J Comput Aided Mol Des (2008) 22:541–551
DOI 10.1007/s10822-008-9195-6

Quantitative Series Enrichment Analysis (QSEA): a novel procedure for 3D-QSAR analysis

Bernd Wendt · Richard D. Cramer

Received: 31 October 2007 / Accepted: 7 February 2008 / Published online: 27 February 2008

Abstract A novel procedure is proposed for 3D-QSAR analysis. The composition of 16 published QSAR datasets has been examined using Quantitative Series Enrichment Analysis (QSEA). The procedure is based on topomer technologies. A heatmap display in combination with topomer CoMFA and a novel series trajectory analysis revealed critical information for the assembly of structures into meaningful series. Global and local centroid structures can be determined from a similarity distance matrix and build the origins for stepwise model building by increasing the similarity radius around the centroid nucleus. The results indicate that the new procedure allows determination of whether compounds belong to an emerging structure-activity relationship and which compounds can be predicted within reliable limits.

Keywords 3D-QSAR · Centroid · CoMFA · Heatmap · Series analysis · Topomer

Introduction

The modeling of quantitative relationships between chemical structures and their biological function (QSAR) is a key methodology in modern drug discovery and development. QSAR methods are usually applied to direct synthesis of new compounds. There exists a variety of QSAR methods representing combinations of various descriptor sets with various data handling schemes. Descriptors are traditionally categorized into 2D-descriptors (Hansch-type, topological indices, hologrames) [1–3] and 3D-descriptors (CoMFA, COMSIA, COMPASS, HINT, a. o.) [4–7].

The standard approach in an optimization project is to generate a QSAR model from all structures that have already been synthesized and tested, and then to use the model to predict for new molecules. Traditionally the main focus in model building has been which method to choose or better which method successfully provides a quantitative relationship. However the composition of the dataset plays an important role as well [8–10]. Model building can fail or even worse good statistical models with poor predictivity may result if structural information is not varied independently and continuously. This phenomenon is usually associated with 2D QSAR methods where inclusion of compounds into the QSAR series is merely controlled by availability of structures and associated activity data. There is a different situation with 3D-QSAR methods such as CoMFA [4]. The composition of the series is highly restricted and involves a lot of human labor (also a lot of subjective bias). A compound will be included in a 3D-QSAR series only if it is possible to align a 3D conformation of its structure onto a 3D model of the data where the 3D model could be a pharmacophore [11] or a receptor binding site [12]. Inclusion of structures to the series is therefore controlled by the outcome of the chosen alignment approach. All structure-activity information of compounds for which activity data would be available but where alignment of its structures to the 3D model failed

Electronic supplementary material The online version of this article (doi:10.1007/s10822-008-9195-6) contains supplementary material, which is available to authorized users.

B. Wendt (✉)
Tripos International, Martin-Kollar-Strasse 17,
Munich 81829, Germany
e-mail: bwendt@tripos.com

R. D. Cramer
Tripos International, 1699 South Hanley Road,
St. Louis, MO 63144, USA

will be missed for QSAR analysis. This can cause model building to fail when the comprised SAR-information of the successfully aligned compounds is too sparse.

It seems that for 2D-QSAR methods the assembly of series is not properly restricted whereas for 3D-QSAR this assembly process is perhaps too heavily restricted. We will propose a new procedure developed from topomer technologies [13–17] for control of the assembly of 3D-QSAR series that will help to overcome these limitations. The workflow of this procedure is schematically presented in Fig. 1. The first step is the creation of a heatmap from a similarity distance matrix followed by the determination of the centroid compound(s). Next the structures of the series are sorted by increasing topomeric distance to the centroid compound. Starting from the first three compounds a topomer CoMFA model is generated and test compounds are predicted that are below a certain threshold distance from the lastly added training set compound. To the training set is iteratively added the next structure of the sorted list until all structures of the series are included in the training set. The new workflow will be exemplified on 16 published datasets.

Methods

Datasets

The datasets used in the present study (see Table 1) are covering two QSAR benchmark datasets, three datasets that have been used mainly with 2D-QSAR methods and 11 CoMFA datasets that have been presented in the original topomer CoMFA publication [17].

The set of 31 steroids with corticosteroid-binding globulin affinities has traditionally been used as benchmark for testing novel QSAR descriptors or data handling schemes [18–20]. The Selwood dataset of 31 Antimycin A1 [21] analogues is also a well known benchmark dataset for QSAR applications and has been used in a number of publications [22–26]. A dataset of D1 antagonists was created from a set of rather diverse structures and was presented in a couple of publications where new 2D-QSAR methods were applied [27, 28]. The dataset of PDGFR-inhibitors from Pandey [29] consisting of 79 piperazinyl-quinazolines have been used in a couple of publications using 2D-QSAR descriptors [30, 31]. The dependent variable used for the current study was the inhibition of ß-PDGFR phosphorylation in absence of human plasma expressed as the logarithm of reported IC50 values. The dataset of human carbonic anhydrase inhibitors (hcaii) generated by Scozzafava [32] has been used as test case in a couple of publications where QSAR models were generated using 2D-descriptors such as topological, geometric and electronic features of molecules [33, 34]. The set was derived from a combinatorial library of 25 sulfonamide compounds reacted with a set of 4 sulfonyl chlorides and 2 acetyl chlorides to yield 150 compounds. The dependent variable used for this study was the logarithm of the Ki value of human carbonic anhydrase II. The analysis of the 11 CoMFA datasets [35–45] chosen for demonstration of our new workflow have partly been described in previous publications [17, 46].

Construction of heatmap and determination of centroid compound

The similarity distance matrices were calculated by the program dbtop [13] using the same structure file as a source for reference as well as candidate structures. A high distance cutoff (999) and a high maximum heavy-atom difference (50) were used to force the calculation of a distance value. All structures of the input file were marked

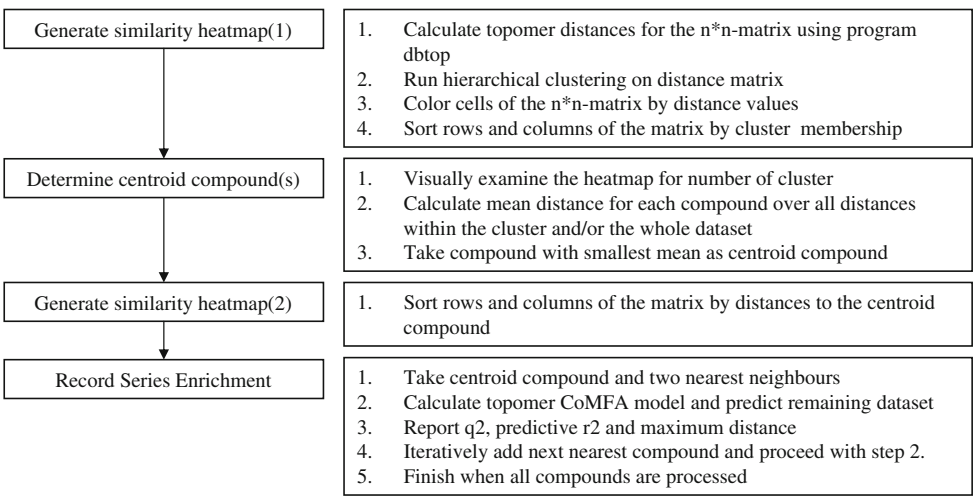


Fig. 1 Summary description of the QSEA protocol

Table 1 Details of the 16 QSAR datasets used in QSEA

Dataset Name	Size	Biological activity	Ref #
Steroids	31	Affinity to corticosteroid globulin	[18]
Selwood	31	Antifalirial activity against D. viteae	[21]
d1	48	Displacing of SCH23390 from Dopamine D1 receptor	[27]
pdgfr	79	Inhibition of b-PDGFR phosphorylation	[29]
hcaii	150	inhibition of human carbonic anhydrase II	[32]
a2a	78	Affinity to Adenosin-A2a receptors	[35]
acest	41	Inhibition of acetylcholin-esterase	[36]
cannab	61	Displacing of WIN-55212-2 from CB1-receptor	[37]
d4	29	D4 receptor antagonism	[38]
flav	38	binding to benzodiazepine site in GABAa receptors	[39]
hiv	25	Inhibition of HIV-1 protease	[40]
5ht3	61	displacing of 5-ht3 from NG 1-8-15 cells	[41]
ice	38	Inhibition of interleukin 1-b converting enzyme	[42]
mao	71	inhibition of monoamine oxidase A	[43]
rvtrans	82	Protection of MT4 cell from HIV-1	[44]
Thrombin	72	Inhibition of thrombin	[45]

with a specific attribute at the splitting bond which for the hcaii set was the sulfonamide bond to force the splitting of the structures at the specified bond [47]. The resulting ascii files were processed by python-scripts to symmetrize the distance matrix (where the smallest of the two distance values was taken) and to allow importing of the values into a SYBYL molecular spreadsheet. The similarity distance values of the topomeric distance matrix were used as descriptor in a hierarchical clustering analysis run in SYBYL using the complete linkage method. Usually the cluster level showing the highest distance to the next level was chosen and the corresponding cluster-IDs were added into a separate column of the spreadsheet. This cluster column was used to sort the rows and columns of the spreadsheet according to cluster membership. The sorted spreadsheet was loaded into Microsoft Excel where the cells of the spreadsheet were filled with temperature colors representing close (=hot) or far (=cold) similarities. The resulting similarity heatmap gives an overview of subsets and centroids within the dataset as exemplified with the heatmap of the hcaii set in Fig. 2. The global centroid compound was determined by the lowest topomeric distance mean over all rows in the spreadsheet. For an overview of the global centroid structures of each dataset and their split bond see Fig. 3. With the global centroid compound on top all rows and columns in the spreadsheet are sorted by increasing topomeric distance to this centroid compound. This representation of the similarity heatmap was used for all datasets. The SLN-input file for the series trajectory was also sorted in the order of the compounds in the heatmap. Local centroids were determined by the lowest topomeric distance mean over all rows of a particular cluster. Series trajectories from local centroids were recorded with SLN-input files starting with the local centroid structure and sorted in order of increasing topomeric distance to this centroid.

Series trajectories

The procedure for generating a topomer CoMFA model is performed by an SPL-script and has been detailed elsewhere [17]. In brief summary the structures and potencies are read from an SLN-file into SYBYL starting from the centroid compound in order of increasing topomeric

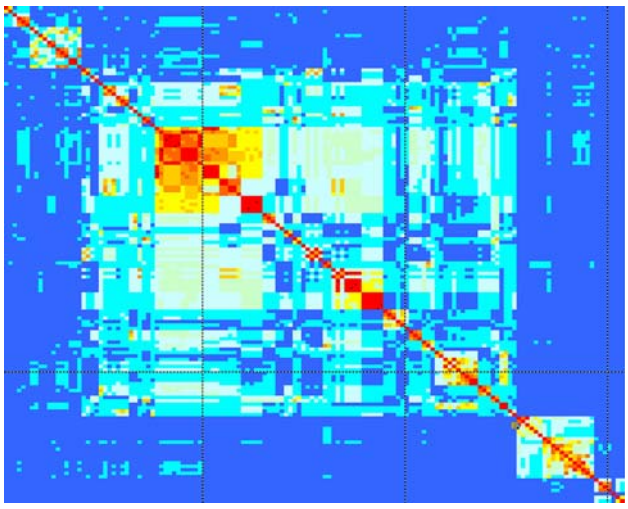


Fig. 2 A similarity heatmap display of the hcaii dataset. All rows and columns are sorted by cluster membership from hierarchical clustering

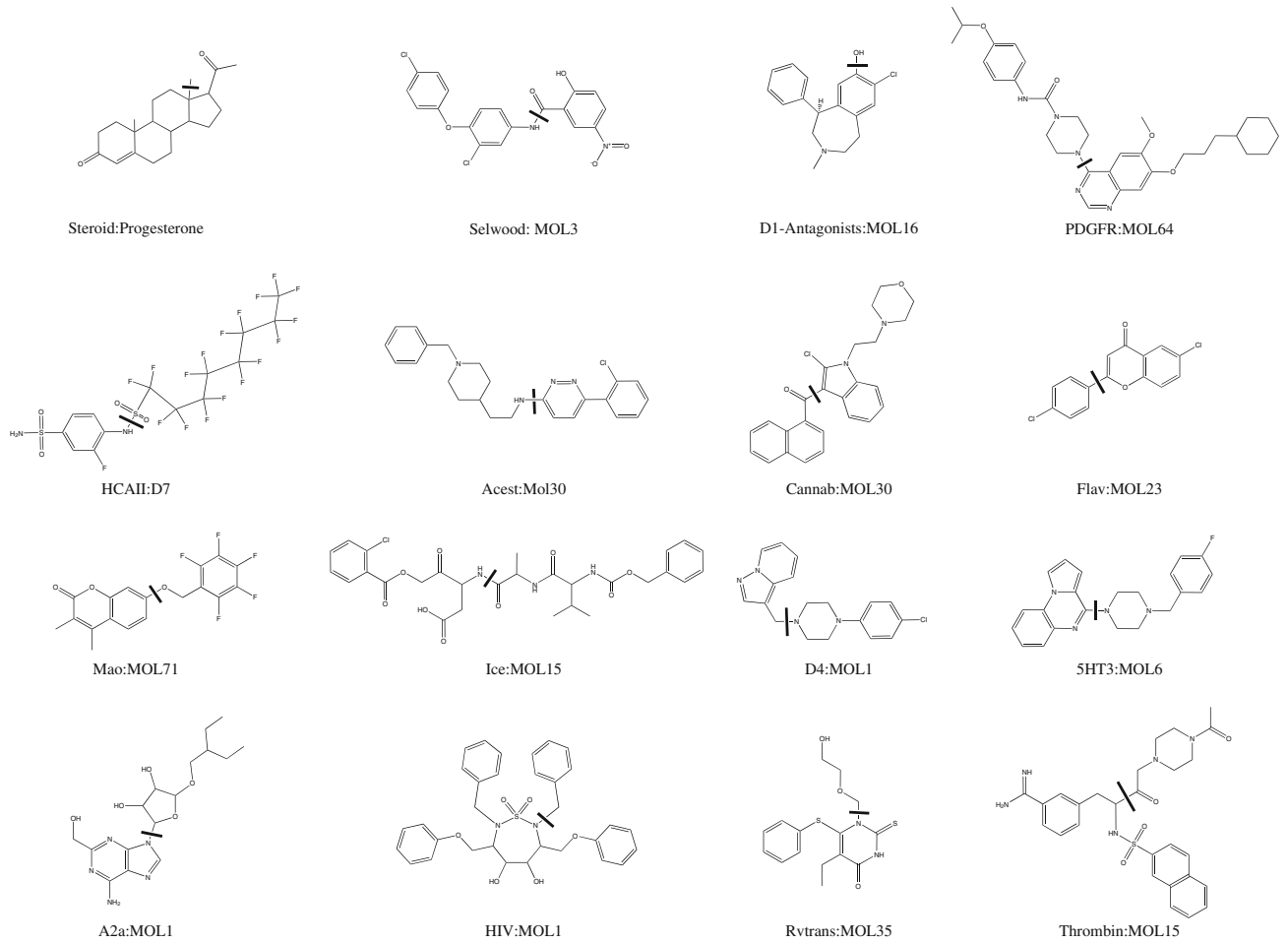


Fig. 3 The centroid compound of each dataset with the split bond marked

distance to this centroid. In each structure one bond is marked for splitting the full structures into two fragments. For each of the two fragments a 3D topomer model is calculated and stored in separate columns of a molecular spreadsheet in SYBYL. PLS-analysis is performed with leave-one-out crossvalidation. The number of PLS-components is chosen as that yielding the first minimum standard error of prediction. Prediction of test compounds from the complete dataset was restricted to groups of compounds based on the ratio of their topomer distance to the centroid to the current training set radius, excluding the less topomer-similar structures, at a cutoff of 1.2× and 1.5× that radius. Each of the test compounds are fragmented and modeled in the same way as the training compounds. The first three structures of the input file are used as nucleus, and then iteratively added to the training set, one by one, the topomerically nearest structure (where nearness is the distance to the centroid compound) followed by generating a topomer CoMFA model and prediction of test compounds within the two cutoff radii. Each trajectory tracks (among other parameters) the q^2 of the topomer CoMFA model, the maximum distance to the cluster centroid within the training set and the predictive r^2 calculated over predicted structures at two different cutoff radii (for definition see [48]).

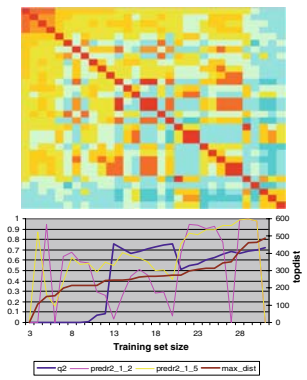
Visual analysis is performed with the series trajectory arranged below the heatmap of the series such that every column in the heatmap is in line with the corresponding structure added to the training set (see Fig. 4).

Results and discussion

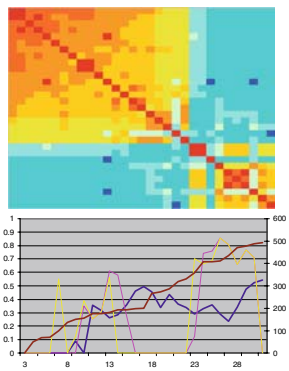
Steroids

Examination of the similarity distance matrix of the steroids reveals the existence of two clusters comprising 17

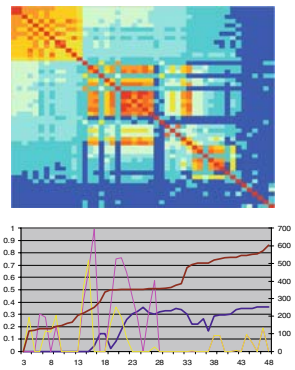
Fig. 4 QSEA-display for each of the 16 datasets. A heatmap is ▶ arranged on top of a series trajectory reporting the progress in q^2, predictive-r^2 (where q^2 is derived from the training set and predictive-r^2 is derived from the two test-sets at different cutoff radii [47]) and maximum topomeric distance to the centroid



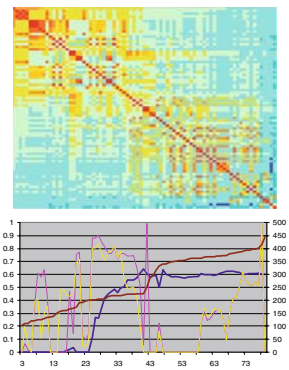
steroids



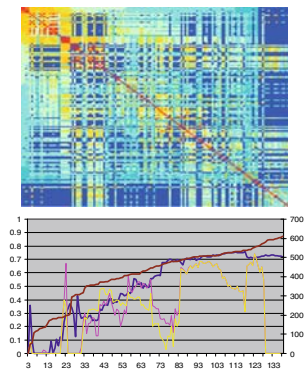
Selwood



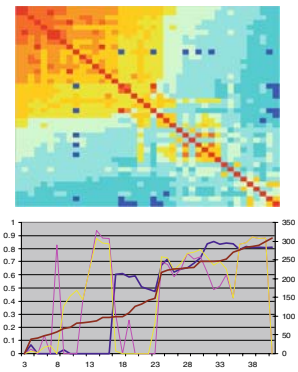
d1



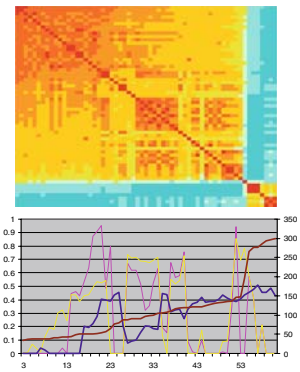
pdgfr



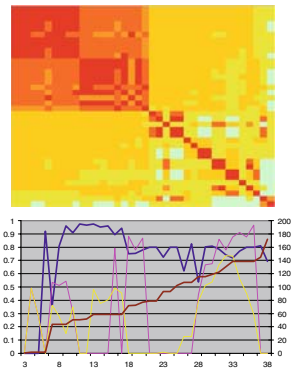
hcaii



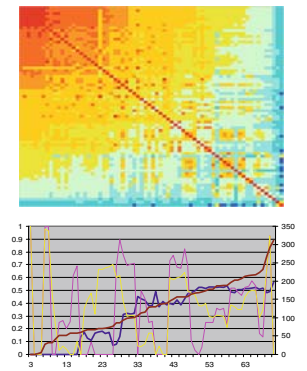
acest



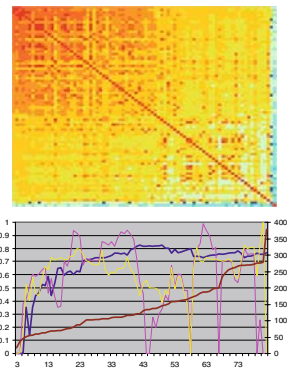
cannab



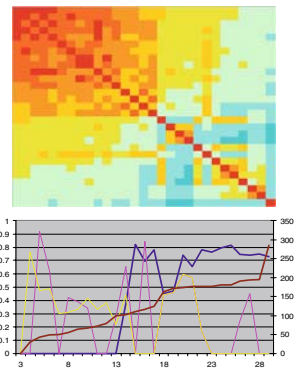
flav



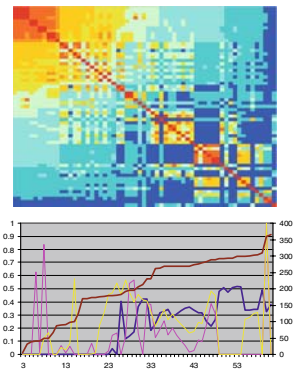
mao



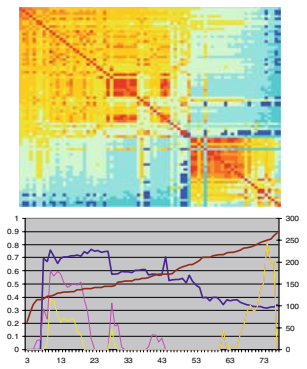
rvtrans



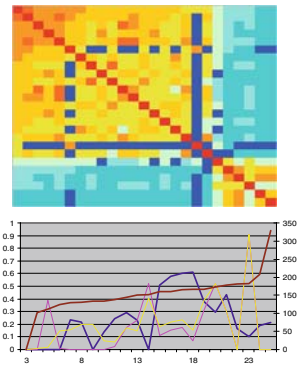
d4



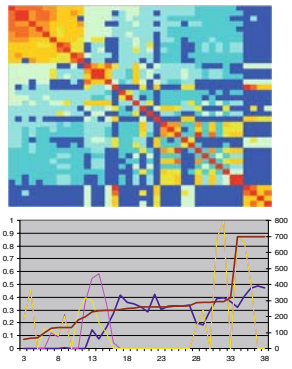
5ht3



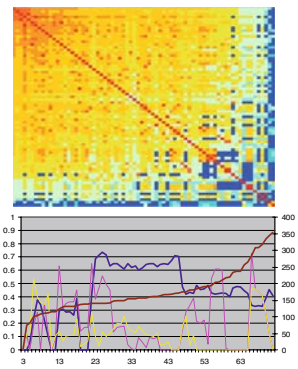
a2a



hiv



ice



thrombin

steroids having a keto-group at C3 and a second cluster of 14 steroids with mainly hydroxyl groups at C3. There is one outlier: aldosterone, mainly because of the missing methyl group at C3. On the basis of the distance matrix progesterone was identified as the global centroid and was taken together with its two most topomerically similar structures (16-methyl-progesterone and norprogesterone) as nucleus for series trajectory analysis. Figure 4 shows the topomeric distance matrix of the 31 steroids together with its series trajectory as well as the corresponding sets of figures for the other 15 datasets. The topomeric distance matrix was sorted according to the sequence of the series trajectory with the global centroid compound at the top followed by the other compounds in order of increasing topomeric distance to this centroid compound. In the trajectory of the steroid dataset the curve of the maximum distance to the cluster centroid shows a smooth increase throughout which reflects the continuous construction of the model on the basis of the individual changes to the progesterone nucleus. Any new changes of methyl-, hydroxyl- and carbonyl- groups at the variant positions (C2, C3, C9, C11, C16, C17 and C19) as well as changes of double-bonds versus single-bonds in the A-ring eventually cause fluctuations of predictive r^2 curves whenever they occur for the first time. Drops in predictive r^2 rebound as soon as these changes occur more often. The q^2 curve reflecting the internal consistency of the steroid set sets in at a training set size of 13 with a q^2 of 0.75 and remains rather stable until the end of the trajectory. Surprisingly, models from smaller training set sizes, i.e. smaller than 13 compounds, that do not show any internal consistency show very good external predictivity instead. The number of predicted compounds is on average 9 with natural minima at the beginning and at the end, and the maximum with 17 compounds at a training set size of 8. Interestingly, at this training set size, the q^2 is negative (−0.3) but the predictive r^2 is highly positive (0.63) with a moderate SDEP of 1.05.

Selwood

The topomeric distance matrix of the Selwood dataset (as shown in Fig. 4) was constructed with all structures split at the central amide bond. The similarity distance matrix resembles that of the steroid dataset. There is one main cluster with 17 amide compounds in upper left hand that is composed of aniline-components coupled with meta-nitrobenzoic components. The other 14 structures in the lower right hand are less related to one another and represent amides originating from a different set of amines (mainly aliphatic chains) and different benzoic-components. There is one outlier MOL31, a benzophenone missing an amide bond.

The trajectory of the Selwood set is split in three segments with fluctuations in the internal consistency (q^2) and external predictivity (pred_r^2). The trajectory shows a maximum in q^2 at training set size of 17, this coincides with a jump in maximum distance to the cluster centroid. The first 17 compounds are exactly the set of compounds of the main cluster in the heatmap therefore this maximum q^2 value represents the completion of a local trend. This first cluster is composed of benzoic acids coupled to anilines with a phenolether group in para-position. After the jump the internal consistency is falling but is rebounding at the end of the trajectory. It seems that the discontinuity in the magnitude of similarity allows more dissimilar structures to be added to the training set, each one introducing multiple changes. The predictivity rebounds first when all these multiple changes have been varied independently by additional structures. A second local trend could be verified by starting another trajectory (results not shown) from a local centroid identified among the 14 loosely bound structures of the lower right hand corner of the similarity distance matrix where a QSAR resulted for 14 compounds with a q^2 of 0.48. This underlines the assumption that the whole Selwood dataset can be viewed as a combination of two distinct subsets. The final segment of the shown trajectory of the Selwood dataset shows high external predictivity and high internal consistency. This indicates that the contributions to the model from both subsets are additive resulting in a global model with a q^2 of 0.55. Figure 5 shows the topomer overlay of all structures of the Selwood set with the amine components on the right and the benzoic components on the left indicating the close relationships among the structures of this set and also showing the separation between the aryl and aliphatic amine components of the two different subsets. Also shown in this figure are the CoMFA contour plots from the final model indicating favourable steric interactions with green contours near the 2- and 3-position of the benzoic component as well as near the para-position of the second phenyl and in the middle of the long aliphatic chain of the amine component. Yellow contours near the beginning and tail of the aliphatic chain indicative of disfavoured steric interactions suggest an optimal length of the chain with 8 carbons. Electrostatic interactions indicated by red contours near the ether-oxygen of the amine component as well red and blue contours near the 5-substituent (nitro favoured/amide disfavoured) on the benzoic component suggest importance of heteroatoms on these positions.

d1

Visual examination of the similarity distance matrix (Fig. 4) reveals three clusters and a number of singletons. The two larger clusters contain 17 benzazepines with

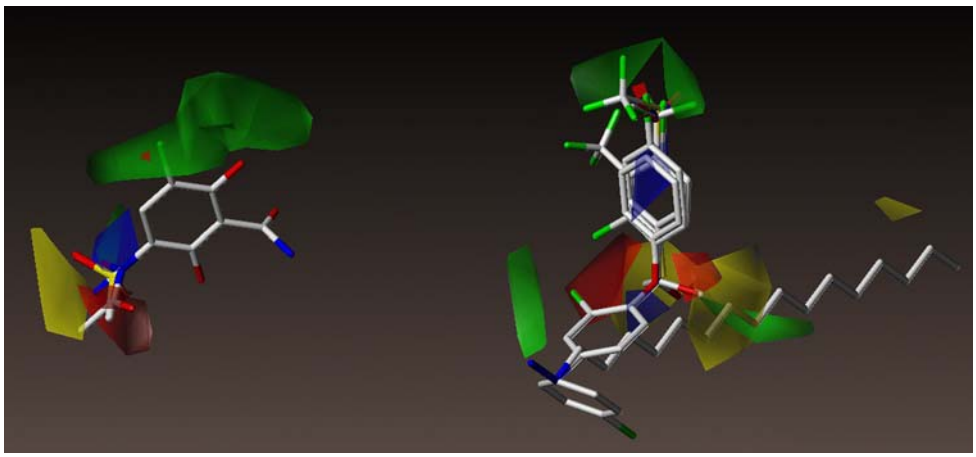


Fig. 5 Topomer overlays of all 31 compounds of the Selwood dataset. *Note*: As a result of topomer alignment there is only one amine component with a long aliphatic chain visible since all the other aliphatic amine components are aligned exactly on top of this longest one and hence invisible

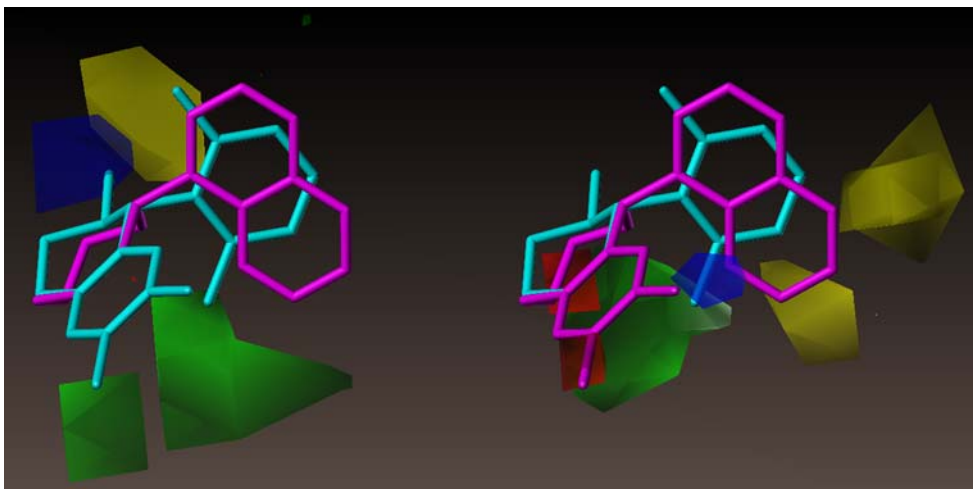


Fig. 6 Topomer overlay of an inactive benzazepine (MOL36, magenta) and an active isoquinoline (MOL14, cyan) with contour plots from an isoquinoline model on the left and from a benzazepine model on the right. Green contours indicate regions where steric interactions are favoured, yellow contours those regions where steric interactions are disfavoured. Negative charge is favoured near red contours (disfavoured near blue contours) and positive charge is favoured near blue contours (disfavoured near red contours)

aromatic substituents in 2-position (cluster 1) and 15 isoquinolines with substituents in 1-position (or 3-position; cluster 2). Both clusters share the substitution pattern on the condensed benzene ring (6-Cl-7-hydroxy). A smaller third cluster contains 5 Aminotetralins with aromatic substituents in 1-position (cluster 3). On a structural basis these three clusters are distantly but clearly related to one another. However, the addition of the remaining13 compounds to the dataset is questionable. Although there might be more distant relationships for a few, it is baffling to include 3 thioxanthenes to the dataset. This might illustrate the situation where mere availability of structures and associated biological activity lead to the inclusion of the data for 2D-QSAR modeling.

The curve for the maximum distance to the centroid in the series trajectory of the d1 set shows two jumps reflecting the composition of the set as displayed in the heatmap. At a training set size of 24 the q^2 reaches a maximum of 0.36 but then falls off and never rebounds. All the three subsets were examined individually by recording series trajectories starting from the local centroids MOL35, MOL5 and MOL24 for clusters 1–3, respectively. The trajectories of MOL35 and MOL5 look similar to the trajectory of the global centroid MOL16, however the trajectory of MOL24 revealed a robust local model with a high q^2 of 0.82. The corresponding topomer CoMFA model reflects a consistent structure-activity-relationship for this set of 17 isoquinolines, a set that basically represents the merger of clusters 2 and 3. Through addition of benzazepines to the training set this relationship vanishes. Basically constant addition of inconsistent data to a consistent relationship is disintegrating that relationship. This situation looks like a limitation of the topomer CoMFA method. However, structural examination of the underlying relationships reveals strength of this method. The conflicting situation is displayed in Fig. 6 where compound MOL14 (cyan) with a pIC50 of 8.43 from the set of isoquinolines and compound MOL36 (magenta) with a pIC50 of 4.48 from the set of benzazepines are shown in their topomeric alignments with CoMFA contours from the model of isoquinolines on the right and from the model of benzazepines on the left. The green contours around the ortho-position of

the pendent phenyl ring of the isoquinoline indicate positive contributions. The underlying SAR comprises a consistent pattern throughout the set. However the same region for the benzazepines seems to be sterically constrained as indicated by the yellow contours. It turns out that for substitutions of the corresponding ring in 2-position of the benzazepines the most active compounds do not have any substitution on that ring, however the least active compounds do have larger groups in that position thereby conflicting with the preferred substitution pattern of the isoquinolines and tetralines.

Lead optimization of d1 antagonists should take special care when comparing results from isoquinolines and benzazepines. Our recommendation would be to handle these two chemotypes with local models rather than global models in order to propose proper directions for synthesis of new compounds. Inclusion of structurally unrelated chemotypes such as thioxanthenes should be avoided.

pdgfr

The topomeric distance matrix was created by splitting all structures at the amide bond connecting the aromatic sidechain to the piperazinylquinazoline scaffold. Overall the structures of the dataset are highly congeneric though the heatmap display (see Fig. 4) suggests a segmentation of the dataset into two larger clusters. It turns out that the main difference between these two clusters is the presence of a basic nitrogen. The 34 structures of the cluster in the lower right hand do bear a basic nitrogen whereas the other 45 structures don't. The heatmap display as well as the series trajectory reminds of the situation with the Selwood dataset since both datasets represent two subsets from which two local models can be built. In fact the local model covering the 34 structures in the lower right hand shows good internal consistency with a q^2 of 0.48, the same applies to the local model of the 45 structures in upper left hand of the matrix with a q^2 of 0.62 that compares well with the global model covering all 79 structures with a q^2 of 0.60. The trajectory shown in Fig. 4 starts from the local centroid MOL64 (a structure missing a basic nitrogen) and shows the buildup of the external predictivity as well as internal consistency. The jump of the maximum distance to the cluster centroid in the trajectory seen at training set size 45 (coinciding with the boundary of the first cluster in the heatmap) signals the completion of the local model and indicates that structures with a positive charge center are added. As a consequence the external predictivity drops down but internal consistency remains stable. After enough SAR data is added, meaning the model is trained for the new trend, the predictivity rebounds. This situation resembles that of the Selwood set representing two local trends with consistent SAR.

hcaii

Figures 2 and 4 show heatmaps of the topomeric distance matrix of 138 compounds (Concord failed to generate 3D coordinates for the 12 thiadiazole-imine structures) sorted by cluster membership and by distances to the centroid compound, respectively. All structures were split at the bond connecting the building blocks of the combinatorial library for the topomeric similarity calculation. From visual inspection of the heatmap in Fig. 2 it turns out that the dataset is composed of numerous clusters. The central core of the matrix comprises 24 compounds that are Sulfonamido-benzenesulfonamides. This central cluster shows distant relationships to the majority of the other compounds. There are two smaller clusters at the top left hand side of the matrix showing no relationships to this central cluster and comprises 6 and 12 compounds containing thiadiazolesulfonamides or benzenesulfonamides with a 5- or 3-atom linker to the second sulfonyl sidechain, respectively. The two other small clusters at the bottom-right hand side of the heatmap comprises 18 and 7 compounds that are benzenesulfonamides with a 2-atom-linker or benzothiazoles with a 1,2 ethanediol linker to the second sulfonyl sidechain, respectively. The global centroid is compound D7, a perfluorooctylsulfonylamido-benzenesulfonamide. The maximum distance curve of the trajectory passes through two steps separated by jumps at training set sizes of 24 and 33 with sharply rising maximum distances to the cluster centroid. Before each jump the predictive r^2 curve is rising and with the jump the curve decreases but the q^2 rises. The drastic jump of the maximum distance causes the cutoff radii used to select topomer-similar structures to rise as well with the consequence of selecting a larger number of compounds for prediction which are in general more dissimilar to the training set compounds. The jump signals completion of a local model indicated by a local maximum of q^2 before the jump. The external predictivity of the local model is low because the chosen topomeric cutoff radii are too high. The similarity gap between the test compounds and the local model is too large. The development of local models can be examined on a structural level as well. Local models at training set sizes 17, 30, 52 and 81 have been examined in more detail. The structures of the first set of 17 structures contain perfluoroalkylsulfonylamido-benzenesulfonamides with different chain lengths and different variations in the 3-position. The next 13 structures introduce aromatic groups in the sulfonyl-sidechain. The next 20 structures introduce 1,3 and 1,2 Amino-sulfonamides with as well as Amino-thiadiazole-sulfonamides and 1,4 Aminosulfonamides with a carboxylic sidechain. With the last level of model expansion benzothiazole compounds are introduced to the training set. The expansion of structural space of the

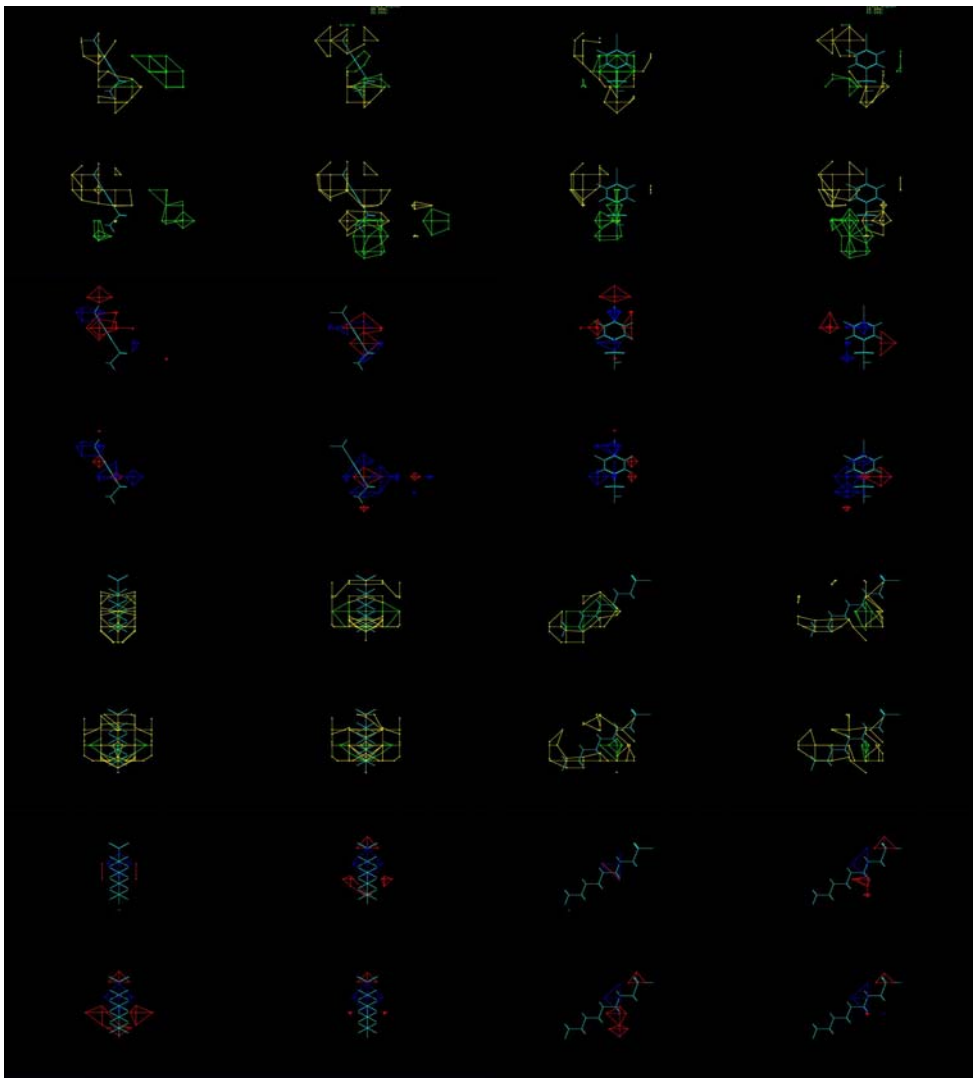


Fig. 7 Orthogonal views (left and right quadrant) of topomer CoMFA results for 4 different models from the hcaii dataset with training set sizes of 17, 30, 52 and 81. Steric contour plots are in rows 1, 2, 5, 6; electrostatic contour plots in rows 3, 4, 7, 8; X1-side of the model in rows 1–4 and X2-side of the model in rows 5–8 with views in column 3 and 4 orthogonal to those in columns 1 and 2, respectively

training set leads to a refinement of the topcomfa model. Figure 7 shows the contour maps of the topcomfa models model_17, model_30, model_52 and model_81 as snapshots of a quartered screen mode where in each quadrant the upper left hand corner is occupied by the contours of model_17, the upper right hand with contours from model_30, the bottom left hand with those of model_52 and the bottom right hand corner with contours of model_81. Orthogonal views are shown for steric and electrostatic contributions and for the X1-side and the X2-side, respectively. By increasing, decreasing or shifting of the polyhedra in Fig. 7 the contours clearly demonstrate the continuous buildup and refinement process of the models starting from the centroid of the central cluster and completing when no further neighbourhood behaviour [49] is available or where structural relationships are no longer present. The boundary of the model generation is reached at a training set size of 100. This coincides with the number of structures showing relationships to the central cluster in the similarity distance matrix. Any addition of another compound to the training set would become more or less arbitrary. Arbitrary selection of compounds is similar to assembly of series for 2D-QSAR methods where there is no restriction such as neighbourhood behaviour or alignment to a 3D model. And that usually leads to loss of external predictivity and internal consistency.

CoMFA-sets

As can be seen from the heatmaps in Fig. 4 the pairwise similarities of the structures of all 11 datasets are much more pronounced compared to the other sets. The global centroid of each dataset was used as nucleus for series trajectory analysis and model building. Most of the trajectories of the 11 CoMFA sets can be assigned to either one of the types of trajectories, either the global model with

consistent SAR as seen with the steroid set (acest, cannab, flav, mao, rvtrans, thrombin), the global model from multiple local trends with consistent SAR as seen with the Selwood set (d4, ice). However, there are also trajectories revealing inconsistencies of the data as seen with 5ht3, a2a and hiv. The inconsistencies identified with the 5ht3 and the a2a dataset could be explained by irregularities of the topomeric alignment protocol. For the 5ht3 set there are structures with a pyrrolidine of a 3-ring-system composed of a central pyrazine with fused with a pyrrolidine and pyridine or benzene that is oriented to the left hand side of the plane, but for other structures this pyrrolidine is oriented to the other side of the plane. For the a2a set there is an irregular treatment of torsions adjacent to alkyne bonds, resulting in the generation of inconsistent conformations of alkyne substituents.

The inconsistencies seen in the series trajectory of the hiv set reflect the challenges of the original authors of the dataset to identify robust SAR trends within a series composed of highly symmetric next to non-symmetric structures [40].

Conclusions

We have developed a novel procedure for 3D QSAR analysis that controls the assembly of a given dataset. A heatmap display of the sorted similarity distance matrix is proposed as a guide for navigation in structural space as well as for identification of cluster centroids. In connection with a series trajectory analysis that records generation of topomer CoMFA models the heatmap allows to monitor the development of local and global trends. Our procedure has also shown to be effective in highlighting inconsistent relationships of series and was helpful in uncovering the structural details underlying the inconsistent trends.

The topomer CoMFA method has been successfully applied to 2D-QSAR as well as 3D-QSAR datasets. The only adjustable parameter for topomer CoMFA is the choice of the splitting bond. From analysis of the 16 datasets it seems that for most of the sets there is a natural choice as to which bond to select. And in fact, we have developed routines for automating this step, the selection of the splitting bond (manuscript in preparation). The results suggest that topomer CoMFA can efficiently bridge the gap between 2D- and 3D QSAR modeling.

The topomeric descriptor provides a similarity metric that is calculated mainly from the same properties as considered for developing the QSAR, the steric CoMFA field of the aligned structure. The centroid structure of a series seems to provide a point in structural space that optimally balances the different changes in steric and electrostatic field values resulting from the various structural modifications occurring within the series. The topomeric distance metric seems to be a sensitive measure of the magnitude of these changes, it can therefore be used to decide whether a compound belongs to an emerging structure-activity relationship and it can be used to control the set of test compounds for which predictions ought to be reliable.

The literature is rich in publications of retrospective QSAR studies where the main emphasis is in presentation of statistical achievements. The literature is poor in prospective publications of successful applications of QSAR such as rational design of new compounds. Rigorous model validation seems to play an important part of published QSAR and may suggest models with high predictive abilities. However, it is not sufficient when QSAR models are being applied for the sake of mining databases and suggesting new compounds for the make-and-test cycle. Rather, being able to analyse QSAR models on a structural level allows chemically intuitive interpretations and hence intensifies discussions among chemists, biologists and QSAR experts paving the way to rationalization of the lead optimization process. The workflow proposed in this publication will enhance QSAR applications in the field of drug discovery and development and will improve the success in the design of new compounds.

References

1. Unger SH, Hansch C (1973) J Med Chem 16:745
2. Kier LB, Hall LH (1976) Molecular connectivity in chemistry and drug research. Academic Press, New York
3. Tong W, Lowis DR, Perkins R, Chen Y, Welsh W, Goddette DW (1998) J Chem Inf Comput Sci 38:669-677
4. Cramer RD, Patterson DE, Bunce JD (1988) J Am Chem Soc 110:5939–5967
5. Klebe G, Abraham U, Mietzner T (1994) J Med Chem 37: 4130–4136
6. Jain AN, Koile K, Chapman D (1994) J Med Chem 37: 2315–2327
7. Kellogg GE, Semus SF, Abraham DJ (1991) J Comp Aid Mol Des 5:545–552
8. Hansch C, Unger SH, Forsythe AB (1973) J Med Chem 16: 1217–1222
9. Wooton R, Cranfield R, Sheppey GC, Goodford PJ (1975) J Med Chem 18:607–613
10. Wold S (1976) Pattern Recognit 8:127–139
11. Martin YC, Bures MG, Danaher EA, DeLazzer J, Lico I, Pavlik PA (1993) J Comput Aid Mol Des 7:83–102
12. DePriest SA, Mayer D, Naylor CB, Marshall GR (1993) J Am Chem Soc 115:5372–5384
13. Topomer technologies consists of Topomer Search, Topomer CoMFA and AllChem, available at Tripos DE, 1699 South Hanley Road, St. Louis, Missouri 63144, USA, at URL http://www.tripos.com
14. Cramer RD, Jilek RJ, Guessregen S, Clark SJ, Wendt B, Clark RD (2004) J Med Chem 47:6777–6791
15. Cramer RD, Clark RD, Patterson DE, Ferguson AM (1996) J Med Chem 39:3060–3069

16. Jilek RJ, Cramer RD (2004) J Chem Inf Comp Sci 44:1221–1227
17. Cramer RD (2003) J Med Chem 46:374–389
18. Robinson DD, Winn PJ, Lyne PD, Richards WG (1999) J Med Chem 42:573–583
19. Stiefl N, Baumann K (2003) J Med Chem 46:1390–1407
20. Peterson SD, Schaal W, Karlen A (2006) J Chem Inf Model 46:355–364
21. Selwood DL, Livingstone DJ, Comley JCW (1990) J Med Chem 33:136–142
22. Rogers RD, Hopfinger A (1994) J Chem Inf Comput Sci 34: 854–866
23. Kubinyi H (1994) Quant Struct-Act Relat 13:285–294
24. Kubinyi H (1994) Quant Struct-Act Relat 13:393–401
25. So S, Karplus M (1996) J Med Chem 39:1521–1530
26. Nicolotti O, Carotti A (2006) J Chem Inf Model 46:264–276
27. Hoffman B, Cho SJ, Zheng W, Wyrick S, Nichols DE, Mailman RB, Tropsha A (1999) J Med Chem 42:3217–3226
28. Oloff S, Mailman RB, Tropsha A (2005) J Med Chem 48: 7322–7332
29. Pandey A, Volkots DL, Seroogy JM, Rose JW, Yu J-C, Lambing JL (2002) J Med Chem 45:3772–3779
30. Khadikar PV, Shrivastava A, Agrawal VK, Srivastava S (2003) Bioorg Med Chem Lett 13:3009–3014
31. Guha R, Jurs PC (2004) J Chem Inf Comput Sci 44:2179–2189
32. Scozzafava A, Menabuoni L, Mincione F, Briganti F, Mincione G, Supuran CT (2000) J Med Chem 43:4542–4551
33. Mattioni BE, Jurs PC (2002) J Chem Inf Comput Sci 42:94–102
34. Lewis RA (2005) J Med Chem 48:1638–1648
35. Rieger JM, Brown ML, Sullivan GW, Linden J, Macdonald TL (2001) J Med Chem 44:531–539
36. Sippl W, Contreras J-M, Parrot I, Rival YM, Wermuth CG (2001) J Comp Aid Mol Des 15:395–410
37. Tetko IV, Kovalishyn VV, Livingstone DJ (2001) J Med Chem 44:2411–2420
38. Lanig H, Utz W, Gmeiner P (2001) J Med Chem 44:1151–1157
39. Huang X, Liu T, Gu J, Luo X, Ji R, Cao Y, Xue H, Wong JT-F, Wong BL, Jiang H, Chen K (2001) J Med Chem 44:1883–1891
40. Schaal WK, Ahlsen A, Lindberg G, Andersson J, Danielson HO, Classon B, Unge T, Samuellson B, Hulten J, Hallberg A, Karlen A (2001) J Med Chem 44:155–169
41. Bureau R, Daveu C, Baglin I, Santos JS-D, Lancelot J-C, Rault SJ (2001) Chem Inf Comput Sci 41:815–823
42. Kulkarni SS, Kulkarni VM (1999) J Med Chem 42:373–380
43. Gnerre C, Catto M, Leonetti F, Weber P, Carrupt P-A, Altomare C, Carotti A, Testa B (2000) J Med Chem 43:4747–4752
44. Hannongbua S, Nivesanond K, Lawtrakul L, Pungpo P, Wolschann P (2001) J Chem Inf Comput Sci 41:848–855
45. Bohm M, Sturtzebecher J, Klebe G (1999) J Med Chem 42: 458–477
46. Cramer RD, Wendt B (2007) J Comp Aid Mol Des 1:23–32
47. The program dbtop usually considers all single acyclic bonds for splitting the molecule into fragments. Topomeric CoMFA requires as input parameter a bond-ID for splitting the structures into fragments. By marking the splitting bond for each structure in the input file it is assured that both methods work on the same set of fragments
48. Tripos Bookshelf 7.3, Tripos Inc.: St. Louis, Missouri, 2007
49. Patterson DE, Cramer RD, Ferguson AM, Clark RD, Weinberger LE (1996) J Med Chem 39:3049–3059