Classification of Kinase Inhibitors Using BCUT Descriptors

Bernard Pirard* and Stephen D. Pickett†

Aventis Pharma, Dagenham Research Centre, Rainham Road South, Dagenham, Essex RM10 7XS, U.K.

Received April 1, 2000

BCUTs are an interesting class of molecular descriptor which have been proposed for a number of design and QSAR type tasks. It is important to understand what kind of information any particular descriptor encodes and to be able to relate this to the biological properties of the molecules. In this paper we present studies with BCUTs for the classification of ATP site directed kinase inhibitors active against five different protein kinases: three from the serine/threonine family and two from the tyrosine kinase family. In combination with a chemometric method, PLS discriminant analysis, the BCUTs are able to correctly classify the ligands according to their target. A novel class of kinase inhibitors is correctly predicted as inhibitors of the EGFR tyrosine kinase. Comparison with other descriptor types such as two-dimensional fingerprints and three-dimensional pharmacophore-based descriptors allows us to gain an insight into the level of information contained within the BCUTs.

INTRODUCTION

High-throughput screening and combinatorial chemistry are now standard tools within the drug discovery process. The growth of these methodologies has had a large impact on computational chemistry, leading to the development of molecular descriptors to characterize and compare compound libraries.[1−5] These descriptors were first developed in the context of designing diverse libraries;[6,7] more recently, they have also been applied to hit follow-up and focused library design.[8−12] It is useful to categorize the descriptors into three main categories:[1−4,13]

(1) 2D fingerprints which encode the presence or absence of certain structural features; (2) property-based descriptors, *C* log *P*, molecular weight, connectivity indices, etc.; (3) 3D pharmacophore-based descriptors.

Several studies have been published[6,7,14] which attempt to assess the relative merits of these different approaches. A common approach to this problem has been to look for enrichment of known actives within a set of selected compounds, compared to random selection. In many cases 2D fingerprints appear to perform remarkably well compared with other methods. However, it is also possible to find cases where 2D fingerprints would not be successful, for example if a tripeptide such as RGD is to be used to search a database of small molecules for RGD mimics.[12] In such a case, 3D pharmacophore descriptors are required to achieve a successful outcome. An alternative approach to descriptor validation has been to look at their effectiveness in a QSAR-type context,[7] introducing the concept of "neighborhood behavior". The reader is referred to several reviews which cover these issues in greater detail.[4,15]

Table 1. Abbreviations Used in This Paper

BCUT	Burden Chemical Abstract Service University of Texas
ATP	adenosine triphosphate
PCA	principal component analysis
PLS DA	partial least-squares discriminant analysis
PC	principal component
PLS	partial least-squares
JNK	c-Jun N-terminal kinase
CDK	cyclin-dependent kinase
SYK	spleen tyrosine kinase
EGFR	epidermal growth factor receptor
T	Tanimoto coefficient
R^2(CUM)	cumulative sum of squares of all the variables explained by all extracted components
Q^2(CUM)	cumulative fraction of the total variation of the dependent variables that can be predicted by all extracted components
R^2X(CUM)	cumulative sum of squares of all the independent variables explained by all extracted components
R^2Y(CUM)	cumulative sum of squares of all the dependent variables explained by all extracted components
MAP	mitogen-activated protein
2D	two-dimensional
3D	three-dimensional

From our perspective, the choice of descriptor depends on a number of factors including the information available at the time, the type of question being asked, and the ease and speed of calculation in relation to the time scale of response required.[4] We also feel it is important that the user understands and can interpret the results from the descriptors being used, thus allowing a more rational choice of descriptor to be made. Most analyses have tended to focus on one biological target at a time. We were interested to explore the effectiveness of BCUT metrics (a list of abbreviations used in this paper is given in Table 1) in discriminating between ligands of different members of a given protein family, the serine/threonine and tyrosine kinases. Comparison with 2D fingerprints from Daylight and multipharmacophore descriptors should also allow us to gain an insight into the type of information the BCUT descriptors encode, thus permitting a better interpretation of the results.

* Corresponding author. Current address: Aventis Pharma Deutschland GmbH Molecular Modeling, Chemical Research G838, D-65926 Frankfurt am Main, Germany. Phone: ++49-69-30583768. FAX: ++49-69331399. E-mail: bernard.pirard@aventis.com.

† Current address: Roche Discovery Welwyn, Broadwater Road, Welwyn Garden City, Hertfordshire, AL7 3AY, U.K.

10.1021/ci000386x CCC: $19.00
Published on Web 10/12/2000

BCUT DESCRIPTORS

BCUT metrics[16] result from an extension of Burden's work.[17] Burden represented the hydrogen-suppressed connection table of a molecule as a symmetrical $N \times N$ matrix with atomic numbers along the diagonal and bonding information in the off-diagonal elements. Pearlman et al. extended this approach to a multidimensional space by considering the highest and lowest eigenvalues of four classes of BCUT matrices, which contain atomic properties significant for ligand−receptor interactions on their diagonals: atomic charges, polarizabilities, H-bond acceptor abilities, and H-bond donor abilities. These properties can also be weighted by atomic surface areas. Burden's suggestion of using bonding information on the off-diagonal elements was extended to functions of topological and interatomic distances. In order to provide the proper balance between diagonal and off-diagonal elements, a scaling factor was also found to be necessary. The combinations of possible diagonal, off-diagonal, and scaling factor choices lead to over 60 possible BCUT descriptors per chemical structure. The DiverseSolutions[18] software has been written to generate and manipulate the descriptors. Smaller subsets of BCUTs (four to six) can be automatically selected by a χ-squared-based "auto-choose" algorithm to best represent the structural diversity of a given compound collection.[16] This low dimension chemistry space can then be partitioned into cells for compound selection and library comparison. Such a procedure has been shown to be able to cluster active compounds[19] in what has been termed a receptor relevant subspace, whose axes correspond to metrics related to affinity for a given receptor. Researchers at Pharmacopeia and Rhône-Poulenc Rorer have used the BCUTs to design and analyze combinatorial libraries.[20,21] In particular, Schnur reported that actives for two related targets, plasmepsin and cathepsin, shared some common regions of the chemistry space, but also had some nonoverlapping regions.[20]

In this paper, we have used the BCUTs in combination with chemometric methods (see Methods) to explore their ability for distinguishing between ligands of five related protein targets, three serine/threonine kinases and two tyrosine kinases. We compare the results obtained from BCUTs with Daylight 2D fingerprints which record the presence of all possible linear substructures up to a length of seven. These fingerprints are commonly used for diversity-related tasks and have been shown to distinguish actives from inactives.[14,15] We also consider multipharmacophore descriptors which encode all possible combinations of three or four pharmacophoric points (aromatic, H-bond acceptor, H-bond donor, hydrophobe, acid, base) and distances for a molecule in a bit-string. Several groups have used multipharmacophore descriptors to design and analyze combinatorial libraries[11,12,22] and in a QSAR context.[23]

Protein kinases are key components of many signal transduction pathways, which are involved in the control of cell growth, metabolism, differentiation, and apoptosis. Signal transduction, via protein kinases, occurs through selective, reversible phosphorylation of serine, threonine, or tyrosine residues of a protein substrate. Both the large number of different protein kinases (the estimated number of mammalian kinases is 2000[24]) and the conservation of structural features within the ATP binding cleft[25,26] have made the discovery of specific ATP site directed kinase inhibitors a challenging task. However, such compounds have been described over the past few years.[25−28] We have considered ATP site directed inhibitors of five protein kinases, three serine/threonine (JNK1, P38, and CDK1) and two tyrosine kinases (SYK and EGFR), and show how BCUTs in combination with chemometric methods are able to distinguish between the different classes of ligands and provide useful information into the different structural requirements of the target proteins.

Table 2. Main Features of the 770 Compound Data Set

class	Ncomp[a]	Ntrain[b]	Ntest[c]	origin
JNK1	277	139	138	in-house
P38	84	43	41	literature[29,30]
CDK1	58	29	29	literature[31,32]
SYK	127	64	63	in-house
EGFR	224	114	110	literature[33−45]
Total	770	389	381	

[a] Total number of compounds in a class. [b] Number of compounds in the PLS DA training set. [c] Number of compounds in the PLS DA test set.

METHODS

Data Sets. We built a collection of 770 potent (maximum $IC_{50} = 1\ \mu M$) ATP site directed inhibitors of five kinases: three serine/threonine (JNK1, P38, and CDK1) and two tyrosine (SYK and EGFR) kinases.[29−45] Compounds appearing in patents, for which IC_{50} values are usually not available, were not considered. In addition, we only included inhibitors for which IC_{50} values were measured by ATP displacement experiments. For PLS DA modeling, each class of inhibitors was split into a training set and a test set of equal size. The main features of this 770 compound data set are given in Table 2.

To investigate further predictive ability of the PLS DA model, we used a model derived from the full 770 compound data set to predict the class membership of 52 novel EGFR inhibitors.[46,47]

BCUTs and Chemometric Methods. SMILES[48] strings provide the input for DiverseSolutions v. 4.0.5[18] which computes 67 BCUTs per compound. The BCUTs were analyzed by PCA and PLS DA. In PCA,[49] the original data matrix is approximated in terms of the products of two smaller matrices, the scores and the loadings matrices. The scores matrix gives a simplified picture of the objects (kinase inhibitors), described in terms of their projection onto the principal components (PCs). The loadings matrix contains the PCs, which are linear combinations of the original variables (BCUTS). The first PC describes the maximum variance among all possible directions, the second component the next largest variation among all directions orthogonal to the first one, and so on. In order to find the minimum number of components necessary for data reproduction within residual errors, the components are added step by step to the model. Analysis of the score plots gives information about the clustering of the objects, while the loading plots contain information about the relative contributions of the variables to the PCs. To gain further insight into the grouping

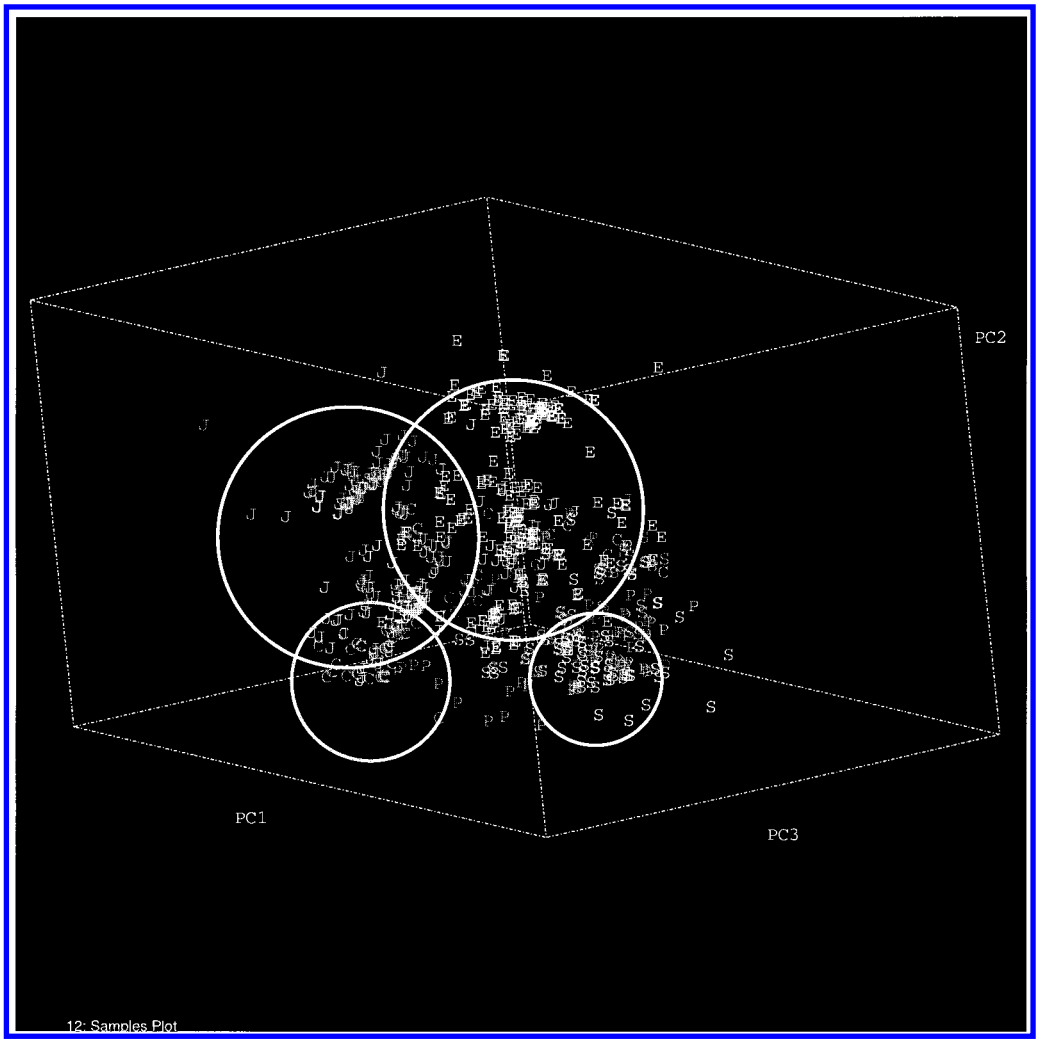

Figure 1. PCA scores plot for PC1, PC2, and PC3. JNK1, P38, CDK1, SYK, and EGFR inhibitors are displayed respectively as J, P, C, S, and E. The four circles are centered on the regions where, as assessed by visual inspection, most of the JNK1, CDK1, SYK, and EGFR inhibitors are found. Compared with the other classes of compounds, the P38 class is more disperse.

of compounds by activity classes, the scores were used as input for hierarchical (Ward's, single, complete, and average linkage) and nonhierarchical (Jarvis−Patrick) cluster analysis techniques.[50] The PCA was performed with either SIMCA 7.0[51] or the QSAR+ module of Cerius[2] 4.0.[52] The clustering was performed within Cerius[2] 4.0.

When groups of objects identified by PCA are poorly separated, PLS DA can be used to resolve them better. PLS DA proceeds in two steps.[53] First, a class membership variable (*Y*) is assigned to each object. For instance, in a three-class system, *Y* is the (1,0,0) vector for members of the first category, the (0,1,0) vector for objects in the second category, etc. In this study, a category consists of inhibitors of a given kinase. A PLS analysis is performed to relate *X* and *Y*. The resulting PLS DA model can be used to predict the class membership of an external set of objects. In the SIMCA 7.0 implementation of PLS DA, cross-validation is performed and an object can be assigned either to one (several) of the defined class(es) or to none of them. For prediction, we have considered an object as belonging to a given class when the corresponding component of its predicted *Y* vector is greater than 0.50.

Daylight Fingerprints. The Daylight fingerprint program[54] was run to generate 2D fingerprints, which were used as input to an in-house agglomerative hierarchical clustering method.[55] This clustering tool tries to maximize the intercluster variance, as the intracluster variance is minimized. Previously, other groups have suggested a pairwise similarity value of $T = 0.85$, at which one can expect that if a structure is active then 80% of the structures with a greater than threshold pairwise similarity to it will be active themselves.[7,15] However, it is not the prime purpose of this paper to explore the utility of Daylight fingerprints in such an approach but rather to use them to explore the structural similarities between the different ligand classes. Therefore, in this study, we considered a lower similarity threshold (*T*) of 0.65 to merge compounds into the same cluster. This is the value we have found useful for clustering structures by type.

Multipharmacophore Descriptors. Details of our methodology have been given elsewhere.[11,56,57] Briefly, 3D structures are generated from SMILES[48] strings using CONCORD.[58,59] The ChemDiverse [60] module of Chem-X (July 98 version) is used to calculate the descriptors. Atom typing is performed using an in-house customization of the standard Chem-X parametrization.[11,56] A systematic search procedure, controlled by in-house scripts, is used for the conformational analysis.[11,57] In this work, we considered both three- and four-point pharmacophores, defined from six of the seven possible user-definable atom types: hydrogen bond

Table 3. PLS DA Classification Results for the Training and Test Sets

	Ncorr[a] JNK1	Ncorr P38	Ncorr CDK1	Ncorr SYK	Ncorr EGFR
training set	136/139 (98)	35/43 (81)	22/29 (76)	49/64 (77)	95/114 (83)
test set	127/138 (92)	30/41 (73)	19/29 (65)	45/63 (71)	101/110 (92)
total	263/277 (95)	65/84 (77)	41/58 (71)	94/127 (74)	196/224 (88)

[a] Number of compounds correctly predicted by the PLS DA model; the percentage of correctly predicted compounds is given in parentheses. A prediction was considered as correct when the class membership variable $Y > 0.50$.

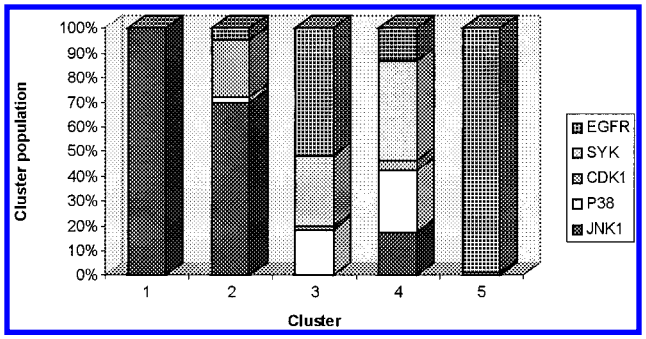

Figure 2. Distribution of the inhibitors from the five different kinases in the clusters identified by Ward's method performed on PC1−PC6.

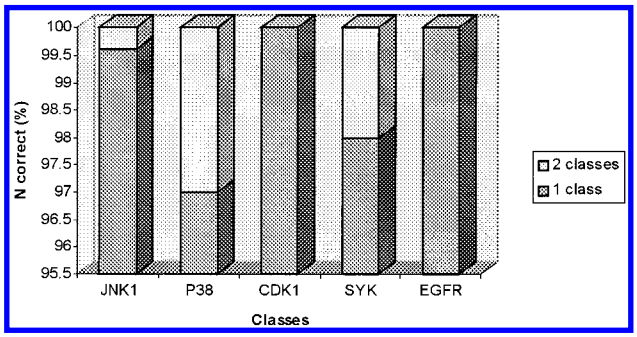

Figure 3. Percentages of correctly predicted compounds (N correct) for the combined training and test sets, which have been assigned to one and two kinase classes by the PLS DA model.

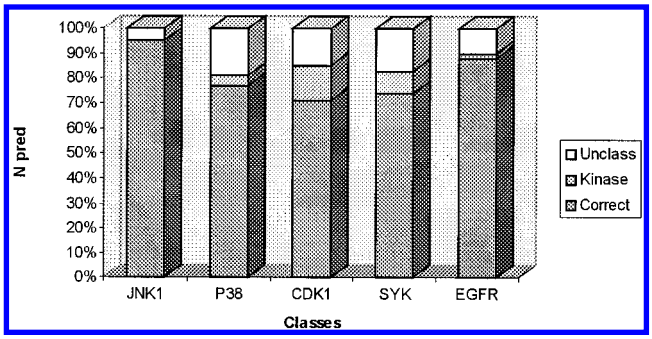

Figure 4. Distribution of the compounds (770 compound set) predicted by the PLS DA model (N pred) for the five classes of kinases.

Table 4. Results of Hierarchical Clustering on Daylight Fingerprints

data set	Ncomp	Nclus[a]	av clus. size[b] (*S*)[c]	Nsing[d]
JNK1	277	21	13.0 (17.0)	4
P38	84	15	6.0 (5.5)	5
CDK1	58	5	12.0 (10.4)	2
SYK	127	12	11.0 (8.3)	2
EGFR	224	28	8.0 (6.4)	2
whole	770	81	9.5 (11.2)	15

[a] Number of clusters for a threshold of $T = 0.65$. [b] Average number of compounds per cluster. [c] Standard deviation for the distribution of the number of compounds. [d] Number of clusters containing only one compound.

donor, hydrogen bond acceptor, aromatic center, hydrophobic point, acid, and base. Hydroxyl groups were included in the definitions of donors and acceptors. The distances between pharmacophore points were binned using the same distance ranges as in previous studies.[11,57] A ChemLib routine has been coded to write individual pharmacophore keys to disk for subsequent postprocessing and gives considerable saving in terms of disk space. The frequency of occurrence of each pharmacophore within a given set of compounds as well as within the combined data set can be recorded.

RESULTS

PCA on BCUTS. PCA on a 770 compound by 67 BCUTs matrix led to a six-component model (R^2(CUM) = 0.817). These six PCs account for 39.6%, 14.0%, 9.5%, 7.8%, 5.7%, and 4.9% of the variance, respectively. Inhibitors of a given kinase tend to group together on the PC1−PC3 score plot (Figure 1). However, the discrimination between some classes (P38 and SYK for instance) is poor. This partial separation is confirmed by the results of the Ward's clustering on PC1−PC6 in Cerius^2 4.0. Figure 2 shows the cluster population statistics for five clusters. Three clusters (clusters 2, 3, and 4) contain a significant proportion of compounds from more than one kinase class. Considering more clusters did not improve the discrimination between the different kinase inhibitors, while the other clustering techniques (Methods) failed to produce a better separation (data not shown).

PLS DA on BCUTs. For a 389 compound training set (Table 2), a five-component PLS DA model (Q^2(CUM) = 0.604, R^2X(CUM) = 0.745, R^2Y(CUM) = 0.620) was obtained. This model has been used to classify both training (cross-validation) and test set compounds. The model was able to classify kinase inhibitors with a success rate greater than 70% in all cases except for the CDK1 test set (65% success rate; Table 3). Nearly all of the correctly classified compounds have been assigned to only one kinase class (Figure 3). For the misclassified compounds, most were not assigned to any class (Figure 4). However, for the CDK1 class, where the lowest success rate (65%, Table 3) has been recorded, nearly half of the errors (8/17) have been misclassified as inhibitors of another kinase, with respectively one, one, three, and three compounds in the JNK1, P38, SYK, and EGFR classes.

Hierarchical Clustering on Daylight Fingerprints. Clustering has been performed on each separate class as well as on the whole set of 770 compounds. Using a threshold value of $T = 0.65$ led to the results given in Table 4. The compounds have been grouped by activity class, though with several clusters within a given activity class.

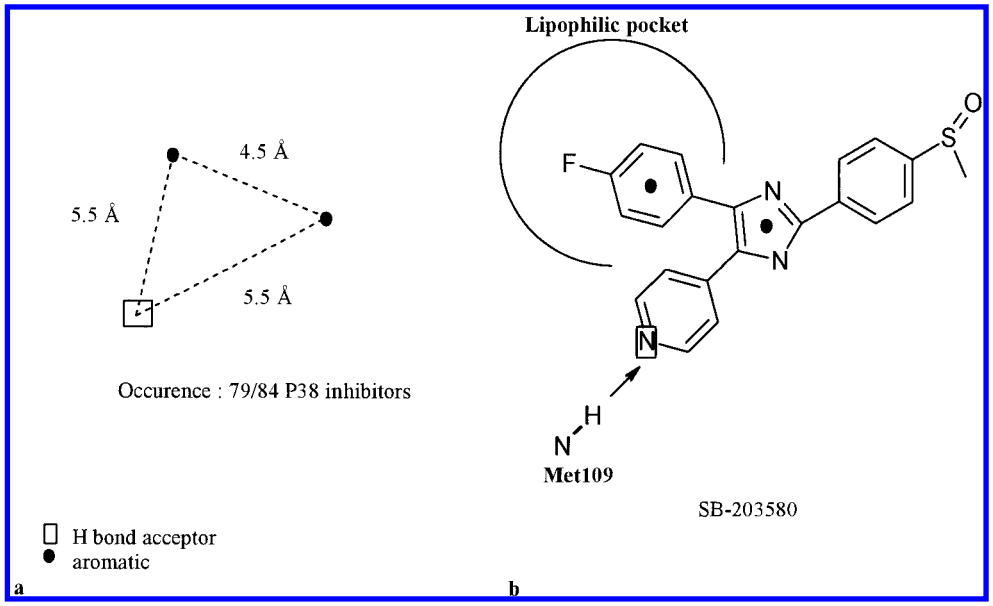

Figure 5. (a) One of the most frequent three-point pharmacophores for P38 inhibitors. (b) Main features of the interaction between SB-203580, a representative of the pyridinylimidazole class of inhibitors, and P38 (PDB code 1a9u). The pharmacophoric features shown in (a) are mapped on SB-203580.

Table 5. Pharmacophore Key Analysis

	JNK1	P38	CDK1	SYK	EGFR	All
Ncomp[a]	277	84	58	127	224	770
N3[b]	12 648	3 687	8 887	16 624	12 393	24 335
N4[b]	77 708	14 777	103 310	171 013	124 085	343 160
Max3[c](%)	99.6	100.0	93.1	77.2	98.3	74.4
Max4[c](%)	98.2	97.6	79.3	70.1	97.3	39.5

[a] Number of compounds. [b] Total number of three (four)-point pharmacophores. [c] Percentage of compounds showing the most frequent three (four)-point pharmacophore.

Multipharmacophore Descriptors Analysis. Both three- and four-point pharmacophore keys were computed for each separate class as well as for the whole set of 770 compounds (Table 5). Within each class of inhibitors, at least one three (four-)point pharmacophore occurs in most of the compounds (Table 5) and is selective compared to the other kinases. The Max 3 and Max 4 values are very similar for most of the kinases, except CDK1 and SYK (Table 5). The pharmacophore descriptors encode information relevant for ligand−protein interaction, as illustrated for a three-point pharmacophore found in 94% of the P38 inhibitors (Figure 5a). In the crystal structures of pyridinylimidazole inhibitor−P38 complexes,[61−63] the pyridine N accepts an H bond from the amide NH of Met109, while the *p*-fluorophenyl ring binds to a lipophilic pocket; the imidazole acts as a scaffold for positioning the other two rings (Figure 5b). This is consistent with literature data which show that replacing the 4-pyridine by a phenyl, a 2-pyridine, or a 3-pyridine reduces potency by more than 500-fold.[30] In addition, the *p*-fluorophenyl ring, which occupies a space near the nonconserved Thr106, confers selectivity for P38 over related MAP kinases such as the JNKs.[61,64] The three-point pharmacophore given in Figure 5a shows all the structural features required for P38 inhibition. An alternative three-point pharmacophore involving the *p*-methylsulfonylphenyl group of SB-203580 occurs less frequently (69 out of 84 compounds). Removal of this group led to only a 3-fold loss in activity.[63] The minor contribution of the *p*-methylsulfonylphenyl moiety to the binding of SB-203580 is also consistent with the weak electron density observed around this group in the crystal structure of the SB-203580/P38 complex.[63] Interestingly, P38 inhibitors where the *p*-methylsulfonylphenyl moiety has been replaced by a nonaromatic group have also been recently described.[65]

Some three- and four-point pharmacophores occur in more than one class of kinase inhibitors (Table 5). For instance, one of the most frequent three-point pharmacophores for pyridinylimidazoles (Figure 6a,b) is also found in a few EGFR inhibitors, such as PD153035 (IC_{50}(EGFR) = 29 pM, Figure 6c). While we have no P38 activity data for this compound, one of its closely related analogues inhibits P38 with an IC_{50} of 5 μM and has been crystallized with P38.[66] This 4-anilinoquinazoline inhibitor exhibits a binding mode similar (Figure 6d) to the one reported for SB-203580 (Figure 5b). However, compared with the SB-203580 *p*-fluorophenyl, the *m*-methylthiophenyl occupies a different region of the lipophilic pocket, with the methylthio group positioned in the vicinity of part of the fluorophenyl.

Comparing with the results from PLS DA, across all kinases, 75% of the compounds misclassified by PLS DA exhibit at least one of the most frequent four-point pharmacophores for the class.

Classification Exercise for Novel EGFR Inhibitors. A five-component PLS DA model (Q^2(CUM) = 0.613, R^2X(CUM) = 0.730, R^2Y(CUM) = 0.622), was derived using the full 770 compound data set. This model was then applied to 52 novel EGFR inhibitors as a test set. The model correctly classifies 48/52 of compounds as belonging to the EGFR class. One compound is classed as JNK1. In contrast, hierarchical clustering on Daylight fingerprints grouped only 15% of the test set compounds (8/52) with training set EGFR inhibitors. A threshold value of 0.65 has been considered for this cluster analysis. Both three- and four-point pharmacophore descriptors have also been computed for the 52 EGFR inhibitors (Table 6). Forty-two and 29 of the three- and four-point pharmacophores respectively which occur in at least 50% of the EGFR test compounds (N3.50 and N4.50,

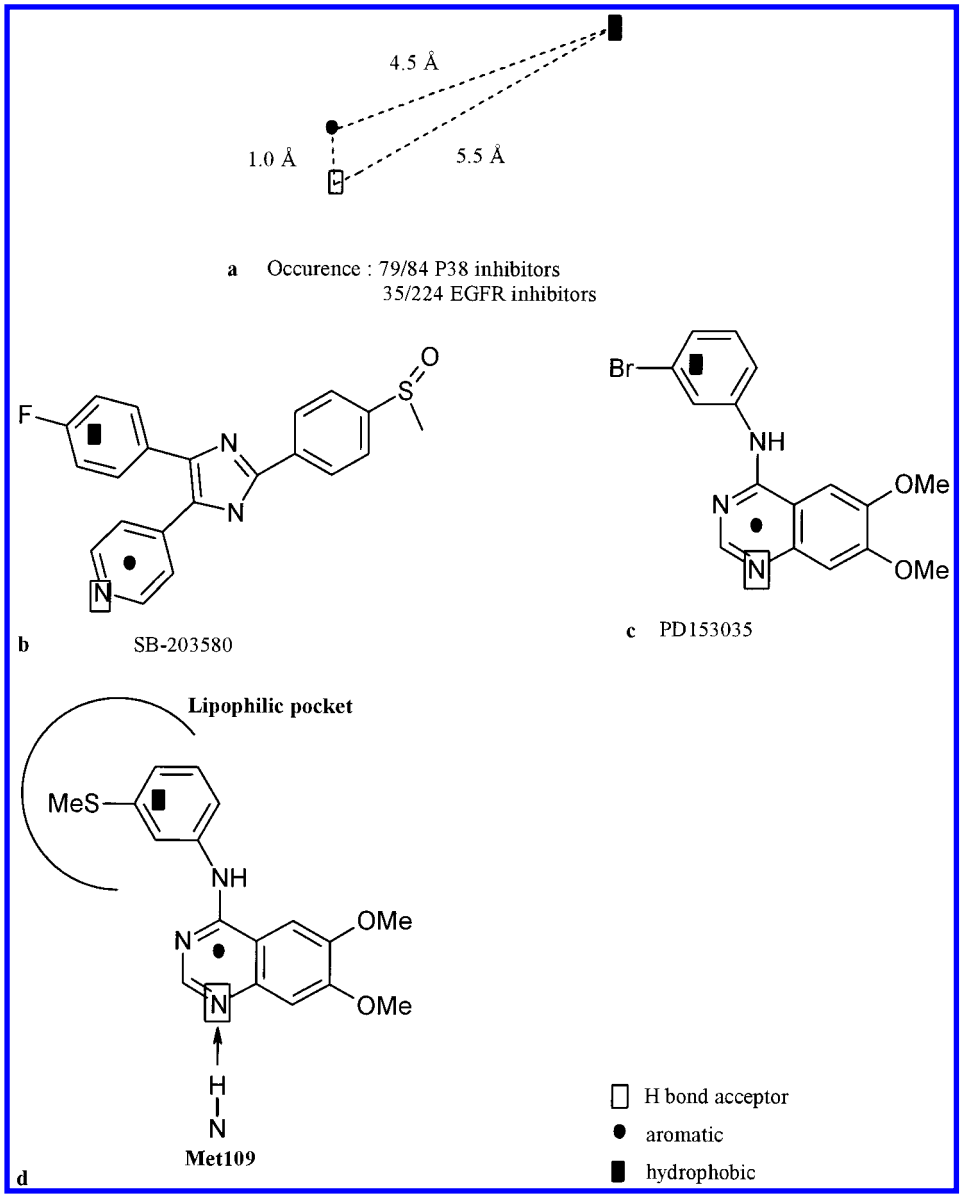

Figure 6. (a) One of the most frequent three-point pharmacophores for P38 inhibitors. (b, c) Pharmacophoric features shown in (a) mapped on SB-203580 and on PD153035. (d) Main features of the interaction between a closely related analogue of PD153035 and P38.

Table 6. Pharmacophore Key Analysis

	Ncomp[a]	N3[b]	N4[b]	Max3[c](%)	Max4[c](%)	N3.50[d]	N4.50[d]
EGFR test	52	1501	4385	100.0	98.0	54	55
EGFR train	224	12393	124085	98.3	97.3	45	37

[a] Number of compounds. [b] Total number of three (four)-point pharmacophores. [c] Percentage of compounds showing the most frequent three (four)-point pharmacophore. [d] Number of three (four)-point pharmacophores found in at least 50% of the compounds.

Table 6), have been found in the training set of 224 EGFR inhibitors. Figure 7 shows one three-point pharmacophore that occurs frequently in both the training set (87%) and the test set (81%) and examples of compounds possessing this pharmacophore. All of these compounds were assigned to the EGFR class by the PLS DA model, whereas they have been grouped in different clusters by the clustering on Daylight fingerprints.

DISCUSSION AND CONCLUSIONS

In this study we have investigated the effectiveness of BCUT descriptors for distinguishing between ligands of five different members of a protein family and compared their performance with two other types of descriptor to help gain an insight into the type of information encoded by the BCUTs. The kinases are an important target for therapeutic intervention, and finding selective inhibitors is crucial. The five proteins studied include three serine/threonine kinases and two tyrosine kinases. PCA using BCUT descriptors was able to achieve a reasonable degree of separation, particularly of JNK1 and EGFR inhibitors from the other classes. It is interesting that there is appreciable overlap between the serine/threonine kinase P38 and the tyrosine kinase SYK within the PCA model.

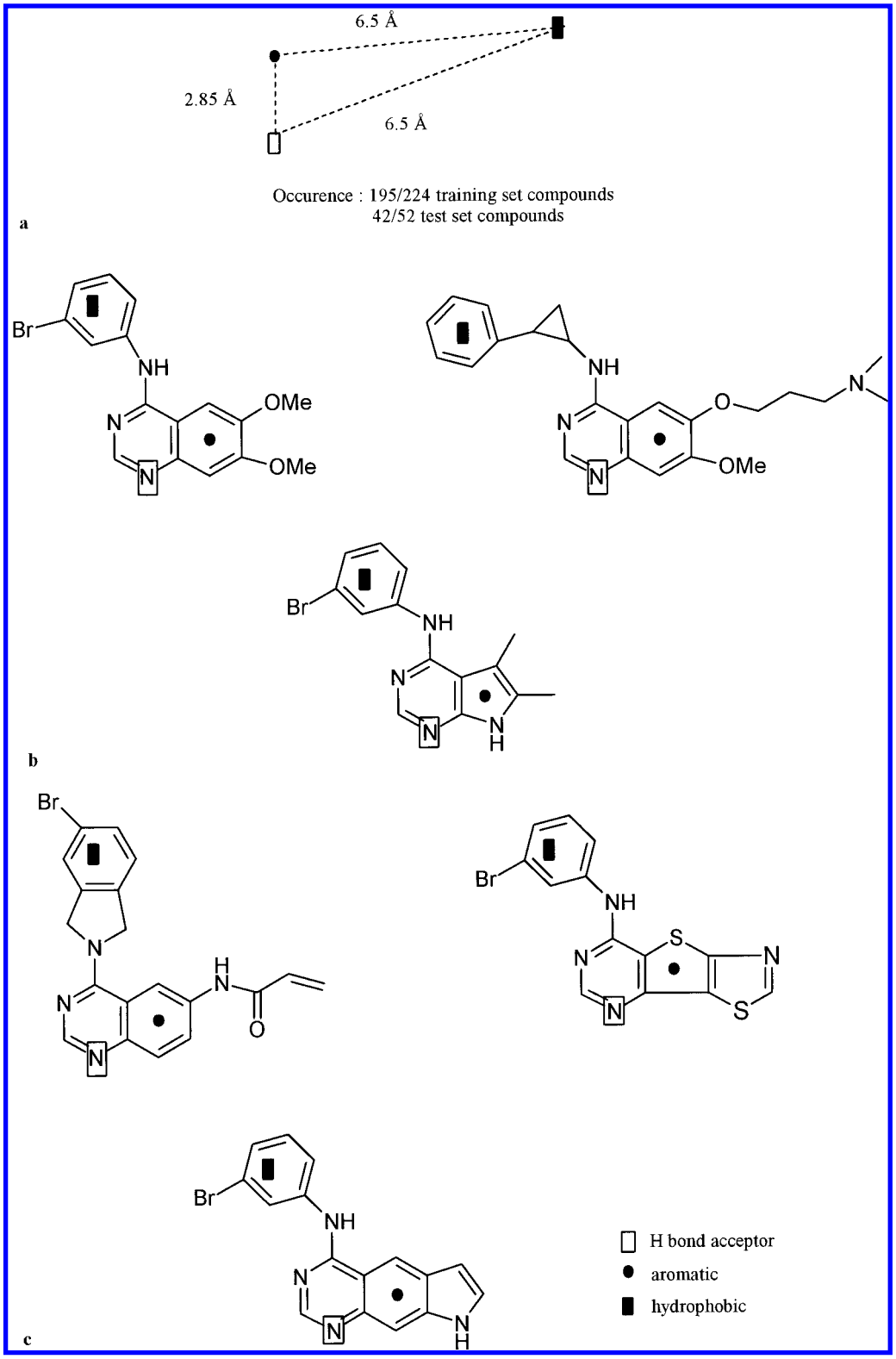

Figure 7. (a) Example of a three-point pharmacophore common to both the training and test set compounds for EGFR. (b) Training and (c) test set compounds showing this pharmacophore.

The application of a higher level chemometric tool, PLS discriminant analysis, gives very satisfactory results when applied to the BCUT descriptors. For the original 770 compound data set, the internal predictivity is good as is the performance on the test sets. The PLS DA model was also applied predictively to 52 EGFR compounds previously unseen. The model correctly classifies 48/52 of the compounds as EGFR inhibitors. To investigate whether this is due solely to some coding of 2D structure and to investigate the structural similarity between the different series of EGFR inhibitors, the compounds were clustered using Daylight fingerprints. Few of the test compounds clustered with the training set of EGFR compounds. This is an important result as it indicates that the BCUTs are encoding more than just the 2D structural properties of the molecules and indeed contain ligand−receptor relevant properties.

The information from the multipharmacophore descriptors is complementary to that obtained from the PLS DA model. Adding information from the multipharmacophore descriptors allows 75% of the compounds misclassified by the original PLS DA model to be correctly assigned. It is interesting to note that going from three- to four-point pharmacophores drastically reduces the percentage of compounds across all classes that share a common pharmacophore (last column of Table 5). However, the percentage of compounds within a class sharing a common pharmacophore is in most cases comparable between three- and four-point pharmacophores. Thus the four-point pharmacophores possess a greater selectivity between classes. Several examples of common pharmacophores are illustrated in Figures 5−7. A similar improvement in discrimination with four-point pharmacophores has been reported by Mason et al. when comparing serine protease inhibitors to the site-derived pharmacophores of the target enzymes.[11]

Our conclusion is that the BCUTs are indeed a useful set of descriptors for design tasks, extracting the information from the connection table in a manner relevant to describing ligand−receptor interactions. However, as indicated in the introduction, interpretation of results is also an issue when choosing appropriate descriptors. It is interesting to look at the loadings plots for the PLS DA model and to try to extract the properties which differentiate the ligand sets. For example, component 4 is highly correlated with the P38 discriminant. The loadings for hydrogen bond donor BCUTs are negatively correlated with this component. This makes sense as all the P38 compounds considered in this study are unusual among kinase inhibitors in only having an acceptor atom to interact with Met109 NH. Most kinase inhibitors mimic ATP to the extent that they form a bidentate donor−acceptor, acceptor−donor interaction with the protein.[34,67−69]

In summary, the BCUTs provide an interesting class of molecular descriptor which can distinguish between ligands of related proteins. While the descriptors may seem esoteric, they can in fact be interpreted in terms of features which are relevant to ligand−protein interactions. The success of the PLS DA model in the classification of unseen EGFR compounds suggests an application for an approach such as that described here in the design of targeted selective kinase libraries and virtual screening of compound collections. The BCUTs are particularly suited to such an approach as they are quick to calculate while containing more information than a standard 2D fingerprint type descriptor. In this sense they may be considered as intermediate between 2D fingerprints and the more computationally intensive 3D pharmacophore-type descriptors.

ACKNOWLEDGMENT

We thank Professor Bob Pearlman for help with the DiverseSolutions program and Dr. David E. Clark for helpful comments on the manuscript.

REFERENCES AND NOTES

(1) Agrafiotis, D. K.; Myslik, J. C.; Salemme, F. R. Advances in Diversity Profiling and Combinatorial Series Design. *Mol. Diversity* **1999**, *4*, 1−22.

(2) Mason, J. S.; Hermsmeier, M. A. Diversity Assessment. *Curr. Opin. Chem. Biol.* **1999**, *3*, 342−349.

(3) Bures, M. G.; Martin, Y. C. Computational Methods in Molecular Diversity and Combinatorial Chemistry. *Curr. Opin. Chem. Biol.* **1998**, *2*, 376−380.

(4) Lewis, R. A.; Pickett, S. D.; Clark, D. E. Computer-Aided Molecular Diversity Analysis and Combinatorial Library Design. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, in press; Vol. 16.

(5) Willett, P. Computational Tools for the Analysis of Molecular Diversity. In *Perspectives in Drug Discovery and Design*; Willett, P., Ed.; KLUWER/ESCOM: Dordrecht, 1997; Vols. 7/8, p 1.

(6) Matter, H.; Pötter, T. Comparing 3D Pharmacophore Triplets and 2D Fingerprints for Selecting Diverse Compounds Subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1211−1225.

(7) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of Molecular Diversity Descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.

(8) Cramer, R. D.; Poss, M. A.; Hermsmeier, A. M.; Caulfield, T. J.; Kowala, M. C.; Valentine, M. T. Prospective Identification of Biologically Active Structures by Topomer Shape Similarity Searching. *J. Med. Chem.* **1999**, *42*, 3919−3933.

(9) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894−2896.

(10) Schuffenhauer, A.; Gillet, V. J.; Willett, P. Similarity Searching in Files of Three-Dimensional Chemical Structures: Analysis of the BIOSTER Database Using Two-Dimensional Fingerprints and Molecular Field Descriptors. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 295−307.

(11) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-Point Pharmacophore Method for Molecular Similarity and Diversity Applications: Overview of the Method and Applications, Including a Novel Approach to the Design of Combinatorial Libraries Containing Privileged Substructures. *J. Med. Chem.* **1999**, *42*, 3251−3264.

(12) Pickett, S. D.; McLay, I. M.; Clark, D. E. Enhancing the Hit-to-Lead Properties of Lead Optimization Libraries. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 263−272.

(13) Livingstone, D. J. The Characterization of Chemical Structures Using Molecular Properties. A Survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195−209.

(14) Brown, R. D.; Martin, Y. C. Use of Structure-Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(15) Brown, R. D. Descriptors for Diversity Analysis. In *Perspectives in Drug Discovery and Design*; Willett, P., Ed.; KLUWER/ESCOM: Dordrecht, 1997; Vols. 7/8, p 31.

(16) Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. In *Perspective in Drug Discovery and Design*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds; KLUWER/ESCOM: Dordrecht, 1998; Vols. 9/10/11, p 339.

(17) Burden, F. R. Molecular Identification Number for Substructure Searches. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225−227.

(18) *DiverseSolutions*, v4.0.5; University of Texas, Austin; distributed by Tripos, Inc.: 1669 S. Hanley Rd., Suite 303, St. Louis, MO 63144.

(19) Pearlman, R. S.; Smith, K. M. Metric Validation and Receptor Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.

(20) Schnur, D. Design and Diversity Analysis of Large Combinatorial Libraries Using Cell-based Methods. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 36−45.

(21) Menard, P. R.; Mason, J. S.; Morize, I.; Bauerschmidt, S. Chemistry Space Metrics in Diversity Analysis, Library Design and Compound Selection. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1204−1213.

(22) McGregor, M. J.; Muskal, S. M. Pharmacophore Fingerprinting. 2. Application to Primary Library Design. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 117−125.

(23) McGregor, M. J.; Muskal, S. M. Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 569−574.

(24) Cohen, P. The Development and Therapeutic Potential of Protein Kinase Inhibitors. *Curr. Opin. Chem. Biol.* **1999**, *3*, 459−465.

(25) McMahon, G.; Sun, L.; Liang, C.; Tang, C. Protein Kinase Inhibitors: Structural Determinants for Target Specificity. *Curr. Opin. Drug Discovery Dev.* **1998**, *1*, 131−146.

(26) Toledo, L. M.; Lydon, N. B.; Elbaum, D. The Structure-Based Design of ATP Site-Directed Protein Kinase Inhibitors. *Curr. Med. Chem.* **1999**, *6*, 775−805.

(27) Garcia-Echeverria, C.; Traxler, P.; Evans, D. B. ATP Site-Directed Competitive and Irreversible Inhibitors of Protein Kinases. *Med. Res. Rev.* **2000**, *20*, 28−57.

(28) Adams, J. L.; Lee, D. Recent Progress towards the Identification of Selective Inhibitors of Serine/Threonine Protein Kinases. *Curr. Opin. Drug Discovery Dev.* **1999**, *2*, 96−109.

(29) de Laszlo, S. E.; Visco, D.; Agarwal, L.; Chang, L.; Chin, J.; Croft, G.; Forsyth, A.; Fletcher, D.; Frantz, B.; Hacker, C.; Hanlon, W.; Harper, C.; Kostura, M.; Li, B.; Luell, S.; MacCoss, M.; Mantlo, N.; O'Neill, E. A.; Orevillo, C.; Pang, M.; Parsons, J.; Rolando, A.; Sahly, Y.; Sidler, K.; Widmer, W. R.; O'Keefe, S. J. Pyrroles and Other Heterocycles as Inhibitors of P38 Kinase. *Bioorg. Med. Chem. Lett.* **1998**, *8*, 2689−2694.

(30) Gallagher, T. F.; Seibel, G. L.; Kassis, S.; Laydon, J. T.; Blumenthal, M. J.; Lee, J. C.; Lee, D.; Boehm, J. C.; Fier-Thompson, S. M.; Abt, J. W.; Soreson, M. E.; Smietana, J. M.; Hall, R. F.; Garigipati, R. S.; Bender, P. E.; Erhard, K. F.; Krog, A. J.; Hofmann, G. A.; Sheldrake, P. L.; McDonnell, P. C.; Kumar, S.; Young, P. R.; Adams, J. L. Regulation of Stress-Induced Cytokine Production by Pyridinylimidazoles; Inhibition of CSBP Kinase. *Bioorg. Med. Chem.* **1997**, *5*, 49−64.

(31) Chang, Y.-T.; Gray, N. S.; Rosania, G. R.; Sutherlin, D. P.; Kwon, S.; Norman, T. C.; Sarohia, R.; Leost, M.; Meijer, L.; Schultz, P. G. Synthesis and Application of Functionally Diverse 2,6,9-Trisubstituted Purine Libraries as CDK Inhibitors. *Chem. Biol.* **1999**, *6*, 361−375.

(32) Schultz, C.; Link, A.; Leost, M.; Zaharevitz, D. W.; Gussio, R.; Sausville, E. A.; Meijer, L.; Kunick, C. Paullones, a Series of Cyclin-Dependent Kinase Inhibitors: Synthesis, Evaluation of CDK1/Cyclin B Inhibiton, and *in Vitro* Antitumor Activity. *J. Med. Chem.* **1999**, *42*, 2909−2919.

(33) Traxler, P.; Green, J.; Mett, H.; Séquin, U.; Furet, P. Use of a Pharmacophore Model for the Design of EGFR Tyrosine Kinase Inhibitors: Isoflavones and 3-Phenyl-4(1H)-quinolones. *J. Med. Chem.* **1999**, *42*, 1018−1026.

(34) Trumpp-Kallmeyer, S.; Rubin, J. R.; Humblet, C.; Hamby, J. M.; Showalter, H. D. H. Development of a Binding Model to Protein Tyrosine Kinases for Substituted Pyrido[2,3-d]pyrimidine Inhibitors. *J. Med. Chem.* **1998**, *41*, 1752−1763.

(35) Rewcastle, G. W.; Murray, D. K.; Elliott, W. L.; Fry, D. W.; Howard, C. T.; Nelson, J. M.; Roberts, B. J.; Vincent, P. W.; Showalter, H. D. H.; Winters, R. T.; Denny, W. A. Tyrosine Kinase Inhibitors. 14. Structure-Activity Relationships for Methyl-amino-Substituted Derivatives of 4-[(3-Bromophenyl)amino]-6-(methylamino)-pyrido[3,4-d]-pyrimidine (PD 158780), a Potent and Specific Inhibitor of the Tyrosine Kinase Activity of Receptors for the EGF Family of Growth Factors. *J. Med. Chem.* **1998**, *41*, 742−751.

(36) Thompson, A. M.; Murray, D. K.; Elliott; W. L.; Fry, D. W.; Nelson, J. A.; Showalter, H. H. D.; Roberts, B. J.; Vincent, P. W.; Denny, W. A. Tyrosine Kinase Inhibitors. 13. Structure-Activity Relationships for Soluble 7-Substituted 4-[(3-Bromophenyl)amino]pyrido[4,3-d]-pyrimidines Designed as Inhibitors of the Tyrosine Kinase Activity of the Epidermal Growth Factor Receptor. *J. Med. Chem.* **1997**, *40*, 3915−3925.

(37) Traxler, P.; Bold, G.; Frei, J.; Lang, M.; Lydon, N.; Mett, H.; Buchdunger, E.; Meyer, T.; Mueller, M.; Furet, P. Use of a Pharmacophore Model for the Design of EGF-R Tyrosine Kinase Inhibitors: 4-(Phenylamino)pyrazolo[3,4-d]pyrimidines. *J. Med. Chem.* **1997**, *40*, 3601−3616.

(38) Rewcastle, G. W.; Bridges, A. J.; Fry, D. W.; Rubin, J. R.; Denny, W. A. Tyrosine Kinase Inhibitors. 12. Synthesis and Structure-Activity Relationships for 6-Substituted 4-(Phenylamino)pyrimido[5,4-d]pyrimidines Designed as Inhibitors of the Epidermal Growth Factor Receptor. *J. Med. Chem.* **1997**, *40*, 1820−1826.

(39) Gibson, K. H.; Grundy, W.; Godfrey, A. A.; Woodburn, J. R.; Ashton, S. E.; Curry, B. J.; Scarlett, L.; Barker, A. J.; Brown, D. S. Epidermal Growth Factor Receptor Tyrosine Kinase: Structure-Activity Relationships and Antitumor Activity of Novel Quinazolines. *Bioorg. Med. Chem. Lett.* **1997**, *7*, 2723−2728.

(40) Myers, M. R.; Setzer, N. N.; Spada, A. P.; Zulli, A. L.; Hsu, C.-Y. J.; Zilberstein, A.; Johnson, S. E.; Hook, L. E.; Jacoski, M. V. The Preparation and SAR of 4-(Anilino), 4-(Phenoxy), and 4-(Thiophenoxy)-Quinazolines: Inhibitors of $p56^{lck}$ and EGF-R Tyrosine Kinase Activity. *Bioorg. Med. Chem. Lett.* **1997**, *7*, 417−420.

(41) Myers, M. R.; Setzer, N. N.; Spada, A. P.; Persons, P. E.; Ly, C. Q.; Maguire, M. P.; Zulli, A. L.; Cheney, D. L.; Zilberstein, A.; Johnson, S. E.; Franks, C. F.; Mitchell, K. J. The Synthesis and SAR of New 4-(N-Alkyl-N-Phenyl)Amino-6,7-Dimethoxyquinazolines and 4-(N-Alkyl-N-Phenyl)amino-Pyrazolo[3,4-d]Pyrimidines, Inhibitors of CSF-1R Tyrosine Kinase Activity. *Bioorg. Med. Chem. Lett.* **1997**, *4*, 421−424.

(42) Traxler, P. M.; Furet, P.; Mett, H.; Buchdunger, E.; Meyer, T.; Lydon, N. 4-(Phenylamino)pyrrolopyrimidines: Potent and Selective, ATP Site Directed Inhibitors of EGF-Receptor Protein Tyrosine Kinase. *J. Med. Chem.* **1996**, *39*, 2285−2292.

(43) Rewcastle, G. W.; Palmer, B. D.; Thompson, A. M.; Bridges, A. J.; Cody, D. R.; Zhou, H.; Fry, D. W.; McMichael, A.; Denny, W. A. Tyrosine Kinase Inhibitors. 10. Isomeric 4-[(3-Bromophenyl)amino]-pyrido[d]-pyrimidines Are Potent ATP Binding Site Inhibitors of the Tyrosine Kinase Function of the Epidermal Growth Factor Receptor. *J. Med. Chem.* **1996**, *39*, 1823−1835.

(44) Thompson, A. M.; Bridges, A. J.; Fry, D. W.; Kraker, A. J.; Denny, W. A. Tyrosine Kinase Inhibitors. 7. 7-Amino-4-(phenylamino)-and 7-amino-4-[(phenylmethyl)amino]pyrido[4,3-d]pyrimidines: A New Class of Inhibitors of the Tyrosine Kinase Activity of the Epidermal Growth Factor Receptor. *J. Med. Chem.* **1995**, *38*, 3780−3788.

(45) Rewcastle, G. W.; Denny, W. A.; Bridges, A. J.; Zhou, H.; Cody, D. R.; McMichael, A.; Fry, D. W. Tyrosine Kinase Inhibitors. 5. Synthesis and Structure-Activity Relationship for 4-[(Phenylmethyl)amino]- and 4-(Phenylamino)quinazolines as Potent Adenosine 5′-Triphosphate Binding Site Inhibitors of the Tyrosine Kinase Domain of the Epidermal Growth Factor Receptor. *J. Med. Chem.* **1995**, *38*, 3482−3487.

(46) Showalter, H. D. H.; Bridges, A. J.; Zhou, H.; Sercel, A. D.; McMichael, A.; Fry, D. W. Tyrosine Kinase Inhibitors. 16. 6,5,6-Tricyclic Benzothieno[3,2-d]pyrimidines and Pyrimido[5,4-b]- and -[4,5-b]indoles as Potent Inhibitors of the Epidermal Growth Factor Receptor Tyrosine Kinase. *J. Med. Chem.* **1999**, *42*, 5464−5474.

(47) Smaill, J. B.; Palmer, B. D.; Rewcastle, G. W.; Denny, W. A.; McNamara, D. J.; Dobrusin, E. M.; Bridges, A. J.; Zhou, H.; Showalter, H. H. D.; Winters, T. R.; Leopold, W. R.; Fry, D. W.; Nelson, J. M.; Slintak, V.; Elliot, W. L.; Roberts, B. J.; Vincent, P. W.; Patmore, S. J. Tyrosine Kinase Inhibitors. 15. 4-(Phenylamino)quinazoline and 4-(Phenylamino)pyrido[d]pyrimidine Acrylamides as Irreversible Inhibitors of the ATP Binding Site of the Epidermal Growth Factor Receptor. *J. Med. Chem.* **1999**, *42*, 1803−1815.

(48) Weiniger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(49) Franke, R.; Gruska, A. Principle Component and Factor Analysis. In *Chemometric Methods in Molecular Design, Methods and Principle in Medicinal Chemistry*; van de Waterbeemd, H., Ed.; VCH Publishers: Weinheim, 1995; Vol. 2, p 113.

(50) Dunbar, J. B. Jr. Cluster-based selection. In *Perspectives in Drug Discovery and Design*; Willett, P., Ed.; KLUWER/ESCOM: Dordrecht, 1997; Vols. 7/8, p 51.

(51) *SIMCA 7.0*; Umetri AB: Umeå, Sweden, 1998.

(52) *Cerius^2 4.0*; Molecular Simulation Inc.: San Diego, CA, 1999.

(53) Multivariate Data Analysis with Emphasis on QSAR; Course organized by Umetri AB: Umeå, Sweden, June 26−28, 1996.

(54) *Daylight Chemical Information Systems Inc.*, version 4.51; Santa Fe, NM, 1997.

(55) Lewis, R. A. Personal communication, September 1997.

(56) Pickett, S. D.; Mason, J. S.; McLay, I. M. Diversity Profiling Using 3D Pharmacophores: Pharmacophores-Derived Queries (PDQ). *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214−1223.

(57) Pickett, S. D.; Luttmann, C.; Guerin, V.; Laoui, A.; James, E. DIVSEL and COMPLIB−Strategies for the Design and Comparison of Combinatorial Libraries Using Pharmacophoric Descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 144−150.

(58) Pearlman, R. S. Rapid Generation of High Quality Approximate 3D Molecular Structures. *Chem. Des. Auto. News.* **1987**, *2*, 1−7.

(59) Balducci, R.; McGarity, C. M.; Rusinko, A. III; Skell, J.; Smith, K.; Pearlman, R. S. *CONCORD*, v4.0.2; University of Texas, Austin; Distributed by Tripos, Inc.: 1669 S. Hanley Rd., Suite 303, St. Louis, MO 63144.

(60) *ChemDiverse*; Oxford Molecular Group plc, The Medawar Centre, Oxford Science Park, Oxford, OX4 4GA U.K.

(61) Wang, Z.; Canagarajah, B. J.; Boehm, J. C.; Kassisa, S.; Cobb, M. H.; Young, P. R.; Abdel-Meguid, S.; Adams, J. L.; Goldsmith, E. J. Structural Basis of Inhibitor Selectivity in Map Kinases. *Structure* **1998**, *6*, 1117−1128.

(62) Wilson, K. P.; McCaffrey, P. G.; Hsiao, K.; Pazharisamy, S.; Galullo, V.; Bemis, G. W.; Fitzgibbon, M. J.; Caron, P. R.; Murcko, M. A.; Su, M. S. S. The Structural Basis for the Specificity of Pyridinylimidazole Inhibitors of P38 MAP Kinase. *Chem. Biol.* **1997**, *4*, 423−431.

(63) Tong, L.; Pav, S.; White, D. M.; Rogers, S.; Crane, K. M.; Cywin, C. L.; Brown, M. L.; Pargelis, C. A. A Highly Specific Inhibitor of Human P38 MAP Kinase Binds to the ATP Pocket. *Nat. Struct. Biol.* **1997**, *4*, 311−316.

(64) Lisnock, J. M.; Tebben, A.; Frantz, B.; O'Neill, E. A.; Croft, G.; O'Keefe, S. J.; Li, B.; Hacker, C.; de Laszlo, S.; Smith, A.; Libby, B.; Liverton, N.; Hermes, J.; LoGrasso, P. Molecular Basis for P38 Protein Kinase Inhibitor Specificity. *Biochemistry* **1998**, *37*, 16573−16581.

(65) Boehm, J. C.; Adams, J. L. New Inhibitors of P38 Kinase. *Exp. Opin. Ther. Patents* **2000**, *10*, 25−37.
(66) Shewchuck, L.; Hassell, A.; Wisely, B.; Rocque, W.; Holmes, W.; Veal, J.; Kuyper, L. F. Binding Mode of the 4-Anilinoquinazoline Class of Protein Kinase Inhibitor: X-ray Crystallographic Studies of the 4-Anilinoquinazolines Bound to Cyclin-Dependent Kinase 2 and P38 Kinase. *J. Med. Chem.* **2000**, *43*, 133−138.
(67) Lamers, M. B. A. C.; Antson, A. A.; Hubbard, R. E.; Scott, R. K.; Williams, D. H. Structure of the Tyrosine Kinase Domain of C-terminal Src Kinase (CSK) in Complex with Staurosporine. *J. Mol. Biol.* **1999**, *285*, 713−725.
(68) Mohammadi, M.; Froum, S.; Hamby, J. M.; Schroeder, M. C.; Panek, R. L.; Lu, G. H.; Eliseenkova, A. V.; Green, D.; Schlessinger, J.; Hubbard, S. R. Crystal Structure of an Angiogenesis Inhibitor Bound to the FGF Receptor Tyrosine Kinase Inhibitor. *EMBO J.* **1998**, *17*, 5896−5904.
(69) Mohammadi, M.; McMahon, G.; Sun, L.; Tang, C.; Hirth, P.; Yeh, B. K.; Hubbard, S. R.; Schlessinger, J. Structures of the Tyrosine Kinase Domain of Fibroblast Growth Factor Receptor in Complex with Inhibitors. *Science* **1997**, *276*, 955−960.

CI000386X