# Exploring Protein Sequence Space Using Computationally Directed Recombination

Thesis by

Michelle Margaret Meyer

In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

California Institute of Technology
Pasadena, California, USA
2006

(Defended May 24, 2006)

# Acknowledgements

My time at Caltech has been a gift. I have had the opportunity to work with many great people who have given me much in terms of both personal and professional development. First I would like to thank my advisor, Frances Arnold. She has given me guidance and support throughout this process but also allowed me the freedom to make my own mistakes as well as achieve my own victories. She has also been very understanding in times of great stress, and truly a champion when it comes to giving feedback quickly. This work would not have been possible without her, and certainly would never have been done in time. My other committee members, Steve Mayo, Doug Rees, and Rob Phillip, have provided many comments that have guided the direction of this research.

The work described here was not conducted in isolation, but with many incredible collaborators. Joff Silberg was an important guide during my first year here, and Allan Drummond, Lisa Hochrein, and Zhongyi Lu have helped me with various experiments. While only the mutagenesis experiments Lisa started made it into this document, all the experiments were essential for realizing what the right direction should be. I also need to thank Jennifer Keefe in the Mayo group for her help in obtaining the circular dichroism measurements.

Aside from all the people that have worked on experiments with me, I also owe a great deal to a group of less experimentally inclined collaborators; the “theoretical people,” who made all of this work possible. Christopher Voigt first told me to write my own code, and made me aware that I could. Jeff Endelman allowed much of this work to take place developing many improvements in our design and analysis methodologies.

Allan Drummond is always willing to talk about crazy ideas. He continues to give me confidence that when I play with numbers, I am making a worthwhile contribution.

Aside from all of my immediate collaborators there have been discussions with many people over the years that have contributed to not just to science but also to maintaining sanity. Karen Wawrousek, Jenny Witman, and Lisa Hochrein are not only good friends but also kept me from falling apart during a difficult time. Chris Otey once reminded me that it is ok, and sometimes better, to just let go of the problem. Jesse Bloom has been a good friend and a great person to hang out with in the lab, especially since we have the same taste in radio stations. Marco Landwehr continues to remind me that I cannot work all the time, so computer games do still have a place. I have had many morning coffee breaks with Jorge Rodriguez and Geethani Bandara that have served as a welcome reason to procrastinate. Finally, conversations with my other officemates not yet named, Mike Chen and Alex Tobias, have reminded me from time to time exactly why we keep coming in every day.

My path, however, did not start when I reached Caltech. There were many people that helped bring me here. My high school biology teacher Brian Pelkey always raised the bar and demanded true excellence. In doing so he showed me how much more there was to science and introduced me to what I consider a fantastic puzzle: how life works. My undergraduate research advisor at Rice University, Seiichi Matsuda not only introduced me to the work that I love, but also taught me to observe carefully and to remember that there is usually a reason *why* an experiment does not work.

My parents, Robert Meyer and Susan Conry Meyer, have always encouraged me to dream big dreams. They have been a constant source of support through this entire

process. My husband Matthew has, in addition to answering the occasional math or programming question, often provided a surrogate sense of humor when mine was sadly absent. He has also served to remind me how important it is to do what you love; otherwise life is just not the same.

# Abstract

Evolution has provided us with many protein sequences. However, these sequences represent a very small fraction of the possible sequences. In the laboratory, scientists have explored areas of sequence space not represented by natural proteins both to better understand natural proteins, and to create new proteins with desirable properties. The principle mechanism used to explore protein sequence space is mutagenesis. However, recombination of homologous genes can also explore regions of sequence space rich with folded and functional proteins.

In this work we demonstrated using a β-lactamase model system that a computation energy function (SCHEMA) can predict which of the chimeras made by recombining distantly related proteins are likely to fold. SCHEMA uses protein sequence and structure information to identify pairwise amino acid interactions disrupted by recombination. Using SCHEMA we designed libraries of chimeric β-lactamases. These libraries were intended to have a high fraction of folded variants, while incorporating many amino acid substitutions compared with the parental proteins. The chimeras in these libraries were characterized to determine whether they retain the parental function and what new substrate specificities could be obtained.

To identify critical variables for determining whether a chimera functions, we used logistic regression analysis to analyze functional and nonfunctional chimeras. From this analysis it is apparent that both two-body (pairwise) and one-body terms play a significant role in determining whether a chimera functions. We also used random mutagenesis to restore functionality to nonfunctional chimeras showing that a thermostabilizing mutation can rescue approximately 5% of the nonfunctional chimeras.

The one-body terms that appear significant for determining whether a chimera functions are not explicitly counted by SCHEMA when predicting chimera folding. To estimate the effects on chimera folding represented by the one-body terms, we developed an additional measure to predict chimera folding based on just the chimera amino acid sequence and a multiple sequence alignment of homologous proteins. This measure is predictive of chimera folding alone, and when combined with the pairwise SCHEMA energy increases the accuracy of the folding predictions compared to SCHEMA.

# Table of Contents

# List of Tables

# List of Figures