A Fast Method of Molecular Shape Comparison: A Simple Application of a Gaussian Description of Molecular Shape

J. A. GRANT*
Zeneca Pharmaceuticals, Mereside, Macclesfield, Cheshire SK10 4TF, England

M. A. GALLARDO
Química Física, Facultad de Ciencias, Universidad de Zaragoza, 50009-Zaragoza, Spain

B. T. PICKUP
Centre for Molecular Materials, Dept. of Chemistry, The University of Sheffield, S3 7HF, England

Received 12 September 1995; accepted 16 January 1996

ABSTRACT

A Gaussian description of molecular shape is used to compare the shapes of two molecules by analytically optimizing their volume intersection. The method is applied to predict the relative orientation of ligand series binding to the proteins, thrombin, HIV protease, and thermolysin. The method is also used to quantify the degree of chirality of asymmetric molecules and to investigate the chirality of biphenyl and the amino acids. The shape comparison method uses the newly described shape multipoles that can also be used to describe the inherent shape of molecules. Some results of calculated shape quadrupoles are given for the ligands used in this work. © 1996 by John Wiley & Sons, Inc.

*Author to whom all correspondence should be addressed.

Journal of Computational Chemistry, Vol. 17, No. 14, 1653–1666 (1996)

CCC 0192-8651 / 96 / 141653-14

Introduction

The shape of a molecule is an important consideration in the design of selective ligands for protein and DNA binding. The well known lock-and-key hypothesis requires a degree of shape complementarity between ligand and receptor. On the other hand, a set of different ligands that all bind to the same receptor site, giving rise to a similar pharmacological response, is expected to possess a degree of molecular shape similarity. To exploit such shape similarity an increasingly useful tool for rational drug design in medicinal chemistry is the method of molecular shape comparison (MSC) that compares the shape of two or more molecules and identifies common spatial features. It is hoped that such comparisons can lead to alternative pharmacophoric models in the process of ligand design. A recent[1] analytical method provides a measure of shape similarity and a way to align molecules such that it is maximized. Application of this method to conformationally flexible angiotensin receptor antagonists identified an experimentally deduced pharmacophore model, as well as several attractive alternative models. The method itself makes use of the most widely used description of molecular shape[2-5] that treats a molecule as a set of intersecting spheres. The exposed surface of these spheres defines the boundary of a molecular volume. However, the actual concept of molecular shape is not so trivial. The complexities and various mathematical treatments of this subject are thoroughly discussed in a recent treatise by Mezey.[6] Here it is emphasized that a rigorous treatment of molecular shape should reflect the quantum nature of molecules, and represent the fuzzy nature of electron charge distributions. The molecular shape complementarily invoked by the lock-and-key approximation, can be understood by considering the dominant role played by the overlap of the peripheral regions of the electron charge distribution. As soon as there is any significant overlap of charge distribution the potential energy becomes strongly repulsive. The peripheral regions of the electron density are therefore very important in assessing the steric effects critical in molecular recognition.

To improve the hard-sphere model, we recently introduced[7] a computationally efficient Gaussian description of molecular shape, which makes use of a generalized coalescence theorem, to compute analytical formulae for molecular volumes, areas, end their nuclear coordinate derivatives. This model accounts for the inherent softness (or fuzziness) associated with electron charge distributions, without being a complex quantum mechanical treatment. In this work we report on the application of this model to the problem of molecular shape comparison, intended to efficiently find the optimal shape similarity between two molecules. An objective of this method is to predict if a molecule with a chemical skeleton different from a known ligand possesses the relevant degree of shape similarity to be considered as a putative ligand for a receptor (probably of unknown structure).

The utility of the Gaussian function, in particular the product theorem, the simple analytical nature of integrals over all space, and the continuity of derivatives,[8] have long been recognized. We briefly describe some relevant applications of the Gaussian function in chemistry. Our model borrows heavily from the ideas introduced to quantum chemistry by Boys,[9] in which Gaussian functions were used to represent atomic orbitals. An empirical Gaussian model of electron density was devised by Diamond[10] as part of a real space refinement procedure to determine protein structures. A very similar representation of molecular electron distributions was implemented by Marshall and Barry,[11] as part of the active analog approach to deduce pharmacophoric patterns from pharmacological data.[12] It is well known that given a fairly smooth function that vanishes rapidly, it is possible to obtain a reasonably accurate expansion in terms of a number of Gaussians.[8,13] This property has been used to represent the conventional Lennard–Jones potential by expanding the functions r^{-12} and r^{-6} in terms of a linear combination of truncated Gaussians.[14] Such a representation leads to a convenient analytical solution of the diffusion equation, and hence a route to globally optimizing the structures of atom clusters[14] and oligopeptides.[15] The inverse distance dependence of the electrostatic potential (r^{-1}) has also been expanded as a linear combination of Gaussian functions as part of a method to compare the electrostatic potentials of different molecules.[16] There are a number of analytical and numerical methods to seek relative orientations of molecules that maximize similarity in some molecular property such as steric shape, electrostatic potential, hydrophobocity, and lipophilicity. There are excel-

lent recent reviews of these techniques by Dean,[17] Good,[18] and Klebe,[19] and in a couple of these methods Gaussians play a central role in the alignment function. SEAL[20] (steric and electrostatic alignment method) uses an empirical function in which a Gaussian (alternatively a Lorentzian) attenuates a property-weighted contribution of all atom pairs between two structures. The very innovative and elegant work of Hodgkin et al.[16] Good and Richards[21] uses atom centered Gaussian functions to approximate *ab initio* electron densities, and using this model analytically aligns molecules by optimizing the Carbo index using a simplex procedure.

Theory and Methods

The Gaussian model of molecular shape[7] uses a representation in which each atomic site i with coordinates $\mathbf{R}_i = (\mathbf{X}_i, \mathbf{Y}_i, \mathbf{Z}_i)$, is given by a spherical Gaussian

$$\rho_i^g(r_i) = p_i \exp(-\alpha_i r_i^2), \tag{1}$$

where the local coordinate

$$r_i = |\mathbf{r}_i| = |\mathbf{r} - \mathbf{R}_i| \tag{2}$$

is defined as a distance vector from the atomic center, and the exponent

$$\alpha_i = \frac{\kappa_i}{\sigma_i^2} \tag{3}$$

is defined using a parameter, σ_i, which is loosely speaking the "radius" of the atom. The dimensionless parameter κ_i can be chosen such that

$$\kappa_i = \frac{\pi}{\lambda_i^{2/3}}, \tag{4}$$

in which case the "atomic volume" of center i is

$$V_i^g = \int d\mathbf{r}_i \; \rho_i^g(r_i) = \frac{4\pi}{3}\sigma_i^3, \tag{5}$$

provided we choose the Gaussian weight (p_i) such that

$$p_i \lambda_i = \frac{4\pi}{3}. \tag{6}$$

We emphasize that the integral appearing in (5) is a volume integral ($d\mathbf{r} = dx\,dy\,dz$), where we use the notation that unlabeled integrals are over the whole of space. The shape of a molecule can be represented using the *shape-density* function

$$\rho^g(\mathbf{r}) = \sum_i \rho_i^g - \sum_{i<j} \rho_i^g \rho_j^g + \sum_{i<j<k} \rho_i^g \rho_j^g \rho_k^g - \sum_{i<j<k<l} \rho_i^g \rho_j^g \rho_k^g \rho_l^g + \cdots. \tag{7}$$

The volume of the molecule can hence be written as

$$V^g = \int d\mathbf{r}\, \rho^g(\mathbf{r}) = \sum_i V_i^g - \sum_{i<j} V_{ij}^g + \sum_{i<j<k} V_{ijk}^g - \cdots, \tag{8}$$

where the multiple summation terms represent the intersection volumes, which must be allowed for in assessing the total volume of a set of intersecting soft spheres. This has a clear precedent in standard hard-sphere methodologies[4, 22]; and although the present work is not aimed at an exact reproduction of hard-sphere volumes, we do want to use a parallel mathematics. Hence the pair intersection volume of two sites i and j is defined as

$$V_{ij}^g = \int d\mathbf{r}\, \rho_i^g \rho_j^g, \tag{9}$$

and similarly for higher order intersections. A convenient alternative representation of the Gaussian shape density is to use the equivalent product formula

$$\rho^g = 1 - \prod_i^{n\,\text{atoms}} (1 - \rho_i^g). \tag{10}$$

Although we have been unable to directly integrate this formula analytically to obtain volumes, it has been integrated (as part of a separate piece of work) by quadrature (numerically)[23] for a set of blocked amino acids, giving volumes almost identical to those obtained from eq. (8). In principle, the shape density could be evaluated on a lattice as part of a CoMFA-type[24] analysis. Such a representation of the steric properties of a molecule has the advantage of being smoothly varying, and because there are no singularities, arbitrary cutoffs are not required to avoid unacceptably large numerical values. In this respect such a function has similar advantages to the Gaussian attenuated functions proposed by Klebe et al.[25] The calcula-

tion of (10) on a grid does not form part of the shape matching algorithm described in the following, although it has proved convenient for the computation of contour representations of molecular shape for graphical display.

The important point about the Gaussian representation given in (8) is that all volumes and intersection volumes can be calculated analytically in terms of simple algebraic formulae that involve nothing more complex than an exponential. For real molecules, with an arbitrary number of atoms this formula would comprise, however, a combinatorial number of product terms, the evaluation of which would be obviously impractical. A simple algorithm is therefore adopted to compute volumes and areas from eq. (8) for an arbitrary number of atoms, which reduces the number of terms computed in (8) by applying a cutoff criterion to define a local neighbor list for each atom. This is fully described in ref. 7. The pair intersection volume is given by

$$V_{ij} = p_i p_j K_{ij} \left(\frac{\pi}{\alpha_i + \alpha_j} \right)^{3/2}, \qquad (11)$$

where

$$K_{ij} = \exp\left(- \frac{\alpha_i \alpha_j R_{ij}^2}{\alpha_i + \alpha_j} \right), \qquad (12)$$

and R_{ij} is the distance between atoms i and j. The existence of simple formulae for volumes implies that volume gradients also have analytic formulae. In particular, the gradient of a volume such as V_{ij} in (11) with respect to nuclear centers i or j is trivial to compute because, for example,

$$\frac{\partial V_{ij}^g}{\partial X_i} = - \frac{2\alpha_i \alpha_j}{\alpha_i + \alpha_j} (X_i - X_j) V_{ij}^g, \qquad (13)$$

which implies that gradients (and higher gradients) are proportional to the volume factors already computed. No extra transcendental function evaluations are required. A complete set of expressions for general intersections, their nuclear coordinate derivatives, and Gaussian areas has already been given.[7]

The main purpose of this article is shape matching. The matching process is essentially a matter of maximizing the intersection volume of two molecules by rigidly translating and rotating one of them with respect to the other. Consider molecules A and B with Gaussian densities

$$\rho_\chi^g = 1 - \prod_{i \in \chi} (1 - \rho_i^g), \quad \chi = A \text{ or } B, \qquad (14)$$

then the molecular intersection volume is

$$V_{AB}^g = \int d\mathbf{r}\, \rho_A^g \rho_B^g. \qquad (15)$$

Expanding the densities (14) using representation (7), we find that it is possible to develop a series

$$V_{AB}^g = \sum_{i \in A,\, j \in B} V_{ij}^g - \sum_{i,\, j \in A,\, k \in B} V_{ijk}^g - \sum_{i,\, j \in B,\, k \in A} V_{ijk}^g + \cdots \qquad (16)$$

in which pair, triple, and higher atom based densities appear, subject to the restriction that the index ranges always cover both molecules. It is important to realize that there are no new volume formulae other than those already presented.

The optimization technique that we have utilized relies upon analytic evaluation of the first and second gradients of the intersection volume with respect to the rigid-body rotations and translations. The rotational part of the problem is solved using a quaternion formulation[26] of the parameters specifying the rotations. This replaces the Euler angles usually used to specify rotations by four real parameters and a constraint. The Euler angles have a plane of singularities that reduce the efficiency of minimization techniques and give rise to artificial saddle points and minima, a problem not encountered when using the purely algebraic quaternion parameters. Our implementation of a singularity-free rigid-body optimization essentially follows that of Markey et al.,[27] although after some experimentation we choose a penalty function suggested by Kearsley[28] to constrain the quaternion parameters to unity. Hence one is thus effectively optimizing in a six-parameter space.

In searching for the optimal shape comparison, optimization from a single initial relative orientation will not necessarily lead to the global maximum in the volume intersection. In principle, any suitable global optimization method (for example Monte Carlo or simulated annealing) can be used to search the rotational/translational space to seek the globally optimal shape match. However, adapting the method of Masek,[1] we find it convenient to define four initial orientations as starting points for optimization. These points are chosen by aligning both molecules to have a common origin

at their respective *shape centroids* that are defined as

$$\bar{\mathbf{R}}_\chi = \frac{\int d\mathbf{r}\, \mathbf{r} \rho_\chi^g}{\int d\mathbf{r}\, \rho_\chi^g} \tag{17}$$

for molecules $\chi = A$ or B. Molecules are then subsequently aligned by first computing a shape "quadrupole" tensor, thus

$$M_{\alpha\beta} = \frac{\int d\mathbf{r}\, \mathbf{r}_\alpha \mathbf{r}_\beta \rho_\chi^g}{\int d\mathbf{r}\, \rho_\chi^g}, \tag{18}$$

where we have used Greek subscripts to indicate Cartesian components (it is straightforward to compute higher order shape multipoles[7,29]). The eigenvectors of the 3×3 matrix in (18) can be made to represent canonical right-handed axes in terms of which of the two molecules can be aligned. From such an alignment, rotation by π radians about any of these axes generates a total of four initial orientations. This procedure provides well-spaced starting points in rotational/translational space, and makes allowance for indeterminate phases of the eigenvectors that result in flipping of *pairs* of axis directions, which represent bona fide right-hand axis systems.

The method of shape comparison that we are proposing requires only two parameters per atom center, namely σ_i and p_i. The radius parameter is normally chosen as the van der Waals radius. The Gaussian weights p_i fix the exponent through constraint (6) and eqs. (3) and (4). It has been demonstrated previously[7] via protein calculations that the value $p_i = 2.70$ ($\lambda = 1.5514$) gives good results for intersection volumes. One can argue that a more suitable p_i value for shape matching would be obtained by taking the limit of V_{ij}^g as the interatom distance becomes zero. In this limit, the intersection volume for equal radius spheres should just be $4\pi\sigma^3/3$. Hence (11) implies

$$V_{ij} = \frac{4\pi\sigma^3}{3} = p_i^2 \left(\frac{\pi}{2\alpha_i}\right)^{3/2}$$
$$\Rightarrow p_i = 2\sqrt{2}, \tag{19}$$

which is close to the value of 2.70 recommended previously.[7] The exact choice of p_i is discussed in the Results section. This ensures that the Gaussian formulation works in the same way as the hard-sphere model in the pair coalescence limit for equal radius spheres. The maximum index obtained after optimization is a measure of goodness of matching for the two molecules. For general comparisons, however, one needs to use a normalization. Although this is arbitrary, it is convenient to use a standard method introduced by Hodgkin and Richards,[30,31]

$$S^{AB} = \frac{2\int d\mathbf{r}\, \rho_A^g \rho_B^g}{\int d\mathbf{r}(\rho_A^2 + \rho_B^2)}, \tag{20}$$

which obviously satisfies the bounds

$$0 \le S^{AB} \le 1. \tag{21}$$

The normalization of the index plays no role in the present work, because we are only interested in the molecular conformations of specifically matched structures, rather than global shape comparisons for large numbers of matched molecules.

Results

To demonstrate the utility of the Gaussian shape matching method, we present a couple of illustrative applications, namely the prediction of the relative alignment of different ligands at a common protein receptor site, and the computation of chiral volumes. We first discuss the computational performance of our method relative to established analytical hard-sphere techniques.

An important feature of our Gaussian description of molecular shape is that there are only two parameters describing each atom. One of these parameters is an atomic radius, and for all of the calculations in this work we use a set of radii based on those given by Connolly[2,32] and shown in Table I. In all of the calculations presented in this section a Gaussian weight of $p = 2.70$ ($\lambda = 1.5514$) has been used. Table II shows various timings for the Gaussian approach in CPU seconds

TABLE I.
Atomic Radii Used for Molecule Calculations.

Atom Type	Radius (Å)
C	1.70
N	1.65
H	1.00
O	1.60
S	1.90
P	1.90
F	1.30

TABLE II.
Performance of Techniques to Calculate Gaussian Volumes and Their Gradients.

Molecule[a]	No. Atoms	Gaussian Vol. (Å^3)	Hard-Sphere Vol. (Å^3)	*T*1 (s)	*T*2 (s)	*T*3 (s)	*T*4 (s)	*T*5 (s)
Ala	22	133	129	0.02	0.02	0.03	14.4	0.00
Helix a	110	1276	1277	0.04	0.05	0.12	41.5	0.01
1mjc	514	5770	5785	0.20	0.27	0.43	266.5	0.18
2int	1047	11848	11864	0.41	0.53	0.93	521.0	0.77
3app	2366	26467	26449	1.18	1.20	2.21	1089.3	4.01
1rve	4046	45001	45074	2.10	2.50	4.75	1804.1	12.6

[a] Molecule and column labels are identified in the text.

for a SGI Indigo II R4000 processor running at 100 MHz. The molecule Ala is the amino acid alanine blocked with acetyl and *N*-methyl at the *N* and C termini, respectively, helix a is a helical segment of interleukin-4 (residues 5–18),[33] and the remaining molecules are proteins identified by their Brookhaven[34] entry code. The columns labeled *T*1, *T*2, and *T*3 are CPU times required to calculate the volume, volume and first gradient, and the latter plus the Hessian, respectively (assuming interatomic distances were precomputed). Column *T*4 gives the time required to calculate the Gaussian shape density (7) at each point on a 65^3 rectangular grid. Such a calculation is not used in the shape matching algorithm, but has proved useful for visualization purposes. We also calculated the function on a rectangular grid in only a few seconds, even for a 6000 atom protein, by introducing a spherical cutoff, without any great loss in accuracy. *T*5 gives the time to build interatomic pairwise distances required for the algorithm. Columns *T*1–*T*4 show that overall our method is approximately linear in the number of atoms, *N*. The dependence in column *T*5, on the other hand is roughly quadratic in *N* (the algorithm uses a neighbor list approach) as expected. The point of separating these parts of the algorithm (interatomic distance calculation, from the rest) is that the neighbor distances are usually available in precomputed form in, for example, molecular mechanics packages in which the Gaussian shape model could be introduced as part of a simple *analytical* area or volume based solvation energy term. It should also be emphasized that the computations of hard-sphere volume derivatives for molecules with thousands of atoms are extraordinarily expensive. Table III shows comparative timings required to overlay copies of identical molecules from a well-defined nonoverlayed starting point (2 Å separation, and a 30° rotation along the long shape–quadrupole axis). The molecules used in this table were already identified, except 7hvp, which is an HIV protease ligand. The overlay is calculated by maximizing the intersection volume with respect to rigid translations and rotations of one of the copies, and convergence is defined when the root mean square (rms) of the gradient is less than 10^{-3}. We find this to be a stringent convergence criteria. For example typical initial rms values of the gradient are $\approx 10^3$; values of $\approx 10^{-1}$ are obtained for pairs of structures that appear from inspection using molecular graphics to exactly align. In all cases the final structure corresponded to an exact overlay of the two identical molecules, demonstrating that the Gaussian overlap model behaves correctly in the limit of a unit rotation/null translation. From the discussion in the Methods section concerning the behavior of the Gaussian product theorem as the interatom distance becomes zero, it can be seen that the exact superposition of identical molecules could not necessarily be expected. However, it seems that consideration of many Gaussian intersections compensates any errors introduced by the Gaussian volume not exactly converging to the maximal hard-sphere intersection volume, as is the case for a pair of Gaussians with a weight other than $p = 2\sqrt{2}$. Investigation of the variation in the

TABLE III.
Comparative Timings for Gaussian and Hard-Sphere Overlay Methods.

Molecule	No. Atoms	FEVAL G	FEVAL HS	Time (s) G	Time (s) HS
Benzene	12	35	77	0.6	17.0
Ala	22	35	99	1.2	49.8
7hvp	61	39	73	9.1	144.6
Helix a	110	50	124	48.2	468.0

behavior of these calculations, and in those described below suggested that the results changed little for Gaussian weights chosen around $p = 2.70 \pm 0.3$. The number of function evaluations required for Gaussian (G) and hard-sphere (HS) methods is given in the column labeled FEVAL of Table III, for a series of molecules similar to that found in Table II. It can be seen that hard-sphere techniques require roughly double the number of iterations. This is probably because the Gaussian volume intersection function is more smoothly varying than the piecewise continuous hard-sphere one. The derivatives of the Gaussian function are therefore better guides for convergence directions on the rigid-body parameter surface. The comparative timings given in Table III indicate a good improvement in performance relative to analogous analytical hard-sphere models.

Having established that the shape-comparison algorithm will efficiently align identical molecules, we now apply the method to the problem of predicting the relative alignment of structurally different ligands binding at a common protein receptor site. Three protein structures were chosen, namely thrombin, HIV protease (HIV PR), and thermolysin. For each protein a set of structurally diverse ligands was selected. They are shown in Figures 1–3 and are identified by the Brookhaven entry code associated with the protein–ligand complex. To obtain the experimental relative orientation of a pair of ligands each in the conformation bound to a given protein, the crystallographic coordinates of the protein–ligand structures were transformed into a common frame by an optimal least-squares alignment of the C^{α} backbone atoms. For a given set of protein–ligand complexes there may be minor conformational or sequence differences between the protein itself in different complexes. The flexible nature of proteins means that the conformation of the binding pocket can be different for different ligands. An example of sequence differences is the case of HIV PR, in which

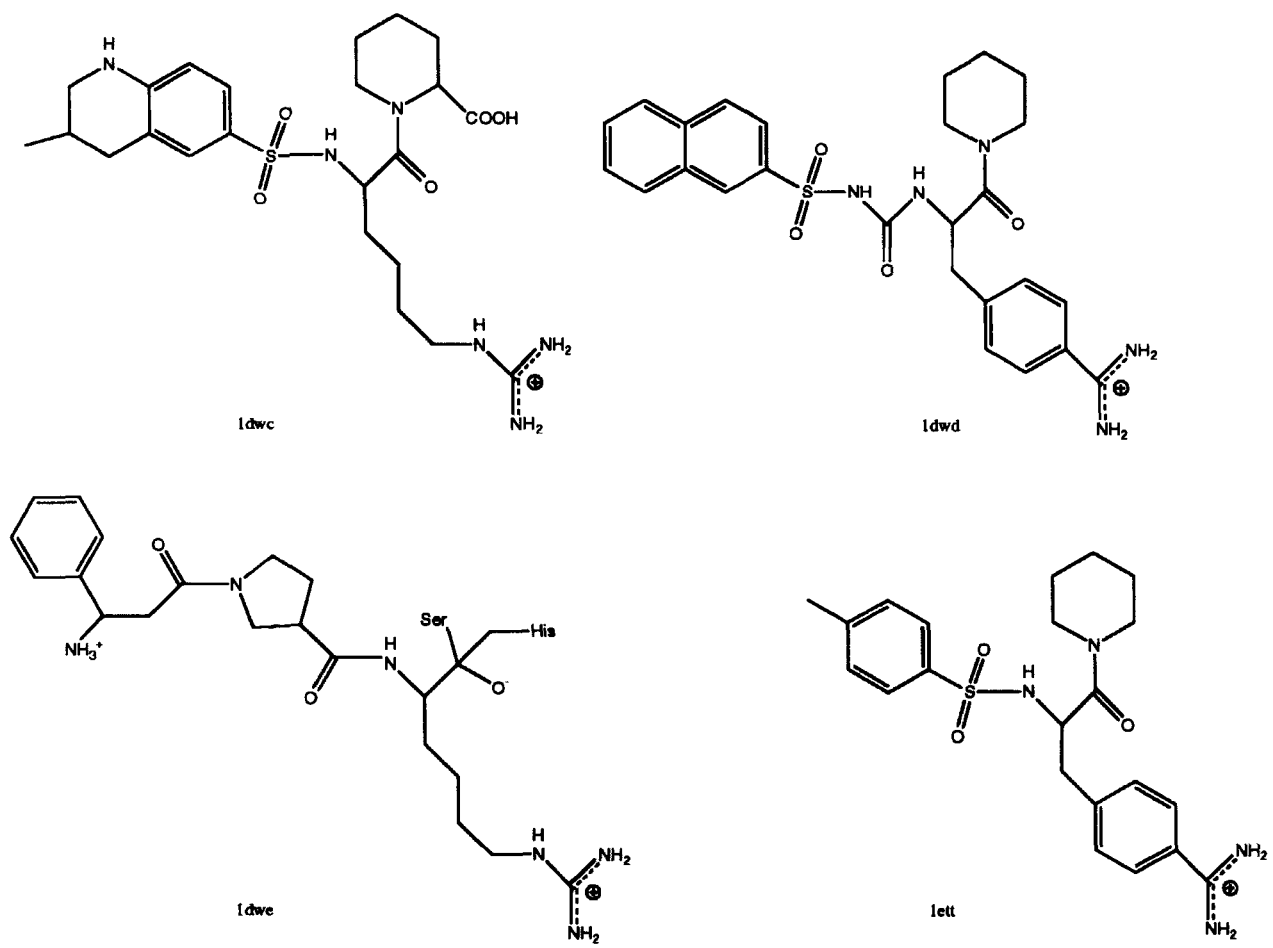
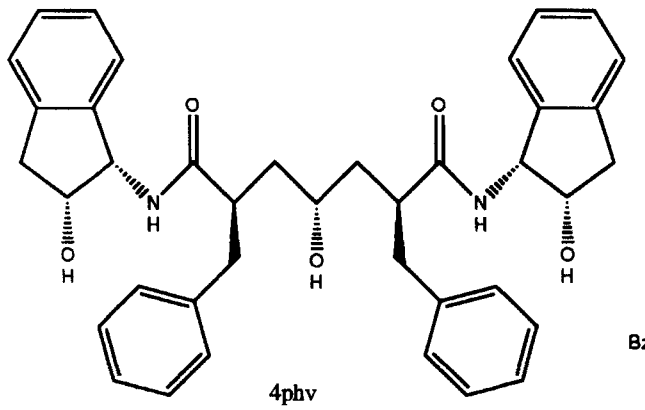
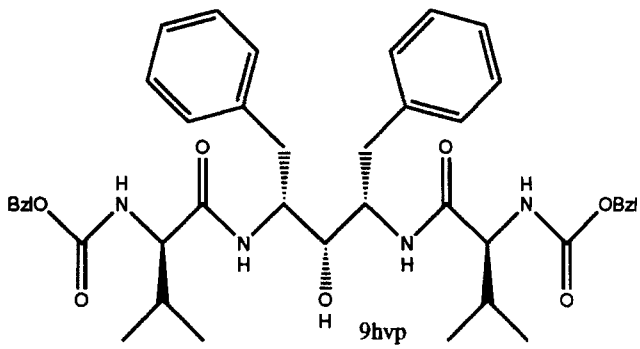


FIGURE 1. Thrombin ligands.

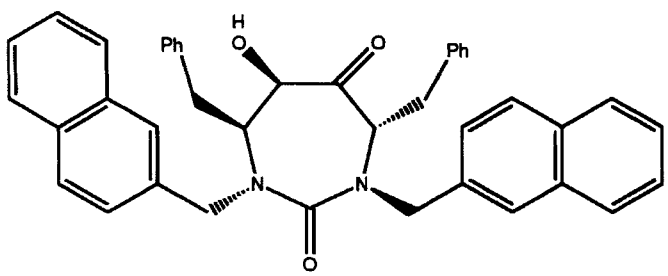
some of the structures derived from a synthetic protein have nondisulfide linked cysteine residues replaced by α-aminobutyric acid. Nonetheless the rms differences between C^{α} backbone atoms of the aligned proteins were always in the range 0.2–1.0 Å, typically ≈ 0.4 Å. The errors were always very small for any pair of the thermolysin structures in which crystals of the complex were obtained by soaking, and the largest error occurred in comparing two thrombin structures obtained from different species. In all cases the errors are very much smaller if only the backbone region around the



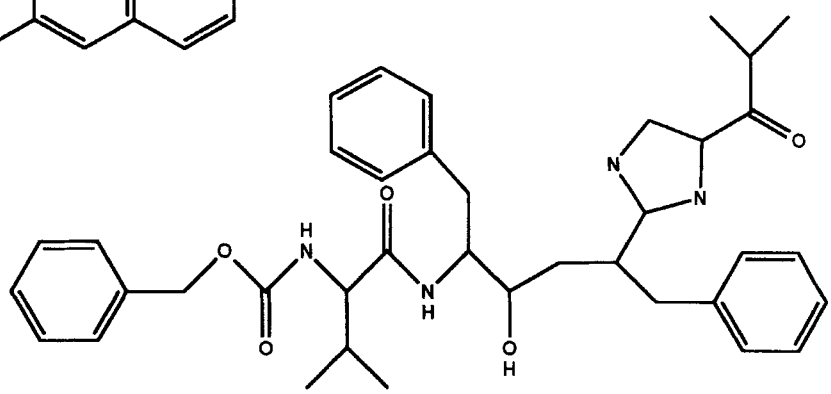
4phv



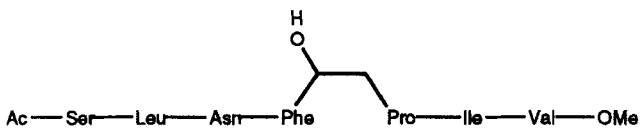
9hvp



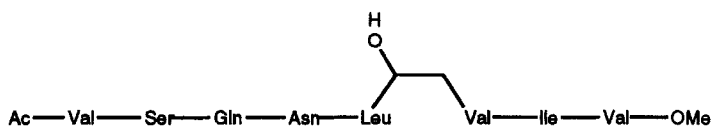
1hvr



1hps



7hvp



8hvp

FIGURE 2. HIV protease ligands.

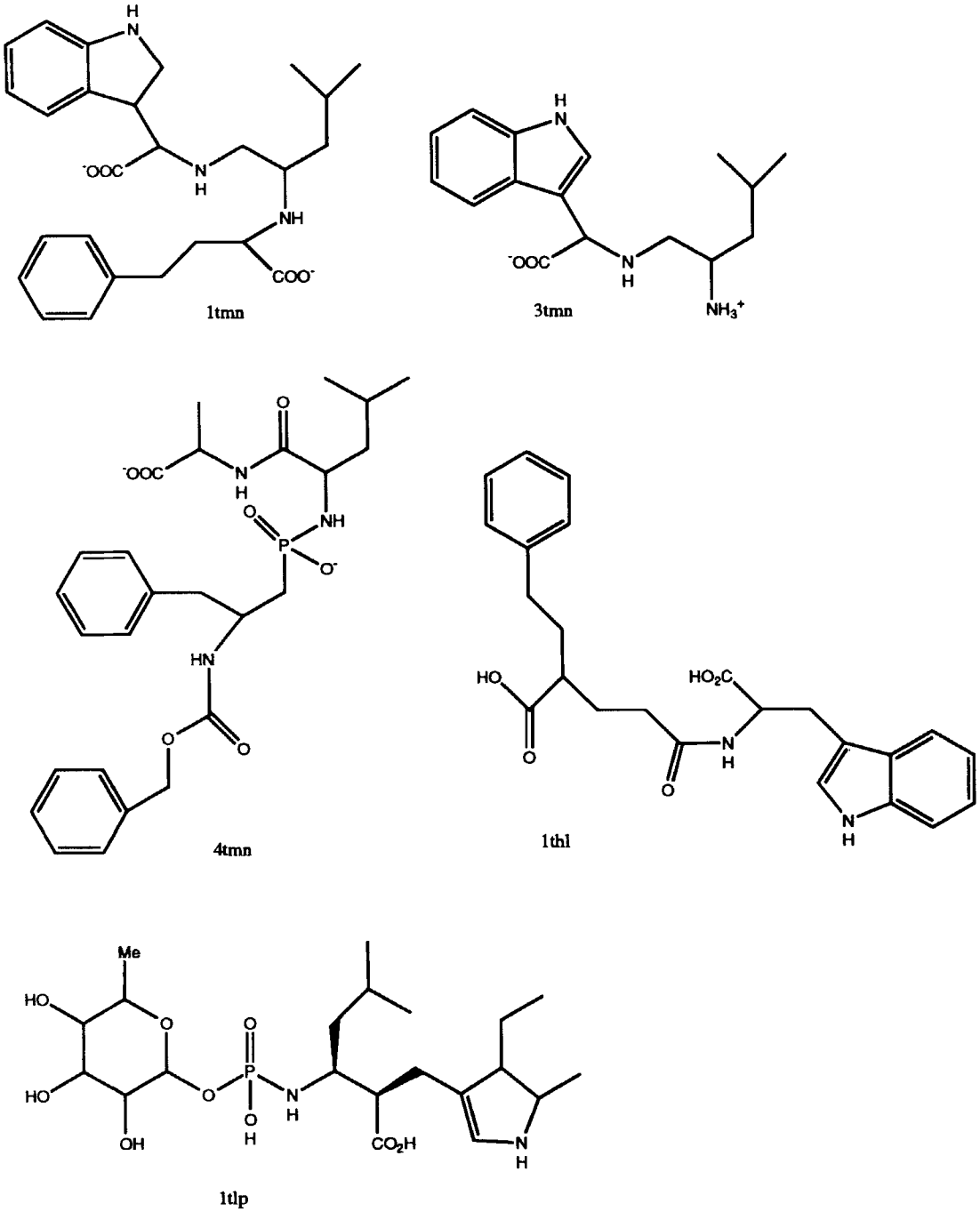


FIGURE 3. Thermolysin ligands.

active site is considered as part of the least-squares alignment procedure. The predicted relative orientation of the ligands is determined purely by using the Gaussian shape comparison procedure described in the previous section. This method makes use of the molecular shape quadrupole to define the initial relative orientation of the ligands, as described in the Methods section. Tables IV–VI show the agreement between the Gaussian (V^g) and the hard-sphere volume (V^{hs}) as well as the eigenvalues of the shape quadrupole (the principal shape quadrupoles), $Q1$, $Q2$, and $Q3$, computed in a frame in which the off-diagonal tensor elements are zero. The thrombin ligands show relatively little variation in each of the components. However, for the ligands of HIV PR (Table V), although

TABLE IV.
Volumes and Principle Shape Quadrupoles for Thrombin Ligands.

Ligand	V^g (Å^3)	V^{hs} (Å^3)	$Q1$ (Å^2)	$Q2$ (Å^2)	$Q3$ (Å^2)
1dwc	394.1	394.6	10.8	6.4	3.3
1dwd	411.6	411.5	14.4	6.2	3.2
1dwe	344.0	344.1	14.3	7.2	2.0
1ett	345.8	344.6	12.0	4.5	3.5

TABLE V.
Volumes and Principle Shape Quadrupoles for HIV PR Ligands.

Ligand	V^g (Å^3)	V^{hs} (Å^3)	$Q1$ (Å^2)	$Q2$ (Å^2)	$Q3$ (Å^2)
4phv	507.5	507.8	22.9	8.4	1.8
7hvp	708.3	706.0	46.6	5.9	2.8
8hvp	714.8	714.5	54.7	5.9	3.3
9hvp	608.4	609.9	36.9	5.7	3.3
1hps	531.1	530.0	29.8	6.7	2.2
1hvr	504.1	504.1	21.5	6.6	2.4

TABLE VI.
Volumes and Principle Shape Quadrupoles for Thermolysin Ligands.

Ligand	V^g (Å^3)	V^{hs} (Å^3)	$Q1$ (Å^2)	$Q2$ (Å^2)	$Q3$ (Å^2)
1tmn	393.9	394.5	13.4	8.6	2.6
3tmn	247.7	248.3	10.6	3.1	1.8
4tmn	415.1	415.0	20.4	6.0	2.4
1thl	386.5	387.4	11.7	8.7	2.6
1tlp	412.9	412.5	13.7	7.3	2.5

there is little variation in components $Q2$ and $Q3$, the component $Q1$ takes a considerable range of values ($\approx$ 20–55 Å^2). This is consistent with the tunnel-like binding pocket of HIV PR being able to bind peptides and peptidelike molecules of various lengths. These results suggest that elements of the shape quadrupole (which are trivial to compute) are capable of characterizing the shape features of ligands. Clearly, one possible application of the shape quadrupole (and higher moments) is as a tool for searching 3-D data bases to find molecules with similar molecular shape.

A comparison between the observed-and predicted relative orientations for different pairs of ligands binding to the chosen proteins is given in Tables VII–IX. These tables report the rms difference between the observed and predicted orientation of the ligand given in the column, relative to the ligand given in the row. The observed orientation refers to the ligand overlap found for two ligands *A* and *B* complexed to a common protein that has been backbone overlayed as explained previously. The predicted orientation, on the other hand, corresponds to the maximal overlap of the ligands *A* and *B* using the Gaussian method. As a check we carried out optimizations starting from the experimental relative orientation, and usually these converged to the best of the minima found from the four starting points described in the Methods section. However, this was not true for the ligand pair 4tmn/1tmn, but in this case the maximal overlap volume was identified using a simple Monte Carlo type search procedure. The simple shape based alignment procedure predicts many of the relative binding modes to an accuracy with an rms < 1.0 Å. There are instances in which the method is not so accurate. For example, in predicting the relative orientation of 1dwe and 1dwc, the shape alignment maximizes the volume intersection by superimposing the proline ring of 1dwe onto the piperidine ring fragment of 1dwc; however, this is not observed experimentally. The worst prediction is for the ligand pair 4tmn/3tmn in which the rms error is 5.7 Å. In this example, a tryptophan ring is present in the 3tmn ligand but

TABLE VII.
rms Comparison of Pairs of Overlapped Thrombin Ligands (See Text).

Ligand	1dwc	1dwd	1dwe	1ett
1dwc	0.00	1.42	1.61	0.57
1dwd	—	0.00	0.33	1.28
1dwe	—	—	0.00	1.45
1ett	—	—	—	0.00

TABLE VIII.
rms Comparison of Pairs of Overlapped HIV PR Ligands (See Text).

Ligand	4phv	7hvp	8hvp	9hvp	1hps	1hvr
4phv	0.00	0.86	0.52	0.36	0.31	0.41
7hvp	—	0.00	0.17	0.37	0.71	1.80
8hvp	—	—	0.00	0.35	0.24	1.40
9hvp	—	—	—	0.00	0.55	0.88
1hps	—	—	—	—	0.00	0.60
1hvr	—	—	—	—	—	0.00

does not overlap with any part of the 4tmn ligand in the experimentally observed relative orientation. The predicted relative orientation involves the largish tryptophan ring structure intersecting with some part of the 4tmn structure. In this example, maximizing volume overlap leads to a very poor prediction of relative binding mode. However, in general the predictions are reasonably accurate and identify the common binding modes of pairs of ligands even when there are no obvious extensive atom–atom correspondences. The shape similarity index S^{AB} of the ligands as defined in eq. (20) is given in Tables X–XII for the various ligand pairs studied. The ligands for this study were picked because they are markedly dissimilar in a chemical sense and the S^{AB} coefficients reported in Tables X–XII reveal this. These examples could not be overlayed using matching based on individual nuclei represented by single points; one requires instead a proper representation of the shape. The main conclusion from our results is that Gaussian shape matching provides a robust predictive tool for ligand binding even for such dissimilar ligands. The molecular graphics representations of overlayed ligands show very clearly how functional groups in dissimilar molecules align in the binding pocket.

The computation of intersection volumes between dissimilar molecules gives rise to the interesting possibility of defining an index of chirality. There has been much interest recently in finding quantitative measures of chirality[6,35–42] as a way of assessing the influence of chirality on ordered bulk phases (liquid crystals), the design of drugs, and asymmetric synthesis, for example. The present work merely involves the overlapping, by maximization of the intersection volume, of the left- and right-hand images of a target chiral molecule. This is essentially the suggestion of Gilat[35] that was implemented using a numerical hard-sphere treatment[36] and used to investigate the potency of chiral drugs.[39,40] The quantity $1 - S^{AB}$ [eq. (20)] was taken as a chirality index for the system. This is trivially shown to be equivalent to considering a normalized difference in intersection volumes between the enantiomeric pair and a nonenantiomeric pair. This procedure has the merit of simplicity as well as efficiency. Other techniques such as the Haussdorff index[37,38] or least squares assess the chirality using pointwise representations of atoms, i.e., they do not fully assess the extent of the molecule. The Haussdorf approach relies on a set theoretic definition of similarity that gives rise to an awkward piecewise continuous function that is difficult to optimize. A similar idea can be used to assess fuzzy symmetry as a development of the ideas of Zabrodsky and Avnir.[41] For a given nonsymmetric molecule, these authors define the nearest geometry with a given exact point group symmetry. This is done using a geometric folding/unfolding procedure. The Gaussian overlay method could then be used to

TABLE IX.
rms Comparison of Pairs of Overlapped Thermolysin Ligands (See Text).

Ligand	1tmn	3tmn	4tmn	1thl	1tlp
1tmn	0.00	0.73	0.60	0.32	0.91
3tmn	—	0.00	5.70	0.33	0.58
4tmn	—	—	0.00	0.45	0.77
1thl	—	—	—	0.00	0.61
1tlp	—	—	—	—	0.00

TABLE X.
Shape Similarity S^{AB} for Pairs of Thrombin Ligands.

Ligand	1dwc	1dwd	1dwe	1ett
1dwc	1.00	0.82	0.69	0.85
1dwd	—	1.00	0.78	0.79
1dwe	—	—	1.00	0.70
1ett	—	—	—	1.00

TABLE XI.
Shape Similarity S^{AB} for Pairs of HIV PR Ligands.

Ligand	4phv	7hvp	8hvp	9hvp	1hps	1hvr
4phv	1.00	0.67	0.64	0.68	0.78	0.78
7hvp	—	1.00	0.83	0.76	0.72	0.68
8hvp	—	—	1.00	0.71	0.61	0.59
9hvp	—	—	—	1.00	0.73	0.61
1hps	—	—	—	—	1.00	0.72
1hvr	—	—	—	—	—	1.00

TABLE XII.
Shape Similarity S^{AB} for Pairs of Thermolysin Ligands.

Ligand	1tmn	3tmn	4tmn	1thl	1tlp
1tmn	1.00	0.72	0.69	0.91	0.88
3tmn	—	1.00	0.60	0.79	0.74
4tmn	—	—	1.00	0.67	0.67
1thl	—	—	—	1.00	0.87
1tlp	—	—	—	—	1.00

give an assessment of symmetry using the index S^{AB} based on the two geometries.

To demonstrate further the utility of the Gaussian shape comparison method, we computed the chiral index $(1 - S^{AB})$ of the biphenyl system, which has an axis of chirality when the dihedral angle between the aromatic rings differs from 0° or 90°. A graph of the chirality index as a function of dihedral angle is given in Figure 4. It can be seen that the chirality index has a maximum at 45° and passes through zero at 0°, 90°, and 180°. This differs from the observation of Osipov et al.[42] who identify a maximal chirality at a twist angle of approximately 30°, although their method treated the molecule as comprising points in space and did not account for the extent of the molecular shape. Unlike the pseudoscalar Osipov index, the shape chirality index is symmetric about the dihedral angle of 90°. This illustrates a potential drawback of a purely scalar chirality index, which cannot be Boltzmann averaged to zero over a range of conformations, as evidently should be the case for biphenyl. We also computed the chiral index for all of the amino acids (blocked with acetyl and *N*-methyl at the N and C terminii, respectively). For the extended structures ($\phi = \psi = 180°$) there is relatively little variation in the chiral index. This is because the maximal volume intersection is achieved by superimposing the enantiomers such that the amino acid side chains approximately overlap with each other, and the peptide backbone of one enantiomer overlaps with the backbone of the other enantiomer, but running in the reverse direction. On the basis of shape this is a reasonable overlap, although one consequence is that the carbonyl group of one backbone chain overlaps with the *NH* moiety of the other chain (and vice versa). For these structures threonine (Thr) had a much larger chiral index (0.11) than serine (0.03) or valine (0.04). This is consistent with the presence of a chiral atom in the Thr side chain. To investigate the dependence of the index on the backbone conformation, we computed it for different backbone dihedral angles chosen to represent local minima observed in peptides and proteins. The results suggest that the most chiral conformers are in the bridge ($\phi = -110$, $\psi = 10.0°$) and helical regions ($\phi = -74$, $\psi = -45°$), whereas the least chiral are in the extended, C5 ($\phi = -150$, $\psi = 150°$) and β ($\phi = -140$, $\psi = 135°$) regions. Given the computational efficiency of the method, it was also possible to compute the chiral index for actual helices. For example we observed that the chiral index converged to a value of ≈ 0.25 for polyalanine helices of various lengths (up to 25 residues), whereas the value of the helical segment of interleukin-4 used in Table III gave a value of 0.35.

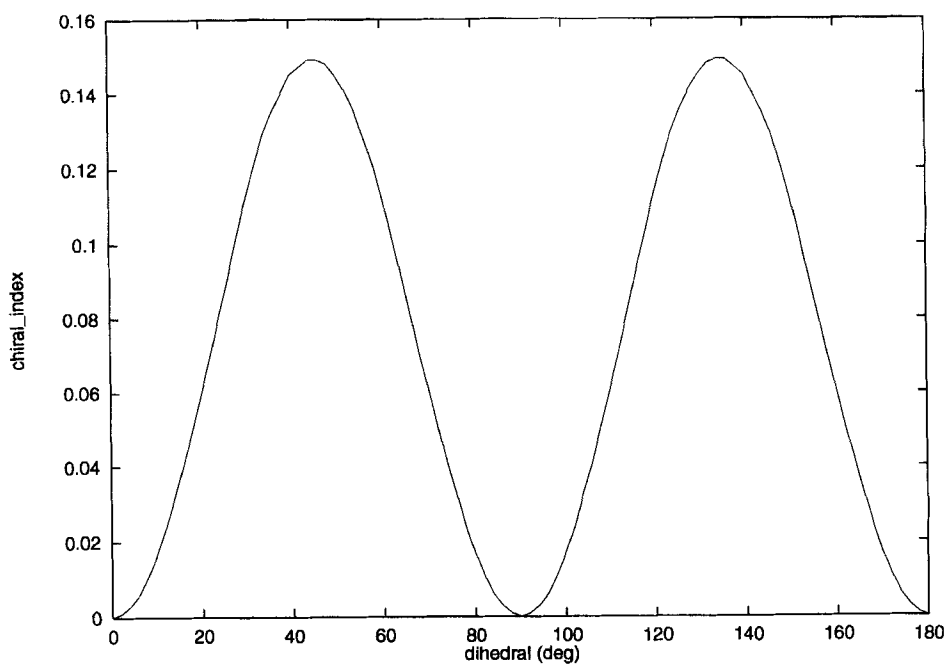


FIGURE 4. Biphenyl chirality diagram.

Conclusions

The present work originated in a scheme that was directed solely toward rapid and accurate calculations of molecular volumes, areas, and their nuclear positional derivatives.[7] This scheme used atom-based Gaussian functions, and introduced a generalized coalescence thereom to compute Gaussian intersections of arbitrary order. Adopting a Gaussian formulation of the inclusion–exclusion formula that is parallel to that used to compute hard-sphere measures then gives values of molecular volumes and areas, which are accurate when compared to hard-sphere volumes and areas to around 1% for small and large molecules. In the present contribution we established that this Gaussian technology can be extended to include the computation of intersection volumes between molecules. An optimization procedure to maximize intersection volumes can then be used to match shapes of dissimilar molecules. We tested our shape matching technique to show that the relative orientation of ligands binding to protein hosts can be predicted on the basis of their volume overlay. The shape matching can also be used to quantify a degree of chirality of asymmetric molecules by overlaying the left- and right-hand forms. A third important idea introduced in this article is that of shape multipoles. The shape multipoles are averages of products of Cartesian coordinates over the Gaussian shape density. The use of the shape centroid and the principal axes of the shape quadrupole provides useful starting frames for shape matching in our new approach. The principal shape-quadrupole moments are indicators of molecular shape that may be useful in QSAR studies of molecular activity. All quantities discussed in this and our previous work[7] can be computed analytically, rapidly, and accurately.

In this work the ligand conformation was taken from the crystallographic experiment. In the general ligand design process, the ligand conformation will be unknown. The efficiency of the Gaussian methodology means that it is feasible to overlay many conformers generated by a search procedure of a ligand (for which an experimental structure is not available) onto either the known structure of another ligand or onto a very rigid ligand. The Gaussian intersection volumes can then be used to assess the suitability of a flexible ligand to bind at a receptor site. There are no difficulties in extending Gaussian shape-matching techniques to include multiple superpositions and to include chemical differences between the atoms in different environments (for example hydrophobic, polar, hydrogen-bond donors/acceptors). This can be achieved by *coloring* atoms, i.e., by computing contributions to intersections that arise from matching atom classes, such as those described.

The importance of shape matching in drug and material design is well understood. It arises because the intermolecular interactions that stabilize the receptor–ligand complex are enthalpically weak and only become effective if the chemical groups involved can approach each other closely, which is favored by shape complementarity. Entropic contributions advantageous to binding involve the loss of complexation water of both the host and the guest and are also favored by shape complementarity (to avoid empty space being filled with water[43,44]). Additional disadvantageous entropic terms involve loss of degrees of freedom. On the whole there is a broad balance between these competing effects and the atoms in the complex tend to be in van der Waals contact. This gives some meaning to the rather jaded lock-and-key concept of binding. Shape matching methods that rely on hard spheres suffer a number of serious defects. These include complicated analytical expressions and gradient discontinuities that lead to slow computation involving complicated algorithms. The Gaussian shape technique obviates all of these disadvantages and offers a more physically realistic description than the traditional hard-sphere model of molecular shape.

Acknowledgments

We thank Dr. Jaroslaw Kostrowicki for very valuable discussions concerning the use of Gaussian functions to represent the unit step function. We also thank Brian Masek, Dave Timms, and Tony Wilkinson for useful discussions concerning molecular shape comparison.

References

1. B. B. Masek, A. Merchant, and J. B. Matthew, *J. Med. Chem.*, **36**, 1230 (1993).
2. M. L. Connolly, *J. Am. Chem. Soc.*, **107**, 1118 (1985).
3. A. Gavezotti, *J. Am. Chem. Soc.*, **105** (1983).
4. K. D. Gibson and H. A. Scheraga, *Mol. Phys.*, **62**, 1247 (1987).
5. F. M. Richards, *Ann. Rev. Biophys. Bioeng.*, **6**, 151 (1977).

6. P. G. Mezey, *Shape in Chemistry*, VCH, New York, 1993.
7. J. A. Grant and B. T. Pickup, *J. Phys. Chem.*, **99**, 3503 (1995).
8. I. Shavitt, *Methods Comp. Phys.*, **2**, 1 (1962).
9. S. F. Boys, *Proc. R. Soc.*, **A200**, 542 (1950).
10. R. Diamond, *Acta Crystallogr.*, **A27**, 436 (1971).
11. G. R. Marshall and C. D. Barry, *Abstr. Am. Crystallogr. Assoc. Honolulu, HI* (1979).
12. G. R. Marshall, C. D. Barry, H. E. Bosshard, R. A. Dammkoehler, and D. A. Dunn, *ACS Symp. Ser.*, **112**, 205 (1979).
13. S. F. Boys and I. Shavitt, *Proc. R. Soc.*, **A254**, 487 (1960).
14. B. J. Cherayil, J. Kostrowicki, L. Piela, and H. A. Scheraga, *J. Phys. Chem.*, **95**, 4113 (1991).
15. J. Kostrowicki and H. A. Scheraga, *J. Phys. Chem.*, **96**, 7442 (1992).
16. E. E. Hodgkin, A. C. Good, and W. G. Richards, *J. Chem. Inf. Comput. Sci.*, **32**, 188 (1992).
17. P. M. Dean, in *Molecular Similarity in Drug Design*, Chapter 1, Wiley, London, 1995.
18. A. C. Good, in *Molecular Similarity in Drug Design*, Chapter 2, Wiley, London, 1995.
19. G. Klebe, *3D QSAR in Drug Design: Theory, Methods and Applications*, ES-COM, Leiden, 1993.
20. S. K. Kearsley, *Tetrahedron, Comput. Methodol.*, **3**, 615 (1990).
21. A. C. Good and W. G. Richards, *J. Chem. Inf. Comput. Sci.*, **33**, 112 (1993).
22. M. Petitjean, *J. Comput. Chem.*, **15**, 507 (1994).
23. B. T. Pickup and N. Powell, Sheffield University Third Year Project Report, unpublished.
24. J. D. Bunce, R. D. Cramer, III, D. E. Patterson, *J. Am. Chem. Soc.*, **110**, 5959 (1988).
25. G. Klebe, U. Abraham, and T. Mietzner, *J. Med. Chem.*, **37**, 4130 (1994).
26. W. R. Hamilton, *Elements of Quaternions*, Chelsea, New York, 1899.
27. B. R. Markey, A. O. Griewank, and D. J. Evans, *J. Chem. Phys.*, **71**, 3449 (1979).
28. S. K. Kearsley, *J. Comput. Chem.*, **11**, 1187 (1990).
29. J. A. Grant and B. T. Pickup, in preparation.
30. E. E. Hodgkin and W. G. Richards, *Int. J. Quantum Chem. Biol. Symp.*, **14**, 105 (1987).
31. E. E. Hodgkin and W. G. Richards, *Chem. Br.*, 1141 (1988).
32. M. L. Connolly, *J. Appl. Crystallogr.*, **16**, 548 (1983).
33. W. J. Cook, B. G. Zahao, R. P. Cameron, S. E. Ealick, R. L. Walter, P. Reichert, T. C. Nagabahushan, P. P. Trotta, C. E. Bugg, and M. R. Walter, *J. Biol. Chem.*, **267**, 20371 (1992).
34. F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, *J. Mol. Biol.*, **112**, 535 (1977).
35. G. Gilat, *J. Phys. A: Math. Gen.*, **22**, L545 (1989).
36. A. M. Meyer and W. G. Richards, *J. Comput.-Aided Mol. Design*, **5**, 427 (1991).
37. A. B. Buda, T. Auf der Heyde, and K. Mislow, *Angew. Chem. Int. Ed. Engl.*, **31**, 989 (1992).
38. N. Weinberg and K. Mislow, *J. Math. Chem.*, **14**, 427 (1993).
39. A. Seri-Levey and W. G. Richards, *Tetrahedron Assym.*, **4**, 1917 (1993).
40. A. Seri-Levy, S. West, and W. G. Richards, *J. Med. Chem.*, **37**, 1727 (1994).
41. H. Zabrodsky and D. Avnir, *J. Am. Chem. Soc.*, **117**, 462 (1995).
42. M. A. Osipov, B. T. Pickup, and D. A. Dunmur, *Mol. Phys.*, **84**, 1193 (1995).
43. K. A. Sharp, A. Nicholls, R. M. Fine, and B. Honig, *Science*, **252**, 106 (1991).
44. K. A. Sharp, A. Nicholls, R. Friedman, and B. Honig, *Biochemistry*, **30**, 9686 (1991).