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

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

*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
Timothy P. Levine: tim.levine@ucl.ac.uk

Supplemental Information
Supplemental Table 1. Classifications of PH-like domains.

Information on our classification of 39 PH-like families (see Figure 2), starting with the major groups that contain overlapping families in some classifications. Our classification is compared to those in Conserved Domain Database (CDD) at NCBI, the Structural Classification of Protein (SCOP) database at scop.berkeley.edu, and Pfam 30.0. RED indicates PH-like families that were missing from the PH-like clan. "X" indicates that the family was missing completely. The penultimate column shows the number of PH-like domains for each family in PDB (using a version non-redundant at 70% identity), both for grouped families (excluding extensive overlap, n=34) or the full list (n=39). The final column shows the one structure we chose to represent the family for searches in PSI-BLAST and HHsearch. Where the structure includes more than one PH-like domain, the segment used is indicated. The bottom lines show the total number of PH-like families in each classification, and information on one domain wrongly assigned as PH-like in Pfam.
## Supplemental Table 2: details of 91 yeast PH-like domains

| A | Known PH-like proteins | systematic gene name | protein length | residues | 1 | 2 | 3 | 4 | 5 | note |
|---|------------------------|----------------------|----------------|----------|-----|---|---|---|---|------|
| Ask1p | Ygr097 | 1146 | 477-725 cPH | cPH cPH cPH | 92.4 | cPH(1/3) | 99.9 | h |
| Alg26p-1 | Ynl189c | 1198 | 165-340 cPH | cPH cPH cPH | 99.2 | cPH(1/2/3) | 99.9 | h |
| Alg26p-2 | Ynl189c | 1198 | 575-682 gram | gram gram | 98.1 | gram | 97.6 | h |
| Avl1p | Yol079w | 1176 | 1045-1172 a | * | * | 100.0 PTB(Ave1) | 100.0 c |
| Bem2p | Yer155c | 2167 | 1780-1957 cPH cPH cPH | 92.7 cPH(1/2/3) | 99.9 | h |
| Bem3p | Ypl115c | 1128 | 630-745 cPH cPH cPH | 99.1 cPH(sp) | 99.9 | h |
| Bol1p | Ybl069w | 980 | 778-895 cPH cPH cPH | 99.5 (PH)(1/2/3) | 99.9 | h |
| Bol2p | Yel114c | 1040 | 769-889 cPH cPH cPH | 99.5 (PH)(1/2/3) | 99.9 | h |
| Bph1p | Ycr032 | 2167 | 1370-1467 BPH BPH BPH | 99.8 Pre- | 93.4 c |
| Bud4p | Yil092w | 1448 | 1300-1420 cPH cPH cPH | 98.9 cPH(1/2/3) | 99.9 | h |
| Ctf12p-2 | Ynl278w | 1063 | 74-214 cPH cPH cPH | 93.1 cPH(1/3) | 99.9 | h |
| Ccr34p | Yab14w | 854 | 464-662 cPH cPH cPH | 98.1 cPH(1/2/3) | 99.7 | h |
| Cle4p | Ynl298w | 842 | 60-183 cPH cPH cPH cPH | 98.9 cPH(1/2/3) | 99.9 | h |
| Dcp1p | Yol149w | 231 | 42-223 + | Dcp1 Dcp1 | 100.0 | Dcp1 | 100.0 c |
| Exo84p | Ybr102c | 753 | 352-443 cPH cPH cPH | 98.6 cPH(1/2) | 99.8 | c |
| Fab1p | Yjl157c | 830 | 444-540 a | * | * | 95.2 cPH(2) | 78.1 h |
| Ira1p | Ybr140c | 3092 | 2204-2302 a | * | * | 99.9 cPH(N1) | 100.0 h |
| Ira2p | Yol081w | 3079 | 2205-2301 a | * | * | 99.9 cPH(N2) | 100.0 h |
| Lam4p | Ynh090c | 1345 | 535-661 gram gram | gram | 97.3 gram | 96.6 h |
| Lam5p | Yil042c | 674 | 223-305 gram gram | gram | 97.9 gram | 95.9 h |
| Lam6p | Yel072w | 693 | 170-270 gram gram | gram | 97.7 gram | 96.4 h |
| Las17p | Yol181 | 633 | 21-134 RBD RBD RBD | 99.5 RBD(2) | 100.0 c |
| Lmo1p | Yln007c | 665 | 462-600 a | * | * | 97.1 cPH(3) | 100.0 h |
| Lof5p | Yln183w | 171 | 50-190 ICIN ICIN | * | * | 99.9 ICIN | 99.8 h |
| Mad1p | Ygr103 | 950 | 40-157 gram gram gram | gram | 97.5 gram | 98.0 h |
| Myo3p | Ykl129c | 1272 | 807-911 a | * | * | 99.7 cPH(4/4g) | 96.3 h |
| Myo5p | Ymr109 | 1219 | 808-911 a | * | * | 99.8 cPH(4/4g) | 97.2 h |
| Num1p | Ydr150 | 2748 | 2572-2687 cPH cPH cPH | 98.1 cPH(1/3) | 99.9 h |
| Num2p | Yol035w | 720 | 602-718 RBD RBD RBD | 99.9 RBD(2) | 100.0 h |
| Nvt2p | Ypr091c | 770 | 125-268 cPH cPH cPH | 99.8 cPH(1/3) | 99.9 h |
| Opy1p-1 | Ybr129c | 328 | 65-155 cPH cPH cPH | 99.8 cPH(1/3) | 99.9 h |
| Opy1p-2 | Ybr129c | 328 | 213-320 cPH cPH cPH | 99.8 cPH(1/3) | 99.9 h |
| Och1p | Yar042 | 1188 | 280-383 cPH cPH cPH | 99.1 cPH(1/2/3) | 99.9 h |
| Och2p | Yel819c | 1293 | 288-390 cPH cPH cPH | 99.3 cPH(1/2/3) | 99.9 h |
| Och3p | Ynl073 | 996 | 220-318 cPH cPH cPH | 99.5 cPH(1/2/3) | 99.9 h |
| Ptc1p | Ypl269w | 869 | 122-217 + | 812 cPH cPH | 98.4 cPH(1/2/3) | 99.9 h |
| Pop3p-1 | Yml069 | 552 | 01-111 a | * | * | 94.2 Pop3N | 100.0 c |
| Pop3p-2 | Yml069 | 552 | 102-232 + | | | 99.9 SSrecog | 100.0 c |
| Pop3p-3 | Yml069 | 552 | 247-347 + | + | + | 96.9 SSrecog | 100.0 c |
| Pop3p-4 | Yml069 | 552 | 365-470 Rtt106 Rtt106 Rtt106 | 99.7 Rtt106C | 100.0 c |
| Psy2p | Ynl021c | 858 | 28-127 a | PH RBD | 99.5 RBD(2) | 100.0 h |
| Rpp1p | Ylr115 | 1063 | 403-581 cPH cPH cPH | 99.3 cPH(1) | 99.8 h |
| Rom1p | Ygr070 | 1155 | 692-822 + | PH cPH | 99.1 cPH(1/3) | 99.6 h |
| Rom2p | Ynl371w | 1356 | 875-1020 + | PH cPH | 99.2 cPH(1/2/3) | 99.5 h |
| Rpn13p | Yln421c | 156 | 20-132 Rpn13 | | | 100.0 | Rpn13 | 100.0 c |
| Rtt106p-1 | Ynl026c | 455 | 70-197 Rtt106 | | | 99.8 SSrecog | 100.0 c |
| Rtt106p-2 | Ynl026c | 455 | 215-305 Rtt106 Rtt106 | 100.0 Rtt106C | 100.0 c |
| See3p | Yae080c | 1136 | 106-223 | PH | | | 100.0 c |
| Spk3p-1 | Ynl257c | 1229 | 308-430 cPH cPH cPH | 99.3 cPH(1/2/3) | 99.9 h |
| Skg3p-1 | Ybr187w | 1026 | 89-230 cPH cPH cPH | 97.3 cPH(1/2/3) | 99.8 h |
| Skg3p-2 | Ynl113w | 1026 | 253-385 cPH cPH cPH | 94.6 cPH(1/3) | 97.4 h |
| Slm1p | Ynl113w | 655 | 03-122 cPH cPH cPH | 99.3 cPH(1/2/3) | 99.9 h |
| Slm1p | Ynl105c | 686 | 467-583 cPH cPH cPH | 98.2 cPH(1/2/3) | 99.9 h |

continued overleaf
### Supplemental Table 2. Description of PH-like domains in yeast

#### A. 73 known PH-like domains in yeast identified by 5 different methods: InterProScan (available through yeastgenome.org), Smart, SuperFamily (methods 1–3) and HHsearch (PDB-to-yeast and yeast-to-PDB =)

| Protein name | Systematic gene name | Protein length | Residues | PSI | Method | p[SS] | Yeast-to-PDB | PDB-to-yeast |
|--------------|----------------------|----------------|----------|-----|---------|-------|-------------|--------------|
| A (cont'd)   |                      |                |          |     | InterProScan ± |     |             |              |
|              |                      |                |          |     | SMART ± |     |             |              |
|              |                      |                |          |     | SUPERFAMILY ± |     |             |              |
|              |                      |                |          |     | PPH (Pfam) ± |     |             |              |
|              |                      |                |          |     | PPH_HHsearch + |     |             |              |

#### B. New PH-like proteins

| Protein name | Systematic gene name | Protein length | Residues | PSI | Method | p[SS] | Yeast-to-PDB | PDB-to-yeast | HHAlign |
|--------------|----------------------|----------------|----------|-----|---------|-------|-------------|--------------|---------|
|              |                      |                |          |     | InterProScan ± |     |             |              |         |
|              |                      |                |          |     | SMART ± |     |             |              |         |
|              |                      |                |          |     | SUPERFAMILY ± |     |             |              |         |
|              |                      |                |          |     | PPH (Pfam) ± |     |             |              |         |
|              |                      |                |          |     | PPH_HHsearch + |     |             |              |         |

#### Notes
- Known positives identified: 62
- Known positives identified: 0
- False positives (non-PH domains identified): 5
- False positives (non-PH domains identified): 0
- PDB-to-yeast: 72
- Yeast-to-PDB: 72
- HHAlign: 72
methods 4 and 5). For methods 1-3, any family specified is indicated (cPH = classical PH), while “+” indicates a generic PH-like identification. § False positives in InterProScan were produced by Gene3D in Cbk1, Fpk1p and Kin82p. HHsearch identified Age1p as a false positive (see Supplemental Figure 4). For method 4, the prob(SS) and family (among the 39 we defined – see Figure 2) of the strongest hit to the yeast domain is indicated. Other notes: c = crystal structure solved, h = homologous to a solved structure, o = omitted by Yu et al. (2004) (64); ° number of PH-like domains within individual proteins takes account of the new domains we identify (see Figure 4).

B. Newly identified PH-like domains in yeast. Details include discoverability by different tools: (1) PSI-BLAST, showing E-value (log₁₀) and number of iterations (#) when a known PH-domain (definitions 1 or 2 including yeast proteins in part A above) is found in the hit list; T indicates only temporarily in hit list, lost before convergence; (2) SMART domain prediction server, showing E-values (log₁₀) for any domains detected; (3) FFAS profile-profile tool, (score is more significant when more negative, threshold = −9) (47); (4) PHYRE2 structural prediction tool (showing probability of PH-like fold, if top hit) (48). HHsearch: method 5 = PDB-to-yeast (prob(SS) and query); method 6 = yeast-to-PDB (prob(SS) and target); method 7 = HHalign result with target in method 6. Prob(SS) values outlined in red were used to make identifications: 4 in PDB-to-yeast, 11 in yeast-to-PDB; one further identification was made by indirect searches (outlined in black).

Other notes: ¶ Bud2p-1 is indicated by SMART as overlapping a false-positive C2 domain; § new true positive in Pkh1p already in SGD (identified by Gene3D); “T” indicates that for Pkh2p and Vid27p-2 the PSI-BLAST hit was only temporary, and was missing at convergence. Grey shading = non-significant hit in Rec114p found using ssw=30%; yellow = false positive for Age1p-N, where the top hit in yeast-to-PDB searches was a thioredoxin, and pairwise alignment of Age1p-N and the PH domain from centaurin (human ARFGEF) had prob[SH]=0.1%. 
Supplemental Figure 1: Structures of strongest hits that are not PH-like domains

Three of the strongest false hits to non-PH-like proteins among PDB-to-PDB HHsearches with 39 PH-like domains (indicated as black squares in Figure 3A): **A.** C-terminus of yeast Get5p (3vej_A) aligned at 14 of its 41 residues to the helix of 4chj_A, prob[SS]=77%; **B.** mid-section of human FANCL (3zqs_A) aligned at 40 of its 186 residues to strands 3–5 of 2kig_A, prob[SS]=83%. **C.** αβ subdomain of NS1 from Dengue virus type 2 (within a multi-domain structure of 350 aa, not shown) aligned at 90 of its 114 residues to all 7 sheets in 4gou_A, prob[SS]=76%. Aligned regions shown in colors (blue = helix, red = sheet, yellow = loop). While some non-PH-like hits are very short and can be identified as non-PH-like by that criterion (A), others cover multiple structural elements and cannot be excluded on the basis of shortness (B & C).
Supplemental Figure 2: Dissection of a false positive in Age1p

A. Alignment showing a strong hit obtained in HHsearch between the PH-like domain of human ArfGEF Centaurin and the N-terminus of Age1p. For interpretation of the alignment, see Figure 5B legend. Three unstructured loops in centaurin of 43, 7 and 8 aa are omitted. The single most conserved residue in PH-like domains is the tryptophan in the helix (asterisk), which is reported as making a very good hit (|) with a glutamine in Age1p.

B. All sequence from the first 102 columns of the Age1p MSA in HHsuite. This part of the MSA includes only 3 of 164 sequences: Age1p (S. cerevisiae) Arf-GEF (Yarrowia) Arf-GEF (Tetraodon).

The N-terminus of Age1p has homologs only in species very closely related to S. cerevisiae, so this part of the MSA should only have one sequence. However, the downstream GEF (Age1p residues 168-307) has 163 homologs, many of which have a domain structure BAR-PH-GEF. Two of these align to part of Age1p 1-130 (A). These give a PH-like appearance to the MSA at its N-terminus that is not present in the original seed (B).
Supplemental Figure 3: New domains found in 132 consecutive yeast genes

132 consecutive verified open-reading frames on the long arm of chromosome IV ("YDR") starting at the centromere (YDR001C-134C) were analyzed by HHsearch. Names in bold indicate proteins (x16) where new domains were documented (x21). New domains are outlined in black and shaded by prob(SS) on a white–red scale; scores and main hit in PDB or Pfam are also given. All new hits were confirmed by pairwise alignments of domains alone. Asterisks indicate that the domain was detected by SMART even though absent from InterProScan data displayed at yeastgenome.org. Known domains in these proteins are shaded as in Figure 4 or in green. Naming: where a protein only has a systematic name, just the suffix is given (i.e. "XYZ w/c" from the full name YDRXYZw/c). Red indicates the proteins (19 = 14%) without prior structural/functional predictions (discounting assignment to a fungal protein family and prediction of transmembrane domains). Seven of these (37%) had newly predicted domains. By contrast, nine of 113 (8%) previously characterised proteins had newly identified domains.
Movies 1-3: PH-like domains share structure
Rotating models of structures in Figure 1.