Identification of Specificity Determining Residues in Peptide Recognition Domains using an Information Theoretic Approach Applied to Large-Scale Binding Maps

Kevin Y. Yip[#1,2] , Lukas Utz[#3] , Simon Sitwell[3] , Xihao Hu[2] , Benjamin E. Turk[4] , Mark Gerstein[*1,5,6] , Philip M. Kim[*3,7,8,9]

[1]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA
[2]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong
[3]Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada
[4]Department of Pharmacology, Yale University, New Haven, CT, USA
[5]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA
[6]Department of Computer Science, Yale University, New Haven, CT, USA
[7]Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada
[8]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada
[9]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada

Email: Kevin Y. Yip - yuklap.yip@yale.edu; Lukas Utz - utz.lukas@gmail.com; Simon Sitwell - ssitwell@gmail.com; Xihao Hu - xhhu@cse.cuhk.edu.hk; Benjamin E. Turk - ben.turk@yale.edu; Mark Gerstein*- mark.gerstein@yale.edu; Philip M. Kim*- pm.kim@utoronto.ca;

[#]These authors contributed equally to this work

[*]Corresponding author

Abstract

Background: Peptide Recognition Domains (PRDs) are commonly found in signaling proteins. They mediate protein-protein interactions by recognizing and binding short motifs in their ligands. Although a great deal is known about PRDs and their interactions, prediction of PRD specificities remains largely an unsolved problem.

Results: We present a novel approach to identifying these *Specificity Determining Residues* (SDRs). Our algorithm generalizes earlier information theoretic approaches to coevolution analysis, to become applicable to this problem. It leverages the growing wealth of binding data between PRDs and large numbers of random peptides, and searches for PRD residues that exhibit strong evolutionary covariation with some positions of the statistical profiles of bound peptides. The calculations involve only information from sequences, and thus can be applied to PRDs without crystal structures. We applied the approach to PDZ, SH3 and kinase domains, and evaluated the results using both residue proximity in co-crystal structures and verified binding specificity maps from mutagenesis studies. Our predictions were found to be strongly correlated with the physical proximity of

residues, demonstrating the ability of our approach to detect physical interactions of the binding partners. Some high-scoring pairs were further confirmed to affect binding specificity using previous experimental results. Combining the covariation results also allowed us to predict binding profiles with higher reliability than two other methods that do not explicitly take residue covariation into account.

Conclusions: The general applicability of our approach to the three different domain families demonstrated in this paper suggests its potential in predicting binding targets and assisting the exploration of binding mechanisms.

Background
Protein Recognition Domains (PRDs) and Specificity Determining Residues (SDRs)

Peptide recognition domains are key elements on proteins with many roles in central signaling pathways [1]. PRDs are involved in many diseases and are a focus of drug discovery efforts [2]. Some viruses co-opt host PRDs via mimicry, emphasizing their relevance to human diseases [3–5]. Representative PRDs include PDZ, SH3 and kinase domains. These domains recognize short (usually around 7-10 amino acids long) motifs in their target proteins. Such motifs often lie in unstructured regions [6] and hence resemble peptides in their conformational flexibility. In particular, PDZ domains recognize hydrophobic C-terminal tails, SH3 domains recognize proline-rich motifs, and kinase domains bind several different classes of motifs; for instance, basophilic kinases bind arginine-rich motifs. As many PRDs are well behaved in solution, they are ideal model systems for studying protein-protein interactions and ligand binding specificity. New approaches, namely phage display [7], protein or peptide arrays [8–10] and oriented peptide libraries [11] have greatly facilitated the measurement of specificities and produced an hitherto unimaginable wealth of binding data. Despite these advances, the residues within the domains that determine their binding specificity have only partly been elucidated. Such Specificity Determining Residues (SDRs) [12,13] can be identified using dedicated experiments involving site-directed mutagenesis and subsequent measurement of specificity. For example, recent studies have directly tested effects of point mutations in kinases [14] and PDZ domains [7]. However, because of the size of the sequence space to be covered, exhaustive experimental search is infeasible. While co-crystal structures of PRDs with the bound ligand are often used to prioritize residues, identification of SDRs remains a complex problem: On the one hand, close proximity to the ligand does not necessarily implicate a residue in specificity determination. On the other hand, a residue that is far away from the ligand can also affect specificity due to secondary

2

effects on the binding site [13, 15]. Hence, computational methods are needed to select and prioritize the positions to be tested through mutagenesis. Meanwhile, the aforementioned wealth of specificity data offers ample resources to be computationally analyzed for information about SDRs.

Currently, there are two major classes of computational approaches for SDR identification: First, there are structure driven approaches, making use of physical properties from protein structures, such as the hydration site free energies displaced by ligand atoms upon binding [16]. The second class of approaches adopts a statistical approach, and identifies SDRs by looking for sites that are conserved across or within functional groups [17], or more conserved in orthologs than paralogs [12]. Our method also takes on a statistical approach, the systematic nature of which enables the potential identification of non-local interactions between residues that are significant for peptide binding. Due to the lack of large-scale binding data previously, most current statistical methods attempt to detect SDRs based only on conservation signals from multiple sequence alignments of the PRDs. These predictions are noisy as the conservation of a residue could be due to roles other than binding specificity determination. Hence, these approaches do not make optimal use of currently available data. Conversely, utilizing the information about the actual bound peptide profiles from the recent large-scale binding datasets has the potential to boost the power of new predictions. In this paper we demonstrate how these new datasets can be incorporated in the search for SDRs. Previously we explored this idea using a proof-of-principle correlation-based method as part of a larger experimental study [14]. Here, we develop a new approach that is based on a novel formalism of mutual information.

Traditional and generalized covariation methods

Our approach is based on the notion that SDRs would covary with the binding specificity. The concept of covariation has first been used to study internal residue contacts in proteins [18, 19]. It has subsequently been used in the identification of energetically coupled residues [20], coevolution of proteins with their interaction partners [21–25], protein fold [26], functionally or structurally coupled residues within a protein [27, 28] and protein fusion [29]. The basic idea is to look for two sites in a multiple sequence alignment (MSA), or a pair of MSAs when studying the interaction between two proteins, that exhibit coordinated changes. Such covariation could suggest functional or structural dependency between the two sites, as there exists evolutionary pressure against their independent evolution. In the context of SDR identification, the covariation between a residue in the PRD and a residue in the bound peptide could imply its role in mediating the binding.

3

Many computational methods have been proposed for identifying covarying sites, including correlation scores [18, 30], Statistical Coupling Analysis (SCA) [20, 31], likelihood methods [32, 33], information theoretic methods [27, 34], independence tests [35, 36] and Bayesian approaches [37]. However, these methods cannot be applied directly to the identification of SDRs, because in a typical protein-peptide binding dataset [13, 38], each PRD does not bind to only one single peptide sequence. Instead, a PRD is associated with a binding profile, often represented in the form of a position weight matrix (PWM). In other words, instead of having two MSAs as in a typical covariation study, SDR identification adopts a more general setting with one MSA of PRDs on the one hand, and a list of respective PWMs on the other hand (Figure 1). Furthermore, a multitude of techniques have been developed in recent studies to handle various issues in the application of the covariation methods [34, 39–42], such as uneven representation of sequences in the MSA and overly diverse sites. These techniques also need to be extended in order to be applied in the current more generalized setting. In this paper we present a novel method to determine SDRs using an information theoretic approach. For each site from the MSA of the PRDs and each site from the aligned peptide PWMs, the method produces a numeric covariation score, which can be used as an indicator of the likelihood that the two sites are involved in binding specificity. Aggregating the scores could also suggest which peptides a PRD is likely to bind. Our approach uses only information from sequences, and thus can be applied to PRDs without known crystal structures.

Methods
Entropy and mutual information

Our method for identifying SDRs is based on an information theoretic measure called mutual information, which has been used in detecting covarying residues in the traditional setting [27, 34]. We first give some general definitions here, and then discuss how they are used in the identification of SDRs in the next subsection. Given a discrete random variable X with a distribution $p(x)$ over the domain of X, $\mathcal{D}(X)$, the entropy of X is defined as

$$H(X) \quad = \quad - \sum_{x \in \mathcal{D}(\mathcal{X})} p(x) \log p(x), \tag{1}$$

where $\log 0$ is defined as 0, the asymptotic value of $\log(x)$ towards 0. In the general definition of entropy, the base of the logarithm can be set to any value. In analyzing protein sequences, where $\mathcal{D}(X)$ is the set of 20 amino acid residues, it is convenient to use 20 as the base so that the entropy value is always between 0
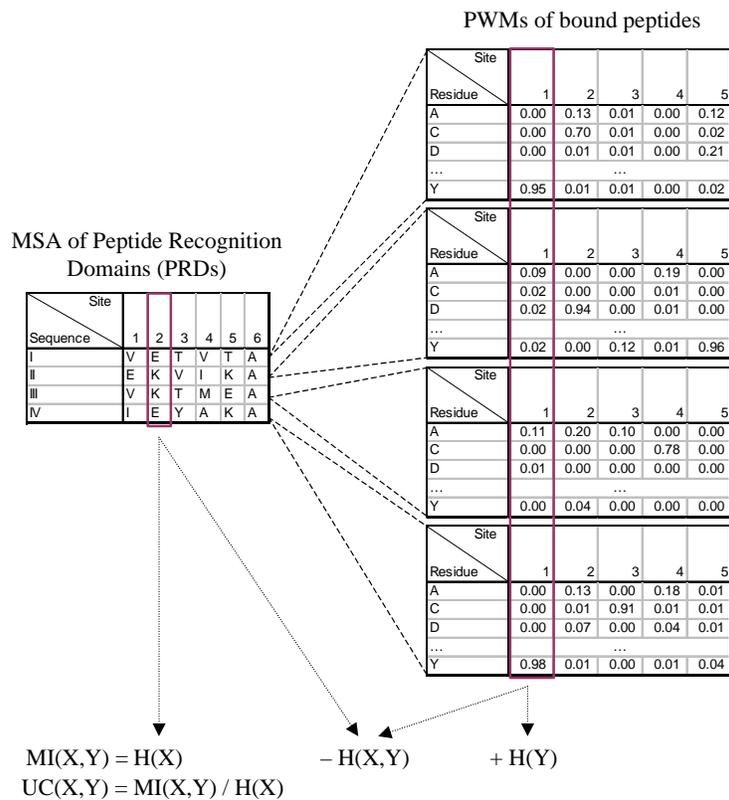
4

PWMs of bound peptides

MSA of Peptide Recognition Domains (PRDs)

Site / Sequence	1	2	3	4	5	6
I	V	E	T	V	T	A
II	E	K	V	I	K	A
III	V	K	T	M	E	A
IV	I	E	Y	A	K	A

Site / Residue	1	2	3	4	5
A	0.00	0.13	0.01	0.00	0.12
C	0.00	0.70	0.01	0.00	0.02
D	0.00	0.01	0.01	0.00	0.21
…			…		
Y	0.95	0.01	0.01	0.00	0.02

Site / Residue	1	2	3	4	5
A	0.09	0.00	0.00	0.19	0.00
C	0.02	0.00	0.00	0.01	0.00
D	0.02	0.94	0.00	0.01	0.00
…			…		
Y	0.02	0.00	0.12	0.01	0.96

Site / Residue	1	2	3	4	5
A	0.11	0.20	0.10	0.00	0.00
C	0.00	0.00	0.00	0.78	0.00
D	0.01	0.00	0.00	0.00	0.00
…			…		
Y	0.00	0.04	0.00	0.00	0.00

Site / Residue	1	2	3	4	5
A	0.00	0.13	0.00	0.18	0.01
C	0.00	0.01	0.91	0.01	0.01
D	0.00	0.07	0.00	0.04	0.01
…			…		
Y	0.98	0.01	0.00	0.01	0.04

$$MI(X,Y) = H(X) \qquad - H(X,Y) \qquad + H(Y)$$
$$UC(X,Y) = MI(X,Y) / H(X)$$

Figure 1: **An illustration of our method for finding specificity determining residues (SDRs).** The sequences of the peptide recognition domains (PRDs) are aligned to form a multiple sequence alignment (MSA). Each PRD has an associated position weight matrix (PWM) of the peptides that it binds, and the PWMs are also aligned. The entropy of each MSA site X and each PWM site Y is computed and combined to give the mutual information (MI) and uncertainty coefficient (UC) scores.

and 1 [43].

Similarly, if we have two random variables X and Y with a joint distribution $p(x, y)$, the joint entropy of them is defined as

$$H(X, Y) \quad = \quad - \sum_{x \in \mathcal{D}(X)} \sum_{y \in \mathcal{D}(Y)} p(x, y) \log p(x, y) \tag{2}$$

Based on the concepts of entropy and joint entropy, the mutual information between two random variables X and Y is defined as follows:

$$MI(X, Y) \quad = \quad H(X) + H(Y) - H(X, Y) \tag{3}$$

Adapting mutual information to the identification of SDRs

Suppose we are given an MSA A with n rows and m columns. Each row corresponds to a PRD sequence and each column corresponds to a site of the alignment. We use A_{ij} to denote the residue at site j of sequence i. The residues at a site can be viewed as a sample drawn from a distribution of residues specific to the site. Let A_j be a random variable for the residues at site j. Using Formula 1, we can calculate the entropy of A_j by replacing $p(x)$ with the sample distribution, $p_j(x)$, defined by the frequency of each residue at the site:

$$p_j(x) \quad = \quad \frac{\sum_{i=1}^{n} \mathbb{1}(A_{ij} = x)}{n} \tag{4}$$

where $\mathbb{1}$ is the indicator function with $\mathbb{1}(true) = 1$ and $\mathbb{1}(false) = 0$.

The entropy can be interpreted as the uncertainty of which residue we would encounter at the site if we randomly draw a sequence from the MSA, where uncertainty here is mathematically quantified by the number of bits needed to encode the information on average. A completely conserved site has an entropy of 0, and indeed there is no uncertainty as the residue being drawn must always be the same. A site with equal probability of all 20 residues has the maximum possible entropy of 1. In general, a more diverse site has a larger entropy.

Similarly, suppose we are given a set of n aligned PWMs W each with w sites. The i-th PWM represents the peptides bound by the i-th PRD in the MSA (Figure 1). Denote $W_{ik}(y)$ as the probability of residue y at the k-th site of the i-th PWM. Let W_k be a random variable for the residues at site k. Again, we can

calculate the entropy of W_k using Formula 1 by replacing $p(x)$ with the expected probability of y in the different PWMs, $p_k(y)$, assuming a uniform distribution of the observed sequences:

$$p_k(y) \quad = \quad \frac{\sum_{i=1}^{n} W_{ik}(y)}{n} \tag{5}$$

Now, we can compute the joint entropy of site j in the MSA and site k in the PWMs using Formula 2, based on the sample distribution of A_j and the probabilities $W_{ik}(y)$:

$$p_{j,k}(x, y) \quad = \quad \frac{\sum_{i=1}^{n} \mathbb{1}(A_{ij} = x) W_{ik}(y)}{n} \tag{6}$$

The joint entropy measures the uncertainty of which two residues we would encounter at MSA site j and PWM site k if we randomly draw a sequence from the MSA and get its corresponding PWM.

Finally, we can compute the mutual information between MSA site j and PWM site k using Formula 3. Since $H(X, Y)$ is the uncertainty of the two sites that persists even when we consider them together, subtracting it from $H(X) + H(Y)$ gives the uncertainty that is eliminated by considering the two sites together. In other words, mutual information measures the information shared by the two sites. A larger mutual information indicates a stronger dependency between them. This could indicate a functional or structural relationship between the two sites. For example, it could suggest that the two sites coevolve in the sense that when the residue at the MSA site is changed, binding strength is restored by having a corresponding change at the PWM site.

Handling uneven sequence representation

In many cases, the input MSA for studying residue covariation has an uneven representation of sequences from different clades. For example, it is common to have more sequences from model organisms or species that are better studied, and fewer sequences from other species. As a result, the MSA could contain many sequences that are highly similar, and few that are significantly different. Each of the highly similar sequences contributes little additional information, but still has an equal amount of influence on the calculation of mutual information under the assumption of a uniform distribution of the observed sequences. Consequently, they could mask the novel information from low abundance sequences.

To counteract this effect, we associate weights with the sequences so that the more unique ones receive higher weights. Statistically, it is equivalent to placing a prior distribution to the observed sequences if the

weights are normalized to take values between 0 and 1 and have a sum of 1. Here we assume there is an external procedure for determining the weights. For instance, one possible way is to construct a phylogenetic tree from the sequences in the MSA, and then recursively distribute the total weight to different branches of the tree, so that each sequence in the crowded branches will receive a smaller share of weights [44]. Suppose the procedure assigns α_i as the weight of sequence i, we can redefine $p_j(x)$, $p_k(y)$ and $p_{j,k}(x, y)$ as follows:

$$p_j(x) = \frac{\sum_{i=1}^{n} \mathbb{1}(A_{ij} = x)\alpha_i}{\sum_{i=1}^{n} \alpha_i} \tag{7}$$

$$p_k(y) = \frac{\sum_{i=1}^{n} \alpha_i W_{ik}(y)}{\sum_{i=1}^{n} \alpha_i} \tag{8}$$

$$p_{j,k}(x, y) = \frac{\sum_{i=1}^{n} \mathbb{1}(A_{ij} = x)\alpha_i W_{ik}(y)}{\sum_{i=1}^{n} \alpha_i} \tag{9}$$

Entropy, joint-entropy and mutual information can then be calculated using these new definitions of probabilities.

Handling uneven site conservation

A potential issue of the mutual information measure is that a pair of unrelated sites could have even higher mutual information than a pair of truly covarying sites if the unrelated sites are individually much more diverse than the covarying sites. This is illustrated by the hypothetical example shown in Table 1. For simplicity, suppose the sequences all have equal weights and the base of logarithms is 2. Sites 1 and 2 are truly covarying. When site 1 changes from the non-polar residue alanine in sequences I and II to the polar residue threonine in sequences III and IV, site 2 simultaneously changes from the non-polar residue valine to the polar residue tyrosine. The entropy of each of the two sites is 1 and the mutual information between them is also 1, the maximum possible value given the two individual entropy values. On the other hand, sites 3 and 4 are random, unrelated sites. The entropy values of them are 2 and 1.5, respectively, and their mutual information is 1.5, which is higher than the mutual information between sites 1 and 2 due to larger entropy values of sites 3 and 4.

To deal with this problem, various kinds of normalization have been proposed to penalize overly diverse sites [34]. We will focus on the uncertainty coefficient [45], which was found to be one of the best normalized mutual information scores in our preliminary study. For an MSA site A_j and a PWM site W_k, the uncertainty coefficient is defined as follows:

Table 1: **A hypothetical example MSA that illustrates the problem of uneven site conservation.** Sites 1 and 2 are truly covarying, while sites 3 and 4 are random and unrelated. However, as sites 3 and 4 have higher entropy values than sites 1 and 2, the resulting mutual information between sites 3 and 4 is also higher than that between sites 1 and 2 (see main text for the calculations).

Sequence / Site	1	2	3	4
I	A	V	A	C
II	A	V	C	D
III	T	Y	D	E
IV	T	Y	E	C

$$UC(A_j, W_k) \quad = \quad \frac{MI(A_j, W_k)}{H(A_j)} \tag{10}$$

We have also tried handling the problem using a statistical test. Specifically, we used mutual information as the test statistic to calculate how unlikely it is to get a mutual information at least as large as the observed one under the null hypothesis that the two sites are independent. The distribution of mutual information can be obtained by permuting the residues of a site, or by using a chi-square approximation. It was shown that when n is large, $(2 \ln 2n) MI(X, Y)$ tends to have a chi-square distribution with $(|\mathcal{D}(\mathcal{X})| - 1)^2 (|\mathcal{D}(\mathcal{Y})| - 1)^2$ degrees of freedom [46, 47]. It turns out that the results based on this statistical test were not better than using the simple normalization approach. We thus focus on the use of the uncertainty coefficient measure in handling uneven site conservation in the remaining of this paper.

Table 1 also demonstrates the tradeoff between the mutual information between two sites and their individual conservation, both of which are indicators of their functional importance. One may try to derive a measure that takes both into account, similar to what the Sequence Harmony method handles both the conservation and similarity of two groups of residues simultaneously, for the problem of identifying important residues that determine the functional differences of protein subfamilies [48]. We leave the derivation of a new covariation measure to a future study.

Predicting the PWMs of bound peptides

One important use of the covariation scores is to contribute towards predicting the PWMs of bound peptides. This can be done by aggregating the detected covariation signals. Suppose we are given a new PRD sequence (the $(n + 1)$-th sequence) of the MSA M without the corresponding PWM of its bound peptides. We would like to predict the PWM based on the $n + 1$ sequences in the MSA and the n known PWMs. We investigate the use of the covariation scores in this problem by comparing a prediction

methods that considers site covariation with two methods that do not.

A simple prediction method that does not take site covariation into account is to perform a weighted averaged of the known PWMs, where the weights are based on the similarity of the new PRD sequence and the original ones. Specifically, the probability of finding residue y at site k of the bound peptides of the new PRD is predicted by the following formula:

$$\hat{W}_{(n+1)k}(y) = \frac{\sum_{i=1}^{n} s(i, n+1) W_{ik}(y)}{\sum_{i=1}^{n} s(i, n+1)}, \tag{11}$$

where $s(i, i')$ is a similarity between sequences i and i' in the MSA, such as their sequence identity:

$$s(i, i') = \frac{\sum_{j=1}^{m} \mathbb{1}(M_{ij} = M_{i'j})}{m} \tag{12}$$

Using the covariation scores, we propose an alternative way to define the similarity function. Each MSA site receives a different weight in the calculation, where the weight depends on the covariation score between the site and the target PWM site k. In other words, the similarity score $s(i, i')$ is replaced with a new score $s_k(i, i')$ that is specific to k:

$$\hat{W}_{(n+1)k}(y) = \frac{\sum_{i=1}^{n} s_k(i, n+1) W_{ik}(y)}{\sum_{i=1}^{n} s_k(i, n+1)} \tag{13}$$

$$s_k(i, i') = \frac{\sum_{j=1}^{m} \mathbb{1}(M_{ij} = M_{i'j}) UC(A_j, W_k)}{\sum_{j=1}^{m} UC(A_j, W_k)}, \tag{14}$$

where the uncertainty coefficient $UC(A_j, W_k)$ is computed based on the n original sequences.

We also investigated if prediction accuracy can be improved by using a more complex model. Specifically, we treated each MSA site as a categorical attribute and trained a regression tree model for each probability value in the PWM. The models were then applied to the new sequence to predict the PWM of its bound peptides. We implemented this method using REPTree of the Weka package [49].

To evaluate the effectiveness of the different methods, we performed left-out validation as follows. Each time, we drew a random sample of PRDs to form the testing set. Each sequence in the testing set took turn to take the role of the $(n + 1)$-th sequence. The sequences not included in the sample formed the training set. These sequences were used to compute covariation scores and train prediction models. The procedure was repeated 1,000 times for PDZ and SH3 and 50 times for kinase (due to the long running time) with different random training-testing splits, and the average performance of the trained models on

the testing sets was recorded. To eliminate near-identical sequences in the training and testing sets, for sequences with 90% or more identity, we kept only one of them in the dataset before making the training-testing splits. As most of the synthetic PDZ sequences (described below) are highly similar, we excluded this dataset from this part of study.

Each predicted PWM was compared to the actual PWM, and a prediction error was computed as the root-mean-square difference between their distributions per site:

$$e(\hat{W}_{(n+1)}, \tilde{W}_{(n+1)}) \quad = \quad \frac{\sum_{k=1}^{w} \sum_{y} [\hat{W}_{(n+1)k}(y) - \tilde{W}_{(n+1)k}(y)]^2}{w}, \tag{15}$$

where $\hat{W}_{(n+1)}$ and $\tilde{W}_{(n+1)}$ are the predicted and actual PWMs for the bound peptides of the testing sequence, respectively, and the inner summation is taken over all 20 amino acid residues.

Results
Application of the method to PDZ, SH3 and kinase domains
Natural PDZ domains

We obtained 33 class I human and worm PDZ domains from a recent large-scale study on the specificity map of PDZ domains [13]. Class I PDZ domains were defined by two positions on the ligand, with the pattern $X[T/S]X\phi_{COOH}$, where X and ϕ represent a residue and a hydrophobe, respectively. In the same study, a number of SDRs were experimentally determined, allowing us to validate our prediction results. We focused on only one class of domains as the sequences in different classes are difficult to align due to divergence. The pairwise sequence identity ranges from 0.13 to 0.87, with an average of 0.28.

The binding profile of each domain, in the form of a PWM, was obtained from phage display experiments that expressed a random library of C-terminal peptides [13]. The domains were then aligned to form an MSA using ClustalW [50] followed by manual corrections of some obvious errors. A phylogenetic tree of the MSA sequences was constructed using Biopython's Nexus module [51], and the tree was used to produce sequence weights according to a described algorithm [44]. The uncertainty coefficient between each MSA site and each site of the peptide PWMs was computed. To reduce noise and eliminate highly conserved sites that provide little information about covariation, we considered only sites with no gaps [52] and the most frequent residue occupying no more than half of the total sequence weights. This filtering was also applied to the other domain families described below.

The remaining unfiltered site pairs were then evaluated in two ways. First, their uncertainty coefficients were compared to their physical distances in the co-crystal structure 2H2C of ligand-bound human ZO-1

PDZ1 domain [53] in PDB [54]. Although proximity does not necessarily mean functional or structural dependency, it is usually used as a quick check in covariation studies [18, 27, 29, 33]. It also provides a complete and unbiased alternative to the more costly experimental validations.

Second, we examined the top-scoring site pairs, and compared them with known SDRs from a mutagenesis study [13] in which ten sites of the ERBB2IP-1 domain were mutated and the corresponding changes of peptide specificity reported. This comparison provides direct evaluation of our SDR identification procedure on the subset of sites that were tested in the assay.

Synthetic PDZ domains

In a recent study, the mutagenesis study in Tonikian et al [13] was extended. A large amount of mutations were introduced at the ten sites, resulting in 61 variations of the Erbin domain that are functional in recognizing some C-terminal heptapeptides [55]. As with the natural PDZ domains, we compared the uncertainty coefficients with the physical distances in the 2H2C PDB structure. Since the synthetic PDZ domains are 100% conserved at sites other than the 10 selected ones, and have a specific set of mutations at the 10 sites introduced by the mutagenesis experiments, their MSA exhibits some statistical properties different from those of the natural PDZ domains.

SH3 domains

We obtained 23 yeast SH3 domains and the corresponding PWMs of the bound peptides from phage display experiments from a recent study [38]. We aligned the PRDs based on a published structural alignment [15], and aligned the peptide PWMs based on both the general PxxP pattern and some published alignments [15, 56]. The pairwise sequence identity ranges from 0.07 to 0.79, with an average of 0.24. We applied the same prediction and evaluation procedures as in the case of PDZ, except that in this case we did not have large-scale mutagenesis data, and therefore the prediction results were only evaluated against the physical distances calculated from the crystal structure 1N5Z in PDB, which contains the yeast Pex13 SH3 domain bound to a Pex14 peptide [57], and the findings of some previous studies.

Kinase domains

We also obtained 149 serine/threonine protein kinase domains from four species (S. cerevisiae, H. sapiens, S. pombe and D. discoideum) and the PWMs of their corresponding bound peptides [14]. The MSA was made using MUSCLE [58] followed by some manual corrections. The prediction results were evaluated

against distances calculated from the crystal structure 1ATP of mouse catalytic subunit of cAMP-dependent protein kinase complexed with Mn-ATP and a peptide inhibitor [59] in PDB, and some findings in previous studies. The pairwise sequence identity ranges from 0.09 to 0.92, with an average of 0.23.

Covariation score correlates with physical proximity and reconfirms previous findings

The covariation score between two sites is found to correlate negatively with the physical distance between them, regardless of the exact definition of the distance measure (see Figures 2, 3 and 4 for the results when the distance between residue centers, alpha carbon atoms and closest atoms minus their van der Waals' radii were used, respectively. All p-values were computed using Fisher transformation [60]). For PDZ, we have also compared the correlations based on several other PDZ structures, and observed similar patterns (Figure 5).
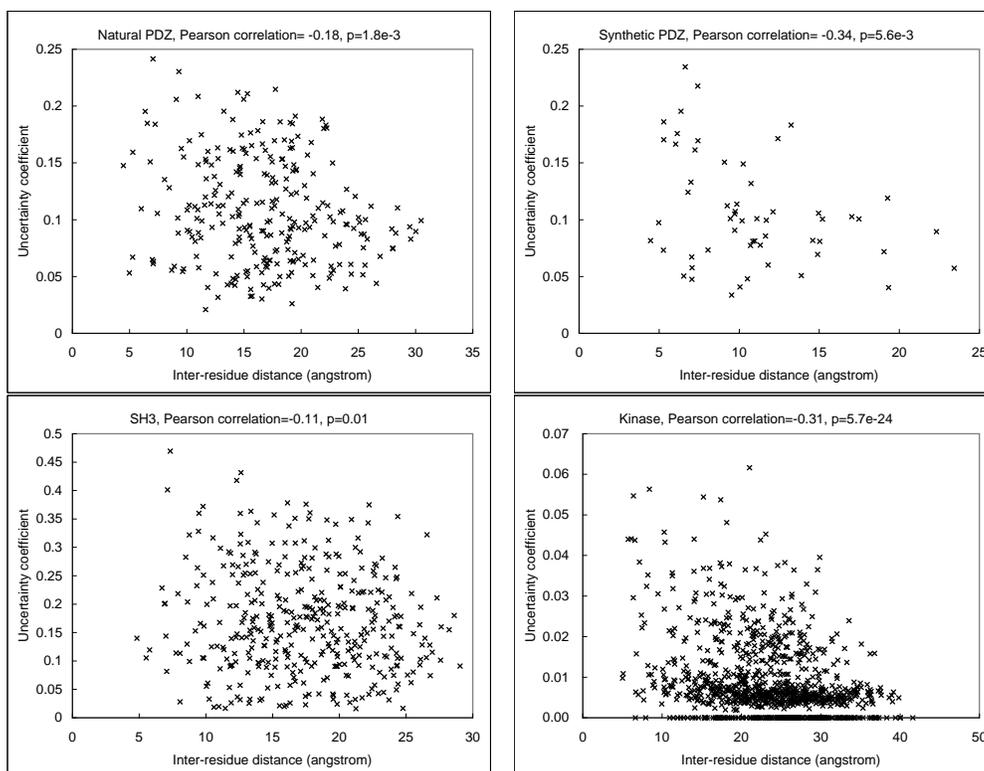
Figure 2: **Correlation between covariation score and physical proximity between each PRD site and each PWM position for the three types of PRDs when distances are computed between residue centers.**

Since low-scoring pairs are more subject to noise, here for each PRD site, we focus on the peptide site that
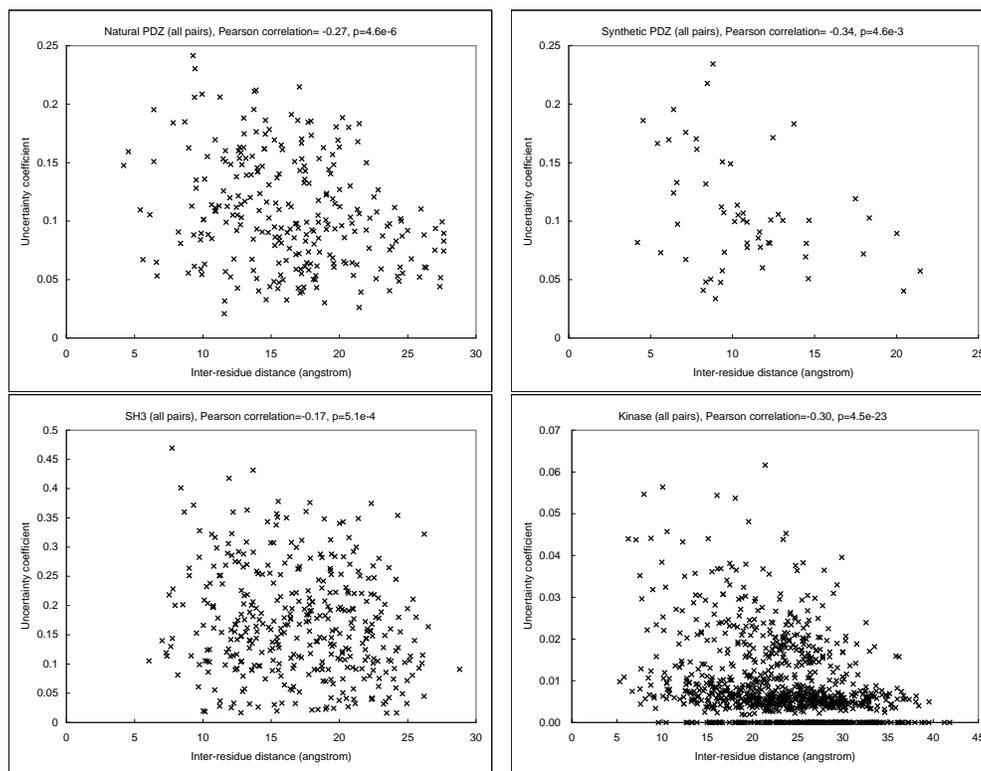
Figure 3: **Correlation between covariation score and physical proximity between each PRD site and each PWM position for the three types of PRDs when distances are computed between alpha carbon atoms.**
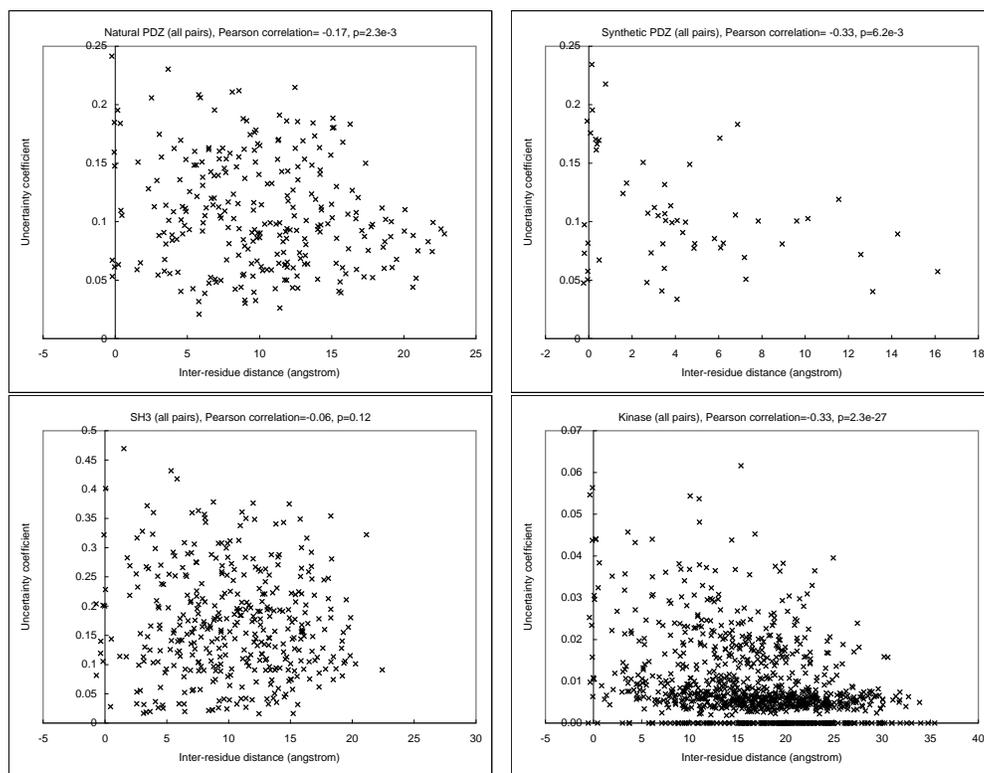
Figure 4: **Correlation between covariation score and physical proximity between each PRD site and each PWM position for the three types of PRDs when distances are computed between the closest atoms minus their van der Waal's radii.**
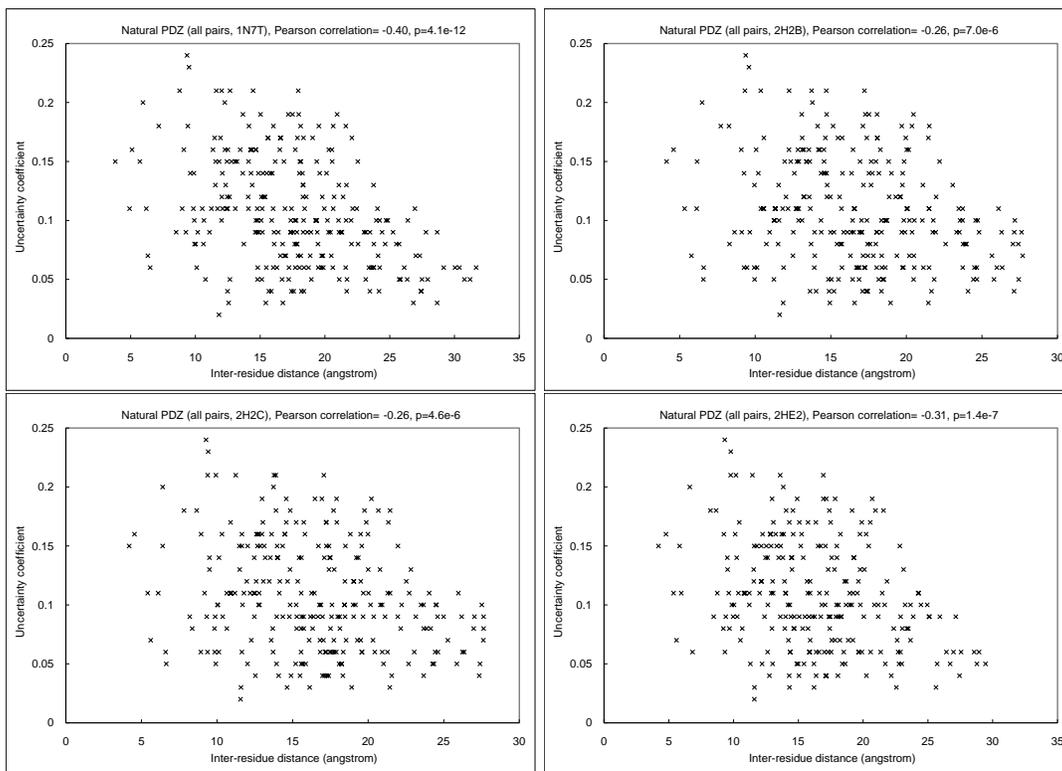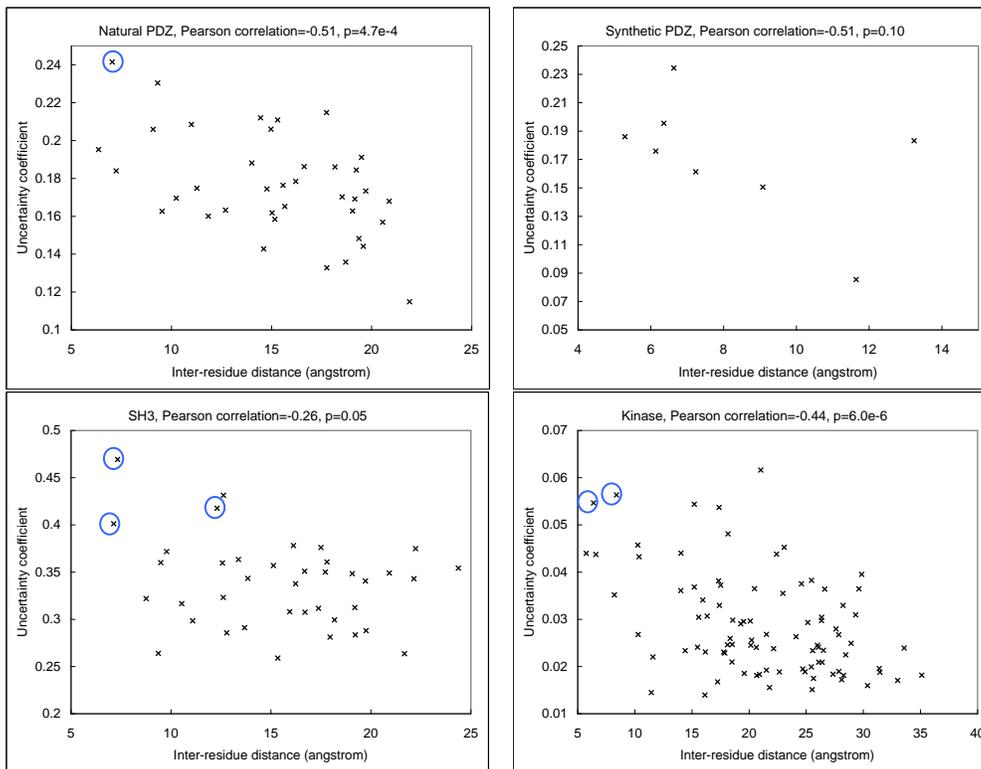
Figure 5: **Correlation between covariation score and physical proximity between each PRD site of a PDZ domain and each PWM position when distances are computed between alpha carbon atoms in four different PDB structures.**

gives the highest uncertainty coefficient with it (Figure 6). The site pairs with the highest covariation scores are listed in Table 2.

Figure 6: **Correlation between covariation score and physical proximity between each PRD site and its *highest-scoring* peptide PWM position for the four types of PRDs.** Selected top-scoring pairs discussed in the text are circled. The correlation for Synthetic PDZ was not statistically significant due to the small number of data points involved.

For natural PDZ domains, the highest-scoring pair (circled) is between Leu60 ($\beta3:\alpha1$-1, structural nomenclature from [53]) of the PDZ domain in 2H2C and position -1 of the binding motif on the bound peptide. These residues are in physical contact in the dimer structure (Figure 7, all visualization produced using VMD [61]). Interestingly, it has been reported that the side chain at $\beta3:\alpha1$-1 can contribute to the recognition of the -1 position of the motif [53], and in the crystal structure from Shank1, a salt bridge is observed between Asp($\beta3:\alpha1$-1) of the PDZ domain and Arg(-1) of the ligand [62]. Our covariation analysis has thus identified these verified SDRs of the PDZ domains *in silico*.

For SH3 domains, the highest-scoring pair is between Asn71 of the SH3 domain and Leu7 (P+2 residue) of the ligand in 1N5Z. These residues are in close physical proximity in the crystal structure (Figure 8). Interestingly, we found that the MSA residues in the first (Asn71), third (Ile73) and fourth (Tyr72)

Table 2: **Site pairs with the highest covariation scores.** For each site on the PRD, only the site on the peptide with the highest score is shown. For synthetic PDZ domains, only 10 sites have variations and among them 2 were filtered, leaving only 8 valid sites. All sites are indexed according to their residue numbers in the reference PDB structures.

| Domain family	Natural PDZ		Synthetic PDZ		SH3		Kinase	
Ref. PDB structure	2H2C		2H2C		1N5Z		1ATP	
Rank	PRD	Peptide	PRD	Peptide	PRD	Peptide	PRD	Peptide
1	Leu60	Trp119	His88	Thr118	Asn71	Leu7	Tyr229	Ile22
2	Val55	Thr117	Ala37	Thr117	Leu34	Leu7	Leu205	Ile22
3	Ala76	Thr117	Ser39	Thr117	Ile73	Leu7	Leu198	Ile22
4	Pro30	Trp119	Leu60	Thr117	Tyr72	Leu7	Lys189	Ile22
5	Ser98	Trp119	Val92	Thr118	Lys35	Leu7	Phe129	Ile22
6	Ala89	Thr117	Ser57	Thr117	Ser44	Leu7	Glu230	Ile22
7	Gln93	Trp119	Asp58	Thr117	Lys61	Leu7	Glu203	Ile22
8	Asp58	Thr117	Phe34	Thr118	Ile68	Leu7	Ile180	Ile22
9	Ala37	Thr117			Leu20	Leu7	Phe187	Arg19
10	Lys103	Trp119			Ala17	Leu7	Cys199	Ile22
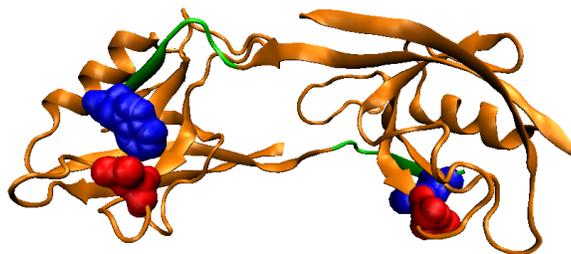
Figure 7: **Top-scoring residue pair for the natural PDZ domain.** The PDZ domain (orange) and the ligand (green) in the biological assembly (dimer) of the PDB structure 2H2C are shown. The top-scoring residue pair between Leu60 (red) on the domain and position -1 of the binding motif on the ligand (blue) are in physical contact.

top-scoring pairs are consecutive in the protein sequence, and two of them are close to Leu7 of the ligand. In a previous study [63], the corresponding residue of Tyr72 on P53BP2, which is an unusual leucine, was hypothesized to cause the protein to bind a peptide very different from its usual ligands. However, mutating the leucine to tyrosine did not affect recognition specificity. As the corresponding residue of Ile73 on P53BP2 is also mutated from the class consensus, and the covariation scores for Asn71 and Ile73 are also high, the three residues may have some combined effects in affecting recognition specificity.

It is also known that the residue corresponding to Glu31 in the RT loop of SH3 domains is a major determinant of the identity of the P-3 residue of the ligand [64, 65]. Since the P-3 residue does not exist in the 1N5Z structure, it is not included in the correlation plots. However, when we checked the covariation scores, indeed the ligand residue having the highest score with Glu31 is the P-3 residue. This observation illustrates the potential of our method to identify SDRs when structural information is not available.
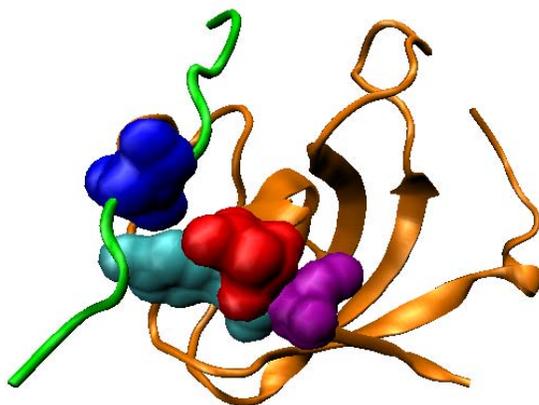

Figure 8: **Top-scoring residue pair for the SH3 domain.** The SH3 domain (orange) and the ligand (green) in the PDB structure 1N5Z are shown. The top-scoring residue pair between Asn71 (red) on the domain and Leu7 of the ligand (blue) is in close physical proximity. Also shown are the MSA residues in the third (Ile73) and fourth (Tyr72) top-scoring pairs.

For protein kinases, the top-scoring pair between Tyr229 of the kinase domain in 1ATP and position +1 of the binding motif in the bound peptide is not physically close. However, the next two pairs (between Leu205 and position +1, and between Leu198 and position +1) are both close in proximity(Figure 9). Both Leu198 and Leu205 were previously found to have hydrophobic interactions with position +1 of the peptide [66, 67]. They are two of the three residues (the third being Pro202, which has the highest covariation score with position +1 as compared to other peptide positions) that form the binding pocket
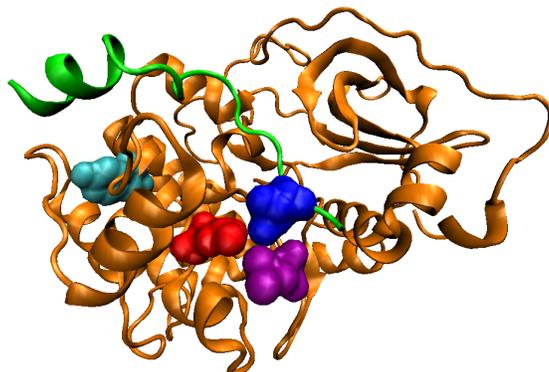
Figure 9: **Top-scoring residue pair for the kinase domain.** The kinase domain (orange) and the ligand (green) in the PDB structure 1ATP are shown. The second top-scoring residue pair between Leu205 (red) on the domain and position 1 of the binding motif on the ligand (blue), as well as the third top-scoring residue pair between Leu198 (purple) and position 1 of the ligand motif, are in close proximity. The MSA residue Tyr229 involved in the top-scoring pair is also shown (cyan).

for position +1 of the peptide, and contributes to the positioning of a significant portion (-3 to +1 positions) of the bound peptide [67].

We found that the pair between Tyr229 and position +1, besides being the top-scoring pair when uncertainty coefficient was used as the covariation score, was also consistently one of the highest-scoring pairs when covariation was measured by other normalized forms of mutual information. These consistent results made us hypothesize that Tyr229 is also involved in determining binding specificity of the kinase domain, and has long-range coupling with position +1 of the bound peptide. In a previous study, this residue was predicted to be involved in linking nucleotide binding and peptide binding in protein kinases [68]. Interestingly, in the study Tyr229 and Leu205 are predicted to belong to the same special network (termed the θ-shaped network) of related residues. It thus might be the case that Tyr229 acts through the residues in the network to affect the recognition of the +1 position of the peptide.

PDZ predictions are consistent with mutagenesis results

We further validated our predictions with the natural PDZ domains by using a mutagenesis dataset from a domain specificity study [13] as described in the Materials and Methods section.

In our covariation calculation, among the 10 mutated sites, 4 were filtered as they were too conserved (Ile36 [β2-1], Ile38 [β2-3], His88 [α2-1] and Val92 [α2-5]). Interestingly, in the mutagenesis study, indeed no significant changes to the peptide PWM could be observed for Ile36 and Ile38. The high conservation of them is thus probably caused by a structural or functional role that is independent of peptide binding.

20

For the remaining 6 sites, we examined their maximum-scoring peptide residues. Four out of the 6 had significant changes of the PWM at the predicted positions in the mutagenesis study (Table 3), including the top-scoring pair among all predictions, between Leu60 and position -1. For the remaining two, changes were also observed, albeit with lower statistical significance.

Table 3: **Validation results of our PDZ predictions.** The 6 sites validated in Tonikian et al. [13] that were not filtered in our calculation are shown, along with their highest-scoring peptide PWM positions. Significant changes of the predicted PWM positions after mutating the PDZ sites were reported for 4 cases, while in the remaining 2 cases some changes were also observed.

| PDZ site in 2H2C	Structure-based nomenclature [53]	Ligand position	Uncertainty coefficient	Distance (Å)	Specificity change
Phe34	$\beta1:\beta2$-7	-3	0.16	15.0	Significant
Ala37	$\beta2$-2	-3	0.20	6.4	Significant
Ser39	$\beta2$-4	-1	0.16	9.5	Significant
Ser57	$\beta3$-4	-3	0.18	7.2	Observed
Asp58	$\beta3$-5	-3	0.21	9.1	Observed
Leu60	$\beta3:\alpha1$-1	-1	0.24	7.0	Significant

The predicted pair between Phe34 and position -3 has a distance of 15.0Å in the PDB structure 2H2C. If the SDRs of different class I PDZ domains are similar, this predicted pair is another example that suggests our covariation method could potentially identify physically distant SDR pairs.

**Covariation scores improve prediction of bound peptide profiles**

As described in the Materials and Methods section, we compared three different methods for predicting the profiles of bound peptides. The prediction results are shown in Figure 10.

In general, the prediction error is lower with a smaller fraction of PRDs left out for testing. This is expected, as having more PRDs in the training set allows for more accurate computation of covariation scores and the construction of more informative prediction models.

In all cases, the weighted average method using covariation scores outperformed regression tree and the weighted average method using uniform scoring. This suggests that the covariation score provided a meaningful way to weight useful features (i.e., PRD sites) for predicting residue distributions of the bound peptides. Interestingly, while the regression tree method also performed feature weighting, in general it performed worse than both weighted average methods. The low performance could be due to over-fitting, as the regression trees are rich in expressive power while the numbers of PRDs in the datasets are small.
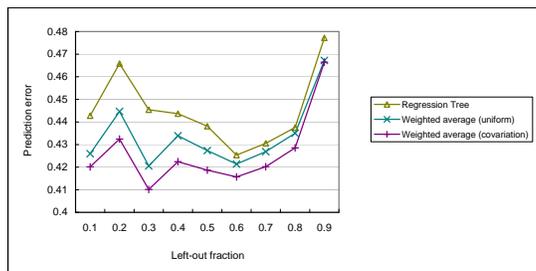
(a) Natural PDZ
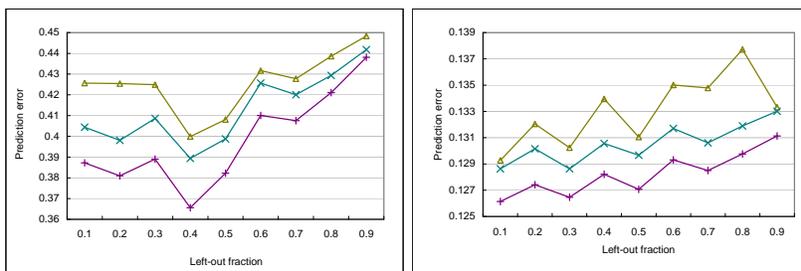
(b) SH3

(c) Kinase

Figure 10: **Results of predicting PWMs of bound peptides.** Three methods for predicting PWMs of bound peptides were compared using data from (a) natural PDZ domains, (b) SH3 domains and (c) kinase domains.

Discussion

We present a novel way to use underutilized data. Our method is a valuable tool for exploring the specificity space of PRDs. Moreover, as the amount of specificity data is increasing swiftly, also due to the advent of next-generation sequencing and its applications to phage display [69], our method will prove even more valuable to make optimal use of this kind of data.

We think our covariation approach can be used in conjunction with other methods to improve SDR predictions. Since most current approaches are based on force fields and structural methods, our method opens up a new perspective for improvement. As shown in the recent PRD specificity prediction challenge of the DREAM4 competition [70], most likely a combination of structural and statistical methods will be most successful at predicting specificities.

One limitation of our approach is that it does not consider possible relationships between different residue pairs. As multiple PRD residues could simultaneously interact with a peptide residue and multiple peptide residues could simultaneously interact with a PRD residue [71], binding specificity could be more accurately modeled by considering covariation signals between residue groups. Furthermore, since a residue could have an indirect covariation signal with another residue through an intermediate residue [42],

performing residue group analysis could help filter out these non-SDR intermediates that have relatively high covariation signals.

Another limitation of our approach is that it fails to identity SDRs that are highly conserved. Indeed we have observed that in PDZ, some highly conserved residues (e.g. α2-1) are physically close to the peptide and have been experimentally shown to affect binding specificity when mutated [13] (Figure 11). On the other hand, some SDRs are not highly conserved, but exhibit strong covariation patterns with peptide residues (e.g. β3:α1-1). Future approaches could improve upon our current method by combining information about conservation with covariation to identity SDRs.
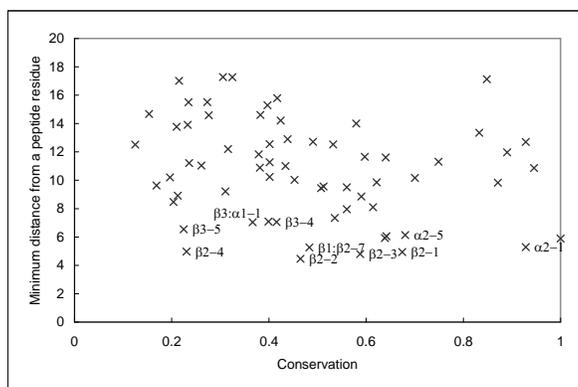

Figure 11: **Conservation and distance to the closest peptide residue of each PDZ domain site. For each site on the domain, we computed the total sequence weight of the sequences having a particular amino acid at the site. The conservation of the site is defined by the maximum of such total weights normalized by the total sequence weight of all sequences in the MSA.**

There are also PRD residues that determine binding but not binding specificity in that if they are mutated, the resulting effect on binding cannot be restored by a second mutation. For instance, if a residue is critical to the protein structure, mutating it could seriously affect the stability of the protein, which in turn affects peptide binding. These residues are likely to be very conserved, and thus would not be ranked high by our method.

A third limitation of our approach, and more generally of all covariation analysis methods based on a multiple sequence alignment, is the dependency on the alignment quality. Different alignments could give very different results, especially for alignments with many gaps. To cope with this issue, we have used a published alignment for SH3, and made some manual corrections to the PDZ and kinase alignments generated by computer programs. Future approaches should try to minimize the effect of alignment quality on the analysis results.

While we have attempted to predict peptide PWMs, it is also possible to predict interactions given a PRD and a peptide. This problem has recently been studied using some large-scale datasets [9,52]. It would be interesting to study how the concept of covariation can be incorporated into these prediction methods.

Conclusions

We have presented a novel approach that utilizes an as of yet underused source of data. We have shown that the covariation scores are consistent with previous findings from both a large-scale study, and other individual experiments. In addition, we have identified a number of candidate SDRs in a ranked list for future experimental validation. In particular, with the top-scoring pairs from natural PDZ domains and kinase domains both verified in previous work, the SH3 top-scoring pairs are good candidates for testing their roles in determining the binding specificity of SH3 domains.

Competing interests

The authors declare they have no competing interests.

Authors contributions

All authors conceived the project and design. KY, LU, SS and PK prepared the data. KY, LU, SS and XH implemented the algorithms, performed the computational experiments, and analyzed the results. KY, LU, MG and PK wrote the paper. All authors read and approved the document.

Acknowledgements

References

1. Pawson T, Nash P: **Assembly of Cell Regulatory Systems through Protein Interaction Domains**. *Science* 2003, **300**(5618):445–452.

2. Pawson T, Linding R: **Network Medicine**. *FEBS Letters* 2008, **582**:1266–1270.

3. Alto NM, Shao F, Lazar CS, Brost RL, Chua G, Mattoo S, McMahon SA, Ghosh P, Hughes TR, Boone C, Dixon JE: **Identification of a Bacterial Type III Effector Family with G Protein Mimicry Functions**. *Cell* 2006, **124**:133–145.

4. Doorbar J: **Molecular Biology of Human Papillomavirus Infection and Cervical Cancer**. *Clinical Science* 2006, **110**:525–541.

5. Carducci M, Licata L, Peluso D, Castagnoli L, Cesareni G: **Enriching the ViralVHost Interactomes with Interactions Mediated by SH3 Domains**. *Amino Acids* 2010, **38**(5):1541–1547.

6. Lam HYK, Kim PM, Mok J, Tonikian R, Sidhu SS, Turk BE, Snyder M, Gerstein MB: **MOTIPS: Automated Motif Analysis for Predicting Targets of Modular Protein Domains**. *BMC Bioinformatics* 2010, **11**(243).

7. Tonikian R, Zhang Y, Boone C, Sachdev SS: **Identifying Specificity Profiles for Peptide Recognition Modules from Phage-Displayed Peptide Libraries**. *Nature Protocols* 2007, **2**(6):1368–1386.

8. Landgraf C, Panni S, Montecchi-Palazzi L, Castagnoli L, Schneider-Mergener J, Volkmer-Engert R, Cesareni G: **Protein Interaction Networks by Proteome Peptide Scanning**. *PLoS Biology* 2004, **2**:e14.

9. Stiffler MA, Chen JR, Grantcharova VP, Lei Y, Fuchs D, Allen JE, Zaslavskaia LA, MacBeath G: **PDZ Domain Binding Selectivity is Optimized across the Mouse Proteome**. *Science* 2007, **317**(5836):364–369.

10. Kaushansky A, Allen JE, Gordus A, Stiffler MA, Karp ES, Chang BH, MacBeath G: **Quantifying Protein-Protein Interactions in High Throughput using Protein Domain Microarrays**. *Nature Protocols* 2010, **5**(4):773–790.

11. Hutti JE, Jarrell ET, Chang JD, Abbott DW, Storz P, Toker A, Cantley LC, Turk BE: **A Rapid Method for Determining Protein Kinase Phosphorylation Specificity**. *Nature Methods* 2004, **1**:27–29.

12. Li L, Shakhnovich EI, Mirny LA: **Amino Acids Determining Enzyme-Substrate Specificity in Prokaryotic and Eukaryotic Protein Kinases**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(8):4463–4468.

13. Tonikian R, Zhang Y, Sazinsky SL, Currell B, Yeh JH, Reva B, Held HA, Appleton BA, Evangelista M, Wu Y, Xin X, Chan AC, Seshagiri S, Lasky LA, Sander C, Boone C, Bader GD, Sidhu SS: **A Specificity Map for the PDZ Domain Family**. *PLoS Biology* 2008, **6**:e239.

14. Mok J, Kim PM, Lam HYK, Piccirillo S, Zhou X, Jeschke GR, Sheridan DL, Parker SA, Desai V, Jwa M, Cameroni E, Niu H, Good M, Remenyi A, Ma JLN, Sheu YJ, Sassi HE, Sopko R, Chan CSM, Virgilio CD, Hollingsworth NM, Lim WA, Stern DF, Stillman B, Andrews BJ, Gerstein MB, Snyder M, Turk BE: **Deciphering Protein Kinase Specificity Through Large-Scale Analysis of Yeast Phosphorylation Site Motifs**. *Science Signaling* 2010, **3**:ra12.

15. Cesareni G, Panni S, Nardelli G, Castagnoli L: **Can We Infer Peptide Recognition Specificity Mediated by SH3 Domains?** *FEBS Letters* 2002, **513**:38–44.

16. Beuming T, Farid R, Sherman W: **High-Energy Water Sites Determine Peptide Binding Affinity and Specificity of PDZ domains**. *Protein Science* 2009, **18**:1609–1619.

17. Lichtarge O, Bourne HR, Cohen FE: **An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families**. *Journal of Molecular Biology* 1996, **257**:342–358.

18. Göbel U, Sander C, Schneider R, Valencia A: **Correlated Mutations and Residue Contacts in Proteins**. *Proteins: Structure, Function, and Bioinformatics* 1994, **18**:309–317.

19. Shindyalov IN, Kolchanov NA, Sander C: **Can Three-Dimensional Contacts in Protein Structures be Predicted by Analysis of Correlated Mutations?** *Protein Engineering* 1994, **7**(3):349–358.

20. Lockless SW, Ranganathan R: **Evolutionarily Conserved pathways of Energetic Connectivity in Protein Families**. *Science* 1999, **286**(5438):295–299.

21. Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE: **Co-evolution of Proteins with their Interaction Partners**. *Journal of Molecular Biology* 2000, **299**:283–293.

22. Pazos F, Valencia A: **In Silico Two-hybrid System for the Selection of Physically Interacting Protein Pairs**. *Proteins: Structure, Function, and Bioinformatics* 2002, **47**:219–227.

23. Ramani AK, Marcotte EM: **Exploiting the Co-evolution of Interacting Proteins to Discover Interaction Specificity**. *Journal of Molecular Biology* 2003, **327**:273–284.

24. Pazos F, Ranea JAG, Juan D, Sternberg MJE: **Assessing Protein Co-evolution in the Context of the Tree of Life Assists in the Prediction of the Interactome**. *Journal of Molecular Biology* 2005, **352**(4):1002–1015.

25. Tiller ERM, Biro L, Li G, Tillo D: **Codep: Maximizing Co-Evolutionary Interdependencies to Disvoer Interacting Proteins**. *Proteins: Structure, Function, and Bioinformatics* 2006, **63**(4):822–831.

26. Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R: **Evolutionary Information for Specifying a Protein Fold**. *Nature* 2005, **437**(7058):512–518.

27. Gloor GB, Martin LC, Wahl LM, Dunn SD: **Mutual Information in Protein Multiple Sequence Alignments Reveals Two Classes of Coevolving Positions**. *Biochemistry* 2005, **44**(19):7156–7165.

28. Fuchs A, Martin-Galiano AJ, Kalman M, Fleishman S, Ben-Tal N, Frishman D: **Co-evolving Residues in Membrane Proteins**. *Bioinformatics* 2007, **23**(24):3312–3319.

29. Halperin I, Wolfson H, Nussinov R: **Correlated Mutations: Advances and Limitations. A Study on Fusion Proteins and on the Cohesin-Dockerin Families**. *Proteins: Structure, Function, and Bioinformatics* 2006, **63**:832–845.

30. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A: **Correlated Mutations Contain Information about Protein-Protein Interaction**. *Journal of Molecular Biology* 1997, **271**:511–523.

31. Shulman AI, Larson C, Mangelsdorf DJ, Ranganathan R: **Structural Determinants of Allosteric Ligand Activation in RXR Heterodimers**. *Cell* 2004, **116**:417–429.

32. Pollock DD, Taylor WR, Goldman N: **Coevolving Protein Residues: Maximum Likelihood Identification and Relationship to Structure**. *Journal of Molecular Biology* 1999, **287**:187–198.

33. Dekker JP, Fodor A, Aldrich RW, Yellen G: **A Perturbation-based Method for Calculating Explicit Likelihood of Evolutionary Co-variance in Multiple Sequence Alignments**. *Bioinformatics* 2004, **20**(10):1565–1572.

34. Martin LC, Gloor GB, Dunn SD, Wahl LM: **Using Information Theory to Search for Co-evolving Residues in Proteins**. *Bioinformatics* 2005, **21**(22):4116–4124.

35. Larson SM, Nardo AAD, Davidson AR: **Analysis of Covariation in an SH3 Domain Sequence Alignment: Applications in Tertiary Contact Prediction and the Design of Compensating Hydrophobic Core Substitutions**. *Journal of Molecular Biology* 2000, **303**:433–4446.

36. Galitsky B: **Revealing the Set of Mutually Correlated Positions for the Protein Families of Immunoglobulin Fold**. *In Silico Biology* 2003, **3**(0022).

37. Dimmic MW, Hubisz MJ, Bustamante CD, Nielsen R: **Detecting Coevolving Amino Acid Sites using Bayesian Mutational Mapping**. *Bioinformatics* 2005, **21**(Suppl. 1):i126–i135.

38. Tonikian R, Xin X, Toret CP, Gfeller D, Landgraf C, Panni S, Paoluzi S, Castagnoli L, Currell B, Seshagiri S, Yu H, Winsor B, Vidal M, Gerstein MB, Bader GD, Volkmer R, Cesareni G, Drubin DG, Kim PM, Sidhu SS, Boone C: **Bayesian Modeling of the Yeast SH3 Domain Interactome Predicts Spatiotemporal Dynamics of Endocytosis Proteins**. *PLoS Biology* 2009, **7**:21000218.

39. Yip KY, Patel P, Kim PM, Engelman DM, McDermott D, Gerstein M: **An Integrated System for Studying Residue Coevolution in Proteins**. *Bioinformatics* 2008, **24**(2):290–292.

40. Halabi N, Rivoire O, Leibler S, Ranganathan R: **Protein Sectors: Evolutionary Units of Three-Dimensional Structure**. *Cell* 2009, **138**:774–786.

41. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T: **Identification of Direct Residue Contacts in Protein-Protein Interaction by Message Passing**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**:67–72.

42. Burger L, van Nimwegen E: **Disentangling Direct from Indirect Co-Evolution of Residues in Protein Alignments**. *PLoS Computational Biology* 2010, **6**:e1000633.

43. Cover TM, Thomas JA: *Elements of Information Theory*. Wiley-Interscience, second edition 2006.

44. Gerstein M, Sonnhammer ELL, Chothia C: **Volume Changes in Protein Evolution**. *Journal of Molecular Biology* 1994, **236**:1067–1078.

45. Press WH, Flannery BP, Teukolsky SA, Vetterling WT: *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press 1986.

46. Goebel B, Dawy Z, Hagenauer J, Mueller JC: **An Approximation to the Distribution of Finite Sample Size Mutual Information Estimates**. In *Proceedings of IEEE International Conference on Communications (ICC'05), Volume 2* 2005:1102–1106.

47. Aktulga HM, Kontoyiannis I, Lyznik LA, Szpankowski L, Grama AY, Szpankowski W: **Identifying Statistical Dependence in Genomic Sequence via Mutual Information Estimates**. *EURASIP Journal on Bioinformatics and Systems Biology* 2007, **2007**:14741.

48. Pirovano W, Feenstra KA, Heringa J: **Sequence Comparison by Sequence Harmony Identifies Subtype-specific Functional Sites**. *Nucleic Acids Research* 2006, **34**(22):6540–6548.

49. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH: **The WEKA Data Mining Software: An Update**. *SIGKDD Explorations* 2009, **11**:10–18.

50. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignmnet through Sequence Weighting, Position-Specific Gap Penalties and Weight matrix Choice**. *Nucleic Acids Research* 1994, **22**(22):4673–4680.

51. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL: **Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics**. *Bioinformatics* 2009, **25**(11):1422–1423.

52. Chen JR, Chang BH, Allen JE, Stiffler MA, MacBeath G: **Predicting PDZ Domain-Peptide Interactions from Primary Sequences**. *Nature Biotechnology* 2008, **26**(9):1041–1045.

53. Appleton BA, Zhang Y, Wu P, Yin JP, Hunziker W, Skelton NJ, Sidhu SS, Wiesmann C: **Comparative Structural Analysis of the Erbin PDZ Domain and the First PDZ Domain of ZO-1. Insights into Determinants Of PDZ Domain Specificity**. *Journal of Biological Chemistry* 2006, **281**(31):22312–22320.

54. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Research* 2000, **28**.

55. Ernst A, Sazinsky SL, Hui S, Currell B, Dharsee M, Seshagiri S, Bader GD, Sidhu SS: **Rapid Evolution of Functional Complexity in a Domain Family**. *Science Signaling* 2009, **2**:ra40.

56. Tong AHY, Drees B, Nardelli G, Bader GD, Brannetti B, Castagnoli L, Evangelista M, Ferracuti S, Nelson B, Paoluzi S, Quondam M, Zucconi A, Hogue CWV, Fields S, Boone C, Cesareni G: **A Combined Experimental and Computational Strategy to Define Protein Interaction Networks for Peptide Recognition Modules**. *Science* 2002, **295**(5553):321–324.

57. Douangamath A, Filipp FV, Klein AT, Barnett P, Zou P, Voorn-Brouwer T, Vega M, Mayans OM, Sattler M, Distel B, Wilmanns M: **Topography for Independent Binding of α-Helical and PPII-Helical Ligands to a Peroxisomal SH3 Domain**. *Molecular Cell* 2002, **10**:1007–1017.

58. Edgar RC: **MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput**. *Nucleic Acids Research* 2004, **32**(5):1792–1797.

59. Zheng J, Trafny EA, Knighton DR, Xuong NH, Taylor SS, Eyck LFT, Sowadski JM: **A Refined Crystal Structure of the Catalytic Subunit of cAMP-Dependent Protein Kinase Complexed with MnATP and a Peptide Inhibitor**. *Acta Crystallographica D* 1993, **49**:362–365.

60. Fisher RA: **On the 'probable error' of a coefficient of correlation deduced from a small sample**. *Metron* 1921, **1**:3–32.

61. Humphrey W, Dalke A, Schulten K: **VMD: Visual Molecular Dynamics**. *Journal of Molecular Graphics* 1996, **14**:33–38.

62. Im YJ, Lee JH, Park SH, Park SJ, Rho SH, Kang GB, Kim E, Eom SH: **Crystal Structure of the Shank PDZ-Ligand Complex Reveals a Class I PDZ Interaction and a Novel PDZ-PDZ Dimerization**. *The Journal of Biological Chemistry* 2003, **278**(48):48099–48104.

63. Espanel X, Sudol M: **Yes-associated Protein and p53-binding Protein-2 Interact through Their WW and SH3 Domains**. *The Journal of Biological Chemistry* 2001, **276**(17):14514–14523.

64. Musacchio A, Saraste M, Wilmanns M: **High-Resolution Crystal Structures of Tyrosine Kinase SH3 Domains Complexed with Proline-Rich Peptides**. *Nature Structural and Molecular Biology* 1994, **1**(8):546–551.

65. Yu H, Chen JK, Feng S, Dalgarno DC, Brauer AW, Schrelber SL: **Structural Basis for the Binding of Proline-rich Peptides to SH3 Domains**. *Cell* 1994, **76**:933–945.

66. Smith CM, Radzio-Andzelm E, Madhusudan, Akamine P, Taylor SS: **The Catalytic Subunit of cAMP-Dependent Protein Kinase: Prototype for an Extended Network of Communication**. *Progress in Biophysics and Molecular Biology* 1999, **71**:313–341.

67. Johnson DA, Akamine P, Radzio-Andzelm E, Madhusudan, Taylor SS: **Dynamics of cAMP-Dependent Protein Kinase**. *Chemical Reviews* 2001, **101**(8):2243–2270.

68. Xu F, Du P, Shen H, Hu H, Wu Q, Xie J, Yu L: **Correlated Mutation Analysis on the Catalytic Domains of Serine/Threonine Protein Kinases**. *PLoS One* 2009, **4**(e5913).

69. Gfeller D, Ernst A, Verschueren E, Vanhee P, Dar N, Serrano L, Sidhu SS, Bader GD, Kim PM: **Uncovering Multiple Specificity in Protein Recognition Modules** 2010. [(submitted)].

70. **DREAM4, Challenge 1: Peptide Recognition Domain (PRD) Specificity Prediction** 2009. [http://wiki.c2b2.columbia.edu/dream09/index.php/D4c1].

71. London N, Movshovitz-Attias D, Schueler-Furman O: **The Structural Basis of Peptide-Protein Binding Strategies**. *Structure* 2010, **18**:188–199.