Published online 23 July 2011 *Nucleic Acids Research, 2011, Vol. 39, No. 19* ***e128***
doi:10.1093/nar/gkr600

Yeast two-hybrid junk sequences contain selected linear motifs

Yun Liu[1], Nicholas T. Woods[2], Dewey Kim[1], Michael Sweet[2], Alvaro N. A. Monteiro[2] and Rachel Karchin[1,*]

[1]Department of Biomedical Engineering and Institute for Computational Medicine, Johns Hopkins University, 3400 N. Charles St, Baltimore, Maryland and [2]Cancer Epidemiology Program, Population Sciences Division, H. Lee Moffitt Cancer Center, Tampa, Florida USA

Received March 14, 2011; Revised July 4, 2011; Accepted July 6, 2011

ABSTRACT

Yeast two-hybrid (Y2H) screenings result in identification of many out-of-frame (OOF) clones that code for short (2-100 amino acids) peptides with no sequence homology to known proteins. We hypothesize that these peptides can reveal common short linear motifs (SLiMs) responsible for their selection. We present a new protocol to address this issue, using an existing SLIM detector (TEIRESIAS) as a base method, and applying filters derived from a mathematical model of SLiM selection in OOF clones. The model allows for initial analysis of likely presence of SLiM(s) in a collection of OOF sequences, assisting investigators with the decision of whether to invest resources in further analysis. If SLiM presence is detected, it estimates the length and number of amino acid residues involved in binding specificity and the amount of noise in the Y2H screen. We demonstrate that our model can double the prediction sensitivity of TEIRESIAS and improve its specificity from 0 to 1.0 on simulated data and apply the model to seven sets of experimentally derived OOF clones. Finally, we experimentally validate one SLiM found by our method, demonstrating its utility.

INTRODUCTION

Most critical cellular processes, including signaling, metabolism and proliferation, depend on the binding of proteins to each other and to small molecules such as nucleic acids, peptides and metals. Many experimental techniques have been developed to study protein binding: yeast two-hybrid (Y2H), phage display, co-immunoprecipitation and co-crystallization of potential binding partners. The Y2H method is based on a fusion between a protein bait of choice and a DNA binding domain (DBD) of a transcription factor (e.g. yeast GAL4), and a fusion between an activation domain (AD) of a transcription factor and a protein prey of choice (1). To perform screenings, a library of cDNA clones is fused to the AD. Upon interaction between the bait and the prey, the DBD and AD are brought in close proximity, leading to transcription of a reporter gene that allows for clone selection. These screenings generally yield from a few to hundreds of hits, which are then sequenced, translated with a computer program, and mapped to known proteins. Strikingly, in any Y2H experiment, up to 67% of the clones lack homology to any known protein and are considered false positives. They are subsequently discarded. These clones represent out-of-frame (OOF) clones generated when cloning restriction-digested cDNAs as a fusion to the AD during library construction. While OOF clones tend to be short (≤20 amino acids) and have amino acid residue compositions that resemble random sampling from the codon table, selected OOF clones show an overrepresentation of hydrophobic residues. These hydrophobic residues generate 'sticky' peptides that bind spuriously to the bait. Thus, OOF clone hits are considered to be false positives (2) in a Y2H screening experiment, an assumption that we re-examine here.

Protein–protein interactions can happen mediated by surface–surface contacts or, alternatively, between protein modular domains and short linear motifs (SLiMs) (3). Modular domains are structurally conserved regions of approximately 100 amino acid residues that can fold independently (4). Specific interaction sites where proteins bind may be identified by the presence of short linear motifs (SLiMs) in the protein sequence. There is considerable evidence that these SLiMs are under evolutionary selection (5–8), although the selection is weaker than selection shaping protein domain evolution (7). It has been suggested that SLiMs may be the products of convergent, rather than divergent evolution (7). SLiMs are known to

*To whom correspondence should be addressed. Tel: +1 410 5165578; Fax: +1 410 5165294; Email: karchin@jhu.edu be more abundant in protein disordered regions or loops (9) but are also found within functional domains (10).

Many bioinformatics methods have been developed or applied to detect SLiMs (6,10–15). Detection methods include statistical analysis of overrepresented sequence patterns, evolutionary conservation, protein interaction data and sequence or structural similarity.

Given the high percentage of OOF clone hits from Y2H, we hypothesize that some of these sequences have specific affinity to the protein bait mediated by SLiMs, provided that the bait is (or contains) a modular domain, and contain SLiMs. These SLiMs may be present in the human proteome but not be over-represented in the Y2H in-frame hits because of their sparsity and because libraries may constitute a biased representation of the proteome due to methodological issues. Thus, they will be missed if SLiM search is limited to the in-frame hits. It is tempting to speculate that OOF sequences may even contain novel SLiMs.

OOF sequences have different amino acid composition and length distributions than natural proteins and peptides, because they were not subject to the same evolutionary pressures for protein functionality. Because existing SLiM detection methods were not designed for OOF sequences, we reasoned that modifications of these methods could be useful to achieve improved detection of SLiMs in the OOF setting.

Here, we describe a new approach to identify signatures of selection (SOS) in OOF sequences by modeling the length distribution of OOF clones that bind to protein module baits in a Y2H experiment. This signature can then be used to predict the presence of SLiMs in OOF clones and several biologically relevant properties. We demonstrate that our SOS model can be used to filter and rerank SLiM hits from bioinformatics methods designed for the in-frame setting. With this approach, and using TEIRESIAS (11) as a base method, we show initial simulation results, which indicate that our protocol may be more effective than existing methods at finding SLiMs in a high-noise OOF setting. With this approach, we are able to double the prediction sensitivity of TEIRESIAS (11) and improve its specificity from 0 to 1.0 on simulated data. We also apply the SOS model and several existing SLiM detection methods to sets of experimentally-derived OOF clones from Y2H experiments in which tandem BRCT domain protein modules from seven human genes were used as baits. SOS, TEIRESIAS and SLiMFinder (12) detected the well-known `SPXF` motif in the OOF clones from the BRCA1 Y2H screen. SOS detected a novel SLiM `KKKKKK` in OOF clones from the LIG4 Y2H screen, which we have experimentally validated with a Y2H direct binding assay.

MATERIALS AND METHODS

Assessment of SLiM detectors on OOF sequences

We developed a Monte Carlo-like algorithm to generate benchmark sets that would mimic the biological production of OOF sequences, either containing SLiMs or not, and for which we would know the correct answer (Figure 1). We applied the algorithm to simulate OOF sequences bound to BRCT tandem domain baits from six human genes, for which we had performed Y2H screens. The algorithm generated OOF sequences by sampling nucleotides according to their background frequencies in the human transcriptome (A = 0.261, C = 0.245, G = 0.245, T = 0.248) NCBI RefSeq database, Version 39, retrieved 26 February 2010 (16). For each artificial sequence, generation was terminated when a stop codon was produced in the correct reading frame. Nucleotide sequences were then translated and the process was repeated until a desired number of sequences was generated. The Monte Carlo-like algorithm currently does not model the redundancy that occurs in real OOF datasets, which is not well

```
GEN-OOF-SEQ($pA,pC,pG,pT$)
$s \leftarrow$ NIL
$c \leftarrow 1$
while $s[c-1] \notin \{TAA,TAG,TGA\}$
do $x \leftarrow Sample(3,\{A,C,G,T\},pA,pC,pG,pT)$
    $s \leftarrow s \cup x$
    $c \leftarrow c+1$
$s' \leftarrow$ Translate(s)
return $s'$

MOTIF-SEARCH($s,t,r,g$)
if $t$ is $Subsequence(s)$ and $r \leq g$
  then return TRUE

MAIN($n,m,o,g,pA,pC,pG,pT$)
 $B \leftarrow$ NIL
 $i \leftarrow 0$
 repeat
         $s \leftarrow$ GEN-OOF-SEQ($pA,pC,pG,pT$)
         $r \leftarrow$ RANDOM$(0,1)$
         $f \leftarrow$ FALSE
         for $j \leftarrow 1$ to $Length(m)$
         do $f \leftarrow$ MOTIF-SEARCH($s,m[j],r,g$) OR $f$

         if $f =$ FALSE AND $r \leq o$
           then $f \leftarrow$ TRUE

         if $f =$ TRUE
           then $B \leftarrow B \cup s$
                $i \leftarrow i+1$

    until $i = n$
 return $B$
```

Figure 1. Monte Carlo-like algorithm to generate an *in silico* benchmark set of OOF sequences from a Y2H screen, with known motifs, desired noise level, and estimated sensitivity of the Y2H screen to the motifs. The resulting sequences simulate OOF clones that bind to a protein module bait of interest in a Y2H screen. Inputs to the algorithm are *n*, the number of desired OOF sequences; *m*, a list of binding motifs known to bind to the bait of interest; *o*, desired noise level; *g*, estimated sensitivity of the Y2H screen to the motifs (e.g. the frequency with which the Y2H will bind an OOF clone if the motif is present). With a slight modification, the algorithm can accept a motif-specific *g* value (*g* may differ among motifs in the *m* list, i.e. a longer motif may be more likely to be picked up in a Y2H screen than a shorter one.) *pA.pC.pG.pT* = background frequencies of nucleotides in (human) transcriptome. We quantify noise as a length-independent selection rate (LISR), expected frequency of a reported binding interaction in a Y2H screen when there is no SLiM present in the clone. At LISR of 5*E*6, only one out of 200 000 clones that do not contain a SLiM will bind to the Y2H bait. At LISR of 5*E*1 one out of every two clones that do not have a SLiM will bind to the bait.

understood biologically and varies widely from library to library.

Positive and negative control sets. First, we generated a positive control set, in which we required that the sampling process yield a well-known SLiM, responsible for binding of the BRCA1 BRCT tandem domains to phosphorylated peptides (pSPXF, where X can be any amino acid). After 100 000 sequence generation attempts, the sampler produced 300 artificial OOF sequences containing this SLiM. To generate the negative control, we repeated the sampling process to yield an equal number of sequences, without filtering for SPXF. These artificial sequence sets model results of Y2H screens with BRCA1 BRCT tandem domain baits in which there was perfect selection for the SPXF SLiM (positive control) and no selection for the SPXF SLiM (negative control). Biologically, SPXF must include a phosphorylated serine. We make the simplifying assumption that the SLiM is phosphorylated *in vivo*, based on the conservation of kinases and their substrates in yeast and mammalian cells.

More realistic artificial OOF sequences. Real Y2H screens are noisy and are unlikely to yield perfect selection for a SLiM of interest. We quantified the noise of a Y2H experiment in terms of LISR (expected frequency of reported binding interaction when true binding SLiM is absent) and (LDSR) or Length-Dependent Selected Rate, the expected frequency of reported binding interaction when true binding SLiM is present, denoted as g in Figure 1. At the lowest noise level, the LISR is $5E^{-6}$, meaning that only one out of 200 000 sequences that do not contain the SLiM will bind to the Y2H bait. At the highest noise level (LISR = $5E^{-1}$) one out of every two sequences that do not have a SLiM will bind to the bait. Furthermore, some OOF sequences may contain more than one SLiM. To make our benchmarks more realistic than the positive and negative controls, we selected 11 SLiMs derived from experimentally validated binding sequences in tandem BRCT domains (17–19) (Supplementary Table S1). The Monte Carlo-like algorithm was used to create benchmarks at LISR ranging from $5E^{-6}$ to $5E^{-1}$ and LDSR of 0.99. For each benchmark, we ran the algorithm until we had 300 sequences containing at least one valdiated SLiM for the BRCT bait of interest. For the BRCA1 and MDC1 benchmarks, more than one selected SLiM was generated in some cases. We provide details of the SLiMs in each benchmark in Supplementary Table S3. Some benchmarks contain hundreds SLiM copies and some contain no SLiMs. While the LISR noise levels were evenly distributed on a log scale, SLiM abundances were not evenly distributed. These seeming discrepancies occurred because because the algorithm incorporates stochastic variation that we expect to see in real OOF datasets. The size 300 was chosen because it is close to the average number of OOF sequences identified in our seven Y2H screens. The choice of LDSR = 0.99 was based on the assumption that if a SLiM is present in an OOF clone, it was highly likely to be picked up by Y2H.

SLiM detection method comparisons. We applied four existing methods of SLiM detection to these *in silico* OOF benchmarks: TEIRESIAS (11) webserver (minimum number of specific amino acids was changed from 3 to 2 since some SLiMs have only two specific positions); GLAM2 (13) (default parameters and each input sequence permuted 100 times to obtain empirical P-values. Hits were assessed as significant only if the score of the hit was greater than at least 99% of the best scores from the permuted sequences); SLiMFinder v4.1 (12) (default parameters and with evolutionary filtering and masking turned off); and the DILIMOT (6) webserver (http://DILIMOT.embl.de) (minimum number of specific amino acids set to 2 as above).

Each method was tested on the positive and negative control sets described above and the 36 'more realistic' sets of sequences containing multiple SLiMs and varying levels of noise.

We assessed:

- *SLiM detection sensitivity*: fraction of datasets out of 34 (one positive control and 33 out of 36 datasets in which the noise level was sufficiently low to produce at least one SLiM) in which the correct SLiM appeared anywhere in the method's ranked list of detected SLiMs.
- *SLiM prediction sensitivity*: fraction of datasets out of 34 (as above) in which the correct SLiM appeared in the top 10 ranked SLiMs identfied by the method.
- *Specificity*: fraction of datasets out of four (one negative control, 3 out of 36 datasets in which the noise level was sufficiently high so that no SLiMs were generated) with no detected SLiMs.

Overall, TEIRESIAS had the highest *detection sensitivity*. SLiMFinder had the highest *prediction sensitivity*. SLiMFinder, GLAM2 and DILIMOT had perfect specificity, while TEIRESIAS had the lowest specificity (Table 1, details in Supplementary Table S2). While the benchmarks are not comprehensive and our conclusions may not generalize beyond these six baits, these tests provided initial feedback about possible limitations, in the OOF setting, of methods that were designed for SLiM detection in in-frame sequences and functional proteins.

Table 1. Initial assessment of SLiM detectors and SOS model on 38 simulated OOF datasets

	TRS	SF	G2	DLM	SOS
Detection sensitivity[a]	0.94	0.59	0.35	0	**0.79**
Prediction sensitivity[b]	0.38	0.59	0.35	0	**0.74**
Specificity[c]	0	1	1	1	**1**

TRS = TEIRESIAS, SF = SLiMFinder, G2 = GLAM2, DLM = DILIMOT. SOS = Signatures of Selection model. The results of using the SOS model to post-process TRS are highlighted in bold.

[a]Fraction with at least one SLiM in which the correct SLiM appeared in the method's ranked list of predicted SLiMs.

[b]Fraction in which the correct SLiM appeared in method's top 10.

[c]Fraction (with no SLiMs) in which no SLiM was reported by the method. See Supplementary Table S2 for more detailed results.

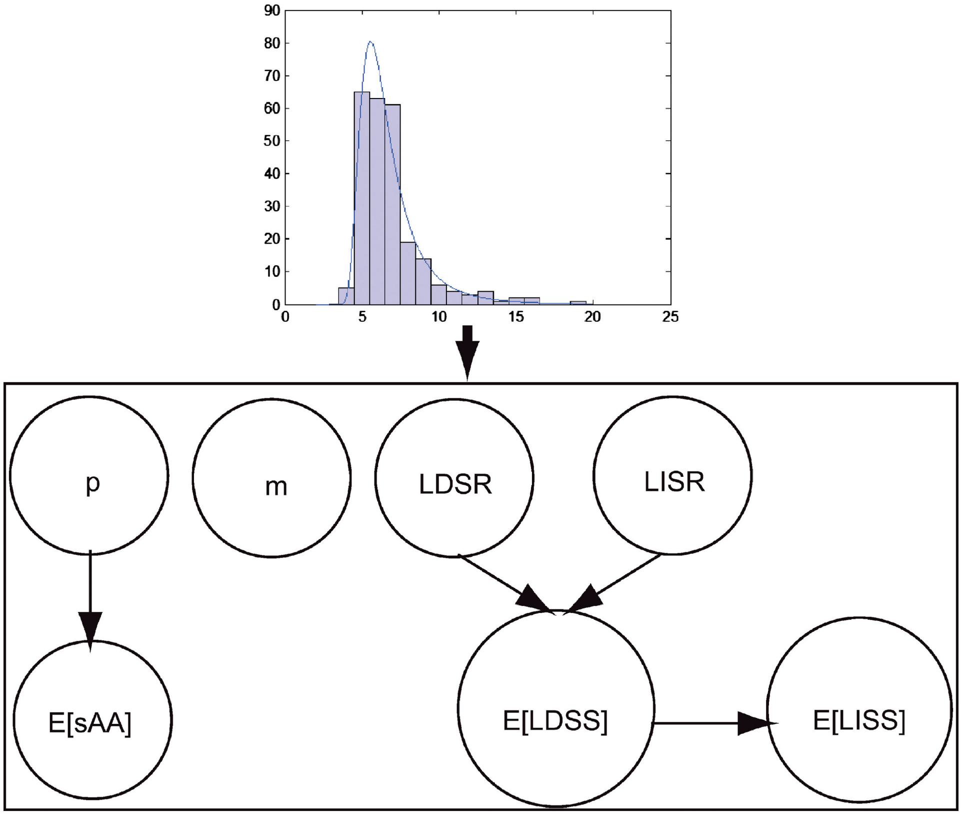

Figure 2. The parameters of the SOS model are fit based on the distribution of bait-bound OOF clone sequence lengths from a Y2H experiment. *p* is the frequency of the most abundant SLiM in the sequences. *m* is the length of the most abundant SLiM, LDSR is the expected fraction of bound OOF clones containing at least one SLiM, LISR is the expected fraction of bound OOF clones with no SLiMs. *E*[*sAA*] is the expected number of selected amino acid residue positions in the most abundant SLiM necessary and sufficient for binding specificity. *E*[*LDSS*] is the expected number of bound OOF sequences with at least one SLiM. *E*[*LISS*] is the expected number of bound OOF sequences that do not contain SLiMs.

Next we describe our new approach (SOS) to improving detection of SLiMs in OOF sequences.

Improving SLiM detection in OOF sequences

We developed a probabilistic model to estimate the presence and several key properties of SLiMs in a collection of OOF sequences. First we describe the model, then discuss its utility in detecting OOF SLiMs when used in conjunction with TEIRESIAS (mathematical and implementation details provided in the Appendix 1).

The SOS probabilistic model. The SOS model is based on the length distribution of OOF sequences bound to a Y2H bait of interest. In the absence of selection for a particular SLiM, the lengths of the OOF sequences follow a geometric distribution, with fixed parameter *s*, which is the probability of a stop codon when there is no evolutionary pressure for functional protein. The probability distribution is on the number *X* of Bernouli trials before the first 'success'. Here each trial is a randomly generated sequence of three nucleotides, or codon, when there is no selection (evolutionary, binding or otherwise) and success is defined as the occurrence of a stop codon. If a sequence contains a selected SLiM, the geometric distribution is modified by adding parameters for: *SLiM frequency* (*p*), *SLiM length* (*m*), the expected frequencies with which the Y2H screen reports a binding interaction when a true binding SLiM is present (LDSR) or when it is absent (LISR). These four parameters are fit using the length distribution of a set of OOF sequences (Figure 2): *E*[*sAA*] (expected number of selected amino acid residue positions necessary and sufficient for binding specificity in the most abundant SLiM), *E*[*LDSS*] (expected number of bound OOF sequences with at least one SLiM) and *E*[*LISS*] (expected number of bound OOF sequences that do not contain SLiMs) are derived from these four parameters (Appendix 1).

Maximum likelihood estimation of parameters. For each dataset of OOF sequences, we applied standard optimization techniques to obtain maximum likelihood estimates (MLE) of the parameters for the (modified geometric) distribution of OOF sequences containing SLiMs under selection.

When applied to our artificial sequence datasets (for which we know the number of SLiMs and the noise levels), we observed that the shape of the four-dimensional negative log likelihood function (NLL) (used for MLE) (Equation A6) corresponded to presence of SLiMs and noise levels. Sharp minimums in the NLL were a signature of SLiM presence and low noise (Figure 3).

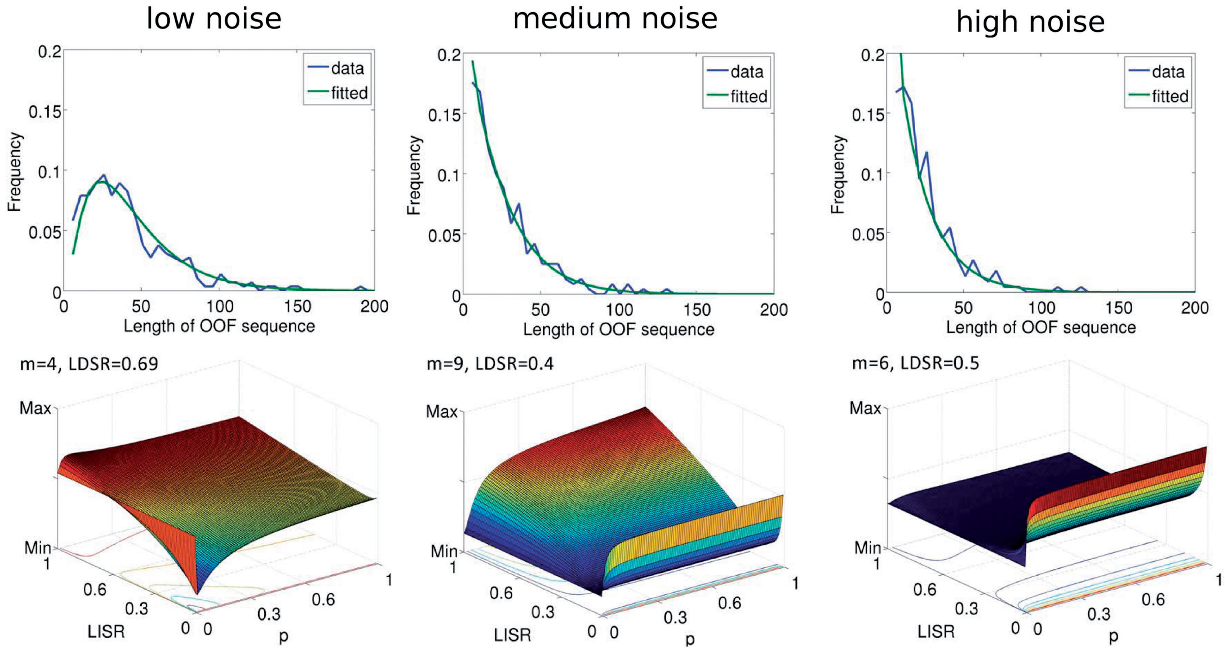

Figure 3. Length distributions of artificial Y2H OOF sequences generated to include one SLiM SFDK and with low, medium and high noise levels (Supplementary Table S3). The impact of adding noise to the sequence generation process can be visualized by slices of the resulting 4D negative log likelihood functions (NLL) (Equation A.6) used to fit the SOS model. Sharp minimums in the NLL are a signature of the presence of SLiMs in the sequences and low noise in the Y2H OOF screen. *Z*-axis shows NLL value. *X*- and *Y*-axes show values of *p* and LISR, respectively. For visualization purposes, the 4D NLL surface is sliced at the MLE estimates of *m* and LDSR. Low noise = LISR = $5E^{-6}$. Medium noise = LISR = $5E^{-4}$. High noise = LISR = $5E^{-1}$ (Figure 2).

```
>1
 QNHRMQWIFKMEMDRRL
>2
LSYNYIVEVKIIITESYLNNQL
>3
GECSISASLKAGPGSAGPIVHLVTSFLPLFLDESKARKAVMESSCTSSCPT EKPARVLPGTGPLPPYSICKQGRSRKRTSPKGSIRI
AIYTGAEKVLMSPSFL ESKDQPEAVGAAAQAVVNWRETGPGFSVRMWI
>4
LSPDQLSHEAGFSLSPTFVPQPSSKAHVAAETCCFCWGQAPLPPAPGASPSFLQPRPLRLDPGPWSLGDSPHPSPTPAPSLML
>5
GQASRNRISLLPSFHLPQGRIPSSPTFPTVGLKTFQRGSCHTHWGKEC
```

Figure 4. Five OOF clones from Y2H experiment with BRCA1 tandem BRCT-domain bait. Examples of clones whose translated protein sequence could not be mapped to the human proteome with BLAST.

We reasoned that by using the parameters of a fitted SOS model, we could post-process the output of a SLiM detection method originally designed for in-frame protein sequence and substantially improve its ability to find SLiMs in Y2H OOF sequences (Figure 4). We focused our efforts on TEIRESIAS because it returned the most comprehensive list of putative SLiMs of the methods tested and thus was most amenable to filtering and re-ranking.

Binomial test for significant SLiMs in OOF sequences. To assess the statistical significance of SLiMs detected in OOF sequences, we designed a one-tailed binomial test using binomial density $B(n,q')$, where n is the number of sequences in a dataset and q' is derived from maximum likelihood fitting of parameters of the SOS model for the OOF sequence collection of interest. The parameter q' is the probability of finding one or more occurrences of a specific SLiM in a sequence of average length in the dataset. It is computed similarly to q but here x is the average sequence length in the dataset (Appendix 1). The binomial test yields a *P*-value for each putative SLiM, which we correct for multiple testing with the Bonferroni method.

Yeast two hybrid screening

To detect pair-wise interactions, we employed a yeast two-hybrid system using the MATCHMAKER Two-Hybrid 3 System (Clontech, Palo Alto, CA, USA). The tandem BRCT domains of each protein of interest was cloned into the pGBKT7 vector as a fusion to GAL4 DBD and transformed into the *Saccharomyces cerevisiae* strain AH109 [MATa] alone or in combination with empty pGADT7 vector. Expression of MDC1 and

TP53BP1 tBRCT consistently led to the generating of very few colonies suggesting toxicity. To circumvent this problem both tBRCT domains of MDC1 and 53BP1 were sub-cloned into pGBT9, which has proven to less toxic in Y2H studies (Clontech). We conducted these two screens in triple drop out medium (SD-Trp-Leu-His), The AH109 transformants containing the bait alone were mated to Y187 [MATα] strain containing a pre-transformed human testes cDNA library (Clontech) and incubated for 8 days. Positive clones were re-streaked into quadruple-selection master plates to confirm growth. Yeast minipreps DNA was amplified by PCR and sequenced. The frame of each sequence was identified using the fusion protein sequence as the reference. Each sequence was then analyzed using pBLAST (20).

Experimental test of predicted novel SLiM

pGBKT7-LIG4-tBRCT (tandem BRCT domains) was co-transformed into yeast strain MaV203 (Invitrogen) along with pACT2-Empty, pACT2-PA2G4 wild-type (WT), or pACT2-PA2G4 6KA mutant (in which all six lysines were mutated to alanines). Colonies expressing both plasmids were selected on synthetic double drop-out media (-Leu, -Trp) and then two colonies were selected and re-streaked on -Leu, -Trp containing 0.2% 5FOA plates and grown for an additional 3 days. Only cells lacking protein–protein interactions between bait and prey grow on this selection, because 5FOA is converted to the genotoxic 5FU when Ura is activated (in cells containing the interaction). (Details in Supplementary Data).

OOF Simulations. For each simulated OOF dataset, we used maximum likelihood fitting of the SOS model to estimate number of specific amino acid residue positions required for binding ($E[sAA]$). If $E[sAA]$ is close to zero, it is likely that no selected SLiM is present in the dataset, thus all hits can be rejected. In practice, we have found that a threshold of $E[sAA] < 0.1$ works well. Otherwise if $E[sAA] \geq 0.1$, sAA and the minimum residue span required for binding (m) can be computed for each SLiM detected (e.g. by TEIRESIAS) and compared to the fitted values of $E[sAA]$ and $E[m]$ for the entire dataset.

Hits are filtered out unless they meet the following criteria:

- a putative SLiM's sAA must be either less than 7 (maximum number of specific residue positions in the ELM database) or the $E[sAA]$ estimated for the entire dataset plus a 'slack' value ζ (Equation 1);

$$1 \leq sAA_{SLiM} \leq min(7, E[sAA] + \zeta) \quad \textbf{(1)}$$

- a putative SLiM's m must be greater than three (minimum residue span in ELM database) and less than 12 (maximum residue span in ELM database) or the $E[m]$ estimated for the entire dataset plus a 'slack' value of ζ (Equation 2).

$$\zeta \leq m_{SLiM} \leq min(12, E[m] + \zeta) \quad \textbf{(2)}$$

sAA_{SLiM} and m_{SLiM} are the number of specific amino acid residues and the minimum residue span required for binding calculated for each putative SLiM. $E[sAA]$ and $E[m]$ are the same parameters estimated for the dataset. $min(a,b)$ represents the minimum of values a and b. Any putative SLiM that did not satisfy both these requirements was rejected. The slack variable ζ is intended to quantify the error ranges in which parameter estimates will be valid. It can be set to a value of choice. We set $\zeta = 3$ (details of ζ estimation in Supplementary Table S4).

Finally, we applied the binomial test to re-rank the remaining SLiMs and applied Bonferroni multiple testing correction. Applying this protocol to the output of TEIRESIAS on 38 simulated datasets yielded a tractable number of hits and achieved the best balance between SLiM detection sensitivity and specificity (Table 1, Supplementary Table S2).

Next, we explored the utility of applying our method to OOF sequences obtained from seven Y2H experiments, in which the baits were tandem BRCT domains.

RESULTS

Experimental OOF sequences from Y2H screens

We collected OOF sequences from seven Y2H experiments and BRCT tandem domains from seven human genes (BRCA1, ECT2, LIG4, BARD1, PAXIP1, MDC1, TP53BP1) were used as bait. The translated protein was BLASTed (20) against human sequences in the NCBI non-redundant protein database. Protein sequences that did not yield any human hits (by a permissive threshold of E-value >2.0) were considered OOF, as were short sequences from 4 to 10 amino acids residues beyond the GAL4 AD. We fit an SOS model to each set of OOF sequences.

We compared the shapes of the 4D NLL functions for each OOF set to the shapes of artificially generated OOF sets containing a single SLiM and varying levels of noise (Figures 3 and 5). For visualization purposes, each NLL function was sliced at its maximum likelihood values of m and LDSR. We observed that the NLL functions of OOF sets for BRCA1 and ECT2 BRCT-tandem domains have the sharp minimums consistent with presence of a SLiM and low noise in the Y2H experiment. In contrast, the NLL functions for the BARD1, PAXIP1, MDC1 and LIG4 (also TP53BP1 not shown) OOF sets have much broader minimums. As expected, the NLL for our positive control also has a sharp minimum and the NLL for our negative control has a very broad minimum. The expected number of selected amino acid residues was highest for BRCA1, ECT2 and LIG4 ($E[sAA] > 1$), lower for BARD1 and PAXIP1 ($0.1 < E[sAA] < 1$) and very low for MDC1 and TP53BP1 ($E[sAA] < 0.1$). Interestingly, the screenings using both MDC1 and TP53BP tandem BRCTs were initially problematic (due to toxicity of the bait) and the method was adjusted to allow the final screenings. The LIG4 set has a high $E[sAA]$, but it also has a high number of sequences expected to bind to the bait which do not contain a SLiM ($E[LISS] = 183.4$) (Table 2).

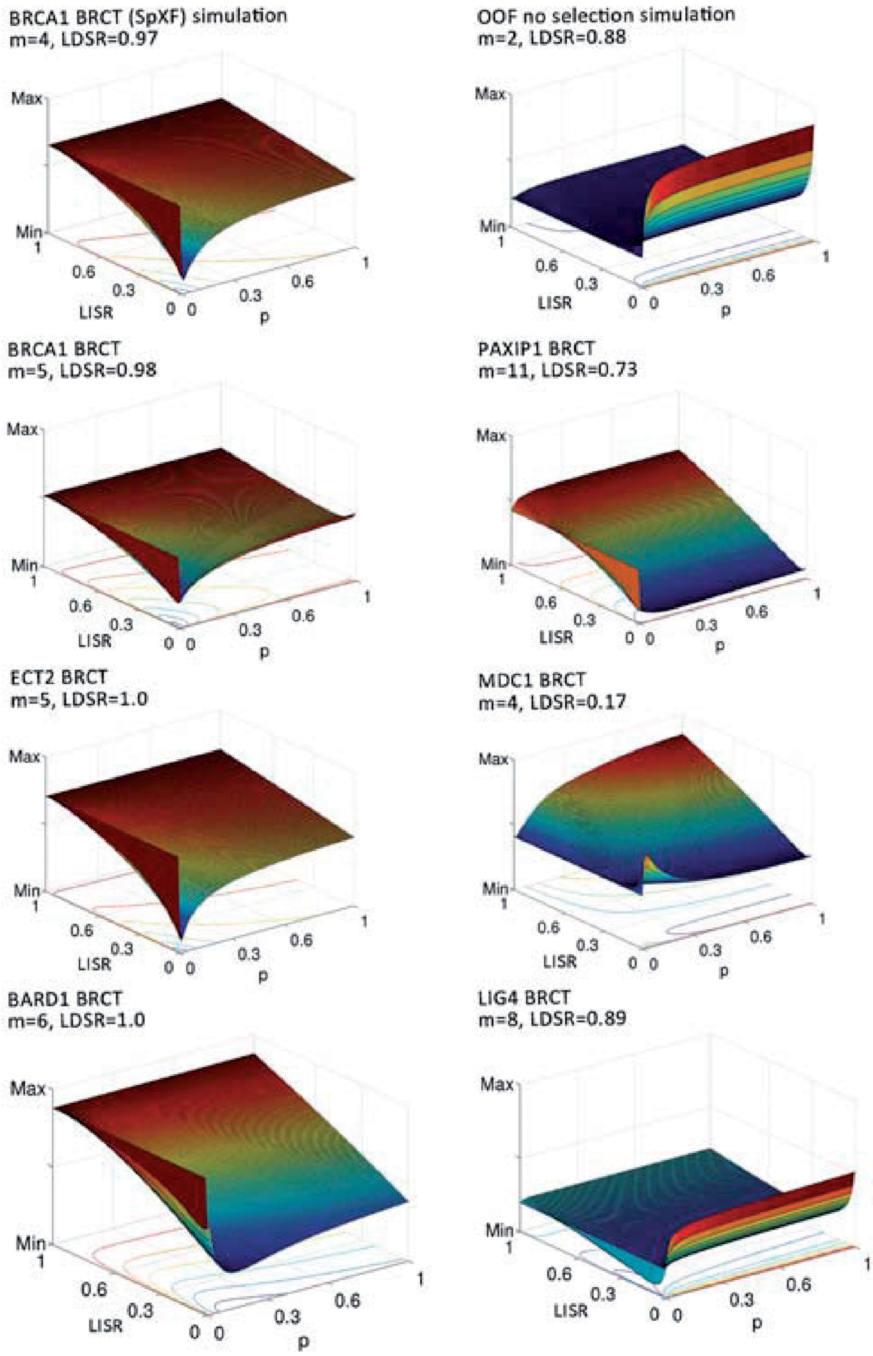

Figure 5. Negative log likelihood function slices after fitting with OOF sequences from six Y2H experiments. BRCT-tandem domains from six human genes were used as bait. NLL slices of positive and negative controls also shown. The *Z*-axis shows values of the four-dimensional NLL function. Each slice shows parameters *p* and *LISR*, with parameters *m* and LDSR fixed at their maximum likelihood estimates. Sharp minimums are a signature of SLiM presence and low noise.

Following the protocol in Figure 6, after fitting SOS models to the seven Y2H OOF sequence sets, we used TEIRESIAS to detect SLiMs for all experiments except MDC1 and TP53BP1. SLiMs with values of *m* and *sAA* within the ranges defined by Equation 1 and Equation 2 were retained and ranked with a one-tailed binomial test (Bonferroni corrected for multiple testing). A number of SLiMs per experiment ranging from over 600 to 2500 remained.

SLiMs in putative interaction partners of the bait

Mapping all of these putative SLiMs onto the human proteome is challenging, because of the very large number of potential binding sites. To narrow the search space, we targeted the in-frame translated clones (identified with BLAST) from each of the seven Y2H experiments. Importantly, while BLAST is able to identify likely interaction partners of each BRCT tandem domain,

Table 2. Fitted MLE parameters demonstrate predicted properties of motifs under selection in seven Y2H screens

Bait	*E*[*sAA*]	*p*	*m*	LDSR	LISR	*E*[*LDSS*]	*E*[*LISS*]
BARD1	0.6	0.15	6	0.99	1.0E-6	366.0	0.0
BRCA1	4.6	1.0E-6	5	0.98	4.0E-6	260.2	63.8
ECT2	3.4	3.9E-5	5	1.00	1.0E-6	117.8	0.2
LIG4	4.6	1.0E-6	8	0.89	1.6E-5	145.6	183.4
MDC1	0.0	0.97	4	0.18	3.8E-2	261.9	9.1
PAXIP1	0.5	0.24	11	0.73	3.4E-2	429.7	18.3
TP53BP1	0.0	1.0	20	0.22	0.47	42.0	139.0

E[*sAA*], expected number of specific amino acid residues in the motif (all non-wildcard positions); *p*, motif frequency; *m*, motif length; *LDSR*, Length-Dependent Selection Rate of Y2H screen; *LISR*, Length-Independent Selection Rate of Y2H screen; *E*[*LDSS*], expected number of Length-Dependent selected sequences in the Y2H screen (Equation A7); *E*[*LISS*], expected number of Length-Independent selected sequences in the Y2H screen.

the SLiM responsible for binding is not identified. We used regular-expression matching to search through the in-frame BLAST hits.

Using this protocol, we discovered several potentially interesting SLiMs in the OOF clones that bound to BRCA1, LIG4 and ECT2, which may point to physical protein–protein interactions. For BRCA1, the well-documented SLiM SPXF was top-ranked by our method (and also by TEIRESIAS). The presence of this SLiM in OOF clones provides a proof-of-concept that linear motifs involved in binding specificity do occur in OOF clones.

For LIG4, the top-ranked SLiM that we mapped onto an interaction partner is KKKKKK mapped onto PA2G4 (aka EBP1), an RNA-binding protein involved in growth regulation and ERBB3 binding (21,22). The only available crystal structure for PA2G4 as of this writing (PDB 2q8k) (23) does not contain electron density for the C-terminal region containing KKKKKK and this region is likely to be disordered, at least in the absence of a binding partner. Interestingly, there is an X-ray crystal structure of the LIG4 tandem BRCT domains in complex with XRCC4 (PDB 3ii6) (24) in which LIG4 and XRCC4 binding is mediated by a lysine-rich region. A helix-loop-helix structure in the linker region that connects the LIG4 tandem BRCT domains with each other (Figure 7A) forms a gate surrounding the coiled-coil of XRCC4 (Figure 7B). Lysines K187 and K188 in XRCC4 are involved in salt bridges at the interface of LIG4 and XRCC4 and are critical to their binding interaction (25). We propose that the KKKKKK SLiM in PA2G4 mediates binding to the LIG4 tandem BRCT domains at the linker region, through a similar mechanism. If this is the case, a PA2G4 dimer is likely required, since two alpha helices containing lysine-rich regions are present in XRCC4 to interact with a single LIG4 tandem BRCT domain.

In order to verify whether the KKKKKK SLiM played a role in the binding between the LIG4 tandem BRCT domains and PA2G4, we tested direct binding using Y2H. We compared binding of the LIG4 tandem BRCT domains to the WT PA2G4 or to a mutated version of PA2G4 in which the lysines in the SLiM were all mutated to alanines. As shown, mutation of the KKKKKK SLiM abrogated binding to LIG4 tandem BRCT domains (Figure 8).

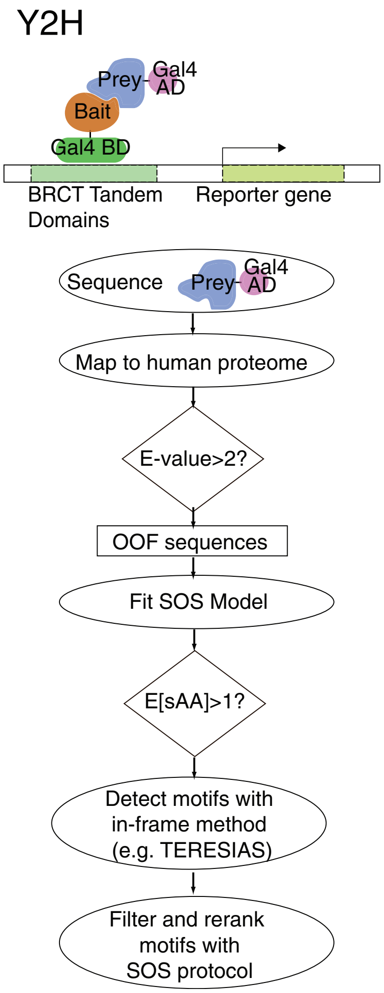

Figure 6. Flow chart of protocol to detect SLiMs in OOF clones from a Y2H experiment. Positive clones are sequenced and analyzed by BLAST (20). Human cDNA sequences that do not translate as human proteins (the OOFs) are used to fit the model. If the expected number of specific amino acids in the sequences (*E*[*sAA*]) > 1.0, an in-frame SLiM detection method (e.g. TEIRESIAS) is applied. We then iterate over the ranked list of putative SLiMs identified by the in-frame method, check whether each SLiM has length and number of specific amino acids and length within a range predicted by the model. Putative SLiMs that pass these criteria are re-ranked with a one-sided binomial *P*-value test and Bonferroni testing correction (data not shown).

For LIG4, the sixth-ranked SLiM PRPRP was mapped onto MAP2K2, a kinase involved in growth factor signal transduction. The SLiM also occurs in a region that could not be resolved by X-ray crystallography (PDB 1S9I), but it is anchored by two ends that are solvent accessible (26). For ECT2, the third-ranked SLiM is PSPXL, which maps onto SI, a glucosidase enzyme and the sixth-ranked SLiM is SXPXXAA, which maps onto PCOLCE, an enzyme involved in processing of procollagen. Both SI and PCOLCE have X-ray crystal structures (PDB 1uap and

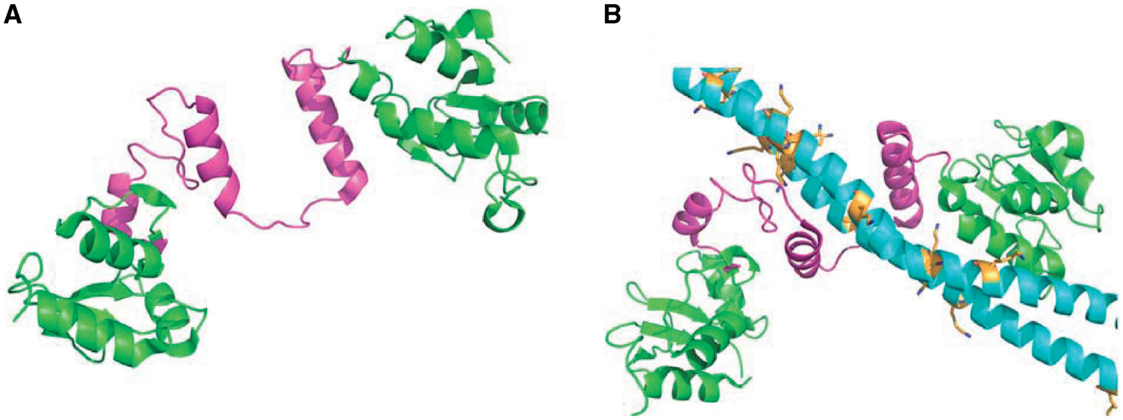

Figure 7. Proposed binding mechanism for the LIG4 BRCT tandem domains and PA2G4 as mediated by the SLiM KKKKKK. (**A**) LIG4 BRCT tandem domains (PDB 3ii6) with the linker region in magenta. (**B**) LIG4 BRCT tandem domains (PDB 3ii6) bound to coiled coil of XRCC4 (in cyan). Lysine residues are shown as orange sticks. Two lysines in XRCC4 (K187, K188) are critical to the binding interaction. If a similar mechanism is involved in the interaction of LIG4 and PA2G4, it would likely require a PA2G4 dimer, since both helices in the XRCC4 coiled-coil contain lysine rich regions. Figures created with PyMOL.

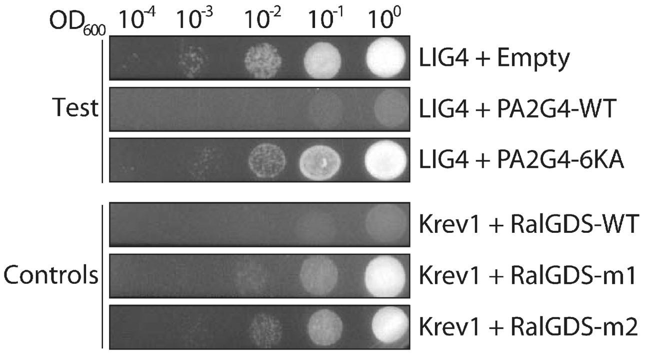

Figure 8. Y2H direct binding assay. Mutation of the KKKKKK SLiM abrogates binding to LIG4 tandem BRCT domains. Only cells lacking protein-protein interactions between bait and prey grow on this selection (-Leu, -Trp containing 0.2% 5FOA plates) because 5FOA is converted to the genotoxic 5FU when Ura is activated in cells containing the interaction. Empty vector and wild-type PA2G4 vector shown as controls. Cells were plated in serial dilutions of a saturated liquid culture.

3lpp) and in both cases the SLiMs were in regions not resolved in the structures (27,28).

Finally, we applied the published SLiM detectors tested on our *in silico* benchmarks to the seven experimental OOF sets. With the exception of the BRCA1 set, in which all methods except DILIMOT identify the well-known SPXF motif, there is very little consensus among methods. SLiMFinder run with relaxed parameters identifies a putative SLiM HPXXL in the LIG4 set, which is also ranked at 109 by our SOS method. This is an interesting hit and warrants further investigation. Details of top hits from all methods are in Supplementary Table S5.

DISCUSSION

We have presented a new approach to extract biologically relevant information from experimental data that is currently discarded after researchers perform Y2H screening of cDNA libraries. The intent of Y2H screening is to discover which gene products bind to a bait of interest. Yet by discarding OOF sequences, which typically comprise up to two-thirds of the hits (2), a researcher may be discarding meaningful information. With our method, it is possible to detect signatures of selection that point to the biological relevance of why sequences do or do not bind to the bait. Although the OOF sequence itself is not part of a gene product, if signatures of selection indicate interaction specificity, then it is very likely to contain a SLiM. Our results suggest that the proposed method can help researchers ascertain if 'junk' sequences from a Y2H experiment are likely to contain SLiMs. This allows experimental work to focus on promising OOF datasets and to abandon efforts on datasets with high noise.

We have developed a novel computational method to detect signatures of selection in OOF sequences. Our method detects signal by considering the sequence lengths in a collection of OOF clones, which can be represented as a list of integers. We derive two probabilistic models to represent: the case of this length distribution when the clones do not contain a SLiM that selects for binding to the Y2H bait; and the case where the clones contain one or more selected SLiMs. The first case is modeled as a simple geometric distribution, where success is the occurrence of a stop codon. The second case is modeled as a modified geometric distribution with additional parameters, representing properties of selected SLiMs. Our basic hypothesis is that if the clones contain selected SLiMs, the length distribution will fit the modified geometric distribution better than the simple geometric distribution, and vice versa. Thus, a likelihood ratio test can be used to assess the presence of selected SLiMs in a collection of OOF clones. In contrast to method such as BLAST (29), the likelihood score is not applied to individual sequences, but to the entire collection of clones. A novel aspect of our method is that it allows a user to reject an entire OOF sequence dataset as unlikely to contain a SLiM so as not to squander resources. We present a protocol that allows a user to first fit the SOS model based on the distribution of sequence lengths in a Y2H OOF dataset and then to assess whether SLiMs are present.

From the experimentalists point of view, it is important to stress that the work described here was limited to datasets recovered in Y2H experiments in which only the 'protein module' domain(s), in this case the BRCT tandem domains, was used as bait. In our experience, OOF clones from Y2H screenings, in general, show an overrepresentation of 'sticky' short hydrophobic peptides that interact non-specifically to the bait. We have found that reducing the bait space to the protein module domain(s) decreases the number of these non-specific hits. Thus, Y2H screens conducted with full length proteins may generate may not be as enriched for potential OOF SLiMs, as screens focused on a protein module domain(s).

In future work, we plan to develop an automated pipeline for the analysis of OOF datasets. The pipeline will map predicted SLiMs back to both the original in-frame dataset and to disordered regions of the human proteome, which are expected to be enriched for SLiMs. Currently, MATLAB and python scripts are available from the authors on request.

CONCLUSION

We believe that the method described here will be an important addition to the search of biologically meaningful SLiMs that are important in signal transduction. Moreover, because the method tests for interactions *in vivo* it may be particularly useful when searching for motifs in which post-translation modifications are required for optimal recognition.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

This work was supported by Agency for Science Technology and Research (Singapore) fellowship (to Y.L.); National Science Foundation (DBI-0845275 to R.K.); National Institutes of Health (CA135877 to R.K.), (CA148112,CA116167,CA11809, to A.M.); US Army Medical Research and Material Command National Functional Genomics Center Project (W81XWH-08-2-0101) (opinions, interpretations, conclusions and recommendations are those of the author(s) and are not necessarily endorsed by the US Army); and Florida Breast Cancer Foundation fellowship (to N.W.). Funding for open access charge: National Science Foundation (DBI-0845275).

Conflict of interest statement. None declared.

REFERENCES

1. Fields,S. and Song,O. (1989) A novel genetic system to detect protein-protein interactions. *Nature*, **340**, 245–246.
2. Vidalain,P.-O., Boxem,M., Ge,H., Li,S. and Vidal,M. (2004) Increasing specificity in high-throughput yeast two-hybrid experiments. *Methods*, **32**, 363–370.
3. Pawson,T. and Nash,P. (2003) Assembly of cell regulatory systems through protein interaction domains. *Science*, **300**, 445–452.
4. Pawson,T. (1995) Protein modules and signalling networks. *Nature*, **373**, 573–580.
5. Yaffe,M.B., Leparc,G.G., Lai,J., Obata,T., Volinia,S. and Cantley,L.C. (2001) A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nat. Biotechnol.*, **19**, 348–353.
6. Neduva,V., Linding,R., Su-Angrand,I., Stark,A., de Masi,F., Gibson,T.J., Lewis,J., Serrano,L. and Russell,R.B. (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol.*, **3**, e405.
7. Neduva,V. and Russell,R.B. (2005) Linear motifs: evolutionary interaction switches. *FEBS Lett.*, **579**, 3342–3345.
8. Zarrinpar,A., Park,S.-H. and Lim,W.A. (2003) Optimization of specificity in a cellular protein interaction network by negative selection. *Nature*, **426**, 676–680.
9. Fuxreiter,M., Tompa,P. and Simon,I. (2007) Local structural disorder imparts plasticity on linear motifs. *Bioinformatics*, **23**, 950–956.
10. Hugo,W., Song,F., Aung,Z., Ng,S.-K. and Sung,W.-K. (2010) SLiM on Diet: finding short linear motifs on domain interaction interfaces in Protein Data Bank. *Bioinformatics*, **26**, 1036–1042.
11. Rigoutsos,I. and Floratos,A. (1998) Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. *Bioinformatics*, **14**, 55–67.
12. Edwards,R.J., Davey,N.E. and Shields,D.C. (2007) SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. *PLoS ONE*, **2**, e967.
13. Frith,M.C., Saunders,N.F.W., Kobe,B. and Bailey,T.L. (2008) Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput Biol.*, **4**, e1000071.
14. La,D. and Livesay,D.R. (2005) MINER: software for phylogenetic motif identification. *Nucleic Acids Res.*, **33**, W267–W270.
15. Tan,S.-H., Hugo,W., Sung,W.-K. and Ng,S.-K. (2006) A correlated motif approach for finding short linear motifs from protein interaction networks. *BMC Bioinformatics*, **7**, 502.
16. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
17. Rodriguez,M., Yu,X., Chen,J. and Songyang,Z. (2003) Phosphopeptide binding specificities of BRCA1 COOH-terminal (BRCT) domains. *J. Biol. Chem.*, **278**, 52914–52918.
18. Burkard,M.E., Maciejowski,J., Rodriguez-Bravo,V., Repka,M., Lowery,D.M., Clauser,K.R., Zhang,C., Shokat,K.M., Carr,S.A., Yaffe,M.B. *et al.* (2009) Plk1 self-organization and priming phosphorylation of HsCYK-4 at the spindle midzone regulate the onset of division in human cells. *PLoS Biol.*, **7**, e1000111.
19. Manke,I.A., Lowery,D.M., Nguyen,A. and Yaffe,M.B. (2003) BRCT repeats as phosphopeptide-binding modules involved in protein targeting. *Science*, **302**, 636–639.
20. Altschul,S.F., Madden,T.L., SchŁffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
21. Squatrito,M., Mancino,M., Donzelli,M., Areces,L.B. and Draetta,G.F. (2004) EBP1 is a nucleolar growth-regulating protein that is part of pre-ribosomal ribonucleoprotein complexes. *Oncogene*, **23**, 4454–4465.
22. Zhang,Y. and Hamburger,A.W. (2005) Specificity and heregulin regulation of Ebp1 (ErbB3 binding protein 1)

mediated repression of androgen receptor signalling. *Br. J. Cancer*, **92**, 140–146.

23. Kowalinski,E., Bange,G., Bradatsch,B., Hurt,E., Wild,K. and Sinning,I. (2007) The crystal structure of Ebp1 reveals a methionine aminopeptidase fold as binding platform for multiple interactions. *FEBS Lett.*, **581**, 4450–4454.
24. Wu,P.-Y., Frit,P., Meesala,S., Dauvillier,S., Modesti,M., Andres,S.N., Huang,Y., Sekiguchi,J., Calsou,P., Salles,B. *et al.* (2009) Structural and functional interaction between the human DNA repair proteins DNA ligase IV and XRCC4. *Mol. Cell. Biol.*, **29**, 3163–3172.
25. Modesti,M., Junop,M.S., Ghirlando,R., van deRakt,M., Gellert,M., Yang,W. and Kanaar,R. (2003) Tetramerization and DNA ligase IV interaction of the DNA double-strand break repair protein XRCC4 are mutually exclusive. *J. Mol. Biol.*, **334**, 215–228.
26. Ohren,J.F., Chen,H., Pavlovsky,A., Whitehead,C., Zhang,E., Kuffa,P., Yan,C., McConnell,P., Spessard,C., Banotai,C. *et al.* (2004) Structures of human MAPKK1 (MEK1) and MEK2 describe novel noncompetitive kinase inhibition. *Nat. Struct. Mol. Biol.*, **11**, 1192–1197.
27. Liepinsh,E., Banyai,L., Pintacuda,G., Trexler,M., Patthy,L. and Otting,G. (2003) NMR structure of the netrin-like domain (NTR) of human type I procollagen C-proteinase enhancer defines structural consensus of NTR domains and assesses potential proteinase inhibitory activity and ligand binding. *J. Biol. Chem.*, **278**, 25982–25989.
28. Sim,L., Willemsma,C., Mohan,S., Naim,H.Y., Pinto,B.M. and Rose,D.R. (2010) Structural basis for substrate selectivity in human maltase-glucoamylase and sucrase-isomaltase N-terminal domains. *J. Biol. Chem.*, **285**, 17763–17770.
29. Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.
30. Coleman,T. and Li,Y. (1994) On the convergence of reflective newton methods for large-scale nonlinear minimization subject to bounds. *Math. Progr.*, **67**, 189–224.
31. Coleman,T. and Li,Y. (1996) An interior, trust region approach for nonlinear minimization subject to bounds. *SIAM J. Optimiz.*, **6**, 418–445.
32. Gould,C.M., Diella,F., Via,A., Puntervoll,P., Gemnd,C., Chabanis-Davidson,S., Michael,S., Sayadi,A., Bryne,J.C., Chica,C. *et al.* (2010) ELM: the status of the 2010 eukaryotic linear motif resource. *Nucleic Acids Res.*, **38**, D167–D180.

APPENDIX 1

Mathematical details

Length distribution of OOF sequences with no SLiM selection. In the absence of selection for a particular SLiM, the lengths of the OOF sequences follow a geometric distribution, with fixed parameter s (Equation A1), which is the probability of a stop codon when there is no evolutionary pressure for functional protein.

We estimated s by taking the frequencies of each nucleotide in the human transcriptome and computing expected stop codon frequency as the sum over products of the possible component nucleotides (TAA, TAG and TGA) ($s = 0.048561$).

$$f(x) = \begin{cases} \frac{(1-s)^x}{0} & x \geq 1 \\ 0 & \text{otherwise} \end{cases} \tag{A1}$$

where x is OOF sequence length and k is the normalizing constant (in practice we find that summing up to length 300 yields sufficient accuracy for normalization).

Length distribution of OOF sequences with SLiM selection. If a sequence contains a SLiM, Equation A1 is modified by adding parameters for: *SLiM frequency* (p), *SLiM length* (m), the expected frequencies with which the Y2H screen reports a binding interaction when a true binding SLiM is present (LDSR) or when it is absent (LISR) (Equation A2).

$$g(x) = \begin{cases} \frac{(1-s)^x(q*LDSR+(1-q)*LISR)}{k'} & x \geq m \\ \frac{(1-s)^x LISR}{k'} & 1 \leq x < m \\ 0 & \text{otherwise} \end{cases} \tag{A2}$$

where x is sequence length, q is the probability that the sequence contains at least one SLiM (Equation A3)

$$q = 1 - (1-p)^{x-m+1} \tag{A3}$$

and k' is a normalizing constant.

SLiM frequency (p) (Equation A4) is defined using nucleotide frequencies from the human transcriptome. Similar to our estimate of s (geometric distribution under no selection), for each position in a SLiM, we sum over the frequencies of all possible codons and all allowed amino acids, then take the product over all positions. We make the simplifying assumption that each position in a SLiM is independent. Given a SLiM without variable length positions, p can be computed by Equation A4.

$$p = h(Y) = \prod_i \sum_j \sum_k \left[\frac{1}{1-s} \prod_l y_{ijkl} \right] \tag{A4}$$

where Y is a SLiM and y_{ijkl} is the frequency of a nucleotide in the human transcriptome. i indexes positions in the SLiM, j indexes the 'allowed amino acids' for each position, k indexes possible codons for each amino acid, and l indexes the component nucleotides for each codon. $\frac{1}{1-s}$ normalizes the codon frequencies so as to exclude stop codons. Allowed amino acids for each position are included in the definition of a SLiM. Where more than one amino acid is allowed, regular expression patterns are used. For example, the frequency of the SLiM `[ST]XXF` is calculated as [0.0966+0.0673]*1*1*0.0310.

To estimate the expected number of specific amino acid residues in a SLiM or *sAA* (amino acid residues in the peptide that are necessary and sufficient for binding), we compute $log_{0.05}(p)$, which gives the expected number of amino acids that are necessary to yield a specific value of p, under the simplifying assumption that all amino acids are equally probable in SLiMs. Wildcard positions do not factor into this calculation since these positions have a frequency of 1.0. Caution should be taken when interpreting *sAA*. For example, a promiscuous SLiM such as `S[YADL] [YAD] [IYD]` has an estimated *sAA* of 2.67 even though 4 positions are being selected for because many positions have multiple accepted amino acids.

SLiM length (m) is the minimum length of the sequence necessary for binding and includes all positions in the regular expression. For example $m = 5$ for the SLiM

(XNPFX) even though specific amino acids are selected for only in the span of the central 3 positions.

Maximum likelihood estimation of parameters. The likelihood function for the modified geometric distribution in Equation A2 is:

$$L(\vec{\theta}) = \prod_i g(\vec{\theta}|x_i) \qquad \textbf{(A5)}$$

where $\vec{\theta} = (p, m, LDSR, LISR)$. The negative log likelihood is:

$$NLL(\vec{\theta}) = -\sum_i log(g(\vec{\theta}|x_i)) \qquad \textbf{(A6)}$$

For each collection of OOF sequences that bind to a protein bait, we used the Trust Region-Reflective Optimization algorithm (30,31), with gradient estimated by central differences, as implemented in Matlab 2007a, to obtain MLE of *p*, LDSR and LISR, using Equation A6 as the objective function.

To increase the tractability of this optimization problem, we did a grid search to find the best fit of the discrete parameter *m*. First, we identified the range of likely SLiM lengths in our sequences, using the ELM database of SLiMs (32) (February 2010 release). Almost all SLiMs in ELM are between 3 and 14 positions in length. We therefore repeated the MLE of *p*, LDSR and LISR for each value of *m* from 3 to 20 and selected the model with the highest likelihood. As positive and negative controls, we repeated the maximum likelihood fitting using our artificial set of simulated BRCA1 BRCT OOF sequences (positive) and the OOF sequences with no selection (negative).

Expected number of length-dependent and length-independent selected sequences from Y2H screen. After obtaining MLE of the parameters, we used Equation A2 to derive the number of LDSS and the number of LISS of a Y2H screen.

$$E[LDSS] = \sum_{x_i \geq m} \frac{q * LDSR}{q * LDSR + (1-q) * LISR} \qquad \textbf{(A7)}$$

The expected number of length-independent selected sequences, $E[LISS]$ is the total number of OOF sequences minus $E[LDSS]$.