ELSEVIER

Journal of Molecular Graphics and Modelling 20 (2001) 111–121

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

Automatic generation of alignments for 3D QSAR analyses

Nicholas E. Jewell [a], David B. Turner [a], Peter Willett [a,*], Graham J. Sexton [b]

[a] *Department of Information Studies and Krebs Institute for Biomolecular Research, University of Sheffield, Western Bank, Sheffield S10 2TN, UK*
[b] *Syngenta, Jealott's Hill Research Station, Bracknell RG42 6EY, UK*

Received 12 June 2000; received in revised form 16 March 2001; accepted 16 March 2001

Abstract

Many 3D QSAR methods require the alignment of the molecules in a dataset, which can require a fair amount of manual effort in deciding upon a rational basis for the superposition. This paper describes the use of FBSS, a program for field-based similarity searching in chemical databases, for generating such alignments automatically. The CoMFA and CoMSIA experiments with several literature datasets show that the QSAR models resulting from the FBSS alignments are broadly comparable in predictive performance with the models resulting from manual alignments. © 2001 Published by Elsevier Science Inc.

Keywords: 3D QSAR; Dataset; CoMFA and CoMSIA experiments

1. Introduction

Current approaches to the design of bioactive molecules make extensive use of 3D QSAR methods [1,2]. These methods seek to establish a statistically significant correlation between experimental biological activity data and structural variables characterising the geometric distribution in 3D space of properties associated with molecular recognition events. Examples of such methods include CoMFA [3], CoMSIA [4], COMPASS [5] and HASL [6].

An important component of many 3D QSAR methods is the need to align the molecules in a dataset as a precursor to the calculation of the structural variables. When all the members of a dataset contain a common structural feature, such as a rigid ring template or an obvious pattern of pharmacophore points, then the alignments can be generated easily using a least-squares fitting procedure. If, however, there is some degree of structural heterogeneity in the dataset then a large amount of time may be required at a molecular graphics terminal to obtain a satisfactory alignment for the dataset. The success of such a manual procedure will be strongly dependent on the experience of the modeller carrying out the alignment and will inevitably involve at least some degree of subjectivity in assessing which atoms (or ring centroids or whatever) should be fitted to which.

This paper describes a fully automated procedure for generating the alignments required for 3D QSAR, specifically the well-known CoMFA and CoMSIA methods. Our approach is based on FBSS (for field-based similarity searching), a program we have developed previously for 3D similarity searching in chemical structure databases, but which we have here applied to the generation of alignments for 3D QSAR. Although both similarity searching and 3D QSAR involve the generation of molecular alignments, the use of calculated molecular similarities as a basis for the latter application is clearly open to debate, since one can visualise cases where the parts of molecules that dominate the calculated similarity scores are not relevant to the important pharmacophoric features for that dataset. Accordingly, the aim of the work reported here is to provide an approach that is complementary to, rather than a replacement for, the manual alignments normally used in QSAR. Specifically, while the automatic alignments could be used directly as the input to a QSAR analysis, we believe that their main value may be as an initial screening mechanism in one of two ways. First, when a new dataset is to be analysed, an automatically-generated set of alignments can be processed using the 3D QSAR method of choice: if this initial, automated analysis results in a predictive QSAR model, then it may be worth the modeller spending time and effort to generate a manual set of alignments. Second, the automatic procedure may suggest non-obvious alignments for consideration by the modeller during a second, more detailed, manual analysis. Here, we focus on the use of the automatically-generated alignments on their own, to justify the potential of the approach; that said, its inherent limitations must always be born in mind.

* Corresponding author. Tel.: +44-114-2222633; fax: +44-114-2780300.
E-mail address: p.willett@sheffield.ac.uk (P. Willett).

1093-3263/01/$ – see front matter
PII: S1093-3263(01)00110-3

The paper is organised as follows. The next section gives a brief introduction to the main features of FBSS and also reports a simple validation experiment that supports the use of FBSS-based alignments in 3D QSAR analyses. The main experimental results, based on CoMFA and CoMSIA analyses of six literature datasets, are described in the third section, where the statistical models resulting from our automated alignments are compared with those resulting from manual alignments, and the paper concludes with a summary of our major findings and suggestions for further work.

2. Use of FBSS

Many different measures have been described for calculating inter-molecular structural similarity [7]. One approach, which derives from the early work of Carbo et al. [8], involves the use of molecular field descriptors. As further developed by Good et al. [9], the approach involves positioning a molecule at the centre of a 3D grid and calculating a molecular field value (such as the molecular electrostatic potential) at each point of the grid. The similarity between two molecules is obtained by aligning the corresponding grids so as to give the best possible fit of the two sets of field values, and then by calculating a similarity coefficient (such as the Carbo index [8]) that reflects the extent of the agreement between the aligned sets of values. This provides a natural, and very elegant, way of quantifying the extent of the relationship between a pair of 3D structures, and has been adopted by many workers (see, e.g. [10–12]). FBSS is a program that uses field-based similarity measures for similarity searching in chemical structure databases and that uses a genetic algorithm (hereafter a GA) to align two molecules' fields so as to maximise the value of the Carbo index [13–16].

In brief, each chromosome in FBSS's GA encodes the rotations and translations that are to be applied to a database structure to align it with the target structure for the similarity search. If no account is taken of conformational flexibility then just the rigid-body rotations are encoded; alternatively, if the molecules are allowed to flex, then the chromosome additionally encodes the torsional rotations [14], although, the experiments reported here consider only rigid molecules. The fitness function is the value of the similarity coefficient resulting from that particular encoded alignment. Three types of Carbo index are calculated in FBSS, using the fast Gaussian approximation procedures described by Good et al. [9,10] for the calculation of electrostatic and steric similarities and using an analogous procedure for the calculation of hydrophobic similarities (these employing a Gaussian version of the molecular lipophilic potential approach of Gaillard et al. [17]). Alignments may be made based on a single field-type, or on any combination of the three types of field; the experiments reported in this paper involved an equally weighted combination. Here, during the execution of the GA, an alignment of a pair of molecules is used to calculate each of the three individual types of field-based similarity and then the fitness for the chromosome encoding that alignment is the mean of the three resulting similarity values.

The effectiveness of FBSS for database searching has been assessed using sets of active structures from the World Drug Index [1] and BIOSTER [2] databases, these experiments demonstrating that the program is capable of identifying sets of bioactive molecules that are very different from those retrieved by conventional similarity measures based on 2D fragment bit-strings [15,16]. In the work reported here, we have used FBSS alignments as the input to a 3D QSAR procedure, and compared the results with those obtained from conventional manual alignments; alternative approaches to the automated alignment of structures for 3D QSAR are described by Jain et al. [5], Parretti et al. [18] and Lemmen et al. [19], inter alia. The use of FBSS for this purpose seems intuitively reasonable, in that FBSS aligns molecules on the basis of field variables that are at least analogous to those that comprise the independent variables in 3D QSAR methods; the experiments reported in the next section provide a range of evidence to justify the use of FBSS for this application.

3. Experimental details and results

3.1. Datasets

The experiments used six datasets from the published literature, these differing in size, degree of heterogeneity and intended biological target. In most of these datasets, the authors have provided both a training-set and a test-set together with the 3D co-ordinates of the modelled structures, and these data were used in our experiments, hence facilitating the comparison of the FBSS-based and the manually-based alignments.

The datasets are as follows:

- The classic set of steroids with binding affinity data towards corticosteroid binding globulin (CBG) that forms a de facto benchmark for the evaluation of any QSAR method [20]. Specifically, we employed the 31 structures (a 21-compound training-set and a 10-compound test-set) reviewed by Wagener et al. [21], with the exception of molecule-31 (one of the test-set compounds that is known to be poorly predicted in QSAR analyses of this dataset) giving a 21-compound training-set and a 9-compound test-set.
- Sicsic et al. [22] report CoMFA analyses of a set of diverse melatonin receptor antagonists. These authors constructed many different alignments and the resulting CoMFA models before selecting an optimum model based on fitting to compound-12 and on the removal of four of

[1] The World Drug Index database is available from Derwent Information at URL: http://www.derwent.co.uk.

[2] The BIOSTER database is available from Synopsys Scientific Systems at URL: http://www.synopsys.co.uk.

1. Select the most active molecule in an n-member QSAR dataset as a template against which the other n-1 molecules are to be aligned.
2. Use FBSS to align each molecule with the template molecule and calculate the mean similarity when averaged over all of the n-1 alignments.
3. Use the resulting aligned dataset as the input to a CoMFA analysis and calculate q^2, *i.e.*, the cross-validated r^2 value, for the resulting QSAR model.
4. Repeat Steps 2 and 3 several times, with each invocation of FBSS having a different parameter setting for its GA so as to control the extent of the search of alignment space that takes place.
5. Calculate the correlation between the mean similarity values and the values calculated in Steps 2 and 3, respectively.

Fig. 1. Validation study for the generation of QSAR alignments using FBSS.

the original training-set compounds. This reduced dataset (44-compound training-set and 9-compound test-set) was used in the experiments here.

- Winn et al. [23] describe the synthesis, testing and structure-activity analysis of a set of non-peptidic endothelin antagonists. This was divided into a training-set of 49 compounds and a test-set of six compounds in a previous HQSAR study [24].
- Böhm et al. report CoMFA and CoMSIA analyses of a set of 88 benzamidine analogues showing selective activity to three separate receptor systems (thrombin, trypsin and factor Xa) in the blood clotting cascade [25], this involving a 72-compound training-set and a 16-compound test-set. The Factor Xa data was not used here as it was found to give uniformly poor QSAR models, however derived.
- Klebe et al. report CoMFA and CoMSIA analyses of a 76 thermolysin inhibitors [26], this involving a 61-compound training-set and a 15-compound test-set.

3.2. Validation study

Our initial experiments were carried out to ascertain the extent to which FBSS-based similarities were related to alignments that could be used to derive good predictive QSAR models. As noted above when describing FBSS, its GA seeks an alignment of two molecules' fields that maximises the Carbo similarity for those molecules, with the inherent assumption that the largest possible similarity corresponds to the most appropriate alignment for similarity searching. As a starting point for the present QSAR application, we make the analogous assumption that the largest possible similarity between a member of the dataset and the most active member corresponds to the most appropriate alignment for submission to the 3D QSAR procedure. The validity of this assumption was tested by means of the simple procedure summarised in Fig. 1.

The classic steroid dataset was the first to be studied with this validation test. FBSS was used to align each member of the dataset with the most active molecule (deoxycortisol) 75 times, and that alignment selected in each case for which the final calculated similarity was the largest. The mean similarity over the 20 compounds (the full training-set less deoxycortisol) was then calculated and the resulting alignments submitted for the CoMFA analyses. These were performed (both here and elsewhere in the paper) using the SYBYL QSAR and Advanced QSAR modules in the SYBYL molecular modelling package with the default CoMFA parameters.[3] Specifically, a grid spacing of 2.0 Å was used in the preliminary validation experiments, with a spacing of 1.0 Å in all the subsequent experiments; an energy cut-off of 30 kcal/mol was used and energy normalisation effected using the CoMFA_STD procedure (in which each variable in a column is divided by the standard deviation of the whole block; steric or electrostatic), with a MIN_SIGMA value of 2.0. Leave-one-out cross-validation was carried out using SAMPLS [27], with the optimum statistical model being defined as the one with the number of components corresponding to the lowest cross-validated standard error (S_{CV}); similar results to those presented here were obtained when the optimum number of components was derived from the "5% rule" (where additional components are selected as long as the q^2 value increases by at least 5% q^2 units through inclusion of that component [28]).

The extent of the space searched by the GA in FBSS is controlled by the size of the population, the number of generations for which the program is run and the selection pressure that is used. This was exploited here with the intention of obtaining a set of sub-optimal solutions so as to demonstrate the existence of a relationship between similarity and q^2. The entire procedure (fitting each compound 75 times to the template to find the best alignment, calculating the mean of the resultant FBSS similarities, carrying out the CoMFA analysis of the dataset using the set of FBSS-derived alignments, and noting the predictivity of the final

[3] SYBYL and CONCORD are produced by Tripos Inc. at URL: http://www.tripos.com.

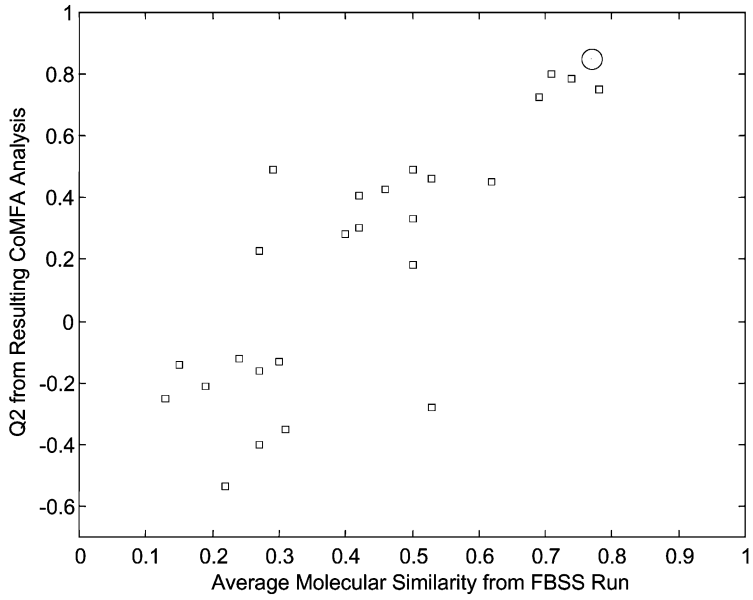


Fig. 2. Plot of the average FBSS molecular similarity against the square of the cross-validated correlation coefficient, q^2, for the steroid dataset. The circle point represents the results obtained with the manual alignments.

statistical model as represented by the q^2 value) was executed with 20 different sets of GA parameters, these covering the ranges 1–5000 for the number of iterations, 10–200 for the population size and 1.1–5.0 for the selection pressure: the resolution of the translations and rotations encoded in the chromosomes were the default values of 1 Å and 1.4° [13]. The 20 pairs of mean FBSS similarities and q^2 values were then plotted to give the scattergram shown in Fig. 2 with, as might be expected, the largest q^2 values corresponding to larger numbers of iterations, larger populations and smaller selection pressures. The correlation between the FBSS similarities and the q^2 values is notable (Spearman coefficient of rank correlation, $\rho = 0.654$, with $P \leq 0.005$), thus, supporting our basic assumption that an alignment with a high FBSS similarity is appropriate for use in a 3D QSAR analysis; indeed, if a correlation is not observed then the dataset is not suitable for alignment using the whole-molecule procedure suggested here. The point marked by a circle in Fig. 2 represents the similarity and q^2 values obtained for the manual alignments provided by Wagener et al. [21], and it will be seen that this point fits well with the correlation obtained from the FBSS alignments. Comparable results are obtained with other datasets in that increasingly close relationships between the similarity and q^2 values are evident as the former is increased by altering the GA parameters; Fig. 3, for example, demonstrates the relationship ($\rho = 0.894$, with $\rho = 0.933$ when discarding any data that has molecular similarity lower than 0.50) for the thrombin dataset.

While there is a clear general trend for high similarities to correspond to high q^2 values, the relationship is not an exact one: for example, there are several pairs of points in Figs. 2 and 3, where a higher similarity corresponds to a lower q^2 value. This is hardly surprising given the global nature of the similarities that are used, with the whole of the molecules that are being compared being involved in the similarity calculation, but does not seem to be a serious problem if the FBSS alignments are to be used in an initial, screening role.

3.3. Evaluation of alignments

The validation study above focused on the effect of GA variations on an entire dataset; we now describe a further, more detailed comparison of the sets of manual and FBSS alignments for the individual molecules in a dataset. As noted above, the FBSS alignments are generated by mapping each molecule, *I*, in a dataset to the most active compound, *A*, in that dataset. By noting the manual alignment of *I* and *A*, it is then possible to calculate the root mean-squared deviation (RMSD) between the sets of heavy atoms for the manual alignment of *I* and the automated alignment of *I*. Each such RMSD value corresponds to one of the FBSS similarity values and these can be plotted to ascertain the relationship, if any, between them. Typical examples of the resulting scattergrams are shown in Figs. 4 and 5, these being for the steroid and endothelin datasets, respectively. In both cases, a well-marked negative correlation is observed ($\rho = 0.786$ and 0.991), thus, showing that the greater the FBSS similarity, the closer the FBSS alignment is to the manual alignment; conversely, those molecules, where it was possible to obtain only a low similarity generally end up in alignments that are quite different from the corresponding manual alignments. This provides further evidence to support the view that high FBSS similarities will be able to provide predictive QSAR models.

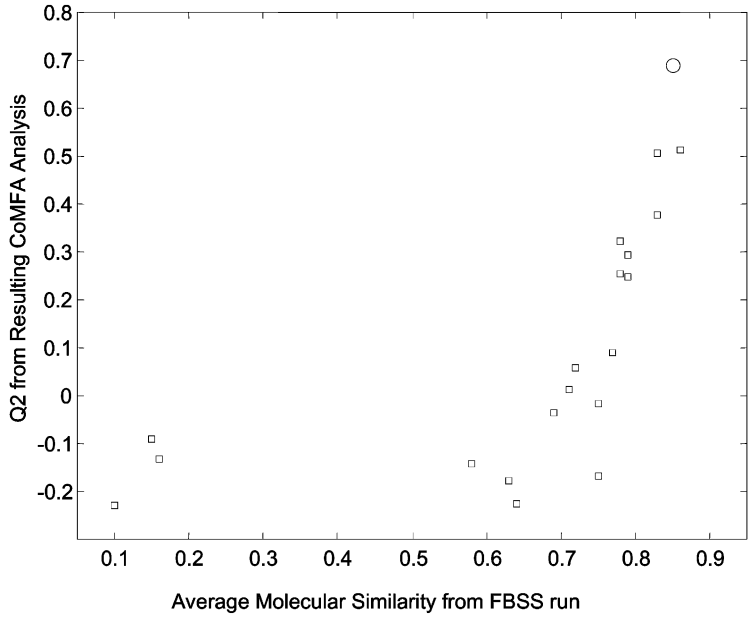


Fig. 3. Plot of the average FBSS molecular similarity against the square of the cross-validated correlation coefficient, q^2, for the thrombin dataset. The circle point represents the results obtained with the manual alignments.

3.4. CoMFA and CoMSIA analyses

Having demonstrated the potential of FBSS for the analysis of our QSAR datasets, the main experiments involved a comparison of the effectiveness of QSAR models derived using manual and FBSS alignments. The FBSS alignments in these runs were generated using 10,000 iterations, a 125-member population and a selection pressure of 1.1, this representing a time of about 30 s to align a pair of structures using a single field-type. The experiments involved both CoMFA and CoMSIA analyses, these being carried out using the standard SYBYL parameter settings: these have been described previously for CoMFA while the CoMSIA analyses used all five of the available field-types (steric,

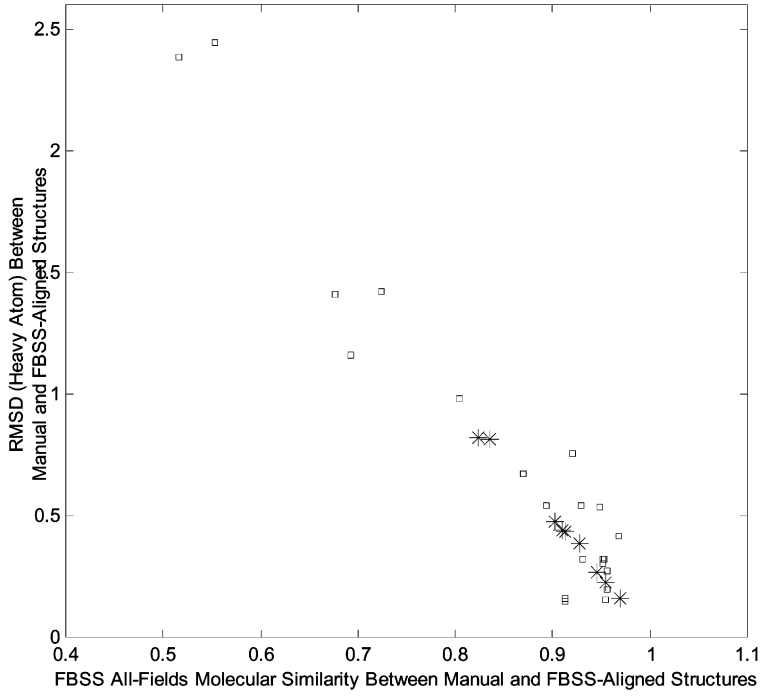


Fig. 4. Plot of the molecular similarity against the RMSD for the steroid dataset (stars represent test-set data, squares represent training-set data).

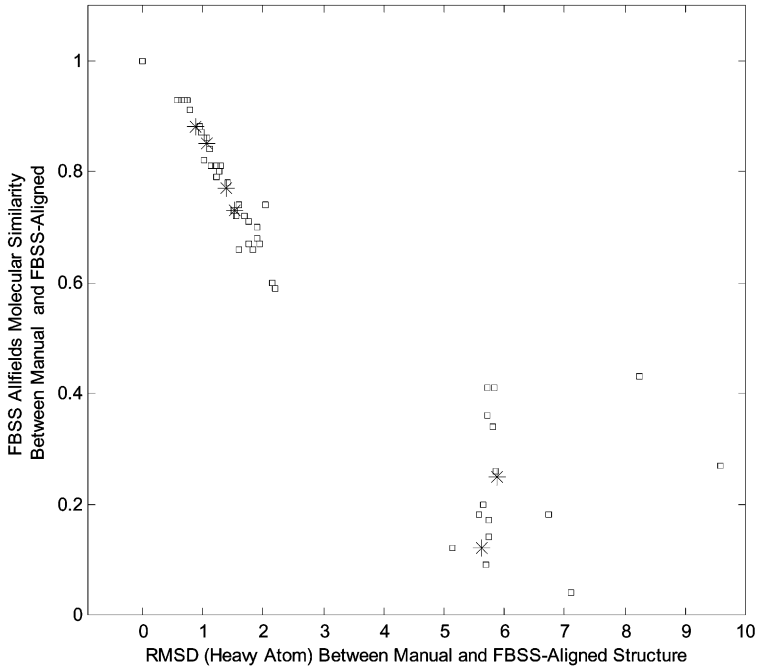


Fig. 5. Plot of the molecular similarity against the RMSD for the endothelin dataset (stars represent test-set data, squares represent training-set data).

electrostatic, hydrophobic, hydrogen bond donor and hydrogen bond acceptor) with the attenuation factor set to the default value of 0.3.

The best CoMFA models (using the default combination of both electrostatic and steric fields) obtained from the analysis of the datasets using the FBSS alignments are listed in Table 1, with the corresponding results from the manual alignments (either those of the original authors or those generated by us for the endothelin dataset) listed in Table 2. An inspection of the manual and automated results suggests that the two approaches yield broadly comparable levels of predictive performance. Thus, if we consider the $pr–r^2$ values, manual is better (to two decimal places) for three of the datasets and automated for three of them; the corresponding figures for the q^2 values are three and two, with one (the melatonin receptor antagonists) being the same.

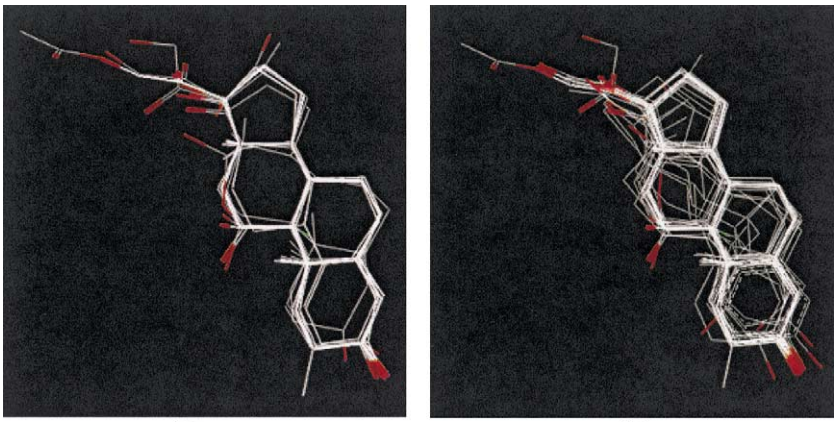
Considering the steroid dataset in more detail, Fig. 6 shows the manual and automated alignments and Figs. 7

Table 1
CoMFA models obtained using FBSS alignments

Dataset	q^2	S_{CV}	N	r^2	S	F	$pr–r^2$
Steroids	0.866	0.466	3	0.982	0.170	315	0.917
Melatonin receptor antagonists	0.717	0.704	5	0.982	0.179	407	0.547
Endothelin antagonists	0.456	2.534	3	0.894	1.121	126	0.852
Thrombin inhibitors	0.514	0.745	5	0.927	0.290	167	0.451
Trypsin inhibitors	0.479	0.659	5	0.936	0.232	192	0.663
Thermolysin inhibitors	0.374	1.694	4	0.874	0.758	98	0.436

Table 2
CoMFA models obtained using manual alignments

Dataset	q^2	S_{CV}	N	r^2	S	F	$pr–r^2$
Steroids	0.851	0.477	2	0.928	0.331	117	0.856
Melatonin receptor antagonists	0.723	0.706	6	0.980	0.191	297	0.790
Endothelin antagonists	0.394	2.618	1	0.553	2.250	58	0.133
Thrombin inhibitors	0.689	0.592	4	0.882	0.365	125	0.478
Trypsin inhibitors	0.619	0.569	6	0.942	0.221	176	0.696
Thermolysin inhibitors	0.639	1.309	6	0.919	0.622	102	0.384



(a) (b)

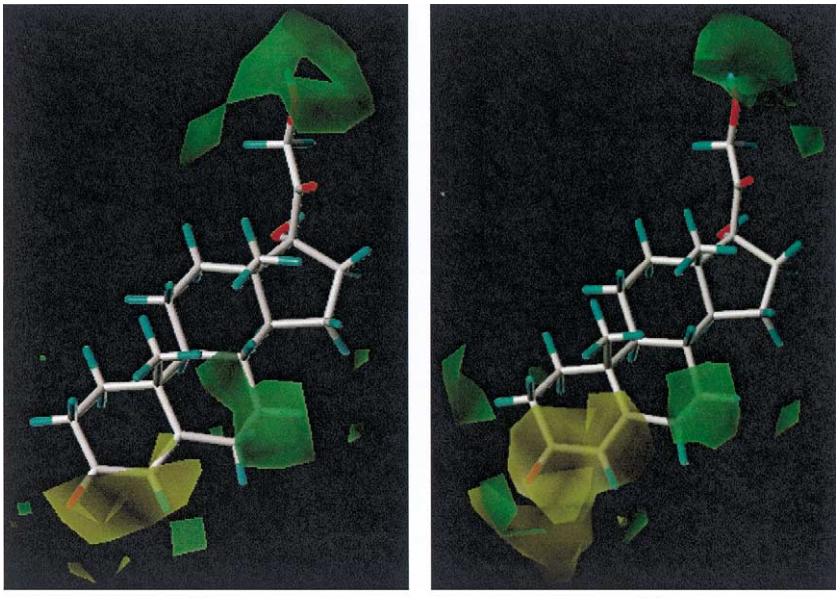
Fig. 6. Alignment of steroid dataset to most active compound using (a) manual and (b) automated methods.

and 8 the corresponding CoMFA maps. It will be seen from Fig. 6 that while the automated alignments are not identical to the manual ones, they are very similar; it is thus, hardly surprising that the steric and electrostatic maps resulting from the alignments (in Figs. 7 and 8, respectively) are very similar in overall shape and would provide broadly comparable levels of information to the drug designer in a realistic CoMFA application.

The sets of automated and manual alignments were then used as the input to CoMSIA analyses. The results obtained are listed in Tables 3 and 4 and are analogous to those obtained in the CoMFA studies, in that the two sets of alignments produce broadly comparable sets of predictive models. Specifically, manual is better (to two decimal places) for four of the datasets and automated for two of them, considering both the $pr–r^2$ values and the q^2 values. Here again, the manual and automated alignments and the resulting maps are very similar. Thus, Fig. 9 shows the alignments for the thrombin dataset, and Fig. 10 the resulting CoMSIA steric maps.

Finally, Fig. 11 shows the alignments for the thermolysin dataset. The FBSS alignments are obviously much less consistent than the manual ones here, and this is reflected in the manual q^2 values being noticeably superior to the automated ones; however, the converse applies if the $pr–r^2$ values are considered.

The experiments thus far have used the published sets of modelled co-ordinates for our six datasets. However, the



(a) (b)

Fig. 7. Steric CoMFA maps for the steroid dataset using (a) manual and (b) automated alignments.

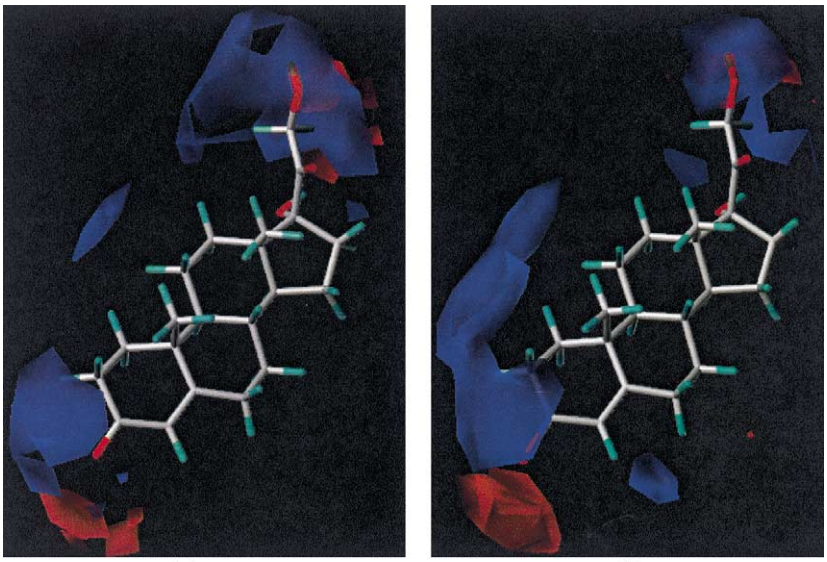


Fig. 8. Electrostatic CoMFA maps for the steroid dataset using (a) manual and (b) automated alignments.

intended use is, as noted previously, as a precursor to a more detailed study, in which case, it is unlikely that fully modelled structures would be available for the generation of the QSAR model. It could thus be argued that the results in Tables 1 and 3 over-estimate the ability of FBSS to generate good predictive QSARs as these tables are based on carefully modelled structures that would not be available in practice. We have hence carried out additional sets of experiments in which the 3D structures for the compounds in our six datasets were generated using CONCORD followed by MOPAC optimisation, thus mirroring the level of structural information that might be expected at the start of an analysis. The results of using these non-modelled structures are summarised in Tables 5 and 6, which can be compared with the results in Tables 1 and 3, respectively. It will be seen that the models developed from alignments using these simpler structures are sometimes poorer than the models developed from the fully modelled structures (e.g. the q^2 values), but they still exhibit significant predictive power (especially in the case of the CoMSIA analyses), thus, validating their use for the screening-like application proposed here.

Table 3
CoMSIA models obtained using FBSS alignments

Dataset	q^2	S_{CV}	N	r^2	S	F	$pr–r^2$
Steroids	0.844	0.553	6	0.995	0.094	513	0.763
Melatonin receptor antagonists	0.731	0.687	5	0.945	0.311	130	0.666
Endothelin antagonists	0.561	2.301	4	0.860	1.298	68	0.926
Thrombin inhibitors	0.561	0.714	6	0.949	0.242	203	0.422
Trypsin inhibitors	0.557	0.599	3	0.833	0.368	123	0.699
Thermolysin inhibitors	0.386	1.692	5	0.908	0.655	109	0.576

Table 4
CoMSIA models obtained using manual alignments

Dataset	q^2	S_{CV}	N	r^2	S	F	$pr–r^2$
Steroids	0.769	0.579	1	0.834	0.490	96	0.898
Melatonin receptor antagonists	0.781	0.620	5	0.920	0.374	87	0.635
Endothelin antagonists	0.424	2.608	3	0.798	1.546	59	0.353
Thrombin inhibitors	0.756	0.532	6	0.950	0.240	208	0.453
Trypsin inhibitors	0.726	0.482	6	0.934	0.234	192	0.929
Thermolysin inhibitors	0.643	1.290	5	0.892	0.709	91	0.377

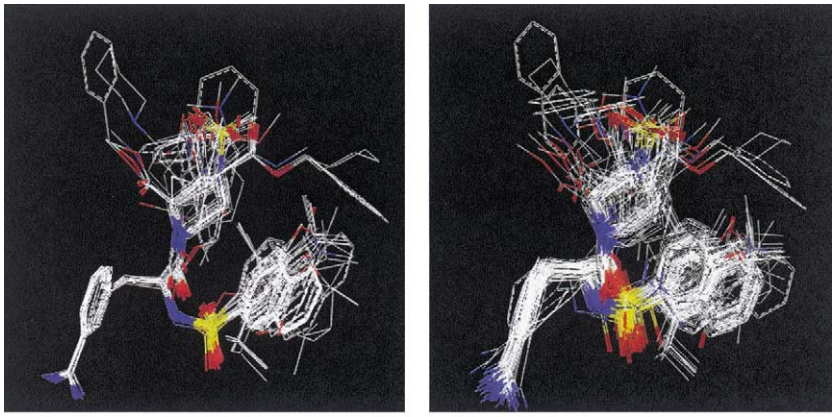


Fig. 9. Alignment of thrombin dataset to most active compound using (a) manual and (b) automated methods.

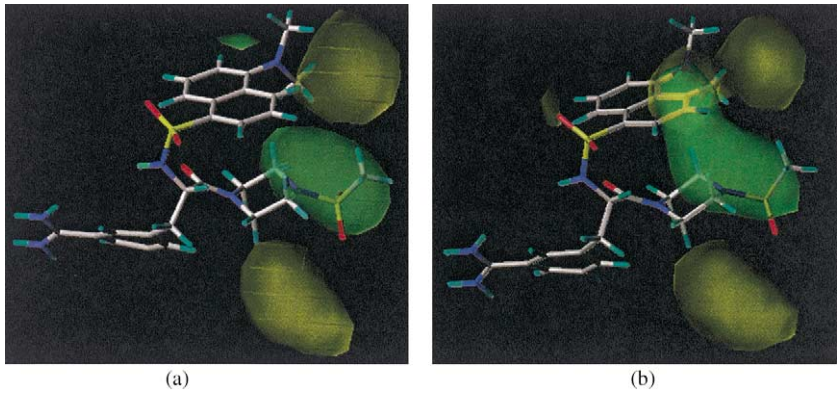


Fig. 10. Steric CoMSIA maps for the thrombin dataset using (a) manual and (b) automated alignments.

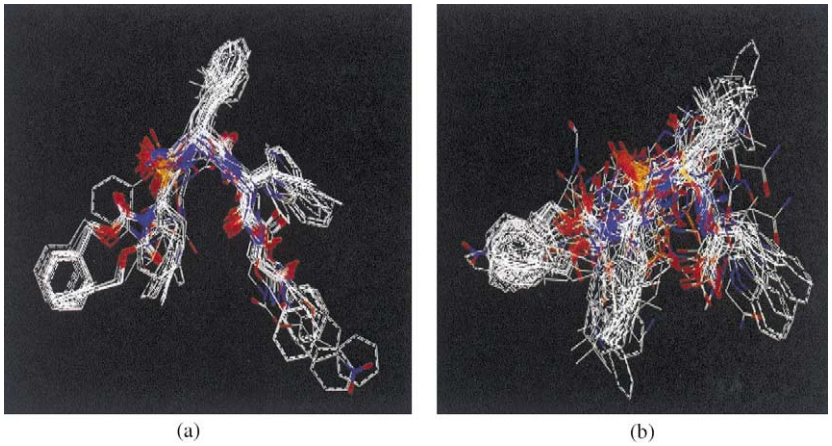


Fig. 11. Alignment of thermolysin dataset to most active compound using (a) manual and (b) automated methods.

Table 5
CoMFA models obtained using FBSS alignments and non-modelled 3D structures

Dataset	q^2	S_{CV}	N	r^2	S	F	$pr–r^2$
Steroids	0.775	0.587	2	0.939	0.305	139	0.640
Melatonin receptor antagonists	0.601	0.836	5	0.962	0.257	194	0.821
Endothelin antagonists	0.294	2.825	1	0.603	2.119	71	0.656
Thrombin inhibitors	0.401	0.834	6	0.944	0.256	182	0.428
Trypsin inhibitors	0.494	0.650	5	0.902	0.287	121	0.748
Thermolysin inhibitors	0.233	1.842	2	0.604	1.323	44	0.423

Table 6
CoMSIA models obtained using FBSS alignments and non-modelled 3D structures

Dataset	q^2	S_{CV}	N	r^2	S	F	$pr–r^2$
Steroids	0.672	0.708	2	0.874	0.439	62	0.839
Melatonin receptor antagonists	0.579	0.847	4	0.859	0.490	60	0.827
Endothelin antagonists	0.394	2.647	2	0.822	1.435	106	0.706
Thrombin inhibitors	0.474	0.782	6	0.931	0.283	146	0.425
Trypsin inhibitors	0.524	0.635	6	0.911	0.274	111	0.730
Thermolysin inhibitors	0.281	1.848	6	0.970	0.378	290	0.270

4. Discussion and conclusions

In this paper, we have described the use of an automated procedure for generating the alignments required by many 3D QSAR methods. Experiments with several CoMFA and CoMSIA datasets demonstrate that our procedure can be used to support conventional, manual approaches to the generation of 3D QSAR models.

The idea of suggesting alignments by automatic means is not a novel one, with the SYBYL field-fit routine having been first described a decade ago [29]. The use of the routine has been reported by several workers (see, e.g. [30–32]), but is quite complex in operation [33], involving the inclusion of weighted field-fit energy penalties as additional parameters in the Tripos force field. More recently, Paretti et al. have described a procedure that is analogous to that reported here [18], but that uses Monte Carlo and simplex procedures for the generation of the alignments, rather than a GA, and PLS analysis of $N \times N$ similarity matrices, rather than CoMFA and CoMSIA. Importantly, their method encompasses full conformational flexibility; however, it has only been applied to a single QSAR dataset for which r^2 and q^2 values are reported.

There are several ways in which our work can be extended. First, and most obviously, we have taken no account of conformational flexibility in the work reported here. FBSS does allow for flexible fitting [14], but while this generally results in better alignments (in the sense that higher Carbo similarity values are obtained) these are normally associated with very highly strained structures. Such conformations can be filtered to some extent by inclusion of an appropriate energy calculation in the GA's fitness function, but this is extremely time-consuming and we have also found that its inclusion has little effect on the quality of the resulting CoMFA models; we are currently considering a more sophisticated approach based on the use of torsion libraries derived from the Cambridge structural database as a component of the GA's fitness function. An alternative approach suggested by a referee might be to pre-sample conformational space for each molecule and then to swap the conformers in and out during the generation of the model, an approach that is both simple to implement and easy to parallelise. Secondly, we should note that the datasets used here are mostly quite simple in nature, thus making it difficult to test one of the main potential benefits of the suggested approach, viz. the possibility of suggesting non-obvious alignments for consideration during a modelling problem. We have, however, recently completed a QSAR analysis of 124 structurally diverse antibacterial phenolics, where the FBSS alignments were noticeably different from manual fitting, whilst demonstrating superior predictive ability [34]; further such datasets need to be analysed to determine the generality of this behaviour. Finally, while we have focused here on the use of FBSS, several other programs have been designed to align pairs of 3D molecules, and we are currently evaluating the effectiveness of several such programs [9,19,35] for the generation of 3D QSAR models.

Acknowledgements

We thank the following: the Biotechnology and Biological Sciences Research Council and Zeneca Agrochemicals for funding; Tripos Inc. for software support; Val Gillet and Russell Viner for helpful discussions on this work; Dominic Ryan for suggesting the RMSD analysis; and the referees for comments on an earlier draft of this paper. The Krebs Institute for Biomolecular Research is a designated

Biomolecular Sciences Centre of the Biotechnology and Biological Sciences Research Council.

References

[1] G. Grecco, E. Novellino, Y.C. Martin, 3D QSAR methods. In: Y.C. Martin, P. Willett (Eds.), Designing Bioactive Molecules: Three-dimensional Techniques and Applications, American Chemical Society, Washington, DC, 1998, pp. 219–252.
[2] H. Kubinyi, G. Folkers, Y.C. Martin (Eds.), 3D QSAR in Drug Design, Kluwer/ESCOM, Leiden, 1998.
[3] R.D. Cramer, D.E. Patterson, J.D. Bunce, Comparative molecular field analysis (CoMFA). Effect of shape on binding of steroids to carrier proteins, J. Am. Chem. Soc. 110 (1988) 5959–5967.
[4] G. Klebe, Comparative molecular similarity indices analysis: CoMSIA, Perspect. Drug Disc. Design 12–14 (1998) 87–104.
[5] A.N. Jain, T.G. Dietterich, R.H. Lathrop, D. Chapman, R.E. Critchlow, B.E. Bauer, T.A. Webster, T. Lozanoperez, COMPASS — a shape-based machine learning tool for drug design, J. Comput.-Aid. Mol. Design 8 (1994) 635–652.
[6] A.M. Doweyko, The hypothetical active-site lattice — in vitro and in vivo explorations using a three-dimensional QSAR technique, J. Math. Chem. 7 (1991) 273–285.
[7] P.M. Dean (Ed.), Molecular Similarity in Drug Design, Chapman & Hall, Glasgow, 1995.
[8] R. Carbó, L. Leyda, M. Arnau, How similar is a molecule to another? An electron density measure of similarity between two molecular structures, Int. J. Quant. Chem. 17 (1980) 1185–1189.
[9] A.C. Good, E.E. Hodgkin, W.G. Richards, The utilisation of Gaussian functions for the rapid evaluation of molecular similarity, J. Chem. Inf. Comput. Sci. 32 (1992) 188–191.
[10] A.C. Good, W.G. Richards, Rapid evaluation of shape similarity using Gaussian functions, J. Chem. Inf. Comput. Sci. 33 (1993) 112–116.
[11] J.D. Petke, Cumulative and discrete similarity analysis of electrostatic potentials and fields, J. Comput. Chem. 14 (1993) 928–933.
[12] J. Mestres, D.C. Rohrer, G.M. Maggiora, MIMIC: a molecular-field matching program. Exploiting applicability of molecular similarity approaches, J. Comput. Chem. 18 (1997) 934–954.
[13] D.J. Wild, P. Willett, Similarity searching in files of three-dimensional chemical structures: alignment of molecular electrostatic potentials with a genetic algorithm, J. Chem. Inf. Comput. Sci. 36 (1996) 159–167.
[14] D.A. Thorner, D.J. Wild, P. Willett, P.M. Wright, Similarity searching in files of three-dimensional chemical structures: flexible field-based searching of molecular electrostatic potentials, J. Chem. Inf. Comput. Sci. 36 (1996) 900–908.
[15] S.K. Drayton, K. Edwards, N.E. Jewell, D.B. Turner, D.J. Wild, P. Willett, P.M. Wright, K. Simmons, Similarity searching in files of three-dimensional chemical structures: identification of bioactive molecules. Internet J. Chem. at URL: http://www.ijc.com/articles/1998v1/37/.
[16] A. Schuffenhauer, V.J. Gillet, P. Willett, Similarity searching in files of three-dimensional chemical structures: analysis of the BIOSTER database using two-dimensional fingerprints and molecular field descriptors, J. Chem. Inf. Comput. Sci. 40 (2000) 295–307.
[17] P. Gaillard, P. Carrupt, B. Testa, A. Boudon, Molecular lipophilicity potential, a tool in 3D QSAR: method and applications, J. Comput.-Aid. Mol. Design 8 (1994) 83–96.
[18] M.F. Parretti, R.T. Kroemer, J.H. Rothman, W.G. Richards, Alignment of molecules by the Monte Carlo optimization of molecular similarity indices, J. Comput. Chem. 18 (1997) 1344–1353.
[19] C. Lemmen, T. Lengauer, G. Klebe, FLEXS: a method for fast flexible ligand superposition, J. Med. Chem. 41 (1998) 4502–4520.
[20] E.A. Coats, The CoMFA steroids as a benchmark dataset for development of 3D QSAR methods, Perspect. Drug Discov. Design 12–14 (1998) 199–213.
[21] M. Wagener, J. Sadowski, J. Gasteiger, Autocorrelation of molecular-surface properties for modeling corticosteroid-binding globulin and cytosolic AH receptor activity by neural networks, J. Am. Chem. Soc. 117 (1995) 7769–7775.
[22] S. Sicsic, I. Serraz, J. Andrieux, B. Bremont, M. Matheallainmat, A. Poncet, S. Shen, M. Langlois, Three-dimensional quantitative structure-activity relationship of melatonin receptor ligands: a comparative molecular field analysis study, J. Med. Chem. 40 (1997) 739–748.
[23] M. Winn, T.W. von Geldern, T.J. Opgenorth, H.-S. Jae, A.S. Tasker, S.A. Boyd, J.A. Kester, R.A. Manei, R. Bal, B.K. Sorensen, J.R. Wu-Wong, W.J. Chiou, D.B. Dixon, E.I. Novosad, L. Hernandez, K.C. Marsh, 2,4-Diarylpyrrolidine-3-carboxylic acids potent ETA selective endothelin receptor antagonists. Part 1. Discovery of A-127722, J. Med. Chem. 39 (1996) 1039–1048.
[24] M. Seel, D.B. Turner, P. Willett, Effect of parameter variations on the effectiveness of HQSAR analyses, Quant. Struct.-Activ. Relat. 18 (1999) 245–252.
[25] M. Böhm, J. Stürzebecher, G. Klebe, Three dimensional quantitative structure-activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa, J. Med. Chem. 42 (1999) 458–477.
[26] G. Klebe, U. Abraham, T. Mietzner, Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity, J. Med. Chem. 37 (1994) 4130–4146.
[27] B.L. Bush, R.B. Nachbar, Sample-distance partial least-squares — PLS optimized for many variables, with application to CoMFA, J. Comput.-Aid. Mol. Design 7 (1993) 587–619.
[28] H. Kubinyi, U. Abraham, Practical problems in PLS analyses. In: H. Kubinyi (Ed.), 3D QSAR in Drug Design, ESCOM, Leiden, 1993, pp. 717–728.
[29] M. Clark, R.D. Cramer, D.M. Jones, D.E. Patterson, P.E. Simeroth, Comparative molecular field analysis (CoMFA). Part 2. Towards its use with 3D-structural databases, Tetrahed. Comput. Methodol. 3 (1990) 47–59.
[30] T.G. Gantchev, H. Ali, J.E. Vanlier, Quantitative structure-activity relationships comparative molecular-field analysis (QSAR/CoMFA) for receptor-binding properties of halogenated estradiol derivatives, J. Med. Chem. 37 (1994) 4164–4176.
[31] R.E. Wilcox, T. Tseng, M.Y.K. Brusniak, B. Ginsburg, R.S. Pearlman, M. Teeter, C. DuRand, S. Starr, K.A. Neva, CoMFA-based prediction of agonist affinities at recombinant D1 versus D2 dopamine receptors, J. Med. Chem. 41 (1998) 4385–4399.
[32] S.S. Kulkarni, V.M. Kulkarni, Three-dimensional quantitative structure-activity relationship of interleukin 1-beta converting enzyme inhibitors: a comparative molecular field analysis study, J. Med. Chem. 42 (1999) 373–380.
[33] S. Dove, A. Buschauer, Improved alignment by weighted field-fit in CoMFA of histamine H-2 receptor agonistic imidazolylpropylguanidines, Quant. Struct.-Activ. Relat. 18 (1999) 329–341.
[34] S. Shapiro, N.E. Jewell, D.B. Turner, Bioactivities of antibacterial phenolics: a comparison of HASL, CoMFA, CoMSIA and EVA, in: Proceedings of the Third European Conference on Computational Chemistry, 4–9 September 2000, Budapest, Hungary.
[35] S.K. Kearsley, J. Smith, SEAL: an alternative method for the alignment of molecular structures — maximising electrostatic and steric overlay, Tetrahedron Comput. Methodol. 6 (1990) 615–633.