Partial Molecular Alignment via Local Structure Analysis

Daniel D. Robinson, Paul D. Lyne, and W. Graham Richards*

Physical and Theoretical Chemistry Laboratory, University of Oxford, South Parks Road, Oxford OX1 3QZ, United Kingdom

Received August 4, 1999

Molecular alignment remains as one of the most problematic aspects of molecular design. A technique is introduced that facilitates the alignment of a range of structures that could not be handled easily using existing alignment procedures. The flexibility of the method is illustrated with a series of test sets. First, an alignment is performed on a series of molecules from a typical 3D-quantitative structure−activity relationship data set. The results of this test show the technique to outperform many existing alignment methodologies based upon the optimization of molecular similarity or molecular overlaps. This test set is then extended to consider the alignment of more structurally diverse inhibitors of HIV-1 reverse transcriptase and HIV-1 protease. Finally, in the most challenging test, a large protein-based inhibitor is matched with a small-molecule mimic. It is believed that the existence of such a versatile alignment technique will prove invaluable in the fields of molecular design and chemical information handling.

INTRODUCTION

The ability to assess quickly and reliably the similarity of a diverse range of molecules has great implications in the areas of drug discovery, molecular recognition, and 3D-quantitative structure−activity relationship (3D-QSAR). A key step in making this assessment is the alignment of the structures in question. The purpose of the alignment is to highlight the areas of commonality between the structures by showing them in the same region of space, presumably as they would be in the active site.

In general, the task of alignment can be broken down into the following steps. First, we have the *template molecule*. Most often this structure is kept fixed in terms of its position in space and its conformation. All of the *working molecules* are aligned with respect to this template molecule by translation, rotation, and in some cases conformational alteration. Depending on the nature of the alignment technique, the template molecule can take the form of either a known molecular structure or a pharmacophore.

Once we have the template and working molecules we need to define a method of comparing the structures so that common regions can be identified as such. Several approaches to this problem have been put forward from a variety of sources and are comprehensively reviewed by Klebe.[1] Some of the earliest work avoided the problem of selecting regions of commonality between structures by attempting to compare the whole of the working molecule with the whole of the template molecule. Carbo introduced one of the first notions of molecular comparison with his similarity index.[2] The Carbo index compared the electron density distribution of two molecules; however, this was soon modified by Hodgkin et al.[3] and Meyer et al.[4] to include the properties of electrostatic potential and molecular shape. Alignment by shape similarity is related to the alignment by common overlap volume,[5] the only difference being that a similarity calculation normalizes the output value to yield a quantity that may be more readily interpreted between different pairs of molecules.

More recent approaches to the alignment problem attempt to section up the template molecule and working molecule by associating attributes to individual atoms or collections of atoms. The work of Jones et al.[6,7] typifies this technique; here the template and working molecule are broken down into features such as H-bond acceptor/donator and ring structures. A genetic algorithm (GA) is then used to select which pairs of features (one from the template molecule and one from the working molecule) are to be compared.

Having arrived at a method of comparing the template and working molecule, it is necessary to have a mechanism for deducing the correct alignment of the two structures. In the case of the whole-molecule similarity based techniques, we attempt to optimize the similarity of the template and working molecules by altering their relative translation and rotation. This method is complicated by the fact that the optimization is an extremely difficult one, as the molecular similarity surface has a large number of local extrema. Furthermore, the optimization fails entirely when the template and working molecule become too structurally diverse as the assumption that the whole of one molecule can be compared with the whole of another molecule is clearly invalid. The more recent approaches tend to do better in this sense as they are attempting to compare areas of the template and working molecules that are chemically alike. In the case of the GA based technique of Jones et al., the alignment is carried out by performing a root-mean-square (RMS) fit between the pairs of features identified by the GA. This technique has been shown to be effective but is complicated by the fact that while the GA is capable of selecting pairs features that are chemically comparable it is not able to select pairs of features that can be geometrically overlaid. As a consequence, the program utilizes a multistage RMS fit where "outliers" of the first stage of alignment are rejected and a

* Corresponding author. E-mail: graham.richards@chem.ox.ac.uk.

10.1021/ci990272p CCC: $19.00
Published on Web 01/19/2000

second fit is carried out using only those features that could be mapped within a certain tolerance of each other.

It is therefore apparent that a method for the overlay of two molecules needs to combine the feature selection process with the molecule overlay stage. It is only by coupling these stages that we can arrive at a technique that is able to make chemically sensible selections of common feature points that are geometrically possible to overlay. It is in the pursuit of this problem that we see our technique having its greatest importance.

THEORY AND METHOD

The problem of finding a full or partial match between two three-dimensional structures has received a great deal of attention not only in the chemical literature but also in other fields. The main motivations stem from the requirement for object recognition in computer vision where a system needs to be able to identify and understand its surroundings in order to interact with them. The applications of a truly reliable system range from robot task planning to military applications for target location and tracking. Further examples of the utility of a general-purpose three-dimensional object-matching algorithm are provided in the medical field where multimodal images are registered to provide a complete view of an organ and any potential disease.

We have based our work on that carried out by Barequet and Sharir;[8] the technique has many parallels with other techniques in chemistry based on geometric hashing[9] and, as demonstrated later, clique detection algorithms.[10,11] However, it tackles the problem of finding features on the molecules to be aligned that are both chemically reasonable and geometrically feasible in a unique and rather elegant manner.

Like in most modern alignment techniques, we begin by breaking down the structures under consideration into a series of smaller fragments. Associated with each fragment is a structure that we refer to as a *footprint*, **f**. A footprint is made up of two parts. The first part is the *position*, **f.Position**. The position specifies the location of the footprint in three-dimensional space and is derived directly from the structure of the molecule. The second part of the footprint is the *descriptor*, **f.Descriptor**. The descriptor is a mathematical representation of the molecular environment about the position of the footprint. Examples of potential descriptors and their required properties will be discussed later; however, at this point it is important to bear in mind that the descriptor must be rotationally invariant. Once this stage is complete, we end up with the template molecule and the working molecule stored as lists of footprints, $\mathbf{F_T}$ and $\mathbf{F_W}$, respectively. Each complete list of footprints describes the whole of the molecule from which it was formed.

The second stage of the process is to combine the sets of footprints from the template and working molecules into a *voting table*, **V**. A voting table is made up of footprints, *voting pairs*, **v**. Each voting pair contains two footprints, one from the template molecule $\mathbf{v.f_T}$ and one from the working molecule $\mathbf{v.f_W}$. Pairs of footprints are added to the voting table if their associated descriptors meet some user-defined similarity threshold. For example, we might permit a pair of footprints to be added to the voting table if they both referred to regions of their respective molecule that were hydrogen bond acceptors. However, for the purpose of alignment, we would not form a hydrogen bond donor and a hydrogen bond acceptor into a voting pair. This initial filtering ensures that any comparisons we make between the two molecules are chemically sensible; however, it is not the purpose of this stage to filter out those comparisons that are not geometrically feasible.

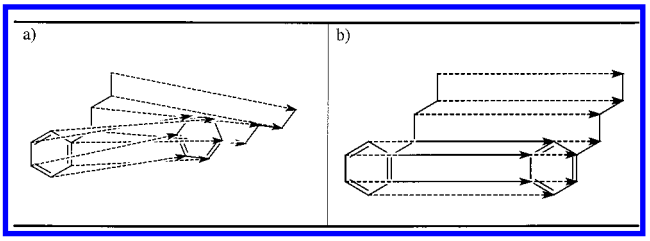

Figure 1. Illustration of the translations generated when (a) the two structures are at the wrong relative rotation and (b) the two structures are at the correct relative rotation.

The final stage of the alignment procedure allows us to simultaneously select those voting pairs that are both chemically sensible and geometrically feasible and also deduce the rotation and translation that aligns the template and working molecule. It is at this final stage that our alignment methodology differs substantially from existing techniques.

In effect we can rotate the working molecule by applying a rotation to the positions of all of its footprints. Let the position $\mathbf{v.f_W.Position_R}$ represent the position of a footprint from the working molecule that has been rotated by the current rotation. This position can be mapped onto the position $\mathbf{v.f_T.Position}$ of the footprint from the template molecule by means of a translation **t** given by eq 1. The value of **t** is a vote for the translation required to map the working molecule onto the template molecule at the current rotation.

$$\mathbf{t} = \mathbf{v.f_T.Position_R} - \mathbf{v.f_W.Position} \qquad (1)$$

Consequently, we can iterate across the complete voting table and generate a list of votes **T** that map each of the respective footprint positions from the working molecule to their counterpart in the template molecule. The three-dimensional distribution of votes within **T** will differ according to the current rotation as shown in Figure 1. Illustrated in Figure 1a are two structures that are rotationally misaligned. As a consequence of this rotational misalignment, the votes for the translations are scattered in essentially random directions. Conversely, when the current rotation aligns the template and working molecules perfectly, Figure 1b, all of the votes in **T** point in the same direction. Thus, when the correct rotation is found, the votes in **T** cluster around the translation required to complete the alignment of the working and template molecules.

Now in the case of Figure 1 the footprints have been artificially chosen to be perfect; that is, the voting table only contains entries that are the correct match for each other. But it should be apparent that if any false entries existed in the voting table they would yield votes that were scattered in direction and thus would not cluster. However, the correct entries in the voting table, those that refer to common features that can be geometrically aligned, will always accumulate into a cluster when the correct rotation is found. Thus we

have a method of aligning molecules that automatically selects the correct feature points to compare with each other. We simply rotate the working molecule, and correspondingly the positions of its footprints, and search for a cluster in **T**.

All that is now required is a method for detecting the presence of a cluster in what may be a sea of randomly distributed false votes. Several methods make themselves immediately apparent, although there are doubtless many other similar methods. First, let us consider a method from nature; let each vote in **T** represent a point in space, and let each of these points have a "gravitational" attraction to the other points represented in **T**. Then the potential of the *i*th vote in **T**, P_i, will be given by eq 2:

$$P_i = \sum_{j}^{j\in\mathbf{T}, j\neq i} \frac{1}{|\mathbf{T}_i - \mathbf{T}_j|} \tag{2}$$

If the *i*th vote is at the center of a tight cluster, then its gravitational potential will take on a maximum value. Thus all we need to do is to find a rotation of the working molecule that leads to one of the votes having the largest possible potential. This was found to be an effective scoring function provided that the value of $|\mathbf{T}_i - \mathbf{T}_j|$ was thresholded to a value τ so that it could never cause a singularity in the mathematics. In practice it was found that preventing $|\mathbf{T}_i - \mathbf{T}_j|$ from falling below 0.5 Å yielded excellent results. A related form of scoring function simply summed across the P_i as shown in eq 3. Once again when a tight cluster was present the value of P took on a maximum value.

$$P = \sum_{i}^{i\in\mathbf{T}} \sum_{j}^{j\in\mathbf{T}, j\neq i} \frac{1}{|\mathbf{T}_i - \mathbf{T}_j|} \tag{3}$$

An alternative scoring scheme represented each point in **T** as a Gaussian function. Then the overall score S for the current rotation is given by eq 4, which effectively calculates the overlap of the Gaussian functions.

$$S = \sum_{i}^{i\in\mathbf{T}} \sum_{j}^{j\in\mathbf{T}, j\neq i} \exp\left(-\frac{|\mathbf{T}_i - \mathbf{T}_j|^2}{\sigma}\right) \tag{4}$$

The parameter σ acts in a manner similar to the parameter τ in the previous scoring functions. Other similar scoring metrics are provided by Barequet and Sharir.

Having identified that a cluster is present in **T**, we need to find its position, and thus the translation needed to complete the alignment of the template and working molecules. Currently we use eq 2 to search for the vote in **T** that has the maximum potential, as this vote is at the center of the tightest cluster of votes and it best represents the position of the cluster and hence the desired translation. However, as noted by Barequet and Sharir, once the correct rotation is known the correct translation can be found by a correlation-based technique.[12]

For clarity the complete alignment procedure is outlined in Figure 2.

1. Calculate footprints $\mathbf{F_T}$ for the template molecule.
2. Calculate footprints $\mathbf{F_W}$ for the working molecule.
3. Form voting table **V**.
   - 3.1. For each footprint $\mathbf{f_T}$ in $\mathbf{F_T}$.
     - 3.1.1. For each footprint $\mathbf{f_W}$ in $\mathbf{F_W}$.
     - 3.1.2. Calculate similarity of $\mathbf{f_T}$**.Descriptor** and $\mathbf{f_W}$**.Descriptor**.
     - 3.1.3. If similarity is greater than a user defined threshold add $\mathbf{f_T}$ and $\mathbf{f_W}$ to **V** as a voting pair.
4. Initialize the best rotation $\mathbf{q_b}$, the best translation $\mathbf{t_b}$, and the current rotation $\mathbf{q_c}$.
5. Calculate the votes.
   - 5.1. For each voting pair **v** in **V**.
     - 5.1.1. Calculate the position **v.f**$_\mathbf{W}$**.Position**$_\mathbf{R}$ from **v.f**$_\mathbf{W}$**.Position** by applying the current rotation $\mathbf{q_c}$.
     - 5.1.2. Calculate the translation **t** that maps **v.f**$_\mathbf{W}$**.Position**$_\mathbf{R}$ to **v.f**$_\mathbf{T}$**.Position** using equation 1. Store **t** in the list of votes **T**.
6. Calculate the score for the current series of votes **T** using either equation 2, 3 or 4.
7. If the score is greater than the current best score then.
   - 7.1. Set $\mathbf{q_b}$ to $\mathbf{q_c}$.
   - 7.2. Find the vote that has the maximum value of P_i (equation 2). Store this vote as the best translation $\mathbf{t_b}$.
8. If we are complete then exit else update $\mathbf{q_c}$ to a new rotation and goto 5.

Figure 2. Outline of the alignment routine.

GENERATING DESCRIPTORS

The type of descriptor that we use dictates the nature of the comparison we perform between the template and working molecule and possibly the nature of the alignment that results. The only requirement for our technique is that the descriptors are rotationally invariant so that matching descriptors can be found even when they are generated from two molecules that are not aligned. It must be stressed that this section does not promote the methods of descriptor generation as being particularly efficient or the best available; it merely demonstrates some of the descriptors that are possible and explains the techniques used to test the methodology that are described in the Results section.

There are clearly many potential methods of generating a rotationally invariant descriptor from a molecule. Existing alignment techniques utilize a variety of parameters from the position of donor/acceptor sites and the presence of rings.[7] There is little doubt that these methods could be added to this alignment technique with a great deal of success. However, to keep the testing of what is a new methodology simple, we have restricted ourselves to considering single atom based descriptors.

The simplest form of atom-based descriptor is clearly the type of atom, carbon, hydrogen, etc. In this case only those footprints whose descriptors were an exact match for each other would be permitted to enter the voting table. Clearly such a crude descriptor is limited in its abilities as it will not be able to distinguish between different carbon atoms in a molecule and consequently the voting table will contain many false entries. These false entries not only act to obscure the cluster of good votes when the correct rotation is found but also slow the routine considerably. Nonetheless a surprisingly large number of cases were successfully handled by this technique, illustrating the reliability of the underlying

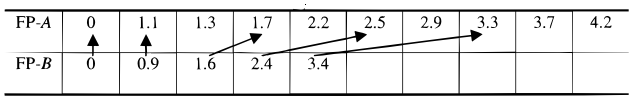

Figure 3. Illustration of the elastic-matching algorithm used in the comparison of distance-based footprints.

methodology and in particular the ability of the scoring functions described previously to locate a cluster in a sea of other data points.

A considerably more reliable descriptor, and the one used for all of the examples given below, involved an analysis of the interatomic distances within a molecule. The distance-based descriptor for the *i*th atom in a molecule is a vector containing the distances of that atom from all other atoms in the molecule. This descriptor is clearly a row (or column) from the molecule's distance matrix. It is known that the distance matrix encodes virtually the entire structure of a molecule enabling the original structure to be regenerated using distance geometry techniques.[13] The comparison of two descriptors generated in this manner needs to be undertaken carefully to avoid any dependency on the atom-numbering scheme used between molecules. This invariance to atom numbering was achieved by sorting all of the distances within a descriptor into ascending order. An unconstrained elastic-matching algorithm as illustrated in Figure 3 was used to compare the sorted distances in each descriptor. Basically this algorithm compares a distance in one descriptor to the closest matching distance in the second descriptor. The sum of the absolute difference between the matched distances provides a numerical measure of how similar the two descriptors are.

There are obvious parallels between alignment using these distance-based descriptors in our technique and alignment methods that attempt to search for the maximum clique between the graphs of the respective molecules. However, in our technique, the largest cluster of votes, which is related to the maximum clique, can always be found by iterating across all possible rotations of the working molecule, a problem of fixed complexity. Conversely, the maximum clique detection problem is in general NP-complete, and while claims have been made for a randomized approach[11] with a reasonable practical running time complexity it is impossible to guarantee finding the best solution by exhaustive evaluation on all but the most trivial of problems. Furthermore, once the maximum clique has been detected, it is still necessary to solve for the transformation that aligns the template and working molecule in a separate step. Although this is a relatively trivial exercise, our alignment technique yields all of this information in a single step.

IMPLEMENTATION DETAILS

To test the methodology described, it has been implemented in Microsoft Visual Basic v5.0[14] on a standard PC under Microsoft Windows NT v4.0. As this implementation was written purely to test the methodology, it has not been optimized to any extent. However, the majority of the examples shown below complete in the order of a few minutes, the longest run time being for the DFKi/TOMI alignment, which required around 9 min of CPU time. Such run times are already similar to the other comparable alignment techniques mentioned but could clearly be improved on in an optimized version written in C++.

The software is a general-purpose implementation of the algorithm described above. A separate program calculates the footprints in the manner described and stores the results for each molecule in a separate file. The run time of this

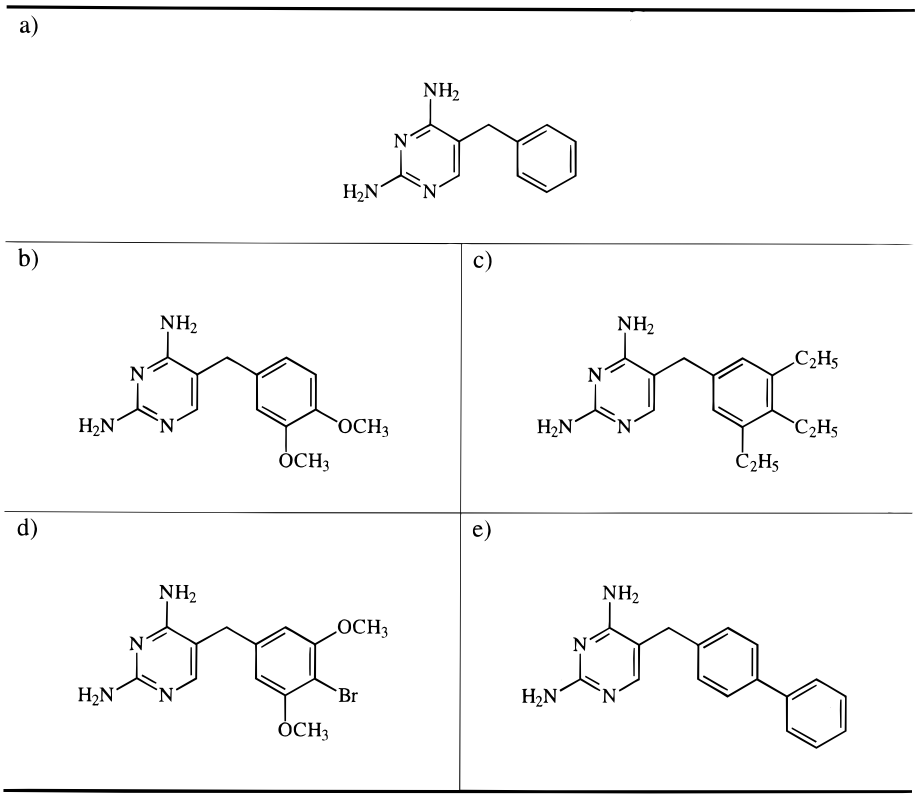

Figure 4. Structures used in the QSAR data set test.

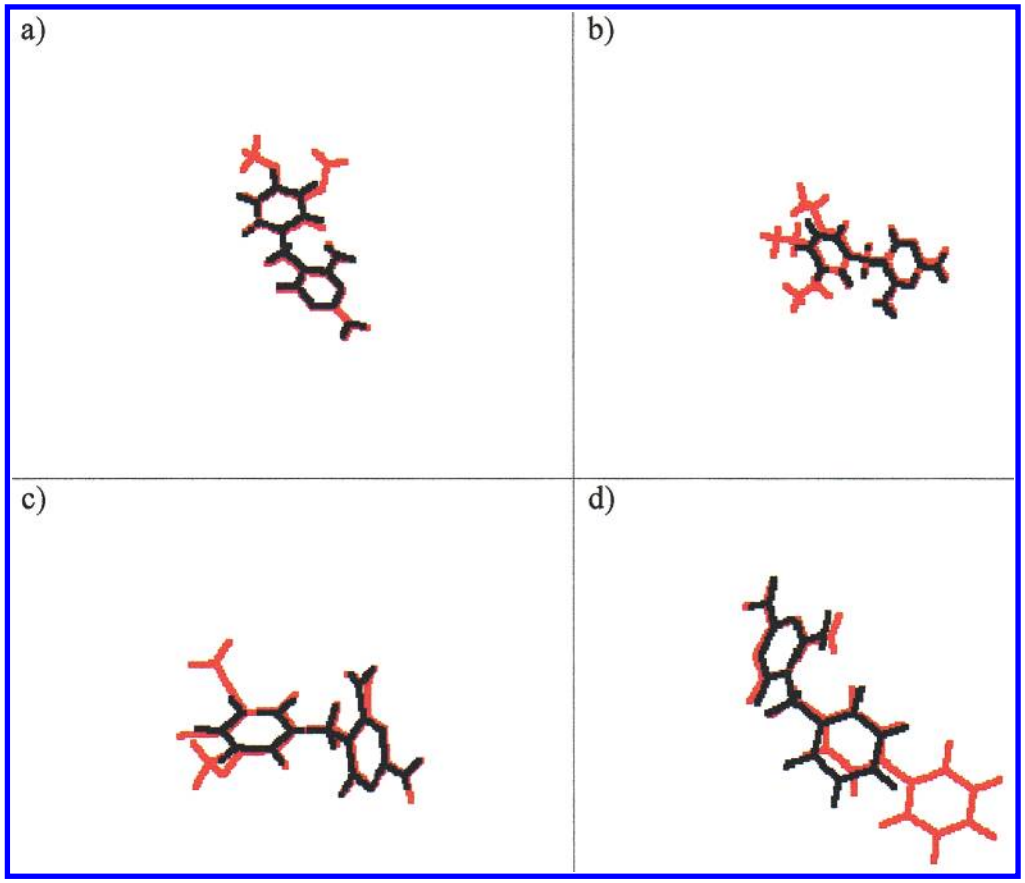

Figure 5. Results from the QSAR data set test.

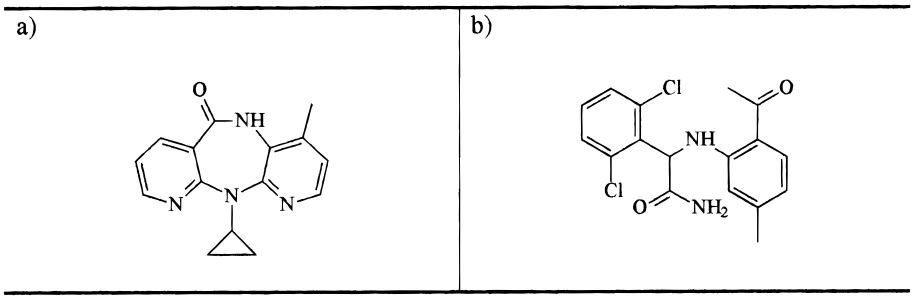

Figure 6. HIV-1 reverse transcriptase inbitors used in the second test set: (a) nevirapine and (b) alpha-APA.

program is negligible for all of the examples shown below. The main program reads the footprints and structures corresponding to the template and working molecule. The user is then required to set the threshold for a pair of descriptors to be considered sufficiently similar to be added to the voting table. In general, it is found that families of molecules require very similar thresholds, and this opens up the possibility that future versions of the software could set the threshold automatically, thereby making the alignment procedure completely automated.

Once the threshold is set, the voting table can be generated and the correct rotation determined. Our initial feeling was to specify the rotation through Euler angles; however, experimentation showed that it was substantially easier to optimize rotations specified as quaternions. Similarly the software was originally written to use a steepest descents class of optimization algorithm similar to that implemented by Barequet and Sharir.[8] In practice it was found that a simple annealing-based optimizer[15] gave superior results with little penalty in the run time. The enhanced performance of

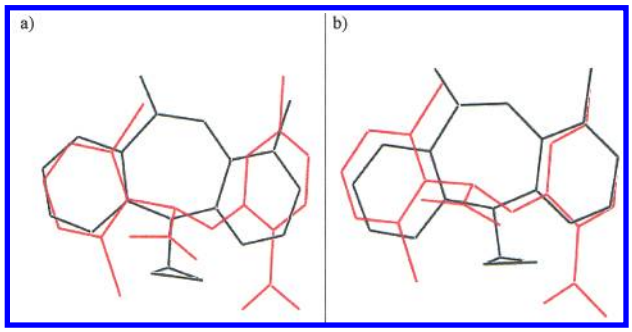

Figure 7. Results of the alignment of the inhibitors of HIV-1 reverse transcriptase: (a) experimental alignment and (b) result of test.

the annealing optimizer can be understood when we consider the presence of many local maxima in the scoring function that would lead a normal optimizer to converge prematurely. The existence of a large number of local maxima in the scoring function should not be taken as a weakness of the scoring function, but rather as an inevitable consequence of

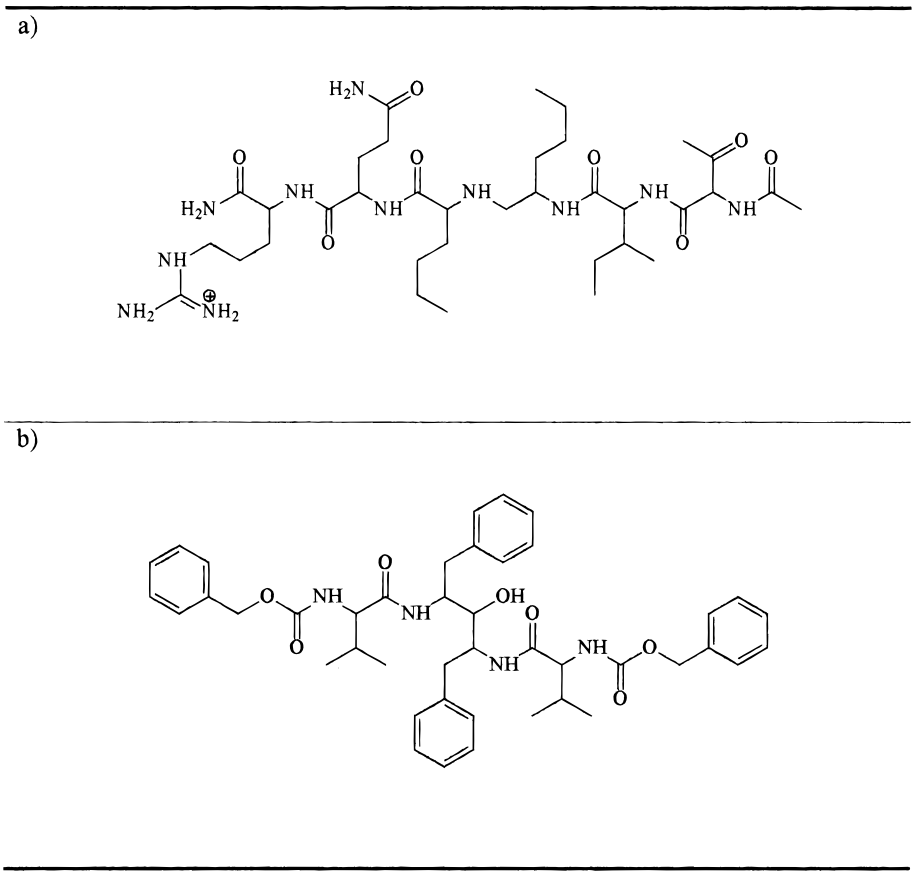

Figure 8. HIV-1 protease inhibitors used in the second test.

the nature of the alignment being undertaken. In many circumstances there will be several acceptable alternative solutions to the alignment owing to the repeat of a structural motif within one of the molecules. To prove this point, a special version of the software has been written that performs a complete analysis across the whole of rotational space. The user can then examine the alignment that corresponds to any particular local maxima in the scoring function. While this program demonstrates that each of the local maxima are plausible alignments, the program requires too much user interaction to be used as an automated molecular alignment system, and as the code is not optimized to any extent it also suffers from an excessively long run time. We currently do not know of a method that is able to locate all maxima within a system without performing a search across the whole of rotation space.

RESULTS

Several test sets have been used during the testing of this algorithm. The test sets are designed to provide a number of challenges for the methodology, ranging from a simple alignment of a low-diversity set of molecules from a standard QSAR set to a highly challenging comparison of a natural protein inhibitor with a small artificial mimic. In all cases the algorithm yielded convincing alignments.

Where possible the quality of the alignment is measured by calculating an RMS deviation value between the molecule being aligned and the position of that molecule determined from crystallographic experiments. However, in the absence of crystal structure data, the results are purely visual. In general, there is no accepted metric for quantifying the quality of an alignment between two structures that do not exhibit an obvious atom correspondence. If such a metric existed, there is little doubt that it would have been employed for the purposes of molecular alignment a long time ago.

Low-Diversity QSAR Set. The first test set is taken from a simple QSAR analysis performed on a series of DHFR inhibitors.[16] As with most QSAR data sets, these molecules exhibit a low range of structural diversity being based around a common structural motif.

Figure 4 shows the structures used for this test. Figure 4a was aligned with the remaining structures in the figure. All of the structures were generated using CAChe[17] and minimized using the AM1 Hamiltonian[18] implemented in MOPAC 6.0.[19]

The resulting alignments are shown in Figure 5. As can be seen, the technique yields an excellent alignment for all of the structure pairs. In the first three cases, this is rather unremarkable; a standard similarity optimizer is able to get similar quality alignments. However, in the case of the alignment shown in Figure 5d, the result is much more interesting. In this case the species being aligned show a considerable difference in size; from a simple ring count the structure in Figure 4e is 50% larger than the structure in Figure 4a. This disparity in size proved too difficult for several of the similarity optimizers we tested, which failed to overlay the common substructure correctly.

High-Diversity Molecule Alignment. Following from the previous test, an attempt was made to reproduce the experimentally determined alignment of more structurally

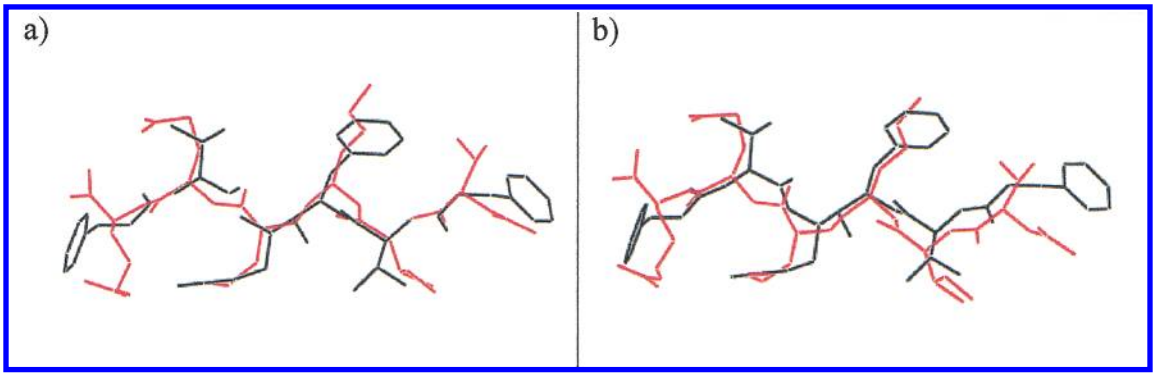

Figure 9. Results of the alignment of the inhibitors of HIV-1 protease: (a) experimental alignment and (b) result of test.

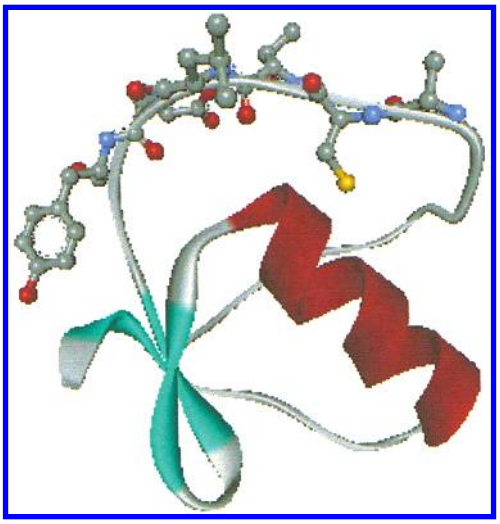

Figure 10. Structure of TOMI. The active residues are displayed in a ball-and-stick representation.

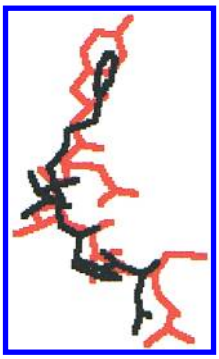

Figure 11. Result of the alignment between DFKi and the active fragment of TOMI.

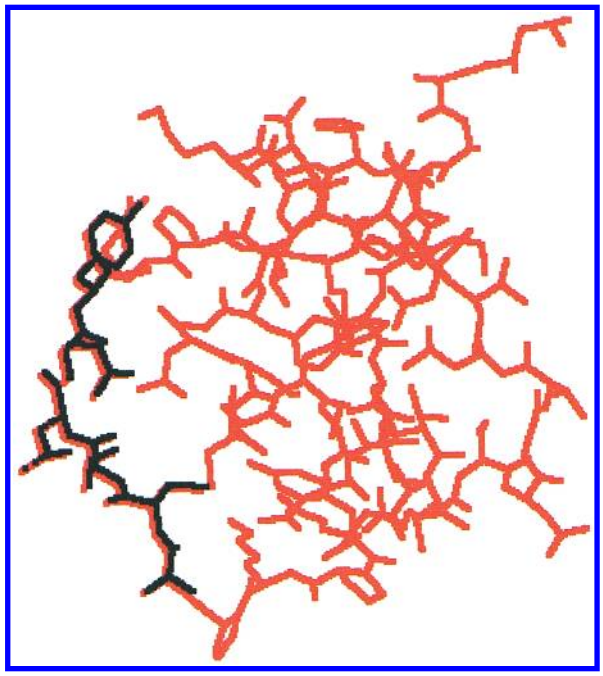

Figure 12. Result of the alignment between the active fragment of TOMI and the complete TOMI structure.

diverse molecules. The crystal structures of HIV-1 reverse transcriptase[20] complexed with two different inhibitors, nevirapine and alpha-APA, were obtained from the Protein Databank.[21] The structures of these inhibitors are shown in Figure 6a,b.

The experimental alignment of these two ligands was determined by performing a RMS fit of backbone atoms of the common HIV-1 reverse transcriptase host. The root-mean-square deviation (RMSD) for this fit was found to be 0.64 Å, showing that the topology of the host is not altered drastically by the presence of the different inhibitors. The resulting experimental alignment is shown in Figure 7(a).

The two ligands were extracted from their complexes and reoriented at random. An attempt was then made to realign the structure of nevirapine with that of alpha-APA. The result of this test is shown in Figure 7b; the RMSD of nevirapine from its alignment in the crystal structure was found to be 0.79 Å. From this figure we can see that the alignment technique has preferentially aligned one of the two common benzene residues at the expense of the remaining structure. To some extent this is to be expected from a partial structure alignment technique and could even be taken as evidence of the methodology's ability to seek out common fragments of molecules in the presence of considerable "noise". However, there is little doubt that the alignment of the two molecules is extremely good and the overlay is sufficiently good for most purposes.

A similar protocol was used with some larger molecules that inhibit the action of HIV-1 protease;[22,23] the structures of these inhibitors are shown in Figure 8a,b. In this case the aligned HIV-1 protease hosts exhibited an RMSD of 0.61 Å; the resulting experimental alignment of the inhibitors is depicted in Figure 9a. Figure 9b shows the results of realigning the two inhibitors from a randomly chosen starting position. In this case the alignment methodology has managed to reproduce the experimental alignment extremely closely, the RMSD of the ligand from Figure 8a relative to its alignment in the crystal structure being 1.05 Å.

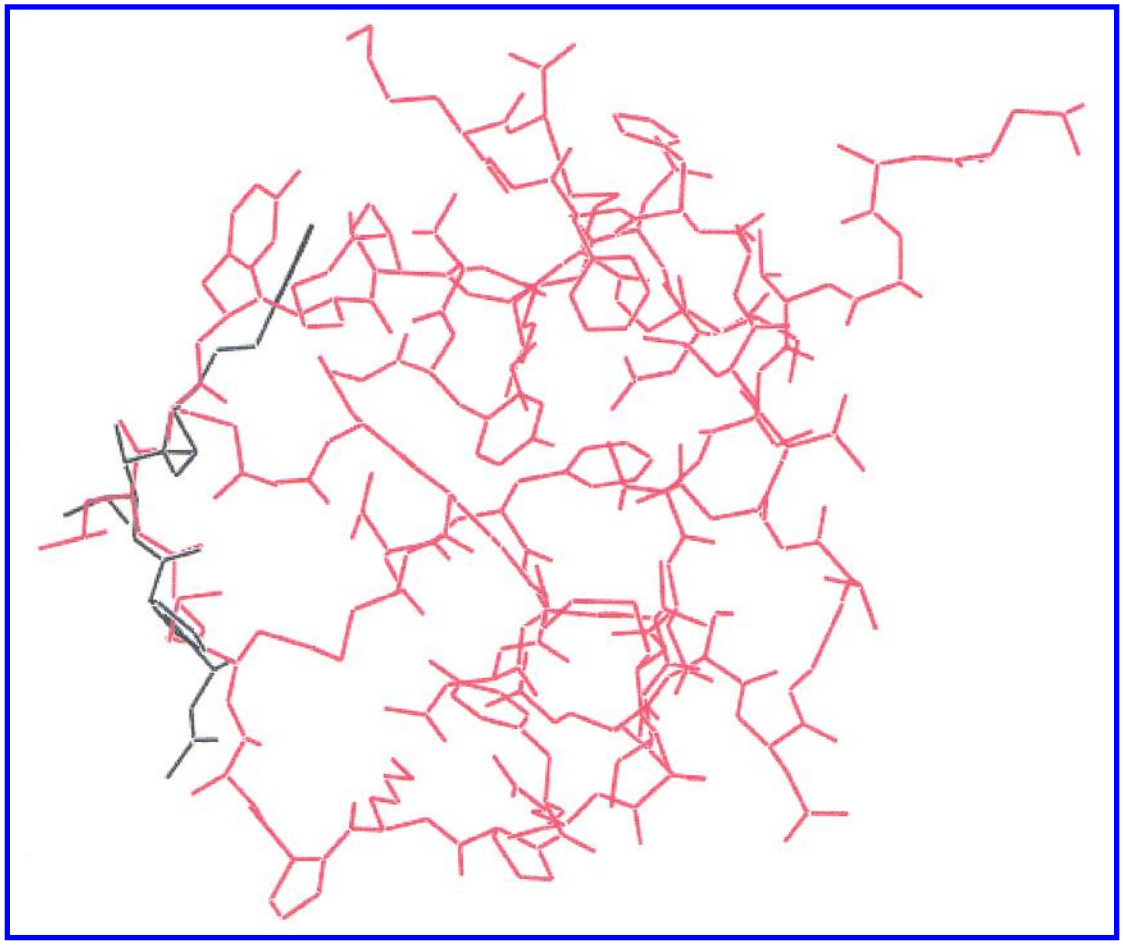

Figure 13. Result of the alignment between DFKi and the complete TOMI structure.

Comparison of a Naturally Occurring Inhibitor with Its Small Molecule Mimic. The final and most challenging test is to compare a naturally occurring inhibitor with a small-molecule mimic to see whether the technique is able to identify and align the small-molecule mimic with the active region of the natural inhibitor.

There have been several attempts to solve this problem, one of the earliest being introduced by Masek et al.[24] with the "molecular skins" approach. In this technique the shapes of the two molecules are represented by a molecular surface of finite thickness, the so-called molecular skins. The molecules are then translated and rotated in an attempt to maximize the intersection of these skins. While there is no doubting the ability of Masek's approach, it has come under criticism from Perkins et al.[25] for its lengthy run times. As an improvement Perkins et al. suggest following a stepwise alignment procedure. First, a set of alignments is generated, each of which exhibits a substantial overlap of the molecular surfaces; these alignments are then ranked by considering the additional requirements of hydrogen bonding and electrostatic similarity. By using a grid-based representation of the molecules under consideration, the technique is able to gain a substantial speed advantage over the molecular skins approach where the molecular surface overlaps are calculated analytically. Finally, Poirrette et al.[26] attempt to generate a simultaneous overlap of several molecular surface properties in order to align large protein structures. In their technique a Connolly surface is generated with additional descriptors showing whether a certain area is attributed with hydrogen-bond-accepting or -donating potential. A genetic algorithm is then used to form the transformation that maps one molecule onto the other with the maximum surface similarity.

In light of these techniques, we have chosen as our test set the comparison of the natural turkey ovomucoid inhibitor (TOMI) with the small DFKi mimic. This test set has become something of a standard for this type of algorithm since it was first considered by Masek et al., and results for all of the techniques described above exist for this test set.

The structures of TOMI bound to human leukocyte elastase[27] and DFKi bound to porcine pancreatic elastase[28] were obtained from the Protein Databank.[21] The excised structure of TOMI is displayed in Figure 10, with the active region highlighted; this region spans Ala15 to Tyr20.

The first test was to ensure that the technique was able to perform an alignment between the active fragment of TOMI and DFKi. It was found that the simple distance-based descriptors described above were not sufficiently general in their ability to represent a structure without setting an excessively large tolerance for two descriptors to be considered similar; doing this yielded a very large voting table that required considerable CPU time to process. Consequently the distances and positions used to form a footprint were rounded to the nearest 0.5 Å. Rounding descriptors in this manner is a standard technique for gaining a degree of generalization.[29] The results of this alignment are shown in Figure 11; as can be seen these are highly convincing. Masek found that these two structures were sufficiently similar for a similarity optimizer to yield a useful result. Visually there is little difference between the alignment generated by a

simple similarity optimizer and the result shown in Figure 11.

Having proved that the alignment technique was able to align the active fragment of TOMI and DFKi, it was necessary to show that the active fragment of TOMI could be identified within the overall TOMI structure. Generating footprints for every atom in the hydrogen-suppressed TOMI structure yielded an enormous amount of data that was not practical to process. To refine the search it was found necessary to limit the footprint generation process to those atoms that were within 10 Å of the α-C of Leu-18. Masek et al. used a similar criterion since Leu-18 was identified by mutagenesis data as being important for the activity of TOMI. In general, we might expect mutagenesis data to be available to give us some notion of which residues are important for the activity. Once this operation was carried out the alignment was facile; the results are shown in Figure 12.

Finally, an attempt was made to probe the structure of TOMI to see where the small-molecule mimic DFKi could be aligned, and whether this alignment corresponded to the active fragment in TOMI. This alignment proved to be the most difficult to carry out as despite our best efforts the optimizer occasionally found itself trapped on substantial local maxima of the scoring function. While these local maxima corresponded to quite plausible alignments of the two structures, they were not the desired result. However, through a series of tests commencing from different start positions, it was possible to determine that the global maxima of the scoring function correctly aligned DFKi with the active fragment of TOMI; this alignment is shown in Figure 13. In this alignment DFKi exhibited an RMSD of 0.6 Å relative to the alignment determined by overlaying the crystal structures of the DFKi and TOMI complexes.

CONCLUSIONS AND FURTHER WORK

The principal purpose of this paper was to illustrate a new methodology for the alignment of three-dimensional structures. Unlike most comparable alignment techniques, we have not separated the stages of feature matching and structural alignment. We have demonstrated how the combination of these two stages yields an elegant algorithm that has the ability to automatically select a group of features from the template and working molecules that are both chemically and geometrically comparable. As it stands our technique has potential applications in the development of more informative 3D-QSAR relationships, and almost certainly could be adapted for use as a search engine for a three-dimensional chemical database.

There is clearly still much work to be done. As the final example has shown, the descriptors currently used are only just adequate to provide discrimination between some structures. We feel that more advanced descriptors, probably based upon additional molecular properties such as electrostatic or hydrophobic parameters, will lead to greater discrimination and even more meaningful and rapid alignments. In addition, no explicit analysis of the effects of flexible groups has been undertaken. It seems reasonable to presume that a flexible molecule could be handled, at least to some extent, by considering a collection of the most probable conformations. These separate conformations could then be aligned to find the closest match for a given template. Alternatively there seems to be no obvious reason the torsional angles in the structures under consideration should not be added as parameters to be optimized, although this approach may be more computationally intensive than considering a collection of likely conformations. Furthermore, there is no apparent obstacle preventing the technique being utilized for docking two structures. All that would appear to be required is the selection of voting pairs that were in some way complementary rather than similar, such as favorable charge or hydrophobicity interactions. All of these topics are currently under investigation and will hopefully be reported upon in the not too distant future.

ACKNOWLEDGMENT

D.D.R. is supported by an EPSRC CASE studentship held in conjunction with Oxford Molecular Group PLC. This work was in part supported by the Wellcome Trust and the National Foundation for Cancer Research.

REFERENCES AND NOTES

(1) Klebe, G. Structural Alignment of Molecules. In *3D-QSAR in Drug Design*; Kubinyi, H., Ed.; ESCOM: Leiden, **1993**; pp 173−199.
(2) Carbo, R.; Leyda, L.; Arnau, M. An Electron Density Measure of the Similarity Between Two Compounds. *Int. J. Quantum Chem.* **1980**, *17*, 1185−1189.
(3) Hodgkin, E. E.; Richards, W. G. Molecular Similarity Based on Electrostatic Potential and Electric Fields. *Int. J. Quantum Chem. Quantum Biol. Symp*. **1987**, *14*, 105−110.
(4) Meyer, A. M.; Richards, W. G. Similarity of Molecular Shape. *J. Comput.-Aided Mol. Des*. **1991**, *5*, 426−439.
(5) Stouch, T. R.; Jurs, P. C. A Simple Method for the Representation, Quantification, and Comparison of the Volumes and Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci*. **1986**, *26*, 4−12.
(6) () Jones, G.; Willet, P.; Glen, R. C. A Genetic Algorithm for Flexible Molecular Overlay and Pharmacophore Elucidation. *J. Comput.-Aided Mol Des*. **1995**, *9*, 532−549.
(7) Jones, G.; Willet, P.; Glen, R. C. GAs for Chemical Structure Handling. In *Principles of QSAR and Drug Design 1: Genetic Algorithms in Molecular Modelling*; Devillers, J., Ed.; Academic Press: New York, **1996**; pp 211−242.
(8) Barequet, G.; Sharir, M. Partial Surface and Volume Matching in Three Dimensions. *IEEE Trans. Pattern Anal. Machine Intell*. **1997**, *19*, (9).
(9) Nussinov, R.; Wolfson, H. J. Efficient Detection of Three-Dimensional Structural Motifs in Biological Molecules by Computer Vision Techniques. *Proc. Natl. Acad. Sci. U.S.A. (Biophysics)* **1991**, *88*, 10495−10499.
(10) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule-Ligand Interactions. *J. Mol. Biol*. **1982**, *161*, 269−288.
(11) Kuhl, F. S.; Crippen, G. M.; Friesen, D. K. A Combinatorial Algorithm for Calculating Ligand Binding. *Comput. Chem*. **1984**, *5*, 24−34.
(12) The correlation function between two molecules A and B can be defined as $COR_{AB}(t_x, t_y, t_z) = \Sigma_x\Sigma_y\Sigma_z\rho_A(x, y, z)\rho_B(x + t_x, y + t_y, z + t_z)$, where ρ is a property of the molecule. To find the optimum translation we simply search for the values of (t_x, t_y, t_z) that maximize COR_{AB}.
(13) Havel, T. F.; Kuntz, I. D.; Crippen, G. M. The Theory and Practice of Distance Geometry. *Bull. Math. Biol*. **1983**, *45*, 665−720.
(14) Microsoft Corporation, One Microsoft Way, Redmond, WA 98052-5234.
(15) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C*, 2nd ed.; Cambridge University Press: New York, 1997.
(16) Richards, W. G. The Dominant Role of Shape Similarity and Dissimilarity in QSAR. *QSAR and Molecule Modelling: Concepts, Computational Tools and Biological Applications*; Sanz, F., Giraldo, J., Manaut, F., Eds.; Prous Science Publishers: Barcelona, **1995**; pp 364−373.
(17) *CaCHE*; Oxford Molecular Ltd.: Magdalen Science Park, Oxford, United Kingdom.
(18) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P.; A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc*. **1985**, *107*, 3902.
(19) Stewart, J. J. P. *Mopac 6.0. QCPE* **1990**, *455*.
(20) Ren, J.; Esnouf, R.; Garman, E.; Somers, D.; Kirby, C. R. I.; Keeling, J.; Darby, G.; Jones, Y.; Stuart, D. I.; Stammers, D. High Resolution

Structures of HIV-1 RT From Four RT-Inhibitor Complexes. *Nat. Struct. Biol.* **1995**, *2*, 293.

(21) Protein databank http://www.pdb.bnl.gov. The structures used for the HIV-1 RT test are stored as 1VRT and 1VRU. The structures used for the HIV-1 protease test are stored as 4HVP and 9HVP. The structure of TOMI used is stored as 1PPF. The structure of DFKi is stored as 4EST.

(22) Miller, M.; Schneider, J.; Sathyanarayana, B. K.; Toth, M. V.; Marshall, G. R.; Clawson, L.; Selk, L.; Kent, S. B. H.; Wlodawer, A. Structure of Complex of Synthetic HIV-1 Protease with a Substrate-Based Inhibitor at 2.3 Angstroms Resolution. *Science* **1989**, *246*, 1149.

(23) Erickson J.; Neidhart, D. J.; van Drie, J.; Kempf, D. J.; Wang, X. C.; Norbeck, D. W.; Plattner, J. J.; Rittenhouse, J. W.; Turon, M.; Widenburg, N.; Kohlbrenner, W. E.; Simmer, R.; Helfrich, R.; Paul, D. A.; Knigge, M. Design, Activity and 2.8 Angstroms Crystal Structure of a C2 Symmetric Inhibitor Complexed to HIV-1 Protease. *Science* **1990**, *249*, 527.

(24) () Masek, B. B.; Merchant, A.; Matthew, J. B. Molecular Skins: A New Concept for Quantitative Shape Matching of a Protein With Its Small Molecule Mimics. *Proteins: Struct., Funct. Genet*. **1993**, *17* (*2*), 193−202.

(25) Perkins, T. D. J.; Mills, J. E. J.; Dean, P. M. Molecular surface-volume and property matching to superpose flexible dissimilar molecules. *J. Comput.-Aided Mol. Des*. **1995**, *9*, 479−490.

(26) Poirrette, A. R.; Artymiuk, P. J.; Rice, D. W.; Willtett, P. Comparison of Protein Surfaces using a Genetic Algorithm. *J. Comput.-Aided Mol. Des*. **1997**, *11*, 557−569.

(27) Bode, W.; Wei, A.; Huber, R.; Meyer, E.; Travis, J.; Neumann, S. X-ray Crystal Structure of the Complex of Human Leukocyte Elastase (PMN Elastase) and the Third Domain of Turkey Ovomucoid Inhibitor. *EMBO J*. **1986**, *5*, 2453−2458.

(28) Takahashi, L. H.; Radhakrishnan, R.; Rosenfield, R. E.; Meyer, E. F.; Trainor, D. A. Crystal Structure of the Covalent Complex formed by Peptidyl α,α-Difluoro-β-keto Amide with Porcine Pancreatic Elastase at 1.78 Å Resolution. *J. Am. Chem. Soc*. **1981**, *111*, 3368−3374.

(29) Nilakantan, R.; Bauman, N.; Venkataraghavan, R. New Method for Rapid Characterisation of Molecular Shapes: Applications in Drug Design. *J. Chem. Inf. Comput. Sci*. **1993**, *33*, 79−85.

CI990272P