Journal of Computer-Aided Molecular Design, **16:** 903–916, 2002.
KLUWER/ESCOM

Flexsim-R: A virtual affinity fingerprint descriptor to calculate similarities of functional groups

Alexander Weber[†,§], Andreas Teckentrup[†] & Hans Briem[†,*]
[†]*Department of Lead Discovery, Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorfer Str. 65, D-88397 Biberach, Germany;* [§]*Department of Pharmaceutical Chemistry, University of Marburg, Marbacher Weg 6, D-35032 Marburg, Germany*

MS received 26 September 2002; accepted in final form 3 February 2003

Key words: molecular descriptors, virtual affinity fingerprints, bioisosterism, molecular similarity, neighborhood behavior, 3D-QSAR, docking

Summary

Methods to describe the similarity of fragments occurring in drug-like molecules are of fundamental importance in computational drug design. In the early phase of lead discovery, they can help to select diverse building blocks for combinatorial compound libraries intended for broad screening. In lead optimization, such methods can guide bioisosteric replacements of one functional group by another or serve as descriptors for QSAR calculations.

In this paper, we outline the development of a novel 3D descriptor, termed Flexsim-R, which is a further extension of our virtual affinity fingerprint idea. Descriptors are calculated based on docking of small fragments such as building blocks for combinatorial chemistry or functional groups of drug-like molecules into a reference panel of protein binding sites.

The method is validated by examining the neighborhood behavior of the affinity fingerprints and by deriving predictive QSAR models for a couple of literature peptide data sets.

Introduction

Molecular descriptors have become indispensable tools in numerous fields of computer-assisted drug discovery, like database searching, combinatorial library design, mining of high-throughput screening data, rational compound purchasing, QSAR etc. [1].

Software packages such as DRAGON [2], SYBYL/Selector [3], TSAR[4] and CERIUS2.Descriptor+ [5] allow the rapid calculation and application of a huge number of molecular descriptors.

In former days, when structural descriptors were primarily used for database searching and QSAR, emphasis was placed on the development of 1D (whole molecule) and 2D (topological fingerprint) descriptors as implemented, for example, in the DAYLIGHT [6] or ISIS [7] software.

Based on the assumption that similar compounds should have similar binding properties at 3D binding sites, in more recent times an increasing number of 3D approaches has been described in the literature. These include the molecular shape, the molecular electrostatic potential, atom distance matrices as well as matches to given pharmacophore models [8]. Recently, we introduced virtual affinity fingerprints based on computational docking of ligands into a set of reference protein binding sites [9–11].

Most of the methods hitherto described focus on the similarity assessment of whole molecules rather than on molecular fragments. As noted by Watson et al. [12], '...the calculation of the similarity and alignment of functional groups contained within ligand molecules has so far been overlooked.' This is quite surprising, considering the fact that important applications, such as building block selection for combinatorial chemistry or the search for bioisos-

*Corresponding author. Current address: Dr. Hans Briem, Schering AG, CDCC/Computational Chemistry, Müllerstr. 178, 13342 Berlin, Germany. E-mail: hans.briem@schering.de

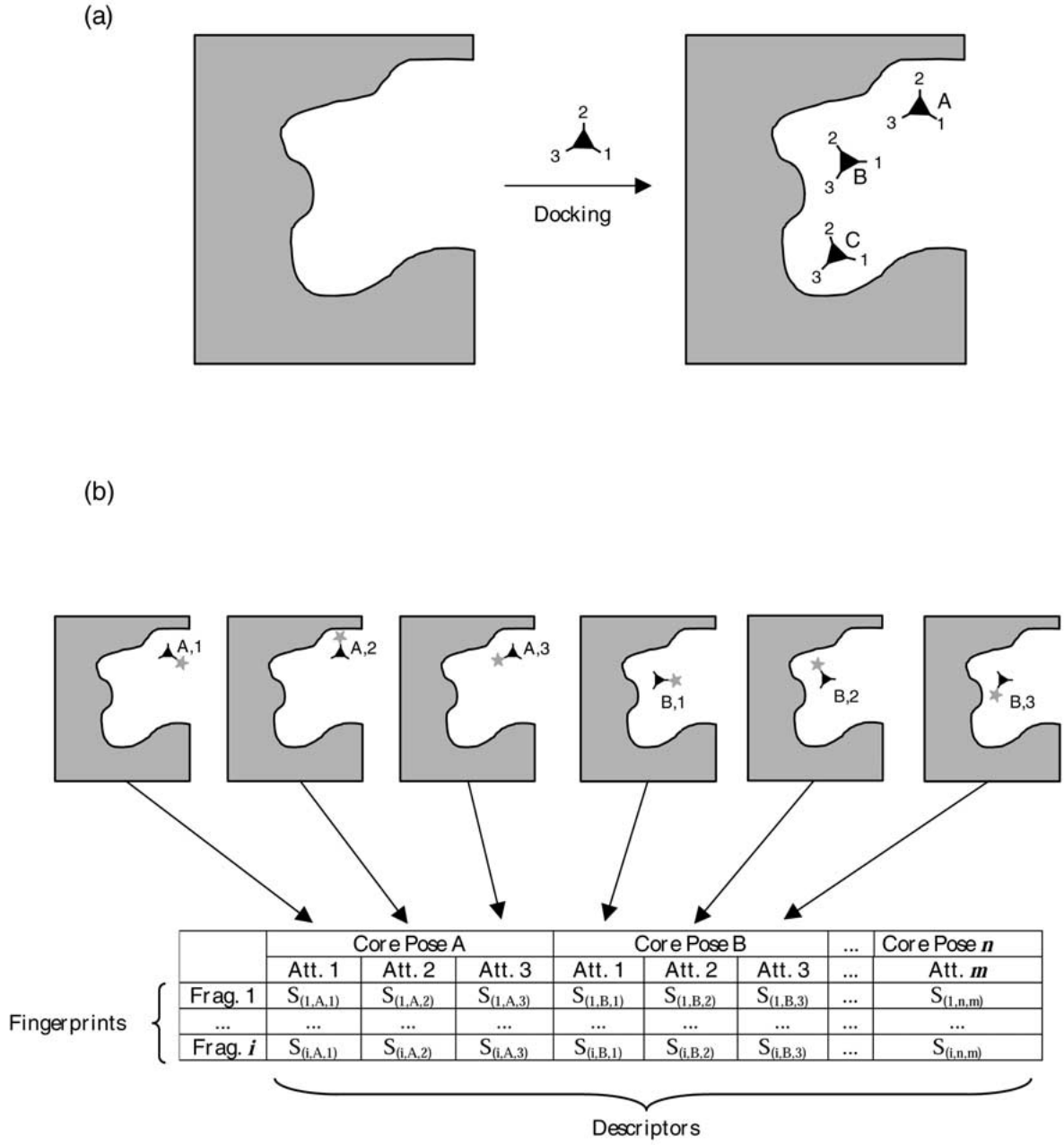


Figure 1. Schematic representation of the Flexsim-R approach. (a) First step: docking of a core molecule (black triangle) with multiple attachment points into the binding pocket of a reference protein, resulting in different 'core poses' (indicated by the letters A–C in this example). (b) Second step: attachment of a fragment (grey asterisk) to each attachment point at each core pose and subsequent constrained docking. For *i* fragments in the data set, the respective docking scores *S* are stored in a data matrix with *n* core poses and *m* attachment points. Additional descriptors can be added to the fingerprint by using additional binding pockets *k* (not shown in the diagram).

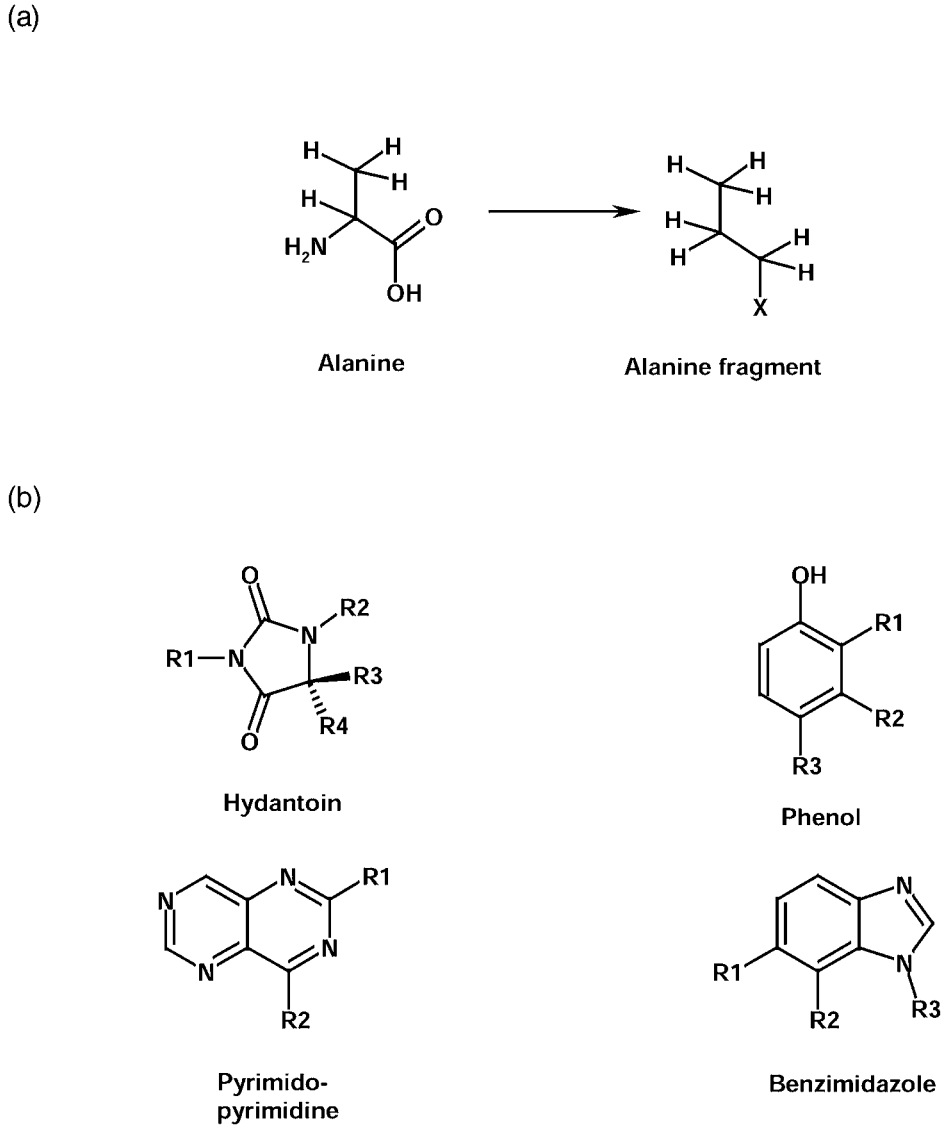


Figure 2. (a) Example of the conversion of an amino acid into a fragment used by Flexsim-R. The attachment point to the core molecule is marked by the letter X. (b) Core molecules used in this study, together with their attachment points (R).

teric functional groups, should rely more on fragment descriptors rather than on whole molecule properties.

Watson et al. proposed an elegant way to describe functional group similarities based on propensity maps calculated from the IsoStar database [13]. A similar idea was published by Anzali et al. [14], who use Kohonen neural networks to visualize and compare molecular electrostatic potentials of functional group pairs known to be bioisosteres.

Leach et al. recently introduced a method called GaP (gridding and partitioning) which tracks typical pharmacophoric features of molecular fragments within a partitioned space [15]. This 3D information is subsequently translated into a binary fingerprint which can be used to compare the fragments or to search for voids in a compound library.

Here, we describe a virtual affinity fingerprint method, termed Flexsim-R (composed of FlexX, similarity and R-groups), which is specifically designed for similarity assessments of small fragments, like R-groups of combinatorial libraries.

In our previous work [9–10], virtual affinity fingerprints were generated by unconstrained docking of whole molecules into a set of different reference

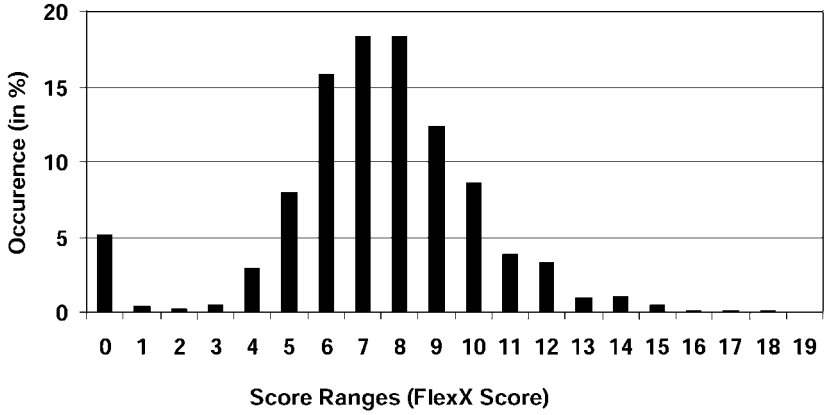


Figure 3. Occurence of FlexX score ranges (difference between lowest and highest score value for each pose) for the following parameter set: all core poses and attachment points for the hydantoin core, with the 20 natural amino acids docked into all seven reference binding pockets.

pockets. Hence, the fingerprint comprises a vector of docking scores with each vector element corresponding to exactly one reference pocket. Since this technique requires the compared molecules to largely fill the reference protein pockets, it is particularly suited for molecules of typical drug-like size.

Obviously, functional groups or building blocks can be much smaller. Unconstrained docking of such fragments into much larger binding pockets often leads to situations where even very similar fragments are placed into different sub-pockets of the protein binding niche. Consequently, fingerprint comparison of such fragment pairs results in a low similarity score.

To cope with this problem, we introduced two major modifications to our previous approach:

(1) When docking, a tether is placed on the connecting bond, i.e. the bond joining the fragment under investigation to the remainder of the molecule. For the intended applications, identifying a connecting bond should not cause any problems, since both in reagent selection for combinatorial chemistry as well as in bioisosteric fragment replacement such a common bond is typically given.
(2) To scan different areas of a protein binding niche, we generate more than one docking orientation per fragment and binding site. Accordingly, one protein contributes to several elements of the resulting virtual affinity fingerprint vector.

The method is validated by calculating virtual affinity fingerprints for the 20 proteinogenic amino acid side chains. These are subsequently used (1) to examine their neighborhood behavior and (2) as QSAR descriptors in PLS calculations on different peptide data sets from the literature.

Methods

Flexsim-R

In general, our approach works in two consecutive stages represented graphically in Figure 1:

First, a core fragment with at least one attachment point is docked into the panel of different reference binding pockets. The *n* highest scoring, non-redundant docking solutions per pocket (with e.g. $n = 50$) are stored as 'core poses'. In this way, different regions of the binding pocket are sampled, giving rise to a number of sub-pockets used for the following fingerprinting of the fragment set (Figure 1a).

It must be noted that the core fragment merely serves as a vehicle to align and tether the fragment set of interest. Instead of using orientation vectors to restrain the fragments, we had to rely on real molecules as core fragments. This is due to the fact that the version of the docking program used in this study (FlexX, version 1.9.0 [16]) does not allow tethering of single bonds only. In future studies, it might be an option to use a different docking program which does not share this limitation.

In the second step, the core poses serve as anchors for docking the fragment test set. For this purpose, each fragment is attached to the core at a suitable attachment point. Next, a docking run of each core + fragment combination, tethered to each precalculated core pose, is performed. The calculated interaction energies (docking scores) are stored in a data matrix. Thus, an *n*-dimensional virtual affinity fingerprint results from *n* core poses (Figure 1b).

In addition to the number of poses per binding site (*n*), the size of the fingerprint can be increased

by multiple (k) binding pockets and by different (m) attachment points of the core, thus giving rise to a fingerprint vector length of

$$L_{fp} = n * k * m \quad (1)$$

Ligand preparation

In this study, we examined four different core molecules whose structures were sketched within the SYBYL molecular modelling package [3]. For 2D to 3D conversion, we used the CORINA program [17]. The 20 natural amino acids were extracted from the standard fragment library of the INSIGHT 2000 molecular modelling package [18]. To be able to handle all 20 proteinogenic amino acids including glycine and proline, we converted the backbone carboxy groups into methyl and assigned one of the methyl hydrogens as attachment point. Since we primarily intended to describe the side chain properties, the backbone amino groups (except for proline) were replaced by hydrogen atoms. The acidic and basic residues (Asp, Glu, Lys, Arg) were treated as charged species with formal charges of -1 and $+1$, respectively. Histidine was treated as neutral species with the ε-proton attached. Figure 2a shows the conversion procedure exemplified by the alanine residue.

Figure 2b lists the structures of the core molecules used in this study, together with their attachment points.

Upon docking, each fragment is attached in turn to each core attachment point, i.e. no combinatorial libraries are docked! Rather, for a core with four attachment points, one particular fragment yields four different docking scores per core pose. Hence, the multiple core attachment points provide a means to arrange the fragments into different directions of the binding pocket, thus covering a broad area of space.

Preparation of the reference binding pockets

Six binding pockets from the PDB [19] as well as one model of a G-protein coupled receptor were selected for this study (Table 1).

These proteins have already been part of the reference set utilized in our previous work on virtual affinity fingerprints where we were able to show the low cross-correlation of docking scores within these pockets [9] (i.e. the selected proteins are reasonably diverse).

Prior to docking, all crystallographic water molecules, bound ligands, and counterions were removed. Binding sites were defined to encompass all residues with atoms closer than 5 Å to the nearest atom of the ligand in the PDB structure.

Docking

The docking runs were performed using version 1.9.0 of the FlexX software [16] and the FlexX$_c$ [27] option for combinatorial library docking. Two different scoring functions were used: the empirical FlexX scoring function as well as the knowledge-based DrugScore function [28].

(a) Calculation of core poses. The default options of FlexX were applied for docking the cores into the binding pocket. The placements within one binding pocket were then clustered using the CLUSTER option in FlexX with a root-mean-square deviation threshold of 2 Å. From each cluster, the energetically most favorable placement was saved as cluster representative. The 50 best representatives from each binding pocket were finally used as core poses. This procedure avoids redundancies and ensures a broad coverage of space within one pocket. For some combinations of core molecules and protein pockets, less than 50 cluster representatives could be found. For this reason and due to the different number of attachment points per core fragment, the resulting fingerprint lengths vary from core to core.

(b) Docking of fragments. Each fragment is attached in turn to each core and docked with the predetermined core poses serving as references. The PERTUBATE option in FlexX was applied, i.e. the coordinates of the core are not strictly fixed, but can rotate and translate to a small extent, allowing the molecule to slightly accommodate to the binding pockets. Default rotation and translation tolerances were applied [29].

The resulting docking scores for each of the seven reference pockets were saved in a score matrix table with the fragments as rows and the core poses as columns. The final full-length fingerprints were generated by concatenating the corresponding rows from each score matrix table (thus, k = 7 in Equation 1).

Although both the core molecule as well as the fragment 'sidechain' contribute to the total docking score, the variations of the core contributions within one particular reference pose are negligible. In other words, almost all of the variance within identical fingerprint positions can be attributed to the docking interaction energy of the fragments only.

Table 1. Reference panel of protein binding pockets

PDB code	Description	Ref.
3DFR	dihydrofolate reductase complexed with methotrexate	20
3CLA	chloramphenicol acetyltransferase complexed with chloramphenicol	21
1DWC	α-thrombin complexed with modified hirudin and argatroban	22
1EED	endothiapepsin complexed with PD125754	23
1POP	papain complexed with leupeptin	24
2TSC	thymidilate synthase complexed with dUMP and an anti-folate	25
–	model of the human 5-HT_{2a} receptor	26

It should be noted that the constrained docking procedure for some core poses may lead to steric penalties for larger fragments, thus resulting in bad docking scores. As the Flexsim-R method is based on relative score differences rather than absolute score values, these bad scores yield an important piece of information to distinguish between sterically more or less demanding moieties. In addition, there was a concern that many of the core poses point either directly into solvent or towards protein, thus allowing no score variation among the docked fragments. As can be concluded from Figure 3 though, only 5% of all poses show a score range of zero (i.e. no variation at all) within the set of the 20 natural amino acids. For the remaining 95% of core poses the corresponding score ranges exhibit a normal distribution.

Correlation analysis

To examine the impact of the choice of core molecules on the results, we performed correlation analyses of the respective affinity fingerprints. As there is no equivalence between the same position number in the fingerprints of different core fragments, a simple one-by-one correlation of the elements of the fragment fingerprints is not possible. Therefore, for each core molecule we first generated affinity fingerprints for all 20 amino acids according to Equation 1. We then calculated the respective euclidean distance matrices for all 190 possible amino acid pairs, normalized to values between 0 (most similar pair) and 100 (most dissimilar pair). Hence, as the size of these euclidean distance matrix tables is identical for each core, the Pearson correlation coefficients r can be determined.

Generation of whole-molecule descriptors

Validation tests for five peptide data sets from the literature (Table 3) were performed. For these tests, it is essential that the amino acid fingerprints generated so far have to be translated into whole-molecule descriptors. This was done by concatenating the amino acid fingerprints in the order of the respective peptide sequences. In general, for the validation runs the concatenation sequence is arbitrary, but has to be consistent throughout the data set.

Neighborhood behavior

To construct neighborhood plots, for each pair of molecules the normalized euclidean distances for the biological activity as well as for the virtual affinity fingerprints were calculated. The descriptor differences were then plotted on the x-axis and the biological differences on the y-axis. An optimal diagonal could be determined which separates the upper left triangle (ULT) of the graph from the lower right trapezoid (LRT). How this diagonal is constructed is explained in detail in ref. [35] and is not repeated here.

For a numerical analysis of the neighborhood plots, we determined the ratio of point density in the LRT to the overall point density, giving rise to an 'enhancement ratio' for each plot. The significance of this enhancement is determined by the χ^2 statistics. Again, for details of the calculations we refer to [35].

Data reduction

In order to reduce the fingerprint dimensionality while retaining as much information content as possible, we applied the following unsupervised variable selection procedure:

(1) For a given descriptor matrix of i rows and L_{fp} columns with $i = 20$ amino acids and $L_{fp} = n * k * m$ variables (see Equation 1), the squared Pearson correlation coefficients (r^2) for each pair of descriptor columns (variables) are calculated.

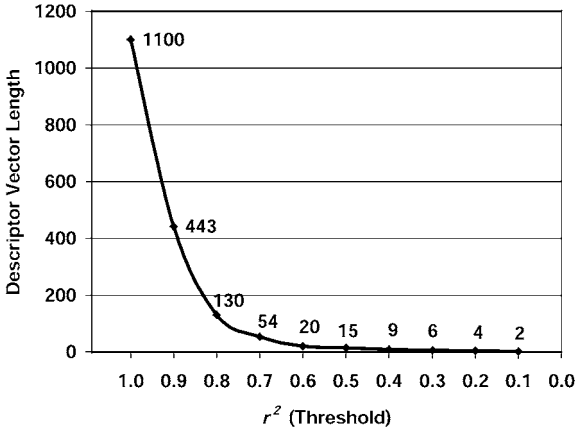


Figure 4. Impact of variable elimination on the size of the virtual affinity fingerprint. The numbers in the graph represent the fingerprint lengths according to the squared Pearson correlation coefficient threshold r^2 applied in the elimination algorithm.

(2) The list of pairs is sorted by r^2 in decreasing order. Then, starting with the most correlated pair, that variable is eliminated which can be better described by a multiple linear regression calculation of the remaining descriptors, i.e. the variable whose information content is retained within the data set to a greater extent. This procedure is repeated until the user-defined threshold for r^2 is reached.

This stepwise elimination algorithm ensures that the remaining data matrix does not contain any descriptor pair with correlation coefficients greater than the threshold. For the validation experiments, we used five different thresholds for r^2: 1.0 (i.e. the full descriptor set), 0.7, 0.5, 0.3 and 0.1.

In addition, we performed a principle component analysis (PCA) on the full descriptor matrix. The PCA implementation of the GOLPE program [36] was used for this purpose.

PLS calculations

For the five peptide data sets, PLS calculations were performed using GOLPE. For cross validation, the *leave random groups out* option in GOLPE was applied with the following parameter settings: 5 random groups per data set, 20 validation runs, maximum of 10 latent variables.

Results

Unsupervised data analysis

In order to check the impact of the choice of core fragments on the results, we compared the fingerprints resulting from four different core molecules (Figure 2b). As explained in the Methods section, we could not correlate the fingerprints directly, but rather compared the normalized euclidean distance matrices from all 190 unique amino acid pairs for each core molecule, respectively. The correlation matrix in Table 2 reveals that the distance matrices for the amino acid data set are highly correlated, i.e. the virtual affinity fingerprints are to a great extent independent from the core molecule employed.

At a first glance this might look surprising, as the four diverse core molecules certainly promote different orientations of the attached fragments upon docking. Obviously, by using a diverse set of binding pockets and various core poses per binding pocket, the fingerprint accumulates enough information to compensate for the different starting orientations.

As a consequence of the equivalence of core molecules, the subsequent validation studies were carried out with a single core structure only. We selected the hydantoin core, as it offers the highest number of different attachment points.

In order to probe the impact of variable selection on the neighborhood behavior and the PLS results, we applied a variable elimination algorithm (see Methods section). In Figure 4, the fingerprint lengths are plotted as a function of the squared Pearson correlation coefficient threshold.

The sharp decline of the curve indicates that the information contained in the full set (all attachment points of the hydantoin core, all possible core poses in all seven binding pockets) is redundant to a great extent, i.e. the full set contains a large number of highly correlated variables. We found that most of the core poses showing low score variation (i.e. the left-hand part of the score variation histogram in Figure 3) are eliminated early on (data not shown).

For the supervised validation studies, we used five different amino acid fingerprint sets:

Set 1: 1,100 descriptors (full set, correlation coefficient threshold $r^2 = 1.0$)
Set 2: 54 descriptors ($r^2 = 0.7$)
Set 3: 15 descriptors ($r^2 = 0.5$)
Set 4: 6 descriptors ($r^2 = 0.3$)
Set 5: 2 descriptors ($r^2 = 0.1$)

Table 2. Pearson correlation coefficients *r* of euclidean fingerprint distance matrices constructed with different core molecules

	Hydantoin	Phenol	Pyrimido-pyrimidine	Benzimidazole
Hydantoin	1.000			
Phenol	0.954	1.000		
Pyrimido-pyrimidine	0.971	0.963	1.000	
Benzimidazole	0.978	0.973	0.987	1.000

Supervised validation studies

To validate the usefulness of our Flexsim-R fingerprints as 3D-QSAR descriptors, we applied them to five peptide data sets from the literature. These data sets have been compiled by Matter to analyze the neighborhood behavior of different frequently used descriptors [37].

Table 3 lists some features of the data sets.

(a) Neighborhood behavior. Introduced by Patterson et al. [35], neighborhood plots are useful means to validate descriptors aimed at molecular diversity. Plotting the pairwise descriptor distances of the molecules in the data set versus the respective differences in the biological response, a useful descriptor is indicated when the point density in the upper left triangle of the graph (ULT) is much lower than in the lower right trapezoid (LRT), i.e. small changes in descriptor space (small structural differences) lead to only small differences in binding affinities. On the other hand, large structural changes do not necessarily have to – and in fact in most cases do not – correspond with large changes in biological behavior. Hence, a neighborhood plot of a 'good' descriptor is characterized by a typical shape with most data points residing in the LRT. For a detailed discussion of neighborhood plots we refer to [35] and [37].

The neighborhood plots for the peptide data sets constructed by our five sets of Flexsim-R fingerprints are shown in Figure 5.

Visual inspection as well as numerical analysis (Table 4) reveals that most data sets show a reasonable to good neighborhood behavior, as indicated by the high enhancement ratios and χ^2 values. Only the enkephalin pentapeptide set (ENK) performs rather poorly.

It is interesting to compare these results with the study performed by Matter [37] who generated neighborhood plots for the same peptide data sets and six different descriptors: We judge the quality of our Flexsim-R approach comparable or even superior to the most useful descriptor in Matter's work – the 'Wold87' descriptor [38] which is derived from experimental physicochemical parameters. The ranking of the peptide data sets is quite similar in both studies with the notable difference that the nonapeptide bradykinin set BR9 did not yield any significant neighborhood behavior with any of the descriptors investigated by Matter, whereas with Flexsim-R an at least moderate neighborhood behavior is achieved.

Furthermore, reduction of the fingerprint length by our variable elimination algorithm has no marked influence on the neighborhood behavior, i.e. the information content even in Set 4 with only 6 docking scores per amino acid is sufficient to constitute a useful descriptor for molecular diversity applications. Only Set 5 with a fingerprint length of just two descriptors performs significantly worse.

PLS studies

Figure 6 summarizes the results of the PLS calculations (q^2 values after *leave random groups out* cross validation, FlexX scoring function).

Several conclusions can be drawn:

(1) For the two dipeptide sets (ACE and BIT) as well as for one of the pentapeptide sets (BRA), significant q^2 values (>0.6) could be obtained, indicating that the Flexsim-R method calculates meaningful QSAR descriptors.

(2) In contrast, the second pentapeptide set (ENK) and in particular the nonapeptide set (BR9) did not yield significant q^2 values. This finding is in agreement with Matter's results, who was not able to find good neighborhood behavior for these two data sets with any descriptor. Obviously, the biological response for very large and flexible molecules can become too complex to be modeled by

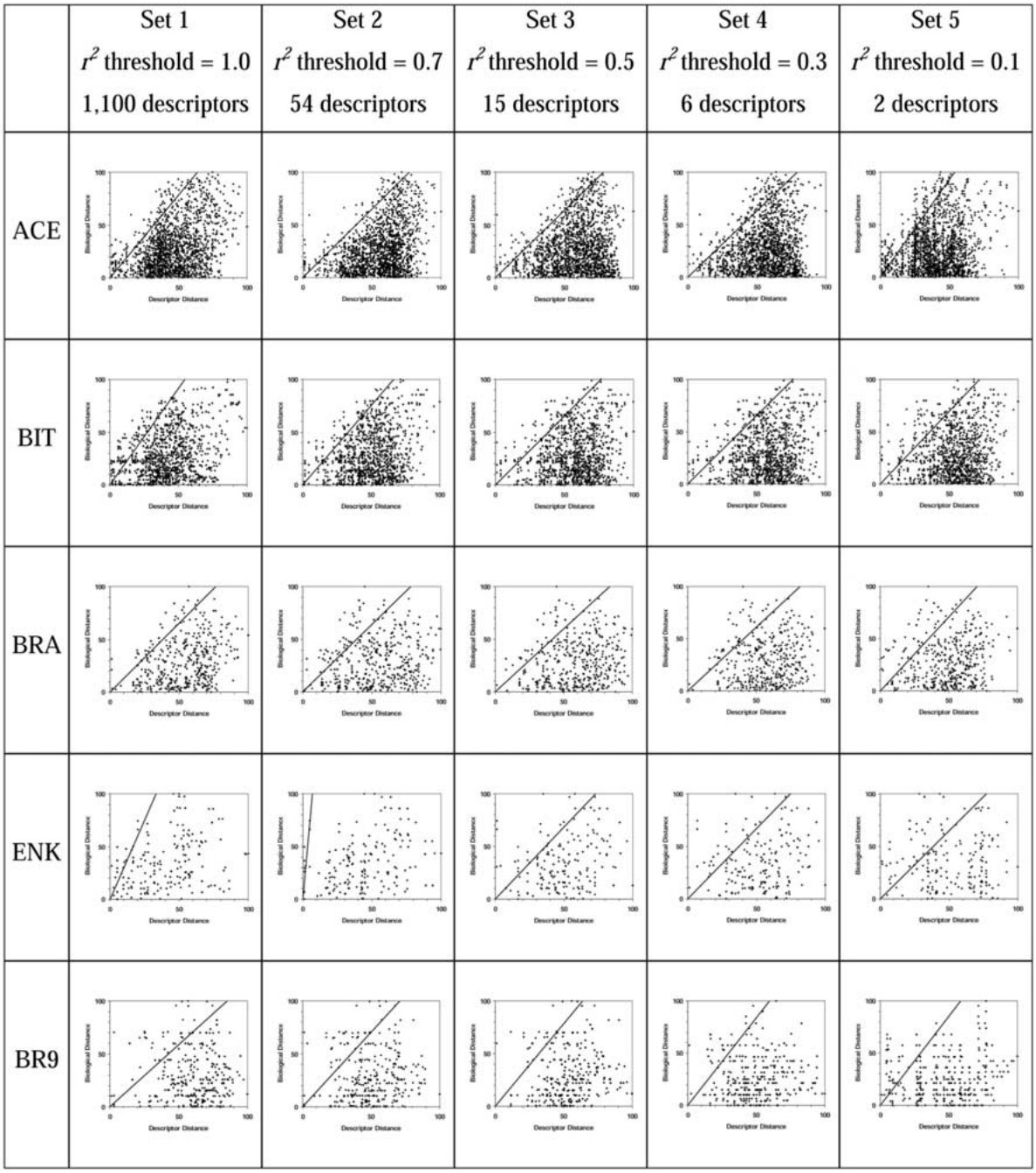


Figure 5. Neighborhood plots of the five peptide data sets (rows) and five different fingerprint lengths (columns). The diagonal line in each graph is constructed to optimize the enhancement ratio, i.e. the ratio of point density in the lower right trapezoid area (LRT) vs. point density in the upper left triangle area (ULT).

Table 3. Peptide data sets used in the validation studies

Abbreviation	Description	n[1]	Peptide length	Ref.
ACE	ACE inhibitors	58	2	30
BIT	Bitter-tasting peptides	48	2	31
BRA	Bradykinin-potentiating peptides	29	5	32
ENK	Enkephalin analogs	19	5	33
BR9	Bradykinin analogs	26	9	34

[1] n = number of compounds per data set

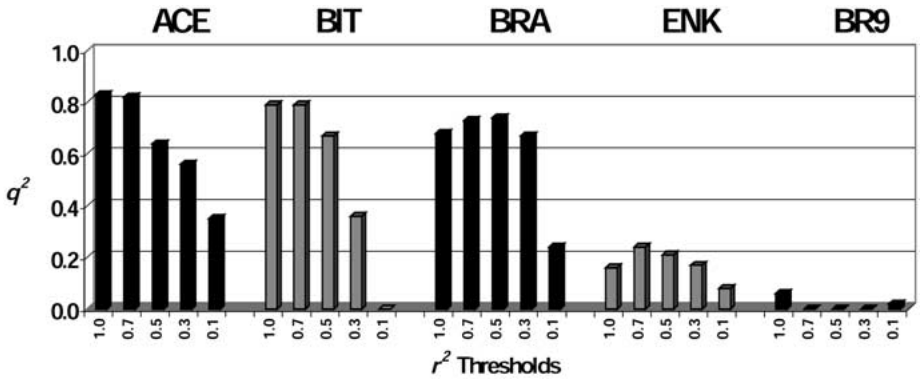


Figure 6. Results of the PLS calculations (*leave random groups out* cross-validated q^2) depending on the fingerprint size (r^2 thresholds).

any linear QSAR calculation, thus indicating a 'natural limit' for the applicability of Flexsim-R. In addition, these two data sets might be too small to achieve statistically significant QSARs.

(3) A ~20-fold reduction of the fingerprint dimensionality by our variable elimination procedure from 1,100 descriptors in Set 1 to only 54 descriptors in Set 2 results in no significant loss of predictive power. Only if we further reduce the size of the fingerprint, q^2 values get worse – though still remaining significant with as few as 6 descriptors (Set 4) for the ACE, BIT and BRA data set. (We have no good explanation for the slight increase of q^2 in the BRA data set upon reduction of the fingerprint size from Set 1 to Set 3. Possibly this increase is due to peculiarities in the PLS procedure.) This indicates that only a few non-redundant core poses are sufficient to cover most of the information in the test data sets.

By comparing the PLS results with the neighborhood behavior outlined above, one can realize a similar rank order in both validation experiments, the only exception being the BR9 data set which performed much worse in the PLS runs than anticipated by its neighborhood behavior.

Hence, we believe that good neighborhood behavior is a necessary, but not commensurate condition for a well behaved QSAR descriptor.

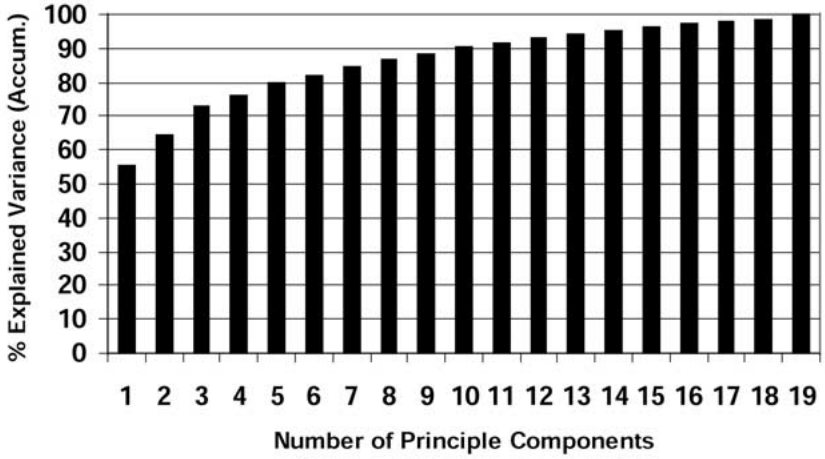
The fact that most of the information is retained by a very small number of poses raised the question whether the docking fingerprints could be replaced by a few principle components.

To answer this question, we performed a principal component analysis (PCA) on Set 1 (i.e. the full data matrix).

The plot in Figure 7a reveals that the first three principle components can explain only 73% of the variance in the data set whereas 14 principle components are required to account for 95% of the explained variance. This indicates that although most of the information content of the Flexsim-R fingerprints might be attributed to just a few straightforward factors, such as steric bulk or hydrogen bonds, there is some evidence for additional components which are hard to describe by typical pharmacophore features.

By subsequently applying the PCAs as QSAR descriptors in a PLS analysis (Figure 7b), we found that the first three PCAs yield q^2 results comparable to Set 4 (the six-element fingerprints) while the best 14 PCAs are sufficient to replace the full fingerprint matrix (Set 1).

(a)



(b)

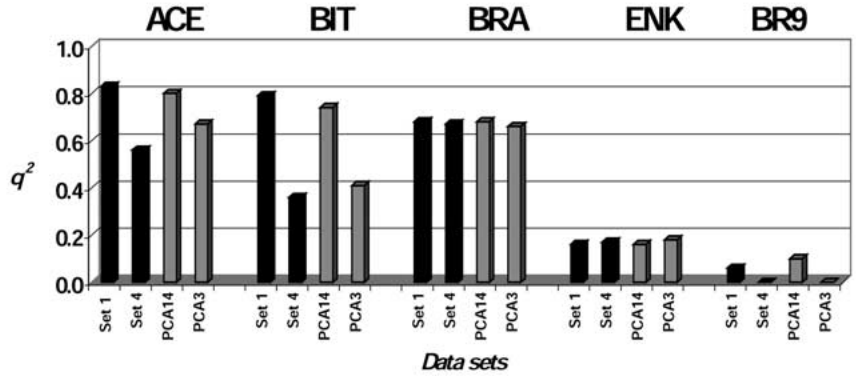


Figure 7. (a) Results of a principle component analysis (PCA) on Set 1. The number of principle components is plotted versus the explained variance (accumulated). (b) Comparison of the PLS results for Set 1 and Set 4 and two PCA data sets (best 3 and best 14 principle components).

Table 4. Neighborhood enhancement results for the five peptide data sets. (χ^2 ratios > 3.84 are significant at the P > 0.95 level and one degree of freedom)

Data Set (r^2 threshold)		Enhancement Ratio	χ^2 Ratio
ACE	Set 1 (1.0)	1.39	172.0
	Set 2 (0.7)	1.56	320.0
	Set 3 (0.5)	1.57	321.2
	Set 4 (0.3)	1.55	301.7
	Set 5 (0.1)	1.25	76.6
BIT	Set 1 (1.0)	1.25	53.1
	Set 2 (0.7)	1.41	125.6
	Set 3 (0.5)	1.53	193.1
	Set 4 (0.3)	1.45	145.3
	Set 5 (0.1)	1.18	25.7
BRA	Set 1 (1.0)	1.51	65.5
	Set 2 (0.7)	1.47	53.5
	Set 3 (0.5)	1.57	78.0
	Set 4 (0.3)	1.54	69.7
	Set 5 (0.1)	1.38	37.5
ENK	Set 1 (1.0)	1.14	2.7
	Set 2 (0.7)	1.02	0.0
	Set 3 (0.5)	1.39	16.8
	Set 4 (0.3)	1.33	12.0
	Set 5 (0.1)	1.27	8.0
BR9	Set 1 (1.0)	1.55	55.6
	Set 2 (0.7)	1.37	28.9
	Set 3 (0.5)	1.37	31.0
	Set 4 (0.3)	1.32	25.8
	Set 5 (0.1)	1.27	18.1

Finally, we investigated how much the results depend on the scoring function used for docking. Therefore, we generated Flexsim-R fingerprints by re-docking the amino acids with the knowledge-based DrugScore [28] function instead of the standard FlexX scoring scheme.

As exemplified in Figure 8, both DrugScore as well as FlexX score yield almost identical PLS results, indicating that the underlying information being picked up is covered by different scoring methods. This finding was not a big surprise to us. Although both scoring functions are derived in a completely different manner (empirically in the case of FlexX score, knowledge-based in the case of DrugScore), it has been shown in numerous applications that both methods are well suited to reflect basic principles of protein-ligand interactions. On the other hand, factors which are not well covered yet by current scoring functions and which hamper an exact estimation of free energies of binding (e.g. treatment of entropy, protein flexibility, water-mediated interactions etc.) may not be of great importance for our given problem of describing fragment similarities.

Discussion and conclusions

The basic idea of the Flexsim-R method can be described as exposing molecular fragments to different binding situations by broadly scanning a reference set of diverse protein binding pockets. Literally, each different protein environment 'visited' by the fragment enables it to 'learn', i.e. the fingerprint expresses the fragment's distinctive character, based on its steric and electrostatic properties. As such, this method is an extension of the (virtual) affinity fingerprint idea which treats molecular similarities from the perspective of protein binding sites rather than from properties derived merely from the ligand structure. In our previous work [11], we have already shown that virtual affinity fingerprints can be viewed as complementary to standard similarity descriptors (e.g. topological fingerprints). Therefore, we believe that Flexsim-R is able to capture resemblances which are not necessarily obvious by just viewing the 2D structures. This will be further investigated in a follow-up study.

We have demonstrated the usefulness of Flexsim-R fingerprints in two validation experiments: From the neighborhood enhancement plots, one can deduce that pairs of molecules with similar fingerprints – indicated by low euclidean distances – mostly show only small differences in their biological response as well. This feature qualifies the virtual affinity fingerprints for bioisosteric replacement experiments where one typically wants to substitute functional groups of lead compounds, e.g. to enhance physicochemical or pharmacokinetic properties, while retaining binding affinity.

The positive results of the PLS calculations for three of the five data sets can be regarded as an even more stringent proof, as these calculations not only focus on similarity, but rather seek a linear relationship between the descriptors and the biological response.

We were able to demonstrate that very few highly uncorrelated protein sub-pockets are sufficient to generate useful affinity fingerprint descriptors, comparable to the first few principle components of a PCA. Nevertheless, our variable selection procedure offers

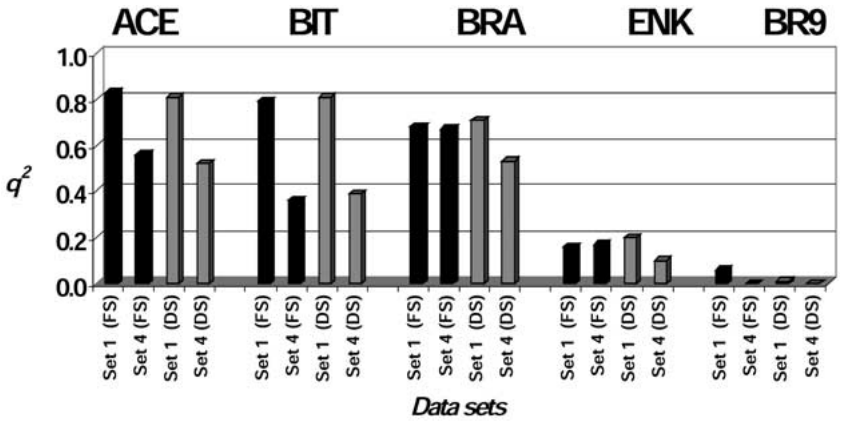


Figure 8. Comparison of PLS calculations using two different scoring functions: FlexX score (FS) and DrugScore (DS). For all five data sets, the q^2 results are very similar. For clarity, only the q^2 values for Set 1 and Set 4 are depicted.

a great advantage over using data set specific principle components: we see a good chance that the sub-pockets identified by the 20 amino acid data set might be used successfully for other fragment sets as well. This will be investigated in a future study. If such a few 'principle sub-pockets' could be identified, calculation of the Flexsim-R fingerprints for other data sets would be an extremely fast procedure.

Several advantages of Flexsim-R over existing techniques can be envisaged:

Compared to the work of Watson et al. [12] who derived functional group similarities from experimentally observed preferences of ligand-protein interactions, no such knowledge-base is required for Flexsim-R. Therefore, our approach can be applied to any fragment and is not limited to functional groups present in the PDB or contained in databases of bioisosteres, such as Bioster [39].

Furthermore, in order to compare 3D molecules a proper alignment technique is required. In our approach, alignment is inherent to the method as all fragments 'grow' from the same attachment point at the fixed core molecule, but are free to adapt to the protein environment upon flexible docking.

Flexsim-R can easily be automated since only 'clipped' fragments, i.e. functional groups containing exactly one attachment point, are required. Thus, the method is ideally suited for the similarity assessment of building blocks in combinatorial library design.

For libraries intended for broad screening in many high-throughput assays, Flexsim-R can be used to maximize the diversity of the library. On the other hand, virtual affinity fingerprints can facilitate the lead-optimization process where bioisosteric replacement of one fragment by a similar one is an important task.

Our method in some aspects resembles the GaP procedure published by Leach et al. [15] (see introductory section): both approaches start by aligning the fragment set onto the common attachment point and the adjacent bond. Then, 3D information is picked up and stored in a numeric vector (fingerprint). The fundamental difference of the two methods lies in the way of assembling the 3D descriptors. In GaP, the location of an explicit set of pharmacophoric features is tracked, whereas Flexsim-R descriptors are based on a more implicit pharmacophore description. It would certainly be interesting to judge the quality of both methods in a comparative study.

Our future work will mainly concentrate on validation studies for non-peptidic data sets. In particular, as discussed above, we want to check how well the most informative poses found for amino acids can be transferred to other groups of fragments.

Furthermore, it is conceivable to use virtual affinity fingerprints to characterize and compare protein binding pockets instead of fragments of ligands. This can be achieved by inverting the score matrix, i.e. a fingerprint for a binding pocket is composed of docking scores from a set of different fragments. As Flexsim-R fingerprints focus on small protein subpockets rather than a complete ligand binding site, we believe that such an approach could complement recently published methods to assess the similarity of protein pockets [40, 41].

Acknowledgements

This work was performed as part of the RELIMO project, which is funded by the German Federal Ministry for Education, Science, Research and Technology (BMBF) under grant No. 0311623.

The authors wish to thank Uta Lessel, Sandra Handschuh, Dirk Gardzella and Herbert Köppen from Boehringer Ingelheim Pharma GmbH & Co. KG for their contributions to this work, and Christoph Sotriffer from the University of Marburg for critical assessment of the manuscript.

We are indepted to Hans Matter from Aventis Pharma for providing the peptide data sets and for helpful discussions.

References

1. Todeschini, R. and Consonni, V., Handbook of Molecular Descriptors, In Mannhold, R., Kubinyi, H. and Timmerman, H. (Eds.) Methods and Principles in Medicinal Chemistry, Wiley-VCH, Weinheim, Germany, 2000.
2. DRAGON, Molecular Modeling Software, v. 2.1, Talete srl, Milan, Italy.
3. SYBYL, Molecular Modeling Software, v. 6.7, Tripos Associates, St. Louis, MO.
4. TSAR, v.3.3, Accelrys Inc., San Diego, CA.
5. CERIUS2, v. 4.7, Accelrys Inc., San Diego, CA.
6. Daylight Chemical Information Systems Inc., Mission Viejo, CA.
7. ISIS, MDL Information Systems Inc., San Leandro, CA.
8. Dean, P.M. (Ed.) Molecular Similarity in Drug Design, Blackie Academic & Professionals, London, U.K., 1995.
9. Briem, H. and Kuntz, I.D., J. Med. Chem. 39 (1996) 3401.
10. Lessel, U.F. and Briem, H., J. Chem. Inf. Comput. Sci., 40 (2000) 246.
11. Briem, H. and Lessel, U.F., Perspect. Drug Discov. Design, 20 (2000) 231.
12. Watson, P., Willett, P., Gillet, V.J. and Verdonk, M.L., J. Comput. Aided Mol. Des., 15 (2001) 835.
13. Bruno, I.J., Cole, J.C., Lommerse, J.P.M., Rowland, R.S., Taylor, R. and Verdonk, M.L., J. Comput. Aided Mol. Des., 11 (1997) 525.
14. Anzali, S., Gasteiger, J., Holzgrabe, U., Polanski, J, Sadowski, J., Teckentrup, A. and Wagener, M., Perspect. Drug Discov. Design, 9/10/11 (1998) 273.
15. Leach, A.R., Green, D.V.S., Hann, M.M., Judd, D.B. and Good, A.C., J. Chem. Inf. Comput. Sci., 40 (2000) 1262.
16. Rarey, M., Kramer, B., Lengauer, T. and Klebe, G., J. Mol. Biol., 261 (1996) 470.
17. Sadowski, J. and Gasteiger, J., Chem. Rev., 93 (1993) 2567.
18. INSIGHT 2000, Molecular Modeling Software, Accelrys Inc., San Diego, CA.
19. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., Nucleic Acids Res., 28 (2000) 235.
20. Filman, D.J., Bolin, J.T., Matthews, D.A. and Kraut, J., J. Biol. Chem., 257 (1982) 13663.
21. Leslie, A.G.W., J. Mol. Biol., 213 (1990) 167.
22. Banner, D.W. and Hadvary, P., J. Biol. Chem., 266 (1991) 20085.
23. Cooper, J.B., Quail, W., Frazao, C., Foundling, S.I., Blundell, T.L., Humblet, C., Lunney, E.A., Lowther, W.T. and Dunn, B.M., Biochemistry, 31 (1992) 8142.
24. Kamphuis, I.G., Kalk, K.H., Swarte, M.B.A. and Drenth, J., J. Mol. Biol., 179 (1984) 233.
25. Montfort, W.R., Perry, K.M., Faumann, E.B., Finer-Moore, J.S., Maley, G.F., Hardy, L., Maley, F. and Stroud, R.M., Biochemistry, 29 (1990) 6964.
26. Oliveira, L., Paiva, A.C.M. and Vriend. G., J. Comput. Aided Mol. Des., 7 (1993) 649. The structure of the human 5-HT_{2a} receptor model has been obtained from http://www.gpcr.org/7tm/seq/all/5HT2A_HUMAN.SW.html.
27. Rarey, M. and Lengauer T., Perspect. Drug Discov. Design, 20 (2000) 63.
28. Gohlke H., Hendlich, M. and Klebe, G., J. Mol. Biol., 295(2) (2000) 337.
29. FlexX manual, available online from: http://cartan.gmd.de/FlexX/export/flexx_ug.pdf.
30. Cushman, D.W., Cheung, H.-S., Sabo, E.F. and Ondetti, M.A., In Horowitz, Z.P. (Ed.), Proceedings of the A.N. Richards Symposium, Urban & Schwarzenberg, Baltimore, MD, 1980, pp. 3-25.
31. Asao, M., Iwamura, H., Akamatsu, M. and Fujita, T., J. Med. Chem., 30 (1987) 1873.
32. Ufkes, J.G.R., Visser, B.J., Heuver, G., and van der Meer, C., Eur. J. Pharmacol., 50 (1978) 119; Ufkes, J.G.R., Visser, B.J., Heuver, G., Wynne, H.J. and van der Meer, C., Eur. J. Pharmacol., 79 (1982) 155.
33. Morley, J.S., Ann. Rev. Pharmacol. Toxicol., 20 (1980) 81
34. Schröder, E., In Erdös, E.G. (Ed.), Handbuch der experimentellen Pharmakologie Vol. 25, Chap. 6.
35. Patterson, D.E., Cramer, R.D., Ferguson, A.M., Clark, R.D. and Weinberger, L.E., J. Med. Chem., 39 (1996) 3049.
36. GOLPE v. 4.5, Multivariate Informetric Analysis srl, Perugia, Italy.
37. Matter, H., J. Peptide Res., 52 (1998) 305.
38. Hellberg, S., Sjöström, M., Skagerberg, B. and Wold, S., J. Med. Chem., 30 (1987) 1126.
39. Bioster database, Accelrys Inc., San Diego, CA.
40. Rosen, M., Liang, S.L., Wolfson, H. and Nussinov, R., J. Mol. Biol., 11 (1998) 263.
41. Schmitt, S., Kuhn, D. and Klebe, G., J. Mol. Biol., 323 (2002) 387.