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 Gβ1 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 Gβ1 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 Gβ1 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.