

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)

©2006

Michelle Margaret Meyer

All Rights Reserved

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

Acknowledgements	iii
Abstract	vi
Table of Contents	viii
List of Tables	ix
List of Figures	x
Chapter I	1
Recombination as an Evolutionary Search Strategy	
Chapter II	8
Library Analysis of SCHEMA-Guided Recombination	
Chapter III	27
Design of Site-Directed Recombination Libraries	
Chapter IV	50
Construction and Characterization of Site-Directed Recombination Libraries	
Chapter V	75
Using Chimeras to Identify Determinants of β -lactamase Function	
Chapter VI	99
Mutagenesis to Restore Chimera Function	
Chapter VII	112
Accuracy of SCHEMA Predictions of Chimera Folding on Different Protein Scaffolds	
Chapter VIII	132
Improving Predictions of Chimera Folding Using Multiple Sequence Alignments	
Appendix I	150
Computer Code	
Appendix II	172
Primers and Oligonucleotides Used for Construction and Analysis of Recombination Libraries.	
Appendix III	180
Characterized Chimeras	
References	195

List of Tables

III-1.	Characteristics of the Constructed Libraries	41
IV-1.	MICs of TEM-1, PSE-4 and SED-1 on β -lactam and Cephalosporin Antibiotics	61
V-1.	Characterized Chimeras Inheriting Blocks 1, 7 and 8 from SED-1	79
V-2.	Energies Assigned to Important Interactions by Logistic Regression Analysis	84
V-3.	Residue-Residue Contacts between Block Pairs and within Each Block	85
V-4.	Length and Sequence Identity between Each Pair of Parental Proteins for Each Block	85
V-5.	Characterized Sets of Chimeras Differing Only by Block 2	89
VI-1.	Randomly Mutated Chimeras	101
VI-2.	Mutations that Rescue Nonfunctional Chimeras	102
VI-3.	Randomly Chosen Chimeras M182T was Introduced Into	107
VII-1.	Comparison of Cytochrome P450 and Lactamase Library Chimera Properties	114
VII-2.	Homologous Structures Used to Calculate E	123
VIII-1.	One- and Two-body Energy Terms Used to Calculate LRA Energies for Cytochromes P450	148
VIII-2.	One- and Two-body Energy Terms Used to Calculate LRA Energies for Lactamases	148
AII-1.	Oligonucleotides for Gene Fragments of RandE:APST and RandEPST Libraries	173
AII-2.	Other Primers Involved with Synthesis of RandE:APST and RandE:PST libraries	175
AII-3.	Primers for Construction of RASPP:PST Using SISDC	176
AII-4.	Half-Library PCR Amplification Primer Sets	178
AII-5.	Probes for DNA Hybridization to Sequence Chimeras	179
AIII-1.	Functional Chimeras from the Naïve Library (RASPP:PST)	180
AIII-2.	Nonfunctional Chimeras from the Naïve Library (RASPP:PST)	182
AIII-3.	Additional Functional Chimeras (RASPP:PST)	187
AIII-4.	Cytochrome P450 Naïve Functional and Nonfunctional Chimeras	188

List of Figures

I-1.	Overview of Site-Directed Recombination	6
II-1.	Pairwise Sequence Identity and RMSD of Crystallized β -lactamases	11
II-2.	Overview of Library Designed Using SCHEMA Profile	12
II-3.	Incorporation of <i>pse-4</i> and <i>tem-1</i> at Different Sequence Positions	13
II-4.	Sequences of Functional Chimeras	15
II-5.	Relationship between <i>E</i> and Chimera Function	17
II-6.	Relationship between <i>m</i> and Chimera Function	18
II-7.	<i>E</i> vs. <i>m</i> for all Possible Chimeras in Library	18
III-1.	Example of RASPP $\langle E \rangle$ vs. $\langle m \rangle$ Curve	31
III-2.	Rank Ordering of Test Libraries by F_{folded}	33
III-3.	Rank Ordering of Test Libraries by m_{folded}	35
III-4.	Comparison of Test Libraries $\langle E \rangle$ vs. $\langle m \rangle$ and $\langle m/E \rangle$ vs. $\langle m \rangle$	36
III-5.	Library Design Using Random Enumeration	40
III-6.	Overview of RandE:APST Library	42
III-7.	RASPP curves for 7, 8, and 9 crossovers between TEM-1, PSE-4, and SED-1	43
III-8.	Library Design Using RASPP	44
III-9.	Overview of RASPP:PST Library	46
IV-1.	Overview of Combinatorial Gene Assembly	51
IV-2.	Overview of Library Construction Using Synthetic Gene Fragments	54
IV-3.	Overview of Library Construction Using SISDC	56
IV-4.	The Characterized Portion of the RASPP:PST Library	59
IV-5.	Distribution of Chimeras in the Theoretical vs. Characterized RASPP:PST Library	60
IV-6.	Positive and Negative Control Measurements for GFP Folding Assay	63
IV-7.	GFP Folding Assay Applied to a Library of Chimeras	63
IV-8.	Chimera Fraction Functional with Respect to <i>m</i> and <i>E</i>	66
IV-9.	Overview of Functional Lactamase Chimeras	67
V-1.	Functional Lactamases Cluster in Sequence Space	77
V-2.	Structural Features Important for Cefotaxime Hydrolysis	81
V-3.	Logistic Regression Analysis of Lactamase Chimeras	84
V-4.	Sequence Alignment of the Parent Proteins at Block 2	88
V-5.	MIC of all Functional Chimeras vs. Functional Chimeras with Block 2 from SED-1	88
V-6.	Periplasmic Extracts of Characterized Chimeras	90
VI-1.	<i>E</i> vs. <i>m</i> for Chimeras Containing M182T Mutation	105
VI-2.	Probability of M182T Rescuing Chimera Function	106
VI-3.	Extrapolated Increase in Fraction of Functional Chimeras if the M182T Mutation was Incorporated into All Chimeras	108
VII-1.	Library Blocks Mapped to Three-Dimensional Structures of TEM-1 (Lactamase) and CYP102A1 (Cytochrome P450)	115
VII-2.	<i>E</i> vs. <i>m</i> Distributions for β -lactamase and Cytochrome P450 Libraries	116
VII-3.	Probability of Folding (P_f) for Lactamase and Cytochrome P450 Libraries with respect to <i>E</i>	117

VII-4.	Different $P_f(E)$ Calculated for Different Groups of Chimeras	118
VII-5.	Total Available Mutual Information and Mutual Information between Chimera Folding and E and m .	120
VII-6.	The Mutual Information between Chimera Folding and E Determined Using Homologous Structures vs. Sequence Identity.	124
VII-7.	Mutual Information between Chimera Folding and E Determined Using Homologous Structures vs. Difference in Chimera Length	125
VII-8.	The Mutual Information between Chimera Folding and E Determined Using $C\alpha$ and Covarying Amino Acids.	127
VIII-1.	Mutual Information between Chimera Folding and E and LRA Energies	133
VIII-2.	Determination of P_{aa}	136
VIII-3.	Distribution of w for Folded and Unfolded Chimeras in Lactamase and Cytochrome P450 Libraries	138
VIII-4.	w vs. m for Folded and Unfolded Chimeras in Lactamase and Cytochrome P450 Libraries	139
VIII-5.	Mutual Information between Chimera Folding and E , $1/w$, and $-w$	140
VIII-6.	w_{blocks} for Cytochrome P450 and Lactamase Library Sequence Blocks	141
VIII-7.	Mutual Information between Chimera Folding and E , $1/w$, and E_w	143
VIII-8.	E vs. m and E_w vs. m for Lactamase and Cytochrome P450 Chimeras	144