International Journal of Biological, Life and Agricultural Sciences
ISSN: 2415-6612
Vol:3, No:6, 2009

Proteins Length and their Phenotypic Potential

Tom Snir, and Eitan Rubin

***Abstract*—**Mendelian Disease Genes represent a collection of single points of failure for the various systems they constitute. Such genes have been shown, on average, to encode longer proteins than 'non-disease' proteins. Existing models suggest that this results from the increased likeli-hood of longer genes undergoing mutations. Here, we show that in saturated mutagenesis experiments performed on model organisms, where the likelihood of each gene mutating is one, a similar relationship between length and the probability of a gene being lethal was observed. We thus suggest an extended model demonstrating that the likelihood of a mutated gene to produce a severe phenotype is length-dependent. Using the occurrence of conserved domains, we bring evidence that this dependency results from a correlation between protein length and the number of functions it performs. We propose that protein length thus serves as a proxy for protein cardinality in different networks required for the organism's survival and well-being. We use this example to argue that the collection of Mendelian Disease Genes can, and should, be used to study the rules governing systems vulnerability in living organisms.

***Keywords*—**Systems Biology; Protein Length.

I. Introduction

MENDELIAN Disease Genes (MDGs) are genes for which an allele has been found causes with high probability and independently of other mutations a phenotype severe enough to be considered a disease. Such genes are the topic of intensive research: associating MDGs to the disease they cause is relatively a straightforward process which yields immediate clinical benefits in diagnosis and family planning, as well as provide insights into the disease underlying mechanism. In addition to those clinical implications, MDGs should also be considered as interesting models in systems biology. Their impact on the phenotype independently of mutations in other genes defines them as examples of single points of failures. This makes MDGs a functionally unique group of genes.

MDGs were shown to differ in some of their properties from genes not shown to be associated with Mendelian diseases (non-MDGs). They have been shown to be evolutionarily more conserved, to share characteristic amino acid composition and to have a wider expression pattern, as well as numerous other special features [1]-[5]. Indeed, differences in these properties were shown to allow novel MDGs to be detected using various machine learning methods [2]-[8]. Similar approaches were also successful in predicting single-knockout lethality in model organisms from gene properties, suggesting that genes that are lethal upon single knockout also define a gene of distinct genes [9].

One of the hallmarks of MDGs is their length. MDGs are more likely to be long, contain more and longer introns and encode longer proteins than non-MDGs. This bias was suggested to result from the increase in the likelihood of longer genes to undergo mutations [2],[10]. Qualitatively, this model suggests that the length effect is derived from the independent likelihood of nucleotides and amino acids to mutate. Quantitatively, the likelihood of a gene $g_{L,I}$ with length L and I introns belonging to the group of disease genes D was suggested to be modeled by:

$$P(g_{L,I} \in D) = p_{\max}(1 - \exp(-\lambda_l L - \lambda_i I)) \quad (1)$$

where λ_l and λ_i are the effective number of mutations per coding nucleotide and per intron, respectively, and $p_{\max}$ is the sum of L/I independent factors that may affect the cataloguing of a gene as being involved in disease [10]. According to so called Lopez-Bigas-Ouzounis model, the propensity of a gene to have an allele is taken to be length-dependent, and the probability for such an allele to have a disease phenotype is taken to be length-independent. The first assumption, that mutational events are independence, is deeply rooted in current understanding of DNA repair and in empirical observations. In contrast, the second assumption, namely that the phenotypic impact of mutations is length-independent, has never been assessed directly.

In the following, we bring evidence in support of the argument that the length of the protein encoded by a given gene correlates with its propensity to present a phenotype. For that purpose, we analyze saturated mutagenesis experiments, and show length affects which cannot be explained with the current model. We thus extend the existing quantitative model for the length-morbidity relation to consider a secondary length effect on the propensity of presenting a phenotype, and discuss the possible sources and meaning of this dependency in the context of systems biology. Our findings suggest that protein length might serve as a proxy for the centrality of a gene and its protein product in the multitude of networks that are essential for the functioning of living organisms.

Tom Snir was with the Shraga Segal department of Microbiology and Immunology, Ben Gurion University, Beer Sheva 84105, Israel. He has since moved to work for Teva Pharmaceutical Industries Ltd, Petach Tikva 49131, Israel .

Eitan Rubin is with the Shraga Segal department of Microbiology and Immunology, as well as the National Institute of Biotechnology in the Negev, Ben Gurion University, Beer Sheva 84105, Israel (phone: +972-8-6479197; email: erubin@bgu.ac.il).

TABLE I
THE LENGTH OF MORBID/LETHAL GENE PRODUCTS IN VARIOUS ORGANISMS

Species	Average protein length Morbid/Lethal?		MW p-value
	Yes	No	
Human	728	478	<2.2e-16
D. melanogaster	835	514	<2.2e-16
C. elegans	568	431	<2.2e-16
S. servicea	524	456	4.092e-07

The length of morbid/lethal gene products in various organisms. For each organism the average length of proteins known to be morbid/lethal ("yes") is compared to proteins not known to be morbid/lethal ("No"). The differences in product length distributions between these gene populations were also compared using the Mann-Witney test ("MW p-value")

II. METHODS

Phenotypic data sources. A list of MDGs were extracted from Morbidmap [11]. For model organisms, lists of lethal and non-lethal genes was obtained from the following databases: FlyBase for *Drosophila melanogaster* [12], WormBase for *Caenorhabditis elegans* [13], and the Saccharomyces genome database, SGD (http://www.yeastgenome.org/) for *Saccharomyces cerevisiae* [14]. For *C. elegans* and *S. cerevisiae*, only those genes that were included in previous saturated mutagenesis surveys were included. Specifically, this was achieved, in the case of the worm, by considering only those genes that are labeled with an RNAi in the annotation field and partitioning them into lethal and non-lethal gene categories based on their having Let, Larval lethal or Emb values listed in the phenotype field (using Wormmart). A similar approach was adopted for the yeast but with lists of lethal and non-lethal genes being generated based on the value of the phenotype properties field ("Viable" or "Non viable"). For *D. melanogaster*, where mutagenesis has reached near-saturation without a major survey, only those genes for which any phenotype was provided were used. Accordingly, FlyBase was first screened for genes for which phenotypic analysis had been performed, using FlyBase with the following query segment: Search Data Set = Alleles, DataBase Field = OBO Phenotype Class (PHC). Genes with the label lethal phenotype in the Phenotype Ontology field were included in the lethal genes list.

Protein length and domain counts. For each species, the databases mentioned above were used to extract the length of every protein variant expressed by each gene. To avoid redundancy, only the longest protein reported to be encoded by a given gene was considered. Similarly for domains, only that variant with the largest number of domains was considered. To obtain protein length, the length of each peptide was determined. Sequences were obtained from the databases mentioned earlier.

Statistical analysis. The significance of differences in length and domain numbers was tested using the Mann-Whitney test as implemented in R [15], using default parameters with the "two-sided" alternative hypothesis.

Programming language: Data processing was performed with the Scriptome (http://sysbio.harvard.edu/csb/resources/computational/scriptome/), a Perl-based end-user programming language. Detailed documentation and some executables are available upon request. Visualization and parameter estimation were performed with R.

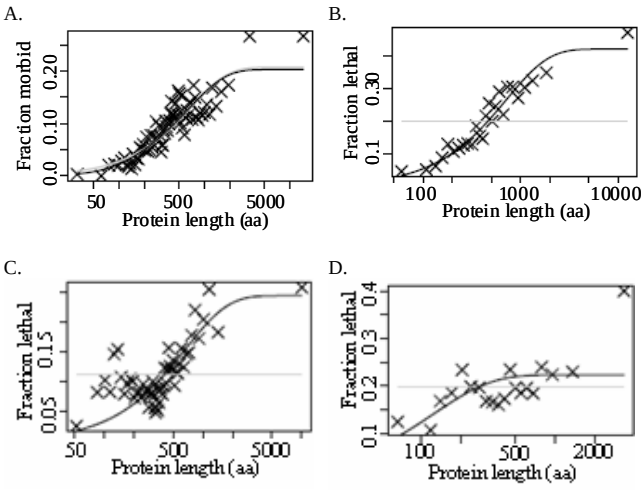

Fig. 1. The effect of protein length on the propensity of a gene to present a lethal or morbid phenotype. The fraction of disease or lethal genes is presented for genes with products of different lengths, using data from four species: Man (A), *D. melanogaster* (B), *C. elegans* (C) and *S. cerevisiae* (D). The best fit of the Lopez-Bigas-Ouzounis model (gray line) and the model developed above, which involves a length effects on phenotype likelihood (black lines) are shown for each species. For a complete description of the equation used in each model and the best fit parameters see Table II. Parameter estimation was achieved with the Nelder-Mead algorithm, as implemented in R.

III. RESULTS

In humans, where the only source of genetic variability is natural mutation, the likelihood of a gene being associated with a disease was suggested to follow equation (1) above. Assuming that lethality is a special case of morbidity, the association between gene features and disease in equation (1) can be replaced by association with lethality (i.e. replacing D with the group of all genes found to be lethal upon single mutation, or K), giving

$$P(g_{L,I} \in K) = p_{\max}(1 - \exp(-\lambda_l L - \lambda_i I)) \qquad (2)$$

It is important to note here that while in this equation is conceptually similar to the in equation (1), its actual value depends on the experimental setup, and the exact definition of lethality, since lethality is often observed only under specific growth conditions. Nevertheless, for a given experiment these should be constant and results between genes should be comparable.

How would this model apply to saturated mutagenesis

experiments? For several model organisms, relatively uniform scoring of phenotypes was achieved for a large number of genes, by methodologically testing lethality for all or nearly all genes. This allowed the application of a simple test of the model described in equation (2). In saturated mutagenesis experiments, the likelihood of every gene to mutate approaches 1, and equation (2) should converge to

$$P(g_{L,I} \in K) = p_{\max} \quad (3)$$

To test this hypothesis, we examined the relation between gene's product length and its propensity to cause a lethal phenotype in model organisms. We consider three species for which phenotypic data is accessible and most complete. In the fly *D. melanogaster*, a collection of 7633 mutations were analyzed, revealing 1498 genes with a lethal phenotype. In the worm *C. elegans*, high-throughput RNAi experiments were con-ducted to test the lethality of >90% of its known genes. A list of 16469 RNAi-analyzed genes was extracted using WormBase, of which 1835 are recorded to have a lethal phenotype. In the yeast *S. cerevisiae*, the results of systemic deletion experiments for 5886 genes were extracted from SGD, of which 1124 were found to have a lethal phenotype. For comparison, human MDGs were also analyzed, as extracted from the Morbidmap curated list of human Mendelian disease genes [11].

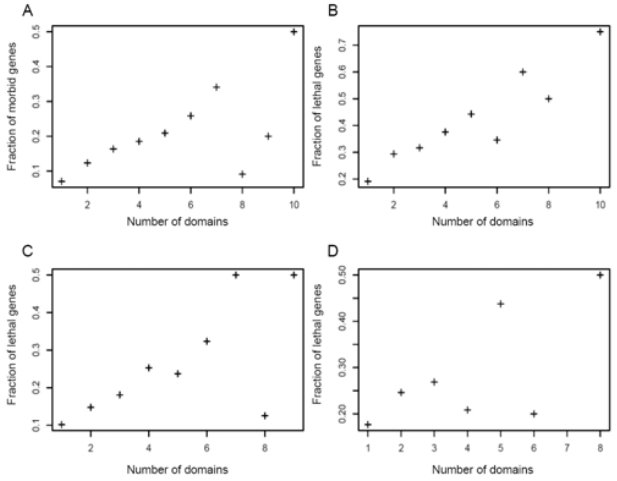

Fig. 2. The effect of domain count on the propensity of a gene to present a lethal or morbid phenotype. Genes from mutagenesis surveys and all human genes were analyzed for the occurrence of PFAM domains in the longest protein they encode. The fraction of genes found to be associated with a lethal or morbid phenotype is shown for genes with different number of PFAM domains.

Using these gene lists, we compared the length distribution of lethal (or morbid) and non-lethal (or non-morbid) genes. We were able to recapture previously reported length differences between morbid and non-morbid genes in man, and show that lethal genes are significantly longer in all 3 of the other species considered (Table I). The probability of a given gene being lethal/morbid was further explored by study-ing the fraction of genes of given length which are lethal/morbid. A clear length dependency is observed in all 4 species (Fig. 1), suggesting that the model presented in equation (2) fails to explain the results observed in saturated mutagenesis experiments.

To accommodate the fact the longer proteins are more likely to have a lethal phenotype even when the length of any gene to be mutated approaches 1, we propose that another term be added to equation (2). Assuming an independent contribution of each amino acid to the likelihood of presenting a phenotype, we obtain the following equation for natural mutation by adding a length-dependent likelihood-of-having – a-phenotype term:

$$P(g_{L,I} \in K) = p_{\max}(1-\exp(-\lambda_l L - \lambda_i I))(1-\exp(-\lambda_p L)) \quad (4)$$

where λ_p is the effective contribution of each encoded amino acid to the likelihood of a gene to present a phenotype. In the case of saturated mutagenesis, where the likelihood of having an allele reaches 1, becomes

$$P(g_{L,I} \in K) = p_{\max}(1-\exp(-\lambda_p L)) \quad (5)$$

This model, as well as the model for natural mutation (equation 4), were fitted to the observed lethal/morbid gene frequencies for the three species considered here (Table II), and compared to the length distribution of morbid genes in humans. The parameters of equations (4) and (5) were fitted to the observed frequencies of morbid/lethal genes (Fig. 2 and Table II). While the Lopez-Bigas-Ouzounis model poorly describes the length dependency of the propensity to mutate in model organisms, equation (5) fits the data reasonably well. For fly, an excellent fit is observed with a correlation coefficient of 0.96 between the observed and predicted frequencies. For worm, a group of relatively short sequences with higher than expected propensity to be lethal is observed, but for longer genes the fit is again very good, with an overall correlation coefficient of 0.75. For yeast, the effect of the length is less pronounced, and the fit of our model is not as significant as it for the other species, but still a correlation coefficient of 0.5 is observed.

One possible explanation for the observed length-function relation could be that lethal genes are more central, i.e. participate in more inter-actions in different cellular networks (e.g. metabolic or signaling networks). According to this hypothesis, lethal genes are expected to encode proteins that contain a higher number of conserved domains, assuming that multiple domains correlate with multiple functions.

Using Biomart, we obtained a list of PFAM identifiers associated with each gene product. In the case of multiple gene products for the same gene, the variant with the most domains was chosen. The number of domains found in lethal and non-lethal genes was compared for the 3 model organisms being studied and for man (Fig. 2). In all 4 species, the fraction of lethal / morbid genes grows with the number of domains. As expected, in all 4 cases, lethal genes are more likely to encode proteins with larger numbers of domains. In

humans, genes encoding a single-domain protein have a 10% chance of being morbid, as compared to 35% for genes encoding a 7-domain protein. Similar trends are also observed in the other species, although the small number of genes with multiple domains in the yeast resulted in high variability for genes with higher domain counts. Overall, these results suggest that domain count is well-correlated with the propensity of causing a lethal/morbid phenotype.

IV. Discussion

Human Mendelian Disease Genes were previously shown to be biased in terms of gene and gene product lengths [2],[3]. The bias in coding region length is largely explained by the increased likelihood of longer genes under-going mutation, directly, via the independent likelihood of a nucleotide to mutate, or indirectly, via the likelihood of a gene to encode un-stable amino acid runs [16]. The rational of this model is very compelling: since Mendelian Disease alleles are eliminated from the population at a high rate, the likelihood of observing a disease allele strongly depends on the rate of novel mutation. This model also lead to some successful predictions: MDGs are predicted to be associated with any gene feature that increases genes' vulnerability to mutation. In accordance with this prediction, MDGs have higher intron numbers (but not length), as compared to non-MDGs [10]. This observation is best explained by each intron splice signal equally and independently contributing to the likelihood of a gene to produce a faulty protein.

In all previous disussion of MDGs, the propensity of a mutated gene to have a phenotype was so far taken to be length-independent. By considering lethal genes in model animals as particular cases of morbidity, we bring evidence that the propensity to have a phenotype is associated with gene product length. We show that there exists a significant length bias between lethal and non-lethal genes in three model organisms, even as the likelihood of mutation in any gene approaches 1 in such experiments. In *D. melanogaster*, *C. elegans* and *S. cerevisiae*, significant differences between the length of lethal and non-lethal genes were observed, with the propensity of presenting a phenotype increasing with length until a maximum is reached (Fig. 1). The observed pattern is qualitatively consistent with the independent contribution of each amino acid in a protein to the likelihood of a phenotype being observed, again approaching a maximal value for long genes. Based on these findings, we propose an extended model for the length-lethality/mobidity relation, one which includes a second exponential term. The resulting model (Equation 4) is reduced to a single-exponent model when the mutation rate approaches one, and can be well fitted to the observed length effect (Fig. 1).

In humans, the extended model does not perform any better than the simpler Lopez-Bigas-Ouzounis model (compare the black and gray fitted lines in Fig. 1A). This could result from the dominance the propensity to mutate over the propensity of gene to present a phenotype. However, it is also possible that both factors are significant and that the lack of improvement in the fit is the result of noise levels and the (relatively) small number of observations. In comparison, in saturated experiments, the proposed model remains length-dependent, and can be fitted very well to the observed fraction of lethal genes at various protein lengths (Figure 1). These results cannot be explained with length independent propensity to have a phenotype.

Why are longer genes more likely to be associated with a lethal phenotype? We hypothesize that this is because longer genes are more likely to serve as functional hubs. A protein may participate in numerous functional networks both inside and outside the cell – signaling pathways, transcription regulation, primary and secondary metabolism, etc. For some networks, gene essentiality has been shown to correlate with the number of connections its product forms [17]. We suggest that having more functions of any kind (e.g. enzymatic activities, regulatory sites, DNA binding sites, etc) would impact the propensity of a protein to be long. While it is possible to imagine compact proteins with high cardinality in this meta-network of biological activities, and indeed examples of such proteins exist, it is reasonable to assume that, on average, added interactions would be associated with increased protein length. Partial support for this hypothesis comes from our analysis of conserved PFAM domain counts. As expected from our model, a clear correlation is observed in most species between the propensity of its absence being lethal/morbid and the number of conserved domains a protein encodes (Figure 2). This suggests that proteins associated with lethal/morbid phenotypes indeed encompass more functions, on average, than do non-lethal/morbid proteins.

Together, we interpret the length-lethality relation in saturated mutagenesis surveys and the length-morbidity relation in natural human populations to suggest that protein length can serve as an indirect proxy for centrality. Highly connected proteins, or "Hubs", are more likely to be encoded by essential genes [18]. Morbidity and lethality are observed in vivo, and may correlate with the over-all connectivity of the gene and it's product in a "super network" that arises from combining all the cellular and intercellular networks essential for an organism's well-being and survival. Morbid and essential genes represent single point of failures in this meta-network.

The analysis of the Mendelian Disease Genes for common features is an active research area, deeply rooted in the traditions of genetics. Originally, geneticist focused on "forward genetics", an approach in which observed phenotypes are studied for their underlying genetic basis. By comparison, systems biology heavily relies on approaches which begin with perturbations of underlying genes in search for rules that explain emerging phenotypic changes (an approach a geneticist will call "reverse genetics". Studies of properties common to all MDGs can foster understanding of the properties of single points of failure in the "super network" of all human networks by starting from phenotypic observations and working backwards to common gene

properties. As this approach is the inverse of typical approaches to systems biology, we propose it should be called "reverse systems biology". Applying this approach to other characteristics of MDGs may help understanding the underlying principles of systems failure in human health.

ACKNOWLEDGEMENT

This work was supported by the Center for Complexity Science, the RICH foundation and the National Institute for Biotechnology in the Negev. We would to thank Prof. Jerry Eichler and unnamed reviewers for useful comments.

REFERENCES

[1] Botstein, D. and Risch, N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease, *Nat Genet*, 33 Suppl, pp 228-237.
[2] Lopez-Bigas, N. and Ouzounis, C.A. (2004) Genome-wide identification of genes likely to be involved in human genetic disease, *Nucleic Acids Res*., 32, pp. 3108-3114.
[3] Kondrashov, F.A., Ogurtsov, A.Y. and Kondrashov, A.S. (2004) Bioinformatical assay of human gene morbidity, *Nucleic Acids Res*., 32, pp. 1731-1737.
[4] Adie, E.A., Adams, R.R., Evans, K.L., Porteous, D.J. and Pickard, B.S. (2005) Speeding disease gene discovery by sequence based candidate prioritization, *BMC Bioinformatics*, 6, pp. 55-88.
[5] Jimenez-Sanchez, G., Childs, B. and Valle, D. (2001) Human disease genes, *Nature*, 409, pp. 853-855.
[6] Oti, M., Snel, B., Huynen, M.A. and Brunner, H.G. (2006) Predicting disease genes using protein-protein interactions, *J Med Genet*. 43, pp. 691-8.
[7] Perez-Iratxeta, C., Bork, P. and Andrade, M.A. (2002) Association of genes to genetically inherited diseases using data mining, *Nat Genet*, 31, pp. 316-319.
[8] Turner, F.S., Clutterbuck, D.R. and Semple, C.A. (2003) POCUS: mining genomic sequence annotation to predict disease genes, *Genome Biol*, 4, pp. R75.
[9] Seringhaus, M., Paccanaro, A., Borneman, A., Snyder, M. and Gerstein, M. (2006) Predicting essential genes in fungal genomes. *Genome Res*., 16, pp 1126-1135
[10] Lopez-Bigas, N., Audit, B., Ouzounis, C., Parra, G. and Guigo, R. (2005) Are splicing mutations the most frequent cause of hereditary disease?, *FEBS Lett*, 579, pp. 1900-1903.
[11] Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. and McKusick, V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*., pp. D514-517.
[12] Drysdale, R. and The FlyBase Consortium, (2008). FlyBase : a database for the Drosophila research community. *Methods Molec. Biol.* 420, pp 45-59
[13] Chen, N. et. al (2005) WormBase: a comprehensive data resource for Caenorhabditis biology and genomics, *Nucleic Acids Res*, 33, pp. D383-389.
[14] Cherry, J.M., Adler, C., Ball, C., Chervitz, S.A., Dwight, S.S., Hester, E.T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S. and Botstein, D. (2006) SGD: Saccharomyces Genome Database. *Nucleic Acids Res.* 26, pp. 73-79.
[15] Ihaka, R. and Gentleman, R. (1996) R: A language for data analysis and graphics, *Journal of Computational and Graphical Statistics* 5, pp. 299-314.
[16] Karlin, S., Chen, C., Gentles, A.j. and Cleary, M. (2002) Associations between human disease genes and overlapping gene groups and multiple amino acid runs. *Proc. Nat. Acad Sci* 99, pp. 17008-17013
[17] Jeong, H., Mason, S.P., Barabási, A..L. and Oltvai, Z.N. (2001) Lethality and centrality in protein networks. *Nature* 411, pp. 41-42.
[18] Batada, N.N., Hurst, L.D. and Tyers, M. (2006) Evolutionary and Physiological Importance of Hub Proteins. *PLoS Comput Biol* 2, pp. e88.

TABLE II
ESTIMATED PARAMETERS FOR TWO MODELS OF LENGTH-PHENOTYPE RELATIONSHIPS

Species	Length independent				Length dependent				
	Eq.	$p_{\max}$	λ_l	R^2	*Eq.*	$p_{\max}$	λ_l	λ_p	R^2
Human	2	0.21	1.5E-3	0.887	4	0.20	0.016	1.5E-3	0.889
D. melanogaster	3	0.20	n/a	n/a	5	0.42	n/a	1.4E-3	0.963
C. elegans	3	0.11	n/a	n/a	5	0.24	n/a	1.5E-3	0.750
S. servicea	3	0.22	n/a	n/a	5	0.22	n/a	8E-3	0.513

The best fit parameters for two possible models of the dependency of morbidity (human) and lethality (fly, worm and yeast) on protein length. Parameters of the appropriate equations (Eq.) were fit for each species, assuming the original Lopez-Bigas-Ouzounis model (Length independent) or the extended model presented here (Length dependent). Where applicable, the correlation coefficient between observed and calculated frequencies is provided (R^2).