Accurate protein stability predictions from homology models

Valanciute, Audrone; Nygaard, Lasse; Zschach, Henrike; Maglegaard Jepsen, Michael; Lindorff-Larsen, Kresten; Stein, Amelie

Published in:
Computational and Structural Biotechnology Journal

DOI:
10.1016/j.csbj.2022.11.048

Publication date:
2023

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY-NC-ND

Citation for published version (APA):
Accurate protein stability predictions from homology models

Audrone Valanciute a,1, Lasse Nygaard a,1, Henrike Zschach b, Michael Maglegaard Jepsen a, Kresten Lindorff-Larsen a,*, Amelie Stein b,*

a Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Copenhagen, Denmark
b Section for Computational and RNA Biology, Department of Biology, University of Copenhagen, Copenhagen, Denmark

1. Introduction

The vast majority of functions in a cell are carried out by proteins. These biomolecules typically need to fold into a stable tertiary structure in order to be functional. Experimental structure determination can provide us with high-resolution details of the arrangement of atoms in the folded protein, however it does not directly reveal the forces holding them together. These have historically been derived from mutagenesis experiments and measurements of their energetic consequences [1–3]. From these studies we know that many changes are detrimental for protein stability [4], although stabilizing mutations have also been observed [5]. In the absence of experimental data, the stability of variant proteins is commonly determined by computational means with so-called assessment of the change in thermodynamic stability upon mutation (ΔAG). These calculations have many practical applications, from protein engineering [6] to identification of disease-causing mutations [7,8]. Hence, a number of methods for ΔAG prediction have been developed. Among the most widely used ones are FoldX [9–11] and Rosetta [12–14]

These methods require a structure of the wild-type protein as input, which in combination with molecular force fields or energy functions in principle allows for the generation of accurate and high-resolution models of the mutant protein. At the same time, though, this requirement represents an important limiting factor for their application. Recent advances in sequencing technology particularly highlight the gap between known sequences and known structures; many new protein families have been found in large-scale sequencing experiments for which no structural information is available, and there are vast differences in coverage between species [15]. For proteins where the structure of another family member is known, homology modeling can be employed to provide a computationally-derived structure model. Briefly, this approach relies on using the solved structure of a homologue of the protein of interest as a template for the backbone of the target protein.

Calculating changes in protein stability (ΔAG) has been shown to be central for predicting the consequences of single amino acid substitutions in protein engineering as well as interpretation of genomic variants for disease risk. Structure-based calculations are considered most accurate, however the tools used to calculate ΔAGs have been developed on experimentally resolved structures. Extending those calculations to homology models based on related proteins would greatly extend their applicability as large parts of e.g. the human proteome are not structurally resolved. In this study we aim to investigate the accuracy of ΔAG values predicted on homology models compared to crystal structures. Specifically, we identified four proteins with a large number of experimentally tested mutations [7,8]. Hence, a number of methods for ΔAG calculations to test. We find that ΔAG-values predicted from homology models compare equally well to experimental ΔAGs as those predicted on experimentally established crystal structures, as long as the sequence identity of the model template to the target protein is at least 40%. In particular, the Rosetta cartesian_ddg protocol is robust against the small perturbations in the structure which homology modeling introduces. In an independent assessment, we observe a similar trend when using ΔAGs to categorize variants as low or wild-type-like abundance. Overall, our results show that stability calculations performed on homology models can substitute for those on crystal structures with acceptable accuracy as long as the model is built on a template with sequence identity of at least 40% to the target protein.
sequence, and using loop modeling and refinement of intramolecular interactions by either physics-based or empirical force fields to generate a high-resolution model [16]. For template-target pairs with sequence identity >70%, models are typically fairly accurate (1 Å–2 Å RMSD), whereas pairs below 30% sequence identity are in the so-called “twilight zone” and often have different 3D structures and distinct functions [16]. However, calculations of the change in thermodynamic stability upon mutation require accurate backbone as well as side chain placement, and the latter may require a higher identity between template and target of around 50% [17]. Model quality depends on the template-target similarity to experimentally resolved structures in terms of torsion angles, core packing and other high-resolution properties [18,19].

If ΔΔG calculations could be performed reliably on computationally generated models of protein structure, this would greatly extend the scope of their applicability, and open up e.g. possibilities of structure-based re-engineering of proteins without an experimentally determined structure. Only about 15% of the human proteome is covered by experimental structures, but with inclusion of homology models, coverage could be doubled or even quadrupled, depending on whether templates with down to 50% or 20% sequence identity are chosen, respectively [20]. Both FoldX [11] and Rosetta [7,16], tools widely used in the community, however, were developed and benchmarked for use with experimentally determined structures so the question remains how applicable they are to computationally-derived structures.

On one hand, stability effects have been shown to be largely consistent when introducing a particular mutation into different homologs in the same family [21]. On the other hand, ΔΔG predictions have been shown to suffer from challenges in reversibility, i.e., the predicted ΔΔG of a V → A mutation is not the inverse of the corresponding A → V mutation, revealing gaps in capturing the underlying biophysical forces and a substantial dependency on the structural input model [22,23].

Here we set out to assess whether ΔΔGs can be accurately predicted using homology models, and to what extent the accuracy is dependent on the sequence identity between template and target. To assess the dependence on target-template similarity systematically, we chose four proteins for which a substantial number of biophysical ΔΔG measurements as well as template protein structures across a range of sequence identities are available. From those four proteins we established a dataset of 344 mutations. For an independent comparison, we also calculate full-saturation ΔΔGs for ~3,500 single amino acid variants of a protein assayed for cellular abundance [24], again using experimental structures, close and distant homology models as starting points. Overall, we find that all tools work well on experimentally resolved structures, and that Rosetta cartesian_ddg performance appears robust on homology models with at least 40% sequence identity.

2. Results and discussion

Our goal was to determine the accuracy of predicted stability changes in proteins calculated using homology models instead of experimentally determined structures. To this end, we compare a set of experimentally-determined stability changes (ΔΔG) to predicted ΔΔG values calculated on either experimentally determined structures or homology models.

For the experimental data we used a curated database of 2971 experimentally-determined stability changes (ΔΔG) [25] which is largely based on the Protherm database [2]. From this set we selected four different proteins based on the following criteria: First, each protein should have a large number of experimentally determined single point ΔG values. Second, a high-resolution crystal structure of the wild-type protein should be available. Finally, we required that for each protein a number of homologues with varying degrees of sequence identity existed, which themselves had experimentally-resolved structures. This latter criterion is central to our objective as it enables us to build homology models of the four proteins using templates with varying degrees of sequence identity. With these criteria we selected human lysozyme (PDB ID 1LZ1), the barley chymotrypsin inhibitor CI2 (2CI2), E. coli RNaseH (2RN2) and sperm whale myoglobin (1BVC) which together have 344 experimentally-determined ΔΔG-values.

For the computationally-derived data, homology models with increasing divergence from the wild-type protein were needed. We divided the range from 100% to 25% sequence identity into 15% intervals and obtained the following five bins: bin1: 100%, bin2: 85%, bin3: 70%, bin4: 55%, bin5: 40%. For each protein and each bin we searched the PDB for homologues and picked the one with a sequence identity closest to the middle of the bin. For example, in the case of sperm whale myoglobin, bin2 was populated using an elephant myoglobin (PDB ID: 1EMY) with a sequence identity of 81% to the sperm whale protein. In this way we populated the full matrix of structures/sequences covering all four proteins and all five bins, apart from bin3 for CI2 for which we could not find a suitable template [Suppl. Table S1].

We used Modeller [26] to build homology models for all four proteins using the 19 different templates. Modeller is widely used and, importantly, allows for direct selection of the template, which ensures that the resulting model’s quality was not influenced by the availability of a crystal structure of the wild-type protein. We show the modeling results for sperm whale myoglobin in Fig. 1 (top row) where we overlay the models to the structure of the wild-type protein; corresponding plots for the remaining proteins can be found in the supporting online material [Suppl. Figs. S1-S3]. As expected, we find the agreement between the crystal structure of the wild-type protein and the template-based models decreases as the sequence of the template diverges from that of the wild-type protein.

We then calculated ΔΔG values with three different structure-based methods: FoldX, Rosetta_ddg_monomer and Rosetta kartesian_ddg. All three methods are widely used for ΔΔG-predictions but, to our knowledge, have never been benchmarked for use with template-based models. For each method and protein we ran ΔΔG calculations on each sequence identity bin plus the wild-type crystal structures and thus predicted the 344 ΔΔG-values corresponding to the protein variants we have experimental data for.

We compared predicted and experimental ΔΔG-values using the Pearson correlation coefficient (r) and the mean absolute error (MAE). For MAEs to be comparable across methods, we first calculated a scaling factor to bring the predicted ΔΔG values to kcal/mol (see Methods). The results are illustrated in Fig. 1 with sperm whale myoglobin as an example. First, we find that all three methods give a comparable agreement to the experimental data when we use a high-resolution crystal structure of the wild-type protein as input for stability predictions (row 1). As expected, we also see that the predicted ΔΔG-values correlate less with experimental data as the sequence identity of the template used to build the homology model decreases (following columns). For example, when we use FoldX on either the crystal structure of sperm whale myoglobin (wild-type structure) or the homology model built on wild boar myoglobin (bin1: 86% sequence identity), we obtain r ~ 0.7 between the predicted and experimental ΔΔG-values. This drops to r ~ 0.6 for bin2 (Asian elephant myoglobin: 81%) and bin3 (Sea turtle myoglobin: 65%) and further for bins 4 and 5. A similar trend was observed for all three methods and all four proteins
(Fig. 1 and S1–S3), though with apparent differences in robustness towards using homology models.

As a baseline for minimal expected $\Delta G$ prediction performance, we developed two simple null models, which are designed to be independent of protein structure and the specific sequence. Both null models are based on the curated experimental $\Delta G$ database mentioned above [25]. The first null model simply assigns the average over all of these experimental measurements, $\Delta G = 1.15$ kcal/mol, to any mutation. Since most stability changes in the set of 344 mutations that we examine are of this order of magnitude, this model achieves a MAE = 1.1 kcal/mol to the experimental $\Delta G$-values. However, the Pearson correlation is very poor ($r = 0$) since this model assigns the same value to all mutations. To introduce more specificity, we designed a second null model in which we again use the average $\Delta G$ over the entire database, but split up by wild-type and mutant residue, so that we obtain 380 averages corresponding to the 380 possible pairs of the 20 different wild-type amino acids with their 19 possible substitutions (Fig. 2). Before applying this null model to our set of mutants, we in practice make two modifications. First, since the 344 $\Delta G$-values we use for validation are a subset of the 2971 values in the full dataset and went into the averages, we leave out all the $\Delta G$-values pertaining to a protein when we apply that null model to that protein. This is particularly important for the many fields in the null model's 20x19 matrix that include only very few mutations. Second, we use the global average of $\Delta G = 1.15$ kcal/mol for any mutations that are not present in the original dataset after leaving out the mutations in the protein that we apply the model to, such as histidine to methionine. When we use this modified null model to predict the stability outcome of the 344 mutations we achieve MAE = 1.1 kcal/mol and $r = 0.42$.

With the (second) null model in hand, we then compared the results from all three structure-based methods and aggregated the results over all four proteins (Fig. 3). As observed for myoglobin (Fig. 1), we find that all three methods perform comparably when using the crystal structure of the wild-type proteins as input. Looking at the correlation coefficients, however, we also find that the methods appear to display different levels of sensitivity to the input structure. Both FoldX and Rosetta $\text{ddg\_monomer}$ appear more sensitive to the input structure than $\text{cartesian\_ddg}$ since the correlation of $\Delta G$-values from homology models with the experimental values is substantially lower than when using the wild-type crystal structures. In particular, when using the homology models from bins 4 and 5 (which in practice corresponds to sequence identities <45 %) both FoldX and Rosetta $\text{ddg\_monomer}$ give results that are at most only slightly better than the (second) structure-independent null model. This decrease in accuracy is also reflected in an increase in the MAE, though we focus less on this value as it is less sensitive to the variation in structure and is also well-captured by the null models. In contrast to that, the results from Rosetta $\text{cartesian\_ddg}$ appear substantially more robust against changes in the input structure. Both the correlation coefficient and MAE vary little across bins 1 to 4 and are comparable to those obtained when using the wild-type crystal structures as input. To summarize, in practice we find that at least for the mutation dataset we used, Rosetta $\text{cartesian\_ddg}$ predicts stability changes with similar accuracy on homology models as when using crystal structures of the actual proteins so long as the template sequence identity is >40 %.

The results detailed above depend on the availability of the crystal structures and homology models of the four proteins, as well as the experimental $\Delta G$ measurements. As an alternative approach, and to check for robustness, we also calculated correlation coefficients using the $\Delta G$ values calculated from the crystal structures as the reference instead of the experimental ones (Fig. S4). Per definition, the crystal structures have $r = 1$, and hence the correlation coefficients from this analysis are higher than those using the experimental data as reference. This observation is expected given remaining biases in stability prediction methods, and in line with a recent study analyzing $\Delta G$ predictions on different homology models, also observing a drop <40 % sequence identity, as well as dependency on the model quality [27]. The results we obtain here confirm the previous results, namely (i) a general decrease in the accuracy of stability predictions as homology models are based on increasingly distant templates and (ii) Rosetta $\text{cartesian\_ddg}$ appearing less sensitive than the other two methods.

The key result obtained above is that the Rosetta $\text{cartesian\_ddg}$ approach to predict stability changes appears to be relatively insensitive to minor differences in the input structure so that it gives fairly accurate results also when applied to template-based models from Modeller. The analyses are, however, based on a biased set of mutations with many mutations being hydrophobic deletion mutations and a high proportion of mutations to alanine or glycine (similar to Fig. 2). For a more unbiased set of single substitutions, we turned to recent data from a type of multiplexed assay of variant effects (MAVE) experiment termed VAMP-seq, performed on a protein called TPMT [24]. Briefly explained, a MAVE involves creating a library of variants and selecting for a property of interest such as cell growth or fluorescence of a reporter protein [28]. In a VAMP-seq experiment the library of protein variants are fused to green fluorescent protein (GFP) and expressed in cell culture. The brightness of each cell's GFP fluorescence will be a function of the abundance of the protein variant [24]. By sorting cells based on their fluorescence, and using DNA sequencing to determine the frequency of every variant before and after sorting, a single experiment provides abundance data for thousands of variants. Since destabilized proteins are degraded by the protein quality control apparatus, a variant protein that has high abundance is expected to have little changes or only stabilizing changes to its thermodynamic stability ($\Delta G$). Vice versa, a variant protein with low cellular abundance may have large changes in its thermodynamic stability compared to the wild-type protein, which is supported by good correlation of melting temperatures for several variants (though proteins can also have low abundance or be degraded for other reasons). We use this relation to evaluate whether $\Delta G$ values calculated on either the crystal structure or the homology models behave in accordance with
the experimental abundance data. In principle the VAMP-seq results are semi-quantitative, but we only use the information whether variants were low- or high-abundance.

In addition to the crystal structure of wild-type TPMT (PDB ID 2BZG) we built homology models of TPMT using templates with 80% identity (2GB4) and 32% identity (3LCC and 1PJZ).
Unfortunately no templates with identities between those values were available. Since the VAMP-seq experiment does not directly probe thermodynamic stability, we instead used a 'receiver operating characteristic' (ROC) curve analysis to relate probe thermodynamic stability, we instead used a 'receiver operating characteristic' (ROC) curve analysis to relate predicted stability values and abundance classification [29]. Specifically, we vary a cut-off for the predicted ΔΔG values, and assess how well it divides the variant proteins into low and high-abundance in terms of true positive rate (sensitivity) and false positive rate (specificity). The area under the curve (AUC) statistic assesses how well the predicted ΔΔG values balance sensitivity and specificity across different thresholds. A ROC curve along the diagonal corresponds to a random prediction method and has AUC = 0.5, whereas a method that can perfectly separate low- and high-abundance variants would have AUC = 1. Our results show that ΔΔG values from Rosetta cartesian_ddg achieve a rather high accuracy both when using the wild-type TPMT structure (AUC = 0.87) and the model based on 2 GB4 (AUC = 0.88) (Fig. 4). Also, as expected we find decreased prediction accuracies when using the two models based on templates with 32% sequence identity (AUC = 0.78 and 0.77, for 3lcc and 1pjz respectively). Thus, the results obtained using these VAMP-seq data confirm the previous observation from the comparison to experimental ΔΔG values, namely that Rosetta cartesian_ddg provides relatively accurate predictions, and that there is little deterioration when the input structure is based on a close homologue. As the VAMP-seq library was designed to target every possible single amino acid substitution, it is much more balanced than data from ProTherm, which is dominated by alanine scanning results [7].

3. Conclusions

The ability to predict changes in thermodynamic stability upon single amino acid substitutions in proteins is important in a wide range of practical and scientific problems. FoldX and the two Rosetta applications ddg_monomer and cartesian_ddg are widely used for that. These methods were parameterized and benchmarked using high-resolution crystal structures, and are recommended to be used with such structures as input. Many proteins of interest however do not have crystal structures available and here template-based (homology) modeling provides an appealing alternative. We asked how accurate the three methods are when used on homology models, and how the results depend on the sequence identity of the template used in model building. While the template-dependency of homology modeling is well-studied, it has to our knowledge not been investigated how sensitive computational stability prediction methods are to the structural noise induced by such models.

We therefore devised and applied a strategy to test how well these three methods work outside their normally recommended application range. Specifically, we chose four proteins for which we have extensive experimental stability measurements, crystal structures of the wild-type protein and templates of varying sequence identities to build homology models. Our analyses show that all three methods perform comparably well when applied to the crystal structures of the four proteins (Fig. 3). Further, the Rosetta cartesian_ddg method appears to be relatively robust to structural noise and errors introduced by homology modeling, at least for models based on templates with sequence identities >40% to the target protein (Fig. 3). This observation was corroborated by analysis of the influence of local sequence similarity, and further by using a ROC analysis to relate predicted stability measurements to high-throughput abundance measurements from VAMP-seq on the protein TPMT, though in this case we could only apply the analysis to a smaller number of templates (Fig. 4).

We further devised two simple null models for stability predictions that effectively require a look-up in a table of average experimental values. While we do not recommend using this model to perform predictions in practical applications, we suggest that it is a useful baseline for ΔΔG benchmark evaluation in general.

Overall, our results suggest that it should be possible to extend the range of applicability of structure-based protein stability predictions to homology models and thus to a much larger number of sequences. For example, about 15% of the human proteome is covered by experimental structures, however the SWISS-MODEL repository contains structural information for ~30% of the human proteome if one includes homology models with template sequence identities >40% [20].

In our work, we only considered single amino acid substitutions, though the ability to model multiple substitutions is important for protein engineering purposes. Previous results suggest that it is difficult to predict non-additive effects of multi-mutants even when using crystal structures as input to for example FoldX [30,31], whereas more computationally demanding free energy calculations achieve better accuracy [31]. Thus, a first step is to achieve higher accuracy for multi-mutants using the faster computational methods, and then examine the accuracy using homology models as input.

We conclude by suggesting that these kinds of analyses should be applied more broadly, and to a wider range of data and proteins, and point out a recent assessment of performance of multiple methods on AlphaFold2 [32] models, compared to measurements from MAVE experiments [33]. Also, it is our hope that future developments of stability prediction methods, as well as other methods to analyze protein structures, will be developed with applications on homology models or predicted structures in mind. Thus, we suggest it should be possible to develop methods that are even more robust to structural noise and possibly supported by structure-independent methods. The first step in such a development would be to better understand the structural origins of why the predictions become less accurate as the sequence identity...
diverges, and why the different methods have different sensitivities to the structural accuracy. If we can increase robustness down to templates with about 30% sequence identity we should be able to analyze about half of the human proteome with these methods.

4. Methods

4.1. Homology modeling

Homology models for human lysozyme, chymotrypsin inhibitor CI2, E. coli RNAseH and sperm whale myoglobin were generated using Modeller's automodel class [26]. The target sequences for the four proteins were obtained from Uniprot [34]. Templates were obtained from a search of the pdball database (obtained from https://salllab.org/modeller/supplemental.html). The search results were binned according to sequence identity in 15% intervals and we selected the template closest to the middle of the interval for each respective bin. For the model building, all settings and parameters were left at their default value. 20 models were generated for high sequence identity bins (>55%) and 1000 models for low sequence identity bins (<55%). Three models were selected by the lowest DOPE scores in each sequence identity bin.

4.2. ΔΔG calculations: FoldX

FoldX version 3.0b4 was used with default and standard parameters. Prior to ddg prediction, the structures were run through the RepairPDB function, as recommended in the FoldX documentation. The FoldX ddg predictions were performed using the BuildModel command of FoldX [10]. For crystal structures, 5 iterations were used as recommended. For homology models, 2 BuildModel iterations were carried out on each of the top 3 models, and the average across those 6 ΔΔGs is reported.

4.3. ΔΔG calculations: Rosetta ddg_monomer

We used Rosetta from October 2018 with git SHA1 ce9cb339991a7e8ca1bc44e6b2b2d8b0a3d557f8. We ran ddg_monomer with the global repacking option (protocol 16, [13]) and the talaris2013 energy function, which has been shown to improve ΔΔG calculation performance [25,35]. We carried out initial constrained minimization as recommended [13]. Average ddGs of 50 iterations are reported for the crystal structure. For homology models, the top 3 models were chosen, and the average across those 6 ΔΔGs is reported.

For each method a scaling factor was established to describe the relationship between the energy units reported by the method and kcal/mol. This was done by comparing the values predicted on the high quality crystal structures of the wild-type proteins against the experimental values and then forcing a linear fit through 0. The slope of this fit is the method’s scaling factor which was applied to all predictions made by the method. We obtained the following scaling factors:

<table>
<thead>
<tr>
<th>Method</th>
<th>Scaling Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rosetta cartesian_ddg</td>
<td>2.469</td>
</tr>
</tbody>
</table>

4.5. Scaling to kcal/mol

For the null models we used the average experimental ΔΔG value of the given amino acid substitution across all the proteins in ProTherm, excluding the values from the protein itself.

4.7. ROC curves

ROC curves were drawn and AUCs calculated with the R package ROCr, directly using ΔΔG values obtained from running Rosetta cartesian_ddg on the homology models as predictor.
variables and low and high abundances classes from [24] as the two underlying true classes.

**CRediT authorship contribution statement**

**Audrone Valanciute:** Investigation, Methodology, Software, Visualization, Writing – review & editing. 
**Lasse Nygaard:** Data curation, Investigation, Methodology, Software, Visualization, Writing – original draft. 
**Henrike Zschach:** Investigation, Validation, Visualization. 
**Michael Maglegaard Jepsen:** Investigation, Methodology. 
**Kresten Lindorf-Larsen:** Conceptualization, Funding acquisition, Resources, Supervision, Writing – original draft. 
**Writing – review & editing. Amelie Stein:** Conceptualization, Funding acquisition, Resources, Supervision, Writing – original draft. 
**Writing – review & editing.**

**Declaration of Competing Interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Acknowledgements**

We thank members of the Lindorf-Larsen and Stein labs for helpful discussions and comments. This work is supported by a Novo Nordisk Foundation, Denmark, Challenge Grant (PRISM, NN18OC0033950 to K.L.-L., A.S.) and the Lundbeck Foundation, Denmark (R272-2017-4528 to A.S.).

**Appendix A. Supplementary data**

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.11.048.

**References**
