PhISCO: a simple method to infer phenotypes from protein sequences

Ayelén S. Hernández Berthet[a]*, Ariel A. Aptekmann[f,i,j]*, Jesús Tejero[b,c,d,e], Ignacio E. Sánchez[f], Martín E. Noguera[g,h]** and Ernesto A. Roman[a,g]****

[a] Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Intendente Güiraldes 2160 - Ciudad Universitaria, 1428EGA C.A.B.A., Argentina.

[b] Heart, Lung, Blood and Vascular Medicine Institute, University of Pittsburgh, Pittsburgh, PA 15261, USA

[c] Division of Pulmonary, Allergy and Critical Care Medicine, University of Pittsburgh, Pittsburgh, PA 15261, USA

[d] Department of Bioengineering, Swanson School of Engineering, University of Pittsburgh, Pittsburgh, PA 15260, USA

[e] Department of Pharmacology and Chemical Biology, University of Pittsburgh, Pittsburgh, PA 15261, USA

[f] Universidad de Buenos Aires. Consejo Nacional de Investigaciones Científicas y Técnicas. Instituto de Química Biológica de la Facultad de Ciencias Exactas y Naturales (IQUIBICEN). Facultad de Ciencias Exactas y Naturales. Laboratorio de Fisiología de Proteínas. Buenos Aires, Argentina

[g] Consejo Nacional de Investigaciones Científicas y Técnicas, Instituto de Química y Fisicoquímica Biológicas, Buenos Aires, Argentina

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

[I] Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, NJ 08873, USA

[J] Institute of Marine and Coastal Sciences, Rutgers University, New Brunswick, NJ 08901, USA

*Both authors contributed equally to this work

**To whom correspondence should be addressed: isanchez@qb.fcen.uba.ar, mnoguera.unq@gmail.com and ernest.roman@gmail.com .

Running title: Inferring phenotypes from sequences.

Statement of Significance. We explored the association between protein sequences and quantitative phenotypes to find residues responsible for phenotype modulation. To this end, we used a simple measure based on the per-position correlation between sequence divergence and phenotype difference. Several function-linked positions were identified in four case studies corresponding to different protein families. Additionally, we used the most highly correlated positions to make quantitative predictions. This extremely simple strategy was useful to identify candidate residues for mutagenesis or other analysis in sequence-function studies and also, for the phenotype prediction of uncharacterized proteins or organisms.

Author contributions. A.S.H.B, A.A.A, I.E.S, M.E.N and E.A.R conceived and designed the study. A.S.H.B, M.E.N and E.A.R performed calculations. A.S.H.B, A.A.A, J.T, I.E.S, M.E.N and E.A.R analyzed data. A.A.A. and E.A.R. contributed code. I.E.S, M.E.N and E.A.R wrote the manuscript. All authors edited the manuscript.

Declaration of Interests. The authors state that they have no conflict of interests.

Acknowledgements. E.A.R. thanks M.D. Ph.D. Lázaro Efraín Gerschenson for always trying to do the best he can with what he has.

Abstract

Protein sequences encode the information necessary for function and folding, however accessing it is not straightforward. Unluckily, although structure and function are tightly linked, the prediction of how specific amino acids contribute to these features is still considerably impaired. Here, we developed PhISCO, Phenotype Inference from Sequence COmparisons, a simple algorithm that finds positions associated with any quantitative phenotype and predicts their values. From a few hundred sequences from four different protein families, we performed multiple sequence alignments and calculated per-position pairwise differences for both the sequence and the observed phenotypes. For the case of Adenylate Kinase we found 3 positions in proximity of the ligand binding sites that are linked to their Optimal Growth Temperature. For microbial rhodopsins, we identified 10 positions close to their chromophore and the lipid-protein interface whose differences are correlated to the maximal absorption wavelength. In the case of myoglobins, 3 positions were identified that have been described as tightly linked to muscular myoglobin concentration. Finally, we identified positions associated with the inhibitory potency of 2 inhibitors of the HIV protease. We showed that strong correlations exist in single positions while an improvement is achieved when the most correlated positions are jointly analyzed. Noteworthy, we performed phenotype predictions using a simple linear model that links per-position divergences of most correlated positions and differences in observed phenotypes. The diversity of the explored systems make PhISCO valuable to find sequence determinants of biological activity modulation and to predict various functional features for uncharacterized members of a protein family.

Introduction

Proteins are essential molecules for cellular metabolism. Conformational dynamics, stability and functional features such as catalytic efficiency or interaction capability are encoded in their sequences. Thus, this information must be accessible from analysis of protein sequences. As an example, coevolutionary analysis from sequences has succeeded in the prediction of native contacts and their concomitant structures with the aid of statistical and bioinformatic tools [1]. Although structure and function are tightly linked, the prediction of functional features in high detail is still considerably impaired.

There is great interest in understanding the mechanisms through which homologous proteins are adapted to the very heterogeneous selective pressures to which they are subjected. A deep analysis of these mechanisms could provide pathways to modulate phenotypic characteristics such as stability, chromophore absorption, catalytic efficiency and ligand affinities. From an applied point of view, this knowledge can be used to design proteins whose properties are optimized to work in the harsh conditions frequently found in industrial processes such as high temperatures or increased salt concentrations. Whereas genetic variation allows the phenotypic variation, a delicate balance of molecular interactions is necessary to preserve the proper stable three-dimensional structure in a functional competent state and consequently, not all the residues across the protein sequence are equally available to functional tuning.

Regarding the existing methodologies to quantitatively predict phenotypes, some of them rely on the structural information available and are time-consuming, require expert supervision, and are poorly accurate[2]. Although with the aid of structure predictors as AlphaFold2 the structure-requirement is no longer an obstacle for most targets, predictions are still not performed with significant accuracy.

On the side of sequence-based methods for linking protein sequences with quantitative phenotypes, much progress has been made [3,4]. Some of these methods associate sequence properties, such as amino acid abundance in specific positions or their physicochemical properties, to a quantitative phenotype such as the organisms' OGT[5], optimal working temperature of proteins [6] or the deleterious impact of mutations in disease[7]. A growing number of these methodologies rely on machine learning, which has been used to predict OGT[8], brightness of fluorescent proteins[4], and in vitro binding affinities[9], among other phenotypes. While these methods are fairly accurate in general, they suffer from parameter interpretation problems and require extensive training tasks for a specific protein family [10,11].

In this work we propose a simple method that we called PhISCO from Phenotype Inference from Sequence COmparisons, to 1) identify protein positions correlated with the acquisition of specific phenotypes and, from these findings, 2) quantitatively predict phenotypes for a target protein sequence. The idea behind this method is that only a limited set of residues is responsible for modulation of functional phenotypes, whereby it is expected that proteins working in similar conditions exhibit similar residues at these positions. We studied the correlation of per-position sequence differences ("sequence divergence") and their corresponding phenotype-associated differences ("phenotype difference") and we illustrate the performance of our method using four case studies of protein families with different associated

properties: a set of adenylate kinases from Archaea with their associated OGT, a set of microbial rhodopsins whose retinal cofactor absorption occurs at different wavelengths [4], a set of myoglobin sequences from mammals and their maximal concentration in muscle[12], and the potency of HIV protease inhibitors for clinical isolates [13]. To our knowledge, only a few methods exploit the sequence information of different organisms with a simple metric that relates the sequence similarity to the proximity of the OGT [12]. Our proposal minimizes calculation requirements with almost no assumption on a minimal set of parameters.

Materials and methods

Identification of protein positions associated to phenotypes

PhISCO is based on pairwise-sequence comparisons of diverse protein families and is described in **Figure 1**. Sequences were downloaded from RefSeq [14] and UNIPROT [15] while their associated phenotypes were retrieved from the correspondingly cited database or publication. Divergence was evaluated in a per-position pairwise manner using an identity matrix, where values of 0 or 1 are assigned to identical or different amino acids, respectively. In our calculations we studied four different quantitative phenotypes: optimal growth temperature of microorganisms, absorption wavelength in bacterial rhodopsins, maximal muscle myoglobin concentration in mammalian divers, and antiviral inhibition potency of clinical HIV isolates. These features were also evaluated in a pairwise manner and the correlation between the absolute value of their differences (*ΔPhenotype*) and the corresponding sequence divergence was studied. The association between these two magnitudes was evaluated independently for each position using the Spearman rank-order correlation coefficient (ρ)[16] (Figure 1A). Then, we ordered the positions from highest to lowest ρ-values and recalculated the correlation by sequentially adding the divergence of each position. Thus, at the beginning of the calculation, the ρ-value corresponds to the position with highest correlation; in the next step, we add the divergence of the second highest correlation to that of the former position and recalculate the ρ-value. We repeated the process until all of the positions are included (Figure 1B). The group of positions with the highest ρ-value were mapped onto the 3D structure of the protein to analyze their functional relevance.

Sequence-based quantitative phenotype predictions

To perform the predictions, 50% of the sequences in each study case were assigned to a *training* and the remaining to a *validation* set. For a given position or set of positions, the relationship between sequence divergence for a pair *i* and *j* from the *training* set and its observed *ΔPhenotype* was modeled according the following linear equation (Figure 1B):

$$\Delta Phenotype, obs(i,j) = a \cdot (Seq\ divergence)(i,j) + b \qquad (1)$$

Each *k* sequence from the *validation* set was compared to the training sequences (*j*) yielding as many pairs of sequence divergence values as the number of sequences in the *training* set (Figure 1C). From these divergences, the values of the predicted *ΔPhenotype* were calculated using the fitted coefficients of Equation 1, according to:

$$\Delta Phenotype, pred(k,j) = a \cdot (Seq\ divergence)(k,j) + b \qquad (2)$$

To do the prediction, the target phenotype was chosen as the one that minimizes the squared-sum of the difference between the predicted *ΔPhenotype* from the linear model (*ΔPhenotype,pred(k,j)*) and the difference between the unknown (*Phenotype,pred(k)*) and each known phenotype from the training set (*Phenotype,obs(j)*) (FIgure 1C):

$$\sum_{j}^{\square} \left\{ \left| Phenotype, obs(j) - Phenotype, pred(k) \right| - \Delta\, Phenotype, pred(k, j) \right\}^2 \qquad (3)$$

The correlation between observed and predicted phenotypes was evaluated using the Pearson correlation coefficient. To test prediction robustness we used 50 random selected training and validation datasets.

Sequence datasets with their associated-phenotypes

To study temperature adaptation in Adenylate kinases (Adk), we used a database of Optimal Growth Temperatures (OGTs) compiled by our group, consisting in Archaeal organisms with fully sequenced genomes, which is an extended version of the data contained in Aptekmann et al. [17]. To this end, OGT values were collected from published articles identified from PubMed searches using the organism's name as a query. When no OGT was available, it was approximated as the mean of the reported minimal and maximal growth temperatures of the organism. Available AdK protein sequences from the organisms in the definitive database were collected using the blastp search tool [18]. Multiple sequence alignments (MSA) were performed using the online server MAFFT with default settings. Columns with more than 10% of gaps were eliminated. A cluster analysis was performed using CLANS [19], a clusterization algorithm based on *E*-values from pairwise-sequence comparisons. As another study case, we studied a set of sequences of microbial rhodopsins with their light-absorption wavelengths obtained from *Karasuma et al.* [4]. The alignment, trimming and clusterization strategies were the same as in the case of AdK. In a third example, we worked with a set of extant mammalians myoglobin sequences, alignment and maximal muscle concentration were obtained from Mirceta et al. [20]. Finally, we studied sequences of HIV protease from clinical isolates and their associated change in the protease inhibitor drug resistance. This information was derived from the Stanford University HIV Drug Resistance Database (HIVdb), as published in Rhee et al. [21] (accessed in February 2022). Due to the high sequence conservation in HIV protease datasets, alignment and gap placement were unnecessary. Phenotypes refer to the level of protease-inhibition resistance, expressed as fold changes of inhibition (the ratio of concentration necessary to inhibit replication of a given strain compared with the wild type). The list of drug resistance positions (DRP) was obtained combining data from the Protease Inhibitors Resistance Notes of HIVdb available at https://hivdb.stanford.edu/dr-summary/resistance-notes/PI and the 2019 edition of the IAS–USA drug resistance mutations list [22].

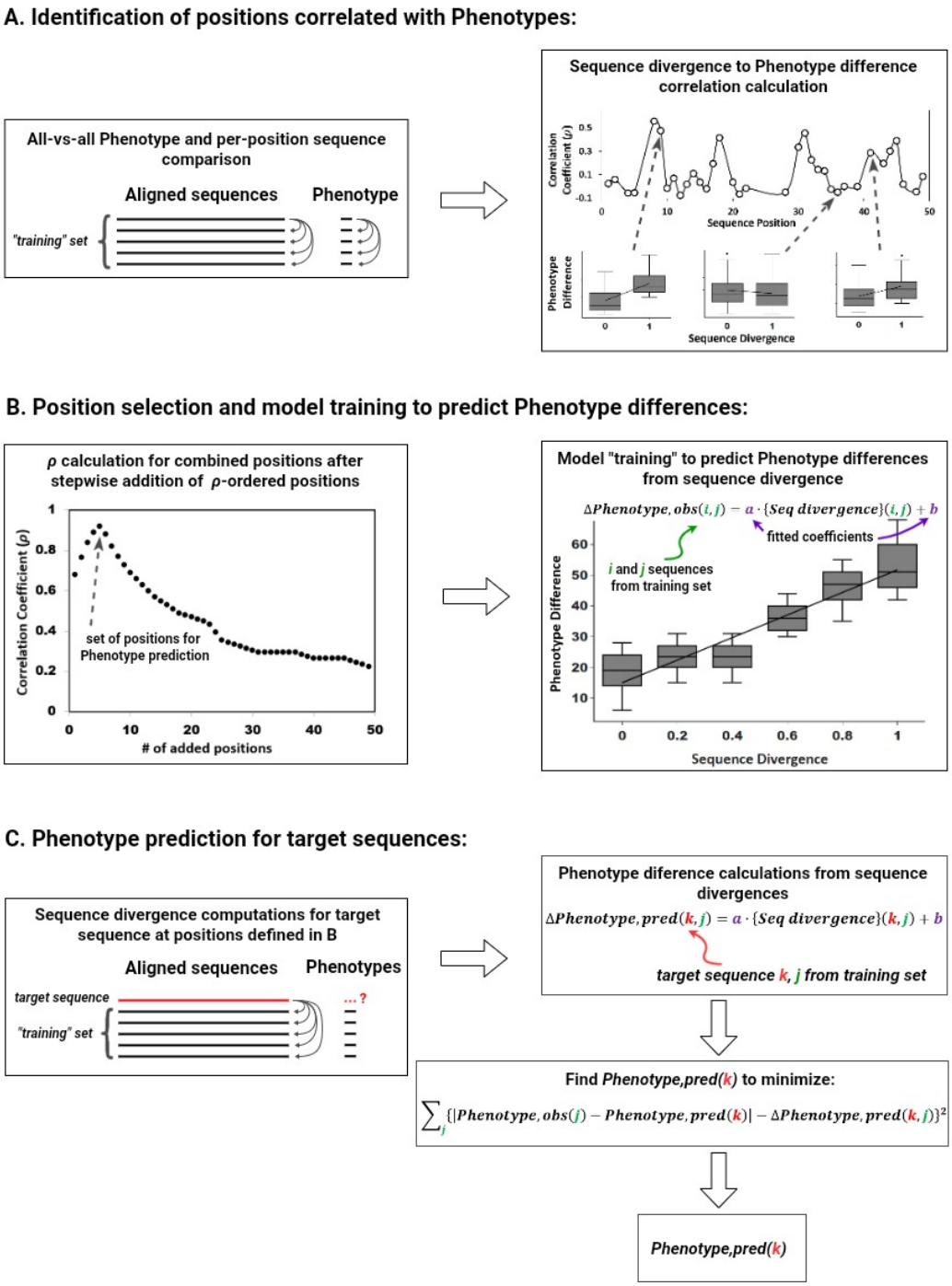

Figure 1. Detailed workflow of PhISCO algorithm. A. Identification of positions correlated with the phenotypes. Sequences from a protein family were aligned and pairwise similarity comparison was performed for the sequence and the observed features, for each position. A Spearman's rank-order correlation analysis was performed and rho values were plotted for each position. B. Position selection and model training to predict phenotype differences. Positions were ordered from highest to lowest *ρ*-values and the correlation was recalculated by sequentially adding the divergence of each position. Thus, at the beginning of the calculation, the *ρ*-value corresponds to the position with highest correlation and the last point corresponds to the whole sequence. C. To do the prediction the target feature was chosen as the one that minimizes the squared-sum of the difference between the predicted ΔPhenotype from

the linear model and the difference between the unknown and each phenotype in the training set sequences. The calculation was repeated for all of the sequences of the validation set.

Results

The case of archaeal adenylate kinases (*ADK*)

Enzymes involved in key-metabolism steps need to properly work in conditions similar to those in which the organisms grow [17]. One of these key enzymes is AdK, which plays a central role in energy metabolism by regulating the ATP concentration [17,23]. We focused on AdK from *Archaea*, since these organisms grow in a wide range of temperatures and there is abundant information about sequences and growing conditions. After applying the clusterization procedure described in *Methods*, we found two groups of sequences: one of 163 and another of 80 sequences (**Figure 2A**). While the minor cluster contains proteins that had been reported as trimeric [24–26], none of the sequences from the major cluster has a reported structure. However, the high identity with bacterial monomeric homologs suggests that this set may also be monomers. In that sense, and since these proteins seem to have different oligomeric states that can redound in differences in their catalytic mechanisms, we focus on the major cluster. An analysis of the whole sequence divergence and their associated *ΔOGT* correlation shows a R^2 = 0.623 ± 0.008 (**Figure S1**). However, a region is evident where although divergence is high, the *ΔOGT* is low. To explore whether there are specific positions that exhibit a tighter link to *ΔOGT* than the whole sequence, we studied the per-position correlation as described in *Methods*. Spearman's rank correlation coefficients showed a widespread distribution with a mean value of 0.17 ± 0.21 (**Figure 2B**). Although most of the positions exhibit correlation coefficients close to the mean, a small subset clearly separates from that value. We evaluated if there exists a group among these positions that further improve the correlation with *ΔOGTs* by correlation recalculation after sequential addition of *ρ*-value ordered positions. Whereas this method to explore the joint correlation is not exhaustive, it provides a path with high chance to find a set of residues that captures most of the sequence-phenotype association. For the case of AdK, we found three positions that combined yielded the highest *ρ*-value (0.83) (**Figure 2C-E, Table 1**). We mapped these positions in the AdK X-ray structure (PDB ID 1AKE) (**Figure 2D**). All of them are located at the ligand binding sites: while position 54 is located in the AMP lid domain, positions 160 and 221 are located in the ATP lid domain (ATPlid) [27]. Computational analysis of coevolution combined with molecular dynamics simulations showed that position 54 is highly coupled to spatially distant positions [28,29]. Regarding positions in the ATPlid, when the mutation of position 160 from I to G is introduced in the *E. coli* AdK ATP binding is improved without significantly protein stability effect but decreasing catalytic efficiency [27]. Also, structure-based simulations and normal mode analysis showed that both positions 160 and 221 are involved in local unfolding events with effects in AdK catalysis [29,30,31]. Given that this reduced set of positions are involved in enzyme functionality and captures most of the correlation, we next tested its capability to predict OGT. The sequence dataset was split in training and validation sets (50 % of randomly picked sequences each one). This random sampling was repeated 50 times to evaluate prediction robustness. A representative prediction plot with the mean R^2 value is shown in **Figure 2F**. Interestingly, the prediction using the three positions gave more accurate results (R^2 = *0.85*) than the whole sequence (R^2 = *0.56*). This finding, besides showing that

those positions are linked to OGT, makes this algorithm very helpful for OGT prediction based on AdK sequences.

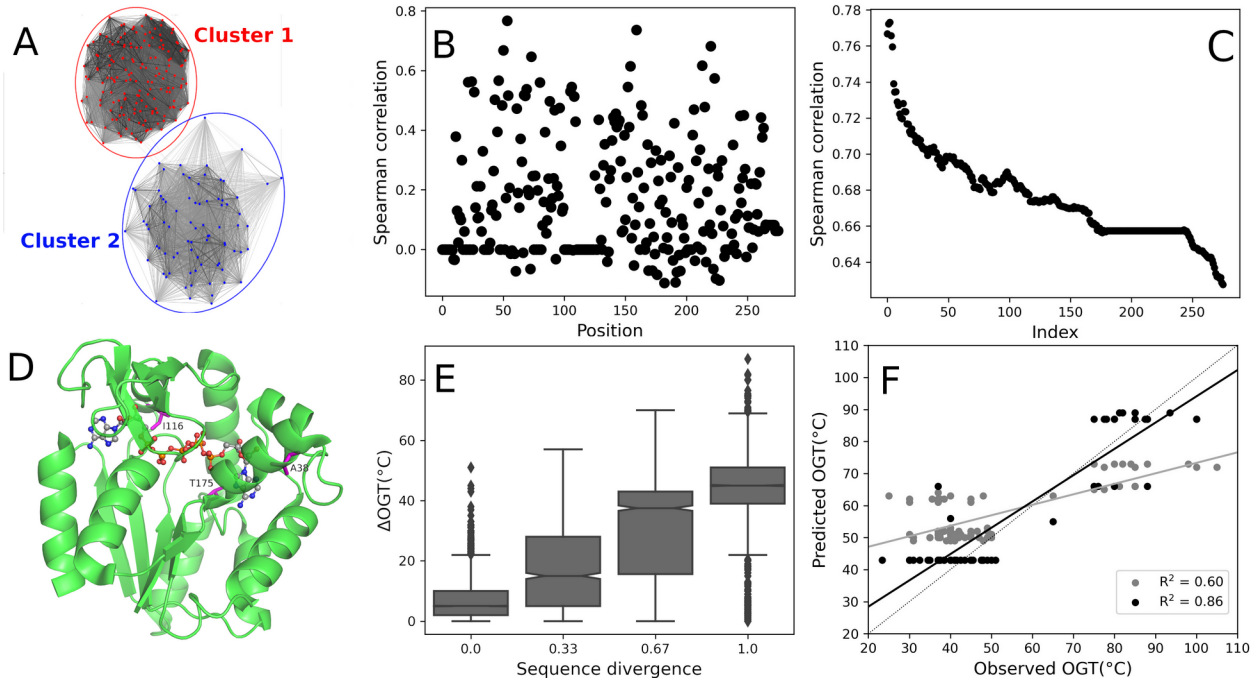

Figure 2. Optimal Growth Temperatures associated with Archaeal Adenylate Kinase (AdK). A. Sequence clusterization according to pairwise similarity threshold. B. Spearman rank-ordered correlated coefficients per position in the MSA. C. Correlation coefficients of the AdK accumulated correlation by addition of positions ranked in a descendent order (black). D. Positions that maximize the correlation mapped in a high resolution structure (PDB ID 1AKE). E. Correlation between the similarity in the positions that maximize the correlation and the *ΔOGT*. F. Prediction of OGT from sequences from the most correlated positions (black) and the whole protein positions (gray).

Table 1. Correlation analysis for the AdK protein family.

Alignment position		54	160	221
Position in PDB 1AKE		38	116	175
ρ-values of the individual positions		0.770	0.740	0.680
Most frequent amino acid residue	**≤ 60°C**	N	L	T
	> 60°C	E	E	M

The case of microbial rhodopsins (BR)

To study if PhISCO can be used in other scenarios, we performed calculations on a diverse set of microbial rhodopsins. They belong to a ubiquitous photoreceptive membrane

protein family that has diverse functions as light-driven ion pumps, light-gated ion channels, photochromatic gene regulators and light-regulated enzymes. These proteins bind all-trans retinal molecules via a protonated Schiff-base linkage and exhibit a variety of specific visible absorption wavelengths. This wide-range color tuning of the retinal is considered to be achieved by optimizing the steric and/or electrostatic interaction with surrounding amino-acid residues [4]. Based on a previous work by *Karasuma et al*.[4] and *Inoue et al*. [,32] we evaluated whether our method was able to identify positions associated with the fine tuning of the absorption-wavelength of microbial rhodopsins. To this end, we worked with the published data containing microbial rhodopsin sequences and associated wavelengths [4]. Analysis of the sequences using CLANS showed three main clusters with similar numbers of elements (**Figure 3A**). The least populated group showed an extremely narrow distribution of absorption wavelengths while the other two spanned broader ranges. Analysis of the major cluster revealed sequences ~30-60% identical that span a distribution of *ΔWavelengths* from 25 nm to 125 nm with a correlation of R^2 = 0.279 ± 0.004 (**Figure S1**). To seek positions that are more tightly linked to the phenotype, we studied the per-position *ρ*-value. The maximum value was 0.47, while the analysis of the joint correlation shows a peak when we used the ten most correlated positions (*ρ*-value = 0.65) (**Figure 3C-E and Table 2**). We mapped these positions in the crystal structure of *Halobacterium salinarum* (PDB ID 1C3W). Within them, positions 71 and 104 are in direct contact to the retinal molecule [4]. Modification of some loops in the Br from *H. salinarum*, involving positions 87, 58 and 51, have been reported to cause a maximal absorption wavelength shift [33]. The remaining positions are not in direct contact to retinal, however, positions 3, 10 and 12 are in direct interaction with the lipids located in the inter-monomer space. Noteworthy, mutation of position 19 was reported to have effects in maximal-absorption wavelength[34]. The sequence dataset was split in training and validation sets (50 % of randomly picked sequences each one). This random sampling was repeated 50 times to evaluate prediction robustness. A representative prediction plot with the mean R^2 value is shown in **Figure 3F**. The chosen set of positions improved the correlation ($R^2 = 0.72$) with respect to the whole sequence ($R^2 = 0.60$) suggesting that a small subset of positions is rich in phenotypic information and can be easily extracted using PhISCO

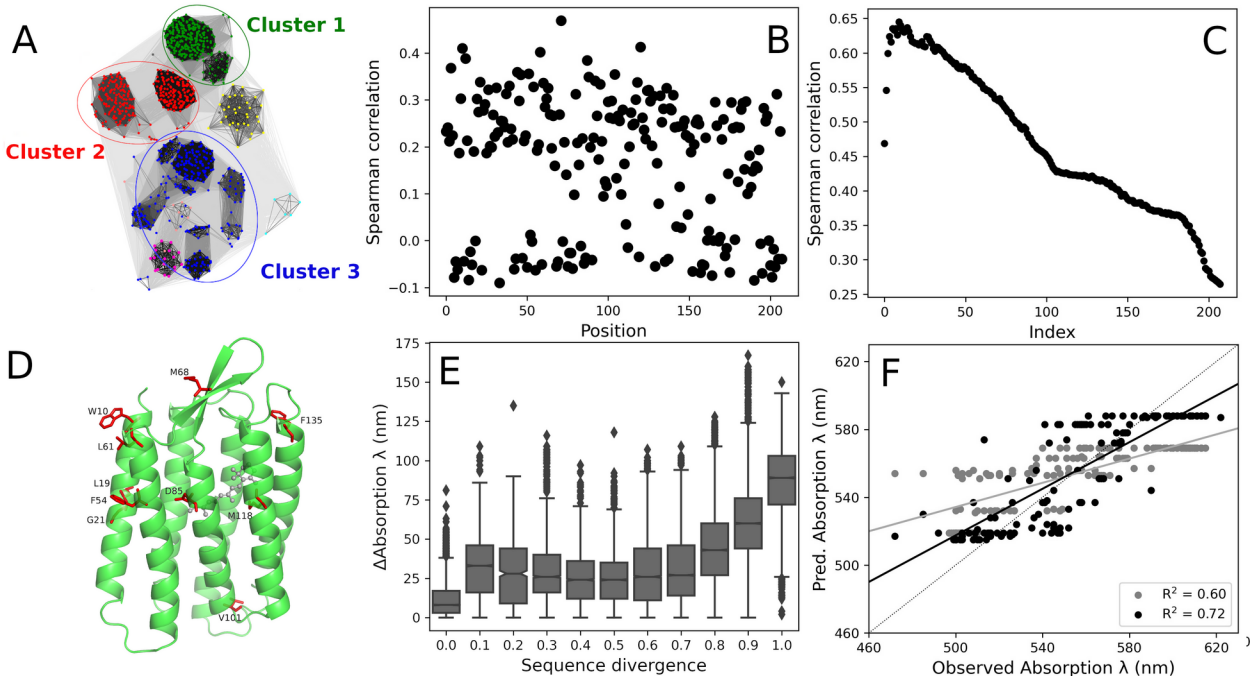

Figure 3. Maximum absorption wavelength of variants of bacterial rhodopsins (Br). A. Sequence clusterization according to pairwise similarity threshold. B. Spearman rank-ordered correlated coefficients per position in the MSA. C. Correlation coefficients of the microbial rhodopsins accumulated correlation by addition of positions ranked in a descendent order (black). D. Positions that maximize the correlation mapped in a high resolution structure (PDB ID 1C3W). E. Correlation between the similarity in the positions that maximize the correlation and the Δ absorption wavelength of the retinal. F. Prediction of absorption wavelength from sequences from the most correlated positions (black) and the whole protein positions (gray).

Table 2. Correlation analysis for the microbial rhodopsin protein family.

Alignment position		3	10	12	44	51	58	71	87	104	120
Position in PDB 1C3W		10	19	21	54	61	68	85	101	118	135
ρ-values of the individual positions		0.37	0.41	0.39	0.36	0.36	0.40	0.47	0.38	0.35	0.41
Most frequent amino acid residue	**≤ 550 nm**	L	G	L	A	A	P	D	A	V	Y
	> 550 nm	W	L	G	F	L	M	N	V	M	F

The case of mammalian myoglobin and its maximal muscle concentration

Aquatic mammals such as whales and seals have evolved mechanisms to hold their breath for periods of over an hour [35,36]. Mammals' diving capacity is tightly linked to the myoglobin concentration in muscle, which can be over 30-fold higher in mammalian elite divers as compared to land species [20,36,37]. In order to increase protein concentration, mammalian myoglobins have undergone extensive modifications to enhance solubility without triggering aggregation processes. Strategies to increase solubility highly rely on the increase of the net surface change [20], although alternative mechanisms such as increasing thermodynamic stability and hydrophobic patch shielding are also involved [38]. To evaluate positions that are linked to changes in myoglobin concentration, we studied a set of 65 mammalian myoglobin sequences from terrestrial and aquatic mammals with available myoglobin concentration data [20,38]. An inspection of the sequences using CLANS showed a single cluster (**Figure 4A**). The pairwise identity of these sequences are between 80% and 90% spanning a range of two orders of magnitude in the Δ*log10[Mb]* with no obvious correlation between them (R^2 = 0.241 ± 0.001) (Figure S1) . To evaluate if there are positions linked to the protein concentrations in the different organisms we studied per-position correlation. Most of the positions were invariant among sequences, then there are only few positions that had higher than zero *ρ-value* (mean value of 0.03 ± 0.07) **(Figure 4B**). The highest individual *ρ-value* was 0.41, while the most correlated positions (62, 101 and 152), when jointly analyzed, improved the *ρ-value* to 0.51 (**Figure 4C-E**). The mutation Q152H is almost exclusively found in diving mammals [20]. A correlation between thermodynamic stability and myoglobin cell concentration has been reported [39],[38]. Studies of the evolution of myoglobin from terrestrial mammals to the elephant seal note mutations in 18 positions. Among them, PhISCO identified positions 62, 101 and 152 [38]. Notably, 152 and 101 side chains are close in the reported structures, such as PDB ID 1ABS (Figure 4D) (Table 3). To perform *log10[Mb]* prediction, the sequence dataset was split in training and validation sets (50 % of randomly picked sequences each one). This random sampling was repeated 50 times to evaluate prediction robustness. A representative prediction plot with the mean R^2 value is shown in **Figure 4F**. Prediction of the *log10[Mb]* using these positions yielded a $R^2 = 0.61$ while using the whole sequence it was $R^2 = 0.44$.

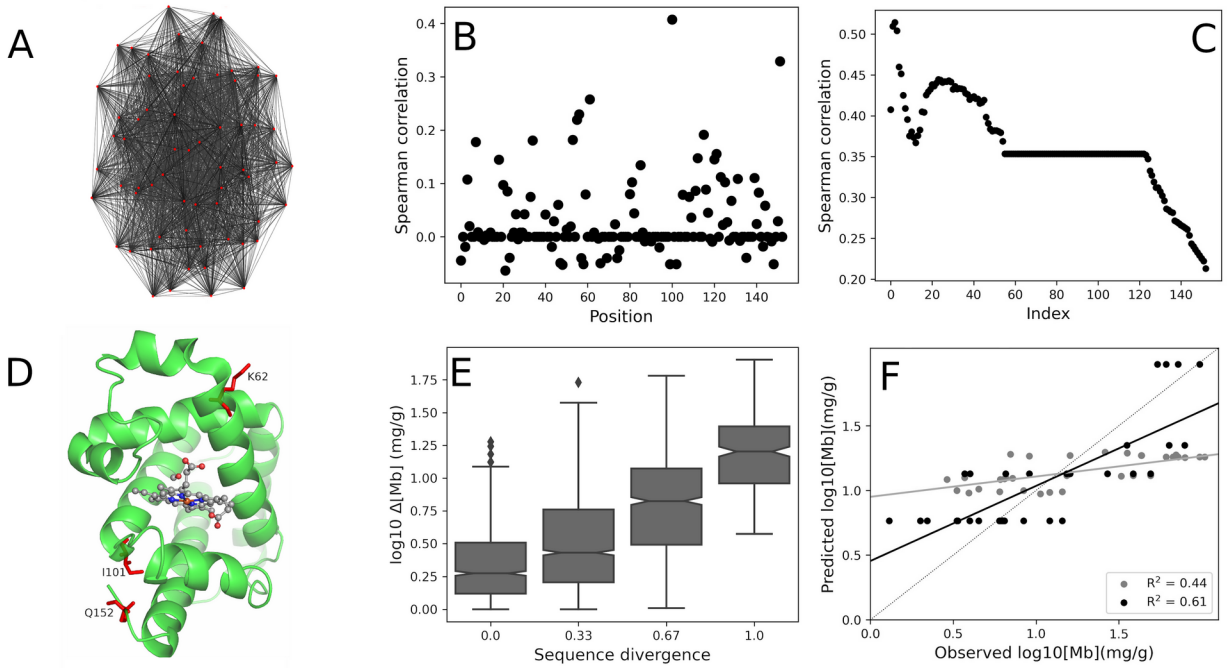

Figure 4. Myoglobin muscle concentrations in mammals (mmMb). A. Sequence clusterization according to pairwise similarity threshold. B. Spearman rank-ordered correlated coefficients per position in the MSA. C. Correlation coefficients of the mmMb accumulated correlation by addition of positions ranked in a descendent order (black). C. Positions that maximize the correlation mapped in a high resolution structure. E. Correlation between the similarity in the positions that maximize the correlation and the logarithm Δconcentration. F. Prediction of absorption wavelength from sequences from the most correlated positions (black) and the whole protein positions (gray).

Table 3. Correlation analysis for the myoglobin protein family.

Alignment position		62	101	152
Position in PDB 1ABS		62	101	152
ρ-values of the individual positions		0.770	0.740	0.680
Most frequent amino acid residue	**[Mb] ≤ 20 mg/g**	K	V	Q
	[Mb] > 20 mg/g	R	I	H

The case of HIV-protease and inhibition potency in clinical isolates

We tested the capability of our algorithm to predict sequence positions associated with antiviral drug resistance in HIV-1 strains. This phenomenon results mainly from prolonged patients-exposure to drug treatments, leading to the fixation and widespread distribution of resistant strains. We focus on the HIV-1 protease (HIV-PR), which is one of the most structurally characterized viral enzymes. HIV-PR is the target of several drugs known as protease inhibitors,

small molecules that are analogs to the transition state of the peptide substrate. Mutations of clinical relevance are observed in around 30% of residues in the HIV-PR sequence, we call those positions resistance determining positions (RDP). We studied the case of two inhibitors: fosamprenavir (FPV) and saquinavir (SQV). Resistance to these inhibitors is associated with different sets of RDP, with some overlapping (Table 4). Using PhISCO with the sequences recovered from HIVdb with their associated inhibition potency for each drug, we found the correlation between positions and the studied phenotype. PhISCO successfully identified most of the RDP for both inhibitors. In the top 20 highest scoring positions, we found 14 out of 18 RDP for FPV and 13 out of 18 for SQV (**Supplementary Table 1**). From the set of RDP relevant for FPV and SQV (**Table 4**)[40], Some of the identified positions are in direct binding to the inhibitor, as observed in positions 47, 50, 82 and 84. Mutations in these positions are mainly I84V, I47V, I50V and V82A, where substitutions result in loss of van der Waals interactions with the inhibitor. This is also the case for position 48 where substitution from G to V causes the valine side chain to impede forming key interactions with the drug. RDP involving more indirect effects were also captured, as involved in substitutions positions 54 and 73, where structural changes are propagated to active site residues, and positions 24, 50 and 53 where structure or stability of the inter-subunit interface is compromised, affecting inhibitor binding.

The availability of abundant data regarding the resistance positions for HIV-PR allowed us to examine the performance of our method using the Receiver Operating Characteristic (ROC) curve and their respective Area Under the Curve (AUC, **Figure 5**). We combined our scores with available data of RDP for each inhibitor, resulting in AUC values of 0.89 and 0.85 for FPV and SQV, respectively. From the 17 experimental RDP for FPV, 14 ranked above the cutoff value taken from the maximal Youden's index of the ROC curve. Globally, our method attained an accuracy for RDP prediction of 90 percent, with a Mathews correlation coefficient (MCC) of 0.71. In the case of SQV, 13 of a total of 18 RDP ranked above the cutoff, with an accuracy of 88 % and MCC 0.67. Noteworthy, for both inhibitors, PhISCO analysis of the positions joint-correlation showed a peak for the first 7 (*ρ-values* of 0.77 and 0.73 for FPV and SQV, respectively). All of them are described as RDP (**Figure S2**). When these positions were used to predict the levels of resistance, we observed correlation between predicted and observed values of R^2 = 0.60 and 0.52 for FPV and SQV, respectively. These values are higher than those obtained using the whole HIV-PR sequence, which renders R^2 = 0.40 and 0.31 for FPV and SQV, respectively (**Figure S2**).

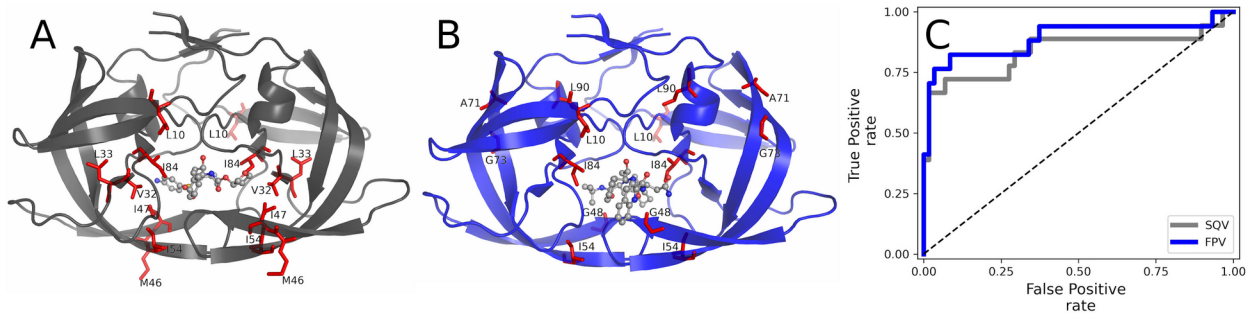

Figure 5. Predicted positions in HIV-PR sequence associated with inhibitor resistance. Most correlated positions (red sticks) for resistance to A. FPV (mapped in the crystal structure for the complex with

amprenavir, the active metabolite of FPV) or B. SQV (mapped in the crystal structure for the complex with this molecule). C. ROC analysis relating the per-position correlation coefficients of sequence divergence versus inhibitor resistance, with the experimentally determined RDP for each inhibitor.

Table 4. Prediction of resistance associated positions in HIV-PR.

		FPV			SQV		
Position	WT res.	Exp. RDP	Mut. Res.*	Prediction	Exp. RDP	Mut. Res.*	Prediction
10	L	+	F, I, R, V	TP	+	I, R, V	TP
11	V	+	I, L	TP	-		FP
20	K	+	T	TP	+	T	TP
24	L	+	I	TP	+	I	TP
32	V	++	I	TP	-		FP
33	L	++	F	TP	-		FP
46	M	++		TP	+	I, L	TP
47	I	++	V, A	TP	-		TN
48	G	-		TN	++	V, M	TP
50	I	++	V	FN	+		FN
53	F	-		FP	+	L	TP
54	I	++	V, A, S, T, L, M	TP	++	V, A, S, T, L, M	TP
62	I	-		TN	+	V	FN
71	A	-		FP	+	V, T	TP
73	G	+	S	TP	+	S, T, C, A	TP
74	T	-	P	TN	+	P	TP
76	L	++	V	TP	-		TN
77	V	-		TN	+	I	FN
82	V	++	A, F, S, T	TP	++	A, F, S, T	TP
83	N	-		TN	+	D	FN
84	I	++	V, A, C	TP	++	V, A, C	TP

88	N	++	S	FN	++	S, D	FN
89	L	+	V	FN	-		TN
90	L	++	M	TP	++	M	TP

+: accessory mutation, ++: major mutation, -: non relevant

TP: true positive, TN: true negative, FP: false positive, FN: false negative
cut off for RDP prediction from ROC analysis
*most frequently found in patients isolates
Exp. RDP: experimentally observed RDP; WT Res. and Mut. Res.: wild-type residue and the most frequently observed mutations.

Discussion

In this work, we found that in four heterogeneous protein families, a limited number of sequence positions are tightly linked to the studied phenotypes. This fact supports the notion that phenotypic adaptation can be achieved by a few sequence modifications. We reached this finding using an extremely simple strategy based on the correlation between sequence divergence and the quantitative phenotype differences, both obtained after pairwise comparisons. We developed a clustering heuristic which is based on the selection of a number of positions based on the *ρ*-value obtained when they are jointly analyzed (Figure 1B). This aided us in defining a set of residues associated with the explored phenotypes. As we show in the *Results* section, most of the positions identified by PhISCO were previously described as associated with function. For three of the studied cases, positions in direct interaction with ligands appear at the top of the scoring list. This is the case for Adk, Br and HIV-PR. Noteworthy, PhISCO also detects positions that are not in close proximity to the functional or catalytic sites. A rather indirect effect captured by our method can be mentioned for myoglobin, where there is a correlation between the net surface charge and its muscle concentration. Position 101 of this protein does not involve charged residues in our dataset, however it can still be associated with myoglobin concentration modulation. Interestingly, in close proximity to position 101 is the charged position 152. This suggests an interaction between both residues, where specific changes at position 101 may be required in order to accommodate a charged residue at position 152 without destabilizing the protein. It should be noted that our methodology for finding sequence positions associated with phenotype variations does not take explicitly into consideration any epistatic effect between sites. For example, the non-additive effect of the mutations at PhISCO-identified positions 89 and 90 (where mutations L to V and L to M occur, respectively) has been thoroughly studied in both a theoretical and experimental manner [41].This may be of particular relevance for phenotype prediction. A similar epistatic effect could be pointed out in the case of position 221 of AdK which is identified by PhISCO. Couñago et al. showed that the equivalent position in the *Bacillus subtilis* AdK (residue 179) is involved in protein adaptation to high temperatures [23]. The mutation T179I has a minor but significant effect in stability in the wildtype background sequence, while it has a significant stabilizing effect when it occurs in association with the Q199R mutation, which has no significant effect by itself.

Since the positions associated with resistance in HIV-PR are extensively characterized, this case gave us the opportunity to test our method in terms of sensitivity and specificity. In the first place, it was observed that most of the experimentally known RDPs appear among those best scored by PhISCO. In addition, assessing the cut-off value provided by the ROC analysis, which represents an equal weight of specificity and sensitivity, it was possible to adequately classify the correlated positions as RDP, with an accuracy of 90 %. Interestingly, the clustering strategy by PhISCO, based on the sequential addition of best scoring positions, resulted in selection of 7 positions, all of them corresponding to RDP. This result suggests that our method to analyze joint correlation favors specificity over sensitivity.

Since PhISCO identify positions tightly linked to phenotypes and, using a linear model, predict unknown phenotypes for target sequences with reasonable accuracy, it suggest that predictions can be performed with no explicit epistatic considerations. These predictions were

evaluated using the determination coefficient (R^2) finding values between 0.6 and 0.9, demonstrating that the residues involved are clearly linked to the studied phenotypes. Moreover, predictions performed using the whole sequences were clearly worse, which reinforces the idea of phenotypes being mostly codified in small subset of residues. It is interesting that the identity matrix implemented is the simplest penalization scheme, however, more complex amino acid substitution matrices may lead to similar or better results, although this was not explored in this work.

The simplicity of this method makes PhISCO useful to perform predictions in a diverse set of possible phenotypes. Also, the balance between speed and accuracy makes PhISCO a valuable tool to explore mutations that can be implemented in protein engineering.

References

1. Sułkowska, J. I., Morcos, F., Weigt, M., Hwa, T. & Onuchic, J. N. Genomics-aided structure prediction. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 10340–10345 (2012).
2. Basu, S. & Sen, S. Turning a mesophilic protein into a thermophilic one: a computational approach based on 3D structural features. *J. Chem. Inf. Model.* **49**, 1741–1750 (2009).
3. Louis, B. B. V. & Abriata, L. A. Reviewing Challenges of Predicting Protein Melting Temperature Change Upon Mutation Through the Full Analysis of a Highly Detailed Dataset with High-Resolution Structures. *Molecular Biotechnology* vol. 63 863–884 Preprint at https://doi.org/10.1007/s12033-021-00349-0 (2021).
4. Karasuyama, M., Inoue, K., Nakamura, R., Kandori, H. & Takeuchi, I. Understanding Colour Tuning Rules and Predicting Absorption Wavelengths of Microbial Rhodopsins by Data-Driven Machine-Learning Approach. *Sci. Rep.* **8**, 15580 (2018).
5. Pinney, M. M. *et al.* Parallel molecular mechanisms for enzyme temperature adaptation. *Science* **371**, (2021).
6. Gamiz-Arco, G. *et al.* Heme-binding enables allosteric modulation in an ancient TIM-barrel glycosidase. *Nature Communications* vol. 12 Preprint at https://doi.org/10.1038/s41467-020-20630-1 (2021).
7. Kashuk, C. S. *et al.* Phenotype-genotype correlation in Hirschsprung disease is illuminated by comparative analysis of the RET protein sequence. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 8949–8954 (2005).
8. Li, G., Rabe, K. S., Nielsen, J. & Engqvist, M. K. M. Machine Learning Applied to Predicting Microorganism Growth Temperatures and Enzyme Catalytic Optima. *ACS Synth. Biol.* **8**, 1411–1420 (2019).
9. Grønning, A. G. B. *et al.* DeepCLIP: predicting the effect of mutations on protein-RNA binding with deep learning. *Nucleic Acids Res.* **48**, 7099–7118 (2020).
10. Otwinowski, J., McCandlish, D. M. & Plotkin, J. B. Inferring the shape of global epistasis. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E7550–E7558 (2018).
11. Figliuzzi, M., Jacquier, H., Schug, A., Tenaillon, O. & Weigt, M. Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1. *Mol. Biol. Evol.* **33**, 268–280 (2016).
12. Sauer, D. B., Karpowich, N. K., Song, J. M. & Wang, D.-N. Rapid Bioinformatic Identification of Thermostabilizing Mutations. *Biophysical Journal* vol. 109 1420–1428 Preprint at https://doi.org/10.1016/j.bpj.2015.07.026 (2015).
13. Jessen, L. E., Hoof, I., Lund, O. & Nielsen, M. SigniSite: Identification of residue-level genotype-phenotype correlations in protein multiple sequence alignments. *Nucleic Acids Res.* **41**, W286–91 (2013).
14. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–45 (2016).
15. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
16. Myers, J. L., Well, A. D. & Lorch, R. F., Jr. *Research Design and Statistical Analysis: Third*

Edition. (Routledge, 2013).
17. Aptekmann, A. A. & Nadra, A. D. Core promoter information content correlates with optimal growth temperature. *Sci. Rep.* **8**, 1313 (2018).
18. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
19. Frickey, T. & Lupas, A. CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* **20**, 3702–3704 (2004).
20. Mirceta, S. *et al.* Evolution of mammalian diving capacity traced by myoglobin net surface charge. *Science* **340**, 1234192 (2013).
21. Rhee, S.-Y. *et al.* Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.* **31**, 298–303 (2003).
22. Wensing, A. M. *et al.* 2019 update of the drug resistance mutations in HIV-1. *Top. Antivir. Med.* **27**, 111–121 (2019).
23. Couñago, R., Chen, S. & Shamoo, Y. In vivo molecular evolution reveals biophysical origins of organismal fitness. *Mol. Cell* **22**, 441–449 (2006).
24. Rusnak, P., Haney, P. & Konisky, J. The adenylate kinases from a mesophilic and three thermophilic methanogenic members of the Archaea. *J. Bacteriol.* **177**, 2977–2981 (1995).
25. Davlieva, M. & Shamoo, Y. Crystal structure of a trimeric archaeal adenylate kinase from the mesophile Methanococcus maripaludis with an unusually broad functional range and thermal stability. *Proteins* **78**, 357–364 (2010).
26. Bönisch, H., Backmann, J., Kath, T., Naumann, D. & Schäfer, G. Adenylate kinase from Sulfolobus acidocaldarius: expression in Escherichia coli and characterization by Fourier transform infrared spectroscopy. *Arch. Biochem. Biophys.* **333**, 75–84 (1996).
27. Olsson, U. & Wolf-Watz, M. Overlap between folding and functional energy landscapes for adenylate kinase conformational change. *Nat. Commun.* **1**, 111 (2010).
28. Armenta-Medina, D., Pérez-Rueda, E. & Segovia, L. Identification of functional motions in the adenylate kinase (ADK) protein family by computational hybrid approaches. *Proteins: Structure, Function, and Bioinformatics* vol. 79 1662–1671 Preprint at https://doi.org/10.1002/prot.22995 (2011).
29. Saavedra, H. G., Wrabl, J. O., Anderson, J. A., Li, J. & Hilser, V. J. Dynamic allostery can drive cold adaptation in enzymes. *Nature* **558**, 324–328 (2018).
30. Miyashita, O., Onuchic, J. N. & Wolynes, P. G. Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proceedings of the National Academy of Sciences* vol. 100 12570–12575 Preprint at https://doi.org/10.1073/pnas.2135471100 (2003).
31. Whitford, P. C., Miyashita, O., Levy, Y. & Onuchic, J. N. Conformational transitions of adenylate kinase: switching by cracking. *J. Mol. Biol.* **366**, 1661–1671 (2007).
32. Inoue, K. *et al.* Exploration of natural red-shifted rhodopsins using a machine learning-based Bayesian experimental design. Preprint at https://doi.org/10.1101/2020.04.21.052548.
33. Kim, J. M., Booth, P. J., Allen, S. J. & Khorana, H. G. Structure and function in bacteriorhodopsin: the role of the interhelical loops in the folding and stability of bacteriorhodopsin. *J. Mol. Biol.* **308**, 409–422 (2001).
34. Structure-Function Studies on Bacteriorhodopsin: IX. Substitutions of Tryptophan Residues

Affect Protein-retinal Interactions in Bacteriorhodopsin. *J. Biol. Chem.* **264**, 14197–14201 (1989).
35. Watwood, S. L., Miller, P. J. O., Johnson, M., Madsen, P. T. & Tyack, P. L. Deep-diving foraging behaviour of sperm whales (Physeter macrocephalus). *J. Anim. Ecol.* **75**, 814–825 (2006).
36. Stewart, B. S. & DeLONG, R. L. *Double migrations of the northern elephant seal, Mirounga angustirostris*. (1995).
37. Berenbrink, M. The role of myoglobin in the evolution of mammalian diving capacity - The August Krogh principle applied in molecular and evolutionary physiology. *Comp. Biochem. Physiol. A Mol. Integr. Physiol.* **252**, 110843 (2021).
38. Isogai, Y. *et al.* Common and unique strategies of myoglobin evolution for deep-sea adaptation of diving mammals. *iScience* **24**, 102920 (2021).
39. Dasmeh, P., Serohijos, A. W. R., Kepp, K. P. & Shakhnovich, E. I. Positively selected sites in cetacean myoglobins contribute to protein stability. *PLoS Comput. Biol.* **9**, e1002929 (2013).
40. Weber, I. T. & Agniswamy, J. HIV-1 Protease: Structural Perspectives on Drug Resistance. *Viruses* **1**, 1110–1136 (2009).
41. Henes, M. *et al.* Molecular Determinants of Epistasis in HIV-1 Protease: Elucidating the Interdependence of L89V and L90M Mutations in Resistance. *Biochemistry* **58**, 3711–3726 (2019).