LU TP 19-26
June 2019

Exploration of an all-atom thermodynamic model to predict site-specific evolutionary rates in proteins

Fábio Dias Correia de Oliveira

Department of Astronomy and Theoretical Physics, Lund University

Master thesis supervised by Ingemar André

Abstract

Understanding the patterns of evolutionary sequence divergence is fundamental for comparative analyses like phylogenetics or genomics. The rate at which the different sites of protein sequences evolve is multifactorial and the causes of variation among them are highly convoluted. Inference methods have been developed to estimate site-specific evolution rates from sequence alignments. Moreover, several molecular traits have been found to correlate with site-specific rates: solvent accessibility, packing density and protein function are some of them. Correlations between rates and predictor variables allow to identify factors that influence rate variation, but they do not provide explicit mechanistic insights into why a given site is variable or conserved. Luckily, the field of protein evolution is amenable to the development of fundamental theory. Hence, mechanistic biophysical models have been proposed to explain the observed rates. Biophysical models are essentially based on protein stability - this is reasonable because stability is related to all the molecular features that correlate with evolutionary rates.

Norn et al. developed an all-atom thermodynamic model to predict site-specific evolutionary rates in proteins. The model has been shown to closely recapitulate the average amino acid substitution rate behaviour. However, the model fails to achieve the same level of accuracy for site-specific rate recapitulation. Several reasons have been put forward so as to explain the weak correlation; but two hold the most interest: propagation of stability prediction errors and the fact that the model relies on only a single protein structure to extrapolate site-rates that emerge within a protein phylogeny. The results obtained in this thesis support the hypothesis defended by Norn et al. that the propagation of stability prediction errors impacts the correlation value but are not enough to explain the average weak correlation; additionally, it is shown that a weighted average of the site-rates of the proteins in a given phylogeny does a better job at recapitulating its empirically inferred site-rates. In consequence, this effectively opens the doors for further adaptation of Norn et al. model to phylogenetic analysis.

Populärvetenskapligt sammanfattning

Proteins are fascinating: they are made up of a simple sequence of amino acids which folds to a complex functional structure. There is, however, a lot of redundancy – different sequences can give rise to the same functional structure. Organisms of different species possess analogous proteins that perform the same task and therefore have a similar structure, but different sequences. Often those analogous proteins have a common ancestor from which they diverged through different evolutionary pressures.

For proteins to evolve the units that compose them, amino acids, must change with time. Depending on their position in the protein, they evolve at different rates. To check which sites in the protein are more conserved and which are more variable multiple sequence alignments can be done. Sequence alignment studies allow to detect patches of the proteins where the amino acid sequence is the same. The conservation of specific amino acids in the same positions in all aligned proteins indicates that they are important. There are two main reasons for some regions to be highly conserved: the first is that those regions are critical for the stability of the functional structure, the second is that that they might be directly involved in the function of the protein. For example, they might be responsible for the binding to a specific substrate. The problem with the approach above is that there is no way to distinguish between the two factors.

With an atomistic model of the protein one can calculate what is the effect on stability caused by changing its amino acid sequence. Sites of the protein that if mutated lead to decreased stability are considered to be conserved and sites that if mutated lead to more stability or a neutral change are considered to be prone to a lot of variation. However, the atomistic model does not explain all the variability in rates that can be extrapolated from sequence alignments. There are two main factors for why this is the case: the stability effects that are calculated are not as accurate as needed and the model does not account for the fact that different proteins responded differently to the same mutations. The main aim of this thesis is to explore the impact of these two factors. Improving this protein sites prediction rate model is of utmost importance. If used together with the sequence alignment method, their synergy provides insight useful to understand when conservation of certain amino acids comes from stability, from functionality or from both. Consequently, it accelerates the process of functional sites identification, which is important for the development of new drugs targeting these sites.

Contents

1 Introduction

1.1 Evolution

Evolution is a change at the population level emergent from mutation, selection and replication of the organisms that compose the population. Mutations are caused by errors in the genetic information that are transfered to the offspring, resulting in population variability. Selection among the variants emerges when some generate descendancy faster than others. Eventually, one of the variants dominates the population and evolution takes place. Albeit simplistic, the iteration of the described events account for much of the diversity in life.

Evolutionary trees, also known as phylogenetic trees or phylogenies, are an attempt to represent evolution. They are a branching diagram depiction of biological entities that are connected through common descent. In a rooted phylogenetic tree, each branching point (node) represents the inferred most recent common ancestor of the species which branched from it, and the edge lengths in some trees may be interpreted as time estimates. In contrast, unrooted trees illustrate only the relatedness of the leaf nodes and do not require the ancestral root to be known or inferred.

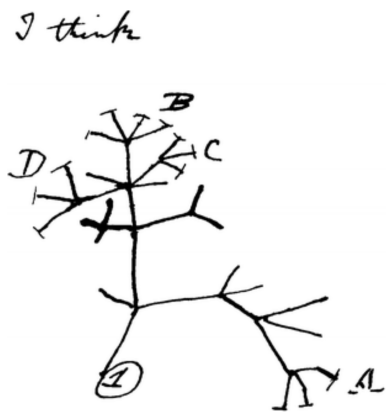

Figure 1: The first evolutionary tree sketched by Darwin (1837) in one of his notebooks [5].

Charles Darwin sketched his first evolutionary tree in 1837. And, perhaps not surprisingly the only illustration in Darwin's The Origin of Species is a phylogeny. In the century following his work, biologists constructed evolutionary relations using similar evidence as did Darwin: morphological differences between species. However, through the lens of a molecular understanding of life, an organism should not be reduced to an aggregate of macroscopic features, but rather, to the interplay of microscopic structures working at the molecular scale. The evolution of macrostructures merely arises from the evolution of the genetical information that encodes for its molecular components. Therefore, it is the genet-

ical differences between organisms that hold the relevant information for the construction of an evolutionary tree.

Even though mutation happens at the nucleotidic level, DNA sequences are not always used to infer evolutionary relations. Oftentimes, it is better to use protein sequences instead. For example, if the proteins being studied are very widely divergent, the likelihood of multiple substitutions at the same DNA base becomes unneglectable. This in turn leads to a higher probability that unobservable intermediate mutations and their respective proteins may have existed, causing the divergence time to be underestimated. Thus, in this case, using DNA data to infer evolutionary trees is very likely to lead to incorrect phylogenetic trees. Moreover, it cannot be forgotten that protein sequences are under selective constraints for protein function, and these are conserved over much longer periods than the individual codon choices. In 1958, Francis Crick foresaw the importance of proteins for the inference of evolutionary relations, as evidenced by his remark [1]:

"Biologists should realize that before long we shall have a subject which might be called 'protein taxonomy', the study of the amino acid sequences of the proteins of an organism and the comparison of them between species. It can be argued that these sequences are the most delicate expression possible of the phenotype of an organism and that vast amounts of evolutionary information may be hidden away within them.".

Nonetheless, a holistic approach is bound to be the key to unlock the evolutionary history of life. From the 1960s onwards, scientists increasingly used diverse types of genetic and molecular data (e.g. allozymes, chromosomes, DNA–DNA hybridisation, nucleotide and amino acid sequences) for phylogenetic inference [13].

From the comparison of sequences of amino acids or nucleotides, it is possible to derive a quantitative measure of relatedness between organisms. A method to compare the different sequences is to align them. The alignment of different sequences is not unique, given that there are different ways to align them. Nonetheless, the goal is always to overlap the sequences in an evolutionarly meaningful way, such that the probability that the sequences are related by a common ancestor is maximized. The final alignment is chosen from the comparison of different possible alignments corresponding to sliding the sequences relative to each other and also inserting gaps. The conceptual background is the idea that over time, the common ancestor sequence changes through a series of substitutions, deletions and insertions [21].

1.2 Protein evolution

Proteins can be thought to inhabit a sequence-space where all proteins are connected to all other through mutational paths. This sequence-space can be constructed in such a way that nearest neighbours only differ by a single amino acid [26]. For an easier visualization

one can imagine that there are only two types of amino acids: 0 and 1. All possible peptides of length three that are made up of this imagined binary amino acids can be mapped into the vertices of a cube (figure 2).

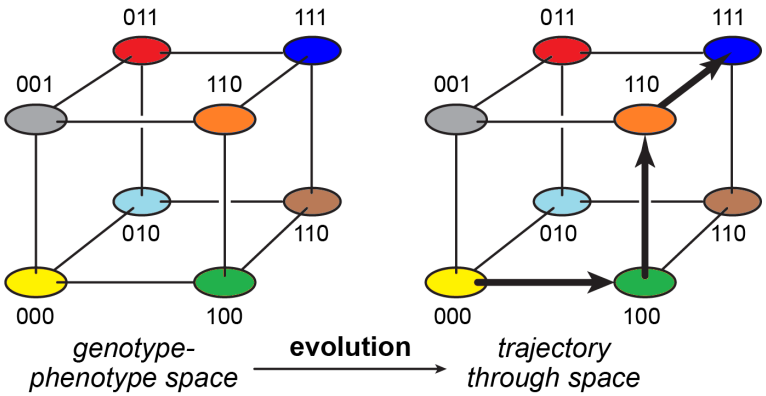

Figure 2: Illustration of the sequence space of a length three peptide composed of binary amino acids.

The transition between states can be modelled as a Markov process. For each site L of the length three peptide the process can be modelled by the rate matrices Q_L:

$$Q_L = \begin{pmatrix} 1 - q_{01}^L & q_{01}^L \\ q_{10}^L & 1 - q_{10}^L \end{pmatrix}, L \in 1, 2, 3 \tag{1.1}$$

where q_{01}^L is the transition rate from amino acid 0 to amino acid 1 at site L. It happens that for two of this imaginary peptides to be connected through an evolutionary trajectory two separate events have to occur: a mutation and the fixation of the new peptide in the population. Thus, the rate of evolutionary transitions can mathematically be written as:

$$q_{01}^L = r_{01} \times f_{01}^L \tag{1.2}$$

where r_{01} is the rate at which mutation from 0 to 1 occurs and f_{01}^L is the probability that the mutation gets fixed in the population.

In the real world there are 64 codons that code for 20 amino acids. Therefore, both the mutational and the evolutionary selection dynamics are much more complex than in the case of the imaginary peptide. All codons are connected to the other 63 codons by mutation events. For simplicity, in figure 3 the different mutational pathways that connect methionine and valine are depicted.

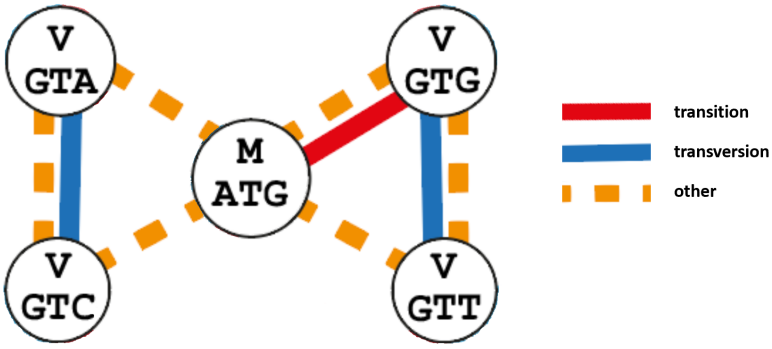

Figure 3: Illustration of the possible ways valine and methionine codons are connected. Codons can be connected by mutations other than transitions and transversions. As a multi-nucleotide mutation can cause the same mutational effect as a transition or transversion all the codons are connected by the dashed orange line.

There are different types of mutations and these have different rates of occurence. The majority of mutations only affect single-base pairs and are called point mutations. In this class of mutations, transitions (pyrimidine to pyrimidine, or purine to purine) happen more frequently than transversions (pyrimidine to purine, or vice-versa). However, and even though they are rarer, mutations that involve multi-nucleotide changes do occur.

It is a complicated matter to understand how a mutation affects the fitness of the organism and therefore the probability of the mutation to fixate in the population. The relation between the organism and one of its proteins is very pragmatic. The protein exists to carry out a function. Thus, the better it is able to perform it, the more beneficial it is for the organism. As structure is a critical determinant of function, mutations that affect the folding stability of the functional structure are deemed to impact the fitness of the organism. However, this is not the only determinant. For example, a mutation that increases the folding stability can be detrimental if it removes the binding specificity that the protein has to have to a certain substrate.

Proteins evolve under multiple biophysical selection pressures that are specific to their function and their chemical environment. Nonetheless, there are overarching tendencies which allow the inference of global transition rates. For example, replacements between arginine (positively charged) and aspartate (negatively charged) are under negative selection and have low rate, whereas replacement between isoleucine and valine (both hydrophobic, aliphatic and very nonreactive) are frequent and have high rate [11]. This global replacement pattern is useful to find homologues, infer phylogenies, reconstruct ancestral sequences, and predict functional sites.

The Markov process can be modelled in two different ways: either as a codon substitution model or as an amino acid substitution model. This means that the rate matrix, Q, would

be a 64 by 64 matrix or a 20 by 20 matrix, respectively. There is a significant methodological difference between the two approaches: a codon substitution model attempts to model the actual processes that underpin amino acid exchanges, while an amino acid substitution model only summarizes the final exchange pattern of amino acids.

1.3 Empirically inferred evolutionary transition rates

Empirical models are models that do not explicitly consider factors that shape protein evolution but attempt to summarize the substitution pattern from large quantities of real data. As evolutionary transition events cannot be measured directly, the transition rates have to be inferred from protein multiple sequence alignments (MSA). Thus, the Markov process is modelled at the amino acid level by a 20 by 20 rate matrix.

Different sites of the same protein do not have the same transition rates between different amino acids. Some sites are slow due to strong constraints, whereas other sites with low evolutionary pressure evolve rapidly. Therefore, if site-independency is assumed, it would be ideal to infer a rate matrix for each site. However, that would render too many parameters to infer. Hence, oftentimes it is assumed that all the sites can be modelled by the same rate matrix Q, albeit for a site-dependent scale factor r_i that allows to model site-rate variability. Thus, the individual transition rate matrix for each site is given by r_iQ. This Q matrix reflects the general biological, chemical and physical properties of amino acids, instead of the specific constraints of a given site.

The transition rates are assumed to be constant under evolution and the evolutionary process is considered to be stationary [34]. Moreover, the process is assumed to be time reversible. Thus it satisfies the detailed balance equations. These equations require that the equilibrium distribution of the process, $\Pi = (\pi_i)$, is such that:

$$q_{ij}\pi_i = q_{ji}\pi_j \tag{1.3}$$

Therefore, $Q = (q_{ij})$ can be rewritten as:

$$\begin{cases} q_{ij} = \pi_j z_{ij}, \text{ for } i \neq j \\ q_{ii} = -\sum\limits_{j \neq i} q_{ij}, \end{cases} \tag{1.4}$$

where $Z = (z_{ij})$ is symmetric, independent of Π and called the exchangeability matrix.

Time and rate are confounded in this inference process - it is not possible to conclude whether a site presents a lot of variability, because it had an enormous amount of time to evolve or because it has evolved under weak evolutionary constraints. Therefore, the rate matrix is scaled so that the rate of substitution is one:

$$\sum_i \sum_{j \neq i} \pi_i q_{ij} = -\sum_i \pi_i q_{ii} = 1 \tag{1.5}$$

This means that a time unit corresponds to one expected substitution per site. Consequently, in molecular phylogenetics, times and branch lengths are measured in average number of substitutions per site. Here the key word is average, since different sites do not evolve at the same rate.

The LG matrix, a general amino acid replacement matrix was estimated using a maximum-likelihood approach [11]. It estimated Q by choosing the Q that maximizes the likelihood that the Markov process explains the empirical data - multiple sequence alignments, phylogenetic trees and site rate variability. This is not trivial, because the empirical data is not observable, it has to be inferred and is interdependent. Therefore, simultaneous inferrence of the empirical data and of the rate matrix would appear to be necessary. However, as that would be too computionally heavy, some heuristic principles are used for it to be possible to calculate the likeliest Q conditional on some previously established evolutionary relatioship between the sequences [11]. Thus, the likelihood that Q explains the phylogenetic tree, the multiple sequence alignment and the site-rate variability is given by:

$$L(\mathbf{Q}; T, \mathrm{MSA}) = \prod_i L(\mathbf{Q}; r_i T, \mathrm{MSA}_i) = \prod_i L(r_i \mathbf{Q}; T, \mathrm{MSA}_i) \tag{1.6}$$

where T is the phylogenetic tree, MSA the multiple sequence alignment and r the site-rate. Index i is the protein site identifier. $L(r_i\mathbf{Q}; T, \mathrm{MSA}_i)$ is computed by applying $e^{\mathbf{Q}t}$ to each branch of T. The LG rate matrix was inferred from a large quantity of evolutionary data: 3912 alignments from Pfam, comprising of $\sim$ 50000 sequences and $\sim$ 6.5 million residues overall [11]. Therefore, it confidently encompasses the average amino acid replacement behaviour.

1.3.1 Empirically inferred site-specific conservation rates

Rate4site is an algorithmic tool that can be used to infer site-specific conservation rates from the topology and branch lengths of the phylogenetic tree (which are estimated from a multiple sequence alignment), as well as the underlying Markov process (LG rate matrix) using empirical Bayesian inference [15].

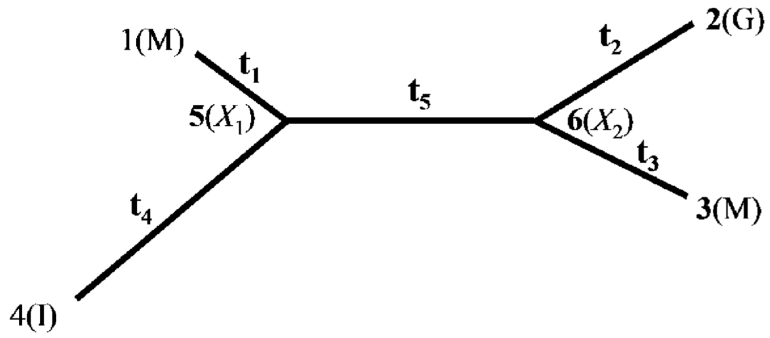

Figure 4: Illustration of the four-taxon unrooted tree used to illustrate the inference of site rates via the empirical Bayesian method implemented on the rate4site program [15]. The leaf nodes are labelled 1 to 4; the internal nodes are labelled 5 to 6. Branch lengths are marked by t_i, where i is the branch identifier. Capital letters in parentheses are one-letter abbreviations for amino acids.

The likelihood that the evolutionary process in figure 4 is explained by rate r is given by:

$$P(\mathrm{MSA}_L|r,T) = \sum_{X_1,X_2\in\{amino\,acids\}} \pi_{X_1} \times P_{X_1,M}(rt_1) \times P_{X_2,G}(rt_2) \times P_{X_2,M}(rt_3) \times P_{X_1,I}(rt_4) \times P_{X_1,X_2}(rt_5) \tag{1.7}$$

where $P_{i,j}(rt)$ is given by $\{exp(rtQ_{LG})\}_{ij}$.

The distribution of site-rates can generally be fitted on a Gamma distribution [30]. Hence, rate4site uses a Gamma distribution determined by the empirical data as the prior distribution for the Bayesian inference process. As the computation of a Bayesian estimate based on a continuous Gamma distribution is computationally impracticable [33], k discrete Gamma categories are used to approximate the continuous Gamma function. Each rate category has equal prior probability $(1/k)$ and the mean of each of them, r_i, is used to represent all the rates within the category. Therefore, the posterior probability for each rate category for site L is given by:

$$P(r|\mathrm{MSA}_L,T) \cong \frac{P(\mathrm{MSA}_L|r,T)P(r)}{\sum_{i=1}^{k} P(\mathrm{MSA}_L|r_i,T)P(r_i)} \tag{1.8}$$

Rate4site uses the average rate for each site L as the representative site-specific rate:

$$E(r|\mathrm{MSA}_L,T) \cong \sum_{i=1}^{k} P(r_i|\mathrm{MSA}_L,T)r_i \tag{1.9}$$

The approximation signs in equations (1.8) and (1.9) are present because the discrete Gamma model is being used.

1.4 First principles calculation of evolutionary transition rates

First principles models are formulated at the codon level and separate mutational biases at the nucleotide level from selective constraints at the amino acid level. They account for features of sequence evolution, such as transition-transversion bias and base or codon frequency biases and make use of physiochemical distances between amino acids to specify nonsynonymous substitution rates.

To explore the root cause of the substitutions rates Norn et al. developed a simple model to calculate amino acid rate matrices [18]. The evolutionary constraints are controlled by a fitness function that reports on the thermodynamic effects of amino acid mutations on a single native sequence. The model is referred to as the Sequence-static Thermostability Evolutionary Dynamics (STED) model.

The basic assumption is that evolution is largely governed by a constraint on proteins to remain folded into active conformations. Therefore, the contribution of a protein to an organism's fitness is assumed to be proportional to the fraction of protein folded in the functional conformation [31]. For a two-state folding model with stability $\Delta G = G_{native} - G_{unfolded}$ the fraction of folded protein is:

$$w = \frac{1}{1 + exp(\Delta G/RT)} \tag{1.10}$$

The assumption was formalized by Norn et al. into a model of protein fitness based on computed stabilities of proteins from an atomistic model. This implies that to assign a fitness contribution ΔG has to be computed for each mutation. This is done by using the folding stability of the native sequence as a global free parameter and by calculating the free energy change resulting from the mutation ($\Delta\Delta G$). Thus, the folding stability of sequence variants is approximated as:

$$\Delta G_j^L = \Delta G_{native} + \Delta\Delta G_{ij}^L \tag{1.11}$$

where ΔG_j^L is the ΔG of the folded state of the variant protein (amino acid j in site L instead of amino acid i) and $\Delta\Delta G_{ij}^L$ is the variation of Gibbs free energy upon substitution of i by j in site L.

The model accounts for nucleotide transition and transversion rate bias, multi-nucleotide codon changes and the number of codons per amino acid as the origin of protein variability. Hence, the rate matrix Q that describes the transitions is a 64 by 64 matrix. The rate of mutation proposal is given by:

$$P_{uv} = \begin{cases} 1 + \gamma, \text{ if single-base pair transversion} \\ \kappa + \gamma, \text{ if single-base pair transition} \\ \gamma, \text{ if else} \end{cases} \tag{1.12}$$

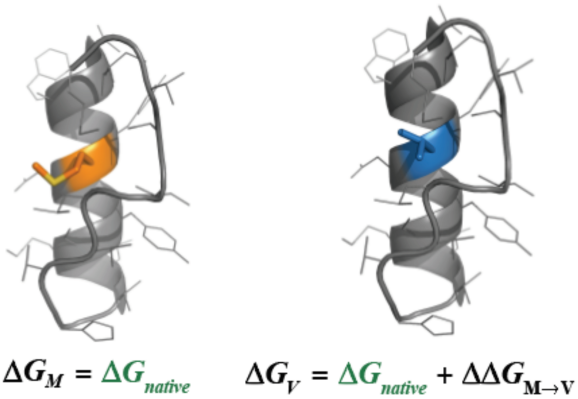

Figure 5: Illustration of the calculation of the ΔG of a variant sequence that only differs from the native by one amino acid (resulting of the mutation from methionine to valine) [18].

where κ is numerically equal to the transition/transversion rate ratio and γ is the rate of whole-codon mutations [18]. u and v represent different codons.

Norn et al. [18] simulated evolution as a Wright-Fisher process. A population evolves under the Wright-Fisher process if, at each time step, all individuals reproduce with a probability proportional to their relative fitness followed by the death of all individuals of the parent generation while maintaining the population constant. Motoo Kimura [9], derived an approximation of the fixation probability of new mutations under the process described above. According to him the fixation probability of a given mutation depends on the relative change in fitness that it causes ($s^L_{uv} = w^L_v/w^L_u - 1$) and on the size of the effective population (population that can generate descendents, N) [9] in the following manner:

$$f^L_{uv} \approx \frac{1 - exp(-2s^L_{uv})}{1 - exp(-4Ns^L_{uv})} \tag{1.13}$$

Therefore, the elements of Q are given by:

$$q^L_{uv} = P_{uv} \times f^L_{uv} \tag{1.14}$$

1.4.1 First principles calculation of site-specific conservation rates

To get insight on the dynamic at the amino acid level, the Markov process of codon substitution is transformed into an amino acid substitution model by grouping synonymous codons. Firstly, the equilibrium frequency of each codon is determined [7] [24]:

$$\pi^L_u = \frac{exp(4Nw^L_u)}{\sum_i exp(4Nw^L_i)} \tag{1.15}$$

where w_u^L is the proportion of folded variant protein with the amino acid corresponding to codon u in site L. Secondly, the elements of the amino acid substitution rate matrix are calculated using the aggregation approach presented by Yang et al. [34]:

$$q_{ij}^L = \sum_{u \in i} \sum_{v \in j} \frac{\pi_v^L}{\pi_i^L} q_{vu}^L \tag{1.16}$$

where the double sum is over all the codons that correspond to i and j amino acids.

The expression above is used for the determination of site-specific amino acid flux matrices and site-rates. The site-specific rate r_L is given by:

$$r_L = \sum_i \sum_{j \neq i} \phi_{ij}^L \tag{1.17}$$

The flux between a pair of amino acids at site L is given by:

$$\phi_{ij}^L = \pi_i^L q_{ij}^L \tag{1.18}$$

Moreover, (1.15) can be averaged across sites to find the mean amino acid replacement behaviour.

1.5 First principles model versus empirical models

The four free parameters of the STED model (ΔG_{native}, N, κ, γ) were optimized by Norn et al. by minimizing the error between model-predicted rate matrices and the LG empirical rate matrix. To find optimal parameters for the STED model a set of 66 non-redundant proteins with diverse folds and a total of 8907 sites was used [18]. The minimum error solution was found to be: $\Delta G_{native} = -6.25\ kcal/mol$, $N = 10^{4.2}$, $\gamma = 0.1$, $\kappa = 2.7$ [18]. This parameterization, referred to as Q-matrix optimized STED model (QOSTED), is consistent with the values found in nature [18]. QOSTED recapitulates the complex pattern of empirical replacement rates summarized by the LG rate matrix remarkably well: it explains 65% of the total variation [18]. Norn et al. concluded that the $\Delta\Delta G$ prediction errors alone are likely to contribute to at least 21% of the variance in the LG matrix that is unexplained by their model.

The site-rates for every site in their benchmark protein dataset were also compared to the rates inferred phylogenetically from multiple sequence alignments using the rate4site program. The effective population size parameter was reoptimized since recapitulation of evolutionary rates of each site is fundamentally different from modeling average substitution patterns. The difference lies in the fact that STED calculates site-rates that reflect the stability constraints imposed by a single protein structure, in contrast rate4site infers site-rates that reflect the empirical evolutionary dynamic of a protein phylogeny. Hence, to achieve optimum correlation with rate4site rates the effective population parameter should

be lower than in QOSTED, since that decreases the selection pressure imposed by the single protein structure from which STED rates are modeled. The optimal effective size for recapitulation of site-rates is $N = 10^{2.2}$, this new parameterization is refered to as rate optimized STED model (ROSTED) [18]. The correlation between rates of individual sites inferred phylogenetically with rate4site and rates of individual sites calculated with ROSTED is low, $\rho^2 = 0.27$ (ρ is the Pearson correlation) [18]. Moreover, it performs comparably to the structural correlate Weighted Contact Number (WCN) [18], which for each site attributes a value equal to the sum of the number of neighbouring sites in the protein weighted by their inverse square distance to the focal site. WCN is a measure of the packing density of each site which has been shown to correlate with site-specific rates [4]. This is a surprising result given the fact that ROSTED calculates site-specific rates considering all-atom detail.

There are several reasons that might explain ROSTED weak performance. The source of errors might come from the $\Delta\Delta G$ predictions. If one considers the propagation of $\Delta\Delta G$ errors to be the only source of error, the expected correlation between true and predicted rates would be $\rho^2 = 0.90$ [18]. However, the core reason for the low correlation is more likely to be due to the different methodology rate4site and ROSTED use to calculate the site-specific rates: rate4site inferred rates emerge as an average property of the multiple sequence alignment while ROSTED rates are modelled conditional only on a single structural environment.

2 Overview

The rates calculated using the ROSTED model have on average a correlation with rate4site inferred rates of $\rho^2 = 0.27$ (ρ is the Pearson correlation). Norn et al. estimated that if the propagation of $\Delta\Delta G_{Rosetta}$ errors was the only factor needed to control for, ρ^2 would only drop to 0.90. However, the degree to which other factors influence the correlation was not estimated by Norn et al., e.g. the effect of the selective pressure being conditioned only on a single structure or the fact that functional fitness constraints are not modelled. The main aim of this thesis work is to further understand how those factors influence the correlation between ROSTED rates and rate4site rates.

Norn et al. suggested a novel method in which the ROSTED model can be used to identify functional sites when predicted rates conflict with empirical rates. According to them, the method can predict up to twice the amount of functional sites than when just using inferred site-rates. As part of this thesis work, Norn et al. method is tested on the model protein $GB1$.

In short, this thesis reports on:

1. the assessment of the impact that the propagation of $\Delta\Delta G_{Rosetta}$ errors have on the recapitulation of rate4site rates by ROSTED through the use of experimental thermodynamic stability data for nearly every single mutant of a small 56-residue protein ($GB1$);

2. the assessment of the impact that the methodological difference between rate4site and ROSTED have on the recapitulation of rate4site rates by ROSTED through the weighted average of the site-rates of the proteins within $GB1$ inferred phylogeny;

3. the identification of $GB1$ functional sites from the comparison of STED rates and rate4site rates and assessment of their reliability by studying the literature on $GB1$.

Rate4site inferred rates are the expected values of the posterior distributions of rates. Norn et al. quantified the error associated with using the expectation value from the site-specific standard deviation in rate. By sampling site-rates within one standard deviation of the expected rate, they concluded that the average correlation between sampled site-rates and the expected rate value for each site would be $\rho^2 = 0.90$ [18]. In this thesis work the sampling was done directly from the posterior distributions of rates. This was done to obtain a more realistic ensemble of the possible combinations of site-rates a phylogeny can have. Nonetheless, the correlations between empirical and inferred rated were estimated both as the correlation with the expected value for the site-rates and as the average correlation value with the sampled rates.

2.1 Using experimental $\Delta\Delta G$ to calculate site-specific exchange rates

The protocol that was used by Norn et al. to predict $\Delta\Delta G$ values is fairly accurate, however it is far from perfect ($\rho^2 = 0.56$ between prediction an experiments [20]). To be able to benchmark the impact of its inaccuracies, an experimental thermodynamic stability data set for nearly every single mutant of a small 56-residue protein ($GB1$) was used [16]. The existence of this data set is greatly useful for an accurate estimation of the accuracy of predicted $\Delta\Delta G$. In this thesis work the correlation of the STED rates calculated with experimental $\Delta\Delta G$ with the rate4site rates was compared to the correlation obtained when using predicted $\Delta\Delta G$ from X-ray crystal structures and from homology models of homologues to $GB1$.

2.2 Averaging together site-specific rates from different proteins of the same phylogeny

The evolutionary rates given by STED for each site are only conditioned on the fitness pressure exerted by one structural environment. However, the rates rate4site infers considers the variation of amino acid in homologous sites of different proteins, thus it incorporates the fitness pressure exerted by the different structures on the phylogeny. To

accommodate this difference in methodology, instead of comparing the rates calculated directly from STED with rate4site, STED site-rates from the different structures should first be aggregated together through a weighted average. Thus, it allows the final rate value for each site to better represent the evolutionary constraints imposed by different structural environments. The weights needed to calculate the average cannot be simply chosen to maximize the correlation of the STED rates with the empirical rates since that would be an artificial fit and therefore would not necessarily reflect the true influence of the different structural backgrounds. In other words, the average it is aimed for should provide a statistic that can be used to estimate a meaningful parameter. At the same time it should be such that similarity by descent does not obscure subtler signals in the data which might have played an important evolutionary role. At the moment there is one tool that appears to be promising in providing the right weights: BranchManger [28]. The process it follows weighs the contribution of each sequence according to its phylogenetic placement, therefore providing a balanced representation of the different structural environments. In this thesis work the STED rates for each of the structures in the phylogeny were compared to the averaged STED rates in their ability to recapitulate the empirical rates given by rate4site.

2.3 Functional sites identification

Norn et al. found many examples where ROSTED correctly predicts site-conservation by detecting sites where mutations result in large energy effects through disruption of well-formed hydrogen bonding networks, formation of cavities in the protein core, or conflicts from steric clashes. However, ROSTED cannot detect site-conservation when it is solely driven by functionality constraints that are not coupled to the stability of the folded state. For those sites having the information from the empirical rates and STED rates allows for an educated guess on the likelihood of the site being functional. If the site is seen to be highly conserved from a multiple sequence alignment, but its STED rate is very high that indicates that the site might be under a high functional constraint. Norn et al. introduced a novel metric, $r_{site}^{ratio} = r_{site}^{STED}/r_{site}^{rate4site}$, which discriminates stability from function for each site; $r_{site}^{ratio} > 1$ indicates a high likelihood of the site being functionally important. There is a vast body of experimental facts and theoretical findings for $GB1$ and thus a literature search was done to reliably investigate whether r_{site}^{ratio} correctly identifies $GB1$ functional sites or not.

3 Methods

3.1 Rate4site program

The sole obligatory input to rate4site is a MSA file. The program then computes a phylogenetic tree that is consistent with the available MSA and calculates the relative evolutionary rate for each site in the MSA. For this thesis work rate4site was run with the following settings: empirical Bayesian method for the inference of rates and LG rate

matrix as the evolutionary model (the remaining settings were left on default).

3.1.1 Multiple sequence alignment

The set of homologous sequences to $GB1$ and their MSA was obtained through the ConSurf web service (default parameters were used). The original set of sequence homologues (48 different sequences) had some sequences with engineered mutations, hence they were removed. Thus, the multiple MSA which was used as input to rate4site had 30 different sequences.

3.1.2 Ensemble

The ensemble has 1000000 sampled sets of site-rates. Each set of site-rates in the ensemble was sampled from the posterior probability distribution with the constraint that the average rate of each set is between 0.9 and 1.1. This constraint was imposed because the site-rates are a measure of the relative evolutionary rate within the phylogeny.

3.2 Experimental stability data

The experimental $\Delta\Delta G$ data set was obtained through domain-wide comprehensive mutagenesis of a simple globular protein, the $B1$ domain of the IgG-binding protein G ($GB1$) by Nisthal et al. [16]. No site was mutated to cysteine (Cys) or tryptophan (Trp) and no mutations were performed at position 43. This would correspond to a thermodynamic stability data of 935 $GB1$ variants, however only 830 variants passed the experimental requirements. Therefore, the used data set corresponds only to those 830 variants. The experimental value for the Gibbs free energy difference between the unfolded state and the folded state of $GB1$ is -4 $kcal/mol$ [16].

For the mutations for which the $\Delta\Delta G$ was not available, an assumption had to be made regarding their influence to the protein fitness. It was assumed that those mutations would be very energetically unfavourable and would therefore not happen. The arbitrary high value of 1000 was assigned to the ΔG of the $GB1$ variants they correspond to. This approach would only be ideal if the mutations that are chosen to be prohibitive are not seen in the multiple sequence alignment. It happens that for sites 2, 12, 33, 38, 42 and 43 those mutations did occurr. Therefore, the rates calculated for those sites are not accurately modeled. The rates obtained using experimental $\Delta\Delta G$ are referred to as experimental rates.

3.3 Calculation of STED model rates

The STED model has four free parameters: ΔG_{native} (folding stability of the native sequence), N (effective population), κ, γ. As in this thesis work the focus was on the individual site-rates, the ROSTED parameterization was used for all parameters except

ΔG_{native}: $N = 10^{2.2}$, $\gamma = 0.1$, $\kappa = 2.7$ [18]. ΔG_{native} was not treated as a global paramater in this thesis work as was by Norn et al. for QOSTED and ROSTED. Here, for each different protein, ΔG_{native} was estimated by maximizing the correlation between rate4site rates and the rates calculated from the STED model.

$$\Delta G^{L}_{j\ optimized} = \Delta\Delta G^{L}_{ij\,Rosetta} + \Delta G_{native} \tag{3.19}$$

$$\Delta G^{L}_{j\ optimized} = \Delta\Delta G^{L}_{ij\,Rosetta} + \Delta G_{native} = E^{L}_{j\ Rosetta} - E^{L}_{i\ Rosetta} + \Delta G_{native} \tag{3.20}$$

ΔG_{native} was not optimized for the calculation of the experimental rates. As the experimental ΔG_{native} was known, it was used instead of the corresponding optimized value.

3.4 Rosetta macromolecular modeling suite

All structural modeling in this thesis work was done with the Rosetta macromolecular modeling suite [12]. The energy function implemented in Rosetta is a combination of physics-based and statistics-based potentials. Rosetta energies are on an arbitrary scale referred to as REU (Rosetta Energy Unit). The Rosetta energy function is protocol specific, which means that the absolute values of Rosetta energies are not necessarily comparable between protocols, but energies within each protocol are comparable among each other. Nonetheless, it is possible to derive a conversion from Rosetta-produced energy values to absolute experimental values. The approach is to derive a line of best fit between experimental values and Rosetta predicted values, which yields a conversation factor.

3.4.1 $\Delta\Delta G$ prediction

The $\Delta\Delta G$ prediction method is based on a modified version of the method presented by Park et al. [20], but with a cutoff in the Lennard-Jones potential set to 6.0 Å. This $\Delta\Delta G$ method samples backbone degrees of freedom for the mutated and neighbouring residues in the sequence and allows repacking of all-side chains in energetic contact ($> 0.1\ kcal/mol$) with the mutated residue - the sole input to the protocol is a PDB file format of a protein structure. For each possible single mutation variant of the inputted protein the output is $E^{L}_{j\ Rosetta}$ (L indicates the site and j the amino acid). Therefore, the $\Delta\Delta G$ difference between the protein variant with amino acid j in site L and the one with amino acid i in site L is:

$$\Delta\Delta G^{L}_{ij\,Rosetta} = E^{L}_{j\ Rosetta} - E^{L}_{i\ Rosetta} \tag{3.21}$$

For this thesis work the used conversion value was calculated by Norn et al. [18] and is equal to 1.947 $REU/kcal/mol$.

3.4.2 Using X-ray crystal structures

The Protein Data Bank (PDB) search tool "Search by Sequences" was used to find which sequences of the MSA provided by the ConSurf web service had a crystal structure. Only 3 out of the 30 sequences had crystal structures deposited in the PDB ($GB1$, $GB2$ and $GB3$). The crystal structures had to be prepared before being used as inputs to the $\Delta\Delta G$ prediction protocol. Each of the structures found in the PDB was adapted to the Rosetta energy function by using the FastRelax protocol as described by Nivon et al. [17] but with Cartesian space sampling. This was done to eliminate potential energy strains that could bias the calculation of $\Delta\Delta G$.

3.4.3 Homology modeling

Homology modeling was used to find the folded structures of the protein sequences in the MSA not represented in the PDB. It was also done for the protein sequences with crystal structures in the PDB, since it is important to verify if there is any difference between the STED site-rates calculated from a homology model and the ones calculated from a X-ray crystal structure of the same protein sequence. The used protocol, RosettaCM, creates a homology model of a protein sequence if given PDB files corresponding to one or more template structures [27]. 1000 homology models were created for each sequence. The model with the lowest energy was chosen as the representative structure. To not bias the modeled structure multiple templates were used. The PDB ID of the templates are: 1PGA, 1IGD, 1QKZ, 2ZW0 and 3U2S.

3.5 Averaging of STED rates

BranchManager attributes a weight to each sequence according to its phylogenetic placement, therefore providing a balanced representation of the different structural environments. Loosely speaking the BranchManager weight on a taxon can be thought of as the fraction of total phylogenetic branch length to which its datum can be attributed. Technically speaking the data is assumed to evolve according to a Brownian motion process on the phylogenetic tree which is conditioned by its leaf nodes. The average value obtained by using the weights BranchManager outputs for each leaf node of the tree corresponds to the expectation of the average Brownian trajectory value. For the two-taxon example in figure 6 the output will simply be $\frac{1}{2}$ for each taxon. In an analogy to physics the average can be thought of as the "phylogenetic center of mass" [28].

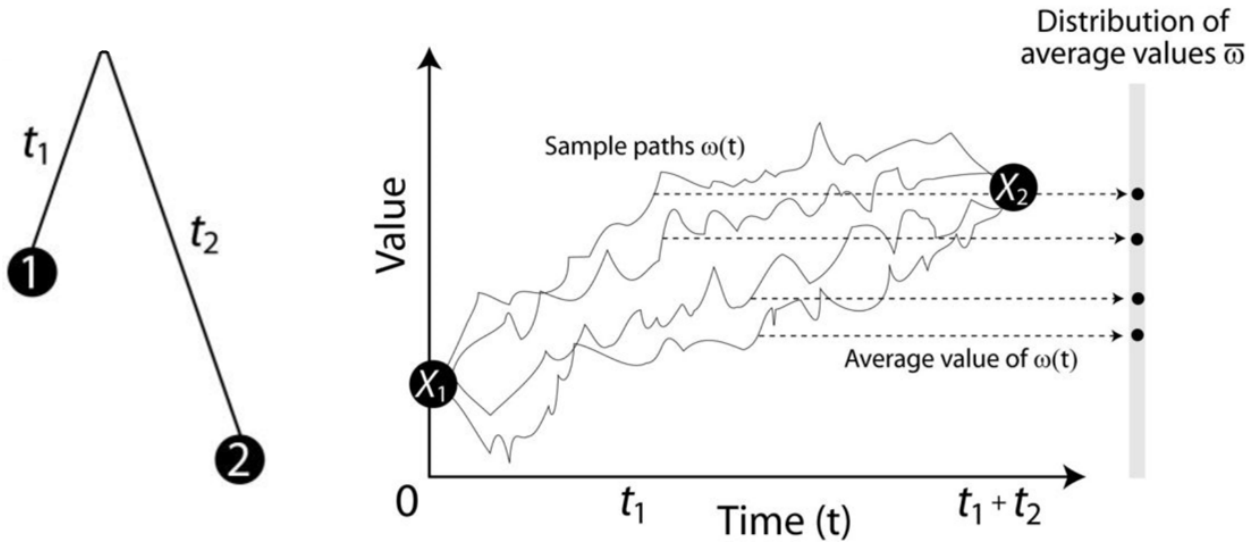

Figure 6: Illustration of how BranchManager calculates the weights for each taxon of a two-taxon unrooted tree [28].

The sole obligatory input to BranchManager is a phylogenetic tree. The phylogenetic tree used was the same rate4site constructed from the MSA it received as input (albeit pruning of some leaf nodes). As some sequences have gaps not all sites have a homologue in the phylogeny. Therefore, the averaging has to be done site by site. A Python script was designed to do the averaging site by site. Firstly, it selects the sequences which have a gap in the site in question. Secondly, it prunes those sequences out of the phylogenetic tree while maintaining the topology and branch lengths (pruning method available in the ETE Toolkit [6]). Thirdly, it feeds the edited phylogenetic to BranchManager. Finally, it returns the weighted average rate for the site.

4 IgG-binding domains of streptococcal protein G

Protein G is found on the surface of streptococci and binds to IgG with high affinity [29] [3]. It is composed of several domains of which two or three (depending on the strain) are IgG-binding domains . The different IgG-binding domains, $B1$, $B2$ and $B3$ all have similar secondary and tertiary structures. $B1$ and $B2$ differ only in two amino acids, $B2$ and $B3$ in four, and $B1$ and $B3$ in six (the differences in sequence can be inspected in the multiple sequence alignment in figure 12). Despite their small sizes, the IgG-binding domains have two separate IgG-binding sites on their surface, each interacting respectively with specific, independent sites on the Fab or Fc fragments of the IgG [25] [2]. They all consist of a four-stranded β-sheet with an α-helix on top. The α-helix and the β-sheet are tightly packed with a hydrophobic core in between.

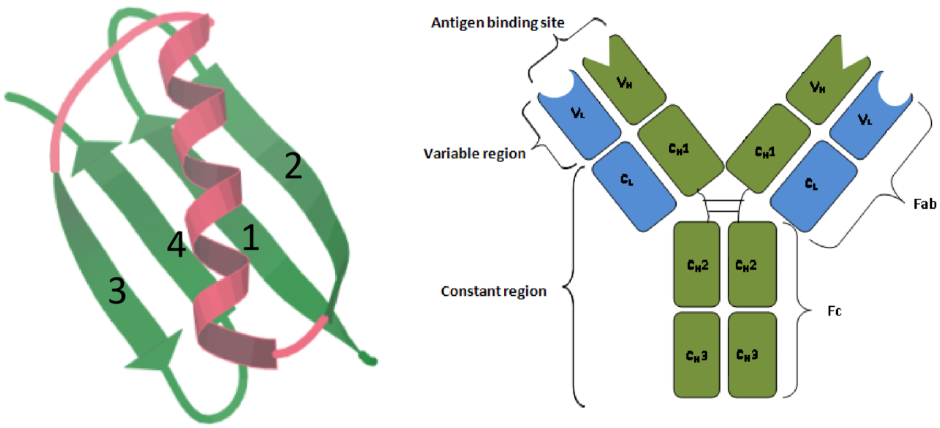

Figure 7: Image to the left: illustration of the folded structure of the IgG-binding domains. The β-strands are ordered from the N-terminal to the C-terminal. Image to the right: illustration of the components of the IgG protein [8].

4.0.1 Binding to Fab fragment

The binding between the Fab fragment and one of the IgG-binding domains is the result of an alignment of β-strands of the two proteins which extends the β-sheet from the IgG domain into protein G. The complex is stabilized by a network of hydrogen-bonds where residues $T11$ to $T17$ play a role [2]. Additionally, a smaller region of contact occurs between the C-terminal end of the α-helix and the first β-strand of the $CH1$ domain of the Fab fragment - it involves residues $Y33$ and $N37$ [2].

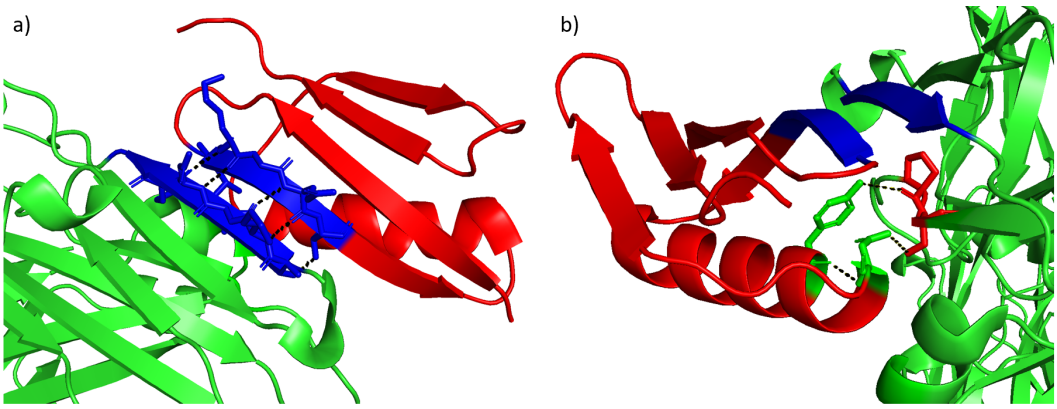

Figure 8: a) Depiction of the alignment of IgG and IgG-binding domain β-strands and of the hydrogen bonds holding them together (IgG in green, IgG-binding domain in red, β-strands responsible for the extension of the β-sheet in blue). b) Depiction of the contact that occurs between residues 33, 37 (in green) and the Fab fragment. The hydrogen bonds are represented as dashed black lines. The images were created from PDB entry 1UWX on PyMol. [23]

4.0.2 Binding to Fc fragment

The complex with the Fc fragment has been shown to be mediated primarily by side-chain contacts between the two proteins [22]. The interface is formed by a double "knobs-into-holes" interaction in which a knob from one of the IgG-binding domains protrudes into a hole in the Fc fragment, and vice versa. $E27$ of $GB1$ fits into a hole on the surface of the Fc fragment. The carboxylate of $E27$ is held in position by a hydrogen bond from the amino group of $K31$, the neighbouring residue, on the surface of $GB1$. The second knob-into-hole interaction is formed by the protusion of N from the surface of the Fc fragment into a hole in $GB1$ bordered by N35, D36, D40, E42 and W43. W43 forms a hydrogen bond with the N on the Fc fragment [25].

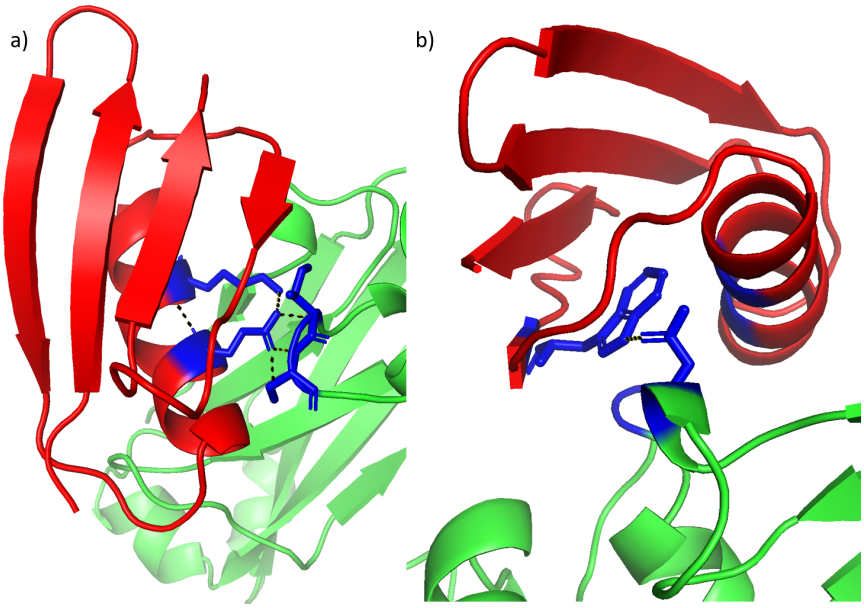

Figure 9: a)Depiction of the protrusion $E27$ of the IgG-binding domain being held by $K31$ and interacting through hydrogen bonds with hole on IgG (IgG in green, IgG-binding domain in red, interacting residues in blue). b) Depiction of the interaction between the protrusion on IgG and the hole on the IgG-binding domain (interacting residues in blue) The hydrogen bonds are represented as dashed black lines. The images were created from PDB entry 1FCC on PyMol [23].

4.0.3 Folding pathway of $GB1$

S. Kmiecik at al. provides a detailed characterization of the folding pathway of $GB1$. They demonstrated that the folding is initiated by the formation of a specific nucleus involving the hydrophobic core residues ($Y3$, $L5$, $F30$, $W43$, $Y45$ and $F52$) [10]. Moreover, they showed that those residues are evolutionarily conserved among proteins that share a similar fold to $GB1$ [10]. They also stress the importance of the early formation of the second hairpin as a key role in $GB1$ folding [10].

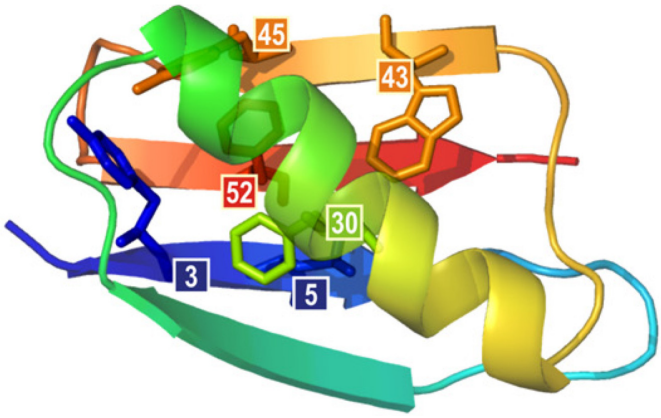

Figure 10: Illustration of $GB1$ with the nucleus residues labeled by their position in the sequence [10].

5 Results and Discussion

5.1 Rate4site

The phylogenetic tree inferred by rate4site and the multiple sequence alignment outputted by the Consurf web service can be seen in figure 11 and figure 12, respectively.

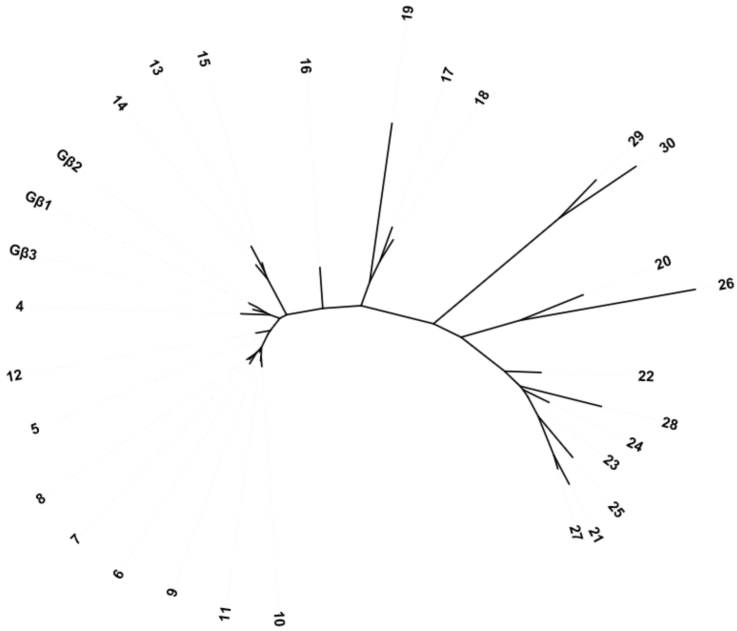

Figure 11: Unrooted phylogenetic tree inferred by rate4site from the multiple sequence alignment (figure 12). The illustration of the phylogeny was designed using the online tool iTOL [14].

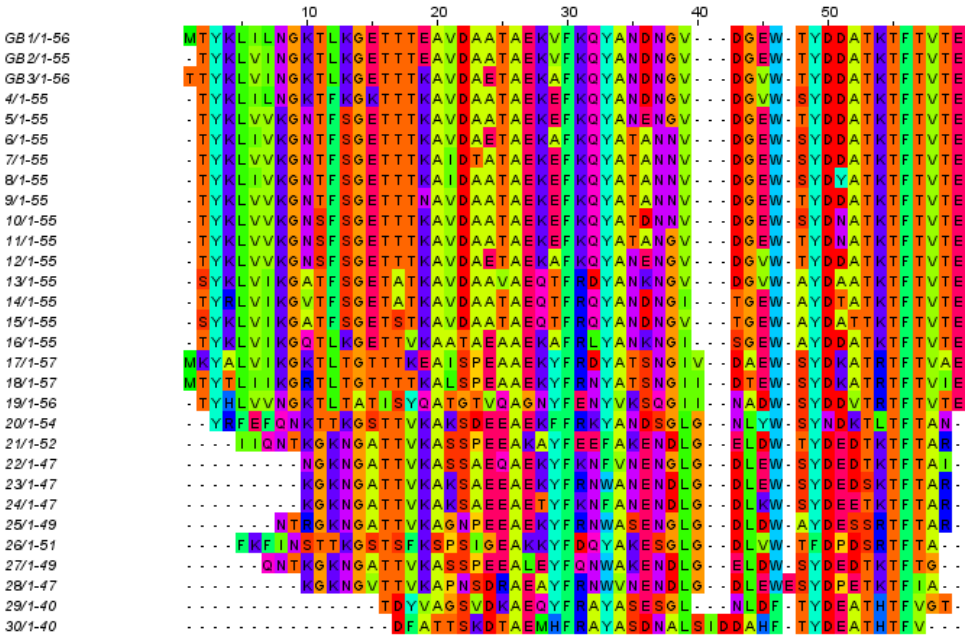

Figure 12: Multiple sequence alignment inferred by rate4site. The black dashes represent the gaps. The illustration of the alignment was obtained by using MSAViewer [32].

From the 30 protein sequences in the MSA only 3 have PDB entries for their X-ray crystal structures. Hence, $GB1$ is represented by a homology modeling structure and two crystal structures (PDB entry 1PGA with 2.07 $\AA$ resolution and PDB entry 1PGB with 1.92 $\AA$ resolution); $GB2$ is represented by a homology modeling structure and one crystal structure (PDB entry 1QKZ with 1.95 $\AA$ resolution); $GB3$ is represented by a homology modeling structure and two crystal structures (PDB entry 1IGD with 1.1 $\AA$ resolution and PDB entry 1PGX with 1.66 $\AA$ resolution). The remaining protein sequences (sequence 4 to 30) are represented only by their homology modeling structures.

The rate value that rate4site infers for each site (average rate calculated using equation (1.9)) is a proxy for the empirical value. That being said, it is important to highlight that the average rate value summarizes the discrete posterior probability distribution (1.8) and the later approximates the most likely distribution of rates for each site based on a Bayesian approach.

To benchmark how well the inferred rate4site site-rates explain the possible set of evolutionary rates that can be sampled from the discrete posterior probability distributions (figure 13), the correlation of each element of the ensemble (generated by sampling the discrete posterior probability distributions) with the inferred site-rates was calculated. The result is summarized in the histogram below (figure 14). The mean correlation value is 0.8 and the standard deviation is 0.05. Therefore, on average the inferred rates for the protein in question explain 64% of the total variation in the ensemble.

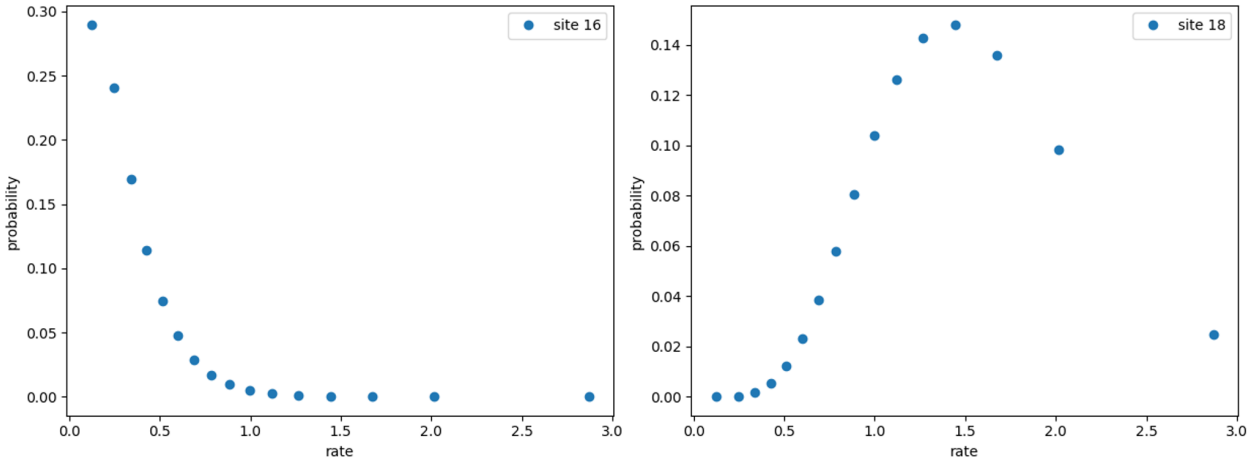

Figure 13: Discrete posterior probability distribution of the site-rate of site 16 and site 18. They highlight the uncertainty inherent to site-rate inference.

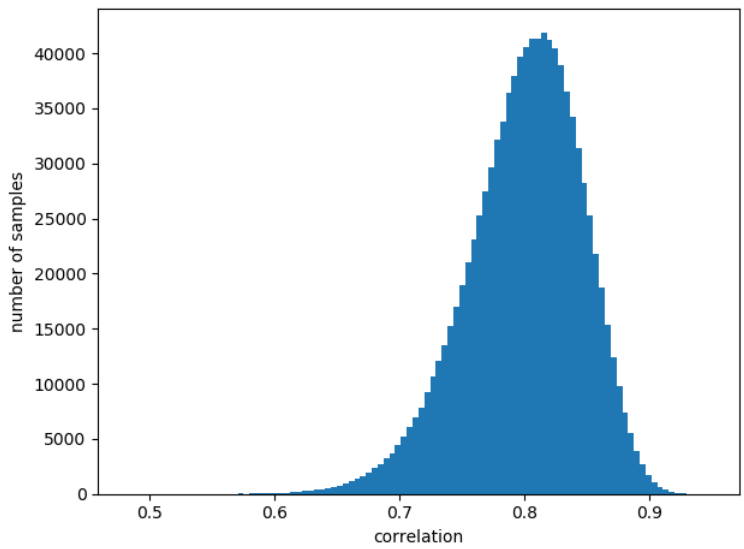

Figure 14: The histogram shows to what degree the average rates represent an ensemble of site-rates. Mean value: 0.8; standard deviation: 0.05. The total number of samples is 1000000 and the total number of bins is 100.

The rates inferred by rate4site together with a representation of the secondary structure of the IgG-binding domain can be seen plotted in figure 15. Additionally, a projection of the site-rates onto to the structure of $GB1$ can be seen in figure 16.

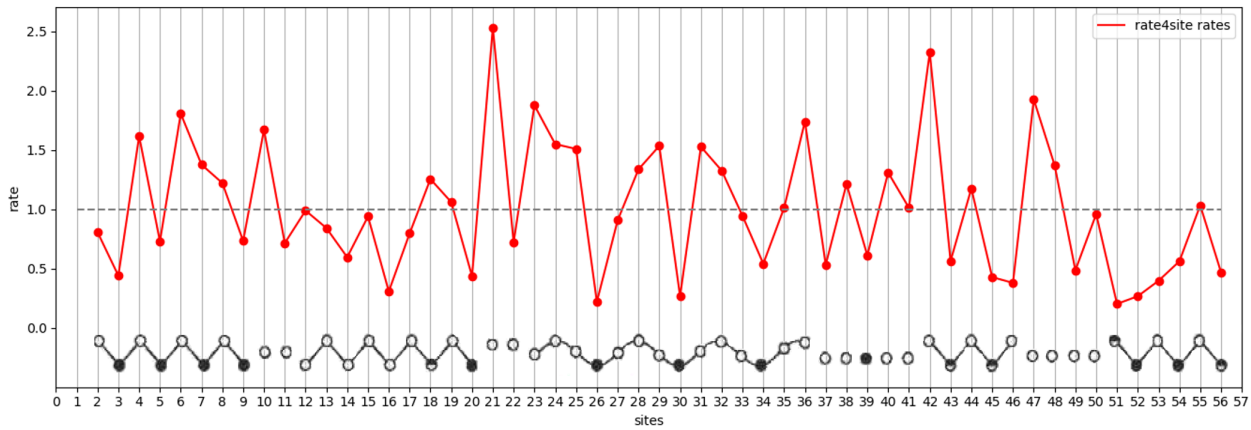

Figure 15: Plot of the site-rate inferred by rate4site for each site in the protein. A circles and lines representation of the secondary structure is included for reference. The fraction of side-chain solvent accessibilities is represented by the filling of the circles - filled circles, < 0.1; half-filled circles, > 0.1 and < 0.4; open circles, > 0.4 [19]. The circles are connected by straight or curved lines to delineate β strands and the α helix, respectively.

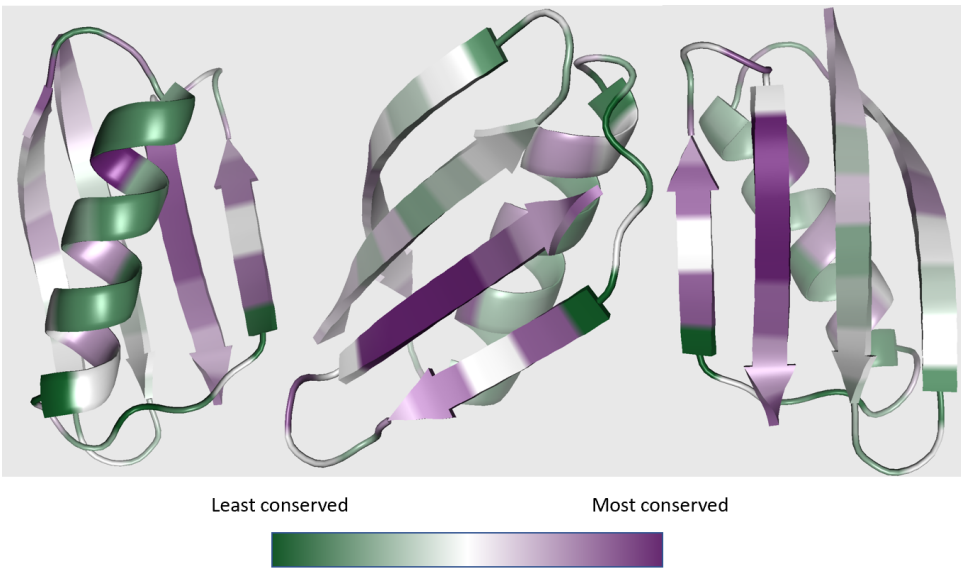

Figure 16: Projection of the site-rates on the structure of $GB1$ obtained through the Consurf web service.

5.2 Experimentally derived rates versus Rosetta derived rates

To illustrate the influence the choice of ΔG_{native} has, it is useful to inspect figure 17. On one hand, if it is chosen such that $\overline{|\Delta\Delta G_{Rosetta}|}/|\Delta G_{native}|$ (the overline indicates average) is too large then the negative impact on fitness will be overestimated, resulting in predicting that the mutation rate of most sites will be too low, on the other hand if it is chosen such that the above proportion is too small the impact of each mutation will be

considered marginal for folding stability, therefore the mutation rates will be overestimated. Nonetheless, the model is not very sensitive to changes around the optimum value.

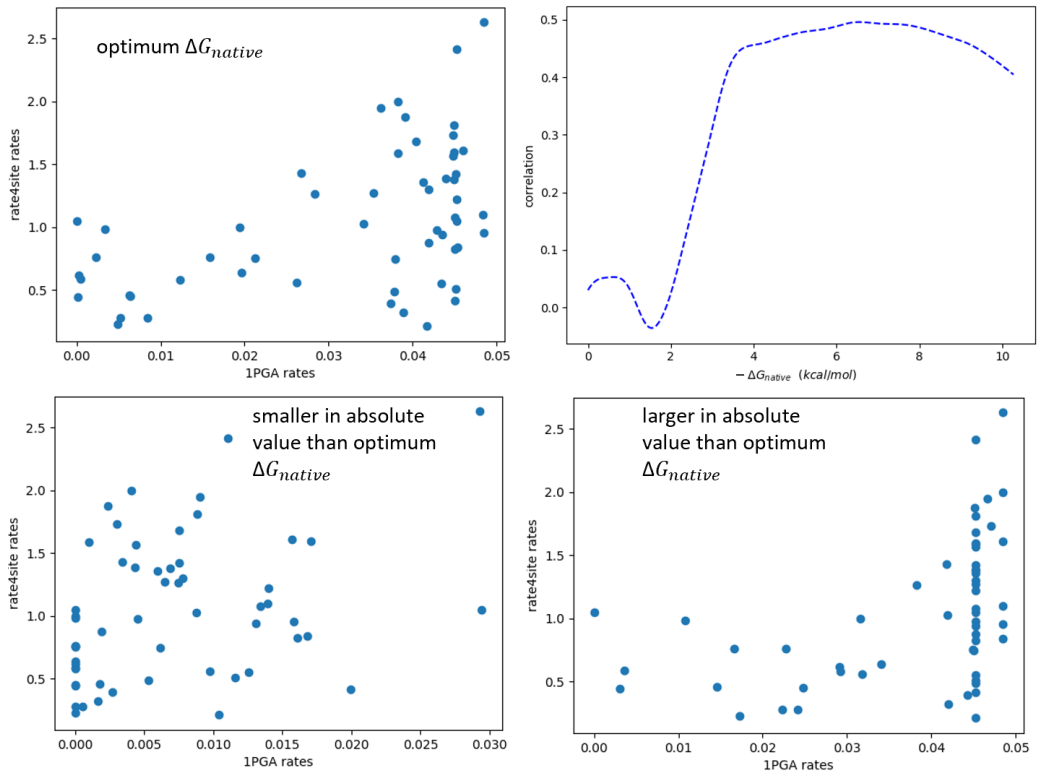

Figure 17: Upper left corner: rate4site rates against 1PGA rates for the optimum ΔG_{native} (approximately -6.55 $kcal/mol$). Bottom left corner: rate4site rates against 1PGA rates for a lower than optimum ΔG_{native} (-3 $kcal/mol$). Bottom right corner: rate4site rates against 1PGA rates for a higher than optimum ΔG_{native} (-9 $kcal/mol$). Upper right corner: Correlation of 1PGA rates and rate4site rates as a function of the model ΔG_{native}. 1PGA rates refer to STED rates calculated using the crystal structure 1PGA in the PDB.

In figure 18 the optimized ΔG_{native} for different structures can be seen. Its average value is -5.90 $kcal/mol$. It approximates very well to the value Norn et al. found to maximize the congruence between the STED and LG rate matrices: -6.25 $kcal/mol$ [18]. Despite the good concordance between the two values and lack of sensitivity of the model to ΔG_{native} near the optimum, the individually optimized native folding stabilities were chosen because they allow to compare the best case scenario of Rosetta rates with the experimental rates (STED rates calculated from the experimental $\Delta\Delta G$ data set). However, there is one aspect that strikes the attention: $\overline{\Delta G_{native}}$ is almost 1.5 times larger than the experimentally measured folding stability for $GB1$. This means that either $\Delta\Delta G_{Rosetta}$ values are overestimating $\Delta\Delta G_{experimental}$ or the correlation with rate4site simply increases when stability constraints are softened.

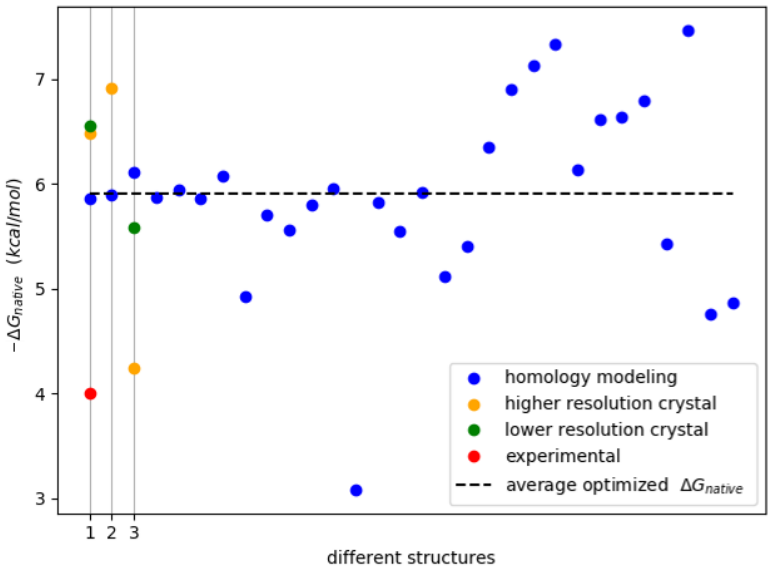

Figure 18: Absolute value of the optimum ΔG_{native} for each different structure. Structures 1, 2 and 3 refer to proteins $GB1$, $GB2$ and $GB3$, respectively.

Inspection of figure 19 shows that there is indeed an overestimation of $\Delta\Delta G$ if the conversion value calculated by Norn et al. (1.947 $REU/kcal/mol$) is used. In the left plot there are several outliers, mostly constituted by proline mutations (depicted in orange). If the outliers are removed the conversion factor decreases from 5.08 to 2.53, which is closer to 1.947.

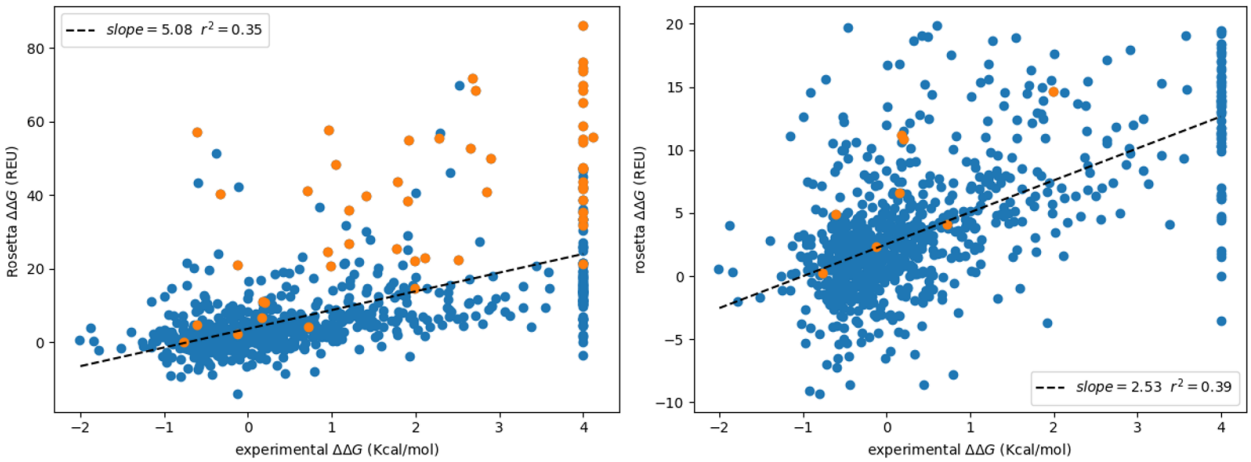

Figure 19: Left plot: $\Delta\Delta G_{Rosetta}$ plotted against $\Delta\Delta G_{experimental}$ and respective least squares linear regression. Right plot: $\Delta\Delta G_{Rosetta}$ plotted against $\Delta\Delta G_{experimental}$ without outliers that have $\Delta\Delta G_{Rosetta} > 20$ REU or $\Delta\Delta G_{Rosetta} < -10$ REU and respective least squares linear regression. Mutations to proline are depicted in orange. All the other mutations are depicted in blue.

It is important to highlight that for STED it is the ability to predict $\Delta\Delta G$ for each of the protein's sites that matters. 28 out of the 55 sites plotted in figure 20 have a correlation above 0.6, which means that at least for those sites $\Delta\Delta G_{Rosetta}$ explains at least 36% to 92% of the variation.

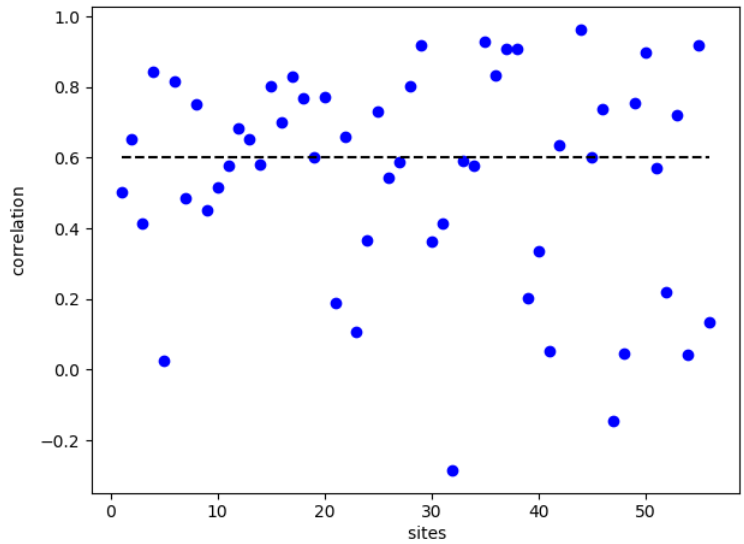

Figure 20: Plot of the correlation between $\Delta\Delta G_{experimental}$ and $\Delta\Delta G_{Rosetta}$ for each of the different sites of the protein. $\Delta\Delta G_{Rosetta}$ in this plot was predicted from structure 1PGA in the PDB. The dashed lines help visualize which plotted points have a correlation above or below 0.6.

Within the group of sequences 1 to 19 in the MSA (figure 12) the only site of sequence 1 that does not have a homologue in all of the group's sequences is site 1. However, from sequence 20 onwards there are more gaps, consequently the only sites of sequence 1 that have a homologue in all of the those sequences are the sites from site 17 to site 53. Therefore, to analyse how the different STED site-rates fare against each other in terms of correlation with the rate4site site-rates, the correlation was calculated in two distinct manners: correlation between rate4site and STED rates considering sites 2 to 56 was calculated for the 19 first sequences (left side plot of figure 21) - type 1; correlation considering sites 17 to 53 for all sequences (right side plot of figure 21) was calculated - type 2.

$$\rho = \begin{cases} \dfrac{\sum_{site=2}^{56}\left(r_{site}^{rate4site}-\overline{r_{site}^{rate4site}}\right)\left(r_{site}^{STED}-\overline{r_{site}^{STED}}\right)}{\sqrt{\sum_{site=2}^{56}\left(r_{site}^{rate4site}-\overline{r_{site}^{rate4site}}\right)^2}\sqrt{\sum_{site=2}^{56}\left(r_{site}^{STED}-\overline{r_{site}^{STED}}\right)^2}}, & \text{if type 1} \\ \\ \dfrac{\sum_{site=17}^{53}\left(r_{site}^{rate4site}-\overline{r_{site}^{rate4site}}\right)\left(r_{site}^{STED}-\overline{r_{site}^{STED}}\right)}{\sqrt{\sum_{site=17}^{53}\left(r_{site}^{rate4site}-\overline{r_{site}^{rate4site}}\right)^2}\sqrt{\sum_{site=17}^{53}\left(r_{site}^{STED}-\overline{r_{site}^{STED}}\right)^2}}, & \text{if type 2} \end{cases} \tag{5.22}$$

$r_{site}^{rate4site}$ is the rate inferred by rate4site for a given site and r_{site}^{STED} is the rate calculated by the STED model for the same site.

In type 1 only 5 out of 24 different structural backgrounds have a higher correlation with rate4site than the correlation experimental rates have with rate4site. On average the experimental rates are 3% better at recapitulating the rate4site rates than the Rosetta rates. For type 2 the experimental rates exhibit the highest correlation and they recapitulate the rate4site rates on average 17% better. Type 1 and type 2 correlations exhibit very different behaviours. This is not unexpected for two reasons: type 2 includes a larger number of different structures than type 1; the residues considered in type 2 are in their majority only evolutionarily constrained by folding stability. The later is a very important difference because it decreases the influence of a confounding aspect that is very present in type 1 correlations - it is impossible to know if the correlation in type 1 is low because of the existence of many competing constraints on evolution or because the $\Delta\Delta G_{Rosetta}$ overestimates or underestimates the true stability variation upon a mutation. There is evidence to believe that using $\Delta\Delta G_{experimental}$ to calculate STED rates meaningfully increases the correlation with rate4site rates. However, due to the confounder described above type 1 underestimates and type 2 overestimates the improvement observed when using $\Delta\Delta G_{experimental}$.

Inspection of figure 22 shows that the correlation decreases when the ensemble is used instead of the inferred rates. This was already expected given that the inferred site-rates only explain 64% of the variation in the ensemble. The right side plot of figure 22 gives statistical confidence to the statement that calculating STED site-rates using $\Delta\Delta G_{experimental}$ is better.

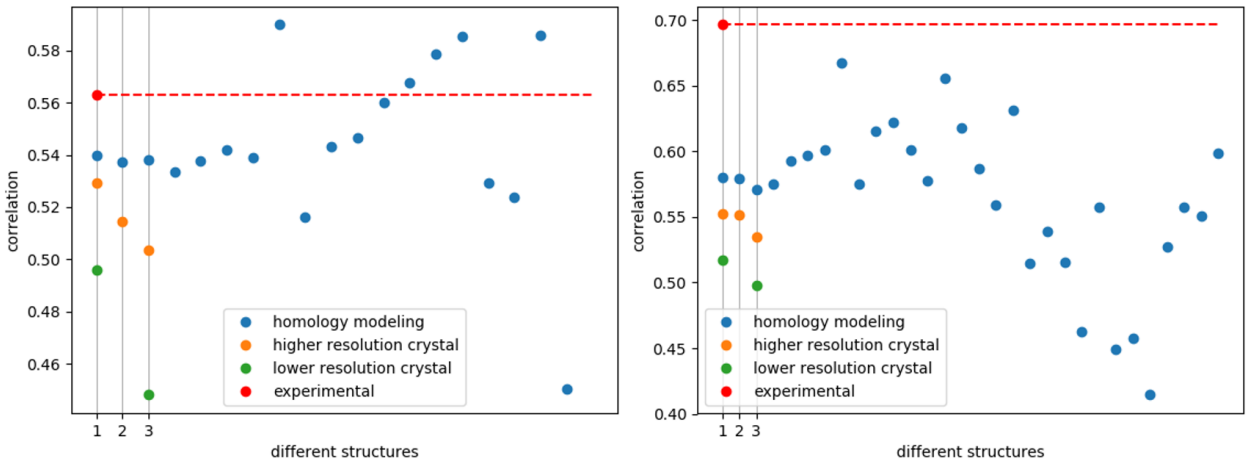

Figure 21: Left plot: type 1 correlation between the STED rates and rate4site inferred rates. Right plot: the same as the left plot but for type 2 correlation. Structures 1, 2 and 3 refer to proteins $GB1$, $GB2$ and $GB3$, respectively. The dashed lines help visualize which plotted points have a correlation above or below the correlation originated by using experimental stability data.

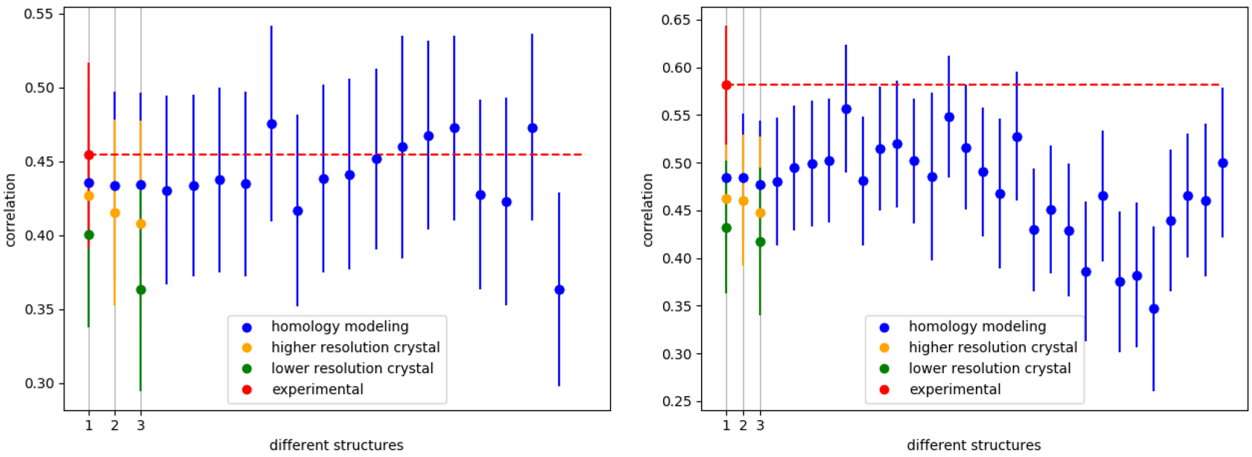

Figure 22: Left plot: average value of type 1 correlation between the STED rates and the elements of the site-rates ensemble sampled from the discrete probability distributions given by rate4site. Right plot: the same as the left plot but for type 2 correlation. The error bars for each site encompass one standard deviation from the mean. The dashed lines help visualize which plotted points have a correlation above or below the correlation originated by using experimental stability data.

To further the hypothesis above the correlation with rate4site rates was calculated similarly to type 1, but without all the residues that play a role in the binding to the Fab fragment. And a second time, but without all the sites involved in binding to IgG. The residues that are responsible for binding to the Fc fragment are under a double constraint:

functionality and stability. Thus, a type 1 correlation only without considering the residues involved in binding to the Fc fragment would not be interesting, as the influence of their removal would get overshadowed by the influence of the Fab binding residues. Figures 23 and 24 show that in both situations the experimental rates fare better than the other rates. Therefore, when controling for the confounding effect, the experimental rates can be said to confidently recapitulate the empirical rates better than the Rosetta rates.

The experimental rates in the case of the left side plot of figures 23 and 24 recapitulate the rate4site rates on average 10% better than the Rosetta rates. This coincides with the value already estimated by Norn et al. [18]. The idiosyncrasies of $GB1$ made the analysis of the impact of using experimental $\Delta\Delta G$ less trivial, as 12 out of 56 of its residues are involved in binding to IgG. This is more than double the average amount of functional residues present in a protein.

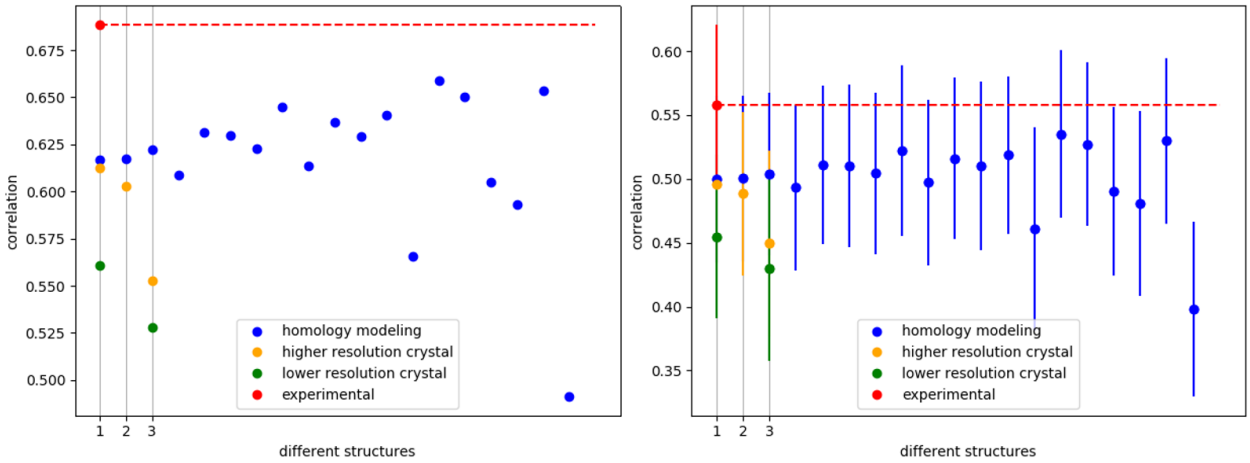

Figure 23: Left plot: correlation of the experimental rates and STED rates with rate4site rates calculated without considering residues 1 and the residues involved in binding to the Fab fragment. Right plot: average value of the correlation between the STED rates and the elements of the site-rates ensemble sampled from the discrete probability distributions given by rate4site without considering the same residues as in the left plot. The error bars for each site encompass one standard deviation from the mean. The dashed lines help visualize which plotted points have a correlation above or below the correlation originated by using experimental stability data.

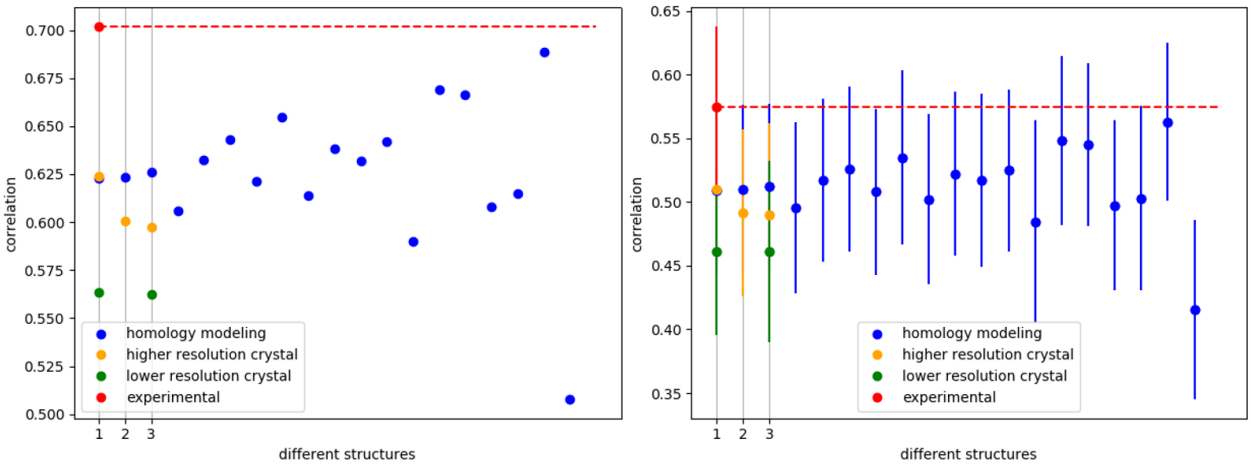

Figure 24: Left plot: correlation of the experimental rates and STED rates with rate4site rates calculated without considering residues 1, the residues involved in binding to the Fab fragment, and to the Fc fragment. Right plot: average value of the correlation between the STED rates and the elements of the site-rates ensemble sampled from the discrete probability distributions given by rate4site without considering the same residues as in the left plot. The error bars for each site encompass one standard deviation from the mean. The dashed lines help visualize which plotted points have a correlation above or below the correlation originated by using experimental stability data.

Surprisingly, the rates derived from structures that were modeled through homology modeling fare better than their X-ray crystal structure counterparts. This might be due to the protocol that was used to relax the X-ray structures to the Rosetta energy function, but this hypothesis has not been tested. Nonetheless, the rates derived from structures that were modeled through higher resolution X-ray structures fared better than their lower resolution counterparts.

5.3 Averaged rates versus Single structure rates

In Figure 25 the weighted average of the rates for each site of $GB1$ over all the homologues can be seen plotted together with the rate4site site-rates and the STED site-rates for each of the structures in the phylogeny. There is a big spread of different rate values that are assigned to each site. This was expected because STED rates are conditioned on different structural backgrounds.

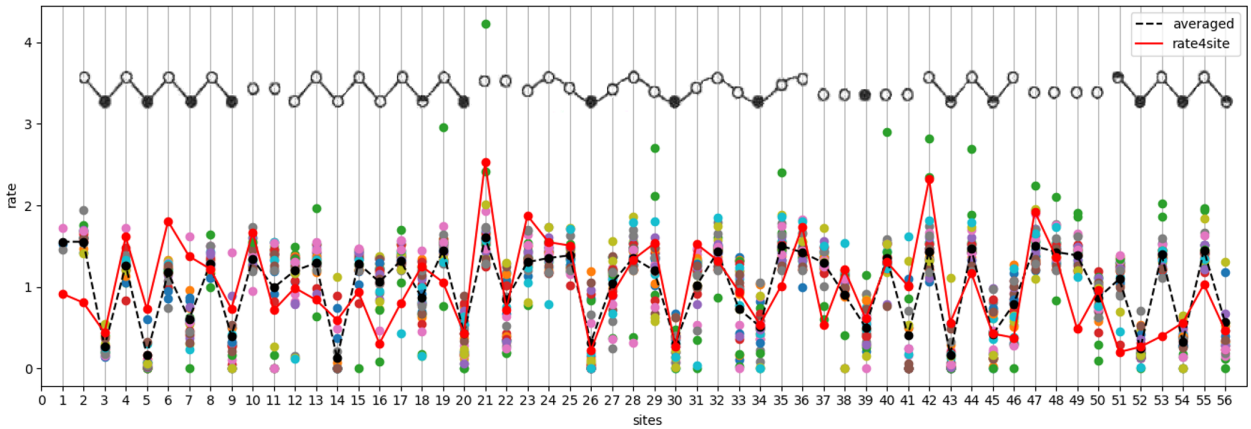

Figure 25: Plot of the site-rates for each of the sites in the protein. The weighted average of the rates for each site of $GB1$ over all the homologues can be seen plotted together with the rate4site site-rates and the STED site-rates for each of the structures in the phylogeny. Each different color of the plotted filled circles represent one of the different structures in the phylogeny. A circles and lines representation of the secondary structure is included for reference. The fraction of side-chain solvent accessibilities is represented by the filling of the circles - filled circles, < 0.1; half-filled circles, > 0.1 and < 0.4; open circles, > 0.4 [19]. The circles are connected by straight or curved lines to delineate β strands and the α helix, respectively.

To be able to conclude if the averaged rates recapitulate the rate4site rates better than the STED rates, type 1 and type 2 methods of calculating correlations were applied (5.22). As type 1 only uses the first 19 sequences in the MSA the phylogenetic tree had to be pruned before being used as input to the BranchManager program. The trees that were used as input are depicted in figure 26.

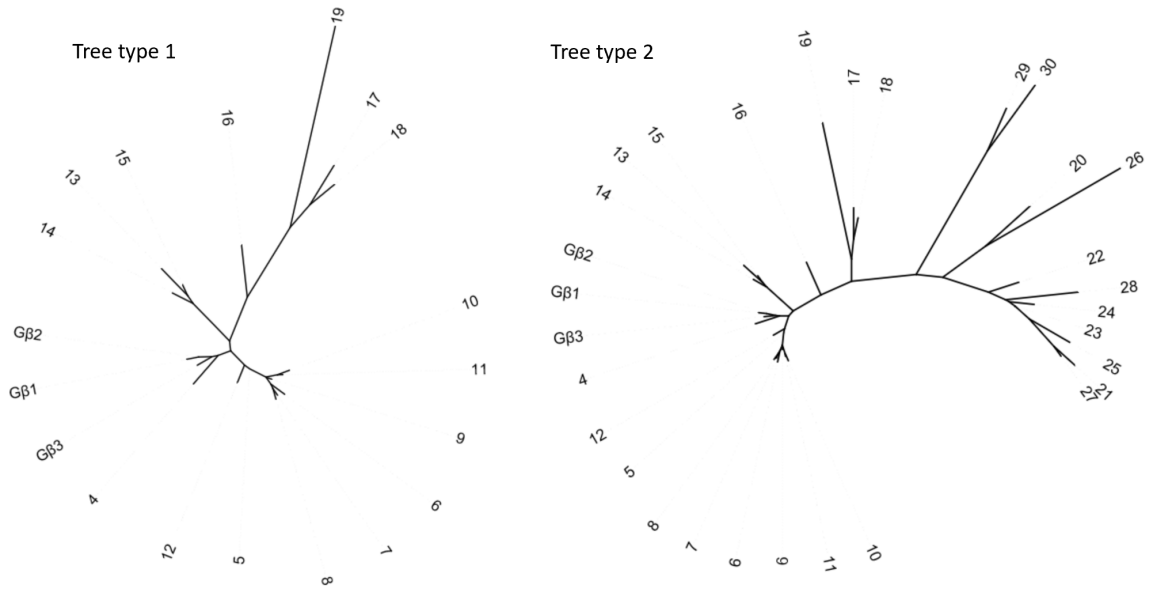

Figure 26: Image to the left: pruned phylogenetic tree used for type 1 correlation calculation. The pruning was done in such a way that the topology and branch lengths of the leaf nodes that are kept are maintained. Image to the right: original phylogenetic tree. It was used for the calculation of type 2 calculation. The phylogenies were designed using the online tool iTOL [14].

In the left side of figure 27 it can be seen that only 3 single-structure derived rates achieve a higher correlation than the averaged rates. To rule out the hypothesis that those rates are the sole responsibles for the high correlation of the averaged rates, a second averaging was done without those structures. As it can be seen in the right side of figure 27 the averaged rates have the highest correlation. The same was done for type 2 correlation calculation, as can be seen from figure 28. The recapitulation of the rate4site rates is on average 5% better explained by the averaged rates in type 1 and 9% better explained in type 2.

From the averaging of rates a correlation improvement emerges. This happens because the averaged rates are a better measure of the phenomena rate4site tries to measure. However, the stochastic process that BranchManager relies on to infer the average is an imperfect reflection of the way proteins evolve. The used averaged rates are an estimate of the rates that a protein at the phylogenetic center of mass would have if the proteins were to evolve according to a Brownian motion process on the tree. However, the process that better approximates the real stochastic nature of evolution is a continuous-time Markov process. Hence, perhaps the average rates would better recapitulate the rate4site rates if the assumption used by BranchManager was that the proteins evolved according to the transition rates of the LG rate matrix.

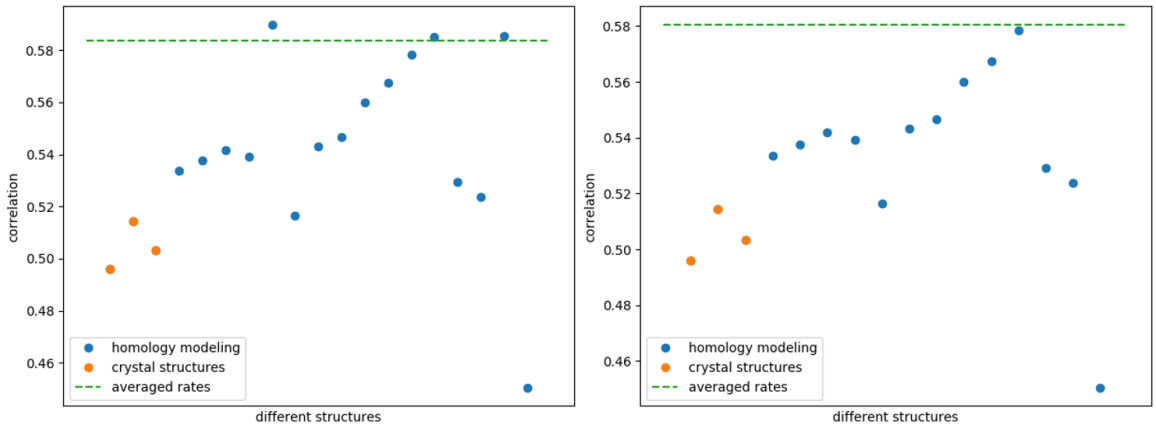

Figure 27: Left plot: type 1 correlation calculation of the STED rates and their weighted averaged. Right plot: the same as in the left plot, but the averaging was done with the leaf nodes 8, 15 and 18 pruned out of the phylogenetic tree.

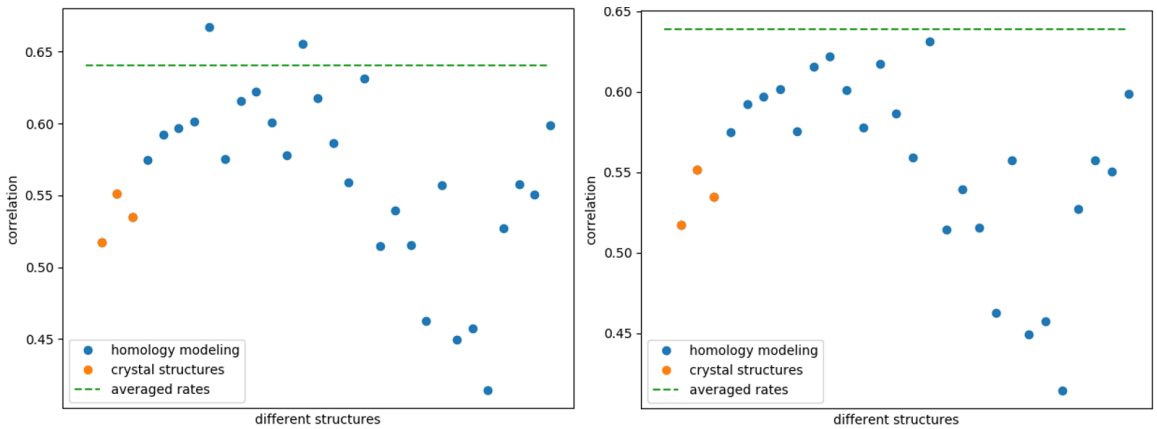

Figure 28: Left plot: type 2 correlation calculation of the STED rates and their weighted average. Right plot: the same as in the left plot, but the averaging was done with the leaf nodes 8 and 14 pruned out of the phylogenetic tree.

An accurate description of how biological sequences evolve is a fundamental prerequisite for phylogenetic analysis [33]. However, current methods do not model heterotachy - that the evolutionary rate of a given site is not always constant throughout evolution. It seems plausible that inference of multiple sequence alignments, phylogenies and site-rates could be improved by the use of STED derived transition rate matrices. In such an application, the likelihood of inferred data could be calculated by simulating the evolutionary trajectories on the phylogenetic tree driven by STED transition rate matrices and conditioned on the leaf nodes. As the states of the process are discrete, sample paths would be piecewise constant rather than continuous. As a result, sample paths could be summarized by the proportion of time spent in each state. The average proportion of time spent in each state would be computed by taking the expectation over all sample paths. The evolutionary transition

rates would not be stationary throughout the process, but would change dependent on the sequence state the process would be at each time - the STED model would calculate the transition rates for each new sequence state.

5.4 STED and functional site prediction

Phylogenetic inference of site-rates cannot distinguish between the different mechanisms that cause sequence conservation. Norn et al. introduced a novel metric, $r_{site}^{ratio} = r_{site}^{STED}/r_{site}^{rate4site}$, which discriminates stability from function for each site. It is based on the assumption that for sites which r_{STED} is high and $r_{rate4site}$ is low, a functional fitness constrain should exist. Under the hypothesis that the sites of $GB1$ involved in binding to Fc and Fab fragments are under a functional fitness constraint, r_{ratio} signal for those sites is expected to be strong.

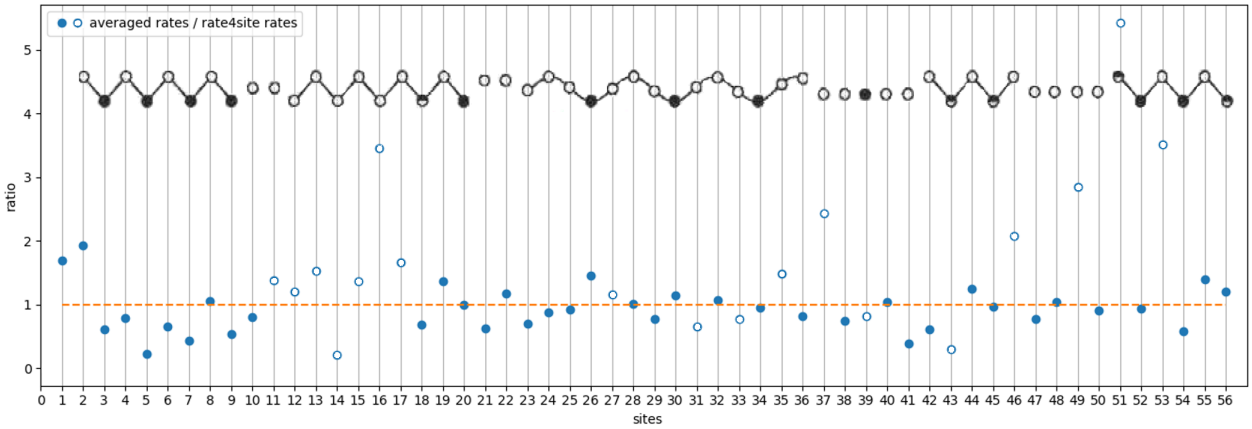

Figure 29: Plot of the ratio between the weighted average of the rates for each site of $GB1$ over all the homologues and rate4site site-rates for each site of the protein. The filled blue circles indicate the sites that are not directly involved in function and the unfilled blue circles indicates the sites that are. A circles and lines representation of the secondary structure is included for reference. The fraction of side-chain solvent accessibilities is represented by the filling of the circles - filled circles, < 0.1; half-filled circles, > 0.1 and < 0.4; open circles, > 0.4 [19]. The circles are connected by straight or curved lines to delineate β strands and the α helix, respectively.

26 out of 56 ratios are larger than one. However, only 13 out of those 26 have a functional justification. That proportion drops to 13 out of 22 if the two sites at the N-terminal and the other two sites and the C-terminal are excluded - it is justified to do it because $GB1$ is part of protein G and so those sites are poorly modeled. Nevertheless, this still means that almost half the sites with $r_{ratio} > 1$ have no functional justification. It should be added that the impact of using a higher threshold value to better control the false positives was not tested in this thesis work. Norn et al. [18] showed that at a false positive rate of 1%,

r_{ratio} predicts 26% of all functional sites while the rate4site inferred rates only predict 12%. Moreover, there are false negatives as well, 5 out of the 30 ratios that are smaller than one are functional sites. The application of STED rates together with inferred rates to detect functional sites does not have to stop here. It is plausible that an intelligent treatment of the data they provide together with other structural information might provide less ambiguous insight to whether a site is functionally important or not.

5.4.1 Fab fragment

Sites 11 to 17 of the second β-strand are involved in the binding to the Fab fragment. Among them only site 14 has a r_{ratio} below one. The binding between the two β-strands can be divided in three sections: first, sites 11 and 12 - the ratio for 11 is slightly higher than the ratio for 12, which might be explained by the fact that site 11 side-chain engages in hydrogen bonds, while site 12 does not present hydrogen bonds with the Fab fragment at all; second, from sites 13 to 15 there is a pattern of hydrogen bonds typical of antiparallel β-sheets which only involves the main-chain atoms, only the main-chain atoms of 13 and 15 form bonds with the Fab fragment, site 14 forms bonds with the first β-strand of $GB1$ - which explains why its r_{ratio} is below one; third, sites 16 and 17 both form hydrogen bonds involving the side-chain atoms. For the smaller region of contact between the C-terminal end of the α-helix and the first β-strand of the $CH1$ domain of the Fab fragment site 33 has a r_{ratio} below one and site 37 shows a high r_{ratio} signal.

5.4.2 Fc fragment

The protrusion on $GB1$ - site 27 - has a weak r_{ratio} signal (only slightly above one). Residue 31 holds residue 27 in place, its r_{ratio} indicates that its conservation is lower than expected, however it is still an important residue. It happens that it can be replaced by amino acid R without compromising the binding to Fc [19], hence the empirical rate is not as low as expected - it can be seen in the multiple sequence alignment (figure 12) that for site 31 most sequences have residue K or R.

The available literature only clarified the role of one of the residues that neighbour the hole in $GB1$ surface (residue 43). According to inspection of 1FCC PDB entry the protusion on Fc surface seems to be fixed to the hole in the B1 domain with the help of side-chain hydrogen bonds of site 43 and 35 and with the help one main chain hydrogen bond of site 39. Site 43 has a weak r_{ratio} signal. This is because it is a tryptophan, the largest of amino acids, therefore mutating it to another amino acid generally comes with a high entropy cost. Hence, for site 43 there is a stability constraint and a functionality constraint acting at the same time. Site 35 has r_{ratio} larger than one.

5.4.3 Second β-hairpin

A different reason for r_{ratio} to be higher than one has to do with the existence of residues that do not play an important role neither in the stability of the folded state nor in its

binding functionality, but play a critical one in the folding process. This seems to be the case for residues 46, 49 and 51 which form a hydrogen bond network in the second β-hairpin and for residue 53 in the last β-strand. According to multiscale modeling simulations the first folding event is the formation of the second β-hairpin, which is strongly stabilized by residues 43, 45 and 52 (all of which can be seen to be highly conserved in figure 25) [10].

6 Conclusion

The results presented in this thesis confirm that using experimental $\Delta\Delta G$ values to calculate STED rates increases their correlation with the inferred rates. The degree of correlation increment depends on the proportion of functional sites. When the correlation is calculated without the functional sites, the experimental rates can recapitulate the inferred rates 10% better than the STED rates that are computed using predicted $\Delta\Delta G$ values. This coincides with the estimation done by Norn et al. [18].

The weighted average of STED rates successfully demonstrated that an accurate description of how biological sequences evolve is fundamental if the STED model is to be used for phylogenetic analysis. In the best scenario, the averaged rates recapitulated the rate4site rates 9% better than the STED rates that were solely modeled on a single leaf node of the phylogeny. This is a motivating result; as the weights that were assigned to each leaf node were calculated based on the assumption that evolutionary trajectories can be modeled as Brownian processes, there is still room for improvement. Hence, it would be an exciting endeavour to explore new ways of applying the STED model to phylogenetic analysis.

The model developed by Norn et al. (the STED model) is a work in progress. This thesis work furthered that both accounting for the propagation of $\Delta\Delta G$ errors and the conditioning on a single structure for the calculation of rates as the only sources of errors does not appear to explain the low correlation that STED site-rates with rate4site rates (on average $\rho^2 = 0.28$ [18]). The experimental rates had on their best case cenario (when controlling for functional sites) a correlation with the rate4site rates of about $\rho^2 = 0.49$. And, the averaged rates had on their best case cenario (type 2 method for calculating correlation) a correlation with the rate4site rates of about $\rho^2 = 0.41$. Therefore, the model should continue to be scrutinized as to fully understand why it does not accurately recapitulate rate4site rates.

The novel metric Norn et al. introduced to detect functional sites is a promising correlate. Only 2 out of 13 of the functional residues of $GB1$ would not be identified by their method. However, there were 13 false positives. It seems likely that the false positives can be narrowed down to a smaller number if more correlates with functionality are added.

7 Acknowledgements

I want to thank my supervisor Ingemar André for his insight and patience while guiding me through the convoluted field of protein evolution.

My family, but specially my parents and my brother for their boundless love. Obrigado por estarem sempre comigo e tornarem este sonho possível.

The proofreaders of this thesis: Killer and Måns. Adoro-te Killer Maria. O que seria de mim sem ti?

Daniel, du är det bästa som någonsin hänt mig. Tack för att du alltid står vid min sida. Utan dig hade det här aldrig varit möjligt.

References

[1] F. H. Crick. On protein synthesis. In *Symp Soc Exp Biol*, volume 12, page 8, 1958.

[2] J. P. Derrick and D. B. Wigley. Crystal structure of a streptococcal protein g domain bound to an fab fragment. *Nature*, 359(6397):752, 1992.

[3] J. P. Derrick and D. B. Wigley. The third igg-binding domain from streptococcal protein g: an analysis by x-ray crystallography of the structure alone and in a complex with fab. *Journal of molecular biology*, 243(5):906–918, 1994.

[4] J. Echave, S. J. Spielman, and C. O. Wilke. Causes of evolutionary rate variation among protein sites. *Nature Reviews Genetics*, 17(2):109, 2016.

[5] T. R. Gregory. Understanding evolutionary trees. *Evolution: Education and Outreach*, 1(2):121, 2008.

[6] J. Huerta-Cepas, F. Serra, and P. Bork. Ete 3: reconstruction, analysis, and visualization of phylogenomic data. *Molecular biology and evolution*, 33(6):1635–1638, 2016.

[7] Y. Iwasa. Free fitness that always increases in evolution. *Journal of Theoretical Biology*, 135(3):265–281, 1988.

[8] S. Kanje. Development of a fab binding protein domain. Master's thesis, KTH, School of Biotechnology (BIO), 2011.

[9] M. Kimura. On the probability of fixation of mutant genes in a population. *Genetics*, 47(6):713, 1962.

[10] S. Kmiecik and A. Kolinski. Folding pathway of the b1 domain of protein g explored by multiscale modeling. *Biophysical journal*, 94(3):726–736, 2008.

[11] S. Q. Le and O. Gascuel. An improved general amino acid replacement matrix. *Molecular biology and evolution*, 25(7):1307–1320, 2008.

[12] A. Leaver-Fay et al. An object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology*, pages 545–574.

[13] M. S. Lee and A. Palci. Morphological phylogenetics in the genomic age. *Current Biology*, 25(19):R922–R929, 2015.

[14] I. Letunic and P. Bork. Interactive tree of life (itol) v4: recent updates and new developments. *Nucleic acids research*, 2019.

[15] I. Mayrose, D. Graur, N. Ben-Tal, and T. Pupko. Comparison of site-specific rate-inference methods for protein sequences: empirical bayesian methods are superior. *Molecular biology and evolution*, 21(9):1781–1791, 2004.

[16] A. Nisthal, C. Y. Wang, M. L. Ary, and S. L. Mayo. Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. *Proceedings of the National Academy of Sciences*, 116(33):16367–16377, 2019.

[17] L. G. Nivón, R. Moretti, and D. Baker. A pareto-optimal refinement method for protein design scaffolds. *PloS one*, 8(4):e59004, 2013.

[18] C. Norn. *An evolutionary basis for protein design and structure prediction.* PhD thesis, Lund University, 2019.

[19] C. A. Olson, N. C. Wu, and R. Sun. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Current Biology*, 24(22):2643–2651, 2014.

[20] H. Park, P. Bradley, P. Greisen Jr, Y. Liu, V. K. Mulligan, D. E. Kim, D. Baker, and F. DiMaio. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *Journal of chemical theory and computation*, 12(12):6201–6212, 2016.

[21] R. Phillips, J. Theriot, J. Kondev, and H. Garcia. *Physical biology of the cell.* Garland Science, 2012.

[22] A. E. Sauer-Eriksson, G. J. Kleywegt, M. Uhlén, and T. A. Jones. Crystal structure of the c2 fragment of streptococcal protein g in complex with the fc domain of human igg. *Structure*, 3(3):265–278, 1995.

[23] L. Schrodinger. The pymol molecular graphics system. *Version*, 1(5):0, 2010.

[24] G. Sella and A. E. Hirsh. The application of statistical physics to evolutionary biology. *Proceedings of the National Academy of Sciences*, 102(27):9541–9546, 2005.

[25] D. J. Sloan and H. W. Hellinga. Dissection of the protein g b1 domain binding site for human igg fc fragment. *Protein science*, 8(8):1643–1648, 1999.

[26] J. M. Smith. Natural selection and the concept of a protein space. *Nature*, 225(5232):563, 1970.

[27] Y. Song, F. DiMaio, R. Y.-R. Wang, D. Kim, C. Miles, T. Brunette, J. Thompson, and D. Baker. High-resolution comparative modeling with rosettacm. *Structure*, 21(10):1735–1742, 2013.

[28] E. A. Stone and A. Sidow. Constructing a meaningful evolutionary average at the phylogenetic center of mass. *BMC bioinformatics*, 8(1):222, 2007.

[29] M. Tashiro and G. T. Montelione. Structures of bacterial immunoglobulin-binding domains and their complexes with immunoglobulins. *Current opinion in structural biology*, 5(4):471–481, 1995.

[30] T. Uzzell and K. W. Corbin. Fitting discrete probability distributions to evolutionary events. *Science*, 172(3988):1089–1096, 1971.

[31] P. D. Williams, D. D. Pollock, B. P. Blackburne, and R. A. Goldstein. Assessing the accuracy of ancestral protein reconstruction methods. *PLoS computational biology*, 2(6):e69, 2006.

[32] G. Yachdav, S. Wilzbach, B. Rauscher, R. Sheridan, I. Sillitoe, J. Procter, S. E. Lewis, B. Rost, and T. Goldberg. Msaviewer: interactive javascript visualization of multiple sequence alignments. *Bioinformatics*, 32(22):3501–3503, 2016.

[33] Z. Yang. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution*, 11(9):367–372, 1996.

[34] Z. Yang, R. Nielsen, and M. Hasegawa. Models of amino acid substitution and applications to mitochondrial protein evolution. *Molecular biology and evolution*, 15(12):1600–1611, 1998.