Using genome-wide measurements for computational prediction of SH2–peptide interactions

Zeba Wunderlich[1] and Leonid A. Mirny[2,*]

[1]Biophysics Program, Harvard University, Cambridge, MA 02138 and [2]Harvard-MIT Division of Health Sciences and Technology and Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Received March 30, 2009; Revised April 29, 2009; Accepted April 30, 2009

ABSTRACT

Peptide-recognition modules (PRMs) are used throughout biology to mediate protein–protein interactions, and many PRMs are members of large protein domain families. Recent genome-wide measurements describe networks of peptide–PRM interactions. In these networks, very similar PRMs recognize distinct sets of peptides, raising the question of how peptide-recognition specificity is achieved using similar protein domains. The analysis of individual protein complex structures often gives answers that are not easily applicable to other members of the same PRM family. Bioinformatics-based approaches, one the other hand, may be difficult to interpret physically. Here we integrate structural information with a large, quantitative data set of SH2 domain–peptide interactions to study the physical origin of domain–peptide specificity. We develop an energy model, inspired by protein folding, based on interactions between the amino-acid positions in the domain and peptide. We use this model to successfully predict which SH2 domains and peptides interact and uncover the positions in each that are important for specificity. The energy model is general enough that it can be applied to other members of the SH2 family or to new peptides, and the cross-validation results suggest that these energy calculations will be useful for predicting binding interactions. It can also be adapted to study other PRM families, predict optimal peptides for a given SH2 domain, or study other biological interactions, e.g. protein–DNA interactions.

INTRODUCTION

In the crowded cellular milieu, the ability of proteins to specifically interact with different molecules, e.g. sites on DNA or short peptides, is essential to virtually every cellular process from DNA replication to cell signaling. The inherent challenges of this problem and nature's apparent success at engineering these specific interactions raise a number of questions, which usually center on either the kinetics of the process or the equilibrium picture: How does recognition happen rapidly and specifically?

Here we study one flavor of the equilibrium problem: how does a peptide-recognition module (PRM) discriminate its particular targets from a large pool of peptides? PRMs are protein domains that recognize specific peptides and are used to mediate protein–protein interactions in many contexts. There are a number of large protein domain families that serve as PRMs. The members of these families are quite similar to each other, but each member can recognize a specific subset of the peptide pool.

Our goal is to gain a physical understanding of this recognition problem. To do this, we construct a potential that describes the interaction energies between a family of PRMs and their peptides. We aim to construct this potential in a way that allows us to apply it to domain–peptide pairs unseen in the construction process, to use the calculated energy to predict whether or not they interact and to gain some insight into the mechanism used to achieve peptide-binding specificity in families of similar PRMs. Our approach uses a minimum amount of prior knowledge about the particular PRM family, so it can be easily applied to new families and can be used both when structures are available and when they are not.

As a model system, we use the SH2 domain family of PRMs, which transmit signals from receptor tyrosine kinases (RTKs) to the cell interior by interacting with the RTK cytoplasmic domains, specifically the tyrosines that are phosphorylated upon RTK activation and their surrounding residues (1). In 2006, MacBeath and coworkers published a protein microarray-based study that quantitatively assessed the interaction strength of almost every human SH2 domain with pY peptides extracted from four ErbB family RTKs (2). The availability of this large amount of data describing the interactions between a variety of SH2 domains and peptides, all measured using the same technology, makes the SH2 domain family an ideal test case for our effort.

*To whom correspondence should be addressed. Tel: +1 617 452 4862; Fax: +617 253 7498; Email: leonid@mit.edu

Other efforts to understand the origins of PRM–peptide specificity usually follow a structural biology-, computational structural biology-, or a bioinformatics-based approach (3–7). Structural biologists have long studied the structures of PRM–peptide complexes, visually analyzing them for positions in the domain and peptide that appear to be important for recognition based on their proximity and likelihood of interaction [e.g. (8–12)]. This approach is sometimes difficult to apply to large amounts of structural data, and it may be difficult to generalize the results to an entire PRM family.

Several computational structural biology approaches have also been applied to the PRM specificity problem. These approaches generally develop energy models of the PRM–peptide interactions based on collections of structural data. The SH2 problem has been studied using a variety of these methods (13–16), most recently in a study in which an empirical energy model was constructed to predict genome-wide targets of nine human SH2 domains (17). Some of these techniques have been quite successful at recapitulating *in vitro* and *in vivo* interaction data, but may be computationally intensive or difficult to apply to domains with no structural data.

A method developed by Ferraro *et al.* (18) is among these structurally based methods of predicting PRM specificity and works by transferring contacts observed in a few available structures into domains and peptides lacking such structural information. Zhang *et al.* (19) integrated a large amount of interaction and structural data to model SH3 and PDZ domain–peptide interactions. They use a combination of neural networks and support vector machines to construct classifiers based on various physical and chemical properties of amino acids. Chen *et al.* (20) have implemented a method developed in collaboration with our group and similar to one presented here.

Our approach is different from these studies in the following aspects: First, along with direct protein–peptide contacts inferred from structural data, we use an information-based approach to detect more distant pairs of residues that nevertheless demonstrate signs of co-evolution (21). Furthermore, we evaluate contributions of such indirect interactions and demonstrate their significant role in providing specificity through ‘indirect readout’. Second, in contrast to machine-learning approaches such as neural networks and support vector machines, our method provides information about the origin of specificity by quantifying the energy contribution of each contact. However, direct comparison of these different methods is complicated by the different quality criteria and domain families used by each study, calling for a community-wide effort, similar to the CASP initiative in protein folding (22).

Bioinformatics-based approaches often center on machine learning methods, which are able to leverage large amounts of data, but do not always give a physical understanding of specificity. Many of these methods seek to calculate position-specific scoring matrices (PSSMs), which give the likelihood of an amino acid occurring at a given position of a bound peptide (23–26). Two notable previous efforts to model PRM–peptide interactions applied a modified Gibb’s sampler (25) and a probabilistic discriminative model (23) to infer the PSSMs of SH3 domains. The former study employed a discriminative prior to find motifs that discriminate between bound and unbound domain–peptide pairs in a yeast two-hybrid SH3 domain interaction network. The latter eliminated the user-defined parameters utilized in the first approach and used a normalization technique to avoid over-fitting of the model to phage display and yeast two-hybrid data. While moderately successful at recapitulating the SH3 interaction network, these approaches are not transferable to unseen domains and do not give insight into the general principles responsible for SH3 peptide recognition. There are also more general models [e.g. (27)], that aim to explain the general properties of protein–protein networks. There have also been successful applications of structure-based threading approaches to other PRM specificity problems, e.g. the recognition of peptides by MHC molecules (28,29), a unique and widely-studied PRM problem (30).

Due to the similarity of the problems, approaches in protein–DNA interactions also merit mention. There are approaches to derive energy potentials from structural studies (31–36), physics-based methods of deriving PSSMs from large scale-binding data (37), and methods to combine structural and sequence data to predict binding sites for a family of transcription factors (38).

Here we develop a physically motivated energy model to describe the interaction energy between SH2 domains and their pY peptides. We measure the performance of the basic model and several variants for predicting the experimentally characterized interactions of a diverse set of SH2 domains and peptides. We find that an amino-acid-based potential with non-specific interaction terms can accurately predict SH2 domain–peptide interactions and that the amino acid-based potential can be used to predict interactions with domains or peptides that were not used to derive the potential. Using structure- and information-based techniques, we find the amino-acid positions in the SH2 domains and pY peptides that confer specificity to the interactions. The technique is sufficiently general that it can be used to predict the peptide partners for new SH2 domains and to model interactions for other PRM families.

MATERIALS AND METHODS

Data preparation

The data is taken from (2). Only the 115 SH2 domains are considered, so the 44 PTB domains are discarded. There are 10 double-SH2 domains, which are also discarded. There are 66 peptides, but we discard all unphosphorylated peptide data points. There are 33 peptides which are singly phosphorylated; two are discarded due to high background binding, leaving 31 peptides and $105 \times 31 = 3255$ interactions. Domain–peptide pairs with measured K_d values less than $2\,\mu M$ are considered bound (198 pairs), all others are considered unbound. The SH2 domains are aligned using MUSCLE (39), and the peptides are aligned by the pY position.

Interaction map construction

To create an energy model to describe domain–peptide interactions, our approach requires two steps: identifying pairs of amino-acid positions in the domain and peptide involved in recognition (*the interaction map*), and deriving an *energy potential* that describes in interaction energy between pairs of amino acids. Here we describe three methods for constructing an interaction map: an information-based method, a structure-based method and a hybrid method. For each construction method, we first define a potential set of domain–peptide amino-acid position pairs—pairs of positions in the domain sequence alignment (made using MUSCLE) and in the peptide sequence alignment (aligned with the pY position as position 0). Then, we order the pairs using some criteria and add them to the interaction map in the order specified by the criteria.

For the information-based interaction maps, all position pairs are considered, and the criteria used to evaluate the pairs are the normalized Kullback–Leibler (KL) divergence between the amino-acid distributions at the position. The general formula for KL divergence of two discrete probability distributions p and q is

$$KL = \sum_{i} p(i) \log\left[\frac{p(i)}{q(i)}\right]$$

In our case, for each possible domain–peptide pair, we let p be the distribution of amino acid–amino acid frequencies in the bound cases and q is the distribution in the unbound cases. For example, to calculate the KL divergence of the 10–0 position pair, we extract the amino-acid pairs present in position 10 of the domain (according to the MUSCLE alignment) and position 0 of the peptide (the pY position). We then tabulate the observed frequencies of each possible amino-acid pair for bound and unbound cases, adding a pseudocount of 1 to each entry to avoiding taking the log of or dividing by 0. Using these empirical p and q distributions, we find the KL divergence. To control for differences in the inherent diversity of the data set at each position pair, we also calculate the KL divergence for a control where the bound/unbound labels are shuffled randomly. We do this 100 times and calculate the mean and standard deviation of the KL divergence for the random shuffles. We then assess the significance of the 'real' KL divergence by calculating a Z-score: $Z = (KL - \langle KL_{control} \rangle)/\sqrt{\text{var}(KL_{control})}$, and pick contacts with the largest Z-scores.

For the structure-based maps, position pairs found to be in 'contact' in SH2–peptide structures are considered. We download all the SH2 structures available in the Protein Data Bank (40) as defined by SCOP (41) (144 structures as of February 2006) and manually sort through the structures, retaining only those bound to a tyrosine-containing peptide. We remove structures with peptides containing two phosphorylated peptides and non-standard amino acids and structures with extraneous domains. For NMR structures, only one model is used, either the averaged model or the first model in ensemble. For structures with multiple identical copies of the domain–peptide pair, only the first domain–peptide combination is retained. After this filtering 38 structures remain, 11 NMR structures and 27 crystal structures. The crystal structures have a resolution of at least 3 Å, with an average of 1.9 Å and a R-factor of 0.23 or better. The PDB codes of these structures are listed in Supplementary Table S2. Using RasMol (42), for each structure, we extract all domain–peptide position pairs in which non-backbone atoms are within a given distance threshold of each other. We vary the threshold from 3.5 to 6.5 Å in 1-Å increments. To construct an interaction map, we add position pairs by the number of structures in which they are found to be in contact.

For the hybrid interaction maps, we only consider position pairs found to be in contact according to the structure criteria and then order the pairs by information-based significance.

To ensure that low-resolution NMR structures do not introduce spurious contacts into the analysis, structure-based and hybrid contact maps were constructed by excluding NMR structures while using best parameter combination (distance cutoff = 5.5 Å, number of contacts = 10). The obtained structure-based contact maps were identical to the maps calculated using all the structures. For the hybrid contact map, 7/10 of the contacts were identical.

Optimization of energy potential

To optimize the energy potential given an interaction map, we adapt the procedure presented in (43). In Equations (1)–(3), we describe our basic energy model. Here, for brevity, let us assume the calculated energy of a domain–peptide pair can be written, as in Equation (3), as:

$$E = d \cdot u.$$

We assume our energy potential u is a 1×400 vector in which each element corresponds to the energy of interaction between a particular amino acid x in the domain and amino acid y in the peptide, hence the $20 \times 20 = 400$ values. In each corresponding element in d, there is a count of the number interactions that involve amino acid x in the domain and y in the peptide.

To derive the optimal energy potential (u^*) for a given interaction map, first, we define a score, Z, which expresses Figure 1A mathematically:

$$Z = \frac{\mu_- - \mu_+}{\sigma_- + \sigma_+},$$

where μ_+ and σ_+ are the mean and standard deviation of the calculated energies of the bound domain–peptide pairs μ_- and σ_- are the same for the unbound pairs. For the basic model, we can write μ_+ and σ_+ as:

$$\mu_+ = \langle d(k) \cdot u \rangle_{k \in \text{all bound pairs}} = \langle d(k) \rangle \cdot u = d_+ \cdot u$$

$$\mu_- = \langle d(l) \rangle_{l \in \text{all unbound pairs}} \cdot u = d_- \cdot u$$

$$\sigma_+^2 = \text{var}(d(k) \cdot u) = u' \cdot \text{cov}(d(k)) \cdot u$$

$$\sigma_-^2 = u' \cdot \text{cov}(d(l)) \cdot u$$

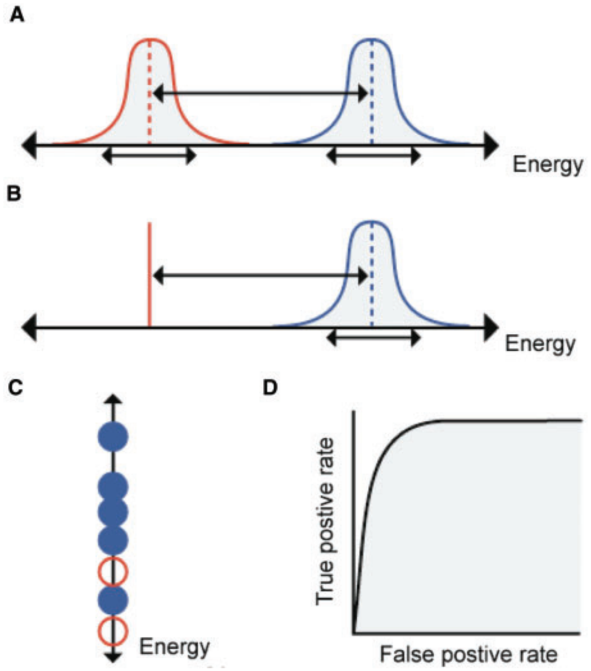

Figure 1. An illustration of the energy potential construction method and the performance metrics. (**A**) Given an interaction map, which specifies the interacting amino acids between the peptide and the domain, we find an energy potential that maximizes the gap between the mean energy of the bound (red) and unbound (blue) peptide–domain pairs, while minimizing the width of these distributions. (**B**) This is analogous to a method used in protein folding that finds, given a map of intramolecular contacts, a potential that maximizes the energy gap between the native structure (red line) and decoys (blue distribution), while minimizing the width of decoy energy distribution. (**C**) We use two metrics to assess the ability of a potential to distinguish between bound and unbound domain–peptide pairs, based on their relative energies. To find the bound rank metric, we order the pairs by increasing energy and calculate the median rank of the bound pairs. In this illustration, the bound pairs are shown as open, red circles on the energy scale and the unbound pairs are filled blue circles. The bound rank here would be median [1, 3] = 2. (**D**) To calculate the ROC AUC, the area under the ROC curve, we construct a ROC curve by plotting the true positive rate versus the false positive rate for various energy thresholds and calculate the area under the curve. The ROC AUC corresponds to the probability that the energy of a randomly chosen bound pair is less than a randomly chosen unbound pair.

Here d_+ and d_- are the average d vectors for the bound and unbound domain–peptide pairs, respectively.

With this, we can rewrite Z as:

$$Z = \frac{d_- \cdot u - d_+ \cdot u}{\sqrt{u' \cdot \mathrm{cov}(d(l)) \cdot u + u' \cdot \mathrm{cov}(d(k)) \cdot u}}$$

$$= \frac{(d_- - d_+) \cdot u}{\sqrt{u' \cdot [\mathrm{cov}(d(l)) + \mathrm{cov}(d(k))] \cdot u}} = \frac{a \cdot u}{\sqrt{u' \cdot B \cdot u}}$$

Here $B = \mathrm{cov}(d(l)) + \mathrm{cov}(d(k))$. For convenience, we optimize Z^2. Setting $\frac{d(Z^2)}{du} = 0$, we get:

$$u^* \propto B^{-1} a$$

Since the performance metrics of a particular energy potential u are unchanged if it is scaled by a constant c or increased by a constant b, the scale of the optimal u is arbitrary.

When adding terms to the energy model the Z-score is modified accordingly and optimized with respect to each term separately. For example, the non-specific energy terms of Equations (5) and (6) are calculated by optimizing the Z-score with respect to the each of energy terms. The resulting terms are proportional to the number of interactions in which each domain/peptide is involved.

Scoring and cross-validation

We use two scores to assess the model performance: bound rank and ROC AUC. As shown in Figure 1C, to calculate the bound rank, the domain–peptide pairs are ordered by the predicted energy. Bound rank is then calculated by taking the median rank of the bound domain–peptide pairs. We report the bound rank score as a ratio of the median bound rank to the number of data points in the test set, which varies depending on cross-validation technique. The best possible bound rank score is $(n_{\mathrm{bound}} + 1)/2$, where n_{bound} is the number of bound domain–peptide pairs. The expected bound rank score for a random model is $(n_{\mathrm{total}} + 1)/2$, where n_{total} is the total number of domain–peptide pairs.

Receiver operating characteristic (ROC) curves are constructed by plotting the true positive rate (TPR) versus the false positive rate (FPR), given all possible energy thresholds. For a given threshold, all domain–peptide pairs below the threshold are predicted to be bound, and all pairs above are predicted to be unbound. The TPR and FPR are then given by:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

Here TP are true positives—domain–peptide pairs both predicted and experimentally found to be bound, TN are true negatives—domain–peptide pairs both predicted and experimentally found to be unbound, FP are false positives—domain–peptide pairs predicted to be bound and experimentally found to be unbound, and FN are false negatives—domain–peptide pairs predicted to be unbound and experimentally found to be bound. The ROC area under the curve (AUC) is calculated by estimating the area under this curve using the trapezoidal rule.

As with all cross-validation techniques, the two methods we use divide the data into two sets: the training set, from which the model parameters are inferred, and the test set, on which the model is applied and tested. The stratified 10-fold cross-validation is done by dividing the data into 10 equal parts with the same proportion of bound and unbound pairs as the total data set. The training set contains 9 of the 10 parts, and the test set is the remaining part. The method uses each tenth as the test set in turn, and averages the result. For some of our tests, we repeat this method 50 times. Other cross-validation tests (3-, 5- and 7-fold) are done accordingly.

To carry out the leave-one-group-out (LOGO) cross-validation, all the domain–peptide pairs involving the given domain or peptide were excluded from the training set and then used as the test set. Neither bound, nor unbound peptides of the excluded domain were excluded during the LOGO cross validation.

RESULTS

Development of a physically-based energy model

The goal of our study is to create a physically based model of the interaction energy between a class of PRMs and their peptides and to use this model to predict whether or not PRM–peptide pairs interact. Here, we use a coarse-grained amino acid-based approach to describe the energy of interaction E. We choose this level of detail because it is easy to transfer to new PRM–peptide pairs and, unlike atomistic potentials, do not require detailed structural information. This level of description also allows us to take into account interactions between residues that are not in direct contact, but that are rather mediated by other structural changes. For each domain–peptide pair, we define E as:

$$E = \sum_i \sum_j \Delta_{ij} U(a(i),a(j)) \qquad \mathbf{1}$$

Here Δ is the interaction map, a matrix in which $\Delta_{ij} = 1$ if amino acid i in the domain interacts (not necessarily directly) with amino acid j in the peptide and $\Delta_{ij} = 0$ otherwise. In order to use the same map for all domain–peptide pairs, we align the PRMs and the peptides (separately), creating a common numbering system for each. U is an energy potential, where $U(x,y)$ is the energy of interaction between amino acids x and y, and $a(i)$ and $a(j)$ are the amino-acid identities at positions i and j of the domain and peptide, respectively. We sum over all positions i in the domain and j in the peptide. We also explore the possibility of using a form of the energy function in which each interacting pair, or contact, (i, j) has its own energy potential [see Equation (4)].

Since we use the same interaction map and potential for each domain–peptide pair, this form is amenable to our goal of being able to apply this potential to new PRMs and peptides. Working in this amino acid-centric manner may also allow us to gain insight into the basis of specificity, since the amino acid is the basic unit of change between proteins.

To rewrite Equation (1) in a more compact form, we let D be a 20×20 matrix

$$D(x,y) = \sum_i \sum_j \delta_{xa(i)} \delta_{ya(j)} \Delta_{ij}, \qquad \mathbf{2}$$

therefore $D(x,y)$ is a count of the number of interactions between an amino acid of type x in the domain and type y in the peptide. Here δ is the Kronecker delta function. Now, if we reshape the 20×20 D matrix into a 1×400 row vector d and reshape the 20×20 matrix U into a 400×1 column vector u, we can rewrite Equation (1) as:

$$E = d \cdot u \qquad \mathbf{3}$$

In order to achieve maximum discriminatory power between the energies of bound and unbound domain–peptide pairs, we aim to find the potential U that maximizes the energy gap between the bound and unbound pairs, while minimizing the variance of the two energy distributions (Figure 1A). The approach is inspired by a method using to derive folding potentials in protein folding field (43), in which the energy potential is designed to maximize the gap between the energy of the native structure and the mean energy of all other structural decoys, while minimizing the variance of energies of the decoys (Figure 1B). Here, the interaction map and energy potential describe intramolecular contacts, rather than intermolecular contacts, but are otherwise completely analogous. This method is similar to an earlier formulation by Goldstein *et al.* (44), and to an approach used to infer PSSMs for protein–DNA recognition that is related to the support vector machine methodology (45).

Another method from protein folding which we might have chosen to optimize U is a perceptron learning method that uses neural networks to solve a system of inequalities that ensure the native structure is lower in energy than all the decoys (46). However, we choose not to use this method because it has been shown to give similar results to the energy gap method (47) and is slightly harder to generalize to the case where one native structure is replaced by a set of bound pairs.

Given Equation (3) and an interaction map Δ, there is an *analytical* solution [similar to (44)] that provides the potential U that maximizes the energy gap between the bound and unbound domain–peptide pairs, explicitly using both positive and negative examples (Materials and Methods section).

Therefore, once the problem is written in this fashion, the challenge is to find the interaction map Δ that gives the optimal separation between energies of the bound and unbound pairs, a problem we explore after defining some performance metrics.

Performance metrics and cross-validation techniques

In order to assess the performance of our model, we use several metrics and cross-validation techniques. (For a more in-depth discussion, see Materials and Methods section). We focus on two metrics, based on their use in similar studies and their intuitive physical interpretation: the bound rank and the area under the receiver operating characteristic curve (ROC AUC).

The bound rank score assesses the clustering of bound pairs to the low end of the predicted energy distribution and is calculated by ordering all the domain–peptide pairs by increasing energy, E and then calculating the median rank of the bound pairs (Figure 1C). We report the bound rank score as a fraction of the median bound rank divided by the number of data points in the test set, which varies depending on the type of cross-validation used. The bound rank gives a sense of how far down an ordered list of energy predictions one should look to find bound domain–peptide pairs. The best possible bound rank score, which is obtained when all the bound pairs have lower energies than all the unbound pairs, is

$(n_{\mathrm{bound}} + 1)/2$, where n_{bound} is the number of bound domain–peptide pairs. The expected bound rank score for a random classifier is $(n_{\mathrm{total}} + 1)/2$, where n_{total} is the total number of domain–peptide pairs.

ROC curves are commonly used to assess how well a classifier categorizes two populations, e.g. in our case how well the predicted interaction energy E separates the bound and unbound pairs. In these problems there is usually a tradeoff between the true positive rate (TPR) and the false positive rate (FPR), and ROC curves summarize this tradeoff by plotting these rates against each other. For example, in this problem, we pick an energy threshold, and we predict that domain–peptide pairs with energies below the threshold are bound. If we raise this threshold we will increase our TPR, the fraction of bound pairs that are correctly classified, but we will also increase our FPR, the fraction of unbound pairs that are incorrectly classified.

The ROC AUC corresponds to the probability that the energy of a randomly chosen bound pair is less than a randomly chosen unbound pair and is calculated by integrating the area under a ROC curve (Figure 1D). It is commonly used to summarize a ROC curve in a single number. The ROC AUC for a random classifier is 0.50 and for a perfect classifier is 1.0.

In order to avoid over fitting and to assess the likely performance of the method on unseen examples, we use two types of cross-validation: 10-fold, in which nine-tenths of the data are used to fit the model, which is then applied to the remaining one-tenth of the data; and leave-one-group-out (LOGO), in which one domain or peptide is left out of the training process and then used as the test data set. We also use 3-, 5- and 7-fold cross-validations to further test how the size of the training set influences the results.

Interaction map optimization

As mentioned above, given an interaction map Δ, there is an analytical way to find the potential U that maximizes the energy gap. So the challenge is to find an interaction map that characterizes the generic binding of SH2 domains to pY peptides. We use three methods—an information-based method, a structure-based method and a hybrid method. While most structure-based approaches use contacts present in the native structure of the complex, our hybrid approach attempts to refine the set of relevant interactions using information-based approaches [similar to correlated mutations, e.g. (48)]. This refinement aims to (i) eliminate the direct interactions that are not contributing to specificity and (ii) identify indirect interactions mediated by structural rearrangements. Some aspects of the hybrid approach are similar in spirit to approaches used by Ferraro *et al.* (18).

In each interaction map construction method, each potential contact—a pair of positions in the aligned collection of SH2 domains and peptides—is ranked by some criteria aimed at assessing its role in recognition. Then we construct interaction maps by taking the top n_{contacts} ranked contacts, where n_{contacts} can be varied, and setting the corresponding elements of Δ to 1. Briefly, the information-based method ranks potential contacts using a normalized measure of the divergence between the amino-acid composition of bound and unbound domain–peptide pairs. The structure-based method ranks the contacts by the number of experimentally determined SH2–peptide structures in which the contact is within some distance cutoff. The hybrid method uses the structure-based method to select potential position pairs and ranks them using the information-based criteria.

In order to find the optimal interaction map, we compare the 10-fold cross-validated ROC AUC and bound rank scores for interaction maps constructed using the three methods and different numbers of contacts (Supplementary Figure S1). The best-performing ('optimal') interaction map is a hybrid interaction map with 10 contacts, using a distance cutoff of 5.5 Å, with ROC AUC of 0.87 (bound rank = 31/325). The information-based technique performs quite well, plateauing at a ROC AUC of 0.82, with five contacts (bound rank = 37/325). Both of the ROC AUC scores and the bound rank scores are well above the scores expected for random classifiers (0.50 and 163/325, respectively). Other cross-validations (3-, 5- and 7-fold) gave similar results on the hybrid interaction map (AUC: 0.85, 0.86, 0.87).

To assess the role of low-resolution NMR structures, we constructed structure-based and hybrid contact maps by excluding NMR structures from consideration (see Methods section). The results for the structure-based map were identical whether or not NMR structures were included and for the hybrid contact map, the results were comparable (cross-validated ROC AUC of 0.84 without the NMR structures, versus 0.87 with the NMR structures).

Physics-based energy potential can predict SH2–peptide interactions

Figure 2A compares the performance of the best information-based map, the optimal hybrid map and a 'standard' energy potential. The ROC curves unambiguously show the energy predictions from the hybrid and information-based interaction maps far outperform a standard energy potential used in protein folding (49). The derived potential is indeed quite different from potentials used for protein folding (43,49), as is evident in Supplementary Figure S4. The failure of the standard folding potential shows that SH2–peptide recognition is based on a very specific set of amino acids interactions and the important role of electrostatic interactions (12,50), that are different from hydrophobic interactions central to protein folding.

In Figure 2B, we plot the energy of bound and unbound peptide–domain pairs, sorted by domain, and it shows that the bound pairs tend to have energies on the low end of the energy spectrum.

Using the hybrid interaction map, we can achieve a TPR of 0.90, predicting 179 of 198 bound pairs correctly, with a FPR of 0.06, predicting 190 of 3057 unbound pairs incorrectly. In other words, at this level about half of the pairs that are identified to be 'bound' actually are. If we use a lower energy threshold, 165 of 198 bound pairs are

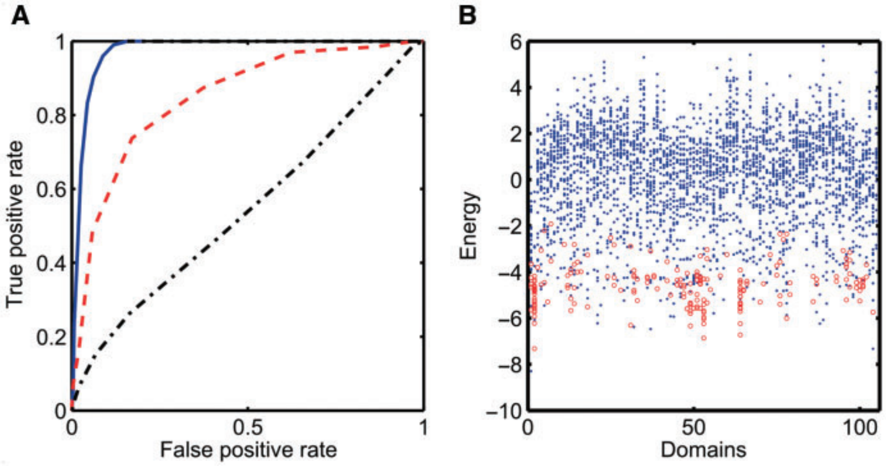

Figure 2. Performance of the basic energy model for predicting domain–peptide interactions. (**A**) Here we compare the ROC curves for the optimal hybrid interaction map (solid blue line), the best information-based interaction map (dashed red line) and a negative control (black dot-dash line), which is a standard energy potential of amino-acid interactions used for protein folding. The hybrid interaction map performs the best, with the information-based map close behind. When applying the control potential (49) to the hybrid interaction map, the ROC AUC is 0.54, which is close to the score of random predictors, 0.50. (**B**) This diagram shows the energy of bound pairs (open red circles) as compared to unbound pairs (filled blue dots), organized into columns by domain. The energy scale is arbitrary. There is a clear trend for the bound pairs to be on the low end of the energy spectrum.

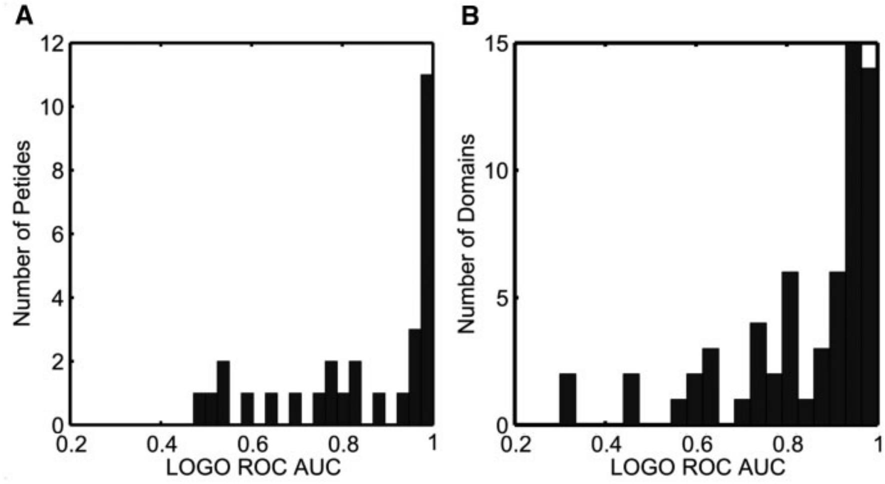

Figure 3. Performance of the basic model on unseen domains and peptides. In order to assess the expected performance of the energy model on peptides or domains not used in the construction process, we perform LOGO cross-validation. To do this, we exclude all the pairs including a particular domain or peptide from the training set used to create the potential and then used the derived potential to predict the interaction energies of the excluded pairs. In (**A**), we plot the ROC AUC scores for excluded peptides and in (**B**) we plot the scores for excluded domains. In both cases, the average ROC AUC is very high, 0.84, indicating that the energy function transfers well to new domains and peptides.

predicted correctly and 135 of 3057 unbound pairs are predicted incorrectly.

Using the optimal interaction map, we carry out LOGO cross-validation to see how the method will fare on unseen domains and peptides. The results are shown in Figure 3, and are quite impressive, with mean ROC AUC values of 0.84 for both unseen domains and peptides. For a number of domains and peptides, the ROC AUC is 1, indicating that the method perfectly separates bound and unbound domain–peptide pairs. (See Supplementary Table S1 for a list of domains and peptides with their ROC AUCs.) Interestingly, the performance of the model on an unseen domain seems to be uncorrelated to the sequence identity of the test domain to its nearest sequence neighbor in the training set when using the whole domain sequence or just the sequences at the contact positions, $r = 0.19$ and 0.30, respectively (Supplementary Figure S2).

To further test the model, we applied the method to data from Songyang *et al.* (51). In this study, the binding preferences of a few SH2 domains for a synthetic library of peptides were assessed using successive rounds of affinity purification. This additional data source allows us to test how well the model does on a set of peptides that are completely unrelated to the ErbB-derived peptides and

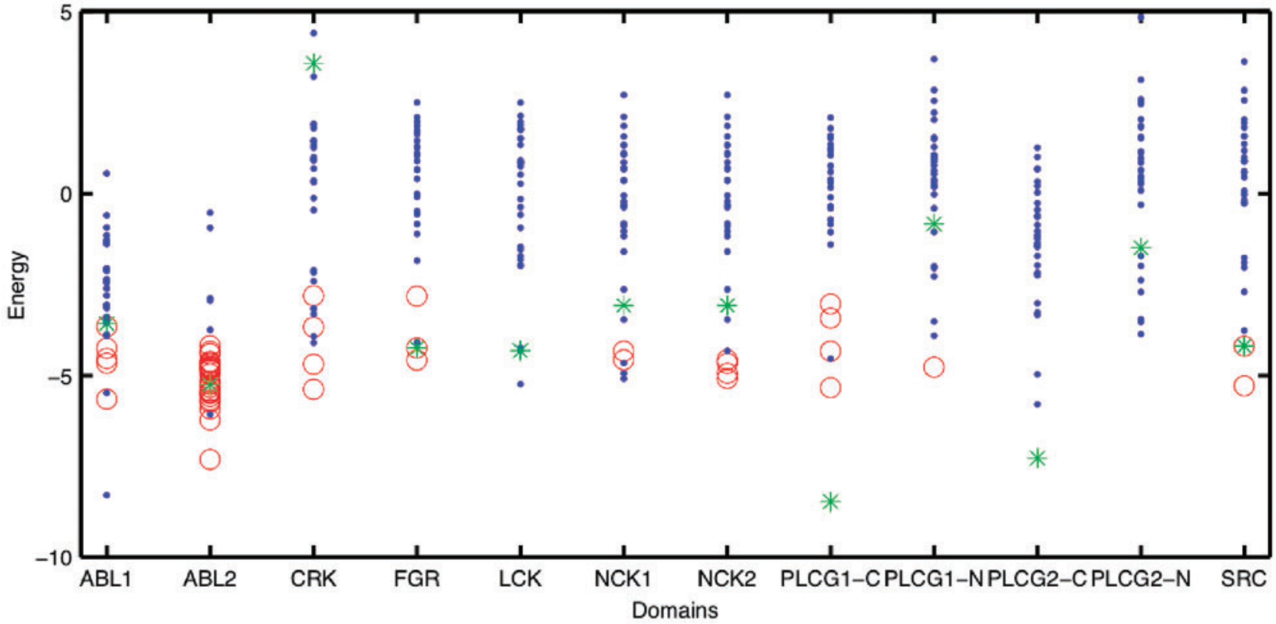

Figure 4. Performance of the basic energy model on new peptides. To test the method on a set of peptides unrelated to the ErbB receptor-derived peptides, we use data from the Songyang *et al.* study, which measured the binding specificity of a few SH2 domains against a synthetic peptide library. Here, we compare predicted energies of domain–peptide interactions for peptides in the Jones *et al.* (2) data set with the preferred peptide found by Songyang *et al.* (51). As in Figure 2, bound domain–peptide pairs are red circles, unbound pairs are blue dots and the preferred peptide for each domain, according to the Songyang *et al.* (51) study, is marked by a green star. For most domains, the preferred peptide (51) has a very low energy, comparable to that of bound peptides in the Jones, et al. data set. This clearly demonstrates that our method is useful for predicting binding specificity to peptides not used in the method construction.

interactions that were measured using a different experimental technique. In Figure 4, we plot the predicted interaction energies for the domains that overlap between the two data sources, for both the ErbB-derived peptides and the preferred peptide, according to Songyang *et al.* (51). This allows us to compare the predicted energies of the preferred peptide and bound ErbB-derived peptides. For the majority of domains, the performance of the method is roughly equivalent on the ErbB-derived and preferred peptides, indicating that there is not a strong bias to make better predictions for the ErbB-derived peptides used in the training set.

Position-specific energy potentials over-fit the data

In the results above, a common energy potential is used for all positions in the interaction map. But given the different structural and chemical contexts of each position, it may be more realistic to describe each position using a separate energy potential. The form of such an energy function is similar to Equation (1), but now the number of parameters in the energy potential is $20^2 \cdot n_{contacts} = 20^2 \cdot 10$—10-fold larger than the example above. Supplementary Figure S3 shows the results of using a mixture of position-specific and common energy potentials:

$$E = \alpha \cdot \sum_i \sum_j \Delta_{ij} U(a(i),a(j)) + (1-\alpha) \sum_i \sum_j \Delta_{ij} U_{ij}(a(i),a(j)) \qquad 4$$

Variables are defined as before; with the *ij* subscript indicating that the energy matrix U_{ij} is specific is to position-pair *ij* and α being a scalar weight. While the use of a position-specific energy map gives excellent results when the entire data set is used (ROC AUC = 0.99), it is not robust to 10-fold cross-validation (ROC AUC = 0.76), indicating that we do not have enough data to infer this increased number of parameters without over-fitting.

Non-specific energy is informative

When examining the data set, it is immediately obvious that the bound pairs are very unevenly distributed over the different domains and peptides—the number of domains or peptides bound by a peptide or domain varies widely. In light of this observation, we try adding another term to the energy function—the non-specific domain (*d*) or peptide (*p*) energy of interaction.

$$E = \alpha \cdot \sum_i \sum_j \Delta_{ij} U(a(i),a(j)) + (1-\alpha)p \qquad 5$$

$$E = \alpha \cdot \sum_i \sum_j \Delta_{ij} U(a(i),a(j)) + (1-\alpha)d \qquad 6$$

Figure 5 shows performance is optimal when α is 0.90 for peptides and 0.80 for domains indicating that a modest contribution from non-specific energy terms gives better predictive performance. However, these non-specific energy terms can only be derived for domains and peptides for which we have data. This means that we can use these terms if we are making predictions for new peptides with a

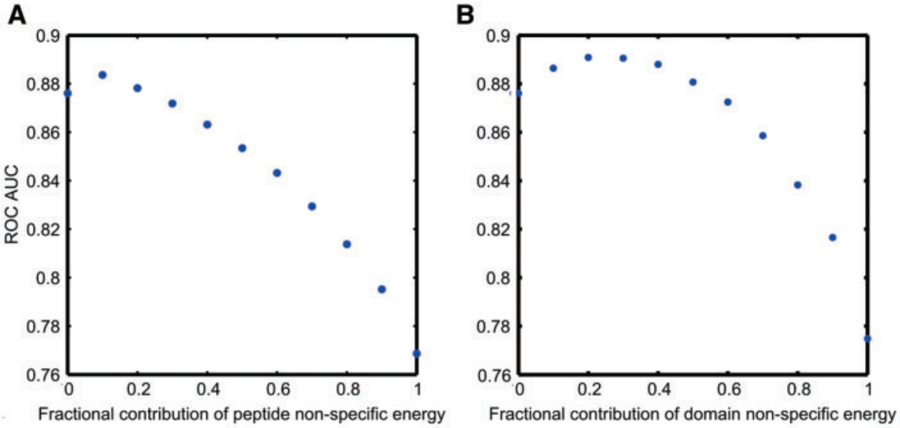

Figure 5. Performance of the basic energy model with a non-specific interaction energy term. Here, we plot the ROC AUC of energy models that include a non-specific energy term for peptides (A) and domains (B). The x-axis corresponds to $1 - \alpha$ in Equations (5) and (6)—α is the weight applied to the basic energy model and $1 - \alpha$ is the weight applied to the non-specific energy term. Here we can see a modest contribution of either domain or peptide non-specific energy terms increases the ROC AUC, and the domain non-specific energy term is more informative than the peptide term.

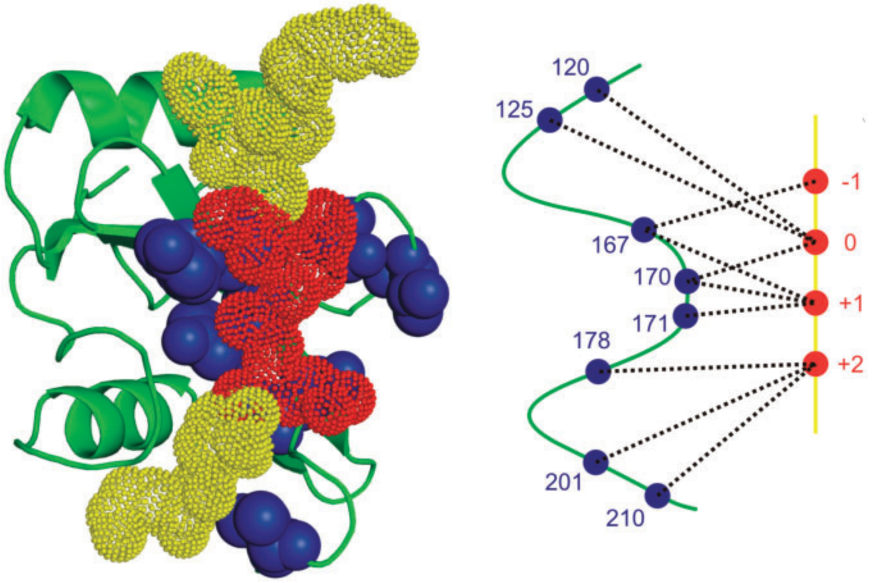

Figure 6. An illustration of the optimal hybrid interaction map. This figure shows a PDB structure of a sample SH2 domain (60), with the domain in a green ribbon depiction and the peptide in a yellow space-filling mesh. The figure was made with MacPyMOL (61). The interacting amino acids are highlighted in space-filling blue (domain) and red (peptide). The cartoon on the right shows the interaction map schematically, where the numbers on the left correspond to positions in the SH2 domain alignment and the numbers on the right correspond to the peptide positions, where position 0 is the location of the phosphorylated tyrosine, and the positions above and below are towards the N- and C-termini, respectively.

characterized domain or new domains with a tested peptide, but these non-specific terms cannot be used to predict interactions between domains and peptides that are both uncharacterized. However, the main (left) term of the energy function, which describes pair-wise interactions, can be applied in any of these cases, even when both the domain and the peptide are uncharacterized.

Interaction map includes unexpected contacts

The results above, Figure 3 in particular, indicate that an amino-acid based potential with a common interaction map can predict the binding of SH2 domains to pY peptides, even when the domain or peptide was not used in the construction of the potential. Now, to understand the origin of specificity, we inspect the optimal interaction map, shown in Figure 6. To review, this map is constructed using the hybrid method, which selects potential contacts by searching for those that are physically close in at least one experimentally determined domain–peptide structure and then ranks them by an information-based criterion. The information-based criterion finds contacts that have different amino-acid composition in the bound and unbound pairs.

While about half of the positions in our interaction map overlap with those identified in a study of SH2 domain selectivity (51), there are two unexpected features of this interaction map. First, there are three contacts with the pY position, numbered as position 0, which is unexpected. Since all peptides have a phosphorylated tyrosine in this position, it cannot be responsible for determining any difference in domain specificity. Therefore, the amino acids identities at the domain positions must specify some sort of general 'stickiness', which is derived separately but is related to the non-specific energy terms. For example, only three of the six most highly bound domains (the N-terminal domains of PIK3R1, PIK3R2 and PIK3R3) contain an alanine residue at position 120, interacting with the pY residue.

The second surprising observation is the lack of contacts with the +3 position of the peptide, which has been previously identified as a mediator of specificity (52). Analysis revealed that the most informative +3 domain contact positions overlap with the +2 positions, but the +2 positions were more informative, which is why they are included in the interaction map. This observation may reflect the choice of peptides included in the data set—they have a more diverse set of +2 than +3 positions—but is nonetheless interesting, since the peptide set is derived from natural peptide sequences and may be a better representative of what they SH2 domains are exposed to in real cell than the random peptides used in phage display experiments or the small set of peptides available in experimentally determined structures of SH2 domain–peptide complexes.

DISCUSSION

One of the significant features of domain–peptide interactions is that large, highly conserved families of domains interact specifically with sets of similar peptides. How is this specificity achieved? The amino-acid level is the natural level on which to study this problem: how do changes in amino-acid sequence alter the specificity of a particular PRM? With this in mind, we created model of domain–peptide interaction energies using an amino acid-based potential. Two major features of our basic model are that (i) it can be applied to SH2 domains or pY peptides that are not used in the model construction and (ii) it is physically interpretable.

It was not obvious from the outset that this level of detail was appropriate to describe the problem, as isotropic contact potentials have been shown to be insufficient for protein folding (43,53), i.e. they do not provide sufficient specificity to identify a native protein structure from an astronomical number of alternatives. Our results show, this energy model is surprisingly effective at predicting domain–peptide interactions. This success may be in part due to the relatively small set of possible peptides that contain a central tyrosine residue, as compared to the large set of possible protein structures.

The results of the LOGO cross-validation show that the model is successful at predicting the interactions involving a new domain or peptide. This is useful in predicting interactions that could not be measured, e.g. those involving domains that are difficult to express, or have not been measured, e.g. interactions with mouse SH2 domains. The LOGO cross-validation performance actually gave insight into experimental considerations, as well. Of the ten domains with the worst LOGO ROC AUC scores, five were difficult to express, and domains that were hard to express generally had below average LOGO ROC AUC scores [A. Gordus, personal communication].

Additionally, a more complicated position-specific energy potential, while excellent at fitting the entire data set, fares poorly under cross-validation, indicating that the data set is not large enough to specify the large number of parameters in the position-specific model. On the other hand, non-specific energy terms, which indicate the general 'stickiness' of the domains and peptides, improve the predictive capability of the model. These results may inform fields of computational biology aimed at predicting protein–protein and protein–DNA interactions, e.g. (38,54–56).

Insights from the interaction map

To make the model generally applicable, we used a single interaction map to describe all SH2 domain–pY peptide interactions. Based on the accuracy of the binding predictions, this simplification seems to work for the majority of cases. This result is somewhat unexpected, especially since others have described several binding modes for SH2–peptide interactions (57).

We tried various interaction map construction strategies based on structural data and information-based techniques. From these efforts come two interesting observations. First, the information-based techniques are quite effective, implying that sequence information is sufficient to build an energy-based classifier of protein–peptide interactions. This is useful for a variety of reasons, e.g. for some domain families, structural data may be sparse or unreliable and processing structural data is labor-intensive. Second, the hybrid technique that combines structural and information-based data is most effective, which is unsurprising since it supplements structural information, which may average over a variety of binding modes, with the information contained in the sequence data. Significant role of information-based contacts, which do not correspond to direct protein-peptide interactions, suggests the importance of the indirect readout in protein-peptide recognition.

By studying the optimal interaction map, we learn two things. First, contacts with the pY position specify a general 'stickiness' of the domain that is informative for predictions. Second, the contacts with the +2 position and +3 position are redundant, but in a peptide population derived from actual protein sequence, the +2 position is more informative. Because we used a different alignment of the SH2 domains, it is difficult to directly compare the interaction map to previous studies. However, an informal comparison shows that there is a significant overlap between the positions in our interaction map and those identified in a seminal study of SH2 domain selectivity (51).

Comparison to previous approaches

Our approach has several advantages over other techniques used to dissect the specificity of domain–peptide interactions and to predict binding. In comparison to traditional structural biology methods, which generally rely on manual, individual analysis of complexes, our method allows us to combine structural data from a number of similar complexes with large-scale proteomic data. Moreover, our method incorporates negative data for domain–peptides pairs that do not interact, as well as positive data from those that do interact. In contrast to some computational structural biology approaches [e.g. (17)], our method does not provide true quantitative predictions of interaction energies, but it is able to provide sufficiently accurate interaction energy rankings to correctly predict which domain–peptide pairs are bound. By using an interaction map in place of direct amino-acid contacts, our method finds amino-acid positions that are directly and indirectly involved in specific recognition and can be applied to members of a PRM family for which there is no structural data.

In comparison to other bioinformatics-based techniques, our approach gives results that are physically interpretable: an interaction map that specifies the positions that are important for recognition and an amino-acid-based energy potential (Supplementary Figure S4). It is also applicable to domains that are not present in the training set, a feature not available in previous studies (23,25).

Future directions

The success of this method opens up a number of future research directions. Most directly, the interaction map and energy potential can be used to scan databases of physiologically phosphorylated peptides [e.g. Phospho.ELM (58)] for potential SH2 interaction partners or to predict likely peptide partners for SH2 domains not included in the study. It would also be interesting to use this approach to predict ‘optimal’ peptides for each SH2 domain and to experimentally verify that these domain–peptide pairs actually interact. This approach is also easily applicable to any PRM family with a large-scale interaction data set, e.g. PTB domains (2) and PDZ domains (59). It would be interesting to attempt an adaptation of this method for protein–DNA interactions, as well.

One obvious drawback of this study is that is does not use the quantitative data available from the SH2 protein microarray experiments (2). As a result, computed scores can be used to predict only binding peptide–SH2 pairs, but not the binding energy. As expected, computed scores do not correlate with measured binding energies (Supplementary Figure S5).

Considering there are only ~200 positive data points in the data set and the comparable number of parameters in the model, modeling the strength of these interactions quantitatively will be challenging and may call for a different approach. In fact, negative data points that do not have an associated quantitative measurement (i.e. peptide–SH2 pairs that do not bind each other) are equally valuable, as they are present in high abundance in the whole-genome dataset (3057 negative versus 190 positive ones).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Grigory Kolesov, Andrew Gordus and Gavin MacBeath for useful discussions and Nickolay Khazanov for useful comments on the manuscript. Z.W. is a Howard Hughes Medical Institute Predoctoral Fellow. L.M. is supported by the NIH-funded National Center for Biomedical Computing (i2b2) Informatics for Integrating Biology & the Bedside. Part of this work took place at the Aspen Center for Physics, Aspen, Colorado.

FUNDING

Funding for open access charge: National Institutes of Health U54LM008748 National Center for Biomedical Computing (i2b2).

Conflict of interest statement. None declared.

REFERENCES

1. Yaffe,M.B. (2002) Phosphotyrosine-binding domains in signal transduction. *Nat. Rev. Mol. Cell Biol.*, **3**, 177–186.
2. Jones,R.B., Gordus,A., Krall,J.A. and MacBeath,G. (2006) A quantitative protein interaction network for the ErbB receptors using protein microarrays. *Nature*, **439**, 168–174.
3. Brannetti,B., Via,A., Cestra,G., Cesareni,G. and Helmer-Citterich,M. (2000) SH3-SPOT: an algorithm to predict preferred ligands to different members of the SH3 gene family. *J. Mol. Biol.*, **298**, 313–328.
4. Grigoryan,G. and Keating,A.E. (2006) Structure-based prediction of bZIP partnering specificity. *J. Mol. Biol.*, **355**, 1125–1142.
5. Hou,T., Chen,K., McLaughlin,W.A., Lu,B. and Wang,W. (2006) Computational analysis and prediction of the binding motif and protein interacting partners of the Abl SH3 domain. *PLoS Comput. Biol.*, **2**, e1.
6. Schleinkofer,K., Wiedemann,U., Otte,L., Wang,T., Krause,G., Oschkinat,H. and Wade,R.C. (2004) Comparative structural and energetic analysis of WW domain-peptide interactions. *J. Mol. Biol.*, **344**, 865–881.
7. Wollacott,A.M. and Desjarlais,J.R. (2001) Virtual interaction profiles of proteins. *J. Mol. Biol.*, **313**, 317–342.
8. Bergamin,E., Wu,J. and Hubbard,S.R. (2006) Structural basis for phosphotyrosine recognition by suppressor of cytokine signaling-3. *Structure*, **14**, 1285–1292.
9. Frese,S., Schubert,W.D., Findeis,A.C., Marquardt,T., Roske,Y.S., Stradal,T.E. and Heinz,D.W. (2006) The phosphotyrosine peptide binding specificity of Nck1 and Nck2 Src homology 2 domains. *J. Biol. Chem.*, **281**, 18236–18245.
10. Hu,J. and Hubbard,S.R. (2006) Structural basis for phosphotyrosine recognition by the Src homology-2 domains of the adapter proteins SH2-B and APS. *J. Mol. Biol.*, **361**, 69–79.
11. Waksman,G., Shoelson,S.E., Pant,N., Cowburn,D. and Kuriyan,J. (1993) Binding of a high affinity phosphotyrosyl peptide to the Src SH2 domain: crystal structures of the complexed and peptide-free forms. *Cell*, **72**, 779–790.
12. Sheinerman,F.B., Al-Lazikani,B. and Honig,B. (2003) Sequence, structure and energetic determinants of phosphopeptide selectivity of SH2 domains. *J. Mol. Biol.*, **334**, 823–841.
13. Henriques,D.A., Ladbury,J.E. and Jackson,R.M. (2000) Comparison of binding energies of SrcSH2-phosphotyrosyl peptides with structure-based prediction using surface area based empirical parameterization. *Protein Sci.*, **9**, 1975–1985.

14. Lee,J.K., Moon,T., Chi,M.W., Song,J.S., Choi,Y.S. and Yoon,C.N. (2003) An investigation of phosphopeptide binding to SH2 domain. *Biochem. Biophys. Res. Commun.*, **306**, 225–230.
15. McLaughlin,W.A., Hou,T. and Wang,W. (2006) Prediction of binding sites of peptide recognition domains: an application on Grb2 and SAP SH2 domains. *J. Mol. Biol.*, **357**, 1322–1334.
16. Suenaga,A., Hatakeyama,M., Ichikawa,M., Yu,X., Futatsugi,N., Narumi,T., Fukui,K., Terada,T., Taiji,M., Shirouzu,M. *et al.* (2003) Molecular dynamics, free energy, and SPR analyses of the interactions between the SH2 domain of Grb2 and ErbB phosphotyrosyl peptides. *Biochemistry*, **42**, 5195–5200.
17. Sanchez,I.E., Beltrao,P., Stricher,F., Schymkowitz,J., Ferkinghoff-Borg,J., Rousseau,F. and Serrano,L. (2008) Genome-wide prediction of SH2 domain targets using structural information and the FoldX algorithm. *PLoS Comput. Biol.*, **4**, e1000052.
18. Ferraro,E., Via,A., Ausiello,G. and Helmer-Citterich,M. (2006) A novel structure-based encoding for machine-learning applied to the inference of SH3 domain specificity. *Bioinformatics*, **22**, 2333–2339.
19. Zhang,L., Shao,C., Zheng,D. and Gao,Y. (2006) An integrated machine learning system to computationally screen protein databases for protein binding peptide ligands. *Mol. Cell Proteomics*, **5**, 1224–1232.
20. Chen,J.R., Chang,B.H., Allen,J.E., Stiffler,M.A. and MacBeath,G. (2008) Predicting PDZ domain-peptide interactions from primary sequences. *Nat. Biotechnol.*, **26**, 1041–1045.
21. Kolesov,G. and Mirny,L.A. (2009) In McDermott,J., Samudrala,R., Bumgarner,R., Montgomery,K. and Ireton,R. (eds), *Computational Systems Biology*. Vol. 541, Springer, New York.
22. Moult,J. (2005) A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol.*, **15**, 285–289.
23. Lehrach,W.P., Husmeier,D. and Williams,C.K. (2006) A regularized discriminative model for the prediction of protein-peptide interactions. *Bioinformatics*, **22**, 532–540.
24. Obenauer,J.C., Cantley,L.C. and Yaffe,M.B. (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.*, **31**, 3635–3641.
25. Reiss,D.J. and Schwikowski,B. (2004) Predicting protein-peptide interactions via a network-based motif sampler. *Bioinformatics*, **20(Suppl 1)**, i274–i282.
26. Wiedemann,U., Boisguerin,P., Leben,R., Leitner,D., Krause,G., Moelling,K., Volkmer-Engert,R. and Oschkinat,H. (2004) Quantification of PDZ domain specificity, prediction of ligand affinity and rational design of super-binding peptides. *J. Mol. Biol.*, **343**, 703–718.
27. Deeds,E.J., Ashenberg,O. and Shakhnovich,E.I. (2006) A simple physical model for scaling in protein-protein interaction networks. *Proc. Natl Acad. Sci. USA*, **103**, 311–316.
28. Altuvia,Y. and Margalit,H. (2004) A structure-based approach for prediction of MHC-binding peptides. *Methods*, **34**, 454–459.
29. Altuvia,Y., Schueler,O. and Margalit,H. (1995) Ranking potential binding peptides to MHC molecules by a computational threading approach. *J. Mol. Biol.*, **249**, 244–250.
30. Lundegaard,C., Lund,O., Kesmir,C., Brunak,S. and Nielsen,M. (2007) Modeling the adaptive immune system: predictions and simulations. *Bioinformatics*, **23**, 3265–3275.
31. Donald,J.E., Chen,W.W. and Shakhnovich,E.I. (2007) Energetics of protein-DNA interactions. *Nucleic Acids Res.*, **35**, 1039–1047.
32. Endres,R.G., Schulthess,T.C. and Wingreen,N.S. (2004) Toward an atomistic model for predicting transcription-factor binding sites. *Proteins*, **57**, 262–268.
33. Havranek,J.J., Duarte,C.M. and Baker,D. (2004) A simple physical model for the prediction and design of protein-DNA interactions. *J. Mol. Biol.*, **344**, 59–70.
34. Morozov,A.V., Havranek,J.J., Baker,D. and Siggia,E.D. (2005) Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Res.*, **33**, 5781–5798.
35. Liu,L.A. and Bader,J.S. (2007) Ab initio prediction of transcription factor binding sites. *Pac. Symp. Biocomput.*, 484–495.
36. Mandel-Gutfreund,Y. and Margalit,H. (1998) Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites. *Nucleic Acids Res.*, **26**, 2306–2312.
37. Kinney,J.B., Tkacik,G. and Callan,C.G. Jr. (2007) Precise physical models of protein-DNA interaction from high-throughput data. *Proc. Natl Acad. Sci. USA*, **104**, 501–506.
38. Kaplan,T., Friedman,N. and Margalit,H. (2005) Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Comput. Biol.*, **1**, e1.
39. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
40. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
41. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
42. Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374.
43. Mirny,L.A. and Shakhnovich,E.I. (1996) How to derive a protein folding potential? A new approach to an old problem. *J. Mol. Biol.*, **264**, 1164–1179.
44. Goldstein,R.A., Luthey-Schulten,Z.A. and Wolynes,P.G. (1992) Protein tertiary structure recognition using optimized Hamiltonians with local interactions. *Proc. Natl Acad. Sci. USA*, **89**, 9029–9033.
45. Djordjevic,M., Sengupta,A.M. and Shraiman,B.I. (2003) A biophysical approach to transcription factor binding site discovery. *Genome Res.*, **13**, 2381–2390.
46. Vendruscolo,M. and Domany,E. (1998) Pairwise contact potentials are unsuitable for protein folding. *J. Chem. Phys.*, **109**, 11101–11108.
47. Vendruscolo,M., Mirny,L.A., Shakhnovich,E.I. and Domany,E. (2000) Comparison of two optimization methods to derive energy parameters for protein folding: perceptron and Z score. *Proteins*, **41**, 192–201.
48. Pazos,F., Helmer-Citterich,M., Ausiello,G. and Valencia,A. (1997) Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.*, **271**, 511–523.
49. Miyazawa,S. and Jernigan,R.L. (1999) Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins*, **34**, 49–68.
50. Grucza,R.A., Bradshaw,J.M., Mitaxov,V. and Waksman,G. (2000) Role of electrostatic interactions in SH2 domain recognition: salt-dependence of tyrosyl-phosphorylated peptide binding to the tandem SH2 domain of the Syk kinase and the single SH2 domain of the Src kinase. *Biochemistry*, **39**, 10072–10081.
51. Songyang,Z., Shoelson,S.E., Chaudhuri,M., Gish,G., Pawson,T., Haser,W.G., King,F., Roberts,T., Ratnofsky,S., Lechleider,R.J. *et al.* (1993) SH2 domains recognize specific phosphopeptide sequences. *Cell*, **72**, 767–778.
52. Kuriyan,J. and Cowburn,D. (1997) Modular peptide recognition domains in eukaryotic signaling. *Ann. Rev. Biophy. Biomol. Struct.*, **26**, 259.
53. Vendruscolo,M. and Domany,E. (1998) Pairwise contact potentials are unsuitable for protein folding. *J. Chem. Phy.*, **109**, 11101–11108.
54. Bock,J.R. and Gough,D.A. (2001) Predicting protein— protein interactions from primary structure. *Bioinformatics*, **17**, 455–460.
55. Gomez,S.M., Noble,W.S. and Rzhetsky,A. (2003) Learning to predict protein-protein interactions from protein sequences. *Bioinformatics*, **19**, 1875–1881.
56. Li,X., Zhong,S. and Wong,W.H. (2005) Reliable prediction of transcription factor binding sites by phylogenetic verification. *Proc. Natl Acad. Sci. USA*, **102**, 16945–16950.
57. Liu,B.A., Jablonowski,K., Raina,M., Arce,M., Pawson,T. and Nash,P.D. (2006) The human and mouse complement of SH2 domain proteins-establishing the boundaries of phosphotyrosine signaling. *Mol. Cell*, **22**, 851–868.
58. Diella,F., Gould,C.M., Chica,C., Via,A. and Gibson,T.J. (2008) Phospho.ELM: a database of phosphorylation sites—update 2008. *Nucleic Acids Res.*, **36**, D240–D244.
59. Stiffler,M.A., Chen,J.R., Grantcharova,V.P., Lei,Y., Fuchs,D., Allen,J.E., Zaslavskaia,L.A. and MacBeath,G. (2007) PDZ domain

binding selectivity is optimized across the mouse proteome. *Science*, **317**, 364–369.

60. Poy,F., Yaffe,M.B., Sayos,J., Saxena,K., Morra,M., Sumegi,J., Cantley,L.C., Terhorst,C. and Eck,M.J. (1999) Crystal structures of the XLP protein SAP reveal a class of SH2 domains with extended, phosphotyrosine-independent sequence recognition. *Mol. Cell*, **4**, 555–561.
61. DeLano,W.L. (2007) MacPyMOL: a PyMOL-based molecular graphics application for MacOS X. DeLano Scientific LLC, Palo Alto, CA.