HHS Public Access

Author manuscript

Methods Enzymol. Author manuscript; available in PMC 2021 July 17.

Published in final edited form as:

Methods Enzymol. 2020 ; 643: 149–179. doi:10.1016/bs.mie.2020.06.001.

The use of consensus sequence information to engineer stability and activity in proteins

Matt Sternke[1,2], **Katherine W. Tripp**[1], **Doug Barrick**[1,*]

[1]T.C. Jenkins Department of Biophysics, Johns Hopkins University, Baltimore, MD, USA

[2]Program in Molecular Biophysics, Johns Hopkins University, Baltimore, MD, USA

Abstract

The goal of protein design is to create proteins that are stable, soluble, and active. Here we focus on one approach to protein design in which sequence information is used to create a "consensus" sequence. Such consensus sequences comprise the most common residue at each position in a multiple sequence alignment (MSA). After describing some general ideas that relate MSA and consensus sequences and presenting a statistical thermodynamic framework that relates consensus and non-consensus sequences to stability, we detail the process of designing a consensus sequence and survey reports of consensus design and characterization from the literature. Many of these consensus proteins retain native biological activities including ligand binding and enzyme activity. Remarkably, in most cases the consensus sequence shows significantly higher stability, as measured by thermal or chemical denaturation, consistent with the statistical thermodynamic model. To understand this stability increase, we compare various features of consensus sequences with the extant MSA sequences from which they were derived. Consensus sequences show enrichment in charged residues (most notably glutamate and lysine) and depletion of uncharged polar residues (glutamine, serine, and asparagine). Surprisingly, a survey of stability changes resulting from point substitutions show little correlation with residue frequencies at the corresponding positions within the MSA, suggesting that the high stability of consensus proteins may result from interactions among residue pairs or higher-order clusters. Whatever the source, the large number of reported successes demonstrates that consensus design is a viable route to generating active and in many cases highly stabilized proteins.

Keywords

Consensus sequences; protein design; protein stability; protein engineering; multiple sequence alignments; bioinformatics

1. Introduction

A major goal of protein design is to generate proteins that have high activity and are well-behaved. Typically, well-behaved proteins have high thermodynamic stability. High stability not only ensures that a protein adopts its target fold, it promotes long "shelf-life" (that is, it

*To whom correspondence should be addressed: barrick@jhu.edu.

remains active over a long period of time) by keeping the protein folded, because unfolded proteins are often inactivated through aggregation, chemical modification, or surface adsorption.

There are several distinct but complementary approaches to protein design, including *de novo* methods (Huang et al., 2016), directed evolution (Arnold, 2015, 2019), and sequence-based phylogenetic approaches (Magliery, 2015; Poole & Ranganathan, 2006; Porebski & Buckle, 2016). *De novo* design typically seeks to generate novel folds, and often uses physics-based energy terms, although local structural preferences from data bases can be included. One successful approach to *de novo* design uses the Rosetta energy function (Alford et al., 2017; Kuhlman, 2019). Rosetta has been highly successful in generating very stable proteins that adopt all sorts of target structures (e.g., Brunette et al., 2015; Lu et al., 2018; Marcos et al., 2018), although these proteins typically lack biological activity. Designing in biological activity into novel folds can be a major challenge.

In contrast, directed evolution and bioinformatic approaches to protein design are restricted to proteins that are found in nature and adopt stable folds. These two methods are more likely to generate active proteins. This is especially true for directed evolution, where activity is often a selected property. Bioinformatic approaches include consensus design and ancestral reconstruction. Although activity is not a selected property in bioinformatic approaches, activity often leaves a strong imprint on protein sequence. Thus, activity is likely to be imparted on proteins that are designed using sequence information from naturally occurring proteins.

In this chapter, we will discuss the consensus approach to protein design, in which a consensus sequence is created from an alignment of many proteins from the same family. The consensus sequence is simply the most frequently occurring residue at each position within the multiple sequence alignment (MSA, Figure 1). Since residues that are important for activity are highly conserved, active sites and binding interfaces are likely to be retained in consensus sequences.

By definition, consensus sequences bear strong resemblance to naturally occurring sequences they are designed from, which we will refer as “extant MSA sequences” in this chapter. Yet, as will be discussed below, consensus sequences are quite different from naturally occurring sequences in terms of overall residue composition. Moreover, consensus design is agnostic to sequence correlations involving two or more residues, since positional frequencies are evaluated independently of residue identities at other positions.[1] It might be expected that these sequence differences would result in decreased stabilities for consensus proteins, compared to the extant MSA sequences. However, as described below, many studies have found consensus proteins to be *more stable* than their naturally occurring counterparts.

In this chapter, we focus on consensus proteins, their design, and their properties, with emphasis on thermodynamic stabilities. We first consider multiple sequence alignments and

[1]Although strong pairwise correlations are likely to be built into consensus sequences.

the resulting consensus sequences using a statistical thermodynamic framework. Next, we describe some methods we use to generate consensus sequences. We then review findings from a number of laboratories on the results of consensus design, which demonstrate consensus proteins adopt their target folds, often maintain biological activities, and usually possess very high thermodynamic stabilities, and compare these findings to those from ancestral reconstruction. Finally, we explore some of the sequence features of consensus proteins, in part to look for clues for the origins of increased stability, and highlight some open questions and future directions.

2. Protein stability and sequence conservation

2.1. A thermodynamic view of the multiple sequence alignment

In this section, we will use a statistical thermodynamic model to interpret multiple sequence alignments, relating the folding free energy to frequencies in an MSA (e.g., Figure 1B). By assuming that each position in the protein is independent of the others, the model provides relationships between the energy and the frequency of each residue at each position in the native state[2]. In this model, each sequence in the alignment is treated as a replica in a thermodynamic ensemble, in the same way that the ensemble method is applied to equilibrium systems in statistical thermodynamics (Hill, 1987; McQuarrie, 1984). This view of multiple sequence alignments and resulting consensus sequences has previously been described by Steipe (Steipe, 2004). The goal of this model is the derivation of a “residue partition function”, which directly relates to the probabilities of each residue at a given position.

Consider sequence variation at position i of a protein L residues in length. The residue at position i can be of any of twenty residue types along with a gap, denoted as the set $r = \{Ala, Cys, \ldots, Tyr, gap\}$. We will consider each of the 20 residues (and the gap) to have an energy that depends on residue identity and the environment in the solvated native protein (Figure 2). Following the standard approach in statistical thermodynamics (Hill, 1987; McQuarrie, 1984), we will build an ensemble of a large number of replicas, each containing a single solvated protein molecule in its native state. We will allow each replica to equilibrate over 21 energy levels corresponding to different residues at position i. This equilibration is equivalent to an alchemical reaction among the different side chains. To connect with the Gibbs free energy, we will also allow the system to equilibrate with a large bath at constant temperature T and pressure p, via heat-flow and volume change for each replica (Barrick, 2017).

After equilibration, the number of replicas containing each of the 21 residues will match equilibrium populations. We will assume that these populations match the frequencies for each residue at position i in a suitably large multiple sequence alignment. Using Lagrange multipliers, we can relate these populations to the free energies of each side chain in the native state through a reaction partition function for position i:

[2]Sequence dependent energetic effects in the denatured state are not considered here. Although denatured-state interactions have been detected in a few systems (Cho et al., 2014), they are likely to be weaker than native-state interactions.

$$\rho_i = \sum_{xaa \in r} \theta_{i,\ xaa} = \sum_{xaa \in r} e^{-\bar{G}_{i,\ xaa}/RT} \tag{1}$$

The quantity $\theta_{i,\ xaa}$ is a partition function for residue xaa at position i.[3] The overall reaction partition function ρ_i is obtained by summing the set of 21 $\theta_{i,\ xaa}$ terms, since each residue is mutually exclusive—a protein chain cannot have two residues at a single position. Note that the statistical weight for each residue is Boltzmann-like, where the energy is the Gibbs free energy.

With this residue partition function, we can calculate the probability of a given residue at position i from the ratio of the statistical weight for that residue divided by the partition function:

$$p_{i,\ xaa} = \frac{\theta_{i,\ xaa}}{\rho_i} = \frac{e^{-\bar{G}_{i,\ xaa}/RT}}{\rho_i} \tag{2}$$

As long as a multiple sequence alignment is large enough to give good statistical averaging, the frequency of residue xaa at position i ($f_{i,\ xaa}$) should be approximately equal to this probability[4]. In principle, Equation 2 provides a way to connect the population to the free energy. However, Equation 2 depends on all 21 residue free energies at position i, not just that of residue xaa. Fortunately, there are 20 such equations[5], so we can combine equations to determine free energies. One particularly useful way of getting free energies in a way that relates to experimentally measured free stabilities is to compute the ratios of p_i values for different pairs of residues:

$$\frac{p_{i,xaa}}{p_{i,yaa}} = \frac{\theta_{i,xaa}/\rho_i}{\theta_{i,yaa}/\rho_i} = \frac{e^{-\bar{G}_{i,\ xaa}/RT}}{e^{-\bar{G}_{i,\ yaa}/RT}} = e^{-\left(\bar{G}_{i,\ xaa} - \bar{G}_{i,\ yaa}\right)/RT} = e^{-\Delta\bar{G}_{i,\ yaa \rightarrow xaa}/RT} \tag{3}$$

Taking the logarithm of Equation 3 gives

$$\Delta\bar{G}_{i,\ yaa \rightarrow xaa} = -RT\ln\left(\frac{p_{i,xaa}}{p_{i,yaa}}\right) \approx -RT\ln\left(\frac{f_{i,xaa}}{f_{i,yaa}}\right) \tag{4}$$

where the approximation on the right-hand side holds as long as frequencies are taken from a large MSA. Ignoring effects of sequence substitution in the denatured state, the free energy

[3] In turn, each $q_{i,\ xaa}$ is obtained by summing over the energy values for each conformation of residue xaa:

$$\theta_{i,\ xaa} = \sum_j \Omega_j e^{-\left(\bar{\varepsilon}_j + p\bar{V}_j\right)/RT} = \sum_j \Omega_j e^{-\bar{h}_j/RT} = \sum_j e^{-\bar{G}_j/RT}$$

where the sum is taken over energy levels with values $\bar{\varepsilon}_i$, replica volumes $\bar{V}_j$, and all available side-chain conformations. W_j is the number of microstates for each term in the sum, and $\bar{h}_i$ is analogous to an enthalpy. See Barrick (2017) for more details.

[4] Within the limits of the model.

[5] Though there seem like 21 equations, one of the 21 $p_{i,\ xaa}$ values is not free to vary, but is represented by the other 20 under the requirement that the probabilities sum to one.

in Equation 4 is equal to the $\Delta\Delta\bar{G}$ value for a substitution (that is, the difference in the folding free energy between xaa and yaa at position *i*).

Finally, we note that an overall "sequence partition function" can be generated by multiplying the corresponding residue partition functions at each position:

$$\sigma = \prod_{i=1}^{L} \rho_i = \prod_{i=1}^{L} \sum_{xaa \in r} \theta_{i,\ xaa} \tag{5}$$

Such multiplication is appropriate, since all positions are assumed to be independent. The resulting product of L single residue partition functions contains a term for every possible sequence (a total of 20^L), and each of these is a product of the statistical weight of each residue at each position in that sequence. The largest term among these 20^L sequence weights will be the one that combines the largest $\theta_{i,\ xaa}$ term at each position. Each of these maximum terms corresponds to the most probable residue at each position, i.e., the consensus sequence. Thus, in an ensemble in which all L positions equilibrated simultaneously, the consensus sequence would be the most probable sequence. Moreover, since each $\theta_{i,\ xaa}$ term is related to the free energy through Equation 1, the maximum probability consensus sequence is the sequence with the lowest free energy residue at each position, and thus the lowest free energy overall.

2.2. A non-thermodynamic view of the multiple sequence alignment

The statistical thermodynamic treatment above makes a number of unrealistic assumptions about protein sequences. One of the most glaring assumptions is that residues are "equilibrated" across an MSA. This is clearly not the case; rather, these sequences all derive from a common ancestor[6] and sequence differences result from mutation under selective pressure along with genetic drift, and many sequence features may be conserved due to a common origin rather than a favorable effect on stability (or fitness, more generally). Another assumption is that the variations in residue frequencies are determined by energetic terms within the native state. Although there is no doubt a strong evolutionary pressure for folding, there are additional pressures for activity, for solubility, and for turn-over. Some of these conserved features are likely to be destabilizing. For example, active site residues have been shown to decrease stability (Shoichet et al., 1995). Likewise, low stability may be beneficial for proteins that require short half-lives. And while proteins that require long half-lives (such as extracellular enzymes that function in harsh environments) may benefit from high thermodynamic stability, examples have been found where long half-life arises from kinetic (rather than thermodynamic) control (Baker et al., 1992; Jaswal et al., 2005).

3. Obtaining an MSA and determining a consensus sequence

One important but easily overlooked feature of consensus sequence design is that any consensus sequence is dependent on the MSA used to generate it, which is at best a random sample of the full biome[7]. Consensus sequences determined from MSAs that contain a small

[6] Ignoring convergent evolution.
[7] And more likely a biased sample, due to taxonomic biases in selecting samples for sequencing.

number of sequences will be highly dependent on the sequences included in the MSA, and will likely provide a poor representation of the sequence family as a whole. Consensus sequences created from large MSAs should provide better representation of the sequence family.

However, even with unlimited sequence data, several steps must be taken to edit an MSA prior to consensus sequence determination. In this section we outline strategies for collecting a set of sequences for consensus design and for editing those sequences to generate an optimal MSA. An outline of the steps in this consensus design strategy is shown in Figure 3A. Several of these steps require subjective decisions, including cutoff values for length variation, sequence identity, inserts, and gaps. Here we provide recommendations based on our own experiences, but these can be easily modified as desired.

3.1 Collecting a sequence set for consensus design

Collecting a high-quality starting MSA is the most important step in obtaining a consensus sequence. There are two strategies for gathering sequences. The easiest strategy is to obtain a sequence set that already exists as an MSA from a database such as Pfam (El-Gebali et al., 2019). As of March of 2020, Pfam contained MSAs for nearly 18,000 protein families and domains. For each family, Pfam offers a small manually-curated “seed” alignment (typically on the order of 10’s or 100’s of sequences) in addition to larger but less well-curated “full” alignments (typically on the order of 1,000’s [and sometimes 10,000’s] of sequences). Other database options that often offer large sequence sets include Interpro (Mitchell et al., 2019) for protein and domain families and SMART (Letunic & Bork, 2018) for domain families only. However, the large sequence sets in these databases are unaligned; thus, an alignment must be generated by the user.

A second strategy for gathering sequences is to build an MSA from a particular query sequence or a small set of sequences using search and alignment tools such as HMMER (Finn et al., 2011) and PSI-BLAST (Altschul et al., 1997). HMMER offers customizable options to identify sequence homologs through iterative searches of various protein sequence databases, using a profile-HMM built on a user-supplied seed sequence or small sequence set. This strategy provides control over the parameters used for homolog identification that is not available when starting with a pre-aligned MSA.

Although a taxonomically broad MSA may be expected to give the best “equilibration” in sequence space, such breadth may not be ideal for all design targets. For example, if distant lineages have evolved specialized subfunctions, including all lineages may compromise activity. Indeed, Magliery and coworkers have shown that although a consensus SOD1 protein constructed from an MSA restricted to eukaryotic sequences showed near-native enzyme activity, a consensus SOD1 from an MSA with both eukaryotic and bacterial sequences had low catalytic activity (Goyal & Magliery, 2018). However, successful consensus protein designs have been created from MSAs that include all domains of life (Sternke et al., 2019; Sullivan et al., 2011) as well as MSAs restricted to a single phylogenetic clade (Goyal & Magliery, 2018; Tripp et al., 2017). The optimal taxonomic distribution of sequences for consensus design likely depends on the target and the goals of

the design. Pfam, Interpro, and SMART each allow filtering sequences by specific taxa if a limited phylogenetic distribution is desired.

Relatedly, computed consensus sequences are dependent on the number of sequences in the starting alignment. For MSAs composed of a small number of sequences, the calculated residue frequency distributions are likely to be rather poor approximations of the true residue frequencies, especially at weakly conserved positions, and the resulting consensus sequence may be a poor representation of the sequence family. Large MSAs should provide a better approximation of frequencies, allowing weak conservation signals to be accurately captured in the consensus sequence. Consistent with this idea, we found that a consensus homeodomain designed from an MSA of 4,571 sequences (cHD2, see Table S1) shows an increased stability (with a decreased folding free energy of −2.7 kcal/mol) over a consensus homeodomain designed from 182 sequences (Sternke et al., 2019; Tripp et al., 2017).[8]

The number of sequences necessary to achieve a good approximation of true residue frequencies is dependent on the extent of conservation among sequences in the MSA. To illustrate this dependence, we randomly sampled subsets of sequences from 5,000 sequence MSAs that have different levels of conservation, and compared the identities of consensus sequences determined from different random samples. We find that subsets of a few hundred sequences generate consensus sequences with about 90% identity to one another, and as expected, the number of sequences necessary to achieve an average of 90% identity decreases as the average conservation among sequences within the MSA increases (Figure 3B). We see similar results if we compare consensus sequences from “synthetic” alignments where MSA sequences are generated by randomly selecting residues at each position using probabilities determined from the residue frequencies of the 5,000 sequence alignment (not shown).

Although consensus sequences are sensitive to the number of sequences used to generate residue frequencies, there are examples where stable consensus sequences have been generated from a small number of sequences. One notable example is a consensus version of a phytase enzyme, where only 13 sequences were used to generate residue frequencies. Surprisingly, the consensus phytase had an increased T_m of 16 °C (Lehmann et al., 2000). Thus, limited sequence availability for a particular protein family need not be a deterrent to attempt consensus protein design, although more sequences is probably better.

3.2 Curating a sequence set

Once a suitable set of sequences is obtained, additional curation steps are often needed to create an alignment of diverse yet non-redundant sequences. Although MSAs obtained from sequence databases can in principle be used directly for consensus sequence design without manipulation, we and others have found that even the curated alignments contain sequence truncations and highly-similar sequences that may bias the resulting consensus sequence (Sullivan et al., 2011). In our studies of consensus proteins, we apply the following steps to curate sequence sets. First, we remove sequence fragments (which often result from terminal truncations) and anomalously long sequences (often from large internal insertions). In our

[8]Note that the sequence of cHD2 differs by two resides on the N-terminus from the consensus sequence in Figure 1B.

studies, we eliminate sequences that deviate by more than 30% from the median sequence length of the alignment. Second, we remove sequences that share high identity to avoid sequencing biases. We have used a clustering program such as CD-HIT (Li & Godzik, 2006) or UCLUST (Edgar, 2010) to cluster sequences at a 90% sequence identity threshold and choose one representative sequence per cluster for inclusion in the final MSA. An alternative strategy is to include all sequences but decrease the weight of sequences that share high identity to many other sequences in the MSA (Morcos et al., 2011; Socolich et al., 2005).

After modifications for length and identity, we generate a new alignment using a sequence alignment program such as MAFFT (Katoh et al., 2002), ClustalW (Larkin et al., 2007), or MUSCLE (Edgar, 2004). In many cases the default parameters in these alignment programs produce quality MSAs. Nonetheless, results should be visually inspected before determining consensus sequences.

3.3 Generating a consensus sequence

The first step in creating a consensus sequence from an MSA is calculating the residue (and gap) frequencies at each position. These frequencies are then used to edit the MSA one last time to eliminate rare insertions (that is, extra residues in a small number of sequences). Rare insertions in an MSA appear as a region where most sequences have a gap, leading to a gap frequency near one. For large MSAs, this occurs at a majority of positions in the alignment. To avoid including rare insertions in a consensus sequence, we eliminate all positions with gap frequencies greater than one-half, and select the most-frequent non-gap residue[9] at all remaining positions in the MSA for the consensus sequence. Another strategy for removing insertions is to eliminate all positions where the gap frequency exceeds the frequency of the most common residue (Porebski & Buckle, 2016), which can fall below 0.5 at some positions. We find that consensus sequences generated using a gap threshold of 0.5 match more closely to the average lengths of sequences in the alignment. As a third strategy for eliminating insertions, if there is a particular extant sequence of interest in an alignment, all MSA sequences can be trimmed to include only positions occupied by residues in that particular sequence.

An attractive feature of consensus sequence design is its ease of implementation. In contrast, *de novo* design and directed evolution are considerably more labor intensive. For example, designing proteins using Rosetta requires considerable expertise, and can be computationally intensive. Directed evolution requires expertise in recombinant DNA and biochemistry, and depends on a biological activity that can be selected for through multiple rounds of selection and amplification. Determining a consensus sequences only requires a curated MSA and basic computer fluency. To assist interested users with the process of designing consensus sequences, we have made available basic Python scripts for cleaning and curating MSAs, calculating residue frequencies from these MSAs, and determining the consensus sequences. These scripts, including examples for how to use them, can be found on Github at:

[9]With a gap cutoff of 0.5, it is possible for a gap to be more frequent than any of the twenty residues at a given position within an MSA. For example, at a position where all twenty residues are equally probable, a gap frequency need only be above 1/21 = 0.0476 to be the most frequent character state. Selecting a gap at such a position would provide a poor representation of the fact that as many as 95% of sequences in the alignment have a residue at that position.

github.com/msternke/protein-consensus-sequence. As a last step in consensus design, the final MSA should be saved in a permanent record, so that sequence features that generated the consensus sequence can be evaluated at a later date. In addition, we recommend publishing the final MSA and the consensus sequence as supplementary material.

4. Survey of consensus proteins

4.1 Thermodynamic stabilities of consensus proteins

In a survey of consensus proteins in the literature, we have found 20 examples in which unique globular protein families have been targets for successful consensus protein designs using the wholesale consensus approach (Table S1), adding to the examples identified in a previous review (Porebski & Buckle, 2016). This large and growing number suggests a high success rate for wholesale consensus design, although it is hard to assess the actual success rate (success / (success + failure)) because failures are not likely to be published. The earliest example we have found is the design of a consensus zinc finger by Berg and coworkers (Desjarlais & Berg, 1993). Another early success includes a consensus phytase that has a T_m 27 °C higher than the average T_m of the extant MSA sequences (and 20 °C higher than the T_m of the most stable extant phytase measured; Lehmann et al., 2000). One surprising aspect of these studies is that they used fairly small MSAs. The consensus zinc finger sequence was determined from 131 sequences, and the consensus phytase sequence was determined from only 13 sequences. These small sequence sets, which are likely the result of the small sequence databases available at the time, are likely to provide a poor representation of true residue frequencies at positions with weak conservation.

In an attempt to see how widely applicable the consensus approach is, we designed a set of seven consensus proteins using MSAs curated as described in Section 3. In these designs, the number of sequences in each MSA ranged from 3500 to 14000. Five out of the seven resulting consensus proteins are more stable than the extant sequences that had been reported in the literature (Sternke et al., 2019), with decreased folding free energies ranging from −1.4 to −8.3 kcal/mol compared to average stabilities of extant sequences. The exceptions to this stability trend are consensus SH3 and PGK. Consensus SH3 is modestly less stable compared to typical extant SH3 domains (increased folding free energy of 0.3 kcal/mol compared to the extant mean free energy). Consensus PGK has a folding free energy nearly as low as those of the most stable extant PGK sequences, though since unfolding transitions are likely to be multistate, these free energies should be interpreted cautiously. This high level of success shows consensus design to be an effective way to stabilize a folded protein, and is significantly less labor intensive than an iterative approach using point substitutions.

Successful consensus design does not appear to be limited by size. Although some stabilized consensus designs are small proteins, such as NTL9 (L=46), albumin binding domain (L=46) and homeodomain (L=57; Jacobs et al. 2015; Tripp et al. 2017; Sternke, Tripp, and Barrick 2019) some consensus designs are quite large, such as PGK (L=392), EF-Tu (L=394), serine protease (L=396), and fungal phytase (L=476; (Cole & Gaucher, 2011; Lehmann et al., 2000; Porebski et al., 2016; Sternke et al., 2019). Thus, it does not appear

that the large networks of interactions present in larger protein domains limits the consensus design approach.

In addition to globular protein targets, several repeat protein families have been the targets of consensus protein design (Table S2). Repeat proteins are particularly amenable to consensus design, given the abundance of sequence information resulting from repeating sequence elements (Aksel et al., 2011; Binz et al., 2003; Main et al., 2003; Mosavi et al., 2002; Parker et al., 2014; Parmeggiani et al., 2008). A number of consensus ankyrin and tetratricopeptide repeat proteins have been designed using MSAs of different origins and sizes; many of these (especially ankyrin repeat proteins) are considerably more stable than corresponding extant MSA proteins, with thermal unfolding midpoints ranging from 50–85 °C (Aksel et al., 2011; Binz et al., 2003; Main et al., 2003; Mosavi et al., 2002). Four-ankyrin-repeat consensus arrays have folding free energies ranging from ranging from −9.5 to −11.4 kcal/mol (Aksel et al., 2011), whereas four-repeat arrays from the Notch ankyrin domain have free energies of 0.3 and −1.8 kcal/mol (Mello & Barrick, 2004). In contrast, leucine-rich repeat and armadillo repeat arrays have proved to be more difficult targets for consensus design, exhibiting poor expression, limited solubility, molten globule characteristics, and noncooperative folding transitions

4.2 Structural features of consensus proteins

Although structural data is not available for many of the consensus proteins in Table S1, the data that is available shows that consensus proteins adopt the same folds as the extant proteins they were designed from. High-resolution crystal structures have been solved for 4 of the 20 globular consensus protein families in Table S1. For the FN3 family, structures of three different consensus sequences have been solved. These structures are all quite similar to structures of extant proteins in the same family. Interestingly, consensus proteins for which crystal structures are available tend to be on the large side, and tend to have charge densities that are closer to densities seen for extant proteins. It is possible that the high charge density of some consensus sequences (see Section 5) interferes with crystallization. In contrast, there are more crystal structures available for consensus repeat proteins; again, these proteins adopt the same structures as extant counterparts (Madhurantakam et al., 2012; Marold et al., 2015).

Though there are few crystal structures of short globular consensus proteins, these proteins tend to give high-quality NMR spectra (as do some of the larger proteins; (Sternke et al., 2019)). These include consensus HD, SH2, SH3, NTL9 (and DHFR, AK, and PGK). Analysis of backbone chemical shifts from these four proteins show that α-helices and β-strands closely match secondary structures of extant MSA sequences for which structural information is available. For consensus HD, a CS-Rosetta structure restrained with backbone chemical shifts and long-range NOEs matches the homeodomain tertiary structure, with an RMSD of 2.14 Å to the *D. melanogaster* engrailed homeodomain (Tripp et al., 2017).

Given that average sequence identities between consensus sequences and extant sequences in the MSA range from 40–60% (Sternke et al., 2019), the observation that consensus proteins adopt folds that are characteristic of their protein families is not a surprise.

Nonetheless, the structural data confirm that consensus design faithfully generates proteins that have the intended fold.

4.3 Activities of consensus proteins

Consensus proteins have been successfully designed from protein families that display a variety of functions, including binding of proteins, peptides, and DNA, and catalysis. Several groups have used consensus proteins as scaffolds to engineer either DNA binding or protein affinity (Binz et al., 2003; Parmeggiani et al., 2008). An early study used consensus zinc fingers, varying the residues in each finger that are known to contact DNA. The consensus zinc fingers have K_D's ranging from 11μM – 2 nM (Desjarlais & Berg, 1993). Similarly, consensus repeat proteins have been used to generate novel binding partners. These studies randomize weakly conserved surface positions to create libraries that can be selected for binding the specific targets. Plückthun and coworkers have used this approach and generated consensus ankyrin repeats that have high binding affinities, with K_D's as low as 4.4 nM (Binz et al., 2004). This strategy makes use of the high stability of consensus repeats and the ability to change the size of the binding surface by varying the number of repeats.

In addition, several consensus proteins designed from families that bind specific ligands have been characterized, and in many cases, retain binding affinity. A consensus homeodomain was shown to bind its cognate DNA sequence with 100-fold higher affinity than a well-studied extant homeodomain (Tripp et al., 2017). A consensus albumin binding domain has been shown to bind albumin from several species with dissociation constants as high as 75 pM (Jacobs et al., 2015). These examples highlight that consensus design may be an effective route to generate tight binders to both native and novel binding partners.

In addition to binding, consensus proteins derived from enzymes have been shown to retain catalytic activity. A set of consensus chorismate mutases showed activities ranging from 2-fold higher to 30-fold lower than the extant sequence used for comparison (Jäckel et al., 2010). Interestingly, consensus β-lactamase has lower activity with penicillin but is more active with third generation antibiotics, demonstrating that the specificity of consensus proteins may differ from extant counterparts (Risso et al., 2014). Consensus DHFR, AK, and PGK all show substantial catalytic activity with K_m values similar to those of extant enzymes (Sternke et al., 2019). However, the k_{cat} values are lower than their mesophilic counterparts, although they are comparable or slightly lower than the thermophilic homologues. A consensus DnaE intein was shown to undergo faster splicing than the extant proteins, and was active in both high denaturant and high temperatures, unlike an extant intein (Stevens et al., 2016). Together these studies present an intriguing idea that may have industrial implications: that the consensus proteins may be a good route for obtaining proteins that are active under harsh conditions.

5. Sequence and structural features of consensus proteins

The general increased stability of consensus proteins compared to extant MSA sequences may reflect a general mechanism underlying consensus stabilization. Extant protein sequences are under selective pressure to be able to find and maintain their native-state

structures in order to perform their respective functions. Finding and maintaining native-state structures requires proteins to be thermodynamically stable. Based on the statistical thermodynamic interpretation of the MSA in Section 2, such a stability requirement for evolved sequences would be expected to be imprinted in the consensus sequence. Thus, we might expect that by comparing consensus and extant protein sequences, we might be able to identify the sequence features that contribute to increased stability. Here we explore the sequence features of consensus sequences, focusing on differences between extant MSA sequences and stabilized consensus sequences derived from these MSA sequences.

5.1 Comparisons of consensus sequences to extant sequences

By definition, a consensus sequence represents the sequence that shares the highest average identity to all sequences in an MSA. The average sequence identity that a consensus sequence shares with its MSA sequences depends on sequence variation within the MSA, but for large well-sampled MSAs (composed of >1,000 extant sequences), average identities between consensus and MSA sequences typically range from 40–60% (Sternke et al., 2019). This is greater than the ~25–45% average identity that is typical of extant sequence pairs within the MSA. However, there are significant sequence differences between the consensus and the most-similar sequence in the founding MSA, with identities ranging from 65–85% (Sternke et al., 2019).

Given the high average identities of consensus sequences to their MSA sequences, it might be expected that consensus sequences have residue compositions similar to these extant sequences. Although some residues in a set of seven consensus sequences that we previously designed (Sternke et al., 2019) have similar average residue frequencies as their respective MSA sequences (e.g., arginine, tryptophan, and tyrosine), other residues have different frequencies (glutamate, lysine and glycine are enriched in the seven consensus sequences; glutamine, serine, and isoleucine are depleted).

This analysis can be extended to consensus sequences reported from other labs by comparing the residue frequencies of the consensus sequences[10] to average residue frequencies from a nonredundant sampling of sequences from the PDB. To obtain this nonredundant sequence, we selected one sequence from each cluster in the weekly *BLASTclust* clustering of all protein chains in the PDB (accessed on 2/4/2020) at a 30% sequence identity threshold (Altschul et al., 1990). Residue frequencies in consensus sequences compared to this PDB reference set shows similar residue biases as described above (Sternke et al., 2019): consensus sequences are enriched in glutamate residues (the larger set also shows some enrichment in lysine and glycine, although these biases are quite variable), and depleted in isoleucine, methionine, asparagine, glutamine, cysteine, and serine residues (Figure 4A; again, the biases can be quite variable).

The individual residue enrichments and depletions are also apparent when residues with similar chemical properties are grouped. The seven consensus proteins we designed using the approach described in Section 3 (Sternke et al., 2019) are an average of two standard

[10]We limited this analysis to consensus sequences designed from MSAs composed of >100 sequence to provide adequate sampling statistics. See Table S1.

deviations above the mean of their MSA sequences in the total proportion of charged residues, and lie two or more standard deviations below the mean of their MSA sequences in the total proportion of polar uncharged residues. The substitutions that result in this enrichment in charged residues and depletion of polar uncharged residues occur most commonly at nonconserved positions on the protein surface (Sternke et al., 2019). We note that charge enrichment is less pronounced for consensus proteins designed by other groups. It should also be noted that some of the largest sequence biases occur in the shortest consensus proteins (e.g. HD and SH3), which is not unexpected since a single substitution generates a larger compositional change in a small protein. It remains unclear which (if any) of these composition differences contribute to the enhanced stability of consensus proteins.

As with sequence features, consensus proteins may show shared structural features that contribute to enhanced thermodynamic stability. Lehmann and coworkers report that a crystal structure of a stabilized consensus phytase shows improved hydrophobic packing in the protein core as well as stabilization of a surface loop relative to structures of less-stable natural phytases (Lehmann et al., 2000). Using ^{15}N spin relaxation measurements by NMR spectroscopy, we found that a stabilized consensus homeodomain shows decreased dynamic motions in loop regions relative to the naturally-occurring *D. melanogaster* Engrailed homeodomain (Tripp et al., 2017), suggestive of a similar loop stabilization. Buckle and coworkers report that a crystal structure of a highly-stabilized consensus FN3 domain shows more hydrogen bonds and salt bridges, a smaller solvent accessible surface area, and a smaller total solvent inaccessible cavity volume than is seen in a set of less-stable natural and designed FN3 domain structures (Porebski et al., 2015). However, the crystal structure of a similarly highly-stabilized consensus serpin determined by Buckle and coworkers shows the consensus serpin shows fewer hydrogen bonds and salt bridges and a larger solvent accessible surface area than a set of less-stable natural serpins (Porebski et al., 2016). A complete understanding of how these structural features may contribute to the enhanced thermodynamic stabilities of consensus proteins will require further structural and thermodynamic studies of a broader set of consensus proteins.

5.2 Comparisons to thermophilic proteins

Like consensus proteins, proteins from thermophilic organisms have high thermodynamic stability (Hollien & Marqusee, 1999; Razvi & Scholtz, 2006). It is possible that both groups of proteins owe their high stabilities to the same compositional or structural features. Thermophilic protein sequences share some of the same residue composition biases (albeit weak ones) as consensus sequences: glutamate and lysine residues are enriched in both sets, whereas cysteine, histidine, glutamine, and serine residues are depleted (Kumar et al., 2000; Y. Li et al., 2010; Sternke et al., 2019). However, thermophilic sequences are enriched in isoleucine, arginine, and tyrosine residues, whereas consensus sequences are either depleted or show no bias for these residues.

It is possible that the partial similarities between consensus and thermophilic sequences may just result from a large number of thermophilic sequences in the starting MSAs. It is often difficult to determine whether sequences in MSAs have thermophilic versus mesophilic origins, since sequence records do not always specify the organism from which the sequence

derives, and even when they do, the environmental preferences of the organism often cannot be determined. However, several highly stable consensus proteins have been derived exclusively from eukaryotic sequences (see Section 3). With just a few exceptions, eukaryotes are mesophiles (and sometimes psychrophiles). For designs from MSAs that contain sequences from thermophilic bacteria and archaea (NTL9, DHFR, AK, and PGK; Table S1), nearly identical consensus sequences (>98%) are obtained when thermophilic sequences are removed from the MSAs (Sternke et al., 2019). Thus, if similar sequence features contribute to thermodynamic stability in consensus and thermophilic proteins, these same biases must also be present (albeit to a lesser extent) in the sequence biases of mesophilic proteins.

5.3 Comparison to ancestral proteins.

Another bioinformatic approach that is similar in some ways to consensus sequence design is ancestral reconstruction, where a set of aligned extant protein sequences is analyzed to deduce a likely phylogenetic tree and the sequences of extinct ancestral proteins within the tree (Hochberg & Thornton, 2017; Merkl & Sterner, 2016; Risso et al., 2018; Wheeler et al., 2016a). It has often (but not always) been observed that ancestral proteins have higher stabilities than extant proteins. For example, a comparison of ancestral β-lactamases with consensus proteins from the same sequence alignments showed higher stabilities for the ancestors than for the consensus proteins, though all were more stable than the extant TEM-1 β-lactamase (Risso et al., 2014). Although all enzymes showed reduced activities towards benzylpenicillin, the ancestors showed increased activities for third-generation antibiotics; one of the consensus proteins showed modest activity increases towards these antibiotics, whereas another did not. Likewise, ancestral DNA gyrases were shown to be more stable than a consensus gyrase and to have higher activity (Satoshi Akanuma et al., 2011). In contrast, a consensus L-threonine 3-dehydrogenase was found to be more stable than the corresponding ancestral version, with both enzymes showing high levels of activity (Nakano et al., 2018). An ancestral nucleotide diphosphate kinase was shown to be more stable than a consensus enzyme from the same set of sequences (S. Akanuma et al., 2013), and two out of three ancestral Ef-TU proteins were shown to be more stable than a consensus Ef-TU (Okafor et al., 2018). A study of ancestral RNaseH proteins showed a deep ancestor that is more stable than mesophilic sequences in the starting alignment, but less stable than thermophilic sequences (Wheeler et al., 2016b).

Though these studies show that ancestral proteins are more often more stable than consensus versions, it is clear that both methods usually lead to increased stability. One simple explanation for the success of both methods is that ancestral and consensus sequences look alike. This similarity is not surprising since both methods extract sequences from a multiple sequence alignment. Tawfik and coworkers have argued that consensus information may be overrepresented in ancestral reconstruction, although this only partly explains the stability enhancement of ancestral sequences (Trudeau et al., 2016). One possible explanation for the increased stability of a majority of ancestral proteins compared to consensus proteins is that ancestral reconstruction is more likely to capture stabilizing coupling between nearby residues (S. Akanuma et al., 2013).

5.4 Correlations of residue frequencies with effects on stability

In the statistical thermodynamic framework presented in Section 2, residue frequencies at each position of an MSA represent equilibrium populations within an ensemble and thus are related to the energies of the residues in the protein native state. As such, substitutions to residues of higher frequency should be stabilizing, whereas substitutions to residues of lower frequency should be destabilizing (Equation 4). Indeed, some studies have found that stability changes resulting from point-substitution are partially correlated with residue conservation (Amin et al., 2004; Di Nardo et al., 2003; Loening et al., 2006; Nikolova et al., 1998; Polizzi et al., 2006; Steipe et al., 1994; Wang et al., 1999), although some of these correlations involve a rather small data set.

To examine the extent of this correlation on a broader scale, we surveyed the ProTherm database (Gromiha et al., 1999) to find proteins where a large number of stability changes resulting from point substitutions have been reported, and where large MSAs could be obtained. The two proteins that best satisfy these criteria are Staphylococcal nuclease (SNase; 514 point substitutions) and Barnase (200 point substitutions). We compared the effects of point substitutions on folding free energies ($\Delta\Delta G°_{H2O}$) to the log ratio of the residue frequencies that define the point substitution, $\ln(f_{i,xaa}/f_{i,yaa})$. Within the framework of the statistical thermodynamic model, this log ratio should be proportional to the free energy difference between the two residues xaa and yaa at position i (see Section 2).

Contrary to expectation, neither protein shows a strong correlation between the stability changes resulting from point substitution and residue frequencies (Figure 4B). Although for SNase substitutions to residues of higher frequencies are stabilizing more often than substitutions to residues of lower frequencies (18% of substitutions compared to 8%), for Barnase substitutions to residues of higher frequencies and substitutions to residues of lower frequencies show similar effects on stability (12% of substitutions compared to 11%). In short, the majority of substitutions towards residues of higher frequencies are destabilizing for both proteins. Furthermore, restricting this analysis to substitutions to consensus residues (substitutions to the residue of greatest frequency at a given position) gives an even weaker correlations (Figure 4C); as with the larger data set, only a minority of these substitutions are stabilizing (21% and 11% for SNase and Barnase). This analysis indicates that at a given position, substitution toward consensus should be destabilizing more frequently than it is stabilizing.

The lack of correlation between residue conservation and effects of point-substitution on stability for SNase and Barnase is inconsistent with the high thermodynamic stability of consensus proteins (Table S1). Although the publication process may impart some bias toward consensus sequence stabilization in the literature (unstable consensus proteins will be likely to be unfolded and unpublishable), an unbiased test of consensus design on seven targets resulted in stability increases in five cases (Sternke et al., 2019), and all seven proteins well-folded. This observation suggests that the success of consensus design in generating proteins with high stability is not primarily a result of publication filter.

The discrepancy between the stability changes associated with point substitutions and wholesale consensus substitution is striking and is worthy of further study. One possible

(though speculative) explanation is that the consensus stability increment is a collective feature of all (or at least a subset of) consensus substitutions, and cannot be attributed to individual substitutions in an extant sequence background. Consistent with this explanation, Magliery and coworkers showed that stability changes resulting from single substitutions toward consensus differ qualitatively from stability changes when those substitutions are combined (Sullivan et al., 2012b). Specifically, TIM constructs that combine 6 and 13 substitutions toward consensus (plus an additional compensatory substitution) were both destabilized compared to the expected summed effect from the individual substitutions. Although this difference goes the opposite direction to the proposed explanation for the ΔΔG values in Figure 4 (where DDG values are mostly destabilizing, and the consensus is assumed to be stabilized[11]), it does demonstrate non-additivity. In some of these cases, evaluation of pairwise sequence correlations showed that destabilizing substitutions toward consensus often show strong coupling with neighboring residues (Sullivan et al., 2012a). An additional testing for this type of synergistic effect could involve large-scale studies of substitutions that traverse sequence space between extant MSA sequences and stabilized consensus sequences via different paths. In addition to resolving the above discrepancy, it is expected that such studies will provide insight into which types of substitutions result in consensus stabilization. These insights might be used to increase protein stability using a relatively small number of substitutions.

6. Summary and open questions

As numerous studies now show, consensus design provides a simple strategy to make proteins that are folded and active, and in many cases show significantly increased stability compared to the extant protein sequences from which they are derived. The increase in stability is consistent with a statistical thermodynamic framework for protein sequence families, although that framework seems at odds with data from point substitutions which show little correlation between residue conservation and stability.

The origins of increased stability in consensus proteins remains an open question; answering this question may provide a route to selectively stabilizing proteins with fewer sequence substitutions than are typically made in a wholesale consensus substitution. The possibility that the consensus stability increment results from nonadditive interactions involving groups of residues is consistent with the discrepancy with the point substitution results, and is worth testing by comparing different paths through sequence space from extant to consensus proteins. If synergy among groups of residues is in fact responsible for consensus stabilization, patterns deduced from covariance analysis within the MSA may provide a route to further stabilization.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

[11]It should be noted that consensus versions of Barnase and SNase have not, to our knowledge, been constructed.

Acknowledgements

We thank current and past members of the Barrick lab for discussions and insights relating to using consensus information to explore protein stability, structure, and function, along with numerous colleagues and students at Johns Hopkins University. Our research in this area as well as support for preparation of this chapter has been provided by the research grant GM60042 to DB from NIGMS/NIH, by the training grant T32 GM008403 from NIGMS/NIH, and by the fellowship F31 GM128295 to MS from NIGMS/NIH.

References

Akanuma S, Nakajima Y, Yokobori S. -i., Kimura M, Nemoto N, Mase T, Miyazono K. -i., Tanokura M, & Yamagishi A (2013). Experimental evidence for the thermophilicity of ancestral life. Proceedings of the National Academy of Sciences, 110(27), 11067–11072. 10.1073/pnas.1308215110

Akanuma Satoshi, Iwami S, Yokoi T, Nakamura N, Watanabe H, Yokobori S, & Yamagishi A (2011). Phylogeny-Based Design of a B-Subunit of DNA Gyrase and Its ATPase Domain Using a Small Set of Homologous Amino Acid Sequences. Journal of Molecular Biology, 412(2), 212–225. 10.1016/j.jmb.2011.07.042 [PubMed: 21819994]

Aksel T, Majumdar A, & Barrick D (2011). The Contribution of Entropy, Enthalpy, and Hydrophobic Desolvation to Cooperativity in Repeat-Protein Folding. Structure, 19(3), 349–360. 10.1016/j.str.2010.12.018 [PubMed: 21397186]

Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, Shapovalov MV, Renfrew PD, Mulligan VK, Kappel K, Labonte JW, Pacella MS, Bonneau R, Bradley P, Dunbrack RL, Das R, Baker D, Kuhlman B, Kortemme T, & Gray JJ (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. Journal of Chemical Theory and Computation, 13(6), 3031–3048. 10.1021/acs.jctc.7b00125 [PubMed: 28430426]

Altschul SF, Gish W, Miller W, Myers EW, & Lipman DJ (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403–410. 10.1016/S0022-2836(05)80360-2 [PubMed: 2231712]

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, & Lipman DJ (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research, 25(17), 3389–3402. 10.1093/nar/25.17.3389 [PubMed: 9254694]

Amin N, Liu AD, Ramer S, Aehle W, Meijer D, Metin M, Wong S, Gualfetti P, & Schellenberger V (2004). Construction of stabilized proteins by combinatorial consensus mutagenesis. Protein Engineering, Design and Selection, 17(11), 787–793. 10.1093/protein/gzh091

Arnold FH (2015). The nature of chemical innovation: New enzymes by evolution*. Quarterly Reviews of Biophysics, 48(4), 404–410. 10.1017/S003358351500013X [PubMed: 26537398]

Arnold FH (2019). Innovation by Evolution: Bringing New Chemistry to Life (Nobel Lecture). Angewandte Chemie International Edition, 58(41), 14420–14426. 10.1002/anie.201907729 [PubMed: 31433107]

Baker D, Sohl JL, & Agard DA (1992). A protein-folding reaction under kinetic control. Nature, 356(6366), 263–265. 10.1038/356263a0 [PubMed: 1552947]

Barrick D (2017). Biomolecular Thermodynamics: From Theory to Application (1st ed.). CRC Press.

Binz HK, Amstutz P, Kohl A, Stumpp MT, Briand C, Forrer P, Grütter MG, & Plückthun A (2004). High-affinity binders selected from designed ankyrin repeat protein libraries. Nature Biotechnology, 22(5), 575–582. 10.1038/nbt962

Binz HK, Stumpp MT, Forrer P, Amstutz P, & Plückthun A (2003). Designing Repeat Proteins: Well-expressed, Soluble and Stable Proteins from Combinatorial Libraries of Consensus Ankyrin Repeat Proteins. Journal of Molecular Biology, 332(2), 489–503. 10.1016/S0022-2836(03)00896-9 [PubMed: 12948497]

Brunette TJ, Parmeggiani F, Huang P-S, Bhabha G, Ekiert DC, Tsutakawa SE, Hura GL, Tainer JA, & Baker D (2015). Exploring the repeat protein universe through computational protein design. Nature, 528(7583), 580–584. 10.1038/nature16162 [PubMed: 26675729]

Cho J-H, Meng W, Sato S, Kim EY, Schindelin H, & Raleigh DP (2014). Energetically significant networks of coupled interactions within an unfolded protein. Proceedings of the National Academy of Sciences, 111(33), 12079–12084. 10.1073/pnas.1402054111

Cole MF, & Gaucher EA (2011). Utilizing natural diversity to evolve protein function: Applications towards thermostability. Current Opinion in Chemical Biology, 15(3), 399–406. 10.1016/j.cbpa.2011.03.005 [PubMed: 21470898]

Desjarlais JR, & Berg JM (1993). Use of a zinc-finger consensus sequence framework and specificity rules to design specific DNA binding proteins. Proceedings of the National Academy of Sciences, 90(6), 2256–2260. 10.1073/pnas.90.6.2256

Di Nardo AA, Larson SM, & Davidson AR (2003). The Relationship Between Conservation, Thermodynamic Stability, and Function in the SH3 Domain Hydrophobic Core. Journal of Molecular Biology, 333(3), 641–655. 10.1016/j.jmb.2003.08.035 [PubMed: 14556750]

Edgar RC (2004). MUSCLE: A multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics, 5(1), 113. 10.1186/1471-2105-5-113 [PubMed: 15318951]

Edgar RC (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 26(19), 2460–2461. 10.1093/bioinformatics/btq461 [PubMed: 20709691]

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, & Finn RD (2019). The Pfam protein families database in 2019. Nucleic Acids Research, 47(D1), D427–D432. 10.1093/nar/gky995 [PubMed: 30357350]

Finn RD, Clements J, & Eddy SR (2011). HMMER web server: Interactive sequence similarity searching. Nucleic Acids Research, 39(suppl_2), W29–W37. 10.1093/nar/gkr367 [PubMed: 21593126]

Goyal VD, & Magliery TJ (2018). Phylogenetic Spread of Sequence Data Affects Fitness of SOD1 Consensus Enzymes: Insights from Sequence Statistics and Structural Analyses. Proteins. 10.1002/prot.25486

Gromiha MM, An J, Kono H, Oobatake M, Uedaira H, & Sarai A (1999). ProTherm: Thermodynamic Database for Proteins and Mutants. Nucleic Acids Research, 27(1), 286–288. 10.1093/nar/27.1.286 [PubMed: 9847203]

Hill TL (1987). An Introduction to Statistical Thermodynamics (unknown edition). Dover Publications.

Hochberg GKA, & Thornton JW (2017). Reconstructing Ancient Proteins to Understand the Causes of Structure and Function. Annual Review of Biophysics, 46(1), 247–269. 10.1146/annurev-biophys-070816-033631

Hollien J, & Marqusee S (1999). A Thermodynamic Comparison of Mesophilic and Thermophilic Ribonucleases H. Biochemistry, 38(12), 3831–3836. 10.1021/bi982684h [PubMed: 10090773]

Huang P-S, Boyken SE, & Baker D (2016). The coming of age of de novo protein design. Nature, 537(7620), 320–327. 10.1038/nature19946 [PubMed: 27629638]

Jäckel C, Bloom JD, Kast P, Arnold FH, & Hilvert D (2010). Consensus Protein Design without Phylogenetic Bias. Journal of Molecular Biology, 399(4), 541–546. 10.1016/j.jmb.2010.04.039 [PubMed: 20433850]

Jacobs SA, Gibbs AC, Conk M, Yi F, Maguire D, Kane C, & O'Neil KT (2015). Fusion to a highly stable consensus albumin binding domain allows for tunable pharmacokinetics. Protein Engineering Design and Selection, 28(10), 385–393. 10.1093/protein/gzv040

Jaswal SS, Truhlar SME, Dill KA, & Agard DA (2005). Comprehensive analysis of protein folding activation thermodynamics reveals a universal behavior violated by kinetically stable proteases. Journal of Molecular Biology, 347(2), 355–366. 10.1016/j.jmb.2005.01.032 [PubMed: 15740746]

Katoh K, Misawa K, Kuma K, & Miyata T (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research, 30(14), 3059–3066. 10.1093/nar/gkf436 [PubMed: 12136088]

Kuhlman B (2019). Designing protein structures and complexes with the molecular modeling program Rosetta. Journal of Biological Chemistry, 294(50), 19436–19443. 10.1074/jbc.AW119.008144

Kumar S, Tsai C-J, & Nussinov R (2000). Factors enhancing protein thermostability. Protein Engineering, Design and Selection, 13(3), 179–191. 10.1093/protein/13.3.179

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, & Higgins DG (2007). Clustal W and Clustal X version 2.0. Bioinformatics, 23(21), 2947–2948. 10.1093/bioinformatics/btm404 [PubMed: 17846036]

Lehmann M, Kostrewa D, Wyss M, Brugger R, D'Arcy A, Pasamontes L, & van Loon APGM (2000). From DNA sequence to improved functionality: Using protein sequence comparisons to rapidly design a thermostable consensus phytase. Protein Engineering, Design and Selection, 13(1), 49–57. 10.1093/protein/13.1.49

Letunic I, & Bork P (2018). 20 years of the SMART protein domain annotation resource. Nucleic Acids Research, 46(D1), D493–D496. 10.1093/nar/gkx922 [PubMed: 29040681]

Li W, & Godzik A (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22(13), 1658–1659. 10.1093/bioinformatics/btl158 [PubMed: 16731699]

Li Y, Middaugh CR, & Fang J (2010). A novel scoring function for discriminating hyperthermophilic and mesophilic proteins with application to predicting relative thermostability of protein mutants. BMC Bioinformatics, 11(1), 62. 10.1186/1471-2105-11-62 [PubMed: 20109199]

Loening AM, Fenn TD, Wu AM, & Gambhir SS (2006). Consensus guided mutagenesis of Renilla luciferase yields enhanced stability and light output. Protein Engineering, Design and Selection, 19(9), 391–400. 10.1093/protein/gzl023

Lu P, Min D, DiMaio F, Wei KY, Vahey MD, Boyken SE, Chen Z, Fallas JA, Ueda G, Sheffler W, Mulligan VK, Xu W, Bowie JU, & Baker D (2018). Accurate computational design of multipass transmembrane proteins. Science (New York, N.Y.), 359(6379), 1042–1046. 10.1126/science.aaq1739

Madhurantakam C, Varadamsetty G, Grütter MG, Plückthun A, & Mittl PRE (2012). Structure-based optimization of designed Armadillo-repeat proteins. Protein Science, 21(7), 1015–1028. 10.1002/pro.2085 [PubMed: 22544642]

Magliery TJ (2015). Protein stability: Computation, sequence statistics, and new experimental methods. Current Opinion in Structural Biology, 33, 161–168. 10.1016/j.sbi.2015.09.002 [PubMed: 26497286]

Main ERG, Xiong Y, Cocco MJ, D'Andrea L, & Regan L (2003). Design of Stable α-Helical Arrays from an Idealized TPR Motif. Structure, 11(5), 497–508. 10.1016/S0969-2126(03)00076-5 [PubMed: 12737816]

Marcos E, Chidyausiku TM, McShan AC, Evangelidis T, Nerli S, Carter L, Nivón LG, Davis A, Oberdorfer G, Tripsianes K, Sgourakis NG, & Baker D (2018). De novo design of a non-local β-sheet protein with high stability and accuracy. Nature Structural & Molecular Biology, 25(11), 1028–1034. 10.1038/s41594-018-0141-6

Marold JD, Kavran JM, Bowman GD, & Barrick D (2015). A Naturally Occurring Repeat Protein with High Internal Sequence Identity Defines a New Class of TPR-like Proteins. Structure, 23(11), 2055–2065. 10.1016/j.str.2015.07.022 [PubMed: 26439765]

McQuarrie DA (1984). Statistical Thermodynamics. Univ Science Books.

Mello CC, & Barrick D (2004). An experimentally determined protein folding energy landscape. Proceedings of the National Academy of Sciences, 101(39), 14102–14107. 10.1073/pnas.0403386101

Merkl R, & Sterner R (2016). Ancestral protein reconstruction: Techniques and applications. Biological Chemistry, 397(1), 1–21. 10.1515/hsz-2015-0158 [PubMed: 26351909]

Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Brown SD, Chang H-Y, El-Gebali S, Fraser MI, Gough J, Haft DR, Huang H, Letunic I, Lopez R, Luciani A, Madeira F, Marchler-Bauer A, Mi H, … Finn RD (2019). InterPro in 2019: Improving coverage, classification and access to protein sequence annotations. Nucleic Acids Research, 47(D1), D351–D360. 10.1093/nar/gky1100 [PubMed: 30398656]

Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, & Weigt M (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences, 108(49), E1293–E1301. 10.1073/pnas.1111471108

Mosavi LK, Minor DL, & Peng Z (2002). Consensus-derived structural determinants of the ankyrin repeat motif. Proceedings of the National Academy of Sciences, 99(25), 16029–16034. 10.1073/pnas.252537899

Nakano S, Motoyama T, Miyashita Y, Ishizuka Y, Matsuo N, Tokiwa H, Shinoda S, Asano Y, & Ito S (2018). Benchmark Analysis of Native and Artificial NAD+-Dependent Enzymes Generated by a Sequence-Based Design Method with or without Phylogenetic Data. Biochemistry, 57(26), 3722–3732. 10.1021/acs.biochem.8b00339 [PubMed: 29787243]

Nikolova PV, Henckel J, Lane DP, & Fersht AR (1998). Semirational design of active tumor suppressor p53 DNA binding domain with enhanced stability. Proceedings of the National Academy of Sciences, 95(25), 14675.

Okafor CD, Pathak MC, Fagan CE, Bauer NC, Cole MF, Gaucher EA, & Ortlund EA (2018). Structural and Dynamics Comparison of Thermostability in Ancient, Modern, and Consensus Elongation Factor Tus. Structure, 26(1), 118–129.e3. 10.1016/j.str.2017.11.018 [PubMed: 29276038]

Parker R, Mercedes-Camacho A, & Grove TZ (2014). Consensus design of a NOD receptor leucine rich repeat domain with binding affinity for a muramyl dipeptide, a bacterial cell wall fragment: Consensus LRR binder of MDP. Protein Science, 23(6), 790–800. 10.1002/pro.2461 [PubMed: 24659515]

Parmeggiani F, Pellarin R, Larsen AP, Varadamsetty G, Stumpp MT, Zerbe O, Caflisch A, & Plückthun A (2008). Designed Armadillo Repeat Proteins as General Peptide-Binding Scaffolds: Consensus Design and Computational Optimization of the Hydrophobic Core. Journal of Molecular Biology, 376(5), 1282–1304. 10.1016/j.jmb.2007.12.014 [PubMed: 18222472]

Polizzi KM, Chaparro-Riggers JF, Vazquez-Figueroa E, & Bommarius AS (2006). Structure-guided consensus approach to create a more thermostable penicillin G acylase. Biotechnology Journal, 1(5), 531–536. 10.1002/biot.200600029 [PubMed: 16892288]

Poole AM, & Ranganathan R (2006). Knowledge-based potentials in protein design. Current Opinion in Structural Biology, 16(4), 508–513. 10.1016/j.sbi.2006.06.013 [PubMed: 16843652]

Porebski BT, & Buckle AM (2016). Consensus protein design. Protein Engineering, Design and Selection, 29(7), 245–251. 10.1093/protein/gzw015

Porebski BT, Keleher S, Hollins JJ, Nickson AA, Marijanovic EM, Borg NA, Costa MGS, Pearce MA, Dai W, Zhu L, Irving JA, Hoke DE, Kass I, Whisstock JC, Bottomley SP, Webb GI, McGowan S, & Buckle AM (2016). Smoothing a rugged protein folding landscape by sequence-based redesign. Scientific Reports, 6(1), 1–14. 10.1038/srep33958 [PubMed: 28442746]

Porebski BT, Nickson AA, Hoke DE, Hunter MR, Zhu L, McGowan S, Webb GI, & Buckle AM (2015). Structural and dynamic properties that govern the stability of an engineered fibronectin type III domain. Protein Engineering, Design and Selection, 28(3), 67–78. 10.1093/protein/gzv002

Razvi A, & Scholtz JM (2006). Lessons in stability from thermophilic proteins. Protein Science, 15(7), 1569–1578. 10.1110/ps.062130306 [PubMed: 16815912]

Risso VA, Gavira JA, Gaucher EA, & Sanchez-Ruiz JM (2014). Phenotypic comparisons of consensus variants versus laboratory resurrections of Precambrian proteins: Consensus vs. Ancestral Proteins. Proteins: Structure, Function, and Bioinformatics, 82(6), 887–896. 10.1002/prot.24575

Risso VA, Sanchez-Ruiz JM, & Ozkan SB (2018). Biotechnological and protein-engineering implications of ancestral protein resurrection. Current Opinion in Structural Biology, 51, 106–115. 10.1016/j.sbi.2018.02.007 [PubMed: 29660672]

Shoichet BK, Baase WA, Kuroki R, & Matthews BW (1995). A relationship between protein stability and protein function. Proceedings of the National Academy of Sciences, 92(2), 452–456. 10.1073/pnas.92.2.452

Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, & Ranganathan R (2005). Evolutionary information for specifying a protein fold. Nature, 437(7058), 512–518. 10.1038/nature03991 [PubMed: 16177782]

Steipe B (2004). Consensus-Based Engineering of Protein Stability: From Intrabodies to Thermostable Enzymes. In Methods in Enzymology (Vol. 388, pp. 176–186). Academic Press. 10.1016/S0076-6879(04)88016-9 [PubMed: 15289071]

Steipe B, Schiller B, Plückthun A, & Steinbacher S (1994). Sequence Statistics Reliably Predict Stabilizing Mutations in a Protein Domain. Journal of Molecular Biology, 240(3), 188–192. 10.1006/jmbi.1994.1434 [PubMed: 8028003]

Sternke M, Tripp KW, & Barrick D (2019). Consensus sequence design as a general strategy to create hyperstable, biologically active proteins. Proceedings of the National Academy of Sciences, 116(23), 11275–11284. 10.1073/pnas.1816707116

Stevens AJ, Brown ZZ, Shah NH, Sekar G, Cowburn D, & Muir TW (2016). Design of a Split Intein with Exceptional Protein Splicing Activity. Journal of the American Chemical Society, 138(7), 2162–2165. 10.1021/jacs.5b13528 [PubMed: 26854538]

Sullivan BJ, Durani V, & Magliery TJ (2011). Triosephosphate isomerase by consensus design: Dramatic differences in physical properties and activity of related variants. Journal of Molecular Biology, 413(1), 195–208. 10.1016/j.jmb.2011.08.001 [PubMed: 21839742]

Sullivan BJ, Nguyen T, Durani V, Mathur D, Rojas S, Thomas M, Syu T, & Magliery TJ (2012a). Stabilizing Proteins from Sequence Statistics: The Interplay of Conservation and Correlation in Triosephosphate Isomerase Stability. Journal of Molecular Biology, 420(4–5), 384–399. 10.1016/j.jmb.2012.04.025 [PubMed: 22555051]

Sullivan BJ, Nguyen T, Durani V, Mathur D, Rojas S, Thomas M, Syu T, & Magliery TJ (2012b). Stabilizing proteins from sequence statistics: The interplay of conservation and correlation in triosephosphate isomerase stability. Journal of Molecular Biology, 420(4–5), 384–399. 10.1016/j.jmb.2012.04.025 [PubMed: 22555051]

Tripp KW, Sternke M, Majumdar A, & Barrick D (2017). Creating a Homeodomain with High Stability and DNA Binding Affinity by Sequence Averaging. Journal of the American Chemical Society, 139(14), 5051–5060. 10.1021/jacs.6b11323 [PubMed: 28326770]

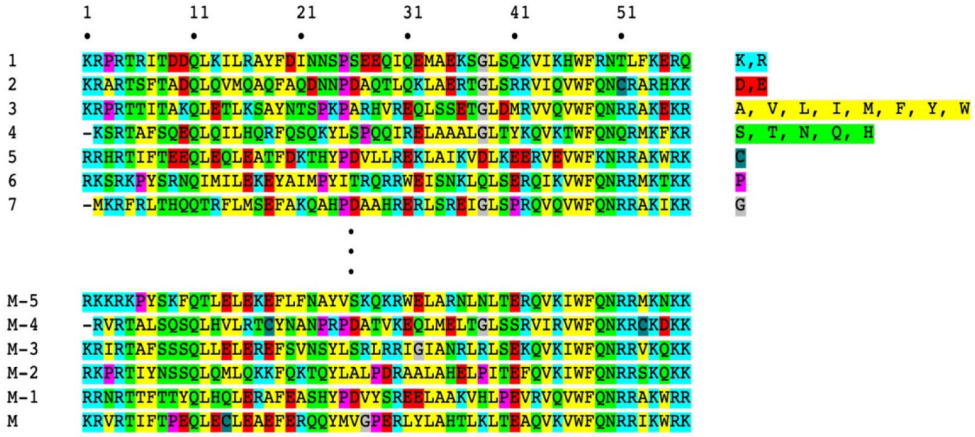
Trudeau DL, Kaltenbach M, & Tawfik DS (2016). On the Potential Origins of the High Stability of Reconstructed Ancestral Proteins. Molecular Biology and Evolution, 33(10), 2633–2641. 10.1093/molbev/msw138 [PubMed: 27413048]

Wang Q, Buckle AM, Foster NW, Johnson CM, & Fersht AR (1999). Design of highly stable functional GroEL minichaperones. Protein Science, 8(10), 2186–2193. 10.1110/ps.8.10.2186 [PubMed: 10548065]

Wheeler LC, Lim SA, Marqusee S, & Harms MJ (2016a). The thermostability and specificity of ancient proteins. Current Opinion in Structural Biology, 38, 37–43. 10.1016/j.sbi.2016.05.015 [PubMed: 27288744]

Wheeler LC, Lim SA, Marqusee S, & Harms MJ (2016b). The thermostability and specificity of ancient proteins. Current Opinion in Structural Biology, 38, 37–43. 10.1016/j.sbi.2016.05.015 [PubMed: 27288744]

A



B

Residue	Position 1	2	3	4	5	6	7	...	57
A	0.029	0.008	0.039	0.001	0.017	0.097	0.018		0.000
C	0.007	0.003	0.005	0.001	0.005	0.016	0.002		0.000
D	0.015	0.000	0.005	0.000	0.002	0.001	0.000		0.000
E	0.025	0.001	0.024	0.002	0.003	0.005	0.000		0.000
F	0.011	0.000	0.007	0.000	0.014	0.010	**0.421**		0.000
G	0.029	0.005	0.018	0.000	0.002	0.002	0.000		0.000
H	0.017	0.005	0.033	0.002	0.031	0.014	0.034		0.007
I	0.019	0.000	0.025	0.000	0.026	0.041	0.066		0.000
K	**0.067**	0.192	**0.178**	0.065	0.049	0.049	0.030		**0.257**
L	0.025	0.000	0.011	0.002	0.034	0.040	0.117	...	0.000
M	0.008	0.002	0.010	0.000	0.006	0.010	0.009		0.000
N	0.032	0.008	0.039	0.002	0.008	0.043	0.000		0.015
P	0.049	0.010	0.091	0.002	0.010	0.030	0.038		0.000
Q	0.026	0.017	0.031	0.005	0.027	0.013	0.000		0.045
R	0.035	**0.379**	0.067	**0.656**	0.075	**0.216**	0.002		0.242
S	0.042	0.008	0.061	0.001	0.022	0.076	0.001		0.000
T	0.031	0.003	0.045	0.001	**0.384**	0.113	0.033		0.000
V	0.029	0.000	0.025	0.000	0.048	0.073	0.010		0.000
W	0.004	0.000	0.000	0.001	0.031	0.000	0.005		0.000
Y	0.021	0.001	0.022	0.022	0.004	0.004	0.099		0.000
gap	0.481	0.358	0.264	0.235	0.202	0.147	0.113		0.435

Consensus:**KRKRTRFTPEQLEELEKEFEKNPYPSREEREELAKELGLTERQVKVWFQNRRAKWKK**

Figure 1. Extracting a consensus sequence from a multiple sequence alignment.

(A) An alignment of M=4,571 homeodomain sequences, each of length L=57 residues. Residues with similar (or unique) chemical features are highlighted to illustrate conservation. Positions with a gap in more than half the sequences are eliminated from the alignment. From such an alignment, the frequency of each type of residue (including remaining “gap residues”) is determined. (B) The consensus sequence (bottom) is obtained by taking the residue with the highest frequency (excluding gaps, see Section 3) at each position (red).

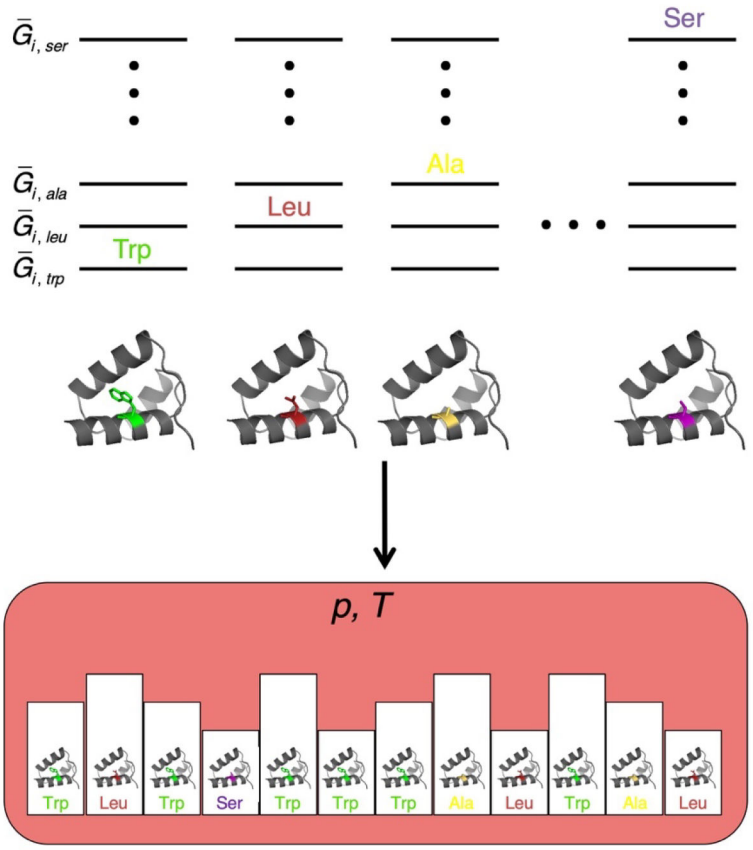


Figure 2. An ensemble model for sequence variation at a specific site in a native protein.

(Top) Gibbs free energy values (where the overbar indicates molar energies) for each of the 20 residues (plus a gap) at position *i* in the native state. Each side-chain free energy is determined by specific interactions between the side chain and the protein and solvent in all conformations available to the side chain. (Bottom) an isothermal isobaric ensemble (defined by the temperature and pressure of a large reservoir, pink) containing single-chain replicas that can exchange heat and volume with the reservoir. Side chains are allowed to equilibrate over the set *r* through alchemical transformation, producing equilibrium populations that reproduce residue frequencies at site *i* in an MSA, and are related to side-chain free energies through Boltzmann-like terms. For simplicity, only the four residues from panel A are shown.

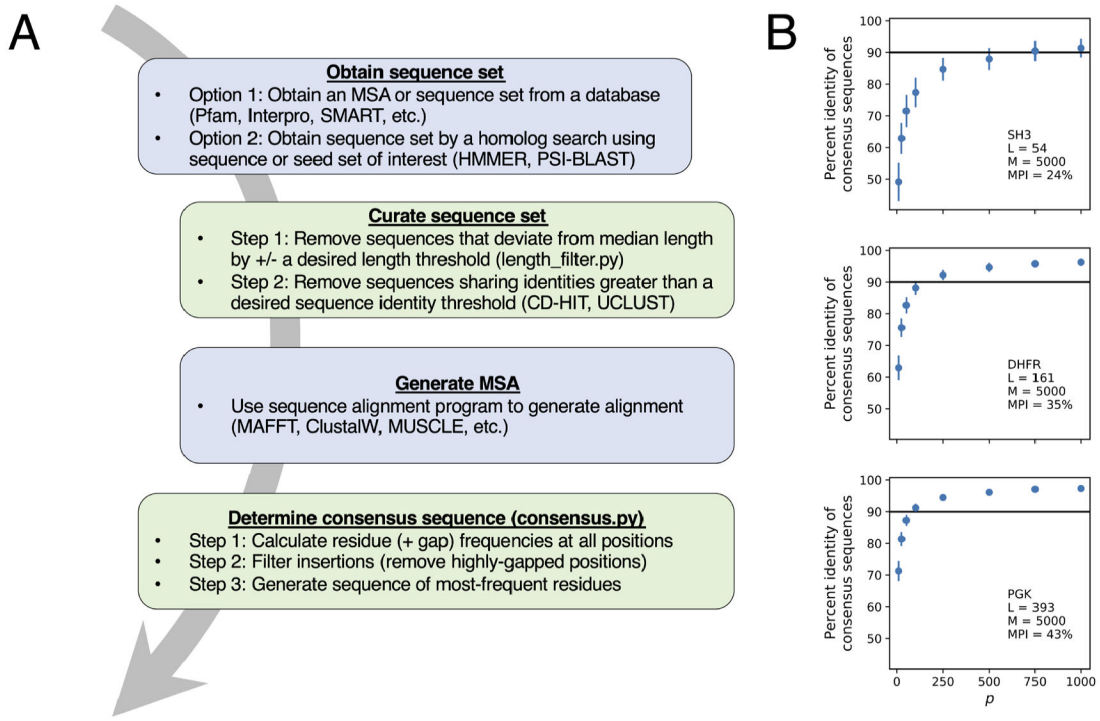


Figure 3. Generating a consensus sequence.

(A) Workflow for consensus sequence design. Major steps involve obtaining sequences, eliminating sequences with atypical length and high sequence identity, (re)aligning sequences, filtering gaps and insertions, and calculating residue frequencies. Examples of available databases and programs are highlighted where applicable. Green boxes contain steps that can be done using Python scripts noted in Section 3.3. (B) Identities among consensus sequences generated from random draws (sampling without replacement) of p sequences from a 5000 sequence MSA for SH3 domain (top), dihydrofolate reductase (middle), and phosphoglycerate kinase (bottom). Points represent the mean pairwise identity of 100 consensus sequences each obtained from random subsets of p sequences, and bars represent one standard deviation. Inset text indicates the length of each protein family (L), the number of sequences in each MSA (M), and the mean pairwise identity (MPI) of sequences in each 5,000 sequence MSA.

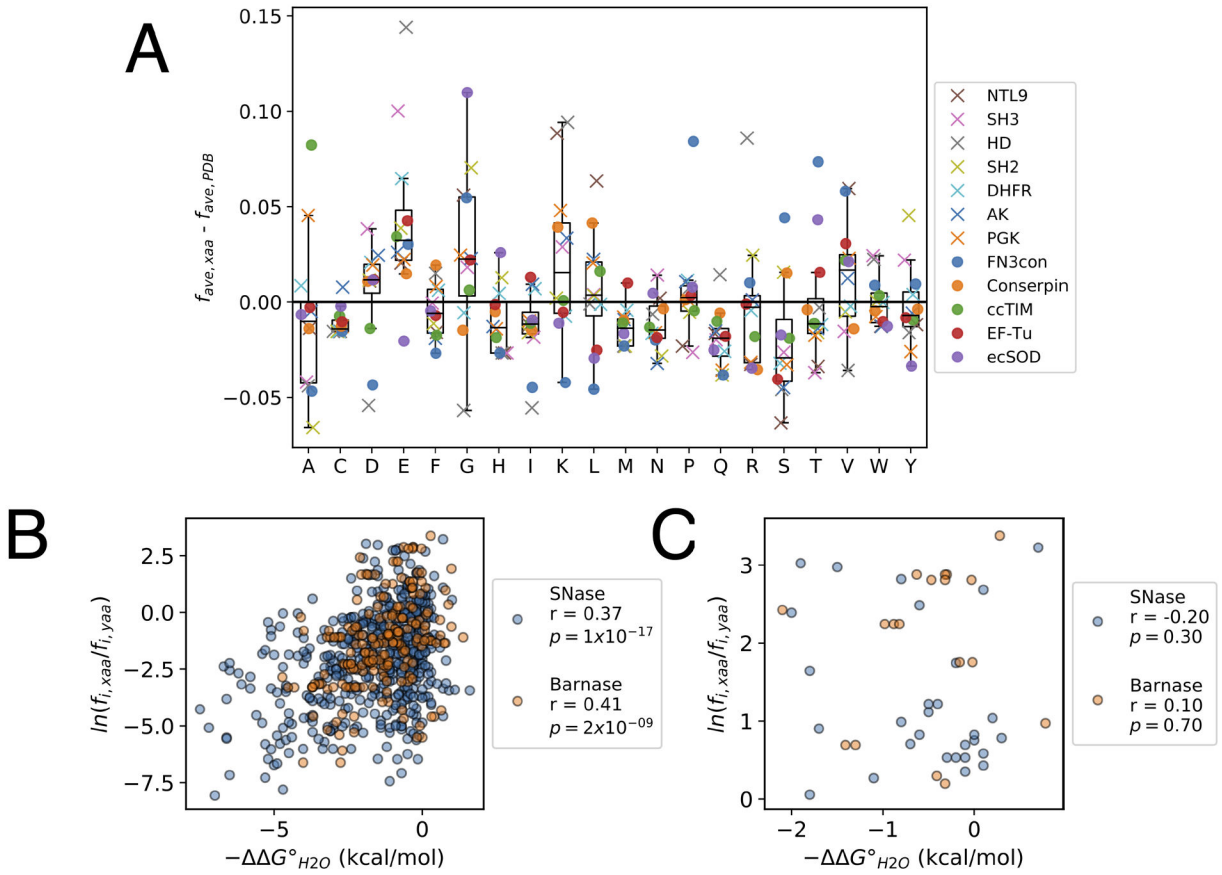


Figure 4. Features of consensus sequences.

(A) Average residue frequencies in consensus compared to extant sequences. Each marker represents the difference in the average frequency of a residue in a consensus sequence from the literature ($f_{avg,\ xaa}$; obtained by dividing the number of residues of a given type by the consensus sequence length L) and the average residue frequency of the same residue among sequences obtained from a nonredundant set from the PDB ($f_{avg,\ PDB}$). X's indicate sequences designed by our lab, circles indicate sequences designed in other labs (Table S1). The analysis above is limited to consensus sequences designed from MSAs with more than 100 sequences. The box indicates the interquartile range (IQR), the line within the box indicates the median, and the whiskers extend to the last data point within +/− 1.5*IQR. (B) and (C) Correlations between folding free energy changes ($\Delta\Delta G°_{H2O}$) from the Protherm database and the Boltzmann-like energies for point substitutions determined from the residue frequencies in MSAs in SNase (blue, n = 514 substitutions) and Barnase (orange, n = 200 substitutions). Pearson correlation coefficients (r) and p values (p) for the correlation coefficients are shown for each distribution. Panel (B) shows all substitutions; panel (C) shows only the subset of the substitutions to consensus residues from panel B.