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

Jose Sergio Hleap and Christian Blouin

Abstract—The Glycoside Hydrolase Family 13 (GH13) proteins have been extensively studied given their importance in industry and evolutionary studies. Despite of being a multicatalytic family, they mainly performed the hydrolysis of starch into smaller carbohydrates. Evolutionarily, they are of interest because of the TIM-barrel fold that all these proteins share. Industrially, they have been subjected to many bioengineering process to improve their function. Here, we introduce a framework to analyse the response to selection of GH13 protein structures given some phylogenetic and dynamic information. We found that the TIM-barrel is not selectable (for thermodynamic stability) since it has been under purifying selection during its recent evolution. We also show a method to rank important residues capable of altering towards an optimum. Overall, this paper demonstrate the feasibility of a framework for the analysis of protein structures in a given fitness landscape.

Index Terms—Geometric analysis, homolog variation, relative deformation, protein structure engineering, protein structure evolution.

1 Introduction

The Glycoside Hydrolase Family 13 (GH13) is a multi-reaction catalytic family [1] acting primarily on α-glucoside linkages of starch. Many industrial applications use members of this family as catalysts [2]. Biologically, it is also of interest since all of its members share a symmetrical TIM-barrel ($(\beta/\alpha)_8$) catalytic domain which is versatile and taxonomically widespread [3].Whether the ubiquity of this fold is a product of convergent or divergent evolution is a matter of debate.However, there is more evidence towards the divergent evolution hypothesis [4].

The GH13 family is of evolutionary interest in the research of plastic folds and how selection might act upon them. Exploring how selection act on protein structures is not a trivial problem. Other fields have been used in organismal biology to explore similar questions . Comparative quantitative genetics, the study of inheritance at the phenotypic level in phylogenetically related organisms [5] is one of them.

This framework is applied to multigenic single traits. That is, a particular trait encoded by a number of different genes that needs to interact in order to "form" the trait (i.e. length of a femur, number of abdominal bristles in an insect, milk production in cattle, etc).

Assuming that protein structures are phenotypes and that their 3D structures respond to both

- *J. S. Hleap was with the Department of Biochemistry & Molecular Biology, Dalhousie University, Halifax, Canada. E-mail: jshleap@dal.ca*
- *Christian Blouin is with the departments of Computer Science and Biochemistry & Molecular Biology, Dalhousie University, Halifax, Canada.*

Manuscript received XXX, 2016; revised XXXXX XX, 2016.

0000–0000/00/$00.00 © 2016 IEEE Published by the IEEE Computer Society

genetics and environment, we can treat them as shape phenotypes. Proteins and other shapes are highly multivariate in nature [6]. To comply with such nature the phenotype can be expressed as [7]:

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

where y represents a vector of observations of multiple traits, 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.

The additive variance (V_A) is the actual contribution from genes to the phenotype. V_A is also the portion of the genetic variance that can be passed from parents to offspring, and be selected. The direct relationship between V_A and selection is expressed in the *breeder's equation* [8]:

$$R = \frac{V_A}{V_P} * S \tag{2}$$

where R is the response to selection of a population of a given trait, S is the selection differential, and $\frac{V_A}{V_P}$ is also called narrow-sense heritability (h^2). Lande and Arnold [9] generalized this for a multivariate case as:

$$\Delta\bar{z} = G\beta \tag{3}$$

where $\Delta\bar{z}$ is a vector of changes in traits, G is the genetic variance-covariance matrix (G-matrix), and β is a vector of selection gradients.

To apply the aforementioned framework, the estimation of a G-matrix is required [10]. To deal with the fact that the number of samples is limited, that inversion of matrices require expensive computation, and that an eigen decomposition of the covariance matrices is required, the restricted maximum likelihood (REML) approach is tipically used to carry out the variance decomposition. When applied to univariate data, it is more accurate since it deals better with 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 is a more robust estimation than REML, with equivalent results in less complex cases [11]. 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, can lead to biases in the estimation.

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

Lynch [12] developed the *phylogenetic mixed model (PMM)*. In this model, the correlation of phylogenetically heritable components is the time to the shared common ancestor (length of the path from the most recent common ancestor among two species and the root) in the phylogeny [13]. The PMM can be described as [12]:

$$\bar{z}_{ci} = \mu_c + a_{ci} + e_{ci} \tag{4}$$

where $\bar{z}_{ci}$ is the observed mean of the trait, μ_c is the grand mean of the cth character over the phylogeny, a_{ci} is the additive (phylogenetic) value the cth character in the ith taxon. Finally, e_{ci} is the residual deviation or Cheverud's specific effect [14] if within species data has been sampled.

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 analysing truly homologous protein structures, since the mean effect on the phenotype is shared by common ancestry. This also means that $\mu_c + a_{ci}$ can be interpreted as the heritable component of the mean phenotype for the ith taxon [12].

The univariate model in equation 4 can be generalized to any number of characters by [12]:

$$\bar{z} = X\mu + a + e \tag{5}$$

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

Here, the phylogenetic effects are the portion of the variation that has been inherited from

ancestral species [14]. It does not only contain the genetic component, but also some environmental contributions given the shared evolutionary history of the taxa [15]. In PMM the ratio between the additive component and the total variance is the heritability (h^2) in an univariate approach. Housworth et al. [15] pointed out that an univariate h^2 in a PMM is actually equivalent to Freckleton et al. [16]'s and Pagel [17]'s phylogenetic correlation (λ).

Martins et al. [18] showed that the PMM is comparable to the phylogenetic generalized least squares (PGLS), spatial auto-regression, and phylogenetic eigenvector regression. They all yield good statistical performance, regardless of the evolutionary model, and even when some of the assumptions where violated.

Despite the robustness of the models, the REML technique 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 [13]. 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 [13]. A discussion of the use of Bayesian MCMC techniques is beyond the scope of this manuscript. We refer the interested reader to Sorensen and Gianola [19] 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 [20]. 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 framework (CQG) in protein structures, we simulated a dataset with variable number 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 is feasible and accurate. By applying of the energy of unfolding (ΔG°) as fitness function to the GH13 family we show how purifying selection have acted on the TIM-barrel and therefore fixing its geometry. We also show how by changing the fitness function the response to selection propensity change accordingly. We also show a proxy for the amount of dynamic deformation happening in the protein given a vector of selection. Overall, we present here a starting framework to explore protein structure evolution and design.

2 METHODS

2.1 GH13 dataset

Given that molecular dynamic simulations are very time consuming, we used a subset of the proteins classified as Glycoside Hydrolases Family 13 (GH13). We randomly selected 35 protein structures from a possible set of 386, but one failed during the MD simulation. A final set of 34 protein structures (Table 1) was used in further analyses.

2.2 Molecular Dynamics (MD) simulations

Each of the 34 protein structure was simulated in solution using the software `GROMACS 4` [21]. The force field modes used for the simulations were GROMOS96 for the protein, and the 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^{TM} GPU.

TABLE 1
Subset of the α-Amylase dataset. PDB codes, taxonomic information, Enzyme comission code and reference of the 34 structures used to estimate the pooled-within covariance matrix

PDB code	Species	EC code	Mutation	Reference
1CGY	*Bacillus circulans*	2.4.1.19	Y195W	[22]
1E3X	*Bacillus amyloliquefaciens*	2.2.1.1	Chimeric	[23]
1G5A	*Neisseria polysaccharea*	2.4.1.4	None	[24]
1GVI	*Thermus* sp.	3.2.1.54	None	[25]
1J0H	*Geobacillus stearothermophilus* TRS40	3.2.1.135	None	[26]
1KB3	*Homo sapiens*	3.2.1.1	R195A	[27]
1KXH	*Pseudoalteromonas haloplanctis*	3.2.1.1	D174N	[28]
1M53	*Klebsiella* sp. LX3	5.4.99.11	None	[29]
1SMA	*Thermus* sp. IM6501	3.2.1.133	None	[30]
1TMQ	*Tenebrio molitor*	3.2.1.1	None	[31]
1UA7	*Bacillus subtilis*	3.2.1	N356Q	[32]
1UD3	*Bacillus* sp. KSM-K38	3.2.1.1	N289H	[33]
1VJS	*Bacillus licheniformis*	3.2.1.1	None	[34]
1W9X	*Bacillus halmapalus*	3.2.1.1	None	[35]
1WZL	*Thermoactinomyces vulgaris* R-47	3.2.1.135	R469L	[36]
1ZJA	*Pseudomonas mesoacidophila*	5.4.99.11	None	[37]
2DIE	*Bacillus* sp. KSM-1378	3.2.1.1	None	[38]
2FH8	*Enterobacter aerogenes*	3.2.1.41	G680L/V882L	[39]
2TAA	*Aspergillus oryzae*	3.2.1.1	None	[40]
2WAN	*Bacillus acidopullulyticus*	3.2.1.41	None	[41]
2Y4S	*Hordeum vulgare*	3.2.1.41	None	[42]
2Z1K	*Thermus thermophilus HB8*	3.2.1.41	NA	NA
2ZE0	*Geobacillus* sp. HTA-462	3.2.1.20	None	[43]
2ZIC	*Streptococcus mutans*	3.2.1.70	N536L	[44]
3AXH	*Saccharomyces cerevisiae*	2.1.1.64/3.2.1.10	E277A	[45]
3CZK	*Xanthomonas axonopodis* pv. glycines	3.2.1.48	E322Q	[46]
3DC0	*Bacillus* sp. KR-8104	3.2.1.1	NA	NA
3EDE	*Flavobacterium* sp. 92	3.2.1.54	T49P	[47]
3GBD	*Serratia plymuthica*	5.4.99.11	None	[48]
3UEQ	*Neisseria polysaccharea*	2.4.1.4	None	[49]
3VM5	*Oryzias latipes*	3.2.1.1	None	[50]
3VM7	*Malbranchea cinnamomea*	3.2.1.73	None	[51]
4E2O	*Geobacillus thermoleovorans* CCB_US3_UF5	3.2.1.1	None	[52]
4GI6	*Rhizobium* sp. MX-45	5.4.99.11	F164L	[53]

*EC number derived from enzyme name using BRENDA [54]

† Inactive mutant

‡ Thermostable α-amylase

II A588 CYS Modelled as oxidised CYS (CSX)

⊎ Truncated

NA Not provided by the RCSB PDB or secondary sources

2.3 Aligning the structures and MD simulations

The alignment of homologous proteins was performed using `MATT` structural aligner [55]. To align the MD simulations a General Procrustes Superimposition was performed using the `R` package `shapes` [56].

2.4 Abstracting protein structures as shapes

On a set of aligned protein structures, the abstraction is performed similar to that in Adams and Naylor [57], [58]. However, those papers do not fully describe the way the abstraction is made. Here we assign a landmark to the residues' centroids defined by (x,y,z):

$$(\frac{1}{A}\sum_{j=1}^{A} X_j, \frac{1}{A}\sum_{j=1}^{A} Y_j, \frac{1}{A}\sum_{j=1}^{A} Z_j) \tag{6}$$

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 itself.

2.5 Pooled-within group covariance matrix estimation

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

1) Align every model within each MD simulation using GPS: Remove extra rotations and translations that could occur during MD simulation.
2) Select an ambassador structure that is closest to the mean structure.
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 the fully homologous sites (gapless columns): Identify shared information among all structures.
5) Concatenate the centroid information for all trajectories
6) Perform a GPS on the entire set of shapes: Bring all pre-aligned structures into the same reference plane.
7) Compute the pooled-within covariance matrix (W) by first computing the deviation from the mean in each class/group (individual homologs in our case):

$$D_k = x_k(\omega) - \bar{x}_{k,s} \tag{7}$$

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

$$F_{l,m} = \sum_{\omega: f(\omega)=s} [D_l] \times [D_m] \tag{8}$$

and finally compute the the pooled-within covariance matrix W by:

$$W = \frac{1}{n-S}\sum_{s=1}^{S} (F_{i,j})_{i,j=1,\cdots,p} \tag{9}$$

where S is a the number of categorical variables describing the groups or species, ω is an instance, were $f(\omega)$ correspond to the class value of the instance, and $\bar{x}_{i,s}$ is the mean of the variable i for individuals belonging to s. 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 P and W.

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

The $\Delta G^{\circ}_{unfold}$ on each model for each protein using was estimated using the command line version of `FoldX` [59]. 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 $\Delta \hat{G^{\circ}_{unfold}} = \frac{\Delta G^{\circ}_{unfold}}{n}$, n 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 of a PCA analysis of the shapes as X and Y axes; $\Delta G^{\hat{\circ}}_{unfold}$ in the Z axis (Supplementary figure A1).

2.7 Propensity to respond to selection

Arnold [60] 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 [61] and Hansen and Houle [62] an approach to measure the angle between β and the predicted response to selection from the multivariate breeders equation, $\Delta\bar{z}$ as:

$$\theta_{\Delta\bar{z}\text{-}\beta} = cos^{-1}\left(\frac{\Delta\bar{z}^T\beta_\lambda}{\sqrt{\Delta\bar{z}\Delta\bar{z}^T}\sqrt{\beta_\lambda\beta_\lambda^T}}\right) \quad (10)$$

$\theta_{\Delta\bar{z}\text{-}\beta}$ would be zero when there is no genetic constraint, whereas an angle of 90° would represent an absolute constraint [63].

All these calculations were computed with in-house python scripts.

3 Results and Discussion

In supplementary materials A and B we have shown that the traditional PMM models and its 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 taking into account. Here, we applied a simple method to overcome this over-parametrization.

4 Overcoming over-parametrization: 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 infeasible. 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 [64]. 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. Because of 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. I have shown the infeasibility of the GLMM to deal with the dimensionality and very large sample size. However, it has been shown that phenotypic (P) matrices can be estimated with more confidence with large sample sizes [65], [66]. It is also shown that in some cases, P can be used as surrogate for G when the two are proportional [64], [67]. To test this, we performed a shape simulation explained in supplementary section A.1. 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 Methods section.

Table 2 show the feasibility and accuracy of the pooled-within species covariance estimation method. Even with highly multivariate data (1024 traits), the memory requirement is manageable (less than 2 Gb). The evaluation is completed in under an hour. The accuracy of the estimation is high, with the estimated G matrix being almost identical to the simulated one, and the estimated MD having over 0.97 correlated response to random vectors to the actual MD. This is a surprising result since this method cannot separate completely the error terms from the genetic and the dynamic component. However, it seems that the split of the error term between the two other components make the error terms negligible. Moreover, it seems that error does not affect significantly the structure of G and MD, allowing them to behave almost identically than the simulated counterparts. Given these results, and the fact that the application to real datasets can only be made with this approach, it seems reasonable to keep using this from this point for-

TABLE 2
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

ward. 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.

4.1 Meaning of the pooled within-structure covariance matrix

It has been widely described that P-matrices can be used as surrogates of G-matrices in cases were they are proportional or sufficiently similar [68]. Proa et al. [69] showed that this assumption can be relaxed if the correlation between G and P is higher than 0.6. In protein structures, we can assume that given the strong selective pressures and long divergence times, the relationship between P and G be standardized. Assuming that this is true in protein structures, the estimated pooled variance-covariance (V/CV) matrices in real datasets might have specific biological meaning. This has been described in Haber [70] for morphological integration in mammals. Following Haber's [70] logic, the translation of the pooled V/CV matrix could be as follows:

1) The within-structure/species (i.e. thermodynamic V/CV matrix) refers to integration of residues in a thermodynamical and functional manner, and also contains information about environmental factors affecting the physical-chemistry of the structure. Haber [70] includes a genetic component for his within population estimation, 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.
2) The among-structure/species (i.e additive or evolutionary V/CV matrix) refers to the concerted evolution of traits given integration and selection [70].

Therefore the pooled within-structure covariance matrix approach is not only the one possible to compute for protein structures, but also biologically meaningful.

5 The Response to Selection in the GH13 Family

As defined in equations 2 and 3, 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 as $\beta = P^{-1}S$, and also known as the vector of selection gradients [71] or directional selection gradient. The second and third elements are summarized in the G matrix. As expressed in equation 3, 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 equation 3 is to compute the long-term selection gradient assuming that G is more or less constant over long periods of time:

$$\beta_\lambda = G^{-1}\Delta\bar{z} \tag{11}$$

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 [61]:

$$f = a + \sum_{i=1}^{n} \beta_i z_i + e_i \tag{12}$$

and so it behaves as the weights of a multiple regression of f on the vector of phenotypes z. Fitness is intuitive in organismal biology and can be represented by the count of the offspring of a given phenotype after an event of selection. In proteins, however, the definition is not as straightforward. We can portray fitness in many different ways, 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. To improve this fitness landscape, f can be defined by ΔG° coupled with a functional measure. Since in proteins function is the main selective trait, including a term accounting for this would create a more realistic fitness surface. In enzymes this can be achieved by using the efficiency or K_{cat}/K_M of each of the enzymes for a given substrate. The fitness function (F) can be expressed as:

$$F(i,s) = \Delta G_i^{\circ} \frac{K_{cat}^{i,s}}{K_M^{i,s}} \tag{13}$$

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 the Michaelis constant of protein i working on substrate s.

As can be seen this is a relative fitness, and its relativity depends on the substrate in which is being computed.

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, equation 13 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 $\Delta G^{\circ}_{unfold}$, but knowing the caveats that this only speaks about structural stability and it has been shown that $\Delta G_{equilibrium}$ or $\Delta G^{\circ}_{unfold}$ are not optimized for during evolution [72].

5.1 Estimating dynamic and genetic variance-covariance matrices in the α-Amylase dataset

The structure depicted with the higher fitness was the model 1 of structure 2TAA, from *Aspergillus oryzae*. Here we are going to assume that evolution optimizes for $\Delta\hat{G}^{\circ}$, despite that the evidence shows that this might not be the case. However, a more comprehensive fitness function can be obtained by experimentally determining enzyme efficiency in a given substrate and by applying equation 13. 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:

$$\Delta\bar{z}_{\varpi} = \mu_{\oplus} - \mu_0 \tag{14}$$

where $\mu_{\oplus}$ is the target or after-selection mean structure and μ_0 is the starting structure 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, giving 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} (equation 11), and replacing $\Delta\bar{z}$ by $\Delta\bar{z}_{\varpi}$. 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 Figure A1), and the model 643 of the structure 4E2O from *Geobacillus thermoleovorans* CCB_US3_UF5 (with the lower fitness in Figure A1) corresponds to the source phenotype. 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, Figure 1 shows the absolute value of the sum of β_λ per residue, standardized from 0 to 1.

Figure 1a shows the selection gradient using the estimated G. Not surprisingly, the selection gradient for the TIM-barrel is very low. This means that there is not much directional selection on this sub-structure. However, it is somewhat surprising that there is not any purifying selection 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 the sequence. Therefore, despite some variation may have occurred at the sequence level, it might not have meaningfully affected the positional information.

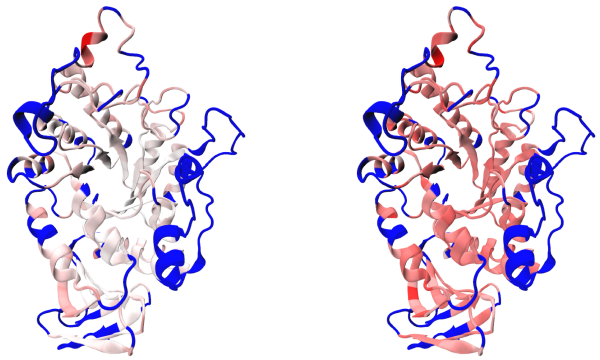
However, one must be cautious with the approach employed in Figure 1 since the signs are missed, thereby ignoring the direction of selection and the correlated response to selection. Nevertheless, this approach allows for a coarse-grained visual exploration of β_{λ_i}. Individual instances identified by this method should be analysed afterwards in each dimension. Table 3 shows the actual values of β_λ for the top 5 positive values (directional selection) and top 5 negative values (purifying selection).

Figure 1b and Table 4 show the mean difference between target and source when effects of correlated dynamic differentials are removed. Given that effectively in equation 11, G acts as a rotation matrix 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 selection. Once G is replaced by M in equation 11, 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 as 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 at one coordinate under negative gradient (Table 4). This can also be inferred by Figure 1b. The values of the dynamic gradient are high but sensi-

TABLE 3
Selection gradient in the top 5 residues. Top panel shows the residues were 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	β_X	β_Y	β_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



(a) (b)

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 depics the non-homologous residues.

TABLE 4
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 cointains the information of residues where at least one of its coordinates is under negative gradient.

ResIndex	Residue	β_X	β_Y	β_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

ble given the definition of fitness. Since I 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 4 and Figure 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 extend 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 than the regular θ. However, manipulating the structure along the dynamics gradient is not feasible, and thus is more applicable to thermodynamic theory of protein structures.

The GH13 dataset $\theta_{\Delta\bar{z}\text{-}\beta}$ 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 selection will drive the source structure towards the target one. The same happens when this approach is applied to M. $\theta^{M}_{\Delta\bar{z}\text{-}\beta}$ is 1.46 degrees, which is almost identical to θ_M. Thus there is 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 to comparative quantitative genetics framework by means of a pooled-within group covariance matrix in a subset of the GH13 proteins is feasible and gives sensible results given the definition of fitness. This definition is essential in the interpretation of the results since it is the one that gives polarity to R_{ϖ}. Therefore, all conclusions about the response to selection and the selection gradient itself must be analysed 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 comes 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 show a small phylogenetic/genetic component to the selection gradient when a less stable structure (4E2O) is analysed with respect to a more stable one (2TAA). In an engineering perspective, this means that most of the changes in shape come from the dynamics. Nevertheless, the small $\theta_{\Delta\bar{z}\text{-}\beta}$ show that most of the changes applied to 4E2O would result directly into increasing the stability towards the one expressed by 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 giving their $\Delta\hat{z}$s. In these cases, the goal will be to shift the position of their centroids given the resulting vector of the three dimensions.

APPENDIX A
SUPPLEMENTARY MATERIAL AND METHODS

All the simulation and test methods used to show the infeasibility of the traditional and bayesian PMM in protein structures can be found in Supplementary material A.

APPENDIX B
SUPPLEMENTARY RESULTS

All the simulation and test results showing the infeasibility of the traditional PMM in protein structures can be found in Supplementary material B.

ACKNOWLEDGMENTS

The authors thank the members of the Blouin Lab for helpful comments and critical review of this 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.

REFERENCES

[1] M. Ben Ali, B. Khemakhem, X. Robert, R. Haser, and S. Bejar, "Thermostability enhancement and change in starch hydrolysis profile of the maltohexaose-forming amylase of bacillus stearothermophilus us100 strain." *Biochem. J.*, vol. 394, no. Pt 1, pp. 51–6, 2006.

[2] R. J. Bothast and M. A. Schlicher, "Biotechnological processes for conversion of corn into ethanol." *Appl. Microbiol. Biotechnol.*, vol. 67, no. 1, pp. 19–25, 2005.

[3] J. A. Gerlt and F. M. Raushel, "Evolution of function in $(\beta/\alpha)_8$-barrel enzymes," *Current opinion in chemical biology*, vol. 7, no. 2, pp. 252–264, 2003.

[4] B. Höcker, C. Jürgens, M. Wilmanns, and R. Sterner, "Stability, catalytic versatility and evolution of the $(\beta/\alpha)_8$-barrel fold," *Current opinion in biotechnology*, vol. 12, no. 4, pp. 376–381, 2001.

[5] S. J. Steppan, P. C. Philips, and D. Houle, "Comparative quantitative genetics: evolution of the g matrix," *Trends in Ecology & Evolution*, vol. 17, no. 7, pp. 320–327, 2002.

[6] C. P. Klingenberg and L. J. Leamy, "Quantitative genetics of geometric shape in the mouse mandible," *Evolution*, vol. 55, no. 11, pp. 2342–2352, 2001. [Online]. Available: http://dx.doi.org/10.1111/j.0014-3820.2001.tb00747.x

[7] R. Thompson, "Estimation of quantitative genetic parameters," *Proceedings of the Royal Society B: Biological Sciences*, vol. 275, no. 1635, pp. 679–686, 2008.

[8] D. S. Falconer and T. F. C. Mackay, *Introduction to quantitative genetics*. Longman, 1996. [Online]. Available: http://books.google.ca/books?id=lrCYQgAACAAJ

[9] R. Lande and S. J. Arnold, "The measurement of selection on correlated characters," *Evolution*, pp. 1210–1226, 1983.

[10] C. Klingenberg, "Quantitative genetics of geometric shape: heritability and the pitfalls of the univariate approach," *Evolution*, vol. 57, no. 1, pp. 191–195, 2003.

[11] A. Blasco, "The bayesian controversy in animal breeding." *Journal of Animal Science*, vol. 79, no. 8, pp. 2023–2046, 2001.

[12] M. Lynch, "Methods for the analysis of comparative data in evolutionary biology," *Evolution*, pp. 1065–1080, 1991.

[13] J. D. Hadfield and S. Nakagawa, "General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters," *Journal of evolutionary biology*, vol. 23, no. 3, pp. 494–508, 2010.

[14] J. M. Cheverud, M. M. Dow, and W. Leutenegger, "The quantitative assessment of phylogenetic constraints in comparative analyses: sexual dimorphism in body weight among primates," *Evolution*, pp. 1335–1351, 1985.

[15] E. A. Housworth, E. P. Martins, and M. Lynch, "The phylogenetic mixed model," *The American Naturalist*, vol. 163, no. 1, pp. 84–96, 2004.

[16] R. P. Freckleton, P. H. Harvey, and M. Pagel, "Phylogenetic analysis and comparative data: a test and review of evidence," *The American Naturalist*, vol. 160, no. 6, pp. 712–726, 2002.

[17] M. Pagel, "Inferring the historical patterns of biological evolution," *Nature*, vol. 401, no. 6756, pp. 877–884, 1999.

[18] E. P. Martins, J. A. F. Diniz-Filho, and E. A. Housworth, "Adaptive constraints and the phylogenetic comparative method: a computer simulation test," *Evolution*, vol. 56, no. 1, pp. 1–13, 2002.

[19] D. Sorensen and D. Gianola, *Likelihood, Bayesian and MCMC methods in quantitative genetics*. Springer, 2002.

[20] P. C. Lambert, A. J. Sutton, P. R. Burton, K. R. Abrams, and D. R. Jones, "How vague is vague?: A simulation study of the impact of the use of vague prior distributions in mcmc using winbugs," *Statistics in medicine*, vol. 24, no. 15, pp. 2401–2428, 2005.

[21] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, "Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation," *Journal of chemical theory and computation*, vol. 4, no. 3, pp. 435–447, 2008.

[22] D. Penninga, B. Strokopytov, H. J. Rozeboom, C. L. Lawson, B. W. Dijkstra, J. Bergsma, and L. Dijkhuizen, "Site-directed mutations in tyrosine 195 of cyclodextrin glycosyltransferase from bacillus circulans strain 251 affect activity and product specificity," *Biochemistry*, vol. 34, no. 10, pp. 3368–3376, 1995.

[23] A. M. Brzozowski, D. M. Lawson, J. P. Turkenburg, H. Bisgaard-Frantzen, A. Svendsen, T. V. Borchert, Z. Dauter, K. S. Wilson, and G. J. Davies, "Structural analysis of a chimeric bacterial α-amylase. high-resolution analysis of native and ligand complexes," *Biochemistry*, vol. 39, no. 31, pp. 9099–9107, 2000.

[24] L. K. Skov, O. Mirza, A. Henriksen, G. P. De Montalk, M. Remaud-Simeon, P. Sarçabal, R.-M. Willemot, P. Monsan, and M. Gajhede, "Amylosucrase, a glucan-synthesizing enzyme from the α-amylase family," *Journal of Biological Chemistry*, vol. 276, no. 27, pp. 25 273–25 278, 2001.

[25] H.-S. Lee, M.-S. Kim, H.-S. Cho, J.-I. Kim, T.-J. Kim, J.-H. Choi, C. Park, H.-S. Lee, B.-H. Oh, and K.-H. Park, "Cyclomaltodextrinase, neopullulanase, and maltogenic amylase are nearly indistinguishable from each other," *Journal of Biological Chemistry*, vol. 277, no. 24, pp. 21 891–21 897, 2002.

[26] H. Hondoh, T. Kuriki, and Y. Matsuura, "Three-dimensional structure and substrate binding of bacillus stearothermophilus neopullulanase," *Journal of molecular biology*, vol. 326, no. 1, pp. 177–188, 2003.

[27] S. Numao, R. Maurus, G. Sidhu, Y. Wang, C. M. Overall, G. D. Brayer, and S. G. Withers, "Probing the role of the chloride ion in the mechanism of human pancreatic α-amylase," *Biochemistry*, vol. 41, no. 1, pp. 215–225, 2002.

[28] N. Aghajari, M. Roth, and R. Haser, "Crystallographic evidence of a transglycosylation reaction: ternary complexes of a psychrophilic α-amylase," *Biochemistry*, vol. 41, no. 13, pp. 4273–4280, 2002.

[29] D. Zhang, N. Li, S.-M. Lok, L.-H. Zhang, and K. Swaminathan, "Isomaltulose synthase (pali) of klebsiella sp. lx3 crystal structure and implication of mechanism," *Journal of Biological Chemistry*, vol. 278, no. 37, pp. 35 428–35 434, 2003.

[30] J.-S. Kim, S.-S. Cha, H.-J. Kim, T.-J. Kim, N.-C. Ha, S.-T. Oh, H.-S. Cho, M.-J. Cho, M.-J. Kim, H.-S. Lee *et al.*, "Crystal structure of a maltogenic amylase provides insights into a catalytic versatility," *Journal of Biological Chemistry*, vol. 274, no. 37, pp. 26 279–26 286, 1999.

[31] S. Strobl, K. Maskos, G. Wiegand, R. Huber, F. X. Gomis-Rüth, and R. Glockshuber, "A novel strategy for inhibition of α-amylases: yellow meal worm α-amylase in complex with the ragi bifunctional inhibitor at 2.5 ǎ resolution," *Structure*, vol. 6, no. 7, pp. 911–921, 1998.

[32] M. Kagawa, Z. Fujimoto, M. Momma, K. Takase, and H. Mizuno, "Crystal structure of *Bacillus subtilis* α-amylase in complex with acarbose," *Journal of Bacteriology*, vol. 185, no. 23, pp. 6981–6984, 2003.

[33] T. Nonaka, M. Fujihashi, A. Kita, H. Hagihara, K. Ozaki, S. Ito, and K. Miki, "Crystal structure of calcium-free α-amylase from bacillus sp. strain ksm-k38 (amyk38) and its sodium ion binding sites," *Journal of Biological Chemistry*, vol. 278, no. 27, pp. 24 818–24 824, 2003.

[34] K. Y. Hwang, H. K. Song, C. Chang, J. Lee, S. Y. Lee, K. K. Kim, S. Choe, R. M. Sweet, and S. W. Suh, "Crystal structure of thermostable alpha-amylase from bacillus licheniformis refined at 1.7 a resolution." *Molecules and cells*, vol. 7, no. 2, pp. 251–258, 1997.

[35] G. J. Davies, A. M. Brzozowski, Z. Dauter, M. D. Rasmussen, T. V. Borchert, and K. S. Wilson, "Structure of a bacillus halmapalus family 13 α-amylase, bha, in complex with an acarbose-derived nonasaccharide at 2.1ă resolution," *Acta Crystallographica Section D*, vol. 61, no. 2, pp. 190–193, Feb 2005. [Online]. Available: http://dx.doi.org/10.1107/S0907444904027118

[36] M. Mizuno, K. Ichikawa, T. Tonozuka, A. Ohtaki, Y. Shimura, S. Kamitori, A. Nishikawa, and Y. Sakano, "Mutagenesis and structural analysis of thermoactinomyces vulgarius r-47 α-amylase ii (tvaii)," *Journal of Applied Glycoscience*, vol. 52, no. 3, pp. 225–231, 2005.

[37] S. Ravaud, X. Robert, H. Watzlawick, R. Haser, R. Mattes, and N. Aghajari, "Trehalulose synthase native and carbohydrate complexed structures provide insights into sucrose isomerization." *J. Biol. Chem.*, vol. 282, no. 38, pp. 28 126–36, 2007.

[38] T. Shirai, K. Igarashi, T. Ozawa, H. Hagihara, T. Kobayashi, K. Ozaki, and S. Ito, "Ancestral sequence evolutionary trace and crystal structure analyses of alkaline α-amylase from bacillus sp. ksm-1378 to clarify the alkaline adaptation process of proteins," *Proteins: Structure, Function, and Bioinformatics*, vol. 66, no. 3, pp. 600–610, 2007.

[39] B. Mikami, H. Iwamoto, D. Malle, H.-J. Yoon, E. Demirkan-Sarikaya, Y. Mezaki, and Y. Katsuya, "Crystal structure of pullulanase: evidence for parallel binding of oligosaccharides in the active site," *Journal of molecular biology*, vol. 359, no. 3, pp. 690–707, 2006.

[40] Y. Matsuura, M. KUSUNOKI, W. HARADA, and M. KAKUDO, "Structure and possible catalytic residues of taka-amylase a," *Journal of Biochemistry*, vol. 95, no. 3, pp. 697–702, 1984.

[41] J. P. Turkenburg, A. M. Brzozowski, A. Svendsen, T. V. Borchert, G. J. Davies, and K. S. Wilson, "Structure of a pullulanase from bacillus acidopullulyticus." *Proteins*, vol. 76, no. 2, pp. 516–9, 2009.

[42] M. B. Vester-Christensen, M. Abou Hachem, B. Svensson, and A. Henriksen, "Crystal structure of an essential enzyme in seed starch degradation: barley limit dextrinase in complex with cyclodextrins." *J. Mol. Biol.*, vol. 403, no. 5, pp. 739–50, 2010.

[43] T. Shirai, V. S. Hung, K. Morinaka, T. Kobayashi, and S. Ito, "Crystal structure of gh13 α-glucosidase gsj from one of the deepest sea bacteria," *Proteins: Structure, Function, and Bioinformatics*, vol. 73, no. 1, pp. 126–133, 2008.

[44] H. Hondoh, W. Saburi, H. Mori, M. Okuyama, T. Nakada, Y. Matsuura, and A. Kimura, "Substrate recognition mechanism of α-1, 6-glucosidic linkage hydrolyzing enzyme, dextran glucosidase from *Streptococcus mutans*," *Journal of molecular biology*, vol. 378, no. 4, pp. 913–922, 2008.

[45] K. Yamamoto, H. Miyake, M. Kusunoki, and S. Osaki, "Steric hindrance by 2 amino acid residues determines the substrate specificity of isomaltase from *Saccharomyces cerevisiae*," *Journal of bioscience and bioengineering*, vol. 112, no. 6, pp. 545–550, 2011.

[46] M.-I. Kim, H.-S. Kim, J. Jung, and S. Rhee, "Crystal structures and mutagenesis of sucrose hydrolase from xanthomonas axonopodis pv. glycines: insight into the exclusively hydrolytic amylosucrase fold." *J. Mol. Biol.*, vol. 380, no. 4, pp. 636–47, 2008.

[47] S. Buedenbender and G. E. Schulz, "Structural base for enzymatic cyclodextrin hydrolysis." *J. Mol. Biol.*, vol. 385, no. 2, pp. 606–17, 2009.

[48] S. Ravaud, X. Robert, H. Watzlawick, R. Haser, R. Mattes, and N. Aghajari, "Structural determinants of product specificity of sucrose isomerases." *FEBS Lett.*, vol. 583, no. 12, pp. 1964–8, 2009.

[49] F. Guérin, S. Barbe, S. Pizzut-Serin, G. Potocki-Véronèse, D. Guieysse, V. Guillet, P. Monsan, L. Mourey, M. Remaud-Siméon, I. André, and S. Tranier, "Structural investigation of the thermostability and product specificity of amylosucrase from the bacterium deinococcus geothermalis." *J. Biol. Chem.*, vol. 287, no. 9, pp. 6642–54, 2012.

[50] K. Mizutani, M. Toyoda, Y. Otake, S. Yoshioka, N. Takahashi, and B. Mikami, "Structural and functional characterization of recombinant medaka fish alpha-amylase expressed in yeast pichia pastoris," *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, vol. 1824, no. 8, pp. 954–962, 2012.

[51] P. Han, P. Zhou, S. Hu, S. Yang, Q. Yan, and Z. Jiang, "A novel multifunctional α-amylase from the thermophilic fungus *Malbranchea cinnamomea*: Biochemical characterization and three-dimensional structure," *Applied Biochemistry and Biotechnology*, vol. 2, no. 170, pp. 420–435, 2013.

[52] S.-C. Mok, A.-H. Teh, J. A. Saito, N. Najimudin, and M. Alam, "Crystal structure of a compact α-amylase from *Geobacillus thermoleovorans*," *Enzyme and microbial technology*, vol. 53, no. 1, pp. 46–54, 2013.

[53] A. Lipski, H. Watzlawick, S. Ravaud, X. Robert, M. Rhimi, R. Haser, R. Mattes, and N. Aghajari, "Mutations inducing an active-site aperture in rhizobium sp. sucrose isomerase confer hydrolytic activity," *Acta Crystallographica Section D: Biological Crystallography*, vol. 69, no. 2, pp. 298–307, 2013.

[54] I. Schomburg, A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn, and D. Schomburg, "Brenda, the enzyme database: updates and major new developments," *Nucleic acids research*, vol. 32, no. suppl 1, pp. D431–D433, 2004.

[55] M. Menke, B. Berger, and L. Cowen, "Matt: local flexibility aids protein multiple structure alignment." *PLoS Comput. Biol.*, vol. 4, no. 1, p. e10, 2008.

[56] I. L. Dryden, "Shapes package," *R Foundation for Statistical Computing, Vienna Contributed package*, 2011.

[57] D. C. Adams and G. J. P. Naylor, "A new method for evaluating the structural similarity of proteins using geometric morphometrics," in *Currents in Computational Molecular Biology*, ser. Frontiers Science Series, S. Miyano, R. Shamir, and T. Takagi, Eds., vol. 30. Tokyo: Universal Academy Press, 2000, pp. 120–121.

[58] ——, "A comparison of methods for assessing the structural similarity of proteins," in *Mathematical Methods for Protein Structure Analysis and Design*, 2003, pp. 109–115.

[59] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano, "The foldx web server: an online force field," *Nucleic Acids Research*, vol. 33, no. suppl 2, pp. W382–W388, 2005.

[60] S. J. Arnold, "Constraints on phenotypic evolution," *American Naturalist*, vol. 61, pp. S85–S107, 1992.

[61] M. Blows and B. Walsh, "Spherical cows grazing in flatland: Constraints to selection and adaptation," in *Adaptation and fitness in animal populations*, J. van der Werf, H.-U. Graser, R. Frankham, and C. Gondro, Eds. Springer, 2009, pp. 83–101. [Online]. Available: http://dx.doi.org/10.1007/978-1-4020-9005-9_6

[62] T. F. Hansen and D. Houle, "Measuring and comparing evolvability and constraint in multivariate characters." *J. Evol. Biol.*, vol. 21, no. 5, pp. 1201–19, 2008.

[63] B. Walsh and M. W. Blows, "Abundant genetic variation+ strong selection= multivariate genetic constraints: a geometric view of adaptation," *Annual Review of Ecology, Evolution, and Systematics*, vol. 40, pp. 41–59, 2009.

[64] G. Marroig and J. M. Cheverud, "A comparison of phenotypic variation and covariation patterns and the role of phylogeny, ecology, and ontogeny during cranial evolution of new world monkeys," *Evolution*, vol. 55, no. 12, pp. 2576–2600, 2001.

[65] J. M. Cheverud, "Quantitative genetic analysis of cranial morphology in the cotton-top (saguinus oedipus) and saddle-back (s. fuscicollis) tamarins," *Journal of Evolutionary Biology*, vol. 9, no. 1, pp. 5–42, 1996.

[66] D. A. Roff, "The estimation of genetic correlations from phenotypic correlations: a test of cheverud's conjecture," *Heredity*, vol. 74, no. 5, pp. 481–490, 1995.

[67] L. J. Revell, L. J. Harmon, R. B. Langerhans, and J. J. Kolbe, "A phylogenetic approach to determining the importance of constraint on phenotypic evolution in the neotropical lizard anolis cristatellus," *Evolutionary Ecology Research*, vol. 9, no. 2, pp. 261–282, 2007.

[68] A. Porto, L. T. Shirai, F. B. de Oliveira, and G. Marroig, "Size variation, growth strategies, and the evolution of modularity in the mammalian skull." *Evolution*, vol. 67, no. 11, pp. 3305–22, 2013.

[69] M. Prôa, P. O'Higgins, and L. R. Monteiro, "Type i error rates for testing genetic drift with phenotypic covariance matrices: a simulation study." *Evolution*, vol. 67, no. 1, pp. 185–95, 2013.

[70] A. Haber, "The evolution of morphological integration in the ruminant skull," *Evolutionary Biology*, vol. 42, no. 1, pp. 99–114, 2015.

[71] M. D. Rausher, "The measurement of selection on quantitative traits: biases due to environmental covariances between traits and fitness," *Evolution*, pp. 616–626, 1992.

[72] J. A. Alfaro, "Capturing the dynamics of protein sequence evolution through site-independent structurally constrained phylogenetic models," Ph.D. dissertation, Department of biochemitry & molecular biology, Dalhousie University, Halifax, Canada, 2014.

Jose Sergio Hleap received a B.Sc (Hon.) and M.Sc. in Biology from the Universidad del Valle in Colombia, in 2005 and 2010 respectively. He received his PhD in Biochemistry and Molecular Biology from Dalhousie University in 2016. He has developed multiple algorithms to analyse protein structures geometrically. He has contributed into a clearer definition of protein domains and the subdomain architecture in protein structures. His research interest include molecular and phenotypic evolution, population genetics, protein structure evolution, and algorithmic methods to help analyse the former subjects.

Christian Blouin Christian Blouin is a professor of Computer Science and Biochemistry at Dalhousie University. His research interest are in the study of molecular evolution at the protein structure level. He is the recipient of a number of teaching awards, and is active in the development and review of academic programmes.