

Accelerating the Interplay Between Theory and Experiment in Protein Design

Thesis by

Alex Nisthal

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2012

(Defended January 31, 2012)

© 2012

Alex Nisthal

All Rights Reserved

Acknowledgements

I'd first like to thank Steve Mayo for all of the support and advice he has given me over the years, including the opportunity to work on large expensive equipment with lots of bells and whistles.

I would also like to thank the rest of my thesis committee (Frances, Doug, and Dave) and Jost Vielmetter at the Protein Expression Center for providing direction when I needed it the most.

As the bridge between two distinct eras of the Mayo Lab, I've witnessed a complete overhaul of students and postdocs. From my time as a learner, I would especially like to thank Oscar Alvizo, Christina Vizcarra, Marie Ary, Rhonda Digiusto, Corey Wilson, Roberto Chica, Heidi Privett, Ben Allen, and Jennifer Keeffe for their time and geniality in teaching me the ways of the lab. Now, as the master, I'd like to thank Matt Moore, Kurt Mou, Alexandria Berry, Toni Lee, Tim Wannier, Bernardo Sosa Padilla Araujo, Mohsen Chitsaz, Gene Kym, Ernest Lee, Grace Lee, Samy Hamdouche, Emzo de los Santos, and Seth Lieblich for making the lab a fun, collaborative, and politically incorrect place to work.

Lastly, I would like to thank my friends for keeping me sane and my family for letting me be whatever I wanted to be.

Abstract

Protein engineering techniques such as directed evolution and structure-based design aim to improve the properties of natural proteins. The next step, the *de novo* insertion of function into previously inert protein scaffolds, is the lofty promise of computational protein design. In order to achieve this goal reliably and efficiently, computational methods can be iteratively improved by cycling between theory and experiment.

Efforts to both accelerate the rate and broaden the information exchanged within protein design cycles form the core of this thesis. Improvements in the throughput of experimental stability determination allowed the thorough assessment of new multi-state and library design tools. Intending to alleviate the fixed backbone, single native state design approximation, the study found constrained molecular dynamics ensembles useful for core repacking applications. The subsequent development of automated liquid handling protocols for common molecular biology techniques brings design experiments to new levels of sample throughput. This technology facilitated the creation of a stability database encompassing every single mutant in a small protein domain. Although constructed to facilitate future computational training efforts, we answer a multitude of questions pertaining to mutational outcomes, distributions, positional sensitivity, tolerance, and additivity in the context of a protein domain.

By expanding the constraints of experimental molecular biology, this work opens up new possibilities in the efforts to train and assay new computational methodologies for protein engineering applications.

TABLE OF CONTENTS

Acknowledgements		iii
Abstract		iv
Table of Contents		v
Figures and Tables		vi
Chapters		
Chapter 1	<i>Introduction</i>	1
Chapter 2	<i>Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles</i>	10
Chapter 3	<i>Automated techniques for the complete site-directed mutagenesis and stability analysis of protein domains</i>	47
Chapter 4	<i>Stability analysis of the complete single mutant library of a protein domain</i>	80
Appendix		
Appendix	<i>High-throughput and automation methods</i>	126

Figures and Tables

Figure 2-1. The core residues of G β 1 designed in this study	35
Figure 2-2. General scheme used to design combinatorial mutation libraries based on computational protein design calculations	36
Table 2-1. Combinatorial libraries designed from sources of structural information	37
Table 2-2. Library coverage	38
Figure 2-3. Fraction-unfolded curves derived from the stability determination of experimental libraries	39
Figure 2-4. Library mutants sorted by experimental stability	40
Table 2-3. Combinatorial libraries designed from the top 16 energy-ranked structures based on two different energy functions	41
Figure 2-5. Library member energies	42
Figure 2-6. Correlation between simulation energy and experimental stability for the cMD-128 library	43
Figure 2-7. Microtiter plate-based stability assay controls	44-5
Figure 3-1. The automated site-directed mutagenesis pipeline	71
Figure 3-2. Visualization of a 96 well plate of SDM products	72
Figure 3-3. Variant construction timeline	73
Figure 3-4. Potential protein unfolding curves	74
Figure 3-5. Precision among experimental measures of protein stability	75
Figure 3-6. Dataset accuracy from literature comparisons	76
Figure 3-7. Point mutant amino acid distributions	77
Figure 4-1. Single mutant stability distribution for the G β 1 domain	108
Table 4-1. Gaussian fitting parameters for the mutational distributions	109
Figure 4-2. Gaussian fits of the G β 1 mutational distribution	110
Figure 4-3. Single mutant stability landscape for the GB1 domain	111
Figure 4-4. Single mutant stability distributions by RESCLASS	112
Figure 4-5. Packing density is linearly correlated with $\Delta\Delta G$ averaged by position	113
Figure 4-6. Amino acid scanning mutagenesis	114
Figure 4-7. Stability distribution of G β 1 by mutant amino acid	115
Figure 4-8. Calculated stability distributions by mutant amino acid	116
Table 4-2. Bioinformatics statistics for selected proteins	117
Table 4-3. Comparing the average $\Delta\Delta G$ of hydrophobic mutations by OSP	118
Table 4-4. Algorithm performance by linear correlation	119
Table 4-5. Algorithm performance by fraction correct	120
Figure 4-9. Complex additivity in core and surface mutation libraries	121
Table 4-6. Identity and stability of additive variants	122

Introduction

Chapter 1

Proteins are biology's workhorse macromolecule, making up about half the dry weight of a typical bacterial cell and responsible for almost every action that occurs inside of it (1). Over the course of natural evolution, proteins have developed prodigious catalytic properties, responsible for a variety of reactions, and exquisite binding activities, key to the cell's signal transduction pathways. Over the last thirty years, all of this functional diversity has become readily available to industrial and clinical biotechnology due to the maturation of recombinant DNA technology. Unfortunately, their application is hindered by nature's handicap: the ability to only select for proteins with activity and stability that provide a biological advantage, and nothing more. This results in a variety of issues for biotechnology, chief among them being the marginal stability of natural proteins. As most organisms on Earth thrive in moderate climates, their proteins have evolved for optimal activity at the same, non-industrially relevant temperatures. In addition, cellular proteins are rapidly turned over in the viscous cytosol, deterring the serendipitous evolution of proteins with long shelf lives under extended *in vitro* conditions. Current and future advances in protein engineering can enrich the number of applicable natural proteins as well as develop customized solutions for current issues in biotechnology.

Protein engineering techniques are centered on two complementary sub-fields, directed evolution and rational structure-based design (2, 3). Directed, or molecular evolution, improves protein properties by making random iterative mutations to a library of sequences and evaluating them either through a direct experimental screen or a functional selection. Larger jumps in sequence space can be achieved by DNA shuffling, a technique that emulates sexual recombination by fragmenting the linear genes of

familial proteins and then stitching them back together (4). This alternative technique overcomes the double-edged sword involved in using conservative random mutagenesis methods to discover novel or dramatic performance enhancements. The major drawback to evolutionary engineering techniques is the application of an appropriate high-throughput screen or selection to sift through the library of mutant sequences. Consequently the adage, “You get what you screen/select for”, is well known to practitioners in the field as substrates or conditions are often altered from those used in the final application for screening purposes. The largest advantage to directed evolution methods is that very little structural information is necessary for isolating enhanced variants, while it is absolutely required for rational design.

Structure-based protein engineering aims to reduce the experimental burden of screening thousands of proteins by rationally predicting desirable mutational effects from structural observations. Efforts in modifying substrate sensitivity found early success (5, 6), and as site-directed mutagenesis techniques improved, the body of literature on mutational tolerance and energetic interactions grew (7, 8). With more scrutiny came an empirical understanding of the difficulty of rational engineering due to the context dependence of mutational effects. This, coupled with the significant amount of experimental data now gathered in the community, sparked computational- and statistical-guided solutions to protein engineering. The ability to evaluate the energy (stability) of structure-sequence pairs *in silico* before doing any bench work represented a tremendous advance in the field. Currently, there are several algorithms in the literature that can predict the stability of a mutation in any globular protein as long as the structure is known (9–12). Fewer methods can efficiently tackle the loftier goals in protein design,

but these advanced software packages have registered several high-profile achievements, including automated redesign (13), extreme thermo-stabilization (14), the design of a novel fold (15), and novel catalysts (16–18).

Despite this success in computational protein engineering and design, the non-robust functioning of these methodologies encumbers their practical use in biotechnology. One example is seen in the muted performance of designed novel catalysts, which leave much to be desired when compared against natural enzymes (19). The poor approximation of the principles important to stability is one factor that dogs both stability prediction and design algorithms, evidenced by the weak-to-moderate linear correlation between calculated and experimental values in recent performance benchmarks (20, 21). Other factors include limited conformational sampling and the absent consideration of explicit non-native states. Due to these issues, the shrewd conjunction of methods in which computational power informs directed evolution screening/selection procedures has proven to be an effective solution to current protein engineering problems (22–24). Going forward, both styles of engineering have much to learn to from each other.

Since its inception, protein design theory has improved through the rigorous cycling between theory and experiment, known as a protein design cycle (25). In order to complete a full cycle, designed sequences had to be synthesized, confirmed by DNA sequencing, translated into protein, purified, and tested before the information gathered could be fed back into the theory. The nature of molecular biology bench work creates bottlenecks at all steps in the design cycle, preventing the rapid iteration of improved protein properties and engineering principles. Commercial solutions for synthesis and

sequencing have improved over the years, but high costs remain an issue. A more economical solution would be to adapt methods from directed evolution, potentially accelerating and broadening the exchange of information between modeling theory and experimental results. Thus, the focus of my graduate work has been to establish experimental high-throughput stability screening methods and subsequently apply them towards the rapid evaluation and improved understanding of proteins.

The second chapter best captures the overall theme of the thesis as we established and applied medium-throughput purification and stability assays to provide a more thorough analysis of core repacking performance when modeling native-state conformational flexibility. Recently developed algorithms for multi-state design (26) and library design generated 24 member libraries from structural inputs such as NMR and molecular dynamics ensembles. The comprehensive experimental stability screening of each library provided insights into the sources of simulation error that crept in from other design approximations. Although a constrained molecular dynamics ensemble produced an entire library of stabilized sequences, issues surrounding the serendipity in library selection prevented our full recommendation of the technique. The large amount of data relative to similar experiments in the literature created an opportunity to discover and discuss the lack of correlation between the calculated and experimental measures of stability. By using high-throughput methodology, we were able to more meticulously validate the applicability of novel computational tools for protein engineering.

Building on the experimental methodology presented in Chapter 2, we raised the bar in the third chapter through the implementation of a liquid handling pipeline that enables the high-throughput construction and stability determination of single-mutant

proteins. Individual automated protocols for the Tecan liquid-handling robot were first developed independently and later strung together in a modular fashion. The methods, better described in the attached robot manual (Appendix), include the automated construction of mutant alleles by PCR site-directed mutagenesis, transformation, and plating of bacterial competent cells, and the expression, purification, and stability determination of mutant proteins. The completed automated pipeline is by no means static, as other sources of protein diversity, such as gene assembly, can easily swap in and take advantage of the high-throughput downstream solutions. To showcase the value and power of the automated system, we carried out a project impossible to achieve through standard bench-top methods: the evaluation of every single mutant of the G β 1 domain. The unbiased, self-consistent nature of the dataset should provide more value toward training next-generation energy functions than what is currently available. Simultaneously, the dense character of the output data coupled with the laboratory's previous work on G β 1 enables an analysis of mutational effects within the context of an entire domain, described in Chapter 4 of this thesis.

The analysis in the last chapter represents insight into mutational outcomes and distributions from the most complete domain-level single mutant stability dataset in the literature. We learn that most single mutations to G β 1 are either neutral or stabilizing, a much discussed topic with implications for protein evolution studies. If we ignore the variants not solubly expressed, the overall distribution can be fit as the sum of core and surface Gaussian distributions. Positional sensitivity to mutation is well predicted by a computational measure of packing density, but better information can likely be gathered from serine scanning mutagenesis. Interestingly, the entire domain was most tolerant of

large hydrophobic residues, a property evidently shared by other, larger proteins. The high-quality dataset can also serve as a benchmark for current stability prediction algorithms. Their lackluster performance should serve as encouragement for the further improvement of energetic approximations. Lastly, the drastic non-additivity seen in variants composed of surface mutations illustrates the knowledge gap that must be bridged before we may reliably and efficiently engineer proteins.

The sum of the work in this thesis is the development and effective use of high-throughput methodology for the rapid testing and improvement of computational theory. As is common in the study of biology, improved technological capabilities lead to more questions, not answers. Nevertheless, the last ten years have seen improved performance from the combination of directed evolution and structure-based design principles. The next ten, hopefully, will strengthen these ties and further realize the benefits protein engineering can bring to biotechnology.

References

1. Voet D & Voet JG (2004) *Biochemistry* (J. Wiley & Sons, New York) 3rd Ed.
2. Chen R (2001) Enzyme engineering: rational redesign versus directed evolution. *Trends in Biotechnology* 19(1):13–14.
3. Arnold FH (2001) Combinatorial and computational challenges for biocatalyst design. *Nature* 409(6817):253–257.
4. Voigt CA, Martinez C, Wang ZG, Mayo SL & Arnold FH (2002) Protein building blocks preserved by recombination. *Nature structural biology* 9(7):553–558.
5. Craik CS, et al. (1985) Redesigning trypsin: alteration of substrate specificity. *Science* 228(4697):291–297.
6. Scrutton NS, Berry A & Perham RN (1990) Redesign of the coenzyme specificity of a dehydrogenase by protein engineering. *Nature* 343(6253):38–43.
7. Fersht AR & Serrano L (1993) Principles of Protein Stability Derived from Protein Engineering Experiments. *Current Opinion in Structural Biology* 3(1):75–83.
8. Eijsink VG, et al. (2004) Rational engineering of enzyme stability. *Journal of Biotechnology* 113(1–3):105–120.
9. 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.
10. Pokala N & Handel TM (2005) Energy functions for protein design: adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. *Journal of Molecular Biology* 347(1):203–227.
11. Yin S, Ding F & Dokholyan NV (2007) Eris: an automated estimator of protein stability. *Nature methods* 4(6):466–467.
12. 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.
13. Dahiyat BI & Mayo SL (1997) De novo protein design: fully automated sequence selection. *Science* 278(5335):82–87.
14. Malakauskas SM & Mayo SL (1998) Design, structure and stability of a hyperthermophilic protein variant. *Nature structural biology* 5(6):470–475.

15. Kuhlman B, et al. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science* 302(5649):1364–1368.
16. Jiang L, et al. (2008) De novo computational design of retro-aldol enzymes. *Science* 319(5868):1387–1391.
17. Rothlisberger D, et al. (2008) Kemp elimination catalysts by computational enzyme design. *Nature* 453(7192):190–195.
18. Siegel JB, et al. (2010) Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science* 329(5989):309–313.
19. Baker D (2010) An exciting but challenging road ahead for computational enzyme design. *Protein Science* 19(10):1817–1819.
20. 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.
21. 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.
22. Voigt CA, Mayo SL, Arnold FH & Wang ZG (2001) Computational method to reduce the search space for directed protein evolution. *Proceedings of the National Academy of Sciences of the United States of America* 98(7):3778–3783.
23. Hayes RJ, et al. (2002) Combining computational and experimental screening for rapid optimization of protein properties. *Proceedings of the National Academy of Sciences of the United States of America* 99(25):15926–15931.
24. Chica RA, Doucet N & Pelletier JN (2005) Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design. *Current Opinion in Biotechnology* 16(4):378–384.
25. Street AG & Mayo SL (1999) Intrinsic beta-sheet propensities result from van der Waals interactions between side chains and the local backbone. *Proceedings of the National Academy of Sciences of the United States of America* 96(16):9074–9076.
26. Allen BD & Mayo SL (2010) An efficient algorithm for multistate protein design based on FASTER. *Journal of computational chemistry* 31(5):904–916.

**Experimental library screening demonstrates the successful
application of computational protein design to large structural
ensembles**

Chapter 2

The text of this chapter was adapted from a manuscript coauthored with Benjamin D. Allen and Stephen L. Mayo.

Nisthal A*, Allen BD*, and Mayo SL (2010) Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. *Proc Natl Acad Sci USA* 107(46):19838–19843. (*Both authors made equal contributions.)

Abstract

The stability, activity, and solubility of a protein sequence are determined by a delicate balance of molecular interactions in a variety of conformational states. Even so, most computational protein design methods model sequences in the context of a single native conformation. Simulations that model the native state as an ensemble have been mostly neglected due to the lack of sufficiently powerful optimization algorithms for multi-state design. Here, we have applied our multi-state design algorithm to study the potential utility of various forms of input structural data for design.

To facilitate a more thorough analysis, we developed new methods for the design and high-throughput stability determination of combinatorial mutation libraries based on protein design calculations. The application of these methods to the core design of a small model system produced many variants with improved thermodynamic stability, and showed that multi-state design methods can be readily applied to large structural ensembles. We found that exhaustive screening of our designed libraries helped to clarify several sources of simulation error that would have otherwise been difficult to ascertain.

Interestingly, the lack of correlation between our simulated and experimentally measured stability values shows clearly that a design procedure need not reproduce experimental data exactly to achieve success. This surprising result suggests potentially fruitful directions for the improvement of computational protein design technology.

Introduction

Protein-engineering efforts based on directed evolution have met with considerable success (1-3). In tandem, structure-based computational protein design (CPD) methods have been developed to allow screening for desirable sequences to be performed *in silico* (4-6). Despite a number of high-profile results that demonstrate the utility of CPD (7-12), the routine computational design of functional proteins remains elusive. Thus, many current efforts focus on the improvement of CPD methodology or on the synergistic application of CPD with experimental high-throughput screening or selection (13).

Although the stability, solubility, and activity of a protein depend on the relative energetic contributions of many conformational states, including ensembles of native, unfolded, and aggregated structures (14), most CPD methods evaluate sequences based on their energies in the context of one fixed backbone structure. This simplification has made design results undesirably sensitive to slight changes in main-chain and side-chain conformation, and has made difficult the selection of sequences with amino acid composition similar to naturally occurring protein. These issues have been approached via the use of high-resolution structural templates, expanded rotamer libraries (15, 16), energy functions with softened repulsive terms (17, 10, 18), iteration between structural refinement and sequence design (10, 19), and amino acid reference energies (10, 20). Although these strategies can help to mitigate the impact of the fixed-backbone approximation, they do not address the fundamental reality that sequence fitness is a function of multiple conformational states.

In a handful of cases, multi-state design (MSD) procedures have been used to find sequences that simultaneously stabilize or destabilize a combination of a few different conformational states (21-23). However, MSD techniques have not yet been applied to native ensembles with many conformational states that might better reflect the flexibility of real proteins. The degree to which various energy functions, rotamer libraries, and structural templates of single-state design (SSD) might be appropriate for this type of MSD calculation is, so far, unknown. We recently developed a framework for MSD that allows for efficient sequence optimization given hundreds of conformational states (24). Here, we have applied this framework to test the applicability of current CPD methods to large structural ensembles, and to investigate whether the use of such ensembles might result in the selection of more desirable sequences by CPD.

The most basic goal of CPD has been to optimize interactions between amino acid side chains to promote thermodynamic stability of the native state. Unfortunately, standard methods for the measurement of protein stability are too laborious to allow the testing of more than a few designed variants, and the top-scoring sequence produced by a new design procedure does not yet sufficiently reflect its general utility. Fortunately, recent progress in laboratory automation has allowed us to construct an efficient pipeline for the basic evaluation of new procedures in CPD. In our scheme, gene libraries are assembled from degenerate oligonucleotides, proteins are expressed and purified in microtiter plates, and liquid-handling robotics assist in the preparation of chemical denaturation series in a 96 well format for assay by tryptophan fluorescence. The integration of these technologies has allowed us to assess the stability of hundreds of

designed protein variants with minimal experimenter intervention and limited incremental expense.

Given several design procedures to evaluate and a high-throughput experimental assay, we needed a general and rigorous method to choose a limited number of representative sequences to test from each design. Although several useful computational protein library design methods have been developed (25-28), none reported so far takes directly into account simulation energies, allows control over library size and possible sets of amino acids, and eschews heuristics that can introduce bias into the libraries it produces. So that our experimental results might better reflect the results of the underlying CPD calculations, we developed a new library design procedure, called Combinatorial Libraries Emphasizing And Reflecting Scored Sequences (CLEARSS), which satisfies these criteria.

We used standard single-state design (SSD) and MSD to redesign the core of the small, stable domain G β 1 based on several sources of structural information, including a crystal structure, an NMR structure, and MD simulations. Our efforts were motivated by a curiosity about the relative merits of different sources of structural data for design, and the hypothesis that use of a structural ensemble might help to correct for design failures observed in SSD. Because the imperfect nature of CPD limits the conclusions that can be drawn from a comparison of single sequences, we developed new methods for the computational design and high-throughput experimental stability determination of combinatorial protein libraries. The results we report here provide simultaneous experimental validation for (1) the application of multi-state protein design methods to large conformational ensembles, (2) the transformation of arbitrary CPD results into

combinatorial mutation libraries, and (3) the experimental stability determination of these libraries by high-throughput gene assembly, protein expression, purification, and screening.

Results and discussion

Designed libraries

To simplify the validation of our multi-state design methods, we applied them to a previously studied set of core positions (Figure 2-1) in a small model system, protein G β 1, and relied on a set of energy functions that previously found stabilized variants of this sequence (17). We assessed these methods by performing designs based on each of the following sources of structural information: a crystal structure (xtal-1), an NMR-constrained minimized average solution structure (NMR-1), an NMR ensemble (NMR-60), a constrained MD ensemble (cMD-128), and an unconstrained MD ensemble (uMD-128). Our new algorithm for library design (Figure 2-2) was then applied to produce degenerate oligonucleotide sequences that reflect quantitatively the amino acid preferences determined by the design calculations. Given the requirements for purified protein of our stability assay, we chose to design and screen a 24-member library based on each structural data source described above.

All five designed libraries comprise relatively conservative sets of mutations away from the wild-type sequence (Table 2-1). The libraries other than uMD-128 share many characteristics in common. Each of these libraries chose only the wild-type amino acid at positions A20, A26, F30, and A34. Every member of each of these four libraries contained the single-mutant Y3F, which previous experiments have shown to be well

tolerated by the structure. These four libraries all allowed the wild-type amino acid at every other position, and all contain the most stable G β 1 core variant previously characterized, Y3F+L7I+V39I (17).

The two NMR libraries were extremely similar to each other: both chose the amino acids FILV at position 52, and directed the remaining diversity to positions 7 and 39. In contrast, xtal-1 and cMD-128 allowed only the wild-type Phe at position 52, and instead allocated diversity towards positions 7, 39, and 54. xtal-1 differs from cMD-128 in that it gave up L7F and V39L to allow L5I. The unconstrained MD ensemble library uMD-128 was the least conservative, specifying a size reversal of two nearby residues via mutations L5A and A34F, and diversity at residue 30, a position untouched in the other libraries.

As shown in Table 2-2, the designed libraries generally succeeded in representing the top-scoring sequences from each design calculation, given the constraints imposed by the genetic code. The exception was the uMD-128 library, which represented only three of the best 100 sequences from the original design calculation. This was caused by an unusual designed sequence list, in which the best-scoring sequence contained a feature (the size reversal mentioned above) that was very uncommon in the remainder of the list.

Experimental characterization of designed libraries

Experimental screening of the xtal-1 library (Figure 2-3A) showed two distinct sets of variants. The 12 library members with wild-type Leu at position 5 all exhibited stabilities similar to or better than the wild-type sequence, while the 12 with Ile at position 5 were all significantly destabilized. Screening of the NMR-based libraries

(Figures 2-3B and 2-3C) showed a similar dichotomy. In each case, the 6 library members with the wild-type Phe at position 52 exhibited wild-type-like stability or better. The remaining 18 variants from each NMR-based library were highly destabilized, and many lacked enough of a pretransition to be fit to the two-state unfolding model.

Evaluation of the MD libraries indicated that all 24 variants from the constrained library, cMD-128, had stability similar to the wild-type protein or better (Figure 2-3D). In contrast, all 24 variants from the uMD-128 library failed to produce any significant change in fluorescence signal across the denaturation series, and likely assume an alternative structure, as discussed below. Sorting the individual sequence members from every library except uMD-128 according to experimental stability (Figure 2-4) shows that the cMD-128 input structural ensemble favored more high-quality sequences (better than wild type) than any other library. Every other designed library specified at least one problematic substitution that rendered many of its sequences destabilized or otherwise unlike the wild type.

Impact of input structural data on designed libraries

Why were apparently destabilizing mutations such as L5I, F52ILV, and A34F chosen by the design procedure? These mutations were all present in high-scoring sequences from the original design calculations, and thus reflect real preferences of the original design procedures, rather than artifacts introduced by the library design process.

The selection of the amino acids FILV at position F52 in the two NMR-based libraries resulted in three quarters of each library being significantly destabilized. In the context of the NMR structures, no Phe rotamer in the library was able to fit perfectly at

position 52, encouraging the selection of smaller amino acids. If the set of rotamers at this position is supplemented with the observed rotamer in each structure, the designs choose to allocate diversity to positions 7 and 39, resulting in libraries similar to xtal-1. This result highlights how dramatically the rotameric approximation can influence the results of a design, despite our biophysical intuition that a solution ensemble might better reflect protein structure than a single crystallographic snapshot. It suggests that, at the very least, rotamers optimized for the wild-type sequence should be included when the goal is to simply find desirable sequences. For this project, we omitted the structurally observed rotamer at each position in order to limit the significant bias towards the wild-type sequence that these rotamers tend to cause. In the context of a real-world protein-engineering project, including these rotamers would have considerably increased our chances of success. Interestingly, this failure of discrete rotamers even occurred in the design of the NMR ensemble, indicating that continuous side-chain optimization may be useful during design, even when allowing conformational flexibility of the main chain.

The L5I mutation, which caused half of the xtal-1 library members to be destabilized relative to the wild-type sequence, may have been selected due to a failure of the softened repulsive contact potential that is used to counteract unrealistic rigidity introduced by the CPD model. The γ methyl group of Ile5 bumps into a Thr residue on an adjacent β strand and is scored as a serious clash using unscaled van der Waals radii, but appears innocuous with the atomic radius scaling factor of $\alpha = 0.9$ that we used for the designs evaluated here (17). Repeating the design calculations with radii scaled by intermediate values such as 0.925 and 0.95 prevents Ile from being chosen at position 5, but also increases the frequency with which smaller residues are chosen at position F52.

Interestingly, the recommendation of $\alpha = 0.9$ is derived from previous experiments based on the same set of G β 1 core positions that were designed here. The earlier work drew conclusions based only on the best-scoring sequences produced by the design calculations, and found no difference between scaling atomic radii by 0.9 or 0.95 (17). Our results indicate that the mutations produced by the design procedure vary significantly with values of α between 0.9 and 0.95 when more sequences are taken into account. Therefore, a more rigorous investigation of appropriate α values for design may be warranted. Although the L5I mutation might also be reasonably attributed to the fixed main chain and discrete rotamers, several good-scoring libraries based on the constrained MD ensemble also contained this mutation (see below). Since the additional conformational diversity provided by the ensemble did not inhibit this design failure, we find explanations related to energy function more plausible.

To analyze the uMD-128 data, it is important to note that our stability assay reports on the environment of the single Trp residue of G β 1. Changes in packing caused by substitutions at other positions could alter the native-state environment of Trp43 enough to flip its side chain out into solution or change its fluorescence properties, crippling our ability to monitor unfolding by fluorescence. This interpretation seems unlikely for the destabilized members of the crystal structure and NMR libraries, for which a partial unfolding transition is clearly indicated by the raw data. However, the members of the uMD-128 library fail to show any such a transition, rendering the validity of our assay suspect in this case.

A constant feature of the uMD-128 library is a size reversal specified by mutations A34F and L5A. The model structures produced by this design were well

packed and contained no obvious flaws such as Trp43 flipping out into the solvent. Previous characterization of several G β 1 variants that include mutation A34F has indicated that these sequences assume oligomeric structures and exhibit altered fluorescence properties (29-31). This suggests that the structural basis for our designs, as well as our fluorescence assay, may be inappropriate for sequences containing this mutation. When we reanalyzed a subset of the uMD-128 variants using circular dichroism, they uniformly displayed wild-type-like secondary structure but lower stability and low levels of protein expression. The previous reports and our new results indicate that the uMD-128 library sequences likely assume structures different from the design target. As target structures move away from experimentally determined structures and greater sequence diversity is enabled (32, 33), more effective negative design strategies may be required to exclude sequences that preferentially adopt alternative conformations.

A recent theoretical analysis of NMR and crystal structures as templates for design has suggested that some individual members of NMR ensembles might be more appropriate templates than others (34). To assess the impact this might have had on our results, we ranked the members of each structural ensemble by DREIDING energy (35) and separately by Rosetta energy (36). We then designed new libraries using only the top 16 energy-ranked structures from each ensemble using each energy ranking (Table 2-3). The two new libraries produced from the NMR structural ensemble were similar to those from the original design; both specified diversity at position 52 and contain destabilized sequences. The library based on the top 16 DREIDING-ranked sequences from the constrained MD ensemble only specifies known non-destabilizing substitutions, whereas

the top 16 Rosetta-ranked structures again gave diversity at position 52. For the unconstrained MD ensemble, the top 16 Rosetta-ranked structures gave a library very similar to that produced by the entire ensemble, and the top 16 DREIDING-ranked structures gave a library of sequences that appear severely overpacked. In total, the libraries produced from the top-ranked sequences were similar to those produced from the full ensembles in four cases, and were worse in the remaining two cases. Based on this post-hoc analysis, our multi-state library design procedure seems robust to the influence of poor templates within each ensemble. However, more sophisticated methods of template selection may ultimately prove more fruitful. For example, it might be interesting to choose a subset of a structural ensemble according to the degree to which individual members are able to recover wild-type-like sequences, and apply MSD to this subset rather than to the entire ensemble.

Influence of the designed library selection method

At this point, it is important to address the degree to which the library design method might affect the conclusions we draw from our experiments. The CLEARSS library design procedure was developed with an understanding that many different combinatorial libraries may similarly represent a given list of scored sequences. It is intended to produce a list of the top-scoring designed combinatorial libraries that satisfy all constraints, and to let the user choose between them. In general, this choice might be influenced by chemical intuition or prior mutational data, and thus partially account for properties of the system that are not modeled during the design procedure. To make our evaluation of input structural data sources as fair as possible, we chose to ignore such

influences and apply an objective strategy based on the energies of the sequences in the libraries. Still, we must ask how the other libraries generated by CLEARSS would have fared in our experimental assay.

Each of the top 20 designed libraries based on the NMR ensemble, and each based on the single average NMR structure, assigned smaller residues than the wild-type Phe to position 52. The remaining diversity of each library was occupied by various combinations of the other mutations present in the xtal-1, NMR-1, and NMR-60 libraries we screened in this work. It seems very likely, then, that the screening of any of the top NMR-based libraries from our designs would have resulted in stabilities similar to those we have reported here. Similarly, all of the top 20 designed libraries based on the unconstrained MD ensemble contained mutations L5A and A34F, and would be expected to exhibit properties similar to uMD-128.

A more interesting case is provided by the designs based on the crystal structure and constrained MD ensemble. Our analysis of the libraries xtal-1 and cMD-128 produced by these designs seems to indicate that cMD-128 was more successful, since a much greater fraction of its members were shown to be highly stable. However, when the top 20 libraries from each design were inspected in aggregate, it became clear that the xtal-1 and cMD-128 designs had produced a variety of libraries, some featuring the destabilizing mutations described above. Both the xtal-1 library and the cMD-128 library were found in the top 20 set of libraries produced by each design. Furthermore, each design produced several libraries with diversity at position 52, like NMR-1 and NMR-60.

The influence of the library design procedure on a comparison between structural inputs can also be assessed by scoring the sequences from each library on each of the

other input structures or ensembles. Histograms of these energies (Figure 2-5) show that each structural input prefers the sequences from its own library over those from other libraries, though often by narrow margins.

These observations, taken in total, suggest that the library design method we used did not unduly influence our optimistic conclusions about the merits of high-quality structural ensembles as inputs for computational protein design.

Approximation in computational protein design

In addition to helping validate the use of multi-state and combinatorial library design methods for computational protein design, our results also reflect unexpectedly on protein design itself. Plots of experimental stability versus simulation energy for the cMD-128 library (Figure 2-6) failed to yield any correlation, despite the apparent success of this design calculation. Likewise, the design calculations for xtal-1 and the NMR libraries failed to predict the pronounced destabilizing effects of mutations L5I or F52L, even though these designs also found a variety of stabilized variants.

An intuitive perspective on the development of CPD methods is that improvements in designed sequences will follow from improvements in our ability to predict or rank experimental stabilities (37). However, recent advances in stability prediction procedures (38, 39) have not yet, to our knowledge, produced the expected benefits to combinatorial protein design. Our results are consistent with a recent assessment of stability prediction methods, which found that the ability to reproduce experimental stability rankings is unnecessary for useful CPD (40). These conclusions

prompt a modified view of the factors that make structure-based design possible in the first place.

Protein structures relax to accommodate mutations, and the computational difficulty of simulating and scoring these relaxed structures has so far rendered intractable the accurate stability ranking of sequence variants with many mutations. Fortunately, this malleability also means that sequences chosen to fit into a rigid protein model, even using approximate energy functions, will likely be tolerated by whatever relaxed structure results from the mutations they contain. In this way, the soft material properties of proteins serve to impede the development of the accurate quantitative protein design methods, but also enable the more qualitative methods we can apply today.

The standard view of CPD has been as a single, rigorously quantitative problem: correct packing of amino acid side chains into a high-resolution template structure leads to a stable and well-behaved designed sequence. However, our analysis supports a revised view of CPD, comprising two distinct problems: (1) to find areas of sequence space that can favorably adopt the target structure, and (2) to avoid areas of sequence space that might favorably adopt alternate structures. The first problem is simply an enhancement of the original formulation of CPD in which we admit that current methods for native-state sequence selection are approximate, and focus on finding areas of sequence space enriched with variants that satisfy the target fold.

The second problem has typically been treated implicitly, as discussed above. The energy function used in this work applies a simple tripeptide model of the reference state for solvation energies, and assumes that all other interactions average out in the unfolded ensemble. However, issues such as those encountered with the uMD-128

library likely cannot be addressed in a general way without the use of explicit competing state models. Such simulations are more difficult than those that model only the native state, in large part because few non-native states have been characterized experimentally. In alpha helical peptide systems where large numbers of undesirable states are readily identifiable, explicit negative design has yielded improvements in structural specificity (41). We hope that general models of unfolded and aggregated states will lead to similar improvements in the design of globular proteins.

Conclusions

We enlisted new methods for the design and screening of combinatorial libraries to test the application of multi-state design procedures to several structural ensembles, and to compare the resulting designs to those based on single structures. Single-state and multi-state designs based on NMR data produced similar sets of libraries; likewise did those based on crystallographic data. Although an MD-based library gave superlative results, we cannot definitively conclude that the use of a structural ensemble provides any particular advantage over a single high-resolution structure for the purposes of design. Nevertheless, this initial success confirms that the energy functions and rotamer libraries developed for single-state modeling are equally applicable for the multi-state design of large structural ensembles.

This work also provides further support in favor of rigorously screening an area of sequence space discovered by simulation, and has helped in vetting our new, general method for library design. For some designs that specified undesired destabilizing mutations, library screening suggested underlying causes for design failure that would not

have been apparent via the ad-hoc testing of individual sequences. Because our library design procedure is specifically intended to faithfully represent its input scored sequence list, and is indifferent to the origin of the list, it should be more useful for the evaluation of new design procedures than its predecessors.

Current design procedures seem to find stable sequences by selecting mutations that are likely to be accommodated by a relaxed version of the template structure, and not by accurately ranking the mutations relative to each other. Given that protein stability and function depend on competing states as well as the native state, the poor agreement we observed between simulated and experimental energies in our successful libraries suggests that future effort towards explicit negative design is warranted.

Materials and methods

Input structural data

Input atomic coordinates for the $\beta 1$ domain of Streptococcal protein G ($G\beta 1$) were taken from the 2.2 Å crystal structure 1pga (42), the 60 member NMR structural ensemble 1gb1, and a constrained, minimized average structure generated from the ensemble 2gb1 (43). Hydrogens (if any) were stripped from each structure, and new hydrogen positions were optimized along with side-chain amide and imidazolium group flips using REDUCE (44). Each structure was then standardized with 50 steps of conjugate gradient minimization using the DREIDING force field (35). An unconstrained 128 member molecular dynamics (MD) ensemble was generated from the minimized crystal structure by running a 12.8 ps MD trajectory at 300 K in vacuum using the DREIDING force field and saving the coordinates every 0.1 ps. The constrained MD

trajectory was generated by the same procedure, using an additional harmonic point restraint with a force constant of $100 \text{ kcal/mol/\AA}^2$ applied to keep C_α atoms near their initial positions. Each MD snapshot was standardized as described above. After standardization, the NMR, unconstrained MD, and constrained MD ensembles exhibited average pairwise main-chain RMSDs of 0.25, 0.84, and 0.12 \AA , respectively.

Sequence design specifications and energy calculations

In the sequence designs, ten core positions of G β 1 (3, 5, 7, 20, 26, 30, 34, 39, 52, and 54), were allowed to assume any of the hydrophobic amino acids A, V, L, I, F, Y, and W. Tryptophan 43 was allowed to change conformation but not amino acid type, so that our fluorescence-based stability assay would not be compromised. Allowed side-chain conformations at the variable positions were taken from the Dunbrack backbone-dependent rotamer library with expansions of ± 1 standard deviation around χ_1 and χ_2 (15). To avoid bias toward the wild-type sequence, this set was not supplemented with the side-chain coordinates from the input structure, except at position 43. All other side chains and the main chain were fixed in the input conformation. Pairwise energies were computed for each structure or ensemble member using energy functions described previously (45, 46), with the polar hydrogen burial term omitted.

Sequence optimization

FASTER was used to find optimized sequences in the single-state design of the crystal structure and the NMR constrained minimized average (47). Multi-state sequence optimization of each ensemble was performed as described (24). The energies of a

sequence in the context of ensemble member were combined into a single score by computing the free energy of the ensemble system at 300 K:

$$A = -kT \log\left(\sum_j e^{-E_j/kT}\right)$$

where each E_j is the energy of the sequence when threaded on member j of the ensemble. While various functions could be used to combine the state energies into a single score, we chose the free energy function over other averaging schemes because it prefers sequences that satisfy multiple states in a physically reasonable way that does not require any particular number of states to be satisfied.

Combinatorial library design

To choose combinatorial sequence libraries for experimental screening, we used a new algorithm reported here (Figure 2-2). Given a list of scored sequences, a list of allowed sets of amino acids, and a range of desired library sizes, the method evaluates all possible combinations of sets of amino acids at different positions that lead to a library with a size in the desired range. Each position in each library is scored by summing the Boltzmann weights of the sequences in the list that contain a library-specified amino acid at that position. The position scores are then summed to give an overall library score. Our algorithm is able to consider all possible libraries because it treats positions independently, and because it ignores amino acid sets that are unnecessarily large in the context of a given position. In this work, we allowed only those sets of amino acids that can be specified by degenerate codons that do not include codons observed with low frequency in *E. coli*. A temperature of 300 K was used in the Boltzmann weighting, and

the target library size was 24. Setting the desired library size to other values, such as 12 or 48, gave libraries composed of the same mutations found in the 24 member libraries.

After applying this algorithm to the lists of sequences produced by the computational designs, we instantiated the 20 best-scoring libraries from each design and rescored all of the amino acid sequences in each library by rotamer optimization. Each library we inspected contained the best-scoring sequence from the design it was based on, although this is not required by the method. From each design, we chose for experimental testing the library in the top 20 with the smallest energy spread between its best-scoring and worst-scoring sequence.

Library construction, expression, and purification

Oligonucleotides (Integrated DNA Technologies) containing ~ 18 bp overlapping segments were assembled via a modified Stemmer method (48) using KOD Hot Start Polymerase (Novagen) to generate full-length streptococcal G β 1 with an N-terminal His₆ tag. Secondary structure content and annealing temperatures were verified by NUPACK (49, 50). For each library, oligonucleotides containing the desired single mutation or degenerate codon were swapped into the assembly mixture. Standard subcloning techniques were performed to first insert the library into the frameshift selection plasmid pInSAlect (51) and finally into an expression plasmid (pET11a). The library was transformed into BL21 Gold DE3 cells (Stratagene) and colonies were picked into 96 well plates for plasmid miniprepping and sequencing (Agencourt Biosciences). Any missing library members were generated by standard quick-change protocols. Sequence-verified library members were pulled from replicated glycerol stocks and inoculated into

Instant TB media (Novagen) in 24 well plates. After overnight incubation at 37°C, cells were pelleted by centrifugation. Pellets were freeze/thawed once and resuspended in 1x CellLytic B (Sigma-Aldrich) lysis buffer before another identical centrifugation step. Cell lysates were loaded onto an equilibrated HIS-Select filter plate (Sigma-Aldrich), washed twice, and eluted with buffer containing 250 mM imidazole, pH 8.

Microtiter plate-based stability determination

Appropriate amounts of GdmCl (Sigma-Aldrich), Milli-Q water, eluted protein, and NaPO₄ buffer, pH 6.5, were added to maintain a fixed volume in each well of 96 well Costar UV transparent flat bottom plates by a Freedom EVO liquid handling robot (Tecan). Adapting a previously reported stability assay, mutant proteins were subjected to a 12 point GdmCl gradient across the columns of the plate, where each row contained a separate denaturation experiment (52). The plates were equilibrated for at least one hour and shaken at 900 rpm on a microtiter plate shaker (Heidolph).

Tryptophan fluorescence measurements were taken on a fluorescence plate reader (Tecan) with a plate stacker attachment. Parameters empirically determined for wild-type Gβ1 were later used for each library assayed. Excitation was performed at 295 nm and emission measured at 341 nm with 10 nm bandwidths. Data were fit as a two-state unfolding transition using the linear extrapolation method (53) in Pylab. The GdmCl concentration at the midpoint of denaturation, C_m , was estimated numerically based on the fraction-unfolded curve fit.

Microtiter plate-based stability assay controls

The fluorescence profiles of the GdmCl gradient and the elution buffer show no effect on the shape of the unfolding transition of wild-type G β 1 (Figure 2-7A). Sample signal below the elution buffer was interpreted as expression failure; any sample whose data could not be fit yet whose signal was above the elution buffer was deemed expressed but unstable, unfolded, or misfolded. In order to test the accuracy of the microtiter plate-based denaturation assay, G β 1 unfolding was monitored by circular dichroism (Aviv Biomedical) and tryptophan fluorescence in a fluorimeter (Photon Technology International). The denaturation profiles from these low-throughput experiments were compared to results from the fluorescence plate reader (Figure 2-7B). The overlapping data points support the use of a two-state unfolding fit during our stability calculations and verify the accuracy of the assay. Next, the unfolding curves from several protein preparations from different concentrations confirmed the assay's precision (Figure 2-7C). These results support some assumptions that the stability determination method described here makes in order to maintain a high level of throughput. First, we never assay for protein concentration before setting up the GdmCl gradient, relying on the fraction-unfolded plot to remove any concentration bias/effects. Second, the high concentration (250 mM) of imidazole in elution buffer is never dialyzed out of the eluted protein solution. Figures 2-7B–C show that these discrepancies in protein preparation have no significant effect on fraction unfolded plots for the wild-type protein.

References

1. Arnold FH (2001) Combinatorial and computational challenges for biocatalyst design. *Nature* 409(6817):253–257.
2. Bershtein S & Tawfik DS (2008) Advances in laboratory evolution of enzymes. *Current Opinion in Chemical Biology* 12(2):151–158.
3. Jackel C, Kast P & Hilvert D (2008) Protein design by directed evolution. *Annual Reviews of Biophysics* 37:153–173.
4. Schueler-Furman O, Wang C, Bradley P, Misura K & Baker D (2005) Progress in modeling of protein structures and interactions. *Science* 310(5748):638–642.
5. Alvizo O, Allen BD & Mayo SL (2007) Computational protein design promises to revolutionize protein engineering. *Biotechniques* 42(1):31–35.
6. Lippow SM & Tidor B (2007) Progress in computational protein design. *Current Opinion in Biotechnology* 18(4):305–311.
7. Dahiyat BI & Mayo SL (1997) De novo protein design: Fully automated sequence selection. *Science* 278(5335):82–87.
8. Malakauskas SM & Mayo SL (1998) Design, structure and stability of a hyperthermophilic protein variant. *Nature Structural Biology* 5(6):470–475.
9. Bolon DN & Mayo SL (2001) Enzyme-like proteins by computational design. *Proceedings of the National Academy of Sciences of the United States of America* 98(25):14274–14279.
10. Kuhlman B et al. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science* 302(5649):1364–1368.
11. Jiang L, et al. (2008) De novo computational design of retro-aldol enzymes. *Science* 319(5868):1387–1391.
12. Rothlisberger D, et al. (2008) Kemp elimination catalysts by computational enzyme design. *Nature* 453(7192):190–195.
13. Chica RA, Doucet N & Pelletier JN (2005) Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design. *Current Opinion in Biotechnology* 16(4):378–384.
14. Shortle D (1996) The denatured state (the other half of the folding equation) and its role in protein stability. *FASEB Journal* 10(1):27–34.
15. Dunbrack RL & Cohen FE (1997) Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Science* 6(8):1661–1681.

16. Lassila JK, Privett HK, Allen BD & Mayo SL (2006) Combinatorial methods for small-molecule placement in computational enzyme design. *Proceedings of the National Academy of Sciences of the United States of America* 103(45):16710–16715.
17. Dahiyat BI & Mayo SL (1997) Probing the role of packing specificity in protein design. *Proceedings of the National Academy of Sciences of the United States of America* 94(19):10172–10177.
18. Grigoryan G, Ochoa A & Keating AE (2007) Computing van der Waals energies in the context of the rotamer approximation. *Proteins* 68(4):863–878.
19. Hu X, Wang H, Ke H & Kuhlman B (2007) High-resolution design of a protein loop. *Proceedings of the National Academy of Sciences of the United States of America* 104(45):17668–17673.
20. Pokala N & Handel TM (2005) Energy functions for protein design: adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. *Journal of Molecular Biology* 347(1):203–227.
21. Havranek JJ & Harbury PB (2003) Automated design of specificity in molecular recognition. *Nature Structural Biology* 10(1):45–52.
22. Ambroggio XI & Kuhlman B (2006) Computational design of a single amino acid sequence that can switch between two distinct protein folds. *Journal of the American Chemical Society* 128(4):1154–1161.
23. Boas FE & Harbury PB (2008) Design of protein-ligand binding based on the molecular-mechanics energy model. *Journal of Molecular Biology* 380(2):415–424.
24. Allen BD & Mayo SL (2010) An efficient algorithm for multistate protein design based on FASTER. *Journal of Computational Chemistry* 31:904–916.
25. Kono H & Saven JG (2001) Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. *Journal of Molecular Biology* 306(3):607–628.
26. Hayes RJ et al. (2002) Combining computational and experimental screening for rapid optimization of protein properties. *Proceedings of the National Academy of Sciences of the United States of America* 99(25):15926–15931.
27. Mena MA & Daugherty PS (2005) Automated design of degenerate codon libraries. *Protein Engineering Design & Selection* 18(12):559–561.
28. Treynor TP, Vizcarra CL, Nedelcu D & Mayo SL (2007) Computationally designed libraries of fluorescent proteins evaluated by preservation and diversity

of function. *Proceedings of the National Academy of Sciences of the United States of America* 104(1):48–53.

29. 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.
30. 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.
31. Jee J, Byeon IJ, Louis JM & Gronenborn AM (2008) The point mutation A34F causes dimerization of GB1. *Proteins* 71(3):1420–1431.
32. Larson SM, England JL, Desjarlais JR & Pande VS (2002) Thoroughly sampling sequence space: large-scale protein design of structural ensembles. *Protein Science* 11(12):2804–2813.
33. Fu X, Apgar JR & Keating AE (2007) Modeling backbone flexibility to achieve sequence diversity: the design of novel alpha-helical ligands for Bcl-xL. *J Mol Biol* 371(4):1099–1117.
34. Schneider M, Fu X & Keating AE (2009) X-ray vs. NMR structures as templates for computational protein design. *Proteins* 77(1):97–110.
35. Mayo SL, Olafson BD & Goddard WA (1990) Dreiding—a Generic Force-Field for Molecular Simulations. *Journal of Physical Chemistry* 94(26):8897–8909.
36. Rohl CA, Strauss CE, Misura KM & Baker D (2004) Protein structure prediction using Rosetta. *Methods Enzymol* 383:66–93.
37. Mendes J, Guerois R & Serrano L (2002) Energy estimation in protein design. *Current Opinion in Structural Biology* 12(4):441–446.
38. 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.
39. Yin S, Ding F & Dokholyan NV (2007) Eris: an automated estimator of protein stability. *Nature Methods* 4(6):466–467.
40. 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 Eng Des Sel* 22(9):553–560.
41. Grigoryan G, Reinke AW & Keating AE (2009) Design of protein-interaction specificity gives selective bZIP-binding peptides. *Nature* 458(7240):859–864.

42. Gallagher T, Alexander P, Bryan P & Gilliland GL (1994) 2 Crystal-Structures of the B1 Immunoglobulin-Binding Domain of Streptococcal Protein-G and Comparison with Nmr. *Biochemistry* 33(15):4721–4729.
43. Gronenborn AM et al. (1991) A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. *Science* 253(5020):657–661.
44. Word JM, Lovell SC, Richardson JS & Richardson DC (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of Molecular Biology* 285(4):1735–1747.
45. Gordon DB, Marshall SA & Mayo SL (1999) Energy functions for protein design. *Current Opinion in Structural Biology* 9(4):509–513.
46. Gordon DB, Hom GK, Mayo SL & Pierce NA (2003) Exact rotamer optimization for protein design. *Journal of Computational Chemistry* 24(2):232–243.
47. Allen BD & Mayo SL (2006) Dramatic performance enhancements for the FASTER optimization algorithm. *Journal of Computational Chemistry* 27(10):1071–1075.
48. Stemmer WP, Cramer 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.
49. Dirks RM & Pierce NA (2003) A partition function algorithm for nucleic acid secondary structure including pseudoknots. *Journal of Computational Chemistry* 24(13):1664–1677.
50. Dirks RM & Pierce NA (2004) An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. *Journal of Computational Chemistry* 25(10):1295–1304.
51. Gerth ML, Patrick WM & Lutz S (2004) A second-generation system for unbiased reading frame selection. *Protein Engineering Design & Selection* 17(7):595–602.
52. Aucamp JP, Cosme AM, Lye GJ & Dalby PA (2005) High-throughput measurement of protein stability in microtiter plates. *Biotechnology and Bioengineering* 89(5):599–607.
53. 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.

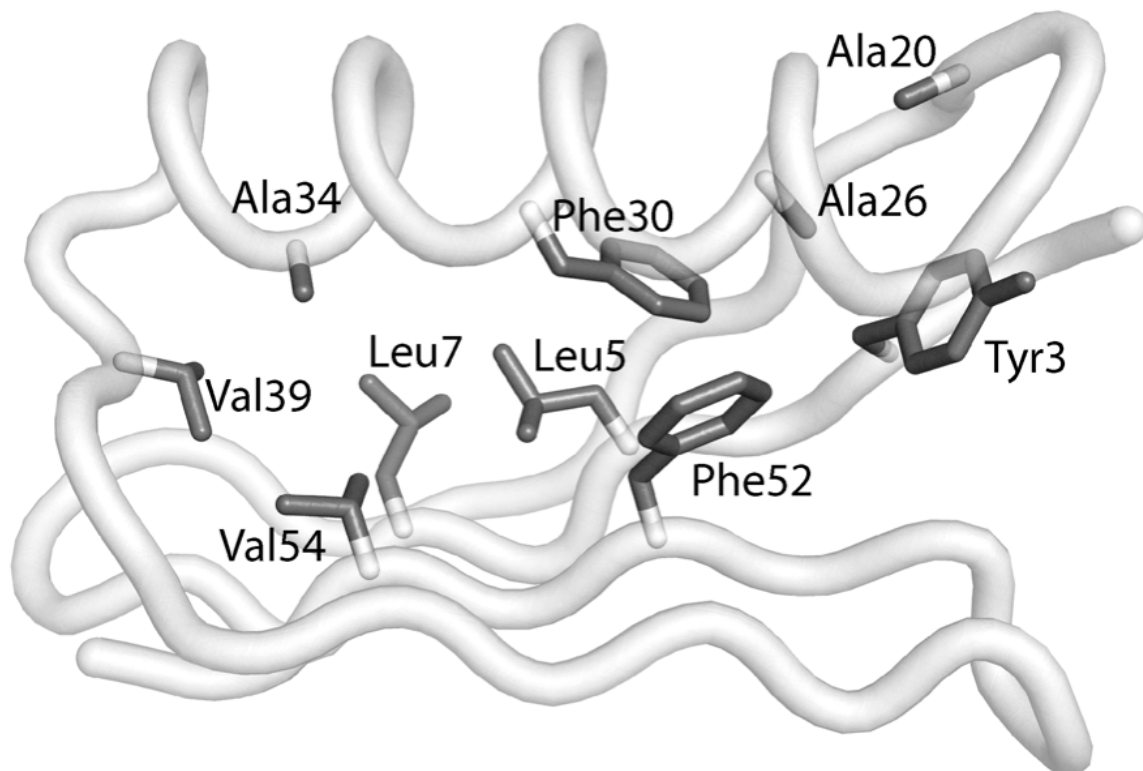


Figure 2-1: The core residues of Gβ1 designed in this study. Each of these positions was allowed to assume various rotamers of the hydrophobic amino acids Ala, Val, Ile, Leu, Phe, Tyr, and Trp. Position Trp43 (not shown) was additionally allowed to change rotamer but not amino acid type. All other side chains and the main chain were fixed in the input conformation for the state being modeled in each case.

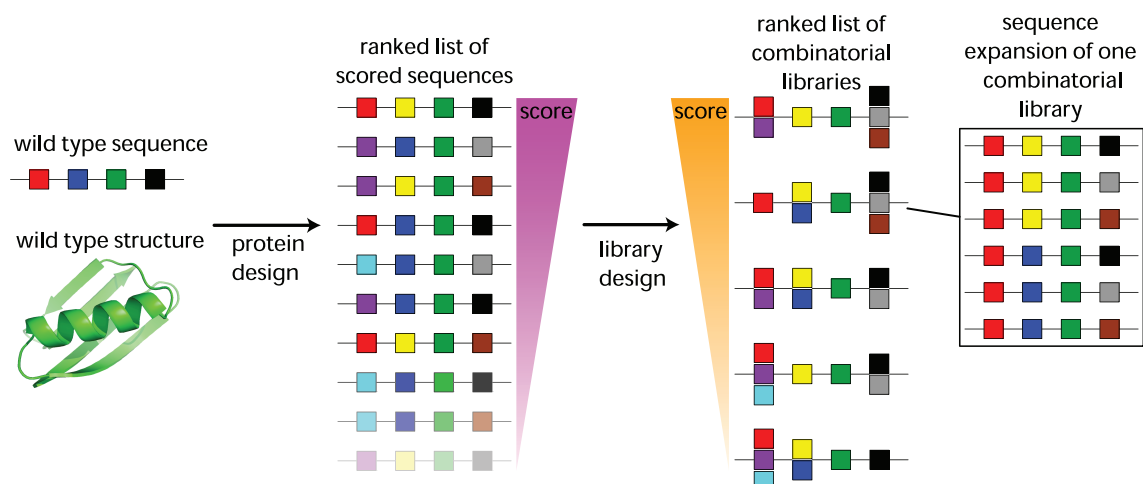


Figure 2-2: General scheme used to design combinatorial mutation libraries based on computational protein design calculations. A line of boxes indicates a protein sequence; each box represents a position in the protein chain. Different colored boxes represent different amino acids. The set of sequences on the far right corresponds to the expansion of a particular combinatorial library into the set of sequences it represents. The energies of the sequences in the expansions are used to decide which combinatorial library to test experimentally, as described in the Methods section.

Residue	WT	xtal-1	NMR-1	NMR-60	cMD-128	uMD-128
3	Y	F	F	F	F	F
5	L	IL	L	L	L	A
7	L	ILV	ILV	IL	FILV	FL
20	A	A	A	A	A	A
26	A	A	A	A	A	A
30	F	F	F	F	F	FIL
34	A	A	A	A	A	F
39	V	IV	IV	ILV	ILV	IL
52	F	F	FILV	FILV	F	F
54	V	IV	V	V	IV	AV

Table 2-1: Combinatorial libraries designed from different sources of structural information. **xtal-1:** library based on single-state design of the crystal structure. **NMR-1:** library based on single-state design of the constrained minimized average NMR solution structure. **NMR-60:** library based on multi-state design of the 60 member NMR structural ensemble. **cMD-128:** library based on multi-state design of the constrained molecular dynamics ensemble. **uMD-128:** library based on multi-state design of the unconstrained molecular dynamics ensemble

	xtal-1	NMR-1	NMR-60	cMD-128	uMD-128
number of top 20 list sequences found in library	8	12	10	8	1
number of top 100 list sequences found in library	15	20	16	16	3

Table 2-2: Library coverage. For each design problem, we report the number of top 20 and top 100 designed sequences from each original list that were represented in each corresponding combinatorial library. The maximum possible number of top 20 sequences that could be represented is 20, whereas the maximum number of top 100 sequences is 24 because each library contains only 24 members.

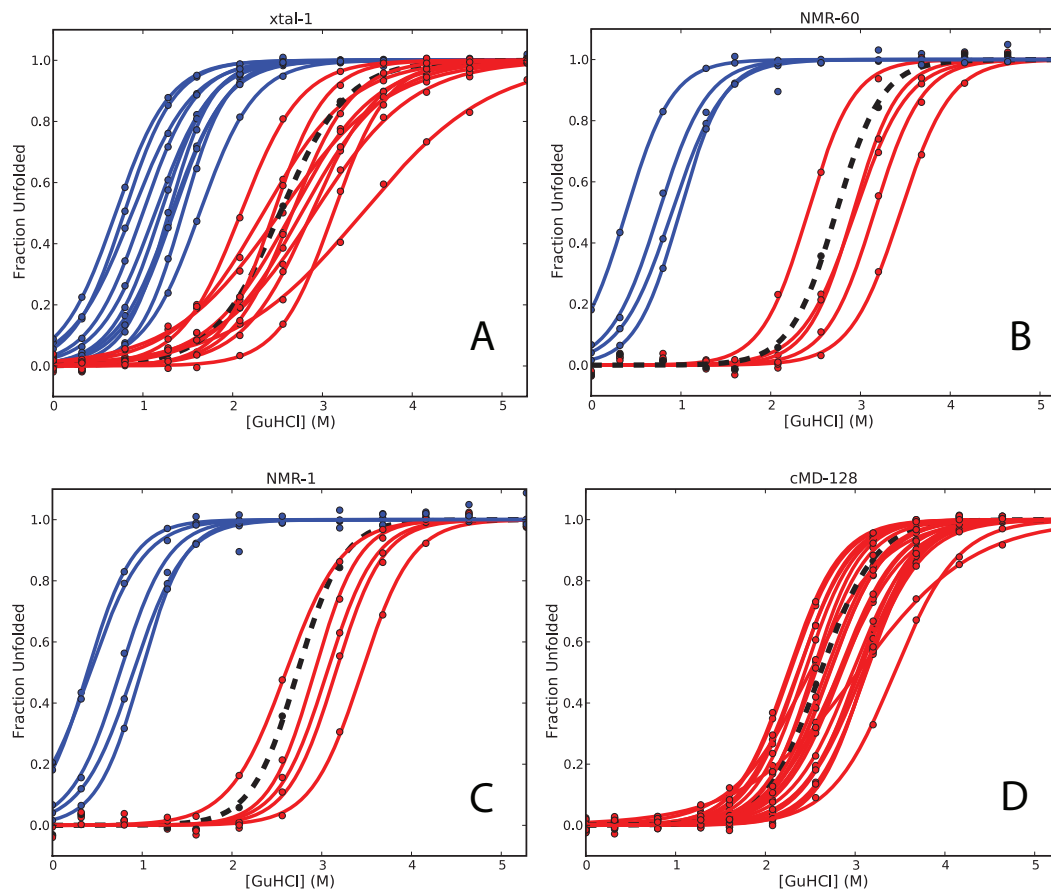


Figure 2-3: Fraction-unfolded curves derived from the stability determination of experimental libraries. The dashed black curve denotes variant Y3F, which is the closest library member to the wild type in terms of sequence, and which is known to have a stability very similar to the wild type. The blue curves denote variants with $C_m < 2.0$ M (“destabilized”) and the red curves denote variants with $C_m > 2.0$ M (“stabilized”). **(A)** xtal-1 library: Destabilized variants feature Leu at position 5 while stabilized variants feature Ile at position 5. Not pictured: variant Y3F+L5I+L7I, which did not give a signal that could be fit to a two-state unfolding model. **(B)** NMR-60 library: Stabilized variants feature Phe at position 52 while destabilized variants lack Phe52 but have Val at position 39. Not pictured: 14 variants that lack Phe at position 52 and which did not give a signal that could be fit to a two-state unfolding model. **(C)** NMR-1 library: Stabilized variants feature Phe at position 52 while destabilized variants lack Phe52 but have Val at position 39. Not pictured: 13 variants that lack Phe at position 52 and which did not give a signal that could be fit to a two-state unfolding model. **(D)** cMD-128 library: Only stabilized variants are present in this library.

Residue	WT	NMR-16D	NMR-16R	cMD-16D	cMD-16R	uMD-16D	uMD-16R
3	Y	F	F	FY	F	W	F
5	L	L	L	L	L	FLV	A
7	L	ILV	IL	IL	IL	I	FL
20	A	A	A	A	A	F	A
26	A	A	A	A	A	A	A
30	F	FILV	F	F	F	FILV	FIL
34	A	A	A	A	A	A	F
39	V	I	IL	ILV	ILV	IV	IL
52	F	FL	FL	F	FL	F	F
54	V	V	ILV	IV	IV	V	AV

Table 2-3: Combinatorial libraries designed from the top 16 energy-ranked structures based on two different energy functions. **NMR-16D**: library based on the top 16 NMR structures ranked by DREIDING energy. **NMR-16R**: library based on the top 16 NMR structures ranked by Rosetta energy. **cMD-16D**: library based on the top 16 constrained MD ensemble structures ranked by DREIDING energy. **cMD-16R**: library based on the top 16 constrained MD ensemble structures ranked by Rosetta energy. **uMD-16D**: library based on the top 16 unconstrained MD ensemble structures ranked by DREIDING energy. **uMD-16R**: library based on the top 16 unconstrained MD ensemble structures ranked by Rosetta energy

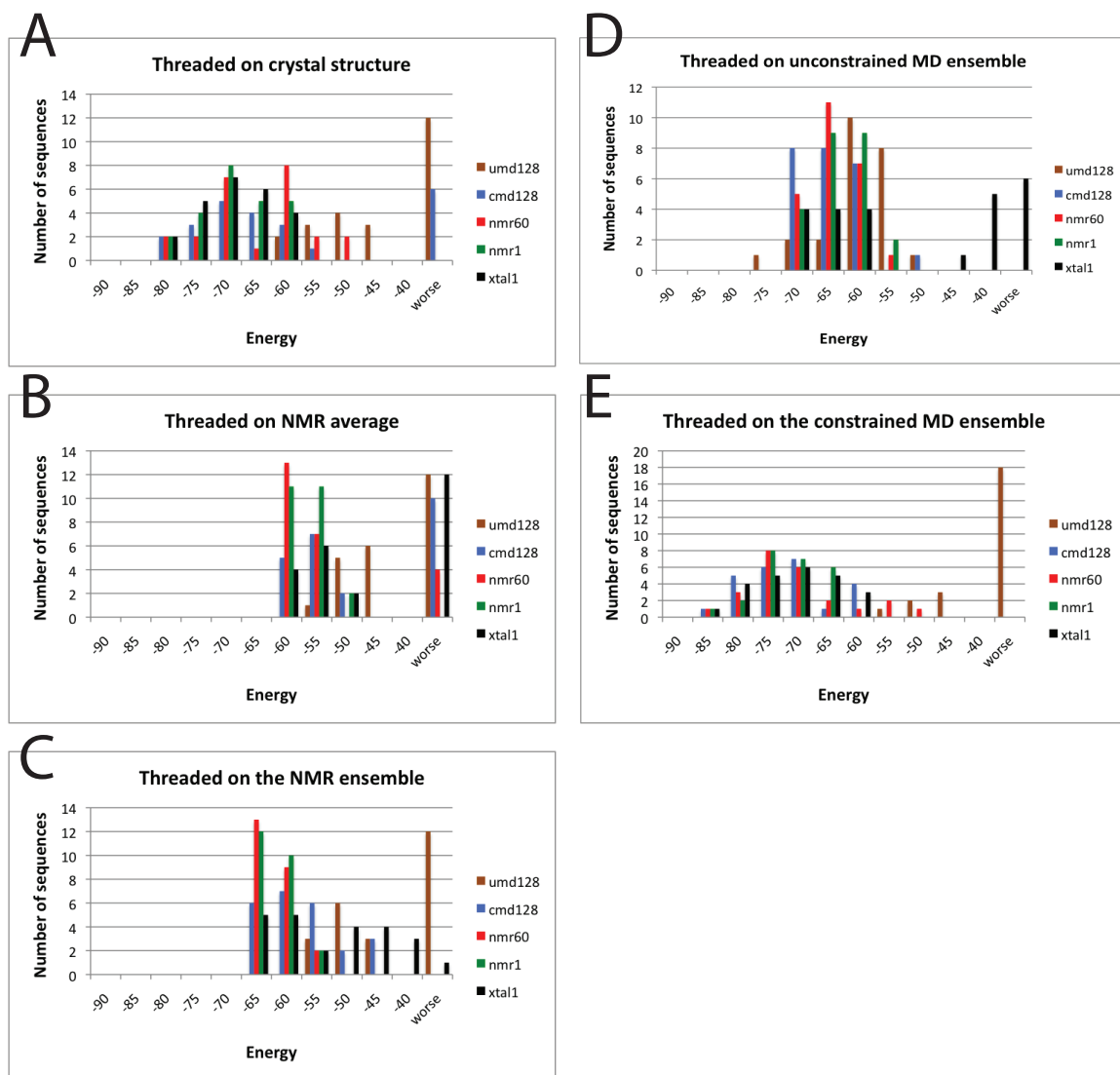


Figure 2-5: Library member energies. Energies of the members of each library when threaded on the structural basis for (A) the xtal-1 library, (B) the NMR-1 library, (C) the NMR-60 library, (D) the cMD-128 library, and (E) the uMD-128 library

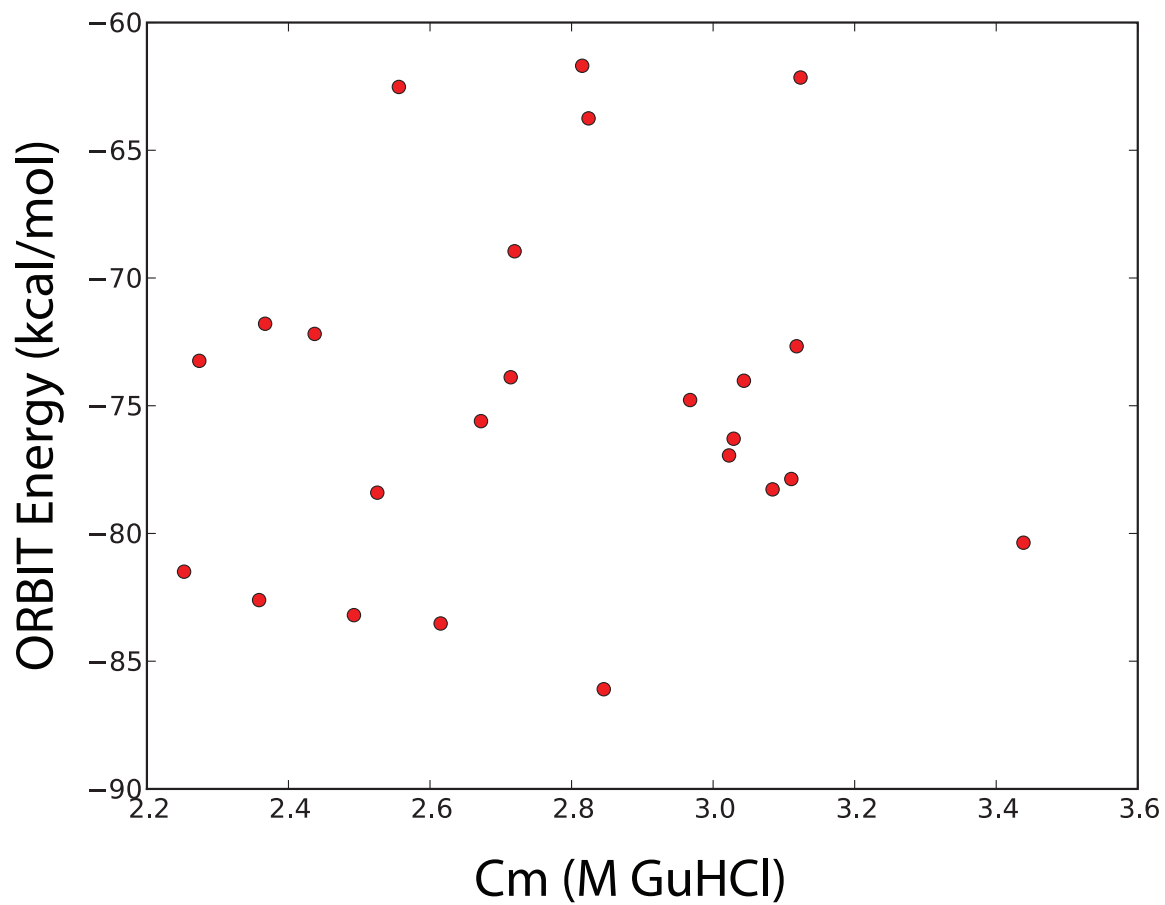


Figure 2-6: Correlation between simulation energy and experimental stability for the cMD-128 library. No correlation was observed between the experimentally measured fitness of the sequences and simulation energies that were used to select them for experimental screening.

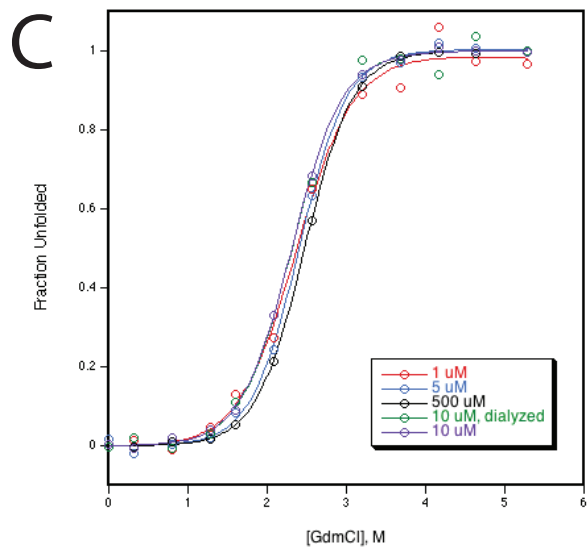
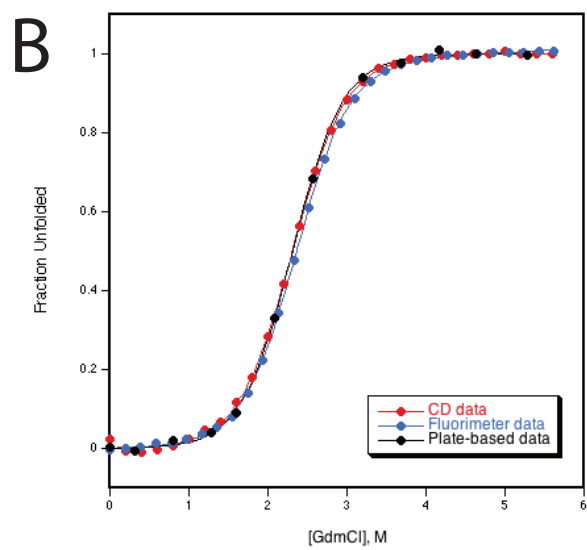
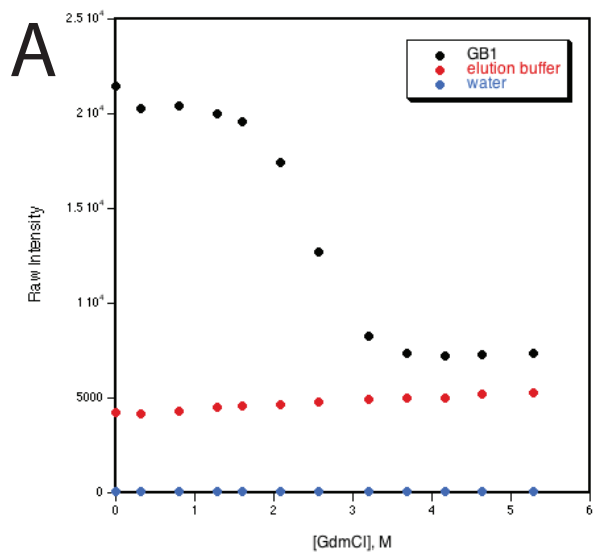


Figure 2-7: Microtiter plate-based stability assay controls. (A) Denaturation gradient and elution buffer fluorescence profiles. G β 1 (black) was expressed in a 5 mL culture, purified, and eluted with 500 μ L of elution buffer (50 μ M NaPO $_4$, 300 mM NaCl, 250 mM imidazole, pH 8). Since each point of the G β 1 denaturation profile contains 35 μ L of eluted protein, the elution buffer profile (red) substitutes protein with 35 μ L of elution buffer. Similarly, the water profile (blue) adds 35 μ L of water to make up the final volume. Each denaturation profile contains an increasing gradient of GdmCl, 50 μ M NaPO $_4$ buffer at pH 6.5, and water. (B) Fraction-unfolded profiles between different modes of detection. CD data (red) measured 5 μ M G β 1 titrated with a 5 μ M G β 1/8 M GdmCl solution in 0.2 M steps at 218 nm. Fluorimeter data (blue) measured 5 μ M G β 1 titrated as in the CD experiment with excitation performed at 295 nm and emission recorded at 341 nm with 4 nm bandwidths. Plate-based data (black) measured 12 separate solutions of 10 μ M G β 1 in response to increasing amounts of 8 M GdmCl with fluorescence parameters identical to the fluorimeter data except for 10 nm bandwidths. All samples were measured at 25°C in 50 μ M NaPO $_4$ buffer at pH 6.5. (C) Fraction-unfolded profiles between different protein preparations. G β 1 was expressed in 100 mL cultures, purified, and diluted to 1, 5, 10, and 500 μ M in 50 μ M NaPO $_4$ buffer at pH 6.5. Another expression culture was dialyzed overnight (Pierce Biotechnology) after purification and diluted to 10 μ M in the same buffer. All measurements were taken on a fluorescence plate reader as described in the text.

**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 G β 1, a 56 residue beta-grasp (ubiquitin-like) fold, which entailed the construction of \sim 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 $\beta 1$ domain of Streptococcal protein G ($G\beta 1$) 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 $G\beta 1$'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 $G\beta 1$ (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 $G\beta 1$ 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 ($T_m = (64.9 + 41 \times (\text{number of gc bp}) - 16.4) / (\text{number of total bp})$) (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 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 G β 1 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 G β 1 domain. This information on the tertiary structure of the protein not only gives thermodynamic details of stability (free energy of unfolding, $\Delta G(\text{H}_2\text{O})$; 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 $\text{dG}(\text{H}_2\text{O})$ 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 $\Delta\Delta G$ -true and $\Delta\Delta G$ -mAVG as seen in Figure 3-5C ($r = 0.89$), supports the application of the $\Delta\Delta G$ -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 G β 1 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 G β 1 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 G β 1 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 (G β 1). 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 G β 1, 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 G β 1 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 μ M 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) $\frac{1}{4}$ 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 troubleshoot 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, Cramer 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.

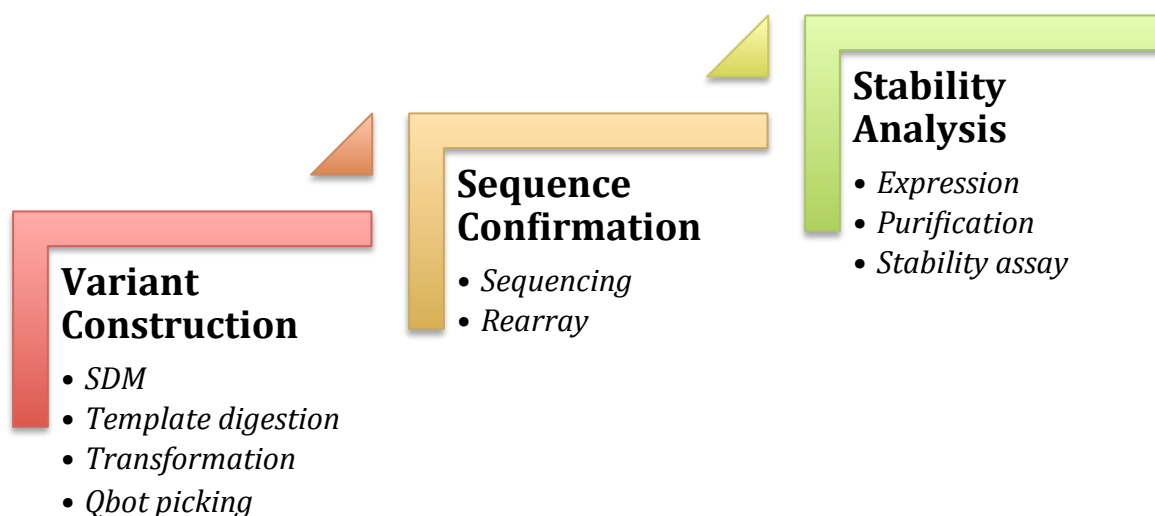


Figure 3-1: The automated site-directed mutagenesis pipeline. The methodology is composed of nine modular protocols that can be grouped into three blocks. The pipeline leads off with variant construction, a block of procedures that takes one mutagenic oligonucleotide per construct and ends with eight 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.

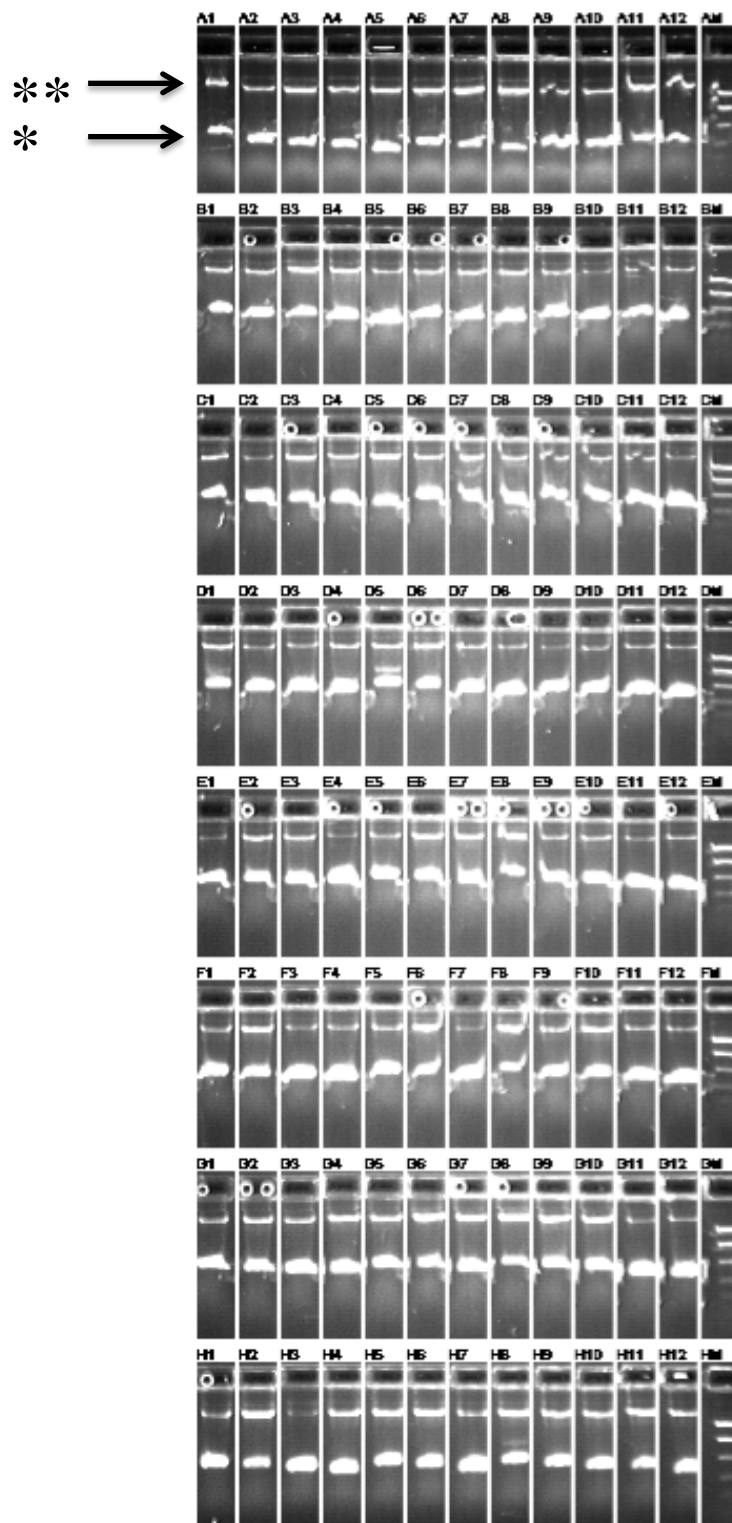


Figure 3-2: Visualization of a 96 well plate of SDM products. Agarose gel electrophoresis of DNA, by E-Gel 96, shows first- (*) and second-step (**) products from the megaprimer method for site-directed mutagenesis.

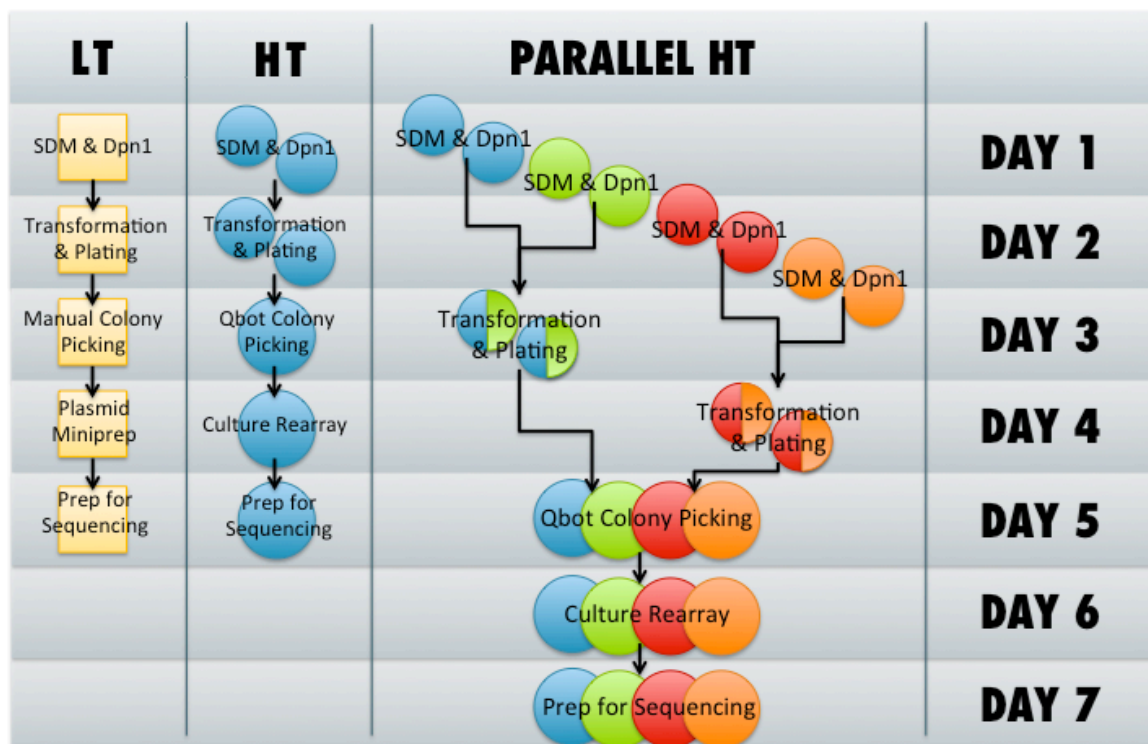


Figure 3-3: Variant construction timeline. Standard bench-top low-throughput (LT) methods are compared against the automated high-throughput (HT) methodology. Although both methods take five days before sending samples for commercial sequencing, the HT method processes 20-fold more reactions. For larger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

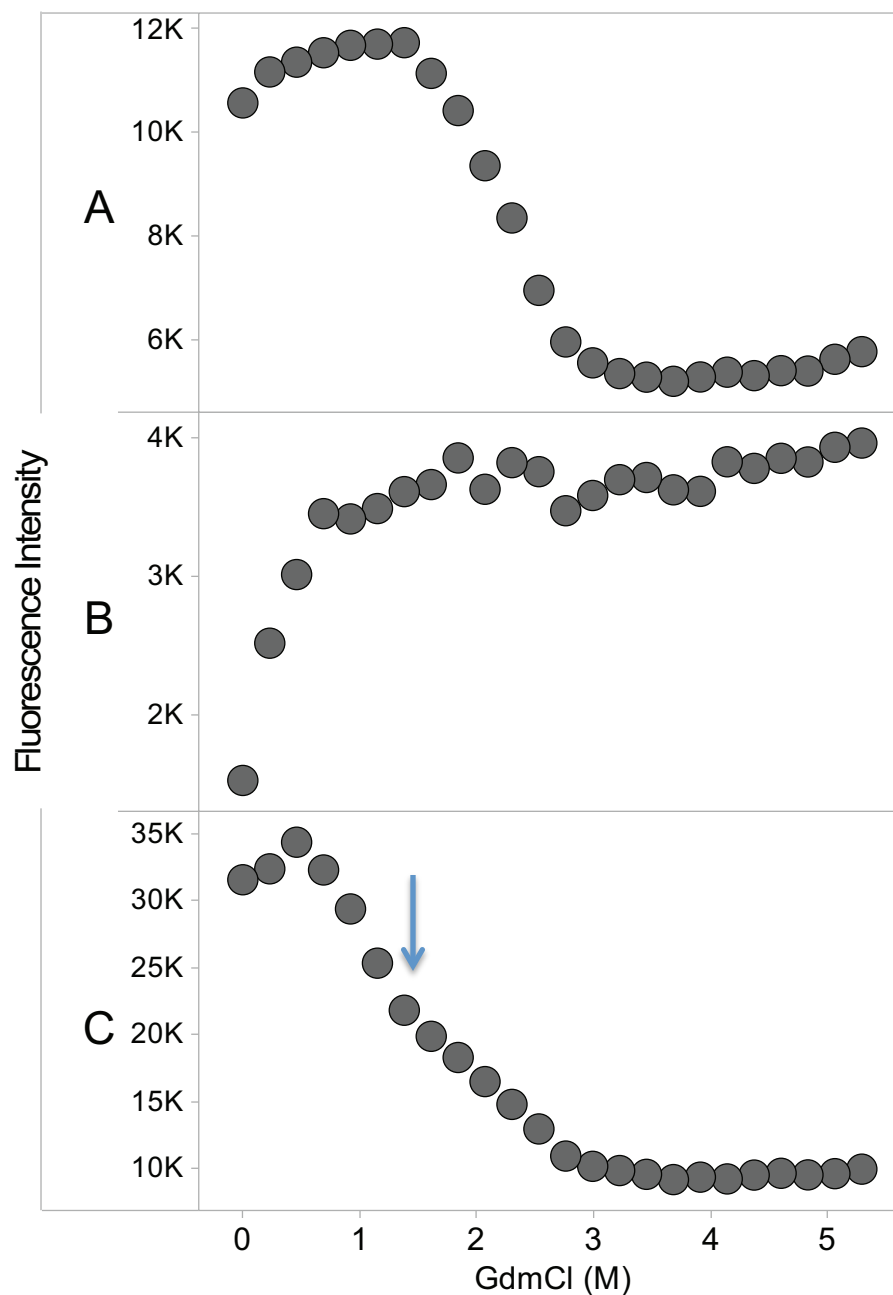


Figure 3-4: Potential protein unfolding curves. Tryptophan fluorescence data, plotted against a guanidinium chloride (GdmCl) gradient, for three examples of data from the high-throughput stability assay. A quality protein, with substantial pre- and post-transition baselines flanking a smooth transition 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.

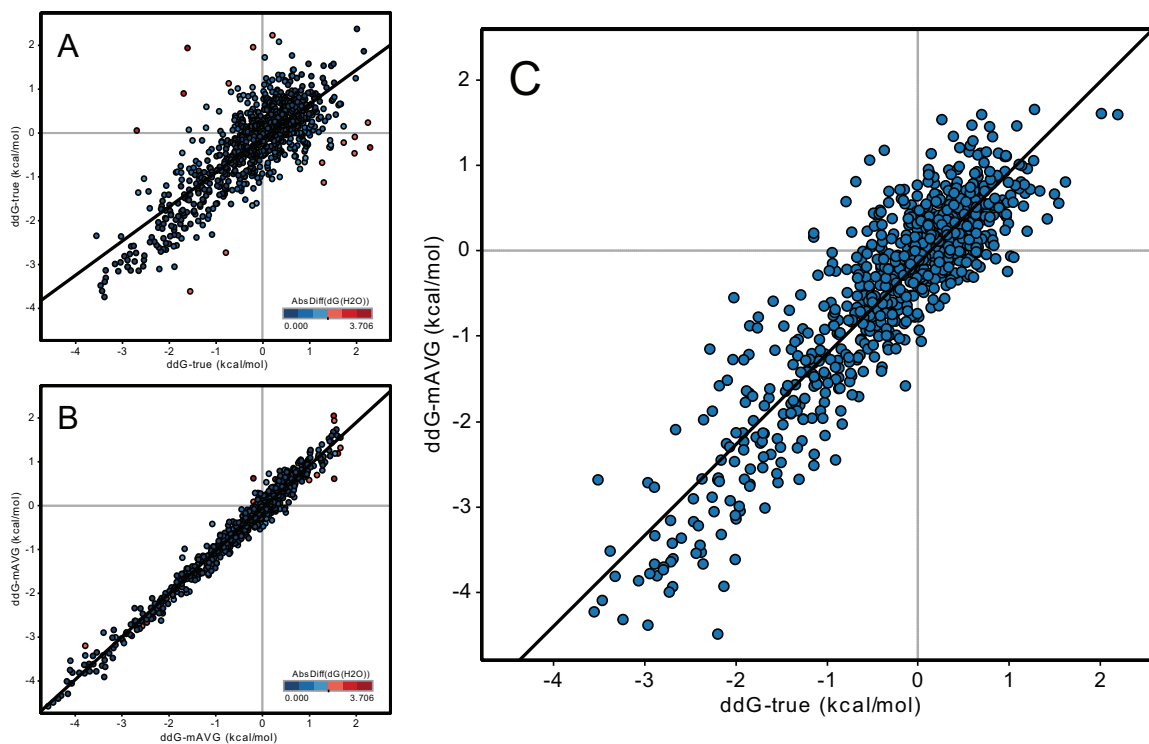
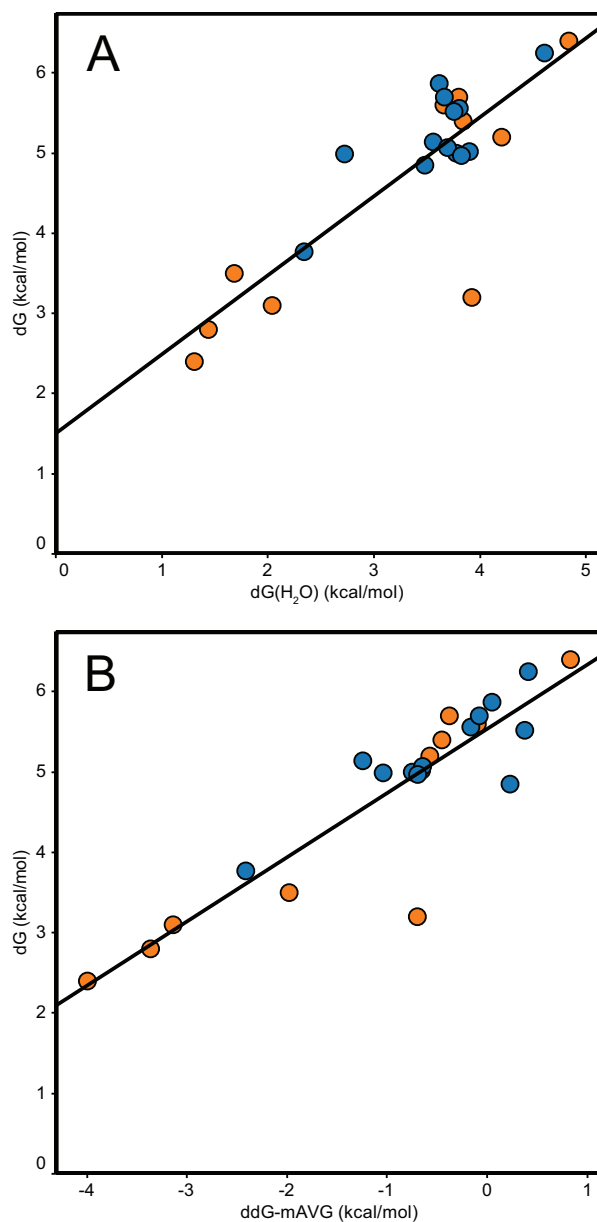


Figure 3-5: Precision among experimental measures of protein stability. Duplicate measurements of $ddG\text{-true}$ (A) and $ddG\text{-mAVG}$ (B) from the complete mutagenesis of G β 1 were correlated against each other. A strong linear relationship exists between $ddG\text{-true}$ and $ddG\text{-mAVG}$ (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.



Reference

- BIOCHEMISTRY 33, 5510-5517 (1994)
- PROTEIN ENG DES SEL 19, 285-289 (2006)

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(H₂O) 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.

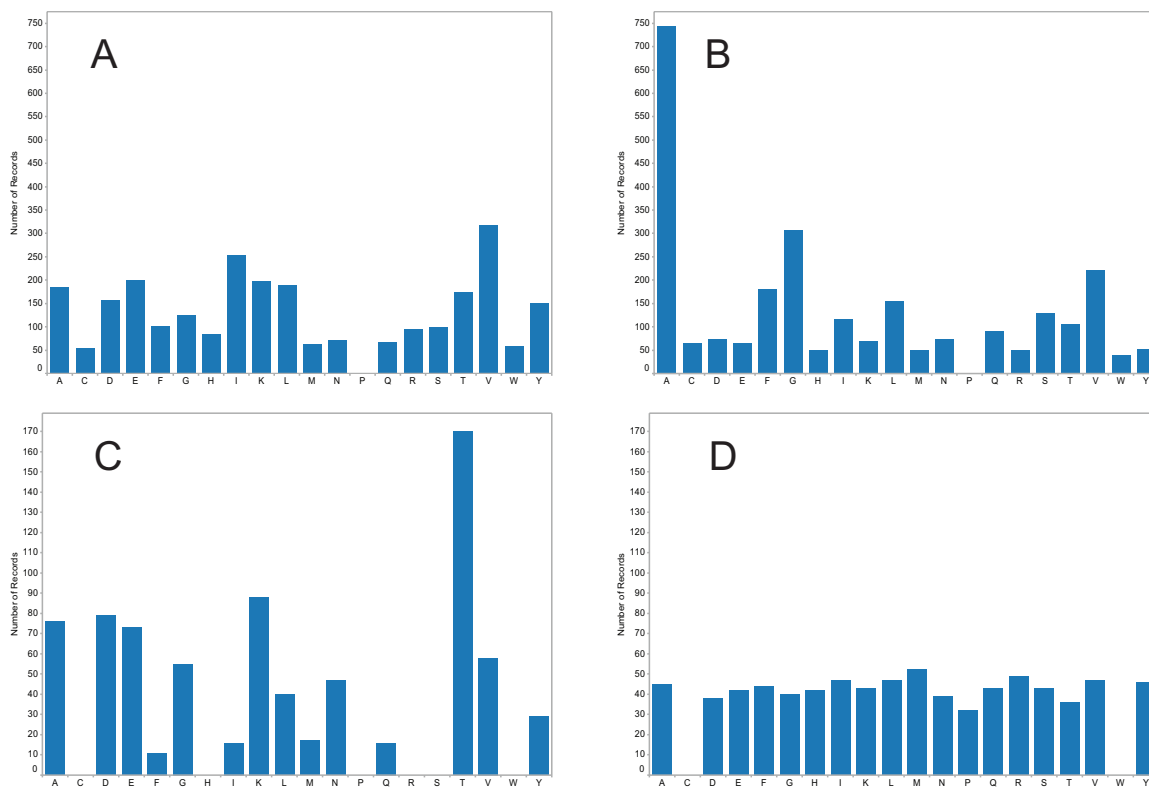


Figure 3-7: Point mutant amino acid distributions. The wild-type amino acid (A) and mutated amino acid (B) distributions are shown for a training set of 2,649 data points (collected from Protherm) used in developing PopMusic 2.0, a protein stability prediction algorithm. The wild-type (C) and mutated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

**Stability analysis of the complete single mutant library of a
protein domain**

Chapter 4

Adapted from a manuscript coauthored with Ernest Lee and Stephen L. Mayo.

Abstract

Proper understanding and prediction of the fitness consequences upon sequence mutation is an outstanding challenge in protein engineering. Although mutational data traditionally has been difficult and expensive to acquire, recent advances in laboratory automation have enabled the thermodynamic evaluation of almost every single mutant in a small 56-residue protein. With a domain-level perspective, we explore mutational outcomes, distributions, positional sensitivity, and mutant amino acid tolerance. Benchmarking current stability prediction methods reveals unbalanced performance across different structural criteria, but a uniform capability to capture the trends of our unbiased dataset. The surprising neutrality of single mutations to the domain contrasts with the stark negative epistasis seen in small, heavily mutated datasets, especially in variants comprised solely of individually stabilizing mutations. The poor predictability of multiple mutants from single mutations indicates that the field must move beyond single global stability measurements in order to truly comprehend the mutational effects of proteins.

Introduction

Protein mutagenesis data has long provided insights into the forces responsible for protein stability and folding (1–5). The protein-engineering field flourished from the ability to test hypotheses by comparing the thermodynamic effect of single- and multiple-amino acid mutations against a wild-type sequence. Abstraction of these concepts into computationally tractable algorithms have pushed the field even further, allowing users to score near or distant mutant sequences (6–9). These tools have also made feasible the computational probing of the mechanisms surrounding domain mutational tolerance and evolvability (10–13).

Despite this success, the principles of stability engineering describe only the general trend of the effects of amino acid mutations. Results are often mixed when applied to any specific problem due to the number of possible secondary and tertiary environments found in proteins (1, 4). A typical solution to this issue of proper context is the acquisition of more and more data in your protein of interest. This was a daunting task in the past, as the efforts required to engineer and purify protein mutants scaled linearly with the number to be made. Recently, technological advancements in laboratory automation and next-generation sequencing (14–16) have lowered this barrier so that the construction effort is identical for orders of magnitude more variants. Here, we analyze thermodynamic data from almost every single-mutant of an entire protein domain obtained through a previous effort in laboratory automation. As experimental data of this magnitude will only become more common, we examine the general utility of large datasets, and explore the performance of today's scoring algorithms.

Rare in the literature due to its laborious nature, domain-level mutagenesis data can provide valuable insight on mutational distributions and average positional and amino acid effects in proteins. Early work on mutational effects was performed on globular proteins such as the globins, lac repressor, lysozyme, staphylococcal nuclease, and barnase. Numbers on the fraction of mutations experimentally found stabilizing or destabilizing varied with each report, suggesting that the distribution of stability effects was unique to each protein. Although a convincing universal distribution for globular proteins was recently reported, most of the evidence was computationally derived (17). Conclusions on positional and amino acid effects from the pioneering studies centered on the importance of core and surface patterning of polar and nonpolar amino acids and secondary structure propensities. But what dominates or decides the sensitivity of a position? And which amino acid is best tolerated by the protein of interest? Our domain-level perspective of single mutant data simultaneously verifies the nature of mutational distributions and illuminates topics on positional sensitivity and amino acid tolerance.

Structure-based stability prediction algorithms translate our general knowledge of non-covalent protein interactions into a context-sensitive output. The degree to which they succeed is entirely dependent upon the constitution of the test dataset. A long necessary independent analysis of prediction algorithms concluded that all of the tested methods did not perform as well as previously reported and frequently failed to capture details (18). The curious fact that the algorithms had previous success in predicting mutational effects was reconciled with their ability to recapitulate general trends in the independent test set. Unfortunately, the test dataset used by Potapov and colleagues was biased towards large-to-small amino acid mutations, a feature very common to datasets

extracted from the popular online aggregate database of stability data, Protherm (19). As most of the current algorithms were trained on datasets sampled from Protherm, their true capabilities may be more underwhelming than reported. Our comprehensive single-mutant library provides a truly unbiased dataset upon which to test the performance of current and future prediction algorithms.

Experimental data on the stability landscape of a protein elicits inquiry into additivity effects and the domain's mutational robustness. Recent work on these topics has put forth that proteins thermodynamically stabilized from wild type or under weak functional selection exhibit a threshold against deleterious mutations that, once exhausted, declines rapidly (11, 12). The larger than expected effect of detrimental mutations after the threshold defines the system as negatively epistatic. We weigh in on this theory with real datasets and, alternatively, determine what effects the combination of multiple *beneficial* mutations might bestow upon a protein. The literature contends that all manner of simple and complex effects can exist under this scenario (20). Does the knowledge of every favorable single mutation enable the engineering of hyper-stabilized proteins?

Using a streamlined laboratory automation method, we constructed 935 single-mutants of the small monomeric domain G β 1, multiplying 55 of 56 positions by 17 of 19 possible mutant amino acids. Variants were purified and assayed for thermodynamic stability by observing tryptophan fluorescence in response to chemical denaturation. Rather than engage in the details of individual mutations, we chose to explore an array of topics important to protein engineering. We experimentally verify previous conceptions about the distribution of mutational effects for an entire protein domain, as well as

describe novel trends previously unexplored due to the lack of data. The unbiased nature of the dataset provided us a terrific avenue upon which to test popular stability prediction algorithms, as well as to guide the mutagenesis of multi-point mutants aimed to explore the epistatic effects of favorable mutations.

Results and discussion

The mutational distribution of G β 1

The overall distribution of $\Delta\Delta G$ effects in the G β 1 domain is shown in Figure 4-1. As a significant fraction of mutations lead to severely destabilized or insoluble proteins, it is not a normal distribution. Thus, we report an overall median $\Delta\Delta G$ of -0.25 kcal/mol with an interquartile range of 1.88 kcal/mol. If neutral mutations are those with a $\Delta\Delta G$ of ± 0.5 kcal/mol then the fractions of positive, neutral, and negative mutations are 16%, 41%, and 43%, respectively. While technically these values indicate that most mutations are destabilizing, 57% of mutations (positive + neutral) would have at worst almost no effect on protein stability. Roughly 12% of mutations tested could not be accurately measured by our stability assay (“unf” in the mutational distribution) and likely represent evolutionary dead ends.

This data illustrates that across the entire domain, the native sequence is robust to point mutations. This topic has seen much debate in the literature with evidence in favor for and against mutational robustness. Discrepancies likely arise (beyond those due to measuring different proteins) from varying stringencies in functional selection and by incorporating only the residues allowed through amber codon suppression. Our stability data likely represents the upper limits of robustness, as layering an activity requirement

upon sequences will lower the yield of neutral and positive mutational outcomes. Omitting unbiased tryptophan and cysteine incorporation, although necessary for the fidelity of the stability assay, also likely skews the fractional outcomes toward mutational tolerance. Finally, because the G β 1 domain is small (56 residues), its ratio of surface-to-buried positions fosters a tolerant distribution that may not be the case for larger proteins with larger cores. Previous measures of a mutation's functional inactivation probability across a protein domain are divergent, with values from 5% (21) to 34% (10). Again, differences in protein identity, selection stringency, and incorporated residues are likely the answer to these inconsistencies. In G β 1's case, having 88% of single mutations available for mutation presents an enormous amount of "safe" potential evolutionary trajectories for stabilization. This finding is somewhat surprising considering the relatively low thermodynamic threshold of G β 1 (~ 4–5 kcal/mol) in comparison to other proteins (up to 15 kcal/mol). Of course, the interactions between subsequent mutations will ultimately decide the feasibility of any particular path, and will likely shrink the potential complexity (22).

Although it appears intuitive that the mutational distributions of any two unrelated monomeric proteins will differ, this was found not to be the case in a recent computational study (17). Using the FoldX algorithm, the authors computed mutational distributions for a large panel of proteins and show that they all follow a similar asymmetric distribution. While future experimental efforts will be able to support or refute the full finding, we can currently verify the presence of the universal distribution in our dataset and confirm the FoldX algorithm's suitability for this type of study. We fitted our dataset (excluding those mutations labeled unfolded) to the bi-Gaussian and

individual core/surface distributions described in (17). In addition, we compare the derived means and variances to those from an identical dataset produced by FoldX3 (Table 4-1). Both sets of data fit the Gaussian equations well and exhibit the overall universal trend of sharper, stabilizing surface distributions with wider, destabilizing core distributions (Figure 4-2). In fact, the exquisite agreements between the calculated and experimental standard deviations provide noteworthy support for a Gaussian description of the surface and core mutational distributions. A two-population t-test to determine the similarity between the calculated and experimental individual Gaussian means finds the core distributions statistically identical ($\alpha = 0.025$, $p = 0.30$) and the surface distributions different ($\alpha = 0.025$, $p = 9.2 \times 10^{-13}$). This discrepancy can be explained by the historically poor treatment of surface energetics in stability algorithms along with small margin of error due to the tighter distribution. As shown in the literature (17, 18, 23) and later in our analysis, it appears that algorithms like FoldX3 can fail to capture specific details but still produce the correct experimental trend.

Positional sensitivity and mutant amino acid tolerance

Efficient visualization of each individual mutant in the library is accomplished through the use of a heat map (Figure 4-3). This perspective makes it abundantly clear that sequence position, not identity of the incorporated mutant amino acid, dominates mutational effects. This phenomenon is supported by the mutational sensitivity of wild-type non-polar amino acids that contribute to the hydrophobic core of the protein. Box plots of the stability effects separated by RESCLASS (Figure 4-4), an algorithm that uses the geometry from a crystal structure to designate core, boundary, or surface positions,

illustrates this behavior extremely well. The 75th percentile of the core mutation distribution sits below the inter-quartile range of the other categories, illustrating the intolerance of core positions to mutation. Only amino acid mutations to proline or glycine serve as exceptions to this result as they are generally deleterious regardless of the position due to their unique phi psi distributions.

Although we know that random core mutations are deleterious to protein stability, what quantitatively determines positional sensitivity? We approached this question by using supervised classification on a large number of attributes to train a linear regression model to predict the average $\Delta\Delta G$ of each position in the G β 1 domain. The best model gives a correlation coefficient of 0.83 and includes weights from a measure of the hydrophobicity of the wild-type amino acid and RESCLASS categories. However, the major contributor to the model was occluded surface packing value (OSP), which alone gives a correlation coefficient of 0.78 (Figure 4-5). As a metric for protein packing, OSP is routinely used to analyze structural datasets and protein folding predictions (24). That its found to be the chief determinant of domain-wide positional sensitivity is not unrealistic, as the result extends previous work on core mutations (25, 26) and makes intuitive sense: heavily occluded amino acids would be less likely to accommodate disruptions in their packing environments. The major outliers to this correlation are Gly41 and Tyr45, whose average $\Delta\Delta G$ values are greatly destabilized in comparison to their OSP. Position 41's sensitivity can be explained by its proximity to Trp43, the reporter amino acid for the stability assay, which will severely limit the allowed mutation types (small or flexible). And while position 45 is located on an outer beta strand, it is only one of two hydrophobic amino acids responsible (the other being Trp43) for

shielding an edge of the protein's core from solvent. Overall, OSP does a superb job in identifying the most sensitive positions to mutation and should be a part of the protein engineer's toolbox.

Complementary to the analysis on positional sensitivity is determining which amino acid scan of the domain best captures the average $\Delta\Delta G$ for each position. Actual experimental data on the system can help alleviate the issues with complex environments that protein packing alone predicted poorly. Whereas alanine mutagenesis is most often used to derive functional hotspots, it is unclear which amino acid can best forecast overall destabilizing, neutral, or stabilizing sites. The result of both a $\Delta\Delta G$ deviation method and a linear ranking method (see Methods) show serine as the highest-ranking amino acid, with a mixture of methionine, threonine, and glutamine rounding out the top four (Figure 4-6). All of these amino acids are non-charged, polar, and fairly amphiphilic in nature, making them reasonable choices for an amino acid stability scan.

If the structure of the protein of interest is available, then adjusting the scan by RESCLASS would likely lead to higher prediction accuracy. In our study, methionine and alanine dominated the core rankings, while threonine, serine, and glutamine topped the rankings for boundary and surface positions. These results again show a preference for uncharged amphiphilic amino acids, along with moderate "like dissolves like" tendencies for core and boundary/surface predictions. Alanine ends up performing respectably well across the core and boundary segments of the protein, although it is a decidedly poor indicator of positional sensitivity on the surface, where the majority of other amino acids outcompete it. In total, the two deviation calculations give similar

results, unifying and strengthening the evidence for small amphiphilic residues such as serine as the first choice for stability scanning mutagenesis.

Instead of averaging by position, we now average by mutant amino acid and ask, which amino acid is best or least tolerated by the domain, and why? Because reasonable solutions to this query require a comprehensive number of mutations per position, our dataset is uniquely positioned to explore this topic. The worst amino acid for general incorporation is proline, followed by glycine (Figure 4-7). This isn't surprising, as the special amino acids are well known to be debilitating to protein stability. Aspartic acid is the third worst incorporated amino acid, most likely due to its highly acidic nature. It also contains the smallest amount of nonpolar atoms in comparison to the other charged amino acids, strengthening its relative charge and snowballing its destabilizing nature. At the other end of the spectrum are the hydrophobic amino acids, and in particular, the large aromatics tyrosine and phenylalanine. Why are these tolerated so well on a domain that, due to its small size, features a much larger surface-to-core ratio than most other proteins?

Close observation of the data in Figure 4-7 shows that among the functionally identical amino acid pairs (D/E and N/Q), the residue carrying an extra methylene was tolerated better across the protein. This, coupled with a high surface-to-core ratio when compared to the average protein, suggested that the G β 1 domain may be unique in its accommodation of hydrophobic mutations, perhaps in an effort to bury more hydrophobic surface area. Since average mutant amino acid data is recapitulated well by Popmusic2 (Figure 4-8) we investigated our hypothesis by calculating the systematic scan of four other proteins, all larger than G β 1, with the Popmusic2 web server. The proteins selected

isolated the effects size, secondary structure composition, and packing density might have on amino acid tolerance (Table 4-2). Amazingly, all four had very similar average mutant amino acid rankings, essentially duplicating the experimental G β 1 results (Figure 4-8). The tolerance to large hydrophobic amino acids across a domain appears to be a general feature of soluble globular proteins.

Native proteins feature very modest amounts of solvent-exposed hydrophobic residues due to the possibility of alternative folded states that better bury the nonpolar surface area. Yet, the very presence of modest amounts of surface hydrophobic residues indicates that the physical mechanism underlying this behavior has some buffer preventing disastrous aggregated outcomes. Single incorporations of nonpolar residues are not likely to alter the native conformation, and as seen by the data, can stabilize the fold. Previous experimental work on staphylococcal nuclease supported the notion that this “reverse hydrophobic effect” is almost nonexistent across single mutations (27). Interestingly, they note that fully exposed positions better tolerate aromatic incorporation than partially buried sites, arguing that mutation sites are still susceptible to steric clashes and packing effects despite being close to the protein surface. We found similar results when the average stability effects of the top four tolerated amino acids (Phe, Tyr, Leu, Ile) in boundary and surface sites were broken down into two populations of packing density. However, when the data was broken down into quartiles, there is a bump in average stability in the partially exposed quartile, providing evidence for preferential packing between the incorporated hydrophobic amino acid and the nonpolar atoms of other native residues near the surface (Table 4-3). In sum, considering the hydrophobic mutability of protein cores (28, 29), the support for partially exposed hydrophobic

clusters, and that the native conformation is unlikely to change because of a single mutation on plastic protein surfaces, non-polar residues offer the best chance at making neutral or stabilizing interactions across a protein domain.

Stability prediction algorithm performance

Three popular prediction algorithms, Popmusic2, FoldX3, and Rosetta, were used to calculate the stability change of the 935 mutations in our domain mutagenesis dataset. Popmusic2 is a reduced-representation statistical energy function trained to recapitulate a large experimental dataset from the Protherm database. FoldX3 is similarly trained, but uses an empirically derived energy function mixed with weighted statistical terms. Rosetta mixes statistical potentials with an all-atom physical potential, and was trained to recover native sequence composition for protein design. Three versions of Rosetta are used, each with increasing amounts of backbone flexibility. The specific details and parameters used for each algorithm are described in the methods. Unfolded mutations for which only approximate data is available were filtered, leaving 825 mutations. Unrealistic predicted energies from the FoldX3 and Rosetta calculations prompted further filtering by removing mutations with abnormally high van der Waals clash or repulsive energies, respectively. Algorithm performance was evaluated by correlation coefficients (Table 4-4) and fraction correct % (Table 4-5). In addition, these metrics are reported for the datasets broken down by volume change, RESCLASS, and polarity change to assess performance by mutation type.

When asked to recapitulate energetic details of the full dataset the hybrid energy functions perform quite poorly, as the purely statistical Popmusic2 method led the pack

with a correlation coefficient of 0.56 (Table 4-4). After filtering mutations with large clashes, FoldX3 shows improved performance while the flexible backbone Rosetta methods achieve the best overall correlations to the dataset. Although full backbone minimization reduces the number of outliers due to repulsive clashes, it is outperformed by constrained minimization, even against our unbiased dataset (Chapter 3). The notion that too much backbone freedom may simultaneously hurt and help structure prediction of a mixed dataset (30) is upheld by our results. When asked to recover the fraction of positive, neutral, and negative mutations in the data (Table 4-5), all of the algorithms perform almost equally. This result speaks to the utility each method has in predicting the correct trend in large datasets (17, 30). That they all do so equally well is both reassuring to users in the field and frustrating to developers looking for avenues of improvement.

Mutations that remove volume ($-Vol\Delta$) are better predicted than those that add volume ($+Vol\Delta$) across all algorithms and both tables. The closer a mutant protein's conformation is to wild type, the better each prediction algorithm performs, as most are capable of only torsion preferences or rotameric flips. The methods that do introduce backbone flexibility perform better, but can be restricted by limited sampling of correct conformations (30). The preference for large-to-small mutations then implies limited structural rearrangements across the domain for this mutation type, a conclusion supported by work on T4 lysozyme (31). The overwhelming number of alanine and glycine mutants ($-Vol\Delta$) in the Protherm database may also partially explain the affinity Popmusic2 and FoldX3 have for this mutation type. A surprising detail is the continued advantage constrained minimization exhibits over unconstrained minimization, even

across the small-to-large mutations ($r = 0.56$ vs. $r = 0.52$) that are expected to introduce sizeable backbone rearrangements.

As a large determinant of mutational sensitivity, tertiary structure can be expected to play a role in algorithm performance. The overwhelming number of destabilizing mutations in the core inflates the fraction of mutations easily predicted (Table 4-5), while mutations closer to the surface are binned at $> 50\%$ accuracy. However, knowledge of a mutant's $\Delta\Delta G$ direction is no guarantee of correlation coefficient accuracy, as shown by the poor performance of core mutants in Table 4-4. The fact that destabilizing variants exist across a larger energetic range than more benign mutations likely promotes this inaccuracy in predicting buried positions.

Breaking down the data by polarity changes highlights the underlying principles that govern the prediction algorithms. The effective van der Waals potential in FoldX3 and Rosetta give these methods an advantage in predicting the core packing effects of nonpolar-to-nonpolar mutations (Table 4-4). Likewise, Rosetta's suboptimal treatment of buried electrostatics (nonpolar-to-polar, $r = 0.34$) and nonpolar exposure (polar-to-nonpolar, $r = 0.42$) is ameliorated by allowing backbone flexibility. No particular method excels in polar-to-polar mutations, likely a result of the lack of explicit solvent from any of the calculations. Differences between algorithm frameworks may be best embodied by comparing the core (nonpolar-to-nonpolar/polar) and surface (polar-to-nonpolar/polar) prediction accuracies of FoldX3 and Popmusic2. Despite very similar training sets, each method's competency lies in inverse structural environments: Core mutations come easier to FoldX3 due to its effective treatment of sterics, while surface mutations are

better captured by Popmusic2 because the statistical nature of its potential can implicitly capture complex multi-body effects.

The differences seen in Table 4-4 between the algorithms are smoothed when the stringency in prediction accuracy is lowered, as in Table 4-5. Only the poor performance in polar-to-polar mutations by FoldX3 and Rosetta is effectively reproduced from the previous metric. Popmusic2 performs admirably in predicting the fraction correct in all polarity change categories. This jack-of-all-trades quality likely stems from the fact that each statistical term is weighted by solvent accessibility, allowing it to grossly fractionate between debilitating core mutations and neutral surface mutations.

Any recommendation on the prediction algorithm of choice must be tempered by the type of question being asked. Queries concerning the specific and accurate stability of particular single-mutants would probably be best estimated by the constrained backbone minimization Rosetta protocol. However, attention should be paid to filter unreasonable repulsive energies and to consider that accuracy can drop with polar-to-polar mutations, buried positions, and mutations that add volume. If computational power is limited or the number of mutations is greater than 10^3 , Popmusic2 is significantly faster and is effective in predicting the trends in the data (Figure 4-8). The actual magnitude of each calculation should be viewed skeptically, unless the mutation involves only polar residues, in which case Popmusic2 performs better than any iteration of Rosetta. FoldX3 serves as the middle ground between the other methods in terms of both speed and accuracy. While observations from mutational trend studies in which FoldX was used are likely to be duplicated by other algorithms, specific values should be taken lightly, especially those involving polar surface mutations.

Additivity of multiple mutations

The wild-type sequence of the G β 1 domain is very tolerant to single mutations, as evidenced by the mutational outcome percentages and heat map distribution reported in the first section. Mutability trials, where single mutations are added until exhausting some threshold, reemphasize the observed leniency surface positions have over those in the core. These calculations assume each mutation is completely additive, an unrealistic assumption for proteins, although they do provide reference points for the mutational load of a perfectly additive system. If the threshold is 4.5 kcal/mol (the $\Delta G(\text{H}_2\text{O})$ of wild-type G β 1 is 4.04 ± 0.4 kcal/mol), then an average of 6.28 ± 4.9 random mutations over 1000 trajectories are needed to break the protein. Predictably, the required number of mutations increases as one progresses through the RESCLASS categories from 2.07 ± 1.1 mutations (core), to 10.27 ± 8.9 mutations (boundary), to a maximum of 17.46 ± 15.0 mutations on the surface. Given the stated evidence, one might assume that making multiple mutations on the surface of G β 1 would be more successful than mutating a similar number of positions in the core.

Data from previous efforts in individually designing the core and surface of G β 1, coupled with our comprehensive dataset, allows us to examine the relative performance of our design procedures as well as the additivity of mutations in different regions of the protein. The core mutant dataset is a compilation of libraries designed from different sources of structural diversity where only hydrophobic amino acids were allowed (23). Each mutant is 2–5 fold away from wild type, above the 2.1 random mutation reference mark, yet more than 80% of the dataset is thermodynamically neutral or better than wild

type (Figure 4-8). This result endorses the all-atom two-body energy function (similar to Rosetta) we used to predict these sequences for modeling hydrophobic core interactions. Rosetta's advantage in accurately predicting nonpolar-to-nonpolar single mutations (Table 4-4) supports this finding. The surface mutant dataset is a single library of mutants, designed using the same energy function as the core study, which aimed to improve overall stability through mutations to the β -sheet surface of the protein (unpublished results). No mutant carrying more than 1 mutation in this library was stabilized from wild type, and a mild inverse relationship exists between stability and the number of mutations (Figure 4-8). In addition, the observed number of mutations isn't remotely close to the random mutation reference point for this segment of the protein, reemphasizing the reduced capacity of physics-based energy functions to capture surface interactions.

The discrepancy in algorithm performance in different protein environments could also be attributed to the nature of additivity in each environment. When the core mutant data is plotted against the sum of individual $\Delta\Delta G$ stability values from our single mutant dataset, the linear trend line ($r = 0.86$) is jilted above the perfect additivity line ($y = x$), signifying that the mutants are destabilized compared to what the simple sum would predict (Figure 4-8). As the majority of mutations are stabilizing, one explanation may be that there is some limiting level of local stabilization that a protein core can reach before interactions elsewhere in the protein become more important globally (20). Conjecture aside, this plot serves as a reference to an identical chart featuring the surface mutant data (Figure 4-8). Here the trend line dramatically intercepts the perfect additivity line, demonstrating much more pronounced non-additivity than in the core mutants. The

particular number of mutations is also very important, as the slope of the trend line radically changes after introducing 3 mutations to the wild-type background, mirroring the largely simple additivity found in double-mutants from another large data study (14). Why is the additivity of multiple mutations so different between the structural environments?

One biophysical explanation would be that the capacity of surface residues to change conformation, along with the greater number of potential interaction partners (solvent molecules) as compared to the situation in the core, allows for the differences seen in mutational non-additivity. Interactions not modeled by traditional protein design software, such as buried waters and extensive hydrogen bond networks along the surface, could be so important to the enthalpy of the protein that they overcame any gain in entropy from joining the bulk solvent. Finally, a parallel can be made between our observations and recent work in linking fitness robustness and epistasis. Bershtein and coworkers describe a quantitative inverse correlation between higher tolerance to mutations and the level of negative epistasis (12). Enzymes (β -lactamase) tested under low fitness (ampicillin) levels exhibited higher degrees of negative epistasis; that is, the effect of mutations after exhausting some threshold level was greater than when under higher levels of fitness stringency. Similarly, surface positions, seemingly tolerant of most single mutations, display markedly stronger non-additivity than core positions. This finding extends the robustness-epistasis theory from describing global, random mutations to capturing particular tertiary-structure effects of proteins. Unfortunately, the ~ 100 multiple mutants in this combined study represent an incredibly small slice of the number

of potential two-, three-, four-, and fivefold mutants possible, and therefore this parallel may only exist for the variants examined here.

The surface mutations in the previous additivity study didn't include any greatly stabilizing single mutations, an element that may have had an effect on the observed epistasis. In fact, the Bershtein theory is reported to represent combined *deleterious* mutational effects. What is the result of combining multiple mutations that are all individually significantly *stabilizing*? Drawing from our comprehensive dataset, we constructed three variants (Table 4-6) by first selecting mutations that each stabilized the G β 1 domain by more than 1 kcal/mol. This led to 33 individual mutations over 16 positions, a great majority of which introduced large hydrophobic residues. We selected one mutation per position with an eye towards limiting the number of incorporated non-polar residues and combined them into a single variant (16-fold). The mutations were then visualized on the 1PGA structure and funneled down to 8 mutations, removing clustered, interacting residues (8-fold). Lastly, 2 more mutations were screened out due to their mutation of special residues, which tend to have entropic effects on protein stability (6-fold).

The three variants were expressed, purified, and assayed for stability in the same way as the entire single mutant dataset. The 16-fold mutant had no soluble expression as determined by the criteria set in Chapter 3. The 6 and 8-fold mutants expressed normally yet both displayed neutral $\Delta\Delta G$ stability values, despite featuring individual mutations that, when summed, should stabilize the protein by more than 8 and 10 kcal/mol, respectively (Table 4-6). Rationally combining multiple efficacious single mutations from our stability map proved not to be a successful avenue for protein stabilization.

Upon investigating the results, all but two of the mutations in the 16-fold mutant occurred in the boundary or surface of the G β 1 domain, and more than half of them involved a polar-to-nonpolar mutation. Introducing that many hydrophobic residues to the protein surface is thought to be dubious for aggregation reasons, unlike single nonpolar incorporations that can be stabilizing (1, 2, 32, 33). Anticipating the 16-fold mutant's issues, the 6- and 8-fold mutants were designed more conservatively, avoiding the insertion of potentially destabilizing interactions and evening the ratio of polar and nonpolar mutations. Mutated sites were selected for their three-dimensional distance from each other, thereby promoting conditions for perfect additivity (20, 34). The subsequent severe non-additivity encountered by the variants suggests very large unfolded state effects are at play. Whether these effects are due to the tertiary location, residue type, number of the mutations, or some combination of these attributes is unknown. Future work on these variants should attempt to definitely explain the source of non-additivity by tracing their potential mutational paths through sequence space.

Structure-based protein engineering and design seeks to modify the properties of proteins through the calculation of folded state energetics. Our analysis of the additivity of mutations in the G β 1 domain demonstrates the difficulty in identifying distant, stabilizing sequences without explicit consideration to the unfolded state. Despite a wild-type sequence experimentally determined to be accommodating of most single mutations, multi-fold variants became increasingly harder to predict, especially if the mutations occurred on the surface. Modern techniques minimize this deficiency in structure-based design by testing libraries of engineered sequences to spread the risk in selecting a mispredicted variant. Although this can be effective (23, 35, 36), novel, efficient, and

effective methods for the unfolded state, among other absent concerns and approximations, will help to usher in truly robust protein design. Concurrently, efforts to capture experimental data on more than just the global stability of mutant proteins deserve attention. For example, multiplexing thermodynamic stability with solubility, expression level, and proteolysis resistance would produce quality high-density datasets that will deepen our understanding of both *in vitro* and *in vivo* stability.

Conclusions

Our aggregate analysis of stability data on every single mutant in G β 1 provided the first experimental look at the mutational distribution of a protein domain. The rather tolerant nature of the protein, especially to hydrophobic residues, illustrates the plasticity of non-core residues and the heavy desire to bury nonpolar surface area. Atomic packing density linearly correlated with positional sensitivity, and scanning with serine, not alanine, served as the best experimental indicator of positional hotspots. The unbiased nature of the dataset provided an even playing field upon which to test popular stability-prediction algorithms. Although Rosetta was the clear performance leader, every method could satisfactorily recapitulate the general trends of the data. Upon examining the additivity of previous design efforts, we learned that non-additivity was prevalent throughout the protein, but especially on the surface. Attempts to utilize the data for rational stability engineering failed in the face of tremendous non-additivity.

Large mutagenesis datasets will only become more common with the maturation of automation technologies and next-generation sequencing. This incoming avalanche of data will provide breadth in complementing traditional in-depth analysis of important

systems. However, the volume of data current cutting-edge technologies can produce (10^5 variants, or all double mutants in a 50 aa protein) still pales in comparison to the enormity of potential sequence space. This is especially troubling considering that ours and another study show that double mutants are fairly additive (14), in contrast to what we see with variants three or more mutations away from wild type. Computational solutions exist that can traverse the ocean of potential sequences, but require high-density datasets from which to train new methods to properly evaluate proteins distant from the starting sequence.

Materials and methods

Dataset

Thermodynamic stability data from 935 single mutants of the $\beta 1$ domain of wild-type Streptococcal protein G ($G\beta 1$) was generated as described in Chapter 3 of this thesis. Briefly, each gene was constructed through laboratory automation and sequence verified. Every protein was expressed, purified, and its chemical melting point (C_m) determined by measuring tryptophan fluorescence in response to a 24-point GdmCl gradient. The $C_{m \text{ Mutant}}$ value was used to calculate $\Delta\Delta G$ from the following equation:

$$\Delta\Delta G = \bar{m} * (C_{m \text{ Mutant}} - C_{m \text{ WT}})$$

where \bar{m} is the average of the wild-type and mutant m -values (37), a parameter obtained from the linear extrapolation method (38) for determining changes in free energy. Using this equation, stabilizing mutations held positive values, while destabilizing mutations held negative values.

$\Delta\Delta G$ distribution fitting

The experimental $\Delta\Delta G$ data was binned into a histogram with 0.5 kcal/mol intervals. In order to match previous work, all stability data was multiplied by -1, making positive values represent destabilizing mutations and vice versa. We used the following Gaussian model to fit the individual core and surface distributions, with F the percent fraction and x the $\Delta\Delta G$ values (17):

$$F_s(x) = \frac{100}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right]$$

The bi-Gaussian fit model was given as a superposition of two Gaussians with different means and variances, and P_1 the fraction of the first Gaussian:

$$F_{bi}(x) = 100 \left\{ \frac{P_1}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{(x - \mu_1)^2}{2\sigma_1^2}\right] + \frac{1 - P_1}{\sqrt{2\pi\sigma_2^2}} \exp\left[-\frac{(x - \mu_2)^2}{2\sigma_2^2}\right] \right\}$$

If the fit is good, the mean and variance of the first Gaussian should correspond with the surface distribution, while the mean and variance of the second Gaussian should correspond with the core distribution. A two-sided t-test was conducted using the Mathematica HypothesisTesting package (Wolfram Research) in order to determine agreements between the means of two given Gaussian distributions (assuming equal and unknown variances). P-values were generated with a 95% confidence interval ($\alpha = 0.025$).

Amino acid scanning analysis

Single mutant data of 19 amino acids across 55 of 56 positions in the G β 1 domain allowed us to identify the best amino acid for experimental stability scanning. We experimented with two ways to rank the deviations between the $\Delta\Delta G$ for a particular

mutation and the average $\Delta\Delta G$ at each position in the protein. Let us define $\delta\Delta\Delta G_{i,a}$ as the deviation of the stability $\Delta\Delta G_i$ for amino acid mutant i from the average stability at a single residue position a where $1 \leq i \leq 19$ and $1 \leq a \leq 56$. This is given by the following equation:

$$\delta\Delta\Delta G_{i,a} = \left| \Delta\Delta G_{i,a} - \frac{\sum_{k=1}^{19} \Delta\Delta G_{k,a}}{19} \right|$$

In the first ranking system, we used a weighting directly proportional to the actual $\delta\Delta\Delta G_{i,a}$ value calculated. We derived $\delta\Delta\Delta G_{i,a}$ for every single possible mutant and then summed the results for each amino acid across all residue positions. The scoring function for each individual amino acid i in this weighted deviation method is given by:

$$score_i = \sum_{k=1}^{56} \delta\Delta\Delta G_{i,k}$$

As the weighted deviation method can be biased by the large stability changes common to core positions versus those on the surface, an alternative deviation method was developed. In this ranking system, we sorted $\delta\Delta\Delta G_{i,a}$ for all amino acids i from smallest to largest, at residue position a . The ranked position of a given mutation i at residue a is defined as $R(\delta\Delta\Delta G_{i,a})$. For a particular residue a , we would assign a score of 1 to the first-ranking single mutation, a score of 2 to the second-ranking single mutation, and so on through the last amino acid. Once completed, we then summed the individual scores $R(\delta\Delta\Delta G_{i,a})$ for each amino acid across all positions to obtain an aggregate score for that particular amino acid. This score for a given amino acid i in this ranked deviation method is given by:

$$score_i = \sum_{k=1}^{56} R(\delta\Delta\Delta G_{i,k})$$

As the actual scores in each method aren't physically relevant, they are normalized with respect to the score determined for alanine incorporation.

Prediction algorithms

The webserver for Popmusic version 2.1, located at <http://babylone.ulb.ac.be/popmusic>, was used by performing a “Systematic” command on the wild-type crystal structure of G β 1 (1PGA).

The latest release of FoldX (version 3.0, beta 5) was retrieved from <http://foldx.org.es>. The crystal structure of G β 1 (1PGA) was prepared by using the “RepairPDB” command to perform Asn, Gln, and His flips, alleviate small Van der Waals' clashes, and optimize wild-type rotamer packing. Every mutation in the dataset was constructed through the “BuildModel” command, and the difference in energy between the WT reference and the corresponding mutant was averaged over five trials.

The latest release of Rosetta (version 3.3) was retrieved from <http://www.rosettacommons.org>. The `ddg_monomer` application was used to generate single mutant stability data from a pre-minimized version of the crystal structure of G β 1 (1PGA). We explicitly followed the available online documentation in order to prepare all necessary input files. Option sets described in the documentation pertain to the various Rosetta iterations tested in this paper (no bb min: low-resolution protocol; cst bb min: high-resolution protocol; full bb min: high-resolution protocol with an empty distance restraints file).

The performance of each algorithm was evaluated by Pearson's correlation coefficient and the fraction correct, defined as the number of correctly categorized

mutants (stabilizing (≥ 0.5 kcal/mol), neutral (< 0.5 kcal/mol and > -0.5 kcal/mol), and destabilizing (≤ -0.5 kcal/mol)) divided by the total number of mutations in the set.

Mutability determination

In order to provide an upper bound on the mutability of the protein we assumed perfect additivity of the $\Delta\Delta G$ values for single mutants. We defined n_{crit} as the critical number of mutations needed to exceed the threshold $\Delta\Delta G$ of 4.5 kcal/mol. For a series of N mutations, the net stability of such a domain is:

$$\Delta\Delta G_{net} = \sum_{k=1}^N \Delta\Delta G_k$$

The probability of stabilizing/destabilizing a protein domain with a random mutation was examined by making each single mutant equally probable. We ran a simulation in Mathematica for 1000 or 10000 trials with 200 random mutations in each trial. This many random mutations were made in order to have a high likelihood that we would reach the unfolding threshold before this value. We found the critical number of mutations n_{crit} by checking the net stability after each successive mutation. When $\Delta\Delta G_{net,k} > 4.5$ kcal/mol, then $n_{crit} = k$.

References

1. Bowie JU, Reidhaar-Olson JF, Lim WA & Sauer RT (1990) Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science* 247(4948):1306–1310.
2. Rennell D, Bouvier SE, Hardy LW & Poteete AR (1991) Systematic mutation of bacteriophage T4 lysozyme. *Journal of Molecular Biology* 222(1):67–88.
3. Matthews BW (1993) Structural and genetic analysis of protein stability. *Annual review of biochemistry* 62:139–160.
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. Markiewicz P, Kleina LG, Cruz C, Ehret S & Miller JH (1994) Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. *Journal of Molecular Biology* 240(5):421–433.
6. Dahiyat BI & Mayo SL (1997) De novo protein design: fully automated sequence selection. *Science* 278(5335):82–87.
7. 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.
8. Kuhlman B, et al. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science* 302(5649):1364–1368.
9. 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.
10. Guo HH, Choe J & Loeb LA (2004) Protein tolerance to random amino acid change. *Proceedings of the National Academy of Sciences of the United States of America* 101(25):9205–9210.
11. Bloom JD, et al. (2005) Thermodynamic prediction of protein neutrality. *Proceedings of the National Academy of Sciences of the United States of America* 102(3):606–611.
12. Bershtein S, Segal M, Bekerman R, Tokuriki N & Tawfik DS (2006) Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* 444(7121):929–932.

13. Tokuriki N & Tawfik DS (2009) Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology* 19(5):596–604.
14. Fowler DM, et al. (2010) High-resolution mapping of protein sequence-function relationships. *Nature methods* 7(9):741–746.
15. Hietpas RT, Jensen JD & Bolon DNA (2011) Experimental illumination of a fitness landscape. *Proceedings of the National Academy of Sciences of the United States of America* 108(19):7896–7901.
16. Araya CL & Fowler DM (2011) Deep mutational scanning: assessing protein function on a massive scale. *Trends in Biotechnology* 29(9):435–442.
17. Tokuriki N, Stricher F, Schymkowitz J, Serrano L & Tawfik DS (2007) The stability effects of protein mutations appear to be universally distributed. *Journal of Molecular Biology* 369(5):1318–1332.
18. 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.
19. 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.
20. Wells JA (1990) Additivity of mutational effects in proteins. *Biochemistry* 29(37):8509–8517.
21. Axe DD, Foster NW & Fersht AR (1998) A search for single substitutions that eliminate enzymatic function in a bacterial ribonuclease. *Biochemistry* 37(20):7157–7166.
22. Weinreich DM, Delaney NF, Depristo MA & Hartl DL (2006) Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* 312(5770):111–114.
23. 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.
24. Fleming PJ & Richards FM (2000) Protein packing: dependence on protein size, secondary structure and amino acid composition. *Journal of Molecular Biology* 299(2):487–498.
25. Serrano L, Kellis JT, Jr., Cann P, Matouschek A & Fersht AR (1992) The folding of an enzyme. II. Substructure of barnase and the contribution of different interactions to protein stability. *Journal of Molecular Biology* 224(3):783–804.

26. Jackson SE, Moracci M, elMasry N, Johnson CM & Fersht AR (1993) Effect of cavity-creating mutations in the hydrophobic core of chymotrypsin inhibitor 2. *Biochemistry* 32(42):11259–11269.
27. Schwehm JM, Kristyanne ES, Biggers CC & Stites WE (1998) Stability effects of increasing the hydrophobicity of solvent-exposed side chains in staphylococcal nuclease. *Biochemistry* 37(19):6939–6948.
28. Axe DD, Foster NW & Fersht AR (1996) Active barnase variants with completely random hydrophobic cores. *Proceedings of the National Academy of Sciences of the United States of America* 93(11):5590–5594.
29. Gassner NC, Baase WA & Matthews BW (1996) A test of the "jigsaw puzzle" model for protein folding by multiple methionine substitutions within the core of T4 lysozyme. *Proceedings of the National Academy of Sciences of the United States of America* 93(22):12155–12158.
30. 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.
31. Baase WA, Liu L, Tronrud DE & Matthews BW (2010) Lessons from the lysozyme of phage T4. *Protein Science* 19(4):631–641.
32. Cordes MH & Sauer RT (1999) Tolerance of a protein to multiple polar-to-hydrophobic surface substitutions. *Protein Science* 8(2):318–325.
33. Taverna DM & Goldstein RA (2002) Why are proteins so robust to site mutations? *Journal of Molecular Biology* 315(3):479–484.
34. Skinner MM & Terwilliger TC (1996) Potential use of additivity of mutational effects in simplifying protein engineering. *Proceedings of the National Academy of Sciences of the United States of America* 93(20):10753–10757.
35. Chica RA, Moore MM, Allen BD & Mayo SL (2010) Generation of longer emission wavelength red fluorescent proteins using computationally designed libraries. *Proceedings of the National Academy of Sciences of the United States of America* 107(47):20257–20262.
36. Guntas G, Purbeck C & Kuhlman B (2010) Engineering a protein-protein interface using a computationally designed library. *Proceedings of the National Academy of Sciences of the United States of America* 107(45):19296–19301.
37. 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.

38. 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.

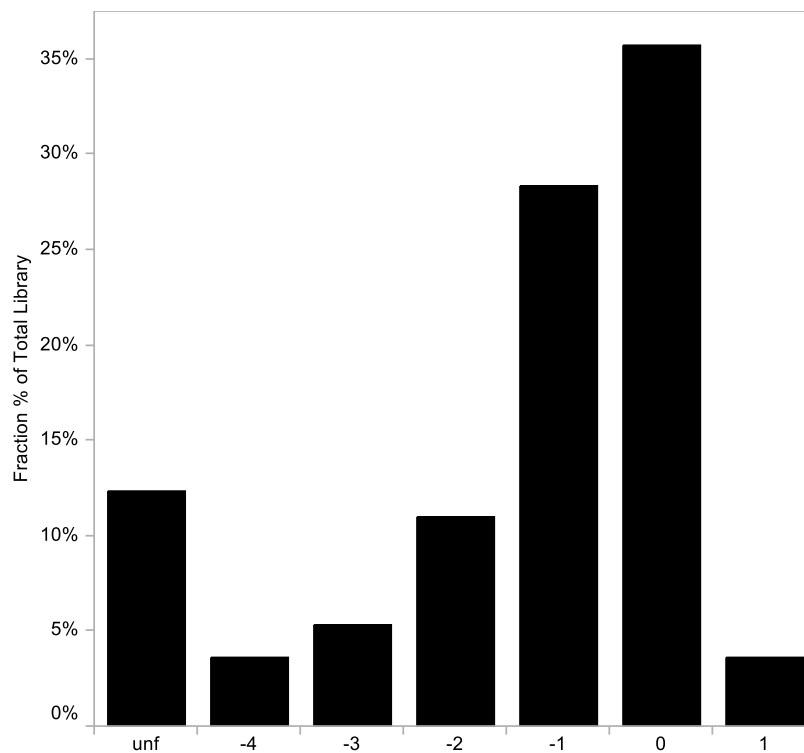


Figure 4-1: Single mutant stability distribution for the G β 1 domain. As an example, if $\Delta\Delta G$ stability data is represented by χ , then the zero bin holds data with values $0 \geq \chi > -1$. The “unf” bin holds mutant stability data that could not be determined, and is likely insoluble or in an alternative conformation.

Type	Experimental data					FoldX3 data				
	μ_1	σ_1	μ_2	σ_2	R^2	μ_1	σ_1	μ_2	σ_2	R^2
Surface ^a	-0.13	0.65	-	-	0.996	-0.53	0.68	-	-	0.995
Core ^a	-	-	0.28	1.23	0.932	-	-	-0.17	1.25	0.901
All	-0.13	0.53	1.01	1.43	0.999	-0.55	0.64	1.02	1.75	0.997

^a Surface and core determination done as described in the methods

Table 4-1: Gaussian fitting parameters for the mutational distributions

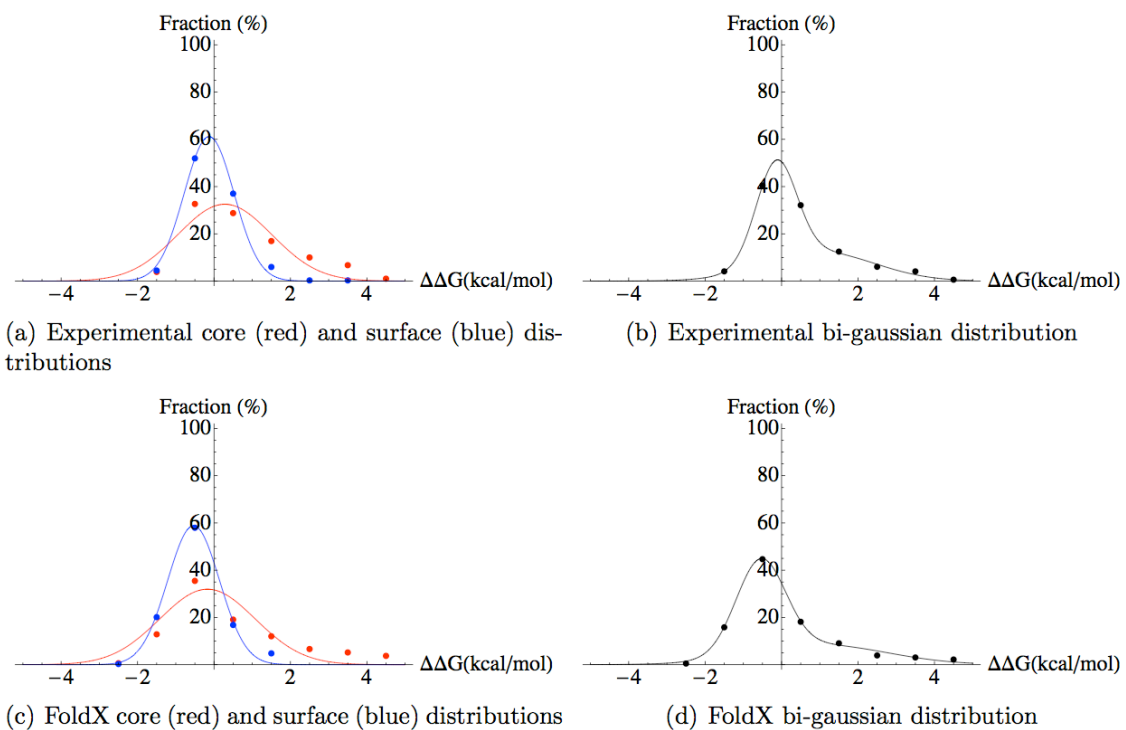


Figure 4-2: Gaussian fits of the G β 1 mutational distribution. The single mutant dataset was calculated using FoldX3, and values corresponding to unfolded data were removed from both datasets. Equation fits are described in the methods. Positive $\Delta\Delta G$ values indicate destabilizing variants and vice versa.

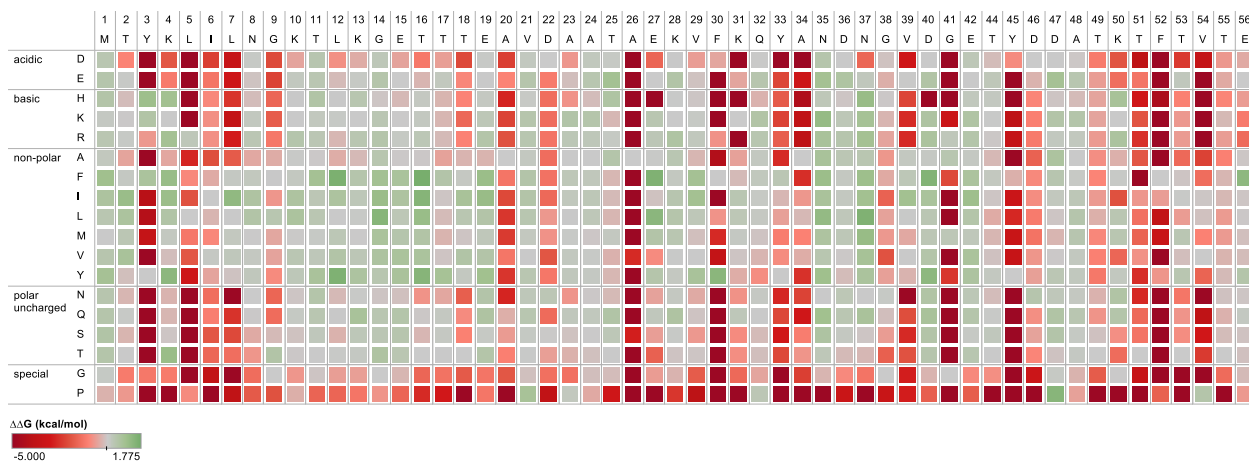


Figure 4-3: Single mutant stability landscape for the Gβ1 domain. Each mutant is colored by its $\Delta\Delta G$ value, where red is destabilizing and green is stabilizing. Self-identity mutations, e.g., M01M, are assigned a zero value and colored gray. Mutant stability data that could not be determined are given the arbitrary value of -5.

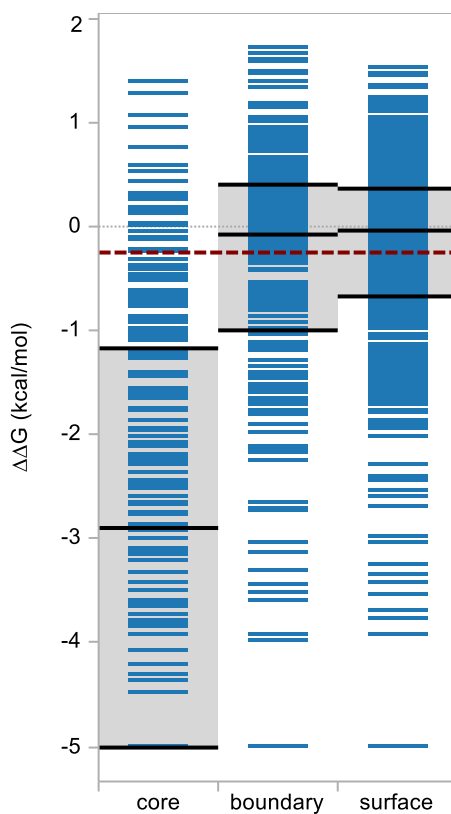


Figure 4-4: Single mutant stability distributions by RESCLASS. The $\Delta\Delta G$ stability distribution for each RESCLASS category is separated into quartiles. The median values for the core, boundary, and surface distributions are -2.89, -0.06, and -0.02, respectively. The red dashed line is the median value for the entire distribution, -0.25. Mutant stability data that could not be determined are given the arbitrary value of -5.

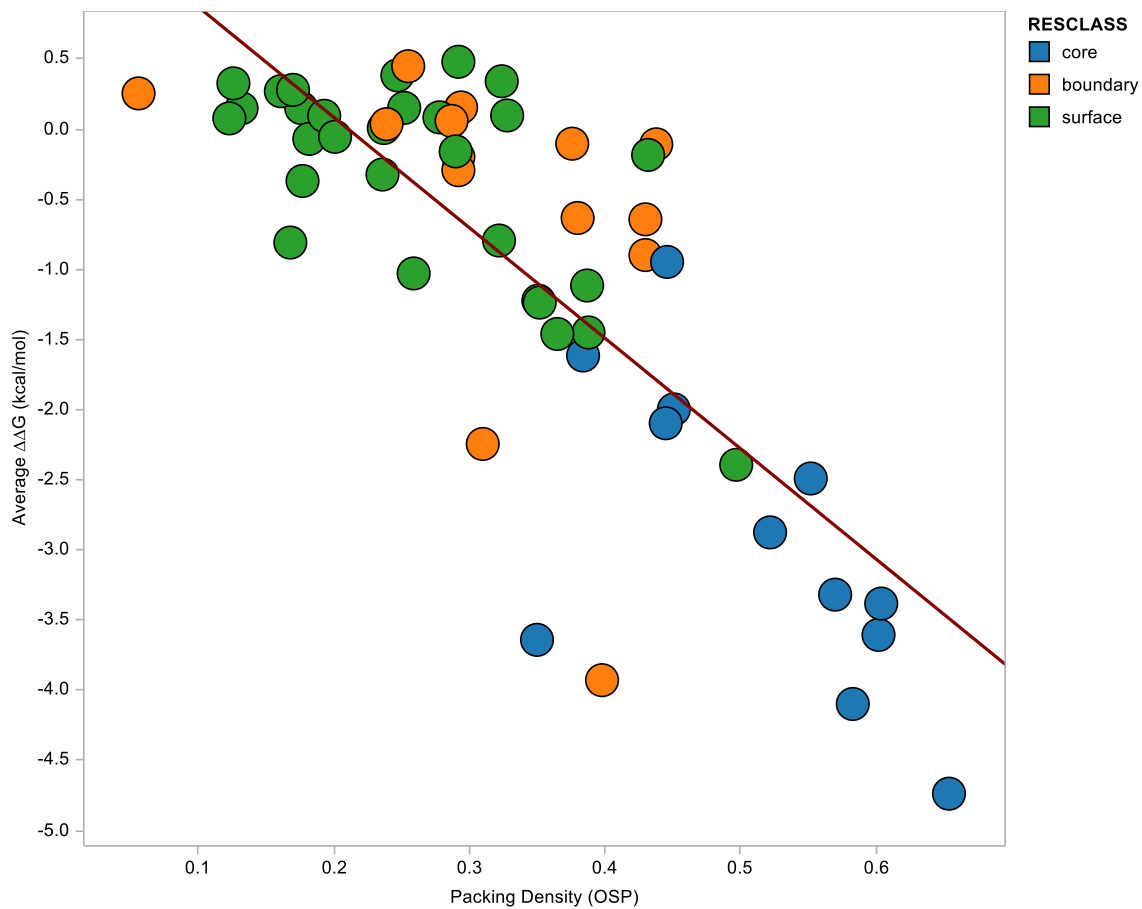


Figure 4-5: Packing density is linearly correlated with $\Delta\Delta G$ averaged by position. Each data point represents a position in the G β 1 domain, and is colored by RESCLASS. The equation for the red trend line is $y = -7.9x + 1.6$, with an r^2 of 0.62.

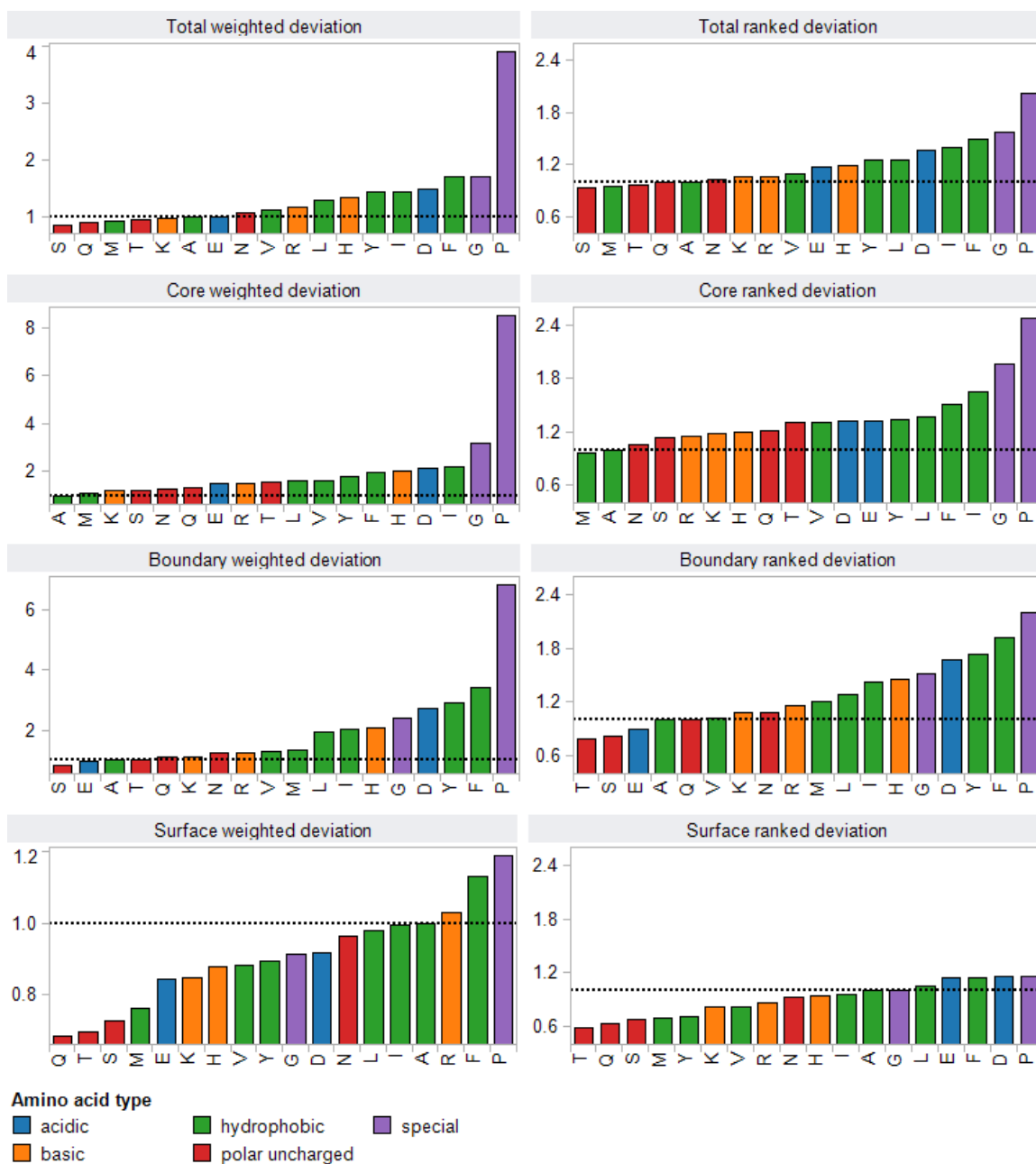


Figure 4-6: Amino acid scanning mutagenesis. The amino acid that best matches the stability at every position in the Gβ1 domain was determined using both ranking and actual kcal/mol weighted deviations (see Methods). The results are normalized to alanine incorporation (black dotted line) to compare against typical functional scanning methodology. The charts describe the overall data and RESCLASS categories and are sorted from the best to worst match. Amino acids are colored by physiochemical type.

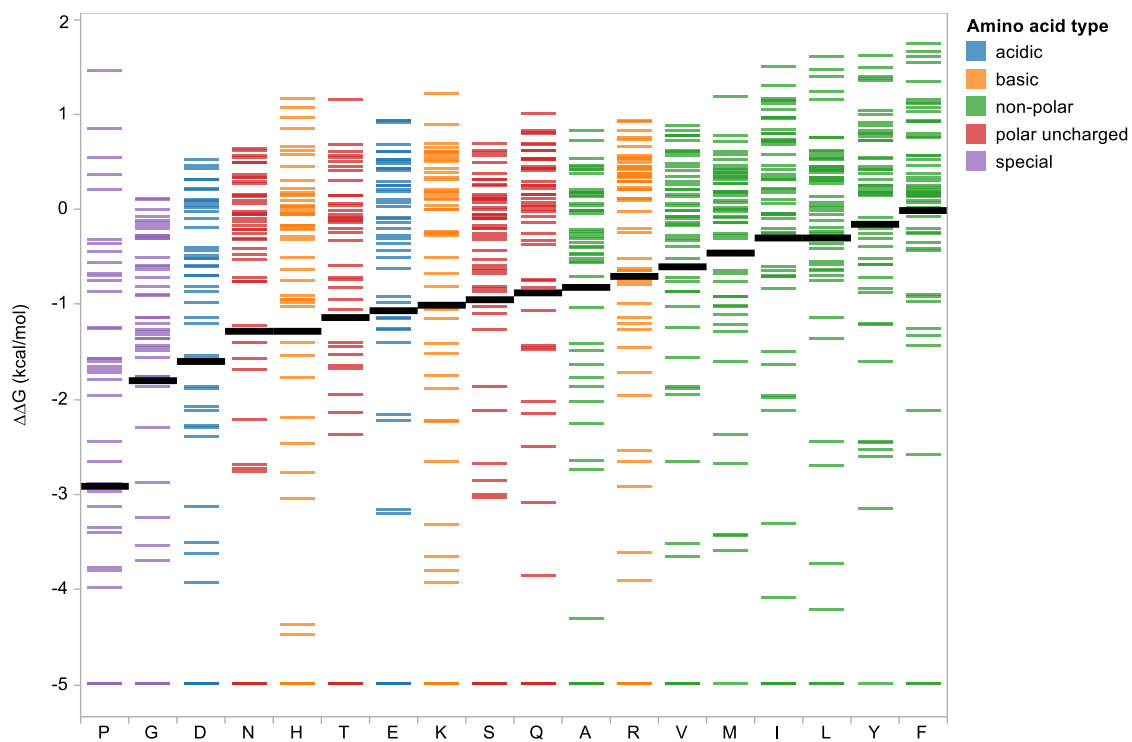


Figure 4-7: Stability distribution of G β 1 by mutant amino acid. Incorporated amino acids are sorted by the average $\Delta\Delta G$ stability effect of that mutation (black reference line) over the G β 1 domain. Amino acids are colored by physiochemical type.

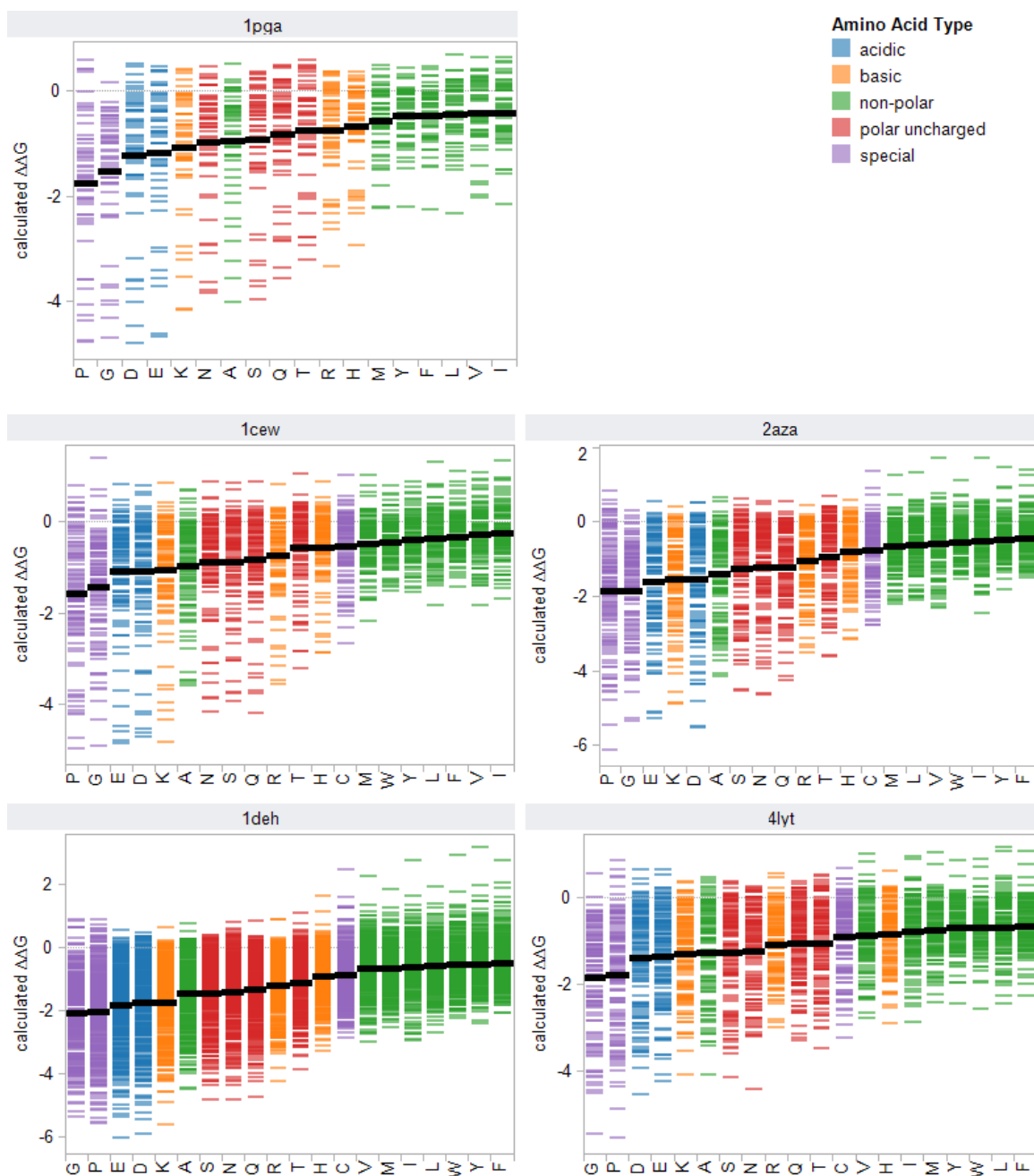


Figure 4-8: Calculated stability distributions by mutant amino acid. Single mutant stability data was calculated with Popmusic2 for the G β 1 domain (1pga), along with four other proteins: cystatin (1cew), azurin (2aza), alcohol dehydrogenase (1deh), and lysozyme (4lyt). Incorporated amino acids are sorted by the average $\Delta\Delta G$ stability effect of that mutation (black reference line) over each domain. Amino acids are colored by physiochemical type.

Variant	Number of Residues	% Helix^a	% Strand^a	Average OSP^b
1pga	56	25	43	0.332
1cew	108	20	48	0.314
2aza	129	16	36	0.368
4lyt	129	41	11	0.376
1deh	374	28	25	0.392

^aSecondary structure was determined through DSSP.

^bResidue packing density (OSP) was averaged over each protein.

Table 4-2: Bioinformatics statistics for selected proteins

OSP percentile rank	Average $\Delta\Delta G^a$ by quartiles	Average $\Delta\Delta G^a$ by halves
1–25	0.382	0.449
26–50	0.516	
51–75	0.097	-0.094
76–100	-0.285	

^a Calculated over Y/F/I/L amino acids only

Table 4-3: Comparing the average $\Delta\Delta G$ of hydrophobic mutations by OSP

	Popmusic2	FoldX3	Rosetta (no bb min)^a	Rosetta (cst bb min)^b	Rosetta (full bb min)^c
All	0.56/825	0.35/825	0.26/819	0.45/825	0.44/825
All w/o clashes	-/-	0.52/742	0.35/747	0.62/810	0.61/824
+VolΔ	0.46/489	0.42/417	0.37/428	0.56/476	0.52/488
-VolΔ	0.6/340	0.53/329	0.35/323	0.62/338	0.64/340
Core	0.28/127	0.31/73	0.3/75	0.26/115	0.22/127
Boundary	0.55/221	0.65/217	0.63/213	0.72/219	0.71/221
Surface	0.54/477	0.41/452	0.35/459	0.57/476	0.57/476
NP\rightarrowNP	0.43/125	0.57/91	0.52/85	0.60/118	0.62/124
NP\rightarrowP	0.49/163	0.62/125	0.34/139	0.64/156	0.53/163
P\rightarrowNP	0.66/249	0.47/240	0.42/236	0.69/248	0.68/249
P\rightarrowP	0.58/288	0.41/286	0.58/287	0.52/288	0.53/288

All entries are tuples of correlation coefficient (r) and number of data points (n); NP: nonpolar; P: polar.

^a No backbone minimization after repacking

^b Constrained backbone minimization after repacking

^c Unconstrained backbone minimization after repacking

Table 4-4: Algorithm performance by linear correlation

	Popmusic2	FoldX3	Rosetta (no bb min)	Rosetta (cst bb min)	Rosetta (full bb min)
All	0.62/935	0.59/935	0.61/918	0.61/935	0.60/935
+VolΔ	0.52/527	0.52/527	0.52/510	0.53/527	0.51/527
-VolΔ	0.75/412	0.69/412	0.72/412	0.71/412	0.70/412
Core	0.78/204	0.84/204	0.86/187	0.83/204	0.82/204
Boundary	0.57/238	0.57/238	0.53/238	0.53/238	0.52/238
Surface	0.58/493	0.50/493	0.55/493	0.53/493	0.54/493
NP\rightarrowNP	0.61/147	0.68/147	0.61/140	0.65/147	0.64/147
NP\rightarrowP	0.66/210	0.72/210	0.68/200	0.74/210	0.67/210
P\rightarrowNP	0.61/272	0.57/272	0.64/272	0.61/272	0.60/272
P\rightarrowP	0.62/306	0.49/306	0.53/306	0.51/306	0.52/306

All entries are tuples of fraction correct and number of data points (n); NP: nonpolar; P: polar.

^a No backbone minimization after repacking

^b Constrained backbone minimization after repacking

^c Unconstrained backbone minimization after repacking

Table 4-5: Algorithm performance by fraction correct

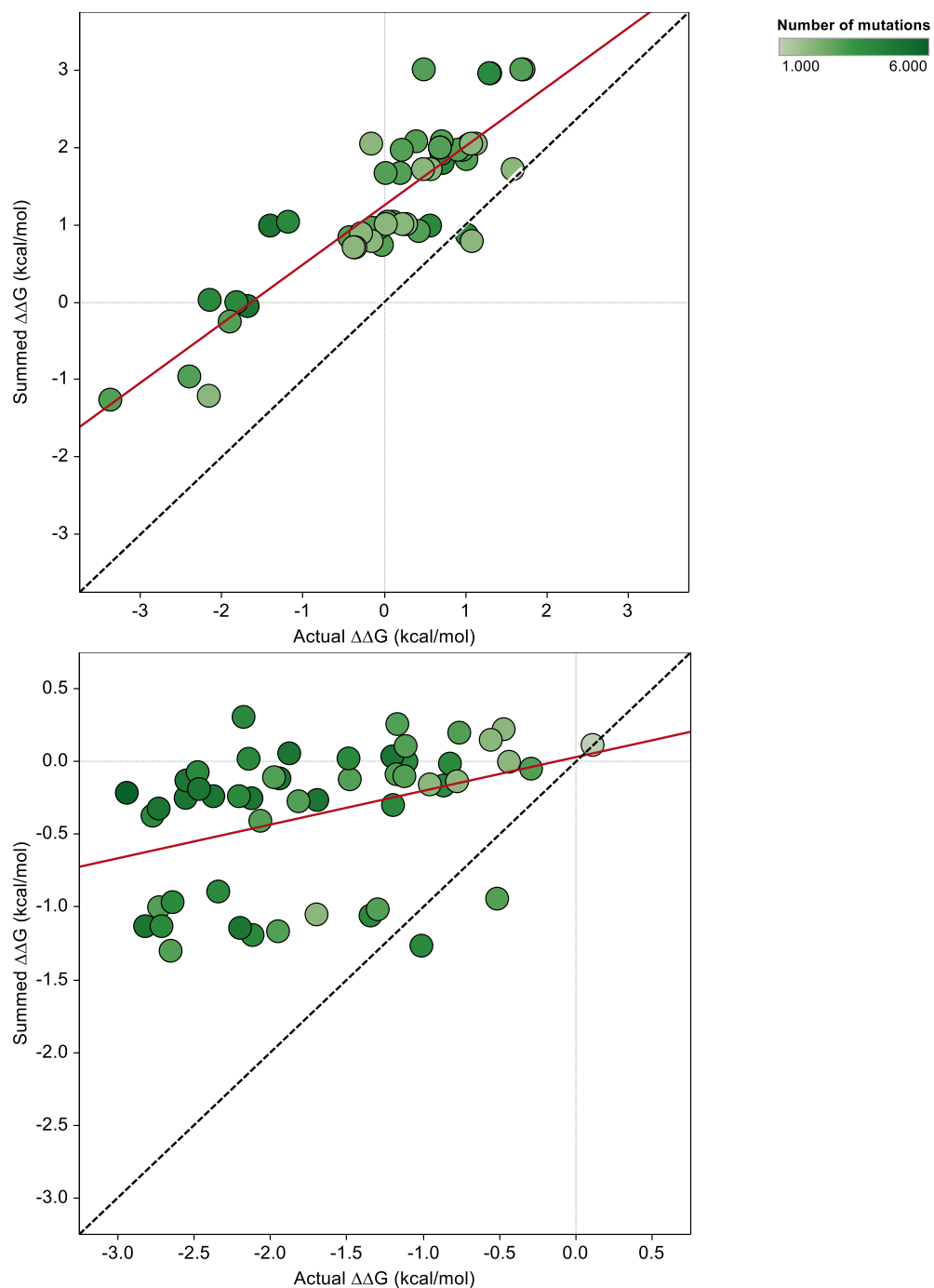


Figure 4-9: Complex additivity in core and surface mutation libraries. Actual $\Delta\Delta G$ is plotted against the sum of single mutation $\Delta\Delta G$ values for core (top) and surface (bottom) mutational libraries. Each data point is colored by the number of mutations from wild-type it carries. The black dashed line is $y = x$, and serves as the indicator for perfect additivity. The r^2 for the red trend lines are 0.74 (top) and 0.15 (bottom).

Variant	Identity	Actual $\Delta\Delta G^a$	Sum $\Delta\Delta G^b$
6-fold	K04T L12Y E27L F30Y N35K E56F	-0.10	8.07
8-fold	6-fold G14L D47P	0.57	10.98
16-fold	8-fold M01F T02I L07I T16I E19I N37L D40F E42I	unfolded	21.21

^a Determined stability through experiment; units in kcal/mol

^b Determined stability by summing the $\Delta\Delta G$ of the individual single mutants; units in kcal/mol

Table 4-6: Identity and stability of additive variants

High-throughput and automation methods

Appendix

High-Throughput and Automation Methods

November 2011

Alex Nisthal | Mayo Lab



Table of Contents

Chapter 1: Robot Operation

Overview	132
The Tecan Freedom EVO 200	132
System Liquid	
Deck Layout	
Power ON	134
Maintenance after Power ON	
Power OFF	135
Maintenance after Power OFF	
Using EVOware 2.1	135
Line Editing	
Worktable Editing	
Running a Script	
Error Messages and Stopping a Script	
Email Notification	
Backing Up/Restoring EVOware	
Basic Operation Checklist	142

Chapter 2: Robot Procedures

Overview	143
Dilute Oligonucleotides (Gene Assembly)	144
Materials	
Methods	
IPIPE and CPEC PCR (Gene Assembly)	147
Materials	
Methods	
CPEC Transformation	154
Materials	
Methods	
QTray Plating w/Sterile Beads	159
Materials	
Methods	

Dilute IDT Oligonucleotides	163
Materials	
Methods	
Site-Directed Mutagenesis	168
Materials	
Methods	
Dpn1 Digestion	174
Materials	
Methods	
Bacterial Transformation	179
Materials	
Methods	
QTray Plating	185
Materials	
Methods	
Culture Rearray	189
Materials	
Methods	
Expression Plate Inoculation	193
Materials	
Methods	
Protein Purification	197
Materials	
Methods	
HT Agarose Gel Electrophoresis	204
Materials	
Methods	

Chapter 3: Site-Directed Mutagenesis

Overview	208
Oligonucleotide Design	209
Experimental Pipeline	210

Chapter 4: Automated Gene Assembly

Overview	213
----------	-----

Python Scripting	215
The Input File	215
Output	216
Test Case: Rational Stabilization of a Designed Enzyme	217

Appendix:

BL21 Gold DE3 competent cells	219
48-Well Qtray Preparation	220
EasyPress Operation	221
E-Gel 96 Operation	222
DNA Ladder for E-Gel 96	224

Chapter 1: Robot Operation

Overview

This chapter provides an introduction to successful operation of the Tecan Freedom EVO liquid handling robot. It outlines the system basics, proper power on/off, maintenance, and software usage. This manual should be used in conjunction with the Tecan-provided EVOware Help, a fantastic resource that explains every setting inside the software.

The Tecan Freedom EVO is located in the Beckman Institute in Room 288. The space is owned by the Protein Expression Center, headed by Dr. Jost Vielmetter. The lab benches and pipettes immediately around the robot can be used to prepare your automation experiments. Full size refrigerators at -20°C and 4°C are also on hand to help with experiment preparation. The center cabinet underneath the desk is full of labware and supplies for the Mayo Lab.

The Tecan Freedom EVO 200

The 2-meter robot sits on a reinforced table inside of positive pressure sterile hood and is composed of three unique arms overlooking a precisely manufactured deck (Figure 1). The left-most arm, the 8-tip Liquid Handler (LiHa), has eight liquid-sensing fixed tips that require a washing step after each use. Each tip is individually addressable and ideal for cherry-picking operations. The middle arm, the Multi Channel Arm 96 (MCA96), uses disposable tips to pipet 96 volumes of liquid at once, making it very useful for serial dilutions or buffer additions. The right arm, the Robotic Manipulator (RoMa), can pick up and move any microtiter plate that conforms to the Society for Biomolecular Science (SBS) standard. Different kinds of carriers (holding places for microtiter plates) populate the deck, and although the deck layout was optimized for our applications, not all carriers are accessible to all of the arms.

System Liquid

All fixed tip systems have a tip wash station that requires water from an outside source. In the general case, this is a large water reservoir that is manually refilled as it empties through the system. Luckily, our Freedom EVO has been customized (courtesy of Dr. Vielmetter) with a reservoir that auto-refills from a local house DI water line. The custom reservoir also has a UV filtration unit installed that prevents microbial growth in the standing water. This setup greatly minimizes weekly maintenance of the system. The custom reservoir feeds sterile water through a single tube into the Freedom EVO through the back; this tube is split into 8 channels (as seen by the 8 syringes on the middle top bar of the instrument) that then exit out of the 8 tips of the LiHa.



Figure 1. The Tecan Freedom EVO 200, as seen in BI288.

The LiHa wash station drains into the house drainage system on the left side of the instrument. With this modification, the Freedom EVO is essentially a closed loop system: water is pulled from the house line into a sterile reservoir, passed through the robot system, flushed through the LiHa, and finally clears out into the drain.

The LiHa arm in our system is outfitted with the Te-Fill accessory that enables it to dispense large volumes of liquid from reservoirs located on the left side of the instrument. The Te-Fill system can be used to switch between five different buffer sources for a variety of extended automated routines. However, the pump on the Te-Fill is slow and was the motivation for acquiring and integrating the peristaltic pump system (described below).

Deck Layout

The deck layout as seen in the EVOware software (Figure 2) is an accurate portrayal of most of the carriers on the deck that are physically accessible by the liquid handling arms. Two room temperature plate hotels, three controlled temperature plate hotels, and four plate stackers along the back of the instrument are not properly represented in the figure.

Directly below the deck, the robot table houses two peristaltic pumps installed and integrated by Dr. Vielmetter. These devices, in combination with specialized labware also developed by Dr. Vielmetter, are able to provide fast-filling reservoirs of any desirable liquid. One useful arrangement of this subsystem is as a tip wash station for the MCA96 in which the primary reservoir is constantly exchanging water (to wash the tips), and the secondary reservoir is exchanging fresh ethanol (to dry the tips). Outside of the hood and to the left is a variable temperature water recirculator that is connected to a specialized 3 position carrier on the deck. The recirculator can cool the carrier down to 4°C or warm it up to 37°C in 30 minutes.

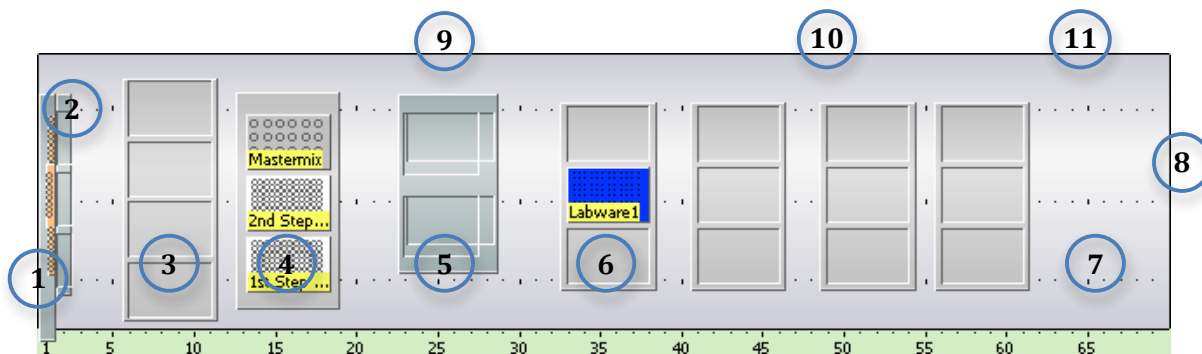


Figure 2. The deck layout, as depicted in the EVOware software. The features on the deck are as follows: (1) LiHa tip wash station, (2) 3 position 100mL trough carrier, (3) 4 position MCA tip box carrier, (4) 3 position MP variable temp carrier, (5) 2 position Te-Vac vacuum system, (6) four 3 position MP flat carriers, (7) 2 position Te-Shake MP shaker, (8) attached Tecan Infinite M1000 plate reader, (9) 4 off-deck stackers, 1 SPE hotel, 1MP hotel, (10) 3 off-deck variable temp hotels, (11) BIO-RAD thermalcycler with automated lid. (MP: microtiter plate, SPE: solid phase extraction)

Power ON

The robot platform and accompanying computer can usually be found with the instrument and computer powered on and the computer at the Microsoft Windows login screen. In this instance, simply login to the computer, double-click the EVOware icon, and login to the EVOware software.

After a cold start (the robot was turned off/on), the robot will need to do a long initialization in which the three arms test their range of motion and move to the left-most edge of the deck. It should be noted that when the robot is turned off, the arms are no longer locked in place and can be moved freely by the user. This 5-minute process is not done when the robot is left on. Instead, the robot will move the LiHa and the RoMa to their home positions, which is a much faster process.

Maintenance after Power ON

Immediately after running EVOware, it is very important to run the *Daily_Flush* maintenance routine. This script not only initializes the instrument (prompting the arms to move to their home positions on a warm start) but also flushes and rinses the LiHa with 25mL of water. The act of flushing the LiHa removes bubbles and restores the system liquid trailing air gaps (STAG) in the eight tips. This is important for accurate liquid dispense operations, especially multi-dispense techniques. After flushing, the maintenance routine moves the LiHa over the second microtiter plate carrier and spreads the tips to allow the user to rubdown each tip with an ethanol-soaked kimwipe. This ethanol rubdown step also helps in maintaining the accuracy of the liquid dispensing system. Finally, check that all the twist knobs associated with the LiHa syringes are finger tightened before use. A common reason for inaccurate liquid handling is loose syringes. After this, the robot is ready for use.

Additional maintenance may be needed if your robotic routine calls for the Te-Vac or the thermalcycler. If using the Te-Vac, check that the waste bucket is less than half-full, or preferably empty it if you have a vacuum heavy routine. If using the thermalcycler, the

sealing pad on the automated lid should be washed with a dilute bleach-soaked kimwipe and subsequently washed twice more with a water-soaked kimwipe.

Power OFF

After your robotic routine is completed, clear the deck of your labware, including plates, tubes, and tips. Throw away used tips boxes and eppendorf tubes. Exit the software, double-click the “shutdown EVOware drivers” icon, and log out of the computer. You may leave both the computer and the robot on for the next user in this state.

To leave the instrument in a true OFF state, after the above is completed press the light green arrow button on the robot located on the bottom far right side. Turn off the computer through the Windows Start menu.

Maintenance after Power OFF

Most if not all routines include a tip wash step for the LiHa after use, so any extra wash steps are unnecessary after a script is finished. If the Te-Vac was used, be sure to clean the carrier and the vacuum carrier using a dilute bleach rinse. If the cooling water recirculator was used, remember to turn it off and refill it with DI water if necessary.

Using EVOware 2.1

Upon execution of the EVOware software, the Startup window asks whether you want to run an existing or maintenance script, create a new script, edit an existing one, or configure the system. If you are just starting your session, select “Run maintenance” and then “Daily_Flush” (Figure 4). This opens the Runtime Controller window, where clicking “RUN” sends the script to the robot to be performed. After the process and the appropriate maintenance is completed, the Runtime Controller will alert you of the successful completion of the script. By clicking “Cancel” after a successful finish, the software will return you to the Startup window.

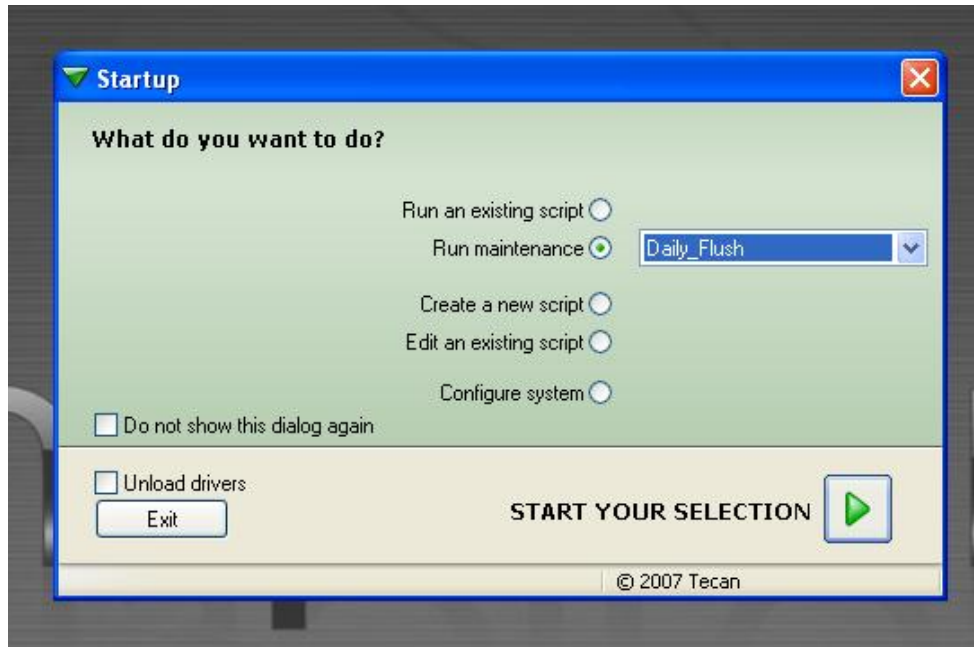


Figure 3. Run the maintenance script "Daily_Flush" from the Startup screen. After clicking "START YOUR SELECTION," click "RUN" at the Runtime Controller window.

After running "Daily_Flush," select "Edit an existing script" to open the list of saved scripts. Select the script of interest (frequently used scripts may be located in the Favorites tab) and open it (Figure 5). Scripts can also be found in separated folders at the bottom of this list. This will open the main functional window of the EVOware software.

Another option is to create a new script and select a template from the Favorites tab. When running one of the methods described later in Chapter 2, this is the preferred option so that any edits made are isolated to that user's script. The majority of users should ignore the other options on the Startup window. Accessing the system configuration is only appropriate for advanced users. Remember, running a script without inspection is decidedly not recommended unless it has been thoroughly debugged and you are an advanced user.

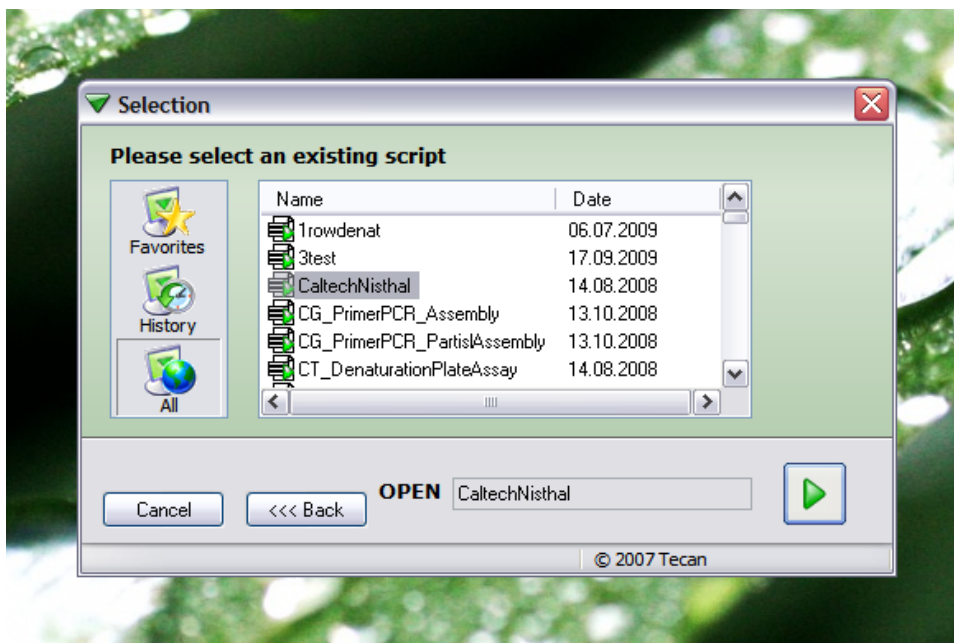


Figure 4. Edit an existing script from the Selection window. Select the script of interest and click the green arrow button.

Line Editing

Along the right side of the screen is the Script Editor window. Here, each line is the equivalent of a line of code that must be performed sequentially. It is extremely important to double check every line of code to make sure the robot is given the correct procedure.

Every available command and popup screen is detailed in the EVOware Help (hit the F1 key or click “Help” along the top of the screen, then “Contents”), so only a brief description of the commonly used “Aspirate” and “Transfer Labware” commands will be provided here.

Many script lines feature information that is useful when scanning a script for possible errors before execution. The “Aspirate” command, for example, shows a cartoon of the tips in use next to a highlighted area of the labware it will pull from (Figure 6). Next to this, the top line shows the amount being aspirated and the liquid class the system will use to do it. The bottom line displays the name of the labware (in quotation marks) followed by the exact positions that will be aspirated (in parentheses). If the command is changed as part of an iterable loop, the number of options that are changing is displayed as well.

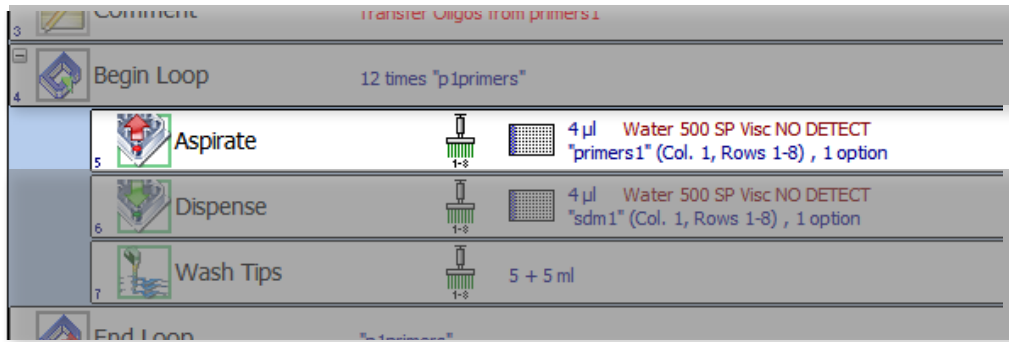


Figure 5. The LiHa "Aspirate" command is highlighted, as seen in the Script Editor window.

The "Transfer Labware" command also contains information that is useful to check before running a particular script. The top line describes a source site using carrier position notation; this is followed by the destination site, which uses the same notation (Figure 7). Finally, the type of grip (narrow or wide) is described. If using a special microtiter plate (such as the PCR plate), a different, user-defined vector is used by the RoMa. All these settings are defined by the user when the script is initially written but can be edited by double-clicking the line of interest.



Figure 6. The RoMa "Transfer Labware" command is highlighted, as seen in the Script Editor window.

Worktable Editing

The Worktable Editor window shows a faithful reconstruction of the physical deck configuration of the robot. Through this graphic interface, the user can assign carrier and labware placements and label labware. In general, this window should be used to populate the deck of the robot with the labware needed for a particular script. However, the user should be aware that EVOware creates duplicate instances of labware that is moved by the RoMa. For example, a PCR plate that is labeled "insert-pcr" will show up with the default label, "Labware1" in other sites where it will eventually end up. For this reason, script authors should label all labware before entering any RoMa commands to help in identifying the duplicates.

As noted earlier, carriers that are not accessible by the liquid handling arms (such as the plate stackers and hotels) are pictured without respect to their numerical position on the robot deck (Figure 2). Their physical positions are only indicated when they have been clicked on by the user.

Interface tips:

- ✓ Double-clicking a piece of labware will open the labeling popup.
- ✓ To quickly create a duplicate of a piece of labware, hold down "Control" while dragging the labware.
- ✓ For advanced users: Right-clicking on labware or carriers opens the option to edit them. This is where new labware or where RoMa vectors to carriers can be taught, respectively.

Running a Script

After a script has been inspected and the user is comfortable proceeding, click the "RUN" icon (green arrow) along the top of the screen. If the latest version of the script isn't saved, the program will prompt you to save it before continuing. Now the Runtime Controller window will pop up; click "RUN" to proceed (Figure 8). Near the top of the Runtime Controller window is a check box for running the full script. If unchecked, specific line numbers corresponding to portions of the script can be entered. This is useful in cases in which a script was programmed to handle two microtiter plates, but you want to process one. The Runtime Controller window also has toggle buttons to open extended view and to open the active logfile viewer.

The extended view window is very handy for checking how far along the robot is in processing your script. The top left window prints out comments that are programmed in the script. The bottom left window features loop counters for any loops you may have in the script.

Once the robot is done, the Runtime Controller window will report the elapsed time and allow the user to view the extended view and logfiles of the completed script. Click "Cancel" to return to the main screen of EVOware. Clicking "New" will allow the user to run the same script again right away, which is generally not recommended.

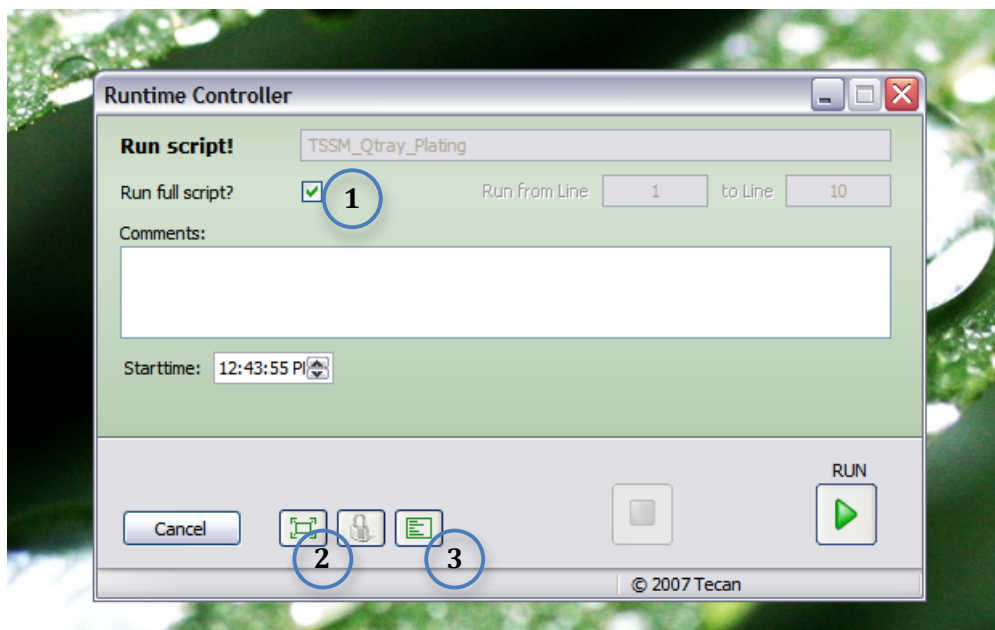


Figure 7. The Runtime Controller window. Important options on this window include: (1) the “Run full script?” toggle, (2) the extended view button, and (3) the logfile viewer button.

Error Messages and Stopping a Script

There are any number of errors that can occur during the execution of a script, but the two most common are hardware crashes and notices of insufficient liquid. Although neither should happen when running verified scripts, they can still creep in if a carrier on the deck has moved or if liquid volumes are adjusted on the fly.

When the software has detected a hardware crash, the script is stopped and the arms lock down into place. EVOware will ask to be placed into virtual mode, where it can no longer communicate with the hardware. In this instance, shut down EVOware and then turn off the instrument. Now the arms should be unlocked and the user can attend to the situation on the deck. Once finished, double-click “Shutdown EVOware drivers,” and then turn on the instrument. Open EVOware and login as usual. You may get messages (that you can cancel/ignore) asking you to zip and send the last logfile to Tecan. Opening a script that has caused a crash will open the “Recover” option in the Run menu. Never use the “Recover” feature of EVOware; it is much better to highlight the remainder of the script you wish to execute. The “Run Direct” command simply skips through the runtime controller; there is no need to use it. The script editor will also highlight which line caused the crash.

An insufficient liquid notice is much easier to handle and is only a factor when using liquid detection. A popup box will ask if you wish to detect again, go to z-max, or aspirate nothing.

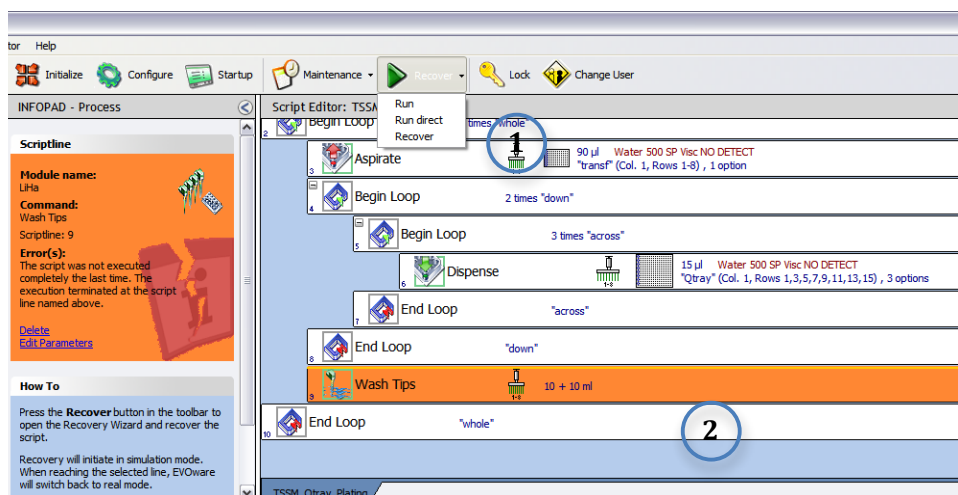


Figure 8. The Run menu in EVOware after an error has occurred. (1) The three options are Run, Run direct, and Recover. (2) The script line that produced the error is also highlighted in orange, and can sometimes be highlighted in red, depending on the severity of the error.

In almost every case, the user should move along and aspirate nothing because moving to z-max may cause a hardware crash and detecting again will give the same error. A note should be made as to why there was not enough volume; sometimes the robot may be attempting to aspirate an unexpectedly viscous liquid or, more likely, a scripting mistake has occurred.

If the user detects a potential upcoming problem in the execution of a script, the Runtime Controller window allows for pausing or stopping a script. After stopping a script, click "Cancel" to return to the main EVOware window. Again, never use the "Recover" feature of EVOware, as it is somewhat unpredictable. Pausing here is identical to pressing the pause button on the robot. Stopping or pausing a script from within the software allows the robot to finish the line it is on before completely stopping and unlocking the shield. If a hardware crash is imminent, it may be necessary to immediately stop the robot by turning it off at the power button.

Email Notification

One of the most useful features in EVOware is the ability to send an email at any point in the script. To do this, you must be entered as a recipient in the EVOware Configuration and then add a "Notification" line to your script.

In the EVOware Configure window, click the "General" tab and then under "Notifications," click "Recipients" and add the name and email of a new user. Then under "Notifications," click "Groups" and add the new user's name as a new group. At the minimum, make the new user and the current automation specialist in the lab members of the newly created group. This configuration allows the automation specialist supervision of new users without flooding other uninvolved users email inboxes. Along the top of the screen, click "Save" and then click "Exit." Restart EVOware to make the changes permanent.

Backing Up/Restoring EVOware

For advanced users only. The home directory for the EVOware installation is "C:\Program Files\TECAN\EVOware."

Complete backups of the EVOware installation can be made by entering the EVOware Configure window and clicking "Backup" at the top of the screen. This will generate a folder with everything needed to restore a broken installation in ".\backup." These backups are also useful for restoring a particular version of a script or for viewing labware/carrier definitions. Every script is saved to a scripts folder in the backup, allowing the user to copy the backup script into the ".\database\scripts" folder in the real EVOware installation directory. Labware and carrier definitions are not saved independently in the backup folder, but instead are saved as one carrier.cfg file that can be copied into the real ".\database" directory.

The software also keeps every saved version of important files so as to immediately recover from a negative change. For example, another user may have edited a carrier definition improperly a few days ago but the last full backup was done two months ago. In that span of time, several positive changes may have been made so that you wish to keep the installation and only fix the carrier file. A user can enter the ".\AuditTrail\configuration" folder and copy the carrier.cfg file from before the negative change was made (minding the dates) and paste it into the actual ".\database" directory.

As in any instance in which a system file is being copied over, remember to create a duplicate of the file in case the older one makes things worse.

Basic Operation Checklist

- ✓ Login to the computer and then into EVOware.
- ✓ Run the maintenance script "Daily_Flush."
- ✓ Check syringe knobs and wipe down the LiHa tips.
- ✓ Create a new script from a template or edit an existing script.
- ✓ Check the script code for possible errors and make adjustments.
- ✓ Populate the physical deck by following the Worktable editor.
- ✓ **Check the script code for possible errors.**
- ✓ Run the script.
- ✓ When finished, clear the deck, close EVOware, and log out of the computer.

Chapter 2:

Robot Procedures

Overview

This chapter covers the important procedures that have been developed on the Tecan Freedom EVO 200. Most of these protocols were developed with performing site saturation mutagenesis in mind, but can be run independently.

Each protocol described here features a brief introduction, followed by a materials list (complete with order numbers), and a methods section. In the methods section, the actual robotic script is reprinted from the basic template procedure and important details are highlighted. The corresponding deck image is also included for every full script shown.

All the procedures were saved as separate EVOware templates. To run a protocol, select “Create a new script” from the Startup Menu and select the template of interest. Then save the opened script with your initials as a prefix. This makes for a much faster startup for novice users who wish to run the default protocol. Also, it keeps the core of the program intact, allowing individual users to make any appropriate changes for their particular experimental setup without affecting other users' scripts. Of course, if a user wishes to run the exact same protocol again later, their EVOware script can be found under “Edit an existing script.”

IMPORTANT

The robot was upgraded and its deck remodeled in the fall of 2010. As such, the details of many scripts described here are incorrect but the concepts remain the same. A handful of scripts were added after the upgrade, are up-to-date, and are marked appropriately.

Dilute Oligonucleotides (Gene Assembly)

*Current Deck Layout

This script performs the replacement method of library assembly as detailed in chapter 4. The methodology was developed in response to issues with the moderate number of high-ranking sequences found in designed degenerate codon libraries. Explicit consideration of each sequence allows for a more direct translation of protein design results into experimental reactions. The Python script described in chapter 4 produces a robotic worklist, written into the .csv format, which is then translated into the robot-readable .gwl format by the EVOware software.

There are three main specification pages when ordering plate oligos from IDT. For the first page, mark “file upload,” “email,” “96-well plate,” and “column” loading. For the second page, mark “25 nmole DNA,” “standard desalting,” “no” CE service, “full yield,” “V-bottom plate,” and shipping “wet.” For the third page, mark “150 μ M,” leave the volume blank, and mark “IDTE buffer pH 7.5.”

These specifications allow users to start automation experiments much faster because there is no need for resuspension, and they keep the plate in a robot-friendly column-loading scheme. The lower limit for ordering oligos in the plate format is 24. A MS Excel template is available in the lab or from IDT to upload the sequence information.

Materials

96-Well IDT Oligonucleotide Plate(s). The oligo plates from IDT arrive with a rubber capmat, lid, and taped up. Dispose of the tape but keep the capmat and the lid.

96-Well V-Bottom Plate(s). These plates, available through Axygen (P-96-450V-C) were chosen for diluting oligos because of their similarity to the plates from IDT. They are ordered non-sterile and autoclaved before use.

Sterile H₂O in 100mL Disposable trough. This trough, available from Tecan (10613049), fits in the carriers located in grids 2 through 5. Fill with high quality water typically used for diluting the oligos.

Microtiter Plate Rubber Capmat. The 96-well V-Bottom plates are sealed with an Axygen ImpermaMat (AM-2ML-RD-IMP) to prevent evaporation. Non-frozen sealed plates may need to be centrifuged down prior to opening due to sample adhesion to the seal.

Methods

Setup

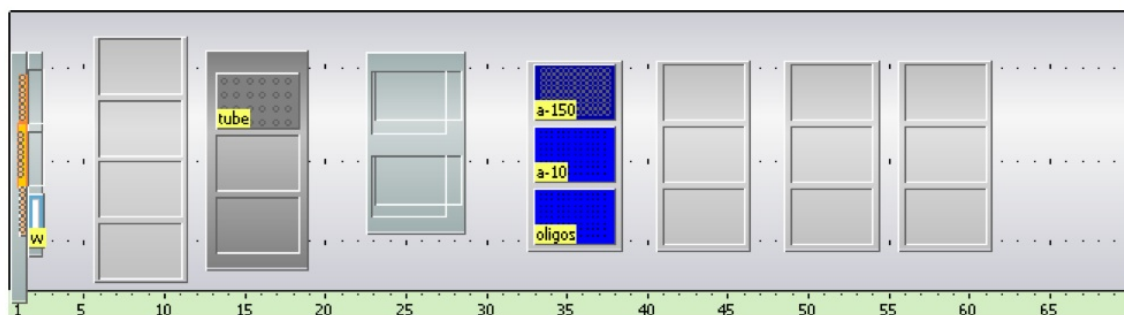
The gene assembly oligo dilution script creates 10uM diluted plates from an original 150uM IDT plate. Volumes can be adjusted to dilute to concentrations more applicable to the user's downstream procedures. A robot worklist, "1A53-8ng.csv", then directs the assembly of oligonucleotides into each well of the oligos collection plate.

Water: 100ml disposable trough

a-150uM: 96-well IDT oligonucleotide plates

a-10uM, oligos: 96-well V-bottom plates









tube: ignore, not used in the method



Sample data from "1A53-8ng.csv"

source position	source label	dest position	dest label	volume
1	a-10	1	oligos	5
2	a-10	1	oligos	5
3	a-10	1	oligos	5
4	a-10	1	oligos	5
5	a-10	1	oligos	5
6	a-10	1	oligos	5
7	a-10	1	oligos	5
8	a-10	1	oligos	5
9	a-10	1	oligos	5
10	a-10	1	oligos	5
11	a-10	1	oligos	5
12	a-10	1	oligos	5
13	a-10	1	oligos	5
14	a-10	1	oligos	5

Procedure

1	Wash Tips		5 + 5 ml
2	Comment	***START 1st PLATE SET***	
3	Comment	Dilute primers into 10uM plate	
4	Begin Loop	8 times "Dispense the dilutee across the 10uM plate"	
5	Aspirate		40 µl Water 500 SP Visc NO DETECT "a-150" (Col. 1, Rows 1-8) , 1 option
6	Dispense		40 µl Water 500 SP Visc NO DETECT "a-10" (Col. 1, Rows 1-8) , 1 option
7	Wash Tips		5 + 5 ml
8	End Loop	"Dispense the dilutee across the 10uM plate"	
9	Begin Loop	2 times "Repeat Water"	
10	Begin Loop	8 times "Dispense the dilutor across the 10uM plate"	
11	Aspirate		280 µl Water 500 SP in Trough "water" (Col. 1, Rows 1-8)
12	Dispense		280 µl Water 500 SP Visc NO DETECT "a-10" (Col. 1, Rows 1-8) , 1 option
13	Wash Tips		2 + 2 ml
14	End Loop	"Dispense the dilutor across the 10uM plate"	
15	End Loop	"Repeat Water"	
16	Worklist Import	C:\Documents and Settings\All Users\Evo_Worklists\Toni\1A53-8ng.csv C:\Documents and Settings\All Users\Evo_Worklists\Toni\1A53-8ng.gwl	
17	Worklist		Load Worklist "C:\Documents and Settings\All Users\Evo_Worklists\Toni\1A53-8ng.gwl" Water 500 SP Vis
18	Worklist	Execute loaded worklist(s)	

IPIPE and CPEC PCR (Gene Assembly)

*Current Deck Layout

The gene assembly scripts described here bring the PIPE/CPEC methods established on the benchtop to the robot. Chapter 4 features a full description of the methodology. Since the IPIPE and CPEC methods are very similar, they are grouped into this one entry.

The input to these scripts is a 96-well plate with mixed oligos. The IPIPE method will then perform assembly PCR followed by amplification with IPIPE outside primers. After the IPIPE method the PCR products can be combined with VPIPE products and transformed into bacteria. It is recommended, however, to continue with the CPEC method and use the resulting product to transform. Much better rates of transformed colonies have been found when using CPEC products over the PIPE reaction.

Materials

96-well BIO-RAD PCR Plate(s) on 96-well cooled block. This PCR plate, available from BIO-RAD (HSP9601), is preferred for automated operations due to its robot-friendly hard shell and full skirt. All automated thermocycler operations have been developed exclusively with this plate. The cooled block, along with the water recirculator, can cool the samples down to 4°C.

96-well V-Bottom Plate(s). These plates, available through Axygen (P-96-450V-C) were chosen for diluting oligos because of their similarity to the plates from IDT. They are ordered non-sterile and autoclaved before use.

Microtiter Plate Rubber Capmat. The 96-well V-Bottom plates are sealed with an Axygen ImpermaMat (AM-2ML-RD-IMP) to prevent evaporation. Sealed plates should be centrifuged down (500xg for 30 sec) prior to opening due to sample adhesion to the seal.

1.5mL eppendorf tubes in 24-well cooled block. These are the lab standard autoclavable tubes available in any biochemistry lab. The cooled block, along with the water recirculator can cool samples down to 4°C.

Mastermix Solution. These solutions vary due to their intended use. In general, the assembly mastermix should include DNA polymerase, dNTPs, water and DNA polymerase buffer. The IPIPE amplification mastermix should add the IPIPE primers to the assembly mastermix. The CPEC mastermix should add linearized vector (VPIPE product) to the assembly mastermix.

Methods

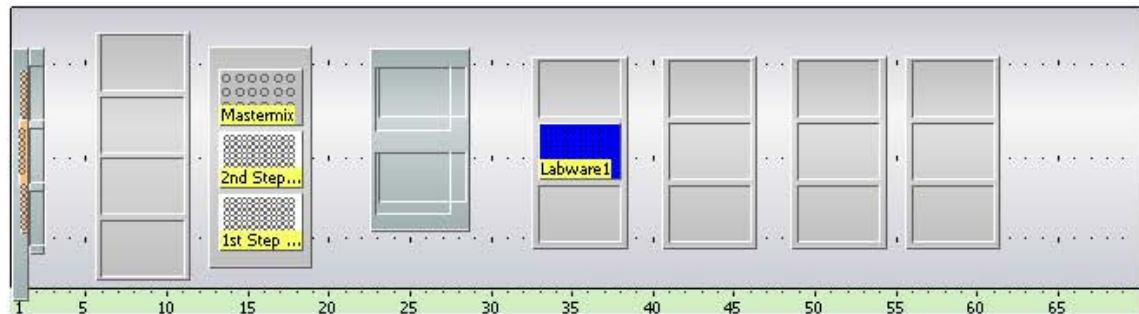
Setup - IPIPE

Turn on the water recirculator at least a half-hour before running the procedure to ensure the 24 and 96-well blocks are cool. Organize the deck as shown in the figure below. The script will transfer oligo mixtures from the V-bottom Plate into the 1st Step PCR Plate, and then add assembly mastermix on top of that. The gripper arm will transfer the plate to the thermalcycler, execute the program, and bring it back. A small amount of product is transferred from the 1st Step to the 2nd Step PCR Plate, IPIPE mastermix is added on top of that, and the thermalcycler is engaged again for the insert amplification.

1st Step, 2nd Step PCR Plate: 96-well BIO-RAD PCR Plates

Mastermix: 1.5mL eppendorf tubes in 24-well cooled block (split the assembly mastermix into C1,D1; split the IPIPE amplification mastermix into C2,D2)

Labware1: 96-well V-Bottom Plate

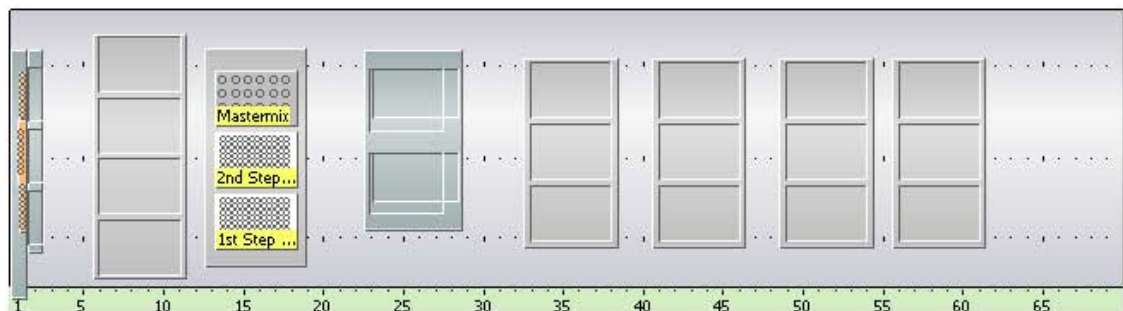


Setup - CPEC

Turn on the water recirculator at least a half-hour before running the procedure to ensure the 24 and 96-well blocks are cool. Organize the deck as shown in the figure below. A small amount of PCR product is transferred from the 1st Step to the 2nd Step PCR Plate and CPEC mastermix is added on top of that. The gripper arm will transfer the 2nd Step PCR Plate to the thermalcycler, execute the program, and bring it back.









1st Step, 2nd Step PCR Plate: 96-well BIO-RAD PCR Plates




Mastermix: 1.5mL eppendorf tubes in the 24-well cooled block (split the CPEC mastermix into C2,D2)




Procedure – IPIPE

1	Comment	Transfer Primers	
2	Begin Loop	12 times "Add Primers"	
3	Aspirate		5 µl Water 500 SP Visc NO DETECT "Labware1" (Col. 1, Rows 1-8) , 1 option
4	Dispense		5 µl Water 500 SP Visc NO DETECT "1st Step PCR Plate" (Col. 1, Rows 1-8) , 1 option
5	Wash Tips		5 + 5 ml
6	End Loop	"Add Primers"	
7	Comment	Add Mastermix	
8	Begin Loop	6 times "mastermix"	
9	Aspirate		20 µl Water 500 SP Visc "Mastermix" (Col. 1, Rows 3,4)
10	Aspirate		20 µl Water 500 SP Visc "Mastermix" (Col. 1, Rows 3,4)
11	Aspirate		20 µl Water 500 SP Visc "Mastermix" (Col. 1, Rows 3,4)
12	Aspirate		20 µl Water 500 SP Visc "Mastermix" (Col. 1, Rows 3,4)
13	Dispense		20 µl Water 500 SP Visc "1st Step PCR Plate" (Col. 1, Rows 1-8) , 1 option
14	Wash Tips		5 + 5 ml
15	End Loop	"mastermix"	
16	Begin Loop	6 times "mastermix"	
17	Aspirate		20 µl Water 500 SP Visc NO DETECT "Mastermix" (Col. 1, Rows 3,4)
18	Aspirate		20 µl Water 500 SP Visc NO DETECT "Mastermix" (Col. 1, Rows 3,4)
19	Aspirate		20 µl Water 500 SP Visc NO DETECT "Mastermix" (Col. 1, Rows 3,4)
20	Aspirate		20 µl Water 500 SP Visc NO DETECT "Mastermix" (Col. 1, Rows 3,4)
21	Dispense		20 µl Water 500 SP Visc "1st Step PCR Plate" (Col. 7, Rows 1-8) , 1 option
22	Wash Tips		5 + 5 ml

23	End Loop	"mastermix"
24	Comment	Move sdm1 to thermocycler
25	ThermalCycler	OpenLid(1)
26	Transfer Labware	Source: Grid '13,' Site '3'; Destination: Grid '63', Site '1'; Narrow (ROMA 1)
27	ThermalCycler	CloseLid(1,1)
28	Comment	Check program name
29	ThermalCycler	StartProgram(1,TONGA,BLOCK,ON,10)
30	ThermalCycler	WaitForProgram(1)
31	ThermalCycler	OpenLid(1)
32	Notification	Send email now Mayo Lab
33	Comment	Move sdm1 out
34	Transfer Labware	Source: Grid '63,' Site '1'; Destination: Grid '13', Site '3'; Narrow (ROMA 1)
35	Comment	Amplification
36	Begin Loop	12 times "add 1st step product"
37	Aspirate	  2 µl Water 500 SP Visc NO DETECT "1st Step PCR Plate" (Col. 1, Rows 1-8) , 1 option
38	Dispense	  2 µl Water 500 SP Visc NO DETECT "2nd Step PCR Plate" (Col. 1, Rows 1-8) , 1 option
39	Wash Tips	 5 + 5 ml
40	End Loop	"add 1st step product"
41	Comment	Dispense Mastermix Again
42	Begin Loop	6 times "mastermix"
43	Aspirate	  23 µl Water 500 SP Visc "Mastermix" (Col. 2, Rows 3,4)
44	Aspirate	  23 µl Water 500 SP Visc "Mastermix" (Col. 2, Rows 3,4)

45	Aspirate		23 µl Water 500 SP Visc "Mastermix" (Col. 2, Rows 3,4)
46	Aspirate		23 µl Water 500 SP Visc "Mastermix" (Col. 2, Rows 3,4)
47	Dispense		23 µl Water 500 SP Visc "2nd Step PCR Plate" (Col. 1, Rows 1-8) , 1 option
48	Wash Tips		5 + 5 ml
49	End Loop		"mastermix"
50	Begin Loop		6 times "mastermix"
51	Aspirate		23 µl Water 500 SP Visc NO DETECT "Mastermix" (Col. 2, Rows 3,4)
52	Aspirate		23 µl Water 500 SP Visc NO DETECT "Mastermix" (Col. 2, Rows 3,4)
53	Aspirate		23 µl Water 500 SP Visc NO DETECT "Mastermix" (Col. 2, Rows 3,4)
54	Aspirate		23 µl Water 500 SP Visc NO DETECT "Mastermix" (Col. 2, Rows 3,4)
55	Dispense		23 µl Water 500 SP Visc "2nd Step PCR Plate" (Col. 7, Rows 1-8) , 1 option
56	Wash Tips		5 + 5 ml
57	End Loop		"mastermix"
58	Comment		2nd Step in Thermocycler
59	ThermalCycler		OpenLid(1)
60	Transfer Labware		Source: Grid '13,' Site '2'; Destination: Grid '63', Site '1'; Narrow (ROMA 1)
61	ThermalCycler		CloseLid(1,1)
62	Comment		Check program name
63	ThermalCycler		StartProgram(1, IPIPE-MM, BLOCK, ON, 10)
64	ThermalCycler		WaitForProgram(1)
65	ThermalCycler		OpenLid(1)
66	Transfer Labware		Source: Grid '63,' Site '1'; Destination: Grid '13', Site '2'; Narrow (ROMA 1)
67	Notification		Send email now Mayo Lab

Procedure – CPEC

1	Wash Tips		5 + 5 ml
2	Comment	Amplification	
3	Begin Loop	12 times "add 1st step product"	
4	Aspirate		2 µl Water 500 SP Visc NO DETECT "1st Step PCR Plate" (Col. 1, Rows 1-8) , 1 option
5	Dispense		2 µl Water 500 SP Visc NO DETECT "2nd Step PCR Plate" (Col. 1, Rows 1-8) , 1 option
6	Wash Tips		5 + 5 ml
7	End Loop	"add 1st step product"	
8	Comment	Dispense Mastermix Again	
9	Begin Loop	6 times "mastermix"	
10	Aspirate		18 µl Water 500 SP Visc "Mastermix" (Col. 2, Rows 3,4)
11	Aspirate		18 µl Water 500 SP Visc "Mastermix" (Col. 2, Rows 3,4)
12	Aspirate		18 µl Water 500 SP Visc "Mastermix" (Col. 2, Rows 3,4)
13	Aspirate		18 µl Water 500 SP Visc "Mastermix" (Col. 2, Rows 3,4)
14	Dispense		18 µl Water 500 SP Visc "2nd Step PCR Plate" (Col. 1, Rows 1-8) , 1 option
15	Wash Tips		5 + 5 ml
16	End Loop	"mastermix"	
17	Begin Loop	6 times "mastermix"	
18	Aspirate		18 µl Water 500 SP Visc NO DETECT "Mastermix" (Col. 2, Rows 3,4)
19	Aspirate		18 µl Water 500 SP Visc NO DETECT "Mastermix" (Col. 2, Rows 3,4)
20	Aspirate		18 µl Water 500 SP Visc NO DETECT "Mastermix" (Col. 2, Rows 3,4)
21	Aspirate		18 µl Water 500 SP Visc NO DETECT "Mastermix" (Col. 2, Rows 3,4)
22	Dispense		18 µl Water 500 SP Visc "2nd Step PCR Plate" (Col. 7, Rows 1-8) , 1 option

23	Wash Tips	 5 + 5 ml
24	End Loop	"mastermix"
25	Comment	2nd Step in Thermocycler
26	ThermalCyder	OpenLid(1)
27	Transfer Labware	Source: Grid '13,' Site '2'; Destination: Grid '63,' Site '1'; Narrow (ROMA 1)
28	ThermalCyder	CloseLid(1,1)
29	Comment	Check program name
30	ThermalCyder	StartProgram(1,CPEC,BLOCK,ON,10)
31	ThermalCyder	WaitForProgram(1)
32	ThermalCyder	OpenLid(1)
33	Transfer Labware	Source: Grid '63,' Site '1'; Destination: Grid '13,' Site '2'; Narrow (ROMA 1)
34	Notification	Send email now Toni Lee

CPEC Transformation

*Current Deck Layout

Although an automated cell-free expression protocol would obviate the need for bacterial transformation, few assays have the prerequisite sensitivity to deal with the low levels of protein expression from cell-free extracts. In most cases, bacterial transformation and subsequent plating will be a necessary step in automated pipelines.

Competent cell manipulation was kept at a minimum to avoid disturbing the cells. The proper setting for the outgrowth step (2 hr shaking at room temperature) was determined empirically.

This method is updated from the older Bacterial Transformation script. Instead of shaking on the robot at room temperature, this method calls for off-robot shaking at 37°C that improves the number of colony forming units.

Materials

CPEC Product. This is typically carried over from the “CPEC PCR (Gene Assembly)” procedure in a 96-well BIO-RAD PCR plate.

1.5mL Eppendorf Tubes in 24-well Cooled Block. These are the lab standard autoclavable tubes available in any biochemistry lab. The cooled block, along with the water recirculator can cool samples down to 4°C.

BL21 Gold DE3 Competent Cells. These homemade chemically competent cells are very effective in transforming plasmid and nicked DNA and not very sensitive to variations in heatshock protocols. For one 96-well plate, thaw 2.4mL of competent cells from the -80°C freezer. Their preparation protocol is included in the appendix of this manual.

96-well BIO-RAD PCR Plate(s) on 96-Well Cooled Block. This PCR plate, available from BIO-RAD (HSP9601), is preferred for automated operations due to its robot-friendly hard shell and full skirt. All automated thermocycler operations have been developed exclusively with this plate. The cooled block, along with the water recirculator, can cool the samples down to 4°C.

96-well Costar Sterile Round Bottom Plate(s). This plate, available from Costar (3788), is preferred for automated operations due to its robot-friendly hard shell.

200uL Robot Tips. These automation tips are available from USA Scientific (TipONE #1188-1700). They are stacked with yellow inserts.

LB in 100mL Trough. These troughs, available from Tecan (10 613 049), only fit in the carrier located at grid 2. Fill with at least 15 mL of lab standard LB liquid media for the transformation of one 96-well plate. Keep warm before use for better results.

Methods

Setup

Turn on the water recirculator at least a half-hour before running the procedure to ensure the 24 and 96-well blocks are cool. Organize the deck as shown in the figure below. The script will transfer a small amount of CPEC product to the cooled PCR plate. Competent cells are aliquoted on top of this product, incubated for 15 min, and then the plate is heatshocked in the thermalcycler. Once the plate is returned, LB media is added to the cells as recovery media. The MCA96 arm then transfers the cells to a sterile 96-well plate for off-robot shaking at 37°C.

Cells: Competent cells split into 2 tubes in C1, D1 of the 24-well cooled block

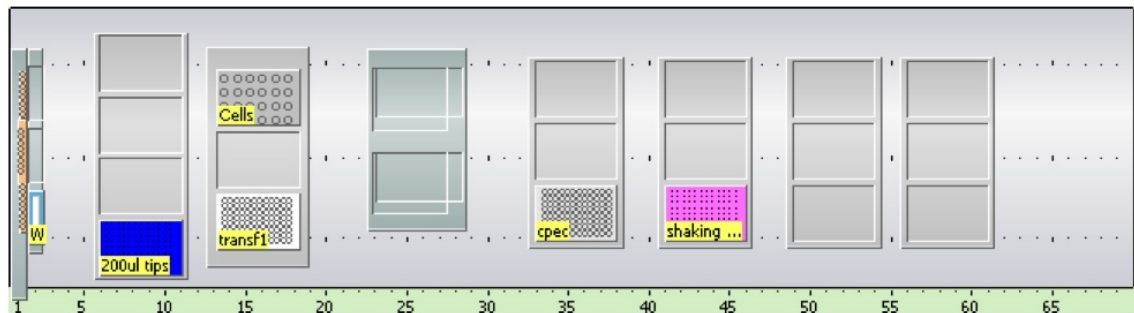
cpec: CPEC PCR product in 96-well BIO-RAD PCR Plate

transf1: 96-well BIO-RAD PCR plate on 96-well cooled block


























shaking plate: 96-well Costar sterile round bottom plate




200uL tips: 200uL disposable tips for the MCA96






Warm LB: LB in 100mL trough



Procedure

1	Comment	Check PCR Plate types
2	Wash Tips	 5 + 5 ml
3	Comment	Add dpn1 Product
4	Begin Loop	12 times "add cpec product"
5	Aspirate	  2 µl Water 500 SP Visc NO DETECT "cpec" (Col. 1, Rows 1-8) , 1 option
6	Dispense	  2 µl Water 500 SP Visc NO DETECT "transf1" (Col. 1, Rows 1-8) , 1 option
7	Wash Tips	 5 + 5 ml
8	End Loop	"add cpec product"
9	Comment	Add Cells
10	Begin Loop	6 times "add cells"
11	Aspirate	  18 µl Water 500 SP Visc "Cells" (Col. 1, Rows 3,4)
12	Aspirate	  18 µl Water 500 SP Visc "Cells" (Col. 1, Rows 3,4)
13	Aspirate	  18 µl Water 500 SP Visc "Cells" (Col. 1, Rows 3,4)
14	Aspirate	  18 µl Water 500 SP Visc "Cells" (Col. 1, Rows 3,4)
15	Dispense	  18 µl Water 500 SP Visc "transf1" (Col. 1, Rows 1-8) , 1 option
16	Wash Tips	 5 + 5 ml
17	End Loop	"add cells"
18	Begin Loop	6 times "add cells"
19	Aspirate	  18 µl Water 500 SP Visc NO DETECT "Cells" (Col. 1, Rows 3,4)
20	Aspirate	  18 µl Water 500 SP Visc NO DETECT "Cells" (Col. 1, Rows 3,4)
21	Aspirate	  18 µl Water 500 SP Visc NO DETECT "Cells" (Col. 1, Rows 3,4)
22	Aspirate	  18 µl Water 500 SP Visc NO DETECT "Cells" (Col. 1, Rows 3,4)

23	Dispense	 18 µl Water 500 SP Visc NO DETECT "transf1" (Col. 7, Rows 1-8) , 1 option
24	Wash Tips	 5 + 5 ml
25	End Loop	"add cells"
26	Comment	Wait 15 minutes for cells and product to incubate
27	Start Timer	1
28	Wait for Timer	Timer 1 : 900 sec
29	Comment	Heat Shock
30	ThermalCycler	OpenLid(1)
31	Transfer Labware	Source: Grid '13,' Site '3'; Destination: Grid '63,' Site '1'; Narrow (ROMA 1)
32	ThermalCycler	CloseLid(1,1)
33	ThermalCycler	StartProgram(1,AN-HS,BLOCK,ON,10)
34	ThermalCycler	WaitForProgram(1)
35	ThermalCycler	OpenLid(1)
36	Comment	Move Plate Out
37	Transfer Labware	Source: Grid '63,' Site '1'; Destination: Grid '13,' Site '3'; Narrow (ROMA 1)
38	Comment	Wait 1 minute to return back to 4 degrees
39	Start Timer	4
40	Wait for Timer	Timer 4 : 60 sec
41	Comment	Add warm LB
42	Begin Loop	3 times "finish LB addition"
43	Aspirate	 400 µl Water 500 SP Visc in Trough "Warm LB" (Col. 1, Rows 1-8)
44	Begin Loop	4 times "add LB over plate"

45	Dispense	  100 µl Water 500 SP Visc NO DETECT "transf1" (Col. 1, Rows 1-8) , 2 options
46	End Loop	"add LB over plate"
47	Wash Tips	 5 + 5 ml
48	End Loop	"finish LB addition"
49	Comment	Transfer to round bottom plate
50	Get DiTis	Grid 6; Site: 4 (DiTi 200ul MCA) Fetch 8 rows and 12 columns
51	Aspirate	 130 µl MCA96 trough "transf1" (Col. 1, Rows 1-8)
52	Dispense	 130 µl MCA96 trough "shaking plate" (Col. 1, Rows 1-8)
53	Drop DiTis	Back to Source
54	Comment	Move To Off Robot Shaker

Qtray Plating w/Sterile Beads

*Current Deck Layout

Directly after the bacterial transformation step is completed, the outgrowth cultures are plated onto 48-segmented Qtrays. This protocol is the most improvised procedure in the entire chapter because the Qtray is not a standard microtiter plate. Load the prepared Qtray (see appendix) with >4 sterile beads in each segment. Follow the placing diagram on the following page.

Materials

Transformed bacterial culture. This is carried over from the “CPEC Transformation” procedure as a 96-well Costar round bottom plate.

96-well Costar Sterile Round Bottom Plate(s). This plate, available from Costar (3788), is preferred for automated operations due to its robot-friendly hard shell.

Qtray(s). This specialty petri plate, available from Genetix (x6029), has 48 subdivisions upon which different bacterial cultures can be plated. The plate was chosen as it is the only input to the Genetix Qbot, a colony picking robot located on the 2nd floor of the Beckman Institute.

Methods

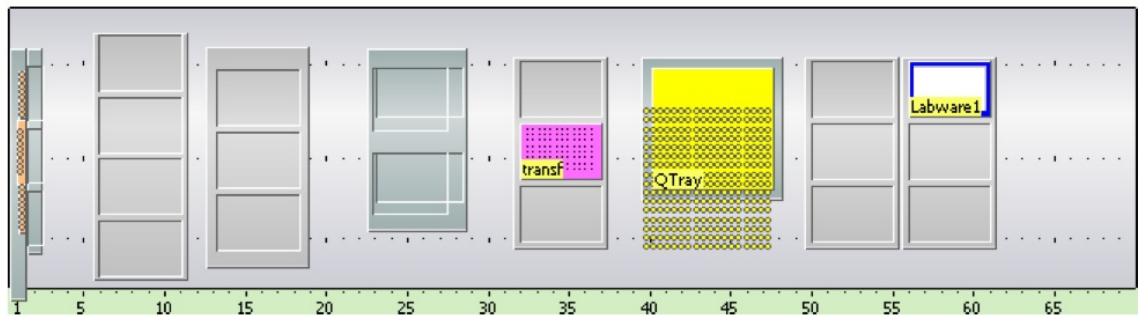
Setup

The script takes less than 5min to complete. Load a Qtray with sterile beads precisely as shown on the following page and organize the deck as seen in the figure below. After the procedure is finished cover the Qtray and shake it. Remove the beads and let the solutions air-dry under a flame. Replace the cover and incubate overnight.

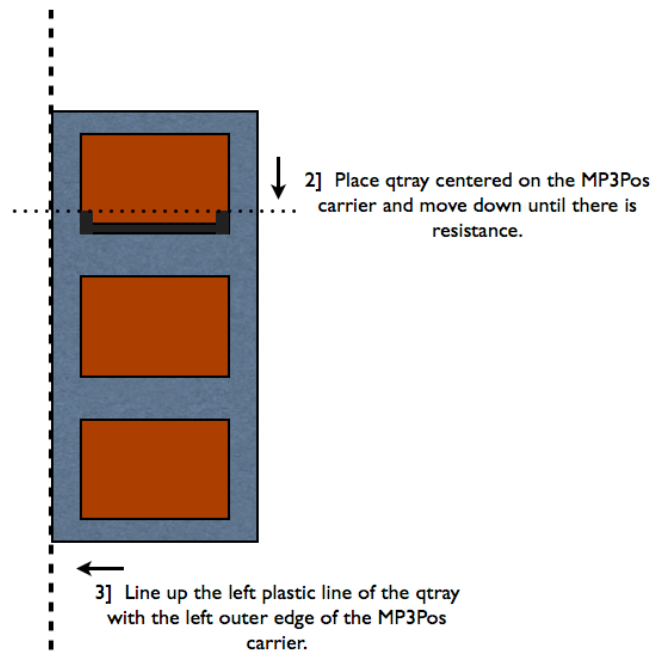
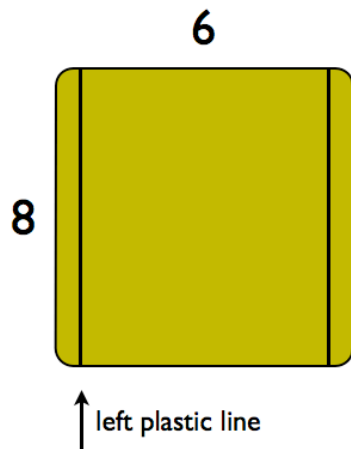
Qtray: 48-segment Qtray

transf: Transformed culture in a 96-well Costar round bottom plate

Labware1: Placeholder reference for arm movements, ignore



1] Orient qtray as shown and take note of the plastic lines along the bottom.



QTray Loading

Procedure

1	Comment	Move the MP 3Pos flat carriers that normally reside at 33 and 41 to 32 and 40
2	Group	Move Roma and MCA out of the way
3	ROMA Vector	Vector "AN Move RoMa Safe-Right", Grid 56, Site 1 (ROMA 1) open, from Safe to End Position
4	Move	 Positioning with global Z-Travel "Labware1" (Col. 1, Rows 1-8)
5	Group End	Move Roma and MCA out of the way
6	Comment	Plate (First 6 Columns)
7	Wash Tips	 5 + 5 ml
8	Begin Loop	6 times "whole"
9	Aspirate	  72 µl Water 500 SP Visc NO DETECT (Asp High) "transf" (Col. 1, Rows 1-8) , 1 option
10	Begin Loop	2 times "down"
11	Begin Loop	3 times "across"
12	Dispense	  12 µl Water 500 SP Visc NO DETECT "QTray" (Col. 1, Rows 1,3,5,7,9,11,13,15) , 3 options
13	End Loop	"across"
14	End Loop	"down"
15	Wash Tips	 10 + 10 ml
16	End Loop	"whole"
17	Comment	Plate (Last 6 Columns)
18	User Prompt	"Switch the Qtray" sound : once
19	Begin Loop	6 times "whole"
20	Aspirate	  72 µl Water 500 SP Visc NO DETECT (Asp High) "transf" (Col. 7, Rows 1-8) , 1 option
21	Begin Loop	2 times "down"
22	Begin Loop	3 times "across"

23	Dispense		12 µl Water 500 SP Visc NO DETECT "QTray" (Col. 1, Rows 1,3,5,7,9,11,13,15) , 3 options
24	End Loop	"across"	
25	End Loop	"down"	
26	Wash Tips		10 + 10 ml
27	End Loop	"whole"	

Dilute IDT Oligonucleotides

Most automation projects start with ordering oligos through Integrated DNA Technologies (IDT) to either generate site-directed mutants or assemble library variants. Luckily, oligos are much cheaper when specified to arrive on a 96-well plate. Of course, certain options such as PAGE purification are unavailable in the plate format, but that feature is prohibitively expensive when ordering the amount of oligos used in automation experiments anyway.

There are three main specification pages when ordering plate oligos from IDT. For the first page, mark “file upload,” “email,” “96-well plate,” and “column” loading. For the second page, mark “25 nmole DNA,” “standard desalting,” “no” CE service, “full yield,” “V-bottom plate,” and shipping “wet.” For the third page, mark “150 μ M,” leave the volume blank, and mark “IDTE buffer pH 7.5.”

These specifications allow users to start automation experiments much faster because there is no need for resuspension, and they keep the plate in a robot-friendly column-loading scheme. The lower limit for ordering oligos in the plate format is 24. A MS Excel template is available in the lab or from IDT to upload the sequence information.

Materials

96-Well IDT Oligonucleotide Plate(s). The oligo plates from IDT arrive with a rubber capmat and lid, and taped up. Dispose of the tape but keep the capmat and the lid.

96-Well V-Bottom Plate(s). These plates, available from Nunc (249944), were chosen for diluting oligos because of their similarity to the plates from IDT. They are ordered non-sterile and are autoclaved before use.

Sterile H₂O in 100mL Disposable Trough. This trough, available from Tecan (10 613 049), only fits in the carrier located at grid 2. Fill with standard molecular biology water for diluting the oligos.

Microtiter Plate Rubber Capmat. The 96-well V-bottom plates are sealed with an Axygen ImpermaMat (AM-2ML-RD-IMP) to prevent evaporation. Non-frozen sealed plates may need to be centrifuged down prior to opening due to sample adhesion to the seal.

Methods

Setup

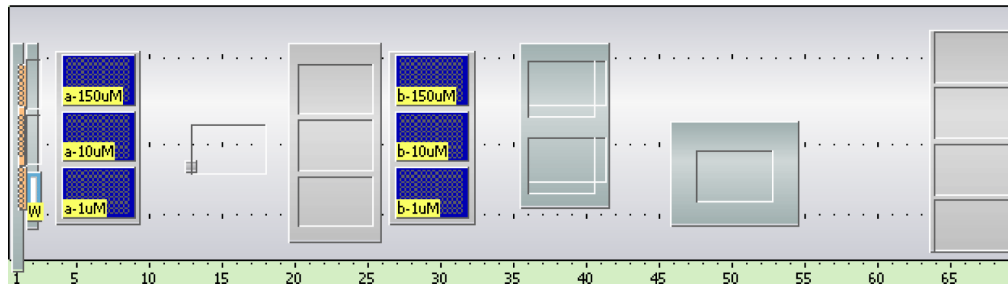
The template script is named “AN_Dilute_IDT_Oligos” and takes roughly 48 min to complete. Fill the water trough to at least 100mL. Organize the deck as shown in the figure below.

This template file creates 10 μ M and 1 μ M diluted plates from two original 150 μ M IDT plates. Volumes can be adjusted to dilute to concentrations more applicable to the user’s downstream procedures.














Water: 100ml disposable trough

a/b-150 μ M: 96-well IDT oligonucleotide plates

a/b-10 μ M, a/b-1 μ M: 96-well V-bottom plates



Procedure

1	Wash Tips		5 + 5 ml
2	Comment	***START 1st PLATE SET***	
3	Comment	Dilute primers into 10uM plate	
4	Begin Loop	12 times "Dispense the dilutee across the 10uM plate"	
5	Mix		3 x 80 µl LiHa Tip Mixing AN "a-150uM" (Col. 1, Rows 1-8) , 1 option
6	Aspirate		18 µl Water 500 SP Visc NO DETECT "a-150uM" (Col. 1, Rows 1-8) , 1 option
7	Dispense		18 µl Water 500 SP Visc NO DETECT "a-10uM" (Col. 1, Rows 1-8) , 1 option
8	Wash Tips		5 + 5 ml
9	End Loop	"Dispense the dilutee across the 10uM plate"	
10	Begin Loop	12 times "Dispense the dilutor across the 10uM plate"	
11	Aspirate		252 µl Water 500 SP in Trough "Water" (Col. 1, Rows 1-8)
12	Dispense		252 µl Water 500 SP "a-10uM" (Col. 1, Rows 1-8) , 1 option
13	Wash Tips		2 + 2 ml
14	End Loop	"Dispense the dilutor across the 10uM plate"	
15	Comment	Dilute primers into 1uM plate	
16	Begin Loop	12 times "Dispense the dilutee across the 1uM plate"	
17	Aspirate		25 µl Water 500 SP Visc NO DETECT "a-10uM" (Col. 1, Rows 1-8) , 1 option
18	Dispense		25 µl Water 500 SP Visc NO DETECT "a-1uM" (Col. 1, Rows 1-8) , 1 option
19	Wash Tips		5 + 5 ml
20	End Loop	"Dispense the dilutee across the 1uM plate"	
21	Begin Loop	12 times "Dispense the dilutor across the 1uM plate"	
22	Aspirate		225 µl Water 500 SP in Trough "Water" (Col. 1, Rows 1-8)
23	Dispense		225 µl Water 500 SP "a-1uM" (Col. 1, Rows 1-8) , 1 option
24	Wash Tips		2 + 2 ml
25	End Loop	"Dispense the dilutor across the 1uM plate"	

See Note 1

See Note 2

See Note 3

26	Comment	***START 2nd PLATE SET***
27	Comment	Dilute primers into 10uM plate
28	Begin Loop	12 times "Dispense the dilutee across the 10uM plate"
29	Mix	 3 x 80 µl LiHa Tip Mixing AN "b-150uM" (Col. 1, Rows 1-8) , 1 option
30	Aspirate	 18 µl Water 500 SP Visc NO DETECT "b-150uM" (Col. 1, Rows 1-8) , 1 option
31	Dispense	 18 µl Water 500 SP Visc NO DETECT "b-10uM" (Col. 1, Rows 1-8) , 1 option
32	Wash Tips	 5 + 5 ml
33	End Loop	"Dispense the dilutee across the 10uM plate"
34	Begin Loop	12 times "Dispense the dilutor across the 10uM plate"
35	Aspirate	 252 µl Water 500 SP in Trough "Water" (Col. 1, Rows 1-8)
36	Dispense	 252 µl Water 500 SP "b-10uM" (Col. 1, Rows 1-8) , 1 option
37	Wash Tips	 2 + 2 ml
38	End Loop	"Dispense the dilutor across the 10uM plate"
39	Comment	Dilute primers into 1uM plate
40	Begin Loop	12 times "Dispense the dilutee across the 1uM plate"
41	Aspirate	 25 µl Water 500 SP Visc NO DETECT "b-10uM" (Col. 1, Rows 1-8) , 1 option
42	Dispense	 25 µl Water 500 SP Visc NO DETECT "b-1uM" (Col. 1, Rows 1-8) , 1 option
43	Wash Tips	 5 + 5 ml
44	End Loop	"Dispense the dilutee across the 1uM plate"
45	Begin Loop	12 times "Dispense the dilutor across the 1uM plate"
46	Aspirate	 225 µl Water 500 SP in Trough "Water" (Col. 1, Rows 1-8)
47	Dispense	 225 µl Water 500 SP "b-1uM" (Col. 1, Rows 1-8) , 1 option
48	Wash Tips	 2 + 2 ml
49	End Loop	"Dispense the dilutor across the 1uM plate"
50	Notification	Send email now Mayo Lab

See Note 4

Notes

1. All mixing steps in the scripts are based on there being at least 150 μ L in each well. Adjust the volume if this is not the case.
2. All “1 option” loop tags instruct the LiHa to move over 1 column in every cycle of the loop.
3. If only 1 plate needs to be diluted, highlight script lines 1 to 25. Alternatively, delete all script lines after line 25.
4. At the end of the dilution script there is a command to send an email to the user. This is useful since the script runs longer than 15 min.

Site-Directed Mutagenesis

The procedure outlined here is based on the article, “A novel megaprimed and ligase-free, PCR-based, site-directed mutagenesis method” by Tseng *et al.*, *Anal. Biochem*, 2008. This method was chosen and optimized as an automated protocol for the total site-saturation mutagenesis of the β 1 domain of protein G. More details on the background of the method can be found in Chapter 3.

Although this site-directed mutagenesis (SDM) protocol was initially developed using New England Biolab (NEB) Hot Start Phusion DNA polymerase, the user can easily change the mastermix and appropriate dispense volumes. The thermalcycler program must be entered directly using the thermalcycler number pad.

Materials

Diluted Oligonucleotide Plate(s). Typically, the diluted mutagenic oligos are in 96-well V-bottom plates at a concentration of $1\mu\text{M}$. These are prepared by running the “Dilute IDT Oligonucleotides” protocol.

96-Well BIO-RAD PCR Plate(s) on 96-Well Cooled Block. This PCR plate, available from BIO-RAD (HSP9601), is preferred for automated operations due to its robot-friendly hard shell and full skirt. All automated thermocycler operations have been developed exclusively with this plate. The cooled block, along with the water recirculator, can cool the samples down to 4°C .

1.5mL Eppendorf Tubes in 24-Well Cooled Block. These are the lab standard autoclavable tubes available in any biochemistry lab. The cooled block, along with the water recirculator, can cool samples down to 4°C .

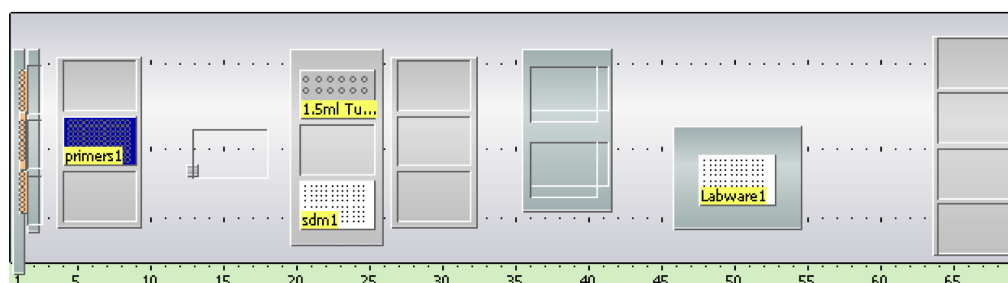
Mastermix Solution. For one 96-well plate, prepare (1) $520\mu\text{L}$ of $5\times$ Phusion HF buffer, (2) $52\mu\text{L}$ of $10\mu\text{M}$ reverse primer, (3) $52\mu\text{L}$ of 10mM each dNTPs, (4) $13\mu\text{L}$ of 50mM MgSO_4 , (5) $13\mu\text{L}$ of $160\text{ng}/\mu\text{L}$ plasmid template, (6) $16.25\mu\text{L}$ of Phusion Hot Start DNA Polymerase, and (7) $1430\mu\text{L}$ of sterile H_2O . This is a $130\times$ mix (total volume: 2.1mL) of a $16\mu\text{L}$ recipe that is combined with $4\mu\text{L}$ of $1\mu\text{M}$ forward primer to obtain a final volume of $20\mu\text{L}$ in each well.

Methods

Setup

The template script is named “AN_SDM_1PLATE” and takes roughly 2.5 hr to complete, including 2 hr and 12 min for the thermalcycler protocol. Turn on the water recirculator at least 30 min before running the procedure to ensure the 24- and 96-well blocks are cool. Evenly split the 2.1 mL mastermix into two 1.5 mL eppendorf tubes. Organize the deck as shown in the figure below. You can ignore the labware with generic names (Labware1, Labware2, etc.) as they are only placeholders for the software.

1.5mL tubes: SDM mix split into two tubes in C1, D1 of the 24-well block
 sdm1: 96-well BIO-RAD PCR plate
 primers1: 96-well V-bottom plate



Troubleshooting

Even after optimizing the melting temperature (T_m) of an oligo, transformable plasmid may not have been generated after the SDM procedure. In that case, several variables can be adjusted. One could vary the 1st step annealing temperature and increase the concentration of template, $MgSO_4$, or enzyme. Of course, there may be negative consequences to changing some of these parameters (see Table 2-1 on the next page).


















The mastermix components and thermalcycler program for a 96-well plate SDM are listed below. The user may use these values to calculate the appropriate amount of mastermix to prepare, depending on the number of columns and plates to be processed.

130×	1×	Mastermix Component	AN-SDM1
520	4	5× HF Phusion buffer	1. 98°C, 30 sec
52	0.4	reverse primer (10 μ M)	2. 98°C, 6 sec
52	0.4	dNTPs (10mM each)	3. 64°C, 15 sec
13	0.1	template (160ng/ μ L)	4. 72°C, 20 sec (go to step 2 for 10×)
13	0.1	$MgSO_4$ (50mM)	5. 98°C, 6 sec
16.25	0.125	HS Phusion enzyme	6. 72°C, 3 min (go to step 5 for 25×)
1430	11	water	7. 72°C, 10 min

Table 2-1. SDM Troubleshooting.

Condition	Pro	Con
1st step anneal (°C)	A temperature gradient may reveal a condition that improves the generation of 1 st step product	None
Increase template	Generates more 2 nd step product (transformable plasmid)	Increases wild-type recovery during sequence verification
Increase MgSO₄	Generates more 1 st and 2 nd step product	May lead to non-specific banding and/or incorrect products
Increase enzyme	Generates more 1 st and 2 nd step product	May lead to non-specific banding and/or incorrect products

Procedure

	1	Wash Tips	 5 + 5 ml
	2	Comment	***SETUP PLATE 1***
See Note 1	3	Comment	Transfer SDM (2 tubes of 1.05mL each = 1 96 well plate)
	4	Aspirate	 96 µl Water 500 SP Visc "1.5ml Tubes" (Col. 1, Rows 3,4)
	5	Aspirate	 96 µl Water 500 SP Visc "1.5ml Tubes" (Col. 1, Rows 3,4)
	6	Aspirate	 96 µl Water 500 SP Visc "1.5ml Tubes" (Col. 1, Rows 3,4)
	7	Aspirate	 96 µl Water 500 SP Visc "1.5ml Tubes" (Col. 1, Rows 3,4)
See Note 2	8	Begin Loop	6 times "mastermix1-6"
	9	Dispense	 16 µl Water 500 SP Visc "sdm1" (Col. 1, Rows 1-8) , 1 option
	10	End Loop	"mastermix1-6"
	11	Wash Tips	 5 + 5 ml
	12	Aspirate	 96 µl Water 500 SP Visc "1.5ml Tubes" (Col. 1, Rows 3,4)
	13	Aspirate	 96 µl Water 500 SP Visc "1.5ml Tubes" (Col. 1, Rows 3,4)
	14	Aspirate	 96 µl Water 500 SP Visc "1.5ml Tubes" (Col. 1, Rows 3,4)
	15	Aspirate	 96 µl Water 500 SP Visc "1.5ml Tubes" (Col. 1, Rows 3,4)
	16	Begin Loop	6 times "mastermix7-12"
	17	Dispense	 16 µl Water 500 SP Visc "sdm1" (Col. 7, Rows 1-8) , 1 option
	18	End Loop	"mastermix7-12"
	19	Wash Tips	 5 + 5 ml
	20	Comment	Transfer Oligos from primers1
See Note 3	21	Begin Loop	12 times "p1primers"
	22	Aspirate	 4 µl Water 500 SP Visc NO DETECT "primers1" (Col. 1, Rows 1-8) , 1 option
	23	Dispense	 4 µl Water 500 SP Visc NO DETECT "sdm1" (Col. 1, Rows 1-8) , 1 option
See Note 4	24	Mix	 2 x 12 µl LiHa Tip Mixing AN "sdm1" (Col. 1, Rows 1-8) , 1 option
	25	Wash Tips	 5 + 5 ml
	26	End Loop	"p1primers"
	27	Comment	Move sdm1 to thermocycler
	28	ThermalCycler	OpenLid(1)

29	Transfer Labware	Source: Grid '20,' Site '3'; Destination: Grid '46', Site '1'; User defined (Narrow' (ROMA 1)
30	ThermalCycler	CloseLid(1,1)
31	Comment	Check program name
32	ThermalCycler	StartProgram(1,AN-SDM1,BLOCK,ON,10)
33	Comment	program will run for ~2hr
34	Group	Put lid on primers1
35	Comment	Put on the lid for overnight
36	ROMA Vector	Vector "Hotel 9Pos Microplate_Narrow_1", Grid 58, Site 1 (ROMA 1) open, from Safe to End Position, grip, from End to Safe Position
37	ROMA Vector	Vector "MP 3Pos Flat: Lid Grip", Grid 4, Site 2 (ROMA 1) move from Safe to End Position, open, from End to Safe Positior
38	Group End	Put lid on primers1
39	ThermalCycler	WaitForProgram(1)
40	Comment	Move sdm1 out
41	ThermalCycler	OpenLid(1)
42	Transfer Labware	Source: Grid '46,' Site '1'; Destination: Grid '20', Site '3'; User defined (Narrow' (ROMA 1)
43	Group	Put lid on sdm1
44	Comment	Put on the lid for overnight
45	ROMA Vector	Vector "Hotel 9Pos Microplate_Narrow_1", Grid 58, Site 2 (ROMA 1) open, from Safe to End Position, grip, from End to Safe Position
46	ROMA Vector	Vector "MP 3Pos Cooled-Lid Grip: PCR plate", Grid 20, Site 3 (ROMA 1) move from Safe to End Position, open, from End to Safe Positior
47	Group End	Put lid on sdm1
48	Notification	Send email now Mayo Lab

See Note 5

See Note 6

See Note 7

See Note 8

Notes

1. The amount of mastermix placed in each tube is key to the script moving forward because the aspirate commands use liquid detection. If it measures an insufficient volume, the script will stop and ask for input from the user. If this happens, click “Go to Z-max,” which will move the tip to the bottom of the tube to aspirate.
2. Script lines 8 and 16 control how many columns of the sdm plate the robot will dispense into. For example, if you wish to only dispense into 9 columns and not 12, change script line 16 to 3 loops. Don’t forget to change the aspiration volumes of lines 12–15 to reflect the smaller amount of mastermix needed. Continuing the example, each aspiration volume would be changed to 48 μ L.
3. Script line 21 controls how many columns of the sdm plate to fill and mix with oligos from the oligo plate. For example, if you only had 9 columns of oligos, change the number in line 21 to 9 loops.
4. The liquid class “LiHa Tip Mixing AN” does a satisfactory job in mixing the oligos into the mastermix. The volume to mix is based on a 20 μ L total volume; adjust this if the script is being changed to accommodate larger total reaction volumes.
5. The Transfer Labware commands used in lines 29 and 42 employ custom made vectors that handle the tricky BIO-RAD PCR plates. The plates are difficult for the RoMa to work with primarily because of the restricted space around the thermalcycler loading area.
6. The actual thermalcycler program call occurs here on line 32. Adjust the program name as necessary.
7. Grouped commands on lines 34–38 and 43–47 pick up a manufacturer agnostic lid (it can fit plates from Greiner, Costar, Nunc, etc.) from sites 1 and 2 of the 9pos microplate hotel and place them on the oligo plate and sdm plate. These lines are useful if this script is going to be run overnight.
8. At the end of the mutagenesis script, there is a command to send an email to the user. This is useful since the script runs longer than 15 min.

Dpn1 Digestion

The Dpn1 digestion step is necessary in the site-directed mutagenesis automated pipeline to remove the wild-type template plasmid. A more efficient digestion step will directly lead to decreased wild-type contamination in the sequencing results and improved recovery of the mutant of interest.

A traditional bench PCR digestion step maintains the Dpn1 enzyme concentration at 2–4% of the reaction mixture. The Dpn1 mix described in the Materials below keeps the enzyme concentration at 8% to digest as much methylated and hemi-methylated wild-type DNA as possible.

The default thermalcycler protocol is 37°C for 2 hr. This can probably be adjusted to 30 min, or even decreased to 5 min if the manufacturer's (NEB) advertising is true.

Materials

Site-Directed Mutagenesis DNA Product on 96-Well Cooled Block. This is typically carried over from the “Site-Directed Mutagenesis” procedure as a 96-well BIO-RAD PCR plate. The cooled block, along with the water recirculator, can cool the samples down to 4°C.

1.5mL Eppendorf Tubes in 24-Well Cooled Block. These are the lab standard autoclavable tubes available in any biochemistry lab. The cooled block, along with the water recirculator, can cool samples down to 4°C.

Dpn1 Mix. For site-directed mutagenesis product in 10µL aliquots, prepare (1) 175µL of 10× buffer 4, (2) 140µL of Dpn1 enzyme, and (3) 185µL of Diluent B for one 96-well plate. When 4µL of this mix is added to 10µL PCR product, buffer 4 and the Dpn1 enzyme attain effective concentrations of 1× and 8%, respectively.

50µL Robot Filter Tips. These automation tips are the ART BioRobotix tips available from Molecular BioProducts (#906-021). They are individually wrapped in green boxes.

Methods

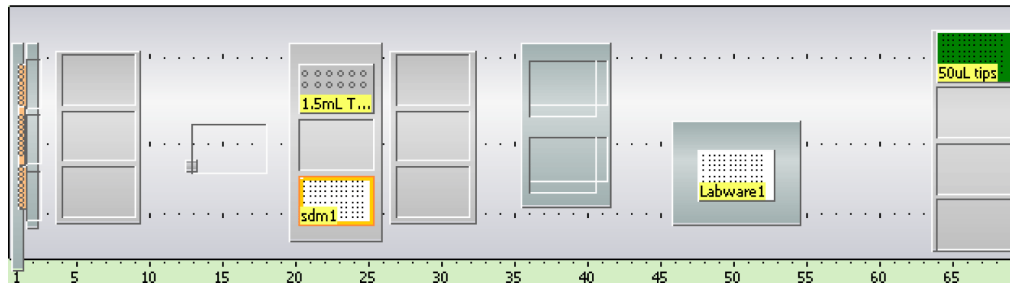
Setup

The template script is named “AN_Dpn1_1PLATE” and takes roughly 2.5 hr to complete, including the 2 hr digestion time. Turn on the water recirculator at least 30 min before running the procedure to ensure the 24- and 96-well blocks are cool. Organize the deck as shown in the figure below. You can ignore the labware with generic names (Labware1, Labware2, etc.), as they are only placeholders for the software.

1.5mL tubes: Dpn1 mix split into two tubes in C1, D1 of the 24-well block

sdm1: 96-well BIO-RAD PCR plate

50µL tips: 50µL robot filter tips (black box)






Advanced Technique:

The Dpn1 digest protocol can be tagged onto a trusted SDM procedure to save time. In addition, a few different techniques can be used to get the SDM samples down to 10µL for efficient digestion. First, 10µL of each sample is needed to run the HT DNA Agarose Gel Electrophoresis procedure. This puts the surplus volume to good use. Second, a more complicated yet time efficient procedure is to aspirate 10µL with the MCA96, dispense that volume into a waste trough, use LiHA multi-pipetting to deliver the Dpn1 mix, and finally mix the samples with the same 50µL robot filter tips used in the first aspirate. This will decrease the procedure’s time by about 25 min.

Procedure

1	Wash Tips	 5 + 5 ml	
2	Comment	Pipet out sdm product, leaving only 10uL	See Note 1
3	Group	MCA96 commands	
4	Group	Move_MCA_Safe_to_Right_Back_Corner	See Note 2
5	Command	"W1PAA,,2362," wait	
6	Command	"W1PAA,10,," wait	
7	Move Vector	Vector "Safe-RightBackCorner", Grid 64, Site 1(MCA 1) move from Safe to End Positior	
8	Group End	Move_MCA_Safe_to_Right_Back_Corner	
9	Get DiTis	Grid 64; Site: 1 (DiTi 50ul Filter MCA) Fetch 8 rows and 12 columns	
10	Move	 Positioning with global Z-Trave "1.5mL Tubes" (Col. 1, Rows 1-4)	See Note 3
11	Aspirate	 10 µl MCA96 trough "sdm1" (Col. 1, Rows 1-8)	
12	Move	 Positioning with global Z-Trave "1.5mL Tubes" (Col. 1, Rows 1-4)	
13	Drop DiTis	Back to Source	
14	Group End	MCA96 commands	
15	Comment	***Dpn1 DIGEST PLATE 1***	
16	Comment	2 tubes of .25ml = 1 96 well plate	
17	Begin Loop	12 times "dpn1addmix"	
18	Aspirate	  4 µl Water 500 SP Visc "1.5mL Tubes" (Col. 2, Rows 3,4)	
19	Aspirate	  4 µl Water 500 SP Visc "1.5mL Tubes" (Col. 2, Rows 3,4)	
20	Aspirate	  4 µl Water 500 SP Visc "1.5mL Tubes" (Col. 2, Rows 3,4)	
21	Aspirate	  4 µl Water 500 SP Visc "1.5mL Tubes" (Col. 2, Rows 3,4)	
22	Dispense	  4 µl Water 500 SP Visc NO DETECT "sdm1" (Col. 1, Rows 1-8) , 1 option	

See Note 4	23	Mix	  3 x 6 µl LiHa Tip Mixing AN "sdm1" (Col. 1, Rows 1-8) , 1 option
	24	Wash Tips	 10 + 10 ml
	25	End Loop	"dpm1addmix"
	26	Comment	Move sdm1 out for incubation
	27	ThermalCycler	OpenLid(1)
See Note 5	28	Transfer Labware	Source: Grid '20,' Site '3'; Destination: Grid '46,' Site '1'; User defined (Narrow. (ROMA 1)
	29	ThermalCycler	CloseLid(1,1)
	30	Comment	Check program name
See Note 6	31	ThermalCycler	StartProgram(1,AN-37INC,BLOCK,ON,10)
	32	ThermalCycler	WaitForProgram(1)
	33	ThermalCycler	OpenLid(1)
	34	Transfer Labware	Source: Grid '46,' Site '1'; Destination: Grid '20,' Site '3'; User defined (Narrow. (ROMA 1)
See Note 7	35	Comment	Put on the lid for overnight
	36	ROMA Vector	Vector "Hotel 9Pos Microplate_Narrow_1", Grid 58, Site 1 (ROMA 1) open, from Safe to End Position, grip, from End to Safe Position
	37	ROMA Vector	Vector "MP 3Pos Cooled-Lid Grip: PCR plate", Grid 20, Site 3 (ROMA 1) move from Safe to End Position, open, from End to Safe Positior
See Note 8	38	Notification	Send email now Mayo Lab

Notes

1. The Group “MCA96 commands,” encompassing lines 3–14, instruct the robot to pick up tips, aspirate 10 μ L from the “sdm1” plate, and drop off the tips (along with the volume) back into the tip box. There is no need to dispense the 10 μ L volume because the tips (and tip box) will not be used for anything else in this script.
2. The Group “Move_MCA_Safe_to_Right_Back_Corner” is a set of commands that orient the MCA96 into a safe position starting from anywhere on the deck. This helps to avoid crashes when executing the “Get DiTi” command.
3. Lines 10 and 12 instruct the MCA96 to move to a far site so that it won’t collide with the thermalcycler as it moves between the MP3Pos carriers and its DiTi carrier.
4. The liquid class “LiHa Tip Mixing AN” does a satisfactory job in mixing the samples with the Dpn1 enzyme. The volume to mix is based on a 14 μ L total volume. Adjust this if the script is being changed to accommodate larger total volumes of product and Dpn1 enzyme.
5. The Transfer Labware commands used in lines 28 and 34 employ custom made vectors that handle the tricky BIO-RAD PCR plates. The plates are difficult for the RoMa to work with primarily because of the restricted space around the thermalcycler loading area.
6. The actual thermalcycler program call occurs here on line 31. Adjust the program name as necessary.
7. Lines 36 and 37 pick up a manufacturer agnostic lid (it can fit plates from Greiner, Costar, Nunc, etc.) from site 1 of the 9pos microplate hotel and place it on the newly Dpn1 digested SDM product plate. These lines are useful if this script is going to be run overnight.
8. At the end of the digestion script, there is a command to send an email to the user. This is useful since the script runs longer than 15 min.

Bacterial Transformation

Although an automated cell-free expression protocol would obviate the need for bacterial transformation, few assays have the prerequisite sensitivity to deal with the low levels of protein expression from cell-free extracts. In most cases, bacterial transformation and subsequent plating will be a necessary step in automated pipelines.

Competent cell manipulation was kept at a minimum to avoid disturbing the cells. The proper setting for the outgrowth step (2 hr shaking at room temperature) was determined empirically.

Currently (3/24/10), the RoMA is having difficulty loading the PCR plate onto the Te-Shake. We therefore recommend running the script to line 47, manually loading the PCR plate, and then running the remainder of the script after line 49.

Materials

Dpn1 Digested Product. This is typically carried over from the “Dpn1 Digestion” procedure as a 96-well BIO-RAD PCR plate.

1.5mL Eppendorf Tubes in 24-Well Cooled Block. These are the lab standard autoclavable tubes available in any biochemistry lab. The cooled block, along with the water recirculator, can cool samples down to 4°C.

BL21 Gold DE3 Competent Cells. These homemade chemically competent cells are very effective in transforming plasmid and nicked DNA and are not very sensitive to variations in heatshock protocols. For one 96-well plate, thaw 2.2mL of competent cells from the -80°C freezer. Their preparation protocol is included in the appendix of this manual.

96-Well BIO-RAD PCR Plate(s) on 96-Well Cooled Block. This PCR plate, available from BIO-RAD (HSP9601), is preferred for automated operations due to its robot-friendly hard shell and full skirt. All automated thermocycler operations have been developed exclusively with this plate. The cooled block, along with the water recirculator, can cool the samples down to 4°C.

LB in 100mL Trough. These troughs, available from Tecan (10 613 049), only fit in the carrier located at grid 2. Fill with at least 15mL of lab standard LB liquid media for the transformation of one 96-well plate. Keep warm before use for better results.

Methods

Setup

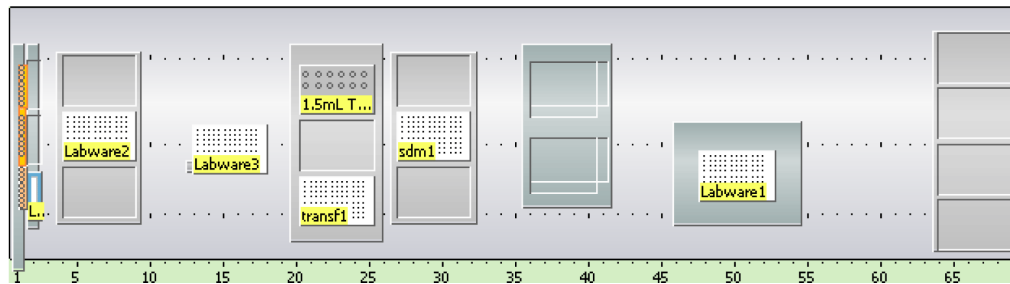
The template script is named “AN_Transformation_1Plate” and takes roughly 2.5 hr to run, including the 2 hr shaking outgrowth period. Turn on the water recirculator at least 30 min before running the procedure to ensure the 24- and 96-well blocks are cool. Organize the deck as shown in the figure below. You can ignore the labware with generic names (Labware1, Labware2, etc.), as they are only placeholders for the software.

1.5mL tubes: Competent cells split into 2 tubes in C1, D1 of the 24-well block

sdm1: Dpn1 digested product





























transf1: 96-well BIO-RAD PCR plate on 96-well cooled block




LB: LB in 100mL trough



Remember, as of 3/24/10, run the script to line 47, manually load the PCR plate onto the Te-Shake, and then run the rest of the script starting from line 49.

Procedure

	1	Wash Tips	 5 + 5 ml
	2	Comment	***Add Dpn1 product to chilled plate***
See Note 1	3	Begin Loop	12 times "add dpn1 product"
	4	Aspirate	  2 µl Water 500 SP Visc NO DETECT (Asp High) "sdm1" (Col. 1, Rows 1-8) , 1 option
	5	Dispense	  2 µl Water 500 SP Visc NO DETECT "transf1" (Col. 1, Rows 1-8) , 1 option
	6	Wash Tips	 5 + 5 ml
	7	End Loop	"add dpn1 product"
	8	Comment	***Add competent cells to plate 1***
See Note 2	9	Comment	Use 2 tubes of 1.1 ml of cells each per 96 well plate
	10	Aspirate	  108 µl Water 500 SP Visc "1.5mL Tubes" (Col. 1, Rows 3,4)
	11	Aspirate	  108 µl Water 500 SP Visc "1.5mL Tubes" (Col. 1, Rows 3,4)
	12	Aspirate	  108 µl Water 500 SP Visc "1.5mL Tubes" (Col. 1, Rows 3,4)
	13	Aspirate	  108 µl Water 500 SP Visc "1.5mL Tubes" (Col. 1, Rows 3,4)
	14	Begin Loop	6 times "add cells 1-6"
	15	Dispense	  18 µl Water 500 SP Visc "transf1" (Col. 1, Rows 1-8) , 1 option
	16	End Loop	"add cells 1-6"
	17	Wash Tips	 5 + 5 ml
	18	Aspirate	  108 µl Water 500 SP Visc "1.5mL Tubes" (Col. 1, Rows 3,4)
	19	Aspirate	  108 µl Water 500 SP Visc "1.5mL Tubes" (Col. 1, Rows 3,4)
	20	Aspirate	  108 µl Water 500 SP Visc "1.5mL Tubes" (Col. 1, Rows 3,4)
	21	Aspirate	  108 µl Water 500 SP Visc "1.5mL Tubes" (Col. 1, Rows 3,4)
	22	Begin Loop	6 times "add cells 7-12"
	23	Dispense	  18 µl Water 500 SP Visc "transf1" (Col. 7, Rows 1-8) , 1 option
	24	End Loop	"add cells 7-12"
	25	Wash Tips	 5 + 5 ml
See Note 3	26	Comment	Wait 15 minutes for cells and product to incubate
	27	Start Timer	1
	28	Wait for Timer	Timer 1 : 900 sec

29	Comment	***Heat Shock Plate 1***	
30	ThermalCycler	OpenLid(1)	
31	Transfer Labware	Source: Grid '20,' Site '3'; Destination: Grid '46,' Site '1'; User defined (Narrow) (ROMA 1)	See Note 4
32	ThermalCycler	CloseLid(1,1)	
33	ThermalCycler	StartProgram(1,AN-HS,BLOCK,ON,10)	See Note 5
34	ThermalCycler	WaitForProgram(1)	
35	ThermalCycler	OpenLid(1)	
36	Comment	move plate to left carrier	
37	Transfer Labware	Source: Grid '46,' Site '1'; Destination: Grid '4,' Site '2'; User defined (Narrow) (ROMA 1)	
38	User Prompt	"add warm LB to trough" sound : no	See Note 6
39	Comment	add warm LB to plate 1	
40	Begin Loop	3 times "add LB over plate"	
41	Aspirate	 400 µl Water 500 SP Visc in Trough "LB" (Col. 1, Rows 1-8)	
42	Begin Loop	4 times "add LB 4x"	
43	Dispense	 100 µl Water 500 SP Visc NO DETECT "Labware2" (Col. 1, Rows 1-8) , 2 options	
44	End Loop	"add LB 4x"	
45	Wash Tips	 5 + 5 ml	
46	End Loop	"add LB over plate"	

See Note 7

47	Comment	***Shake the plate for 2 hrs***
48	Transfer Labware	Source: Grid '4,' Site '2'; Destination: Grid '13', Site '1'; Narrow (ROMA 1)
49	Te-Shake Shaker	SetFrequency(1800)
50	Begin Loop	12 times "shakeit"
51	Te-Shake Shaker	Start(1)
52	Start Timer	2
53	Wait for Timer	Timer 2 : 60 sec
54	Te-Shake Shaker	Stop()
55	Start Timer	3
56	Wait for Timer	Timer 3 : 180 sec
57	Te-Shake Shaker	Start(1)
58	Start Timer	2
59	Wait for Timer	Timer 2 : 60 sec
60	Te-Shake Shaker	Stop()
61	Start Timer	4
62	Wait for Timer	Timer 4 : 300 sec
63	End Loop	"shakeit"
64	Notification	Send email now Mayo Lab

See Note 8

Notes

1. Script line 3 controls how many columns of the “sdm1” plate to add to the “transf1” plate. For example, if you only had 9 columns of digested product, change the number in line 3 to 9 loops.
2. The amount of volume in the 1.5mL tubes is key to the script proceeding properly. If there is an input error, hit “Move tips to Z-Max” to aspirate from the bottom of the tube. Script lines 14 and 22 control how many columns of the “transf1” plate the robot will dispense into. For example, if you wish to only dispense into 9 columns and not 12, change script line 22 to 3 loops. Don’t forget to change the aspiration volumes of lines 10–13 and 18–21 to reflect the smaller amount of competent cells needed. Continuing the example, change each aspiration volume to 54 μ L.
3. Common in most transformation protocols, script lines 26–28 incubate the plate for 15 min at 4°C.
4. The Transfer Labware commands used in lines 31 and 37 employ custom made vectors that handle the tricky BIO-RAD PCR plates. The plates are difficult for the RoMa to work with primarily because of the restricted space around the thermalcycler loading area.
5. The actual thermalcycler program call occurs here on line 33. Adjust the program name as necessary.
6. Script line 38 is a user prompt that stops the script and only continues when a user clicks OK in the software. Ideally, this is when the user will take warm LB and add it to the 100mL trough.
7. The Te-Shake is addressed by a Start command, then a timer, and finished by a Stop command. In lines 49–63, the Te-Shake’s frequency is set (1800 is the maximum), followed by commands for 12 cycles of 1 min shake, 3 min rest, 1 min shake, and 5 min rest. This was programmed to keep the cells suspended in media over 2 hr.
8. At the end of the transformation script there is a command to send an email to the user. This is useful since the script runs longer than 15 min.

QTray Plating

Immediately after the bacterial transformation step is completed, the outgrowth cultures are plated onto 48-segmented Qtrays. This protocol is the most improvised procedure in the entire chapter because the Qtray is not a standard microtiter plate. After finding a consistent location for the Qtray on the deck, a simple and convenient spreading method is used, which pipettes just enough culture in the plate segment to both spread by gravity and dry in a reasonable amount of time.

Materials

Transformed bacterial culture. This is typically carried over from the “Bacterial Transformation” procedure as a 96-well BIO-RAD PCR plate.

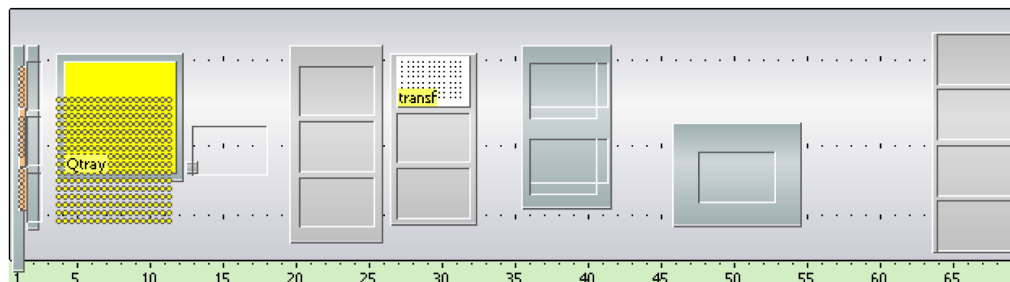
Qtray(s). This specialty petri plate, available from Genetix (x6029), has 48 subdivisions upon which different bacterial cultures can be plated. The plate was chosen because it is the only input to the Genetix Qbot, a colony picking robot located on the 2nd floor of the Beckman Institute.

Methods

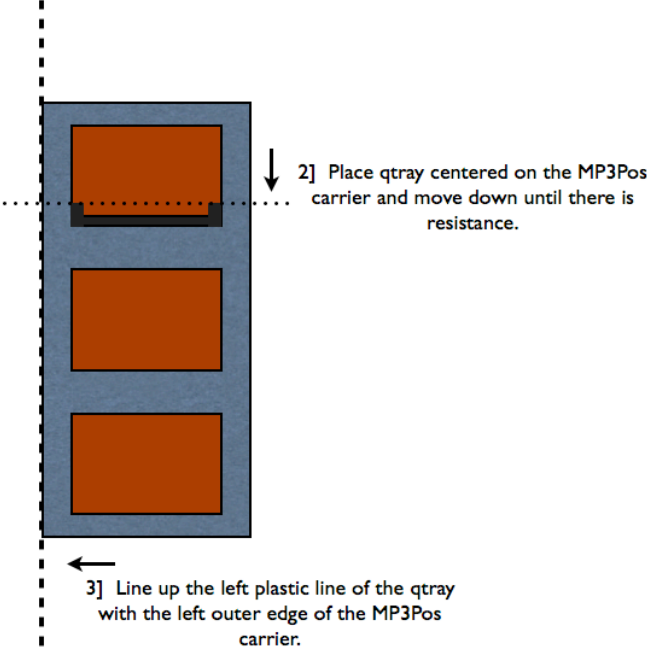
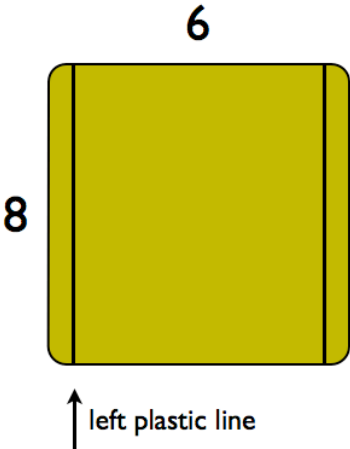
Setup

The template script is named “AN_qtrayplating” and takes less than 5 min to complete. Load the Qtray precisely as shown on the following page and organize the deck as seen in the figure below. After the procedure is finished, cover the Qtray and shake slightly to spread the drops in each segment. Remove the cover and let the solutions air dry. Replace the cover and incubate overnight.

Qtray: 48-segment Qtray
transf: 96-well BIO-RAD PCR plate



1] Orient qtray as shown and take note of the plastic lines along the bottom.







QTray Loading

Procedure

See Note 1

See Note 2

1	Wash Tips		5 + 5 ml
2	Begin Loop		6 times "whole"
3	Aspirate		72 µl Water 500 SP Visc NO DETECT (Asp High) "transf" (Col. 1, Rows 1-8) , 1 option
4	Begin Loop		2 times "down"
5	Begin Loop		3 times "across"
6	Dispense		12 µl Water 500 SP Visc NO DETECT "Qtray" (Col. 1, Rows 1,3,5,7,9,11,13,15) , 3 options
7	End Loop		"across"
8	End Loop		"down"
9	Wash Tips		10 + 10 ml
10	End Loop		"whole"

Notes

1. Change the loop “whole” to reflect the number of columns you wish to plate. Typically, leave this at 6, the maximum number of columns you can plate onto a 48-segment Qtray.
2. Double click this LiHa aspirate command on line 4 and change the starting column from column 1 to column 7 when plating the second half of “transf.” Save the changes and run the altered protocol.

Culture Rearray

The colony picking robot, or Qbot, is located on the 2nd floor of the Beckman Institute and is available for supervised use by all members in the Mayo Lab. Contact Autumn (email: qiuy@caltech.edu) to schedule a session. The preferred output plate of the Qbot is a Genetix 384-well plate. However, it is beneficial for us to generate 96-well glycerol stock plates because it gives the experimentalist more volume to pull from when it is time to inoculate expression plates. For the site-directed mutagenesis experimental pipeline, 8 colonies are picked per reaction (48 different reactions per 384-well plate) and 2 colonies are sent for initial sequencing to Agencourt Bioscience. Hence, this robot protocol takes 2 colony cultures per reaction from a 384-well plate and rearrays them into 2 96-well plates.

Materials

384-Well Genetix Picking Culture Plate(s). These plates, available from Genetix, are used exclusively with the Qbot to inoculate picked cultures into liquid media. They can be filled with LB/antibiotic on the colony picking day by the Qfill instrument, located in the Qbot room. For this protocol, the plate should contain overnight cultures.

96-Well Round-Bottom Plates(s). These polystyrene (PS) plates, available from Falcon (351177), are used to send samples to Agencourt Bioscience for sequencing and to keep as frozen stock from which to inoculate downstream volumes of media. These plates were chosen because of their optical clarity, which allows visual checks of bacterial growth, and for their sterility right out of the box. They are typically sealed with aluminum sticker seals. Fill with 250 μ L of LB/10% glycerol before use on the robot.

LB With 10% Glycerol. Four 96-well plates (2 384-well plates, or looking back, 1 96-well site-directed mutagenesis plate) require about 100mL of this media. We recommend making 1L in a 2L Corning bottle by mixing 100mL of glycerol with 25g of LB granules and filling with 900mL of MilliQ water.

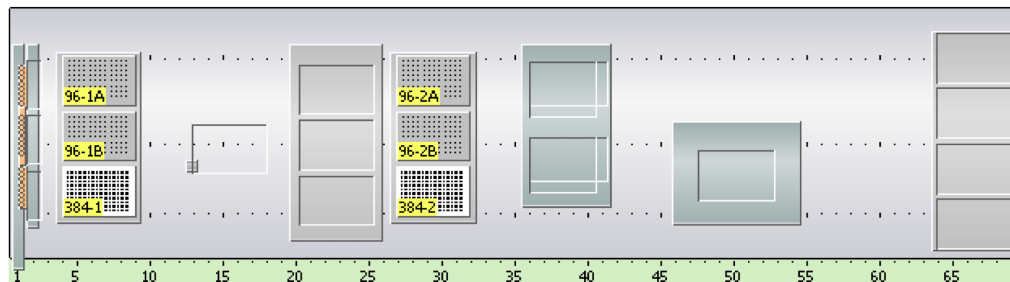
Methods

Setup

The template script is named “AN_Qbot_rearray” and takes roughly **X** min to complete. It will inoculate the first two picked colony cultures from each reaction of the 384-well plate into a fresh 96-well plate. It then creates a duplicate 96-well plate to send for sequencing and continues to do the same procedure on another 384-well plate. Organize the deck as shown in the figure below.

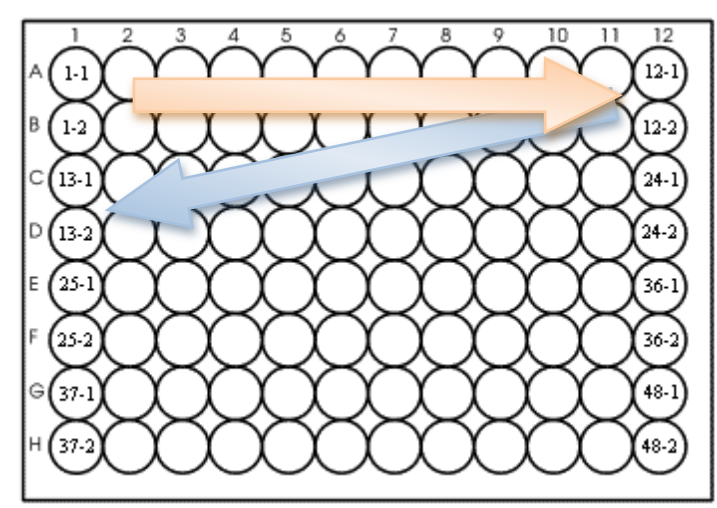
96-1A, 96-1B, 96-2A, 96-2B: 96-well round-bottom plates

384-1, 384-2: 384-well Genetix picking culture plates



Colony Placement:

Execution of the script inoculates the first two colony cultures (X-1, X-2) from the 384-well plate into the 96-well array shown below. The 96-well plate uses the column-style numbering from the 48-well Qtray the 384-well plate was picked from. The arrows are provided as guides for possible manual inoculations and the subsequent sequencing sample identification.














Procedure

See Note 1

See Note 2

See Note 3

See Note 4

1	Wash Tips		5 + 5 ml
2	Comment	rearray 1st 384-well plate	
3	Begin Loop	2 times "duplicate plate"	
4	Begin Loop	12 times "across plate"	
5	Aspirate		15 µl >> Water 500 SP Visc NO DETECT << "384-1" (Col. 1, Rows 1,3) , 1 option
6	Aspirate		15 µl >> Water 500 SP Visc NO DETECT << "384-1" (Col. 2, Rows 1,3) , 1 option
7	Aspirate		15 µl >> Water 500 SP Visc NO DETECT << "384-1" (Col. 1, Rows 2,4) , 1 option
8	Aspirate		15 µl >> Water 500 SP Visc NO DETECT << "384-1" (Col. 2, Rows 2,4) , 1 option
9	Dispense		15 µl Water 500 SP Visc NO DETECT "96-1A" (Col. 1, Rows 1-8) , 2 options
10	Wash Tips		5 + 5 ml
11	End Loop	"across plate"	
12	End Loop	"duplicate plate"	
13	Comment	rearray 2nd 384-well plate	
14	Begin Loop	2 times "duplicate plate"	
15	Begin Loop	12 times "across plate"	
16	Aspirate		15 µl >> Water 500 SP Visc NO DETECT << "384-2" (Col. 1, Rows 1,3) , 1 option
17	Aspirate		15 µl >> Water 500 SP Visc NO DETECT << "384-2" (Col. 2, Rows 1,3) , 1 option
18	Aspirate		15 µl >> Water 500 SP Visc NO DETECT << "384-2" (Col. 1, Rows 2,4) , 1 option
19	Aspirate		15 µl >> Water 500 SP Visc NO DETECT << "384-2" (Col. 2, Rows 2,4) , 1 option
20	Dispense		15 µl Water 500 SP Visc NO DETECT "96-2A" (Col. 1, Rows 1-8) , 2 options
21	Wash Tips		5 + 5 ml
22	End Loop	"across plate"	
23	End Loop	"duplicate plate"	
24	Notification	Send email now Mayo Lab	

Notes

1. This loop on script line 3 controls the creation of a duplicate plate to send to Agencourt for sequencing.
2. Script lines 5–8 control which colony cultures to pick from out of the 384-well plate. Currently, they are pulling the top two colony cultures from each reaction. If you wanted colony cultures 3 and 4, line 5 should read (Col. 1, Rows 5,7), line 6 should read (Col. 2, Rows 5,7), line 7 should read (Col. 1, Rows 6,8) and line 8 should read (Col. 2, Rows 6,8). Also notice that the liquid class name is encapsulated by ">> <<" notations, indicating a custom liquid class. In this case, the liquid class "Water 500 SP Visc NO DETECT" is prefixed with a tip mixing routine to resuspend bacterial culture that may have settled. This feature was implemented in this way to avoid having an extra four Tip Mixing script lines in the code before each aspiration block.
3. The comment on line 13 separates the script into two parts, indicating the completion of the first 384-well plate.
4. At the end of the rearray script, there is a command to send an email to the user. This is useful since the script runs longer than 15 min.

Expression Plate Inoculation

For experimental assays that require a large amount of protein, it is very beneficial to grow 5mL bacterial cultures in 24-well plates as opposed to 1mL cultures in 96-well deep-well plates. To assist with this task, a robotic procedure was developed to inoculate 24-well plates from a 96-well liquid stock plate. Incubate the 24-well plates for 16 hr with shaking and at the appropriate temperature for optimal protein expression.

Materials

96-Well Frozen and Liquid Stock Plate(s). Typically, 96-well frozen stocks are kept in Falcon round-bottom PS plates (351177) with aluminum seals. These are then replicated into identical plates filled with 200 μ L of LB/antibiotic and grown statically overnight. This liquid stock plate is then used to inoculate the 24-well plates.

24-Well Expression Plate(s), Lid(s), and Capmat(s). These round-bottom Whatman plates (7701-5110) are made of polypropylene (PP), arrive sterile, and can be autoclaved and reused. The BugStopper capmats from Whatman (7704-0014) are porous to air, autoclavable, and keep the wells sealed from each other during incubated shaking.

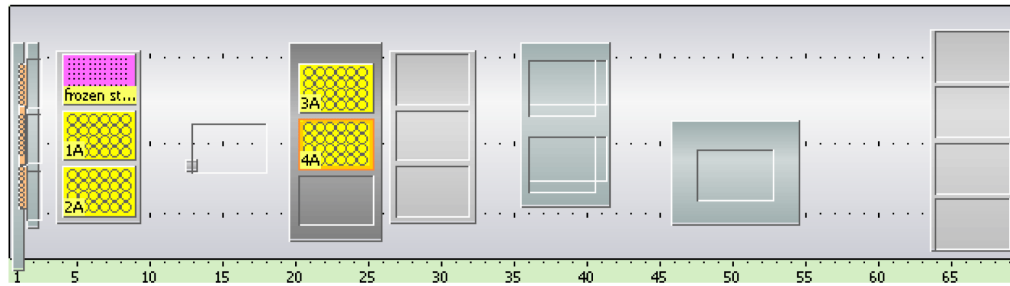
Auto-Induction TB Media and Antibiotic. This media, available from Novagen (71491-4), obviates the need for an induction agent such as IPTG and can be prepared in the autoclave or microwave. Any leftover media is refrigerated. For 1L, add 1 packet (60g) to 10mL of glycerol and fill with water. Add the appropriate antibiotic before dispensing into the 24-well plates.

Methods

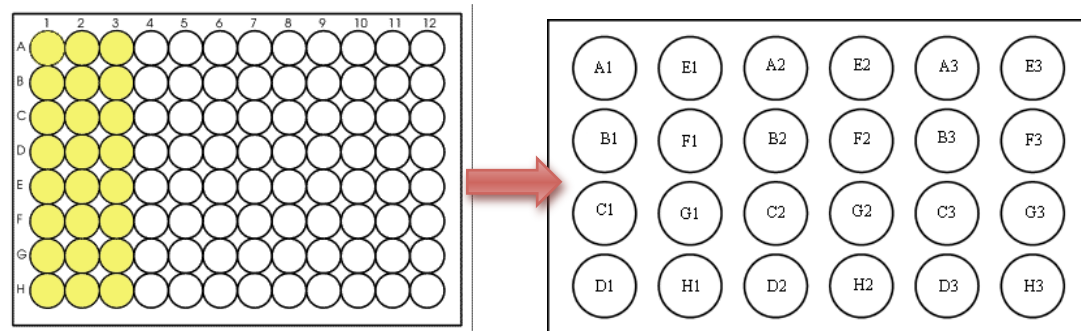
Setup

The template script is named “AN_96inoculate_1x5mL” and takes roughly 10 min to complete. Organize the deck as shown in the figure below.

Frozen/liquid stock plate: 96-well Falcon round-bottom PS plate
 1A,2A,3A,4A: 24-well Whatman round-bottom plates












Example:



Every three columns of the 96-well plate is moved to a 24-well plate following the labeling shown above.

Procedure

1	Wash Tips		5 + 5 ml
2	Comment	inoculate 1 96well plate into 4 24well plates	
3	Comment	inoculate plates 1A and 2A	
4	Begin Loop	2 times "diffplate"	
5	Begin Loop	3 times "aspdisp"	
6	Mix		2 x 75 µl LiHa Tip Mixing AN "frozen stock" (Col. 1, Rows 1-8) , 2 options
7	Aspirate		25 µl Water 500 SP Visc NO DETECT "frozen stock" (Col. 1, Rows 1-8) , 2 options
8	Dispense		25 µl Water 500 SP Visc NO DETECT submerged dispense "1A" (Col. 1, Rows 1-4) , 2 options
9	Dispense		25 µl Water 500 SP Visc NO DETECT submerged dispense "1A" (Col. 2, Rows 1-4) , 2 options
10	Wash Tips		10 + 10 ml
11	End Loop	"aspdisp"	
12	End Loop	"diffplate"	
13	Comment	inoculate plate 3A	
14	Begin Loop	3 times "aspdisp"	
15	Mix		2 x 75 µl LiHa Tip Mixing AN "frozen stock" (Col. 7, Rows 1-8) , 1 option
16	Aspirate		25 µl Water 500 SP Visc NO DETECT "frozen stock" (Col. 7, Rows 1-8) , 1 option
17	Dispense		25 µl Water 500 SP Visc NO DETECT submerged dispense "3A" (Col. 1, Rows 1-4) , 1 option
18	Dispense		25 µl Water 500 SP Visc NO DETECT submerged dispense "3A" (Col. 2, Rows 1-4) , 1 option
19	Wash Tips		10 + 10 ml
20	End Loop	"aspdisp"	
21	Comment	inoculate plate 4A	
22	Begin Loop	3 times "aspdisp"	
23	Mix		2 x 75 µl LiHa Tip Mixing AN "frozen stock" (Col. 10, Rows 1-8) , 1 option
24	Aspirate		25 µl Water 500 SP Visc NO DETECT "frozen stock" (Col. 10, Rows 1-8) , 1 option
25	Dispense		25 µl Water 500 SP Visc NO DETECT submerged dispense "4A" (Col. 1, Rows 1-4) , 1 option
26	Dispense		25 µl Water 500 SP Visc NO DETECT submerged dispense "4A" (Col. 2, Rows 1-4) , 1 option
27	Wash Tips		10 + 10 ml
28	End Loop	"aspdisp"	

See Note 1

See Note 2

See Note 3

See Note 4

Notes

1. The mixing steps are based on there being at least 100 μ L in each well of the stock plate. It is OK to be more aggressive and add more cycles if the cells have precipitated.
2. The “2 options” loop tag on script lines 6–9 inform the LiHa to (1) move over 1 column every “aspdisp” and (2) either move over 3 columns in every “diffplate” in lines 6 and 7 or move down 1 labware in every “diffplate” in lines 8 and 9. The first option is consistent with the “1 option” tag in script lines 15–18 and 23–26.
3. The liquid class has been defined to dispense liquids under the presumed volume level in the destination plate. This minimizes cross-contamination from ricochet droplets.
4. The next two 24-well plates (plates 3A and 4A) are inoculated in separate blocks of script code, which helps to visualize what is going on.

Protein Purification

Automated protein purification has long been a major goal of high-throughput methodology, hence the existence of several commercial applications. Several companies offer 96-well nickel filter plates and readily available lysis and elution reagents. Much of the procedure described here was developed by balancing the simplicity of ordering materials against the cost of those materials.

Purification protocols vary from protein to protein, and the procedure here is no different. All conditions were developed with the β 1 domain of protein G, and preparation of other protein domains will likely need some optimization. In addition, the vacuum steps in the procedure are not entirely robust and currently require human supervision. This procedure is fed directly from the “Expression Plate Inoculation” protocol described earlier. This signifies that the unpacking scheme used to inoculate 24-well plates is used to transfer the cell lysates to the 96-well nickel filter plate.

Materials

Expression Cultures With Lids. Typically, automation experiments use 24-well round-bottom Whatman PP plates (7701-5110) filled with 5mL cultures of bacteria. Autoclavable BugStopper capmats from Whatman (7704-0014) are used as lids during overnight growth.

Centrifuge Plate Adapters. The Sorvall Legend RT centrifuge with a swing-out rotor in Broad 140 (autoclave room) has four hanging bucket adapters for microtiter plates.

Lysis Buffer. For four 24-well plates, make 45mL of lysis buffer consisting of (1) 4.5mL of 10 \times CellLytic B (Sigma-Aldrich), (2) 0.9mL of 10mg/mL lysozyme (Sigma-Aldrich), (3) 9 μ L of HC benzonase (Sigma-Aldrich), (4) 0.45mL of elution buffer, and (5) enough equilibration buffer to fill to 45mL.

His-Select Filter Plate(s). These nickel plates, available from Sigma-Aldrich (H0413), are the most expensive reagent in the purification process. However, the purity in all purification schemes performed so far has been excellent.

96-Well Reservoir Plate. Available from Seahorse Scientific (#S30014), these are pyramid-bottom PP reservoirs.

Equilibration, Wash, and Elution Buffers in 100mL Troughs. These troughs, available from Tecan (10 613 049), only fit in the carrier located at grid 2. To purify one 96-well plate, prepare 57.6mL of equilibration buffer (50mM NaPO₄ buffer, 300mM NaCl at pH 8), 115.2mL of wash buffer (equilibration buffer + 5mM imidazole at pH 8), and 48mL of elution buffer (equilibration buffer + 250mM imidazole at pH 8). Be sure to have at least 15% more volume than required in the appropriate troughs.

200 μ L Robot Tips. These automation tips are available from USA Scientific (TipONE #1188-1700). They are stacked with yellow inserts.

96-Well Collection Plate(s) and Plate Seal(s). These plates are manufactured by Axygen and are available from VWR (P-DW-11-C-S). They are 1mL deep-well PP plates and arrive sterilized. These plates fit and work well with the Tecan-provided spacer #2 and are used to minimize cross contamination during elution. The plastic sticker seal is then applied to prevent evaporation.

Methods

Setup

The template script is named “AN_ProtPurif_1x5ml_96” and takes roughly 1 hr to complete. Organize the deck as shown in the figure below.

1A,2A,3A,4A: Expression cultures in 24-well Whatman round-bottom plates

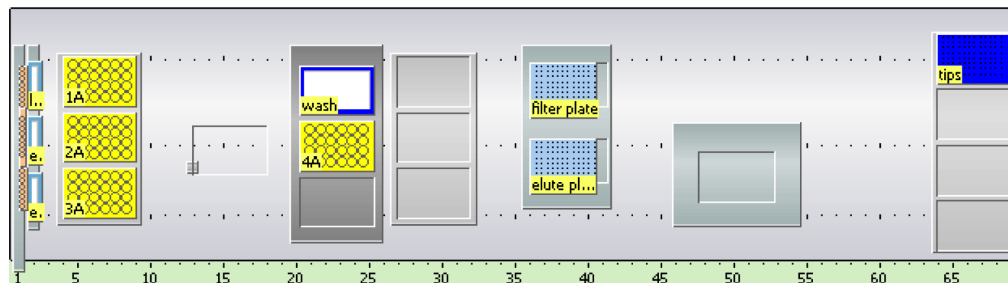
Lysis/equilibration/elution buffers: Stored in 100mL troughs at grid 2

Wash buffer: 96-well reservoir plate

Filter plate: 96-well His-Select filter plate, on top of separation block

Elute plate: 96-well collection plate on top of spacer #2

















Tips: 200 μ L robot tips (blue box)










Prior to the robot procedure:

1. Centrifuge 24-well expression cultures at 3060 \times g and dump the supernatant.
2. Freeze plates at -20°C for at least 15 min. Bring back to 25°C . Once pellets are sufficiently thawed, proceed with the robot protocol.

Procedure

	1	Wash Tips		10 + 10 ml
	2	Comment		Lyse all plates on left carrier
See Note 1	3	Begin Loop		3 times "lyse all plates on carrier"
	4	Begin Loop		3 times "lyse plate"
	5	Aspirate		400 µl Water 500 SP Visc in Trough "lysis buffer" (Col. 1, Rows 1-8)
	6	Dispense		400 µl Water 500 SP Visc NO DETECT "1A" (Col. 1, Rows 1-4), 2 options
See Note 2	7	Mix		5 x 440 µl LiHa Lysate Resuspension AN "1A" (Col. 1, Rows 1-4), 2 options
	8	Dispense		400 µl Water 500 SP Visc NO DETECT "1A" (Col. 2, Rows 1-4), 2 options
	9	Mix		5 x 440 µl LiHa Lysate Resuspension AN "1A" (Col. 2, Rows 1-4), 2 options
	10	Wash Tips		10 + 10 ml
	11	End Loop		"lyse plate"
	12	End Loop		"lyse all plates on carrier"
	13	Comment		Lyse plate on middle carrier
	14	Begin Loop		3 times "lyse plate"
	15	Aspirate		400 µl Water 500 SP Visc in Trough "lysis buffer" (Col. 1, Rows 1-8)
	16	Dispense		400 µl Water 500 SP Visc NO DETECT "4A" (Col. 1, Rows 1-4), 1 option
	17	Mix		5 x 440 µl LiHa Lysate Resuspension AN "4A" (Col. 1, Rows 1-4), 1 option
	18	Dispense		400 µl Water 500 SP Visc NO DETECT "4A" (Col. 2, Rows 1-4), 1 option
	19	Mix		5 x 440 µl LiHa Lysate Resuspension AN "4A" (Col. 2, Rows 1-4), 1 option
	20	Wash Tips		10 + 10 ml
	21	End Loop		"lyse plate"
	22	Comment		Add equilibration buffer to filter plate and vacuum
	23	Begin Loop		12 times "across plate"
	24	Begin Loop		2 times "600ul"
	25	Aspirate		300 µl Water 500 SP Visc in Trough "equil buffer" (Col. 1, Rows 1-8)
	26	Dispense		300 µl Water 500 SP Visc NO DETECT "filter plate" (Col. 1, Rows 1-8), 1 option
	27	End Loop		"600ul"
	28	Wash Tips		5 + 5 ml
	29	End Loop		"across plate"

30	Comment	run equilibrium buffer
31	Vacuum Separation	ApplyVacuumRear(700)
32	Start Timer	1
33	Wait for Timer	Timer 1 : 25 sec
34	Vacuum Separation	VentRear()
35	Vacuum Separation	DeactivateSystem()
36	Comment	Add cell lysates to filter plate and vacuum
37	Group	cell lysate pipetting
38	Comment	plates1-3
39	Begin Loop	3 times "carrier"
40	Begin Loop	3 times "addlysate"
41	Mix	 5 x 440 µl LiHa Lysate Resuspension AN "1A" (Col. 1, Rows 1-4) , 2 options
42	Aspirate	 425 µl Water 500 SP Visc NO DETECT "1A" (Col. 1, Rows 1-4) , 2 options
43	Mix	 5 x 440 µl LiHa Lysate Resuspension AN "1A" (Col. 2, Rows 1-4) , 2 options
44	Aspirate	 425 µl Water 500 SP Visc NO DETECT "1A" (Col. 2, Rows 1-4) , 2 options
45	Dispense	 425 µl Water 500 SP Visc NO DETECT "filter plate" (Col. 1, Rows 1-8) , 2 options
46	Wash Tips	 10 + 10 ml
47	End Loop	"addlysate"
48	End Loop	"carrier"
49	Comment	plate4
50	Begin Loop	3 times "addlysate"
51	Mix	 5 x 440 µl LiHa Lysate Resuspension AN "4A" (Col. 1, Rows 1-4) , 1 option
52	Aspirate	 425 µl Water 500 SP Visc NO DETECT "4A" (Col. 1, Rows 1-4) , 1 option
53	Mix	 5 x 440 µl LiHa Lysate Resuspension AN "4A" (Col. 2, Rows 1-4) , 1 option
54	Aspirate	 425 µl Water 500 SP Visc NO DETECT "4A" (Col. 2, Rows 1-4) , 1 option
55	Dispense	 425 µl Water 500 SP Visc NO DETECT "filter plate" (Col. 10, Rows 1-8) , 1 option
56	Wash Tips	 10 + 10 ml
57	End Loop	"addlysate"
58	Group End	cell lysate pipetting




See Note 3




See Note 4

See Note 5

See Note 6

See Note 7

59	Comment	run lysate
60	Vacuum Separation	ApplyVacuumRear(150)
61	Start Timer	1
62	Wait for Timer	Timer 1 : 60 sec
63	Vacuum Separation	VentRear()
64	Vacuum Separation	DeactivateSystem()
65	User Prompt	"Check if all lysate went through." sound : no
66	Comment	Add Wash Buffer
67	Begin Loop	2 times "2x wash step"
68	Group	Move_MCA_Safe_to_Right_Back_Corner
69	Command	"W1PAA,,2362," wait
70	Command	"W1PAA,10,," wait
71	Move Vector	Vector "Safe-RightBackCorner", Grid 64, Site 1(MCA 1), move from Safe to End Positior
72	Group End	Move_MCA_Safe_to_Right_Back_Corner
73	Get DiTis	Grid 64; Site: 1 (DiTi 200ul MCA) Fetch 8 rows and 12 columns
74	Move	 Positioning with global Z-Trave "wash" (Col. 1, Rows 1-4)
75	Begin Loop	3 times "wash buffer"
76	Aspirate	 200 µl MCA96 trough "wash" (Col. 1, Rows 1-8)
77	Dispense	 200 µl MCA96 trough "filter plate" (Col. 1, Rows 1-8)
78	End Loop	"wash buffer"
79	Drop DiTis	Back to Source
80	Comment	run wash buffer
81	Vacuum Separation	ApplyVacuumRear(500)
82	Start Timer	1
83	Wait for Timer	Timer 1 : 30 sec
84	Vacuum Separation	VentRear()
85	Vacuum Separation	DeactivateSystem()
86	End Loop	"2x wash step"

87	User Prompt	"Wipe filter plate nozzles if necessary." sound : no
88	Comment	move filter plate to front Vac position
89	Comment	Pickup block at rear position
90	ROMA Vector	Vector "Te-VacS_Narrow_1", Grid 36, Site 2 (ROMA 1) open, from Safe to End Position, grip, from End to Safe Position
91	Comment	Deliver block to front positior
92	ROMA Vector	Vector "Te-VacS_Narrow_1", Grid 36, Site 5 (ROMA 1) move from Safe to End Position, open, from End to Safe Positior
93	Move ROMA	Move to home position (ROMA 1)
94	Comment	Add elution buffer to filter plate and vacuum
95	Begin Loop	12 times "across plate"
96	Begin Loop	2 times "500ul"
97	Aspirate	 250 µl Water 500 SP Visc in Trough "elute buffer" (Col. 1, Rows 1-8)
98	Dispense	 250 µl Water 500 SP Visc NO DETECT "elute plate" (Col. 1, Rows 1-8) , 1 option
99	End Loop	"500ul"
100	Wash Tips	 5 + 5 ml
101	End Loop	"across plate"
102	Comment	run elution buffer
103	Vacuum Separation	ApplyVacuumFront(300)
104	Start Timer	1
105	Wait for Timer	Timer 1 : 45 sec
106	Vacuum Separation	VentFront()
107	Vacuum Separation	DeactivateSystem()
108	User Prompt	"Check if all elution volumes went through." sound : no
109	Notification	Send email now Mayo Lab

See Note 8

See Note 9

Notes

1. The “2 options” loop tags in lines 6–9 refer to the loops defined in lines 3 and 4. The loop in line 3, “lyse all plates on carrier,” moves through the labware on the current carrier. The loop in line 4, “lyse plate,” moves through every 2 columns on the 24-well plates. Script lines 14–21 define an isolated “lyse plate” loop for the fourth 24-well plate on the middle carrier at grid 20.
2. All mixing steps use the special liquid class “LiHa Lysate Resuspension AN” developed exclusively for this protein purification procedure. Lysates are mixed immediately after the lysis buffer dispense command and again before being added to the nickel filter plate. These steps do an adequate to superb job of resuspending the pellets.
3. All the Vacuum Separation commands in the script follow the same structure. The first (line 31) tells which of the two blocks to pull from (rear or front) and at what pressure (mbar). Then the script counts off a timer, after which the pressure is vented (line 34) and the vacuum is turned off (line 35). All vacuum pressure and timer parameters were determined empirically and can and should be adjusted depending on what the user sees during a run.
4. The loops within the cell lysate pipetting steps are identical to those described in Note 1.
5. Line 65 (and later on, lines 87 and 108) has user prompts that stop the run, unlock the shield, and require the user to click OK. Due to the non-robustness of the current purification protocol, these prompts are necessary.
6. The Group “Move_MCA_Safe_to_Right_Back_Corner” is a set of commands that orient the MCA96 into a safe position starting from anywhere on the deck. This helps to avoid crashes when executing the “Get DiTis” command.
7. Line 74 instructs the MCA96 to move to a far site so that it won’t collide with the thermalcycler as it moves between the MP3Pos carriers and its DiTi carrier.
8. Lines 89–93 are very specific RoMa vectors that were developed to move the separation block from the rear to the front position.
9. At the end of the purification script there is a command to send an email to the user. This is useful since the script runs longer than 15 min.

HT Agarose Gel Electrophoresis

In order to aid with the high-throughput visualization of nucleic acid product, Invitrogen has developed 48- and 96-well agarose gel systems for use on liquid handling platforms. The 96-well version was purchased and adapted for use on the Tecan Freedom EVO. Check the Invitrogen website for questions concerning what percentage agarose is appropriate for the size of the products being separated.

Invitrogen provides the MS Windows program “E-Editor” that aids in visualization by aligning EGel images into a variety of convenient arrays.

Materials

Nucleic Acid Product. Typically, this is DNA product from a site-directed mutagenesis procedure in a 96-well BIO-RAD PCR plate.

E-Gel Low Range Quantitative DNA Ladder. This ladder is from Invitrogen (12373-031). For one 96-well agarose gel, 200 μ L of a 1:1 mixture of ladder and water is required.

E-Gel 96 2% Agarose Gel(s). These are available in packs of eight from Invitrogen (G7008-02). This version uses ethidium bromide for staining, but Invitrogen also offers non-carcinogenic versions.

Mother Ebase. This is the base (EB-M03) for the 48- and 96-well agarose and PAGE gels that Invitrogen offers. They can be daisy chained, but we currently only have one. Plug it into the power strip underneath the deck.

50 μ L Robot Filter Tips. These automation tips are the ART BioRobotix tips available from Molecular BioProducts (#906-021). They are individually wrapped in green boxes.

1.5mL Eppendorf Tubes in 24-Well Cooled Block. These are the lab standard autoclavable tubes available in any biochemistry lab placed in the 24-well cooled block. For this method, the block doesn't need to be cold.

Sterile H₂O in 96-Well Low Profile Reservoir Plate. Available from Seahorse Scientific (S300-18), these are pyramid-bottom PP reservoirs. Although the low profile is not required, the method was developed when only this reservoir was available. Fill with standard sterile DI water.

Methods

Setup

The template script is named “AN_EGel_load_econ” and takes roughly 20 min to complete, including the 12-min electrophoresis. Organize the deck as shown in the figure below. Plug in the Mother Ebase and make sure it is set to “EG” for E-Gel. Open the E-Gel package and load the E-Gel. The display on the Mother Ebase should switch to a 12-min timer.

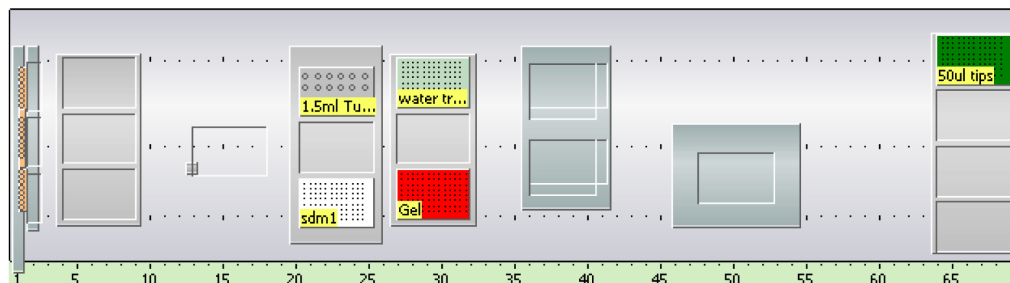
1.5ml tubes: Put the 1:1 DNA ladder/water mix into A4 of the 24-well block

sdm1: 96-well BIO-RAD PCR plate

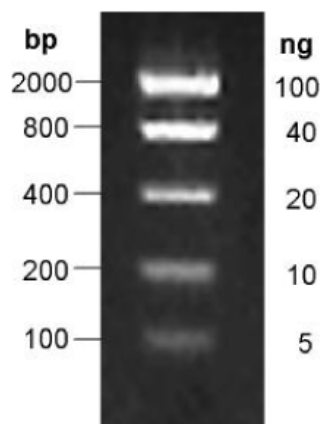
Water trough: 96-well low profile reservoir plate

Gel: Mother Ebase loaded with an E-Gel 96





50µL tips: 50µL robot filter tips (black box)



E-Gel Low Range Quantitative DNA Ladder



Procedure

1	Group	Move_MCA_Safe_to_Right_Back_Corner	
2	Command	"W1PAA,,2362," wait	
3	Command	"W1PAA,10," wait	
4	Move Vector	Vector "Safe-RightBackCorner", Grid 64, Site 1(MCA 1) move from Safe to End Positor	
5	Group End	Move_MCA_Safe_to_Right_Back_Corner	
6	Comment	Dilute the DNA samples 1:1 inside the tip and dispense into the EGe	
7	Get DiTis	Grid 64; Site: 1 (DiTI 50ul MCA) Fetch 8 rows and 12 columns	
8	Move	 Positioning with global Z-Trave "water trough" (Col. 1, Rows 1-8)	
9	Aspirate	 10 µl MCA96 trough "water trough" (Col. 1, Rows 1-8)	
10	Aspirate	 10 µl MCA96 trough (2nd aspirate) "sdm1" (Col. 1, Rows 1-8)	
11	Dispense	 20 µl MCA96 Diti ege "Gel" (Col. 1, Rows 1-8)	
12	Move	 Positioning with global Z-Trave "water trough" (Col. 1, Rows 1-8)	
13	Drop DiTis	Back to Source	
14	Comment	Add the DNA Ladder to the eGel96; 200uL of 1:1 water/ladder	
15	Wash Tips	 5 + 5 ml	
16	Aspirate	 20 µl Water 500 SP Visc NO DETECT "1.5ml Tubes" (Col. 4, Row 1)	
17	Aspirate	 20 µl Water 500 SP Visc NO DETECT "1.5ml Tubes" (Col. 4, Row 1)	
18	Aspirate	 20 µl Water 500 SP Visc NO DETECT "1.5ml Tubes" (Col. 4, Row 1)	
19	Aspirate	 20 µl Water 500 SP Visc NO DETECT "1.5ml Tubes" (Col. 4, Row 1)	
20	Aspirate	 20 µl Water 500 SP Visc NO DETECT "1.5ml Tubes" (Col. 4, Row 1)	
21	Aspirate	 20 µl Water 500 SP Visc NO DETECT "1.5ml Tubes" (Col. 4, Row 1)	
22	Aspirate	 20 µl Water 500 SP Visc NO DETECT "1.5ml Tubes" (Col. 4, Row 1)	
23	Aspirate	 20 µl Water 500 SP Visc NO DETECT "1.5ml Tubes" (Col. 4, Row 1)	
24	Dispense	 20 µl Water 500 SP Visc NO DETECT "Gel" (Col. 13, Rows 1-8)	
25	Wash Tips	 5 + 5 ml	

See Note 1

See Note 2

See Note 3

See Note 4

Notes

1. The Group “Move_MCA_Safe_to_Right_Back_Corner” is a set of commands that orient the MCA96 into a safe position starting from anywhere on the deck. This helps to avoid crashes when executing the “Get DiTiS” command.
2. Lines 8 and 12 instruct the MCA96 to move to a far site so that it won't collide with the thermalcycler as it moves between the MP3Pos carriers and its DiTi carrier.
3. The liquid class “MCA96 trough (2nd aspirate)” used in line 10 was developed to allow the MCA96 to introduce a small airgap between successive aspirations. By doing this, the script obviates the need for a separate dilution plate and instead dispenses 10µl each of water and DNA product directly into the E-Gel.
4. After the script is finished, check the E-Gel for any missed dispenses. If everything is satisfactory, run the 12-min electrophoresis program.

Chapter 3: Site-Directed Mutagenesis

Overview

Although there are numerous site-directed mutagenesis (SDM) procedures available, it can be difficult to select and effectively employ the protocol that best serves your application. In this case, an automated SDM protocol should be cost-effective, simple in terms of enzymatic steps, and robust enough to avoid manual intervention. The method described by Tseng *et al*, *Anal. Biochem*, 2008 titled “A novel megaprimed and ligase-free, PCR-based, site-directed mutagenesis method” satisfies all of these prerequisites. The single mutagenic primer halves oligo costs when compared against the standard quickchange method. The absence of phosphorylation or ligase steps simplifies the overall procedure and with some tweaking the method can be robust enough to not incur a large number of repeat reactions.

Some modifications were made to the procedure described in the paper in order to further accommodate the needs of automated protocols. The most important change was to the choice of polymerase from Pfu Turbo to NEB’s Hot Start Phusion Polymerase. The new enzyme is faster (speeds reaction times from 8 hours to 2 hours), advertised as higher fidelity and its Hot Start feature prevents it from modifying template and primer DNA before the initial melting step. Polymerase choice has a large effect on the successful outcome of this particular SDM protocol, and therefore should be investigated for each new system. A minor modification to the overall procedure was the usage of smaller mutagenic oligos which can function as well as larger oligos so long as the annealing temperature has been optimized. The financial benefits of ordering hundreds of shorter oligos are quickly realized.

The SDM protocol is outlined below in Figure 1. First, megaprimers are amplified by annealing the mutagenic and constant flanking oligos to the template. Then, full length product is generated by annealing the megaprimers to the template and completing the extension. Finally, a Dpn1 digestion step is necessary to remove the parental wild-type template.

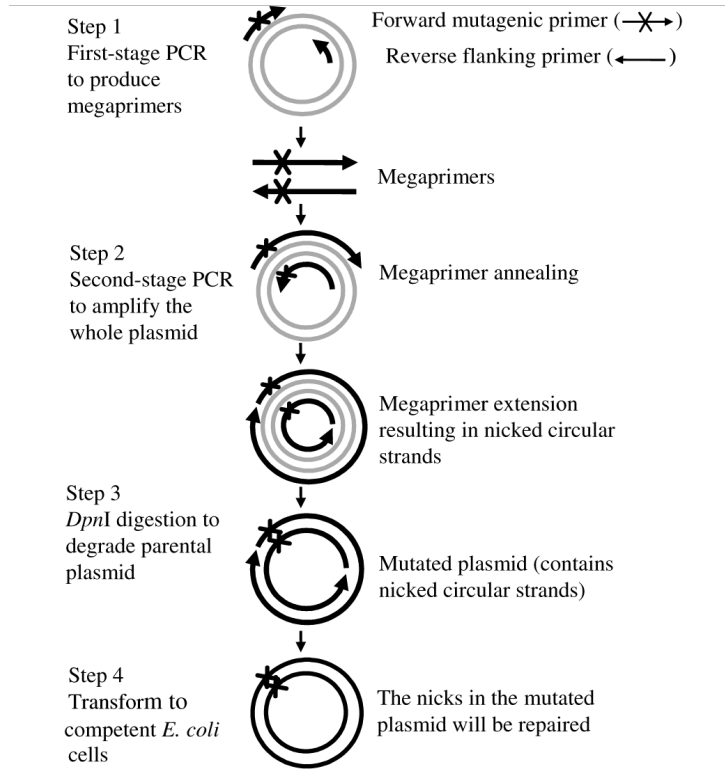


Fig. 1. Schematic diagram of the megaprimed and ligase-free site-directed

Figure 1. Schematic of the megaprimed and ligase-free SDM method. Adapted from Tseng *et al*, *Anal. Biochem.* 2008.

Oligonucleotide Design

In the original paper, an oligo T_m calculation is described that includes a “%mismatch” variable in order to account for the number of basepairs that won’t anneal exactly to the template strand. Initial oligo design work was guided by this T_m value, and five amino acid scans (A, V, S, T, and M) of the small model system GB1 were tested. Upon experimental construction of the variants it was noticed that the calculated T_m values had no bearing on successful amplification of the mutagenic megaprimer and therefore subsequent full length mutated plasmid. Instead, successful amplification and mutagenesis correlated better with the T_m values calculated by IDT, done by the nearest neighbor method. In fact, a later review of the data showed that a simpler T_m calculator was almost as good as the complex nearest neighbor method in terms of predicting successful amplification.

The first few amino acid scans were designed manually in an excel spreadsheet. In order to facilitate the entire high-throughput scheme, a python script was developed (oligotm.py) that scans a given nucleotide sequence and generates a mutagenic oligo for each codon in the sequence while keeping the T_m within a user-defined range. This range should reflect the annealing temperature used in the first step of the SDM reaction. The T_m calculators implemented so far are the basic T_m calculator ($\text{oligo } t_m = 64.9 + 41 * (\text{number of gc bp}) - 16.4 / (\text{number of total bp})$) and the mismatch T_m calculator ($\text{oligo } t_m = 81.5 + 0.41 * (\text{oligo gc}\%) - 675 / (\text{number of total bp}) - \text{mismatch}\%$). The script

output can easily be tailored to allow for a simple copy/paste operation into an IDT excel order form, thus simplifying oligonucleotide design.

As is the script is ideal for site-directed scanning mutagenesis, facilitating the design of oligos for small projects such as alanine scans all the way to larger projects like total site saturation mutagenesis of a protein domain. Alternatively, automated oligonucleotide design for codon saturation mutagenesis schemes (all codons at non-continuous sites) can be adapted from the existing functions defined in the python script.

Experimental Pipeline

The methods for automated site-directed mutagenesis were developed in response to the desire for a database containing stability data for every single-mutant of the GB1 domain. The project was initiated in the spring of 2009 and saw completion by the summer of 2010. A simple schematic of the experimental pipeline is shown in Figure 2. Protocols were first developed as separate modules and later strung together into an efficient system. Over the course of the project, clustered groups of samples fed to the experimental pipeline provided opportunities to individually refine and optimize each module. Currently, 192 single mutants can be constructed and sent for confirmation sequencing in five days with minimal experimentalist strain.

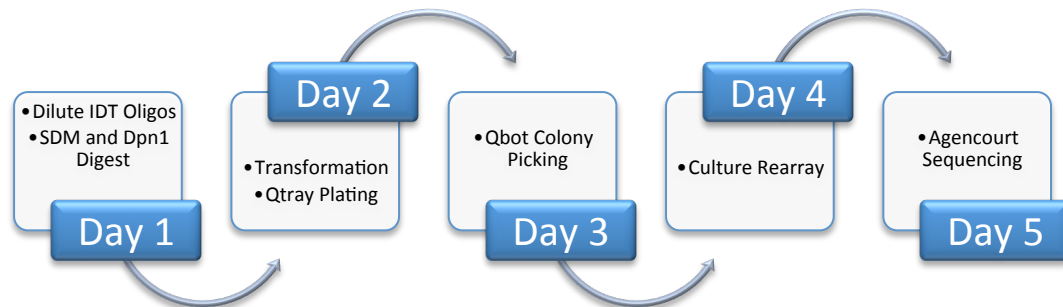


Figure 2. The 5 day SDM experimental pipeline. The majority of the procedure is automated by stringing together the relevant robotic procedures described in chapter 2.

Day 1

Allow the oligos delivered from IDT to come to room temperature. Prepare and run the protocols “Dilute IDT Oligos”, “Site-Directed Mutagenesis”, and “Dpn1 Digest”. For maximum efficiency, generate “2plate” versions of SDM and Dpn1 Digest and then combine them into one ~8hr script that can be run overnight.

Day 2

In the afternoon of the second day, prepare and run the protocols “Bacterial Transformation” and “Qtray Plating”. For maximum efficiency, generate a “2plate” version of Bacterial Transformation.

Day 3

Colony picking can be performed manually at the bench or robotically by contacting Autumn (qiuy@caltech.edu) in the Beckman Institute (BI) to assist with the Qbot colony picker. If done robotically, on the morning of the third day bring the Qtray plates and an appropriate volume of LB/antibiotic (153.6ml for 4 Qtrays) to the BI. Incubate the picked 384-well Genetix plates overnight.

Day 4

On the morning of the fourth day prepare and run the “Culture Rearray” protocol for the four 384-well plates. Incubate the generated 96-well plates for 12 hours.

Day 5

On the morning of the fifth day, prepare the 96-well plates for shipping by first professionally labeling them, then affixing aluminum seals and finally freezing them with dry ice. Fill a Styrofoam box with freezer packs, the plates, and the Agencourt order form and ship it via Fedex. In approximately one week, a representative will contact with data downloading instructions.

The fantastic increase in experimental productivity provided by these automation methods should not be overlooked. As an example, consider the time courses shown in Table 1. Before the introduction of automation methods to the lab, only handfuls of single mutants could be generated at any single time for stability studies. It is unrealistic for one experimentalist to routinely construct and verify any more than 5 mutants with these manual methods, especially considering the large time commitment required in performing stability determination by circular dichroism. The robotics assisted method for constructing single mutants presented here, coupled with the plate-based stability assay described in Allen, Nisthal and Mayo, 2010 is able to generate and analyze 192 mutants in about the same amount of time. The limiting step in this procedure is the week long turnaround time of confirmation sequencing. However, this turnaround time should remain constant if more mutants are sent for sequencing, ensuring the superb scalability of the automated methods.

Table 1. Time Required to construct and measure stabilities for single mutants. The lab procedure in 2007 (manual construction) was only feasible for a handful of mutants. The current state of automation in the lab allows for ~40x more mutants to be constructed and analyzed in roughly the same amount of time.

Procedural Step	Manual Construction (5 mutants)	Robotics Assisted Construction (192 mutants)
SDM and Dpn1 Digest	1.5 days	Overnight
Transformation and Plating	0.5 day	0.5 day
Colony Picking	0.5 day	0.5 day
Protein Expression and Purification	1.5 days	1 day
Sequencing Turnaround	1 day	7 days
Dialysis	Overnight	Overnight
Stability Determination	5 days	1 day
Total Time Required	10 days	10 days

Chapter 4: Automated Gene Assembly

Co-written with Samy Hamdouche

Overview

The chief benefit to developing automation technology in the lab is to expedite the interplay between theory and experiment. Having the ability to take the results from a protein design calculation and in a minimal amount of days produce viable, sequence verified mutant genes should greatly improve the productivity of the lab.

The project was first approached from the experimental side, as it was necessary to determine whether it was possible to perform the molecular biology needed to construct and clone genes into an expression plasmid in an automated fashion. After failing to adapt the complex gene construction methods detailed in Allen *et al*, PNAS 2010 for automation, more simplified protocols were tested. To assist with constructing multiple variants the DNAworks source code was acquired and installed onto our local computer cluster. By ordering completely overlapping oligonucleotides (as designed by DNAworks), performing a PCR assembly reaction, and then amplifying the product with flanking primers we were able to recover the correct sequence-verified gene more than half the time. The next step required automating molecular cloning for further protein expression and production.

The traditional cloning methods of restriction enzymes and overnight ligation reactions are already not very robust when done on the bench, so in their stead we first adapted the PCR cloning method called PIPE, for polymerase incomplete primer extension (Klock *et al*, Proteins, 71, 2008). As shown in Figure 4-1, the PCR products from an insert amplification (IPIPE) and a plasmid linearization/amplification (VPIPE) are mixed together and transformed directly into cells, without any modifications. Any wild-type background is eradicated by limiting the choice of plasmid to those containing a suicide gene such as *ccdB*, where upon transformation the gene product kills the cell. Although this method performs admirably under benchtop conditions (~40 colonies

per 300uL transformation), attempts to transfer that success to the robotic platform have failed to produce enough colonies in a robust fashion. To address this problem a second PCR cloning method was introduced into the lab called CPEC, or circular polymerase extension cloning (Quan and Tian, PLoS ONE, 2009). Figure 4-2 shows that the method is essentially a site-overlap extension (SOE) PCR reaction that uses the IPIPE and VPIPE

products as templates for each other in order to create a fully assembled plasmid. Benchtop testing with this method has shown a direct correlation between number of PCR cycles and amount of transformed bacteria, with 15 cycles of CPEC producing ~800 colonies from a 300uL transformation. The primary reason for the twenty-fold improvement in transformable material is that where the PIPE procedure uses the minor product of a PCR reaction (incomplete PCR products) to form complete plasmid, the CPEC method's major product is our desired end product.

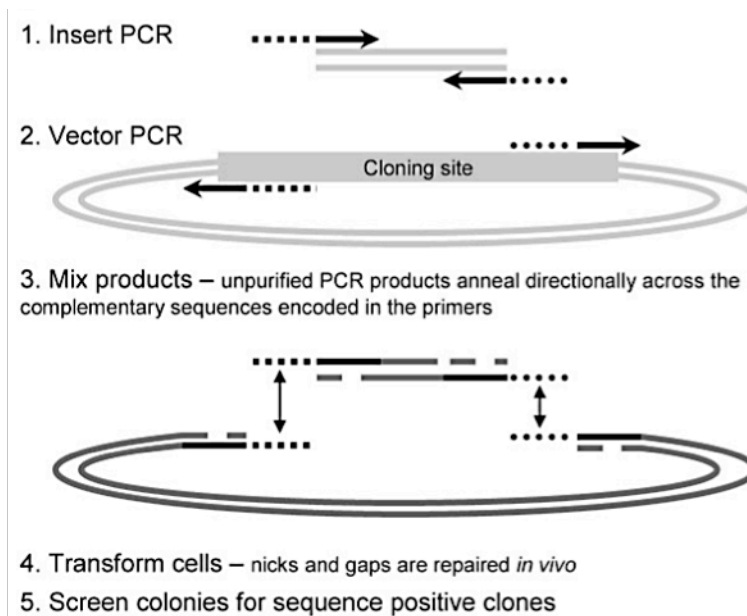


Figure 4-1. Schematic of the polymerase incomplete primer extension method, or PIPE. Adapted from Klock *et al*, 2008.

Plugging in the transformation, plating and picking routines developed for automated site-directed mutagenesis (Chapter 3) completes the experimental pipeline for gene assembly. Detailed robotic protocols for the steps mentioned above can be found in Chapter 2. Finally, a significant amount of scripting was required in order to bridge the gap between the sequence lists produced by typical protein design software and robot-compatible worklists. We now describe our method for taking sequence lists from standard or library designs and ultimately outputting an IDT oligonucleotide order form and an oligo assembly worklist for the robot.

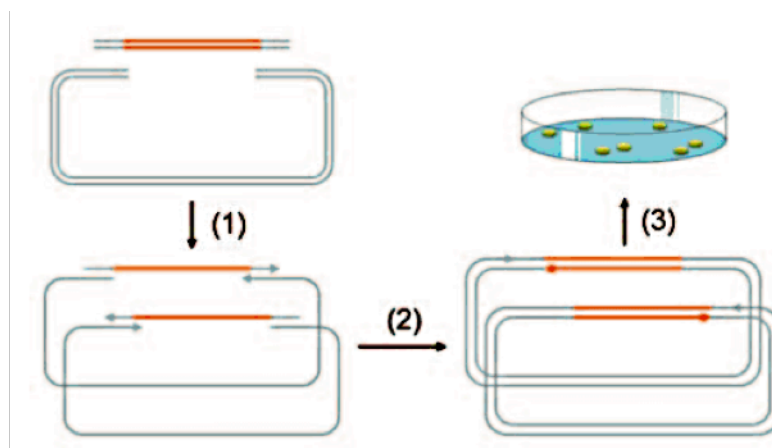


Figure 4-2. Schematic of Circular Polymerase Extension Cloning, or CPEC. Adapted from Quan and Tian, PLoS ONE, 2009.

Python Scripting: From the Computer Screen to Reality

Once a standard or library design calculation is made, a python script is available to generate both an IDT order form for the required oligonucleotides, and the robot work list that assembles the oligonucleotides into the pre-mixes for gene assembly. There are several options for how the order form and worklist is generated, including whether the sequence list is generated from standard or library design output, whether a replacement oligonucleotide or degenerate oligonucleotide method of library construction is desired, and whether the wild-type oligonucleotides have previously been designed by DNAworks and are arrayed onto a 96-well plate. All of these options are specified in the input file to the script. A call to the script is of the form

```
> python geneassem.py input.txt
```

where the argument “input.txt” simply designates the name of the input file.

The Input File

The input file specifies the parameters for the python script and consists of five lines. The parameter of the first line can either be “previous” or “new.” The argument “previous” specifies that the wild-type oligonucleotides have already been constructed using non-gapped design by DNAworks and are arrayed column-wise onto a 96-well plate. In this case, the second line of the input file specifies the name of the DNAworks output file for the wild-type sequence, and the solution number from the DNAworks output that is used, separated by whitespace. The argument “new” specifies that the script should design the wild-type oligonucleotides and generate an order form including both wild-type and mutant oligos. In this case, the second line of the input file is simply the wild-type amino-acid sequence from which oligos are to be designed. The third line of the input file is the name of the PDB file for the wild-type protein (i.e. from input to PHOENIX). The parameter of the fourth line can either be “standard” or “library,” specifying whether sequences are taken from standard or library PHOENIX design output, respectively. In the former case, the PHOENIX output files “design.out” and “design.phoenix” are expected in the same

directory as the script. In the latter case, the library number desired is also specified on the fourth line, with a whitespace separator, and the PHOENIX library design output file “design.out” is expected in the same directory as the script. Finally, parameter on the fifth line can be either “replacement” or “degenerate,” designating whether the replacement oligonucleotide (separate assembly and construction of each individual mutant) or degenerate oligonucleotide (libraries that sample multiple mutations at a position are assembled via degenerate oligonucleotides) method of library construction is used, respectively.



Figure 4-3. Top: an example input file, where wild-type oligos have previously been designed and plated. Bottom: an example input file, where the wild-type oligos have not been designed or ordered.

Output

The output of “geneassem.py” is two files. If the replacement oligonucleotide method is used, the files will be named “idt_order.xls” and “wl.csv.” The file “idt_order.xls” is the excel-formatted IDT order form for the replacement oligonucleotides on the required number of 96-well plates needed to construct the library (which includes wild-type oligos if the “new” parameter is specified). The second output file is the robot-formatted work

list, "wl.csv." In the case that the wild-type oligos have previously been plated, the work list directs the robot to assemble the pre-mixes for all the mutants in the library from the wild-type oligonucleotides in 96-well plates (labeled "wt plate *," where "*" denotes the plate number) and the replacement oligonucleotides on separate 96-well plates (labeled "mut plate *," where "*" denotes the plate number). In the case that the wild-type oligonucleotides are ordered along with the mutant oligos, the work list directs the robot to assemble the pre-mixes for all the mutants in the library from the wild-type oligonucleotides and the replacement oligonucleotides from the 96-well plates ordered as per "idt_order.xls" (labeled "idt plate *," where "*" denotes the plate number).

In the case that the degenerate oligonucleotide method is used, output of "geneassem.py" will be the files "idt_order_deg.xls" and "wl_deg.csv," which are the IDT order form and the robot work list, respectively, in which degenerate oligonucleotides are used.

Test Case: Rational Stabilization of a Designed Enzyme

An example of this powerful methodology is exhibited in the recent construction and analysis of two degenerate oligonucleotide libraries that attempted to stabilize the HG-2 enzyme reported in Privett *et al* while maintaining its activity (in preparation).

The two libraries came from a design that repacked the hydrophobic space between the inner beta sheets and outer alpha helices of the enzyme's fold. One library used the output design structure from the enzyme design calculation as a structural input (deslib) while the other library used the experimentally determined crystal structure of the designed enzyme (xtallib). Each 32-member library was designed and chosen using the protein design software and parameters as described in Allen *et al*, PNAS 2010.

First, the wildtype HG-2 enzyme was constructed in DNAworks and its oligos were ordered from IDT. Then, as in Figure 4-3, the appropriate inputs were submitted to libseqgen.py, followed by writedeg.py, and finally wlggen.py. The oligos necessary to construct the two libraries were ordered using the output idt_order.xls file and once the materials were received, the libraries were assembled on the robot using the output wl.csv file.

A PCR assembly reaction was then performed on this pre-mix of oligos, producing a great number of different construct sizes that can be visualized as a smear on an agarose gel. The correctly sized construct was then amplified out of the smear by running an IPIPE PCR reaction with the appropriate flanking oligos. An agarose gel confirmed a single band of the correct size was amplified after this reaction. Mixing the IPIPE PCR product in a 1:1 ratio with previously prepared VPIPE product and subsequently transforming the solution with chemically competent cells produced an adequate amount of colonies for simple molecular cloning. An additional transformation was performed the next day to get the required amount of colonies for sequencing a 4x oversampling of the library size.

Analysis of the sequencing results showed that we had recovered ~75% of the members of each library. These mutants were then tested for stability by first following the inherent tryptophan fluorescence (chemical denaturation) and then by adding the fluorescent dye SYPRO Orange (thermal denaturation). Unfortunately, no stabilizing mutants were found in the libraries.

Including the weeklong turnaround required for high-throughput sequencing, the experimental construction and verification of the two libraries spanned less than one month. We would not expect the time to increase significantly when considering the simultaneous construction of many more libraries.

Appendix

BL21 Gold DE3 Competent Cell Preparation

This strain grows extremely fast and is useful for both plasmid and protein expression and recovery. When transforming mutagenic protein libraries into this strain of competent cells be sure to keep a glycerol stock of the picked colonies; this simplifies and expedites protein expression after plasmid sequencing as there is no need to perform miniprep or transformation procedures.

On the day of competent cell preparation, be sure to clear space in the cold room for all 4°C manipulations and to autoclave a 1L centrifuge container. Alternatively, prechill all pipets, volumes, and containers to -20°C and perform all procedures on ice.

MgCl₂ solution. 100mM MgCl₂/water, filter sterilized. Store at 4°C.

CaCl₂ solution. 100mM CaCl₂/15% glycerol/water, filter sterilized. Store at 4°C.

1. Start a 10mL overnight culture in LB at 37°C for a 1L batch.
2. Seed a 1L volume with the 10mL overnight culture and shake at 37°C until OD600 reaches 0.4-0.6. All further operations need to be done at 4°C.
3. Spin down the culture at 2500xg for 20 min at 4°C in a liter centrifuge container. Pour off the supernatant and add 100mL (1/10th of initial 1L culture) of the MgCl₂ solution. Resuspend the pellet gently, ideally by shaking at 200-250 rpm on ice. Split the volume between two pre-chilled 50mL tubes.
4. Incubate the resuspended cells at 4°C for 30 min.
5. Spin down the cultures at 2500xg for 20 min at 4°C. Pour off the supernatant and add 20mL (1/50th of initial 1L culture) of the CaCl₂/glycerol solution. Resuspend the pellet gently, ideally by shaking/inversion/nutating.
6. Divide the cells into 1.2mL aliquots in 1.7mL sterile eppendorf tubes. Don't use a repeater pipet. Flash freeze the aliquots in liquid nitrogen. Store at -80°C.

48-Well Qtray Preparation

The Genetix 48-well QTray (x6029) is the perfect petri dish for high-throughput bacterial transformation experiments. The 8 row by 6 column format is amenable to the Freedom EVO's LiHa, and two Qtrays represent exactly one 96-well plate. They are prepared in much the same way as normal petri dishes.

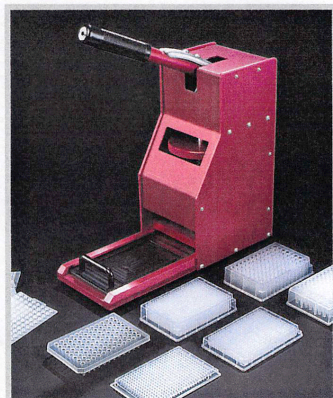
1. Prepare the Qtrays ideally the day before you expect to plate the transformation reactions. Turn on the 55°C water bath located at the end of the Bay 3/4's sink.
2. For four Qtrays, or conversely, 2 96-well Site-Directed Mutagenesis plates, autoclave 1L of LB agar (mix 25g of LB broth miller with 15g of Bacto-Agar and add 1L of MilliQ water).
3. Move the flask out of the autoclave and into the prepared water bath. After ~25 min, the LB agar should be cool enough to add antibiotic and pour into the Qtrays.
4. Turn on your Bunsen burner. Add antibiotic, mix, and then using a 50mL sterile pipet add 200mL of LB agar to the open Qtray. The best location found so far is the bench in Bay 2 or Bay 3 as it is more level than the bench in Bay 1. Flame off any bubbles, replace the 48-well divider, and then put on the Qtray lid in a skewed position to allow air exchange.
5. Let the plates cool to room temperature and dry, ideally overnight in this position. When storing the plates before or after transformation, make sure to keep them upside down to prevent moisture from collecting on the agar surface. If the plate was stored in the cold room, remove any moisture on the surface of the agar before transformation by removing the lid and letting them air-dry under a Bunsen burner for ~30min. This same drying process is also highly recommended after plating.

EasyPress Operation

product manual



EasyPress, silicone sealing mat applicator



Axygen EasyPress is a manual press which provides simple and effective application of Axymats to 96 and 384 microplates, deep well plates, microtiter tube systems and PCR plates. The EasyPress can handle any size plate of varying heights (384 PCR to 2 ml deep well blocks) by quick adjustment of the large thumbwheel.

The procedure is simple

Compatible with the following Axygen Multiwell Plates and Sealing Mats

- 2.0ml deep well plates, 96- round wells
- 2.0ml deep well plates, 96- square wells
- 1.1ml deep well plates, 96- round wells
- 600µl deep well plates, 96- round wells
- 120µl microwell plates, 384- square wells
- 240µl microwell plates, 384- square wells
- ImpermaMat for 384 well, 120µl and 240µl, microplates with square wells
- ImpermaMat for 96 well, 1.1ml and 2.0ml, deep well plates with round wells
- Axymat for 600µl 96 well deep well plates
- Axymat for 1.1ml and 2ml 96 well round deep well plates
- Axymat for 2ml 96 well square deep well plates
- Axymat for 384 well, 120µl and 240µl, microplates with square wells

Request an Axygen catalog for full technical ordering information

Pre-adjustment of EasyPress



To adjust for the first time or for a new a plate, lower the handle and place a plate on the 'plate slide' with no Axymat.



Adjust the large thumbwheel (spin thumbwheel to right side to lower the press head and left to elevate) until the press head is above the height of the plate and slide the plate into the press.



Adjust the thumbwheel to lower the press head until the press head makes light contact with the plate, return the handle the top position. You are now ready to seal plates.



Caution, improper adjustment may cause damage to the EasyPress, do not adjust thumbwheel to exert excess force when attempting to seal plates.

Operation of EasyPress



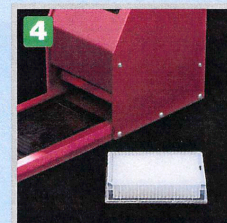
With the 'plate slide' in the out position, place the appropriate Axymat on top of the plate, align 2 - 3 corners of the Axymats and place the plate with the Axymat into the 'plate slide'.



Move the 'plate slide' into the press.



Holding the outer edge of the 'base plate', gently pull down on the handle to the bottom of its stroke and then return the handle to the top. Slide the plate tray out of the EasyPress and remove your plate.



You are now ready to continue sealing plates of the same size.

E-Gel 96 Operation

E-Gel® 96 Gels

Catalog nos. G7008-01, G7008-02

25-0419 Version I; 18 April 2005

QUICK
REFERENCE
CARD

Instructions are provided below for using E-Gel® 96 Gels with the E-Base™. For more details, refer to the E-Gel® Technical Guide available at www.invitrogen.com or contact Technical Service.

Preparing Samples

- Use 20-100 ng DNA per band for samples containing one unique band or up to 500 ng per lane for samples containing multiple bands.
- Prepare DNA samples in a **total sample volume of 20 µl for E-Gel® 96 gels** in deionized water or loading buffer (recommended final loading buffer concentration is 10 mM Tris-HCl; 1 mM EDTA, pH 7.5; 0.005% bromophenol blue; and 0.005% xylene cyanol FF).
- Dilute **high salt samples** (samples with >50 mM NaCl, >100 mM KCl, >10 mM acetate ions, >10 mM EDTA), 2- to 20-fold in deionized water, TE, or loading buffer in final volume of 20 µl.

Selecting Program On E-Base™

The recommended program for E-Gel® 96 gel is EG and the run time is 12 minutes.

1. Plug the Mother E-Base™ into an electrical outlet. Connect the Daughter E-Base™ to a Mother E-Base™ or another Daughter E-Base™ connected to a Mother E-Base™.
2. Press and release the pwr/prg (power/program) button on the base to select program EG.

Note: The E-Gel® 96 gels are also compatible with the E-Gel® 96 base available previously from Invitrogen. For using E-Gel® 96 gels with E-Gel® 96 base, refer to the E-Gel® Technical Guide.

Loading and Running E-Gel® 96 Gels

Load each gel within 30 minutes of removing gel from the package and run within 15 minutes of loading.

1. Remove gel from the package and remove plastic comb from the gel.
2. Slide gel into the two electrode connections on the Mother or Daughter E-Base™. If gel is properly inserted, a fan in the base begins to run, a red light illuminates, and digital display shows 12 minutes.

3. Load 20 µl prepared DNA sample into the well. Keep all sample volumes uniform. Load samples manually, with a multichannel pipettor, or use robotic loading devices (8-, 12-, 96-tip).

Note: To ensure proper sample loading with robotic loading device, align the robotic tip assembly (see E-Gel® Technical Guide for details).

4. Load appropriate DNA markers in the marker wells. Be sure the marker salt concentration is similar to that of the adjacent samples.

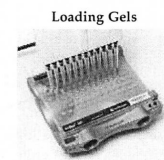
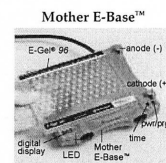
1% gel: E-Gel® 96 High Range DNA Marker

2% gel: E-Gel® Low Range Quantitative DNA Marker

5. Load 20 µl sample buffer containing the same salt concentration as the sample into any empty wells.

6. To begin electrophoresis, press and release the pwr/prg button on the E-Base™. The red light changes to green.

7. At the end of the run (signaled with a flashing red light and rapid beeping), press and release the pwr/prg button to stop the beeping.



1600 Faraday Avenue • Carlsbad • CA 92008
Toll Free: 800 955 6288 • F: 760 602 6500
tech_service@invitrogen.com

 **invitrogen™**

Contact Information for Other Countries:
See our Website: www.invitrogen.com

E-Gel® 96 Gels

Catalog nos. G7008-01, G7008-02
25-0419 Version I; 18 April 2005

QUICK
REFERENCE
CARD

Instructions are provided below for using E-Gel® 96 Gels with the E-Base™. For more details, refer to the E-Gel® Technical Guide available at www.invitrogen.com or contact Technical Service.

Preparing Samples

- Use 20-100 ng DNA per band for samples containing one unique band or up to 500 ng per lane for samples containing multiple bands.
- Prepare DNA samples in a **total sample volume of 20 µl for E-Gel® 96 gels** in deionized water or loading buffer (recommended final loading buffer concentration is 10 mM Tris-HCl; 1 mM EDTA, pH 7.5; 0.005% bromophenol blue; and 0.005% xylene cyanol FF).
- Dilute **high salt samples** (samples with >50 mM NaCl, >100 mM KCl, >10 mM acetate ions, >10 mM EDTA), 2- to 20-fold in deionized water, TE, or loading buffer in final volume of 20 µl.

Selecting Program On E-Base™

The recommended program for E-Gel® 96 gel is EG and the run time is 12 minutes.

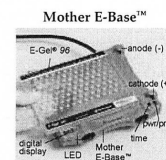
1. Plug the Mother E-Base™ into an electrical outlet. Connect the Daughter E-Base™ to a Mother E-Base™ or another Daughter E-Base™ connected to a Mother E-Base™.
2. Press and release the pwr/prg (power/program) button on the base to select program EG.

Note: The E-Gel® 96 gels are also compatible with the E-Gel® 96 base available previously from Invitrogen. For using E-Gel® 96 gels with E-Gel® 96 base, refer to the E-Gel® Technical Guide.

Loading and Running E-Gel® 96 Gels

Load each gel within 30 minutes of removing gel from the package and run within 15 minutes of loading.

1. Remove gel from the package and remove plastic comb from the gel.
2. Slide gel into the two electrode connections on the Mother or Daughter E-Base™. If gel is properly inserted, a fan in the base begins to run, a red light illuminates, and digital display shows 12 minutes.
3. Load 20 µl prepared DNA sample into the well. Keep all sample volumes uniform. Load samples manually, with a multichannel pipettor, or use robotic loading devices (8-, 12-, 96-tip).
Note: To ensure proper sample loading with robotic loading device, align the robotic tip assembly (see E-Gel® Technical Guide for details).
4. Load appropriate DNA markers in the marker wells. Be sure the marker salt concentration is similar to that of the adjacent samples.
1% gel: E-Gel® 96 High Range DNA Marker
2% gel: E-Gel® Low Range Quantitative DNA Marker
5. Load 20 µl sample buffer containing the same salt concentration as the sample into any empty wells.
6. To begin electrophoresis, press and release the pwr/prg button on the E-Base™. The red light changes to green.
7. At the end of the run (signaled with a flashing red light and rapid beeping), press and release the pwr/prg button to stop the beeping.



1600 Faraday Avenue • Carlsbad • CA 92008
Toll Free: 800 955 6288 • F: 760 602 6500
tech_service@invitrogen.com

 **invitrogen™**

Contact Information for Other Countries:
See our Website: www.invitrogen.com

DNA Ladder for E-Gel 96



E-Gel® Low Range Quantitative DNA Ladder

Cat. No. 12373-031

Conc. 175 ng/10 µl

Introduction

The E-Gel® Low Range Quantitative DNA Ladder is designed for use as a molecular weight standard and for estimating the amount of DNA.

The important features of the ladder are listed below:

- Consists of five linear double-strand DNA fragments (100-2000 bp)
- Suitable for estimating molecular weight and quantity of DNA fragments in agarose gel electrophoresis
- Provided in a ready-to-use format
- Visualized by ethidium bromide staining

Specifications

Contents: 1 ml of E-Gel® Low Range Quantitative DNA Ladder. Sufficient ladder is provided for 100 applications.

Storage Buffer: 8 mM Tris-HCl, pH 7.5; 1 mM EDTA; 5% glycerol; and 0.005% Orange G.

Storage: Store at room temperature or 4°C.

Directions for Use

The E-Gel® Low Range Quantitative DNA Ladder is provided in a ready-to-use format. There is no need to heat the ladder prior to loading. Use loading volumes listed below to obtain the best results:

Gel	Ladder	Water	Total Volume
Standard agarose	10 µl	—	10 µl
E-Gel® 96	10 µl	10 µl	20 µl
E-Gel® double comb			
Marker lane	5 µl	5 µl	10 µl
Sample lane	10 µl	10 µl	20 µl
E-Gel® single comb	10 µl	10 µl	20 µl

Part No. 12373031.PPS

Doc. Rev. 081602

This product is distributed for laboratory research only. CAUTION: Not for diagnostic use. The safety and efficacy of this product in diagnostic or other clinical uses has not been established.

For technical questions about this product, call the Invitrogen Tech-Line™ U.S.A. 800 955 6288



E-Gel[®] Low Range Quantitative DNA Ladder

Cat. No. 12373-031

Conc. 175 ng/10 μ l

Introduction

The E-Gel[®] Low Range Quantitative DNA Ladder is designed for use as a molecular weight standard and for estimating the amount of DNA.

The important features of the ladder are listed below:

- Consists of five linear double-strand DNA fragments (100-2000 bp)
- Suitable for estimating molecular weight and quantity of DNA fragments in agarose gel electrophoresis
- Provided in a ready-to-use format
- Visualized by ethidium bromide staining

Specifications

Contents: 1 ml of E-Gel[®] Low Range Quantitative DNA Ladder.
Sufficient ladder is provided for 100 applications.

Storage Buffer: 8 mM Tris-HCl, pH 7.5; 1 mM EDTA; 5% glycerol;
and 0.005% Orange G.

Storage: Store at room temperature or 4°C.

Directions for Use

The E-Gel[®] Low Range Quantitative DNA Ladder is provided in a ready-to-use format. There is no need to heat the ladder prior to loading. Use loading volumes listed below to obtain the best results:

Gel	Ladder	Water	Total Volume
Standard agarose	10 μ l	--	10 μ l
E-Gel [®] 96	10 μ l	10 μ l	20 μ l
E-Gel [®] double comb			
Marker lane	5 μ l	5 μ l	10 μ l
Sample lane	10 μ l	10 μ l	20 μ l
E-Gel [®] single comb	10 μ l	10 μ l	20 μ l

Part No. 12373031.PPS

Doc. Rev. 081602

This product is distributed for laboratory research only. CAUTION: Not for diagnostic use. The safety and efficacy of this product in diagnostic or other clinical uses has not been established.

For technical questions about this product, call the Invitrogen Tech-Line[™] U.S.A. 800 955 6288