Bioinformation
www.bioinformation.net

open access
Hypothesis

Length constraints of multi-domain proteins in metazoans

Sarah Middleton, Timothy Song, Sudhir Nayak*

Department of Biology, The College of New Jersey, 2000 Pennington Rd., Ewing, NJ 08628; Sudhir Nayak-E-mail:nayak@tcnj.edu Phone: 609-771-2659 ; *Corresponding Author

Received March 06, 2010; accepted April 09, 2010; published April 30, 2010

Abstract:
The increasing number of annotated genome sequences in public databases has made it possible to study the length distributions and domain composition of proteins at unprecedented resolution. To identify factors that influence protein length in metazoans, we performed an analysis of all domain-annotated proteins from a total of 49 animal species from Ensembl (v.56) or EnsemblMetazoa (v.3). Our results indicate that protein length constraints are not fixed as a linear function of domain count and can vary based on domain content. The presence of repeating domains was associated with relaxation of the constraints that govern protein length. Conversely, for proteins with unique domains, length constraints were generally maintained with increased domain counts. It is clear that mean (and median) protein length and domain composition vary significantly between metazoans and other kingdoms; however, the connections between function, domain content, and length are unclear. We incorporated Gene Ontology (GO) annotation to identify biological processes, cellular components, or molecular functions that favor the incorporation of multi-domain proteins. Using this approach, we identified multiple GO terms that favor the incorporation of multi-domain proteins; interestingly, several of the GO terms with elevated domain counts were not restricted to a single gene family. The findings presented here represent an important step in resolving the complex relationship between protein length, function, and domain content. The comparison of the data presented in this work to data from other kingdoms is likely to reveal additional differences in the regulation of protein length.

Background:
The proteome of an organism is broadly defined as the sum total of all proteins expressed by its genome **[1]**. The sequencing and annotation of large numbers of genomes have generated a considerable amount of raw material for the investigation of proteins and their component domains. Given the breadth of data available, a powerful tool to identify the underlying principles that govern the length or domain content of proteins is the comparative analysis across species. This type of analysis has been used to identify the functional and evolutionary constraints that control the domain content and length of proteins. Perhaps the most interesting revelation from the comparative analysis of proteins is the combinatorial nature of all proteomes. The majority of proteins across all kingdoms are built from a limited inventory of domains that are combined in different ways **[2, 3]**. The duplication, deletion, fission, and fusion of single domains and groups of domains have been described in all kingdoms and vary by gene family **[2, 4, 5]**. The exact mechanism behind the duplications and exchange remain elusive; however, it has been hypothesized that functional constraints may favor the duplication of certain domains or groups of domains **[4, 6]**. While fusion of domains and whole proteins are by far the most commonly occurring evolutionary events, many of the metazoan-specific protein families have highly variable numbers of repeated domains that are related to their function **[2]**. For proteins that contain repeating domains, indels (insertion and/or deletion events) are the most common followed by the repetition of multiple domains **[4, 5]**. Since the addition, deletion, or fusion of two domains can have profound impact on the function of a protein, the evolutionary history of any given species is reflected in the unique combination of domains that make up its proteome.

Another important area of investigation has been the identification of factors that influence protein length. Significant protein length differences, even among clusters of orthologous groups (COGs), have between well described for eukaryotes, bacteria, and archaea **[7, 8]**. The lengths of proteins, even within a given species, can vary by several orders of magnitude from less than 100 amino acids to greater than 30,000 amino acids. In metazoans, many of the larger multi-domain proteins are formed as the result of the combinatorial rearrangement of domains present in all kingdoms **[2, 7, 8]**. Regardless of the etiology of the variation, the fusion and expansion of domains contributes significantly to the length of proteins in all species. It is reasonable to assume that larger proteins involved in complex biological processes are likely to have more domains; however, the predictors of protein length at the domain level are not well understood. For example, given proteins with the same number of domains, how does the domain content (repetitive – vs- non-repetitive) of the protein influence protein length? Similarly, what are the biological processes that select for proteins with increased numbers of domains? To address these outstanding questions, we used data from 49 metazoan species in conjunction with the structured Gene Ontology (GO) vocabulary to determine the relationship between protein length, domain content, and function.

Methodology:
Data and Tools:
All raw data from this work and algorithms used to filter or perform calculations are available on our web site (http://www.tcnj.edu/~nayaklab/domains.htm). All algorithms were developed using Java JDK 6 (http://java.sun.com/javase/downloads/index.jsp) and are provided under the terms of the GNU General Public License as published by the Free Software Foundation (version 3) (www.gnu.org/copyleft/gpl.html).

Length distribution and domain distribution of proteins:
To determine the length and domain distribution of proteins in metazoans and validate our approach, we obtained information for 843,871 domain-annotated proteins from 49 animal species in Ensembl (v.56) or EnsemblMetazoa (v.3) **[9]**. The length distribution for proteins over intervals of 100 amino acid was obtained by calculating the number of proteins in each interval relative to the total number of proteins for each species **(Figure 1A)**. The distribution of metazoan proteins grouped by phylum and class showed remarkable similarity over all intervals tested. These data are similar to previously published results using other methods and suggest that our data set provides a precise representation of the metazoan protein length distribution **[7, 8, 10]**. Proteins were scored for the presence of Interpro (IPR) annotated domains in order to determine the distribution of multi-domain proteins in metazoans **(Figure 1B) [11]**. When proteins with multiple isoforms were encountered, the variant with the most domains was selected in all cases. Much like the length distributions, the domain distributions in all metazoans tested were remarkably similar **(Figure 1B).**

Domain content influences protein length:
Large proteins involved in multiple cellular processes are more likely to contain multiple domains. To identify the relationship between length and domain content, we compared protein length to domain count across all proteins **(Figure 1C).** Interestingly, the relationship between these two properties is not a simple linear polynomial; rather, the best-fit line is a weak power function ($y = 393.23e^{0.0672x}$), suggesting that length constraints become increasingly relaxed as proteins begin to acquire more domains. To further investigate this relationship, we assessed length and domain information based on the presence or absence of repeat domains **(Figure 1D).** For this analysis, a repeat domain was defined as any domain that occurred two or more times on a protein. Repeat-containing proteins (RCP) and non-repeat-containing proteins (NRCP) were analyzed separately to identify any differences in

protein length based on repetitive character. When proteins with less than five domains were considered, NRCPs were found to be significantly longer than RCPs. In contrast, when proteins with more than eight domains were considered, NRCPs were found to be significantly shorter than RCPs.

Identification of functions that favor multi-domain proteins:
We took advantage of the structured Gene Ontology (GO) **[12]** vocabulary to identify biological processes, cellular components, or molecular functions that favor multi-domain proteins. GO IDs associated with each protein were obtained from Ensembl (v.56) and EnsemblMetazoa (v.3) BioMarts. A domain distribution was generated for each GO ID and graphed as a fraction of the subset of the proteins annotated with that ID **(Figure 2).** This was done iteratively for all GO IDs with more than five proteins. Examples of GO IDs found to have a substantially higher than average domain-per-protein counts are shown in **Figure 2 (A, B, C).**

Discussion:

Length constraints are relaxed as domain count increases:
With the availability of multiple animal proteomes in public databases it is clear that protein length can vary over several orders of magnitude within a given proteome, yet the distribution of protein lengths across metazons are very similar. Interestingly, our results indicate that protein length constraints are not fixed as a linear function of domain count; rather, overall protein length constrains are relaxed with increasing domain count. When proteins become larger than approximately 500 amino acids (or 5 domains) each additional domain will, on average, increase the length of the protein by more than the previous domain. Several mechanisms (recombination, exon shuffling) for the expansion of proteins size have been previously suggested **[6]**, and it is possible that some of these same mechanisms could be involved in the relaxation of length constraints. For example, the relaxed constraints could reflect the increased contribution of recombination when the underlying genomic sequence of a protein reaches some critical size **[13]**. Another possibility is that relaxed constraints may reflect a selection for a specific number of domains in certain protein families **[6]**. Regardless, at some point, the length of a protein is likely to be limited by the evolutionary constraints that act to minimize sequences with no required functional roles because energetic of the costs associated with processing multi-domain proteins **[14, 15]**.

The presence of repeat domains influences protein length:
To determine the relationship between domain content and protein length, we separated proteins based on the presence or absence of repeated domains. Curiously, we found that repeat-containing proteins (RCP) were significantly shorter than non-repeat-containing proteins (NRCP) when less than five domains were present; however, RCP were significantly longer when eight or more domains were present. It is unlikely that this difference in growth rate is result of length differences between the two types of domains themselves, since the average lengths (NRCP = 133.7+/-98.9, RCP = 127.1+/-93.0) and medians (NRCP = 113, RCP = 103) of repeat domains and unique domains were not significantly different in the species studied. Interestingly, while the average size of RCP increased steadily with domain count, the average size of NRCP remained between 400 and 600 residues regardless of the number of domains. Possible explanations could be that genes that contain repeat domains tend to have extra sequences that increase their length-to-domain ratio **[6]** or that proteins that contain domains that repeat may be easier to fold, thus mitigating the costs associated with processing and allowing them to grow larger.

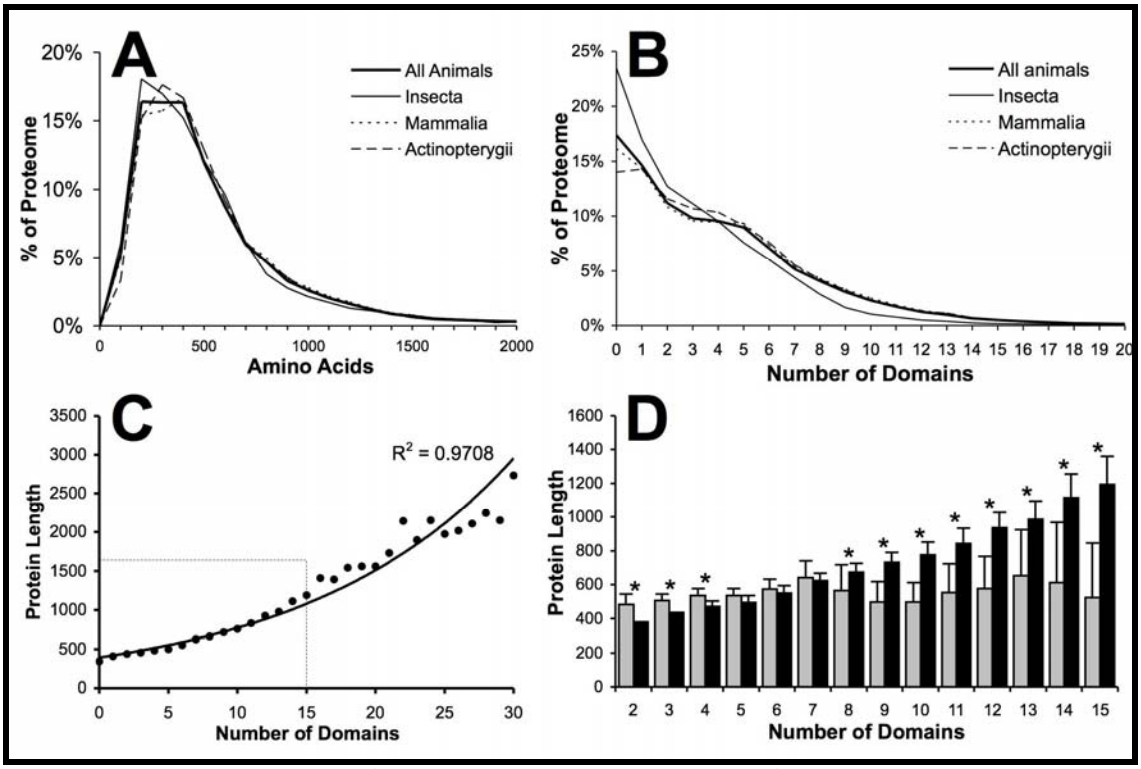

Figure 1: Length and domain distributions for metazoan proteins. **A)** Length distribution of metazoan proteins. Proteins >2000 amino acids were excluded (1.7%) for illustration. **B)** Domain distribution of metazoan proteins. Proteins with >20 domains were excluded (0.5%) for illustration. **C)** Number of domains versus protein length. Proteins with >30 domains were excluded (<0.1%) for illustration. Equation of $y = 393.23e^{0.0672x}$. **D)** Length constraints vary based on domain content. Proteins with >15 domains were excluded (1.7%) for illustration. Grey bars = non-repeat-containing proteins. Black bars = repeat-containing proteins. "*" = p<0.0001 Mann-Whitney.

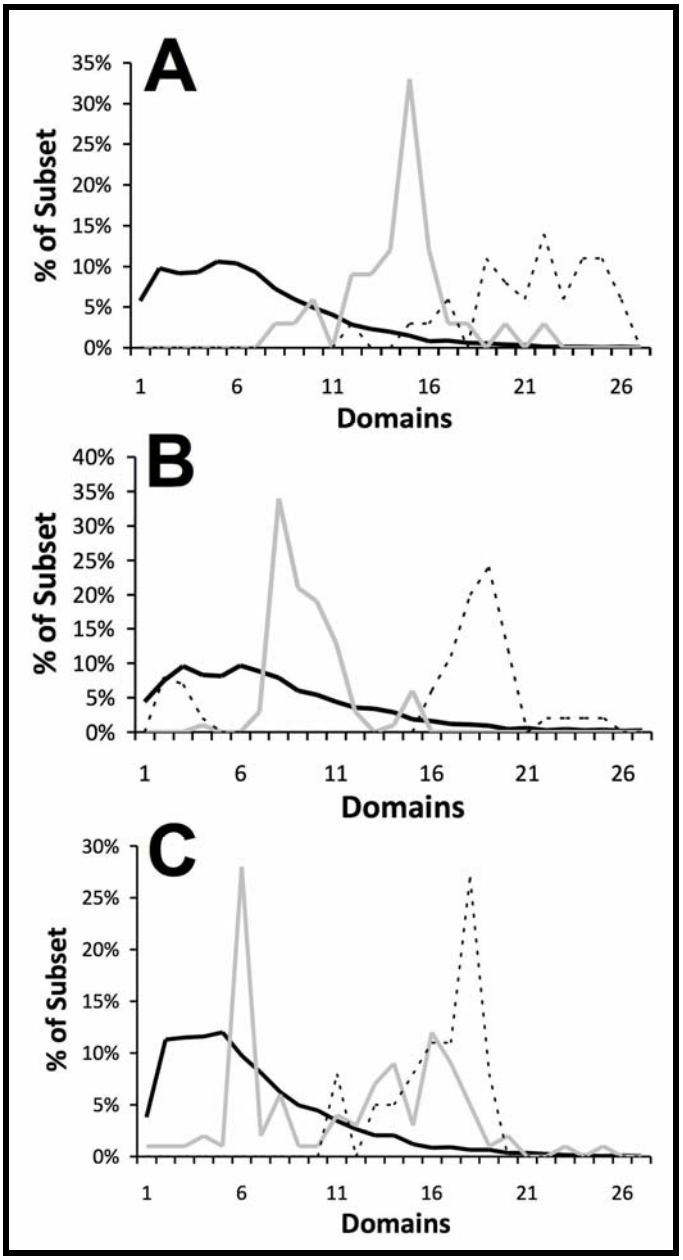

Figure 2: GO ID domain distribution. **A)** GO ID domain distribution within biological process. Black line = Domain distribution of all GO IDs in biological process, gray line = Domain distribution in GO:0001834, and dashed line = Domain distribution in GO:0002316. **B)** GO ID domain distribution within cellular component. Black line = Domain distribution of all GO IDs in cellular component, gray line = Domain distribution in GO:0000235, and dashed line = Domain distribution in GO:0001527. **C)** GO ID domain distribution within molecular function. Black line = Domain distribution of all GO IDs in molecular function, gray line = Domain distribution in GO:0000155, and dashed line = Domain distribution in GO:0004087. A total of 10636 GO IDs were used, with 6429 in biological processes, 964 in cellular component, and 3243 in molecular function.

Proteins related by biological function have similar domain counts:

It is clear that domain combinations vary dramatically across kingdoms and may have played a role in the evolution of metazoan-specific functions **[2]**. In general, it is thought that evolution selects for the proteins with the minimal number of amino acids needed to perform their function because they are more efficient to produce **[16]**. From this we would expect that proteins with fewer domains would be biosynthetically favored, and indeed, our analysis of the metazoan domain distribution corroborates that proteins are more likely to have fewer domains than many **(Figure 1B).** However, as demonstrated in **Figure 2,** certain proteins involved in a specific biological process (A), cellular component (B), or molecular function (C) often strongly favor the presence of many domains because it is beneficial or essential to their function. As expected, many of the GO IDs associated with elevated domain-per-protein counts were composed of closely related proteins. For example, GO:0002316 **(Figure 2A)** is associated only with the Plcg2 gene family. This is not always the case, however, as other GO IDs were associated with several unrelated families, all favoring multi-domain proteins. GO:0000235 **(Figure 2B)** contains 27 gene products from a variety of families and species, all of which are involved in astral microtubules and favor the incorporation of approximately 8 domains.

Limitations:
The analysis presented in this work has several limitations. For example, we restricted our dataset to domain-annotated metazoan proteins available from Ensembl (v.56) and EnsemblMetazoa (v.3). While this data set is extensive, it could be further refined by the addition of information from other databases or the inclusion of more species. In addition, an important future direction could be the inclusion of positional information for each domain to assess the contribution of N- and C- terminal changes and internal duplications **[4]**.

Conclusion:
The need for large, multi-domain proteins in complex cellular processes is likely to be balanced by the evolutionary constraints that govern protein length. In species that have evolved complex biological processes, proteins tend to be longer than in simpler organisms; this can clearly be seen when comparing eukaryotic and prokaryotic proteins **[8]**. In order to elucidate the determinants of protein length in metazoans, we analyzed proteome data from 49 species to identify common patterns. Our results indicate that the presence of repeating domains is associated with a relaxation of the constraints that govern protein length; however, for proteins with unique domains, length constraints were generally maintained. In addition, we identified biological processes, molecular functions, and cellular components that favor proteins with increased numbers of domains. This study lays important groundwork for future research into the global constraints on protein evolution and the relationship between length, function, and domain content.

References:
[1] M Wilkins, *Expert Rev Proteomics*. (2009) **6**: 599 [PMID: 19929606]
[2] G Apic J. Gough, *et al. J Mol Biol*. (2001) **310**: 311 [PMID: 11428892]
[3] P Djian, *Cell*. (1998) **94**: 155 [PMID: 9695944]
[4] A. K. Bjorklund D. Ekman, *et al. J Mol Biol*. (2005) **353**: 911 [PMID: 16198373]
[5] SK Kummerfeld & SA Teichmann, *Trends Genet*. (2005) **21**: 25 [PMID: 15680510]
[6] AK Bjorklund D. Ekman, *et al. PLoS Comput Biol*. (2006) **2**: e114 [PMID: 16933986]
[7] L Brocchieri & S Karlin, *Nucleic Acids Res*. (2005) **33**: 3390 [PMID: 15951512]
[8] J Zhang, *Trends Genet*. (2000) **16**: 107 [PMID: 10689349]
[9] TJ Hubbard BL Aken *et al. Nucleic Acids Res*. (2009) **37**: D690 [PMID: 19033362]
[10] G Apic J. Gough *et al.*, *Bioinformatics*. (2001) **17** Suppl **1**: S83 [PMID: 11472996]
[11] R Apweiler TK Attwood *et al. Nucleic Acids Res*. (2001) **29**: 37 [PMID: 11125043]
[12] MA Harris J. Clark, *et al. Nucleic Acids Res*. (2004) **32**: D258 [PMID: 14681407]
[13] EM Marcotte M. Pellegrini, *et al. J Mol Biol*. (1999) **293**: 151 [PMID: 10512723]
[14] FU Hartl & M Hayer-Hartl, *Science*. (2002) **295**: 1852 [PMID: 11884745]
[15] DJ Lipman A. Souvorov, *et al. BMC Evol Biol*. (2002) **2**: 20 [PMID: 12410938]
[16] MJ Dufton, *J Theor Biol*. (1997) **187**: 165 [PMID: 9237887]

Edited by P. Kangueane
Citation: Middleton *et al*. Bioinformation 4(10): 441-444 (2010)