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 3, 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

A R T I C L E   I N F O
Article history:
Received 15 August 2022
Received in revised form 22 November 2022
Accepted 23 November 2022
Available online 25 November 2022

Keywords:
Protein stability
ΔAG
Protein variant
Mutation

A B S T R A C T
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 ΔGs and templates for homology modeling across a broad range of sequence identities, and selected three methods for ΔAG calculations to test. We find that ΔAG-values predicted from homology models compare equally well to experimental ΔGs 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.

© 2022 Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

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...
2. Results and discussion

Our goal was to assess the accuracy of predicted stability changes in proteins calculated using homology models instead of experimentally determined structures. For this, we compare a set of experimentally determined structures to predict the AAV values calculated for each homology model. We used an established database [13] containing 100,000 protein structures and their corresponding AAV values. For the experimental set, we used a curated database of 2971 proteins, covering the human proteome.

First, each protein should have a large number of experimentally determined single-point AAC values, reflecting the high-resolution structure of the wild-type protein. Second, a high-resolution structure of the wild-type protein is required for the homology model. Finally, we required that for each protein, a number of homologues were available with varying degrees of sequence identity. We selected four different proteins based on their sequence identity with the wild-type protein: human lysozyme, whale myoglobin, and human cytochrome C.

We chose four different proteins based on the following criteria: (1) they have experimentally determined structures, (2) they cover the range of sequence identities, and (3) they are central to our objective as they enable us to build homology models with varying degrees of sequence identity. We selected human lysozyme (PDB ID 1LZ1), whale myoglobin (PDB ID 1EMY), and human cytochrome C (PDB ID 1QMM). The sequence identity of the template used to build the homology model is determined by the sequence identity of the template used to build the homology model.

The wild-type protein structure is compared to the homology model, and the resulting model is used to predict the AAV values. We used three different structure-based methods: FoldX, Rosetta, and a homology model-building method, to calculate the AAV values. For MAEs, we used the Pearson correlation coefficient to compare the predicted AAV values. The results are illustrated in Fig. 1 with stem-and-leaf plots. The results are consistent with the known correlation coefficients, as expected.

We used a homology model-building method, which we call the "machine-learning" method, to predict AAV values. For MAEs, we used linear regression to compare the predicted AAV values. The results are illustrated in Fig. 2 with scatter plots. The results are consistent with the known correlation coefficients, as expected.
Target structure: Sperm whale myoglobin (PDB 1BVC)
As a baseline for minimal expected ΔΔG 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 ΔΔG database mentioned above [25]. The first null model simply assigns the average over all of these experimental measurements, ΔΔG = 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 ΔΔG-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 ΔΔG 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 mutations, we in practice make two modifications. First, since the 344 ΔΔG-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 ΔΔG-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 ΔΔG = 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 cartesian_ddg appear more sensitive to the input structure than cartesian_ddg since the correlation of ΔΔG-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 cartesian_ddg 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 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 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 ΔΔG measurements. As an alternative approach, and to check for robustness, we also calculated correlation coefficients using the ΔΔG 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 ΔΔG 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 cartesian_ddg appearing less sensitive than the other two methods.

The key result obtained above is that the Rosetta 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 TPM2 [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 (ΔΔG). 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 ΔΔG values calculated on either the crystal structure or the homology models behave in accordance with

(Fig. 1 and S1–S3), though with apparent differences in robustness towards using homology models.

...
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).

**Fig. 2.** Mutations in the curated ProTherm dataset split by wild-type and target amino acid. These data are used in the null models and are a superset of the test data. Each entry lists the number of mutations of that type, and the color indicates the average change in stability (in kcal/mol).

**Fig. 3.** Correlation of ddGs calculated on the original crystal structure as well as homology models with decreasing sequence identity to the experimentally determined values. Turquoise, Rosetta cartesian_ddg, blue, Rosetta ddg_monomer, violet, FoldX. The dashed line indicates the average performance on the original crystal structures. Note that bin2 contains two models based on NMR structures, see also Suppl. Fig. S4. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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.
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 ΔΔG 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 each (3LCC, blue, AUC = 0.778; 1Pjz, cyan, AUC = 0.768). Performance of a random classifier is indicated by the diagonal line (AUC = 0.5). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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 ΔΔG 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 Cl2, 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://salilab.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 sequence 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 ΔΔG prediction, the structures were run through the RepairPDB function, as recommended in the FoldX documentation. The FoldX ΔΔG 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 ce9cb33999a17e8ca1bc44efb2b2d8b0a3d557f8. 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, 20 ΔΔG iterations were performed on each, and the overall average is reported.

Flags for constrained minimization:
- in::file::fullatom.
- ignore_unrecognized_res.
- fa_max_dis 9.0.
- ddg::harmonic_ca_tether 0.5.
- ddg::constraint_weight 1.0.
- ddg::sc_min_only false.

Flags for ddg_monomer:
- ddg::weight_file soft_rep_design.
- ddg::local_opt_only false.
- constraints::cst_file ca_dist_restraints.cst.
- fa_max_dis 9.0.
- ddg::min_cst true.
- ddg::iterations 20.
- ddg::mean false.
- ddg::min true.
- ddg::sc_min_only false.
- ddg::output_silent true.
- ddg::ramp_rejective true.
- ddg::mut_only.

4.4. ΔΔG calculations: Rosetta cartesian_ddg

We used Rosetta from October 2018 with git SHA1 ce9cb33999a17e8ca1bc44efb2b2d8b0a3d557f8. For cartesian_ddg [12,14] we first performed constrained relaxation, then carried out 3 iterations of the cartesian_ddg protocol, and reported average ΔΔGs over these. For homology models, we carried out one iteration each over the top 3 homology models, and reported the average over these.

Flags for relax:
- relax::constrain_relax_to_start_coords.
- ignore_unrecognized_res.
- missing_density_to_jump.
- ex1 – ex2.
- relax::min_type lbfgs_armijo_nonmonotone.
- flip_HNQ – no_optH false.
- relax::coord_constrain_sidechains – relax::cartesian.
- beta.
- score::weights beta_nov16_cart.

Flags for cartesian_ddg:
- fa_max_dis 9.0.
- ddg::dump_pdbds false.
- ddg::iterations 1.
- score::weights beta_nov16_cart.
- missing_density_to_jump.
- ddg::mut_only.
- ddg::bnnbrs 1.
- beta.cart.
- ex1 – ex2.
- ddg::legacy false.
- optimize_proline.

4.5. Scaling to kcal/mol

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:

- FoldX: 1/1.039.
- Rosetta cartesian_ddg: 1/2.469 - note that a factor of 1/2.9 has previously been reported [14].
- Rosetta ddg_monomer: 1/1.233.

4.6. Null models

Data was extracted from a curated version of ProTherm [25]. In addition to the published curation, we noticed two apparent self-mutations with non-zero ΔΔG, T53T in protein G and K33K in ubiquitin. We traced both back to articles reporting the ΔΔG with reference to the stability of the sequence with alanine at the position of interest [36,37] as opposed to the more common convention of reporting with respect to wild-type. We removed entries from both articles from the dataset before creating the null models.

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**