Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability

Graphical Abstract

Highlights

- A new computational method is used to stabilize five recalcitrant proteins
- Designed variants show higher expression and stability with unmodified function
- A designed human acetylcholinesterase variant expresses solubly in bacteria
- The method is fully automated and implemented on a webserver

Authors

Adi Goldenzweig, Moshe Goldsmith, Shannon E. Hill, ..., Joel L. Sussman, Dan S. Tawfik, Sarel J. Fleishman

Correspondence
dan.tawfik@weizmann.ac.il (D.S.T.), sarel@weizmann.ac.il (S.J.F.)

In Brief

Heterologous expression of proteins and their mutants often results in misfolding and aggregation. Goldenzweig et al. (2016) developed an automated algorithm for protein stabilization requiring minimal experimental testing; for instance, the five tested variants of human acetylcholinesterase showed ≥100-fold higher soluble bacterial expression and higher melting temperatures than wild-type.

Goldenzweig et al., 2016, Molecular Cell 63, 337–346
July 21, 2016 © 2016 The Author(s). Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.molcel.2016.06.012
Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability

Adi Goldenzweig,1 Moshe Goldsmith,1 Shannon E. Hill,2 Or Gertman,3 Paola Laurino,1 Yacov Ashani,1,4 Orly Dym,4 Tamar Unger,4 Shira Albeck,4 Jaime Prilusky,4 Raquel L. Lieberman,2 Amir Aharoni,3 Israel Silman,6 Joel L. Sussman,4 Dan S. Tawfik,1,2 and Sarel J. Fleishman1,4
1Department of Biomolecular Sciences, Weizmann Institute of Science, Rehovot 7610001, Israel
2School of Chemistry & Biochemistry, Georgia Institute of Technology, Atlanta, GA 30332-0400, USA
3Department of Life Sciences, Ben-Gurion University of the Negev, P.O.B. 653, Beer-Sheva 8410501, Israel
4Israel Structural Proteomics Center, Weizmann Institute of Science, Rehovot 7610001, Israel
5Bioinformatics & Biological Computing Unit, Weizmann Institute of Science, Rehovot 7610001, Israel
6Department of Neurobiology, Weizmann Institute of Science, Rehovot 7610001, Israel
*Correspondence: dan.tawfik@weizmann.ac.il (D.S.T.), sarel@weizmann.ac.il (S.J.F.)

http://dx.doi.org/10.1016/j.molcel.2016.06.012

SUMMARY

Upon heterologous overexpression, many proteins misfold or aggregate, thus resulting in low functional yields. Human acetylcholinesterase (hAChE), an enzyme mediating synaptic transmission, is a typical case of a human protein that necessitates mammalian systems to obtain functional expression. We developed a computational strategy and designed an AChE variant bearing 51 mutations that improved core packing, surface polarity, and backbone rigidity. This variant expressed at ~2,000-fold higher levels in E. coli compared to wild-type hAChE and exhibited 20°C higher thermostability with no change in enzymatic properties or in the active-site configuration as determined by crystallography. To demonstrate broad utility, we similarly designed four other human and bacterial proteins. Testing at most three designs per protein, we obtained enhanced stability and/or higher yields of soluble and active protein in E. coli. Our algorithm requires only a 3D structure and several dozen sequences of naturally occurring homologs, and is available at http://pross.weizmann.ac.il.

INTRODUCTION

Most natural proteins are only marginally stable (Magliery, 2015). Thus, taken out of their natural context, through either overexpression, heterologous expression, or changes in environmental conditions, many proteins misfold and aggregate. The most general origin of overexpression challenges is low stability of the protein’s native, functional state relative to alternative nonfunctional or aggregation-prone states. By designing variants with more favorable native-state energy, yields of soluble and functional protein obtained by heterologous overexpression can be dramatically increased, alongside other merits such as longer storage and usage lifetimes and enhanced engineering potential.

Engineering stable protein variants is a widely pursued goal. Methods based on phylogenetic analysis (Lehmann et al., 2000; Steipe et al., 1994) and structure-based rational or computational design (Borgo and Havranek, 2012; Jacak et al., 2012; Korkegian et al., 2005; Lawrence et al., 2007) yielded proteins with improved stability and higher functional expression (Magliery, 2015). Individual mutations, however, contribute little to stability (typically ≤1 kcal/mol) (Zhao and Arnold, 1999), whereas stabilizing large and poorly expressed proteins typically requires many mutations. However, since even a single severely destabilizing mutation can undermine the benefit accruing from all others, high prediction accuracy is essential. Despite improvements in accuracy, existing approaches have a relatively high probability of inadvertently introducing disruptive mutations (false-positive predictions) (Borgo and Havranek, 2012; Magliery, 2015; Potapov et al., 2009). Published efforts to stabilize large proteins therefore either incorporate only a few predicted stabilizing mutations (typically ≤4) at each experimental step or use library approaches to identify optimal combinations of stabilizing mutations (Khersonsky et al., 2011; Sullivan et al., 2012; Trudea et al., 2014; Whitehead et al., 2012; Wijma et al., 2014). Such approaches are laborious and impractical for proteins without established medium-to-high throughput screens, let alone for proteins of unknown function. To address the demand for stabilizing large, recalcitrant proteins by a wide range of researchers, who lack background in computational design, we developed an automated algorithm based on atomistic Rosetta modeling and phylogenetic sequence information. We specifically aimed to develop a general method that minimizes false-positive predictions to ensure that only a few variants need to be experimentally tested to achieve high functional yields (ideally, just one variant). We applied this algorithm to four different enzymes and one protein of unknown function. In each case, up to five variants were designed as the default output, encoding from 9 to 67 mutations relative to wild-type. These
variants exhibited enhanced bacterial expression yields and stability, without sacrificing or altering activity.

DESIGN

Computational Mutation Scanning Minimizes False-Positive Mutations

To address the challenge of designing variants with a large net stabilizing effect, yet without modifying their function, our computational process began by scanning the natural sequence diversity. For every target wild-type sequence, we generated a sequence alignment, from which we computed a position-specific substitution matrix (PSSM) (Altschul et al., 2009); the PSSM represents the log-likelihood of observing any of the 20 amino acids at each position. At every amino acid position, “allowed” mutations were defined as those with a favorable PSSM score (≥0). The rationale for restricting the allowed sequence space through the alignment scan is that, in general, deleterious mutations are purged by natural selection; purging is not absolute, however, and less favorable amino acids may occur at certain positions. Indeed, mutation to the most frequently observed amino acid often increases stability (the consensus effect) (Lehmann et al., 2000; Magliery, 2015; Steipe et al., 1994). However, our approach does not implement the consensus per se. Foremost, the alignment scan eliminates mutations that are rare or never seen in the natural diversity, rather than strictly selecting the most frequent amino acid. Next, we applied Rosetta computational mutation scanning (Whitehead et al., 2012), wherein each “allowed” mutation from the previous step was singly modeled against the background of the wild-type structure, and the energy difference between the wild-type and the single-point mutant was calculated ($\Delta G_{\text{calc}}$). By computing the effects of each mutation singly, rather than in combination with others, we restricted design choices to mutations that were likely to make additive contributions to stability, rather than nonadditive and context-dependent contributions, thereby minimizing the risk of false positives. We thereby defined the space of “potentially stabilizing” mutations as mutations with $\Delta G_{\text{calc}} \leq -0.45$ Rosetta energy units (R.e.u.); we chose this cutoff rather than 0 R.e.u. to further lower the risk of false positives, as indicated by the systematic evaluation described below. As a final step, we used Rosetta combinatorial sequence design to find an optimal combination of mutations within the space of potentially stabilizing mutations. Additionally, we used lower $\Delta G_{\text{calc}}$ cutoffs to select potentially stabilizing mutations prior to combinatorial sequence design, thereby generating several designs for experimental testing (see Data S1 available online).

The choice of mutations at Gly416 in human acetylcholinesterase (hAChE) illustrates the role of these two filters (alignment scan and computational mutation scanning) in pruning false positives (Figure 1A). Position 416 is located on a partially exposed helical surface, where the small and flexible amino acid Gly is likely to destabilize hAChE. Indeed, in the alignment of AChE homologs, Gly is infrequent, and His is the most prevalent amino acid. Modeling shows, however, that in the specific context of hAChE, His adopts a strained side-chain conformation; in contrast, Gln, the third most prevalent amino acid, is predicted to be most stabilizing owing to its high helical propensity and favorable hydrogen bonding with Tyr503. The combined filter therefore favors Gln over His for downstream design calculations.

To systematically evaluate the ability of our filtering method to identify stabilizing mutations, we compared its output to published experimental data on the stability effects of single-point mutations in the enzymes fungal endoglucanase Cel5A and yeast triosephosphate isomerase (TIM) (Figure 1B; Table 1) (Sullivan et al., 2012; Trudeau et al., 2014). The two filters (PSSM and Rosetta mutational scanning using $\Delta G_{\text{calc}} \leq -0.45$ R.e.u. as cutoff) eliminated all the severely destabilizing mutations (and 99.6% of all destabilizing mutations) and retained one-third and two-thirds of the experimentally verified stabilizing
mutations in Cel5A and TIM, respectively. Thus, for both enzymes, our method identified a large fraction of the known stabilizing mutations, while excluding false positives entirely. The energy cutoff emphasizes the importance of minimizing false-positive predictions, that is, mutations that are experimentally destabilizing but are computationally assessed as favorable; for instance, in the case of Cel5A, using $\Delta G_{\text{calc}} \leq 0$ R.e.u. as cutoff would correctly predict four additional stabilizing mutations but would introduce eight more destabilizing mutations. False-negative predictions—that is, mutations that are stabilizing in experiment but are considered destabilizing by modeling—could not be attributed to a single modeling error (Table S1). Specifically, out of 22 false negatives in Cel5A, eight were filtered because they are unlikely according to the sequence alignment (PSSM score <0), and four showed $\Delta G_{\text{calc}}$ in the range $-0.45$–$0$ R.e.u., thereby missing the energy threshold by a small margin. Of the ten remaining false negatives, three occurred at core positions, where energy calculations typically penalize mutations, and two were mutations to proline. Improvements in the energy function and the conformation–relaxation procedure may in the future increase accuracy, and the analysis reported here provides a benchmark for such improvements.

**RESULTS**

**Designed hAChE with Nearly 2,000-Fold Higher Bacterial Expression and Intact Activity**

We applied our design strategy to human acetylcholinesterase (hAChE), a 60 kDa enzyme responsible for terminating synaptic transmission at cholinergic synapses by rapidly hydrolyzing the neurotransmitter acetylcholine (ACh) (Sussman et al., 1991). Although AChE is potentially useful for detection and detoxification of organophosphate nerve agents and pesticides, hAChE is currently produced by costly procedures involving purification from blood erythrocyte membranes, or using plant or mammalian expression systems. Previous attempts to express hAChE in bacteria resulted in extremely low levels of soluble and active protein (Fischer et al., 1993). AChE’s active site is located at the bottom of a deep gorge that penetrates half way (20 Å) into the enzyme, and mutations along the gorge can reduce ACh hydrolysis rates by up to 1,000-fold (Ordentlich et al., 1995). To increase the stability and expression levels of hAChE without altering its activity, we imposed a further restriction on the allowed sequence space of the newly designed hAChE: in all Rosetta modeling simulations, the side-chain conformations of amino acids within 8 Å of the reversible inhibitor E2020, which spans the full length of the active-site gorge (Cheung et al., 2012), had to remain as in the native hAChE structure (Table S2). The latter restriction, in combination with the two above-described filters, dramatically reduced the sequence space available for design. The theoretical sequence space for hAChE, a 550-residue enzyme, is $10^{700}$, a formidable number inaccessible even for advanced modeling algorithms. The size of the reduced sequence space, by contrast, was $10^{31}$ sequences, equivalent to complete computational design of a 24-amino acid peptide, a challenge solved already in the 1990s (Dahiyat and Mayo, 1997). We noted that the reduced sequence space led to convergence of combinatorial sequence optimization to identical, or nearly identical, sequences for any given $\Delta G_{\text{calc}}$ cutoff; this convergence, which is not usual in computational design (Fleishman et al., 2011), is a prerequisite for reproducibility and usage by nonexperts.

Given AChE’s large size, we designed five alternatives using different $\Delta G_{\text{calc}}$ thresholds, with 17–67 mutations relative to hAChE (Table S3; Data S2), and subjected them to experimental testing. Synthetic genes encoding wild-type hAChE and the five designs were optimized for bacterial translation efficiency, fused to the C terminus of thioredoxin, and expressed in E. coli Shufflet7express cells to facilitate disulfide-bond formation. In SDS-PAGE gels of supernatant fractions from bacterial lysates, the AChE band overlapped with other bands, precluding visual quantification. Nevertheless, because the designed variants’ specific activity was found to be nearly identical to the wild-type’s (Table 2), we could quantify the relative levels of soluble and active enzyme by measuring AChE activity in crude lysates. The five designs showed $\geq 100$-fold higher lysate ACh-hydrolysis rates compared to hAChE, with the best design, dAChE4 (Figure 2A; 51 mutations), exhibiting an almost 2,000-fold higher rate (Figure 2B). Furthermore, due to its extremely low soluble expression, hAChE could not be purified from crude cell lysates, whereas the designed variants were readily purified to homogeneity, with dAChE4 yields in standard shaker flasks approaching 2 mg protein per liter of bacterial culture (Table S4).

The designed AChE mutations are scattered throughout the enzyme and show typical characteristics of stabilizing mutations, including improved core packing, higher backbone rigidity, and increased surface polarity (Figure 2A). In agreement with the design strategy and the higher levels of soluble and functional enzyme, we observed increased resistance to heat inactivation of up to 20°C relative to the E. coli-expressed hAChE (Figure 2C) and to hAChE expressed in mammalian cells (Table 2). The designs hydrolyzed ACh at rates that are within a 2-fold margin of hAChE rates, and displayed inactivation-rate constants by the nerve agent VX that are nearly identical to hAChE (the largest deviation was observed for dAChE3, which exhibited a 2.5-fold lower inactivation rate; Table 2).

The above observations of nearly identical activity profiles of the designed and wild-type AChEs suggested that the designed enzymes’ active sites are essentially identical to that of hAChE. To verify this, we conducted crystallization trials on dAChE4, the design exhibiting the highest bacterial-expression yields. We noted that, in contrast to various natural AChEs studied by us, large crystals formed already within a few days and more…

---

**Table 1. Computational Mutation Scanning versus Experimental Point Mutations in Fungal Endoglucanase 5A**

<table>
<thead>
<tr>
<th>Mutations</th>
<th>Stabilizing</th>
<th>Destabilizing</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>True predictions</td>
<td>12 (35%)</td>
<td>230 (99.6%)</td>
<td>242</td>
</tr>
<tr>
<td>False predictions</td>
<td>22 (65%)</td>
<td>1 (0.4%)</td>
<td>23</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>231</td>
<td>265</td>
</tr>
</tbody>
</table>

*aMutations were experimentally assayed by heat inactivation (Trudeau et al., 2014; raw data kindly provided by D. Trudeau). bMutations were predicted as stabilizing if they exhibited $\Delta G_{\text{calc}} < -0.45$ R.e.u. and destabilizing otherwise.*
reproducibly. dAChE4’s structure was solved at 2.6Å resolution, thus yielding, to the best of our knowledge, the first structure of an AChE expressed in a prokaryote (Table S5). dAChE4’s structure is very similar to that of wild-type hAChE, with a 0.7 Å root-mean-square deviation (rmsd) over Cα atoms. Residues at the catalytic gorge aligned particularly well, with an all-atom rmsd of only 0.125Å (Figure 2D). Thus, despite 51 mutations relative to hAChE (dAChE1), our designed mutant can be used to reproducibly target the AChE active site.

Other Designed Enzymes Show High Stability and/or Soluble Expression

We applied our algorithm to two other human enzymes—the histone deacetylase SIRT6, the human DNA methyltransferase Dnmt3a, and a bacterial phosphotriesterase dubbed PTE. Being the only enzyme known to have actually evolved to degrade organophosphates (OPs), PTE is of considerable biotechnological potential for decontamination and detoxification of OPs, including nerve agents. However, PTE detoxifies most nerve agents at rates too slow for practical applications. Wild-type PTE was previously engineered for higher expression (Roodveldt and Tawfik, 2005) and hydrolysis rates toward various OPs (Bigley et al., 2015; Cherry et al., 2013). Introduction of functionaltering mutations, however, destabilized the enzyme, as is often the case for laboratory-evolved enzymes (Tokuriki et al., 2008); thus, stabilization is a prerequisite for further engineering (Bloom et al., 2009). We designed three alternatives using different ΔGcat thresholds, encoding 9–28 mutations relative to wild-type PTE (Table S3), and subjected them to experimental testing. Synthetic genes encoding the three designs and PTE-S5, a published variant encoding three mutations, which displays ~20-fold higher expression levels compared to wild-type (Roodveldt and Tawfik, 2005), were optimized for bacterial translation efficiency, fused to maltose-binding protein, and expressed in E. coli GG448 cells. All three designs displayed increased levels of soluble, functional enzyme compared to PTE-S5, which already displays higher expression levels than wild-type (Table 3). Two of the three designs showed ~10°C higher tolerance to heat inactivation relative to PTE-S5, with no significant change in activity toward the OP substrate paraoxon (Table 3; Figure S2A). Another noteworthy outcome of stabilization design was increased metal affinity. PTE is a metalloenzyme bearing two active-site metals, typically Zn^{2+} (Benning et al., 2001). Directed evolution of wild-type PTE for higher expression (PTE-S5) led to a significant decrease in metal affinity—a major practical drawback for applications in conditions in which Zn^{2+} cannot be supplemented. The designed variant dPTE2, which contained 19 mutations and exhibited the highest tolerance to heat inactivation, also exhibited a marked increase in metal affinity compared to PTE-S5, approaching the affinity of wild-type PTE (Figures 3A and S2B; Table 3).

SIRT6 is an ADP-ribosylase and NAD^{+}-dependent deacetylase that removes acyl groups from acylated lysines, thereby regulating several essential cellular processes (Kugel and Mostoslavsky, 2014). The low bacterial-expression levels of human SIRT6 (hSIRT6), and its weak deacetylation activity relative to SIRT1, limit its study. The designed variant (dSIRT6) contained 11 mutations relative to hSIRT6 (Tables S2 and S3). It was expressed in E. coli and purified to homogeneity with yields of 20 mg per liter culture, an ~5-fold increase relative to hSIRT6, and its deacetylase activity was 60% higher than hSIRT6’s (Figures S3A and S3B). We subsequently incorporated the design mutations on the background of an engineered hSIRT6 variant dubbed E1, which contains three mutations relative to hSIRT6 (our unpublished data). E1 exhibits increased deacetylation activity compared to hSIRT6 but 3-fold lower expression levels in human cell lines. In contrast, the designed variant dE1 recapitulates hSIRT6’s expression levels in human cell lines while maintaining high deacetylation activity (Figure 3B). Given the beneficial effects of SIRT6 overexpression on longevity in mice (Kanfi et al., 2012), our designed mutant can be used to

<table>
<thead>
<tr>
<th>AChE Variant</th>
<th>Mutat°</th>
<th>Normalized Activity°</th>
<th>Inactivation Temperature (°C)</th>
<th>AChE Hydrolysis $k_{cat}/K_m$ × 10⁹ (M⁻¹ min⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>hAChE (HEK293)</td>
<td>–</td>
<td>–</td>
<td>50.6 ± 0.3</td>
<td>7.92 ± 0.15</td>
</tr>
<tr>
<td>hAChE (bacterial)</td>
<td>1</td>
<td>44 ± 0.4</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>dAChE1</td>
<td>17</td>
<td>119 ± 20</td>
<td>60.5 ± 0.4</td>
<td>6.48 ± 0.71</td>
</tr>
<tr>
<td>dAChE2</td>
<td>30</td>
<td>280 ± 40</td>
<td>61.5 ± 0.5</td>
<td>2.65 ± 0.52</td>
</tr>
<tr>
<td>dAChE3</td>
<td>42</td>
<td>308 ± 44</td>
<td>62.3 ± 0.2</td>
<td>0.177 ± 0.010</td>
</tr>
<tr>
<td>dAChE4</td>
<td>51</td>
<td>1770 ± 258</td>
<td>62.3 ± 0.3</td>
<td>0.071 ± 0.007</td>
</tr>
<tr>
<td>dAChE5</td>
<td>67</td>
<td>637 ± 134</td>
<td>61.1 ± 0.2</td>
<td>0.104 ± 0.010</td>
</tr>
</tbody>
</table>

°Number of amino acid mutations relative to wild-type hAChE.
°Activity in crude lysates of cells expressing the AChE variants from 250 ml E. coli cultures.
°The temperature at which 50% of activity was retained. Enzyme samples were incubated for 30 min at varying temperatures and tested for AChE activity after cooling.
°Not determined.
Figure 2: Design of a Stable hAChE variant and Its Functional Expression in Bacteria

(A) The structural underpinnings of stabilization in the designed variant dAChE4. Wild-type hAChE is shown in blue and 51 mutated positions, which are distributed throughout dAChE4, are indicated by orange spheres. Thumbnails highlight stabilizing effects of selected mutations.

(B) Bacterial lysate activity levels of designed AChEs normalized to hAChE activity. Crude lysates were derived from 250 ml flasks (medium scale) or 0.5 ml E. coli cultures grown in a 96-well plate (small scale). The higher activity levels in the designed variants reflect higher levels of soluble, functional enzyme. (C) Designed AChE variants (colored lines) show higher resistance to heat inactivation compared to hAChE (black). Residual activities following incubation at different temperatures were measured in bacterial lysates and normalized to the activity in nontreated lysates.

See also Figure S1 and Tables S2–S5.

establish a correlation between SIRT6’s deacetylase rates and its physiological function.

In the case of human Dnmt3a (hDnmt3a), soluble and active fractions of the human enzyme can be obtained by E. coli expression, but enzyme activity is very low. We tested one design dDnmt3a containing 14 mutations relative to hDnmt3a (Tables S2 and S3). In contrast to all other designs, the designed dDnmt3a showed significantly lower expression levels than hDnmt3a in E. coli (Figures S3C and S3D). Nonetheless, it exhibited nearly 10-fold higher specific activity (Figure 3C), suggesting that dDnmt3a preparations contain a significant fraction of soluble but inactive or poorly active enzyme. Therefore, although Dnmt3a failed to meet our design method’s success criteria (higher stability and soluble expression), it significantly increased the protein’s functional yield.

A Webserver for Protein Stabilization

Encouraged by the consistently positive results presented above, we implemented the algorithm as a webserver, called the Protein Repair One Stop Shop (PROSS, http://pross.weizmann.ac.il). Following their request for assistance in solving a critical expression and stability question regarding the human myocilin OLF domain (hMyoc-OLF), two of the authors (S.E.H. and R.L.L.) were granted unsupervised access to the webserver. OLF domains are found in extracellular proteins of multicellular organisms. A number of OLF domains have been implicated in human disease, but the function(s) of most OLF domains remains elusive. The best-studied OLF domain is that of hMyoc-OLF, in which more than 100 nonsynonymous mutations are implicated in inherited forms of open-angle glaucoma. Mutations documented in patients lead to destabilized myocilin protein that forms cytotoxic aggregates (Burns et al., 2010; Yam et al., 2007). S.E.H. and R.L.L. therefore posited that mutations that confer enhanced stability might reduce hMyoc-OLF’s propensity to misfold at physiological temperatures.

Three hMyoc-OLF structures (PDB: 4WXQ, 4WXS, and 4WXU) were submitted to PROSS, producing seven designs for each structure with 5–25 mutations each. From all suggested point mutations, S.E.H. and R.L.L. manually derived one variant (dMyoc-OLF) comprising 21 mutations. Although hMyoc-OLF binds Ca^{2+} (Donegan et al., 2012), the physiological role of Ca^{2+} binding is still unknown; nevertheless, the Ca^{2+}-binding positions remained unchanged, reflecting the high sequence and structural constraints acting on them. Furthermore, none of the design mutations have been implicated in human disease,
Table 3. Stability and Kinetic Parameters of PTE Variants

<table>
<thead>
<tr>
<th>Variant</th>
<th>Mut&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Normalized Activity&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Inactivation Temperature (°C)</th>
<th>k&lt;sub&gt;cat&lt;/sub&gt; × 10&lt;sup&gt;d&lt;/sup&gt; (min&lt;sup&gt;-1&lt;/sup&gt;)</th>
<th>k&lt;sub&gt;cat&lt;/sub&gt;/K&lt;sub&gt;M&lt;/sub&gt; × 10&lt;sup&gt;d&lt;/sup&gt; (min&lt;sup&gt;-1&lt;/sup&gt;M&lt;sup&gt;-1&lt;/sup&gt;)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTE-S5</td>
<td>3 1</td>
<td></td>
<td>Lysate: 50.9 ± 0.7, 52.4 ± 0.2; Purified: 7.5 ± 0.3</td>
<td>0.101 ± 0.023</td>
<td>0.970 ± 0.076</td>
</tr>
<tr>
<td>dPTE1</td>
<td>9 2.0</td>
<td></td>
<td>ND&lt;sup&gt;c&lt;/sup&gt;</td>
<td>ND&lt;sup&gt;c&lt;/sup&gt;</td>
<td>ND&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>dPTE2</td>
<td>19 6.1</td>
<td></td>
<td>59.2 ± 0.7, 62.0 ± 0.2</td>
<td>0.060 ± 0.014</td>
<td>0.70 ± 0.05</td>
</tr>
<tr>
<td>dPTE3</td>
<td>28 2.3</td>
<td></td>
<td>47.0 ± 1.3</td>
<td>ND&lt;sup&gt;c&lt;/sup&gt;</td>
<td>ND&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>a</sup>Number of mutations relative to wild-type PTE.
<sup>b</sup>Fold increase in activity in crude E. coli lysates relative to PTE-S5.
<sup>c</sup>Chelator, 50 μM 1,10-phenanthroline.
<sup>d</sup>Kinetic parameters for paraoxon.
<sup>e</sup>ND, not determined.

supporting the notion that the disease-associated mutations are destabilizing.

dMyoc-OLF and hMyoc-OLF were expressed in E. coli fused to maltose-binding protein (MBP-OLF). The designed variant gave an order of magnitude higher yield than hMyoc-OLF (Table S6). In addition, whereas the purified hMyoc-OLF comprises a mixture of a soluble MBP-OLF-aggregate and the properly folded monomer (Burns et al., 2010), the designed variant was almost exclusively a soluble MBP-OLF-aggregate and the properly folded monomer.

In addition, whereas the purified hMyoc-OLF comprises a mixture of a soluble MBP-OLF-aggregate and the properly folded monomer (Burns et al., 2010), the designed variant was almost exclusively a soluble MBP-OLF-aggregate and the properly folded monomer. The designed variant gave an order of magnitude higher yield than hMyoc-OLF (Table S6). In addition, whereas the purified hMyoc-OLF comprises a mixture of a soluble MBP-OLF-aggregate and the properly folded monomer (Burns et al., 2010), the designed variant was almost exclusively a soluble MBP-OLF-aggregate and the properly folded monomer.

DISCUSSION

Most existing stability-design methods focus on one element or another of protein stability. For instance, some methods introduce disulfides or prolines, while others optimize core packing (Borgo and Havranek, 2012; Korkegian et al., 2005) and increase surface polarity (Magliery, 2015) and charge (Lawrence et al., 2007). Our design algorithm, by contrast, selects for all amino acid mutations that optimize the protein’s computed energy, subject to constraints inferred from homologous sequences. It thereby generates designs that improve a range of molecular parameters that are associated with stability (Table 4) while preserving function. We find, furthermore, that in the various proteins studied here, the algorithm implemented solutions that appear to selectively tackle specific defects of each protein. For instance, 17 of the 51 mutations in design dACHE4 impact core positions, a very large number compared to previous design studies that targeted protein cores (Borgo and Havranek, 2012; Korkegian et al., 2005). Indeed, natural AChE structures contain unusually large core cavities (Koellner et al., 2000). In design dPTE2, by contrast, only two core mutations were implemented, and the net negative charge increased by 13 units, shifting the isoelectric point (pI) from a near-neutral value of 6.7 for the wild-type to 5.0; indeed, near-neutral pl is often associated with poor solubility, and surface “supercharging” has been used by others to improve thermal stability (Lawrence et al., 2007). The stability considerations, however, do not come at the expense of functional constraints. In Dnmt3a, for example, the positive net charge and basic pI are fully conserved as expected for a DNA-binding protein. We conclude that the combination of evolutionary constraints, selection of mutations that individually contribute to computed stability, and combinatorial sequence optimization within the space of these mutations can address a broad range of stability defects in various proteins without affecting their original molecular functions.

The case studies discussed here represent a diverse group of proteins belonging to different fold families (TIM barrel, β-propeller, and α/β hydrolase fold) and displaying different activities. In each case, the wild-type protein and variants previously engineered using traditional methods suffer from low stability, as manifested by intolerance to high temperature and function-modifying mutations or by low solubility, heterologous expression, cofactor affinity, or specific activity. That a single fully automated method that only considers phylogenetic constraints and optimizes the native state’s energy is able to address these diverse challenges without sacrificing function demonstrates that the underlying source of these problems is the native state’s marginal stability. Furthermore, the stability-design method has the potential to optimize proteins used as scaffolds for de novo binder or enzyme design, where low stability and expression levels have been limiting (Fleishman et al., 2011; Khersonsky et al., 2011).

Limitations

The stability design algorithm requires a few dozen unique sequence homologs and an atomic structure of the target protein. With the burgeoning of sequence databases, the first requirement is likely to be met with few exceptions; the second, however, may be an impediment, particularly since low protein yields and stability negatively impact structure determination. To expand beyond experimentally determined structures, homology models may be used, and it remains to be seen whether they provide sufficient accuracy for atomistic stability-design calculations.

EXPERIMENTAL PROCEDURES

Computational Procedures

Phylogenetic Sequence Constraints

For every query sequence, homologous sequences were collected using the BLASTP algorithm (Altschul et al., 1990) on the nonredundant (nr) database,
with the maximum number of hits set to 500 and the e value to 10^-4. All other parameters were set to default values. Hits were excluded if they covered less than 60% of the query sequence or if their sequence identity to the query sequence was lower than 30% (28% for PTE due to the low diversity of homologs at >30% sequence identity). The remaining sequences were clustered using cd-hit (Li and Godzik, 2006), with a clustering threshold of 90% and default parameters.

MUSCLE (Edgar, 2004) was used with default parameters to derive a multiple sequence alignment (MSA) from the clustered sequences. Gaps in the alignment between the query and any homolog often occur within loop regions and may reflect differences in the local backbone conformation. To reduce alignment uncertainty, we detected secondary-structure elements in the query protein structure (using DSSP [Kabsch and Sander, 1983]) and eliminated subsequences in homologs that contained gaps in loop segments that intervene between any two secondary-structure elements relative to the query. In effect, for every loop region this procedure generated a specific alignment that only comprised homologous sequences with no insertions or deletions relative to the query. Since this feature of the stability-design algorithm had not yet been implemented when we designed AChE and Dnmt3a, these two proteins were designed with default MUSCLE alignments, containing information from all homologs in loop regions.

Given the alignments of each query sequence, we computed a position-specific scoring matrix (PSSM) using the PSI-BLAST algorithm (Altschul et al., 2009). The PSSM represents the log probability of observing each of the 20 amino acids at each position in the query. Non-negative PSSM scores are considered likely to occur in evolution and define the space of allowed mutations.

Structure-Based Constraints
To prevent activity loss, residues within 8 Å of small-molecule ligands observed in the active site (hAChE and Dnmt3a), or within 5 Å of metal cofactors (PTE), DNA chains (Dnmt3a) and protein ligands (SIRT6) were held fixed throughout all Rosetta simulations. In homo-oligomer structures (PTE and AChE), residues within 5 Å of the oligomer interface were held fixed. For PTE and SIRT6, some of the active-site constraints were defined using an alignment to another, ligand-bound structure, namely residues within 8 Å of DPJ in 4NP7 for PTE, and residues within 5 Å of chain F in 3ZG6 for SIRT6
Table 4. Structure and Sequence Features of Designed Variants Compared to Their Wild-Type Counterparts

<table>
<thead>
<tr>
<th>Design Variant</th>
<th>aa</th>
<th>Mut</th>
<th>Core</th>
<th>Salt Bridges</th>
<th>H Bonds</th>
<th>X → Pro</th>
<th>Charge N/P</th>
<th>pl(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>dACHES1</td>
<td>17 (3%)</td>
<td>9 (53%)</td>
<td>3</td>
<td>14</td>
<td>1/2</td>
<td>5.5</td>
<td>5.5</td>
<td></td>
</tr>
<tr>
<td>dACHES2</td>
<td>30 (6%)</td>
<td>15 (50%)</td>
<td>3</td>
<td>11</td>
<td>4/4</td>
<td>5.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dACHES3</td>
<td>42 (8%)</td>
<td>16 (38%)</td>
<td>5</td>
<td>18</td>
<td>6/6</td>
<td>5.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dACHES4</td>
<td>51 (10%)</td>
<td>17 (33%)</td>
<td>3</td>
<td>10</td>
<td>7/8</td>
<td>5.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dACHES5</td>
<td>67 (13%)</td>
<td>28 (28%)</td>
<td>10</td>
<td>15</td>
<td>7/9</td>
<td>5.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dpTE1</td>
<td>9 (3%)</td>
<td>0 (0%)</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>6/-2</td>
<td>5.3</td>
<td>6.7</td>
</tr>
<tr>
<td>dpTE2</td>
<td>19 (6%)</td>
<td>2 (11%)</td>
<td>8</td>
<td>9</td>
<td>11/-2</td>
<td>5.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dpTE3</td>
<td>28 (9%)</td>
<td>18 (18%)</td>
<td>12</td>
<td>7</td>
<td>16/-3</td>
<td>4.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dSIRT6</td>
<td>280 (11%)</td>
<td>1 (9%)</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>0/0</td>
<td>8.3</td>
<td>8.3</td>
</tr>
<tr>
<td>dDmnt3</td>
<td>272 (14%)</td>
<td>0 (0%)</td>
<td>4</td>
<td>7</td>
<td>1</td>
<td>1/1</td>
<td>8.4</td>
<td>8.6</td>
</tr>
<tr>
<td>dMyc-OLF</td>
<td>259 (21%)</td>
<td>2 (10%)</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>-2/1</td>
<td>5.4</td>
<td>5.0</td>
</tr>
</tbody>
</table>

(1) Number of amino acids in the PDB structure (per monomer); the design algorithm ignores all protein segments that are absent from the structure.
(2) Number of amino acid mutations in the contrast to wild-type; in brackets, as percentage of total number of amino acids.
(3) Number of mutations in protein core. In brackets, as percentage of total number of mutations.
(4) Difference compared to wild-type structure. Positive values indicate higher number in design than wild-type.
(5) Number of mutations to proline.
(6) Mutations to negative (D/E) and positive (K/R) amino acids are indicated to the left and right of the slash, respectively.
(7) Isoelectric point: des refers to designed variant, and WT refers to wild-type.

For DNA and protein sequences, see Data S2.

AUTHOR CONTRIBUTIONS
I.S., and S.J.F. The results were discussed and the manuscript written by A.G., D.S.T., and S.J.F. with input from all authors.

ACKNOWLEDGMENTS

We thank David Schreiber for analyzing structural features of design models; Ravit Netzer, Gideon Lapidoth, and Christoffer Norn for suggestions; and our collaborators for testing the PROSS server. Research in the Fleishman laboratory is supported by the Israel Science Foundation (ISF) through an individual grant; the Center for Research Excellence in Structural Cell Biology, and the joint ISF-UGC program, a Starter’s Grant from the European Research Council, a Career Development Award from the Human Frontier Science Program and a Marie Curie Reintegration Grant, the Minerva Foundation, an Alon Fellowship, and a charitable donation from Sam Switzer and Family. Funding by a DTRA project grant (HDTRA1-11-C-0026) to D.S.T. and NIH (R01EY021205) to R.L.T. are gratefully acknowledged. S.J.F. is a Martha S. Sagon Career Development Chair. D.S.T. is the Nella and Leon Benoziyo Professor of Biochemistry. The collaboration between the Fleishman and Tawfik laboratories is also supported by the Rothschild-Caesaria Foundation.

Received: March 15, 2016
Revised: May 18, 2016
Accepted: June 7, 2016
Published: July 14, 2016

REFERENCES


