SUPPLEMENTARY FILE

Protein intrinsic disorder negatively associates with gene age in different eukaryotic lineages

Sanghita Banerjee[1, 2,*], Sandip Chakraborty[2]

[1]Machine Intelligence Unit, Indian Statistical Institute, 203 Barrackpore Trunk Road, Kolkata 700108, India.

[2]Biology Department, University of Nevada, Reno, Nevada, United States of America.

[*] To whom correspondence should be addressed.

Emails:

Sanghita Banerjee: banerjee.sanghita@gmail.com
Sandip Chakraborty: sandipc@unr.edu

SUPPLEMENTARY TABLES AND FIGURES

Table S1. Percentage of disordered residues in proteins encoded by the classes of youngest and oldest genes

Organism	**Youngest genes**			**Oldest genes**			**Fold Change**	***P*-value**
	Mean	**Median**	***N***	**Mean**	**Median**	***N***		
H. sapiens	35.66	31.45	121	8.22	4.82	680	6.52	$1.06{\times}10^{-36}$
D. melanogaster	38.90	36.89	2526	7.33	3.88	679	9.51	$2.33{\times}10^{-17}$
C. elegans	26.07	13.34	4984	6.08	3.31	687	2.21	$1.52{\times}10^{-14}$
A. thaliana	24.59	14.47	7086	6.51	3.16	2081	4.58	$1.90{\times}10^{-23}$
S. cerevisiae	27.63	17.92	1865	6.55	3.29	627	8.40	$5.11{\times}10^{-11}$

The fold change was measured using the median values. *P*-values correspond to Mann–Whitney *U* test.

Table S2. Percentage of disordered residues in proteins encoded by the classes of young and old genes using Gene Ages dataset.

Disorder Prediction Tool	Organism	Young genes			Old genes			*P*-value
		Mean	Median	*N*	Mean	Median	*N*	
IUPred	*H. sapiens*	35.23	29.63	560	22.62	14.30	14581	1.01×10^{-9}
	D. melanogaster	30.75	22.19	633	20.64	10.19	7873	2.31×10^{-22}
	C. elegans	25.96	14.93	193	17.80	9.21	9451	0.062 (NS)
	S. cerevisiae	45.60	42.95	651	33.20	28.60	4165	1.43×10^{-58}
FoldIndex	*H. sapiens*	40.68	38.79	560	32.19	28.36	14581	1.77×10^{-6}
	D. melanogaster	40.66	39.29	633	31.88	27.12	7873	2.46×10^{-13}
	C. elegans	40.07	36.80	193	31.58	27.56	9451	4.0×10^{-6}
	S. cerevisiae	42.33	39.04	651	35.71	23.19	4165	3.11×10^{-11}

P-values correspond to Mann–Whitney *U* test. NS denotes non-significant.

Table S3. Correlation between protein connectivity and gene ages.

Organism	**ρ**	***P*-value**	***N***
Homo sapiens	0.267	1.45×10^{-27}	11641
Drosophila melanogaster	0.381	3.05×10^{-88}	6272
Caenorhabditis elegans	0.308	2.16×10^{-114}	7300
Arabidopsis thaliana	0.278	1.27×10^{-105}	14697
Saccharomyces cerevisiae	0.341	1.06×10^{-19}	5261

Spearman’s correlation coefficients (ρ) correspond to the correlation between the number of protein interactions and evolutionary gene ages. *P*-values correspond to the Spearman’s correlation test.

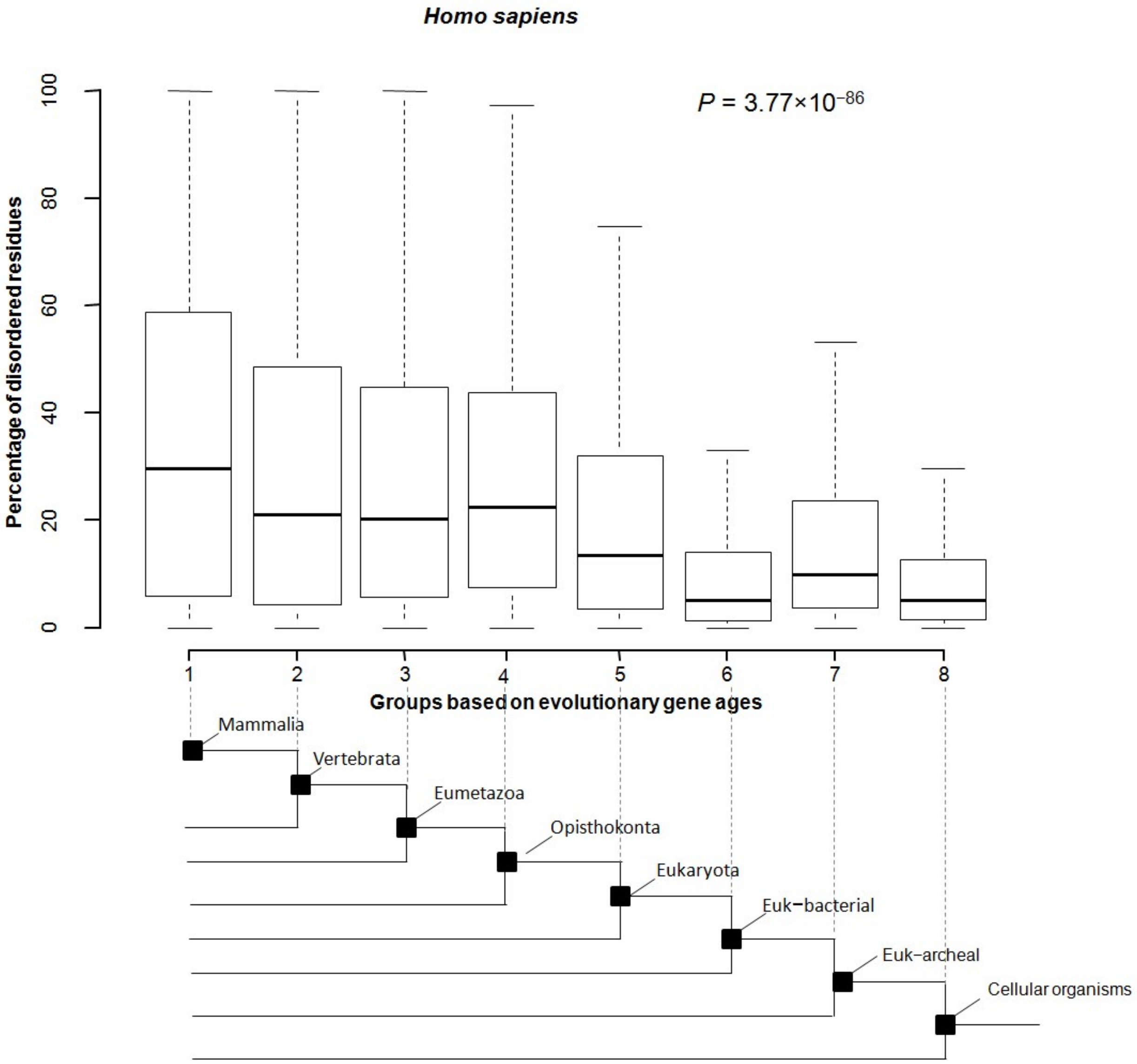

Homo sapiens
P = 3.77×10⁻⁸⁶
Percentage of disordered residues
0
20
40
60
80
100
1
2
3
4
5
6
7
8
Groups based on evolutionary gene ages
Mammalia
Vertebrata
Eumetazoa
Opisthokonta
Eukaryota
Euk–bacterial
Euk–archeal
Cellular organisms

Figure S1A

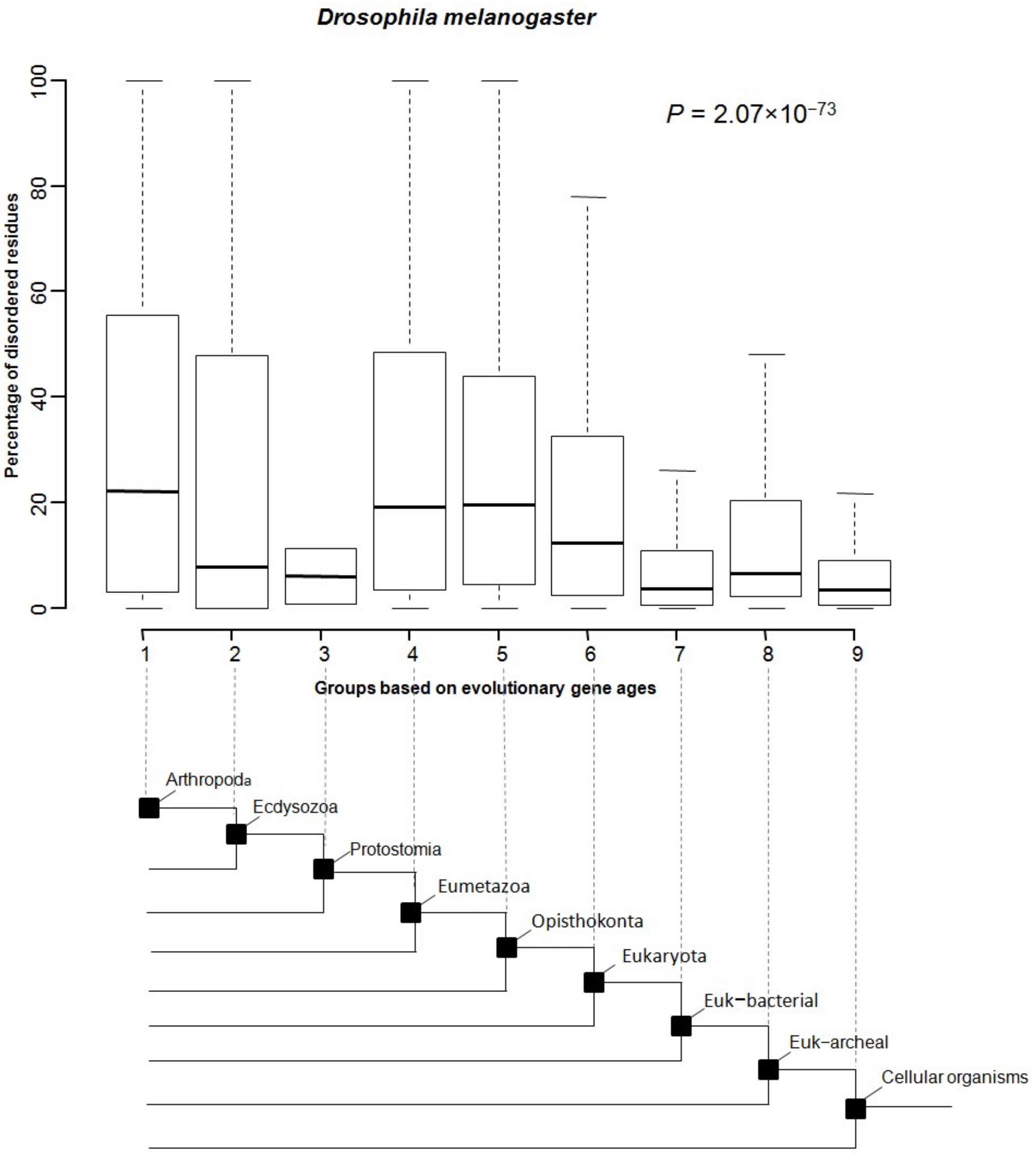
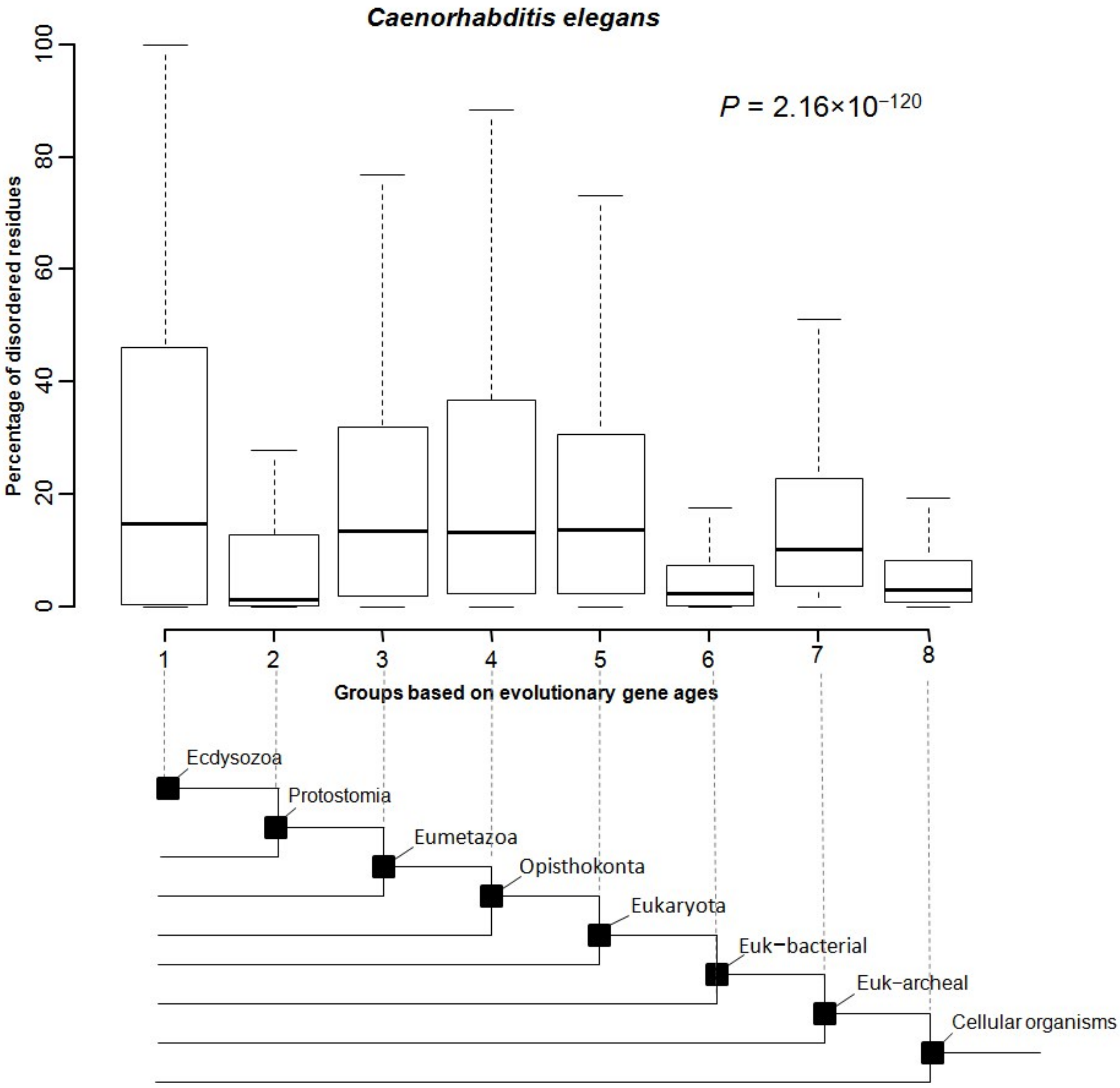

Drosophila melanogaster
$P = 2.07 \times 10^{-73}$
Percentage of disordered residues
0
20
40
60
80
100
1
2
3
4
5
6
7
8
9
Groups based on evolutionary gene ages
Arthropoda
Ecdysozoa
Protostomia
Eumetazoa
Opisthokonta
Eukaryota
Euk−bacterial
Euk−archeal
Cellular organisms

Figure S1B



Caenorhabditis elegans
P = 2.16×10−120
Percentage of disordered residues
0
20
40
60
80
100
1
2
3
4
5
6
7
8
Groups based on evolutionary gene ages
Ecdysozoa
Protostomia
Eumetazoa
Opisthokonta
Eukaryota
Euk−bacterial
Euk−archeal
Cellular organisms

Figure S1C

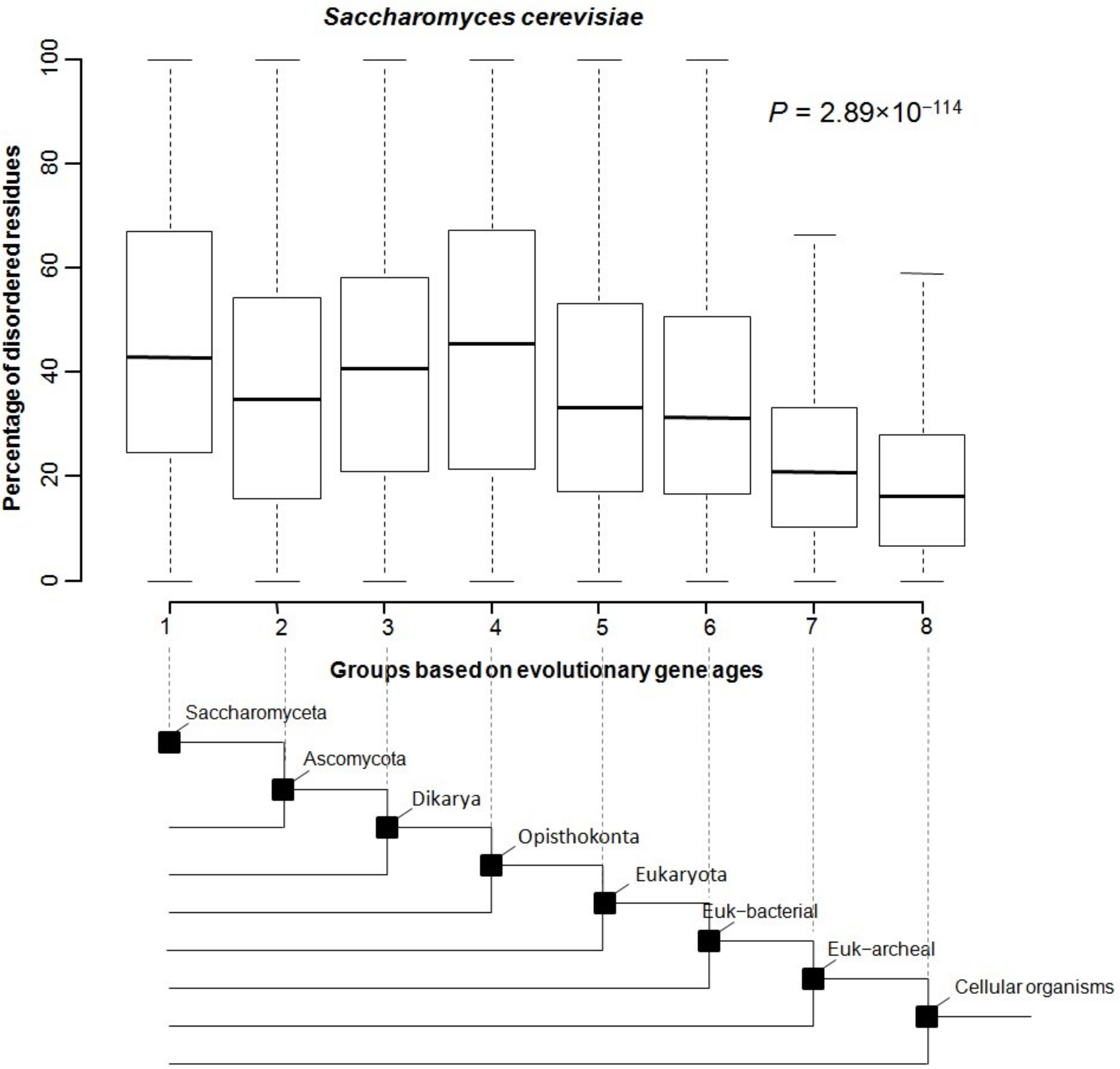


Figure S1D

Figures S1A−S1D. Distribution of intrinsic disorder across different evolutionary gene age classes using Gene Ages dataset. Gene age classes are numbered from 1 to 8, where 1 denotes the youngest or lineage-specific genes and 8 denotes the oldest gene group. The complete phylostratigraphy is shown below the boxplots with Mammalia (in Fig. S1A), Arthopoda (in Fig. S1B), Ecdyzoa (in Fig. S1C) and Saccharomyceta (in Fig S1D) as the focal lineages. The broken lines connect each gene age class to its corresponding phylostratum (shown in square box) –a node in the phylogenetic hierarchy at which the group of genes first originate. *P*-values correspond to Kruskal-Wallis test.

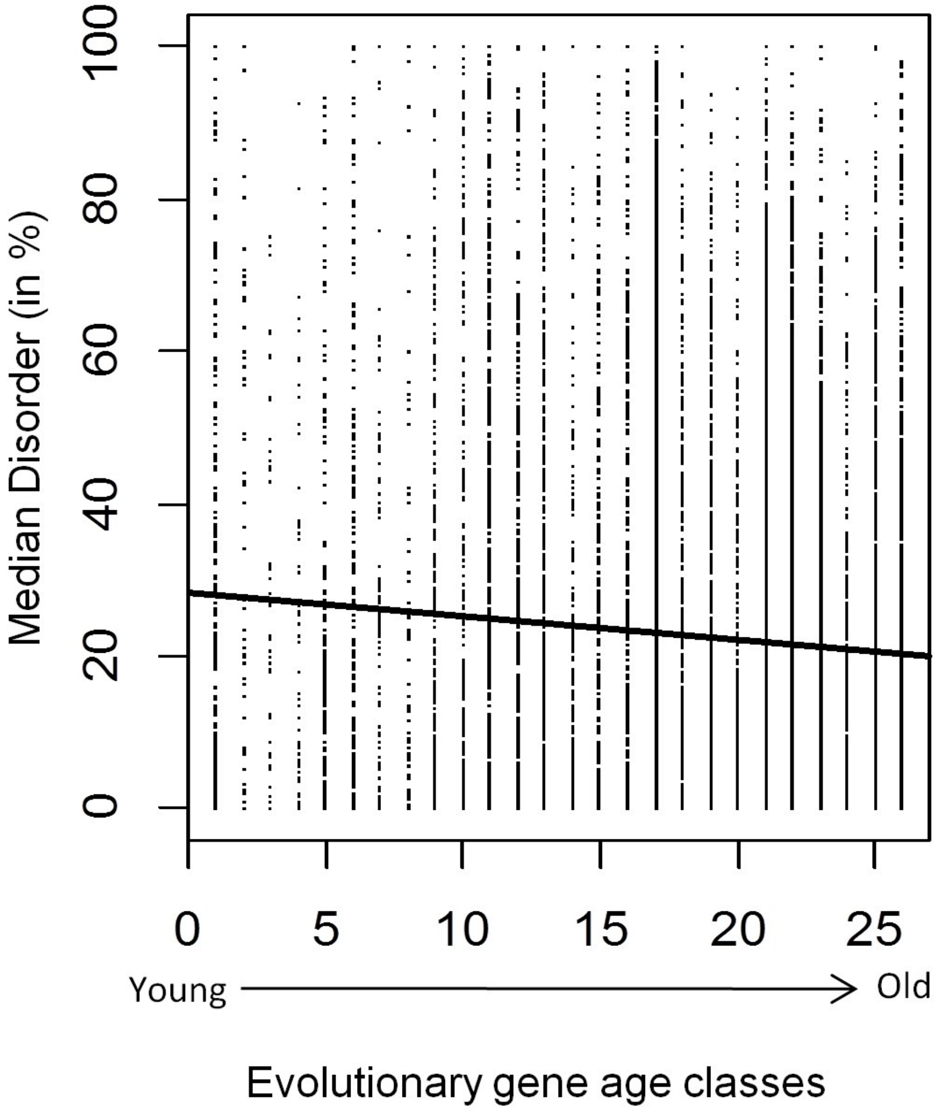


Figure S2. Changes in the intrinsic disorder with increasing evolutionary gene ages. Linear regression line has been fitted to exhibit the pattern of change in the level of intrinsic disorder across evolutionary gene age classes corresponding to the human genome. Evolutionary gene ages positively correspond with the values (0-26) denoting gene age classes, i.e. youngest genes are represented with 0.

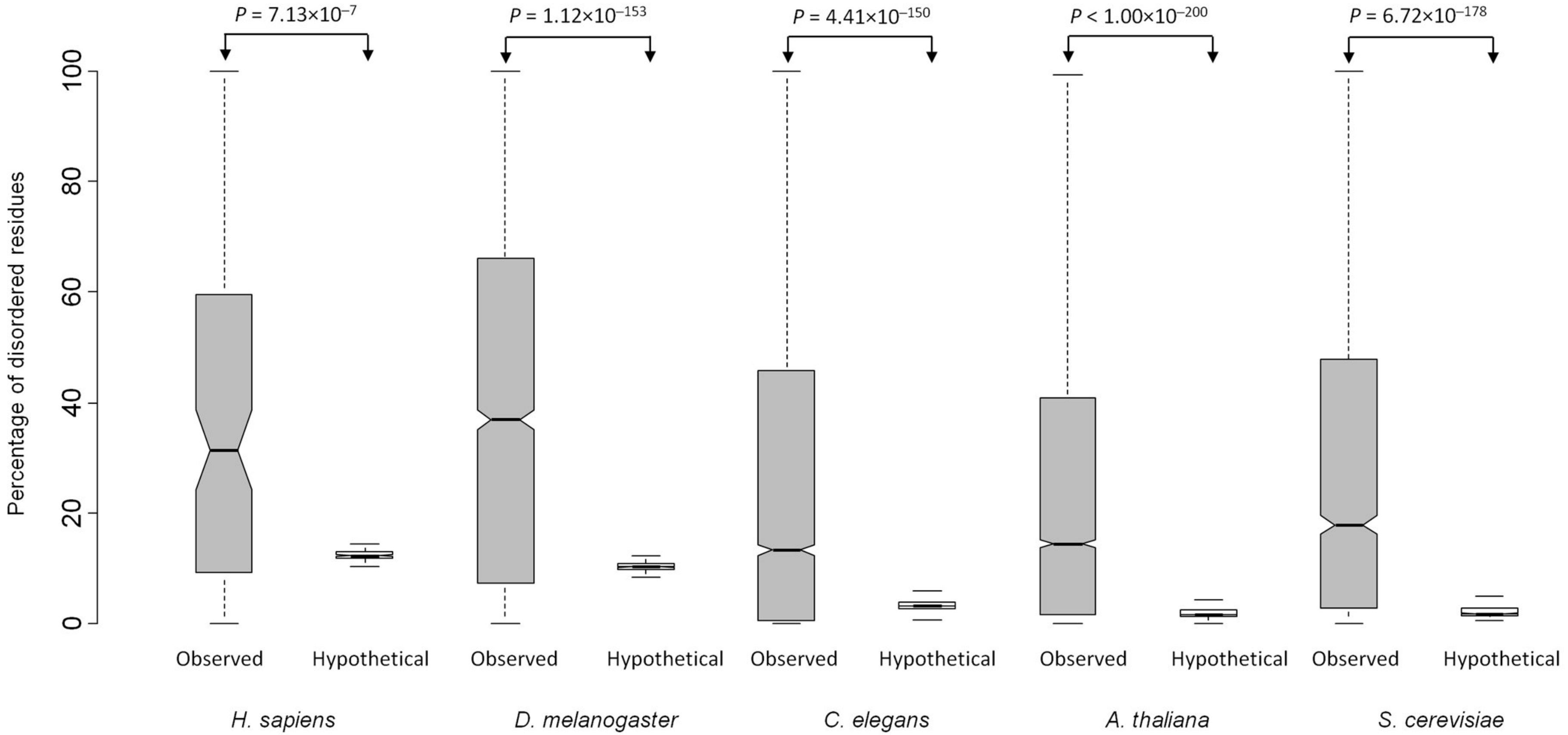


Figure S3. Distribution of the percentage of disordered residues between the set of young proteins (Observed) and the hypothetical ones. *P*-values correspond to Mann-Whitney *U* test.

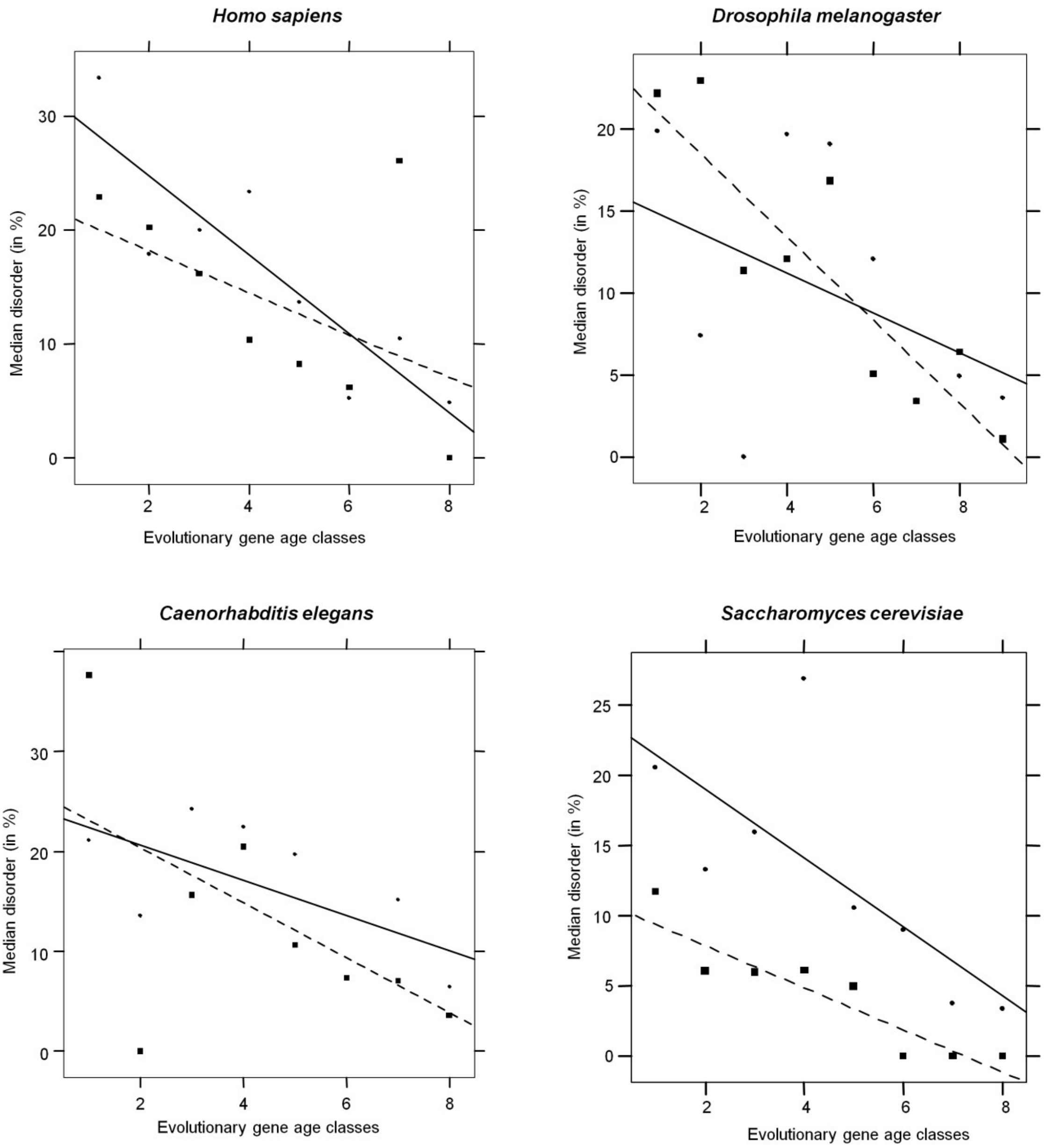


Figure S4. Change in the degree of disorder within hub (square dots) and non-hub (circular dots) proteins across different evolutionary ages using Gene Ages datasets. Median values of the percentage of disordered residues corresponding to hub and non-hub proteins have been plotted against each gene age class. Evolutionary gene ages denote the time of origin of the group of genes estimated in million years. For all the organisms, linear regression lines have been fitted for both the classes of hub (smooth line) and non-hub proteins (broken line).

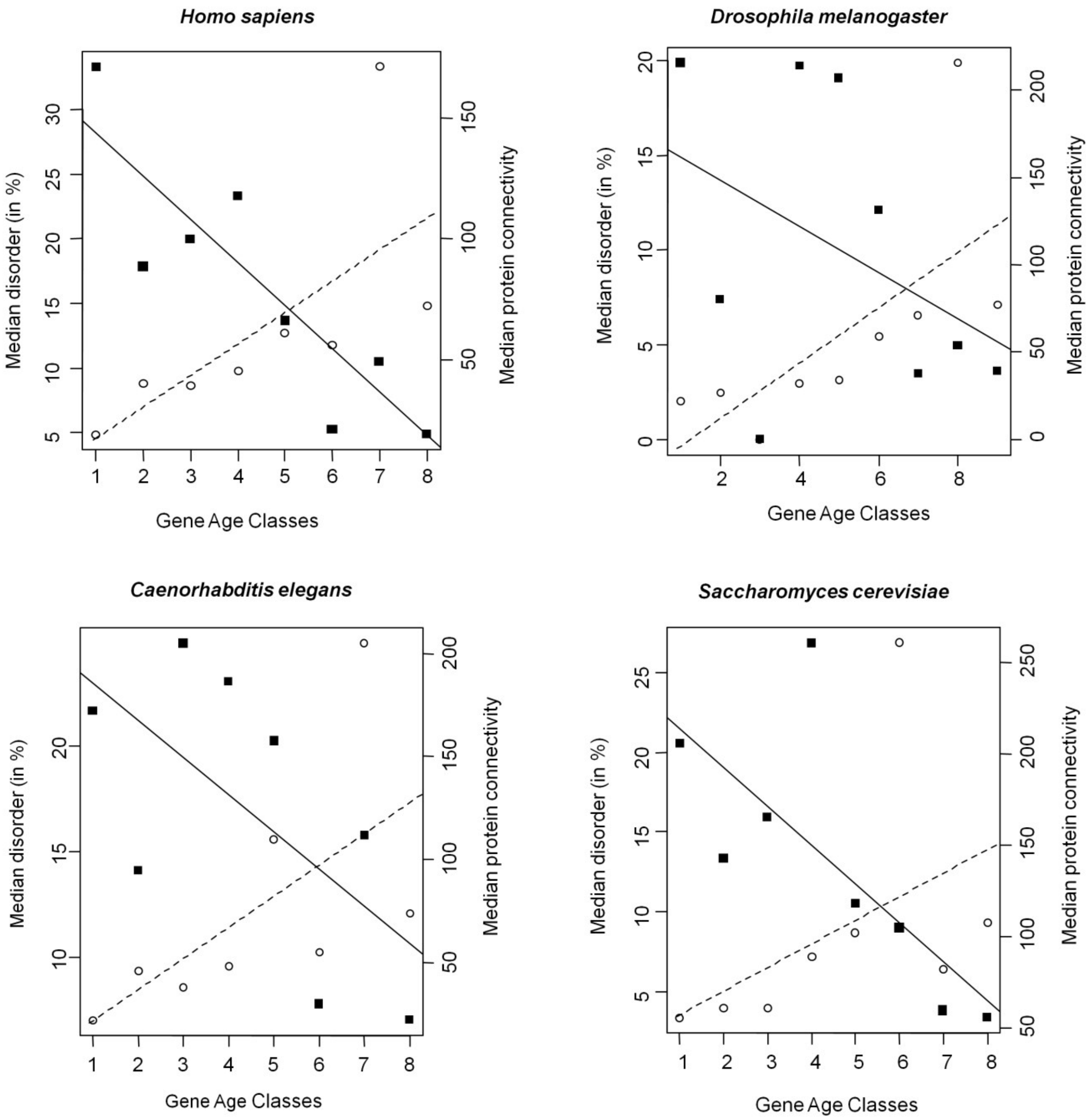


Figure S5. Association between the degree of disorder (square dots) and protein connectivity (circular dots) with evolutionary gene ages using Gene Ages datasets. Median values of the percentage of disordered residues (left y-axis) and protein connectivity (right y-axis) have been plotted against each gene age class. Evolutionary gene ages denote the time of origin of the group of genes estimated in million years. For all the organisms, linear regression lines have been fitted to represent the change in intrinsic disorder (smooth line) and protein connectivity (broken line) simultaneously, across different evolutionary ages.