J Mol Model (2008) 14:49–57
DOI 10.1007/s00894-007-0251-2

ORIGINAL PAPER

Pattern recognition based on color-coded quantum mechanical surfaces for molecular alignment

Brian D. Hudson · David C. Whitley · Martyn G. Ford · Martin Swain · Jonathan W. Essex

Received: 27 July 2007 / Accepted: 19 October 2007 / Published online: 24 November 2007

Abstract A pattern recognition algorithm for the alignment of drug-like molecules has been implemented. The method is based on the calculation of quantum mechanical derived local properties defined on a molecular surface. This approach has been shown to be very useful in attempting to derive generalized, non-atom based representations of molecular structure. The visualization of these surfaces is described together with details of the methodology developed for their use in molecular overlay and similarity calculations. In addition, this paper also introduces an additional local property, the local curvature (C_L), which can be used together with the quantum mechanical properties to describe the local shape. The method is exemplified using some problems representing common tasks encountered in molecular similarity.

Keywords Molecular modeling · Molecular similarity · Pattern recognition · QSAR · Quantum mechanics

Introduction

Recent publications [1–4] have introduced the concept of local molecular properties projected onto a surface as a generalized method for describing the physical properties of molecules. These surfaces are composed of an isodensity, shrink-wrapped molecular surface using the electron density from semi-empirical quantum mechanical calculations. Upon these surfaces are projected the values of a set of four local properties; these being the Molecular Electrostatic Potential (V_L), the local Ionization Energy (I_L), the local Electron Affinity (E_L) and local polarizability (a_L). The calculation of these properties has been implemented in the program ParaSurf and is fully described in Lin and Clark [4].

This molecular representation has proved to be extremely useful in describing a range of chemical properties. The inclusion of local representations of reactivity and polarizability allows the methodology to tackle a wider range of chemical problems than is available if only the van der Waals and electrostatic properties are considered and is thus significantly different from current computational technologies for performing similarity calculations related to high throughput screening, *in-silico* docking and QSAR.

Crucially, the methods are not atom based. The atoms merely form a scaffold about which the surfaces exist. The use of quantum mechanical methods to calculate both the surfaces and the properties overcomes a common problem in that a, sometimes arbitrary, choice of a suitable atom type is often necessary in atom based approaches. This, together with the underlying physical basis of the methodology, leads to a more generalized paradigm for molecular representations. A further advantage of the quantum mechanical approach is that all of the properties are derived directly from the electron density matrix.

In this paper we present a method for the visualization of these molecular surfaces and an algorithm for performing molecular alignment and similarity calculations based on them. The color coded surfaces are well suited to this approach, since as well as providing a more general view of the properties of a molecule, they are highly suitable for use

B. D. Hudson (✉) · D. C. Whitley · M. G. Ford
Centre for Molecular Design, Institute of Biomedical and Biomolecular Sciences, University of Portsmouth, Portsmouth PO1 2EG, UK
e-mail: brian.hudson@port.ac.uk

M. Swain · J. W. Essex
School of Chemistry, University of Southampton, Highfield, Southampton SO17 1BJ, UK

in the visual, machine-learning based pattern recognition techniques commonly used in areas such as manufacturing quality control.

A number of methods for molecular alignment based on atomic structure analysis have been developed. These methods include geometrical hashing [5], Hopfield neural networks [6, 7], and the quantum similarity superposition algorithm (QSSA) [8, 9]. One method of interest is the topogeometrical superposition approach (TGSA) [10, 11]. TGSA, like a number of other methods, finds a molecular substructure common to the set of molecules being aligned, such as atomic diads and triads. The alignment is performed by matching substructure pairs located at relatively similar positions. The most limiting feature of TGSA is that different molecules must have some degree of similarity; TGSA was designed to align homogenous sets of molecules [10].

Generalized procrustes analysis (GPA) is often used in the social and behavioral sciences to align configurations from different analyses to assess them with respect to each other. Procrustes analysis allows four different transformations to be applied to configurations: translations, reflections, rotations, and scaling. Of these four transformations only translations and rotations are relevant when aligning molecules. However, this method requires the choice of a set of common atoms for the alignment [12, 13]. An advantage of the GPA method is that all molecules in the dataset are aligned together to form a consensus alignment, rather than a series of pair-wise alignments as is common in other methods. In addition, atom based approaches often decompose molecules into entities composed of atoms and bonds, and so such methods are often not easily transferred to molecular surfaces.

A method that has been developed for molecular surfaces uses fuzzy set theory [14] with the surfaces described by properties such as curvature, local lipophilicity, electrostatic potential and the hydrogen bonding density. Linguistic variables are used to generate surface patches or domains, and these are used to search and compare molecular surfaces. Another surface based approach that uses spherical harmonics to represent the surface has been developed for proteins [15, 16] and small molecules [17, 18]. A major advantage of this approach is that the analytical representation of the surface allows rotations to be calculated very quickly.

Barequet and Sharir [19], working within the fields of computer vision and pattern recognition, developed a general method for finding a full or partial alignment between three-dimensional objects. Robinson *et al.* [20] successfully used this method to align small molecules represented by atoms and bonds. In this paper we have built on the work of these authors and applied the method to molecular surfaces.

The method is exemplified using test problems representing some of the common tasks encountered in molecular similarity problems. The features of a typical molecular surface are discussed particularly with regard to the color coding and visualization, and the overlay of pharmacologically similar molecules.

Methods

Surface alignment

The superposition algorithm used is a modified version of that described in Robinson *et al.* [20]. This algorithm is outlined below with the notation following that of the original paper.

We consider that we have two surfaces. The first, called **F**, is fixed in space and used as a template, while the other, called **M**, is able to move and it is this surface that we are attempting to align to **F**. The basic approach is to describe the surface in terms of a series of *footprints*. Each footprint **f** is composed of two parts; the first is the cartesian coordinates, **f.coords**, that can be transformed by a rotation, whereas the second, some molecular property **f.desc**, is independent of the coordinate system. The surfaces calculated by the ParaSurf program consist of a surface mesh that is defined by discrete points. Each of these surface points is suitable for use as a footprint. **f.desc** is used to describe local features, or characteristics, of a surface. Example surface features could be based on the surface's curvature, and would indicate whether the surface has a maximum, a minimum or a saddle region at each surface point. Features like this do not change when the surface is rotated.

When aligning two different molecular surfaces the problem is to identify which surface features, or footprints, should be aligned with each other. A naive approach would need to consider aligning every footprint on surface **F** with every footprint on surface **M**. However, when large numbers of possible alignments are considered, performing the surface alignment by rotating and translating all the footprints is computationally expensive; we therefore need to be more intelligent with the alignments that we consider. We can do this by first comparing the descriptors for every footprint on **F** with every footprint on **M**. If the descriptors are similar then we know that we are comparing similar surface regions on the two surfaces.

More formally, we fill a voting table with a list of voting pairs composed of two footprints, one from each surface, with similar descriptor values. Now consider that each **f.desc** has a single, floating point value. Then we can compare descriptors from each surface by calculating the

absolute difference, **abs**, between two footprints **F.f** and **M.f** from surfaces **F** and **M** respectively:

$$\mathbf{abs} = |\mathbf{F.f.desc} - \mathbf{M.f.desc}| \quad (1)$$

The more similar the descriptors, the lower the value of **abs**. The voting table is populated by calculating **abs** for each possible voting pair. It is then possible to fill the voting table by selecting a number **N** of the most similar voting pairs by choosing a suitable value for **abs**. By changing the value of **N** we can control the size of the alignment problem, which in turn allows us to influence the speed and accuracy of the method.

The transformation needed to align surface **M** to **F** is given by a rotation **R** and a translation **T**. We want to find **T** and **R**, and we do this by applying a series of rotations **r** to surface **M**, each **r** transforming **M** to **M(r)**. The problem now is to deduce which **r** is the optimum **R**, and then given that **R** is known, what is the best **T** needed to optimize the alignment between **F** and **M**.

The optimum transformation is deduced in the following manner. For each rotation **r**, a list of translations is calculated. The translation list is composed of all translations **t** for each voting pair in the voting table. As each voting pair contains a footprint from each surface, these translations are calculated using:

$$t = F.f.coord - M(r).f.coord \quad (2)$$

This corresponds to the vector linking the surface point on the fixed surface, **F.f.coord,** with the other surface point of the voting pair on the moving surface after application of the rotation, **M(r).f.coord**.

Using the translation list it is possible to locate a cluster of similar translations **t**. At this stage we assume that the center of this cluster gives us the best translation for the current rotation. There are several methods that can be used to find a cluster in the translation list. Here a method that is analogous to a gravitational potential has been used. Each **t** is treated like a body in space, with the position of the body given by the vector components of **t**, so that the body *i* with the highest potential $P_i(\mathbf{r})$ will be at the center of the densest cluster:

$$P_i(r) = \sum_{j \in L(r), j \neq i} \frac{1}{|t_i - t_j|} \quad (3)$$

For every **r**, $\mathbf{P_i(r)}$ is calculated, and the maximum value of $\mathbf{P_i(r)}$ gives the optimum rotation **R**. For $\mathbf{P_i(r)}$, the **t** at the center of the densest cluster will give us a suitable value for **T**. Following Robinson *et al.* [20], a cutoff is used in the calculation of $\mathbf{P_i(r)}$ such that, if $|\mathbf{t_i} - \mathbf{t_j}| < 0.5$, $|\mathbf{t_i} - \mathbf{t_j}|$ is set to 0.5 A.

The surfaces and their physical properties are generated by the ParaSurf program, which uses quantum mechanics to calculate the electron densities at, or near, the molecular surface. The electron densities are used to define the shape of the surface and to compute the four descriptors that define the local surface properties.

Although the four descriptors are useful for comparing the physical properties of molecular surfaces, it is possible to define a fifth descriptor that is more suitable for comparing molecular shapes. This fifth descriptor is a simple representation of the surface curvature, and is based on an analysis of the surface points calculated by ParaSurf. Rotationally invariant footprint descriptors play an essential role in the accuracy and efficiency of this alignment method. If we know which parts of the two molecules need to be aligned with each other then we are already very near a solution. It is also important for applications in drug design that the physical properties of molecules can be compared effectively, so that similar molecules can be identified. This is particularly important as two molecules with a similar shape can have very different chemical activities.

The curvature descriptor is defined in the following manner:

1. Calculate the center of mass of the surface, **COM**.
2. For each surface point **p**, find the *N* nearest surface points. Here *N* may be set to a value equal to up to 100% of the total number of surface points.

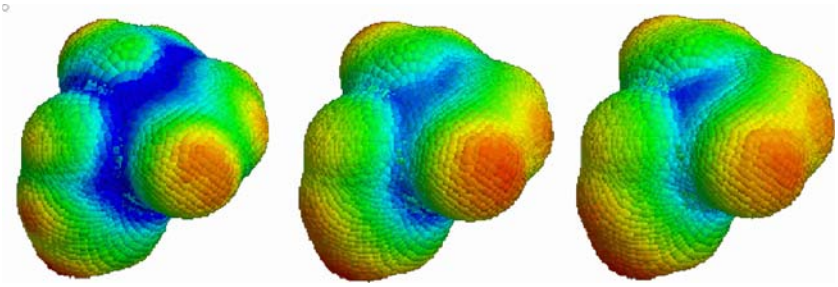

Fig. 1 The curvature descriptors created using 10% (left), 45% (center) and 100% (right) of the nearest surface points for a surface of 4038 points

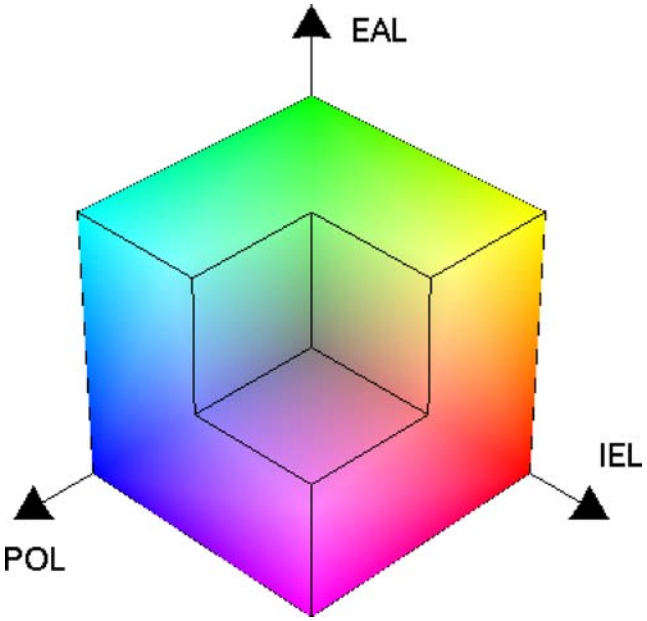

Fig. 2 Additive RGB color scheme

3. Define a triangle given by **COM**, **p**, and a nearby surface point **n**.
4. Calculate the cosine of the angle *A* defined by **COM-p-n**.
5. Repeat the above procedure for all *N* closest points, and calculate the average value of cos*A*. This gives a description of the local curvature at the point **p** i.e., a number $\mathbf{p_{curv}}$ with a value such that $-1 < \mathbf{p_{curv}} < 1$. For a convex region, $p_{curv} > 0$ and for a concave region, $p_{curv} < 0$.

Figure 1 shows an example surface colored according to $\mathbf{p_{curv}}$ (the CRV descriptor). In this figure blue represents areas where the surface is concave ($p_{curv} < 0$) and red represents convex surfaces ($p_{curv} > 0$). The figure also shows the effect of the value of N on the representations. If all points are used in the calculation of the curvature the representations are highly localized. If smaller numbers (e.g. 45%) are used the descriptor is more diffuse. The default value used is 100%. The curvature descriptor is used, in the applications presented here, to define the voting table **V**.

Table 1 Property ranges for drug datasets

	Chembank	Drugs
N	4515	73
IEL mean (median)	475.76 (472.36)	464.74 (458.70)
IEL sd	54.70	53.81
IEL range	366.36 to 585.16	357.12 to 572.36
EAL mean (median)	−91.35 (−96.06)	−90.26 (−95.22)
EAL sd	20.18	18.93
EAL range	−50.99 to −131.71	−52.4 to −128.12
POL mean (median)	0.29 (0.29)	0.3 (0.3)
POL sd	0.04	0.03
POL range	0.21 to 0.37	0.24 to 0.36

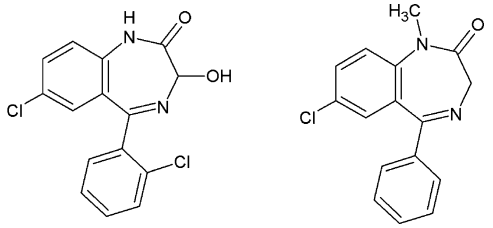

Fig. 3 Molecular structures of Lorazepam and Diazepam

The problem of scoring the matches between individual molecules remains. The gravitational potential, given by Eq. (3), is effective in picking out the optimum translation and rotation, but because different pairs of surfaces have different numbers of voting pairs, this equation is not so useful for identifying, from a dataset of surfaces, the most similar pair. We have used Eq. (4), based on the Coulomb relationship, to compare alignments using each of the five descriptors. Here the sum is over all surface points N_i and N_j of the two surfaces, $\mathbf{\Phi_i}$ is the descriptor value for each surface point, and $\mathbf{r_{ij}}$ is the distance between the two surface points *i,j* using a cutoff of 0.001 Å to avoid overflows in Eq. (4).

$$\mathbf{E_{prop}} = \frac{\mathbf{1}}{\mathbf{N_i N_j}} \sum_{\mathbf{i=1}}^{\mathbf{N_i}} \sum_{\mathbf{j=1}}^{\mathbf{N_j}} \frac{\mathbf{\Phi_i \Phi_j}}{\mathbf{r_{ij}^2}} \tag{4}$$

The individual scores for each of the descriptors are combined to give a total score using Eq. (5). In practice the weights in Eq. (5) have been set to unity.

$$\mathbf{E_{total}} = \sqrt{\mathbf{w_{mep}E_{mep}^2} + \mathbf{w_{iel}E_{iel}^2} + \mathbf{w_{eal}E_{eal}^2} + \mathbf{w_{pol}E_{pol}^2} + \mathbf{w_{crv}E_{crv}^2}} \tag{5}$$

The algorithm has been implemented in the software program ParaMatch. The general features of the algorithm are described below.

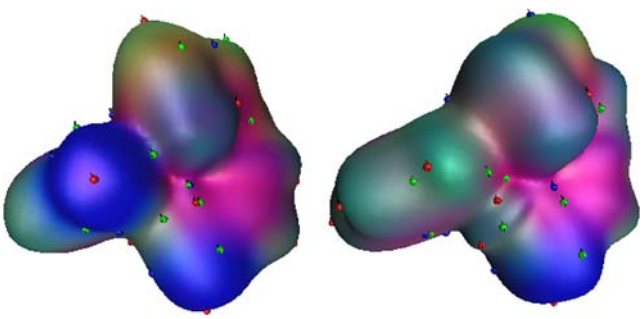

Fig. 4 Molecular surfaces for Lorazepam (left) and Diazepam (right)

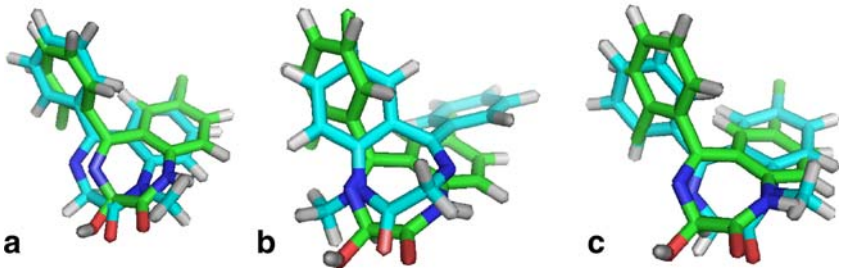

Fig. 5 Overlays of Lorazepam and Diazepam. (**a**) Hex canonical; (**b**) ParaMatch using CRV; (**c**) ParaMatch using MEP

Figure 1 uses 4038 surface points, and this results in over 16 million possible voting pairs. If such a large number of voting pairs are used in the alignment procedure it will take a very long time to find a solution. However, if too few voting pairs are used poor alignments may result. This may happen because the points involved in the voting pairs sample the overall surface shapes badly. Poorly sampled surfaces can be avoided by specifying that the surface points used in the voting table are separated by a minimum distance. The default value for this distance is 2.0 Å. This typically reduces the number of surface points by a factor of over 100.

Surface visualization

To visualize the properties simultaneously, an encoding of the properties onto the RGB color scale was performed. To do this, the local ionization energy (I_L), local electron affinity (E_L) and local polarizability (a_L) are range scaled to values between 0 and 1 and these values are assigned to the red, green, and blue channels respectively using the additive RGB color scheme. Figure 2 shows a schematic representation of the additive RGB scheme.

Unfortunately, this simple scheme has some problems. Firstly, the colors are scaled relative to the internal range of an individual molecule. This highlights differences within the molecule and is useful when considering, for instance, chemical substitution patterns. However, it makes comparisons between molecules that do not have similar ranges of the properties impossible. Secondly, a single functional group can significantly skew the ranges and hence change the colors over the whole molecule. Most noticeable is replacement of hydrogen atom in a molecule with a highly polarizable group such as chlorine. This correctly shows a blue region round the chlorine but has the effect of reducing the blue in other parts of the molecule and thus distorting the overall colors of the molecules.

As a result of this we have developed an absolute color coding scheme. This uses representative property ranges for a number of typical drug molecules. Properties were calculated for two datasets, a small database of 73 commonly prescribed drugs [21] and a larger dataset of 4,515 drugs taken from the Chembank bioactives database [22]. Table 1 shows the descriptive statistics for these datasets for the three local properties. It can be seen in this table that the statistics are similar for both datasets, indicating that the drugs dataset can be used as a representative drug set. It is also clear that there is little difference between the mean and the median of the properties indicating that the use of the mean value is acceptable. In order to avoid the overdue influence of extreme points, the range was calculated as twice the standard deviation about the mean. The values of the chembank dataset were used as the absolute RGB color coding scheme.

The electrostatic potential (V_L) is treated differently. Since there are only three color channels available and the features of the electrostatic potential of molecules are fairly well understood, we chose to encode only the local maxima and minima of this property. These are superimposed onto the surface as solid balls with blue balls representing minima and red maxima of V_L. This representation is a means to depict all four properties in a single figure.

In the following examples, structures of all molecules were generated using CORINA [23]. Quantum mechanical procedures, including geometry optimization of the CORINA structures, were calculated using the AM1 hamiltonian using VAMP 9.0 [24]. The molecular surfaces, comprising the isodensity surface and the projected values of the local properties, were calculated using the default parameters in

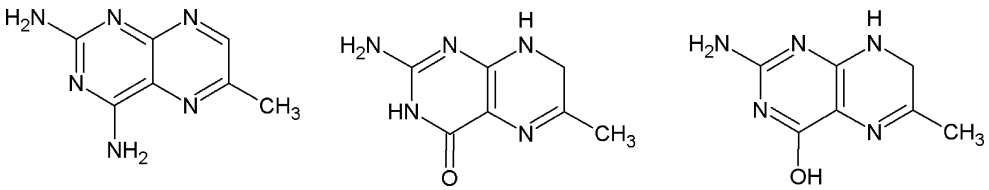

Fig. 6 Molecular structures of heterocyclic rings in Dihydrofolate (DHF) and Methotrexate (MTX)

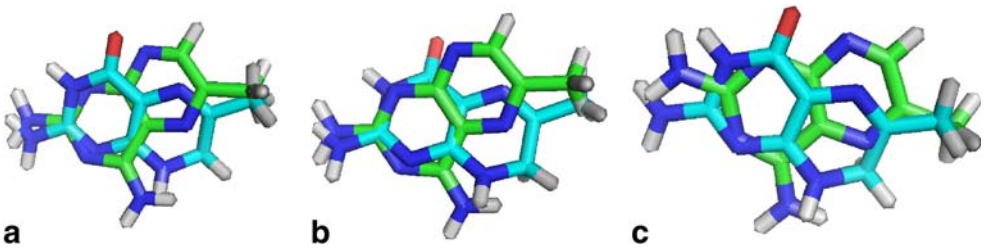

Fig. 7 Overlays of DHF and MTX in the keto form. (**a**) Hex canonical; (**b**) ParaMatch using CRV; (**c**) ParaMatch using MEP

ParaSurf [25]. Graphical representations were generated using freely available software including geomview [26] and pyMol [27]. Alignments using the above algorithm were performed using ParaMatch with both the curvature descriptor and the molecular electrostatic potential. In all cases the minimum number of voting pairs was set to 200 and the angular increment used in the rotational search was set to 5 degrees.

As a comparison, the molecular alignments produced using ParaMatch, are compared with the equivalent results produced using the program Hex [15]. Hex is a molecular alignment tool based on spherical harmonic representations of molecular surfaces. In this work, the canonical alignment function of Hex, where the molecules are aligned by maximizing the fit between the first six spherical harmonic coefficients of the expansion of the molecular surface [21], was used. This is equivalent to aligning the longest axis of the molecules along the X axis, the largest orthogonal axis along the Y axis and the largest axis orthogonal to both along the Z axis.

Results

Example 1: Benzodiazepines

The first example is a relatively simple overlay between two chemically similar molecules, Lorazepam and Diazepam (Fig. 3).

The molecular surfaces produced from the ParaSurf local properties are shown in Fig. 4.

The general features of these chemically and pharmacologically similar molecules are obvious in the representations. Pure blue represents a region of high local polarizability and it can be clearly seen that Lorazepam and Diazepam share a common polarizability only region. Lorazepam has an additional such region. The use of the quantum mechanically derived potentials and the RGB coding scheme in these representations leads to very different visual models of the chemical nature of the molecules than the more common atom centered coulomb model would produce.

Since the representations give a reasonable description of the ParaSurf surfaces, the major application of ParaMatch is to align molecules based on their local surface properties and to assess the degree of similarity between the molecules in that alignment. The calculations were performed using both the CRV and MEP descriptors to align the surfaces.

Figure 5a-c shows the overlays. In these figures the fixed molecule is represented with the carbon atoms in green and the moving molecule has the carbons in cyan. Figure 5a is the canonical alignment from Hex. This alignment is similar to that which would be produced using an atom-based alignment of the common rings, modified by the presence of the methyl group in Diazepam and the extra chloro group in Lorazepam. Figure 5b shows the ParaMatch alignment based on the CRV descriptor. It can be seen that the positions of the phenyl rings have been reversed compared to the canonical alignment. The score for this alignment, using the weighted sum of all five properties (Eq. (5)) is 33.6. The alignment based on the MEP (Fig. 5c) is more similar to the canonical alignment but the atoms are displaced somewhat reflecting a better fit between the values of the MEP. The score for this alignment is 33.8, which shows a slightly better fit than that obtained with the CRV descriptor.

Example 2: DHFR inhibitors

In this example, we have repeated the examples described in Kearsley *et al*. [28]. In this, the heterocyclic rings of two molecules, which bind to the enzyme dihydrofolatereductase (DHFR), were aligned using the steric and

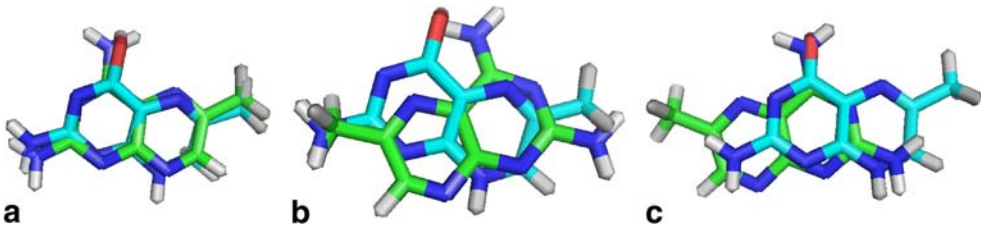

Fig. 8 Overlays of DHF and MTX in the enol form. (**a**) Hex canonical; (**b**) ParaMatch using CRV; (**c**) ParaMatch using MEP

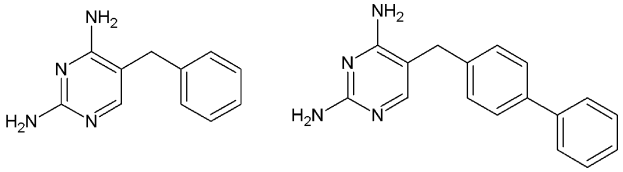

Fig. 9 Molecular structures of low-diversity DHFR structures

electrostatic alignment (SEAL) algorithm. The first of these was the natural substrate dihydrofolate (DHF) and the second was a highly effective inhibitor, methotrexate (MTX). A problem with this analysis is that only the keto form of the DHF ring was studied, whereas the enol form is equally likely. Indeed it is the enol form of DHF which is often portrayed in chemical depictions. This tautomer problem is significant for all 3D methods of alignment, and has yet to be resolved. The chemical structures of the rings are shown in Fig. 6.

The two molecules were therefore investigated using ParaMatch. Figures 7a-c show the Hex canonical and the ParaMatch CRV and MEP alignments of MTX and DHF in the keto form respectively. Figures 8a-c show the equivalent alignments of the enol form of DHF.

Crystallographic structures of both of these molecules bound to *L. Caseii* DHFR have long been a test bed for molecular modeling applications. According to Kearsley, there are two possible overlays for the heterocyclic rings in these two molecules. The chemically intuitive alignment overlays the fused 6,6 ring systems, with the oxygen of DHF matched to the ortho amino of MTX. This is the alignment given by Hex when DHF is in the enol form (Fig. 8a).

However, the crystallographically observed alignment differs. Here the oxygen of DHF is aligned with the pyridine-like nitrogen in the other ring. This alignment is seen in the Hex and both of the ParaMatch alignments (Figs. 7a-c). The consistency of these matches with the observed alignment would lend support to the use of the keto form by Kearsley. It is of interest to note here that both the ParaMatch alignments using the enol form have the rings swapped over (i.e., the connecting linkers, represented here by the methyl group, are at opposite ends of the aligned molecules). It is also interesting that the ParaMatch alignments, using either the CRV or MEP descriptors, are essentially the same, suggesting that the fit is good both in steric and electrostatic terms.

The second DHFR alignment is taken from Robinson *et al*. [20]. This particular alignment proved problematic for many similarity optimizers due to the larger size of the para substituent of the phenyl ring. The structures are in Fig. 9.

The Hex canonical alignment (Fig. 10a) fails on this example. The di-aminopyrimidine rings are located at the opposite ends of the molecules in the alignment. This would probably be corrected by using the spherical harmonic coefficients of the MEP expansion (rather than the expansion of the radial function used here). Indeed, a combination of the two expansions may give the correct alignment. However, there is another problem relating to the centers of gravity of the molecules. The spherical harmonic expansions assume that the centers of gravity of the two molecules are coincident. Thus, the shorter molecule lies within the confines of the larger, despite the fact that it is only at the extremities that they differ.

This illustrates an important feature of ParaMatch, i.e., that overlaying the centers of gravity is not a requirement. This allows for the possibility of partial matching of the molecular features, which can be of vital importance in some overlaying exercises. An obvious example would be in alignment prior to a 3D QSAR approach such as CoMFA.

Example3 : HIV-1 reverse transcriptase inhibitors

This example is the high-diversity alignment, from Robinson *et al.* [20], describing the two non-nucleoside inhibitors of HIV Reverse Transcriptase, α−APA and Nevirapine. The structures are shown in Fig. 11.

The major feature of the crystallographic alignment is the counter-intuitive match of the cyclopropyl ring in NEV with the amide group in APA. This would suggest that the shape features of this receptor are more important than electrostatic effects.

The Hex canonical alignment (Fig. 12a) gives a good description of the two diverse molecules based on their shape. The ParaMatch alignment based on the CRV descriptor shows a similar behavior. Neither of these alignments shows the correct orientation of the cyclopropyl and amide groups. Of more interest is the ParaMatch alignment based on the MEP. This shows the cyclopropyl ring of NEV being on the same side of the alignment as the amide group of APA. The translation makes this a very

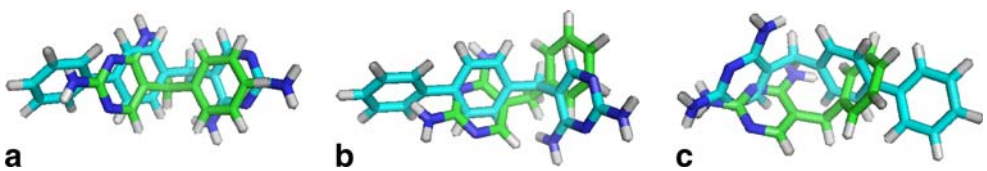

Fig. 10 Overlays of low-diversity DHFR structures. (**a**) Hex canonical; (**b**) ParaMatch using CRV; (**c**) ParaMatch using MEP

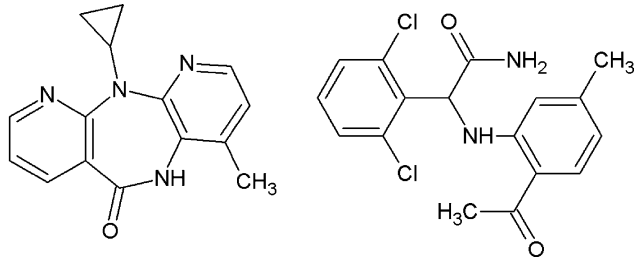

Fig. 11 Molecular Structures of Nevirapine and alpha-APA

poor overlay, but it suggests that the combination of descriptors, as discussed above, may give rise to a chemically meaningful result.

Discussion

The ParaMatch software described here shows promise in addressing the issue of alignment of molecules for virtual high throughput screening or 3D QSAR. This is particularly the case when used in conjunction with very fast alignment procedures such as Hex. This observation also suggests another role for ParaMatch in high throughput alignment studies. The need for speed in these calculations is critical due to the large numbers of structures needing to be assessed. The canonical alignment in Hex gives a very good approximation to the alignment in a short amount of CPU time (owing to the use of computationally efficient rotation functions). However, it will always superimpose the molecular centers of gravity. It is proposed that ParaMatch could be used in an optimization mode in which the Hex alignments could be tweaked using ParaMatch. This would not require the extensive systematic search of the ParaMatch alignments reported here and would therefore significantly reduce the computational requirement. At the simplest level, ParaMatch could be run with the rotation increment set to zero, i.e., ParaMatch would perform a simple translational optimization. In practice, it is envisioned that a small number of rotations would be performed, at small angle increments, about the Hex solution.

The identification of the matching voting pairs could be improved. Currently, this is done using a simple tolerance of the difference in the value of the property (Eq. (1)). It is not clear, for instance, whether two minima in similar positions but with different depths would be recognized as a voting pair in the current implementation. A better choice of the voting pairs would also improve the algorithm. An example would be to use the critical points (maxima and minima) of the surface properties as the initial grid from which the voting pairs are selected. Thus only chemically significant surface points would be considered for inclusion in the voting table. In the current implementation, many of these potential voting pairs will have been selected from the mid points of the property distribution. This should also improve the computational efficiency, since the number of potential voting pairs would be dramatically reduced.

The program currently performs the alignment using an exhaustive grid search about the cartesian axes. This over-represents possible alignments near the poles. A better option would be to use a table of angles resulting in points equidistant on the surface of a sphere. This should again lead to computational speed up. Note that, if ParaMatch is being used solely to optimize a local orientation from Hex then this would be of limited value.

The selection of voting pairs for use in the gravitational potential calculation relies on a tolerance (Eq. (1)) and the cut off applied to Eq. (3) (the latter to stop very close matches from dominating). The values used here are taken directly from reference [20]. Since these values relate to different footprints, (atomic co-ordinates), the values currently being used could be investigated and possibly optimized for property based footprints.

The major factor that determines the computational cost is the exhaustive search. At resolutions sufficient to get reliable, reproducible solutions, the CPU time is very long (hours) which is inappropriate for virtual high throughput screening. An optimization methodology could be developed to overcome this requirement. One possibility would be a coarse search followed by a local optimization such as a simplex or steepest descent method. A more sophisticated approach using simulated annealing or genetic algorithm techniques may also be possible. However, this is a complex problem, which is beyond the scope of the present work.

ParaMatch currently only performs alignments based on a single property. It would be useful to match on weighted

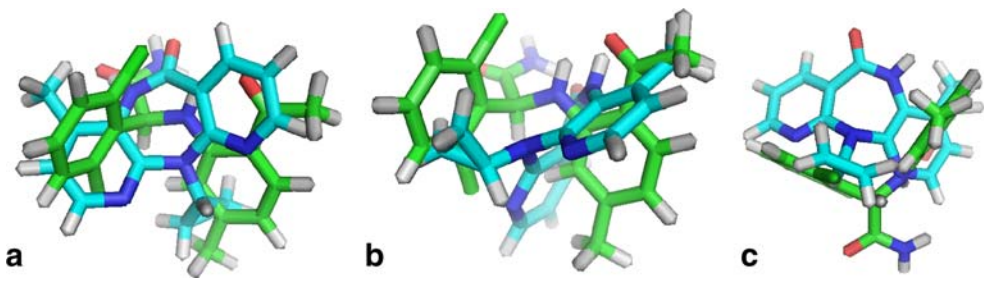

Fig. 12 Overlays of Nevirapine and alpha-APA. (**a**) Hex canonical; (**b**) ParaMatch using CRV; (**c**) ParaMatch using MEP

combinations of properties. For ParaMatch this is problematic as, unlike the spherical harmonic coefficients, the properties have different ranges.

Conclusions

The combination of the molecular surfaces with the pattern recognition algorithm shows utility in a number of typical tasks performed in the field of molecular overlay and superposition. The method is general in that it is not atom-based, uses quantum mechanics to derive its properties and, with the inclusion of reactivity parameters, is applicable in a wide range of chemical problems. In addition, it could be made more accurate by going to a higher level of theory, without any fundamental changes to the underlying algorithm, although it should be noted that ParaSurf is currently restricted to semi-empirical hamiltonians.

The use of local descriptions, such as those used by ParaSurf, would appear to have significant advantages over global approximations, and certainly over artificial atomistic models. The pattern recognition program ParaMatch can use these descriptions to perform molecular alignments. The advantages of this approach include molecular alignments where the center of gravity superposition approximation is not appropriate. The software can also be used to fine tune the results of highly computationally efficient structure alignment methods (such as Hex) at little extra cost.

The method has been shown to have potential in providing a generalized pattern recognition methodology applied to physically realistic quantum mechanical properties. The methods suffer, as do all 3D similarity methodologies, from the problems of conformational change and tautomerism. These issues are currently under consideration.

Acknowledgements BDH would like to acknowledge an Engineering & Physical Sciences Research Council (EPSRC) Basic Technology Initiative Proof of Concept award grant ref GR/S71477/01 and a Biological and Biophysical Sciences Research Council (BBSRC) Follow-on-Fund grant ref BB/E525985/1. We would also like to acknowledge many fruitful discussions with Prof T Clark, Prof WG Richards, Dr D Ritchie and Dr V Venkatraman.

References

1. Ehresmann B, Martin B, Horn AHC, Clark T (2003) J Mol Model 9:342–347
2. Clark T (2004) J Mol Graph Model 22:519–525
3. Ehresmann B, deGroot MJ, Alex A, Clark T (2004) J Chem Inf Comput Sci 44:658–668
4. Lin J, Clark T (2005) J Chem Inf Model 45:1010–1016
5. Nussinov R, Wolfson HJ (1991) Proc Natl Acad Sci USA 88:10495–10499
6. Arakawa M, Hasegawa K, Funatsu K (2003) J Chem Inf Comput Sci 43:1390–1395
7. Arakawa M, Hasegawa K, Funatsu K (2003) J Chem Inf Comput Sci 43:1396–1402
8. Bultinck P, Carbo-Dorca R, Van Alsenoy C (2003) J Chem Inf Comput Sci 43:1208–1217
9. Bultinck P, Kuppens T, Girone X, Carbo-Dorca R (2003) J Chem Inf Comput Sci 43:1143–1150
10. Girones X, Robert D, Carbo-Dorca R (2001) J Comput Chem 22:255–263
11. Girones X, Carbo-Dorca R (2004) J Comput Chem 25:153–159
12. Commandeur JJF, Kroonenberg PM, Dunn WJ (2004) J Chemometr 18:37–42
13. Kroonenberg PM, Dunn WJ, Commandeur JJF (2003) J Chem Inf Comput Sci 43:2025–2032
14. Exner T, Keil M, Brickmann J (2002) J Comput Chem 23:1176–1187
15. Ritchie DW, Kemp GJL (1999) J Comput Chem 20:383–395
16. Ritchie DW, Kemp GJL (2000) Proteins: Struct Funct Genet 39:178–194
17. Cai W, Zhang M, Maigret B (1998) J Comput Chem 19:1805–1815
18. Cai W, Shao X, Maigret B (2002) J Mol Graph Model 20:313–328
19. Barequet G, Sharir M (1997) IEEE Trans Pattern Anal Machine Intell 19:1–21
20. Robinson DD, Lyne PD, Richards WG (2000) J Chem Inf Comput Sci 40:503–512
21. Mavridis L, Hudson BD, Ritchie DR (2007) J Chem Inf Model 47:1787–1796
22. Chembank, http://chembank.broad.harvard.edu
23. Gasteiger J, Rudolph C, Sadowski J (1990) Tetrahedron Comp Method 3:537–547
24. Clark T, Alex A, Beck B, Burkhardt F, Chandrasekhar J, Gedeck P, Horn AHC, Hutter M, Martin B, Rauhut G, Sauer W, Schindler T (2005) VAMP 9.0, Erlangen
25. ParaSurf, Cepos insilico, http://www.ceposinsilico.com
26. geomview, http://www.geomview.org
27. pymol, DeLano Scientific, Palo Alto, CA, USA. http://pymol.sourceforge.net
28. Kearsley SK, Smith GM (1990) Tetrahedron Comp Method 3:615–633