Protein Conformational Diversity Correlates with Evolutionary Rate

Diego Javier Zea,[1] Alexander Miguel Monzon,[1] Maria Silvina Fornasari,[1] Cristina Marino-Buslje,[2] and Gustavo Parisi*,[1]

[1]Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Bernal, Argentina

[2]Fundación Instituto Leloir, Ciudad Autónoma de Buenos Aires, Argentina

***Corresponding author:** E-mail: gusparisi@gmail.com.

Associate editor: James McInerney

Abstract

Native state of proteins is better represented by an ensemble of conformers in equilibrium than by only one structure. The extension of structural differences between conformers characterizes the conformational diversity of the protein. In this study, we found a negative correlation between conformational diversity and protein evolutionary rate. Conformational diversity was expressed as the maximum root mean square deviation (RMSD) between the available conformers in Conformational Diversity of Native State database. Evolutionary rate estimations were calculated using 16 different species compared with human sharing at least 700 orthologous proteins with known conformational diversity extension. The negative correlation found is independent of the protein expression level and comparable in magnitude and sign with the correlation between gene expression level and evolutionary rate. Our findings suggest that the structural constraints underlying protein dynamism, essential for protein function, could modulate protein divergence.

Key words: **conformational diversity, evolutionary rate, protein evolution.**

The study of protein evolutionary rates is a central issue to understand the mechanisms underlying protein molecular evolution. Protein evolutionary rates are generally estimated by the number of nonsynonymous nucleotide changes per site (d*N*) in the coding genes of orthologous proteins. Several factors have been associated to the modulation of the evolutionary rate such as amino acid composition (Tourasse and Li 2000), codon adaptation index (Rocha and Danchin 2004), functional importance of the protein (Wilson et al. 1977), expression level (Pal et al. 2006), number of protein interactions (Fraser et al. 2002), protein stability (Zeldovich et al. 2007), and protein length (Marais and Duret 2001) (for a review see Pal et al. [2006] and cites therein [Pal et al. 2006]). However, it was established that the gene expression level, measured in mRNA transcripts per cell, is the property showing one of the strongest, pervasive, and consistent correlation between genomic data and evolutionary rate (Drummond et al. 2005).

Most proteins require proper structural arrangement to be biologically active. The conservation of protein fold during evolution imposes constraints to sequence divergence modulating the site-specific substitution pattern of residues. Several studies have correlated structural constraints with evolutionary rates. One of the strongest signal found was that solvent-exposed residues evolve faster than those buried (Franzosa and Xia 2009) or that transmembrane regions in membrane proteins evolve slower than their extramembrane regions (Oberai et al. 2009). Other studies have found that neither secondary structure nor protein fold have strong correlation with evolutionary rate, attributing to protein structural constraints as much as 10% of the evolutionary rate in proteins (Bloom et al. 2006). Alternatively, the results by Wilke and Drummond (2010) indicated that structural constraints could play a major role modulating evolutionary rates lessening the influence of the biological function of the protein. The aforementioned studies were performed considering a single structure to describe the native state of proteins. However, it is well established that native state of proteins is better represented by an ensemble of different conformers in dynamical equilibrium (Tsai et al. 1999). The concept of conformational ensemble is a central key to explain essential properties of proteins such as (Boehr et al. 2006; del Sol et al. 2009; Hilser 2010; Ma and Nussinov 2010), enzyme and antibody promiscuity (James et al. 2003), signal transduction (Smock and Gierasch 2009), and protein–protein recognition (Yogurtcu et al. 2008). In this work, we have studied the influence of conformational diversity on protein evolutionary rate.

To study this relationship, we have used human proteins contained in the Conformational Diversity of Native State (CoDNaS) database (Monzon et al. submitted), which is a collection of redundant protein structures that can be taken as snapshots of protein dynamism (Zoete et al. 2002; Best et al. 2006). Human orthologous sequences were used to estimate evolutionary rates in 16 species as they share at least 700 orthologous proteins each (supplementary table S1, Supplementary Material online).

We found that the maximum RMSD100 between conformers shows a monotonic nonlinear correlation with d*N* with a mean Spearman's rank correlation coefficient (SCC) among the 16 species of −0.135 and a standard error (SE) of 0.007 (table 1). A similar result is found using d*N*/d*S* (mean over the

16 pairs of comparison −0.147 ± 0.008, see supplementary table S3, Supplementary Material online). All correlations were significant at the 0.05 level after correction by false discovery rate (FDR). This correlation is in the same order that the one between d*N* and expression level in our data set (mean SCC of −0.189, SE: 0.007) and the reported for human (−0.163, $P < 0.001$) (Drummond and Wilke 2008) (figs. 1 and 2). Our results suggest that increasing of structural differences between the conformers describing the native state impose more constraints on the protein sequence divergence reducing the evolutionary rate. This observation is related with the finding that the presence of conformational diversity modulates the sequence substitution pattern (Juritz 2013). We further demonstrated that this correlation does not depend on protein expression level (partial SCC between maximum RMSD100 and d*N* is −0.15 with SE: 0.01, for given expression level). Furthermore, the partial correlation analysis between d*N* and maximum RMSD100 for a given number of conformers per protein yielded similar results to the raw correlation (−0.147, SE: 0.007) showing no bias due to the number of conformers per protein. Finally, as different structural similarity measurements have been developed, we also found a similar correlation between d*N*, d*N*/d*S*, and TM score (Zhang and Skolnick 2004) (0.118 ± 0.007 and 0.133 ± 0.008, respectively).

Table 1. Total and Partial Spearman's Rank Correlation Coefficients in the 16 Pairs of Comparisons.

X	d*N*			
Y	RMSD100	RMSD100		Expression
Z		Conformers	Expression	
Ailuropoda melanoleuca	−0.14***	−0.16***	−0.17***	−0.19***
Anolis carolinensis	−0.14***	−0.13***	−0.13*	−0.15**
Bos taurus	−0.13***	−0.15***	−0.15**	−0.21***
Callithrix jacchus	−0.09**	−0.10**	−0.08**	−0.21**
Canis familiaris	−0.11***	−0.13***	−0.12**	−0.18***
Cavia porcellus	−0.13***	−0.13***	−0.14***	−0.17***
Gallus gallus	−0.16***	−0.17***	−0.13*(*)	−0.18***
Felis catus	−0.17***	−0.19***	−0.19**(*)	−0.16**
Equus caballus	−0.14***	−0.16***	−0.16**(*)	−0.22***
Loxodonta africana	−0.14***	−0.16***	−0.15**(*)	−0.21***
Macaca mulatta	−0.11***	−0.11**(*)	−0.14**	−0.18***
Monodelphis domestica	−0.11**	−0.13**	−0.12*(*)	−0.17**
Mus musculus	−0.13***	−0.14***	−0.14**	−0.25***
Pan troglodytes	−0.20***	−0.20***	−0.24***	−0.15***
Pongo pygmaeus abelii	−0.14***	−0.16***	−0.22***	−0.17***
Rattus norvegicus	−0.12***	−0.13***	−0.11*	−0.24***

NOTE.—RMSD100, the maximum normalized RMSD between conformers of a protein; expression, for expression level measured by the mRNA level of the protein. Significance levels in parentheses disappear after FDR correction.

*$P < 0.05$.

**$P < 0.01$.

***$P < 0.001$.

We found a significant negative relationship between the degree of conformational diversity of a protein and its evolutionary rate. Unfortunately, at the moment, available data allowed the study using only human proteins with enough statistical confidence. Our results suggest a key role of structural constraints maintaining the conformational ensemble of the native state of the protein. It is interesting to note that as protein function is close related with protein dynamism, our results could also suggest the indirect influence of protein function on the rate of evolution.

Materials and Methods

Proteins with different degrees of conformational diversity were obtained from CoDNaS database (Monzon et al. submitted). CoDNaS is a redundant collection of crystallographic structures for the same protein that could be taken as a collection of different conformers. It includes a total of 70,467 PDB structures, representing a set of 9,398 monomeric proteins of the PDB database. The degree of conformational

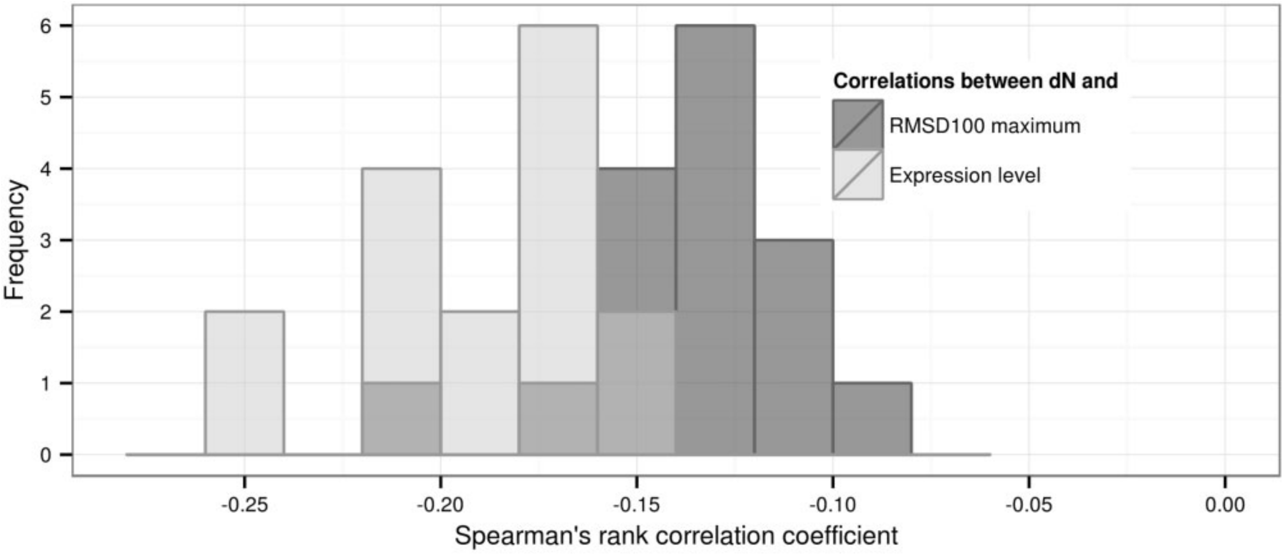

FIG. 1. Histograms for distribution of Spearman's rank correlation coefficients between d*N* and RMSD100 maximum (dark gray) and expression level (light gray) in the 16 species studied.

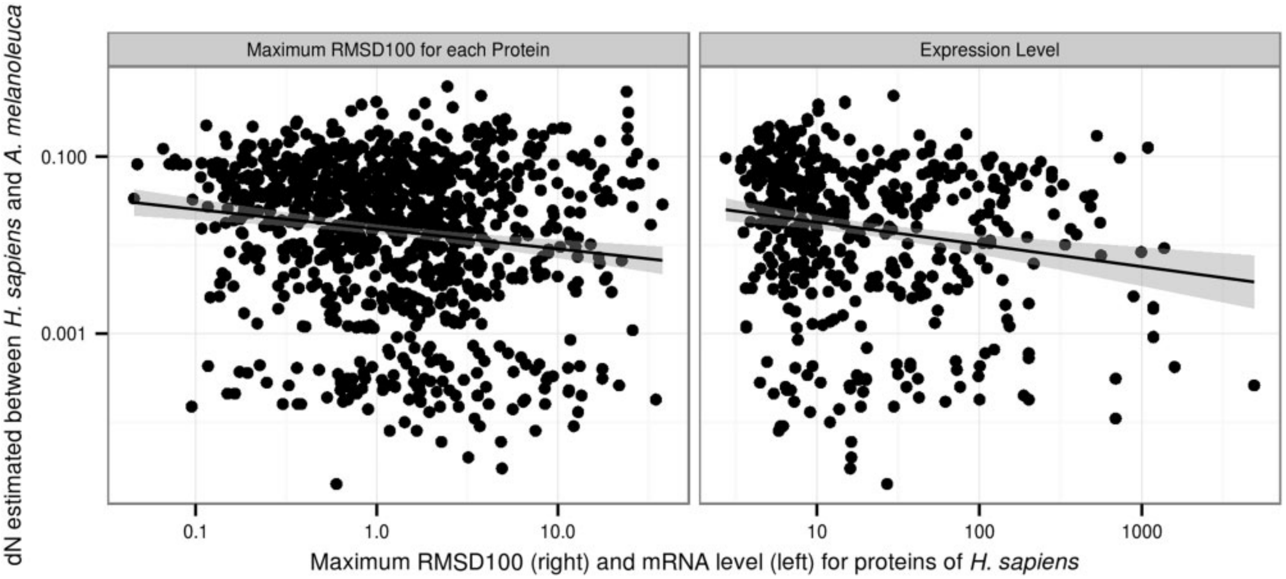

Fig. 2. Relationship between maximum RMSD100 (right) and mRNA level (left) with d*N* is linear in logarithmic (on base 10) scale. As an example, representations correspond to the comparison between human and giant panda (950 and 515 proteins comparison for conformational diversity and expression level, respectively). The linear correlation between conformational diversity and d*N* in logarithmic scale gives a Pearson correlation coefficient of −0.143 with a *P* value of 9×10^{-6} and for expression level −0.172 and *P* value of 8.31×10^{-5}.

diversity, measured as the maximum RMSD between available conformers was normalized to RMSD100 for all proteins with more than 40 residues (Carugo and Pongor 2001). Each protein entry was linked using its Uniprot Accessions with OMA database to obtain the corresponding orthologs. Only 1:1 orthologs were selected (Altenhoff et al. 2011). We used codeml from PAML 4.5 for pairwise d*N* and d*N*/d*S* estimations using model 0 for codons (Yang 2007). Protein-aligned sequences were taken as templates to get codon alignments using the program pal2nal (v14) (Suyama et al. 2006). Because of the low SCC between d*N* and RMSD100, we used those proteins that share more than 700 orthologs to *Homo Sapiens* proteins to achieve a statistical power close to 80% at 5% of significance. Thus, SCCs were estimated for 14,301 and 7,706 pairs of orthologous proteins coming from 16 species for RMSD100 and expression level, respectively (supplementary table S1, Supplementary Material online). The data set includes 1,094 human proteins with 5,592 PDB entries in CoDNaS (supplementary table S2, Supplementary Material online). We considered all the different structures for the same protein as putative conformers except those with annotated mutations. Human mRNA levels were obtained from U133A/GNF1H array signals from bioGPS (Su et al. 2004), averaged as the geometric mean signal across all normal adult tissues. All statistical analyses were performed with R (http://www.R-project.org/), and the “ppcor package” was used to calculate partial correlations. *P* values were corrected by FDR in all cases (Benjamini and Hochberg 1995).

Supplementary Material

Supplementary tables S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by grants PIP CONICET (112-200801-02849) and UNQ (1004/11) to G.P. and M.S.F. D.Z. has a Type II fellowship from CONICET.

References

Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C. 2011. OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res.* 39:D289–D294.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B (Methodol).* 57:289–300.

Best RB, Lindorff-Larsen K, DePristo MA, Vendruscolo M. 2006. Relation between native ensembles and experimental structures of proteins. *Proc Natl Acad Sci U S A.* 103:10901–10906.

Bloom JD, Drummond DA, Arnold FH, Wilke CO. 2006. Structural determinants of the rate of protein evolution in yeast. *Mol Biol Evol.* 23: 1751–1761.

Boehr DD, McElheny D, Dyson HJ, Wright PE. 2006. The dynamic energy landscape of dihydrofolate reductase catalysis. *Science* 313: 1638–1642.

Carugo O, Pongor S. 2001. A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Sci.* 10: 1470–1473.

del Sol A, Tsai CJ, Ma B, Nussinov R. 2009. The origin of allosteric functional modulation: multiple pre-existing pathways. *Structure* 17:1042–1050.

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, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341–352.

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

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. *Science* 296: 750–752.

Hilser VJ. 2010. Biochemistry. An ensemble view of allostery. *Science* 327: 653–654.

James LC, Roversi P, Tawfik DS. 2003. Antibody multispecificity mediated by conformational diversity. *Science* 299:1362–1367.

Juritz E, Palopoli N, Fornasari S, Fernandez Alberti S, Parisi G. 2013. Protein conformational diversity modulates sequence divergence. *Mol Biol Evol.* 30(1):79–87.

Ma B, Nussinov R. 2010. Enzyme dynamics point to stepwise conformational selection in catalysis. *Curr Opin Chem Biol.* 14: 652–659.

Marais G, Duret L. 2001. Synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*. *J Mol Evol.* 52: 275–280.

Oberai A, Joh NH, Pettit FK, Bowie JU. 2009. Structural imperatives impose diverse evolutionary constraints on helical membrane proteins. *Proc Natl Acad Sci U S A.* 106: 17747–17750.

Pal C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. *Nat Rev Genet.* 7:337–348.

Rocha EP, Danchin A. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol Biol Evol.* 21: 108–116.

Smock RG, Gierasch LM. 2009. Sending signals dynamically. *Science* 324: 198–203.

Su AI, Wiltshire T, Batalov S, et al. (13 co-authors). 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A.* 101:6062–6067.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34:W609–W612.

Tourasse NJ, Li WH. 2000. Selective constraints, amino acid composition, and the rate of protein evolution. *Mol Biol Evol.* 17:656–664.

Tsai CJ, Ma B, Nussinov R. 1999. Folding and binding cascades: shifts in energy landscapes. *Proc Natl Acad Sci U S A.* 96:9970–9972.

Wilke CO, Drummond DA. 2010. Signatures of protein biophysics in coding sequence evolution. *Curr Opin Struct Biol.* 20:385–389.

Wilson AC, Carlson SS, White TJ. 1977. Biochemical evolution. *Annu Rev Biochem.* 46:573–639.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.

Yogurtcu ON, Erdemli SB, Nussinov R, Turkay M, Keskin O. 2008. Restricted mobility of conserved residues in protein-protein interfaces in molecular simulations. *Biophys J.* 94:3475–3485.

Zeldovich KB, Chen P, Shakhnovich EI. 2007. Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proc Natl Acad Sci U S A.* 104:16152–16157.

Zhang Y, Skolnick J. 2004. Scoring function for automated assessment of protein structure template quality. *Proteins* 57:702–710.

Zoete V, Michielin O, Karplus M. 2002. Relation between sequence and structure of HIV-1 protease inhibitor complexes: a model system for the analysis of protein flexibility. *J Mol Biol.* 315:21–52.