Design of a Protein with Improved Thermal Stability by an Evolution-Based Generative Model

Tian, Pengfei; Lemaire, Adrien; Senechal, Fabien; Habrylo, Olivier; Antonietti, Viviane; Sonnet, Pascal; Lefebvre, Valerie; Marin, Frederikke Isa; Best, Robert B.; Pelloux, Jerome; Mercadante, Davide

Published in:
Angewandte Chemie International Edition

DOI:
10.1002/anie.202202711

Publication date:
2022

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

Document license:
CC BY-NC-ND

Citation for published version (APA):
Design of a Protein with Improved Thermal Stability by an Evolution-Based Generative Model

Pengfei Tian,* Adrien Lemaire, Fabien Sénéchal, Olivier Habrylo, Viviane Antonietti, Pascal Sonnet, Valérie Lefebvre, Frederikke Isa Marin, Robert B. Best, Jérôme Pelloux,* and Davide Mercadante*

Abstract: Efficient design of functional proteins with higher thermal stability remains challenging especially for highly diverse sequence variants. Considering the evolutionary pressure on protein folds, sequence design optimizing evolutionary fitness could help designing folds with higher stability. Using a generative evolution fitness model trained to capture variation patterns in natural sequences, we designed artificial sequences of a proteinaceous inhibitor of pectin methylesterase enzymes. These inhibitors have considerable industrial interest to avoid phase separation in fruit juice manufacturing or reduce methanol in distillates, averting chromatographic passages triggering unwanted aroma loss. Six out of seven designs with up to 30 % divergence to other inhibitor sequences are functional and two have improved thermal stability. This method can improve protein stability expanding functional protein sequence space, with traits valuable for industrial applications and scientific research.

Introduction

The development of efficient computational approaches for protein engineering carries enormous biotechnological and medical interest.[1] Even though more and more protein structures are being solved, especially with the recent breakthrough of AlphaFold,[2] systematically improving protein stability through protein engineering remains a big challenge. However, the proteins which have been experimentally characterized as stable and functional are only an extremely tiny subset of those which are theoretically predicted to be stably folded, suggesting rich potential for design.[3] The design and discovery of thermostable protein structures purely based on physics principles have made remarkable progress thanks to advanced techniques such as FoldX[4] and Rosetta.[5] However, predicting the stability and function of sequences with high divergence is still a challenging task,[6] often with several rounds of design iterations having to be applied.[7]

Multi-residue mutational patterns within sets of protein homologous sequences provide important clues of biophysical restraints, useful to maintain the 3D structure and function of proteins. Previously, evolution-based generation of new sequences has mostly been obtained by either randomly recombining fragments of proteins homologous to a target design[8] or via consensus sequence design,[9] in which sequences are generated using information concerning the probability of finding certain residues along a sequence aligned with multiple homologous. Both these approaches ignore global statistical dependency between residues, deemed crucial for protein fitness.[10] Recent studies showed that a statistical model can capture the complex pattern of variants between positions by accounting for both residue conservation and covariation.[11] These kinds of models are very powerful for predicting residues
which are close to each other on the 3D structure and for reconstructing epistatic effects between residues. A few deep learning methods have also been devised to model protein fitness landscapes e.g. variational autoencoders (VAEs),[12] generative adversarial networks[13] and autoregressive models,[14] and they can be used to design proteins with nature-like functions.[14a,15] However, the application of these generative models to guide the design of highly divergent sequences with enhanced stability while remaining functional, has not been demonstrated.

The rapidly expanding number of design algorithms has also prompted the development of statistical frameworks able to meaningfully score algorithm effectiveness within the purpose of increasing protein fitness.[16] In this work, we build the statistical model of protein fitness landscape using a Potts-type model, which has been shown to be better at capturing higher-order mutational statistics than other methods such as VAEs.[17] Using a similar approach, we have previously successfully designed sequences leading to three different small protein folds (GA, GB and SH3), all featuring 56 amino acids.[18] However, none of the designed variants had improved stability over the wild type. It has also remained unclear if this method could be used to design functional proteins with improved stability of larger size, since the number of foldable sequences rapidly becomes a much smaller fraction of the total number of possible sequences with increasing protein length.[19]

In this work, we address this question by exploring the high fitness sequence space of a pectin methylesterase enzymes inhibitor (PMEI) from Actinidia chinensis (kiwifruit), here dubbed PMEI_{Act}, which is 150 amino acids in length. PMEI_{Act} has been previously co-crystallized bound to a pectin methylesterase enzyme (PME) from Solanum lycopersicum (tomato), in a 1:1 complex (PDB accession code: 1xg2, shown in Figure 1a), in which the inhibitor engages three α-helices of its four-helix bundle, to obstruct the substrate’s binding groove of the target PME (Figure 1a).[20]

PMEIs have been identified in many different species of plants, where they finely tune the activity of PMEs by direct competition with the enzymes’ substrate.[19] PMEs hydrolyze methyl esterified galacturonic acids in homogalacturonan (HG)-type pectins: the major polysaccharide component of the plant cell wall.[20] HGs are synthesized as highly methyl esterified polymers (i.e. >80 %) and their subsequent de-methyl esterification can lead to distinct physico-chemical properties of the plant cell wall.[21] Variations in the expression levels of PMEI genes are known to trigger dramatic changes in the chemistry of pectin with important effects on plant development.[22] In plants, PMEIs are encoded by large multigenic families, which questions the functionality of their interaction with different PME isoforms and the existence of specific PME-PMEI pairs.[19,23] A PMEI from one species can interact with PMEs from other species, as shown in the co-crystallized PME:PMEI complex featuring a tomato PME in complex with a PMEI from kiwifruit.[20] This high degree of promiscuity provides the opportunity to particularly explore the effect of mutations at the binding interface versus the rest of a PMEI structure.

Hence, new designs of PMEIs can be used to both understand the plant life cycle as tools to probe plant-cell wall physiology, or as industrial tools during the production of fruit juices that are rich in pectins,[24] or to control the quantity of methanol produced by PMEs in processes like distillation. In the case of grape distillates, a second distillation featuring a de-methylating chromatographic passage is indeed needed to achieve methanol concentrations below the legal limit but it is also responsible for the loss of aroma compounds, ultimately downgrading the quality of the final product.[25] Although PMEIs hold a considerable interest for both industrial applications and a deeper understanding of plant physiology, to the best of our knowledge there has not been any attempt to engineer PMEIs before, but only to optimize molecular biology protocols and use the naturally expressed PMEI from kiwifruit (PMEI_{Act}) to target methanol reduction in

Figure 1. Complex between PME and PMEI and UMAP analysis. a) Structure of the pectin methylesterase inhibitor (PMEI) from Actinidia chinensis bound to a pectin methylesterase (PME) from Solanum lycopersicum (PDB accession code: 1xg2). The protein structures are colored in grey and green for PME and PMEI respectively. PMEI residues sitting 0.6 nm from the PME binding interface are colored in pink. b) Structure of the PMEI with predicted co-evolving residues linked by purple lines. c) Homologous sequences (blue) of PMEI_{Act} are projected onto two dimensions by UMAP analysis. Designed variants 1–4 and variants 5–7 are mapped as pink and yellow, respectively. Green and purple points are representative artificial sequences extracted from Monte Carlo sampling trajectories (Figure 2).
distillates. Eventually, the design of some PMEIs bearing a higher structural stability or specificity profiles, can be beneficial to both industry and research, with the latter that can adopt designed PMEIs to probe the homeostasis of plant cell wall polysaccharides and their interplay with different PME isoforms. The ability to understand plant cell wall physiology is indeed particularly restrained by the combinatorial explosion of the PME:PMEI pairs arising from the tens of isoforms expressed in plants and their binding promiscuity.

Results and Discussion

Sequence Design of PMEI Variants

The statistical model that we use for the design of PMEIs’ artificial sequences employs two sets of parameters: amino acid conservation (b) and higher-order inter-residue covariance (W), which describe the probability of finding that a certain protein sequence \( \{X_i\}, \) which belongs to the PMEI family (blue in Figure 1c) based on the following pattern of sequence variation:

\[
P(\{X_i\}, \{X_j\}, \ldots, \{X_L\}) = \frac{1}{Z} \exp \left\{ \sum_{i<j} W_{ij}(X_i, X_j) + \sum_i b(X_i) \right\}
\]

where \( i \) and \( j \) correspond to different residue positions, and \( L \) is the total length of the protein. \( b \) and \( W \) can be obtained by the pseudo-likelihoods method (Figure 2a), or using Boltzmann machines. As illustrated in Figure 2b, residues with the strongest covariance signal are proximal in the 3D structure, suggesting that the model is effectively able to capture the pattern of the homologous sequences (see Supporting Information for details). A likelihood estimator is then used to calculate the likelihood, or probability (fitness score), of a sequence \( P(x) \), which is then linked to an evolutionary Hamiltonian energy \( E_{EH} \), as:

\[
E_{EH}(x) = -\ln P(x)
\]

Simulated annealing Monte Carlo simulations (Figure 2c) are employed to generate sequences with higher-order mutational statistics resembling the natural sequences of the PMEI protein family (Figure 2d). Sequence generation was achieved by randomly mutating a single amino acid at a time, with the mutation accepted or rejected based on the Metropolis criterion (see Supporting Information for details). Sequences featuring an \( E_{EH} \) lower than the starting sequence possibly translate into a higher fitness than the natural sequence, as suggested in the green region of Figure 2d showing that there are many sequences with lower predicted energy than the WT. The mutational landscape is therefore explored by keeping near the minimum described by the \( E_{EH} \).

To challenge the robustness of the fitness model, we selected sequences with high divergence from the wild type (WT). The fraction of sequence identity between designed and natural sequences is kept lower than \( \approx 70 \% \) (Table 1) by adding a penalty term proportional to the similarity of each generated sequence to that of the wild-type (WT) PMEI\(_{1xg2}\) (see Supporting Information for details).

To disentangle the interplay between PMEI fitness and functional behavior we envisioned two sequence designs: one where the binding interface (pink in Figure 1a) was not mutated within a radius of 0.6 nm from the interface of the binding PME, and one where every residue was allowed to mutate.

For the chosen PMEI (PEMI\(_{1xg2}\)), we found a large number of generated sequences to have lower evolutionary energy than the wild-type, suggesting that they may be more stable, in agreement with our earlier work on different protein folds. Before a full experimental characterization however, we estimated the expected thermal stability of the designed sequences, by obtaining an estimate of their unfolding time from molecular dynamics (MD) simulations.

The role of MD simulations in the designed pipeline is to filter the range of sequences identified in the first step, to
Table 1: Sequence properties of the designed proteins and activity obtained from differential scanning fluorimetry experiments and biochemical characterization.

<table>
<thead>
<tr>
<th>Mutants/WT</th>
<th>IDwt [%][a]</th>
<th>IDns [%][b]</th>
<th>$E_{\Delta V}(k_B T)$[c]</th>
<th>$T_m$ [°C][d]</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>–</td>
<td>–</td>
<td>−341</td>
<td>45.5</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variant 1</td>
<td>67</td>
<td>67</td>
<td>−351</td>
<td>–</td>
<td>Yes</td>
</tr>
<tr>
<td>Variant 2</td>
<td>69</td>
<td>69</td>
<td>−367</td>
<td>–</td>
<td>Yes</td>
</tr>
<tr>
<td>Variant 3</td>
<td>66</td>
<td>66</td>
<td>−368</td>
<td>49.1</td>
<td>Yes</td>
</tr>
<tr>
<td>Variant 4</td>
<td>70</td>
<td>70</td>
<td>−361</td>
<td>54.7</td>
<td>Yes</td>
</tr>
<tr>
<td>Variant 5</td>
<td>23</td>
<td>75</td>
<td>−456</td>
<td>–</td>
<td>Yes</td>
</tr>
<tr>
<td>Variant 6</td>
<td>24</td>
<td>79</td>
<td>−454</td>
<td>–</td>
<td>Yes</td>
</tr>
<tr>
<td>Variant 7</td>
<td>23</td>
<td>77</td>
<td>−432</td>
<td>–</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[a] Identity to the wild type. [b] Identity to closest natural sequence from the training data set. [c] Evolutionary energy. [d] Folding temperature.

The Experimental Assessment of the Designed PMEI Variants Shows the Retention of Functionality and an Increased Thermal Stability

The experimental characterization of the designed variants was obtained for 6 out of 7 designs as one of the sequences could not be expressed (variant 7). For variants 1–6 we assessed fitness in terms of functional efficiency and/or increased thermal stability, when we obtained variants with enough purity to confidently estimate the $T_m$ of the designs through biophysical methods. Thus, alongside the functional assessment of the designed sequences in terms of their ability to inhibit a target PME from *Citrus sinensis* (orange), when able to sufficiently purify the expressed variants, we experimentally assessed the melting temperature ($T_m$).

Biochemical assays were used to assess whether the designed variants can be considered as bona fide PMEIs. For this purpose, the selected variants were expressed in the yeast *Pichia pastoris* using codon-optimized sequences (Table S1). The choice of *Pichia pastoris* to produce the designed amino acidic sequences (Table S2) has the advantage to achieve high extraction yields of an expressed target, in line with the need to target high yields if this method is to be used for larger scale production of functional proteins. In this expression system, heterologous proteins can indeed be directly secreted in the growth medium.

Importantly, the expression system used here does not natively encode PMEIs so that PME inhibition of the exogenously expressed PMEIs can be undoubtedly obtained through a series of dilutions of the concentrated supernatants as showed in Figure 3a. Concentrated supernatants that correspond to growth media of centrifugated cultures of 6 out of 7 designed variants, together with the WT template PMEI (PEML15213773), showed to be efficiently inhibiting the target PME, albeit showing different potency (Figure 3a).

The halo present on the gel, which reveals de-methyl esterified pectin, is absent in every sample where the supernatant of a PMEI variant was mixed at dilutions of 1:5 v/v with the orange PME (see comparison of denaturated supernatant/DSN vs native supernatant/SN). However, var-
The thermal stability of the expressed variants was obtained by both differential scanning fluorimetry (DSF) and circular dichroism (CD). DSF confirmed that both variants 3 and 4 have a $T_m$ higher than PMEI$_{tag2}$ with estimated values of 49.1 °C and 54.7 °C, respectively, compared to 45.5 °C of the wild-type template (Figure 3d,e). The DSF profiles show a single thermal unfolding transition suggesting that eventual polydispersity of the sample due to glycosylated isoforms (Figure S3) are not differentially contributing to the estimated $T_m$. At 30 °C the CD signal of both variants 3 and 4 is the one typical of a-helical protein, as expected from the resolved structure of the target PMEI$_{tag2}$ showing that the designs retain a fold similar to the PMEI$_{tag2}$, while CD experiments performed at 90 °C show considerable loss of signal for PMEI$_{tag2}$, variants 3 and 4 (Figure 3e). Importantly, CD unfolding curves confirmed the trends uncovered by DSF, showing a higher $T_m$ for variant 3 and variant 4 compared to the WT (Figure S4).

**Molecular Dynamics Simulations Pinpoint the Determinants of Increased Thermal Stability**

An analysis of the number of contacts in the collected MD trajectories of the simulated designed and experimentally characterized sequences, showed that the design targeted non-conservative mutations of charged as well as hydrophobic and polar residues (Figure 4). From an analysis of the co-crystallized PME:PMEI complex, it can be seen that the complex is predominantly stabilized by intermolecular hydrogen bonds and four salt bridges. To gain insight into how stabilization of the PMEI monomer is achieved, we analyzed the interactions sampled in the MD simulations. An analysis of the intramolecular salt bridges sampled in the simulations revealed that variants 1–6 had all an increased
and hydrophobic interactions (Figure 4a). Interestingly, mutations targeting the insertion of polar amino acids mostly concentrate at the N-terminal end of the four-helix bundle, the furthest away from the binding interface.

Non-conservative mutations featuring the insertion of hydrophobic or charged amino acids mostly concentrate in the proximity of the four-helix bundle core or in helices α3 and α4 that fall in proximity of the binding interface (Figure 4b). Overall, while the total number of mutations for all the variants with unmutated binding interfaces is similar, the relative changes in salt-bridging, polar and hydrophobic residues is considerably different (Figure 4c).

**Conclusion**

In this work we were able to design functional PMEIs with an increased fitness measurable with a higher melting temperature. The obtained increases of up to \( \approx 10^\circ C \) in melting temperature would be expected to significantly improve the applicability of the designs for industrial applications. The effectiveness of a design obtained by exploring the sequence space has recently been incorporated within historically structure-based protocols,[32] with the development of tools expanding the scope of sequence search towards the understanding of protein folding.[33] Our study indicates that the fitness landscape only built on sequence information allows the design of functional and more thermostable sequences, which readily diverge from a template—a result that is much harder to achieve with purely structure-based protocols. Since the sequence space is enormously large, MD simulations can be applied as a further step to narrow down the number of sequences that can then be experimentally tested: enhancing the success rate of the design pipeline and reducing the cost for the experimental characterization of new variants.

MD simulations are known to be a computationally expensive method as they are commonly employed to investigate molecules at a high level of detail, with solvent molecules often explicitly accounted for. Nevertheless, the large advancements in computing power and GPU-acceleration make the scan of mean unfolding times for a large number of proposed protein variants, an easy target that, depending on computing capabilities, is an effort of a few days to a few weeks.

While sampling the stability of the designs via MD simulations depends from the available computational power, more generally, limitations deriving from the quality of the models for the designed variants are being overcome by the development of robust and evolutionary-based structure prediction strategies, such as AlphaFold.[3]

6 out 7 of the designed variants were successfully expressed and found to be active, suggesting a high success rate of our evolution-based workflow in creating new functional sequences. Assuming that the threshold of the evolutionary energy is at the value of the WT PME1_{1g2} (vertical dotted line in Figure 2d), the total amount of PME1-like sequences with higher fitness (shown in the green area of Figure 2d) are estimated to be 86[3] (see paragraph

![Figure 4. Contacts sampled in the molecular dynamics simulations of PME1_{1g2} and the designed variants. a) Probability density function of the number of salt bridges (left panel), hydrophobic contacts (central panel) and hydrogen bonds (right panel). b) The Cα carbons showing the position of the substitutions in the modelled structures are shown as red and blue spheres for negatively and positively charged residues, respectively. Hydrophobic and polar residues substitutions are colored black and green respectively. PMEI structures are shown as ribbons. c) Number of mutations in all the designed variants for positively (blue) and negatively (red) charged, hydrophobic (black) and polar (green) amino acids. The total number of mutations in the design is shown by the magenta bar.](image-url)
on estimation sequence entropy in the Supporting Information): this suggests that there are plenty of sequences to explore for improved function and stability. The high sequence similarity of the designs in this work is a result of the simulated annealing driving the sequences towards the global energy minimum (fitness maximum). We are mainly interested in the sequences with the lowest energy since they are more likely to be stable and have high fitness. To avoid sequences being too similar we imposed a maximum of 85% sequence identity between designs. These designs only occupy very narrow sequence space in contrast to the vast space of all the possible sequences. In future work, we will explore the robustness of the method by applying a lower sequence similarity limit.

While the ability to apply designed sequences in industrial workflows is yield-limited, a limitation to the discovery of new active and more thermally stable sequences is ultimately linked to expressing and purifying designs at an appreciable level of purity. The expression system proposed here, together with the protocol used to direct the expressed sequence in the expression medium, aims at maximizing production yields to increase batch production. Further development of the design pipelines could be then concentrated on optimizing the presence of specific sequence motifs that could improve the success rate of heterologous expression in systems such as <em>Pichia pastoris</em>.

Considering the peculiar ability of eukaryotic expression systems to modify recombintant constructs with complex post-translational modifications, a potential development of the design pipeline could move towards the incorporation of additional energy terms to favor or disfavor the introduction of specific motifs in the designed sequences. Moreover, with the possibility to explicitly consider the coevolutionary information between PME and PMEI, this method can be further optimized to design sequences with higher binding affinity by sampling the sequence space of PMEI conditioned on the given sequence of PME. Our framework also allows to implement the deliberate insertion or avoidance of glycosylation sites or other post-translational modifications that can increase the stability of a targeted fold, to achieve the desired stability of functional proteins and generate sequences that are fitter for harsher micro-environments.

**Acknowledgements**

DM thanks the School of Chemical Sciences of The University of Auckland, the Nectar Cloud computing infrastructure, and the New Zealand eScience high-performance computing infrastructure (NeSI) for financial and logistic support. This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD (http://biowulf.nih.gov). JP gratefully acknowledges the funding from the Agence Nationale de la Recherche (ANR) and the “Région Hauts de France”. Open Access publishing facilitated by The University of Auckland agreement via the Council of Australian University Librarians.

**Conflict of Interest**

The authors declare no conflict of interest.

**Data Availability Statement**

The data that support the findings of this study are available from the corresponding author upon reasonable request.

**Keywords:** Coevolution · Molecular Dynamics Simulations · Monte Carlo Simulations · Potts Models · Protein Design