Electronic Supplementary Material (ESI) for Molecular BioSystems

The Origins and Evolution of Ubiquitination Sites

Tzachi Hagai, Ágnes Tóth-Petróczy, Ariel Azia and Yaakov Levy

Supporting Information

Figure S1: Phylogenetic trees and list of species used for rate calculation and ancestral reconstruction

For computation of the evolutionary rates, and the ancestral reconstructions we used orthologs from the following lists: Vertebrates (19 species): Tetraodon (T.nigroviridis), Fugu (T.rubripes), Zebrafish (D.rerio), Frog (X.tropicalis) ,Chicken (G.gallus) , Platypus (O.anatinus), Opossum (M.domestica), Macaque (M.mulatta), Orangutan (P.pygmaeus), Human (H.sapiens), Chimp (P.troglodytes), Guinea Pig (C.porcellus), Rat (R.norvegicus), Mouse (M.musculus), Cow (B.taurus), Dog (C.familiaris), Horse (E.caballus), (O.latipes), Stickleback (G.aculeatus); Yeast (16 species): S.cerevisiae, S.paradoxus, S.mikatae, S.bayanus, C.glabrata, S.castellii, K.waltii, K.lactis, A.gossypii, D.hansenii, C.albicans, Y.lipolytica, M.grisea, N.crassa, F.graminearum, S.pombe.

Trees used are shown below:

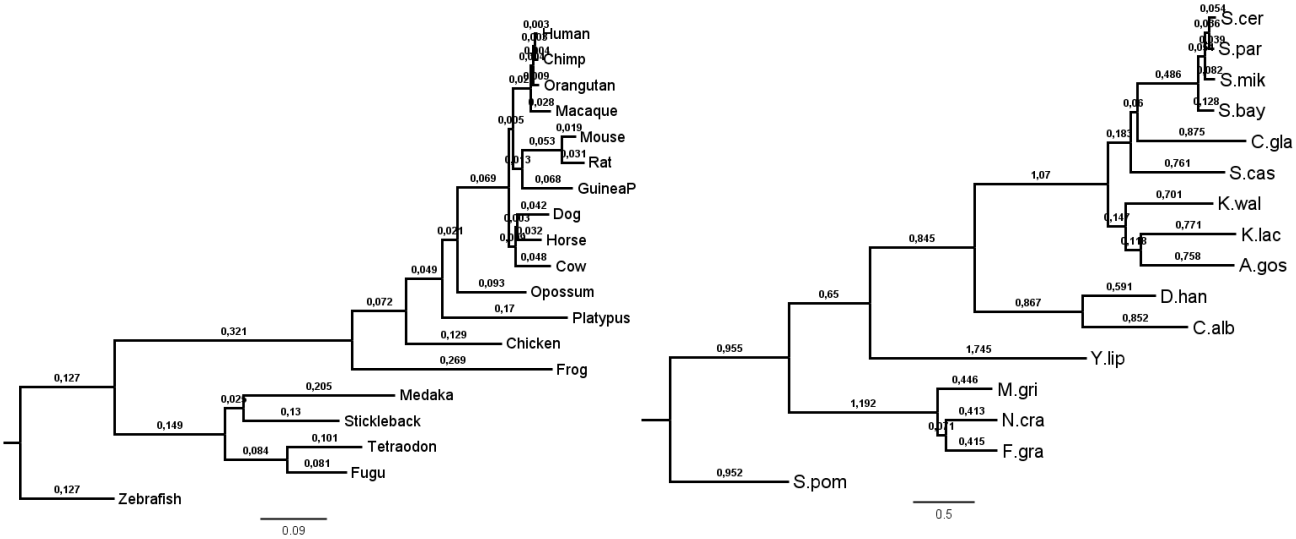


Table S1: Details of modification sites and proteins used

Phylogeny	Number of proteins	Number of modification sites
Mammals	416	452 ubiquitination sites 366 SUMOyltion sites 327 acetylation sites 1510 phospho-serine sites 425 phospho-threonine sites 188 phospho-tyrosine sites
Yeast	210	301 ubiquitination sites 32 SUMOyltion sites 8 acetylation sites * 564 phospho-serine sites 172 phospho-threonine sites 9 phospho-tyrosine sites *

* Due to the relatively low numbers of acetylation and phosphor-tyrosine sites that were extracted in yeast, these sites were not included in the analysis. In addition, any ambiguous lysine residue was discarded from all analyses.

** In both yeast and mammals, methylation sites were not included in the analysis due to their low numbers.

Table S2: a detailed list of all the ubiquitination and SUMOylation substrates

Attached as a separate excel file.

Table S3: Details of evolutionary analysis of ubiquitinated proteins in mammalian set

The distributions of evolutionary rates of ubiquitination, SUMOylation, acetylation, and phosphorylation sites and their analogous unmodified residues (lysine, serine, threonine and tyrosine) were compared to test for significant conservation. The unmodified residue value was calculated using the median of the unmodified analogous residues in the respective proteins (e.g. in the case of ordered ubiquitination sites, for each site the respective median of the unmodified lysine residues in ordered regions was taken).

Ubiquitination sites are significantly more conserved than their unmodified counterparts (P = 6.4E-4 for ordered regions, P = 0.047 for disordered regions, Kolmogorov-Smirnov test). These sites were further divided into two groups in two different analyses: (1) Fast and slow evolving proteins: In the fast-evolving set, sites are significantly more conserved than their unmodified residues (P = 1.4E-6 for the ordered set and P = 2.1E-3 for the disordered set). (2) Functional mutagenesis and MS-identified ubiquitination sites: In the functional-mutations set, ubiquitination sites are significantly more conserved than their unmodified counterparts (P = 2.5E-8 for the ordered set and P = 4.5E-4 for the disordered set). SUMOylation and acetylation sites were not found to be significantly more conserved (either the unmodified set had lower evolutionary rates or the P-value was higher than 0.05). Phosphorylation sites showed similar results except of Ser and Thr phosphosites in disordered regions (P = 2.2E-7 and 6.4E-4 respectively).

Site type	**Ordered regions: p-value**	**Disordered regions: p-value**
ub sites (entire set)	6.4E-4	P = 0.047
ub sites (functional mutagenesis set)	2.5E-8	4.5E-4
ub sites (fast evoving set)	P = 1.4E-6	P = 2.1E-3
SUMO sites	0.1647	0.0475

Table S4: Details of prokaryote species used in analysis and orthologs annotation process

1.	E. coli
2.	B.burgdorferi.DSM.4680
3.	P.furiosus
4.	P.aerophilum
5.	B.subtilis
6.	V.cholerae.ATCC.39315
7.	T.maritima
8.	T.acidophilum
9.	A.fulgidus
10.	N.meningitidis.A
11.	A.aeolicus
12.	C.muridarum
13.	C.tepidum
14.	M.mazei
15.	S.coelicolor
16.	F.nucleatum.nucleatum
17.	R.conorii
18.	N.equitans
19.	G.sulfurreducens
20.	S.pneumoniae.TIGR4
21.	S.solfataricus.DSM.1617
22.	P.gingivalis.W83
23.	H.pylori.51
24.	S.usitatus
25.	M.tuberculosis.F11
26.	D.ethenogenes
27.	R.baltica
28.	T.thermophilus.HB27
29.	K.cryptofilum
30.	H.salinarum
31.	P.marinus.AS9601
32.	A.pernix

Genomes were downloaded from the EBI site (http://www.ebi.ac.uk/integ8/).

Orthologs were found by running the InParanoid program[1], on pairs of genomes (each pair containing one eukaryote species and one prokaryote species). The eukaryote species used in this analysis were: H.sapiens, B.taurus, M.musculus, R. norvegicus, G. gallus and S.cerevisiae. In order to account for the large separation

between eukaryote and prokaryote species, the default parameters of InParanoid were changed as follows: the matrix used was BLOSUM45 (instead of BLOSUM62), the cutoff score (score_cutoff parameter) was reduced from 40 to 35, and the overall sequence overlap cutoff (seq_overlap_cutoff parameter) was reduced from 0.5 to 0.4.

Table S5: Details on evolutionary analysis of ubiquitinated proteins in a longer evolutionary timescale

In order to analyze the evolution of modification sites in longer evolutionary timescales, we extracted a subset of the mammalian and the yeast proteins which have a high number of orthologous across the entire eukaryotic kingdom (at least 80 species) – 90 proteins in total. We repeated our analysis regarding the evolutionary rates of their ubiquitination sites using all their orthologs, and found it again similar to the unmodified lysine residues (p-values>>0.05 for both ordered and disordered sets, Kolmogorov-Smirnov test). Since most of these proteins are highly conserved – appearing in most of the eukaryote species, the evolutionary rates of most of their residues is very slow (they are slower than the yeast or mammalian set by two-three orders of magnitude, and the set's median is 0.006-0.0084), which results in almost immeasurable differences in evolutionary rates.

Site type	**Number of sites**	**p-value (Kolmogorov-Smirnov test)**
Ordered ub sites	89	0.202
Disordered ub sites	15	0.308

Figure S2: Flow of work

Stages of research (left column) and the methods / programs / databases used in each stage (right column):

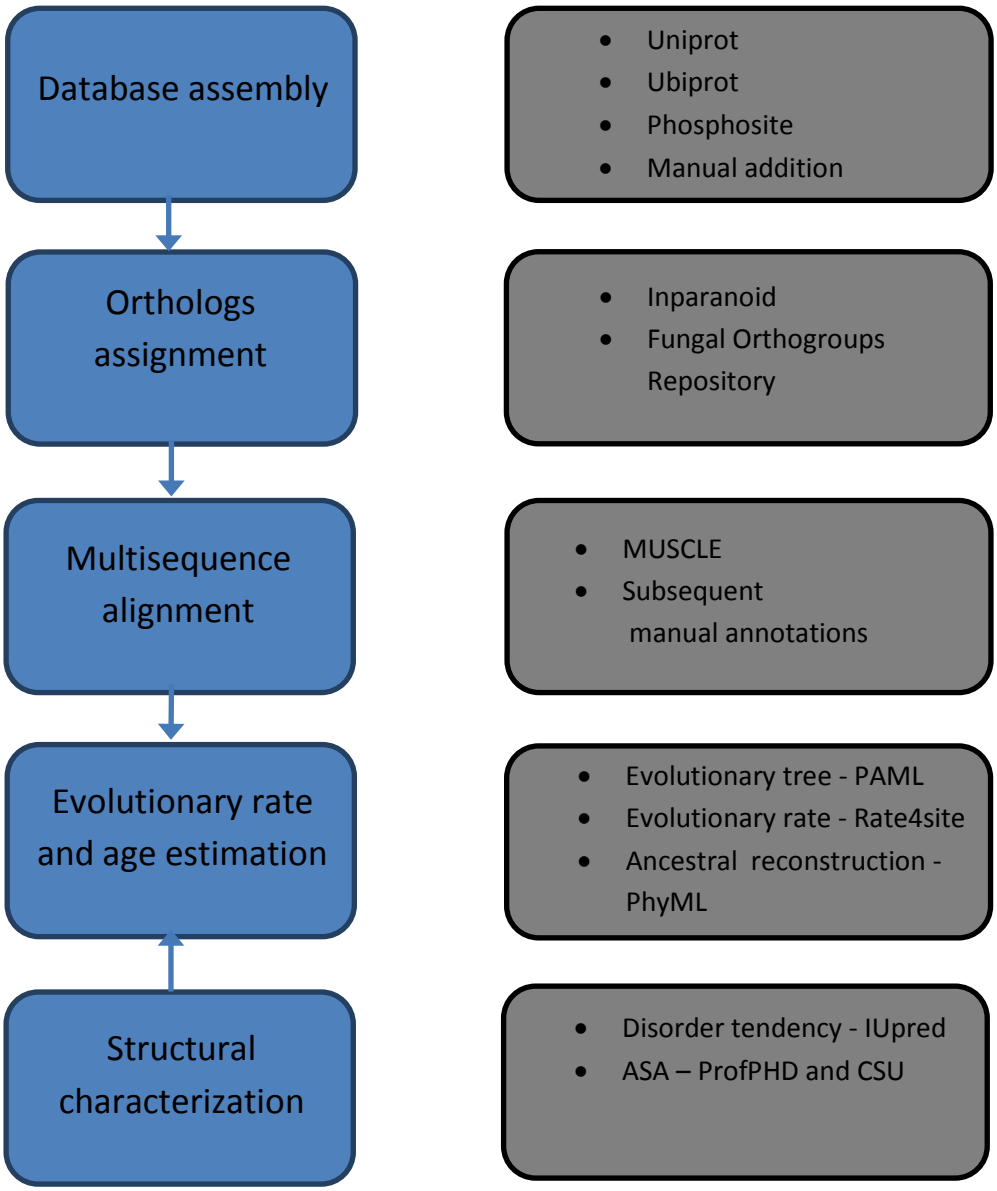


Figure S3: evolutionary rates of yeast modification sites

Mean evolutionary rates of ubiquitination, SUMOylation, and phosphorylation sites and their analog unmodified residues (lysine, serine, threonine) in ordered regions (A) and disordered regions (B) of yeast proteins. Rates were inferred from rate4site[2], using the same tree for all the proteins, and indicate how fast residues evolve in comparison to the tree (the higher the value – the faster they evolve). Numbers of sites are indicated above each column. The unmodified residue value was calculated using the median of the unmodified analogous residues in the respective proteins. In both ordered regions(A) and disordered regions (B), ubiquitination and SUMOylation sites were found to have rates which are not significantly more conserved than unmodified lysine residues (rates of modified lysine residues are higher on average than the unmodified counterparts, and p-values are higher than 0.05, Kolmogorov-Smirnov test). Phosphorylation sites in ordered regions have on average higher rates than unmodified residues (p value = 2.9E-9 and 0.01 for phosphor-Ser and phosphor-Thr sites, Kolmogorov-Smirnov test). In disordered regions phosphor-Ser sites are significantly more conserved (p value = 1.1E-7, Kolmogorov-Smirnov test), but phosphor-Thr sites are not (p value = 0.2, Kolmogorov-Smirnov test).

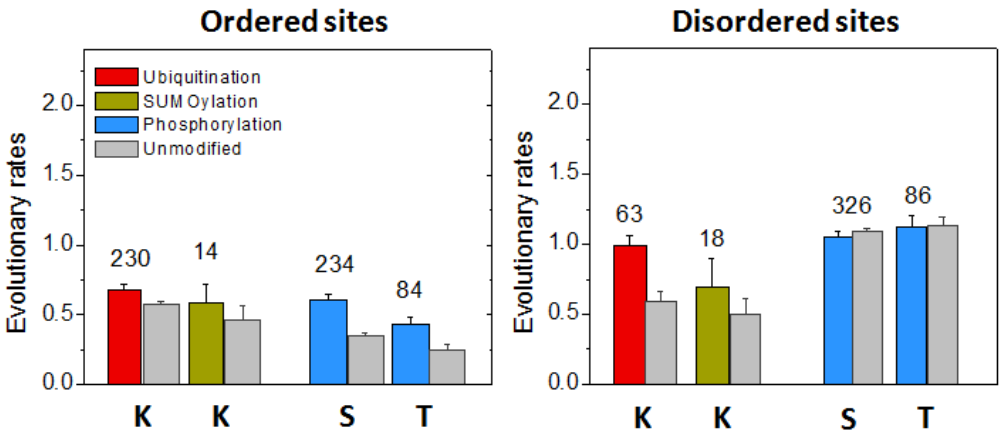


Figure S4: age distribution of ubiquitinated and unmodified Lys residues in yeast

The inferred age of ordered and disordered ubiquitination sites and analogous unmodified Lys residues. The age was inferred by ancestral reconstruction, using the tree and the given multi-sequence alignment. Each lysine was assigned an age based on the most ancient node at which it was predicted to exist using the PAML program[3]. Numbers of nodes are indicated on the tree (to the left), and the percentages of lysine residues in each node are shown to the right.

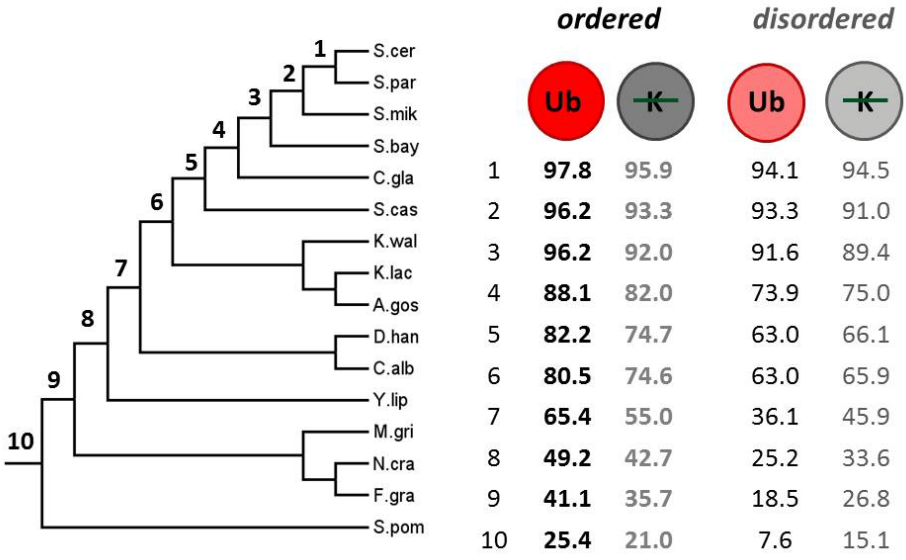


Figure S5: evolutionary rates of ubiquitinated mammalian substrates separated by their mean evolutionary rates and by their method of characterization

Mean evolutionary rates of ubiquitination sites and their analog unmodified lysine residues in ordered and disordered region of mammalian proteins separated by their mean average rate – slow-evolving and faster evolving proteins and by their method of characterization – highthroughput MS or functional mutagenesis. Numbers of sites are indicated above each column. The unmodified residue value was calculated using the median of the unmodified analogous residues in the respective proteins. In the slow-evolving set (average rate $\leq$ 0.5), ubiquitination sites have higher rates than unmodified lysine residues (p value = 0.17 and 0.008 for ordered and disordered regions, respectively), in the faster-evolving set the trend is opposite and the ubiquitination sites are significantly more conserved than their unmodified counterparts (p value = 1.4E-6 and 0.002 for ordered and disordered regions, respectively). In the functional-mutations set ubiquitination sites are significantly more conserved than their unmodified counterparts (p-value=2.5E-8 for the ordered set and p-value=4.5E-4 for the disordered set). In the MS-based set ubiquitination sites are as conserved as the unmodified lysine set (p-value=0.52 for the ordered set and p-value=0.75 for the disordered set).

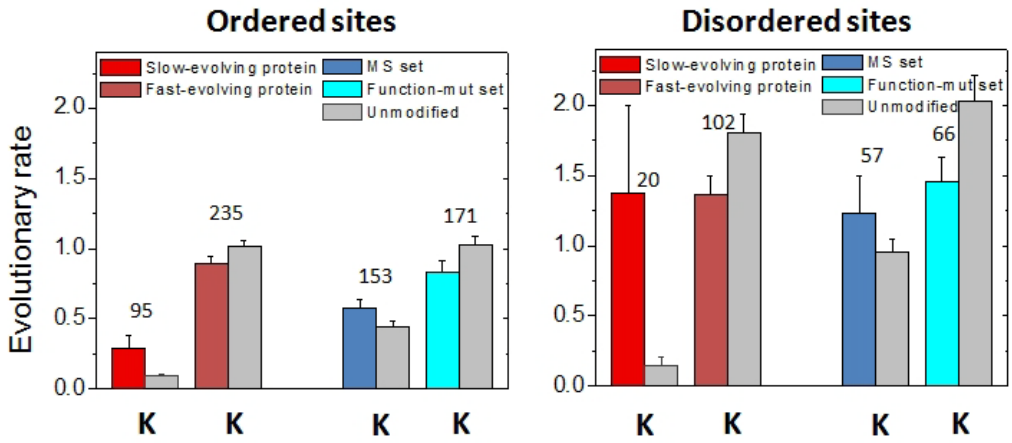
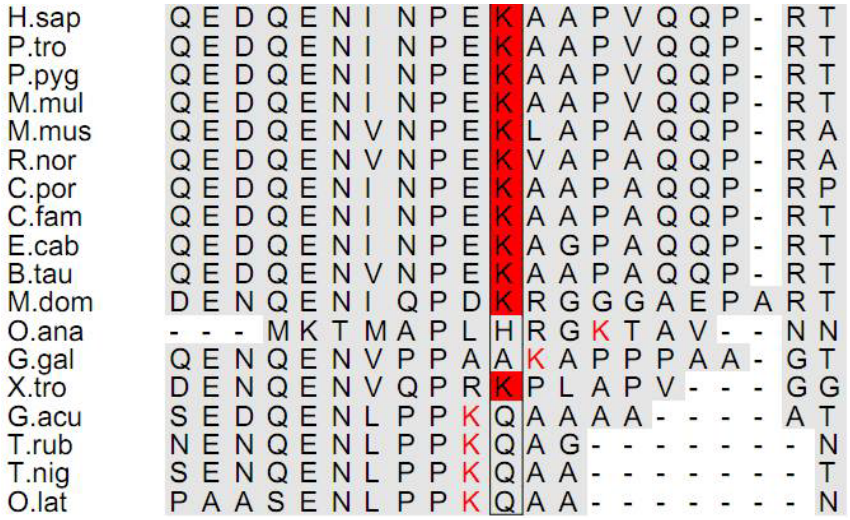
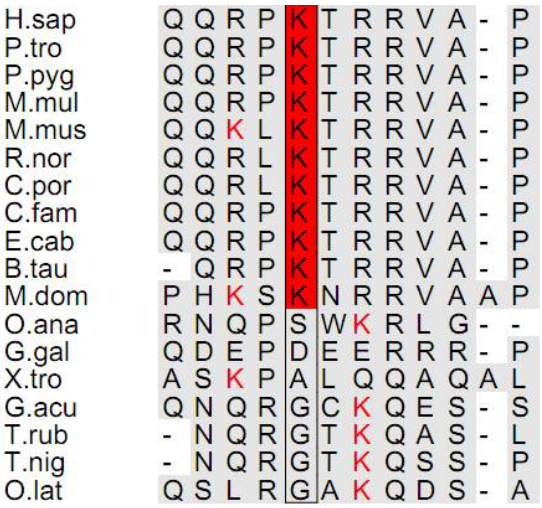


Figure S6: Putative cases of shifting sites

A. Evolution of ubiquitination of Cyclin A2 on K37



B. Evolution of ubiquitination of Cyclin A2 on K68



C. Evolution of ubiquitination of p21 on K154,k161 and K163

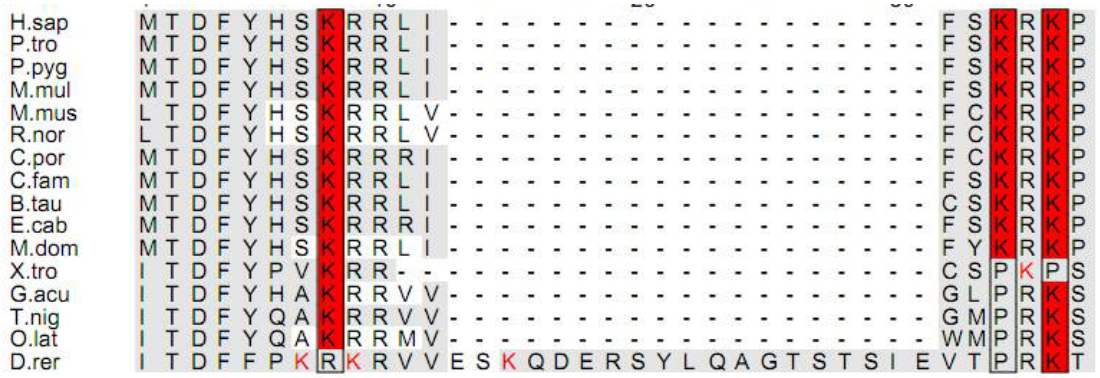
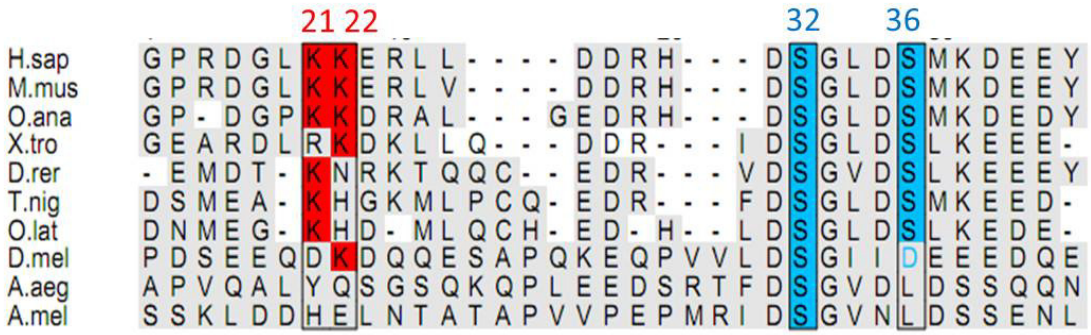


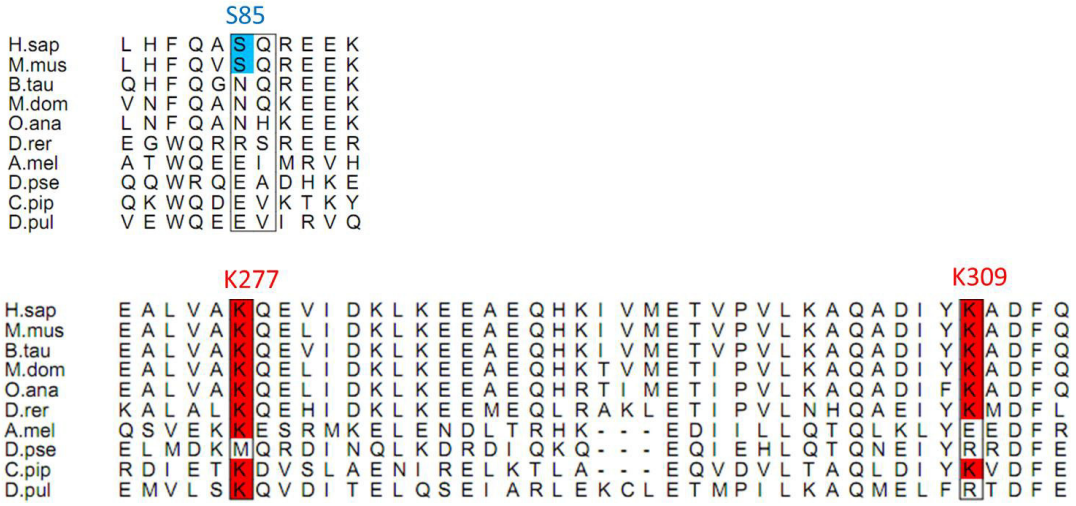
Figure S7: Additional examples of evolution of crosstalk in PTMs

A. The evolution of phosphorylation-mediated ubiquitination of Iκbα



Phosphorylation events on Ser32 and Ser 36 lead to ubiquitination on either Lys21 or Lys22 which mediates proteasomal degradation.

B. The evolution of phosphorylation, SUMOylation and ubiquitination crosstalk of NEMO



SUMOyaltion of NEMO on Lys277 or Lys309 leads to phosphorylation by ATM on Ser85, this in turn leads to the replacement of SUMO1 by mono-ubiquitin on the same residues.

Figure S8: evolution of phosphorylation at Ser38 of Activation-induced cytidine deaminase

Human	E	T	Y	L	C	Y	V	V	K	R	R	D	S	A	T	S	F	S	L	D	F	G	Y	L
Chimp	E	T	Y	L	C	Y	V	V	K	R	R	D	S	A	T	S	F	S	L	D	F	G	Y	L
Orangutan	E	T	Y	L	C	Y	V	V	K	R	R	D	S	A	T	S	F	S	L	D	F	G	H	L
Macaque	E	T	Y	L	C	Y	V	V	K	R	R	D	S	A	T	S	F	S	L	D	F	G	H	L
Mouse	E	T	Y	L	C	Y	V	V	K	R	R	D	S	A	T	S	C	S	L	D	F	G	H	L
Rat	E	T	Y	L	C	Y	V	V	K	R	R	D	S	A	T	S	F	S	L	D	F	G	H	L
GuineaPig	E	T	Y	L	C	Y	V	V	K	R	R	D	S	A	T	S	C	S	L	D	F	G	H	L
Dog	E	T	Y	L	C	Y	V	V	K	R	R	D	S	A	T	S	F	S	L	D	F	G	H	L
Horse	E	T	Y	L	C	Y	V	V	K	R	R	D	S	A	T	S	F	S	L	D	F	G	H	L
Cow	E	T	Y	L	C	Y	V	V	K	R	R	D	S	P	T	S	F	S	L	D	F	G	H	L
Opossum	E	T	Y	L	C	Y	V	V	K	R	R	D	S	A	T	S	F	S	L	D	F	G	Y	L
Platypus	E	T	Y	L	C	Y	V	V	K	R	R	D	S	A	T	S	F	S	L	D	F	G	H	L
Chicken	E	T	Y	L	C	Y	V	V	K	R	R	D	S	A	T	S	C	S	L	D	F	G	Y	L
Frog	E	T	Y	L	C	Y	I	V	K	R	R	Y	S	S	V	S	C	A	L	D	F	G	Y	L
Stickleback	E	T	Y	L	C	F	V	V	K	R	R	V	G	P	D	S	L	S	F	D	F	G	H	L
Fugu	E	T	Y	L	C	F	V	V	K	R	R	V	G	P	D	T	L	T	F	D	F	G	H	L
Tetraodon	E	T	Y	L	C	F	V	V	K	R	R	V	G	P	D	T	L	T	F	D	F	G	H	L
Medaka	E	T	Y	L	C	F	V	V	K	R	R	V	G	P	E	S	L	S	F	D	F	G	H	L
Zebrafish	E	T	Y	L	C	F	V	V	K	R	R	I	G	P	D	S	L	S	F	D	F	G	H	L

Phosphorylation of Ser38, which was shown to be functionally important, is absent in fish species, but a negatively-charged residue nearby was shown to replace it[4].

References:

1. Berglund AC, Sjolund E, Ostlund G, Sonnhammer EL (2008) InParanoid 6: eukaryotic ortholog clusters with inparalogs. Nucleic Acids Res 36: D263-266.
2. Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 18 Suppl 1: S71-77.
3. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586-1591.
4. Basu U, Wang Y, Alt FW (2008) Evolution of phosphorylation-dependent regulation of activation-induced cytidine deaminase. Mol Cell 32: 285-291.