Three-Dimensional Pattern Recognition from Molecular Distance Minimization[†]

Michel Petitjean[‡]

Institut de Topologie et de Dynamique des Systemes, (CNRS, URA34), University Paris 7,
1 rue Guy de la Brosse, 75005 Paris, France

Received April 6, 1996[⊗]

A procedure optimizing the alignment of two three-dimensional molecules is presented. It uses atomic numbers, positions and radii, and optionally computed atomic charges. Neither bonds nor connectivities are required. The dissimilarity between two molecules is measured with a molecular distance, which is minimized under all rotations and translations with a Newton-like algorithm. This molecular distance is the usual norm of a two-components vector. One of the components is the electronic molecular distance, and the other is the protonic molecular distance. Both the electronic and the protonic component are computed with the same algorithm, which assumes a homogeneous spheres model. The resulting minimized norm is shown to be an intrinsic molecular distance. When a family of more than two compounds is involved, the intrinsic molecular distances matrix of the family is built. Various applications are presented, including comparisons of X-ray crystallographic data compounds, maximal common 3D-substructure searching, comparisons of geometry and charge calculations, and quantitative chirality measurement. The optimal alignments are more easily obtained when large atomic radii are selected. The charge calculation algorithm have only a little influence on the results.

INTRODUCTION

Molecular similarity procedures are needed for various purposes, as structural alignment of proteins,[1−13] chemical reactivity studies,[14−20] three-dimensional molecular template recognition,[21,22] 3D-structure−activity relationships,[23−25] pharmacophore identification, and drug design.[26−30] There are two types of molecular similarity: the shape similarity and the electronic similarity. Most shape similarity procedures are based upon interatomic distances.[31−39] Even when the two molecules have the same number of atoms, a rigorous treatment of the distances matrix leads to heavy combinatorial algorithms. Thus, efforts have been provided to find correspondences between subsets of atoms.[40−43] When the pairwise correspondence between two sets of same cardinality is known, the optimal superposition problem is analytically solvable with the Procuste algorithm.[44,45] All these geometric algorithms assumed that a molecule is a set of points. A homogeneous full body model has also been proposed.[44] Very few chemical or biological situations involve only shape similarity. Most situations involve electronic interactions. Thus, various similarity procedures based upon electronic properties have been proposed. They differ by the similarity criterion to be optimized. The use of density functions was proposed by Carbo et al.,[46,47] with the following index: $R = \langle f_1|f_2\rangle/(\langle f_1|f_1\rangle * \langle f_2|f_2\rangle)^{1/2}$, $\langle f_1|f_2\rangle$ being the functional scalar product of the density functions (i.e., the volumic integral of $f_1 * f_2$ extended over the whole space). The Carbo index was computed by Good et al.[48,49] using Gaussian approximations of the densities. The Carbo index, formally analogous to the cosine of two vectors, is not sensitive to the magnitude (e.g., f_1, $-f_1$, and $2*f_1$ are not distinguished). Hodgkin and Richards[50] proposed to use instead $H = 2*\langle f_1|f_2\rangle/(\langle f_1|f_1|\rangle + \langle f_2|f_2\rangle)$, f_1 and f_2 being either the electrostatic potential of the molecules or their electrostatic field. The functional norm $||f_1-f_2|| = \langle f_1-f_2|f_1-f_2\rangle^{1/2}$ is magnitude sensitive and was used by Carbo, Leyda, and Arnau[46] with density functions, but they did not retain it because they prefer an index weakly depending on molecular dimensions. Once f_1 and f_2 were defined, any of these electronic similarity criterions leads to the same alignment when $\langle f_1|f_2\rangle$ is maximized under all rotations and translations, because $\langle f_1|f_1\rangle$ and $\langle f_2|f_2\rangle$ are invariants. These criterions have their associated geometric equivalent, which are formally obtained when f_1 and f_2 are the indicator functions of the two molecules ($f_1 = 1$ inside the molecule 1 and $f_1 = 0$ elsewhere): $\langle f_1|f_1\rangle = \mathrm{V1}$, $\langle f_2|f_2\rangle = \mathrm{V2}$ and $\langle f_1|f_2\rangle = \mathrm{V12}$, V1, V2, and V12 being the respective volumes of the molecules 1 and 2 and their intersection. Thus, $\mathrm{V12}/(\mathrm{V1}*\mathrm{V2})^{1/2}$ was used by Hermann and Herron,[27] and its discrete approximation was used by Meyer and Richards,[51] who used also a discrete approximation of $2*\mathrm{V12}/(\mathrm{V1}+\mathrm{V2})$. The geometric equivalent of the functional norm $||f_1-f_2||$ was also used:[44] $(\mathrm{V1}+\mathrm{V2}\text{-}2*\mathrm{V12})^{1/2}$. The SEAL algorithm introduced by Kearsley and Smith[52] involve both electronic and steric aspects. It minimizes the summation over all atom-pairs (i,j) of $(E*Q_\mathrm{i}*Q_\mathrm{j} + S*V_\mathrm{i}*V_j)*\exp(-A*R_{ij})$, Q_i and Q_j being the partial atomic charges, V_i and V_j being the atomic volumes, and E, S, and A being constants to be empirically set by the user. Some similarity procedures perform an optimal alignment using a first criterion and then compute the similarity index with an other criterion[48,49,53] or limit the alignment search to rotations only.[54]

Thus, not all methods indeed perform a full search of the six-tuple of alignment parameters (three translational and three rotational), and it was recognized that this full search was in 1985 beyond the capacity of the fastest computers.[55] The main reason comes from the need to get sufficiently

[†] Key words: molecular similarity; optimal 3D-molecular alignments; maximal common 3D-substructure; intrinsic molecular distance; sorted distances matrix algorithm; chirality index.

[‡] E-mail: petitjean@itodys.jussieu.fr. Phone: +33 1 44 27 48 57; FAX +33 1 44 27 68 14.

[⊗] Abstract published in *Advance ACS Abstracts,* September 1, 1996.

S0095-2338(96)00370-8 CCC: $12.00

accurate values of the similarity measure to ensure a good convergence of the minimization algorithm: one accurate computation of a density function or an electrostatic potential is time-consuming, then the numeric evaluation of the volumic integral requires a large number of functions values, and the volumic integral has to be computed many times. The need of a fast algorithm involving both steric and electronic aspects without empirical parameter settings (except, of course, empirical parameters set to compute input data) has motivated the development of the method described hereafter. The dissimilarity criterion is a distance. As a consequence, when two compounds 1 and 2 are found to be closed from a reference compound 3, the triangle inequality permits one to know that 1 is indeed closed to 2. It is a major defect of many procedures that the similarity criterion bears no proximity information.

THEORY

The electronic similarity between two molecules 1 and 2 is basically related to the comparison of their density functions, or their square root, which are dimensionally equivalent to wave functions. Let f_1 and f_2 be the respective square roots of the densities function of 1 and 2. The simplest molecular distance criterion is the functional distance $D(1,2) = \langle f_1 - f_2 | f_1 - f_2 \rangle$. As shown previously,[44] the minimized distance between compounds also has the mathematical properties of a distance over the space of the compounds and is called[44] an intrinsic molecular distance, denoted ID(1,2). The term "intrinsic" means that the distance between two compounds is independent from their two respective coordinates systems.

The volumic integral of a density always exists. This ensures the existence of D and ID. This existence is not established for the electrostatic potential or field or for the density itself. The shape concept can be viewed as a consequence of the radial decreasing of the electron density: outside a limiting surface, the density is forced to be null. Atomic radii can be derived from this concept.[56] It means that the comparison between density functions limited to the union of atomic spheres handles both electronic and shape similarity aspects. The practical output of many quantum chemistry softwares is not an accurate density but rather the set of atomic charges located at the centers of the atomic spheres. The electrostatic potentials and fields are computed from this discrete tridimensional set of dot charges. Outside the union of the atomic spheres, a homogeneous distribution of each atomic charge over its associated atomic sphere would lead to the same potential and field. This simplified model is much easier to handle than the full density model and avoids the major drawback of the dot charge model: the ID of two identical dot charge sets located on two spatial positions sets is not a continuous function of the atomic positions. E.g., two dipoles ($-$e;$+$e) have not their ID tending to zero when their intercharge distance tends to equality. Moreover, the homogeneous sphere approximation is a more realistic model than the dot charge approximation.

Another important feature of the model is the distinction between negative and positive charge, despite the known additivity of charge effects over the molecular potential. Let us consider two monoatomic molecules, such as neon and argon. In the additive model, the partial atomic charges are null, and their ID is null, although the sizes of the molecules are different. The ID between argon and the anywhere zero density is also null. It means that molecules differing only by null partial charge atoms cannot be distinguished: the size and the shape concepts are altered. The null partial charge atoms here are called "transparent", because they can be added or removed from a molecule without modifying its intrinsic distance to another molecule.

This transparence is not related to the homogeneous spheres model but is a consequence of the use of partial atomic charges in the additive model. Moreover, when the computed partial charges are close to zero, as for long alkyl chains, an important part of the molecule becomes quasi-transparent. Now, assume a nonadditive model, such that each atom bears its complete electronic charge (i.e., the partial charge plus the atomic number). Thus there are two charge distributions in a molecule: the electronic charge distribution (containing all the negative charges) and the protonic charge distribution (containing all the positive charges). That leads to a two-component vector model: one component is the electronic molecular distance, the other is the protonic molecular distance. Thus, the length of this vector distance is also a molecular distance (see appendix 1) which therefore induces[44] an intrinsic molecular distance. If we intend to compare only the electronic distribution, the protonic component is set to zero, but this does not reflect properly some situations: e.g., two isoelectronic ions 1 and 2, as Cl^- and K^+, would have $ID^2\ (1,2) = 2{*}N{*}e{*}(1 - (R_1/R_2)^{3/2})$, $N{*}e$ being the common electronic charge, and R_1,R_2 the respective atomic radii. Because R_1 is close to R_2, the electronic component of ID(1,2) is close to zero. Assuming $R_1 = R_2$, the full vector model leads to $ID(1,2) = 2{*}e^{1/2}/((N{+}1)^{1/2} + (N{-}1)^{1/2})$. This distance decreases when the atomic number increases: it means that there is more similarity between K^+ and Cl^- than between Na^+ and F^-. This is coherent with the fact that the same charge difference is smaller when it is referred to a big ion. The major drawback of the homogeneous spheres vector model is its physical sense: although the homogeneous electronic charge distribution in the atomic sphere may be viewed as a rough approximation of the density, the homogeneous protonic charge is unrealistically distributed. Nevertheless, this model is preferred to the dot charge model, because the dot charge model leads to a discontinuous ID (see previously the dipole example), and using molecular potentials or fields does not avoid such singularities.

Despite the fact that partial charges are more important than total charges in terms of chemical properties, a major interest of the vector model comes from working with better precision over the atomic charges, and thus better precision over similarity measures, as shown as follows: the partial atomic charge Q_i of an atom i is known with a relative precision dQ_i/Q_i. dQ_i often has the same magnitude than Q_i, and sometimes, even the sign of Q_i is not sure. It means that similarity calculations using the partial charges has a poor precision. N_i being the atomic number, this precision becomes $dQ_i/(N_i + Q_i)$, with $N_i \gg Q_i$. Thus, some robustess is expected, even when $dQ_i > Q_i$. The limit situation is reached when no charge calculation is performed: the approximation $Q_i = 0$ is expected to be acceptable because $Q_i \ll N_i$. In this limit situation, the electronic and the protonic components are equal. In the general situation, they are computed with the same routine: only the atomic charge is different.

The density associated with either of these two components is computed as the sum of the individual atomic densities. The atomic density of any i atom can be written as $1(i)*w_i$, where w_i is the ratio of the atomic charge to the atomic volume, and $1(i)$ is the indicator function of i. The volumic integrals $\langle f_1|f_1\rangle$ and $\langle f_2|f_2\rangle$ and all the w_{i1} and w_{i2} coefficients are invariant under rotations and translations. For a particular rotation and translation, the product of the square roots of the overall densities is a piecewise constant function of the three spatial coordinates, and its volumic integral computation requires to partition the space in pieces so that all the indicator functions are constant in a piece, each piece being a spheres intersection. Obtaining the list and the volume of any intersection of spheres is a problem analytically solvable.[57] Although the number of sphere intersections is small for a single molecule, it becomes very high when two molecules are superimposed,[44] and the computing times required to perform alignments precludes any practical use. Thus, we keep the bicomponent vector model and the homogeneous spheres assumption, but we replace for each molecule the square of the overall density by the summation of the squares of the individual atomic densities. Let g be this function. It must be pointed out that q is specific of the molecule as is the square of the density, (or as would have been the potential or the field). Both are dimensionally equivalent to a wave function, and the two intrinsic molecular distances are equivalent (see Appendix 2). It means that, when two molecules are similar for one of the intrinsic distances, they are also similar for the other. When a molecule is monoatomic, $f = g$.

The summation symbols over i_1 and i_2 in the $\langle g_1|g_2\rangle$ expression are now writable out of the volumic integral: $\langle g_1|g_2\rangle$ is the double sum over i_1 and i_2 of $(w_{i1}*w_{i2})^{1/2}*V(i_1,i_2)$, $V(i_1,i_2)$ being the volume of the intersection of atomic spheres i_1 and i_2. This leads to practical computing times roughly divided by 10^6. A part of this ratio is due not only to the number of intersections to compute but also to a drastic simplification of the program. Some applications are presented hereafter. The minimization algorithm is a Newton-like method: the function, its gradient, and hessian are all analytically computed. The program is written in FORTRAN and runs on DEC AlphaStation 2100, IBM RS6000, and CRAY Y-MP-EL. Performing a minimization requires usually some seconds for molecules with up to a hundred atoms, and some tenths of minutes for proteins with about a thousand atoms. These times are several times larger on the CRAY Y-MP-EL, due to poor vectorization. All computing times given hereafter refer to the AS2100 computer. The absolute precision on the squared intrinsic distance is better than 10^{-6} electron charge unit.

APPLICATION TO CRYSTALLOGRAPHIC DATA

The anthramycin methyl ether monohydrate ($C_{17}H_{19}N_3O_4,H_2O$) is an antibiotic derivative with antitumor activity[58,59] having three entries in the Cambridge Crystallographic Database:[60] ANTMYC, ANTMYCO1, and ANTMYCO3. No coordinates were stored for ANTMYC. ANTMYCO1 ($C_{17}H_{20}N_3O_5$) has 45 3-tuples of coordinates, and ANTMYCO3 ($C_{17}N_3O_5$) has only 25 3-tuples of coordinates, because it is hydrogen depleted. They are displayed in Figure 1. Except for this graphical display, the associated connectivity information is never read. The atom labeling of the two entries are different, and no correspondence is read.

The geometries of these two entries are submitted to the similarity procedure described in the theory section. The atomic radii were taken from Gavezzotti's paper.[61] The computed atomic charges were first set to zero (the effects of charge calculations are examined further), and five minimizations were performed. The overall computing time was 13 s. The global minimum was reached three times, and the squared minimized distance expressed in electron charge unit was $\mathrm{ID}^2 = \langle g_1 - g_2|g_1 - g_2\rangle = 97.372$. This value is small relative to the squared maximized distance: $\langle g_1|g_1\rangle + \langle g_2|g_2\rangle = 1590.694$. Their ratio, called here the squared distance index, is equal to 0.061. The pertinence of the optimal alignment was not evaluated graphically, because it requires reading the connectivity informations, and we assume to have as few chemical data as possible. It was evaluated as follows. The 45*25 interatomic distances array between the 45 atoms of ANTMYCO1 and the 25 atoms of ANTMYCO3 was sorted by increasing values. It was observed that the 24 lowest interatomic distances, expressed in angstroms, vary smoothly from 0.006 to 0.082, the next jumps to 0.627, and the remaining vary smoothly from 0.831 to 14.494. Outputing the sequential numbering of the coupled atoms showed that the 24 best correspond to a pairwise correspondence between all the non-hydrogen atoms, except one oxygen in each molecule. In fact, this oxygen, which was labeled O5 in both entries, pertained to the water molecule of the hydrate, and it can be verified in Figure 1 that, compared to the anthramycin, their relative location is completely different. The pairwise correspondence was then input to the Procuste similarity algorithm,[44] and the minimized quadratic mean distance between the 24-tuples of atoms was equal to 0.032 Å, with a maximum of 0.063 Å: the optimal alignments of the 24-tuples of atoms with the two similarity procedures are very close.

Although ANTMYCO1 and ANTMYCO3 indeed have a common substructural fragment containing 24 atoms, there is some interest to see how the procedure works when the same molecule is matched with itself. Five minimizations were performed to align ANTMYCO1 with itself, and five were performed to match ANTMYCO3 with itself. The perfect alignment was indeed reached respectively two and four times, and the other minimizations has lead to local minima. The squared intrinsic distances were both less than $0.5*10^{-6}$ electron charge unit. The average minimization took about 2.5 s.

SELECTION OF ATOMIC RADII

The value of an atomic radius can be derived from various concepts. Several definitions and measures were discussed by Bondi.[62] The selection of values depends on the application:molecular volume computation,[61,63,64] structure–activity studies,[65] or conformational analysis,[66–68] and even for a particular application several values are available. Our purpose is not to discuss the best approach, but rather to see whether this is crucial or not for the similarity procedure. Neither the atomic hybridations nor the connectivity information are read by the procedure. Thus, the radii proposed by Gavezzotti[61] are suitable, because they do not depend on the chemical environment of the atoms: $R(H) = 1.17$, $R(C) = 1.75$, $R(N) = 1.55$, $R(O) = 1.40$, $R(F) = 1.30$, $R(Cl) = 1.77$, $R(Br) = 1.95$, $R(I) = 2.10$. The set can be coherently completed by Bondi's values:[62] $R(P) = 1.80$ and $R(S) = 1.80$. This set is characterized by large values. At its opposite, the set of atomic radii stored[69] in the Cambridge Crystallographic Database for the anthramycin data (see previous section) is characterized by small values: $R(H) = 0.23$ and $R(C) = R(N) = R(O) = 0.68$ (it is not a van der Waals radii set, but it is suitable for numeric trials). The CSD entries ANTMYCO1 and ANTMYCO3 were matched with the set of small radii. The computed atomic charges were set to zero, and ten minimizations were performed. The squared intrinsic distance expressed in electron charge unit was $ID^2 = 92.264$ (the squared maximized distance was 695.479). As for the set of large radii, the correspondence array between the 24 non-hydrogen atoms of the anthramycin moiety was obtained. The pairwise correspondences were identical. The last of the 24 lowest distances was 0.080, and the next lowest was 0.624. Thus, the set of small radii has also permitted a successful optimal alignment. However, nine of the ten minimizations have lead to a local minima, and the optimal alignment was obtained only once. Matching the two molecules with themselves failed within ten minimizations each. It was successful for other molecules (smaller or larger). Other tests (results not reported) have shown that the set of large atomic radii always leads to reach the global minimum more frequently (but the absolute precisions are the same). The large radii set is thus retained.

MAXIMAL COMMON 3D-SUBSTRUCTURE SEARCHING

Once performed with some optimal alignment, it is useful to know which part of the each molecule is indeed similar. This problem is called the maximal common 3D-substructure searching. Manual alignments need graphic facilities and lead to subjective conclusions. Many competitive methods are available,[31−39] mainly based upon distances matrix. Most require empirical parameter settings, as distances range between atom-pairs,[35−37] and are sometimes complex to handle.[41] Thus, it is proposed here to search the maximal common 3D-substructure with the following algorithm, which avoids these drawbacks.

Let n_1 and n_2 be the respective number of atoms of two aligned molecules, the alignment being performed by any method. The n_1*n_2 interatomic distances array between the n_1 atoms of the first molecule and the n_2 atoms of the second molecule is sorted by increasing values. Starting with the first couple of atoms (i.e., those having the smallest interatomic distance), the array is read until a couple of atoms has at least one member previously encountered in the same molecule. The process is stopped without recording this last couple. Let n be the number of couples output by the algorithm. Obviously, n cannot be null and the second occurrence of a member of a couple is ensured after reading at most $\min(n_1,n_2)$ couples: the outputs of the algorithm are the two respective n-membered 3D-substructures of the two molecules. As a byproduct, the two sets of n atoms are pairwise associated, and their alignment can be geometrically refined with the Procuste algorithm.[44]

The pertinence of this algorithm is obvious for the anthramycin CSD entries ANTMYCO1 ($n_1 = 45$) and ANTMYCO3 ($n_2 = 25$) (see results in the two preceding sections): for both sets of atomic radii, the maximal common 3D-substructure is the hydrogen depleted anthramycin moiety ($n = 24$: the last atom of ANTMYCO3 is the oxygen of the hydrate part), and the associated distances distribution show an unambiguous jump between couples 24 and 25. For other molecules alignments, the mathematical conditions ensuring the existence of this jump are not known.

The optimality of the alignment is not needed to get a maximal common 3D-substructure, but it is expected to lead to a pertinent one. The following example uses one of the local minima obtained when an ethanol conformer is matched to itself (randomly rotated and translated) with the geometric algorithm based upon the intrinsic distance[44,70] $ID = V_1 + V_2 - 2*V_{12}$ (see input data in ref 44): the aligned conformer data are in Table 1). The nine first sorted distances were 0.006, 0.007, 0.009, 0.009, 0.009, 0.010, 0.307, 0.307, and 1.060 (next is again 1.060) and were associated to the respective couples (8,6), (6,8), (1,2), (5,7), (7,5), (2,1), (3,4), (4,3), and (9,4). Thus, the maximal common 3D-substructure has $n = 8$ atoms and corresponds, as seen from Table 1, to the alignment of the ethanoates chains: H-CH_2-CH_2-O versus O-CH_2-CH_2-H. The jump in the distance distribution is clearly observed.

The sorted distances matrix algorithm can also be used to evaluate the quality of any manual or automated alignments on a graphic workstation: the user performs rotations and translations, and the computer updates the maximal common 3D-substructure and displays useful values, as n and the distance values. Additional facilities can be provided, as highlighting the n atoms, storing, and recalling alignments, and so on.

COMPARISONS OF CHARGE CALCULATIONS

The electronic molecular similarity between parts of molecules is of course useful to get common 3D-substructures, but it is also useful to compare outputs of quantum chemistry softwares: when both the geometry and the computed atomic charges are different, it is a quantitative measure of the difference between the results. The anthramycin CSD entries ANTMYCO1 and ANTMYCO3 used in the preceding sections were hydrogen completed using the "fillvalence * H" command of the SYBYL software and submitted to all the five charge calculation methods available with the SYBYL software,[71] discarding whether they are adequate to anthramycin or not: Del Re, Gasteiger and Marsili, Gasteiger and Hückel, Hückel, and Pullmann. The two molecules ($C_{17}H_{19}N_3O_4$,H_2O) with null atomic charges are also considered. The squared intrinsic molecular distances between these two sets of six conformers are in Table 2: they range between 86.433 and 92.632 electron charge unit. The squared distance index varies from 0.049 to 0.052, and the maximal common 3D-substructure was always the same set, containing $n = 40$ atoms. Its associated jump in the sorted interatomic distances matrix was from 0.580 to 0.628 Å, depending on the conformers. The six atoms not included in the common 3D-substructure were for all 12 conformers: the three atoms of the water molecule, the hydrogen of the phenol function, the hydrogen of the aromatic secondary amine function, and one hydrogen of the amide group (see Figure 1). The whole molecule (46 atoms) is not the maximal common 3D-substructure, because the

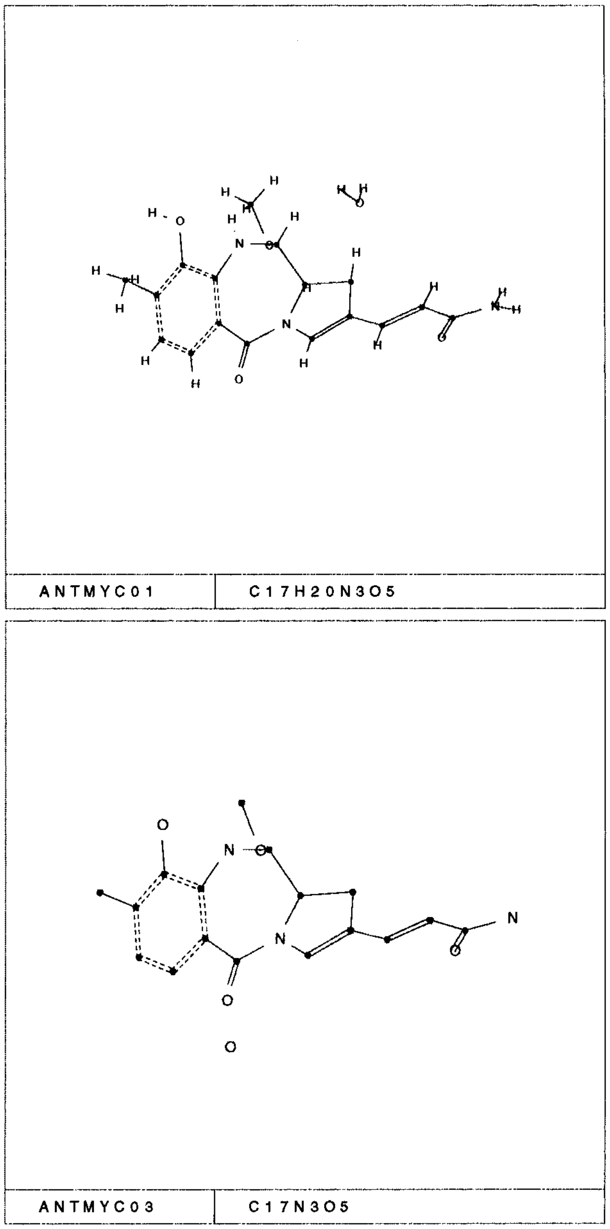

Figure 1. Display of the CSD entries ANTMYCO1 and ANTMYCO3.

water molecule does not have the same relative location, and because the hydrogens added by the fillvalence command to get Ar-OH, Ar-NH, and $CONH_2$ groups could have reasonably more than one possible orientation.

The similarity procedure has the robustess expected in the theory section. It was possible to work without charge calculation, because it has lead to the same maximal common 3D-substructure. Thus, neither quantum chemistry software nor connection table is required. On the other hand, the similarity procedure is sensitive enough to compare the computed charges for different geometries: it is seen from Table 1 that the largest differences observed when the charge calculation method is the same, arises with the Pullmann and the Del Re methods, and that the smallest arise for the Hückel method and when no charge computation is performed. Since the Hückel method applies only to π-systems, it means that the lack of charge information increases the similarity between compounds. When each conformer is matched with itself (i.e., the geometry is the same and the atomic charges differ), the largest differences arise between: Pullmann or Del Re methods, with Hückel or no-charge methods (see Table 3). It shows that the Pullmann and Del Re methods bear more charge information than the other methods, but, of course, no conclusion is drawn for the pertinence of this charge information.

Although sensitive to the charge calculation method, the robustness of the procedure can be now quantitatively evaluated as follows: when the geometry is the same and the charge calculation differs, the highest squared intrinsic distance was 0.274 (see Table 3), but when the charge calculation is the same and the geometry differs slightly, the lowest squared intrinsic distance was about 300 times greater: 86.433 (see Table 1). Thus, when the procedure is used as a 3D-substructure search algorithm rather than an electronic similarity procedure, the charge calculation is not required.

QUANTITATIVE CHIRALITY MEASUREMENT

The need of a quantitative chirality measurement has been stressed,[72] and it was noticed[51,73] that this concept can be derived from the geometric similarity concept. Meyer and Richards[51] propose to measure the similarity of two enantiomers 1 and 2 by maximization of their common volume V_{12} and use of the normalized chiral index $X = 1 - \max(V_{12})/V$, $V = V_1 = V_2$ being the common volume. Zabrodsky and Avnir[72] use the minimized squared quadratic mean distance between paired atoms. In fact, their algorithm, which they named the Pose algorithm, is the Procuste algorithm already cited in this paper.

These similarity measures are particular instances of geometric intrinsic molecular distances. The maximization of V_{12} corresponds to the minimization of the molecular distance[44] $V_1 + V_2 - 2*V_{12}$, which is itself the three-dimensional instance of the distance introduced by Dinghas[70] and is also the square of the usual distance $\langle f_1 - f_2 | f_1 - f_2 \rangle$ between the indicator functions f_1 and f_2 of the molecules. The minimization of the squared quadratic mean distance between paired atoms corresponds to the minimization of the three-dimensional instance of the Procuste distance, which is induced by the norm derived from the scalar product $\langle X_1 | X_2 \rangle = \mathrm{trace}(X_1^t \cdot X_2)$ defined over the $(n,3)$-matrices vector subspace, n being the number of atom pairs.

A generalization can be done as follows. Any molecular distance (electronic or geometric) induces an intrinsic molecular distance and then a distance index, which is the minimized distance between the molecules, normalized to their maximized distance when this maximum is known. If not known, no normalization is performed. The chirality index is defined as the squared distance index between the enantiomers.

E.g., let us consider the distance induced by the norm derived from the usual functional scalar product. When f_1 and f_2 are positive functions, the upper bound of the distance is reached when the molecules have no intersection, and $X^2 = \min(\langle f_1 - f_2 | f_1 - f_2 \rangle)/(\langle f_1 | f_1 \rangle + \langle f_2 | f_2 \rangle)$ induces a chiral index taking values in [0;1]. When f_1 and f_2 are the indicator functions, the chiral index $1 - \max(V_{12})/V$ is retrieved. When f_1 and f_2 take signed values, either $\langle f_1 | f_2 \rangle$ takes at least one positive value and X takes values in [0;1], or $\langle f_1 | f_2 \rangle$ is negative for any location of the molecules. In this situation, the existence of the volumic integrals $\langle f_1 | f_1 \rangle$ and $\langle f_2 | f_2 \rangle$ ensures that $||f_1||*||f_2||$ tends to be zero when one of the molecules

Table 1. Atomic Positions of the Aligned Ethanol Conformers

atom	*x*	*y*	*z*	*x*	*y*	*z*
C	−0.955 564	−0.381 617	−0.001 092	0.341 045	0.421 684	−0.001 601
C	0.334 426	0.428 483	−0.002 272	−0.955 656	−0.377 586	0.006 759
O	1.451 866	−0.462 967	−0.002 642	−1.007 703	−1.205 300	−1.157 517
H	−1.018 444	−1.035 707	−0.901 342	1.227 691	−0.253 665	−0.001 866
H	−1.846 044	0.288 783	−0.003 112	0.410 471	1.076 172	0.897 966
H	−1.019 464	−1.031 687	0.902 008	0.404 163	1.073 192	−0.903 719
H	0.405 456	1.072 593	0.905 138	−1.843 605	0.296 695	−0.007 621
H	0.404 476	1.072 253	−0.910 002	−1.016 015	−1.037 171	0.903 767
H	2.243 296	0.049 863	0.013 318	−1.821 490	−1.682 051	−1.149 133

Table 2. Squared Intrinsic Molecular Distances in Electron Charge Unit between the Two Anthramycin Conformers Derived from Hydrogen Completed CSD Entries ANTMYCO1 (Columns) and ANTMYCO3 (Lines)[a]

	NC	DR	GM	GH	HU	PU
NC	86.433	89.437	88.358	88.420	86.459	89.480
DR	89.839	92.592	91.565	91.616	89.819	92.632
GM	88.630	91.435	90.375	90.426	88.601	91.460
GH	88.701	91.495	90.435	90.484	88.668	91.518
HU	86.498	89.458	88.370	88.428	86.488	89.480
PU	89.856	92.606	91.564	91.613	89.815	92.623

[a] The atomic charge calculation methods are NC = no charges, DR = Del Re, GM = Gasteiger and Marsili, GH = Gasteiger and Hückel, HU = Hückel, and PU = Pullmann.

Table 3. Squared Intrinsic Molecular Distances Matrix in Electron Charge Unit for Two Anthramycin Conformers Derived from Hydrogen Completed CSD Entries ANTMYCO1 (Upper Triangle Matrix) and ANTMYCO3 (Lower Triangle Matrix)[a]

	NC	DR	GM	GH	HU	PU
NC		0.260	0.146	0.161	0.018	0.274
DR	0.244		0.035	0.034	0.233	0.011
GM	0.139	0.033		0.002	0.109	0.031
GH	0.153	0.032	0.002		0.121	0.028
HU	0.018	0.219	0.104	0.114		0.225
PU	0.259	0.011	0.029	0.026	0.212	

[a] The atomic charge calculation methods are NC = no charges, DR = Del Re, GM = Gasteiger and Marsili, GH = Gasteiger and Hückel, HU = Hückel, and PU = Pullmann.

is located infinitely far from the other, and from Cauchy−Schwarz inequality we get that the lower bound of $|\langle f_1|f_2\rangle|$ is zero and that $X = 1$. The distance index X, and therefore the chiral index, is always taking values over [0;1]. When f_1 and f_2 take signed values, the alternate normalization is to the squared maximized distance: $Y^2 = \min(\langle f_1 - f_2|f_1 - f_2\rangle)/\max(\langle f_1 - f_2|f_1 - f_2\rangle)$. The maximized distance is reached when $f_1 = -f_2$, leading to a squared distance index and its derived chiral index upper bounded by 1/4. Exhibiting two positive functions maximizing the distance index X or Y is difficult, and, apart from very few exceptions (as the most achiral triangle for the Hausdorff chirality measure[73]), the most achiral enantiomers couples maximizing X^2 or Y^2 are not known.

The intrinsic distance derived from $\langle X_1|X_2\rangle = \text{trace}(X_1{}^t\cdot X_2)$ is restricted to molecules having the same number of atoms. Since $\langle \mathrm{X}_1 - \mathrm{X}_2|X_1 - X_2\rangle$ is not upper bounded, the induced chiral index is simply the squared intrinsic distance normalized to the number of atom pairs.

The continuity properties of X are those of the intrinsic distance. The need of a continuous chiral index was pointed in the paper of Zabrodsky and Avnir,[72] and the need of a continuous intrinsic distance was mentioned in the theory section of this paper (see the dipole example). For symmetry groups presenting more than two ideal situations, the intrinsic distances matrix should be built. However, an evaluation criteria to select which chiral index is better for some application remains an open problem.

ALIGNMENTS OF HEAVY MOLECULES: PROTEINS AND DNA STRANDS

The computing times needed to align small molecules (less than 100 atoms) are ca. a few seconds. In order to evaluate the computing times needed for large molecules, two protein alignments were made. All atomic positions were obtained from the Protein Data Bank.[74] The first trial was to match the insulin (code 4INS; 102 amino acids residues; 829 non-hydrogen atoms) with a random translated and rotated copy of itself. The squared intrinsic distance was found to be less than $0.5*10^{-6}$, and the average minization has taken about 30 min. The maximal common substructure contained all the 829 atoms.

The second trial was to match the model 1 and 2 of the cytokin (code 1HUN; 138 amino acids residues; 2128 atoms; hydrogens are recorded). The squared intrinsic distance was about 11 704 electron charge units (the squared distance index was about 0.148), and the average minimization has taken about 4 h. The maximal common substructure contained only 1417 atoms, although the 2D-maximal common substructure is the set of all the 2128 atoms. Then, the alignment between the two models of the cytokins was also performed with the Procuste algorithm.[44] The minimized quadratic mean interatomic distance (i.e., between an atom of the first model of cytokine and its associated equivalent of the second model) is 1.587 Å, and the worst interatomic distance is 9.973 Å. The quadratic mean interatomic distance and the worst interatomic distance associated with the ID^2 = 11 704 electrons are respectively 1.589 and 9.989 Å. These values are close to those obtained with the Procuste algorithm. They show that the two models of cytokine cannot be perfectly superimposed, due to the rigidity of the molecular model. The alignment associated with ID^2 = 11 704 electrons is displayed in Figure 2. It shows clearly that the method is effective, and, at the difference from the Procuste algorithm, it produces the pairwise atomic correspondence as an output rather than reading this correspondence as an input.

Alignments of ADN strands were also performed successfully. E.g., aligning two decamers took about 5 min per minimization. The graphical displays have shown that the backbones were primarily aligned, rather than the monomer units.[75]

Table 4. STX/TTX: Distribution of the 10, 100, and 1000 Minimized Distances

squared minimized distance or squared minimized distance range	multiplicity of the minimum or number of distinct single minima in a range		
269.399553		2	38
281.987438	1	14	83
293.498743		6	91
298.237193		3	34
299.561896−331.711953		2	33
331.740622		16	166
334.256694−348.007001			6
352.575166			3
352.692156			1
353.667855			1
354.382221	1	6	75
356.063862	1	8	61
361.802589			1
364.798302	1	4	39
364.844460−366.796168		2	19
368.193310		1	19
368.756989		1	1
369.702326			1
370.308774	1	2	7
370.544914		2	31
371.222408−391.708377	1	2	15
396.478265	1	1	4
397.272980−412.043015		1	20
412.495148	1	2	6
413.183599−1115.572841	2	25	245

DISCUSSION OF THE LOCAL MINIMA PROBLEM

The minimization procedure leads to local minima, even for small molecules. Thus, several minimizations must be performed to reach indeed the global minimum. An example of distribution of the minima is presented for the saxitoxin (STX) and tetrodotoxin (TTX) sodium channel inhibitors. These toxins have quite different 2D formulas and are known to have some shape similarity.[76,77] The atomic coordinates are available from the literature.[54] Minimizations were performed to get 10, then 100, and then 1000 minima. The associated distributions are in Table 4, and the nine most frequently observed minima are displayed in Figure 3. All of these attractive minima are observable when only 100 minimizations are done, but it seems that 10 minimizations do not suffice. It should be pointed that these distributions are dependent on the random starting point selection algorithm, at least because there are several ways to define what is a random translation. The random rotation was an isotropic random axis associated to a random angle uniformly distributed over [O;Pi]. Then, the random translation was such that a random point of the moved molecule is translated to a random point of the fixed molecule, the random point of a molecule being defined such that the probability that it has to appear in a region of the space is proportional to the overall positive charge density weighed by the total positive charge of the molecule plus the overall negative charge density weighed by the total negative charge of the molecule.

The STX/TTX alignment obtained for the smallest minimized distance has not superimposed the guanidinium group of the toxins, although the latter is known to be responsible for the common biological activity.[77] However, the eighth most frequently observed minimum shows this superimposition (see Figure 3). It means that it is sometimes useful to look at the local minima, and not only at the global minima, to get pertinent information. In the particular example of STX/TTX alignments, it is stressed that the similarity was recognized by the biologists to be a steric one[76,77] and not an electrostatic one. Thus, the volumic similarity criterion which was successfully applied to STX/TTX[44] was more adequate than the similarity criterion presented in this paper. Despite the fact that these difficulties were not encountered for the other molecules, the presentation of the STX/TTX example was useful to illustrate that no similarity procedure is universal, because there are various similarity concepts.

CONCLUSION

A procedure optimizing the electronic and steric alignment of three-dimensional molecules has been presented. The molecular distance is minimized under all rotations and translations within some seconds for molecules with some tenths of atoms. No connectivity information is required. This is useful when incomplete data obtained from NMR or X-ray experiments are used. Large atomic radii are better suitable, because the optimal alignments are easier get. Each atomic radius is chemical-context independent, because only atomic positions and computed atomic charges are read.

There is little difference when the atomic charges are computed with different quantum chemistry methods, and the procedure works even with null atomic charges. In this situation, the procedure acts as a 3D-substructure search algorithm rather than an electronic similarity program. The lack of atomic charges decreases slightly the intrinsic molecular distance: computing the atomic charges is therefore useful, but not needed. This is a major advantage when a test compound is matched against a large database (as the Cambridge Crystallographic): the charge calculation step is not needed. However, it has been shown that the intrinsic molecular distance between compounds is indeed sensitive to charge calculations and is an adequate electronic similarity tool.

The electronic and the protonic parts of the molecular distance are computed on the basis of an homogeneous spheres model. Although physically unrealistic, this model was retained because it leads to intrinsic molecular distances matrices easily computable and interpretable.

The selection of random starting rotations and translations has lead the Newton-like minimizer to be trapped in various local minima. It does not mean that the minimizer is inefficient, but rather that the overall procedure could be a black box performing as Newton steps as needed to obtain the global minimum. Using such a black box precludes the detection of possible special situations for which some local minima is chemically or biologically pertinent. There is thus some interest in looking at local minima.

Once some alignment is performed, the maximal common 3D-substructure is obtained using the sorted distances matrix algorithm. The pairwise atomic correspondence between the maximal 3D-substructures of two aligned molecules is a byproduct of this algorithm. This algorithm can be applied for alignments obtained with any other procedure, but the pertinence of the common 3D-substructure is not ensured. Comparing X-ray crystallographic data compounds, or geometry and charge calculations, and performing quantitative chirality measurements are some applications presented in this paper. The molecules used in the present study have some tenths of atoms. They were selected to exemplify how the procedure works rather than for their chemical or biological interest.

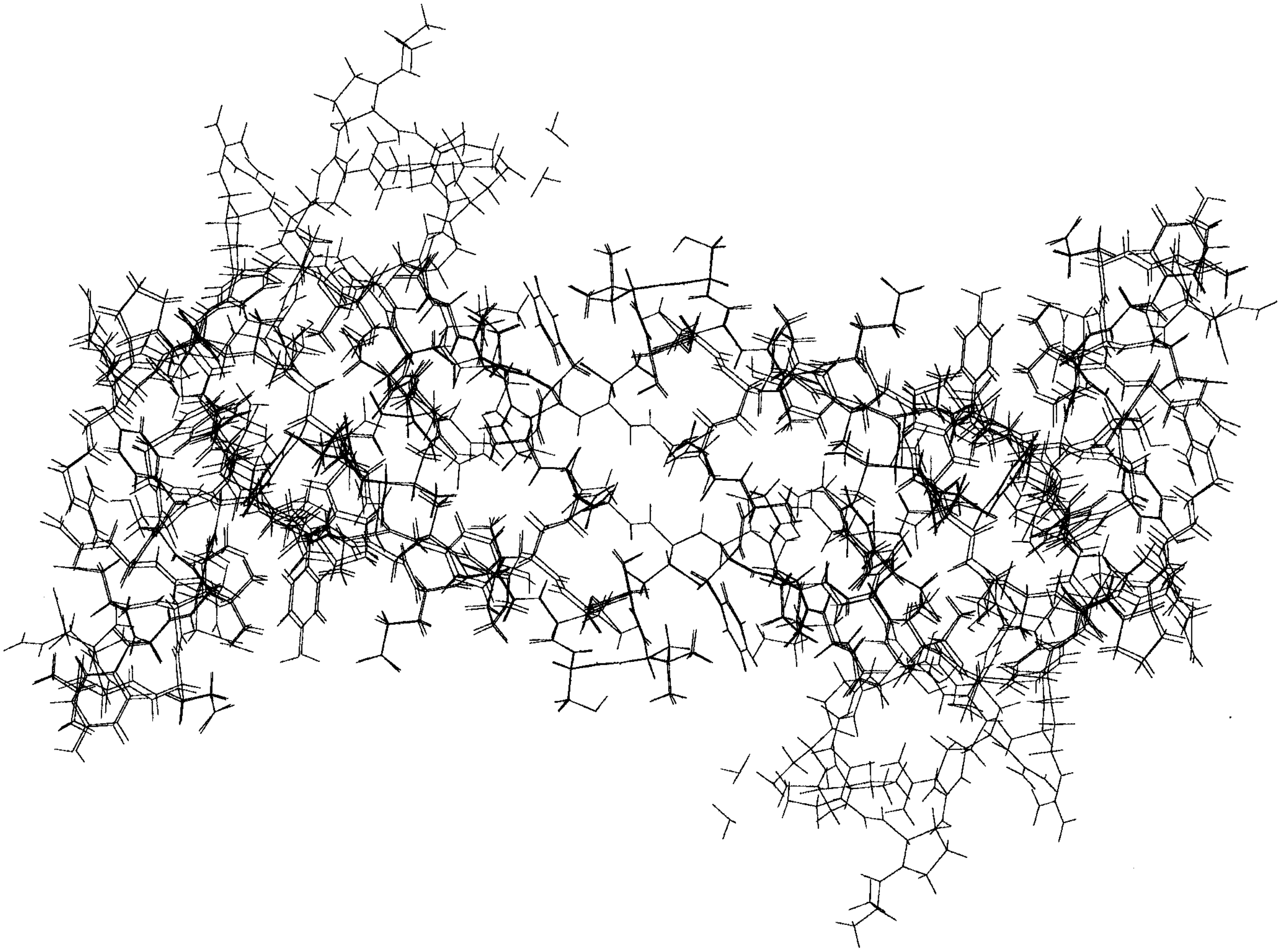

Figure 2. Optimal alignment of cytokin models 1 and 2.

When proteins with thousands of atoms are matched, the computing times increase to several hours. Although the optimal alignments between two models of the same protein are correctly performed, a part of the maximal common 3D-substructure is not recognized by the sorted distances matrix algorithm. This can be due to the rigidity of the molecular model, but the question arises how to find some criterion to decide when two 3D-models of the same 2D-chemical are indeed identical or not.

ACKNOWLEDGMENT

This work was connected with the project number 940430 at the IDRIS (Orsay) and CNUSC (Montpellier) computing centers. I thank V. Alessandrini, M. Allavena, and G. Urbach for obtaining the associated computing resources. I thank T. Couesnon for formation upon the Cambridge Structural Database, and I thank V. Fabart for formation upon the SYBYL software. I also thank Mrs. Samuel for bibliographic assistance.

APPENDIX 1: PROPERTY OF THE *n*-TUPLE OF DISTANCES IN A METRIC SPACE

The following lemma is proven here: Let $D_1,D_2,,...,D_n$ a finite set of distances defined over the same space S. Lemma: $D = (D_1*D_1 + D_2*D_2 + ...D_n*D_n)^{1/2}$ is a distance over S.

Obviously, it can be recurrently shown that, if the lemma stands for $n = 2$, it stands for any n. Assume now that $n = 2$: $D*D = D_1*D_1 + D_2*D_2$ is a positive function such that $D(x,y) = 0 \Leftrightarrow D_1(x,y) = 0$ and $D_2(x,y) = 0 \Leftrightarrow x = y$. D is symmetric. Let us define N_1, N_2, and $N = N_1 + N_2$

$$N_1 = D_1^{\ 2}(x,y) - D_1^{\ 2}(x,z) - D_1^{\ 2}(y,z) - 2*D_1(x,z)*D_1(y,z)$$

$$N_2 = D_2^{\ 2}(x,y) - D_2^{\ 2}(x,z) - D_2^{\ 2}(y,z) - 2*D_2(x,z)*D_2(y,z)$$

Squaring the triangle inequalities for D_1 and D_2 shows that N cannot be positive. Assume that some x, y, z do not satisfy the triangle inequality for D

$$D(x,y) > D(x,z) + D(y,z)$$

Then

$$D^2(x,y) - D^2(x,z) - D^2(y,z) > 2*D(x,z)*D(y,z)$$

The first term of this inequality is rewritten with N

$$N + 2*D_1(x,z)*D_1(y,z) + 2*D_2(x,z)*D_2(y,z) > 2*D(x,z)*D(y,z)$$

and because N is not positive

$$D_1(x,z)*D_1(y,z) + D_2(x,z)*D_2(y,z) > D(x,z) + D(y,z)$$

This inequality is squared

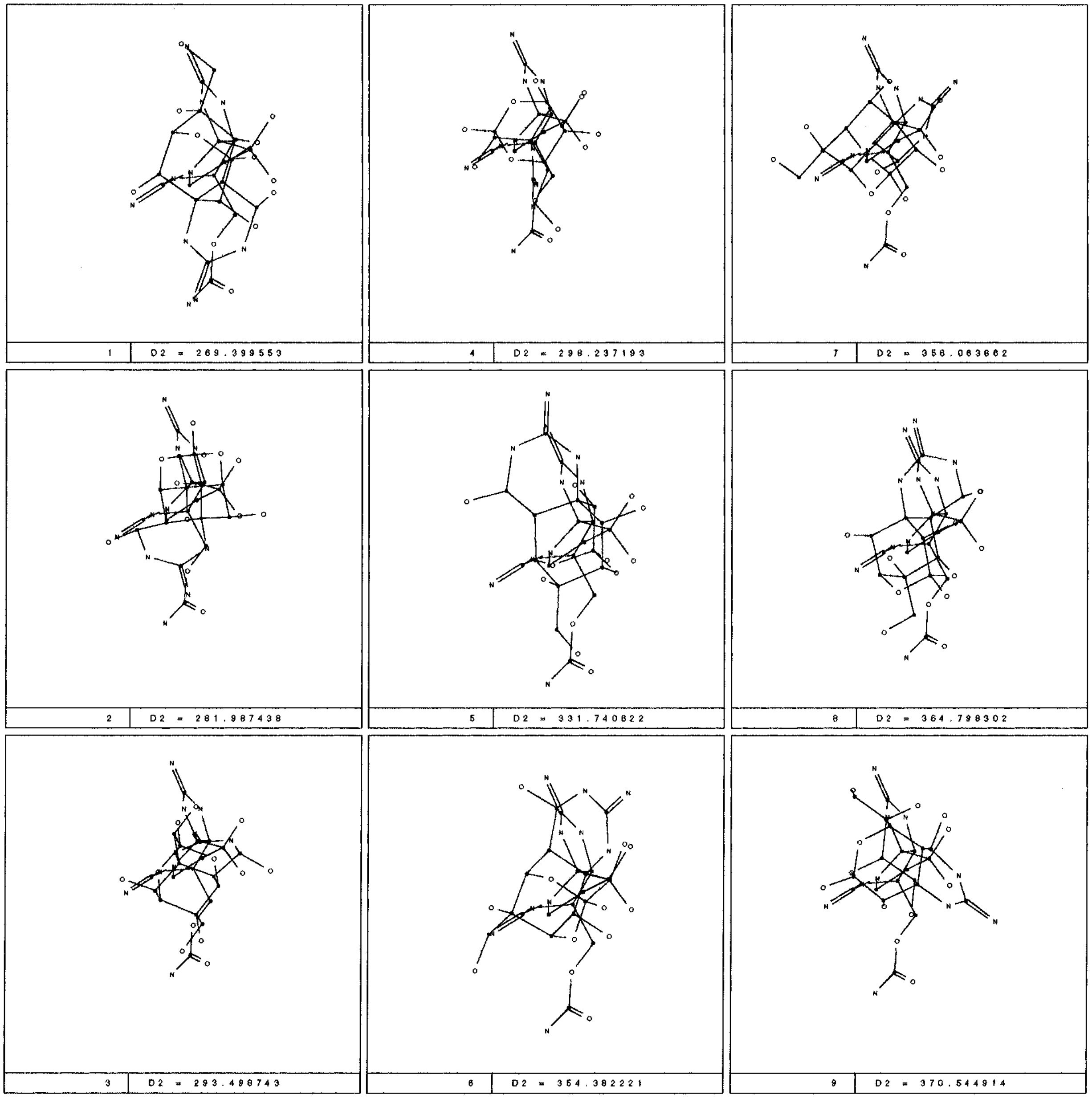

Figure 3. STX/TTX: the nine most frequent alignments obtained for 1000 minimizations.

$$D_1^2(\mathrm{x,z})*D_1^2(y,z) + D_2^2(x,z)*D_2^2(y,z) + 2*D_1(x,z)*D_1(y,z)*D_2(x,z)*D_2(y,z) > (D_1^2(x,z) + D_2^2(x,z))*(D_1^2(y,z) + D_2^2(y,z))$$

After some reordering, we get $-(D_1(x,z)*D_2(y,z) - D_2(x,z)*D_1(y,z))^2 > 0$. No x,y,z can lead to this impossible inequality. The triangle inequality always stands for D, which is indeed a distance over S.

APPENDIX 2: TOPOLOGICAL EQUIVALENCE BETWEEN MOLECULAR METRICS

We assume the homogeneous spheres model. Let n_1 and n_2 be the respective number of atomic spheres of the two molecules. Let us consider first the electronic components of the two intrinsic molecular distances defined in the theory section. $D_f^2 = \langle f_1 - f_2 | f_1 - f_2 \rangle$ and $D_g^2 = \langle g_1 - g_2 | g_1 - g_2 \rangle$, the functional scalar product being the volumic integral over the whole space, with

$$f_1 = \Big(\sum_{i_1=1}^{i_1=n_1} w_{i1}*1(il)\Big)^{1/2},\quad f_2 = \Big(\sum_{i_2=1}^{i_2=n_2} w_{i2}*1(i_2)\Big)^{1/2}$$

$$g_1 = \sum_{i_1=1}^{i_1=n_1} (w_{i1}*1(i_1))^{1/2},\quad g_2 = \sum_{i_2=1}^{i_2=n_2} (w_{i2}*1(i_2))^{1/2}$$

w_i denoting the ratio of the total electronic charge of the spherical atom i to its volume, and $1(i)$ denoting the indicator function of i (i.e., $1(i) = 1$ inside the atomic sphere and $1(i) = 0$ outside). Let $h_1 = f_1^2$, $h_2 = f_2^2$, and $D_h^2 = \langle h_1 - h_2 | h_1 - h_2 \rangle$.

The following two lemmas are needed to prove that D_f and D_g are tending to zero together: Lemma 1: $D_f \to 0 \Leftrightarrow D_h \to 0$. Lemma 2: $D_h \to 0 \Leftrightarrow D_g \to 0$.

Proof of Lemma 1. f_1 and f_2 are piecewise constant nonnegative functions of the spatial coordinates, and $(h_1 - h_2)^2 = (f_1 - f_2)^2*(f_1 + f_2)^2$, with $(f_1 + f_2) = 0$ if and only if $f_1 = f_2 = 0$, i.e., outside the union of the atomic spheres of both molecules. Thus, $A = \min(f_1 + f_2)$ exists and is a strictly positive number inside this union. $B = \max(f_1 + f_2)$ exists also. Outside the area of the space for which $h_1 = h_2 = f_1 = f_2 = 0$, we have $0 < A^2*(f_1-f_2)^2 \leqslant (h_1 - h_2)^2 \leqslant B^2*(f_1-f_2)^2$, and then $A^2 * \langle f_1 - f_2|f_1 - f_2\rangle \leqslant \langle h_1 - h_2|h_1 - h_2\rangle \leqslant B^2*\langle f_1 - f_2|f_1 - f_2\rangle$, and therefore $D_f \rightarrow 0 \Leftrightarrow D_h \rightarrow 0$.

Proof of Lemma 2. The following assumption is made: there is a universal positive constant z, which can be arbitrarily small, such that no molecule has two atomic spheres for which $(V_i + V_j - 2*V_{ij})/(V_i + V_j) < z$, V_i and V_j being their respective volumes, and V_{ij} being the volume of their intersection. This assumption means that no molecule contains two atomic spheres with very close centers and radii. It is obviously true for physical systems, even when the radii are different.

We start first from $D_h^2 = \langle h_1 - h_2|h_1 - h_2\rangle \rightarrow 0$. Let us consider the union of the $n_1 + n_2$ atomic spheres of the two molecules. Assume that at least one of the n_1 spheres of the molecule 1 is such that its volume minus the volume of its intersection between the union of n_2 spheres of the molecule 2 is not tending to zero: the function to be integrated over this area is equal to h_1^2 and its contribution to the volumic integral $\langle h_1 - h_2|h_1 - h_2\rangle$ cannot tend to zero. Thus, any one of the n_1 spheres is covered, may be except an area with volume tending to zero, by the union of the n_2 spheres of the other molecule, and conversely. As a consequence, the geometric distance between the molecules (i.e., the volume of their union minus the volume of their intersection) is tending to zero. Let us consider the first molecule. There is at least one sphere not included in the union of the $n_1 - 1$ remaining ones (if not, it would mean that all spheres have same radius and center): apart a quasi-null volume area, its nonincluded part is covered by at least one sphere of the second molecule. The existence of a universal constant z ensuring that there is not two of the n_2 spheres of the molecule 2 with infinetely close center and radii implies that at most one of the n_2 spheres of the molecule 2 can cover the nonincluded part of the sphere of the molecule 1: the indicator function of this sphere of the molecule 2 is equal to the indicator function of the associated sphere of the molecule 1, apart over a volume tending to zero.

Remark: this is a consequence of geometric properties of spheres. It is not true for any class of shapes, as parallelepipeds: the union of two parallelepipeds can cover exactly a third one without being all three perfectly superimposed. When the union of two spheres covers exactly a third one, they are at least two perfectly superimposed, and it remains true when the covering is restricted to a non-null volume containing a non-null area of the limiting surface of the covered sphere.

Let us consider the associated spheres and their common part not included in the rest of the first molecule. This common part has a volume not tending to zero, and, except perhaps a quasi-null fraction of this volume, it is not included in the rest of the second molecule (if not, it would mean that the covering was achieved with more than one sphere of the molecule 2). The part of the intersection of the two associated spheres which is not included in the union of the $n_1 + n_2 - 2$ remaining spheres thus has a contribution to the volumic integral which cannot tend to zero unless the difference of atomic densities tends to zero.

If both molecules are monoatomic, we have $D_g \rightarrow 0$. If not, let h_1' and h_2' be the respective functions associated with the molecule 1 and 2 after removal of the two associated spheres: $\langle h_1' - h_2'|h_1' - h_2'\rangle$ and $\langle h_1 - h_2|h_1 - h_2\rangle$ are differing by a quantity tending to zero. Thus $\langle h_1' - h_2'|h_1' - h_2'\rangle$ is also tending to zero. If only one molecule would be monoatomic, it would imply that all atomic densities of the other have to tend to zero: it is impossible from our definition of the atomic charges. The result can now be applied until all couples of associated spheres are removed, and we get $D_g \rightarrow 0$.

Let $V(i_1,i_2)$ be the volume of the intersection of the atomic spheres i_1 and i_2. The volumic integral of $1(i_1)*1(i_2)$ over the whole space is equal to $V(i_1,i_2)$. We obtain

$$D_h^{\,2} = \sum_{i_1=1}^{i_1=n_1}\sum_{j_1=1}^{j_1=n_1} w_{i1}*w_{j1}*V(i_1,j_1) + \sum_{i_2=1}^{i_2=n_2}\sum_{j_2=1}^{j_2=n_2} w_{i2}*w_{j2}*V(i_2,j_2) - 2*\sum_{i_1=1}^{i_1=n_1}\sum_{i_2=1}^{i_2=n_2} w_{i1}*w_{i2}*V(i_1,i_2)$$

from the expansion of the scalar product. The expression of D_g is the same, except that each coefficient w_i is replaced by its square root.

When each w_i coefficient is interchanged with its square root, the same proof can be used to obtain conversely: $D_g \rightarrow 0 \Rightarrow D_h \rightarrow 0$. Remark: replacing either in D_g or in D_h each w_i by any strictly positive function of w_i leads to the same conclusions (e.g.: the geometric homogeneous interpenetrating spheres model formally obtain when $w_i = 1$ is set).

Using both lemmas 1 and 2, we have: $D_f \rightarrow 0 \Leftrightarrow D_g \rightarrow 0$. The same result is obtained for the protonic components of the two intrinsic molecular distances. Thus, the two full intrinsic molecular distances are topologically equivalent.

REFERENCES AND NOTES

(1) Rustici, M.; Lesk, A. M. Three-Dimensional Searching for Recurrent Structural Motifs in Data Bases of Protein Structures. *J. Comput. Biol.* **1994**, *1*, 121−132.

(2) Lesk, A. M. In Encyclopedia of Computer Science and Technology: Computational Molecular Biology; Kent, A., Williams, J. G., Hall, C. M., Kent, R., Eds.; Marcel Dekker Inc.: New York, 1994; Vol. 31, Supplement 16, pp 101−165.

(3) Taylor, W. R. Remotely Related Sequences and Structures: Analysis and Predictive Modelling. *Trends Biotechnol.* **1994**, *12*, 154−158.

(4) Taylor, W. R. Protein Structure Modelling from Remote Sequence Similarity. *J. Biotechnol.* **1994**, *35*, 281−291.

(5) Holm, L.; Sander, C. Structural Alignment of Globins, Phycocyanins and Colicin A. *FEBS Lett.* **1993**, *315*, 301−306.

(6) Yee, D. P.; Dill, K. A. Families and the Structural Relatedness among Globular Proteins. *Prot. Sci.* **1993**, *2*, 884−899.

(7) Orengo, C. A.; Taylor, W. R. A Local Alignment Method for Protein Structure Motifs. *J. Mol. Biol.* **1993**, *233*, 488−497.

(8) Orengo, C. A.; Brown, N. P.; Taylor, W. R. Fast Structure Alignment for Protein Databank Searching. *Prot. Struct. Func. Gen.* **1992**, *14*, 139−167.

(9) Orengo, C. A.; Taylor, W. R. A Rapid Method of Protein Structure Alignment. *J. Theor. Biol.* **1990**, *147*, 517−551.

(10) Sali, A.; Blundell, T. L. Definition of General Topological Equivalence in Protein Structures. A Procedure Involving Comparison of Properties and Relationships through Simulated Annealing and Dynamic Programming. *J. Mol. Biol.* **1990**, *212*, 403−428.

(11) Taylor, W. R.; Orengo, C. A. Protein Structure Alignment. *J. Mol. Biol.* **1989**, *208*, 1−22.

(12) Matthews, B. W.; Rossmann, M. G. In Methods in Enzymology; Academic Press: London, 1985; Vol. 115B, pp 397−420.

(13) McLachlan, A. D. Rapid Comparison of Protein Structure. *Acta Crystallogr. A* **1982**, *A38*, 871−873.

(14) Ponec, R. Topological Aspects of Chemical Reactivity. On the Similarity of Molecular Structures. *Collec. Czec. Chem. Comm.* **1987**, *52*, 555−562.

(15) Ponec, R. Topological Aspects of Reactivity, Substituent Effects on the Validity of Woodward-Hoffmann Rules. *Z. Phys. Chem.* **1989**, *270*, 365−376.

(16) Ponec, R. Similarity Approach to Chemical Reactivity. A Simple Criterion for Discriminating between One-Step and Stepwise Reaction Mechanisms in Pericylic Reactivity. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 805−811.

(17) Ponec, R. Electron Pairing and Chemical Bonds. *Coll. Czech. Chem. Comm.* **1994**, *59*, 505−516.

(18) Ponec, R.; Strnad, R. Position Invariant Index for Assessment of Molecular Similarity. *Croat. Chem. Acta* **1993**, *66*, 123−127.

(19) Ponec, R.; Strnad, R. Topological Aspects of Chemical Reactivity. Electron Correlation in the Course of Chemical Reactions. *Coll. Czech. Chem. Comm.* **1993**, *58*, 1751−1760.

(20) Ponec, R.; Strnad, R. The Least Motion Principle, Concertedness and the Mechanisms of Pericyclic Reactions. A Similarity Approach. *Coll. Czech. Chem. Comm.* **1994**, *59*, 75−88.

(21) Islam, S. A.; Wolf, C. R.; Lennard, M. S.; Sternberg, M. J. E. A Three-Dimensional Molecular Template for Substrates of Human Cytochrome P450 Involved in Debrisoquine 4-Hydroxylation. *Carcinogenesis* **1991**, *12*, 2211−2219.

(22) Good, A. C.; Hodgkin, E. E.; Richards, W. G. Similarity Screening of Molecular Data Sets. *J. Comput.-Aided Mol. Design* **1992**, *6*, 513−520.

(23) Good, A. C.; So, S.-S.; Richards, W. G. Structure−Activity Relationships from Molecular Similarity Matrices. *J. Med. Chem.* **1993**, *36*, 433−438.

(24) Calder, J. A.; Wyatt, J. A.; Frenkel, D. A.; Casida, J. E. CoMFA Validation of the Superposition of Six Classes of Compounds which Block GABA Receptors Non-Competitively. *J. Comput.-Aided Mol. Design* **1993**, *7*, 45−60.

(25) Cramer, 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.

(26) van de Waterbeemd, H.; Carrupt, P. A.; Testa, B. Similarities of Pharmacophoric Patterns Revealed by the MEP of Metoclopramide, Molindone and Piquidone, a Subgroup of Dopamine D-2 Receptor Antagonists. *J. Mol. Gr.* **1986**, *4*, 51−55.

(27) Hermann, B.; Herron, D. K. OVID and SUPER: Two Overlap Programs for Drug Design. *J. Comput.-Aided Mol. Design* **1991**, *5*, 511−524.

(28) Martin, Y. C.; Bures, M. G.; Danaher, E. A.; DeLazzer, J.; Lico, I.; Pavlik, P. A. A Fast New Approach to Pharmacophore Mapping and Its Application to Dopaminergic and Benzodiazepine Agonists. *J. Comput.-Aided Mol. Design* **1993**, *7*, 83−102.

(29) Dean, P. M. Molecular Similarity. In 3D QSAR in Drug Design: Theory, Methods and Applications; Kubinyi, H., Ed.; ESCOM Science Publishers B.V.: Leiden, The Netherlands, 1993; pp 150−172.

(30) Molecular Similarity in Drug Design; Dean, P. M., Ed.; Blackie Academic & Professional, Chapman & Hall: London, 1995.

(31) Jakes, S. E.; Watts, N.; Willett, P.; Bawden, D.; Fisher, J. D. Pharmacophoric Pattern Matching in Files of 3D Chemical Structures: Evaluation of Search Performance. *J. Mol. Gr.* **1987**, *5*, 41−48.

(32) Brint, A. T.; Willett, P. Pharmacophoric Pattern Matching in Files of 3D Chemical Structures: Comparison of Geometric Searching Algorithms. *J. Mol. Gr.* **1987**, *5*, 49−56.

(33) Brint, A. T.; Davies, H. M.; Mitchell, E. M.; Willett, P. Rapid Geometric Searching in Protein Structures. *J. Mol. Gr.* **1989**, *7*, 48−53.

(34) Pepperrell, C. A.; Willett, P. Techniques for the Calculation of Three-Dimensional Structural Similarity Using Interatomic Distances. *J. Comput.-Aided Mol. Design* **1991**, *5*, 455−474.

(35) Pepperrell, C. A.; Poirrette, A. R.; Willett, P.; Taylor, R. Development of an Atom Mapping Procedure for Similarity Searching in Databases of Three-Dimensional Chemical Structures. *Pesticide Science* **1991**, *33*, 97−111.

(36) Willett, P. A Review of Three-Dimensional Chemical Structure Retrieval Systems. *J. Chemometrics* **1992**, *6*, 289−305.

(37) Wild, D. J.; Willett, P. Similarity Searching in Files of Three-Dimensional Chemical Structures. Implementation of Atom Mapping on the Distributed Array Processor DAP-610, the MasPar MP-1104, and the Connection Maching CM-200. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 224−231.

(38) Bath, P. A.; Poirrette, A. R.; Willett, P.; Allen, F. H. Similarity Searching in Files of Three-Dimensional Chemical Structures Comparison of Fragment-Based Measures of Shape-Similarity. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 141−147.

(39) Clark, D. E.; Jones, G.; Willett, P. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Comparison of Conformational-Searching Algorithms for Flexible Searching. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 197−206.

(40) Danziger, D. J.; Dean, P. M. The Search for Functional Correspondences in Molecular Structure between Two Dissimilar Molecules. *J. Theor. Biol.* **1985**, *116*, 215−224.

(41) Barakat, M. T.; Dean, P. M. Molecular Structure Matching by Simulated Annealing. III. The Incorporation of Null Correspondences into the Matching Problem. *J. Comput.-Aided Mol. Design* **1991**, *5*, 107−117.

(42) Papadopoulos, M. C.; Dean, P. M. Molecular Structure Matching by Simulated Annealing. IV. Classification of Atom Correspondences in Sets of Dissimilar Molecules. *J. Comput.-Aided Mol. Design* **1991**, *5*, 119−133.

(43) Feuilleaubois, E.; Fabart, V.; Doucet, J. P. Implementation of the Three-Dimensional-Pattern Search Problem on Hopfield-Like Neural Networks. *SAR QSAR Environ. Res.* **1993**, *1*, 97−114.

(44) Petitjean, M. Geometric Molecular Similarity from Volume-Based Distance Minimization: Application to Saxitoxin and Tetrodotoxin. *J. Comput. Chem.* **1995**, *16*, 80−90.

(45) Hurley, J. R.; Cattell, R. B. The Procrustes Program:Producing Direct Rotation to Test a Hypothesized Factor Structure. *Behav. Sci.* **1962**, *7*, 258−262.

(46) Carbo, R.; Leyda, L.; Arnau, M. How Similar is a Molecule to Another? An Electron Density Measure of Similarity between Two Molecular Structures. *Int. J. Quant. Chem.* **1980**, *17*, 1185−1189.

(47) Carbo, R.; Calabuig, B. Concepts and Applications of Molecular Similarity; Johnson, M. A., Maggiora, G. M., Eds.; Wiley: New-York, 1990; Chapter 6, pp 147−171.

(48) Good, A. C.; Hodgkin, E. E.; Richards, W. G. Utilization of Gaussian Functions for the Rapid Evaluation of Molecular Similarity. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 188−191.

(49) Good, A. C.; Richards, M. G. Rapid Evaluation of Shape Similarity Using Gaussian Functions. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 112−116.

(50) Hodgkin, E.; Richards, W. G. Molecular Similarity Based on Electrostatic Potential and Electric Field. *Int. J. Quant. Chem. Quantum Biol. Symp.* **1987**, *14*, 105−110.

(51) Meyer, A. Y.; Richards, W. G. Similarity of Molecular Shape. *J. Comput.-Aided Mol. Design* **1991**, *5*, 427−439.

(52) Kearsley, S. K.; Smith, G. M. An Alternative Method for the Alignment of Molecular Structures:Maximizing Electrostatic and Steric Overlap. *Tetr. Comput. Method* **1990**, *3*, 615−633.

(53) Namasivayam, S.; Dean, P. M. Statistical Method for Surface Pattern-Matching between dissimilar Molecules: Electrostatic Potentials and Accessible Surfaces. *J. Mol. Gr.* **1986**, *4*, 46−50.

(54) Dean, P. M.; Chau, P. L. Molecular Recognition: Optimized Searching Through Rotational 3-space for Pattern Matches on Molecular Surfaces. *J. Mol. Gr.* **1987**, *5*, 152−158.

(55) Lattam, E. Methods in Enzymology; Academic Press: London, 1985; Vol. 115B, pp 55−77.

(56) Fernandez Pacios, L. Atomic Radii Scales and Electron Properties Deduced from the Charge Density. *J. Comput. Chem.* **1995**, *16*, 133−145.

(57) Petitjean, M. On the Analytical Calculation of van der Waals Surfaces and Volumes: Some Numerical Aspects. *J. Comput. Chem.* **1994**, *15*, 507−523.

(58) Mostad, A.; Romming, C.; Storm, B. Structure of the DNA Complexing Agent Anthramycin. *Acta Chem. Scand. B* **1978**, *B32*, 639−645.

(59) Arora, S. K. Structural Investigations of Mode of Actions of Drugs. II. Molecular Structure of Anthramycin Methyl Ether Monohydrate. *Acta Crystallogr. B* **1979**, *B35*, 2945−2948.

(60) Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1EZ, United Kingdom.

(61) Gavezzotti, A. The Calculation of Molecular Volumes and the Use of Volume Analysis in the Investigation of Structured Media and of Solid-State Organic Reactivity. *J. Am. Chem. Soc.* **1983**, *105*, 5220−5225.

(62) Bondi, A. Van Der Waals Volumes and Radii *J. Phys. Chem.* **1964**, *68*, 441−451.

(63) Richards, F. M. Methods in Enzymology; Academic Press: London, 1985; Vol. 115B, pp 440−464.

(64) Meyer, A. Y. More on the Size of Molecules. *Struct. Chem.* **1990**, *1*, 265−279.

(65) Moriguchi, I.; Kanada, Y.; Komatsu, K. van der Waals Volume and the Related Parameters for Hydrophobicity in Structure−Activity Studies. *Chem. Pharm. Bull.* **1976**, *24*, 1799−1806.

(66) Scott, R. A.; Scheraga, H. A. Conformational Analysis of Macromolecules. III. Helical Structures of Polyglycine and Poly-L-Alanine. *J. Chem. Phys.* **1966**, *45*, 2091−2101.

(67) Ramachandran, G. N.; Sasisekharan, V. Advances in Protein Chemistry; Academic Press: New York, 1968; Vol. 23, pp 283−438.

(68) Hopfinger, A. J. Conformational Properties of Macromolecules; Academic Press: New York, 1973; Chapter 2, p 39.

(69) Cambridge Structural Database System; Vol. 3, Appendix 15, pp 15−20. Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1EZ, United Kingdom.

(70) Dinghas, A. Über das Verhalten der Entfernung zweier Punktmengen bei gleichzeigtiger Symmetrisierung derselben. *Arch. Math.* **1957**, *8*, 46−51.

(71) SYBYL Molecular Modeling Software Version 6.1, SYBYL Command Manual; p 361. SYBYL Theory Manual; Chapter 2.3, pp 64−69. TRIPOS Inc., 1699 S. Hanley Road, St. Louis, MO 63144-2913.

(72) Zabrodsky, H.; Avnir, D. Continuous Symmetry Measures. 4. Chirality, *J. Am. Chem. Soc.* **1995**, *117*, 462−473.

(73) Weinberg, N.; Mislow, K. Distance Functions as Generators of Chirality Measures. *J. Math. Chem.* **1993**, *14*, 427−450.

(74) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures. *J. Mol. Biol.* **1977**, *112*, 535−542.

(75) Petitjean, M.; Cordier, C.; Dodin, G. Unpublished results.

(76) Ritchie, J. M.; Rogart, R. B. The Binding of Saxitoxin and Tetrodotoxin to Excitable Tissue. *Rev. Physiol. Biochem. Pharm.* **1977**, *79*, 2−50.

(77) Hille, B. The Receptor for Tetrodotoxin and Saxitoxin. A Structural Hypothesis. *Biophys. J.* **1975**, *15*, 615−619.

CI9603700