J Mol Model (2008) 14:59–67
DOI 10.1007/s00894-007-0252-1

ORIGINAL PAPER

Effect of steric molecular field settings on CoMFA predictivity

Ruchi R. Mittal · Ross A. McKinnon · Michael J. Sorich

Received: 28 May 2007 / Accepted: 25 October 2007 / Published online: 24 November 2007

Abstract Steric molecular field can be represented in a number of ways in comparative molecular field analysis (CoMFA). This study aimed to investigate whether the choice of steric molecular field settings significantly influences the predictive performance of CoMFA and, if so, which is the best. The three-dimensional quantitative structure activity relationship (3D-QSAR) models based on Lennard–Jones, indicator, parabolic and Gaussian steric fields were compared using 28 datasets taken from the literature. The analysis of the predictive ability of these models (cross validated R^2) indicates that steric fields in which the value drops off quickly with distance (i.e. Lennard–Jones and indicator fields) tend to perform better than the Gaussian version, which has a slower and smoother decrease. Furthermore, depending on the steric field type used, the field sampling density (i.e. grid spacing) has a variable influence on the predictive ability of the models generated.

Keywords COMFA · CoMSIA · 3D-QSAR · Molecular shape · Steric molecular field

Introduction

Lead optimisation, in which a chemical showing promise is modified to improve its usefulness as a drug, is a vital component of the drug discovery process. Quantitative structure–activity relationship (QSAR) methods can facilitate this process by elucidating the chemical characteristics that are favourable and unfavourable through statistical analysis of a series of chemical analogues.

Three-dimensional (3D)-QSAR techniques such as comparative molecular field analysis (CoMFA) [1] and comparative molecular similarity index analysis (CoMSIA) [2] are popular [3] due to their ability to generate both highly predictive and easily interpretable models. The process of undertaking a 3D-QSAR can generally be broken down into three parts:

1. Molecular alignment: The alignment or superimposition of the molecules involves deciding on a common pattern of receptor binding so that all of the molecules can be placed in this pattern [1].
2. Calculation and sampling of molecular fields: Force fields are used in CoMFA to describe the interactions that typically occur between a ligand and the target macromolecule. The forces primarily responsible for ligand–protein interactions include the steric (also known as dispersion or van der Waals), electrostatic, hydrogen-bonding and hydrophobic molecular fields. The aligned molecules are placed within a 3D grid or lattice of points. This grid is used as a means of sampling the interaction between the individual molecules (of the aligned set) and various probes placed at each of the lattice points of the grid [1, 4].
3. Analysis of molecular fields: Typically, partial least squares regression (PLSR) is used to determine the linear function that maps molecular field values into the binding affinity of the ligand [1].

At each stage there are a multitude of technical options, the choice of which can significantly influence the utility of the models generated. Furthermore, there is little good

Electronic supplementary material The online version of this article (doi:10.1007/s00894-007-0252-1) contains supplementary material, which is available to authorized users.

R. R. Mittal · R. A. McKinnon · M. J. Sorich (✉)
Sansom Institute, School of Pharmacy and Medical Sciences,
University of South Australia,
Adelaide SA 5000, Australia
e-mail: michael.sorich@unisa.edu.au

evidence to suggest which options are best, thereby leaving these complicated choices to the end user.

This study aimed to evaluate the importance of choices available for calculating steric molecular fields on the prediction accuracy of the models generated. In particular:

1. The choice of steric field type (Fig. 1).
2. The effect of attenuation factor value for the steric Gaussian field.
3. The effect of lattice spacing for sampling of steric fields.
4. The effect of column filtering prior to PLSR.

Lennard–Jones potential

The standard CoMFA method uses the Lennard–Jones (LJ) 6–12 potential to calculate the steric interaction between a probe carbon atom (placed at different grid points) and a molecule [5]. This is calculated by summing the interaction between the probe atom and each atom in the molecule.

$$LJ(x,y,z) = \sum_{atoms} \sqrt{E_{probe} \times E_{atom}} \left\{ \frac{1.0}{\left(\frac{d}{R_{probe}+R_{atom}}\right)^{12}} - \frac{2.0}{\left(\frac{d}{R_{probe}+R_{atom}}\right)^{6}} \right\}$$

Where E_{atom} is the Van der Waals constant of the molecule atom (kcal mol^{-1}), E_{probe} is the van der Waals constant of the probe atom (kcal mol^{-1}), d the distance between probe atom and molecule atom (Å), R_{probe} the van der Waals radius of the probe atom (Å), and R_{atom} the van der Waals radius of the molecule atom (Å).

The LJ potential is sharp near the van der Waals forces of the molecule (Fig. 1). Generally the resultant energy values for these attractive and repulsive forces range from −10 kcal mol^{-1} to infinity at the atomic centres. To avoid infinite values some arbitrary cut-off values are assigned to lattice points where the calculated value is higher than the specified threshold. In SYBYL [6], a default cut-off of 30 kcal mol^{-1} is used [7, 8].

Indicator potential

Indicator fields are a variation of the LJ 6–12 potential, in which the value of the field can take one of only two possible values. For the steric indicator field, the value assigned to a grid point is 0 if the steric potential falls below the defined cut-off (default 30 kcal mol^{-1}). If the steric potential is at or above the defined cut-off, the grid point value is set equal to the cut-off. 3D-QSAR models based on the indicator potential can be calculated very fast, even for a small grid spacing, as most of the points have zero variance and are dropped from the PLSR analysis [5, 9].

Parabolic potential

The parabolic field is a squared transformation of the standard LJ 6–12 potential with the original sign retained [5, 10].

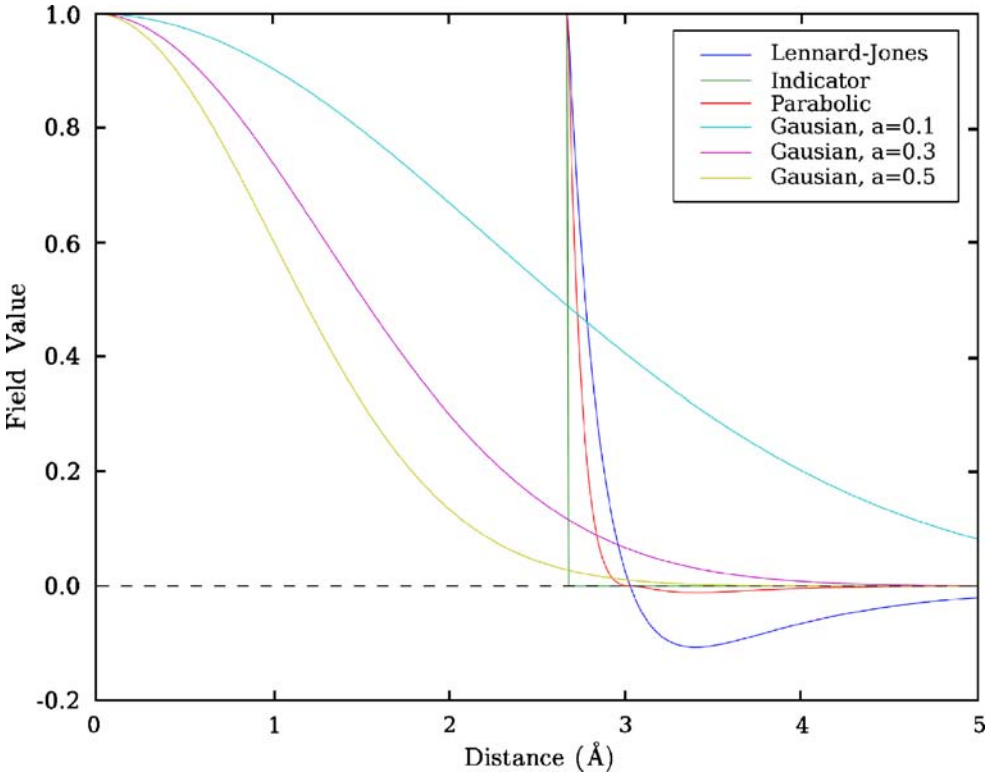

Fig. 1 Shape of various steric functions for a simulated example involving the normalised interaction of two carbon atoms [5]

Gaussian potential

CoMFA and CoMSIA are basically the same process, differing only in their calculation of molecular fields. Unlike CoMFA, CoMSIA uses Gaussian functions to calculate molecular field potentials.

$$G_{steric}(x,y,z) = \sum_{atoms} R_{probe} \times R_{atom} \times e^{-\alpha d^2}$$

Where G_{steric} is the similarity index at a particular grid point, summed over all atoms of the molecule; R_{probe} the van der Waals radius of the probe atom (Å), R_{atom} the van der Waals radius of the molecule atom (Å), α the attenuation factor (default value used is 0.3); and d the distance between probe atom and molecule atom (Å).

Comparatively, the Gaussian steric potential can be calculated both inside and outside the molecule. Additionally, this functional form is relatively smoother—hence not requiring arbitrary cut-off values (see Fig. 1)—and results in contiguous contours. The only adjustable parameter in the Gaussian function is the attenuation factor (α) [2, 5]. This controls the steepness of the function and a default value of 0.3 is commonly used. A larger value for α results in a steeper function, and thus the field contains more local information on the molecular property (i.e. global molecular features become less important) [2, 5].

Data and methods

In total, 28 pre-aligned molecular datasets were sourced from the literature. The 3D aligned molecular structures and the experimentally derived activity values of these datasets were either requested from the respective author or were extracted from the publication's supporting information. Table 1 lists all the QSAR datasets used and their respective references.

All the modelling and analyses were carried out on an SGI Octane system using SYBYL 7.1 [6]. The 3D-QSAR models based on LJ, indicator, parabolic and Gaussian steric fields were generated for each dataset using partial least squares regression (PLSR) using QSAR and the

Table 1 Datasets used in the comparisons of steric molecular field settings

Dataset	Description	*N*	Reference
ACE	Inhibitors of angiotensin converting enzyme	114	[15]
ACHE	Inhibitors of acetyl-cholinesterase	111	[15]
BZR	Inhibitors of benzodiazepine receptor	163	[15]
COX2	Inhibitors of cyclooxygenase-2	322	[15]
DHFR	Inhibitors of rat dihydrofolate reductase	397	[15]
GPB	Inhibitors of glycogen phosphorylase b	66	[15]
THERM	Inhibitors of thermolysin	76	[15]
THR	Inhibitors of thrombin	88	[15]
COMT	Inhibitors of catechol-O-methyltransferase	92	[31]
HIVPR	Inhibitors of human immunodeficiency virus (HIV-1) protease	113	[32]
MX	Mutagenicity of mutagen X analogues	29	[14]
DR	Antagonists of dopamine receptor	38	[33, 34]
GHS	Growth hormone secretagogue mimics	31	[35]
PLA2	Inhibitors of Phospholipase A2	11	[35]
YOPH	Inhibitors of Yersinia protein tyrosine phosphatase	39	[36]
STEROIDS	Binding of steroids to carrier proteins	21	[1, 6]
PTC	Phase-transfer asymmetric catalysts	40	[37]
RYR	Binding of ryanoids to the ryanodine receptor	18	[6, 38]
HIVRT	Inhibition of HIV-1 reverse transcriptase	101	[39, 40]
AI	Steroid aromatase inhibitors	78	[6,41]
ARB	Non-peptide angiotensin II receptor antagonists	28	[42]
MT	MT1 and MT2 melatonin receptor ligands	56	[43]
KOA	Kappa opioid antagonists	39	[44]
TCHK	Inhibition of *Trypanosoma cruzi* hexokinase	42	[45]
ERB	Estrogen receptor binders	123	[46]
PDE	PDE-IV inhibitors	29	[47]
CBRA	Cannabinoid CB1 receptor agonists	32	[48]
ATA	Anti-tuberculosis agents	72	[49]

Advanced COMFA modules in SYBYL. The following CoMFA settings were used:

- TRANSFORM (CoMFA field class): NONE (LJ potential), INDICATOR (Indicator potential), SQUARED (Parabolic potential).
- FIELD_TYPE: STERIC_ONLY
- VOLUME_AVG_TYPE: NONE
- STERIC_ENERGY_MAX (cut-off): 30 kcal mol^{-1}
- SWITCH_FCN (the type of transition in the cut-off region): "YES" for Tripos standard field; "NO" for indicator and parabolic fields.
- REGION_SETTINGS: Fields were sampled at a density (i.e. grid spacing) of 1 Å, 1.5 Å & 2 Å with an extension of 4 Å in all directions from the aligned molecules. The probe atom has van der Waals properties of sp^3 carbon atom (c.3).

The attenuation factor, the only variable setting in the Gaussian field, was varied from 0.1 to 0.5. The region settings were the same as those used in the steric CoMFA fields.

The leave-one-out cross-validated R^2 $\left(R^2_{cv}\right)$ was used to measure the predictive ability of each model [11]. In the leave-one-out method, one compound is removed from the dataset and its activity is predicted using the model derived from the compounds remaining in the dataset [12].

$$R^2_{CV} = \frac{\sum \left(Y_{pred} - Y_{mean}\right)^2}{\sum \left(Y_{obs} - Y_{mean}\right)^2}$$

Two main parameters in the calculation of a PLS regression model are the number of components (also known as latent variables) and the column filtering (minimum standard deviation). Components are the mutually orthogonal linear combination of descriptive independent column data. Column filtering is used to remove grid points that have a small standard deviation prior to PLS regression analysis [6].

The maximum number of components was set at six and a column filtering of 2.0 kcal mol^{-1} was applied to all the PLS analyses to elevate the signal-to-noise ratio by dropping lattice points with energy variation less than this threshold. The above models were also assessed using column filtering of 1.0 and 3.0 kcal mol^{-1}. All of the above procedures (QSAR table generation, molecular field calculations, and PLSR) were automated with the use of in-house SYBYL programming language (SPL) scripts.

The R^2_{cv} values for each dataset and steric molecular field setting were statistically analysed using SPSS 12.0 and R [13] statistical software. Paired Wilcoxon signed ranks test were carried out to determine if a statistically significant difference occurred when a variety of steric potential, grid spacing, attenuation factor and column filtering were used. Differences with $P<0.05$ were considered statistically significant. In order to assess the importance of column filtering settings, the median R^2_{cv} values (across the datasets) for the various steric field settings were compared using paired Wilcoxon signed ranks tests.

Hierarchical clustering using the Euclidean distance function was undertaken using the R language. Prior to clustering, the 24 different steric molecular field settings (each a unique combination of a specific steric potential and the grid spacing) were ranked for each dataset based on the R^2_{cv}. Thus, datasets that are most similar (i.e. clustered together) tend to have the most predictive models generated from the same steric molecular field settings (and least predictive models generated from the same steric molecular field settings).

Results

There are 24 unique steric molecular field settings—a combination of the steric potential functional form (LJ, indicator, parabolic and five Gaussian fields differing in the attenuation factor) and the grid spacing (1 Å, 1.5 Å and 2 Å). For each steric molecular field setting a model was generated and assessed for each dataset. The median R^2_{cv} across the 28 datasets, assessed for the 24 different steric molecular field settings, are displayed in Fig. 2. Additionally, Supplementary Fig. 1 shows the box plots representing the distribution of R^2_{cv} obtained with the different models.

When all 24 different settings are ranked based on their overall ability to generate the most predictive models, the indicator steric field at a grid spacing of 1 Å was the most predictive. The improvement of the indicator field with 1 Å spacing over the other steric field settings was statistically significant, other than for the LJ and parabolic fields with 1 Å or 1.5 Å spacing.

CoMFA vs CoMSIA steric fields

The different combinations of steric potential and grid spacing were compared statistically in a pairwise manner and the full set of results can be found in Supplementary Table 1. In general, the CoMFA steric field potentials (LJ, indicator and parabolic) were consistently better than the CoMSIA Gaussian steric field potentials. For example, at a grid spacing of 1 Å, the LJ field was more predictive than the standard Gaussian steric field (attenuation factor of 0.3) in a highly statistically significant fashion (95% CI of R^2_{cv} difference: 0.07–0.13, $P<10^{-5}$). Similar results were found at other grid spacing values and for comparisons between the other CoMFA steric potentials (indicator and parabolic) and the standard Gaussian steric potential.

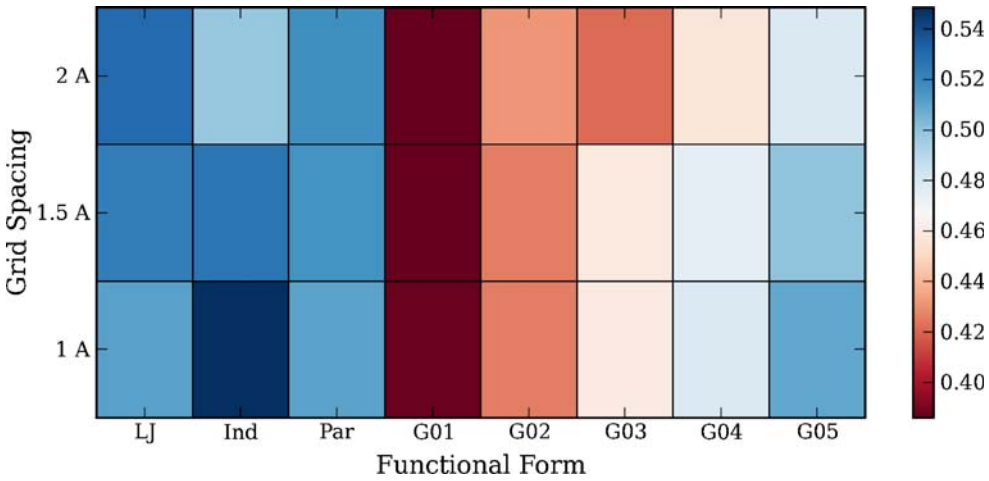

Fig. 2 Colour map showing the effect of grid spacing and steric field functional form on predictive performance of comparative molecular field analysis (CoMFA). The median R^2_{cv} of 28 datasets is represented by a colour ranging from *red* (worst predictive ability) to *blue* (best predictive ability). *LJ* CoMFA Lennard-Jones molecular field, *Ind* CoMFA Indicator steric molecular field, *Par* CoMFA Parabolic steric molecular field, *Gα* CoMSIA Gaussian steric molecular field with attenuation factor of α

Analysis for the two models, Indicator and Gaussian (α=0.3) at grid spacing 1.5 Å, were repeated using leave-many-out cross-validation (random partitioning of the dataset into 70% training and 30% test set) in SYBYL for all datasets. These results were of a similarly high statistical significance (95% CI of R^2_{cv} difference: 0.05–0.13, $P<$ 0.0005).

CoMFA steric fields

In general, there were no statistically significant differences in the predictive ability between the CoMFA steric fields (LJ, indicator and parabolic) when equivalent grid spacings were used. The only exception was a minor improvement of the LJ field over the indicator steric fields at a grid spacing of 2 Å (95% CI of R^2_{cv} difference: 0.004–0.036, $P<0.05$).

CoMSIA Gaussian steric fields

The value of the Gaussian steric field attenuation factor was varied from 0.1 to 0.5 in steps of 0.1. The distribution of R^2_{cv} across the datasets tested is shown in Fig. 3 using a grid spacing of 1.5 Å (and more generally in Supplementary Fig. 1). The predictive ability of these models varied in a highly statistically significant manner (see Supplementary Table 1 for all statistical comparisons). A consistent trend was found, in which the higher attenuation factor values of the Gaussian steric potential resulted in a higher average predictive ability. The default attenuation factor value in SYBYL is 0.3 and this is used in almost all published CoMSIA studies. It was demonstrated that, with a grid spacing of 1.5 Å, a steric Gaussian attenuation factor value of 0.3 significantly improved predictive ability over an attenuation factor value of 0.1 (95% CI of R^2_{cv} difference: 0.03–0.09, $P<10^{-5}$). Additionally, increasing the attenuation factor from 0.3 to 0.5 further improved predictive performance (95% CI of R^2_{cv} difference: 0.03–0.07, $P<$ 0.0005). However, even at the optimized steric Gaussian attenuation factor value of 0.5, the CoMFA steric potentials (LJ, indicator and parabolic) generated models with statistically significantly superior R^2_{cv} values at grid spacings of 1 Å and 1.5 Å. At a grid spacing of 2 Å, the Gaussian potential was still inferior but the differences were no longer statistically significant.

Grid spacing

The influence of grid spacing was statistically significant only between 1 Å and 2 Å for the CoMFA steric fields, with the 1 Å grid spacing resulting in a higher R^2_{cv}. This effect was strongest for the indicator field (95% CI of R^2_{cv} difference: 0.03–0.10, $P<0.0005$). At 1 Å grid spacing, the indicator steric field is as good, if not better, than the LJ field. At 2 Å, the predictive performance of the indicator steric is greatly reduced and is statistically significantly worse than the LJ field with 2 Å grid spacing. There were no statistically significant differences between the grid spacing of the Gaussian steric fields at any attenuation factor values assessed.

Clustering of datasets and steric field settings

Specific steric field settings (the combination of steric potential and grid spacing) were hierarchically clustered based on similarity for the datasets they predicted well (i.e. high R^2_{cv}) and poorly (i.e. lower R^2_{cv}). CoMFA steric field types cluster separately to CoMSIA (Gaussian) steric field potentials. Within these clusters, the steric field settings cluster by grid spacing.

Datasets were clustered into three main groups. The first and largest group contains datasets that are predicted better by the CoMFA steric field types. The second group contains datasets that are predicted as well or better by the CoMSIA steric field types with a high attenuation factor ($\geq$=0.3). The final group is the smallest and contains a couple of datasets that are predicted best by CoMSIA steric field types with a lower attenuation factor.

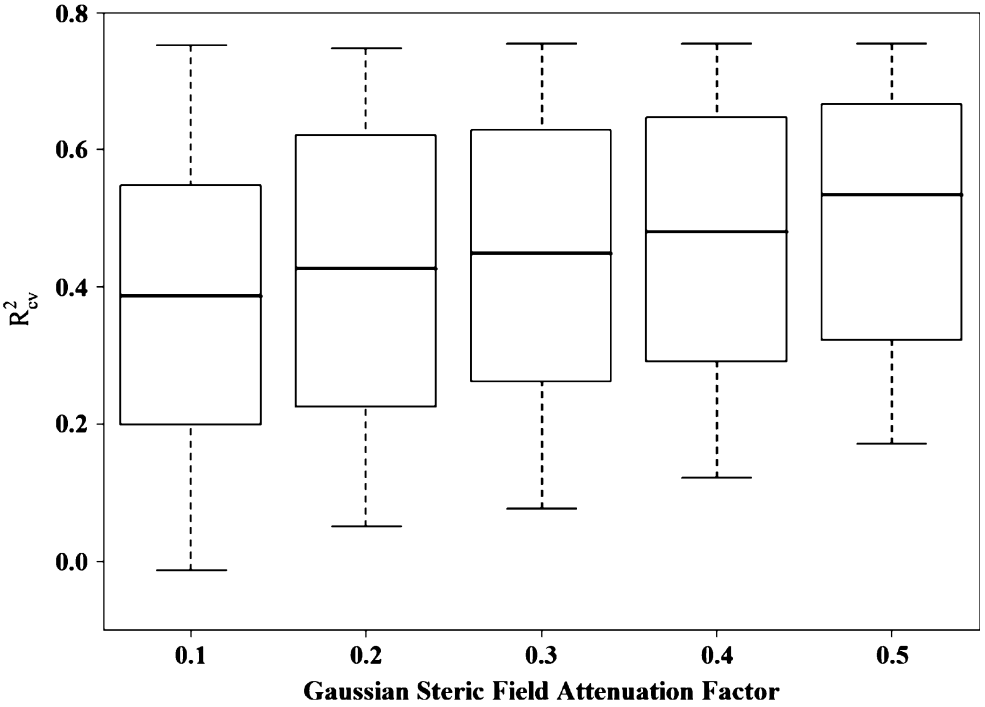

Fig. 3 Box plots displaying the distribution of R^2_{cv} across 28 datasets for models generated using Gaussian steric fields with different attenuation factors at a grid spacing of 1.5 Å

Column filtering

All the models were reanalysed with column filtering of 1.0 and 3.0 kcal mol^{-1}. No statistical difference was noted between column filtering at 2.0 kcal mol^{-1} and 3.0 kcal mol^{-1} (95% CI of R^2_{cv} difference: 0.007–(−0.002), P= 0.263). However, there was a statistical difference between the R^2_{cv} distribution obtained with column filtering of 1.0 and 2.0 (95% CI of R^2_{cv} difference: 0.017–0.007, $P<10^{-5}$), and additionally between column filtering of 1.0 and 3.0 (95% CI of R^2_{cv} difference: 0.02–0.009, $P<10^{-5}$). Generally, the results obtained with a column filtering of 1.0 were found to be better than with column filtering of 2.0 or 3.0. Supplementary Fig. 1 displays the median values of the 24 models for all datasets for different column filtering values.

Discussion

CoMFA vs CoMSIA steric fields

The CoMFA steric fields (LJ, indicator and parabolic) generally have a higher average R^2_{cv} compared to the Gaussian steric field used in CoMSIA. It has been hypothesised that Gaussian fields are less susceptible to small variations in molecular alignment and hence likely to perform better [2, 14, 15]. However, the results of this study indicates that steric potentials with strong distance dependence (i.e. LJ, indicator and parabolic fields) produce better results than Gaussian functions with a smooth transition. The default attenuation factor used in CoMSIA analyses is 0.3. On the basis of the results reported here, it would seem prudent for future CoMSIA analyses to trial an attenuation factor of 0.5 for the steric Gaussian field in order to optimise model prediction accuracy.

CoMFA steric fields

The indicator steric field with 1 Å lattice spacing resulted in the most predictive model for a number of the datasets. Parabolic field settings with 1 Å and 1.5 Å grid spacing also showed strong predictability for some datasets. There are a number of studies that indicate that the indicator and parabolic fields tend to perform better than the LJ field [10, 16–19], while others contradict this [20, 21]. It is likely that the conflicting results are a consequence of the small number of datasets compared. These differences can be explained by the results presented here. The difference between indicator and LJ steric fields is fairly small at a grid spacing of 1 Å or 1.5 Å and hence for some datasets the indicator will be slightly better and for others LJ will be slightly better. Larger differences can be seen at a grid spacing of 2 Å as the performance of the indicator variable deteriorates quickly with a sparser molecular field sampling.

CoMSIA steric fields

The results of the present investigation demonstrate that Gaussian steric fields with a higher attenuation factor yield

better R^2_{cv} values. Statistically, there is a significant effect on the calculated R^2_{cv} values for the same field sampling density upon changing the attenuation factor ($0.1 \leq \alpha \leq 0.5$). This indicates that, for most datasets, it is more important to capture local steric information rather than global molecular shape features. However, as shown in Fig. 4, there are a small number of datasets for which capture of the global steric information (using a Gaussian field with a lower attenuation factor) results in improved predictivity over alternatives that take predominantly local steric information into account. This additionally demonstrates the heterogeneity of datasets and underlines why it is not possible to choose one setting that is optimal for every dataset, but only the best on average.

Grid spacing

A number of studies have been undertaken using a very limited number of datasets to understand the issue of grid spacing [22–25]. These have generally indicated that grid spacing does not dramatically affect CoMFA results. Here we have demonstrated that there is a statistically significant effect of grid spacing on prediction accuracy for certain steric field settings.

Lennard Jones, indicator and parabolic steric fields (CoMFA steric fields) were prone to variations in prediction accuracy with changes to grid sampling density. Although not statistically significant, a lattice spacing of 1.5 Å results in models with decreased predictive accuracy compared to a lattice spacing of 1 Å. There were statistically significant differences between 1 Å and 2 Å and between 1.5 Å and 2 Å for the CoMFA field types in general, but especially for the indicator fields, which require denser sampling.

Gaussian steric fields (CoMSIA steric fields) were fairly resistant to variations in grid spacing. Although there were no statistically significant differences in model predictive accuracy among the various grid spacings, a few trends were observed. Analysis with a lattice spacing of 2 Å tends to perform better when the attenuation factor is 0.3 or less. When the attenuation factor was 0.4 or higher, the

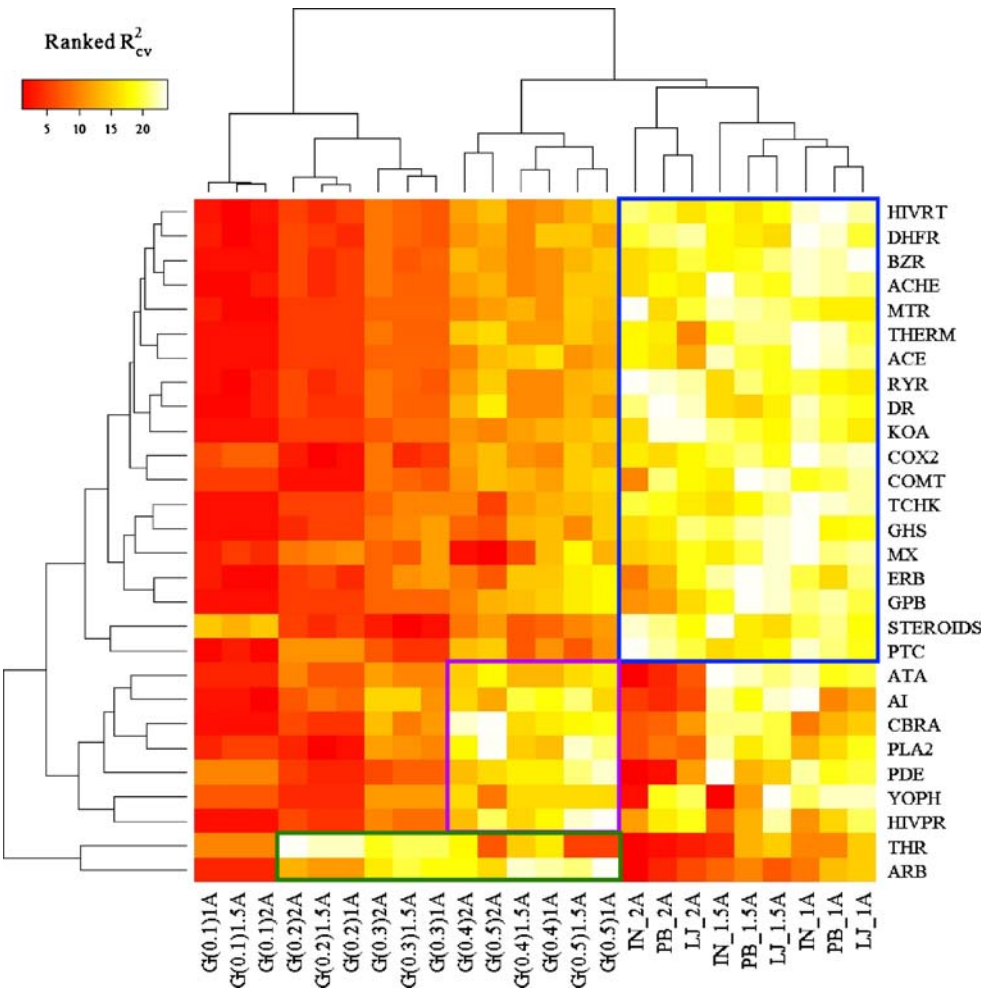

Fig. 4 Heat map and hierarchical clustering of R^2_{cv} rankings of each steric molecular field setting for each dataset. The higher the R^2_{cv} ranking (the brighter the colour on the heat map), the better the steric molecular field setting (combination of specific steric potential and specific grid spacing value) is able to generate a predictive model for the respective dataset. The row labels refer to specific datasets (see Table 1 for details) and the column labels refer to specific steric molecular field settings, which are a combination of a steric potential and a grid spacing value. For example 'G (0.1)_1A' indicates 'Gaussian steric field with an attenuation factor of 0.1 and a grid spacing of 1 Å' and 'PB_2A' refers to 'Parabolic steric field with a grid spacing of 2 Å'

behaviour was similar to the COMFA fields in that model, with lower lattice spacings performing better. For $\alpha \leq 0.3$, these results were apparently consistent with those of a previous study carried out by Hou et al. [26].

Thus, grid spacing has the greatest effect on CoMFA steric fields, especially indicator fields. This makes sense because the steeper the function (i.e. how quickly it changes with distance), the more important it is to have denser sampling to accurately capture information on the field.

Clustering of datasets and steric field settings

The clustering of steric settings by CoMFA/CoMSIA field type and grid spacing indicates that both of these features influence the predictive ability of the models generated. Additionally, the clustering of datasets indicates that the majority of datasets (the uppermost cluster or Fig. 4) are modelled best using a CoMFA steric field (LJ or indicator). Although this is a useful general rule, the other two clusters indicate that for some datasets the CoMSIA Gaussian steric fields are as good or better. This demonstrates the difficulty of treating all datasets as equivalent. Clearly, it is not feasible to say that using CoMFA steric fields will result in the optimal prediction accuracy for any dataset, only that it is the most likely to give the best results on average. A future goal will be to understand what is different between the datasets that makes them more suitable for one method than another.

Column filtering

Generally, grid points with low statistical variance can be eliminated without compromising the results. The default column filtering used in SYBYL is 2.0 kcal mol^{-1}. A past study suggested that the higher values of column filtering influence the precision of the CoMFA results and the number of latent variables [7, 27, 28]. Kim proposed an optimum limit of column filtering of between 0.05 and 2.0 kcal mol^{-1} [7]. The present study demonstrates that a column filtering of 1.0 resulted in a statistically significant improvement in R^2_{cv} over a column filtering of either 2.0 (default) or 3.0. This indicates that, in some cases, the default minimum standard deviation may be too high, resulting in filtering of useful molecular information.

Limitations of the study

The number of datasets used here to compare the effects of changes in steric field settings was much greater than in any previous similar study. The current study was thus able to demonstrate a number of relatively small differences between the choice of steric field and grid spacing. Nevertheless, it would still be advantageous to increase the sample size to allow detection of even smaller differences in prediction accuracy in a statistically significant manner.

It is possible that the differences between steric fields and grid spacing shown here would be altered when steric fields are used in combination with other fields (e.g. electrostatic or hydrophobic). This may result from redundancy between the molecular information captured by different field types [29]. Further studies will be required to study the importance of this. Nevertheless, it is useful to analyse steric fields in isolation in order to interpret the effect of different field functions on the steric information captured at different sampling densities.

Even though the effects of various parameters such as lattice spacing, transformation, attenuation factor and column filtering were studied, there is still scope to further expand the current study to investigate other factors such as smoothing functions, cut-off values, and grid orientation.

Conclusions

It is evident from previous literature studies that methodical variations of COMFA options, e.g. lattice spacing, cut-off values, and box size, affect the predictive accuracy of the models generated [10, 30]. This study evaluated the various steric field settings available for CoMFA and CoMSIA analyses. In total, 24 models were generated for each of the 28 different datasets by varying the grid spacing, transformation, Gaussian function attenuation factor, and column filtering.

Analysis of the predictive ability of these models (R^2_{cv}) indicates that steric fields in which the values drop off quickly with distance (i.e. LJ, indicator and parabolic fields) tend to perform better than the Gaussian version, which has a slower and smoother decrease. There was no statistically significant differences in model predictivity between LJ, indicator and parabolic steric fields. Gaussian fields were influenced by varying the attenuation factor. Fields with higher attenuation factor yielded better R^2_{cv} values.

Field sampling density (i.e. grid spacing) has a relatively minor influence on the predictive ability of Gaussian steric fields, whereas LJ, indicator and parabolic steric fields are more influenced by such changes. Models with a grid size of 1.5 Å, produced results comparable to 1 Å. Indicator fields with a grid spacing of 1 Å performed best on average, but there is significant variation with regards to which field type and sampling density are best for each individual dataset.

Although variation of column filtering (minimum sigma) had minor effects on the results, a column filtering of 1.0 kcal mol^{-1} gave higher results than column filtering of 2.0 kcal mol^{-1} or 3.0 kcal mol^{-1}, indicating the default minimum variation may result in filtering of useful information for some datasets and steric settings.

Acknowledgements We are grateful to the authors who contributed datasets for analysis in this study.

References

1. Cramer RD, Patterson DE, Bunce JD (1988) J Am Chem Soc 110:5959–5967
2. Klebe G, Abraham U, Mietzner T (1994) J Med Chem 37:4130–4146
3. Zhu LL, Hou TJ, Chen LR, Xu XJ (2001) J Chem Inf Comput Sci 41:1032–1040
4. Klebe G (1998) Perspect Drug Discov 12:87–104
5. Tripos Bookshelf 7.1 TI, St. Louis, MO
6. SYBYL 7.1 (2005) Tripos International, St. Louis, MO
7. Kim K (1995) Comparative molecular field analysis (CoMFA) In: Dean P (ed) Molecular similarity in drug design, 1st edn. Blackie, London, pp 291–331
8. Kubinyi H, Folkers G, Martin YC (1998) 3D QSAR in drug design. Kluwer, Dordrecht
9. Kroemer RT, Hecht P, Guessregen S, Liedl KR (1998) Perspect Drug Discov 12:41–56
10. Peterson SD, Schaal W, Karlen A (2006) J Chem Inf Model 46:355–364
11. Hawkins DM, Basak SC, Mills D (2003) J Chem Inf Model 43:579–586
12. Cunningham SL, Cunningham AR, Day BW (2005) J Mol Model 11:48–54
13. R Development Core Team (2006) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna
14. Bang SJ, Cho SJ (2004) Bull Korean Chem Soc 25:1525–1530
15. Sutherland JJ, O'Brien LA, Weaver DF (2004) J Med Chem 47:5541–5554
16. Dixit A, Kashaw SK, Gaur S, Saxena AK (2004) Bioorg Med Chem 12:3591–3598
17. Gaur S, Prathipati P, Saxena M, Saxena AK (2004) Med Chem Res 13:746–757
18. Pandey G, Kashaw SK, Saxena AK (2004) Med Chem Res 13:677–686
19. Kroemer RT, Hecht P (1995) J Comput Aided Mol Des 9:205–212
20. Ducrot P, Legraverend M, Grierson DS (2000) J Med Chem 43:4098–4108
21. Pajeva IK, Wiese M (1998) Quant Struct-Act Rel 17:301–312
22. Jung M, Kim H (2001) Bioorg Med Chem Lett 11:2041–2044
23. Wang BL, Ma N, Wang JG, Ma Y, Li ZM, Li YH (2004) Acta Phys-Chim Sin 20:577–581
24. Zhu LL, Xu XJ (2002) Acta Phys-Chim Sin 18:1087–1092
25. Kunick C, Lauenroth K, Wieking K, Xie X, Schultz C, Gussio R, Zaharevitz D, Leost M, Meijer L, Weber A, Jorgensen FS, Lemcke T (2004) J Med Chem 47:22–36
26. Hou TJ, Li YY, Liao N, Xu XJ (2000) J Mol Model 6:438–445
27. Kim KH (1993) Use of the hydrogen bond potential function in comparative molecular field analysis (CoMFA): an extension of CoMFA. In: Kubinyi H (ed) 3D QSAR in drug design: theory, methods and applications, vol 1. ESCOM, Leiden, pp 245–251
28. Martin YC, Kim KH, Lin CT (1996) Comparative molecular field analysis: CoMFA. In: Charton M (ed) Advances in quantitative structure–property relationships, vol 1. JAI Press, Greenwich, CT, pp 1–52
29. Dias MM, Mittal RR, McKinnon RA, Sorich MJ (2006) J Chem Inf Model 46:2015–2021
30. Cramer III RD, Depriest SA, Patterson DE, Hecht P (1993) The developing practice of comparative molecular field analysis. In: Kubinyi H (ed) 3D QSAR in drug design: theory methods and applications, 1st edn. ESCOM, Leiden, pp 443–485
31. Tervo AJ, Nyronen TH, Ronkko T, Poso A (2003) J Comput Aided Mol Des 17:797–810
32. Tervo AJ, Nyronen TH, Ronkko T, Poso A (2004) J Chem Inf Comput Sci 44:807–816
33. Bostrom J, Bohm M, Gundertofte K, Klebe G (2003) J Chem Inf Comput Sci 43:1020–1027
34. Melville JL, Hirst JD (2004) J Chem Inf Comput Sci 44:1294–1300
35. Wang RX, Gao Y, Liu L, Lai LH (1998) J Mol Model 4:276–283
36. Hu X, Stebbins CE (2005) Bioorg Med Chem 13:1101–1109
37. Melville JL, Lovelock KRJ, Wilson C, Allbutt B, Burke EK, Lygo B, Hirst JD (2005) J Chem Inf Model 45:971–981
38. Welch W, Ahmad S, Airey JA, Gerzon K, Humerickhouse RA, Besch HR, Ruest L, Deslongchamps P, Sutko JL (1994) Biochemistry 33:6074–6085
39. Hannongbua S, Lawtrakul L, Sotriffer C, Rode B (1996) Quant Struct-Act Rel 15:389–394
40. Luco JM, Feretti FH (1997) J Chem Inf Comput Sci 37:392–401
41. Sulea T, Oprea TI, Muresan S, Chan SL (1997) J Chem Inf Comput Sci 37:1162–1170
42. Belvisi L, Bravi G, Catalano G, Mabilia M, Salimbeni A, Scolastico C (1996) J Comput Aided Mol Des 10:567–582
43. Rivara S, Mor M, Silva C, Zuliani V, Vacondio F, Spadoni G, Bedini A, Tarzia G, Lucini V, Pannacci M, Fraschini F, Plazzi PV (2003) J Med Chem 46:1429–1439
44. Li W, Tang Y, Zheng YL, Qiu ZB (2006) Bioorg Med Chem 14:601–610
45. Hudock MP, Sanz-Rodriguez CE, Song YC, Chan JMW, Zhang YH, Odeh S, Kosztowski T, Leon-Rossell A, Concepcion JL, Yardley V, Croft SL, Urbina JA, Oldfield E (2006) J Med Chem 49:215–223
46. Marini F, Roncaglioni A, Novic M (2005) J Chem Inf Model 45:1507–1519
47. Srivani P, Kiran K, Sastry GN (2006) Indian J Chem A 45:68–76
48. Salo OMH, Savinainen JR, Parkkari T, Nevalainen T, Lahtela-Kakkonen M, Gynther J, Laitinen JT, Jarvinen T, Poso A (2006) J Med Chem 49:554–566
49. Nayyar A, Malde A, Jain R, Coutinho E (2006) Bioorg Med Chem 14:847–856