Structure-based statistical modeling and analysis of peptide affinity and cross-reactivity to human senile osteoporosis OSF SH3 domain

Wei Zhang | Biao Zhong | Yulin Zhan | Congfeng Luo

1 Department of Orthopaedics, Shanghai Sixth People’s Hospital East Affiliated to Shanghai University of Medicine & Health Sciences, Shanghai, China
2 Department of Orthopaedics, Shanghai Jiao Tong University Affiliated Sixth People’s Hospital, Shanghai, China

Correspondence
Biao Zhong, Department of Orthopaedics, Shanghai Jiao Tong University Affiliated Sixth People’s Hospital, Shanghai, China.
Email: biao.zhong@139.com
Yulin Zhan, Department of Orthopaedics, Shanghai Sixth People’s Hospital East Affiliated to Shanghai University of Medicine & Health Sciences, Shanghai, China.
Email: bondwe@163.com

Funding information
SUMHS Seed Foundation Project, Grant/Award Number: HMSF-16-21-014

Abstract
Human osteoclast-stimulating factor (OSF) induces osteoclast formation and bone resorption in senile osteoporosis by recruiting multiple signaling complexes with cognate interacting partners through its N-terminal Src homology 3 (SH3) peptide-recognition domain. The domain can recognize and bind to the polyproline regions of its partner proteins, rendering a broad ligand specificity and cross-reactivity. Here, the structural basis and physicochemical property of peptide affinity and cross-reactivity to OSF SH3 domain were investigated systematically by using an integration of statistical analysis and molecular modeling. A structure-based quantitative structure-activity relationship method called cross-nontbonded interaction characterization and statistical regression was used to characterize the intermolecular interactions involved in computationally modeled domain-peptide complex structures and then to correlate the interactions with affinity for a panel of collected SH3-binding peptide samples. Both the structural stability and generalization ability of obtained quantitative structure-activity relationship regression models were examined rigorously via internal cross-validation and external test, confirming that the models can properly describe even single-residue mutations at domain-peptide complex interface and give a reasonable extrapolation for the mutation effect on peptide affinity. Subsequently, the best model was used to investigate the promiscuity and cross-reactivity of OSF SH3 domain binding to its various peptide ligands. It is found that few key residues in peptide ligands are primarily responsible for the domain affinity and selectivity, while most other residues only play a minor role in domain-peptide binding affinity and stability. The peptide residues can be classified into 3 groups in terms of their contribution to ligand selectivity: key, assistant, and marginal residues. Considering that the key residues are very few so that many domain interacting partners share a similar binding profile, additional factors such as in vivo environments and biological contexts would also contribute to the specificity and cross-reactivity of OSF SH3 domain.
KEYWORDS
osteoclast-stimulating factor, peptide cross-reactivity, senile osteoporosis, SH3 domain, statistical modeling

1 | INTRODUCTION

Senile osteoporosis (SO) is a major health problem characterized by compromised bone strength predisposing patients to an increased risk of fracture, with significant economic consequences and adverse impacts on the quality of life. Hip fractures are the most devastating complication of SO, are likely to increase exponentially with an increasingly aged population, are associated with high recurrence rate, and lead to significant morbidity and mortality. Human osteoclast-stimulating factor (OSF) is one of the most promising targets for therapeutic intervention in SO, which is an intracellular protein produced by osteoclasts and, in a positive feedback manner, induces osteoclast formation and bone resorption in SO by recruiting multiple signaling complexes with its cognate interacting partners. The OSF contains an N-terminal modular peptide-recognition domain of Src homology 3 (SH3) that can recognize and bind to the polyproline regions of its partner proteins and then integrate the protein into osteoclast-specific signaling pathways.

Short polyproline peptide segments have been widely used as the OSF SH3 ligands, as they can mimic the interacting sites of the domain’s cognate partner proteins. However, these peptides show a broad specificity and high cross-reactivity when binding to the domain, largely limiting the application of designing therapeutic peptides to disrupt the peptide-mediated interactions of between human OSF and its partners in SO therapy. Recently, computational modeling and statistical analysis have been widely used to characterize the binding behavior and energetics of peptide ligands to their cognate domain receptors. In the present study, the structural basis and physicochemical property of OSF SH3 cross-reactivity were investigated systematically by using an integration of statistical analysis and molecular modeling. Structure-based quantitative structure-activity relationship (QSAR) methodology was used to characterize the intermolecular interactions involved in domain-peptide complex structures and to correlate the interactions with affinity for a panel of collected SH3-binding peptide samples. The built regression model was then used to investigate the affinity and cross-reactivity of OSF SH3 domain binding to its various peptide ligands. We also discussed the molecular mechanism and biological implication underlying peptide affinity and specificity for the domain.

2 | METHODS AND MATERIALS

2.1 | SH3-binding peptide data set

Landgraf et al have used phage display and SPOT synthesis to screen various peptides in the human proteome binding to a variety of SH3 domains. In the procedure, peptides matching the defined patterns were synthesized at high density on cellulose membranes by SPOT synthesis technology, and the membranes were probed with the corresponding SH3 domains fused to glutathione S-transferase. The intensity of each SPOT analysis experiment was assessed by a quantitative Boehringer light unit (BLU), which is partially correlated with domain-peptide dissociation constant and thus can be considered as a semiquantitative measure of peptide affinity to SH3 domains. Previously, a distinct panel of 2015 SH3-binding decapeptide ligands with measured BLU values was collected by Liu et al from the Landgraf peptide set. These peptides were partitioned into a training set consisting of 883 samples for developing regression models and a test set containing 1132 samples for validating the developed models, where the training peptides have at least 2 experimental results, while the test peptides were only assayed in a single time.

2.2 | Computational modeling of domain-peptide complex structures

Computational modeling of SH3 domain complex structures with its peptide ligands was performed using a protocol described previously. The NMR solution structure of amphiphysin SH3 in complex with chikungunya virus nsP3 peptide was retrieved from the protein data bank (PDB) database with accession ID 5I22. The nsP3 (STVPVAPPRRRRGRNLTR) is a 17-mer peptide, from which the 10-residue core binding sequence TVPVPAPPRR in complex with SH3 domain was extracted as template to model the domain complexes with other peptide ligands (Figure 1). The sequence is a class 2 peptide that contains the consensus motif PXXPX+ (where P is proline residue, + is positive charged residue, and X is any residue). In the procedure, nsP3 peptide in complex with SH3 domain was virtually mutated to, 1-by-1, the 2015 decapeptides compiled in our peptide set by using SIDEpro program, which were then subjected to a round of fast molecular
mechanics minimizations to remove bad atom contacts and overlapping. The minimized structure models of the 2015 domain-peptide complexes will be used in subsequent analysis.

2.3 | Cross-nonbonded interaction characterization and statistical regression

Previously, He and coworkers\(^{18}\) described a methodology called atomic cross-nonbonded interaction analysis to extract atomic-level nonbonded interaction information at domain-peptide complex interface. Here, the atomic cross-nonbonded interaction analysis method was extended to cover more detailed atom types and more diverse nonbonded interactions involved in domain-peptide binding. The extended version is termed as the cross-nonbonded interaction characterization and statistical regression (CICSR) that can apperceive the exquisite structural features of domain-peptide complex interface. In CICSR, the protein atom types were assigned with quantum chemical topology–based computation.\(^{19}\) Molecular electron densities of 20 amino acids and their smaller derivatives were partitioned into a set of 760 topological atoms; each atom was characterized by 7 atomic properties and subjected to cluster analysis, that is, C, H, O, N, and S. From the respective dendrograms, totally 42 protein atom types, including 21 carbons, 7 hydrogens, 2 nitrogens, 6 oxygens, and 6 sulfurs, were identified. The cross-interactions between the 42 atom types of domain and peptide in a complex can generate \(42^2 = 1764\) inter-atom type pairs at the complex interface. Then, 4 nonbonded interaction potentials that are primarily responsible for biomolecular recognition, including van der Waals, electrostatic, hydrophobic, and entropic, are assigned for each pair, thus resulting in \(1764 \times 4 = 7056\) structural descriptors to characterize the intermolecular interaction profile of a domain-peptide complex. These descriptors are used as independent variables and then correlated to dependent variable, ie, the domain-peptide binding affinity, by using the chemometric method of partial least squares regression (PLSR)\(^{20}\) on a panel of collected SH3-peptide binding samples. The PLSR algorithm consists of outer relations (independent variable \(x\) and dependent variable \(y\)) and an inner relation linking both variables:

\[
x_{ij} = \sum_{a=1}^{A} v_{ia} p_{aj} + e_{ij},
\]

\[
y_{ik} = \sum_{a=1}^{A} u_{ia} q_{ak} + g_{ik}.
\]

The latent variables \(v\) and \(u\) are correlated through the inner relation given below, which leads to the estimation of the \(y\) from the \(x\).

\[
\hat{u} = bt
\]

Instead of traditionally used physical potential functions such as Coulomb, Lennard-Jones, and empirical hydrophilic field, we herein used a quasipotential known as distance-dependent Gaussian function to calculate intermolecular nonbonded potentials. Previously, the distance-dependent Gaussian function has been demonstrated to work fairly well in the widely used QSAR methodology of comparative molecular similarity index analysis.\(^{21}\) Here, a numerical method proposed by Zhou et al\(^{22}\) was engaged to analyze the van der Waals, electrostatic, hydrophobic, and entropic interplay between domain and peptide:

\[
U(T) = \sum_{i=1}^{n} \sum_{j=1}^{m} \eta_{i} \cdot \eta_{j} \exp(-\alpha \cdot d_{ij}),
\]
where \( n \) and \( m \) are the numbers of atoms in domain and peptide. \( d_{ij} \) is the distance between atoms \( i \) and \( j \). \( \eta \) is an atomic attribute, which can be assigned with steric, polar, aquatic, and flexible parameters to characterize van der Waals, electrostatic, hydrophobic, and entropic effects in domain-peptide interaction, respectively; all parameters including charge, size, hydrophobicity, and flexibility for different protein atoms were taken from previous reports.\(^23,24\) \( \alpha \) is an attenuation factor, and, generally speaking, larger \( \alpha \) results in stronger attenuation of the distance-dependent function, thus focusing on local molecular structure, whereas smaller \( \alpha \) causes interatomic interaction more sensitive to distance change, thereby largely overlooking the global structural property. Here, \( \alpha \) was set to 6.908 so that the interatomic interaction can be attenuated to 1/1000 at a standard distance.

2.4 | Statistical validation

The coefficient of determination of fitting \((r_{\text{fit}}^2)\) in training set cannot reflect the predictive power of statistical regression models. Therefore, internal cross-validation and external test validation were used in this study to examine the stability and predictability of the CICSR-based QSAR models. Since cross-validation is thought as the most objective that can always yield a unique result for a given benchmark data set, and hence has been increasingly used and widely recognized by investigators to examine the accuracy of various predictors,\(^25\) the coefficient of determination \((q_{\text{cv}}^2)\) of tenfold cross-validation was used to test modeling stability and reliability for the training set. In addition, Golbraikh and Tropsha have argued that the high value of \( q_{\text{cv}}^2 \) appears to be the necessary but not the sufficient condition for the model to have a high predictive power and the external validation is the only way to establish a reliable model.\(^26\) Here, the predictive coefficient of determination \((r_{\text{prd}}^2)\) and the root mean square error of prediction \((r_{\text{rmsp}})\) on test set were used as a rigorous measure of predictive power for the built QSAR models. The \( r^2, q^2, r_{\text{prd}}^2 \), and \( r_{\text{rmsp}} \) can be expressed as follows:\(^27\)

\[
r_{\text{fit}}^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i^{\text{fit}})^2}{\sum_{i=1}^{n} (y_i - \bar{y}_t)^2},
\]

\[
q_{\text{cv}}^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i^{\text{cv}})^2}{\sum_{i=1}^{n} (y_i - \bar{y}_t)^2},
\]

\[
r_{\text{prd}}^2 = 1 - \frac{\sum_{i=1}^{m} (y_i - \hat{y}_i^{\text{prd}})^2}{\sum_{i=1}^{m} (y_i - \bar{y}_e)^2},
\]

\[
r_{\text{rmsp}} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i^{\text{prd}})^2},
\]

where \( n \) and \( m \) are the numbers of peptide samples in, respectively, training and test sets; \( y_i \) is the experimentally determined affinity of peptide \( i \); \( \bar{y}_t \) and \( \bar{y}_e \) are the average values of \( y_i \) over, respectively, training and test sets; and \( \hat{y}_i^{\text{fit}} \), \( \hat{y}_i^{\text{cv}} \), and \( \hat{y}_i^{\text{prd}} \) are the estimated affinities for peptide \( i \) by fitting, cross-validation, and external prediction, respectively.
3.1 CICSR-based statistical modeling of domain-peptide binding

Previously, a number of QSAR modeling studies have been performed on relating peptide sequence information with affinity, which cannot directly characterize the structure architecture and noncovalent interaction of domain-peptide complexes. Here, we used the complex structure-based CICSR method in conjunction with PLSR regression to statistically correlate the nonbonded effects with binding affinity for a panel of 2015 SH3-binding decapeptides. The complex structures of SH3 domain with these peptide ligands were 1-by-1 modeled computationally with virtual mutagenesis and energy minimizations based on the crystal template of amphiphysin SH3 in complex with chikungunya virus nsP3 peptide (PDB: 5I22), which then were subjected to CICSR analysis and characterization, resulting in thousands of nonbonded descriptors for each peptide sample. Here, we used PLSR to ascertain the linear relationship between the peptide structures (CICSR descriptors) and their binding affinities (log10BLU values). The linear PLSR method possesses a good computational efficiency and can be readily interpreted. First, the 4 groups of 1764 descriptors characterizing the electrostatic, van der Waals, hydrophobic, and entropic nonbonded interactions at domain-peptide complex interfaces were separately adopted to perform statistical modeling by using PLSR regression, resulting in 4 CICSR-based QSAR models, and their statistics are tabulated in Table 1. As can be seen, 3 to 5 significant latent variables were extracted from the crude nonbonded descriptors in PLSR modeling, which show a moderate linear correlation with peptide binding affinity, with coefficient of determination of fitting ($r^2_{fit}$) and cross-validation ($q^2_{cv}$) range between 0.44 and 0.58 and between 0.24 and 0.41, respectively. In addition, the 4 built regression models were then predicted on external test set, exhibiting a very modest predictability and generalization ability of these models, with coefficient of determination of prediction ($r^2_{prd}$) and root mean square error of prediction (rmsp) range between 0.27 and 0.38 and between 0.804 and 0.940, respectively. All these come together to suggest that none of a single nonbonded effect can dominate the whole binding process of domain-peptide complexes, and peptide affinity should be codetermined by all the 4 nonbonded factors.

In this respect, we rebuilt the statistical regression model using all the 4 nonbonded descriptors, ie, totally 7056 independent variables. As might be expected, the fitting ability and internal stability of the rebuilt CICSR-based QSAR model were improved considerably to $r^2_{fit} = 0.72$ and $q^2_{cv} = 0.53$, respectively. In particular, the external predictive power, which is measured by $r^2_{prd} = 0.55$ and rmsp = 0.672, was increased substantially, which stratifies the criterion $q^2_{cv} > 0.5$ and $r^2_{prd} > 0.5$ for a predictive QSAR model recommended by Tropsha et al. The model that contains 7 significant latent were extracted by PLSR, which can explain 72% variance for dependent variables by fitting, can predict 53% of the variance for dependent variables by cross-validation, and can cover 55% variance for test dependent variables (Table 1). As shown in Figure 2, the scatter plots of estimated versus experimental affinities for 883 SH3-binding peptides in training set and predicted versus experimental affinities for 1132 SH3-binding peptides in test set exhibit a good linear consistence between calculated and experimental affinities for both the training and test peptides, indicating that the CICSR-based QSAR model built using all the 7056 nonbonded descriptors in conjunction with PLSR regression possesses a good internal fitting ability and strong external predictive power.

| Nonbonded Descriptor | Regression Method | ND | NL | $r^2_{fit}$ | $q^2_{cv}$ | $r^2_{prd}$ | rmsp |
|----------------------|------------------|----|----|-------------|-----------|-------------|------|
| E                   | PLSR             | 1764 | 4  | 0.52        | 0.37      | 0.34        | 0.832|
| V                   | PLSR             | 1764 | 3  | 0.47        | 0.28      | 0.31        | 0.895|
| H                   | PLSR             | 1764 | 4  | 0.58        | 0.41      | 0.38        | 0.804|
| S                   | PLSR             | 1764 | 5  | 0.44        | 0.24      | 0.27        | 0.940|
| E + V + H + S       | PLSR             | 7056 | 7  | 0.72        | 0.53      | 0.55        | 0.672|
| E + V + H + S       | IC-PLSR          | 5829 | 6  | 0.65        | 0.58      | 0.59        | 0.641|

Abbreviations: CICSR, cross-nonbonded interaction characterization and statistical regression; E, electrostatic; H, hydrophobic; IC, intracorrelation; PLSR, partial least squares regression; NL, number of significant latent variables extracted from the independent variables; ND, number of independent variables used in regression; QSAR, quantitative structure-activity relationship; S, entropic; V, van der Waals.

Training set consists of 883 peptides, while test set contains 1132 peptides.
It is possible that not all the 7056 descriptors are really informative and some of them may be autocorrelated and noisy. To identify the most informative descriptors and to exclude those autocorrelated descriptors, we performed intracorrelation (IC) analysis for the 7056 descriptors. In the procedure, the Pearson correlation was calculated systematically across all the pairs among these descriptors and then removed those descriptors if the Pearson correlation coefficients of their pairs are larger than 0.8. Consequently, a total of 5829 descriptors with low IC were selected, which were then used to rebuild the CICSR-based QSAR model (termed as IC-PLSR regression). With the rebuilt model, the estimated/predicted versus experimental affinities for the 883 training peptides and 1132 test peptides as well as statistics are shown in Figure 3 and listed in Table 1. Evidently, most of peptide sample scatters are evenly distributed around the line fitted through these scatters, and no obvious bias and outlier can be observed in both the training and the test sets. In contrast to PLSR models built with all the 7056 variables, the internal stability ($q_{cv}^2$) and external predictive power ($r_{prd}^2$) of IC-PLSR model rebuilt with the 5829 selected variables are improved considerably, with $q_{cv}^2$ and $r_{prd}^2$ increase from 0.53 and 0.55 (PLSR model) to 0.58 and 0.59 (IC-PLSR model), respectively, although the fitting ability ($r_{fit}^2$) of IC-PLSR model shows a slight decrease from 0.72 to 0.65 as compared to PLSR model. Next, plot of the distance to IC-PLSR model in independent variable space was examined to analyze rebuilding ability of the model for each training sample. It is
revealed that the most peptides can well be rebuilt on the space by 6 significant latent variables as only few samples (< 5%) whose normalized distance is out of the critical value at 5% significance level, indicating that only little information loss in independent variables when performing the IC-PLSR modeling. In addition, the scoring scatters of training peptides distributed in the top 2 latent variable spaces of IC-PLSR model suggested that most sample points are within the Hotelling T2 ellipse at 5% significance level, with only few outliers.

### 3.2 Peptide cross-reactivity to OSF SH3 domain

The OSF can interact with an array of its cognate protein partners through its SH3 domain to orchestrate cellular signaling networks. Previously, Han et al identified 3 interacting partners of OSF SH3 domain, ie, proto-oncogene tyrosine-protein kinase.

| Peptide     | Source Protein | Binding Assay Method     | Predicted log10BLU |
|-------------|----------------|--------------------------|--------------------|
| 35SQTPSKPASA | c-Src          | Immunoblot analysis      | 3.2                |
| 482YRMPCPEC | c-Src          | Immunoblot analysis      | 2.7                |
| 485PCPECESL | c-Src          | Immunoblot analysis      | 3.5                |
| 192SFLPPPPPM | SMN            | Yeast 2-hybrid screening | 2.9                |
| 193FLPPPPMPF | SMN            | Yeast 2-hybrid screening | 3.8                |
| 195PPPMPGPR | SMN            | Yeast 2-hybrid screening | 3.4                |
| 204RLFGKPKL | SMN            | Yeast 2-hybrid screening | 2.3                |
| 214FNGPPPPP | SMN            | Yeast 2-hybrid screening | 1.8                |
| 215NGPPPPPP | SMN            | Yeast 2-hybrid screening | 3.9                |
| 216GPPPPPPP | SMN            | Yeast 2-hybrid screening | 3.4                |
| 217PPPVPPPP | SMN            | Yeast 2-hybrid screening | 4.1                |
| 218PPPPPPP | SMN            | Yeast 2-hybrid screening | 3.2                |
| 219PPPPPPP | SMN            | Yeast 2-hybrid screening | 2.7                |
| 220PPPPPPH | SMN            | Yeast 2-hybrid screening | 3.5                |
| 231CWLPPFPSG | SMN           | Yeast 2-hybrid screening | 3.0                |
| 234PPPSGPPI | SMN            | Yeast 2-hybrid screening | 3.8                |
| 238SGPPPPPP | SMN            | Yeast 2-hybrid screening | 2.6                |
| 241PPPPPPC | SMN            | Yeast 2-hybrid screening | 3.8                |
| 244PPPPPP | SMN            | Yeast 2-hybrid screening | 2.9                |
| 245PPPPICPD | SMN            | Yeast 2-hybrid screening | 1.7                |
| 35SQQPLLPHS | Sam68          | Phage-ELISA              | 2.6                |
| 62QPPLLPPSA | Sam68          | Phage-ELISA              | 3.2                |
| 82APTPLLPSA | Sam68          | Phage-ELISA              | 4.0                |
| 292GAAPPPPPV | Sam68          | Phage-ELISA              | 3.7                |
| 293AAPPPPPV | Sam68          | Phage-ELISA              | 3.4                |
| 295PPPPVRGR | Sam68          | Phage-ELISA              | 3.1                |
| 331RGVPPPTVR | Sam68          | Phage-ELISA              | 2.7                |
| 353QRIPLPPPA | Sam68          | Phage-ELISA              | 3.6                |
| 355IPPPPAPE | Sam68          | Phage-ELISA              | 2.2                |
| 357LPPPAPETY | Sam68          | Phage-ELISA              | 2.6                |
| 424KAPPARVKG | Sam68          | Phage-ELISA              | 3.9                |

Abbreviations: BLU, Boehringer light unit; c-Src, proto-oncogene tyrosine-protein kinase; ELISA, enzyme-linked immunosorbent assay; OSF, osteoclast-stimulating factor; Sam68, Src-associated in mitosis 68 kD; SH3, Src homology 3; SMN, survival motor neuron.
tyrosine-protein kinase (c-Src); survival motor neuron (SMN); and Src-associated in mitosis, 68 kD (Sam68), from which totally 31 octapeptide segments that contain linear polyproline motif PXXP recognized by SH3 domain were determined. Considering that the CICSR-based QSAR model built in this study was based on SH3-binding decapetide ligands, we herein extended these 8- to 10-mer peptides by adding a residue separately at N- and C-termini of each peptide. The primary sequence of full-length proteins c-Src, SMN, and Sam68 were retrieved from the UniProt database with accession IDs P12931, Q16637, and Q07666, respectively. Consequently, the corresponding 31 decapetides were obtained (Table 2), and their complex structures with OSF SH3 domain were modeled computationally using peptide docking and virtual mutagenesis based on the domain crystal structure (PDB: 1ZLM), which were then used to predict their binding affinity with the IC-PLSR regression model. The predicted log10BLU values are listed in Table 2 and shown as a histogram in Figure 4. As can be seen, most of these peptides exhibit a moderate affinity towards the domain, with log10BLU range between 1.8 and 4.1, with average value and standard error of 3.2 and 0.6, respectively. It is suggested that the OSF SH3 domain exhibits a broad specificity and strong cross-reactivity for its potential peptide ligands, and no predominant selectivity of the domain can be found for specific peptide binders. In addition, it is evident in Figure 4 that the domain appears to have no bias for the 3 cognate protein partners c-Src, SMN, and Sam68, with a basically consistent affinity for peptides from different source proteins, although there are very few outliers that were inferred as good (log10BLU > 4) or bad (log10BLU < 2) domain binders. In fact, it is known that many peptide-recognition domains, including SH3, PDZ, and WW, have a high promiscuity and cross-reactivity for a variety of their potential peptide ligands. Thus, these domains have been recognized as versatile protein modules participating in molecular regulation of diverse biological events. Previously, by investigating the cross-reactivity of Src SH3 domain, He et al found that the molecular interaction is not the only term to select right binding partners for specific SH3 domains, and other factors such as cellular context and the rest of the SH3-containing proteins may play important roles in reducing their ligand cross-reactivity. This can readily explain the paradox that many peptide-recognition domains have broad specificity at molecular level but high specificity at cellular level.

Next, the Sam68 peptide 424KAPPARPVKG433, which was predicted theoretically as a strong binder of OSF SH3 domain (log10BLU = 3.9) and has been determined experimentally to have a high affinity (Kd = 3.2 μM) to the domain previously, was selected to analyze the effect of systematic single mutations at each residue of the peptide. In the procedure, the native amino acids at each peptide residue were systematically mutated to other 19 types by using SIDEpro program, and the peptide affinity changes upon the mutations were predicted using the CICSR-based QSAR model (IC-PLSR). Consequently, a systematic single-point mutation profile was defined for the peptide (Figure 5), where the gray, green, blue, and red boxes represent no mutation, neutral mutation, unfavorable mutation, and favorable mutation in terms of the affinity changes, respectively. Evidently, most mutations would not influence the peptide binding affinity substantially, with −0.5 < log10BLU change < 0.5, suggesting a high cross-reactivity between the native peptide and its most mutants for the domain. However, there are also few residue mutations that may cause considerable effect on peptide binding, with log10BLU changes >1 or <−1. According to the analysis, peptide residues can be classified into
FIGURE 5  The predicted affinity changes of Src-associated in mitosis 68 kD peptide 424KAPPARPVKG433 to osteoclast-stimulating factor Src homology 3 domain upon systematic single mutations at the 10 peptide residues. The mutations are colored in gray (no mutation), green (neutral mutation), blue (unfavorable mutation), and red (favorable mutation) in terms of the affinity changes.

3 groups, ie, key residues (P427, P430, and K432); assistant residues (A425, P426, A428, R429, and V431); and marginal residues (K424 and G433). The 3 key residues correspond to the core binding motif PXXPX± shared by class 2 peptides, which are primarily responsible for the domain-peptide binding affinity and specificity. Thus, a number of mutations, especially those charged and polar mutations, on the 3 key residues would largely impair peptide binding potency. In addition, there are also some assistant residues that play a relatively important role in the domain-peptide binding; mutation of these residues can cause favorable or unfavorable effect to the binding. Moreover, the 2 marginal residues are separately located at the N- and C-termini of the peptide. Visual analysis of domain-peptide complex structure revealed that the 2 residues are out of peptide-binding pocket and far away from the domain, thus contributing very limitedly to the binding.

ACKNOWLEDGMENT

This work was supported by the SUMHS Seed Foundation Project (HMSF-16-21-014).

ORCID

Biao Zhong http://orcid.org/0000-0002-7419-0893

REFERENCES

1. Atik OS, Uslu MM, Eksioglu F, Satana T. Etiology of senile osteoporosis: a hypothesis. Clin Orthop Relat Res. 2006;443:25-27.
2. Duque G, Demontiero O, Troen BR. Prevention and treatment of senile osteoporosis and hip fractures. Minerva Med. 2009;100:79-94.
3. Reddy S, Devlin R, Menaa C, et al. Isolation and characterization of a cDNA clone encoding a novel peptide (OSF) that enhances osteoclast formation and bone resorption. J Cell Physiol. 1998;177:636-645.
4. Zhou Y, Liu J, Zhou Q, Liu C, Dong Q. The peptide-mediated interactions between human osteoclast-stimulating factor and its partner proteins in osteoporosis: which binds to which? Int J Pept Res Ther. 2017;23:61-67.
5. Han S, Liu Q, Wang F, Yuan Z. Targeting the SH3 domain of human osteoclast-stimulating factor with rationally designed peptoid inhibitors. J Pept Sci. 2016;22:533-539.
6. Yang C, Zhang S, He P, Wang C, Huang J, Zhou P. Self-binding peptides: folding or binding? J Chem Inf Model. 2015;55:329-342.
7. Yang C, Zhang S, Bai Z, et al. A two-step binding mechanism for the self-binding peptide recognition of target domains. Mol Biosyst. 2016;12:1201-1213.
8. Yao ZS, Han SP, Jin RZ. ROCK-I derived bi-functional peptides with antibacterial infection activity. Int J Pept Res Ther. 2016;22:407-412.
9. Xu Y, Huang R, Gu J, Jiang W. Derivation of inhibitory peptides to target the cardiac troponin C–I interaction as potential therapeutics for heart failure. Int J Pept Res Ther. 2017;23:387-392.
10. Bai Z, Hou S, Zhang S, Li Z, Zhou P. Targeting self-binding peptides as a novel strategy to regulate protein activity and function: a case study on the proto-oncogene tyrosine protein kinase c-Src. J Chem Inf Model. 2017;57:835-845.
11. Landgraf C, Panni S, Montecchi-Palazzi L, et al. Protein interaction networks by proteome peptide scanning. PLoS Biol. 2004;2:e14
12. Liu L, He D, Yang S, Xu Y. Applying chemometrics approaches to model and predict the binding affinities between the human amphiphysin SH3 domain and its peptide ligands. Protein Pept Lett. 2010;17:246-253.
13. Wu G, Zhang ZL, Fu CJ, Lv FL, Tian FF. Proteome-wide inference of human endophilin 1-binding peptides. Protein Pept Lett. 2012;19:1094-1102.
14. Han S, Liu Q, Dong J, Yuan Z. Chemometrics-based identification, characterization, and modeling of the intermolecular interaction between human osteosarcoma SH3 domain and its peptide ligands. *J Chemometr.* 2017;31: e2892.

15. Berman HM, Westbrook J, Feng Z, et al. The protein data bank. *Nucleic Acids Res.* 2000;28:235-242.

16. Sakseka K, Permi P. SH3 domain ligand binding: what’s the consensus and where’s the specificity? *FEBS Lett.* 2012;586:2609-2614.

17. Nagata K, Randall A, Baldi P. SIDEpro: a novel machine learning approach for the fast and accurate prediction of side-chain conformations. *Proteins.* 2012;80:142-153.

18. He P, Wu W, Wang HD, et al. Why ligand cross-reactivity is high within peptide recognition domain families? A case study on human c-Src SH3 domain. *J Theor Biol.* 2014;340:30-37.

19. Popelier PLA, Aicken FM. Atomic properties of amino acids: computed atom types as a guide for future force-field design. *ChemPhysChem.* 2003;4:824-829.

20. Wold S, Sjostrom M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemometr Intel Lab Syst.* 2001;58:109-130.

21. Klebe G, Abraham U, Mietzner T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J Med Chem.* 1994;37:4130-4146.

22. Zhou P, Tian F, Li Z. Quantitative structure-property relationship studies for collision cross sections of 579 singly protonated peptides based on a novel descriptor as molecular graph fingerprint (MoGF). *Anal Chim Acta.* 2007;597:214-222.

23. Ruan Z, Wang H, Ren Y, et al. Pseudo receptor probes: a novel pseudo receptor-based QSAR method and application into studies on a new kind of selective vascular endothelial growth factor-2 receptor inhibitors. *Chemometr Intel Lab Syst.* 2008;92:157-168.

24. He P, Wu W, Yang K, et al. Exploring the activity space of peptides binding to diverse SH3 domains using principal property descriptors derived from amino acid rotamers. *Biopolymers.* 2011;96:288-301.

25. Konovalov DA, Llewellyn LE, Vander Heyden Y, Coomans D. Robust cross-validation of linear regression QSAR models. *J Chem Inf Model.* 2008;48:2081-2094.

26. Golbraikh A, Tropsha A. Beware of $r^2$! *J Mol Graph Model.* 2002;20:269-276.

27. Ren Y, Chen X, Feng M, Wang Q, Zhou P. Gaussian process: a promising approach for the modeling and prediction of peptide binding affinity to MHC proteins. *Protein Pept Lett.* 2011;18:670-678.

28. Zhou P, Tian F, Wu Y, Li Z, Shang Z. Quantitative sequence-activity model (QSAM): applying QSAR strategy to model and predict bioactivity and function of peptides, proteins and nucleic acids. *Curr Comput Aided Drug Des.* 2008;4:311-321.

29. Zhou P, Chen X, Wu Y, Shang Z. Gaussian process: an alternative approach for QSAM modeling of peptides. *Amino Acids.* 2010;38:199-212.

30. Lin H, Yan T, Wang L, Guo F, Ning G, Xiong M. Statistical design, structural analysis, and in vitro susceptibility assay of antimicrobial peptoids to combat bacterial infections. *J Chemometr.* 2016;30:369-376.

31. Tropsha A, Gramatica P, Gombar VK. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb Sci.* 2003;22:69-77.

32. UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 2017;45:D158-D169.

33. Zhou P, Hou S, Bai Z, et al. Disrupting the intramolecular interaction between proto-oncogene c-Src SH3 domain and its self-binding peptide PPII with rationally designed peptide ligands. *Artif Cells Nanomed Biotechnol.* 2017; https://doi.org/10.1080/21691401.2017.1360327. [Epub ahead of print]

34. Li SS. Specificity and versatility of SH3 and other proline-recognition domains: structural basis and implications for cellular signal transduction. *Biochem J.* 2005;390:641-653.

**How to cite this article:** Zhang W, Zhong B, Zhan Y, Luo C. Structure-based statistical modeling and analysis of peptide affinity and cross-reactivity to human senile osteoporosis OSF SH3 domain. *Journal of Chemometrics.* 2018;32:e2967. [https://doi.org/10.1002/cem.2967]