# Evolutionary Rate Covariation of Domain Families

**Author: Brandon Jernigan**

**A Thesis Submitted to the Department of Chemistry and Biochemistry in Partial Fulfillment of the Bachelors of Science Degree in Biochemistry**

**Department of Chemistry and Biochemistry**
**University of Arizona**
**Spring 2017**

**Research Faculty Member _____ Date: _____**

**Biochemistry Faculty Advisor _____ Date: _____**

# Abstract

Evolutionary rate covariation (ERC) is a phylogenetic measure of the evolutionary relationship between pairs of proteins. As proteins evolve over time, their rate of evolution (dN/dS) may vary. ERC measures how closely the evolutionary rates of two proteins match over a phylogeny. Proteins known to interact directly or indirectly tend to have higher ERC, because they typically experience similar evolutionary pressures within each lineage. Much is known about ERC at the whole protein level, but little is known at the domain level. Because individual functions of a protein are often performed by distinct domains, a focus on the domain level is expected to provide a clearer relationship between specific functions and ERC. Here we investigate ERC within and between domain families. In particular, we identify domain families with high ERC and investigate potential biochemical explanations.


# Introduction

As proteins evolve over time, their rate of evolution will change depending on influences from the environment. Proteins that experience the same influences at the same times will show evolutionary rate covariation (ERC) [Clark, N. L., Alani, E., & Aquadro, C. F. (2012)]. This can be represented by comparing the phylogenetic trees of two proteins. The branch lengths represent the evolutionary rate of a protein, and each of the branches represents a time-span in which that protein evolved. Two proteins that share faster or slower than average evolutionary rates within the same branches will have higher ERC than those that do not (Figure 1).

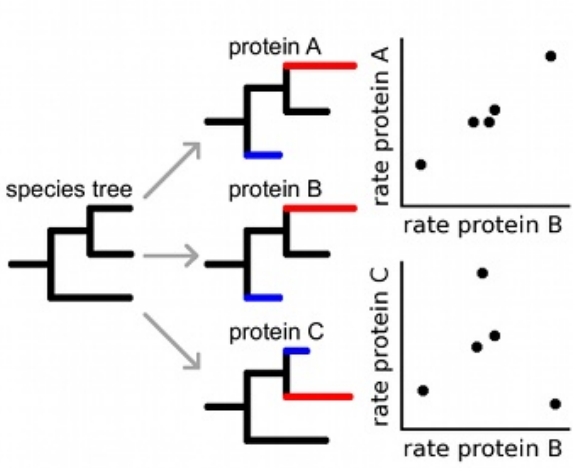


**Figure 1.** Compared to the average species tree across all proteins in the proteome, Proteins A and B have higher than average evolutionary rates (red) in their top branches. Their other branches also match, both showing average rates in the middle branch (black), and slower rates in the bottom branch (blue). This corresponds to a high ERC, shown in the scatter plot at the top right by a high correlation between their evolutionary rates. Proteins B and C show differing evolutionary rates in their corresponding branches, so have a low ERC, shown in the scatter plot in the bottom right [Figure from Clark, N. L., Alani, E., & Aquadro, C. F. (2012)].

Previous work with ERC has shown relationships between whole proteins. ERC can be used to infer which proteins are related to a disease, even when they haven't been implicated with other methods [Priedigkeit N, Wolfe N, Clark NL (2015)]. It can also be used to find new

members of functional biochemical networks [Findlay GD, Sitnik JL, Wang W, Aquadro CF, Clark NL, et al. (2014)]. However, less effort has been gone into exploring ERC at the domain level.

In most cases a protein will contain at least one domain which acts as a functional subunit [Bork Peer(1991)]. Since the function of a protein is often dependent upon the domain(s) present, isolating the ERC of the domains should reveal a more distinct relationship between a function and ERC. Because not every domain in any given proteome has been linked to its function, the best currently available way to explore ERC on the domain level is through domain families, which have been identified on the whole proteome level in databases like Pfam [Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., … Punta, M. (2014)].

We calculated the ERC of all the domains in the *S. cerevisiae* proteome using data from 13 yeast species. Yeast was chosen since previous work had been done with yeast ERC on the protein level with the data made available so that we could check our methods [Clark, N. L., Alani, E., & Aquadro, C. F. (2012)]. The domain ERC values were then grouped into their respective domain families and underwent a t-test to determine the significant values. These values were then visualized in a network to see how groups of domain families were interrelated. Groups of domain families with significant high ERC between them were then examined for function, evolutionary rate, and number of representative domains to explore potential explanations for their grouping.

# Methods

**Proteomes**

Proteomes for 13 yeast species were obtained from NCBI RefSeq [Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2005)], which is an annotated, curated database that provides one non-redundant record of the proteome for each species. The 13 yeast species used were: *Saccharomyces cerevisiae, Candida glabrate, Kluyveromyces thermotoleran, Kluyveromyces lactis, Eremothecium gossypii, Candida albicans, Candida dubliniensis, Candida tropicalis, Meyerozyma guilliermondii, Clavispora lusitaniae, Lodderomyces elongisporus, Debaryomyces hansenii*, and *Scheffersomyces stipites*.

**Orthologs**

Orthologs for each S. cerevisiae protein within all other yeast proteomes were found using Inparanoid [O'Brien, K. P., Remm, M., & Sonnhammer, E. L. L. (2005)]. Inparanoid is a tool that finds the best BLAST pairwise similarity scores to detect orthologs from among the various candidate proteins. The need for this tool comes from the difficulty in separating inparalogs from outparalogs. Outparalogs come from gene duplication events that occur before a speciation event, whereas inparalogs occur after a speciation event. Outparalogs are likely to have a more diversified function than inparalogs, so it is useful to exclude them and then only select the highest confidence ortholog candidate from the inparalog group.

**Multiple Sequence Alignment**

Multiple Sequence Comparison by Log-Expectation (MUSCLE) [Edgar, R. C. (2004)] was used to align the each of the protein ortholog sequences in the 13 yeast species. This alignment was made so that those sections of each sequence which were likely to have directly descended from a common ancestral section could be compared.

**Protein Families**

From the alignments, each set of orthologous protein sequences was then subdivided into its protein domains and linker regions using annotations from the Protein Families Database for *S. cerevisiae* (Pfam) [Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., … Punta, M. (2014)]. The domain family that each domain belonged to was recorded so that it could be identified in later analysis.

**Phylogenetic Trees**

The aligned sequences of each set of protein ortholog domains were then analyzed using the program Phylogenetic Analysis by Maximum Likelihood (PAML) [Ziheng Yang (2007)], which estimated the evolutionary rate based on a comparison of the synonymous and nonsynonymous substitution rate (dN/dS). After a phylogenetic tree topology containing all yeast species was input, PAML produced a phylogenetic tree for every domain, with the branch lengths scaled to the evolutionary rate within that batch

**Evolutionary Rate Covariation**

An R script provided by Nathan Clark was then used to calculate the ERC between every pair of domain phylogenetic trees. This script analyzed every set of trees with the same species topology, producing an average tree. Each of individual phylogenetic tree was then compared to the average tree to determine to what extent its branch lengths deviated from that of the average tree. These deviations were then compared for every pair of trees, and the correlation between them was the ERC.

**Comparison**

The protein domains were then grouped by domain family, allowing the ERC values between pairs of domain families to be analyzed. Any pairs of domains within the same protein were removed, since these would likely have elevated ERC values simply by being physically linked. All ERC values between pairs of domain families underwent a one-sided Student's t-test to determine whether their mean was significantly (p-value = 0.05) higher than 0. Since there were greater than 8,000 pairs of protein families with tests being run, it was necessary to correct for multiple testing. The Bonferroni correction was used for this purpose, yielding 32 pairs of protein families with significant, mean ERC greater than 0.2(Table 1, in appendix).

**Network Analysis**

All pairs of protein domain families with an significant ERC higher than 0.2 were then visualized in a network using NetworkX. The nodes in the network were assigned to each domain family

and the edge lengths were scaled by 1 – ERC, so that connected nodes that were closer together represented those with higher ERC values between them.

# Results

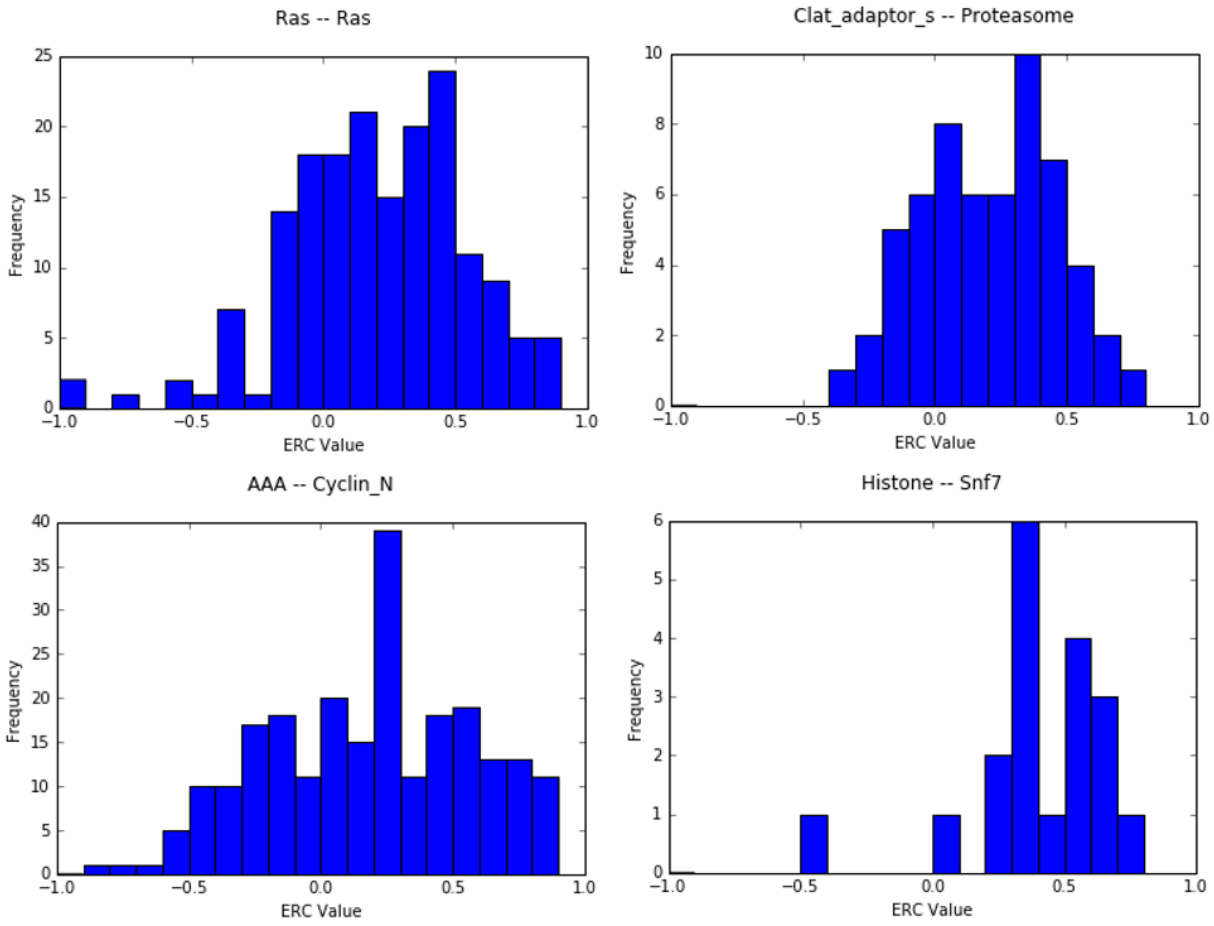


**Figure 2**: Sample histograms of ERC values between domains within each family. These distributions were shown to be statistically significant. The mean ERC values for the domain family pairs shown are 0.2168 (Ras—Ras), 0.2021 (Clat_adaptor_s—Proteasome), 0.2025 (AAA—Cyclin), and 0.3943 (Histone—Snf7).

ERC values were calculated for pair of domains and the domains were grouped by domain family. It was necessary to determine which pairs of domain families had elevated ERC between them, and which of these values was significant. For this reason, the distribution between each domain family underwent a t-test. Since there were thousands of distributions being tested, this presented a multiple comparisons problem. When many simultaneous tests are being made, it becomes more likely that there will be a test with a significant p-value due to random sampling error alone (a false positive) so it is necessary to decrease the cutoff value using a multiple comparisons correction. The Bonferroni correction is the most conservative correction, and was chosen to ensure higher confidence that the ERC values obtained were significant.

Out of a set of 8844 pairs of domain families, the distributions for 138 were shown to have a statistical significance of 0.05 using a one-sided t-test after application of the Bonferroni correction for multiple comparisons (Figure 2). Those with an average ERC value above 0.2 were selected for further analysis, of which there were 32.

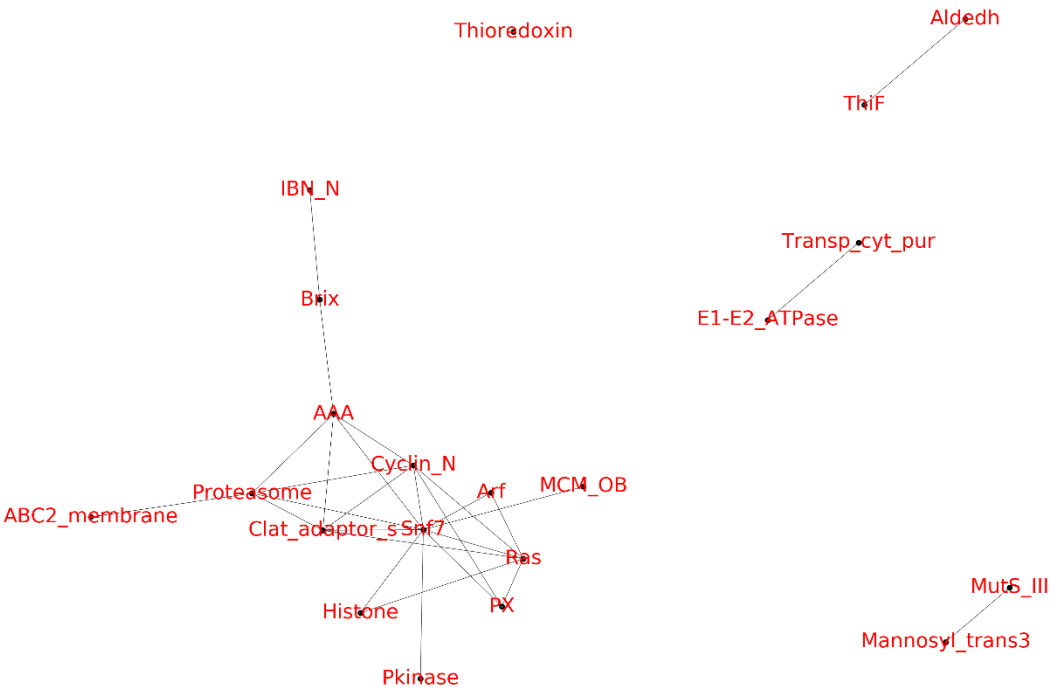


**Figure 3**: Network of protein domain families with significant ERC values greater than 0.2. Each node represents a domain family, and the edges are scaled by 1 - ERC value so that pairs with higher ERC are closer together than those with lower ERC.

Many of the 32 pairs of domain families with significant, high ERC values shared partners, so these pairs were graphed in a network to see how they were interrelated. It was expected that there would be several clusters of domain families all related by a shared function.

When the network was made, it was seen that there was only one major cluster of pairs of domain families with high ERC with significant values after the Bonferroni correction (Figure 3). The seven most highly connected domain families in the network were: proteasome subunit, Ras subfamily, protein kinase domain, ATPases associated with diverse cellular activities, cyclins, clathrin adaptors, and SNF7. These domain families did not all share the same general functions, but instead represent several distinct functions ranging from signaling, regulation, cell division, and transport.

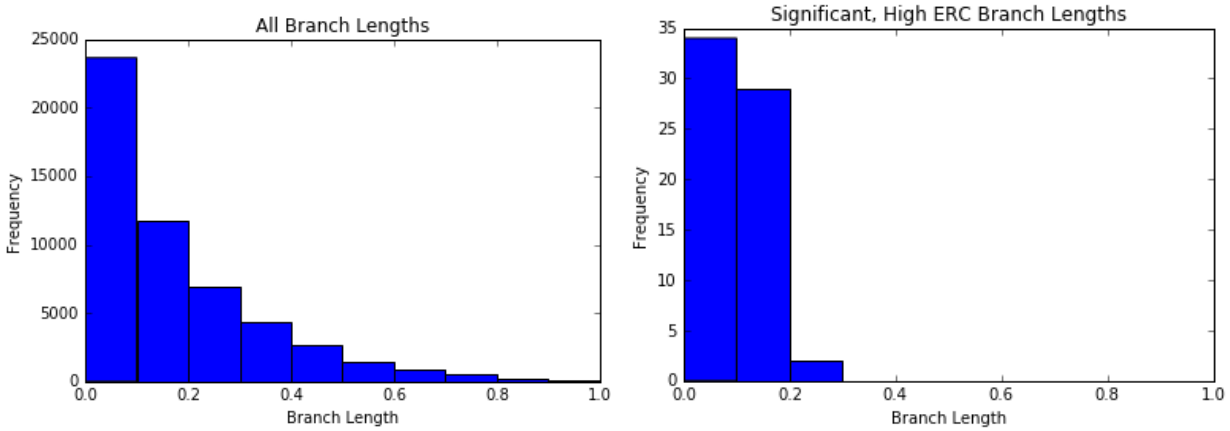


**Figure 4**: Comparison of phylogenetic tree branch lengths between all domains and those with significant ERC values greater than 0.2. Shorter branch lengths represent a lower evolutionary rate, and longer branches represent a higher evolutionary rate. The mean branch length for all domains is 0.1955, while that for significant, high ERC domains is 0.1022.

Since it seemed likely that this cluster of domain families was related by being conserved rather than sharing a function, an additional test was designed to test this hypothesis. Those domain families which are conserved will evolve at a slower rate than average. This corresponds to a shorter branch length in their phylogenetic trees.

The branch lengths for the whole proteome were collected, as were those belonging to the domain families with significant, high ERC. Compared to the overall mean phylogenetic tree branch length of 0.1955, those domains with significant, high ERC values had a mean branch length of almost half the size, 0.1022 (Figure 4). The phylogenetic tree branch lengths represent the evolutionary rate, so the cluster that appears in the network in figure 3 had a much smaller than average evolutionary rate.

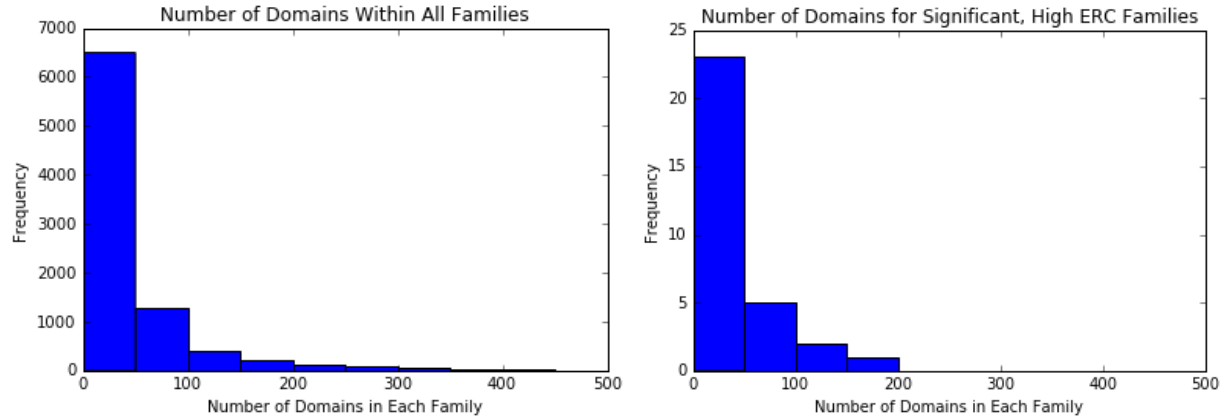


**Figure 5**: Comparison of the number of domains in each domain family for the entire dataset, and those with significant, high ERC. The mean number of domains for the entire dataset is 55, while that for significant, high ERC is 45.

The number of instances of a domain within a family can increase the likelihood that a value will be significant after a t-test and Bonferroni correction. This could affect the types of domains that appear in the result, and was a potential reason for these particular domain families appearing after these analyses. However, it turned out that the mean number of domains per family overall was 55, but the mean domains per family in the significant, high ERC families was smaller than that, at 45 (Figure 5).

# Discussion

The ERC between protein domain families was compared (Figure 2), and the significant values with an ERC greater than 0.2 were graphed using NetworkX (Figure 3). This analysis revealed a single large cluster of protein families with elevated ERC between its members. Before these results were discovered it was expected that multiple clusters of domain families would appear. Each cluster would be associated by a related function or other known relationship. Instead, this single cluster does not appear to have a coherent functional relationship between its constituents.

One plausible reason for this is that these domain families are related by being highly conserved. The seven most highly connected families were mainly associated with cell regulation, signaling, and division. Additionally, the mean branch length of these families was much smaller than most others on average (Figure 4). This indicates that these families did evolve more slowly. Perhaps by consistently showing lower than average evolutionary rates along each of its branches, this group of families showed high ERC without having an explicit function relationship. In this case, the relevant evolutionary pressure leading to high ERC would be the pressure to remain the same, which all highly conserved sequences feel.

Another potential explanation for this single cluster of non-functionally-related proteins was that they all were overrepresented in the proteome, and thus were the only ones to survive the conservative Bonferroni correction of p-values. To test this, the average number of representatives for domain families in the entire proteome was compared to that of the significant, high ERC families (Figure 5). Those in the entire proteome had an average of 55 representatives per family, while those with significant, high ERC had an average of 45. This indicates that overrepresentation was not the cause of their significance after the t-test and Bonferroni correction.

In the future, it would be interesting to do a gene ontology (GO) analysis of the genes for each protein which contain the domain families with significant, high ERC. This would provide a way to obtain functional annotations for these genes, and allow the functions associated with each domain family to be more fully mapped. In addition, since the Bonferroni correction is extremely conservative, determining if a more appropriate, less conservative multiple

comparisons correction existed may provide a greater number of domain families and clusters of domain families. GO analysis, combined with a less conservative correction would allow our inferences to be further tested, and would likely reveal additional relationships which don't appear in our results here.

It would also be worthwhile to look at specific domains with known functions. When looking at protein domain families, there is a lot of ambiguity when relating a function to a domain type since many domain families are present in a variety of proteins with divergent functions. If this analysis could be done between all domains, it could provide a more precise picture of the relationship between ERC and function. The reason this is currently difficult is that there are no large databases documenting the function of every domain on every protein in the proteome. If such a database is developed such an analysis may become possible. It may also be revealing to look at a subset of the proteome where much has already been cataloged about the functions of each domain, such as a well-studied biochemical pathway.

It would also be worthwhile to look at the whole protein ERC values among conserved proteins to see if they also show low ERC values. This would further confirm the inference that proteins don't necessarily have to interact or share a function to have a high ERC value between them.

# References


Clark, N. L., Alani, E., & Aquadro, C. F. (2012). Evolutionary rate covariation reveals shared functionality and coexpression of genes. *Genome Research*, *22*(4), 714–720. http://doi.org/10.1101/gr.132647.111

Priedigkeit N, Wolfe N, Clark NL (2015) Evolutionary Signatures amongst Disease Genes Permit Novel Methods for Gene Prioritization and Construction of Informative Gene-Based Networks. PLOS Genetics 11(2): e1004967. https://doi.org/10.1371/journal.pgen.1004967

Findlay GD, Sitnik JL, Wang W, Aquadro CF, Clark NL, et al. (2014) Evolutionary Rate Covariation Identifies New Members of a Protein Network Required for *Drosophila melanogaster* Female Post-Mating Responses. PLOS Genetics 10(1): e1004108. https://doi.org/10.1371/journal.pgen.1004108

Bork Peer(1991), Shuffled domains in extracellular proteins, FEBS Letters, 286, doi: 10.1016/0014-5793(91)80937-X

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., … Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Research*, *42*(Database issue), D222–D230. http://doi.org/10.1093/nar/gkt1223

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *33*(Database Issue), D501–D504. http://doi.org/10.1093/nar/gki025

O'Brien, K. P., Remm, M., & Sonnhammer, E. L. L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Research*, *33*(Database Issue), D476–D480. http://doi.org/10.1093/nar/gki107

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. http://doi.org/10.1093/nar/gkh340

Ziheng Yang (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* ; 24 (8): 1586-1591. doi: 10.1093/molbev/msm088

# Appendix

| Domain Family Pair | ERC Value | P-Value | Number of Domains |
|---|---|---|---|
| Snf7--Snf7 | 0.5097 | 9.184E-05 | 15 |
| Thioredoxin--Thioredoxin | 0.4925 | 8.538E-03 | 25 |
| Brix--Brix | 0.4429 | 3.380E-04 | 14 |
| Iso_dh--Snf7 | 0.4399 | 3.020E-02 | 18 |
| MCM_OB--Snf7 | 0.4312 | 2.148E-03 | 13 |
| Histone--Snf7 | 0.3944 | 4.237E-02 | 19 |
| PX--Snf7 | 0.3872 | 1.837E-04 | 37 |
| PX--Clat_adaptor_s | 0.3845 | 2.623E-02 | 19 |
| Cyclin_N--Snf7 | 0.3633 | 2.684E-05 | 31 |
| Ras--Snf7 | 0.3577 | 2.605E-12 | 72 |
| IBN_N--zf-met | 0.3452 | 5.496E-02 | 24 |
| E1-E2_ATPase--Transp_cyt_pur | 0.3164 | 4.144E-02 | 26 |
| Arf--LSM | 0.2959 | 2.355E-04 | 49 |
| AAA--Snf7 | 0.2951 | 1.295E-11 | 104 |
| Ras--Arf | 0.2942 | 1.177E-03 | 49 |
| Snf7--Proteasome | 0.2822 | 8.256E-03 | 37 |
| PX--Ras | 0.2742 | 1.356E-09 | 110 |
| Cyclin_N--PCI | 0.2670 | 9.046E-03 | 42 |
| Histone--Ras | 0.2527 | 1.240E-03 | 58 |
| Iso_dh--Ras | 0.2482 | 3.259E-02 | 49 |
| Cyclin_N--Ras | 0.2444 | 1.591E-03 | 84 |
| Proteasome--Clat_adaptor_s | 0.2426 | 5.446E-02 | 26 |
| ThiF--Aldedh | 0.2417 | 5.359E-02 | 29 |
| Ras--Brix | 0.2328 | 3.316E-03 | 70 |
| Ras--Clat_adaptor_s | 0.2271 | 4.768E-02 | 43 |
| Pkinase--Snf7 | 0.2261 | 4.437E-11 | 258 |
| Ras--Ras | 0.2168 | 2.936E-10 | 178 |
| Actin--LSM | 0.2111 | 2.252E-06 | 88 |

| | | | |
|---|---|---|---|
| Adaptin_N--Ras | 0.2108 | 8.198E-02 | 66 |
| Ras--AAA | 0.2069 | 2.124E-20 | 329 |
| AAA--Brix | 0.1999 | 1.719E-03 | 99 |
| MCM_OB--Cyclin_N | 0.1988 | 7.157E-02 | 35 |
| MCM_OB--AAA | 0.1979 | 4.032E-05 | 65 |
| Ras--Proteasome | 0.1912 | 4.502E-07 | 135 |
| PX--Proteasome | 0.1905 | 7.571E-02 | 76 |
| Histone--AAA | 0.1898 | 1.411E-03 | 101 |
| Proteasome--Proteasome | 0.1836 | 4.461E-05 | 91 |
| AAA--Clat_adaptor_s | 0.1824 | 3.275E-02 | 61 |
| PCI--RRM_1 | 0.1796 | 1.957E-03 | 108 |
| Cyclin_N--AAA | 0.1762 | 8.079E-05 | 145 |
| Aminotran_1_2--Mito_carr | 0.1729 | 1.579E-03 | 105 |
| PX--AAA | 0.1689 | 1.057E-06 | 190 |
| AAA--Proteasome | 0.1673 | 2.038E-08 | 207 |
| Pkinase--Brix | 0.1619 | 9.130E-06 | 274 |
| LSM--LSM | 0.1606 | 1.072E-02 | 91 |
| Pkinase--Cyclin_N | 0.1574 | 8.346E-13 | 659 |
| Pkinase--Ras | 0.1495 | 8.239E-17 | 767 |
| AAA--AAA | 0.1494 | 7.881E-17 | 496 |
| Pkinase--zf-met | 0.1492 | 7.201E-04 | 280 |
| Pkinase--Arf | 0.1439 | 1.387E-02 | 238 |
| GTP_EFTU_D2--Ras | 0.1349 | 1.251E-02 | 98 |
| PX--RRM_1 | 0.1346 | 3.677E-03 | 215 |
| Actin--RRM_1 | 0.1272 | 1.737E-03 | 219 |
| Pkinase--Histone | 0.1225 | 1.552E-02 | 320 |
| RhoGAP--AAA | 0.1171 | 7.591E-02 | 118 |
| RRM_1--LSM | 0.1147 | 7.167E-06 | 297 |
| Pkinase--PCI | 0.1146 | 2.924E-03 | 448 |
| MFS_1--MFS_1 | 0.1134 | 3.480E-02 | 423 |
| Ras--RRM_1 | 0.1116 | 1.994E-05 | 370 |
| Pkinase--PH | 0.1053 | 6.358E-05 | 600 |
| Pkinase--RhoGAP | 0.1019 | 3.430E-02 | 398 |
| Cyclin_N--WD40 | 0.0957 | 2.019E-02 | 369 |
| IBN_N--WD40 | 0.0894 | 5.378E-03 | 270 |
| Pkinase--MMR_HSR1 | 0.0893 | 1.014E-02 | 572 |
| Pkinase--AAA | 0.0880 | 2.914E-10 | 1435 |
| Pkinase--Proteasome | 0.0860 | 1.334E-02 | 591 |
| PX--WD40 | 0.0837 | 7.526E-03 | 466 |
| Pkinase--LSM | 0.0757 | 3.007E-03 | 635 |
| AAA--RRM_1 | 0.0731 | 8.032E-03 | 554 |
| Actin--WD40 | 0.0708 | 8.826E-02 | 455 |
| RRM_1--RRM_1 | 0.0668 | 9.493E-03 | 709 |

| | | | |
|---|---|---|---|
| Pkinase--Pkinase | 0.0588 | 4.252E-13 | 4450 |
| Pkinase--RRM_1 | 0.0545 | 1.870E-03 | 1603 |
| WD40--WD40 | 0.0475 | 1.422E-08 | 2082 |
| WD40--DEAD | 0.0403 | 3.219E-02 | 1100 |

**Table1.** All pairs of domain families with significant ERC values greater than 0.