Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Molecular Graphics and Modelling 23 (2004) 199–209

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

Alignment of three-dimensional molecules using an image recognition algorithm

Nicola J. Richmond[a,*], Peter Willett[a], Robert D. Clark[b]

[a] Department of Information Studies, Krebs Institute for Biomolecular Research,
University of Sheffield, Western Bank, Sheffield, S10 2TN, UK
[b] Tripos Inc., 1699 South Hanley Road, St. Louis, MO 63144, USA

Received 27 August 2003; received in revised form 6 April 2004; accepted 8 April 2004

Available online 18 August 2004

Abstract

This paper describes a novel approach, based on image recognition in two dimensions, for the atom-based alignment of two rigid molecules in three dimensions. The atoms are characterised by their partial charges and their positions relative to the remaining atoms in the molecule. Based on this information, a cost of matching a pair of atoms, one from each molecule, is assigned to all possible pairs. A preliminary set of intermolecular atom equivalences that minimises the total atom matching cost is then determined using an algorithm for solving the linear assignment problem. Several geometric heuristics are described that aim to reduce the number of atom equivalences that are inconsistent with the 3D structures. Those that remain are used to calculate an alignment transformation that achieves an optimal superposition of atoms that have a similar local geometry and partial charge. This alignment is then refined by calculating a new set of equivalences consisting of atom pairs that are approximately overlaid, irrespective of partial charge. A range of examples is provided to demonstrate the efficiency and effectiveness of the method.

Keywords: Linear assignment; Molecular alignment; Molecular overlay; Molecular superimposition

1. Introduction

The alignment of two molecules is one of the most important and common tasks in chemoinformatics. For example, aligning several structurally disparate molecules that exhibit a common bioactivity, so as to identify the topological or geometrical arrangement of the common features and hence to obtain a putative pharmacophore pattern. Further such applications include aligning the top-ranked nearest neighbours in a similarity search with the target structure, so as to facilitate visual inspection of the search output; aligning the surface of a small molecule with the binding site of a protein to determine the extent of the complementarity and hence of the probability that the small molecule might function as a ligand; aligning a set of molecules to some structural template as a prelude to a full 3D QSAR analysis; and aligning the reactant and product molecules in a chemical reaction so as to identify the substructural change that has occurred.

It is hardly surprising that this wide range of possible applications has spurred the development of many different alignment methods. An excellent review of the field is provided by Lemmen and Lenguaer [1], and new procedures continue to appear for this purpose [2,3]. One may describe the available techniques in terms of three broad and interrelated characteristics: the type of representation, which determines the types of feature that are to be aligned (e.g., a 2D connection table, 3D atomic coordinates, a smoothed bounded distance matrix, a Connolly surface, a molecular electrostatic potential field); the type of algorithm used to optimise the alignment (e.g., clique detection, genetic algorithm, simulated annealing, tabu search); and the criteria used to determine how good the alignment obtained is (e.g., size of a maximum common subgraph, root mean squared deviation between paired points, multi-dimensional scaling-like stress factor, force field energy calculation, or Dice similarity coefficient).

In this paper, we describe a new alignment method that has been developed for superimposing molecules represented by

* Corresponding author. Current address: Cheminformatics, GlaxoSmithKline, Medicines Research Centre, Gunnels Wood Road, Stevenage, SG1 2NY, UK.
E-mail address: nicola.j.richmond@gsk.com (N.J. Richmond).

1093-3263/$ – see front matter © 2004 Published by Elsevier Inc.
doi:10.1016/j.jmgm.2004.04.004

their atomic coordinates in 3D space. Pairs of atoms, one from each molecule, are tentatively matched using an algorithm for solving the linear assignment problem, the objective being to calculate an alignment transformation that achieves the best superposition of atoms with similar local geometry and partial charge. This preliminary atom matching is checked for geometrical consistency, edited as appropriate and then refined by matching atoms based exclusively on their spatial arrangement and not on their atomic characteristics.

Our approach yields crisp atom-based alignments, qualitatively similar to manual atom-based alignments, but which are often quite different from those obtained using other automated techniques. Furthermore, the method allows for a degree of atomic mismatch between structures that would otherwise complicate or preclude alignment by substructure alone.

The remainder of the paper is structured as follows. In Section 2, we provide a detailed description of the algorithm, which includes an introduction to the linear assignment problem (LAP), followed by a discussion of the basic alignment method; in particular we focus on how matching atoms in one molecule to atoms in another molecule is an instance of the LAP. We also describe how one can calculate an alignment transformation whose application minimises the sum of the distances between the atoms in one molecule and their respective matches in the other molecule. As the set of atom equivalences may contain outliers, we describe several heuristics that filter inappropriate matches in order to achieve an optimal overlay. In Section 3, we demonstrate the effectiveness and efficiency of the method with several examples. The paper concludes, in Section 4, with a summary of our findings and suggestions for future work.

2. Methodology

The objective of the alignment algorithm is to superimpose atoms in one molecule M onto similar atoms in another molecule N. It consists of four steps.

Step 1. Identify a set of candidate atom equivalences (a_i, b_j), where the a_i are atoms in M, the b_j are atoms in N and a_i is matched with b_j if they are both similar in terms of local geometry and partial charge.

Step 2. Filter the set of candidate atom equivalences by identifying and discarding those pairs that cannot be overlaid by any alignment transformation required for the majority.

Step 3. Calculate an alignment transformation that superimposes the molecules so that pairs in the filtered set of atom equivalences are overlaid.

Step 4. Refine the alignment by calculating a new set of atom equivalences—matching a_i with b_j if the distance between a_i and b_j is less than a user-defined threshold, and then computing a second alignment transformation based on this new set.

2.1. The linear assignment problem

The linear assignment problem is one of the most important in combinatorial optimisation and can be defined informally as the task of assigning n computers to n jobs so that all jobs are carried out as efficiently as possible [4]. Every computer must be assigned a job, and no job can be assigned more than once. This can be formulated mathematically as follows.

Given an $n \times n$ cost matrix $C = (c_{ij})$, where $c_{ij} \in \mathbb{R}$, $c_{ij} \geq 0$ is the cost of assigning row i to column j, find a one-to-one assignment of the rows to the columns (r_i, c_j) so that the total cost $\sum_{i=1}^{n} c_{ij}$ of the row–column assignment is minimised. Equivalently, find a permutation p of the integers $\{1, \ldots, n\}$, such that $\sum_{i=1}^{n} c_{ip(i)}$ is minimised.

For example, suppose that C is the positive integer-valued cost matrix given by

$$\begin{pmatrix} 1 & 6 & 2 & 5 \\ 2 & 1 & 0 & 4 \\ 1 & 3 & 2 & 3 \\ 0 & 5 & 1 & 6 \end{pmatrix},$$

whose *ij*th entry is the cost for computer i to complete job j. Then by inspection, one optimal assignment is computer 1 to job 1, computer 2 to job 2, computer 3 to job 4 and computer 4 to job 3. More formally, the solution to the LAP, with cheapest cost 6, is given by the permutation p_1. Note that there is an alternative optimal assignment given by the permutation p_2, also with cost 6. So solutions to the linear assignment problem are not necessarily unique.

$$p_1 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 4 & 3 \end{pmatrix}, \; p_2 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 4 & 1 \end{pmatrix}.$$

In this example, there are just $4! = 24$ possible permutations, so one could adopt a brute force approach and try every possible permutation to determine which yields the lowest cost. However, in practice, we could be dealing with very large cost matrices, so one requires a sophisticated algorithm to solve the LAP. There are many such algorithms described in the literature. We have chosen the Jonker–Volgenant shortest augmenting path algorithm which is of complexity $O(n^3)$ [5].

2.2. The basic alignment method

The method we have developed is inspired by a recent computer vision paper [6] in which Belongie et al. describe a technique for matching 2D shapes represented by a set of points sampled from the shape contours. Their approach consists of three phases. The first phase is to identify a one-to-one correspondence between the points describing shape A and the points describing shape B. The second phase is to calculate a morphing transformation that maps the points on shape B, to their corresponding points on shape

A. The third and final phase is to determine the similarity of the two shapes by measuring the sum of the matching errors between corresponding points and the magnitude of the morphing transformation. We proceed with a brief description of the relevant aspects of the shape matching procedure.

2.2.1. Matching 2D shapes

Each shape is represented by a discrete set of n points $\{p_1, p_2, \ldots, p_n\}$, $p_i \in \mathbb{R}^2$, sampled from the internal or external contours on the shape, obtained by an edge detector. Obviously, the greater the number of points, the more accurate the description of the shape.

Each point p_i of a shape is described by a coarse histogram $\mathcal{H}_i$, whose jth bin $\{h_{ij}\}$, is the number of points in the jth sector of a log polar grid centred about p_i. The objective is to match points, one from each shape, that have similar histograms. The cost of matching point $p(A)_i$ of shape A to point $p(B)_j$ of shape B is calculated as

$$c_{ij} = \frac{1}{2}\sum_{k=1}^{N} \frac{(h(A)_{ik} - h(B)_{jk})^2}{h(A)_{ik} + h(B)_{jk}},$$

where $h(A)_{ik}$ and $h(B)_{jk}$ denote the kth bins of the normalised histograms of shapes A and B, respectively.

If two histograms are very similar, their respective points will have a low cost of being matched. Thus, given the set of costs c_{ij} between all possible pairs, the aim is to minimise the total cost of matching points on shape A to points on shape B, subject to the constraint that each point on shape A must be matched to a different point on shape B. This is an instance of the linear assignment problem with cost matrix $C = (c_{ij})$. If the number of sample points on the two shapes is different, then the cost matrix is squared up by adding sufficiently many rows or columns, each element of which is a large dummy number.

For the purposes of the discussion, we illustrate the shape matching procedure by comparing the two forms of the letter “A” shown in Fig. 1(a) and (b). The log polar grid, Fig. 1(c), used to compute the histograms has 5 bins for $\log r$ and 12 bins for θ, giving a total of 60 bins. The histograms are calculated by placing the centre of the grid over each point, and then counting the number of the remaining points that lie in each sector. The histograms for the points marked by $\bigcirc$, $\triangleleft$ and $\diamond$ are shown in Fig. 1(d)–(f) respectively, where the darker the value, the higher the number of points in the corresponding sector. Notice the similarity between the histograms of $\bigcirc$ and $\diamond$, which are relatively similar points. In contrast, the histogram for $\triangleleft$ is quite different. The equivalences corresponding to the optimal solution of the linear assignment problem are shown in Fig. 1(g). Note that for the problem of character recognition, an absolute polar coordinate system is required, in order to avoid recognising, for example, $>$ as $<$ or M as W.

As in the shape matching problem considered by Belongie et al., we are aligning objects, defined by a discrete set of points—the atoms. However, since all inter-atomic distances and angles must be strictly conserved, we are restricted in the types of transformations that are allowed. Hence any alignment transformation we calculate must be an isometry, preserving molecular chirality, and so cannot involve a reflection—a morphing transformation in this situation would not be appropriate.

Despite these differences, we adopt the basic idea of the shape matching approach by: first determining a set of one-to-one correspondences between the atoms in one molecule and the atoms in the other molecule; and second, using these correspondences to calculate an isometry, known as a Procrustes transformation, that superimposes the first molecule onto the second molecule so that equivalent atoms are overlaid.

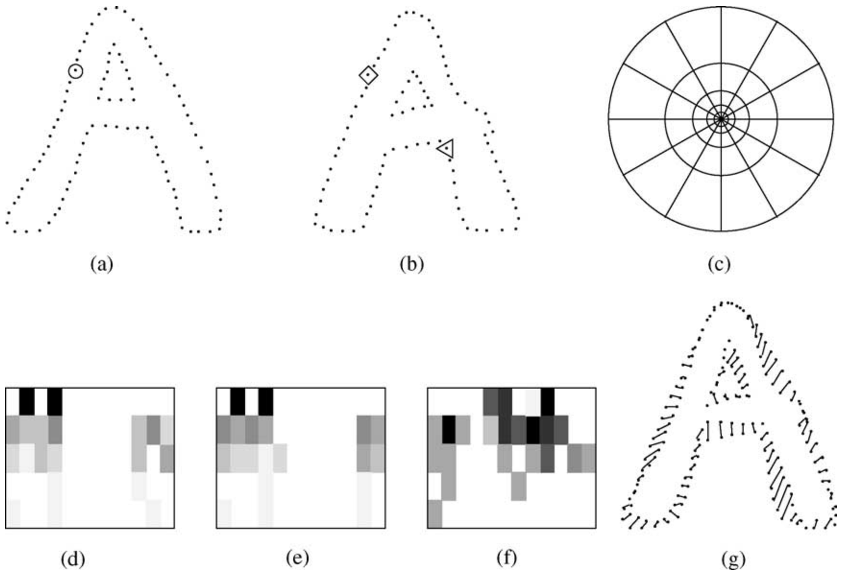


Fig. 1. Shape matching in 2D.

2.2.2. Determining atom equivalences

The first modification we make is to the shape context descriptor, i.e. the set of histograms, one for each point, binned by a log polar grid (r, θ), centred about the point of interest. Since we are working in three dimensions, the obvious generalisation is a set of histograms, one for each atom, binned by a log spherical grid (r, θ, ϕ) which is centred about the atomic coordinate of interest. However, since molecules can be presented in any 3D orientation, this binning system is inappropriate as we cannot sensibly define the x, y and z axes for the grid. So we adopt a simple radial approach as follows.

Let M be a molecule represented by the set $\mathcal{A} = \{a(M)_1, \ldots, a(M)_n\}$ of its constituent atoms $a(M)_i = (p(M)_i, c(M)_i)$, where $p(M)_i \in \mathbb{R}^3$ is the atomic coordinate and $c(M)_i \in \mathbb{R}$, the partial charge. Each atom $a(M)_i$ is described by the 20-binned histogram $\mathcal{H}(M)_i = \{h(M)_{ij}\}$, whose jth bin is the number of atoms in the molecule that are at a distance of between $j - 1$ and j Å from $a(M)_i$. Note that this descriptor does not make any use of bond information and is completely invariant under translation and rotation.

The cost of matching two atoms is a weighted sum of the absolute difference of their partial charges and the difference in their histograms. More formally, the cost c_{ij} of matching atom $a(M)_i$ in molecule M with atom $a(N)_j$ in molecule N is defined by

$$c_{ij} = w_1|c(M)_i - c(N)_j| + \sum_{k=1}^{K} \frac{(h(M)_{ik} - h(N)_{jk})^2}{h(M)_{ik} + h(N)_{jk}},$$

where $h(*)_{ik} = |\{a(*)_j \in \mathcal{A} : k-1 \le |p(*)_j - p(*)_i| < k\}|$ and $w_1 \in \mathbb{R}$ is the charge weight, 10 by default.

The cost function can be defined using any atomic property, or indeed, any suitable function thereof. However, we have chosen to use atomic partial charge as it contains implicit information about the connectivity of the atoms and substantial pharmacophoric information about the heteroatoms, in particular with regard to hydrogen bond donor and acceptor strengths. Pharmacophoric features could also be used, as is done in the GASP method [7]. However, molecules typically contain few such features and since these features tend to be properties of substructures, they must be represented by a centroid. Consequently, the resulting align-

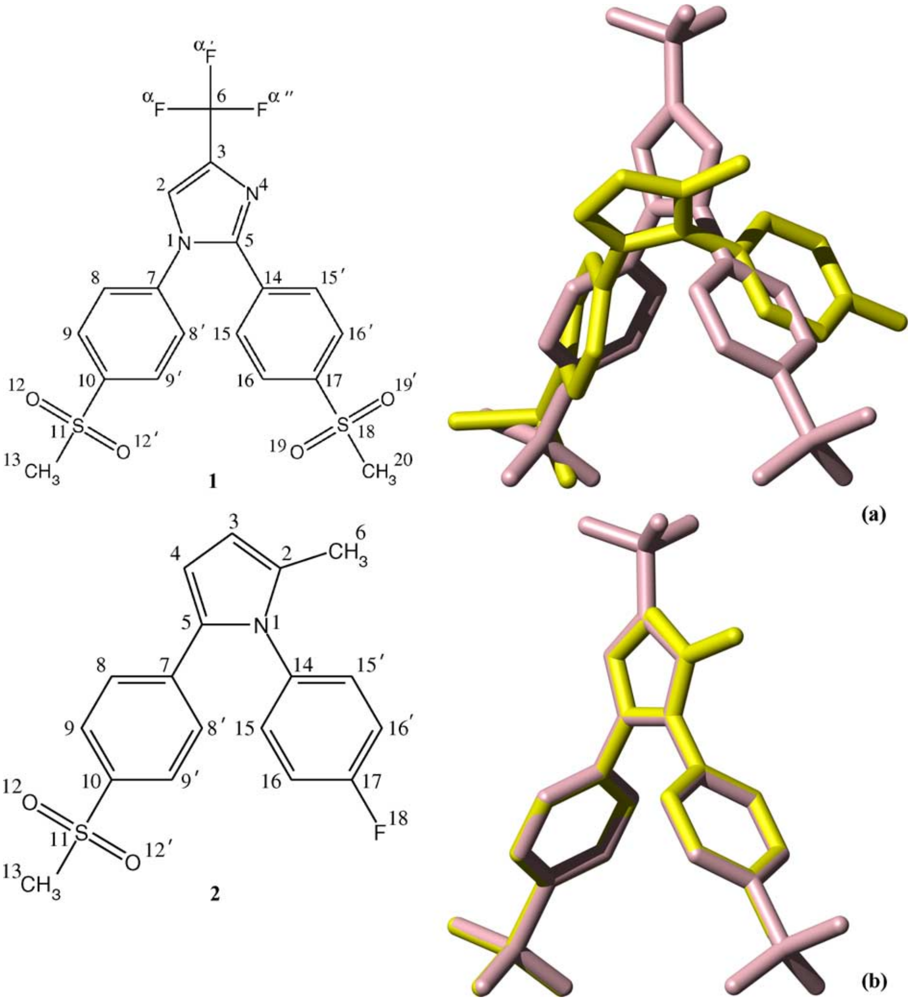


Fig. 2. Aligning COX2 inhibitors with filters.

ments may lack the distinctive crispness of those obtained using atomic partial charge.

Recall that the linear assignment problem is the task of assigning each row to a unique column of a square matrix so as to minimise the sum of the row–column assignments. In most cases, M and N will have a different number of atoms, so without loss of generality, we will assume that M is the larger molecule and define the cost matrix, $C = (c_{ij})$, with either *ij*th entry the cost of matching atom $a(M)_i$ in molecule M to atom $a(N)_j$ in molecule N, or a large positive real number DUMMY, typically 100 000, if j exceeds the number of atoms in N. Again, as in the shape matching procedure, the set of atom equivalences is a solution to the LAP with cost matrix C.

Invariably, not all the atom equivalences will be appropriate, so before the alignment transformation is calculated, we apply several filters to the solution to ensure that only the best matches are overlaid. For clarity and ease of exposition, we first describe the algorithm used to calculate the reflection-free, Procrustes transformation before discussing, in Section 2.3, the filters we have developed.

2.2.3. The alignment transformation

Using an appropriate set of atom equivalences, we calculate an alignment transformation that superimposes one complete molecule onto the other, so that identified equivalent atoms are optimally overlaid. Since we are dealing with rigid molecules, this transformation must preserve the molecular conformation and stereochemistry. So we require a Procrustes transformation that does not include a reflection.

Let $\{(a(M)_1 \sim a(N)_1), \ldots, (a(M)_r \sim a(N)_r)\}$ be the filtered set of atom equivalences. We calculate the Procrustes transformation using the algorithm described by Rohlf and Slice [8], of complexity $O(r)$. Let P_M and P_N be the $r \times 3$ matrices whose jth rows are the coordinates of atoms $a(M)_j$ and $a(N)_j$ respectively.

Step 1. Apply translations T_M to P_M and T_N to P_N, such that the $a(M)_i$ and $a(N)_i$ are centred about the origin.

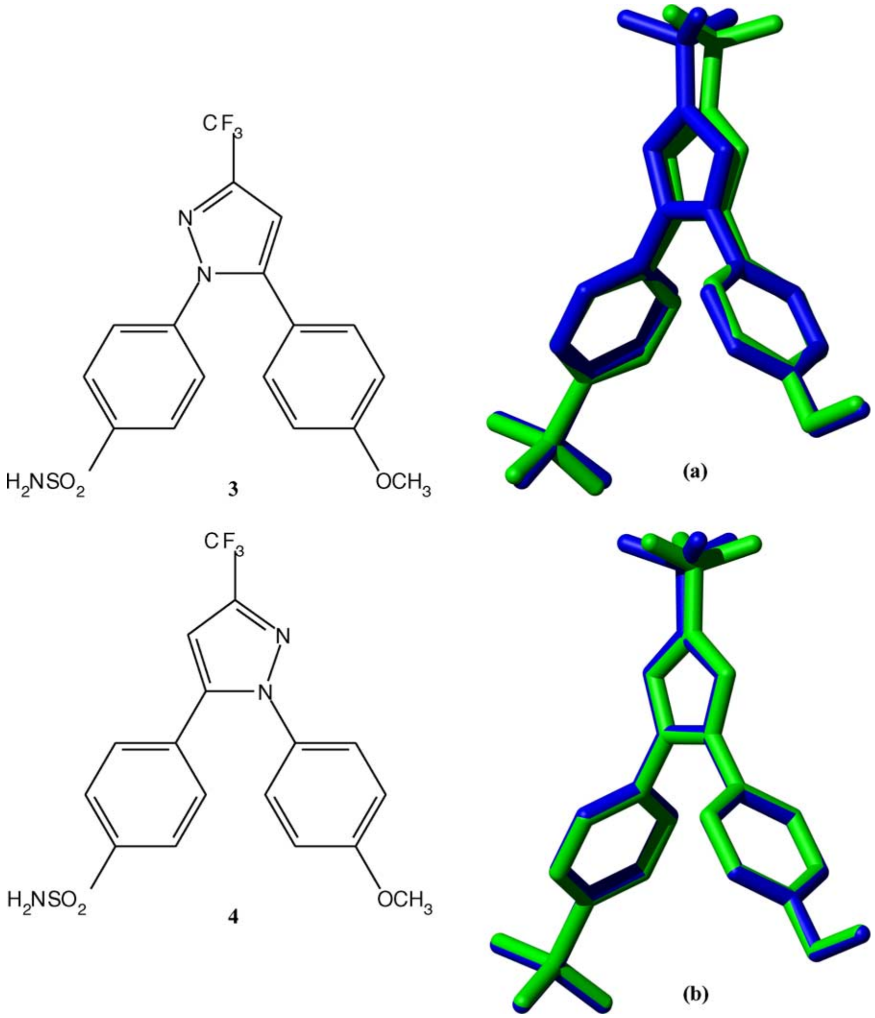


Fig. 3. Alignment of COX2 inhibitors with refinement.

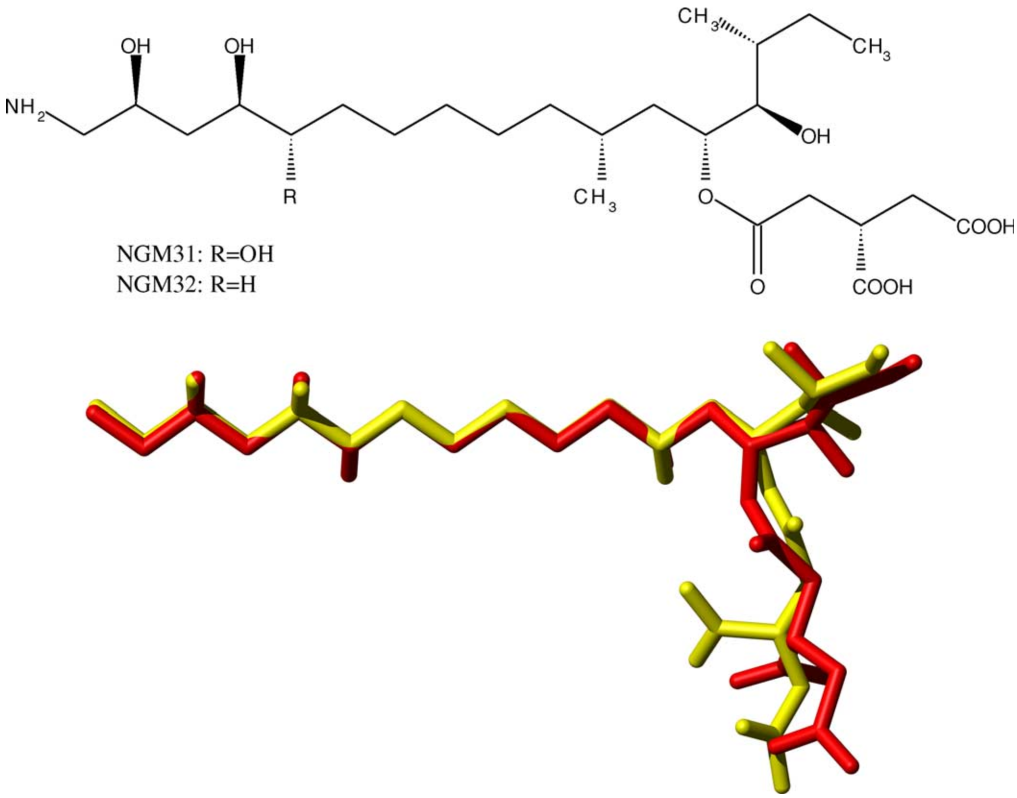


Fig. 4. Alignment of NGM31 (red) with NGM32 (yellow) in approximately 0.03 seconds. The non-superposition on the right is due to the different conformations.

We will denote the translated coordinates of M and N by P'_M and P'_N respectively.

Step 2. Calculate the singular-value decomposition of $P'^{t}_{M}P'_{N}$, which is of the form $U\Sigma V^{t}$, where U and V are 3×3 orthogonal matrices, $\Sigma = (\sigma_{ii})$ is a 3×3 diagonal matrix of positive eigenvalues with $\sigma_{ii} \le \sigma_{jj}$ if and only if $i < j$, and where t denotes the transpose operator.

Step 3. Calculate $H = VU^{t}$.

Step 4. If the determinant of $P'^{t}_{M}P'_{N}$ is negative, multiply the last column of V by -1, then return to *Step 3* and terminate.

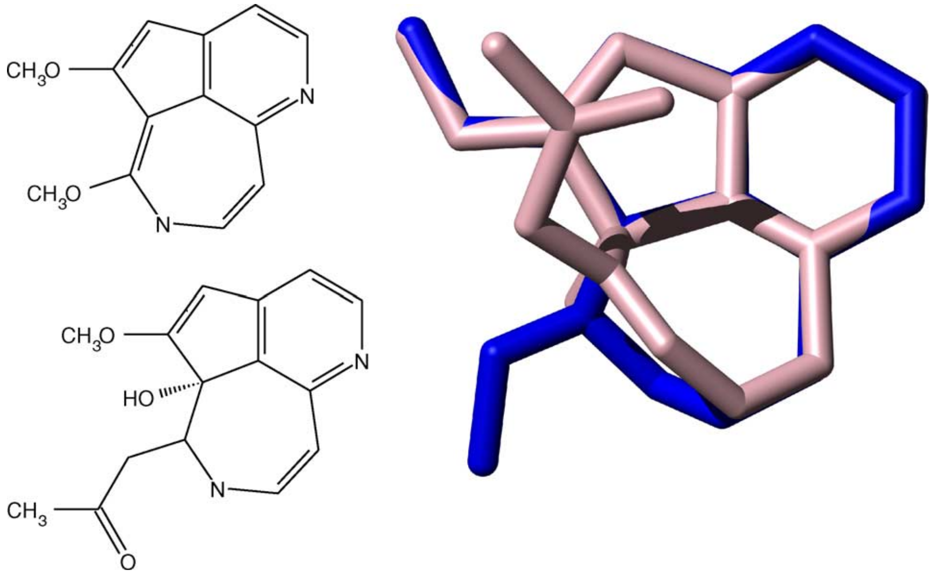
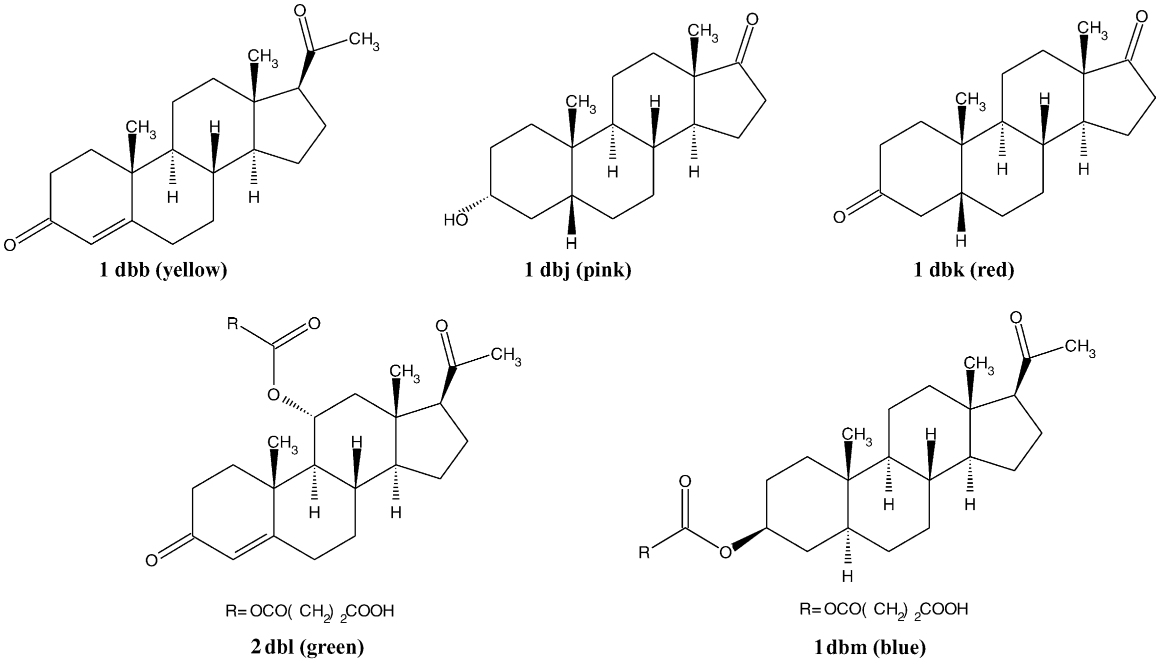


Fig. 5. Alignment of FYG37 (cyan) with MKM92 (magenta) in approximately 0.007 seconds.



1 dbb (yellow)

1 dbj (pink)

1 dbk (red)

R=OCO(CH_2) $_2$COOH

2 dbl (green)

R=OCO(CH_2) $_2$COOH

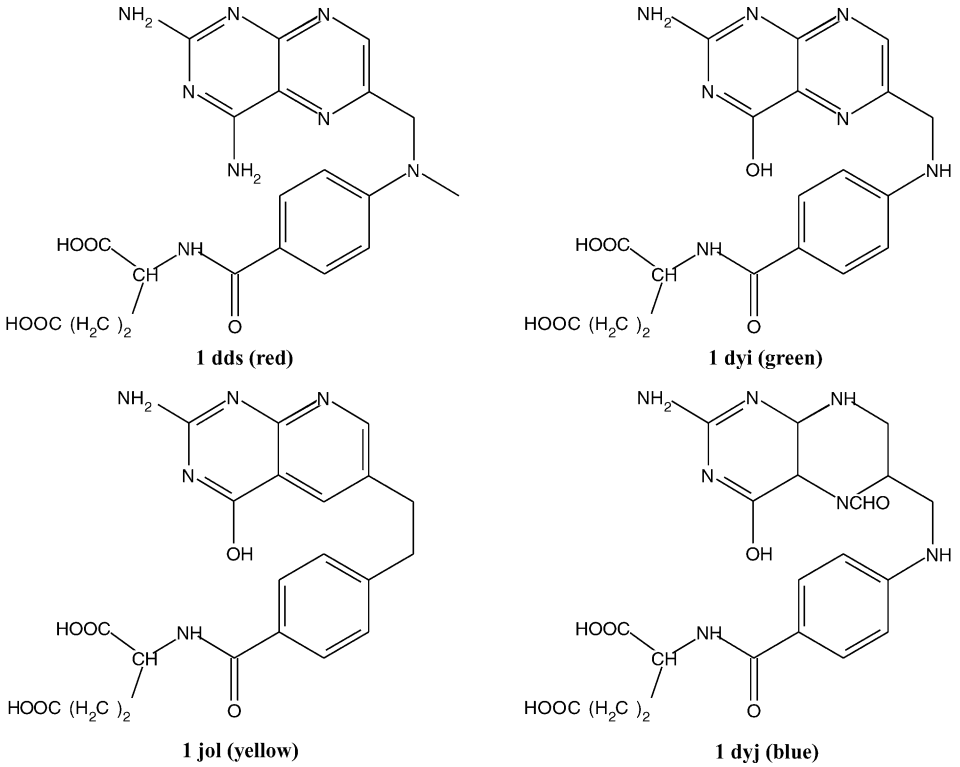
1 dbm (blue)

Fig. 6. Ligands from mouse immunoglobulin complexes.

To superimpose N onto M so that the atoms $a(N)_i$ and $a(M)_i$ are overlaid, we simply translate N by T_N, rotate by H and then translate by T_M^{-1}. The result is an alignment of M and N, such that the sum of the distances between the atoms identified as being equivalent, is minimised.

2.3. Filtering the set of atom equivalences

The Procrustes transformation is calculated based on a specific set of atom equivalences. If, as is often the case, this set contains pairs of atoms that have been inappropriately matched in terms of geometry, then the resulting transfor-



1 dds (red)

1 dyi (green)

1 jol (yellow)

1 dyj (blue)

Fig. 7. Ligands from *Escherichia coli* dihydrofolate reductase complexes.

mation will not yield a maximal overlay of atoms. Consider the case when a molecule has a high degree of symmetry, for example benzene. Each carbon atom in the ring is indistinguishable from the next. So the cost matrix for aligning benzene with any other molecule will have identical columns and hence many optimal LAP solutions.

In many cases, we can discard such geometrically inappropriate atom equivalences from the set by applying a series of filters. The most obvious, is to discard the trivial atom equivalences whose matching cost is higher than DUMMY, since a matching cost of DUMMY will generally indicate that one of the pair is a necessary "dummy" atom required because the two molecules differ in size.

Another filter aims at discarding those pairs that do not respect atomic distance. More precisely, suppose that $(a(M)_i \sim a(N)_i)$ and $(a(M)_j \sim a(N)_j)$ are two atom equivalences featuring in the linear assignment solution. Then, if the distance between $a(M)_i$ and $a(M)_j$ is much different from the distance between $a(N)_i$ and $a(N)_j$, one can conclude that no isometry can overlay both $a(M)_i$ and $a(N)_i$ in addition to $a(M)_j$ and $a(N)_j$. This approach is similar to the "refine" procedure of the Ullmann algorithm,

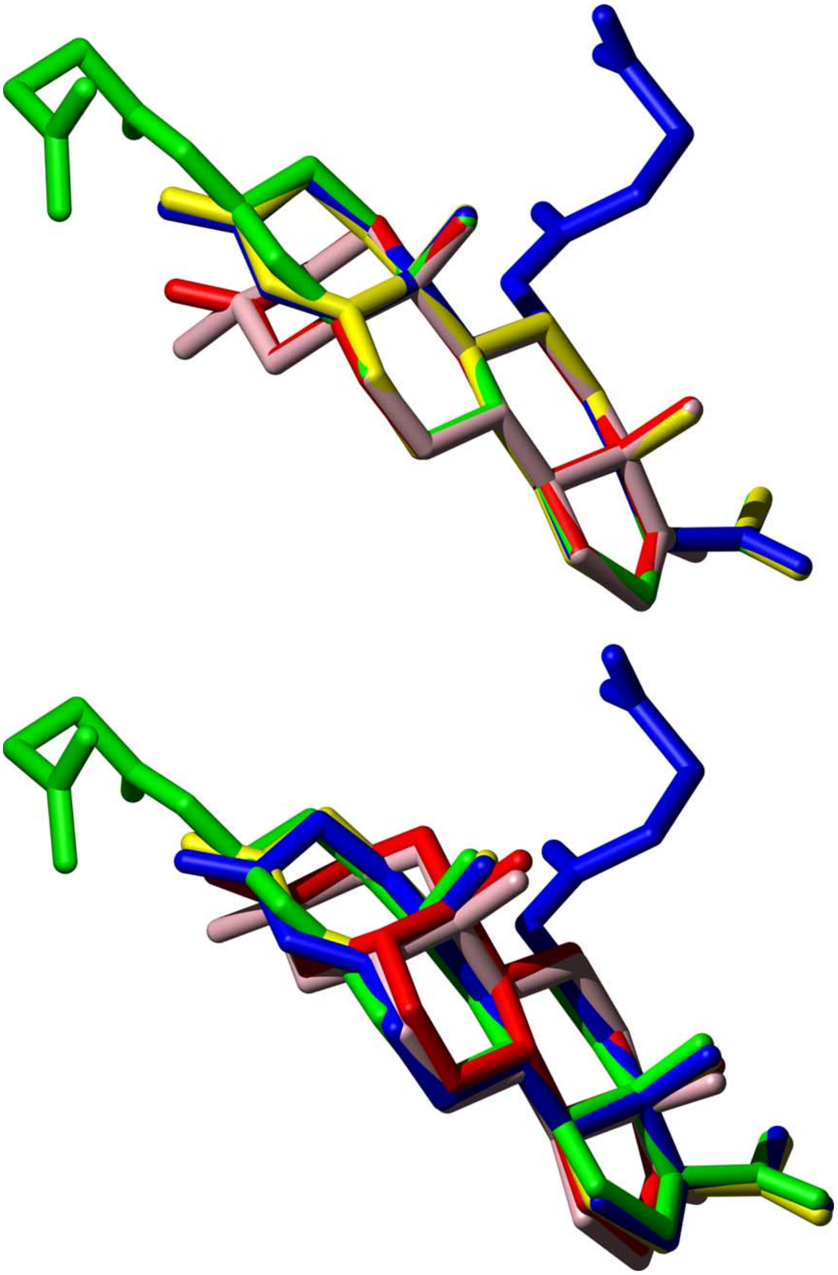


Fig. 8. A comparison of the alignments obtained using our method (top) and Field-Fit (bottom) of ligands aetiocholanolone (1dbj), 5-β-androstane-3,17-dione (1dbk), 5-α-pregnane-3-β-ol-hemisuccinate (2dbl), progesterone-11-α-ol-hemisuccinate (1dbm), with reference ligand progesterone (1dbb) from mouse immunoglobulin complexes, all with minimised crystal structure conformations, in respective average times of 0.012 and 20 second per pair.

when applied to the problem of 3D substructure searching [9].

Let $\mathcal{E} = \{(a(M)_1 \sim a(N)_1), \ldots, (a(M)_r \sim a(N)_r)\}$ be the subset of the LAP solution consisting of all atom equivalences with a matching cost lower than DUMMY. To decide which equivalences from $\mathcal{E}$ do not respect atomic distance, we calculate the $r \times r$ difference matrix $D = (d_{ij})$ whose rows and columns are indexed by the $a(M)_i$ and $a(N)_j$ respectively, and whose *ij*th entry is given by $d_{ij} = ||a(M)_j - a(M)_i| - |a(N)_j - a(N)_i||$. We score atom equivalence $(a(M)_i \sim a(N)_i)$ by counting the number of entries in the *i*th row of D that are greater than zero, up to some tolerance threshold, typically 0.1 Å. The whole set is then ranked according to this score—low being good—and the equivalences that feature at the bottom of the list are discarded.

To determine the degree of symmetry of a molecule, we calculate the cost matrix of a molecule with itself to identify pairs of atoms that are similar to each other. As we want to catch local symmetry, the difference in histograms is only calculated for the first five bins. If the *ij*th entry of the cost matrix is zero, up to some small tolerance. typically 0.1, then atoms i and j of the molecule will potentially be indistinguishable from each other. Any atom equivalences featuring i and j will then be discarded.

The need to compensate for degeneracy before calculating an alignment transformation is well illustrated by considering the difficulties of superimposing vicinal diaryl inhibitors of the COX2 enzyme [10,11]. The COX2 inhibitors **1** and **2**, shown in Fig. 2, have a high degree of symmetry. The method correctly identifies that atoms $8, 8', 9, 9', 12, 12', 15, 15', 16, 16', 19, 19', \alpha, \alpha''$ all have some symmetry (see Fig. 2(1)). Hence any pairs featuring the above atoms will be filtered from the set and thus, will not have an influence on the calculation of the Procrustes transformation. Without applying these filters, we achieve the unsatisfactory alignment of **1** and **2**, shown in Fig. 2(a), obtained from atom equivalences (1,4), (2,2), (3,7), (4,1), (5,5), (6,6), (7,14), (8,3), $(8', 9')$, (9,8), $(9', 16')$, (10,10), (11,11), $(12, 12')$, $(12', 12)$, (13,13), (14,9), $(15, 15')$, $(15', 15)$, (16,16), $(16', 8')$, (17,17), $(18, \infty)$, $(19, \infty)$, $(19', \infty)$, $(20, \infty)$, (α, ∞), $(\alpha', 18)$, (α'', ∞). The more satisfactory alignment, shown in Fig. 2(b), is obtained from atom equivalences (10, 10), (11, 11), (13, 13), (17, 17) after applying the filters we have just described.

Once the molecules have been superimposed, the last step is to refine the alignment. This is achieved by recalculating the set of atom equivalences, where an atom is now paired up with its nearest transformed neighbour, if the distance between the two is less than a certain threshold, typically 0.7 Å. A second Procrustes transformation is then computed based on this new set of atom equivalences and applied to the second molecule.

In most cases, the filters are sufficient to achieve the desired alignment and so this procedure is often unnecessary. However, as the example of the alignment of **3** and **4** shown in Fig. 3(a) illustrates, inappropriate atom matches can be missed, resulting in an overlay that is slightly flawed. Essentially, the refinement step calculates an adjustment that can enhance the crispness of the overlay, as depicted in Fig. 3(b).

3. Results and discussion

Further test cases and typical timings for the method described here include alignments of molecules, with CONCORD conformations, from the Dictionary of Natural Products [12] and ligands, with CONCORD and/or minimised crystal structure conformations, from complexes in

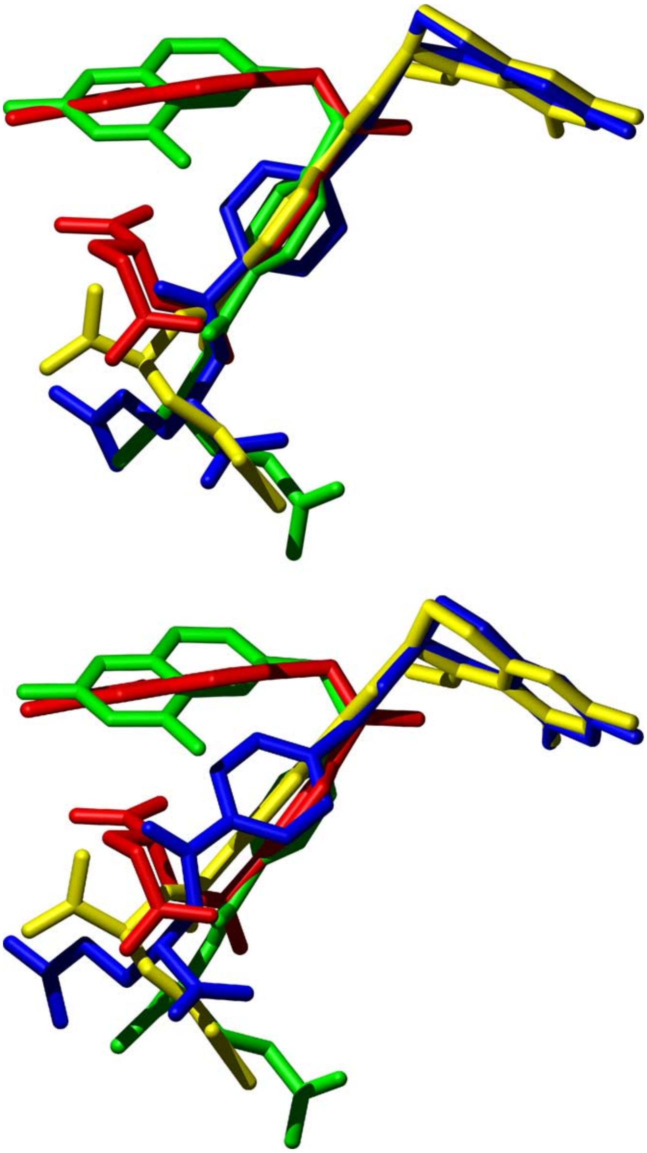


Fig. 9. A comparison of the alignments obtained using our method (top) and Field-Fit (bottom) of ligands methotrexate (1dds) with folate (1dyi), 5-formyl-6-hydrofolic acid (1jol) with 5,10-dideazafolate (1dyj) and methotrexate (1dds) with 5-formyl-6-hydrofolic acid (1jol) from *E. coli* dihydrofolate reductase complexes, all with minimised crystal structure conformations, in respective average times of approximately 0.023 and 25 second per pair.

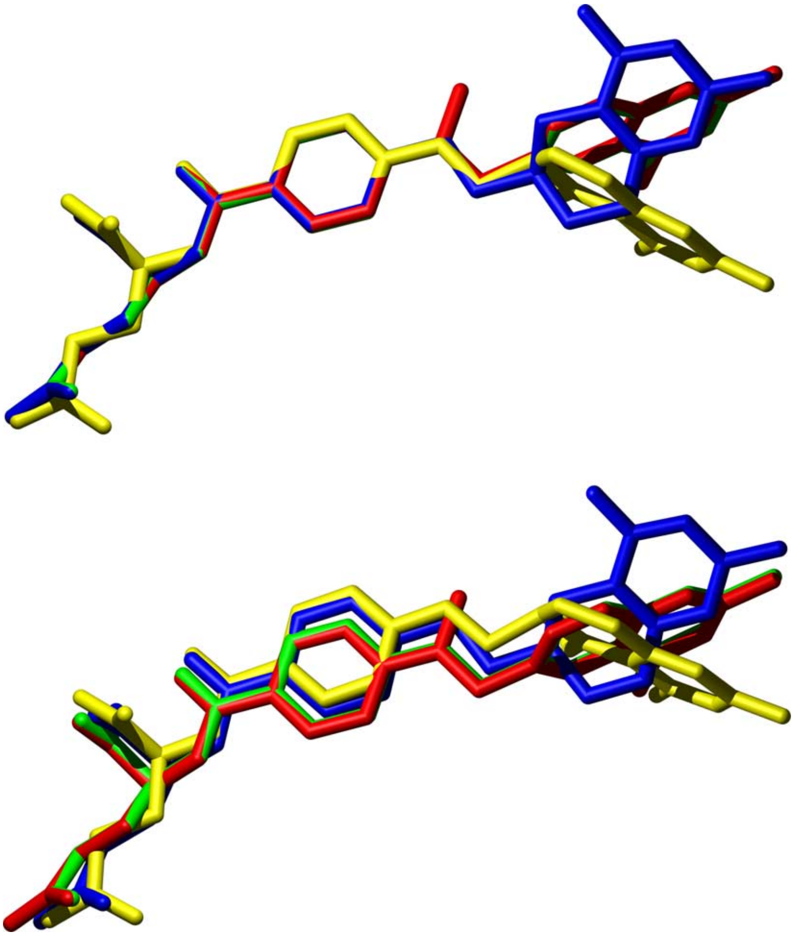


Fig. 10. A comparison of the alignments obtained using our method (top) and Field-Fit (bottom) of ligands from the *E. coli* dihydrofolate reductase dataset with CONCORD conformations, in respective average times of approximately 0.023 and 27 second per pair.

the Protein Data Bank [13]. All partial charges are calculated using the Gasteiger method in SYBYL [14]. We show images of the alignments in Figs. 4–10, and also include timing information. All experiments, including those using SYBYL, have been carried out on a Pentium III, 800 MHz processor, with 640 MB RAM.

For comparison, we have also aligned the same structures using SYBYL's "Field-Fit" tool, which is an automatic flexible alignment method for use with the 3D QSAR model builder CoMFA [15]. The "Field-Fit" algorithm automatically aligns a dataset by minimising the RMS difference in steric and electrostatic molecular fields, as sampled at points on a 3D lattice. By using molecular fields, "Field-Fit" explicitly avoids the need to identify atom equivalences between molecules, as is often the case with other 3D alignment methods, and thus provides a basis of comparison with the algorithm described in this paper. Since "Field-Fit" requires as input, a dataset with substantial overlap, we have used our alignments as a starting point.

The examples we show demonstrate that our method generates effective structural alignments in a very efficient manner, typically requiring approximately two hundredths of a second per alignment. Despite the much greater computational requirement, typically three orders of magnitude, the "Field-Fit" alignments shown in Figs. 8–10 (bottom) are noticeably less crisp than those shown in Figs. 8–10 (top).

There are however, two outstanding issues. The first and more fundamental is that the Jonker-Volgenant linear assignment algorithm computes one optimal solution, whereas there could potentially be many. What would be preferable is an algorithm that returns all optimal solutions, and perhaps even some that are modestly suboptimal, from which the most appropriate set of atom equivalences can be selected. One such algorithm, described by Lawler [16], has been implemented and tested, but experiments demonstrate that it is excessively time-consuming and that the filters applied in *Step 2* of the algorithm are adequate, at least for the many examples that we have considered to date.

The second issue is probably common to all alignment methods, in that the code may fail to generate a sensible alignment of molecules that have a low degree of similarity. Indeed this is a more general problem with any type of similarity measure, alignment based or not: it is only appropriate to compare molecules if there is some structural resemblance between them. However, since the method only requires as few as three good atom equivalences to calculate an alignment transformation, it is very effective at superimposing molecules that differ substantially in size and also promotes the overlaying of atoms that represent homologous substitutions such as the methylene, nitrogen and oxygen in the piperidine, piperazine and morpholine groups, respectively.

4. Conclusions

In this paper, we have described a new method for aligning pairs of rigid 3D molecules, characterised by their atomic coordinates and partial charges. Elaborating on an algorithm described by Belongie et al. for 2D shape matching, the method provides an effective and efficient approach to 3D molecular alignment.

Three obvious applications suggest themselves: the mapping of multiple active molecules to identify their common pharmacophoric features; the use of the method for virtual screening by aligning a bioactive target structure with each molecule in a database, then ranking those structures in decreasing order of some similarity coefficient derived from the computed alignment; and the use of the method for 3D QSAR by fitting molecules in a dataset to a common template. We are currently investigating such applications, as well as ways to incorporate conformational flexibility into the procedure: this work will be reported shortly.

Finally, we wish to point out that there are many applications in chemoinformatics that involve the identification of similar structural characteristics. If a sensible cost matching function can be defined, then these tasks can often be formulated as instances of the linear assignment problems. There are other situations where one may wish to exploit a linear assignment algorithm. For example, in the RASCAL algorithm [17], the optimal assignment cost associated with matching pairs of nodes from two different graphs is used to decide if an expensive clique detection should be carried out. We believe that linear assignment is a powerful and general mechanism for investigating structural and substructural equivalences.

Acknowledgements

We would like to thank Tripos Inc., for their generous funding and Dr. Martin Bohl (Tripos Inc.), Dr. Gareth Jones (Arena Pharmaceuticals Inc.) and Prof. Ian Dryden (University of Nottingham) for helpful discussions. All 2D chemical sketches were produced using Chemtool [18] and all 3D figures were prepared with the program MOLMOL [19] and POV-Ray. The Krebs Institute for Biomolecular Research is a designated centre of the Biotechnology and Biological Sciences Research Council.

References

[1] C. Lemmen, T. Lengauer, Computational methods for the structural alignment of molecules, J. Comput. Aid. Mol. Des. 14 (2000) 215–232.

[2] M. Arakawa, K. Hasegawa, K. Funastsu, Novel alignment method of small molecules using the Hopfield Neural Network, J. Chem. Inf. Comput. Sci. 43 (2003) 1390–1395.

[3] F. Melani, P. Gratteri, M. Adamo, C. Bonaccini, Field interaction and geometrical overlap: a new simplex and experimental design based computational procedure for superposing small ligand molecules, J. Med. Chem. 46 (2003) 1359–1371.

[4] R.E. Burkard, E. Dragoti-Çela, Linear assignment problems and extensions, in: Z. Du, P. Pardalos (Eds.), Handbook of Combinatorial Optimization I, Kluwer Academic Publishers, Dordrecht, 1999, pp. 75–149.

[5] R. Jonker, A. Volgenant, A shortest augmenting path algorithm for dense and sparse linear assignment problems, Computing 38 (1987) 325–340.

[6] S. Belongie, J. Malik, J. Puzicha, Matching shapes and object recognition using shape contexts, IEEE Trans. Pattern Anal. Machine Intelligence 24 (2002) 509–522.

[7] G. Jones, P. Willett, R.C. Glen, A genetic algorithm for flexible molecular overlay and pharmacophore elucidation, J. Comput. Aid. Mol. Des. 9 (1995) 523–549.

[8] F.J. Rohlf, D. Slice, Extensions of the Procrustes method for the optimal superimposition of landmarks, Systematic Zool. 39 (1990) 40–59.

[9] P. Willett, Searching for pharmacophoric patterns in databases of three-dimensional chemical structures, J. Mol. Recog. 8 (1995) 290–303.

[10] P. Chavatte, S. Yous, C. Marot, N. Baurin, D. Lesiur, Three-dimensional quantitative structure-activity relationships of cyclo-oxygenase-2 (COX-2) inhibitors: a comparative molecular field analysis, J. Med. Chem. 44 (2001) 3223–3230.

[11] R.D. Clark, Boosted leave-many-out cross-validation: the effect of training and test set diversity on PLS statistics, J. Comput. Aid. Mol. Des. 17 (2003) 265–275.

[12] The Dictionary of Natural Products, available from Chapman & Hall/CRC, http://www.crcpress.com.

[13] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucleic Acids Research 28 (2000) 235–242.

[14] SYBYL is available from Tripos, Inc., 1699 S. Hanley Road, St. Louis, MO 63144, http://www.tripos.com.

[15] M. Clark, R.D. Cramer, D.M. Jones, D.E. Patterson, P.E. Simeroth, Comparative Molecular Field Analysis (CoMFA). Part II. Toward its use with 3D structural databases, Tetrahed. Comput. Method. 3 (1990) 47–59.

[16] E.L. Lawler, A procedure for computing the *k* best solutions to discrete optimization problems and its applications to the shortest path problem, Management Sci. 18 (1972) 401–407.

[17] J.W. Raymond, E.J. Gardiner, P. Willett, RASCAL: calculation of graph similarity using maximum common edge subgraphs, Comput. J. 45 (2002) 631–644.

[18] http://ruby.chemie.uni-freiburg.de/ martin/chemtool/chemtool.html1.

[19] R. Koradi, M. Billeter, K. Wüthrich, MOLMOL: a program for display and analysis of macromolecular structures, J. Mol. Graphics 14 (1996) 51–55.