J Comput Aided Mol Des (2006) 20:751–762
DOI 10.1007/s10822-006-9085-8

ORIGINAL PAPER

Feature-map vectors: a new class of informative descriptors for computational drug discovery

Gregory A. Landrum · Julie E. Penzotti · Santosh Putta

Received: 7 June 2006 / Accepted: 30 September 2006 / Published online: 5 January 2007

Abstract In order to develop robust machine-learning or statistical models for predicting biological activity, descriptors that capture the essence of the protein–ligand interaction are required. In the absence of structural information from X-ray or NMR experiments, deriving informative descriptors can be difficult. We have developed feature-map vectors (FMVs), a new class of descriptors based on chemical features, to address this challenge. FMVs, which are derived from the conformational models of a few actives, are low dimensional, problem specific, and highly interpretable. By using shape-based alignments and scoring with chemical features, FMVs can combine information about a molecule's shape and the pharmacophores it can match. In five validation studies, bag classifiers built using FMVs have shown high enrichments for identifying actives for five diverse targets: CDK2, 5-HT_3, DHFR, thrombin, and ACE. The interpretability of these descriptors has been demonstrated for CDK2 and 5-HT_3, where the method automatically discovers the standard literature pharmacophore.

Keywords Chemical features · Descriptor · Machine learning · Molecular shape · Pharmacophores

Introduction

Many computational approaches to aid drug discovery utilize three-dimensional arrangements of chemical features that are intended to capture the essential interactions between a potential drug and the binding site of a drug target. These chemical features can either be expressed as a continuous field, as in methods such as CoMFA and CoMSIA [1, 2], or as discrete points, as in pharmacophore or feature-map based methods [3–6]. These maps of conserved chemical features are often derived from a set of structurally aligned compounds, as implemented in programs such as Catalyst [7], MOE [8], and FlexS [9]. New compounds can be aligned to these maps to evaluate their potential for activity. However, the measure of fit for a new compound is typically reduced to a single number, such as its CoMFA score, an RMSD for a pharmacophore alignment, or a calculated overlap with (or similarity to) the feature map [5, 10]. This conglomerate score is useful for rank-ordering compound alignments, but it discards much of the detailed information provided by the feature map. In many cases the feature-map score alone does not provide enough information to develop a robust and successful model to predict activity. More flexible algorithms, such as machine-learning or statistical methodologies, are needed for model building, and these techniques require informative descriptors in order to be successful.

Molecular docking algorithms, another well-established tool used to screen compounds for their

Electronic Supplementary material Supplementary material is available in the online version of this article at http://dx.doi.org/10.1007/s10822-006-9085-8 and is accessible for authorized users.

G. A. Landrum (✉)
Novartis Institutes For BioMedical Research,
CH-4002 Basel, Switzerland
e-mail: gregory.landrum@novartis.com

J. E. Penzotti
5731, 16th Ave NE, Seattle, WA 98105, USA

S. Putta
Telik Inc., 3165 Porter Drive, Palo Alto, CA 94304, USA

potential to exhibit a desired activity, generally measure the quality of each pose using scoring functions. The development of scoring functions capable of accurately and consistently ranking docked compounds has proven problematic [11], so several groups have combined standard docking methods with statistical or machine-learning methods as a strategy to improve scoring [12–15]. Springer et al. combined terms from docking scoring schemes with novel descriptors to build a random forest classifier that provided a significant enrichment of true binders versus decoy poses when compared to using the docking score alone. Other studies have demonstrated that using a naïve Bayes classifier and 2D fingerprints can improve upon the compound rankings from multiple docking algorithms [12, 14].

In this paper we answer the question: can machine-learning methodologies applied to informative descriptors improve the predictivity of molecular alignments? We present a new class of descriptors derived from molecular alignments, feature-map vectors (FMVs), that capture all available information about a compound's overlap with a feature map. We also describe an algorithm for calculating FMVs that, in contrast to docking, requires no structural information about the target protein. FMVs are simply defined as the vector of numbers containing the contributions of each point in the feature map to an alignment's overall feature-map score (Fig. 1). The sum of all elements in the FMV for a compound is equal, by definition, to its feature-map score. The scoring method, which uses the overlap between Gaussian functions located on feature points in the feature map and on the aligned molecule, is described in detail in the Methods section. In this contribution we use a recently published conformation-mining algorithm to align compounds and to construct feature maps [16]. However, any molecular alignment method can be used to derive FMV descriptors.

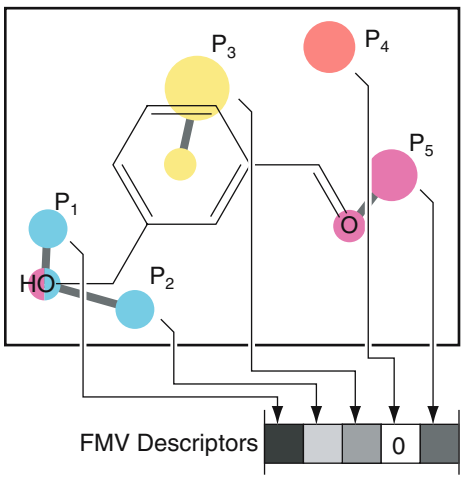

Fig. 1 Sketch illustrating the definition of a feature-map vector (FMV) for a feature map with five points. The large circles labeled P1 through P5 indicate chemical features in the map, color coded by type. The smaller spheres indicate the chemical features in the structure aligned to the feature map also color-coded by feature type. Thick gray lines connect features in the molecule and the feature points in the map that they score against. The red P4 feature has no corresponding feature in the compound and thus the FMV value corresponding to this feature is 0

Feature-map vectors have a number of traits that make them very useful. Perhaps one of the most important of these is that *FMVs are directly interpretable*—an individual element in the vector corresponds to the contribution of a given point in the feature map to a compound's score. So by examining large values in an active compound's FMV we can easily identify where that compound's chemical features align well with the feature map of interest. This is particularly useful because *feature-map vectors are problem specific*: like a three-dimensional pharmacophore, FMVs contain information about a particular target's binding site. Using a set of descriptors that is tuned to a particular problem provides significant advantages in predictive model building. The specificity of the FMV descriptors allows us to use a comparatively small number of feature points, which means that *FMVs are low dimensional*. This eliminates the need to do feature-set selection on our descriptors and decreases the likelihood that model-building algorithms will overfit to the data set. Their low dimensionality also allows the calculation of meaningful pairwise distances, making FMVs suitable for use in similarity-based methods. Finally, *FMVs are derived directly from the process of scoring a molecular alignment to a feature map* and therefore do not require any additional computation. However, the speed and efficiency of the alignment algorithm used determines the utility of this approach for virtual screening of very large databases. To avoid the expense of computing alignments, Renner and Schneider transformed similar Gaussian-based pharmacophore models into a correlation vector of probabilities for use in rapid database screening [5]. While this approach proved useful for ranking compound databases and allowed scaffold hopping, it requires the optimization of a large number parameters and provides results that are less suited for building predictive models and interpreting biological data. In contrast, FMVs, which require explicit compound alignments and thus can be more computationally demanding, are simple but powerful descriptors to use for modeling. In the present study, we describe the application of FMVs to five diverse pharmaceutical targets: CDK2, 5-HT_3, ACE, DHFR, and thrombin.

Methods

Generating the feature-map vector for a probe compound requires a reference feature map to score against, a method for aligning the probe to that feature map, and, when the probe has multiple conformations, a way to identify the conformation which should be used to calculate the FMV descriptors. In each of the validation studies reported here a recently published algorithm, conformation mining, was used both to generate a reference feature map from a small number of diverse actives and to identify the top-scoring conformation for FMV generation [16]. After providing an overview of our validation protocol in the next section, conformation mining and the subshape-alignment algorithm it uses [10] are briefly summarized in the following two subsections. In the fourth subsection, further details of the calculation of FMVs are presented. The final two subsections contain an overview of the methods used to build predictive models and validate them using literature datasets

Validation protocol

In order to validate the performance of FMVs, the following protocol was followed for each data set:

1. A set of 3–4 reference actives was chosen to define the feature map. In the examples where crystal structure data were available, actives were chosen for which bound-ligand crystal structures are available; these structures were used to compare the performance of FMVs generated via conformation mining and those generated using crystal data.
2. Conformation mining was applied to the reference actives in order to generate an alignment reference and feature map.
3. FMVs were generated for the remaining actives and inactives using the alignment reference and feature map from step 2 and the subshape alignment algorithm.
4. The FMVs from step 3 were used as inputs to either a machine learning algorithm or for similarity picking.

Subshape alignment and feature-map scoring

The subshape-alignment algorithm, which was developed to enable molecules of different sizes to be aligned to each other, was used for molecular alignment in this work [10]. The subshape methodology, in addition to allowing molecule-to-molecule comparisons, provides a means of computing multi-molecule alignments: a probe molecule can be aligned to a collection of molecules by using their combined shape (the union of their individual shapes); the alignment algorithm has no difficulty finding the steric similarities between these large combined shapes and smaller probe molecules.

The subshape method provides alignments between conformations. *Feature maps*—collections of chemical feature points in three-dimensional space—are then used to score these alignments [10, 16]. Each point in the feature map is represented by a Gaussian function whose position is derived from the conformations used to generate the target shape. The score for an alignment of a probe conformation is computed by summing the overlaps of Gaussian functions centered at each of the probe's feature locations with the Gaussians of a similar feature type in the map. A standard set of six feature types was used: positive and negative ionizable groups, hydrogen-bond donors and acceptors, aromatic groups, and, for the ACE example, zinc binders [7].

Just as shapes can be combined to provide multi-molecular alignments, a combined feature map can be constructed from the feature maps of a set of aligned conformations to give information about feature overlap with multiple molecules. Two feature maps are combined by increasing the weights of the Gaussians at each feature point based on the distance to the closest feature point of the same type in the second map. Feature points of the same type that lie within a cutoff distance are coalesced to form a single new point with an averaged weight and location [16].

Conformation mining

When working with active compounds which have co-crystal structures available, constructing an alignment reference and combined feature map for use with subshape alignment is straightforward by using protein-based alignment techniques. However, when structural information is not known, a method is needed to identify conformations of the actives that are close to the binding conformation. Our recently published conformation-mining algorithm does this by discovering and exploiting the common steric and chemical features of a set of active compounds, each of which has multiple conformations [16]. Conformation mining sifts through the conformations for each compound looking for such steric and feature commonalities using subshape alignments scored with feature maps. The top-scoring alignment of a compound's conformations is used to generate FMVs.

Generating feature-map vectors

In general terms, the feature-map vector for an aligned conformer is generated by looping over each feature in the feature map and calculating its overlap with each equivalent feature in the aligned conformer (Fig. 1). Overlaps are calculated using weighted Gaussian spheres located at each feature point (Eq. 1)

$$\mathrm{FMV}_i = w_i \cdot \sum_j \exp\left(-\frac{d_{ij}^2}{\sigma_i^2}\right) \quad (1)$$

where the loop is over all features in the aligned conformation of the same type as point i in the feature map, w_i and σ_i are the Gaussian weight and width parameters of feature-map point i, and d_{ij} is the Euclidean distance between feature-map point i and feature j in the aligned conformation. When training machine-learning models, the total feature-map score is added as an extra element to the FMV.

Feature-map vectors can be generated using alignments from any alignment algorithm. In the validation studies reported here, conformation-mining is used to select and align the best conformation (as measured by total feature-map score) from the full set of each molecule's CONAN [17] conformations (generated with a target number of 100 conformations per stereoisomer). Although the computation of FMVs themselves takes almost no time, the computational expense of the subshape-alignment algorithm used in conformation mining introduces practical limits on the size of the dataset that can be examined using this exact methodology without the use of a cluster or grid. For example, although generating FMVs for the 752 5-HT_3 actives used in this study took less than 5 s on a Linux workstation (2.8 GHz Pentium D CPU, 1 GB RAM), the subshape alignments required 1020 s. This total runtime (~1.3 s/compound) is not prohibitively high, but it would require multiple computers to do large-scale virtual library screening.

Conformational analysis, subshape alignment, conformation mining, and generation of feature-map vectors was carried out using the CombiCode software available from Deltagen Research Labs.

Machine-learning methodology

Algorithms for building and testing models

A large body of literature exists demonstrating the power of using ensembles of individual predictive models when learning from real-world data [18–22]. Ensembles have proven to be very flexible and highly resistant to overfitting. In this study we used bagging, a well-established approach based on building ensembles of decision trees for classification [23]. Each decision tree in the bag classifier was built from a local training set—a random subset of the full data set. The individual trees were constructed using the ID3 algorithm [19] with an automated process to quantize the descriptor values [24, 25]. When choosing from sets of compounds, predicted actives were ranked based on their vote margin according to the classifier (e.g. the fraction of trees predicting the compound to be active).

In addition to providing great flexibility and a resistance to overfitting, one significant advantage of ensemble predictive models is that their performance can be tested without the use of a holdout set. By using out-of-bag error estimation, it is possible to obtain good estimates of a model's accuracy based solely on the training data set [26]. Out-of-bag error estimation takes advantage of the fact that each decision tree in the bag classifier is built using a subset of the data in the training set: when a point in the training set is screened, only those trees that were not trained using that point are allowed to vote on its activity. In effect, the combination of bagging and out-of-bag error estimation allows us to use the entire data set as both the training and holdout sets.

Statistics and shuffle tests

Because building a bag classifier involves a random process—the formation of random subsets of the data to build the individual models—any assessment of model performance should be carried out multiple times to ensure that the results are statistically valid. To this end, all experiments involving predictive models in this work were repeated ten times; results are reported as means with 90% confidence levels.

In order to measure the potential impact of overfitting on the results, we used a series of shuffle tests. In these experiments, the activity values in the data set are randomly permuted and the model-building and testing processes are repeated with the same parameters used to build the original models. These shuffle tests, where all physical connection between the descriptors and activity values is removed, establish baseline values for what might be expected from models that have memorized the data.

Validation data

The application of computational approaches to identify potential actives from synthetic datasets such as those used in this work for 5-HT_3, thrombin, and ACE requires a great deal of caution to ensure that the results are statistically valid and scientifically reasonable. A recent study has beautifully highlighted this risk by demonstrating that descriptors based solely on atom counts, containing no structural information, can perform as well as or better than more complex similarity methods such as DOCKSIM and Unity fingerprints [27]. We control for this risk by comparing the results from bag classifiers built using FMV descriptors to two different similarity approaches for selecting actives: 2D similarity using Tanimoto similarities between Daylight-like topological fingerprints and FMV similarity using Euclidean distances between individual feature-map vectors (neglecting the total feature-map score when calculating the distance). When computing similarity of an individual compound to multiple reference compounds using either similarity metric, its highest similarity is used.

While we follow established tradition and report enrichment factors relative to random picking, it is also imperative to compare the performance of our computationally expensive three-dimensional method to that which is possible using only topological similarity. The ACE results, below, provide an illustration of why this comparison is important: 2D similarity picking performs as well as the models built using FMV descriptors.

There are numerous conventions for picking points along the enrichment curve to assess performance, with earlier stopping points usually leading to better enrichment values. In the interest of simplicity, in this work we report enrichments at the ''halfway point''—where half of the actives in the dataset have been picked. The halfway point provides a reasonable measure of how well a method is doing at selecting actives without exaggerating the results (as earlier stopping points can).

Results

We performed a number of experiments to validate and establish the utility of feature-map vectors for building predictive models for biological activity. These experiments included three targets for which X-ray crystal structure information is available for the ligands used to build the feature maps (CDK2, DHFR, and thrombin), one target for which limited structural information is available (ACE), and one target without structural data (5-HT_3). The CDK2 and thrombin ligand sets were used in our previous work to validate the conformation mining algorithm [16].

CDK2

Cyclin-dependent kinases (CDKs) act as regulators of the progression of eukaryotic cells through their cell division cycle [28, 29]. This connection to cell division has made CDKs natural targets for proliferative diseases such as cancer. One member of the CDK family, CDK2, has also shown promise as a target for preventing chemotherapy-induced alopecia (hair loss) [30]. CDK2, a target with a number of X-ray crystal structures and a large amount of experimental data available in the literature, is a natural starting place for our validation of FMVs.

In order to generate the conformation-mining alignment reference and feature map, three diverse CDK2 ligands were selected from a collection of 11 aligned crystal structures. The compounds chosen, which were also used in our earlier conformation mining study [16], are shown in Table 1 and Fig. 2.

The pool of compounds used for model building and testing were taken from a dataset of more than 16,000 compounds published in 2003 [33]. This dataset includes two main classes of compounds: a diverse set of screening library compounds (marked ''DivScreen'') and a collection of compounds that were either

Table 1 Comparison of the conformational models of the CDK2 ligands and their crystal conformers

Molecule	Number of Conan conformations	Closest (Å)	Farthest (Å)	Reference
1FVV	17	0.796	2.59	[30]
1OIR	94	0.66	3.94	[31]
1H01	90	1.08	4.27	[32]

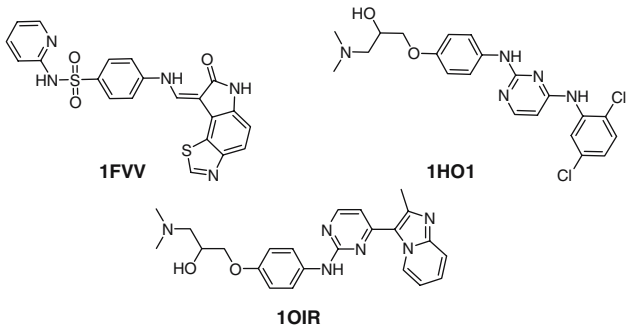

Fig. 2 Structures of the CDK2 inhibitors used in this study

synthesized (marked ''SynScreen'') or acquired (marked ''SimScreen'') specifically to target CDK2 activity. As a result the ''SynScreen'' and ''SimScreen'' sets contain active and inactive compounds that share some features known to be important for kinase activity. This ''pre-enrichment'' makes this part of the dataset more challenging for model building and a good test for a new descriptor. Following the approach we used in earlier work on building predictive models for this dataset [34], we concentrated on these ''SimScreen'' and ''SynScreen'' compounds and did not use the screening library compounds or those marked ''Moderately Active''. We were left with a set of 3464 compounds that had been synthesized and screened as part of a therapeutic project. This pool of compounds includes 161 actives (4.65%).

Enrichment curves and summaries for picking actives from the CDK2 dataset are shown in Fig. 3 and Table 2. These compare the performance of models

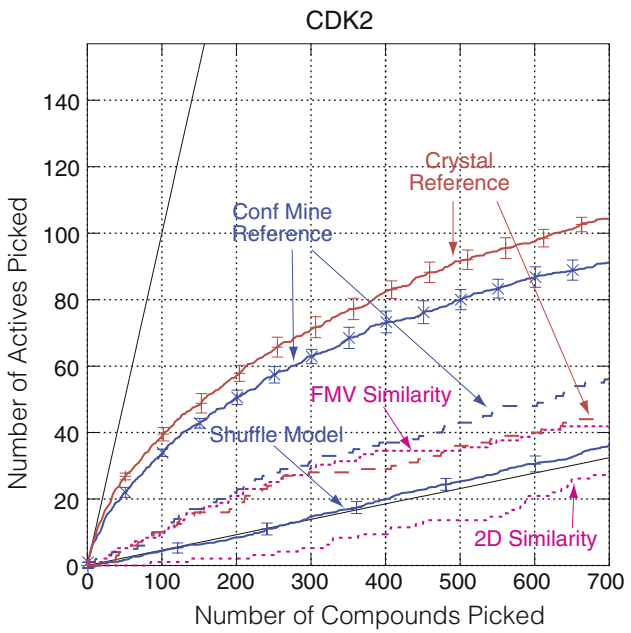

Fig. 3 Enrichment plots for the CDK2-picking experiments. Curves are shown for picking compounds using: bag classifiers built using FMVs based on crystal structures (solid red line) and conformation mining (solid blue line); total feature-map score based on crystal structures (dashed red line) and conformation mining (dashed blue line); and similarity to the actives used to build the FMVs (dotted lines). The two solid black lines indicate perfect performance (number of actives picked = number of compounds picked) and expected random performance (determined by the ratio of actives to inactives in the dataset). The performance of a shuffle-test model built using the conformation-mining FMVs, statistically indistinguishable from the random-picking results, is shown as a solid blue line. The error bars indicate 90% confidence levels

Table 2 Summary of the performance of the CDK2 experiments

	Enrichment at 50% of Actives (90% confidence level)
Conf-Mining FMVs	3.43 (0.13)
Crystal FMVs	4.54 (0.18)
Conf-Mining FeatMap Score	1.54
Crystal FeatMap Score	1.31
FMV Similarity	1.17
2D Similarity	0.74

(bag classifiers) using FMV descriptors to picking using similarity metrics and feature-map scores.

The FMV-based models are far more effective at finding actives than the other approaches—enrichments from the models are two to three times those from similarity or feature-map scores. The models have more information to work with in finding actives and are able to use that information in more sophisticated ways than simply summing FMV elements or looking at distances in either 2D fingerprint or FMV space. The remarkably poor (worse than random) performance of the 2D similarity approach for this dataset is because the compounds being screened, which were tested as part of a therapeutic project, are not topologically similar to the three actives used as probes. Figure 4 shows two representative inactives that were highly ranked in the 2D similarity screen of the CDK2 dataset using 1FVV as a probe. The first active, also shown in Fig. 4, is not encountered until pick 87. The CDK2 dataset simply does not contain many compounds that are ''chemically'' similar to 1FVV or the other probe ligands.

In our earlier work, we found that the alignments and feature maps from conformation mining are very

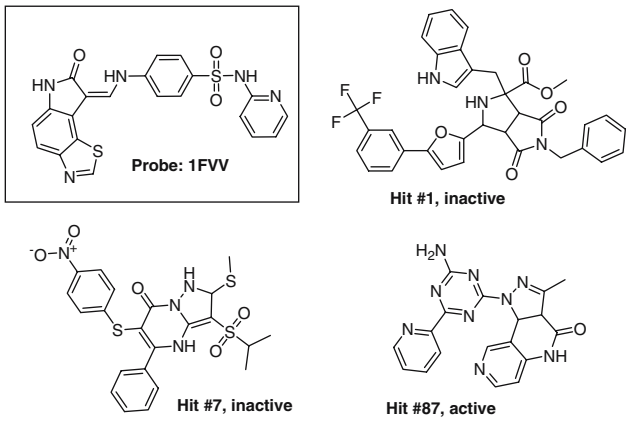

Fig. 4 Representative hits from searching the CDK2 data set for compounds similar to 1FVV. Hit #87 is the first active retrieved from the data set

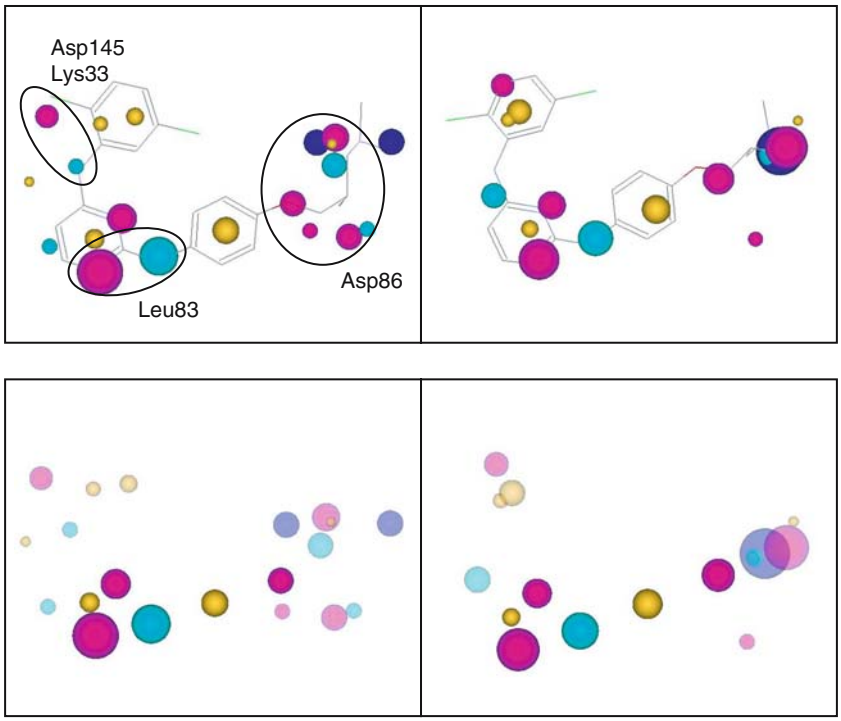

Fig. 5 *Top*: Comparison of the feature maps from crystal-structure data (left) and conformation mining (right) used to generate FMVs for CDK2 ligands. The best alignment of a conformation of compound 1HO1 is shown in each pane. Feature points are indicated by spheres with radius proportional to the feature's weight in the feature map. The colors of the spheres indicate feature type: purple: hydrogen-bond acceptor; cyan: hydrogen-bond donor; yellow: aromatic; blue: positively ionizable. *Bottom*: The same feature maps with the most-important features (see text) highlighted (opaque)

similar to those from the crystal structures of CDK2 ligands [16]. Given the similarity in feature maps (Fig. 5 top), it comes as little surprise that models built using the FMVs generated from conformation mining and those from crystal structures show similar performance. The reasons for the somewhat higher accuracy of the models built using FMVs derived from crystal-structure alignments are difficult to pinpoint exactly, but they could be connected to the artificial ''restrictions'' seen in the Asp86 region of the conformation-mining feature map. The crystal conformers show substantial flexibility in this region where they protrude from the binding cleft and are involved in interactions with solvent or bound water [30, 31, 35, 36]. Conformation mining, which seeks and emphasizes steric similarity between conformations, favors more extended conformations in this region. This artificial steric constraint in the Asp86 region may negatively impact other parts of the alignment in the conformation mining results and yield models that are somewhat less predictive.

By examining the individual decision trees that make up a bag classifier and collecting statistics for the FMV components that are frequently observed at higher levels of the trees, we can obtain an indication of which points in the feature map are most useful for building models for predicting CDK2 activity. The bottom two panes of Fig. 5 show the most important points in the CDK2 feature maps from the crystal structures (left) and conformation mining (right)—the correspondence here is perfect.

The donor–acceptor pair that interacts with Leu83 leads the list of important feature points. This ''two-point pharmacophore'', commonly observed in kinase/inhibitor complexes, is followed by a number of other feature points in the same region which are most likely responsible for properly orienting the donor and acceptor to interact with the Leu83 carbonyl O and amide N. The aromatic feature in this region is also involved in hydrophobic interactions with Ala131 and Leu134. The other important feature points are a hydrogen-bond acceptor in the Asp86 region and an aromatic feature between the Asp86 and Leu83 regions. It is interesting to note that the feature points in the region responsible for interactions with Asp145 and Lys33 do not play an important role in the models built using either the crystal structure or the conformation mining descriptors. Previous studies have shown that this region of the binding pocket is very probably associated with a structure-activity relationship [31, 36]. However for the CDK2 compounds in this study,

the degree of interaction with Asp145 and Lys33 is not useful in discriminating actives from inactives. In fact, removing these points from the descriptor set (results not shown) yields models that are statistically indistinguishable from those presented here. There are two primary reasons for this discrepancy. First, in our dataset the criterion for activity (taken from the original study: IC50 < 25 μM or percent inhibition > 50% at 10 μM) is probably too coarse to allow the models to detect these more subtle SARs. Second, and more importantly, we are building models to distinguish between actives and inactives in a set of compounds that were screened as part of a therapeutic project: these compounds all have conformers that place features in favorable positions in the Asp145/Lys33 region.

5-HT_3

5-HT_3 is a nervous-tissue bound (''M'' type) serotonin receptor that has been a successful target for developing anti-emetics [37] (including granisetron, ondansetron, and dolasetron). Although 5-HT_3 is distinguished from the other 5-HT receptors by the fact that it is a ligand-gated ion channel receptor not a G-protein coupled receptor, it still has not proven amenable to crystallization and no structural data are available. Although the crystal structures available for the other targets studied here were used only for validation of the FMV results, 5-HT_3 provides a demonstration of how feature-map vectors perform when structural data for the target are absent.

Starting with a pool of 752 antagonists of 5-HT_3 collected by Hert et al. [38], a set of four diverse compounds were picked for use in conformation mining (Fig. 6). The compounds were selected by clustering in a topological fingerprint space and selecting cluster centroids. The same diversity picking procedure was employed to select a set of 50 distinct actives to use in model building/validation. There is no overlap between the 4 actives used for conformation mining and the 50 actives used for modeling. Because we do not have screening data containing 5-HT_3 inactives available, we were forced to select a set of ''presumed inactives'' from a screening library of drug-like compounds (the Bionet library [39]) and assume that they were inactive. In order to make the experiment more realistic, we only selected compounds that are similar in molecular weight to those in the active pool (i.e. that have a molecular weight within 100 g/mol of the mean active molecular weight) and that contain a positively ionizable group (a key feature for 5-HT_3 antagonists). A random subset of 1400 of the compounds in Bionet meeting these two criteria was selected to serve as inactives. The final data set contained 1450 compounds (3.4% active).

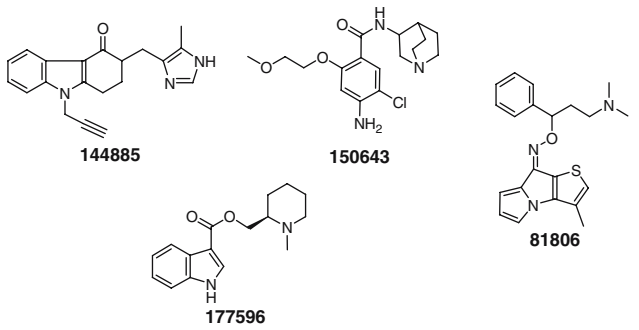

Fig. 6 Structures of the 5-HT_3 antagonists used to derive an alignment reference and feature map

In order to control for bias in the results as a result of using diversity selection to choose active compounds for our model building, a second set of experiments was done using a randomly selected set of 50 5-HT_3 actives with the same set of 1400 inactives. The results of these experiments are also reported.

The modeling results for 5-HT_3 are summarized in Fig. 7 and Table 3. First, note that although the model built using 50 diverse actives outperforms that using 50 random actives, the difference is not large compared to the relative performances of the models derived using other approaches. The inherent diversity of the dataset is such that using an explicit diversity criterion to select actives does not dramatically improve the model-building results. As one would expect, 2D similarity performs slightly better on the dataset using 50 random actives (not shown in Fig. 7) than on the dataset using 50 diverse actives, but the 2D approach is still not as efficient as FMV similarity.

On the 5-HT_3 dataset, simply picking compounds based on feature-map score performs very well up to about 20 actives picked (40% of the actives), then this curve falls off relative to the predictive models. The score-based enrichment at 50% of actives picked (7.4, or 2.5 relative to fingerprint similarity) is quite high, but it compares poorly to the excellent enrichment from the models (28.7, 9.8 relative to fingerprint similarity). Although the fingerprint similarity-based picking scheme leads to real enrichment over random picking in this synthetic dataset, it is substantially outperformed by both feature-map scores and the FMV-based models. This indicates that the results observed are not solely due to topological differences between the active and inactive compounds and that the FMV-based models are learning features important for binding across multiple classes of compounds.

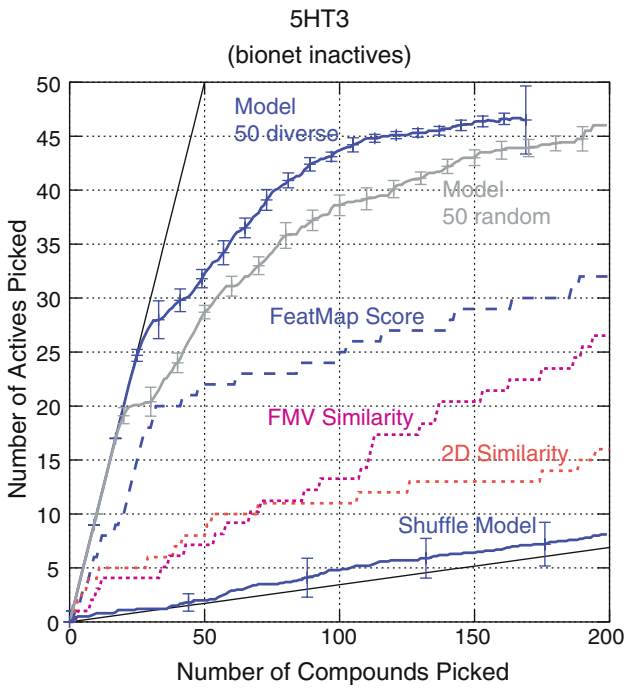

Fig. 7 Enrichment curves for the 5-HT_3 experiments. Curves are shown for picking compounds using: a bag classifier built with conformation-mining FMVs using both 50 diverse actives in the training set (solid blue line) and 50 random actives in the training set (solid gray line); total feature-map score (dashed blue line); similarity to the actives used to build the FMVs in FMV space (dotted purple line); and Tanimoto similarity using topological fingerprints (dotted red line). The two solid black lines indicate perfect performance (number of actives picked = number of compounds picked) and expected random performance (determined by the ratio of actives to inactives in the dataset). The performance of a shuffle-test model, statistically indistinguishable from random picking, is shown as a solid blue line. The error bars indicate 90% confidence levels

Table 3 Summary of the performance of the 5-HT_3 experiments

	Enrichment at 50% of Actives (90% confidence level)
Conf-Mining FMVs	28.7 (0.8)
FeatMap Score	7.4
FMV Similarity	3.87
Fingerprint Similarity	2.93

The quality of the FMV-based predictive models was further tested by evaluating their performance on the 694 actives that were not used to build the model. This screen generates a hit rate of 90.4(0.4)%, i.e. it correctly identifies about 90% of these holdout actives. This provides further confidence that the FMV-based models are learning the features that are important for activity, not just memorizing the datasets.

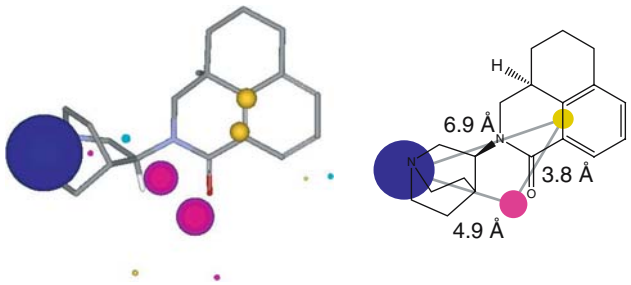

Fig. 8 *Left*: The feature map for 5-HT_3 ligands with a conformation of compound (*S*,*S*)-**37** from Reference [39] as aligned by the 5-HT_3 alignment reference. Unlike in Figure 5, where the feature spheres were scaled by their weights in the feature map, in this Figure feature spheres are scaled by the magnitude of the corresponding FMV elements. *Right*: a 2D sketch of (*S*,*S*)-**37** with important feature positions and distances between them indicated. The acceptor (purple) and aromatic (yellow) features are averages of the corresponding points in the left part of the Figure, see text

In order to compare our results to previous studies of 5-HT_3, we generated the feature-map vector for one of the compounds used in Ref. [40] to generate a pharmacophore for 5-HT_3, (*S*,*S*)-**37** p*K*i = 10.4, drawn at the right in Fig. 8. We followed the standard approach for this: after generating conformations for (*S*,*S*)-**37** we aligned them to the 5-HT_3 alignment reference using the subshape algorithm, scored the alignments using the feature map, and picked the top alignment to generate FMVs. The best alignment is shown at the left in Fig. 8 along with the FMV, which is visualized by scaling the radius of each feature-map point by the corresponding entry in the FMV. In this depiction, large features indicate large values in the FMV.

This arrangement of features is strikingly similar to the 5-HT_3 pharmacophore identified by other researchers [40, 41]. The right-hand side of Fig. 8 shows a sketch of this pharmacophore superimposed on a 2D drawing of the molecule. Because the feature map has multiple points for the acceptor and aromatic features, we report average distances in the Figure. These distances are very close to those seen in the previous work: our aromatic—positively ionizable distance, 6.9 Å, falls between those reported by Clark et al. 7.4 Å and Hibert et al. 6.7 Å. Our acceptor—positively ionizable distance is a bit shorter, 4.9 Å versus 5.2 Å in both previous works, while our aromatic–acceptor distance, 3.8 Å, is longer than the value from Hibert et al., 3.3 Å (Clark *et al.* do not report an aromatic–acceptor distance). It is reassuring to see that our shape-based alignment approach, combined with FMVs, can reproduce these literature pharmacophores so closely.

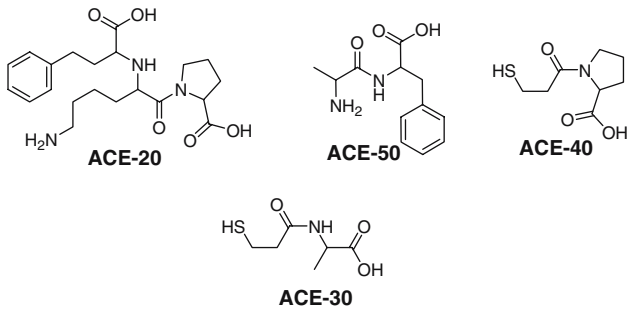

Fig. 9 Structures of the ACE inhibitors used to derive an alignment reference and feature map

ACE

Angiotensin-converting enzyme (ACE) is a zinc-containing metallopeptidase that converts angiotensin I to angiotensin II [42]. Because angiotensin II is a vasoconstrictor and ACE also inactivates the vasodilator bradykinin, it has a powerful effect on regulation of blood pressure. ACE is also a regulator of fertility [43] and renal functions [44].

To build models for predicting ACE inhibition we used a set of 64 known ACE inhibitors from the literature [45], along with a set of presumed inactives from Bionet. In order to carry out conformation mining and construct a feature map, a set of four topologically diverse actives was selected by clustering in topological fingerprint space and choosing cluster centroids. These inhibitors are shown in Fig. 9. The inactive pool was again constrained to consist of only compounds in the same molecular weight range as the actives that also contain a key feature (a zinc binder for this metalloprotein). Our final data set contained 1321 compounds (4.84% active).

The enrichment curves for picking actives from the ACE dataset are shown in Fig. 10. Although both the FMV-based models and feature-map score picking methods perform very well, the meaningfulness of these results is called into serious question by the performance of picking using 2D similarity. In this dataset the actives are so closely related to each other topologically that 2D similarity is all that is required to pick them from a pool of diverse inactives. So, although the FMV-based models produce excellent enrichment values, 19.2(0.2) at 50% picking, there is just no way to tell if the 3D method is learning anything about the 3D structure or if it is just reproducing 2D features.

A further cautionary note about this particular dataset is warranted. When 2D similarity picking is used to identify actives from the full set of almost

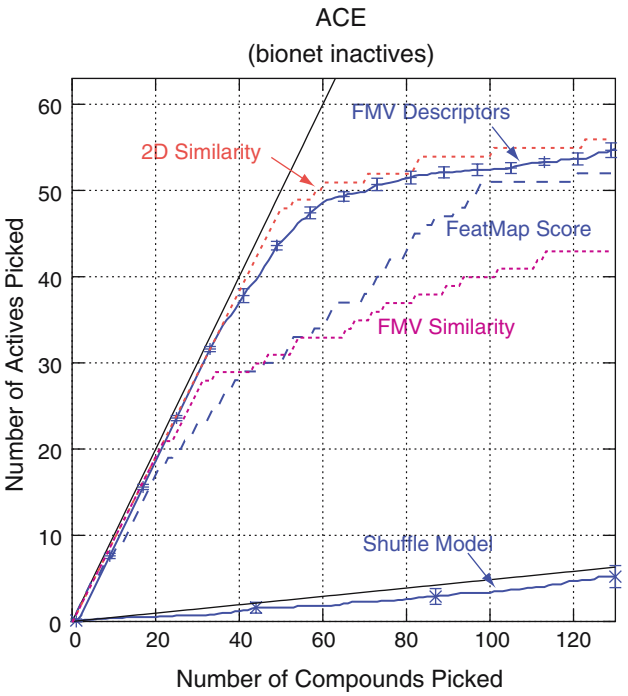

Fig. 10 Enrichment curves for the ACE experiments. Curves are shown for picking compounds using: a bag classifier built using FMVs based on conformation mining (solid blue line); total feature-map score (dashed blue line); similarity to the actives used to build the FMVs in FMV space (dotted purple line); and Tanimoto similarity using topological fingerprints (dotted red line). The two solid black lines indicate perfect performance (number of actives picked = number of compounds picked) and expected random performance (determined by the ratio of actives to inactives in the dataset). The performance of a shuffle-test model, statistically indistinguishable from random picking, is shown as a solid blue line. The error bars indicate 90% confidence levels

34,000 Bionet compounds (results not shown here), the performance is even more impressive. Using a set of three literature actives (captopril, enalapril, and lisinopril) as the probes gives perfect picking at 50% of actives picked—an enrichment of 556 over random. Even out as far as 80% of actives picked, the enrichment is still 314. This set of actives, seeded into a large pool of unrelated compounds, make 2D similarity picking look like a very impressive strategy indeed.

Additional validation experiments: DHFR and thrombin

Two additional experiments using targets where crystal structures are available were carried out to validate the utility of FMVs for building predictive models for biological activity. A detailed description of these two

experiments and their results, which are summarized here, is available in the supplemental material.

The first supplemental experiment used a literature collection of compounds with IC50 values for inhibition of dihydrofolate reductase (DHFR) [46]. An alignment reference and feature map were created by applying conformation mining to three DHFR inhibitors with available crystal structures (PDB codes 1HFQ [47], 1S3U [48], and 1KMS [49]). Models built using FMVs do an excellent job of identifying actives from the DHFR data set, showing an enrichment of almost 12 relative to random picking (almost 7 relative to picking using 2D similarity). The FMVs from conformation mining yield models that are as accurate as those from the crystal conformers at 50% of actives picked.

The second supplemental experiment used a collection of 140 literature actives for thrombin [50] and a pool of 1400 assumed inactives from Bionet that contain a key feature (a positively ionizable group) and that are in the same molecular weight range as the actives. Conformation mining was applied to four inhibitors where bound-ligand crystal structures are available (PDB codes 1ETS [51], 1ETR [51], 1A61 [52], and 1UVT [53]). Once again the models built using FMV descriptors show high enrichments relative to both random picking—10.9(0.1) at 50% of actives—and 2D similarity—15.4(0.1) at 50% actives.

Conclusions

Feature-map vectors, a new descriptor based on chemical feature maps, have been defined and explained. The validation studies presented here demonstrate that FMVs generated using conformation mining as an alignment strategy can be used to build robust predictive models that identify diverse active compounds with high enrichments and significantly improved enrichments compared to the feature map score alone. Using the CDK2 and 5-HT_3 datasets, we have demonstrated that FMVs encode features that are largely independent of the topology and that the resulting models are capable of learning the features that are important for binding. In addition, the straightforward nature of the FMV descriptors—each element of an FMV corresponds to a molecule's score against a particular chemical feature – expedites analysis and interpretation of the predictive models themselves. Each descriptor identified as important by the models indicates a specific component of the chemical feature map that is important for distinguishing actives from inactives.

Although the validation studies presented here have focused on using FMVs to build models for classification of compounds as either active or inactive, we believe that the descriptors contain sufficient information to be useful for building regression models. Future work will focus on using FMVs to develop QSARs and applying FMVs to the problems of lead optimization and predicting selectivity.

Acknowledgements The authors would like to thank Erin Bradley (Sunesis Inc.) for providing the aligned CDK2 crystal structures and Christian Lemmen (BioSolveIT Gmbh) for providing the ACE and thrombin datasets.

References

1. Cramer RD 3rd, Patterson DE, Bunce JD (1988) J Am Chem Soc 110:5959
2. Klebe G, Abraham U, Mietzner T (1994) J Med Chem 37:4130
3. Guner O (ed) (2000) Pharmacophore perception, development, and use in drug design. International University Line, La Jolla
4. Eksterowicz JE, Evensen E, Lemmen C, Brady GP, Lanctot JK, Bradley EK, Saiah E, Robinson LA, Grootenhuis PD, Blaney JM (2002) J Mol Graph Model 20:469
5. Renner S, Schneider G (2004) J Med Chem 47:4653
6. Putta S, Lemmen C, Beroza P, Greene J (2002) J Chem Inf Comput Sci 42:1230
7. Greene J, Kahn S, Svoj H, Sprague P, Teig S (1994) J Chem Inf Comput 34:1297
8. *MOE, Molecular Operating Environment*, Chemical Computing Group
9. Lemmen C, Lengauer T, Klebe G (1998) J Med Chem 41:4502
10. Putta S, Eksterowicz J, Lemmen C, Stanton R (2003) J Chem Inf Comput Sci 43:1623
11. Warren G, Webster Andrews C, Capelli A-M, Clarke B, LaLonde J, Lambert M, Lindvall M, Nevins N, Semus S, Senger S, Tedesco G, Wall I, Woolven J, Peishoof C, Head M (2005) J Med Chem ASAP http://dxdoiorg/101021/jm050362n
12. Klon AE, Glick M, Davies JW (2004) J Chem Inf Comput Sci 44:2216
13. Klon AE, Glick M, Davies JW (2004) J Med Chem 47:4356
14. Klon AE, Glick M, Thoma M, Acklin P, Davies JW (2004) J Med Chem 47:2743
15. Springer C, Adalsteinsson H, Young MM, Kegelmeyer PW, Roe DC (2005) J Med Chem 48:6821
16. Putta S, Landrum GA, Penzotti JE (2005) J Med Chem 48:3313
17. Smellie A, Stanton R, Henne R, Teig S (2003) J Comput Chem 24:10
18. Webb A (2002) Statistical pattern recognition. John Wiley & Sons, Hoboken
19. Mitchell T (1997) Machine learning. McGraw-Hill, New York
20. Dietterich TG (1997) AI Mag 18:97
21. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP (2003) J Chem Inf Comput Sci 43:1947
22. Landrum GA, Penzotti JE, Putta S (2004) Mat Res Soc Symp Proc 804:JJ115

23. Breiman L (1996) Machine Learning 24:123
24. Fayyad UM, Irani KB (1992) Machine Learning 8:87
25. Fayyad UM, Irani KB (1993) Multi-interval discretization of continuous-valued attributes for classification learning. 13th International Joint Conference on Artificial Intelligence, Morgan-Kaufmann, pp 1022–1027
26. Out-of-Bag Estimation, UC Berkeley Department of Statistics, ftp://ftpstatberkeleyedu/pub/users/breiman/OOBestimationpsZ
27. Bender A, Glen RC (2005) J Chem Inf Model 45:1369
28. Norbury C, Nurse P (1992) Annu Rev Biochem 61:441
29. Sherr CJ (1996) Science 274:1672
30. Davis ST, Benson BG, Bramson HN, Chapman DE, Dickerson SH, Dold KM, Eberwein DJ, Edelstein M, Frye SV, Gampe RT Jr, Griffin RJ, Harris PA, Hassell AM, Holmes WD, Hunter RN, Knick VB, Lackey K, Lovejoy B, Luzzio MJ, Murray D, Parker P, Rocque WJ, Shewchuk L, Veal JM, Walker DH, Kuyper LF (2001) Science 291:134
31. Anderson M, Beattie JF, Breault GA, Breed J, Byth KF, Culshaw JD, Ellston RP, Green S, Minshull CA, Norman RA, Pauptit RA, Stanway J, Thomas AP, Jewsbury PJ (2003) Bioorg Med Chem Lett 13:3021
32. Beattie JF, Breault GA, Ellston RP, Green S, Jewsbury PJ, Midgley CJ, Naven RT, Minshull CA, Pauptit RA, Tucker JA, Pease JE (2003) Bioorg Med Chem Lett 13:2955
33. Bradley EK, Miller JL, Saiah E, Grootenhuis PD (2003) J Med Chem 46:4360
34. Landrum GA, Penzotti JE, Putta S (2004) *eChemInfo 2004*
35. Bramson HN, Corona J, Davis ST, Dickerson SH, Edelstein M, Frye SV, Gampe RT Jr, Harris PA, Hassell A, Holmes WD, Hunter RN, Lackey KE, Lovejoy B, Luzzio MJ, Montana V, Rocque WJ, Rusnak D, Shewchuk L, Veal JM, Walker DH, Kuyper LF (2001) J Med Chem 44:4339
36. Breault GA, Ellston RP, Green S, James SR, Jewsbury PJ, Midgley CJ, Pauptit RA, Minshull CA, Tucker JA, Pease JE (2003) Bioorg Med Chem Lett 13:2961
37. Gozlan H, In Olivier B, van Wijngaarden I, Soudijn W (eds) (1997) Serotonin receptors and their ligands, Elsevier, Amsterdam
38. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A (2004) J Chem Inf Comput Sci 44:1177
39. Bionet Screening Compounds Database, Key Organics Limited, http://wwwkeyorganicsltduk/screeninhtm
40. Clark RD, Miller AB, Berger J, Repke DB, Weinhardt KK, Kowalczyk BA, Eglen RM, Bonhaus DW, Lee CH, Michel AD et al (1993) J Med Chem 36:2645
41. Hibert MF, Hoffmann R, Miller RC, Carr AA (1990) J Med Chem 33:1594
42. de Gasparo M, Catt KJ, Inagami T, Wright JW, Unger T (2000) Pharmacol Rev 52:415
43. Hagaman JR, Moyer JS, Bachman ES, Sibony M, Magyar PL, Welch JE, Smithies O, Krege JH, O'Brien DA (1998) Proc Natl Acad Sci USA 95:2552
44. Kessler SP, deS Senanayake P, Scheidemantel TS, Gomos JB, Rowe TM, Sen GC (2003) J Biol Chem 278:21105
45. Fink C (1996) Exp Opin Ther Pat 6:1147
46. Sutherland JJ, O'Brien LA, Weaver DF (2003) J Chem Inf Comput Sci 43:1906
47. Cody V, Galitsky N, Luft JR, Pangborn W, Blakley RL, Gangjee A (1998) Anticancer Drug Des 13:307
48. Cody V, Luft JR, Pangborn W, Gangjee A, Queener SF (2004) Acta Crystallogr D Biol Crystallogr 60:646–55
49. Klon AE, Heroux A, Ross LJ, Pathak V, Johnson CA, Piper JR, Borhani DW (2002) J Mol Biol 320:677–93
50. Stahl M, Rarey M, Klebe G (2001) In: Lengauer T (ed) Bioinformatics: from genomes to drugs, VCH, Weinheim, pp 137–170
51. Brandstetter H, Turk D, Hoeffken HW, Grosse D, Sturzebecher J, Martin PD, Edwards BF, Bode W (1992) J Mol Biol 226:1085
52. St Charles R, Matthews JH, Zhang E, Tulinsky A (1999) J Med Chem 42:1376
53. Engh RA, Brandstetter H, Sucher G, Eichinger A, Baumann U, Bode W, Huber R, Poll T, Rudolph R, von der Saal W (1996) Structure 4:1353