Elucidation of Characteristic Structural Features of Ligand Binding Sites of Protein Kinases: A Neural Network Approach

Tomoko Niwa*

Discovery Research Laboratories, Nippon Shinyaku Co., Ltd., 14, Nishinosho-Monguchi-cho, Kisshoin, Minami-ku, Kyoto, 601-8550 Japan

Received December 6, 2005

Protein kinases play important roles in regulating cellular signal transduction and other biochemical processes, and they are attractive targets for drug discovery programs in many disease areas. Most kinase inhibitors under development as drugs act by directly competing with ATP at the ATP-binding site of the kinase. There are more than 500 protein kinases, and the ATP-binding site is highly conserved among them. Therefore selectivity is an essential requirement for clinically effective drugs, and understanding the structural characteristics of ATP-binding sites is of crucial importance. The objective of the present study was to elucidate the structural characteristics of the adenosine-binding site of four major kinase groups, AGC (PKA, PKG, and PKC families), CaMK (calcium/calmodulin-dependent protein kinases), CMGC (CDK, MAPK, GSK3, and CLK families), and TK (tyrosine kinases). To do this, we classified the kinases into groups by using feed-forward multilayer perceptron (MLP) neural networks and structural, electronic, and hydrophobic descriptors of the amino acids at the adenosine-binding site. A total of 275 kinases were classified in two ways: (1) kinases belonging to a certain group were distinguished from those not belonging to that group, and (2) all of the kinases were classified into four groups. More than 85% of the kinases were correctly classified by both methods. Trained neural networks clarified which amino acids and which properties characterize the adenosine-binding site of each group, and the results were visualized by molecular graphics. Comparison of the modeled neural networks and the distributions of amino acids provided more detailed information on the structural characteristics of each group. Application of the present results to drug development is also discussed.

INTRODUCTION

Protein kinases catalyze the transfer of the γ-phosphoryl group from ATP to the hydroxyl groups of protein side chains, and they play critical roles in regulating cellular signal transduction and other biochemical processes. They are very attractive targets for today's drug discovery and development, and many pharmaceutical companies are intensively developing kinase inhibitors that may have therapeutic value.[1] Recent success with the BCR-ABL tyrosine kinase inhibitor Imatinib in the treatment of chronic myeloid leukaemia is a good example.[2] Most kinase inhibitors that have been tested in clinical trials act by directly competing with ATP at the ATP-binding site of the kinase. There are more than 500 protein kinases,[3] and the ATP-binding site is highly conserved among them. Therefore, selectivity is an essential requirement for clinically effective drugs, and an understanding of the structural differences among the ATP-binding sites of kinases is of crucial importance.

X-ray crystallography is a promising method for understanding the structural and physicochemical characteristics of the ATP-binding sites of protein kinases, but a sufficient number of X-ray crystallographic structures is not yet available. The only data which are generally available for kinases is the amino acid sequence. Sequence similarity is widely used for classifying proteins and predicting biological activities. Indeed, the sequence similarity in the catalytic domain of kinases has been successfully used to classify these enzymes into groups such as AGC (PKA, PKG, and PKC families), CaMK (calcium/calmodulin-dependent protein kinases), CMGC (CDK, MAPK, GSK3, and CLK families), and TK (tyrosine kinases).[4,5] However, it is difficult to directly elucidate the structural characteristics of the ATP-binding site from amino acid sequences alone. The objectives of the present study were to develop a procedure to overcome this difficulty by using physicochemical descriptors of amino acids and neural network modeling and to investigate the differences in the structural characteristics of the adenosine-binding sites among different groups of protein kinases. Since most kinase inhibitors do not interact with the phosphate-binding region of the ATP-binding site, the amino acids in the adenosine-binding site were selected for investigation.

We have previously proposed a hydrophobic descriptor, π_b, for amino acids;[6] this descriptor is derived from the 1-octanol/water partition coefficients of oligopeptides measured by Akamatsu et al.[7,8] To examine the applicability of our π_b values, we performed quantitative structure−activity relationships (QSAR) studies on biologically active peptides and found that the use of π_b, along with steric and electronic descriptors of the amino acids, produced excellent results.[6] In addition to the biological activities of peptides, these physicochemical descriptors rationalized well the structural stabilities of proteins and the β-sheet propensities of the amino acids.[6,9] These findings led us to investigate the use

* Corresponding author phone: +81-75-321-9010; fax: +81-75-321-9038; e-mail: t.niwa@po.nippon-shinyaku.co.jp.

10.1021/ci050528t CCC: $33.50
Published on Web 08/22/2006

of physicochemical descriptors of the amino acids for the structural analysis of protein kinases. The structural analysis of kinases is a challenging task for which powerful solutions are required. We have already applied neural networks to solve difficult problems in the evaluation of human intestinal absorption and the prediction of biological targets from chemical structures,[10,11] and in this study we apply neural networks to elucidate the structural characteristics of the adenosine-binding sites of protein kinases.

As mentioned above, kinases have been classified into groups based on the sequence similarity in the catalytic domain. These groupings indicate that eukaryotic protein kinase (ePK) domain phylogeny reflects substrate specificity and/or mode of regulation and could therefore serve as a useful classification tool. In this study, kinases were classified into four major groups, AGC, CaMK, CMGC, and TK, to elucidate the positions as well as the properties of the amino acids that allow discrimination between the different kinase groups. To our knowledge, this is the first attempt to apply neural networks and physicochemical descriptors of amino acids to the classification of protein kinases and the elucidation of the structural factors differentiating the groups of protein kinases. This newly developed method permitted us to identify and compare the structural characteristics of the adenosine-binding site in the various kinase groups. The kind of information we obtained is potentially helpful at every stage of drug development, from the selection of targets to the optimization of inhibitors.

DATA PREPARATION AND CLASSIFICATION PROCEDURE

Selection of Groups of Kinases. Hanks and Quinn[4] aligned the homologous catalytic-domain sequences of 65 distinct protein kinases from diverse eukaryotes and constructed a phylogenetic tree. They showed that the eukaryotic protein kinase (ePK) domain phylogeny reflects the substrate specificity and/or mode of regulation of protein kinases and serves as a useful classification tool. Later, an extended ePK domain alignment containing 390 sequences was made publicly available; this alignment includes four major groups, namely, the AGC, CaMK, CMGC, and TK groups.[12] The completion of the human genome sequence prompted the determination of the full complement of human protein kinases. Kostich et al.[13] used public GenBank records and BLAST searches and put together a collection of 510 potentially unique human ePKs.

A catalog of the protein kinase complement of the human genome, known as the human kinome, recently published by Manning et al.[3] includes 518 human ePKs and three new groups, CK1 (casein kinase 1), STE (homologues of yeast Sterile 7, Sterile 11, and Sterile 20 kinases), and TKL (tyrosine kinase-like kinases), as shown in Table 1. The AGC, CaMK, CMGC, and TK groups were selected for classification in our study, and the three new groups were excluded for the following reasons. First, according to the phylogenetic tree generated by Manning et al.,[3] the AGC, CaMK, CMGC, and TK groups form distinct branches, whereas the CK1, STE, and TKL groups do not form such distinct branches. Because this is our first attempt to use neural networks and physicochemical descriptors together to elucidate the structural characteristic of protein kinases, we selected only those groups that form distinct branches. Second, to obtain reliable classification models, the sizes of the groups should be comparable. Third, inhibitors of kinases belonging to the AGC, CaMK, CMGC, and TK groups are now under active development as potential therapeutic agents, so that these kinase groups are of particular interest to us.

Table 1. Major Groups of Human Kinases

group[a]	no. of families	no. of subfamilies	no. of individual kinases
AGC	14	21	63
CaMK	17	33	74
CK1	3	5	12
CMGC	8	24	61
other	37	39	83
STE	3	13	47
TK	30	30	90
TKL	7	13	43

[a] According to ref 3.

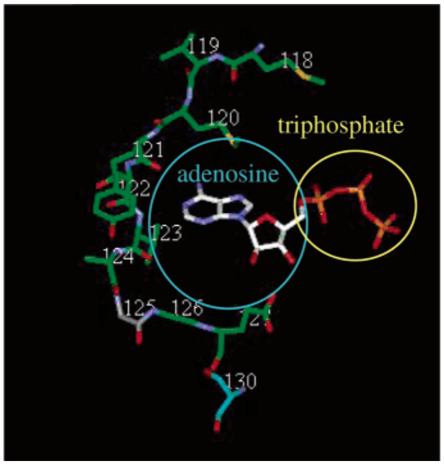


Figure 1. The amino acids in the adenosine-binding site of the X-ray structure of cAMP-dependent protein kinase (PDB ID: 1ATP), a representative protein kinase. The amino acids used for classification in this study are shown in green or blue, and the single amino acid not used for classification (at position 125) is shown in gray. The numbering of the amino acids is the same as that in the PDB file of 1ATP.

Selection of Amino Acids. Alignments of amino acid sequences are critically important, and without highly reliable alignments it is not feasible to elucidate the positions and physicochemical properties of the amino acids characterizing the groups into which the kinases are classified. We selected the amino acids in the adenosine-binding site for the following reasons. First, the adenosine-binding site is the most conserved region of the catalytic domain. Second, there is some ambiguity in the alignments of the amino acids at the phosphate-binding site, rendering it unsuitable for our classification studies. And third, most kinase inhibitors bind to the adenosine-binding site. A representative kinase, cAMP-dependent protein kinase (Protein Data Bank (PDB) ID: 1ATP), is shown in Figure 1, and the numbering of the amino acids is the same as in the PDB file. The amino acids at positions 120−127 are located close to the adenosine moiety of bound ATP and form the binding site for adenosine. We included the amino acids at positions 118 and 119 because of their proximity to the amino acid at position 120. Similarly, the amino acid at position 130 (shown in blue in Figure 1) was included because it is located close to the amino acid at position 127.

Table 2. Kinases Used for 3D Alignment

group[a]	PDB ID	kinase
ACG	1ATP	cAMP-dependent protein kinase
ACG	1O6K	PKB kinase
CaMK	1PHK	phosphorylase kinase
CaMK	1NY3	MAP kinase-activated protein kinase 2
CMGC	1CM8	MAP kinase p38
CMGC	1FIN	cyclin-dependent kinase 2
TK	1IR3	insulin receptor tyrosine kinase
TK	1JQH	IGF-1 receptor kinase

[a] According to ref 3.

Sequence Alignments. The sequence alignments were taken from Manning et al.[3] and were checked against the 3D alignments of the X-ray structures of the kinases listed in Table 2. All the kinases listed in this table included ATP analogues as ligands in the crystal structure. They were aligned by the combinatorial extension (CE) method.[14,15] 1CM8 and 1FIN, which belong to the CMGC group, are one amino acid shorter than the kinases belonging to the other three groups, and there is a gap at position 125 (shown in gray in Figure 1). A gap was placed at position 125. For the other amino acids, the alignments of Manning et al.[3] were used without modification. When the number of amino acids at the adenosine-binding site of any kinase was greater than the number of amino acids at the adenosine-binding site of cAMP-dependent protein kinase, that kinase was excluded from the analysis in order to avoid ambiguous alignments. The final selection consisted of 275 kinases; the numbers of kinases in the AGC, CaMK, CMGC, and TK groups were 60, 64, 61, and 90, respectively (Table 3).

Selection of Physicochemical Descriptors. During our quantitative structure−activity relationships (QSAR) studies on biologically active peptides,[6] we found that several physicochemical descriptors effectively rationalized the biological activities of the peptides. The descriptors were also shown to be useful to rationalize the structural stabilities of proteins and the β-sheet propensities of the amino acids.[9] On the basis of these results, the physicochemical descriptors listed in Table 4 were selected for use in the present study.

The first descriptor, π_b, represents the hydrophobicity of the amino acids. Akamatsu et al.[7,8] have reported the 1-octanol/water partition coefficients for the *N*-acetyl dipeptide and tripeptide amides in their neutral forms. We analyzed these partition coefficients by multiple regression analysis and defined our π_b descriptors representing the hydrophobicity of the amino acids.[6]

The second and third descriptors, δHc and L, represent structural properties. We have previously found that the β-sheet propensities of amino acids are highly correlated with the nuclear magnetic resonance (NMR) chemical shift of the α-carbon of the amino acids, δHc.[9] δHc was chosen for our classification studies because the β-sheet propensities of the amino acids represent structural properties of the amino acids, and the adenosine-binding site includes an extended β-sheet-like structure. The Sterimol descriptors developed by Verloop et al.[16] are a popular set of steric descriptors in QSAR studies. The descriptors comprise L, B1, and B5, which are measures of, respectively, the length, minimal width, and maximum width, of a substituent. The Sterimol descriptors used in the present study were those calculated by Verloop et al.[16] Because B1 is highly correlated with δHc and B5 is correlated with L, L was selected to represent the dimensional properties of the amino acids.

The fourth descriptor, isoelectric point (pI), expresses the electronic properties of the amino acids.[17] The pI of an amino acid is the pH at which it has an equal number of positive and negative charges and thus carries no net charge. Amino acids with basic side chains have high pI values, and those with acidic side chains have low pI values.

Neural Network Modeling. Neural networks were modeled with STATISTICA Neural Networks Release 4.0E by StatSoft, Inc., run on a Pentium III desktop computer.[18] All networks were modeled with the "Intelligent Problem Solver" implemented by STATISTICA Neural Networks, which automates the choice of the network architecture and input descriptor selection by comparing multiple potential solutions with different network types combined with different selections of input descriptors. Input descriptors were selected by using a variety of search, regularization, and sensitivity techniques.[18] Multilayer perceptron (MLP) neural networks with softmax outputs were trained by the back-propagation algorithm and the conjugate gradient method.[19] Each network had one input layer, one hidden layer, and one output layer. The number of units in the input layer is the number of input descriptors, while the number of units in the output layer is the number of output categories. The number of units in the hidden layer is then automatically determined. A training set (50%), a verification set (25%), and a test set (25%) were randomly selected for each round of modeling. An MLP neural network was trained on the training data set. To prevent the neural network from overfitting the training data, it was evaluated on its ability to make correct predictions of the verification data set. The predictive power of the trained neural network was then checked with the test set.[19]

Classification Procedures. The details of the neural network modeling procedures we used are shown in Figure 2. In step 1, the training (50%), verification (25%), and test (25%) sets were randomly selected. With the Intelligent Problem Solver,[18] all neural network architecture and the input descriptors were automatically trained, and 10 neural networks with different architectures and input descriptors were produced. Because we intended to investigate the positions and properties of the amino acids governing the classification, we selected from the 10 modeled neural networks a neural network having fewer than eight input descriptors. When more than one neural network was modeled with fewer than eight input descriptors, the best-performing neural network was selected. Because different data sets usually gave somewhat different neural networks, step 1 was repeated seven times, and the seven resulting neural networks were selected for the next step.

Some input descriptors were used in all seven neural networks obtained in step 1, but some appeared in only a few of them. To obtain the relevant descriptors, in step 2 we selected those descriptors that appeared five or more times in the seven neural networks obtained in step 1. Obtaining the relevant descriptors is the major objective of this study, since they reveal which amino acids and which properties are important for the classification. (The results are shown in Figures 3 and 4 and Table 5.)

In step 3, the classification abilities of the relevant descriptors were tested. For each data set prepared in step 1, the Intelligent Problem Solver provided 10 networks, and the

Table 3. Kinases Used for Classification

AGC (60)[a]
AKT1, AKT2, AKT3, SGK, SGK3, SGK2, PKCa, PKCb, PKCg, PKCh, PKCe, PKCd, PKCt, PKCi, PKCz, PKN1, PKN2, PKN3, MSK1, MSK2, p70S6K, p70S6Kb, RSK1, RSK2, RSK3, RSK4, PRKX, PRKY, PKACa, PKACb, PKACg, PKG1, PKG2, MAST1, MAST4, MAST2, MAST3, MRCKb, MRCKa, DMPK2, DMPK1, ROCK2, ROCK1, NDR1, NDR2, LATS1, LATS2, CRIK, BARK1, BARK2, GPRK4, GPRK5, GPRK6, RHOK, GPRK7, PDK1, YANK3, YANK2, YANK1, MASTL
CaMK (64)
CaMK1a, CaMK1d, CaMK1g, CaMK1b, CaMK4, VACAMKL, PSKH1, PSKH2, DCAMKL1, DCAMKL2, DCAMKL3, CaMK2a, CaMK2d, CaMK2b, CaMK2g, CASK, PHKg1, PHKg2, MAPKAPK2, MAPKAPK3, MAPKAPK5, MNK1, MNK2, CHK2, STK33, AMPKa1, AMPKa2, BRSK2, BRSK1, MARK2, MARK1, MARK3, MARK4, QIK, SIK, QSK, NuaK1, NuaK2, MELK, NIM1, SNRK, HUNK, TSSK3, TSSK2, TSSK1, TSSK4, PIM3, PIM1, PIM2, PASK, CHK1, DAPK1, DAPK3, DAPK2, DRAK1, DRAK2, caMLCK, SgK085, skMLCK, smMLCK, TTN, Trio, Trad, Obscn
CMGC (61)
CDK2, CDK3, CDC2, CDK5, PCTAIRE1, PCTAIRE2, PCTAIRE3, PFTAIRE1, PFTAIRE2, CDK4, CDK6, CDK10, PITSLRE, CDK7, CCRK, CHED, CRK7, CDK9, CDK8, CDK11, Erk1, Erk2, Erk5, NLK, JNK1, JNK3, JNK2, p38a, p38b, p38g, p38d, Erk7, Erk3, Erk4, MAK, ICK, MOK, CDKL2, CDKL3, CDKL1, CDKL4, CDKL5, GSK3A, GSK3B, CLK1, CLK4, CLK2, CLK3, SRPK2, MSSK1, SRPK1, DYRK1B, DYRK1A, DYRK2, DYRK3, DYRK4, HIPK1, HIPK2, HIPK3, HIPK4, PRP4
TK (90)
ABL, ARG, BTK, BMX, TEC, TXK, ITK, HCK, LYN, LCK, BLK, SRC, YES, FYN, FGR, FRK, BRK, SRM, CSK, CTK, FER, FES, FGFR2, FGFR3, FGFR1, FGFR4, RET, KDR, FLT1, FLT4, FMS, KIT, PDGFRa, PDGFRb, FLT3, TIE2, TIE1, ALK, LTK, ROS, INSR, IGF1R, IRR, DDR1, DDR2, TRKB, TRKC, TRKA, MUSK, ROR1, ROR2, CCK4, AXL, MER, TYRO3, MET, RON, RYK, ACK, TNK1, EGFR, HER2/ErbB2, HER4/ErbB4, HER3/ErbB3, SYK, ZAP70, EphA3, EphA5, EphA4, EphA6, EphA7, EphB1, EphB2, EphB3, EphB4, EphA8, EphA2, EphA1, EphB6, EphA10, FAK, PYK2, JAK1, TYK2, JAK2, JAK3, LMR1, LMR3, LMR2, SuRTK106

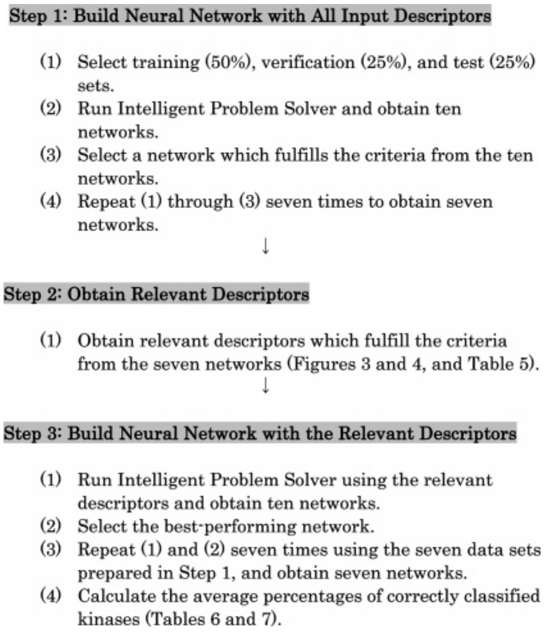
[a] The figures in parentheses are the number of kinases in each group.

Table 4. Physicochemical Descriptors of Amino Acids Used for Classification

amino acid		π_b[a]	δHc[b]	L[b]	pI[c]
A	Ala	0.16	7.3	2.87	6.00
C	Cys	0.56	14.4	4.47	5.05
D	Asp	−0.03	9.2	4.74	2.77
E	Glu	−0.05	11.4	5.97	3.22
F	Phe	1.72	13.9	4.62	5.48
G	Gly	0.00	0.0	2.06	5.97
H	His	0.10	10.2	5.23	7.59
I	Ile	1.37	16.1	4.92	6.02
K	Lys	0.53	10.9	6.89	9.74
L	Leu	1.47	10.1	4.92	5.98
M	Met	0.94	10.4	6.36	5.74
N	Asn	−0.59	8.0	4.58	5.41
P	Pro	0.76	17.8	4.11	6.30
Q	Gln	−0.58	10.6	6.11	5.65
R	Arg	−0.50	11.1	7.82	10.76
S	Ser	−0.27	13.1	3.97	5.68
T	Thr	0.01	16.7	4.11	5.66
V	Val	0.95	17.2	4.11	5.96
W	Trp	1.97	13.2	7.68	5.89
Y	Tyr	1.06	13.9	4.73	5.66

[a] Taken from ref 6. [b] Taken from ref 16. [c] Taken from ref 17.

best-performing network was selected as a representative network. Using the seven data sets prepared in step 1, we modeled seven neural networks. (The average percentages of the correctly classified kinases for the seven neural networks are shown in Tables 6 and 7.)



Step 1: Build Neural Network with All Input Descriptors

(1) Select training (50%), verification (25%), and test (25%) sets.
(2) Run Intelligent Problem Solver and obtain ten networks.
(3) Select a network which fulfills the criteria from the ten networks.
(4) Repeat (1) through (3) seven times to obtain seven networks.

↓

Step 2: Obtain Relevant Descriptors

(1) Obtain relevant descriptors which fulfill the criteria from the seven networks (Figures 3 and 4, and Table 5).

↓

Step 3: Build Neural Network with the Relevant Descriptors

(1) Run Intelligent Problem Solver using the relevant descriptors and obtain ten networks.
(2) Select the best-performing network.
(3) Repeat (1) and (2) seven times using the seven data sets prepared in Step 1, and obtain seven networks.
(4) Calculate the average percentages of correctly classified kinases (Tables 6 and 7).

Figure 2. Flowchart of classification procedure.

RESULTS

Classification into Two Categories. We first investigated whether an MLP neural network could discriminate between kinases belonging to a certain group and those not belonging to that group. To do so, we divided the data set comprising

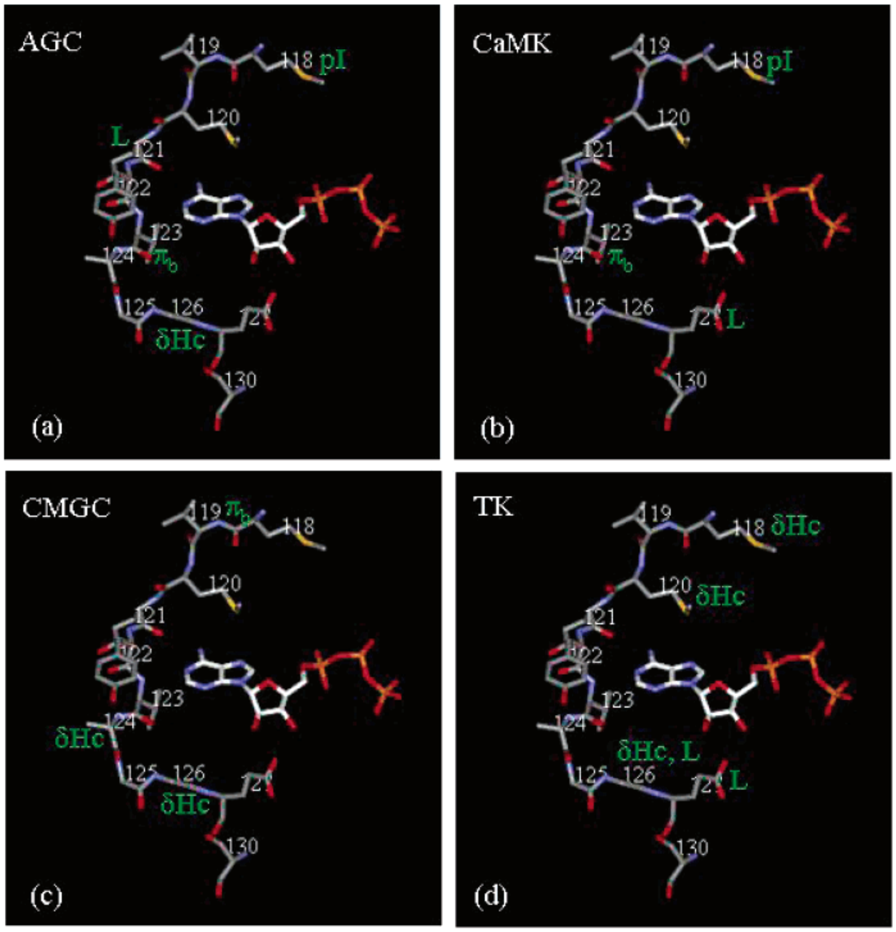


Figure 3. Results for classification into two categories. The relevant descriptors are shown in green. π_b is the hydrophobic descriptor of the amino acids, δHc is the nuclear magnetic resonance (NMR) chemical shift of the α-carbons of the amino acids, L is the Verloop steric descriptor representing the length of the amino acid side chains, and pI is the isoelectric point of the amino acids.

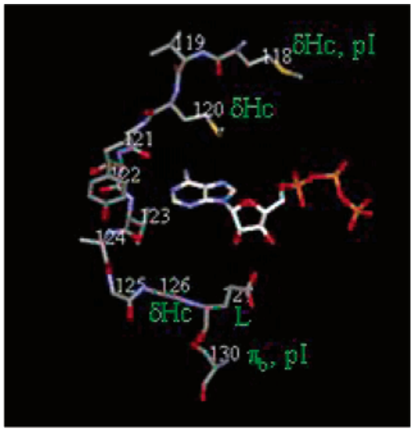


Figure 4. Results for classification into four categories. Details are as in the caption to Figure 3.

275 kinases into two categories. For example, for the AGC group, one category consisted of kinases belonging to the AGC group of 60 kinases, and the other included the remaining 215 kinases. The classification used the nine amino acids shown in green in Figure 1 (positions 118−124, 126, and 127), and thus the input data consist of 36 (9 × 4) descriptors. The significant descriptors among these were selected with the Intelligent Problem Solver (step 1 in Figure 2).

Table 5. Relevant Descriptors and the Architecture of Trained Neural Networks

kinase group	relevant descriptors[a]	hidden units[b]
	Classification into Two Output Categories	
AGC	pI(118), L(121), π_b(123), δHc(126)	4−7 (6)
CaMK	pI(118), π_b(123), L(127)	3−7 (4)
CMGC	π_b(119), δHc(124), δHc(126)	2−5 (3)
TK	δHc(118), δHc(120), δHc(126), L(126), L(127)	3−8 (7)
	Classification into Four Output Categories	
ALL	δHc(118), pI(118), δHc(120), δHc(126), L(127), δHc(130), pI(130)	6−13 (9)

[a] The relevant descriptors are shown, with the positions of the amino acids in parentheses. For example, pI(118) refers to the pI descriptor for amino acid 118. [b] The numbers of hidden units are shown. Figures in parentheses are the average values for seven neural networks. The number of units in the input layer is the number of relevant descriptors, while the number of units in the output layer is the number of output categories. Each network has one input layer, one hidden layer, and one output layer.

The seven different data sets gave seven somewhat different neural networks; some input descriptors appeared in all seven neural networks, and some appeared in only a few. The relevant descriptors were obtained by choosing those descriptors that appeared in five or more neural networks (step 2). These descriptors show some interesting features (Figure 3 and Table 5). First, the relevant descriptors differ

Table 6. Classification Performance of Trained Neural Network with Two Output Categories

	training set[a]		verification set[a]		test set[a]		all[a]	
kinase group	C+[b]	C−[b]	C+	C−	C+	C−	C+	C−
AGC	91.4	92.3	90.5	91.8	90.5	90.5	91.0	91.7
CaMK	85.7	87.3	85.7	85.6	83.0	87.8	85.0	87.0
CMGC	98.1	98.1	98.1	97.9	97.1	98.4	97.9	98.1
TK	93.9	95.1	93.7	94.7	93.0	94.1	93.7	94.7

[a] The percentages of correctly classified kinases are shown. [b] For AGC, C+ means the percentage of kinases predicted to belong to the AGC group. C− is the percentage of the kinases predicted not to belong to the AGC group. For the other kinase groups, C+ and C− are similarly defined.

Table 7. Classification Performance of the Trained Neural Networks with Four Output Categories

kinase group	training set[a]	verification set[a]	test set[a]	all[a]
AGC	99.0	96.2	91.4	96.4
CaMK	89.7	86.6	85.7	87.9
CMGC	100.0	99.0	98.2	99.3
TK	98.1	94.4	91.8	95.6

[a] The percentages of correctly classified kinases are shown.

from group to group. This clearly shows that the structural and physicochemical factors characterizing the binding site differ among the groups. Second, the descriptor that appeared most often is δHc. The structural factors expressed by δHc will be discussed later. Third, pI, π_b, and L also appeared often, so that they too worked well in the classification of the kinases.

In step 3, the percentages of the correctly classified kinases were calculated with the relevant descriptors selected as described above to check the classification abilities of the molded networks. As Table 6 shows, 91.4% of the AGC kinases in the training set were correctly classified, and 92.3% of the remaining kinases in the training set were predicted not to belong to the AGC group. More than 90.5% of the kinases were successfully classified not only for the verification set but also for the test set. The differences between the percentages for the training and test tests were small; therefore, it is unlikely that overfitting on the training data had occurred. Similar results were obtained for the other groups. The classification performance was satisfactory despite the fact that the modeling procedures were simple and fast.

The kinase domains (also known as catalytic domains) consist of approximately 250−300 amino acid residues, and these 250−300 residues have been used to classify kinases into groups.[4,5] However, in our neural network modeling studies we used only nine amino acid residues at the adenosine-binding site. Considering the small number of amino acid residues used, our classification procedures performed well.

Classification into Four Categories. MLP neural networks having four output groups were successively trained, and the results are listed in Table 7. Except for the differences in the number of the output categories, the classification procedures were the same as those described above for classification into two categories. Preliminary modeling suggested that the 36 descriptors used in the classification of the kinases into two categories were not sufficient for classification into four categories, so the amino acid at position 130 (shown in blue in Figure 1) was also included. This amino acid is located close to the amino acid at position 127, which in turn is in close contact with the ribose moiety of adenosine. The significant descriptors were selected from the resulting pool of 40 (10 × 4) input descriptors. As with classification into two categories, descriptors that appeared five or more times in the seven selected networks were selected (Figure 4 and Table 5). More than half of the descriptors in Figure 3 appeared in Figure 4, but new descriptors also appeared.

Only one MLP neural network successfully classified the 275 kinases into four categories, namely, the AGC, CaMK, CMGC, and TK groups. The percentage of kinases correctly classified was always greater than 85% and usually greater than 90% (Table 7). Even though we excluded the amino acids in the phosphate-binding site, and included only the amino acids in the adenosine-binding site, we could still satisfactorily classify the kinases. The classification performance with the external test set is an excellent metric to evaluate the quality of a trained MLP neural network. The differences between the classification performances for the verification and test sets were small, and this clearly demonstrates the high predictive power of our networks.

DISCUSSION

Although the interpretation of neural networks is generally difficult because of their complex structures, we have tried to elucidate the structural and physicochemical properties characterizing the adenosine-binding sites of each protein kinase group by comparing the distribution of amino acids at each position in the sequence at the adenosine-binding site. These data are provided as Supporting Information, with the distributions expressed as percentages.

Characteristic Structural Features of the AGC Group. The relevant descriptors for the AGC group were pI(118), L(121), π_b(123), and δHc(126), where the figures in parentheses are the positions of the amino acids in the sequence. While 87% of the amino acids at position 118 have alkyl side chains (Leu, Ile, and Val) in the CaMK, CMGC, and TK groups (hereafter referred to as the non-AGC group), only 38% of the amino acids at this position in the AGC group have alkyl side chains. Phe and Met constitute 37% and 25%, respectively, of the amino acids at position 118 in the AGC group and 1% and 11%, respectively, in the non-AGC group. Because the pI values of Phe and Met are lower than those of Leu, Ile, and Val, the greater acidity of the amino acids at this position characterizes the adenosine-binding site of the AGC group. Most of the amino acids at position 121 are acidic in both the AGC and the non-AGC groups; thus the percentages of Asp and Glu are 32% and 58%, respectively, in the AGC group and 7% and 82%, respectively, in the non-AGC group. The side chain of Asp is shorter than that of Glu, and the presence of the shorter acidic side chain is a noteworthy feature of the AGC group. At position 123, hydrophobic amino acids, such as Leu, Ile, and Val, occur more frequently in the AGC group than in the non-AGC group, and hydrophilic amino acids, such as Ala and Cys, occur more frequently in the non-AGC group than in the AGC group. The amino acids at this position are more hydrophobic in the AGC group than in the non-AGC group. While the identity of the amino acid at position 126 varies greatly among members of the CMGC group, it is always

Gly in the AGC group and it is usually Gly in the CaMK and TK groups. The δHc value for Gly is zero, and δHc for the other amino acids varies from 7.3 to 17.8. Thus δHc mainly reflects the presence or absence of Gly, which presumably distinguishes the AGC group from the CMGC group.

Characteristic Structural Features of the CaMK Group. The relevant descriptors for the CaMK group were pI(118), π_b(123), and L(127). The amino acids at position 118 are hydrophobic ones in both the CaMK and the non-CaMK groups, but Phe occurs in this position only in the non-CaMK group. The absence of the more acidic Phe at this position is a noteworthy feature of this group. While most of the amino acids at position 123 in the non-CaMK group are hydrophobic ones such as Met, Leu, Ile, and Val, the amino acids at this position in the CaMK group include more hydrophilic amino acids such as Ala, Glu, and Gly, in addition to the hydrophobic ones. The lower hydrophobicity of the amino acids at position 123 characterizes the adenosine-binding site of the CaMK group. The amino acids most frequently appearing at position 127 are Glu in the CaMK group and Asp in the non-CaMK group. The side chain of Glu is longer than that of Asp, and, furthermore, the proportion of amino acids with shorter side chains such as Ala and Ser is greater in the non-CaMK group than in the CaMK group. The length of the side chains at this position differentiates the CaMK group from the non-CaMK group.

Characteristic Structural Features of the CMGC Group. π_b(119), δHc(124), and δHc(126) were found to be relevant descriptors for the CMGC group. As for the amino acids at position 119, the percentages of Val, Leu, and Ile in that position were 80%, 3%, and 7%, respectively, in the CMGC group and 57%, 7%, and 32%, respectively, in the non-CMGC group. Because Val is less hydrophobic than Leu or Ile, the proportion of the less hydrophobic amino acids is greater in the CMGC group than in the non-CGMG group. The δHc values are large for Pro, Thr, and Val. The CMGC group has these amino acids at position 124 less frequently than does the non-CMGC group. The δHc value for Gly is zero, and Gly constitutes 18% of the amino acids at position 124 in the CMGC group but only 1% of the amino acids at this position in the non-CMGC group. We have previously shown that δHc is mainly correlated with the widths of the amino acid side chains.[9] The width of the amino acid side chain at position 124 differentiates the CMGC group from the non-CMGC group. Having amino acids with larger δHc values at position 126 is a distinguishing feature of the CMGC group.

Characteristic Structural Features of the TK Group. The relevant descriptors for the TK group were δHc(118), δHc(120), δHc(126), L(126), and L(127). Ile and Val have large δHc values, and Ile and Val constitute 47% and 21%, respectively, of the amino acids at position 118 in the TK group and 12% and 3%, respectively, of the amino acids at this position in the non-TK group. Thus the adenosine-binding site of the TK group is characterized by amino acids with large δHc values at position 118. Similarly, the TK group contains amino acids with large δHc values at position 120 (where Thr constitutes 46% of amino acids and Val 10%), while the non-TK group does not (Thr constitutes only 4% of amino acids at position 120 and Val only 2%). Again, differences in the widths of the amino acid side chains at position 120 distinguish the TK group from the non-TK group. The δHc value for the amino acids at position 126 can be interpreted for the TK group as for the AGC and CMGC groups. In addition to δHc, L was also found to be a relevant descriptor. There is great variation among the amino acids at this position, so that it is difficult to clarify the meaning of L. It is possible that the length of the amino acid side chains works cooperatively with other structural features to distinguish the TK group from the non-TK group. The TK group has amino acids with shorter side chains, such as Ala, Cys, and Ser, at position 127, whereas the non-TK group has no Ala or Cys, and only 3% Ser, at this position. The presence of amino acids with shorter side chains is a distinguishing feature of the adenosine-binding sites of the TK group.

Roles of Descriptors. The following summary of the roles of descriptors is based on the results we obtained in this study. pI served to reflect small differences in the electronic effects of hydrophobic amino acids. This may be because the adenosine-binding site, especially the adenine-binding part of it, is hydrophobic, and the number of amino acids with ionizable side chains is small. In our preliminary studies on the neural network classification of GPCRs (unpublished results), pI represented the electronic effects of ionizable amino acids. The physicochemical meanings of L and π_b are clear, and they represented well the lengths and hydrophobicities of amino acids, respectively. The descriptor which appeared the most frequently in the neural networks was δHc, and it reflected not only the presence or absence of Gly but also the widths of the side chains. As an alternative to δHc, we tried to use a descriptor reflecting the presence or absence of Gly and a steric descriptor reflecting the widths of the side chains such as Sterimol B1,[16] but we could not obtain improved models. We conclude that δHc performs well by working differently in different situations.

We found no remarkable differences in the properties expressed by the relevant descriptors. In addition, the number of relevant descriptors was small for all groups. This is reasonable because all kinases recognize the same ligand, ATP, and therefore the structural features of the binding sites would be expected to be very similar among all kinases. Nevertheless, we successfully classified kinases and elucidated the characteristic features of their adenosine-binding sites. Amino acid descriptors represented well the structural features of the adenosine-binding sites. In neural networks, selected amino acid descriptors worked “cooperatively” to classify kinases. Thus our study shows that the combined use of amino acid descriptors and neural networks is a powerful method of discriminating between structurally very similar binding sites.

Application to Drug Development. There are more than 500 protein kinases,[3] so that selectivity is an essential requirement for clinically effective drugs. Choosing targets suitable for selective inhibitors is thus critically important. The structural characteristics of the CaMK group are rather similar to those of the AGC group (Figure 3a,b). Therefore it may not be easy to identify inhibitors which discriminate between the adenosine-binding sites of the CaMK and the AGC groups. To develop inhibitors specifically targeting either of these groups, it is advisable to identify additional amino acids, for example in the phosphate-binding site, with which inhibitors can specifically interact. In contrast, there are many physicochemical properties characterizing the TK group, which suggests that tyrosine kinases represent good targets for drug development.

The graphical representations we obtained (Figures 3 and 4) are also helpful in the design of selective inhibitors. The amino acids labeled in Figure 3d are candidate amino acids to be considered in the design of selective kinase inhibitors specific for the TK group. For example, the inclusion of amino acid 120 in the TK group suggests that drugs with functional groups which specifically interact with this amino acid will be promising. Indeed, a hydrogen-bonding interaction between the amino group of Imatinib and the Thr at position 120 of Abl kinase is essential for the action of that drug.[20] To design kinase inhibitors which are specific for a certain kinase in the TK group, the amino acids not labeled in Figures 3d and 4 are the candidate amino acids which should be considered, and the introduction of functional groups which interact with, for example, amino acids 124 and 125 would be expected to yield interesting results. These are only a few examples, and there are many other ways to generate ideas from our results at every stage of drug development from target selection to inhibitor optimization.

Kinase inhibitors can bind to either the active or the inactive forms of their target or to both; for example, Imatinib binds only to the inactive conformation of the Abl catalytic domain,[20] whereas Tarceva binds to the active conformation of the EGFR catalytic domain.[21] The structural differences between the active and inactive conformations are mainly in the phosphate-binding site. For classification, we excluded amino acids in the phosphate-binding site and included only amino acids in the adenosine-binding site. Therefore our classification results should be applicable to both the inactive and active conformations. Furthermore, the kinases which were misclassified are also worth considering. The sequence alignment is based on the amino acids in the entire catalytic domain of the kinases, whereas the present classification is based solely on the amino acids in the adenosine-binding site of the catalytic domain. It is thus reasonable to expect that some kinases would not be correctly classified. Such kinases may be useful in the search for other targets for existing kinase inhibitors as well as in the identification of kinases whose inhibition could cause undesirable effects.

Amino Acid Descriptors. Hydrophobic, steric, and electronic descriptors are widely used in QSAR studies, and it is not surprising that only four descriptors, π_b, δHc, L, and pI, were sufficient to express the properties of the amino acids in the adenosine-binding site. Calculating a large number of descriptors for amino acids is not difficult, and selecting the appropriate descriptors from a pool of descriptors might produce improved results. Nevertheless, limiting the number of input descriptors has certain advantages. First, different data sets usually give somewhat different neural networks. With a large number of input descriptors, neural networks would be modeled with various descriptors, and identifying the relevant descriptors might be difficult. Second, to understand the differences among the kinase groups, descriptors having clear physicochemical meanings are of great help. Third, we hope that the set of representative descriptors we have selected will be useful in future studies. As the results show, the four descriptors performed satisfactorily. Thus π_b, δHc, L, and pI can be regarded as a basic set of descriptors which express the properties of amino acids in proteins.

Neural Network Modeling. There are many kinds of neural networks, such as Hopfield, Kohonen, and radial basis function (RBF) neural networks as well as multilayer perceptrons.[19] This is our first attempt to use neural networks and the physicochemical descriptors of amino acids for the classification of protein kinases. That is why we selected MLP neural networks, the most popular type of neural network, and trained them automatically. Despite our use of such simple modeling procedures, we obtained very informative results. This demonstrates the power of using MLP neural networks and amino acid descriptors together. Of course, the application of other modeling methods is also worth considering. The objective of this study was to elucidate the factors governing the classification of protein kinases but not necessarily to obtain networks with very high classification abilities. To obtain such networks, improvement would be needed in several areas. For example, increasing the number of amino acids selected for classification, the number of input descriptors, and the complexity of the architectures would be beneficial.

General Applications. In this study, we used physicochemical descriptors of amino acids to elucidate the structural characteristics of kinases. It is the physicochemical properties of amino acids that govern many biologically important interactions, such as protein−ligand and protein−protein interactions. The descriptors we used here represent well the physicochemical properties of amino acids. These descriptors are expected to be applicable to the analysis of various kinds of interactions involving proteins.

Of course, modeling methods are not limited to neural networks. Since the descriptors used here are simple and easy to handle, they are applicable to various methods, such as self-organization maps (SOM), multidimensional scaling (MDS), support vector machines (SVM), and decision trees. For example, genome-wide comparison of the characteristic structural features of ligand-binding sites and investigations of the patterns of interaction at protein−protein interfaces which make use of our four descriptors should also prove informative. Therefore the method we present in this study is expected to be widely applicable.

CONCLUSION

The combined use of physicochemical descriptors and neural networks has allowed us to identify characteristic structural and physicochemical properties of the adenosine-binding sites of protein kinases from aligned sequences and to elucidate the differences in the structural characteristics of various kinase groups. Comparison of the modeled neural networks and the distributions of amino acids at certain positions in the sequence at the adenosine-binding sites provided more-detailed information on the structural characteristics of the binding sites. A graphical presentation of our results is of assistance at the various stages of drug development. Our newly developed method is also expected to be widely applicable in studies of protein−ligand and protein−protein interactions.

ACKNOWLEDGMENT

We thank Dr. Gerald E. Smyth for his careful reading of the manuscript and for his useful comments during its preparation.

Supporting Information Available: Tables of the distribution of amino acids at each position in the adenosine-binding site of protein kinase groups. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Noble, M. E.; Endicott, J. A.; Johnson, L. N. Protein kinase inhibitors: insights into drug design from structure. *Science* **2004**, *303*, 1800−1805.

(2) Capdeville, R.; Buchdunger, E.; Zimmermann, J.; Matter, A. Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug. *Nat. Rev. Drug Discovery* **2002**, *1*, 493−502.

(3) Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The protein kinase complement of the human genome. *Science* **2002**, *298*, 1912−1934.

(4) Hanks, S. K.; Quinn, A. M.; Hunter, T. The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science* **1988**, *241*, 42−52.

(5) Hanks, S. K.; Quinn, A. M. Protein kinase catalytic domain sequence database: identification of conserved features of primary structure and classification of family members. *Methods Enzymol.* **1991**, *200*, 38−62.

(6) Sotomatsu-Niwa, T.; Ogino, A. Evaluation of the hydrophobic parameters of the amino acid side chains of peptides and their application in QSAR and conformational studies. *THEOCHEM* **1997**, *392*, 43−54.

(7) Akamatsu, M.; Katayama, T.; Kishimoto, D.; Kurokawa, Y.; Shibata, H.; Ueno, T.; Fujita, T. Quantitative analyses of the structure-hydrophobicity relationship for *N*-acetyl di- and tripeptide amides. *J. Pharm. Sci.* **1994**, *83*, 1026−1033.

(8) Akamatsu, M.; Fujita, T. Quantitative analyses of hydrophobicity of di- to pentapeptides having un-ionizable side chains with substituent and structural parameters. *J. Pharm. Sci.* **1992**, *81*, 164−174.

(9) Niwa, S. T.; Ogino, A. Multiple regression analysis of the beta-sheet propensity of amino acids. *THEOCHEM* **1997**, *419*, 155−160.

(10) Niwa, T. Prediction of biological targets using probabilistic neural networks and atom-type descriptors. *J. Med. Chem.* **2004**, *47*, 2645−2650.

(11) Niwa, T. Using general regression and probabilistic neural networks to predict human intestinal absorption with topological descriptors derived from two-dimensional chemical structures. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 113−119.

(12) Hanks S. K. http://www0.nih.go.jp/mirror/Kinases/pkr/pk_catalytic/pk_hanks_seq_align_long.html (accessed May 2006).

(13) Kostich, M.; English, J.; Madison, V.; Gheyas, F.; Wang, L.; Qiu, P.; Greene, J.; Laz, T. M. Human members of the eukaryotic protein kinase family. *Genome Biol.* **2002**, *3*, research0043.1−0043.12.

(14) Shindyalov, I. N.; Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **1998**, *11*, 739−747.

(15) Guda, C.; Lu, S.; Scheeff, E. D.; Bourne, P. E.; Shindyalov, I. N. CE-MC: A multiple protein structure alignment server. *Nucleic Acids Res.* **2004**, *32*, W100−103.

(16) Fauchere, J. L.; Charton, M.; Kier, L. B.; Verloop, A.; Pliska, V. Amino acid side chain parameters for correlation studies in biology and pharmacology. *Int. J. Pept. Protein Res.* **1988**, *32*, 269−278.

(17) Zimmerman, J. M.; Eliezer, N.; Simha, R. The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.* **1968**, *21*, 170−201.

(18) *STATISTICA Neural Networks Release 4.0E*; StatSoft, Inc.: Tulsa, OK.

(19) Bishop, C. M. *Neural Networks for Pattern Recognition*; Oxford University Press: New York, 1995; pp 116−193.

(20) Schindler, T.; Bornmann, W.; Pellicena, P.; Miller, W. T.; Clarkson, B.; Kuriyan, J. Structural mechanism for STI-571 inhibition of abelson tyrosine kinase. *Science* **2000**, *289*, 1938−1942.

(21) Wood, E. R.; Truesdale, A. T.; McDonald, O. B.; Yuan, D.; Hassell, A.; Dickerson, S. H.; Ellis, B.; Pennisi, C.; Horne, E.; Lackey, K.; Alligood, K. J.; Rusnak, D. W.; Gilmer, T. M.; Shewchuk, L. A unique structure for epidermal growth factor receptor bound to GW572016 (Lapatinib): relationships among protein conformation, inhibitor off-rate, and receptor activity in tumor cells. *Cancer Res.* **2004**, *64*, 6652−6659.

CI050528T