proposed by No et al. was used to calculate HFED tensor. The descriptors were used for the similarity index calculations between substituted benzenes and between lead compounds of HIV-1 protease inhibitors. The substituted benzenes are grouped according to the similarity indices. The grouping seems reasonable from the viewpoint of a chemical sense. The lead fragments of the HIV-1 protease inhibitors have a high similarity among themselves though their chemical formulas are not very similar, the lead fragments are diverse. Although the chemical formulas are diverse, the spatial distribution of the physical properties around the molecules is similar. The descriptors have high discriminating power in the similarity calculation between the molecules.

I. INTRODUCTION

Similarity is actually a very old concept that has been with us from the earliest times, the first Greek contribution is a variety of similarity called analogy. Over a period of time the concept gradually changed in meaning coming to refer ultimately to direct comparison between two sets of relationships characterizing two different systems.[1]

In the study of various phenomena in nature, the concept of similarity plays a fundamental role. Ultimately, the recognition and analysis of molecular similarity serve as the basis of understanding of molecular structure and properties and represent the first steps in the development of theoretical models explaining chemical behavior. In this role, molecular similarity is the foundation of predictive models in chemistry.[2]

Molecular similarities and molecular properties are strongly related. Comparing the properties of molecules is an important task for detecting molecular similarities and differences. In the past, similarity has been studied qualitatively in almost every discipline of knowledge. Recently, chemists have devoted considerable attention to devising several useful quantitative measures of similarity.[1−3]

The particular shape characteristics of molecules provide important clues concerning their interactions and reactivity; shape similarity can be used for interpreting many chemical and biochemical processes. The numerical representation of shape similarity and complementarily play an increasingly important role in many fields of chemistry and biochemistry, including quantum pharmacology and computer aided drug design (CADD).[4−6]

Physical, chemical, or biological properties of a compound depend on the three-dimensional (3-D) arrangements of the atoms in a molecule.[6] The analysis of such structure−property relationships should therefore take into account the 3-D structure of the molecule. This has become standard usage in qualitative molecular modeling studies of substrate and receptor interaction. Chemical species can be represented in manifold ways and can yield numerical parameters, through the descriptors.[1] Such parameters can be obtained from (i) physicochemical or thermodynamic properties, such as the octanol−water partition coefficient or the heat of formation, (ii) topological indices,[7] such as the Wiener indices[8,9] or structural indices, (iii) quantum-mechanical parameters, such as momentum space electron density[2] or shape descriptor,[4−6,10] and (iv) complexity indices, such as the Bertz index.[11]

Shape is now realized to be one of the most fundamental conceptions of the chemistry lock and key idea that a drug interacts with its receptor just as a key fits into and opens a lock. Most approaches rely upon molecular descriptors, which are numerical values representing selected features of the compounds. Each structure is represented as a scalar or vector of such numerical descriptors and thus represented by a point in a high-dimensional space with coordinates equal to (or related to) the corresponding descriptor values. Many different types of descriptors have been presented in the literature. In rational ligand design, consideration should be given to a combination of steric, electrostatic, and hydrophobic factors. Each of these plays its part in deciding the optimum arrangement of ligands in a binding site.[12−14]

[†] Department of Chemistry, Soong Sil University.
[‡] Member of the Center for Molecular Science, Korea.
[§] CAMD Research Center, Soong Sil University.
[⊥] Chong Kun Dang Research Institute.

10.1021/ci980224p CCC: $18.00
Published on Web 04/28/1999

Steric factors are readily assessed by a number of methods, for example, the volume and surface area of a set of related molecules. Molecular shape as represented by common steric volume and surface area, is the major factor contributing to affinity of inhibitors.[15,16] The importance of molecular electrostatic potential (MEP) in long-range ligand−receptor interactions has long been recognized. MEP mapping has been widely used to explain electrostatic interaction of a variety of molecules with their environments.[2,14,17] The following two types of intermolecular interactions are usually dominant in ligand−receptor binding: (i) electrostatic, e.g., hydrogen bonding and (ii) hydrophobic, introduced by the aqueous chemical environment.

Since most physical, chemical, and biological phenomena take place in solvents, the calculation of the free energy of solvation is important in understanding the phenomena occurring in solution. In our previous work, a practical dielectric continuum model to calculate the Solvation Energy of Density (SFED) was proposed.[18] SFED is a useful tool to investigate the solvent binding affinity distribution around a solute. The empirical Hydration Free Energy Density (HFED) is expressed by a linear combination of the physical properties calculated with net atomic charges, polarizabilities, dispersion coefficients of the atoms in a molecule, and solvent accessible surface (SAS).

In this work, we develop some molecular descriptors with effective atomic charges, molecular electrostatic potential (MEP), polarizabilities, and dispersion coefficients of the atoms in molecules and with HFED. In order to keep the information of the spatial distributions of the above physical quantitative in molecules, we introduced second momentum as well as the scalar and first momentum of the physical properties. To test the reliability of the descriptors in the similarity calculation, we calculated similarity indices between some aromatic molecules and between some HIV- 1 protease inhibitors.

II. CALCULATIONS

The descriptors developed in this work are general descriptors which do not contain topological information. It is believed that the degree of similarity between two molecules can be described by the similarity in (i) the electron distribution of the molecules and (ii) the fields produced by the molecules. In this work, as an approximation, the electron distribution in a molecule was described with point charges. A molecule was represented by the descriptors which was calculated with the volume, surface area, distribution of hydrophilic and hydrophobic sites, electric fields, degree of polarization, and degree of dispersion interaction of the molecule.

Since the physical properties of a molecule, which are the measure of the similarity, are the results of the interaction of the molecule with its environments, the fields produced by the molecule were used for the calculation of the descriptors. For the calculation of the fields around a molecule (i.e., electrostatic field, HFED, polarization and dispersion interaction fields), the same grid model, which was proposed to calculate the HFED, was introduced.

A. Computation of the Physical Properties of Solutes. We adopt a model of previous effort to calculate the HFED around a molecule.[18] In the model, each molecule has three domains as described in Figure 1. The inner domain is constructed by the overlap of the van der Waals spheres of the atoms in the molecules. Outside of the inner domain, a shell with thickness r_s, an effective solvent radius, is added. The center of the solvent cannot penetrate this shell, the inner shell. The surface of this shell corresponds to the solvent accessible surface. Outside of this shell, a shell with thickness r_c, a critical shell thickness, is added, the outer shell.

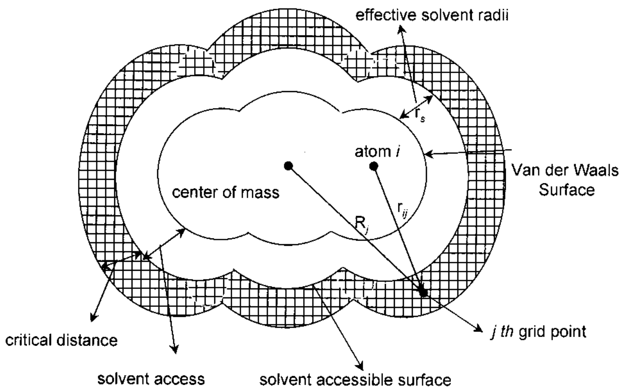

Figure 1. The geometrical parameters introduced in the model are described.

In order to calculate physical properties around the molecule, grid points are generated inside of the outer shell. Since the computing time is linearly depends on the number of grid points, the grid point density was determined as eight points/Å^3 by compromising between the computing time and the accuracy of the results. The computational details are described in our previous work.[18]

Volume (V_s) and surface area (A_s) of a molecule are important descriptors for describing the molecular shape. Since the shape of the molecule can be defined by its environments, *i.e.*, solvent, it seems physically realistic to calculate the surface area and molecular volume from the number of the grid points, N_p, around the molecule. The surface area is proportional to the number of the grid points, N_p.

$$A_s \propto N_p \tag{1}$$

The molecular volume can be approximated as follows

$$V_s \propto \sum_{j}^{N_p} R_j \tag{2}$$

where R_j is the distance between the center of mass of the molecule and the grid point i.

The molecular electrostatic potential (MEP) at point j, V_j, is evaluated using simple point charges, atom centered effective point charge, Q_i.

$$V_j = \sum_{i=1}^{n_A} \left(\frac{Q_i}{r_{ij}} \right) \tag{3}$$

where n_A is the number of atoms in the molecule and r_{ij} is the distance between atom i and grid point j. The net atomic charge Q_i was calculated with the Modified Partial Equalization of Orbital Electronegativity (MPEOE) method.[19,20]

For the calculation of the degree of polarization of the solute by its environment, the polarizability of the molecule was expressed as a sum of the atomic polarizabilities of the atoms in the molecule. The atomic polarizability is centered on each atom center. This effective atomic polarizability was calculated with an empirical atomic polarizability calculation method proposed by No et al.[21] In the method, Charge Dependent Effective Atomic Polarizability (CDEAP), the effective atomic polarizability of atom i, α_i, was expressed as a linear function of the net atomic charge, Q_i,

$$\alpha_i = \alpha_{i,0} - a_i Q_i \tag{4}$$

where $\alpha_{i,0}$ and a_i are the effective atomic polarizability of neutral atom and the charge coefficient, respectively. Details of the CDEAP method are well described elsewhere in our previous paper.[21]

Dispersion interaction of a solute with its environments is important for describing the physical property of the solute. The dispersion coefficients of atom i were calculated with the following Slater−Kirkwood formula.[22]

$$D_i = \frac{3}{4}\left[\frac{eh}{me^{1/2}}\right]\frac{{\alpha_i}^2}{(\alpha_i/N_i)^{1/2}} \tag{5}$$

where N_i is the number of effective electrons of the atom i. α_i is the atomic polarizabilities of the atom i. The other symbols have their usual meaning.

In our previous work,[18] the free energy of hydration was expressed as a function of the physical properties of solutes. The hydration free energy density (HFED) was calculated at each grid point, g_j in Figure 1, and the hydration free energy was obtained by summing over the HFED within critical distance, r_c.

$$\Delta G_{\text{hyd}} = \sum_{j=1}^{N_p} g_j \tag{6}$$

HFED at point j, g_j, was described as follows

$$g_j = \left[C_1\left|\sum_{i=1}^{n_A}\frac{Q_i}{r_{ij}}\right| + \left\{C_2\sum_{i=1}^{n_A}\frac{{Q_i}^2}{r_{ij}} + C_3\sum_{i=1}^{n_A}\frac{\alpha_i}{{r_{ij}}^3} + C_4\sum_{i=1}^{n_A}\frac{D_i}{{r_{ij}}^6}\right\}\right] + \frac{1}{N_p}\left(C_5\sum_{j}^{N_p}R_j + C_6\right) \tag{7}$$

where C's are the linear coefficients which were determined with experimental hydration free energy data through an optimization procedure.

B. Calculation of Descriptors from the Physical Properties. Several descriptors were derived with g_j and the components of g_j. Using abbreviations, eq 7 is simplified as follows:

$$g_j = C_1|V_j| + C_2|QV_j| + C_3A_j + C_4\Delta_j + \overline{C_5}V + \overline{C_6} \tag{8}$$

At each grid point j, one can obtain physical properties, g_j, $|V_j|$, $|QV_j|$, A_j, and Δ_j. Each quantity can be a good representation of the property of the solute. Where $C_1 = -0.3475 \times 10^{-1}$, $C_2 = -0.76\overline{86} \times 10^{-2}$, $C_3 = -0.2252 \overline{\times} 10^{-4}$, $C_4 = -0.5299 \times 10^{-3}$, $\overline{C_5} = 0.1869 \times 10^{-4}$, and $\overline{C_6} = 1.23$.

Table 1. The Atomic Species Introduced in This Work and the Atomic Radius That Were Optimized for HFED[l8] Model

atomic species	definition	atomic radius (Å)
C1	SP3 carbon	2.074
C2	SP2 carbon in conjugate system	1.994
C3	SP2 carbon in carbonyl, alkene	1.725
C4	amide carbon	1.725
H1	aliphatic H	1.474
H2	H bonded to aromatic system	1.579
H3	hydroxyl H in alcohol	1.175
H4	hydroxy H in carboxylic acid	1.220
H5	H bonded to amide N	1.165
H6	H bonded to amine N	1.130
O1	SP2 oxygen in carboxylic acid	1.764
O2	SP3 hydroxyl oxygen	1.489
N1	SP2 trivalent conjugated nitrogen	1.589
N2	SP2 divalent conjugated nitrogen	1.349
N3	SP3 nitrogen in amine	1.689
N4	SP2 nitrogen in amide	1.435
S1	SP2 conjugated sulfer	1.994
F1	fluoride	1.672
C11	chloride	2.95
Brl	bromide	2.420
I1	iodide	2.549

In order to describe the three-dimensional distribution of the physical quantities g, V, QV, A, and Δ, not only scalar and vector but also tensor of the physical quantities were introduced as descriptors. The elements of the tensor in Cartesian coordinates are obtained as

$$X_{xx} = 0.5\sum_{j=1}^{N_p} X_j(2x_j^2 - y_j^2 - z_j^2) \tag{9}$$

$$X_{xy} = 1.5\sum_{j=1}^{N_p} X_j x_j y_j \tag{10}$$

where X_j, is one of the physical quantities g_j, V_j, QV_j, A_j, and Δ_j. X_{xx} and X_{xy} are the xx and xy components of tensor X, $\underline{X}'$, and (x_j, y_j, z_j) is the position of the j point from the center of mass origin in arbitrary Cartesian coordinates. The tensors, $\underline{X}'$s, are diagonalized through a unitary transformation.

$$\underline{X} = \underline{R}^+\,\underline{X}'\,\underline{R} \tag{11}$$

The electric moments of the solutes were also used as descriptors. Both dipole and quadrupole moments were calculated, and the quadrupole moments are diagonalized. The polarizability quadrupole moments were also calculated with the atomic polarizabilities. The procedure is the same with the case of electric quadrupole moment calculation. The point charges are replaced with the CDEAPs of the atoms in a molecule.

Surface area, volume, and (surface area/volume) were also used as descriptors. In Table 2, the descriptors introduced in this work are summarized.

C. Build the Similarity Indices (Measure Similarity). In order to test the discriminating power of the description, the descriptors were introduced for similarity calculation among a group of molecules. As a criterion of the similarity between two molecules, Good's linear form similarity index[23]

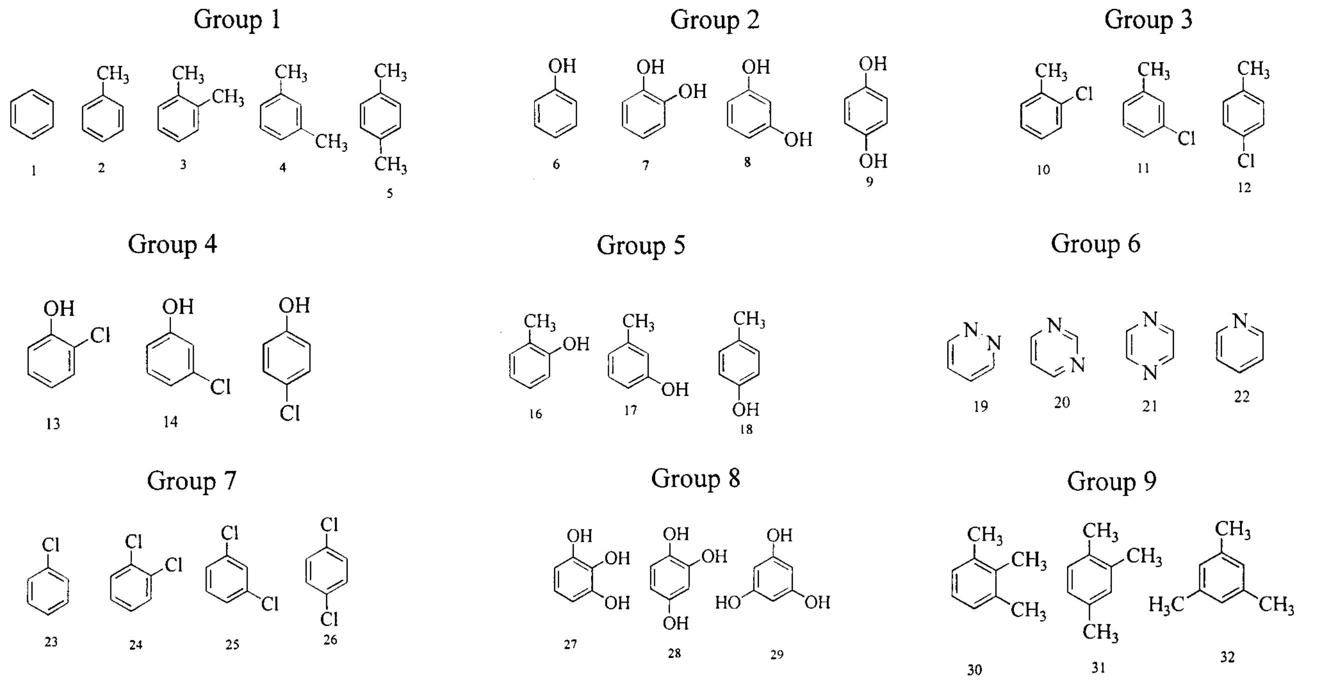

Figure 2. Aromatic compounds introduced in the similarity index calculation. The minimum energy conformation of each compound was obtained through energy minimization with CVFF.

Table 2. The Descriptors Used in This Work Are Summarized and Described

symbol	description
$\underline{\epsilon R}$	$X_j = \Sigma_i Q_i/r_{ij}$, calculate second momentum using eqs 9 and 10
$\underline{g}$	$X_j = g_j$, calculate second momentum using eqs 9 and 10
$\underline{\|V\|}$	$X_j = \|\Sigma_i Q_i/r_{ij}\|$, calculate second momentum using eqs 9 and 10
$\underline{\|QV\|}$	$X_j = \|\Sigma_i Q_i^2/r_{ij}\|$, calculate second momentum using eqs 9 and 10
$\underline{A}$	$X_j = \Sigma_i \alpha_i/r_{ij}^3$, calculate second momentum using eqs 9 and 10
$\underline{\|\Delta\|}$	$X_j = \Sigma_i D_i/r_{ij}^6$, calculate second momentum using eqs 9 and 10
ΣV_j,	$\Sigma_j \Sigma_i Q_i/r_{ij}$
ΣQV_j	$\Sigma_j \Sigma_i Q_i^2/r_{ij}$
$\Sigma \Delta_j$	$\Sigma_j \Sigma_i \alpha_i/r_{ij}^3$
ΣA_j	$\Sigma_j \Sigma_i D_i/r_{ij}^6$
S	surface area
V	molecular volume
S/V	molecular surface area/molecular volume
ΔG_{hyd}	free energy of hydration
$\vec{\mu}$	total dipole moment
$\underline{Q_{\mathrm{elec}}}$	$X_j = X_i, = Q_i$, calculate second momentum using eqs 9 and 10
$\underline{\alpha_{\mathrm{mol}}}$	$X_j = X_i = \alpha_i$, calculate second momentum using eqs 9 and 10

was used. Similarity relations are measured as a difference of descriptors between molecule *A* to *B*. The similarity of the *k*th physical property L^k (*A*; *B*) between molecules A and B was described as

$$L^k (A; B) = 1 - \left(\frac{|X_A^k - X_B^k|}{|X_A^k|} \right) \quad \text{if } (X_A^k > X_B^k) \tag{12}$$

where X_A^k and X_B^k are the *k*th descriptor of molecules *A* and *B*. Since L_{AB}^k is normalized, it takes the value in the range [0,1] for each molecular pair. The total similarity, the linear form of Good's similarity index, between molecules *A* and *B* was obtained as

$$L(A; B) = \frac{1}{w_k N_{\mathrm{d}}} \sum_{k=1}^{N_{\mathrm{d}}} w_k L^k (A; B) \tag{13}$$

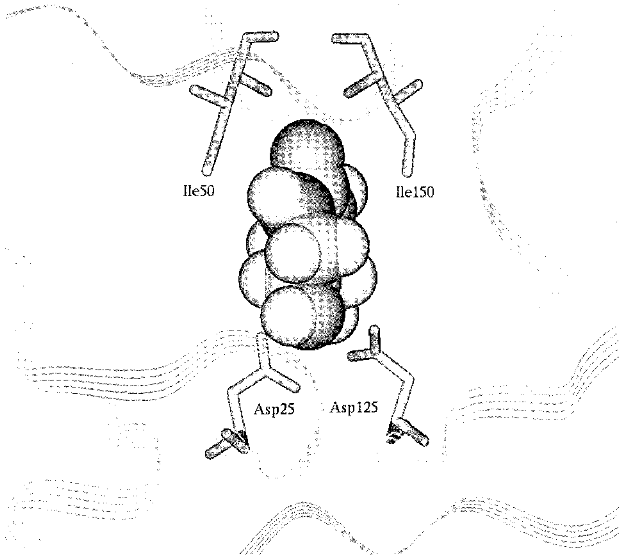

Figure 3. Ribbon back bone trace of the HIV-1 protease taken from the x-ray crystal complex with compound **5** in Figure 4. The catalytic sites Asp25/125, Ile50/150 are shown.

where w_k and N_{d} are the weighting factors for the *k*th descriptor and the number of the descriptors.

The descriptors were used for the classification of two groups of molecules. The molecules in each group were classified. The groups are benzenes with several kinds of functional groups and the lead fragments of the peptidomimetic HIV-1 protease inhibitors. The molecules are grouped according to the similarities among the molecules.

D. Application to Aromatic Compounds. Thirty-two aromatic compounds which have various functional groups, in Figure 2, were used for the similarity index calculations. The geometries of the molecules were obtained through energy minimization with CVFF.[24,25] All the geometry optimizations were carried out with DISCOVER/INSIGHT.[26]

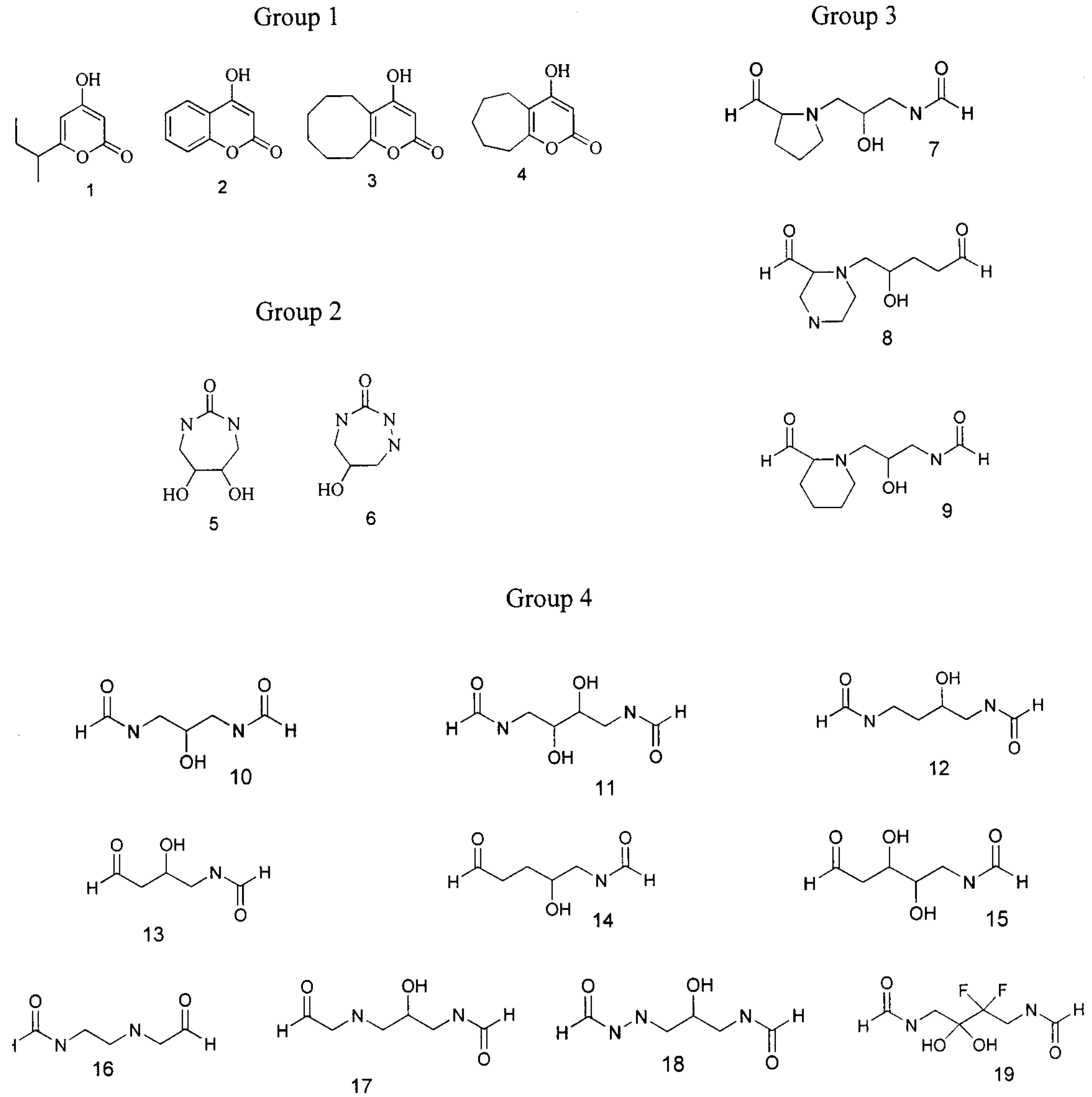

Figure 4. The lead fragments of HIV-1 protease inhibitors which were taken from Protein Data Bank (PDB) are shown. The geometry of the lead compounds are not energy minimized and the same geometry with PDB data was used.

E. Application to HIV-1 Protease Inhibitor. Overall, peptidomimetic compounds bind to HIV-1 protease in an extended conformation with their amide group involved in conserved hydrogen bonds with the active site residues and water (Figure 3). All the peptidomimetic inhibitors have a catalytic water bind to Ile50/Ile150 residues in HIV-1 protease, in the case of nonpeptidic inhibitors (pyrones (**1**−**4**)) and cyclic ureas (**5** and **6**)), a carbonyl group takes the place of the water. Therefore one water molecule was put to each peptidomimetic inhibitor, compounds **7**−**19**. The position of the water was taken from X-ray crystallographic data. HIV-1 protease has C_2 symmetry, thus many inhibitors have C_2 symmetry. Asp25, Asp125, Ile50, Ile150 residues are arranged in C_2 symmetry at the binding pocket (Figure 3).[27−29]

In this work, we are interested in the similarity between the fragment of lead compounds which bind to the center of a binding pocket. The structures of the HIV-1 protease inhibitors were collected from Protein Data Bank (PDB). From the PDB inhibitors structures, the side chains which bound to *P1*, *P1′*, *P2*, *P2′*, *P3*, and *P3′* pockets were removed, and the broken bonds were terminated by hydrogens. The geometries of the fragments were not optimized. The lead fragments are summarized in Figure 4.

III. RESULTS AND DISCUSSION

A. Aromatic Molecules. In Table 3, the similarity indices for the aromatic molecules are listed in a matrix. Since all the model compounds have benzene or benzene-like rings and the size of the compounds are not much different, the volume and the surface area descriptors contribute evenly to all the similarity indices. For this reason, all the elements obtained are larger than 0.36.

According to the similarity indices, benzene is discriminated from the methylated benzenes, groups 1 and 9 compounds. All the methylated benzenes have high similarities to each other. The physical properties of *o*-dimethylbenzene may be closer to toluene than those of *p*-dimethylbenzene because *S*(toluene;*o*-dimethylbenzene) = 0.892 and *S*(*o*-dimethylbenzene;*p*-dimethylbenzene) = 0.758.

Hydroxylbenzenes, groups 2 and 8 compounds, are highly similar to each other. As in the case of the methylated

Table 3. The Similarity Index Matrix for the 32 Aromatic Molecules

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1.000	0.649	0.615	0.576	0.553	0.527	0.452	0.412	0.441	0.563	0.517	0.514	0.513	0.443	0.422	0.492
2	0.649	1.000	0.892	0.860	0.758	0.684	0.590	0.530	0.565	0.743	0.663	0.629	0.686	0.590	0.552	0.720
3	0.615	0.892	1.000	0.887	0.789	0.682	0.595	0.532	0.563	0.776	0.686	0.646	0.685	0.609	0.570	0.721
4	0.576	0.860	0.887	1.000	0.871	0.693	0.607	0.532	0.554	0.759	0.743	0.673	0.710	0.639	0.604	0.751
5	0.553	0.758	0.789	0.871	1.000	0.716	0.640	0.567	0.567	0.746	0.776	0.705	0.742	0.666	0.611	0.756
6	0.527	0.684	0.682	0.693	0.716	1.000	0.812	0.709	0.687	0.683	0.647	0.590	0.789	0.726	0.690	0.864
7	0.452	0.590	0.595	0.607	0.640	0.812	1.000	0.858	0.751	0.622	0.594	0.549	0.706	0.769	0.700	0.784
8	0.412	0.530	0.532	0.532	0.567	0.709	0.858	1.000	0.877	0.589	0.572	0.548	0.621	0.723	0.721	0.698
9	0.441	0.565	0.563	0.554	0.567	0.687	0.751	0.877	1.000	0.631	0.609	0.604	0.658	0.681	0.701	0.670
10	0.563	0.743	0.776	0.759	0.746	0.683	0.622	0.589	0.631	1.000	0.873	0.777	0.784	0.712	0.688	0.675
11	0.517	0.663	0.686	0.743	0.776	0.647	0.594	0.572	0.609	0.873	1.000	0.874	0.763	0.710	0.744	0.654
12	0.514	0.629	0.646	0.673	0.705	0.590	0.549	0.548	0.604	0.777	0.874	1.000	0.683	0.665	0.734	0.610
13	0.513	0.686	0.685	0.710	0.742	0.789	0.706	0.621	0.658	0.784	0.763	0.683	1.000	0.816	0.747	0.803
14	0.443	0.590	0.609	0.639	0.666	0.726	0.769	0.723	0.681	0.712	0.710	0.665	0.816	1.000	0.849	0.783
15	0.422	0.552	0.570	0.604	0.611	0.690	0.700	0.721	0.701	0.688	0.744	0.734	0.747	0.849	1.000	0.723
16	0.492	0.720	0.721	0.751	0.756	0.864	0.784	0.698	0.670	0.675	0.654	0.610	0.803	0.783	0.723	1.000
17	0.467	0.676	0.688	0.714	0.732	0.861	0.823	0.729	0.674	0.672	0.655	0.604	0.812	0.844	0.768	0.905
18	0.450	0.635	0.659	0.703	0.733	0.829	0.805	0.729	0.696	0.712	0.698	0.634	0.811	0.856	0.804	0.834
19	0.555	0.572	0.543	0.533	0.529	0.639	0.565	0.586	0.605	0.610	0.610	0.563	0.532	0.502	0.570	0.575
20	0.550	0.536	0.506	0.500	0.508	0.690	0.658	0.623	0.626	0.571	0.542	0.501	0.611	0.584	0.576	0.618
21	0.530	0.570	0.544	0.534	0.544	0.690	0.763	0.690	0.599	0.518	0.466	0.427	0.589	0.666	0.570	0.645
22	0.595	0.586	0.558	0.550	0.560	0.743	0.645	0.603	0.606	0.624	0.590	0.544	0.628	0.615	0.599	0.667
23	0.579	0.757	0.735	0.726	0.692	0.662	0.608	0.588	0.642	0.901	0.839	0.791	0.755	0.700	0.683	0.661
24	0.507	0.652	0.681	0.662	0.634	0.593	0.550	0.562	0.625	0.795	0.770	0.771	0.661	0.608	0.683	0.587
25	0.487	0.635	0.667	0.724	0.773	0.662	0.629	0.601	0.608	0.798	0.861	0.815	0.709	0.737	0.742	0.688
26	0.480	0.595	0.616	0.671	0.743	0.596	0.620	0.573	0.528	0.686	0.726	0.747	0.652	0.734	0.663	0.641
27	0.373	0.519	0.532	0.531	0.545	0.666	0.776	0.834	0.783	0.576	0.579	0.568	0.606	0.693	0.748	0.693
28	0.378	0.489	0.497	0.492	0.513	0.635	0.749	0.863	0.835	0.596	0.594	0.555	0.612	0.700	0.726	0.642
29	0.368	0.485	0.493	0.485	0.492	0.623	0.752	0.840	0.784	0.539	0.539	0.530	0.560	0.646	0.691	0.631
30	0.564	0.785	0.854	0.881	0.810	0.661	0.589	0.521	0.549	0.727	0.723	0.673	0.676	0.621	0.581	0.708
31	0.530	0.748	0.769	0.851	0.921	0.699	0.643	0.563	0.560	0.709	0.746	0.697	0.717	0.662	0.600	0.777
32	0.575	0.766	0.828	0.823	0.786	0.630	0.565	0.497	0.532	0.684	0.675	0.647	0.633	0.591	0.544	0.678

	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
1	0.467	0.450	0.555	0.550	0.530	0.595	0.579	0.507	0.487	0.480	0.373	0.378	0.368	0.564	0.530	0.575
2	0.676	0.635	0.572	0.536	0.570	0.586	0.757	0.652	0.635	0.595	0.519	0.489	0.485	0.785	0.748	0.766
3	0.688	0.659	0.543	0.506	0.544	0.558	0.735	0.681	0.667	0.616	0.532	0.497	0.493	0.854	0.769	0.828
4	0.714	0.703	0.533	0.500	0.534	0.550	0.726	0.662	0.724	0.671	0.531	0.492	0.485	0.881	0.851	0.823
5	0.732	0.733	0.529	0.508	0.544	0.560	0.692	0.634	0.773	0.743	0.545	0.513	0.492	0.810	0.921	0.786
6	0.861	0.829	0.639	0.690	0.690	0.743	0.662	0.593	0.662	0.596	0.666	0.635	0.623	0.661	0.699	0.630
7	0.823	0.805	0.565	0.658	0.763	0.645	0.608	0.550	0.629	0.620	0.776	0.749	0.752	0.589	0.643	0.565
8	0.729	0.729	0.586	0.623	0.690	0.603	0.588	0.562	0.601	0.573	0.834	0.863	0.840	0.521	0.563	0.497
9	0.674	0.696	0.605	0.626	0.599	0.606	0.642	0.625	0.608	0.528	0.783	0.835	0.784	0.549	0.560	0.532
10	0.672	0.712	0.610	0.571	0.518	0.624	0.901	0.795	0.798	0.686	0.576	0.596	0.539	0.727	0.709	0.684
11	0.655	0.698	0.610	0.542	0.466	0.590	0.839	0.770	0.861	0.726	0.579	0.594	0.539	0.723	0.746	0.675
12	0.604	0.634	0.563	0.501	0.427	0.544	0.791	0.771	0.815	0.747	0.568	0.555	0.530	0.673	0.697	0.647
13	0.812	0.811	0.532	0.611	0.589	0.628	0.755	0.661	0.709	0.652	0.606	0.612	0.560	0.676	0.717	0.633
14	0.844	0.856	0.502	0.584	0.666	0.615	0.700	0.608	0.737	0.734	0.693	0.700	0.646	0.621	0.662	0.591
15	0.768	0.804	0.570	0.576	0.570	0.599	0.683	0.683	0.742	0.663	0.748	0.726	0.691	0.581	0.600	0.544
16	0.905	0.834	0.575	0.618	0.645	0.667	0.661	0.587	0.688	0.641	0.693	0.642	0.631	0.708	0.777	0.678
17	1.000	0.893	0.554	0.625	0.690	0.653	0.666	0.589	0.690	0.670	0.713	0.675	0.649	0.683	0.728	0.653
18	0.893	1.000	0.548	0.629	0.671	0.657	0.688	0.631	0.727	0.677	0.700	0.688	0.642	0.685	0.722	0.644
19	0.554	0.548	1.000	0.797	0.660	0.791	0.594	0.587	0.561	0.473	0.586	0.563	0.560	0.523	0.510	0.499
20	0.625	0.629	0.797	1.000	0.772	0.845	0.549	0.485	0.532	0.465	0.560	0.559	0.542	0.487	0.487	0.465
21	0.690	0.671	0.660	0.772	1.000	0.731	0.502	0.439	0.502	0.548	0.608	0.597	0.605	0.512	0.526	0.515
22	0.653	0.657	0.791	0.845	0.731	1.000	0.602	0.531	0.588	0.531	0.543	0.547	0.527	0.536	0.539	0.513
23	0.666	0.688	0.594	0.549	0.502	0.602	1.000	0.793	0.761	0.647	0.568	0.589	0.535	0.703	0.665	0.648
24	0.589	0.631	0.587	0.485	0.439	0.531	0.793	1.000	0.701	0.591	0.590	0.583	0.564	0.638	0.601	0.612
25	0.690	0.727	0.561	0.532	0.502	0.588	0.761	0.701	1.000	0.824	0.571	0.579	0.530	0.721	0.765	0.684
26	0.670	0.677	0.473	0.465	0.548	0.531	0.647	0.591	0.824	1.000	0.541	0.560	0.504	0.666	0.741	0.666
27	0.713	0.700	0.586	0.560	0.608	0.543	0.568	0.590	0.571	0.541	1.000	0.861	0.888	0.516	0.540	0.488
28	0.675	0.688	0.563	0.559	0.597	0.547	0.589	0.583	0.579	0.560	0.861	1.000	0.888	0.485	0.511	0.462
29	0.649	0.642	0.560	0.542	0.605	0.527	0.535	0.564	0.530	0.504	0.888	0.888	1.000	0.476	0.492	0.453
30	0.683	0.685	0.523	0.487	0.512	0.536	0.703	0.638	0.721	0.666	0.516	0.485	0.476	1.000	0.823	0.903
31	0.728	0.722	0.510	0.487	0.526	0.539	0.665	0.601	0.765	0.741	0.540	0.511	0.492	0.823	1.000	0.810
32	0.653	0.644	0.499	0.465	0.515	0.513	0.648	0.612	0.684	0.666	0.488	0.462	0.453	0.903	0.810	1.000

benzenes, *o*-dihydroxybenzene is more similar to phenol than *p*-dihydroxybenzene. Monohydroxybenzene, phenol, has a low similarity with trihydroxylbenzenes. Phenol is similar to methylated phenols group 5 compounds.

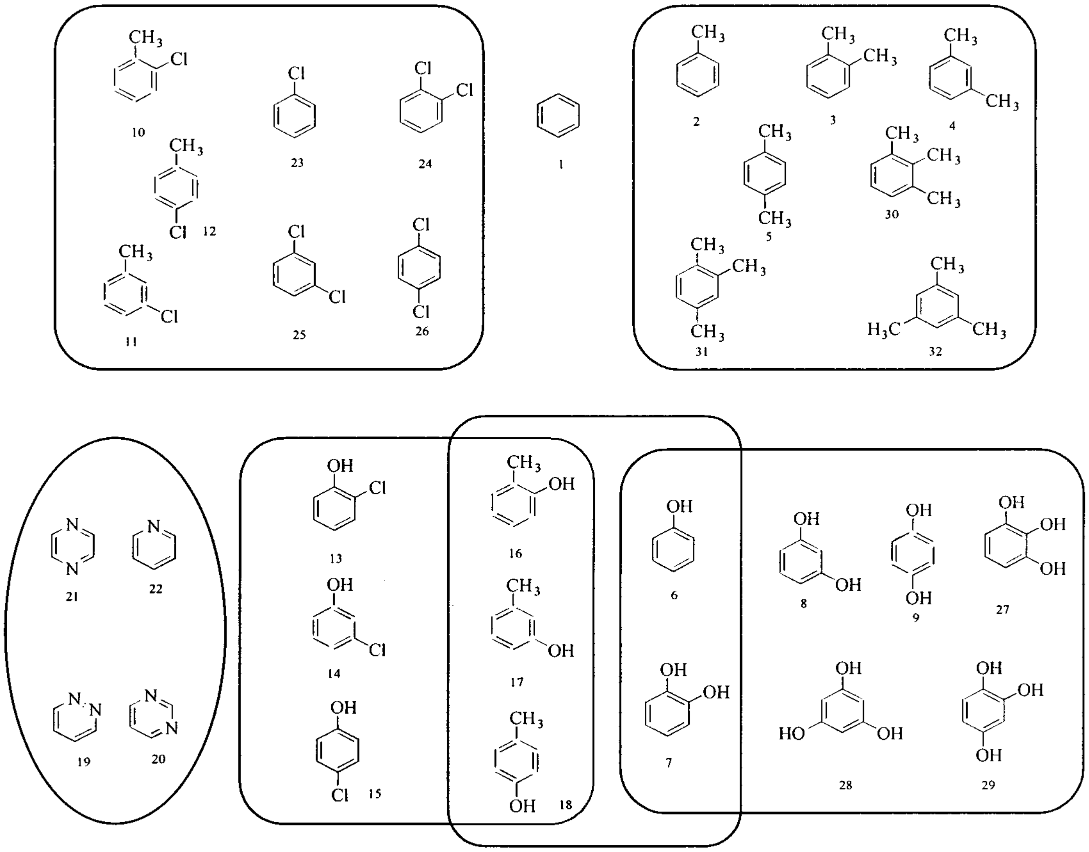

Figure 5. The grouping of the aromatic compounds according to their similarity index.

Table 4. Similarity Matrix for 19 HIV-1 Protease Inhibitors

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	1.000	0.734	0.719	0.730	0.624	0.649	0.528	0.497	0.498	0.563	0.583	0.573	0.612	0.621	0.608	0.677	0.590	0.564	0.530
2	0.734	1.000	0.846	0.846	0.624	0.645	0.592	0.591	0.578	0.601	0.589	0.588	0.603	0.639	0.584	0.683	0.663	0.580	0.603
3	0.719	0.846	1.000	0.942	0.643	0.672	0.635	0.617	0.604	0.605	0.591	0.595	0.615	0.636	0.581	0.721	0.681	0.617	0.619
4	0.730	0.846	0.942	1.000	0.679	0.674	0.625	0.597	0.583	0.599	0.599	0.594	0.619	0.643	0.591	0.727	0.673	0.641	0.613
5	0.624	0.624	0.643	0.679	1.000	0.815	0.656	0.596	0.590	0.667	0.651	0.669	0.818	0.740	0.723	0.824	0.685	0.736	0.580
6	0.649	0.645	0.672	0.674	0.815	1.000	0.642	0.624	0.619	0.702	0.695	0.703	0.833	0.797	0.766	0.881	0.746	0.690	0.590
7	0.528	0.592	0.635	0.625	0.656	0.642	1.000	0.844	0.872	0.743	0.817	0.782	0.668	0.751	0.689	0.673	0.763	0.757	0.779
8	0.497	0.591	0.617	0.597	0.596	0.624	0.844	1.000	0.893	0.740	0.790	0.783	0.665	0.754	0.686	0.657	0.761	0.674	0.744
9	0.498	0.578	0.604	0.583	0.590	0.619	0.872	0.893	1.000	0.762	0.817	0.804	0.675	0.750	0.704	0.646	0.762	0.710	0.825
10	0.563	0.601	0.605	0.599	0.667	0.702	0.743	0.740	0.762	1.000	0.874	0.911	0.789	0.879	0.877	0.741	0.868	0.825	0.756
11	0.583	0.589	0.591	0.599	0.651	0.695	0.817	0.790	0.817	0.874	1.000	0.915	0.737	0.845	0.803	0.731	0.821	0.820	0.820
12	0.573	0.588	0.595	0.594	0.669	0.703	0.782	0.783	0.804	0.911	0.915	1.000	0.783	0.862	0.846	0.743	0.845	0.802	0.783
13	0.612	0.603	0.615	0.619	0.818	0.833	0.668	0.665	0.675	0.789	0.737	0.783	1.000	0.845	0.846	0.858	0.813	0.764	0.618
14	0.621	0.639	0.636	0.643	0.740	0.797	0.751	0.754	0.750	0.879	0.845	0.862	0.845	1.000	0.891	0.825	0.869	0.805	0.707
15	0.608	0.584	0.581	0.591	0.723	0.766	0.689	0.686	0.704	0.877	0.803	0.846	0.846	0.891	1.000	0.773	0.834	0.768	0.680
16	0.677	0.683	0.721	0.727	0.824	0.881	0.673	0.657	0.646	0.741	0.731	0.743	0.858	0.825	0.773	1.000	0.770	0.745	0.613
17	0.590	0.663	0.681	0.673	0.685	0.746	0.763	0.761	0.762	0.868	0.821	0.845	0.813	0.869	0.834	0.770	1.000	0.806	0.729
18	0.564	0.580	0.617	0.641	0.736	0.690	0.757	0.674	0.710	0.825	0.820	0.802	0.764	0.805	0.768	0.745	0.806	1.000	0.741
19	0.530	0.603	0.619	0.613	0.580	0.590	0.779	0.744	0.825	0.756	0.820	0.783	0.618	0.707	0.680	0.613	0.729	0.741	1.000

Methylated chlorobenzenes (group 3) and chlorinated benzenes (group 7) have high similarity. It seems that $-CH_3$ and $-Cl$ play a similar role when they are bonded to benzene. Chlorinated toluenes and mono- or dichlorinated benzenes are highly similar. It may be said that the methyl group can be replaced with chlorine without much change in the physical properties of a molecule.

The compounds in group 6 have low similarity with benzene though their molecular shape is very similar to benzene. Compounds **19** and **21** show low similarity.

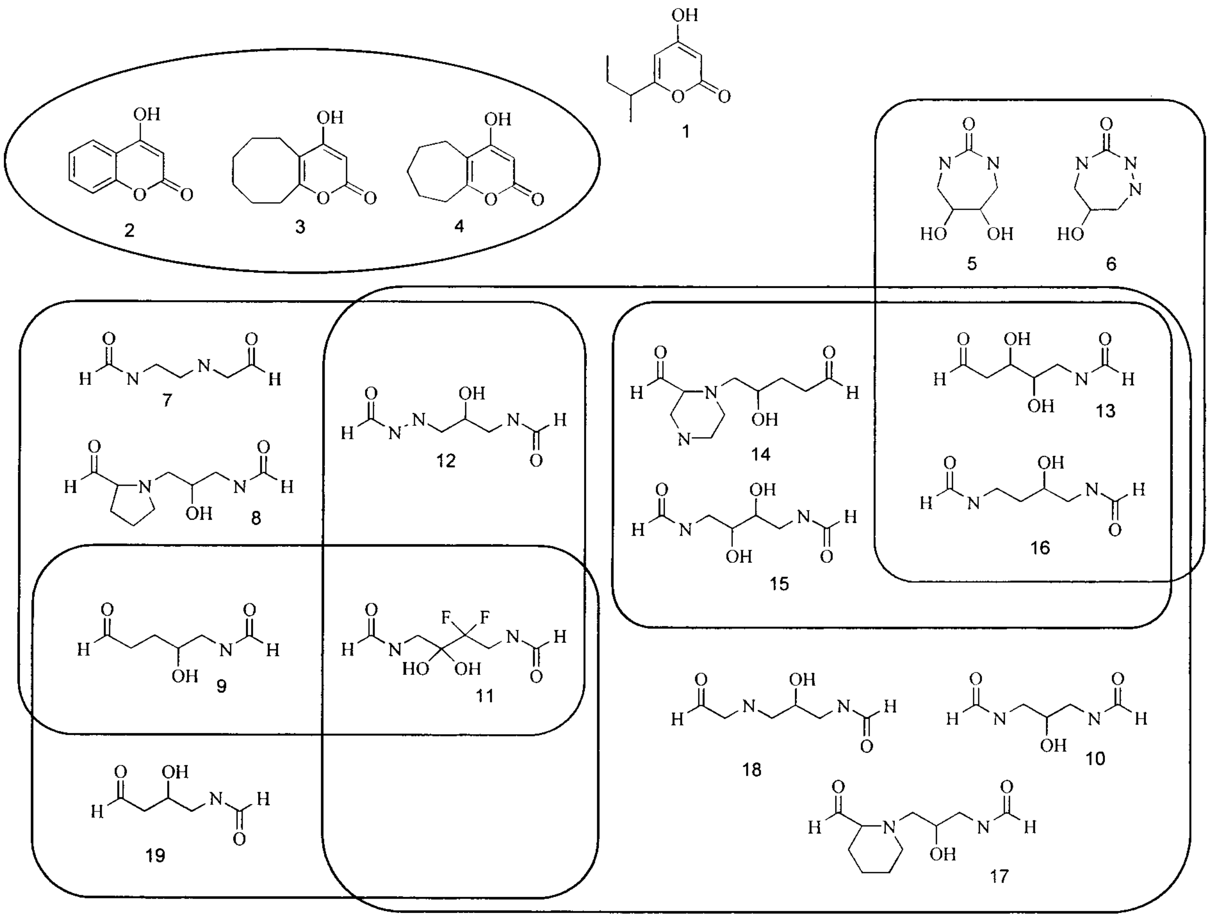

Figure 6. The lead compounds of the HIV-1 protease inhibitors are grouped according to their similarity index.

Compounds in group 8 and group 9 have low similarity although their molecular shapes are very similar. In general, ortho- and para-substituted molecules are less similar compared with ortho- and meta- or meta- and para-substituted molecules.

With the calculated similarity indices, the compounds in Figure 2 are regrouped according to their similarity as in Figure 5.

B. HIV-1 Protease Inhibitors. In Table 4, the similarity indices between the HIV protease inhibitors in Figure 4 are listed. Though the chemical formulas of the compounds are chemically not very similar, the similarity indices are relatively larger than those in Table 3. Since the inhibitors are bound to the same binding pocket in the protease, they may have similar shape and physical properties. This is the reason why the similarity indices obtained are relatively high.

Group 1 consists of pyrone derivatives that are nonpeptidic compounds, compounds **2**−**4** are similar to each other, and compound **1** has low similarity with compounds **2**−**4**.

Cyclic urea derivatives, compounds **5** and **6**, have close similarity with a few peptidomimetic molecules (**13** and **16**). Though the chemical formula of the compounds **13** and **16** are very different from the cyclic urea, their three-dimensional structure and physical properties are similar to the cyclic urea.

Group 3 molecules (**7**, **8**, **9**) which have five- or six-membered piperidine or piperazine in the lead frame form a class and are similar to compounds **11**, **12**, and **19** in spite of the big difference in chemical formula.

The compounds in group 4, **10**−**19**, form four classes which are overlapped. Compounds **13**−**16** are similar to each other. Compounds **13** and **14** form a class with compounds **5** and **6**, and compounds **14** and **15** form a class with compounds **10**−**12**, **17**, and **18**. Though compound **19** looks quite different from the other compounds, according to the calculations, it is similar to compounds **9** and **11**.

Most of peptidomimetic molecules are similar to each other. The similarity indices were distributed from 0.7 to 1.0. Since the inhibitors have the same binding pocket in protease, the inhibitors must have similar binding patterns to each other.

In Figure 6, the compounds are clustered according to the similarity indices. In the representation of the clustering, some of the similarity relations are not expressed. Compound **1** forms a cluster itself and compounds **2**−**4** form a cluster. The other clusters are (5,6,13,16), (13,14,15,16), (10,11,-12,14,15,17,18), (7,8,9,11,12), and (9,11,19).

IV. CONCLUSION

Three-dimensional descriptors which were expressed in tensor form were proposed. Especially, the tensor calculated with the free energy of hydration density was newly introduced in this work. With the spatial distribution of the physical properties of solute, electrostatic potential, dispersion interaction, polarization interaction, etc., scalar, first and second moments were calculated and were used as descriptors. Thirty-two descriptors were introduced for the calculation of the similarity between molecules.

With the newly developed descriptors, the similarity indexes were calculated for two groups of molecules. These groups are the benzenes with several kinds of functional groups and the HIV-1 protease inhibitors. In both groups, the calculated similarity indexes seem physically realistic. The HIV-1 protease inhibitors show high similarity among themselves though their chemical formulas are quite different. Because they bind the same binding pocket of the protease, the high similarities seem physically realistic.

Since, in this work, MPEOE charges were used for the HFED calculation, for the groups of molecules their accurate MPEOE charges are not available, the HFED cannot be accurately calculated. For example, for boron, aluminum, and silicon containing molecules, some heterocyclic compounds and molecular ions the HFED cannot evaluated accurately. For the ionic molecules, especially which form HB with water, some modification of the method is necessary for accurate HFED calculation.

We will extend the HFED calculation method to calculate 1-octanol solvation free energy density calculation (1-OSFED) and to the hydrophobicity density (HpD). Both densities will be used for the similarity calculation in following works.

ACKNOWLEDGMENT

This work was supported by the Korean Science and Engineering Foundation (KOSEF-96-0501-04-01-3 and KOSEF-965-0305-053-2), STEPI/FOTD (1997), and Korea Research Center for Theoretical Physics and Chemistry.

REFERENCES AND NOTES

(1) Sen, K. D. *Molecular Similarity I,II*; Springer-Verlag: Berlin, Heideberg, 1995.

(2) Carbó, R. *Molecular Similarity and Reactivity*; Kluwer Academic Publishers: Netherlands, 1995.

(3) Rouvray, D. H. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 580.

(4) Walker, P. D.; Maggiora, G. M.; Johnson, M. A.; Petke, J. D.; Merey, P. D. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 568.

(5) Mezey, P. G. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 650.

(6) Schuur, J. H.; Slelzer, P.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334.

(7) Rouvray, D. H. *J. Comput. Chem.* **1987**, *8*, 470.

(8) Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 2636.

(9) Wiener, H. *J. Chem. Phys.* **1947**, *15*, 766.

(10) Mezey, P. G. *Shape in Chemistry*; VCH Press: NewYork, London, 1993.

(11) Bertz, S. H. *J. Am. Chem. Soc.* **1981**, *103*, 3599.

(12) Nilakantan, R.; Bauman, N.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 79.

(13) Randic, M. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 373.

(14) Bone, R. G. A.; Villar, H. O. *J. Mol. Graphics* **1995**, *13*, 201.

(15) Gibson, K. D.; Scheraga, H. A. *Mol. Phys.* **1987**, *62*, 1247.

(16) Connolly, M. L. *J. Am. Chem. Soc.* **1985**, *107*, 1118.

(17) Wild, D. J.; Willett, P. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 159.

(18) No, K. T.; Kim, S. G.; Cho, K. H.; Scheraga, H. A. *Biophys. Chem.* In press.

(19) No, K. T.; Grant, J. A.; Jhon, M. S.; Scheraga, H. A. *J. Phys. Chem.* **1990**, *94*, 4740.

(20) No, K. T.; Grant, J. A.; Scheraga, H. A. *J. Phys. Chem.* **1990**, *94*, 4732.

(21) No, K. T.; Cho, K. H.; Jhon, M. S.; Scheraga, H. A. *J. Am. Chem. Soc.* **1993**, *115*, 2005.

(22) Slater, J. C.; Kirkwood, J. G. *Phys. Rev.* **1931**, *37*, 682.

(23) Good, A. C. *J. Mol. Graphics* **1992**, *10*, 114.

(24) Hagler, A. T.; Meienhofer, J., Ed.; Academic Press: New York, **1985**; Vol. 7, p 213.

(25) Hagler, A. T.; Lifson, S.; Dauber, P. *J. Am. Chem. Soc.* **1979**, *101*, 5122.

(26) InsightII(950)/Discover 2.9.5 Molecular Modeling Software; Biosym Technologies Inc.: San Diego, CA, 1995.

(27) Lunney, E. A.; Hagen, S. E.; Domagala, J. M.; Humblet, C.; Kosinski, J.; Tait, B. D.; Warmus, J. S.; Wilson, M.; Ferguson, D.; Hupe, D.; Tummino, P. J.; Baldwin, E. T.; Bhat, T. N.; Lia, B.; Erickson, J. M. *J. Med. Chem.* **1994**, *37*, 2664.

(28) Hulten, J.; Bonham, N. M.; Nillroth, U.; Hansson, T.; Zuccarello, G.; Bouzide, A.; Åqvist, J.; Classon, B.; Danielson, U. H.; Karlén, A.; Kvarnström, I.; Samulesson, B.; Hallberg, A. *J. Med. Chem.* **1997**, *40*, 885.

(29) Skulnick, H. I.; Johnson, P. D.; Aristoff, P. A.; Morris, J. K.; Lovasz, K. D.; Howe, W. J.; Watenpaugh, K. D.; Janakiraman, M. N.; Anderson, D. J.; Reischer, R. J.; Schwartz, T. M.; Banitt, L. S.; Tomich, P. K.; Lynn, J. C.; Horng, M.; Chong, K.; Hinshaw, R. R.; Dolak, L. A.; Seest, E. P.; Schwende, F. J.; Rush, B. D.; Howard, G. M.; Toth, L. N.; Wilkinson, K. R.; Kakuk, T. J.; Johnson, C. W.; Cole, S. L.; Zaya, R. M.; Zipp, G. L.; Possert, P. L.; Dalga, R. J.; Zhong, W.; Williams, M. G.; Romines, K. R. *J. Med. Chem.* **1997**, *40*, 1149.

CI980224P