RESEARCH ARTICLE

The response to selection in Glycoside Hydrolase Family 13 structures: A comparative quantitative genetics approach

Jose Sergio Hleap[1,2¤]*, **Christian Blouin**[3]

1 Department of Human Genetics, McGill University, Montreal, QC, Canada, **2** SQUALUS Foundation, Cali, Colombia, **3** Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada

¤ Current address: office 7209, 740 Dr Penfield Ave, Montreal, QC H3A 0G1, Montreal, QC, Canada
* jshleap@squalus.org

Abstract

The Glycoside Hydrolase Family 13 (GH13) is both evolutionarily diverse and relevant to many industrial applications. Its members hydrolyze starch into smaller carbohydrates and members of the family have been bioengineered to improve catalytic function under industrial environments. We introduce a framework to analyze the response to selection of GH13 protein structures given some phylogenetic and simulated dynamic information. We find that the TIM-barrel (a conserved protein fold consisting of eight *α*-helices and eight parallel *β*-strands that alternate along the peptide backbone, common to all amylases) is not selectable since it is under purifying selection. We also show a method to rank important residues with higher inferred response to selection. These residues can be altered to effect change in properties. In this work, we define fitness as inferred thermodynamic stability. We show that under the developed framework, residues 112Y, 122K, 124D, 125W, and 126P are good candidates to increase the stability of the truncated *α*-amylase protein from *Geobacillus thermoleovorans* (PDB code: 4E2O; *α*-1,4-glucan-4-glucanohydrolase; EC 3.2.1.1). Overall, this paper demonstrates the feasibility of a framework for the analysis of protein structures for any other fitness landscape.

Introduction

The Glycoside Hydrolase Family 13 (GH13) is a multi-reaction catalytic family and its members perform hydrolysis, transglycosylation, isomerization [1], condensation, and cyclization reactions [2], and even animal amino acid transport [3] with no glycolase activity [4]. The initial definition for this family was formulated in the early 1990's [5–7]. According to this definition, a member of this family must [6]: 1) hydrolyse or form (by transglycosylation) *α*-glucosidic linkages; 2) have four conserved amino-acidic regions [8]; 3) contain the catalytic triad: Asp, Glu, and Asp.; and 4) have a TIM-barrel-like fold in its structure.

Since then, the number of family members have increased [9] to include *α*-1,1-, *α*-1,2-, *α*-1,3- and *α*-1,5-glucosidic linkages [1]. Also, the number of conserved regions has been updated to 7 [10, 11].

OPEN ACCESS

Citation: Hleap JS, Blouin C (2018) The response to selection in Glycoside Hydrolase Family 13 structures: A comparative quantitative genetics approach. PLoS ONE 13(4): e0196135. https://doi.org/10.1371/journal.pone.0196135

Editor: Israel Silman, Weizmann Institute of Science, ISRAEL

Received: October 30, 2017

Accepted: April 7, 2018

Published: April 26, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files. Some source data are available from the PDB database (accession numbers provided in the supplements).

Funding: NSERC partially financed the project of this article under the grant No. 120504858. This work was also partially supported by the Departamento Administrativo de Ciencia y Tecnología - Colciencias (Colombia) through the CALDAS scholarship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

The catalytic activity and substrate binding residues in the GH13 family members occur at the C-termini of the β-strands and in the loops that extend from these strands [12]. The catalytic site includes aspartate as a catalytic nucleophile, glutamate as an acid/base, and a second aspartate for stabilization of the transition state [13]. The catalytic triad plus an arginine residue are conserved in this family across all catalytic members [14, 15]. The GH13 family has many characterized enzymes with diverse functions which are summarized and clustered in the CAZY database [16]. GH13 is a highly diverse family in function and is also ubiquitous, being found in all kingdoms of life [15]. The GH13 family has been subdivided into over 40 subfamilies [17, 18] by their sequence motif and enzyme specificity [9], but they are all related both in sequence and structure. To date, thousands of sequences comprise this family for which hundreds of structures have been solved with more than 30 different enzymatic specificities [9]. Many comprehensive reviews on their mechanisms, sequences, abundance, phylogeny and concept have been performed [12, 15, 19–24].

Part of the interest in researching this family lies in its industrial importance [25, 26]. The GH13 family is the target of engineering efforts, focusing on factors such as thermal and alkaline stability [27–30], specific activity [27, 31], and other diverse biochemical properties that are important to the industrial context [28, 32, 33]. Many strategies have been used to engineer this family including different "rational design" approaches [34] such as B-fitter [35], proline theory [36], PoPMuSiC-2.1 [37], and sequence consensus [34]. However, to our knowledge, there has been no attempt to leverage both phylogenetic and molecular dynamics signals to quantify the potential of a structure in response to selection.

Exploring how selection acts on protein structures is not a trivial problem. One approach is to assume that protein structures are shape phenotypes and that their 3D structures respond to both genetic and environmental factors, thereby falling under a quantitative genetics framework. Proteins and other shapes are highly multivariate in nature [38], and the model for their phenotype (y) can be expressed as [39]:

$$y = Xb + Za + e \tag{1}$$

where X and Z represent design matrices for the fixed and random effects in vectors b and a respectively, and e is the residual component that cannot be explained by the model. Here, y is the phenotype of one structure and contains the x, y, and z coordinates of each homologous residue. The residue's homology is inferred with respect to the rest of the structures being analyzed. For a protein structure t that has 100 homologous residues, the length of y_t is 300. The more detailed explanation of the abstraction of the protein structure as a shape can be seen in section "Abstracting protein structures as shapes". With this model, the phylogenetic contribution to phenotype can be estimated. In a multivariate setting such estimation is called the G-matrix, or genetic variance-covariance matrix, which summarizes the genetic contribution and the interaction of all traits. In the example above, G is a 300 by 300 matrix.

Lande and Arnold [40] proposed a multivariate strategy to estimate the response to selection given G as:

$$\Delta\bar{z} = G\beta \tag{2}$$

where $\Delta\bar{z}$ is a vector of changes in traits, and β is a vector of selection gradients. The latter quantity is the effect of a particular trait on the relative fitness, and therefore depends on its definition. Here we define fitness of a molecule as its function. In enzymes, for example, this term could include the stability, effectiveness, and efficiency necessary for the protein to perform the required function. Then, the selection gradient can be understood as the change in fitness when the trait (in this case geometry) varies.

To apply the framework, the estimation of a G-matrix is required [41]. To deal with the fact that the number of samples is limited, this inversion of matrices requires expensive computation and an eigen-decomposition of the covariance matrices carried out using the restricted maximum likelihood (REML) approach is typically employed to perform the variance decomposition. When applied to univariate data, REML is more accurate than maximum likelihood methods because it better handles missing data (i.e. unknown parents, arbitrary breeding designs, etc. . .) and can account for selection processes. However, REML has good properties only asymptotically. The reliability of the estimates is questionable when data is scarce. One way to deal with complex cases that might bias the REML estimates is to use Bayesian inference of the animal model. This approach uses Markov chain Monte Carlo simulations and produces more robust estimations than REML, with equivalent results in less complex cases [42]. This robustness assumes that the Bayesian model has enough information in the prior probability distribution. A given set of priors considerably affect the estimation of the variance components. In particular, uninformative priors, such as flat priors (all parameters are equally likely), can lead to biases in the estimation.

Lynch's comparative quantitative genetic model: Applications to protein structures

Lynch [43] developed the *phylogenetic mixed model (PMM)*. In this model, the correlation of phylogenetically heritable components is the length of the path from the most recent common ancestor among two species and the root of the phylogenetic tree (i.e. time to the shared common ancestor) in the phylogeny [44]. The PMM can be described as [43]:

$$\bar{z} = X\mu + a + e \tag{3}$$

where X is an $np \times p$ incidence matrix, p being the number of traits and n the number of observations.

An assumption of the model is that μ_c is shared among all taxa in the phylogeny. This is a sensible assumption to make when analyzing truly homologous protein structures, since the mean effect on the phenotype is shared by common ancestry. This also means that μ_c+ a_{ci} can be interpreted as the heritable component of the mean phenotype for the ith taxon [43].

Here, the phylogenetic effects are the portion of the variation that has been inherited from ancestral species [45]. It does not only contain the genetic component, but also some environmental contributions given the shared evolutionary history of the taxa [46]. In PMM the ratio between the additive component and the total variance is the heritability (h^2) in a univariate approach. Housworth et al. [46] pointed out that a univariate h^2 in a PMM is actually equivalent to Freckleton et al. [47]'s and Pagel [48]'s phylogenetic correlation (λ).

Despite the robustness of the models, the REML technique, employed to estimate them, has two major drawbacks: assumption of normality of the data and high sample size requirements. It is widely known that REML poorly estimates genetic correlation when overparameterized (multi-trait inference), when the sample size is small (Martins, personal communication), and when the normality assumption is violated [44]. These violations can be handled in a Bayesian framework using Markov Chain Monte Carlo techniques. In such techniques, the higher complexity of the joint probability calculation needed for the likelihood estimation can be broken down in lower dimensional conditionals. From those conditionals the MCMC sampling can be performed and marginal distributions can be extracted [44]. A discussion of the use of Bayesian MCMC techniques is beyond the scope of this work. We refer the interested reader to Sorensen and Gianola [49] for a good description of likelihood and Bayesian methods in quantitative genetics.

Despite its strengths, the Bayesian framework also has weaknesses. The most important one is that it requires proper and informative priors. Uninformative priors lead to biases with high variation in results. The sensitivity to the choice of prior distribution should always be assessed [50]. Given that in evolutionary biology datasets the amount of knowledge on the estimator is scarce, well-informed priors are normally not available and by informing priors with partial information, the estimation can become ill-conditioned.

To explore the feasibility of a comparative quantitative genetics (CQG) framework in protein structures, we simulated a dataset with variable numbers of traits and observations. We show that the current implementations of the CQG framework are not feasibly applied to the dimensionality required for protein structures. We devised a method that functions as a proxy for the CQG framework and show that it is feasible and accurate. By applying this framework using the energy of unfolding (ΔG°) as fitness function to the GH13 family, we are able to show how purifying selection has fixed the geometry of the TIM-barrel. We also demonstrate how, by changing the fitness function, the response to selection propensity changes accordingly. Finally, a proxy for the amount of dynamic deformation happening in the protein, given a vector of selection, is explored.

Overall, we present here a starting framework to explore protein structure evolution and design. This approach has the potential to inform researchers on the potential of a given structural family to respond under selective pressures. This is especially important in protein engineering and structural evolution. It also has the potential to inform possible pathways to be visited in a given fitness landscape. This approach can narrow down the structural variants that a particular structural group followed (or could/can potentially followed) in evolution or engineering, and therefore become an important tool in structural biology inquiring.

Materials and methods

GH13 dataset

The GH13 family has 111 structures reported in the CAZY database, with 386 representatives in the PDB. Given that molecular dynamic simulations are very time consuming, we used a subset of the proteins classified as Glycoside Hydrolases Family 13 (GH13). A stratified selection (from each CAZY structural grouping) of 35 protein structures was performed maximizing taxonomic diversity (i.e. avoiding sampling the same species). The structure of the maltotetraose-forming exo-amylase from *Pseudomonas stutzer* (PDB code 1GCY) failed during the MD simulation due to structural abnormalities in the crystal, perhaps steric clashes, which induced unacceptably large forces, causing the integrator to fail. After lengthy energy minimizations, the structure could not be resolved and therefore was removed from the analysis. We chose to remove the structure instead of artificially modifying it, in order to avoid bias in the data. A final set of 34 protein structures (Table A1 in S1 File) was used in further analyses.

Molecular dynamics (MD) simulations

Each of the 34 protein structures was simulated in solution using the software `GROMACS 4` [51]. The force field modes used for the simulations were GROMOS96 for the protein, and SPCE for the water molecules. Data were collected every two picoseconds for at least 40 nanoseconds, discarding the first 10 nanoseconds of simulation to achieve stability. This process was performed using a workstation with 24 CPU cores and an NVIDIA TESLA™ GPU.

The analysis of these simulations will provide information on the flexibility (or within protein variance) of the protein, as opposed to the analysis across homologs which would provide phylogenetic information (or between structures variance). By 40 ns all proteins analyzed have achieved equilibrium and therefore most of the intrinsic variance has been captured.

Aligning the structures and MD simulations

The alignment of homologous proteins was performed using `MATT` software [52]. To align the snapshots from MD simulations a General Procrustes Superimposition (GPS) was performed using the `R` package `shapes` [53].

Abstracting protein structures as shapes

On a set of aligned protein structures, the abstraction is performed in a similar way to that in Adams and Naylor [54]. However, they do not fully describe the abstraction. Here we assign a landmark to the centroids of residues defined by:

$$\left(\frac{1}{A}\sum_{j=1}^{A} X_j, \frac{1}{A}\sum_{j=1}^{A} Y_j, \frac{1}{A}\sum_{j=1}^{A} Z_j\right) \tag{4}$$

where A will be the number of heavy atoms (C, O, N) that constitute the side chain of a residue including the alpha carbon (C_α). This procedure takes into account only the homologous residues. It captures the variance of both the backbone and the side chain. In the case of glycine, the centroid is the C_α.

Once the structure is abstracted as a shape, the resulting n (number of observations) by l (number of coordinates of homologous residues) matrix is referred to as the phenotypic matrix (P). For example, let us assume that we have a protein structure composed of 150 residues. Let's imagine that 100 different taxa share an ortholog of this protein. After aligning the protein structures let's assume that 100 residues are homologous across all 100 taxa. The resulting phenotypic matrix (P) will be composed of 100 rows of observations (n) and 300 coordinates (l). These dimensions correspond to the x, y, and z axis of each of the 100 homologous residues. To estimate the variation of this phenotype, the phenotypic variance V_P can be estimated by computing the variance-covariance matrix of P as $V_P = var(P)$, or G in a multivariate scenario.

Pooled-within group covariance matrix estimation

After the MD simulations up to 500 samples per simulation were obtained. The estimation of the pooled-within covariance matrix was performed as follows:

1. Align every model within each MD simulation using General Procrustes Superimposition (GPS): Remove extra rotations and translations that could occur during MD simulation.
2. Select an ambassador structure that is closest to the mean structure (the geometrical mean of the dataset).
3. Align all ambassadors using MATT flexible structure aligner to identify homologous sites: Multiple structure alignment to identify structural homology.
4. Extract the centroid of fully homologous sites: Identify shared information among all structures, as explained in section "Aligning the structures and MD simulations".
5. Concatenate the centroids' three dimensions for all trajectories.
6. Perform a GPS on the entire set of shapes to bring all pre-aligned structures into the same reference plane.

7. Compute the pooled-within-species covariance matrix (W) by first computing the deviation from the mean in each class/group (individual homologs in our case) as:

$$D_k = x_k(\omega) - \bar{x}_{k,s} \tag{5}$$

then computing the sum over the classes of the products of D_k as:

$$F_{l,m} = \sum_{\omega : f(\omega) = s} [D_l] \times [D_m] \tag{6}$$

Finally, compute the pooled-within covariance matrix:

$$W = \frac{1}{n - S} \sum_{s=1}^{S} (F_{i,j})_{i,j=1,\cdots,p} \tag{7}$$

where S is the number of categorical variables describing the groups or species, ω is an instance where $f(\omega)$ corresponds to the class value of the instance, and $\bar{x}_{i,s}$ is the mean of the variable i for individuals belonging to s. Finally, n is the sample size.
Here, W contains the covariance matrix of the within-homolog (i.e. Molecular dynamic data). To estimate the evolutionary component of P, the between structures/species covariance matrix (B) has to be taken into account. B will be simply the difference between the V_P and W.

Estimating $\Delta G^{\circ}_{unfold}$ as proxy for fitness

The $\Delta G^{\circ}_{unfold}$ on each model for each protein was estimated using the command line version of `FoldX` [55]. It is important to notice that the computed $\Delta G^{\circ}_{unfold}$ is not comparable in proteins of different size, therefore we computed the average $\Delta G^{\circ}_{unfold}$ per residue as:

$$\Delta \hat{G}^{\circ}_{unfold} = \frac{\Delta G^{\circ}_{unfold}}{r} \tag{8}$$

r being the number of residues. With this $\Delta \hat{G}^{\circ}_{unfold}$ as proxy for fitness we can try to explore the fitness surface. To do this, we used the first two principal components (PC) of a PC analysis of the shapes as X and Y axes; $\Delta \hat{G}^{\circ}_{unfold}$ in the Z axis (S1 Fig).

Is important to state here that the fitness is a relative quantity and depends on the objective of the analysis. Here we chose $\Delta G^{\circ}_{unfold}$ for ease of computation, but this function only measures structural stability. In proteins, function is the main selective trait. However, function in proteins depends on several properties of the structure such as the aforementioned stability, as well as the Michaelis constant (for enzymes), activation energy, free energy of the system, among others. In this manuscript we will assume that the fitness function we are modeling is stability, and therefore $\Delta G^{\circ}_{unfold}$ works as a good proxy for it.

Propensity to respond to selection

Arnold [56] showed that, despite high additive variances, G might not be aligned with the fitness surface. This implies that even though β_λ can be non-zero, the response to selection might send the phenotype in a different direction than the fitness surface. Blows and Walsh [57] and Hansen and Houle [58] developed an approach to measure the angle between β and the predicted response to selection from the multivariate breeders equation, $\Delta\bar{z}$ as:

$$\theta_{\Delta\bar{z}\text{-}\beta} = cos^{-1}\left(\frac{\Delta\bar{z}^{T}\beta_{\lambda}}{\sqrt{\Delta\bar{z}\Delta\bar{z}^{T}}\sqrt{\beta_{\lambda}\beta_{\lambda}^{T}}}\right). \tag{9}$$

$\theta_{\Delta\bar{z}-\beta}$ would be zero when there is no genetic constraint, whereas an angle of 90° would represent an absolute constraint [59].

In simpler terms, $\theta_{\Delta\bar{z}-\beta}$ will tell us how responsive to selection the protein is. Let's imagine we are trying to push one particular protein structure towards a higher fitness by selection. If $\theta_{\Delta\bar{z}-\beta}$ is low (close to zero), most of our selection effort will result in a shift in the structures towards the desired goal. If $\theta_{\Delta\bar{z}-\beta}$ is closer to 90°, a very small fraction of the selection effort will go towards the desired outcome. $\theta_{\Delta\bar{z}-\beta}$ is therefore measuring the propensity of the structure to respond to selection.

Results and discussion

In Supplementary Material and Methods, and in Supplementary Results (S1 File) we have shown that the traditional PMM models and their Bayesian counterparts are not feasible when the number of traits and observations are in the order of those obtained in protein science when MD simulations are taken into account. Here, we applied a simple method to overcome this over-parameterization.

Overcoming over-parameterization: Approaching the G-matrix by means of the P-matrix

Given the previous results, the estimation of the G-matrix within the Lynch's PMM is not feasible. This is not a new observation since in comparative evolutionary biology it is widely known that accurate measures of *G* are difficult or impossible to obtain [60]. This pattern is even more evident when dimensionality is high. On average, protein structures are composed of over 200 residues in a three-dimensional system, which means over 600 variables. Also, the sample size at the species level is typically small. For these reasons, a full and stable estimation of the *G*-matrix is not possible. However, an increased number of samples can be achieved by means of molecular dynamic simulations. This increases *n* considerably depending on the length of the simulation. We have shown the infeasibility of the GLMM to deal with the dimensionality and very large sample size. However, it has been shown that phenotypic (V_P) covariance matrices can be estimated with more confidence with large sample sizes [61]. It is also shown that in some cases, V_P can be used as surrogate for *G* when the two are proportional [60, 62]. To test this, we performed a shape simulation explained in S1 File. The simulation was performed with 500 replicates as molecular dynamics snapshots, 100 taxa, and the traits were varied from 2 to 1024 in a geometric series increase. Since the within-homolog matrix structure is known, a pooled-within covariance matrix (*W*) was computed as exposed in the section "Pooled-within group covariance matrix estimation".

Table 1 shows the feasibility and accuracy of the pooled-within species co-variance estimation method. Here the Cheverud's Random Skewer (RS) test [61, 63] implemented in the `R` package `phytools` [64] were used to test the accuracy. A discussion of the appropriateness of the usage of this metric can be found in S1 File and references therein.

Even with highly multivariate data (1024 traits), the memory requirement is manageable (less than 2 Gb), the evaluation is completed in under an hour, and the accuracy of the estimation is high. The estimated G-matrix is almost identical to the simulated matrix in most of the

Table 1. Accuracy and feasibility of the pooled-within covariance estimation. Memory (Mb), time (sec) and accuracy (random skewer correlation) of the pooled-within covariance estimation approach. RS_B corresponds to the random skewer test for the phylogenetic covariance and RS_W to the dynamic component.

Traits	Time (secs.)	Memory (Mb)	RS_B		RS_W	
			p-val	ρ	p-val	ρ
2	0.60	182.9	0.002	1.000	0.021	0.999
4	0.80	238.2	0.000	0.999	0.007	0.952
8	1.00	387.6	0.000	0.998	0.000	0.983
16	1.82	407.5	0.000	0.998	0.000	0.963
32	6.08	428.5	0.000	0.998	0.000	0.966
64	20.32	465.9	0.000	0.999	0.000	0.953
128	91.14	539.4	0.000	0.999	0.000	0.947
256	341.90	686.8	0.000	0.999	0.000	0.950
512	1342.36	982.2	0.000	0.999	0.000	0.938
1024	5268.82	1843.7	0.000	0.999	0.000	0.937

https://doi.org/10.1371/journal.pone.0196135.t001

runs, and the estimated *MD* have over 0.97 correlated responses to random vectors than the actual *MD*. This is a surprising result since this method cannot completely separate the error terms from the genetic and dynamic components. However, the split of the error term between the two other components can render it negligible. Moreover, it seems that error does not significantly affect the structure of *G* and *MD*, allowing them to behave almost identically in comparison to the simulated counterparts. Given these results, and the fact that the application to real datasets can only be made with this approach, it is reasonable to keep using the described method from this point forward. However, the biological and evolutionary meaning of this approach is less clear than in the other methods since there is no explicit use of a phylogeny.

Meaning of the pooled within-structure covariance matrix. V_P-matrices can be used as surrogates of G-matrices in cases were they are proportional or sufficiently similar [65]. Proa et al. [65] showed that this assumption can be relaxed if the correlation between *G* and $V_P \geq 0.6$. In protein structures, we can assume that given the strong selective pressures and long divergence times, the relationship between V_P and *G* is standardized. Assuming that this is true in protein structures, the estimated pooled variance-covariance (V/CV) matrices in real datasets might have a specific biological meaning. This was described in Haber [66] for morphological integration in mammals. Following Haber's [66] logic, the within-structure/species (i.e. thermodynamic V/CV) matrix refers to integration of residues in a thermodynamic and functional manner. It also contains information about environmental factors affecting the physical chemistry of the structure. Haber [66] includes a genetic component for his estimation of the within population variation, since populations follow a filial design. Our data, on the other hand, have a controlled amount of genetic component given that the sampling is done in a time series instead of a static population. Our approach would be more related to an estimation of within repeated measures design.

The among-structure/species (i.e additive or evolutionary V/CV) matrix refers to the concerted evolution of traits given integration and selection [66].

Response to selection in the GH13 family

As defined in Eq 2, the response to selection of a phenotype depends on the within-species change in mean due to selection, the correlation between different traits, and the amount of heritable component of the shape. The first component can be referred to as $\beta = V_P^{-1}S$, and also known as the vector of selection gradients [67] or directional selection gradient. The second and third elements are summarized in the G-matrix. As expressed in Eq 2, this covariance matrix represents the genetic component of the variation in the diagonal, and the correlated response of every trait to each other in the off-diagonal.

Another extension from Eq 2 is to compute the long-term selection gradient assuming that G is more or less constant over long periods of time:

$$\beta_{\lambda} = G^{-1}\Delta\bar{z} \tag{10}$$

Here $\Delta\bar{z}$ would be proportional to the differences in mean between two diverging populations.

It is important to stress the relationship between these concepts and fitness. Given that fitness (w) is directly related to selection, its mathematical relationship can be expressed as $f = a + \sum_{i=1}^{n}\beta_i z_i + e_i$ [57], and so it behaves as the weight of a multiple regression of f on the vector of phenotypes z.

In proteins, the definition of fitness is not trivial, and can vary depending on the hypothesis being tested. If the analysis is done comparatively (i.e. across different protein structures from different sources), a fitness analysis including exclusively structural measures, such as Gibbs free energy (ΔG), can be misleading. The fitness surface that can arise from this data would only represent departures from every individual native state. Nevertheless, ΔG and the energy of unfolding (ΔG°), are important measures to determine the stability of the protein which is important for the fitness of a protein structure. The stability of the structure allows it to perform a function and is therefore under selection because it is necessary for the particular biochemical function [68]. We are aware that there is a limitation to the protein structure stability role in fitness. To improve this fitness landscape, f can be defined by ΔG° coupled with a functional measure. In proteins, function is the main selective trait; therefore, including a term accounting for this would create a more realistic fitness surface. In enzymes this can be achieved by using the K_{cat}/K_M for each of the enzymes for a common substrate. The fitness function (F) can be expressed as:

$$F(i,s) = \Delta G_i^{\circ}\frac{K_{cat}^{i,s}}{K_M^{i,s}} \tag{11}$$

where ΔG_i° is the free energy of unfolding of the structure i, $K_{cat}^{i,s}$ is the turnover number for structure i in substrate s, and $K_M^{i,s}$ is the Michaelis constant of protein i working on substrate s.

In the case of the α-amylase family (GH13), one might try to apply the framework developed in previous sections and try to estimate the response to selection of a subset of them. However, Eq 11 cannot be applied since the information of the relative efficiency given a common substrate is not consistently available across all proteins in the dataset. For this reason we are going to work exclusively with $\Delta G^{\circ}_{unfold}$, keeping in mind two caveats, 1) that $\Delta G^{\circ}_{unfold}$ only represents structural stability and 2) that it has been shown that $\Delta G_{equilibrium}$ or $\Delta G^{\circ}_{unfold}$ are not optimized for during evolution [69].

Estimating dynamic and genetic variance-covariance matrices in the $\boldsymbol{\alpha}$-amylase dataset. The structure depicted with the higher fitness was model 1 of the Taka-amylase A structure (PDB code 2TAA; EC 3.2.1.1; α-1,4-glucan 4-glucanohydrolase; henceforth referred to with its PDB code) (S1 Fig), from *Aspergillus oryzae* assuming $\Delta\hat{G}^{\circ}$ as fitness. The model 1 of structure 2TAA can be assumed to be the result of the goal of selection. The realized response to selection $\Delta\bar{z}_{\varpi}$ can be defined as $\mu_{\oplus} - \mu_0$, where $\mu_{\oplus}$ is the target or after-selection mean structure and μ_0 is the starting or before-selection structure. To estimate $\Delta\bar{z}_{\varpi}$ it is essential to have the fitness defined based on the questions to be asked, given that the interpretation of the realized response to selection depends on it.

In an engineering perspective, let's assume that $\mu_{\oplus}$ is the mean of a population of structures with the desired stability. On the other hand, μ_0 is the mean of a population of structures created by a desired vector. One might ask the question of how does μ_0 have to change towards the stability of $\mu_{\oplus}$. This can be achieved by computing β_λ (Eq 10), and replacing $\Delta\bar{z}$ with $\Delta\bar{z}_{\varpi}$. In the particular case of the GH13 dataset, let's assume that the model 1 of the structure 2TAA is the desired phenotype (with the higher fitness in S1 Fig), and the model 643 of the α-amylase protein (PDB code: 4E2O; α-1,4-glucan-4-glucanohydrolase; EC 3.2.1.1; henceforth referred to by its PDB code) from *Geobacillus thermoleovorans* CCB_US3_UF5 (with the lower fitness in S1 Fig) corresponds to the source phenotype. We have selected the protein with the lowest stability (as $\Delta G^{\circ}_{unfold}$; S1 Fig), a functional yet truncated form of an α-amylase lacking both the N- and C-terminal domains [70], as the source phenotype and pose the hypothetical scenario in which we wish to improve its stability towards the more stable structure of the Taka-amylase A (PDB code 2TAA; EC 3.2.1.1; α-1,4-glucan 4-glucanohydrolase). The latter structure was shown to be the most stable (as per $\Delta G^{\circ}_{unfold}$; S1 Fig) of the set. In our framework, however, there is no requirement to use the extremes of the fitness distribution, as any gradient from the source to the target will suffice to asses the response to selection in that particular hypothesis.

In the posed scenario, β_λ would have a length corresponding to the dimensions of the shape. In the GH13 case 297 homologous residues were identified, which means that these shapes have a dimensionality of 891 traits. This dimension-per-dimension output is important since it reflects the amount of pressure in each dimension per each residue. However, it makes the visualization more difficult. For the sake of visualization simplicity, Fig 1 shows the absolute value of the sum of β_λ per residue, standardized from 0 to 1.

Fig 1A shows the selection gradient using the estimated *G*. Not surprisingly, the selection gradient for the TIM-barrel is low. This means that there is not much directional selection in this sub-structure. However, it is somewhat surprising that there is not any purifying selection

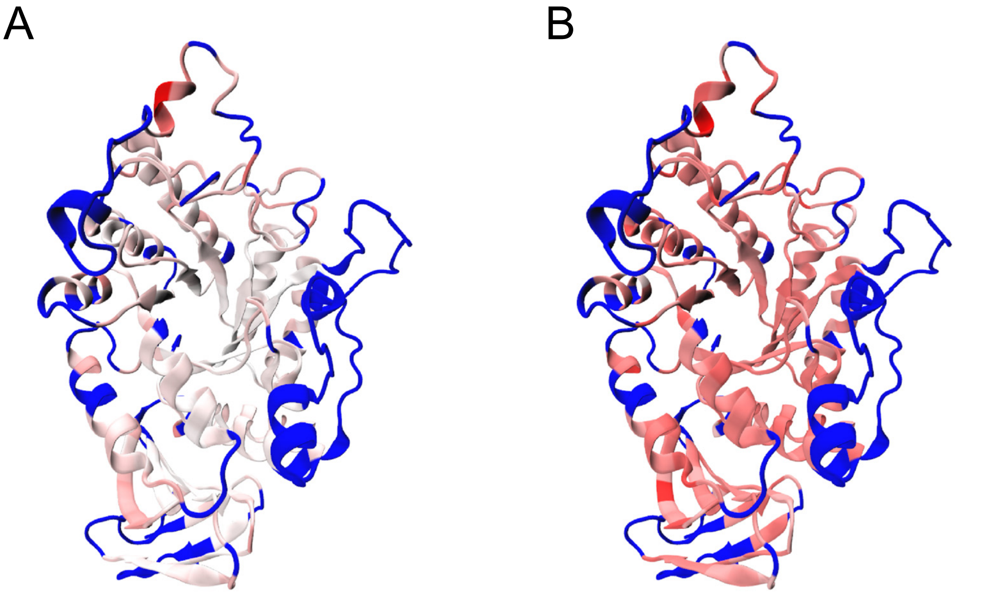

Fig 1. $\sum_{i=x,y,z}|\beta_{\lambda_i}|$ **rendered in the source structure 4E2O.** White represents the lowest magnitude (0), while red the highest (1). Blue depicts the non-homologous residues. A. Selection gradient on *G*. B. Dynamic gradient on *M*.

https://doi.org/10.1371/journal.pone.0196135.g001

Table 2. Selection gradient in the top 5 residues. Top panel shows the residues where at least one of its coordinates is under directional selection and the sum of their absolute values is the highest. Bottom panel contains the information of residues where at least one of its coordinates is under purifying selection, and the sum of the raw values are the lowest.

ResIndex	Residue	$\boldsymbol{\beta}_X$	$\boldsymbol{\beta}_Y$	$\boldsymbol{\beta}_z$	$\Delta\bar{z}_X$	$\Delta\bar{z}_Y$	$\Delta\bar{z}_Z$
				Directional			
112	TYR	-5.225	1.082	11.138	-5.106	2.043	10.248
122	LYS	12.333	-2.321	-0.964	12.452	-1.360	-1.854
124	ASP	14.28	-6.963	-10.036	14.399	-6.002	-10.926
125	TRP	18.001	-0.984	0.336	18.121	-0.022	-0.554
126	PHE	11.53	-0.833	3.253	11.650	0.128	2.363
				Purifying			
80	HIS	-5.580	-2.148	4.023	-5.461	-1.187	3.13
121	THR	2.508	-4.644	-5.731	2.627	-3.683	-6.621
223	TYR	-0.010	-7.631	-7.634	0.110	-6.670	-8.524
358	SER	-8.647	-3.461	1.963	-8.527	-2.500	1.073
394	GLU	-4.561	-0.449	-4.002	-4.442	0.512	-4.892

https://doi.org/10.1371/journal.pone.0196135.t002

either. This can be explained by the fixation of the trait in the evolution. Since the TIM-barrel is a widespread sub-structure that has been strongly selected during evolution, it might have reached a point of fixation of its geometry. Therefore, the G-matrix shows little covariation among these residues since the geometric variability is also low. It is important to stress here that the phenotype measured is the geometry of the structure more than that of the sequence. Therefore, despite some variation that may have occurred at the sequence level, it might not have meaningfully affected the positional information.

However, one must be cautious with the approach employed in Fig 1 since the signs are overlooked, thereby ignoring the direction of selection and the correlated response to selection. Nevertheless, this approach allows for a coarse-grained visual exploration of β_{λ_i}. Individual instances identified by this method should be analysed afterwards in each dimension. Table 2 shows the actual values of β_λ for the top 5 positive values (directional selection) and top 5 negative values (purifying selection).

Fig 1B and Table 3 show the mean difference between target and source when effects of correlated dynamic differentials are removed. Given that effectively G acts as a rotation matrix in Eq 10 to remove the selection differentials, one may posit that the same can be achieved with the dynamic (M) matrix. This concept is more difficult to interpret than the actual response to

Table 3. Dynamics gradient in the top 5 residues. Top panel shows the residues where at least one of its coordinates is under positive gradient. Bottom panel contains the information of residues where at least one of its coordinates is under a negative gradient.

ResIndex	Residue	$\boldsymbol{\beta}_X$	$\boldsymbol{\beta}_Y$	$\boldsymbol{\beta}_z$	$\Delta\bar{z}_X$	$\Delta\bar{z}_Y$	$\Delta\bar{z}_Z$
				Directional			
117	LEU	13.028	37.149	11.848	2.130	3.521	4.437
125	TRP	29.019	33.605	6.857	18.121	-0.022	-0.554
126	PHE	22.548	33.755	9.774	11.650	0.128	2.363
262	LYS	12.972	38.081	11.412	2.073	4.454	4.001
367	LEU	13.590	34.561	15.609	2.692	0.933	8.197
				Purifying			
124	ASP	25.297	27.625	-3.515	14.399	-6.002	-10.926
223	TYR	11.008	26.958	-1.113	0.110	-6.670	-8.524

https://doi.org/10.1371/journal.pone.0196135.t003

selection. Once G is replaced by M in Eq 10, we might call it *dynamic gradient* to differentiate it from the selection gradient already explained. In this case, if the gradient is zero for a given trait, this can be interpreted that the dynamic component of the phenotype does not contribute significantly to the difference in shape for that particular trait. In the case of non-zero gradients, these can be interpreted as contributions of the dynamics to the differential, either towards the target (positive gradient) or away from the target (negative gradient).

In the GH13 subset, most dynamic gradients were positive having only two residues that had one coordinate under a negative gradient (Table 3). This can also be inferred by Fig 1B. The values of the dynamic gradient are high but sensible given the definition of fitness. Since we defined fitness as the energy of unfolding (ΔG°), most of the information used to select the target and source structures comes from stability, and therefore thermodynamic information. The results depicted in Table 3 and Fig 1B suggest that most of the variation that explains the difference in phenotype between the structure 4E2O and 2TAA is contained within the molecular dynamic component rather than the approximation to the phylogenetic component.

Orientation of *G*. The GH13 θ was 1.4 degrees, which means that the direction of optimal response is 1.4 degrees away from the total genetic variation of 99% explained by the projection. According to this, the *Geobacillus thermoleovorans* structure is susceptible to the selection in the actual direction of the fitness landscape towards the structure of *Aspergillus oryzae* to achieve maximum stability. The extent of such change is given by $\Delta\bar{z}$, which means that the centroid position of the residue i should be displaced by $\vec{v} = (\Delta\bar{z}_{ix}, \Delta\bar{z}_{iy}, \Delta\bar{z}_{iz})$.

In the case of the dynamics, the same approach can be taken. Here, θ_M was 1.5 degrees which means that the optimal dynamic response is 1.5 degrees away from the optimal response. This can be interpreted in a similar way as that of the regular θ. However, manipulating the structure along the dynamics gradient is not feasible.

The GH13 dataset $\theta_{\Delta\bar{z}-\beta}$ was 0.3. This means that the genetic constraints on 4E2O are not affecting the direction of selection. This posits the possibility that a strong directional selection will drive the source structure towards the target. The same pattern happens when this approach is applied to M. $\theta^{M}_{\Delta\bar{z}-\beta}$ is 1.46 degrees, which is almost identical to θ_M. Thus, there are almost no within-variation or dynamic constraints to the vector of response given the dynamic gradient.

Concluding remarks

We have introduced the application of the approximation of comparative quantitative genetics framework, by means of a pooled-within group covariance matrix in a subset of the GH13 proteins, and demonstrated this application is feasible and provides sensible results, given the definition of fitness. This definition is essential in the interpretation of the results since it is the interpretation that gives polarity to R_{ϖ}. Therefore, all conclusions about the response to selection and the selection gradient itself must be analyzed under this light.

The usage of M in the determination of the dynamic gradient could be controversial. This is due to the fact that, in the partition of the phenotypic variance, M is expected to be the environmental variance plus an error term. However, since the source data for the estimation of G and M come from repeated measures by MD, M contains information about the thermodynamics and folding stability of the protein. It is therefore also contributing to selection.

It is important to stress the fact that this is an approximation to the true G and true M, since we have shown in previous sections that these cannot be estimated given the dimensionality of the phenotype. However, we have shown that the pooled-within group approach gives consistent results.

We have also shown that, in a stability perspective, the TIM-barrel shows a small phylogenetic/genetic component to the selection gradient when a less stable structure (4E2O) is analyzed with respect to a more stable structure (2TAA). In an engineering perspective, this means that most of the changes in shape come from the dynamics. Nevertheless, the small $\theta_{\Delta\bar{z}-\beta}$ shows that most of the changes applied to 4E2O would directly result in increasing the stability towards the structure 2TAA. 4E2O is a truncated protein, and therefore some loss of stability is expected. It seems that residues 112Y, 122K, 124D, 125W, and 126P, are good candidates to increase the stability of the molecule given their $\Delta\hat{z}$s. In these cases, the goal will be to shift the position of their centroids by the resulting vector of the three dimensions.

Ultimately, we have shown a framework to analyse protein structures' response to selection. This framework have incredible potential in industry (protein engineering), structural biology, and evolutionary biology, since allow us to narrow the search space within a given fitness landscape and potentially predict the extent of the propensity of a protein structure to be selected towards a desired target.

Supporting information

S1 File. Supplementary note. Supplementary methods (A) and results (B) testing the feasibility of the of the linear mixed model and Bayesian mixed model implementations.
(PDF)

S1 Fig. Fitness surface of the MD simulations in the 34 structures dataset. Fitness in the Z axis is defined as $\Delta\hat{G}^{\circ}$.
(PNG)

Acknowledgments

The authors thank the members of the Blouin Lab for helpful comments and critical review of this manuscript. We also thank Jitka M. Krejci for the editorial and language revision of the manuscript. This study was funded by NSERC through the grant No. 120504858. This work was partially supported by the Departamento Administrativo de Ciencia y Tecnología—Colciencias (Colombia) through the CALDAS scholarship.

Author Contributions

Conceptualization: Jose Sergio Hleap, Christian Blouin.

Data curation: Jose Sergio Hleap.

Formal analysis: Jose Sergio Hleap.

Funding acquisition: Christian Blouin.

Investigation: Jose Sergio Hleap.

Methodology: Jose Sergio Hleap.

Project administration: Christian Blouin.

Resources: Christian Blouin.

Software: Jose Sergio Hleap.

Supervision: Christian Blouin.

Validation: Jose Sergio Hleap.

Visualization: Jose Sergio Hleap.

Writing – original draft: Jose Sergio Hleap.

Writing – review & editing: Jose Sergio Hleap, Christian Blouin.

References

1. MacGregor EA, Janeček Š, Svensson B. Relationship of sequence and structure to specificity in the α-amylase family of enzymes. Biochimica et Biophysica Acta (BBA)—Protein Structure and Molecular Enzymology. 2001; 1546(1):1–20. https://doi.org/10.1016/S0167-4838(00)00302-2
2. Ben Ali M, Khemakhem B, Robert X, Haser R, Bejar S. Thermostability enhancement and change in starch hydrolysis profile of the maltohexaose-forming amylase of Bacillus stearothermophilus US100 strain. Biochem J. 2006; 394(Pt 1):51–6. https://doi.org/10.1042/BJ20050726 PMID: 16197365
3. Janeček Š, Svensson B, Henrissat B. Domain evolution in the α-amylase family. Journal of molecular evolution. 1997; 45(3):322–331. https://doi.org/10.1007/PL00006236 PMID: 9302327
4. Fort J, Laura R, Burghardt HE, Ferrer-Costa C, Turnay J, Ferrer-Orta C, and Usón I, Zorzano A, Fernández-Recio J, Orozco M, Lizarbe MA, Palacín A. The structure of human 4F2hc ectodomain provides a model for homodimerization and electrostatic interaction with plasma membrane Journal of Biological Chemistry. 2007; 282(43): 31444–31452 PMID: 17724034
5. Henrissat B. A classification of glycosyl hydrolases based on amino acid sequence similarities. Biochemical Journal. 1991; 280(2):309–316. https://doi.org/10.1042/bj2800309 PMID: 1747104
6. Takata H, Kuriki T, Okada S, Takesada Y, Iizuka M, Minamiura N, et al. Action of neopullulanase. Neopullulanase catalyzes both hydrolysis and transglycosylation at alpha-(1—-4)-and alpha-(1—-6)-glucosidic linkages. Journal of Biological Chemistry. 1992; 267(26):18447–18452. PMID: 1388153
7. Jespersen HM, Ann MacGregor E, Henrissat B, Sierks MR, Svensson B. Starch-and glycogen-debranching and branching enzymes: prediction of structural features of the catalytic (β/α) 8-barrel domain and evolutionary relationship to other amylolytic enzymes. Journal of protein chemistry. 1993; 12 (6):791–805. https://doi.org/10.1007/BF01024938 PMID: 8136030
8. Nakajima R, Imanaka T, Aiba S. Comparison of amino acid sequences of eleven different α-amylases. Applied Microbiology and Biotechnology. 1986; 23(5):355–360. https://doi.org/10.1007/BF00257032
9. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic acids research. 2009; 37 (suppl 1):D233–D238. https://doi.org/10.1093/nar/gkn663 PMID: 18838391
10. Janecek S. New conserved amino acid region of alpha-amylases in the third loop of their (beta/alpha) 8-barrel domains. Biochemical Journal. 1992; 288(Pt 3):1069. https://doi.org/10.1042/bj2881069 PMID: 1471979
11. Janeček Š. Sequence Similarities and Evolutionary Relationships of Microbial, Plant and Animal α-amylases. The FEBS Journal. 1994; 224(2):519–524.
12. Svensson B. Protein engineering in the α-amylase family: catalytic mechanism, substrate specificity, and stability. Plant molecular biology. 1994; 25(2):141–157. https://doi.org/10.1007/BF00023233 PMID: 8018865
13. Uitdehaag JC, Mosi R, Kalk KH, van der Veen BA, Dijkhuizen L, Withers SG, et al. X-ray structures along the reaction pathway of cyclodextrin glycosyltransferase elucidate catalysis in the α-amylase family. Nature Structural & Molecular Biology. 1999; 6(5):432–436. https://doi.org/10.1038/8235
14. Machovič M, Janeček Š. The invariant residues in the α-amylase family: just the catalytic triad. Biologia. 2003; 58(6):1127–1132.
15. Svensson B, Janeček v. Glycoside Hydrolase Family 13; 2014. CAZypedia: http://www.cazypedia.org/.
16. Terrapon N, Lombard V, Drula E, Coutinho PM, Henrissat B. The CAZy Database/the Carbohydrate-Active Enzyme (CAZy) Database: Principles and Usage Guidelines. In: A Practical Guide to Using Glycomics Databases. Springer; 2017. p. 117–131.
17. Stam MR, Danchin EGJ, Rancurel C, Coutinho PM, Henrissat B. Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of alpha-amylase-related proteins. Protein Eng Des Sel. 2006; 19(12):555–62. https://doi.org/10.1093/protein/gzl044 PMID: 17085431
18. Janeček Š, Svensson B, MacGregor EA. α-Amylase: an enzyme specificity found in various families of glycoside hydrolases. Cellular and molecular life sciences. 2014; 71(7):1149–1170. https://doi.org/10.1007/s00018-013-1388-z PMID: 23807207
19. Janeček Š. How many conserved sequence regions are there in the α-amylase family. Biologia. 2002; 57(Suppl 11):29–41.

20. MacGregor EA. An overview of clan GH-H and distantly related families. Biologia. 2005; 60(Suppl 16):5–12.

21. El Kaoutari A, Armougom F, Gordon JI, Raoult D, Henrissat B. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. Nature reviews Microbiology. 2013; 11(7):497–504. https://doi.org/10.1038/nrmicro3050 PMID: 23748339

22. Zhang Q, Han Y, Xiao H. Microbial α-amylase: A biomolecular overview. Process Biochemistry. 2017; 53:88–101. https://doi.org/10.1016/j.procbio.2016.11.012

23. Janeček Š, Gabriško M. Remarkable evolutionary relatedness among the enzymes and proteins from the α-amylase family. Cellular and Molecular Life Sciences. 2016; 73(14):2707–2725. https://doi.org/10.1007/s00018-016-2246-6 PMID: 27154042

24. Božić N, Lončar N, Slavić MŠ, Vujčić Z. Raw starch degrading α-amylases: an unsolved riddle. Amylase. 2017; 1(1):12–25.

25. Gupta R, Gigras P, Mohapatra H, Goswami VK, Chauhan B. Microbial α-amylases: a biotechnological perspective. Process Biochemistry. 2003; 38(11):1599–1616. https://doi.org/10.1016/S0032-9592(03)00053-0

26. Bothast RJ, Schlicher MA. Biotechnological processes for conversion of corn into ethanol. Appl Microbiol Biotechnol. 2005; 67(1):19–25. https://doi.org/10.1007/s00253-004-1819-8 PMID: 15599517

27. Lu Z, Wang Q, Jiang S, Zhang G, Ma Y. Truncation of the unique N-terminal domain improved the thermos-stability and specific activity of alkaline α-amylase Amy703. Scientific reports. 2016; 6:22465. https://doi.org/10.1038/srep22465 PMID: 26926401

28. Dey TB, Kumar A, Banerjee R, Chandna P, Kuhad RC. Improvement of microbial α-amylase stability: strategic approaches. Process Biochemistry. 2016; 51(10):1380–1390. https://doi.org/10.1016/j.procbio.2016.06.021

29. Tang SY, Le QT, Shim JH, Yang SJ, Auh JH, Park C, et al. Enhancing thermostability of maltogenic amylase from Bacillus thermoalkalophilus ET2 by DNA shuffling. FEBS Journal. 2006; 273(14):3335–3345. https://doi.org/10.1111/j.1742-4658.2006.05337.x PMID: 16857016

30. Ghollasi M, Ghanbari-Safari M, Khajeh K. Improvement of thermal stability of a mutagenised α-amylase by manipulation of the calcium-binding site. Enzyme and microbial technology. 2013; 53(6):406–413. https://doi.org/10.1016/j.enzmictec.2013.09.001 PMID: 24315644

31. Ranjani V, Janeček Š, Chai KP, Shahir S, Rahman RNZRA, Chan KG, et al. Protein engineering of selected residues from conserved sequence regions of a novel Anoxybacillus α-amylase. Scientific reports. 2014; 4:5850. https://doi.org/10.1038/srep05850 PMID: 25069018

32. Li C, Du M, Cheng B, Wang L, Liu X, Ma C, et al. Close relationship of a novel Flavobacteriaceae α-amylase with archaeal α-amylases and good potentials for industrial applications. Biotechnology for biofuels. 2014; 7(1):18. https://doi.org/10.1186/1754-6834-7-18 PMID: 24485248

33. André I, Potocki-Véronèse G, Barbe S, Moulis C, Remaud-Siméon M. CAZyme discovery and design for sweet dreams. Current opinion in chemical biology. 2014; 19:17–24. https://doi.org/10.1016/j.cbpa.2013.11.014 PMID: 24780275

34. Chen A, Li Y, Nie J, McNeil B, Jeffrey L, Yang Y, et al. Protein engineering of Bacillus acidopullulyticus pullulanase for enhanced thermostability using in silico data driven rational design methods. Enzyme and microbial technology. 2015; 78:74–83. https://doi.org/10.1016/j.enzmictec.2015.06.013 PMID: 26215347

35. Reetz MT, Carballeira JD. Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. Nature protocols. 2007; 2(4):891–903. https://doi.org/10.1038/nprot.2007.72 PMID: 17446890

36. Liu L, Deng Z, Yang H, Li J, Shin Hd, Chen RR, et al. In silico rational design and systems engineering of disulfide bridges in the catalytic domain of an alkaline α-amylase from Alkalimonas amylolytica to improve thermostability. Applied and environmental microbiology. 2014; 80(3):798–807. https://doi.org/10.1128/AEM.03045-13 PMID: 24212581

37. Dehouck Y, Kwasigroch JM, Gilis D, Rooman M. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC bioinformatics. 2011; 12(1):151. https://doi.org/10.1186/1471-2105-12-151 PMID: 21569468

38. Klingenberg CP, Leamy LJ. Quantitative genetics of geometric shape in the mouse mandible. Evolution. 2001; 55(11):2342–2352. https://doi.org/10.1111/j.0014-3820.2001.tb00747.x PMID: 11794792

39. Thompson R. Estimation of quantitative genetic parameters. Proceedings of the Royal Society B: Biological Sciences. 2008; 275(1635):679–686. https://doi.org/10.1098/rspb.2007.1417 PMID: 18211869

40. Lande R, Arnold SJ. The measurement of selection on correlated characters. Evolution. 1983; p. 1210–1226. https://doi.org/10.1111/j.1558-5646.1983.tb00236.x PMID: 28556011

41. Klingenberg CP. Quantitative genetics of geometric shape: heritability and the pitfalls of the univariate approach. Evolution. 2003; 57(1):191–195. https://doi.org/10.1554/0014-3820(2003)057%5B0191:QGOGSH%5D2.0.CO;2 PMID: 12643582

42. Blasco A. The Bayesian controversy in animal breeding. Journal of Animal Science. 2001; 79(8):2023–2046. https://doi.org/10.2527/2001.7982023x PMID: 11518211

43. Lynch M. Methods for the analysis of comparative data in evolutionary biology. Evolution. 1991; p. 1065–1080. https://doi.org/10.1111/j.1558-5646.1991.tb04375.x PMID: 28564168

44. Hadfield JD, Nakagawa S. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. Journal of evolutionary biology. 2010; 23(3):494–508. https://doi.org/10.1111/j.1420-9101.2009.01915.x PMID: 20070460

45. Cheverud JM, Dow MM, Leutenegger W. The quantitative assessment of phylogenetic constraints in comparative analyses: sexual dimorphism in body weight among primates. Evolution. 1985; p. 1335–1351. https://doi.org/10.2307/2408790 PMID: 28564267

46. Housworth EA, Martins EP, Lynch M. The phylogenetic mixed model. The American Naturalist. 2004; 163(1):84–96. https://doi.org/10.1086/380570 PMID: 14767838

47. Freckleton RP, Harvey PH, Pagel M. Phylogenetic analysis and comparative data: a test and review of evidence. The American Naturalist. 2002; 160(6):712–726. https://doi.org/10.1086/343873 PMID: 18707460

48. Pagel M. Inferring the historical patterns of biological evolution. Nature. 1999; 401(6756):877–884. https://doi.org/10.1038/44766 PMID: 10553904

49. Sorensen D, Gianola D. Likelihood, Bayesian and MCMC methods in quantitative genetics. Springer; 2002.

50. Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. How vague is vague?: A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. Statistics in medicine. 2005; 24(15):2401–2428. https://doi.org/10.1002/sim.2112 PMID: 16015676

51. Hess B, Kutzner C, van der Spoel D, Lindahl E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. Journal of chemical theory and computation. 2008; 4 (3):435–447. https://doi.org/10.1021/ct700301q PMID: 26620784

52. Menke M, Berger B, Cowen L. Matt: local flexibility aids protein multiple structure alignment. PLoS Comput Biol. 2008; 4(1):e10. https://doi.org/10.1371/journal.pcbi.0040010 PMID: 18193941

53. Dryden IL. Shapes package. R Foundation for Statistical Computing, Vienna Contributed package. 2011;.

54. Adams DC, Naylor GJP. A Comparison of Methods for Assessing the Structural Similarity of Proteins. In: Mathematical Methods for Protein Structure Analysis and Design; 2003. p. 109–115.

55. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucleic Acids Research. 2005; 33(suppl 2):W382–W388. https://doi.org/10.1093/nar/gki387 PMID: 15980494

56. Arnold SJ. Constraints on phenotypic evolution. American Naturalist. 1992; 61:S85–S107. https://doi.org/10.1086/285398

57. Blows M, Walsh B. Spherical Cows Grazing in Flatland: Constraints to Selection and Adaptation. In: van der Werf J, Graser HU, Frankham R, Gondro C, editors. Adaptation and fitness in animal populations. Springer; 2009. p. 83–101. Available from: http://dx.doi.org/10.1007/978-1-4020-9005-9_6.

58. Hansen TF, Houle D. Measuring and comparing evolvability and constraint in multivariate characters. J Evol Biol. 2008; 21(5):1201–19. https://doi.org/10.1111/j.1420-9101.2008.01573.x PMID: 18662244

59. Walsh B, Blows MW. Abundant genetic variation+ strong selection = multivariate genetic constraints: a geometric view of adaptation. Annual Review of Ecology, Evolution, and Systematics. 2009; 40:41–59. https://doi.org/10.1146/annurev.ecolsys.110308.120232

60. Marroig G, Cheverud JM. A comparison of phenotypic variation and covariation patterns and the role of phylogeny, ecology, and ontogeny during cranial evolution of New World monkeys. Evolution. 2001; 55 (12):2576–2600. https://doi.org/10.1111/j.0014-3820.2001.tb00770.x PMID: 11831671

61. Cheverud JM. Quantitative genetic analysis of cranial morphology in the cotton-top (Saguinus oedipus) and saddle-back (S. fuscicollis) tamarins. Journal of Evolutionary Biology. 1996; 9(1):5–42. https://doi.org/10.1046/j.1420-9101.1996.9010005.x

62. Revell LJ, Harmon LJ, Langerhans RB, Kolbe JJ. A phylogenetic approach to determining the importance of constraint on phenotypic evolution in the neotropical lizard Anolis cristatellus. Evolutionary Ecology Research. 2007; 9(2):261–282.

63. Cheverud JM, Marroig G. Comparing covariance matrices: random skewers method compared to the common principal components model. Genetics and Molecular Biology. 2007; 30(2):461–469. https://doi.org/10.1590/S1415-47572007000300027

64. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution. 2012; 3(2):217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x

65. Prôa M, O'Higgins P, Monteiro LR. Type I error rates for testing genetic drift with phenotypic covariance matrices: a simulation study. Evolution. 2013; 67(1):185–95. https://doi.org/10.1111/j.1558-5646.2012.01746.x PMID: 23289571

66. Haber A. The Evolution of Morphological Integration in the Ruminant Skull. Evolutionary Biology. 2015; 42(1):99–114. https://doi.org/10.1007/s11692-014-9302-7

67. Rausher MD. The measurement of selection on quantitative traits: biases due to environmental covariances between traits and fitness. Evolution. 1992; p. 616–626. https://doi.org/10.1111/j.1558-5646.1992.tb02070.x PMID: 28568666

68. Bloom JD, Wilke CO, Arnold FH, Adami C. Stability and the evolvability of function in a model protein. Biophysical Journal. 2004; 86(5):2758–2764. https://doi.org/10.1016/S0006-3495(04)74329-5 PMID: 15111394

69. Alfaro JA. Capturing the dynamics of protein sequence evolution through site-independent structurally constrained phylogenetic models. Department of biochemitry & molecular biology, Dalhousie University. Halifax, Canada; 2014.

70. Mok S-C, Teh A-H, Saito JA, Najimudin N, Alam M. Crystal structure of a compact α-amylase from *Geobacillus thermoleovorans*. Enzyme and microbial technology. 2013; 53(1): 46–54. https://doi.org/10.1016/j.enzmictec.2013.03.009 PMID: 23683704