Systematic Discovery of New Recognition Peptides Mediating Protein Interaction Networks

Victor Neduva[1], Rune Linding[1], Isabelle Su-Angrand[1], Alexander Stark[1], Federico de Masi[1], Toby J. Gibson[1], Joe Lewis[1], Luis Serrano[1], Robert B. Russell[1,2*]

1 European Molecular Biology Laboratory, Heidelberg, Germany **2** European Molecular Biology Laboratory–European Bioinformatics Institute, Hinxton, United Kingdom

Many aspects of cell signalling, trafficking, and targeting are governed by interactions between globular protein domains and short peptide segments. These domains often bind multiple peptides that share a common sequence pattern, or "linear motif" (e.g., SH3 binding to PxxP). Many domains are known, though comparatively few linear motifs have been discovered. Their short length (three to eight residues), and the fact that they often reside in disordered regions in proteins makes them difficult to detect through sequence comparison or experiment. Nevertheless, each new motif provides critical molecular details of how interaction networks are constructed, and can explain how one protein is able to bind to very different partners. Here we show that binding motifs can be detected using data from genome-scale interaction studies, and thus avoid the normally slow discovery process. Our approach based on motif over-representation in non-homologous sequences, rediscovers known motifs and predicts dozens of others. Direct binding experiments reveal that two predicted motifs are indeed protein-binding modules: a DxxDxxxD protein phosphatase 1 binding motif with a K_D of 22 μM and a VxxxRxYS motif that binds Translin with a K_D of 43 μM. We estimate that there are dozens or even hundreds of linear motifs yet to be discovered that will give molecular insight into protein networks and greatly illuminate cellular processes.

Citation: Neduva V, Linding R, Su-Angrand I, Stark A, de Masi F, et al. (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. PLoS Biol 3(12): e405.

Introduction

Protein interactions are central to all cellular processes. At the molecular level they can occur in a variety of ways. Probably the best known involve specific contacts between globular domains (~100–200 residues) present in the interacting proteins. These are seen in many different contexts ranging from different subunits in large molecular machines (e.g., RNA polymerase II [1]), to more transient interactions (e.g., cyclins binding to CDK2 [2]).

However not all interactions are mediated by pairs of globular domains. Many involve the binding of a domain in one protein to short regions (approximately three to eight residues) in another [3,4]. These regions often show a particular sequence pattern, or "linear motif," which captures the key residues involved in function or binding [5]. Linear motifs are critical to many processes including signal transduction (e.g., SH3 domains bind PxxP [6]), gene expression (e.g., Groucho→WRPW [7]) and DNA replication (e.g., PCNA→QxxxxxFF [8]).

In contrast to domains, which are readily detectable by sequence comparison, linear motifs are difficult to discover due to their short length, a tendency to reside in disordered regions in proteins, and limited conservation outside of closely related species. To date they have typically been found by time-consuming experiments, meaning that only a few hundred motifs are known compared to thousands of domains that might bind them. Although it is at present difficult to estimate just how many such interaction motifs exist, it is likely that many interactions are mediated by those not yet discovered. Here we perform the first systematic attempt to discover new motif candidates and their corresponding binding partners using results of genome-scale interaction datasets.

Results

Methodology

Our central hypothesis is that proteins with a common interaction partner will share a feature that mediates binding, either a domain or a linear motif. In the absence of a shared domain, a linear motif could well be the only common sequence feature and might thus be detectable simply by virtue of over-representation, which is the basis of our approach (Figure 1).

Given a set of proteins sharing an interaction partner we first remove sequence regions unlikely to contain linear motifs: globular domains, trans-membrane segments, coiled-coils, collagen regions, and signal peptides. This is justified because only 15% of known linear motifs [5] occur within

Received August 9, 2005; Accepted September 27, 2005; Published November 15, 2005
DOI: 10.1371/journal.pbio.0030405

Abbreviations: PP1, protein phosphatase 1; S_{cons}, the product of all binomial probabilities for sets of orthologous proteins from a given set of genomes

Academic Editor: Rowena Matthews, University of Michigan, United States of America

*To whom correspondence should be addressed. E-mail: russell@embl.de

these regions, and including them can give rise to misleading motif signals, particularly if common domains are found in more than one protein in a set. Most importantly, this avoids the detection of repetitive, purely structural patterns, such as β-turns, coiled-coil heptads, or collagen repeats, because these are unlikely to occur in the unstructured parts of proteins that remain after this filtering. We also compare all sequences in a set to each other and leave only one representative of any homologous segments. We do this in order to measure over-representation that is not the result of homology; our assumption is that each of the remaining instances of a particular motif has arisen convergently and is thus an independent observation. We specifically avoid removing regions of low complexity because linear motifs frequently occur within them.

We then find all three to eight residue motifs in the remaining sequence [9], and score their over-representation as the binomial probability *(P)* of seeing them randomly in a similar set of sequences (see Materials and Methods). This allows multiple observations of an otherwise insignificant motif to become statistically significant by over-representation, and readily accounts for sets of different sizes and composition. For example, the SH3-binding pattern RxPxxP readily occurs in about one out of 20 randomly selected proteins, but its occurrence in seven sequences in a set of nine becomes highly significant. We also compute *P* for all closely related species based on whether or not the same motifs are seen in the corresponding orthologues, and multiply these to give a final score (S_{cons}*;* see Materials and Methods).

We applied our approach to interacting sets of proteins from *Saccharomyces cerevisiae, Drosophila melanogaster, Caenorhabditis elegans,* and *Homo sapiens* [10–14]. For the first three species, these datasets are from yeast two-hybrid screens; human data comes from the human Proteome Resource Database (HPRD) [14] and consists of hand-curated interactions extracted from the literature (see Protocol S1). For each dataset, we constructed a control by selecting random sets of proteins of a similar length and number, and performed the same calculations. We then defined a confidence threshold (*p*-value < 0.001) for S_{cons} for each dataset (see Materials and Methods). Note that this threshold does not necessarily reflect the accuracy in terms of identifying *binding* motifs, only that the particular sequence pattern reported is very unlikely to arise by chance. It is possible that patterns can arise for other reasons, including localization signals or other sequence features common to protein performing similar function.

Known motifs come in different flavours, for instance canonical SH3-binding motifs (PxxP) are embellished with different amino acids, which determine the specific SH3-containing protein they bind (e.g., RxPxxP and PxxPxK). The sets of proteins above (i.e., those sharing an interaction partner) are appropriate for finding such motif flavours because each protein containing a particular instance of a domain (e.g., SH3) is considered separately. However, it is also beneficial to detect more general motifs specific to a domain family. To do this we simply merge sets if the common binding partners shared a particular domain. We refer to these as "domain" sets in the sections that follow (see Protocol S1 and Figure S1).

Benchmark

The Eukaryotic Linear Motif resource (ELM) [5] contains a curated set of experimentally validated instances of binding motifs (i.e., their location in a particular protein). This provides several pertinent sets of proteins to test the approach, namely each set of proteins containing a known instance of a particular motif (e.g., all PxxP motif–containing sequences known to interact with SH3 domains). Of 58 different sets, 22 contained at least four non-homologous instances of the motif, and could be used to test our approach. We ran the procedure on each set and monitored where the known motif (or a variant) was found in the list of all motifs ranked according to S_{cons}. Despite many thousands of possibilities, the approach detected the correct motif as the very best ranked for 14 out of 22 and among the top ten for an additional three (Table1). Applying the confidence threshold left eleven correct motifs at first rank, and no false predictions (see legend to Table1). Inspection showed that those motifs that were either missed or scored poorly were

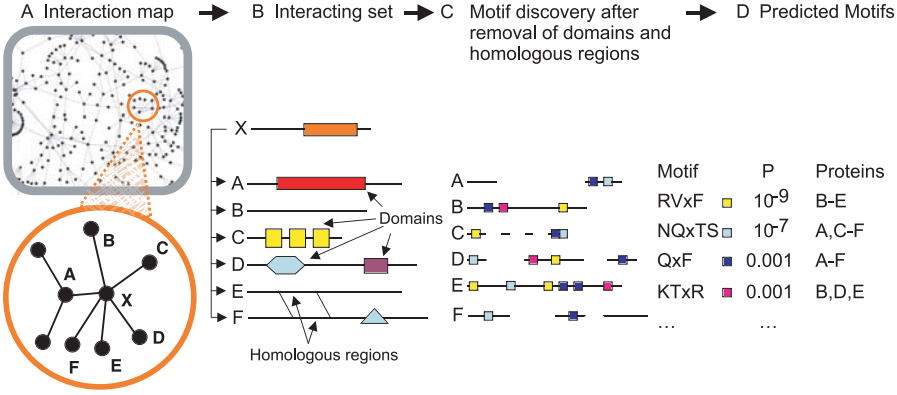

Figure 1. Schematic of the Linear Motif Discovery Strategy

Interaction maps are probed for interaction sets (A): Partners of proteins with multiple interactions are clustered together when there are no known sequence features present (B). Domains and homologous regions are then identified (B) and removed prior to running exhaustive pattern discovery (C) to produce a list of motifs ranked by their probabilities *P* (D). Hypothetical motifs are shown as coloured squares in (C) and (D). "Proteins" in (D) gives the set of proteins containing at least one copy of the motif.

DOI: 10.1371/journal.pbio.0030405.g001

Table 1. Detection of Known Linear Motifs in Experimentally Verified Sets from the Eukaryotic Linear Motif

Name→Motif (or Name:Motif) [K_D Range][a]	Initial Motifs[b]	Best Correct (Rank)[c]	S_{cons}	Fraction with Motif[d]
Retinoblastoma→(LI)xCx(DE)	24,581	LxCxE (1)[e]	4.4×10^{-37}	14/24
Cyclin→(RK)xLx$_{\{0-1\}}$(FYLIVMP) [0.19 μM] [55]	13,179	KRRL (15)	9.7×10^{-18}	3/19
14–3–3 (type 1)→R(SFYW)xSxP [0.15 μM] [56]	225	RSxSxP (1)[e]	4.2×10^{-31}	3/4
14–3–3 (type 3)→(RHK)(STALV)x(ST)x(PEDSIF)	1656	RSxSxE (6)	1.97×10^{-20}	7/12
γ-adaptin→(DE)(DES)xFx(DE)(LVIMFD)	392	DDxFxxF (1)[e]	2.35×10^{-24}	3/4
SH3 (type 2) domain→PxxPx(KR) [0.45–142 μM] [38]	1,406	PPxxPxR (1)[e]	2.23×10^{-41}	4/7
CtBP→Px(DEN)L(VAST) [2.5–45 μM] [57]	26,892	DxPxDL (1)[e]	2.45×10^{-50}	8/25
Integrin→RGD	154	R.DV (2)[f]	1.95×10^{-19}	3/8
TRAF2→(PSAT)x(QE)E	191	PxQE (1)[f]	2.32×10^{-12}	4/7
TRAF6→PxE	121	PQE (1)[f]	6.5×10^{-11}	3/7
HP-1→PxVx(LM)	5,287	KVPxVxL (4)	4.31×10^{-20}	3/7
N-glycosylation:NxC	53	—	—	0/4
Golgi-to-ER signal:(KRH)(DENQ)EL	8	KDEL (1)[e]	4.67×10^{-32}	3/5
PCNA→Qxx(ILM)xx(FHM)(FHM) [0.1–60 μM] [58]	1,505	QxxxxxFF (1)[e]	7.6×10^{-104}	11/24
SUMO-1→(VILAFP)Kx(EDNGP)	13,722	—	—	0/14
Dynein light chain→(KR)xTQT	117	KxTQT (1)[e]	8.96×10^{-30}	3/4
Groucho/TLE→(WFY)RP(WFY)	5,324	WRP (1)[e]	0.0	21/33
Clathrin box→L(ILM)x(ILMF)(DE)	778	LIxLD (1)[e]	1.01×10^{-49}	3/5
EH 1→Fx(IV)xx(IL)(ILM)	16,676	FxIxNI (1)[e]	0.0	4/83
NRBOX→LxxLL	27,874	—	—	1/18
Endosome sorting signal:(DER)xxxL(LVI)	1,471	ExxxLL (22)	1.4×10^{-17}	5/12
Mannosylation site:WxxW	21	DGxW (1)[f]	2.1×10^{-20}	3/6

This table shows the results of a benchmark using the 22 cases from the Eukaryotic Linear Motif (ELM) [5] resource having at least four non-homologous instances of the given motif.

[a]Motifs and their associated proteins or domains are listed (variable positions are shown in parenthesis). Interactions are denoted by an arrow (→), modifications and targeting by a colon (:). When available, these are followed by the range of known affinities for these motifs and their interaction partners (note that the cyclin/peptide affinity is for a longer, 18 amino acid sequence containing the motif).

[b]The total number of motifs initially discovered in the set.

[c]The best motif matching the known consensus (the rank is given in parenthesis).

[d]The number of proteins having the motif shown over the total number of non-homologous proteins in the set.

[e]The motif scored above the confidence threshold (5.0×10^{-21}).

[f]The correct motif is ranked first.

A long dash indicates that the known pattern was not detected among the 100 best-ranked motifs.

CtBP, C-terminus binding protein; ER, endoplasmic reticulum; Groucho/TLE, Groucho/transducin-like enhancer-of-split family; HP-1, heterochromatin associated protein 1; NRBOX, nuclear receptor box; TRAF, tumor necrosis factor (TNF) receptor-associated factor.

DOI: 10.1371/journal.pbio.0030405.t001

generally highly degenerate in nature (e.g., the sumoylation site (VILAFP)Kx(EDNGP)).

Motifs in Genome-Scale Interaction Sets

Considering the genome-scale interactions, each dataset produced a number of protein sets sharing a common interaction partner: yeast, 191; fly, 632; nematode, 367; and human, 1,986. Only a small fraction of these produced one or more confident motifs (as assessed by the binomial probability): yeast, 11; fly, 26; nematode, 27; and human, 112. In all cases, known motifs were among those produced, though to varying degrees: yeast, 1 (domain set); fly, 9; nematode, 4 (domain set); and human, 48 (all significant motifs from the protein set are given in Protocol S1 and Table S1; all motifs, including those with poorer significance, are available at http://lmd.embl.de). Figure 2 shows a summary of the 26 motifs found in the fly set, highlighting the nine rediscoveries of known motifs (including one likely nuclear localization signal). The better results in human data (i.e., 48/112) are undoubtedly because the hand-curated interactions (HPRD, [14]) contain fewer errors than those from the comparatively noisy high-throughput yeast two-hybrid screens for the other organisms. Here we found motifs spanning virtually the full range of those known (SH3→PxxP, 14–3–3 proteins→RxxSxP, Clathrin→LDxL, etc.), in addition to several that appear to be novel.

Inspection showed that known motifs were typically missed because the sets contained too few sequences with the correct motif to reach significance. For example, in yeast interaction data, just four out of 23 proteins interacting with the protein phosphatase 1 (PP1) domain contained the established (RK)VxF motif. A similar situation occurred with WW domains, where no more than three instances of known motifs were found among their interaction partners. It could be that certain motifs are just too rare in the interacting set to be detected. However, it is also well established that the yeast two-hybrid system, particularly when applied in genome-screens, can miss known interactions [15] and, moreover, make false predictions that cloud the signal from true motifs. The prediction accuracy and coverage will certainly increase when more comprehensive and reliable interaction data become available. The error prone nature of the underlying yeast two-hybrid data for yeast, fly, and nematode might be expected to yield inconsistencies (i.e., different motifs for the same protein) when comparing predictions from different species. Encouragingly, however, we found very few of these, and indeed in one case (see PP1, in Experimental Testing of New Motifs), we think the apparent inconsistency corresponds to two distinct motifs that bind to the same protein, each detected in a different species.

Many of the motifs detected in the protein sets were also found when interaction partners were pooled owing to the presence of a common domain (domain sets). Frequently a more general motif was found in the domain than in the protein sets. For example, in the fly we identified the

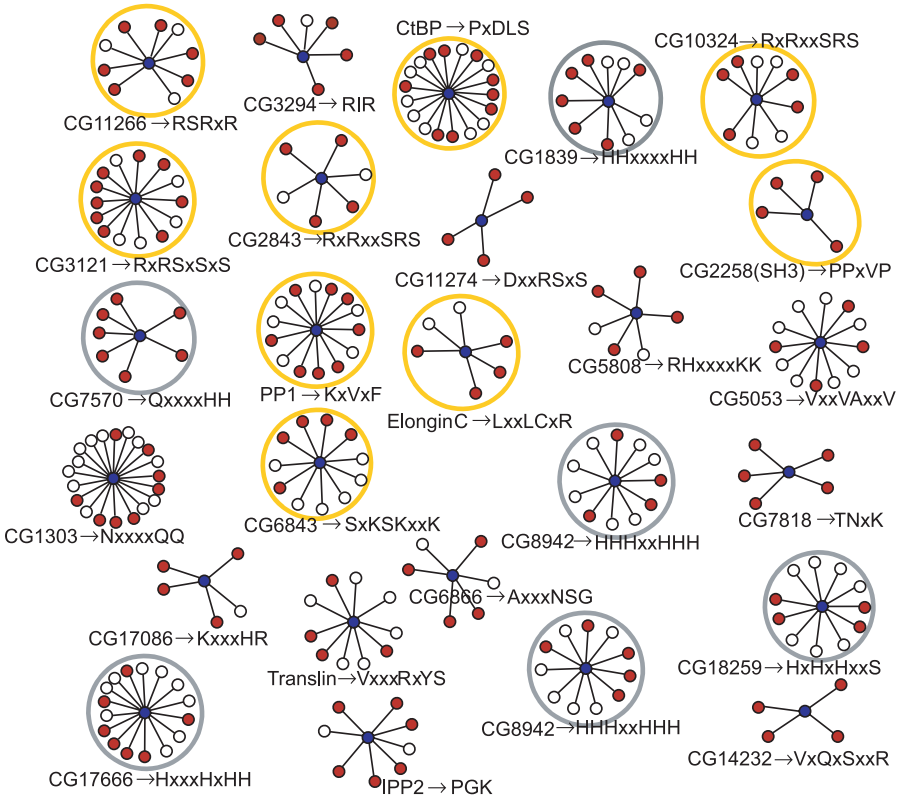

Figure 2. Overview of Motifs Found in the Fly

Significant predictions from the yeast two-hybrid set for the fly. Blue dots in the center of each cluster represent proteins with four or more interaction partners (red and white dots) containing at least one confidently predicted motif (p-value < 0.001; $S_{cons} \leq 8 \times 10^{-15}$). Partner proteins containing the motif are represented by red dots, whereas proteins lacking the motif are indicated by white dots. Clusters are labelled as gene name→detected motif. Yellow circles enclose known motifs: SH3→PxxP [38], PP1→RVxF [22], C-terminal binding protein (CtBP)→PxDLS [52], SR splicing factors RS-rich segments [53], and CG6843→SxKSKxxK, a likely nuclear localization signal. The Translin→VxxxRxYS motif was experimentally tested (Figure 3). The grey circles enclose clusters with low-complexity patterns. Two additional known motifs were also found in the fly using more relaxed criteria than those used for the other motifs in the figure: Groucho→WRPW [7] and Dynein light chain→TQT [26] as the variant A(TI)QT(DE). The latter was also identified as significant in the domain sets. Proteins are denoted either by their FlyBase accession codes or protein names when available.

DOI: 10.1371/journal.pbio.0030405.g002

canonical TQT motif as the binding site for Dynein light chain domains, but found the more specific pattern A(TI)QT(DE) for the specific Dynein homologue Cdlc2. Other motifs were seen only in one of the protein or domain sets. For example, in yeast a correct SH3 motif was only found in the domain sets, because no single SH3-domain protein had a sufficient number of interaction partners for the motifs to be found with significance. The reverse was true in the fly in which the correct SH3 motif was found only in the protein set, because the domain set had too many proteins lacking the canonical motif, meaning that the signal was lost.

There is also a problem of ambiguity in both protein and domain sets. For multi-domain proteins predicted to bind a motif, it is not possible to discern which domain is mediating the interaction. This can be partly resolved by considering domain sets, but even here there are still some examples where genuine motifs were predicted for the wrong domains. Inspection showed that this was either because the process selected the wrong domain of a frequently co-occurring pair (e.g., SH2 domains predicting to bind SH3 ligands) or selected an activator/inhibitor of the correct binding domain (e.g., protein phosphatase inhibitor 2 (IPP2) binding to PP1-like RVxF-like motifs [16]). The latter highlights the possibility of the yeast two-hybrid system identifying indirect interactions [17]. These domain ambiguities can likely only be resolved by experiment.

Experimental Testing of New Motifs

For a selection of new protein→motif associations, we tested direct binding via fluorescence anisotropy, using labelled peptides corresponding to the regions containing the predicted motifs (Protocol S1). Because the motifs we have predicted might be the lowest common denominator of what could be a slightly longer binding region, we included two additional residues to the N- and C-terminus of each peptide (extracted from the original sequence). We first selected candidate motifs in yeast, fly, and nematode based on the feasibility of expressing and purifying the common interaction partner (i.e., the protein predicted to bind the motif). Of the 55 significant novel motif predictions, only 13 contained a single globular domain, were not excessively long ($\leq$ 650 amino acids), and lacked long regions of predicted disorder or low complexity. From these we selected five that had available clones and established purification protocols. These spanned a range of novelty, ranging from variations on a known motif, to those for which there was some supporting, but not direct, evidence in the literature, to those lacking any

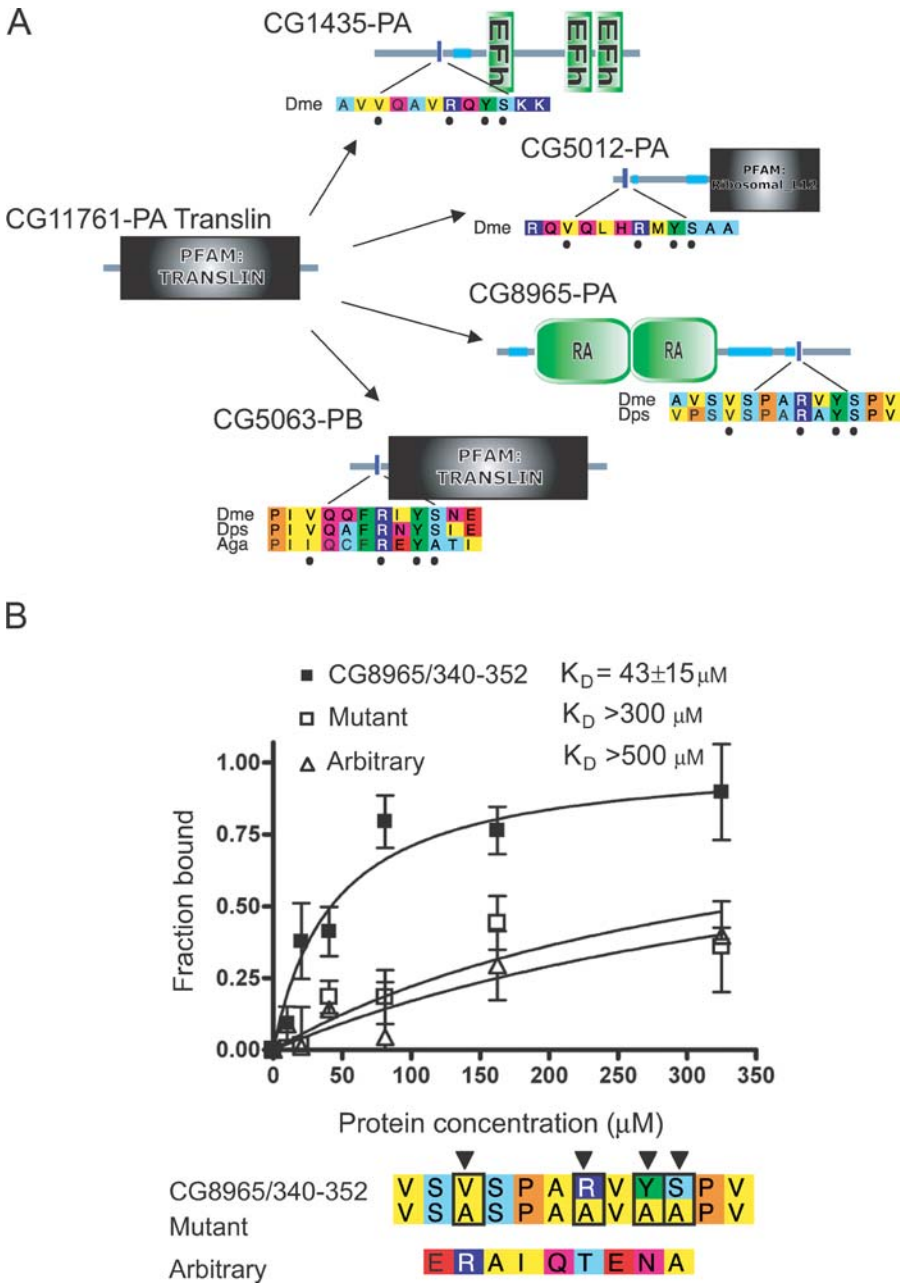

Figure 3. A Novel Fly VxxxRxYS Motif That Binds Translin

(A) Translin (left) shown surrounded by interaction partners containing the predicted motif VxxxRxYS. Proteins are shown as lines with domains (labelled shapes), predicted coiled coils (light blue/green segments), and the location of motifs (blue vertical bars). Sequences for the motif-containing region are shown aligned to the best homologues in closely related species. Amino acids are coloured according to residue type: blue, positive; red, negative; light blue, small; yellow, hydrophobic; green, aromatic; magenta, polar; and orange, proline. Those constituting the predicted motif are denoted by circles. Aga, *Anopheles gambiae;* Dme, *D. melanogaster;* Dps, *D. pseudoobscura*.

(B) Saturation curves, showing bound fraction (fluorescently labelled peptides at saturation) as a function of Translin concentration. Polarization values (mP) at zero concentration and B_{max} were normalised to give the bound fraction. K_D was computed by non-linear regression on values from three independent experiments. The lower panel shows the alignment of the native and mutated peptides together with the arbitrary peptide (selected randomly). Black triangles show positions specifying the motif (VxxxRxYS). The alignment is coloured as described in (A).

DOI: 10.1371/journal.pbio.0030405.g003

additional support. Of the five selected, we could obtain clones for four and could purify three.

We tested a highly significant Translin→VxxxRxYS motif found in fly data (Figure 3A). Translin is a protein thought to be involved in chromosomal rearrangements, and binds double-stranded RNA and DNA [18,19]. The fluorescence polarisation assay shows that it binds the peptide motif specifically compared to a mutated counterpart, or randomly selected peptides (Figure 3B). The affinity of binding ($K_D = 43 \pm 15$ μM) is within the range typical for known linear motifs when considered in isolation (5–150 μM; see Table 1). Mutated controls or arbitrarily chosen peptides do not show specific binding (Figure 3B; note that the apparent linear increase in both is due to the high protein concentrations

reached). We can only speculate what role this motif plays in modulating Translin function. However, there are several precedents for interaction motifs playing critical regulatory roles by binding to other DNA- or RNA-binding proteins, such as PCNA [20] or CtBP [21].

We also tested a DxxDxxxD motif found in 10 of 12 interaction partners of yeast protein phosphatase 1 (PP1, Figure 4A). Eight are well-known PP1 interactors, and five contain the canonical RVxF PP1-binding motif. Fluorescence polarisation shows that a peptide corresponding to the region in Scd5 binds specifically to PP1 ($K_D = 22 \pm 5$ μM), compared to arbitrary peptides (Figure 4B). Inspection of other PP1-binding proteins [22] reveals that 12 of 33 also contain the new motif, with an additional 15 containing a more relaxed pattern (permitting Glu). Deletions of the canonical RVxF motif do not always disrupt PP1-binding, and have led others to suggest additional binding sites [23]. Interestingly, deletions of some segments containing this new motif can affect PP1 binding in other proteins [22]. Other support comes from pull-down studies, which identified a similar region (RVRLDDDDE) critical for the Cdk5–PP1 interaction [24] and the recent crystal structure of human PP1 bound to a myosin-targeting subunit MYPT1, which led the authors to propose a positively charged surface on which a similar acidic stretch could interact [25]. Interestingly, the mutated control appears to retain some affinity, probably owing to the presence of additional negative charges that have not been mutated to alanine, and indeed a near match to the motif (DxxxExxD) is still present in the mutant. Arbitrary peptides did not show any specific binding (Figure 4B).

Lastly, we tested a variant of the well-known Dynein light chain binding motif. The canonical motif has a consensus sequence (KR)xTQT and mediates interactions important for cell trafficking [26]**.** We found the canonical motif in the fly, but noticed a variant, IQTE, among three partners of Cdlc2 (Dynein light chain 2), which is similar to one present in the protein swallow from *D. pseudoobscura* [27]. We could detect no binding of the Cdlc2 to fluorescently labelled peptides over a range of protein concentrations (5–400 μM). Surprisingly, a true instance of the motif, known to bind Cdlc2 in vivo and in vitro [27], also did not give a signal using this procedure, suggesting that the experimental assay might not be suitable for Dynein light chain interactions (see Protocol S1).

Other Promising Predicted Motifs

For other predictions, we scrutinized the literature for previous experiments hinting that the motifs could be genuine. For example, among several interesting predictions in the fly was an Elongin C→LxxLCxR motif, which has been described previously only as part of a longer sequence called the SOCS box [28]. Only three of four interacting proteins with the motif contain the full SOCS box. The protein lacking it (CG18171) is not well understood; the interaction has not been reported apart from the genome screen. Deletion and mutagenesis experiments have shown that this region is important for the interaction with Elongin C [29]. Our finding agrees with this and further suggests that the motif could be sufficient on its own for mediating the interaction.

We found the motif SxPxxxS in 11 of 17 interaction partners of the nematode MAP-kinase lit-1 involved in *wnt* signalling and morphogenesis (Figure 5). These include three well-known regulators/interactors of lit-1: two nuclear proteins (wrm-1 and mom-4) and another morphogenesis protein (pop-1). Deletions have already demonstrated that regions containing the motif are critical for lit-1 binding (yellow boxes in Figure 5): a 148 N-terminal segment in wrm-1 [30], a 21-residue stretch in mom-4 [31], and a 45-residue region just six residues N-terminal to the motif in pop-1 [32] all disrupt lit-1 binding.

Among several compelling new motifs in human was a T(PL)QP motif predicted to bind to the transcription factor PC4 (Positive Cofactor 4). PC4 binds double-stranded DNA and promotes the assembly of the preinitiation complex via a mechanism that is not fully understood [33,34]. The five proteins containing the putative motif all participate in transcription, but share no common globular domain that could mediate binding to PC4. Such a proline-rich motif could be a good candidate to bind one of the several aromatic patches on the surface of the PC4 protein [35].

Low-Complexity Linear Motifs

In both fly and nematode, several very significant motifs arose from regions of low sequence complexity (i.e., dominated by a few amino acids). These included examples already known to mediate interactions, and others not described previously, including His-, Ser-, Lys-, and Glu/His-rich motifs. We could find no motifs like these in random sets, which suggested that they are not the result of the general prevalence of low-complexity regions within proteins, but just what they mean is an open question. They might well be true, biologically meaningful interactions, and indeed for some sets the proteins show similarities in function. This idea is supported by the fact that many known motifs, including the protein/RNA binding RS/SR motifs [36], the Tudor domain→$(RG)_n$ [37], and SH3→poly-proline ligands [38], are themselves low complexity. Alternatively, they could be the result of some artefact of the yeast two-hybrid system. The last possibility is supported by the fact that we found fewer such motifs in the less error-prone human data.

How Many Protein–Motif Interaction Pairs Are Still to Be Found?

Both our experiments and those done previously suggest that many of our findings are genuine motifs that have not yet been reported. This raises the question as to how many new interaction motifs there are yet to be discovered. An estimate can come by considering what fraction of the previously known motifs we found and extrapolating this to the new discoveries. For example, in fly we predict 26 motifs of which nine are known, from a total number of roughly 60 that are known in this organism [5]. If we assume that *all* the remaining motifs are correct, and assume an equal distribution of motifs in fly proteins not seen in the yeast two-hybrid data (4,683/13,833), we estimate 334 additional motifs (the equivalent number for human is 405). Even the more modest assumption of between 10%–20% of the predicted motifs being correct (roughly the fraction for which we could see direct binding experimentally, which is clearly a lower estimate) gives estimates of 33–67 new motifs in fly (40–80 in human). There are very likely dozens to hundreds of new motifs to be discovered, which will correspond potentially to thousands of individual binding sites. To date we have just scratched the surface of what is likely a sophisticated network of peptide-mediated interactions in the cell.

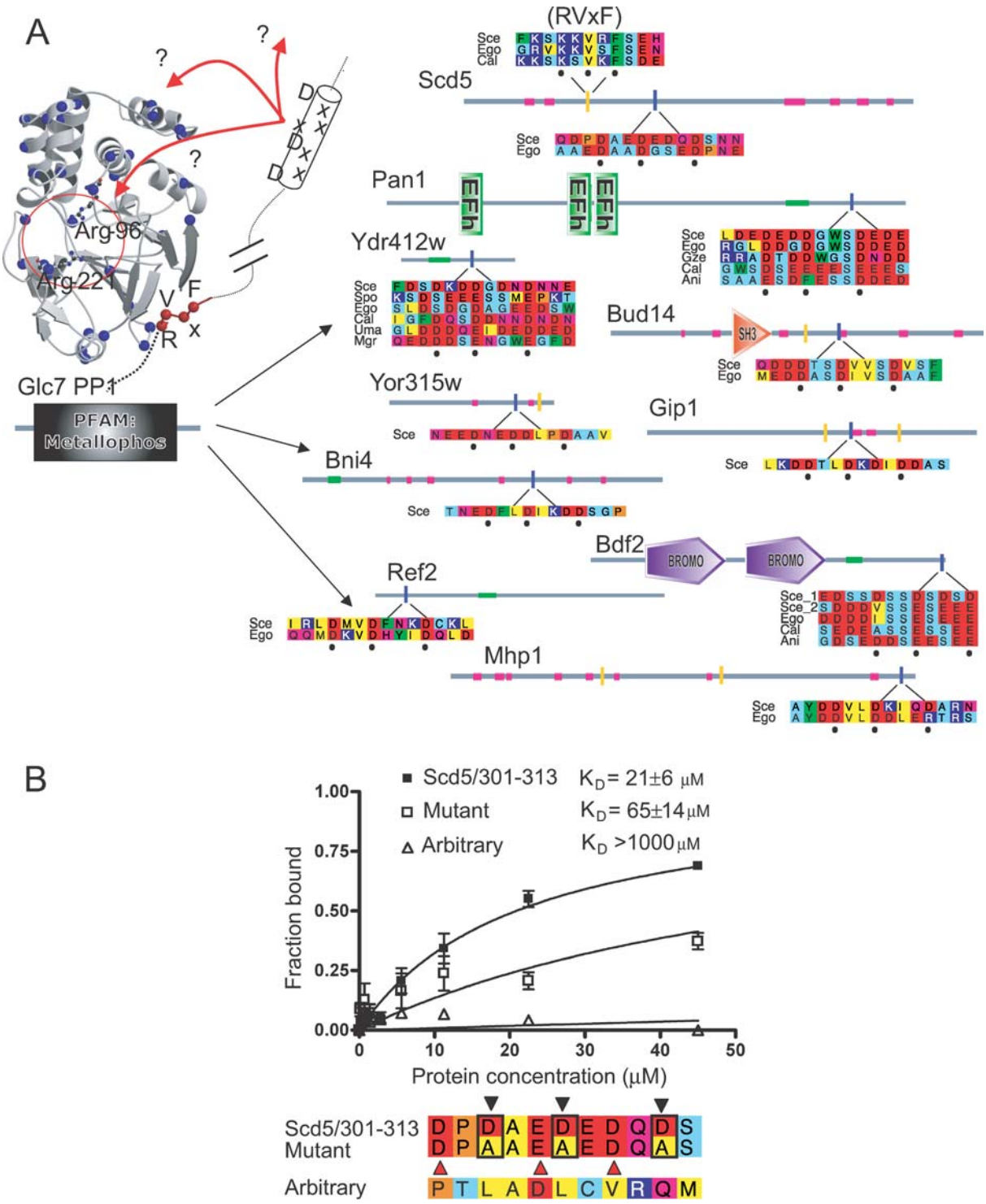

Figure 4. An Acidic Yeast PP1 Binding Motif

(A) PP1 (Glc7) with the set of interaction partners containing the DxxDxxxD motifs. Details are as for Figure 3A. Here the location of RVxF motifs (defined as matches to (RK)x_{0-1}(VI)x(FW)) are shown as yellow bars, and low-complexity regions are magenta. The figure also shows the structure of PP1 bound to RVxF (red spheres) [54] with a hypothetical helix containing the motif. Blue spheres show the location of Arg or Lys residues, and the active site is circled with critical Arginines shown in ball-and-stick. Red arrows show hypothetical interactions of the motif either with sites on PP1 or elsewhere. Ani, Aspergillus nidulans; Cal, Candida albicans; Ego, Eremothecium gossypii; Gze, Gibberella zeae; Mgr, Magnaporthe grisea; Sce, S. cerevisiae; Spo, Schizosaccharomyces pombe; Str, Salinospora tropicalis; Uma, Ustilago maydis; Xla, Xenopus laevis.

(B) Saturation curves, showing bound fraction as a function of PP1 concentration. The polarization values (mP) were normalized to an extrapolated B_{max} because B_{max} could not be reached experimentally. Other details are as given in Figure 3B. Red triangles in the lower panel show the location of the near match to the motif in the mutated sequence.

DOI: 10.1371/journal.pbio.0030405.g004

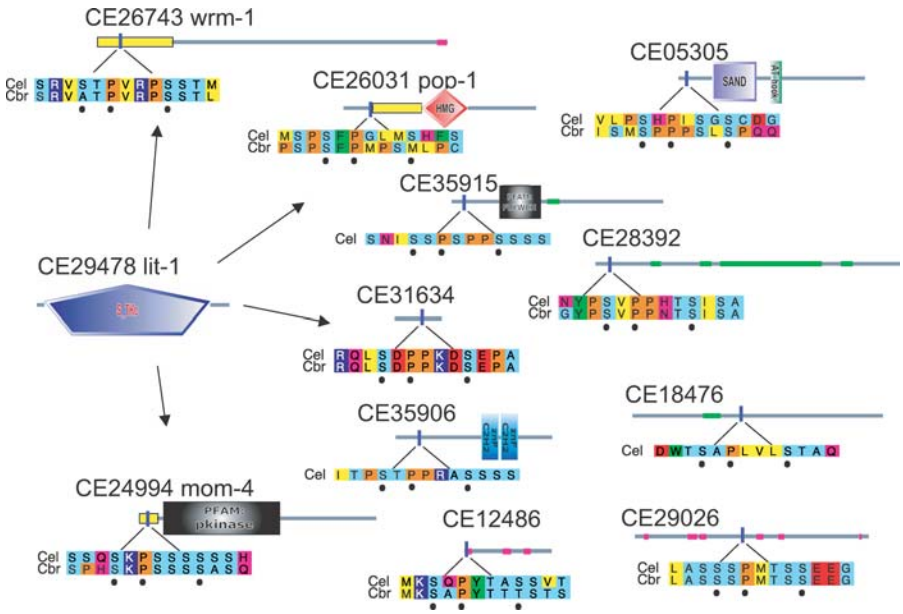

Figure 5. A Lit-1 MAP Kinase SxPxxxS Motif

The MAP kinase lit-1 surrounded by its interaction partners containing the SxPxxxS motif. Details are as for Figure 3. Yellow boxes show the location of deletion mutants known to affect the interaction. Cbr, *C. briggsae;* Cel, *C. elegans*.

DOI: 10.1371/journal.pbio.0030405.g005

Discussion

Many studies continue to highlight the importance of networks mediated by linear motifs [39,40], and each new discovery opens new lines of research into critical aspects of cell function [41]. We have shown here that these very simple features can be detected successfully, even in error-prone data, provided they occur with a sufficient frequency in otherwise unrelated proteins. The approach need not be restricted to protein-protein interactions. It can also be applied in other contexts: Any set of proteins or nucleic acids can be probed for short sequences responsible for a common biological feature (cellular location, modifications, etc.).

Both globular domains and linear motifs are modular in the sense that they are reused in different functional contexts, but they probably differ in how they arise. Domain shuffling involves duplication of part of a gene and its insertion into another. In contrast, the short length of linear motifs makes them likely to arise convergently in proteins by evolutionary drift [42]. This suggests that there are probably many near matches to the motifs just waiting for an appropriate point mutation to induce a function. They are, in effect, powerful switches for nature to explore during the evolution of complex functions. In this regard they are highly similar to transcription factor–binding sites [43,44] or microRNA target sequences [45]. In all three cases, molecular recognition is mediated by very short and fast-evolving sequences that are relatively unspecific in isolation, with more than one often being required for function. Identifying the correct sequence is a true needle-in-a-haystack problem, for nature and computational techniques alike.

New motifs are a treasure trove for investigations to deduce the molecular details of protein–protein interactions, particularly to understand those not mediated by domains alone. Given the essential regulatory functions of the motifs already known, we expect our new discoveries to have a profound impact on understanding the complex network of macromolecular interactions that exists in all living cells.

Materials and Methods

For proteins in all sets, we identified domains using SMART [46], including domains from Pfam-A [47]. We also removed regions showing similarity between members in a set of sequences using BLAST ($E \leq 0.001$) [48], which removes the redundant measurements. We used TEIRESIAS [9] to detect all non-overlapping motifs of three to eight residues, requiring at least two identical positions. The method essentially detects all motifs of a variable length (i.e., three to eight) in which positions can either be specified as a particular amino acid, or represented by a wildcard (i.e., "x"). We did not allow for conservative substitutions (e.g., D/E), and ignored any motif that occurred in fewer than three sequences in the set.

We assessed the significance of a particular motif occurring a certain number of times within a set of sequences (interaction set) using the binomial distribution:

$$P(n|M) = \binom{M}{n} p^n (1-p)^{M-n} \quad (1)$$

where p is the probability of seeing the motif in a background database, n is how often the motif was seen in the set of proteins, and M the size of the set.

The probability (p) was computed as a frequency of the motif in the background database of 15,000 randomly selected proteins. These proteins were taken from the SWISSPROT [49] and were subjected to the same filtering procedure as the test protein sets.

Values agree well with intuition: Motifs that are complex and thus rare need only be observed a few times to be significant, for example, the motif PxVPLR occurring in four out of 21 proteins gives a probability of 10^{-11}. More common motifs must be seen more often to reach the same significance; for example, the VxxR (a subset of the first motif) must be seen in 19 out of 21 to reach a similar probability.

True instances of linear motifs are typically conserved across closely related species [42]. It is thus an advantage to use the information from the same (i.e., orthologous) protein in multiple genomes. Information from orthologues can be readily combined into a single value (S_{cons}), which is the product of all binomial probabilities from the genomes considered:

$$S_{cons} = P_1 \times P_2 \times P_3 ... P_n \quad (2)$$

This procedure will decrease the final value (and thus increase the significance) for all conserved motifs, but will have no effect if the motifs (or indeed the orthologues) are missing. The combined value is no longer a true probability, because the motifs from related species are not independent, but rather are a measure of likelihood of a conserved motif to occur at random in the set. To estimate significance we thus compare the values to those generated from random sets of proteins. These combined values greatly improve the sensitivity and specificity of the procedure: More known motifs are recovered and fewer clearly false predictions are made.

To get confidence thresholds for S_{cons}, we created 50 random sets of sequences of the same number and length as seen in the interaction sets for each organism using the complete proteomes. We then ran the complete procedure for each random set and computed the distribution of S_{cons}, which gave thresholds (p-value $<$ 0.001) for each dataset: 3.0×10^{-17} for yeast, 7.5×10^{-14} for nematode, 8.0×10^{-15} for fly, and 7.0×10^{-38} for human. The differences between the thresholds are due largely to differences in the number and similarity of closely related species with complete genomes available: Four substantially similar genomes were available for human but only one for the fly and nematode.

We extracted orthologues from the STRING database [50] and aligned those using MUSCLE [51] with default parameters. We considered only closely related species because known instances of linear motifs are rarely conserved outside of them. We considered orthologues in the four other completely sequenced yeast genomes *(Kluyveromyces lactis, Ashbya gossypii, Debaryomyces hansenii,* and *Candida glabrata)* for yeast *(S. cerevisiae)* motifs, *D. pseudoobscura* for fly *(D. melanogaster), C. briggsae* for nematode *(C. elegans),* and *Mus musculus, Rattus norvegicus, Gallus gallus,* and *Fugu rubripes* for motifs found in human *(H. sapiens)* proteins.

The Linear Motif Discovery (LMD) program and all data related to this paper are available online (http://lmd.embl.de).

Supporting Information

Protocol S1. Supplementary Information

Found at DOI: 10.1371/journal.pbio.0030405.sd001 (289 KB PDF).

Figure S1. Schematic of Discovery Process of Linear Motifs Recognized by Protein Domains

Found at DOI: 10.1371/journal.pbio.0030405.sg001 (77 KB PDF).

Table S1. All Significant Motifs from the Protein Sets and the Rediscovered Motif:Domain Sssociations for Yeast, Fly, Nematode, and Human Interaction Datasets

Found at DOI: 10.1371/journal.pbio.0030405.st001 (173 KB PDF).

Acknowledgments

We are grateful to the other members of the Eukaryotic Linear Motif (ELM) consortium, particularly Rein Aasland, Pål Puntervoll, and Morten Mattingsdal (University of Bergen, Norway) for advice, to Lars Juhl Jensen (European Molecular Biology Laboratory [EMBL]) and Ewan Birney (European Bioinformatics Institute [EBI], Hinxton, United Kingdom) for detailed help with the statistics, to Hugo Ceulemans and Mathieu Bollen (Katholieke Universiteit Leuven, Belgium) for advice and PP1 clones, Sean Hooper (EMBL) for NetView used in Figure 2, and Christian von Mering (EMBL) for help defining orthologues. We thank Patrick Aloy and Peer Bork (EMBL) for a critical reading of the manuscript. Part of this work was supported by a grant from the European Commission.

Competing interests. The authors have declared that no competing interests exist.

Author contributions. VN and RBR conceived and designed the experiments. VN and ISA performed the experiments. VN and RBR analyzed the data. RL, AS, FdM, TJG, JL, and LS contributed reagents/materials/analysis tools. VN, RL, AS, and RBR wrote the paper. ■

References

1. Poglitsch CL, Meredith GD, Gnatt AL, Jensen GJ, Chang WH, et al. (1999) Electron crystal structure of an RNA polymerase II transcription elongation complex. Cell 98: 791–798.
2. Jeffrey PD, Russo AA, Polyak K, Gibbs E, Hurwitz J, et al. (1995) Mechanism of CDK activation revelated by the structure of a cyclinA-CDK2 complex. Nature 376: 313–320.
3. Pawson T, Scott JD (1997) Signaling through scaffold, anchoring, and adaptor proteins. Science 278: 2075–2080.
4. Sudol M (1998) From Src Homology domains to other signaling modules: Proposal of the 'protein recognition code'. Oncogene 17: 1469–1474.
5. Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, et al. (2003) ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res 31: 3625–3630.
6. Pawson T, Nash P (2003) Assembly of cell regulatory systems through protein interaction domains. Science 300: 445–452.
7. Paroush Z, Finley RL Jr, Kidd T, Wainwright SM, Ingham PW, et al. (1994) Groucho is required for *Drosophila* neurogenesis, segmentation, and sex determination and interacts directly with hairy-related bHLH proteins. Cell 79: 805–815.
8. Prelich G, Stillman B (1988) Coordinated leading and lagging strand synthesis during SV40 DNA replication in vitro requires PCNA. Cell 53: 117–126.
9. Rigoutsos I, Floratos A (1998) Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. Bioinformatics 14: 55–67.
10. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, et al. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403: 623–627.
11. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, et al. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A 98: 4569–4574.
12. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, et al. (2003) A protein interaction map of *Drosophila melanogaster*. Science 302: 1727–1736.
13. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, et al. (2004) A map of the interactome network of the metazoan *C. elegans*. Science 303: 540–543.
14. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, et al. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res 13: 2363–2371.
15. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, et al. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417: 399–403.
16. Terry-Lorenzo RT, Elliot E, Weiser DC, Prickett TD, Brautigan DL, et al. (2002) Neurabins recruit protein phosphatase-1 and inhibitor-2 to the actin cytoskeleton. J Biol Chem 277: 46535–46543.
17. Aloy P, Russell RB (2002) The third dimension for protein interactions and complexes. Trends Biochem Sci 27: 633–638.
18. Han J, Gu W, Hecht N (1995) Testis-brain RNA-binding protein, a testicular translational regulatory RNA-binding protein, is present in the brain and binds to the 3′ untranslated regions of transported brain mRNAs. Biol Reprod 53: 707–717.
19. Aoki K, Suzuki K, Sugano T, Tasaka T, Nakahara K, et al. (1995) A novel gene, Translin, encodes a recombination hotspot binding protein associated with chromosomal translocations. Nat Genet 10: 167–174.
20. Warbrick E (2000) The puzzle of PCNA's many partners. Bioessays 22: 997–1006.
21. Chinnadurai G (2002) CtBP, an unconventional transcriptional corepressor in development and oncogenesis. Mol Cell 9: 213–224.
22. Bollen M (2001) Combinatorial control of protein phosphatase-1. Trends Biochem Sci 26: 426–431.
23. Chang JS, Henry K, Wolf BL, Geli M, Lemmon SK (2002) Protein phosphatase-1 binding to scd5p is important for regulation of actin organization and endocytosis in yeast. J Biol Chem 277: 48002–48008.
24. Agarwal-Mawal A, Paudel HK (2001) Neuronal Cdc2-like protein kinase (Cdk5/p25) is associated with protein phosphatase 1 and phosphorylates inhibitor-2. J Biol Chem 276: 23712–23718.
25. Terrak M, Kerff F, Langsetmo K, Tao T, Dominguez R (2004) Structural basis of protein phosphatase 1 regulation. Nature 429: 780–784.
26. Lo KW, Naisbitt S, Fan JS, Sheng M, Zhang M (2001) The 8-kDa dynein light chain binds to its targets via a conserved (K/R)XTQT motif. J Biol Chem 276: 14059–14066.
27. Schnorrer F, Bohmann K, Nusslein-Volhard C (2000) The molecular motor dynein is involved in targeting swallow and bicoid RNA to the anterior pole of *Drosophila* oocytes. Nat Cell Biol 2: 185–190.
28. Kamura T, Sato S, Haque D, Liu L, Kaelin WG Jr, et al. (1998) The Elongin BC complex interacts with the conserved SOCS-box motif present in members of the SOCS, ras, WD-40 repeat, and ankyrin repeat families. Genes Dev 12: 3872–3881.
29. Zhang JG, Farley A, Nicholson SE, Willson TA, Zugaro LM, et al. (1999) The conserved SOCS box motif in suppressors of cytokine signaling binds to elongins B and C and may couple bound proteins to proteasomal degradation. Proc Natl Acad Sci U S A 96: 2071–2076.
30. Rocheleau CE, Yasuda J, Shin TH, Lin R, Sawa H, et al. (1999) WRM-1 activates the LIT-1 protein kinase to transduce anterior/posterior polarity signals in *C. elegans*. Cell 97: 717–726.
31. Shin TH, Yasuda J, Rocheleau CE, Lin R, Soto M, et al. (1999) MOM-4, a MAP kinase kinase kinase-related protein, activates WRM-1/LIT-1 kinase to

transduce anterior/posterior polarity signals in *C. elegans*. Mol Cell 4: 275–280.
32. Lo MC, Gay F, Odom R, Shi Y, Lin R (2004) Phosphorylation by the beta-catenin/MAPK complex promotes 14–3–3-mediated nuclear export of TCF/POP-1 in signal-responsive cells in *C. elegans*. Cell 117: 95–106.
33. Kretzschmar M, Kaiser K, Lottspeich F, Meisterernst M (1994) A novel mediator of class II gene transcription with homology to viral immediate-early transcriptional regulators. Cell 78: 525–534.
34. Ge H, Roeder RG (1994) Purification, cloning, and characterization of a human coactivator, PC4, that mediates transcriptional activation of class II genes. Cell 78: 513–523.
35. Brandsen J, Werten S, van der Vliet PC, Meisterernst M, Kroon J, et al. (1997) C-terminal domain of transcription cofactor PC4 reveals dimeric ssDNA binding site. Nat Struct Biol 4: 900–903.
36. Shen H, Kan JLC, Green MR (2004) Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly. Molecular Cell 13: 367–376.
37. Sprangers R, Groves MR, Sinning I, Sattler M (2003) High-resolution X-ray and NMR structures of the SMN Tudor domain: Conformational variation in the binding site for symmetrically dimethylated arginine residues. J Mol Biol 327: 507–520.
38. Musacchio A (2002) How SH3 domains recognize proline. Adv Protein Chem 61: 211–268.
39. Yaffe MB, Leparc GG, Lai J, Obata T, Volinia S, et al. (2001) A motif-based profile scanning approach for genome-wide prediction of signaling pathways. Nat Biotechnol 19: 348–353.
40. Landgraf C, Panni S, Montecchi-Palazzi L, Castagnoli L, Schneider-Mergener J, et al. (2004) Protein interaction networks by proteome peptide scanning. PLoS Biol 2: E14.
41. Benes CH, Wu N, Elia AE, Dharia T, Cantley LC, et al. (2005) The C2 domain of PKCdelta is a phosphotyrosine binding domain. Cell 121: 271–280.
42. Neduva V, Russell RB (2005) Linear motifs: Evolutionary interaction switches. FEBS Lett 579: 3342–3345.
43. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, et al. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. Nature 434: 338–345.
44. Ettwiller LM, Rung J, Birney E (2003) Discovering novel cis-regulatory motifs using functional networks. Genome Res 13: 883–895.
45. Stark A, Brennecke J, Russell RB, Cohen SM (2003) Identification of *Drosophila* microRNA targets. PLoS Biol 1: E60.
46. Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, et al. (2002) Recent improvements to the SMART domain-based sequence annotation resource. Nucleic Acids Res 30: 242–244.
47. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, et al. (2002) The Pfam protein families database. Nucleic Acids Res 30: 276–280.
48. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
49. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31: 365–370.
50. von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, et al. (2003) STRING: A database of predicted functional associations between proteins. Nucleic Acids Res 31: 258–261.
51. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.
52. Schaeper U, Boyd JM, Verma S, Uhlmann E, Subramanian T, et al. (1995) Molecular cloning and characterization of a cellular phosphoprotein that interacts with a conserved C-terminal domain of adenovirus E1A involved in negative modulation of oncogenic transformation. Proc Natl Acad Sci U S A 92: 10467–10471.
53. Cazalla D, Zhu J, Manche L, Huber E, Krainer AR, et al. (2002) Nuclear export and retention signals in the RS domain of SR proteins. Mol Cell Biol 22: 6871–6882.
54. Egloff MP, Johnson DF, Moorhead G, Cohen PT, Cohen P, et al. (1997) Structural basis for the recognition of regulatory subunits by the catalytic subunit of protein phosphatase 1. EMBO J 16: 1876–1887.
55. Lee C, Chang JH, Lee HS, Cho Y (2002) Structural basis for the recognition of the E2F transactivation domain by the retinoblastoma tumor suppressor. Genes Dev 16: 3199–3212.
56. Stomski FC, Dottore M, Winnall W, Guthridge MA, Woodcock J, et al. (1999) Identification of a 14–3–3 binding sequence in the common beta chain of the granulocyte-macrophage colony-stimulating factor (GM-CSF), interleukin-3 (IL-3), and IL-5 receptors that is serine-phosphorylated by GM-CSF. Blood 94: 1933–1942.
57. Molloy DP, Milner AE, Yakub IK, Chinnadurai G, Gallimore PH, et al. (1998) Structural determinants present in the C-terminal binding protein binding site of adenovirus early region 1A proteins. J Biol Chem 273: 20867–20876.
58. Bruning JB, Shamoo Y (2004) Structural and thermodynamic analysis of human PCNA with peptides derived from DNA polymerase-delta p66 subunit and flap endonuclease-1. Structure (Camb) 12: 2209–2219.