Journal of Computer-Aided Molecular Design, **15:** 587–612, 2001.
KLUWER/ESCOM

FLASHFLOOD: A 3D Field-based similarity search and alignment method for flexible molecules

Michael C. Pitman[a,*], Wolfgang K. Huber[b,†], Hans Horn[b], Andreas Krämer[b], Julia E. Rice[b] & William C. Swope[b]
[a]*IBM T. J. Watson Research Center, Yorktown Heights NY 10598, USA*
[b]*IBM Almaden Research Center, San Jose CA 95120*

Received 6 October 2000; accepted 20 April 2001

Key words: context-adapted similarity measure, drug design, field-based similarity searching, flexible alignment, fragment assembly, molecular fragmentation, molecular property fields, molecular similarity, molecular superposition, structural alignment

Summary

A three-dimensional field-based similarity search and alignment method for flexible molecules is introduced. The conformational space of a flexible molecule is represented in terms of fragments and torsional angles of allowed conformations. A user-definable property field is used to compute features of fragment pairs. Features are generalizations of CoMMA descriptors [1] that characterize local regions of the property field by its local moments. The features are invariant under coordinate system transformations. Features taken from a query molecule are used to form alignments with fragment pairs in the database. An assembly algorithm is then used to merge the fragment pairs into full structures, aligned to the query. Key to the method is the use of a context adaptive descriptor scaling procedure as the basis for similarity. This allows the user to tune the weights of the various feature components based on examples relevant to the particular context under investigation. The property fields may range from simple, phenomenological fields, to fields derived from quantum mechanical calculations. We apply the method to the dihydrofolate/methotrexate benchmark system, and show that when one injects relevant contextual information into the descriptor scaling procedure, better results are obtained more efficiently. We also show how the method works and include computer times for a query from a database that represents approximately 23 million conformers of seventeen flexible molecules.

Introduction

The problem of aligning a group of flexible molecules to a particular query conformation (molecular superposition) remains an important problem in computer-assisted drug design. Many approaches to this common end have been reported [2–16]. In the absence of structural information regarding the ligand-receptor or ligand-enzyme complex, structural alignment is a way of both elucidating important features responsible for activity [17, 18] and a means of finding new molecules with similar or better activity [3, 19, 11].

When one is attempting to elucidate spatial and chemical information about the nature of the host-ligand interaction, one often begins with the alignment of a series of active compounds based on some kind of alignment rule. Unfortunately, this process is riddled with difficulties and assumptions about the relevant conformations, relevant features, importance of internal strain, the role of hydrogen bonds, electrostatics, solvation, hydrophobicity, as well as more profound concerns such as whether compounds in a data set even bind at the receptor site via the same mechanism (for an enjoyable discussion, see [20]). It is clear that no single method for alignment will settle these

*Author to whom correspondence should be addressed: E-mail: pitman@watson.ibm.com
†Current address: DKFZ Heidelberg, Germany w.huber@dkfz-heidelberg.de

issues across widely varying contexts. Our interest, accordingly, is in a system that allows the incorporation of context-specific information to balance the above considerations in a manner suitable to the problem at hand. A consequence is that one-time calibrations of similarity measures are inappropriate. We hold that the basis of a similarity search should be tunable to the particular context being considered.

Several superposition methods reported are field-based [3, 4, 9, 13]. An attractive aspect of field-based approaches is the potential for incorporating high levels of electronic structure theory into the description of the field. Apart from the difficulties and expense of deploying high level quantum mechanical calculations, we regard the design of a system that can utilize the results of such calculations for use in similarity analysis as, at the very least, forward looking. The method we describe is aimed at providing a systematic way of treating field-based similarity searching, without being confined to a particular field definition. In the property field section, we describe a very simple example property field that we have chosen for this initial study. The point we would like to emphasize is that our method allows for fields ranging from simplistic, phenomenological fields, like the one employed here, to fields derived from quantum mechanical calculations.

Having said that, we turn to concerns of conformational space and the expense of quantum mechanical calculations. In general, the conformational space for drug-like molecules can become quite appreciable. Some methods represent the conformational space of a molecule as a collection of rigid fragments with preselected torsions [21]. Other approaches prepare a database of representative conformations [22, 23], or compute conformations on the fly [12]. The Fragment pairs section introduces our approach, based on fragmenting molecules into more manageable partitions. We introduce and motivate our choice for the smallest irreducible unit of characterization as the fragment *pair* rather than the fragment. We show how conformational space becomes more manageable with our treatment.

Setting out to address flexible superposition via potentially sophisticated property fields, we must give special attention to reducing the number of similarity evaluations that are to be performed, while maintaining some degree of confidence that the space has been covered. We have attempted to preprocess as much as possible, while still leaving enough tunability to adapt to the context of the investigation. We seek a practical tradeoff of the size of conformational space, and the need for fragment pairs as large as possible to maximize the relevance of the computed property fields.

Our approach [24] decomposes the conformational space of molecules to fragments. Then, to minimize boundary effects, we compute the property field on *pairs* of fragments. From the computed property fields of the fragment pairs, several *features* are sampled and stored. Features are generalizations of CoMMA descriptors [1] that characterize local regions of the property field by its local moments. They are invariant under coordinate system transformations. To query the database for molecules that are similar to a particular molecule, the *query molecule*, features are calculated for the query molecule, and fragment pairs that contain a sufficient number of similar features are retrieved. The key point is that, due to the coordinate system invariance of the features, the retrieval can happen *without* any alignment, or optimization over rotational and translational degrees of freedom. The alignment of retrieved fragment pairs on the query is determined by a pose clustering procedure from the individual feature-feature-correspondences. Finally, to construct full aligned candidate molecules, the retrieved fragment pairs are assembled by an incremental buildup procedure, similar in principle to ones used in docking [25] and *de novo* design [26]. Our goal in this paper is to outline the method and provide *proof of concept* of the algorithm by applying it to a small database of seventeen molecules, representing 23 million conformations.

Fragment pairs

There are two types of complexity in a database of three-dimensional molecular structures: first, the *conformational variety* of individual molecules, and second, in the case of a virtual combinatorial library, the *combinatorial variety* that results from the possibility of synthesizing a large number of different molecules from a small number of reagents. Generally, the total number of three-dimensional structures grows exponentially both with the number of rotatable bonds and the number of reagents. In this section we discuss how to efficiently represent and store such a database.

A molecule can be partitioned into a *fragment graph*. This is an acyclic graph that consists of *fragment nodes* connected by *rotatable bond edges*. Within a fragment node, there may be zero, one or several rotatable bonds, as well as other degrees of freedom such as ring conformations. Given a molecule, there

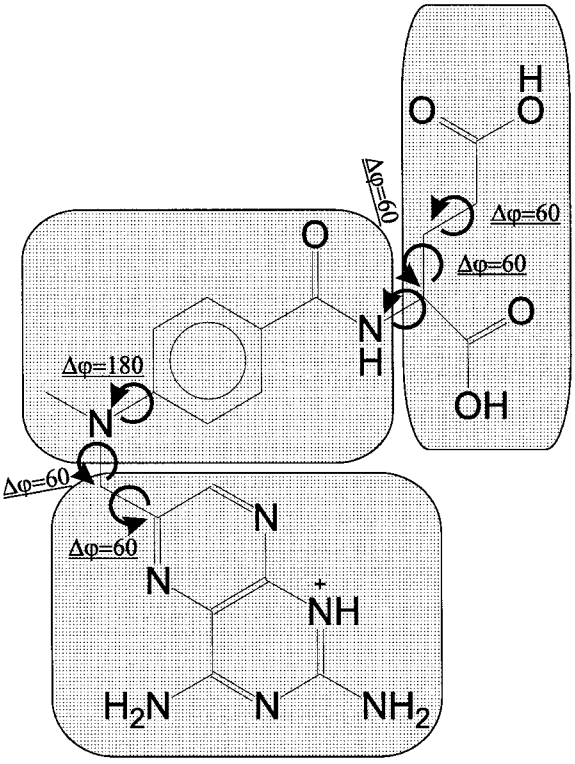

Figure 1. Fragmentation and conformational expansion of methotrexate. Single bonds are sampled with a resolution of $\Delta\varphi = 60°$, the aniline bond with $\Delta\varphi = 180°$. The molecule is partitioned into three fragments.

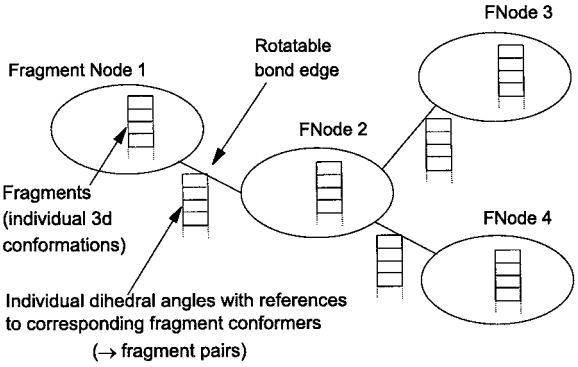

Figure 2. Schematic representation of a fragment graph.

are in general multiple possible ways of partitioning it into a fragment graph. Typically, we have in mind that a fragment node consist of about 10 heavy atoms, for example an aromatic ring plus some substituents. A sample partition is shown in Figure 1. The substructure represented by a fragment node can in general assume several different conformations. We call a specific conformation of a fragment node a *fragment*. A *fragment pair* consists of two neighboring fragments connected by a rotatable bond at a specific dihedral angle. A schematic representation of a fragment graph is depicted in Figure 2.

Fragment pairs are the fundamental entities of our approach: property fields are defined and calculated on fragment pairs, and the similarity search is based on rotationally invariant features that are calculated from those property fields. Property fields will be discussed in The property field section. The assumption that underlies our use of fragment pairs is that a property field calculated from an isolated fragment pair is a good local approximation of the property field of the corresponding region of the composite molecule.

Conceptually, the use of fragment pairs is equivalent to using overlapping fragments, with the overlap being about half their size. This has advantages both for the recognition and for the assembly steps. First, the fragmentation locally distorts the property field in those places where the molecule is cut. By using fragment pairs, the regions around the fragment joints are always in the interior of at least one fragment pair, such that meaningful local descriptors can be calculated for them. Further, an aligned database molecule is constructed by assembling fragment pairs that have one fragment in common, and which both locally match the query with compatible orientations. Thus, the relevant dihedral range of the connecting rotatable bonds – determined, e.g., from steric and energetic criteria – is already available in the precomputed fragment pairs.

Obviously, this approach is suitable both for conventional molecule libraries, as well as for virtual libraries supporting combinatorial chemistry approaches. The efficiency of the fragment pair representation is best discussed by way of an example. As indicated in Figure 1, the conformational space of methotrexate is spanned essentially by six rotatable bonds. If five of them are sampled in steps of 60°, and one at 180°, the total number of conformations is

$$C_{\text{mol}} = 6^5 \times 2 = 15552. \tag{1}$$

In practice, fewer conformations have to be considered because some are sterically forbidden. Similar to (1), C_{fp1}, the number of fragment pairs from the lower and middle fragment node, and C_{fp2}, the number of fragment pairs from the middle and upper fragment node are

$$\begin{aligned} C_{\text{fp1}} &= 6 \times 6 \times 2 = 72, \\ C_{\text{fp2}} &= 2 \times 6 \times 36 = 432. \end{aligned} \tag{2}$$

The total number of fragment pairs is therefore $72 + 432 = 504$. Furthermore, the size of a fragment pair is in this example only about $\frac{2}{3}$ of the size of the whole molecule, with a corresponding smaller number of local descriptors, so that the fragment pair representation in this case needs about 50 times less storage than the brute force enumeration.

More generally, for a molecule consisting of n fragments, each of which has C_{frag} conformations, and which are connected by rotatable bond edges that are sampled in C_{rbe} steps, the total number of conformations is

$$C_{\text{mol}} = C_{\text{frag}}^{n} \times C_{\text{rbe}}^{n-1}, \tag{3}$$

where $n - 1$ is the number of rotatable bond edges. In comparison, the total number of fragment pairs is

$$C_{\text{fp}} = (n-1) \times C_{\text{frag}}^{2} \times C_{\text{rbe}} \tag{4}$$

Note that (3) grows exponentially with n, whereas (4) only depends linearly on n.

The property field

The basis for the three-dimensional similarity searching and alignment are two property fields $\mu(\mathbf{r})$ and $\rho(\mathbf{r})$. The method makes no assumptions about these fields, except for the requirement that $\mu(\mathbf{r})$ is scalar and positive. The following presentation will also assume that $\rho(\mathbf{r})$ is a scalar field, but this can be straightforwardly extended to multiple scalar, vector or tensor fields.

Both fields $\mu(\mathbf{r})$ and $\rho(\mathbf{r})$ are used to identify similar regions in query and database molecules. Their geometrical alignment however is performed solely on the basis of the field $\mu(\mathbf{r})$.

A simple property field can be defined as:

$$\begin{aligned} \mu(\mathbf{r}) &= \frac{1}{\left(\sqrt{2\pi}\sigma\right)^3} \sum_{j=1}^{N_{\text{atoms}}} A_j \exp \frac{-\left(\mathbf{r}-\mathbf{a}_j\right)^2}{2\sigma^2}, \\ \rho(\mathbf{r}) &= \mu(\mathbf{r}) - \overline{\mu}. \end{aligned} \tag{5}$$

Here, the jth atom is located at $\mathbf{a}_j$, and its electronegativity, A_j is given according to the Allred scale [27]. We used the following values: $A_{\text{C}} = 2.6$, $A_{\text{O}} = 3.4$, $A_{\text{N}} = 3.0$, $A_{\text{H}} = 2.2$, $A_{\text{P}} = 2.2$, $A_{\text{S}} = 2.6$, $A_{\text{F}} = 4.0$, $A_{\text{Cl}} = 3.2$. σ is a parameter that controls the range of the Gaussian smearing function. Throughout the present work, $\sigma = 0.5$ Å was used. The rationale was to choose a value as big as possible, but small enough for the property field not to be too uniform and nonspecific, $\overline{\mu}$ is the average of $\mu(\mathbf{r})$ over all space. $\mu(\mathbf{r})$ is positive and $\rho(\mathbf{r})$ is analogous to a neutral charge distribution.

Another possible choice of a property field is [1]

$$\begin{aligned} \mu(\mathbf{r}) &= \frac{1}{\left(\sqrt{2\pi}\sigma\right)^3} \sum_{j=1}^{N_{\text{atoms}}} M_j \exp \frac{-\left(\mathbf{r}-\mathbf{a}_j\right)^2}{2\sigma^2} \\ \rho(\mathbf{r}) &= \frac{1}{\left(\sqrt{2\pi}\sigma\right)^3} \sum_{j=1}^{N_{\text{atoms}}} Q_j \exp \frac{-\left(\mathbf{r}-\mathbf{a}_j\right)^2}{2\sigma^2}, \end{aligned} \tag{6}$$

where M_j is the atomic mass and Q_j is the atomic charge computed by considering the 'fraction of ionic character' of each bond in the molecule, a model described, for example, in reference 28.

These example fields are by no means intended to offer new insight into processes that underlie the chemistry of the presented applications. We have chosen the property field defined by Equation 5 for the application presented in the application section. Purely because of its simplicity, in order to exemplify and prototype the method. Still, despite its simple-mindedness, it performs adequately for the present application, and thus may serve in the future as a base level, against which more elaborate fields may be benchmarked.

Obviously, the choice of the property fields will have a great effect both on the selectivity and on the efficiency of the search. The point is that the preferred choice depends on the application and on the questions asked. The exploration of different alternative fields is intended to be part of the process of adapting the method to a certain domain.

Possible fields are by no means restricted to 'smeared out' atomic properties. The following method is prepared and intended to make use of fields derived from quantum mechanical calculations.

Descriptors and feature generation

Given the property fields $\mu(\mathbf{r})$ and $\rho(\mathbf{r})$, we now construct a set of local, rotationally invariant, moment-based descriptors. If the property fields of the query molecule and a database fragment pair are similar, then these descriptors will have similar values. Since the descriptors are rotationally invariant, no alignment is necessary, and the comparison can be performed very quickly.

The similarity of the descriptors alone is a necessary, but not a sufficient criterion for the similarity of the fields. However, together with the descriptors we

store information on their relative positions and orientations within the query and database structures. When a database structure has enough descriptors similar to the query, the relative positions and orientations of the descriptors are compared. Only if these are also consistent, the two structures are considered similar, and an approximate alignment is deduced from this information. Note that in order to obtain the alignment, no explicit (and costly) optimization of a property field overlap function with respect to translation and rotation operators has to be performed. However, if desired, such an optimization can afterwards be applied to a small set of promising candidates, starting from near-optimal initial conditions. These ideas will be explained in more detail later.

The first step in the construction of the descriptors is the partitioning of the volume occupied by the structure into overlapping *scoops*. If the property fields are defined by smearing out atomic properties A_j, as is the case in the two examples in the Property field section, then this is done as follows: Let $\{\vec{s}_k\}$ be a set of points within or around the structure, such that the spheres with radius R around these points provide a highly overlapping coverage of the relevant regions of the property fields. We call each of these spheres a scoop. Define a ramping function

$$\Delta(a) = \begin{cases} 1 - a/R & a \le R, \\ 0 & a > R \end{cases} \tag{7}$$

and a window function

$$h(r) = \begin{cases} 1 & r \le R, \\ 0 & r > R. \end{cases} \tag{8}$$

Then the 'attenuated atomic properties' that contribute to the kth scoop are given by

$$A_j^{(k)} = \Delta\left(\left|\mathbf{a}_j - \mathbf{s}_k\right|\right) \cdot A_j. \tag{9}$$

and the k-th scoop's property field is

$$\mu_k(\mathbf{r}) = h\left(|\mathbf{r} - \mathbf{s}_k|\right) \cdot \frac{1}{\left(\sqrt{2\pi}\sigma\right)^3} \sum_j^{N_{\text{atoms}}} A_j^{(k)} \exp \frac{-\left(\mathbf{r} - \mathbf{a}_j\right)^2}{2\sigma^2}, \tag{10}$$

and correspondingly for $\rho_k(\mathbf{r})$, as in Equation 5. The intention of the ramping function 7 is that the property fields μ_k and ρ_k (and therefore the descriptors) are continuous functions of the location of the scoop center s_k.

For general property fields $\mu(\mathbf{r})$ and $\rho(\mathbf{r})$, that are not obtained by smearing out atomic properties, scoop property fields can simply be obtained by setting

$$\mu_k(\mathbf{r}) = h\left(|\mathbf{r} - \mathbf{s}_k|\right) \cdot \mu(\mathbf{r}), \tag{11}$$

and correspondingly for $\rho_k(\mathbf{r})$.

In the present work, the set of points $\mathbf{s}_k$ was the set of all atom positions, and $R = 3$ Å. With this choice, the scoops are objects of intermediate size, larger than a functional group, but smaller than a fragment pair. Typically, a scoop contains 6 to 8 non-hydrogen atoms.

Having defined a set of scoops and their associated local property fields μ_k and ρ_k, we now construct rotationally invariant descriptors. There will be a descriptor consisting of 16 real numbers for each scoop. To simplify the notation, in the following we drop the index k that numbers the scoops, and replace the continuum fields by the discretized versions μ_i and ρ_i, defined on a grid of points $\{\mathbf{r}_i \mid i = 1, \ldots, N\}$. For the grid, we use a face-centered cubic lattice [29] of unit cell length $\Delta R = R/18$ within each scoop of radius R. The grid spacing ΔR has been determined by varying the grid orientation with respect to the atoms in the scoop, and making sure that the resulting descriptors, as described below, do not significantly depend on the orientation.

The zeroth moments of the fields are

$$M = \sum_{i=1}^{N} \mu_i, \quad Q = \sum_{i=1}^{N} \rho_i. \tag{12}$$

In the case of the property field defined by Equation 6, M and Q correspond to total mass and total charge within the scoop.

The center of the μ-field is defined by

$$\mathbf{c}_\mu = \frac{1}{M} \sum_{i=1}^{N} \mu_i \mathbf{r}_i, \tag{13}$$

and the center of the ρ-field by

$$\mathbf{c}_\rho = \begin{cases} \frac{1}{Q}\mathbf{b} & \text{if } |Q| > \varepsilon_Q, \\ \frac{1}{3b^2}\left(\mathbf{B}\mathbf{b} - \frac{\mathbf{b}^t\mathbf{B}\mathbf{b}}{4b^2}\mathbf{b}\right) & \text{otherwise,} \end{cases} \tag{14}$$

where $\mathbf{b}$ and $\mathbf{B}$ are dipole and quadrupole moment of the ρ-field with respect to the origin of the laboratory

coordinate system,

$$\mathbf{b} = \sum_{i=1}^{N} \rho_i \, \mathbf{r}_i, \quad \mathbf{B} = \sum_{i=1}^{N} \rho_i \left(3\mathbf{r}_i \, \mathbf{r}_i{}^t - r_i^2 \mathbb{1} \right). \tag{15}$$

and the superscript t stands for transposition. In the case of the property field defined by Equation 6, $\mathbf{c}_\mu$ is the center of mass, and the first line of Equation 14 is the center of charge. The center of charge is only defined if the scoop has a net charge, i.e., if the charge is larger than the threshold ε_Q. Otherwise, the second line of Equation 14 calculates the center of dipole [1].

The inertial tensor $\mathbf{J}$ and a cubic vector $\mathbf{j}$ with respect to $\mathbf{c}_\mu$, the center of μ, are defined as

$$\mathbf{J} = \sum_{i=1}^{N} \mu_i \left(r_i'^2 \, \mathbb{1} - \mathbf{r}_i{}' \, \mathbf{r}_i{}'^t \right) \tag{16}$$

$$\mathbf{j} = \sum_{i=1}^{N} \mu_i \, r_i'^2 \, \mathbf{r}_i{}', \tag{17}$$

where $\mathbf{r}_i{}' = \mathbf{r}_i - \mathbf{c}_\mu$. Similarly, we define dipole moment $\vec{p}$ and quadrupole moment $\mathbf{Q}$ of the ρ-field with respect to $\mathbf{c}_\rho$, the center of ρ:

$$\vec{p} = \sum_{i=1}^{N} \rho_i \, \mathbf{r}_i{}'' \tag{18}$$

$$\mathbf{Q} = \sum_{i=1}^{N} \rho_i \left(3\mathbf{r}_i{}'' \, \mathbf{r}_i{}''^t - r_i''^2 \mathbb{1} \right), \tag{19}$$

where $\mathbf{r}_i{}'' = \mathbf{r}_i - \mathbf{c}_\rho$. The point is now that we express the quantities 17–19 in a uniquely defined scoop-internal coordinate system, so that they no longer depend on the arbitrary choice of the laboratory frame. Therefore, the descriptors of different scoops can be compared without prior alignment. The axes of the scoop-internal coordinate system are given by the eigenvectors of the inertial tensor $\mathbf{J}$:

$$\mathbf{J}\mathbf{v}_n = J_n \mathbf{v}_n, \quad n = 1, 2, 3. \tag{20}$$

The positive numbers J_n are the inertial moments. The vectors $\mathbf{v}_n$ can be arranged into the columns of an orthonormal matrix $\mathbf{V}$. An arbitrary vector is then transformed from the laboratory frame to the internal frame by left-multiplying it with $\mathbf{V}^t$.

In order to uniquely define the internal coordinate system, we need to fix (i) the ordering, and (ii) the sensing (i.e., the signs) of the coordinate axes.

The ordering is defined by $J_1 \leq J_2 \leq J_3$. If two or three of the eigenvalues of $\mathbf{J}$ are degenerate, then there are no unique eigenvectors, but rather two- or three-dimensional eigenspaces, there is no well-defined inertial coordinate system, and the corresponding scoop is not used to generate a descriptor. Degeneracy in this context means that two eigenvalues are equal to within some threshold that may depend on the choice of property field. The degeneracy condition is

$$J_2/J_1 < 1 + \epsilon_J \quad \text{or} \quad J_3/J_2 < 1 + \epsilon_J. \tag{21}$$

To fix the sensing, we define the asymmetry vector

$$\boldsymbol{\alpha} = \mathbf{V}^t \mathbf{j}, \tag{22}$$

where $\mathbf{j}$ is the cubic vector from Equation 17, and choose the signs of the axes according to the following table:

$\lvert\alpha_1\rvert \geq \lvert\alpha_2\rvert \geq \lvert\alpha_3\rvert$	$\mathbf{V}$ is replaced by $(\operatorname{sgn}\alpha_1 \cdot \mathbf{v}_1, \operatorname{sgn}\alpha_2 \cdot \mathbf{v}_2, s \cdot \mathbf{v}_3)$
$\lvert\alpha_1\rvert \geq \lvert\alpha_3\rvert > \lvert\alpha_2\rvert$	$\mathbf{V}$ is replaced by $(\operatorname{sgn}\alpha_1 \cdot \mathbf{v}_1, s \cdot \mathbf{v}_2, \operatorname{sgn}\alpha_3 \cdot \mathbf{v}_3)$
$\lvert\alpha_2\rvert \geq \lvert\alpha_1\rvert \geq \lvert\alpha_3\rvert$	$\mathbf{V}$ is replaced by $(\operatorname{sgn}\alpha_1 \cdot \mathbf{v}_1, \operatorname{sgn}\alpha_2 \cdot \mathbf{v}_2, s \cdot \mathbf{v}_3)$
$\lvert\alpha_2\rvert \geq \lvert\alpha_3\rvert > \lvert\alpha_1\rvert$	$\mathbf{V}$ is replaced by $(\operatorname{sgn}\alpha_1 \cdot \mathbf{v}_1, s \cdot \mathbf{v}_2, \operatorname{sgn}\alpha_3 \cdot \mathbf{v}_3)$
$\lvert\alpha_3\rvert \geq \lvert\alpha_1\rvert \geq \lvert\alpha_2\rvert$	$\mathbf{V}$ is replaced by $(s \cdot \mathbf{v}_1, \operatorname{sgn}\alpha_2 \cdot \mathbf{v}_2, \operatorname{sgn}\alpha_3 \cdot \mathbf{v}_3)$
$\lvert\alpha_3\rvert \geq \lvert\alpha_2\rvert > \lvert\alpha_1\rvert$	$\mathbf{V}$ is replaced by $(s \cdot \mathbf{v}_1, \operatorname{sgn}\alpha_2 \cdot \mathbf{v}_2, \operatorname{sgn}\alpha_3 \cdot \mathbf{v}_3)$

The sign $s \in \{-1, 1\}$ is determined by requiring right-handedness,
i.e. $\det \mathbf{V} = 1$. The table above represents a unique choice of the axes sensing depending only on the cubic vector $\mathbf{j}$.

If two, or three components of α are within some chosen threshold of zero, one is left with two or even four different equally admissible choices for the axes sensing. For descriptors stored in the database, we simply take an arbitrary choice. For the query features that are to be matched against the database, a descriptor is generated for each admissible choice of axes sensing. For the present data, we used the duplication condition $\alpha_n / R J_n < 0.02$.

As a result, the CoMMA descriptor X_k for the kth scoop, which we call a *feature*, consists of the following $d = 16$ real numbers:

Quantity	Components
Total 'mass' M	1
Total 'charge' Q	1
Sorted eigenvalues $J_1 < J_2 < J_3$ of the inertial tensor $\mathbf{J}$	3
Dipole vector $\mathbf{V}^t \vec{p}$	3
Quadrupole tensor $\mathbf{V}^t \mathbf{Q} \mathbf{V}$	5
Spatial vector between the centers $\mathbf{V}^t (\mathbf{c}_\mu - \mathbf{c}_\rho)$	3

If the ρ-field is more complex than a scalar field, e.g., a combination of several scalars, or a vector or a tensor field, then the first and second moments **p** and **Q** will have a correspondingly greater number of components, and d, the number of components of the descriptor X_k is larger than 16.

Since the vector and tensor quantities in the descriptor are expressed in the internal coordinate system, which only depends on local properties of the molecule, the descriptor is completely independent of the laboratory frame. This has two crucial implications: first, two scoop descriptors can be compared against each other without prior alignment, and second, if two descriptors are found to be similar, the two corresponding molecules can be locally aligned by simply overlaying the internal coordinate systems of the two scoops.

Descriptor scaling and quantization

The scoop descriptors $X_k = (X_{k1}, \ldots, X_{kd})$ cannot be used straightforwardly to define similarity between scoops. This is both because the components $X_{k1}, X_{k2}, \ldots$ have different physical units, *and* because the different components may be more or less important for the similarity search at hand.

For example, it may be important in a particular application of this method to recognize and distinguish between some specific set of structural motifs or chemical functional groups in identifying fragment pairs from a database to align on a query molecule. In a different application of the method, it may be more important instead to recognize and distinguish a different set of motifs, or regions of electrical polarity. Clearly, a useful method should be able to *adapt* to what is meaningful to the user in assessing similarity.

We therefore define a distance measure that weights the different descriptors according to their importance as derived from a training set of descriptors. The training data consists of a number of sample descriptors that are categorized into *groups*. From this, the method 'learns' two types of descriptor variation: important variations are those that occur systematically between descriptors from different groups. They are used to define the distance measure. The descriptors within a group are considered similar, and the distance measure is made to ignore this type of variations. While there is a vast repertoire of methods from the disciplines of classification and pattern recognition to address this task, we presently simply use Fisher's linear discriminant analysis [30]. It provides a linear mapping from the d-dimensional descriptor space into a lower-dimensional space:

$$Y_k = W^t X_k, \tag{23}$$

where W is a $d \times p$ matrix with $p \leq d$. The discriminant matrix W is calculated from the user-supplied classification of a sample of descriptors X_k, into $p+1$ groups. The within-group scatter matrix S_w is the average covariance matrix of descriptors that are in the same group, and the between-group scatter matrix S_b is the covariance of the group centroids [30]. The discriminant matrix W is defined through the maximization of the criterion function

$$\mathcal{J}(W) = \frac{|W^t S_b W|}{|W^t S_w W|}, \tag{24}$$

This optimization criterion selects W such that the distance between the Y_k within the same group are minimized, whereas the distances between the groups' centroids are maximized. Numerically, W is calculated using a generalized eigenvalue routine [31].

To allow for a fast database lookup of stored descriptors that are similar to a query, we quantize descriptor space. Descriptors that fall within the same compartment are considered similar, and are matched. Those that fall into different compartments do not match. Presently, the compartments are defined by a rectilinear grid in *discriminant space*, i.e., in the space of the Y_k. The grid spacings s_j are chosen such that a grid cell can accommodate the typical scatter within a group. This is done by a heuristic that sets the bin width s_j to four times the largest standard deviation of the j-th component ($j = 1, \ldots, p$) within an individual group, and the right end of the leftmost bin to the minimum over all groups. A sample distribution of features in discriminant space is shown in Figure 5.

The present scaling and quantization method is fairly simple and robust, but it has a number of shortcomings. The main restrictions are the linearity of the mapping (23), the fact that the quantization cells all have the same size and shape, and the possibility of missing scoop similarities because the descriptors happen to fall across a cell boundary. Improved methods will be investigated in future work.

Nevertheless, the present method does accomplish a context-adapted descriptor calibration and similarity measure, utilizing a user-defined training set of descriptors. Missed matches as a consequence of cell boundaries are not as likely in the low dimension in which the grid is defined as they would be for

higher-dimensional situations. Since there is redundancy in the set of scoops characterizing a molecule, a meaningful alignment will be recovered if just a fraction of them are actually matched. A virtue of linear mappings, like (23), in comparison to, for example, neural networks, is that they have a limited number of parameters and require only a relatively small training set, which is important for practical applications. Furthermore, the mapping (23) and the subsequent discretization are exactly invariant under overall linear transformations of the descriptors. This means that the calibration and quantization scheme does not depend on whether, for example, length is measured in meter or in Angstrom, and mass in gram or amu.

Group definition

The training set is chosen according to the following criteria: the different groups should be selected to contain examples of structural or functional units that are relevant for the similarity search. Within the individual groups, members should represent the kinds of variation that occur in the descriptor database, like experimental uncertainty in bond lengths and angles, different environments of functional groups, and different conformations deemed irrelevant for the problem at hand. An example for this is presented in the Application to Dihydrofolate and Methotrexate section.

The feature database

The previous section gave us a context-adapted mapping from descriptor space to an integer set of feature keys. These keys are simply the indices that label the cells in discriminant space. The use of a hash table and fast integer key lookup methods supports efficient queries against large databases, with access times largely independent of database size.

In principle, compared to a method that would use a detailed, fully continuous distance metric, the quantization of the descriptors to produce the keys causes a loss of sensitivity (more false negatives), and a loss of selectivity (more false positives) in the similarity search. This is because true positives could be lying just across the bin boundaries and similarly true negatives could still be in the same bin but almost a bin width away. However, the set of descriptors that describe a molecule is highly redundant, so that an incomplete set of scoop matches still leads to a complete alignment of the molecules. Furthermore, false positive matches are reduced in the subsequent clustering step (see Feature correspondence section), since they do not occur consistently.

The *feature database* is the product of two inputs: a set of descriptors, generated from all conformations of a selection of fragment graphs, and the context-adapted descriptor-to-key mapping. The feature database is a hash table, whose keys are given by evaluating the mapping on the descriptors. An entry in the hash table consists of a reference back to a fragment pair, and a description of the internal frame associated with the descriptor. The explicit values of the descriptors are not stored in the feature database. The thrifty use of memory by the feature database is crucial for the scaleability of our method.

The calculation of the descriptor set for all the fragment pairs of the fragment graphs is generally the most expensive part of the method, however the descriptor set needs to be computed only once, in a preprocessing step, and is then stored. Application of the descriptor-to-key mapping is fairly cheap and quick. As the domain context is varied, the descriptor-to-key mapping will change, and different feature databases can be created to query against. Finally, a given feature database can be queried very rapidly from the keys of query features, with any number of queries, including differing conformations of the same molecule, as well as differing molecules.

Feature correspondence

Having prepared the feature database, we now describe how it is used to align fragment pairs to the query molecule. The basic idea is to calculate scoops, descriptors, and keys for the query in the same way as for the fragment pairs stored in the database. Each query key then accesses the matching entries in the feature database. Each pair of query and database scoops that have the same key is called a correspondence. By overlaying the internal coordinate axes of such a scoop pair, each correspondence implies a certain alignment of a stored fragment pair onto the query. Whereas such a single correspondence might be coincidental, a significant alignment – which we call a *hypothesis* – is inferred when several independent correspondences from different regions of the query and fragment pair support the same relative orientation of the two. This assumption is the basis of pose-clustering methods [32]. A selected number of the strongest hypotheses will be passed on to the next step, the assembly, described in the Assembly section.

The feature correspondence process therefore consists of three steps: construction of all correspondences by keyed access from the query to the feature database, clustering of the correspondences, and construction of the hypotheses as average alignments of the significant clusters.

The internal coordinate system that is associated with a feature is specified by its origin $\mathbf{c}_\mu$ and an orthonormal system of inertial eigenvectors $\mathbf{V}$, see Equations 13 and 20. The transformation of laboratory frame coordinates $\mathbf{x}$ into internal coordinates is given by

$$T : \mathbf{x} \mapsto \mathbf{V}^t(\mathbf{x} - \mathbf{c}_\mu). \tag{25}$$

Given a correspondence between a query feature X_q and a stored feature X_s, and the two associated transformations T_q and T_s, the transformation that aligns the stored fragment pair coordinates onto the query is

$$T_{qs} = T_q^{-1} \circ T_s : \mathbf{x} \mapsto \mathbf{c}_{\mu,q} + \mathbf{V}_q \mathbf{V}_s^t (\mathbf{x} - \mathbf{c}_{\mu,s}). \tag{26}$$

This is a putative alignment of the fragment pair represented in the database onto the query molecule, based on a single feature correspondence. It would be impractical and prohibitively expensive to evaluate and score each such alignment separately. In the next step, therefore, the set of all putative alignments is divided into clusters. In order to perform clustering, a metric in the space of transformations 26 is required. We are using

$$\begin{aligned} d(T_{qs}, T_{q's'}) &= \left| T_{qs}(\mathbf{x}_0) - T_{q's'}(\mathbf{x}_0) \right| \\ &\quad + 2\alpha \tan\left(\tfrac{1}{2} d_{\text{rot}}\right) \qquad (27) \\ d_{\text{rot}} &= \operatorname{acos}\left(\tfrac{1}{2}(\operatorname{Tr}\{\mathbf{V}_{q'} \mathbf{V}_{s'}^t \mathbf{V}_s \mathbf{V}_q^t\} - 1)\right) \qquad (28) \end{aligned}$$

if s and s' are from the *same* fragment pair, and $d(T_{qs}, T_{q's'}) = \infty$ if they are from different fragment pairs. $\mathbf{x}_0$ is the center of geometry of the fragment pair coordinates, and the first term on the right hand side of Equation 27 is simply the Euclidean distance between the transformed fragment pair centers. $\mathbf{V}_{q'} \mathbf{V}_{s'}^t \mathbf{V}_s \mathbf{V}_q^t$ is the rotation part of the transformation $T_{q's'} \circ T_{qs}^{-1}$ that maps a set of coordinates transformed by T_{qs} onto the one transformed by $T_{q's'}$, and d_{rot} is the magnitude of the angle of that rotation. The function $f(x) = 2 \tan \frac{x}{2}$ is close to zero for small angles, but becomes large when x approaches π. We use it to make sure that transformations with very different rotations are considered very far apart. The parameter α provides a measure of the relative weighting of orientation and translation for the transformation distance. Here, we chose $\alpha = 3$ Å for all fragment pairs.

Having defined a metric, we can now proceed to the clustering step. We use the hierarchical clustering method G03ECF implemented in the NAG library [33], with the complete-linkage distance updating method. Hierarchical clustering was chosen over partitional, because the latter is cumbersome for non-Euclidean metrics like (27), and the complete-linkage method because it produces the most compact clusters, which is desirable for the subsequent averaging [34]. The only parameter of the clustering step is the distance level d_{clust} at which the dendrogram is cut. For the data presented in Application to Dihydrofolate and Methotrexate, after some investigation [35], $d_{\text{clust}} = 3$ Å was used. Note that if d_{clust} is too large, what should be distinct clusters will be merged and the average transformation will not be representative of any transformation of these distinct clusters. On the other hand, if d_{clust} is too small, the number of members in each cluster is small and no clusters emerge with sufficient signal.

In the hypothesis building step, the largest and therefore most significant clusters are selected, and average transformations are calculated. Consider a cluster of n transformations $T_1, \ldots, T_n$, each one represented by

$$T_i : \mathbf{x} \mapsto \mathbf{t}_i + \mathbf{R}_i \mathbf{x}, \qquad i = 1, \ldots, n. \tag{29}$$

Representations (29) and (26) are related by $\mathbf{R} = \mathbf{V}_q \mathbf{V}_s^t$ and $\mathbf{t} = \mathbf{c}_{\mu,q} - \mathbf{R}\mathbf{c}_{\mu,s}$. The average rotation is calculated as

$$\mathbf{R}_{\text{av}} = \mathbf{U}\mathbf{W}, \qquad \text{where } \mathbf{U}\mathbf{D}\mathbf{W} = \sum_{i=1}^{N} \mathbf{R}_i \tag{30}$$

is a singular value decomposition (SVD) of $\sum_i \mathbf{R}_i$ [36]. This is well-defined as long as the rotations in the cluster are not too different. In the situation where rotations vary widely, the notion itself of an average rotation becomes disputable. The average translation vector $\mathbf{t}_{\text{av}}$ is calculated by requiring that the fragment pair center $\mathbf{x}_0$ under the average transformation is the arithmetic average of the fragment pair center under the individual transformations in the cluster,

$$\mathbf{t}_{\text{av}} = \frac{1}{n} \sum_{i=1}^{n} [\mathbf{t}_i + \mathbf{R}_i \mathbf{x}_0] - \mathbf{R}_{\text{av}} \mathbf{x}_0. \tag{31}$$

The average transformation for the cluster is therefore $T : \mathbf{x} \mapsto \mathbf{t}_{\text{av}} + \mathbf{R}_{\text{av}} \mathbf{x}$.

In the present work, all feature correspondences carry an equal weight both in the clustering and averaging steps. The consideration of some correspondences being more significant than others will be picked up in future work.

Summing up, for each query feature, the descriptor-to-key mapping forms a correspondence with all stored features that have the same key value. The set of all such pairs is clustered according to the metric in transformation space. There might be correspondences involving many different stored fragment pairs, but the correspondences within one cluster only refer to the same fragment pair. The largest clusters, according to a user-defined minimum cluster size n_{clust} are selected. For each of them, the average transformation is calculated and applied to the coordinates of the fragment pair whose features it was derived from. The result is a set of hypotheses, that is, fragment pairs aligned on the query.

Assembly

The assembly process begins with this set of hypotheses, i.e., fragment pairs positioned in the query frame according to the transformations derived in the clustering procedure. These fragment pairs may belong to any molecule in the database. The objective is to merge these fragment pairs into complete, aligned structures. The assembly algorithm was designed to be able to process as much in parallel as possible. A detailed discussion and analysis, however, will be the subject of another paper.

For the present application, we allow only fragment pairs belonging to the same fragment graph to be merged. The structures that result from the assembly process will therefore be conformers of those molecules that were used to build the database. However, if fragment pairs from different fragment graphs are allowed to merge, new structures may be created through the assembly. We shall take up this issue when we apply the approach to *de novo* design methods and focused library design. Furthermore, we can also allow pieces of larger structures to qualify as matches to the query.

The assembly is an iterative process, where each iteration step begins with a set of fragment pairs and/or partially assembled structures, and produces a set of partially or fully assembled structures that have increased in size by one fragment. Whereas the overall iteration loop runs sequentially, within a step, many fragment pairs or partial structures can be processed in parallel.

At the beginning of each step, fragment pairs and any partial structures from previous steps are grouped together and referred to as *bases*. In other words, a base is a group of adjacent fragments, that have been aligned to the query, and connected with specific torsional angles of the rotatable bonds. All atom coordinates are expressed in the query frame. In addition, a base carries a record of which fragment nodes it contains, and which rotatable bondedges may be used for growth.

The first phase of an assembly step is termed the *base expansion* phase. In this phase, the list of potential merges of fragment pairs with bases is determined. It results from consideration of the *fronts* of the growing bases. The front of a base is defined as the set of rotatable bond edges between a matched and an unmatched fragment node on the fragment graph. Any fragment pair from the set of hypotheses that includes both a frontal rotatable bond edge and matches the fragment conformer present in the base can potentially merge with the base. A merge between a base and fragment pair may occur if the root mean square distance computed between corresponding atoms in the fragments that are common to the base and the fragment pair is less than a threshold. Note that during the assembly process a hypothesis fragment pair can be used for merges many times, onto multiple growing bases.

There are different ways to merge a base and a fragment pair that have close, but not identical atom positions of the common fragment. First, one could leave the base fixed and superpose the fragment pair onto the base's fragment. In the other extreme, one could leave the fragment pair fixed and superpose the base onto the fragment in the fragment pair. We are using an intermediate option, which is constructed by joining the base and the fragment pair, and then determining the least-squares alignment [37] of the unique corresponding atoms in the original base and the grown one, and in the fragment pair and the grown base. The atoms of the common fragment are not included in in this calculation. Note that, due to the merging, the orientation and position of a base depends not only on the fragment pairs that went into it, but also on the order in which they were merged. Therefore, multiple identical bases can be produced with slightly different positions and orientations relative to the query molecule.

After each merge, a *bump check* is performed, which checks whether atoms from the newly added fragment are too close to those that were in the base before. We presently use a threshold of 1.7 Å for non-bonded atom-atom distance [38].

After that, a *shape screen* is applied, which prevents the base from growing too far outside the query's volume. The screen is performed simply by verifying that no atom in the base is more than a threshold distance from the closest atom in the query. This threshold presently defaults to 3.5 Å (see Table 1).

Bases that pass the shape screen are then scored against the query. The default scoring function is a simple Carbo function [39],

$$Q = \frac{\int\limits_{\mathbb{R}^3} \left(\sum\limits_{j=1}^{N_q} h(\vec{x} - \vec{a}_j) \right) \left(\sum\limits_{j=1}^{N_b} h(\vec{x} - \vec{b}_j) \right) d^3\vec{x}}{\left(\int\limits_{\mathbb{R}^3} \left(\sum\limits_{j=1}^{N_q} h(\vec{x} - \vec{a}_j) \right)^2 d^3\vec{x} \right)^{\frac{1}{2}} \left(\int\limits_{\mathbb{R}^3} \left(\sum\limits_{j=1}^{N_b} h(\vec{x} - \vec{b}_j) \right)^2 d^3\vec{x} \right)^{\frac{1}{2}}} \tag{32}$$

where $\mathbf{a}_j$ is the atom coordinate of the jth query atom, $\mathbf{b}_j$ is the atom coordinate of the j^{th} atom in the base, and $h(\mathbf{x}) = \exp(-x^2/2\tau^2)$ is a normalized three dimensional Gaussian density with $\tau = 1$ Å. N_q and N_b are the numbers of atoms in the query molecule and the base, respectively. Scoring functions can vary widely in their sophistication [40]. We have chosen (**??**) for its simplicity.

A base that passes certain criteria qualifies as a *candidate*, meaning that it is considered a successful match to the query. A base is considered a candidate if it has some minimum number of fragments, passes all applicable checks and/or has a sufficient score with respect to the scoring function in use.

Bases that have no opportunities for growth are removed from consideration in subsequent iterations of the assembly process. The bases remaining are used as input to the next iteration of the assembly phase. The assembly process is terminated when there are no bases left for consideration, i.e., when there are no bases left to grow. Note that the procedure is exhaustive in that all candidate alignments are produced if they qualify with respect to the the bump check, shape screen and scoring criteria. The method is designed to attempt to produce the better candidates early. In the application of the method that is described below, all possible qualifying candidate alignments were produced that could be constructed from the aligned fragment pairs submitted to the assembly phase.

Table 1. Experimental parameters

Parameter	Value	Equation
Scoop center placement	nuclear locations	(9), (11)
Scoop radius R	3 Å	(7), (8)
Smearing parameter σ	0.5 Å	(10)
Scoop internal grid spacing ΔR	R/18	
Asymmetry threshold for $\alpha_n / R J_n$	0.02	(22)
Degeneracy threshold ϵ_J	0.04	(21)
Transformation metric parameter α	3 Å	(27)
Cluster distance threshold d_{clust}	3 Å	
Assembly bump check	1.7 Å	
Assembly merge threshold	3.0 Å	
Assembly shape screen	3.5 Å	

Application to dihydrofolate and methotrexate

In this section we will explore applications of the present method to the case of dihydrofolate (DHF) and methotrexate (MTX). This system has been the subject of study for numerous methods of superposition [3, 41, 6] as well as docking [25, 42]. It has become a benchmark for such methods, as it exhibits an interesting aspect that the most reasonable superposition from the perspective of topology is different from the superposition that can be deduced from aligning the enzyme parts of the crystal structures of ligand-enzyme complexes of DHF and MTX bound to dihydrofolate reductase (DHFR) [16]. The latter superposition can be understood in terms of electrostatics and hydrogen-bonding sites [20]. The MTX-DHF system thus serves as a probe to characterize how such methods weight the importance of topological similarity, electrostatic similarity, and potential non-bonded interactions. We chose it as our case study to characterize how our method performs, using the very simple property field detailed in the Property field section.

To this end, we describe two experiments, which will illustrate the importance of feature classification. The first considers the alignment of a rigid conformation of MTX on a rigid conformation of DHF. Both conformations are extracted from the crystal structures of the molecules bound to DHFR. Features are computed at atom centers and grouped according to a classification scheme derived with knowledge of the respective binding modes observed in the crystal structures. This *functional-based* grouping (see Figure 3) is compared to the case where all features are classified equally, and every query feature is compared

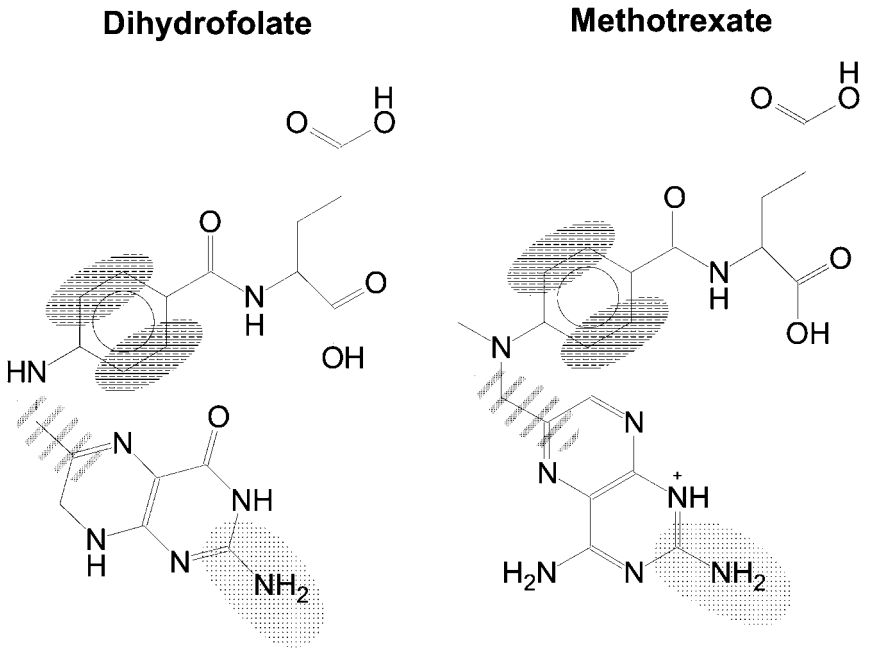

Figure 3. The definition of the three groups for the descriptor scaling, cf. Descriptor scaling and quantization. The first group consists of the descriptors derived from scoops around the guanidine of the pteridine system. There are four such scoops for DHF, and four for MTX, therefore the first group consists of 8 descriptors. Similarly, the second group is made up of 8 descriptors from the chain linking the pteridine and benzamide rings, and the third group of 16 descriptors from around the *p*-amino-benzamide ring.

to every stored feature. Taking DHF as the query, MTX is then aligned under the two conditions of classification, and the results are compared. We should emphasize up front, that knowledge of the particular binding mode is not a requirement for application of the present method. If, however, one has such information, one should be able to use it to better characterize the meaning of similarity in a similarity search.

As a second experiment we use DHF as a query molecule on a database that represents the full conformational space of MTX. The features are classified using two schemes. The first scheme is the functional-based grouping described above. The second scheme assumes no knowledge of the binding modes and simply groups features by the element type (C, N, or O) of the central atom. Comparison of the results of using these two schemes will illustrate the value of injecting relevant context into the feature classification scheme. We compare the workload consequent from these two feature classification schemes. We then conclude with the top scoring alignments assembled from the fragment pairs of MTX. As we shall see, reasonable results are obtained with both grouping schemes. However, when features are classified with the functional-based, i.e., context-derived, scheme, the results are obtained more efficiently and are of higher quality, as assessed by the Carbo scoring function or by examination of the resulting alignments on DHF.

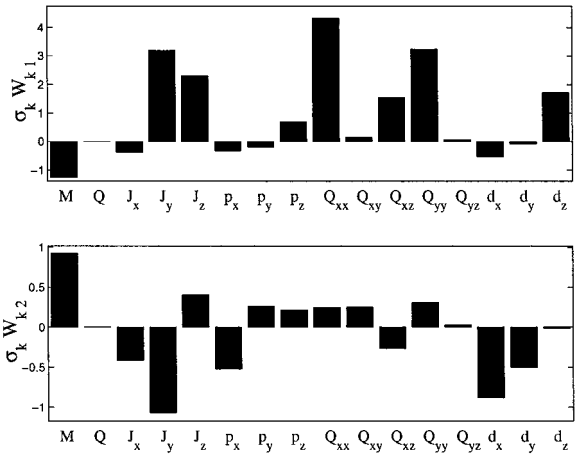

Figure 4. The Fisher discriminant matrix W, see Equation 23. W represents the linear mapping from the 16-dimensional local scoop descriptors, described in the Descriptors and feature generation section , to the 2-dimensional discriminant space in which the similarity measure is defined. In order to make the components W_{kj} comparable, they have been multiplied by the standard deviation σ_k of the kth descriptor ($k = 1, \ldots, 16$), sampled over the total pool of features.

Rigid Body Superposition

Structures of DHF and MTX were extracted from the crystal structures with PDB identifiers **1dhf** and **4dfr** respectively. **1dhf** provides the coordinates of DHF bound to the human form of the enzyme DHFR [43]. **4dfr** provides the coordinates of MTX bound to an E. Coli strain [44] of DHFR. No attempts were made to optimize the structures with quantum or classical methods. Thus, the ring distortion and out-of-plane bending as observed in the crystal structures were left intact. Hydrogen atoms were added using Cerius2 [45].

Features were computed according to the parameters in Table 1. Placing scoops at atomic nuclei produced 53 scoops for DHF and 56 for MTX. Feature groups were defined using the functional-based classification scheme illustrated in Figure 3. These group definitions are somewhat similar to those defined in a QSAR study of DHFR inhibitors by Hopfinger [46]. They emphasize regions of the molecule important for binding and avoid features from parts of the structures that appear unessential for activity.

The loadings of the two Fisher discriminants derived from the functional grouping are shown in Figure 4. The first discriminant has dominant loadings in Q_{xx} and Q_{yy}, as well as the inertial J_y and J_z components of the feature. Large Q components with the present property field arise when the differences in electronegativity with respect to the average lie further from their center of dipole. The second discriminant is most sensitive to M, the integral of the electronegativity property field over the scoop, and differences in J_z from J_x and J_y, which may be seen as a crude measure of planarity of field within the scoop. Differing training sets will lead to a different feature weighting in discriminant space.

The atom-centered features of both DHF and MTX, once projected onto the Fisher discriminants resulting from the functional grouping, are partitioned into bins. The size of the bins in each dimension is set to four times the standard deviation of the largest group. The resulting partitioning scheme is shown in Figure 5 and Table 2. The feature labels refer to the atoms on which the feature scoops were centered. Figure 6 shows the atom labels for DHF (**1**) and MTX (**2**). Inspection of the grouping reveals the features of the designated groups are in fact clustered in the same region of discriminant space. However, rectilinear quantization has its problems. There is no easy way with this method to segregate all of the desired features together in distinct bins.

DHF has 53 atoms, so with scoops located on atomic centers, this results in 53 features. However, 6 features fail to pass the threshold for sufficient non-degeneracy and are removed. For the construction of correspondences between the query and database features, multiple features are generated (see Descriptors and feature generation) for scoops in the query molecule (DHF) that lack sufficient asymmetry to define the sense of their internal axes. This results in the addition of 25 more features for DHF, for a total of 72. Of the 56 scoops for MTX, 10 fail to pass the threshold for sufficient non-degeneracy, leaving 46 features. The total number of correspondences without grouping is therefore the product of the number of features in DHF and MTX, or $72 \times 46 = 3312$, as shown in Table 3. With the functional grouping, the total number of correspondences is only 194.

We see that in spite of its problems, a partitioning of discriminant space is necessary. A brute force comparison of all query features with all candidate features will become prohibitively expensive from a computational perspective when applied to larger problems. Furthermore, examination of the large clusters in the full correspondence case (see Table 3) indicates that *noise* from spurious members in these clusters contaminates the average transformation of the clusters. Ill-defined transformations from such clusters lead to alignments where few structural elements are aligned well.

In both groupings, when one passes the clusters to the assembler (which in this case only applies shape screens and scoring since the molecules are already assembled), *essentially* the same alignments result. This will not be the case when considering the flexibility of MTX. The important point to emphasize here, is that by injecting relevant knowledge into the feature classification scheme, one arrives at the answer with less work. This point becomes critical when larger databases of molecules with more conformational degrees of freedom are considered. This is more than a timing issue: in practice, one often must apply cutoffs and limits to the search. Arriving at an answer efficiently at a small scale may sometimes translate into whether one sees it at all at a larger scale.

A result that is fairly robust with respect to choice of clustering parameters is the production of two alignments, one where the benzamide rings are aligned, and the other with the observed crystallographic alignment of the pteridine rings indicating superposition of the

Table 2. To define similarity, the discriminant space depicted in Figure 5 is partitioned into rectangular bins. Each bin corresponds to a cell in the table. The features are labeled by the name of the central atom of the corresponding scoop (1 = DHF, 2 = MTX), as shown in Figure 6

		1H53				1C18
	2028 2H36	1C2 2N3 1N3 2023 106 2H35 1031 1H33	1029 2H50 1H52 2H55		2032	
	2N1	1022 2C2 1H34 2H34 2H37	2N6	1C12 2H42	2H43	2C15 1H48 2033 2H54 2H52
	1N4 2N11 1C17 2H46 1C19 1C20 1032 1H35 1H44	1N8 2C5 1028 2C10 2H56		1C10	1C9 2H40 1C26 1H37 1H50	1C13 2C9 1C25 2C12 1H39 2C29 1C30 2N14 1H40 2H39 1H49 2N24 1H51 2C26 2C30 2C31 2H48 2H53
	1N1 2N4 1C5 2N8 1C16 2C17 1C21 2C18 1H41 2C20 1H42 2C21 1H45 2027 2H44 2H45 2H47	1N11 2H38 1N14 1N23 1H38 1H43 1H46	1H36	2H41		1C27 2C13 1C24 2C25 2H51 1H47 2H49
1C15	1C7		2C19			2C7 2C16 2C22

regions of the molecules that bind to DHFR. The presence of these two alignments stems from two origins in conflict. One alignment arises from the strength of an exact substructural match, namely the *p*-aminobenzamide. The other, which is the experimentally observed alignment of the pteridine rings, is due to the similarity in locations of chemical functional groups.

Balancing how exact substructural matches should be weighted with respect to similarity in functional group definition is an issue that must be dealt with in any superposition algorithm or similarity scoring function. Exact substructure matches are sometimes relevant, even if trivial. One certainly could not fault an algorithm for scoring such a correspondence high. Rather, one must balance the relative importance *exact* substructural matches have with respect to *similar* substructures. Since there is no universal answer, this should be addressed by the context. In the present case, the two form separate and distinct clusters in transformation space, and are thus presented as alternative alignments.

Flexible Superposition

To query DHF against a database that represents the full conformational space of MTX, the MTX structure was fragmented, and the conformations of each fragment and fragment pair were tabulated. The fragmentation is illustrated in Figure 1, which also shows that the conformational resolution was 60° for the five single bonds and 180° for the aniline bond. Fragment

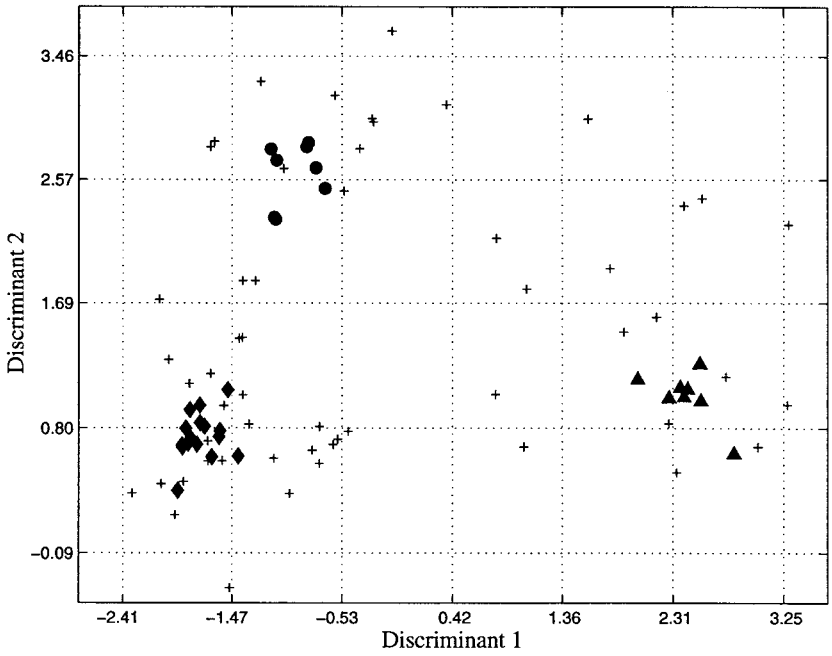

Figure 5. The distribution of features after mapping into a 2-D space by way of discriminant analysis, with the functional grouping (see Figure 3 and text). The features marked with diamonds, triangles, and circles represent the three groups. The remaining features are marked with plus symbols. The query is represented by 53 features, and a single conformation of the database molecule (the crystallographic one) is represented by 56 features.

conformers and fragment pairs that are produced using these angle resolutions were analysed for the presence of non-bonded atom pairs that are closer than 1.7 Å, and these conformations were eliminated from the tabulation [38]. The non-bonded atom threshold reduces the number of conformers of fragment C from 36 to 27, the number of A–B fragment pairs from 72 to 48, and the number of B–C fragment pairs from 324 to 234. The list of inter-fragment torsion angles is appended to the original angle present before fragmentation. The results are shown in Table 5. Thus, one of the 49 A–B fragment pairs and one of the 235 B–C fragment pairs has the original angle found in the crystal structure.

The cluster distributions in Table 4 show the importance of injecting relevant context into the classification scheme. As can be seen from the total number of correspondences constructed in each case, the work load of the correspondence step is markedly higher for the elemental grouping scheme than in the functional grouping scheme. Inspection of the partitioning of the elemental grouping revealed a decidedly less rational distribution. Bins were occupied with more random groupings of features compared to the functional scheme.

Furthermore, there is a difference in the types of alignments of the fragment pairs containing the *p*-amino-benzamide ring and the pteridine ring that are produced by the two schemes. In the functional grouping scheme we see three types of alignments: one with the *p*-amino-benzamide ring of MTX tightly aligned onto that of the DHF; one with the pteridine ring of MTX tightly aligned onto that of DHF in a way compatible with the DHFR binding implied by the crystal structures; and one with the pteridine ring of MTX tightly aligned onto that of DHF in a way compatible with good steric overlap of the rings. In contrast, with the element grouping scheme we see only alignments where the *p*-amino-benzamide ring of MTX is tightly aligned onto that of the DHF. With the set of parameters shown in Table 1, tight alignments of the pteridine rings are not observed, although they can be produced by varying the parameters of the clustering. Not surprisingly, it appears that with the element grouping scheme, the ability to recognize an exact substructural match on the *p*-amino-benzamide ring is greater than the ability to recognize either a functional group or steric match of the pteridine rings. It should be added, however, that even with the element grouping scheme, although alignments based on superposition

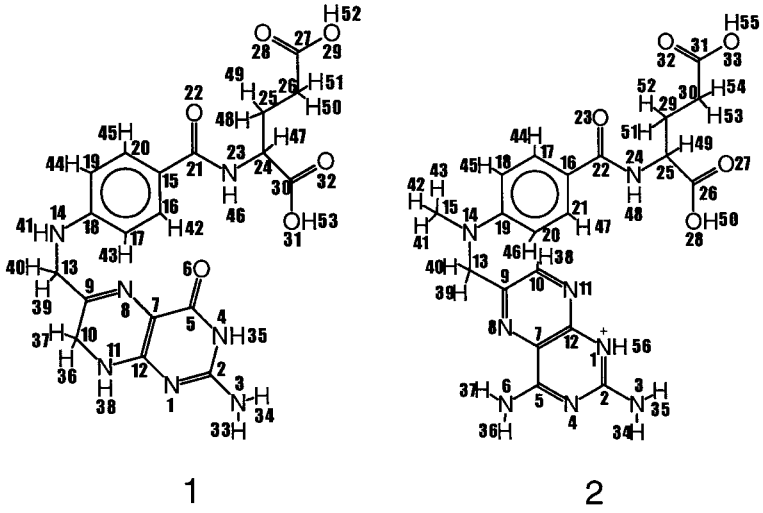

Figure 6. Dihydrofolate **1** and methotrexate **2** with atom numbers shown. Table 2 refers to feature scoops centered on the atoms indicated by the labels.

Table 3. Distribution of the correspondence cluster sizes for rigid body superposition of the crystal structure of MTX onto that of DHF. Left column: features were not grouped, all query features were corresponded to all database features. Right column: functional feature grouping (see text), only similar features were corresponded. The total number of correspondences with grouping is only about one twelfth of the number without grouping, resulting in a significant decrease in computer time. While the cluster size distribution suggests that the grouping eliminates a number of those correspondences that went into the large clusters in the all-to-all case, detailed inspection of the individual high-voting clusters shows that the same alignments are produced in both cases

	Full correspondence	Functional grouping
Votes	Occurrences	Occurrences
16	1	
11	4	
10	1	
9	6	1
8	7	-
7	16	-
6	34	-
5	70	-
4	123	4
3	272	13
2	414	27
1	330	76
Total clusters	1278	121
Total correspondences	3312	194

Table 4. Distribution of the correspondence cluster sizes for flexible superposition of MTX onto the crystal structure of DHF. Middle column: features were grouped according to the element of the atom at the center of the scoop (C, O or N). Right column: features were grouped according to the chemical function of the molecular region in which the scoop is defined. The two grouping schemes imply different similarity measures. The importance of a cluster is given by the number of its members, a larger number indicating more correspondences consistently supporting the same alignment. The total number of correspondences following from functional grouping is only about half the number of correspondences from elemental grouping, which results in a significant decrease in workload by the clustering algorithm

	Element grouping	Functional grouping
Votes	Occurrences	Occurrences
11		14
10		9
9	15	85
8	126	19
7	129	43
6	210	235
5	410	249
4	448	600
3	1358	1466
2	6635	6391
1	31157	11340
Total clusters	40488	20451
Total correspondences	55649	35037

Table 5. Statistics for the fragment and fragment pair partitioning of MTX. The 35 fragment conformers and 284 rotatable bond edge records allow for the representation of more than 15,500 full molecular conformations, see Equation 1

Fragment node	Fragment conformers	Rotatable bond edge	Number of fragment pairs
A	6	A–B	49
B	2	B–C	235
C	27		
Total	35	Total	284

of the pteridine ring of MTX onto that of DHF are not observed, we do see both the binding mode and steric overlap *orientations* of the pteridine ring, but these alignments are produced by correspondences from regions of the molecules away from the pteridine ring and happen to survive shape screening in the assembly phase.

Table 6. The number of bases that survive each assembly phase for the flexible superposition of MTX onto the crystal structure of DHF. The assembly step constructs candidate alignments from the local fragment pair alignments. For both grouping schemes, all clusters with five or more votes are considered for the assembly

Assembly phase	Functional grouping	Elemental grouping
Selected correspondence clusters	654	890
Potential merge pairs	91092	103128
Actual merges	13886	25816
After bump check	11194	21576
After shape screen	436	428

All clusters with votes of five or more are used to produce the starting bases of the assembly step. Since, for this example, completed molecules of MTX are produced when two fragment pairs are merged, the assembly process completes after one iteration. Table 6 shows the work load of the different phases of the assembly process. While the work load is similar for the two grouping schemes, it will turn out that the quality of the produced candidate set is better with the functional grouping scheme.

The assembly process produced 428 candidate alignments with the element grouping scheme and 436 with the functional grouping scheme. All alignments were scored using the function of Equation 32 and clustered into sets based on similarity of score

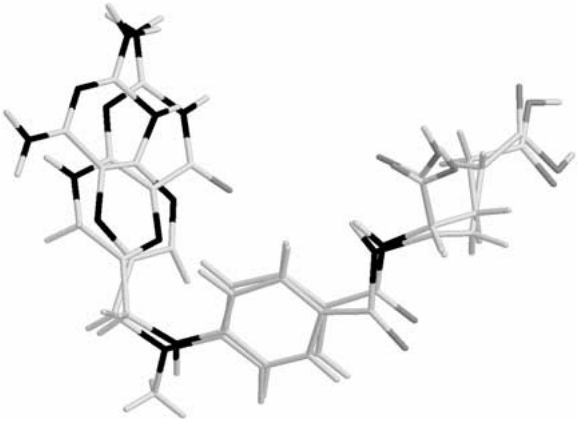

Figure 7. A flexible alignment of MTX onto the crystal structure of DHF generated with FLASHFLOOD. Most of the feature correspondences that led to this alignment are around the aromatic rings, therefore this substructure is almost exactly aligned. The conformation of MTX is different from the one found in the crystal structure. Using the scoring function from Equation **??**, this is the highest scoring alignment, $Q = 0.95$.

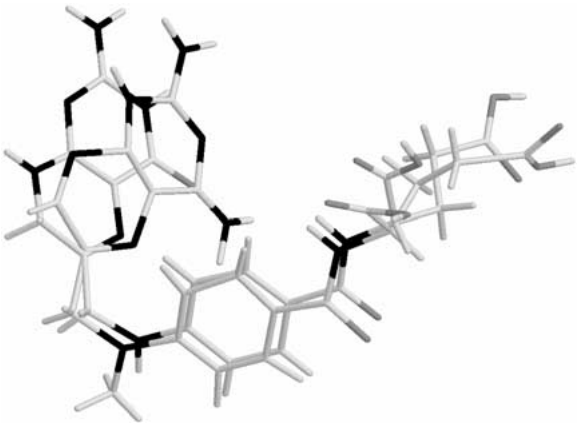

Figure 8. A flexible alignment of MTX onto the crystal structure of DHF generated with FLASHFLOOD. This alignment is mainly based on the aromatic ring substructure match. In contrast to Figure 7, where the pteridine ring alignment agrees with the crystal structure, here the pteridine ring is flipped by 180°. $Q = 0.90$.

and similarity of binding mode. The scores ranged from 0.42 to 0.95. Figures 7–9 show three notable alignments from the three highest scoring sets formed from the 436 alignments that resulted from using the functional grouping scheme.

Figure 7 shows the highest scoring of the 436 alignments, with a score of $Q = 0.95$. The *p*-aminobenzamide rings are strongly aligned and the aliphatic chains are actually aligned more closely than is found from the two crystal structures. It is interesting to speculate about why the chains were not observed in this conformation in the crystal structures. Perhaps it can

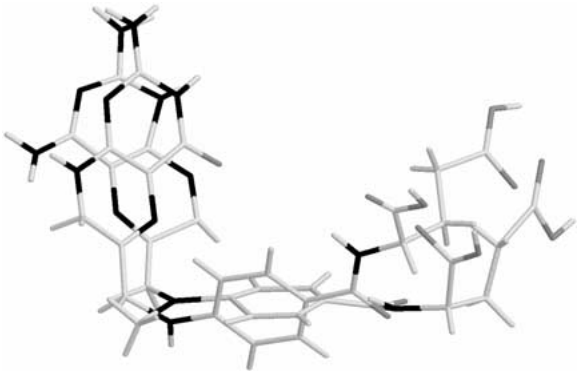

Figure 9. A flexible alignment of MTX onto the crystal structure of DHF generated with FLASHFLOOD. This alignment is based on a similarity match of the pteridine ring, and it is the one closest to that implied by comparison of the two crystal structures. Note that this is not a substructure match. The planes of the aromatic rings are skewed with respect to each other. $Q = 0.74$.

be attributed to the fact that **4dfr** is an E. Coli form and **1dhf** is the human form of DHFR, and the two sequences differ in a nontrivial way, resulting in a different site geometry. Another reason might be that these chains are solvent exposed, and are thereby influenced by packing interactions from crystallization of the complex. Of course, these kinds of considerations do not factor into a superposition analysis because the enzyme is not considered.

While the alignments represented in Figures 7 and 8 both strongly align the *p*-amino-benzamide rings, the pteridine rings are in opposite orientations. In Figure 7, the pteridine alignment is indicative of the observed binding, whereas in Figure 8, with a score of $Q = 0.90$, the pteridine rings are in a sterically similar orientation. Due to the strong effect the benzamide has on the average transformation that aligned the respective fragment pairs, the correct hydrogen bonding groups on the pteridine ring are not as incident to the corresponding groups on DHF as they are in the alignment shown in Figure 9, with a score of $Q = 0.74$. Of the three alignments shown, this one has the best overlap of the pteridine ring, and is reasonably close to the observed binding. This favorable pteridine ring alignment comes at the cost of the overlap of the benzamide ring.

The alignments produced by the element grouping scheme were similar to those shown in Figures 7 and 8.

Selectivity

We now turn to the question of how well the method works using these operating parameters when other molecules are added to the database. Two groups of molecules will be added to the database: a group of DHFR inactives and a second group of DHFR inhibitors. By comparing the selectivity of MTX to the eight DHFR inactives we will first examine selectivity in a broad context. Examination of selectivity in a finer context will be conducted by comparing MTX to the eight DHFR inhibitors.

Table 7 gives information about the sixteen additional molecules that were added to the database containing MTX. The set for this experiment is comprised of eight inhibitors of DHFR [47] and eight inhibitors of other enzymes indicated below, referred to collectively as the DHFR inactives. The eight DHFR inhibitors are labeled according to the compound number in Crippen's original work [47]. The initial coordinates for the eight DHFR inhibitors shown in Figure 10 were obtained by energy minimization [48]. The eight DHFR inactives shown in Figure 10, are ligands whose coordinates were extracted from PDB files of the ligand bound to a protein. Four are Thermolysin inhibitors (from the PDB files **1tlp**, **1tmn**, **3tmn**, **5tmn**), two are Carboxypeptidase inhibitors (**1cbx**, **7cpa**), and the remaining two are Glycogen Phosphorylase inhibitors (**3gpb**, **4gpb**).

Though small, this group exhibits important features. The size on the inactives range from the small and rigid sugars that are Glycogen Phosphorylase inhibitors, to the larger flexible Thermolysin inhibitors. The size of the eight DHFR inhibitors are consistently smaller than MTX, and their activity spans a wide range [47], as shown in the last column of Table 11.

Using a torsion angle resolution consistent with the methotrexate characterization described above, the number of conformations represented ranges from a few dozen to over 10 million for the DHFR inactives. The number of conformations represented for the DHFR inhibitors total under a thousand for the set, due to their limited flexibility. Thus, with a fairly small set of molecules, we have created a virtual conformational database where the DHFR inhibitors are highly under-represented.

We used the same conditions as were described in the Application to Dihydrofolate and Methotrexate section. All single bonds were sampled at 60° intervals. Terminal groups such as carboxy, isopropyl, and phenyl were left rigid. Amides were left in a *trans*

Table 7. Fragmentation statistics on the additional molecules added to the database

PDB source	Rotatable bonds	Approximate conformers represented	Fragment conformers	Fragment pairs	Total features
DHFR Inactives					
1cbx	3	216	12	103	2781
1tlp	8	1679616	54	3272	138866
1tmn	9	10077696	68	2922	107601
3tmn	4	1296	28	350	15050
5tmn	8	1679616	68	4219	174957
7cpa	9	10077696	144	1735	82231
3gpb	2	36	6	11	319
4gpb	2	36	7	19	532
4dfr	5	15552	35	284	10185
Total for Inactives	–	23531760	422	12915	532522
DHFR Inhibitors					
1	2	36	7	21	798
2	2	36	7	21	756
4	2	36	4	9	360
6	3	216	8	53	2173
61	3	216	8	55	2475
64	3	216	8	51	2448
67	2	36	7	21	735
68	2	36	8	21	714
Total for Inhibitors	–	828	57	252	10459
Combined Total	–	23532588	479	13167	542981

configuration. All property field settings were used as specified in Table 1.

Ligands from **1tlp**, **1tmn**, **7cpa**, and **5tmn** were partitioned into four fragment nodes; and the ligands from **3tmn**, **1cbx**, **3gpb**, **4gpb**, and the eight DHFR inhibitors were partitioned into two fragment nodes[49]. The fragmentation statistics for the sixteen molecules are shown along with those for methotrexate in Table 7. In this table, except for methotrexate (**4dfr**) the approximate number of conformers represented is simply the number of torsional angles represented for each bond (in this case 6, since the torsional angle resolution is 60°) raised to the number of rotatable bonds. Ethyl groups in esters **61** and **64** were held fixed in an extended conformation. Note that this represents only approximately the number of conformers because some, that are sterically forbidden, are removed by the bump check, and others are added, corresponding to the inter-fragment torsion angles present in the original crystal or energy minimized structures. For discussion on methotrexate statistics, see the Fragment pairs section, Table 5 and Equation 1.

The number of fragment pairs is the number that have passed a bump check of 1.7 Å [38]. Features were computed at each atom center for each fragment pair. One can see from the data in the table that there is considerable savings in breaking the larger structures up and characterizing fragment pairs rather than molecular conformations. Due to the single partitioning of the DHFR inhibitors, only the bump checks reduced the number of fragment pairs used for feature extraction.

Feature computation can be expensive, but it is done only once, when molecular information is added

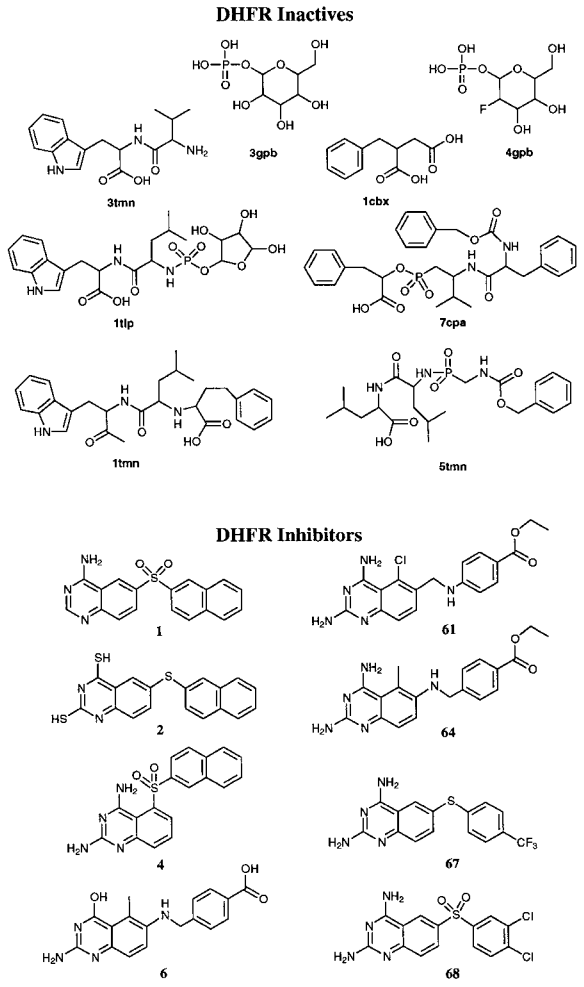

Figure 10. The sixteen additional structures added to the database containing MTX. Each of the DHFR inactives structures were extracted from the PDB source file indicated.

to the database. Its cost will vary widely depending on complexity of the field computation and grid resolution, if the field is computed on a grid. Once the features are computed, however, timings for the alignment and assembly stages of a query are independent of the field complexity. Feature computation with the present property field and operational settings averaged about 4 features per second on a workstation [50].

Selectivity During Alignment Stage

It is critical for the efficiency and scalability of the method that the burden on the assembly stage is kept to a minimum. The results shown in Tables 8–11 indicate a considerable degree of selectivity before the assembly stage for MTX and the eight DHFR inhibitors relative to the DHFR inactives. The feature classification scheme selects which correspondences will be made. There are a total of 542 981 features in the database, and 72 features on the query molecule. This gives 39 094 632 possible correspondences. Using the functional grouping scheme, and screening out degenerate features, only 1 421 086 or about 3.7% of the total possible correspondences were considered for MTX and the eight inactives. If we include the other eight inhibitors, this number only rises to 1 458 843. The correspondences of MTX and the eight inhibitors are only 0.19% of the total. Each correspondence has an associated transformation which is clustered as described above. Clustering produced 1 053 312 transformation clusters in total. These transformation clusters are ranked according to the number of correspondences they contain. From this pool of transformation clusters, each of which describes a fragment pair aligned on the query molecule, one can select for assembly either the top ranking clusters from *each* molecule, or the top ranking clusters from the entire pool of clusters. For either choice, one sees a high degree of selectivity for MTX, which has been achieved by the choice of property field and functional grouping scaling scheme.

Table 8 and Figure 11 illustrate the selectivity of DHF for MTX with respect to the eight inactives in the database. Table 8 shows each molecule's distribution of correspondence cluster sizes for alignments of its fragment pairs onto DHF. Figure 11 gives the corresponding proportions from each of the eight inactives and MTX for each cluster size. Excluding the set of eight DHFR inhibitors, there were 1 421 086 feature correspondences. Only 35 037, or 2.5%, of them were from MTX. Similarly, this gives 1 036 401 transformation clusters, of which only 20 451, or 2.0%, were from MTX. However, if one considers only the larger clusters, which represent more significant alignments, one sees that MTX quickly dominates the distribution. For example, if one selects clusters with six or more transformations from the entire pool of transformations, MTX represents 405 of the 533, or 76%; seven or more gives 170 out of 190 or 89%; and eight or more exclusively selects MTX. Even though MTX represents a small fraction of the total correspondences considered, clustering clearly reveals MTX fragment pairs as the best ones to use in the assembly stage. Subsequent assembly of these MTX fragment pairs leads to the results seen in the previous sections. The additional 23 million conformations represented

Table 8. Each molecule's distribution of correspondence cluster sizes for fragment pair alignments. Selectivity for MTX with respect to the eight DHFR inactives is illustrated. This is a direct consequence of the choice of property field and contextual scaling provided by the functional grouping scheme. Selecting the highest ranking transformation clusters, such as all those with at least six correspondences, across all molecules almost exclusively selects MTX (**4dfr**) for assembly

	DHFR Inactives									
Cluster size	**1cbx**	**1tlp**	**1tmn**	**3tmn**	**5tmn**	**7cpa**	**3gpb**	**4gpb**	**4dfr**	Total
11	–	–	–	–	–	–	–	–	14	14
10	–	–	–	–	–	–	–	–	9	9
9	–	–	–	–	–	–	–	–	85	85
8	–	–	–	–	–	–	–	–	19	19
7	–	–	4	–	2	14	–	–	43	63
6	1	15	21	8	0	63	–	–	235	343
5	4	1224	743	194	10	1887	–	–	249	4311
4	67	4574	1590	556	983	3380	–	–	600	11750
3	245	15434	10231	1592	16655	7947	–	–	1466	53570
2	1620	52362	42092	4667	72091	42599	41	61	6391	221924
1	5061	190624	132235	22714	248973	132622	225	519	11340	744313
Total clusters	6998	264233	186916	29731	338714	188512	266	580	20451	1036401
Total correspondences	9330	366156	257341	40066	447116	265092	307	641	35037	1421086

by the 13 000 fragment pairs of the molecules in Figure 10 do not affect the result that MTX is primarily selected from the database.

Table 9 shows the corresponding results for the DHFR inhibitors. Despite the low number of conformations relative to the inactives, the DHFR inhibitors scored consistently high (see Tables 10 and 11) and were within the selection threshold of 6 votes or higher, as seen in Table 9. If one selects clusters with 6 or more transformations from the entire pool of transformations, the inhibitors (including MTX) represent 487 of the 606. Even with clusters representing 5 or more transformations, the inhibitors (including MTX) represent 1031 of the 5221 – and this is out of a total of 1 061 703 transformation clusters. Selection of the larger clusters for assembly therefore appears to be a valid means of significantly reducing the aligned fragment pairs prior to assembly. Given the elementary level of consideration for medicinal chemistry, the dramatic reduction in the number of clusters to consider is encouraging. It is clear that such specific pre-screening is essential for larger datasets, particularly when the final scoring function used is expensive.

Selectivity During Assembly Stage

If, however, we look at the top ranking transformation clusters from *each* molecule individually, will there still be selectivity for MTX? Table 10 shows the timing and load data for the assembly of the fragment pairs MTX and the eight inactives. Table 11 gives the data for the eight inhibitors. Based on the size of the transformation clusters, the top ranking fragment pair alignments were selected for each molecule separately. For each molecule, enough cluster sizes were considered to yield at least 500 fragment pair alignments. This resulted in the number of initial alignments shown in Tables 10 and 11. For molecules small enough to be defined by a single fragment pair, the assemble stage will yield as candidates the aligned fragment pairs, provided they pass the shape screen. This is the case for **3tmn**, **1cbx**, **3gpb** and **4gpb** and the DHFR inhibitors (excluding MTX). For larger molecules, both partial and full assemblies will yield candidates, again aligned to the corresponding region of the query molecule, provided they pass the bump check, merge threshold and the shape screen. However, most of the larger molecules did not have sufficient similarity to DHF for a complete or partial assembly of the molecule. In fact, only MTX showed any merges of fragment pairs. Note that assembling conformers requires the combination of high ranking fragment pairs *and* their proximal placement. So, on examination of Tables 10 and 11, we see that the bulk of the time for the Thermolysin and Carboxypeptidase

Table 9. Distribution of correspondence cluster sizes for the DHFR inhibitors. Selectivity for the DHFR inhibitors with respect to DHFR inactives is illustrated by comparison with Table 8. Note that although correspondences involving the inhibitors form larger clusters that inactives, they are still less favored than MTX (**4dfr**). Inspection reveals that clusters greater than 6 are mainly due to the alignment of the flexible chain. This part of DHF and MTX is mostly solvent exposed, as seen in the crystal structures. Corresponding structure that would have occupied the same space as the flexible chain is not present in the eight inhibitors

DHFR inhibitors									
Cluster size	**1**	**2**	**4**	**6**	**61**	**64**	**67**	**68**	Total
7	–	–	–	1	–	1	–	–	2
6	3	3	5	10	9	13	11	17	71
5	35	9	7	79	53	60	37	24	304
4	89	57	25	119	148	115	45	26	624
3	237	183	89	255	325	252	160	109	1610
2	638	559	254	1226	1152	1021	453	477	5780
1	1287	1319	662	4022	3762	3422	1248	1189	16911
Total clusters	2289	2130	1042	5712	5449	4884	1954	1842	25302
Total correspondences	3823	3277	1602	8177	7952	7065	3065	2796	37757

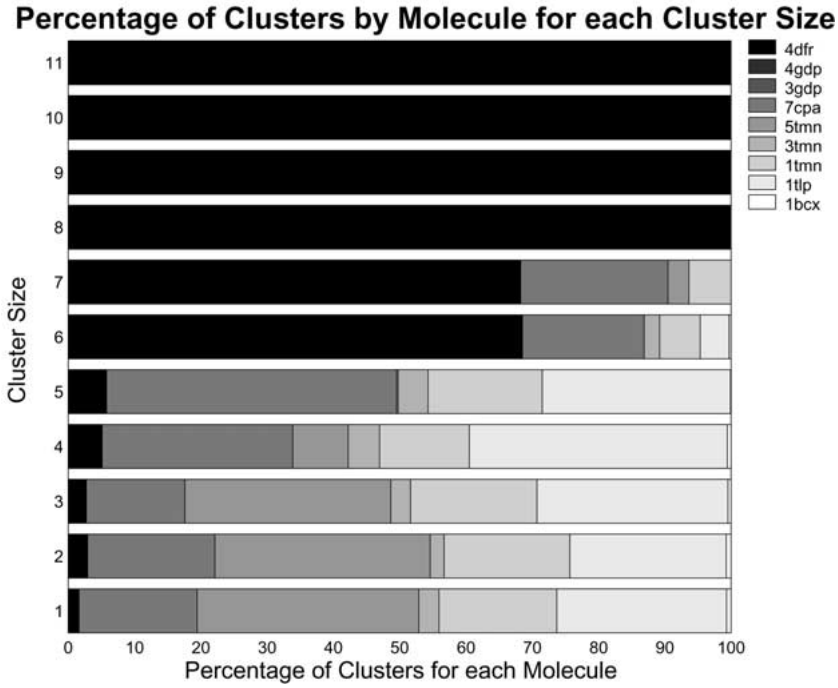

Figure 11. Shown here is a 100% stacked bar chart, derived from Table 8, for the proportion of transformation clusters for each cluster size for each of the eight DHFR inactives and MTX (**4dfr**). One sees that for small cluster sizes, transformations of fragment pairs from MTX represent a small fraction of the total number of clusters of that size. For the larger clusters, however, transformations involving MTX dominate. Selecting the largest clusters for assembly, therefore, almost exclusively focuses resources on MTX.

inhibitors is spent indexing features and clustering transformations during the fragment pair alignment stage, where most of the potential assembly load is screened out. For MTX, however, most of the time is spent on assembly due to the large number of fragment pair alignments that could be merged, screened for shape and scored. All assembly runs were exhaustive for the selected fragment pairs.

From Tables 10 and 11 we see that four of the DHFR inactives and all of the DHFR inhibitors had alignments of conformers that met the criteria for candidate selection, and, the assembly results again strongly favor MTX over any of the other molecules investigated.

The last column of Table 10 shows the highest Carbo scores of the candidates. We can see, however, that of the few candidate conformer alignments submitted for final evaluation, the Carbo score ranks MTX well above the rest. The difference in size and shape of the DHFR inactives, **3tmn**, **1cbx**, **3gpb** and **4gpb** from DHF results in a much lower Carbo score.

Inspection of the alignments of the DHFR inhibitors reveal that in each case, alignments similar to those exhibited by MTX in Figures 7, 8 and 9 were observed. Compounds **2** and **6** have an hydrogen bond acceptor that allows a binding mode analogous to DHF, which is also observed.

The scores for the eight DHFR inhibitors are given in Table 11. These scores are on average higher than for the DHFR inactives. In fact, only one of the alignments of **3tmn** has a Carbo score (0.66) that is comparable to those of the DHFR inhibitors. Inspection of this alignment of **3tmn** to DHF however reveals that **3tmn** is aligned to the branched chain of DHF – a similarity not important for binding to DHFR. In fact, the highest scoring alignments of two of the other DHFR inactives (**3gpb** and **4gpb**) also align along the branched chain of DHF. Such scores could well be screened out by filters favoring the known binding mode of DHFR inhibitors, since the branched chain region is known from crystal structures to be solvent exposed. In terms of analyzing the Carbo function scores, we should remember that these scores are based entirely on shape similarity and do not take into account any other considerations relevant to medicinal chemistry (e.g., consideration of functional group compatibility, as in SQ [3]). Given this, we would not expect the Carbo scores to more than qualitatively represent the difference between these molecules in their chemical similarity to DHF, and would certainly not expect the scores to mirror the experimental binding affinities reported for the DHFR inhibitors. Furthermore, we note that the Carbo function penalizes somewhat for mismatched shape. This is particularly relevant for the DHFR inhibitors (excluding MTX) which have no structure to fill the volume corresponding to the chain of DHF. This is reflected in the scores of the best alignments, which are around 0.6.

In summary, even though we selected the top ranking fragment pair alignments for *each* molecule, fragment pairs did not match across the entire volume of the query, so there were no multiple fragment pairs to be considered for merging to form more complete conformers (and, therefore, higher scoring candidates) for molecules other than MTX.

We did not use knowledge of the structure of DHFR at any phase of the screening process, since our aim was to develop a similarity searching application. Furthermore, it was not our intention here to further refine or optimize the binding mode. Our results show that from a modest flexible dataset, we are able to reduce the number of conformation and alignments to a few hundred to examine, from several million.

Summary

We have presented a method for the field-based alignment of flexible molecules. The method offers a systematic way to use arbitrary property fields for the purposes of similarity search and alignment. Context specific information can be used to scale the relative importance of high dimensional descriptors derived from the property field under study. The method is designed to operate on overlapping parts, or *fragment pairs*, of molecules, and utilizes an efficient method of conformational space representation.

We applied the method to the benchmark dihydrofolate-methotrexate system, and have demonstrated that by injecting context specific knowledge into the feature classification scheme, one arrives at the reasonable alignments more efficiently. For the present study we used a very simple property field. Even with this simple property field the appropriate inclusion of context in the definition of similarity allowed the production of alignments consistent with the binding modes present in the crystal structures in both rigid and flexible treatments of conformational space.

Finally, we have looked in detail at how the method works in performing a query for alignments on dihydrofolate from a database of seventeen molecules

Table 10. This table shows the alignment and assembly timings [50], the top Carbo scores, and the total number of candidates for each ligand investigated. The statistics are for the case of flexible alignment of the eight DHFR inactives and MTX onto the DHF query molecule. The initial alignments for each molecule represent those from the largest clusters

PDB source	Alignment time, s	Assembly time, s	Initial alignments	merges	candidates	Top score
1cbx	19	22	1399	-	450	0.48
1tlp	489	10	1773	-	-	-
1tmn	409	6	1095	-	-	-
3tmn	55	8	690	-	81	0.66
5tmn	571	9	996	-	-	-
7cpa	545	15	1964	-	-	-
3gpb	2	7	266	-	129	0.43
4gpb	3	12	580	-	180	0.42
4dfr	48	209	654	13886	436	0.95

Table 11. This table shows the alignment and assembly timings [50], the top Carbo scores, and the total number of candidates for the eight DHFR inhibitors investigated. All structures in this series had only one fragment pair, and so no merging was required. The last column is the experimental binding energy [47]

Compound number	Alignment time, s	Assembly time, s	Initial alignments	candidates	Top score	ΔG binding [47] (Kcal $\cdot$ mol^{-1})
1	60	8	1002	37	0.67	-5.8
2	50	6	811	26	0.65	-6.0
4	50	9	1034	33	0.66	-6.5
6	75	15	1690	44	0.66	-6.5
61	60	5	487	17	0.77	-12.8
64	70	16	1338	17	0.81	-13.1
67	46	5	706	30	0.63	-13.4
68	47	5	653	37	0.61	-13.4

representing approximately 23 million conformations. The method exhibits a high degree of selectivity for alignments of methotrexate as well as other dihydrofolate reductase inhibitors. The selectivity is apparent at initial alignment and assembly stages, and is reflected in the resulting Carbo scores.

Future work will investigate a broader range of molecular systems, more sophisticated property fields such as those derived from quantum mechanical calculations, a more detailed look at the effects of context and the training set used to define context, more optimal parameter settings, and will characterize scalability and performance of the method with respect to both the number of molecules and the size of their respective conformational spaces.

Acknowledgements

The authors would like to acknowledge the significant contributions of Daniel E. Platt to some of the early embodiments of this work, and to B. David Silverman for many insightful discussions. The contributions of Blake G. Fitch and Robert S. Germain to the system archetecture will prove to be essential for the scalability of the method. The authors would like to thank Andrea Califano for creating the context and support for the initial work. The authors would also like to acknowledge the referees for many helpful suggestions made during the review process that helped complete and clarify the presentation of the material.

References

1. Silverman, B.D. and Platt, D.E., J. Med. Chem., 39 (1996) 2129.
2. Lemmen, C., Lengauer, T. and Klebe, G., J. Med. Chem., 41 (1998) 4502.
3. Miller, M.D., Sheridan, R.P. and Kearsley, S.K., J. Med. Chem., 42 (1999) 1505.
4. Lemmen, C., Hiller, C. and Lengauer, T., J. Comput. Aid. Mol. Des., 11 (1997) 357.
5. Klebe, G., Mietzner, T. and Weber, F., J. Comput. Aid. Mol. Des., 8 (1994) 751.
6. Kearsley, S.K. and Smith, G.M., J. Comput. Aid. Mol. Des., 8 (1994) 565.
7. McMartin, C. and Bohacek, R.S., J. Med. Chem., 42 (1999) 1505.
8. Handschuh, S., Wagener, M. and Gasteiger, J., J. Chem. Inf. Comput. Sci., 38 (1998) 220.
9. Mestres, J., Rohrer, D.C. and Maggiora, G.M., J. Mol. Graph. Modeling, 15 (1997) 114.
10. Martin Y.C., Bures, M.G., Danaher, E.A., DeLazzer, J., Lico, J. and Pavlik, P.A., J. Comput. Aid. Mol. Des., 7 (1993) 83.
11. Willett, P., J. Mol. Recog., 8 (1995) 290.
12. Jones, G., Willett, P. and Glen, R.C., J. Comput. Aid. Mol. Des., 9 (1995) 532.
13. Thorner, D.A., Wild, D.J., Willett, P. and Wright, P.M., J. Chem. Inform. Comput. Sci., 36 (1996) 900.
14. Wild, D.J. and Willett, P., J. Chem. Inform. Comput. Sci., 36 (1996) 159.
15. Thorner, D.A., Willett, P., Glen, R.C., Wright, P.M. and Taylor, R., J. Comput. Aid. Mol. Des., 1 (1997) 163.
16. Klebe, G., in Kubinyi, H. (ed.), 3D QSAR in Drug Design, ESCOM, Leiden, 1993, pp. 173–199.
17. Kim, K.H., list of comfa references 1993–1997, In Kubinyi, H., Folkers, G. and Martin, Y.C. (eds), 3D QSAR in Drug Design, Vol. 3, Kluwer, Dordrecht, 1998, pp. 317–338.
18. Klebe, G., comparative molecular similarity indicies analysis, in Kubinyi, H., Folkers, G. and Martin, Y.C. (eds), 3D QSAR in Drug Design, Vol. 3. Kluwer, Dordrecht, 1998, pp. 87–104.
19. Good, A.C. and Mason, J.S., three-dimensional structure database searches, in Lipkowitz, K.B. and Boyd, D.B. (eds), Reviews in Computational Chemistry, Vol. 7, chapter 2. VCH Publishers, New York, NY, 1996, pp. 67–117.
20. Kubinyi, H., similarity and dissimilarity: A medicinal chemist's view, in Kubinyi, H., Folkers, G. and Martin, Y.C. (eds), 3D QSAR in Drug Design, Vol. 3.Kluwer, Dordrecht, 1998, pp. 317–338.
21. Klebe, G. and Mietzner, T., J. Comput. Aid. Mol. Des., 8 (1994) 583.
22. Kearsley, S.K., Underwood, D.J., Sheridan, R.P. and Miller, M.D. J. Comput. Aid. Mol. Des., 8 (1994) 565.
23. Hahn, M., J. Chem. Inf. Comput. Sci., 37 (1996) 80.
24. Patents have been filed and are pending for several aspects of the work described in this paper. Please refer to U.S. Patent filing YOR8-1999-0949 and other references therein.
25. Rarey, M., Kramer, B., Lengauer, T. and Klebe, G., J. Mol. Biol., 261 (1996) 470.
26. Böhm, H.J., J. Comput. Aided. Mol. Des., 6 (1992) 61.
27. Douglas, B.E., McDaniel, D.H. and Alexander, J., Concepts and Models of Inorganic Chemistry. John Wiley & Sons, New York, NY, 1983.
28. Karplus, M. and Porter, R.N., Atoms and Molecules. W. A. Benjamin, Inc., Menlo Park, CA, 1971.
29. Other types of lattices can also be used.
30. Duda, R.O. and Hart, P.E., Pattern Classification and Scene Analysis. John Wiley & Sons, New York, NY, 1973.
31. IBM, Poughkeepsie, NY. Engineering and Scientific Subroutine Library for AIX, Version 3, Guide and Reference, 1997.
32. Stockman, G., Comput. Vision Graphics Image Proc., 40 (1987) 361.
33. NAG Ltd., Oxford, UK, The NAG Fortran Library Manual, Mark 16, 1993.
34. Jain, A.K. and Dubes, R.C., Algorithms for Clustering Data. Prentice Hall, New York, NY, 1988.
35. The choice for d_{clust} used in the present work was made based on a number of experiments with the DHF-MTX system. However, results were relatively insensitive to values in the range of 2.5 to 4.0 Å. This value is probably appropriate for studies using the same clustering algorithm, tranformation distance metric, with fragments that are approximately 10 heavy atoms in size, and for nuclear placement of scoops.
36. Curtis, W.D., Janin, A.L. and Zikan, K., a note on averaging rotations, in IEEE Virtual Reality Annual International Symposium, pp. 377–385. IEEE, 1993.
37. Arun, K.S., Huang, T.S. and Blostein, S.D., Least–square fitting of two 3-d point sets. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-9 (5), 1987, pp. 698–700.
38. The value of 1.7 Å chosen for the bump check parameters used in this work is rather small. Therefore, some high energy conformations are used as fragment pairs and also some high energy candidates are produced during assembly. The value was chosen to screen out only the worst cases of steric overlap in order to assess the method with a larger work load. The performance of the method improves if this parameter is increased.
39. Carbo, R., Leyda, L. and Arnaua, M., Int. J. Quant. Chem., 17 (1980) 1185.
40. Good, A.C. and Richards, W.G., Explicit calculation of 3d molecular similarity, in Kubinyi, H., Folkers, G. and Martin, Y.C. (eds), 3D QSAR in Drug Design, Vol. 2. Kluwer, Dordrecht, 1996, pp. 321–338.
41. Lemmen, C. and Lengauer, T., J. Comput. Aid. Mol. Des., 11 (1997) 357.
42. Lorber, D.M. and Shoichet, B.K. Prot. Sci., 7 (1998): 938.
43. Prendergast, N.J., DeCamp, T.J., Smith, P.L. and Freisheim, J.H., Biochemistry, 27 (1988) 3664.
44. Bolin, J.T., Filman, D.J., Matthews, D.A., Hamlin, R.C. and Kraut, J., J. Biol. Chem., 257 (1982) 13650.
45. Program Cerius2, distributed by Molecular Simulations Inc., 9685 Scranton Rd. San Diego CA 92121-3752.
46. Dunn, W.J. III, Hopfinger, A.J., Cantana, C. and Duraiswami, C., J. Med. Chem., 39 (1996) 4825.
47. Crippen, G.M., J. Med. Chem., 23 (1980) 599.
48. The initial structures were generated using the FlashFlood user interface and were Energy-minimized using the Dreiding force field[51] with the Mulliken suite of programs[52].
49. The atom numbering for the non-hydrogen atoms was kept the same as in the corresponding pdb files. For the nine molecules studied, the bonds that were cut to define fragments are as follows (pdb label: {(atom number)–(atom number),. . . }): **1cbx**: {3–5} , **1tlp**: {1–2, 15–16, 23–24} , **1tmn**: {1–13, 14–15, 22–25}, **3tmn**: {8–9} , **5tmn**: {12–13, 17–18, 24–25}, **7cpa**: {9–10, 18–41, 23–26}, **3gpb**: {1–7}, **4gpb**: {1–7}, and **4dfr**: {11–12, 24–25}. The internal torsional angles that were rotated are as folows: **1cbx**: {2–3, 5–6} , **1tlp**: {2–12, 16–17, 16–19, 27–28, 24–27} **1tmn**: {1–2, 2–3, 13–14, 14–

17, 21–22, 25–26}, **3tmn**: {12–13, 9–12, 2–3}, **5tmn**: {3–4, 11–12, 16–17, 17–20, 25–28}, **7cpa**: {23–24, 36–41, 18–40, 10–11, 10–39, 2–32}, **3gpb**: {5–6}, **4gpb**: {5–6}, and **4dfr**: {9–13, 14–19, 25–29, 29–30}. All angles were sampled at 60° with the exception of 14–19 in **4dfr**, which was sampled at 180°. For the additional DHFR eight inhibitors, the structural parameters of C-S-C groups in molecules **2** and **67** were constrained to the experimental values of CS bond lengths and CSC bond angles in dimethylsulfide [53]. Structural parameters of C-SO2-C groups in molecules **1**, **4** and **68** were constrained to the experimental values of CS and SO bond lengths and CSC, CSO and OSO bond angles in dimethylsulfone [53].

50. The timings were produced running the program on an IBM RS/6000 43P Model 260 workstation, which has a Power3 CPU running at 200 MHz and a 4MB L2 Cache.
51. Mayo, S.L., Olafson, B.D. and Goddard III, W.A., J. Phys. Chem., 94 (1990) 8897.
52. Mulliken 2.0: Rice, J.E., Horn, H., Lengsfield III, B.H., McLean, A.D., Carter, J.T., Replogle, E.S., Barnes, L.A., Maluendes, S.A., Lie, G.C., Gutowski, M., Rudge, W.E., Sauer, P.A., Lindh, R., Andersson, K., Chevalier, T.S., Widmark, P.-O., Bouzida, D., Pacansky, J., Singh, K., Gillan, C.J., Carnevali, P., Swope, W.C., Liu, B., IBM Almaden Research Center, San Jose CA, 1996.
53. Lide, D.R., CRC Handbook of Chemistry and Physics, 75th edition, 1994, pp. 9–31.