Journal of Molecular Evolution (2021) 89:257–268
https://doi.org/10.1007/s00239-021-10003-7

ORIGINAL ARTICLE

Estimating the Influence of Physicochemical and Biochemical Property Indexes on Selection for Amino Acids Usage in Eukaryotic Cells

Giovani B. Fogalli[1] · Sergio R. P. Line[1]

Received: 29 May 2020 / Accepted: 10 March 2021 / Published online: 24 March 2021

Abstract

Proteins can evolve by accumulating changes on amino acid sequences. These changes are mainly caused by missense mutations on its DNA coding sequences. Mutations with neutral or positive effects on fitness can be maintained while deleterious mutations tend to be eliminated by natural selection. Amino acid changes are influenced by the biophysical, chemical, and biological properties of amino acids. There is a multiplicity of amino acid properties that can influence the function and expression of proteins. Amino acid properties can be expressed into numerical indexes, which can help to predict functional and structural aspects of proteins and allow statistical inferences of selection pressure on amino acid usage. The accuracy of these analyses may be compromised by the existence of several numerical indexes that measure the same amino acid property, and the lack of objective parameters to determine the most accurate and biologically relevant index. In the present study, the *gradient consistency test* was used in order to estimate the magnitude of directional selection imparted by amino acid biochemical and biophysical properties on protein evolution.

Keywords Gradient consistency test · Amino acid properties · Amino acid usage · Protein evolution

Introduction

A protein can evolve by accumulating changes on its amino acid sequences. These changes are mainly caused by missense mutations on its DNA coding sequences, where a single nucleotide change may result in a codon that codes for a different amino acid. Amino acid gain and loss in protein evolution follow nearly neutral theoretical expectations (Hurst et al. 2006; Jordan et al. 2005; McDonald et al. 2006). Mutations with neutral or positive effect on fitness can be maintained while deleterious mutations tend to be eliminated by natural selection. Protein evolution seems to be mainly constrained by selection against misfolding and misinteractions with other molecules (Echave and Wilke 2017). Although most missense mutations seem to have an effect on protein stability, mutations that affect function seem to be restricted to few sites within protein sequences (DePristo et al. 2005). An analysis of 4000 amino acid substitutions in the lac repressor XV showed that most replacements do not interfere with phenotype (Suckow et al. 1996). Likewise, it has been predicted that most amino acid polymorphisms in the human proteome are functionally neutral (Ng and Henikoff 2003; Choi et al. 2012). In fact, both adaptive and non-adaptive evolution are mainly caused by substitutions between similar amino acids (Bergman and Eyre-Walker 2019).

The selection of variants in coding sequences may be influenced by the biophysical, chemical, and biological properties of amino acids (Rudnicki et al. 2014). Estimating the magnitude of directional selection imparted by a specific property of amino acids is not a simple task. There is a multiplicity of amino acid properties that can influence the function and expression of proteins. It is possible, however, that in some proteins the frequency of amino acids are mainly influenced by a single property, while in other proteins the frequency of amino acids may be influenced by multiple and diverse characteristics (Suckow et al. 1996; Wei et al. 2010). Based on theoretical and experimental approaches, diverse amino acid characteristics and properties have been quantified into numerical indexes (Kawashima et al. 2008). Quantitative

Handling Editor: **Ashley Teufel**.

✉ Sergio R. P. Line
serglin@unicamp.br

[1] Department of Biosciences, Piracicaba Dental School, University of Campinas, Campinas, Brazil

indexes allow a more precise statistical inference on the estimation of selection forces acting on protein function and evolution.

The accuracy of these analyses, however, may be compromised by the fact that there are usually several numerical indexes that measure the same amino acid property, and the lack of objective parameters to determine the most accurate and biologically relevant index. In general terms, the magnitude of directional selection on a characteristic can be estimated by the coefficient of the regression of phenotypic values of traits and fitness (Lande and Arnold 1983). Accordingly, the strength of natural selection forces acting on a specific intrinsic characteristic to influence the frequency of amino acids would be estimated by calculating the correlation between quantitative scores that estimate the strength of the characteristic on each of the 20 amino acids and their frequency on proteins. This approach will produce a different estimation for each amino acid. Swire (2007) has integrated this reductionist approach by developing the *gradient consistency test,* which can detect signatures of selection independently from the analysis of proteins expression levels or constraint data. It relies on the estimation of interprotein gradients in amino acid usage to detect the signature of selective evolution caused by a specific amino acid property. This method was originally developed to analyze selection on amino acid biosynthetic cost. It was demonstrated that the frequency of specific low biosynthetic cost amino acids tends to increase as the frequency of other low-cost amino acids increase. The rate of increase for each amino acid is given by the slope of the regression line obtained between the frequency of the focal amino acid (i.e., amino acid used for frequency calculation, see *y*-axis of Figs. S1–S5 of Online Resource 3) and the mean per residue of sites not occupied by the focal amino acid of a specific index. The *gradient consistency test* score is obtained by the correlation between the absolute shift in usage (slope) for each amino acid and the respective index that measures the strength (ex: biosynthetic cost) of the characteristic.

Since the *gradient consistency test* scores are dependent on several analyses (mean per residue index, slope obtained with ordinary least square regression, and the final score is obtained with Pearson correlation coefficient), we have analyzed and compared the effect of alternative procedures (median per residue index, slope obtained with robust regression, and final score obtained with Kendall correlation coefficient) on its performance. The *gradient consistency test* (Swire 2007) was applied on 555 indexes that measure amino acid properties. These analyses allowed identification of indexes significantly associated with the amino acid composition of proteins and detection of signatures of selective evolution caused by diverse amino acid properties.

Materials and Methods

Proteins

Amino acid sequences of *Bos taurus (cow, class Mammalia), Caenorhabditis elegans (class Secernentea), Homo sapiens (human, class Mammalia), Loxodonta africana (elephant, class Mammalia), Mus musculus (mouse, class Mammalia), and Saccharomyces cerevisiae* (class Saccharomycetes) were downloaded from the Ensembl/Biomart database (https://www.ensembl.org/biomart/martview/). These taxa include 4 vertebrate, 1 invertebrate multicellular, and 1 unicellular organisms. The four mammalian species included (Mus, Homo, Bos, Loxodonta) present a wide range of body mass. The reason for this was that metabolic rate has been shown to be related with body mass (White and Seymour 2003) and also protein evolution (Gillooly et al. 2007). In case there were more than one peptide for the same protein only the largest was used for further analysis.

Amino Acid Indexes

Numerical values of 544 indexes representing various physicochemical and biochemical properties of amino acids and pairs of amino acids were downloaded from the AAindex database (Kawashima et al. 2008, https://wwwgenomejp/aaindex/); Eleven indexes related to amino acids biosynthetic cost and fitness were also included in the analysis, totaling 555 indexes (Online Resource 1).

The Gradient Consistency Test (Adapted from Swire, 2007)

This test assumes that if a specific amino acid characteristic exerts selective pressure on amino acid choice then the low-score amino acids are expected to be found in proteins mostly composed of other low-score amino acids (Swire 2007). The calculation of the *gradient consistency test* score can be divided in 3 steps:

Step 1. Raw shift in usage. The rate of increase for each amino acid is given by the slope of the regression line between the frequency of the focal amino acid and the mean per residue of the indexes of sites not occupied by the focal amino acid of each protein (unbiased mean). The exclusion of the focal amino acid would give a mean cost per residue that is independent of the occurrence of this amino acid.

Step 2. Slope standardization. It is obtained by the following calculation:

Standardized slope = slope * *abs*(*y*/*x*) where *y* is the mean per residue of the relevant 19 amino acid types of the indexes of all proteins and *x* is the mean percentage usage of the focal amino acid in all proteins. The *abs* refers to the absolute value as many indexes have both positive and negative values.

Step 3. Obtaining the variation in gradients. The *gradient consistency test* score is obtained by the correlation between the absolute shift in usage (slope) for each amino acid and the respective index that measures the amino acid strength for the characteristic.

Since the gradient consistency test scores are dependent on several analyses (mean per residue index, slope obtained with ordinary least square regression, and final score obtained with Pearson correlation coefficient), we have analyzed and compared the effect of alternative procedures on the *gradient consistency test* scores:

1. Comparing mean and median per residue index. *The m*ean values used by Swire (2007) can be influenced by outliers, therefore, extremely low or high index values could potentially produce false positives or negative gradient scores.
2. Comparing the slope regression method (least square versus robust regression). The ordinary least square method for linear regression (LSQ), where slope can be obtained using the *lm(y ~ x)$coefficients[[2]]* function of The R Project for Statistical Computing (R Core Team 2013). This method is influenced by outliers, which can influence the slope values leading to non-representative slope values. The robust regression using the M method (ROB) is more robust to outliers. The slope was obtained using the *rlm(y ~ x, method = "M")$coefficients[[2]]* function of the MASS package of R (Venables and Ripley 2002).
3. The *gradient consistency test* score (Step 3) for cost synthesis was originally obtained with linear Pearson correlation. However, it is possible that other indexes may produce a non-linear correlation pattern. In these cases, the final scores would be better estimated by a non-linear correlation test. Therefore, besides Pearson's we have also calculated Kendall's correlation coefficient.

Codes and Statistical Analysis

The statistical analysis and figures were done using R (version 4.0.2) and genome processing was done using Ruby programming language (version 2.5.1, Line et al. 2014). Codes were run on Ubuntu version 18.04.5. The *p* values for each taxa were filtered by false discovery rate (FDR) using the *qvalue* package of the R statistical software (Storey et al. 2020). The Ruby and R files used for obtaining the gradient consistency test score and figures are on Online Resource 2.

Results

Comparing the Use of Mean or Median Per Residue Index

In order to obtain slope values for each amino acid, Swire (2007) used the unbiased mean (excluding the focal amino acid) per residue synthesis cost. Since the mean can be influenced by outliers we also tested the more robust median values. Therefore, the effect of unbiased median and mean per residue index (555 indexes) on the *gradient consistency test* scores were tested in three distinct species, representing vertebrate (*H. sapiens,* Fig. 1), invertebrate (*C. elegans,* Fig. S8 Online Resource 3), and unicellular (*S. cerevisiae,* Fig. S9 Online Resource 3) organisms. Our analyses showed that the gradient consistent scores obtained with mean and median per residue indexes were highly correlated in the 3 taxa analyzed. Pearson's r ranged from 0.85 to 0.91. Due to the high correlation scores, all further analyses were performed with the mean per residue index following the original protocol of Swire (2007).

Comparing Least Squares (LSQ) and Robust Regression (ROB) Line Fitting Slopes on the *Gradient Consistency Test* Scores

H. sapiens proteins were used to obtain the slopes of the best fit lines using the LSQ and ROB regression methods. Our analyses showed that the gradient consistent scores obtained with LSQ and ROB were fairly similar (Online Resource 4). When the *gradient consistency test* scores obtained with ROB and LSQ were used, respectively, as dependent and independent variables in least square linear regression, the coefficient of determination (R2) was 0.98 when the final score (Step 3) was obtained with Pearson and 0.97 when Step 3 was obtained with Kendall's correlation (Fig. 2a, b). In both cases, the slope of lines was 1.03. The scatter plots between the frequency of each amino acid and the mean per residue of the indexes of sites not occupied by the focal amino acid of each protein (unbiased mean), as well as the slopes obtained with LSQ and ROB for a hydrophobicity index (NADH010103) is shown in Figs. S1–S5 (Online Resource 3). It can be noted that slopes obtained with LSQ and ROB are fairly similar. The scatter plots between amino acid index (*x*-axis) and the standardized slope obtained with LSQ and ROB are shown, respectively, in figures S6 and S7 (Online Resource 3).

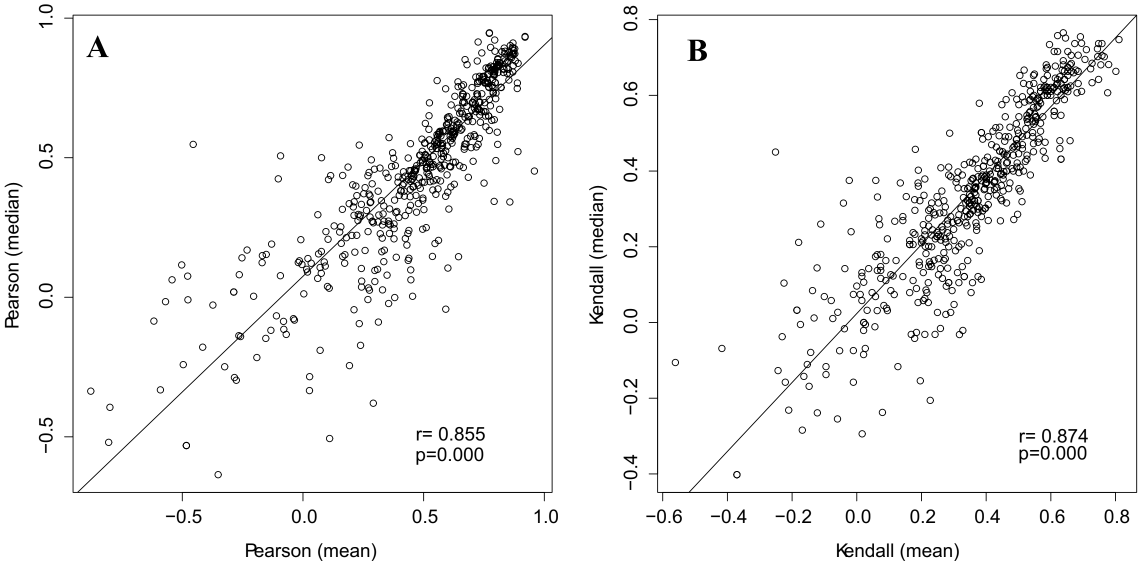

Fig. 1 Comparison of the use of mean or median per residue index to obtain slope values for each amino acid and its influence on the *gradient consistency test* scores. **a** G*radient consistency test* scores obtained with Pearson's correlation. **b** *Gradient consistency test* scores obtained with Kendall's correlation. Mean values are shown on x-axis and median values are shown on *y*-axis. Each point represents the *gradient consistency test* score of an amino acid property index. Analyses were performed in *H. sapiens* proteome. Note that the *scores* obtained with mean and median per residue indexes were highly correlated

Comparing the Use of Pearson and Kendall Correlation Methods to Obtain the Final *Gradient Consistency Test* Score

The *gradient consistency test* score for cost biosynthesis was originally obtained with linear Pearson correlation (Swire 2007). However, it is possible that other indexes may produce a non-linear correlation which would produce low scores. It is also possible that linear Pearson's correlation could be spuriously high due to the presence of high leverage data points. In these cases, the final scores would be better estimated by a non-linear correlation test. Therefore, besides Pearson's we have also calculated Kendall's correlation coefficient. When scores obtained when Kendall and Pearson were, respectively, used as dependent and independent variables, the R2 of least square regressions were 0.67 when slope was estimated with LSQ, and 0.71 when slope was estimated with ROB (Fig. 2c, d). These results show that the linear and non-linear correlation methods tend to produce similar *gradient consistency test* scores.

Physicochemical Indexes Associated with Selection for Amino Acid Usage

For selection purposes, an index was considered to have a highly significant influence on amino acid selection when *q* values for the 6 species were lower than 0.005 in both Pearson and Kendall correlation tests in the two methods used for slope determination (LSQ and ROB). This threshold provides a more strict and accurate selection of significant results (Johnson 2003). One hundred and ten significant indexes were selected. The most frequent indexes selected were water solubility (hydropathy, hydrophobicity, $n=35$), solvent partition ($n=11$), side chain characteristic ($n=9$), biosynthetic cost ($n=5$), and flexibility ($n=3$) (Fig. 3 and Fig S10, Online Resource 3). An interesting aspect is that the score variation among species, measured by the standard deviation, was inversely correlated with the mean of the *gradient consistency test* score. Indexes with highest scores tended to have the smallest variations among species (Fig. 4). Therefore, indexes with highest influence on protein evolution exhibited similar *gradient consistency test* scores across species.

The Gradient Consistency Test is Capturing the Effect of Directional Selection of the Amino Acid Indexes

In order to show the strength of the *gradient consistency test* in capturing the effect of directional selection on amino acid properties, two simulation analyses were performed. For these analyses, 6 indexes that represent distinct and significant amino acid properties and were among the highest scores obtained were selected, hydrophobicity (NADH010103), long-range non-bonded energy per atom (OOBM770103), 8 A contact number (NISK800101), flexibility (VINM940103), side chain orientational preference (RACS770103), and biosynthetic costs (SN15).

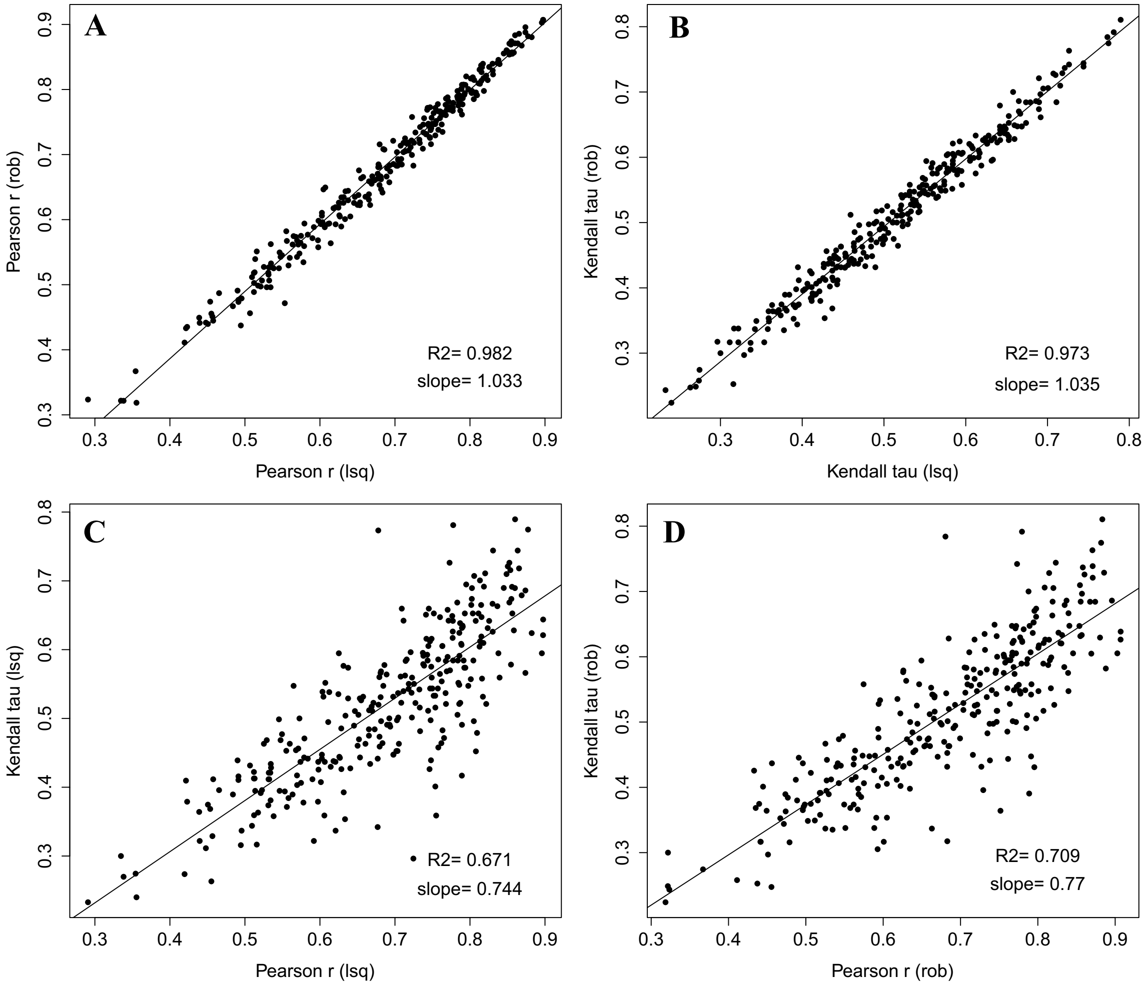

Fig. 2 Comparison of the use of Pearson and Kendall correlation methods to obtain the final *gradient consistency test* score. **a** G*radient consistency test* scores obtained with Pearson's correlation with the use of robust (rob) and least square (lsq) regression methods for slope determination. **b** G*radient consistency test* scores obtained with Kendall's correlation with the use of robust (rob) and least square (lsq) regression methods for slope determination. **c** *Gradient consistency test* scores obtained with Pearson's (*x*-axis) and Kendall's (*y*-axis) correlation with the use least square (lsq) regression method for slope determination. **d** *Gradient consistency test* scores obtained with Pearson's (*x*-axis) and Kendall's (*y*-axis) correlation with the use robust (rob) regression method for slope determination. Each point represents the *gradient consistency test* score of an amino acid property index. Note that the *gradient consistent test* scores obtained with Pearson's and Kendall's methods were highly correlated

In the first simulation experiment, 2000 proteins with random amino acid frequency and distribution were generated. Proteins were formed by 1001 amino acids. Random generated proteins were submitted to the *gradient consistency test.* Experiments were repeated 5 times. Wild-type protein *gradient consistency test* scores were always higher than random generated proteins when both Pearson (Fig. 5a) and Kendall (Fig. 5b) correlation methods were used to obtain the scores.

In the second simulation experiment, the strength of the *gradient consistency test* in capturing the effect of directional selection on amino acid properties was assessed by randomly reshuffling the amino acid indexes values previously to obtain the *gradient consistency test* scores. The reshuffling was repeated 500 times. G*radient consistency test* scores with unshuffled index values were always higher than shuffled when both Pearson (Fig. 6) and Kendall correlation methods (Fig S11, Online Resource 3) were used in Step 3.

The Strength of Gradient Consistency Test Can Vary Among Protein Subgroups.

Proteins can be grouped according to the predominance of amino acids with similar properties. Accordingly, a protein may have a predominance of hydrophobic or hydrophilic amino acids, or present an equilibrium of its constituents. Although these characteristics can influence its chemical properties and function, little is known about its influence on protein evolution. In order to assess the influence of this

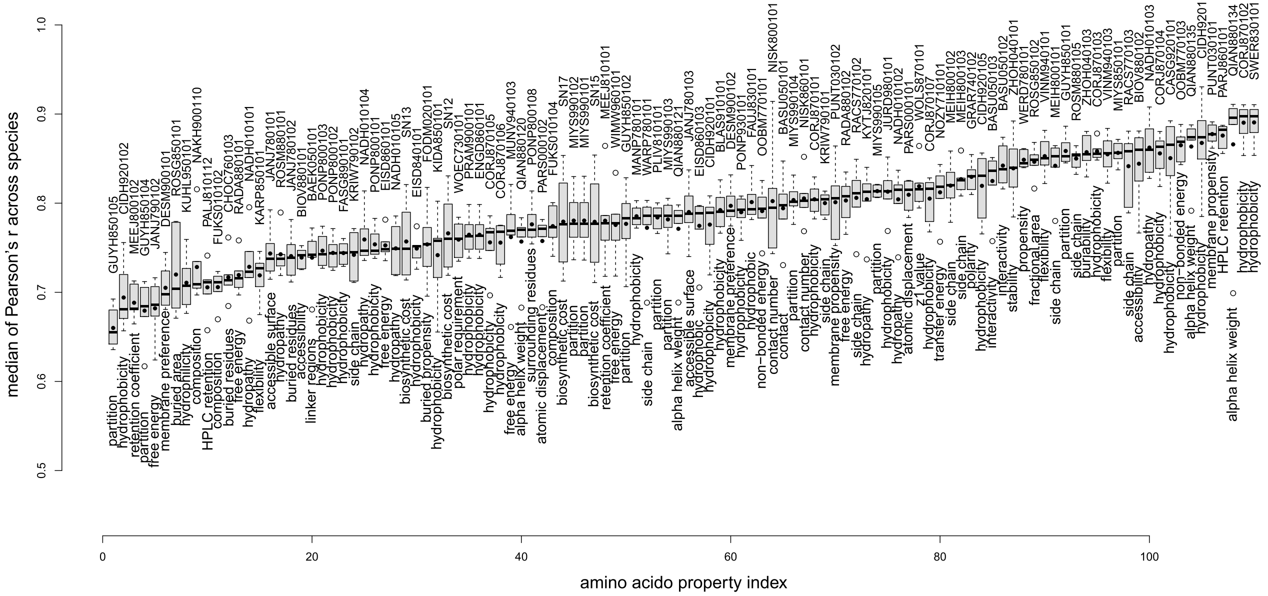

Fig. 3 Indexes with highest *gradient consistency test* scores. Boxplot showing the median and interquartile range of absolute values of *gradient consistency test* scores obtained with Pearson's correlation for the 6 taxa analyzed. Only indexes with q values lower than 0.005 for the six species were listed. Slopes were obtained with least square regression method. The index code and the property are shown above and below the boxplot, respectively. The mean scores are represented by black dots

aspect, human proteins were divided into subgroups according to the mean value of the amino acid property index of each protein. The *gradient consistency tests* were performed with the 6 indexes previously analyzed. Results show that directional selection estimated by the gradient consistency test can vary according to the subgroup characteristic. Subgroups with lowest and/or highest means tended to present the lowest absolute scores values (i.e., first and last columns, Fig. 7 and Fig S12, Online Resource 3), while subgroups with intermediate means frequently exhibited absolute scores higher than when the analysis was performed with all proteins (Fig. 7 and Fig S12, Online Resource 3). For biosynthesis cost (SN15), 8 A Contact Number (NISK800101), and hydrophobicity (NADH010103), non-significant score ($p > 0.05$) was observed only in the group of proteins with the highest mean synthesis cost. An opposite trend was observed with long-range non-bonded energy per atom (OOBM770103), side chain orientational preference (RACS770103), and flexibility (VINM940103) where selection on amino acid usage was weaker in proteins with low mean values (Fig. 7 and Fig S12, Online Resource 3).

Indexes That Estimate Distinct Amino Acid Properties May Be Correlated

Distinct amino acid properties may be related (Duan and Zhou 2005; Raiford et al. 2008). Therefore, we have analyzed possible correlations between the 6 properties selected. Significant correlations were found between biosynthetic costs (SN15) and 8 A Contact Number (NISK800101, Pearson = 0.53, $p = 0.017$); biosynthetic costs (SN15) and hydrophobicity (NADH010103, Pearson = 0.55, $p = 0.012$); flexibility (VINM940103) and side chain orientational preference (RACS770103, Pearson = 0.79, $p = 3e-05$); hydrophobicity (NADH010103) and 8 A contact number (NISK800101, Pearson = 0.91, $p = 3e-08$, Fig. 8). The 8 A Contact Number index (NISK800101) exhibited a significant correlation with the hydrophobicity scale NADH010103 that was based on solvent accessibility (Naderi-Manesh et al. 2001). This correlation can be explained by the fact that the contact energy between amino acids is also related to their solvent accessibility nature (Ma and Wang 2015).

Discussion

Protein structure, function, and evolution are mainly determined by its amino acid sequence. The amount and distribution of amino acids in a protein is ultimately determined by their physicochemical properties. Amino acid properties have been estimated through a large number of experiments and theoretical studies. These properties have been translated into numerical indexes, which have been used in a wide ranging of research areas such as protein subcellular localization, (Sarda et al. 2005), evolution (Abriata et al. 2015), and protein structure prediction (Pokarowski et al.

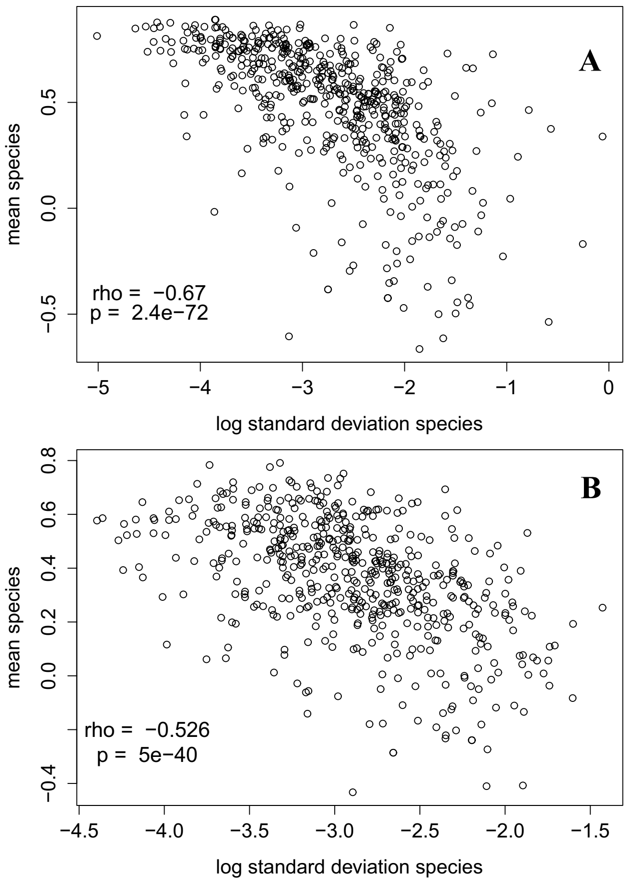

Fig. 4 Indexes with highest *gradient consistency test* scores tend to have the lowest standard deviations among taxa. **a** Dot plot showing the log of standard deviation (x-axis) versus the mean (y-axis) of *gradient consistency test* scores obtained with Pearson's correlation among the 6 species analyzed. **b** Dot plot showing the log of standard deviation (x-axis) versus the mean (y-axis) of *gradient consistency test* scores obtained with Kendall's correlation among the 6 species analyzed. The negative correlation between the two variables shows that higher scores were associated with smaller interspecies variations

2005). The results of the *gradient consistency test* rely upon the slope determination of bivariate regression analysis. There are several methods for line-fitting, which can produce lines with different slopes. Additionally, there are no precise parameters to evaluate which line-fitting method will reflect the most realistic and accurate biological association. In general, regression methods will produce lines with fairly similar slopes in the absence of outliers and leverage points. In order to overcome the influence of these factors, two different regression methods for line-fitting were used in our analysis. Gradient consistency test scores were considered significant only if both methods produce *q* values < 0.005. This procedure selected only amino acid properties whose slopes follow a concordant pattern among the methods used, increasing the reliability of our analysis. Our results showed that the use of least squared (LSQ) or robust regression (ROB) methods for slope determination did not significantly change the final score. These results indicate that outliers do not significantly influence the slopes of the regression lines, and both LSQ and ROB are equally suitable for the analyses. Likewise, the unbiased median and mean per residue index did not affect significantly the *gradient consistency test* scores.

In the gradient consistency test described by Swire (2007), the final score was given by the Pearson's correlation between the slope values and the indexes of the 20 amino acids. In the present work, both Pearson and Kendall correlations were used. The inclusion of the non-linear Kendall's correlation occurred due to the large-scale processing and statistical analyses performed, where 555 amino acid indexes, 2 slope methods, and 6 species were analyzed, with a total of 6660 tests. Spuriously high Pearson correlation coefficients may occur due to the presence of leverage and extreme data values. These abnormal cases do not significantly interfere with Kendall's Tau. In these analyses, Kendall's correlation was chosen over the more popular non-linear Spearman's correlation as it has a smaller gross error sensitivity and a smaller asymptotic variance (Croux and Dehon 2010).

In a broad sense, the indexes selected are related to two major aspects: protein topology (hydropathicity, contact number, flexibility, long-range non-bond energy per atom, side chain orientation) and biosynthesis cost. Hydropathicity is a physicochemical property that is relevant for the initial folding of polypeptides (Dyson et al. 2006), this parameter permits distinguishing between peptides with transmembrane α-helices and β-sheets (Simm et al. 2016). Hydropathicity indexes SWER830101 and NADH010103 presented the highest scores among all selected indexes when Pearson's r and Kendall's tau were used in Step 3, respectively. These results show that this property is of prime relevance for amino acid selection. The SWER830101 index was based on the observed statistical frequency of amino acid replacements among related structures (Sweet and Eisenberg, 1983). Its values were normalized with a mean of 0 and a standard deviation of 1. The NADH010103 index is based on prediction of solvent accessibility of amino acid residues in various states (Naderi-Manesh et al. 2001). It was obtained by the application of information theory from a single amino acid position and pair-information for a window of seventeen amino acids around the desired residue. In both indexes, the more hydrophobic amino acids have positive values while more hydrophilic are negative.

The 8 A contact number (NISK800101, Nishikawa and Ooi 1980) is a local packing density parameter, which refers to the number of C atoms around the distance range of 8 Angstroms of C-alpha atoms of the focal amino acid. The C-alpha is the first C bonded to the carbonyl C atom. The contact number is a measure of the exposition

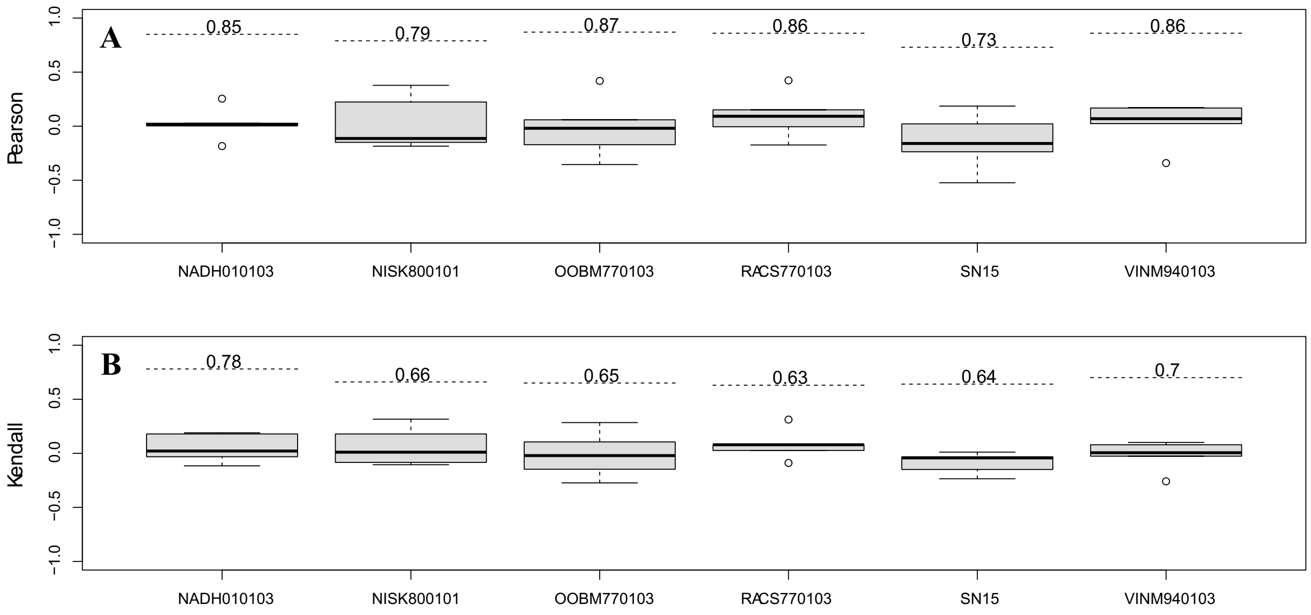

Fig. 5 The strength of the *gradient consistency test* in capturing the effect of directional selection on amino acid properties was assessed by generating 2000 proteins containing 1001 amino acids with random frequency and distribution. The boxes show the median and interquartile ranges of 5 distinct assays. The *gradient consistency test* scores performed with wild-type human proteins (dashed lines) were always higher than random generated proteins when both Pearson (**a**) and Kendall (**b**) correlation methods were used to obtain the scores

of amino acids to the local environment. The 8A contact number is correlated to the amino acid solvent accessible area (Pollastri et al. 2002). Contact number is an important parameter used for prediction of secondary structure of proteins (Heffernan et al. 2017), and was associated with protein evolution (Yeh et al. 2014; Shahmoradi and Wilke 2016).

The long-range non-bond energy per atom (OOBM770103, Oobatake and Ooi 1977) refers to non-covalent interactions between atoms. These interactions can be mediated by forces produced by electrostatic interactions, salt bridges, hydrogen bonds, van der Waals, and other weaker interactions among amino acids. The OOBM770103 index was calculated using the atomic coordinates obtained by X-ray crystallography of 16 proteins. The non-covalent interactions are critical in maintaining the tertiary and quaternary structures of proteins (Prasad et al. 2019). Long-range contact energy has been positively correlated with evolvability (Yan et al. 2014).

The normalized flexibility parameters (B-values) for each residue surrounded by none rigid neighbors (VINM940102, Vihinen et al. 1994) are related to protein structural stability. Protein structures are highly dynamic (Teilum et al. 2009). Many biological processes such as antigen–antibody receptor-ligand binding and enzyme catalysis are dependent on the capacity of proteins to permit conformational structural changes. Although flexible proteins or protein domains tend to evolve at a faster rate (Brown et al. 2011), the conservation of their flexibility indicates that this property is of key importance in proteins function (Forcelloni and Giansanti 2020) and evolution (Martin and Vila 2020).

The RACS770103 index was obtained by the analysis of the distribution of distances of each type of amino acid from the center of mass in a sample of 13 proteins (Rackovsky and Scheraga 1977). It was based on the orientational preference given by the ratio of occurrence in two orientations of the amino acids side chain.

Indexes with the highest scores tended to have the smallest variations among species. This result indicates that the *gradient consistency test* can select and rank the most relevant indexes related to amino acid selection (i.e., the higher the relevance the lower the interspecies variation). The fact that several amino acid properties were highly significant is supported by previous experimental analysis on beta-lactamase enzymes showing that substitutions in over one-third of the residues can be quantitatively modeled by monotonic dependencies on amino acid descriptors and predictions of changes in folding stability. Amino acid volume and steric hindrance are major constraints of evolution on the protein core; hydrophobicity and solubility are more relevant underneath the surface, while salt bridges and polar networks act on the protein surface. Amino acid flexibility also provides additional constraints at many locations. (Abriata et al. 2015).

The low score and non-significant *p* value (*p* value > 0.05) observed in the subgroup of proteins with the highest mean synthesis cost (SN15) is consistent with the selection towards lower cost proteins. It is likely that evolution of high-cost

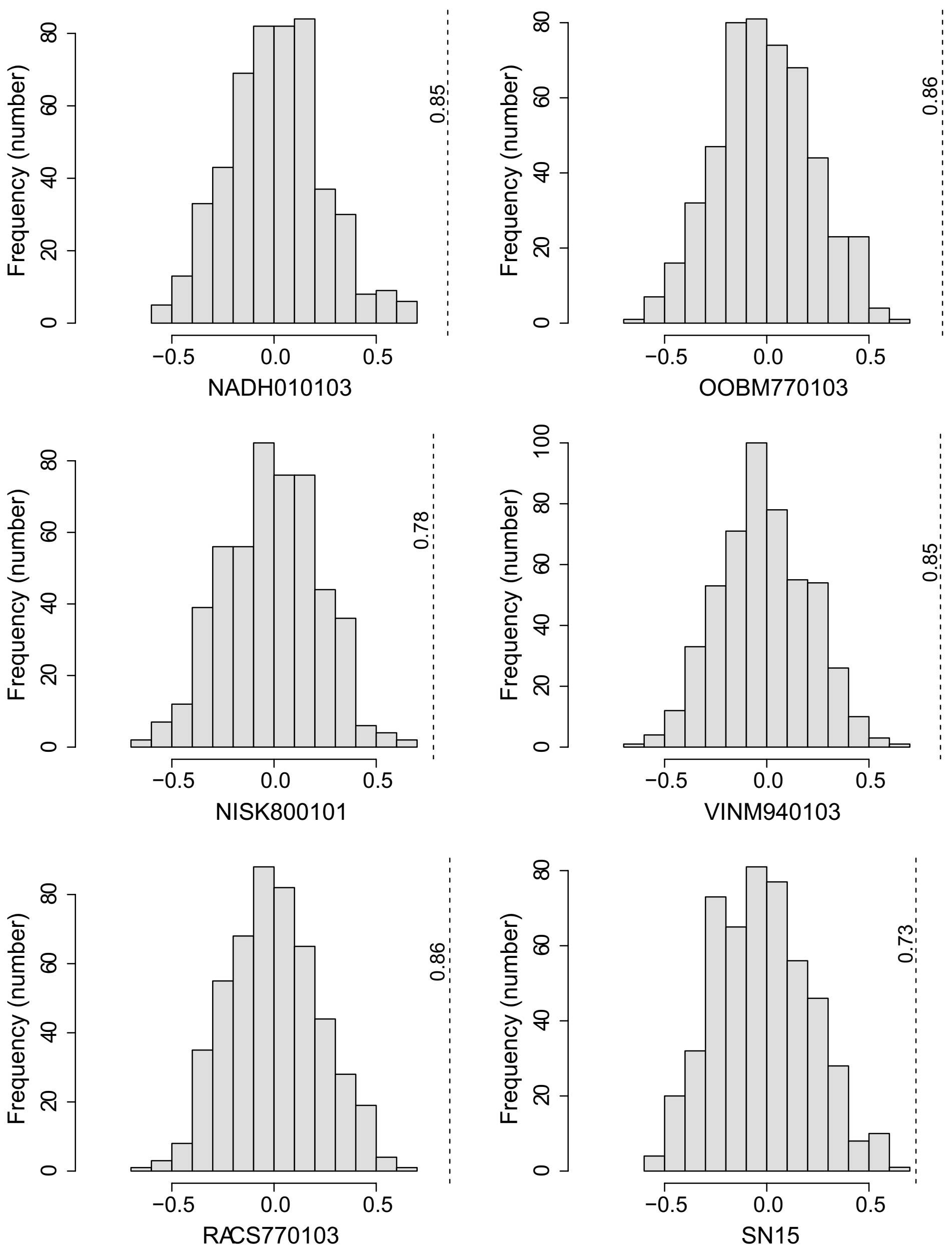

Fig. 6 The strength of the *gradient consistency test* in capturing the effect of directional selection on amino acid properties was assessed by randomly reshuffling the indexes values previously to obtain the *gradient consistency test* scores. Reshuffling was repeated 500 times. For these analyses, 6 indexes that represent distinct and significant amino acid properties were selected, hydrophobicity (NADH010103), long-range non-bonded energy per atom (OOBM770103), 8 A contact number (NISK800101), flexibility (VINM940103), side chain orientational preference (RACS770103), and biosynthetic costs (SN15). Pearson's correlation scores are shown on x-axis and the number of occurrences are shown on y-axis. Vertical dashed lines represent the scores obtained with unshuffled data. Note that unshuffled scores were always higher than shuffled

proteins might be constrained by other properties. The SN15 index values showed no significant correlation with long-range non-bonded energy per atom (OOBM770103), normalized flexibility parameters (*B* values) for each residue surrounded by none rigid neighbors (VINM940102), and side chain orientational preference (RACS770103), and only moderate correlations with the hydrophobicity (NADH010103) and 8 A contact number (NISK800101) indexes. The indexes NADH010103 and NISK800101 values were highly correlated (Pearson $r=0.91$). Similar to the SN15 in the NADH010103 and NISK800101 indexes, the subgroups with highest means presented the lowest *gradient consistency test* scores with non-significant *p* values (Fig. 7 and Fig S11, Online Resource 3). These two indexes showed a significant and positive correlation (see Fig. 8). This high linear correlation can be explained by the fact that the hydrophobicity NADH010103 index was estimated based on prediction of solvent accessibility (Naderi-Manesh et al. 2001) and 8A contact number is correlated to the amino acid solvent accessible area (Pollastri et al. 2002). In fact, in globular proteins the densely packed sites (i.e., high contact areas) are frequently highly hydrophobic (Rose and Roy 1980). Hydrophobic regions tend to be buried in the dense core of globular proteins (Perunov and England 2014).

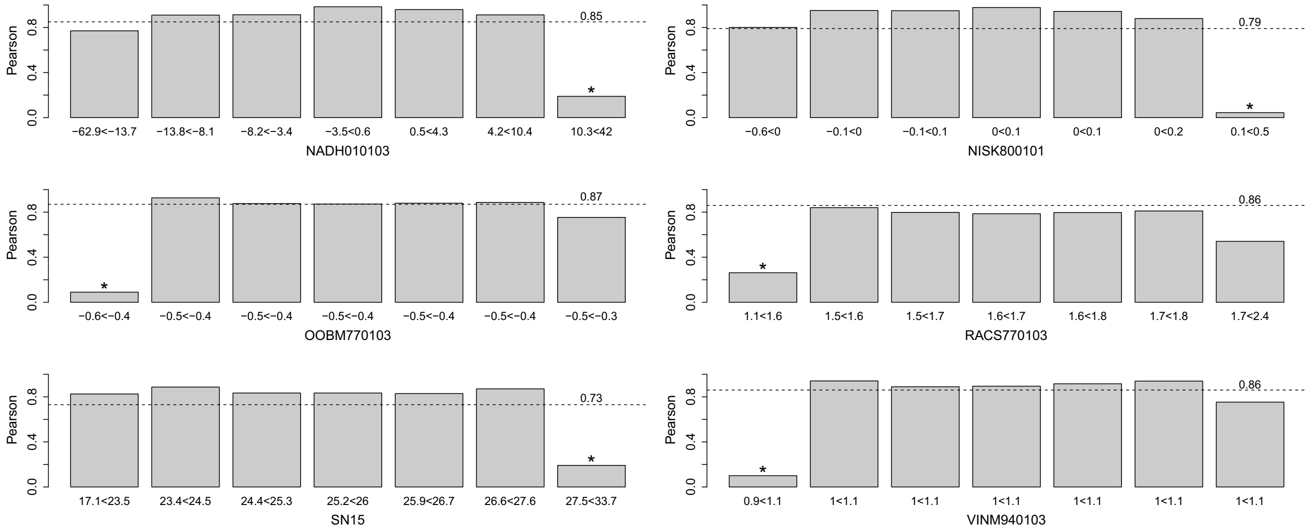

Fig. 7 The strength of *gradient consistency test* can vary among protein subgroups. For these analyses, 6 indexes that represent distinct and significant amino acid properties were selected, hydrophobicity (NADH010103), long-range non-bonded energy per atom (OOBM770103), 8 A contact number (NISK800101), flexibility (VINM940103), side chain orientational preference (RACS770103), and biosynthetic costs (SN15). Human proteins were divided into subgroups according to the mean value of the index of each protein. Subgroups with lowest and/or highest means presented the lowest absolute score values, while subgroups with intermediate means frequently exhibited absolute scores higher than when the analysis was performed with all proteins. In biosynthesis cost (SN15), the lowest absolute score was observed in the group of proteins with the highest mean synthesis cost. An opposite trend was observed with long-range non-bonded energy per atom (OOBM770103), 8 A contact number (NISK800101), and side chain orientational preference (RACS770103), where selection on amino acid usage was weaker in proteins with low mean values. The lowest absolute scores for hydrophobicity (NADH010103) and flexibility (VINM940103) were observed in both extremes. Horizontal dashed lines mark the value of the absolute scores of the whole (undivided) protein set. The number of proteins and index range in each subgroup is shown above and below the bars, respectively. The *gradient consistency test* scores obtained with Pearson's correlation are shown on the y-axis. *$p > 0.05$

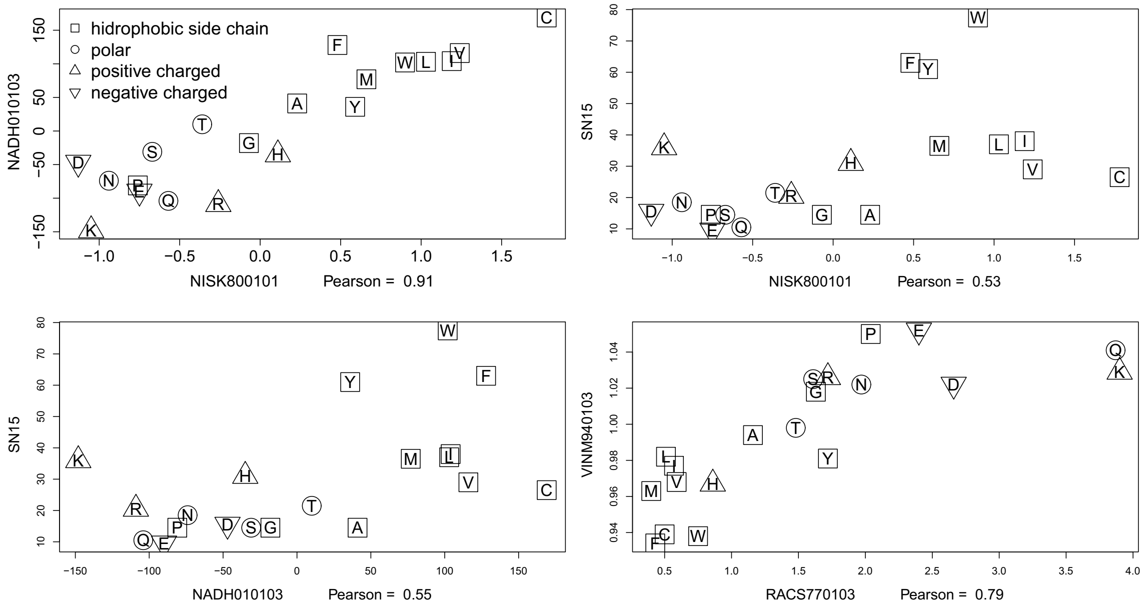

Fig. 8 Indexes that estimate distinct amino acid properties may be correlated. Scatterplot showing Pearson's correlations between indexes estimating different amino acid properties. Significant correlations ($p < 0.05$) were found between hydrophobicity (NADH010103) and 8 A contact number (NISK800101); biosynthetic costs (SN15) and 8 A contact number (NISK800101); biosynthetic costs (SN15) and hydrophobicity (NADH010103); flexibility (VINM940103) and side chain orientational preference (RACS770103)

The *gradient consistency test* gives an estimate of the strength of directional selection on a specific amino acid property (Swire 2007). This test assumes that if a specific amino acid property affects protein evolution then amino acids with high/low property index scores are expected to be found in proteins mostly composed of other high/low index score amino acids significantly more often than they are found in proteins mostly composed of low/high amino acids. This method was devised to detect synthesis cost selection. In order to check if the *gradient consistency test* would also perform with other amino acid properties, we have performed simulation analysis in 6 indexes with highly significant *gradient consistency test* scores that estimate distinct properties analyses. These analyses showed that for all properties the wild-type data always performed better than random amino acid-generated proteins or shuffled index values. The *gradient consistency test* does not depend on data of protein expression levels or phylogenetic comparison of sequences among species. Since there may be several indexes for a same amino acid property, the identification and ranking of indexes may allow a more precise estimation of amino acids property indexes that drive protein evolution. This aspect may be particularly interesting in the analysis of amino acid properties that constrains protein evolution in processes characterized by rapid proliferation and mutational rates such as bacterial (Bosshard et al 2019), viral proliferation (Korber et al. 2020), and cancer (Zhang et al. 2018). Besides evolutionary studies, the selection and use of indexes with high scores may allow a more accurate analyses of the effects of amino acid properties in protein structure and function.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00239-021-10003-7.

Funding GBF is supported by a fellowship from Sao Paulo Research Foundation-FAPESP (Grant 2018/23038–7). SRPL is supported by a grant from Brazilian Council for Scientific and Technological Development-CNPq (Grant 305783/2018–1).

Declaration

Conflict of interest The authors declare that they have no competing interests.

References

Abriata LA, Palzkill T, Dal Peraro M (2015) How structural and physicochemical determinants shape sequence constraints in a functional enzyme. PLoS ONE 10(2):e0118684. https://doi.org/10.1371/journal.pone.0118684

Bergman J, Eyre-Walker A (2019) Does adaptive protein evolution proceed by large or small steps at the amino acid level? Mol Biol Evol 36(5):990–998. https://doi.org/10.1093/molbev/msz033

Bosshard L, Peischl S, Ackermann M, Excoffier L (2019) Mutational and selective processes involved in evolution during bacterial range expansions. Mol Biol Evol 36(10):2313–2327. https://doi.org/10.1093/molbev/msz148

Brown CJ, Johnson AK, Dunker AK, Daughdrill GW (2011) Evolution and disorder. Curr Opin Struct Biol 21(3):441–446. https://doi.org/10.1016/j.sbi.2011.02.005

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP (2012) Predicting the functional effect of amino acid substitutions and indels. PLoS ONE 7(10):e46688. https://doi.org/10.1371/journal.pone.0046688

Croux C, Dehon C (2010) Influence functions of the Spearman and Kendall correlation measures. Stat Methods Appl 19:497–515. https://doi.org/10.1007/s10260-010-0142-z

DePristo MA, Weinreich DM, Hartl DL (2005) Missense meanderings in sequence space: a biophysical view of protein evolution. Nat Rev Genet 6(9):678–687. https://doi.org/10.1038/nrg1672

Duan MJ, Zhou YH (2005) A contact energy function considering residue hydrophobic environment and its application in protein fold recognition. Genomics Proteomics Bioinformatics 3(4):218–224. https://doi.org/10.1016/s1672-0229(05)03030-5

Dyson HJ, Wright PE, Scheraga HA (2006) The role of hydrophobic interactions in initiation and propagation of protein folding. Proc Natl Acad Sci USA 103(35):13057–13061. https://doi.org/10.1073/pnas.0605504103

Echave J, Wilke CO (2017) Biophysical models of protein evolution: understanding the patterns of evolutionary sequence divergence. Annu Rev Biophys 46:85–103. https://doi.org/10.1146/annurev-biophys-070816-033819

Forcelloni S, Giansanti A (2020) Evolutionary forces and codon bias in different flavors of intrinsic disorder in the human proteome. J Mol Evol 88(2):164–178. https://doi.org/10.1007/s00239-019-09921-4

Gillooly JF, McCoy MW, Allen AP (2007) Effects of metabolic rate on protein evolution. Biol Lett. https://doi.org/10.1098/rsbl.2007.0403

Heffernan R, Yang Y, Paliwal K, Zhou Y (2017) Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. Bioinformatics 33(18):2842–2849. https://doi.org/10.1093/bioinformatics/btx218

Hurst LD, Feil EJ, Rocha EP (2006) Protein evolution: causes of trends in amino-acid gain and loss. Nature 442(7105):E11–2; discussion E12. https://doi.org/10.1038/nature05137.

Johnson VE (2003) Revised standards for statistical evidence. Proc Natl Acad Sci USA 110:19313–19317. https://doi.org/10.1073/pnas.1313476110

Jordan IK, Kondrashov FA, Adzhubei IA, Wolf YI, Koonin EV, Kondrashov AS, Sunyaev S (2005) A universal trend of amino acid gain and loss in protein evolution. Nature 433(7026):633–638. https://doi.org/10.1038/nature03306

Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M (2008) AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 36(Database issue):D202–5. https://doi.org/10.1093/nar/gkm998.

Korber B, Fischer WM, Gnanakaran S et al (2020) Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell 182(4):812-827.e19. https://doi.org/10.1016/j.cell.2020.06.043

Lande R, Arnold SJ (1983) The measurement of selection on correlated characters. Evolution 1:109–124. https://doi.org/10.1111/j.1558-5646.1983.tb00236.x

Line SRP, Souza AP, Mofatto LS (2014) Large scale statistical analysis of genome data with Ruby and R: skipping interface libraries. EMBNet J. https://doi.org/https://doi.org/10.14806/ej.20.0.753

Ma J, Wang S (2015) AcconPred: predicting solvent accessibility and contact number simultaneously by a multitask learning framework under the conditional neural fields model. Biomed Res Int 2015:678764. https://doi.org/10.1155/2015/678764

McDonald JH (2006) Apparent trends of amino acid gain and loss in protein evolution due to nearly neutral variation. Mol Biol Evol 23(2):240–244. https://doi.org/10.1093/molbev/msj026

Martin OA, Vila JA (2020) The marginal stability of proteins: How the jiggling and wiggling of atoms is connected to neutral evolution. J Mol Evol 88(5):424–426. https://doi.org/10.1007/s00239-020-09940-6

Naderi-Manesh H, Sadeghi M, Arab S, Moosavi Movahedi AA (2001) Prediction of protein surface accessibility with information theory. J Proteins 42:452–459. https://doi.org/10.1002/1097-0134(20010301)42:4%3c452::aid-prot40%3e3.0.co;2-q

Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31(13):3812–3814. https://doi.org/10.1093/nar/gkg509

Nishikawa K, Ooi T (1980) Prediction of the surface-interior diagram of globular proteins by an empirical method. Int J Peptide Protein Res 16:19–32. https://doi.org/10.1111/j.1399-3011.1980.tb02931.x

Oobatake M, Ooi T (1977) An analysis of non-bonded energy of proteins. J Theor Biol 67:567–584. https://doi.org/10.1016/0022-5193(77)90058-3

Perunov N, England JL (2014) Quantitative theory of hydrophobic effect as a driving force of protein structure. Protein Sci 23(4):387–399. https://doi.org/10.1002/pro.2420

Pollastri G, Baldi P, Fariselli P, Casadio R (2002) Prediction of coordination number and relative solvent accessibility in proteins. Proteins 47:142–153. https://doi.org/10.1002/prot.10069

Pokarowski P, Kloczkowski A, Jernigan RL, Kothari NS, Pokarowska M, Kolinski A (2005) Inferring ideal amino acid interaction forms from statistical protein contact potentials. Proteins 59:49–57. https://doi.org/10.1002/prot.20380

Prasad VK, Otero-de-la-Roza A, DiLabio GA (2019) PEPCONF, a diverse data set of peptide conformational energies. Sci Data 6:180310. https://doi.org/10.1038/sdata.2018.310.

R Core Team (2013) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Rackovsky S, Scheraga HAT (1977) Hydrophobicity, hydrophilicity, and the radial and orientational distributions of residues in native proteins. Proc Natl Acad Sci USA 74:5248–5251. https://doi.org/10.1073/pnas.74.12.5248

Raiford DW, Heizer EM Jr, Miller RV, Akashi H, Raymer ML, Krane DE (2008) Do amino acid biosynthetic costs constrain protein evolution in Saccharomyces cerevisiae? J Mol Evol 67(6):621–630. https://doi.org/10.1007/s00239-008-9162-9

Rose GD, Roy S (1980) Hydrophobic basis of packing in globular proteins. Proc Natl Acad Sci USA 77(8):4643–4647. https://doi.org/10.1073/pnas.77.8.4643

Rudnicki WR, Mroczek T, Cudek P (2014) Amino acid properties conserved in molecular evolution. PLoS ONE 9(6):e98983. https://doi.org/10.1371/journal.pone.0098983

Sarda D, Chua GH, Li K-B, Krishnan A (2005) pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties. BMC Bioinformatics 6:152. https://doi.org/10.1186/1471-2105-6-152

Shahmoradi A, Wilke CO (2016) Dissecting the roles of local packing density and longer-range effects in protein sequence evolution. Proteins 84(6):841–854. https://doi.org/10.1002/prot.25034

Simm S, Einloft J, Mirus O, Schleiff E (2016) 50 years of amino acid hydrophobicity scales: revisiting the capacity for peptide classification. Biol Res 49(1):31. https://doi.org/10.1186/s40659-016-0092-5

Storey JD, Bass AJ, Dabney A, Robinson D (2020) qvalue: Q-value estimation for false discovery rate control. R package version 2.22.0. http://github.com/jdstorey/qvalue.

Suckow J, Markiewicz P, Kleina LG, Miller J, Kisters-Woike B, Müller-Hill B (1996) Genetic studies of the Lac repressor XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. J Mol Biol 261(4):509–523. https://doi.org/10.1006/jmbi.1996.0479

Sweet RM, Eisenberg DT (1983) Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. J Mol Biol 171:479–488. https://doi.org/10.1016/0022-2836(83)90041-4

Swire J (2007) Selection on synthesis cost affects interprotein amino acid usage in all three domains of life. J Mol Evol 64:558–571. https://doi.org/10.1007/s00239-006-0206-8

Teilum K, Olsen JG, Kragelund BB (2009) Functional aspects of protein flexibility. Cell Mol Life Sci 66(14):2231–2247

Venables WN, Ripley BD (2002) Modern applied statistics with S, Fourth edition. Springer, New York. ISBN 0-387-95457-0. http://www.stats.ox.ac.uk/pub/MASS4/.

Vihinen M, Torkkila E, Riikonen PT (1994) Accuracy of protein flexibility predictions. J Proteins 19:141–149. https://doi.org/10.1002/prot.340190207

Wei Q, Wang L, Wang Q, Kruger WD, Dunbrack RL Jr (2010) Testing computational prediction of missense mutation phenotypes: functional characterization of 204 mutations of human cystathionine beta synthase. Proteins 78(9):2058–2074. https://doi.org/10.1002/prot.22722

White CR, Seymour RS (2003) Mammalian basal metabolic rate is proportional to body mass 2/3. Proc Natl Acad Sci USA 100(7):4046–4049. https://doi.org/10.1073/pnas.0436428100

Yan W, Sun M, Hu G, Zhou J, Zhang W, Chen J, Chen B, Shen B (2014) Amino acid contact energy networks impact protein structure and evolution. J Theor Biol 355:95–104. https://doi.org/10.1016/j.jtbi.2014.03.032

Yeh SW, Huang TT, Liu JW, Yu SH, Shih CH, Hwang JK, Echave J (2014) Local packing density is the main structural determinant of the rate of protein sequence evolution at site level. BioMed Res Int 2014:572409. https://doi.org/10.1155/2014/572409

Zhang H, Wang Y, Li J, Chen H, He X, Zhang H, Liang H, Lu J (2018) Biosynthetic energy cost for amino acids decreases in cancer evolution. Nat Commun 9(1):4124. https://doi.org/10.1038/s41467-018-06461-1