COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

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*

ARTICLE INFO

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
ΔΔG
Protein variant
Mutation

ABSTRACT

Calculating changes in protein stability (ΔΔG) 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 ΔΔGs 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 ΔΔG values predicted on homology models compared to crystal structures. Specifically, we identified four proteins with a large number of experimentally tested ΔΔGs and templates for homology modeling across a broad range of sequence identities, and selected three methods for ΔΔG calculations to test. We find that ΔΔG-values predicted from homology models compare equally well to experimental ΔΔGs 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 ΔΔGs 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.

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 (ΔΔG). These calculations have many practical applications, from protein engineering [6] to identification of disease-causing mutations [7,8]. Hence, a number of methods for ΔΔG 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

[*] Corresponding authors.
E-mail addresses: lindorff@bio.ku.dk (K. Lindorff-Larsen), amelie.stein@bio.ku.dk (A. Stein).
[1] These authors contributed equally.

https://doi.org/10.1016/j.csbj.2022.11.048
2001-0370/

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 sequence identity. A number of tools have been developed to assess the quality of an individual homology model based on 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.

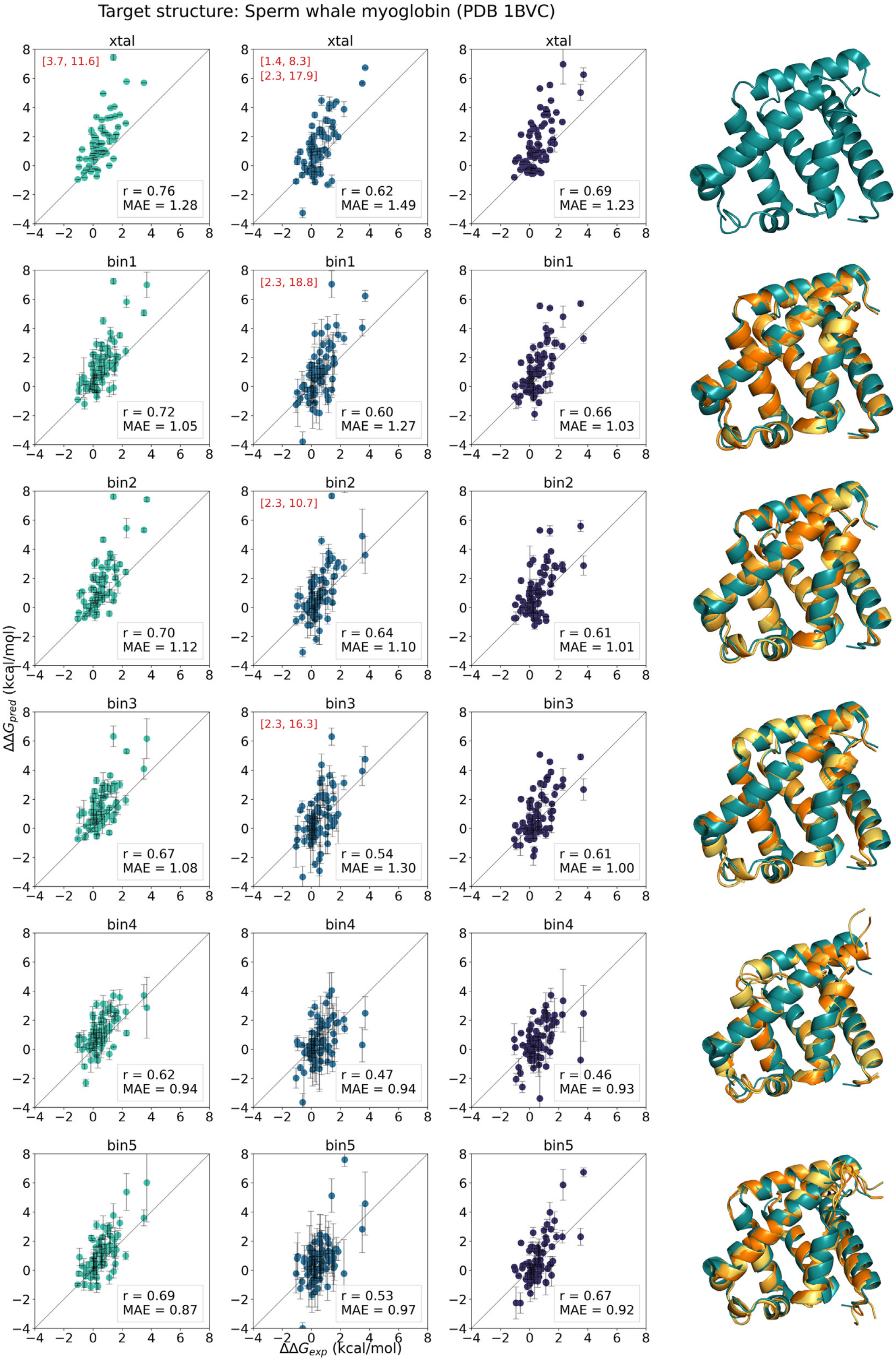
For the experimental data we used a curated database of 2971 experimentally-determined stability changes (ΔΔG) [25] which is largely based on the Protherm database [2]. From this set we selected four different proteins based on the following criteria: First, each protein should have a large number of experimentally determined single point ΔΔG 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 %,85 %]; bin2: (85 %,70 %]; bin3: (70 %,55 %]; bin4: (55 %,40 %]; bin5: (40 %,25 %]. 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, 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 `cartesian_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


Target structure: Sperm whale myoglobin (PDB 1BVC)
xtal
[3.7, 11.6]
r = 0.76
MAE = 1.28
xtal
[1.4, 8.3]
[2.3, 17.9]
r = 0.62
MAE = 1.49
xtal
r = 0.69
MAE = 1.23
bin1
r = 0.72
MAE = 1.05
bin1
[2.3, 18.8]
r = 0.60
MAE = 1.27
bin1
r = 0.66
MAE = 1.03
bin2
r = 0.70
MAE = 1.12
bin2
[2.3, 10.7]
r = 0.64
MAE = 1.10
bin2
r = 0.61
MAE = 1.01
bin3
r = 0.67
MAE = 1.08
bin3
[2.3, 16.3]
r = 0.54
MAE = 1.30
bin3
r = 0.61
MAE = 1.00
bin4
r = 0.62
MAE = 0.94
bin4
r = 0.47
MAE = 0.94
bin4
r = 0.46
MAE = 0.93
bin5
r = 0.69
MAE = 0.87
bin5
r = 0.53
MAE = 0.97
bin5
r = 0.67
MAE = 0.92
$\Delta\Delta G_{pred}$ (kcal/mol)
$\Delta\Delta G_{exp}$ (kcal/mol)
−4 −2 0 2 4 6 8

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

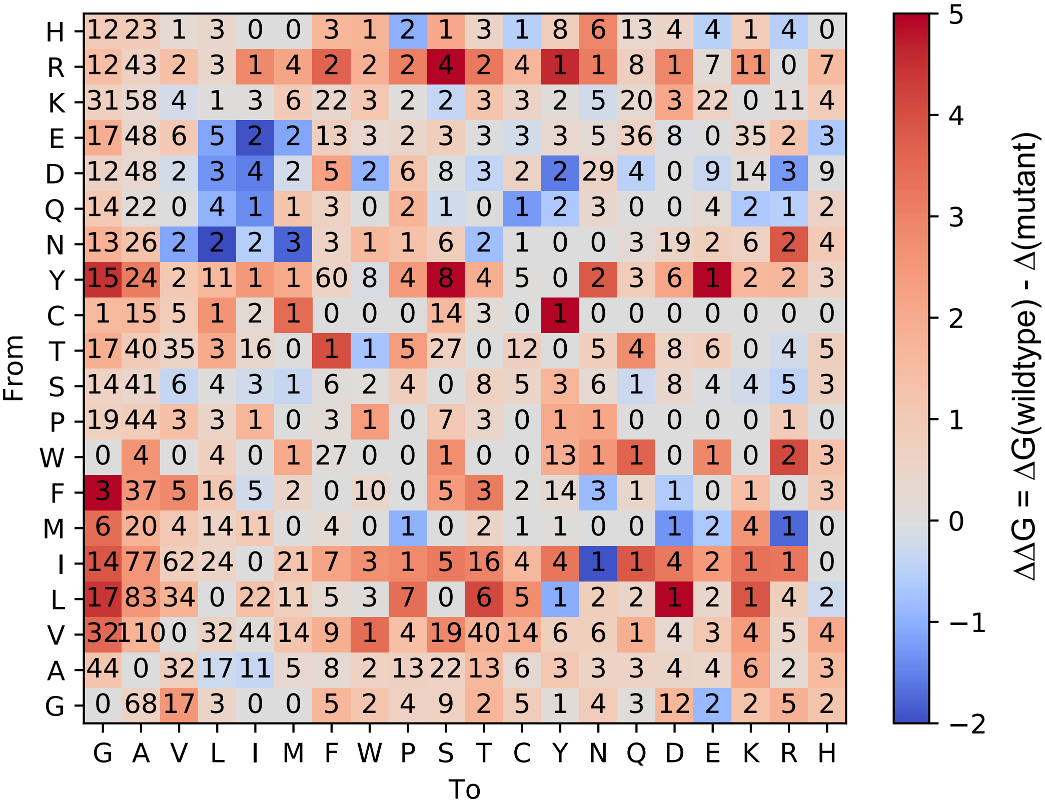
As a baseline for minimal expected ΔΔG prediction performance, we developed two simple null models, which are designed to be independent of protein structure and the specific sequence. Both null models are based on the curated experimental ΔΔG database mentioned above [25]. The first null model simply 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 `ddg_monomer` 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 `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 `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 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 (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. Comparison of computationally-derived ΔΔG values with experimental values on sperm whale myoglobin using three different ΔΔG-prediction methods and either the crystal structure or a homology model. Each field shows a scatter plot comparing predicted and experimental ΔΔG values, including the corresponding Pearson's correlation coefficient (r) and the mean absolute error (MAE; kcal/mol). If present, values outside the [-4,8] kcal/mol range on either axis are listed in red in the top left corner of each plot. First row: ΔΔG values computed on the crystal (xtal, 1BVC) structure of the wild-type protein. Following rows: ΔΔG values computed on homology models with decreasing sequence identity between target and template. First column (turquoise): ΔΔG values predicted by Rosetta `cartesian_ddg`. Second column (blue): ΔΔG values predicted by Rosetta `ddg_monomer`. Third column (violet): ΔΔG values predicted by FoldX. Last column: Superimposition of the wild-type protein crystal structure with the homology model. Agreement between the crystal structure and the model deteriorates as the sequence identity between template and target decreases (top to bottom, bin1: (100%,85%]; bin2: (85%,70%]; bin3: (70%,55%]; bin4: (55%,40%]; bin5: (40%,25%]). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



From \ To	G	A	V	L	I	M	F	W	P	S	T	C	Y	N	Q	D	E	K	R	H
H	12	23	1	3	0	0	3	1	2	1	3	1	8	6	13	4	4	1	4	0
R	12	43	2	3	1	4	2	2	2	4	2	4	1	1	8	1	7	11	0	7
K	31	58	4	1	3	6	22	3	2	2	3	3	2	5	20	3	22	0	11	4
E	17	48	6	5	2	2	13	3	2	3	3	3	3	5	36	8	0	35	2	3
D	12	48	2	3	4	2	5	2	6	8	3	2	2	29	4	0	9	14	3	9
Q	14	22	0	4	1	1	3	0	2	1	0	1	2	3	0	0	4	2	1	2
N	13	26	2	2	2	3	3	1	1	6	2	1	0	0	3	19	2	6	2	4
Y	15	24	2	11	1	1	60	8	4	8	4	5	0	2	3	6	1	2	2	3
C	1	15	5	1	2	1	0	0	0	14	3	0	1	0	0	0	0	0	0	0
T	17	40	35	3	16	0	1	1	5	27	0	12	0	5	4	8	6	0	4	5
S	14	41	6	4	3	1	6	2	4	0	8	5	3	6	1	8	4	4	5	3
P	19	44	3	3	1	0	3	1	0	7	3	0	1	1	0	0	0	0	1	0
W	0	4	0	4	0	1	27	0	0	1	0	0	13	1	1	0	1	0	2	3
F	3	37	5	16	5	2	0	10	0	5	3	2	14	3	1	1	0	1	0	3
M	6	20	4	14	11	0	4	0	1	0	2	1	1	0	0	1	2	4	1	0
I	14	77	62	24	0	21	7	3	1	5	16	4	4	1	1	4	2	1	1	0
L	17	83	34	0	22	11	5	3	7	0	6	5	1	2	2	1	2	1	4	2
V	32	110	0	32	44	14	9	1	4	19	40	14	6	6	1	4	3	4	5	4
A	44	0	32	17	11	5	8	2	13	22	13	6	3	3	3	4	4	6	2	3
G	0	68	17	3	0	0	5	2	4	9	2	5	1	4	3	12	2	2	5	2

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).

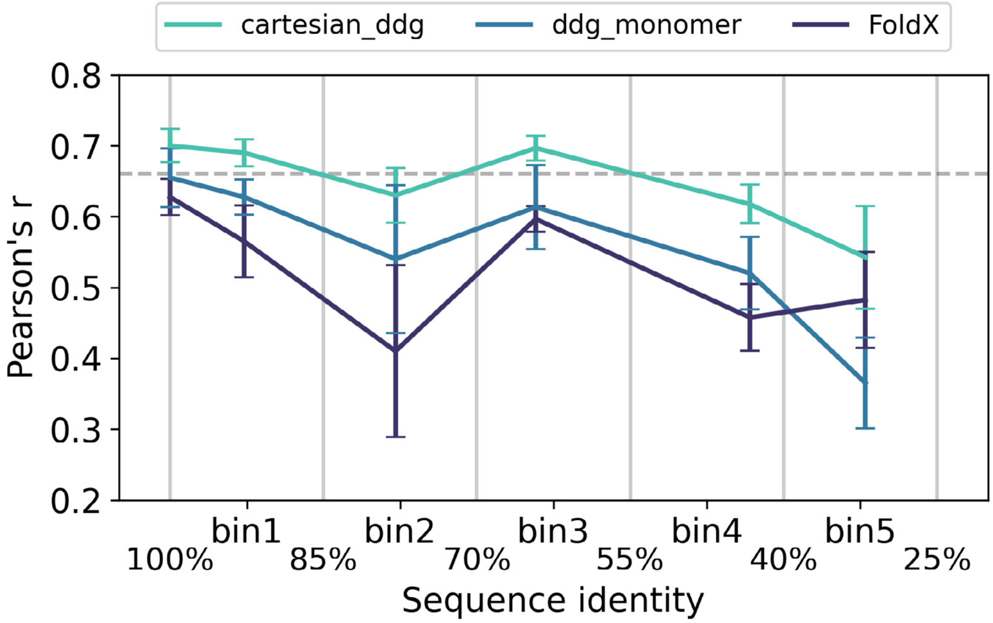


Fig. 3. Correlation of ddGs calculated on the original crystal structure as well as homology models with decreasing sequence identity to the experimentally determined values. Turquoise, Rosetta `cartesian_ddg`, blue, Rosetta `ddg_monomer`, violet, FoldX. The dashed line indicates the average performance 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.

In addition to the crystal structure of wild-type TPMT (PDB ID 2BZG) we built homology models of TPMT using templates with 80 % identity (2 GB4) and 32 % identity (3LCC and 1PJZ).

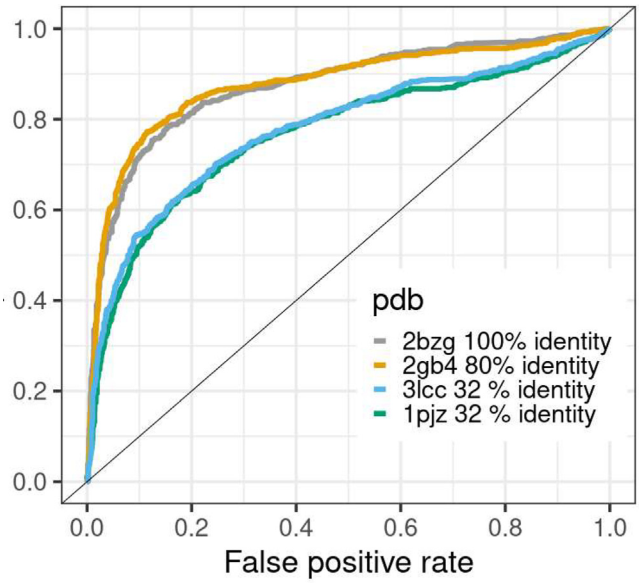


Fig. 4. Receiver Operating Characteristic (ROC) curve of classifying variants of human thiopurine methyltransferase (TPMT) into high and low abundance based on the ΔΔG calculated using Rosetta `cartesian_ddg`, compared to experimentally determined abundance by VAMPseq [24]. ΔΔG values were calculated on a crystal structure of TPMT (2BZG, gray, AUC = 0.873), a homology model built on a template with 80 % sequence identity (2 GB4, orange, AUC = 0.877), and two homology models built 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 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://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 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. $\Delta\Delta G$ 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 $\Delta\Delta$Gs is reported.

4.3. $\Delta\Delta G$ calculations: Rosetta ddg_monomer

We used Rosetta from October 2018 with git SHA1 ce9cb339991a7e8ca1bc44efb2b2d8b0a3d557f8. We ran `ddg_monomer` with the global repacking option (protocol 16, [13]) and the talaris2013 energy function, which has been shown to improve $\Delta\Delta$G 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 $\Delta\Delta$G 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_repulsive true.
−ddg::mut_only.

4.4. $\Delta\Delta G$ calculations: Rosetta cartesian_ddg

We used Rosetta from October 2018 with git SHA1 ce9cb339991a7e8ca1bc44efb2b2d8b0a3d557f8. For `cartesian_ddg` [12,14] we first performed constrained relaxation, then carried out 3 iterations of the cartesian_ddg protocol, and reported average $\Delta\Delta$Gs over these. For homology models, we carried out one iteration each over the 3 top 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_pdbs false.
−ddg::iterations 1.
−score::weights beta_nov16_cart.
−missing_density_to_jump.
−ddg::mut_only.
−ddg::bbnbrs 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 $\Delta\Delta$G, T53T in protein G and K33K in ubiquitin. We traced both back to articles reporting the $\Delta\Delta$G 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 $\Delta\Delta$G 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 $\Delta\Delta$G 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 Lindorff-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 Lindorff-Larsen and Stein labs for helpful discussions and comments. This work is supported by a Novo Nordisk Foundation, Denmark, Challenge Grant (PRISM, NNF18OC0033950 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

[1] Dill KA. Dominant forces in protein folding. Biochemistry 1990;29 (31):7133–55.
[2] Gromiha MM, An J, Kono H, Oobatake M, Uedaira H, Sarai A. ProTherm: thermodynamic database for proteins and mutants. Nucleic Acids Res 1999;27 (1):286–328.
[3] Nikam R, Kulandaisamy A, Harini K, Sharma D, Michael Gromiha M. ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. Nucleic Acids Res 2021;49(D1):D420–4.
[4] Tokuriki N, Tawfik DS. Stability effects of mutations and protein evolvability. Curr Opin Struct Biol 2009;19(5):596–604.
[5] Shoichet BK, Baase WA, Kuroki R, Matthews BW. A relationship between protein stability and protein function. *Proc Natl Acad Sci USA* 1995;92 (2):452–6.
[6] Goldenzweig A, Fleishman SJ. Principles of protein stability and their application in computational design. Annu Rev Biochem 2018;87 (June):105–29.
[7] Stein A, Fowler DM, Hartmann-Petersen R, Lindorff-Larsen K. Biophysical and mechanistic models for disease-causing protein variants. Trends Biochem Sci 2019;44(7):575–88.
[8] Yue P, Li Z, Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol 2005;353(2):459–73.
[9] Gerasimavicius L, Liu X, Marsh JA. Identification of pathogenic missense mutations using protein stability predictors. Sci Rep 2020;10(1):15387.
[10] Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 2002;320(2):369–87.
[11] Radusky LG, Luis Serrano. "pyFoldX: enabling biomolecular analysis and engineering along structural ensembles." *bioRxiv*. 2021 10.1101/2021.08.16.456210.
[12] Frenz B, Lewis SM, King I, DiMaio F, Park H, Song Y. Prediction of protein mutational free energy: benchmark and sampling improvements increase classification accuracy. Front Bioeng Biotechnol 2020;8(October):558247.
[13] Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins 2011;79(3):830–88.
[14] Park H, Bradley P, Jr PG, Liu Y, Mulligan VK, Kim DE, et al. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. J Chem Theory Comput 2016;12 (12):6201–12.
[15] Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. Nat Biotechnol 2012;30(11):1072–80.
[16] Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 2000;29:291–325.
[17] Chung SY, Subbiah S. A structural explanation for the twilight zone of protein sequence homology. Structure 1996;4(10):1123–2117.
[18] Kryshtafovych A, Barbato A, Monastyrskyy B, Fidelis K, Schwede T, Tramontano A. Methods of model accuracy estimation can help selecting the best models from decoy sets: assessment of model accuracy estimations in CASP11. Proteins 2016;84 Suppl 1(September):349–69.
[19] Shen M-Y, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci: Publ Protein Soc 2006;15(11):2507–24.
[20] Bienert S, Waterhouse A, de Beer TAP, Tauriello G, Studer G, Bordoli L, et al. The SWISS-MODEL Repository-New Features and Functionality. Nucleic Acids Res 2017;45(D1):D313–9.
[21] Ashenberg, Orr, L. Ian Gong, Jesse D. Bloom. Mutational effects on stability are largely conserved during protein evolution. *Proc Natl Acad Sci USA* 2013; 110 (52): 21071–76.
[22] Johansson KE, Johansen NT, Christensen S, Horowitz S, Bardwell JCA, Olsen JG, et al. Computational redesign of Thioredoxin is hypersensitive toward minor conformational changes in the backbone template. J Mol Biol 2016;428 (21):4361–77.
[23] Thiltgen G, Goldstein RA. Assessing predictors of changes in protein stability upon mutation using self-consistency. PLoS One 2012;7(10):e46084.
[24] Matreyek KA, Starita LM, Stephany JJ, Martin B, Chiasson MA, Gray VE, et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. Nat Genet 2018;50(6):874–82.
[25] Conchúir Ó, Shane KA, Barlow RA, Pache NO, Kundert K, O'Meara MJ, et al. A web resource for standardized benchmark datasets, metrics, and Rosetta protocols for macromolecular modeling and design. PLoS One 2015;10(9): e0130433.
[26] Webb, Benjamin, and Andrej Sali. "Comparative protein structure modeling using MODELLER." *Curr Protocols Bioinformatics* 2016; 54 (June): 5.6.1–5.6.37.
[27] Pak MA, Ivankov DN. Best templates outperform homology models in predicting the impact of mutations on protein stability. *Bioinformatics* 2022. https://doi.org/10.1093/bioinformatics/btac515.
[28] Fowler DM, Stephany JJ, Fields S. Measuring the activity of protein variants on a large scale using deep mutational scanning. Nat Protoc 2014;9(9):2267–84.
[29] Jepsen MM, Fowler DM, Hartmann-Petersen R, Stein A, Lindorff-Larsen K. Chapter 5 - Classifying Disease-Associated Variants Using Measures of Protein Activity and Stability. In: Pey AL, editor. Protein Homeostasis Diseases. Academic Press; 2020. p. 91–107.
[30] Hamborg L, Granata D, Olsen JG, Roche JV, Pedersen LE, Nielsen AT, et al. Synergistic stabilization of a double mutant in chymotrypsin inhibitor 2 from a library screen in *E. Coli*. Commun Biol 2021;4(1):980.
[31] Werner M, Gapsys V, de Groot BL. One plus one makes three: triangular coupling of correlated amino acid mutations. J Phys Chem Lett 2021;12 (12):3195–201.
[32] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596 (7873):583–659.
[33] Akdel M, Pires DEV, Pardo EP, Jänes J, Zalevsky AO, Mészáros B, et al. A Structural biology community assessment of AlphaFold2 applications. Nat Struct Mol Biol 2022;29(11):1056–67.
[34] The UniProt Consortium, The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res 2017. https://doi.org/10.1093/nar/gkw1099.
[35] O'Meara MJ, Leaver-Fay A, Tyka MD, Stein A, Houlihan K, DiMaio F, et al. Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. J Chem Theory Comput 2015;11(2):609–22.
[36] Loladze VV, Makhatadze GI. Both Helical propensity and side-chain hydrophobicity at a partially exposed site in alpha-helix contribute to the thermodynamic stability of ubiquitin. Proteins 2005;58(1):1–6.
[37] Smith CK, Withka JM, Regan L. A thermodynamic scale for the beta-sheet forming tendencies of the amino acids. Biochemistry 1994;33(18):5510–57.