QSAR Analysis of SH2-Binding Phosphopeptides: Using Interaction Energies and Cross-Correlation Coefficients

Taesung Moon, Jin-Soo Song, Jin Kak Lee, and Chang No Yoon*

Bioanalysis and Biotransformation Research Center, Korea Institute of Science and Technology, P.O. Box 131, Cheongryang, Seoul 130-650, Korea

Received April 16, 2003

Quantitative structure−activity relationships (QSAR) analyses were carried out on the SH2-phosphopeptide complexes using multiple linear regressions. The residue−residue interaction energies and cross-correlation coefficients were used as descriptors. Since the number of descriptors was very large (602 for interaction energies and 951 for cross-correlation coefficients), the stepwise addition method was applied for the multiple linear regressions. The residue−residue interaction energies were good descriptors for structure−activity relationships. The high r^2 regression models were achieved by using interaction energy. In addition, the concerted atomic motions, which show the dynamic properties during the SH2-phosphopeptide interaction, were used as descriptors. They were identified by the cross-correlation coefficients for atomic displacement. The best regression model, derived by using four cross-correlation coefficients, gave a high r^2 value of 0.925. This suggests that the dynamic properties showing concerted atomic motions can be used as good descriptors in QSAR study.

INTRODUCTION

The src homology 2 (SH2) domains mediate specific protein−protein interactions through binding to specific, phosphotyrosine-containing peptide sequences in their targets. Tyrosine phosphorylation of cellular target proteins mediates normal signal transduction and plays a significant role in regulation of cell growth, differentiation, and oncogenesis.[8,21,22] The cellular level of tyrosine phosphorylation is maintained by kinase and phosphatase. Tyrosine kinase activity is implicated in the activation of growth factor receptor such as EGF-R (epidermal growth factor receptor) and PDGF-R (platelet-derived growth factor receptor). The binding of growth factor to its surface receptor leads to activation of intrinsic protein kinase activity which is followed by tyrosine autophosphorylation of target substrate proteins.[21] In addition to the kinase catalytic domain (SH1; src homology-1), tyrosine kinases of the src family have highly conserved amino acid motifs named SH2 (src homology-2) and SH3 (src homology-3) which are noncatalytic domains.[9] The SH3 domain is a distinct motif which may modulate interactions with the cytoskeleton and membrane together with SH2 domain. The transmission of growth factor-mediated signals, for example, depends on the sequence-specific recognition of phosphorylated tyrosines by SH2 domains. Many studies report that they are involved in mediating protein−protein interactions in the downstream of growth factor receptors, including GAP (Ras GTPase-activating protein), PIK (phosphatidylinocitol 3′-kinase), and phospholipase C-γ[3]. The function of the SH2 domain is less clear. The SH2 may play a role in the regulation of kinetic activity itself. When Tyr-527 at the c-terminus of src is phosphorylated, it suppresses the kinase activity.[1,5] Substitution of Tyr-527 with phenylalanine increases kinase activity.[4] Many crystal and solution structures of SH2 have been reported,[6,7,13,14,16,20,23,24] and crystal structures of the SH2 from src and lck with a phosphopeptide containing pYEEI motif have been determined.[6,24] These studies show the presence of two pockets. One pocket binds the pY (phosphotyrosine) with ion-paring, hydrogen-bonding, and amino-aromatic interactions. The other pocket binds the pY+3 (isoleucine) with hydrophobic interactions. The peptide is anchored by insertion of pY and pY+3 side chains into their pockets. In the low affinity phosphopeptide/v-src complex, the pY+3 do not insert to the binding pocket, and the peptide main chain is displaced from the surface of the domain.[24] Quantitative studies of the relative affinities of synthetic phosphopeptides for SH2 domain have been carried out.[17−19] In this work the quantitative structure−activity relationships (QSAR) studies were carried out to obtain future insight into the relationships between the structure and biological activity of the phosphotyrosine-containing peptides for Lck SH2 domain. The multiple linear regression methods were performed to find the relationships. A data set of 30 phosphopeptides[17] that bind SH2 domain was used for analysis.

EXPERIMENTAL METHODS

Published data on 30 pY324 phosphopeptides for GST-Lck SH2[17] were used for this study (Table 1). Of the 30 phosphopeptides the 25 ones that do not have an experimental error were used in the regression. The activity of phosphopeptides is expressed as IC_{50} values by multiplying relative affinities of the original paper by 2.87 μM.[17]

Conformation Sampling. For the construction of the phosphopeptide-binding structures, the crystal structure of pY324 SH2/phosphopeptide complex[6] was used as the template for other 29 structures. Each structure was built by

* Corresponding author phone: +82-2-958-5068; fax: +82-2-958-5059; e-mail: cody@kist.re.kr.

10.1021/ci034073o CCC: $25.00
Published on Web 07/23/2003

Table 1. Sequences and Relative Affinities of Altered PY324 Phosphopeptides for GST-LckSH2[17]

no.	phosphopeptide	phosphopeptide sequence	relative affinity	IC_{50} (μM)	$-\log IC_{50}$
1	ac-py324+1a	Ac-pYAEIPI	42	120.540	3.9189
2	ac-py324+1d	Ac-pYDEIPI	13	37.310	4.4282
3	ac-py324+1e	Ac-pYEEIPI	1.7	4.879	5.3117
4	ac-py324+1f	Ac-pYFEIPI	18	51.660	4.2868
5	ac-py324+1g	Ac-pYGEIPI	>47	>134.890	~3.8700
6	ac-py324+1i	Ac-pYIEIPI	18	51.660	4.2868
7	ac-py324+1k	Ac-pYKEIPI	≫47	≫134.890	~3.8700
8	ac-py324+1l	Ac-pYLEIPI	47	134.890	3.8700
9	ac-py324+1m	Ac-pYMEIPI	26	74.620	4.1271
10	ac-py324+1q	Ac-pYQEIPI	5.9	16.933	4.7713
11	ac-py324+1s	Ac-pYSEIPI	17	48.790	4.3117
12	ac-py324+4a-NH2	Ac-pYEEIA-NH2	3.2	9.184	5.0370
13	py324	EPQpYEEIPIYL	1.0	2.870	5.5421
14	py324+2i	EPQpYEIIPIYL	1.6	4.592	5.3380
15	py324+3a	EPQpYEEAPIYL	10	28.700	4.5421
16	py324+3e	EPQpYEEEPIYL	6.1	17.507	4.7568
17	py324+3l	EPQpYEELPIYL	4.8	13.776	4.8609
18	py324+3m	EPQpYEEMPIYL	2.7	7.749	5.1108
19	py324+3q	EPQpYEEQPIYL	24	68.880	4.1619
20	py324+3v	EPQpYEEVPIYL	2.2	6.314	5.1997
21	py324+4n	EPQpYEEINIYL	3.5	10.045	4.9981
22	py355	ASQVpYFTYDPYSE	260	746.200	3.1271
23	py697	GGVDpYKNIHLE	430	1234.100	2.9086
24	py708	LEKKpYVRRDSG	≫760	≫2181.200	~2.6613
25	py723	VDTpYVEMRPVS	>760	>2181.200	~2.6613
26	py730	CVSpYVVPTKAD	270	774.900	3.1108
27	py745	IGSpYIERDVTG	760	2181.200	2.6613
28	py809	NDSNpYIVKGNA	>760	>2181.200	~2.6613
29	py936	NHIpYSNLANSS	120	344.400	3.4629
30	py969	QPNNpYQFS	140	401.800	3.3960

the replacement of corresponding residue, and the orientation of the side chain was adjusted manually. All 30 structures were then subjected to molecular dynamics simulations to get the optimum conformations. Prior to molecular dynamics, these complexes were immersed in a water box and energy-minimized using steepest descents for 200 steps and conjugate gradients for 300 steps with periodic boundary conditions. By this procedure short contacts in the molecules were relieved, and the water positions were adjusted. The minimized coordinates were then used as a starting point for 25-ps molecular dynamics simulations at 300 K using periodic boundary conditions. The initial velocities were taken from a Maxwell−Boltzmann distribution for target temperature. The leapfrog algorithm was used to integrate the equations of motion with an integration time step of 1 fs. The minimum energy conformations during simulations were sampled and minimized for QSAR analysis. All the molecular mechanics and dynamics calculations were carried out using Discover2.98 program (Accelrys Inc.) with CVFF (Consistent Valence Force Field) using a nonbonded cutoff of 9.0 Å running on Silicon Graphics O2 workstation.

Statistical Analysis. The multiple linear regression is as follows

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \epsilon_i \quad i = 1, 2, ..., n \tag{1}$$

where the observed values of the y_i's are the dependent variables, the χ_{i1}'s, χ_{i2}'s, ..., χ_{ik}'s are the sets of the k descriptors, β_0, β_1, ..., β_k are the regression coefficients, and the ϵ_i's are independently distributed normal errors. Since the number of descriptors was much larger than that of dependent variables, the stepwise addition method was applied for the multiple linear regression analysis. The subsets having high r^2 were listed in the order of decreasing r^2. Then, one descriptor was added to the subsets and the new subsets were listed again. This procedure was repeated to reach a given number of descriptors. The multicollinearity among descriptors was identified using variance inflation factor (VIF).[10] The VIF for the ith regression coefficient is expressed as

$$\mathrm{VIF} = \frac{1}{1 - r_i^2} \tag{2}$$

where r_i^2 is the coefficient produced by regressing the descriptor x_i against the other descriptors, the x_j ($j \neq i$). The models of which VIF is greater than 10 were not considered. The quality of the model is quantitated in terms of r^2 which is defined as

$$r^2 = 1.0 - \frac{\Sigma(y_{\mathrm{calc}} - y_{\mathrm{actual}})^2}{\Sigma(y_{\mathrm{actual}} - y_{\mathrm{mean}})^2} \tag{3}$$

where y_{calc}, y_{actual}, and y_{mean} are the calculated, actual, and mean values of the target property, respectively.

The residue−residue interaction energies and cross-correlation coefficients[2,11,12] were used as descriptors in multiple linear regression analysis. The interaction energies were calculated from the minimum energy conformations during simulations, and cross-correlation coefficients were calculated from 15 to 25 ps molecular dynamics trajectories. Only the phosphopeptide-binding residues of SH2 domain (Table 2) and four residues of phosphopeptide (pY, pY+1, pY+2, and pY+3) were considered in the calculation of interaction energies and cross-correlation coefficients. Before regression analysis the descriptor sets were transformed so that each

Table 2. Phosphopeptide-Binding Sites of SH2 Domain

residues of phosphopeptide	binding sites of protein
PY (Pty)	Arg12 Arg32 Ser34 Glu35 Ser36 Thr37 Ser42 Ser44 His58 Tyr59 Lys60
PY+1	Lys57 Tyr59
PY+2	Arg62
PY+3	Tyr59 Ile71 Ser72 Arg74 Tyr87 Asp92 Gly93 Leu94

Table 3. 10 High r^2 Models Derived by Using Residue−Residue Interaction Energies[a]

			variables and coefficients				
model	r^2	*F* value	A	B	C	D	intercept
1	0.937	74.502	Et_{44-57} −0.687	Et_{pY1-57} −0.016	Ec_{pY3-87} +0.311	Ec_{pY3-93} +0.292	+3.466
2	0.936	73.438	Et_{pY1-32} −0.047	Et_{pY1-37} −0.912	Ec_{pY3-87} +0.247	Ec_{pY-wat} −0.014	+3.232
3	0.936	73.413	Et_{pY1-37} −0.912	Ec_{pY1-32} −0.047	Ec_{pY3-87} +0.247	Ec_{pY-wat} −0.014	+3.233
4	0.936	73.321	Ec_{pY1-32} −0.047	Ec_{pY1-37} −0.906	Ec_{pY3-87} +0.247	Ec_{pY-wat} −0.014	+3.238
5	0.936	73.081	Et_{pY1-32} −0.047	Ec_{pY1-37} −0.907	Ec_{pY3-87} +0.247	Ec_{pY-wat} −0.014	+3.237
6	0.935	72.056	Et_{44-57} −0.685	Ec_{pY1-57} −0.016	Ec_{pY3-87} +0.307	Ec_{pY3-93} +0.289	+3.497
7	0.934	70.618	Et_{pY1-32} −0.047	Et_{pY1-37} −0.907	Et_{pY-wat} −0.014	Ec_{pY3-87} +0.243	+3.140
8	0.934	70.594	Et_{pY1-37} −0.909	Et_{pY-wat} −0.014	Ec_{pY1-32} −0.047	Ec_{pY3-87} +0.243	+3.141
9	0.934	70.335	Et_{pY1-32} −0.047	Et_{pY-wat} −0.014	Ec_{pY1-37} −0.904	Ec_{pY3-87} +0.243	+3.146
10	0.934	70.309	Et_{pY-wat} −0.014	Ec_{pY1-32} −0.047	Ec_{pY1-37} −0.903	Ec_{pY3-87} +0.243	+3.147

[a] Et and Ec are total and coulombic interaction energies, respectively. $[-\log(IC_{50}) = Coef_A \cdot A + Coef_B \cdot B + Coef_C \cdot C + Coef_D \cdot D + Intercept]$.

set had minimum value of 0.0 and then divided by its average value. The descriptor sets which have a standard deviation below a certain threshold (0.1) were eliminated. The final 602 energy terms and 951 cross-correlation coefficient terms were used in the linear regression analysis.

RESULTS AND DISCUSSION

Interaction Energy. Table 3 shows the good regression models (10 high r^2 models) in which the r^2 values are in the range of 0.934−0.937. These models were derived by using four interaction energy terms. The interaction energy terms included the interactions between SH2 domain and its binding phosphopeptide as well as the interactions between residues in SH2 domain. The best regression model is as follows

$$-\log(IC_{50}) = \underset{(\pm 0.116)}{-0.687}\,[Et_{44-57}] \underset{(\pm 0.001)}{-0.016}\,[Et_{pY1-57}] \underset{(\pm 0.037)}{+0.311}\,[Ec_{pY3-87}] \underset{(\pm 0.050)}{+0.292}\,[Ec_{pY3-93}] \underset{(\pm 0.126)}{+3.466} \quad (4)$$

$$N = 25;\ r^2 = 0.937;\ F = 74.502;\ s = 0.202$$

where Et_{44-57} is the total energy between Ser44 and Lys57 of SH2 domain, Et_{pY1-57} is the total energy between pY+1 of phosphopeptide and Lys57 of SH2 domain, Ec_{pY3-87} is the Coulombic energy between pY+3 of phosphopeptide and Tyr87 of SH2 domain, and Ec_{pY3-93} is the Coulombic energy between pY3 of phosphopeptide and Gly93 of SH2 domain. The interaction between residues in SH2 domain (Et_{44-57}) was selected as a good descriptor since the phosphopeptide induced the conformational change of its binding sites in SH2 domain and changed the interactions between binding site residues.

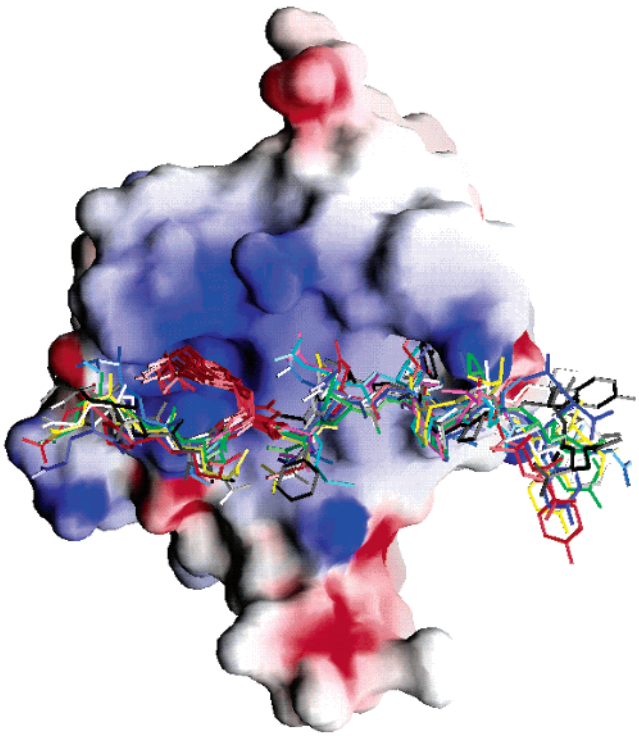

Figure 1. The SH2 domain and superimposed structures of binding phosphopeptides; the surface of SH2 domain was generated by the grasp program and the pY residue of each peptide was colored red.

Of 602 interaction energy terms, only 11 terms were selected in 10 high r^2 models (Table 3). The interaction pairs between SH2 domain and phosphopeptide were pY+1-Arg32, pY+1-Thr37, pY+1-Lys57, pY+3-Tyr87, and pY+3-Gly93. This fact shows that Arg32, Thr37, Lys57, Tyr87, and Gly93 may play an important role in the interactions with phosphopeptide. The interaction pair in SH2 domain was Ser44-Lys57. It is thought that the binding peptide induces the conformational change of binding sites and that the conformational change can be estimated by interaction energy. It is interesting that the interaction energies between pY (phosphotyrosine) and water molecules were selected as descriptors (Table 3), which suggests that water molecules play an important role in the binding of phosphopeptides. The residue−residue interaction energy was suggested as a good descriptor for the structure−activity relationships of ligand−macromolecule complexes.[17] In this study, the high r^2 regression models were also achieved by using interaction energy descriptors. As shown in the location of phosphopeptides within SH2 domain (Figure 1), the orientation of side chains of phosphopeptides is different each other, which is considered by the interaction energy terms in QSAR analysis.

Cross-Correlation Coefficients. The constituent atoms of protein domain may be expected to execute concerted, well-correlated motions. A molecular dynamics trajectory contains information about the dynamical motions of atoms.

Table 4. 10 High r^2 Models Derived by Using Residue−Residue Cross-Correlation Coefficients[a]

			variables and coefficients				
model	r^2	F value	A	B	C	D	intercept
1	0.925	61.885	Cbb_{35-60} +2.733	Css_{pY1-12} +3.013	Csb_{pY3-12} −2.510	Cbs_{pY2-58} −2.746	+3.361
2	0.908	49.316	Cbb_{35-60} +2.714	Css_{pY1-12} +2.886	Css_{pY3-35} −1.971	Cbs_{pY2-58} −2.720	+3.352
3	0.905	47.506	Caa_{57-94} +2.238	Cbb_{pY-92} +1.225	Css_{pY1-12} +5.138	Css_{pY1-71} −4.174	+2.142
4	0.904	47.013	Caa_{pY3-35} −3.276	Caa_{pY1-58} −3.422	Cbb_{35-59} +3.870	Css_{pY1-12} +3.741	+3.960
5	0.901	45.578	Caa_{57-94} +2.214	Css_{pY1-12} +5.324	Css_{pY1-71} −4.377	Cbs_{pY1-87} +1.357	+2.148
6	0.901	45.288	Cbb_{pY3-12} −2.088	Cbb_{35-36} −2.994	Cbb_{36-60} +3.815	Cbs_{pY3-32} −2.037	+5.858
7	0.900	45.099	Caa_{57-94} +2.217	Css_{pY1-12} +5.307	Css_{pY1-71} −4.717	Csb_{pY1-92} +1.450	+2.174
8	0.899	44.514	Caa_{57-94} +2.573	Css_{pY1-12} +5.557	Css_{pY1-71} −5.314	Css_{pY1-87} +1.851	+2.059
9	0.897	43.470	Cbb_{36-60} +2.663	Css_{pY1-12} +3.987	Csb_{pY1-12} −2.735	Cbs_{pY3-32} −2.370	+2.989
10	0.896	42.897	Caa_{57-94} +1.796	Cbb_{pY-93} +1.512	Css_{pY1-12} +5.209	Css_{pY1-71} −3.839	+2.133

[a] The subscripts a, b, and s mean all atom, backbone, and side chain, respectively. [$-\log(IC_{50}) = Coef_A \cdot A + Coef_B \cdot B + Coef_C \cdot C + Coef_D \cdot D$ + Intercept].

Table 5. 10 High r^2 Models Derived by Using Interaction Energies and Correlation Coefficients[a]

			variables and coefficients				
model	r^2	F value	A	B	C	D	intercept
1	0.961	124.031	Ec_{pY1-42} −1.120	Caa_{pY3-32} −2.211	Caa_{72-87} −1.751	Cbb_{36-60} +1.972	+4.176
2	0.961	122.173	Ec_{pY1-42} −1.182	Caa_{pY3-32} −2.041	Cbb_{36-60} +1.821	Cbb_{72-87} −1.566	+4.057
3	0.959	116.600	Et_{pY1-42} −1.162	Caa_{pY3-32} −2.048	Cbb_{36-60} +1.870	Cbb_{72-87} −1.576	+4.009
4	0.959	116.241	Et_{pY1-42} −1.099	Caa_{pY3-32} −2.218	Caa_{72-87} −1.755	Cbb_{36-60} +2.022	+4.129
5	0.958	113.125	Ec_{pY1-42} −1.221	Caa_{pY3-32} −1.801	Caa_{36-60} +1.559	Cbb_{72-87} −1.946	+4.060
6	0.957	110.292	Ec_{pY1-42} −1.245	Cbb_{pY3-32} −1.977	Cbb_{36-60} +1.704	Cbb_{72-87} −1.449	+3.883
7	0.954	103.648	Et_{pY1-42} −1.224	Cbb_{pY3-32} −1.981	Cbb_{36-60} +1.755	Cbb_{72-87} −1.458	+3.832
8	0.953	101.521	Et_{pY1-42} −1.200	Caa_{pY3-32} −1.797	Caa_{36-60} +1.589	Cbb_{72-87} −1.964	+4.015
9	0.948	90.723	Ec_{pY1-42} −1.186	Cbb_{pY3-32} −1.978	Cbb_{36-60} +1.922	Cbb_{71-87} −1.137	+3.761
10	0.948	90.229	Ec_{pY1-12} −0.021	Caa_{pY3-32} −1.938	Caa_{36-62} +1.459	Cbb_{35-42} +1.890	+3.027

[a] Et, Ec, and C are total energy, Coulombic energy, and correlation coefficient, respectively. The subscripts a, b, and s mean all atom, backbone, and side chain, respectively. [$-\log(IC_{50}) = Coef_A \cdot A + Coef_B \cdot B + Coef_C \cdot C + Coef_D \cdot D$ + Intercept].

Concerted atomic motions can be identified by analyzing the cross-correlation coefficients for atomic displacements[2,11,12] defined between two atoms i and j by the expression

$$C_{ij} = (\hat{r}_i \cdot \hat{r}_j)/(\hat{r}_i^2 \cdot \hat{r}_j^2)^{1/2} \quad (5)$$

where $\hat{r}_i$ is the displacement from the mean position of ith atom. Cross-correlation coefficients range from a value of −1 (completely anticorrelated motions) to a value of +1 (completely correlated motions). Deviations from 1 (or −1) imply that their motions are less correlated (or anticorrelated).

Table 4 shows the 10 high r^2 models derived by using four cross-correlation coefficient terms. The r^2 values are in the range of 0.896−0.925 which suggests that the concerted atomic motions can be good descriptors for structure−activity relationships. In this study the cross-correlation terms include cross correlations between SH2 domain and its binding phosphopeptide as well as between residues in SH2 domain. The best regression model is as follows

$$-\log(IC_{50}) = \underset{(\pm 0.323)}{2.733}\,[Cbb_{35-60}] + \underset{(\pm 0.324)}{3.013}\,[Css_{pY1-12}] - \underset{(\pm 0.339)}{2.510}\,[Csb_{pY3-12}] - \underset{(\pm 0.369)}{2.746}\,[Cbs_{pY2-58}] + \underset{(\pm 0.186)}{3.361} \quad (6)$$

$$N = 25;\ r^2 = 0.925;\ F = 61.885;\ s = 0.221$$

where Cbb_{35-60} is the backbone-backbone cross correlation between Glu35 and Ser60 of SH2 domain, Css_{pY1-12} is the side chain−side chain cross correlation between pY+1 of phosphopeptide and Arg12 of SH2 domain, Csb_{pY3-12} is the side chain-backbone cross correlation between pY+3 of phosphopeptide and Arg12 of SH2 domain, and Csb_{pY3-12} is the side chain−backbone cross correlation between pY+3 of phosphopeptide and Arg12 of SH2 domain. The cross correlation between residues, Glu35 and Ser60, in SH2 domain (Cbb_{35-60}) was selected as a good descriptor. Since they are located in pY binding pocket, it seems that the concerted atomic motions in pY binding pocket are likely to be mediated by the binding peptide and play an important role in the peptide binding.

Of 951 cross-correlation terms, only 20 terms were selected in 10 high r^2 models. The correlation pairs between SH2 domain and phosphopeptide were Arg12-pY+1, Arg12-pY+3, Arg32-pY+3, Glu35-pY+3, His58-pY+1, His58-pY+2, Ile71-pY+1, Tyr87-pY+1, Asp92-pY, Asp92-pY+1, and Gly93-pY. This result suggests that Arg12, Arg32, Glu35, His58, Ile71, Tyr87, Asp92, and Gly93 may play an important role in the peptide binding in terms of the concerted motions. It is noticeable that the cross correlations between the residues which are far in distance are found in the high r^2 models, for example, Arg12-pY+3, Asp92-pY, and Gly93-pY. It is unable to give an explanation about how the long-distance cross correlations have relation with the binding affinity. However, it is at least thought that these long-distance cross correlations play a certain role in the binding of phosphopeptide. The correlation pairs in SH2 domain were Glu35-Ser36, Glu35-Tyr59, Glu35-Lys60, Ser36-Lys60, and Lys57-Leu94. Most of them were the correlations within pY binding pocket except Lys57-Leu94. This fact implies that the concerted atomic motions in pY binding pocket are important in the peptide binding. The concerted motion between Lys57 (pY+1 binding pocket) and Leu94 (pY+3 binding pocket) is thought to be mediated by the binding peptide.

Interaction Energy and Cross-Correlation Coefficients. In this section, the interaction energies and cross-correlation coefficients (1553 descriptors) were used in the regression analysis. The highest r^2 value of 0.961 was achieved by using four descriptors (Table 5), which shows the regression was improved than the cases that the interaction energies and

cross-correlation coefficients were used separately. The 10 high r^2 models derived by using four descriptors are shown in Table 5 in which the r^2 values are in the range of 0.948−0.961. The best regression model is as follows where Ec_{pY1-42}

$$-\log(IC_{50}) = \underset{(\pm 0.089)}{-1.120}\,[Ec_{pY1-42}] \underset{(\pm 1.185)}{-2.211}\,[Caa_{pY3-32}] \underset{(\pm 0.241)}{-1.751}\,[Caa_{72-87}] \underset{(\pm 0.233)}{+1.972}\,[Cbb_{36-60}] \underset{(\pm 0.105)}{+4.176} \quad (7)$$

$$N = 25;\ r^2 = 0.961;\ F = 124.031;\ s = 0.159$$

is the Coulombic energy between pY+1 of phosphopeptide and Ser42 of SH2 domain, Caa_{pY3-32} is the all atom−all atom cross correlation between pY+3 of phosphopeptide and Arg32 of SH2 domain, Caa_{72-87} is the all atom−all atom cross correlation between Ser72 and Tyr87 of SH2 domain, and Cbb_{36-60} is the backbone−backbone cross correlation between Ser36 and Lys60 of SH2 domain. The interaction energies and cross correlations between SH2 domain and binding peptide as well as the cross correlations within peptide binding pockets of SH2 domain were selected in this model. It is suggested that both of the interaction energy and concerted atomic motions are important in this binding mode.

Of 1553 interaction energy and cross-correlation terms, only 13 terms were selected in 10 high r^2 models (Table 5). The interaction energy pairs were pY+1-Arg12 and pY+1-Ser42, and cross-correlation pairs were Glu35-Ser42, Ser36-Lys60, Ser36-Arg62, Ile71-Tyr87, Ser72-Tyr87, and pY+3-Arg32. It is interesting that the cross correlations selected in high r^2 models were mostly between the residues within peptide binding pockets of SH2 domain, while the interaction energies were between SH2 domain and the phosphopeptide. Compared with the models derived using only interaction energy terms, the interaction energies between the phosphopeptide and water molecules were not found in 10 high r^2 models. Since the concerted motions between residues within binding pockets can be mediated by the binding peptide, it is reasonable that the cross correlations between the pocket residues might be changed by the binding peptide, and these terms can be used in the explanation of the binding affinity. This study suggests that the concerted atomic motions between receptor and ligand or within the binding site of a receptor are able to affect the ligand binding, and so the cross correlations can be used in the prediction of binding affinity.

CONCLUSIONS

In this study, the residue−residue interaction energies and cross-correlation coefficients were used in multiple linear regression analysis. Since the number of descriptors was very large (602 for interaction energies and 951 for cross correlations), the stepwise addition method was applied for the multiple linear regressions. When a large number of descriptors are used in the stepwise addition method, it is expected that there are many combinations of descriptors which give high r^2 values. But 10 high r^2 models derived by using four descriptors were used in further discussion. The residue−residue interaction energy was suggested as a good descriptor for structure−activity relationships of ligand−macromolecule complexes.[16] The high r^2 regression models of which the r^2 values were in the range of 0.934−0.937 were also achieved by using four interaction energy descriptors in this work. In addition, the concerted atomic motions were used as descriptors for structure−activity relationships study here. The molecular dynamics simulations give a lot of information about various molecular motions which can be good descriptors. Of this information, only the concerted atomic motions were used in QSAR analysis. The concerted motions were described by cross-correlation coefficients for atomic displacement. The best regression model derived by using four descriptors of cross-correlation coefficients gave high r^2 value of 0.925. This suggests that the dynamic properties obtained by molecular simulations can be used as good descriptors in QSAR study.

ACKNOWLEDGMENT

This work was supported by National Research Laboratory program of Ministry of Science & Technology, Korea.

REFERENCES AND NOTES

(1) Amrein, K. E.; Sefton, B. M. Mutation of a site of tyrosine phosphorylation in the lymphocyte-specific tyrosine protein kinase, p56lck, reveals its oncogenic potential in fibroblasts. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 4247−4251.
(2) Brooks, C. L. III.; Karplus, M.; Pettitt, B. M. *Proteins: a theoretical perspective of dynamics, structure, and thermodynamics*; John Wiley: New York, 1988.
(3) Cantley, L. C.; Auger, K. R.; Carpenter, C.; Duckworth, B.; Granziani, A.; Kapeller, R.; Soltoff, S. Oncogenes and signal transduction. *Cell* **1991**, *64*, 281−302.
(4) Cartwright, C. A.; Eckhart, W.; Simon, S.; Kaplan, P. L. Cell transformation by pp60c-src mutated in the carboxy-terminal regulatory domain. *Cell* **1987**, *49*, 83−91.
(5) Cooper, J. A.; Gould, K. L.; Cartwright, C. A.; Hunter, T. Tyr527 is phosphorylated in pp60c-src: implications for regulation. *Science* **1986**, *231*, 1431−1434.
(6) Eck, M. J.; Shoelson, S. E.; Harrison, S. C. Recognition of a high-affinity phosphotyrosyl peptide by the Src homology-2 domain of p56lck. *Nature* **1993**, *362*, 87−91.
(7) Eck, M. J.; Atwell, S. K.; Shoelson, S. E.; Harrison, S. C. Structure of the regulatory domains of the src-family tyrosine kinase Lck. *Nature* **1994**, *368*, 764−769.
(8) 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.
(9) Koch, C. A.; Anderson, D.; Morgan, M. F.; Ellis, C.; Pawson, T. SH2 and SH3 domains: elements that control interactions of cytoplasmic signaling proteins. *Science* **1991**, *252*, 668−674.
(10) Myers, R. H. Classical and modern regression with applications. PWS/KENT, Boston, 1990.
(11) McCammon, J. A.; Gelin, B. R.; Karplus, M. Dynamics of folded proteins. *Nature* **1977**, *267*, 585−590.
(12) McCammon, J. A.; Harvey, S. C. *Dynamics of Proteins and Nucleic Acids*; Cambridge University Press: Cambridge, 1987.
(13) Makol, V.; Baumann, G.; Keller, T. H.; Manning, U.; Zurini, M. G. M. The crystal structures of the SH2 domain of p56lck complexed with two phosphopeptides suggest a gated peptide binding site. *J. Mol. Biol.* **1995**, *246*, 344−355.
(14) Narula, S. S.; Yuan, R. W.; Adams, S. E.; Green, O. M.; Green, J.; Philips, T. B.; Zydowsky, L. D.; Botfield, M. C.; Hatada, M.; Laird, E. R.; Zoller, M. J.; Karas, J. L.; Dalgarno, D. C. Solution structure of C-terminal SH2 domain of the human tyrosine kinase syk complexed with a phosphotyrosine pentapeptide. *Structure* **1995**, *3*, 1061−1073.
(15) Ortiz, A. R.; Pisabarro, T.; Gago, F.; Wade, R. C. Prediction of drug binding affinities by comparative binding energy analysis. *J. Med. Chem.* **1995**, *38*, 2681−2691.
(16) Pascal, S. M.; Singer, A. U.; Gish, G.; Yamazaki, T.; Shoelson, S. E.; Pawson, T.; Kay, L. E.; Forman-kay, J. D. Nuclear magnetic resonance structure of an SH2 domain of phospholipase C-gamma1 complexed with a high affinity binding peptide. *Cell* **1994**, *77*, 461−472.
(17) Payne, G.; Stolz, L. A.; Pei, D.; Band, H.; Shoelson, S. E.; Walsh, C. T. The phosphopeptide-binding specificity of Src family SH2 domains. *Chem. Biol.* **1991**, *1*, 99−105.

(18) Piccione, E.; Case, R. D.; Domchek, S. M.; Hu, P.; Chaudhuri, M.; Backer, J. M.; Schlessinger, J.; Shoelson, S. E. Phosphatidylinocitol 3-kinase p85 SH2 domain specificity defined by direct phosphopeptide/SH2 domain binding. *Biochemistry* **1993**, *32*, 3197−3292.
(19) Songyang, Z.; Shoelson, S. E.; Chaudhuri, M.; Gish, G.; Pawson, T.; Haser, W. G.; King, F.; Roberts, T.; Ratnofsky, S.; Lechleider, R. J.; Neel, B. G.; Birge, R. B.; Fajardo, J. E.; Chou, M. M.; Hanafusa, H.; Schaffhausen, B.; Cantley, L. C. SH2 domains recognize specific phosphopeptide sequences. *Cell* **1993**, *72*, 767−778.
(20) Tong, L.; Warren, T. C.; King, J.; Rose, R. B. J.; Jakes, S. Crystal structures of the human p56lck SH2 domain in complex with two short phosphotyrosyl peptides at 1.0 Å and 1.8 Å resolution. *J. Mol. Biol*. **1996**, *256*, 601−610.
(21) Ullrich, A.; Schlessinger, J. Signal transduction by receptors with tyrosine kinase activity. *Cell* **1990**, *61*, 203−212.
(22) Williams, L. T. Signal transduction by the platelet-derived growth factor receptor. *Science* **1989**, *243*, 1564−1570.
(23) Waksman, G.; Kominos, D.; Robertson, S. C.; Pant, N.; Baltimore, D.; Birge, R. B.; Cowburn, D.; Hanafusa, H.; Mayer, B. J.; Overduin, M.; Resh, M. D.; Rios, C. B.; Silverman, L.; Kuriyan, J. Crystal structure of the phosphotyrosine recognition domain SH2 of v-src complexed with tyrosine-phosphorylated peptides. *Nature* **1992**, *358*, 646−653.
(24) Waksman, G.; Shoelson, S. E.; Pant, N.; Cowburn, D.; Kuriyan, J. Binding of high affinity phosphotyrosyl peptide to the Src SH2 domain: crystal structures of the complexed and peptide-free forms. *Cell* **1993**, *72*, 779−790.

CI034073O