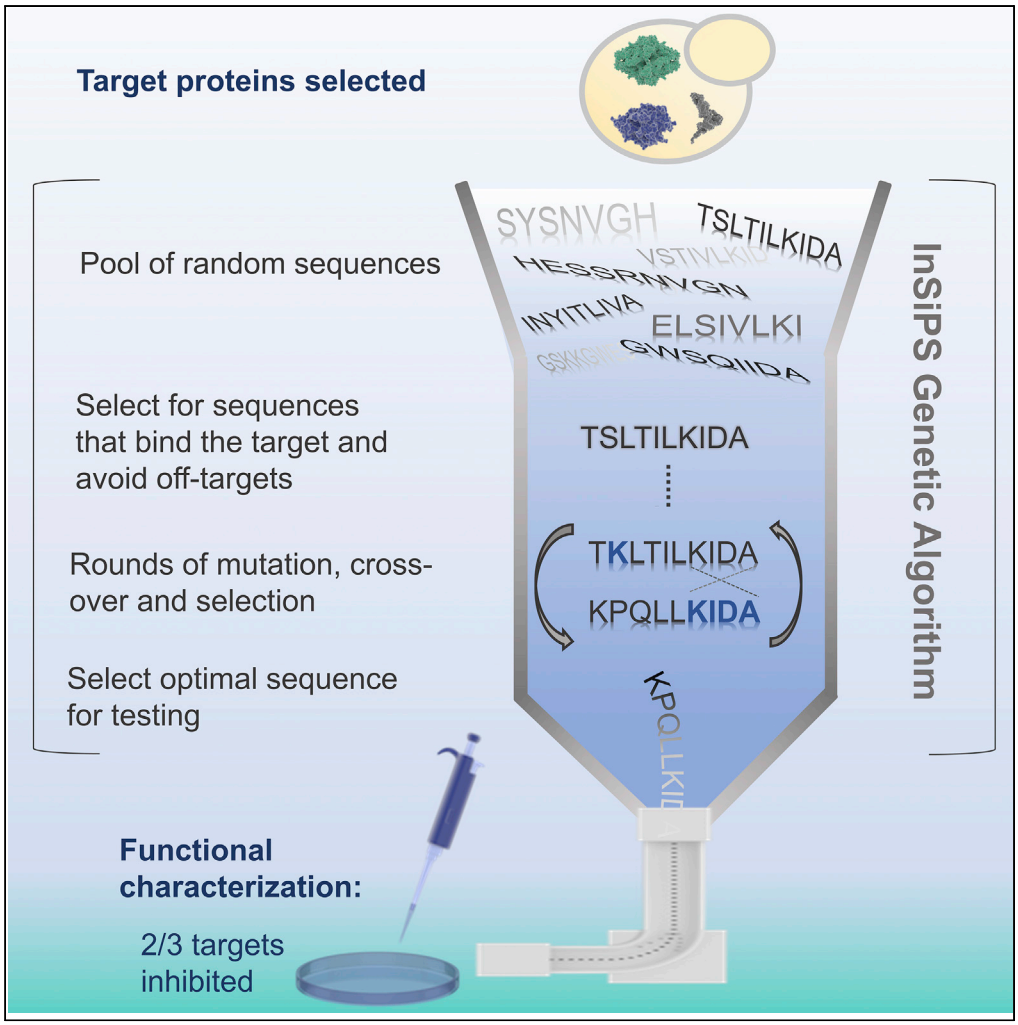
iScience

CellPress

Article

In Silico Engineering of Synthetic Binding Proteins from Random Amino Acid Sequences



Daniel Burnside, Andrew Schoenrock, Houman Moteshareie, ..., Frank Dehne, Kyle K. Biggar, Ashkan Golshani

ashkan_golshani@carleton.ca

HIGHLIGHTS

InSiPS engineers synthetic binding proteins (SBPs) using primary protein sequence

SBPs are designed to a bind a target protein and avoid "off-target" interactions

Binding and functional inhibition of two of three target proteins in yeast is demonstrated

Our new approach offers advantages over alternative tools that rely on 3D models

Burnside et al., iScience *11*, 375–387
January 25, 2019
https://doi.org/10.1016/j.isci.2018.11.038

iScience

CellPress

Article

In Silico Engineering of Synthetic Binding Proteins from Random Amino Acid Sequences

Daniel Burnside,[1,2,8] Andrew Schoenrock,[3,8] Houman Moteshareie,[1,2] Mohsen Hooshyar,[1] Prabh Basra,[1] Maryam Hajikarimlou,[1,2] Kevin Dick,[4] Brad Barnes,[3] Tom Kazmirchuk,[1] Matthew Jessulat,[5] Sylvain Pitre,[3] Bahram Samanfar,[1,6] Mohan Babu,[5] James R. Green,[4] Alex Wong,[1] Frank Dehne,[3] Kyle K. Biggar,[1,7] and Ashkan Golshani[1,2,7,9,*]

SUMMARY

Synthetic proteins with high affinity and selectivity for a protein target can be used as research tools, biomarkers, and pharmacological agents, but few methods exist to design such proteins *de novo*. To this end, the *In-Silico* Protein Synthesizer (InSiPS) was developed to design synthetic binding proteins (SBPs) that bind pre-determined targets while minimizing off-target interactions. InSiPS is a genetic algorithm that refines a pool of random sequences over hundreds of generations of mutation and selection to produce SBPs with pre-specified binding characteristics. As a proof of concept, we design SBPs against three yeast proteins and demonstrate binding and functional inhibition of two of three targets *in vivo*. Peptide SPOT arrays confirm binding sites, and a permutation array demonstrates target specificity. Our foundational approach will support the field of *de novo* design of small binding polypeptide motifs and has robust applicability while offering potential advantages over the limited number of techniques currently available.

INTRODUCTION

Proteins are diverse macromolecules that form intricate and complex protein-protein interaction (PPI) networks through selective affinity binding. These properties have driven an expansion in the field of protein and peptide design over the past decade (Kang and Saven, 2007; Pantazes et al., 2011; Saven, 2010; Yu et al., 2014). Specifically, the ability to design synthetic proteins that can bind, label, or inhibit a specified target with high affinity are of primary importance and have the potential to replace antibodies and chemical compounds in a wide range of applications. Current methods to develop engineered binding proteins include peptide aptamer selection (Colombo et al., 2015), directed evolution of display systems (Goldflam and Ullman, 2015), and computational methods, the majority of which modify naturally occurring protein folds rather than designing novel structures *ab initio* (Benjamin Stranges and Kuhlman, 2013; Karanicolas and Kuhlman, 2009; Mikut et al., 2016; Schreiber and Fleishman, 2013).

Computational protein design (CPD) can allow for the *in silico* evaluation of amino acid sequences on a scale that goes beyond the constraints of many laboratory approaches (Chica et al., 2005). Natural proteins represent only an infinitesimal portion of potential functional sequences, limiting the scope of most current CPD techniques (Woolfson et al., 2015). Many protein targets lie beyond the reach of natural protein folds or current approaches to developing binding peptides, and searching randomized sequences space has been shown to successfully yield novel functional binding proteins (Cherkasov et al., 2009; Devlin et al., 1990). It is thought that true large-scale *de novo* protein design can expand beyond the confines of biologically derived molecules into the vast space of "never-born proteins" (Li et al., 2013; Luisi et al., 2006). This unexplored sequence potential coupled with the fact that many protein-based therapeutics have been shown to be effective and well-tolerated in clinical trials (Craik et al., 2013; Otvos and Wade, 2014) have made peptides a quickly expanding category of US Food and Drug Administration (FDA)-approved drugs over the past 20 years (Fosgerau and Hoffmann, 2015; Kaspar and Reichert, 2013).

In addition to peptide therapeutics, CPD has been used in recent years for developing ligand-binding proteins (Tinberg et al., 2013), nanobiotechnology (Wilson, 2015), *de novo* enzyme design, and the development of antibody mimetics (Lao et al., 2014). Much of the recent focus has been on developing proteins to

[1]Department of Biology, Carleton University, Ottawa, ON K1S5B6, Canada

[2]Ottawa Institute of Systems Biology, Carleton University, Ottawa, ON K1S5B6, Canada

[3]School of Computer Science, Carleton University, Ottawa, ON K1S5B6, Canada

[4]Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S5B6, Canada

[5]Department of Biochemistry, Research and Innovation Centre, University of Regina, Regina, SK S4S 0A2, Canada

[6]Ottawa Research and Development Centre (ORDC), Agriculture and Agri-Food Canada, Ottawa, ON K1A 0C5, Canada

[7]Institute of Biochemistry, Carleton University, Ottawa, ON K1S5B6, Canada

[8]These authors contributed equally

[9]Lead Contact

*Correspondence: ashkan_golshani@carleton.ca

https://doi.org/10.1016/j.isci.2018.11.038

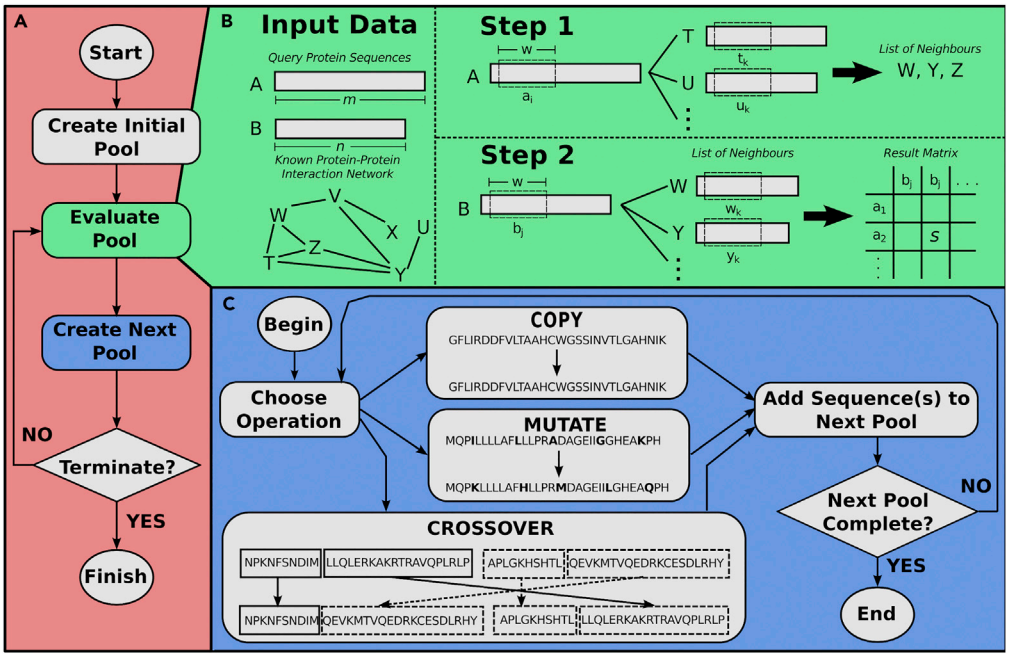


Figure 1. An Overview of the InSiPS Genetic Algorithm

(A) An initial pool of random protein sequences 150 aa in length is created. Next, the primary loop is entered: sequences in the pool are evaluated, and subsequent generations are created. This process repeats for a minimum of 250 generations until a high-fitness peptide is produced.

(B) PPI prediction. Sequences generated by InSiPS are evaluated using the Protein-Protein Interaction Prediction Engine (PIPE) (Pitre et al., 2008). PIPE requires a validated global PPI network as input. Step 1: protein *A* is compared with all proteins in the known PPI network. A sliding window is used both on *A* and the proteins in the PPI network until some segment of *A*, starting at position *i*, matches a segment of some protein *T* in the network. All known interactors of *T* (neighbors) are put into a list to be used in the next step. Step 2: protein *B* is compared with the proteins in the neighbor's list in the same manner. When a segment of *B*, starting at position *j*, is found to match a segment of a protein from this list, the result matrix is incremented at position (*i*,*j*). This matrix represents all the segments in proteins *A* and *B* that co-occur in experimentally validated PPIs and is used to predict if *A* and *B* interact. The interaction algorithm assigns a predicted interaction score between 0 and 1. Any pair scoring over 0.51 is predicted to interact with a specificity of 99.5%. The fitness of a protein sequence is calculated based on predicted interactions with targets-non-targets.

(C) Generating the next generation of candidate sequences. First, a copy, mutate, or crossover operation is randomly chosen with a preset probability proportional to the fitness of a sequence as calculated in (B). This process is repeated until the next generation is complete. The algorithm is terminated after a minimum of 250 generations when the fitness score does not improve over 50 consecutive generations.

replace targeted antibody therapies (Huang et al., 2013; Takeuchi et al., 2014). Rationally designed synthetic proteins with high affinity or specificity for a chosen target may become an important alternative to antibody-based biological drugs, which experience numerous limitations including ineffective pharmacokinetics, a relatively large size, immunological complications, ethical questions, and high production costs. Computational tools that can effectively design novel proteins specifically architected to interact with a wide range of targets are now beginning to emerge (Chevalier et al., 2017; Viart et al., 2016).

We present a powerful massively parallel computational tool that designs high-affinity binding proteins for a given target. This tool is the first of its kind as it employs a unique genetic algorithm, actively minimizes off-target interactions during the design process, and does not employ docking models or require information on the 3D structure of the target. The *In-Silico* Protein Synthesizer (InSiPS) algorithm begins with a pool of random amino acid sequences and, over many generations of fitness-based selection followed by mutation and crossover events, converges on sequences that are predicted to interact with a specified target and minimize interactions with non-targets (other proteins in the environment) (Figure 1). InSiPS uses the co-occurrence of small interacting motif pairs (Pitre et al., 2012; Schoenrock et al., 2014) to predict PPIs and intelligently design proteins with desired interaction profiles. In this way, previously "undruggable"

proteins (Ostrem and Shokat, 2016), and those that lack a well-recognized binding pocket, may be targeted by this method.

The InSiPS algorithm evaluates the predicted affinity and specificity (fitness) of hundreds of thousands of sequences over hundreds of generations, meaning upward of a billion predictions are made in a single run. This scale is difficult to achieve when using competing methods that rely on detailed 3D protein configuration data (Lewis and Kuhlman, 2011) and are thus limited by the computational restraints of working with docking models (Pierce et al., 2014).

Other CPD methods have employed sequence-based approaches to design proteins. For example, Fisher et al. (2011) utilized binary patterning of alternating polar and non-polar residues to yield biologically functional proteins in *Escherichia coli* (Fisher et al., 2011). Keating and others have developed CLEVER and CLASSY, a method of cluster expansion that maps a complex function of atomic 3D coordinates from structure-based models of protein energetics to more simple linear functions of sequence. This change dramatically speeds up scoring and has been used to successfully design highly specific synthetic protein ligands against multiple basic-region leucine zipper transcription factor families (Grigoryan et al., 2009; Negron and Keating, 2013). However, InSiPS differs from these methods and other sequence-based approaches as it uses sequential optimization of binding via a genetic algorithm without any predetermined pattern or use of structural considerations, instead relying on conserved short linear binding motifs.

Conserved short linear motifs are known to mediate PPIs in a manner that is unique from the more classically accepted interactions between large, rigid domain structures (Chica et al., 2009, 2005; Davey et al., 2012). The more flexible linear motifs are ubiquitous across higher eukaryotes and are proposed to be capable of re-wiring PPI networks through the loss or gain of these functional modules (Neduva and Russell, 2006). Our algorithm uses primary protein sequences and experimentally validated interaction networks to screen for the co-occurrence of such motifs common to known protein pairs. This technique has been used to accurately predict global PPI networks in a variety of organisms including *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Mus musculus*, and humans (with a precision of 82.1%) (Pitre et al., 2012; Schoenrock et al., 2014).

As a proof of concept, we aimed to design synthetic binding proteins (SBPs) that could functionally inhibit non-essential endogenous proteins in the yeast *Saccharomyces cerevisiae*. For ease of experimentation, 18 initial targets were chosen that fit our desired criteria of localizing to the cytoplasm, being of moderate size (<1,500 amino acids [aa]) and relatively steady abundance (500–5,000 molecules per cell) in addition to possessing readily observable phenotypes when the encoding gene is deleted or the protein product is functionally inhibited. Of 18 targets, 3 with varying fitness scores were selected for wet-laboratory experimentation: (1) Psk1, a serine/threonine kinase, which plays a role in regulating sugar metabolism; (2) Pin4, a protein involved in G2-M phase progression following DNA damage; and (3) Rmd1, a protein involved in meiotic nuclear division. The engineered SBPs were specifically designed to avoid interactions with all other yeast cytoplasmic proteins (designated as non-targets). A length of 150 aa was chosen for this project to ensure that the majority of predicted interaction motifs are included within the SBP sequence. Only a subsection of this 150-aa polypeptide, which also contains a 6xHIS tag for affinity purification, is likely required for binding. We evaluated the ability of anti-Psk1, anti-Pin4, and anti-Rmd1 to bind or inhibit their respective targets using a series of phenotypic assays and binding experiments. See Tables S2 and S3 for a complete list of InSiPS results for all target proteins considered in this study.

Our results validate the ability of this approach to computationally engineer unique SBPs. Because InSiPS does not begin with a template, but rather a pool of random sequences, the algorithm has the potential to search sequence space beyond biological barriers and is not constrained to naturally occurring sequences. Moreover, because of efficient parallelization of the algorithm, hundreds of thousands of predictions can be made during each "generation" of the genetic algorithm, allowing strong selective pressure to be applied to maximize binding affinity and specificity. Ultimately, InSiPS-engineered proteins may be useful for research, for biotechnology, or as pharmacological agents.

RESULTS

InSiPS Designs SBPs against Three Target Proteins

A preliminary run of InSiPS was used to evaluate initial SBP designs against 18 yeast targets. Our genetic algorithm assigns each candidate protein sequence a fitness (interaction) score between 0 and 1, balancing

Synthetic Binding Protein	Fitness	Target Score	Max Non-Target Score	Max Non-Target	Average Non-Target Score	Closest Yeast Homolog to SBP
Anti-Psk1	0.465	0.718	0.352	Ubi4p	0.072	Mmp1p
Anti-Pin4	0.380	0.630	0.398	Cdc39p	0.0797	Esl1p
Anti-Rmd1	0.344	0.563	0.389	Sec14p	0.132	YAP1801p

Table 1. InSiPS Predictions Suggest High Affinity of SBPs for Target Proteins and No Predicted Off-target Interactions

Higher interaction scores indicate an increased likelihood of interaction, and a score of >0.51 would be deemed likely to interact at 99.5% specificity. The fitness function weighs scores between the SBP and both targets and off-targets {Fitness(SBP) = [1 - MAX(score(Non-targets)) x score(Target)}. A higher fitness function indicates an increased likelihood of a protein having specificity for the target. The target score represents the score between the SBP and the target protein. The max non-target is score between the SBP and the next most likely off-target interaction (no interactions predicted). The average non-target score is the averaged score between the SBP and all proteins localized to the cytoplasm. The highest scoring non-target protein and yeast protein most homologous to the SBP are also listed (see Figure S1 for alignment and Table S1 for sequences of synthetic proteins).

affinity for the target with specificity (i.e., reduced binding to non-targets; see Transparent Methods for more details). In all cases, InSiPS was able to design SBPs with substantially stronger predicted affinity for the designated target than the highest likely "non-target" protein. In addition, all SBPs produced very low average non-target scores (Table 1), highlighting the selectivity of SBPs, and showed limited sequence homology to known yeast proteins (Figure S1).

Of the three designed SBPs selected for wet-laboratory experimentation, anti-Psk1 demonstrated the highest fitness (0.465) followed by anti-Pin4 (0.380) and anti-Rmd1 (0.344). All three anti-target proteins showed relatively strong interaction scores against their respective targets (Psk1, 0.718; Pin4, 0.630; Rmd1, 0.563). In all cases, the scores of all the non-target proteins are below the threshold of 0.51 at which the algorithm would predict an interaction, meaning no off-target interactions are predicted. InSiPS appears to work better for some targets than others. Anti-Psk1 exceeds the other two anti-target proteins in terms of maximizing target score and minimizing of the max non-target score. InSiPS was able to effectively design binding proteins for all targets considered (see Tables S2 and S3), but only three targets were selected for wet-laboratory experimentation.

To evaluate the novelty of the anti-target proteins, the sequences were compared against the yeast proteome using BlastP. The results show limited sequence similarity to yeast proteins (Figure S1). Anti-Psk1 most closely resembles the known Psk1 interactor Mmp1 and aligns with 34% coverage and a maximum 52% identity over 38 aa. Anti-Pin4 has sequence homology to a single yeast protein, Esl1, a known interactor of Pin4 with 29% sequence identity over a region representing only 3% of the total protein sequence. Anti-Rmd1 was found to have significant sequence homology with two proteins, Sec14 with a maximum 52% sequence identity over 29 aa, and Pmt5 with a maximum 28% sequence identity. However, neither of these proteins is known to interact with Rmd1.

Anti-Psk1 and Anti-Pin4 Show Functional Inhibition of Targets *In Vivo*

We hypothesized that the expression of our anti-target proteins may inhibit the function of the target proteins if biologically significant binding occurs *in vivo*. To this end, we expressed the SBPs in *S. cerevisiae* and performed three assays that examined conditional viability or growth rate and effects on protein distribution (Figure 2).

To test if the anti-Psk1 SBP can functionally inhibit Psk1, we induced oxidative stress and compared viability and growth rate to wild-type (WT) and *Δpsk1* strains as the loss of *PSK1* is known to increase sensitivity to UV light and oxidizing agents (Hanway et al., 2002). Anti-Psk1 expression phenocopies the *Δpsk1* mutant, consistent with inhibition of protein function (Figure 2). As seen in Figures 2A and 2C, *Δpsk1* is sensitive to UV irradiation and exposure to H_2O_2, phenotypes that were also seen when anti-Psk1 is expressed. UV exposure decreased viability in *Δpsk1* by 85% and by 83% when anti-Psk1 is expressed. To determine if the sensitivity observed was the result of the anti-Psk1 protein and not other factors, we expressed anti-Psk1 in the *Δpsk1* strain and observed no significant alteration to viability. Growth curve analysis showed that exposure to H_2O_2 decreased the growth rate of cells expressing anti-Psk1 relative to WT cells, which strongly resembles the *Δpsk1* phenotype (Figure 2C). In addition, expression of anti-Psk1 resulted in a

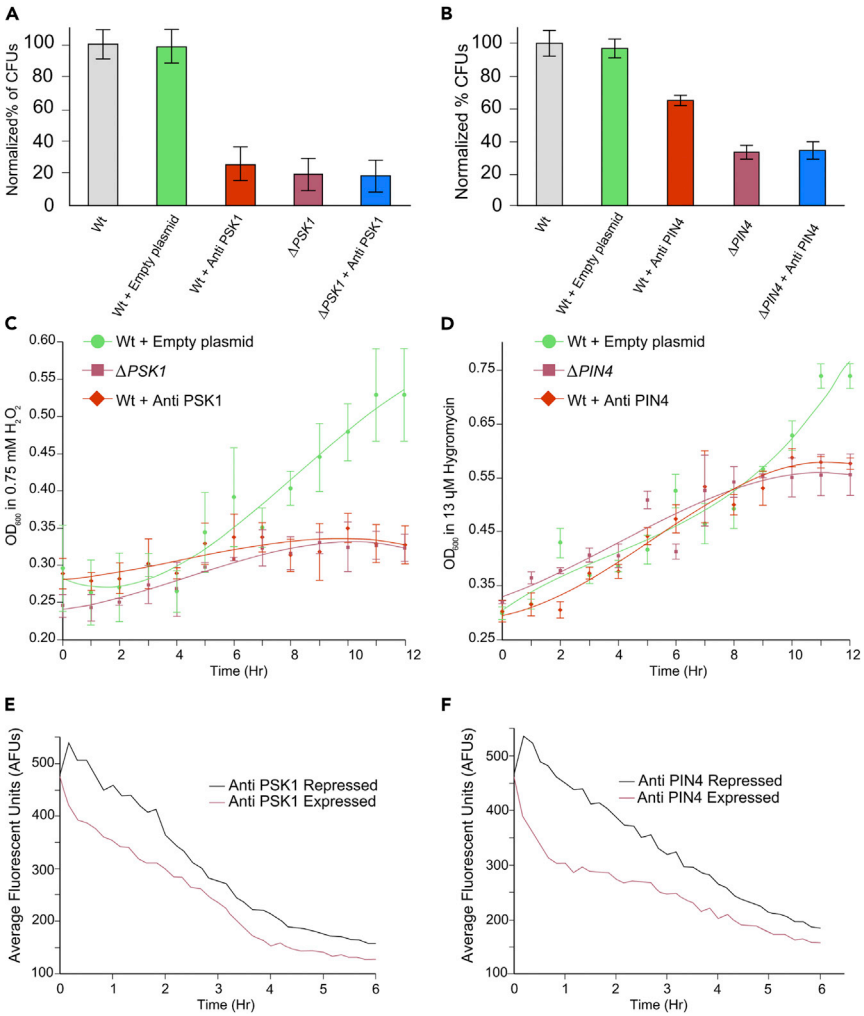


Figure 2. Strains Expressing Anti-Psk1 and Anti-Pin4 Can Phenocopy Deletion Mutants of the Target Proteins and Alter Target Protein Expression or Assembly

(A and B) Viability of cells under strain-specific stress condition shows that anti-Psk1 and anti-Pin4 expression can produce phenotypes that resemble loss of function mutants of the target proteins. (A and B) Average normalized colony-forming unit (CFU) counts from four trials are displayed as mean ± SD. Stress conditions in trial were (A) exposure to UV light for 30 s for the anti-Psk1 trial and (B) exposure to cycloheximide (65 ng/mL) for the Pin4 trial.

(C and D) Expression of anti-target SBPs produces growth defects under strain-specific stress conditions that resemble deletion of the target. Three replicates of each culture or condition were grown for 12 h in liquid YPG (Yeast extract/Peptone/Galactose) + drug, media and OD600 was measured hourly. This experiment was repeated three times. Error bars represent SD among replicates, and a polynomial line of best fit is presented. (C) *Δpsk1* sensitivity to H_2O_2 resembles the phenotype of strains expressing anti-Psk1. Cells were grown in media containing 0.75 mM H_2O_2. (D) *Δpin4* sensitivity to 13 μM hygromycin resembles the phenotype of strains expressing anti-Pin4.

(E and F) Observed alteration of fluorescence profile of GFP-tagged targets when anti-target proteins are expressed. Aliquots of WT + anti-target protein cells from the same culture were used to inoculate complete media with either 4% galactose (where anti-target SBP is expressed) or 4% glucose (where anti-target SBP is repressed), and overall fluorescent signal from three independent cultures for each condition were measured over time and normalized to the growth rate. See also Figure S2 showing that anti-Pin4 can cause sensitivity to arsenite.

significant change (net –decrease) in the fluorescent signal of GFP-tagged-Psk1 protein, suggesting possible aggregation or degradation of the target (Figure 2E).

We performed the same assays as above to test for any functional inhibition of Pin4 by the anti-Pin4 protein. The deletion of *PIN4* is known to cause yeast cells to become sensitive to inhibitors of protein

synthesis such as cycloheximide and hygromycin B (Brown et al., 2006). Culturing in the presence of 65 ng/mL cycloheximide decreased viability in *Δpin4* by 67%, WT+ anti-Pin4 by 35%, and *Δpin4*+ anti-Pin4 by 34% (Figure 2B). Growth curves show similar sensitivity of WT+ anti-Pin4 and *Δpin4* to hygromycin B (Figure 2D). These results show that anti-Pin4 sensitizes cells to translational inhibitors, suggesting functional inhibition of Pin4. In addition, cells expressing anti-Pin4 become sensitive to arsenite in a manner similar to those expressing *Δpin4* (Figure S2). The expression of anti-Pin4 decreased the net fluorescent signal of GFP-tagged Pin4. These results together suggest partial functional inhibition of target protein function.

The third target selected for experimentation, Rmd1, was analyzed to detect if phenotypic changes can be produced through expression of the anti-Rmd1 protein. However, no significant alteration to conditional viability or sensitivity to β-mercaptoethanol or L-1,4-dithiothreitol exposure similar to *Δrmd1* was observed and no significant change to the fluorescent profile of GFP-tagged Rmd1 was detected when anti-Rmd1 was expressed. For these reasons, we chose not to experiment further using the anti-Rmd1 peptide as the aim of this project was to demonstrate functional inhibition through binding. This observation suggests that not all designed anti-target peptides are functionally effective.

Yeast Two-Hybrid Analysis Indicates Binary Interactions between Target/Anti-target Proteins

Together, the results displayed in Figure 2 suggest that anti-Psk1 and anti-Pin4 possibly bind to and alter the endogenous functionality of their respective targets. To confirm that binary PPIs between Psk1/anti-Psk1 and Pin4/anti-Pin4 occur *in vivo*, we performed a series of yeast two-hybrid (Y2H) assays. Three reporter genes were present in our Y2H strain, which are all induced by Gal4 reconstitution, and three independent reporter assays were employed to test for interactions (see Transparent Methods for details).

In this way, the reconstitution of GAL4 by a physical interaction between the bait (target) and prey (anti-target) will induce growth on minimal media lacking uracil, activate lacZ activity, and provide resistance to 3-aminotriazole. All three reporter assays showed binding signals indicating physical interactions between the two target-anti-target combinations *in vivo* (Figure 3). Figure 3C indicates that binding affinity between anti-Pin4/Pin4 may be lower than the anti-Psk1/Psk1 affinity as this β-galactosidase assay is the most quantifiable of the three Y2H assays employed.

Peptide SPOT Array Analysis Shows that Binding on Targets Occurs at Predicated Loci

Positive results in our Y2H assays further support binding between target and anti-target proteins in a biological system. To probe these interactions *in vitro* and to further evaluate target interaction sites, a walking peptide SPOT array was used (Figure 4) (Jia et al., 2005). Because InSiPS predicts regions on both the target and anti-target proteins responsible for binding, it provides a starting point to probe the binding regions. Using InSiPS-predicted interaction sites, we designed walking peptide arrays that probed the predicted interaction site of the target and flanking regions using 18-aa motifs shifting at single amino acid intervals. As seen in Figures 4A and 4B, very specific residues on both interacting partners are proposed to facilitate binding (dark green regions). Interactions between both anti-Psk1/Psk1 and anti-Pin4/Pin4 were shown to occur within or directly adjacent to Protein-Protein Interaction Prediction Engine-predicted interaction regions (Figure 4). This further supports the biological validity of our CPD method for engineering binding proteins and demonstrates the specificity of the engineered proteins.

A very specific site was predicted to mediate the Psk1/anti-Psk1 interaction. This interaction region spans residues 1209–1247 in Psk1 and residues 47–69 (22 amino acids long) on the anti-Psk1 designed protein. Peptides spanning the Psk1 target region and flanking residues were probed with purified 6xHis-tagged anti-Psk1. As expected our observations using a walking array indicate an overlap between the predicted interaction site and the site identified by the walking array.

Interestingly, the predicted interacting region on anti-Psk1 residues is in an area that does not have significant sequence homology to Mmp1 (Figures S1 and S3), the protein in the yeast proteome with the greatest sequence homology to anti-Psk1. Here we see how InSiPS demonstrates its ability to focus on specific small interacting motifs that facilitate direct protein binding.

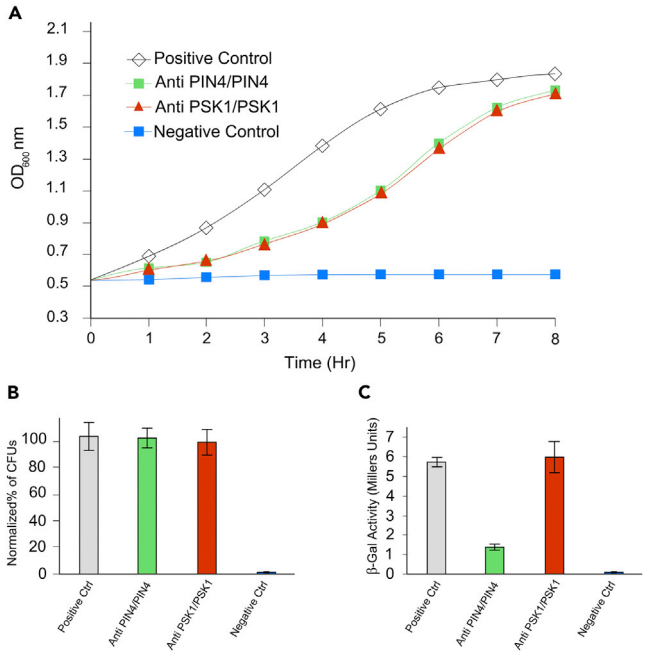


Figure 3. Yeast Two-Hybrid Assay Indicates Physical Interactions between Target and Anti-target Proteins

(A and B) (A) Positive Y2H results using uracil reporter assay. A growth curve in minimal media lacking uracil shows that Pin4/anti-Pin4 and Psk1/anti-Psk1 bait-prey combinations grow better than the negative control strain, indicating a PPI between bait and prey proteins through expression of the URA3 reporter. Triplicate trials produced similar positive results, but the results from a single trial are shown. (B) Positive Y2H result for Pin4/anti-Pin4 and Psk1/anti-Psk1 bait-prey combinations based on resistance to 3-aminotriazole (3-AT). Normalized colony-forming unit (CFU) counts for triplicate trials on minimal media lacking histidine +25 mM 3-AT resistance in test bait-prey combinations and in the positive control are presented as mean ± SD.

(C) Positive Y2H result for Pin4/anti-Pin4 and Psk1/anti-Psk1 bait-prey combinations using a β-galactosidase reporter. Miller units are used to quantify β-galactosidase activity by measuring the hydrolysis of ortho-Nitrophenyl-β-galactoside (ONPG) spectrophotometrically. Relative β-gal activity (fold change) from triplicate trials is shown relative to negative control ± SD. The Psk1/anti-Psk1 interaction produced a stronger signal than Pin4/anti-Pin4.

Walking SPOT array analysis of the predicted Pin4 interaction region indicated a single motif that facilitates the observed binding between Pin4 and anti-Pin4, which again corresponded closely to the region predicted by InSiPS. A single region spanning residues 440–466 demonstrated binding affinity for the anti-target protein in the walking array (Figure 4D). This further supports the premise that InSiPS can successfully engineer proteins that bind through short interaction motifs.

Specific binding of both anti-Psk1 and anti-Pin4 synthetic proteins to short subsequences within the target protein was observed in the walking peptide SPOT array (Figures 4C and 4D). We further studied the interaction between Psk1/anti-Psk1 as this peptide demonstrated the highest fitness score (Table 1), strongest Y2H signal (Figure 3C), and greatest functional inhibition of the target (Figures 1A and 1B).

The walking array results showed that the highest binding intensity occurred between anti-Psk1 and a truncated version of Psk1 spanning residues 1207–1224 (Figure 4C). We probed this interacting region using a permutation array to determine which residues are the most essential for binding by making single amino acid substitutions at all positions and monitoring the effect on binding affinity.

Permutation Array Shows High Specificity of Subsequence in the Psk1-Binding Motif

Our permutation analysis showed that amino acids 1218–1224 on Psk1 appear to be the most essential to the interaction (Figure 5C). This region that has three identical residues with the highest scoring predicted off-target interactor Ubi4, 1219K, 1221L, and 1223D (Figures 5A and 5D). However, the residues that have the greatest influence on the affinity between Psk1/anti-Psk1 binding are two histidines at 1220H and 1222H such that any substitution at these loci abolishes the interaction (Figure 5C). Notably, on Ubi4, the

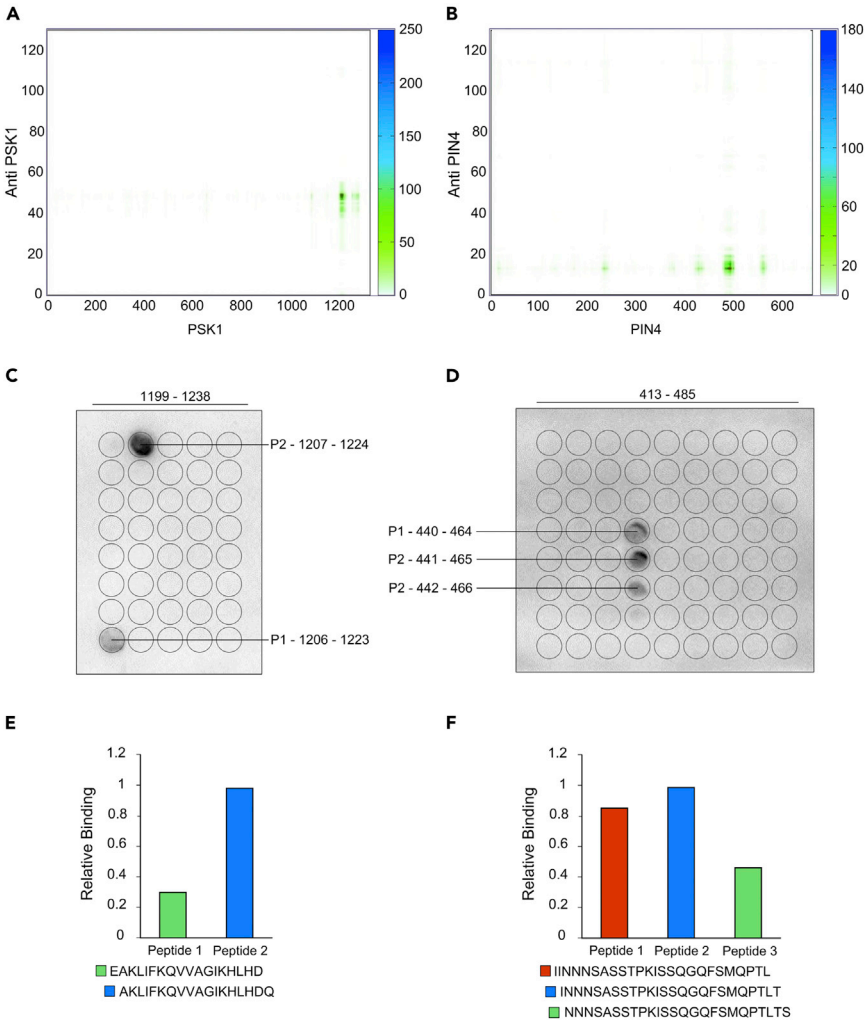


Figure 4. Walking Peptide SPOT Arrays Indicate Specific Binding Regions

SPOT arrays containing 18-aa-long printed peptides corresponding to subsequences from within the predicted interaction regions of target protein at single amino acids intervals.

(A and B) Predicted interaction matrices highlight the predicted interaction regions between target (x axis) and anti-target (y axis). (A) The anti-Psk1/Psk1 interaction site was predicted to occur between residues 1209–1246 of the PSK1 protein. (B) The Pin4/anti-Pin4 interaction site was predicted to occur between residues 472–506 on Pin4.

(C and D) SPOT arrays of predicted target binding sites and flanking regions probed with 6xHis-tagged anti-target proteins followed by detection using an anti-His antibody. (C) Specific binding of the anti-Psk1 protein to the target was detected between amino acids 1204–1228 (D) Specific binding of the anti-Pin4 protein to the target was detected between amino acids 436–458.

(E and F) Relative binding of SPOT array peptides indicates highly specific binding regions with highest relative binding. See also Figure S3 for analysis of the anti-Psk1 predicted interaction site.

off-target protein predicted to be the most likely to interact with anti-Psk1, the corresponding residues are glutamine (Q) and glutamate (E). Importantly, both single-amino-acid substitutions, H→Q and H→E, demonstrated significantly decreased affinity for anti-Psk1 protein suggesting relatively lower affinity for the Ubi4 (Figure 5D). The specificity of our peptides is shown by the limited number of acceptable substitutions at these two key loci. Despite the sequence homology to this region on Ubi4, key residues 1222H and 1224H are expected to limit binding by anti-Psk1.

To gauge the affinity of the anti-Psk1 protein against the predicted Psk1-binding site, fluorescent peptides corresponding to the Psk1 1207–1224 amino acid sequence were synthesized and purified for binding

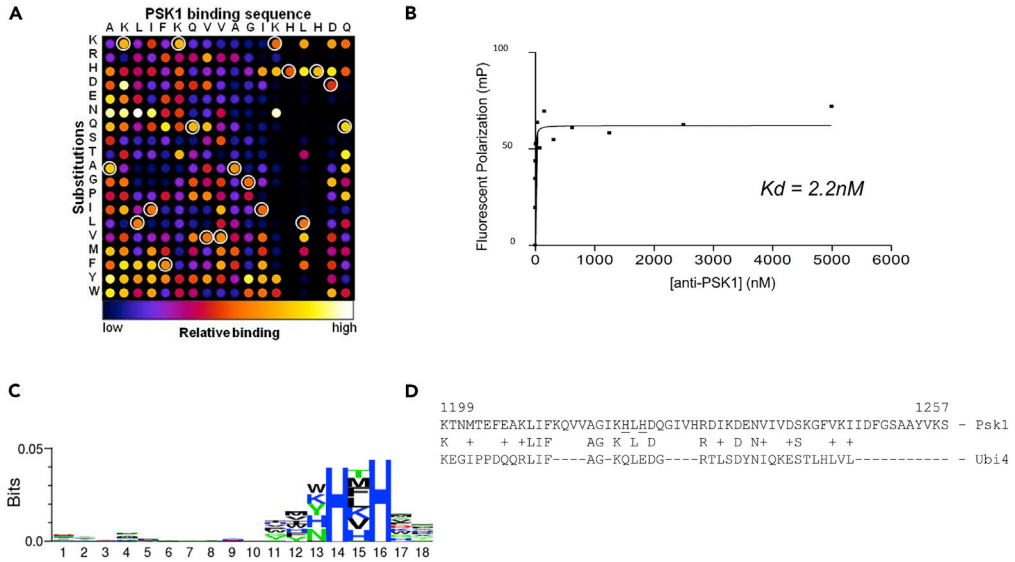


Figure 5. Characterization of Psk1 Interaction Motif and Binding Affinity

(A) Permutation array based on relative binding of the anti-Psk1 binding peptide to the Psk1 interaction motif (1207–1224). Peptides that correspond to the WT Psk1 sequence are outlined in white. Spot coloring is based on the relative binding intensity for each permutated position.

(B) Binding curve of anti-Psk1 with the Psk1-binding motif. Shown in the diagram are equilibrium isotherms for the Psk1(1207–1224) peptide from fluorescent polarization.

(C) Anti-Psk1 recognition motif based on a position-specific scoring matrix created from the permutation array.

(D) Sequence homology between the anti-Psk1-binding site of Psk1 and closest homolog Ubi4. Histidine residues at positions 14 and 16 of the anti-Psk1-binding motif are not conserved in the Ubi4 protein sequence.

studies in solution. As shown in Figure 5B, binding of the Psk1 (1207–1224) peptide to anti-Psk1 assumed saturable patterns with a *Kd* value of 2.2 nM. Together these results demonstrate the ability of InSiPS to engineer a small synthetic binding peptide with limited sequence homology to natural proteins that can bind with high affinity to the designated targets.

DISCUSSION

We have designed and validated a unique computational tool, InSiPS, for engineering proteins that bind specified protein targets with high affinity. InSiPS offers multiple strengths over current CPD tools because of its unique methodology. By functioning as a genetic algorithm and analyzing thousands of candidate sequences over hundreds of generations, InSiPS can engineer high-confidence binding proteins, which limits the need for costly and time-intensive speculative laboratory trials using classical enrichment techniques. Because the algorithm analyzes PPIs using patterns of primary amino acid sequences, it is not limited to working on proteins with known high-resolution 3D structures or domains (Zhang et al., 2013) and is free of the requirement for known 3D pockets. In this way it expands our ability to tackle diverse protein targets. InSiPS also has the unique advantage of actively avoiding interactions with thousands of off-target proteins. This robust algorithm has broad applicability and could theoretically be used to target various proteins. These features, coupled with the wide variability of potential future applications (for example, in the areas of biomarkers and pharmacological agent development), make InSiPS a powerful CPD tool.

The efficacy of the InSiPS algorithm is demonstrated here through the successful engineering, production, and biological analysis of two SBPs, anti-Psk1 and anti-Pin4, which successfully bind to respective targets Psk1 and Pin4. Each binding protein was the manifestation of hundreds of generations of the genetic algorithm and upward of a billion PPI predictions. Both SBPs had limited sequence homology to endogenous yeast proteins (Figures S1 and S2) and demonstrated high fitness scores (0.465, 0.380), predicting strong affinity for their target proteins and low affinity for all other proteins localized to the yeast cytoplasm. Consecutive biological assays verified the ability of these proteins to bind their targets both *in vivo* and *in vitro* and inhibit their natural biological functions.

Initial experiments indicated that the expression of anti-target proteins anti-Psk1 and anti-Pin-4 could functionally inhibit target proteins and produce phenotypes that resembled Δ*psk1* and Δ*pin4*, respectively (Figures 2A–2D). Y2H analysis indicated binary interactions between target and anti-target proteins *in vivo* (Figure 3), and a peptide SPOT array was probed *in vitro* to identify specific binding motifs on Psk1 and Pin4 responsible for mediating interactions (Figure 4). These results suggest that binary interactions are occurring between anti-Psk1/Psk1 and anti-Pin4/Pin4 within predicted target regions, but further study is needed to understand binding dynamics.

The anti-Psk1/Psk1 interaction was further probed as it exhibited the highest fitness score (Table 1), binding affinity (Figure 5), Y2H signal (Figure 3C), and functional inhibition of the target (Figures 1A and 1B). Permutation analysis showed that a single amino acid change to the target binding site can prevent binding of anti-target proteins (Figure 5D). This demonstrates the ability of the genetic algorithm to "evolve" proteins that possess significant binding specificity by searching beyond natural sequence configurations. Fluorescence polarization of the Psk1-binding motif and anti-Psk1 protein demonstrated a saturable pattern with a *Kd* value of 2.2 nM (Figure 5B). The strong affinity of the anti-Psk1 protein for its target coupled with the specificity observed in the permutation analysis lends further support to the ability of the InSiPS algorithm to efficiently design protein sequences with desired interaction properties.

The current study furthers the emerging field of *de novo* binding protein or peptide design, which strives to explore beyond natural protein sequence space and create functional high-specificity proteins that often are not found in nature. The majority of previous approaches to protein engineering have involved intelligent manipulation of naturally occurring proteins, but CPD is now quickly entering a new era of *de novo* design. Most major advances in *de novo* CPD over the past few years have focused on mastering protein folding and structure prediction and using these principles to design simple novel protein structures (Huang et al., 2016). For the most part, these newly designed proteins are structural with limited functionality and have simply served to lay the groundwork for future endeavors. Fundamental protein structures such as barrels (Huang et al., 2015; Thomson et al., 2014), helical bundles (Huang et al., 2014), protein nanomaterials (King et al., 2012), and oligomers (Boyken et al., 2016) have been developed and may eventually be used for a variety of future applications as we delve further into sequence space and take advantage of the scalability and specificity of polypeptides. However, at this point, very few novel functional polypeptides have been developed computationally *de novo*.

Proteins or peptides with short highly specific binding motifs, like the ones developed in this study, could be developed to function as aptamers, research tools, biomarkers, pharmacological agents, or more. Peptide aptamers continue to be developed for a variety of industrial (Colombo et al., 2015) and medicinal (Hanley-Bowdoin and Lopez-Ochoa, 2015) applications, but the field of using proteins as pharmacological agents has expanded significantly over the past 5–7 years. Peptide therapeutics can be tailored to maximize compatibility, stability, and potency with relative ease, and there are currently over 60 available FDA-approved peptide drugs with another 500+ progressing through the development stage (Kaspar and Reichert, 2013; Rafferty et al., 2016).

Other potential applications of this methodology include the development of biosensors for previously unreachable biomarkers or the intelligent design of peptide aptamers. Certain small protein targets such as the medically relevant angiotensinogenase renin have proven difficult to probe with traditional approaches but can be detected using peptide-based aptamers developed using cDNA display techniques (Biyani et al., 2016). Peptides designed to function as therapeutics can outperform relatively larger (~150 kDa) antibodies in terms of bioavailability, tumor penetration, and production efficiency and have been used to develop binding assays, cancer therapy, drug delivery, and *in vivo* imaging (Yu et al., 2017). Peptides, which were previously thought of as ineffective pharmacologically due to poor delivery mechanisms and rapid degradation or clearance (Otvos and Wade, 2014), are changing pre-conceptions as synthetic peptide production decreases in cost, new delivery systems are developed, and biological production using vectors such as viruses or genetic manipulation via CRISPR-Cas9 become more realistic possibilities.

To summarize, we have presented a unique CPD tool that employs a massively parallel genetic algorithm and PPI prediction tool to engineer binding peptides against protein targets. We demonstrated that two synthetic proteins engineered by InSiPS can bind endogenous yeast proteins at predicted motifs and inhibit functionality. Further work will examine if the technique can work in other species and explore

the range of potential targets. We invite those interested in using InSiPS to contact the corresponding author. By combining constructive techniques such as InSiPS with modeling technologies, future techniques will aim to develop highly stable, specific binding proteins for a range of applications.

Limitations of the Study

Our method has furthered the field of *de novo* computational design of short binding polypeptides and offers promising preliminary findings, but more work is required to expand the reach and efficiency of the method. The current framework restricts InSiPS to working on annotated proteins within known PPI networks. The technology may be applicable in other systems, but this has not yet been shown. The algorithm also does not directly consider protein stability but combines aspects of known functional motifs to confer bioactivity. Because InSiPS functions in the realm of flexible linear motifs, limited structural considerations are required. Future work will examine incorporating additional structural prediction tools to predict stability and folding patterns.

Another limitation may be the applicability of our approach to tackle different targets. In the current study, we started with three potential targets. In our first attempt, we generated high-scoring anti-peptides with limited off-target interactions for two of three targets. As each round of computation begins with a random pool of sequences, further iterations of InSiPS may identify suitable anti-peptides for Rmd1. However, with low sample numbers it remains difficult to speculate about the broad applicability of our approach to different proteins. InSiPS may also be constrained if the desired target is a member of a protein family with highly similar sequences or shares significant sequence similarity with other proteins in the cell. Last, the specificity of these proteins remains unclear despite being engineered to avoid off-target interactions. Only indirect and predicted evidence that our anti-target proteins avoid off-target interactions is provided in this article. We did not observe significant changes in observable phenotypes when the target protein is deleted and the anti-target protein is expressed (Figures 2A and 2B) and show specificity through the binding observed in the SPOT arrays (Figures 4 and 5).

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Transparent Methods, three figures, three tables, and one data file and can be found with this article online at https://doi.org/10.1016/j.isci.2018.11.038.

ACKNOWLEDGMENTS

The authors would like to acknowledge Dr. Shelley Hepworth for providing plasmids for Y2H analysis. The authors would also like to thank Dr. Michael Downey for providing GFP strains. Research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

AUTHOR CONTRIBUTIONS

Conceptualization, A.S., D.B., F.D, A.G., A.W., J.R.G., B.B, K.D., B.S., and S.P.; Methodology, A.S., D.B., H.M., K.K.B, P.B., M.Hajikarimlou., A.W., M.J., and M.B.; Software, A.S. and F.D.; Validation, D.B. and A.S.; Investigation, D.B., H.M., A.S., P.B., M.Hajikarimlou., M.Hooshyar., and T.K.; Resources, A.G., F.D., A.W., K.K.B., and M.B.; Data Curation, A.S. and D.B; Writing – Original Draft, D.B. and A.S.; Writing – Reviewing and Editing, D.B., A.S., J.R.G., K.K.B., H.M., A.W., B.S., A.G., and F.D.; Visualization, H.M., D.B., K.K.B., and A.S.; Supervision, A.G. and F.D.; Project Administration, A.G. and F.D.; Funding Acquisition, F.D., A.G., K.K.B., A.W., M.B., J.R.G., and B.S.

DECLARATIONS OF INTERESTS

A.G. and F.D. are co-founders of Designed Biologics Inc.

Received: June 14, 2018
Revised: October 19, 2018
Accepted: November 28, 2018
Published: January 25, 2019

REFERENCES

Benjamin Stranges, P., and Kuhlman, B. (2013). A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds. Protein Sci. *22*, 74–82.

Biyani, M., Kawai, K., Kitamura, K., Chikae, M., Biyani, M., Ushijima, H., Tamiya, E., Yoneda, T., and Takamura, Y. (2016). PEP-on-DEP: a competitive peptide-based disposable electrochemical aptasensor for renin diagnostics. Biosens. Bioelectron. *84*, 120–125.

Boyken, S.E., Chen, Z., Groves, B., Langan, R.A., Oberdorfer, G., Ford, A., Gilmore, J.M., Xu, C., DiMaio, F., Pereira, J.H., et al. (2016). De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. Science *352*, 680–687.

Brown, J.A., Sherlock, G., Myers, C.L., Burrows, N.M., Deng, C., Wu, H.I., McCann, K.E., Troyanskaya, O.G., and Brown, J.M. (2006). Global analysis of gene function in yeast by quantitative phenotypic profiling. Mol. Syst. Biol. *2*, 2006.0001.

Cherkasov, A., Hilpert, K., Jenssen, H., Fjell, C.D., Waldbrook, M., Mullaly, S.C., Volkmer, R., and Hancock, R.E.W. (2009). Use of artificial intelligence in the design of small peptide antibiotics effective against a broad spectrum of highly antibiotic-resistant superbugs. ACS Chem. Biol. *4*, 65–74.

Chevalier, A., Silva, D.-A., Rocklin, G.J., Hicks, D.R., Vergara, R., Murapa, P., Bernard, S.M., Zhang, L., Lam, K.-H., Yao, G., et al. (2017). Massively parallel de novo protein design for targeted therapeutics. Nature *550*, 74.

Chica, C., Diella, F., and Gibson, T.J. (2009). Evidence for the concerted evolution between short linear protein motifs and their flanking regions. PLoS One *4*, e6052.

Chica, R.A., Doucet, N., and Pelletier, J.N. (2005). Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design. Curr. Opin. Biotechnol. *16*, 378–384.

Colombo, M., Mizzotti, C., Masiero, S., Kater, M.M., and Pesaresi, P. (2015). Peptide aptamers: the versatile role of specific protein function inhibitors in plant biotechnology. J. Integr. Plant Biol. *57*, 892–901.

Craik, D.J., Fairlie, D.P., Liras, S., and Price, D. (2013). The future of peptide-based drugs. Chem. Biol. Drug Des. *81*, 136–147.

Davey, N.E., Van Roey, K., Weatheritt, R.J., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H., and Gibson, T.J. (2012). Attributes of short linear motifs. Mol. Biosyst. *8*, 268–281.

Devlin, J.J., Panganiban, L.C., and Devlin, P.E. (1990). Random peptide libraries: a source of specific protein binding molecules. Science *249*, 404–406.

Fisher, M.A., Mckinley, K.L., Bradley, L.H., Viola, S.R., and Hecht, M.H. (2011). De novo designed proteins from a library of artificial sequences function in escherichia coli and enable cell growth. PLoS One *6*, 15364.

Fosgerau, K., and Hoffmann, T. (2015). Peptide therapeutics: current status and future directions. Drug Discov. Today *20*, 122–128.

Goldflam, M., and Ullman, C.G. (2015). Recent advances toward the discovery of drug-like peptides de novo. Front. Chem. *3*, 69.

Grigoryan, G., Reinke, A.W., and Keating, A.E. (2009). Design of protein-interaction specificity gives selective bZIP-binding peptides. Nature *458*, 859–864.

Hanley-Bowdoin, L., and Lopez-Ochoa, L. (2015). Peptide aptamers that bind to the rep proteins of ssDNA viruses. US Pat. 9,102,705.

Hanway, D., Chin, J.K., Xia, G., Oshiro, G., Winzeler, E.A., and Romesberg, F.E. (2002). Previously uncharacterized genes in the UV- and MMS-induced DNA damage response in yeast. Proc. Natl. Acad. Sci. U S A *99*, 10605–10610.

Huang, P.-S., Oberdorfer, G., Xu, C., Pei, X.Y., Nannenga, B.L., Rogers, J.M., DiMaio, F., Gonen, T., Luisi, B., and Baker, D. (2014). High thermodynamic stability of parametrically designed helical bundles. Science *346*, 481–485.

Huang, P.-S., Feldmeier, K., Parmeggiani, F., Fernandez Velasco, D.A., Höcker, B., and Baker, D. (2015). De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. Nat. Chem. Biol. *12*, 29–34.

Huang, P.-S., Boyken, S.E., and Baker, D. (2016). The coming of age of de novo protein design. Nature *537*, 320–327.

Huang, Y., Jiang, Y., Wang, H., Wang, J., Shin, M.C., Byun, Y., He, H., Liang, Y., and Yang, V.C. (2013). Curb challenges of the "Trojan Horse" approach: smart strategies in achieving effective yet safe cell-penetrating peptide-based drug delivery. Adv. Drug Deliv. Rev. *65*, 1299–1315.

Jia, C.Y.H., Nie, J., Wu, C., Li, C., and Li, S.S.-C. (2005). Novel Src homology 3 domain-binding motifs identified from proteomic screen of a pro-rich region. Mol. Cell. Proteomics *4*, 1155–1166.

Kang, S.G., and Saven, J.G. (2007). Computational protein design: structure, function and combinatorial diversity. Curr. Opin. Chem. Biol. *11*, 329–334.

Karanicolas, J., and Kuhlman, B. (2009). Computational design of affinity and specificity at protein-protein interfaces. Curr. Opin. Struct. Biol. *19*, 458–463.

Kaspar, A.A., and Reichert, J.M. (2013). Future directions for peptide therapeutics development. Drug Discov. Today *18*, 807–817.

King, N.P., Sheffler, W., Sawaya, M.R., Vollmar, B.S., Sumida, J.P., Andre, I., Gonen, T., Yeates, T.O., and Baker, D. (2012). Computational design of self-assembling protein nanomaterials with atomic level accuracy. Science *336*, 1171–1174.

Lao, B.B., Drew, K., Guarracino, D.A., Brewer, T.F., Heindel, D.W., Bonneau, R., and Arora, P.S. (2014). Rational design of topographical helix mimics as potent inhibitors of protein-protein interactions. J. Am. Chem. Soc. *136*, 7877–7888.

Lewis, S., and Kuhlman, B. (2011). Anchored design of protein-protein interfaces. PLoS One *6*, e20872.

Li, Z., Yang, Y., Zhan, J., Dai, L., and Zhou, Y. (2013). Energy functions in de novo protein design: current challenges and future prospects. Annu. Rev. Biophys. *42*, 315–335.

Luisi, P.L., Chiarabelli, C., and Stano, P. (2006). From never born proteins to minimal living cells: two projects in synthetic biology. Orig. Life Evol. Biosph. *36*, 605–616.

Mikut, R., Ruden, S., Reischl, M., Breitling, F., Volkmer, R., and Hilpert, K. (2016). Improving short antimicrobial peptides despite elusive rules for activity. Biochim. Biophys. Acta *1858*, 1024–1033.

Neduva, V., and Russell, R.B. (2006). Peptides mediating interaction networks: new leads at last. Curr. Opin. Biotechnol. *17*, 465–471.

Negron, C., and Keating, A.E. (2013). Multistate Protein Design Using CLEVER and CLASSY, pp. 171–190.

Ostrem, J.M.L., and Shokat, K.M. (2016). Direct small-molecule inhibitors of KRAS: from structural insights to mechanism-based design. Nat. Rev. Drug Discov. *15*, 771–785.

Otvos, L., and Wade, J.D. (2014). Current challenges in peptide-based drug discovery. Front. Chem. *2*, 1–4.

Pantazes, R., Grisewood, M., and Maranas, C. (2011). Recent advances in computational protein design. Curr. Opin. Struct. Biol. *21*, 467–472.

Pierce, B.G., Wiehe, K., Hwang, H., Kim, B.H., Vreven, T., and Weng, Z. (2014). ZDOCK server: Interactive docking prediction of protein-protein complexes and symmetric multimers. Bioinformatics *30*, 1771–1773.

Pitre, S., Alamgir, M., Green, J.R., Dumontier, M., Dehne, F., and Golshani, A. (2008). Computational methods for predicting protein-protein interactions. Adv. Biochem. Eng. Biotechnol. *110*, 247–267.

Pitre, S., Hooshyar, M., Schoenrock, A., Samanfar, B., Jessulat, M., Green, J.R., Dehne, F., and Golshani, A. (2012). Short co-occurring polypeptide regions can predict global protein interaction maps. Sci. Rep. *2*, 1–10.

Rafferty, J., Nagaraj, H., McCloskey, A.P., Huwaitat, R., Porter, S., Albadr, A., and Laverty, G. (2016). Peptide therapeutics and the pharmaceutical industry: barriers encountered translating from the laboratory to patients. Curr. Med. Chem. *23*, 4231–4259.

Saven, J.G. (2010). Computational protein design: advances in the design and redesign of biomolecular nanostructures. Curr. Opin. Colloid Interface Sci. *15*, 13–17.

Schoenrock, A., Samanfar, B., Pitre, S., Hooshyar, M., Jin, K., Phillips, C.A., Wang, H., Phanse, S., Omidi, K., Gui, Y., et al. (2014). Efficient prediction of human protein-protein interactions at a global scale. BMC Bioinformatics *15*, 1–22.

Schreiber, G., and Fleishman, S.J. (2013). Computational design of protein-protein interactions. Curr. Opin. Struct. Biol. *23*, 903–910.

Takeuchi, T., Popiel, H.A., Futaki, S., Wada, K., and Nagai, Y. (2014). Peptide-based therapeutic approaches for treatment of the polyglutamine diseases. Curr. Med. Chem. *21*, 2575–2582.

Thomson, A.R., Wood, C.W., Burton, A.J., Bartlett, G.J., Sessions, R.B., Brady, R.L., and Woolfson, D.N. (2014). Computational design of water-soluble -helical barrels. Science *346*, 485–488.

Tinberg, C.E., Khare, S.D., Dou, J., Doyle, L., Nelson, J.W., Schena, A., Jankowski, W., Kalodimos, C.G., Johnsson, K., Stoddard, B.L., et al. (2013). Computational design of ligand-binding proteins with high affinity and selectivity. Nature *501*, 212–216.

Viart, B., Dias-Lopes, C., Kozlova, E., Oliveira, C.F.B., Nguyen, C., Neshich, G., Chávez-Olórtegui, C., Molina, F., and Felicori, L.F. (2016). EPI-peptide designer: a tool for designing peptide ligand libraries based on epitope-paratope interactions. Bioinformatics *32*, 1462–1470.

Wilson, C.J. (2015). Rational protein design: developing next-generation biological therapeutics and nanobiotechnological tools. Wiley Interdiscip. Rev. Nanomed. Nanobiotechnol. *7*, 330–341.

Woolfson, D.N., Bartlett, G.J., Burton, A.J., Heal, J.W., Niitsu, A., Thomson, A.R., and Wood, C.W. (2015). De novo protein design: how do we expand into the universe of possible protein structures? Curr. Opin. Struct. Biol. *33*, 16–26.

Yu, K., Liu, C., Kim, B.-G., and Lee, D.-Y. (2014). Synthetic fusion protein design and applications. Biotechnol. Adv. *33*, 155–164.

Yu, X., Yang, Y.-P., Dikici, E., Deo, S.K., and Daunert, S. (2017). Beyond antibodies as binding partners: the role of antibody mimetics in bioanalysis. Annu. Rev. Anal. Chem. *10*, 293–320.

Zhang, Q., Petrey, D., Deng, L., Qiang, L., Shi, Y., and Thu, C. (2013). Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature *495*, 556–560.

ISCI, Volume *11*

Supplemental Information

In Silico Engineering of Synthetic Binding Proteins from Random Amino Acid Sequences

Daniel Burnside, Andrew Schoenrock, Houman Moteshareie, Mohsen Hooshyar, Prabh Basra, Maryam Hajikarimlou, Kevin Dick, Brad Barnes, Tom Kazmirchuk, Matthew Jessulat, Sylvain Pitre, Bahram Samanfar, Mohan Babu, James R. Green, Alex Wong, Frank Dehne, Kyle K. Biggar, and Ashkan Golshani

A Anti-Psk1

- Mmp1p [Saccharomyces cerevisiae CEN.PK113-7D]

Score	Expect	Method	Identities	Positives	Gaps
36.2 bits(82)	0.015	Compositional matrix adjust.	14/27(52%)	18/27(66%)	0/27(0%)

```
Anti Psk1  65   MAQCAPEEACQYPVRRSYGLHATNCIE  91
               + QC  E +CQYPV  SY LHA+  I+
MMP1      120  VVQCGAELSCQYPVSGSYALHASRFID  146
```

B Anti-Pin4

- Esl1p [Saccharomyces cerevisiae S288c]

Score	Expect	Method	Identities	Positives	Gaps
28.1 bits(61)	0.60	Composition-based stats.	11/38(29%)	17/38(44%)	0/38(0%)

```
Anti Pin4 65   AHTKMGAAQNYDCKLYFGLKTQIWVHFCVQCLQAETNN  102
               A    G+ +NY+C LYF +    W+       +   NN
Esl1p    933   AAASKGSDENYNCTLYFVIDATSWLRHFAHIFKLAKNN  970
```

C Anti-Rmd1

- phosphatidylinositol/phosphatidylcholine transfer protein SEC14 [Saccharomyces cerevisiae S288c]

Score	Expect	Method	Identities	Positives	Gaps
36.6 bits(83)	6e-04	Compositional matrix adjust.	15/29(52%)	16/29(55%)	0/29(0%)

```
Anti Rmd1     2   NMVWEIAQVVQYRLPMCCSWGPHDTEQKC  30
                  N+VWE   VVQYRLP C     H  E  C
Sec14       146  NLVWEYESVVQYRLPACSRAAGHLVETSC-- 174
```

- putative dolichyl-phosphate-mannose-protein mannosyltransferase PMT5 [Saccharomyces cerevisiae S288c]

Figure S1. BlastP analysis of anti-target proteins with closest yeast homolog shown, related to Table 1. Closest yeast homolog of anti-Psk1 (A), anti-Pin4 (B) and anti-Rmd1 (C) are shown.

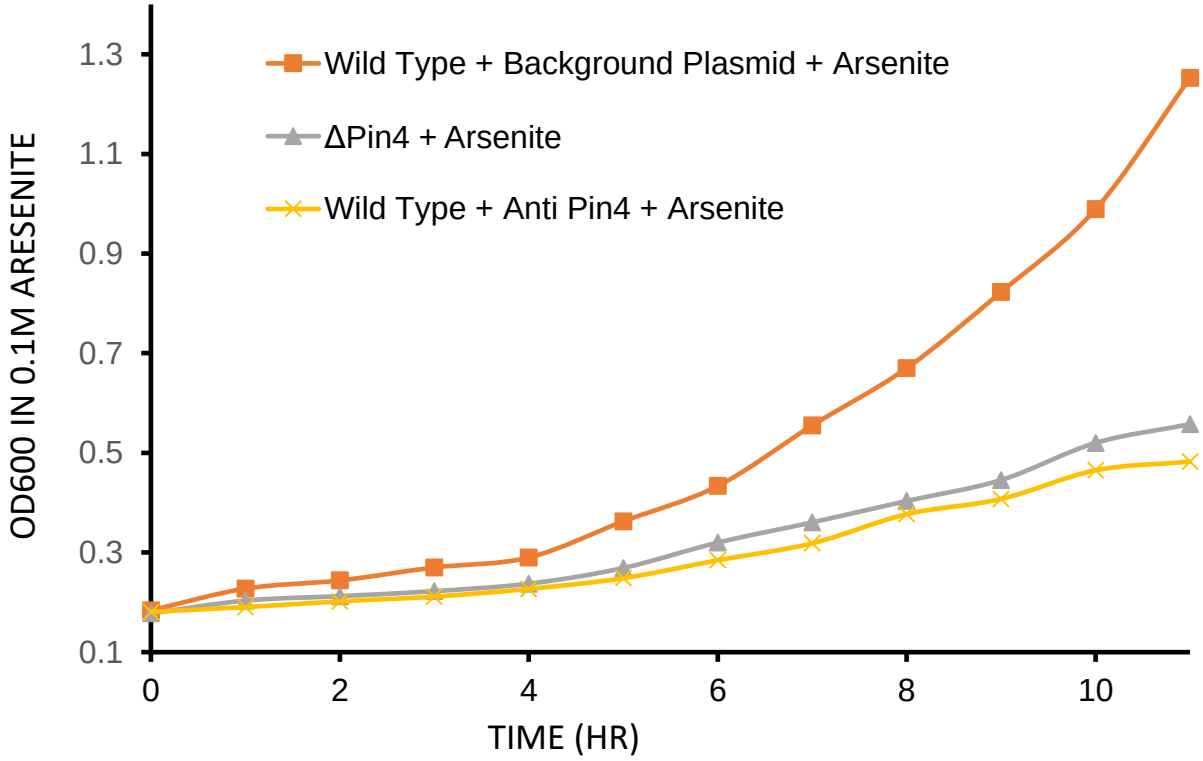


Figure S2. Drug sensitivity growth curve of the anti-Pin4 protein causes sensitivity to arsenite (0.1M) similar to *Δpin4*, related to Figure 2. A single large culture (40mL) of each condition is represented but multiple repeats demonstrated the same trend.

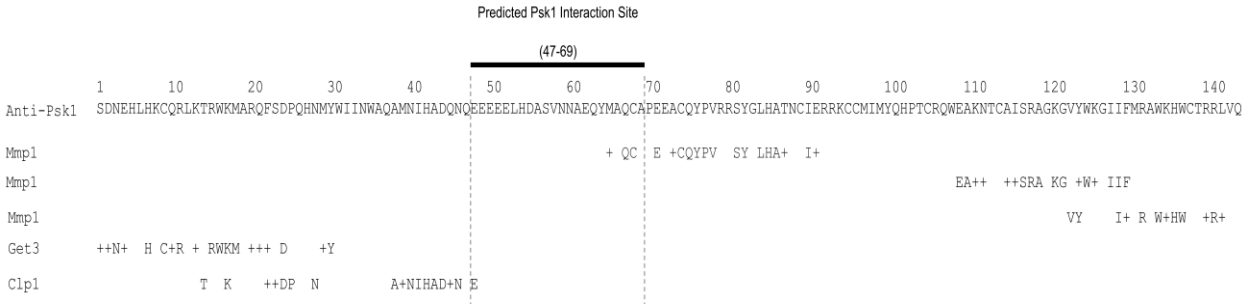


Figure S3. Sequence similarity of anti-Psk1 to known yeast proteins show that there is limited sequence homology within the predicted Psk1 binding site, related to Figures 4-5.

Table S1. InSiPS output results for three targets chosen for wet-lab experimentation including protein sequences, related to Table 1. Also included are predicted fitness (specificity), interaction scores for the target protein, highest scoring off-target protein, and the average non-target PIPE score. A protein pair with a PIPE score over 0.51 is predicted to interact with a specificity of 99.5%.

Target Protein	Fitness	Target Score	Max non-target score	Avg non-target score	Sequence of anti-target protein
Psk1	0.465	0.718	0.352 Ubi4	0.072	SDNEHLHKCQRLKTRWKMARQFSDPQHNMYWII NWAQAMNIHADQNQEEEEELHDASVNNAEQYM AQCAPEEACQYPVRRSYGLHATNCIERRKCCMIM YQHPTCRQWEAKNTCAISRAGKGVYWKGIIFMR AWKHWCTRRLVQ [17.3kD)
Pin4	0.380	0.630	0.398 Cdc39	0.0797	IFIYGDRLLDQRSQQCEQQRPKGQDHHLTPFANTS RKDAEKHFFCPETFTYHLSVCTNPAGAMNAHTK MGAAQNYDCKLYFGLKTQIWVHFCVQCLQAETN NQDADIQAYMIPAHRGLRYKSPHALPLSLGKRYF CKTGKSWEA [16.6kD]
Rmd1	0.344	0.563	0.389 YAP1801	0.132	VNMVWEIAQVVQYRLPMCCSWGPHDTEQKCPE ACQFAQRIAPGDGRRGAFGKFHRGFFGRHVAHCR NPWFDSCASLNVTVQPMEPHLTRFTGKKKWVVQ VETKHQKPTLAQQHVTQPQQKPGAQPLKWEMRF DRHWICNNGARF [16.8kD}

Table S2. InSiPS outputs for all 18 candidate target sequences, related to Table 1, Table S3.

Target name	Systematic gene name	Fitness	Target score	Max non-target score	Average non-target Score
REI1	YBR267W	0.437109	0.66573458	0.34341903	0.03132127
CHK1	YBR274W	0.570826	0.70807537	0.19383371	0.019953153
BDH1	YAL060W	0.290459	0.54265767	0.4647474	0.236912524
AIM2	YAL049C	0.276894	0.5333423	0.48083291	0.139402474
HSM3	YBR272C	0.358195	0.59402891	0.39700788	0.125126232
ACS1	YAL054C	0.45144	0.64881096	0.3042048	0.015965355
RPL19B	YBL027W	0.329424	0.54180512	0.39198845	0.076419325
PIN4	YBL051C	0.313386	0.55347628	0.43378601	0.144189497
PSK1	YAL017W	0.541232	0.71515927	0.24320134	0.041448462
RMD1	YDL001W	0.399594	0.67404023	0.4071663	0.088788771
PSK2	YOL054W	0.433372	0.65443715	0.33779394	0.025854975
HEK2	YBL032W	0.66108	0.84123403	0.21415452	0.018977927
CCR4	YAL021C	0.464692	0.67289423	0.30941335	0.059046125
CDC24	YAL041W	0.364046	0.62862367	0.4208845	0.114832926
FUN11	YAL036C	0.301512	0.58082879	0.48089416	0.079736457
YPT10	YBR264C	0.283617	0.54491094	0.47951721	0.265006925
EFM2	YBR271W	0.22569	0.4165458	0.45818566	0.291757241
PRD1	YCL057W	0.24819	0.45788308	0.45796182	0.315611118

Table S3: Sequences of 18 anti-target proteins generated in the original InSiPS run, related to Table 1.

Target Protein	Sequence anti-target protein
REI1	HHHHHHVMKPAEFAVQAIYKLAPVARRMHCYVLAWTDSQNTIISRNAINGVIHADAHDCGVSNPNFGD SARFGNGQVCIVCNEQMKLLTKHPWNVVQHRNMGCPMKTEMCERMPGYMEPCFHQFNPASMPDD SDDWRDAQCEPAGKINGL
CHK1	HHHHHHAHFIHGSPLAIEAIGAALAYGCKFMFIHAPLYLCYSDWRVTGMDFHDMKEPAWAHVSAGVPY SAPAPCYHTYSIMMVGYLAGKHIFANREPWPWYVPYSRTHAGQRVFPTDAPIWEDDQRERMGAAQQ SKWIGRRPDKMSGKAF
BDH1	HHHHHHQNQQRACQPMPQHQQQHHAADQLQARYRWQLEPWHEWAWCAACARGYYETHVVCITCF CIGMAKPHISYTERHERPSKCAAWSRDALTAKFARTAKASKCACNLYMHCPLYRNCPQAQTCPIYRCD ELGAVKTHFCAEAWETSW
AIM2	HHHHHHSATQADSTAHESTKFHNSKNADQPHTCYFPMMCHGAYFEGWPPCCKWYLQHNEHPGHYR THAHIDHCTSNETRKKKHHQHHIKKYRQMHTDGIRCSWRNKDESKTGPVTPVRLKNAHSKSCYGFCH FNNQMQPCSNYKRPSSGN
HSM3	HHHHHHQQAGMQSQQQQQEQLCKEEPQTKAGPEYESNHAFAQQQVNMCAHGKCAAFWQTKNDWI QMFDDPTTRCSPCHWAADPAEVQEEAVAADIKIWHCDKQCQCCSRMSDGIPHRIVCCFYTDDSRYAK FKSVYDGFGFTSMMIVPCS
ACS1	HHHHHHIFIWISDKVSWMNCRFAGDKQNAHTYSCMHFQEPMSEMWLKLLMYTARFANTEPRGVYTH EARPVWTNIKDYVFNAAICPAWREKGGNNAANSMENHYVGACAVESNECIKPHCVPALWFYCKFCHQ FTEPHMHPSAYILAKGT
RPL19B	HHHHHHRCPPHSRKGQLLFRHAFADREQQGRNQWPYQNRNWKARYEACPKKEMQVTVKLLWCEM NMAWHCQQAIRGRHHIISHNQLWHDEEEFGIPCAMEVNQLRIHFNWLFSLYVPFGGLMPPHLAFQNR MHCLEKEGTKFIPAMEYFA
PIN4	HHHHHHTMACYQWRQMHERQHHHRENKTVAAVEMCHCLMTWAYHKARWDCRECGQTGVPFHQW QHHQQQARQHSQKHRAALYREDMMFFHFYNHPNHLVHQVTQMTQYWAHFIPRSHPWEPATCQSNIL VCCKNTCYTMAHTVRTTDAEW
PSK1	HHHHHHEQVHRQVHDKERAQAAVGTWLFALCDLCSYGGWFCCCFCRKRVRILVFAFKHHRICRREW WRWCTKKMNIHMIPTQVVMVMIEPHIMALYMGCQTGFIECDFYMGRFRSQAHFVHMFFAQMVWQAK CTQQDWYSRINRKKNDHL
RMD1	HHHHHHHCECDQPVHAIAYQLYVLVEFCDHGAMEKYMMTRWHAALREVLGEICTDWGGDYTQAHAQ HAEAEAECMMHGIAAHPYFIIQWPTFHDQEQTYNIDDTGAGACCKNHAAEHTCKDEPEWPKNFRWG QFAGPIHCCQGYADTAQI
PSK2	HHHHHHATQCDPQYFQQDAPPNQGELCKWQIHVYFRFQAYWERPHHAEETHAAWRRWWVQNKTM KVVLPKMHRGVGKSGEHRLMMWALECQYVMTARWFAKFGGTKAGMIFIMKRNYKLRCHPCFEIWAT AKNRMPWNFPGQGGSDSVTV
HEK2	HHHHHHSVKANCSYFWAKYNYILTAVWATNEAERHTRRAEWKDFYHWGLIKAPVSVSINPECTTANE EYFPWWFLYKNRAGQYPTGPPAMAPHAKSPQTMQSKQAEIGMMHPNVFAREWHHYAPPQNEMFMS ARFLMQGKTPCQYGGRIR
CCR4	HHHHHHFYCTHLQDRWAMIGWGCLPPHNCTGFAENEDEFPLAISLKNKYRKVKFCFLLYCLCTCPIEY AFTGHEQIAARGQQRSHQTQYYKPGMHVGWRKSPTKFFLYYKQPGRLCQGMTCVMVMQEQHQENL DWQQKEKHRKASSQQYD
CDC24	HHHHHHSMARCFESYMDNVMRPDRIWHELWYAGVNDMWAAFQCCPWCTWFQCEVHYLYVITCALK VARCWEAVDHYWPFVTAAWAQQLAVPFMQEQQHLQTHDQEDDLMHQKLIKICWLKNAQHPGAFPAL AECFCYWKPVVANTLSVAL
FUN11	HHHHHHSHAPYHKQASGDRHPSIRFWANMQLFLFCESYQKAACSAVVCFPLVCTYNEAMNAWVRVP EIFLHRPWYPNILATRDPCNYYTNFELGSSRLLVNCGRRQNWVWEEEGAERCHPMMSAHLVDTHRPI YGQSGQVHHCSDDNDNG
YPT10	HHHHHHFLMARAAFNYWYMGYPPIFITIQAVIQEENGDQHFQQQEHGQPGEQTAQMQQHQQPEQQH MKTEMAACPSDGLGAACPTWFHEGTGCPRCKFPARGQIFAKCHRTWMQWDMFVLQWAYTLEFWQ CSGEHDMCLFFRNICHAVAY
EFM2	HHHHHHIPAQCKLWSTLTAKQCFNCGQTNGTRCYRCGQFCHKIRDCPEARDCYHCGNGALVFCYWK NNCHERAFRYACPWLPTADVGAPWKCPDGGSQCEHNCDAAWLTWAVQSIQTHNQPKQGPQQQEQ RNQNERQQSHVKQTTGQKEY
PRD1	HHHHHHAYDDYMYHDHFHNNHEHPHQLMCKLRDYAHYLTWCERQGQAQCLCDARVDVRQRMEYC PFFEGYREQCKFGPHCTQSACSYRPPTHQCKCNCACQGCPEAQCCTCHSYSQDGGGSTQQQEPPH HQPAQWQQQDPNTQVDNAPYE

Transparent methods

InSiPS Algorithm

The InSiPS algorithm aims to design protein sequences that are predicted to interact specifically with a given target. Formally, a problem is presented to InSiPS in the form of a target protein sequence and a set of non-target protein sequences (off-targets). For each protein design problem addressed in this manuscript, InSiPS was provided the sequences of the given target protein and the sequences of 1701 off-target yeast proteins known to localize to the cytoplasm. InSiPS then uses a genetic algorithm to design functional peptides. It starts with an initial population of 1000 randomly generated sequences 150aa in length.

Next, the fitness of each individual is calculated based on the predicted interaction scores of the given sequence and the target/non-targets (see next section in Transparent Methods for details on the fitness function). Once all of the sequences in the current population have been evaluated, the next generation (pool) is created. Three operations are used to build the next pool: copy, mutate, crossover. The copy operation simply selects a sequence from the current pool and places it in the next pool. The mutate operation selects a sequence from the current pool and randomly mutates each amino acid in said sequence with a fixed probability and places the resulting sequence in the next pool. The cross-over operation selects two sequences from the current pool, chooses a random point at which to split the two sequences exchanges the ends of the sequences and places the resulting new sequences in the pool of the next generation. Each operation needs to select at least one sequence from the current population and this is done randomly with each sequence's probability of being selected being proportional to its fitness. The probability of selecting the copy, mutation or cross-over operation as well as the rate of mutation can all be manually set. These parameters were set to 0.1, 0.4, 0.5 and 0.05, respectively, for all runs described in this manuscript.

This process of evaluation and creation was repeated for a minimum of 250 generations at which point the algorithm did not terminate until the fitness score of the best individual sequence observed did not improve for a further 50 generations. Three SBPs amenable to wet lab testing in our facility were chosen and re-run three times using random seeds and the highest scoring SBP out of all runs was chosen for follow-up investigations. An arbitrary length of 150aa was chosen for this proof of principle project as predicted sites can vary in length and because a larger polypeptide is more likely to result in functional inhibition when bound. A length of 150 includes the majority of predicted sites (up to a 114aa predicted site +/- 20aa required for prediction + 6XHIS tag).

Fitness Function

InSiPS employs a well-established protein-protein interaction prediction tool known as PIPE (Protein Interaction Prediction Tool) (Pitre et al., 2008, 2012) to predict the likelihood of sequences physically interacting with targets and non-targets. PIPE has successfully predicted proteome wide interaction maps in several species including *S. cerevisiae*, *Schizosaccharomyces pombe, Caenorhabditis elegans*, and humans at an extremely high specificity (99.95%) (Pitre et al., 2012). The method is purely sequence based and has been independently shown to outperform competing PPI prediction methods in recall precision (Park, 2009) with a nearly negligible false positive rate of 0.05%.

To assign a fitness to a given protein sequence produced by InSIPS, PIPE is used to evaluate the predicted interaction profile of each sequence generated (likelihood of interacting with the target vs non-targets). The resulting prediction scores are combined into a single score using the InSiPS fitness function. This function rewards target binding and penalizes predicted off-target interactions. The fitness of a given sequence calculated as follows:

$$\text{Fitness(seq)} = [1 - \text{MAX(PIPE(seq, } \textit{non-targets}\text{))} \times \text{PIPE (seq, } \textit{target}\text{)}]$$

For each target, InSiPS generates an ordered list of sequences which maximize this fitness function. Importantly, this list generally includes several unique polypeptides that have limited homology to known proteins.

Implementation Details

InSiPS is a massively parallel algorithm designed to run on IBM's Blue Gene/Q supercomputer. It is implemented in the C programming language and consists of a two-level master-slave / all-slaves architecture. The master process is responsible for all of the genetic algorithm functionality while the slave processes are responsible for executing the PIPE algorithm on generated protein sequences and target/non-target proteins. This aspect is implemented using MPI, a communication protocol allowing the processes to communicate with one another.

The master process starts by generating a pool of random protein sequences to act as the first generation (pool). This population then needs to be evaluated and are then sent individually to the slave processes. The results are then returned as a list of PIPE scores between the candidate sequence and all of the target and non-target proteins. The master process uses this data to assign each sequence in the current pool a fitness score and creates the next pool, as discussed above, before reevaluation happens. All of these functions are performed in parallel within the master process through the use of openMP, a shared memory multiprocessing API.

Although the master process performs most of the important functionality of InSiPS, the bulk of the computation time is spent performing PPI predictions. This is what the slave processes are responsible for. At the start of an InSiPS run, each slave process requests work from the master. It then receives a candidate protein sequence and is then responsible for applying the PIPE algorithm on this sequence and all of the target/non-target proteins. Once complete the results are sent back to the master process the slave process waits to receive another sequence to analyze. Again, all of these predictions are done in parallel within the slave process in an all-slaves manner through the use of openMP.

A single run on InSiPS will evaluate hundreds of thousands of potential sequences, leading to hundreds of millions of individual PPI predictions needing to be computed. A typical run of InSiPS would use 512 Blue Gene/Q nodes, each capable of supporting 64 computational threads. This leads computational processes consisting of 512 MPI processes and over 32,000 computational threads.

Target Protein Selection

18 endogenous yeast proteins with the following characteristics were chosen as experimental candidate targets;

i. localized to cytoplasm
ii. between 500-1500aa in size
iii. moderate abundance (3,000-10,000 copies per cell)
iv. mutants demonstrate readily observable phenotypes (drug sensitivies/resistance)

InSiPS generated an optimized anti-target protein for each of the 18 yeast cytoplasmic protein targets. Three proteins of varying fitness most amenable to laboratory testing (Rmd1, Pin4 and Psk1) were each re-run three more times using different random seeds. The best results from these runs were forwarded for experimental validation.

Vectors

Output sequences from InSiPS for anti-Rmd1, anti-Pin4, and anti-Psk1 (Table 1) were synthetized by GeneArt®, a division of Invitrogen/ThermoFisher and cloned into the pYES2 expression cassette. The open reading frame (ORF) of the synthetic proteins is under the control of a GAL1 promoter and a 6X His tag was added to the C-terminus. The cassette was also sub-cloned into pBI881 for use in yeast-two hybrid and affinity purification assays. The p416 plasmid which contains a LacZ expression cassette was used as a background plasmid control for drug sensitivity analysis. Protein purification was performed using the pET28a *E. coli* expression vector.

Strains

Drug sensitivity assays were performed using variations of *S. cerevisiae* S288C (MAT a orfΔ::kanMX4 his3Δ1 leu2Δ0 met15Δ0 ura3Δ0), described in (Winzeler et al 1999). GFP experiments were performed in an EY0986 (MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0) background. Yeast-2-hybrid experiments were performed in MAV203 (MaV203 (MATanti leu2–3, 112, trp 1–901, his3Δ200, ade2–101, gal4Δ, gal80Δ, SPAL10::URA3, GAL1::lacZ, HIS3UAS GAL1::HIS3@LYS2, can1R, cyh2R). Protein purification was performed in BL21 *E. coli*.

Growth Media

S. cerevisiae strains were maintained in complete yeast media YP (Yeast extract, Peptone, 2-4% sugar) or autotrophic media [yeast nitrogen base, an amino acid supplement lacking the appropriate amino acid (-Leu lacking leucine, -Ura lacking uracil, -Trp lacking tryptophan, or –Leu/-Trp lacking leucine and tryptophan) and 2%-4% of various sugars]. Expression of synthetic proteins was achieved using media supplemented with 2% galactose. Synthetic complete (SC) media (yeast nitrogen base, and a complete amino acid supplement + 2-4% various sugars) was used as a complete media when comparing to strains grown in autotrophic media. *E. coli* strains were grown in LB media with 30μg/mL kanamycin for selection and supplemented with 500μM IPTG for induction of protein expression.

Drug Sensitivity Colony Count Analysis

S288C strains containing pYES2 plasmids expressing either anti-Psk1 or anti-Pin4 or the p416 background plasmid were grown to saturation in 5mL of –Ura liquid medium for 48hr shaking at 30ºC. A WT strain containing a p416 plasmid was included as an addition WT control. S288C WT and strains with either *PSK1* or *PIN4* deleted were grown in SC liquid medias. Cultures were then serially diluted in sterile distilled water from 10^{-1} to 10^{-4}. 100μL of the 10^{-4} dilution was spread on YP-galactose agar plates for the control condition and plates were grown at 30ºC for 48 hours. For the experimental condition, 100μL of the 10^{-4} dilution was spread on YP-galactose plates containing with 65ng/mL cyclohexamide for the Pin4 experiment and for anti-Psk1, the plates were exposed to 30 seconds of UV radiation using the sterilization bulb in a biological safety cabinet. Colony counts are represented as the ratio of colony forming units (CFUs) counted on the experimental condition plates over the number of CFUs control plates normalized to the ratio of the WT + background plasmid strain in the same conditions. Mean and standard deviation from triplicate trials are presented.

Drug Sensitivity Growth Curve Analysis

S288C strains containing pYES2 plasmids expressing either anti-Psk1 or anti-Pin4 or the p416 background plasmid were grown to saturation in 5mL of –Ura liquid medium for 48hr shaking at 30ºC. A WT strain containing a p416 plasmid was included as an addition WT control. S288C strains with either *PSK1* or *PIN4* deleted were grown in SC liquid medias. The p416 plasmid expresses β-galactosidase under a Gal4 promoter to ensure that overexpression off of a plasmid does not contribute to any observed phenotypes. Overnight cultures were collected by centrifugation at 5000 RPM then washed in dH_2O and resuspended in YPGalactose. The cell density of this suspension was estimated by obtaining OD_{600} values from 10X diluted samples. Aliquots from these samples were then added to culture tubes containing 5mL of YPGalactose with and without drug. Equal volumes of this cell suspension were added to the appropriate culture tubes that starting densities were consistent at OD_{600}= ~0.3. Three cultures without drug and three cultures containing drug were inoculated in each trial and 3 trials were completed. Cultures were grown in appropriate media over 12 hours with OD_{600} readings of 200µL aliquots taken every 60 minutes. Drug conditions for anti -Psk1 was and 0.75 mM H_2O_2. Drug conditions for anti-Pin4 were hygromycin 13µM and 0.1mM arsenite.

GFP Analysis

EY9086-derived strains with GFP-tagged targets (Psk1 and Pin4) were obtained from the yeast-GFP collection (Thermo Fisher) and analyzed to identify any alteration to the fluorescent signal when anti-target proteins were expressed. Cultures were grown for 48hr shaking at 30ºC to saturation in 5ml of –Ura +2% raffinose. 10µL of each saturated culture was used to inoculate wells containing 190µL of either induction or repression media. Induction was achieved using -Ura + 2% raffinose + 2% galactose and repression media contained -Ura + 2% raffinose + 2% glucose. Fluorescent determinations were made every 10 minutes using a BioTek FL600 microplate reader at 30ºC shaking orbitally prior to readings and plotted over 6 hours. Average fluorescent units were normalized to cell density readings (OD_{600}). Raw data is provided as a Supplemental Data file.

Yeast-2-Hybrid Construct Preparation

The ORFs of anti-Pin4 and anti-Psk1 were amplified from the synthesized pYES2 expression plasmids using primers with unique recognition sites to facilitate cloning into the prey (GAL4-TA) plasmid pBI-881 (Kohalmi et al., 2002). The ORF of *PSK1* and *PIN4* were amplified from commercially available BG1805 yeast overexpression plasmids obtained from GE-Dharmacon (Gelperin et al., 2005) using primers containing XhoI and NotI unique restriction sequences and cloned into the bait (GAL4-DB) plasmid pB880. The anti-Pin4, anti-Psk1 inserts were cloned into pBI881 using SalI-NotI recognition

sequences. Primer sequences are found below. Transformation was completed using the LiAC/PEG method using carrier ssDNA. Transformants containing pBI880 were selected on –leu plates and plated on –ura plates to ensure plasmids did not self-initiate. Double-transformants containing both the appropriate pBI881 and BI880 plasmids were selected on –leu/-trp minimal media plates.

F-primer: Psk1
5’ AGC ACC ↓TCG AGC CCC TAC ATC GGT GCT TCC AAC CTC TCA GAA CAT

R-primer: Psk1
5’ ACT GAG CGG CCG CCA TCA TCA AAT AAC CAA CCA TTT GTC GTT ATT

F-primer: Pin4
5’ AGC ACC TCG AGC GAG ACC AGT TCT TTT GAG AAT GCT CCT CCT GCA

R-primer: Pin4
5’ ACT GAG CGG CCG CCA TTA TTA CCA TAG ATT CTT CTT GTT TTG GTT

F-primer: Anti-Pin4
5’ AGC AGC AGC AGC AGC AGC AGC AGC AGC ACC TCG AGC ATT TTC ATC TAC GGT GAT AGA TTA TTG GAT CAA

R-primer: Anti-Pin4
5’ ACT AGC GGC CGC TTA TTA ATG ATG GTG ATG GTG GTG ATG ATG GTG ATG ATG ATG GGC

F-primer: Anti-Psk1
5’ AGC AGC AGC AGC AGC AGC ACC ↓TCG AGC TCT GAT AAT GAA CAC TTG CAT AAG TGC CAA

R-primer: Anti-Psk1
5’ ACT AGC ↓GGC CGC TTA TTA ATG ATG GTG ATG GTG GTG ATG ATG GTG ATG ATG ATG CTG

Y2H Assays

Analysis of Y2H interactions was performed using three reporter assays which measure the expression of designated ORFs under the control of unique promoter sites influenced by the presence of a reconstituted Gal4 transcription factor. In MAV203, Gal4 activates transcription of URA3, HIS3 (competitively inhibited by 3-aminotriazole), and Beta-Galactosidase. The first reporter assay monitors the ability of

pBI880(bait)/pBI881(prey) double-transformants to grow in auxotrophic media lacking uracil. Growth in –ura suggests activation at the *SPO13* promoter which has 10 Gal4p binding sites driving *URA3* expression (Vidal et al., 1996). The other two reporter genes are controlled by the GAL1 promoter, which drives the expression of either *HIS3* that is competitively inhibited by 3-aminotriazole or lacZ which cleaves ONPG to produce a quantifiable signal.

Firstly, to measure URA3 expression, a growth curve was constructed using liquid minimal media lacking uracil and containing 2% galactose. 5mL cultures were grown to saturation in –leucine/-tryptophan and used to inoculate 40mL of –Ura+ 2% galactose media. Cultures were grown at 30ºC shaking at 160RPM over 8 hours. Samples were taken every 60 minutes and OD_{600} was read. Secondly, CFU counts on media containing 25mM 3-AT were performed. Again, 5mL cultures were grown to saturation in –leu/-tryp. 10^{-4} dilutions were plated on synthetic complete (SC) media and 10-3 dilutions were plated on SC+3AT. Results are expressed as the number of colonies on SC+3AT plates (x10) over the number of colonies on SC and the mean +/- standard deviation. Thirdly, Beta-galactosidase activity was measured using the liquid culture ONPG substrate assay with 30 minutes selected as an endpoint as previously described (Clonetech Laboratories Inc, 2009). Fresh 5mL cultures are incubated at 30ºC shaking until an OD_{600} of 0.5-0.8 is reached. 1.5mL aliquots (3X) are taken, cells are pelleted resuspended in Z-buffer with ONPG for 30 minutes before stopping with Na_2CO_3.

Protein Purification

Proteins were expressed from a *E. coli* with a 6X-HIS tag were affinity purified using Ni-NTA agarose resin using a Bio-Rad gravity flow column according to manufacturer protocols. 750mL of LB media containing 30μg/mL of kanamycin was inoculated with BL21 + pET28a expressing either anti-Psk1 or anti-Pin4 and grown to OD_{600} ~0.4 and then induced with 500μM IPTG and grown overnight at 16ºC. Cells were harvested via centrifugation, flash frozen, and lysed using lysozyme and sonication. Lysate was run through 1mL Ni-NTA agarose columns, washed with PBS + 40mM imidazole wash buffer and eluted with PBS + 250mM elution buffer. Polyacrylamide electrophoresis identified fractions with high purity and a Bio-Rad® Bradford assay was used to determine protein concentration.

Peptide SPOT arrays

Individual peptides were synthesized at 0.1-mmol scale on a Multipep RSi peptide synthesizer (Intavis Inc.) using standard Fmoc (*N*-(9-fluorenyl) methoxycarbonyl) chemistry. For fluorescein labeling, an appropriate amount of 5-(and-6)-carboxyfluorescein succinimidyl ester was added to a peptide resin, and the coupling reaction was allowed to proceed for 1 h at room temperature. Upon cleavage of a peptide from the resin using trifluoroacetic acid, the fluorescein-labeled peptide was separated from the unlabeled

peptide by HPLC on a C_{18} column. Identities of the peptides were confirmed by mass spectrometry. Peptides were arranged in to a SPOT array on a cellulose membrane (Jia et al., 2005).

All peptide SPOT arrays were blocked with 5% BSA in TBST (0.1 M Tris-HCl, pH 7.4, 150 mM NaCl, and 0.1% Tween 20) for 1 h. Purified protein was added directly in the blocking buffer to a final concentration of 1 μM and incubated with the SPOT array at room temperature for 1 h. The array was then washed 3 × 5 min with TBST before a rabbit anti-His antibody (1:4000 dilution in TBST; Cat# ab3553, abcam) was added. The membrane was allowed to incubate at room temperature for 30 min prior to 3 × 5-min washes with TBST. After final 3 × 5-min washes in TBST, the SPOT arrays were visualized by enhanced HRP-based chemiluminescence.

Fluorescent polarization

A varied amount of a purified 6xHis-anti-Psk1 was titrated to a fluorescent peptide solution in 20 mM PBS, pH 7.0, 100 mM NaCl. The mixtures were allowed to incubate in a dark environment for 30 min prior to fluorescent anisotropy measurements at 20 °C. Binding curves were generated by fitting the isothermal binding data to a hyperbola nonlinear regression model using Prism 3.0 (GraphPad Software, Inc., San Diego, CA), which also produced the corresponding dissociation constants (*Kd*).

Supplemental References:

Clonetech Laboratories Inc (2009). Clonetech - Yeast Protocols Handbook.

Gelperin, D., White, M., and Wilkinson, M. (2005). Biochemical and genetic analysis of the yeast proteome with a movable ORF collection. Genes &.

Jia, C.Y.H., Nie, J., Wu, C., Li, C., and Li, S.S.-C. (2005). Novel Src Homology 3 Domain-binding Motifs Identified from Proteomic Screen of a Pro-rich Region. Mol. Cell. Proteomics *4*, 1155–1166.

Kohalmi, S., Nowak, J., and Crosby, W. (2002). The yeast two-hybrid system (London: Taylor and Francis Ltd.).

Park, Y. (2009). Critical assessment of sequence-based protein-protein interaction prediction methods that do not require homologous protein sequences. BMC Bioinformatics.

Pitre, S., North, C., Alamgir, M., Jessulat, M., Chan, A., Luo, X., Green, J.R., Dumontier, M., Dehne, F., and Golshani, A. (2008). Global investigation of protein-protein interactions in yeast Saccharomyces cerevisiae using re-occurring short polypeptide sequences. Nucleic Acids Res. *36*, 4286–4294.

Pitre, S., Hooshyar, M., Schoenrock, A., Samanfar, B., Jessulat, M., Green, J.R., Dehne, F., and Golshani, A. (2012). Short Co-occurring Polypeptide Regions Can Predict Global Protein Interaction Maps. Sci. Rep. *2*, 1–10.

Schoenrock, A., Samanfar, B., Pitre, S., Hooshyar, M., Jin, K., Phillips, C. a, Wang, H., Phanse, S., Omidi, K., Gui, Y., et al. (2014). Efficient prediction of human protein-protein interactions at a global scale. BMC Bioinformatics *15*, 1–22.

Vidal, M., Braun, P., Chen, E., Boeke, J.D., and Harlow, E. (1996). Genetic characterization of a mammalian protein-protein interaction domain by using a yeast reverse two-hybrid system. Proc. Natl. Acad. Sci. U. S. A. *93*, 10321–10326.