Global analysis of multi-mutants to discover stabilizing amino acid substitutions

Kristoffer E. Johansson, Kresten Lindorff-Larsen* and Jakob R. Winther**

Linderstrøm-Lang Centre for Protein Science, Section for Biomolecular Sciences, Department of Biology, University for Copenhagen, Ole Maaloes Vej 5, DK-2200 Copenhagen N, Denmark

*To whom correspondence may be addressed: kristoffer.johansson@bio.ku.dk, lindorff@bio.ku.dk, jrwinther@bio.ku.dk

Abstract

The identification of stabilizing amino acid substitutions in proteins is a key challenge in protein engineering. Advances in biotechnology have enabled assaying of thousands of protein variants in a single high-throughput experiment, and more recent studies use such data in protein engineering. We present a Global Multi-Mutant Analysis (GMMA) that exploits the presence of multiply-substituted variants to identify individual substitutions that stabilize the functionally-relevant state of a protein. GMMA identifies substitutions that stabilize in different sequence contexts that thus may be combined to achieve improved stability. We have applied GMMA to >54,000 variants of green fluorescent protein (GFP) each carrying 1-15 amino acid substitutions. The method is transparent with a physical interpretation of the estimated parameters and related uncertainties. We show that using only this single experiment as input, GMMA is able to identify nearly all of the substitutions previously reported to be beneficial for GFP folding and function.

Introduction

A major challenge in practical uses of proteins is the engineering of protein stability while at the same time maintaining the function of the protein. New developments in biotechnology are continuously applied to address this challenge with high-throughput methods currently in focus. Synthesis, screening and sequencing may today be performed for thousands of protein variants in parallel via Multiplexed Assay of Variant Effects (MAVE) also known as deep mutational scanning [1, 2]. Such experiments can identify loss-of-function variants with high accuracy, but are often unable to gauge more subtle effects on stability. Stabilizing, neutral or mildly destabilizing substitutions are likely to have a minor, if any, detectable impact on protein function and are therefore more difficult to identify from such experiments. This has in general limited their direct applicability in protein engineering.

When the screened library of variants contains doubly- or multiply-substituted variants, statistical models have been used to investigate how the effect of individual substitutions combine in the read-out for observed variants. Global fitness models typically consider additive single-substitution effects in a latent space and transformed these to the assayed quantity via various non-linear functions to describe the data generated by a MAVE [3, 4]. One particularly relevant study further showed that a thermodynamic model could be used to improve the mechanistic understanding and to quantify the effects of single-substitutions on protein binding and structural stability [5]. This model was shown to capture structural stability well [6] and a similar approach was successfully applied to fit deleterious effects of multiply-substituted variants [4].

Here, we present a generally applicable method, Global Multi-Mutant Analysis (GMMA), that allows for the identification of amino acid substitutions that have a general stability-enhancing effect with little dependency on the sequence context, and thus substitutions with the potential to be additively combined for further enhanced stabilization. We demonstrate that single-substitution effects, in general, may be informed by multiply-substituted protein variants, which we here, for simplicity, refer to as multi-mutants. Since the analysis is based on a functional read-out of diverse sequence contexts, the identification of stabilizing substitutions is carried out while considering the assayed function of the protein. The central idea is to identify stabilizing substitutions by their ability to compensate destabilizing ones. Specifically, while a stabilizing substitution may not have a measurable effect in the background of an already stable protein, it can be identified in the background of one or more destabilizing substitutions. The concept is similar to the "partner potentiation" principle formulated previously [7], but here generalized to multi-mutants and with a method to understand and handle the parameter estimation challenges that arises with random multi-mutant libraries.

We have applied our analysis to an experiment that reports the fluorescence of >54,000 variants of green fluorescent protein (GFP) each containing 1–15 of the total 1,879 unique single-amino-acid substitutions observed in the experiment [3]. The GFP variants were generated using error-prone PCR (epPCR) and expressed in fusion with a red-fluorescent protein in order to correct for variations in expression levels. Then, fluorescence activated cell sorting (FACS) was used to divide the cells into eight fractions based on level of green fluorescence, and each fraction was sequenced to reconstruct a measure of fluorescence of each genotype. We have applied GMMA to the results of this single experiment to estimate the effects on stability of 1,107 single-substitutions, and identify stabilizing substitutions that do not compromise function and are directly applicable to engineering studies. As a validation of the approach, we identity a large number of substitutions that are known to be beneficial in improved GFP variants.

Results

Biophysical concept

The identification of stabilizing single-amino-acid substitutions in an already stabile protein, when based on a functional assay, is challenging because these may not improve the assayed function substantially. Also, stability assays based on general structural properties, e.g. circular dichroism, may not report on the particular state of the protein that is relevant for function. In order to address this situation, we define the states of a protein by the assayed property and consider an equilibrium between the active state, A, and an inactive state, D:

$$A \rightleftharpoons D \tag{1}$$

In analogy with traditional measurements of protein stability [8], we can probe this equilibrium by perturbing the system and measuring a deactivation transition. In the context of protein engineering the relevant change is to the protein sequence. Thus, we consider a "variable protein" and probe the stability related to Eq. 1 in amino acid sequence space. With this, the equilibrium does not describe a particular protein variant but rather a system that may be perturbed by changing the sequence of the protein.

The equilibrium is associated with a free energy of activation, denoted $\Delta G_v = G_v^A - G_v^D$ for a variant $v \in \{1 \ldots V\}$ in a library of V amino acid sequence variants that will each contain one or more

individual amino acid *substitutions*. From this free energy of activation, here simply referred to as the stability, we can calculate the fraction of active protein which is assumed to be proportional to the functional readout (see Methods). This probing of the equilibrium in sequence space has some analogy to conventional denaturation experiments where the equilibrium of a "fixed protein" is probed by changes e.g. in temperature or solvent composition [8].

The equilibrium Eq. 1 implies that amino acid substitutions stabilize or destabilize the protein via mechanisms that allow for compensation by other substitutions. Such effects are often found to be mostly independent and with additive free energies [7, 9, 10]:

$$\Delta G_v = \Delta G_{\mathrm{wt}} + \sum_{s \in v} \Delta\Delta G_s \qquad (2)$$

Here, ΔG_{wt} is the stability of the "wild type" (reference) sequence and $\Delta\Delta G_s$ is the stability effect of the substitution *s*. We note that our approach is not limited to additive effects and that couplings may be included in Eq. 2, as long as the data warrants estimation of these. Rather, the additivity in this particular model could be viewed as the desired output: We wish to identify those stabilizing substitutions that can compensate several different destabilized variants and thus be stabilizing additively in diverse backgrounds (Fig. 1a). This formulation is similar to previous global models [3-5, 10], but here applied to identify stabilizing substitutions and with a thermodynamic interpretation.

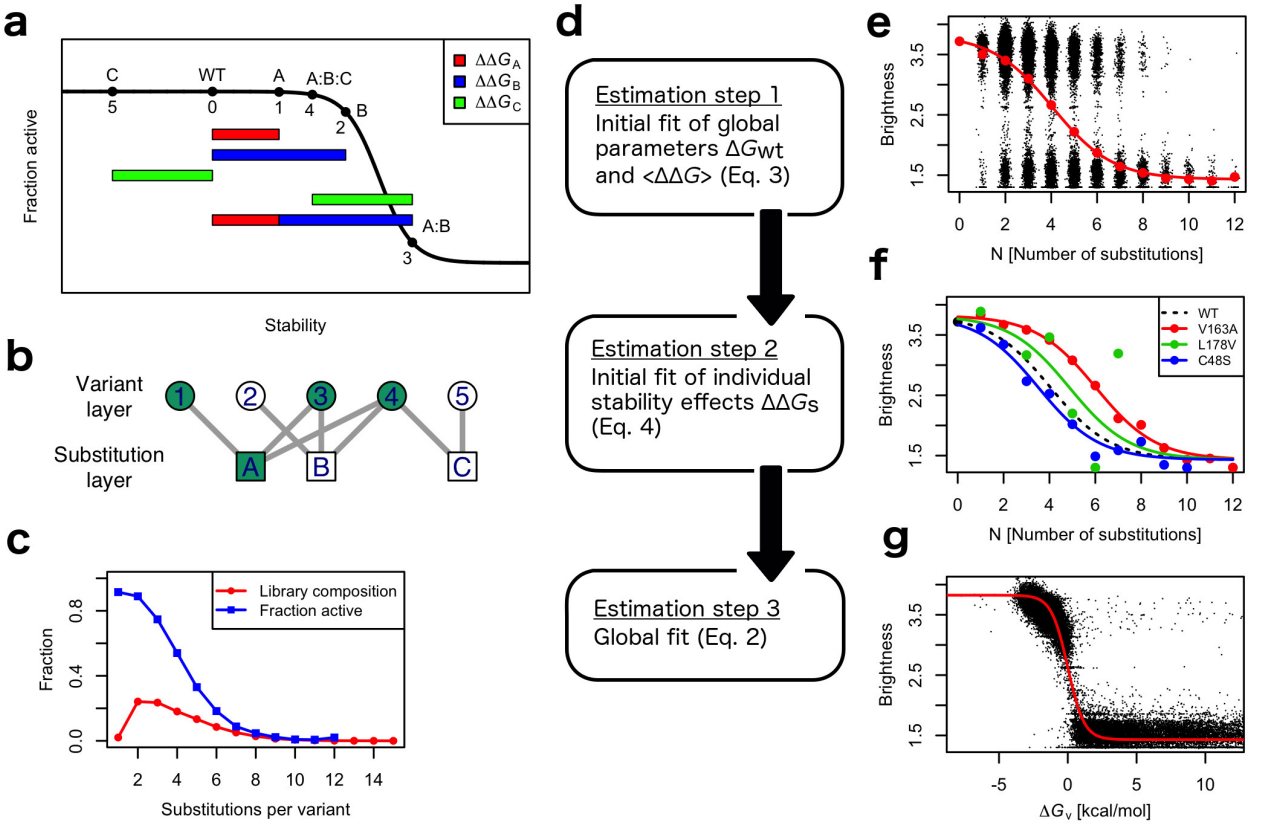

Figure 1. **(a)** A schematic outline of the GMMA approach. Consider a protein (WT) with five variants (named 1–5) that are composed of one or more of three substitutions (named A–C). The lengths of the colored bars represent the magnitude of the additive stability effects. All single-substitution

variants show wild-type-like activity, with the most destabilized variant, B, only being slightly less active than the wild type. While both variants A and B are active, the double mutant A:B is inactive. Thus, we infer the additive effect of A and B are both mildly destabilizing. Substitution C on its own does not appear to affect activity. However, when it is introduced into the inactive A:B background (to form A:B:C) it is able to compensate the loss of function, and we thus infer that substitution C is stabilizing with the magnitude of the green bar. (**b**) The five variants and the three substitutions may be represented in a bipartite graph, and as an example we highlight (in green) the subset of protein variants used to estimate the effect of substitution A. (**c**) Multi-mutant composition of the GFP library (red) shown together with the fraction of active variants (blue). The high population of variants in the transition region between 2 and 6 amino acid substitutions makes this excellent for GMMA. **(d)** Three-step estimation strategy. The three steps relate to panels e, f and g respectively **(e)** Fit of initial global parameters (red line) to the average brightness of variants with a given number of substitutions (red points) using Eq. 3. The GFP library is shown with random jitter in the horizontal coordinate to ease visualization (black dots). **(f)** Three examples of initial fits of individual stability effects (colored lines) compared to the WT fit from panel e (dashed line). Each $\Delta\Delta G_s$ is fit to the average *N*-mutant brightness (points) calculated only using the subset of variant that contains *s*. Mildly destabilizing substitutions, like C48S, lowers the endurance towards additional generally destabilizing substitutions and thus the system inactivates with fewer substitutions and the curve shifts left (blue line) compared to the wild-type (dashed line). On the other hand, stabilizing substitutions, like V163A, may be identified by their ability to increase the endurance towards substitutions, thus shifting the curve to the right (red line). Substitutions observed in few variants are difficult to fit, here demonstrated with L178V observed in only 14 variants compared to 677 and 194 variants for V163A and C48S respectively. **(g)** Result of the final global optimization of all stability effects (red line) to all variants (black dots).

The GMMA thus comprises a set of *V* equations, each describing the activity of a variant, and carrying a number of parameters, $\Delta\Delta G_s$, together with the global wild-type stability, ΔG_{wt}, and baseline parameters. For such a system of equations, it is important to have more data (variants) than parameters (substitutions), and this is possible with a multi-mutant library where a set of substitutions may be mixed in many different ways to make a larger set of variants. It is also important that all equations are coupled, and this may be tested by analyzing an undirected bipartite graph in which the protein variants constitute one layer of nodes (Fig. 1b, circles) and the unique substitutions the other layer (Fig. 1b, squares). This substitution-variant-graph formalism may be used to study many aspects of the multi-mutant library (see discussion section and supplementary Fig. S3). For example, the degree distribution of variant nodes gives the distribution of the number of substitutions in the variants (Fig. 1c), which shows that most variants contain two to six substitutions. The fraction of active variants per *N*-mutant (Fig. 1c, blue line) shows that the substitutions are in general destabilizing and that the stability of GFP is approximately 4 when measured in units of “general substitution effects”. GMMA identifies stabilizing substitutions by compensation of this general destabilization, and it is important that the variant library has an *N*-mutant distribution that covers the transition region where the system loses activity.

The inactive state, *D*, results from amino acid substitutions that are deactivating by reversible means, irrespective of mechanism, and is thus somewhat broadly defined. On the other hand, we term deactivating substitutions that cannot be compensated by stabilizing substitutions, e.g. removal of a crucial functional side chain, as *irreversible fatal substitutions* (IFS). IFS may be related

to the assayed function of the protein, stability or folding hotspots, or for GFP, related to the chromophore maturation reaction. Since they are dominantly deleterious for the active system, IFS will artificially appear as highly destabilizing although nothing may be inferred about their effect on conventional structural stability. This highlights a distinct advantage that GMMA will only identify substitutions that stabilize the active system defined by the assay.

In the interpretation of the GFP data it is important to realize that the destabilization-via-substitutions is qualitatively different from conventional *in vitro* unfolding experiments where unfolding is induced after the irreversible maturation. Indeed, one of the best-known enhanced variants of GFP is named *superfolder GFP* (sfGFP) and this name refers to the ability to fold and mature before creating non-fluorescent inclusion bodies [11]. For GMMA, the considered equilibrium, Eq. 1, may likewise lie *before* the maturation reaction with only the active state being able to mature. Furthermore, the physical situation may closer resemble a steady state where active protein is removed from the equilibrium by irreversible maturation, possibly in a chaperone dependent fashion [12], and inactive protein is removed, e.g. by protease degradation or aggregation. In this scenario, maturation kinetics could also influence the apparent stabilities and thus the outcome of GMMA. The sfGFP variant was selected to complement folding when GFP was destabilized by fusion to a poorly folding protein – an approach which indeed has some analogy to GMMA [11]. Here, we assume that the simple equilibrium Eq. 1. approximates the situation well and keep in mind that the absolute stabilities estimated here should not be compared directly to denaturation experiments of mature GFP.

Model estimation

The global fit of the effects of the individual substitutions is complex for at least two reasons. First, the global parameters including baselines (Fig. 1a, black line) are estimated simultaneously with all individual variant stabilities (Fig. 1a, abscissa values). The trade-off between adjusting the curve or the points may result in a highly rough optimization surface. Second, many substitutions are observed in only few variants and may be poorly determined with greater uncertainties that are combined with otherwise well-determined substitution effects. In order to address these complexities, we have developed a three-step estimation procedure that relies on initial estimates of individual stability effects (Fig. 1d). The first two steps focus on achieving initial estimates of global parameters (Fig. 1e) and individual stability effects (Fig. 1f) while the final global optimization in step three is only allowed to converge to the nearest optimum (Fig. 1g).

In the first step, an initial global wild-type stability, ΔG_{wt}, is estimated together with the average effect of a substitution, $\langle \Delta\Delta G \rangle$ (Fig. 1e). This average effect replaces the sum in Eq. 2:

$$\Delta G_v \approx \Delta G_{\mathrm{wt}} + N_v \langle \Delta\Delta G \rangle \qquad (3)$$

where, N_v is the number of substitutions in variant *v*. While this assumption is rather crude, it is only used in the estimates of initial values in step two, and it becomes robust with an increasing number of variants of each *N*-mutant (i.e. a variant with *N* substitutions). The average stability effect, $\langle \Delta\Delta G \rangle$, may be biased by IFS (which are in principle infinite) and these are therefore not used in the fit of ΔG_{wt} and $\langle \Delta\Delta G \rangle$ in steps one and two (see Methods).

In the second step, stability effects of individual substitutions are initially estimated by a similar approach (Fig. 1f). For each substitution, *s*, $\Delta\Delta G_s$ is estimated from the subset of variants that contains that substitution. The effect is included in Eq. 3 by replacing one term of the average stability with the specific stability effect of *s* to be estimated:

$$\Delta G_v \approx \Delta G_{\mathrm{wt}} + \Delta\Delta G_s + (N_v - 1)\langle\Delta\Delta G\rangle \quad (4)$$

A robust fit to Eq. 4 requires a sufficient number of *N*-mutants (to estimate the average brightness) for several different values of *N*, but in contrast to above we only consider the subset of variants that contains the substitution *s*. To illustrate the required number of variants per substitution, figure 1f shows three examples of fits to Eq. 4 where it is clear that the 14 observed variants containing L178V result in average values of the *N*-mutant brightness that do not fit the model well. In contrast, the stabilizing V163A and destabilizing C48S fit the model well with 677 and 194 observed variants, respectively. The three-step estimation strategy ensures that effects that are poorly estimated due to few observations, typically < 10-20 depending on the distribution on *N*-mutants (supplementary Fig. S1), do not affect the initial estimation for substitutions with good data. When all initial stability effects have been estimated individually, we perform a global optimization in step three using damped least-squares optimization from which uncertainties are calculated (see Methods).

We estimate the stability of the "wild-type" sequence (avGFP + F64L) to be -2.8 kcal/mol. This is substantially smaller in magnitude than values < -10 kcal/mol reported from unfolding experiments of αGFP (avGFP + Q80R:F99S:M153T:V163A); however, unambiguously determining the stability of GFP is challenged by at least one folding intermediate of stability -3.7 kcal/mol [13]. The discrepancy may also be explained by *in vitro* unfolding experiments being qualitatively different from the equilibrium probed in the GFP data considered here. As discussed above, GMMA probes the stability of the pre-mature GFP fold and its ability to mature irreversibly in the context of a cell, whereas unfolding experiments deactivates the mature protein *in vitro*. Since maturation is spontaneous, the pre-mature structure is likely less stable compared to the mature protein and thus, GMMA could probe a less stable structure compared to unfolding experiments of the mature protein.

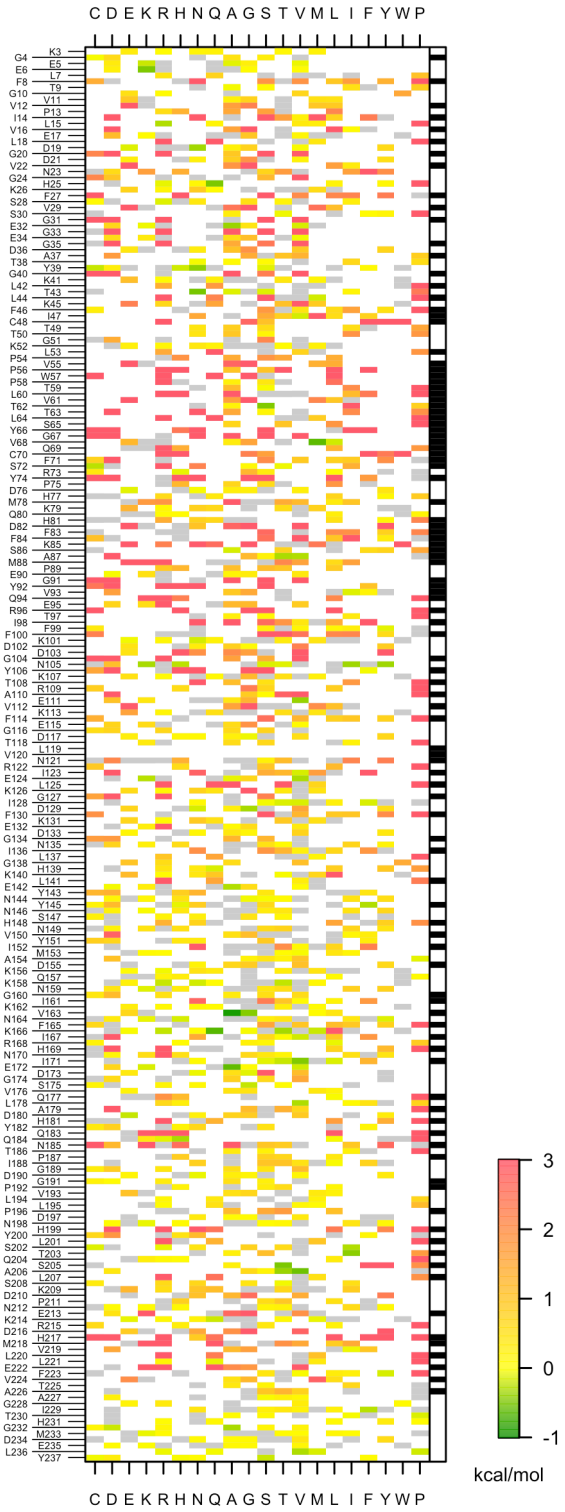

Figure 2. Heatmap showing the 1107 single-substitution effects estimated by GMMA from the multi-mutant library of GFP. Green indicates a stabilizing substitution, yellow are substitutions with close-to-zero effect, and orange-red indicate destabilizing substitutions. Substitutions in gray are observed in the library but poorly estimated and white substitutions are not observed in the data. The rightmost column shows the solvent exposed residues in white and buried residues in black.

GFP substitution effects

Using GMMA we obtain accurate estimates for 1,107 substitution effects (59% of the 1,879 present in the library) with 80% found to be destabilizing and 8% stabilizing (Fig. 2). The majority of the accurately estimated effects are found at solvent exposed positions (712/1,107; 64%) which includes most of the stabilizing (63/83; 76%; Fig. 3) and almost all of the substitutions with insignificant effect (126/140; 90%). We find an enrichment of stabilizing substitutions at positions with solvent exposed Phe, Ile and Leu, so that 2/11 (18%) substitutions from Phe, 4/18 (22%) from Ile and 5/35 (14%) from Leu are stabilizing, compared to a total of 63/712 (9%) stabilizing substitutions at surface positions. Perhaps more surprising, surface positions with wild-type Glu have a significant higher fraction of stabilizing substitutions, 12/71 (17%), as compared to Asp with only 5/90 (6%) substitutions being stabilizing. This may suggest a qualitative difference in the role of these two negatively charged amino acids (that otherwise have the same transitions in the codon table) in the context of a beta-barrel.

While most stabilizing substitutions are found at solvent exposed positions, a notable exception is the buried strand position V163 for which the smaller side chains Ala and Gly are among the most stabilizing in our analysis with V163A as the most highly stabilizing (0.94 ± 0.04 kcal/mol). Also notable is that positions with wild-type Ile show a relatively high fraction of stabilizing substitutions (10/68; 15%) with relatively more at exposed positions (4/18; 22%) compared to those that are buried (6/50; 12%). All 42 substitutions at all ten Pro residues in GFP are found to be destabilizing, which is not surprising considering the structural role of Pro. Other positions that only show highly destabilizing substitutions include the chromophore positions Y66 and G67 and the maturation-related sites R96 and E222 [14].

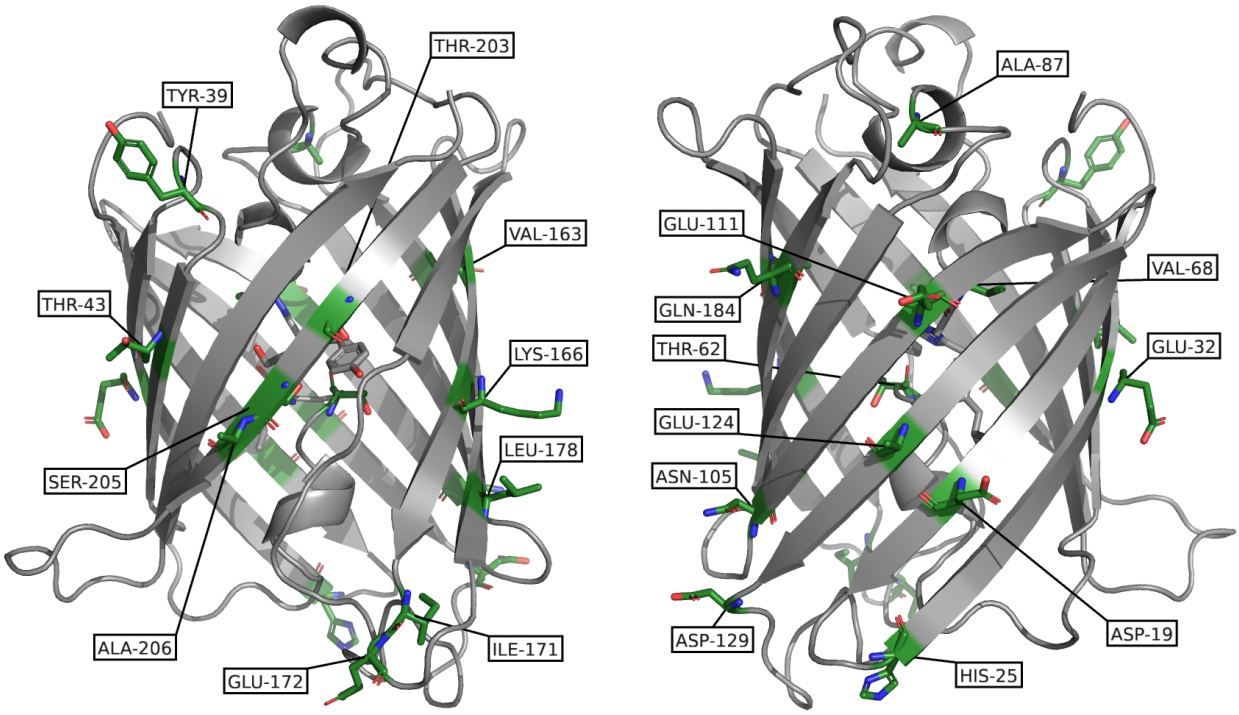

Figure 3. Positions of the top 30 stabilizing substitutions shown on the structure of GFP (PDB ID 1EMM). Some positions have more substitutions within top 30 and positions of rank 9, 25 and 29 (E3, G232 and L236) are not resolved in this structure.

Analysis of known GFP substitutions

To examine whether the substitutions we identify using GMMA indeed have stabilizing and generally background-insensitive effects, we compare these to substitutions that are known to be beneficial for GFP [15]. We focus on enhanced variants that are relevant for our "wild-type" reference sequence (avGFP + F64L) and consider the substitutions that constitute superfolder GFP (sfGFP) [11], T-Sapphire GFP (tsGFP) [16], a split GFP (splitGFP) [17], a computationally-optimized GFP known as des11 [12] and a tryptophan chromophore variant called nowGFP [18] (supplementary Table S1). We expect substitutions from these to be stabilizing or have an insignificant effect in our GMMA. A one-to-one comparison is not possible because substitutions in these variants have typically not been characterized individually. Additionally, we also consider 136 substitutions found in 147 structures of GFP, selected in the PDB to have >90% identity to our wild-type sequence. Since these variants have all been expressed and crystalized, we expect that these substitutions in general do not destabilize GFP substantially, i.e. slightly lower expectations compared to substitutions from the enhanced variants mentioned above.

As expected, the substitutions that constitute sfGFP, tsGFP, splitGFP, nowGFP and des11 are in general found to be stabilizing or with insignificant effect (Fig. 4 and supplementary Fig. S2). We find that 19 of the top 30 stabilizing substitutions obtained from GMMA are described in the literature including 4 of the 10 sfGFP substitutions, 2 of the 5 from tsGFP, 7 of the 17 from splitGFP and 5 of the 20 from nowGFP (Fig. 3). In general, the substitutions known to be favorable tend to rank high in GMMA and within the top 100 most stabilizing, 46 are known from previous studies and cover 39% of all the known substitutions included here (supplementary Fig. S2). Known substitutions that are estimated by GMMA to be destabilizing may indeed also be accurate because they were originally selected for other purposes than stability. Perhaps most notable are the chromophore substitution S65T, selected for spectral properties, and splitGFP C48S, introduced to avoid cysteine oxidation in extracellular environments. Likewise, the substitutions Q80R (sfGFP) and H231L (nowGFP, tsGFP and des11), here estimated to be slightly destabilizing, are historical substitutions caused by early PCR errors which are still present in some synthetic genes of GFP variants [19]. The most destabilizing substitutions present in nowGFP either interact directly with the Trp chromophore (V150A) or are reported to indirectly support the chromophore (N146I, Y151N and L207Q) [20, 21]. Thus, the destabilizing effect of these substitutions may indeed be accurate with the Tyr chromophore in our wild-type. The only other sfGFP substitution found to be slightly destabilizing in our analysis, namely F99S with 0.26 ± 0.04 kcal/mol, emerged from an early optimization of GFP using DNA shuffling and was speculated to remove a hydrophobic Aequorin interaction site [22]. This is sometimes replaced with F99T [23] whereas our analysis suggests that F99Y would be a better choice. None of the known substitutions were estimated to make the protein unstable ($\Delta\Delta G > \Delta G_{wt}$) and only four substitutions found in the PDB were estimated to destabilize by more than half of the wild-type stability.

Interestingly, while des11 has 11 substitutions, a 12th substitution (V68G) also resulted from the computational analysis, but was later excluded based on experimental assessment [12]. In agreement with this, GMMA estimates V68G to be highly destabilizing (1.6 ± 0.4 kcal/mol; more than twice the average effect) and highlights another substitution, V68M, known from nowGFP and estimated by GMMA to be the third most stabilizing with -0.65 ± 0.08 kcal/mol.

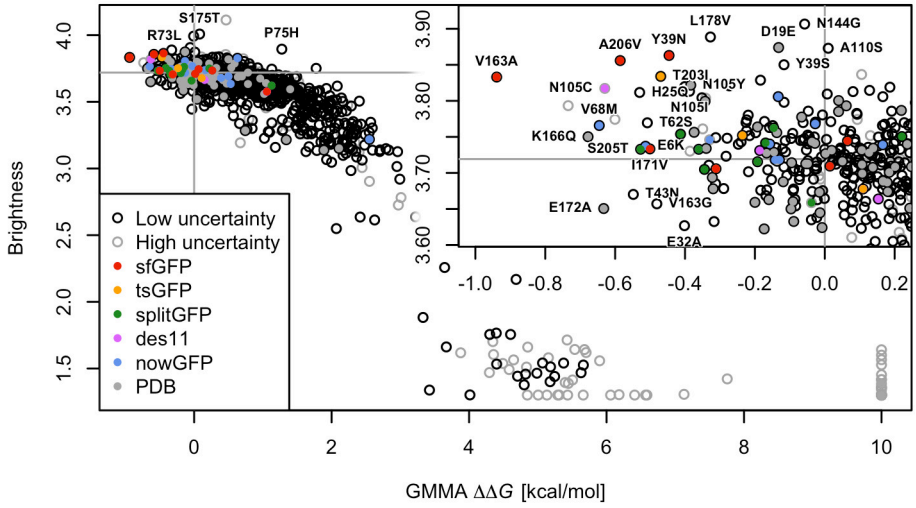

Figure 4. Brightness versus GMMA stabilities for single amino acid variants. GMMA estimates with large uncertainties are shown in grey circles. Substitutions known from optimized versions of GFP are shown in color. The vertical lines mark $\Delta\Delta G$=0 and the horizontal lines the brightness of the wild-type sequence. The insert shows the highly stabilizing, high brightness region and is enriched in known GFP variants.

The stability effects estimated by GMMA correlate strongly with the observed brightness of singly-substituted variants (Fig. 4; Spearman's correlation coefficient, r_s=-0.77). This, however, is mainly caused by the fact that the averaged FACS readout and GMMA analysis agree on destabilizing/low brightness variants. In contrast, when the goal is the more challenging one of identifying stabilizing variants we see a much weaker correlation between the observed brightness and GMMA (r_s=-0.26 for variants identified by GMMA as significantly stabilizing; see Methods) because individual stabilizing variants have only modest effects when introduced in an already stable background. Thus, GMMA is able to separate this population according to their stability effects with the sfGFP, splitGFP and des11 substitutions estimated to be stabilizing or with a minor effect (Fig. 4, insert).

Discussion

We have demonstrated the ability of GMMA to identify stabilizing substitutions that are insensitive to other substitutions and facilitate the function of the protein as defined by a specific assay. In relation to protein engineering, it is interesting that the chromophore substitution S65T is found to be destabilizing by GMMA even though this has previously been selected as enhancing in a fluorescence assay similar to the one considered here [24]. Indeed, in protein engineering, it is not rare that function-enhancing substitutions compromise stability as it is seen here. While S65T specifically enhances function, it is, however, found to decrease the endurance towards additional generally destabilizing substitutions and thus deemed destabilizing by GMMA. Whereas the former function-enhancing effect may be identified from a single-mutant analysis as in traditional selection approaches, the latter stability effect results from the multi-mutant analysis which is insensitive to the level of function of the single-mutant as long as this is active. This indicates that the catalogue of GMMA effects may be complemented by single-mutant effects from traditional screening approaches, or interestingly, the single-mutant read-out from the same assay as used by the GMMA if this is within the desired accuracy.

For the practical construction of a GMMA multi-mutant library, different approaches may be considered. The GFP multi-mutant library which we have analyzed was generated by epPCR. The variant-substitution-graph (Fig. 1b) indicated the presence of "hub-substitutions" that are observed in more variants than expected from random (supplementary Fig. S3). These could emerge both from biases in the codon table, and from sequential rounds of epPCR where early substitutions are inherited in the following rounds. Most notably is N121S that is observed in 982 variants together with 847 other substitutions (i.e. more than half of the 1,879 unique single-amino-acid substitutions), and results in a substantial parameter correlation with the global stability parameter in the initial fits. A narrower degree distribution, e.g. a more homogenous distribution around 100-200 variants per substitution (compared to supplementary Fig. S3), could be beneficial for GMMA, although this would not always solve the problem that arises when observing either only inactive or only active variants for a given substitution. Several technologies are available for making random, semi-random or defined DNA libraries that may be explored for designing GMMA libraries. However, we speculate that a library with a broad range of *N*-mutants (i.e. containing variants with both many and few substitutions) may provide a beneficial range of backgrounds in which each substitution is observed. The random connections between variants and substitutions gives an entangled graph where systematic biases from specific side-chain interactions are avoided, presumably making the approximation of additivity (Eq. 1) more reliable. Also, while a library consisting only of *N*-mutants for a fixed value of *N* could have a stability distribution that lies around the inactivation transition, it might be more sensitive to systematic biases and estimation artifacts. For example, consider a hypothetical data set of 4-mutants only. Here, an increase, δG, in the wild-type stability may be fully compensated by a decrease in all substitution effects of $\delta G/4$ to make an equally good fit of Eq. 2 (an isoline in the optimization hyper-surface). Furthermore, for the estimation strategy presented here, a broad range of *N*-mutants is necessary in the initial stability estimations (Fig. 1e and 1f). Finally, we highlight that GMMA in the presented case appears robust with a random library and thus offers a cost-effective approach for protein engineering given the low cost of epPCR.

Conclusions

We have here introduced an approach to extract information about individual substitutions through a global analysis of noisy, high-throughput measurements on randomly generated multi-mutants. We argue that for such experiments, the effect of single amino acid substitutions may be better determined by analysis of multi-mutant variants and have presented a Global Multi-Mutant Analysis (GMMA) that implements this. Important features of the multi-mutant library are that the distribution of stabilities lies around the inactivation transition and that the variant-substitution-graph is connected with many more variants than substitutions.

Because GMMA works by finding stabilizing substitutions that can compensate effects of several different destabilizing substitutions it is particularly suitable for identifying substitutions with additive effects. This makes the method ideal for protein engineering and we find indications that GMMA may complement measurements of single-mutant function which is particular interesting for enzyme engineering. The robustness of the method makes it cost-effective and the presented results are obtained from a single high-throughput experiment based on a random genetic library and a relatively simple assay. Thus, it should be applicable to any protein which is amenable to a simple high-throughput screen, and, importantly, can be applied to an already very stable starting

point. This makes the method applicable to systems that are difficult to study via traditional optimization approaches, and does not require access to automated high-throughput facilities.

References

[1] Fowler DM, Fields S (2014) “Deep mutational scanning: A new style of protein science” Nat Methods 8, 801-807.
[2] Kinney JB, McCandlish DM (2019) “Massively Parallel Assays and Quantitative Sequence–Function Relationships”. Annu Rev Genom Hum Genet 20, 99–127.
[3] Sarkisyan KS, Bolotin DA, Meer MV, Usmanova DR, Mishin AS, Sharonov GV, Ivankov DN, *et al*. (2016) “Local fitness landscape of the green fluorescent protein” Nature 533, 397-401.
[4] Starr TN, Greaney AJ, Hilton SK, Crawford KHD, Navarro MJ, Bowen JE, Tortorici MA *et al.* (2020) ”Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding” Cell 182, 1295-1310.
[5] Otwinowski, J (2018) “Biophysical inference of epistasis and the effects of mutations on protein stability and function”. Mol Biol Evol 35, 2345–2354.
[6] Nisthal A, Wang CY, Ary ML, Mayo SL (2019) “Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis” Proc Natl Acad Sci USA 116, 16367–16377
[7] Araya CL, Fowler DM, Chen W, Muniez I, Kelly JW, Fields S (2012) “A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function” Proc Natl Acad Sci USA 109, 16858–16863.
[8] Lindorff-Larsen K, Teilum K (2020) “Linking thermodynamics and measurements of protein stability” arXiv:2010.12281.
[9] Ashenberg O, Gong LI, Bloom JD (2013) “Mutational effects on stability are largely conserved during protein evolution” Proc Natl Acad Sci USA 110, 21071-21076.
[10] Otwinowski J, McCandlish DM, Plotkin JB (2018) “Inferring the shape of global epistasis” Proc Natl Acad Sci USA 115, 7550–7558.
[11] Pédelacq J-D, Cabantous S, Tran T, Terwilliger TC, Waldo GS (2006) “Engineering and characterization of a superfolder green fluorescent protein” Nat Biotechnol 24, 79-88.
[12] Bandyopadhyay B, Goldenzweig A, Unger T, Adato O, Fleishman SF, Unger R (2017) “Local energetic frustration affects the dependence of green fluorescent protein folding on the chaperonin GroEL” J Biol Chem 292, 20583–20591.
[13] Huang J, Craggs TD, Christodoulou J, Jackson SE (2007) “Stable Intermediate States and High Energy Barriers in the Unfolding of GFP” J Mol Biol 370, 356–371.
[14] Sniegowski JA, Lappe JW, Patel HN, Huffman HA, Wachter RM. Base catalysis of chromophore formation in Arg96 and Glu222 variants of green fluorescent protein. J Biol Chem 280: 26248–26255, 2005.
[15] Lambert TJ (2019) “FPbase: A community-editable fluorescent protein database”. Nat Methods 16, 277–278.
[16] Zapata-Hommer O, Griesbeck O (2003) “Efficiently folding and circularly permuted variants of the Sapphire mutant of GFP” BMC Biotechnol 3, 5.
[17] Do K, Boxer SG (2011) “Thermodynamics, kinetics, and photochemistry of beta-strand association and dissociation in a split-GFP system", J Am Chem Soc 133, 18078-18081.
[18] Sarkisyan KS, Goryashchenko AS, Lidsky PV, Gorbachev DA, Bozhanova NG, Gorokhovatsky AY, Pereverzeva AR (2015) “Green fluorescent protein with anionic tryptophan-based chromophore and long fluorescence lifetime”. Biophysical J 109, 380-389.

[19] Tsien RY (1998) "The green fluorescent protein". Annu Rev Biochem 67, 509–544.
[20] Heim R, Tsien RY (1996) "Engineering green fluorescent protein for improved brightness, longer wavelengths and fluorescence resonance energy transfer" Curr Biol 6, 178–182.
[21] Sarkisyan KS, Yampolsky IV, Solntsev KM, Lukyanov SA, Lukyanov KA, Mishi AS (2012) "Tryptophan-based chromophore in fluorescent proteins can be anionic" Sci Rep 2, 608.
[22] Cremeri A, Whitehorn EA, Tate E, Stemmer WPC (1996) "Improved green fluorescent protein by molecular evolution using DNA shuffling" Nat Biotechnol 14, 315-319.
[23] Valbuena FM, Fitzgerald I, Strack RL, Andruska N, Smith L, Glick BS (2020) "A photostable monomeric superfolder green fluorescent protein" Traffic 21, page 534–544
[24] Cormack BP, Valdivia RH, Falkow S (1996) "FACS-optimized mutants of the green fluorescent protein". Gene 173, 33-38.
[25] Rost B, Sander C (1994) "Conservation and prediction of solvent accessibility in protein families". Proteins: Structure, Function, and Bioinformatics 20, 216–226.

Acknowledgements

We thank Prof. Amelie Stein for discussions and insights. This work was supported by Independent Research Fund Denmark and the PRISM (Protein Interactions and Stability in Medicine and Genomics) centre funded by the Novo Nordisk Foundation (NNF18OC0033950).

Contributions

KEJ and JRW conceptualized the method. KEJ, KLL and JRW developed the method and analyzed the data. KEJ implemented the software and wrote the manuscript with contributions from all authors.

Methods

We assume that the observed activity, *F*, of a variant, *v*, is proportional to the fraction of the protein found in the active state A:

$$F \propto f_v = \frac{[\mathrm{A}]}{[\mathrm{A}] + [\mathrm{D}]} = \frac{1}{1 + \exp[\Delta G_v / RT]} \tag{5}$$

where [A] and [D] are the concentrations of active and inactive protein respectively, *R* is the gas constant and *T* the temperature. Following the original report of the data [3], the log fluorescence, or brightness *F*, is assumed proportional to the fraction of active protein. We carried out all data analyses using the *R project for statistical computing* with packages *minpack.lm* and *igraph*, and our code for GMMA is available from https://github.com/KULL-Centre/papers/tree/master/2020/multi-mutant-analysis-Johansson-et-al.

Initial estimation

The initial estimation of stability effects considers the *average* effect of substitutions, $\langle \Delta\Delta G \rangle$, which is sensitive to highly destabilizing substitutions. Thus, we first detect *irreversible fatal substitutions* (IFS) that are inactive in all contexts and have an unlikely pattern of activity among multi-mutants, e.g. many inactive double mutants in the case of GFP (see supplementary Appendix 1). Variants containing nonsense mutations are also generally expected to be IFS. We exclude 62 nonsense IFS and 51 missense IFS from the initial estimation together with the 2,310+4,851 variants that contain these nonsense and missense IFS substitutions respectively.

The wild-type stability, ΔG_{wt}, is first estimated together with the average effect of all substitutions, $\langle \Delta\Delta G \rangle$, using Eq. 6 which is simply a combination of Eqs. 3 and 5. These are fitted to the average brightness of the *N*-mutants, $\langle F \rangle_N$, i.e. the average brightness of double-mutants, triple-mutants, etc.

$$\langle F \rangle_N = \frac{\alpha_{\mathrm{A}} + \alpha_{\mathrm{D}} \exp[-(\Delta G_{\mathrm{wt}} + N\langle \Delta\Delta G \rangle)/RT]}{1 + \exp[-(\Delta G_{\mathrm{wt}} + N\langle \Delta\Delta G \rangle)/RT]} \tag{6}$$

Here, a constant baseline, α_{D}, for the brightness of the inactive variants is fitted whereas a constant baseline for active variants, α_{A}, is not fitted independently but calculated from ΔG_{wt} and the brightness of the wild-type sequence, F_{wt}, during fitting:

$$\alpha_{\mathrm{A}} = F_{\mathrm{wt}} + (F_{\mathrm{wt}} - \alpha_{\mathrm{D}}) \exp[-\Delta G_{\mathrm{wt}}/RT] \tag{7}$$

This makes α_{A} less sensitive to noise and outliers in the variant readout by relying on the data point (ΔG_{wt}, F_{wt}) which is experimentally well-determined with 3,645 barcodes in the high-throughput assay, i.e. individual observations of wild-type nucleotide sequence (2,444) or synonymous sequences [3]. A standard error of each parameter is calculated as the square root of the diagonal of the inverse Hessian matrix. The initial wild-type stability is estimated to $\Delta G_{\mathrm{wt}} = -1.8 \pm 0.2$ kcal/mol, the average effect of substitutions $\langle \Delta\Delta G \rangle = 0.45 \pm 0.04$ kcal/mol, and $\alpha_{\mathrm{D}} = 1.43 \pm 0.03$ (Fig. 1e). This results in $\alpha_{\mathrm{A}} = 3.8$.

In step 2, we use the values of ΔG_{wt}, $\langle \Delta\Delta G \rangle$ and the baseline parameters from step 1, to estimate initial values of the individual substitutions, $\Delta\Delta G_s$, from the subset of variants that contains the substitution, *s*, using Eqs. 4 and 6 (Fig. 1f). The use of a fixed value of $\langle \Delta\Delta G \rangle$ makes the initial estimates robust and self-consistent for the global fit, and the approach is further supported by the observation that variant effects are mostly independent of the details of the background [9]. Since we always only change the background with a single substitution, $\langle \Delta\Delta G \rangle$ is not expected to vary much and in practice less than the uncertainty that results from small sets of variants. We require that all subsets have a diversity of at least 3 different multi-mutants that spans the transition region and gives a fit with standard error < 1.0 and an absolute deviation < 5.0 log fluorescence units. With this, we find that 56% of substitutions have sufficient data for this initial estimation. We then afterwards estimate the remaining initial $\Delta\Delta G_s$ values from these well-determined effects.

Graph analysis

We pre-process the data for the global multi-mutant analysis, and build a bipartite graph by assigning protein variants to one layer of nodes and another layer of individual substitutions. All variant nodes are linked to the substitution nodes that the variant is composed of. We check that all nodes are connected in the graph. If a subset of variants is composed of a subset of substitutions that does not occur in the rest of the variants, this graph becomes disconnected from the rest, and GMMA may be carried out on the subset alone. Since a single-mutant library is fully disconnected, GMMA cannot be applied to this. Single mutants do not inform the global fit more than any other variants and indeed 130 stability effects are estimated without the single mutant being observed. Substitution nodes with a single link represents substitutions that contributes one parameter and only one data point to the global analysis, referred to as hanging nodes. These do not inform the global optimization and, thus, the effect of these are calculated after the global optimization. In summary, the graph is cleaned for 7 disconnected node pairs, and 257 hanging substitution nodes together with 255 dependent variants nodes. Finally, the graph is checked to ensure that no pair of

substitutions only occurs together as these would make them impossible to distinguish and should be reparametrized as a single effect. The graph analysis is independent of the initial estimates and considers all data, including IFS and non-sense substitutions.

Global estimation

With input from both the initial estimation and the graph analysis, we perform the global optimization using the same damped non-linear least squares algorithm as above. We optimize all stability effects $\{\Delta\Delta G_s\}$ and the wild-type stability ΔG_{wt} to the nearest optimum (Fig. 1g). The baselines are fixed at the values determined in the initial fit (1.4 and 3.8). The stability effects are limited to the range -5 to 10 kcal/mol for robustness. With analytically calculated gradients, the global optimization of 1,616 parameters from 53,763 data points took 4-5 hours on a normal laptop.

The fit has a reduced chi-square $\tilde{\chi}^2 = 4.4$ which indicates that some parts of the data do not fit the model, however, we note that the risk of overfitting is in general low. One contribution to an elevated $\tilde{\chi}^2$ could be the use of the distribution of the observed brightness of the wild-type sequence as a proxy for the variation of all variants, which may indeed have higher uncertainties. Since our aim is robust identification of additive and stabilizing substitutions, we accept that other models may indeed explain the data better, e.g. by including higher order terms in Eq. 2.

Uncertainties

In the global analysis, we estimate the uncertainties from the covariations in the inverted Hessian matrix. We calculate two error measures to judge the uncertainty and reliability of the stability effects. The first, δ_s, is calculated using the log-fluorescence measurement uncertainty, reported to be $\delta F_{\text{wt}} = 0.11$ for the wild-type sequence [3], propagated via the covariance matrix diagonal and multiplied by 3 to get the 99.7% percentile (and to somewhat compensate for the expectation that the variants have higher uncertainty than the wild type):

$$\delta_s = 3 \left| \frac{\partial \Delta\Delta G_s}{\partial F} \right| \delta F_{\text{wt}} \tag{8}$$

A resampling experiment suggests that this measure captures the estimation accuracy well and we report this uncertainty as the stability error with ± notation (supplementary Fig. S1).

The second uncertainty, $\delta_{\Delta\Delta Gs}$, is used to filter out unreliable stability effects and is calculated from:

$$\delta_{\Delta\Delta Gs}{}^2 = \left(\frac{\partial \Delta\Delta G_s}{\partial F} \right)^2 \frac{\text{RSS}_s}{\text{DOF}_s} \frac{1}{V_s} \tag{9}$$

Again, the derivative is from the diagonal of the covariance matrix, RSS_s is the residual sum-of-squares of the variants used to estimate substitution *s*, DOF_s is the number of degrees-of-freedom and V_s is the number of variants used to estimate substitution *s*. The last factor gives an error-of-the-mean type of uncertainty that compensates for the case where a lucky fit of few variants is penalized. The number of degrees of freedom for a substitution assumes that a uniform fraction of parameters is estimated together with $\Delta\Delta G_s$ from the V_s data points

$$\text{DOF}_s = V_s\left(1 - {}^{S}/_{V}\right) - 2$$

where *S* and *V* are the total number of substitutions and variants respectively. We set a relatively conservative threshold and mark 772 (61%) stability effects with $\delta_{\Delta\Delta Gs} > 0.05$ kcal/mol as unreliably (a value that can be compared to $\langle \Delta\Delta G \rangle \approx 0.5$ kcal/mol). This low threshold has been set by manual inspection of plots that show the fit of each substitution to its respective subset of

variants (similar to Fig. 1f). We use a hard threshold here to facilitate a clear discussion. However, in a specific application, substitutions could be judged individually based on both uncertainty measures and such plots, since many of the 772 poorly estimated substitutions are still informative. Notably, 222 (12% of all substitutions) are exclusively or predominantly observed in inactive variants and are therefore only represented in the flat region of the model. Thus, all of these may reliably be identified as destabilizing or even well determined on a range, even though the reported point estimate itself is highly uncertain. Of the remaining 550 (29%) substitutions with uncertain effects, the majority (398, 21%) are caused by poor statistics with five or fewer observed variants. The conservative threshold does exclude some substitutions with more than 10 observations (61 or 3%), that are potentially stabilizing, e.g. L221V, Q80K or E6A, and may be interesting depending on the application.

Classification

For the sake of discussion and early IFS identification, we classify all variants as either active or inactive. Variants with log fluorescence below 2.7, half-way between maximum and minimum observed log fluorescence in the original data, are assigned as inactive and the rest as active.

We mark substitutions with low uncertainty in the scores from GMMA (high/low uncertainty classification described above with the uncertainty calculation) as *significantly stabilizing* if the effect plus uncertainty is less than zero, *destabilizing* if the effects minus uncertainty is greater than zero and *insignificant* otherwise. Substitutions with high uncertainty are marked as destabilizing if the effect is more destabilizing than the wild-type stability and marked as unknown otherwise.

Solvent exposure categories are exposed or buried according to DSSP [25].