Determinants of the rate of molecular evolution in fission yeast

Simon Harnqvist[1*]
[1] Department of Biology, University of York, YO10 5DD, United Kingdom
[*] Corresponding author: seh589@york.ac.uk

ABSTRACT

Background
What determines the rate of gene and protein sequence evolution is a central question in evolutionary genomics. Although many of the variables that influence conservation across most species are known, the relative importance of each one is still debated. In the model organism fission yeast *(Schizosaccharomyces pombe)*, the determinants of the rate of sequence evolution have yet to be determined. In most organisms, gene expression levels are the most important variable, but network centrality in the protein-protein interactions network is also believed to be significant. Functional importance has, counterintuitively, often been found to be less important. Many other variables are known to have some impact.

Results
This study examines the relationships between numerous variables and gene sequence evolution rate in fission yeast. Partial least squares (PLS), a nonparametric multivariate regression technique, is employed to provide a quantitative indication of the influence of each variable. Network centrality and gene expression are found to be the two most important variables impacting constraint, with functional importance and gene/protein size each playing lesser yet significant roles. There is considerable heterogeneity within variable groups such as gene ontology categories and amino acid composition. The PLS model explains less than a third of variance in constraint, which is shown to be mostly due to the data rather than choice of regression technique.

Conclusion
This study provides further evidence of the general importance of both network centrality and gene expression as determinants of constraint. However, the majority of variance in constraint in fission yeast remains unexplained by this study, suggesting that the rate of molecular evolution is largely stochastic.

BACKGROUND
The question of which variables determine the rate of sequence evolution in both genes and proteins is arguably one of the most central in evolutionary genomics. While there is a long list of variables that are believed to influence the rate of sequence evolution, the importance of each has still not been explored in fission yeast. The importance of each variable has been quantitatively examined – notably in humans by Alvarez-Ponce and colleagues (1) – using multivariate regression techniques to estimate the influence of each variable. Because of the importance of fission yeast (*Schizosaccharomyces pombe)* in genetic research, it is important

to investigate whether the drivers of molecular evolution differ in fission yeast compared to other organisms.

When the question of why some genes evolve faster than others was first explored in the 1960s, the answer based on the neutral theory of molecular evolution (2) was that the most important genes (those with highest *functional importance)* would be those to evolve the most slowly. This hypothesis is intuitive; the more important a protein is, the stronger the negative consequences of harmful mutations, and therefore the stronger the selective pressure to limit sequence changes. With the rise of molecular biology which has allowed gene knockout experiments, contradictory evidence both an in favour and against this hypothesis has been generated. Whereas early work in the 1970s (3) found no such relationship once duplicate genes had been removed, various other studies have found functional importance to be important in mice (4), bacteria (5) and baker's yeast (*Saccharomyces cerevisiae)* (6). The preponderance of evidence appears to be that functional importance is influential, but not the most important in any major model organism (4).

Instead, gene expression level has arisen as the seemingly most important factor determining sequence conservation – this is true overall in most organisms (4) and a strong correlation has indeed been shown in fission yeast (7). There are at least four hypotheses as to why this is, all well summarised by Zhang and Yang (4). The first is the protein misfolding avoidance hypothesis (8), which argues that highly expressed proteins are under higher selective pressure to avoid toxic misfolding, as the consequences would be worse the more of the protein is produced. Avoiding misfolding requires translational robustness, which constrains sequence evolution. The second hypothesis is the protein misinteraction avoidance hypothesis (9), reasoning that the cost of random misinteractions increases with protein concentration because a higher number of molecules will take part, meaning that the selective pressure to prevent such interactions is higher for more highly expressed proteins. The mRNA folding requirement hypothesis (10), suggests that mutations in highly expressed mRNAs would have worse consequences in the strongly folded mRNAs of highly expressed genes in the more weakly folded mRNAs of less expressed genes. Finally, the expression cost hypothesis (11,12), proposes that because a mutation reducing a protein's fitness by a fraction k is equivalent to a loss of the same fraction k protein molecules, the fitness loss to the organism would be greater with highly expressed proteins. It is likely that several of these mechanisms contribute at least to some extent to the correlation between constraint and expression, making it apparent why this correlation is so strong, relatively speaking.

Connectivity in the protein-protein interactions network is also often listed as an influential determinant of constraint (1,13–15), although one study that used the baker's yeast, worm, and fly genomes (16) failed to detect any such correlation. The relationship reported in most studies is that more connected proteins (i.e., they have either more direct connections or are on more signalling paths than most proteins) are evolving more slowly. This, as first suggested by Ingram (14), is generally thought to be because mutations that disrupt interactions of highly connected proteins would be more detrimental to the organism than mutations that prevent the binding of less connected ones.

Additionally, a host of variables influencing the rate of sequence evolution in a range of organisms has been reported before, including but not limited to protein length (17), pleiotropy (16), intron presence (18), cellular location (19), and codon bias (20); Zhang and Yang (4) provide a comprehensive review of the various variables implicated in sequence evolution. The present study uses publicly available sequence evolution and functional genomics data mainly from Grech and colleagues (21), with additional data from the PomBase (22) and STRING (23) databases, to investigate some of the variables that may influence gene

sequence evolution. At time of writing, there appears to be no readily available data on protein sequence evolution in *S. pombe*; this article does nonetheless discuss sequence evolution in both protein and DNA sequences.

RESULTS AND DISCUSSION

To determine which factors are the most influential in determining the rate of molecular evolution, various functional genomics data were retrieved from Grech and colleagues (21), PomBase (22), and STRING 11 (23). These included general gene and protein data (location, size, length, charge, amino acid composition), expression data, gene ontology data (cellular process, biochemical function, and cellular component), interactors in the protein-protein network, essentiality and knockout fitness data, and data on constraint, measured as phyloP score (24). Gene ontology "slims" (broad categories) were retrieved using GO Mapper (25). Correlations were determined between each continuous variable and constraint, as was the mean difference in constraint for each categorical (binary encoded) variable between present/true (1-encoded) and absent/false (0-encoded). To model constraint by all other variables, a partial least squares regression (PLS) model was trained, chosen because it is a regression technique suitable with noisy and correlated variables (26). To estimate the importance of each variable in the model, variable importance in projection (VIP) scores were calculated. VIP can be described as the influence each variable has on the weights of each model component, accounting for the variance explained by the component (27). For reference, a rule of thumb in variable selection is that variables with a VIP score below 1 can be removed (27,28).

Network centrality is the most significant determinant of constraint, with gene expression second

Somewhat surprisingly, degree centrality (the number of direct interactors in the protein-protein interactions network) explains most variance with a VIP score of 3.65 (fig 1). It also correlates the strongest with constraint (Kendall's $\tau = 0.29$, p.adj. = 3.2×10^{-205}, fig 2), further suggesting the importance of this variable. Betweenness centrality, which can be thought of as the importance a node has in relaying information in a network (29), was found to be less important, with a VIP score of only 1.70, but nonetheless also with a strong correlation in this context ($\tau = 0.21$, p.adj. = 2.0×10^{-111}), the latter presumably due to high collinearity with degree centrality. The high importance of network centrality is consistent with previous findings in *Saccharomyces cerevisiae* (15) and even more so with more recent research in humans (1) showing that network centrality may be the most significant determinant of constraint.

Less surprisingly is that the variable so often found to be most influential in many organisms (4), gene expression levels, is the second most important. With a comparatively high Kendall rank correlation coefficient (τ) of 0.25 (p.adj. = 2.0×10^{-155}) but a variable importance in projection score of 2.47, this variable is not quite as important as network centrality. Notably, the influence of protein expression is limited, whether measured as correlation and its importance in projection, showing that effect of expression on sequence evolution rate appears to be determined at the transcriptional level.

As was described above, multiple mechanistic reasons why these variables are so significant have been proposed, and indeed this study helps further confirm that both network centrality and gene expression levels are significantly associated with sequence evolution constraint in general, and in fission yeast in particular.

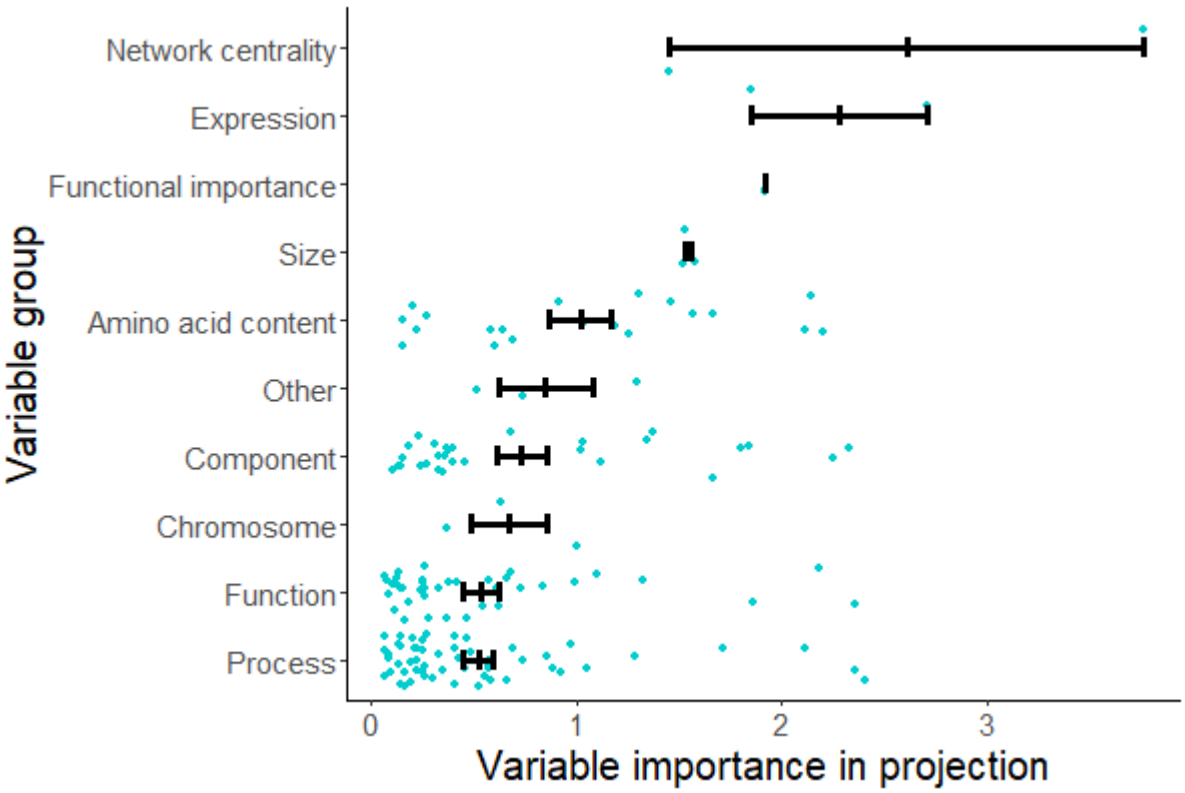

Figure 1. Variable importance projection scores per variable from partial least squares regression of constraint, grouped into variable groups. *Degree centrality, gene expression, essentiality, protein expression, and knockout fitness are some of the most important variables in the PLS regression model. Variable groups that comprise more than a few variables, i.e. process, function, component, and amino acid content, show considerable heterogeneity in importance. Error bars show mean ± standard error.*

Functional importance and sequence length are influential determinants of the rate of evolution

As previously discussed, one would intuitively expect that the importance of a gene for fitness would be the most important determinant of constraint. While this is not true in the fission yeast protein coding genome, and indeed not in most genomes studied (4), both essentiality (VIP = 1.97) and knockout fitness (VIP = 1.89) appear to be important still. Essential genes have a constraint of on average 0.10 phyloP units higher (p.adj. = 3.3×10^{-59}) than the average gene, a 26% increase (fig. 3). Knockout fitness shows the strongest negative correlation of all variables (τ = -0.17, p.adj. = 8.7×10^{-72}) with constraint, meaning that lower constraint (faster evolution) correlates with higher knockout fitness. In other words, less important genes can undergo more sequence change, just as one would expect. These findings are consistent with other studies in various organisms which have showed that functional importance is indeed important, just not the most important of the variables.

Sequence length – measured as gene or protein length – also correlates positively with conservation (τ = 0.12, p.adj. = 5.4×10^{-38} for gene length), with moderate VIP scores of about 1.6 for each of gene length (1.64), protein length (1.61), and protein mass (1.60). This relationship has been reported in several organisms, including humans (17) and *Saccharomyces cerevisiae* (17, 30). Lipman and colleagues (17) suggested that this is because of selective pressure to make proteins as short as they can be without disrupting function in order to reduce

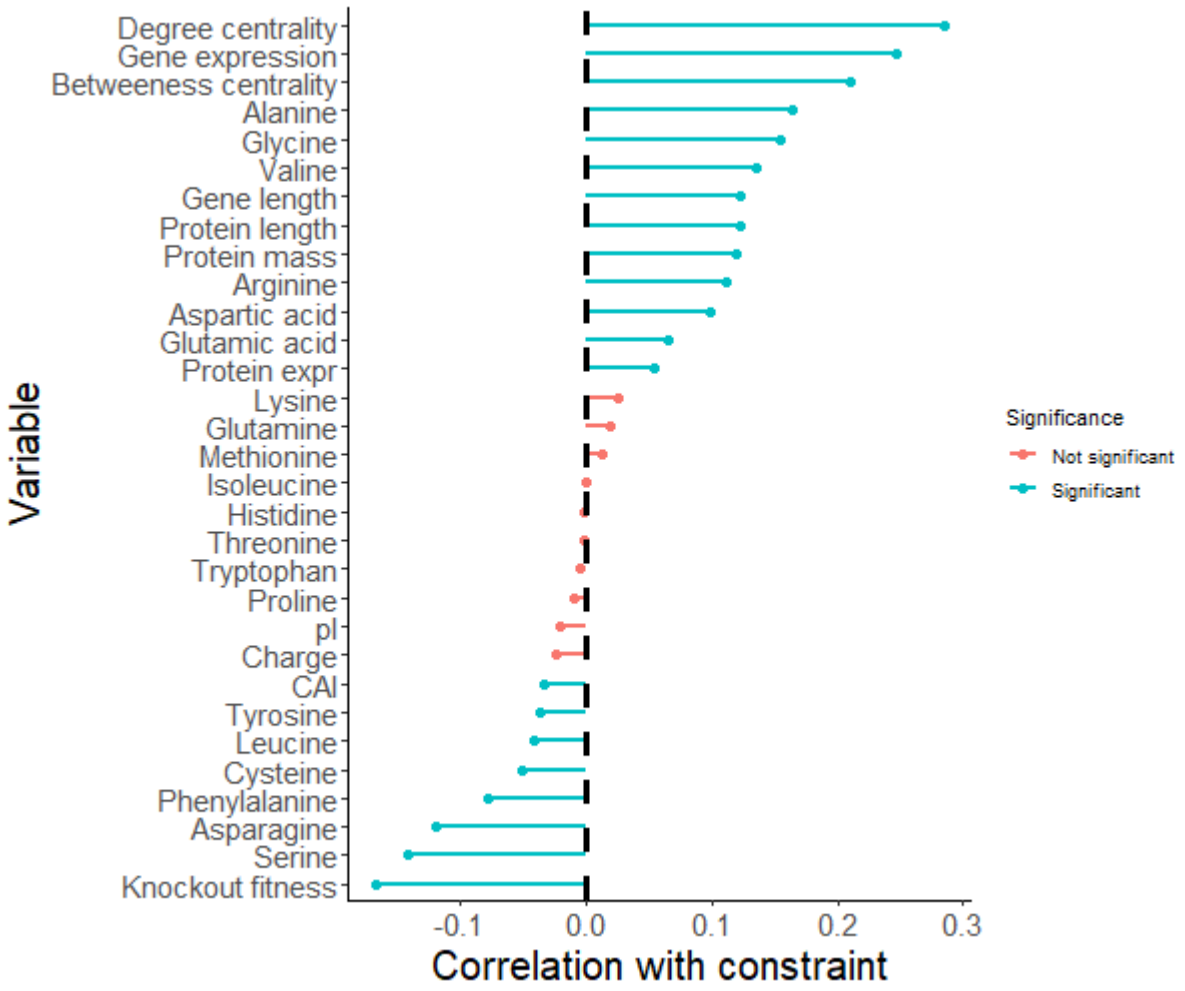

Figure 2. Kendall rank coefficients for correlations between constraint and each continuous variable. *The strongest correlations are with network centrality and gene expression variables, with the strongest correlations all being positive rather than negative (anti-correlations). Quasi-continuous variables (e.g., gene length) have been treated as continuous. Significance threshold p.adj. < 0.05, with Bonferroni correction for multiple tests.*

the cost of translation, meaning that both proteins with high functional importance and proteins with specific functional roles are likely to be longer. Length is probably a proxy for several confounding variables.

Functional role and amino acid content display large heterogeneity in contribution to constraint

As is clear from figures 1 and 3, some gene ontology categories contribute strongly to increased or decreased constraints, whereas most categories appear to have a negligible influence. This may be because GO categories might be less well defined, resulting in a bias in measured constraint towards well-defined categories, although one would expect a considerable heterogeneity for biological reasons as well. Table 1 shows the GO slims with the ten highest variable importance in projection scores (the ten most predictive slims). Three of these relate to translation; this is not very surprising given the complexity of the transcriptional machinery, as one would expect that even minor mutations could result in cell death. Other GO slims with a high level of constraint are nitrogen metabolic processes, which includes the

synthesis of nucleosides and amino acids, and complexing, likely due to need to conserve structure that allows complex formation.

Table 1. Gene ontology (GO) slims with the ten highest variable importance in projection (VIP) scores.

GO Slim	VIP score	Constraint difference (phyloP)	p. adj.
Process: Cellular nitrogen compound metabolic process	2.25	0.10	2.0×10^{-60}
Process: Translation	2.17	0.18	3.3×10^{-52}
Component: Protein-containing complex	2.13	0.11	1.4×10^{-59}
Function: Structural constituent of ribosome	2.06	0.23	5.0×10^{-35}
Component: Ribosome	2.06	0.22	4.7×10^{-35}
Process: Biosynthetic process	1.98	0.10	1.4×10^{-48}
Component: Intracellular	1.95	0.11	4.3×10^{-34}
Function: Structural molecule activity	1.94	0.19	2.0×10^{-30}
Component: Cell	1.89	0.13	3.5×10^{-38}
Component: Cytosol	1.82	0.09	6.9×10^{-43}

Constraint difference refers to the difference in constraint (phyloP) when the variable is 1 (e.g. the gene has an intracellular function) from when the variable is 0 (e.g. the gene does not have an intracellular function). P.adj. = Bonferroni corrected p-values.

The large heterogeneity in variable importance is also true for amino acid composition, as seen in figures 1 and 2. The changeabilities of amino acids are known to differ depending on the structural requirements of the protein domain – particularly, there is a considerable difference in amino acids composition between transmembrane and other protein domains (31). It is known that alanine ($\tau = 0.16$, p.adj. = 4.1×10^{-68}, VIP = 2.22) and glycine ($\tau = 0.15$, p.adj. = 5.0×10^{-61}, VIP = 2.15), which relatively speaking strongly correlate with conservation, are highly enriched in Low Complexity Regions (LCRs) of proteins, which are known to be highly conserved (32). That said, some of the amino acids that show no correlation with constraint (e.g. glutamine, $\tau = 0.01$, p.adj. = 1, VIP = 0.67) or even display negative correlation (e.g. serine, $\tau = -0.14$, p.adj. = 1.7×10^{-51}, VIP = 1.81), are also common in LCRs (33). Regardless, what is clear is that amino acid composition serves as a proxy for protein domains and regions that are under different selective pressures, in turn affecting constraint.

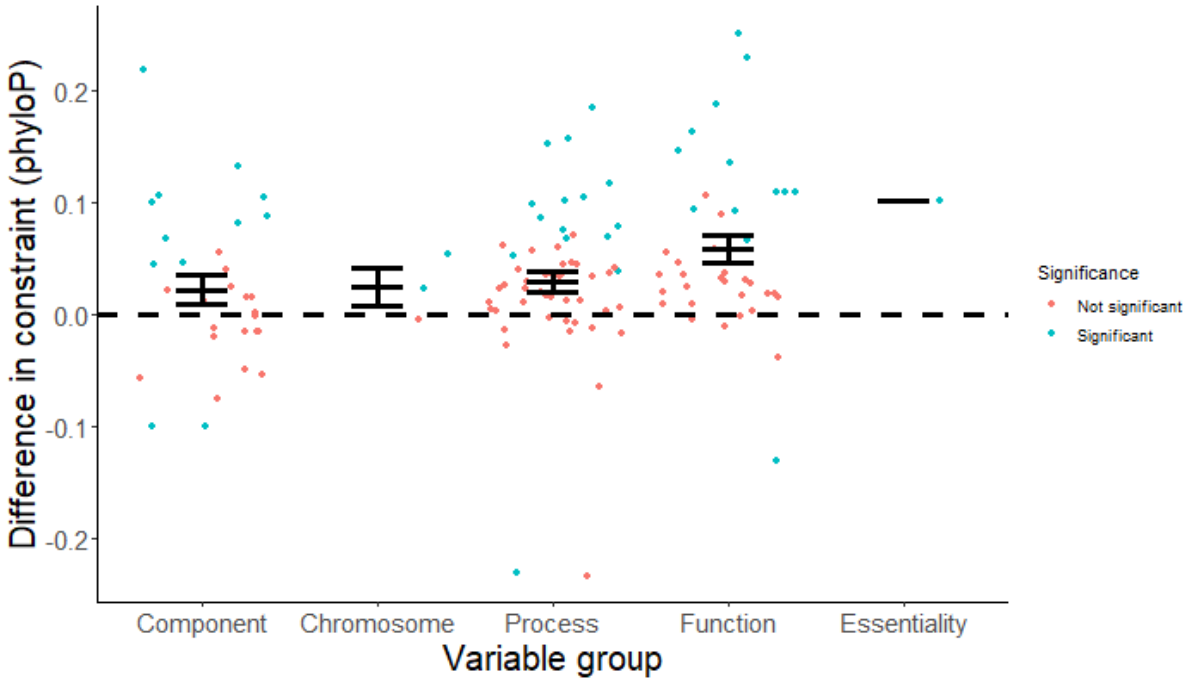

Figure 3. Difference in constraint between the presence and absence of each binary (one-hot encoded) variable, grouped by variable groups. *Each data point shows the mean of constraint for those genes for which the variable is true/1 (e.g., the gene is on chromosome 1) minus the value of constraint for those genes for which the variable is false/0 (e.g., the gene is* not *on chromosome 1). There is considerable heterogeneity in the difference in constraint between different variables; in most cases, the difference is not statistically significant. Significance threshold p.adj. < 0.05, with Bonferroni correction for multiple tests.*

Most of the variance in sequence evolution rate remains unexplained by these variables
Combined in a partial least squares model, these variables only explain 29.5% of the variance in constraint at the optimum number of components (3). This shows that the model fails to adequately explain most of the variation in constraint. This could be due to a poor choice of modelling technique, but comparison with principal components regression (PCR), a highly similar method to PLS, shows that PLS explains more variance at far fewer components, with comparable root mean squared error (table 2). A random forest (34) regression model, a state-of-the-art machine learning technique suitable for high-dimensional noisy data, shows better performance – but still fails to explain over 60% of the variance. Note that random forests are not necessarily suitable for robust and unbiased variable importance estimation (35), which is why variable importance estimation for the RF model has not been done here.

While the fit of the PLS model is not ideal, none of these three techniques manage to produce a model that explains even half of the variance in conservation. Given the high flexibility of random forest models in particular, it does appear unlikely that any common regression technique would allow construction of a model explaining much more than 50% variance. In other words, the high unexplained variance is primarily a property of the data and is not just due to model choice. Notably, multivariate models of protein evolutionary rates in humans (1) and *Arabidopsis* (36) have only been able to explain between 11% and 18% of variance in rates.

Table 2. A comparison of the performance of three regression techniques applied to sequence evolution constraint prediction.

Regression model	RMSE	Variance in constraint explained	Number components
Partial least squares (PLS)	0.176	29.5%	3
Principal components (PCR)	0.169	27.2%	32
Random Forest (RF)	0.155	38.5%	

RMSE = root mean squared error. Variance explained calculated as 1 – mean squared error of prediction, divided by variance in constraint. Evaluation was performed on the 20% test set.

One possible explanation for the low proportion of variance explained is poor data quality – particularly of interactions and functional role data. Another possible reason is that some variables have not been explored, or not explored in sufficient depth. For instance, this study has analysed what the impact of which chromosome a gene is on has on constraint, ignoring finer details such as where on the chromosome – this data is available but difficult to include in modelling without overfitting– or location in the 3D genome. Similarly, individual GO terms rather than GO slims could be considered, but this would lead to difficulties in interpretation as well as a problems with overfitting.

The more interesting explanation for the poor fit is that the rate of sequence evolution for a gene in the protein coding genome of fission yeast may be largely stochastic. This would be consistent with the neutral theory of molecular evolution (2), which states that most genetic variation is due to random drift. In this scenario rates of evolution should show considerable heterogeneity due to chance, resulting in large statistical noise. It appears likely that this would account for a considerable part of the missing variance. Determining how much of this is due to data quality, how much is due to missing variables, and how much is due to stochastics may prove very difficult to determine. A systematic comparison across multiple species could provide useful insights about the variation in unexplained variance; this in turn might indicate the roles of data availability and quality, both of which vary greatly between organisms.

CONCLUSION

Unlike what has been found in most organisms (4), it appears that network centrality measured as number of interactors (degree centrality) is the single variable that most strongly influences the rate of molecular evolution in fission yeast. Given that research from the past few years in the human genome (1) has shown the high importance of network centrality as a variable reducing the rate of evolution, it is possible that this finding might be generally true across many organisms. Perhaps, until recently, this conclusion has been obscured by poor quality proteomics data – whereas high quality gene expression data has been available for longer. Only a minority of variance in constraint is explained by all the variables combined, which is shown to be in part due to the limitations of the PLS regression technique, but largely due to the data itself. It consequently appears that the rate of molecular evolution in fission yeast is influenced by variables not considered here, and that stochastics may have a highly important role as well.

METHODS

Data sources

The gene list (n = 5609) of most protein coding genes in the fission yeast genome, with data on constraint, gene and protein expression expression, gene length, chromosome, essentiality and solid media fitness, was retrieved from Grech and colleagues (21), downloaded manually from their article's Figshare repository. Interactome data were retrieved from STRING using FTP download. Gene ontology data were retrieved using the GO Term Mapper (25), with the gene list from Grech et al. as input. All other data, including amino acid composition and protein length and size, were retrieved via FTP from PomBase (22). The links to the datasets are available below (see "Original data sources").

Data pre-processing

Gene ontology (GO) annotation terms as well as chromosome data were one-hot encoded; using the mltools R package (37); all other variables were already numeric. Amino acid composition data were scaled from number of nucleotides to proportion of protein. Betweenness and degree centrality measures were calculated for each gene based on protein interactions data with the igraph (38) R package, using a minimum interaction score of 0.400. Missing data were imputed using the missForest R package (39). All work was carried out in R 4.0.0 (40) with extensive use of the Tidyverse packages (41).

Partial least squares regression modelling

Partial least squares regression (PLS/PLSR) was chosen to model constraint by all other variables because it is suitable for multivariate non-parametric modelling with noisy and non-orthogonal variables (26,27). It also allows easy interpretation of variable influence on the model by calculating the variable importance in projection (VIP) score for each variable (27). The model was trained with the pls R package (42) using all variables except gene name and constraint as independent variables (X), with constraint as the dependent variable (y) to be predicted. All data were centred and scaled. A 20% random holdout set was used for validation to guard against overfitting, and ten-fold cross-validation was used in training. The optimum number of components was chosen with the selectNcomp function in the pls package.

Model comparisons

As comparison with the PLS model, principal component regression (PCR) and random forest (RF) models were trained. The PCR model was trained with the pls package (42), whereas the Random Forest model was trained using randomForest (43). The number of components to keep in the PCR model with selectNcomp, as for the PLS model. The random forest model was trained using 500 trees with mtry set to number of variables / 3 (defaults). Variance explained by each model was calculated as 1 – mean squared error of prediction, divided by the variance in true y.

Variable importance in projection

VIP scores were calculated using the PLSVarSel package (27). Take the following PLS model, where X is the matrix of independent variables, y is the vector of dependent variable, T, P, and b are vectors of predictors, and f and E are errors (example from 28):

$$X = TP^T + E$$
$$y = Tb + f$$

For this model, the VIP (v_j) for each variable is found by (28):

$$v_j = p \sum_{k=1}^{h} \left(SS(b_k t_k)(\frac{w_{jk}}{||w_k||})^2 \right) / \sum_{k=1}^{h} SS(b_k t_k)$$

Where SS stands for the sum of squares, k is the component, h is the total number of latent variables, b_k explains the relationship between t_k and y, and t_k is a the kth column vector of T, W is the matrix of the model weights, which are found such that ||f|| is minimised. (28) As previously mentioned, the rule of thumb is that variables with a VIP of less than 1 can be excluded from a model (28) – however, the purpose here is not variable selection, but importance inference.

Statistical comparisons

Kendall correlations were calculated between constraint and each continuous variable. For categorical one-hot encoded variables, the Mann-Whitney-Wilcoxon test was used to determine significance of the difference in constraint when each variable was true or present (1-encoded) and when it was false or absent (0-encoded). All p-values were adjusted using Bonferroni correction.

DATA AND CODE AVAILABILITY

Original data sources

The dataset from Grech et al. was retrieved from Figshare (https://figshare.com/articles/dataset/gene-based_data/6265748). Protein data was retrieved from PomBase (ftp://ftp.pombase.org/pombe/Protein_data/PeptideStats.tsv and ftp://ftp.pombase.org/pombe/Protein_data/aa_composition.tsv). Interactions data were retrieved from STRING v 11 (https://stringdb-static.org/download/protein.links.v11.0/4896.protein.links.v11.0.txt.gz).

Data availability

All data are provided both in their original format and processed on Figshare (https://doi.org/10.6084/m9.figshare.c.5263523.v1).

Code availability

All R scripts are available both in the Figshare repository and on GitHub (https://github.com/simonharnqvist/pombe)

ACKNOWLEDGEMENTS

Thanks to Daniel Jeffares for his support and advice – this article would not have been written without him. No specific funding was awarded for this work.

SH is an employee of GlaxoSmithKline at time of publication.

REFERENCES

1. Alvarez-Ponce D, Feyertag F, Chakraborty S. Position Matters: Network Centrality Considerably Impacts Rates of Protein Evolution in the Human Protein–Protein Interaction Network. Genome Biology and Evolution. 2017 Jun;9(6):1742–56.

2. Kimura M. Evolutionary Rate at the Molecular Level. Nature. 1968 Feb;217(5129):624–6.

3. Wilson AC, Carlson SS, White TJ. Biochemical Evolution. Annual Reviews in Biochemistry. 1977;46:573–639.

4. Zhang J, Yang J-R. Determinants of the rate of protein sequence evolution. Nat Rev Genet. 2015 Jul;16(7):409–20.

5. Jordan IK, Rogozin IB, Wolf YI, Koonin EV. Essential Genes Are More Evolutionarily Conserved Than Are Nonessential Genes in Bacteria. Genome Research. 2002;12:962–8.

6. Hirsh AE, Fraser HB. Protein dispensability and rate of evolution. Nature. 2001 Jun;411(6841):1046–9.

7. Mata J. Correlations Between Gene Expression and Gene Conservation in Fission Yeast. Genome Research. 2003 Dec 1;13(12):2686–90.

8. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. Why highly expressed proteins evolve slowly. Proceedings of the National Academy of Sciences. 2005 Oct 4;102(40):14338–43.

9. Yang J-R, Liao B-Y, Zhuang S-M, Zhang J. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. Proceedings of the National Academy of Sciences. 2012 Apr 3;109(14):E831–40.

10. Park C, Chen X, Yang J-R, Zhang J. Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. Proceedings of the National Academy of Sciences. 2013 Feb 19;110(8):E678–86.

11. Cherry JL. Expression Level, Evolutionary Rate, and the Cost of Expression. Genome Biology and Evolution. 2010 Oct 25;2(0):757–69.

12. Gout J-F, Kahn D, Duret L, Paramecium Post-Genomics Consortium. The Relationship among Gene Expression, the Evolution of Gene Dosage, and the Rate of Protein Evolution. Pritchard JK, editor. PLoS Genet. 2010 May 13;6(5):e1000944.

13. Fraser HB, Hirsh AE. Evolutionary rate depends on number of protein-protein interactions independently of gene expression level. BMC Evolutionary Biology. 2004;5.

14. Ingram M. Gene Evolution and the Haemoglobins. Nature. 1961;181(4766):704–8.

15. Jovelin R, Phillips PC. Evolutionary rates and centrality in the yeast gene regulatory network. Genome Biol. 2009;10(4):R35.

16. Hahn MW, Kern AD. Comparative Genomics of Centrality and Essentiality in Three Eukaryotic Protein-Interaction Networks. Molecular Biology and Evolution. 2005 Apr;22(4):803–6.

17. Lipman DJ, Souvorov A, Koonin EV, Panchenko AR, Tatusova TA. The relationship of protein conservation and sequence length. BMC Evolutionary Biology. 2002;10.

18. Carmel L, Rogozin IB, Wolf YI, Koonin EV. Evolutionarily conserved genes preferentially accumulate introns. Genome Research. 2007 Jun 13;17(7):1045–50.

19. Liao B-Y, Weng M-P, Zhang J. Impact of Extracellularity on the Evolutionary Rate of Mammalian Proteins. Genome Biology and Evolution. 2010 Jan 1;2:39–43.

20. Drummond DA, Wilke CO. Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. Cell. 2008 Jul;134(2):341–52.

21. Grech L, Jeffares DC, Sadée CY, Rodríguez-López M, Bitton DA, Hoti M, et al. Fitness Landscape of the Fission Yeast Genome. Molecular Biology and Evolution. 2019 Aug 1;36(8):1612–23.

22. Lock A, Rutherford K, Harris MA, Hayles J, Oliver SG, Bähler J, et al. PomBase 2018: user-driven reimplementation of the fission yeast database provides rapid and intuitive access to diverse, interconnected information. Nucleic Acids Research. 2019 Jan 8;47(D1):D821–7.

23. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019 Jan 8;47(D1):D607–13.

24. Gagliano SA, Barnes MR, Weale ME, Knight J. A Bayesian Method to Incorporate Hundreds of Functional Characteristics with Association Evidence to Improve Variant Prioritization. PLOS ONE. 2014 May 20;9(5):e98122.

25. Lewis-Sigler Institute for Integrative Genomics. Generic gene ontology (GO) term mapper [Internet]. Princeton.edu. Available from: https://go.princeton.edu/cgi-bin/GOTermMapper

26. Cramer RD. Partial Least Squares (PLS): Its strengths and limitations. Perspectives in Drug Discovery and Design. 1993 Dec;1(2):269–78.

27. Mehmood T, Liland KH, Snipen L, Sæbø S. A review of variable selection methods in Partial Least Squares Regression. Chemometrics and Intelligent Laboratory Systems. 2012 Aug;118:62–9.

28. Chong I-G, Jun C-H. Performance of some variable selection methods when multicollinearity is present. Chemometrics and Intelligent Laboratory Systems. 2005 Jul;78(1–2):103–12.

29. Freeman LC. A Set of Measures of Centrality Based on Betweenness. Sociometry. 1977 Mar;40(1):35.

30. Bloom JD, Drummond DA, Arnold FH, Wilke CO. Structural determinants of the rate of protein evolution in yeast. Mol Biol Evol. 2006 Sep;23(9):1751–61.

31. Tourasse NJ, Li W-H. Selective Constraints, Amino Acid Composition, and the Rate of Protein Evolution. Molecular Biology and Evolution. 2000 Apr 1;17(4):656–64.

32. Ntountoumi C, Vlastaridis P, Mossialos D, Stathopoulos C, Iliopoulos I, Promponas V, et al. Low complexity regions in the proteins of prokaryotes perform important functional roles and are highly conserved. Nucleic Acids Research. 2019 Nov 4;47(19):9998–10009.

33. Radó-Trilla N, Albà Mm. Dissecting the role of low-complexity regions in the evolution of vertebrate proteins. BMC Evolutionary Biology. 2012 Aug 24;12(1):155.

34. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. J Chem Inf Comput Sci. 2003 Nov;43(6):1947–58.

35. Strobl C, Boulesteix A-L, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics. 2007 Dec;8(1):25.

36. Yang L, Gaut BS. Factors that Contribute to Variation in Evolutionary Rate among Arabidopsis Genes. Molecular Biology and Evolution. 2011 Aug 1;28(8):2359–69.

37. Gorman B. mltools: Machine Learning Tools [Internet]. 2018. Available from: https://CRAN.R-project.org/package=mltools

38. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal. 2006;Complex Systems:1695.

39. Stekhoven DJ, Buhlmann P. MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012 Jan 1;28(1):112–8.

40. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2020. Available from: https://www.R-project.org/

41. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the tidyverse. Journal of Open Source Software. 2019;4(43):1686.

42. Mevik B-H, Wehrens R, Liland KH. pls: Partial Least Squares and Principal Component Regression [Internet]. 2019. Available from: https://CRAN.R-project.org/package=pls

43. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18–22.