

Original Research Article

Improving antibody thermostability based on statistical analysis of sequence and structural consensus data

Lei Jia^{*,†} , Mani Jain and Yaxiong Sun

Discovery Research, Amgen, Thousand Oaks, CA 91320, USA

Received: May 7, 2022; Revised: June 21, 2022; Accepted: July 12, 2022

ABSTRACT

Background: The use of Monoclonal Antibodies (MAbs) as therapeutics has been increasing over the past 30 years due to their high specificity and strong affinity toward the target. One of the major challenges toward their use as drugs is their low thermostability, which impacts both efficacy as well as manufacturing and delivery.

Methods: To aid the design of thermally more stable mutants, consensus sequence-based method has been widely used. These methods typically have a success rate of about 50% with maximum melting temperature increment ranging from 10 to 32°C. To improve the prediction performance, we have developed a new and fast MAbs specific method by adding a 3D structural layer to the consensus sequence method. This is done by analyzing the close-by residue pairs which are conserved in >800 MAbs' 3D structures.

Results: Combining consensus sequence and structural residue pair covariance methods, we developed an in-house application for predicting human MAb thermostability to guide protein engineers to design stable molecules. Major advantage of this structural level assessment is in significantly reducing the false positives by almost half from the consensus sequence method alone. This application has shown success in designing MAb engineering panels in multiple biologics programs.

Conclusions: Our data science-based method shows impacts in Mab engineering.

Statement of Significance: A data science-based method which can accurately predict antibody thermostability in high throughput. The method can also guide protein engineering by indication key residues that affect thermostability.

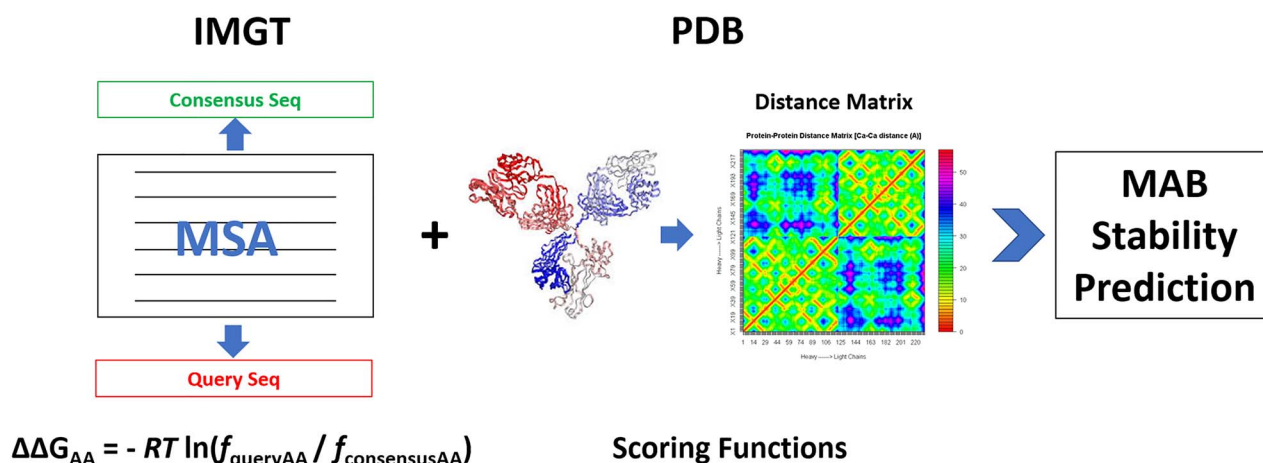
*To whom correspondence should be addressed. Lei Jia, Tel: 1-619-984-7473; Email: leijiachem@gmail.com

†Present Address: Bristol Myers Squibb, 10300 Campus Point Dr, San Diego, CA 92121

© The Author(s) 2022. Published by Oxford University Press on behalf of Antibody Therapeutics. All rights reserved. For Permissions, please email: journals.permissions@oup.com

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Graphical Abstract



KEYWORDS: monoclonal antibody; thermostability; consensus; structure; covariance

INTRODUCTION

Monoclonal Antibodies (MAbs) have become one of the most important classes of therapeutics in various disease areas. Half of the top 10 bestselling drugs in year 2019 are MAbs [1]. Thermostability is a basic biophysical property of MAbs. And it can be a major concern in the development of protein therapeutics due to its impact on both efficacy as well as manufacturing and delivery [2]. Low thermostability can cause MAbs to denature and aggregate [3], which can further lead to loss in binding potency [4], lower purity in manufacturing [5], and shortened shelf life [6]. Optimizing MAbs' thermostability attribute is among the fundamental protein engineering processes for a therapeutic MAb discovery project [7–11]. A high-throughput and accurate *in silico* prediction method can be helpful to eliminate liable molecules as early as possible and design mutations to improve thermostability of lead molecules.

There are multiple approaches to predict protein thermostability and engineer protein to improve its stability [12, 13]. Protein consensus sequence-based methodology has been applied to improve protein thermostability for over 20 years [14–18]. In general, the success rate for this statistics-based method is about 50% with maximum melting temperature increase ranging between 10 and 32°C [19]. Since MAbs all fold into conserved structures [20], we can push the consensus method to structural level and study the covariance relationship between residue pairs on 3D protein structures. Structural level assessment helps to significantly decrease false positives (FPs) from the consensus sequence method alone.

Combining consensus sequence and structural residue pair covariance methods, we developed an in-house application for predicting human MAb thermostability to guide protein engineers to design more stable molecules. The consensus method was trained by ~25 K and ~12 K human heavy and light chain variable region sequences,

respectively, from The International ImMunoGeneTics Information System (IMGT) <http://www.imgt.org/> [21]. The structural covariance method was trained by over 800 curated high-resolution MAb crystal structures. A scoring system was developed to evaluate pairwise residue interaction with confidence in consideration. Guided by data science and artificial intelligence, the application, which consists of about 1 500 lines of python codes, was incorporated in our antibody engineering workflow. It has shown success in designing MAb engineering panels in multiple biologics programs. Further development areas include enriching the training data for human MAb prediction (improve accuracy based on statistical significance), developing predictive models for other species' antibodies e.g. camelid heavy chain only antibody, and seeking applications for multi-specific antibody engineering.

METHODS

Sequence-based consensus scoring

MAb sequences from IMGT were used as the starting point. We specifically focused on the variable domains of human MAbs. From IMGT (as of year 2016), we obtained 35 614 human VH, 7 674 human VK, and 5 430 human VL sequences. The human germline sequence definition was taken from the V BASE, <https://www2.mrc-lmb.cam.ac.uk/vbase/>. All antibody sequences from IMGT were assigned to a germline type and germline family based on sequence alignment to the germline sequences. To have a cleaner germline annotation, for each sequence being used for consensus calculation, we set up 80% sequence similarity threshold as a filter to construct the consensus sequence data base. The filtering process yielded 25 220 VH, 7 190 VK, and 4 789 VL sequences. Table 1 shows the number of sequences in each germline family.

Table 1. Number of sequences in each germline family

Germline family	No. of sequences	Germline family	No. of sequences	Germline family	No. of sequences
VH1	3 111	VK1	3 412	VL1	1 307
VH2	619	VK2	1 049	VL2	1 039
VH3	15 610	VK3	2 143	VL3	1 465
VH4	3 490	VK4	516	VL4	154
VH5	1 005	VK5	29	VL5	110
VH6	1 231	VK6	41	VL6	300
VH7	154			VL7	132
				VL8	206
				VL9	38
				VL10	38

The consensus sequences were calculated based on three different levels: 1, all sequences in VH, VK, or VL sequence database; 2, all sequences at the germline family level; 3, all sequences at the germline level. To calculate the consensus sequence, the sequences were annotated and aligned following an Amgen in-house numbering scheme, which is similar to the AHO numbering scheme. At each residue position, the amino acid with the highest frequency was defined as consensus amino acid.

The idea behind the consensus sequence-based protein stability engineering method is that “a conserved residue is more likely to be stabilizing than a random mutation at that same position”. For a given amino acid in the query sequence (the sequence that needs to be evaluated for thermostability), we defined the consensus score as a free energy change ($\Delta\Delta G_{AA}$) by using Boltzmann distribution theory:

$$\Delta\Delta G_{AA} = -RTLn(f_{queryAA}/f_{consensusAA}) \quad (1)$$

Where R is the Boltzmann constant, T is the temperature (298 K), $f_{queryAA}$ is the frequency of the query amino acid at a given position, $f_{consensusAA}$ is the frequency of the consensus amino acid at that same position. $\Delta\Delta G_{AA}$ measures the effect of a single residue change to consensus amino acid on the stability of the antibody molecule and, a positive $\Delta\Delta G_{AA}$ reflects increased stability. The consensus score for the whole antibody sequence ($\Delta\Delta G_{Seq}$) is the sum of the consensus score of each individual amino acid’s consensus score as shown in equation (2). $\Delta\Delta G_{Seq}$ measures the overall effect of all the single amino acid changes to consensus residues on the stability of the MAb and the higher the $\Delta\Delta G_{Seq}$ is, the more stable the MAb will be.

$$\Delta\Delta G_{Seq} = \sum \Delta\Delta G_{AA} \quad (2)$$

Applying the consensus analysis can be based on the whole IMGT database level, germline family level or germline level, ranging from high level to low level sequence coverage. The higher the level (e.g. at IMGT database

level), the more sequences are available in the gene pool, and thus the more statistically meaningful results can be obtained. On the other hand, the higher the level, the sequences used to calculate consensus are less specific to the query sequence (lower sequence homology). Thus, the results may not be accurate. The scope of the analysis can be based on Complementarity-Determining Region (CDR), Framework Region (FR), whole Fv, or the whole Fab region. Note that engineering CDR region residues has a high risk of affecting Mabs’ binding affinity. The consensus method works best for the FR region as FR region residues provide structural stability to the MAb. For the purpose of this study, we focused on the germline family level (the sequence homology and statistical confidence are appropriate at this level), FR region (providing structural stability), and single point mutations (providing cleaner validation). Hence, in our Results section, the $\Delta\Delta G_{Seq}$ (ddG) values reflect the effect of mutating one position in the query sequence to a consensus or germline residue.

Metric to select the best ddG cutoff

ddG cutoff serves as our confidence threshold for predicted thermostabilizing mutations. If for a query sequence, the ddG score is higher than the ddG cutoff value, we suggest the protein engineers with a high confidence that the mutation to consensus residue can be thermostabilizing and can help to improve the melting temperature. The predictions from the consensus sequence method can fall under four categories as shown in Fig. 1. True Positives (TP): mutations being predicted to have ddG values greater than or equal to the ddG cutoff. And experimentally observed melting temperature (T_m) difference between the mutant and the query sequence is greater than or equal to the dT_m cutoff. FPs: mutations being predicted to have ddG values greater than or equal to the ddG cutoff. But experimentally observed melting temperature difference between the mutant and the query sequence is less than the dT_m cutoff. True Negatives (TNs): mutations being predicted to have ddG values less than the ddG cutoff. And experimentally observed melting temperature difference between the mutant and the query sequence is lower than the dT_m cutoff. False Negatives (FN): mutations being predicted to

Table 2. Different metrics being explored for selecting the best ddG cutoff.

Metric	Formulae	Criteria for choosing ddG cutoff ^{note}
Precision	TP/(TP + FP)	Not defined
Metric 1	TP*(TN-FP)/(FP*TN)	> 0
Metric 2	TP*(TN-FP)	≥ 0
Primary metric (P Metric)	TP + TP*(TN-FP)	> 0
Fine-tuning metric (F Metric)	(TP-TN)/(TN-FP)	≥ 0

Note: Criteria are based on navigation from lowest to highest ddG cutoff values

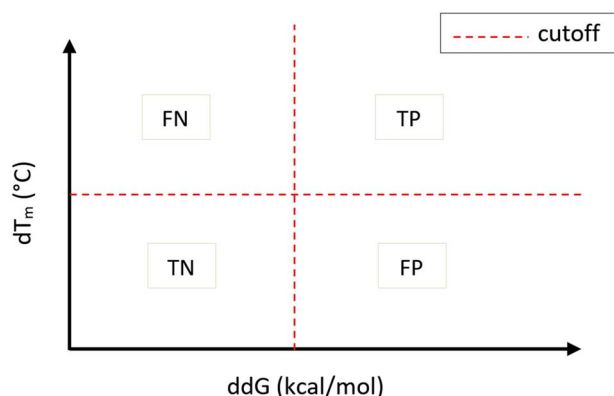


Figure 1. The x - y plot is showing the four categories into which the predictions from the consensus sequence method can fall. The x -axis represents the ddG values predicted by consensus sequence method and y -axis shows the experimentally observed melting temperature difference between the mutant and reference molecule. The red dotted lines represent the cutoff values for ddG and dT_m .

have ddG values less than the ddG cutoff. But experimentally observed melting temperature difference between the mutant and the query sequence is greater than or equal to the dT_m cutoff.

The dT_m cutoff is dependent on the experimental error in T_m measurements. The idea behind the metric to select the best ddG cutoff is based on maximizing TP, minimizing FP and at the same time maximizing value for true predictions (TP + TN). We care about TP the most, and then reducing FP and then maximizing our total true predictions. We devised and explored four different metrics along with precision as shown in Table 2.

Finally, we decided on using the two metrics, the primary metric (P metric) and the fine-tuned metric (F metric) in conjunction with one another. The first step was to titrate the ddG cutoff values and sort them, we used a step size of 0.1. The criteria for choosing the best ddG cutoff were defined as we navigate from the lowest to the highest ddG cutoff. The selection of the best ddG cutoff was defined with respect to the ddG value at which P metric > 0 ($ddG_{cutoff(i)}$) and if the value of F metric was < 0 at $ddG_{cutoff(i)}$, then we would pick a ddG cutoff one before the one which had P metric > 0 i.e. $ddG_{cutoff(i-1)}$. Otherwise, if F metric was positive at $ddG_{cutoff(i)}$, we would pick that as the best ddG cutoff. This procedure is described in the following pseudocode. The key point is to traverse possible ddG cutoff values from the lowest to the highest, and finer

titrations of ddG cutoff gave us a better cutoff value.

$$ddG_{cutoff} \in \{0, 0.1, \dots, \max(ddG)\}$$

$$\text{if } P \text{ metric} > 0 \text{ at } ddG_{cutoff(i)}$$

$$\text{if } F \text{ metric} < 0 \text{ at } ddG_{cutoff(i)} \text{ then select } ddG_{cutoff(i-1)}$$

$$\text{elif } F \text{ metric} \geq 0 \text{ at } ddG_{cutoff(i)} \text{ then select } ddG_{cutoff(i)}$$

Consensus structure-based MAb residue-pair covariance analysis

We obtained 841 crystal structures of human antibody variable domain from IMGT 3D structure database. Pairwise residue distance matrix was calculated for each of the Fv structure using only FR residues. We used the following two criteria to flag a pair of residues as close-by: 1. only calculated amino acid residue pairs which are separated by more than two amino acids (to avoid analyzing adjacent pairs and loop tip at the beta-hairpin); 2. picked minimum distance between all side chain heavy atoms of the amino acid pairs, used 4 Å distance as a cutoff. Residue pairs which had minimum distance between any side chain heavy atom < 4 Å were flagged as close-by residue pairs. The residue numbers on the variable domain were uniform for all 841 MABs based on the Amgen in-house numbering system. After all close-by residue pairs were calculated for all 841 antibody structures, we used 100 occurrences as a cutoff to mark consensus close-by residue pairs. This yielded 257 close-by residue pairs and we used these for assessing the residue synergy in the query antibody.

Close-by residue pairs are preferred if the two residues had opposite charge leading to columbic interaction, or both having high hydrophobicity for favored van der Waals interactions (packing), or both were aromatic for favored stacking interaction. To quantify the favorable interactions, we developed a scoring system:

$$\text{Charge score} = \text{charge score}(\text{residue 1}) * \text{charge score}(\text{residue 2}) \quad (3)$$

$$\text{Hydrophobicity score} = -1 * \text{hydrophobicity score}(\text{residue 1}) * \text{hydrophobicity score}(\text{residue 2}) \quad (4)$$

$$\begin{aligned} \text{Aromaticity score} &= -1 * \text{aromaticity score} \\ &(\text{residue 1}) * \text{aromaticity score}(\text{residue2}) \end{aligned} \quad (5)$$

$$\begin{aligned} \text{Total score} &= \text{charge score} + \text{hydrophobicity score} \\ &+ \text{aromaticity score} \end{aligned} \quad (6)$$

$$\begin{aligned} \text{dScore} &= \text{total score}(\text{query sequence}) \\ &- \text{total score}(\text{reference sequence}) \end{aligned} \quad (7)$$

The charge and hydrophobicity scores were defined by amino acid charge and hydrophobicity in literature [22, 23]. The aromaticity score was based on Gasser [23] and defined with our empirical adjustment as following (tryptophan: 1, phenylalanine and tyrosine: 0.8, histidine: 0.3, arginine: 0.3, and the rest of amino acids were 0). The reference sequence could either be the germline sequence or the consensus sequence. Similar to energy, the lower the dScore was, the more favorable the residue pair was, thus the more favorable the antibody which contains such residue pair was. One residue could be presented in multiple close-by residue pairs. For protein engineering purpose, each residue's impact on protein stability is summed up by their contribution to every close-by residue pairs which contained it with adjustment of a confidence level:

$$\begin{aligned} \text{Sum dScore}(\text{residue } A) &= \sum_{i=\text{residue pair } i}^n \text{dScore} \\ &(\text{residue } A)_i * \text{confidence}_i \end{aligned} \quad (8)$$

Confidence was defined as the ratio of number of times a pair of residues is present as close-by residue pairs to the total 841 structures:

$$\begin{aligned} \text{confidence} &= \\ &\frac{\text{Number of structures which include a closeby residue pair}}{\text{Number of total structures (841)}} \end{aligned} \quad (9)$$

The final output of the MAb residual covariance analysis was the sum dScore (equation (8)). Compared with the reference MAb, if this score was a positive number, it meant that the residue was less stable than the residue in the reference MAb. A mutation to consensus residue was suggested to stabilize the query MAb.

This method of combining the consensus sequence method with consensus structure-based MAb residue-pair covariance analysis was implemented in Python programming language. A graphical user interface was also developed with Pipeline Pilot for deployment. It required input Fv domain sequence of the query MAbs in

fasta format (annotated and aligned following the Amgen numbering scheme). The workflow of the in-house Pipeline Pilot tool included two steps, the first step was to calculate the consensus sequence and the ddG score (equation (1)) for all residues in the query sequence in the FR region, and the second step was to perform the structural filtering. The tool output a csv file with thermostabilizing mutation suggestions for each query MAb including the ddG score. The users had an optional flag to turn off structural filtering i.e. retrieve mutation suggestions based only on the consensus sequence method. By default, structural filtering was turned on and the output only included thermostabilizing mutations which had passed structural filtering.

RESULTS

Evaluate the consensus sequence method with published data

To evaluate the consensus sequence prediction method, we used a stability engineering study of the single chain fragment variable domain (scFv) of an antibody published by Monsellier *et al.* [24] In that study the authors used consensus method to predict point mutations of scFv to improve thermostability. Experimental free energy change (ddG) was reported to assess the computational method. Table 3 shows the experimental ddG, calculated ddG from that paper and from our method for the mutants. Our calculated ddG values were similar to those reported in the paper and both calculated ddG values per variant were in the same positive direction as in the experimental ddG, which indicated good prediction outcome. Through this comparison, we validated the consensus sequence method that we implemented.

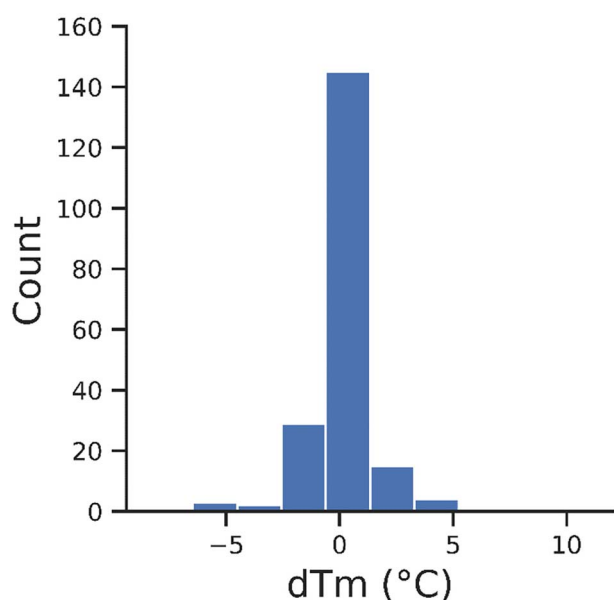
Select the optimal ddG cutoff for thermostability prediction and MAb engineering

We noted that the consensus sequence stability method did not have high accuracy to confidently predict the actual melting temperature (T_m) and experimental ddG directly. Protein engineers desired to have a high throughput prediction method to predict variants' stability and propose what mutations they can make to improve stability. So, we further developed this method to be a classification tool to meet the protein engineers' need. For classification, we need to determine a ddG cutoff value to confidently predict the amino acid stability at a given position and identify the amino acid position which can be engineered to improve stability. We focused on the FR of the MAb Fv and analyzed only single point mutations for cleaner validation.

We used a set of 234 internal MAb thermostability data in T_m representing 201 single point mutation pairs to determine the optimal ddG cutoff. The dT_m cutoff used in our work was 0.5°C, it was chosen based on the distribution of dT_m of our experimental data as shown in Fig. 2 and suggestion from the analytical scientist who performed the measurements (according to the experimental error bar in T_m measurements by Differential Scanning Fluorimetry). We devised four metrics namely Metric 1, Metric 2, P metric, and F metric as shown in Table 2. We compared the TP,

Table 3. Experimental and calculated ddG (in kcal/mol) to validate our consensus sequence method

Variants	Mutations	Experimental ddG	Calculated ddG from Monsellier <i>et al.</i>	Calculated ddG from our work
L1	L-Q40P,L-K42Q	0.6	3.1	4.2
L2	L-Q45K	4.1	1.3	0.9
L3	L-K74T	2.9	0.7	1.1
L4	L-N76S	0.6	1.5	1.7
L5	L-G84A, L-S85T	2.5	2.3	3.5
H1	H-S15G	0.1	1	0.7
H2	H-S61E, H-A62K, H-L63F	1.6	2.1	2.5
H3	H-H83T, H-T84S, H-D85E	0.9	5.9	6.7

**Figure 2.** The frequency distribution of dT_m for 201 single point mutation pairs in our dataset.

FP, TN, and FN statistics for different ddG cutoff values using these four metrics and the precision as shown in Table S1. The criteria for selecting the optimal ddG cutoff using each of these five metrics are shown in Table 2. For the precision, we were not able to come up with any definite criteria based on which we could choose the optimal ddG cutoff.

If we navigated from the lowest to the highest ddG cutoff values at the fixed dT_m cutoff of 0.5°C , we observed that TP and FP would reduce, but TN and FN would increase as shown in Table S1. Since we cared the most about maximizing TP, we started the navigation from the lowest ddG cutoff as TP would be the highest. The second goal was to minimize FP as much as possible without sacrificing many TP. Based on this rationale, we chose P metric and F metric together as defined in the Methods section. For further information, refer to the analysis shown in Table S1 and Fig. S1. Based on our P metric and F metric, we

selected 1.7 kcal/mol as the optimal ddG cutoff for the consensus sequence method. The selected optimal ddG cutoff depended on the data set and prediction confidence goal. So, in another use case, this value can be different from the value we obtained. In the next section, we demonstrated that this cutoff value could be further optimized based on another use case (adding consensus structural filter). In future, with more data, this cutoff can be further optimized.

Consensus structure-based MAb residue-pair analysis can significantly decrease the false positive prediction rate

Human MAb has a conserved structural fold. Described by IMGT's Colliers de Perles illustration, the variable domain of heavy and light chains each has nine anti-parallel beta-sheets, which form the FR of the variable domain. Three out of four loops in connecting the beta-sheets build up the CDR region. The consensus based on a set of MAb crystal structures can provide a general representation of their overall structural features. This is the rationale behind the consensus structure-based MAb residue-pair method. This method is high throughput, but is not highly accurate to predict the stability directly. So, we only used this method as a filter on top of the consensus sequence method. We demonstrated that this structural method can significantly decrease the FP of the consensus sequence method prediction.

We used the same set of 234 MAb thermostability data in T_m representing 201 single point mutation pairs for validating the consensus structure-based MAb residue pair method. About 154 single point mutation pairs out of 201 passed the structural filter i.e. they had a positive sum dScore as described in equation (8). For practical application, we optimized the ddG cutoff again using these 154 data points as the protein engineers would only be looking at the mutations which passed the structural filter. By applying P metric and F metric on these 154 single point mutation pairs at the fixed dT_m cutoff of 0.5°C , we obtained the optimal ddG cutoff of 1.3 kcal/mol (for further details, refer to the Table S2 and Fig. S2). For our further analysis, we used this practical optimal ddG cutoff of 1.3 kcal/mol.

Table 4. Performance comparison between without and with structural filtering.

Statistics	Without Structural Filtering	With Structural Filtering
TP	60	51
FP	67	38
TN	47	38
FN	27	27
Precision	0.47	0.57

Note: Precision measures the number of correct positive predictions made, it is defined as $TP/(TP + FP)$. Maximum possible value for precision can be 1

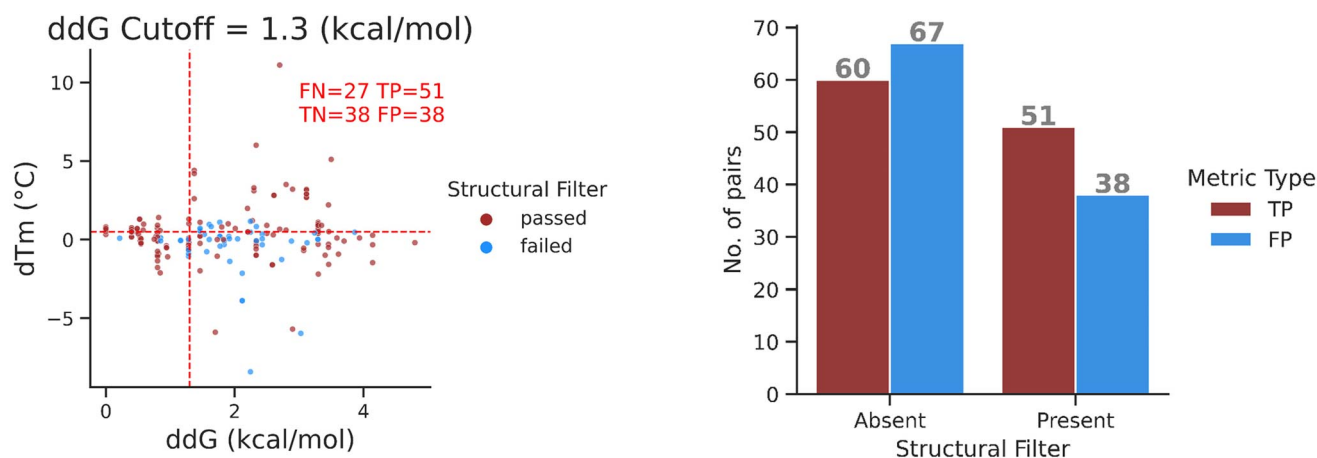


Figure 3. Scatter plot (x - y plot) of dT_m VS ddG . X -axis represents the ddG values computed from the consensus sequence method and y -axis represents the dT_m values (difference of experimental T_m between the parent and the mutant molecules). The single point mutation pairs are colored based on whether they pass the structural filter or not. The single point mutation pairs which pass the structural filter are classified into TP, FP, TN, and FN based on the dT_m cutoff of 0.5°C and the ddG cutoff of 1.3 kcal/mol. The red dotted lines represent the cutoff values for ddG and dT_m .

Figure 3 is a scatter plot of dT_m V. S. ddG for all 201 single point mutation pairs. dT_m describes the difference of experimental T_m between the mutant and the parental molecules. ddG is predicted from the consensus sequence method. We also computed confusion matrix and precision comparing the performance between without and with structural filtering (Table 4). Since we cared most about maximizing TP and minimizing FP, we would aim to have as high precision as possible. We could observe that structural filtering lead to precision improvement. To visualize the effect of consensus structure-based MAb residue-pair method in reducing FP, we plotted a bar plot as shown in Fig. 4. We could see structural filtering helped in significantly reducing the FP without sacrificing much of the TP. In addition, we observed a consistent effect of structural filtering on reducing number of FP at different ddG cutoffs.

After examining the FP cases, we found out that one primary reason to cause FP without structural filtering was due to conflicts in the local structural environment. For example, we observed A to R or S to R mutations which were favored by sequence consensus (R was the consensus amino acid at this position). However, the location of this residue was buried inside the protein and the large side

Figure 4. Bar plot showing the effect of structural filtering on reducing FP. TP and FP statistics in this plot are based on the optimal ddG cutoff of 1.3 kcal/mol and dT_m cutoff of 0.5°C .

chain of R cannot fit in its location. In another example, we observed V to G mutation which was favored by sequence consensus. But the V to G mutation created a void in the core of the Fv structure and thus destabilized the protein. In a similar example, an F to S mutation was favored by sequence consensus, but such mutation removed a key structurally favored hydrophobic stacking interaction. With those examples, we demonstrated that structure information could significantly improve the thermostability prediction accuracy.

One feature of our method was to implement structure information without spending extra computing resource to model the query Mabs' structure. The structure information was encoded in the method via pretraining by using a statistically significant amount of Mab crystal structures. As the available MAb experimental structures grow, the consensus structure-based MAb close-by residue pairs can be retrained with higher confidence. The scoring system can be further improved with consideration of more detailed molecular interaction. After reach a high level of accuracy, we can use the scoring system to rank molecules' stability. The method can help to flag any residue pairs in the query molecule that may be liable for stability. This is equivalent to building homology models and manually examining the structures. But our method does not require homology modeling and the manual modeling efforts, so it is more systematic and higher throughput than the traditional thermostability engineering methods. It is suitable to deal with

a large number of Mab sequences at industry scale (see Discussion for more details).

DISCUSSION

Fast stability engineering method in comparison to homology modeling approach

One key advantage of our consensus sequence method in combination with consensus structure-based MAb residue pair covariance analysis is the fast calculation speed and a capability to process a large amount of MAb sequences in high throughput. On average, it takes 19.5 s per MAb molecule. In a typical MAb engineering workflow, a protein engineer may choose homology modeling approach to obtain the same information which can be obtained by residue-pair covariance analysis. However, it usually takes 3–30 min to build a reasonable homology model per MAb molecule. And manually examining the three-dimensional models is even more time consuming. So, the homology modeling approach is not suitable for a large-scale MAb stability engineering task. Our method was pretrained by a set of nearly 1 000 high resolution crystal structure of MAbs. It took advantage of conserved structural fold of MAbs to yield the consensus close-by residue pairs. This close-by residue pair information was then used to evaluate the fitness of stability of query MAb based on its sequence. So, our method leverages pretrained structural information to process only sequence data to achieve the goal of fast calculation speed.

Factors affecting prediction accuracy

The accuracy of consensus sequence prediction and design depends on the following four factors: 1. The gene pool, i.e. the sequence database that is used for multiple sequence alignment (MSA). The larger and more comprehensive the gene pool is, the better chance that the consensus residue can be truly representative. 2. Sequence homology; 3. Sequence count in MSA. 2 and 3 are correlated. The higher the sequence homology is e.g. use germline subset in our application, the fewer the sequence counts can be used in MSA. Therefore, the sequences being selected for MSA are more similar to the query sequence and the consensus sequence being yielded from MSA can be more accurate. On the other hand, the fewer number of the sequences in the MSA are, the statistical significance would be lower. So, factors 2 and 3 are a tradeoff. The default option for selecting sequences for MSA is germline family, which is a balance between all sequences in IMGT and very few sequences in a given germline. 4. The bias from sequence alignment algorithm being used for MSA. For MAb application, several well-developed numbering schemes e.g. IMGT, Chothia, Kabat, and Aho are available. Those numbering systems generate MSA by using different rules.

Species-specific application

To make the method more accurate, we developed it with certain species consideration. We deployed the method for human MAbs at first. This is our most common use case.

All consensus and germline sequences are specific to human MAbs. The consensus method can be further developed for other species given the sequence data are available. Mouse MAb development is the second common use case. We obtained mouse MAb sequences and germline information from IMGT and further developed the method for predicting mouse antibody thermostability.

Non-antibody applications

The consensus sequence-based stability prediction and design originated from non-antibody applications. The foundation of this method is MSA of a set of sequences with certain level of homology cutoff. In the case of antibody, since the homology of sequences is already very high within the same species, it is possible to include all sequences from repertoire like IMGT. For higher homology cutoff, germline family and germline subset selections are available. For non-antibody applications, selecting the proper homology cutoff is critical to ensure the validity of MSA and for accurately identifying the consensus residues.

CONCLUSION

The goal of our work was to develop a method to help mitigate the liability of thermostability associated with MAbs. We combined consensus sequence method with consensus structure-based MAb residue pair covariance analysis to predict thermostabilizing mutations of the query MAb. The theoretical ground of our method is based on the idea that conserved structural fold of MAbs yield consensus close-by residue pairs. This residue pair information is applied to significantly reduce the FP by almost half compared with the consensus sequence-based method alone. Major advantages of our data science-based method are improved accuracy compared with the consensus sequence method alone, faster computation, as well as high-throughput capability compared with homology modeling-based approaches. Future areas of development include enriching the training data for human MAb prediction which can further improve accuracy due to the higher statistical significance, developing predictive models for other species' antibodies e.g. camelid heavy chain only antibody, and extending applicability to multi-specific antibody engineering.

SUPPLEMENTARY DATA

Supplementary Data are available at ABT Online

CONFLICT OF INTEREST STATEMENT

Lei Jia was an employee of Amgen Inc. Mani Jain and Yaxiong Sun are employees of Amgen Inc.

DATA AVAILABILITY

None proprietary data are available in supporting information as well as on public websites such as <https://www.imgt.org/> and <https://www.rcsb.org/>

FUNDING

None.

ETHICS AND CONSENT STATEMENT

Consent is not required.

ANIMAL RESEARCH STATEMENT

Not applicable.

ACKNOWLEDGEMENT

We would like to acknowledge our colleagues and friends who provided help for our works in this manuscript: Darren Bates, Hannah Catterall, Igor D'angelo, Bram Estes, Fernando Garces, Kevin Graham, Marissa Mock, Austin Rice, Daniel Yoo, and Adam Zalewski.

REFERENCES

- Blankenship, K. The top 20 drugs by global sales in 2019. (2020) <https://www.fiercepharma.com/special-report/top-20-drugs-by-global-sales-2019>.
- Le Basle, Y, Chennell, P, Tokhadze, N *et al.* Physicochemical stability of monoclonal antibodies: a review. *J Pharm Sci* 2020; **109**: 169–90.
- Ewert, S, Huber, T, Honegger, A *et al.* Biophysical properties of human antibody variable domains. *J Mol Biol* 2003; **325**: 531–53.
- McConnell, AD, Spasojevich, V, Macomber, JL *et al.* An integrated approach to extreme thermostabilization and affinity maturation of an antibody. *Protein Eng Des Sel* 2013; **26**: 151–64.
- Bondos, SE, Bicknell, A. Detection and prevention of protein aggregation before, during, and after purification. *Anal Biochem* 2003; **316**: 223–31.
- Weiss, WF IV, Young, TM, Roberts, CJ. Principles, approaches, and challenges for predicting protein aggregation rates and shelf life. *J Pharm Sci* 2009; **98**: 1246–77.
- Knappik, A, Ge, L, Honegger, A *et al.* Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides. *J Mol Biol.* 2000; **296**: 57–86.
- Seeliger, D, Tosatto, SCE. Development of scoring functions for antibody sequence assessment and optimization. *PLoS One* 2013; **8**: e76909.
- Couto, JR, Christian, RB, Peterson, JA *et al.* Designing human consensus antibodies with minimal positional templates. *Cancer Res* 1995; **55**: 5973s–5977s.
- Ewert, S, Honegger, A, Plückthun, A. Stability improvement of antibodies for extracellular and intracellular applications: CDR grafting to stable frameworks and structure-based framework engineering. *Methods* 2004; **34**: 184–99.
- Rouet, R, Lowe, D, Christ, D. Stability engineering of the human antibody repertoire. *FEBS Lett* 2014; **588**: 269–77.
- Kulshreshtha, S, Chaudhary, V, Goswami, GK *et al.* Computational approaches for predicting mutant protein stability. *J Comput Aided Mol Des* 2016; **30**: 401–12.
- Samish, I. *Computational Protein Design*. New York, NY: Springer, 2017
- Steipe, B, Schiller, B, Plückthun, A *et al.* Sequence statistics reliably predict stabilizing mutations in a protein domain. *J Mol Biol* 1994; **240**: 188–92.
- Lehmann, M, Pasamontes, L, Lassen, SF *et al.* The consensus concept for thermostability engineering of proteins. *Biochim Biophys Acta* 2000; **1543**: 408–15.
- Steipe, B. Consensus-based engineering of protein stability: from intrabodies to thermostable enzymes. *Methods Enzymol* 2004; **388**: 176–86.
- Sullivan, BJ, Nguyen, T, Durani, V *et al.* Stabilizing proteins from sequence statistics: the interplay of conservation and correlation in triosephosphate isomerase stability. *J Mol Biol* 2012; **420**: 384–99.
- Aerts, D, Verhaeghe, T, Joosten, H-JJ *et al.* Consensus engineering of sucrose phosphorylase: the outcome reflects the sequence input. *Biotechnol Bioeng* 2013; **110**: 2563–72.
- Porebski, BT, Buckle, AM. Consensus protein design. *Protein Eng Des Sel* 2016; **29**: 245–51.
- Ruiz, M, Lefranc, MP. IMGT gene identification and colliers de Perles of human immunoglobulins with known 3D structures. *Immunogenetics* 2002; **53**: 857–83.
- Lefranc, MP, Giudicelli, V, Duroux, P *et al.* IMGT R, the international ImMunoGeneTics information system R 25 years on. *Nucleic Acids Res* 2015; **43**: D413–22.
- Black, SD, Mould, DR. Development of hydrophobicity parameters to analyze proteins which bear post- or cotranslational modifications. *Anal Biochem* 1991; **193**: 72–82.
- Gasser, C. Amino acid properties. (2010) <http://www.mcb.ucdavis.edu/courses/bis102/AAProp.html>.
- Monsellier, E, Bedouelle, H. Improving the stability of an antibody variable fragment by a combination of knowledge-based approaches: validation and mechanisms. *J Mol Biol* 2006; **362**: 580–93.