Florida International University

FIU Digital Commons

FIU Electronic Theses and Dissertations | University Graduate School

3-28-2019

Computational Analysis of Large-Scale Trends and Dynamics in Eukaryotic Protein Family Evolution

Joseph Boehm Ahrens
Florida International University, jahre002@fiu.edu

Follow this and additional works at: https://digitalcommons.fiu.edu/etd

 Part of the Applied Statistics Commons, Bioinformatics Commons, Biostatistics Commons, Computational Biology Commons, Evolution Commons, Genomics Commons, Molecular Biology Commons, Molecular Genetics Commons, Statistical Models Commons, and the Theory and Algorithms Commons

Recommended Citation

Ahrens, Joseph Boehm, "Computational Analysis of Large-Scale Trends and Dynamics in Eukaryotic Protein Family Evolution" (2019). *FIU Electronic Theses and Dissertations*. 4039.
https://digitalcommons.fiu.edu/etd/4039

This work is brought to you for free and open access by the University Graduate School at FIU Digital Commons. It has been accepted for inclusion in FIU Electronic Theses and Dissertations by an authorized administrator of FIU Digital Commons. For more information, please contact dcc@fiu.edu.

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

COMPUTATIONAL ANALYSIS OF LARGE-SCALE TRENDS AND DYNAMICS IN EUKARYOTIC PROTEIN FAMILY EVOLUTION

A dissertation submitted in partial fulfillment of

the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOLOGY

by

Joseph B. Ahrens

2019

To: Dean Michael R. Heithaus
College of Arts, Sciences and Education

This dissertation, written by Joseph B. Ahrens, and entitled Computational Analysis of Large-Scale Trends and Dynamics in Eukaryotic Protein Family Evolution, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

_____
Timothy Collins

_____
Heather Bracken-Grissom

_____
Giri Narasimhan

_____
Prem Chapagain

_____
Jessica Siltberg-Liberles, Major Professor

Date of Defense: March 28, 2019

The dissertation of Joseph B. Ahrens is approved.

_____
Dean Michael R. Heithaus
College of Arts, Sciences and Education

_____
Andrés G. Gil
Vice President for Research and Economic Development
and Dean of the University Graduate School

Florida International University, 2019

© Copyright 2019 by Joseph B. Ahrens

All rights reserved.

DEDICATION

To my family, my friends and all of the people who reminded me to occasionally go outside in the process of completing this dissertation.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Jessica Liberles, for all of her support and guidance over the course of my graduate career at Florida International University. Thanks also for helping me to attend multiple academic conferences, as well as an excellent molecular evolution workshop. I would also like to thank Dr. Timothy Collins and Dr. Heather Bracken-Grissom for setting aside so much of their time to talk phylogenetics with me. Thanks to Dr. Giri Narasimhan for helping me with algorithm development, and Dr. Prem Chapagain for teaching me important biophysics concepts. Thanks to Jordon Rahaman for all of his help with writing, debugging and testing countless lines of code. Thanks to Dr. Laura Timm for helping me format, proofread and prepare this document, and for generally being available when I needed help. Thanks to graduate students Janelle Nunez-Castilla and Kyoko Nakamura for all of their help with teaching and writing. Thanks to Luis Nassar for his help with sequence data collection. Thanks also to the Instructional & Research Computing Center (IRCC) at Florida International University for providing HPC computing resources, and especially to Cassian D'Cunha and Mengxing Cheng for HPC support. Much of my work was supported by a Doctoral Evidence Acquisition Fellowship (Summer 2018) Dissertation Year Fellowship (Fall 2018 through Spring 2019) from Florida International University.

ABSTRACT OF THE DISSERTATION

COMPUTATIONAL ANALYSIS OF LARGE-SCALE TRENDS AND DYNAMICS IN EUKARYOTIC PROTEIN FAMILY EVOLUTION

by

Joseph B. Ahrens

Florida International University, 2019

Miami, Florida

Professor Jessica Siltberg-Liberles, Major Professor

The myriad protein-coding genes found in present-day eukaryotes arose from a combination of speciation and gene duplication events, spanning more than one billion years of evolution. Notably, as these proteins evolved, the individual residues at each site in their amino acid sequences were replaced at markedly different rates. The relationship between protein structure, protein function, and site-specific rates of amino acid replacement is a topic of ongoing research. Additionally, there is much interest in the different evolutionary constraints imposed on sequences related by speciation (orthologs) versus sequences related by gene duplication (paralogs).

A principal aim of this dissertation is to evaluate and characterize several broad trends in eukaryote protein evolution. To this end, I use sequence-based computational predictors of protein structure (intrinsic disorder and protein secondary structure) and protein function (predicted functional domains), in addition to Bayesian phylogenetic inference methods, to analyze thousands of homologous protein sequence clusters from four eukaryotic lineages: animals, plants, fungi and protists. Using these data, I performed large-scale factorial analyses, testing the correlation between protein structure/function

and rates of sequence evolution. The combined results of these analyses somewhat corroborate the findings of previous research in the field, but they also illuminate a subtle interaction among multiple drivers of protein sequence evolution, which is consistently observed across multiple eukaryote groups. Furthermore, using the results of Bayesian phylogenetic analysis on real and simulated protein sequence alignments, I show that orthologous and paralogous proteins exhibit significantly different overall patterns of sequence divergence, indicating that paralogs tend to evolve under relaxed selective pressure.

The acquisition of homologous biological sequence clusters is a prominent component of computational biological research. To assist in the identification of protein families within large sequence databases, I implement a simple, graph-based single-linkage clustering procedure, and I demonstrate its capacity to recover homologous subunits of the Rpt regulatory ring in the 26S proteasome complex.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

The following chapters have been or will be submitted for publication and are formatted according to journal specifications:

CHAPTER II

Ahrens, J., Dos Santos, H. G., & Siltberg-Liberles, J. (2016). The Nuanced Interplay of Intrinsic Disorder and Other Structural Properties Driving Protein Evolution. *Molecular Biology and Evolution*, *33*(9), 2248–56. http://doi.org/10.1093/molbev/msw092

CHAPTER III

Ahrens, J., Rahaman, J., & Siltberg-Liberles, J. (2018). Large-Scale Analyses of Site-Specific Evolutionary Rates across Eukaryote Proteomes Reveal Confounding Interactions between Intrinsic Disorder, Secondary Structure, and Functional Domains. *Genes*, *9*(11), 553. http://doi.org/10.3390/genes9110553

CHAPTER IV

Ahrens, J., Teufel, A. I., & Siltberg-Liberles, J. Evaluation of Site-Specific Rate Heterogeneity Reveals Significant Differences in Sequence Divergence Patterns Between Orthologous and Paralogous Proteins in Both Animals and Plants. Formatted for submission to *Molecular Biology and Evolution*.

CHAPTER V

Ahrens, J., Rahaman, J., & Siltberg-Liberles, J. Acquisition of Homologous Protein Sequence Clusters from Local Databases Using a Simple, Graph-Based Single-Linkage Clustering Procedure. Formatted for submission to *Molecular Biology and Evolution*.

CHAPTER I

INTRODUCTION

Molecular data—the nucleotide and amino acid sequences extracted from living organisms—has enhanced our understanding of virtually every aspect of biology. The biological sequence databases where molecular data is archived have grown substantially in the post-genomic era, and institutions like the European Bioinformatics Institute (EBI) and the National Center for Biotechnology Information (NCBI), are in a constant race to keep up with ever-increasing quantities of nucleotide and amino acid sequences (Cook et al. 2016; Agarwala et al. 2018). The resulting availability of large-scale molecular datasets, containing a multitude of complete genomes from divergent organisms, has improved our understanding of the complex functional relationships among genes (see Chen and Coppola 2018) and the evolutionary origins of living things (see Koonin 2010; Telford et al. 2015). Additionally, functional annotation (i.e., the assignment of gene functions to sequences of unknown function) can be greatly improved by considering gene function in an evolutionary context—rather than simply assigning function using sequence similarity, it is important to consider the particular evolutionary history (phylogeny) relating genes of known function to genes of unknown function (Eisen 1998; Eisen and Wu 2002). The analysis of whole-genome evolution, coupled with the unique evolutionary histories of individual genes, has transformed into a rapidly-maturing field of biology known as phylogenomics (Eisen and Fraser 2003).

Nucleotide substitutions within protein-coding genes can result in changes (replacements) in the individual residues of their translated amino acid sequences. However, a wide range of structural and functional constraints are known to govern protein sequence evolution (Echave et al. 2016). Thus, as a protein sequence evolves, the individual amino acids occupying each position (site) in the sequence can be replaced

over time at very different rates. Most of the statistical (likelihood-based) phylogenetic inference applications assume that sites in a protein evolve independently, according to a fixed rate matrix (see Felsenstein 1973). To account for site rate heterogeneity (i.e., differences in the relative speed of site-specific evolution), site rates are often assumed to be drawn from a discrete gamma distribution (Yang and Kumar 1996). Accounting for rate heterogeneity using a gamma distribution has been shown to greatly improve the accuracy of phylogenetic inference (Yang and Kumar 1996).

Considerable work has been done in recent decades to better understand the association between the structure/function of a given protein sub-sequence (region) and the site-specific rates of amino acid replacement within that region (see Echave et al. 2016). One of the primary drivers of rate heterogeneity in protein evolution is solvent exposure, and sites which are exposed to environmental solvents are often more variable than internal, buried residues (Perutz et al. 1965; Kimura and Ohta 1974; Franzosa and Xia 2009). Another notable driver is local packing density, since spatially proximal residues that form a large number of stabilizing contacts tend to be more conserved than residues that form fewer stabilizing contacts (Franzosa and Xia 2009; Yeh, Liu, et al. 2014; Yeh, Huang, et al. 2014). Additionally, protein regions exhibiting intrinsic disorder—an extreme form of conformational flexibility—experience more amino acid replacements than ordered regions (Brown et al. 2002) and the residue replacements in disordered regions are less biochemically conservative (Brown et al. 2010). Notably, studies of the above nature often concentrate on one-way or "single-factor" effects: the relationship between sequence evolutionary rates and a single aspect of protein structure. However, a single-factor methodology can easily prove problematic—if the confounding

effects of interacting variables are ignored, the measured effect of one structural property (on evolutionary rate) may stem from a failure to control for additional variables. For instance, (Huang et al. 2014) found that the positive correlation between flexibility and evolutionary rate is diminished after controlling for local packing density.

Protein scientists have conjectured for decades that, in addition to rate heterogeneity, site-specific shifts in amino acid replacement rates over time (i.e., sites exhibiting heterotachy) are a common feature of protein molecular evolution (Fitch and Markowitz 1970; Fitch 1971; Lopez et al. 2002). In fact, many statistical models of sequence evolution have been proposed to more directly account for heterotachy (e.g., Fitch and Markowitz 1970; Tuffley and Steel 1998; Galtier 2001). Additionally, it is thought that widespread shifts in site-specific evolutionary rates between two related protein sequences can indicate acquired structural or functional differences (Gu 1999; Gaucher et al. 2002). However, as mentioned previously, most contemporary phylogenetic inference applications—in the interest of computational tractability—assume that sites in a protein evolve independently, and do not explicitly account for heterotachy. Still, further work has shown that site-specific shifts in amino acid replacement rates can be inferred using existing statistical frameworks. For instance, (Abhiman et al. 2006) showed that changes in the shape parameter (α) of the inferred discrete gamma distribution of site rates is an indicator of functional divergence.

Sequence clustering (the agglomeration of similar sequences into subgroups or "clusters" within a database) is an indispensable component of large-scale protein data science. While clustering can be useful for simply creating reduced, non-redundant or reference protein databases like UniRef (Suzek et al. 2007), it is also quite useful for

identifying homologous groups of proteins which may be of interest to researchers studying gene/protein family evolution (see Huerta-Cepas et al. 2008) and genome functional annotation (Eisen 1998; Eisen and Wu 2002). While there are many sequence clustering applications available (e.g., Li et al. 2003; Li and Godzik 2006; Miele et al. 2011; Hauser et al. 2013), they tend to be optimized for the task of whole-database clustering (i.e., partitioning an entire database into sequence clusters based on a pre-defined similarity threshold). As such, researchers who are interested in a particular gene family have limited options when mining sequence databases.

The aim of this dissertation is, in part, to explore and analyze some of the overarching trends in eukaryotic protein sequence evolution. Additionally, I illustrate how a particular form of graph-based sequence clustering (single-linkage clustering) can be used for targeted identification of inclusive, presumably homologous protein sequence groups without clustering an entire database. In the following two chapters, I use large-scale sequence clustering analysis, combined with phylogenetic inference methods and sequence-based structural and functional predictors, to shed light on a subtle interaction of structural and functional factors driving site-specific protein sequence evolution. Afterward, I show that there are significant differences in patterns of sequence divergence between orthologous genes (related by speciation) and paralogous genes (related by gene duplication) in both plant and animal proteins. Finally, I describe an implementation of a simple, graph-based single-linkage clustering algorithm which is capable of identifying protein families for downstream evolutionary analysis.

The Nuanced Interplay of Intrinsic Disorder and Other Structural Properties Driving Protein Evolution

In the second chapter of this dissertation, I present a large-scale analysis of site-specific evolutionary rates across thousands of multiple sequence alignments of metazoan proteins. I used the single-linkage clustering program BLASTClust (Altschul et al. 1990) to generate thousands of homologous protein sequence clusters for my study. Rather than relying on publicly-available 3D protein structural data, I used sequence-based structural prediction methods on every protein sequence in my dataset to detect conserved intrinsic disorder, secondary structure, and functional domains within alignments, which allowed me to analyze a large number of proteins with unknown structure and function. I then used Bayesian phylogenetic inference combined with empirical Bayesian site rate estimation to analyze rate heterogeneity in millions of amino acid alignment sites. The primary aim of my study is to better understand the relationship between structural properties and the evolutionary rates of protein residues, particularly with regard to intrinsic disorder. I also employed a factorial experimental design to investigate the possibility of statistical interactions among the three factors under evaluation.

Large-Scale Analyses of Site-Specific Evolutionary Rates across Eukaryote Proteomes Reveal Confounding Interactions between Intrinsic Disorder, Secondary Structure, and Functional Domains

In the third chapter, I present an extended evaluation of the structural factors studied in Chapter II, where I analyze protein sequence datasets representing four divergent eukaryotic lineages: metazoans (animals), plants, saccharomycete fungi and

alveolate protists. Here, I used the same sequence-based predictors employed in Chapter II to identify protein family alignment sites with conserved intrinsic disorder, secondary structure and functional domain predictions. I also applied the same multifactor statistical analyses used in Chapter II to measure the effects of these structural and functional factors on site-specific rates of sequence evolution. Additionally, I used the combined results of structural prediction and gene ontology (GO) term analysis (Ashburner et al. 2000) to identify and characterize "disordered-structured" sites (with both intrinsic disorder and secondary structure propensity), which exist in low abundance in all four eukaryotic groups studied here. The aim of this chapter is to discern whether there are statistically significant, and broadly consistent forces driving eukaryotic protein evolution.

Evaluation of Site-specific Rate Heterogeneity Reveals Significant Differences in Sequence Divergence Patterns between Orthologous and Paralogous Proteins in Both Animals and Plants

In the fourth chapter, I present a large-scale study where I evaluate differences in sequence divergence patterns between alignments of orthologous and paralogous protein sequences found in metazoans (animals) and plants. For this work, I utilized thousands of sequence clusters (taken from the animal and plant datasets in Chapter III) representing either putative orthologous relationships (i.e., sequences arising from a speciation event) or paralogous relationships (i.e., sequences arising from gene duplications). Using sequence-based phylogenetic analyses, I establish a correlation between sequence alignment divergence (total branch length of the phylogenetic tree) and the α parameter

of the sequence alignment's inferred discrete gamma rate distribution. I also develop and describe simple computational protein sequence simulation methods which reproduce the correlations observed in real protein sequence data. Finally, I show that the correlation between divergence (tree length) and rate heterogeneity (α) is significantly different between orthologous and paralogous genes in both plants and animals, and I discuss the potential implications of this difference.

Acquisition of Homologous Protein Sequence Clusters from Local Databases Using a Simple, Graph-Based Single-Linkage Clustering Procedure

A wide range of bioinformatics utilities are available for the task of sequence clustering: the agglomeration of similar biological sequences into subgroups or "clusters." In the fifth chapter of this dissertation, I outline a simple computational procedure for defining a single-linkage cluster in a protein sequence database using a combination of pairwise sequence identity and a bi-directional measurement of sequence alignment quality. Additionally, I describe a straightforward implementation of this clustering procedure using a combination of the Python programming language (Rossum and Guido 1995) and BLAST (Altschul et al. 1990). I benchmark the performance of said implementation using the same database of metazoan (animal) proteomes from Chapters II, III and IV. Finally, via a combination of phylogenetic inference and sequence-based structural/functional predictions, I demonstrate that our procedure can recover large, divergent protein families, using sequences from the regulatory ring (Rpt proteins) of the 26S proteasome complex as a notable example result. Leveraging the combined results of phylogenetic, structural and functional analysis, I also summarize the evolutionary

history and paralog-specific structural/functional divergence observed in the single-linkage Rpt sequence cluster.

Intellectual Merit

The work described in this dissertation constitutes several novel contributions to the field. The large protein sequence datasets constructed in Chapters II and III contain tens of thousands of aligned protein sequence clusters, replete with inferred evolutionary histories and sequence-based predictions of intrinsic disorder propensity, secondary structures and functional domains. These datasets can serve as a springboard for a plethora of future computational studies for years to come. In Chapter II, I describe a complex interaction of three structural and functional factors (intrinsic disorder, secondary structure and functional domain involvement) driving site-specific protein sequence evolution. My follow-up study in Chapter III confirms that the trends identified in animal protein evolution are consistently observed in three other eukaryotic groups as well (plants, protists and fungi). This analysis is, to my knowledge, the largest and most comprehensive of its kind. Further, by evaluating a combination of structural predictions, I was able to identify a conserved subset of protein sequence sites found in all four of the eukaryotic lineages I studied ("disordered-structured" sites), many of which appear to function through real-time alternations between intrinsic disorder and secondary structure. The protein sequence simulation work I conducted (as part of my analysis in Chapter IV) illustrates a simple method for incorporating heterotachy into simulated sequence data. Moreover, my analysis of real protein sequences indicates that there are significantly different sequence divergence patterns between orthologous and paralogous

genes over evolutionary time scales which, plausibly, result from differences in their selective/functional constraints. Lastly, I describe a single-linkage clustering procedure, as well as a simple implementation of said procedure, capable of mining individual homologous groups of protein sequences from large sequence databases. The clustering procedure is useful for mining individual protein families from databases which are too large to be clustered in their entirety.

LITERATURE CITED

Abhiman S, Daub CO, Sonnhammer ELL. 2006. Prediction of Function Divergence in Protein Families Using the Substitution Rate Variation Parameter Alpha. Mol. Biol. Evol. 23:1406–1413.

Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, Bolton E, Bourexis D, Brister JR, Bryant SH, Canese K, et al. 2018. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 46:D8–D13.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene Ontology: tool for the unification of biology. Nat. Genet. 25:25–29.

Brown CJ, Johnson AK, Daughdrill GW. 2010. Comparing models of evolution for ordered and disordered proteins. Mol. Biol. Evol. 27:609–621.

Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, Williams CJ, Dunker AK. 2002. Evolutionary rate heterogeneity in proteins with long disordered regions. J. Mol. Evol. 55:104–110.

Chen J, Coppola G. 2018. Bioinformatics and genomic databases. In: Handbook of clinical neurology. Vol. 147. p. 75–92.

Cook CE, Bergman MT, Finn RD, Cochrane G, Birney E, Apweiler R. 2016. The European Bioinformatics Institute in 2016: Data growth and integration. Nucleic Acids Res. 44:D20-6.

Echave J, Spielman SJ, Wilke CO. 2016. Causes of evolutionary rate variation among protein sites. Nat. Rev. Genet. 17:109–121.

Eisen JA. 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res. 8:163–167.

Eisen JA, Fraser CM. 2003. Phylogenomics: Intersection of Evolution and Genomics. Science. 300:1706–1707.

Eisen JA, Wu M. 2002. Phylogenetic analysis and gene functional predictions: phylogenomics in action. Theor. Popul. Biol. 61:481–487.

Felsenstein J. 1973. Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters. Syst. Zool. 22:240.

Fitch WM. 1971. The nonidentity of invariable positions in the cytochromes c of different species. Biochem. Genet. 5:231–241.

Fitch WM, Markowitz E. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. 4:579–593.

Franzosa EA, Xia Y. 2009. Structural determinants of protein evolution are context-sensitive at the residue level. Mol. Biol. Evol. 26:2387–2395.

Galtier N. 2001. Maximum-Likelihood Phylogenetic Analysis Under a Covarion-like Model. Mol. Biol. Evol. 18:866–873.

Gaucher EA, Gu X, Miyamoto MM, Benner SA. 2002. Predicting functional divergence in protein evolution by site-specific rate shifts. Trends Biochem. Sci. 27:315–321.

Gu X. 1999. Statistical Methods for Testing Functional Divergence after Gene Duplication. Mol. Biol. Evol 16:1664–1674.

Hauser M, Mayer CE, Söding J. 2013. kClust: fast and sensitive clustering of large protein sequence databases. BMC Bioinformatics 14:248.

Huang T-T, del Valle Marcos ML, Hwang J-K, Echave J. 2014. A mechanistic stress model of protein evolution accounts for site-specific evolutionary rates and their relationship with packing density and flexibility. BMC Evol. Biol. 14:78.

Huerta-Cepas J, Bueno A, Dopazo J, Gabaldón T. 2008. PhylomeDB: a database for genome-wide collections of gene phylogenies. Nucleic Acids Res. 36:D491-6.

Kimura M, Ohta T. 1974. On some principles governing molecular evolution. Proc. Natl. Acad. Sci. U. S. A. 71:2848–2852.

Koonin E V. 2010. The origin and early evolution of eukaryotes in the light of phylogenomics. Genome Biol. 11:209.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659.

Lopez P, Casane D, Philippe H. 2002. Heterotachy, an Important Process of Protein Evolution. Mol. Biol. Evol. 19:1–7.

Miele V, Penel S, Duret L. 2011. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics 12:116.

Perutz MF, Kendrew JC, Watson HC. 1965. Structure and function of haemoglobin. J. Mol. Biol. 13:669–678.

Rossum, Guido. 1995. Python reference manual.

Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23:1282–1288.

Telford MJ, Budd GE, Philippe H. 2015. Phylogenomic Insights into Animal Evolution. Curr. Biol. 25:R876–R887.

Tuffley C, Steel M. 1998. Modeling the covarion hypothesis of nucleotide substitution. Math. Biosci. 147:63–91.

Yang Z, Kumar S. 1996. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. Mol. Biol. Evol. 13:650–659.

Yeh S-W, Huang T-T, Liu J-W, Yu S-H, Shih C-H, Hwang J-K, Echave J. 2014. Local packing density is the main structural determinant of the rate of protein sequence evolution at site level. Biomed Res. Int. 2014:572409.

Yeh S-W, Liu J-W, Yu S-H, Shih C-H, Hwang J-K, Echave J. 2014. Site-specific structural constraints on protein sequence evolutionary divergence: local packing density versus solvent exposure. Mol. Biol. Evol. 31:135–139.

CHAPTER II

THE NUANCED INTERPLAY OF INTRINSIC DISORDER AND OTHER STRUCTURAL PROPERTIES DRIVING PROTEIN EVOLUTION

ABSTRACT

Protein evolution often occurs at unequal rates in different sites along an amino acid chain. Site-specific evolutionary rates have been linked to several structural and functional properties of proteins. Previous analyses of this phenomenon have involved relatively small datasets and, in some cases, the interaction among multiple structural factors is not evaluated. Here, we present the results of a large-scale phylogenetic and statistical analysis, testing the effects and interactions of three structural properties on amino acid replacement rates. We used sequence-based computational methods to predict (i) intrinsic disorder propensity, (ii) secondary structure, and (iii) functional domain involvement across millions of amino acid sites in thousands of sequence alignments of metazoan proteins. Our results somewhat corroborate earlier findings that intrinsically disordered sites tend to be more variable than ordered sites, but there is considerable overlap among their rate distributions, and a significant confounding interaction exists between intrinsic disorder and secondary structure. Notably, protein sites that are consistently predicted to be both intrinsically disordered and involved in secondary structures tend to be the most conserved at the amino acid level, suggesting that they are highly constrained and functionally important. In addition, a significant interaction exists between functional domain involvement and secondary structure. These findings suggest that multiple structural drivers of protein evolution should be evaluated simultaneously in order to get a clear picture of their individual effects as well as any confounding interactions among them.

INTRODUCTION

Protein evolution is commonly modeled in a "site-specific" manner, where the individual amino acid sites in a polypeptide are assumed to evolve independently at different rates. Understanding the driving forces of site-specific rate variation is a challenging but crucial endeavor in which many researchers are currently engaged (Echave et al. 2016). The primary drivers of rate heterogeneity appear to be (i) solvent exposure, where residues exposed to environmental solvents are more variable than internal, buried residues (Perutz et al. 1965; Kimura and Ohta 1974; Franzosa and Xia 2009), and (ii) local packing density, where spatially proximal residues that form a large number of stabilizing contacts tend to be more conserved (Franzosa and Xia 2009; Yeh, Huang, et al. 2014; Yeh, Liu, et al. 2014). In addition, research suggests that intrinsically disordered protein regions experience more amino acid replacement than ordered regions (Brown et al. 2002) and that residue replacement in disordered regions is less biochemically conservative (Brown et al. 2010). These results appear to be compatible with the abiding notion that ordered regions tend to be more crucial to protein structure and function than disordered regions, though there are prominent counterexamples to this trend (Brown et al. 2002; van der Lee et al. 2014).

Many studies relating protein structural properties to sequence evolution have focused on one-way or "single-factor" effects. This approach is appealing from a modeling standpoint: a strong correlation between a single structural property (e.g., packing density, intrinsic disorder, or solvent accessibility) and replacement rate would be more straightforward in terms of explanatory power. A serious issue with this methodology, however, is that the confounding effects of interacting variables are

ignored. Thus, the perceived effect of one structural property on evolutionary rate might actually be caused by an unbalanced representation of additional factors that were not controlled for. This is apparently the case with some metrics of structural flexibility, since the positive correlation between flexibility and evolutionary rate disappears after controlling for local packing density (Huang et al. 2014).

Here, we present a large-scale analysis of site-specific evolutionary rates across thousands of multiple sequence alignments of metazoan proteins. Using structural prediction methods, Bayesian phylogenetic inference, and empirical Bayesian rate estimation, we were able to analyze millions of amino acid sites for the presence of intrinsic, disorder, secondary structure, and functional domains. The aim of this study was to better understand the relationship between the structural properties and evolutionary rates of protein residues, particularly in the case of intrinsic disorder. Importantly, we have predicted the structural properties of every sequence in each alignment in order to locate structurally conserved amino acid sites for downstream analysis. Furthermore, we employed a factorial design to investigate the possibility of statistical interactions among the three structural factors being studied.

RESULTS

Clustering Analysis and Phylogeny

Clustering analysis yielded a total of 13,003 clusters containing between 5 and 600 sequences (fig. 1). 11,973 of these clusters were of sufficient quality (30% minimum sequence identity, 50% minimum alignment coverage) to be used in the phylogenetic analysis (fig. 1). The species composition of these 11,973 clusters showed considerable

variation. 1,029 clusters were species-specific, containing only protein sequences from a single species (fig. 2), while 5,893 others indicated ortholog groups, having only one sequence per species. Notably, the majority of clusters had between 5 and 12 species. Many of these were specific to either the mammalian lineage (695), or the arthropod lineage (458). A small number of clusters (32) contained at least one sequence from all 25 species in the dataset (figs. 3 and 4).

Most of the Bayesian phylogenetic inference analyses reached a very low convergence diagnostic (average standard deviation of split frequencies <0.005) prior to running for their allotted 5 million generations. Only 35 analyses that ran for a full 5 million generations ended with an average standard deviation of split frequencies higher than 0.01, the typical "stop" value recommended by the program authors (Ronquist et al. 2011).

Structural Prediction

Of the 7,990,416 aligned sites from all sequence clusters used in the phylogenetic analysis, 5,898,946 (~74%) did not contain any gap characters. 3,214,254 of these nongapped sites were predicted to be consistently ordered (i.e., every sequence in an alignment was predicted to be ordered at a particular site), while 993,937 sites were predicted to be consistently disordered. 1,879,338 sites were predicted to have conserved secondary structure (either always α-helix or always β-strand) and 2,664,147 were conserved coil sites. 2,391,699 fell unanimously inside of predicted functional domains, while 2,899,814 were in linker regions. In total, there were 2,969,226 sites with conserved predictions for all three of the above factors (disorder propensity, secondary structure, and domain involvement), making them ideal for 2^3 factorial analysis.

Statistical Analysis

Ordered sites had a median amino acid transition rate of −0.58, whereas the median rate for disordered sites was −0.25 (fig. 5 A). Similarly, sites predicted to be in secondary structures had lower median transition rates than coil sites (−0.57 vs. −0.40, respectively; fig. 5 B), and sites in functional domains had lower median transition rates than sites in linker regions (−0.61 vs. −0.32, respectively; fig. 5 C). Mann–Whitney tests of all three differences indicated high statistical significance ($P < 2 \times 10^{-16}$).

A Kruskal–Wallis test of all eight factor-level combinations indicated a significant difference in rate distributions ($P < 2 \times 10^{-16}$). Post hoc evaluation (fig. 6) indicated that only 4 of the 28 pairwise comparisons were not significantly different ($\alpha = 0.05$, corrected for multiple comparisons). All four nonsignificant pairwise differences involved comparisons with rate distributions represented by comparatively small sample sizes (table 1). In corroboration with the Mann–Whitney test results, the sites predicted to be disordered, lacking in secondary structure, and outside of domains had the highest median transition rate (−0.10; mean = 0.18). In contrast, the set of ordered sites with secondary structure in domains had a lower median transition rate (−0.63; mean = −0.37). The lowest median rate was observed in sites predicted to be disordered, but also involved in secondary structures and domains (−0.70; mean = −0.47) (table 1).

The 2^3 factorial analysis supported the results of the nonparametric analyses in terms of only the three main factor effects (table 2). All three differences in rate means (disordered sites vs. ordered sites, coil sites vs. structured sites, and linker sites vs. domain sites) were statistically significant ($P < 2 \times 10^{-16}$). Nonetheless, the overall fit of the factorial model was low (adjusted $R^2 \sim 0.05$). In addition, a significant interaction

($P < 2 \times 10^{-16}$) was observed between disorder propensity and secondary structure, as well as between secondary structure and domain involvement. Trace plots indicate that the effect of disorder propensity is confounded by the effect of secondary structure (fig. 7 A). In other words, sites predicted to be disordered have higher amino acid replacement rates on average than ordered sites, provided that they are not involved in secondary structures. In contrast, ordered sites predicted to be involved in secondary structures have higher average replacement rates than disordered, structured sites, which have lower average replacement rates than any other structural group (overall mean = −0.39; −0.47 for sites in domains and −0.35 for sites in linkers). In addition, the effect of domain involvement (i.e., the difference in mean amino acid replacement rates between sites in domains and linkers) is larger in sites where there is no predicted secondary structure (fig. 7 B). A significant higher-order interaction was also detected among all three structural factors, though the confounding effects of the disorder–structure interaction are still present in both domain and linker sites.

DISCUSSION

Clustering Analysis and Phylogeny

Our analysis returned a relatively large number of sequence groups representing between 5 and 12 species (fig. 3). This result is expected due to the inherent taxonomic sampling bias present in well-curated genomic databases and, by extension, the proteome set selected for our study. Many (but not all) of the clusters in this species range contain arthropod-specific genes (i.e., the five arthropods in our dataset), mammal-specific genes (six of the vertebrates) or vertebrate-specific genes (12 vertebrates in total; see fig. 4).

Any clusters with more than 12 species represent genes found across multiple phyla that are still sufficiently conserved to form a group based on our linkage cutoffs. Similarly, the large number of clusters containing sequences from only a single species (fig. 2) are partly the result of taxonomic bias. Many of these species-specific clusters contain a species that is the sole representative of a phylum in our dataset (e.g., *Strongylocentrotus purpuratus*, *Amphimedon queenslandica*, *Caenorhabditis elegans*). Others contain a species that is highly divergent from the other members of their phylum (e.g., *Branchiostoma floridae*, *Daphnia pulex*). It is likely that some of the genes in these clusters are also present in other species, not included here, that are more closely related to these taxa.

The number of species with well-annotated proteomes is steadily growing, and an increasing number of animal phyla (e.g., annelids, mollusks, and flatworms) can now be represented, at least by a single taxon, in large-scale metazoan studies using existing databases (UniProt Consortium 2014). Still, several large animal phyla contain highly divergent lineages that are hundreds of millions of years old (e.g., Echinodermata, Cnidaria, Porifera, Mollusca, Annelida), yet they are sparsely represented in curated proteome databases. Exploring metazoan protein evolution more completely will require a more even representation of multiple taxonomic groups. Future efforts to curate animal proteomes should focus on underrepresented groups such as Cnidaria (corals, anemones, and jellyfish), Annelida (leeches, earthworms, and polychaetes), and Mollusca (bivalves, gastropods, cephalopods, etc.).

Structural Prediction

Although the majority of gap-free alignment sites in our dataset were predicted to be ordered, nearly 1 million (~17%) were predicted to be conserved disordered sites. 186,026 (~19%) of these disordered sites were in conserved disordered regions at least 30-amino acids long. Overall, our results indicate that a non-negligible percentage of disordered metazoan protein sites are shared among species and among paralogous genes.

Interestingly, about 118,596 (~6%) of the sites predicted to be involved in secondary structures were also consistently predicted to be disordered. Statistical analysis of these sites revealed that they have the lowest average evolutionary rates (table 1). The meaning of this prediction combination (structured yet disordered) is unclear. These may be sites that have high propensities for secondary structure formation but are nonetheless disordered. For instance, they may belong to protein regions that alternate between intrinsic disorder and ordered secondary structure, as is sometimes the case in allostery (Motlagh et al. 2014), and in molecular recognition features or MoRFs (Yan et al. 2016). The IUPred algorithm predicts disorder propensity by estimating the potential for each residue to form stabilizing contacts with local residues within a predefined amino-acid window (Dosztányi et al. 2005). Some of these sites may be stabilized via contacts with residues outside of this window, or they may be exposed to solvents and form very few contacts with other residues in their respective proteins. Given their high level of average sequence conservation, future studies should focus on elucidating the functions of these conserved, disordered, and structured sites.

Statistical Analysis

Previous work suggests that flexible, intrinsically disordered regions of proteins experience higher rates of amino acid replacement than ordered sites (Brown et al. 2002) and that variable regions of protein alignments are difficult to study because they often contain missing residues resulting from insertions/deletions, and that disordered regions might be particularly affected by alignment gaps (Brown et al. 2011). Our study is limited to protein sites with consistent structural predictions and without missing characters, so nonconserved structural regions are not considered, as their putative structure is unclear. Still, our results indicate that, on average, (i) intrinsically disordered sites tend to evolve faster than ordered sites, (ii) sites in coil regions tend to evolve faster than sites that are involved in secondary structures, and (iii) sites in linker regions tend to evolve faster than sites within functional domains. However, in all three cases, there is considerable overlap in the rate distributions of the sites being compared (fig. 5). Notably, previous research indicates that intrinsic disorder can occur in either conserved or variable regions of a protein, and that the relative sequence conservation of a disordered region is correlated with protein function (Bellay et al. 2011). This finding is consistent with our results, in that we observe a broad range of evolutionary rates associated with intrinsically disordered sites in our dataset.

The significant interaction terms in the factorial analysis indicate that individual effects of disorder propensity, secondary structure, and domain involvement provide an incomplete picture of the driving forces behind protein evolution. For example, the shift in transition rates between ordered and disordered sites is larger outside of functional domains. Trace plots also indicate that both ordered and disordered sites tend to

experience more sequence conservation in functional domains than in linker regions. For sites within secondary structures, the overall effect of disorder propensity is actually reversed: sites predicted to be disordered tend to be more conserved than ordered sites.

Despite strongly significant main effects and interaction terms, the overall fit of the factorial model was quite low (adjusted $R^2 \sim 0.05$). This value highlights the subtle but important distinction between the statistical significance and the practical significance of our results. Because of the large overlap in rate distributions among different structural site categories, the factorial model only explains a small percentage of the total variance in the dataset, and thus has very poor predictive power. Therefore, the claim that intrinsically disordered sites tend toward higher amino acid replacement rates appears valid, but the notion that rapidly evolving protein regions are most likely disordered is not supported here.

The relationships between various structural properties of proteins and site-specific evolutionary rates are currently a topic of great interest (Echave et al. 2016). Recent studies have indicated fairly strong correlations between specific structural properties and amino acid replacement rates (Franzosa and Xia 2009; Yeh, Huang, et al. 2014). Our results highlight the importance of considering combinations of structural factors in future studies in order to account for their interactions when estimating evolutionary rates. Moreover, our large-scale analysis illuminates a subtle interplay between sequence evolution and structural properties across a diverse range of metazoan proteins.

METHODS

Data Collection

Complete, canonical (single isoform: only one protein representative per gene) proteomes for 24 metazoan taxa and one choanoflagellate (fig. 4) were retrieved from the 2014_4 release of the Uniprot Reference Proteome Set (Suzek et al. 2007; UniProt Consortium 2014). These taxa were selected to represent important divergence events in the evolutionary history of metazoans, with special emphasis placed on chordates and, to a lesser extent, arthropods. In addition, we specifically included many model organisms, such as *Mus musculus*, *Danio rerio*, *Drosophila melanogaster*, and *C.elegans*. For each taxon, protein sequences shorter than 30 amino acids, as well as sequences with unidentified residues (marked by an "X" symbol), were excluded from the dataset.

Clustering Analysis

We used the clustering program BLASTClust from BLAST2.26 (Altschul et al. 1990) to sort our database into groups of sequences for which phylogenies could be constructed. Clusters were formed based on pairwise similarities in amino acid sequences (≥40% sequence identity) as well as similarities in aligned sequence lengths (the BLAST alignment footprint must cover ≥90% of both proteins). BLASTClust uses a single-linkage algorithm, meaning that sequences were added to a cluster if they were sufficiently similar to any sequence already in that cluster. The aforementioned combination of similarity cutoffs was chosen because it produced a large number of relatively small clusters, without highly divergent sequences that could negatively impact phylogenetic analyses (fig. 1). At the same time, these clusters were generally inclusive enough to reconstruct phylogenies illustrating the relationships of many homologous

proteins separated by speciation events (orthologs) and, where applicable, the relationships among duplicated genes (paralogs).

Phylogenetic Analysis

Each cluster containing between 5 and 600 full-length protein sequences was aligned via MAFFT v7.123b (Katoh and Standley 2013) using the "local-pair" strategy (L-INS-i algorithm for a maximum of 1,000 iterations). This strategy was deemed to be the most accurate iterative protein alignment method, both by the program authors and by the results of a recent benchmark study (Thompson et al. 2011). Aligned clusters were checked for sequences that might confound the quality of the multiple sequence alignment and, consequently, the results of phylogenetic analyses. Clusters containing any sequence pair with less than 30% global sequence identity, or any sequence covering <50% of the length of the multiple sequence alignment, were excluded from phylogenetic analysis. Large clusters (containing >600 sequences) were excluded as well. Lastly, any clusters of sequences with nonstandard or ambiguous amino acid characters were omitted from further analysis.

Phylogenetic trees were constructed using the Bayesian inference method implemented in MrBayes 3.2.2 (Ronquist et al. 2012). We used the mixed-model approach in MrBayes to estimate the model type and parameters used in each analysis. A gamma distribution among rates (four discrete categories) was applied to each alignment. Each analysis (two independent runs, four chains per run) was allowed to run either for 5 million generations or until the average standard deviation of split frequencies fell below 0.005. Runs were summarized into 50% majority-rule consensus trees, with the first 25% of trees being discarded as burn-in. We used the gene–species tree reconciliation software

Notung 2.6 (Durand et al. 2005; Vernot et al. 2007) to root each gene tree. The species tree used to determine root placements was based on the NCBI Taxonomy Common Tree (Benson et al. 2009; Sayers et al. 2009), with two important exceptions (fig. 1). First, the unresolved portion of the metazoan tree containing Porifera, Placozoa, and Eumetazoa was altered to make Porifera basal to the latter two, in accordance with the conclusion of a recent review of metazoan phylogeny (Dohrmann and Wörheide 2013). Second, we positioned the lancelet *B.floridae* basal to Urochordata and Vertebrata based on genomic evidence (Gee 2008; Putnam et al. 2008).

Site-specific evolutionary rates were estimated for each sequence alignment using the program Rate4Site 3.0.0 (Mayrose et al. 2004). Rates were estimated using the empirical Bayesian approach assuming a gamma distribution with 16 rate categories. The rooted, 50% majority-rule consensus trees described above were used as input phylogenies and no further refinement or branch length optimizations were performed. All rate estimations were computed using the amino-acid substitution matrix developed by Jones et al. (1992). Site-specific rate estimates were transformed to normalized *Z* -scores (the default normalization procedure within Rate4Site) with mean equal to 0 and standard deviation equal to 1, allowing for the distinction between slow-evolving sites (negative values) and fast-evolving sites (positive values) across multiple alignments.

Structural Prediction

Because the number of protein sequences in our dataset far exceeds the number of proteins with empirically determined structural information, structural characterization of each protein was achieved using well-established, sequence-based predictors. Amino acid sites from each alignment were categorized according to three binary structural factors:

intrinsic structural disorder propensity (i.e., is the site predicted to be ordered or disordered?), secondary structure (is the site part of a secondary structure or a coil?), and domain involvement (is the site part of a domain or linker region?). Intrinsic disorder, secondary structure, and domain predictions were obtained for full-length proteins in order to preserve the structural context of individual amino acids. Only sites that could be consistently categorized for all proteins in each multiple sequence alignment were used in statistical analyses. Sites containing gap characters or non-conserved predictions (e.g., disordered in some sequences but ordered in others) were excluded.

Intrinsic structural disorder (i.e., the low propensity to form stable intramolecular contacts) was predicted with IUPred 1.0 (Dosztányi et al. 2005) using the option for detecting long disordered regions. IUPred was specifically developed for the *de novo*prediction of intrinsically unfolded protein regions via estimated energy content, without assuming disorder conservation in related sequences. By default, a score >0.5 indicates a propensity toward disorder, with a maximum score of 1.0 indicating an extreme propensity to be in a disordered state. We instead used a cut-off of 0.4 for the binary conversion of disorder predictions, as this threshold is purportedly more accurate when predicting disordered regions in experimentally verified disordered proteins (Fuxreiter et al. 2007; Xue et al. 2009).

Secondary structure was predicted by PSIPRED 3.4 using default parameters (Jones 1999). Profiles for each sequence were generated with PSI-BLAST (Altschul et al. 1997) using the filtered version of the UniRef90 database as of April 2015. PSIPRED converts profiles of evolutionarily related proteins into secondary structure propensities (helix, strand, or coil), and returns the most probable state for each site. The accuracy of

these predictions in a single sequence has been estimated at roughly 80% when compared with empirical information from the Protein Data Bank (Bernstein et al. 1977; Bryson et al. 2005). All sites that were consistently predicted to have identical secondary structure (either only α-helix or only β-strand across all sequences) were classified as structured. Sites with no predicted secondary structure in any sequence were classified as coils.

Functional domains were predicted using the Pfam database (version 27) (Finn et al. 2014) by aligning each sequence to a hidden Markov Model profile with predefined gathering thresholds. Sites unanimously predicted to fall within Pfam-A domains (based on envelope coordinates) were considered "domain" sites, while those with no predicted Pfam-A domains were considered linker sites.

Statistical Analysis

From all sites that were 100% conserved for at least one structural property (order or disorder, secondary structure or coil, domain or linker), three individual datasets were assembled (table 1). Based on these datasets, the individual effects of disorder, secondary structure, and domain involvement (considering only one factor at a time while ignoring the others) on amino acid replacement rates were evaluated using a nonparametric Mann–Whitney test (fig. 5). For each of the structural factors listed above, exactly two levels were considered. Sites that were 100% conserved for all three structural properties were categorized according to the eight possible factor-level combinations and used as treatment groups. Statistical analysis of these groups was accomplished using a Kruskal–Wallis test followed by post hoc, nonparametric pairwise comparisons from the "pgirmess" package including multiple comparison correction (Siegel and Castellan 1988) in R (R Core Team 2012). In addition, all sites included in the Kruskal–Wallis test

were analyzed under an unbalanced factorial model to evaluate the significance of each factor effect as well as their interaction terms. An unbalanced (type III) multifactor analysis of variance was performed on the model using tools from the "car" library (Fox et al. 2015) in the R programming language (R Core Team 2012).

ACKNOWLEDGMENTS

The authors would like to acknowledge the Instructional & Research Computing Center (IRCC) at Florida International University for providing HPC computing resources that have contributed to the research results reported within this paper, web: http://ircc.fiu.edu.

LITERATURE CITED

Altschul SF Gish W Miller W Myers EW Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 245:403–410.

Altschul SF Madden TL Schäffer AA Zhang J Zhang Z Miller W Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Bellay J Han S Michaut M Kim T Costanzo M Andrews BJ Boone C Bader GD Myers CL Kim PM. 2011. Bringing order to protein disorder through comparative genomics and genetic interactions. Genome Biol. 12:R14.

Benson DA Karsch-Mizrachi I Lipman DJ Ostell J Sayers EW. 2009. GenBank. Nucleic Acids Res. 37:D26–D31.

Bernstein FC Koetzle TF Williams GJ Meyer EF Brice MD Rodgers JR Kennard O Shimanouchi T Tasumi M. 1977. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol. 112:535–542.

Brown CJ Johnson AK Daughdrill GW. 2010. Comparing models of evolution for ordered and disordered proteins. Mol Biol Evol. 27:609–621.

Brown CJ Johnson AK Dunker AK Daughdrill GW. 2011. Evolution and disorder. Curr Opin Struct Biol. 21:441–446.

Brown CJ Takayama S Campen AM Vise P Marshall TW Oldfield CJ Williams CJ Dunker AK. 2002. Evolutionary rate heterogeneity in proteins with long disordered regions. J Mol Evol. 55:104–110.

Bryson K McGuffin LJ Marsden RL Ward JJ Sodhi JS Jones DT. 2005. Protein structure prediction servers at University College London. Nucleic Acids Res. 33:W36–W38.

Dohrmann M Wörheide G. 2013. Novel scenarios of early animal evolution—is it time to rewrite textbooks? Integr Comp Biol. 53:503–511.

Dosztányi Z Csizmók V Tompa P Simon I. 2005. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol. 347:827–839.

Durand D Halldórsson B Vernot B. 2005. A hybrid micro-macroevolutionary approach to gene tree reconstruction. In: Miyano S Mesirov J Kasif S Istrail S Pevzner P Waterman M, editors. Research in computational molecular biology, vol. 3500. Berlin: Springer. p. 250–264.

Echave J Spielman SJ Wilke CO. 2016. Causes of evolutionary rate variation among protein sites. Nat Rev Genet. 17:109–121.

Finn RD Bateman A Clements J Coggill P Eberhardt RY Eddy SR Heger A Hetherington K Holm L Mistry J, et al.. 2014. Pfam: the protein families database. Nucleic Acids Res. 42:D222–D230.

Fox J Weisberg S Adler D Bates D Baud-Bovy G Ellison S Firth D Friendly M Gorjanc G Graves S, et al.. 2015. Package “car” 2.0-25: companion to applied regression.

Franzosa EA Xia Y. 2009. Structural determinants of protein evolution are context-sensitive at the residue level. Mol Biol Evol. 26:2387–2395.

Fuxreiter M Tompa P Simon I. 2007. Local structural disorder imparts plasticity on linear motifs. Bioinformatics 23:950–956.

Gee H. 2008. Evolutionary biology: the amphioxus unleashed. Nature 453:999–1000.

Huang T-T del Valle Marcos ML Hwang J-K Echave J. 2014. A mechanistic stress model of protein evolution accounts for site-specific evolutionary rates and their relationship with packing density and flexibility. BMC Evol Biol. 14:78.

Jones DT. 1999. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 292:195–202.

Jones DT Taylor WR Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 8:275–282.

Katoh K Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30:772–780.

Kimura M Ohta T. 1974. On some principles governing molecular evolution. Proc Natl Acad Sci U S A. 71: 2848–2852.

Mayrose I Graur D Ben-Tal N Pupko T. 2004. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Mol Biol Evol. 21:1781–1791.

Motlagh HN Wrabl JO Li J Hilser VJ. 2014. The ensemble nature of allostery. Nature 508: 331–339.

Perutz MF Kendrew JC Watson HC. 1965. Structure and function of haemoglobin. J Mol Biol. 13:669–678.

Putnam NH Butts T Ferrier DEK Furlong RF Hellsten U Kawashima T Robinson-Rechavi M Shoguchi E Terry A Yu J-K, et al.. 2008. The amphioxus genome and the evolution of the chordate karyotype. Nature 453:1064–1071.

R Core Team. 2012. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Ronquist F Huelsenbeck JP Teslenko M. 2011. mb3.2_manual.pdf. MrBayes version 3.2 Man. Tutorials Model S.

Ronquist F Teslenko M van der Mark P Ayres DL Darling A Höhna S Larget B Liu L Suchard MA Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 61:539–542.

Sayers EW Barrett T Benson DA Bryant SH Canese K Chetvernin V Church DM DiCuccio M Edgar R Federhen S, et al.. 2009. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 37:D5–D15.

Siegel S Castellan NJ. 1988. Non parametric statistics for the behavioural sciences. MacGraw Hill Int. p. 213–214.

Suzek BE Huang H McGarvey P Mazumder R Wu CH. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23:1282–1288.

Thompson JD Linard B Lecompte O Poch O. 2011. A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. PLoS One 6:e18093.

UniProt Consortium 2014. UniProt: a hub for protein information. Nucleic Acids Res. 43:D204–D212.

van der Lee R Buljan M Lang B Weatheritt RJ Daughdrill GW Dunker AK Fuxreiter M Gough J Gsponer J Jones DT, et al.. 2014. Classification of intrinsically disordered regions and proteins. Chem Rev. 114:6589–6631.

Vernot B Stolzer M Goldman A Durand D. 2007. Reconciliation with non-binary species trees. Comput Syst Bioinformatics Conf. 6:441–452.

Xue B Oldfield CJ Dunker AK Uversky VN. 2009. CDF it all: consensus prediction of intrinsically disordered proteins based on various cumulative distribution functions. FEBS Lett. 583:1469–1474.

Yan J Dunker AK Uversky VN Kurgan L. 2016. Molecular recognition features (MoRFs) in three domains of life. Mol Biosyst. 12:697–710.

Yeh S-W Huang T-T Liu J-W Yu S-H Shih C-H Hwang J-K Echave J. 2014. Local packing density is the main structural determinant of the rate of protein sequence evolution at site level. Biomed Res Int. 2014:572409.

Yeh S-W Liu J-W Yu S-H Shih C-H Hwang J-K Echave J. 2014. Site-specific structural constraints on protein sequence evolutionary divergence: local packing density versus solvent exposure. Mol Biol Evol. 31:135–139.

Tables

Table 1: Conserved Property.

Single Factor Group[1]	Sample Size	Mean	Median
Order (O)	3214254	-0.28234	-0.5769
Disorder (D)	993937	0.06282	-0.2548
Structure (s)	1879338	-0.29828	-0.5713
Coil (c)	2664147	-0.06323	-0.3979
Domain (d)	2391699	-0.30871	-0.6098
Linker (l)	2899814	-0.00982	-0.3244
Combinations[2]			
O-s-d	701380	-0.37081	-0.6325
O-s-l	456067	-0.2617	-0.5351
O-c-d	615698	-0.3356	-0.6444
O-c-l	415903	-0.18029	-0.505
D-s-d	36632	-0.46661	-0.7023
D-s-l	75897	-0.35199	-0.6175
D-c-d	88210	-0.18499	-0.5277
D-c-l	579439	0.181966	-0.1033

[a] Single factor group: sites with at least one 100% conserved structural property.

[b] Factor combinations: sites for which all three structural properties are 100% conserved; e.g., D-s-d are sites with conserved disorder (D) AND secondary structure (s) AND domain involvement (d).

Table 2. ANOVA table for 2^3 factorial analysis.

Source	Sum of Squares	df[1]	F	P(>F)[2]
Disorder (D)	320	1	392.8311	<2e-16
Secondary structure (s)	2053	1	2523.6104	<2e-16
Domain (d)	325	1	399.0750	<2e-16
disorder:secondary structure (D:s)	1457	1	1790.6429	<2e-16
disorder:domain (D:d)	1	1	0.8488	0.3569
Secondary structure:domain (s:d)	1189	1	1462.1611	<2e-16
disorder:secondary structure:domain (D:s:d)	695	1	853.8174	<2e-16
Residuals	2415212	2969218		

[1]df: Degrees of freedom

[2]P(>F): p-value calculated from the F statistic

Figure Captions

Figure 1. Scatterplot showing sequence similarity and alignment quality within clusters of various sizes (5–600). *Y* -axis depicts the minimum pairwise sequence identity per group; *x* -axis shows the minimum alignment coverage (sequence length/total alignment length) found in each aligned cluster. Grey rectangle encloses clusters used in phylogenetic analyses. Cluster sizes (the number of sequences in each cluster) are indicated by the shade and size of each point.

Figure 2. Number of species-specific clusters (i.e., containing proteins from only a single species) for each species in the database. Clusters depicted in this plot were included in phylogenetic analyses.

Figure 3. Bar graph showing the distribution of species representation (total number of species found in each cluster) across all clusters with at least five sequences used for phylogenetic analyses. Bars depict the total number of clusters (*y* -axis) containing the indicated number of species (*x* -axis).

Figure 4. Species tree showing the purported evolutionary relationships among the taxa used in this study. Nodes labeled with an asterisk (*) are not supported by the NCBI Common Taxonomy Tree. Bar graph (right) shows the number of proteins sampled from each species. Each bar is divided into sections illustrating the number of sequences found in clusters of various size ranges. All phyla represented in the tree are labeled, along with

several important lower taxonomic groups. Dashed grey line intersects lineages thought to be present during the Cambrian Explosion.

Figure 5. Violin plots showing distributions of normalized evolutionary rates for (*A*) ordered ($n = 3{,}214{,}254$) versus disordered ($n = 993{,}937$) sites, (*B*) structured ($n = 1{,}879{,}338$) versus coil ($n = 2{,}664{,}147$) sites, and (*C*) domain ($n = 2{,}391{,}699$) versus linker ($n = 2{,}899{,}814$) sites. Violins indicate the estimated kernel density of each distribution (bandwidth = 0.4). Boxplots are drawn inside each violin with median values indicated as a white dot. *Y* -axis indicates evolutionary rates, normalized as *Z* -scores. (See Methods for details regarding evolutionary rate estimation and normalization.)

Figure 6. Violin plots showing distributions of normalized evolutionary rates for all factor/level combinations considered in the factorial analysis. Factor levels are indicated using three letters separated by hyphens, where the first letter denotes ordered ("O") or disordered ("D") sites, the second letter denotes sites within secondary structures ("s") or coils ("c"), and the third denotes sites in domains ("d") or linker regions ("l"). The upper diagonal of the matrix below the plots indicates whether there is a significant difference between group pairs (dark grey cells are significant at $P < 0.05$, whereas light grey cells are not).

Figure 7. Interaction plots illustrating statistically significant ($P < 2e^{-16}$) interactions between (*A*) disorder propensity versus secondary structure and (*B*) domain involvement versus secondary structure. In both plots, secondary structure is represented as the trace

factor and the *Y* -axis represents mean normalized evolutionary rates. Note the change in slope sign in plot A.

Figures

Figure 1

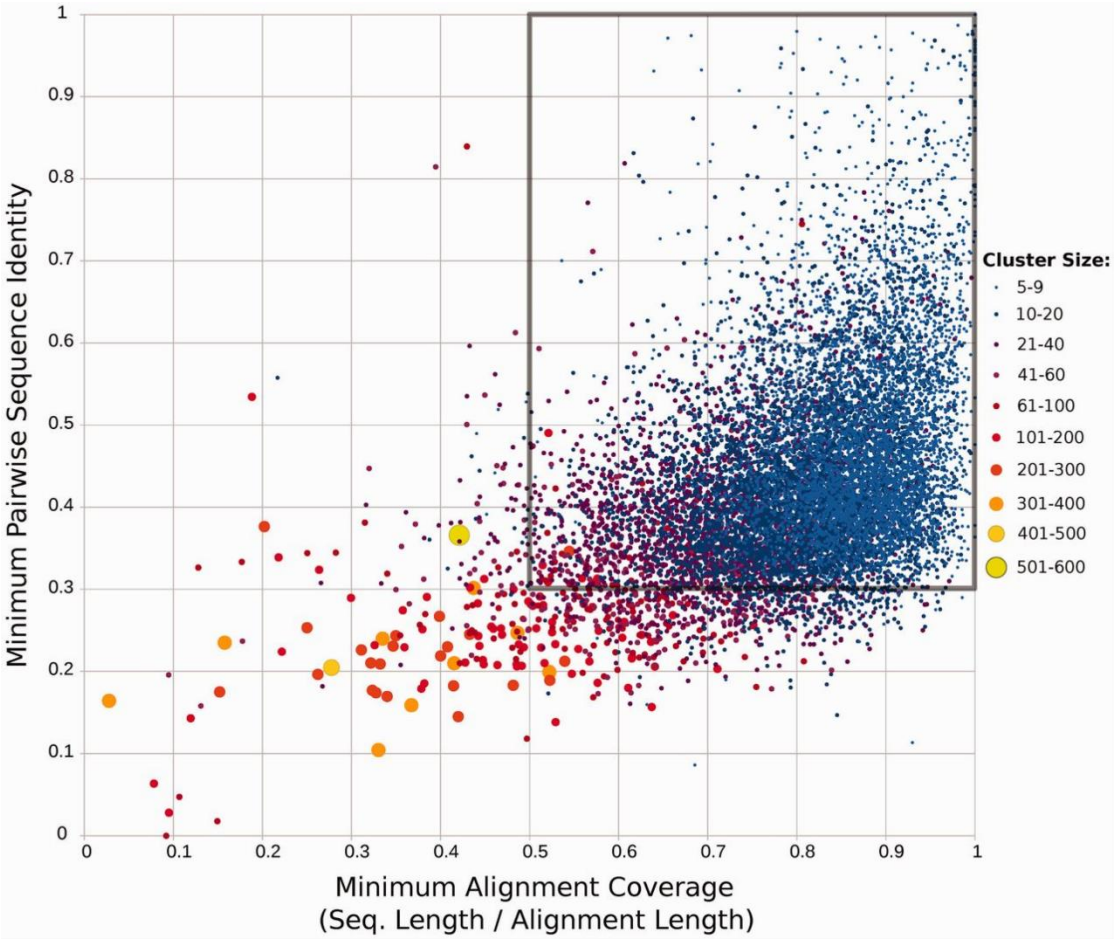


Figure 2

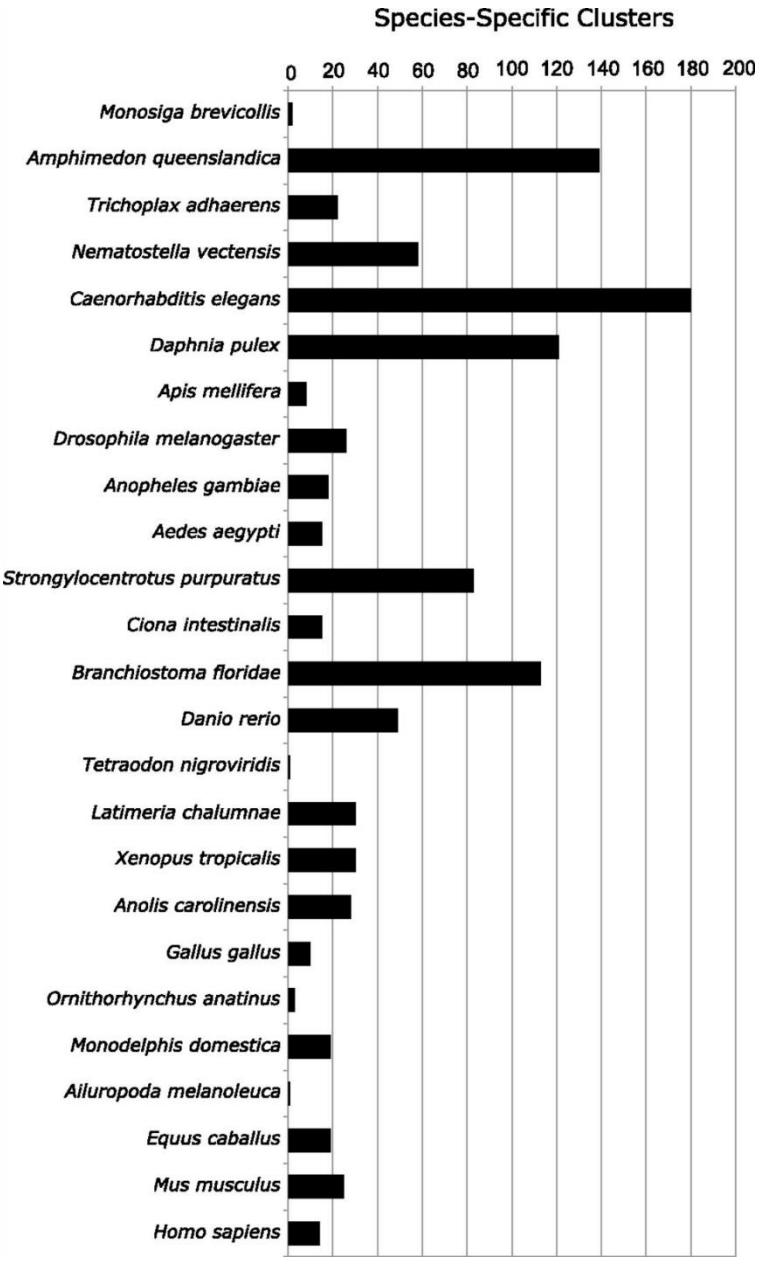

Species-Specific Clusters
0
20
40
60
80
100
120
140
160
180
200
Monosiga brevicollis
Amphimedon queenslandica
Trichoplax adhaerens
Nematostella vectensis
Caenorhabditis elegans
Daphnia pulex
Apis mellifera
Drosophila melanogaster
Anopheles gambiae
Aedes aegypti
Strongylocentrotus purpuratus
Ciona intestinalis
Branchiostoma floridae
Danio rerio
Tetraodon nigroviridis
Latimeria chalumnae
Xenopus tropicalis
Anolis carolinensis
Gallus gallus
Ornithorhynchus anatinus
Monodelphis domestica
Ailuropoda melanoleuca
Equus caballus
Mus musculus
Homo sapiens

Figure 3

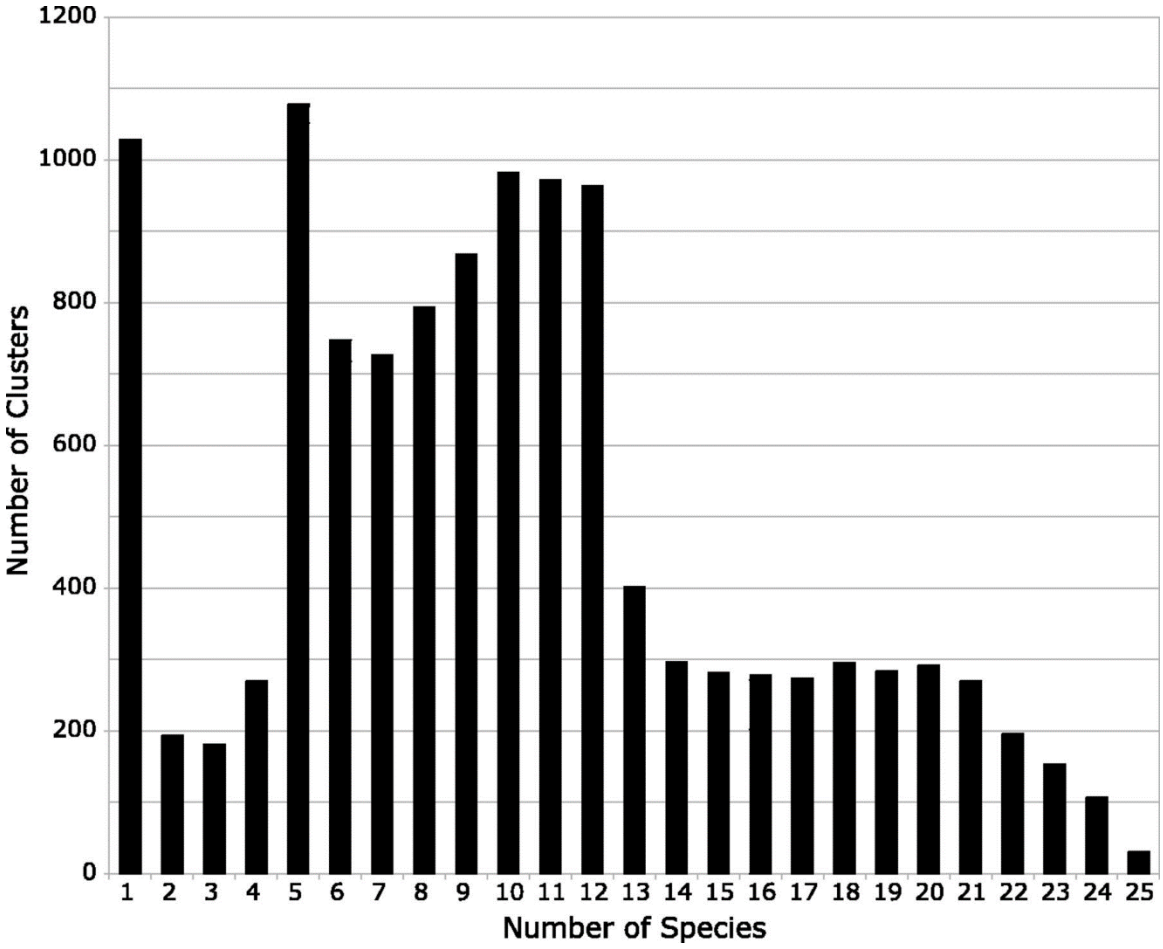

1200
1000
800
600
400
200
0
Number of Clusters
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
Number of Species

Figure 4

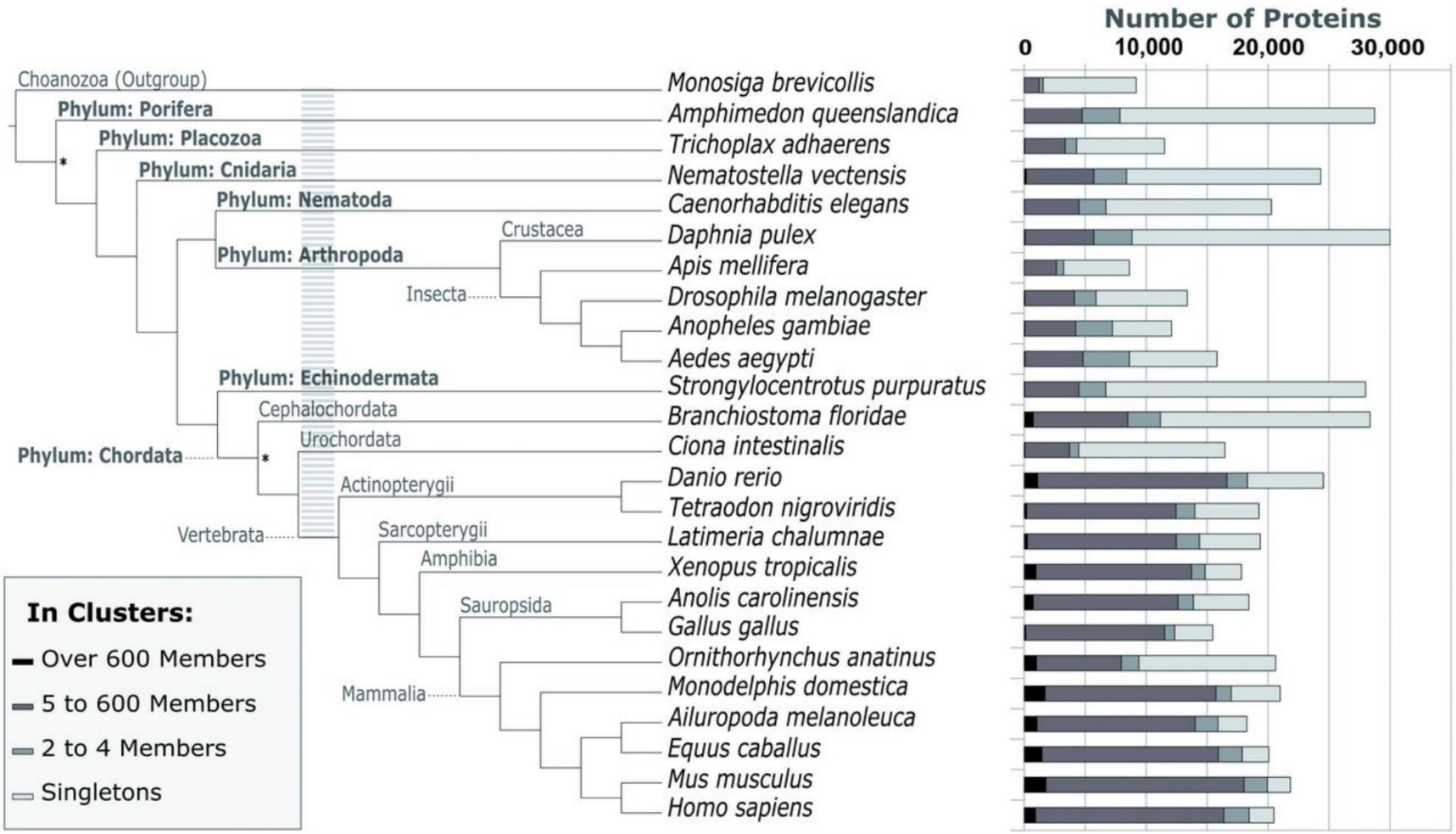

Number of Proteins
0
10,000
20,000
30,000
Choanozoa (Outgroup)
Phylum: Porifera
Phylum: Placozoa
Phylum: Cnidaria
Phylum: Nematoda
Phylum: Arthropoda
Crustacea
Insecta
Phylum: Echinodermata
Cephalochordata
Urochordata
Phylum: Chordata
Actinopterygii
Vertebrata
Sarcopterygii
Amphibia
Sauropsida
Mammalia
Monosiga brevicollis
Amphimedon queenslandica
Trichoplax adhaerens
Nematostella vectensis
Caenorhabditis elegans
Daphnia pulex
Apis mellifera
Drosophila melanogaster
Anopheles gambiae
Aedes aegypti
Strongylocentrotus purpuratus
Branchiostoma floridae
Ciona intestinalis
Danio rerio
Tetraodon nigroviridis
Latimeria chalumnae
Xenopus tropicalis
Anolis carolinensis
Gallus gallus
Ornithorhynchus anatinus
Monodelphis domestica
Ailuropoda melanoleuca
Equus caballus
Mus musculus
Homo sapiens
In Clusters:
Over 600 Members
5 to 600 Members
2 to 4 Members
Singletons

Figure 5

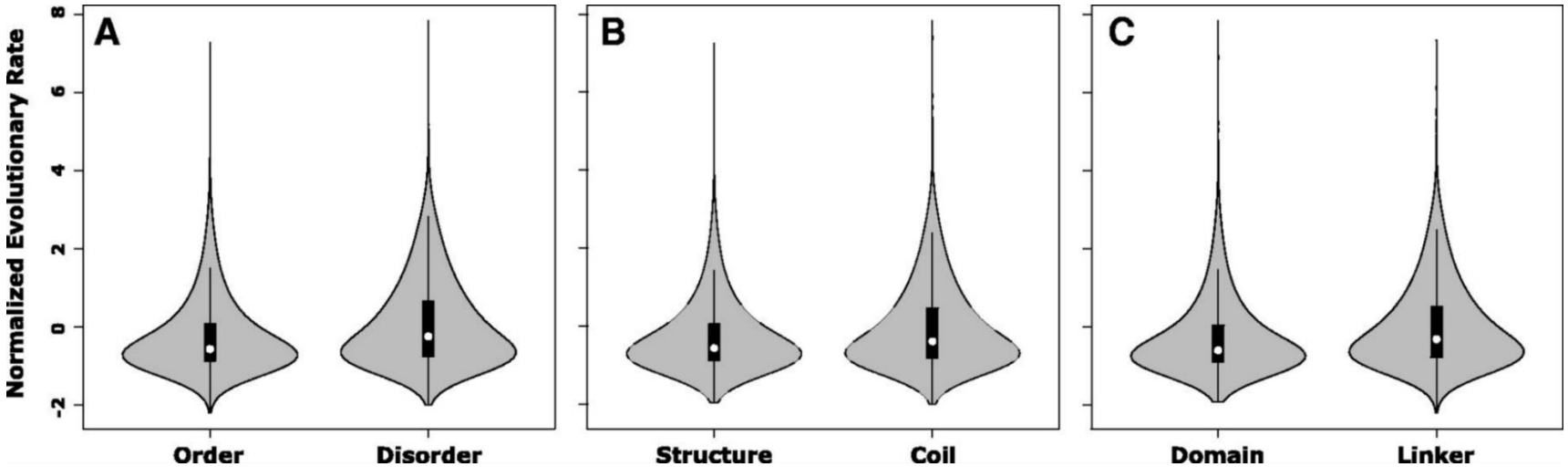

A
B
C
Normalized Evolutionary Rate
8
6
4
2
0
-2
Order
Disorder
Structure
Coil
Domain
Linker

Figure 6

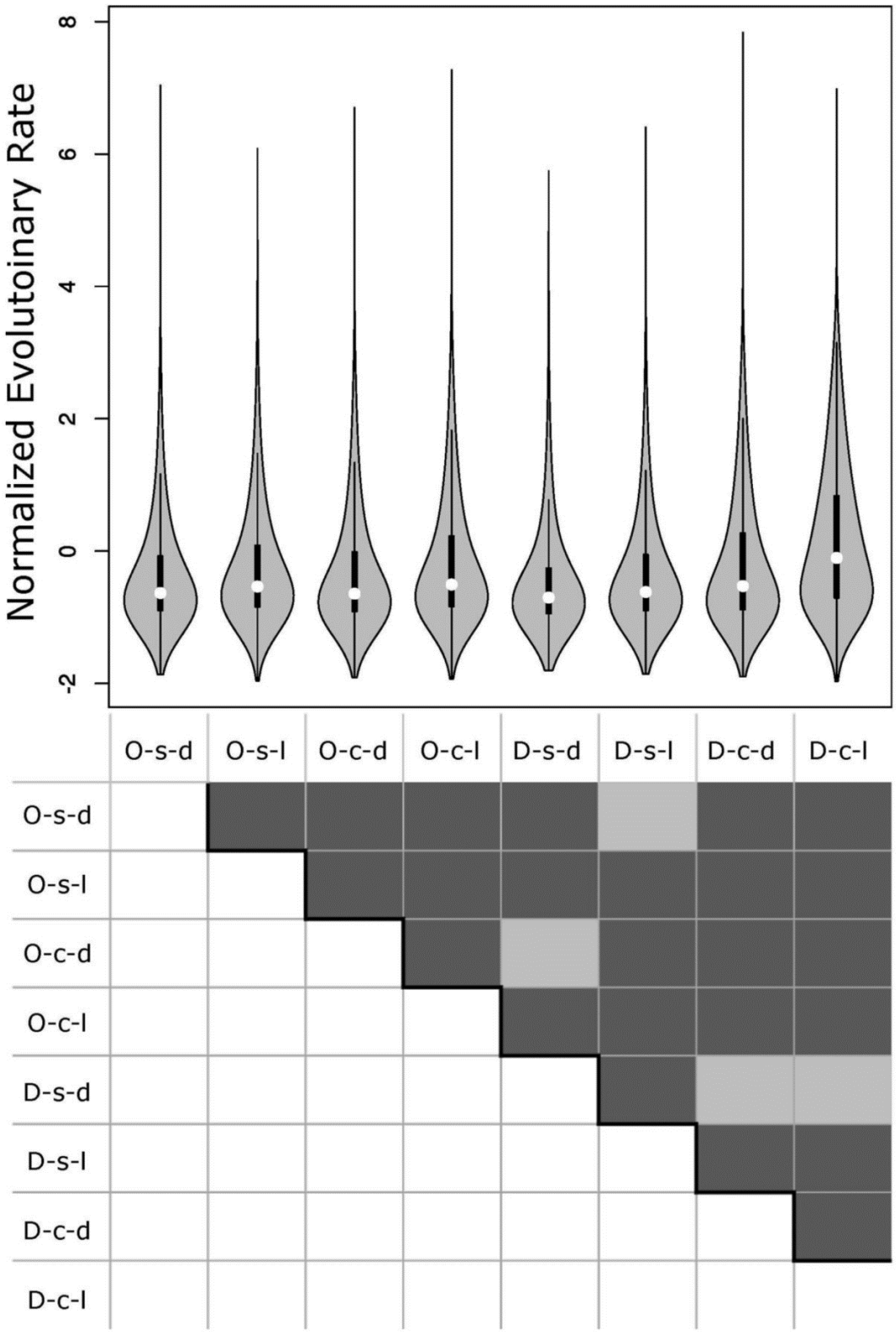
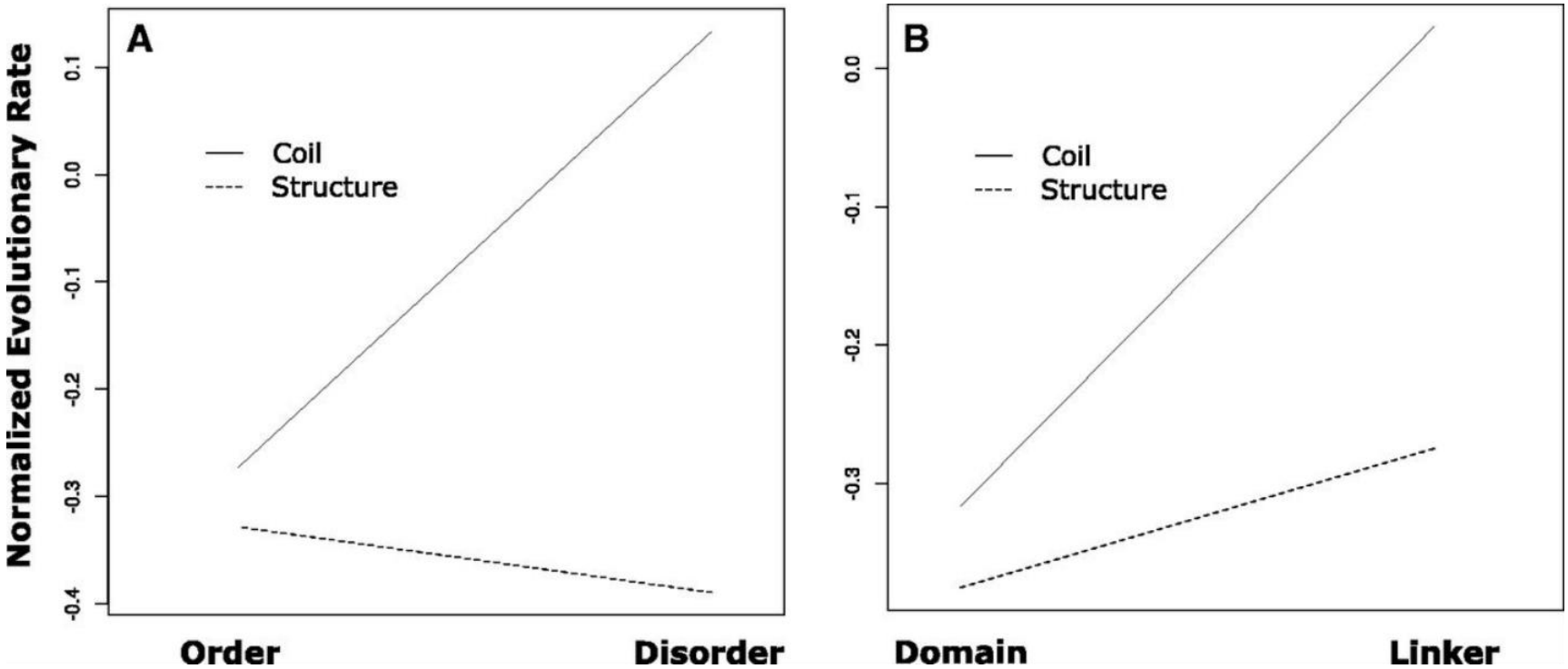

Normalized Evolutoinary Rate
8
6
4
2
0
-2
O-s-d
O-s-l
O-c-d
O-c-l
D-s-d
D-s-l
D-c-d
D-c-l

Figure 7


A
Normalized Evolutionary Rate
0.1
0.0
-0.1
-0.2
-0.3
-0.4
Coil
Structure
Order
Disorder
B
0.0
-0.1
-0.2
-0.3
Coil
Structure
Domain
Linker

CHAPTER III

LARGE-SCALE ANALYSES OF SITE-SPECIFIC EVOLUTIONARY RATES ACROSS EUKARYOTE PROTEOMES REVEAL CONFOUNDING INTERACTIONS BETWEEN INTRINSIC DISORDER, SECONDARY STRUCTURE, AND FUNCTIONAL DOMAINS

ABSTRACT

Various structural and functional constraints govern the evolution of protein sequences. As a result, the relative rates of amino acid replacement among sites within a protein can vary significantly. Previous large-scale work on Metazoan (Animal) protein sequence alignments indicated that amino acid replacement rates are partially driven by a complex interaction among three factors: intrinsic disorder propensity; secondary structure; and functional domain involvement. Here, we use sequence-based predictors to evaluate the effects of these factors on site-specific sequence evolutionary rates within four eukaryotic lineages: Metazoans; Plants; Saccharomycete Fungi; and Alveolate Protists. Our results show broad, consistent trends across all four Eukaryote groups. In all four lineages, there is a significant increase in amino acid replacement rates when comparing: (i) disordered vs. ordered sites; (ii) random coil sites vs. sites in secondary structures; and (iii) inter-domain linker sites vs. sites in functional domains. Additionally, within Metazoans, Plants, and Saccharomycetes, there is a strong confounding interaction between intrinsic disorder and secondary structure—alignment sites exhibiting both high disorder propensity and involvement in secondary structures have very low average rates of sequence evolution. Analysis of gene ontology (GO) terms revealed that in all four lineages, a high fraction of sequences containing these conserved, disordered-structured sites are involved in nucleic acid binding. We also observe notable differences in the statistical trends of Alveolates, where intrinsically disordered sites are more variable than in other Eukaryotes and the statistical interactions between disorder and other factors are less pronounced.

INTRODUCTION

Nucleotide substitutions within protein-coding genes can produce downstream changes (amino acid replacements) within the sequences of their translated expression products (proteins). Consequently, protein molecular evolution entails the replacement of amino acid residues at various positions (sites) within a protein's primary structure (sequence) over time. The relative rates of amino acid replacement may vary significantly among sequence sites, and accounting for rate heterogeneity greatly increases the accuracy of phylogenetic reconstruction based on molecular evolutionary models [1]. This phenomenon has attracted considerable research examining the relationship between protein structure/function and site-specific rates of protein sequence evolution (see Echave et al. [2] for a review).

Several structural and functional properties of proteins are now known to drive overall rates of protein sequence evolution as well as site-specific evolutionary rates within a protein sequence. In particular, sites with a large number of stabilizing contacts (high local packing density) tend to evolve slowly [3,4], and sites with high solvent exposure tend to evolve faster than buried sites [3,5,6]. At the whole-sequence level, there is a strong negative correlation between gene expression level and the rate of protein sequence evolution [7]. Brown et al. [8] also found that proteins with long intrinsically disordered regions (IDRs) tend to experience higher overall levels of amino acid replacement than ordered proteins.

Previously, Ahrens et al. [9] used sequence-based predictors to show that site-specific evolutionary rates in Metazoan (Animal) proteins are partially governed by an interaction among three factors: intrinsic disorder propensity; secondary structure; and

functional domain involvement. A strong statistical interaction was detected between conserved intrinsic disorder and conserved secondary structure, and sites which were predicted to be both intrinsically disordered and involved in secondary structures ("disordered-structured" sites) had lower mean rate scores than any other structural category [9,10].

Here, we present an evaluation of the structural factors studied by Ahrens et al. [9] across large-scale protein sequence datasets representing four eukaryotic lineages: Metazoans; Plants; Saccharomycete Fungi; and Alveolate Protists. We used the sequence-based predictors employed in Ahrens et al. [9] on hundreds of thousands of sequences to identify protein family alignment sites with conserved intrinsic disorder, secondary structure and functional domain predictions, and we applied multifactor statistical analyses to measure the effects of these structural/functional factors on site-specific rates of sequence evolution. Despite the moderate error inherent in structural prediction, our results indicate that there are statistically significant, and broadly consistent forces driving eukaryotic protein evolution. Furthermore, proteins with conserved disordered-structured sequence sites are found in all four Eukaryote lineages and appear to be important for nucleic acid binding, as well as various other fold-upon-binding events.

MATERIALS AND METHODS

2.1. Data Collection

We collected protein sequence data from canonical reference proteomes made available by the UniProt Consortium [11]. These proteomes are useful for evolutionary

analysis because, for alternatively spliced genes, only a single protein isoform is chosen to represent each gene locus. We used this data to construct four large-scale protein datasets containing important model organisms from four divergent eukaryotic lineages: Metazoans (Animals), Plants, Alveolate Protists, and Saccharomycete Fungi (see Appendix 1). To represent Metazoan proteins, we used the 24 Metazoan proteomes (plus the *Monosiga brevicollis* proteome) described in Ahrens et al. [9]. We collected 22 Plant proteomes from the February 2015 release of the UniProt Reference Proteome set, and downloaded two additional proteomes (*Oryza sativa* and *Volox carteri*) directly from UniProt in April of 2016. All of the 44 Alveolate Protist proteomes, as well as the 49 proteomes from Saccharomycete Fungi, were taken from the UniProt Reference Proteome set released in July of 2016. In all four datasets, we excluded any protein sequences that (i) were less than 30 amino acids in length or (ii) contained X characters (indicating missing sequence data) prior to sequence clustering.

2.2. Clustering and Multiple Sequence Alignment

Sequence clustering was accomplished by running the graph-based single-linkage program BLASTClust from BLAST v2.2.26 [12] on each of the four datasets described above. We used two criteria (pairwise sequence identity and sequence overlap) to establish linkage: two sequences were grouped in the same cluster if (i) their pairwise sequence identity was at least 40% and (ii) the length of their BLAST alignment footprint (the region of sequence overlap) was at least 90% the length of the longer sequence. The motivation for this permissive clustering approach was to obtain inclusive clusters of homologous protein sequences that were suitable for multiple sequence alignment and subsequent downstream analyses. Clusters containing between 10 and 300 sequences

were aligned with MAFFT v7.123b (Animals) and v7.313 (Plants, Protists, Fungi) using the local pairwise alignment strategy and a maximum of 1000 iterations [13]. Sequence alignments were used for downstream evolutionary analysis if the following conditions were met: (i) the minimum pairwise sequence identity ($1 - p$-distance) of any two sequences in the alignment was at least 30%; (ii) every sequence was at least 50% the length of the full sequence alignment; (iii) none of the sequences contained ambiguous characters or non-standard amino acids; (iv) less than 90% of alignment sites were conserved (invariant) at the amino acid level; and (v) at least four sequences in each alignment were unique.

2.3. Evolutionary Analysis

We inferred phylogenetic trees using the MPI-enabled version of MrBayes 3.2.2 [14] with tree-bisection-reconnection (TBR) moves disabled. Each analysis used the mixed-model approach (substitution matrix treated as a free parameter) and a four-category gamma distribution among site rates. Analyses were run for 5,000,000 generations, or until the average standard deviation of split frequencies fell below 0.005. Majority-rule consensus trees were constructed for each alignment, discarding the initial 25% of trees as burn-in. To infer site-specific rates of sequence evolution, we used a modified version of the program Rate4site [15] which prints the entire alignment-wide distribution of rate scores rather than only the values associated with a particular reference sequence. Multiple sequence alignments and their associated consensus trees were used as inputs and evaluated under a sixteen-category gamma-distributed model. To more directly measure the values of interest (i.e., the relative site-wise rates of amino acid residue replacement), and in consideration of recent developments in the field [16,17],

site rates were scored based on the equal-probability matrix proposed by Jukes and Cantor [18] rather than the default matrix proposed by Jones et al. [19]. We used the empirical Bayesian method of rate inference implemented in Rate4site, and site rates were normalized as z-scores with mean = 0.0 so that in all alignments, positive scores indicated faster sites while negative scores indicated slower sites.

2.4. Structural Prediction

As in Ahrens et al. [9], we predicted the intrinsic disorder propensity, secondary structure and functional domains of all sequences in each alignment using sequence-based computational tools. Intrinsic disorder propensity was evaluated using the long disorder prediction method implemented in IUPred 1.0 [20]. The accuracy of IUPred-long varies from 62% against DisProt [21] to 85% against IDEAL [22] using the intended cut-off of 0.5 [23]. However, IUPred has greater accuracy against DisProt using a cut-off of 0.4 [24,25]. Here, sequence sites with a propensity score above 0.4 were considered intrinsically disordered, in accordance with previous studies [9,24,25]. Secondary structures (α-helices, β-strands and random coils) were predicted using PSIPRED 3.4 [26] based on sequence profiles generated with PSIBLAST [27] against a filtered version of the UniRef90 database [28]. Previous benchmarks indicate that when based on sequence profiles, PSIPRED predicts secondary structure with >80% accuracy [29,30]. Functional domains were predicted using the Pfam database [31], and all sequence regions outside of functional domains were considered inter-domain linkers. All binary predictions were mapped onto their corresponding protein family alignment sites, and only alignment sites with conserved predictions were considered for statistical analysis.

2.5. Gene Ontology

From each Eukaryote dataset, sequence clusters containing disordered-structured alignment sites (i.e., sites where every sequence in the alignment was predicted to be intrinsically disordered as well as involved in either an α-helix or β-strand) were reserved for gene ontology analysis. Sequences from these alignments corresponding to *Homo sapiens* (Metazoans), *Arabidopsis thaliana* (Plants), *Saccharomyces cerevisiae* (Saccharomycetes) or *Plasmodium falciparum* (Alveolates) were collected and analyzed using the Panther webserver [32,33].

2.6. Statistical Analysis

Each alignment site was labelled based on the predicted structural properties of all sequences in the alignment. A site was called "disordered" if the IUPred score for every sequence at that site was above 0.4, and "ordered" if every score was below 0.4. Similarly, a site was considered "structured" if PSIPRED indicated that either (i) every sequence fell within an alpha helix or (ii) every sequence fell within a beta strand, and it was labelled "coil" if all sequences fell within random coils at that site. Finally, sites were called "domain" sites when all sequences fell within a predicted Pfam domain and "linker" sites when none of them fell within a Pfam domain. Sites containing any number of gap characters were excluded from further evaluation.

All statistical analysis and visualization was performed in the R programming language [34,35] as well as the "matplotlib" module [36] available in the Python programming language [37]. In each of the four eukaryotic datasets, nonparametric Mann-Whitney tests were used to compare normalized rates of sequence evolution observed in ordered vs. disordered sites, structured vs. coil sites, and domain vs. linker

sites found across all sequence alignments. Additionally, based on the above criteria, many alignment sites could be labeled according to all three structural properties (e.g., disordered/coil/linker). Following a Kruskal–Wallis test, nonparametric multiple pairwise significance tests ($\alpha = 0.05$) were performed to compare the rate distributions of all factor-level combinations (e.g., disordered/coil/linker vs. disordered/coil/domain) in all four datasets via the "kruskalmc" method available in the "pgirmess" package [38] in R. Using the "car" package developed by Fox and Weisberg [39], these sites were also incorporated into an unbalanced (type III) factorial analysis of variance (ANOVA) with zero-sum contrasts to evaluate the statistical interaction among intrinsic disorder, secondary structure and functional domain involvement. The relationship between cluster disorder content (fraction of disordered alignment sites) and mean rate scores within disordered-structured alignment sites was analyzed via Loess regression and visualized in the "ggplot2" library [40].

RESULTS

3.1. Clustering and Phylogenetics

Across all four Eukaryote datasets, single-linkage clustering via BlastClust [12] produced 25,871 clusters containing between 10 and 300 sequences (see Appendix 1). After multiple sequence alignment, 22,395 (87%) of these clusters were suitable for downstream phylogenetic inference and site-wise evolutionary rate inference (Figure 1; see Methods: Clustering and Multiple Sequence Alignment for suitability criteria). These sequence alignments contained a total of 14,011,483 sites, of which 9,202,935 (66%)

contained no gap characters. Refer to Table 1 for more information relating to individual datasets.

Nearly all of the 22,395 phylogenetic analyses in MrBayes [14] converged in less than 5,000,000 generations. Only 204 (<1%) of the analyses ran for 5,000,000 generations without reaching an average standard deviation of split frequencies (ASDSF) of less than 0.01, the convergence diagnostic value recommended by the program authors [41], while 21,952 (98%) reached an ASDSF of less than 0.005.

3.2. Structural Prediction

IUPred results [20] indicated that 847,431 of the 9,202,935 gap-free sites were conserved disordered alignment sites (i.e., sites where every sequence in an alignment was intrinsically disordered) and 5,551,255 were conserved ordered sites. Relative to the number of gap-free sites, the percentages of conserved disordered alignment sites in Metazoans (11.6%), Plants (8.2%), Saccharomycetes (6.4%), and Alveolates (9.7%) were consistently low (see Table 1). PSIPRED [26] indicated 3,216,527 conserved structured sites (sites where every sequence fell within either an α-helix or a β-strand) and 3,474,440 conserved coil sites, and Pfam [31] indicated 3,972,117 conserved domain sites and 4,132,983 conserved linker sites. Furthermore, 4,206,014 sites could be consistently labeled according to all three binary factors (e.g., all sequences predicted to be disordered/coil/linker at a particular site), making them suitable for multiple pairwise comparison and factorial ANOVA.

3.3. Statistical Analysis

Mann-Whitney tests indicated that in all four eukaryotic datasets, disordered sites had higher median amino acid replacement rate scores than ordered sites

(Δ_{median_rate} Metazoans: =+0.28, Plants: +0.33, Saccharomycetes: +0.29, Alveolates: +0.75). Similarly, coil sites had higher median rate scores than structured sites (Δ_{median_rate} Metazoans: +0.11, Plants: +0.12, Saccharomycetes: +0.03, Alveolates: +0.15) and linker sites had higher median scores than domain sites (Δ_{median_rate}Metazoans: +0.25, Plants: +0.30, Saccharomycetes: +0.25, Alveolates: +0.27). All median differences in all datasets were highly statistically significant ($p < 2.2 \times 10^{-16}$), but opposing rate distributions (e.g., order vs. disorder) exhibited large overlaps in their range of values (Figure 2). Notably, Mann-Whitney tests considering only sites from clusters where opposing structural properties co-occur (e.g., disordered and ordered sites found within the same alignment) were statistically significant as well ($p< 2.2 \times 10^{-16}$). Kruskal-Wallis tests comparing the eight factor-level combinations were statistically significant in all four datasets ($p < 2.2 \times 10^{-6}$), and most of the 28 post hoc multiple pairwise comparisons were also significant (corrected $p < 0.05$; see Appendix 3).

In addition to statistically significant main effects (all $p < 10^{-5}$), parametric factorial analyses for all four datasets showed statistically significant interaction terms (all $p < 2 \times 10^{-16}$). First-order interactions were particularly large between disorder and secondary structure where the effect of disorder was reversed across three of the four datasets: in Metazoans, Plants, and Saccharomycetes, alignment sites predicted to be both disordered and involved in secondary structures (disordered-structured sites) have lower mean rate scores than ordered, structured sites (Figure 3). A similar phenomenon is observed in the disorder x domain interaction in Plants: disordered sites in functional domains tend to be more conserved than ordered domain sites (Figure 3). Higher-order interactions (disorder × structure × domain) were also detected in all four datasets (all p <

2×10^{-16}). Correlation coefficients (adjusted R^2values) were low in all four models (Metazoans: 0.04, Plants: 0.03, Saccharomycetes: 0.02, Alveolates: 0.06).

Loess regression indicated a negative correlation between sequence evolutionary rates of disordered-structured sites and the overall disorder content (fraction of disordered sites) in their respective alignments (Figure 4). This trend is less pronounced in Alveolate alignments than in the other three datasets.

3.4. Gene Ontology of Proteins with Disordered-Structured Sites

Analysis of GO (gene ontology) terms in PantherDB [32,33] revealed similar patterns in sequences containing conserved disordered-structured sites within all four eukaryotic lineages. Of the GO annotations found for sequences with conserved disordered-structured sites in *Homo sapiens* (Metazoans), *Arabidopsis thaliana* (Plants), *Saccharomyces cerevisiae* (Saccharomycetes), and *Plasmodium falciparum* (Alveolates), the majority had molecular functions associated with binding (53.3%, 43.0%, 40.5%, and 41.8%, respectively) or catalytic activity (30.8%, 39.5%, 39.6%, and 38.8%, respectively). Additionally, the majority of identified biological processes within these four taxa were either cellular processes (30.1%, 35.3%, 35.4%, and 37.6%, respectively) or metabolic processes (23.9%, 34.9%, 32.1%, and 34.3%, respectively) and the majority of associated cellular components were cell parts (38.9%, 42.0%, 40.2%, and 39.6%, respectively), organelles (30.4%, 32.0%, 30.3%, and 30.8%, respectively) and macromolecular complexes (17.2%, 20.2%, 24.4%, and 24.8%, respectively). In all four taxa, a large fraction of protein classes identified for sequences from alignments with conserved disordered-structured sites were nucleotide-binding proteins (24.3%, 32.1%, 37.5%, and 37.4%, respectively) compared to sequences from alignments lacking

conserved disordered-structured sites (11.7%, 15.5%, 17.1%, and 24.8%, respectively). Refer to Appendix 2 for GO term results for sequences with conserved disordered-structured sites from all four representative taxa.

DISCUSSION

4.1. Clustering and Phylogenetics

Previous work by Ahrens et al. [9] highlighted the inherent difficulty of taxon sampling when working with curated molecular datasets—such as the Uniprot Reference Proteome Database [42]—because the bias toward well-studied model organisms is phylogenetically uneven (see Appendix 1). Indeed, there are large percentages of: (i) Vertebrates in the Metazoan dataset (48%); (ii) Angiosperms (flowering Plants) in the Plant dataset (75%); (iii) *Saccharomyces* congeners in the Saccharomycete dataset (20.5%); and (iv) *Plasmodium* congeners in the Alveolate dataset (33%). This phylogenetic unevenness can create downstream biases, wherein the sequence clusters suitable for evolutionary analysis primarily depict relationships among well-represented taxa (Vertebrates, Angiosperms, etc.).

When considering only a single dataset (e.g., Metazoans), it is difficult to determine whether a statistical analysis is biased toward trends in well-represented taxa (e.g., Vertebrates) or truly reflective of more general trends in molecular evolution. By independently analyzing multiple divergent lineages, our statistical results show that there are broad, generally consistent trends across several eukaryotic groups (i.e., in the relationship between structural/functional factors and sequence evolutionary rate) despite the phylogenetic unevenness inherent within the individual datasets.

4.2. Structural Prediction

Previous research has revealed that intrinsic disorder is more prevalent in eukaryotic proteins than either Bacteria or Archaea [43,44,45,46]. Rather than simply acting as flexible linkers, some eukaryotic IDR's occur within functional domains and are crucial to the functions of their associated proteins [47], and many functional IDR's undergo disorder-to-order transitions in the process of binding to neighboring proteins or nucleotide molecules [48]. Thus, the three factors evaluated in this study (intrinsic disorder, secondary structure, functional domains) appear to be intricately connected and overlapping: intrinsic disorder can occur within functional domains, and transient secondary structures may form within IDR's to facilitate interactions with other biomolecules. In this light, the combined results of conserved intrinsic disorder, secondary structure and functional domain predictions in an evolutionary context (i.e., multiple sequence alignment sites) appear to be very useful for detecting biologically important sequence regions within proteins.

While sequence-based predictors are not perfectly accurate, our in-silico assignment of three binary states to individual alignment sites (order/disorder, structure/coil, and domain/linker) allowed us to study a wide range of protein alignments from several eukaryotic lineages, including many alignments containing sequences where experimentally-determined structural data is not available. Our analysis workflow (site rate inference, structural prediction, statistical analysis) was applied consistently, such that data arising from different alignments, and different Eukaryote datasets, are directly comparable. Furthermore, by limiting statistical analyses to only gap-free alignment sites with conserved structural predictions, we avoided many error-prone alignment regions as

well as inconsistent (and possibly inaccurate) structural assignments. Also, evaluating all combinations of the three binary factors inferred by predictors, we have identified an interesting category of evolutionarily conserved alignment sites (i.e., disordered-structured sites). Notably, such an interplay of structural factors cannot be readily identified via publicly-available experimental data from the Protein Data Bank (PDB) [49], since structural assignments are not provided for regions of intrinsic disorder, where electron density is missing.

4.3. Gene Ontology

In prior work on Metazoan protein alignments, Ahrens et al. [9] proposed that disordered-structured sites may be involved in the kinds of disorder-to-order transitions commonly associated with molecular recognition features (MoRFs), wherein the ordered state often adopts secondary structure upon binding to another protein molecule [50,51]. Similar disorder-to-order transitions are important in many nucleic acid binding proteins, especially RNA-binding proteins [52,53,54]. The disorder propensity of these binding regions is thought to confer high specificity, while still allowing binding partners to easily dissociate when necessary [52].

Based on protein class GO terms in our four reference taxa, a large percentage of sequences containing conserved disordered-structured sites are in fact nucleic acid binding proteins (see Appendix 2). Interestingly, a large number of hydrolase proteins also had conserved disordered-structured sites, and there is evidence that some hydrolases rely directly on intrinsic disorder to function. Ubiquitin C-terminal hydrolase activity, for example, is mediated by a disorder-to-order transition within its active site [55,56].

The low amino acid replacement rates we observed in disordered-structured sites suggest selective constraint, likely resulting from the functional importance of transient secondary structure within regions of many eukaryotic proteins [51,52,53,54]. Hence, the joint output of intrinsic disorder and secondary structure predictors in a conserved evolutionary context (i.e., consistent predictions across multiple related sequences) may be useful for identifying protein sites where transitions between disorder and secondary structure are required for protein function.

4.4. Intrinsic Disorder in Alveolates

Other researchers have observed that the proteomes of many Alveolate Protists, particularly multi-host pathogens in the clade Apicomplexa, possess a high abundance of proteins with long disordered regions [56,57] and a high fraction of disordered residues in general [46]. Mohan et al. [57] predicted long disordered regions (>30 residues) in most of the protein sequences from the Apicomplexan pathogens *Toxoplasma gondii* (87.8–89.8%) as well as members of the genus *Plasmodium* (75.3–82.5%). Pancsa and Tompa [46] showed that the overall percentage of disordered sites within *T. gondii* proteins was higher than any of the other 193 Eukaryotes they examined, and the disorder percentages of *Plasmodium* spp. proteins were more similar to those of multicellular Eukaryotes (Metazoans, Plants, and Fungi) than other Alveolates. Among the alignment sites containing no gap characters, we observed percentages of conserved disordered sites (6.4–11.6%) that were markedly lower than the overall percentages reported in previous studies [46,57]. Such a disparity is expected, though, since the total number of disordered sites in a given protein sequence exceeds the number of sites with conserved disorder across several related sequences [58].

In the case of membrane and secreted proteins, intrinsic disorder in Apicomplexan parasites has a potential dual function: (i) the reduction of antibody binding affinity and (ii) the facilitation of promiscuous attachment to various host cells [59]. Many potential vaccine targets in *Plasmodium* are intrinsically disordered [60], and the erythrocyte binding-like proteins in P. falciparum appear to lack transient secondary structures even when recognizing and binding to cell surface receptors during host invasion [61]. Our results indicate that disordered sites in Alveolate proteins also experience higher amino acid replacement rates than other Eukaryotes, and disordered-structured sites in Alveolates are less conserved at the sequence level than in Metazoans, Plants, or Saccharomycetes (Figure 2, Figure 3 and Figure 4). However, recent work has shown that increased rates of protein sequence evolution in disordered regions can result from high positive selection (i.e., an increase in non-synonymous nucleotide substitutions) rather than relaxed purifying selection [62], so the relatively high replacement rates we observed in Alveolate disordered sites may actually be driven by increased pressure for innovation to avoid host recognition and/or to make novel host interactions. Ultimately, these results suggest that developing effective drugs and vaccines targeting Apicomplexan parasites could prove especially difficult, and require a deeper understanding of drug interactions within disordered protein regions.

4.5. Statistical Analysis

Across four large-scale molecular datasets, spanning four divergent eukaryotic lineages (Animals, Plants, Fungi, and Protists), we found mostly consistent, statistically significant relationships between three structural/functional factors and site-specific rates of amino acid replacement. By using the equal-probability model from Jukes and Cantor

[18] to evaluate rate scores, our results merit a natural, intuitive interpretation—intrinsically disordered sequence sites are more variable than ordered sites, sites in random coils are more variable than sites within secondary structures, and sites in inter-domain linkers are more variable than sites in functional domains. Furthermore, factorial ANOVA indicated widespread confounding interactions among all pairwise combinations of the three factors we tested, as well as significant higher-order interactions beyond what can be observed in trace plots (Figure 3). In fact, the least significant (i.e., highest) p-value observed in any factorial ANOVA corresponded to a main effect term (intrinsic disorder in Plants: $p = 4.13 \times 10^{-6}$), while all other terms across all analyses were highly significant ($p < 2.2 \times 10^{-16}$). Nonetheless, the first-order interactions appear to follow largely similar patterns in each dataset. One notable exception is the disorder x structure interaction in Alveolates which, although statistically significant, lacks the sign reversal observed in the other three lineages (i.e., disordered-structured sites are more variable on average than ordered, structured sites). Additionally, the disorder x domain interaction seen in Plant sites, where disordered sites within domains tend to be more conserved than ordered domain sites, is less pronounced (but still significant) in the other datasets.

Importantly, the statistical significance of these results (indicated by p-values) is consistently high, but the predictive power of the associated factorial models (indicated by correlation coefficients) is consistently low. The residual variance contributing to low model fit can also be seen in the large amount of overlap between the opposing distributions of rate scores (order vs. disorder, structure vs. coil, and domain vs. linker) in every dataset (Figure 2). Hence, it is appropriate to conclude based on our results that

ordered sites, for instance, tend to evolve more slowly than disordered sites, but the likelihood that a particular conserved site is ordered is not necessarily high, and said likelihood clearly depends on additional site-specific factors as well (i.e., secondary structure and functional domain involvement). Future large-scale analyses incorporating additional structural factors (e.g., relative solvent exposure) may detect stronger statistical interactions with higher correlations to amino acid replacement rates.

The negative correlation between alignment disorder content (the fraction of disordered sites in an aligned sequence cluster) and the mean relative rate scores of disordered-structured sites within a given alignment suggests that latent structural factors at the sequence level also govern observed rates of amino acid replacement (Figure 4). Such effects are likely nontrivial, considering the unbalanced nature of the site-wise factors discussed here. The prevalence of disordered-structured sites is generally low compared to ordered, structured sites or disordered random coils, and many protein sequences essentially lack intrinsic disorder entirely. Joint analysis of several sequence-level and site-level factors (e.g., via hierarchical linear modelling) may provide deeper insight into the forces driving amino acid replacement.

The complex network of structural and functional properties governing protein (and therefore gene) sequence evolution is a topic of active research [2,63]. To this end, previous work on intrinsic disorder has uncovered similar trends regarding protein sequence conservation [8,9], and much stronger correlations between other protein structural properties and sequence evolutionary rate (e.g., contact number and packing density) have also been observed [2,3,4,64]. Nonetheless, to our knowledge, the results described here represent the most comprehensive evidence for widespread, large-scale

structural and functional drivers of eukaryotic sequence evolution to date (Appendix 1 [65,66]). Furthermore, they reinforce the notion that several factors interact, often in subtle ways, to influence molecular evolution.

FUNDING

This work was supported by a Dissertation Year Fellowship to J.B.A. from Florida International University.

ACKNOWLEDGMENTS

The authors would like to acknowledge the Instructional & Research Computing Center (IRCC) at Florida International University for providing HPC computing resources that have contributed to the research results reported within this paper, web: http://ircc.fiu.edu. We are also thankful to Janelle Nunez-Castilla and Luis Nassar for additional computational support during the completion of this study.

LITERATURE CITED

1. Yang, Z.; Kumar, S. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. *Mol. Biol. Evol.* **1996**, *13*, 650–659.
2. Echave, J.; Spielman, S.J.; Wilke, C.O. Causes of evolutionary rate variation among protein sites. *Nat. Rev. Genet.* **2016**, *17*, 109–121.
3. Franzosa, E.A.; Xia, Y. Structural determinants of protein evolution are context-sensitive at the residue level. *Mol. Biol. Evol.* **2009**, *26*, 2387–2395.
4. Yeh, S.-W.; Huang, T.-T.; Liu, J.-W.; Yu, S.-H.; Shih, C.-H.; Hwang, J.-K.; Echave, J. Local packing density is the main structural determinant of the rate of protein sequence evolution at site level. *BioMed Res. Int.* **2014**, *2014*, 572409.
5. Perutz, M.F.; Kendrew, J.C.; Watson, H.C. Structure and function of haemoglobin. *J. Mol. Biol.* **1965**, *13*, 669–678.

6. Kimura, M.; Ohta, T. On some principles governing molecular evolution. *Proc. Natl. Acad. Sci. USA* **1974**, *71*, 2848–2852.

7. Zhang, J.; Yang, J.-R. Determinants of the rate of protein sequence evolution. *Nat. Rev. Genet.* **2015**, *16*, 409–420.

8. Brown, C.J.; Takayama, S.; Campen, A.M.; Vise, P.; Marshall, T.W.; Oldfield, C.J.; Williams, C.J.; Dunker, A.K. Evolutionary rate heterogeneity in proteins with long disordered regions. *J. Mol. Evol.* **2002**, *55*, 104–110.

9. Ahrens, J.; Dos Santos, H.G.; Siltberg-Liberles, J. The nuanced interplay of intrinsic disorder and other structural properties driving protein evolution. *Mol. Biol. Evol.* **2016**, *33*, 2248–2256.

10. Ahrens, J.B.; Nunez-Castilla, J.; Siltberg-Liberles, J. Evolution of intrinsic disorder in eukaryotic proteins. *Cell. Mol. Life Sci.* **2017**, *74*, 3163–3174.

11. Bateman, A.; Martin, M.J.; O'Donovan, C.; Magrane, M.; Alpi, E.; Antunes, R.; Bely, B.; Bingley, M.; Bonilla, C.; Britto, R.; Bursteinas, B.; et al. UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **2017**, *45*, D158–D169.

12. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410.

13. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780.

14. Ronquist, F.; Teslenko, M.; van der Mark, P.; Ayres, D.L.; Darling, A.; Höhna, S.; Larget, B.; Liu, L.; Suchard, M.A.; Huelsenbeck, J.P. MrBayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **2012**, *61*, 539–542.

15. Mayrose, I.; Graur, D.; Ben-Tal, N.; Pupko, T. Comparison of site-specific rate-inference methods for protein sequences: Empirical Bayesian methods are superior. *Mol. Biol. Evol.* **2004**, *21*, 1781–1791.

16. Spielman, S.J.; Kosakovsky Pond, S.L.; Yeager, M. Relative evolutionary rates in proteins are largely insensitive to the substitution model. *Mol. Biol. Evol.***2018**, *35*, 2307–2317.

17. Sydykova, D.K.; Wilke, C.O. Theory of measurement for site-specific evolutionary rates in amino-acid sequences. *bioRxiv* **2018**, 411025.

18. Jukes, T.H.; Cantor, C.R. Evolution of protein molecules. *Mamm. Protein Metab.* **1969**, *3*, 21–132.

19. Jones, D.T.; Taylor, W.R.; Thornton, J.M. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **1992**, *8*, 275–282.

20. Dosztányi, Z.; Csizmók, V.; Tompa, P.; Simon, I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.* **2005**, *347*, 827–839.

21. Sickmeier, M.; Hamilton, J.A.; LeGall, T.; Vacic, V.; Cortese, M.S.; Tantos, A.; Szabo, B.; Tompa, P.; Chen, J.; Uversky, V.N.; et al. DisProt: The database of disordered proteins. *Nucleic Acids Res.* **2007**, *35*, D786–D793.

22. Fukuchi, S.; Sakamoto, S.; Nobe, Y.; Murakami, S.D.; Amemiya, T.; Hosoda, K.; Koike, R.; Hiroaki, H.; Ota, M. IDEAL: Intrinsically disordered proteins with extensive annotations and literature. *Nucleic Acids Res.* **2012**, *40*, D507–D511.

23. Di Domenico, T.; Walsh, I.; Tosatto, S.C. Analysis and consensus of currently available intrinsic protein disorder annotation sources in the MobiDB database. *BMC Bioinform.* **2013**, *14*, S3.

24. Fuxreiter, M.; Tompa, P.; Simon, I. Local structural disorder imparts plasticity on linear motifs. *Bioinformatics* **2007**, *23*, 950–956.

25. Xue, B.; Oldfield, C.J.; Dunker, A.K.; Uversky, V.N. CDF it all: Consensus prediction of intrinsically disordered proteins based on various cumulative distribution functions. *FEBS Lett.* **2009**, *583*, 1469–1474.

26. Jones, D.T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **1999**, *292*, 195–202.

27. Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402.

28. Suzek, B.E.; Huang, H.; McGarvey, P.; Mazumder, R.; Wu, C.H. UniRef: Comprehensive and non-redundant UniProt reference clusters. *Bioinformatics***2007**, *23*, 1282–1288.

29. Drozdetskiy, A.; Cole, C.; Procter, J.; Barton, G.J. JPred4: A protein secondary structure prediction server. *Nucleic Acids Res.* **2015**, *43*, W389–W394.

30. Buchan, D.W.A.; Ward, S.M.; Lobley, A.E.; Nugent, T.C.O.; Bryson, K.; Jones, D.T. Protein annotation and modelling servers at University College London. *Nucleic Acids Res.* **2010**, *38*, W563–W568.

31. Finn, R.D.; Bateman, A.; Clements, J.; Coggill, P.; Eberhardt, R.Y.; Eddy, S.R.; Heger, A.; Hetherington, K.; Holm, L.; Mistry, J.; et al. Pfam: The protein families database. *Nucleic Acids Res.* **2014**, *42*, D222–D230.

32. Thomas, P.D.; Campbell, M.J.; Kejariwal, A.; Mi, H.; Karlak, B.; Daverman, R.; Diemer, K.; Muruganujan, A.; Narechania, A. PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res.* **2003**, *13*, 2129–2141.

33. Mi, H.; Dong, Q.; Muruganujan, A.; Gaudet, P.; Lewis, S.; Thomas, P.D. PANTHER version 7: Improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res.* **2010**, *38*, D204–D210.

34. Ihaka, R.; Gentleman, R. R: A Language for data analysis and graphics. *J. Comput. Graph. Stat.* **1996**, *5*, 299–314.

35. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Development Core Team: Vienna, Austria, 2011; Volume 1, p. 409.

36. Hunter, J.D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95.

37. Rossum, G. *Python Reference Manual*; Centrum voor Wiskunde en Informatica (CWI): Amsterdam, The Netherlands, 1995.

38. Siegel, S.; Castellan, N.J. *Nonparametric Statistics for the Behavioral Sciences*; McGraw-Hill: New York, NY, USA, 1988; ISBN 0070573573.

39. Fox, J.; Weisberg, S. *An R Companion to Applied Regression*, 2nd ed.; Sage: Thousand Oaks, CA, USA, 2011.

40. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*; Springer: Berlin/Heidelberg, Germany, 2009; ISBN 9780387981413.

41. Ronquist, F.; Huelsenbeck, J.P.; Teslenko, M. MrBayes Version 3.2 Manual: Tutorials and Model Summaries. 2011. Available online: mrbayes.sourceforge.net/mb3.2_manual.pdf (accessed on 19 October 2018).

42. The UniProt Consortium UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **2017**, *45*, D158–D169.

43. Dunker, A.K.; Obradovic, Z.; Romero, P.; Garner, E.C.; Brown, C.J. Intrinsic protein disorder in complete genomes. *Genome Inform.* **2000**, *11*, 161–171.

44. Ward, J.J.; Sodhi, J.S.; McGuffin, L.J.; Buxton, B.F.; Jones, D.T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **2004**, *337*, 635–645.

45. Xue, B.; Dunker, A.K.; Uversky, V.N. Orderly order in protein intrinsic disorder distribution: Disorder in 3500 proteomes from viruses and the three domains of life. *J. Biomol. Struct. Dyn.* **2012**, *30*, 137–149.

46. Pancsa, R.; Tompa, P. Structural disorder in eukaryotes. *PLoS ONE* **2012**, *7*, e34687.

47. Chen, J.W.; Romero, P.; Uversky, V.N.; Dunker, A.K. Conservation of intrinsic disorder in protein domains and families: II. functions of conserved disorder. *J. Proteome Res.* **2006**, *5*, 888–898.

48. Iakoucheva, L.M.; Brown, C.J.; Lawson, J.D.; Obradović, Z.; Dunker, A.K. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.***2002**, *323*, 573–584.

49. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

50. Yan, J.; Dunker, A.K.; Uversky, V.N.; Kurgan, L. Molecular recognition features (MoRFs) in three domains of life. *Mol. Biosyst.* **2016**, *12*, 697–710.

51. Mohan, A.; Oldfield, C.J.; Radivojac, P.; Vacic, V.; Cortese, M.S.; Dunker, A.K.; Uversky, V.N. Analysis of molecular recognition features (MoRFs). *J. Mol. Biol.***2006**, *362*, 1043–1059.

52. Dyson, H.J. Roles of intrinsic disorder in protein-nucleic acid interactions. *Mol. Biosyst.* **2012**, *8*, 97–104.

53. Varadi, M.; Zsolyomi, F.; Guharoy, M.; Tompa, P. Functional advantages of conserved intrinsic disorder in RNA-Binding proteins. *PLoS ONE* **2015**, *10*, e0139731.

54. Wang, C.; Uversky, V.N.; Kurgan, L. Disordered nucleiome: Abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea. *Proteomics* **2016**, *16*, 1486–1498.

55. Misaghi, S.; Galardy, P.J.; Meester, W.J.N.; Ovaa, H.; Ploegh, H.L.; Gaudet, R. Structure of the ubiquitin hydrolase UCH-L3 complexed with a suicide substrate. *J. Biol. Chem.* **2005**, *280*, 1512–1520.

56. Fong, J.H.; Shoemaker, B.A.; Garbuzynskiy, S.O.; Lobanov, M.Y.; Galzitskaya, O.V.; Panchenko, A.R. Intrinsic disorder in protein interactions: Insights from a comprehensive structural analysis. *PLoS Comput. Biol.* **2009**, *5*, e1000316.

57. Mohan, A.; Sullivan, W.J., Jr.; Radivojac, P.; Dunker, A.K.; Uversky, V.N. Intrinsic disorder in pathogenic and non-pathogenic microbes: Discovering and analyzing the unfoldomes of early-branching eukaryotes. *Mol. Biosyst.* **2008**, *4*, 328–340.

58. Bellay, J.; Han, S.; Michaut, M.; Kim, T.; Costanzo, M.; Andrews, B.J.; Boone, C.; Bader, G.D.; Myers, C.L.; Kim, P.M. Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome Biol.* **2011**, *12*, R14.

59. Feng, Z.-P.; Zhang, X.; Han, P.; Arora, N.; Anders, R.F.; Norton, R.S. Abundance of intrinsically unstructured proteins in *P. falciparum* and other apicomplexan parasite proteomes. *Mol. Biochem. Parasitol.* **2006**, *150*, 256–267.

60. Guy, A.J.; Irani, V.; MacRaild, C.A.; Anders, R.F.; Norton, R.S.; Beeson, J.G.; Richards, J.S.; Ramsland, P.A. Insights into the immunological properties of intrinsically disordered malaria proteins using proteome scale predictions. *PLoS ONE* **2015**, *10*, e0141729.

61. Blanc, M.; Coetzer, T.L.; Blackledge, M.; Haertlein, M.; Mitchell, E.P.; Forsyth, V.T.; Jensen, M.R. Intrinsic disorder within the erythrocyte binding-like proteins from *Plasmodium falciparum*. *Biochim. Biophys. Acta Proteins Proteom.* **2014**, *1844*, 2306–2314.

62. Afanasyeva, A.; Bockwoldt, M.; Cooney, C.R.; Heiland, I.; Gossmann, T.I. Human long intrinsically disordered protein regions are frequent targets of positive selection. *Genome Res.* **2018**, *28*, 975–982.

63. Siltberg-Liberles, J.; Grahnen, J.A.; Liberles, D.A. The evolution of protein structures and structural ensembles under functional constraint. *Genes* **2011**, *2*, 748–762.

64. Yeh, S.-W.; Liu, J.-W.; Yu, S.-H.; Shih, C.-H.; Hwang, J.-K.; Echave, J. Site-specific structural constraints on protein sequence evolutionary divergence: Local packing density versus solvent exposure. *Mol. Biol. Evol.* **2014**, *31*, 135–139.

65. Sayers, E.W.; Barrett, T.; Benson, D.A.; Bryant, S.H.; Canese, K.; Chetvernin, V.; Church, D.M.; DiCuccio, M.; Edgar, R.; Federhen, S.; et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2009**, *37*, D5–D15.

66. Benson, D.A.; Karsch-Mizrachi, I.; Lipman, D.J.; Ostell, J.; Sayers, E.W. GenBank. *Nucleic Acids Res.* **2009**, *37*, D26–D31.

Tables

Table 1: Dataset-specific information for nonparametric analysis.

Dataset	Metazoans	Plants	Saccharomycetes	Alveolates
Clusters	6938	8266	4494	2697
Sequences	130632	198081	122132	44060
Total Alignment Sites	4677490	4703587	2990109	1640297
Gap-free sites	3217225	2851827	1954761	1179122
Ordered Sites	1819695	1706275	1223656	801629
Disordered Sites	373639	234853	125047	113892
Structured sites	1062380	1014001	722444	417702
Random coil sites	1314563	1064725	670357	424795
Domain sites	1436746	1175745	936813	422813
Linker sites	1368702	1289830	817371	657080
Median Order Rate	−0.599	−0.625	−0.6188	−0.605
Median Disorder Rate	−0.3155	−0.2916	−0.3271	0.1426
Median Structure Rate	−0.5787	−0.6262	−0.5935	−0.605
Median Coil Rate	−0.4682	−0.5013	−0.5603	−0.4542
Median Domain Rate	−0.62345	−0.6679	−0.6353	−0.629
Median Linker Rate	−0.3698	−0.3718	−0.3902	−0.3569

Figure Captions

Figure 1. Scatterplots showing minimum pairwise sequence identity (fraction of matching aligned characters) and minimum alignment coverage (seq. length/alignment length) for all Metazoan, Plant, Saccharomycete, and Alveolate clusters used in analyses.

Figure 2. Split violin plots showing differences in normalized site-specific rates of amino acid replacement in: (**a**) ordered vs. disordered sites; (**b**) structured vs. coil sites; and (**c**) domain vs. linker sites within four eukaryotic datasets. Middle dashed lines indicate medians and outer dashed lines indicate quartiles.

Figure 3. Trace plots illustrating first-order interactions among all site-wise binary factor levels: order (Order) and intrinsic disorder (Disorder), secondary structures (Structure) and random coils (Coil), functional domains (Domain) and interdomain linkers (Linker). Trace factors (solid vs. dashed lines) are indicated to the right of each row of plots. Vertical columns of plots correspond to each of the four datasets (indicated) above. Y-axes represent mean normalized evolutionary rates.

Figure 4. Scatterplot showing the disorder content of clusters (fraction of disordered alignment sites) against the mean rate of sequence evolution among sites predicted to be both disordered and structured. Only sequence clusters containing disordered/structured sites are shown. Trend lines were constructed for each of the four eukaryotic datasets using Loess regression. Note that the Alveolate trend line (dashed) is consistently higher than other lineages.

Figures

Figure 1

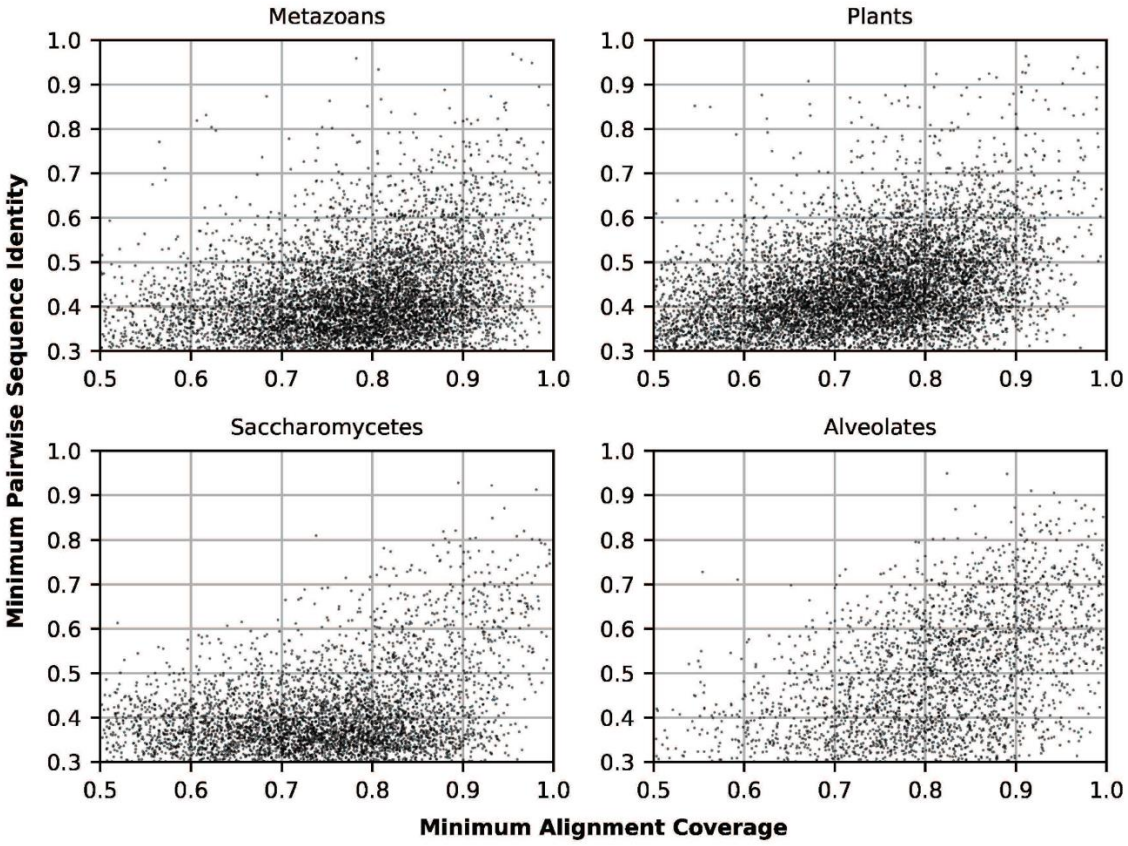


Figure 2

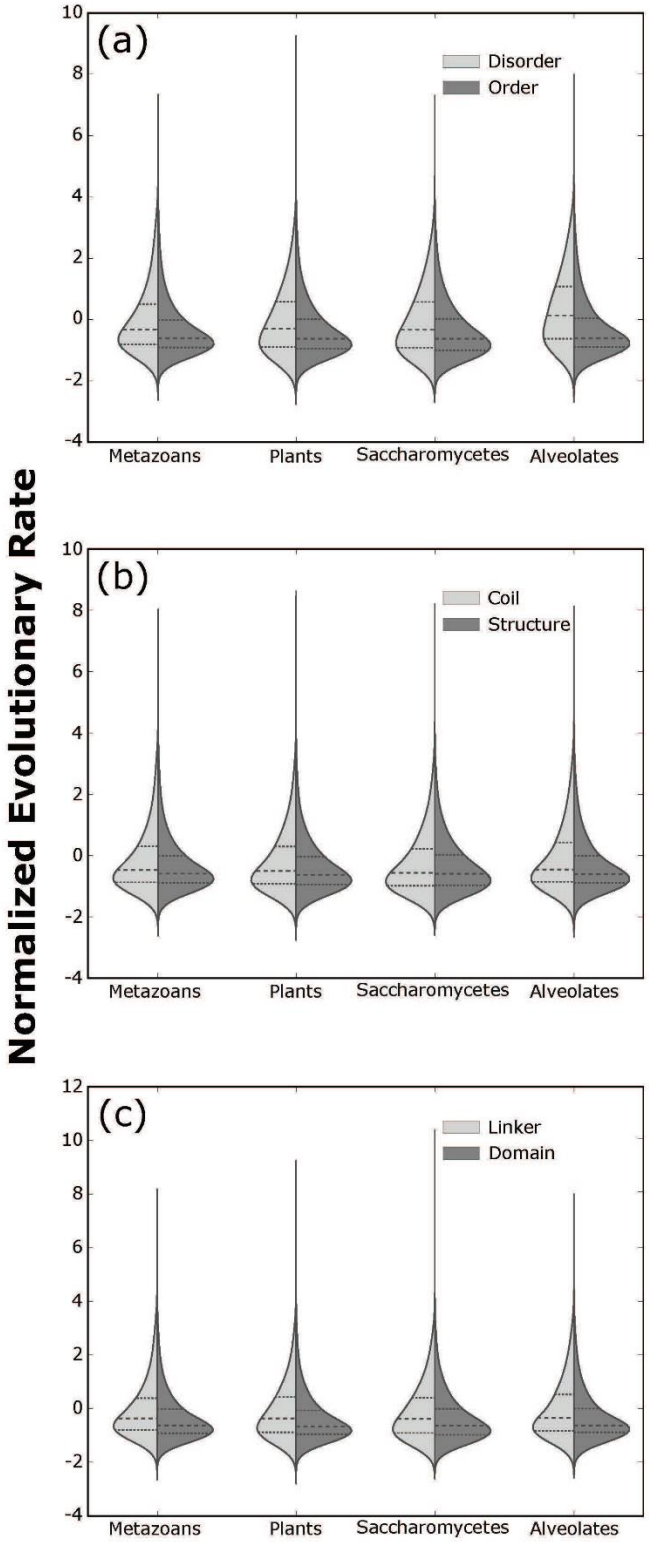

(a)
Disorder
Order
(b)
Coil
Structure
(c)
Linker
Domain
Normalized Evolutionary Rate
Metazoans
Plants
Saccharomycetes
Alveolates

Figure 3

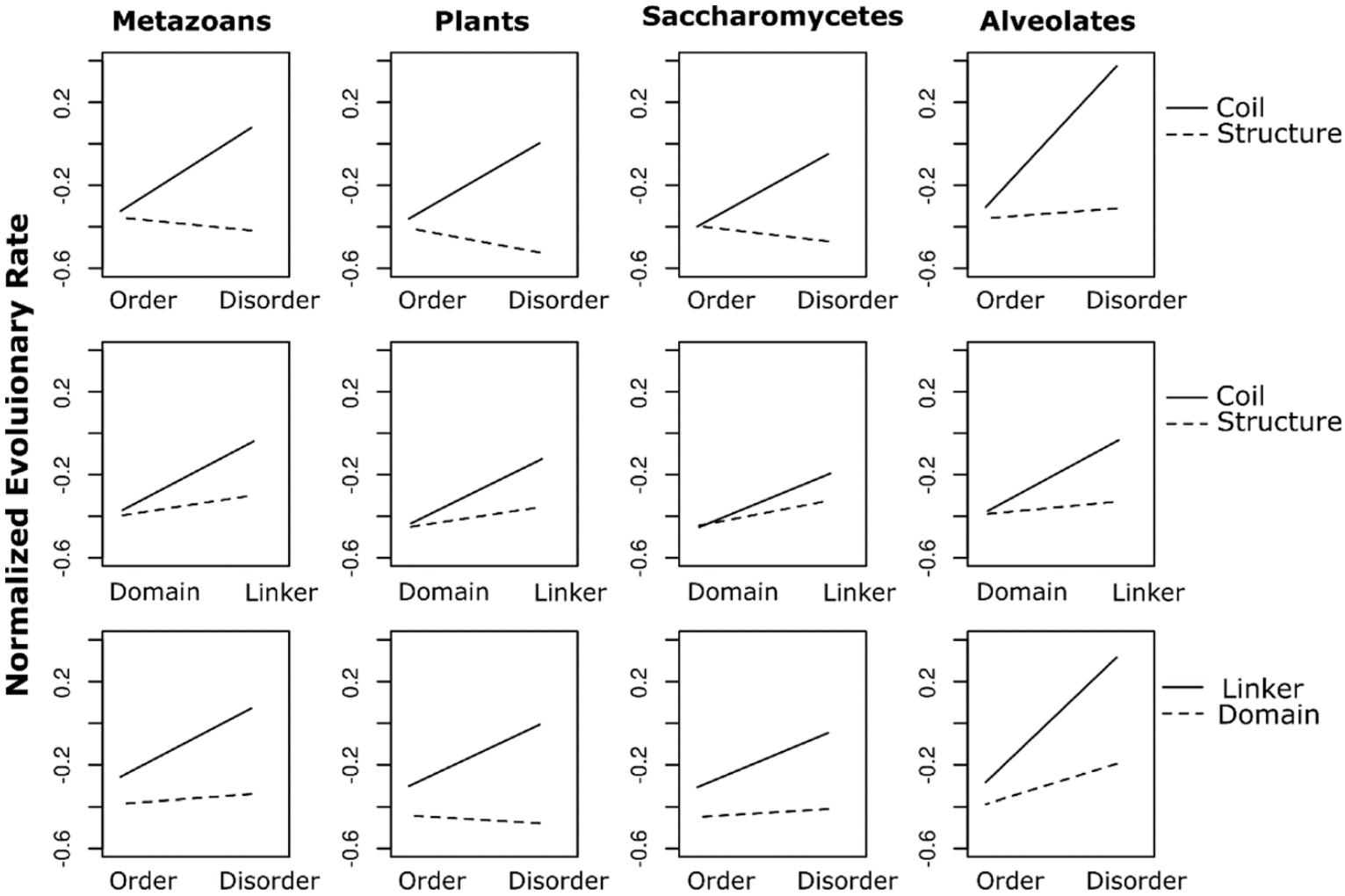
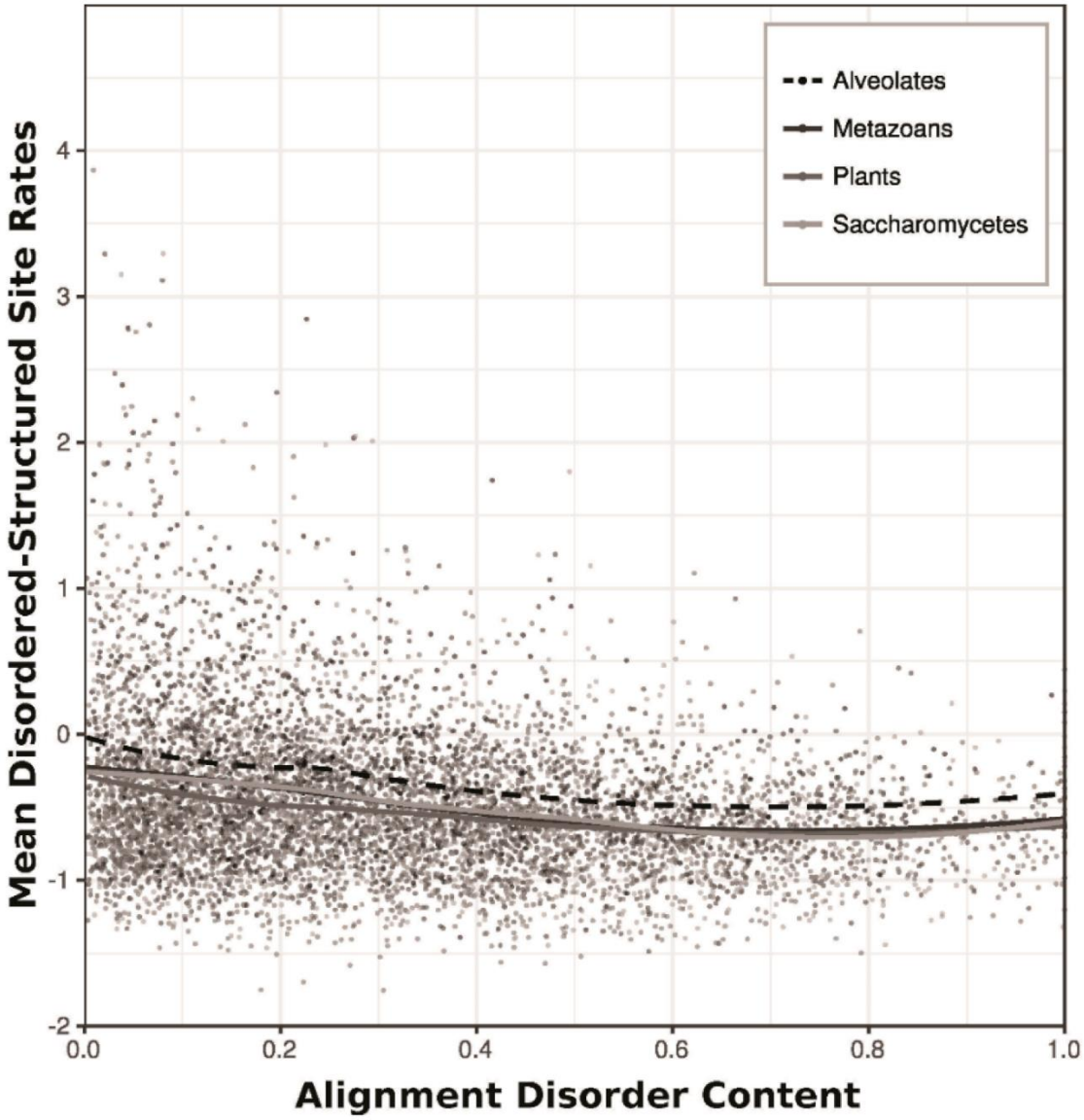

Metazoans
Plants
Saccharomycetes
Alveolates
Normalized Evoluionary Rate
0.2
-0.2
-0.6
Order
Disorder
Domain
Linker
Coil
Structure
Linker
Domain

Figure 4


Alveolates
Metazoans
Plants
Saccharomycetes
Mean Disordered-Structured Site Rates
4
3
2
1
0
-1
-2
0.0
0.2
0.4
0.6
0.8
1.0
Alignment Disorder Content

Appendix Captions

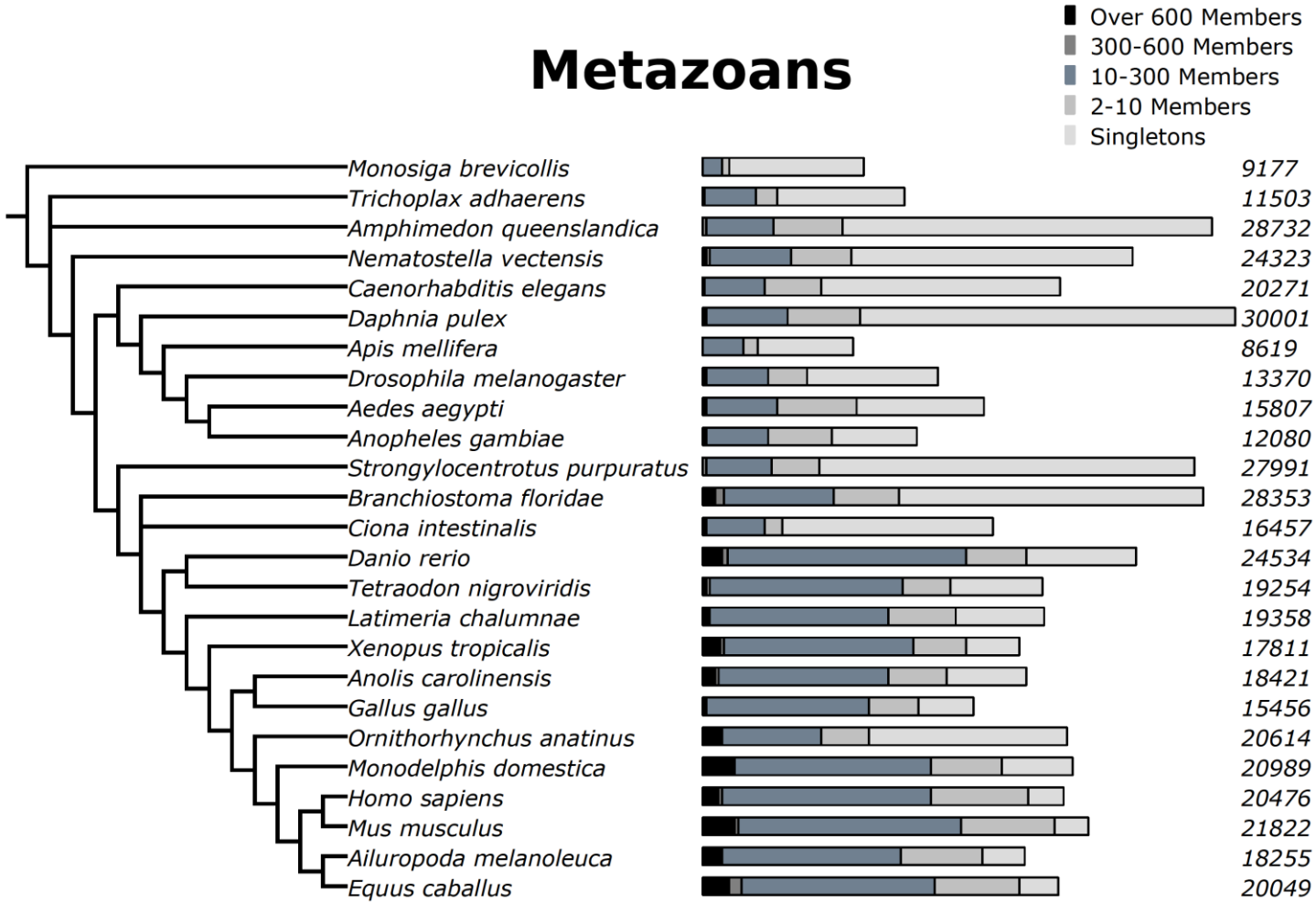
Appendix 1. Dataset information: Cladograms representing evolutionary relationships among the Metazoans, Plants, Saccharomycetes and Alveolates used in this study. Stacked bars (drawn to scale) indicate the fractions of protein sequences from each reference proteome that fall within sequence clusters of various sizes (see individual legends) after clustering analysis. Total proteome sizes (number of sequences in each proteome) are indicated to the right of each stacked bar. Cladograms were drawn according to the NCBI Common Taxonomy Tree [65,66].

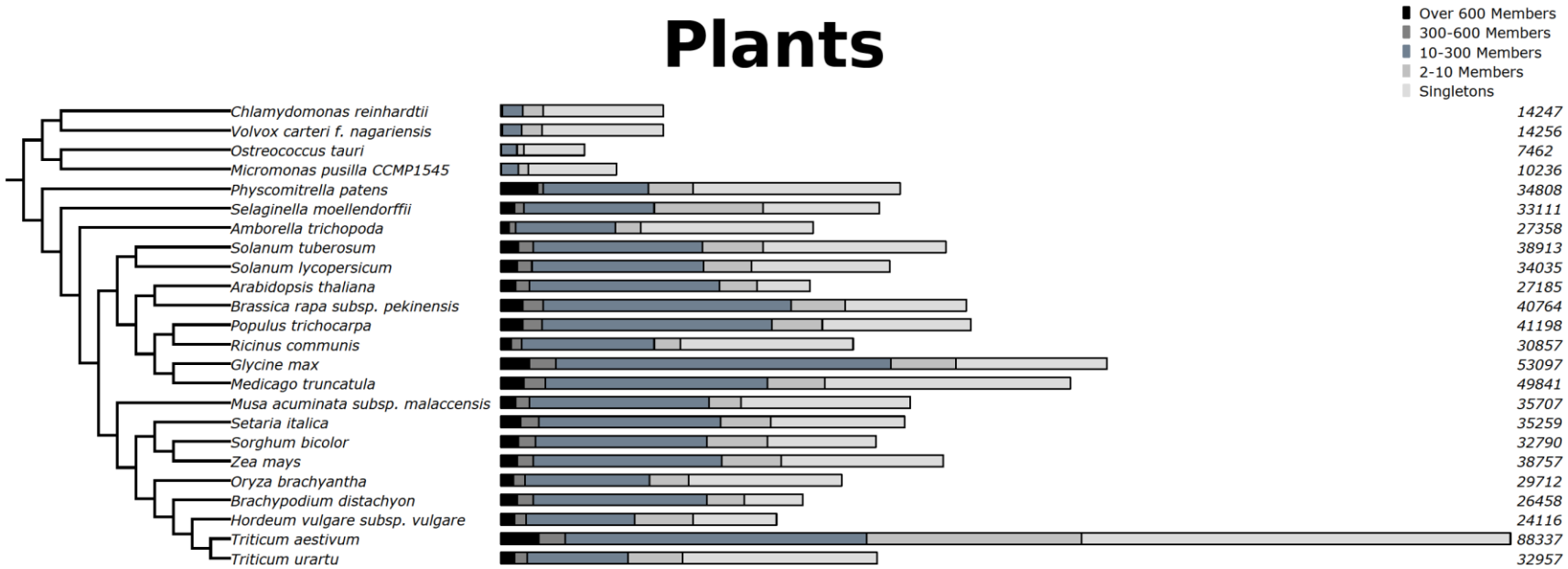
Appendix 2. GO Term results: Gene ontology results obtained from PantherDB [32,33] for protein sequences containing conserved disordered-structured sites.

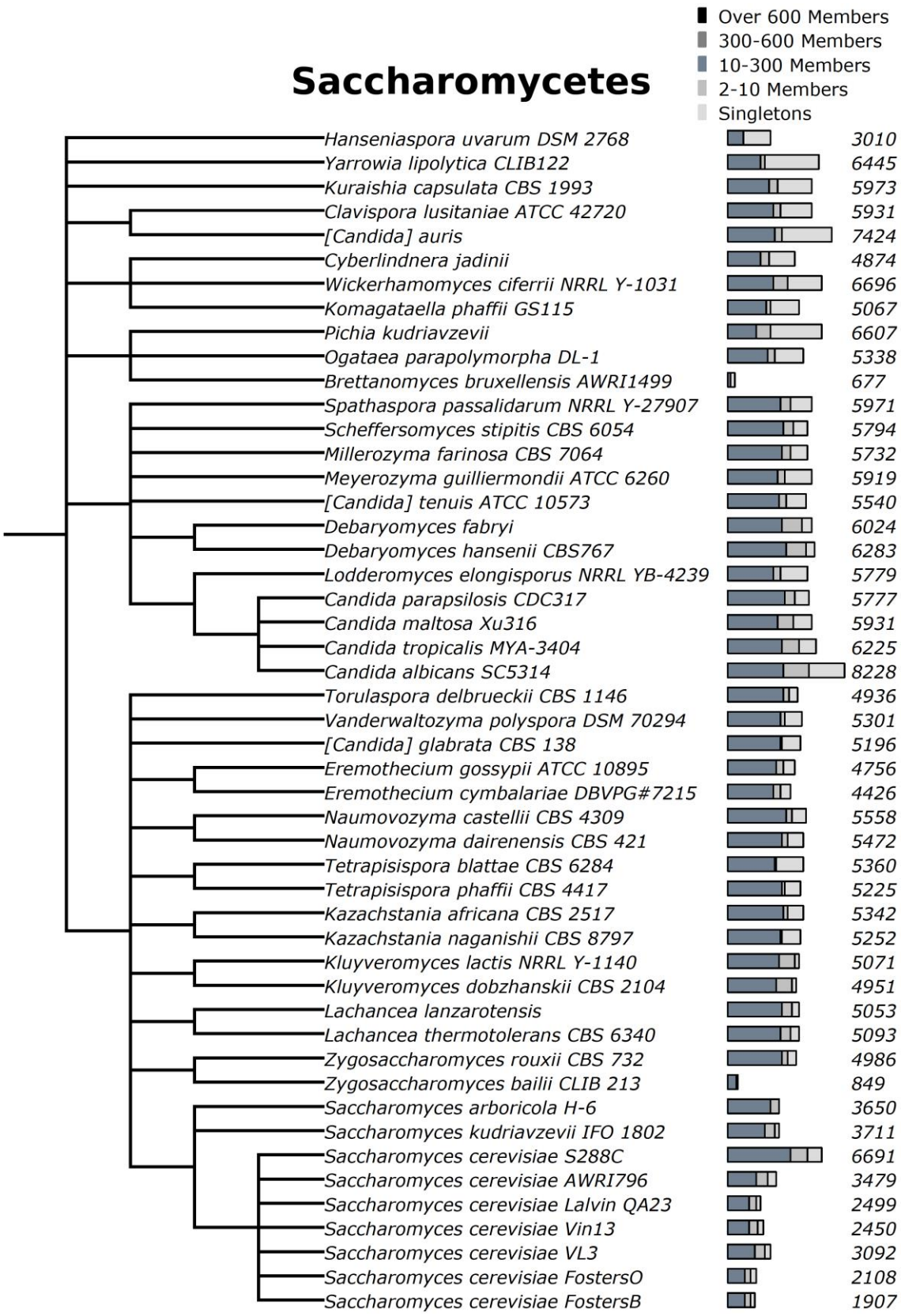
Appendix 3. Nonparametric *post hoc* multiple comparison results.

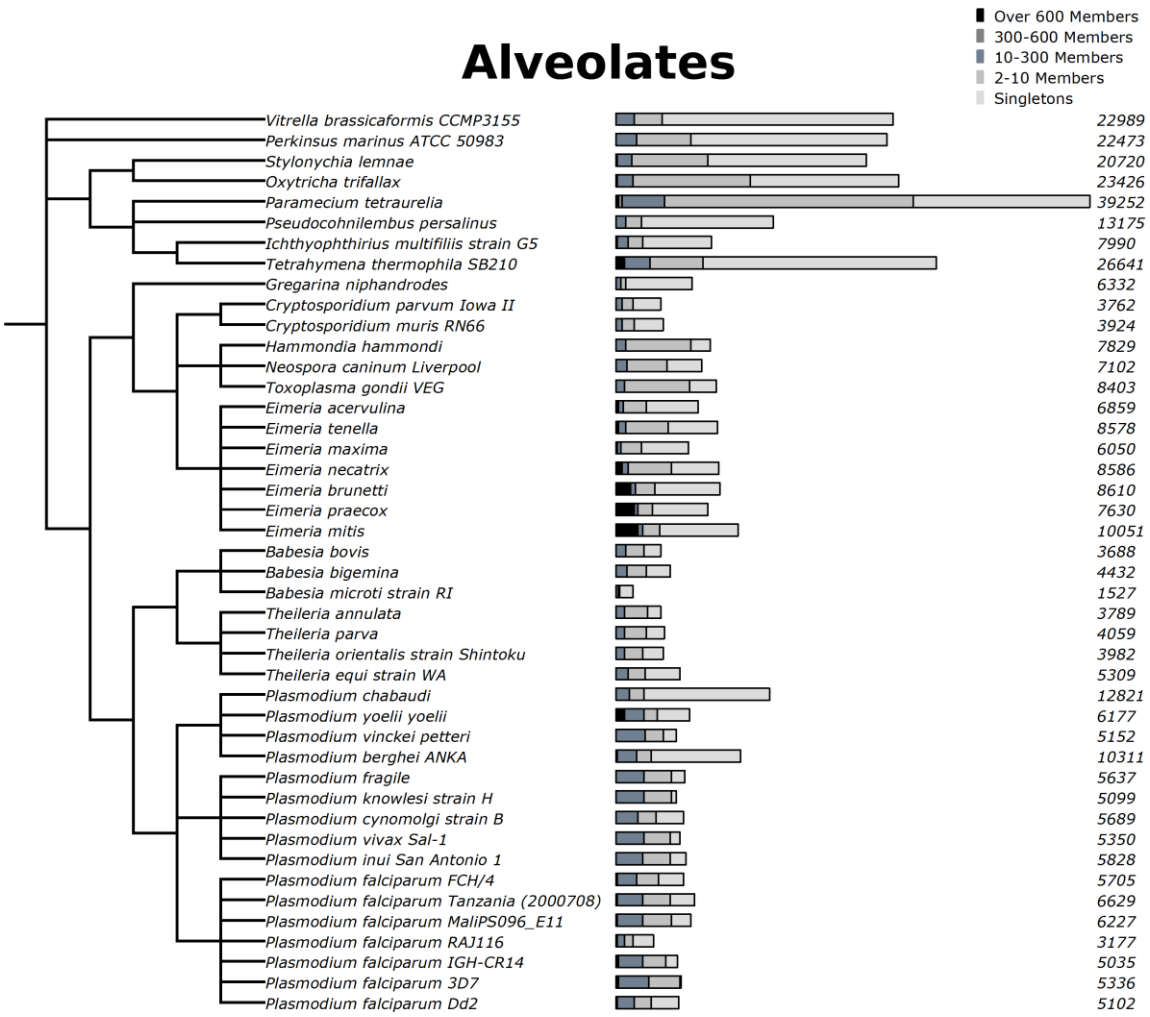
Appendices

Appendix 1




Plants
Over 600 Members
300-600 Members
10-300 Members
2-10 Members
Singletons
Chlamydomonas reinhardtii 14247
Volvox carteri f. nagariensis 14256
Ostreococcus tauri 7462
Micromonas pusilla CCMP1545 10236
Physcomitrella patens 34808
Selaginella moellendorffii 33111
Amborella trichopoda 27358
Solanum tuberosum 38913
Solanum lycopersicum 34035
Arabidopsis thaliana 27185
Brassica rapa subsp. pekinensis 40764
Populus trichocarpa 41198
Ricinus communis 30857
Glycine max 53097
Medicago truncatula 49841
Musa acuminata subsp. malaccensis 35707
Setaria italica 35259
Sorghum bicolor 32790
Zea mays 38757
Oryza brachyantha 29712
Brachypodium distachyon 26458
Hordeum vulgare subsp. vulgare 24116
Triticum aestivum 88337
Triticum urartu 32957


Saccharomycetes
Over 600 Members
300-600 Members
10-300 Members
2-10 Members
Singletons
Hanseniaspora uvarum DSM 2768 3010
Yarrowia lipolytica CLIB122 6445
Kuraishia capsulata CBS 1993 5973
Clavispora lusitaniae ATCC 42720 5931
[Candida] auris 7424
Cyberlindnera jadinii 4874
Wickerhamomyces ciferrii NRRL Y-1031 6696
Komagataella phaffii GS115 5067
Pichia kudriavzevii 6607
Ogataea parapolymorpha DL-1 5338
Brettanomyces bruxellensis AWRI1499 677
Spathaspora passalidarum NRRL Y-27907 5971
Scheffersomyces stipitis CBS 6054 5794
Millerozyma farinosa CBS 7064 5732
Meyerozyma guilliermondii ATCC 6260 5919
[Candida] tenuis ATCC 10573 5540
Debaryomyces fabryi 6024
Debaryomyces hansenii CBS767 6283
Lodderomyces elongisporus NRRL YB-4239 5779
Candida parapsilosis CDC317 5777
Candida maltosa Xu316 5931
Candida tropicalis MYA-3404 6225
Candida albicans SC5314 8228
Torulaspora delbrueckii CBS 1146 4936
Vanderwaltozyma polyspora DSM 70294 5301
[Candida] glabrata CBS 138 5196
Eremothecium gossypii ATCC 10895 4756
Eremothecium cymbalariae DBVPG#7215 4426
Naumovozyma castellii CBS 4309 5558
Naumovozyma dairenensis CBS 421 5472
Tetrapisispora blattae CBS 6284 5360
Tetrapisispora phaffii CBS 4417 5225
Kazachstania africana CBS 2517 5342
Kazachstania naganishii CBS 8797 5252
Kluyveromyces lactis NRRL Y-1140 5071
Kluyveromyces dobzhanskii CBS 2104 4951
Lachancea lanzarotensis 5053
Lachancea thermotolerans CBS 6340 5093
Zygosaccharomyces rouxii CBS 732 4986
Zygosaccharomyces bailii CLIB 213 849
Saccharomyces arboricola H-6 3650
Saccharomyces kudriavzevii IFO 1802 3711
Saccharomyces cerevisiae S288C 6691
Saccharomyces cerevisiae AWRI796 3479
Saccharomyces cerevisiae Lalvin QA23 2499
Saccharomyces cerevisiae Vin13 2450
Saccharomyces cerevisiae VL3 3092
Saccharomyces cerevisiae FostersO 2108
Saccharomyces cerevisiae FostersB 1907


Alveolates
Over 600 Members
300-600 Members
10-300 Members
2-10 Members
Singletons
Vitrella brassicaformis CCMP3155 22989
Perkinsus marinus ATCC 50983 22473
Stylonychia lemnae 20720
Oxytricha trifallax 23426
Paramecium tetraurelia 39252
Pseudocohnilembus persalinus 13175
Ichthyophthirius multifiliis strain G5 7990
Tetrahymena thermophila SB210 26641
Gregarina niphandrodes 6332
Cryptosporidium parvum Iowa II 3762
Cryptosporidium muris RN66 3924
Hammondia hammondi 7829
Neospora caninum Liverpool 7102
Toxoplasma gondii VEG 8403
Eimeria acervulina 6859
Eimeria tenella 8578
Eimeria maxima 6050
Eimeria necatrix 8586
Eimeria brunetti 8610
Eimeria praecox 7630
Eimeria mitis 10051
Babesia bovis 3688
Babesia bigemina 4432
Babesia microti strain RI 1527
Theileria annulata 3789
Theileria parva 4059
Theileria orientalis strain Shintoku 3982
Theileria equi strain WA 5309
Plasmodium chabaudi 12821
Plasmodium yoelii yoelii 6177
Plasmodium vinckei petteri 5152
Plasmodium berghei ANKA 10311
Plasmodium fragile 5637
Plasmodium knowlesi strain H 5099
Plasmodium cynomolgi strain B 5689
Plasmodium vivax Sal-1 5350
Plasmodium inui San Antonio 1 5828
Plasmodium falciparum FCH/4 5705
Plasmodium falciparum Tanzania (2000708) 6629
Plasmodium falciparum MaliPS096_E11 6227
Plasmodium falciparum RAJ116 3177
Plasmodium falciparum IGH-CR14 5035
Plasmodium falciparum 3D7 5336
Plasmodium falciparum Dd2 5102

Appendix 2

Molecular Function				
#	GO Term	Count	% Proteins	% Terms
Homo sapiens				
1	binding (GO:0005488)	1127	34.40%	53.30%
2	catalytic activity (GO:0003824)	652	19.90%	30.80%
3	structural molecule activity (GO:0005198)	101	3.10%	4.80%
4	transporter activity (GO:0005215)	97	3.00%	4.60%
5	signal transducer activity (GO:0004871)	56	1.70%	2.60%
6	receptor activity (GO:0004872)	55	1.70%	2.60%
7	translation regulator activity (GO:0045182)	25	0.80%	1.20%
8	channel regulator activity (GO:0016247)	2	0.10%	0.10%
Arabidopsis thaliana				
1	binding (GO:0005488)	574	22.20%	43.00%
2	catalytic activity (GO:0003824)	527	20.40%	39.50%
3	structural molecule activity (GO:0005198)	114	4.40%	8.50%
4	transporter activity (GO:0005215)	44	1.70%	3.30%
5	translation regulator activity (GO:0045182)	38	1.50%	2.80%
6	receptor activity (GO:0004872)	15	0.60%	1.10%
7	signal transducer activity (GO:0004871)	12	0.50%	0.90%
8	antioxidant activity (GO:0016209)	11	0.40%	0.80%
Saccharomyces cerevisiae				
1	binding (GO:0005488)	274	31.50%	40.50%
2	catalytic activity (GO:0003824)	268	30.80%	39.60%
3	structural molecule activity (GO:0005198)	78	9.00%	11.50%
4	transporter activity (GO:0005215)	26	3.00%	3.80%
5	translation regulator activity (GO:0045182)	23	2.60%	3.40%
6	signal transducer activity (GO:0004871)	4	0.50%	0.60%

7	receptor activity (GO:0004872)	3	0.30%	0.40%
8	antioxidant activity (GO:0016209)	1	0.10%	0.10%
Plasmodium falciparum				
1	binding (GO:0005488)	157	29.50%	41.80%
2	catalytic activity (GO:0003824)	146	27.40%	38.80%
3	structural molecule activity (GO:0005198)	36	6.80%	9.60%
4	transporter activity (GO:0005215)	18	3.40%	4.80%
5	translation regulator activity (GO:0045182)	17	3.20%	4.50%
6	receptor activity (GO:0004872)	1	0.20%	0.30%
7	antioxidant activity (GO:0016209)	1	0.20%	0.30%

Biological Process				
#	GO Term	Count	% Proteins	% Terms
Homo sapiens				
1	cellular process (GO:0009987)	1533	46.80%	30.10%
2	metabolic process (GO:0008152)	1218	37.20%	23.90%
3	cellular component organization or biogenesis (GO:0071840)	529	16.20%	10.40%
4	biological regulation (GO:0065007)	369	11.30%	7.20%
5	developmental process (GO:0032502)	362	11.10%	7.10%
6	localization (GO:0051179)	329	10.00%	6.50%
7	response to stimulus (GO:0050896)	314	9.60%	6.20%
8	multicellular organismal process (GO:0032501)	244	7.50%	4.80%
9	immune system process (GO:0002376)	64	2.00%	1.30%
10	biological adhesion (GO:0022610)	54	1.60%	1.10%
11	locomotion (GO:0040011)	42	1.30%	0.80%
12	reproduction (GO:0000003)	31	0.90%	0.60%
13	rhythmic process (GO:0048511)	5	0.20%	0.10%
14	growth (GO:0040007)	4	0.10%	0.10%
Arabidopsis thaliana				
1	metabolic process (GO:0008152)	903	34.90%	35.30%

2	cellular process (GO:0009987)	894	34.50%	34.90%
3	cellular component organization or biogenesis (GO:0071840)	294	11.40%	11.50%
4	localization (GO:0051179)	164	6.30%	6.40%
5	response to stimulus (GO:0050896)	134	5.20%	5.20%
6	biological regulation (GO:0065007)	104	4.00%	4.10%
7	developmental process (GO:0032502)	32	1.20%	1.30%
8	reproduction (GO:0000003)	20	0.80%	0.80%
9	multicellular organismal process (GO:0032501)	11	0.40%	0.40%
10	rhythmic process (GO:0048511)	2	0.10%	0.10%
Saccharomyces cerevisiae				
1	cellular process (GO:0009987)	478	54.90%	35.40%
2	metabolic process (GO:0008152)	434	49.80%	32.10%
3	cellular component organization or biogenesis (GO:0071840)	226	25.90%	16.70%
4	localization (GO:0051179)	95	10.90%	7.00%
5	response to stimulus (GO:0050896)	59	6.80%	4.40%
6	biological regulation (GO:0065007)	55	6.30%	4.10%
7	reproduction (GO:0000003)	4	0.50%	0.30%
8	developmental process (GO:0032502)	1	0.10%	0.10%
Plasmodium falciparum				
1	cellular process (GO:0009987)	275	51.60%	37.60%
2	metabolic process (GO:0008152)	251	47.10%	34.30%
3	cellular component organization or biogenesis (GO:0071840)	103	19.30%	14.10%
4	localization (GO:0051179)	54	10.10%	7.40%
5	response to stimulus (GO:0050896)	24	4.50%	3.30%
6	biological regulation (GO:0065007)	17	3.20%	2.30%
7	reproduction (GO:0000003)	4	0.80%	0.50%
8	locomotion (GO:0040011)	2	0.40%	0.30%
9	developmental process (GO:0032502)	1	0.20%	0.10%
Cellular Component				

#	GO Term	Count	% Proteins	% Terms
Homo sapiens				
1	cell part (GO:0044464)	1248	38.10%	38.90%
2	organelle (GO:0043226)	976	29.80%	30.40%
3	macromolecular complex (GO:0032991)	553	16.90%	17.20%
4	membrane (GO:0016020)	295	9.00%	9.20%
5	extracellular region (GO:0005576)	60	1.80%	1.90%
6	cell junction (GO:0030054)	31	0.90%	1.00%
7	synapse (GO:0045202)	29	0.90%	0.90%
8	extracellular matrix (GO:0031012)	16	0.50%	0.50%
Arabidopsis thaliana				
1	cell part (GO:0044464)	936	36.20%	42.00%
2	organelle (GO:0043226)	712	27.50%	32.00%
3	macromolecular complex (GO:0032991)	450	17.40%	20.20%
4	membrane (GO:0016020)	118	4.60%	5.30%
5	extracellular region (GO:0005576)	9	0.30%	0.40%
6	nucleoid (GO:0009295)	1	0.00%	0.00%
7	cell junction (GO:0030054)	1	0.00%	0.00%
Saccharomyces cerevisiae				
1	cell part (GO:0044464)	484	55.60%	40.20%
2	organelle (GO:0043226)	365	41.90%	30.30%
3	macromolecular complex (GO:0032991)	294	33.80%	24.40%
4	membrane (GO:0016020)	61	7.00%	5.10%
5	extracellular region (GO:0005576)	1	0.10%	0.10%
Plasmodium falciparum				
1	cell part (GO:0044464)	243	45.60%	39.60%
2	organelle (GO:0043226)	189	35.50%	30.80%
3	macromolecular complex (GO:0032991)	152	28.50%	24.80%
4	membrane (GO:0016020)	27	5.10%	4.40%

5	extracellular region (GO:0005576)	3	0.60%	0.50%

Protein Class				
#	GO Term	Count	% Proteins	% Terms
Homo sapiens				
1	nucleic acid binding (PC00171)	529	16.20%	24.30%
2	transcription factor (PC00218)	367	11.20%	16.90%
3	hydrolase (PC00121)	209	6.40%	9.60%
4	enzyme modulator (PC00095)	178	5.40%	8.20%
5	cytoskeletal protein (PC00085)	129	3.90%	5.90%
6	transferase (PC00220)	122	3.70%	5.60%
7	membrane traffic protein (PC00150)	87	2.70%	4.00%
8	signaling molecule (PC00207)	79	2.40%	3.60%
9	transporter (PC00227)	55	1.70%	2.50%
10	receptor (PC00197)	53	1.60%	2.40%
11	ligase (PC00142)	47	1.40%	2.20%
12	transfer/carrier protein (PC00219)	45	1.40%	2.10%
13	calcium-binding protein (PC00060)	40	1.20%	1.80%
14	chaperone (PC00072)	31	0.90%	1.40%
15	oxidoreductase (PC00176)	29	0.90%	1.30%
16	cell adhesion molecule (PC00069)	29	0.90%	1.30%
17	transmembrane receptor regulatory/adaptor protein (PC00226)	27	0.80%	1.20%
18	defense/immunity protein (PC00090)	26	0.80%	1.20%
19	cell junction protein (PC00070)	26	0.80%	1.20%
20	extracellular matrix protein (PC00102)	22	0.70%	1.00%
21	structural protein (PC00211)	16	0.50%	0.70%
22	isomerase (PC00135)	15	0.50%	0.70%
23	lyase (PC00144)	11	0.30%	0.50%
24	viral protein (PC00237)	1	0.00%	0.00%
25	surfactant (PC00212)	1	0.00%	0.00%

Arabidopsis thaliana				
1	nucleic acid binding (PC00171)	403	15.60%	32.10%
2	hydrolase (PC00121)	175	6.80%	13.90%
3	transferase (PC00220)	112	4.30%	8.90%
4	transcription factor (PC00218)	97	3.70%	7.70%
5	oxidoreductase (PC00176)	64	2.50%	5.10%
6	enzyme modulator (PC00095)	63	2.40%	5.00%
7	cytoskeletal protein (PC00085)	56	2.20%	4.50%
8	transporter (PC00227)	45	1.70%	3.60%
9	membrane traffic protein (PC00150)	41	1.60%	3.30%
10	lyase (PC00144)	39	1.50%	3.10%
11	ligase (PC00142)	36	1.40%	2.90%
12	isomerase (PC00135)	30	1.20%	2.40%
13	chaperone (PC00072)	22	0.90%	1.80%
14	transfer/carrier protein (PC00219)	19	0.70%	1.50%
15	calcium-binding protein (PC00060)	17	0.70%	1.40%
16	defense/immunity protein (PC00090)	7	0.30%	0.60%
17	signaling molecule (PC00207)	6	0.20%	0.50%
18	structural protein (PC00211)	6	0.20%	0.50%
19	receptor (PC00197)	6	0.20%	0.50%
20	extracellular matrix protein (PC00102)	3	0.10%	0.20%
21	transmembrane receptor regulatory/adaptor protein (PC00226)	3	0.10%	0.20%
22	cell adhesion molecule (PC00069)	3	0.10%	0.20%
23	storage protein (PC00210)	2	0.10%	0.20%
Saccharomyces cerevisiae				
1	nucleic acid binding (PC00171)	227	26.10%	37.50%
2	hydrolase (PC00121)	68	7.80%	11.20%
3	transferase (PC00220)	50	5.70%	8.30%
4	enzyme modulator (PC00095)	44	5.10%	7.30%
5	transcription factor (PC00218)	39	4.50%	6.40%

6	transporter (PC00227)	25	2.90%	4.10%
7	ligase (PC00142)	23	2.60%	3.80%
8	membrane traffic protein (PC00150)	22	2.50%	3.60%
9	cytoskeletal protein (PC00085)	20	2.30%	3.30%
10	oxidoreductase (PC00176)	20	2.30%	3.30%
11	lyase (PC00144)	14	1.60%	2.30%
12	transfer/carrier protein (PC00219)	13	1.50%	2.10%
13	chaperone (PC00072)	10	1.10%	1.70%
14	calcium-binding protein (PC00060)	7	0.80%	1.20%
15	isomerase (PC00135)	7	0.80%	1.20%
16	signaling molecule (PC00207)	5	0.60%	0.80%
17	cell junction protein (PC00070)	4	0.50%	0.70%
18	receptor (PC00197)	4	0.50%	0.70%
19	defense/immunity protein (PC00090)	1	0.10%	0.20%
20	structural protein (PC00211)	1	0.10%	0.20%
21	storage protein (PC00210)	1	0.10%	0.20%
Plasmodium falciparum				
1	nucleic acid binding (PC00171)	143	26.80%	37.40%
2	hydrolase (PC00121)	48	9.00%	12.60%
3	enzyme modulator (PC00095)	34	6.40%	8.90%
4	transferase (PC00220)	28	5.30%	7.30%
5	cytoskeletal protein (PC00085)	18	3.40%	4.70%
6	ligase (PC00142)	17	3.20%	4.50%
7	membrane traffic protein (PC00150)	15	2.80%	3.90%
8	transporter (PC00227)	14	2.60%	3.70%
9	transcription factor (PC00218)	13	2.40%	3.40%
10	calcium-binding protein (PC00060)	8	1.50%	2.10%
11	transfer/carrier protein (PC00219)	8	1.50%	2.10%
12	chaperone (PC00072)	8	1.50%	2.10%
13	isomerase (PC00135)	8	1.50%	2.10%

14	oxidoreductase (PC00176)	5	0.90%	1.30%
15	lyase (PC00144)	3	0.60%	0.80%
16	extracellular matrix protein (PC00102)	2	0.40%	0.50%
17	signaling molecule (PC00207)	2	0.40%	0.50%
18	cell junction protein (PC00070)	2	0.40%	0.50%
19	structural protein (PC00211)	2	0.40%	0.50%
20	receptor (PC00197)	2	0.40%	0.50%
21	transmembrane receptor regulatory/adaptor protein (PC00226)	1	0.20%	0.30%
22	defense/immunity protein (PC00090)	1	0.20%	0.30%

Pathway				
#	GO Term	Count	% Proteins	% Terms
Homo sapiens				
1	Wnt signaling pathway (P00057)	75	2.30%	6.90%
2	Gonadotropin-releasing hormone receptor pathway (P06664)	64	2.00%	5.90%
3	Integrin signalling pathway (P00034)	40	1.20%	3.70%
4	PDGF signaling pathway (P00047)	38	1.20%	3.50%
5	CCKR signaling map (P06959)	37	1.10%	3.40%
6	Inflammation mediated by chemokine and cytokine signaling pathway (P00031)	36	1.10%	3.30%
7	Angiogenesis (P00005)	35	1.10%	3.20%
8	Alzheimer disease-presenilin pathway (P00004)	30	0.90%	2.80%
9	Cadherin signaling pathway (P00012)	30	0.90%	2.80%
10	Huntington disease (P00029)	26	0.80%	2.40%
11	FGF signaling pathway (P00021)	25	0.80%	2.30%
12	EGF receptor signaling pathway (P00018)	25	0.80%	2.30%
13	Apoptosis signaling pathway (P00006)	23	0.70%	2.10%
14	Transcription regulation by bZIP transcription factor (P00055)	20	0.60%	1.90%
15	TGF-beta signaling pathway (P00052)	20	0.60%	1.90%
16	Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha mediated pathway (P00027)	18	0.50%	1.70%
17	T cell activation (P00053)	17	0.50%	1.60%

18	Cytoskeletal regulation by Rho GTPase (P00016)	17	0.50%	1.60%
19	p53 pathway (P00059)	16	0.50%	1.50%
20	Parkinson disease (P00049)	16	0.50%	1.50%
21	General transcription regulation (P00023)	15	0.50%	1.40%
22	Endothelin signaling pathway (P00019)	15	0.50%	1.40%
23	B cell activation (P00010)	15	0.50%	1.40%
24	VEGF signaling pathway (P00056)	14	0.40%	1.30%
25	Interleukin signaling pathway (P00036)	14	0.40%	1.30%
26	Alzheimer disease-amyloid secretase pathway (P00003)	13	0.40%	1.20%
27	p53 pathway feedback loops 2 (P04398)	13	0.40%	1.20%
28	Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha mediated pathway (P00026)	13	0.40%	1.20%
29	Toll receptor signaling pathway (P00054)	12	0.40%	1.10%
30	Nicotinic acetylcholine receptor signaling pathway (P00044)	12	0.40%	1.10%
31	Ras Pathway (P04393)	11	0.30%	1.00%
32	Muscarinic acetylcholine receptor 1 and 3 signaling pathway (P00042)	10	0.30%	0.90%
33	Insulin/IGF pathway-protein kinase B signaling cascade (P00033)	10	0.30%	0.90%
34	Axon guidance mediated by netrin (P00009)	9	0.30%	0.80%
35	Ubiquitin proteasome pathway (P00060)	9	0.30%	0.80%
36	Hypoxia response via HIF activation (P00030)	9	0.30%	0.80%
37	5HT2 type receptor mediated signaling pathway (P04374)	9	0.30%	0.80%
38	Oxidative stress response (P00046)	8	0.20%	0.70%
39	Notch signaling pathway (P00045)	8	0.20%	0.70%
40	Thyrotropin-releasing hormone receptor signaling pathway (P04394)	8	0.20%	0.70%
41	p38 MAPK pathway (P05918)	8	0.20%	0.70%
42	DNA replication (P00017)	8	0.20%	0.70%
43	Synaptic vesicle trafficking (P05734)	7	0.20%	0.60%
44	Ionotropic glutamate receptor pathway (P00037)	7	0.20%	0.60%
45	Interferon-gamma signaling pathway (P00035)	7	0.20%	0.60%
46	Oxytocin receptor mediated signaling pathway (P04391)	7	0.20%	0.60%
47	Histamine H1 receptor mediated signaling pathway (P04385)	7	0.20%	0.60%

48	PI3 kinase pathway (P00048)	6	0.20%	0.60%
49	Hedgehog signaling pathway (P00025)	6	0.20%	0.60%
50	Angiotensin II-stimulated signaling through G proteins and beta-arrestin (P05911)	6	0.20%	0.60%
51	Circadian clock system (P00015)	6	0.20%	0.60%
52	Cell cycle (P00013)	6	0.20%	0.60%
53	Axon guidance mediated by Slit/Robo (P00008)	5	0.20%	0.50%
54	Alpha adrenergic receptor signaling pathway (P00002)	5	0.20%	0.50%
55	mRNA splicing (P00058)	5	0.20%	0.50%
56	GABA-B receptor II signaling (P05731)	5	0.20%	0.50%
57	Insulin/IGF pathway-mitogen activated protein kinase kinase/MAP kinase cascade (P00032)	5	0.20%	0.50%
58	Nicotine pharmacodynamics pathway (P06587)	5	0.20%	0.50%
59	Blood coagulation (P00011)	5	0.20%	0.50%
60	Beta2 adrenergic receptor signaling pathway (P04378)	5	0.20%	0.50%
61	Beta1 adrenergic receptor signaling pathway (P04377)	5	0.20%	0.50%
62	TCA cycle (P00051)	4	0.10%	0.40%
63	De novo pyrimidine ribonucleotides biosythesis (P02740)	4	0.10%	0.40%
64	Metabotropic glutamate receptor group II pathway (P00040)	4	0.10%	0.40%
65	Metabotropic glutamate receptor group III pathway (P00039)	4	0.10%	0.40%
66	Vasopressin synthesis (P04395)	4	0.10%	0.40%
67	P53 pathway feedback loops 1 (P04392)	4	0.10%	0.40%
68	Opioid proopiomelanocortin pathway (P05917)	4	0.10%	0.40%
69	Opioid proenkephalin pathway (P05915)	4	0.10%	0.40%
70	Dopamine receptor mediated signaling pathway (P05912)	4	0.10%	0.40%
71	Adrenaline and noradrenaline biosynthesis (P00001)	3	0.10%	0.30%
72	Heme biosynthesis (P02746)	3	0.10%	0.30%
73	Plasminogen activating cascade (P00050)	3	0.10%	0.30%
74	Muscarinic acetylcholine receptor 2 and 4 signaling pathway (P00043)	3	0.10%	0.30%
75	Metabotropic glutamate receptor group I pathway (P00041)	3	0.10%	0.30%
76	Endogenous cannabinoid signaling (P05730)	3	0.10%	0.30%
77	JAK/STAT signaling pathway (P00038)	3	0.10%	0.30%

78	p53 pathway by glucose deprivation (P04397)	3	0.10%	0.30%
79	General transcription by RNA polymerase I (P00022)	3	0.10%	0.30%
80	Pyruvate metabolism (P02772)	3	0.10%	0.30%
81	Cortocotropin releasing factor receptor signaling pathway (P04380)	3	0.10%	0.30%
82	5HT4 type receptor mediated signaling pathway (P04376)	3	0.10%	0.30%
83	5HT3 type receptor mediated signaling pathway (P04375)	3	0.10%	0.30%
84	5HT1 type receptor mediated signaling pathway (P04373)	3	0.10%	0.30%
85	Axon guidance mediated by semaphorins (P00007)	2	0.10%	0.20%
86	Methylmalonyl pathway (P02755)	2	0.10%	0.20%
87	Isoleucine biosynthesis (P02748)	2	0.10%	0.20%
88	De novo purine biosynthesis (P02738)	2	0.10%	0.20%
89	ATP synthesis (P02721)	2	0.10%	0.20%
90	Vitamin D metabolism and pathway (P04396)	2	0.10%	0.20%
91	Glycolysis (P00024)	2	0.10%	0.20%
92	Succinate to proprionate conversion (P02777)	2	0.10%	0.20%
93	FAS signaling pathway (P00020)	2	0.10%	0.20%
94	5-Hydroxytryptamine degredation (P04372)	2	0.10%	0.20%
95	SCW signaling pathway (P06216)	1	0.00%	0.10%
96	GBB signaling pathway (P06214)	1	0.00%	0.10%
97	DPP signaling pathway (P06213)	1	0.00%	0.10%
98	DPP-SCW signaling pathway (P06212)	1	0.00%	0.10%
99	BMP/activin signaling pathway-drosophila (P06211)	1	0.00%	0.10%
100	Pyridoxal-5-phosphate biosynthesis (P02759)	1	0.00%	0.10%
101	Mannose metabolism (P02752)	1	0.00%	0.10%
102	Lipoate_biosynthesis (P02750)	1	0.00%	0.10%
103	Fructose galactose metabolism (P02744)	1	0.00%	0.10%
104	De novo pyrimidine deoxyribonucleotide biosynthesis (P02739)	1	0.00%	0.10%
105	Cysteine biosynthesis (P02737)	1	0.00%	0.10%
106	Arginine biosynthesis (P02728)	1	0.00%	0.10%
107	Vitamin B6 metabolism (P02787)	1	0.00%	0.10%

108	Valine biosynthesis (P02785)	1	0.00%	0.10%
109	Anandamide degradation (P05728)	1	0.00%	0.10%
110	Opioid prodynorphin pathway (P05916)	1	0.00%	0.10%
111	Salvage pyrimidine ribonucleotides (P02775)	1	0.00%	0.10%
112	Enkephalin release (P05913)	1	0.00%	0.10%
113	Pyridoxal phosphate salvage pathway (P02770)	1	0.00%	0.10%
114	Proline biosynthesis (P02768)	1	0.00%	0.10%
115	Beta3 adrenergic receptor signaling pathway (P04379)	1	0.00%	0.10%
116	Pentose phosphate pathway (P02762)	1	0.00%	0.10%
Arabidopsis thaliana				
1	General transcription regulation (P00023)	17	0.70%	5.50%
2	Transcription regulation by bZIP transcription factor (P00055)	16	0.60%	5.20%
3	Ubiquitin proteasome pathway (P00060)	10	0.40%	3.30%
4	Huntington disease (P00029)	9	0.30%	2.90%
5	PDGF signaling pathway (P00047)	8	0.30%	2.60%
6	DNA replication (P00017)	8	0.30%	2.60%
7	Wnt signaling pathway (P00057)	7	0.30%	2.30%
8	Parkinson disease (P00049)	7	0.30%	2.30%
9	De novo purine biosynthesis (P02738)	7	0.30%	2.30%
10	Nicotinic acetylcholine receptor signaling pathway (P00044)	7	0.30%	2.30%
11	Ras Pathway (P04393)	7	0.30%	2.30%
12	EGF receptor signaling pathway (P00018)	7	0.30%	2.30%
13	Adrenaline and noradrenaline biosynthesis (P00001)	6	0.20%	2.00%
14	Histidine biosynthesis (P02747)	6	0.20%	2.00%
15	Heme biosynthesis (P02746)	6	0.20%	2.00%
16	Muscarinic acetylcholine receptor 2 and 4 signaling pathway (P00043)	6	0.20%	2.00%
17	Insulin/IGF pathway-mitogen activated protein kinase kinase/MAP kinase cascade (P00032)	6	0.20%	2.00%
18	Gonadotropin-releasing hormone receptor pathway (P06664)	5	0.20%	1.60%
19	Interleukin signaling pathway (P00036)	5	0.20%	1.60%
20	Inflammation mediated by chemokine and cytokine signaling pathway (P00031)	5	0.20%	1.60%

21	TCA cycle (P00051)	4	0.20%	1.30%
22	Oxidative stress response (P00046)	4	0.20%	1.30%
23	Ionotropic glutamate receptor pathway (P00037)	4	0.20%	1.30%
24	Tryptophan biosynthesis (P02783)	4	0.20%	1.30%
25	FGF signaling pathway (P00021)	4	0.20%	1.30%
26	FAS signaling pathway (P00020)	4	0.20%	1.30%
27	S-adenosylmethionine biosynthesis (P02773)	4	0.20%	1.30%
28	p53 pathway (P00059)	3	0.10%	1.00%
29	De novo pyrimidine ribonucleotides biosythesis (P02740)	3	0.10%	1.00%
30	Muscarinic acetylcholine receptor 1 and 3 signaling pathway (P00042)	3	0.10%	1.00%
31	Metabotropic glutamate receptor group II pathway (P00040)	3	0.10%	1.00%
32	Arginine biosynthesis (P02728)	3	0.10%	1.00%
33	Metabotropic glutamate receptor group III pathway (P00039)	3	0.10%	1.00%
34	Integrin signalling pathway (P00034)	3	0.10%	1.00%
35	Thyrotropin-releasing hormone receptor signaling pathway (P04394)	3	0.10%	1.00%
36	Oxytocin receptor mediated signaling pathway (P04391)	3	0.10%	1.00%
37	p38 MAPK pathway (P05918)	3	0.10%	1.00%
38	Opioid proopiomelanocortin pathway (P05917)	3	0.10%	1.00%
39	Opioid prodynorphin pathway (P05916)	3	0.10%	1.00%
40	Opioid proenkephalin pathway (P05915)	3	0.10%	1.00%
41	Dopamine receptor mediated signaling pathway (P05912)	3	0.10%	1.00%
42	Pyruvate metabolism (P02772)	3	0.10%	1.00%
43	Cortocotropin releasing factor receptor signaling pathway (P04380)	3	0.10%	1.00%
44	Beta3 adrenergic receptor signaling pathway (P04379)	3	0.10%	1.00%
45	Beta2 adrenergic receptor signaling pathway (P04378)	3	0.10%	1.00%
46	Beta1 adrenergic receptor signaling pathway (P04377)	3	0.10%	1.00%
47	5HT4 type receptor mediated signaling pathway (P04376)	3	0.10%	1.00%
48	5HT3 type receptor mediated signaling pathway (P04375)	3	0.10%	1.00%
49	5HT2 type receptor mediated signaling pathway (P04374)	3	0.10%	1.00%
50	5HT1 type receptor mediated signaling pathway (P04373)	3	0.10%	1.00%

51	Apoptosis signaling pathway (P00006)	2	0.10%	0.70%
52	O-antigen biosynthesis (P02757)	2	0.10%	0.70%
53	Methionine biosynthesis (P02753)	2	0.10%	0.70%
54	CCKR signaling map (P06959)	2	0.10%	0.70%
55	Lysine biosynthesis (P02751)	2	0.10%	0.70%
56	Fructose galactose metabolism (P02744)	2	0.10%	0.70%
57	TGF-beta signaling pathway (P00052)	2	0.10%	0.70%
58	PI3 kinase pathway (P00048)	2	0.10%	0.70%
59	Adenine and hypoxanthine salvage pathway (P02723)	2	0.10%	0.70%
60	Insulin/IGF pathway-protein kinase B signaling cascade (P00033)	2	0.10%	0.70%
61	p53 pathway feedback loops 2 (P04398)	2	0.10%	0.70%
62	Vitamin D metabolism and pathway (P04396)	2	0.10%	0.70%
63	General transcription by RNA polymerase I (P00022)	2	0.10%	0.70%
64	Pentose phosphate pathway (P02762)	2	0.10%	0.70%
65	Pyridoxal-5-phosphate biosynthesis (P02759)	1	0.00%	0.30%
66	N-acetylglucosamine metabolism (P02756)	1	0.00%	0.30%
67	Methylmalonyl pathway (P02755)	1	0.00%	0.30%
68	Isoleucine biosynthesis (P02748)	1	0.00%	0.30%
69	mRNA splicing (P00058)	1	0.00%	0.30%
70	Glutamine glutamate conversion (P02745)	1	0.00%	0.30%
71	Notch signaling pathway (P00045)	1	0.00%	0.30%
72	Chorismate biosynthesis (P02734)	1	0.00%	0.30%
73	Biotin biosynthesis (P02731)	1	0.00%	0.30%
74	Synaptic vesicle trafficking (P05734)	1	0.00%	0.30%
75	Ascorbate degradation (P02729)	1	0.00%	0.30%
76	Allantoin degradation (P02725)	1	0.00%	0.30%
77	Interferon-gamma signaling pathway (P00035)	1	0.00%	0.30%
78	Vitamin B6 metabolism (P02787)	1	0.00%	0.30%
79	Valine biosynthesis (P02785)	1	0.00%	0.30%
80	Hypoxia response via HIF activation (P00030)	1	0.00%	0.30%

81	Hedgehog signaling pathway (P00025)	1	0.00%	0.30%
82	Succinate to proprionate conversion (P02777)	1	0.00%	0.30%
83	Serine glycine biosynthesis (P02776)	1	0.00%	0.30%
84	Salvage pyrimidine ribonucleotides (P02775)	1	0.00%	0.30%
85	Pyrimidine Metabolism (P02771)	1	0.00%	0.30%
86	Pyridoxal phosphate salvage pathway (P02770)	1	0.00%	0.30%
87	Circadian clock system (P00015)	1	0.00%	0.30%
88	Cell cycle (P00013)	1	0.00%	0.30%
89	Phenylethylamine degradation (P02766)	1	0.00%	0.30%
90	Peptidoglycan biosynthesis (P02763)	1	0.00%	0.30%
91	5-Hydroxytryptamine degredation (P04372)	1	0.00%	0.30%
Saccharomyces cerevisiae				
1	Transcription regulation by bZIP transcription factor (P00055)	16	1.80%	8.00%
2	General transcription regulation (P00023)	15	1.70%	7.50%
3	Parkinson disease (P00049)	14	1.60%	7.00%
4	Ubiquitin proteasome pathway (P00060)	13	1.50%	6.50%
5	Apoptosis signaling pathway (P00006)	10	1.10%	5.00%
6	Wnt signaling pathway (P00057)	10	1.10%	5.00%
7	Nicotinic acetylcholine receptor signaling pathway (P00044)	7	0.80%	3.50%
8	Glycolysis (P00024)	6	0.70%	3.00%
9	De novo purine biosynthesis (P02738)	5	0.60%	2.50%
10	TCA cycle (P00051)	4	0.50%	2.00%
11	EGF receptor signaling pathway (P00018)	4	0.50%	2.00%
12	DNA replication (P00017)	4	0.50%	2.00%
13	Isoleucine biosynthesis (P02748)	3	0.30%	1.50%
14	Heme biosynthesis (P02746)	3	0.30%	1.50%
15	PDGF signaling pathway (P00047)	3	0.30%	1.50%
16	Muscarinic acetylcholine receptor 2 and 4 signaling pathway (P00043)	3	0.30%	1.50%
17	Muscarinic acetylcholine receptor 1 and 3 signaling pathway (P00042)	3	0.30%	1.50%
18	Metabotropic glutamate receptor group II pathway (P00040)	3	0.30%	1.50%

19	ATP synthesis (P02721)	3	0.30%	1.50%
20	Valine biosynthesis (P02785)	3	0.30%	1.50%
21	Huntington disease (P00029)	3	0.30%	1.50%
22	FGF signaling pathway (P00021)	3	0.30%	1.50%
23	CCKR signaling map (P06959)	2	0.20%	1.00%
24	Leucine biosynthesis (P02749)	2	0.20%	1.00%
25	Toll receptor signaling pathway (P00054)	2	0.20%	1.00%
26	De novo pyrimidine ribonucleotides biosythesis (P02740)	2	0.20%	1.00%
27	Arginine biosynthesis (P02728)	2	0.20%	1.00%
28	Tryptophan biosynthesis (P02783)	2	0.20%	1.00%
29	Endothelin signaling pathway (P00019)	2	0.20%	1.00%
30	Cell cycle (P00013)	2	0.20%	1.00%
31	B cell activation (P00010)	2	0.20%	1.00%
32	Pentose phosphate pathway (P02762)	2	0.20%	1.00%
33	SCW signaling pathway (P06216)	1	0.10%	0.50%
34	GBB signaling pathway (P06214)	1	0.10%	0.50%
35	DPP signaling pathway (P06213)	1	0.10%	0.50%
36	DPP-SCW signaling pathway (P06212)	1	0.10%	0.50%
37	BMP/activin signaling pathway-drosophila (P06211)	1	0.10%	0.50%
38	Gonadotropin-releasing hormone receptor pathway (P06664)	1	0.10%	0.50%
39	Pyridoxal-5-phosphate biosynthesis (P02759)	1	0.10%	0.50%
40	Angiogenesis (P00005)	1	0.10%	0.50%
41	Alzheimer disease-amyloid secretase pathway (P00003)	1	0.10%	0.50%
42	Mannose metabolism (P02752)	1	0.10%	0.50%
43	Lipoate_biosynthesis (P02750)	1	0.10%	0.50%
44	p53 pathway (P00059)	1	0.10%	0.50%
45	mRNA splicing (P00058)	1	0.10%	0.50%
46	Histidine biosynthesis (P02747)	1	0.10%	0.50%
47	VEGF signaling pathway (P00056)	1	0.10%	0.50%
48	Tetrahydrofolate biosynthesis (P02742)	1	0.10%	0.50%

49	TGF-beta signaling pathway (P00052)	1	0.10%	0.50%
50	Cysteine biosynthesis (P02737)	1	0.10%	0.50%
51	Notch signaling pathway (P00045)	1	0.10%	0.50%
52	Chorismate biosynthesis (P02734)	1	0.10%	0.50%
53	Synaptic vesicle trafficking (P05734)	1	0.10%	0.50%
54	Metabotropic glutamate receptor group III pathway (P00039)	1	0.10%	0.50%
55	Ionotropic glutamate receptor pathway (P00037)	1	0.10%	0.50%
56	Vitamin B6 metabolism (P02787)	1	0.10%	0.50%
57	Inflammation mediated by chemokine and cytokine signaling pathway (P00031)	1	0.10%	0.50%
58	Thyrotropin-releasing hormone receptor signaling pathway (P04394)	1	0.10%	0.50%
59	Oxytocin receptor mediated signaling pathway (P04391)	1	0.10%	0.50%
60	Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha mediated pathway (P00026)	1	0.10%	0.50%
61	Hedgehog signaling pathway (P00025)	1	0.10%	0.50%
62	Sulfate assimilation (P02778)	1	0.10%	0.50%
63	General transcription by RNA polymerase I (P00022)	1	0.10%	0.50%
64	Enkephalin release (P05913)	1	0.10%	0.50%
65	Dopamine receptor mediated signaling pathway (P05912)	1	0.10%	0.50%
66	FAS signaling pathway (P00020)	1	0.10%	0.50%
67	Histamine H2 receptor mediated signaling pathway (P04386)	1	0.10%	0.50%
68	Pyruvate metabolism (P02772)	1	0.10%	0.50%
69	Histamine H1 receptor mediated signaling pathway (P04385)	1	0.10%	0.50%
70	Pyridoxal phosphate salvage pathway (P02770)	1	0.10%	0.50%
71	Purine metabolism (P02769)	1	0.10%	0.50%
72	Beta2 adrenergic receptor signaling pathway (P04378)	1	0.10%	0.50%
73	Beta1 adrenergic receptor signaling pathway (P04377)	1	0.10%	0.50%
74	5HT2 type receptor mediated signaling pathway (P04374)	1	0.10%	0.50%
75	5HT1 type receptor mediated signaling pathway (P04373)	1	0.10%	0.50%
Plasmodium falciparum				
1	Ubiquitin proteasome pathway (P00060)	8	1.50%	10.10%
2	Huntington disease (P00029)	5	0.90%	6.30%

3	Transcription regulation by bZIP transcription factor (P00055)	4	0.80%	5.10%
4	General transcription regulation (P00023)	4	0.80%	5.10%
5	DNA replication (P00017)	4	0.80%	5.10%
6	Parkinson disease (P00049)	4	0.80%	5.10%
7	Wnt signaling pathway (P00057)	3	0.60%	3.80%
8	Nicotinic acetylcholine receptor signaling pathway (P00044)	3	0.60%	3.80%
9	p53 pathway (P00059)	2	0.40%	2.50%
10	mRNA splicing (P00058)	2	0.40%	2.50%
11	General transcription by RNA polymerase I (P00022)	2	0.40%	2.50%
12	Methylcitrate cycle (P02754)	2	0.40%	2.50%
13	Tryptophan biosynthesis (P02783)	2	0.40%	2.50%
14	CCKR signaling map (P06959)	2	0.40%	2.50%
15	De novo pyrimidine ribonucleotides biosythesis (P02740)	2	0.40%	2.50%
16	Pyruvate metabolism (P02772)	2	0.40%	2.50%
17	Gonadotropin-releasing hormone receptor pathway (P06664)	2	0.40%	2.50%
18	De novo purine biosynthesis (P02738)	1	0.20%	1.30%
19	Purine metabolism (P02769)	1	0.20%	1.30%
20	Alzheimer disease-presenilin pathway (P00004)	1	0.20%	1.30%
21	Interleukin signaling pathway (P00036)	1	0.20%	1.30%
22	Alzheimer disease-amyloid secretase pathway (P00003)	1	0.20%	1.30%
23	Adrenaline and noradrenaline biosynthesis (P00001)	1	0.20%	1.30%
24	Insulin/IGF pathway-protein kinase B signaling cascade (P00033)	1	0.20%	1.30%
25	Inflammation mediated by chemokine and cytokine signaling pathway (P00031)	1	0.20%	1.30%
26	p53 pathway feedback loops 2 (P04398)	1	0.20%	1.30%
27	Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha mediated pathway (P00027)	1	0.20%	1.30%
28	Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha mediated pathway (P00026)	1	0.20%	1.30%
29	Vitamin D metabolism and pathway (P04396)	1	0.20%	1.30%
30	Ras Pathway (P04393)	1	0.20%	1.30%
31	FGF signaling pathway (P00021)	1	0.20%	1.30%
32	FAS signaling pathway (P00020)	1	0.20%	1.30%

33	ATP synthesis (P02721)	1	0.20%	1.30%
34	TCA cycle (P00051)	1	0.20%	1.30%
35	Endothelin signaling pathway (P00019)	1	0.20%	1.30%
36	EGF receptor signaling pathway (P00018)	1	0.20%	1.30%
37	PI3 kinase pathway (P00048)	1	0.20%	1.30%
38	PDGF signaling pathway (P00047)	1	0.20%	1.30%
39	Heme biosynthesis (P02746)	1	0.20%	1.30%
40	Cell cycle (P00013)	1	0.20%	1.30%
41	Cadherin signaling pathway (P00012)	1	0.20%	1.30%
42	Muscarinic acetylcholine receptor 2 and 4 signaling pathway (P00043)	1	0.20%	1.30%
43	Salvage pyrimidine ribonucleotides (P02775)	1	0.20%	1.30%

Appendix 3

Nonparametric *Post Hoc* Multiple Comparison Results (TRUE = significant; Corrected Alpha = 0.05)				
Comparison	Metazoans	Plants	Saccharomycetes	Alveolates
Disordered-Coil-Domain vs. Disordered-Coil-Linker	TRUE	TRUE	TRUE	TRUE
Disordered-Coil-Domain vs. Disordered-Structure-Domain	FALSE	FALSE	FALSE	TRUE
Disordered-Coil-Domain vs. Disordered-Structure-Linker	TRUE	TRUE	FALSE	TRUE
Disordered-Coil-Domain vs. Ordered-Coil-Domain	TRUE	TRUE	TRUE	FALSE
Disordered-Coil-Domain vs. Ordered-Coil-Linker	TRUE	TRUE	TRUE	FALSE
Disordered-Coil-Domain vs. Ordered-Structure-Domain	TRUE	TRUE	TRUE	TRUE
Disordered-Coil-Domain vs. Ordered-Structure-Linker	TRUE	TRUE	TRUE	TRUE
Disordered-Coil-Linker vs. Disordered-Structure-Domain	TRUE	TRUE	TRUE	TRUE
Disordered-Coil-Linker vs. Disordered-Structure-Linker	TRUE	FALSE	TRUE	TRUE
Disordered-Coil-Linker vs. Ordered-Coil-Domain	TRUE	TRUE	TRUE	TRUE
Disordered-Coil-Linker vs. Ordered-Coil-Linker	TRUE	TRUE	TRUE	TRUE
Disordered-Coil-Linker vs. Ordered-Structure-Domain	TRUE	TRUE	TRUE	TRUE
Disordered-Coil-Linker vs. Ordered-Structure-Linker	TRUE	TRUE	TRUE	TRUE
Disordered-Structure-Domain vs. Disordered-Structure-Linker	TRUE	TRUE	FALSE	TRUE
Disordered-Structure-Domain vs. Ordered-Coil-Domain	FALSE	TRUE	FALSE	TRUE
Disordered-Structure-Domain vs. Ordered-Coil-Linker	TRUE	TRUE	FALSE	TRUE
Disordered-Structure-Domain vs. Ordered-Structure-Domain	TRUE	TRUE	TRUE	FALSE
Disordered-Structure-Domain vs. Ordered-Structure-Linker	TRUE	TRUE	TRUE	TRUE
Disordered-Structure-Linker vs. Ordered-Coil-Domain	TRUE	TRUE	FALSE	TRUE
Disordered-Structure-Linker vs. Ordered-Coil-Linker	FALSE	TRUE	FALSE	FALSE
Disordered-Structure-Linker vs. Ordered-Structure-Domain	FALSE	TRUE	TRUE	TRUE
Disordered-Structure-Linker vs. Ordered-Structure-Linker	TRUE	TRUE	TRUE	FALSE
Ordered-Coil-Domain vs. Ordered-Coil-Linker	TRUE	TRUE	FALSE	TRUE
Ordered-Coil-Domain vs. Ordered-Structure-Domain	TRUE	TRUE	TRUE	TRUE
Ordered-Coil-Domain vs. Ordered-Structure-Linker	TRUE	TRUE	TRUE	TRUE
Ordered-Coil-Linker vs. Ordered-Structure-Domain	TRUE	TRUE	TRUE	TRUE
Ordered-Coil-Linker vs. Ordered-Structure-Linker	TRUE	TRUE	TRUE	TRUE
Ordered-Structure-Domain vs. Ordered-Structure-Linker	TRUE	TRUE	FALSE	TRUE

CHAPTER IV

EVALUATION OF SITE-SPECIFIC RATE HETEROGENEITY REVEALS SIGNIFICANT DIFFERENCES IN SEQUENCE DIVERGENCE PATTERNS BETWEEN ORTHOLOGOUS AND PARALOGOUS PROTEINS IN BOTH ANIMALS AND PLANTS

ABSTRACT

Heterotachy—the change in sequence evolutionary rate over time—is a common feature of protein molecular evolution. Decades of research has shed some light on the conditions under which heterotachy occurs, and there is evidence that evolutionary rate shifts are correlated with changes in protein function. Here, we present a large-scale, computational analysis using thousands of protein sequence alignments from metazoan (animal) and plant proteomes, representing genes related either by orthology (speciation events) or paralogy (gene duplication). We use the results of sequence-based phylogenetic analyses to establish a correlation between sequence alignment divergence (tree length) and the estimated shape parameter (α) of the alignment's inferred rate distribution. We also describe and implement simple, computational simulation methods which largely reproduce the patterns we observed in real protein data. Our simulation results indicate that sequence divergence and the α parameter are positively correlated when sequences evolve with heterotachy, meaning that inferred site rate distributions tend to become more uniform as sequence alignments become more divergent. Tree length and α are also correlated in both orthologous and paralogous genes. However, the rate of α increase is markedly higher in paralogous protein alignments than in orthologous alignments, which is consistent with the widely-held view that paralogous proteins are evolving under relaxed selective pressure promoting functional divergence, and hence experiencing more evolutionary rate fluctuations than orthologous proteins. We discuss these findings in the context of the ortholog conjecture, a long-standing assumption in molecular evolution, which posits that protein sequences related by orthology tend to be more functionally conserved than paralogous proteins.

INTRODUCTION

Homologous pairs of protein-coding genes within multicellular eukaryotes are typically related in one of two broad ways: orthology or paralogy. Orthologous genes are observed in different organisms and are related via speciation events, whereas paralogous genes may be found within a single organism and are related via gene duplications. A long-standing notion in molecular evolution, known as the "ortholog conjecture," posits that orthologous genes tend to have very similar functions, whereas paralogous genes can change function (Koonin 2005). The rationale for the conjecture is that single-copy protein-coding genes cannot easily alter their function without decreasing the overall fitness of an organism. However, following a gene duplication event, one or more gene copies (paralogs) may change function (neofunctionalize) in response to shifting selective pressures associated with concomitant changes in gene dosage (Hughes and Liberles 2008; Ahrens et al. 2017). Retained gene copies often initially undergo subfunctionalization (i.e., complementary partial loss of their ancestral functions), and this may eventually lead to complete functional change at a later time (Rastogi and Liberles 2005; Teufel et al. 2016). Nonetheless, the most common outcome of a small-scale gene duplication (i.e., duplication of only a small segment of the overall genome) is the inactivation (pseudogenization) of any additional gene copies (Lynch and Conery 2000).

Tests of the ortholog conjecture using gene ontology (GO) terms (Ashburner et al. 2000) have produced controversial results (see Nehrt et al. 2011; Chen and Zhang 2012; Rogozin et al. 2014). More careful GO-term-based studies indicated that orthologs are indeed less functionally divergent than paralogs (see Altenhoff et al. 2012) and

alternative lines of experimental evidence in favor of the ortholog conjecture (e.g., gene expression level) have also been presented (Chen and Zhang 2012; Rogozin et al. 2014). Still, in many studies, the explicit phylogenetic context of the data is not considered (see Dunn et al. 2018). Moreover, the expected results under an appropriate null hypothesis (i.e., no association between homology type and propensity for functional divergence) is not always specified (see Studer and Robinson-Rechavi 2009).

Nearly fifty years ago, Fitch (1971) speculated that as a protein sequence changes, the subsets of invariant sites (where mutations cannot occur) and variable sites (where new mutations are accepted) shift as well. The phenomenon is now more generally referred to as heterotachy: a lineage-specific shift in amino acid replacement rates over time (Lopez et al. 2002). Further work has indicated that shifts in site-specific amino acid replacement rates (i.e., heterotachy) are associated with functional changes in proteins (Gu 1999; Gaucher et al. 2002). In particular, Philippe et al. (2003) found that novel mutations in previously invariant sites are strong signifiers of functional change.

Likelihood-based methods of phylogenetic inference have been a component of molecular evolutionary studies throughout much of the 21st century. Following the development of an efficient dynamic programming strategy for likelihood computation (Felsenstein 1973), a multitude of software applications became available for general use among molecular biologists. Some programs employ maximum-likelihood strategies to provide a point estimate of jointly optimized model parameters (tree topology, branch lengths, etc.) whereas others like MrBayes (Ronquist et al. 2012) use Markov chain Monte Carlo algorithms to jointly estimate the posterior distributions of model parameters. A common feature of these methods is the underlying framework in which

likelihood scores are calculated, wherein molecular sequences are assumed to evolve along a branching, continuous-time Markov chain whose state transitions are governed by a fixed, instantaneous rate matrix.

Statistical models of sequence evolution have been proposed (e.g., Fitch and Markowitz 1970; Tuffley and Steel 1998; Galtier 2001) to directly account for heterotachy (Lopez et al. 2002). However, in the interest of computational tractability, heterotachy is typically ignored during phylogenetic inference, and sites are assumed to evolve independently, according to a single rate matrix. To account for fixed differences in the relative speed of site-specific evolution (rate heterogeneity), site rates are often assumed to be drawn from a discrete gamma distribution with a shape parameter $\alpha = \beta$ such that the mean rate in the distribution is 1.0 and the variance of the distribution is $1/\alpha$ (Yang 1996). Notably, though, shifts over time in the overall distribution of site rates—which can be measured via differences in the gamma rate distribution's α parameter among groups of related genes—often indicate changes in gene function (Abhiman et al. 2006).

Here, we present the results of a large-scale study evaluating differences in sequence divergence patterns between alignments of orthologous and paralogous protein sequences found in metazoans (animals) and plants. We use the results of sequence-based phylogenetic analyses to establish a correlation between sequence alignment divergence (total tree length) and the α parameter of the alignment's inferred gamma rate distribution. We also describe simple, computational simulation methods which reproduce the patterns we observed in real protein data. Our goal is to illustrate that common phylogenetic inference methods (which to not directly account for heterotachy)

can still be used to detect varying levels of heterotachy in large-scale datasets. Further, we describe a staightforward statistical test of the ortholog conjecture with i) a clearly defined null hypothesis and ii) a dataset in an explicit phylogenetic context, which can be applied to large molecular sequence datasets even when no GO term annotations exist.

RESULTS

Protein Sequence Data

In the metazoan dataset taken from Ahrens et al. (2016) and Ahrens et al. (2018) (see Figure 1), we identified 5893 sequence alignments containing putative orthologs (i.e., exactly one protein sequence per represented species) and 1028 alignments of Type I paralogs (i.e., several protein sequences from exactly one species). Additionally, 1133 new (Type II) paralog alignments were extracted from protein family alignments (containing a mixture of orthologous and paralogous sequences) in the original metazoan dataset (Figure 1). In the plant dataset (see Figure 1) from Ahrens et al. (2018), we identified 1295 putative ortholog alignments and 823 Type I paralog alignments, and 4623 Type II paralog alignments were extracted from protein family clusters. (See Methods and Figure 2 for details on Type I vs. Type II paralog alignments.)

Phylogenetic Analysis

Of the 1800 simulated multiple sequence alignments that we analyzed in MrBayes (Ronquist et al. 2012), only 7 (0.4%) did not reach an average standard deviation of split frequencies (ASDSF) below 0.005 within 5,000,000 generations. Further, only 3 of these 7 analyses did not reach an ASDSF below 0.01, the convergence diagnostic threshold recommended by the program authors (Ronquist et al. 2011). Only 32 (0.4%) of the 9082

metazoan alignments failed to reach an average standard deviation of split frequencies (ASDSF) below 0.005 within 5,000,000 generations, and only 24 did not reach an ASDSF below 0.01. Similarly, 20 (0.3%) of the 7564 plant alignments failed to reach an average standard deviation of split frequencies (ASDSF) below 0.005 within 5,000,000 generations, and only 8 did not reach an ASDSF below 0.01

Statistical Analysis

Regression analyses, relating normalized estimated tree length (i.e., mean tree length divided by the number of terminal nodes) to the shape parameter α of the inferred 4-category gamma distribution among site rates, show that MrBayes (Ronquist et al. 2012) consistently predicts α values which are close to the true (i.e., predefined) value when 1) alignments are simulated under a fixed gamma rate distribution with no heterotachy and 2) the α value is low (Figure 3A). However, when site rates are allowed to vary along the length of a phylogenetic tree (Figure 3B-C), we observe a positive correlation between the normalized estimated (mean) tree length and the estimated (mean) α parameter of the site rate distribution. Additionally, the degree to which site rates may vary (determined by a scaling constant *C* in each heterotachy model; see methods for details) significantly impacts the correlation between tree length and the inferred α parameter. In other words, the more site rates are allowed to vary along the phylogenetic tree (used to simulate the sequence alignment), the more uniform the inferred distribution of site rates will appear (as α increases, individual site rates become increasingly similar).

Logarithmic regressions also indicated that, in both metazoan and plant sequence alignments, there is a positive correlation between the estimated tree length of the

inferred phylogeny (corresponding to each sequence alignment), and the inferred α parameter of the site rate distribution (Figure 4). Moreover, interaction tests show that paralogous gene clusters tend to have significantly higher estimated α parameters, relative to their normalized estimated tree length, than orthologous alignments ($p < 0.05$). This is true in metazoans when considering the original (Type I) paralogous alignments taken from Ahrens *et al*. (2016) and Ahrens *et al*. (2018), as well as new (Type II) paralogous alignments extracted from protein family clusters used in those previous studies. In plants, the difference in regression lines between Type II paralogs and putative orthologs is not as statistically significant ($p = 0.08$).

Regression plots comparing inferred phylogenetic tree lengths to actual phylogenetic trees (used to simulate protein sequence alignments in this study) suggest that MrBayes (Ronquist et al. 2012) tends to underestimate tree lengths as the actual phylogenetic trees become very divergent (i.e., as the true total tree length increases). However, when the true phylogenetic trees are relatively small (e.g., tree lengths between 0.0 – 10.0), the inferred tree lengths become more accurate. Notably, the majority of inferred phylogenies associated with actual protein sequence data (from metazoans and plants) have estimated tree lengths less than 10.0.

DISCUSSION

Protein Sequence Data and Phylogenetic Analysis

The infeasibility of maintaining taxonomic evenness (i.e., a similar degree of representation for each species) in large-scale datasets has been discussed previously in Ahrens et al. (2016) and reiterated in Ahrens et al. (2018). Much of this difficulty arises

from the inherent phylogenetic unevenness of publicly available protein sequence databases, which tend to be enriched with model organism proteomes, primarily stemming from a relatively small number of clades (e.g., arthropods, chordates, angiosperms). The result is that, when using agglomerative techniques (e.g., single-linkage clustering) to group sequence data into homologous clusters, organisms from underrepresented clades (e.g., echinoderms, poriferans, nematodes) will be underrepresented in orthologous sequence clusters as well. Indeed, our dataset contains several times more human (*Homo sapiens*) and mouse (*Mus musculus*) sequences in orthologous alignments (5355 and 5324, respectively) than *Caenorhabditis elegans* or *Amphimedon queenslandica* sequences (456 and 616, respectively), most likely because *C. elegans* and *A. queenslandica* are the only representatives of their respective phyla (Nematoda and Porifera).

In addition to the uneven taxonomic representation within orthologous sequence alignments, paralogous alignments exhibit apparent bias as well (see Figure 1). This is also largely attributable to the phylogenetic unevenness of the datasets, since paralogs of more closely related species are more likely to be found together in Type II paralog groups (i.e., extracted from clusters originally containing a mixture of orthologs and paralogs) than in Type I paralog groups (species-specific groups taken directly from the original datasets). Conversely, paralogs from species with few close relatives (and hence, few close orthologs) tend be more well represented in Type I paralog groups than in Type II groups.

Even as the number of publicly-available proteomes continues to increase, phylogenetic unevenness will ultimately remain a feature of many large-scale molecular

datasets for a variety of reasons. For one, the relative diversity of extant phyla varies widely in nature. *Trichoplax adhaerens* is nearly the only extant representative of the phylum Placozoa (Eitel et al. 2018), whereas the true number of extant species in the phylum Arthropoda—which includes insects, crustaceans and arachnids—has proven difficult to even estimate (Stork et al. 2015). Furthermore, variation in proteome size as well as differences in gene duplication history determine the number and size of homologous sequence groups which can be constructed for a particular taxon. Nonetheless, we observe consistent differences in sequence divergence patterns between orthologous and paralogous proteins in both plant and animal sequence alignments, despite the inherent phylogenetic unevenness in both datasets.

Statistical Analysis of Simulated Data

A range of augmented statistical evolutionary models—attempting to account for heterotachy explicitly—have been available for several decades (see Philippe et al. 2003). These include the simple covarion model (Fitch and Markowitz 1970; Tuffley and Steel 1998) as well as the more complex "covarion-like" model proposed by Galtier (2001), wherein site rates are allowed to change over the length of a phylogeny. Both of these models introduce relatively few parameters for the sake of computational tractability, and studies have revealed that accounting for heterotachy in protein evolution is advantageous for phylogenetic reconstruction (Lopez et al. 2002) as well as detecting positive selection (Siltberg and Liberles 2002) and functional divergence (Gaucher et al. 2001; Philippe et al. 2003).

The two models we implemented here to simulate sequence evolution with heterotachy (see Methods) were designed to clearly illustrate the effects of heterotachy on

phylogenetic tree inference when site-specific amino acid replacement rates are assumed (by the inference software) to be constant over the length of the tree (i.e., when heterotachy is ignored). In simulations where site rates are constant and gamma-distributed, MrBayes (Ronquist et al. 2012) quite accurately predicts the true α shape parameter of the gamma distribution even when i) the number of simulated rate categories (16) exceeds the number of allowed categories for inference (4) and ii) the tree lengths used to simulate the data are very large (Figure 3A). However, when heterotachy is introduced to the simulations, we observe a positive correlation between sequence divergence (tree length / number of sequences) and the estimated α value. This effect increases with elevated levels of heterotachy (Figure 3).

The positive correlation we observe between sequence divergence and the inferred α parameter provides crucial insight into the expected behavior of phylogenetic inference software when the model of sequence evolution is misspecified. Essentially, when rates of amino acid replacement are allowed to change among lineages, a single alignment site may actually be governed by a complex mixture of replacement rates (i.e., in different lineages and at different times), and when the software infers the replacement rate at a particular alignment site, it actually provides a point estimate of this mixture. Thus, as phylogenetic tree length and heterotachy are increased, site-specific rate estimates tend to appear more uniform (i.e., they tend toward the mean rate score), meaning that the inferred α parameter of the gamma distribution becomes large (i.e., the distribution centers more tightly around the mean value).

Importantly, even the models of sequence evolution accounting for heterotachy tend to make unrealistic simplifying assumptions. For instance, fitness effects of site-

specific amino acid replacements are partly governed by concomitant replacements at neighboring sites within the same protein sequence (Fitch and Markowitz 1970; Goldstein and Pollock 2016), as well as replacements in other protein sequences (Gao and Zhang 2003; Breen et al. 2012). Such epistatic interactions form the basis of contemporary "mechanistic" models of protein sequence evolution (see Pollock et al. 2017), where even sequences evolving under purifying selection are constrained by a constantly shifting set of site-specific amino acid preferences and replacement rates (Pollock et al. 2012). Our results show that phylogenetic inference, under a statistical model that fails to properly account for heterotachy, displays a particular relationship between measurable values (sequence divergence and α parameter) when heterotachy is in fact a component of sequence simulation. Future simulation work incorporating more realistic parameters (e.g., variation in site-specific amino acid preferences, epistatic effects, etc.) may further illuminate the effects of model misspecification during phylogenetic inference.

Statistical Analysis of Protein Sequence Data

Previous studies have largely indicated that paralogous proteins diverge in function more readily than orthologous proteins (Altenhoff et al. 2012; Chen and Zhang 2012; Rogozin et al. 2014). Furthermore, prior work has demonstrated a link between protein functional divergence and site-specific shifts in sequence evolution (Gaucher et al. 2002; Philippe et al. 2003). Statistical methods have even been developed to evaluate functional divergence based on measurements of molecular evolution (e.g., Gu 1999; Gaucher et al. 2002; Siltberg and Liberles 2002; Gu 2003; Abhiman et al. 2006; Gu et al. 2013).

Notably, the link between site-specific rate shifts and functional divergence is not entirely clear. Gribaldo et al. (2003) assert that heterotachy is actually a common feature of neutral (i.e., non-adaptive) sequence evolution, but that functional divergence is specifically associated with sudden amino acid replacements in highly conserved sites, which they refer to as "constant but different" (CBD) sites. Studer and Robinson-Rechavi (2010) showed that indicators of functional divergence (e.g., CBD sites) can be found among both orthologous and paralogous proteins, and they did not observe a difference between the two sequence groups.

Our results show that, similar to the protein simulations discussed above, there is a relationship between sequence divergence (tree length / number of sequences) and the estimated α parameter of the gamma rate distribution in real protein sequence alignments (Figure 4). Additionally, in both metazoans and plants, as sequences diverge, the inferred α value increases more quickly in paralogous alignments than in orthologous alignments. This implies, based on our simulations, that heterotachy is more prevalent on average in paralogous alignments than in orthologous alignments. Given that heterotachy is often associated with functional change (e.g., Gaucher et al. 2002; Abhiman et al. 2006), these results provide compelling evidence in favor of the ortholog conjecture, across several thousand groups of homologous sequences, from two divergent eukaryotic lineages.

In both the metazoan and plant datasets, we also observe a difference in divergence patterns between Type I paralogous alignments (taken directly from the original dataset) and Type II paralogous alignments (extracted from mixed clusters of orthologous and paralogous sequences), wherein Type I alignments tend have higher α values than Type II alignments (Figure 4). While the biological distinction between these

two alignment types is not entirely clear, a possible explanation for this discrepancy in α values is that a large number of the Type I paralogous alignments contain sequences corresponding to small-scale, lineage-specific gene duplications, many of which will eventually become pseudogenes (Wagner 1998; Lynch and Conery 2000). Whereas Type II alignments originally contained a mixture of multiple species, often implying gene duplication events which precede several speciation events, there is no direct evidence (within our datasets) suggesting that Type I paralogs are found in other lineages (although as discussed above, this is partly an artifact of phylogenetically uneven datasets).

Per the ortholog conjecture, functional divergence is often associated with duplicated genes, where clades of orthologous genes (ortholog groups) retain similar structure and function, but paralogs differ (see Dos Santos and Siltberg-Liberles 2016). However, noteworthy counterexamples to this trend can also be observed. For example, the tumor suppressor protein, p53, is part of a family of three paralogs (p53, p63 and p73), but a recent study found strong evidence of ongoing functional divergence among p53 orthologs, as well as more sequence divergence (inferred via branch length) among the p53 orthologs than between the other two paralogs (p63 and p73) in the family (dos Santos et al. 2016). While we observe consistent differences in the large-scale divergence patterns of orthologous and paralogous sequences, the overlapping regions of our results (Figure 4) are consistent with previous findings (see Studer and Robinson-Rechavi 2010; dos Santos et al. 2016) indicating that both orthologs and paralogs can undergo functional divergence over time.

The criteria we used to delineate orthologous and paralogous sequence alignments in the present study are relatively simple. Given the nature of our alignments (i.e., single-linkage sequence clusters) and the scale of our dataset (thousands of alignments), it was not feasible to account for the potential misidentification of out-paralogs (i.e., paralogs from different species) as orthologs (see Koonin 2005), nor could we reliably distinguish ohnologs (paralogs in a whole genome duplication) from small-scale paralogs (resulting from small-scale duplication). Nonetheless, our analysis constitutes i) a statistical test of the ortholog conjecture against a clear null hypothesis (i.e., no difference in divergence pattern) where ii) the data were evaluated in an explicit phylogenetic context (see Dunn et al. 2018) and iii) our analysis cannot be impacted by biased functional annotation (see Altenhoff et al. 2012; Chen and Zhang 2012; Rogozin et al. 2014). Additionally, our results show an apparent difference in sequence divergence patterns between orthologs and paralogs which, to our knowledge, has not been previously reported in the literature.

METHODS

Protein Sequence Data Collection

Clusters of homologous protein sequences from 24 metazoan (animal) species (plus the choanoflagellate *Monosiga brevicollis*) as well as 24 plant species were taken from previous datasets used in Ahrens et al. (2016) and Ahrens et al. (2018). These clusters were originally generated using the graph-based single-linkage clustering program BLASTClust (Altschul et al. 1990) with a pairwise sequence identity threshold of 40% and a pairwise length threshold of 90%. We used MAFFT (Katoh and Standley 2013) to align all clusters containing at least 5 sequences. Any alignments with i) a

minimum pairwise sequence identity of at least 30% (but less than 100%) and ii) a minimum alignment coverage (ratio of sequence length to alignment length) greater than 50% were retained for further analysis.

Many of the alignments from the above datasets contained exactly one protein sequence from each represented species (e.g., 5 protein sequences from 5 different species). These "non-redundant" alignments were classified as putative orthologs (Figure 2). Other alignments were "species-specific," containing several different protein sequences from exactly one species, and were classified as Type I paralogs (i.e., a paralog group identified directly in the initial set of alignments). The remaining multispecies alignments contained protein sequences wherein some species were represented more than once (e.g., 10 sequences from only 5 different species), indicating that the sequences were connected by a mixture of orthologous and paralogous relationships (Figure 2). Within these mixed alignments, any subset of 5 or more sequences originating from a single species was extracted and classified as a Type II paralog group (Figure 2). These new paralogous alignments were edited via Trimal (Capella-Gutierrez et al. 2009) to eliminate any gap-only sites which were created when the sequences from other species were removed.

Sequence Alignment Simulation

A total of 200 phylogenetic trees, each containing 20 terminal nodes (leaves), were generated via the birth-death model implemented in Dendropy (Sukumaran and Holder 2010) using a mean birth rate of 1.0 (s.d.: 0.1) and a mean death rate of 0.5 (s.d., 0.1). Each tree was randomly rescaled by a factor between 0.0 and 3.0 to produce a sample of phylogenies exhibiting a wide range of tree lengths. Using these phylogenies,

1800 sequence alignments, each containing 1600 gap-free sites, were simulated in the Pyvolve module developed by Spielman and Wilke (2015). All simulated alignments were generated using the fixed amino acid rate matrix developed by Jones et al. (1992).

The first 600 simulations introduced site-specific rate heterogeneity wherein site rates were drawn from a discrete (16-category) gamma distribution. In other words, these alignments were simulated under the JTT + Γ model (see Yang and Kumar 1996; Darriba et al. 2011). The shape parameter α of the discrete gamma distribution was fixed at one of three values (α = 0.5, 1.0, 5.0) for each simulation, resulting in three sets (of 200 alignments) generated under differing degrees of rate heterogeneity.

The next 600 simulations introduced heterotachy (i.e., variation in lineage-specific site rates) using a simple random walk model. Initial site rates at the root node of each tree were drawn from a 16-category discrete gamma distribution with shape parameter α = 0.5. At every descendant node in the tree, each site rate R_i was iteratively modified by drawing a value V_i from a normal distribution, with mean equal to 0.0 and variance equal to the product of the node's branch length L_i and an additional constant C. The new site rate was then set to $(R_i + V_i)$ *mod* 10.0, effectively imposing a random walk over the range [0.0, 10.0]. Similar to the first 600 simulations, the random walk constant C was set to one of three values (C = 0.05, 0.5, 1.0) for each simulation, resulting in three sets of alignments generated under differing degrees of heterotachy.

The final 600 simulations introduced heterotachy by allowing pairs of sites to randomly exchange rates along the length of each tree. Again, initial (root node) site rates were drawn from a 16-category gamma distribution (α = 0.5). However, at every descendant node, N_i randomly-chosen site pairs were allowed to swap rate scores, where

N_i is a random integer drawn from a Poisson distribution with rate parameter λ equal to the product of the node's branch length B_i and a constant C. Again, the rate parameter constant C was set to one of three values (C = 10.0, 50.0, 200.0) for each simulation, resulting in alignments generated under differing degrees of heterotachy. Notably, this method of introducing heterotachy differs from the random walk simulations (described previously) in that all nodes in each tree are guaranteed to have gamma rate distributions with identical shape parameters (i.e., α = 0.5), but the exact arrangement of fast or slow site rates may differ substantially at each individual tree node.

Phylogenetic Analysis

Phylogenetic analyses for each protein sequence alignment, as well as for all simulated alignments, were performed in MrBayes 3.2.2 (Ronquist et al. 2012) with tree-bisection-reconnection (TBR) moves disabled. Phylogenies were estimated using the "mixed-model" approach (variable matrix plus gamma-distributed site rates), and the shape parameter α of an underlying 4-category gamma distribution was estimated from the data. Each analysis was run for 5,000,000 generations or until the average standard deviation of split frequencies fell below 0.005. After discarding the first 25% of generations as "burn-in," we recorded the mean estimated tree length as well as the mean estimated gamma shape parameter (α) for each analysis.

Statistical Analysis

Regression analyses and interaction tests were performed in R (Ihaka and Gentleman 2012) to examine the correlation between normalized tree length (i.e., the estimated mean tree length divided by the number of terminal nodes) and the estimated mean gamma shape parameter α in each sequence alignment. Separate regression

analyses were performed for i) ortholog groups, ii) original (Type I) paralog groups and iii) new (Type II) paralog groups identified in the animal and plant datasets, respectively. Regression analyses were also performed for all subgroups of simulated protein alignments described above. Visualization of all regression analyses was accomplished using the ggplot2 library (Wickham 2009).

LITERATURE CITED

Abhiman S, Daub CO, Sonnhammer ELL. 2006. Prediction of Function Divergence in Protein Families Using the Substitution Rate Variation Parameter Alpha. Mol. Biol. Evol. 23:1406–1413.

Ahrens J, Rahaman J, Siltberg-Liberles J. 2018. Large-Scale Analyses of Site-Specific Evolutionary Rates across Eukaryote Proteomes Reveal Confounding Interactions between Intrinsic Disorder, Secondary Structure, and Functional Domains. Genes (Basel). 9:553.

Ahrens J, Dos Santos HG, Siltberg-Liberles J. 2016. The Nuanced Interplay of Intrinsic Disorder and Other Structural Properties Driving Protein Evolution. Mol. Biol. Evol. 33:2248–2256.

Ahrens JB, Nunez-Castilla J, Siltberg-Liberles J. 2017. Evolution of intrinsic disorder in eukaryotic proteins. Cell. Mol. Life Sci. 74:3163–3174.

Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C. 2012. Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs.Eisen JA, editor. PLoS Comput. Biol. 8:e1002514.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene Ontology: tool for the unification of biology. Nat. Genet. 25:25–29.

Benson DA Karsch-Mizrachi I Lipman DJ Ostell J Sayers EW. 2009. GenBank. Nucleic Acids Res. 37:D26–D31.

Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA. 2012. Epistasis as the primary factor in molecular evolution. Nature 490:535–538.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973.

Chen X, Zhang J. 2012. The Ortholog Conjecture Is Untestable by the Current Gene Ontology but Is Supported by RNA Sequencing Data.Ouzounis CA, editor. PLoS Comput. Biol. 8:e1002784.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164–1165.

Dunn CW, Zapata F, Munro C, Siebert S, Hejnol A. 2018. Pairwise comparisons across species are problematic when analyzing functional genomic data. Proc. Natl. Acad. Sci. U. S. A. 115:E409–E417.

Eitel M, Francis WR, Varoqueaux F, Daraspe J, Osigus H-J, Krebs S, Vargas S, Blum H, Williams GA, Schierwater B, et al. 2018. Comparative genomics and the nature of placozoan species.Tyler-Smith C, editor. PLOS Biol. 16:e2005359.

Felsenstein J. 1973. Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters. Syst. Zool. 22:240.

Fitch WM. 1971. The nonidentity of invariable positions in the cytochromes c of different species. Biochem. Genet. 5:231–241.

Fitch WM, Markowitz E. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. 4:579–593.

Galtier N. 2001. Maximum-Likelihood Phylogenetic Analysis Under a Covarion-like Model. Mol. Biol. Evol. 18:866–873.

Gao L, Zhang J. 2003. Why are some human disease-associated mutations fixed in mice? Trends Genet. 19:678–681.

Gaucher EA, Gu X, Miyamoto MM, Benner SA. 2002. Predicting functional divergence in protein evolution by site-specific rate shifts. Trends Biochem. Sci. 27:315–321.

Gaucher EA, Miyamoto MM, Benner SA. 2001. Function-structure analysis of proteins using covarion-based evolutionary approaches: Elongation factors. Proc. Natl. Acad. Sci. 98:548–552.

Goldstein RA, Pollock DD. 2016. The tangled bank of amino acids. Protein Sci. 25:1354–1362.

Gribaldo S, Casane D, Lopez P, Philippe H. 2003. Functional Divergence Prediction from Evolutionary Analysis: A Case Study of Vertebrate Hemoglobin. Mol. Biol. Evol. 20:1754–1759.

Gu X. 1999. Statistical Methods for Testing Functional Divergence after Gene Duplication. Mol. Biol. Evol 16:1664–1674.

Gu X. 2003. Functional divergence in protein (family) sequence evolution. Genetica 118:133–141.

Gu X, Zou Y, Su Z, Huang W, Zhou Z, Arendsee Z, Zeng Y. 2013. An Update of DIVERGE Software for Functional Divergence Analysis of Protein Family. Mol. Biol. Evol. 30:1713–1719.

Hughes T, Liberles DA. 2008. Whole-Genome Duplications in the Ancestral Vertebrate Are Detectable in the Distribution of Gene Family Sizes of Tetrapod Species. J. Mol. Evol. 67:343–357.

Ihaka R, Gentleman R. 2012. R: A Language for Data Analysis and Graphics. J. Comput. Graph. Stat. 5:299–314.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8:275–282.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 30:772–780.

Koonin E V. 2005. Orthologs, paralogs, and evolutionary genomics. Annu. Rev. Genet. 39:309–338.

Lopez P, Casane D, Philippe H. 2002. Heterotachy, an Important Process of Protein Evolution. Mol. Biol. Evol. 19:1–7.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155.

Nehrt NL, Clark WT, Radivojac P, Hahn MW. 2011. Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals.Rzhetsky A, editor. PLoS Comput. Biol. 7:e1002073.

Philippe H, Casane D, Gribaldo S, Lopez P, Meunier J. 2003. Heterotachy and Functional Shift in Protein Evolution. IUBMB Life (International Union Biochem. Mol. Biol. Life) 55:257–265.

Pollock DD, Pollard ST, Shortt JA, Goldstein RA. 2017. Mechanistic Models of Protein Evolution. In: Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts. Cham: Springer International Publishing. p. 277–296.

Pollock DD, Thiltgen G, Goldstein RA. 2012. Amino acid coevolution induces an evolutionary Stokes shift. Proc. Natl. Acad. Sci. 109:E1352–E1359.

Rastogi S, Liberles DA. 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evol. Biol. 5:28.

Rogozin IB, Managadze D, Shabalina SA, Koonin E V. 2014. Gene family level comparative analysis of gene expression in mammals validates the ortholog conjecture. Genome Biol. Evol. 6:754–762.

Ronquist F, Huelsenbeck JP, Teslenko M. 2011. mb3.2_manual.pdf. MrBayes version 3.2 Man. Tutorials Model S.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice across a Large Model Space. Syst. Biol 61:539–542.

dos Santos HG, Nunez-Castilla J, Siltberg-Liberles J. 2016. Functional Diversification after Gene Duplication: Paralog Specific Regions of Structural Disorder and Phosphorylation in p53, p63, and p73.Roemer K, editor. PLoS One 11:e0151961.

dos Santos HG, Siltberg-Liberles J. 2016. Paralog-Specific Patterns of Structural Disorder and Phosphorylation in the Vertebrate SH3–SH2–Tyrosine Kinase Protein Family. Genome Biol. Evol. 8:2806–2825.

Sayers EW Barrett T Benson DA Bryant SH Canese K Chetvernin V Church DM DiCuccio M Edgar R Federhen S, et al.. 2009. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 37:D5–D15.

Siltberg J, Liberles DA. 2002. A simple covarion-based approach to analyse nucleotide substitution rates. J. Evol. Biol. 15:588–594.

Spielman SJ, Wilke CO. 2015. Pyvolve: A Flexible Python Module for Simulating Sequences along Phylogenies.Robinson-Rechavi M, editor. PLoS One 10:e0139047.

Stork NE, McBroom J, Gely C, Hamilton AJ. 2015. New approaches narrow global species estimates for beetles, insects, and terrestrial arthropods. Proc. Natl. Acad. Sci. U. S. A. 112:7519–7523.

Studer RA, Robinson-Rechavi M. 2009. How confident can we be that orthologs are similar, but paralogs differ? Trends Genet. 25:210–216.

Studer RA, Robinson-Rechavi M. 2010. Large-Scale Analysis of Orthologs and Paralogs under Covarion-Like and Constant-but-Different Models of Amino Acid Evolution. Mol. Biol. Evol. 27:2618–2627.

Sukumaran J, Holder MT. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571.

Teufel AI, Liu L, Liberles DA. 2016. Models for gene duplication when dosage balance works as a transition state to subsequent neo- or sub-functionalization. BMC Evol. Biol. 16:45.

Tuffley C, Steel M. 1998. Modeling the covarion hypothesis of nucleotide substitution. Math. Biosci. 147:63–91.

Wagner A. 1998. The fate of duplicated genes: loss or new function? BioEssays 20:785–788.

Wickham H. 2009. Ggplot2: elegant graphics for data analysis. Springer

Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. 11:367–372.

Yang Z, Kumar S. 1996. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. Mol. Biol. Evol. 13:650–659.

Figure Captions

Figure 1. Phylogenetic trees showing the 24 animal species plus M. brevicollis (top) and the 24 plant species (bottom) used in this study. Columns to the right of each species show the number of clusters (separated by cluster type) in which each species can be found. Phylogenies are based on the NCBI Common Taxonomy Tree (Sayers et al. 2009; Benson et al. 2009).

Figure 2. Illustration of the three types of sequence clusters used in this study. Putative orthologs (top-left) are homologous sequence clusters with exactly one sequence per species (e.g., S1-S5). Type I paralogs (top-right) are sequence clusters in which all sequences correspond to the same species (e.g., S1). Many sequence clusters contained a mixture of orthologous and paralogous genes (bottom-left). If at least 5 of the genes in such a cluster corresponded to the same species, they were extracted and placed in a Type II paralog cluster (bottom-right).

Figure 3. Loess regressions showing the relationship between the mean phylogenetic tree length (normalized by the number of sequences in each cluster) and the mean estimated α parameter (of the gamma rate distribution) for simulated sequence datasets with A) no heterotachy (fixed $\alpha = 0.5,1.0,5.0$), B) a random-walk heterotachy model ($\alpha = 0.5$) and C) a rate-swap heterotachy model ($\alpha = 0.5$). A parameter "C" is used (fig B,C) to control the degree of heterotachy in each simulation, and larger values of C indicate more heterotachy. Note that the estimated α parameter is close to the true value when sequences are simulated without heterotachy and the fixed α parameter is small (A). In all

other cases, there is a positive correlation between tree length and α, which increases with increasing heterotachy. Grey bands indicate 95% confidence intervals for each regression.

Figure 4. Log regressions showing the relationship between the mean phylogenetic tree length (normalized by the number of sequences in each cluster) and the mean estimated α parameter (of the gamma rate distribution) for all three cluster types in animals (top) and plants (bottom). Grey bands indicate 95% confidence intervals for each regression line. Note that in animals and plants, the line corresponding to orthologous sequence clusters is significantly lower, over most of the chart range, than both Type I and Type II paralog cluster regression lines.

Figures

Figure 1

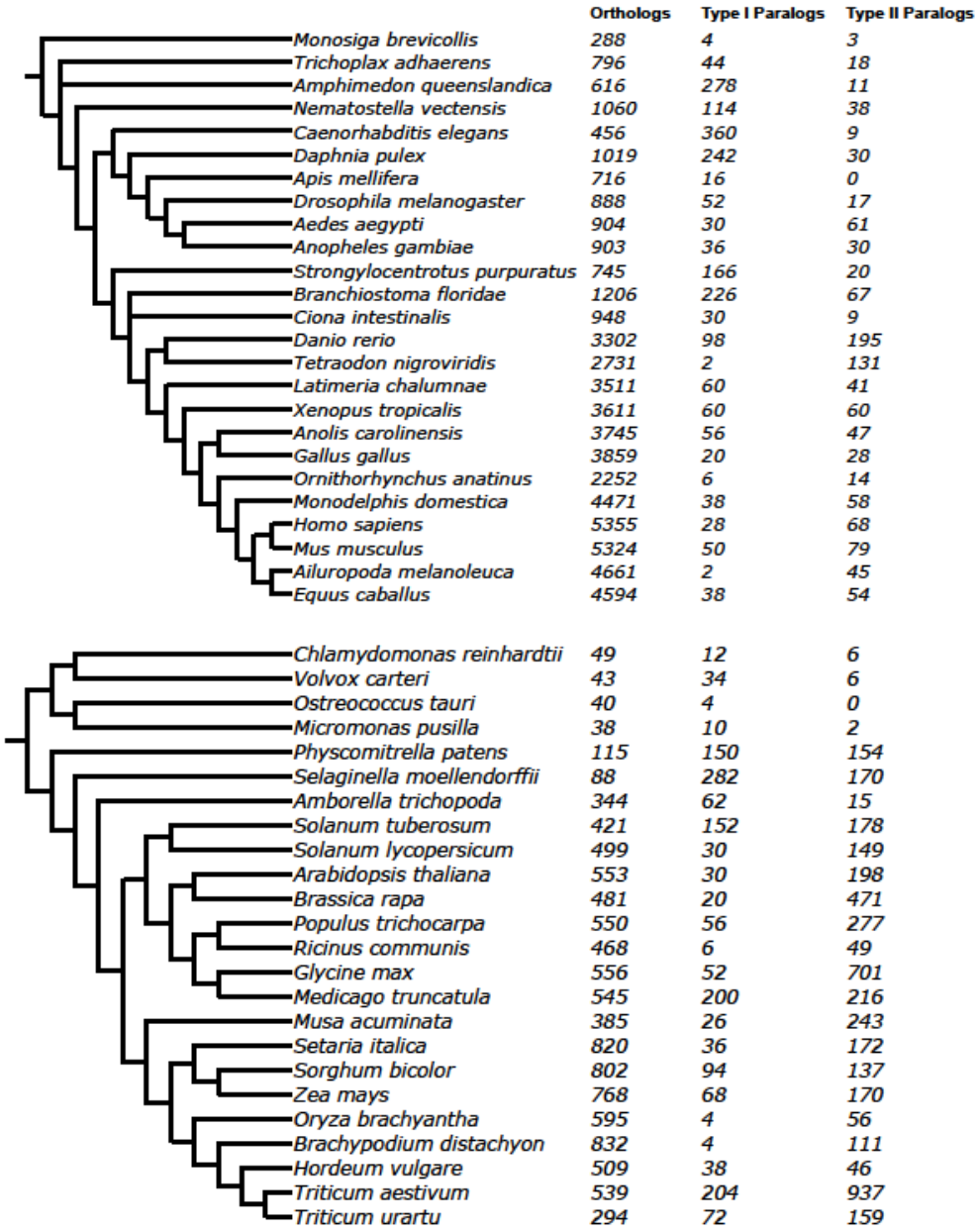


Figure 2

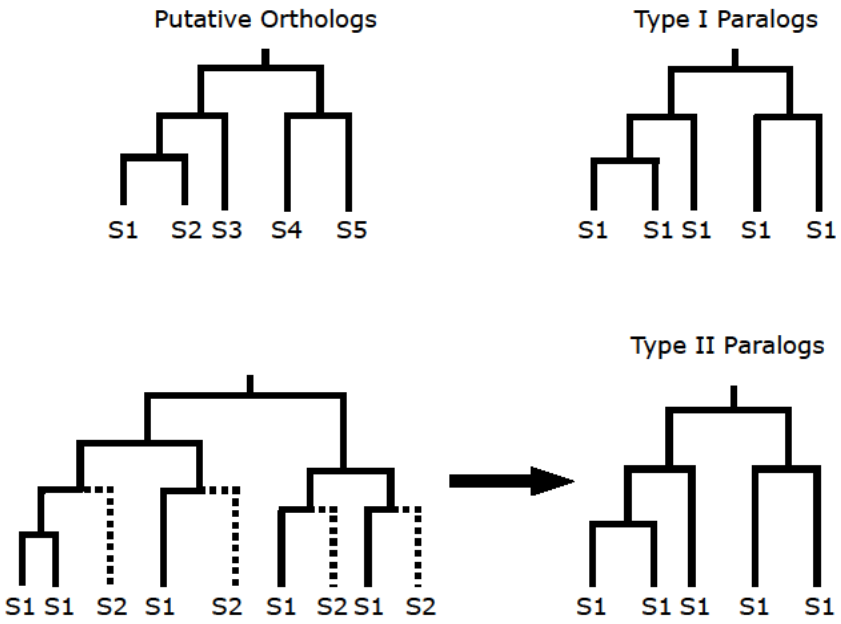

Putative Orthologs
S1 S2 S3 S4 S5
Type I Paralogs
S1 S1 S1 S1 S1
Type II Paralogs
S1 S1 S2 S1 S2 S1 S2 S1 S2
S1 S1 S1 S1 S1

Figure 3

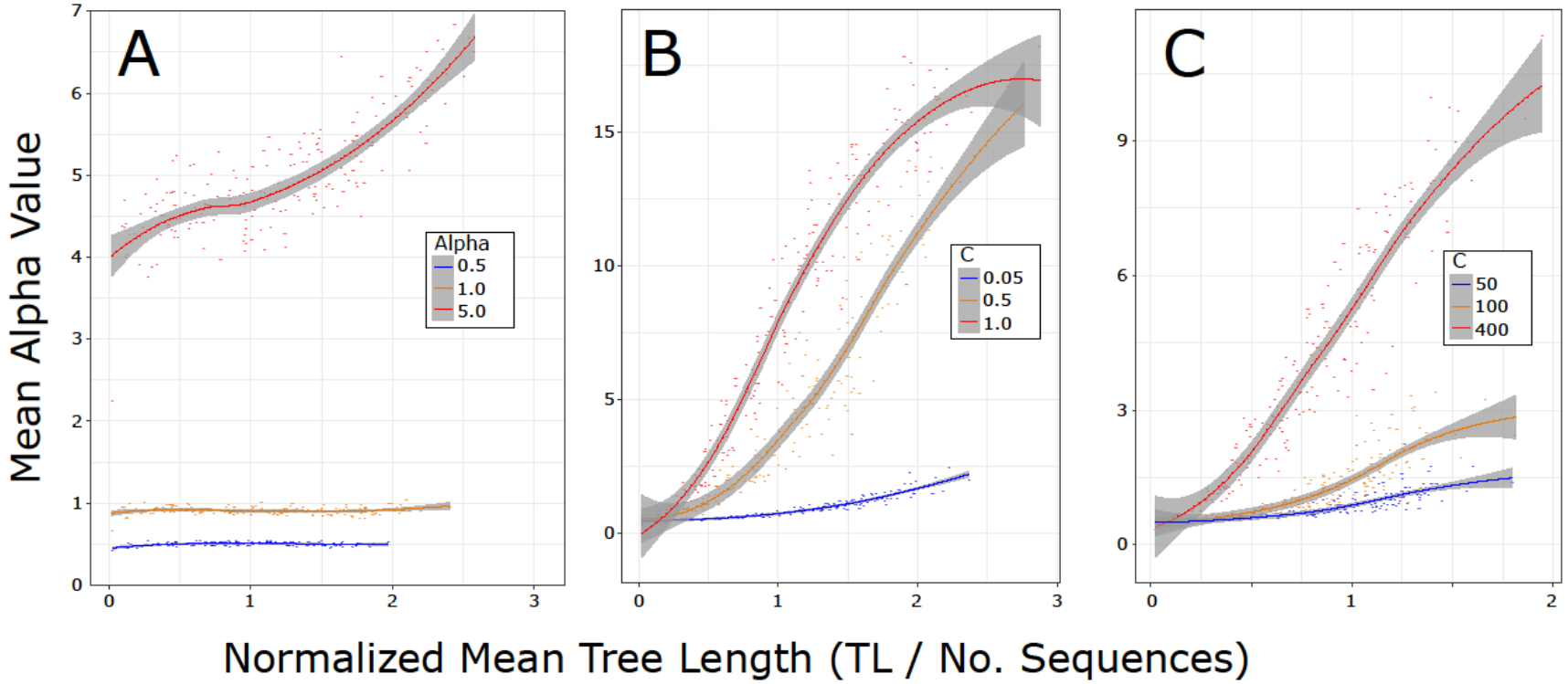
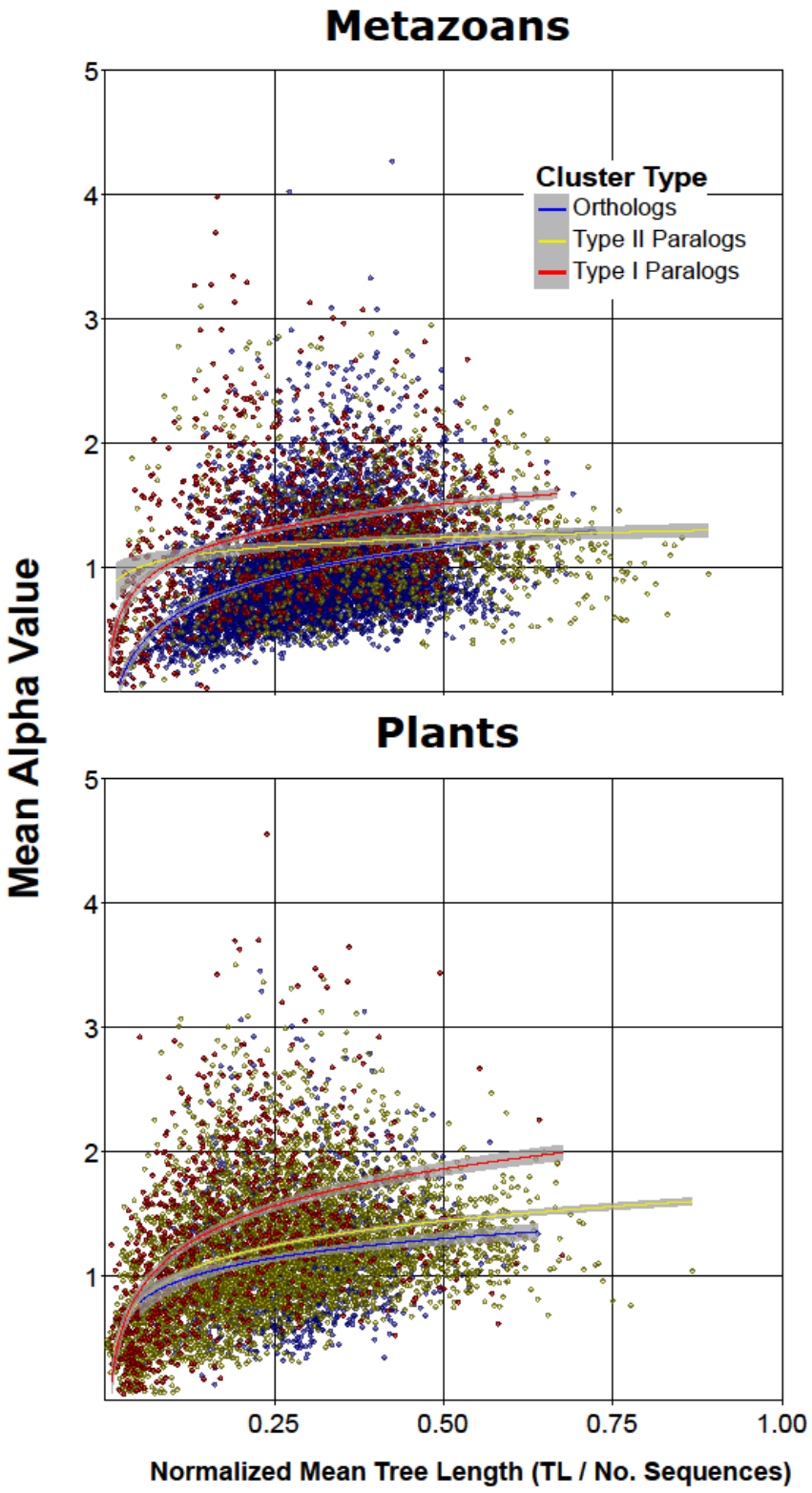

A
B
C
Mean Alpha Value
Normalized Mean Tree Length (TL / No. Sequences)
Alpha
0.5
1.0
5.0
C
0.05
0.5
1.0
C
50
100
400

Figure 4


Metazoans
Plants
Cluster Type
Orthologs
Type II Paralogs
Type I Paralogs
Mean Alpha Value
1
2
3
4
5
0.25
0.50
0.75
1.00
Normalized Mean Tree Length (TL / No. Sequences)

CHAPTER V

ACQUISITION OF HOMOLOGOUS PROTEIN SEQUENCE CLUSTERS FROM LOCAL DATABASES USING A SIMPLE, GRAPH-BASED SINGLE-LINKAGE CLUSTERING PROCEDURE

ABSTRACT

The identification of homologous groups of gene sequences is useful in a wide range of biological research applications, including phylogenetic inference and genome functional annotation. Numerous utilities are available to computational biologists for the task of sequence clustering: the agglomeration of similar biological sequences into subgroups or "clusters." Many of these clustering applications are designed to operate on entire sequence databases at once, and they often utilize incremental, greedy clustering strategies to work effectively on large databases. However, in cases where one is only interested in a small group of homologous sequences (e.g., a protein family), such large-scale applications may be too inflexible and time-consuming (since one has to cluster an entire database just to obtain the group of sequences they want to study). Here, we present a simple, graph-based single-linkage clustering procedure which uses an iterative search-and-filter approach to identify just one cluster of similar protein sequences based on a set of user-defined starting sequences and similarity cutoffs. We describe a simple implementation of this procedure involving the Basic Local Alignment Search Tool (BLAST) and the BioPython library for the Python programming language. We also benchmark the performance of our implementation (runtime relative to cluster size) using 49 sequences from a eukaryote proteome database (24 animals + Monosiga brevicollis) and compare our results to an existing single-linkage application. Our results indicate that the composition of a single-linkage cluster is quite sensitive to the sequence alignment strategy employed to establish linkage. Additionally, we show that our clustering procedure can easily be used to recover subunits (Rpt proteins) of the eukaryotic regulatory ring of the 26S proteasome from all of the species in our benchmark database.

We use phylogenetic inference and sequence-based structure/function prediction methods to show that this sequence cluster contains a diverse (but homologous) set of protein sequences suitable for evolutionary analysis.

INTRODUCTION

The pace of biological sequence data collection (i.e., nucleotide and amino acid sequences from biological organisms) has significantly increased in the post-genomic era. As a result, the institutions responsible for storing and maintaining publicly-available sequence data, such as the European Bioinformatics Institute (EBI) and the National Center for Biotechnology Information (NCBI), must constantly increase their digital storage capacity, as well as their data accession and visualization tools, to accommodate ever-growing databases (Cook et al. 2016; Agarwala et al. 2018). The curation of such large sequence databases remains a crucial task for bioinformaticians, and the scalability of manual gene annotation in well-curated databases such as Swiss-Prot (Bairoch and Apweiler 2000) has been called into question (see Baumgartner et al. 2007, but see also Poux et al. 2017). Still, the availability of large-scale molecular datasets has already greatly enhanced our ability to study the complex functional relationships among genes (see Chen and Coppola 2018), illuminate the evolutionary origins of present-day organisms (see Koonin 2010; Telford et al. 2015) and determine/annotate the functional roles of homologous genes (Eisen 1998; Eisen and Wu 2002). In cases where comprehensive manual gene annotation is not required, canonical protein sequence databases (i.e., containing one representative sequence per gene), such as the canonical proteome dataset maintained by UniProt (Bateman et al. 2017), provide a useful starting

point for a diverse range of molecular studies, from comparative genomics to experimental molecular studies.

The total number of unique protein sequences ever to exist constitutes an extremely small fraction of protein sequence space: the set of all possible amino acid sequences (Salisbury 1969). This is partly because of the short time period in which life has existed (relative to the number of possible protein sequences), but also because of the ruggedness of the protein fitness landscape and, hence, the small number of acceptable amino acid replacements at any given moment in time (Smith 1970; Povolotskaya and Kondrashov 2010). The net result is that the portion of sequence space represented by real proteins exhibits a notable pattern (see Buchholz et al. 2017), where homologous proteins form "clusters" of similar sequences surrounded by large empty regions (i.e., sequences which are not found in living organisms). The identification of these clusters of biological sequences within large databases remains an ongoing challenge in computational biology, but efforts have been made to provide databases of sequence clusters replete with resolved phylogenies (evolutionary relationships) for future studies (see Huerta-Cepas et al. 2008). Many applications have also been developed to assist researchers in clustering nucleotide and protein databases by themselves. Commonly, these applications produce graph-based clusters, wherein members of a given cluster are viewed as nodes/vertices in a connected edge-weighted linkage graph. Often, the edge weights in the linkage graph represent a measurement (or a combination of measurements) of sequence similarity (Figure 1).

Because clustering a large number of biological sequences becomes computationally demanding, many of the applications designed to cluster entire databases

are primarily optimized for speed. For instance, the programs kClust (Hauser et al. 2013) and CD-Hit (Li and Godzik 2006) employ pre-filters to minimize the number of pairwise alignments necessary to establish linkage. Both kClust and CD-Hit also utilize a greedy, incremental strategy to construct linkage graphs quickly. Essentially, all sequences in the database are ranked in descending order of length, and the longest sequence is chosen as the representative of the first cluster. All other sequences are then compared to the representative sequence, and any sequences which meet user-defined linkage cutoffs (e.g., pairwise identity) are grouped with the representative. The longest remaining sequence (i.e., of the sequences which do not yet belong to a cluster) is then compared to the other remaining sequences, and the process is repeated until all sequences have been assigned to a group. The result of this process is that all sequences in a given cluster are directly linked to a single reference sequence (Figure 1 A-B). Importantly, clustering strategies such as these are order-dependent, meaning that sorting the sequences in ascending order of length instead, and then beginning with the shortest sequence as a representative, may result in a different set of sequence clusters.

An early form of graph-based sequence clustering, known as single-linkage clustering, was implemented in the (now deprecated) program BLASTClust (Altschul et al. 1990) as well as the newer program SiLiX (Miele et al. 2011). Single-linkage clusters are essentially defined as follows: 1) any two sequences, ***A*** and ***B***, are linked if they are sufficiently similar, according to a predefined set of linkage cutoffs (e.g., pairwise identity, coverage, alignment score, etc.) and 2) if ***A*** and ***B*** are linked, and ***B*** and ***C*** are linked, then ***A***, ***B*** and ***C*** are members of a single-linkage cluster. While single-linkage methods such as BLASTClust are more inclusive than the representative-based clustering

methods described above (Figure 1C), they are computationally slower (to the point of being infeasible when clustering a database containing more than a few million sequences), and currently-available programs are only designed to cluster entire databases.

Interestingly, if single-linkage cutoffs are defined appropriately, such that ***A*** and ***B*** are homologs and ***B*** and ***C*** are homologs, then ***A*** and ***C*** can be considered transitive homologs, and the overall cluster can be considered a group of homologous sequences (but see Miele et al. 2011 for important caveats). Additionally, single linkage clusters are non-overlapping and order-independent, meaning that in a particular database at a particular linkage cutoff, sequence ***A*** only belongs to one single-linkage cluster, and the composition of a given cluster does not depend on the order in which the clusters were constructed. Thus, within a very large sequence database, it is possible to define a single-linkage cluster without clustering the rest of the sequences in the database.

Here, we discuss a simple procedure for defining a single-linkage cluster in a protein sequence database using a combination of pairwise sequence identity and a bi-directional measurement of sequence alignment quality. We describe a straightforward implementation of this procedure using the Python programming language and the BLAST program (Altschul et al. 1990), and we benchmark its performance on a database of metazoan (animal) proteomes with a choanoflagellate outgroup species (Monosiga brevicollis) originally constructed by (Ahrens et al. 2016). Finally, we use phylogenetic inference and sequence-based structural/functional predictions to demonstrate that our procedure can recover large, divergent protein families, using sequences from the Rpt regulatory ring of the 26S proteasome complex as a notable example result. The

evolutionary history and paralog-specific structural/functional divergence observed in our single-linkage Rpt cluster are also discussed in context of recent discoveries relating to the diversification of the Rpt protein family.

RESULTS

Single-linkage Clustering Procedure

We developed an iterative single-linkage clustering procedure for amino acid sequence data using the BLAST sequence search program (Altschul et al. 1990) as well as the BioPython library (Cock et al. 2009) in the Python programming language (Rossum 1995). The procedure works by performing BLAST searches against a local BLAST-formatted database on a collection of $\boldsymbol{n} \geq 1$ user-defined starting sequences, which can either be taken directly from the local database or supplied externally, in a FASTA-formatted file. For each BLAST query, results can be filtered by expectation threshold (E-value) or a maximum number of BLAST search hits (i.e., subject sequences) can be specified beforehand. Then, each subject sequence is aligned to the query sequence to measure their pairwise sequence identity, and the alignment footprint coverage relative to the longer of the two sequences (i.e., the number of overlapping residues in the alignment divided by the length of the longer sequence). If both of these measurements are higher than the cutoff values specified by the user, the subject and query sequences are considered linked, and (if it is not already included in the single linkage cluster) the subject is added to the list of queries, so that the above procedure can be performed again, using the subject sequence as a BLAST query. Eventually, the entire list of BLAST queries will be searched (including the linked sequences which were

added by the procedure), and the final list of (already searched) queries are the members of the single-linkage cluster (see methods for a more detailed description).

The procedure also allows linkage measurements to be obtained from a variety of different alignment strategies. The simplest strategy (BLAST-SL) uses the original BLAST alignment (query vs. subject) to test for linkage. Because BLAST alignments are sub-optimal, a more exhaustive strategy (BLAST-BD) can be used to also check the reverse alignment (query vs. subject and subject vs. query) to see if the linkage measurements improve (i.e., BLAST-BD performs "bi-directional" alignment). Additionally, optimal alignments can be generated using the Smith-Waterman (SW) local alignment algorithm (Smith and Waterman 1981) implemented in BLAST, or the Needleman-Wunsch (NW) global alignment algorithm (Needleman and Wunsch 1970) implemented in BioPython. Because there can exist more than one optimal NW alignment for a given pair of sequences, the first 1000 NW alignments reported by BioPython are all measured and, if any one of them produces satisfactory identity/coverage measurements, the sequences are considered linked.

Benchmark Analyses

To evaluate the runtime of our clustering procedure under different parameter settings, and to test its performance against BLASTClust (Altschul et al. 1990), we selected 49 sequences from a database used in (Ahrens et al. 2016) and again in (Ahrens et al. 2018) (see methods for details). We created single-linkage clusters containing each of the 49 sequences using a 40% minimum sequence identity cutoff and a 90% alignment footprint coverage cutoff (i.e., the number of overlapping residues divided by the length of the longer sequence in the alignment must be at least 0.9), and we employed three

distinct sequence alignment strategies to establish linkage (Figure 2, 3). All 49 sequences aggregate into single-linkage clusters containing > 1 member in BLASTClust using a 40% identity cutoff and a 90% alignment coverage cutoff. However, 16 of the 49 sequences are identified as singletons (clusters with only one member) in our procedure when the BLAST-SL strategy is used to establish linkage. Further, 15 of the 16 sequences are still identified as singletons using BLAST-BD. This indicates that the alignment coverage threshold used by BLASTClust (which includes gaps when measuring alignment footprint length) can result in more inclusive single-linkage clusters than our procedure (which does not include gaps in the alignment footprint length). However, when using the SW + NW strategy to establish linkage, only 4 sequences are still identified as singletons, and 39 sequences are found in larger clusters than the single-linkage groups identified by BLASTClust.

Final runtimes for analyses resulting in singletons were relatively short (ranging from 4.2 to 46.1 seconds) regardless of the sequence alignment strategy used to establish linkage. The total runtime increases with cluster size, and clustering analyses that were run using the optimal sequence alignment strategy (SW + NW) show a tendency toward longer total runtimes than analyses using only BLAST-SL or BLAST-BD (Figure 2). Notably, the variance in runtime increases with cluster size as well. For instance, the longest analysis runtime (442,289.4 seconds; 122.9 hours) resulted in a cluster containing 112 sequences, whereas the analysis resulting in the largest cluster (887 sequences) terminated in far less time (22,727.8 seconds; 6.3 hours).

Analysis of the Rpt Protein Family

One of the clusters we obtained from our benchmark test contains 142 members of a protein family comprising the heterohexameric ring (Rpt) of the 19S regulatory particle (RP) in the 26S proteasome complex. After 15,000,000 generations, Bayesian MCMC analysis of the aligned sequence cluster in MrBayes (Ronquist et al. 2012) reached an average standard deviation of split frequencies (ASDSF) of 0.014. Although the observed ASDSF is slightly higher than the target convergence diagnostic (0.01) recommended by the program authors (Ronquist et al. 2011), the resulting 50% majority-rule consensus tree contained 6 highly-supported clades (posterior probability = 1.0) containing all 6 subunits of the eukaryotic Rpt ring.

Genes from all of the 25 species in the target BLAST database are represented in the Rpt ring cluster, and 10 of the species are represented exactly once in each of the 6 Rpt clades (Figure 4). Of the 6 species with multiple genes in at least one clade, only Strongylocentrotus purpuratus is not represented in all 6 clades (2 genes in Rpt4 clade plus Rpt1, Rpt5, Rpt6). The species with the highest number of genes in the cluster is *Drosophila melanogaster*—in addition to identifying all 6 subunits of the Rpt ring, we found three additional genes (Rpt3R, Rpt4R, Rpt6R) which appear to have arisen from more recent gene duplications (Figure 4).

Results from IUPRed (Dosztányi et al. 2005) revealed marked differences in site-specific intrinsic disorder propensities (i.e., the potential of each site to form stabilizing contacts) among the six clades we identified (Figure 5). For instance, the majority of sequences in clade containing Rpt2 genes possess an extended n-terminal region with high disorder propensity (low potential to form stabilizing contacts). Members of the

Rpt1 clade have two predicted domains: an AAA domain and a C-terminal AAA+ Lid domain, both of which are associated with ATPase activity. Most of the remaining sequences also contained a predicted oligonucleotide binding (OB) domain, with the exception of one sequence in the Rpt6 clade (corresponding to *Amphimedon queenslandica*), a second gene in the Rpt2 clade (from *Caenorhabditis elegans*) and a third sequence in the Rpt3 clade (from *Trichoplax adhaerens*). Lastly, a single gene in the Rpt5 clade from *Monosiga brevicollis* also contained a predicted N-terminal tRNA pseudouridine synthase D (TruD) domain.

DISCUSSION

Single-linkage Clustering Procedure

Many graph-based sequence clustering applications are currently available for database analysis, including reference-based clustering applications such as CD-Hit and kClust (Li and Godzik 2006; Hauser et al. 2013), single-linkage clustering methods like BLASTClust (Altschul et al. 1990) and Silix (Miele et al. 2011), and non-deterministic Markov clustering procedures like OrthoMCL (Li et al. 2003). These applications are often optimized for clustering entire sequence databases, and a description of their overall clustering procedure (see Hauser et al. 2013) indicates that this is their intended use case. By contrast, the clustering procedure we present here is not intended for partitioning an entire database into sequence clusters. Rather, the advantage of our procedure (relative to whole-database approaches) is that a particular single-linkage cluster, at a particular linkage threshold, can be defined in a sequence database without expending the computational resources (time or memory) to define all of the other clusters at that same

threshold. This means that, if a researcher is only interested in mining a database to identify a particular homologous gene cluster (or a small number of gene clusters), the size of the target database is not as problematic (i.e., because the clustering procedure only needs to spend computational resources defining the groups of interest). Additionally, the user input for our procedure does not have to be a sequence from the target database (external query sequences can be supplied in a FASTA file), and the linkage cutoffs, BLAST pre-filters (e-value, number of hits) and alignment strategies can be set individually for each sequence cluster. Ultimately, these features allow for efficient and flexible acquisition of sequence clusters from large databases.

Benchmark Analyses

Because most sequence clustering applications focus on partitioning entire databases according to a pre-defined set of similarity cutoffs, their benchmark analyses also tend to evaluate database-level runtimes, or the time it takes the application to cluster a database of a given size (see Li et al. 2003; Li and Godzik 2006; Miele et al. 2011; Hauser et al. 2013). As mentioned previously, our procedure is more suitable for generating individual clusters within a database, so we focused instead on the relationship between cluster size and analysis runtime, as well as the way this relationship may change under more or less time-consuming alignment strategies. Our results show that individual runtimes can vary widely for specific single-linkage clusters, but runtimes generally increase with cluster size (Figure 2). Further, the longest runtimes in our benchmark correspond to large clusters (> 50 sequences) that were formed using optimal SW and NW alignment strategies. This makes sense because a BLAST search must be performed for every member of each single-linkage cluster (to identify potentially linked

sequences), and each BLAST search (Altschul et al. 1990) is computationally expensive, as it entails short word (k-mer) matching, local alignment, and E-value approximation across an entire database. Optimal sequence alignment (SW and NW) is also time-consuming, and up to 1,000 NW alignments may need to be evaluated per query/subject pair.

While BLASTClust includes gap characters when measuring alignment length and, subsequently, alignment footprint coverage, our procedure considers only the number of overlapping residues in a pairwise alignment (where neither sequence contains a gap character) when computing coverage. This means that our results are not directly comparable to BLASTClust, and indeed, many of the sequences which fall within single-linkage clusters in BLASTClust (at 40% identity, 90% coverage) are identified as singletons with our procedure using the same percentage cutoffs (but a more stringent form of coverage measurement). However, MAFFT alignments (Katoh and Standley 2013) of the 4 BLASTClust clusters containing sequences that we identified as singletons, even under the permissive SW + NW alignment strategy, appear to have relatively poor alignment quality, as the minimum alignment coverages (see Figure 6) of these clusters are all below 75%. Other developers have warned that clustering strategies permitting low alignment coverage can result in groups of sequences which have very different domain architectures (Miele et al. 2011), making the clusters unsuitable for many downstream analyses (alignment, phylogenetic inference, etc.). In light of this, while the coverage measurement used by BLASTClust is more sensitive (i.e., inclusive), we feel that our coverage measurement (considering only overlapping residues) is more appropriate for the task of agglomerating homologous sequences when their explicit

evolutionary history (phylogeny) is of interest. That being said, alignment strategy also appears to play a significant role in cluster composition and downstream multiple sequence alignment quality. Multiple sequence alignment of the non-singleton clusters identified by our procedure indicate that the SW + NW alignment strategy tends to produce clusters of more divergent sequences (Figure 7), so accounting for pairwise alignment strategy is important when determining appropriate linkage cutoffs.

Analysis of the Rpt Protein Family

The 26S proteasome is a large (roughly 2.5 megadalton) molecular machine that degrades protein sequences which have been labeled with ubiquitin (Voges et al. 1999; Komander and Rape 2012). In eukaryotes, it is composed of two main subunits: a 20S core particle (CP) and a 19S regulatory particle (RP) which together form the 26S proteasome complex (Voges et al. 1999; Bard et al. 2018). The 19S RP can be divided further into a 9-subunit "lid" complex (Lander et al. 2012; Lasker et al. 2012) and a 10-subunit "base" complex, which attaches to one (or both) open ends of the CP, and is responsible for unfolding and translocating targeted proteins to the interior of the CP for degradation (Lander et al. 2012; de la Peña et al. 2018).

The base complex of the 26S proteasome in eukaryotes includes 6 paralogous subunits (Rpt proteins, forming a ring in the order Rpt1, Rpt2, Rpt6, Rpt3, Rpt4, Rpt5), nearly all of which we were able to recover using our single-linkage clustering strategy (Figure 4). Recent structural determination via cryogenic electron microscopy (cryo-EM) has shown that these Rpt subunits of the lid complex are arranged in a "spiral-staircase" fashion and undergo substantial conformational changes while translocating a substrate (i.e., the protein targeted for degradation) to the internal proteolytic chamber of the CP

(de la Peña et al. 2018). The ensemble of cryo-EM structures elucidated by de la Peña et al. (2018) also reveals that the N-terminal regions of the individual Rpt subunits perform a diverse range of specific tasks, from the delineation of a path guiding the substrate into the CP for degradation (via Rpt2) to the stabilization of the β-hairpin structure in the ubiquitin-binding RPN11 lid subunit (via Rpt5).

Notably, the N-termini of the Rpt proteins in our single linkage cluster exhibit relatively fast rates of amino acid replacement, as well as variation in both length and intrinsic disorder propensity (Figure 5). However, there is considerable conservation in the overall disorder profiles (i.e., arrangement of ordered and disordered sites) within each of the six major clades in our phylogeny (Figure 5), and every member of the Rpt1 clade is missing the OB domain prediction from PFAM, which is found in nearly all other sequences in the phylogeny (though all 142 sequences contain at least part of the aligned region where the OB domain is predicted to be found). The appropriate placement of the root for the Rpt protein family phylogeny is unclear (see Wollenberg and Swaffield 2001; Fort et al. 2015) and because there is no discernible outgroup sequence in our dataset (i.e., the most divergent taxon, *Monosiga brevicollis*, is found in all 6 clades), we are displaying our consensus phylogeny as a midpoint-rooted tree. Nonetheless, the strong overall structural conservation seen in sequences within each main clade of our phylogeny (Figure 5), as well as the apparent structural differences among sequences in different clades, corroborate recent findings that different Rpt subunits play unique, complementary roles in the 26S proteasome.

The fact that some species are not represented in all 6 clades (Figure 4) does not necessarily imply that those Rpt subunits were lost in their respective organisms, because

the database we used for this benchmark omitted a subset of sequences from each organism's proteome, specifically any sequences that i) were less than 30 amino acids in length or ii) contained "X" characters, indicating ambiguous or unknown sequence data (see Ahrens et al. 2016). However, based on our results, many species certainly possess more than 6 copies of the Rpt gene, and in particular, *Anopheles gambiae*, *Drosophila melanogaster*, and *Caenorhabditis elegans* possess additional gene copies that are highly divergent in sequence, as measured by large tree distances between gene copies (Figure 4). The duplicated Rpt/RptR genes in D. melanogaster have been studied previously (Belote and Zhong 2009), and it appears that at least some of the duplicated 26S proteasome genes are crucial for normal spermatogenesis (Zhong and Belote 2007; Belote and Zhong 2009). Given the divergent Rpt gene copies we observe here in A. gambiae and C. elegans, it is possible that additional Rpt gene copies are a more general feature of ecdysozoans (i.e., arthropods and nematodes).

CONCLUSION AND FUTURE DIRECTIONS

Our aim here was to i) describe a simple but flexible single-linkage clustering procedure which can be used for targeted acquisition of homologous protein sequences for downstream analysis (alignment, phylogenetic inference, etc.), ii) provide benchmark data to illustrate how the procedure performs under various scenarios and iii) to show that the procedure can identify inclusive sequence groups representing divergent protein families (i.e., the Rpt ring of the 26S proteasome). In the future, we plan to release a standalone version of the clustering procedure which will serve as a flexible data-mining tool for other researchers.

METHODS

Algorithm and Implementation

We used the Basic Local Alignment Search Tool, or BLAST (Altschul et al. 1990), as the basis of our sequence search strategy. The remainder of the procedure was implemented in the Python programming language (Rossum 1995) using the Biopython library (Cock et al. 2009). The clustering protocol we implemented is based on a simple, graph-based single-linkage algorithm utilizing an iterative search-filter strategy which can be summarized as follows:

1. Let $\boldsymbol{Q}$ be a container for homologous sequences such that each sequence corresponds to a unique label (header). Initialize $\boldsymbol{Q}$ with user-defined sequences. User input can be a FASTA-formatted sequence file and/or a list of sequence accession codes from a local BLAST-formatted database $\boldsymbol{D}$ (see 3). $\boldsymbol{Q}$ will serve as the query queue for performing BLAST searches.
2. Let $\boldsymbol{C}$ be a separate container for labeled sequences. Initialize $\boldsymbol{C}$ as an empty container ($\boldsymbol{C} = \boldsymbol{\emptyset}$). (As BLAST searches are performed on queries $\boldsymbol{q_i} \in \boldsymbol{Q}$, they will be moved to $\boldsymbol{C}$ and eventually $\boldsymbol{C}$ will contain all sequences that belong to the single-linkage cluster.)
3. Let $\boldsymbol{D}$ be a set of labeled sequences in a BLAST-formatted database such that a substring of each label is recognized as a unique accession code. (Note that $\boldsymbol{Q}$ may contain a mixture of database sequences $\boldsymbol{q_i} \in \boldsymbol{D}$ and external query sequences $\boldsymbol{q_i} \notin \boldsymbol{D}$).
4. Let $\boldsymbol{H_i}$ be the set of all hit sequences $\boldsymbol{h_j} \in \boldsymbol{H_i}$ returned from a BLAST search against $\boldsymbol{D}$ using $\boldsymbol{q_i} \in \boldsymbol{Q}$ as a query sequence. Based on user BLAST specifications, $\boldsymbol{H_i}$ can be

pre-filtered to contain only the first $\boldsymbol{n}$ sequences or only sequences with a specified maximum e-value.

5. Let $\boldsymbol{f_{ij}}$ be the alignment footprint for $\boldsymbol{q_i} \in \boldsymbol{Q}$ vs. $\boldsymbol{h_j} \in \boldsymbol{H_i}$ (the number of residues that overlap when $\boldsymbol{q_I}$ and $\boldsymbol{h_j}$ are aligned).
6. Let $\boldsymbol{S(q_i,h_j)}$ be a function evaluating a user-defined set of conditions establishing whether or not a query sequence $\boldsymbol{q_i} \in \boldsymbol{Q}$ is sufficiently similar to a hit sequence $\boldsymbol{h_j} \in \boldsymbol{H_i}$. Allowable conditions include i) minimum pairwise sequence identity and ii) minimum alignment footprint coverage, i.e., $\boldsymbol{min(f_{ij}/q_i, f_{ij}/h_j)}$. In cases where $\boldsymbol{q_i} \in \boldsymbol{D}$ and $\boldsymbol{h_j} \in \boldsymbol{D}$, the bidirectional optimum sequence identity and alignment footprint coverage may also be considered (i.e., the values obtained when using $\boldsymbol{h_j}$ as a query to find the hit $\boldsymbol{q_i}$ in $\boldsymbol{D}$ may be used to establish linkage instead). Also, sequence identity and alignment footprint coverage may be computed from the first 1,000 optimal global alignments generated using the global alignment algorithm by Needleman and Wunsch (1970) (NW) as implemented in BioPython. Finally, an alignment can be generated using the optimal local alignment strategy developed by Smith and Waterman (1981) (SW) as implemented in BLAST. $\boldsymbol{S(q_i,h_j)}$ returns *True* if and only if all user-defined conditions (e.g., sequence identity > 40%, alignment footprint coverage > 90%) are simultaneously met for a single alignment of $\boldsymbol{q_i}$ vs. $\boldsymbol{h_j}$ (or $\boldsymbol{h_j}$ vs. $\boldsymbol{q_i}$).
7. While $\boldsymbol{Q} \neq \boldsymbol{\emptyset}$:

    Move one $\boldsymbol{q_i} \in \boldsymbol{Q}$ to $\boldsymbol{C}$ (i.e., $\boldsymbol{C} \leftarrow \boldsymbol{q_i}$)

    BLAST $\boldsymbol{q_i}$ against $\boldsymbol{D}$ to produce $\boldsymbol{H_i}$

    For each $\boldsymbol{h_j} \in \boldsymbol{H_i}$:

If $\boldsymbol{S(q_i,h_j)}$ = *True* and $\boldsymbol{h_j} \notin (\boldsymbol{C} \cup \boldsymbol{Q})$:

$$\boldsymbol{Q} \leftarrow \boldsymbol{h_j}$$

In essence, the procedure begins by performing a BLAST search for a sequence $\boldsymbol{q_i} \in \boldsymbol{Q}$ against the target database $\boldsymbol{D}$ and moving that searched sequence into $\boldsymbol{C}$. Any hit sequences which are linked to a given query based on $\boldsymbol{S(q_i,h_j)}$, but are currently neither in $\boldsymbol{Q}$ nor $\boldsymbol{C}$, are then added to $\boldsymbol{Q}$. The procedure repeats until $\boldsymbol{Q}$ is empty (i.e., $\boldsymbol{Q} = \boldsymbol{\emptyset}$), at which point $\boldsymbol{C}$ will contain all sequences in the single-linkage cluster at the linkage cutoff defined by $\boldsymbol{S(q_i,h_j)}$.

Importantly, if the linkage cutoffs are sufficiently relaxed (e.g., sequence identity: 0%, alignment footprint coverage: 0%), the above algorithm (see 7) will create a single-linkage cluster which contains every sequence in the target database $\boldsymbol{D}$. To avoid the creation of overly-inclusive clusters (and unnecessarily long run times) we have included a safeguard in our implementation, wherein the user can specify a maximum number of members in $\boldsymbol{C}$. If this upper bound is reached, the procedure halts and only returns the sequences found thus far.

Benchmark Tests

To evaluate the performance of our implementation of the above single-linkage clustering procedure, we used a BLAST-formatted database containing 25 eukaryotic proteomes (24 animals plus *Monosiga brevicollis*) which were originally used in Ahrens et al. (2016) and Ahrens et al. (2018). The database was filtered to exclude sequences which i) were less than 30 amino acids long or ii) contained "X" characters (i.e., ambiguous or missing sequence data). We selected 49 accession codes to initialize 49

single-sequence query queues (see ***Q*** above) and formed 49 single-linkage clusters using a minimum pairwise sequence identity threshold of 40% and a minimum alignment footprint coverage threshold of 90%. In Ahrens et al. (2016) and Ahrens et al. (2018), sequences were agglomerated into single-linkage clusters via BLASTClust (Altschul et al. 1990) using the similar thresholds (40% identity, 90% length), but BLASTClust includes internal alignment gaps when calculating the length of the alignment footprint, whereas our procedure only considers the alignment "footprint" to be the number of aligned non-gap characters. Thus, the footprint coverage threshold used in our procedure is actually more stringent than the one used by BLASTClust, so even though the numerical thresholds appear to be the same, the resulting single-linkage clusters are not necessarily identical.

For each of the 49 analyses, we recorded 1) the final number of sequences in the single-linkage cluster and 2) the run time required to produce the cluster. To evaluate the effect of sequence alignment strategies on cluster size and run time, we ran each single-linkage analysis three times. In the first run (BLAST-SL), linkage was determined based only on the initial BLAST sequence alignment. The second run (BLAST-BD) used the initial BLAST alignment as well as the alignment produced by using the hit sequence ($hj \in Hi$) as a search string to find the original query sequence ($qi \in Q$) in the database (note that because BLAST alignments are sub-optimal, these two alignments may not be identical, and hence may produce different measurements of sequence identity and alignment footprint coverage). The third run (SW + NW) considered an optimal local aligment produced by the Smith-Waterman (SW) algorithm, as well as all of the first 1000 optimal global alignments produced by the Needleman-Wunsch (NW) algorithm. In

all three runs, we used the same linkage cutoff (40% identity, 90% alignment footprint coverage). Clusters containing more than 1 sequence were aligned with MAFFT (Katoh and Standley 2013).

Analysis of the Rpt Protein Family

One of the single-linkage clusters obtained from our benchmark analysis contained a protein family whose members comprise the subunits of the Rpt regulatory ring of the eukaryotic 26S proteasome complex. We aligned the protein sequences in this cluster with MAFFT (Katoh and Standley 2013) using a local-pair alignment strategy. Bayesian Markov Chain Monte Carlo (MCMC) phylogenetic analysis was run for 15,000,000 generations in MrBayes3.2.2 (Ronquist et al. 2012) using a mixed-model strategy and assuming a 4-category gamma distribution among alignment site rates. We then inferred a 50% majority-rule consensus phylogenetic tree, discarding the initial 25% of generations as burn-in, and displayed the topology using a midpoint rooting strategy. Posterior probabilities for each clade (i.e., the fraction of trees in the MCMC chain which contain a given clade) were mapped to the corresponding internal nodes of the phylogeny.

For each sequence in the cluster, we inferred site-specific intrinsic disorder propensities using IUPred (Dosztányi et al. 2005) to predict long disordered regions. These propensities were then mapped onto the sequence alignment to help detect changes in disorder propensity among orthologous groups of sequences within the protein family. Functional domains were also inferred for each sequence using PFAM (Finn et al. 2014). Alignment site rates were inferred using the empirical Bayesian method implemented in the program Rate4Site (Mayrose et al. 2004). Data visualization was accomplished using

a combination of the ETE3 (Huerta-Cepas et al. 2016) and matplotlib (Hunter 2007)

Python libraries.

LITERATURE CITED

Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, Bolton E, Bourexis D, Brister JR, Bryant SH, Canese K, et al. 2018. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 46:D8–D13.

Ahrens J, Rahaman J, Siltberg-Liberles J. 2018. Large-Scale Analyses of Site-Specific Evolutionary Rates across Eukaryote Proteomes Reveal Confounding Interactions between Intrinsic Disorder, Secondary Structure, and Functional Domains. Genes (Basel). 9:553.

Ahrens J, Dos Santos HG, Siltberg-Liberles J. 2016. The Nuanced Interplay of Intrinsic Disorder and Other Structural Properties Driving Protein Evolution. Mol. Biol. Evol. 33:2248–2256.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

Bairoch A, Apweiler R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 28:45–48.

Bard JAM, Goodall EA, Greene ER, Jonsson E, Dong KC, Martin A. 2018. Structure and Function of the 26S Proteasome. Annu. Rev. Biochem. 87:697–724.

Bateman A, Martin MJ, O'Donovan C, Magrane M, Alpi E, Antunes R, Bely B, Bingley M, Bonilla C, Britto R, et al. 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 45:D158–D169.

Baumgartner WA, Cohen KB, Fox LM, Acquaah-Mensah G, Hunter L, Hunter L. 2007. Manual curation is not sufficient for annotation of genomic databases. Bioinformatics 23:i41-8.

Belote JM, Zhong L. 2009. Duplicated proteasome subunit genes in Drosophila and their roles in spermatogenesis. Heredity (Edinb). 103:23–31.

Buchholz PCF, Fademrecht S, Pleiss J. 2017. Percolation in protein sequence space.Mehrotra R, editor. PLoS One 12:e0189646.

Chen J, Coppola G. 2018. Bioinformatics and genomic databases. In: Handbook of clinical neurology. Vol. 147. p. 75–92.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422–1423.

Cook CE, Bergman MT, Finn RD, Cochrane G, Birney E, Apweiler R. 2016. The European Bioinformatics Institute in 2016: Data growth and integration. Nucleic Acids Res. 44:D20-6.

Dosztányi Z, Csizmók V, Tompa P, Simon I. 2005. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J. Mol. Biol. 347:827–839.

Eisen JA. 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res. 8:163–167.

Eisen JA, Wu M. 2002. Phylogenetic analysis and gene functional predictions: phylogenomics in action. Theor. Popul. Biol. 61:481–487.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, et al. 2014. Pfam: the protein families database. Nucleic Acids Res. 42:D222-30.

Fort P, Kajava A V, Delsuc F, Coux O. 2015. Evolution of proteasome regulators in eukaryotes. Genome Biol. Evol. 7:1363–1379.

Hauser M, Mayer CE, Söding J. 2013. kClust: fast and sensitive clustering of large protein sequence databases. BMC Bioinformatics 14:248.

Huerta-Cepas J, Bueno A, Dopazo J, Gabaldón T. 2008. PhylomeDB: a database for genome-wide collections of gene phylogenies. Nucleic Acids Res. 36:D491-6.

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol. Biol. Evol. 33:1635–1638.

Hunter JD. 2007. Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. 9:90–95.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 30:772–780.

Komander D, Rape M. 2012. The Ubiquitin Code. Annu. Rev. Biochem. 81:203–229.

Koonin E V. 2010. The origin and early evolution of eukaryotes in the light of phylogenomics. Genome Biol. 11:209.

de la Peña AH, Goodall EA, Gates SN, Lander GC, Martin A. 2018. Substrate-engaged 26 S proteasome structures reveal mechanisms for ATP-hydrolysis–driven translocation. Science 362:eaav0725.

Lander GC, Estrin E, Matyskiela ME, Bashore C, Nogales E, Martin A. 2012. Complete subunit architecture of the proteasome regulatory particle. Nature 482:186–191.

Lasker K, Forster F, Bohn S, Walzthoeni T, Villa E, Unverdorben P, Beck F, Aebersold R, Sali A, Baumeister W. 2012. Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. Proc. Natl. Acad. Sci. 109:1380–1387.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659.

Mayrose I, Graur D, Ben-Tal N, Pupko T. 2004. Comparison of site-specific rate-inference methods for protein sequences: Empirical Bayesian methods are superior. Mol. Biol. Evol. 21:1781–1791.

Miele V, Penel S, Duret L. 2011. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics 12:116.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48:443–453.

Poux S, Arighi CN, Magrane M, Bateman A, Wei C-H, Lu Z, Boutet E, Bye-A-Jee H, Famiglietti ML, Roechert B, et al. 2017. On expert curation and scalability: UniProtKB/Swiss-Prot as a case study.Kelso J, editor. Bioinformatics 33:3454–3460.

Povolotskaya IS, Kondrashov FA. 2010. Sequence space and the ongoing expansion of the protein universe. Nature 465:922–926.

Ronquist F, Huelsenbeck JP, Teslenko M. 2011. mb3.2_manual.pdf. MrBayes version 3.2 Man. Tutorials Model S.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice across a Large Model Space. Syst. Biol 61:539–542.

Rossum, Guido. 1995. Python reference manual.

Salisbury FB. 1969. Natural selection and the complexity of the gene. Nature 224:342–343.

Smith JM. 1970. Natural selection and the concept of a protein space. Nature 225:563–564.

Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. J. Mol. Biol. 147:195–197.

Telford MJ, Budd GE, Philippe H. 2015. Phylogenomic Insights into Animal Evolution. Curr. Biol. 25:R876–R887.

Voges D, Zwickl P, Baumeister W. 1999. The 26S proteasome: a molecular machine designed for controlled proteolysis. Annu. Rev. Biochem. 68:1015–1068.

Wollenberg K, Swaffield JC. 2001. Evolution of Proteasomal ATPases. Mol. Biol. Evol. 18:962–974.

Zhong L, Belote JM. 2007. The testis-specific proteasome subunit Pros 6T of D. melanogaster is required for individualization and nuclear maturation during spermatogenesis. Development 134:3517–3525.

Figure Captions

Figure 1. Examples of linkage graphs connecting objects (e.g., sequences) to form graph-based clusters. Reference-based clusters (A, B) are created by using a reference object (black node) to identify and group similar objects (grey nodes). Note that the members of a reference-based cluster depend on the choice of reference object. By contrast, a single-linkage cluster (C) is the connected graph that results from joining all pairs of similar objects together (effectively, every object is treated as a reference). This clustering strategy can be used to identify much larger groups than a reference-based clustering strategy using the same similarity threshold.

Figure 2. Scatterplot showing the relationship between analysis runtime (y axis) and the number of sequences identified in a single-linkage cluster (x-axis). Results are shown for analyses using three different alignment strategies: i) the initial BLAST alignment of a hit sequence to the query sequence (BLAST-SL), ii) the additional reverse BLAST alignment of the query sequence to the hit sequence (BLAST-BD) or iii) the optimal alignments produced by the local Smith-Waterman algorithm and the global Needleman-Wunsch algorithm (SW + NW). Note that both the x-axis and y-axis are log-scaled.

Figure 3. Jitterplots showing the number of sequences clustered in each of the 49 benchmark analyses using i) the legacy single-linkage program BLASTClust, ii) the initial BLAST alignment of a hit sequence to the query sequence (BLAST-SL), iii) the additional reverse BLAST alignment of the query sequence to the hit sequence (BLAST-BD), iv) the optimal alignments produced by the local Smith-Waterman algorithm and

the global Needleman-Wunsch algorithm (SW + NW) and v) the first linked sequences identified by the initial BLAST search of the clustering procedure. Note that the y-axis is log-scaled.

Figure 4. 50% majority-rule consensus tree (scale bar: bottom) showing inferred relationships between the 142 sequences identified in a benchmark single-linkage cluster containing the Rpt regulatory ring of the 26S proteasome complex. Labels indicate species names, UniProt codes, and gene annotations. Branch support values (posterior probability) are given for basal nodes of the tree. Note that all 6 main clades (containing subunits of the heterohexameric Rpt ring) are well-supported.

Figure 5. Structural and functional information obtained for sequences in the Rpt ring single-linkage cluster. Top: Site-specific sequence evolutionary rates are shown over a heatmap (below) displaying the IUPred intrinsic disorder propensity of each site in each sequence (higher values indicate higher disorder). Top-left: phylogenetic tree with color-coded terminal nodes indicating sequences with known functional annotations corresponding to Rpt1 (red), Rpt6 (orange), Rpt4 (yellow), Rpt2 (green), Rpt3 (blue) and Rpt5 (purple). Bottom-left: phylogeny indicating the 5 additional sequences (red) found in our optimal (SW + NW) single linkage cluster, which are not found in the original BLASTClust cluster. Bottom heatmap indicates predicted functional domains superimposed over amino acid sequence data (dark grey). Scale bar (bottom-left corner) indicates tree length.

Figure 6. Scatterplot showing sequence alignment quality of the 49 BLASTClust sequence clusters containing the 49 benchmark sequences used in our analysis. Y-axis indicates the minimum pairwise sequence identity between any two sequences in a given cluster. X-axis indicates the minimum alignment coverage (sequence length divided by the number of sites in the multiple sequence alignment) in each cluster. One of our benchmark sequences (red) was only identified as a singleton using the BLAST-SL method (BLAST-BD and SW + NW recovered additional members). Grey dots are clusters containing a benchmark sequence which was also identified as a singleton using BLAST-BD. Black dots are clusters containing a benchmark sequence which our clustering method identified as a singleton using all three alignment strategies (including SW + NW). Note that BLASTClust uses a more permissive calculation of pairwise coverage (including many alignment gaps), and the 4 single-linkage groups which our method failed to recover have relatively low alignment quality (low minimum alignment coverage).

Figure 7. Scatterplot showing sequence alignment quality of the non-singleton clusters produced using different pairwise alignment strategies. Y-axis indicates the minimum pairwise sequence identity between any two sequences in a given cluster. X-axis indicates the minimum alignment coverage (sequence length divided by the number of sites in the multiple sequence alignment) in each cluster. Note that the alignment qualities of clusters produced using the SW + NW strategy tend to be lower than the other two strategies.

Figures

Figure 1

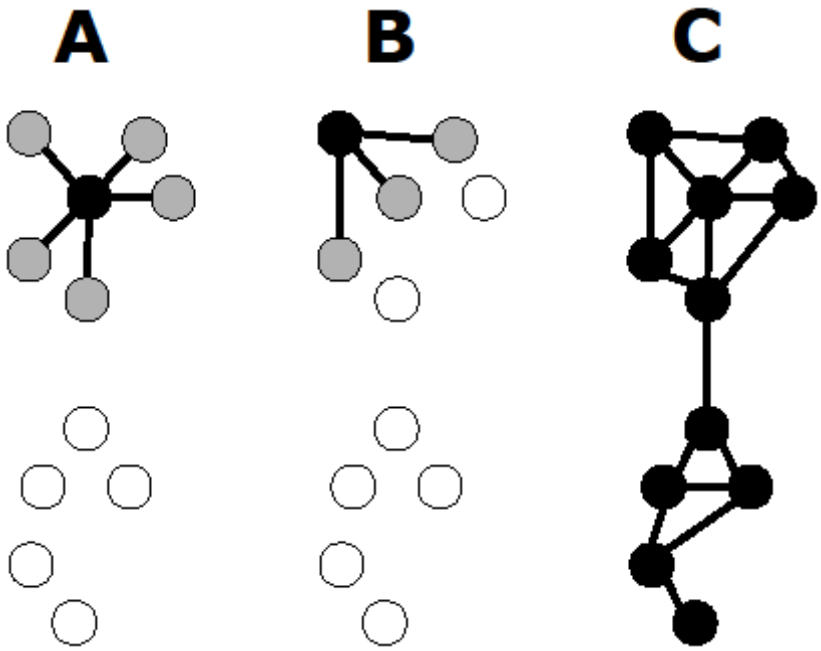


Figure 2

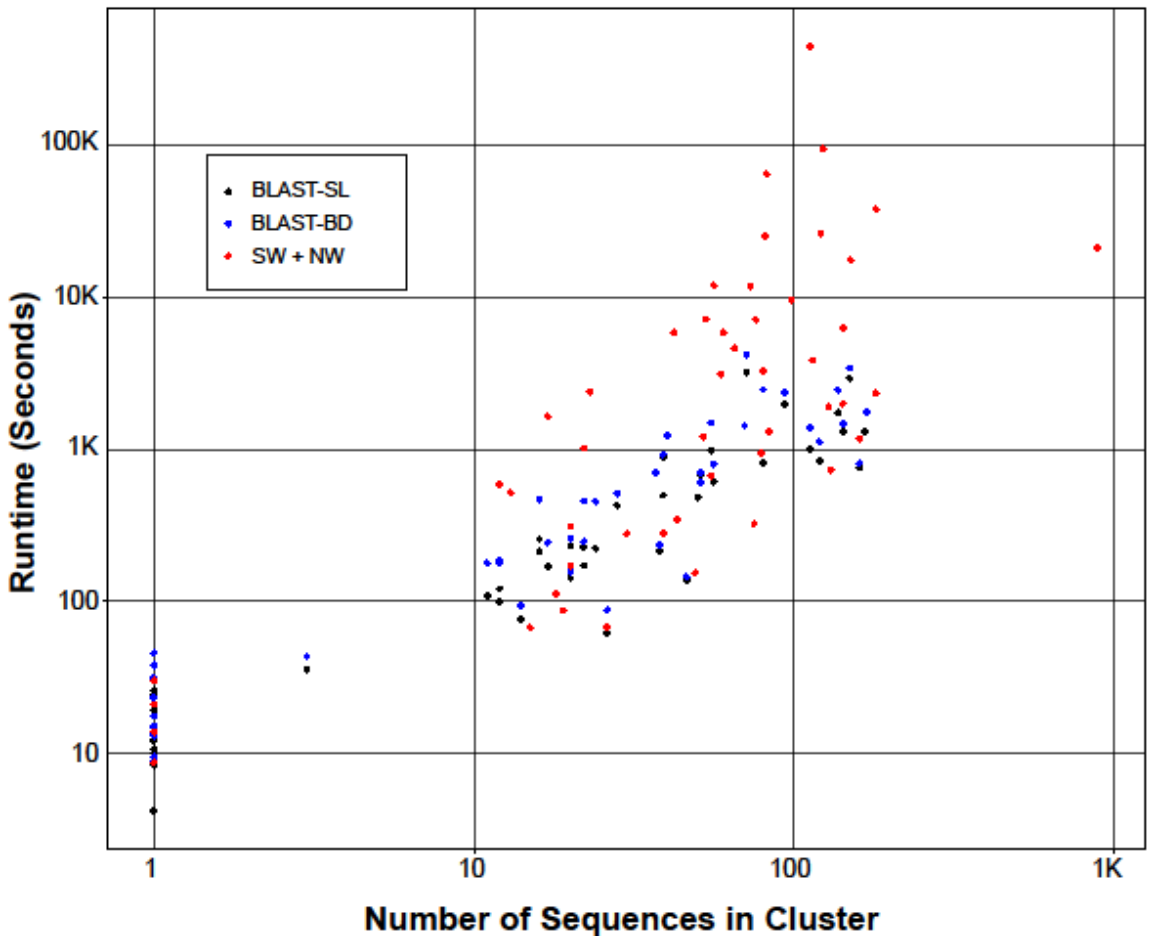

BLAST-SL
BLAST-BD
SW + NW
Runtime (Seconds)
100K
10K
1K
100
10
1
10
100
1K
Number of Sequences in Cluster

Figure 3

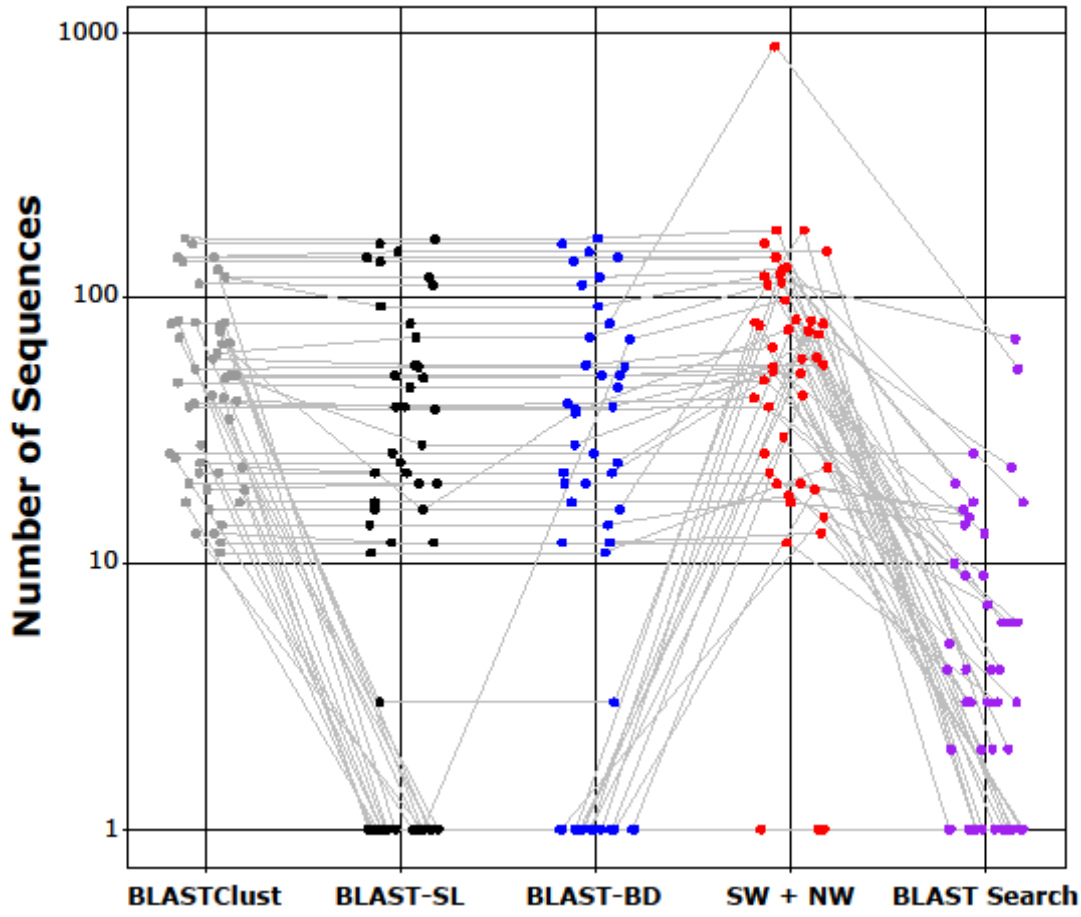

Number of Sequences
1000
100
10
1
BLASTClust
BLAST-SL
BLAST-BD
SW + NW
BLAST Search

Figure 4

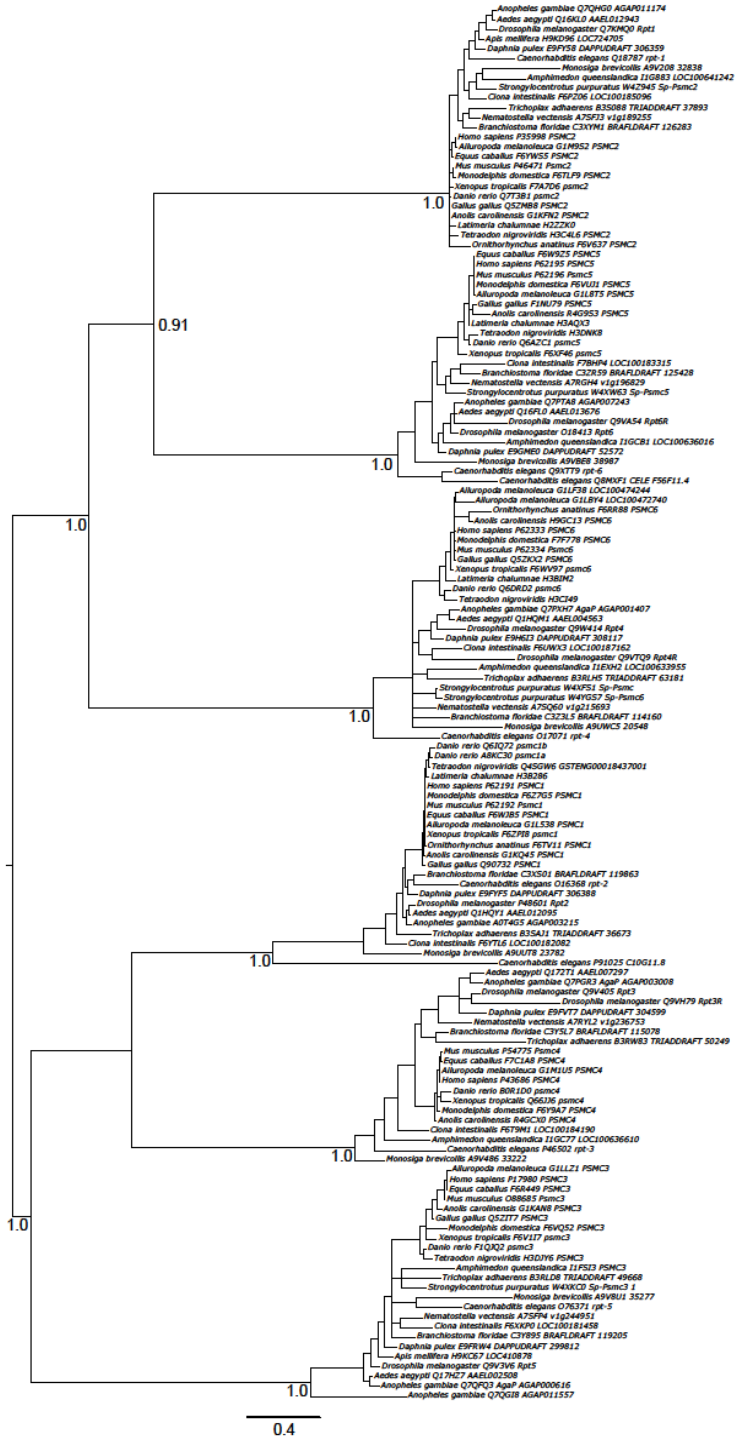

Anopheles gambiae Q7QHG0 AGAP011174
Aedes aegypti Q16KL0 AAEL012943
Drosophila melanogaster Q7KMQ0 Rpt1
Apis mellifera H9KD96 LOC724705
Daphnia pulex E9FY58 DAPPUDRAFT 306359
Caenorhabditis elegans Q18787 rpt-1
Monosiga brevicollis A9V208 32838
Amphimedon queenslandica I1G883 LOC100641242
Strongylocentrotus purpuratus W4Z945 Sp-Psmc2
Ciona intestinalis F6PZ06 LOC100185096
Trichoplax adhaerens B3S088 TRIADDRAFT 37893
Nematostella vectensis A7SFI3 v1g189255
Branchiostoma floridae C3XYM1 BRAFLDRAFT 126283
Homo sapiens P35998 PSMC2
Ailuropoda melanoleuca G1M9S2 PSMC2
Equus caballus F6YWS5 PSMC2
Mus musculus P46471 Psmc2
Monodelphis domestica F6TLP9 PSMC2
Xenopus tropicalis F7A7D6 psmc2
Danio rerio Q7T3B1 psmc2
Gallus gallus Q5ZMB8 PSMC2
Anolis carolinensis G1KFN2 PSMC2
Latimeria chalumnae H2ZZK0
Tetraodon nigroviridis H3CAL6 PSMC2
Ornithorhynchus anatinus F6V637 PSMC2
Equus caballus F6W9Z5 PSMC5
Homo sapiens P62195 PSMC5
Mus musculus P62196 Psmc5
Monodelphis domestica F6VU31 PSMC5
Ailuropoda melanoleuca G1LBT5 PSMC5
Gallus gallus F1NU79 PSMC5
Anolis carolinensis R4G9S3 PSMC5
Latimeria chalumnae H3AQX3
Tetraodon nigroviridis H3DNK8
Danio rerio Q6AZC1 psmc5
Xenopus tropicalis F6XF46 psmc5
Ciona intestinalis F7BHP4 LOC100183315
Branchiostoma floridae C3ZR59 BRAFLDRAFT 125428
Nematostella vectensis A7RGH4 v1g196829
Strongylocentrotus purpuratus W4XWE3 Sp-Psmc5
Anopheles gambiae Q7PTA8 AGAP007243
Aedes aegypti Q16FL0 AAEL013676
Drosophila melanogaster Q9VA54 Rpt6R
Drosophila melanogaster O18413 Rpt6
Amphimedon queenslandica I1GCB1 LOC100636016
Daphnia pulex E9GME0 DAPPUDRAFT 52572
Monosiga brevicollis A9V8E8 38987
Caenorhabditis elegans Q9XTT9 rpt-6
Caenorhabditis elegans Q8MXF1 CELE F56F11.4
Ailuropoda melanoleuca G1LT38 LOC100474244
Ailuropoda melanoleuca G1LBY4 LOC100472740
Ornithorhynchus anatinus F6RR88 PSMC6
Anolis carolinensis H9GC13 PSMC6
Homo sapiens P62333 PSMC6
Monodelphis domestica F7F778 PSMC6
Mus musculus P62334 Psmc6
Gallus gallus Q5ZJX2 PSMC6
Xenopus tropicalis F6WV97 psmc6
Latimeria chalumnae H3BIN2
Danio rerio Q6DRD2 psmc6
Tetraodon nigroviridis H3CI49
Anopheles gambiae Q7PXH7 AgaP AGAP001407
Aedes aegypti Q1HQM1 AAEL004563
Drosophila melanogaster Q9NV14 Rpt4
Daphnia pulex E9H6I3 DAPPUDRAFT 308117
Ciona intestinalis F6UWX3 LOC100187162
Drosophila melanogaster Q9VTQ9 Rpt4R
Amphimedon queenslandica I1EXH2 LOC100633955
Trichoplax adhaerens B3RLH5 TRIADDRAFT 63181
Strongylocentrotus purpuratus W4XFS1 Sp-Psmc
Strongylocentrotus purpuratus W4YGS7 Sp-Psmc6
Nematostella vectensis A7SQ60 v1g215693
Branchiostoma floridae C3Z3L5 BRAFLDRAFT 114160
Monosiga brevicollis A9UWC5 20548
Caenorhabditis elegans O17071 rpt-4
Danio rerio Q6IQ72 psmc1b
Danio rerio A8KC30 psmc1a
Tetraodon nigroviridis Q4SGW6 GSTENG00018437001
Latimeria chalumnae H3B286
Homo sapiens P62191 PSMC1
Monodelphis domestica F6Z7G5 PSMC1
Mus musculus P62192 Psmc1
Equus caballus F6WJB5 PSMC1
Ailuropoda melanoleuca G1L538 PSMC1
Xenopus tropicalis F6ZP88 psmc1
Ornithorhynchus anatinus F6TV11 PSMC1
Anolis carolinensis G1KQ45 PSMC1
Gallus gallus Q90732 PSMC1
Branchiostoma floridae C3X501 BRAFLDRAFT 119863
Caenorhabditis elegans O16368 rpt-2
Daphnia pulex E9FYF5 DAPPUDRAFT 306388
Drosophila melanogaster P48601 Rpt2
Aedes aegypti Q1HQY1 AAEL012095
Anopheles gambiae A8T4G5 AGAP003215
Trichoplax adhaerens B3SAJ1 TRIADDRAFT 36673
Ciona intestinalis F6YTL6 LOC100182082
Monosiga brevicollis A9UUT8 23782
Caenorhabditis elegans P91025 C10G11.8
Aedes aegypti Q172T1 AAEL007297
Anopheles gambiae Q7PGR3 AgaP AGAP003008
Drosophila melanogaster Q9V405 Rpt3
Drosophila melanogaster Q9VH79 Rpt3R
Daphnia pulex E9FVT7 DAPPUDRAFT 304599
Nematostella vectensis A7RYL2 v1g236753
Branchiostoma floridae C3YSL7 BRAFLDRAFT 115078
Trichoplax adhaerens B3RW83 TRIADDRAFT 50249
Mus musculus P54775 Psmc4
Equus caballus F7C1A8 PSMC4
Ailuropoda melanoleuca G1M1U5 PSMC4
Homo sapiens P43686 PSMC4
Danio rerio B0R100 psmc4
Xenopus tropicalis Q66JJ6 psmc4
Monodelphis domestica F6Y9A7 PSMC4
Anolis carolinensis R4GCX0 PSMC4
Ciona intestinalis F6TWH1 LOC100184190
Amphimedon queenslandica I1GC77 LOC100636610
Caenorhabditis elegans P46502 rpt-3
Monosiga brevicollis A9V486 33222
Ailuropoda melanoleuca G1LLZ1 PSMC3
Homo sapiens P17980 PSMC3
Equus caballus F6R449 PSMC3
Mus musculus O88685 Psmc3
Anolis carolinensis G1KAN8 PSMC3
Gallus gallus Q5ZIT7 PSMC3
Monodelphis domestica F6VQ52 PSMC3
Xenopus tropicalis F6V1I7 psmc3
Danio rerio F1QJQ2 psmc3
Tetraodon nigroviridis H3DJY6 PSMC3
Amphimedon queenslandica I1FSI3 PSMC3
Trichoplax adhaerens B3RLD8 TRIADDRAFT 49668
Strongylocentrotus purpuratus W4XKC9 Sp-Psmc3 1
Monosiga brevicollis A9V8U1 35277
Caenorhabditis elegans O76371 rpt-5
Nematostella vectensis A7SFP4 v1g244951
Ciona intestinalis F6XKP0 LOC100181458
Branchiostoma floridae C3Y895 BRAFLDRAFT 119205
Daphnia pulex E9FRW4 DAPPUDRAFT 299812
Apis mellifera H9KC57 LOC410878
Drosophila melanogaster Q9V3V6 Rpt5
Aedes aegypti Q17HZ7 AAEL002508
Anopheles gambiae Q7QFQ3 AgaP AGAP000616
Anopheles gambiae Q7QGI8 AGAP011557
1.0
0.91
1.0
1.0
1.0
1.0
1.0
1.0
0.4

Figure 5

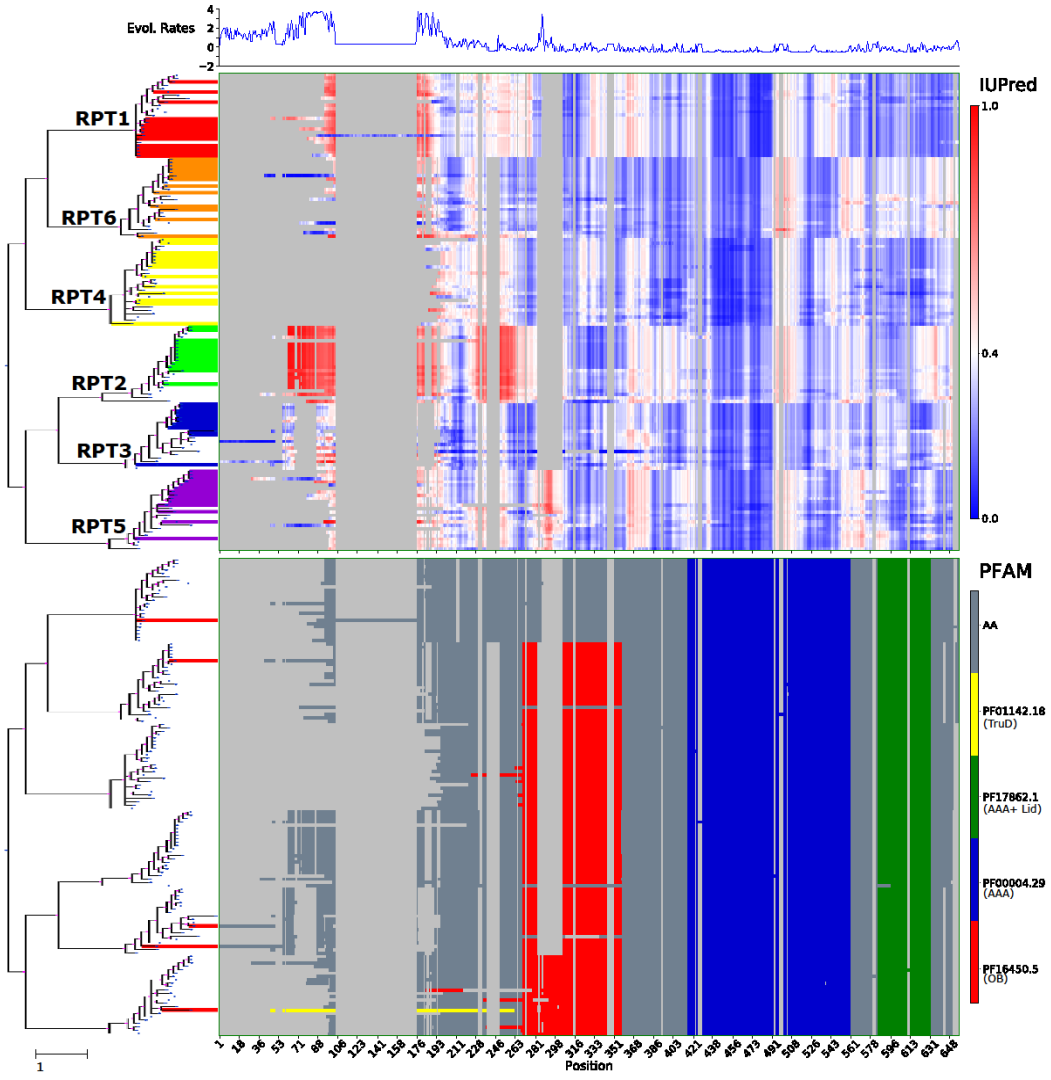

Evol. Rates
4
2
0
−2
RPT1
RPT6
RPT4
RPT2
RPT3
RPT5
IUPred
1.0
0.4
0.0
PFAM
AA
PF01142.18
(TruD)
PF17862.1
(AAA+ Lid)
PF00004.29
(AAA)
PF16450.5
(OB)
1
1 18 36 53 71 88 106 123 141 158 176 193 211 228 246 263 281 298 316 333 351 368 386 403 421 438 456 473 491 508 526 543 561 578 596 613 631 648
Position

Figure 6

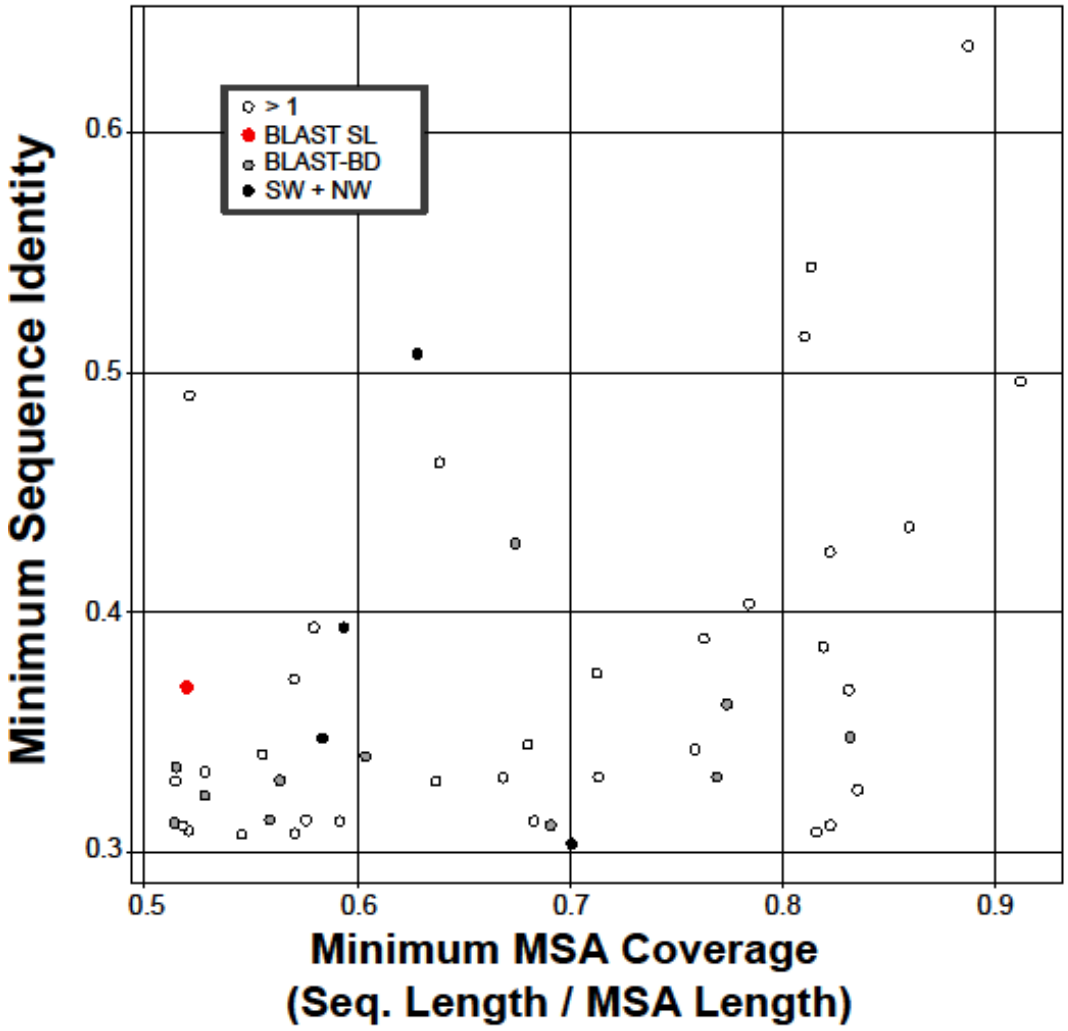
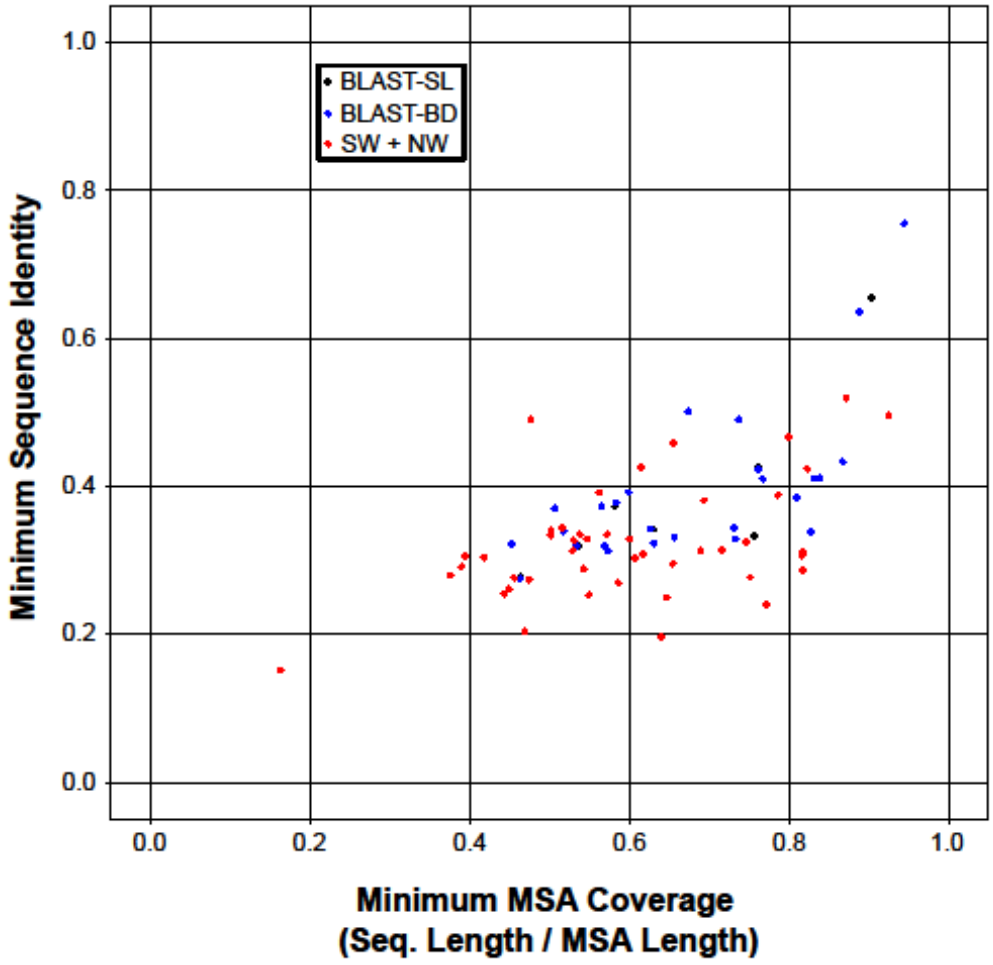

> 1
BLAST SL
BLAST-BD
SW + NW
Minimum Sequence Identity
0.6
0.5
0.4
0.3
0.5
0.6
0.7
0.8
0.9
Minimum MSA Coverage
(Seq. Length / MSA Length)

Figure 7


BLAST-SL
BLAST-BD
SW + NW
Minimum Sequence Identity
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Minimum MSA Coverage
(Seq. Length / MSA Length)

CHAPTER VI

CONCLUSIONS AND FUTURE DIRECTIONS

The Nuanced Interplay of Intrinsic Disorder and Other Structural Properties Driving Protein Evolution

The link between the structural properties of proteins and their site-specific rates of amino acid replacement will likely remain a topic of intense research, and other studies have already indicated fairly strong correlations between specific structural properties and replacement rates (Franzosa and Xia 2009; Yeh et al. 2014). Here, I describe the results of a large-scale analysis of metazoan protein evolution, examining the association between site-specific protein evolution and three factors: intrinsic disorder propensity, secondary structure and functional domain involvement. In designing this study, I explicitly tested for the possibility of non-additive statistical interactions, and indeed, the factorial model reveals significant interactions between all three of the factors I measured. Hence, the results of this study illustrate that it is important to consider all possible combinations of structural factors, and also to account for the possibility of non-additive effects that interacting structural/functional factors may have on rates of sequence evolution. In this case, doing so exposes a nuanced interplay between multiple drivers of sequence evolution.

A curious additional discovery in the course of this study was that a fraction of sequence alignment sites were consistently predicted to be both intrinsically disordered and involved in a conserved secondary structure (either an α-helix or a β-strand). Notably, these sites tend to have very low rates of amino acid replacement. At the time, I termed these puzzling predictions “disordered-structured” sites, and speculated that they may correspond to regions of proteins that alternate between secondary structure and intrinsic disorder.

Large-Scale Analyses of Site-Specific Evolutionary Rates across Eukaryote Proteomes Reveal Confounding Interactions between Intrinsic Disorder, Secondary Structure, and Functional Domains

While my initial factorial analysis of animal proteins revealed a surprising relationship between protein structure, function and sequence evolution, it was unclear whether those trends could be extrapolated to other eukaryotic lineages. In this follow-up study, I examine the same relationships between intrinsic disorder, secondary structure, functional domain involvement and sequence evolutionary rates in animals, plants, alveolate protists and saccharomycete fungi. Here, I report largely consistent trends in the relationships between structural/functional factors and rates of sequence evolution. In fact, in all four lineages, I find that i) ordered sites experience lower average amino acid replacement rates than disordered sites, ii) sites in secondary structures have lower average replacement rates than sites in random coils, and iii) sites in functional domains have lower average replacement rates than sites in inter-domain linker regions. Furthermore, the non-additive statistical interactions I initially reported in metazoans are observed in the other three eukaryotic lineages as well.

Notably, the alveolate dataset is somewhat of an exception to an overarching trend, in that the confounding interaction between disorder and secondary structure is less pronounced than in the other three lineages (though it is still statistically significant) and there also appears to be a larger difference in amino acid replacements between ordered and disordered sites in alveolate proteins. The alveolate dataset I studied here includes pathogenic organisms from the clade Apicomplexa, including several strains of the malaria-inducing parasite *Plasmodium falciparum*. Other researchers have noted that

apicomplexan proteomes contain a large number of long disordered regions (Mohan et al. 2008; Fong et al. 2009), and many of the potential vaccine targets in *Plasmodium* are known to be intrinsically disordered (Guy et al. 2015). Moreover, the erythrocyte binding-like proteins in *P. falciparum*, which are responsible for attaching to the surface of blood cells during host invasion, are intrinsically disordered even while binding to cell surface receptors (Blanc et al. 2014). Together, these findings suggest that intrinsic disorder may play a uniquely important functional role in many alveolate protists. This also suggests that future work will be required to better our understanding of interactions within disordered protein regions, in order to develop effective drugs/vaccines against malaria.

I initially speculated that the conserved disordered-structured alignment sites found in animal proteins may be associated with molecular recognition features or MoRFs: regions of protein sequences which must alternate between (unbound) intrinsic disorder and (bound) secondary structure in order to properly function (Mohan et al. 2006; Yan et al. 2016). Following up on this speculation, I analyzed the gene ontology (GO) terms (Ashburner et al. 2000) associated with protein sequences which contained conserved disordered-structured sites. A high fraction of these sequences have GO functional annotations corresponding to nucleic acid binding, and there is substantial evidence that nucleic acid binding proteins exhibit disorder-to-order transitions, upon binding to nucleotide targets, which are similar to the transitions observed in MoRFs (Dyson 2012; Varadi et al. 2015; Wang et al. 2016). A large fraction of hydrolase proteins also contain disordered-structured sites, and there is some evidence to suggest

that hydrolase proteins also rely on disorder-to-order transitions for proper function (Misaghi et al. 2005; Fong et al. 2009).

Ultimately, although many of the trends I describe here are i) statistically significant and ii) consistently observed across the four eukaryotic lineages I studied, the overall predictive power of the resulting statistical models is quite low. In other words, based on my results, it is somewhat appropriate to claim that intrinsically disordered protein sites have faster average rates of sequence evolution than ordered sites, but because of the high overlap in site rate distributions among ordered and disordered sites, there is not necessarily a large probability that a particular fast-evolving protein site is intrinsically disordered. Additionally, because of the strong confounding interaction observed in animals, plants and saccharomycetes, the expected rate of sequence evolution at a disordered site depends crucially on other structural factors (e.g., whether the site is involved in a secondary structure).

Evaluation of Site-specific Rate Heterogeneity Reveals Significant Differences in Sequence Divergence Patterns between Orthologous and Paralogous Proteins in Both Animals and Plants

The relationship between inferred sequence divergence (i.e., the normalized lengths of a phylogenetic tree corresponding to a multiple sequence alignment) and the α parameter of the inferred gamma rate distribution among alignment sites is something which, to my knowledge, has not been explicitly evaluated before. In this study, I show using simulated evolutionary scenarios that the inferred α value of a sequence alignment correlates positively with tree length in cases where heterotachy contributes to protein

sequence evolution. Furthermore, the slope of this correlation increases with higher levels of heterotachy. Interestingly, in alignments of real protein data, there is a significant difference in the relationship between tree length and α when considering i) clusters of orthologous sequences (related by speciation) and ii) clusters of paralogous sequences (related by gene duplication). Specifically, the difference in divergence patterns implies that paralogous protein evolution entails significantly more heterotachy than orthologous protein evolution.

The ortholog conjecture (Koonin 2005), the hypothesis that orthologous genes tend to be more functionally similar than paralogous genes, has been tested numerous times in recent years (see Nehrt et al. 2011; Altenhoff et al. 2012; Chen and Zhang 2012; Rogozin et al. 2014). Here, I have shown that orthologous proteins do exhibit a different sequence divergence pattern than paralogous genes, a difference which is consistent with my finding that they experience significantly less heterotachy overall. There is also a theoretical basis for the notion that variation in site-specific rates of evolution (i.e., heterotachy) is associated with changes in protein function (Gu 1999; Gaucher et al. 2002; Abhiman et al. 2006). Therefore, given that i) heterotachy is associated with functional change and ii) heterotachy is apparently significantly more prevalent in paralogous phylogenies than in orthologous phylogenies, it stands to reason that iii) on average, paralogous proteins are significantly more functionally divergent than orthologous proteins, when overall sequence divergence is considered in an evolutionary (i.e., phylogenetic) context. In this light, the results I present here are compelling evidence in favor of the ortholog conjecture, as they corroborate the hypothesis that genes

related by duplication tend to be more functionally divergent than genes related by speciation.

Acquisition of Homologous Protein Sequence Clusters from Local Databases Using a Simple, Graph-Based Single-Linkage Clustering Procedure

The intended use-case of most sequence clustering methods is the partitioning of entire sequence databases. As such, their developers tend to emphasize their performance at the whole-database scale (see Li and Godzik 2006; Miele et al. 2011; Hauser et al. 2013). The graph-based single-linkage clustering procedure I describe here is instead intended to produce only one sequence cluster at time, and is thus not designed to operate efficiently as a whole-database clustering procedure. However, for researchers interested in only a single gene cluster (or a relatively small number of gene clusters) the small-scale nature of my procedure affords a degree of flexibility which is not available in whole-database clustering programs. For instance, rather than being limited to producing clusters of sequences within a target database, my procedure allows the use of external query sequences (i.e., sequences not found in the target database). Additionally, because the time required to define just one single-linkage cluster (minutes to hours) is typically much smaller than that required to cluster an entire database in BLASTClust (Altschul et al. 1990) (hours to days), users may define clusters using several different query sequences, linkage cutoffs and alignment strategies in a relatively short amount of time, and without expending unnecessary resources.

A curious feature of single-linkage clustering is that, if linkage cutoffs are set such that any two sequences which are directly linked can be considered homologs (e.g.,

40% protein sequence identity, 90% coverage), then all members of a single linkage cluster are effectively "transitive" homologs (i.e., ***A*** is homologous to ***B*** and ***B*** is homologous to ***C***, so ***A*** is homologous to ***C***). An important caveat of this feature is that the alignment coverage must be stringent enough to avoid clustering sequences with, for example, incongruent domain architectures (see Miele et al. 2011). Nonetheless, I show here that using a moderately stringent set of linkage cutoffs (40% protein sequence identity, 90% alignment footprint coverage, BLAST E-value filter of 10^{-6} and a maximum of 500 hits per BLAST search), it is possible to recover a single-linkage cluster representing a mostly-complete protein family—in this case, the proteins comprising the heterohexameric Rpt ring of the 26S proteasome complex. The Rpt protein family is known to be ancient, and the six primary gene duplications (giving rise to the heterohexamer) are believed to have occurred prior to the divergence of plants and animals, and possibly even prior to the last eukaryote common ancestor or LECA (Fort et al. 2015). Multiple sequence alignment indicates that protein sequence, structure and function is conserved among orthologous sequences in this group, but there is apparent divergence in structure (i.e., differing regions of intrinsic disorder) and function (differences in predicted functional domains) among paralogous proteins. These findings are consistent with recent cryogenic electron microscopy (CryoEM) work, indicating that subunits of the Rpt heterohexamer play unique roles in stabilizing the active proteasome complex and guiding targeted proteins to the core particle for degradation (de la Peña et al. 2018).

Future Directions

Sequence-based predictors of protein structure are still being developed, and newer prediction methods achieve higher accuracy by jointly estimating several structural factors at once (see Yang et al. 2017). Future work, which considers the joint effects of a larger number of (more accurately-predicted) structural/functional protein properties, may uncover a clearer association between protein sequence, structure and function. Further, as statistical methods become more efficient, hierarchical linear models (e.g., considering the disorder propensity of a site in addition to the overall disorder content of the sequence) will likely reveal a more complete picture of how protein structure drives sequence evolution.

Given the observed correlation between sequence divergence (i.e., tree length) and rate heterogeneity (i.e., the α parameter of the gamma distribution), it is possible that there also exists a positive correlation between site-specific heterotachy (i.e., the amount of rate variation at a particular alignment site across different lineages) and the uncertainty (variance) in the corresponding site rate estimate. Rather than focusing only on global measurements of rate heterogeneity (i.e., α values), future Bayesian analyses should estimate the joint posterior distributions of individual site rates as well. Such distributions may be helpful for identifying alignment sites which are evolving under markedly different evolutionary constraints in different lineages.

While my goal in Chapter V was primarily to outline a targeted clustering procedure, and present a straightforward implementation of said procedure, future optimizations to this implementation (e.g., concurrency/parallelization) have the potential to greatly accelerate its performance. At the moment, the implementation I have

described performs reasonably efficiently and is well-suited to the task of mining local protein sequence databases for protein families. This makes the implementation suitable for a wide range of research projects, from phylogenetic inference to homology detection to functional annotation.

LITERATURE CITED

Abhiman S, Daub CO, Sonnhammer ELL. 2006. Prediction of Function Divergence in Protein Families Using the Substitution Rate Variation Parameter Alpha. Mol. Biol. Evol. 23:1406–1413.

Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C. 2012. Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs.Eisen JA, editor. PLoS Comput. Biol. 8:e1002514.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene Ontology: tool for the unification of biology. Nat. Genet. 25:25–29.

Blanc M, Coetzer TL, Blackledge M, Haertlein M, Mitchell EP, Forsyth VT, Jensen MR. 2014. Intrinsic disorder within the erythrocyte binding-like proteins from *Plasmodium falciparum*. Biochim. Biophys. Acta - Proteins Proteomics 1844:2306–2314.

Chen X, Zhang J. 2012. The Ortholog Conjecture Is Untestable by the Current Gene Ontology but Is Supported by RNA Sequencing Data.Ouzounis CA, editor. PLoS Comput. Biol. 8:e1002784.

Dyson HJ. 2012. Roles of intrinsic disorder in protein-nucleic acid interactions. Mol. Biosyst. 8:97–104.

Fong JH, Shoemaker BA, Garbuzynskiy SO, Lobanov MY, Galzitskaya O V, Panchenko AR. 2009. Intrinsic disorder in protein interactions: insights from a comprehensive structural analysis. PLoS Comput. Biol. 5:e1000316.

Fort P, Kajava A V, Delsuc F, Coux O. 2015. Evolution of proteasome regulators in eukaryotes. Genome Biol. Evol. 7:1363–1379.

Franzosa EA, Xia Y. 2009. Structural determinants of protein evolution are context-sensitive at the residue level. Mol. Biol. Evol. 26:2387–2395.

Gaucher EA, Gu X, Miyamoto MM, Benner SA. 2002. Predicting functional divergence in protein evolution by site-specific rate shifts. Trends Biochem. Sci. 27:315–321.

Gu X. 1999. Statistical Methods for Testing Functional Divergence after Gene Duplication. Mol. Biol. Evol 16:1664–1674.

Guy AJ, Irani V, MacRaild CA, Anders RF, Norton RS, Beeson JG, Richards JS, Ramsland PA. 2015. Insights into the Immunological Properties of Intrinsically Disordered Malaria Proteins Using Proteome Scale Predictions.Mantis NJ, editor. PLoS One 10:e0141729.

Hauser M, Mayer CE, Söding J. 2013. kClust: fast and sensitive clustering of large protein sequence databases. BMC Bioinformatics 14:248.

Koonin E V. 2005. Orthologs, paralogs, and evolutionary genomics. Annu. Rev. Genet. 39:309–338.

de la Peña AH, Goodall EA, Gates SN, Lander GC, Martin A. 2018. Substrate-engaged 26 S proteasome structures reveal mechanisms for ATP-hydrolysis–driven translocation. Science 362:eaav0725.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659.

Miele V, Penel S, Duret L. 2011. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics 12:116.

Misaghi S, Galardy PJ, Meester WJN, Ovaa H, Ploegh HL, Gaudet R. 2005. Structure of the Ubiquitin Hydrolase UCH-L3 Complexed with a Suicide Substrate. J. Biol. Chem. 280:1512–1520.

Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, Dunker AK, Uversky VN. 2006. Analysis of Molecular Recognition Features (MoRFs). J. Mol. Biol. 362:1043–1059.

Mohan A, Sullivan Jr WJ, Radivojac P, Dunker AK, Uversky VN. 2008. Intrinsic disorder in pathogenic and non-pathogenic microbes: discovering and analyzing the unfoldomes of early-branching eukaryotes. Mol. Biosyst. 4:328.

Nehrt NL, Clark WT, Radivojac P, Hahn MW. 2011. Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals.Rzhetsky A, editor. PLoS Comput. Biol. 7:e1002073.

Rogozin IB, Managadze D, Shabalina SA, Koonin E V. 2014. Gene family level comparative analysis of gene expression in mammals validates the ortholog conjecture. Genome Biol. Evol. 6:754–762.

Varadi M, Zsolyomi F, Guharoy M, Tompa P. 2015. Functional Advantages of Conserved Intrinsic Disorder in RNA-Binding Proteins. PLoS One 10:e0139731.

Wang C, Uversky VN, Kurgan L. 2016. Disordered nucleiome: Abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea. Proteomics 16:1486–1498.

Yan J, Dunker AK, Uversky VN, Kurgan L, Fischer E, Lemieux UR, Spohr U, Dunker AK, Garner E, Guilliot S, et al. 2016. Molecular recognition features (MoRFs) in three domains of life. Mol. BioSyst. 12:697–710.

Yang Y, Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, Wang J, Sattar A, Zhou Y. 2017. SPIDER2: A Package to Predict Secondary Structure, Accessible Surface Area, and Main-Chain Torsional Angles by Deep Neural Networks. In: Methods in molecular biology (Clifton, N.J.). Vol. 1484. p. 55–63.

Yeh S-W, Huang T-T, Liu J-W, Yu S-H, Shih C-H, Hwang J-K, Echave J. 2014. Local packing density is the main structural determinant of the rate of protein sequence evolution at site level. Biomed Res. Int. 2014:572409.

VITA

JOSEPH B AHRENS

2006-2010	B.S., Biology Texas A&M University College Station, Texas
2010-2011	MARB IDP Graduate Assistantship Texas A&M University at Galveston Galveston, Texas
2010-2013	M.S., Marine Biology Texas A&M University at Galveston Galveston, Texas
2017-2019	Doctoral Candidate Florida International University Miami, Florida
2018	Doctoral Evidence Acquisition Fellow Florida International University Miami, Florida
2018-2019	Dissertation Year Fellow Florida International University Miami, Florida

PUBLICATIONS

Ahrens J, Rahaman J, Siltberg-Liberles J. (2018). Large-Scale Analyses of Site-Specific Evolutionary Rates across Eukaryote Proteomes Reveal Confounding Interactions between Intrinsic Disorder, Secondary Structure, and Functional Domains. Genes, 9(11), 553. doi:10.3390/genes9110553

Ahrens, J. B., Nunez-Castilla J. & Siltberg-Liberles, J. (2017). Evolution of intrinsic disorder in eukaryotic proteins. Cellular and Molecular Life Sciences, 74(17), 3163-3174. doi:10.1007/s00018-017-2559-0

Ahrens, J., dos Santos, H. G. & Siltberg-Liberles, J. (2016). The Nuanced Interplay of Intrinsic Disorder and other Structural Properties Driving Protein Evolution. Molecular Biology and Evolution, 33(9), 2248-56. doi:10.1093/molbev/msw092

Ahrens, J. B., Kudenov, J. D., Marshall, C. D., & Schulze, A. (2014). Regeneration of posterior segments and terminal structures in the bearded fireworm, Hermodice carunculata (Annelida: Amphinomidae). Journal of Morphology, 275(10), 1103–12. doi:10.1002/jmor.20287

Ahrens, J. B., Borda, E., Barroso, R., Paiva, P. C., Campbell, A. M., Wolf, A., Nugues, M.M., Rouse, G, Schulze, A. (2013). The curious case of Hermodice carunculata (Annelida: Amphinomidae): evidence for genetic homogeneity throughout the Atlantic Ocean and adjacent basins. Molecular Ecology, 22(8), 2280–91. doi:10.1111/mec.12263