Supplementary Information

Table of Content

1. Supplementary Text

Mixture of PWMs

The multiple PWMs model is mathematically defined as a mixture of PWMs. A mixture of K PWMs is described by the model parameters θ_{li}^{k}, which correspond to the probability of having amino acid i and position l according to the k^{th} PWM, and the mixing coefficients π^{k} quantifying the weight of each PWM ($\sum_{k=1}^{K}\pi^{k}=1$) (Bailey and Elkan, 1994) (Bishop, 2006). The score of a peptide $X=(x_1...x_L)$ of length L is then given by:

Equation 1

$$P(X\mid\theta^{1},...,\theta^{k},\pi)=\sum_{k=1}^{K}\pi^{k}P(X\mid\theta^{k})=\sum_{k=1}^{K}\pi^{k}\prod_{l=1}^{L}\theta_{lx_l}^{k}$$

Given a dataset $D = \{X^1,...,X^N\}$ of N peptides, the different parameters of the mixture of PWMs can be learned using a Maximum Likelihood (ML) approach with Dirichlet prior (Bailey *et al*, 1994) (Bishop, 2006).

The log-likelihood function reads:

Equation 2

$$\log\{P(D \mid \theta,\pi)P(\theta)\} = \sum_{n=1}^{N} \log\left\{\sum_{k=1}^{K} \pi^k P(X^n \mid \theta^k)\right\} + \log(P(\theta))$$

with $\theta = \{\theta_{li}^k\}$. $P(\theta) = \frac{1}{Z}\prod_{k=1}^{K}\prod_{l=1}^{L}\prod_{i=1}^{20}\left(\theta_{li}^k\right)^{\alpha_{li}^k}$ is the usual Dirichlet prior with hyper-parameters α_{li}^k and Z is the normalizing factor. The right-hand side of 2 can be maximized using the Expectation-Maximization (EM) algorithm (Bailey *et al*, 1994). EM optimization is computationally very efficient since its complexity scales linearly with the number of binding peptides and therefore the algorithm can be applied on very large datasets.

Choosing the prior

The Dirichlet distribution used for $P(\theta)$ is the conjugate prior for the multinomial distribution underlying the PWMs and thus provides the most natural and common choice of prior. Intuitively, such prior represents the fact that we want to exclude values of θ_{li}^k too close to 0. The hyper-parameters α_{li}^k quantify how much uncertainty we expect in our data. For this reason we decided to take them proportional to the entropy in each alignment column: $\alpha_{li}^k = -0.2 \cdot \sum_{j=1}^{20} \tilde{p}_l^k(j)\log_{20}(\tilde{p}_l^k(j))$, where $\tilde{p}_l^k(j) = \frac{p_l^k(j) + 0.1}{1 + N \cdot 0.1}$ is the regularized probability at each position (the regularization parameter, chosen here as 0.1, is needed so that positions completely specific for one

residue still have $\alpha_{li}^{k} \neq 0$). If K=1, the hyper-parameters are equivalent to pseudo-counts that are often used with single PWMs. In addition, for any missing amino acid (i.e. N- or C-terminal gaps) in the peptide alignments, we add *1/20* to the values of α_{li}^{k}. In this way the missing amino acids are correctly treated as *a priori* uncertainties in the data.

Changing K

We tried to change the value of K to explore the sensitivity of the model with respect to this parameter. Adding one cluster for each domain, we observed that, in 38% of the cases, one of the components of the mixture model was given a weight lower than 0.05, suggesting that it is not much relevant. In many other cases some of the multiple PWMs were clearly redundant. In terms of cross-validation performance (see Table S1), we did not observe significant changes with larger K.

Average entropy before and after clustering

The entropy at a given position in a set of aligned peptide binding to a domain quantifies the specificity observed at this position (low values correspond to highly specific positions). It is defined as $-\sum_{i=1}^{20} p(i)\log_{20}(p(i))$, where *p(i)* is the probability of amino acid *i* at this position. To compute the global average entropy before clustering, the entropies were first averaged over all positions for each domain and then the average over all domains was taken. To compute the global average entropy after clustering, the entropies were averaged over all position for each cluster and each cluster was weighted by its mixing coefficient (π^k) before averaging over all domains. The p-values were computed by reshuffling the clusters 10^4 times for each domain, in order to ensure that the observed decrease in entropy did not result from a

scaling effect (smaller sets of peptides are statistically more likely to have lower entropies).

Hidden Markov Models

We used standard HMMs (Bishop, 2006) with both emission and transition probabilities depending on the amino acid positions along the peptides. To allow for comparison with multiple PWMs, the number of hidden units was taken as the number of clusters and the EM optimization was seeded with emission probabilities corresponding to the different residue frequencies in each clusters. The same prior distributions as in the mixture model were used. We note that mixtures of PWMs and HMMs are mathematically similar objects (mixtures of PWMs can be seen as a constrained HMMs in which the choice of a hidden state is done once for all in the beginning and transitions from one hidden state to another one are not allowed). Of course, single PWMs can also be seen as trivial HMMs with only one hidden state. As such, scores given by single PWMs, multiple PWMs and HMMs can be readily compared between each other.

Artificial Neural Networks

Standard three-layer ANNs (Bishop, 2006) as implemented in R were trained for each domain, with the same number of hidden units as in the mixture of PWMs. The decay parameter was taken as 0.0005. To generate negative data, the PDZ and SH3 domains were grouped into different specificity classes, following previously identified clusters (Tonikian *et al*, 2008), (Tonikian *et al*, 2009). The negative data for a domain in one specificity cluster were sampled from domains of the same family but in other specificity classes, ensuring that none of the negatives appear in the set of positives (Miller *et al*, 2008). For the three WW domains, such strategy could not be applied

and the negative data were randomly generated. To achieve a balanced training set, a ratio of five negatives for each positive was used throughout this work (Miller *et al*, 2008).

Cross-validation

Ten-fold cross-validation was used to compare the different models describing the specificity of PDZ, SH3 and WW domains. Since no negative data have been used to train both multiple PWMs and single PWMs, we first applied the cross-validation technique described in (Sharon et al, 2008). We start by randomly dividing each set of phage peptides into 10 groups of similar size. 9 groups were used to train the multiple PWMs and the single PWMs, including the initial clustering step. The probability for each peptide in the testing sets was then computed. This allows us to compare the log-likelihood of the testing sets for each model. The cross-validation score is defined as:

$$CV_{\text{score}} = \left\langle \frac{\log\left\{P(X^{\text{test}} \mid \text{mixture of PWMs})\right\} - \log\left\{P(X^{\text{test}} \mid \text{single PWMs})\right\}}{\mid X^{\text{test}} \mid} \right\rangle$$

where the average is taken over the ten peptide sets, each being used once as testing data. Results are shown in Figure S2.

To compare with HMMs and ANNs, we used ROC curve analysis. Both the positives and negatives data were split into ten groups for each domain. Nine positive and negative groups were used for training the model (although for multiple PWMs and HMMs, only positive training data are used in the training step) and the remaining positive and negative groups were used for testing. AROC values listed in Supplementary Table S1 correspond to average and standard-deviation over ten possible choices for the testing sets. The very high cross-validation AROC values

(compared to typical values found in other studies) can be partly explained by the nature of our data, since phage display allows for a very dense and accurate sampling of the specificity space of Modular Protein Domains binding to short peptides.

Blind testing with other experimental datasets

SH3 domains

The SPOT data used in this work were retrieved from (Tonikian et al, 2009) and (Wiedemann et al, 2004). The SPOT technique allows for printing peptides on a chip and interactions with a domain are measured as light intensity of GST-domain fusion protein. It is recognized that very low intensity signals are difficult to quantitatively interpret. We therefore restricted our comparison to proteins detected in the SPOT array with a light intensity BLU>10^5 (Wiedemann et al, 2004). The correlation between the light intensity and the scores given by these three models (since ANNs are not a probabilisitic models, the ANN scores cannot be compared directly with the one from the other models). We also used Y2H data published in (Tonikian et al, 2009). For all interactions detected in the Y2H assay, the scores for single PMWs, multiple PWMs, HMMs and ANNs were computed by scanning over the whole length of the protein and choosing the highest scoring position as the predicted binding site. Since Y2H experiments only provide binary information about interactions, we used the scores obtained with the four models to rank all yeast proteins and tested the enrichment in Y2H interactions among the high-ranking proteins, using a strategy similar to the one used in Gene Set Enrichment Tests (Subramanian et al, 2005). One way of representing enrichment is to use the so-called running sum where for each rank the curve goes up if the protein was found in the Y2H experiment and down if it was not (the increments and decrements are chosen so

that the curve always reaches 0 at the end). Figure S4 shows the final results. In this case ANNs performed worst. This can be understood by observing that ANNs as trained in this work are slightly more likely than multiple PWMs or HMMs to find by chance a peptide ranking high within a set of randomly generated peptides, possibly because of the difficulty of having negative training data that correctly represent the entire space of negative interactions. Although this probability is small in absolute value, it becomes amplified when scanning over the whole length of a protein and taking the highest scoring region. This issue was not present in the other validation experiments, which only consisted in small peptides.

PDZ domains

SPOT data are available for ERBIN and MLLT4 PDZ domains (Wiedemann et al, 2004), and the correlations with single PWMs, multiple PWMs and HMMs scores are displayed in Table S2. Here again, multiple PWMs outperform single PWMs. A benchmarking set of positive interactions was also retrieved from PDZbase (see Table S3). To generate ROC curves, negative interactions were chosen as all other possible interactions between these proteins (i.e., the complement of positive interactions), except for DLG1#1 with HTR2C and HTR2A, and LRRC7#1 with CTNND2, as some evidence indicates possible interactions, although the domain mediating these interactions is not known. AROC values for multiple PWMs, HMMs and ANNs are displayed in Table S4.

Structural Analysis with FoldX

Binding energies and amino acid preferences where analyzed with FoldX (Schymkowitz *et al*, 2005) at 0.15 Ionic strength and pH7.0. In Figure 5D the recent structure of DLG1 PDZ1 (PDB: 2WL7) with the docked ligand was used. Amino acid

preferences at each position were computed starting from the docked ligand (EETDIW) and scanning each position with each possible residue. The heatmap in Figure 5 shows the FoldX predicted binding energies of each possible mutant.

For DLG2 PDZ3 (PDB: 2HE2), SCRIB PDZ1 (PDB: 2W4F) and MPDZ PDZ10 (PDB: 2OPG), a similar strategy was applied and amino acid preferences are shown in Figure S6. The canonical specificity is approximately recovered, especially for DLG2 PDZ3 and SCRIB PDZ1. In particular, the Ser/Thr at -2 is given a very good score. This is a characteristic feature of canonical PDZ binding modes, where the polar group of Ser/Thr at -2 makes an H-bond with a conserved His in the PDZ domain (Doyle *et al*, 1996). However, the non-canonical specificities derived from phage are often not detected in the structural analysis. To explore the correlations observed in phage, we also generated all possible double-mutants and computed the difference between the predicted energy of the double mutant and the sum of each single mutant. However, as expected from the single mutant analysis, we could not see any clear correlations between the over-represented pairs of residues in the phage data and the ones with better FoldX energies for the double mutant compared with the sum of single mutants (data not shown). We also note that non-canonical specificities have not been observed either in a related study were large datasets of computationally generated random peptides have been positioned in the binding sites of PDZ domains and specificity has been predicted with Rosetta (Smith and Kortemme, 2010).

2. Supplementary Tables

Table S1: Cross-validation AROC values for multiple PWMs, HMMs and ANNs. The last two columns show the results of the multiple PWMs model with one more cluster. 'sd' stands for standard deviations

	Domain	Multiple PWMs AROC	sd	HMM AROC	sd	ANN AROC	sd	Multiple PWMS (K+1) AROC	sd
Human PDZ	DLG1#1	0.973	0.061	0.973	0.061	0.953	0.082	0.977	0.049
	DLG2#3	1	0	0.995	0.014	0.984	0.03	1	0
	ERBIN#1	0.996	0.009	0.995	0.011	0.997	0.007	0.998	0.004
	HTRA2#1	0.877	0.082	0.909	0.094	0.97	0.05	0.850	0.108
	INADL#2	0.994	0.02	0.994	0.02	0.988	0.04	0.988	0.026
	INADL#3	0.945	0.102	0.951	0.105	0.952	0.08	0.943	0.102
	LRRC7#1	0.983	0.037	0.987	0.031	0.979	0.033	0.982	0.041
	MAGI1#5	0.974	0.061	0.978	0.06	0.979	0.048	0.970	0.064
	MAGI3#4	0.997	0.004	0.998	0.003	0.997	0.007	0.996	0.005
	MLLT4#1	0.872	0.063	0.873	0.065	0.909	0.038	0.881	0.048
	MPDZ#1	0.933	0.05	0.932	0.047	0.962	0.035	0.931	0.050
	MPDZ#10	0.993	0.014	0.992	0.018	0.998	0.006	0.992	0.020
	MPDZ#2	0.976	0.035	0.979	0.028	0.983	0.025	0.972	0.047
	MPDZ#5	0.968	0.064	0.964	0.077	0.981	0.037	0.966	0.071
	MPDZ#9	0.98	0.065	0.975	0.075	0.992	0.025	0.981	0.061
	SCRIB#1	0.993	0.009	0.994	0.009	0.987	0.018	0.994	0.009
	TJP2#3	0.989	0.017	0.989	0.017	0.988	0.016	0.986	0.021
Worm PDZ	C33B4.3#1	0.996	0.005	0.996	0.006	0.995	0.008	0.995	0.009
	C43E11.6a#1	0.993	0.013	0.991	0.015	0.991	0.014	0.993	0.011
	C52A11.4#6	0.979	0.03	0.979	0.03	0.983	0.017	0.982	0.034
	F25H2.2#1	0.962	0.049	0.961	0.049	0.972	0.035	0.960	0.047
	K01A6.2#3	0.986	0.046	0.986	0.046	1	0	0.983	0.054
	K01A6.2#5	1	0	1	0	0.995	0.014	1	0
	Y38C1AB.4#2	0.992	0.013	0.992	0.013	0.998	0.005	0.991	0.013
Yeast SH3	Abp1	0.996	0.013	0.998	0.006	1	0	0.998	0.006
	Bbc1	0.988	0.016	0.993	0.013	0.998	0.006	0.991	0.012
	Bem1#2	0.898	0.095	0.883	0.105	0.966	0.031	0.895	0.094
	Boi1	0.976	0.028	0.984	0.026	1	0	0.974	0.032
	Boi2	0.977	0.058	0.986	0.031	0.995	0.014	0.982	0.044
	Lsb1	0.986	0.017	0.986	0.016	0.993	0.01	0.985	0.018
	Lsb3	0.981	0.031	0.985	0.03	0.995	0.016	0.985	0.026
	Myo5	1	0	0.995	0.016	0.995	0.016	0.990	0.032
	Nbp2	0.968	0.062	0.967	0.055	0.997	0.008	0.961	0.074
	Pex13	0.997	0.005	0.996	0.008	1	0	0.996	0.011
	Pin3	0.968	0.039	0.971	0.04	0.999	0.001	0.972	0.039
	Rvs167	0.985	0.022	0.985	0.025	0.997	0.01	0.985	0.021
	Sla1#1_2	0.945	0.05	0.963	0.042	0.954	0.054	0.938	0.059
Human WW	NEDD4#3	0.987	0.031	0.985	0.044	1	0	0.987	0.037
	YAP1#2	0.999	0.001	0.998	0.007	1	0	0.999	0.003
	ARHGAP27#3	0.999	0.001	0.999	0.002	1	0	0.999	0.001

Table S2: Correlation between the SPOT signal and the scores for single PWMs, multiple PWMs and HMMs.

	Domains	Single PWM	p-value	Mixture of PWMs	p-value	HMM	p-value
Yeast	Abp1	0.41	1.7×10^{-11}	0.43	3.6×10^{-12}	0.45	1.2×10^{-13}
SH3	Bbc1	0.21	3.1×10^{-3}	0.24	9.5×10^{-4}	0.25	5.0×10^{-4}
	Bem1#2	0.24	9.6×10^{-2}	0.31	3.3×10^{-2}	0.31	2.9×10^{-2}
	Boi1	0.52	1.1×10^{-5}	0.52	1.1×10^{-5}	0.37	2.8×10^{-3}
	Boi2	0.59	2.6×10^{-2}	0.59	2.5×10^{-2}	0.60	2.2×10^{-2}
	Lsb1	0.34	1.3×10^{-7}	0.44	2.5×10^{-12}	0.39	8.4×10^{-10}
	Lsb3	0.62	0	0.64	0	0.65	0
	Myo5	0.70	0	0.73	0	0.64	0
	Nbp2	0.41	3.9×10^{-9}	0.48	1.1×10^{-12}	0.48	2.5×10^{-12}
	Pex13	0.33	4.2×10^{-7}	0.35	5.7×10^{-8}	0.36	2.2×10^{-8}
	Pin3	-0.03	7.0×10^{-1}	0.15	3.3×10^{-2}	0.13	6.5×10^{-2}
	Rvs167	0.45	2.3×10^{-4}	0.54	5.3×10^{-6}	0.55	4.3×10^{-6}
Human	ERBIN	0.32	7.49×10^{-7}	0.42	7.83×10^{-11}	0.38	2.3×10^{-9}
PDZ	MLLT4	0.26	4.4×10^{-4}	0.30	4.37×10^{-5}	0.34	2.0×10^{-6}
	Average	**0.40**		**0.45**		**0.43**	

Table S3: Benchmarking interactions for Human PDZ domains retrieved from PDZbase (Beuming *et al*, 2005).

PDZ domain	Interacting protein	C-terminus
LRRC7#1	CNKSR2	syiethv
DLG1#1	ATP2B2	hsletsl
DLG1#1	ATP2B4	qsletsv
DLG1#1	PBK	ealetdv
MLLT4#1	PVRL3	srrewyv
MLLT4#1	EPHA7	hgtgiqv
MLLT4#1	PVRL4	ngrghlv
MLLT4#1	PVRL2	srravyv
MLLT4#1	EPHB2	qiqsvev
MLLT4#1	BCR	ilfstev
MLLT4#1	EPHB6	qqgsvev
MLLT4#1	PVRL1	skkewyv
MLLT4#1	EPHB3	qtlpvqv
MPDZ#10	HTR2B	eeqvsyv
MPDZ#10	SSTR3	tmrisyl
MPDZ#10	KIT	llvhddv
MPDZ#10	HTR2C	serissv
MPDZ#10	HTR2A	nekvscv
MPDZ#10	PLEKHA2	nirtsdv
MPDZ#10	PLEKHA1	slpvsdv
SCRIB#1	ARHGEF7	awdetnl
ERBIN#1	ARVCF	qpvdswv
ERBIN#1	PKP4	gspdswv
ERBIN#1	CTNND2	aspdswv

Table S4: AROC values for the benchmarking set of Table S3 for multiple PWMs, HMMs and ANNs.

Domains	**Multiple PWMs AROC**	**p-value**	**HMM AROC**	**p-value**	**ANN AROC**	**p-value**
DLG1#1	0.86	0.02	0.85	0.03	0.81	0.05
LRRC7#1	1	0.04	1	0.04	0.91	0.13
MPDZ#10	0.92	0.0009	0.92	0.0009	0.93	0.0001
ERBIN#1	0.79	0.06	0.79	0.06	0.73	0.11
MLLT4#1	0.62	0.17	0.64	0.14	0.75	0.023
SCRIB#1	0.96	0.07	0.96	0.07	1	0.04

Table S5: Aligned phage display peptides for the three WW domains. X on the left or on the right of the sequence stands for terminal gaps in the alignment.

ARHGAP27#3	NEDD4#3	YAP2#1
KEALPVYPSXXXXXX	FGLPEYVWIXXXXX	KLMMVTRLLPSYXXXXXXX
HVSLPPYIDXXXXXX	KLPPLYGSLLXXXX	KLRMVPRLLPSYXXXXXXX
RGPAPLYWESLXXXX	NSPGYYSLLVXXXX	KLSMVTRLLPSYXXXXXXX
GDMLPSYASHFXXXX	YLPPSYLELFXXXX	KLTMATPLLPSYXXXXXXX
QVELPIYESLYXXXX	GKLPTYQDNWXXXX	KLTMDTRPLPPYXXXXXXX
RDFLPTYEVLFXXXX	NSLPFYWFVTGXXX	KLTMFTRLLPSYXXXXXXX
LVALPSYLMSVXXXX	NSLPFYWFLTGXXX	KLTMLTRLLPSYXXXXXXX
GYGLPAYQVIVXXXX	NSLPFYWLLTGXXX	KLTMVARLLPSYXXXXXXX
CLVVPSYQEWLHXXX	PFLPAYSISSSXXX	KLTMVMRLLPSYXXXXXXX
DTGLPAYVRCTTXXX	AFPPNYRALFATXX	KLTMVTQLLPPYXXXXXXX
GANLPLYTESLVXXX	AGPPNYSELFATXX	KLTMVTQLLPSYXXXXXXX
LAGPPSYLSWERXXX	AGWFTYVELPEYXX	KLTMVTQLPPSYXXXXXXX
LGVLPSYHVGQSXXX	AIPPNYSELFATXX	KLTMVTRLPPPYXXXXXXX
MGRLPAYGEMLSXXX	ATLPGYPLFNFVXX	KLTMVTRLPPSYXXXXXXX
MVGLPSYYATNSXXX	AVPPDYSELFATXX	KLTMVTRLVPSYXXXXXXX
RGLLPYYREVVGXXX	AVPPHYSELFATXX	KLTMVTRPLPSYXXXXXXX
RGVLPSYGVVDIXXX	AVPPNYIEPFATXX	KLTMVTWLLPSYXXXXXXX
TTLPPPYRSYCMXXX	AVPPNYNELFATXX	KLTRVTRLLPSYXXXXXXX
TTLTPPYSSYCMXXX	AVPPNYRELFAMXX	KLTTVTRLLPSYXXXXXXX
XATLPGYPLFNFVXX	AVPPNYRELLATXX	NLTMVTRLLPSYXXXXXXX
XFALPSYPGQQHVXX	AVPPNYSALFATXX	XXGTGRASPPAYLVXXXXX
XLVVPSYDSWLTHXX	AVPPNYSELFARXX	XXVGDRRLPVYYESXXXXX
XWILPSYESGMSYXX	AVPPNYSELFGTXX	XXXGVIRVLPGYSEFXXXX
XWYLPTYGEAKSYXX	AVPPNYSELFPTXX	XXXNPGRLLPLYQVQXXXX
XXALPNYVTAMWQRX	AVPPNYSELFSTXX	XXXVGLSPLPSYVGFXXXX
XXDLPVYLTAMWQLX	AVPPNYSEMFATXX	XXXXIKRLPPSYVEFMXXX
XXGLPAYPCKDYLCX	AVPPNYSEQFATXX	XXXXLNRQLPGYWAQDXXX
XXLLPDYVTAMWQRX	AVPPNYSKLFATXX	XXXXSAILPPDYFQMYXXX
XXPLPAYRSAMSSYX	AVPPSYSELFATXX	XXXXSVHILPTYDLRLXXX
XXRLPFYSEGVWRMX	GMLPPYSMSDICXX	XXXXVTARPPVYVSYYXXX
XXRLPHYYSVIQMGX	GVPPNYSELFATXX	XXXXWRAVLPSYQVAIXXX
XXSLLAYSRLPAYIX	IGPVYMDYTYEGXX	XXXXXFRASPEYVYGSIXX
XXSLPPYYLSPPPMX	RGLPVYSTILWMXX	XXXXXGGILPSYTVWSMXX
XXVLPDYDMAMWQRX	RKLPPYWFWRANXX	XXXXXGRALPQYVCESYXX
XXVLPDYITAMWQRX	RKLPSYWFWRANXX	XXXXXGRALPQYVCGSCXX
XXVLPDYVAAMWQRX	SKFPSYWFWRANXX	XXXXXKTLPPPYSSYCMXX
XXVLPDYVMAMWQGX	SKLPSYLFWRANXX	XXXXXRGFAPLYRALQLXX
XXVLPDYVMAMWQRX	TGLPGYRDYINVXX	XXXXXRGLAPHYRALQLXX
XXVLPDYVRAMWQRX	TVPPNYSELFATXX	XXXXXRGLAPLYRALQRXX
XXVLPDYVRARGKGX	VTLPQYSALREVXX	XXXXXRGLAPLYRTLQLXX
XXVLPDYVRAWWQGX	WGAPAYYQIEDVXX	XXXXXRGLAPLYWALQLXX
XXVLPDYVTAMGQRX	XALPPYVTELLPTX	XXXXXRGLAPVYRALQLXX
XXVLPDYVTAMSQRX	XAVPDYVTAMWQRX	XXXXXRGLGPLYRALQLXX
XXVLPDYVTAMWERX	XKVPSYTVQLAQEX	XXXXXRGLPPLYRALQLXX
XXVLPDYVTAMWHRX	XMPPGYPVDCSAAX	XXXXXRRSAPGYYTLIQXX
XXVLPDYVTAMWQRX	XPLPLYTNVECLLX	XXXXXTTLLPPYSSYCMXX
XXVLPDYVTSMWQRX	XSLPDYVTAMWQRX	XXXXXTTLPPHYSRYCMXX
XXVLPNYSMVSSSFX	XSLPIYAMSGTSWX	XXXXXTTLPPPYRRYCMXX
XXVLPNYWAVEEFSX	XTVPSYHPLWEMEX	XXXXXTTLPPPYRRYSMXX
XXWLPAYYAPTLPWX	XVLPAYVTALWQTX	XXXXXTTLPPPYRSYCMXX

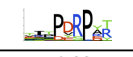
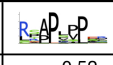
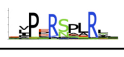
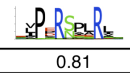
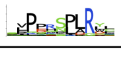
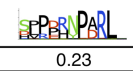
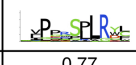
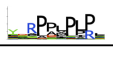
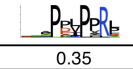
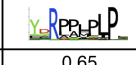
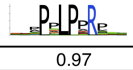
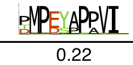
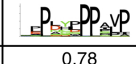
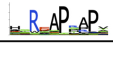
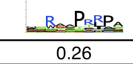
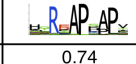
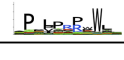
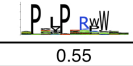
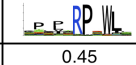
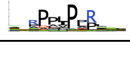
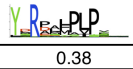
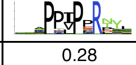
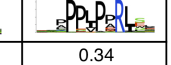
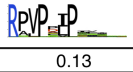
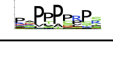
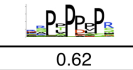
Table S5 (second part):

ARHGAP27#3	NEDD4#3	YAP2#1
XXXLPEYVSDRGWEL	XVLPAYVTAMWQRX	XXXXXTTLPPPYSRYCMXX
XXXLPFYPGMINAAQ	XVLPDYVRAMWQRX	XXXXXTTLPPPYSSYCIXX
XXXLPGYPHQYNDSQ	XVLPDYVTAMSQRX	XXXXXVRQLPEYHAMNVXX
XXXLPGYTDFPAWAI	XVVPAYHAMVHNEX	XXXXXXAVPPDYSELFATX
XXXLPHYWAQDKVIV	XVVPAYHDMAHNEX	XXXXXXAVPPNYSALFATX
XXXLPIYGHAVDITV	XXPPLYQQIRSVDF	XXXXXXAVPPNYSELFAAX
XXXLPKYMGPILVDS	XXPPMYQQFCSVDF	XXXXXXAVPPNYSELLATX
XXXLPLYADRGPVVI	XXPPMYQQFRPVDF	XXXXXXAVPPNYSELSATX
XXXLPLYSNFYGFSV	XXPPMYQQFRSDGF	XXXXXXAVPPNYSVLFATX
XXXLPPYSISNIGYW	SRLPMYGDVLMXXX	XXXXXXAVPQYYRALCATX
XXXLPRYSRATSAIL		XXXXXXAVSPNYSELFATX
XXXLPSYQRHALGFV		XXXXXXCKLPSYWFWRANX
XXXLPTYGSHHCMML		XXXXXXGKLPSYWFWRANX
XXXLPVYGSWAMDQT		XXXXXXGMLPLYGDIYATX
XXXLPVYWLGNEVVQ		XXXXXXRALPIYSYLASNX
XXXQLVYTALPGYEL		XXXXXXRALPRYTDIGTAX
XYGLPGYSDIVNIXX		XXXXXXRLAPLYYDVVKGX
YGWLPAYSNTQXXXX		XXXXXXRLPPSYWWRDFWX
		XXXXXXRLPVYMETMNPGX
		XXXXXXRLSPPVYSESWWX
		XXXXXXRSLPVYEASIDAX
		XXXXXXRVLPNYSRAIYYX
		XXXXXXSKLPSYWFWRAHX
		XXXXXXSKLPSYWFWRASX
		XXXXXXSKLPSYWFWRVNX
		XXXXXXSLLPPYGYNHTLX
		XXXXXXTVPPNYSELFATX
		XXXXXXXALPDYVTAMWQG
		XXXXXXXALPEYVTAMWQR
		XXXXXXXALVVLGLPGYVW
		XXXXXXXAVPAYVNAMLQP
		XXXXXXXAVPAYVRAMCQR
		XXXXXXXAVPDYVTAMWQR
		XXXXXXXGLPDYVTAMWQR
		XXXXXXXHLPAYGNLPFFV
		XXXXXXXVLPDYVKTMRQR
		XXXXXXXVLPDYVTAIWQR
		XXXXXXXVLPDYVTAMWKR
		XXXXXXXVLPDYVTAMWQG
		XXXXXXXVLPDYVTAMWQW
		XXXXXXXVLPDYVTSMWQR
		XXXXXXXVLPHYVTAMWQR
		XXXXXXXXPPAYRVITGNI
		XXXXXYHILPGYSNMPWXX
		XXXYGWRALPAYEGLXXXX
		XXYMSTGLLPSYSLXXXXX

3. Supplementary Figures

A

Yeast SH3 domains

Domain	Phage Peptides	single PWM	Mixture of PWMs		
Abp1	82				
mixing coefficients			0.79	0.21	
Bbc1	101				
mixing coefficients			0.95	0.05	
Bem1 #2	31				
mixing coefficients			0.22	0.52	0.26
Boi1	115				
mixing coefficients			0.81	0.19	
Boi2	22				
mixing coefficients			0.23	0.77	
Lsb1	67				
mixing coefficients			0.35	0.65	
Lsb3	139				
mixing coefficients			0.97	0.03	
Myo5	19				
mixing coefficients			0.22	0.78	
Nbp2	46				
mixing coefficients			0.26	0.74	
Pex13	106				
mixing coefficients			0.55	0.45	
Pin3	98				
mixing coefficients			0.38	0.28	0.34
Rvs167	136				
mixing coefficients			0.13	0.87	
Sla1 #1/2	29				
mixing coefficients			0.62	0.38	

B

Human WW domains

ARHGAP27#3	68				
			0.67	0.27	0.06
NEDD4#3	60				
			0.40	0.23	0.15
			0.15	0.07	
YAP1#2	96				
			0.41	0.15	0.14
			0.11	0.10	0.09

Figure S1: (previous page) Sequence logos for all SH3 and WW domains exhibiting multiple specificity. Both the single PWM and the multiple PWMs together with their weights are shown.

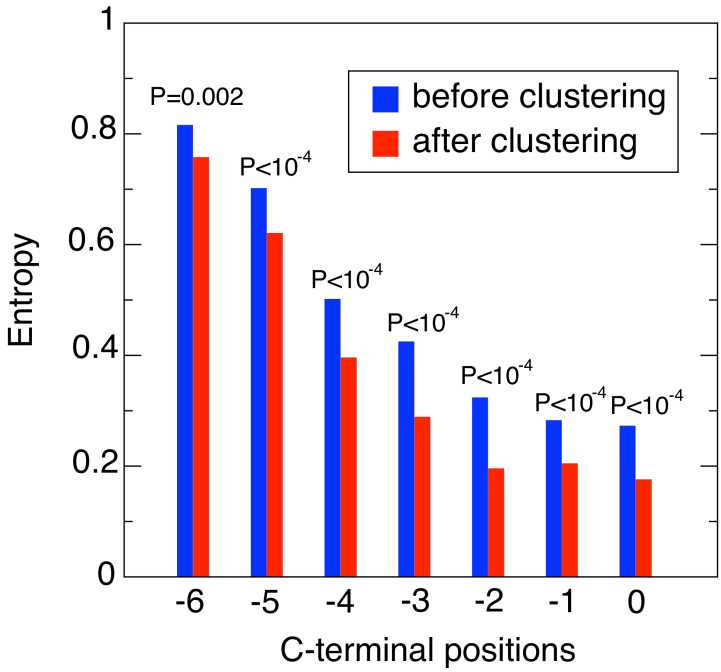


Figure S2: Comparison of the entropy at each position before and after clustering the peptides. Blue bars show the average entropy over all worm and human PDZ domain interacting peptide sets at each position. Red bars show the average entropy over all clusters at each position (i.e., the entropy of each cluster has been weighted by the cluster weight before averaging over all domains). P-values were generated by randomizing the clusters.

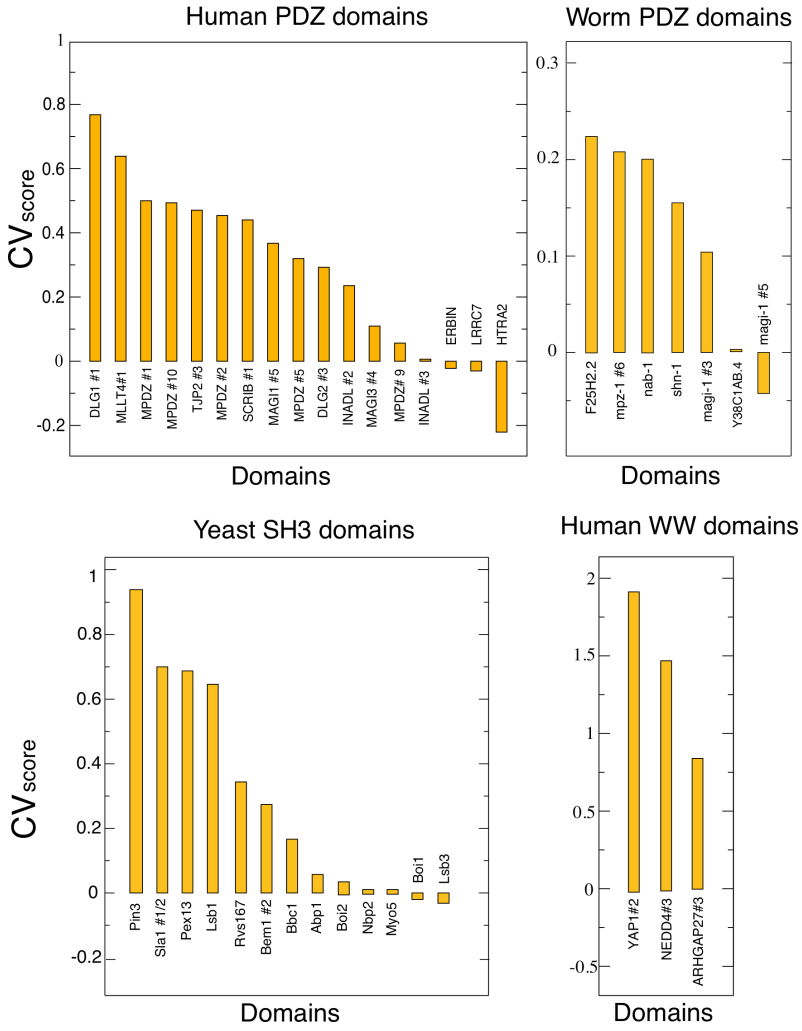


Figure S3: Cross-validation scores for domains displaying multiple specificity. Positive values mean that the peptides in the testing sets are given a higher score with the multiple PWMs than with the single PWM.

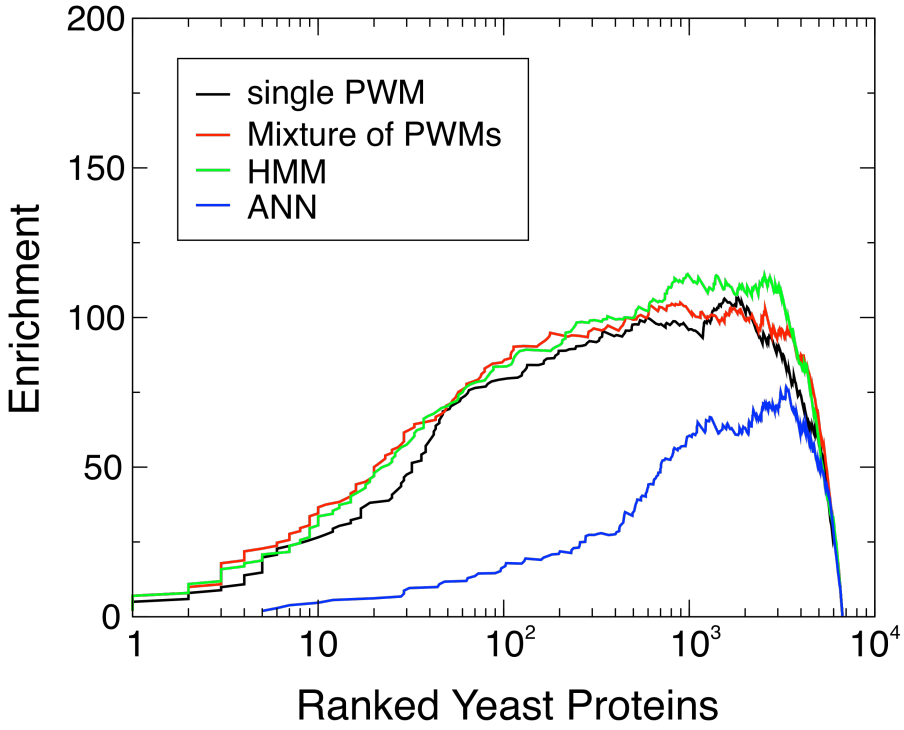


Figure S4: Comparison with Y2H data. The different curves show the enrichment in Y2H interactions among the best scoring Yeast proteins. Black curve corresponds to single PWM scoring, red curve to multiple PWMs, green curve to HMMs and blue curve to ANNs.

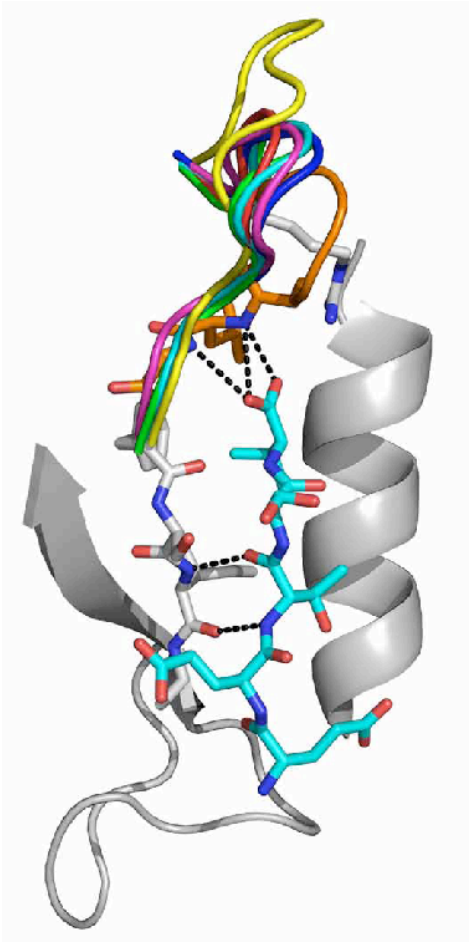


Figure S5: Structural alignment of PDZ domains with a displaced carboxylate binding loop. The canonical position of the loop is shown in orange and comes from the structure of DLG3 PDZ1 in complex with a C-terminal ligand (PDB: 2I1N). The yellow loop corresponds to par-6 (PDB: 1X8S), the green one to DLG2 PDZ1 (PDB: 2WL7), the blue one to DLG1 PDZ2 (PDB: 2X7Z), the red one to DLG3 PDZ2 (PDB: 2FE5), the cyan one to SYNJ2BP (PDB: 2JIK) and the pink one to MAGI1 PDZ4 (PDB: 2Q9V).

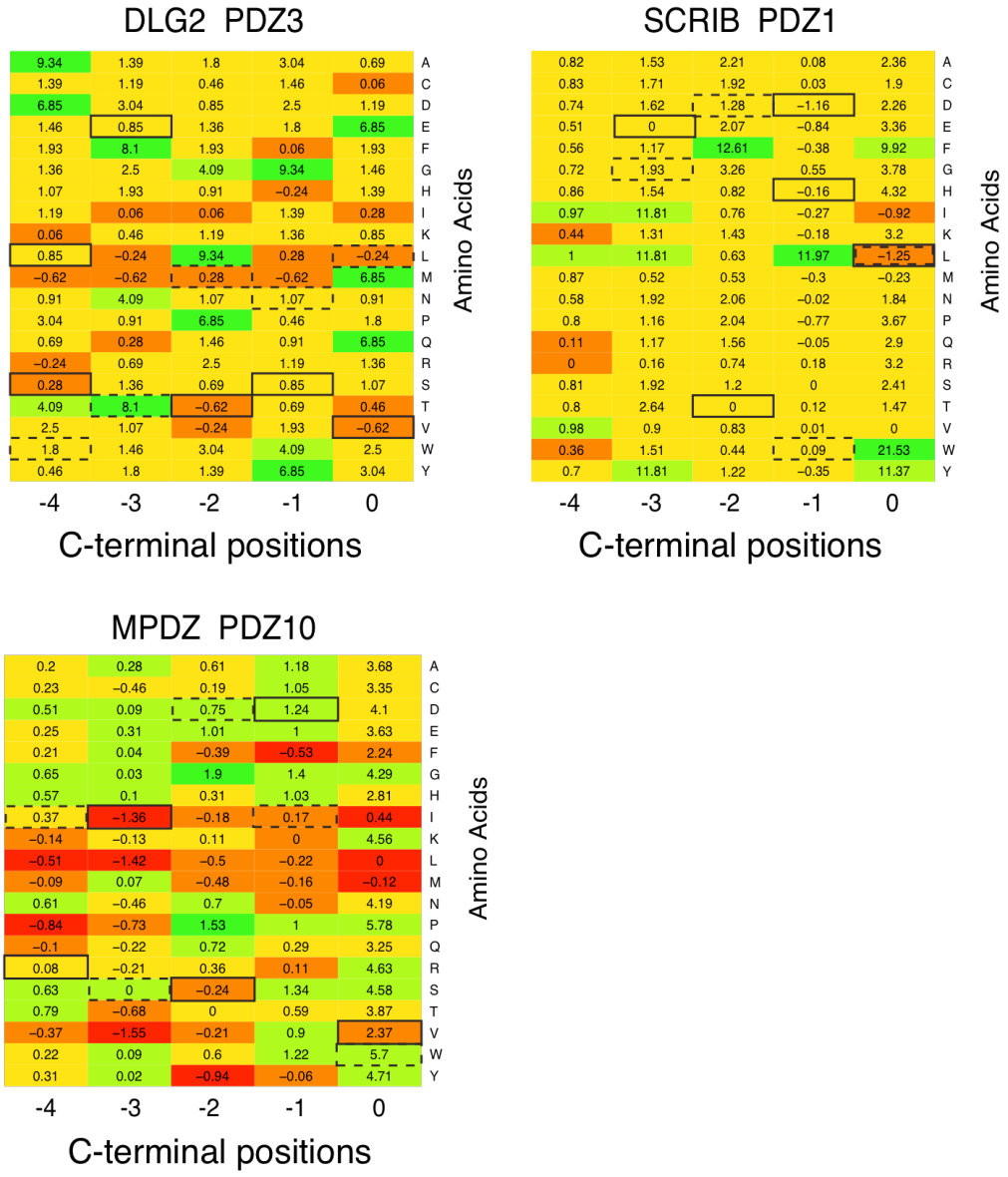


Figure S6: Structure-based predictions of residue preferences for DLG2 PDZ3, MPDZ PDZ10 and MPDZ PDZ10. Numbers on the heatmap correspond to predicted free-energies of the single mutants. Residues following the canonical specificity observed in phage are shown with the black boxes and the ones following the non-canonical specificity are shown in dashed boxes.

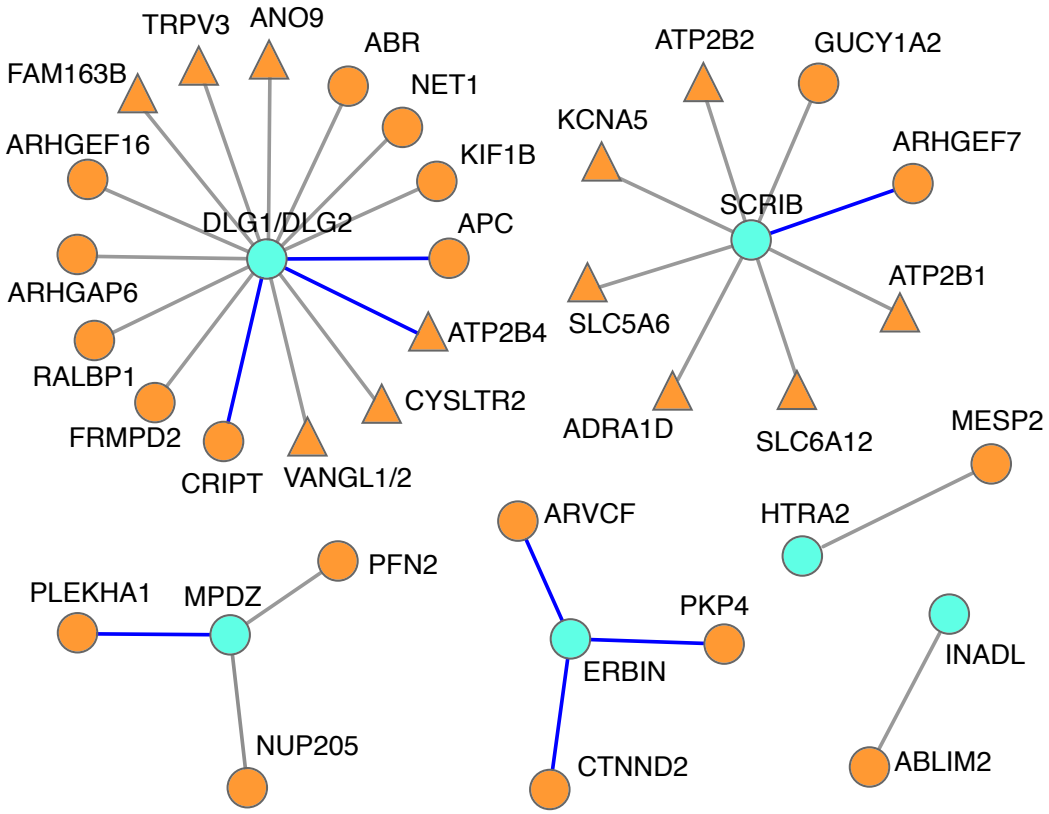


Figure S7: Human protein interaction predictions for multi-specific PDZ domains. The network displays all interactions predicted by the multiple PWMs (see Material and Methods). Blue edges correspond to interactions known from previous studies. Triangles represent membrane proteins. Circles stand for non-membrane proteins. Cyan nodes correspond to PDZ-containing proteins (preys in MYTH).

4. References

Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2:** 28-36.

Beuming T, Skrabanek L, Niv MY, Mukherjee P, Weinstein H (2005) PDZBase: a protein-protein interaction database for PDZ-domains. *Bioinformatics* **21:** 827-828.

Bishop CM (2006) *Pattern Recognition and Machine Learning*, Vol. 1. Singapore: Springer.

Doyle DA, Lee A, Lewis J, Kim E, Sheng M, MacKinnon R (1996) Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. *Cell* **85:** 1067-1076.

Miller ML, Jensen LJ, Diella F, Jorgensen C, Tinti M, Li L, Hsiung M, Parker SA, Bordeaux J, Sicheritz-Ponten T, Olhovsky M, Pasculescu A, Alexander J, Knapp S, Blom N, Bork P, Li S, Cesareni G, Pawson T, Turk BE, Yaffe MB, Brunak S, Linding R (2008) Linear motif atlas for phosphorylation-dependent signaling. *Sci Signal* **1:** ra2.

Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L (2005) The FoldX web server: an online force field. *Nucleic Acids Res* **33:** W382-388.

Sharon E, Lubliner S, Segal E (2008) A feature-based approach to modeling protein-DNA interactions. *PLoS Comput Biol* **4:** e1000154.

Smith CA, Kortemme T (2010) Structure-Based Prediction of the Peptide Sequence Space Recognized by Natural and Synthetic PDZ Domains. *J Mol Biol*.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102:** 15545-15550.

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 (2009) Bayesian modeling of the yeast SH3 domain interactome predicts spatiotemporal dynamics of endocytosis proteins. *PLoS Biol* **7:** e1000218.

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 (2008) A specificity map for the PDZ domain family. *PLoS Biol* **6:** e239.

Wiedemann U, Boisguerin P, Leben R, Leitner D, Krause G, Moelling K, Volkmer-Engert R, 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.