Science Signaling AAAS

www.sciencesignaling.org/cgi/content/full/5/215/rs1/DC1

Supplementary Materials for

Proteome-Wide Discovery of Evolutionary Conserved Sequences in Disordered Regions

Alex N. Nguyen Ba, Brian J. Yeh, Dewald van Dyk, Alan R. Davidson, Brenda J. Andrews, Eric L. Weiss, Alan M. Moses*

*To whom correspondence should be addressed. E-mail: alan.moses@utoronto.ca

Published 13 March 2012, *Sci. Signal.* **5**, rs1 (2012)
DOI: 10.1126/scisignal.2002515

This PDF file includes:

Fig. S1. Schematic of the phylo-HMM approach.
Fig. S2. Regions with no conserved segments are not detected by the phylo-HMM approach.
Fig. S3. Newly identified KEN box in Spt21 and Cbk1 interaction motif in Ssd1 are conserved in further yeast species.
Fig. S4. Simulation of protein evolution.
Fig. S5. Performance of the phylo-HMM approach on literature-curated short linear motifs.
Fig. S6. Binding of FxFP peptides to Cbk1.
Fig. S7. Phylogenetic tree of species used for this study.
Tables S1 to S4 legends
Table S5. Annotation of top 20 clusters from different distance metrics of predicted short conserved sequences.

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencesignaling.org/cgi/content/full/5/215/rs1/DC1)

Table S1 (Microsoft Excel format). Predictions on the yeast proteome by the phylo-HMM approach.
Table S2 (Microsoft Excel format). Literature-curated characterized short linear motifs.
Table S3 (Microsoft Excel format). Enrichment analysis of motifs matching known consensus sequences.
Table S4 (Microsoft Excel format). Clusters of similar short conserved sequences.

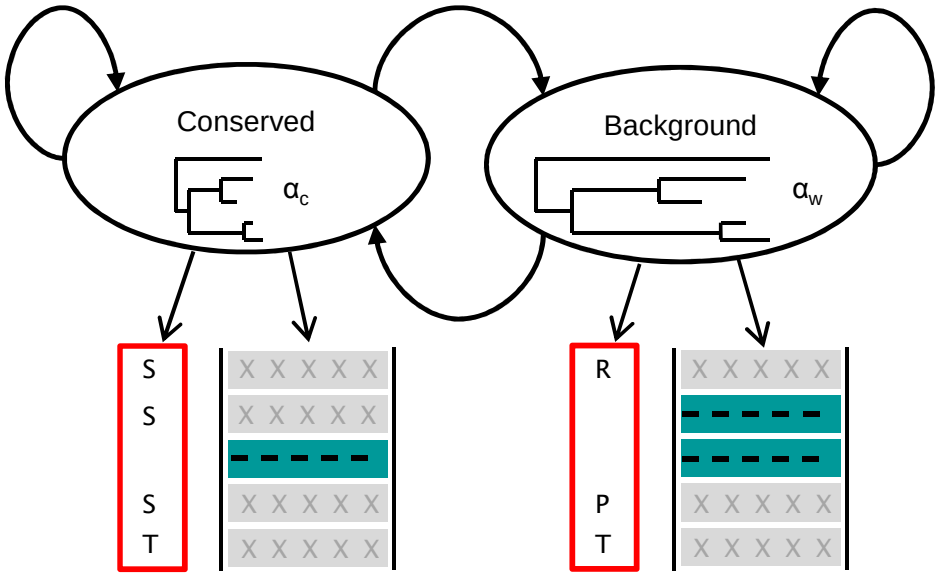


Fig. S1. Schematic of the phylo-HMM approach. In the phylo-HMM framework, a column of the sequence alignment is assumed to belong to either a conserved state ("Conserved") or a background state ("Background") and probabilities of observing alignment columns in each state depend on a phylogenetic model of protein evolution with a rate parameter α.

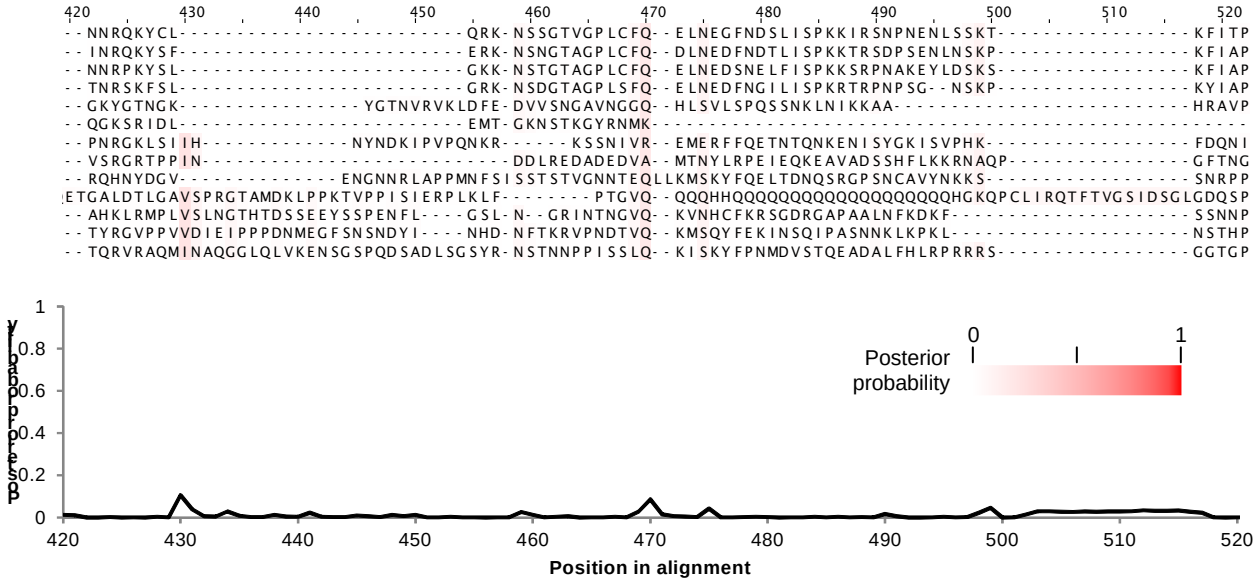


Fig. S2. Regions with no conserved segments are not detected by the phylo-HMM approach. Posterior trace of the region 420-520 in the alignment of Swi5. No locally conserved segments were identified. The region shown corresponds to position 266-322 in *S.cerevisiae*. Red color intensity represents the posterior probability of the conserved state.

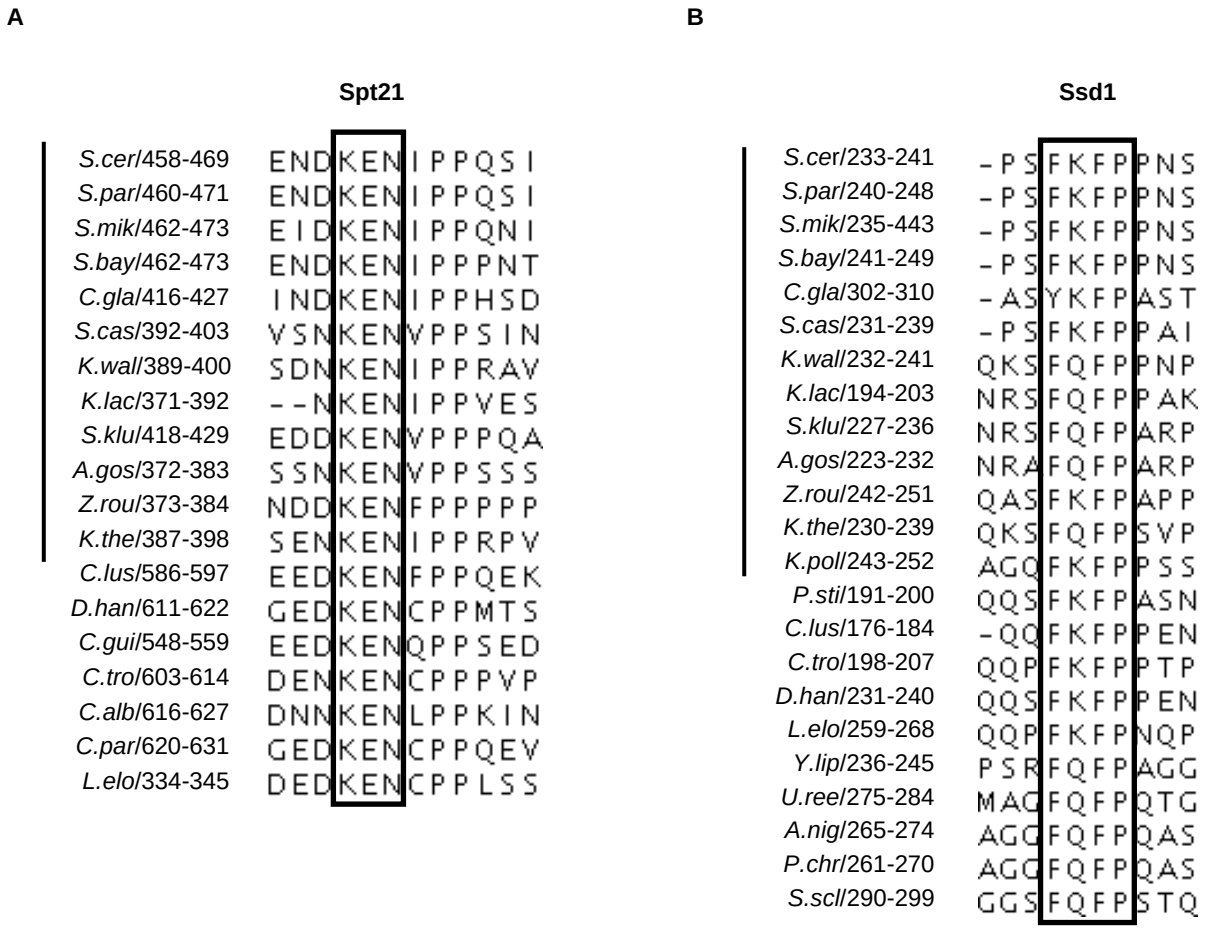


Fig. S3. Newly identified KEN box in Spt21 and Cbk1 interaction motif in Ssd1 are conserved in further yeast species. (**A**) Alignment of the KEN box (in rectangle) in Spt21 with distantly related yeast species. We note that the KEN motif in the Candida clade does not align with species used in our study when performing a whole gene alignment. However, the motif is well conserved. (**B**) Alignment of one of the Cbk1 interaction motif (in rectangle) in Ssd1 with distantly related yeast species. In both panels, species used for the phylo-HMM analysis are labelled with a vertical black bar.

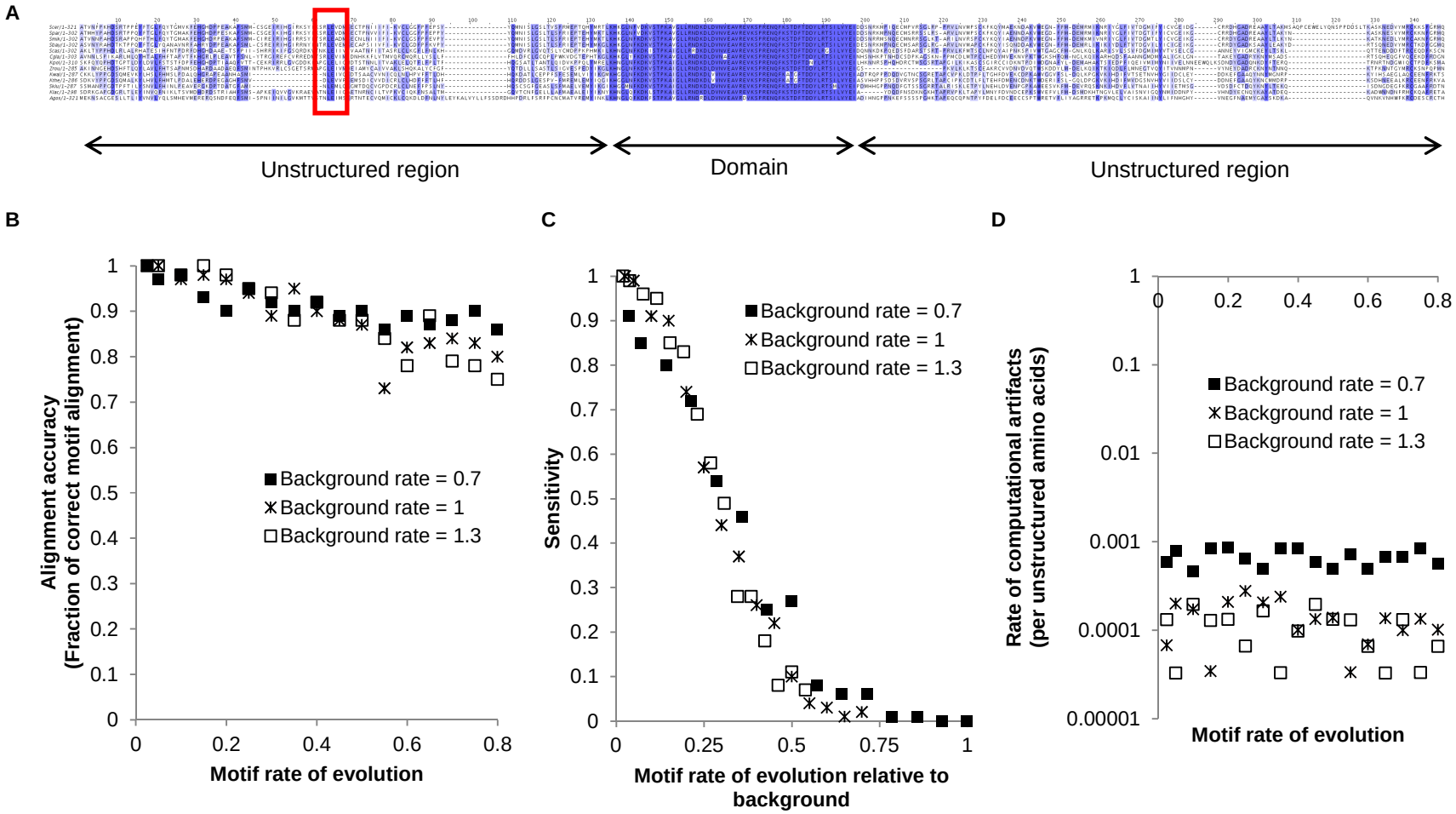


Fig. S4. Simulation of protein evolution. (**A**) Simulations of proteins evolution were performed by randomly generating a protein sequence containing a motif in an unstructured region. An alignment of a typical simulated protein is shown with the motif, properly aligned, boxed in red. (**B**) Alignment accuracy of the motif using MAFFT depends on the background rate of evolution and on the rate of motif evolution. (**C**) The sensitivity of the phylo-HMM on simulated data shows strong dependence on the relative rate of evolution of the motif to the rate of evolution of the background. (**D**) The rate of computational artifacts of the phylo-HMM on simulated data is dependent on the background rate of evolution. Each point represents the results from 100 simulated proteins.

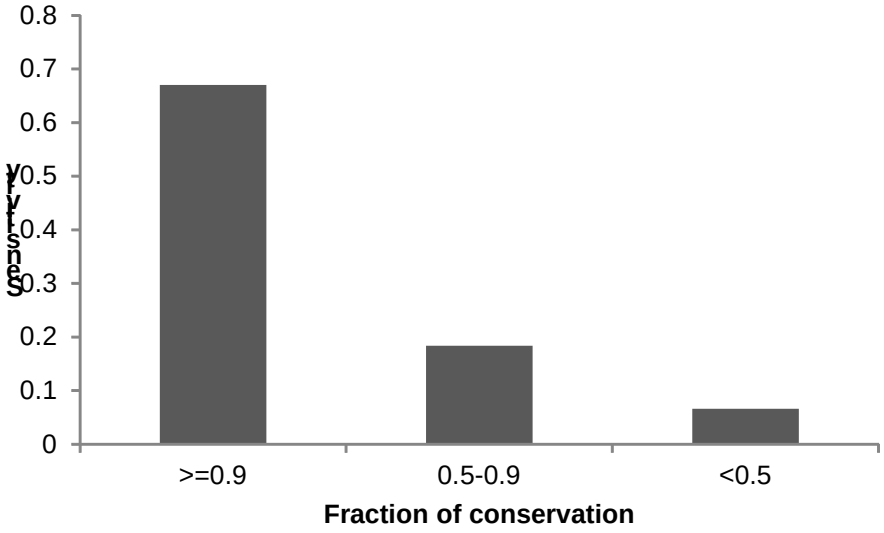


Fig. S5. Performance of the phylo-HMM approach on literature-curated short linear motifs. On characterized short linear motifs, the phylo-HMM performs best on conserved regulatory sequences. Data shown here only includes motifs with consensus sequences (no localization signals). Regulatory motifs were binned on the basis of the fraction of species in which the consensus sequence could be found (Fraction of conservation) and sensitivity of the phylo-HMM was calculated for each bin (see Methods).

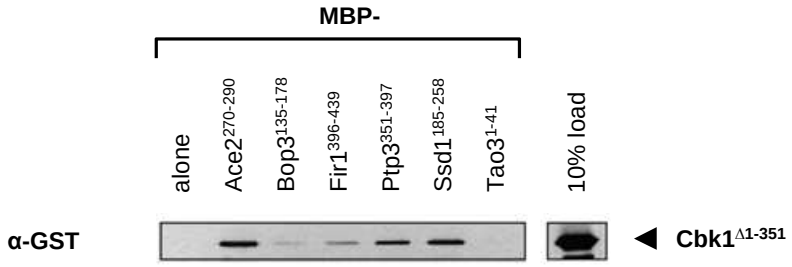


Fig. S6. Binding of FxFP peptides to Cbk1. Fragments from proteins identified in the [YF]xFP cluster (Fig. 5C) were expressed as maltose-binding protein (MBP) fusions and immobilized on amylose resin. The beads were assayed for binding to GST-tagged Cbk1 ($Cbk1^{\Delta 1-351}$) in a pulldown assay. At the exposure time shown in Figure 6, the lane showing 10% of the GST-tagged Cbk1 input (on a different nitrocellulose membrane but imaged at the same time and incubated with the same conditions) is overexposed. A reduced exposure of the assay is shown in this figure for comparison.

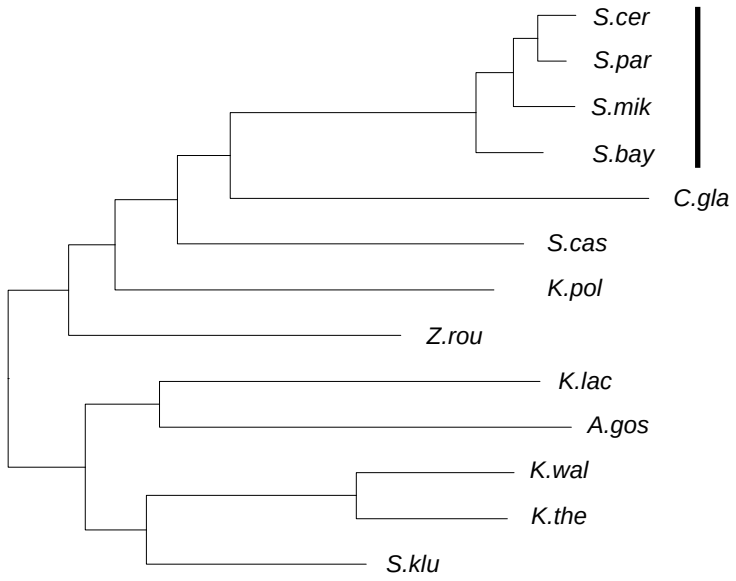


Fig. S7. Phylogenetic tree of species used for this study. Protein sequences from related yeast species were used to predict short conserved sequences by our phylo-HMM approach. The vertical black line indicates the four closest species of yeast that were used to obtain the amino acid substitution model. The branch lengths were estimated from random concatenations of proteins (See Methods) and the Newick format tree representation is the following: ((((((((Scer: 0.0317431, Spar: 0.022837): 0.0200533, Smik: 0.0499671): 0.0302187, Sbay: 0.0545286): 0.1984531, Cgla: 0.3382801): 0.042494, Scas: 0.2796456): 0.0506147, Kpol: 0.3061641): 0.0374508, Zrou: 0.2684242): 0.0489862, ((Klac: 0.3075333, Agos: 0.3326945): 0.0600818, ((Kwal: 0.1277869, Kthe: 0.1223815): 0.1697185, Sklu: 0.1779484): 0.0492655): 0.0622827);

Legends to Tables S1-S4

Table S1. Predictions on the yeast proteome by the phylo-HMM approach. Short conserved sequences were predicted on the yeast proteome using our phylo-HMM approach. Each motif is represented by a unique identifier, as well as the gene that contains it (systematic name), its position and sequence within the *S. cerevisiae* protein.

Table S2. Literature-curated characterized short linear motifs. A set of short linear motifs in yeast proteins was curated from the literature. Each short linear motif is defined by the gene that contains it (standard name and systematic name), the reported function of the motif within the gene, the regulator that interacts with the motif, its position within the protein and the sequence of amino acids reported. The reference for this information is included for each short linear motif. The following columns indicate whether the motif is found in unstructured regions (by our definition) and whether the motif was predicted by the phylo-HMM. If the motif could be defined by a consensus sequence and was found in unstructured regions, the fraction of species that contained this consensus sequence is shown. For short linear motifs that matched a known consensus, if this fraction was above 90%, we considered the consensus sequence to be conserved. For short linear motifs that did not match a known consensus, in some cases we inspected the alignment by eye and indicated these as 'considered conserved' if we could identify them in many species.

Table S3. Enrichment analysis of motifs matching known consensus sequences. Three tables describe the functional enrichment analysis of proteins (labeled by their systematic name) with motifs matching three known consensus sequences. The first table describes whether proteins known to be phosphorylated and known not to be phosphorylated by the cyclin-dependent kinase Cdc28 contain a phylo-HMM prediction with a conserved phosphorylation site consensus sequence ([ST]Px[RK]). The second table lists the proteins with an FG-dipeptide sequence, whether they are FG-repeat containing nucleoporins, whether they are involved in non-nuclear protein transport or sorting, and whether they contain a phylo-HMM prediction with a conserved FG-dipeptide sequence. The last table indicates the proteins that contain a predicted KEN-box consensus sequence, and the experimental line of evidence that describes APC/C-mediated degradation.

Table S4. Clusters of similar short conserved sequences. Three tables show the results from three different cluster analyses that used different metrics to assign sequence similarity between motifs. The first table describes the results using an all-by-all pairwise distance measure (described in Methods) between the predicted segments. The second table describes the results after first extending either side of the motifs by five residues before applying the all-by-all distance measure. Using the same extended motifs, the third table describes the results considering only the top 10 most similar motifs as similar. Each motif is represented by its identifier, the sequence that was used in the cluster analysis, the gene that contains the motif, its position within the *S. cerevisiae* protein, and the cluster where it is found.

Table S5. Annotation of top 20 clusters of predicted from different distance metrics of predicted short conserved sequences. Three tables show annotations of the top 20 clusters of each three cluster analyses that used three different metrics to assign sequence similarity between motifs (see Table S4). Each set of annotation includes a sequence logo representing the pattern formed by the motifs, whether this pattern is known and notes regarding functional enrichments. The first table describes the results using an all-by-all pairwise distance measure (described in Methods) between the predicted segments. The second table describes the results after first extending either side of the motifs by five residues before applying the all-by-all distance measure. Using the same extended motifs, the third table describes the results considering only the top 10 most similar motifs as similar.

Table S5-1. Patterns obtained from clustering of pairwise sequence distance between conserved sequences. Motif profiles were aligned using the average pairwise substitution score derived from the empirical substitution matrix described in Methods. Pairs of profiles (m_i and m_j) were then clustered using the Smith-Waterman score ($S[m_i,m_j]$) and divided by the square root of the aligned region length ($l_{i,j}$). Edges were pruned if the score exceeded a threshold of $\min(7.7, S[m_i,m_i]/\mathrm{sqrt}(l_{i,i}))$.

Cluster number	Known	Motif	Notes
1	Yes		Enriched in cell-cycle proteins Probably proline-directed phosphorylation sites
2	Yes		Enriched in nuclear proteins Probably related to nuclear localization signals
3	Yes		Enriched in nucleoporins GLFG-repeat motif
4	Yes		Enriched in cell-cycle proteins Probably proline-directed phosphorylation sites
5	Yes		Proline-rich motif
6	Yes		Basophilic-kinase phosphorylation sites
7	No		
7	No		
8	No		
9	Yes		Probably related to nuclear localization signals
10	Yes		Enriched in endocytosis genes EH-interacting motif

10	Yes		Proline-rich motif
11	Yes		Similar to Ime2 phosphorylation sites
12	Yes		Probably related to nuclear localization signals
13	No		
14	No		
15	Yes		Probably related to nuclear localization signals
16	No		
17	Yes		Proline-rich motif
17	Yes		Probably related to proline-directed phosphorylation sites
18	Yes		Glutamine-repeat
19	No		
19	No		
20	No		

Table S5-2. Patterns obtained from clustering of pairwise sequence distance between conserved sequences. Each predicted conserved sequences were first extended on both side and trimmed (See Methods). Motif profiles were aligned and scored as in the previous section of this table.

Cluster number	Known	Motif	Notes
1	Yes		Enriched in nucleoporins GLFG-repeat motif
2	Yes		Proline-rich motif
3	Yes		Enriched in nuclear localization Probably related to nuclear localization signal
4	Yes		Enriched in endocytosis genes EH-interacting motif
5	Yes		Enriched in cell-cycle proteins Probably proline-directed phosphorylation sites
6	Yes		Enriched in nucleoporins FxFG-repeat motif
7	Yes		Enriched in cell-cycle proteins Probably proline-directed phosphorylation sites
7	No		
8	No		Enriched in vesicle and nuclear membrane proteins, enriched in protein transport process. Probably related to NPF motif
9	No		
10	Yes		Probably basophilic-kinase phosphorylation sites
10	Yes		Probably related to nuclear localization signal
11	Yes		Probably related to nuclear localization signals
12	Yes		Probably related to nuclear localization signals

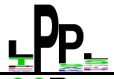
12	Yes		Probably related to nuclear localization signals
12	Yes		Basophilic-kinase phosphorylation sites
12	Yes		Probably proline-directed phosphorylation sites
13	Yes		Probably acidophilic-kinase phosphorylation sites
13	No		
14	Yes		Enriched in ER localization ER-localization signal
15	Yes		KEN-box APC/C degradation signal
16	Yes		Proline-rich motif
16	Yes		Probably proline-directed phosphorylation sites
17	No		Enriched in amino acid permeases
17	No		
17	No		
18	No		
19	Yes		Probably proline-directed phosphorylation sites
20	Yes		Enriched in mitochondrial localization Mitochondrial targeting signal

Table S5-3. Patterns obtained from clustering of conserved sequences and their top ten most similar conserved sequences, without allowing matches to paralogs or to the same protein. Each predicted conserved sequences were first extended on both side and trimmed (See Methods). Motif profiles were aligned using the average pairwise substitution score derived from the empirical substitution matrix described in Methods. These profiles were then clustered using the Smith-Waterman score ($S[m_i,m_j]$). The top ten most similar profiles were taken.

Cluster number	Known	Motif	Notes
1	Yes		Resembles motif in vacuolar proteins in yeast
2	Yes		Proline-rich motif
3	Yes		Enriched in ER localization ER-localization signal
4	Yes		Proline-rich motif of class II (PxxPx+)
5	Yes		Enriched in mitochondrial localization Mitochondrial targeting signal
6	Yes		Probably related to nuclear localization signal
6	Yes		PCNA-interacting motif
6	No		
7	No		
8	Yes		Probably basophilic-kinase phosphorylation sites
9	No		Enriched in Cbk1 interactors
10	Yes		Cbk1 phosphorylation motif
11	No		
12	Yes		Disulfide isomerase motif
13	Yes		Probably related to nuclear localization signal
14	Yes		eIF4e binding site
15	Yes		EH-interacting motif
16	No		N-terminal motif

17	No	No particular motif	
18	No		
19	No		
19	Yes		Disulfide isomerase motif
19	No		
20	Yes		FxFG-repeat motif