JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

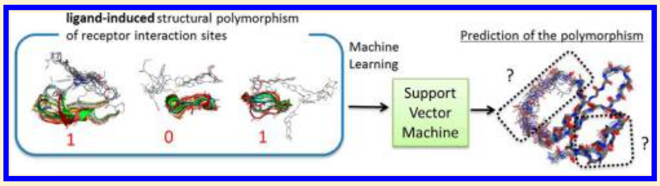
Prediction of Ligand-Induced Structural Polymorphism of Receptor Interaction Sites Using Machine Learning

Daisuke Takaya,[†] Tomohiro Sato,[†] Hitomi Yuki,[†] Shunta Sasaki,[†] Akiko Tanaka,[†] Shigeyuki Yokoyama,[†,‡] and Teruki Honma[†,*]

[†]RIKEN Systems and Structural Biology Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan

[‡]Department of Biophysics and Biochemistry, Graduate School of Science, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

S *Supporting Information*


ligand-induced structural polymorphism
of receptor interaction sites
1
0
1
Machine
Learning
Support
Vector
Machine
Prediction of the polymorphism
?
?

ABSTRACT: Protein functions are closely related to their three-dimensional structures. Various degrees of conformational changes in the main and side chains occur when binding with other molecules, such as small ligands or proteins. The ligand-induced structural polymorphism of proteins is also referred to as "induced-fit", and it plays an important role in the recognition of a particular class of ligands as well as in signal transduction. We have developed new prediction models that discriminate conformationally fluctuant residues caused by ligand-binding. The training and test data sets were obtained from the Protein Data Bank. The induced-fit residues were judged based on the Z values of the Cα atom distances in each protein cluster. Moreover, we introduced various descriptors, such as the number of residues, accessible surface area (ASA), depth of the residue, and position-specific scoring matrix (PSSM), which were obtained from the 2D- or 3D-structural information for the protein. After the optimization of the parameters by 5-fold cross validation, the best prediction model was applied to some well-known induced-fit target proteins to verify its effectiveness. Especially in the validation for the DFG motif of a protein kinase family, we succeeded in the prediction of the DFG-out possibility from only the DFG-in conformation of each kinase structure.

■ BACKGROUND

Protein functions are closely related to the structural features derived from the three-dimensional coordinates.[1] Therefore, it is important to experimentally determine the protein structure coordinates by using biological methods, such as X-ray crystallographic analysis and NMR spectroscopy. The experimentally determined structures are registered in the Protein Data Bank.[2] In general, proteins bearing the same amino acid sequence should form the same three-dimensional structure in the same solution environment. However, when binding other molecules, such as small ligands or proteins, various levels of conformational changes in the main and side chains have been observed. This phenomenon is called "induced-fit", and it plays important roles in signal transduction and specific ligand recognition. For example, it is well-known that some protein kinases, such as Abl and p38, recognize the Type II kinase inhibitors with an unusual conformational change of a Phe residue in the DFG motif and/or the G-loop.[3] In addition, the structural change of the binding loop of Eglin C plays an important role in the inhibition of trypsin.[4] The domain structure of the glutamine-binding protein changes in the presence of glutamine.[5] Thus, structural changes are often observed in analyses of protein function and intracellular signaling.

To investigate protein functions accompanied by dynamic structural changes, experimental protein structure data including multiple conformations, such as NMR solution structures and X-ray crystal structures determined under different conditions, should be obtained. In terms of computational simulations, to predict or analyze the motion of a protein structure, molecular dynamics (MD) simulation is one of the more powerful methods. MD has been utilized for various purposes, including conformational searches as well as calculations of the binding free energy of ligands.[6,7] However, a MD simulation requires huge computation power and a lot of time. At least a 10 nano to micro second order is needed to sufficiently simulate a dynamic structural change of a protein, and the duration of the simulation requires weeks to months, with hundreds of CPU cores.[8,9] In addition, performing adequate MD simulations is not a straightforward endeavor because suitable parameter selections and setups of calculation conditions, such as the preparation of an initial structure, are required.

Under such circumstances, in a protein–ligand docking study, the flexibility of the receptor is represented by reducing the repulsion term of the protein–ligand interaction in the scoring functions, such as the modified Lennard–Jones potential term (called a soft protein surface)[10] or by using a protein structure ensemble. If the conformational change is very small, then the soft protein surface works; however, the method cannot manage large conformational changes, such as the DFG-loop alteration in protein kinases.[11–13] In the conventional docking programs, the soft surface is applied to all protein atoms, but this is rather

Received: September 26, 2012

problematic for the rigid parts of protein atoms in some docking programs, such as GLIDE version 5.7 in the Maestro 2011 suite.[14] In contrast, the ensemble method can consider a large conformational change. However, to sufficiently consider all conformational changes, vast numbers of conformations must be prepared as the protein structure ensemble for many binding site residues. To efficiently use the residue-based soft surface and ensemble methods for induced-fit docking, predictions of the flexibility of each receptor residue or atom are critically important for efficient induced-fit docking. Recently, we proposed a new docking system called GENIUS,[15] which uses multiple protein coordinates, such as NMR structures. On the basis of the clustering results of the NMR structures, the atom−atom collision term was implemented to represent the induced-fit of the receptor. The GENIUS docking system was applied to the in silico screening of HCV NS3-4A protease, and we successfully discovered nonpeptidic inhibitors considering induced-fit.[15] In GENIUS, at least two experimentally determined coordinates are required to employ the induced-fit mode. In these induced-fit docking programs, it is quite useful to predict the flexible receptor region (induced-fit region) from only one receptor structure and to apply a soft van der Waals potential or a low atom−atom collision term to the region. In this study, we developed a prediction method for induced-fit regions to facilitate investigations of protein function and in silico screening.

Recently, statistical prediction models based on experimental databases, such as the Protein Data Bank, have been reported in the field of protein functional prediction. For example, discrimination models of the binding sites between protein−protein,[16] protein−RNA,[17] or protein−GTP[18] based on the three-dimensional structures have been reported. In the prediction models, high predictive accuracy (Protein−protein: sensitivity = 60.6%, specificity = 53.4%, MCC = 0.243. Protein−RNA: MCC 0.31 to 0.45. Protein-GTP: MCC = 0.70, precision = 0.93, recall = 0.73, ACC = 83.98%) was achieved. To construct accurate prediction models, the preparation of sufficiently diverse data sets and selection of suitable descriptors are critically important. The success of these previous analyses suggested that the data sets obtained from the experimental data in the PDB have the potential to predict phenomena related to structure−function relationships.

In the statistical prediction field, machine learning methods, such as support vector machine (SVM), random forest, and naive Bayesian classifier, have been preferentially used. Among these methods, SVM is regarded as one of the most popular and effective machine learning algorithms for pattern recognition, classification, and regression. SVM models can nonlinearly discriminate two classes of the prediction state by mapping data vectors to a very high-dimensional descriptor space and finding the hyperplane that separates the two classes with the largest margin.[19] In the field of protein function prediction, many prediction methods based on these machine learning techniques have been applied, including the above-mentioned binding site prediction[16−18] and the protein fold recognition problems.[20]

For RNA binding site prediction, Kumar and co-workers built prediction models based on the RNA binding residues obtained from the Protein Data Bank. In addition, the scalar values derived from the PSSM (position specific scoring matrix[21]) were used as the descriptors of the SVM input data. The RNA binding residues were predicted by using overlapping patterns of different sequence lengths (window sizes) for each sequence. In addition, the GTP binding site prediction was also performed using similar descriptors. For the prediction of protein−protein interaction regions, Nan and co-workers built prediction models based on the properties of the protein−protein interaction residues (core interface residues). Upon construction of the prediction model, the scalar values derived from the PSSM and physico−chemistry properties (i.e., number of ionized atoms, hydrophobicity, and solvent accessible area) were used as descriptors for the SVM model. These studies commonly employed PSSM as the alternative indicator to measure the amino acid similarities. In fact, Kumar and co-workers reported that the prediction performance using the PSSM descriptor is better than that using simple sequence identity for RNA-binding site prediction. In the case of the prediction of a protein−protein interaction site on a flexible loop, structures, including Ramachandran angles, crystallographic B-factors, and relative accessible surface area, were used as the descriptors.[22]

In this study, we constructed discrimination models of receptor atom motions caused by interactions with ligands (induced-fit region) using machine learning techniques. Atomic motion occurs at both the side chains and main chains of a protein. Because the side chain flexibility is largely affected by the main chain flexibility, we first started to construct the main chain flexibility prediction models. We defined the criterion of the induced-fitting residues, based on the Z-score of the RMSD of the Cα atoms, calculated from the structural ensemble of the same protein. Subsequently, the descriptors for model building were optimized to maximize the prediction performance. Finally, the constructed models were applied to well-known induced-fit proteins, such as the kinase family with the reported characteristic conformational changes called “DFG-out” and “DFG-in”, protein−protein interactions (PPI), and GPCR to assess the effectiveness of our proposed prediction model.

■ MATERIALS AND METHODS

Overview. This study aimed to predict the induced-fit-involved residues and the non-induced-fit residues using the protein coordinates, such as a PDB formatted file, as input. Machine learning is one of the most powerful methods based on experimental data if the training data set is sufficient. Therefore, the collection of a suitable data set is important for accurate prediction. First, all of the protein coordinates were collected from the PDB. These coordinates were clustered for classifying induced-fit and non-induced-fit residues using the Z-scores based on the Cα atom RMSD in the same position of the amino acid sequence alignment. Second, the labeled data set was divided into a training set and test set. The training set was used for the parameter optimization and prediction model construction, and the test set was used for the prediction performance assessment.

Preparation of the Protein Data Set Involved in Induced-Fit. The PDB contains structures determined by X-ray crystal analyses, NMR, and electron microscopy. Among them, because the NMR structures were determined under conditions in an aqueous solution, the structural fluctuation would be large. Therefore, only the structures determined by X-ray crystal structure analyses were used to explicitly determine the flexibility of the residues. As of April 8, 2010, 149,012 three-dimensional structures were extracted from the PDB. To remove the redundant proteins from the data set, representative subsets were acquired based on PISCES.[23] PISCES is a protein sequence culling algorithm to produce subsets of sequences from huge data sets, such as the PDB. The following PISCES conditions were used: nonredundant by 95% sequence homology, resolution ≤ 2.0 Å, and R factor ≤ 0.3. This process generated 10,285 representative clusters. Moreover, for each cluster, the chains consisting of more than 500 residues or including fewer than 5 protein structures were discarded from each subset. In this study, to detect the ligand-binding site, the clusters that did not include any ligand molecule were also discarded from the data set. Apo structures were permitted if the cluster included at least one ligand-binding state protein. The ligand molecules were required to satisfy the conditions that the molecular weight was less than 600 but greater than 150, and the number of heavy atoms was not less than 10. Under these conditions, 404 clusters were obtained from the initial clusters. The obtained clusters were checked to ensure that proteins with 95% sequence identity or better would not be present in the other clusters (6637 chains). This was called the "NR95 data set" in this study. On the other hand, to avoid unfair prediction dependent on merely the sequence identity, the cluster sets in which the sequence identity did not exceed 30%, called the "NR30 data set" (268 clusters, 4426 chains), was prepared for the validation of the prediction model. A representative protein for each cluster was selected, based on the number of residues missing in the cluster member; that is, the protein structure with the smallest number of missing residues was chosen as the representative. In this study, "missing residue" means a residue that partially or totally lacks atomic coordinates because of experimental issues.

Criteria for Detection of the Induced-Fit Residues. All of the amino acid residues in a protein structure move to some degree under various conditions because of inherent flexibility and disordered regions as well as ligand binding. However, our goal was to limit the protein movement to ligand-induced structural polymorphisms. Among the movements, we defined "induced-fit movement" by the following method. First, on the basis of the sequence-based alignment (using MOE[24]), each cluster member protein, except for the representative protein, was fitted to the representative one, and the ligand binding site was detected. To estimate how much flexibility the residue has in the clustered proteins, we introduced a normalized indicator, rather than an absolute value of the geometric distance. The normalization enables the detection of relatively more flexible regions in the protein cluster than the average flexibility of the protein cluster. Because the average flexibility levels of the protein clusters are quite different from each other, the normalized indicator can adequately detect significantly flexible regions on the whole protein structure surface. The normalized indicator was defined to estimate the flexibility of the i-th residue ($1 \leq i \leq N$), where N means the number of residues of the representative protein. For the i-th residue, the average of all possible distances among the Cα atoms in the same alignment position was calculated (dist_i). The determination of an induced-fit residue was performed for the residues within 8 Å around all of the superposed ligands in the coordinate system. Using the dist_i values, the Z-score of the i-th residue (Z-score_i) was calculated according to the following formula

$$\text{Z-score}_i = (\text{dist}_i - \text{dist}_{\text{mean}})/\text{dist}_{\text{stddev}}$$

where $\text{dist}_{\text{mean}}$ means the average of the distribution of dist_i in the alignment, and $\text{dist}_{\text{stddev}}$ means the standard deviation of the dist_i. The optimal threshold of the Z-score was determined by the following procedure. First, for well-known protein−ligand complexes with flexible loops that cause induced fit, we detected the residues with a Z-score above a certain criterion and visually inspected whether the remarkable residues matched the experimentally observed flexible loops. Moreover, we considered whether the distribution of the categorized residues in the condition of the tentative Z-score satisfied its biological correspondence.

Machine Learning. In this study, SVM$^{\text{light}}$[25] was used for SVM model building to predict the induced-fit residues. SVM$^{\text{light}}$ is one of the implementations of support vector machine developed by Joachimes et al. In the construction of models by SVM$^{\text{light}}$, some kernels have already been prepared, such as linear, polynomial, RBF (radial basis function), and sigmoid. Because the prediction performance was comparatively better than the other kernels in our preliminary examinations, the RBF kernel was used in this study. In addition, the RBF kernel has many configurable parameters that affect the prediction performance, such as the gamma and cost factors. We determined the optimized parameters by testing combinations of these parameters for various ranges of $0.0001 \leq \text{gamma} \leq 1$ and $1 \leq \text{cost factor} \leq 50$).

Setting of the Residue Window as the Prediction Unit. To predict the induced-fit residues, an objective central residue was defined in turn. The sequentially continuous residues around the central residue were defined as the prediction unit (called the "window" in Figure 1). The descriptors used for the SVM input data were calculated per the window unit. This method is similar to those reported by Kumar et al.[17] and Nan et al.[16] To improve the predictive performance, a suitable window size was determined because the number of descriptors depends on the window size. Next, the descriptors based on the physicochemical properties and the coordinates of the residues (including the $\text{AW}_{\text{central}}$ residue in Figure 1) were calculated. For example, Kumar and co-workers used the sequentially continuous residues (AW_n) as the prediction unit.[17] On the other hand, Li's group calculated the descriptors of the structurally nearest neighbor residues (AE_n) around the residues in the window.[16] Therefore, in our prediction model, the descriptors were also calculated for the nearest residue (AE_n) from the geometric centroid of the AW_n residue in the window.

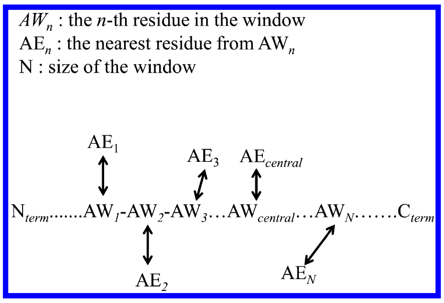

AW_n : the n-th residue in the window
AE_n : the nearest residue from AW_n
N : size of the window
AE_1
AE_3
$AE_{central}$
N_{term}.......AW_1-AW_2-AW_3...$AW_{central}$...AW_N.......C_{term}
AE_2
AE_N

Figure 1. Prediction objective residue ($\text{AW}_{\text{central}}$) and prediction unit for descriptor calculation. The descriptors were calculated for all of the residues (AW_n) in the windows (AW_N). In the case where the 3D-extension mode is valid, the descriptors are calculated for the extended residues (AE_N).

Selection of Descriptors. In this study, for the query sequence, the prediction model can calculate a predicted value for each residue. The adopted descriptors for model building were mainly divided into two classes. First, a physicochemical property based on the residue unit was used to describe the physicochemical environment. For example, the physicochemical properties include the number of hydrophobic bonds and hydrogen bonds around the objective residue. Second, a statistical property based on an exhaustive amino acid database, such as NCBI NR, was used. In this case, only the position specific scoring matrix (PSSM[21]) was used as the statistical property. The PSSM descriptors express amino acid sequence conserved qualities. All of the descriptors were calculated using the computational functions implemented by MOE[24] and PSI-BLAST[21] and are described below. An abbreviated word in parentheses means that the symbols are used subsequently in figures in the Results and Discussion Section.

1. PSSM Obtained from NCBI NR and PSI-BLAST (PSSM). PSSM was calculated for each residue in the query sequence. PSSM shows the mutation frequency of each amino acid residue at the specific position in the arrangement. It was previously reported that PSSM is an alternative descriptor substituting for the frequencies of amino acid residues.[26] Thus, 20 descriptors were calculated for one residue. In the PSI-BLAST search process, the default parameters were used, except the number of iterations was set to 3. The non-redundant NCBI NR created by CD-HIT[27] was used for the database.

2. Accessible Surface Area (ASA). The 2D descriptors calculated from the amino acid sequence are independent from the protein coordinates. However, the 3D descriptors depend on the protein coordinates, and can describe structurally unique features, as compared with the 2D descriptors. Generally, the proportion of water exposure varied in each residue environment. Therefore, the exposed surface area was calculated for each residue in the query sequence. Thus, one descriptor was calculated for one residue. The value was divided by the maximum value for standardization between 0 and 1.

3. Depth of Residue in the Protein (depth). In a previous protein sequence alignment study, an indicator of the depth of an amino acid residue in the protein surface improved the performance of the alignment.[28] Therefore, we introduced the depth parameter as the descriptor. We simply calculated the Euclidean distance between the gravity point of the residue and the gravity point of all residues. One descriptor was calculated for each residue. The residues with small depth values tended to be located close to the core. The values were divided by the maximum value for standardization between 0 and 1.

4. Secondary Structure Probability (SSPROB). The amino acid residues in a folded protein form a characteristic hydrogen bond network known as the "secondary structure", such as an α-helix and a β-sheet. These characteristic structures influence the protein stability. Therefore, the secondary structure class of the residues in the query sequence was calculated based on Thompson et al.[29] and implemented in MOE. Thus, three descriptors (i.e., probability of α-helix, β-sheet, and none) were calculated for each residue. The values were divided by the maximum value for standardization between 0 and 1.

5. Hydrophobicity of Residue (HYD). The hydrophobicity of a residue is physically important in protein folding and stability.[30] Therefore, the hydrophobicity of the residue in the query sequence was introduced. The hydrophobic scale developed by Eisenberg et al.[31] was calculated for each residue by MOE. Thus, one descriptor was calculated for each residue. The values were divided by the maximum value for standardization between 0 and 1.

6. Number of Residues Around the Objective Residue (count residue). The shape of a protein is defined by the 3D structure. Similar flexible regions tend to form similar shapes. Therefore, we took the local shape of the residue into consideration. We simply counted the number of hydrophobic, hydrophilic, and other residues within 8.0 Å from the gravity point of the residue. If the residue has a larger value, then the residue is located in a densely surrounded region of the protein, such as the core. When the hydrophobic scale value was more than 0.5, the residue was classified as a hydrophobic residue. In addition, if the value was less than −0.5, then the residue was classified as a hydrophilic residue. In this study, Ala, Ile, Leu, Met, Phe, Trp, and Val were judged as hydrophobic residues. Arg, Asn, Asp, Gln, Glu, and Lys were judged as hydrophilic residues. Cys, Gly, His, Pro, Ser, Thr, and Tyr were judged as neutral residues. Thus, three descriptors were calculated for each residue. The values were divided by the maximum value for standardization between 0 and 1.

7. Cavity Around the Residue (CAV). To estimate the concave cavity around the residue, dummy atoms were generated in the receptor coordinates by using the MOE Alpha site finder. The default conditions for the generation of the dummy atoms were used. The number of dummy atoms that were within 3.4 Å from the gravity point of the heavy atom of the residue was included in the descriptor. Thus, if the residues were located in a deep position in a cavity of the receptor structure, then they would contact many dummy atoms. One value was calculated for each residue. The value was divided by the maximum value for standardization between 0 and 1.

8. Charge of the Residue (charge). Electrostatic interactions contribute to the stability of protein folding. To estimate the influence of the electrostatic interaction, the Gasteiger charge was assigned to each atom in the objective residues. In this study, each descriptor was calculated per residue. The average value of the charge was calculated as the electrostatic interaction energy. Thus, one value was calculated for each residue. The value was divided by the maximum value for standardization between 0 and 1.

9. Molecular Weight of the Residue (MW). The molecular weights of the heavy atoms of each amino acid residue in the query sequence were summed. Thus, one value was calculated for each residue. The value was divided by the maximum value for standardization between 0 and 1.

10. Temperature Factor of the Residue (tempfactor). The temperature factor experimentally obtained by the X-ray crystal structure analysis expresses the mobility in the crystal lattice. Because it is likely that the residues with high temperature factors would have greater probabilities of participating in induced-fit, the values were added to the descriptors. The average of the temperature factors of the atoms of the objective residue was calculated for each residue. Thus, one value was calculated for each residue. The value was divided by the maximum value for standardization between 0 and 1.

11. Number of Rotatable Bonds in the Residue (RB) and Number of Bonds (adeg). Rotatable bonds are related to the flexibility of the residue and entropic contribution. Therefore, the number of rotatable bonds was introduced to the descriptors. The summation of the rotatable bonds for each residue was included in the descriptor. Thus, one value was calculated for each residue, as "rb". Moreover, the number of bonds formed by each heavy atom of the residue was included in the descriptor. Thus, one value was calculated for each residue as "adeg". These values were divided by the maximum value for standardization between 0 and 1.

12. Number of Aromatic Ring Systems (ringbond). Hydrophobic effects play a role in protein stability. Therefore, the influence of aromatic rings was introduced to estimate these interactions. For a simple, qualitative estimation, the number of residues bearing an aromatic ring system, such as Phe, Tyr, and Trp, was included in the descriptor. Thus, one descriptor was calculated for each residue. The value was divided by the maximum value for standardization between 0 and 1.

13. Connectivity Index (chideg). The Chi connectivity index developed by Kier and Hall[32,33] was calculated for each residue by using MOE. The value was divided by the maximum value for standardization between 0 and 1.

Descriptors Based on Window Unit. The descriptors mentioned in the above section were calculated per residue unit. However, cross-sectional descriptors also exist. For example, the dihedral angle consisting of four atoms across residues is a typical representative descriptor. Thus, the following cross-sectional descriptors based on the window unit were introduced to our descriptors.

14. Number of Hydrogen Bonds in the Window Unit (FRG_H). It is supposed that relatively rigid loop structures would be formed by hydrogen bonding interactions in the segment, as compared to the free coil states. Moreover, the loop structure would contribute to the local stability. Therefore, the number of hydrogen bonds within the fragment was calculated for each window unit. The number of descriptors was same as the number of residues in the window. The value was divided by the maximum value of each descriptor for standardization between 0 and 1.

15. Torsion Angle in the Window Unit (FRG_TOR). It assumed that the torsion angles (i.e., ϕ, φ, ω) have characteristic tendencies based on the flexibility of the fragment in a similar manner to rotatable bonds. Thus, the torsion angles in the segment were included in the descriptors. Three values were calculated for each residue. The values were divided by the maximum value of each descriptor for standardization between 0 and 1.

3D Extension Descriptors. In the previously mentioned cases, only the sequentially continuous residues were considered for the calculation of the descriptors. However, sequentially separated but neighboring residues in the 3D coordinates can influence the flexibility of the target residue. Therefore, to extend our descriptor set from 2D to 3D, the same number of residues as that of the window unit was considered as the target residues to calculate the descriptors. The extended residues were chosen based on the closest distance from each residue in the window unit. The calculated descriptors of the extended residues were same as those previously mentioned in the "Descriptors Based on Window Unit" section (Figure 1).

Training Set and Test Set for Prediction. The data including the residues that satisfied the induced-fit residue criteria were categorized as the positive cases (2982 residues) for machine learning. In addition, the other data were used as the negative cases (15,736 residues). From the negative data, the same number as the positive cases was randomly extracted. The positive cases and extracted negative cases were randomly divided into five subsets. For optimizing the SVM parameters (i.e., gamma and cost factors), one subset was chosen for the test set, and the remaining four subsets were used as the training set. Moreover, to determine the threshold of the SVM model, 5-fold cross validation was performed using the obtained SVM parameters with the best prediction performance in this optimization step. We used MCC (described in the next section) as the measure of performance. All of the prediction performances, including averages and SD values, were calculated by 5-fold cross validation.

Measures for Induced-Fit Prediction Models. To evaluate the prediction performance of the discrimination models, kappa statistics,[34] MCC (Matthew's correlation coefficient),[35] SN (sensitivity), SP (specificity), ACC (accuracy), PPV (positive predictive value), and NPV (negative predictive value) were employed. These quality measurements have been widely used in other machine learning research.[17] The definitions of these measures are provided below

$$\kappa = \frac{\mathrm{Pr}(o) - \mathrm{Pr}(e)}{1 - \mathrm{Pr}(e)}$$

where

$$\mathrm{Pr}(o) = \frac{(\mathrm{TP} + \mathrm{TN})}{N}$$

$$\mathrm{Pr}(e) = \frac{(\mathrm{TN} + \mathrm{FN})(\mathrm{TN} + FP) + (\mathrm{FP} + \mathrm{TP})(\mathrm{FN} + \mathrm{TP})}{N^2}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FN})(\mathrm{TN} + \mathrm{FN})}}$$

$$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}$$

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}$$

Preparation of Well-Known Flexible Proteins for External Validation. To estimate the performance of our prediction models, well-known induced-fit proteins were manually collected. One of the most famous protein groups is the protein kinases, which are involved in phosphorylation to control the functions of their target proteins. Therefore, the MOE.2010 Kinase Database Search[24] was used for selecting the kinase targets for validation. This database provided information about the conserved regions (G-Loop, hinge, alpha C Helix, DFG-motif) based on the kinase coordinates and alignments registered in the PDB. Moreover, the structural classification of the DFG motif (i.e., DFG-out or DFG-in) was also assigned under the condition that if the atoms of the Phe residue in the DFG motif were within 5.5 Å of the alpha C helix, then it was classified as the DFG-in state. Samples of the kinase family proteins that can adopt the DFG-out and DFG-in states were searched using this database. The other applicable proteins, GPCR and Eglin-C, were selected by visual inspection of the targets and the literature.

■ RESULTS AND DISCUSSION

The scheme of the construction and validation of the prediction model is summarized in Figure 2. First, the whole PDB data set was prepared. The data set was clustered by sequence identity to determine the criterion of whether induced-fit occurs or not. Subsequently, the clustered data set was used for the training of SVM discrimination models. The SVM descriptors were classified into two categories, the physicochemical properties (hydrogen bonds, hydrophobicity, etc.) and statistical properties

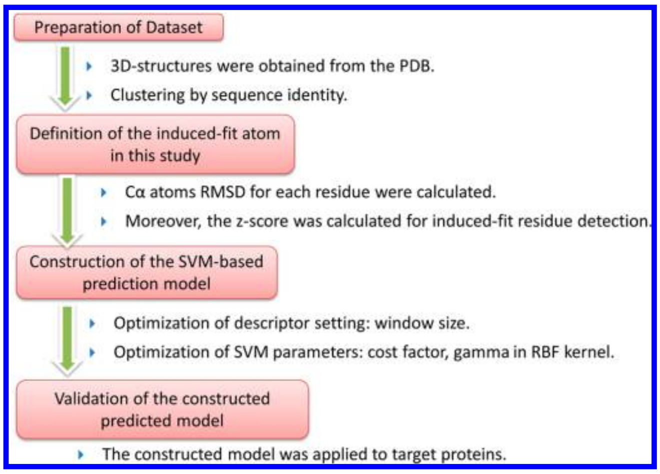

Preparation of Dataset
3D-structures were obtained from the PDB.
Clustering by sequence identity.
Definition of the induced-fit atom in this study
Cα atoms RMSD for each residue were calculated.
Moreover, the z-score was calculated for induced-fit residue detection.
Construction of the SVM-based prediction model
Optimization of descriptor setting: window size.
Optimization of SVM parameters: cost factor, gamma in RBF kernel.
Validation of the constructed predicted model
The constructed model was applied to target proteins.

Figure 2. Overview of the construction of the prediction model and its application in this study.

related to sequence (PSSM), as specified in the Materials and Methods section. The SVM parameters, such as cost factor and gamma, were optimized to maximize the prediction performance, such as ACC and MCC. Finally, the best constructed model was applied to well-known induced-fit targets.

Application of Criteria for Induced-Fit Residues. In consideration of a protein structure, to obtain the appropriate Z-score value for the judgment of induced-fit, we looked at the regions identified by the Z-score values in each cluster for well-known induced-fit proteins. In the data set, under the previously mentioned conditions, the p38 MAP kinase cluster included 20 protein structures. The kinase family members maintain their flexibility especially in the ligand-binding region, corresponding to each kind of subfamily. Therefore, the kinase family was a suitable target for visual inspection.

In the p38 MAP kinase cluster (Figure 3), three regions are shown, under the conditions that the Z-score ≥ 2.0 and 1.0 ≤ Z-score < 2.0, respectively. Our visual inspection revealed that the ensemble in the (c) motif was more rigid than the other two motifs. Moreover, motifs (a) and (b) obviously formed flexible loops and were observed to undergo significant conformational changes upon ligand interaction.

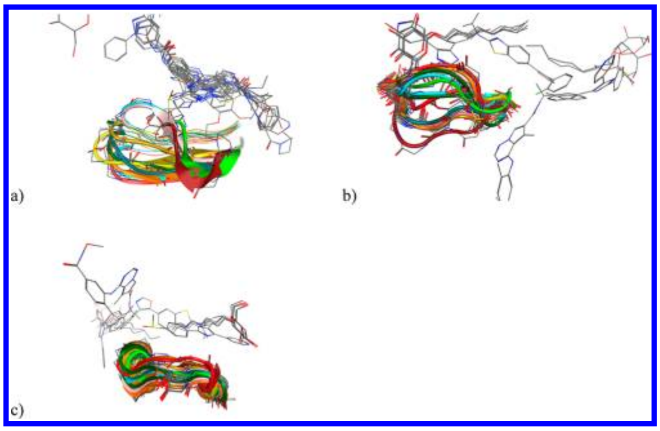


Figure 3. Regions of induced-fit amino acid residues (a,b) or non-induced-fit amino acid residues (c) in the p38 map kinase family cluster judged by Z-score. Protein segments shown by ribbons include consecutive categorized residues: (a) Gly31, Ser32, Gly33, Ala34, Tyr35, Gly36 (Z-score ≥ 2.0). (b) Met198 and His199 (Z-score ≥2.0). (c) Lys248, Ile250, Ser251, Ser252, Glu253, Ser254, Arg256 and Asn257 (1.0 ≤ Z-score < 2.0). This figure includes 20 protein structures. PDB codes: 2FST.X, 2FSO.X, 2FSL.X, 2FSM.X, 2NPQ.A, 2BAJ.A, 2BAK.A, 2BAL.A, 2BAQ.A, 3L8X.A, 1IAN.A, 2RG5.A, 2RG6.A, 2QD9.A, 3BX5.A, 3C5U.A, 3BV2.A, 3BV3.A, 1A9U.A, and 2EWA.A.

We adopted the Z-score = 2.0 as the threshold to discriminate the two class of induced-fit.

Ratios of Induced-Fit Classification for Each Residue Type. As previously mentioned, the Z-score based on the Cα atom RMSD was introduced to detect the induced-fit residues. To validate the threshold value for the classification of whether a residue was induced-fit positive or negative, the distribution of the percentages of these classified residues was plotted as a histogram using a Z-score ≥2 (Figure 4).

The numbers of positive and negative residues were 2982 and 15,736, respectively. In the case of Gly, the frequency of the positive residues was greater than that of the negative residues. In fact, the Gly-rich region, such as the G-loop, in the protein kinase families easily moved to fit ligands. Because the Gly residue has no side chain, the main chain seems to be relatively more flexible than those of the other residues, which are restricted by their side chains. Surprisingly, Pro is frequently positive, even though it has a rigid covalent bond between the main and side chains. In contrast, in the cases of the hydrophobic residues, Ile, Leu, and Val, the ratio of each positive is significantly lower than that of each negative. These residues tend to form the rigid protein core because they were classified as hydrophobic residues, and the protein core would not easily interact with a ligand. In contrast, in the case of Cys, the ratio of negative cases is greater than that of positive cases. A disulfide bond would be involved in the stability of the local environment around the Cys residue in some cases. From these analyses, a Z-score value of 2.0 would be adequate to distinguish the two types of residues and to satisfy the biological correspondence. Therefore, the threshold of the Z-score value was set to 2.0 in this study.

Optimization of Window Size SVM parameters. Window size is one of the main factors for achieving high prediction accuracy because this value determines the quantity of the PSSM descriptors used for the machine learning. However, increasing the number of residues in one window unit makes it difficult to describe the local environment and also directly affects the computational cost, such as the calculation speed and memory occupation. To determine the optimized value of the window size, the prediction performance was plotted for each window size for 17, 23, and 29, using the NR95 cluster data set (Figure S1Supporting Information) (described in the Materials and Methods section).

The highest prediction performance was recorded under the conditions with a window size of 29 and a prediction value of −0.6 as the threshold for discrimination. However, the maximum MCC values using window sizes of 23 (MCC = 0.543 ± 0.032, threshold = −0.6) and 29 (MCC = 0.553 ± 0.032, threshold = −0.6) were almost the same and were slightly better than that of the window size of 17 (MCC = 0.506 ± 0.028, threshold = −0.6). As the number of descriptors increased [1241 (window size: 17), 1679 (window size: 23), and 2117 (window size: 29)], the construction of one model required a long computer processing time and a huge amount of memory (over 1 gigabyte). We concluded that sufficient precision performance was obtained for the prediction with the window size of 23. Therefore, in this study, the optimal window size was set to 23.

In SVM classification, the calculated prediction value was applied to the two class separation. Therefore, the threshold of this value serves as an important parameter of the induced-fit

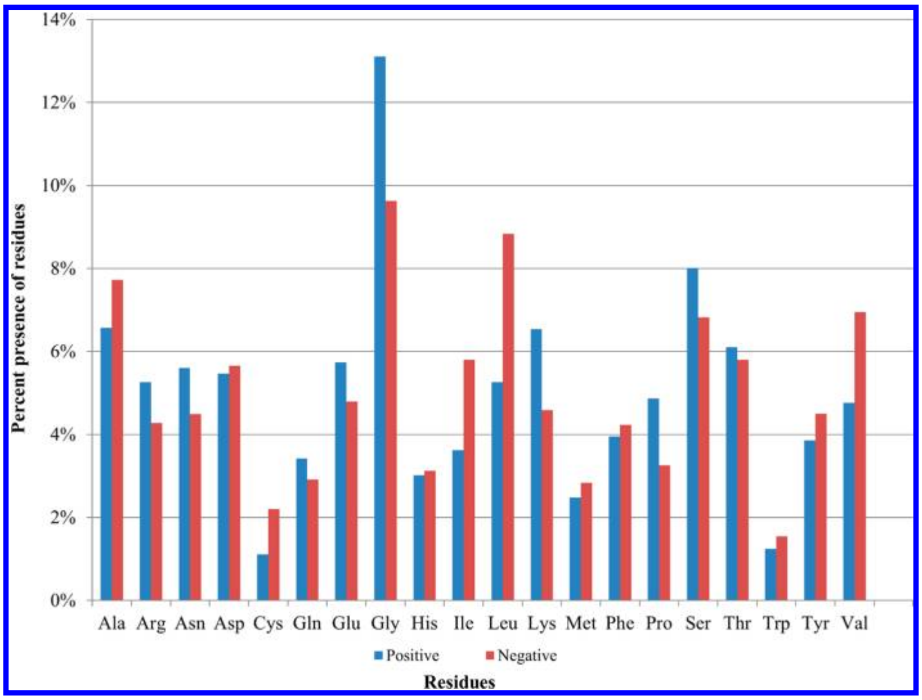


Figure 4. Distribution of ratios of the induced-fit classification. "Positive" label means the residue was classified as an induced-fit residue. "Negative" label means the residue was non-induced-fit.

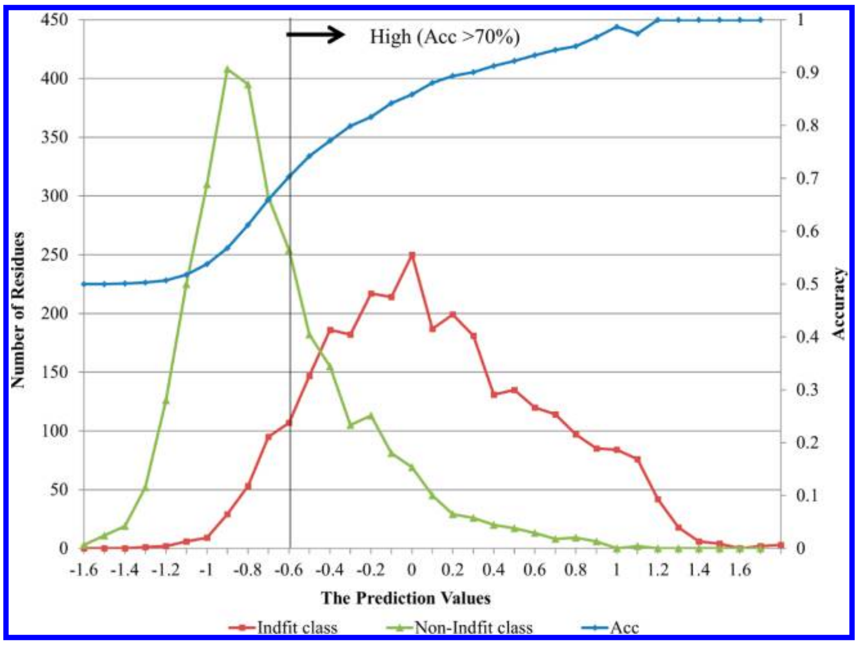


Figure 5. Distribution of the prediction values of induced-fit cases and non-induced-fit cases in the 5-fold cross validation test. "Indfit class" and "Non-indfit class" were defined for each residue in the test set based on the Z-score (≥2).

classification. As described in the Materials and Methods section, the gamma value and cost factor in SVM learning ranged from 0.000001 to 1.0 and 1 to 1000, respectively. The maximum MCC = 0.560 (SD = 0.030) for the NR95 data set was observed under the conditions of gamma = 0.001, cost factor = 5, and threshold = −0.6 under the 5-fold cross validation. In addition, other measures for prediction performance were calculated: SN = 0.849 (SD = 0.024), SP = 0.705 (SD = 0.016), ACC = 0.777 (SD = 0.015), PPV = 0.742 (SD = 0.011), and kappa = 0.554 (SD = 0.029) (Figure S2, Supporting Information). In the case of the NR30 cluster data set under the same parameters, comparable prediction performance was achieved: MCC = 0.525 (SD = 0.025), SN = 0.838 (SD = 0.019), SP = 0.759 (SD = 0.012), ACC = 0.759 (SD = 0.012), PPV = 0.724 (SD = 0.012), and kappa =0.519 (SD = 0.025) (Figure S2, Supporting Information).

In the range of threshold values between −0.7 (MCC = 0.540) and −0.4 (MCC = 0.546), almost the same prediction performances for MCC were observed. The sensitivity was dominant when less than the threshold = −0.6, and the specificity was dominant when more than the threshold = −0.4. Therefore, if a false positive prediction is permitted, then the threshold less than −0.6 should be used. These threshold values would be used depending on the prediction situation and the prediction target families.

Distribution of the Induced-Fit State for Each Prediction Value. The prediction value distributions of the induced-fit (indfit) and non-induced-fit (nonindfit) classes are shown in Figure 5. At the prediction value of −0.1, the peak of the indfit-fit class was observed. At the prediction value of −1.0, the peak of the non-induced-fit class was observed. In terms of the positive precision (accuracy in positively predicted samples), that at the prediction value of −0.6 was 70%.

Size of Data Set and Descriptor Selection. In the knowledge-based prediction approach using machine learning, both the quality and quantity of the training data set are important. Generally, when more information is included in the training data set, better prediction performance is expected; however, there is a concern that good prediction performance would only be obtained from sequential similarities that tend to cause induced-fit. We examined the effects of the absence of similar sequences using the NR30 data set in addition to the NR95 data set. In the NR30 data set (4426 protein chains), the homologous proteins (identity ≥ 30%) were removed from each representative sequence, and thus, the number of selected proteins was fewer than that of the NR95 data set (6637 protein chains) (described in the Materials and Methods section). In the validation, the same learning parameters (i.e., gamma, cost factor) were applied to construct models based on the NR30 and NR95 data sets. The prediction performances based on the NR30 data set are plotted in Figure S2 of the Supporting Information. As compared with the prediction performance of the NR95 data set, the maximum MCC and ACC values were slightly reduced by 0.035 (0.525 to 0.560) and 0.018 (0.759−0.777) at a threshold = −0.6, respectively. The quantity of similar sequences in the training and test sets slightly affects the predictive performance. Even when the similar sequences were removed from the training and test sets, a drastic decrease in the predictive performance was not observed in our validation. Therefore, we used prediction modes based on only NR95 for further validation of the actual targets in the following section.

In order to validate the contributions of the individual descriptor sets to the induced-fit prediction, the prediction models constructed by using each type of descriptor set were applied to the induced-fit prediction. The MCC and ACC values of each descriptor set at the threshold = −0.6 are summarized in Table 1.

Table 1. MCC and ACC at Threshold = −0.6 in Each Descriptor Set

	NR95 (threshold = −0.6)			
	MCC_{ave}	MCC_{SD}	ACC_{ave}	ACC_{SD}
ALL[a]	0.560	0.030	0.777	0.015
count residue	0.424	0.019	0.702	0.008
ASA	0.401	0.014	0.690	0.006
depth	0.359	0.024	0.668	0.010
sequence only[b]	0.359	0.019	0.666	0.009
PSSM	0.332	0.013	0.654	0.007
SSPROB	0.213	0.026	0.594	0.012
FRG_TOR	0.172	0.041	0.580	0.019
HYD	0.140	0.034	0.562	0.016
FRG_H	0.107	0.035	0.551	0.017
RB	0.093	0.042	0.546	0.021
tempfactor	0.088	0.015	0.518	0.004
adeg	0.077	0.035	0.537	0.017
CAV	0.061	0.014	0.529	0.007
MW	0.055	0.041	0.527	0.021
chideg	0.039	0.040	0.519	0.020
ringbond	0.038	0.020	0.519	0.010
charge	0.001	0.024	0.501	0.011

[a]ALL: All of the descriptors were used for constructing the prediction model. [b]Sequence only: A combination of PSSM, SSPROB, and HYD descriptors was used.

The best prediction performance was based on only using the "count residue" descriptor set, which means the number of classified amino acid residues around the objective residue (MCC = 0.424, ACC = 0.702). The contributions of "ASA" (degree of exposure, MCC = 0.401, ACC = 0.690) descriptor sets and the "depth" (distance from the centroid of the residue, MCC = 0.359, ACC = 0.668) followed. "Count residue", "ASA", and "depth" express the position of the amino acid residue in three-dimensional space. Particularly, the "count residue" descriptor set expresses the local three-dimensional environment of each residue. Among the 2D descriptor sets, the PSSM descriptor set showed moderately fine prediction performance (MCC = 0.332, ACC = 0.654). Additionally, the prediction performance of the combination of all descriptors calculated from only the amino acid sequence (i.e., PSSM, SSPROB, and HYD) was assessed for a better understanding of the 2D descriptors ("sequence only", Table 1). The prediction performance was slightly improved (MCC = 0.359, ACC = 0.666) compared to that using only the PSSM descriptor. Although the prediction performances of the PSSM model and "sequence only" model are inferior to those of the 3D descriptor models, "count residue", "ASA", and "depth", if an experimental structure is not available for the objective protein, then these sequence-based models can be used for induced-fit prediction.

On the other hand, the models based on the "charge", "MW", and "chideg" descriptor sets showed poor performance (MCC = 0.039− 0.055, ACC = 0.501−0.527). Surprisingly, the temperature factor model also exhibited low performance in spite of the fact that temperature factors fluctuate in crystal structures. This suggested that induced-fit is influenced by other aspects compared to simple disorder prediction. Although some differences in the prediction performance were observed for these descriptor set models, all of the descriptor sets contributed to the prediction performance to some degrees because the ACC values exceeded 0.50. We supposed that all of the descriptor sets prepared in this study are somehow related to induced-fit. We used the RBF kernel of SVM to build the models. In our models, not only the simply linear combination of the descriptors but also the cooperative effects of the descriptors are expressed, as suggested by Kinning et al.[36] Therefore, all of the descriptors were employed in our prediction model for the following application stage.

Applications to External and Well-Known Induced-Fit Examples. *1. ECL Loop and Helix Motion of a GPCR.* The GPCR (G protein-coupled receptor) family proteins are located on the cell membrane and play important roles in controlling signal transduction from an extracellular domain to an intracellular one. GPCRs are also known as popular target proteins of various diseases, such as cancer, cardiac dysfunction, diabetes, central nervous system disorders, obesity, inflammation, and pain.[37] Recent X-ray crystallography analyses revealed that the extracellular loop (ECL) is involved in the substrate recognition by the GPCR.[38,39] Moreover, the helix domain in the trans-membrane region exhibits a drastic conformational change. Here, our prediction model was applied to a GPCR to verify the precision of the induced-fit prediction. The prediction values are shown in Table 2.

Table 2. Predicted Values for Each Residue in Turkey Beta1 Adrenergic Receptor (PDB code: 2VT4.A)[a]

residue	prediction value
Met179	−0.707
His180	**-0.330**
Trp181	**-0.168**
Trp182	**0.078**
Arg183	**0.358**
Asp184	**0.582**
Glu185	**0.692**
Asp186	**0.817**
Pro187	**0.866**
Gln188	**0.753**
Ala189	**0.472**
Leu190	**0.457**
Lys191	**0.190**
Cys192	**0.136**
Tyr193	**-0.044**
Gln194	**0.059**
Asp195	**-0.315**
Pro196	**-0.161**
Gly197	**-0.528**
Cys198	**-0.598**
Cys199	−0.731
Asp200	−0.897
Phe201	−0.730
Val202	−0.713
Thr203	−0.706
Asn204	−0.718
Arg205	−0.643
Ala206	**-0.478**
Tyr207	−0.700
Ala208	−0.770

[a] If the value was more than the cutoff of −0.6 (shown in bold type), then the residue was a predicted induced-fit residue. The underlined residues belong to the ECL2 region. All of the prediction values are listed in Table S1 of the Supporting Information.

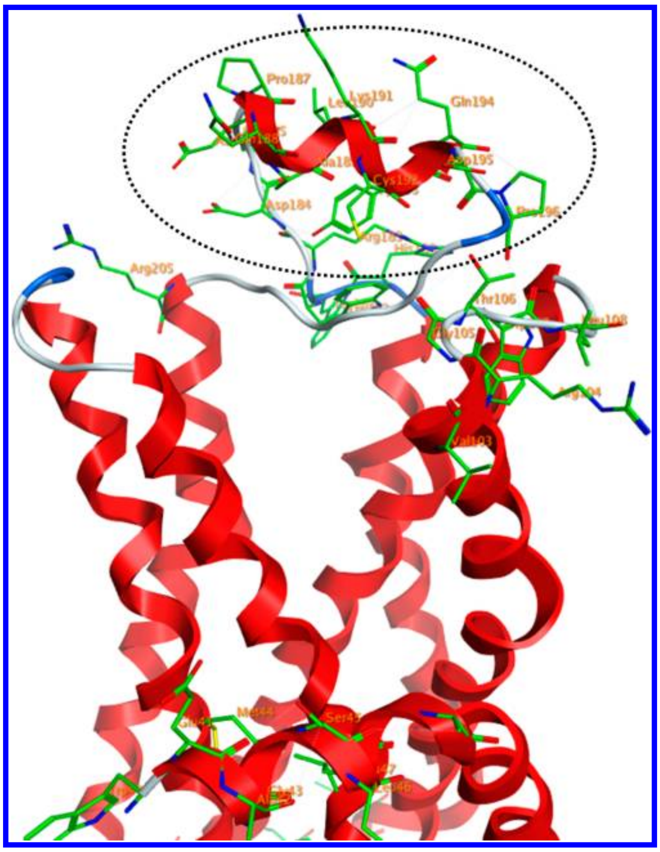


Figure 6. Turkey beta1 adrenergic receptor (PDB code: 2VT4) shown by ribbon style. The predicted induced-fit residues are shown by stick models with orange labels. The α-helix region above indicates the ECL2 region, enclosed in a dotted circle.

The region encircled with the dashed line in Figure 6 indicates the induced-fit residues predicted by our prediction model. This region is known as ECL2, and it adopts various conformations for ligand binding. Thus, the prediction successfully detected the ECL domain as the induced-fit region.

It is also known that GPCR activation is accompanied by a conformational change of the large helix of the trans-membrane domain. However, the induced-fit residues in this region could not be predicted by the current model. In this study, because the prediction model was built according to the induced-fit "residue" based on the RMSD of the Cα atom, large conformational changes, such as domain motions, would be difficult to detect by only residue-based information. Moreover, the hydrophilic and hydrophobic residue allocations of membrane proteins are different from those of globular proteins because of the hydrophobic environment of the trans-membrane regions. The construction of prediction models using a training data set corresponding to the membrane protein is needed for the GPCR target in a future study.

2. DFG Motifs of Protein Kinases. The protein kinase family members (PKs) are ATP-dependent enzymes that phosphorylate Ser, Thr, and Tyr residues in various substrates. The phosphoryl group transfer reaction is frequently involved in the signaling and control of intracellular protein functions. Because the abnormal expression of PKs can cause cancer, PKs are one of the main targets of innovative drug development. In PKs, various conformational changes in the presence of ligands have often been observed, and they contribute to enhancing the kinase selectivity of the ligands. For example, the easily moved loop motif located in the kinase active site is known to adopt the "DFG-in" (active form for phosphorylation) and "DFG-out" (inactive form) conformations.[11−13] Here, we assessed whether our prediction model could detect the DFG motif and discriminate between kinases that form only DFG-in and those that form both DFG-in and DFG-out.

Prediction of the DFG-Out State from the DFG-In State Structure. Basically, our model was assumed to apply to the static coordinates (such as X-ray structures) to predict the dynamics of the protein. First, we verified the constructed prediction model to detect the DFG-out state using only the usual active state of the kinase structure (i.e., DFG-in state). The kinase family data set was obtained from the MOE kinase DB described in the Materials and Methods section. The data for the two categories were abstracted from all kinase data. The first category consists of the DFG-in structures in that the DFG-state was reported as only the DFG-in state in the MOE kinase DB (group "A" in Figure 7). The second group consists of the DFG-in structures for which both the DFG-in and DFG-out states are known, such as Abl, p38, and insulin kinase (group "B" in Figure

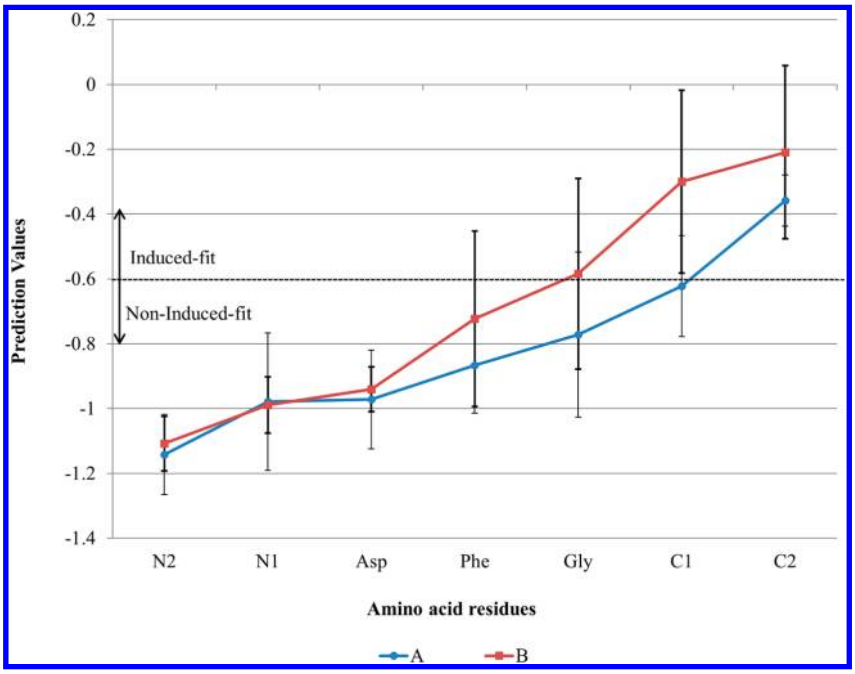


Figure 7. Predicted induced-fit values in the DFG-in motif residues of the kinase search by MOE kinase search 2010. (A) Only using DFG-in motif families ($N = 14$) reported in MOE kinase DB 2010. (B) Using both DFG-in and DFG-out motif families ($N = 7$) reported in MOE kinase DB 2010. The subfamilies including more than 10 chains were selected for the calculation. If the value was more than the cutoff of −0.6, then the residue was a predicted induced-fit residue (shown by a bold letter). The red-colored residues belong to the DFG-motif. (These values are also available in Table S2 of the Supporting Information.).

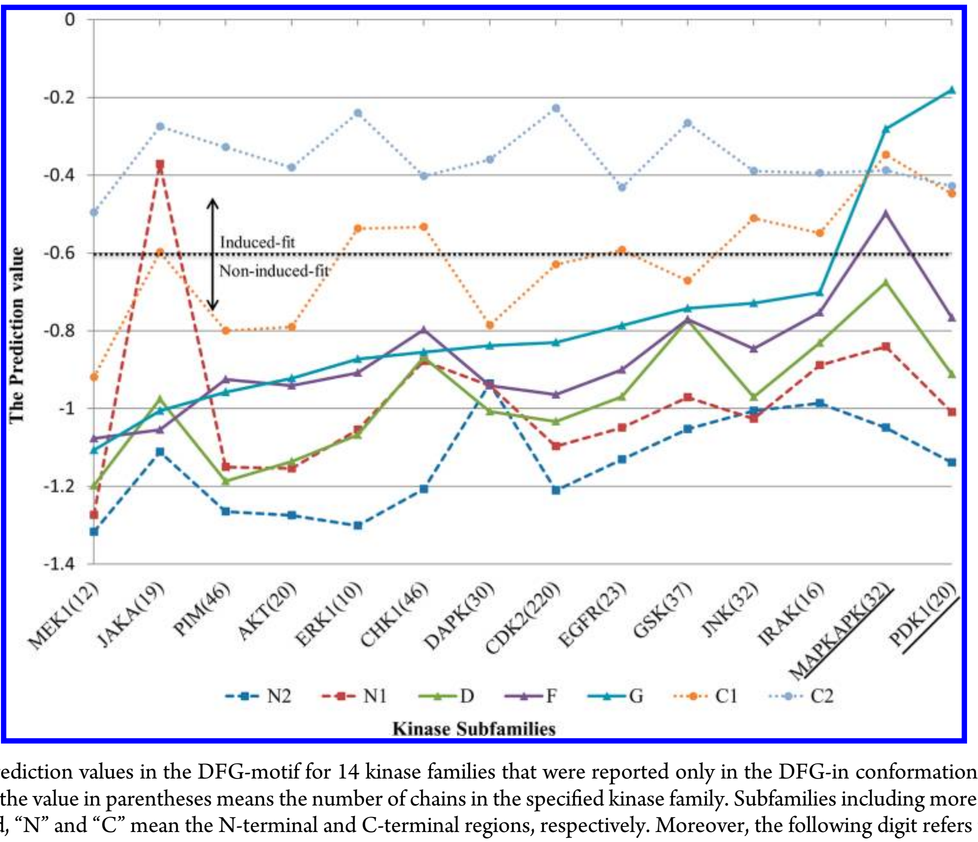


Figure 8. Average prediction values in the DFG-motif for 14 kinase families that were reported only in the DFG-in conformation. In the labels of the kinase family names, the value in parentheses means the number of chains in the specified kinase family. Subfamilies including more than 10 chains were plotted. In the legend, "N" and "C" mean the N-terminal and C-terminal regions, respectively. Moreover, the following digit refers to the distance from the DFG-motif. The two underlined subfamilies on the right are the probable DFG-out candidates predicted by our prediction model.

7). The prediction model was applied to these two groups to compare the averages of the prediction values.

The distributions of the predicted values of the two categories showed significant changes at the residues Phe, Gly, C1, and C2. In particular, there is a major difference between the two groups at the C1 residue. It successfully predicted the occurrence of DFG-out. At Gly and C1 of the B group, the prediction values were −0.584 and −0.300, respectively, which are over the

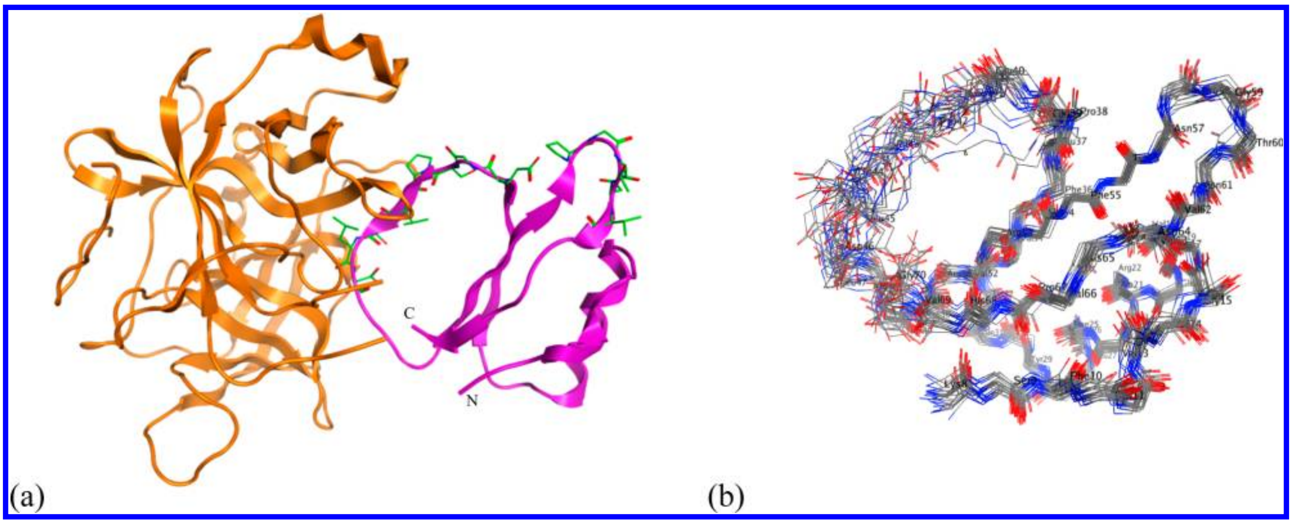


Figure 9. Application of our induced-fit prediction model to the Eglin C structures. (a) Structures of chymotrypsin and Eglin C are shown by orange and magenta ribbon styles, respectively. The residues with positive prediction values are shown by green stick models. The structure of Eglin C was determined by X-ray crystallography (PDB code: 1ACB.I). (b) Ensemble of 25 Eglin C structures determined by NMR (PDB code: 1EGL). Only main chain atoms of the corresponding residues to the X-ray structure (i.e., from Lys8 to Gly70) are shown for comparison toFigure 9a.

threshold (i.e., −0.6) to judge an induced-fit positive residue, while the corresponding values of the A group were −0.772 and −0.622, respectively. At the Phe and C2 positions, the prediction values of the B group were also superior to those of the A group. These results indicated that the DFG-motif motion is caused by the effects of the residues after the Gly residue, rather than the direct motion of the Phe residue in the DFG-motif.

Here, to predict the potential of the conformational change from DFG-in to DFG-out, our induced-fit prediction model was applied to the PKs with DFG-motifs that were experimentally reported to form only the DFG-in state, according to X-ray crystallographic analyses. The DFG-in state-only PKs were collected from the MOE kinase DB 2010. At this time, the PKs lacking Asp, Phe, and Gly residues in the DFG-motif and not belonging to the human species were removed from the data set. Finally, 14 PKs were obtained and used for the prediction (Figure 8).

As a result, the predicted values of Gly in most of the PKs, except for MAPKAPK and PDK1, showed non-induced-fit, and the results were reasonable because the structures of those PKs are known to adopt the DFG-in state. Interestingly, the PDK1 structures included only the DFG-in motif in MOE 2010. Recently, in MOE 2011, the structures of the DFG-out state (PDB codes: 3NAX.A,[40] 3QC4.A, and 3QC4.B[41]) have been newly registered. In spite of the fact that our prediction model was based on the PDB data as of 2009, the model successfully predicted the DFG-out state, including the prediction value of Gly (−0.181) in PDK1 without using any PDB 2011 information.

As with PDK1, the predicted values of MAPKAPK (Human mitogen-activated protein kinase) were relatively higher than those of the other kinases. In addition, our model detected the Phe residue as an induced-fit residue within the threshold = −0.6. Therefore, the MAPKAPs are expected to form the DFG-out conformation caused by the ligand interaction. In fact, in the same family, MNK (MAPK-interacting kinase), which shares 30.8% sequence identity with MAPKAP (PDB codes: 3MW2[42] and 2AC3,[43] respectively), was also reported to form the DFG-out state structure in MOE 2010. In the future, we predict that the DFG-out state of MAPKAP will be reported.

Although the default threshold was applied to the entire protein families, the prediction performance in the range of the threshold (−0.6 to ∼−0.4) around the threshold = −0.5 was almost the same, and specializations for specificity (SP) and sensitivity (SN) were observed. Therefore, exploring a suitable threshold for each protein family would also be a useful solution for more precise prediction.

3. Protein−Protein Interaction (PPI) Target. In this study, we have described a method to predict the flexible regions of protein−small molecule interactions. PPIs have recently been reported as target proteins for next generation drug discovery. Protein motion induced by another protein molecule is also referred to as induced-fit. For example, Eglin C is a PPI-involved protein that inhibits various proteases, such as cathepsin G and chymotrypsins. The structure of Eglin C consists of both the rigid core and binding loop.[4] Moreover, various conformations of the binding loop have been experimentally observed by NMR (Figure 9). Thus, PPI can also cause the conformational changes involved in induced-fit as in the interaction with a small molecule. Here, we determined whether our prediction model could predict the induced-fit region of the binding loop. The model was applied to Eglin C (PDB code: 1ACB.I[44]) in the complex structure of chymotrypsin−Eglin C (PDB code: 1ACB) (Figure 9), and the prediction values are plotted in Figure 10.

The predicted induced-fit residues were mostly in agreement with the flexible residues determined by NMR. Especially, the binding loop of Eglin C was predicted with good precision for the second motif shown in Figure 10. The region around the N-terminus (first motif) also has a positive prediction value. Generally, the N-terminal region is naturally unstable and actually flexible as observed in the NMR ensemble (Figure 9). The average prediction values of the first and second motifs were 0.131 and 0.059, respectively. Both values were much higher than the threshold value (i.e., −0.6). As a result, although the model based on the binding site around the small molecules data set was applied, the predictions were reasonable for the PPI target. This result indicated that the common induced-fit mechanism would be functional with both small molecules and macromolecules. Currently, the prediction values do not show the degree of motion but the possibility of flexibility at the defined threshold.

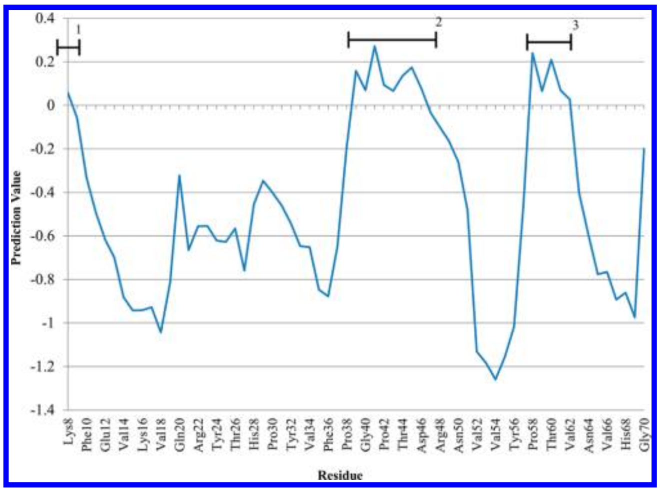


Figure 10. Predicted values for each residue in Eglin C (PDB code: 1ACB.I). The black lines show residues with positive prediction values. Especially, the second motif is also known as the binding loop. (These values are available in Table S3 of the Supporting Information.).

The next objective is to assess the degree of motion upon ligand binding using regression models.

CONCLUSION

In this study, machine learning models using SVM and selected descriptors were built based on the experimentally obtained induced-fit data. The results validated that the best constructed model could predict the main chain residues that are involved in induced-fit caused by ligand binding by outside test sets of protein kinases, protein−protein interaction targets, and GPCR. To predict the induced-fit region, amino acid residues were classified into two categories (induced-fit residue and non-induced-fit residue) based on the Z-score of the RMSD for the Cα atoms in the same position in the alignment. In this study, a Z-score = 2.0 was used to detect main chain motions. The classification of the induced-fit residues can be changed in the range of the Z-score. In a future study, suitable Z-score thresholds for various induced-fit situations, such as loop motion and large domain motion, can be used. As descriptors for machine learning, we examined sequence-based descriptors, such as PSSM and physicochemical properties based on the 3D environment. From the results of validation by individual descriptor-based prediction models, no model was superior to the all descriptor-based model. Both descriptor categories contributed to the prediction performance. In these descriptors, the number of surrounding residues and depth of position, classified in the physicochemical descriptors, were significantly more important than the sequence-based PSSM. Therefore, the three-dimensional structural data are required for more accurate predictions for induced-fit. The descriptor-based model recorded high prediction performance (ACC = 0.777 and MCC = 0.560). In external validations, our prediction model was effective in detecting the induced-fit regions for the PPI binding loop and GPCR extracellular loop. In the validation using kinases, our model successfully discriminated kinases with only the DFG-in state and kinases with both the DFG-in and DFG-out states from only the DFG-in state structures. The results suggested that the DFG-motif motion is caused by the residues following the Gly residue, rather than the direct motion of the Phe residue in the DFG-motif.

However, the model was not applicable to changes with drastic main chain movement (domain motion), such as the helix allocation of a GPCR and rearrangement of a side chain hydrogen bond network. We believe that this issue can be resolved by preparing data sets with classifications of side chain and domain motions. Currently, our model can only predict the possibility of induced-fit with moderate main chain motions. To assess the extent of the induced-fit motion, more specific data sets and descriptors are required. In a future study, our prediction model for the induced-fit residues can be applied to enhance docking-based virtual screening by considering the flexibility of the detected induced-fit regions.

ASSOCIATED CONTENT

Ⓢ Supporting Information

Figure S1: Distributions of MCC and ACC for each window size using the NR95 data set. Figure S2: Distribution of all of the performance indicators using the NR95 and NR30 data set. Table S1: All of the predicted induced-fit values for each residue in Turkey beta1 adrenergic receptor. Table S2: Predicted induced-fit values in the DFG-in motif residues from the kinase search. Table S3: All of the predicted induced-fit values for each residue in Eglin C. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Telephone: +81-45-503-9433. Fax: +81-45-503-9432. E-mail: honma@gsc.riken.jp.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported in part by the Targeted Protein Research Programs from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

REFERENCES

(1) Thornton, J. M.; Todd, A. E.; Milburn, D.; Borkakoti, N.; Orengo, C. A. From structure to function: Approaches and limitations. *Nat. Struct. Biol.* **2000**, *7*, 991−994.

(2) Berman, H. M. The Protein Data Bank: A historical perspective. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **2008**, *64*, 88−95.

(3) Pargellis, C.; Tong, L.; Churchill, L.; Cirillo, P. F.; Gilmore, T.; Graham, A. G.; Grob, P. M.; Hickey, E. R.; Moss, N.; Pav, S.; Regan, J. Inhibition of p38 MAP kinase by utilizing a novel allosteric binding site. *Nat. Struct. Biol.* **2002**, *9*, 268−272.

(4) Hyberts, S. G.; Goldberg, M. S.; Havel, T. F.; Wagner, G. The solution structure of eglin c based on measurements of many NOEs and coupling constants and its comparison with X-ray structures. *Protein Sci.* **1992**, *1*, 736−751.

(5) Sun, Y.-J.; Rose, J.; Wang, B.-C.; Hsiao, C.-D. The Structure of glutamine-binding protein complexed with glutamine at 1.94 Å resolution: Comparisons with other amino acid binding proteins. *J. Mol. Biol.* **1998**, *278*, 219−229.

(6) Bakan, A.; Bahar, I. Computational generation inhibitor-bound conformers of p38 map kinase and comparison with experiments. *Pac. Symp. Biocomput.* **2011**, *16*, 181−192.

(7) Soliva, R.; Gelpí, J. L.; Almansa, C.; Virgili, M.; Orozco, M. Dissection of the recognition properties of p38 MAP kinase. Determination of the binding mode of a new pyridinyl-heterocycle inhibitor family. *J. Med. Chem.* **2007**, *50*, 283−293.

(8) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. Long-timescale molecular dynamics simulations of protein structure and function. *Curr. Opin. Struct. Biol.* **2009**, *19*, 120−127.

(9) Karplus, M.; Kuriyan, J. Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (19), 6679−6685.

(10) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.

(11) Yi, L.; Gray, N. S. Rational design of inhibitors that bind to inactive kinase conformations. *Nat. Chem. Biol.* **2006**, *2*, 358−364.

(12) Zuccotto, F.; Ardini, E.; Casale, E.; Angiolini, M. Through the "Gatekeeper Door": Exploiting the active kinase conformation. *J. Med. Chem.* **2010**, *53*, 2681−2694.

(13) Zhang, J.; Yang, P. L.; Gray, N. S. Targeting cancer with small molecule kinase inhibitors. *Nat. Rev. Cancer* **2009**, *9*, 28−39.

(14) *Maestro*, version 9.2; Schrödinger, LLC: New York, 2011.

(15) Takaya, D.; Yamashita, A.; Kamijo, K.; Gomi, J.; Ito, M.; Maekawa, S.; Enomoto, N.; Sakamoto, N.; Watanabe, Y.; Arai, R.; Umeyama, H.; Honma, T.; Matsumoto, T.; Yokoyama, S. A new method for induced fit docking (GENIUS) and its application to virtual screening of novel HCV NS3−4A protease inhibitors. *Bioorg. Med. Chem.* **2011**, *19* (22), 6892−6905.

(16) Li, N.; Sun, Z.; Jiang, F. Prediction of protein−protein binding site by using core interface residue and support vector machine. *BMC Bioinf.* **2008**, *9*, 553.

(17) Kumar, M.; Gromiha, M. M.; Raghava, G. P. S. Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins: Struct., Funct., Bioinf.* **2008**, *71*, 189−194.

(18) Chauhan, J. S.; Mishra, N. K; Raghava, G. P. S. Prediction of GTP interacting residues, dipeptides and tripeptides in a protein from its evolutionary information. *BMC Bioinf.* **2010**, *11*, 301.

(19) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer-Verlag: New York, 1995.

(20) Han, S.; Lee, B.; Yu, S. T.; Jeong, C.; Lee, S.; Kim, D. Fold recognition by combining profile-profile alignment and support vector machine. *Bioinformatics* **2005**, *21* (11), 2667−2673.

(21) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25* (17), 3389−3402.

(22) Hwang, H.; Vreven, T.; Whitfield, T. W.; Wiehe, K.; Weng, Z. A machine learning approach for the prediction of protein surface loop flexibility. *Proteins: Struct., Funct., Bioinf.* **2011**, *79* (8), 2467−2474.

(23) Wang, G.; Dunbrack, R. L., Jr. PISCES: A protein sequence culling server. *Bioinformatics* **2003**, *19*, 1589−1591.

(24) *Molecular Operating Environment (MOE)*, 2011.10; Chemical Computing Group, Inc.: Montreal, QC, Canada, 2011.

(25) Joachims, T. Making Large-Scale SVM Learning Practical. In *Advances in Kernel Methods-Support Vector Learning*; Schölkopf, B., Burges, C., Smola, A., Ed.; MIT Press: Cambridge, MA, 1999; pp 41−56.

(26) Kakuta, M.; Nakamura, S.; Shimizu, K. Prediction of protein−protein interaction sites using only sequence information and using both sequence and structural information. *IPSJ Digital Courier* **2008**, *4*, 217−227.

(27) Li, W.; Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22* (13), 1658−1659.

(28) Zhou, H.; Zhou, Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins: Struct., Funct., Bioinf.* **2005**, *58*, 321−328.

(29) Thompson, M. J.; Goldstein, R. A. Predicting protein secondary structure with probabilistic schemata of evolutionarily derived information. *Protein Sci.* **1997**, *6*, 1963−1975.

(30) Pace, C. N.; Shirley, B. A.; McNutt, M.; Gajiwala, K. Forces contributing to the conformational stability of proteins. *FASEB J.* **1996**, *10* (1), 75−83.

(31) Eisenberg, D.; Schwarz, E.; Komaromy, M.; Wall, R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.* **1984**, *179* (1), 125−142.

(32) Hall, L. H.; Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure−Property Modeling. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; John Wiley and Sons, Inc.: Hoboken, NJ, 2007; *2*, pp 367−422.

(33) Hall, L. H.; Kier, L. B. The nature of structure−activity relationships and their relation to molecular connectivity. *Eur. J. Med. Chem.* **1997**, *4*, 307−312.

(34) Carletta, J. Assessing agreement on classification tasks: The kappa statistic. *Comput. Linguist.* **1996**, *22* (2), 249−254.

(35) Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A. F.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics* **2000**, *16* (5), 412−424.

(36) Kinnings, S. L.; Liu, N.; Tonge, P. J.; Jackson, R. M.; Xie, L.; Bourne, P. E. A machine learning-based method to improve docking scoring functions and its application to drug repurposing. *J. Chem. Inf. Model.* **2011**, *51* (2), 408−419.

(37) Lappano, R.; Maggiolini, M. G protein-coupled receptors: Novel targets for drug discovery in cancer. *Nat. Rev. Drug Discov.* **2011**, *10* (1), 47−60.

(38) Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G. F.; Thian, F. S.; Kobilka, T. S.; Choi, H.-J.; Kuhn, P.; Weis, W. I.; Kobilka, B. K.; Stevens, R. C. High-resolution crystal structure of an engineered human β2-adrenergic G protein-coupled receptor. *Science* **2007**, *318* (5854), 1258−1265.

(39) Warne, T.; Serrano-Vega, M. J.; Baker, J. G.; Moukhametzianov, R.; Edwards, P. C.; Henderson, R.; Leslie, A. G. W.; Tate, C. G.; Schertler, G. F. X. Structure of a β1-adrenergic G-protein- coupled receptor. *Nature* **2008**, *454* (7203), 486−491.

(40) Nagashima, K.; Shumway, S. D.; Sathyanarayanan, S.; Chen, A. H.; Dolinski, B.; Xu, Y.; Keilhack, H.; Nguyen, T.; Wiznerowicz, M.; Li, L.; Lutterbach, B. A.; Chi, A.; Paweletz, C.; Allison, T.; Yan, Y.; Munshi, S. K.; Klippel, A.; Kraus, M.; Bobkova, E. V.; Deshmukh, S.; Xu, Z.; Mueller, U.; Szewczak, A. A.; Pan, B. S.; Richon, V.; Pollock, R.; Blume-Jensen, P.; Northrup, A.; Andersen, J. N. Genetic and pharmacological inhibition of PDK1 in cancer cells: Characterization of a selective allosteric kinase inhibitor. *J. Biol. Chem.* **2011**, *286* (8), 6433−6448.

(41) Erlanson, D. A.; Arndt, J. W.; Cancilla, M. T.; Cao, K.; Elling, R. A.; English, N.; Friedman, J.; Hansen, S. K.; Hession, C.; Joseph, I.; Kumaravel, G.; Lee, W. C.; Lind, K. E.; McDowell, R. S.; Miatkowski, K.; Nguyen, C.; Nguyen, T. B.; Park, S.; Pathan, N.; Penny, D. M.; Romanowski, M. J.; Scott, D.; Silvian, L.; Simmons, R. L.; Tangonan, B. T.; Yang, W.; Sun, L. Discovery of a potent and highly selective PDK1 inhibitor via fragment-based drug discovery. *Bioorg. Med. Chem. Lett.* **2011**, *21* (10), 3078−3083.

(42) Argiriadi, M. A.; Ericsson, A. M.; Harris, C. M.; Banach, D. L.; Borhani, D. W.; Calderwood, D. J.; Demers, M. D.; DiMauro, J.; Dixon, R. W.; Hardman, J.; Kwak, S.; Li, B.; Mankovich, J. A.; Marcotte, D.; Mullen, K. D.; Ni, B.; Pietras, M.; Sadhukhan, R.; Sousa, S.; Tomlinson, M. J.; Wang, L.; Xiang, T.; Talanian, R. V. 2,4-Diaminopyrimidine MK2 inhibitors. Part I: Observation of an unexpected inhibitor binding mode. *Bioorg. Med. Chem. Lett.* **2010**, *20* (1), 330−333.

(43) Jauch, R.; Jäkel, S.; Netter, C.; Schreiter, K.; Aicher, B.; Jäckle, H.; Wahl, M. C. Crystal structures of the Mnk2 kinase domain reveal an inhibitory conformation and a zinc binding site. *Structure* **2005**, *13* (10), 1559−1568.

(44) Frigerio, F.; Coda, A.; Pugliese, L.; Lionetti, C.; Menegatti, E.; Amiconi, G.; Schnebli, H. P.; Ascenzi, P.; Bolognesi, M. Crystal and molecular structure of the bovine alpha-chymotrypsin-eglin c complex at 2.0 Å resolution. *J. Mol. Biol.* **1992**, *225* (1), 107−123.