PhISCO: a simple method to infer phenotypes from protein sequences

Ayelén S. Hernandez-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.

Index

Abstract, keywords and abbreviations……3
Introduction……..5
Results……..……..……8
- PhISCO overview……,...……..8
- The case of archaeal adenylate kinases (ADK).....….....9
- The case of microbial rhodopsins (BR) ……11
- The case of mammalian myoglobin and its maximal muscle concentration ……..12
- The case of HIV-protease and inhibition potency in clinical isolates ……...14

Discussions……...16
- Analysis of identified positions……16
- Comparison with state-of-the art methodologies……..18
- Concluding remarks…….……..19

Materials and methods……..21
- Sequence datasets with their associated-phenotypes…… 21
- Phenotype predictions ……...……..22
- Receiver operating characteristic (ROC) equations ……23
- Data analysis and scripts deposition ……24

References……..26
Plots and tables:
- Figure 1. Detailed workflow of PhISCO algorithm……31
- *Figure 2. Optimal Growth Temperatures associated with Archaeal Adenylate Kinase (AdK). ……..33*
- *Figure 3. Maximum absorption wavelength of variants of bacterial rhodopsins (Br)……..34*
- Figure 4. Myoglobin muscle concentrations in mammals (mmMb)……...35
- Figure 5. Predicted positions in HIV-PR sequence associated with inhibitor resistance. …….36
- Table 1. Correlation analysis for the adk protein family……..37
- Table 2. Correlation analysis for the microbial rhodopsin protein family……..38

- table 3. Correlation analysis for the mammals myoglobin protein family…….39
- *table 4. Prediction of resistance-associated positions in HIV-PR. ……....40*

Supplementary information document:

- Table S1. Top 20 scoring positions for resistance of HIV-PR to PI ……3
- Table S2. Comparison between PhISCO and SigniSite performance1……..4
- Figure S1. Whole sequence divergence and *Δphenotypes*……..*5*
- Figure S2. Inhibitory potency of Fosamprenavir (FPV) and Saquinavir (SQV) in clinical isolates of HIV-protease (Hp)……6
- *Figure S3. Control calculations for PhISCO's joint correlation analysis*……..*7*
- *Figure S4. Positions involved in different traits associated with myoglobin*……....*8*
- *Figure S5. Performance of PhISCO predictions in 100 random selected training and validation datasets for the ADK case*…….. *9*
- *Figure S6. Performance of PhISCO predictions in 100 random selected training and validation datasets in the microbial rhodopsins case*……..*10*
- *Figure S7. Performance of PhISCO predictions in the 100 random selected training and validation datasets in the myoglobin case*……..…..*11*
- *Figure S8. Performance of PhISCO predictions in the 100 random selected training and validation datasets for HIV-PR inhibition with FPV*……*12*
- *Figure S9. Performance of PhISCO predictions in the 100 random selected training and validation datasets for HIV-PR inhibition with SQV*……....*13*
- *Figure S10. Performance of PhISCO in an alignment with 100 sequences and different number of positions (open triangles) and with 100 of positions and different number of sequences (black circles)*……...*14*

Supplementary information file:

- PHISCO.ipynb (a jupyter-notebook with the code to run PHISCO)
- Files are deposited in: https://doi.org/10.5281/zenodo.7761213.

Abstract

Although protein sequences encode the information for folding and function, understanding their link is not an easy task. Unluckily, 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. We found that from 3 to 10 positions, depending on the studied case, were enough to identify positions associated with the phenotypes and perform quantitative predictions of them. Here we show that these strong correlations can be found using individual 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 and differences in observed phenotypes. We also show that although extremely simple, predictions are comparable to the state-of-art methodologies which, in most of the cases, are far more complex. All of the calculations are obtained at a very low information cost since the only input needed is a multiple sequence alignment of protein sequences with their associated quantitative phenotype. The diversity of the explored systems makes PhISCO a valuable tool to find sequence determinants of biological activity modulation and to predict various functional features for uncharacterized members of a protein family.

Keywords

Phenotype prediction, adaptation, protein function, sequence information, sequence divergence.

Abbreviations

PhISCO: Phenotype Inference from Sequence COmparisons, OGT: Optimal Growth Temperature, MSA: Multiple Sequence Alignment, Adk: Adenylate kinase, HIVdb: Stanford University HIV Drug Resistance Database, RDP: resistance determining positions, BR: microbial rhodopsin, ATPlid: ATP lid domain, AMPlid: AMP lid domain, HIV-PR: HIV-1 protease, SQV: saquinavir, FQV: fosamprenavir, AUC: area under the curve, MCC: Mathews correlation coefficient, ROC: Receiver Operating Characteristic.

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.

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.

In this sense, variations in the sequences may be reflected in variations in observed phenotypes. However, usually, the changes in the observed phenotypes do not appear in an additive way ('additive' means that the effect of each individual mutation accumulates as a simple sum of effects). In many cases, accumulation of mutations does not have a linear counterpart in the observed trait. This type of phenomena is called epistasis and usually named

as a non-additive effect of mutations. A reasonable hypothesis is mentioned in Otwinowski J. et al. posing that mutations have an additive effect in an unobserved trait which is related in a non-linear way to the observed trait [9].

Multiple existing algorithms aim at the quantitative prediction of phenotypes. Some of them take a protein structure as a starting point, while others are sequence-based. Structure-based predictions of protein stability are time-consuming, require expert supervision, and are moderately accurate [2]. Although with the aid of structure predictors as AlphaFold2 the structure-requirement is no longer an obstacle for most targets, predictions of functional properties are still not always performed with significant accuracy. Sequence-based prediction algorithms attempt to associate amino acid abundances and/or physical properties at specific sequence positions with phenotype values such as the optimal working temperature of a protein/organism [3-5], the impact of a mutation in disease [6], binding affinity [7] or the spectroscopic properties of a protein [8]. Since the nature of the problems tackled is quite diverse, it comes to no surprise that individual methods are tailored towards specific questions. As a consequence their applicability is very limited [9]-[10]. Their accuracy is quite diverse ranging from moderately low to high and many of these methods are based on machine learning techniques requiring extensive training data sets and suffer from parameter interpretation problems [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 in a protein family, it is expected that protein variants working in similar conditions exhibit similar residues at these positions. To search these residues, we performed calculations to correlate per-position sequence differences (“sequence divergence”) and the corresponding phenotype differences of the involved

sequences. At this point, two considerations are important to support this notion: i) the phylogenetically distant organisms would obtain similar residues due to a convergence phenomenon and ii) these positions would have a "dominant" effect acting in an independent manner from the sequence background (epistasis). Given the absence of epistasis, the historical contingency in phenotype modulation would be absent. Then, although different lineages could develop independent adaptation strategies, the same sequence changes would alter the phenotype in a similar way regardless of the lineage in which they operate.

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 which has been studied by several methods, a set of microbial rhodopsins whose retinal cofactor absorption occurs at different wavelengths [8], a set of myoglobin sequences from mammals and their maximal concentration in muscle, 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.

Results

PhISCO overview

In this work we developed an algorithm that calculates per-position sequence divergences in the alignment as well as the associated phenotype differences (Figure 1A). Divergences are calculated using an identity matrix, where values of 0 or 1 are assigned to identical or different amino acids, and phenotype differences are treated as the corresponding absolute values (ΔPhenotypes in Figure 1 and equations thereof). Association between these two magnitudes is evaluated position by position using the Spearman rank-order correlation coefficient (ρ) [37], from which the positions more correlated with phenotypes can be extracted and studied. In addition, a set of positions highly correlated with the phenotypes can be used for the prediction of the phenotype value of uncharacterized sequences. For this end, we developed a specific heuristic (Figure 1B) to explore the existence of a set of positions that, when analyzed jointly, renders a higher correlation than when individually analyzed. Briefly, the correlations between sequence divergence and phenotype differences are calculated for sets of positions that increase in size by the addition of one position at a time, starting from the most correlated individual position up to the least. Thus, at the beginning of the calculation, the ρ-value corresponds to the position with highest correlation; in the next step, it sums the divergence of the second highest correlated position to that of the former position and recalculates the ρ-value. The process is repeated until all of the positions are included (Figure 1B).

The group of positions with highest ρ-value are then used for phenotype prediction of target sequences. For this end the training sequences (those with already known phenotypes) are used for pairwise comparisons as described above using the group of positions with highest ρ-value, and the correlation between sequence divergence and ΔPhenotype is calculated using a linear equation (Figure 1C). The equation coefficients derived from the linear regression are then used to extrapolate the ΔPhenotype from the divergence values at the specified positions of the target sequence to those of the training set. Finally, the target phenotype is derived from minimization of a sum of squares calculated from extrapolated ΔPhenotypes and those calculated from the difference in the unknown (target) phenotype and the known phenotypes of the training set. A detailed description of the minimization procedure and equations is included in the Methods section. In the rest of the results section we discuss the application of this algorithm to four studied cases and analyze the 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 [14]. One of these key enzymes is AdK, which plays a central role in energy metabolism by regulating the ATP concentration [14,15]. 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. To constrain the analysis to sequences from organisms subjected to similar adaptive strategies we performed a clusterization analysis using CLANS as described in Methods section. 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 [16–18], 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**). 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 (38 in *E.coli* sequence, Table 1) is located in the AMP lid domain (AMPlid), positions 160 (116 in *E.coli* sequence, Table 1) and 221 (175 in *E.coli* sequence, Table 1) are located in the ATP lid domain (ATPlid) [19]. Computational analysis of coevolution combined with molecular dynamics simulations showed that residue 38 is highly coupled to spatially distant positions [20]-[21]. Regarding positions in the ATPlid, when the mutation of residue 116 from I to G is introduced in the *E. coli* AdK ATP binding is improved without significantly protein stability effect but decreasing catalytic efficiency [19]. Also, structure-based simulations and normal mode analysis showed that both residues 116 and 175 are involved in local unfolding events with effects in AdK catalysis [21,22]-[23]. 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.

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 [8]. Based on a previous work by *Karasuma et al*.[8] and *Inoue et al*. [24] 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 [8]. 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 $R^2 = 0.279 \pm 0.004$ (**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 (85 in *H. salinarum* sequence, Table 2) and 104 (118 in *H. salinarum* sequence, Table 2) are in direct contact to the retinal molecule [8]. Modification of some loops in the Br from *H. salinarum*, involving positions 87, 58 and 51 (101, 68 and 61 in *H. salinarum* sequence, respectively, Table 2), have been reported to cause a maximal absorption wavelength shift [25]. The remaining positions are not in direct contact to retinal, however, positions 3, 10 and 12 (10, 19 and 21 in *H. salinarum* sequence, respectively, Table 2) are in direct interaction with the lipids located in the inter-monomer space. Noteworthy, mutation of position 10 was reported to have effects in maximal-absorption wavelength [26]. 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.

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 [27]-[28]. 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 [28,29]-[30]. 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 [29], although alternative mechanisms such as increasing thermodynamic stability and hydrophobic patch shielding are also involved [31]. 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 [29,31]. 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 *p-value* (mean value of 0.03 ± 0.07) **(Figure 4B**). The highest individual *p-value* was 0.41, while the most correlated positions (62, 101 and 152), when jointly analyzed, improved the *p-value* to 0.51 (**Figure 4C-E**). The mutation Q152H is almost exclusively found in diving mammals [29]. A correlation between thermodynamic stability and myoglobin cell concentration has been reported [31]-[32]. 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 [31]. 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 $R^2 = 0.44$.

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 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 16 RDP for FPV and 12 out of 15 for SQV (**Supplementary Table 1**). From the set of RDP relevant for FPV and SQV (**Table 4**) [33], 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 0.94 and 0.89 for FPV and SQV, respectively. From the 16 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 91 percent, with a Mathews correlation coefficient (MCC) of 0.75. In the case of SQV, 12 of a total of 15 RDP ranked above the cutoff, with an accuracy of 90 percent and MCC 0.69. 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**). To predict the levels of resistance we divided the input sequences in training and validation sets by randomly choosing 50% of the sequences for the training set and the rest for the validation set (repeated 50 times). 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**).

Discussion

Analysis of the identified positions

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. However, positions linked to non-additivity effects may be captured. The simplicity of the heuristic chosen to find positions make PhISCO

unable to differentiate between additive and non-additive positions. For example, in the case of HIV-PR 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 [34]. This may be of particular relevance for phenotype prediction. Moreover, figure S2B and S2E suggest that the association between divergence and the differences in the antiviral drug inhibition is not linear, giving room to further improvements in the phenotype prediction by including epistatic effects. A similar epistatic effect could be pointed out in the case of residue 175 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 [15]. 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. Although the found positions have been previously documented to be involved in enzyme function modulation and that these positions are used with success to predict the phenotype, conclusions should be taken with care. Due to the phylogenetic structure of the organisms involved in the analysis, some neutral mutations may arise as a consequence of a fixation in the different lineages. The actual version of PhISCO is not taking into account this effect.

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.

Comparison to the state-of-the art methodologies

Although there is a myriad of available methods that do phenotype predictions, our proposed algorithm is one of the few that can be used for a diverse set of quantitative phenotypes. We designed PhISCO as a simple and interpretable algorithm for the prediction of quantitative phenotypes. Compared to other existing methods, both PhISCO and Signisite [13] allow obtaining positions associated with the phenotype by studying the relationship between a set of aligned sequences and their corresponding phenotype, expressed as a continuous quantifiable variable. Signisite analyzes the residue composition at a given position based on the magnitude of its associated phenotype; briefly, it orders the sequences according to their phenotype and analyzes the abundance of each amino acid according to its position in the ranking. On the other hand, PhISCO identifies positions performing a correlation of its sequence divergence with its phenotype difference. The performance of both methods can be compared in the case of HIV-PR, where the relevant positions for modulating the phenotype are known. As we can see in Table S2, both methods perform similarly in terms of accuracy and specificity, with PhISCO's sensitivity being remarkably higher when using the cutoff derived from the ROC analysis. Although both methods are simple and fast, PhISCO adds a new capability which is the ability to make predictions of the phenotype value for sequences with unknown phenotype values. We compared our method with some of the available state-of-art strategies. For OGT prediction, the method of Li et. al. [11] which is a machine learning algorithm based on full proteome information, attains a RMSE (root-mean squared error) value of 2.16 °C and R^2 of 0.95. The multiple linear regression method reported by Sauer et. al. [12], which uses several genome-derived features, attains a RMSE of 4.93 °C and R^2 of 0.767. Even though both methods perform better than PhISCO in OGT prediction (RMSE of 8.5 °C and R^2 of 0.85), it should be remembered that the performance obtained by this method is attained by the solely use of a protein family (Figure S5). In the case of predictions of absorption wavelengths of

bacterial rhodopsins, PhISCO performs with a MAE (mean-absolute error) of 17.3 nm and R^2 of 0.72, which is in the same order as for the machine learning-based method published by Karasuyama et al. (MAE of 7.8 nm and R^2 of 0.774 and MAE of 12.4 nm and R^2 of 0.8594, for 2 different rhodopsin subgroups). Although performance is better in the case of the Karasuyama et al., it comes with a cost of expensive training of thousands of parameters of a machine learning algorithm. Also, this training is performed for a specific protein family as a difference with PhISCO which is easily tunable for different protein families and phenotypes (Figure S6). Moreover, when studying the inhibitory efficiency for molecules that bind the HIV-PR we found that Beerenwinkel et al. developed the Gene2Pheno [42] algorithm based on support vector machines and perform predictions with a MSE = 0.204 and R^2 = 0.71 for SQV inhibitor. PhISCO yielded a MSE = 0.21 with a R^2 = 0.73. Although the cited publication is 20 years old, we were not able to find a recent software that performs this prediction for a similar dataset. It should be remembered at this point that most of the algorithms perform position finding or phenotype classification. In this sense, PhISCO do both position finding and phenotype prediction. For the case of myoglobin, no prediction method of its muscular concentration is available as far as we know.

Concluding remarks

Since PhISCO identify positions tightly linked to phenotypes and, using a linear model, predict unknown phenotypes for target sequences with reasonable accuracy, it suggests that predictions can be performed with no explicit epistatic considerations. PhISCO uses a minimal set of parameters for predictions (the two coefficients of linear regression) which reflect an acceptable accuracy level even when a very limited number of data in the training set is available. Also, 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.

Overall, PhISCO is a generalist algorithm that performs at a level comparable to machine-learning based methods without the need for extensive training. Also the method is really simple since predictions are obtained at a very low information cost with results comparable to the state-of-art methodologies which are, in most of the cases, far more complex. In sum, we consider PhISCO a successful first-generation interpretable algorithm for the prediction of quantitative phenotypes and the associated sequence positions.

Materials and methods

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. [14]. 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 [38]. Multiple sequence alignments (MSA) were performed using the online server MAFFT with default settings. Columns with more than 10% of gaps were eliminated. As another study case, we studied a set of sequences of microbial rhodopsins with their light-absorption wavelengths obtained from *Karasuma et al.* [8]. The alignment and trimming strategies were the same as in the case of AdK. In a third example, we worked with a set of extant mammalian myoglobin sequences whose alignment and maximal muscle concentration were obtained from Mirceta et al. [29]. 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. [40] (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 [41].

Cluster analysis for all the studied cases were performed using CLANS [39], a clusterization algorithm based on *E*-values from pairwise-sequence comparisons, using a Network based method with global average.

Phenotype predictions

With the aim to evaluate the performance of PhISCO's phenotype prediction, the input sequences were assigned to two groups: a training set and 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:

$$\Delta Phenotype,\ obs(i,j) = a\ \{Seqdivergence\}(i,j) + b \qquad (1)$$

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 . From these divergences, the values of the predicted ΔPhenotype were calculated using the fitted coefficients of Equation 1, according to:

$$\Delta Phenotype,\ pred(k,j) = a\ \{Seqdivergence\}(k,j) + b \qquad (2)$$

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)):

$$\frac{\Sigma}{j}((|Phenotype,obs(j)-Phenotype,pred(k)|)-\Delta Phenotype,pred(k,j))^2 \qquad (3)$$

Receiver operating characteristic (ROC) equations

For ROC analysis in the HIV-PR case, the number of correctly predicted RDPs and non RDPs are indicated as true positives (TP) and true negatives (TN), whereas the number of incorrectly predicted RDPs and non RDPs are indicated as false negatives (FN) and false positives (FP), respectively. From these values, we calculated Sensitivity and Specificity, as defined below:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \qquad (4)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \qquad (5)$$

The receiver operator characteristic (ROC) curve was derived from these measures calculated at different cutoff values (for review see [43]) and from that we obtained the area under the ROC curve (AUC). An "optimal" cutoff point was calculated from the maximal Youden's index (J), defined as:

$$J = \text{Sensitivity} + \text{Specificity} - 1 \qquad (6)$$

Finally, the accuracy and the Matthews correlation coefficient were calculated as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \qquad (7)$$

$$\text{MCC} = \frac{TPxTN - FPxFN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \qquad (8)$$

Data analysis and scripts deposition

The calculations scripts were developed using standard python libraries and an implementation using Jupyter Notebooks. Both these scripts and the MSAs are deposited in the following repository: https://doi.org/10.5281/zenodo.7761213.

Accession numbers

PDB ID **1AKE**, PDB ID **1C3W**, PDB ID **1ABS,** PDB ID **3OXC,** PDB ID **3EKV.**

Acknowledgements. This work was funded by the *CONICET* (PIP 2017) and *Agencia Nacional de Promoción de la Investigación, el Desarrollo Tecnológico y la Innovación* (PICT 2016 0014). E.A.R. thanks M.D. Ph.D. Lázaro Efraín Gerschenson for always trying to do the best he can with what he has.

References

[1] J.I. Sułkowska, F. Morcos, M. Weigt, T. Hwa, J.N. Onuchic, Genomics-aided structure prediction, Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 10340–10345.

[2] S. Basu, S. Sen, Turning a mesophilic protein into a thermophilic one: a computational approach based on 3D structural features, J. Chem. Inf. Model. 49 (2009) 1741–1750.

[3] B.B.V. Louis, L.A. Abriata, 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. 63 (2021) 863–884. https://doi.org/10.1007/s12033-021-00349-0.

[4] M.M. Pinney, D.A. Mokhtari, E. Akiva, F. Yabukarski, D.M. Sanchez, R. Liang, T. Doukov, T.J. Martinez, P.C. Babbitt, D. Herschlag, Parallel molecular mechanisms for enzyme temperature adaptation, Science. 371 (2021). https://doi.org/10.1126/science.aay2784.

[5] G. Gamiz-Arco, L.I. Gutierrez-Rus, V.A. Risso, B. Ibarra-Molero, Y. Hoshino, D. Petrović, J. Justicia, J.M. Cuerva, A. Romero-Rivera, B. Seelig, J.A. Gavira, S.C.L. Kamerlin, E.A. Gaucher, J.M. Sanchez-Ruiz, Heme-binding enables allosteric modulation in an ancient TIM-barrel glycosidase, Nature Communications. 12 (2021). https://doi.org/10.1038/s41467-020-20630-1.

[6] C.S. Kashuk, E.A. Stone, E.A. Grice, M.E. Portnoy, E.D. Green, A. Sidow, A. Chakravarti, A.S. McCallion, Phenotype-genotype correlation in Hirschsprung disease is illuminated by comparative analysis of the RET protein sequence, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 8949–8954.

[7] A.G.B. Grønning, T.K. Doktor, S.J. Larsen, U.S.S. Petersen, L.L. Holm, G.H. Bruun, M.B. Hansen, A.-M. Hartung, J. Baumbach, B.S. Andresen, DeepCLIP: predicting the effect of mutations on protein-RNA binding with deep learning, Nucleic Acids Res. 48 (2020) 7099–

7118.

[8] M. Karasuyama, K. Inoue, R. Nakamura, H. Kandori, I. Takeuchi, Understanding Colour Tuning Rules and Predicting Absorption Wavelengths of Microbial Rhodopsins by Data-Driven Machine-Learning Approach, Sci. Rep. 8 (2018) 15580.

[9] J. Otwinowski, D.M. McCandlish, J.B. Plotkin, Inferring the shape of global epistasis, Proc. Natl. Acad. Sci. U. S. A. 115 (2018) E7550–E7558.

[10] M. Figliuzzi, H. Jacquier, A. Schug, O. Tenaillon, M. Weigt, Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1, Mol. Biol. Evol. 33 (2016) 268–280.

[11] G. Li, K.S. Rabe, J. Nielsen, M.K.M. Engqvist, Machine Learning Applied to Predicting Microorganism Growth Temperatures and Enzyme Catalytic Optima, ACS Synth. Biol. 8 (2019) 1411–1420.

[12] D.B. Sauer, N.K. Karpowich, J.M. Song, D.-N. Wang, Rapid Bioinformatic Identification of Thermostabilizing Mutations, Biophysical Journal. 109 (2015) 1420–1428. https://doi.org/10.1016/j.bpj.2015.07.026.

[13] L.E. Jessen, I. Hoof, O. Lund, M. Nielsen, SigniSite: Identification of residue-level genotype-phenotype correlations in protein multiple sequence alignments, Nucleic Acids Res. 41 (2013) W286–91.

[14] A.A. Aptekmann, A.D. Nadra, Core promoter information content correlates with optimal growth temperature, Sci. Rep. 8 (2018) 1313.

[15] R. Couñago, S. Chen, Y. Shamoo, In vivo molecular evolution reveals biophysical origins of organismal fitness, Mol. Cell. 22 (2006) 441–449.

[16] P. Rusnak, P. Haney, J. Konisky, The adenylate kinases from a mesophilic and three thermophilic methanogenic members of the Archaea, J. Bacteriol. 177 (1995) 2977–2981.

[17] M. Davlieva, Y. Shamoo, Crystal structure of a trimeric archaeal adenylate kinase from the mesophile Methanococcus maripaludis with an unusually broad functional range and

thermal stability, Proteins. 78 (2010) 357–364.

[18] H. Bönisch, J. Backmann, T. Kath, D. Naumann, G. Schäfer, Adenylate kinase from Sulfolobus acidocaldarius: expression in Escherichia coli and characterization by Fourier transform infrared spectroscopy, Arch. Biochem. Biophys. 333 (1996) 75–84.

[19] U. Olsson, M. Wolf-Watz, Overlap between folding and functional energy landscapes for adenylate kinase conformational change, Nat. Commun. 1 (2010) 111.

[20] D. Armenta-Medina, E. Pérez-Rueda, L. Segovia, Identification of functional motions in the adenylate kinase (ADK) protein family by computational hybrid approaches, Proteins: Structure, Function, and Bioinformatics. 79 (2011) 1662–1671. https://doi.org/10.1002/prot.22995.

[21] H.G. Saavedra, J.O. Wrabl, J.A. Anderson, J. Li, V.J. Hilser, Dynamic allostery can drive cold adaptation in enzymes, Nature. 558 (2018) 324–328.

[22] O. Miyashita, J.N. Onuchic, P.G. Wolynes, Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins, Proceedings of the National Academy of Sciences. 100 (2003) 12570–12575. https://doi.org/10.1073/pnas.2135471100.

[23] P.C. Whitford, O. Miyashita, Y. Levy, J.N. Onuchic, Conformational transitions of adenylate kinase: switching by cracking, J. Mol. Biol. 366 (2007) 1661–1671.

[24] K. Inoue, M. Karasuyama, R. Nakamura, M. Konno, D. Yamada, K. Mannen, T. Nagata, Y. Inatsu, K. Yura, O. Béjà, H. Kandori, I. Takeuchi, Exploration of natural red-shifted rhodopsins using a machine learning-based Bayesian experimental design, (n.d.). https://doi.org/10.1101/2020.04.21.052548.

[25] J.M. Kim, P.J. Booth, S.J. Allen, H.G. Khorana, Structure and function in bacteriorhodopsin: the role of the interhelical loops in the folding and stability of bacteriorhodopsin, J. Mol. Biol. 308 (2001) 409–422.

[26] Structure-Function Studies on Bacteriorhodopsin: IX. Substitutions of Tryptophan Residues Affect Protein-retinal Interactions in Bacteriorhodopsin, J. Biol. Chem. 264 (1989) 14197–

14201.

[27] S.L. Watwood, P.J.O. Miller, M. Johnson, P.T. Madsen, P.L. Tyack, Deep-diving foraging behaviour of sperm whales (Physeter macrocephalus), J. Anim. Ecol. 75 (2006) 814–825.

[28] B.S. Stewart, R.L. DeLONG, Double migrations of the northern elephant seal, Mirounga angustirostris, 1995.

[29] S. Mirceta, A.V. Signore, J.M. Burns, A.R. Cossins, K.L. Campbell, M. Berenbrink, Evolution of mammalian diving capacity traced by myoglobin net surface charge, Science. 340 (2013) 1234192.

[30] M. Berenbrink, 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 (2021) 110843.

[31] Y. Isogai, H. Imamura, S. Nakae, T. Sumi, K.-I. Takahashi, T. Shirai, Common and unique strategies of myoglobin evolution for deep-sea adaptation of diving mammals, iScience. 24 (2021) 102920.

[32] P. Dasmeh, A.W.R. Serohijos, K.P. Kepp, E.I. Shakhnovich, Positively selected sites in cetacean myoglobins contribute to protein stability, PLoS Comput. Biol. 9 (2013) e1002929.

[33] I.T. Weber, J. Agniswamy, HIV-1 Protease: Structural Perspectives on Drug Resistance, Viruses. 1 (2009) 1110–1136.

[34] M. Henes, K. Kosovrasti, G.J. Lockbaum, F. Leidner, G.S. Nachum, E.A. Nalivaika, D.N.A. Bolon, N. Kurt Yilmaz, C.A. Schiffer, T.W. Whitfield, Molecular Determinants of Epistasis in HIV-1 Protease: Elucidating the Interdependence of L89V and L90M Mutations in Resistance, Biochemistry. 58 (2019) 3711–3726.

[35] N.A. O'Leary, M.W. Wright, J.R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C.M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V.S. Joardar, V.K. Kodali, W. Li, D. Maglott, P.

Masterson, K.M. McGarvey, M.R. Murphy, K. O'Neill, S. Pujar, S.H. Rangwala, D. Rausch, L.D. Riddick, C. Schoch, A. Shkeda, S.S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R.E. Tully, A.R. Vatsan, C. Wallin, D. Webb, W. Wu, M.J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T.D. Murphy, K.D. Pruitt, Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation, Nucleic Acids Res. 44 (2016) D733–45.

[36] UniProt Consortium, UniProt: a worldwide hub of protein knowledge, Nucleic Acids Res. 47 (2019) D506–D515.

[37] J.L. Myers, A.D. Well, R.F. Lorch Jr, Research Design and Statistical Analysis: Third Edition, Routledge, 2013.

[38] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.

[39] T. Frickey, A. Lupas, CLANS: a Java application for visualizing protein families based on pairwise similarity, Bioinformatics. 20 (2004) 3702–3704.

[40] S.-Y. Rhee, M.J. Gonzales, R. Kantor, B.J. Betts, J. Ravela, R.W. Shafer, Human immunodeficiency virus reverse transcriptase and protease sequence database, Nucleic Acids Res. 31 (2003) 298–303.

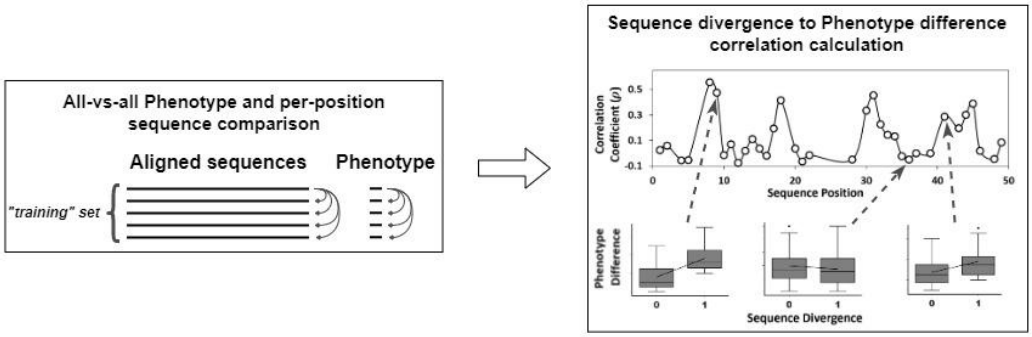
[41] A.M. Wensing, V. Calvez, F. Ceccherini-Silberstein, C. Charpentier, H.F. Günthard, R. Paredes, R.W. Shafer, D.D. Richman, 2019 update of the drug resistance mutations in HIV-1, Top. Antivir. Med. 27 (2019) 111–121.

[42] L.E. Jessen, I. Hoof, O. Lund, M. Nielsen, SigniSite: Identification of residue-level genotype-phenotype correlations in protein multiple sequence alignments, Nucleic Acids Res. 41 (2013) W286–91.

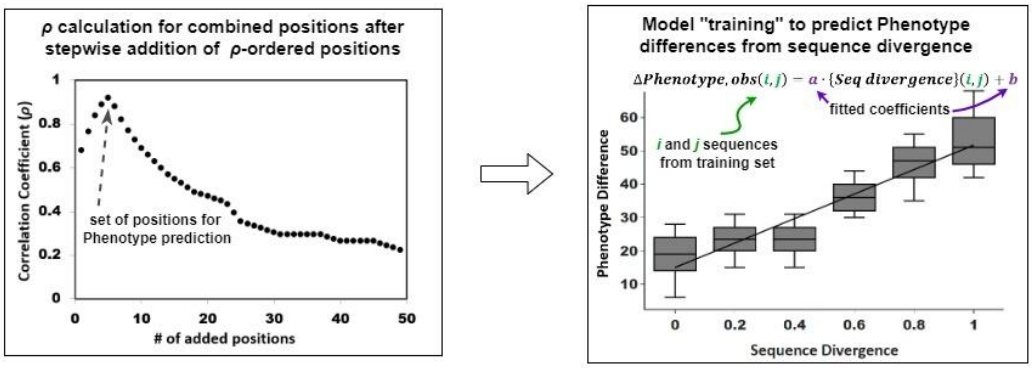
[43] P. Sonego, A. Kocsor, S. Pongor, ROC analysis: applications to the classification of biological sequences and 3D structures, Brief. Bioinform. 9 (2008) 198–209.

Figures

A. Identification of positions correlated with Phenotypes:



B. Position selection and model training to predict Phenotype differences:



C. Phenotype prediction for target sequences:

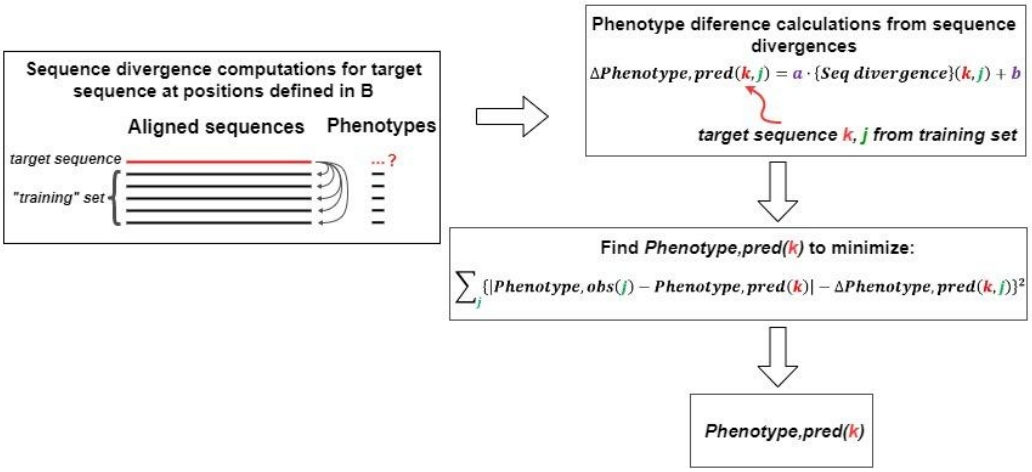


Figure 1. Detailed workflow of PhISCO algorithm. A. Identification of positions correlated with the phenotypes. Sequences from a protein family were aligned and pairwise similarity comparison was performed for the sequence and the observed features, for each position. A Spearman's rank-order correlation 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.

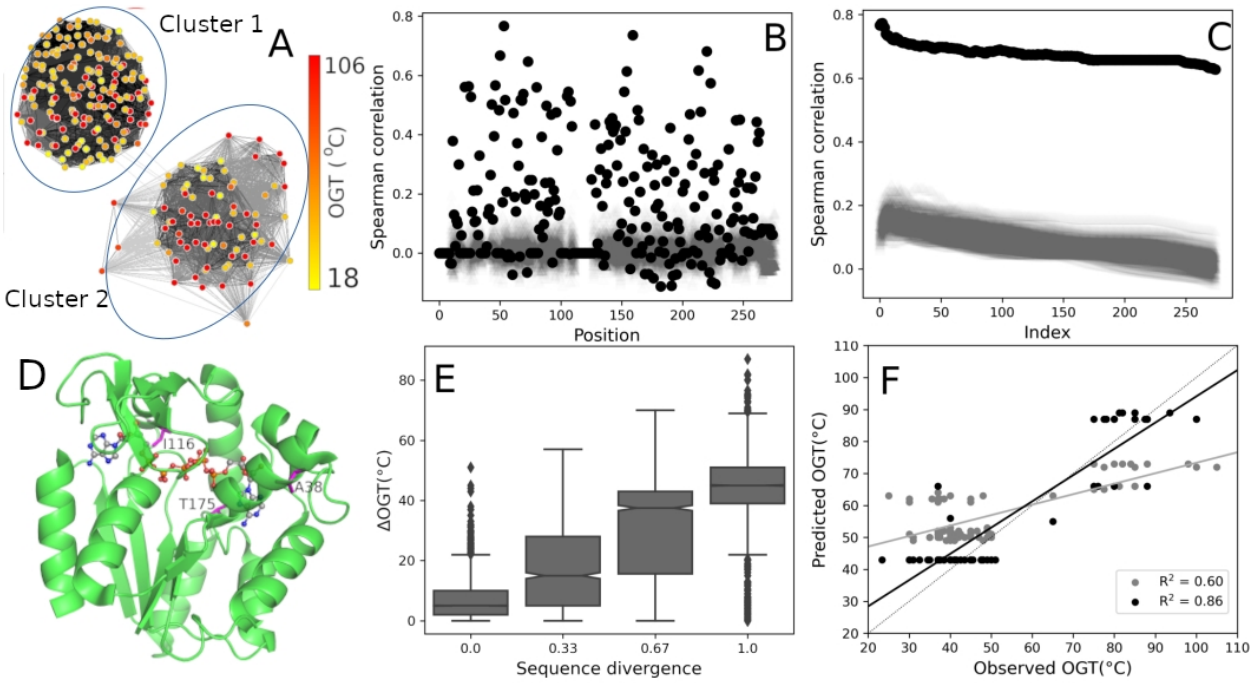


Figure 2. *Optimal Growth Temperatures associated with Archaeal Adenylate Kinase (AdK). A. Sequence clusterization according to pairwise similarity threshold. Cluster 1 has 163 sequences while cluster 2 has 80. B. Spearman rank-ordered correlation coefficients per position in the MSA (black filled circles) and control calculations after randomization of phenotype values (100 rounds of random assignments of phenotypes to sequences, grey points). C. Joint correlation coefficients of the AdK by addition of positions ranked in a descendent order (black filled circles) or by adding the same positions in the same order but after their phenotype values have been randomly permuted (100 rounds, gray points). D. Positions found in the joint correlation analysis mapped in a high resolution structure (PDB ID 1AKE). E. Distribution of the phenotype differences at different sequence divergence values. (R2 = 0.86 from a linear fit of ΔOGT vs sequence divergence). F. Prediction of OGT from sequences for the most correlated positions (black) and the entire sequence (gray).*

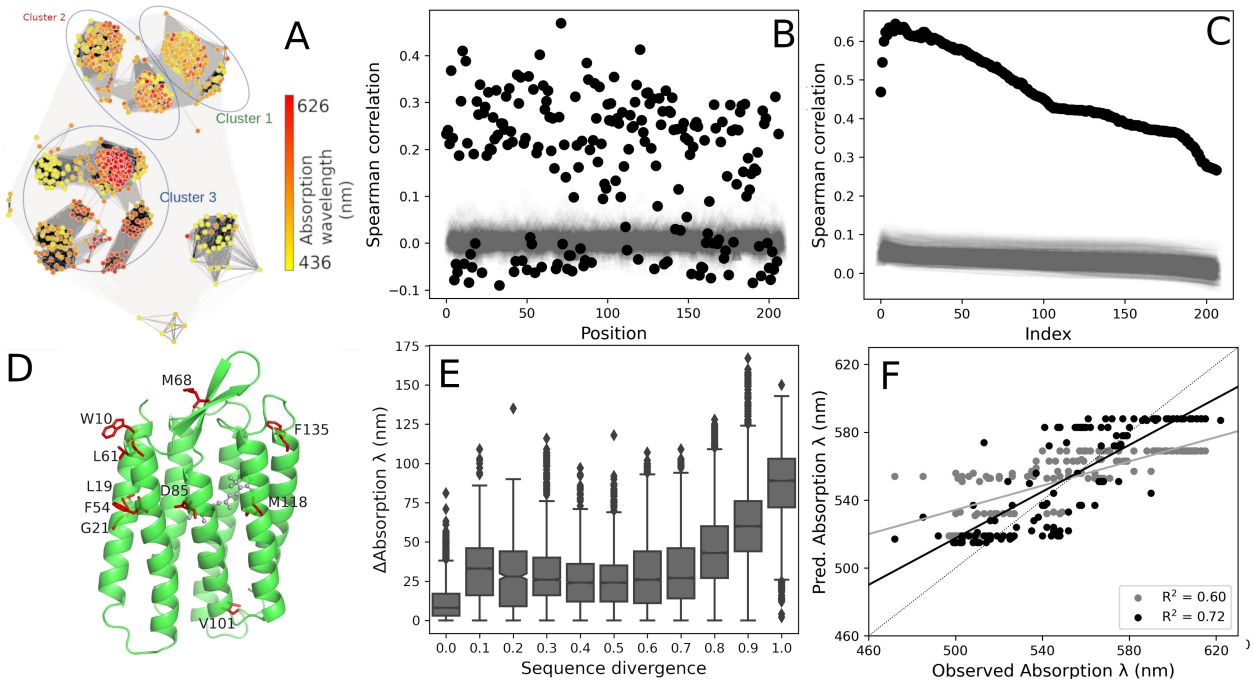


Figure 3. *Maximum absorption wavelength of variants of bacterial rhodopsins (Br). A. Sequence clusterization according to pairwise similarity threshold. Cluster 1 has 284 sequences, cluster 2 has 281 sequences and cluster 3 has 245 sequences. B. Spearman rank-ordered correlation coefficients per position in the MSA (black filled circles) and control calculations after randomization of phenotype values (100 rounds of random assignments of phenotypes to sequences, grey points). C. Joint correlation coefficients of the microbial rhodopsins by addition of positions ranked in a descendant order (black filled circles) or by adding the same positions in the same order but after their phenotype values have been randomly permuted (100 rounds, gray points). D. Positions found in the joint correlation analysis mapped in a high resolution structure (PDB ID 1C3W). E. Distribution of the phenotype differences at different sequence divergence values (R2 = 0.58 from a linear fit of ΔAbsorption wavelength vs sequence divergence). F. Prediction of absorption wavelength from sequences for the most correlated positions (black) and the entire sequence (gray).*

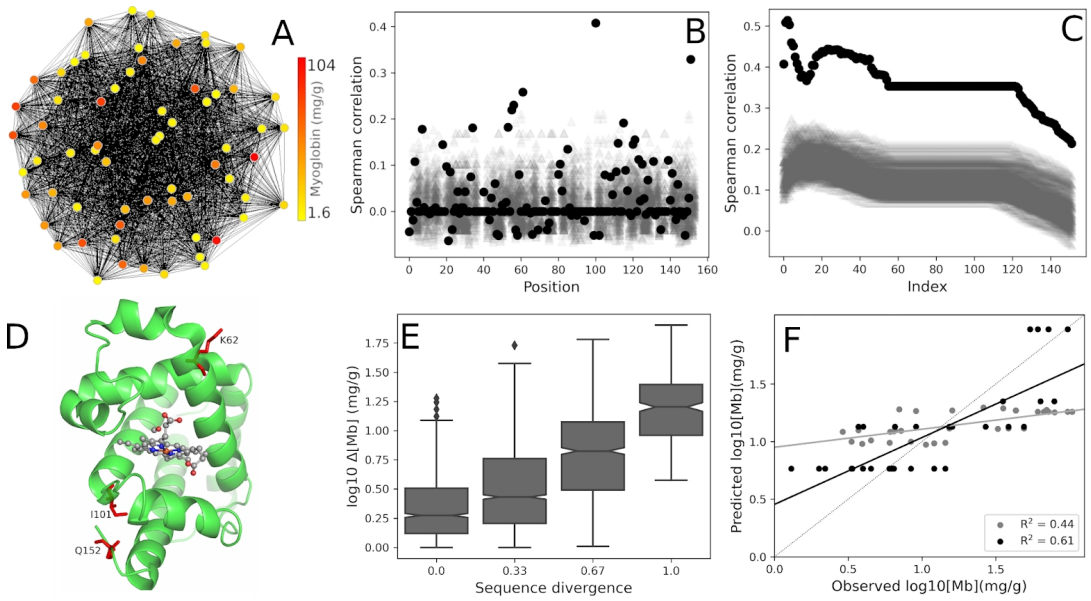


Figure 4. Myoglobin muscle concentrations in mammals (mmMb). A. Sequence clusterization according to pairwise similarity threshold. B. Spearman rank-ordered correlation coefficients per position in the MSA (black filled circles) and control calculations after randomization of phenotype values (100 rounds of random assignments of phenotypes to sequences, grey points). C. Joint correlation coefficients of the mmMb by addition of positions ranked in a descendent order (black filled circles) or by adding the same positions in the same order but after their phenotype values have been randomly permuted (100 rounds, gray points). D. Positions found in the joint correlation analysis mapped in a high resolution structure (PDB ID 1ABS). E. Distribution of the phenotype differences at different sequence divergence values (R2 = 0.55 from a linear fit of Δ log10[Mb] vs sequence divergence). F. Prediction of mmMb from sequences for the most correlated positions (black) and the entire sequence (gray).

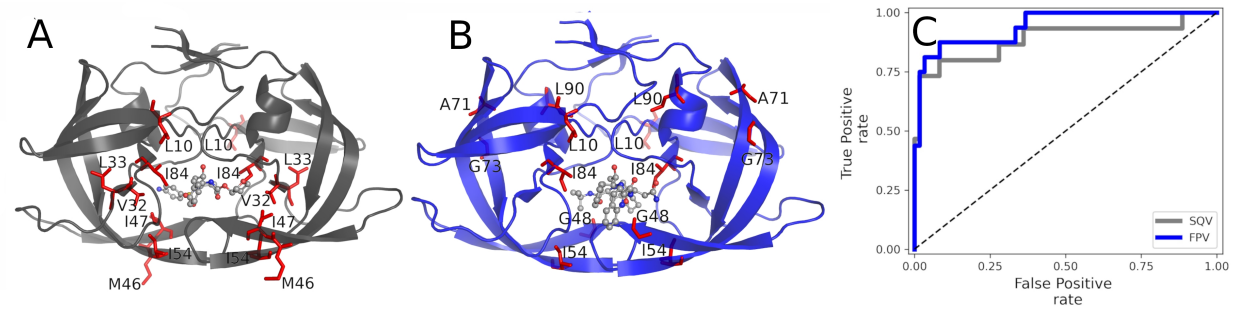


Figure 5. Predicted positions in HIV-PR sequence associated with inhibitor resistance. Most correlated positions (red sticks) for resistance to: A. FPV (mapped in the crystal structure for the complex with amprenavir, the active metabolite of FPV) or B. SQV (mapped in the crystal structure for the complex 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. The cut-off values obtained using the ROC analysis were 0.1700 for FPV and 0.1366 for SQV.

Tables

Table 1

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

Table 2

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

Table 3

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

Table 4

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

		FPV			SQV		
Position	WT res.	Exp. RDP[1]	Mut. Res.[2]	Pred. Status[3]	Exp. RDP[1]	Mut. Res.[2]	Pred. Status[3]
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	++	I, L	TP	+	I, L	TP
47	I	++	V, A	TP	-		TN
48	G	-		TN	++	V, M	TP
50	I	++	V	FN	-		TN
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
76	L	++	V	TP	-		TN
77	V	-		TN	+	I	FN
82	V	++	A, F, S, T	TP	++	A, F, S, T	TP
84	I	++	V, A, C	TP	++	V, A, C	TP
88	N	-		TN	++	S, D	FN
89	L	+	V	FN	-		TN
90	L	++	M	TP	++	M	TP

[1]Exp. RDP: experimentally observed RDP: '+': accessory mutation, '++': major mutation, '-': non relevant.
[2]Mutations most frequently found in patients' isolates.
[3]Prediction status: 'positives' are the positions whose PhISCO's rho value is above the cutoff from the ROC analysis. 'TP': true positive (correctly predicted positives), 'TN': true negative (correctly predicted negatives, i.e. experimentally confirmed RDPs whose rho values are below the cutoff), 'FP': false positive (incorrectly predicted positives), FN: false negative (incorrectly predicted negatives)