

1 **SHORT COMMUNICATIONS**

2  
3 **N-glycoproteins exhibit a positive expression level–**  
4 **evolutionary rate correlation**

5  
6 **Felix Feyertag<sup>1,\*,#</sup>, Patricia M Berninsone<sup>1</sup> and David Alvarez-Ponce<sup>1,\*</sup>**

7  
8 <sup>1</sup>Department of Biology, University of Nevada, Reno, Reno, Nevada, USA.

9  
10 **\*Corresponding authors:**

11 David Alvarez-Ponce, Department of Biology, University of Nevada, Reno. 1664 N. Virginia  
12 Street, Reno, NV 89557. Email: dap@unr.edu.

13 Felix Feyertag, Department of Biology, University of Nevada, Reno. 1664 N. Virginia Street,  
14 Reno, NV 89557. Email: ffeyertag@unr.edu.

15  
16  
17 **Short title:** N-glycoproteins exhibit positive E-R correlation.

18  
19 **Keywords:**

20 Rates of evolution,  $d_N/d_S$ , E–R anticorrelation, N-linked glycosylation.

21  
22  
23 **Acknowledgements:**

24 The authors are grateful to Xiaowei Jiang for helpful discussion. This work was supported by a  
25 grant from the National Science Foundation (MCB 1818288), funds from the University of  
26 Nevada, Reno, and by Pilot Grants from Nevada INBRE (P20GM103440) and the Smooth Muscle  
27 Plasticity COBRE from the University of Nevada, Reno (5P30GM110767-04), both funded by the  
28 National Institute of General Medical Sciences (National Institutes of Health), awarded to DAP.

29  
30  

---

# Present address: Target Discovery Institute, Nuffield Department of Medicine, University of  
Oxford, Oxford OX3 7FZ, United Kingdom.

## Abstract

The different proteins of any proteome evolve at enormously different rates. One of the primary factors influencing rates of protein evolution is expression level, with highly expressed proteins tending to evolve at slow rates. This phenomenon, known as the expression level–evolutionary rate (E–R) anticorrelation, has been attributed to the abundance-dependent deleterious effects of misfolding or misinteraction. We have recently shown that secreted proteins either lack an E–R anticorrelation or exhibit a significantly reduced E–R anticorrelation. This effect may be due to the strict quality control to which secreted proteins are subject in the endoplasmic reticulum (which is expected to reduce the rate of misfolding and its deleterious effects) or to their extracellular location (expected to reduce the rate of misinteraction and its deleterious effects). Among secreted proteins, N-glycosylated ones are under particularly strong quality control. Here we investigate how N-linked glycosylation affect the E–R anticorrelation. Strikingly, we observe a *positive* E–R correlation among N-glycosylated proteins. I.e., N-glycoproteins that are highly expressed evolve at faster rates than lowly expressed N-glycoproteins, in contrast to what is observed among intracellular proteins.

## Introduction

A large disparity between the evolutionary rates of different proteins is observed within all studied proteomes. While selective constraints cause some proteins to remain virtually unaltered over long evolutionary periods, other proteins quickly accumulate amino acid changes that may lead to change in function (Zuckerandl and Pauling 1965; Dickerson 1971; Li, et al. 1985). One of the key questions in Evolutionary Biology is to explain the fundamental causes for these differences in rates of protein evolution. Gene expression level appears to be a major factor affecting rates of protein evolution, with highly expressed proteins tending to evolve slowly (Pál, et al. 2001; Drummond, et al. 2005; Drummond, et al. 2006), a trend known as the expression level – evolutionary rate (E–R) anticorrelation.

Even though a number of hypotheses have been proposed to explain the E–R anticorrelation, its reasons remain largely unclear (for review, see Herbeck and Wall 2005; Pál, et al. 2006; Rocha 2006; Alvarez-Ponce 2014; Zhang and Yang 2015). The translational robustness hypothesis postulates that highly expressed proteins are under strong purifying selection to maintain their ability to fold into their native conformation, even in the presence of translation errors, as the costs of misfolding (refolding, degradation, resynthesis, aggregation and cytotoxicity) are abundance-dependent (Drummond, et al. 2005). A second hypothesis, the protein misfolding avoidance hypothesis, takes into account not only translation error-induced misfolding, but also misfolding in the absence of translation errors (Yang, et al. 2010). The protein misinteraction avoidance hypothesis postulates that highly expressed proteins are under strong selection to avoid unspecific interactions with other molecules, thereby reducing disruption of molecular pathways among other deleterious effects (Yang, et al. 2012).

We have recently found that the E–R anticorrelation is markedly reduced or even non-existent among secreted proteins (Feyertag, et al. 2017). This may be attributed to the strong quality control that takes place along the secretory pathway, which is expected to reduce the likelihood of protein misfolding and/or its deleterious effects, or to the spatial confinement of secreted proteins (to the endoplasmic reticulum, the Golgi apparatus and eventually the extracellular matrix), which is expected to reduce the likelihood of misinteraction and/or its deleterious effects.

A subset of secreted proteins undergo N-linked glycosylation, a post-translational modification that makes them subject to additional layers of quality control. In the lumen of the endoplasmic reticulum, the enzyme oligosaccharyltransferase transfers oligosaccharides containing terminal glucose residues to selected asparagine residues that are part of the consensus sequence N-X-S/T. The lectins calnexin and calreticulin bind such oligosaccharides, thereby retaining N-glycosylated proteins in the endoplasmic reticulum. During protein maturation, glucose residues are trimmed, freeing N-glycoproteins from calnexin and calreticulin and allowing them to move to the Golgi. However, if the “overseer” enzyme UDP-glucose:glycoprotein glucosyltransferase recognizes an N-glycoprotein that is not properly folded, it catalyzes the reglucosylation of their oligosaccharides, and calnexin and calreticulin can bind them again. As a result, N-glycosylated proteins only leave the endoplasmic reticulum once they are properly folded (for review, see Ellgaard and Frickel 2003; D'Alessio, et al. 2010). In addition, calnexin and calreticulin act as chaperones (for review, see Williams 2006). Therefore, the rates of misfolding and the fitness costs associated to misfolding are expected to be reduced among N-glycoproteins, which led us to hypothesize that N-glycoproteins should be particularly free from the E–R anticorrelation..

## Results and Discussion

For each human protein-coding gene, we identified its most likely ortholog in the mouse genome (orthologs were found for 16,581 genes) and estimated the rate of protein evolution from the nonsynonymous to synonymous divergence ratio ( $d_N/d_S$ ). We also retrieved protein abundance data (for 20 human tissues, as well as integrated values for the entire body) from the PaxDB database (Wang, et al. 2015), and mRNA expression data (for 32 human tissues), from the HumanAtlas database (Uhlen, et al. 2015). Proteins were classified as N-glycoproteins if N-linked glycosylated sites were available in the GlycoProtDB database, which includes experimentally determined N-linked glycosylation sites for proteins expressed in ascites fluids, ovary, stomach and serum. By combining these datasets, 460 proteins were classified as N-glycoproteins while the remaining 16,121 were deemed to lack N-glycosylation (henceforth referred to as “non-N-glycoproteins”).

In line with previous studies (Pál, et al. 2001; Drummond, et al. 2005; Drummond, et al. 2006; Feyertag, et al. 2017), we observed a strong negative correlation between rates of protein evolution

and whole-body protein abundances when analyzing all human proteins ( $\rho = -0.2485$ ,  $p = 8 \times 10^{-218}$ ,  $n = 15,571$ ). We next repeated our analyses for non-N-glycosylated proteins and for N-glycosylated proteins separately. Strikingly, while non-N-glycoproteins maintain the strong negative correlation ( $\rho = -0.2668$ ,  $p = 7 \times 10^{-245}$ ,  $n = 15,132$ ), N-glycoproteins exhibited a strong *positive* correlation between protein abundance levels and rates of evolution ( $\rho = 0.2897$ ,  $p = 6 \times 10^{-10}$ ,  $n = 439$ ; Figure 1). Similar observations were made when we investigated the correlation between rates of evolution and protein abundance levels in 20 human tissues separately: among non-N-glycoproteins, the correlation was significantly negative in 19 of the tissues (the only exception being plasma, for which the correlation was weakly positive), while among N-glycoproteins the correlations were always positive, significantly in 18 of the tissues (Figure 2, Table S1).

Similar, albeit less pronounced, differences were observed when we considered the correlation between evolutionary rates and mean mRNA abundance levels (averaged across 32 human tissues): among non-N-glycoproteins, a strong negative correlation was observed ( $\rho = -0.2301$ ,  $p = 5.44 \times 10^{-189}$ ,  $n = 15,809$ ), while a non-significant positive correlation was observed among N-glycoproteins ( $\rho = 0.0505$ ,  $p = 0.2829$ ,  $n = 454$ ). When we analyzed mRNA abundance levels in 32 human tissues separately, the correlation was always significantly negative among non-N-glycoproteins, while among N-glycoproteins the correlation was higher than among N-glycoproteins in 29 out of 32 tissues, and significantly positive in one tissue (Table S1).

Sites predicted to be involved in N-linked glycosylation have been shown to appear in regions under positive selection and co-evolution (Jiang, et al. 2015; Jiang, et al. 2017). We thus considered the possibility that N-glycoproteins under positive selection might drive these observations (if N-glycoproteins under positive selection tended to be highly expressed). However, similar results were obtained when we removed proteins with signatures of positive selection from our analyses (Table S2).

As all N-glycoproteins follow the secretory pathway (i.e., they are either secreted or membrane proteins), we decided to analyze extracellular and membrane proteins separately, obtaining similar results in both cases (Tables S3 and S4). The correlation between evolutionary rates and whole

body protein abundance levels was significantly positive among extracellular N-glycoproteins ( $\rho = 0.3393$ ,  $p = 8.63 \times 10^{-10}$ ,  $n = 310$ ), and non-significantly positive among membrane N-glycoproteins ( $\rho = 0.0649$ ,  $p = 0.4866$ ,  $n = 117$ ).

We next considered whether these observations may be a by-product of confounding factors. Potentially confounding factors include those that (a) differ between N-glycoproteins and non-N-glycoproteins and (b) affect the E–R correlation. We evaluated the differences between N-glycoproteins and non-N-glycoproteins in terms of several potentially confounding factors, separately for extracellular and membrane proteins. Investigated factors included protein abundance, protein expression breadth (number of tissues in which the protein is found), mRNA expression, mRNA expression breadth (number of tissues in which a gene is expressed), protein length, codon adaptation index (CAI), and number of protein-protein interactions (PPIs). Among extracellular proteins, we observed that N-glycoproteins had significantly higher protein abundance (Wilcoxon test,  $Z = 16.36$ ,  $p = 3.71 \times 10^{-60}$ ), mRNA expression ( $Z = 13.02$ ,  $p < 10^{-300}$ ), protein expression breadth ( $Z = 14.98$ ,  $p = 9.28 \times 10^{-51}$ ), mRNA expression breadth ( $Z = 7.85$ ,  $p = 4.16 \times 10^{-15}$ ), protein length ( $Z = 9.13$ ,  $p = 6.79 \times 10^{-20}$ ), and number of PPIs ( $Z = 7.12$ ,  $p = 1.05 \times 10^{-12}$ ) (Table S5). The same observations were made among membrane proteins (Table S6).

For each of these factors, we performed a test to determine whether they may be impacting our results. For each N-glycoprotein in our dataset, we randomly selected a non-N-glycoprotein with a similar value of the factor under consideration and with the same subcellular location (extracellular space or membrane). As a result, the N-glycoproteins and the randomly selected non-N-glycoproteins exhibit a virtually identical distribution for the potentially confounding factor. This resulted in datasets ranging in size  $n = 303$ – $324$  for extracellular proteins and  $n = 116$ – $124$  for membrane proteins. Remarkably, among the selected subsets of secreted N-glycoproteins, the E–R correlation was usually negative, and never significantly positive (Table S7), implying that none of the studied factors is responsible for the positive E–R correlation observed among N-glycoproteins. Similar results were obtained when analyzing membrane proteins (Table S8).

We next considered the potential effect of gene duplication. N-glycosylated proteins are significantly more likely to have paralogs (86.6%) compared to non-N-glycosylated proteins

(69.88%, Fisher test  $p = 1.4 \times 10^{-16}$ ). Non-N-glycoproteins always had significant negative correlations between evolutionary rates and both mRNA expression and protein abundance among both singletons (protein abundance:  $\rho = -0.4040$ ,  $p = 5.38 \times 10^{-176}$ ; mRNA expression:  $\rho = -0.3825$ ,  $p = 9.37 \times 10^{-166}$ ) and duplicates (protein abundance:  $\rho = -0.2394$ ,  $p = 1.42 \times 10^{-138}$ ; mRNA expression:  $\rho = -0.1932$ ,  $p = 2.50 \times 10^{-93}$ ). Among N-glycoproteins, duplicates showed positive correlations (protein abundance:  $\rho = 0.2962$ ,  $p = 3.93 \times 10^{-9}$ ; mRNA expression:  $\rho = 0.0649$ ,  $p = 0.1994$ ), while singletons showed a significant positive correlation between protein abundance and evolutionary rate ( $\rho = 0.2895$ ,  $p = 0.0261$ ) and a non-significant negative correlation between mRNA expression and evolutionary rate ( $\rho = -0.0662$ ,  $p = 0.6125$ ).

Finally, we compared pairs of highly-expressed and lowly-expressed paralogous genes. For each gene family with more than one member encoding non-N-glycosylated proteins ( $n = 1527$  families), we compared the rates of evolution of the protein with the highest protein abundance with that of the lowest. In agreement with prior results (Drummond, et al. 2005), highly expressed proteins tended to evolve slower (median  $d_N/d_S$  for highly expressed proteins: 0.0827, median  $d_N/d_S$  for lowly expressed proteins: 0.1080; Wilcoxon signed rank test,  $p = 8 \times 10^{-22}$ ). In addition, the most highly expressed protein exhibited the lowest  $d_N/d_S$  value in 911 out of the 1527 families (59.7% of cases; binomial test,  $p = 4.5 \times 10^{-14}$ ). However, the opposite trend was observed when we repeated the analyses on N-glycosylated proteins: highly expressed proteins exhibit a higher median rate of evolution (median  $d_N/d_S$  for highly expressed proteins: 0.1295, median  $d_N/d_S$  for lowly expressed proteins: 0.1000; one-tailed Wilcoxon signed rank test:  $p = 0.0308$ ), and in 49 of the 81 gene families with more than one N-glycosylated protein the most highly expressed protein was also the fastest-evolving (60.5% of cases; one-tailed binomial test,  $p = 0.0374$ ).

In summary, we have shown that among proteins containing N-glycosylated sites, highly expressed proteins tend to evolve faster than lowly expressed ones. This observation is in stark contrast to the strong negative E–R correlation observed among intracellular proteins. Even though we observed a number of differences between N-glycoproteins and non-N-glycoproteins—in terms of mRNA and protein abundance and tissue breadth, protein length, and number of PPIs—, these differences do not seem to account for the positive E–R correlation observed in N-glycoproteins. The trend is not due to highly expressed N-glycoproteins being under positive selection, either. As

stated above, our initial hypothesis was that the strong quality control N-glycoproteins are subject to should attenuate or remove the E–R anticorrelation typical to other proteins. However, a positive E–R correlation was not expected and we do not know of any mechanism promoting such a trend. It is possible that one or multiple mechanisms, yet to be discovered, promote such a positive E–R correlation in all proteins, but that these mechanisms are silenced in non-N-glycoproteins due to weak quality control.

## **Material and methods**

Human protein coding genes were downloaded from the ENSEMBL database (release 62; Cunningham, et al. 2015). For each protein the longest protein was chosen for analysis. The most likely ortholog in the mouse genome was identified using a best reciprocal BLAST hit approach (using BLASTP and an  $E$ -value cut-off of  $10^{-10}$ ). Pairs of orthologous sequences were aligned using ProbCons 1.12 (Do, et al. 2005), and the resulting alignments were used to guide the alignment of the corresponding coding sequences (CDSs). Signal peptides and proteolyzed domains were identified using the UniProt database (The UniProt Consortium, 2017) and removed from our analyses, given their fast rates of evolution. PAML (model M0; Yang 2007) was used to estimate the  $d_N/d_S$  ratio. Expanded alignments with 10 species (human, chimpanzee, gorilla, orangutan, macaque, mouse, rat, cow, dog, and opossum) were used to detect positive selection using PAML (M7 vs. M8 test; Yang 2000). Alignments were filtered as in Chakraborty and Alvarez-Ponce (2016) before positive selection analyses. Extracellular and membrane location data were obtained from the MetazSecKB database (Meinken, et al. 2015). N-linked glycosylation data was retrieved from the GlycoProtDB database (version 2016-12-05) (Kaji, et al. 2012). Protein abundances for the entire body and for 20 individual tissues/organs were obtained from the PaxDB database, version 4 (Wang, et al. 2015). Messenger RNA abundance data for 32 tissues/organs were obtained from the HumanAtlas database, version 16.1 (Uhlen, et al. 2015). Codon Adaptation Index values were computed using the EMBOSS package (Rice, et al. 2000). Protein–protein interaction data were retrieved from the BioGRID database, version 3.4.133 (Chatr-Aryamontri, et al. 2015). Paralogous proteins were identified using ENSEMBL BioMart (Guberman, et al. 2011). Nonparametric statistical analyses, including Wilcoxon and Spearman rank correlations tests, were conducted using R (<https://www.r-project.org>) and JMP (SAS Institute, Inc., NC, USA).



## References

- Alvarez-Ponce D. 2014. Why proteins evolve at different rates: The determinants of proteins' rates of evolution. In: MA F, editor. *Natural Selection: Methods and Applications*. London: CRC Press (Taylor & Francis). p. 126-178.
- Chakraborty S, Alvarez-Ponce D. 2016. Positive Selection and Centrality in the Yeast and Fly Protein-Protein Interaction Networks. *Biomed Res Int* 2016:4658506.
- Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L, et al. 2015. The BioGRID interaction database: 2015 update. *Nucleic Acids Res* 43:D470-478.
- Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2015. Ensembl 2015. *Nucleic Acids Res* 43:D662–D669.
- D'Alessio C, Caramelo JJ, Parodi AJ. 2010. UDP-Glc:glycoprotein glucosyltransferase-glucosidase II, the ying-yang of the ER quality control. *Semin Cell Dev Biol* 21:491-499.
- Dickerson RE. 1971. The structures of cytochrome c and the rates of molecular evolution. *J Mol Evol* 1:26-45.
- Do CB, Mahabhashyam MS, Brudno M, Batzoglou S. 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res* 15:330-340.
- Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A* 102:14338-14343.
- Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol* 23:327-337.
- Ellgaard L, Frickel EM. 2003. Calnexin, calreticulin, and ERp57: teammates in glycoprotein folding. *Cell Biochem Biophys* 39:223-247.
- Feyertag F, Berninsone PM, Alvarez-Ponce D. 2017. Secreted Proteins Defy the Expression Level-Evolutionary Rate Anticorrelation. *Mol Biol Evol* 34:692-706.
- Guberman JM, Ai J, Arnaiz O, Baran J, Blake A, Baldock R, Chelala C, Croft D, Cros A, Cutts RJ, et al. 2011. BioMart Central Portal: an open database network for the biological community. *Database (Oxford)* 2011:bar041.
- Herbeck JT, Wall DP. 2005. Converging on a general model of protein evolution. *Trends Biotechnol* 23:485-487.
- Jiang X, Feyertag F, Meehan CJ, McCormack GP, Travers SA, Craig C, Westby M, Lewis M, Robertson DL. 2015. Characterizing the Diverse Mutational Pathways Associated with R5-Tropic Maraviroc Resistance: HIV-1 That Uses the Drug-Bound CCR5 Coreceptor. *J Virol* 89:11457-11472.
- Jiang X, Feyertag F, Robertson DL. 2017. Protein structural disorder of the envelope V3 loop contributes to the switch in human immunodeficiency virus type 1 cell tropism. *PLoS One* 12:e0185790.
- Kaji H, Shikanai T, Sasaki-Sawa A, Wen H, Fujita M, Suzuki Y, Sugahara D, Sawaki H, Yamauchi Y, Shinkawa T, et al. 2012. Large-scale identification of N-glycosylated proteins of mouse tissues and construction of a glycoprotein database, GlycoProtDB. *J Proteome Res* 11:4553-4566.

- Li WH, Wu CI, Luo CC. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol* 2:150-174.
- Meinken J, Walker G, Cooper CR, Min XJ. 2015. MetazSecKB: the human and animal secretome and subcellular proteome knowledgebase. *Database (Oxford)* 2015:bav077.
- Pál C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. *Genetics* 158:927-931.
- Pál C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. *Nat Rev Genet* 7:337-348.
- Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16:276-277.
- Rocha EP. 2006. The quest for the universals of protein evolution. *Trends Genet* 22:412-416.
- The UniProt Consortium. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 45: D158-D169
- Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, et al. 2015. Proteomics. Tissue-based map of the human proteome. *Science* 347:1260419.
- Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. 2015. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* 15:3163-3168.
- Williams DB. 2006. Beyond lectins: the calnexin/calreticulin chaperone system of the endoplasmic reticulum. *J Cell Sci* 119:615-623.
- Yang JR, Liao BY, Zhuang SM, Zhang J. 2012. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proc Natl Acad Sci U S A* 109:E831-840.
- Yang JR, Zhuang SM, Zhang J. 2010. Impact of translational error-induced and error-free misfolding on the rate of protein evolution. *Mol Syst Biol* 6:421.
- Yang Z. 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *J Mol Evol* 51:423-432.
- Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586-1591.
- Zhang J, Yang JR. 2015. Determinants of the rate of protein sequence evolution. *Nat Rev Genet* 16:409-420.
- Zuckermandl E, Pauling L. 1965. Molecules as documents of evolutionary history. *J Theor Biol* 8:357-366.

**Figure legends**

**Fig. 1. Correlation between protein abundance and rates of evolution for non-N-glycoproteins and N-glycoproteins.** A) All proteins; B) membrane proteins; C) extracellular proteins. Lines represent regression lines. Spearman correlation coefficients are shown.  $\omega$ , nonsynonymous to synonymous divergence ratio ( $d_N/d_S$ ); PPM, parts per million; \*,  $p < 0.05$ ; \*\*\*,  $p < 0.001$ .

**Fig. 2. Spearman correlations between rates of protein evolution and protein abundances in 20 human tissues, for non-N-glycoproteins and N-glycoproteins.** A) All proteins; B) membrane proteins; C) extracellular proteins. Tissues are listed from the lowest to the highest non-N-glycoprotein  $p$  value.