Automated techniques for the complete site-directed mutagenesis and stability analysis of protein domains

Chapter 3

Adapted from a manuscript coauthored with Stephen L. Mayo.

Abstract

The development of scoring functions for predicting protein stabilities requires large amounts of high-quality data. All current general-purpose stability prediction software was trained on the ProTherm database, an aggregate dataset of all stability data reported in the literature. While extremely useful, the datasets extracted from the database suffer from the following limitations: (1) data collected are from a wide variety of labs, experimental assays, and conditions, (2) the mutational distribution is biased towards large-to-small mutations, and (3) only positive measurements are reported, ignoring insoluble/unfolded sequences. To address these concerns, we initiated a large-scale project to facilitate the systematic construction of every single mutant of any particular protein domain. We developed high-throughput automation technology and established an experimental pipeline for the ordering, mutagenesis, sequence verification, expression, purification, and stability analysis of single-site protein mutants. The first domain we processed was Gβ1, a 56 residue beta-grasp (ubiquitin-like) fold, which entailed the construction of ~ 1000 single-site variants. This dataset, managed by relational database software, is already significant as it contains precise and accurate data on a large number of both folded and unfolded protein sequences. It is anticipated that single mutant data will be periodically added from domains on the order of 100–250 residues and featuring vastly different folds. The unique features of the current dataset are expected to directly benefit the optimization and validation of future stability prediction potentials.

Introduction

Site-directed mutagenesis has long been a potent tool for elucidating the principles governing protein function (1). Much of the knowledge we have today on protein stability was determined by introducing point mutants into specific positions and correlating the change in molecular structure with the accompanying change in free energy (2–6). This understanding of the forces behind protein stability has in exchange allowed insights into disease mechanisms (7) and unlocked the field of protein engineering and design (8, 9). Recently, much interest has surrounded the ability to predict the stability of protein mutants from their wild-type structure in order to minimize the experimental burden of constructing and evaluating mutants (10–15). Fresh critical analysis however, shows that most methods perform equally, and all have plenty of room for improvement (16, 17). One possible reason for this consistent lack of accuracy may be that although the algorithms and molecular force fields differ significantly in their approach, almost every method performs some statistical analysis on experimental data, and every method that does so acquires its dataset from the same thermodynamic stability database, ProTherm.

Amassing its data from the scientific literature, the ProTherm database is a valuable repository of experimentally determined stability data (18). At the time this chapter was written the database website boasted 24,875 entries from 716 unique proteins, retrieved from 1,846 scientific articles. However, in order to serve as training data for stability prediction, the number of data points are commonly culled to a smaller collection totaling in the low thousands due to low-quality data or the lack of wild-type crystal structures. These datasets, although useful, suffer from three major limitations.

First, the experimental conditions under which the data is collected varies not only in pH and temperature, which is known to alter the free energy determination (19), but also in the methods of stability determination between different laboratories. Second, the distribution of mutations is overwhelmingly skewed toward small hydrophobic amino acids, with mutations to alanine more than twice as common as those to any other amino acid. This bias is no fault of ProTherm itself, but is instead a byproduct of the value alanine scanning mutagenesis provides to the scientific community. Finally, because it is not common practice to report mutations that completely impede protein folding, stability prediction efforts are hampered by the complete absence of this class of potentially valuable data. In order to overcome the deficiencies of the current datasets, we propose that a database containing stability data collected under a unified automated protocol would greatly benefit the prediction community.

A concerted effort to acquire protein stability information could improve efforts in prediction training by maintaining consistent experimental conditions, keeping a uniform mutational distribution, and providing much-needed data on non-folded sequences. Here we report the development of an automated platform and database for the site-directed mutagenesis and stability analysis of protein domains. Drawing both inspiration and methodology from structural genomics (20), the described procedures utilize liquid-handling robotics to efficiently and rapidly construct, validate, and assay very large numbers of protein mutants. To demonstrate the capability of our platform, every possible single mutant of a protein domain was constructed and analyzed.

Results and discussion

Experimental system

The protein chosen for this study was the β1 domain of Streptococcal protein G (Gβ1) primarily because it has already been well studied by ours and other laboratories in the protein engineering field (21–25). Some reasons for these levels of interest include Gβ1's small size, high amount of secondary structure, and the fact that the wild-type sequence is very well behaved. The last point is especially important when adapting protein purification and analysis protocols for automation, where it may be difficult to reveal and understand strange results. Although the literature does contain examples of bizarre behavior in Gβ1 (26–28), we feel it is an advantage knowing these details ahead of time before developing methodology and conducting a project of this size.

The wild-type sequence of Gβ1 is 56 amino acid residues long, so a complete site-directed mutagenesis project would involve constructing 1064 single mutants. However, the tryptophan residue at position 43 (W43) was left untouched due to that residue's critical importance for measuring intrinsic fluorescence in the stability assay. Similarly, no cysteine or tryptophan residues were inserted as point mutants to avoid potential oligomerization (disulfide bridges) and analysis (multiple tryptophan residues could mask W43's ability to report folded-ness in the stability assay) issues. After these considerations, 935 point mutants were constructed and analyzed.

Automation scheme

The experimental pipeline (Figure. 3-1), starts from mutagenic oligonucleotides and generates high-quality protein stability data from sequence-confirmed site-directed mutagenesis (SDM) products. With equal estimated costs, we employed explicit site-directed oligos over degenerate site-saturation oligos, as it would be much simpler for the former method to recover single mutants not found in the initial round of sequencing. A protocol featuring mutation confirmation by restriction analysis (29) was not considered due to the higher fidelity of sequencing and the potential difficulty of incorporating identical restriction sites at every position. All liquid-handling steps are performed by a customized robotics platform (described in the Materials and methods), which ensures that each SDM reaction is individually addressable at any time as it moves between 24, 48, 96, and 384 well microplate formats. Each step in the pipeline was developed independently and then later strung together as modular parts for production experiments.

Variant construction

The initial step of the automation scheme begins with site-directed mutagenesis, a mature technology that has seen widespread use because of its tremendous utility in protein science and the availability of easy-to-use commercial kits (Stratagene). However, because kits are cost-prohibitive in large volumes, an in-house method was developed from the existing literature. Most reports improve upon the classical Stratagene Quik-Change method by avoiding primer-dimers and vary in the number and specific design of mutagenic oligonucleotides (30–34). An ideal automated SDM method should be cost-effective, require a minimal amount of simple enzymatic steps, and robust enough to avoid manual intervention. The megaprimer-based method described by Tseng et al. best satisfies these criteria as it requires only one unique oligonucleotide per mutagenesis reaction, is completely PCR-based, and was reported to produce more colonies at similar mutagenesis efficiencies when compared against the Quick-Change method (34). Briefly, the mega-primer method combines a forward mutagenic primer with a static reverse flanking primer in an initial PCR reaction to generate large megaprimers that then anneal to the template plasmid in a second PCR reaction to generate the full-length mutagenized nicked circular plasmid. The parental template is then degraded by Dpn1 digestion and the reaction is transformed directly into bacterial cells. All liquid-handling manipulations during variant construction take place on the robot in 96 well PCR plates.

Further optimizations were made to improve the applicability of the megaprimer method for automation. Although the megaprimer protocol already halves the expenditure on oligonucleotides because each reaction requires only a single unique primer, we employed shorter mutagenic oligos (~ 25 bases) than previously described because the cost-savings adds up in a large automation project. The reaction was sped up upon switching from Pfu Turbo to Hot-start Phusion DNA polymerase. This also made the setup more automation-friendly as the Affibody-based Hot-start feature prevents non-specific amplification and primer/template degradation (35, 36). Small-scale experiments showed that the primer melting temperature (T_m) correlated better with a basic T_m calculator (T_m = (64.9+41×(number of gc bp)−16.4)/(number of total bp)) (37, 38), improving reliability over the mismatch method used in the original paper. The 96 well agarose gel in Figure 3-2 shows the performance of the final optimized two-step SDM method.

Less viscous percent solutions of Dpn1 were used to perform template digestion. Although bacterial transformation was very simple to automate through the use of an integrated PCR machine, the cell recovery period had to be done off-line, as no automated solution could match the performance of high-speed shaking at 37°C. Plating the cultures after transformation proved serendipitously simple to automate, as the eight-channel liquid-handling arm (LiHa) on the robot can separate its tips into a range of distances, allowing for elegant column-to-column transfers between a PCR plate and a 48 well LB agar Qtray (Genetix), using a matrix of liquid drops spotted onto each well to aid in spreading. Combined with beads previously dispensed by hand onto the LB agar, 96 well plates of bacterial transformations are plated onto two 48 well Qtrays in less than 10 minutes. After traditional overnight incubation, the Qtrays are are picked by a dedicated Qbot colony picker (Genetix). As described in the methods, eight colonies for each segment of the 48 well Qtray are picked into 384 well LB glycerol plates, creating a one-to-one correspondence between Qtrays and high-density glycerol stock plates. The liquid-handling robot then re-arrays two cultures per reaction in 96 well plates for high-throughput commercial miniprep and sequencing (Beckman Genomics).

The throughput of an automated system is stunning when compared against what can be done manually. The speedup in variant construction leading up to sequence confirmation is achieved primarily by the robot's ability to parallelize work on large numbers of samples without making mistakes common to human laboratory workers. Figure 3-3 shows that in the same amount of time (5 days), a single robot user can perform roughly up to two orders of magnitude more mutagenesis reactions than someone at the bench. Also, because of the low time requirement each day and the fact that the chronological spacing of the procedures are a requirement of the bacterial cells and not of the robot itself, multiple runs through the experimental pipeline are possible by staggering the operations one day apart. After an initial development run with a small selection of Gβ1 mutants, four runs of variant construction (a total of 768 mutants) were performed in 7 days, demonstrating the power of automation.

Sequence confirmation

Analysis of the sequencing results can give insight into the mutagenesis efficiency of the method. In an initial run of variant construction, four colonies per reaction were sent for sequencing, successfully recovering 45 of the 49 constructs sought. However, 41 of the 49 constructs would have been found had we only sent two colonies for sequencing. The savings afforded by halving the number of sequencing requests more than made up for the miniscule drop in recovery rate. This modification was adopted throughout the rest of the project, and 96 well plates sent for sequencing had recovery rates between 80–90%. In addition, the percentage of colonies coming back as wild type dropped as more experience was gained in performing the SDM methodology. The percentage of colonies with non-mutated sequences fluctuated between 6 and 30% with an average of 17%, increasing when adding plasmid template to the mutagenesis reaction and decreasing when using more concentrated Dpn1 enzyme during template digestion. The parameters reported in the methods represent a qualitative balance between the cost of the Dpn1 enzyme, success of the SDM reaction, and mutant recovery rate.

After all of the 935 constructs were confirmed by sequencing, the cultures containing successful mutants were re-arrayed sequentially and by mutant amino acid type. This allows the entire library of mutants to be stored on just ten 96 well

LB/glycerol plates, where they are easily accessible for both humans and robots to replicate from and perform further experiments.

Stability analysis

The final block of the automation scheme probes the thermodynamic stability of the point mutant library after over-expression in auto-induction media and Ni-NTA purification. As described in the methods, the robot performed all liquid-handling operations except for the wash and elution steps of the purification. This was necessary, as automated vacuum methods could not replicate the reliability and speed provided by manually loaded centrifugation during filter-plate purification. Future robot layouts needing to perform solid-phase extraction would benefit more from an integrated centrifuge than a vacuum station. Lowering the imidazole concentration in the elution step and diluting the purified protein fivefold obviated buffer exchange, which would otherwise be necessary to remove the harsh conditions found in protein elutions after hexahistidine-based purification.

The automated plate-based stability assay developed here is a considerable improvement over the first iteration of this method (25), as it is faster to setup and process, while simultaneously maintaining precision and doubling the number of measurements from twelve to twenty-four. Where the old method would have gathered 1152 data points (12 data points over 96 proteins) in 5 hours, the new method gathers 2304 data points (24 data points over 96 proteins) in 4 hours, a 2.5x increase in efficiency.

By measuring the intrinsic fluorescence in response to a chemical gradient that unfolds the protein, the stability assay probes the environment around the single buried tryptophan at position 43 (W43) in the Gβ1 domain. This information on the tertiary structure of the protein not only gives thermodynamic details of stability (free energy of unfolding, $\Delta G(H_2O)$; denaturation concentration at 50% unfolded, C_m; slope of the denaturation curve, *m*-value), but can also shed light on the foldedness and oligomeric state of the purified protein. Figure 3-4 shows fluorescence data for three examples from the single mutant library of well-folded, unfolded, and likely oligomeric proteins. Since stability data from oligomeric and unfolded proteins are not amenable to curve fitting analysis, every 24 point measurement was annotated with a comment describing protein quality to simplify data cleaning. Of the 935 mutants analyzed, 100 proteins had unfolding transitions consistent with very unstable, completely unfolded, or oligomeric characteristics. These records, although missing proper thermodynamic parameters, still provide valuable information on mutations that substantially disrupt a protein's native fold. This type of negative data is typically unreported in the literature and therefore missing from datasets extracted from ProTherm.

In order to get a measure of data quality, duplicate records with measurable thermodynamic parameters were correlated against each other using ddG, the difference between the free energies of the wild-type and mutant proteins. Two different measures of ddG were examined, one taking the difference between the fitted $dG(H_2O)$ values given by the linear extrapolation method (LEM) for stability analysis (39) (ddG-true, Figure 3-5A), and the other taking the difference between C_m values and multiplying by the average of the wild-type and mutant m-values (ddG-mAVG, Figure 3-5B). The latter calculation is much more precise (r = 0.78 versus r = 0.99) as advocated in the literature (40, 41), and removes any uncertainties concerning the non-linear dependence of free energy on denaturant concentration, a potential issue when using the linear extrapolation method (41). However, this simplification is only valid when single mutations are not expected to greatly affect the m-value or the stability of the mutant protein. The strong linear relationship between ddG-true and ddG-mAVG as seen in Figure 3-5C (r = 0.89), supports the application of the ddG-mAVG assumption for this dataset.

Data tracking and analytics

An important factor that supports and facilitates the proper operation of the automated scheme is the usage of relational database software. Early in the development of the project, a need arose for a data management solution to tackle the volume of oligonucleotide, mutation, and stability information already being generated as well as that on the horizon. The database marketplace has many chemoinformatic solutions for the pharmaceutical industry, but painfully few options exist for handling protein-centric mutational studies. Taking cues from an inventive solution (42) to the problem, we developed in-house databases in Access 2010 (Microsoft) to house records detailing the construction and experimental stability of the protein mutant library.

The focus of the experimental construction database (ecDB) is to maintain records of every mutation attempt and to track the mutants that have been recovered versus those that haven't been confirmed by sequencing. For each construction attempt, records are kept detailing protocol parameters for the SDM, template digestion, and transformation methods. This detailed metadata was helpful while optimizing mutagenic oligo designs and the SDM protocol. To facilitate reconstruction attempts, SQL queries identified those mutants still missing after sequencing and provided robot-friendly location information of the required mutagenic oligo. After sequence confirmation, another query identified the first instance of each mutation from the sequencing plates and reported its robot-friendly location for the re-array procedure.

With the entire library located in a manageable number of 96 well plates, the experimental stability database (estabDB) was designed to keep records on protein purification attempts as well as the resultant raw and fitted stability data. Like ecDB, it stores detailed metadata for the expression, purification, and stability assay protocols. Although the raw stability data was fitted and analyzed outside the database and later imported, future database iterations using open source MySQL will enable on-the-fly analysis and recording of fitted data. Data cleaning routines made use of a data comment system, made necessary by the concerns conveyed in Figure 3-4, to quickly filter denaturation curves containing outliers or depicting potentially oligomeric or unfolded proteins. To enable future in-depth analysis of the stability data, queried results not only contain standard thermodynamic stability details but also separate the mutation label (e.g., V29A) into fields for the wild-type amino acid, position, and mutant amino acid. In this way ancillary tables containing information on the individual amino acids and protein domain positions can be related to the stability results, permitting the investigation and rationalization of advanced queries such as “How many proteins were stabilized over wild-type and featured steric volume loss by mutation?” or “Which surface-exposed positions on the protein were most accepting of mutations?” An in-depth analysis to queries of this nature is the focus of the next chapter in this thesis.

Automated gene assembly

The modularity of the developed automation scheme allows the pipeline to be easily adapted to other protein engineering methods, such as gene assembly. Automated gene assembly overcomes the multitude of ways to design self-assembling oligonucleotides into a full-length gene constructs (43–49). Recently the technology has garnered a great deal of attention for its utility in synthetic biology (46, 48) and has already been previously adapted for robotic automation (49). If appropriate methods to assemble and insert a gene of interest into a plasmid were available, the current site-directed mutagenesis pipeline could fill in the rest of the necessary molecular biology. Two methods that may satisfy the demands of automated gene assembly, oligo design by DNAworks (44) and gene cloning by circular polymerase extension cloning (CPEC) (50), are currently undergoing laboratory testing. A promising small-scale experiment has shown that 80% of a 60 member individually assembled gene library was successfully recovered after sequencing 3 colonies per construct (results not shown).

Future data deposition

The methods developed to perform the complete site-directed mutagenesis of Gβ1 will continue to be employed in future domain mutagenesis projects. A growing database of stability and eventually activity information will be that much more valuable for stability prediction and a better understanding of the complexity of protein physics. In choosing the next few proteins to undergo the mutagenesis treatment, a handful of characteristics will be considered. First and foremost, target proteins must be compatible with our plate-based stability assay. This requires a high-resolution crystal structure in order to identify (preferably single) tryptophans buried away from solvent that can act as a fluorescence reporter for foldedness. Second, proteins on the smaller side of the structural continuum are preferred over multi-domain behemoths because of the lower price tag of a mutagenesis effort and the increased likelihood of a two-state cooperative transition during denaturation. The latter criterion is required for proper application of the linear extrapolation method used to estimate thermodynamic parameters such as the free energy of unfolding and the slope of the denaturation curve. Lastly, domains recognized as superfolds (51) under the CATH classification of protein architectures (52) are preferable as it makes the acquired information more applicable to greater proportions of natural proteins. Also, superfolds are also more likely to have thermophilic homologs that could provide interesting perspectives on mutagenesis data from their mesophilic counterparts.

Whereas the production and characterization of the Gβ1 single mutant library took over a year because of the simultaneous development of the automated methodology, future projects should see completion in considerably less time. As depicted in Figure 3-3, staggering the experimental modules allows for a large number of mutants to be constructed and sent for sequencing all at once. Drawing from this experience, a majority of mutants can be attained in the first wave of mutagenesis, but the subsequent production of the remainder of the library can markedly slow the entire procedure. Nevertheless it is not unreasonable to expect that the complete single mutant mutagenesis of an entire domain of 100 residues (~ 1700 point mutants, excluding Cys and Trp) be constructed and analyzed in three months time under our current automation scheme. Future modifications to increase the throughput of the scheme could include: 1) the integration of a 384 well PCR machine for faster mutagenesis/template digestion, and 2) integration of an automated centrifuge for avoiding the manual intervention now necessary during protein purification. Unfortunately, because of the appeal for our stability assay that can produce high-quality thermodynamic data but requires significant amounts of protein, high-throughput fluorescent dye-based thermal scanning (53) or *in vitro* transcription and translation methods are not appropriate (54).

External database comparisons

To ensure the accuracy of the automated method, the stability of a small collection of previously determined point mutants of Gβ1 were correlated against values from our database. The test-set, retrieved from ProTherm, was comprised of mutants from a proline-scanning mutagenesis study and a site saturation mutagenesis study, the former being previously performed in our lab. Remarkably, the combined test-set gave correlation coefficients of $r = 0.84$ and $r = 0.88$ (Figure 3-6) when correlated against our ddG-true and ddG-AVGm data, respectively, despite reporting thermodynamic data from dissimilar experimental methods. This result affirms the validity of the automated site-directed mutagenesis method and the ancillary high-throughput stability assay.

In addition to providing self-consistent and seemingly accurate data, our experimental method has provided a dataset with a unique composition when compared against those previously used for energy function training and stability prediction testing. Although a recent training set used in the development of the stability prediction algorithm PopMusic 2.0 has a fairly even distribution over wild-type identity amino acids

(Figure 3-7A), the mutated amino acid distribution is heavily skewed towards alanine incorporation (Figure 3-7B). One might then presume that any stability ranking potential trained on this data might perform remarkably well on predicting large-to-small mutations, but fail to accurately predict the effects of other types of amino acid substitutions. Published datasets from ProTherm used in the training of other stability prediction algorithms have almost identical distributions (results not shown). In contrast, our dataset is unbiased in its distribution (Figure 3-7D) of mutant amino acids because of the nature of the mutagenesis project. Unfortunately, a similar impartiality is not evident in the wild-type amino acid distribution (Figure 3-7C) as this is dependent on the wild-type amino acid composition of only one system (Gβ1). Future deposits of stability data from other systems will help to ameliorate this issue.

Conclusions

We have developed automated methods to construct, validate, and analyze very large numbers of protein point mutants. Our automated platform sees massive gains in throughput over traditional bench-top methods by employing liquid-handling robotics to boost the number of samples performed during each run and parallelize the number of concurrent runs through the experimental pipeline. Each pass of the mutagenesis routine can expect to recover 80–90% of the desired sequences. Using this platform, 935 variants of Gβ1, comprising almost every single mutation possible, were constructed and assayed for thermodynamic stability. The precision and accuracy of the improved high-throughput stability assay is comparable to existing lower throughput methods, and all relevant data and metadata has been stored in relational databases that proved useful for data tracking and later for data cleaning and analysis. In large part because of the success of the methodology, the employment of this automated platform will not end with this lone mutagenesis project.

Our experimental pipeline, built modularly, can be adapted for use in projects featuring functional enzymatic assays or even repurposed for automated gene assembly. The volume of thermodynamic stability data collected will grow in spurts and jumps as more domains are processed by the total site-directed mutagenesis method. And ultimately this is where the automated system can make a fundamental impact on protein science: by reporting higher quality and more diverse mutational stability data than what is already publicly available, it is expected that substantial progress in stability prediction and understanding of protein physics will follow.

Materials and methods

Liquid handling robotics

A 2 meter Freedom EVO (Tecan) liquid-handling robot was used to automate the great majority of the experimental pipeline. The instrument includes an eight-channel fixed-tip liquid-handling arm, a 96 disposable-tip single-channel liquid-handling arm, and a robotic plate-gripping arm. The robot's deck features a fast-wash module, a refrigerated microplate carrier, a microplate orbital shaker, a SPE vacuum system, an integrated PTC-200 PCR machine (Bio-Rad Laboratories), stacks and hotels for microplates and an integrated Infinite M1000 microplate reader (Tecan).

Variant construction and enrichment conditions

The Gβ1 gene, with an N-terminal hexahistidine tag, was inserted into pET11a under control of an IPTG inducible T7 promoter. Mutagenic oligonucleotides were ordered from Integrated DNA Technologies in a 96 well format (150 uM concentration, 25 nmole scale) and purified by standard desalting. The site-directed mutagenesis reaction was performed in two parts: 1) the diluted mutagenic oligonucleotide was mixed with a mastermix solution composed of Hot-start Phusion DNA polymerase (NEB), GC Phusion buffer, dNTPs, the plasmid template, and the non-mutagenic flanking oligonucleotide, and 2) ¼ of the first step product was mixed with a similar mastermix solution that omits the flanking oligonucleotide. The PCR cycling conditions for the two parts were: 1) a 30 sec preincubation at 98°C followed by 15 thermocycling steps (98°C, 8 sec; 62°C, 15 sec; 72°C, 20 sec), and 2) a 30 sec preincubation at 98°C followed by 25 thermocycling steps (98°C, 8 sec; 72°C, 3 min) followed by a final extension step at 72°C for 5 min.

Reactions were often diagnosed by E-Gel 96 (Invitrogen) electrophoresis systems, with loading performed by the liquid-handling robot. A bright band corresponding to the size of the template plasmid indicated a successful second-step reaction. Samples could be troubleshot by observing the desired first-step product, the amplified megaprimer. If the reactions performed well they would be subjected to an 8%-by-volume Dpn1 (NEB) digestion reaction (37°C, 1 hour) in order to remove the parental template plasmid.

Bacterial manipulation and sequence verification

Dpn1 digested products were mixed with homemade chemically competent BL21 Gold DE3 cells (55) in a 20 uL total reaction volume, and incubated at 4°C for 10 min. After heatshock (42°C, 45 sec) on the PCR machine, the bacterial transformations were recovered by adding 100 uL of SOC media, and shaken off robot at 1200 rpm for 1 hour at 37°C on a microplate shaker (Heidolph).

The transformations were plated by the liquid-handling robot onto 48 well LB agar Qtrays (Genetix) and spread by sterile beads (55). The Qtrays were incubated for 14 hours at 37°C. For each mutagenesis reaction eight colonies were picked by a colony-picking robot (Qbot, Genetix) into 384 well plates (Genetix) filled with LB/10% glycerol. The 384 well receiving plates were incubated overnight at 37°C, after which 2 of the 8 cultures per mutagenesis reaction were used to inoculate 96 well microplates containing LB/10% glycerol. These 96 well glycerol stock plates were grown overnight at 37°C, replicated, and sent to Beckman Genomics for sequencing.

After analyzing the sequencing data, missing library members could be recovered either by sending more picked colonies from the 384 well receiving plate, or by redoing the entire mutagenesis reaction with different PCR conditions. The 96 well E-Gel results were critical in informing the subsequent optimization that should take place. Once all of the mutants were constructed, work-lists were generated for the liquid-handling robot to cherry-pick from the replicated 96 well glycerol stock plates and inoculate into column-arrayed 96 well master stock plates containing LB/10% glycerol.

Protein expression and purification

Small volumes from replicated master stock plates were used to inoculate 5 mL of Instant TB auto-induction media (Novagen) in 24 well round-bottom plates (Whatman). The 24 well plates were incubated overnight, shaking at 250 rpm, at 37°C. The expression cultures were then pelleted, lysed with a sodium phosphate lysis buffer solution (pH 8) containing CelLytic B (Sigma Aldrich), lysozyme, and HC Benzonase (Sigma Aldrich). Lysates were then added directly to 96 well His-Select Ni-NTA resin filter plates (Sigma Aldrich) and processed off-robot by centrifugation. His-tagged protein was washed and eluted in sodium phosphate buffer (pH 8) containing 0 mM and 100 mM imidazole, respectively. Protein samples were diluted fivefold into sodium phosphate buffer (pH 6.5), thereby diluting the amount of imidazole in each sample.

Plate-based stability assay

Large volumes of a 24-point gradient of GdmCl in sodium phosphate buffer (pH 6.5) were constructed using graduated cylinders and dispensed into 96 well deep-well plates by a multi-channel pipettor. These reagent reservoirs, along with the liquid-handling robot, greatly simplified and sped up the stability assay previously described (25). Each stability assay was comprised of twenty-four individual solutions containing 1 part purified protein to 4 parts GdmCl/buffer solution, and measured by the integrated plate reader for tryptophan fluorescence (Ex: 295 nm, Em: 341 nm). The assay employed 384 well UVstar plates (Greiner) that allowed 16 different protein mutants to be measured per plate, thus requiring 6 of these plates per 96 well master stock plate. Measurements were made in duplicate. Data was analyzed as described previously (25).

References

1. Knowles JR (1987) Tinkering with enzymes: what are we learning? *Science* 236(4806):1252–1258.

2. Alber T (1989) Mutational effects on protein stability. *Annual review of biochemistry* 58:765–798.

3. Pace CN (1990) Measuring and increasing protein stability. *Trends in Biotechnology* 8(4):93–98.

4. Fersht AR & Serrano L (1993) Principles of Protein Stability Derived from Protein Engineering Experiments. *Current Opinion in Structural Biology* 3(1):75–83.

5. Matthews BW (1993) Structural and genetic analysis of protein stability. *Annual review of biochemistry* 62:139–160.

6. Vieille C & Zeikus GJ (2001) Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiology and molecular biology reviews : MMBR* 65(1):1–43.

7. Sunyaev S, Lathe W, 3rd & Bork P (2001) Integration of genome data and protein structures: prediction of protein folds, protein interactions and "molecular phenotypes" of single nucleotide polymorphisms. *Current Opinion in Structural Biology* 11(1):125–130.

8. Baltzer L & Nilsson J (2001) Emerging principles of de novo catalyst design. *Current Opinion in Biotechnology* 12(4):355–360.

9. Bolon DN, Voigt CA & Mayo SL (2002) De novo design of biocatalysts. *Current Opinion in Chemical Biology* 6(2):125–129.

10. Guerois R, Nielsen JE & Serrano L (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of Molecular Biology* 320(2):369–387.

11. Masso M & Vaisman, II (2008) Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. *Bioinformatics* 24(18):2002–2009.

12. Benedix A, Becker CM, de Groot BL, Caflisch A & Bockmann RA (2009) Predicting free energy changes using structural ensembles. *Nature methods* 6(1):3–4.

13. Dehouck Y, et al. (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics* 25(19):2537–2543.

14. Ozen A, Gonen M, Alpaydan E & Haliloglu T (2009) Machine learning integration for predicting the effect of single amino acid substitutions on protein stability. *BMC structural biology* 9:66.

15. Kellogg EH, Leaver-Fay A & Baker D (2011) Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins* 79(3):830–838.

16. Potapov V, Cohen M & Schreiber G (2009) Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein engineering, design & selection* 22(9):553–560.

17. Khan S & Vihinen M (2010) Performance of protein stability predictors. *Human mutation* 31(6):675–684.

18. Kumar MD, et al. (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Research* 34(Database issue):D204–206.

19. Pace CN, Laurents DV & Thomson JA (1990) pH dependence of the urea and guanidine hydrochloride denaturation of ribonuclease A and ribonuclease T1. *Biochemistry* 29(10):2564–2572.

20. Chandonia JM & Brenner SE (2006) The impact of structural genomics: expectations and outcomes. *Science* 311(5759):347–351.

21. Minor DL & Kim PS (1994) Context is a major determinant of beta-sheet propensity. *Nature* 371(6494):264–267.

22. Gronenborn AM, Frank MK & Clore GM (1996) Core mutants of the immunoglobulin binding domain of streptococcal protein G: stability and structural integrity. *FEBS letters* 398(2–3):312–316.

23. Dahiyat BI & Mayo SL (1997) De novo protein design: fully automated sequence selection. *Science* 278(5335):82–87.

24. Malakauskas SM & Mayo SL (1998) Design, structure and stability of a hyperthermophilic protein variant. *Nature structural biology* 5(6):470–475.

25. Allen BD, Nisthal A & Mayo SL (2010) Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. *Proceedings of the National Academy of Sciences of the United States of America* 107(46):19838–19843.

26. Kirsten Frank M, Dyda F, Dobrodumov A & Gronenborn AM (2002) Core mutations switch monomeric protein GB1 into an intertwined tetramer. *Nature structural biology* 9(11):877–885.

27. Byeon IJ, Louis JM & Gronenborn AM (2003) A protein contortionist: core mutations of GB1 that induce dimerization and domain swapping. *Journal of Molecular Biology* 333(1):141–152.

28. Jee J, Byeon IJ, Louis JM & Gronenborn AM (2008) The point mutation A34F causes dimerization of GB1. *Proteins* 71(3):1420–1431.

29. Carapito R, Gallet B, Zapun A & Vernet T (2006) Automated high-throughput process for site-directed mutagenesis, production, purification, and kinetic characterization of enzymes. *Analytical Biochemistry* 355(1):110–116.

30. Shenoy AR & Visweswariah SS (2003) Site-directed mutagenesis using a single mutagenic oligonucleotide and DpnI digestion of template DNA. *Analytical Biochemistry* 319(2):335–336.

31. Chiu J, March PE, Lee R & Tillett D (2004) Site-directed, Ligase-Independent Mutagenesis (SLIM): a single-tube methodology approaching 100% efficiency in 4 h. *Nucleic Acids Research* 32(21):e174.

32. Zheng L, Baumann U & Reymond JL (2004) An efficient one-step site-directed and site-saturation mutagenesis protocol. *Nucleic Acids Research* 32(14):e115.

33. Liu H & Naismith JH (2008) An efficient one-step site-directed deletion, insertion, single and multiple-site plasmid mutagenesis protocol. *BMC biotechnology* 8:91.

34. Tseng WC, Lin JW, Wei TY & Fang TY (2008) A novel megaprimed and ligase-free, PCR-based, site-directed mutagenesis method. *Analytical Biochemistry* 375(2):376–378.

35. Nord K, et al. (1997) Binding proteins selected from combinatorial libraries of an alpha-helical bacterial receptor domain. *Nature biotechnology* 15(8):772–777.

36. Wikman M, et al. (2004) Selection and characterization of HER2/neu-binding affibody ligands. *Protein engineering, design & selection* 17(5):455–462.

37. Wallace RB, et al. (1979) Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. *Nucleic Acids Research* 6(11):3543–3557.

38. Sambrook J & Russell DW (2001) *Molecular cloning : a laboratory manual* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.) 3rd Ed.

39. Santoro MM & Bolen DW (1988) Unfolding free energy changes determined by the linear extrapolation method. 1. Unfolding of phenylmethanesulfonyl alpha-chymotrypsin using different denaturants. *Biochemistry* 27(21):8063–8068.

40. Pace CN (1986) Determination and analysis of urea and guanidine hydrochloride denaturation curves. *Methods in enzymology* 131:266–280.

41. Myers JK, Pace CN & Scholtz JM (1995) Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Science* 4(10):2138–2148.

42. Vielmetter J, Tishler J, Ary ML, Cheung P & Bishop R (2005) Data management solutions for protein therapeutic research and development. *Drug discovery today* 10(15):1065–1071.

43. Stemmer WP, Crameri A, Ha KD, Brennan TM & Heyneker HL (1995) Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene* 164(1):49–53.

44. Hoover DM & Lubkowski J (2002) DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Research* 30(10):e43.

45. Gao X, Yo P, Keith A, Ragan TJ & Harris TK (2003) Thermodynamically balanced inside-out (TBIO) PCR-based gene synthesis: a novel method of primer design for high-fidelity assembly of longer gene sequences. *Nucleic Acids Research* 31(22):e143.

46. Tian J, et al. (2004) Accurate multiplex gene synthesis from programmable DNA microchips. *Nature* 432(7020):1050–1054.

47. Young L & Dong Q (2004) Two-step total gene synthesis method. *Nucleic Acids Research* 32(7):e59.

48. Villalobos A, Ness JE, Gustafsson C, Minshull J & Govindarajan S (2006) Gene Designer: a synthetic biology tool for constructing artificial DNA segments. *BMC Bioinformatics* 7:285.

49. Cox JC, Lape J, Sayed MA & Hellinga HW (2007) Protein fabrication automation. *Protein Science* 16(3):379–390.

50. Quan J & Tian J (2009) Circular polymerase extension cloning of complex gene libraries and pathways. *PLoS ONE* 4(7):e6441.

51. Orengo CA, Jones DT & Thornton JM (1994) Protein superfamilies and domain superfolds. *Nature* 372(6507):631–634.

52. Cuff AL, et al. (2011) Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Research* 39(Database issue):D420–426.

53. Lavinder JJ, Hari SB, Sullivan BJ & Magliery TJ (2009) High-throughput thermal scanning: a general, rapid dye-binding thermal shift screen for protein engineering. *Journal of the American Chemical Society* 131(11):3794–3795.

54. Katzen F, Chang G & Kudlicki W (2005) The past, present and future of cell-free protein synthesis. *Trends in Biotechnology* 23(3):150–156.

55. Klock HE & Lesley SA (2009) The Polymerase Incomplete Primer Extension (PIPE) method applied to high-throughput cloning and site-directed mutagenesis. *Methods in Molecular Biology* 498:91–103.

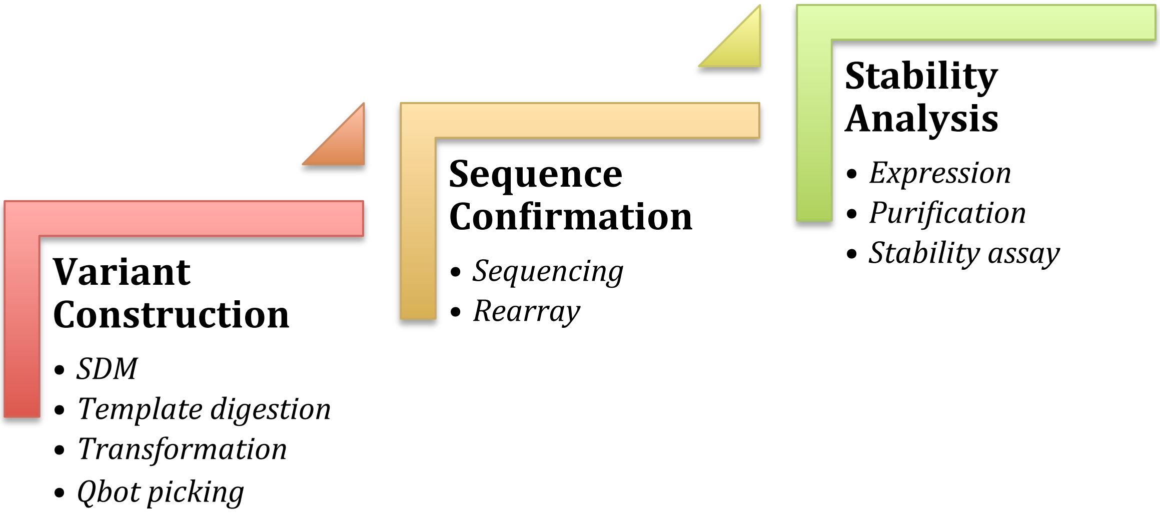

Figure 3-1: The automated site-directed mutagenesis pipeline. The methodology is composed of nine modular protocols that can be grouped into three blocks. The pipeline leads off with variant construction, a block of procedures that takes one mutagenic oligonucleotide per construct and ends with eight colonies per mutagenesis reaction. The next block, sequence confirmation, rearrays all of the constructs only after being validated by sequencing. The last block of protocols, stability analysis, takes from rearrayed, confirmed plates of mutants to generate high-quality data.

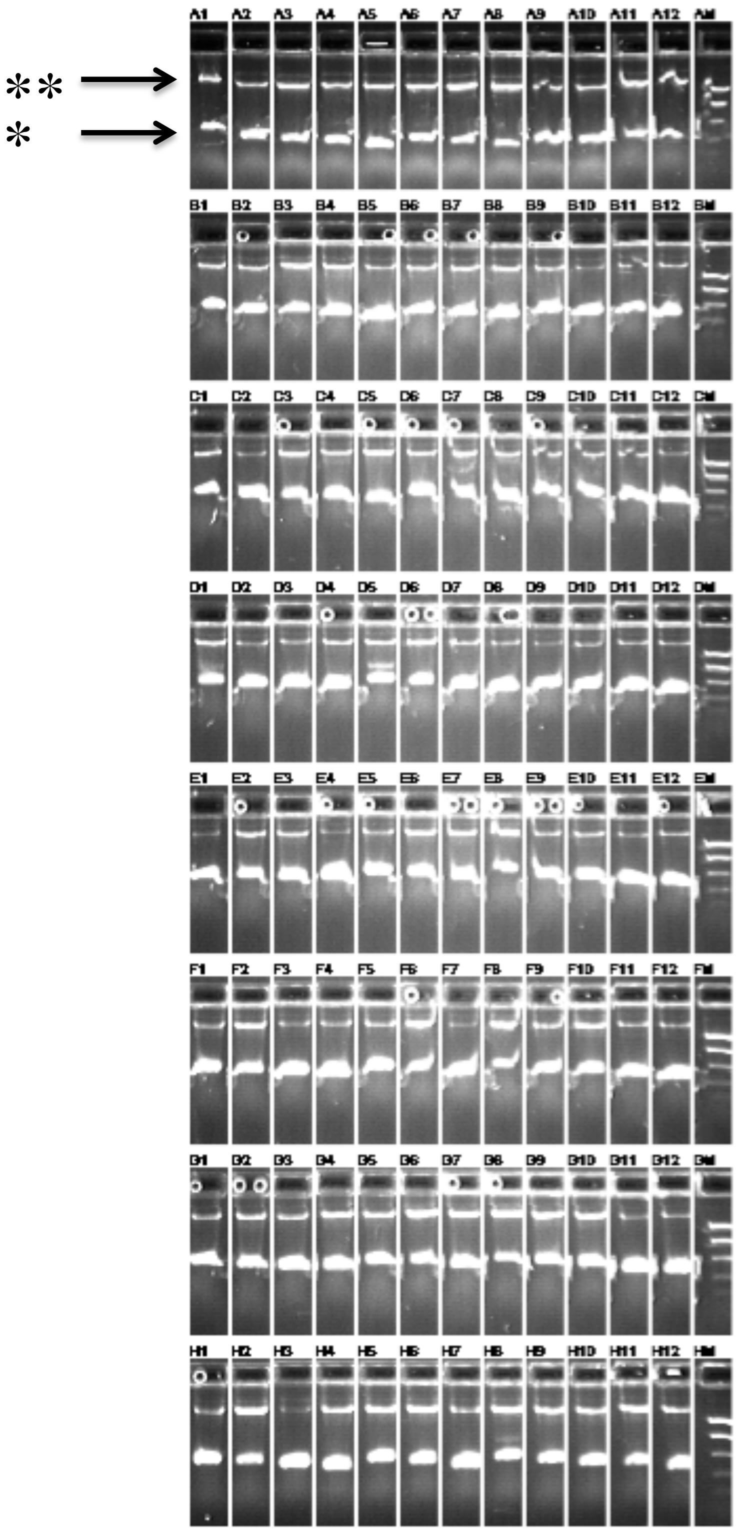

Figure 3-2: Visualization of a 96 well plate of SDM products. Agarose gel electrophoresis of DNA, by E-Gel 96, shows first- (∗) and second-step (∗∗) products from the megaprimer method for site-directed mutagenesis.

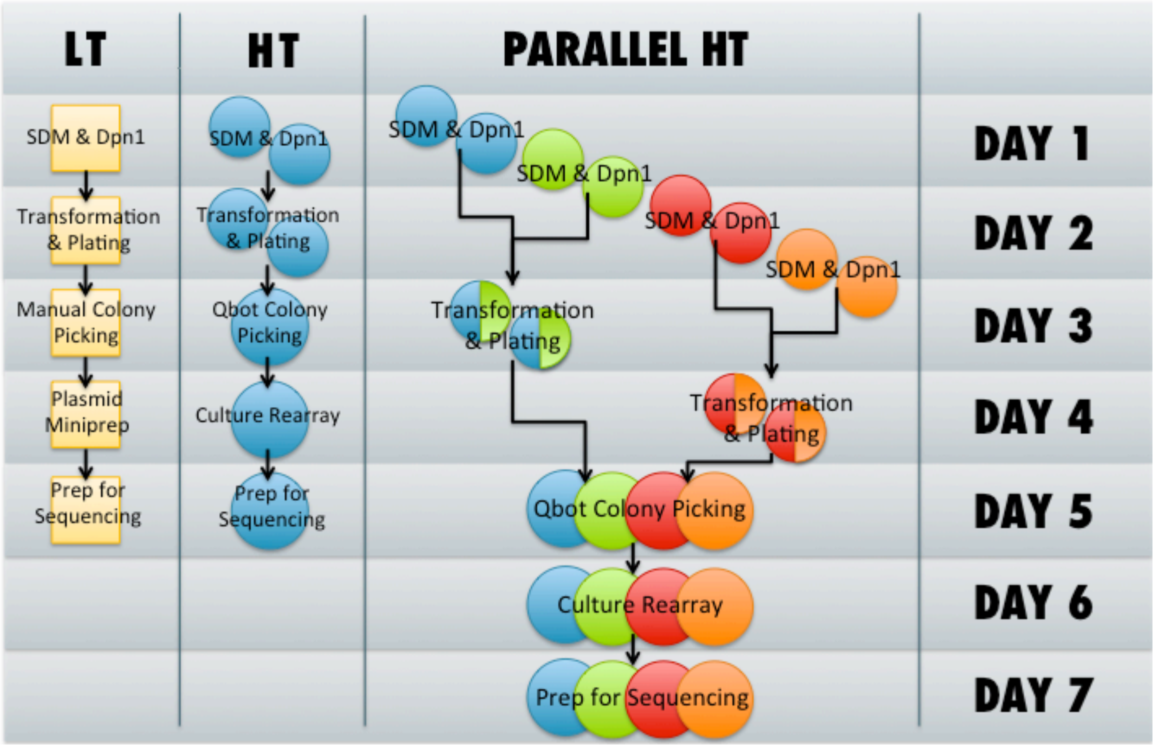

Figure 3-3: Variant construction timeline. Standard bench-top low-throughput (LT) methods are compared against the automated high-throughput (HT) methodology. Although both methods take five days before sending samples for commercial sequencing, the HT method processes 20-fold more reactions. For larger projects, the HT pipeline can be parallelized, processing 4-fold more reactions in just seven days.

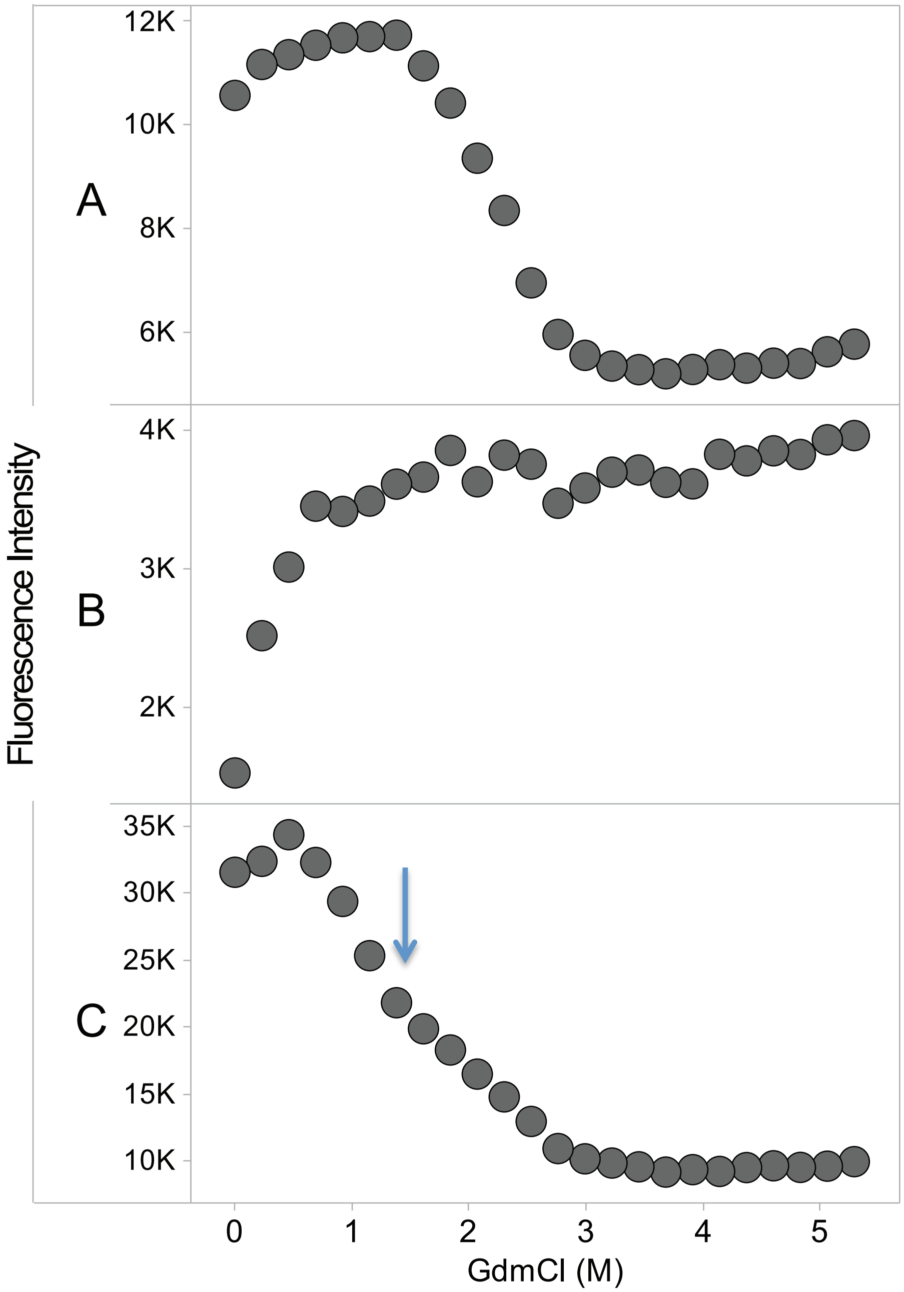

Figure 3-4: Potential protein unfolding curves. Tryptophan fluorescence data, plotted against a guanidinium chloride (GdmCl) gradient, for three examples of data from the high-throughput stability assay. A quality protein, with substantial pre- and post-transition baselines flanking a smooth transition is pictured in plot **A**. Unfolded or non-expressed protein, with very low fluorescence intensity increasing in value, is pictured in plot **B**. Plots of non two-state or oligomeric proteins can exhibit various characteristics, but plot **C** shows one example with an inflection (arrow) in the transition region, violating the two-state assumption.

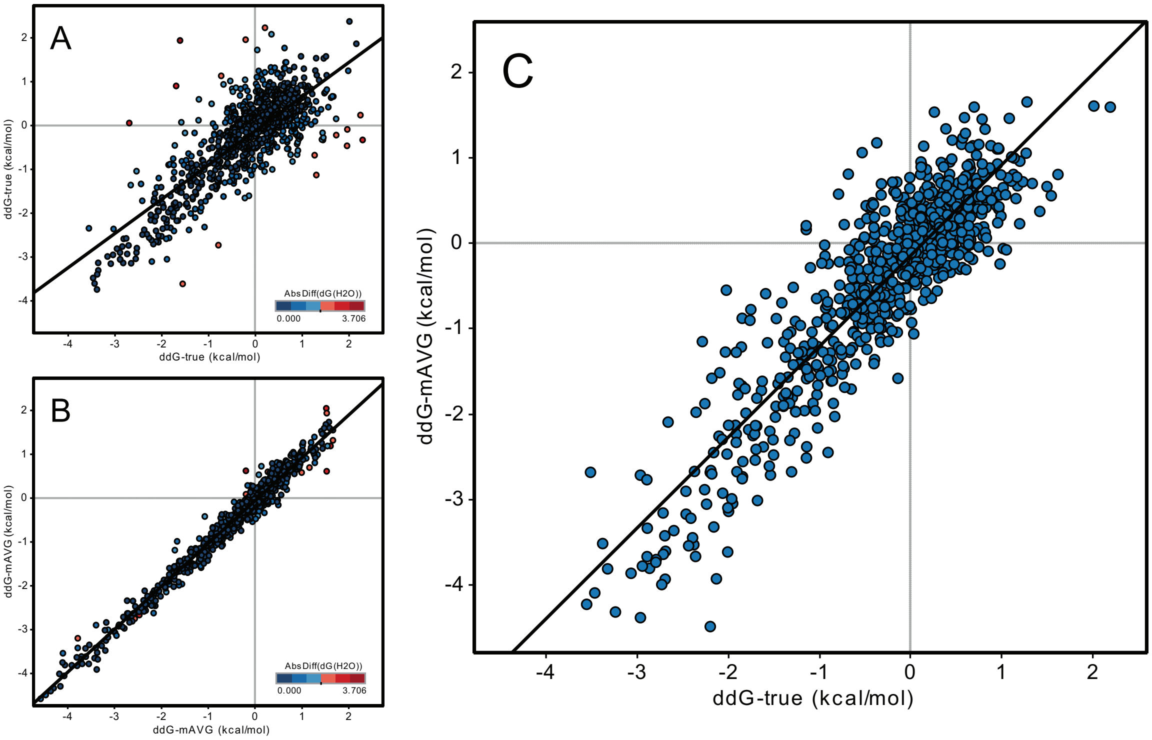

Figure 3-5: Precision among experimental measures of protein stability. Duplicate measurements of ddG-true **(A)** and ddG-mAVG **(B)** from the complete mutagenesis of Gβ1 were correlated against each other. A strong linear relationship exists between ddG-true and ddG-mAVG **(C)**. Linear trend lines are in solid black. Each data point in A and B is colored by the absolute difference in dG(H_2O) measurements.

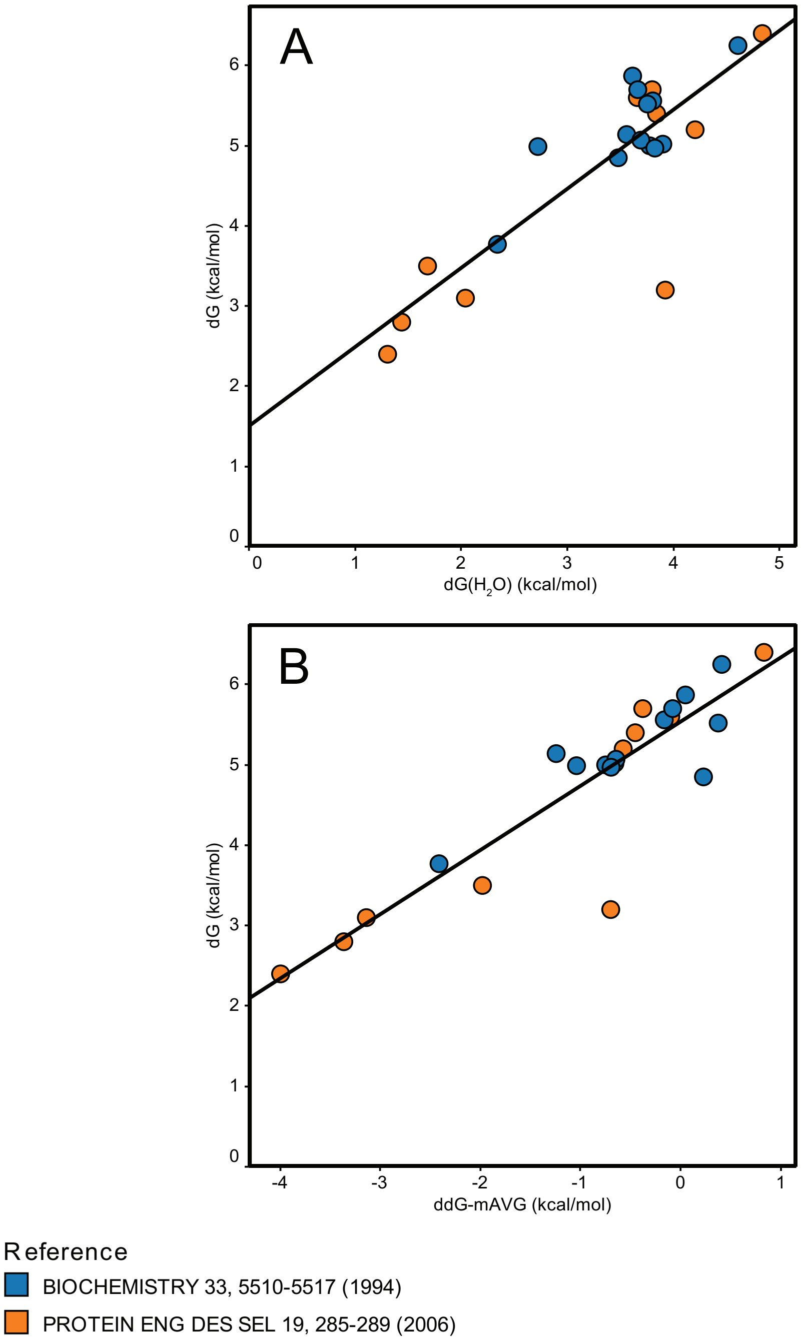

Figure 3-6: Dataset accuracy from literature comparisons. Stability data on a subset of single mutants of Gβ1 were collected from Protherm and correlated against **(A)** dG(H_2O) and **(B)** ddG-mAVG values from our automated site-directed mutagenesis library. The previously reported data differ in unfolding method (thermal against chemical denaturation) and experimental conditions (pH 5.2–5.5 against pH 6.5) from our dataset. Linear trend lines are in solid black.

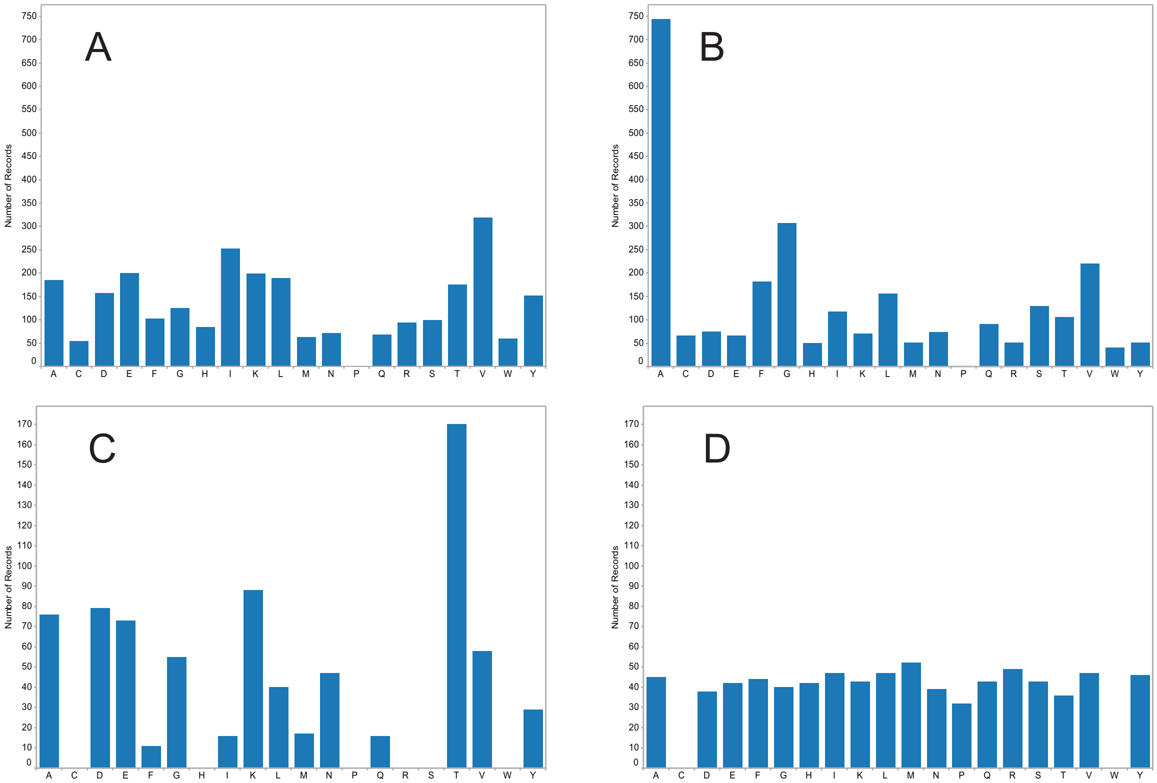

Figure 3-7: Point mutant amino acid distributions. The wild-type amino acid **(A)** and mutated amino acid **(B)** distributions are shown for a training set of 2,649 data points (collected from Protherm) used in developing PopMusic 2.0, a protein stability prediction algorithm. The wild-type **(C)** and mutated **(D)** amino acid distributions for the current version of our mutagenesis database (775 data points)