UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

TRABAJO FIN DE MÁSTER

Improving phylogenetic inference of protein evolution through a structurally constrained protein evolution model based on torsional normal modes

Máster Universitario en Bioinformática y Biología Computacional

Autor: CUDDIHY, Tadhg

Tutor: BASTOLLA, Ugo
Ponente: MARTÍNEZ MUÑOZ, Gonzalo
Departamento de Bioinformática (CSIC-UAM)

FECHA: febrero de 2021

ABSTRACT

Aims: The goal of this project is to develop and test a new method for predicting the change of a known protein structure produced by a given amino acid mutation. This prediction consists of two parts: (1) Modelling the perturbation induced by the mutation, which was done in the host laboratory and whose parameters the present work aims to determine; and (2) Computing the structural change produced by this perturbation as the linear response of the protein, as suggested by the structurally constrained model of protein evolution developed by Julián Echave. For this second part the Torsional Network Model (TNM) developed in the host laboratory was applied. The grand-goal is to predict the fitness change associated with the mutation and integrate it with the Stability Constrained models of Protein Evolution model previously developed in the host lab to improve phylogenetic inference of protein evolution. In this work, predicted and observed structural changes and RMSD are used to determine optimal parameters on a training set and assess the model on a test set.

Method: A set of protein pairs differing in one mutated amino acid were used for this project. Several filters were applied to ensure, as far as possible, that the observed structural change was mainly due to the mutation. For instance, the structures must contain the same ligands in order to reduce the possibility that the structural change arises from ligand binding. Proteins were divided into a training set and a test set. Three mutation parameters were optimized using Jarratt's method of successive parabolic interpolation to minimize the error of the predicted RMSD of the training set.

Results: The initial results suggested that overfitting was taking place and that it was necessary to regularize the optimization by imposing an additional condition on the direction of the structural change. The regularized parameters avoided unrealistic negative parameters, improved the prediction of the direction of the structural change and yielded an acceptable error on the predicted RMSD of the test set.

Conclusion: The optimized parameters produced acceptable results, although the regularization could be further improved with additional work.

Key Words: Stability Constrained models of Protein Evolution, Torsional Network Model

ACKNOWLEDGEMENTS

I would like to take this time to thank my tutor, Ugo, for helping me with this project and the other members of his lab. I would also like to thank my partner for his support and encouragement while I undertook this Masters.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ANM	Anisotropic Network Model
DOF	Degrees of Freedom
ENM	Elastic Network Model
LFNM	Low Frequency Normal Modes
MUT	Mutant protein
NMA	Normal Mode Analysis
PDB	Protein Data Bank
RMSD	Root Mean Square Deviation
RMSE	Root Mean Square Error
SCPE	Stability Constrained models of Protein Evolution
SPI	Successive Parabolic Interpolation
TNM	Torsional Network Model
WT	Wild-Type protein

1 INTRODUCTION

1.1 PROTEIN EVOLUTION

Proteins have evolved over time to perform specialized tasks throughout different systems. They achieve this through natural selection which acts on the random mutations occurring in the replicative process (insertions/deletions, copy number variations or single-point mutations) [1]. Natural selection also acts with varying strengths on different protein sites, leading to site variation of evolutionary rates. This was discussed by Echave, Wilke *et al.*, leading to two main conclusions: (1) Empirical substitution models assume that rates of evolution are the same at all positions, which is incorrect. (2) Stability, which is easier to predict than function, may rationalize the site dependence of evolutionary rates, which are strongly correlated with the number of native contacts that influence stability [2]. These conclusions form the basis for the development of Stability Constrained models of Protein Evolution (SCPE) as a method to study protein evolution.

1.2 STUDYING PROTEIN EVOLUTION WITH SCPE

The SCPE model models the fitness associated with a protein sequence as the probability that the protein is in its folded state [3], which is a function of its folding free energy (ΔG) (Eq. 1):

$$f = \frac{1}{1 + exp^{(\Delta G/T)}} \qquad \text{(Eq. 1)}$$

Fitness (f) is a sigmoidal function of stability (ΔG). For a large range of stability values, this means fitness is very close to either 0 or 1, i.e., the model is effectively neutral [4]. In the SCPE model developed in the host laboratory, the folding free energy (ΔG) is computed through a simple model of contact interactions that takes into account the known structure of the native state and the statistical ensemble of compact misfolded conformations [3].

This model maintains computational simplicity by assuming that the evolution of sites occurs independently (i.e., it does not consider the effect that a change to a site has on one another). However, the model still reflects the average constraints imposed by other protein sites, in a mean-field spirit, and predicts that the resulting substitution process is different on each protein site. The evolutionary rates resulting from this model are shown to be lowest at buried sites with many native contacts that contribute more to protein stability and are less tolerant to mutations [3]. Therefore, these sites experience stronger evolutionary pressure. In comparison with empirical substitution processes, the site-specific substitution processes predicted by the model increase the likelihood of the inferred evolutionary process and allow for the reconstruction of more realistic ancestral sequences.

This model presents, however, several limitations: While empirical data shows that site-specific substitution rates decrease with the number of contacts, sites with an intermediate number of contacts present the highest substitution rates in this model, which is incorrect. This model is also too tolerant to mutations, in particular those at sites with an intermediate

number of contacts [3]. This is a result of the model assuming that protein structure is perfectly conserved and does not consider the structural changes caused by mutations.

To study the differential tolerance to mutations of different protein sites, Echave modeled the effect single-point mutations have on the structural change of the protein by introducing a new Linear Forced Elastic Network Model (LFENM). This study found that protein structures evolve along the lowest normal modes and are predicted as the linear response of an Elastic Network Model (ENM) to the perturbation caused by the mutation, as observed when proteins undergo conformational changes induced by ligand binding [5]. It should be noted that the LFENM applies random perturbations at specific sites as opposed to perturbations caused by the single-point mutation. To fill this gap, the host laboratory developed a detailed mathematical model of the perturbations caused by each of the 210 possible single-point mutations.

Normal Mode Analysis (NMA) can also be used to analyze functional protein motions. NMA approximates harmonic potential between interacting atoms to describe their interactions. This was first proposed by Tirion [6] and, when used in conjunction with ENMs, has three important characteristics:

1. They assume that the structure in the PDB is a minimum of the free energy function.
2. Interactions are minimally frustrated, i.e., each interacting pair is at the distance where the interaction energy is minimal.
3. The harmonic approximation is used to approximate this energy with a quadratic function.

Over time, many variations of ENMs have been developed, such as the coarse grained version of Tirion's approach, the Anisotropic Network Models (ANM) and the Gaussian Network Model (GNM) [7]. This also includes the program developed for this project, which is an ENM in torsion angle space model known as the Torsional Network Model (TNM).

1.3 STUDYING PROTEIN EVOLUTION WITH TNM

The TNM program used in this project was first developed in the host lab by Mendez and Bastolla [8] and uses the protein backbone torsional angles as the degrees of freedom. This is similar to other torsion angle space models where NMA is performed using the torsional angle space [9], [10]. The TNM program performs NMA by allowing only the rotation around the alpha C-N (phi) and rotation around alpha C-C (psi) angles to vary while fixing all other degrees of freedom [11]. Moreover, interactions and kinetic energy are computed using all atoms in the protein, not only the alpha carbons.

TNMs have numerous advantages as detailed by Mendez and Bastolla [8]:

1. They better predict the displacement of atoms without additional computational costs. This applies to all atoms and not just alpha carbon.
2. They are faster at computation, compared to ANMs. This is due in part to the fact that they use fewer degrees of freedom. This results in diagonalizing a smaller matrix for computing the normal modes.

3. The covalent geometry is conserved up to the first order. This is achieved by applying small amplitude perturbations and has numerous advantages in constructing protein-like structures.

The TNM program developed by the host lab can be used to predict the conformational change arising from a given perturbation. To predict the structural effect of a mutation, it is still necessary to model how an amino acid mutation is mapped into a structural perturbation. The host laboratory developed this model taking into account three types of effects (on amino-acid size, on contact stability and on contact distance), extracting their parameters from the PDB. However, there is still a need to optimize the three mutation parameters that allow to combine these three types of effects. This is the goal of the present work.

1.4 HOW TO TACKLE THE PROBLEM OF OPTIMIZATION

1.4.1 Parabolic Optimization

The goal of this work is to find the set of three parameters that minimize the quadratic error between observed and predicted RMSD. There are numerous strategies that can be used for solving optimization problems. One such method utilized in this project is Jarratt's method of Successive Parabolic Optimization (SPI) [12]. Given three values, all in the same direction and with the other parameters fixed, SPI determines a candidate optimum value by fitting a parabola to the three points. The SPI method is then applied recursively in the three directions. There are numerous benefits to using Jarratt's method to optimize the predicted RMSD, including: (1) The method is superlinear to the order of $\alpha = 1.325$, meaning as iteration continues, the optimization becomes faster [12]. (2) This method requires three starting points, which for the optimization problem presented in this work, are easily generated.

1.4.2 Modelling the RMSD

It has been shown that sequence similar proteins can have dissimilar protein structures [13], and even when no mutation is present, proteins with the same sequence can predict a non-zero observed RMSD. Therefore, the Mean Square Deviation (MSD) was modeled between two proteins (the square of the RMSD) as the sum of the MSD in the absence of mutations and the MSD due to the mutation (Eq. 2). The MSD was summed because the mean is a linear operation whereas the square root is not linear. The TNM program predicts the RMSD due to the mutation, therefore Eq. 2 can be rewritten as Eq. 3, whose slope and constant can be computed by performing the linear regression.

$$RMSD_{observed}{}^2 = RMSD_{no\ mutation}{}^2 + RMSD_{mutation}{}^2 \qquad \text{(Eq. 2)}$$

$$RMSD_{observed}{}^2 = constant + \left(slope^2 * RMSD_{mutation}{}^2\right) \qquad \text{(Eq. 3)}$$

$RMSD_{mutation}$ is output by the TNM program, where it is computed based on the effect of the mutation on all the structural contacts formed by the mutated residue. Three types of effects are modelled: based on change of amino acid size, change of stability, and change of

optimal distance. These changes are obtained from the statistics of the PDB, and are scaled by three mutation parameters that are optimized in the present work.

1.5 MOTIVATION FOR THE PROJECT

As described above, mutations play an important role in evolution. These mutations can have a positive or negative effect on the proteins, thus on the organism as a whole. Studying these mutations and their effects on a protein can be very complex, hence the requirement for numerous mathematical models and methods to tackle this problem. Increasing the accuracy of the predicted mutant structure increases our understanding on how a given mutation changes the structure and function of a protein and from this can improve the ability to develop drug targets for the mutant.

The problem this project aims to solve is to reliably predict the RMSD between the wild-type and the mutant structure. This is achieved by optimizing the mutation parameters (size, stability, distance). Improving the predicted RMSD improves the similarity of the predicted mutant to the actual mutant.

Once the mutation parameters are optimized, the TNM program will allow improved predicted mutant atomic coordinates, which in turn is one step closer to predicting a more accurate mutant structure.

1.6 OBJECTIVE OF THE PROJECT

The objective of this project is to improve phylogenetic inference of protein evolution by integrating the structurally constrained model developed by Julian Echave [14] with the SCPE model previously developed in the host lab [15]. This can be achieved by reliably predicting the size of the structural change (predicted RMSD).

The work presented here aims at testing the mathematical model that represents how a single-point mutation perturbs the known native structure of the protein. This perturbation is applied to predict the mutated structure as the linear deformation of the TNM in response to the perturbation, and the comparison between the predicted and observed RMSD is adopted to assess the model.

The objective will be achieved by optimizing the mutation parameters of size, stability and distance in the mutation model. In order to determine if these parameters are sufficiently optimized, the predicted RMSD, calculated based on the mutation parameters, will be compared to the observed RMSD. Ideally the predicted RMSD will be similar to that of the observed RMSD while maintaining a reasonably low standard error.

2 MATERIALS AND METHODS

2.1 USING THE TNM PROGRAM

The TNM can operate in two different modes:

- Mode 1: Optimizing the mutation parameters for each pair independently, resulting in the observed and predicted RMSD being the same.
- Mode 2: Manually setting the mutation parameters used by the TNM program, resulting in a different observed and predicted RMSD for each pair. The performance of these parameters was tested and optimized as part of the main goal of this project.

The TNM program works by providing it with two or more protein files in PDB format and an input file containing the desired conditions to test (e.g., maximum number of mutations allowed between proteins, minimum RMSD allowed to continue calculations, etc.). In the absence of this input file (which is not required for the correct functioning of the program), the default parameters provided in the program are used.

The TNM models the perturbation associated to a mutation based on the contacts that the mutated amino acid forms in the native structure. Each contact contributes a force that is oriented along the direction of the contact, and whose signed value depends on the combination of the change of the size, stability and optimal distance between the wild-type pair and the mutated pair of interacting residues. Size, stability and distance are pre-computed from a representative set of the PDB and tabulated for the 210 possible amino acid pairs, so that the only free parameters are the coefficients of the three components of the force.

The program functions by pulling the protein structure information, following these main processes:

1. The sequences are aligned, using Needleman-Wunsch alignment method, and the number of mutations between each protein is calculated. If the number of mutations exceeds what is specified in the input file, the program exits.
2. The program checks that the reference atoms exist. The center mass is placed at the origin, and the cartesian axes are reoriented along the principal axes.
3. The topology is analyzed. At this stage, the location of the mutation/s between the two proteins and the number of bonds for building the molecule are recorded and tested to ensure they meet the minimum threshold. The observed RMSD of the conformational change is calculated based on the alpha carbon. If the RMSD does not meet the minimum specified in the input file, the program exits.
4. A number of ENM computations are performed. These include:
    a. The interactions between the atoms, calculated using the Gō model.
    b. The degrees of freedom, calculated on both the main and side chains. Depending on the input parameter specified, the main

degrees of freedom will be used later in the computation of protein dynamics in the harmonic approximation.
   c. TNM normal modes, computed using the degrees of freedom.
   d. Normal modes, selected based on eigenvalues and collectivity. Those not selected are discarded.
5. The conformational change is then analyzed. At this point, the predicted RMSD is calculated based on the input mutation parameters.

2.2 PROTEIN DATA INPUT FILE

The raw data protein input file (bc-95.out), containing all proteins with 95% sequence identity, was downloaded from the PDB supplemental archive:

https://www.rcsb.org/pages/download/ftp

This file is composed of proteins with chains longer than 20 amino acids. Each line in the file represents a new cluster. The PDB generates this file by running blastclust with the following parameters:

-c param_file.txt[-e 0.01] -p T -b T -S 90

At the time of this project, there were 60,590 clusters and a total of 472,717 proteins.

2.3 GENERATING THE TRAIN AND TEST FILES

To ensure the structural changes being examined were due to single-point mutations and not resulting from other factors (such as structural changes resulting from different ligands), only wild-type and mutant pairs from the same family, containing the same ligands and pairs separated by one mutation, were examined (Figure 1). Additionally, proteins whose structures are determined by nuclear magnetic resonance (NMR) were discarded. Structures determined by NMR are less compact, which results in the TNM predicting larger flexibility compared to structures determined by X-ray crystallography.

The list of pairs was processed through the TNM program. The TNM computes the ratio between the observed conformational change at a position and the structural fluctuation predicted at the same position due to thermal motion. This computation is used to select mutants for which the conformational change is likely to be due to the mutation rather than changed experimental conditions. This was achieved by imposing the following criteria:

- Observed RMSD > 0.4
- Mutation: max. ratio ≥ 1.2
- No-mutation: max. conformational change ≤ 1
- Average ratio Mutation > Average ratio No-mutation

Pairs that failed to meet these criteria were discarded.

In order to ensure that the observed structural change resulted from the single-point mutation, the conformational change observed needed to be significant. A threshold of 1 was

set for the change not due to the mutation and a threshold of 1.2 was set for the change due to the mutation. If the non-mutation conformational change is larger than the mutation conformational change, this may be caused by factors other than single-point mutation.

The final protein pair list was split roughly 80/20 into a train and test file, ensuring all pairs within the same cluster of sequence-related proteins were kept in the same file.

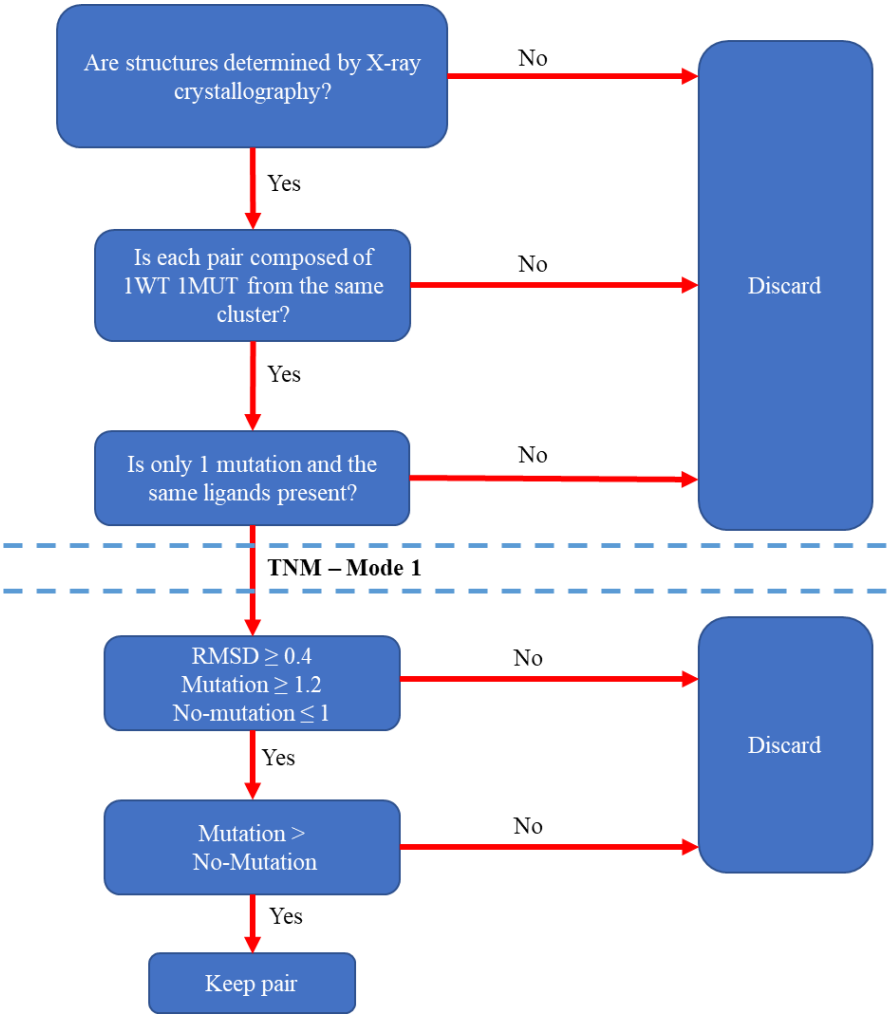

Figure 1 – Pair Creation Requirements. All created pairs were required to be from the same cluster and have the same ligands. Only single-point mutations were analyzed in this project to ensure the changes observed could be associated with that specific mutation.

2.4 Optimization of Mutation Parameters

To generate the starting values for the optimization, the training set was passed through the TNM program in Mode 1, which optimizes the three mutation parameters (size, stability and distance) per each protein pair individually by maximizing the cosine between the observed and predicted direction of the conformational change and equating observed and predicted RMSD. The mean values of the optimized parameters for all protein pairs were chosen as starting parameters, and Standard Error of the Mean (SEM) (Eq. 4) was used to generate the three initial values in the parabolic optimization method. Subsequently, the optimization of

the parameters was disabled and the TNM was run with the same mutation parameters for all proteins. After each run, the linear regression was performed using the python package sklearn [16] over the squared observed and predicted RMSD. The slope of the regression line, returned from sklearn, was used to scale the parameters and SEM (Eq. 5).

$$SEM = \frac{\sigma_{parameter}}{\sqrt{n}} \quad \text{(Eq. 4)}$$

$$Scaled\ parameter = parameter * \sqrt{slope} \quad \text{(Eq. 5)}$$

One parameter at a time was optimized using SPI while the other two parameters remained fixed. The optimization starts by using the scaled mean and scaled SEM (Figure 2). Since the mean quadratic error of the linear fit is proportional to 1-r^2, where r is the correlation coefficient of the fit, the correlation coefficient was maximized as the function of the parameters.

In the SPI method, a parabola is fitted from three data points consisting of three values of the parameter to optimize ([param - SEM, param, param + SEM]) and the associated score (the correlation coefficient computed from the TNM results).

In order to maximize the scoring function $f(xi)$, the coordinates were fitted to a parabola using the equations below, where x is the parameter value.

$$(i = 1, 2, 3) \qquad f(xi) = a * xi^2 + b * xi + c \quad \text{(Eq. 6)}$$

$$2ax_4 + b = 0 \quad \text{(Eq. 7)}$$

The new estimated parameter, $\{x_4\}$ was computed as:

$$a = \frac{x_3y_2 - x_3y_1 - x_2y_3 + x_2y_1 + x_1y_3 - x_1y_2}{x_3{x_2}^2 - x_3{x_1}^2 - x_2{x_3}^2 + x_2{x_1}^2 + x_1{x_3}^2 - x_1{x_2}^2} \quad \text{(Eq. 8)}$$

$$b = \frac{y_1{x_2}^2 - y_1{x_3}^2 - y_2{x_1}^2 + y_2{x_3}^2 + y_3{x_1}^2 - y_3{x_2}^2}{x_1{x_2}^2 - x_1{x_3}^2 - x_2{x_1}^2 + x_2{x_3}^2 + x_3{x_1}^2 - x_3{x_2}^2} \quad \text{(Eq. 9)}$$

$$x4 = -\frac{b}{2a} \quad if\ a < 0,$$
$$x4 = Argmax\big(f(x1), f(x2), f(x3)\big) \quad if\ a > 0 \quad \text{(Eq. 10)}$$

The optimization is stopped when $f(xi)$ no longer increases for three successive iterations in the three directions corresponding to the three parameter types (Figure 2). The parameters prior to the three successive iterations are considered the optimized parameters.

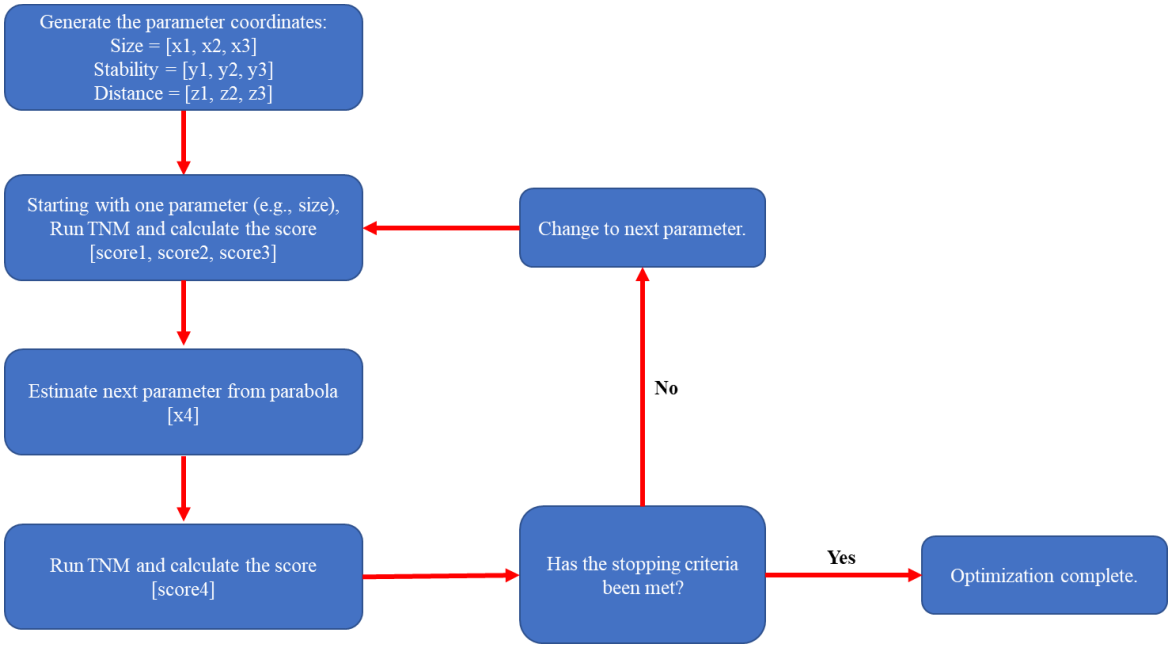

Figure 2 – Method for Optimization of Parameters. Optimization occurs one parameter at a time. In the event there are three successive iterations where the score remains the same or decreases without showing any sign of recovery, optimization will stop returning the optimized parameters. Otherwise, the next parameter is optimized until this condition is achieved.

The parameters were scaled (Eq. 5) prior to testing on the test set. Using the test set, the scaled parameters were then run through the TNM in Mode 2. The un-normalized Root Mean Square Error (RMSE) on the RMSD, expressed in Angstrom, was calculated on the test set and on the training set from the following equations:

$$RMSD_{predicted} = \sqrt{constant + \left(slope^2 * {RMSD_{mutation}}^2\right)} \quad \text{(Eq. 11)}$$

$$RMSE = \sqrt{\frac{\sum\left(RMSD_{observed} - RMSD_{predicted}\right)^2}{n}} \quad \text{(Eq. 12)}$$

3 RESULTS

3.1 INITIAL STARTING VALUES

Table 1 shows the initial mean and scaled values calculated with the training set. These values were used as the starting parameters for optimization.

Parameter Type	Mutation Parameter			SEM		
	Size	Stability	Distance	Size	Stability	Distance
Mean	100.86	250.77	144.09	20.53	56.19	62.79
Scaled	11.65	28.96	16.64	2.37	6.49	7.25

Table 1 – Starting Values. The TNM was run in Mode 1 with the mean parameters. The slope of the regression line (0.1155) was used to generate the scaled parameters and scaled SEM. The scaled parameters are the starting parameters for SPI.

3.2 OPTIMIZING THE MUTATION PARAMETERS

3.2.1 Optimizing for Correlation Coefficient of the RMSD

Table 2 details the initial optimization results. The optimized parameters are indicated in Iteration 20 ([-4.46, 95.71, 2.62]) with a correlation of 0.775, after which the correlation decreases. Using the slope of the regression line, the scaled parameters were run with the training set (Table 3). A large positive correlation was shown with an RMSE of 0.158.

Nevertheless, this set of optimized parameters showed some problems. First of all, the optimal parameter related with the amino acid size is negative, which is contrary to physical intuitions since it means that, if an amino acid is mutated with a larger one, amino acids in contact are pulled towards it instead of being pushed away. The TNM program computes the cosine between the predicted and observed conformational change. Consistent with the negative value of the size parameter, it was observed that the cosine was on the average negative (see the last column of Table 2), i.e., the optimal parameters predict conformational changes of approximately correct magnitude (RMSD) but with qualitatively wrong directions. Since the same RMSD can be obtained with two conformational changes in a direction and the opposite one, this result suggests that the good fit of the RMSD was obtained not because of a reasonable physical model but because of overfitting.

Overfitting is very common in ill-defined optimization problems in which there are correlated explanatory variables such as the size, stability and distance that are used to predict the RMSD. Very frequently, overfitting yields optimized parameters that may be contrary to physical intuition, as the obtained negative size parameter. To address overfitting issues, it is necessary to regularize the optimization problem imposing additional conditions that reduce the optimal score that can be achieved and reduce the risk of overfitting. A very popular regularization scheme is Tikhonov regularization or ridge regression. However, regularization requires fixing an optional parameter that weights the regularizing condition,

which is often arbitrary (even if some objective recipes have been proposed, also in the host laboratory) and it can be cumbersome when the computational burden is high as in the present case. Therefore, it was decided to choose the simplest regularization, i.e. reducing the number of parameters.

Optimize Correlation - Train Set					
Iteration	Mutation Parameter			Correlation	Cosine
	Size	Stability	Distance		
0	14.86	28.96	16.64	0.619	-0.0149
1	14.86	19.00	16.64	0.692	-0.0066
2	14.86	19.00	18.12	0.687	-0.0089
3	21.76	19.00	18.12	0.708	0.0007
4	21.76	9.97	18.12	0.723	-0.0013
5	21.76	9.97	16.34	0.725	0.0008
6	20.25	9.97	16.34	0.728	-0.0006
7	20.25	1.71	16.34	0.738	-0.0008
8	20.25	1.71	13.38	0.738	0.0033
9	9.85	1.71	13.38	0.730	-0.0116
10	9.85	100.00	13.38	0.694	-0.0553
11	9.85	100.00	5.80	0.734	-0.0395
12	-2.27	100	5.8	0.770	-0.0470
13	-2.27	100.43	5.8	0.769	-0.0470
14	-2.27	100.43	4.06	0.770	-0.0449
15	-3.73	100.43	4.06	0.773	-0.0435
16	-3.73	99.69	4.06	0.774	-0.0435
17	-3.73	99.69	4.14	0.774	-0.0430
18	-4.46	99.69	4.14	0.772	-0.0429
19	-4.46	95.71	4.14	0.772	-0.0430
20	-4.46	95.71	2.62	0.775	-0.0411
21	-5.5	95.71	2.62	0.774	-0.0403
22	-5.5	97.04	2.62	0.773	-0.0403
23	-5.5	97.04	3.39	0.772	-0.0412

Table 2 – Results from SPI (Optimize Correlation). For each iteration, only one parameter was optimized at a time, in the sequence of size, stability, distance. Optimization stops when the correlation no longer increases for three successive iterations. The optimized parameters are Iteration 20. The cosine column contains the average cosine for each iteration. The cosine results are presented to four significant figures in order to assess small changes between each iteration.

Optimize correlation - Train Set		
Scaled Optimized Parameters ***[-4.32, 92.77, 2.54]***	**Observed**	**Model**
RMSD	0.801	0.823
SEM	0.033	0.019
Correlation	0.774	
RMSE	0.158	

Table 3 – Scaled Optimized Parameters (Optimize Correlation). Results from the training set when the scaled optimized parameters are used.

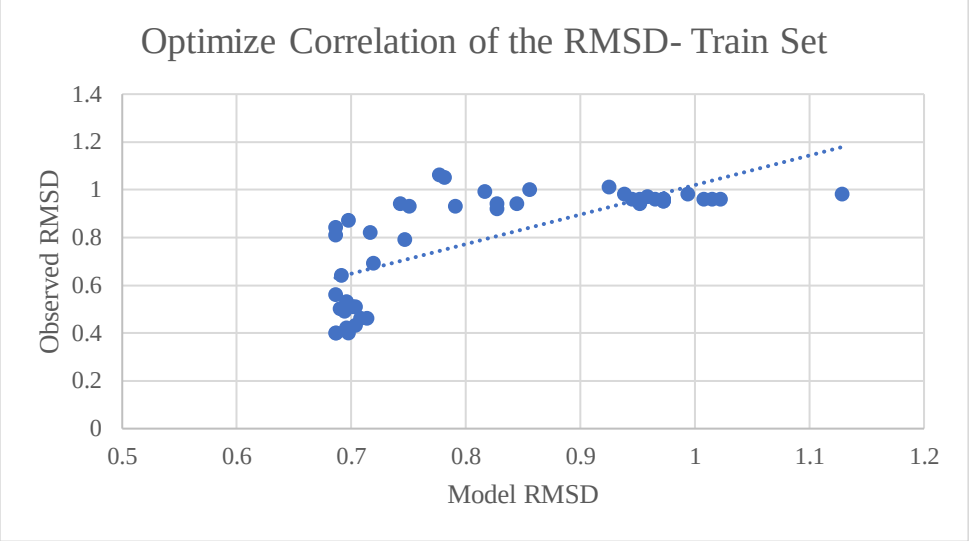

Figure 3 – Observed RMSD and the Model RMSD (Optimize Correlation)

3.2.2 Optimizing for Correlation Coefficient of the RMSD (2 Parameters)

Each one of the three mutation parameters was eliminated and the best results were obtained when eliminating the stability parameter. Table 4 shows the results for the two parameter optimization. In this run, the optimized parameters are Iteration 1, after which the correlation coefficient stops increasing. All iterations in this run have a negative cosine. The scaled parameters returned an RMSE of 0.165 (Table 5), slightly higher than the three parameter run. Figure 4 shows the fit, which compared to the previous three parameter run (Figure 3), is very similar. Analyzing the results, the optimal iteration had a negative cosine, suggesting the model is not realistic.

2 Parameter - Train Set					
Iteration	Mutation Parameter			Correlation	Cosine
	Size	Stability	Distance		
0	131.66	0	144.09	0.74	-0.007
1	131.66	0	112.70	0.741	-0.002
2	131.66	0	112.70	0.741	-0.002
3	131.66	0	112.70	0.741	-0.002

Table 4 – 2 Parameter Optimization

2 Parameter Optimization – Train Set		
Scaled Optimized Parameters ***[23.3, 0, 25.5]***	**Observed**	**Model**
RMSD	0.801	0.821
SEM	0.033	0.018
Correlation	0.738	
RMSE	0.165	

Table 5 – 2 Parameter Optimization. Results of the scaled parameters run with the training set.

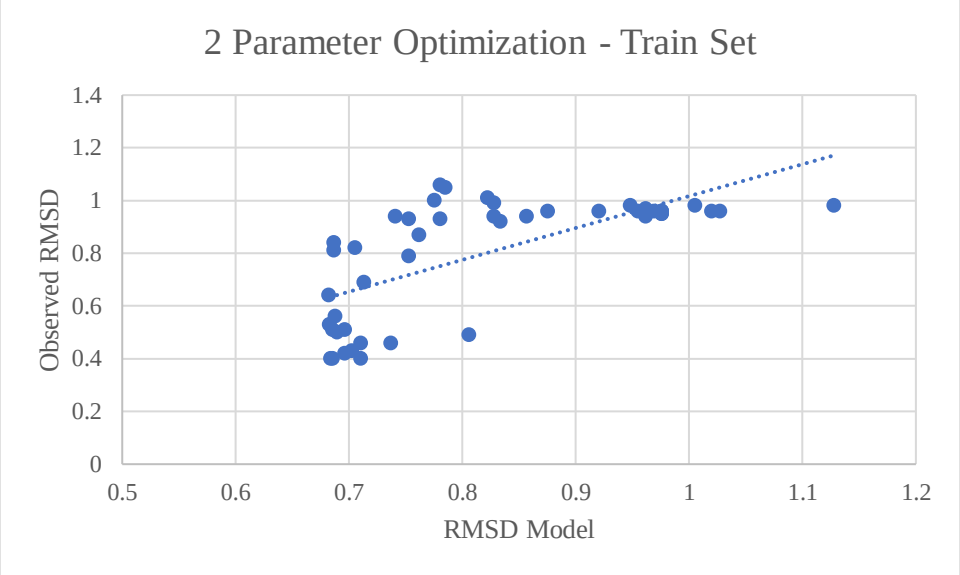

Figure 4 – 2 Parameter Optimization

3.2.3 Optimizing for a Positive Cosine

Since having a large cosine between the observed and predicted conformational change not only provides a more realistic physical model of the mutation but it may also achieve a good fit between the observed and predicted scale of the conformational change (i.e. the RMSD), it was decided to maximize the average cosine over the training set (Table 6). It is noteworthy that the average cosine becomes rapidly positive and keeps increasing, and all the parameters stay positive, which indicates that the resulting models are more realistic. Nevertheless, the correlation between observed and predicted RMSD reaches the maximum value at Iteration 4. Since our goal is to have a good fit between observed and predicted RMSD, Iteration 4 was chosen, where the correlation was reasonably high (0.65) and the average cosine was positive.

Optimize Cosine – Train Set					
Iteration	**Mutation Parameter**			**Correlation**	**Cosine**
	Size	**Stability**	**Distance**		
0	14.02	28.96	16.64	0.598	-0.019
1	14.02	22.47	16.64	0.661	-0.012
2	14.02	22.47	9.39	0.602	0.005
3	16.39	22.47	9.39	0.629	0.010
4	16.39	21.79	9.39	0.647	0.010
5	16.39	21.79	2.14	0.473	0.032
6	14.02	21.79	2.14	0.383	0.036
7	14.02	28.28	2.14	0.186	0.043

Table 6 - Results from SPI (Optimize Cosine). For each iteration, only one parameter was optimized at a time, in the order of size, stability, distance. Iteration 4 was considered the finishing point due to having a positive cosine and followed by three successive iterations of decreasing correlation. The cosine column contains the average cosine for each iteration. The cosine results are presented to four significant figures in order to assess small changes between each iteration.

All optimized parameters are positive in this run. The scaled optimized parameters from Iteration 4 (Table **7**) yielded a RMSE of 0.18 Angstrom, an accuracy similar to that of the previous runs. However, the correlation (0.635) was lower.

Optimize Cosine – Train Set		
Scaled Optimized Parameters ***[42.87, 57.00, 24.56]***	**Observed**	**Model**
RMSD	0.801	0.824
SEM	0.033	0.015
Correlation	0.635	
RMSE	0.184	

Table 7 – Training Set Results Using Scaled Optimized Parameters (Optimize Cosine). Table shows results using the scaled optimized parameters.

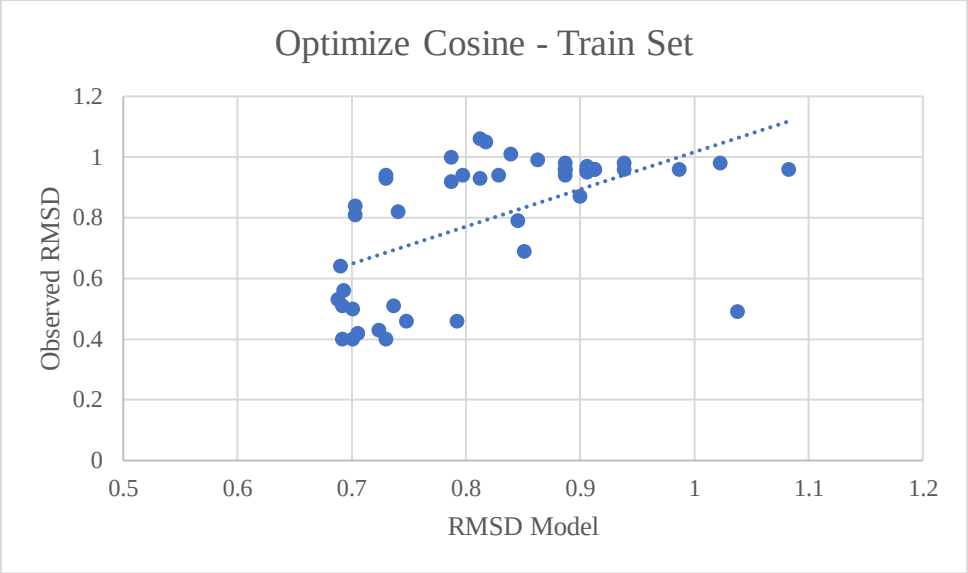

Figure 5 – Observed RMSD and the RMSD Model (Optimize Cosine)

3.2.4 Optimizing for Combination of Correlation and Cosine

The decrease of the correlation coefficient for higher average cosine observed in Table 6 suggests that optimizing the cosine is not sufficient to obtain a good fit between the observed and predicted RMSD. Therefore, it was decided to optimize a combination of the correlation coefficient plus the average cosine, adopting the following scoring function:

$$\text{Score} = \text{Correlation coefficient} + \lambda * \text{Average cosine} \qquad (\text{Eq. } 13)$$

The correlation coefficient is the quantity to be optimized, and the term containing λ can be interpreted as a regularization. Optimizing the correlation corresponds to $\lambda = 0$ (no regularization), whereas optimizing only the cosine corresponds to the $\lambda \to \infty$ limit. As can be seen from the first step (Iteration 0 to Iteration 1) in Table 6, the correlation coefficient increased roughly 10 times more than the average cosine, suggesting that $\lambda = 10$ could be a good estimation for the regularization parameter.

Table 8 shows the results for optimizing Eq. 13 with $\lambda = 10$ (combination column). However, the observed behaviour was not different from the run where only the cosine was optimized, since the correlation coefficient reached its maximum at Iteration 4 and decreased in the following iterations. Moreover, after a few iterations some parameters became negative, suggesting that the resulting models are not very realistic. Iteration 7 was considered the best combination of the correlation coefficient (0.707) and the average cosine. The optimized parameters [22.92, 2.01, 2.14] were scaled using the slope of the regression line and run on the training set. A strong positive correlation of 0.705 and an RMSE of 0.173 were observed (Table 9).

The scaled parameters [35.48, 3.11, 3.31] from Iteration 7 were considered the optimized parameters and tested on the test set.

Optimize Combination – Train Set						
Iteration	**Mutation Parameter**			**Correlation**	**Cosine**	**Combination**
	Size	**Stability**	**Distance**			
0	16.26	28.96	16.64	0.621	-0.010	0.525
1	16.26	13.47	16.64	0.711	-0.004	0.668
2	16.26	13.47	9.39	0.701	0.008	0.778
3	20.55	13.47	9.39	0.702	0.011	0.808
4	20.55	6.98	9.39	0.724	0.010	0.824
5	20.55	6.98	2.14	0.687	0.029	0.981
6	22.92	6.98	2.14	0.689	0.030	0.987
7	22.92	2.01	2.14	0.707	0.028	0.990
8	22.92	2.01	-131.40	0.582	0.060	1.178
9	25.29	2.01	-131.40	0.579	0.061	1.190
10	25.29	-4.48	-131.40	0.578	0.063	1.206

Table 8 – Results from SPI (Optimize Combination)

Optimize Combination – Train Set		
Scaled Optimized Parameters ***[35.48, 3.11, 3.31]***	**Observed**	**Model**
RMSD	0.801	0.823
SEM	0.033	0.017
Correlation	0.705	
MSE	0.173	

Table 9 - Training Set Results Using Scaled Optimized Parameters (Optimize Combination)

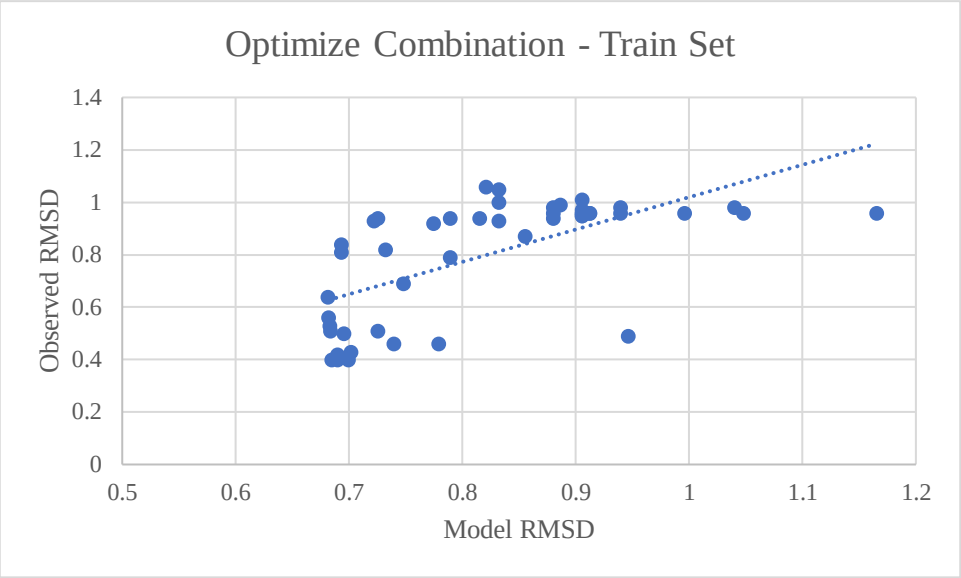

Figure 6 – Observed RMSD and the RMSD Model (Optimize Combination)

3.3 TESTING THE OPTIMIZED PARAMETERS

The scaled optimized parameters were tested on the test set (Table 10). The correlation on this set was -0.220 and the RMSE was 0.239 Angstrom. Although the correlation was negative, upon further investigation, the pairs in test set showed a very limited range of RMSD. This would suggest that it is very difficult to fit this small variation of the order of less than 0.1 Angstrom. The low RMSE indicates that the accuracy of the prediction may be acceptable for our purposes.

Scale Optimize Parameters – Test Set		
Scaled Optimized Parameters ***[35.48, 3.11, 3.31]***	**Observed**	**Model**
RMSD	0.464	0.699
SEM	0.013	0.001
Correlation	-0.220	
RMSE	0.239	

Table 10 – Test Set Results. The model RMSD was calculated using the slope and intercept of the optimized iteration from the training set (Iteration 7).

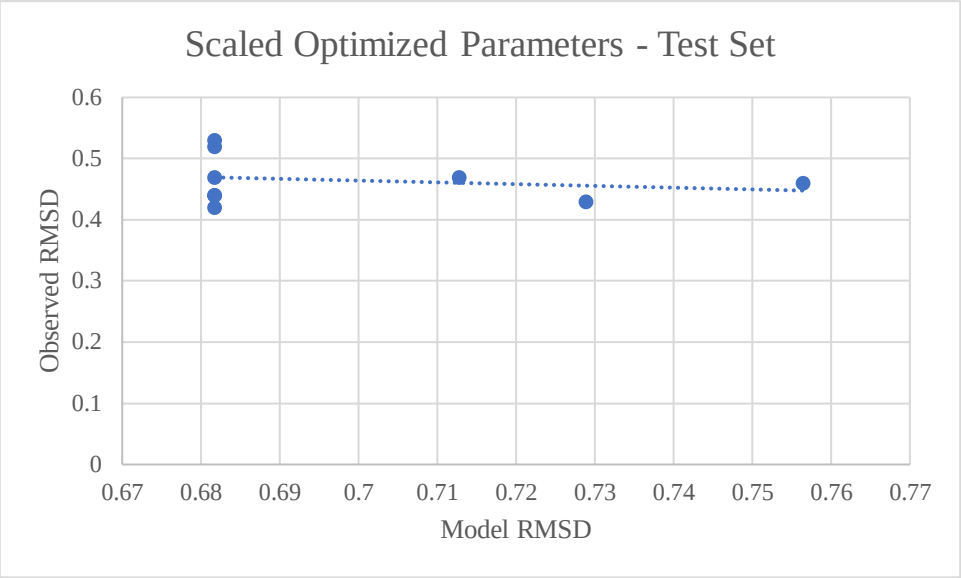

Figure 7 – Test Results Using Scaled Optimized Parameters

4 DISCUSSION

Maximizing the correlation coefficient of the RMSD (lambda = 0) yielded good results on the training set. However, it resulted in a negative optimized parameter (size) and a negative average cosine for the optimized iteration. Optimizing for correlation coefficient of the RMSD was problematic, since similar results can be obtained also with unphysical moves (negative cosines, negative parameters). This optimization resulted in a negative size parameter, which may have been generated to compensate for overestimation of the effect of the other parameters, suggesting that the model is overfitted to the data, as it is frequent when there are correlated explanatory variables. Removing the stability parameter and rerunning the optimization resulted in positive parameters but returned a negative average cosine and very low correlation.

To try and overcome overfitting in the model, the cosine was optimized in order to achieve predicted conformational changes that are directed in the same direction as the observed ones. The optimized model produced more physical models of mutations, however, compared to previous runs, the correlation of the RMSD was lower.

The two scoring functions (Eq. 13) were then combined. In this equation, the correlation is the quantity that we aim to fit and the average cosine is a regularization condition that imposes more realistic conformational changes. As an initial estimation, lambda=10 was chosen as the regularization parameter. The optimization returned positive average cosines and better correlation of the RMSD compared to just optimizing the cosine alone. However, the cosine and the correlation soon started to increase in opposite directions, which resulting in choosing suboptimal parameters where the correlation was sufficiently good. These parameters returned a strong correlation on the training set but a poor one on the test set. However, the range of RMSD in the test set was very limited, making it difficult to obtain good correlations. The RMSE was 0.24 Angstrom, which is acceptably good for our purposes.

There is potential to further refine the lambda, however, the optimized parameters returned an RMSE on the training (0.172) and test (0.239) set that was considered low. This, combined with the fact that the optimized parameters had a positive average cosine, would suggest that the model can be accepted and the mutation parameters of size, stability and distance show sufficiently good performances.

5 CONCLUSIONS AND FUTURE WORK

Optimization of the mutation parameters proved to be a more complex problem than originally thought because of problems of overfitting due to correlated explanatory variables and unphysical moves. Although optimizing for a positive cosine yielded positive parameters, they showed lower correlation on the training set compared to optimizing for the correlation of the coefficient. The combination of the two strategies resulted in a model that was more realistic and had a low RMSE. The chosen parameters are size = 35.48, stability = 3.11, distance = 3.31.

The parameters obtained as a result of this project can be used as the initial starting values to test the fitness function (Eq. 1), with a modified version of the TNM program, to compute the effect of all possible mutations as opposed to just single-point mutations tested in this work.

6 REFERENCES

[1] J. Zhang and J.-R. Yang, "Determinants of the rate of protein sequence evolution," *Nat. Rev. Genet.*, vol. 16, no. 7, pp. 409–420, Jul. 2015, doi: 10.1038/nrg3950.

[2] J. Echave, S. J. Spielman, and C. O. Wilke, "Causes of evolutionary rate variation among protein sites," *Nat. Rev. Genet.*, vol. 17, no. 2, pp. 109–121, Feb. 2016, doi: 10.1038/nrg.2015.18.

[3] M. J. Jimenez, M. Arenas, and U. Bastolla, "Substitution Rates Predicted by Stability-Constrained Models of Protein Evolution Are Not Consistent with Empirical Data," *Mol. Biol. Evol.*, vol. 35, no. 3, pp. 743–755, Mar. 2018, doi: 10.1093/molbev/msx327.

[4] A. W. R. Serohijos and E. I. Shakhnovich, "Merging molecular mechanism and evolution: theory and computation at the interface of biophysics and evolutionary population genetics," *Curr. Opin. Struct. Biol.*, vol. 26, pp. 84–91, Jun. 2014, doi: 10.1016/j.sbi.2014.05.005.

[5] J. Echave, "Evolutionary divergence of protein structure: The linearly forced elastic network model," *Chem. Phys. Lett.*, vol. 457, no. 4, pp. 413–416, May 2008, doi: 10.1016/j.cplett.2008.04.042.

[6] M. M. Tirion, "Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis," *Phys. Rev. Lett.*, vol. 77, no. 9, pp. 1905–1908, Aug. 1996, doi: 10.1103/PhysRevLett.77.1905.

[7] "Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model | Elsevier Enhanced Reader." https://reader.elsevier.com/reader/sd/pii/S000634950176033X?token=64F954B07C96F45CD6312F0974E20A1B356B374D04BF9E67F0456D76E7A97F2F534DB125E169608F6B84E1C97D390D49 (accessed Aug. 30, 2020).

[8] R. Mendez and U. Bastolla, "Torsional Network Model: Normal Modes in Torsion Angle Space Better Correlate with Conformation Changes in Proteins," *Phys. Rev. Lett.*, vol. 104, no. 22, p. 228103, Jun. 2010, doi: 10.1103/PhysRevLett.104.228103.

[9] J. K. Bray, D. R. Weiss, and M. Levitt, "Optimized Torsion-Angle Normal Modes Reproduce Conformational Changes More Accurately Than Cartesian Modes," *Biophys. J.*, vol. 101, no. 12, pp. 2966–2969, Dec. 2011, doi: 10.1016/j.bpj.2011.10.054.

[10] J. R. Lopéz-Blanco, J. I. Garzón, and P. Chacón, "iMod: multipurpose normal mode analysis in internal coordinates," *Bioinformatics*, vol. 27, no. 20, pp. 2843–2850, Oct. 2011, doi: 10.1093/bioinformatics/btr497.

[11] H. G. Dos Santos, J. Klett, R. Méndez, and U. Bastolla, "Characterizing conformation changes in proteins through the torsional elastic response," *Biochim. Biophys. Acta BBA - Proteins Proteomics*, vol. 1834, no. 5, pp. 836–846, May 2013, doi: 10.1016/j.bbapap.2013.02.010.

[12] P. Jarratt, "An iterative method for locating turning points," *Comput. J.*, vol. 10, no. 1, pp. 82–84, Jan. 1967, doi: 10.1093/comjnl/10.1.82.

[13] M. Kosloff and R. Kolodny, "Sequence-similar, structure-dissimilar protein pairs in the PDB," *Proteins*, vol. 71, no. 2, pp. 891–902, May 2008, doi: 10.1002/prot.21770.

[14] J. Echave and F. M. Fernández, "A perturbative view of protein structural variation," *Proteins Struct. Funct. Bioinforma.*, vol. 78, no. 1, pp. 173–180, 2010, doi: 10.1002/prot.22553.

[15] U. Bastolla, Y. Dehouck, and J. Echave, "What evolution tells us about protein physics, and protein physics tells us about evolution," *Curr. Opin. Struct. Biol.*, vol. 42, pp. 59–66, Feb. 2017, doi: 10.1016/j.sbi.2016.10.020.

[16] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.