Characterization of Domain–Peptide Interaction Interface: A Case Study on the Amphiphysin-1 SH3 Domain

Tingjun Hou1, Wei Zhang2, David A. Case2 and Wei Wang1* 

1Department of Chemistry and Biochemistry, University of California at San Diego, La Jolla, CA 92093, USA
2Department of Molecular Biology, The Scripps Research Institute, La Jolla, CA 92037, USA

Received 17 October 2007; received in revised form 14 December 2007; accepted 20 December 2007
Available online 3 January 2008

Many important protein–protein interactions are mediated by peptide recognition modular domains, such as the Src homology 3 (SH3), SH2, PDZ, and WW domains. Characterizing the interaction interface of domain–peptide complexes and predicting binding specificity for modular domains are critical for deciphering protein–protein interaction networks. Here, we propose the use of an energetic decomposition analysis to characterize domain–peptide interactions and the molecular interaction energy components (MIECs), including van der Waals, electrostatic, and desolvation energy between residue pairs on the binding interface. We show a proof-of-concept study on the amphiphysin-1 SH3 domain interacting with its peptide ligands. The structures of the human amphiphysin-1 SH3 domain complexed with 884 peptides were first modeled using virtual mutagenesis and optimized by molecular mechanics (MM) minimization. Next, the MIECs between domain and peptide residues were computed using the MM/generalized Born decomposition analysis. We conducted two types of statistical analyses on the MIECs to demonstrate their usefulness for predicting binding affinities of peptides and for classifying peptides into binder and non-binder categories. First, combining partial least squares analysis and genetic algorithm, we fitted linear regression models between the MIECs and the peptide binding affinities on the training data set. These models were then used to predict binding affinities for peptides in the test data set; the predicted values have a correlation coefficient of 0.81 and an unsigned mean error of 0.39 compared with the experimentally measured ones. The partial least squares–genetic algorithm analysis on the MIECs revealed the critical interactions for the binding specificity of the amphiphysin-1 SH3 domain. Next, a support vector machine (SVM) was employed to build classification models based on the MIECs of peptides in the training set. A rigorous training-validation procedure was used to assess the performances of different kernel functions in SVM and different combinations of the MIECs. The best SVM classifier gave satisfactory predictions for the test set, indicated by average prediction accuracy rates of 78% and 91% for the binding and non-binding peptides, respectively. We also showed that the performance of our approach on both binding affinity prediction and binder/non-binder classification was superior to the performances of the conventional MM/Poisson–Boltzmann solvent-accessible surface area and MM/generalized Born solvent-accessible surface area calculations. Our study demonstrates that the analysis of the MIECs

*Corresponding author. E-mail address: wei-wang@ucsd.edu.

Abbreviations used: SH3, Src homology 3; MIEC, molecular interaction energy component; MM, molecular mechanics; SVM, support vector machine; PSSM, position-specific scoring matrix; MD, molecular dynamics; PB, Poisson–Boltzmann; PBSA, Poisson–Boltzmann solvent-accessible surface area; GB, generalized Born; hAmph1, human amphiphysin-1; PLS, partial least squares; GBSA, generalized Born solvent-accessible surface area; GA, genetic algorithm; G/PBSA, genetic algorithm-based partial least squares; LOO, leave one out; UME, unsigned mean error; BLU, Boehringer light unit; RBF, radial basis function.
between peptides and the SH3 domain can successfully characterize the binding interface, and it provides a framework to derive integrated prediction models for different domain–peptide systems.

© 2008 Elsevier Ltd. All rights reserved.

Keywords: energetic component analysis; molecular interaction energy component, MIEC; SH3 modular domain; support vector machine, SVM; partial least squares, PLS

Introduction

Protein–protein interactions play essential roles in regulating many key biological processes. Pioneering work on the Src homology 3 (SH3) domain by several groups showed that protein–protein interactions are often mediated by modular domains that bind to short peptides with specific sequence motifs.1-3 Thereinafter, many other peptide recognition domains have also been characterized for their roles in signal transduction by mediating weak and transient protein–protein interactions.4-5 Understanding the binding specificity and identifying interaction partners of modular domains are thus critical in reconstructing protein interaction networks.

High-throughput technologies to identify protein–protein interactions, such as yeast two-hybrid and complex purification followed by mass spectrometry, often miss these domain-mediated interactions because they are weak and transient.6 Such experimental techniques as peptide library and SPOT synthesis have been developed to successfully determine the peptide motifs recognized by modular domains.7 The accuracy of the motifs is still limited by the coverage of all possible peptides, and the prediction power of the consensus motif for interacting partners of individual domains is not satisfactory. Tong et al. combined yeast two-hybrid and phage display peptide motif analyses to successfully predict interacting partners for SH3 domains in the budding yeast genome.8 Stiftler et al. trained a variation of a position-specific scoring matrix (PSSM) on interaction pairs determined by protein array to discriminate the binding specificity of mouse PDZ domains.9 The success of these two studies showed the strength of integrating information from diverse sources.

Computational methods have also been developed to predict domain–peptide interactions.10-20 For example, the SH3-SPOT method builds a position-specific contact frequency matrix based on protein–peptide contacts in a number of crystal structures of SH3/peptide and SH3/protein complexes.11 The matrix was then used to calculate the probability of a peptide binding to a specific SH3 domain. Recently, machine learning algorithms, such as artificial neural network and support vector machine (SVM), were introduced to predict the SH3 domain binding peptides based on contact information.12,13 Training these classifiers usually requires data for numerous SH3 domains because the number of possible combinations of contacts is huge. On the other hand, these methods are computationally efficient and can be used for quick proteome screening. In these approaches, a limited number of available crystal complex structures of domain–peptide pairs are used to define residue contact pairs and such contact information is transferred directly to those domain–peptide pairs with no solved complex structure based on sequence alignment. Obviously, structural information encoded in the contact matrix is crude because three-dimensional conformational flexibility is not considered for each individual domain–peptide pair, and the physiochemical properties of contact pairs are only roughly considered by dividing the 20 amino acids into several groups. The contact matrix is always sparse because of the large number of combinations of amino acids (20 × 20 = 400) or amino acid groups for each residue contact pair and the relatively small number of known domain–peptide interaction pairs. The sparse matrix introduces noise to the predictors and makes it difficult to train the models.

Alternative computational methods have also been developed to incorporate structural information in a more sophisticated way and to consider domain–peptide interaction based on physical chemistry. Wollacott and Desjarlais reported a method based on peptide design strategies to derive a PSSM for peptides binding to a modular domain, and this PSSM was then applied to scan protein sequence databases.14 Hou et al. took a virtual mutagenesis approach to mimic the ideal sampling coverage of all possible peptides in the peptide library approach.15 The complex structures were optimized by molecular dynamics (MD) simulation, and the PSSM for binding peptides of the Abl SH3 domain was calculated using the molecular mechanics/Poisson–Boltzmann solvent-accessible surface area (MM/PBSA) method. This PSSM was then used to screen the sequence database, and most known binders of the Abl SH3 domain could be identified. McLaughlin et al. proposed an approach to investigate the binding motifs of the Grb2 and SAP SH2 domains by integrating binding free energy estimation and peptide sequence analysis.16 The binding motifs of the SAP and Grb2 SH2 domains, derived from the binding class obtained by clustering peptides using the sequence and the calculated binding free energy, agreed well with those determined through experimental studies. These structure-based approaches do not need a large amount of binding affinity data to train the model, but the
quality of the modeled structures and the accuracy of the free energy calculations are critical for the success of these methods.

To truly incorporate the structural information of each peptide in the prediction model and to reduce the sensitivity of the model to the accuracy of the free energy calculations, we herein propose a new approach to characterize the binding interface between a modular domain and its binding peptides based on an energetic component analysis. First, each domain–peptide complex was modeled from a template structure by side chain mutation; this modeled structure was optimized using MM minimization. Second, the molecular interaction energy components (MIECs), including van der Waals, electrostatic, and desolvation free energies, for all residue pairs, including both inter- (domain–peptide) and intramolecular (peptide residues) pairs, between domains and peptides were computed using the MM/generalized Born (MM/GB) decomposition analysis. The MIECs were then encoded into a matrix. This MIEC matrix represents the energetic characteristics of the binding interface and can be used to either predict binding affinities for peptides in the proteome or classify proteomic peptides into the binder and non-binder categories.

There are several advantages of using MIECs to study domain binding specificity. First, the complex structure between each individual peptide and the domain is modeled and optimized, in which conformational flexibility is considered. Second, since MIECs describe the local interaction interface, they are less sensitive to the inaccuracy of structure modeling and free energy calculation compared with the approaches that rank peptides purely based on binding free energy. Third, unlike the sparse contact matrix, the MIEC matrix is fully filled because the interactions between residue pairs are represented by the energy terms regardless of amino acid type. In training classifiers, such a fully filled MIEC matrix is much more informative and more robust to noise or error than a sparse contact matrix.

We performed a proof-of-concept study on the human amphiphysin-1 (hAmph1) SH3 domain and 884 peptide ligands. Amphiphysins, types 1 and 2, are involved in clathrin-mediated endocytosis, actin function, and signaling pathways. Amphiphysin-1 has an SH3 domain at its C-terminus that interacts, with high affinity and specificity, with a class II SH3 domain binding site, PSKPNR, in the prolinerich C-terminus of dynamin. In the current study, we calculated MIECs, and, based on these, we predicted (1) peptide binding affinities using linear regression [partial least squares (PLS)] models and (2) binding and non-binding peptides for the SH3 domain using an SVM. Compared with the conventional MM/PBSA and MM/generalized Born solvent-accessible surface area (GBSA) calculations, our method showed higher prediction accuracies on both binding affinity prediction and binder/non-binder classification, which suggested that MIECs can successfully capture the binding characteristics of the SH3/peptide system.

Results and Discussion

MIECs

In the current work, the SH3/peptide binding interface was characterized by the MIECs between the hAmph1 SH3 domain and the binding peptides. First, the important residues on the SH3 binding surface that are close to the binding peptides and may contribute to binding specificity were identified (Fig. 1 and Table 1). For example, for Arg5 in the peptide PLPRRPPRAA, the important SH3 residues are Phe8, Glu9, Ala10, Ala11, Asn12, Glu15, Trp39, Pro62, Asn64, and Phe65 (Fig. 1 and Table 1). Then, the interactions between the peptide residue and the important residues in SH3 were calculated using free energy decomposition analysis (see Materials and Methods). In total, there were 57 residue–

![Fig. 1. The Connolly surface for the important residues of the hAmph1 SH3 domain for peptide binding. The binding peptide PLPRRPPRAA is shown in stick form; the residue Arg of the peptide at P5 is shown in yellow; the surface for the important residues of SH3, which may form effective interactions with Arg5 of the peptide, is shown in blue, and the surface for the other residues of SH3 is shown in red. The figure was generated using InsightII.](image-url)
residue interaction pairs to characterize the interaction between the hAmph1 SH3 domain and the peptides. In addition, 9 pairs between adjacent peptide residues were also included to characterize the structural features of the peptide. Thus, the MIECs for all SH3/peptide complexes are described by a $66 \times m$ matrix, a $132 \times m$ matrix, and a $198 \times m$ matrix ($m$ represents the number of peptides) when one type, two types, and three types of MIECs, respectively, were used (see Fig. 2). Compared with other approaches,\textsuperscript{10–20} the MIEC matrix characterizes the interaction interface of SH3/peptide in a universal way in the sense that it is derived based on free energy calculation and does not depend on amino acid type. Favorable free energy is the determining factor for whether an amino acid or an amino acid type can fit to the binding interface.

Since the free energy decomposition can still be noisy and some interaction pairs may not be informative, we applied two statistical techniques to analyze the MIEC matrix and demonstrated that the MIECs successfully capture the local environment of the interaction interface. Similar concepts have been implemented in CoMFA\textsuperscript{26} and its variant COMBINE,\textsuperscript{27} which build prediction models by analyzing the molecular fields around ligands or the molecular interaction fields between ligands and proteins. We further generalized this approach to study the domain–peptide interactions. We first built linear regression models using PLS and genetic

---

**Fig. 2.** Scheme of the procedure to build the prediction models based on MIECs: (1) model the SH3/peptide complexes based on virtual mutagenesis and GB-based MM minimizations; (2) determine the important SH3 domain residues that form effective interactions with the peptides (one residue in peptide is shown as a pink ball, and the SH3 residues forming contacts with the peptide residue are shown as green balls); (3) generate the SH3/peptide MIECs using the MM/GB free energy decomposition analysis (the columns of the table represent the 66 residue-residue interaction pairs; columns 1–57 represent the SH3/peptide interaction pair, and columns 58–66 represent the 9 pairs between adjacent peptide residues; the four important SH3 residues that may form effective interaction with the P1 peptide residue are shown in red as an example; and (4) apply statistical methods to analyze the MIEC matrix and construct the correlation or classification models.
algorithm (GA)-based PLS (G/PLS) based on the MIECs. We showed that these models could predict peptide binding affinities better than the conventional MM/PBSA and MM/GBSA. Since in many biological problems it is sufficient to just separate binding peptides from non-binders, we next trained a set of SVM classifiers to classify peptides into binder and non-binder categories. Cross-validation showed that the SVM classifiers based on the MIECs achieved high specificity and sensitivity.

**Prediction of binding affinities by analyzing the MIECs using linear regression**

We first examined whether binding affinity for a given peptide can be predicted based on its MIECs. For this purpose, we built linear regression models using PLS between the MIECs and the binding affinities of peptides measured by Landgraf et al. 28

**PLS regression models**

We investigated the prediction performances of different combinations of three types of MIECs (i.e., van der Waals, electrostatic, and polar contribution to desolvation energy). We built three models using van der Waals and electrostatic MIECs (model 1), van der Waals and polar MIECs (model 2, which is the sum of the electrostatic and the polar desolvation energy), and all three types of MIECs (model 3). For each combination, 10 PLS models were derived from using the first principal component to using all the top 10 principal components. The dependence of the q values [the regression coefficient based on leave-one-out (LOO) cross-validation] on the number of principal components was used to guide the selection of the optimal number of principal components. For the q values of the 10 PLS models, a saddle-like curve was observed. The saddle point corresponds to the optimal numbers of principal components, which are 4, 4, and 6 for models 1, 2, and 3, respectively (Table 2). The statistical significance of these three models can be verified by the high correlation coefficients of LOO cross-validation for the training set (Table 2).

It should be noted that the interaction energies between the adjacent residues in peptides were also included to consider structural characteristics of the binding peptides. Our calculations showed that inclusion of these interaction terms indeed improved the prediction accuracy. For example, if the interaction fields of the adjacent peptide residues were not considered, the statistical performance of the best PLS model, model 2 in Table 2, would become a little worse (N_Pc=5, q=0.776).

**Selecting the most informative MIECs**

The numbers of the informative MIEC terms in models 1, 2, and 3 used by PLS regression are 120, 109, and 183 (after the removal of non-informative terms using a cutoff), respectively. These terms are a significant portion of the total 198 MIECs. It is possible that not all these terms are really informative and that some of them may even just introduce noise. Therefore, in order to identify the most crucial MIEC terms, we used a genetic algorithm (GA) to search for the most informative MIEC combinations (see Materials and Methods). The G/PLS correlation models were built by using one to five principal components. For each G/PLS calculation, 100 best models were saved, and the one with the highest cross-validation coefficient was selected for further analysis. Considering that the solutions given by GA may be trapped into local minima, three independent runs were performed for each principal component; the best models for the principal components are shown in Table 3. It is obvious that, after the optimization of G/PLS, the number of the MIEC terms could be greatly decreased without affecting the models. For example, the numbers of the MIEC terms in model 1 were reduced from 120 to 25, 25, 36, 43, and 39 using one, two, three, four, and five principal components, respectively. Compared with the q = 0.785 in model 1, models 4–8 achieved a cross-validation correlation coefficient of about 0.81.

The selection of the best model was carried out on the basis of the prediction for the 442 peptides in the test set. The prediction unsigned mean errors (UMEs) and the correlation between the predicted affinities and the experimental values for the test set are listed in Table 3. Table 3 shows that several models have similar prediction accuracies. For example, the UMEs for models 4, 10, 15, and 17 are 0.398, 0.398, 0.400, and 0.398, respectively. It is thus difficult to decide which model is the best. It has been proposed that the outputs of multiple models could be averaged to gain more reliable results than those obtained using a single best model. 29,30 We therefore averaged the prediction results of the best four models (models 4, 10, 15, and 17) to get the consensus prediction. The UME and r_{test} of the consensus prediction are 0.391 and 0.807, respectively, which are superior to those obtained using a single model. The correlation between the experimental affinities and the consensus predictions for the tested peptides is shown in Fig. 3a, and the distribution of the prediction errors for the peptides in the test set is shown in Fig. 3b. Of the 442 tested

**Table 2.** Prediction performances for various linear regression models using different MIEC terms

<table>
<thead>
<tr>
<th>Model</th>
<th>MIEC terms</th>
<th>n</th>
<th>N_Pc</th>
<th>N</th>
<th>r^2</th>
<th>q^2</th>
<th>PRESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ΔE_{vdw}, ΔE_{ele}</td>
<td>442</td>
<td>4</td>
<td>120</td>
<td>0.821</td>
<td>0.785</td>
<td>125.677</td>
</tr>
<tr>
<td>2</td>
<td>ΔE_{vdw}, ΔE_{ele}</td>
<td>442</td>
<td>4</td>
<td>109</td>
<td>0.825</td>
<td>0.787</td>
<td>124.847</td>
</tr>
<tr>
<td>3</td>
<td>ΔE_{vdw}, ΔE_{ele}, ΔE_{gb}</td>
<td>442</td>
<td>6</td>
<td>183</td>
<td>0.829</td>
<td>0.784</td>
<td>125.870</td>
</tr>
</tbody>
</table>

^a Number of peptides in the training set.

^b Number of principal components.

^c Number of MIEC terms in the model.

^d Correlation coefficient of the linear regression.

^e Correlation coefficient of LOO cross-validation.

^f Predictive error sum of squares.
peptides, 327 have absolute prediction errors smaller than 0.5 (74.0%), 84 have absolute prediction errors between 0.5 and 1.0 (19.0%), 26 have absolute prediction errors between 1.0 and 1.5 (5.9%), and 5 have absolute prediction errors larger than 1.5 (1.1%). Overall, about 95% peptides in the test set can be well predicted with absolute prediction errors smaller than 1.0.

Characterization of the binding interface by MIECs

We examined the MIEC terms selected by the final G/PLS models because they may provide insight into understanding the binding specificity of domain–peptide interactions. Figure 4 shows the loadings of the X variables for the first component of model 10 as an example, where loading reflects the contribution of each xᵢ to the first principal component. The loading plot can highlight the interaction pairs important for binding. The residue–residue interaction pairs with absolute loadings larger than 0.6 are labeled. As shown in Fig. 4, the first principal component of model 10 is predominantly defined by the interactions of the following residue–residue pairs: P2–Phe8, P2–Gln8, P3–Asn64, P3–Asp7, P4–Phe65, P5–Pro62, P6–Asn64, P8–Ala11, P8–Asp34, P8–Gln15, P9–Gln35, and P9–Gly38. Moreover, three peptide residue–residue pairs, P2–P3, P4–P5, and P5–P6, also have contributions. All the selected intermolecular interaction terms correspond to interactions of residues in the SH3 active site and modulate the specific binding of the peptide. Of these residue–residue interaction pairs, some are especially important, such as P8–Ala11, P8–Asp14, and P8–Glu15. These results are consistent with our previous calculations, which showed that Arg at this position can form effective interactions with five SH3 domain residues, Asp14, Gln15, Asp34, Gln35, and Trp39. Of these five residues, Asp14 and Gln15 are the most important, because they can form more favorable interactions with the peptide. For the binding peptides, the residues at positions P1, P7, and P10 are not involved in these important residue–residue pairs. The terminal residues at P1 and P10 are dynamically flexible and may not form strong and specific interactions with SH3. Examining the complex structure revealed that the side chain of the residue at P7 points toward the solvent and does not form direct interactions with SH3.

Comparison with the MM/PBSA and MM/GBSA methods

To further evaluate the performance of our method, we applied the MM/PBSA and MM/GBSA methods to compute the binding free energies for all 884 peptides in the training and test sets. Because of the large number of peptides, it would be very costly to run MD simulations to optimize each complex structure and to obtain snapshots for free energy calculations. Instead, we estimated the binding free energy for each peptide using the single minimized complex structure and calculated the correlation coefficients (r) between the predicted binding free energies or the individual energy components and the experimental affinities (Fig. 5). The binding free energies calculated by MM/PBSA (r = 0.67) correlate slightly better than those calculated by MM/PBSA (r = 0.60), compared with correlation coefficients larger than 0.78 by the MIEC–PLS models that were also built based on the single minimized complex structure. We also analyzed the correlation between the measured affinities and the individual energy components. Compared with the van der Waals interactions (r = 0.34), the electrostatic interactions (r = 0.45) show a better, albeit still weak, correlation. It is interesting to observe that the polar contributions of desolvation free energy calculated by PB and GB have the same correlations with the binding affinities (r = 0.43), which can be explained by the very high linear correlation between ΔG_{polar} of PB and that of GB (data not shown).

We then built a model by only reweighting the van der Waals, electrostatic, and desolvation components of the MM/GBSA energy function without

---

**Table 3. Prediction performances of the G/PLS models**

<table>
<thead>
<tr>
<th>Model</th>
<th>MIEC terms</th>
<th>n</th>
<th>NᵢC</th>
<th>N</th>
<th>r</th>
<th>q</th>
<th>PRESS</th>
<th>r_{test}</th>
<th>UME_{test}</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>ΔF_{vdw}, ΔF_{ele}</td>
<td>442</td>
<td>1</td>
<td>25</td>
<td>0.817</td>
<td>0.807</td>
<td>114.100</td>
<td>0.796</td>
<td>0.398</td>
</tr>
<tr>
<td>5</td>
<td>ΔF_{vdw}, ΔF_{polar}</td>
<td>442</td>
<td>2</td>
<td>25</td>
<td>0.823</td>
<td>0.809</td>
<td>113.172</td>
<td>0.796</td>
<td>0.402</td>
</tr>
<tr>
<td>6</td>
<td>ΔF_{vdw}</td>
<td>442</td>
<td>3</td>
<td>36</td>
<td>0.830</td>
<td>0.807</td>
<td>114.336</td>
<td>0.797</td>
<td>0.403</td>
</tr>
<tr>
<td>7</td>
<td>ΔF_{vdw}</td>
<td>442</td>
<td>4</td>
<td>43</td>
<td>0.836</td>
<td>0.807</td>
<td>114.077</td>
<td>0.793</td>
<td>0.405</td>
</tr>
<tr>
<td>8</td>
<td>ΔF_{vdw}</td>
<td>442</td>
<td>5</td>
<td>39</td>
<td>0.838</td>
<td>0.807</td>
<td>114.103</td>
<td>0.787</td>
<td>0.410</td>
</tr>
<tr>
<td>9</td>
<td>ΔF_{vdw}</td>
<td>442</td>
<td>1</td>
<td>28</td>
<td>0.811</td>
<td>0.791</td>
<td>122.614</td>
<td>0.789</td>
<td>0.417</td>
</tr>
<tr>
<td>10</td>
<td>ΔF_{vdw}</td>
<td>442</td>
<td>2</td>
<td>25</td>
<td>0.824</td>
<td>0.803</td>
<td>116.073</td>
<td>0.799</td>
<td>0.398</td>
</tr>
<tr>
<td>11</td>
<td>ΔF_{vdw}</td>
<td>442</td>
<td>3</td>
<td>26</td>
<td>0.827</td>
<td>0.798</td>
<td>118.896</td>
<td>0.799</td>
<td>0.405</td>
</tr>
<tr>
<td>12</td>
<td>ΔF_{vdw}</td>
<td>442</td>
<td>4</td>
<td>34</td>
<td>0.837</td>
<td>0.814</td>
<td>110.681</td>
<td>0.796</td>
<td>0.405</td>
</tr>
<tr>
<td>13</td>
<td>ΔF_{vdw}</td>
<td>442</td>
<td>5</td>
<td>34</td>
<td>0.840</td>
<td>0.810</td>
<td>112.514</td>
<td>0.795</td>
<td>0.409</td>
</tr>
<tr>
<td>14</td>
<td>ΔF_{vdw}</td>
<td>442</td>
<td>1</td>
<td>18</td>
<td>0.809</td>
<td>0.798</td>
<td>118.934</td>
<td>0.796</td>
<td>0.407</td>
</tr>
<tr>
<td>15</td>
<td>ΔF_{vdw}</td>
<td>442</td>
<td>2</td>
<td>28</td>
<td>0.822</td>
<td>0.806</td>
<td>114.888</td>
<td>0.794</td>
<td>0.400</td>
</tr>
<tr>
<td>16</td>
<td>ΔF_{vdw}</td>
<td>442</td>
<td>3</td>
<td>33</td>
<td>0.831</td>
<td>0.805</td>
<td>115.264</td>
<td>0.802</td>
<td>0.406</td>
</tr>
<tr>
<td>17</td>
<td>ΔF_{vdw}</td>
<td>442</td>
<td>4</td>
<td>40</td>
<td>0.835</td>
<td>0.803</td>
<td>116.190</td>
<td>0.789</td>
<td>0.398</td>
</tr>
<tr>
<td>18</td>
<td>ΔF_{vdw}</td>
<td>442</td>
<td>5</td>
<td>41</td>
<td>0.836</td>
<td>0.800</td>
<td>117.709</td>
<td>0.802</td>
<td>0.408</td>
</tr>
</tbody>
</table>

*a* Correlation between the experimental log(BLU) and the predicted values for the test set.

*b* UME for the test set.

---
performing pairwise residue decomposition. The correlation coefficient of the model is 0.69, which is slightly better than the total MM/GBSA (−0.67), but it is still much worse than that of the model based on the pairwise decomposed energies. This calculation shows that only reweighting the energy components of the MM/GBSA energy terms may not effectively remove the noise introduced by some residue–residue pairs. In contrary, weighting each pairwise interaction component in the MIEC model can better characterize the local environments. For example, the dielectric constant inside a protein or at the interface of protein–protein interaction depends on the local chemical environment. In the total MM/GBSA model, a universal dielectric constant is used for the entire complex, and such a constant value may introduce noise to some residue–residue interactions. Without reweighting as in the total MM/GBSA model, the noise will be included in the total binding free energy. In our model, the weights of these noisy pairwise interactions are small; thus, the prediction accuracy is improved.

Distinguishing binders and non-binders based on MIECs

MIECs coupled with PLS regression can give quantitative predictions for binding affinities of peptides, which require a training set of peptides with measured binding affinities. In reality, such quantitative data are not always available. Rather, peptides are often only known as binders or non-binders to a domain. Also, in many situations, it is sufficient to just classify peptides into binder and non-binder groups for a domain. We therefore examined the usefulness of MIECs in this peptide classification problem. Based on the experimentally measured Boehringer light unit (BLU) values, the 884 peptides were divided into two classes (binder and non-binder). Our goal was to correctly identify the class for a given peptide using the MIECs.

Classification models using SVM

SVM is a widely used classification technique in biology, and we took it as the classification model in this study. Given the data set, the proper kernel function must be chosen to construct the best classifier. This selection is very important because the kernel function determines the sample distribution in the higher-dimension feature space. There is no gold standard in choosing the form and parameters of the optimal kernel function, but studies have shown that the radial basis function (RBF) and polynomial kernel functions perform better than others in most cases. We examined the performances for SVMs using four kernel functions and different MIEC terms (Table 4). Based on the cross-validation results, the linear kernel has much lower sensitivity compared with the other three kernels, which have comparable SE, SP, Q+, Q−, and C values. The RBF kernel achieved the highest Matthews correlation coefficient C, which suggests the best classification, no matter what MIEC terms were used. With the use of the RBF kernel, SVM using all three MIECs has the highest C value (Table 4). This predictor can successfully identify 77.8% of peptides in the binder class and 91.0% of peptides in the non-binder class on average in the 1000 runs of threefold cross-validation.

We should point out that the data set we used is quite unbalanced: the number of binders is only 41, which is less than 5% of the total sample. Therefore, different penalty parameters were applied on different classes [see Eq. (8)] in training SVMs. The importance of using different penalty parameters was demonstrated by examining the prediction accuracy of SVM using the same weight parameters for binder and non-binder classes: the prediction accuracy of the non-binder class is 100%, but that of
the binder class is 0%. The calculation suggests that the SVM classifier was strongly biased to the non-binder class when the same weight parameters were used for different classes.

The SVM can give very good predictions for the non-binder class, indicated by the accuracy rates of 91.7% and 91.0% for the non-binder class in the training and test sets, respectively. Despite the high prediction accuracy for the non-binder class, since the data set is unbalanced and contains many more non-binders than binders, 72 non-binding peptides were identified as binders, and they are false positives. We suggest that other criteria may be used to filter out these false positives when predicting interaction partners of modular domains in a proteome. For example, because of the functional constraint of forming protein–protein interactions, the peptides binding to modular domains tend to be conserved across species and can thus be used as a filter to reduce false positives in the prediction.16

On the other hand, it is important to note that an unbalanced data set should be used to train SVM when predicting interacting partners for modular domains in a proteome because the binders only occupy a small portion compared with the non-binders. We investigated the performance of SVM using a balanced training data set. Following the procedure of Zhang et al.,13 we retrained an SVM classifier using a balanced data set of 41 binders and 41 randomly selected non-binders. The prediction accuracies for binders and non-binders are 90.2% and 85.4%, respectively. We then applied this classifier to predict all samples in the data set. The prediction accuracy for the 841 peptides in the non-binder class is only 623/841 (74.1%), which means that more than 200 non-binders were classified as binders. Obviously, the model trained on the balanced data set generated many more false positives.

Comparison with the MM/GBSA methods for classification

Finally, we examined the performance of MM/GBSA on peptide classification. The distributions of the binding free energies calculated by MM/GBSA for binders and non-binders are shown in Fig. 6. Student’s t test was employed to evaluate the significance of the difference between the means. The p value associated with the difference in the mean binding energies of the 41 known peptide binders versus the 843 non-binders was 1.05e−19 at the 95% confidence level, indicating that the two distributions are significantly different. To compare the performance of classification using MM/GBSA and MIECs, we used a cutoff of −67.0 kcal/mol to separate binders and non-binders. A peptide was considered to be a binder if ΔGbind was smaller than −67.0 kcal/mol; otherwise, it was considered to be a non-binder. This cutoff was chosen so that MM/GBSA and MIEC–SVM classifiers had a similar prediction accuracy for the binder class (78.0% versus 77.8%). Of the 843 non-binders, 694 had a ΔGbind > −67.0 kcal/mol and were correctly classified. On the other hand, 149 non-binding peptides with a ΔGbind < −67.0 kcal/mol were predicted to be binders, and they are false positives, compared with 72 misclassified non-binding peptides by SVM based on MIECs. MM/GBSA can generate two distributions of binding free energies for binders and non-binders, but MIEC-based SVM models still
Characterizing Domain–Peptide Interactions

Fig. 5 (legend on previous page)
have a much lower false-positive rate in peptide classification.

**Conclusions**

In the present study, we proposed a method to characterize the binding interface between SH3 domain and its binding peptides using MIECs. The MM/GB decomposition analysis was used to generate the MIECs, including van der Waals, electrostatic, and polar contribution to desolvation energy. We demonstrated the usefulness of MIECs to capture domain–peptide interaction features on peptide binding affinity prediction using PLS regression and peptide classification using SVM. Compared with the conventional MM/PBSA and MM/GBSA calculations, MIECs coupled with statistical analysis are more robust to noise or error in free energy calculation and less sensitive to the structure modeling accuracy.

Compared with the other bioinformatics methods to predict domain–peptide interactions, our approach provides a more rigorous and sophisticated way to capture the energetic patterns of the binding interface and thus the binding specificity between domains and peptides. The fully filled MIEC matrix is a compact encoding of the residue–residue interactions on the interface and avoids the overfitting problem that is more likely to occur in training classification models on the sparse contact matrix. Conformational preference of peptides is considered by calculating intrapeptide MIECs, and the conformational flexibility of residues on the interface is, at least partially, considered by modeling of each domain–peptide complex followed by MM optimization. These features distinguish our approach from those approaches that, for example, use simple energetic measurement for residue–residue contacts to build classification models. More importantly, our approach can reveal biophysical properties that dominate the domain–peptide binding. Of course, modeling complex structures and calculating MIECs take more computer hours than the other bioinformatics approaches. Given the faster and faster computers we have, we expect that our approach will become more and more useful.

On both binding affinity prediction and peptide classification, MIECs coupled with statistical analysis are superior to ranking peptide binding affinities calculated by MM/GBSA and MM/PBSA. This result is not as surprising as it looks at first glance. MM/GBSA or MM/PBSA calculation, if conformational entropy is not included, considers the interactions, including van der Waals, electrostatic, and desolvation energy, between all inter- and intramolecular pairs of protein and peptide residues. Some of these terms may be noisy and inclined to error due to, for example, inefficient sampling or inaccurate force field parameterization. PLS or SVM works as an additional filter to select the interaction pairs and MIEC terms that are most informative to binding affinity prediction or peptide classification: PLS assigns larger weights to the informative $x_i$ values, and SVM uses informative $x_i$ values to generate a hyperplane for classification in the higher-dimension space. As long as the overall pattern of the interaction interface is captured by the modeled complex structure and MIEC calculations, statistical analysis of PLS or SVM is resistant to noise or error in some of the interaction pairs and thus, as shown in our analysis, is able to achieve higher prediction accuracy.
Characterizing Domain–Peptide Interactions

Materials and Methods

Data set

Landgraf et al. combined phage display and SPOT synthesis to determine the peptides in the yeast proteome binding to eight yeast SH3 domains and those in the human proteome binding to two human SH3 domains.28 The SPOT signal intensity of each peptide was quantitatively measured in BLUs. For the hAmph1 SH3 domain, the BLU values for the 884 peptides analyzed by replicate experiments were used in this study to approximate binding affinity. The 884 peptides were randomly divided into two equal groups as the training and test sets. All peptides are 10 residues long, and the positions for these 10 residues are referred to as P1 to P10 from the N-terminus to the C-terminus.

Modeling the SH3/peptide complexes

The structure of the hAmph1 SH3 domain was modeled based on the crystal structure of the rat amphiphysin-2 SH3 domain by homology modeling and optimized by 3-ns MD simulations in our previous study.31 The predicted model of the hAmph1 SH3 domain complexed with the peptide PPLRRPRAA was used as the template to construct all hAmph1/peptide complexes.32 The template peptide PPLRRPRAA was mutated to the other sequences by using the sander program.33 Energy minimization for each complex was carried out using the sander program in AMBER9.034 and the AMBER03 force field.35 The solvent effect was considered by using the GB model (gb = 2) implemented in sander.36 The maximum number of minimization steps was set to 3000. The first 500 steps were performed with the steepest descent algorithm, whereas the rest of the steps were performed with the conjugate gradient algorithm. The cutoff for non-bonded interaction was 10 Å, and the convergence criterion for the rms of the Cartesian elements of the energy gradient was 0.2 kcal/mol/Å. The minimized structures for all SH3/peptide complexes were saved for further analysis.

Calculating the MIECs

For each complex, the minimized conformation was used in the following analysis: First, the important SH3 residues for peptide binding were identified, which were defined as those within 8 Å of each peptide residue in the template structure of the hAmph1 SH3 domain complexed with the peptide PPLRRPRAA. In total, 57 residue–residue interaction pairs between the peptide and the SH3 domain were identified (Table 1), and the interaction between each residue–residue pair was computed using the MM/GB energy decomposition protocol in AMBER.36,37 In addition, we also calculated the MIECs for the 9 residue pairs between adjacent peptide residues because they may characterize the conformational preference of the peptide. The MIEC calculations, including read-in of the SH3/peptide complexes, definition of atom types in the GB calculation, and assignment of the force field parameters, were automatically carried out using the gsoap program (which will be released in AMBER10 in early 2008).37 The MIECs for each important residue–residue pair include (a) electrostatic interaction (coulombic interaction), (b) van der Waals interaction, and (c) polar contribution to desolvation free energy, referred to as the polar desolvation energy, which is calculated by the following equation:38

$$\Delta G_{gb} = -\frac{1}{2} \sum_{ij} q_i q_j \exp \left( -\frac{r_{ij}}{\kappa R_i R_j} \right) \left( 1 - \frac{e^{-r_{ij}/\kappa R_i R_j}}{r_{ij}/\kappa R_i R_j} \right)$$

where \(r_{ij}\) is the distance between atoms \(i\) and \(j\), \(q_i\) and \(q_j\) are the partial charges of atoms \(i\) and \(j\), respectively; \(R_i\) and \(R_j\) are the effective Born radii of atoms \(i\) and \(j\), respectively; \(\kappa\) is the Debye–Hückel screening parameter, and \(\varepsilon\) is the solvent dielectric constant. The smooth function \(f_{gb}^{st}\) is calculated by:

$$f_{gb}^{st} = \left[ r_{ij}^2 + R_i R_j \exp \left( -r_{ij}/\kappa R_i R_j \right) \right]^2$$

The cutoff for calculating \(\Delta E_{vdw}\) and \(\Delta E_{ele}\) was set to 18.0 Å. A distance-independent interior dielectric constant of 1 was used to calculate \(\Delta E_{ele}\). The charges used in the GB calculations were taken from the AMBER03 force field.35 The values of interior dielectric and exterior dielectric constants were set to 1 and 80, respectively. The GB parameters developed by Tsui and Case were used.30

The above three MIECs of the 66 residue pairs were calculated for each peptide under consideration. The interaction between the hAmph1 SH3 domain and the peptide was thus characterized by a variable \(X\) (Fig. 2). Depending on which interaction fields are included (e.g., van der Waals alone or both van der Waals and electrostatic), the dimension of \(X\) is 66, 132 (66 × 2), or 198 (66 × 3). The logarithmic (base 10) BLU, log(BLU), value for the 884 peptides was the response variable \(Y\) in the statistical analysis subsequently described.

PLS regression models

For each dimension or component of \(X\) (i.e., predictor), \(x_i\) was normalized with a zero mean and unit standard deviation. A standard deviation of 0.02 before normalization was used as a heuristic cutoff to remove uninformative \(x_i\). PLS analysis was then conducted on the input matrix to generate a correlation model between \(X\) and \(Y\). PLS regression is particularly useful in predicting a set of dependent variables from a large set of independent variables.39 PLS regression searches for a set of principal components that performs a simultaneous decomposition of \(X\) and \(Y\) with a constraint that these components explain as much of the covariance between \(X\) and \(Y\) as possible. It is followed by a regression step where the decomposition of \(X\) is used to predict \(Y\).41

An initial PLS model was built using all \(x_i\) values, and the optimal number of the principal components was chosen as that produced the most predictive models, as judged by the LOO cross-validation regression coefficient. Here, a G/PLS in Cerius2 was applied to identify the best prediction models by automatically removing these noisy independent variables.41 PLS regression searches for a set of principal components that performs a simultaneous decomposition of \(X\) and \(Y\) with a constraint that these components explain as much of the covariance between \(X\) and \(Y\) as possible. It is followed by a regression step where the decomposition of \(X\) is used to predict \(Y\).41

The values of interior dielectric and exterior dielectric constants were set to 1 and 80, respectively. The GB parameters developed by Tsui and Case were used.30

The above three MIECs of the 66 residue pairs were calculated for each peptide under consideration. The interaction between the hAmph1 SH3 domain and the peptide was thus characterized by a variable \(X\) (Fig. 2). Depending on which interaction fields are included (e.g., van der Waals alone or both van der Waals and electrostatic), the dimension of \(X\) is 66, 132 (66 × 2), or 198 (66 × 3). The logarithmic (base 10) BLU, log(BLU), value for the 884 peptides was the response variable \(Y\) in the statistical analysis subsequently described.
linear polynomial terms were used in the PLS models. In G/PLS calculations, PLS models with one to five principal components were systematically examined. The performance of a model was evaluated by LOO cross-validation using the training set and by its prediction accuracy on the test set.

Classification models based on an SVM

The 884 peptides were separated into two classes according to the experimental BLU values. If the BLU value is larger than 10,000, a peptide was considered as a binder (positive); otherwise, it was in the negative class (non-binder). An SVM was trained on these data to classify peptides based on the normalized MIECs. Namely, there are 884 peptides in a 198-dimension MIEC space (if all three types of MIECs were used). The goal of SVM is to find a hyperplane to separate binding and non-binding peptides with a maximum margin. We used the LIBSVM program developed by Chang and Lin in this study to train SVM.

The performance of the SVM models was assessed by sensitivity SE, specificity SP, prediction accuracy for positives (binders) Q₂, prediction accuracy for negatives (non-binders) Q₁, and Matthews correlation coefficient C. The Matthews correlation coefficient C ranges from 0 to 1 (C = 1 means perfect prediction).

\[
SE = \frac{TP}{TP + FN} \quad (3) \\
SP = \frac{TN}{TN + FP} \quad (4) \\
Q₂ = \frac{TP}{TP + FP} \quad (5) \\
Q₁ = \frac{TN}{TN + FN} \quad (6) \\
C = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \quad (7)
\]

where true positive TP, true negative TN, false positive FP, and false negative FN are defined following the convention.

A rigorous training-validation procedure was designed to assess the statistical significance of the SVM predictors. The whole data set was randomly divided into three groups with equal sizes. Two groups were used for training, and the third group was used for testing (threefold cross-validation). This procedure was run 1000 times, and SE, SP, Q₂, Q₁, and C were calculated for the test subset in each run.

Considering that only about 5% of the peptides are defined as binders in the positive class, the numbers of data points in positive and negative classes are quite unbalanced. Therefore, in LIBSVM, when we use different penalty parameters, the objective function becomes:

\[
\begin{aligned}
\min_{w,b,k} & \frac{1}{2} w^T w + C \sum \xi_i + \sum \xi_i \\
\text{subject to } & y_i(w \phi(x_i) + b) \geq 1 - \xi_i, \\
& \xi_i \geq 0, i = 1, \ldots, I.
\end{aligned} \quad (8)
\]

where \(C = k, C = k, C\) are the penalty parameters of error terms for the positive class and the negative class, respectively, and \(k, k, k\) are the weight parameters for the positive and negative classes, respectively. Here a higher weight \(k\) was given for the binder class \(k = 14\), while \(k = 1\).

Calculating binding free energies for peptides using MM/PBSA and MM/GBSA

Based on the single minimized complex structure, the binding free energy for each peptide was calculated using the MM/PBSA and MM/GBSA methods.

\[
\Delta G_{\text{bind}} = G_{\text{complex}} - G_{\text{protein}} - G_{\text{ligand}} = \Delta E_{\text{MM}} + \Delta G_{\text{polar}} + \Delta G_{\text{non-polar}} - T \Delta S \quad (9)
\]

where \(\Delta E_{\text{MM}}\) represents the change of MM potential energy upon ligand binding; \(\Delta G_{\text{polar}}\) and \(\Delta G_{\text{non-polar}}\) represent the polar component and the non-polar component of solvation free energy, respectively; and \(-T\Delta S\) is the conformational entropy change, which was not considered in this study because of the high computational cost.

\(\Delta E_{\text{MM}}\) was calculated using the \textit{sander} program in AMBER9.0. In MM/PBSA calculations, \(\Delta G_{\text{polar}}\) was computed using the \textit{pbsa} program in AMBER9.0 to solve the PB equation. The grid size for the PB calculations was 0.5 Å, and the values of interior and exterior dielectric constants were set to 1 and 80, respectively. In MM/GBSA calculations, \(\Delta G_{\text{polar}}\) was computed using the GB model with the parameters developed by Tsui and Case. \(\Delta G_{\text{non-polar}}\) was estimated based on the solvent-accessible surface area (SASA) as \(\Delta G_{\text{non-polar}} = 0.0072 \times \text{SASA}\).

Acknowledgements

Simulations were performed on the Linux cluster at the University of California at San Diego Center for Theoretical Biological Physics. This work was partially supported by the National Science Foundation Physics Frontier Centers-sponsored Center for Theoretical Biological Physics (grant nos. PHY-0216576 and PHY-0225630). T.H. is supported by a Center for Theoretical Biological Physics postdoctoral scholarship. We thank Prof. J. Andrew McCammon for providing access to the Cerius2 and InsightII molecular simulation packages.

References

4. Castagnoli, L., Costantini, A., Dall’armi, C., Gonfloni,
Characterizing Domain–Peptide Interactions


