Using HHsearch to tackle proteins of unknown function – a pilot study with PH domains

David R. Fidler[1*], Sarah E. Murphy[1*], Katherine Courtis[1], Pantelis Antonoudiou[1], Rana El-Tohamy[1], Jonathan Ient[1] and Timothy P. Levine[1¶]

[1] Department of Cell Biology, UCL Institute of Ophthalmology, 11-43 Bath Street, London EC1V 9EL, UK
* these authors contributed equally to this work
¶ to whom correspondence should be addressed

Contact emails
David Fidler: david.r.fidler@googlemail.com, or david.fidler.12@ucl.ac.uk
Sarah Murphy: sarah.murphy.14@ucl.ac.uk
Katherine Courtis: katherine.courtis.10@ucl.ac.uk
Pantelis Antonoudiou: pantelis.antonoudiou@new.ox.ac.uk
Rana El Tohamy: tohamy_rana@hotmail.com
Jonathan Ient: joncient@googlemail.com
Timothy P. Levine: tim.levine@ucl.ac.uk

Running title: HHsearch finds new PH domains

KEYWORDS: GRAM domains, Gyp7p, pleckstrin homology (PH) domains, profile-profile search, Saccharomyces cerevisiae, secondary structure prediction, structural bioinformatics, TBC1D15, Vps13p, YJL181C, YJR030C, YJL016W.

ABSTRACT

Advances in membrane cell biology are hampered by the relatively high proportion of proteins with no known function. Such proteins are largely or entirely devoid of structurally significant domain annotations. Structural bioinformaticians have developed profile-profile tools such as HHsearch (online version called HHpred), which can detect remote homologies that are missed by tools used to annotate databases. Here we have applied HHsearch to study a single structural fold in a single model organism as proof of principle. In the entire clan of protein domains sharing the pleckstrin homology domain fold in yeast, systematic application of HHsearch accurately identified known PH-like domains. It also predicted 16 new domains in 13 yeast proteins many of which are implicated in intracellular traffic. One of these was Vps13p, where we confirmed the functional importance of the predicted PH-like domain. Even though such predictions require considerable work to be corroborated, they are useful first steps. HHsearch should be applied more widely, particularly across entire proteomes of model organisms, to significantly improve database annotations.

INTRODUCTION

An important way to determine the function of an unknown protein is to identify homology to a protein with known function. However, a significant minority (~20%) of protein sequences shows no obvious homology at the level of primary structure (sequence alone) to any other protein[1]. Tools such as Basic Local Alignment Search Tool (BLAST) begin to perform poorly when sequence identities within a single domain drop below 30%[2]. To overcome this problem, sequence-sequence searches have been supplemented with profile-sequence searches such as PSI-BLAST[3], which increase sensitivity by searching with a profile that contains statistical information from a multiple sequence alignment (MSA) created from the initial query. Even more sensitive tools carry out profile-profile searches, which extract family-wide information for both query and target[4-6]. Such tools involve the construction of profiles for all members of target sequence libraries, which requires considerable computational effort. Therefore, profile-profile searches tend to be restricted to a small range of targets, most often the Protein Databank of solved structures (PDB). Profiles can implicitly encode aspects of secondary structure[7], but some tools go further by explicitly including secondary structure[8-10], either relying on solved structures or predicting secondary structural elements[11,12]. There has been a steep decline in the discovery of completely new folds[6], which suggests that many regions of unknown function will turn out to be distant structural homologues of known domains. It may be possible to detect these distant structural homologies by applying the most sensitive tools to all protein regions of unknown function. HHsearch is a powerful profile-profile alignment tool[13] that has not been applied widely or systematically.

Pleckstrin homology (PH) domains and their structural homologues, PH-like domains, are one of the most common folds in eukaryotes[14]. Although PH-like domains have no unified function, they have typical docking sites for other proteins or inositide lipids[15]. They are often involved in targeting peripheral membrane proteins involved in traffic and signaling[16]. Hence, prediction of new PH-like domains may identify residues that are functionally important for intracellular traffic. Since the initial discovery of classical PH domains, newly solved PH-like structures have unexpectedly been found, both early on[17], and at least ten times since[18-27]. Each of these discoveries added a new family of PH-like domains to a growing PH-like clan[28]. PH-like domains are therefore a good example of a structural fold with high sequence divergence.

The model organism *S. cerevisiae* is one of the most highly studied, especially in systems biology. Yet for 10-13% of yeast proteins there is no functionally or structurally relevant domain information (TL, unpublished observation). Progress in finding functions for yeast proteins with no recognizable domain is slower than expected[1]. As an example that can be worked through in depth, we used HHsearch to study the PH-like domain clan (sometimes called superfamily) in yeast. After bench-marking HHsearch against PH-like domains with solved structures, we searched for PH-like domains in yeast. HHsearch was more accurate than tools used to annotate databases, and it detected some new domains easily. We found even greater sensitivity by seeding searches with yeast protein sequences, although this required high volume automated offline searches. Overall, we discovered 16 new yeast PH-like domains. One of these is at the C-terminus of Vps13p, a conserved membrane contact site protein implicated in sorting along the endocytic pathway. The presence of this one PH-like domain was verified by demonstrating both that it has a specific targeting activity and that it is required for one aspect of the function of full length Vps13p. This study of PH-like domains in yeast shows how the problem of proteins without domain information could be significantly reduced by systematic application of HHsearch.

RESULTS and DISCUSION

Curating PH-like domains – what we knew already

PH-like domains are defined by a combination of structure and sequence alignment (Box 1). Different PH-like domain families are not linked by sequence alignment. For example, classical PH domains and GRAM domains have very similar alpha-carbon backbones (Figure 1 and Movies 1-3), but no PSI-BLAST search seeded with a classical PH domain makes a significant alignment (*i.e.* there is no hit) with GRAM domains and *vice versa*. Each PH-like family is presumed to have diverged from a single ancestral protein. In addition there is some evidence that the whole clan diverged from a prokaryotic origin[26]. PH-like families (Supplemental Table 1) have been defined by a variety of profile-sequence tools (Box 2), but given that there is a lag between discoveries and upgrading databases, we repeated these definitions.

Defining PH-like domains starts with mining the literature, the PDB for solved structures, and sequence databases. For a standard definition of the PH-like clan we used PSI-BLAST to study PDB, which contains >200 different PH-like domains.

We described the minimal set of 39 PH-like sequences that identified the complete set of PH-like domains in PDB as strong hits in PSI-BLAST searches. This minimal set indicates the presence of 39 families in the PH-like clan (Figure 2A and Supplemental Table 1). This classification is more complete than definitions of the PH-like clan in databases (Box 2). The largest family consists of the classical PH domains (~50% of all sequences). Three families have approximately 10% of the domains each: FERM-C, PTB and RanBD. Non-significant hits indicated that some families share weak homology (Figure 2A). However, the main feature was the presence of 20 small families (≤3 members) that were completely isolated from all others (Figure 2A). The lack of cross-detection between most families indicates that if other PH-like domains existed, they might easily be undetectable by PSI-BLAST.

HHsearch - more sensitive than PSI-BLAST

Profile-profile tools such as HHsearch (Box 3) have previously detected many new domains[29-39], so we asked if HHsearch can improve on PSI-BLAST for PH-like domains. We first studied how HHsearch detects PH-like domains in PDB, where we are certain if hits are true or false. This allowed us to determine key parameters relating to both specificity and sensitivity. Importantly, for this exercise HHsearch was set to be unaware of solved structure. HHsearch was seeded with sequences from our minimal set of 39 PH-like families.

Specificity (avoiding false positives): HHsearch provides a list of alignments with decreasing probability of shared structure (prob[SS], Box 3). Previous work with HHsearch has found that a threshold of prob[SS] ≥80% is sufficient to detect many true positives[29-39]. In our 39 searches,1300 hits scored prob[SS] ≥80%, of which 1298 were PH-like domains. Among all non-PH-like hits scoring prob[SS] ≥5%, the observed error rate was far less than indicated by prob[SS] (Figure 3A). This indicates that the prob[SS] metric is a conservative estimate of accuracy. Alignments to non-PH-like domains differed from true positives by being shorter (Figure 3B). In addition, non-PH-like domains had weaker secondary structural alignment (data not shown). These alignments were either to a single PH-like structural element, or to longer portions of the PH-like structure (Supplemental Figure 1). We conclude from this benchmarking that a threshold prob[SS] ≥85% will ensure close to 100% specificity.

Sensitivity (avoiding false negatives): To compare the sensitivities of HHsearch and PSI-BLAST, we compared their ability to join the PH-like clan together. The 39 searches with HHsearch produced 1200 hits to PH-like domains in PDB with prob[SS]≥85%, which is many more positive hits than PSI-BLAST produced (n=350). While 20 families were unlinked by PSI-BLAST, HHsearch left just 4 unlinked families (Figure 2B). For example, the previously isolated myosin-1c TH1 domain (4r8g_A) produced a hit to classical PH domains with prob[SS]=96%. HHsearch even linked GRAM domains to classical PH domains (Figure 2B). The detection of homologies by HHsearch that PSI-BLAST could not find indicates that it might also to detect new PH-like domains in proteins of unknown structure.

PDB-to-yeast: easy detection of some new domains

We next turned from searching PDB, where structure has been solved, to searching the yeast proteome where we have no structural information for 10-13% of 5,900 proteins. Thorough curation of current literature identified 73 PH-like domains in yeast. We noted that current database annotations lag behind this, containing from 48 to 62 (66-85%) of these domains (Supplemental Table 2A).

We used HHsearch to search a target HMM library containing all yeast proteins with the 39 PH-like sequences defined above. Among hits with prob[SS] ≥85%, there were 71 of the 73 known yeast PH-like domains (97%, Figure 2C). In addition, there were five hits with prob[SS]≥85% to regions not previously identified as PH-like. To investigate these predictions we turned to sequence homology. This might appear paradoxical, since by definition these regions have not been detected by profile-sequence tools. However, these profiles have typically been seeded on mammalian sequences, which biases the profile (Box 4). Thus, more information may be obtained about yeast proteins if they are used to seed profiles. We seeded PSI-BLAST with all five regions. Four sequences from Caf120p, Lam1p, Sip3p and Vid27p produced top hits that were PH-like domains, and other tools also identified these regions as PH-like (Supplemental Table 2B). Therefore, our PDB-to-yeast searches identified four new PH-like domains.

The 5th region identified was at the N-terminus of the ARF-GEF Age1p. When this was seeded into HHsearch or other tools there was no match to PH domains or any other strong hit (Supplemental Table 2B). Therefore, it is a false positive hit,. We next investigated how such a strong false positive arose (asterisk in Figure 3B). We found an explanation in the way the MSA for Age1p was created. The MSA was made for full length Age1, which aligns with >100 other ARF-GEFs. The N-terminus of the MSA is spuriously dominated by PH-like domains because of three factors: (i) the N-terminus of Age1p is essentially unique; possibly it contains an defunct relic of a PH domain. (ii) most of the other ARF-GEFs in the MSA have a classical PH domain upstream of their GEF domain. (iii) the Age1 N-terminal region shares non-significant sequence features with parts of two of these PH domains. This led to the inclusion of two genuine PH sequences in the MSA, even though they are divergent from Age1p (Supplemental Figure 2). The N-terminus of the Age1p MSA therefore aligns with classical PH domains. This kind of false positive can be avoided by repeating all positive alignments obtained by PDB ⇔ whole protein with just the identified domain (*i.e.* PDB ⇔ domain). To do this, we made MSAs for all five newly identified regions with minimal flanks (3-5 aa), and carried out pairwise alignments with the relevant PH-like domains by HHalign (see Methods). This confirmed that only the four new regions in Caf120p, Lam1p, Sip3p and Vid27p are PH-like domains (Figure 4A).

Yeast-to-PDB: a step change in sensitivity

A key finding in this work is that initiating a tool such as PSI-BLAST with yeast sequences strongly matched them to defined PH-like domains, even though the same hits were not seen in the other direction with PSI-BLAST or any other profile sequence tool. This indicates that profiles centred on yeast sequences may contain critical information that is missed by profiles seeded on non-yeast sequences. Since HHsearch makes profiles of both query and target, its results are potentially symmetrical (Box 4). However, standard usage of HHsearch is not entirely symmetrical. Profiles for queries use more iterations (up to eight, see Box 3). By comparison, target libraries were constructed with two iterations, and the model organism libraries were made several years ago, when there were fewer solved sequences and fewer sequenced divergent genomes. Since PSI-BLAST benefits so much from being initiated on yeast sequences, we asked if HHsearch would also benefit from being seeded with yeast sequences.

We remade the entire yeast MSA library using 8 iterations (see Methods). We then used each of the yeast MSAs in turn to search the PDB library. This identified 88 yeast regions where the top hit was **both (1)** a solved >200 PH-like domain **and (2)** had a prob[SS]≥85%. This group contained 72 of the 73 known yeast PH-like domains (99%). The 73rd known domain also matched PH-like domains in PDB, but at prob[SS] =79%. The 16 other regions included all four of the new regions found above and 12 further candidate PH-like domains, with matches in the range of prob[SS] 88–100%. We tested if these were true positives by carrying out pairwise alignments. These showed strong hits to known PH-like domains for 11 of the regions. One other region, in Pib2p, was a false positive for the same reasons as Age1p, in that matches to PH-like domains were only obtained with longer segments (*e.g.* the whole protein) not the domain (data not shown). We predict that the remaining 11 matches are all PH-like domains (Figure 4A). Three of the proteins have no previous structural information: Yjl016wp, is a twin PH protein, the 3rd in a family also containing Caf120p and Skg3p, so we named it Tph3p, for twin PH domain-3. Yjl181cp and Yjr030cp are paralogs that are distant Ran binding domain homologs, so we named them Rbh1/2p.

Overall, seeding HHsearch with yeast proteins led to a clear increase in sensitivity for HHsearch. In practice, rather than constructing a proteome-wide library for unknown-to-known searches, as we did here, users could produce a list of unknowns enriched for true positives by using the hits with prob[SH]<85% in PDB-to-yeast searches (Figure 3A). Among the 11 domains we found by reverse yeast-to-PDB searches, five had moderately high prob[SS] (>60%) in the more easily achieved PDB-to-yeast searches. These would require little work to find by yeast-to-PDB searches among a very small number of false positives. However, four of the 11 domains had prob[SS]<5% in PDB-to-yeast searches, meaning that they were not in the enriched list. This indicates that maximum sensitivity might require the construction of entire libraries, as we have done.

Indirect use of HHsearch for maximum sensitivity

One way to enhance the sensitivity of HHsearch without having to construct a library is to define intermediate sequences that are strong hits to both a PDB structure and to a yeast region of unknown function. Linking an unknown to the gold-standard database indirectly is called transitive searching[40,41], and theoretically it involves a massive increase in computation. For PH-like domains in budding yeast, we carried out a simplified form of indirect searching by finding all PH-like homologues for the 88 yeast PH-like domains in two other fungi: *U. maydis* and *S. pombe*. Proteome-wide libraries of these species are available within HHsearch. Using these fungal domains, we identified one extra PH-like domain (Vid27p-1, Figure 4A). Importantly, these indirect searches easily identify all 11 of the domains that we found by yeast-to-PDB searches (data not shown).

Use of HHsearch increased the number of PH-like domains in yeast by >20% (73 → 89). Whether this involved offline use of HHsearch, or a combination of strategies for the online server HHpred (Workflows described in Box 5), ultimate sensitivity required indirect searches. The newly described sequences add five further families to the PH-like clan. Three of the new families are widely distributed throughout evolution, and all of these are involved in intracellular traffic: Gyp7, late endosomal RabGAP in all eukaryotes including 5 human homologs: TBC1D15/16/17 RUTBC1/2; Vid27, vacuolar import and degradation in fungi and plants; Vps13, a membrane contact site protein affecting lipid metabolism in all eukaryotes. Thus, predictions in yeast can have broader significance.

The final option we explored was to test different settings within HHsearch. To identify conserved folds with highly diverged sequence, we raised the secondary structure weighting (ssw) within each alignment. In HHsearch, this has a default of 11% (Box 3). Increasing ssw to 30% produced no major improvement in overall accuracy in identifying new PH-like domains in yeast (data not shown). However, it led to one extra region in Rec114p being identified weakly, although it was below the false positive threshold. In yeast-to-PDB searches, the top hit for this region in PDB was a PH-like domain, so this is a possible new PH-like domain (Figure 4B).

Support for the predicted PH-like domain in Vps13p

One of the most typical functions of PH-like domains is organelle targeting, binding proteins or lipids. In a previous survey of 33 yeast classical PH domains, only nine had this function[42], but it is also applicable to other PH-like families [43]. To determine if the newly described PH-like domains target membranes, we expressed GFP-PH fusions for domains in eight domains implicated in membrane traffic/function: Bud2p (both domains), Gyp7p, Lam1p, Pkh2p, Tph3p-1, Tph3p-2, and Vps13p (Figure 4A). We also expressed the false positive in Age1p. None of these showed any specific cellular localisation (data not shown). We next constructed dimers (in the form GFP-PH-PH) for six of these PH-like domains (all except Tph3-1 and Tph3p-2) and Age1, since dimers can reveal weak membrane targeting[44,45]. All constructs remained diffusely cytosolic except for the Vps13 PH-dimer, which faintly localised to rings at the bud neck of some cells (Figure 5A). This suggests that the region we identified in Vps13p is involved in intracellular targeting, and may be functionally important.

Full-length Vps13p (3144 aa) has intracellular localisations to multiple membrane contact sites, including vacuole and mitochondrial patches (vCLAMP), endosome-mitochondrial contacts and nucleus vacuole junctions (NVJ) [46-48], but as yet no specific targeting domains have been tested. We determined the role of the predicted PH-like domain by constructing a mutant version predicted to inactivate it. A desirable strategy would be to mutate putative ligand binding sites in the predicted variable loops sited at β1-β2, β3-β4 and β6-β7[42]. However, Vps13p has no conserved residues in these loops (Figure 5B). Therefore, we constructed two other mutants: (a) Vps13ΔPH, lacking the entire PH domain (deletion of 3029-3144); (b) a double point mutation of conserved hydrophobic residues in the predicted alpha helix ("LIAA" = L3125A I3129A, Figure 5B). We predicted that the hydrophobic residues would stabilise the PH domain core, similar to the WxxxØ motif (Ø = hydrophobic residue) in classical PH domains, so the LIAA mutant may partially unfold[49]. Cells expressed the PH(LIAA) dimer construct at a lower level than wild-type dimer (Figure 5C). This is consistent with reduced protein stability caused by partial unfolding.

We next studied the effect of the two Vps13 mutations on intracellular distribution. Vps13-EGFP showed a complex intracellular distribution: in log phase it was largely diffuse, with the majority of cells containing puncta, some of which were close to the vacuole (Figure 6A); by comparison in early stationary phase Vps13-EGFP targeted the NVJ (Figure 6B). The two Vps13p mutants showed a marginal reduction in punctate targeting (Figure 6C/E), but considerable loss of NVJ targeting, which was partial for Vps13LIAA-EGFP (Figure 6D) and complete for Vps13ΔPH-EGFP (Figure 6F). Thus, the extreme C-terminus of Vps13p appears to play a role in targeting, particularly to the NVJ in early stationary phase.

We next looked for a functional role of the proposed PH-like domain in Vps13p. We tested if plasmid-borne Vps13-EGFP, Vps13LIAA-EGFP and Vps13ΔPH-EGFP rescue sorting of carboxypeptidase-Y (CPY), a vacuolar enzyme that is subject to a strong vacuolar protein sorting (*vps*) defect (and hence is secreted) in Δ*vps13* cells[50]. Wild-type Vps13 tagged at the C-terminus with EGFP fully corrected the defect as shown previously[48]. In contrast, Vps13ΔPH-EGFP and Vps13LIAA in Δ*vps13* produced significantly stronger signals than wild-type Vps13-EGFP, stronger for Vps13ΔPH-EGFP than Vps13LIAA, although both were significantly weaker than empty plasmid (Figure 6G). Thus, both constructs in which the predicted PH-like domain was mutated elicited partial rescue. This shows that the C-terminus of Vps13p is functionally important, and this function requires the predicted PH-like domain to be intact. Related PH-like domains are widely predicted at the extreme C-terminus of VPS13, for example in all four human homologs and in some plant homologs (data not shown), but this not the most highly conserved part of the C-terminus[48]. Without the prediction by HHsearch, the PH-like domain might be overlooked for an adjacent glycine-rich domain that is far more conserved (Figure 6H)[51].

Beyond PH domains to the whole proteome

Is the increase in domain discovery seen for PH-like domains likely to apply to other domains? To begin to examine this we surveyed 132 contiguous open reading frames from on the right arm of yeast chromosome IV, representing >2% of the yeast proteome. MSAs for all ORFs were used to search PDB, and regions not previously known to contain domains were examined. 21 new domains were predicted in 16 (12%) of the proteins (Supplemental Figure 3). As with PH-like domains, many of these new domains would also be discoverable by systematically initiating PSI-BLAST with the yeast sequence (prob[SS]≥99% n=9). Although some of the newly discovered domains do not strongly determine function (*e.g.* the ARM repeat in Rrp1p), other predictions are more functionally significant (*e.g.* intasome in Fob1p, vacuolar polyphosphate polymerase in Ydr089wp, a DNA-binding domain in Ydr124wp). Seven newly predicted domains were found in the 19 proteins with no prior structural/functional information, a discovery rate of 37%. Rolled out across the entire yeast proteome HHsearch could identify ~900 new domains, and 250-300 of these would be in the proteins that currently have no structural/functional information.

CONCLUSION

Current protein domain annotations can be significantly improved with profile-profile tools, to catch up with advances in the literature, to focus searches on proteins of unknown function and to maximize sensitivity. For the PH-like clan, HHsearch identified the known domains and also extended their number by 20%. Elsewhere in yeast, application of HHsearch may be particularly good for proteins of unknown function. Yeast proteins are very good subjects for this tool because many organisms exist that diverged from it at different times in the last 10^9 years, so homologues can be found at different evolutionary distances to provide information about functionally critical residues in yeast proteins. In the future we plan to analyze the yeast proteome systematically, annotating all the knowns and discovering many unknowns. Meanwhile HHpred is an extremely powerful online resource for discovering remote homologies.

MATERIALS and METHODS

Database files

Files downloaded (source, date generated): InterProScan (Saccharomyces Genome Database, December 2015), HHsuite 2.0.15 (Tuebingen Toolkit, June 2012) including yeast template HMMs (May 2008), PDB ("nr70", *i.e.* reduced in number so that no sequence shares more than 70% identity with any other, Tuebingen Toolkit), and Interpro 34.0.

PSI-BLAST

Searches were carried out in two stages at Tuebingen Toolkit. Firstly the PDB sequence was seeded in searches of sequences from NCBI (nr70, $p \le 0.005$, iterations = 20). If new sequences were still added in the final of 10 iterations, a further round of 10 was initiated, until no more sequences were added (max 40 iterations, ≤4000 sequences). Aligned sequences were then submitted to a single round of BLAST using PDB as the target database.

HHblits: This was run on its own or as part of HHsearch with up to 8 iterations and an e-value threshold for inclusion in MSA = 0.001–0.01.

HHsearch: Use was both online ("HHpred") at http://toolkit.tuebingen.mpg.de[52] and locally in R. Settings were default except: 8 iterations of HHblits, "maximum accuracy" alignment was turned on, e-value = 0.01, ≥1000 matches were returned, lower limit on prob[SS]=0%.

To allow secondary structure to be weighted higher than the standard 11%, we edited line 853 of "hhhitlist.C":

```
  hit.score_aass=(hit.logPval<-10.0? hit.logPval: log(-log(1-hit.Pval)))/0.45–fmin(lamda*hit.score_ss,
  fmax(0.0,0.2*(hit.score-8.0)))/0.45–3.0
```

The underlined term in was reduced to:

```
  hit.score_aass=(hit.logPval<-10.0? hit.logPval: log(-log(1-hit.Pval)))/0.45–lamda*hit.score_ss/0.45–3.0
```

This change removed a cap on the contribution of secondary structure.

HHalign: Domain sequences from both PDB and yeast proteins with 3-5 flanking residues where available were submitted to HHblits, and the resulting "representative alignments", deleting the secondary structure prediction and confidence entries, were compared in HHalign run with standard settings.

Yeast HMM library (Ychunk200): Total yeast protein sequence was obtained from downloads.yeast-genome.org/sequence/S288C_reference/orf_protein/. The problem highlighted by Age1p (Supplemental Figure 2) was reduced by dividing proteins into regions of 200 residues (overlapping by 100), producing ~26,000 "chunks". MSAs and HMMs were created for all chunks (8 iterations, ssw=11%) to create a "Ychunk200" library that is available as an FTP download on request.

Yeast-to-PDB searches: All HMMs in Ychunk200 were submitted as queries for HHsearch using the PDB library (June 2013) as target. Top hits to PDB entries that contain PH-like domains were then filtered for the presence of multiple domains, and if these were found alignments were curated by hand to include only those hits where the alignment includes the PH-like domain.

Residue conservation across the C-terminus of Vps13p: ConSurf was used with standard settings for the final 995 aa of Vps13p[53]. Results were normalized so that the 50th centile is zero and the interquartile distance (75th – 25th) = 1. Values were smoothed by taking a rolling average of x ±9 (n=19). To check that an internal domain did not force mis-alignment at the extreme C-terminus, the final 116 aa were submitted separately with no significant difference.

Plasmids: All plasmids were based on pRS416 (CEN, *URA3*), and used the 168 bp fragment from the promoter of *PHO5* for moderately strong, constitutive expression. PH domains were cloned as monomers in the form GFP — LGSAPVMAS — cloned sequence — SR*, and dimers were separated by either SR or SS. Residues used were Bud2p-1=1-163, Bud2p-2=168-319, Gyp7p=1-182, Pkh1p=562-674, Pkh2p=833-966, Tph3p-1=1- 177, Tph3p-2=180-561, Vps13p=3028-3144, and Ysp1p-2=568-717. Full length and truncated Vps13p constructs were cloned by gap repair to clone either Vps13-EGFP (1-3144 aa), Vps13ΔPH-EGFP (1-3028 aa), or Vps13LIAA-EGFP (1-3144, L3125A and I3129A). All PCR products and non-gap repaired segments were checked by sequencing.

Microscopy: Cells grown at 30°C to mid log phase, or 16 hours thereafter for early stationary phase, were immobilized between slide and coverslip, and visualized on a Leica AOBS SP2 confocal microscope at room temperature, using identical settings for images comparing different constructs.

Vacuolar Protein Sorting of CPY: Haploid BY4741 yeast lacking *VPS13* were transformed with full length and mutated Vps13-EGFP plasmids, as well as a plasmid expressing GFP-GFP. The parent wild-type strain, also carrying GFP-GFP, was included as positive control. Ten-fold dilutions of cells were spotted on selective medium (10^4, 10^3, 10^2 per spot) and grown for 24 hours before being overlaid with a nitrocellulose membrane for 16 hours. This was washed extensively, and then processed to detect CPY[50], using monoclonal anti-CPY (Invitrogen).

Authors' contributions DRF adapted the program hhhitlist.C, constructed Ychunk200, and carried out all offline HHsearches. SEM assayed CPY sorting, and carried out cloning and imaging. KC created and analyzed match lists, and created display items. PA, RE-T and JI created and analyzed match lists. TPL conceived the study, created match lists, and created display items. All authors contributed to drafting of the manuscript, and read and approved the final manuscript.

Acknowledgments We would like to thank Johannes Söding for assistance in using HHsearch, Kate Bowers for very kindly providing anti-CPY, and Matt Hayes for critical reading of the manuscript. SEM has a BBSRC iCASE studentship (BB/L016028). PA and RE-T were funded by summer studentships from the Biochemical Society and the Mycological Society respectively. The funding bodies had no role in the design, execution, analysis or writing up of this work. The authors declare that they have no conflicts of interest.

REFERENCES

1. Pena-Castillo L, Hughes TR. Why are there still over 1000 uncharacterized yeast genes? Genetics 2007;176:7-14.
2. Brenner SE, Chothia C, Hubbard TJ. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci U S A 1998;95:6073-6078.
3. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389-3402.
4. Panchenko AR. Finding weak similarities between proteins by sequence profile comparison. Nucleic Acids Res 2003;31:683-689.
5. Yona G, Levitt M. Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. J Mol Biol 2002;315:1257-1275.
6. Jaroszewski L, Li Z, Krishna SS, Bakolitsa C, Wooley J, Deacon AM, Wilson IA, Godzik A. Exploration of uncharted regions of the protein universe. PLoS Biol 2009;7:e1000205.
7. Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 2001;310:243-257.
8. Tang CL, Xie L, Koh IY, Posy S, Alexov E, Honig B. On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles. J Mol Biol 2003;334:1043-1062.
9. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics 2005;21:951-960.
10. Krishnadev O, Srinivasan N. AlignHUSH: alignment of HMMs using structure and hydrophobicity information. BMC Bioinformatics 2011;12:275.
11. Blundell TL, Sibanda BL, Sternberg MJ, Thornton JM. Knowledge-based prediction of protein structures and the design of novel molecules. Nature 1987;326:347-352.
12. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195-202.
13. Mariani V, Kiefer F, Schmidt T, Haas J, Schwede T. Assessment of template based protein structure predictions in CASP9. Proteins 2011;79 Suppl 10:37-58.
14. Mayor LR, Fleming KP, Muller A, Balding DJ, Sternberg MJ. Clustering of protein domains in the human genome. J Mol Biol 2004;340:991-1004.
15. Orengo CA, Swindells MB, Michie AD, Zvelebil MJ, Driscoll PC, Waterfield MD, Thornton JM. Structural similarity between the pleckstrin homology domain and verotoxin: the problem of measuring and evaluating structural similarity. Protein science : a publication of the Protein Society 1995;4:1977-1983.
16. Lemmon MA. Pleckstrin homology (PH) domains and phosphoinositides. Biochem Soc Symp 2007;10.1042/BSS0740081:81-93.
17. Zhou MM, Ravichandran KS, Olejniczak EF, Petros AM, Meadows RP, Sattler M, Harlan JE, Wade WS, Burakoff SJ, Fesik SW. Structure and ligand recognition of the phosphotyrosine binding domain of Shc. Nature 1995;378:584-592.
18. Begley MJ, Taylor GS, Kim SA, Veine DM, Dixon JE, Stuckey JA. Crystal structure of a phosphoinositide phosphatase, MTMR2: insights into myotubular myopathy and Charcot-Marie-Tooth syndrome. Mol Cell 2003;12:1391-1402.
19. Gervais V, Lamour V, Jawhari A, Frindel F, Wasielewski E, Dubaele S, Egly JM, Thierry JC, Kieffer B, Poterszman A. TFIIH contains a PH domain involved in DNA nucleotide excision repair. Nat Struct Mol Biol 2004;11:616-622.
20. Jin R, Junutula JR, Matern HT, Ervin KE, Scheller RH, Brunger AT. Exo84 and Sec5 are competitive regulatory Sec6/8 effectors to the RalA GTPase. EMBO J 2005;24:2064-2074.
21. D'Angelo I, Welti S, Bonneau F, Scheffzek K. A novel bipartite phospholipid-binding module in the neurofibromatosis type 1 protein. EMBO Rep 2006;7:174-179.
22. Teo H, Gill DJ, Sun J, Perisic O, Veprintsev DB, Vallis Y, Emr SD, Williams RL. ESCRT-I core and ESCRT-II GLUE domain structures reveal role for GLUE in linking to ESCRT-I and membranes. Cell 2006;125:99-111.
23. VanDemark AP, Blanksma M, Ferris E, Heroux A, Hill CP, Formosa T. The structure of the yFACT Pob3-M domain, its interaction with the DNA replication factor RPA, and a potential role in nucleosome deposition. Mol Cell 2006;22:363-374.
24. Troffer-Charlier N, Cura V, Hassenboehler P, Moras D, Cavarelli J. Functional insights from structures of coactivator-associated arginine methyltransferase 1 domains. EMBO J 2007;26:4391-4401.

25. Baek K, Knodler A, Lee SH, Zhang X, Orlando K, Zhang J, Foskett TJ, Guo W, Dominguez R. Structure-function study of the N-terminal domain of exocyst subunit Sec3. J Biol Chem 2010;285:10424-10433.

26. Xu Q, Bateman A, Finn RD, Abdubek P, Astakhova T, Axelrod HL, Bakolitsa C, Carlton D, Chen C, Chiu HJ, Chiu M, Clayton T, Das D, Deller MC, Duan L*, et al.* Bacterial pleckstrin homology domains: a prokaryotic origin for the PH domain. J Mol Biol 2010;396:31-46.

27. Kemble DJ, Whitby FG, Robinson H, McCullough LL, Formosa T, Hill CP. Structure of the Spt16 middle domain reveals functional features of the histone chaperone FACT. J Biol Chem 2013;288:10188-10194.

28. Blomberg N, Baraldi E, Nilges M, Saraste M. The PH superfold: a structural scaffold for multiple functions. Trends Biochem Sci 1999;24:441-445.

29. Knutson BA, Broyles SS. Expansion of poxvirus RNA polymerase subunits sharing homology with corresponding subunits of RNA polymerase II. Virus Genes 2008;36:307-311.

30. Pei J, Grishin NV. Prediction of a caspase-like fold in Tannerella forsythia virulence factor PrtH. Cell Cycle 2009;8:1453-1455.

31. Bateman A, Finn RD, Sims PJ, Wiedmer T, Biegert A, Soding J. Phospholipid scramblases and Tubby-like proteins belong to a new superfamily of membrane tethered transcription factors. Bioinformatics 2009;25:159-162.

32. Kopec KO, Alva V, Lupas AN. Homology of SMP domains to the TULIP superfamily of lipid-binding proteins provides a structural basis for lipid exchange between ER and mitochondria. Bioinformatics 2010;26:1927-1931.

33. Knutson BA. Insights into the domain and repeat architecture of target of rapamycin. J Struct Biol 2010;170:354-363.

34. Hayes MJ, Bryon K, Satkurunathan J, Levine TP. Yeast homologues of three BLOC-1 subunits highlight KxDL proteins as conserved interactors of BLOC-1. Traffic 2011;12:260-268.

35. Knutson BA, Hahn S. Yeast Rrn7 and human TAF1B are TFIIB-related RNA polymerase I general transcription factors. Science 2011;333:1637-1640.

36. Pei J, Mitchell DA, Dixon JE, Grishin NV. Expansion of type II CAAX proteases reveals evolutionary origin of gamma-secretase subunit APH-1. J Mol Biol 2011;410:18-26.

37. Levine TP, Daniels RD, Gatta AT, Wong LH, Hayes MJ. The product of C9orf72, a gene strongly implicated in neurodegeneration, is structurally related to DENN Rab-GEFs. Bioinformatics 2013;29:499-503.

38. Levine TP, Daniels RD, Wong LH, Gatta AT, Gerondopoulos A, Barr FA. Discovery of new Longin and Roadblock domains that form platforms for small GTPases in Ragulator and TRAPP-II. Small GTPases 2013;4:62-69.

39. Hirst J, Schlacht A, Norcott JP, Traynor D, Bloomfield G, Antrobus R, Kay RR, Dacks JB, Robinson MS. Characterization of TSET, an ancient and widespread membrane trafficking complex. eLife 2014;3:e02866.

40. Gerstein M. Measurement of the effectiveness of transitive sequence comparison, through a third 'intermediate' sequence. Bioinformatics 1998;14:707-714.

41. Pei J, Grishin NV. The P5 protein from bacteriophage phi-6 is a distant homolog of lytic transglycosylases. Protein science : a publication of the Protein Society 2005;14:1370-1374.

42. Yu JW, Mendrola JM, Audhya A, Singh S, Keleti D, DeWald DB, Murray D, Emr SD, Lemmon MA. Genome-wide analysis of membrane targeting by S. cerevisiae pleckstrin homology domains. Mol Cell 2004;13:677-688.

43. Murley A, Sarsam RD, Toulmay A, Yamada J, Prinz WA, Nunnari J. Ltc1 is an ER-localized sterol transporter and a component of ER-mitochondria and ER-vacuole contacts. J Cell Biol 2015;209:539-548.

44. Gillooly DJ, Morrow IC, Lindsay M, Gould R, Bryant NJ, Gaullier JM, Parton RG, Stenmark H. Localization of phosphatidylinositol 3-phosphate in yeast and mammalian cells. Embo J 2000;19:4577-4588.

45. Levine TP, Munro S. Targeting of Golgi-specific pleckstrin homology domains involves both PtdIns 4-kinase-dependent and -independent components. Curr Biol 2002;12:695-704.

46. Park JS, Neiman AM. VPS13 regulates membrane morphogenesis during sporulation in Saccharomyces cerevisiae. J Cell Sci 2012;125:3004-3011.

47. Lang AB, John Peter AT, Walter P, Kornmann B. ER-mitochondrial junctions can be bypassed by dominant mutations in the endosomal protein Vps13. J Cell Biol 2015;210:883-890.

48. Park JS, Thorsness MK, Policastro R, McGoldrick LL, Hollingsworth NM, Thorsness PE, Neiman AM. Yeast Vps13 promotes mitochondrial function and is localized at membrane contact sites. Mol Biol Cell 2016;27:2435-2449.

49. Ingley E, Hemmings BA. Pleckstrin homology (PH) domains in signal transduction. J Cell Biochem 1994;56:436-443.

50. Rothman JH, Hunter CP, Valls LA, Stevens TH. Overproduction-induced mislocalization of a yeast vacuolar protein allows isolation of its structural gene. Proc Natl Acad Sci U S A 1986;83:3248-3252.

51. Grille S, Zaslawski A, Thiele S, Plat J, Warnecke D. The functions of steryl glycosides come to those who wait: Recent advances in plants, fungi, bacteria and animals. Prog Lipid Res 2010;49:262-288.

52. Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 2005;33:W244-248.

53. Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res 2010;38:W529-533.

54. Haslam RJ, Koide HB, Hemmings BA. Pleckstrin domain homology. Nature 1993;363:309-310.

55. Doerks T, Strauss M, Brendel M, Bork P. GRAM, a novel domain in glucosyltransferases, myotubularins and other putative membrane-associated proteins. Trends Biochem Sci 2000;25:483-485.

56. Hanawa-Suetsugu K, Kukimoto-Niino M, Mishima-Tsumagari C, Akasaka R, Ohsawa N, Sekine S, Ito T, Tochio N, Koshiba S, Kigawa T, Terada T, Shirouzu M, Kohda D, Nishikimi A, Fukui Y*, et al.* Structural basis of functional complex formation between DOCK2 and ELMO1 for Rac activation in lymphocyte chemotaxis. DOI:102210/pdb3a98/pdb 2010.

57. Lu Q, Li J, Ye F, Zhang M. Structure of myosin-1c tail bound to calmodulin provides insights into calcium-mediated conformational coupling. Nat Struct Mol Biol 2015;22:81-88.

58. Im YJ, Raychaudhuri S, Prinz WA, Hurley JH. Structural mechanism for sterol sensing and transport by OSBP-related proteins. Nature 2005;437:154-158.

59. Gough J, Chothia C. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. Nucleic Acids Res 2002;30:268-272.

60. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 2016;44:D279-285.

61. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A. Pfam: clans, web tools and services. Nucleic Acids Res 2006;34:D247-251.

62. Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14:755-763.

63. Lees J, Yeats C, Redfern O, Clegg A, Orengo C. Gene3D: merging structure and function for a Thousand genomes. Nucleic Acids Res 2010;38:D296-300.

64. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 2011;39:W29-37.

65. Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, Hill DP, Kania R, Schaeffer M, St Pierre S, Twigger S, White O, Rhee SY. Big data: The future of biocuration. Nature 2008;455:47-50.

66. Hokanson DE, Laakso JM, Lin T, Sept D, Ostap EM. Myo1c binds phosphoinositides through a putative pleckstrin homology domain. Mol Biol Cell 2006;17:4856-4865.

67. Feeser EA, Ignacio CM, Krendel M, Ostap EM. Myo1e binds anionic phospholipids with high affinity. Biochemistry 2010;49:9353-9360.

68. Chen KY, Tsai PC, Hsu JW, Hsu HC, Fang CY, Chang LC, Tsai YT, Yu CJ, Lee FJ. Syt1p promotes activation of Arl1p at the late Golgi to recruit Imh1p. J Cell Sci 2010;123:3478-3489.

69. Jian X, Gruschus JM, Sztul E, Randazzo PA. The pleckstrin homology (PH) domain of the Arf exchange factor Brag2 is an allosteric binding site. J Biol Chem 2012;287:24273-24283.

70. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D*, et al.* InterPro: the integrative protein signature database. Nucleic Acids Res 2009;37:D211-215.

71. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Karra K, Krieger CJ*, et al.* Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res 2012;40:D700-705.

72. Remmert M, Biegert A, Hauser A, Soding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 2012;9:173-175.

73. Biegert A, Mayer C, Remmert M, Soding J, Lupas AN. The MPI Bioinformatics Toolkit for protein sequence analysis. Nucleic Acids Res 2006;34:W335-339.

74. Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A. FFAS03: a server for profile--profile sequence alignments. Nucleic Acids Res 2005;33:W284-288.

75. Kelley LA, Sternberg MJ. Protein structure prediction on the Web: a case study using the Phyre server. Nature protocols 2009;4:363-371.

LEGENDS

Fidler *et al.* Figure 1.

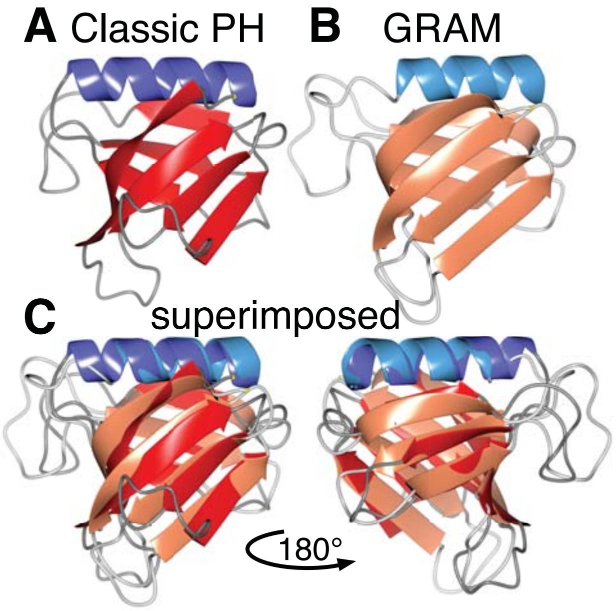


Figure 1: PH-like domains that share no significant sequence have highly similar folds.
Ribbon diagrams of the $C\alpha$ backbones of: **A.** classical PH domain in PEPP1 (1upq_A, residues 54-152); **B.** GRAM domain of MTMR2 (1lw3_A, residues 83-183); **C.** the two structures superimposed (root mean square distance = 1.9 Å across 86 residues) with views of both sides of the beta sandwich. Models colored by secondary structure: red = sheet, blue = helix; strong colors = classical PH, weak colors = GRAM. For more detail in A-C, see **Movies 1–3** respectively. Domains in pleckstrin were initially identified as a family of homologues ~100-120 aa[(54)], and several structures were solved soon after[(15)]. GRAM domains were identified as a family of sequences ~70 aa long[(55)] which the first solved structure identified as beta sheets 1–5 of a complete PH-like domain[(18)].

Fidler *et al.* Figure 2.

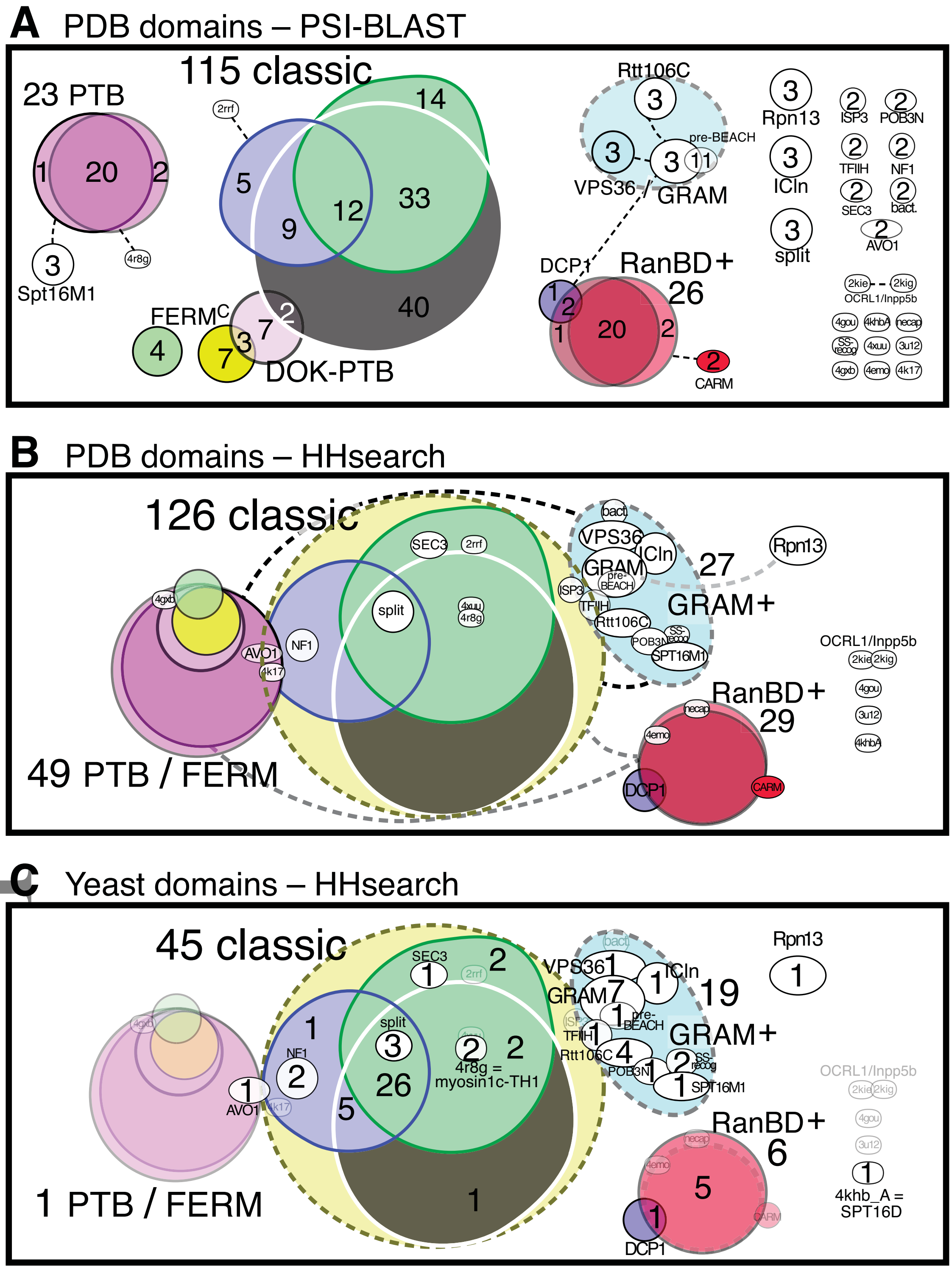


Figure 2: Family grouping of PH domains in PDB and yeast

A. ~240 PH-like domains in PDB were divided into 39 families by PSI-BLAST. The size of the coloured shapes and extent of overlap correlates with numbers of domains. Family names are standard abbreviations, and PDB identifiers are listed in Supplemental Table 1. Weak homologies, shown by dashed lines, were deduced from the presence of non-significant hits (p>0.005) that occurred with no false positive hits above them in hit lists. They lead to weak family grouping indicated by dashed outlines. Classical PH domains (~50% of all domains) are made up to three overlapping groups, and they have some overlap with some phosphotyrosine binding (PTB) domains. A high degree of overlap is seen between two PTB groups and in two Ran-binding domain (RanBD) groups (EVH1, WH1). 20 independent families contain one, two or three domains with no links to larger families. **B.** The same domains and families were analyzed by HHsearch. Domains that produce hits to each other with prob[SS]≥85% are grouped together. 16 of the 20 unlinked small families are now included in larger groups. PTB are fused with FERM-C domains; also GRAM domains are fused with nine other families. Some domains produce hits to both classical PH domains and either PTB/FERM-C domains or GRAM domains, leading to partial fusion of these three groups. Dashed lines indicate incomplete overlap (prob[SH] = 50-80% varying colour grey to black). **C.** The families arranged according to the HHsearch analysis from B were populated with 73 yeast PH-like domains described in the literature (details in Supplemental Table 2A). Empty families are shown in faint outline.

Fidler *et al.* Figure 3.

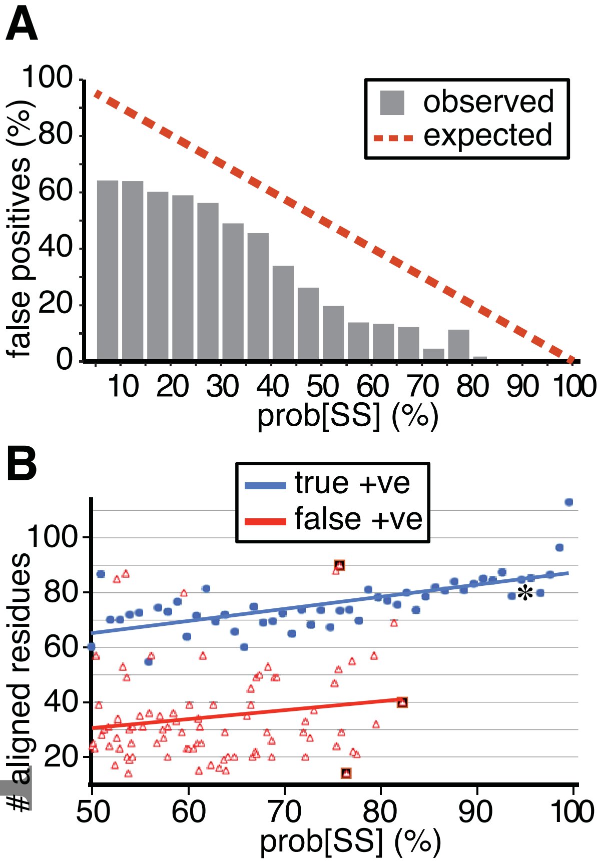


Figure 3. Properties of PH-like and non-PH-like hits

A. Specificity of HHsearch at different levels of prob[SS]. Hit lists with prob[SS]≥5% from 39 PDB-to-PDB searches were merged and scanned for the occurrence of non-PH-like domains. At each level of prob[SS], the rate of non-PH-like hits was less than predicted from the prob[SS] metric. No non-PH-like domain scored prob[SS]>85%. **B.** The relationship between prob[SS] (≥ 50%) and the number of aligned residues (COLs, see Box 3), both for true positives (showing means for every prob[SS] centile (total n=2200, median 22 hits per centile) and for non-PH-like hits (showing 90 individual occurrences). In both groups COLs increased with prob[SS]. Although COLs was lower for false positives, there were some exceptions. Non-PH-like hits similarly tended to have lower secondary structural similarity scores (data not shown). Three strong false positives (black squares) are described in detail in Supplemental Figure 1. Asterisk indicates the position of Age1p, a strong false positive in yeast (Supplemental Figure 1).

Figure 4 Fidler et al.

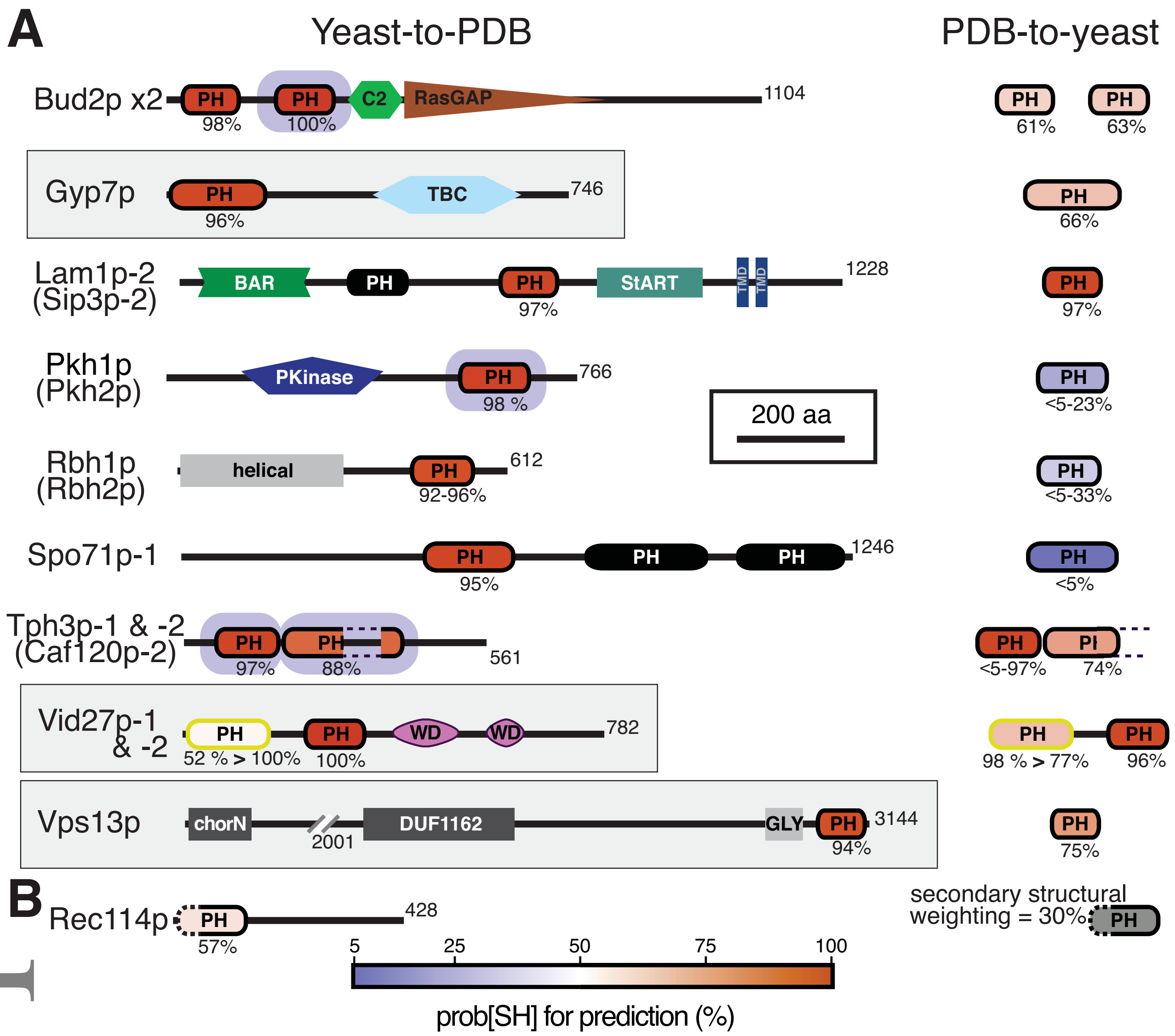
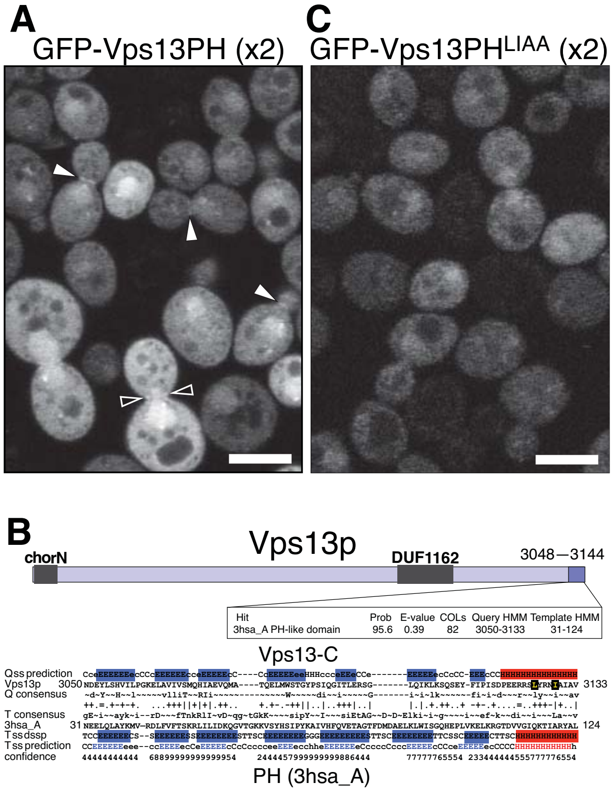


Figure 4: New PH-like domains identified in yeast

A. 16 strongly predicted new PH-like domains in yeast, shown in the context of 9 full-length proteins. 4 paralogs that share highly similar domain patterns have been omitted. New domains (black outline) are shaded according prob[SS] of strongest hits (blue–red graded scale). Main section shows yeast-to-PDB hits; right-hand section shows PDB-to-yeast hits. Where paralogs are reported in the same line, both prob[SS] values are reported, but the diagram and shading belong to the hit with higher prob[SS]. For the one domain predicted by indirect alignment (yellow outline, Vid27p-1), prob[SS] values are given for both searches. Light grey boxes (Gyp7p, Vid27p Vps13p) indicate new PH-like families present widely in eukaryotic evolution. Bud2p-2, Pkh1/2p, Tph3p and Caf120p (domains with green surrounds) have homologues with PH-like domains at the same position, so these new discoveries are to some extent expected. Accompanying domains include other, known PH-like domains (in Lam1p/Sip3p and Spo71p - shaded black) as well as C2, RasGAP, DH, TBC, BAR, StART, and others as follows: domain of unknown function=DUF; transmembrane domain=TMD, protein kinase=PKinase, chorein-N domain in VPS13=chorN, WD40=WD, glycine-rich=GLY. The N-terminus of Rbh1p (and Rbh2p) contains a helical domain of unknown function with homology to RhoGEFs. **B.** One PH-like domain tentatively identified with low prob[SH]. This was first found with PDB-to-yeast searches using increased secondary structural weighting, hence the usual prob[SS] value scale does not apply. Details of all newly predicted PH-like domains are in Supplemental Table 2B.

Fidler *et al.*, **Figure 5**



Hit	Prob	E-value	COLs	Query HMM	Template HMM
3hsa_A PH-like domain	95.6	0.39	82	3050-3133	31-124

```
Qss prediction     CceEEEEEEecCCcEEEEEEcceEEEEEcC----CcEEEEEEeeHHHccceEEeCCe-------EEEEEecCcCC-EEEcCCHHHHHHHHHHHHHHH
Vps13p        3050 NDEYLSHVILPGKELAVIVSMQHIAEVQMA----TQELMWSTGYPSIQGITLERSG-------LQIKLKSQSEY-FIPISDPEERRSLYRNIAIAV 3133
Q consensus        ~d~Y~~H~~l~~~~~vlliT~~RIi~~~~~~-------~~~~~W~~~~~di~~i~~~~~G-------i~i~lk~~~~~~-fi~i~d~~~r~~ly~~i~~av
                   ++.=.+-+-+.  ...+++||+|+|.++..    +.+-.-++||..|..+++|..|       |.|.+.++..+ -+-.+....=..+++-|+..|
T consensus        gE~i~~ayk~i--rD~~~fTnkRlI~vD~qg~tGkK~~~~sipY~~I~~~siEtAG~~D~D~Elki~i~g~~~~~i~~ef~k~~di~~i~~~La~~v
3hsa_A          31 NEELQLAYKMV-RDLFVFTSKRLILIDKQGVTGKKVSYHSIPYKAIVHFQVETAGTFDMDAELKLWISGQHEPLVKELKRGTDVVGIQKTIARYAL 124
T ss dssp          TCCEEEEEEECS--SEEEEEEESSEEEEEEEESTTSCEEEEEEEEGGGEEEEEEEEEESTTSCEEEEEEEETTCSSCEEEEECTTSCHHHHHHHHHHHH
T ss prediction    CCEEEEEEeee--ccEEEEecCeEEEEEcCcCccccceeEEEecchheEEEEEEeCccccCcccEEEEEcCCceEEEEEecCCCCHHHHHHHHHHHh
confidence         44444444444  6889999999999954    2444457999999999886444          77777765554 23344444455577777655 4
```

Figure 5: Intracellular targeting by the PH-like domain of Vps13p
A. GFP-Vps13-PH-PH (dimer) weakly targets the bud neck, seen as linear targeting across the neck of small-to-medium buds (filled arrowheads), and dots either side of occasional larger buds (hollow arrowheads). The minor nuclear enrichment is nonspecific, being seen with all other PH monomers and dimers (data not shown). **B.** Vps13p 3028-3144 as query (Q, top 3 lines) aligned with a target (T) hit from the solved structure 3hsa_A, a bacterial protein (*Shewanella amazonensis*, bottom 5 lines). The middle line indicates which residues align, where: "|" is a very good alignment, "+" is good, "." is neutral, "-" is bad, and "=" is a clash. For both Q & T, the secondary structure prediction (ss prediction) is in three states, E for sheet (blue), H for helix (red) and C for unstructured loop (black), with prediction confidence shown for target. The target also has "ss_dssp" showing its solved structure. The box above shows statistics on the hit, including prob[SS] and COLs. L3125 and I3129 (highlighted in yellow) are partially conserved residues (lower case in consensus) that align with 3hsa_A ("+" and "|" respectively). Alignment made by HHalign. **C**. GFP-tagged dimeric Vps13-PH(LIAA) (L3125A and I3129A) accumulates in cells to a much lesser extent than wild-type. Scale bars 5 µm.

Fidler *et al.*, **Figure 6**

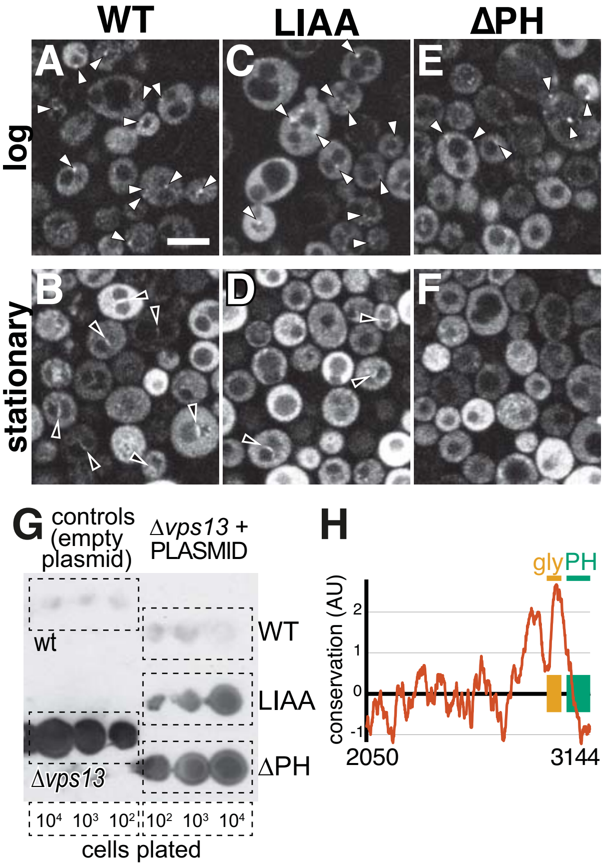


Figure 6: Intracellular targeting by PH-like domain of Vps13p
A-F. Vps13-EGFP constructs, A/B: wildtype (WT), C/D: L3125A I3129A (LIAA) and E/F: 1-3028 (ΔPH) in Δ*vps13* cells in log phase (A/C/E) or early stationary phase (B/D/F). Sites of localization include intracellular puncta (filled arrowheads) in log phase, and the nucleus vacuole junction (NVJ, open arrowheads) in stationary phase. Scale bar (shown in A only) 5 µm. **G.** Overlay blots to detect secreted CPY. Controls: wild-type cells and Δ*vps13* cells show no secretion and maximal secretion respectively. Rescue of Δ*vps13* was tested for Vps13-EGFP plasmids as in A-F. Results are representative of 6 similar experiments. **H.** Residue conservation in the C-terminal 1095 residues of Vps13p, calculated by ConSurf and scaled as described in Methods. The most conserved part of the Vps13 C-terminus is a glycine rich domain (orange) of no known function[51], not the predicted PH domain (green).

<u>**BOXES (and Glossary – terms in blue)**</u>

Box 1: Defining a PH-like domain

PH-like domains are the fourth most common fold in the human proteome[14]. Their 8 structural elements (βββββββα) span 100–120 aa, making an orthogonal, slightly splayed beta sandwich (4+3) capped by the helix (Figure 1, Movies 1 and 2). Excluding redundant sequences, >200 PH-like domain structures have been solved. Structures have confirmed earlier identifications made solely on the basis of remote sequence homology in most[56,57] but not all[58] cases.

The definition of a PH-like domain we use is **either: (1)** by structural similarity to the region occurring twice in pleckstrin. The peptide backbone follows the characteristic βββββββα□pattern closely; **or (2)** by significant sequence homology to a domain that fits definition (1). This may require indirect alignment (profile-sequence) rather than just sequence-sequence. Both the classical PH domain in PEPP1 and the GRAM domain in MTMR2 meet definition 1 (Figure 1, Movies 1-3), while in yeast the classical PH domain in Boi1p and the GRAM domain in Ymr1p (orthologous to MTMR2) meet definition 2.

Finding that a sequence has homology to a PDB entry documented as having a PH-like fold goes some way to meet definition 2. However, PDB entries may contain multiple different domains, so it is important to check that a sequence shares homology with the PH-like domain portion of the PDB sequence.

Box 2: Profile-sequence tools currently used to define domains

All profile-sequence tools are seeded with one or multiple sequences to find new hits, these being sequences that align more closely than the E-value threshold. The profile is a statistical representation of the different residues across the multiple sequence alignment (MSA) made from seed plus hits. The profile is then used to initiate another round of searching, and because it contains more sequence information than the previous round it may detect more hits. The whole process iterates through multiple rounds until no more sequences are added.

Profile-sequence tools differ in choice of seeds, in particular how broad a net to cast. For example, the tool SuperFamily only uses seeds with solved structure, while Pfam starts with any recognizable domain[59,60]. Tools also differ in the curation of the profiles obtained, with Pfam using human intervention to group structurally related families into clans[61]. Tools also vary in the way profiles calculate the MSA. The simple way, which minimizes computation, scores all alignments using fixed gap opening/extension penalties. A more sensitive, but computationally harder, technique is to convert the MSA into a hidden Markov model (HMM), which codes insertions or deletions at each position in a profile-specific way[9,62]. HMM profile-sequence searches are used in Pfam and Gene3D[63,64].

Domain definitions are updated regularly (*e.g.* Pfam now version 30, where the PH domain clan is CL0266, see pfam.xfam.org/clan/CL0266), but new discoveries are not disseminated to databases instantly[65]. Upgrades lag behind the literature and curation is not perfect. For example, Pfam describes a domain in unconventional myosins as "Myosin_TH1", but lacks any link to the PH-like fold[57,66,67]. In addition, some domains that do not meet the PH-like definition because homology is too weak are nevertheless accepted. The basis for this is typically that the remainder of the protein sequence is highly homologous to a protein that does contain a PH-like domain, and it is thought that the two proteins are full-length homologues. For example Syt1p has a Sec7-homology guanine exchange factor (GEF) domain, and related GEFs in other species have a classical PH domain. The non-significant hit in Syt1p (E-values: SMART= 0.002, Pfam not found (>10), SuperFamily =0.05) has been accepted as PH-like, because the GEF domain homology strongly enhances the likelihood of a PH-like domain existing[68,69].

Yeast is one of the most highly annotated organisms. The Saccharomyces Genome Database (SGD) is annotated by InterProScan, which is a resource from the European Bioinformatics Institute (Hinxton, UK). InterPro incorporates all of Pfam, SuperFamily, GENE3D and many other profile-sequence tools[70,71]. For PH-like domains in yeast, InterPro provides useful data, but its output is not close to 100% sensitive, only including 61 out of 73 domains (Supplemental Table 2). Also, it is not entirely specific, as it includes three false positives (Supplemental Table 2).

Box 3: Profile-profile searches with HHsearch (HHpred).

HHsearch is freely accessible profile-profile search software that is run either offline, or on a webserver called HHpred. The program is powerful[13], fast and flexible, with many settings that can be varied by users.

HHsearch has three stages[9]: **(1)** starting with a single seed or aligned multiple seeds ("query"), MSAs are made by multiple iterations of HHblits, which is more sensitive than PSI-BLAST[72]. **(2)** PSIPRED predicts one of three structural states (helix = H, sheet =E, loop = C) for each position in the MSA[12]. **(3)** the query HMM containing both sequence and predicted secondary structure is used to search through pre-made libraries of target HMMs built before-hand. Freely available libraries (at http://toolkit.tuebingen.mpg.de/) include both "knowns" such as PDB, and curated domains from Pfam, SuperFamily, Gene3D(CATH) and other profile tools, and also "unknowns", *i.e.* proteomes, both prokaryotes (>30) and eukaryotes (x9): human, mouse, fruit fly, nematode, malarial parasite, thale cress and three fungi: budding yeast (*S. cerevisiae*), fission yeast (*S. pombe*) and maize blast (*U. maydis*). The two libraries used mostly for this study are PDB, which contains ~35,000 entries (non-redundant at 70% sequence identity) updated weekly, and yeast, which has 5870 entries. A third library constructed in-house was used for off-line searches (see Yeast-to-PDB section, and Methods**)**.

Important variables that can be set by the user for each of the three stages are:
(A) number of iterations. The defaults number of iterations for building the query MSA is 3. Libraries of targets have been built with 2 iterations. Iterating more times (maximum number = 8) enhances sensitivity, but takes longer. Searches designed specifically to identify only orthologues within a large family of homologues may benefit from using 1 iteration, deliberately reducing sensitivity.
(B) scoring secondary structure. There is an option to not score secondary structure at all. Run off-line, it is also possible to vary the proportion of the overall alignment weighted to secondary structure. Default is 11%, *i.e.* sequence alignment dominates.
(C) using solved structure. If the PDB is being used as either query or target, the user can choose to use solved secondary structure states (reduced to 3 states) instead of predicted structure.

HHsearch returns up to ~20,000 hits ordered in terms of descending probability of shared structure (prob[SS]) with the query sequence. For each hit, apart from the prob[SS] value, there is a precise alignment (see Figure 5B) that can be summarized by three important values: (i) the statistical expectation (E-value) of achieving the hit in terms of sequence alone given the size of the library, equivalent to the HHblits E-value; (ii) the 2° structural similarity (2Ssim), which is independent of sequence homology; and (iii) the number of columns aligned between query and target (COLs), which can be as few as 6. We applied a length threshold that the alignment should cover 20% of the domain, *i.e.* 20 aligned columns, and we found that shorter hits were often to a single structural element (Supplemental Figure 1A). Hits that can be detected by PSI-BLAST typically have a prob[SS]≥99% and an E-value of 10^{-10}, which reflects the greater sensitivity of HHblits than PSI-BLAST.

Box 4 : Asymmetry in searches (Q: "How do I get where I want to go?" A:"I wouldn't start from here")

Profile-profile tools in a trivial way might be symmetric. So long as profiles are made the same way for use as queries and targets, comparing them either way round (query→target or target→query) makes no difference. On the other hand profile-sequence tools are dependent on where they start because the profile represents a large group of sequences centred on the query. Thus, a profile-sequence tool searching in one direction (*e.g.* known-to-unknown) will probably not make the same hits as the same tool searching in the other direction (*e.g.* unknown-to-known).

One way to overcome the problem of profile bias in favor of where it started is to manipulate the growing MSA to attempt to eradicate any trace of its origin. HHblits does this by filtering an MSA (thousands of sequences) to only pick a small group (~100) that are chosen to maximize diversity(72).

Box 5: Workflows in HHpred

A. To find a fold/function for an unknown protein:

1. Carry out protein-to-PDB searches. All hits above a fairly generous threshold (prob[SS]>50%, COLs>20) are worth considering further.
2. Carry out domain-to-PDB searches to exclude false positives (Supplemental Figure 2).
3. For every interesting region, confirm the significance of the hit by comparing its strength to that obtained by the strongest false positives in PDB-to-PDB searches (set to detect predicted structure only).

B. To maximally extend the structural homologues for a known fold in a proteomes with an HMM library (as for PH-like domains here)

1. Use PSI-BLAST to define families that ensure representation of all families with the fold within the overall clan.
2. Set a threshold for prob[SS] that avoids false positives, using HHsearch to carry out PDB-to-PDB searches with one sequence from each family (with HHsearch option set to ignore known structure, and use predicted secondary structure).
3. Seed PDB-to-unknown HHsearch searches with these sequences, looking in the proteome of choice (one per family), identifying positive hits with the threshold obtained above.
4. Investigate non-significant hits, especially those with relatively high COLs and 2Ssim (Box 3), with unknown-to-PDB searches.
5. Link further hits to PDB entries by indirect searches in related organisms. First, use unknown-to-related proteome searches to identify hits that have no overt homologues in the proteome of choice. Use these hits to carry out related proteome-to-PDB searches.
6. Confirm all new matches with domain only searches.

A more complex alternative for the confirmatory steps (A2 and B6) is to create HMMs for the query and target regions with 8 iterations, and align them pairwise with HHalign(73).

Glossary

Clan: a group of domain families that share the same fold but cannot be linked by standard sequence alignment tools (*e.g.* PSI-BLAST) alone. Sometimes called superfamilies.
E-value: each alignment between query and target is assigned an e-value, which is the number of hits as good as the one obtained that would be expected to occur randomly given the size of the database being searched. A threshold is chosen, (here 0.001) so that alignments more statistically unlikely than this are considered significant.
HHalign: pairwise application of HHsearch for direct comparison of two MSAs.
HHblits: Builds profiles from a single query via multiple rounds of searching, differing from PSI-BLAST in several ways that increase speed and sensitivity.
HHpred: online server for HHsearch.
HHsearch: profile-profile tool that explicitly weights a proportion of an alignment to aligned (predicted) secondary structure. MSAs can be built by either PSI-BLAST or HHblits.
Hit: target in database that aligns with an e-value more statistically significant (i.e. lower) than the chosen threshold.
HMM: Hidden Markov Model: a way to represent MSAs, coding features including penalizing insertions and deletions in a profile-specific way; more computationally complex than a simple profile
MSA: multiple sequence alignment: sequences with multiple small regions of local homology are aligned across a large region by inserting gaps. MSAs can be dominated by large numbers of highly related sequences (*e.g.* mammalian orthologues). This problem can be addressed by filtering MSAs for redundancy, reducing non-diversity.
Pfam: a tool that identifies domains broadly, even without known structure/function. Pfam domains are constructed as HMMs. They are curated, reliable, and when structurally related families are found, they are grouped into clans on the basis of structural homology confirmed by HHsearch(60).
Profile: statistical representation of the residues across an MSA, sometimes represented as a protein "logo". Simple profiles apply inflexible, standard rules, for example for insertions/deletions.

PSI-BLAST: Position-specific iterated basic local alignment search tool. Builds profiles from a single query via multiple rounds of BLAST.

Query (or seed) : sequence from which searches are initiated

SuperFamily: a narrow specificity tool for identifying domains similar to known structures, similar to SCOP.

Target: group of sequences curated into a database among which homologs are being sought. Database size varies from relatively small (all solved structures, predicted proteins in individual genomes) to very large (all proteins in all genomes). Size can be reduced by excluding redundant sequences sharing more than a pre-determined level of sequence identity (20-90%).