# Rapid Evaluation of Molecular Shape Similarity Index Using Pairwise Calculation of the Nearest Atomic Distances

Takayuki Kotani* and Kunihiko Higashiura

Institute of Bio-Active Science, Nippon Zoki Pharmaceutical Company Ltd., Kinashi, Yashiro-cho, Kato-gun, Hyogo 673-1461, Japan

Received July 21, 2001

Rapid evaluation method for obtaining molecular shape similarity index using pairwise calculation of the nearest atomic distance respected to the template atoms was investigated. This method for calculations of similarity indices remarkably reduced required time compared with hitherto methods (especially 2 or 3 orders of magnitude faster than the previous grid-based evaluation technique) and gave without clear loss of preciseness. The potential of these improvements and possible further enhancements are discussed.

## INTRODUCTION

Molecular similarity calculations[1] are now being widely applied in molecular applications of drugs and agricultural chemicals. Though 3D quantitative structure−activity relationships (QSAR) analysis such as CoMFA[2] and CoMSIA[3,4] have been widely used for drug design with an increase in efficacy, how superpose of the molecules is one of the difficult problems to obtain reproducible and precise 3D QSAR models. In addition to the above reason, medicinal chemists are often inspired by new target molecules that should be synthesized during superposition of the lead compounds. A number of evaluation methods have been presented, one of the major techniques being the Carbó similarity index:[5−9]

$$R_{AB} = \frac{\int P_A P_B dv}{(\int P_A{}^2 dv)(\int P_B{}^2 dv)}$$

Molecular similarity $R_{AB}$ is detected from structural properties $P_A$ and $P_B$ of the two molecules being compared. The numerator measures property overlap, while denominator normalizes the similarity result. As originally applied by Carbó, quantum mechanically derived electron density is used as the structural property $P$.

Mayer modified the index to permit the evaluation of molecular shape similarity.[10] The mechanics of similarity evaluation are the same as those applied to electrostatic potential and electric filed calculations.[11−16] The molecules are surrounded by a rectilinear grid, and the structural property is evaluated at each intersection. For shape, every grid point is tested to see whether it falls inside the van der Waals surface of each molecule. The results are then applied to the following equation:

$$S_{AB} = \frac{B}{(T_A T_B)^{1/2}}$$

$B$ is the number of grid points falling inside both molecules, while $T_A$ and $T_B$ are the total number of grid points falling inside each individual molecule. The use of extremely fine grids (0.2 Å separation) in conjunction with this technique makes for prolonged calculation times, restricting its utility to SAR calculations rather than molecular alignment.[17]

While grid-based similarity evaluation techniques are common, their numerical foundations impart inherent drawbacks. The largest of these problems is that, to gain computation speed, the grids employed are normally coarse, with the consequence that resulting evaluations of spatial properties are somewhat rough. In particular, the similarity optimization through the modification of relative molecular position is coarse and rough. Good et al. reported the shape similarity index using the Gaussian function approximation to atomic orbital electron density shown as the following equation[17−19]

$$R_{AB} = \sum_{i=1}^{n}\sum_{j=1}^{m} \int (G_a{}^i + G_a{}^i + G_a{}^i)(G_x{}^j + G_y{}^j + G_z{}^j)dv / (\sum_{i=1}^{n}\sum_{i=1}^{n} (\int (G_a{}^i + G_a{}^i + G_a{}^i)^2 dv)^{1/2} (\sum_{j=1}^{m}\sum_{j=1}^{m} (\int (G_x{}^j + G_y{}^j + G_z{}^j)^2 dv)^{1/2})$$

where $G^j{}_Z = \gamma_Z e^{-\alpha_Z (r-R_j)^2}$ and $R_j$ is the nuclear coordinate position of atom $j$. The equation expands into a series of two center Gaussian overlap integrals. The two center integrals have a simple form made up of exponent values and atom center distances.[20] For example

$$\int e^{-\alpha_Z (r-R_j)^2} e^{-\alpha_Z (r-R_j)^2} dv = \left(\frac{\pi}{\alpha_1 + \alpha_2}\right)^{3/2} \exp\left(\frac{-\alpha_1 \alpha_2}{\alpha_1 + \alpha_2} \left| R_i - R_j \right|^2\right)$$

The similarity calculation can thus be broken into a succession of readily calculable exponent terms. As a result of this, it is possible to evaluate similarity rapidly and

* Corresponding author: fax: +81-795-42-5332; e-mail: t-kotani@nippon-zoki.co.jp.

10.1021/ci010068d CCC: $22.00 
Published on Web 01/09/2002

analytical. The results confirmed that the analytical Gaussian functions produced similarity values compatible with grid-based calculations, with a 2 orders of magnitude increase in evaluation speed.

Kearsley et al. proposed a discrete formula for the evaluation of molecular similarity, called the SEAL. The value $A_F$ is the similarity score for a particular alignment.[21,22]

$$A_F = -\sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij} e^{-\alpha r_{ij}^2}$$

The subscript *i* runs over *m* atoms in the first structure, and the subscript *j* runs over *n* atoms in the second structure. In the exponential, the $r^2$ term is the distance between atom *i* of the first structure and atom *j* of the second structure. The exponential form was chosen to attenuate the influence of each atom in the structure being aligned with respect to the atoms in the other structure. Parameter α adjusts this range of influence. The $w_{ij}$ preexponential factor weights each distance interaction for a pair of structure.Recently, McMahon et al. reported optimization of molecular similarity index using gradient methods.[23] Its performance for Carbó index was compared with the simplex optimization. In the vast majority cases, this method is approximately an order of magnitude faster. The Gaussian function similarity index calculations are reported to improve calculation times compared with the earlier methods such as grid-based similarity index calculations.[17−19] However, they are still a time-consuming process and require a high performance computer.

## MOLECULAR SHAPE SIMILARITY INDEX USING PAIRWISE CALCULATION OF THE NEAREST ATOMIC DISTANCES

We investigated rapid similarity indices using consideration only for the nearest atomic distances of each template atom (Figure 1). This technique is somewhat coarse and rough but extremely rapid. The strategy for calculation of similarity indices is shown in the following equation

$$S_{AB} = \frac{\sum_{i=1}^{N_A} I_i}{N_A}$$

where $N_A$ is the number of atoms of the template molecule A. The procedures described here superpose the smaller of two molecules (defined as that with fewer atoms and designated molecule B) onto the larger molecule (molecule A). The larger molecule A is selected as a template due to normalize the similarity index value in the range of 0−1. $I_i$ is a distance dependent value defined as

$$I_i = \begin{cases} 1.0 & \text{if } r_i \le rlim1 \\ 0.5 & \text{if } rlim1 < r_i \le rlim2 \end{cases}$$

where

$$rlim1 = \alpha_1 \frac{vdW_i + vdW_j}{2}$$

$$rlim2 = \alpha_2 \frac{vdW_i + vdW_j}{2}$$

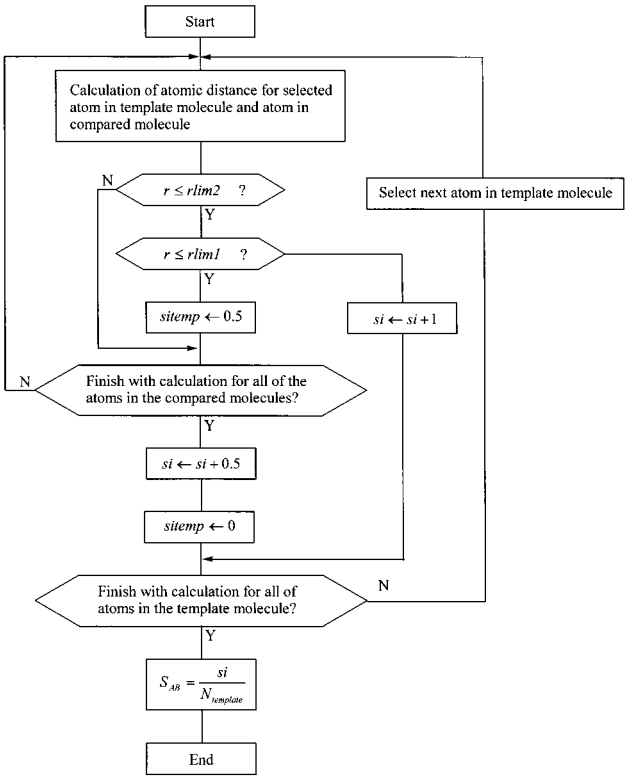



**Figure 1.** Flowchart of Method A.

$r_i$ is the distance from atom *i* of template molecule A to the nearest atom *j* on shape query B and $(vdW_i + vdW_j)/2$ is the mean of van der Waals radii of atom *i* and *j*. This method has great advantage for reducing calculation time because if atomic distance between atoms *i* and *j* is less than *rlim1*, then further atomic distances based on atom *j* are not required to calculate. Namely all atomic distance pairs are not needed to calculate and it helps to save calculation times.

## SIMILARITY CALCULATIONS

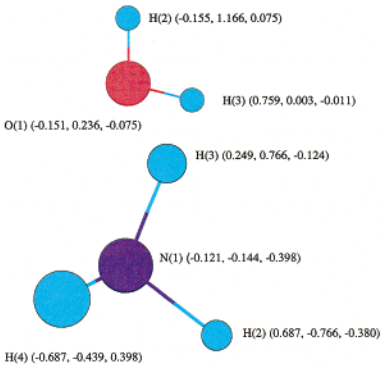
To compare the behavior of the new shape similarity routines (Method A) with the earlier grid-based routines (Meyer's method, Method B) and the Gaussian function approximation (Good's method, Method C), four separate studies were undertaken.

**Study 1.** Optimization of $\alpha_1$ and $\alpha_2$ were carried out by stepwise translations of aldosterone. Ranges of 0.5−0.8 for $\alpha_1$ and $\alpha_2$ of 1.0−1.3 were investigated. Aldosterone was moved every 0.15 Å for x, y, and z axis from the complete overlay orientation. The obtained similarity indices were compared with that of Method B. Similarity indices calculated by Method C were also compared with that of Method B. For each run, correlation constant $r^2$ and standard deviation (SD) with Method B were calculated.

**Study 2.** Random rotations (−180 to 180 °) and translations (−3.0 to 3.0 Å) were then applied systematically to a calculation of similarity indices for Methods A, B, and C. Four hundred randomly oriented conformations of aldosterone were generated by each 20 randomly oriented conformations in steps of 0.05 similarity indices calculated by Method A.

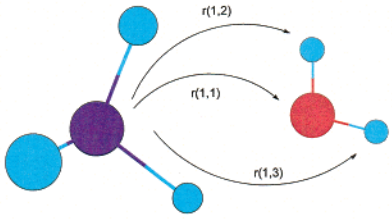
**Chart 1**

The typical procedure for calculation of similarity index is represented as follows using water molecule (Mol1) and ammonia molecule (Mol2).




1) Decide template molecule by comparing atom numbers of molecules.
Nmol1=3
Nmol2=4
Ammonia molecule (mol1) is used for template molecule.

2) Calculation of similarity index of mol2

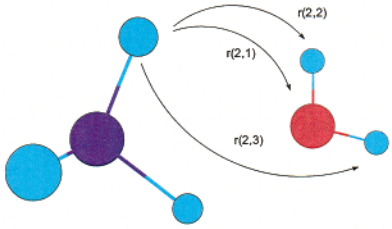



3) Calculation of atomic distance of the 1st atom of mol2 and the 1st atom of mol1.

i) For reducing calculation time, atomic distance r(1,1) is calculated when all of the atomic distances of two molecules for x-axis, y-axis, z-axis are within threshold rlim11 (average of vdW radius of the 1st atom of mol2 and 1st atom of mol1 x 1.3).
vdw(N)=1.54
vdw(O)=1.40
rlim11=1.3*(1.54+1.40)/2=1.911 (average of vdW radius of the 1st atom of mol2 and the 1st atom of mol1 x 1.3)
Abs(x2-x1)=abs(-0.121-(-0.151))=0.030<1.911
Abs(y2-y1)=abs(-0.144-(0.236))=0.380<1.911
Abs(z2-z1)=abs(-0.398-(-0.075))=0.323<1.911

ii) Temporary similarity value of 1st atom of mol2, sitemp(1), is decided by r(1,1).
r(1,1)=sqrt(0.030^2+0.380^2+0.323^2)=0.500<0.882 (average of vdW radius x 0.6)
sitemp(1)=1
Go to next step because calculation of distance between the 1st atom of mol2 and the rest atoms of mol1 are not required.

iii) Same procedure is carried out for the 2nd atom of mol2. Sitemp(2) is obtained by stepwise calculation of atomic distances.




According to the 2nd atom of mol2 and 1st atom of mol1
vdW(H)=1.20
rlim21=1.3*(1.200+1.400)/2=1.690 (average of vdW radius x 1.3)
r(2,1)=1.34
0.780 (average of vdW radius x 0.6) < r(2,1) < rlim21
sitemp(2)=0.5

iv) Calculations are continued for the 2nd atom of mol2 and the 2nd atom of mol1.
rlim22=1.3*(1.200+1.200)/2=1.560 (average of vdW radius x 1.3)
Abs(x2-x1)=abs(0.687-(-0.155))=0.842<1.690
Abs(y2-y1)=abs(-0.766-1.116)=1.822>rlim22
Go to the next atomic pair.
Keep stemp(2)=0.5

v) Calculations are continued for the 2nd atom of mol2 and the 3rd atom of mol1.
rlim23=1.3*(1.200+1.200)/2=1.560 (vdW 半径の平均 x 1.3)
r(2,3)=0.856
0.720 (average of vdW radius x 0.6) < r(2,3) < rlim23
stemp(2)=0.5

vi) Same procedures are continued and sitemps are for the 3rd and 4th atoms.
sitemp(3)=1
sitemp(4)=0.5

vii) Calculation of summation of sitemp.
sitemp=sitemp(1)+sitemp(2)+sitemp(3)+sitemp(4)=1+0.5+1+0.5=3

4) Optimization of sitemp (for example simplex optimization)

5) Normalization of similarity by dividing of sitemp by Nmol2.
Si=sitemp/Nmol2=3/4=0.75

**Table 1.** Screening of Constants for Method A[a]

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha_1$ | 0.5 | 0.6 | 0.6 | 0.6 | 0.6 | 0.7 | 0.7 | 0.7 | 0.8 | Method B |
| $\alpha_2$ | 1.3 | 1.1 | 1.2 | 1.3 | 1.4 | 1.0 | 1.1 | 1.2 | 1.0 | |
| $r^2$ ($n = 20$) | 0.974 | 0.984 | 0.984 | 0.987 | 0.986 | 0.979 | 0.982 | 0.982 | 0.981 | 0.993 |

[a] Template aldosterone was compared with 20 stepwise-translated conformations of aldosterone.

**Study 3.** Method A was applied to obtain optimum superposition of two molecules. We used Taxol as an example and simplex optimization for the above tactics.

**Study 4. FK-506**[24] was applied to superpose to the second **FK-506** using simplex optimization. For each study, Method B was undertaken using a 0.2 Å grid separation. Method B was calculated using H VDW fitted function. The location of local stationary points is a well-known falling of the simplex optimization method. For that reason, suboptimal alignments are generated in some cases. To avoid for falling into such a suboptimal location, we generated 14 starting conformations by sequential rotation of the compared molecule and carried out simplex optimizations for respective 14 runs. On each simplex, 120° rotations and 3.0 Å translations were applied for initial movement of the second molecule, while on the combined optimization, 20° rotations and 1.0 Å translations were used for the precise optimization procedure. Convergence tolerances on rotations and translations are 1.0° and 0.1 Å, respectively. All calculations were carried out on a standard 600 MHz PC. All programs were written in Fortran language and were not fully optimized.

## RESULTS AND DISCUSSION

**Studies 1 and 2.** The screening of the optimum constants $\alpha_1$ and $\alpha_2$ seems to be around 0.6 and 1.3, respectively, since wide ranges of constants showed high correlation with Method C (Table 1). Using these values gave agreeable results in a wide range of similarity index as shown in Figure 2. Further optimization of the $\alpha_1$ and $\alpha_2$ must be required as

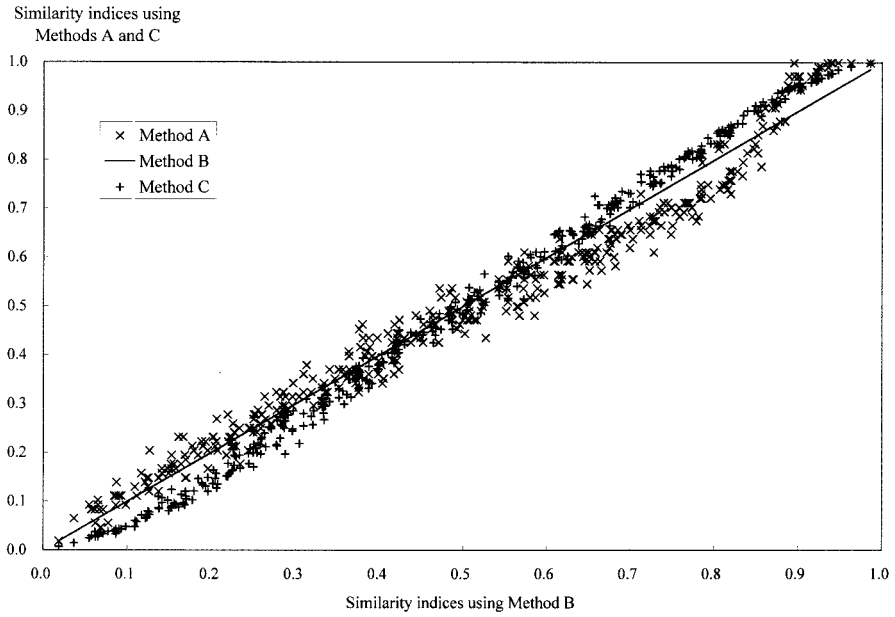



**Figure 2.** Comparison of similarity indices.

**Table 2.** Results of $r^2$ and Standard Deviation

| | Method A | Method C |
|---|---|---|
| $r^2$ ($n = 400$) | 0.971 | 0.977 |
| SD | 0.043 | 0.038 |

structural variation is limited; the behavior of Method A gave almost compatible results with Methods B and C (Table 2).

**Studies 3 and 4.** Results showed that Method A is effective for superposition of large molecules (Tables 3 and 4 and Figures 3−6). In Figures 3−6, all hydrogen atoms were removed for clarity. Method B consumes exhaustive times to reach final optimized conformer. We found that Method A gave more than about 5000-fold faster than Method B and more 100-fold faster than Method C. Especially large molecules such as taxol and **FK-506** tend to improve not only calculation times but also accuracy by increasing its denominator. Unfortunately the limitation of

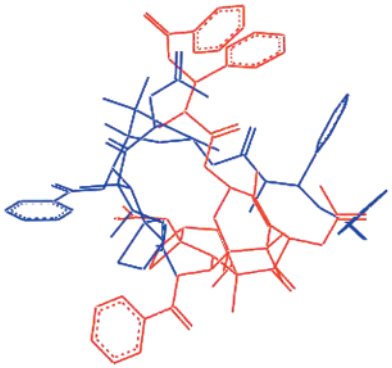


**Figure 3.** Starting conformations of taxol for simplex optimization. (Hydrogen atoms were removed for clarity.)

Method A seems to be not suitable to obtain precise superposed conformation. In the case of study 3, when we compared single point calculation of similarity indices for

**Table 3.** Relative Speeds of Simplex Optimization of Taxol with the Similarity Evaluation Techniques[a]

| | Method A | Method B | Method C | combination of Methods A and C |
|---|---|---|---|---|
| final similarity index | 0.938 (0.919) | 0.986 | 1.000 | 1.000 |
| iterations | 1219 | 1743 | 3533 | 1285 |
| calculation time (s) | 2.15 | 18783.86 | 1003.16 | 20.62 |
| ratio | 8736.68 | 1.00 | 18.72 | 910.95 |
| calculation time par iterations (s) | 0.002 | 10.777 | 0.284 | |

[a] Parentheses of final similarity index in Method A shows similarity index of the optimized conformation evaluated by Method C.

**Table 4.** Relative Speeds of Simplex Optimization of **FK-506** with the Similarity Evaluation Techniques[a]

| | Method A | Method B | Method C | combination of Methods A and C |
|---|---|---|---|---|
| final similarity index | 1.000 (0.961) | 0.995 | 0.996 | 1.000 |
| iterations | 1217 | 1579 | 2674 | 1269 |
| calculation time (s) | 3.09 | 18257.80 | 1109.01 | 29.02 |
| ratio | 5908.67 | 1.000 | 16.46 | 629.15 |
| calculation time par iterations (s) | 0.003 | 11.563 | 0.415 | |

[a] Parentheses of final similarity index in Method A shows similarity index of the optimized conformation evaluated by method C.

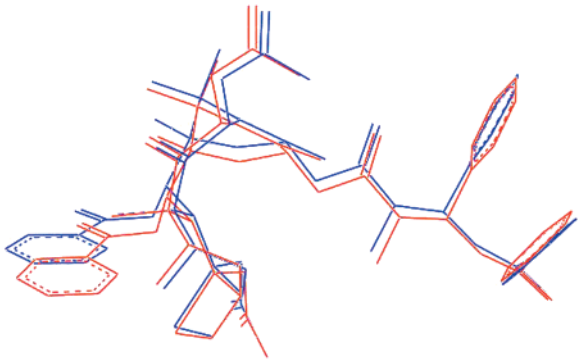


**Figure 4.** Superposition of taxol by Method A.

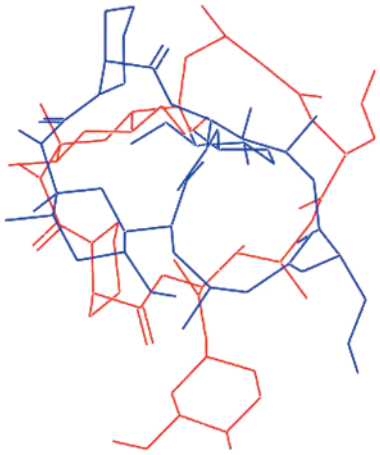


**Figure 5.** Starting conformation of **FK-506** for simplex optimization.

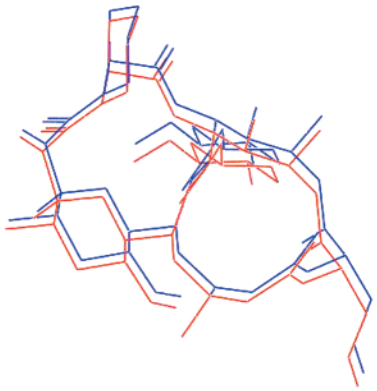


**Figure 6.** Superposition of **FK-506** by Method A.

superposed conformation by Method A, 0.938 and 0.913 were obtained by Methods A and C, respectively. RMSD (root mean squire of deviation) of this model is 0.648 Å. This limitation is explainable since the means of van der Waals radii of these atoms were used as a discriminant to accomplish rapid calculation. With another point of view, the nearest atomic distance based calculation is restricting its utility to rough molecular alignment rather than SAR calculations. Nevertheless, complementary using the earlier methods is one way to solve this limitation, i.e. the optimized conformation obtained by the nearest atomic distance based calculation is extremely suitable for the starting conformation of the earlier precise alignment methods such as Method B and/or Method C to remarkably reduced calculation times with high calculation accuracy. As expected, combination of the Methods A and C indicated about quadruple faster than application of Method C itself with the same optimized conformation. Another possibility for method A is utilization of the first screening of similarity/dissimilarity compounds, especially big molecules.

## CONCLUSIONS

Here we report a new calculation method of similarity indices and its application toward an optimized superposition of molecule. The results obtained with new functions are generally found to be similar with the earlier similarity evaluation methods according to the single point calculation. Unfortunately, simplex optimization using this new method was coarse and rough, while combinations of the earlier evaluation method gave remarkably gain calculation times to perform optimized conformations. The use of this nearest atomic distance based similarity evaluation should therefore greatly enhance the flexibility of the calculations which may be undertaken with respect to the optimization of the molecular shape.

The more raid calculation of optimized molecular similarity will enable faster database searching. If our method of optimization was to be incorporated in the search software, then the savings in time could be utilized by performing more in-depth calculations on the smaller sets of molecules.

Shape may be a useful descriptor for the repulsive force between receptor and ligand, with electrostatic potential being the main contribution to the attractive force. Comparison of both molecular properties may lead to more accurate (in terms of correlation with experimentally determined data) similarity calculations. Our methods were also observed in this area.

## ACKNOWLEDGMENT

We would like to thank Professor Miki Akamatsu (Kyoto University) for her useful comments.

## REFERENCES AND NOTES

(1) Good, A. C.; Richards, W. G. Explicit calculation of 3D molecular similarity. *Perspect. Drug Discov. Design* **1998**, *9/11*, 321−338.
(2) Cramer III, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(3) Klebe, G.; Abraham, U.; Mietzner, T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* **1994**, *37*, 4130−4146.
(4) Klebe, G.; Abraham, U. Comparative molecular similarity index analysis (CoMSIA) to study hydrogen-bonding properties and to score combinatorial libraries. *J. Comput. Aided Mol. Des.* **1999**, *13*, 1−10.
(5) Carbó, R.; Leyda, L.; Arnau, M. How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *Int. J. Quantum Chem.* **1980**, *XVII*, 1185−1189.
(6) Carbó, R.; Domingo, L. LCAO-MO similarity measures and taxonomy. *Int. J. Quantum Chem.* **1987**, *XXXII*, 517−545.
(7) Constans, P.; Amat, L.; Fradera, X.; Carbó-Dorca, R. Quantum Molecular Similarity Measures (QMSM) and the Atomic Shell Approximation (ASA). In *Advances in Molecular Similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press: Greenwich, CT, 1996; Vol. 1, pp 187−211.
(8) Amat, L.; Carbó-Dorca, R. Quantum similarity measures under atomic shell approximation: first-order density fitting using elementary Jacobi rotations. *J. Comput. Chem.* **1997**, *18*, 2023−2039.
(9) Amat, L.; Carbó-Dorca, R. Molecular electronic density fitting using elementary Jacobi rotations under atomic shell approximation. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1188−1198.
(10) Meyer, A. Y.; Richards, W. G. Similarity of molecular shape. *J. Comput. Aided Mol. Des.* **1991**, *5*, 427−439.

(11) Hodgkin, E. E.; Richards, W. G. A semiempirical method for calculating molecular similarity. *J. Chem. Soc., Chem. Commun.* **1986**, *17*, 1342−1344.
(12) Hodgkin, E. E.; Richards, W. G. Molecular Similarity Based on Electrostatic Potential and Electric Field. *Int. J. Quantum Chem.* **1987**, *14*, 105−110.
(13) Burt, C.; Richards, W. G. Molecular similarity: the introduction of flexible fitting. *J. Comput. Aided Mol. Des.* **1990**, *4*, 231−238.
(14) Burt, C.; Richards, W. G.; Huxley, P. The application of molecular similarity calculations. *J. Comput. Chem.* **1990**, *11*, 1139−1146.
(15) Richards, A. M. Quantitative comparison of molecular electrostatic potentials for structure−activity studies. *J. Comput. Chem.* **1991**, *12*, 959−969.
(16) Thorner, D. A.; Wild, D. J.; Willett, P.; Wright, P. M. Calculation of structural similarity by the alignment of molecular electrostatic potentials. *Perspect. Drug Discov. Design* **1998**, *9/11*, 301−320.
(17) Good, A. C.; G., R. W. Rapid evaluation of shape similarity using Gaussian function. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 112−116.
(18) Good, A. C. The calculation of molecular similarity: alternative formulas, data manipulation and graphical display. *J. Mol. Graph.* **1992**, *10*, 144−151, 162.
(19) Drayton, S. K.; Edwards, K.; Jewell, N.; Turner, D. B.; Wild, D. J.; Willett, P.; M., W. P.; Simmons, K. Similarity searching in files of three-dimensional chemical structures: Identification of bioactive molecules. *Internet J. Chem.* at URL http://www.ijc.com/articles/1998v1/37.
(20) Szabo, A.; Ostland, N. S. In *Modern Quantum Chemistry*; Macmillan: New York, 1982; pp 410−412.
(21) Kearsley, S. K.; Smith, G. M. An alternative method for the alignment of molecular structures: maximizing electrostatic and steric overlap. *Tetrahedron Comput. Methodol.* **1990**, *3*, 615−633.
(22) Klebe, G.; Mietzner, T.; Weber, F. Different approaches toward an automatic structural alignment of drug molecules: applications to sterol mimics, thrombin and thermolysin inhibitors. *J. Comput. Aided Mol. Des.* **1994**, *8*, 751−778.
(23) McMahon, A. J.; King, P. L. Optimization of Carbó molecular similarity index using gradient methods. *J. Comput. Chem.* **1997**, *18*, 151−158.
(24) Atomic coordinates of **FK-506** were extracted from 1BKF, which hydrogens were generated by Chem3D (CambridgeSoft Corporation). Itoh, S.; DeCenzo, M. T.; Livingston, D. J.; Pearlman, D. A.; Navia, M. A. Conformation of **FK506** in X-ray structures of its complexes with human recombinant FKBP12 mutants. *Bioorg. Med. Chem. Lett.* **1995**, *5*, 1983−1988.

CI010068D