CHEMMEDCHEM

DOI: 10.1002/cmdc.200700255

Modeling and Selection of Flexible Proteins for Structure-Based Drug Design: Backbone and Side Chain Movements in p38 MAPK

Jyothi Subramanian, Somesh Sharma, and Chandrika B-Rao*[a]

Receptor rearrangement upon ligand binding (induced fit) is a major stumbling block in docking and virtual screening. Even though numerous studies have stressed the importance of including protein flexibility in ligand docking, currently available methods provide only a partial solution to the problem. Most of these methods, being computer intensive, are often impractical to use in actual drug discovery settings. We had earlier shown that ligand-induced receptor side-chain conformational changes could be modeled statistically using data on known receptor–ligand complexes. In this paper, we show that a similar approach can be used to model more complex changes like backbone flips and loop movements. We have used p38 MAPK as a test case and have shown that a few simple structural features of ligands are sufficient to predict the induced variation in receptor conformations. Rigorous validation, both by internal resampling methods and on an external test set, corroborates this finding and demonstrates the robustness of the models. We have also compared our results with those from an earlier molecular dynamics simulation study on DFG loop conformations of p38 MAPK, and found that the results matched in the two cases. Our statistical approach enables one to predict the final ligand-induced conformation of the active site of a protein, based on a few ligand properties, prior to docking the ligand. We can do this without having to trace the step-by-step process by which this state is arrived at (as in molecular dynamics simulations), thereby drastically reducing computational effort.

Introduction

Computational methods play a crucial role in modern drug discovery projects. Both ligand and structure-based virtual screening techniques are widely used. Virtual screening of library compounds using docking and structure-based de novo drug design (SBDD) are both heavily dependent on accuracy of docking of small molecules to protein binding sites. Although the vast majority of molecular docking programs currently in vogue take into account the flexibility of the ligand, docking methods that also incorporate the flexibility of the protein, are still in their infancy and are computationally demanding.[1,2] Time and again, studies have shown that docking results are extremely sensitive to the protein conformation selected.[3,4] However, the receptor conformational changes that accompany ligand binding, ranging all the way from a local rotation of a few side chains to whole domain rearrangements, prove to be a major impediment to the development of truly predictive docking methods. Many major pharmaceutical R&D companies are actively trying to find a solution to the problem of ligand-induced conformational changes in proteins.

Techniques that are currently applied to deal with receptor flexibility during docking fall into two major classes—ensemble methods and molecular dynamics simulations (MDS). Ensemble docking methods utilize an ensemble of predefined receptor conformations. The binding energy of the ligand is either assessed against all the receptor models in a cross-docking protocol, or, multiple receptor models are averaged and the single averaged structure is used for the docking.[5,6] In MDS, the receptor conformation is allowed to change dynamically during the docking simulation. For computational tractability, the specific degrees of freedom are limited by either identifying the optimal side-chain torsional angles during the docking procedure or by using a rotamer library to represent the preferred side-chain orientations.[7,8] Recently, a new hybrid approach has been reported and implemented in the Glide software where the ligands are docked and the receptor conformations are sampled in an iterative fashion.[9,10] All these techniques are computationally demanding.

We had recently reported a new technique to model ligand-induced side-chain conformational changes in the cyclin dependent kinases (CDKs).[11] In this approach, side chains contributing to the conformational variability in the binding site were identified using receptor–ligand X-ray crystal data. Linear models were then developed to identify ligand properties that maximally influence these side-chain conformations. A few simple properties of the ligands were seen to account for more than 70% of the variation in the side-chain conformations. These models were validated and shown to be useful for predicting the best CDK crystal structure for docking of new ligands. This approach can be called quantitative structure conformation relationship (QSCR) analysis and is similar in spirit to

[a] *J. Subramanian, S. Sharma, C. B-Rao*
Nicholas Piramal Research Centre
1 Nirlon Complex, Off Western Express Highway, Goregaon(E), Mumbai-400063 (India)
Fax: (+91) 22-3081-8036
E-mail: chandrika@nicholaspiramal.co.in

Supporting information for this article is available on the WWW under http://www.chemmedchem.org or from the author.

the QSAR approach used for predicting ligand activity. The difference is that we are predicting the effect of ligand on structural conformation of key residues in the active site of the protein instead of the biological activity of the ligand.

As protein side chains typically exhibit greater mobility than the backbone, modeling side-chain flexibility was an important first step towards modeling receptor flexibility. However, it is generally agreed that predicting backbone level conformational changes is more complex, and in this paper, we examine whether a similar statistical approach based on available ligand-receptor data could also be used for modeling backbone level receptor conformational changes.

p38 mitogen activated protein kinase (p38 MAPK), belonging to the class of serine-threonine MAP kinases, seemed to be a good system to address this question. Small molecule inhibition of p38 MAPK has emerged as a promising therapeutic strategy for the treatment of inflammatory diseases such as rheumatoid arthritis as well as other diseases such as cancer, diabetes, atherosclerosis, and Alzheimer's disease. Recently, new classes of p38 MAPK inhibitors have been identified and the structural basis of the inhibition has been reviewed.[12]

Structurally, p38 MAPK is folded into the bilobal structure typical of most protein kinases (Figure 1). The ATP binding site is found in the deep hydrophobic cleft between the two lobes. Previous evidence from X-ray crystallographic structures of other kinases suggests that the adenine ring of the ATP directly interacts with the p38 hinge region residues His 107 and Met 109 to form a pair of hydrogen bonds. ATP competitive inhibitors mimic the binding of ATP while also taking advantage of additional binding regions that are not utilized by ATP.

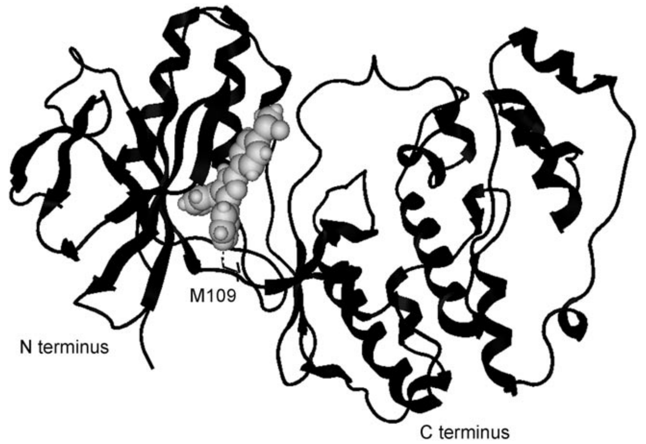


Figure 1. Structure of p38 MAPK bound to SB203580, a small molecule inhibitor (PDBID: 1A9U).[14] SB203580 is depicted as a spacefill model. The hinge region residue Met 109 is shown.

In the process of identification of potent and selective p38 MAPK inhibitors, it was found that[13] some inhibitors induce a peptide flip of the Gly 110 residue at the hinge region that enables an additional H-bond interaction with Gly 110. The ability of an inhibitor to induce a Gly 110 peptide flip upon binding results in an improved selectivity of the inhibitor for p38 over other MAP kinases like extracellular signal-regulated protein kinases (ERK) and c-Jun N-terminal kinase (JNK) that have larger residues than Gly at the 110 position. Presence of larger residues makes the required peptide flip upon binding more energetically unfavorable.[13] The second major conformational change induced by some p38 MAPK inhibitors is the structural rearrangement of the kinase's conserved Asp-Phe-Gly (DFG) motif, shifting the phenylalanine (Phe 169) side chain from its usual, buried location (the DFG-in conformation), to a location ~10 Å away that sterically interferes with ATP binding (the DFG-out conformation). This movement of the Phe 169 side chain leaves a vacant hydrophobic pocket that could be filled by the hydrophobic groups in the inhibitors. Inhibitors that induce a DFG flip tend to show very high affinity and exhibit slow binding kinetics relative to other p38 inhibitors.[14] Apart from these backbone level conformational changes, binding of some inhibitors cause a major change in the conformation of the Tyr 35 side chain.[15] The extent of variation in ligand bound p38 MAPK crystal structures can be seen in the figure in the table of contents graphic.

The correlation between the physicochemical and structural descriptors of the p38 MAPK ligands and the conformational changes in p38 MAPK receptor induced by these ligands is the subject of this study. Using ligand–receptor co-crystal data, these conformational changes are statistically modeled as functions of ligand descriptors. Through rigorous validation, the robustness and predictive power of the models is demonstrated. It is proposed that these models can be used to predict the probable conformational changes on binding to new ligands and hence can be used to select the best p38 MAPK crystal structure for docking of new ligands.

Materials and Methods

Protein and ligand structures

Crystal structures of the protein–ligand complexes used in this study were obtained from the protein data bank (PDB). Twenty five p38 MAPK crystal structures were downloaded. These included all wild-type human p38α protein structures that were co-crystallized with ligands. Structures with mutations were not considered. The crystal structure resolution varied from 1.75 Å to 2.80 Å. The set of protein structures used in this study is given in Table S1, Supporting Information along with the crystal structure resolutions, the original references, and the chemical structures of the bound ligands. The status of the DFG loop and of Gly 110 (flip/no flip) is also mentioned in this table.

Protein and ligand preparation.

The protein and ligand structures were prepared as per standard procedures[16] using MOE 2005-06 software.[17] When the p38 MAPK crystal structure contained multiple chains, only one of the chains involved in ligand binding was retained. Solvent and small molecules other than the ligand were removed. The ligand was corrected for any structural errors. Hydrogens were added to all the atoms and the structures of the protein–ligand complexes were energy minimized using the CHARMM[18] force field after fixing all nonhydrogen atoms to their crystallographic positions. The minimization was carried out to an RMS gradient of 0.01 kcal mol^{-1}. From each protein–ligand complex, the ligands were then extracted and stored in an MOE 2005-06 database for further analysis.

Conformational differences in the binding site

Without loss of generality, the twenty five p38 MAPK structures were brought to a common frame of reference by aligning to the 1A9U p38 MAPK structure, which was arbitrarily taken as the standard. The alignment was done on the basis of all protein residues using MOE 2005-06 software. The binding site variations in the different co-crystal structures were compared. The major backbone level conformational changes in the ATP binding site of p38 MAPK—the Gly110 flip and the flip in the DFG loop conformation—are shown in Figure 2.

To quantify the backbone flip in the hinge region, the backbone ψ-angle of Met 109 (ψ_{109}) was used. Negative values of ψ_{109} indicate a flip in the backbone whereas positive values of ψ_{109} indicate absence of a backbone flip. To quantify the changes in the DFG loop conformation, two atoms that lie on either side of this loop—the C$_{\alpha}$s of Lys 53 and Asn 114—were chosen. These residues were chosen as reference points because their positions were found to be invariant in the set of p38 MAPK structures. The C$_{\alpha}$s of the residues were chosen so as to prevent confounding effects from minor ($<$1 Å) side-chain movements. The distances between these atoms and the C4 carbon of Phe 169 provided a numerical measure of variability for the shift in the DFG loop conformation and the relative positioning of the Phe 169 side chain. The Phe 169 residue was chosen because of its bulk. Shifting of this residue creates space that could potentially be filled with ligand atoms. The example shown in Figure 3 a illustrates the differences in the distances for the DFG-in and the DFG-out conformations. The average Lys 53–Phe 169 (LysPhe) distances are 15.51 Å and 9.68 Å for the DFG-in and DFG-out conformations respectively.

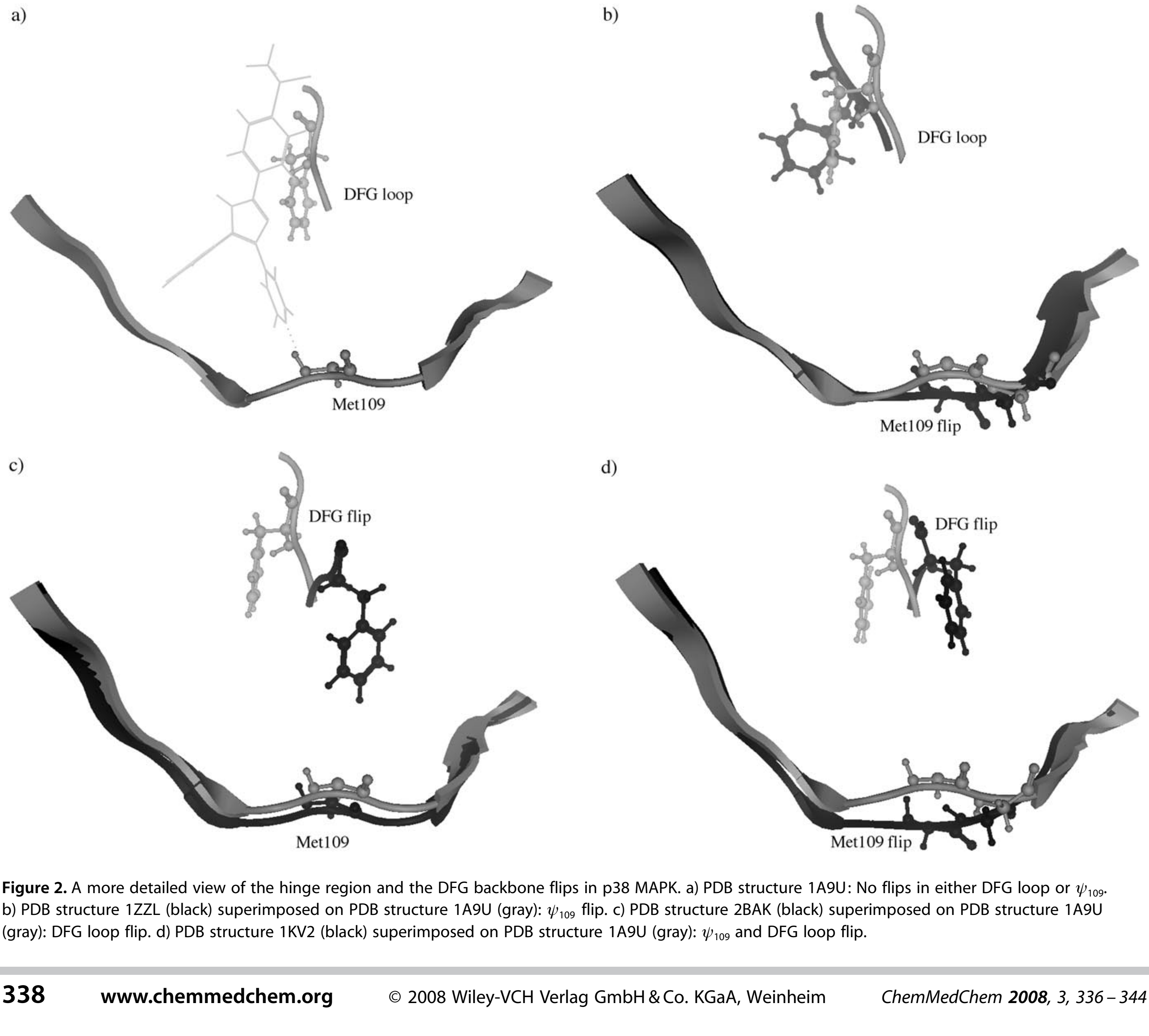


Figure 2. A more detailed view of the hinge region and the DFG backbone flips in p38 MAPK. a) PDB structure 1A9U: No flips in either DFG loop or ψ_{109}. b) PDB structure 1ZZL (black) superimposed on PDB structure 1A9U (gray): ψ_{109} flip. c) PDB structure 2BAK (black) superimposed on PDB structure 1A9U (gray): DFG loop flip. d) PDB structure 1KV2 (black) superimposed on PDB structure 1A9U (gray): ψ_{109} and DFG loop flip.

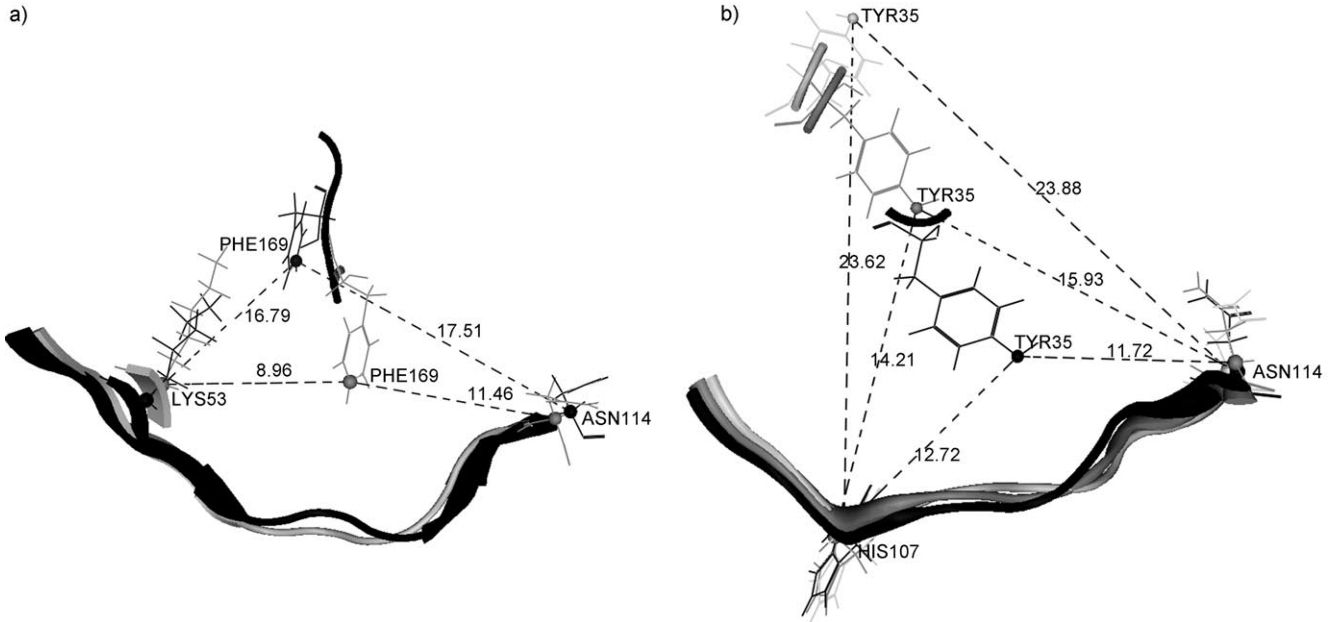


Figure 3. Calculation of inter-residue distances to quantify conformational changes in p38 MAPK. a) Two p38 MAPK structures - DFG-in (black) and DFG-out (gray) showing variation in the distances due to movement of the DFG loop. b) Three p38 MAPK structures showing the variation in distances due to changes in the Tyr 35 conformation.

The average Asn 114–Phe 169 (AsnPhe) distances are 18.54 Å and 11.54 Å for the DFG-in and DFG-out conformations, respectively.

With regard to side chains, Tyr 35 shows significant conformational variability. To quantify this variability, again two invariant atoms, the C_αs of His 107 and Asn 114, were chosen and the distances between these atoms and the OH of Tyr 35 were measured. The example shown in Figure 3 b illustrates the variation in the His 107–Tyr 35 (HisTyr) and Asn 114–Tyr 35 (AsnTyr) distances for different positions of the Tyr 35 side chain.

Descriptor calculations

To avoid any bias in descriptor selection, all 2D descriptors, as well as all 3D descriptors that do not depend on the frame of reference of the molecule, were calculated for all the ligands using the descriptor calculation module of MOE 2005-06 software. This set of 241 descriptors[19] is the same set that was used in our earlier study[11] and includes physical property descriptors, subdivided surface area descriptors, atom and bond count descriptors, Kier and Hall connectivity descriptors, Kappa shape indices, adjacency and distance matrix descriptors, pharmacophore feature descriptors, partial charge descriptors, potential energy descriptors, surface area, volume and shape descriptors, and conformation-dependent charge descriptors. Additionally, the 166 MACCS (Molecular ACCess System) keys[20] representing various structural features of the molecules were also calculated.

Model building

The correlation coefficients between the descriptors and the dihedral ψ_{109}, LysPhe, AsnPhe, HisTyr, and AsnTyr distances were computed. For each of these variables, descriptors showing less than 50 % correlation were discarded. In the resulting set of descriptors, if two descriptors had more than 85 % correlation, one of them was removed.

To find the optimum combination of descriptors that could explain most of the variation in the inter-residue distances, regression models connecting the inter-residue distances and ψ_{109} with the ligand descriptors were fitted. In each case two models were fitted, one with the best correlated topological descriptors and the other with the best correlated MACCS keys. The models were derived using stepwise selection of variables. From a full model including all descriptors selected from the previous paragraph, at each step, descriptors with SE-scaled coefficients having values less than 0.01 were excluded from the model if they caused a decrease in cross-validated R^2, till there were no more variables with such coefficients.

Nonlinear regression methods were used where linear methods performed poorly.

Results

Conformational analysis of the active site

The distance variations due to conformational changes in the active site are summarized in Figure 4. Structures with DFG-out conformation are found to have small LysPhe and AsnPhe distances and large HisTyr and AsnTyr distances. But the structures that are in DFG-in conformation are found to have large

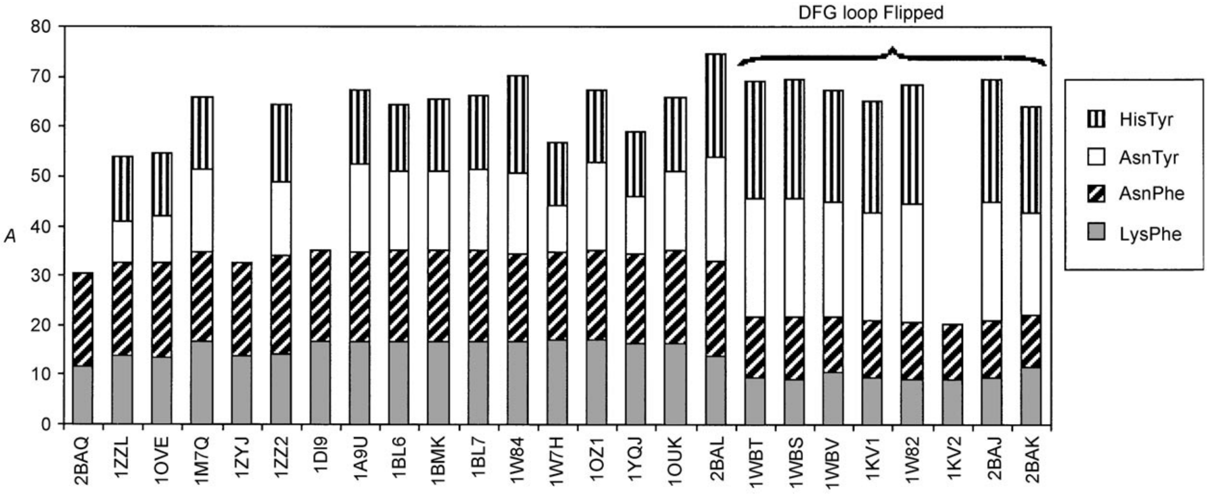


Figure 4. Variation in the distances corresponding to changes in the backbone and side-chain conformations. The lines for Tyr 35 related distances show breaks because the Tyr 35 side chain position was not elucidated in the corresponding crystal structures.

LysPhe and AsnPhe distances and variable HisTyr and AsnTyr distances. There is no association seen between either the DFG conformation or the Tyr 35 conformation with the hinge region peptide (Gly 110) conformation.

Correlation between ligand structure and binding-site conformation. Models were built using both the topological descriptors and the MACCS keys. As models based on MACCS keys gave better model-fit statistics overall, only the results from MACCS keys are presented in the main paper. The results of the model fit for topological descriptors are given in the Supporting Information (Tables S3 and S4). MACCS keys with more than 50 % correlation to the binding site parameters are shown in Table 1. The optimal regression models were derived using these descriptors by stepwise selection. It was found that linear models were sufficient to explain the variation in the active site distances. However, in the case of the peptide dihedral, linear models did not give good predictions. Hence, the hinge region peptide dihedral was coded as a binary variable, with 1 indicating a flip ($\psi_{109} < 0$) and, 0 indicating no flip ($\psi_{109} > 0$). Descriptors were then used to discriminate between these two possibilities using a nonlinear binary QSAR (MOE 2005.06)[21] model. The regression models are shown in Table 2. The structural features encoded by MACCS keys occurring in the models are summarized in Table 3.

Validation

To study the sensitivity of the models to changes in the training set data, a rigorous resampling-based cross-validation procedure was adopted.[22] Leave-many-out cross-validations were performed by leaving out a random selection of 1/6th, 1/3rd, and $^1/_2$ of the data set. Regression models were built each time from the descriptors in Table 2. The results of these cross validations are given in Table 4 and Table 5.

For further validation, five p38 MAPK crystal structures that were not used in the training set were downloaded from the PDB. The PDB IDs of this set of protein structures is given in Table S2 of the Supporting Information along with their crystal structure resolutions, the original references, and the status of the DFG loop and hinge region peptide. This table also gives the chemical structures of the bound ligands. As MACCS keys gave better models than the physicochemical and topological descriptors (compare R^2 and q^2 in

Table 1. MACCS fingerprints with more than 50 % correlation.

Correlation	Fingerprints correlated with AsnPhe distance	Fingerprints correlated with LysPhe distance	Fingerprints correlated with AsnTyr distance	Fingerprints correlated with HisTyr distance	Fingerprints correlated with ψ_{109} flip
(0.6, 0.75]			MACCS(133)	MACCS(133)	
(0.5, 0.6]	MACCS(84)		MACCS(151), MACCS(156)	MACCS(66), MACCS(110), MACCS(132), MACCS(151), MACCS(154)	
[−0.6, −0.5)	MACCS(37), MACCS(57), MACCS(95), MACCS(149), MACCS(151), MACCS(160)	MACCS(57), MACCS(43), MACCS(109), MACCS(131), MACCS(149), MACCS(156), MACCS(158)	MACCS(144)	MACCS(62)	
[−0.75, −0.6)	MACCS(135), MACCS(66), MACCS(43), MACCS(109), MACCS(110), MACCS(133), MACCS(156)	MACCS(110), MACCS(133), MACCS(135), MACCS(117), MACCS(154), MACCS(66)			MACCS(154), MACCS(117), MACCS(110)

Table 2. Best linear model for inter-residue distances based on correlated MACCS fingerprints.[a]

Best model	$R^2(q^2)$	Relative importance
LysPhe = 17.25−3.29*MACCS(110)−1.40*MACCS(133)	0.73 (0.68)	MACCS(110) (1.00), MACCS(133) (0.87)
AsnPhe = 18.82−1.85*MACCS(110)−1.54*MACCS(109)−4.32*MACCS(66)	0.79 (0.62)	MACCS(66) (1.00), MACCS(109) (0.99), MACCS(110) (0.58)
AsnTyr = 13.30 + 1.28*MACCS(156) + 1.66*MACCS(133)−1.33*MACCS(144)	0.69 (0.55)	MACCS(144) (1.00), MACCS(156) (0.72), MACCS(133) (0.63)
HisTyr = 14.52 + 9.13*MACCS(66) + 3.19*MACCS(132)	0.78 (0.70)	MACCS(66) (1.00), MACCS(132) (0.95)

[a] A binary discrimination model was used to classify ψ_{109} as flipped/not flipped. MACCS(154) and MACCS(117) were used to discriminate between presence and absence of a flip. The accuracy of the prediction was 92% and the cross-validated accuracy was also 92%.

Table 3. Structural definition of the descriptors.

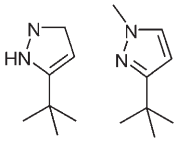
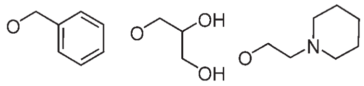
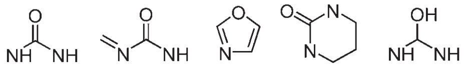
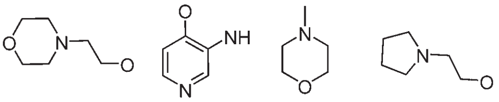
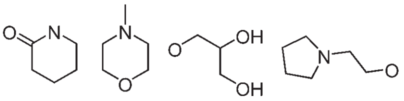
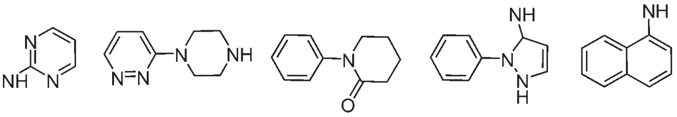
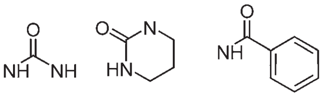
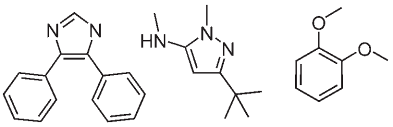
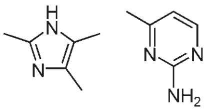
Descriptor	Explanation	Sample structural features present in inhibitors
MACCS(66)	#CX4 bonded to>=3 carbons	
MACCS(109)	#O attached to CH2	
MACCS(110)	#O 1C from an N	
MACCS(117)	#N 2C from an O	
MACCS(132)	#O 2 bonds from CH2	
MACCS(133)	#N nonring bonded to a ring	
MACCS(154)	#O in C=O	
MACCS(144)	#atoms separated by (!:):(!:)	
MACCS(156)	#XN where coordination number of X>=3	

Table S4 and Table 2), validation was carried out only for the models developed using MACCS keys. The active site distances and ψ_{109} class were predicted for these new structures using the models in Table 2. The predictions are given in Table 6.

Discussion

Numerous computational studies have demonstrated that the results of docking and virtual screening are extremely sensitive to the conformation of protein used.[2,3,4] However, due to the large protein conformational space to be sampled and/or dependence on biophysical theory to determine conformational changes, currently available methods that account for protein flexibility in the docking protocol, like ensemble docking and MDS, are computer intensive. Ensemble methods also suffer from the disadvantage of not providing protein conformations specific to the query ligand. Our method presented in this paper can be used along with MDS to reduce the computational effort and obtain more reliable results.

Table 4. Cross-validation results for the prediction of distances related to DFG flip and Tyr35 side-chain movement.

Model	Data points left out	Cross-validated R^2	RMSE [Å]
LysPhe	1	0.68	1.75
	1/6th	0.69	1.70
	1/3rd	0.66	1.77
	1/2	0.66	1.79
AsnPhe	1	0.62	2.12
	1/6th	0.72	1.75
	1/3rd	0.71	1.76
	1/2	0.71	1.78
AsnTyr	1	0.55	3.32
	1/6th	0.53	3.35
	1/3rd	0.57	3.21
	1/2	0.33	4.03
HisTyr	1	0.70	2.39
	1/6th	0.75	2.18
	1/3rd	0.72	2.28
	1/2	−0.54	5.40

Table 5. Cross validation results for the prediction of ψ_{109} peptide flip.

Data Points Left out	Accuracy [%]		
	Total	Flip	No Flip
1	92	83	100
1/6th	92	83	92
1/3rd	88	83	92
1/2	88	92	85

Table 6. Predictions for the test set data along with RMSEs of distance predictions and accuracy of peptide flip predictions.

PDB ID	LysPhe [Å]	AsnPhe [Å]	AsnTyr [Å]	HisTyr [Å]	Peptide Flip
	(Actual)Pred	(Actual)Pred	(Actual)Pred	(Actual)Pred	(Actual)Pred
1OUY	(16.47)	(18.24)	(16.28)	(15.83)	(1)
	12.55	16.97	14.97	10.36	0
1W83	(11.61)	(10.46)	(25.50)	(23.57)	(0)
	11.15	12.34	21.83	19.40	0
1WBN	(12.45)	(10.04)	(25.27)	(23.45)	(0)
	11.15	11.11	21.67	23.65	0
1WBW	(16.27)	(18.42)	(9.20)	(12.48)	(1)
	15.84	17.28	15.36	13.26	1
2GFS	(16.17)	(17.74)	(8.58)	(13.04)	(0)
	14.44	15.74	21.70	21.38	1
	RMSE = 2.02 Å	RMSE = 1.52 Å	RMSE = 6.9 Å	RMSE = 5.02 Å	Accuracy = 60 %

Working on the assumption that similar ligands induce similar conformational changes in the receptor, we had earlier developed a novel technique based on known ligand–receptor data and applied it to quantify ligand-induced side chain conformational changes in the ATP binding site of CDKs. The predicted conformational changes were validated using test data sets and molecular simulations, which showed good agreement with the predictions. This approach was also a computationally inexpensive and reliable way to identify optimal crystal structures for docking new ligands.[11] In the present study, we have extended our findings and demonstrated that a similar approach can also be used to handle the more complex problem of predicting backbone-level conformational changes. This has been illustrated using the prediction of ligand-induced conformational changes in p38 MAPK. Our results show that all the significant conformational changes taking place on ligand binding to p38 MAPK—the peptide flip in the hinge region, the flip in the DFG loop, and the movement of the Tyr 35 side chain—can be modeled by this technique in a reliable and computationally inexpensive manner.

Among the methods that attempt to solve the problem of receptor flexibility, only ensemble methods that take into consideration multiple protein conformations (for example, FlexE[5] or IFREDA[23]) and to a lesser extent, MDS, can account for backbone conformational changes. However, the problems of choice of structures and that of combining the results from multiple dockings are not yet optimally addressed in ensemble methods. Moreover, for the specific case of p38 MAPK, it was found that the hinge region peptide flip was not reproduced by IFREDA, possibly due to the large potential barrier of this flip.[23]

We have used distances from invariant atoms in the active site to quantify the backbone conformational changes. The variations in these distances are then correlated to the variations in the physicochemical and structural properties of the ligands. To avoid any bias in the analysis, we started with a large set of descriptors and identified crucial descriptors that are well correlated with the variation in the receptor conformational changes by stepwise selection of descriptors. Comparison of the regression models built using physicochemical descriptors with the models built using structural descriptors provide evidence that structural descriptors are better correlated with the conformational variations in p38MAPK. This is in contrast to what we had observed in the case of CDKs (unpublished data).

The structural fingerprints MACCS(110) and MACCS(133) were identified to be the most important for explaining the variation in the LysPhe distances and the structural fingerprints MACCS(66), MACCS(109), and MACCS(110) were identified to be the most important for explaining the variation in the AsnPhe distances. As these descriptors have a negative coefficient in the linear model (Table 2), we conclude that these structural features are important in making the conformational change from DFG-in to DFG-out in p38 MAPK. A comparison of these structural features with the crystal structure binding mode of the ligands reveals that when the receptor is in the DFG-out conformation, the structural feature encoded by MACCS(110) is important for making H-bond interactions with the Asp 168 residue and the structural features encoded by

MACCS(133) and MACCS(66) are important for hydrophobic interactions in the space created by the movement of the Phe 169 residue. The structures encoded by MACCS(109) are seen to make interactions in the solvent exposed portions of the receptor.

The structural fingerprints MACCS(156), MACCS(133), and MACCS(144) were found to be significant for explaining the variation in the AspTyr distance and the structural fingerprints MACCS(66) and MACCS(132) were found to be significant for explaining the variation in the HisTyr distance. We had earlier noted the association between the DFG-out conformation and the position of the Tyr 35 side chain. It is probably this association that is reflected by the two common descriptors (MACCS-(133), MACCS(66)) for the prediction of the DFG-loop conformations and the prediction of the Tyr35 side-chain conformation.

For the prediction of the flip in the hinge region Gly 110 peptide, MACCS(154) and MACCS(117) were found to be important. These descriptors pertain to the formation of an additional H-bond interaction with the backbone of Gly 110.

The models described in this report were validated using resampling based cross-validation and also on an independent test set. In the case of the models for DFG flip and hinge region peptide flip, the results from the cross-validations (Table 4 and Table 5) are robust with respect to changes in the training set. However, the models for the Tyr 35 side chain show poor validation statistics on successive removal of chunks of data (Table 4). This phenomenon can be attributed to the high conformational variability of this side chain—as more and more of the data points are removed, the data set becomes increasingly nonrepresentative. As noted before (Figure 4), this variability in the Tyr 35 side-chain position is especially high for the p38 MAPK structures that have the DFG-in conformation. The relatively poor predictability for Tyr 35 is also reflected in the results of the test set data (Table 6) where we see that the AsnTyr and HisTyr predictions show higher RMSE as compared to LysPhe or AsnPhe predictions. We also note from this table that for the p38 MAPK structure 2GFS, with poor predictions for the HisTyr and AsnTyr distances, the DFG loop is in an "in" conformation.

In a recently published molecular dynamics study of p38 MAPK, the movement of the DFG loop on ligand binding was analyzed. It was observed in this study that apart from the commonly found DFG-in and DFG-out conformations, other stable intermediate conformations are also formed.[24] The authors refer to these conformations as pseudo DFG-in and pseudo DFG-out conformations. Such conformations are actually seen in the set of crystal structures (see table of contents graphic). In our method, instead of considering just two discrete cases of DFG-in and DFG-out, we consider the actual positional variation of the Phe side chain and hence we would be able to recover the intermediate conformations also.

As with any statistical model fitting, the model developed here is likely to be sensitive to the number and quality of protein crystal structures in the training set. Quality of training set includes not only statistical issues but also issues involved in the reliability of crystal structures. The quality and reliability of the model is expected to improve with more and better quality ligand–receptor data. The advantage of our method is that given the 3D structure data on a few ligand-bound protein complexes, one can predict the binding site geometry for any other ligand. Determining the mobile and invariant atoms in the protein is the most time-consuming step in the model building process. Once this decision is made, the subsequent steps such as minimization of the protein–ligand complex, calculation of descriptors, construction and validation of regression models do not take more than a couple of days on a Xeon workstation. Prediction of the conformational changes for a new ligand based on the model is almost instantaneous.

It is well known that the results of molecular dynamics simulations are very much dependent on the starting conformation of the protein–ligand complex. Our method would be useful in determining a good protein conformation that is likely to be induced by a new ligand. The ligand can then be docked keeping this protein conformation rigid. This can be followed by dynamics simulations to refine the poses, taking structural and energetic considerations, and effects of solvent and ions, into account. This computational sequence would help us first fix the approximate crystal structure for docking, the docking would take care of the ligand conformational changes as a result of binding to protein and the final step of MDS would take care of any additional factors affecting conformational changes in the receptor. This computational scheme includes the possibility of prediction of novel binding modes for new ligands. Thus, the present work, while giving an insight into the ligand-based determinants of induced-fit, also offers a novel and computationally efficient way of dealing with receptor-flexibility during docking and structure-based ligand design.

Keywords: bioinformatics · drug design · p38 MAPK · predictive models · protein flexibility

[1] E. M. Krovat, T. Steindl, T. Langer, *Curr. Comput.-Aided Drug Des.* **2005**, *1*, 93–102.
[2] H. A. Carlson, *Curr. Opin. Chem. Biol.* **2002**, *6*, 447–452.
[3] C. W. Murray, C. A. Baxter, *J. Comput.-Aided Mol. Des.* **1999**, *13*, 547–562.
[4] M. P. Thomas, C. McInnes, *J. Med. Chem.* **2006**, *49*, 92–104.
[5] H. Clausen, C. Buning, M. Rarey, T. Lengauer, *J. Mol. Biol.* **2001**, *308*, 377–395.
[6] S. Y. Huang, X. Zou, *Proteins Struct. Funct. Genet.* **2007**, *66*, 399–421.
[7] H. Alonso, A. A. Bliznyuk, J. E. Gready, *Med. Res. Rev.* **2006**, *26*, 531–568.
[8] D. Sivanesan, R. V. Rajnarayanan, J. Doherty, N. Pattabiraman, *J. Comput.-Aided Mol. Des.* **2005**, *19*, 213–228.
[9] W. Sherman, T. Day, M. P. Jacobson, R. A. Friesner, R. Farid, *J. Med. Chem.* **2006**, *49*, 534–553.
[10] H. Alonso, A. A. Bliznyuk, J. E. Gready, *Med. Res. Rev.* **2006**, *26*, 531–568.
[11] J. Subramanian, S. Sharma, C. B-Rao, *J. Med. Chem.* **2006**, *49*, 5434–5441.
[12] S. T. Wrobleski, A. M. Doweyko, *Curr. Top. Med. Chem.* **2005**, *5*, 1005–1016.
[13] C. E. Fitzgerald, S. B. Patel, J. W. Becker, P. M. Cameron, D. Zaller, V. B. Pikounis, S. J. O'Keefe, G. Scapin, *Nat. Struct. Biol.* **2003**, *10*, 764–769.
[14] C. Pargellis, L. Tong, L. Churchill, P. F. Cirillo, T. Gilmore, A. G. Graham, P. M. Grob, E. R. Hickey, N. Moss, S. Pav, J. Regan, *Nat. Struct. Biol.* **2002**, *9*, 268–272.
[15] Z. Wang, B. J. Canagarajah, J. C. Boehm, S. Kassisa, M. H. Cobb, P. R. Young, S. Abdel-Meguid, J. L. Adams, E. J. Goldsmith, *Structure* **1998**, *6*, 1117–1128.
[16] M. Kontoyianni, L. M. McClellan, G. S. Sokol, *J. Med. Chem.* **2004**, *47*, 558–565.

[17] Chemical Computing Group. www.chemcomp.com.

[18] A. D. MacKerell, Jr., M. Feig, C. L. Brooks III, *J. Comput. Chem.* **2004**, *25*, 1400–1415.

[19] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, **2000**.

[20] J. L. Durant, B. A. Leland, D. R. Henry, J. G. Nourse, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.

[21] P. Labute, *Pac. Symp. Biocomput.* **1999**, *4*, 444–455.

[22] S. Geisser, *J. Am. Stat. Assoc.*. **1975**, *70*, 320–328.

[23] C. N. Cavasotto, R. A. Abagyan, *J. Mol. Biol.* **2004**, *337*, 209–225.

[24] T. Frembgen-Kesner, A. H. Elcock, *J. Mol. Biol.* **2006**, *359*, 202–214.

Received: September 14, 2007
Revised: October 15, 2007
Published online on December 14, 2007