Peptide Recognition Module Networks: Combining Phage Display with Two-Hybrid Analysis to Define Protein-Protein Interactions

**Gary D. Bader,[4] Amy Hin Yan Tong,[1] Gianni Cesareni,[2]
Christopher W. Hogue,[5] Stanley Fields,[3] and Charles Boone[1]**

*[1]Banting and Best Department of Medical Research and Department of Molecular and
Medical Genetics, University of Toronto, Toronto, Ontario, Canada
[2]Department of Biology, University of Rome Tor Vergata,
00133 Rome, Italy
[3]Howard Hughes Medical Institute, Departments of Genome Sciences and
Medicine, University of Washington, Seattle, Washington
[4]Samuel Lunenfeld Research Institute, Mount Sinai Hospital,
Toronto, Ontario, Canada, and
[5]Department of Biochemistry, University of Toronto, Toronto, Canada*

Introduction

Many of the protein-protein interactions of macromolecular signaling complexes are mediated by domains that function as recognition modules to bind specific peptide sequences found in their partner proteins [1]. For example, SH3, WW, and EVH1 domains bind to proline-rich peptides [2–4], EH domains bind to peptides containing the NPF motif [5,6], and SH2 and FHA domains bind to peptides phosphorylated on Tyr and Thr, respectively [7,8]. For particular modules within the same family, specificity is determined by critical residues in the binding partner flanking the core peptide motif [9,10]. A major challenge is to construct protein-protein interaction networks in which every module within the predicted proteome of a sequenced organism is linked to its cognate partner.

To address this problem, we developed a four-step strategy for the derivation of protein-protein interaction networks mediated by peptide recognition modules [11–13]. First, the consensus sequences for preferred ligands for each peptide recognition module are defined by isolating 10 to 20 different peptide ligands from screens of phage display libraries. Second, the consensus sequences resulting from the phage display experiments are used to computationally derive a protein-protein interaction network that links each peptide recognition module to proteins containing a preferred peptide ligand. Third, a protein-protein interaction network is experimentally derived via large-scale two-hybrid analysis.

Fourth, the intersection of the predicted and experimental networks is determined.

As a test of this strategy, we constructed a protein interaction network for the SH3 domains of the yeast *Saccharomyces cerevisiae*. The SH3 domain is one of the more commonly used protein recognition modules. In fact, over 1500 different SH3 domains have been identified in the protein databases of eukaryotic organisms [14]. The yeast proteome contains a total 28 SH3 domains, found in 24 different proteins [15], the majority of which had been implicated in signal transduction (Bem1, Boi1, Boi2, Cdc25, Sdc25, and Sho1) or reorganization of the actin cytoskeleton (Abp1, Bud14, Cyk3, Hof1, Myo3, Myo5, Rvs167, Sla1) [16,17]. A set of eight SH3 proteins remained to be characterized (Bbc1, Bzz1, Nbp2, Yfr024c, Ygr136w, Yhl002w, Ypr154w, and Ysc84).

We were able to express 24 different SH3 domains in a soluble form as glutathione S-transferase (GST)-SH3 fusion proteins in *Escherichia coli*. Because some of the SH3 domains did not select a ligand from the nonapeptide library, we were able to obtain a consensus sequence for only a subset of 20 different SH3 domains. Most SH3 domains bind to a core PxxP ligand motif (P = proline, x = any amino acid), with particular residues that occur on either side of the core determining binding specificity. Two general classes of SH3 ligands have been defined; class I peptides conform to the general consensus RxxPxxP (R = arginine) and class II peptides conform to PxxPxR [2], Most of the yeast SH3 domains selected peptides that aligned to yield a class I or class II consensus ligand, with one to six domain-specific residues constrained outside the PxxP motif (Table I).

The consensus sequences were used to search the yeast proteome for proteins that contained potential SH3 ligands. Because hundreds of the predicted yeast proteins contain an SH3 class I and class II consensus ligand, we used a position specific scoring matrix (PSSM) to rank the peptides present in yeast proteins based upon their similarity to the peptides selected from the phage display libraries. The peptides within the top 20% of the PSSM scores captured most of the literature-validated SH3 domain interactions, and therefore this set was considered as potential ligands. The predicted protein-protein interactions were imported into the Biomolecular Interaction Network Database (BIND) [18] and formatted for visualization in the Pajek package [19], a program originally designed for visualization of social networks. The resulting phage display protein-protein interaction network contained 394 interactions among 206 proteins (Fig. 1A). Proteins are represented as nodes on the graph and the interactions represented as edges connecting the nodes.

Proteins found within highly connected subgraphs can be extracted from more complex networks by using graph

Table I Consensus Sequence of Yeast SH3 Peptide Ligands

	Class I	Class II	Unusual
Bem1-1			P P x V x P Y
Fus1			R x x R st st S l
Abp1		rk x x **p** x x **P** x rk P x w #	
Myo3	P x @ p P **P** x x **P**		
Myo5	P x @ p P **P** x x **P**		
Pex13	R x l **P** x # **P**		
Sla1-3	h R x p **P** x p **P**		
Sho1	s kr x L **P** x x **P**		
Ygr136w	R x rk **#@** x l **P**	**P** x # **P** x R p	
Ypr154w	@ kr R P p **#** x l **P**	**P** P # **P** x R P	
Yhl002w	y R p # **P** x x **P**		f R x x x h Y t
Ysc84		**P** x L **P** x R	
Yfr024c		**P** p L **P** x R P	
Rvs167	R x # **P** x p **P**	**P** P # **P** P R	
Bzz1-1	K kr x P **P** p x **P**		
Bzz1-2	kr kr p P **P** P p **#** P		
Bbc1	R kr x **P** x p **P**	**P** kr # **P** x R P	
Boi1	R x x **P** x x **P**	p **P** R x **P** r R #	
Boi2		p **p** R n **P** x R #	
Nbp2	P x R P a **P** x x **P**		

The consensus peptides were derived from an alignment of the selected phage-display peptides (x, any amino acid; lowercase letters, residues conserved in 50 to 80% of the selected peptides; uppercase letters, residues conserved in more than 80% of the selected peptides). Abbreviations for the amino acid residues are as follows: A, Ala; H, His; K, Lys; L, Leu; N, Asn; P, Pro; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; Y, Tyr; #, hydrophobic residues; @, aromatic residues. The consensus sequences corresponding to Class I peptides, first column, Class II peptides, second column; unaligned, third column.

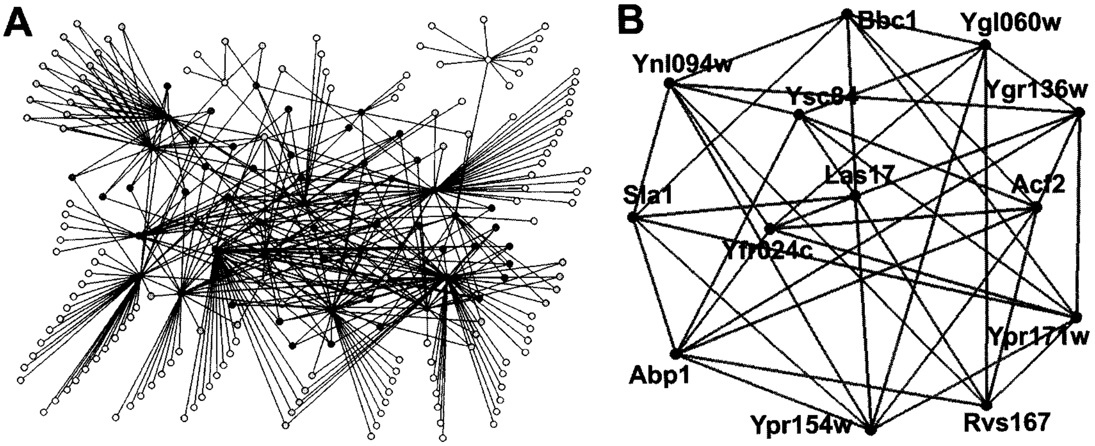

Figure 1 (A) Yeast SH3 domain protein-protein interaction network predicted via phage display selected peptides; 394 interactions and 206 proteins are shown; a network with each gene name labeled is included in the supplementary material [7]. The proteins are colored according to their k-core value (six-core=black, five-core=cyan, four-core=blue, three-core=red, two-core=green, one-core=yellow), identifying subsets of interconnected proteins in which each protein has at least k interactions. By definition, lower core numbers encompass all higher core numbers (e.g. a four-core includes all the nodes in the four-core, five-core, and six-core). The interactions of the six-core subgraph are highlighted in red. (B) The six-core subgraph derived from the phage display protein-protein interaction network, expanded to allow identification of individual proteins. The six-core subset contains eight SH3 domain proteins (Abp1, Bbc1, Rvs167, Sla1, Yfr024c, Ysc84, Ypr154w, and Ygr136w) and five proteins predicted to bind at least six different SH3 domains (Las17, Acf2, Ypr171w, Ygl060w, and Ynl094w).

theoretical algorithms. The phage display network contained a highly connected six-core subgraph, in which each protein has at least six interactions with the other proteins in the subgraph (Fig. 1B). Because the phage display network represents an integration of all potential interactions and does not take into account temporal expression or protein localization information, the six-core is subject to various biological interpretations. It may represent a single complex, provided all the proteins are co-expressed *in vivo* and all of the interactions occur simultaneously; however, it may represent multiple dimers or other oligomers, each of which forms independently under some cellular state. In any case, the presence of a highly connected core suggests a functional association between the interacting proteins. We examined 1,000 random model networks, in which a similar number of random proteins were linked to each SH3 domain. The model networks were not as highly connected as the phage display network and at most contained a four-core subgraph, indicating that the six-core within the phage display network was unlikely to occur by chance. Indeed, the six-core contains a number of functionally related proteins. At the center of the six-core is Las17, the yeast homolog of human Wiscott-Aldrich syndrome protein, which binds and activates the Arp2/3 actin nucleation complex and assembles the filamentous actin of yeast cortical actin patches [20–23]. The six-core also contains Acf2, a protein required for Las17-dependent reconstitution of actin assembly *in vitro* [24] and a set of proteins that were either implicated previously in the endocytotic role of cortical actin patches (Abp1, Sla1, Rvs167) [25–27] or found to localize to cortical patches (Bbc1, Ysc84, Ynl094w, and Ypr171w) [11,28]. Thus, the construction of a protein-protein interaction network from *in vitro* peptide binding information and the graphical analysis of its connectivity revealed known components of the yeast cortical actin patch complex.

To construct a two-hybrid protein-protein interaction network for comparison to the phage display network, we screened 18 SH3 domain baits against conventional two-hybrid libraries and an ordered genome-wide array of yeast Gal4 activation domain–open reading frame fusions [29]. The results from these screens were assembled into a network containing 233 interactions and 145 proteins (Fig. 2A). Only a subset of the interactions within the phage-display network and the two-hybrid network are expected to overlap. In particular, the phage display and two-hybrid methodologies will lead to different sets of false positives, which should exclude them from the overlap network. A total of 59 interactions in the phage display network also occurred in the two-hybrid network (Fig. 2B). All of the overlap interactions are mediated directly by SH3 domains; the precise ligand of the binding partner was predicted by the phage display analysis. Three lines of evidence suggest that the interactions within the overlap network are meaningful. First, the phage display network was highly enriched for overlap interactions when compared to the random model networks, which contained an average of 0.84 overlap interactions (SD = 1.01). Second, the overlap network was enriched for interactions validated previously in the literature, over three-fold compared to the two-hybrid network and over five-fold compared to the phage display network. Third, a focused analysis of the proline-rich peptides within Las17 revealed that the phage display ligand analysis consistently predicted the ligand fragment that showed the strongest binding.

Future experiments of this type may be able to achieve better results by optimizing specific steps. For example, some false positives in the phage display approach undoubtedly arise because the predicted ligand peptide is, in fact, buried in the core of the protein. This aspect of the analysis could be improved by assessing surface accessibility with a program

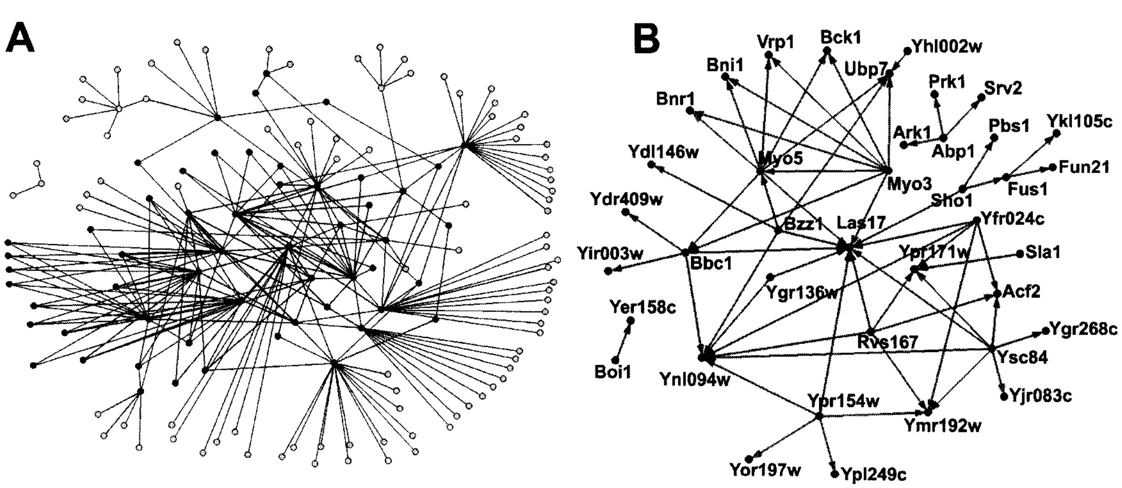

Figure 2 (A) Two-hybrid SH3 domain protein-protein interaction network. Two-hybrid results, based largely on screens with SH3 domains as baits, generated a network containing 233 interactions and 145 proteins. Proteins are colored according to their k-core (see Fig. 1A). The largest core of the two-hybrid network is a single four-core (blue nodes). Interactions common to the phage display network are highlighted in red. (B) Overlap of the protein-protein interaction networks derived from phage display and two-hybrid analysis. Expanded view of the common elements of the phage display and two-hybrid protein-protein interaction networks, 59 interactions, and 39 proteins. All of these interactions are predicted to be mediated directly by SH3 domains. The arrows point from an SH3 domain protein to the target protein. Additional evidence to support the relevance of several of these interactions is provided as supplementary material.

such as PHDacc [30], or homology models [31] of the protein could be scanned. Another means to improve proteome scanning would use a specificity and sensitivity analysis to assess what PSSM score threshold would retain the largest number of physiologically relevant interactions (true positives) and discard as many potential false positive interactions as possible. In this case, false positives can be defined operationally as those not identified within the literature or the yeast two-hybrid network. Thus, the optimization could be based on maximizing overlap with the yeast two-hybrid network or a set of confirmed interactions from a literature-based benchmark.

The overlap step could be improved in a number of ways. While the reasons for the false-positives and false-negatives of yeast two-hybrid screens seem satisfyingly orthogonal to those of the phage display predicted network, other protein interaction experimental methods, such as co-immunoprecipitation coupled with mass spectrometry [32,33], should also be evaluated. The current network representation, with a single node corresponding to a protein and a single edge corresponding to an interaction, could be much improved by making it probabilistic. The attachment of a probability value as a weight on the edges could enter into the overlap calculation to result in a more realistic model. For instance, a weight value on an edge could be high if the interaction has been characterized by several different methods, or found by multiple laboratories. These highly probable edges could be made to appear in the weighted combination of networks; in this fashion, "textbook" interactions would be included even if they were not found by both the phage display and two-hybrid derived networks. A review by Gerstein *et al.* (2002) addresses some of these points in more detail [12]. A better visualization tool that could draw networks with probabilistic information and allow one to examine parameter changes (for example, in the PSSM score threshold) in real-time would complement these method improvements and facilitate evaluation of the results.

Many of these future improvements depend on the availability of a literature-based benchmark, a manually curated collection of high-quality, expert-validated interactions. Sources of more stringently validated interactions are MIPS [17], YPD [34] and PreBIND [33]. Collecting these together in a nonredundant set creates a benchmark of over 3,300 protein-protein interactions for yeast. Because some experimental methods are more likely to yield physiologically relevant information (for example, interactions detected with full length proteins expressed at native levels), the literature benchmark could also include a reliability score for each record.

A set of over 15,000 unique protein interactions collected for yeast from the literature and from all available large-scale studies contained 519 interactions involving 364 proteins in which one interaction partner has an SH3 domain [18]. Because many of these proteins are highly conserved, it will be of interest to determine the extent to which the connectivity of the network is conserved. The prospects for applying this interaction network mapping approach to other organisms are reasonable; for example, *Caenorhabditis elegans* has only 99 SH3 domains in 77 proteins, according to the SMART database, whereas the mouse has on the order of 327 SH3 domains in 172 proteins. A map of peptide-binding module-mediated interaction networks across organisms will provide a powerful dataset to study the specificity of domain-mediated interactions, the evolution of complexity, and the biology that these interactions dictate. Finally, the systematic analysis of binding properties and protein-protein interaction networks for peptide recognition modules will enable the development of sets of dominant interfering small molecules for systematic functional interrogation of the network [35].

References

1. Pawson, T. and Scott, J. D. (1997). Signaling through scaffold, anchoring, and adaptor proteins. *Science* **278**(5346), 2075–2080.
2. Cesareni, G. *et al.* (2002). Can we infer peptide recognition specificity mediated by SH3 domains? *FEBS Lett.* **513**(1), 38–44.
3. Fedorov, A. A. *et al.* (1999). Structure of EVH1, a novel proline-rich ligand-binding module involved in cytoskeletal dynamics and neural function. *Nat. Struct. Biol.* **6**(7), 661–665.
4. Macias, M. J., Wiesner, S., and Sudol, M. (2002). WW and SH3 domains, two different scaffolds to recognize proline-rich ligands. *FEBS. Lett.* **513**(1), 30–37.
5. de Beer, T. *et al.* (1997). Molecular mechanism of NPF recognition by EH domains. *Nat. Struct. Biol.* **7**(11), 1018–1022.
6. Salcini, A. E. *et al.* (1997). Binding specificity and in vivo targets of the EH domain, a novel protein-protein interaction module. *Genes Dev.* **11**(17), 2239–2249.
7. Moran, M. F. *et al.* (1990). Src homology region 2 domains direct protein-protein interactions in signal transduction. *Proc. Natl. Acad. Sci. USA* **87**(21), 8622–8626.
8. Durocher, D. and Jackson, S. P. (2002). The FHA domain. *FEBS Lett.* **513**(1), 58–66.
9. Paoluzi, S. *et al.* (1998). Recognition specificity of individual EH domains of mammals and yeast. *EMBO J.* **17**(22), 6541–6550.
10. Panni, S., Dente, L., and Cesareni, G. (2002). In vitro evolution of recognition specificity mediated by SH3 domains reveals target recognition rules. *J. Biol. Chem.* **277**(24), 21666–21674.
11. Tong, A. H. *et al.* (2002). A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* **295**(5553), 321–324.
12. Gerstein, M., Lan, N., and Jansen, R. (2002). Proteomics integrating interactomes. *Science* **295**(5553), 284–287.
13. Legrain, P. (2002). Protein domain networking. *Nat. Biotechnol.* **20**(2), 128–129.
14. Mayer, B. J. SH3 domains: complexity in moderation. *J. Cell Sci.* **114**(Pt. 7), 1253–1263.
15. Letunic, I. *et al.* (2002). Recent improvements to the SMART domain-based sequence annotation resource. *Nucl. Acids Res.* **30**(1), 242–244.
16. *http://genome-www.stanford.edu/Saccharomyces/.*
17. Mewes, H. W. *et al.* (2002). MIPS: a database for genomes and protein sequences. *Nucl. Acids Res.* **30**(1), 31–34.
18. Bader, G. D. *et al.* (2001). BIND—the Biomolecular Interaction Network Database. *Nucl. Acids Res.* **29**(1), 242–245.
19. *http://vlado.fmf.uni-lj.si/pub/networks/pajek/.*
20. Winter, D., Lechler, T., and Li, R. (1999). Activation of the yeast Arp2/3 complex by Bee1p, a WASP-family protein. *Curr. Biol.* **9**(9), 501–504.
21. Madania, A. *et al.* (1999). The *Saccharomyces cerevisiae* homologue of human Wiskott-Aldrich syndrome protein Las17p interacts with the Arp2/3 complex. *Mol. Biol. Cell* **10**(10), 3521–3538.
22. Lechler, T., Shevchenko, A., and Li, R. (2000). Direct involvement of yeast type I myosins in Cdc42-dependent actin polymerization. *J. Cell Biol.* **148**(2), 363–373.
23. Evangelista, M. *et al.* (2000). A role for myosin-I in actin assembly through interactions with Vrp1p, Bee1p, and the Arp2/3 complex. *J. Cell Biol.* **148**(2), 353–362.
24. Lechler, T. and Li, R. (1997). In vitro reconstitution of cortical actin assembly sites in budding yeast. J. **138**(1), 95–103.
25. Lila, T. and Drubin, D. G. (1997). Evidence for physical and functional interactions among two *Saccharomyces cerevisiae* SH3 domain proteins, an adenylyl cyclase-associated protein and the actin cytoskeleton. *Mol. Biol. Cell* **8**(2), 367–385.
26. Colwill, K. *et al.* (1999). In vivo analysis of the domains of yeast Rvs167p suggests Rvs167p function is mediated through multiple protein interactions. *Genetics* **152**(3), 881–893.
27. Ayscough, K. R. *et al.* (1999). Sla1p is a functionally modular component of the yeast cortical actin cytoskeleton required for correct localization of both Rho1p-GTPase and Sla2p, a protein with talin homology. *Mol. Biol. Cell* **10**(4), 1061–1075.
28. Drees, B. L. *et al.* (2001). A protein interaction map for cell polarity development. *J. Cell Biol.* **154**(3), 549–571.
29. Uetz, P. *et al.* (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae. Nature* **403**(6770), 623–627.
30. Rost, B., Sander, C., and Schneider, R. (1994). PHD—an automatic mail server for protein secondary structure prediction. *Comput. Appl. Biosci.* **10**(1), 53–60.
31. Pieper, U. *et al.* (2002). MODBASE, a database of annotated comparative protein structure models. Nucl. Acids Res. **30**(1), 255–259.
32. Gavin, A. C. *et al.* (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**(6868), 141–147.
33. Ho, Y. *et al.* (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**(6868), 180–183.
34. Costanzo, M. C. *et al.* (2001). YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucl. Acids Res.* **29**(1), 75–79.
35. Oneyama, C., Nakano, H., and Sharma, S. V. (2002). UCS15A, a novel small molecule, SH3 domain-mediated protein-protein interaction blocking drug. *Oncogene* **21**(13), 2037–2050.