ABSTRACT

Title of Document: USING RECOMBINANT PCR TO STUDY SEQUENCE POLYMORPHISMS IN A FAMILY OF COMPACT ALBUMIN BINDING DOMAINS

David Anthony Rozak, Doctor of Philosophy, 2005

Directed By: Dr. Philip N. Bryan, Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute.

Sixteen homologs of a compact albumin binding domain were previously identified in six proteins and four bacterial species. These domains, which exhibit varied affinities for different albumins, have been shown to support bacterial growth *in vitro*, and may contribute to host specificity. This dissertation describes the development of a robust PCR-based recombination technique, which is applied to representatives of the albumin binding domain to identify and understand the impact of sequence polymorphisms on domain stability and function. Analysis of phage-selected recombinants highlights the potential impact of multiple mutations in stabilizing the selected domains and improving albumin binding through gains in hydrophilic surface area, direct modifications to the binding interface, and subtle changes in the position of the third helix. The most common mutant was encoded by three fourths of the selected phage and exhibited 5 and 10 fold increases in human and guinea pig albumin binding constants compared to the wild type streptococcal domain (G148-

GA3). This study serves to validate further the application of *in vitro* recombination and phage display in the analysis of sequence polymorphisms. The recombination technique itself is shown to be well suited for producing multiple recombination events among compact heterologous domains and appears to offer several advantages over traditional DNA shuffling techniques.

USING RECOMBINANT PCR TO STUDY SEQUENCE POLYMORPHISMS IN A FAMILY OF COMPACT ALBUMIN BINDING DOMAINS

By

David Anthony Rozak

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2005

Advisory Committee:
Dr. Philip Bryan, Chair
Dr. Jeffery Davis, Dean's Representative
Dr. Douglas Julin
Dr. George Lorimer
Dr. John Orban
Dr. Daniel Stein

Dedication

To Michelle, Nicholas, and Duncan.

Acknowledgements

There are many at the University of Maryland and the Center for Advanced Research in Biotechnology who made this work possible. I am particularly grateful for the unwavering confidence and support of my advisor, Phil Bryan, who encouraged me to develop and pursue the recombinogenic strategy described here. Dr. Bryan's guidance was further enhanced by the careful attention John Orban, Douglas Julin, George Lorimer, Daniel Stein, and Jeffery Davis gave to my research while graciously serving on my advisory committee. I'd like to specifically recognize my colleagues Kathryn Fisher, Patrick Alexander, and Biao Ruan for their friendship and support throughout this endeavor. Special thanks go to Yanan He and Yihong Chen who obtained the highly informative NMR structure for one of the phage-selected domains discussed in this paper. Finally, the calorimetric work described in these pages would not have been possible without the patient assistance of Fred Schwarz, Chittoor Swaminathan, and Masanori Osawa. This work was financed by NIH grant GM62154.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

amp	Ampicillin
amp^R	Ampicillin resistance
BSA	Bovine serum albumin
CD	Circular dichroism
CIP	Calf intestinal alkaline phosphatase
C_P	Heat capacity
DSC	Differential scanning calorimetry
G	Gibbs free energy
GA	Protein G-related albumin binding
GPSA	Guinea pig serum albumin
H	Enthalpy
H-D	Hydrogen-deuterium
HSA	Human serum albumin
IgG	Immunoglobulin G
IPTG	Isopropyl β-D-thio-galactopyranoside
ITC	Isothermal titration calorimetry
NMR	Nuclear magnetic resonance
ORF	Open reading frame
OR-PCR	Offset recombinant polymerase chain reaction
PCR	Polymerase chain reaction
PSD	Phage-selected domain
S	Entropy
SDS-PAGE	Sodium dodecyl sulfate-polyacrylamide gel electrophoresis
T	Temperature

TD	Template domain
T_m	Melting temperature
UV	Ultra-violet
X-Gal	5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside

Chapter 1: Motivations for Developing and Applying a Novel Recombinogenic Technique to a Family of Albumin Binding Domains

Exploring Natural Sequence Space

Many families of homologous proteins describe rich sequence spaces, which encode biologically significant variations in protein structure and function. While many of the sequence polymorphisms defined by families of homologous proteins are likely to represent functionally neutral mutations, some encode significant biochemical changes that support the specific lifestyle of the host organism. Efforts to identify and understand functional polymorphisms could support predictions about the biochemical properties of uncharacterized domains and guide molecular biologists in engineering mutants with desirable traits.

The GA albumin binding module (de Chateau and Bjorck 1994), which is found on the surface of several bacterial pathogens, provides an excellent example of a protein family with polymorphisms that encode a broad spectrum of protein behavior. The three helix 46 amino acid albumin binding module has been associated with 16 different domains found in six proteins and four bacterial species (Johansson, de Chateau et al. 1995; Johansson, de Chateau et al. 1997). Experiments suggests that the domain supports bacterial growth *in vitro*, possibly by scavenging albumin-bound nutrients (de Chateau, Holst et al. 1996). As a result, the various affinities for albumin species that have been identified in native GA domains (Johansson, Frick et al. 2002) may play a role in supporting bacterial tropisms.

Structural and competitive binding studies of two albumin binding domains reveal significant differences in the backbone dynamics and albumin binding capabilities of these homologs. Specifically, hydrogen-deuterium (H-D) exchange data for the third albumin binding domain (G148-GA3) of streptococcal protein G and the single albumin binding domain (ALB8-GA) found in the *Finegoldia magna* (formerly *Peptostreptococcus magnus*) PAB protein suggest that the former has a more dynamic structure than the latter (Johansson, Nilsson et al. 2002). Competitive binding experiments reveal that G148-GA3 can efficiently bind a much broader range of albumins than ALB8-GA and have lead the researchers to propose that the flexible G148-GA3 domain may somehow contribute to its wide affinity for albumins from different species (Johansson, Nilsson et al. 2002).

And yet, despite the availability of NMR structural data for both domains (Johansson, de Chateau et al. 1995; Johansson, de Chateau et al. 1997; Johansson, Frick et al. 2002) and a recently obtained crystal structure of ALB8-GA complexed with HSA (Lejon, Frick et al. 2004), little is known about the impact of sequence polymorphisms on the abilities of these two domains to bind different species of albumins. Even less is known about the manners in which sequence polymorphisms encode novel functionality within the remaining fourteen domains of the GA module.

Recombinogenic Analysis of Sequence Polymorphisms

One appealing strategy for unraveling the phenotypic impact of polymorphisms found among members of the GA module and other protein families involves shuffling family homologs to produce a library of randomly recombined constructs.

These recombinant sequences can be probed by panning for functional mutants displayed on the surfaces of filamentous phage. Specifically, Zhao and Arnold used DNA shuffling and phage selection to determine which residues in an engineered subtilisin contributed to enhanced thermostability when compared with the wild-type protein (Zhao and Arnold 1997). The team showed that positive functional mutations occurred in a high percentage of the selected recombinant population whereas neutral mutations occurred in about half the selected samples and deleterious mutations were largely absent from screened sequences. However, Zhao and Arnold's initial attempt at using recombinogenics to unravel the associations between sequence polymorphisms and protein function was narrowly focused on two DNA species with a high degree of homology. Similar analysis of the polymorphisms defined by the 16 members of the GA module presents a far greater challenge, largely because existing recombinogenic techniques are ill equipped to promote efficient recombination of the compact three-helix domains.

Cumulative evidence suggests that the shuffling technique developed by Stemmer (1994) and employed by Zhao and Arnold in the above experiment is likely to encounter difficulties when applied to compact heterologous stretches of DNA similar to those contained in the GA module and other protein families. These troubles arise from a dramatic increase in the likelihood of homoduplex formation and improper fragment assembly as the complexity of the shuffled sequence space grows.

DNA Shuffling and Related Strategies

DNA shuffling involves a two step process in which homologous genes are randomly fragmented with DNase I and reassembled via primerless PCR to form a library of shuffled genes with novel assortments of genetic markers. As researchers used Stemmer's original technique (Stemmer, 1994) to shuffle increasingly heterologous sets of genes, they became aware of a propensity for fragments to form homoduplexes with their own kind rather than annealing to fragments from the other species to form novel constructs. This process, which leads to the reassembly of native parental genes, has been termed parental recombination. After experiencing an overwhelming preference for parental recombination when DNA shuffling was applied to genes encoding two proteins with 84% sequence identity (Kikuchi, Ohnishi et al. 1999), Kikuchi successfully experimented with protocols involving restriction enzyme digests (Kikuchi, Ohnishi et al. 1999) and fragmentation of single stranded phagemid DNA (Kikuchi, Ohnishi et al. 2000) to reduce the likelihood of homoduplex formation during PCR assembly. In the first protocol, Kikuchi cut and religated mixed homologs at internal restriction sites to reduce the chance of parental recombination by avoiding the use of DNase I and PCR. Kikuchi's second technique succeeded by fragmenting complementary single stranded DNA from two homologs to ensure that primerless PCR progressed by assembling the complementary homologs. Others were similarly driven by parental recombination to create Incremental Truncation for the Creation of Hybrid enzYmes (ITCHY) (Ostermeier, Nixon et al. 1999), a DNase I enhanced version of ITCHY, known as SCRATCHY (Lutz, Ostermeier et al. 2001), and Degenerate Oligonucleotide Gene Shuffling

(DOGS) (Gibbs, Nevalainen et al. 2001) as alternatives to Stemmer's shuffling strategy. ITCHY relies on an exonuclease to create nested blunt-end fragments, which can be ligated to form recombinant genes with single crossover events. SCRATCHY applies Stemmer's DNA shuffling protocol to ITCHY products in an effort to increase the number of crossover events. DOGS uses pairs of degenerate primers to amplify homologous gene segments. These overlapping amplification products are then reassembled into a full length gene with primerless PCR.

Experimentally validated computer models suggest that Stemmer's DNA shuffling technique is also limited in its ability to create multiple crossover events in compact heterologous domains. Independent attempts to model shuffling reactions revealed a strong propensity for DNA fragments under 20 nt in length to produce out-of-order assemblies that encoded dysfunctional proteins and interfere with subsequent amplification and cloning efforts (Moore, Maranas et al. 2001; Moore and Maranas 2002; Maheshri and Schaffer 2003). One model predicted that half of all 15 nt fragments formed out-of-order assemblies when annealing took place at 55°C and that these results were largely independent of sequence homology (Moore, Maranas et al. 2001). Another model suggests that out-of-order assembly further increases when annealing temperatures were reduced to avoid parental recombination described above (Maheshri and Schaffer 2003).

In addition to promoting misassembly, computer models predict that gene fragments of the size required to promote multiple crossover events in small domains significantly reduced the fraction of full-length products generated during primerless PCR. Only 0.01% of all DNA fragments with an average fragment size of 100 nt and

minimum fragment size of 50 nt contributed to the production of full-length genes (Maheshri and Schaffer 2003). No reassembly is expected when the average fragment size falls below 50 bp for a 2.2 kb gene that encodes a protein with 80% sequence identity (Maheshri and Schaffer 2003).

Given these limitations, Maheshri's model predicts that a mix of 20-50 bp fragments will generate no more than 2.7 observable crossover events per kb (Maheshri and Schaffer 2003)—a value that is in close agreement with the 2.3 non-silent crossover events per kb observed *in vitro* under identical conditions (Zhao and Arnold 1997) and is ill-suited for efforts to effectively recombine members of the GA module and other small globular proteins.

Although ITCHY, SCRATCHY, DOGS, and the other techniques mentioned above successfully overcome the problems of parental recombination, they do little to promote high-densities of crossover events or avoid the limitations associated with accurate assembly of recombinant genes. ITCHY is incapable of creating more than one crossover event per sequence. SCRATCHY, which was developed to exceed the single crossover event produced by ITCHY through the application of Stemmer's DNA shuffling technique to ITCHY products (Lutz, Ostermeier et al. 2001), probably suffers from similar fragment size constraints as Stemmers original approach. Both DOGS and Kikuchi's restriction digest protocol rely on predefined crossover points, which are unable to produce a random assortment of high density crossover events (Kikuchi, Ohnishi et al. 1999; Gibbs, Nevalainen et al. 2001). Even Kikuchi's more flexible phagemid strategy is subject to the same primerless PCR dynamics that severely limit the fraction of full-length gene assemblies from small DNA fragments

(Kikuchi, Ohnishi et al. 2000). Another approach, referred to here as recombinant PCR, may offer solutions to many of the difficulties encountered by related DNA shuffling strategies.

Recombinant PCR

Recombinant PCR is an alternative strategy that offers an appealing option for *in vitro* recombination of heterologous sequences because it does not require DNA fragmentation prior to PCR assembly. Rather, these related PCR-based techniques exploit the premature termination of polymerization reactions to produce nested oligonucleotides that can be extended on homologous templates during subsequent rounds of PCR.

The natural recombinogenic nature of PCR was first recognized soon after the advent of the amplification technology and has been documented on multiple occasions since then (Saiki, Gelfand et al. 1988; Meyerhans, Vartanian et al. 1990; Yang, Wang et al. 1996; Bradley and Hillis 1997). However, the relatively low fraction of recombinant fragments produced during the PCR amplification of mixed alleles made the technique ill suited for DNA recombination applications.

The ability of PCR to generate recombinant fragments is limited by the extent to which partially extended primers are (a) terminated within the recombinant region and (b) complexed with heterologous templates during subsequent elongation rounds. Standard PCR amplification protocols limit the likelihood of heterologous recombination by employing long extension phases and high concentrations of unextended primers, which compete with fewer partially extended primers to bind the

templates. Researchers have shown that when reactions are modified to reduce each of these effects they produce significantly elevated levels of recombinant DNA.

For example, the Staggered Extension Process (StEP) achieved a 39% chimeric recombination rate between markers separated by 113 nt by using a 5 second combined annealing-elongation phase to increase the likelihood that polymerization will be halted multiple times within the recombinant region (Zhao, Giver et al. 1998). StEP also increased the chance that partially extended primers will anneal to templates during subsequent rounds by significantly reducing the initial primer concentration.

Taking a somewhat different approach, Judo achieved a 21% chimeric recombination frequency between markers separated by 287 nt when he added an additional high temperature annealing phase that would favor template interactions with the longer partially extended primers over shorter unextended primers (Judo, Wedel et al. 1998). While Zhao and Judo independently achieved high recombination rates for recombinant PCR without resorting to the additional fragmentation step that can prove problematic for DNA shuffling, they did not probe the extent to which their technologies can be successfully applied to sequences of decreasing size and homogeneity.

Research Overview

This dissertation describes my efforts to develop a novel PCR-based strategy, which is capable of generating multiple recombination events among compact heterologous domains, and apply it to a functional analysis of the GA sequence space.

In an effort to devise a recombination technique that is robust enough to handle recombination among multiple heterologous domains of the type described by the GA module I conceived of and tested a recombinant strategy that produces unparalleled results by placing the recombinant region near one end of the amplicon in a typical polymerase chain reaction. This approach, which is further described and characterized in Chapter 2, has a number of advantages over previously reported recombination techniques. Specifically, by locating the recombination region near one end of the amplicon I was able to:

- Achieve a high recombination rate without resorting to the highly abbreviated elongation phases or lower primer concentrations applied by StEP.

- Take advantage of long stretches of identical template to significantly reduce the preference for homoduplex formation and avoid the chance of out-of-sequence assembly.

- Exploit the accumulation of recombinant template populations during the PCR amplification process to increase the overall recombination rate and generate multiple crossover events within a compact recombinant region.

Using a *lacZ* reporter system, I show that in a typical amplification reaction *Pfu* polymerase generated chimeric crossover events in 13% of the population when markers were separated by only 70 nt. The fraction of recombinant sequences reached 42% after six consecutive rounds of PCR, a value close to the 50% expected from a fully shuffled population. When homology within the compact recombinant region was reduced to 82%, the recombination frequency dropped by nearly half for a single

amplification reaction and crossover events were clustered towards one end of the domain. Surprisingly, recombination frequencies for template populations with high and low sequence homologies converged after just four rounds of PCR, suggesting that the exponential accumulation of chimeric molecules in the PCR mix serves to promote recombination within heterologous domains.

Chapter 3 describes the calorimetric analysis of wild-type G148-GA3 folding and albumin binding reactions to obtain the most complete set of thermodynamic state functions for any of the native GA domains. These thermodynamic data were needed to understand the impact of subsequent recombinogenic studies on the domain. Calorimetry shows that when buffered at pH 7.0 the 46-amino acid three helix domain melts at 72°C and exhibits marginal stability (-15 kJ/mol) at 37°C. G148-GA3 unfolding is characterized by small contributions to entropy from non-hydrophobic forces and a low ΔC_p (1.1 kJ/(deg·mol)). Isothermal titration calorimetry reveals that the domain has evolved to optimally bind human serum albumin near 37°C with a binding constant of 1.4×10^7 M^{-1}. Analysis of G148-GA3 thermodynamics suggests that the domain experiences atypically small per residue changes in structural dynamics and heat capacity while transiting between folded and unfolded states.

Finally, in Chapter 4 I demonstrate the application of OR-PCR to the recombinogenic analysis of GA module polymorphisms by using the technique to shuffle seven synthetic homologs that represent much of the natural GA sequence space. Phage display is used to probe the resulting library for members that show simultaneous improvements to human and guinea pig serum albumin binding.

Analyses of selected mutants suggest that domain stabilizing mutations indiscriminately improved GA binding for both species of albumin.

Based on the experiments described in this dissertation, it is possible to conclude that recombinogenic analysis of protein family polymorphisms is a valuable technique for identifying and understanding the functional impact of natural mutations. However, this approach is possible only with the advent of a robust recombinogenic technique such as OR-PCR introduced in the following chapter.

Chapter 2: Offset Recombinant PCR Provides a Simple but Effective Method for Shuffling Compact Heterologous Domains*

Introduction

The intrinsic ability of PCR to generate recombinant products from mixed homologous template populations was recognized as early as 1988 when researchers observed the appearance of chimeric products during the amplification of two alleles with the Klenow fragment of DNA polymerase I (Saiki, Gelfand et al. 1988). Similar results were later described for amplification reactions involving *Taq* and *Vent* polymerases (Meyerhans, Vartanian et al. 1990; Yang, Wang et al. 1996; Bradley and Hillis 1997). In each of these cases recombination occurred at a relatively low frequency, making the phenomenon more of an inconvenience for researchers interested in amplifying allelic DNA than an effective mechanism for *in vitro* recombination. Despite the frequent reliance of DNA shuffling strategies on *in vitro* polymerization reactions as part of a multi-step process, researchers have yet to embrace PCR itself as an effective recombination technique.

PCR-based recombination is thought to occur when a primer is extended first on one template and then another to form a chimeric molecule with a distribution of genetic markers that differs from either of the parent templates. In order for this process to play an appreciable role in the amplification reaction, primers must

* The contents of this chapter were largely derived from a paper by the author and advisor that appeared in *Nucleic Acids Research* (Rozak and Bryan 2005).

regularly be terminated between template polymorphisms during one PCR cycle and reanneal to a template with a different assortment of genetic markers during a later cycle. A successful *in vitro* recombination technique known as the staggered extension process (StEP) uses a highly abbreviated annealing/elongation phase to generate nested primers and promote crossover events along the full length of the template (Zhao, Giver et al. 1998). However, because the process requires many cycles to fully extend a single primer, StEP fails to achieve the exponential product growth that is characteristic of PCR.

The efficient reannealing of partially extended primers to new templates is another important factor in the formation of chimeric PCR products. During each annealing phase unextended primers, partially extended primers, and full-length templates compete with one another to form DNA duplexes. PCR amplification mixes are generally saturated with excess amounts of unextended primers to efficiently prime the exponentially growing template populations. In contrast, partially extended primers, which are needed to promote recombination, are relatively few—especially in the early stages of PCR—and unable to effectively compete with unextended primers for a limited number of templates. Given these unfavorable conditions, partially extended primers are more likely to accumulate in the reaction mix than contribute to the formation of chimeric products. By adding a separate high-temperature annealing phase, which presumably favors complexes involving the elongated primers over unextended primers, Judo was able to achieve a noticeable increase in the recombination frequency for *Vent* and *Taq* polymerases respectively (Judo, Wedel et al. 1998). Significantly, Judo's modified PCR protocol used a 45 s

elongation phase as opposed to the highly abbreviated one employed by StEP. While Judo's longer extension time supported exponential growth it is unclear how it affected the distribution of crossover points within the amplicon.

Judo relied on altered polymerase cycling conditions to promote PCR recombination. However, it should be possible to achieve similar results by simply positioning the recombinant region towards one end of the amplicon. Careful consideration of PCR chemistry suggests that primers extended significantly beyond the amplicon's midpoint should be unaffected by competition from large concentrations of unextended primers. Rather, nearly extended forward and reverse primers can freely anneal to one-another without competing with unextended primers for full-length templates (**Figure 1**). Furthermore, by locating the recombinant region on one end of the amplicon, it may be possible to generate an adequate distribution of crossover events in the targeted region without resorting to highly abbreviated elongation times. In short, standard PCR could be used as an effective *in vitro* recombination technique for offset regions.

This chapter describes the use of a *lacZ* reporter system to characterize PCR-induced recombination between markers that are located at one end of the amplicon. By varying reaction conditions I was able to explore the effects of cycle number, extension time, and sequence homology on recombination frequencies near product ends. These results suggest that this strategy—referred to here as offset recombinant PCR (OR-PCR)—offers a simple but effective approach for generating recombinant libraries of compact heterologous domains.

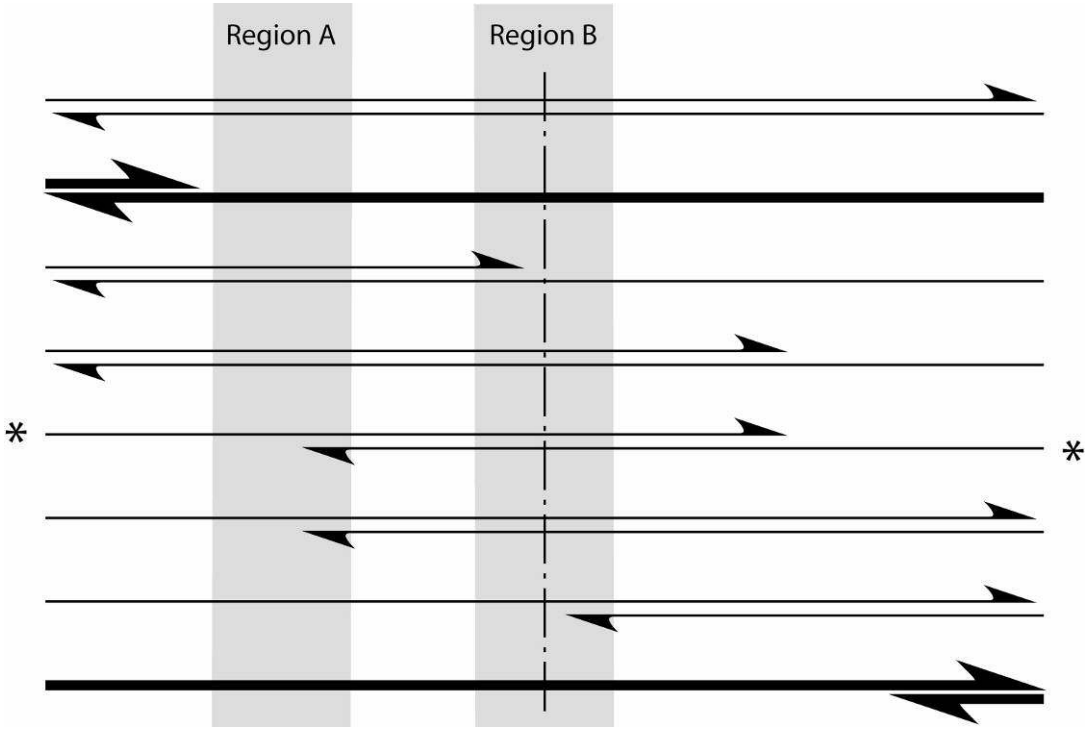


Figure 1. Anticipated Impact of DNA Duplex Formation on PCR-Mediated Recombination. Primers and templates are depicted as half-arrows pointing towards their 3' ends. The complexes favored by high primer and template concentrations are rendered in bold. Assuming incomplete elongation during earlier PCR cycles, an annealing phase can result in the formation of DNA duplexes between unextended primers, partially extended primers, and completed templates. Due to competition for full-length templates from excess amounts of unextended primers, duplexes are less likely to involve template-template pairs and templates paired with partially extended primers. However, those primers that have been sufficiently extended beyond the product's midpoint during an earlier elongation cycle (indicated by asterisks) are more likely to form duplexes with their counterparts in the reverse direction. This suggests that primers terminated between markers near the product end (Region A) are more likely to reanneal and extend to form chimeras than primers terminated at or before the product's center (Region B).

Experimental Results

The lacZ Reporter System

I first sought to create a simple reporter system in which phenotype rescue frequencies could be used to monitor recombination among *lacZ* alleles with knockout mutations. Site directed mutagenesis was used to eliminate the β-galactosidase phenotype by placing pairs of adjacent ochre stop codons at two different points within the *lacZ* open reading frame (ORF). As expected, cells transformed with the pUC19-03 (ochre mutations at β-galactosidase positions 35 and 36) and pUC19-05 (ochre mutations at positions 9 and 11) mutants failed to produce blue colonies. These pUC19 constructs are diagrammed in **Figure 2A** and collectively referred to as set A mutants.

Two variants of pUC19-03 and pUC19-05, which contain a total of 8 silent point mutations in the 82 nt region, were also created to explore PCR recombination among heterologous stretches of DNA. These variants, which were designated pUC19-06 and pUC19-07 and are collectively referred to as set B mutants, exhibit an 82% DNA sequence homology to one another within the recombinant region (**Figure 2A**).

Three primers were designed to amplify mixed populations of the pUC19 mutants. Per **Figure 2B**, when the pUC19 mutants are amplified with P1 and P2, the 82 nt recombinant region is located towards one end of a 329 bp product. However, when pUC19 populations are amplified using P1 and P3, the recombinant region is centered on a 511 bp product. These two primer combination were used in this study to compare recombination frequencies between offset and centered markers.

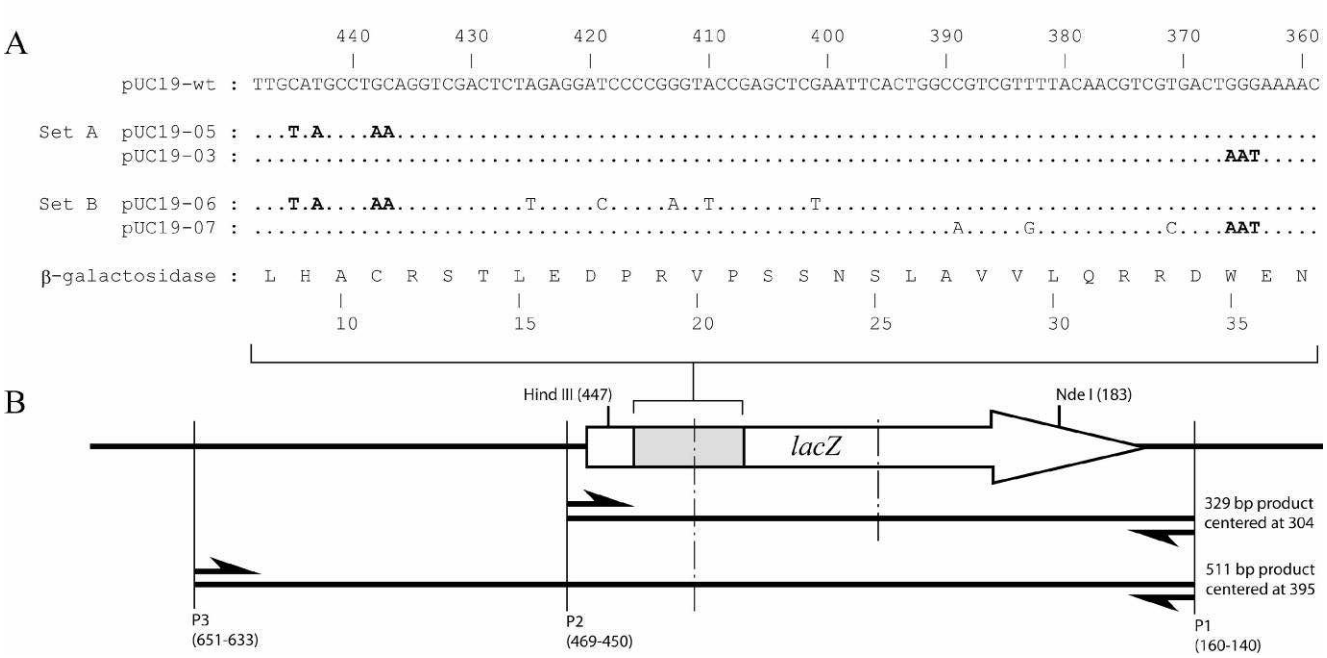


Figure 2. The *lacZ* Reporter System Used to Test OR-PCR Performance. (**A**) Point mutations for each of the pUC19 constructs used in the PCR recombination experiments. Although all nucleotide sequences in this chapter are presented in terms of the *lacZ* coding strand, they are numbered in accordance with their complements for consistency with the published pUC19 sequence. Nucleic acid mutations that lead to the creation of ochre stop codons in the *lacZ* ORF are shown in bold. All other mutations have no impact on the wild-type β-galactosidase amino acid sequence, which is aligned with the nucleotide sequences at the bottom of the frame. (**B**) Locations of primers, restriction sites, and the 82 bp recombinant region (shaded box) with respect to the *lacZ* ORF on the pUC19 vector. Dashed lines mark center-points for the 329 bp and 511 bp PCR products. Half-arrows represent the arrangements of forward and reverse primers (P1, P2, and P3) on the two amplicons.

Paired stop codons were designed into each of the pUC19 mutants to reduce the chance of phenotype recovery by random point mutations. To validate this approach pUC19-03 and pUC19-05 constructs were amplified separately using P1 and P2 primers and one-minute elongation times. The PCR products were then ligated back into a pUC19 *LacZ* $^{-}$ construct and used to transform TG-1 *E. coli* strains via heat shock. Fewer than one in 1,000 colonies transformed with pUC19-03 or pUC19-05 PCR products recovered the blue *lacZ*$^{+}$ phenotype through simultaneous point mutations in both of the paired stop codons.

Phenotype Rescue as a Function of OR-PCR Cycle

Equal mixtures of set A mutants were amplified with P1 and P2 to observe phenotype rescue via the recombination of offset markers as function of OR-PCR cycle number. Reaction mixtures and cycling conditions were selected to reproduce a typical amplification reaction. These cycling conditions included a one-minute elongation phase to promote complete extension of the 329 bp PCR products based on a previously reported *Pfu* elongation rate of 25 bases/s (Takagi, Nishioka et al. 1997). Identical amounts of PCR product, as determined by OD_{260}, were taken from the thermocycler on odd cycles and ligated into pUC19 for transformation and screening. Absorbance readings confirmed an exponential growth in PCR products. As shown in **Figure 3**, phenotype rescue was undetected in sampled colonies until the 11^{th} cycle and reached a frequency of 0.11 ± 0.02 blue colonies per sampled population after 30 cycles. The higher ratios of blue colonies generated during the second half of the cycling reaction are consistent with observations of PCR-induced

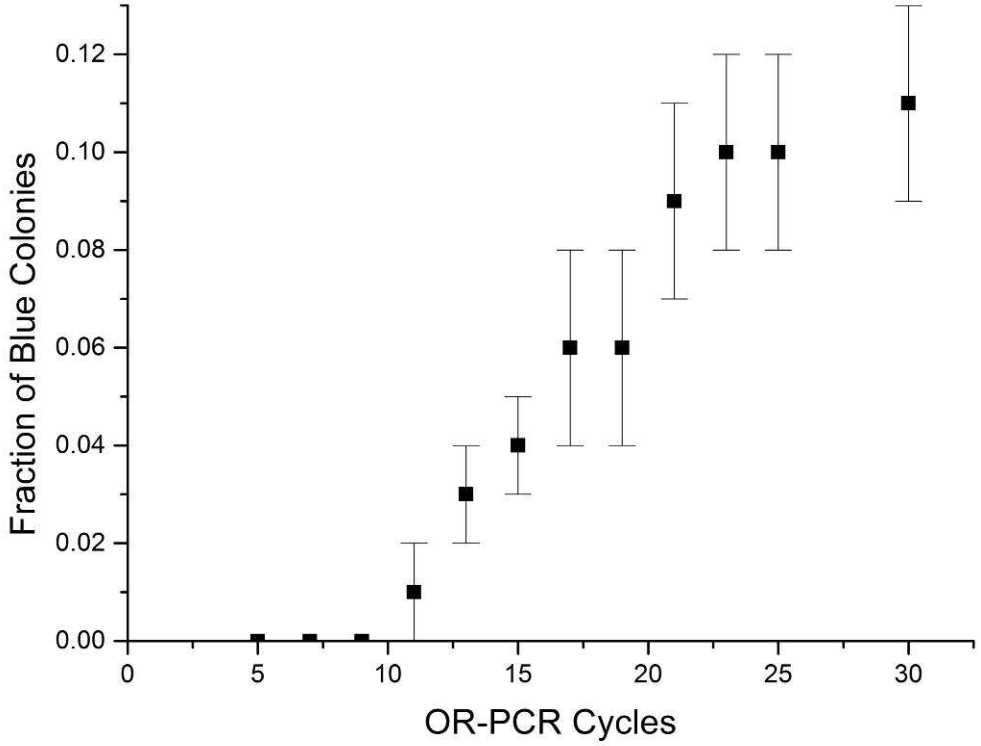


Figure 3. Effect of OR-PCR Cycle Number on Phenotype Rescue. Phenotype rescue frequencies observed after different numbers of cycles (*c*) in a typical amplification reaction (30s at 95°C; (30 s at 95°C; 30 s at 55°C, 60 s at 75°C) x *c*). Error bars represent the statistical error inherent in sampled population sizes.

recombination reported elsewhere (Meyerhans, Vartanian et al. 1990; Judo, Wedel et al. 1998).

Phenotype Rescue as a Function of Elongation Time

Amplification reactions were run with elongation times ranging from one to 120 s in order to observe the effect on OR-PCR. After 30 cycles, equal concentrations of each PCR product were ligated into pUC19 *lacZ*$^-$ vectors and transformed into *E. coli* to observe phenotype rescue. The results, which are plotted in **Figure 4**, suggest that

elongation times of less than 40 s dramatically increase the frequency of phenotype rescue to as much as 18%. Within the limits of experimental error, no significant difference was observed in phenotype rescue frequencies for elongation times between one and 40 s. This broad range of effective polymerase extension times suggests that the hyper-attenuated 5 s 55°C annealing/elongation phase employed by StEP (Zhao, Giver et al. 1998) is not necessary to promote distributed recombination events among offset markers. Unless otherwise stated, reactions described in the rest of this chapter were conducted with 15 s elongation times.

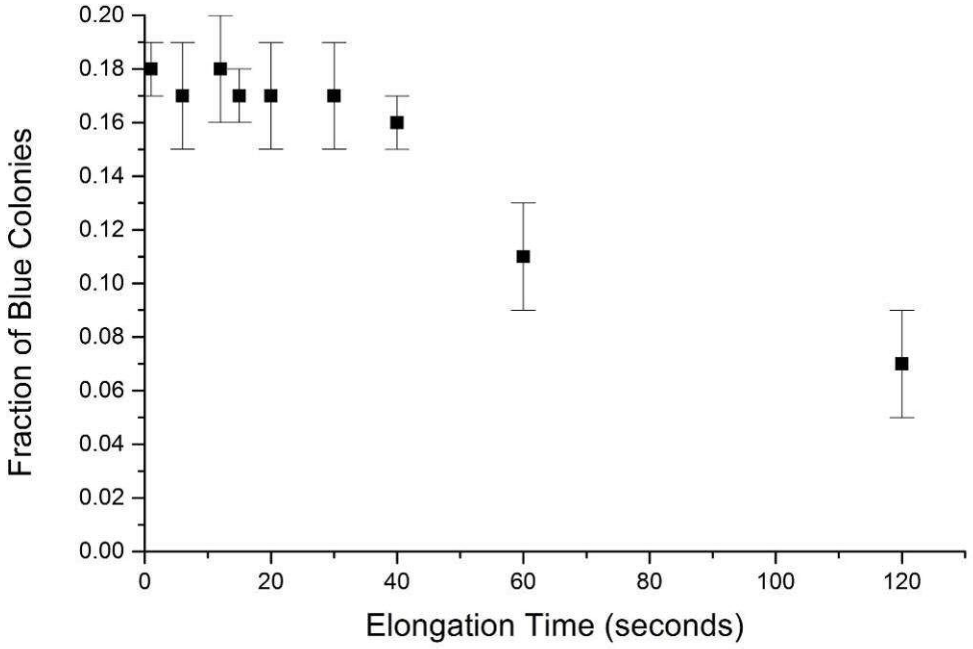


Figure 4. Effect of OR-PCR Elongation Time on Phenotype Rescue. Phenotype rescue frequencies observed for different elongation times (*t*) in a typical amplification reaction (30s at 95°C; (30 s at 95°C; 30 s at 55°C, *t* s at 75°C) x 30). Error bars represent the statistical error inherent in sampled population sizes.

Comparing Phenotype Rescue Frequencies for Centered and Offset Markers

In order to compare phenotype rescue frequencies for centered and offset markers, the amplification reaction described above was run with 15 s elongation times using P1 and P3 primers to generate a 511 bp product with the 82 nt recombinant region centered approximately 255 bp from each end (**Figure 2B**).

Because P1 and P3 are equidistant from the recombinant region, the 15 s elongation phase should be equally effective in terminating both primers within the 82 nt region to promote observable crossover events. However, primers terminated within this region will also be within 40 nt of the amplicon's center, reducing the chance they can avoid competition from unextended primers by annealing to one another. These less favorable reannealing conditions should have an observable effect on the phenotype rescue frequency.

Indeed, when PCR products were ligated into the pUC19 *lacZ*⁻ vector and expressed in *E. coli* they exhibited a phenotype rescue frequency of 0.14 ± 0.1 compared to the frequency of 0.17 ± 0.1 observed for the P1- and P2-primed templates with an offset recombinant region. Because elongation times have not changed and P1 is just as likely to be terminated within the 82 nt recombinant region as in previous experiments, the drop in phenotype rescue frequency is likely attributable to the reduced ability of center-terminated P1 and P3 primers to form extendable complexes.

The negative impact of the centered markers on primer-mediated recombination is even more pronounced when one considers that both P1 and P3 are equally likely to be terminated in the centered recombinant region. Therefore, the phenotype rescue

frequency observed for centered markers is twice that attributable to the reannealing and extension of P1 or P3 alone. This should be compared to experiments involving offset markers where elongation conditions made it highly unlikely that P2 would be terminated within the neighboring recombinant region and phenotype rescue frequencies were almost exclusively attributable to the action of P1. This analysis suggest that P1 is nearly two and a half times as effective at generating recombinants among offset markers than centered markers.

Similar results were observed when this experiment was repeated with one-minute elongation times. Under these more typical amplification conditions centered markers generated a phenotype rescue frequency of 0.11 ± 0.02, which fails to surpass frequencies generated by offset markers under identical reaction conditions, despite the advantage of both forward and reverse primers having equal chances of terminating between centered markers and contributing to recombination.

OR-PCR Recombination of Heterologous DNA

A major goal of this study was to probe the extent to which OR-PCR can promote recombination among heterologous stretches of DNA. I explored these limits by amplifying equal quantities of set B mutants (**Figure 2A**), which exhibit 82% homology within the 82 nt recombinant region, with P1 and P2 primers and 15 s elongation times. Cells transformed with set B-derived OR-PCR products produced phenotype rescue frequencies of 0.11 ± 0.01 compared to 0.17 ± 0.01 observed when set A mutants were amplified under identical conditions.

In order to study the distribution of crossover events between the closely spaced nucleotide polymorphisms found among set B constructs, plasmid DNA was purified and sequenced from sixteen colonies exhibiting the blue *lacZ* $^{+}$ phenotype. The distribution of sequence polymorphisms (**Figure 5**) suggests that crossovers occurred primarily in the half of the recombinant region closest to the P1 primer. Per these results, *Pfu* polymerase is able to elongate partially extended P1 primers that reanneal to their homolog in this region despite the presence of mismatched bases within as few as 4 nucleotides from the 3' terminal. A similar tolerance for 3' mismatches in heteroduplexed DNA has been reported elsewhere for *Taq* polymerase (Kwok, Chang et al. 1995). However, *Pfu* polymerase appeared to be incapable of extending oligonucleotides that traversed more than half of the 82 nt heterologous region, possibly due to a decrease in local annealing temperature that results from an accumulation of 3' mismatches. If this is the case, the adverse effects of 3' heteroduplex instability on annealing and elongation might be mitigated by the use of lower annealing temperatures.

A careful review of DNA chromatographs revealed that half the samples represented in **Figure 5** were derived from colonies containing two alleles of the *lacZ* gene. The mixed alleles most likely resulted from heteroduplex formation during PCR. In each of these cases, one allele represented the *lacZ*$^{+}$ gene with its dominant blue phenotype, as depicted in **Figure 5**, while the other appeared to be a non-recombinant *lacZ*$^{-}$ gene identical to that of the parental pUC19-07 construct. pUC19-06 derivatives were conspicuously absent from all of the sequenced samples.

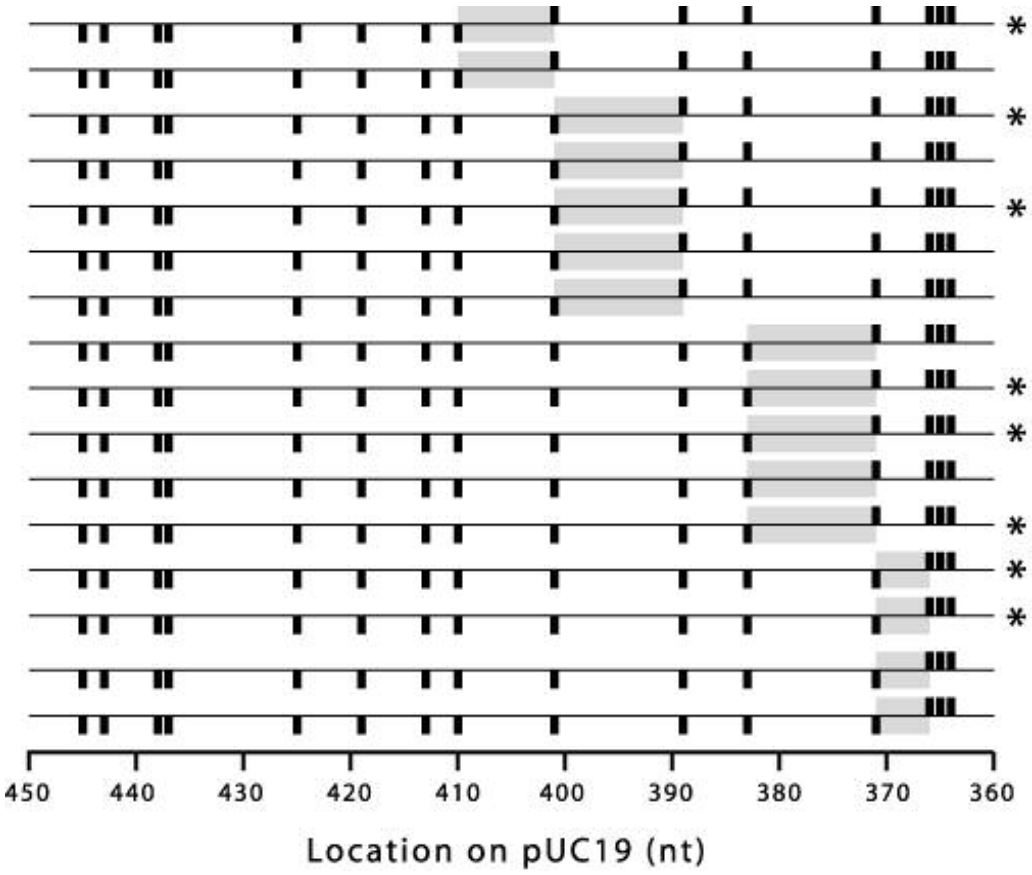


Figure 5. Recombinant *lacZ* Sequences After One Round of OR-PCR. Assortment of single nucleotide polymorphisms in 16 DNA sequences derived from blue colonies transformed with set B recombinants. pUC19-06 derived markers are indicated by upward ticks while pUC19-07 derived markers are depicted with downward ticks. Crossover events (shaded boxes) are concentrated in the half of the recombinant region closest to the P1 primer. Asterisks indicate those sequences derived from colonies that also appeared to contain the non-recombinant pUC19-07 construct.

Heteroduplex formation is to be expected from PCR amplification of mixed templates. However, the presence of mixed alleles in only half of the sequenced samples and complete absence of stop codons at β-galactosidase positions 9 and 11

was unforeseen. Random assortment suggests that nearly all of the sequenced samples should show the less common recombinant *lacZ*$^{+}$ gene paired with one of the more abundant *lacZ*$^{-}$ mutants. Furthermore, the *lacZ*$^{-}$ mutants should represent an equal distribution of the stop codon mutations found in both pUC19-06 and pUC19-07.

The low incidence of mixed alleles and complete absence of pUC19-06 derived N-terminal stop codons in the sequenced samples can be explained by the possibility that plasmids containing N-terminal stop codons are not well maintained in transformants. In this manner, the propagation of cells containing *lacZ*$^{+}$ genes paired with *lacZ*$^{-}$ genes possessing N-terminal stop codons would lead to the eventual loss of the unstable *lacZ*$^{-}$ variant—whose truncated products may serve to exhaust cell resources and significantly impede growth—leaving only *lacZ*$^{+}$ plasmids in the sampled population. Because cellular dynamics appear to alter the distribution of pUC19 mutants, observed phenotype rescue frequencies probably differ from actual OR-PCR recombination frequencies.

Estimating OR-PCR Recombination Frequencies Via Serial Amplification Reactions

Equal mixtures of set A mutants were subjected to consecutive rounds of OR-PCR in order to assess the degree to which observed phenotype rescue frequencies differ from OR-PCR recombination frequencies. I hoped that serial amplification reactions would lead to an accumulation of recombinant products and reveal an asymptotic approach to a maximum phenotype rescue frequency that could be used to estimate the recombination frequency.

Serial reactions were performed by taking a 2 μL aliquot from a completed 30-cycle OR-PCR and transferring it to a fresh reaction buffer for another round of thermocycling. A 75 ng sample of the product from each serial reaction was ligated into the pUC19 vector and transformed into *E. coli* to measure phenotype rescue frequency. The results of these experiments, which are reported as solid squares in **Figure 6**, show that more than half of the colonies exhibit the blue *lacZ*$^+$ phenotype after being transformed with DNA derived from six consecutive reactions. This is significantly above the maximum phenotype rescue frequency of 43.75% expected from random duplexes of recessive *lacZ*$^-$ and dominant *lacZ*$^+$ alleles if the latter can not exceed one quarter of a fully shuffled set A population.[*] The higher than expected phenotype rescue frequency may be due to sequence-specific biases introduced during DNA recovery or post-transformational processing in *E. coli*. The absence of N-terminal stop codons observed above suggests that some sequence-specific selection is taking place. However, further efforts to explore the mechanisms behind these phenomena would exceed the scope of this study and have little impact on the conclusions presented here.

Even in the absence of a clearly articulated mechanism for the post-recombinant fate of *lacZ* DNA it is possible to arrive at an initial estimate of the rescue frequency by fitting experimental data with a statistical model that allows for the observed bias.

[*] No more than 25% of the DNA in a fully shuffled population of pUC19-03 and pUC19-05 constructs is expected encode a functional *lacZ*$^+$ gene. Assuming the dominant *lacZ*$^+$ allele only needs to be present in one strand of the heteroduplexed DNA used to transform *E. coli*, the chance of finding a blue colony can be computed from **Equation 3** below by setting E_1 and E_2 to 0.25.

The recombination frequency for a 30-cycle amplification reaction can be expressed as the probability *p* that a single-stranded DNA product contains an assortment of genetic markers different from those found in the initial population. When *p* is small, it provides a reasonable estimate of the probability that a DNA strand will have undergone a single chimeric recombination event during the 30-cycle reaction to form

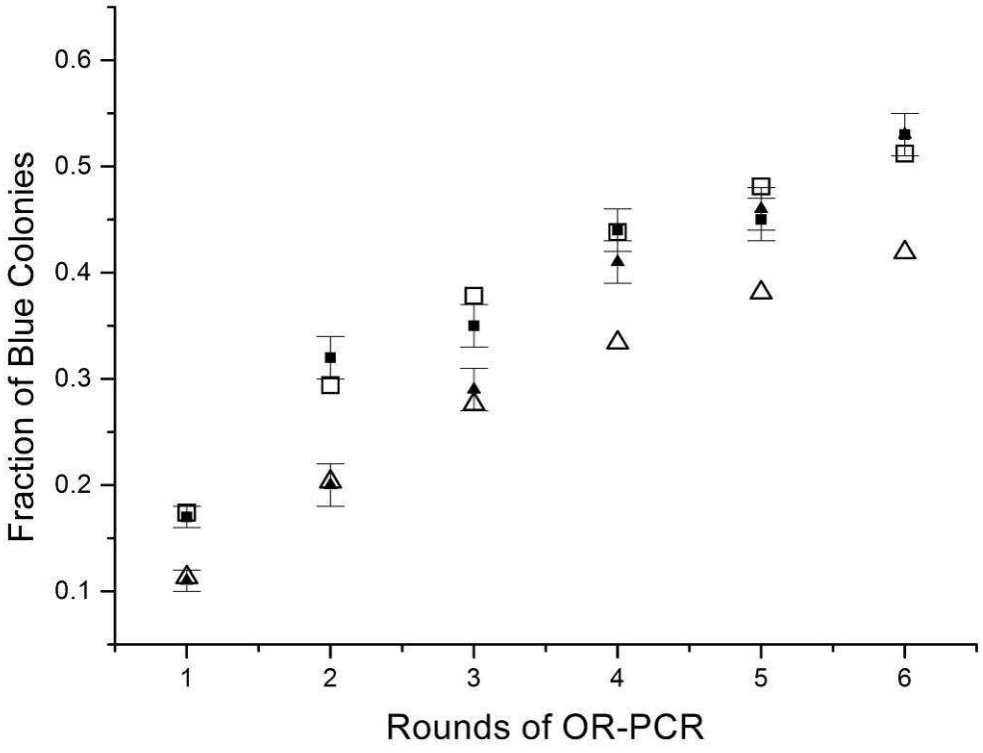


Figure 6. Phenotype Rescue During Consecutive Rounds of OR-PCR. Frequencies are derived from reactions involving set A (solid boxes) and set B (solid triangles) template mixes. Error bars express the statistical error inherent in the sample size. Phenotype rescue frequencies predicted by **Equation 6** are shown for $c = 1.36$, $p = 0.13$ (open boxes), and $p = 0.08$ (open triangles). These values were selected to produce the best fit to experimental data.

a sequence that differs from either of the molecules that contributed to its formation. Taking *p* as a good approximation of the chimeric recombination frequency for a single DNA polymer during a 30-cycle amplification reaction, one can use the binomial expansion to compute the chance a sequence undergoes *k* chimeric recombination events during *n* serial OR-PCR amplifications:

$$b(p,n,k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{1}$$

Only those sequences that have undergone an odd number of chimeric recombination events (*k = 1, 3, 5, ...*) during *n* consecutive rounds of PCR will contain an assortment of stop codons different than those found in the original population. This subset of recombinant sequences will either contain knockout mutations from both constructs or none at all. Since these two species should be relatively uniform in number, only half of the sequences that have undergone an odd number of chimeric recombination events will lack both sets of knockout mutations and represent functional *lacZ* genes. Therefore, the chance of a DNA strand encoding a functional β-galactosidase protein after *n* serial rounds of PCR can be expressed as *Z(p,n)* where:

$$Z(p,n) = \tfrac{1}{2} \sum_{k=1,3,5\ldots} b(p,n,k) \tag{2}$$

Assuming the largely homologous PCR products randomly anneal to one another at the end of the amplification reaction, it is likely that each cell will be transformed with heteroduplexed pUC19 mutants representing two distinct alleles. The dominant blue phenotype will be observed in colonies if at least one of the two pUC19 strands

from the original cell codes for a functional *lacZ* gene. If events E_1 and E_2 represent the incorporation of *lacZ*$^+$ sense and antisense strands into a pUC19 heteroduplex, the probability *B* of finding a blue colony on the plate can be expressed as

$$B = P\{E_1 \cup E_2\} = P\{E_1\} + P\{E_2\} - P\{E_1E_2\} \quad (3)$$

where *P* is the probability of a given event or set of events occurring. Assuming

$$P\{E_1\} = P\{E_2\} = Z(p,n) \quad (4)$$

this expression simplifies to

$$B(p,n) = 2Z(p,n) - Z(p,n)^2 \quad (5)$$

Finally, **Equation 5** is multiplied by the constant *c* to reflect the observed bias towards *lacZ*-encoding heteroduplexes. In the absence of further data on the cellular fate of recombinant DNA, *c* provides a reasonable approximation of the impact post-recombinant factors have on the fraction of blue colonies, provided these factors act in a manner that is largely independent of *n* and *p*.

$$B(c,p,n) = c[2Z(p,n) - Z(p,n)^2] \quad (6)$$

By setting *B(c,p,n)* equal to the experimentally determined fraction of blue colonies for *n* serial offset recombination reactions, it is possible to fit the equation to the experimental data in **Figure 6** by varying *c* and *p,* which respectively impact the asymptotic height and rate of ascent for the curve. **Figure 6** shows a reasonably tight

fit to experimental data from the serial recombination of pUC19-03 and pUC19-05 for p = 0.13 and c = 1.36. Since the error bars in **Figure 6** are based solely on statistical uncertainty inherent in the sample size, it is understandable that some values for $B(c,p,n)$ fall slightly outside of these ranges, possibly due to procedural errors, which are not reflected in the error estimates.

While p represents the fraction of the population containing a reassortment of terminal markers after one 30-cycle OR-PCR, the recombinant fraction is as high as 0.42 after 6 consecutive rounds of OR-PCR as computed by $2Z(p = 0.13, n = 6)$.

Effects of Serial Amplification Reactions on Heterologous Recombination

In order to observe the effects of serial OR-PCR on heterologous recombination I repeated the preceding experiment using the set B constructs. The results, which are plotted as solid triangles in **Figure 6**, show a distinct convergence of phenotype rescue frequencies for set A and B recombinants after only four rounds of PCR. Although one would expect phenotype rescue frequencies for each set to asymptotically approach a common maximum as the markers become evenly distributed among members of the population, the plots converge well before either show signs of leveling off at the maximum phenotype rescue rate of 0.6 predicted by **Equation 6** when c = 1.36. When I attempt to fit **Equation 6** to the data obtained from set B recombinants, experimentally derived rescue frequencies appear to undergo a clear transition during PCR rounds 3 and 4 between the data sets described by $B(c = 1.36, p = 0.08, n)$ and $B(c = 1.36, p = 0.13, n)$. One possible explanation for this phenomena lies in the growing fraction of recombinant templates and partially extended primers carried over from one round to the next. During consecutive rounds

of OR-PCR, set B-derived templates and primers increasingly represent a more homologous mix of pUC19-06 and pUC19-07 markers. This may serve to increase the average homology of heteroduplexed pairs and make it more likely that a partially extended P1 primer will form an extendable heteroduplex with other members of the population.

Plasmid DNA was sequenced from 14 blue and 11 white colonies that had resulted from the six consecutive rounds of PCR performed on set B plasmids. To avoid sequencing mixed alleles, plasmid samples obtained from the original colonies were transformed back into *E. coli* and purified from individual plated colonies prior to sequencing. As a result, all of the sequences depicted in **Figure 7** appear to have been derived from monoallelic samples.

As expected, sequences isolated entirely from blue colonies (**Figure 7A**) lacked the stop codons associated with pUC19-06 and pUC19-07 mutants and showed evidence of at least one crossover event. Two of the sequences revealed three distinct crossover events. Consistent with the limitations placed on k in **Equation 2**, even numbers of crossover events were not observed among these sequences.

Only half of the sequences isolated from white colonies showed signs of recombination (**Figure 7B**). Each of these appeared to have undergone two recombination events. Furthermore, none of the twelve sequences had a stop codon at the position 11 of the *lacZ* gene and only two had stop codons at position 9. The low incidence of N-terminal stop codons among white colonies reinforces the earlier hypothesis that pUC19 mutants with the N-terminal stop codons are poorly retained in the bacteria.

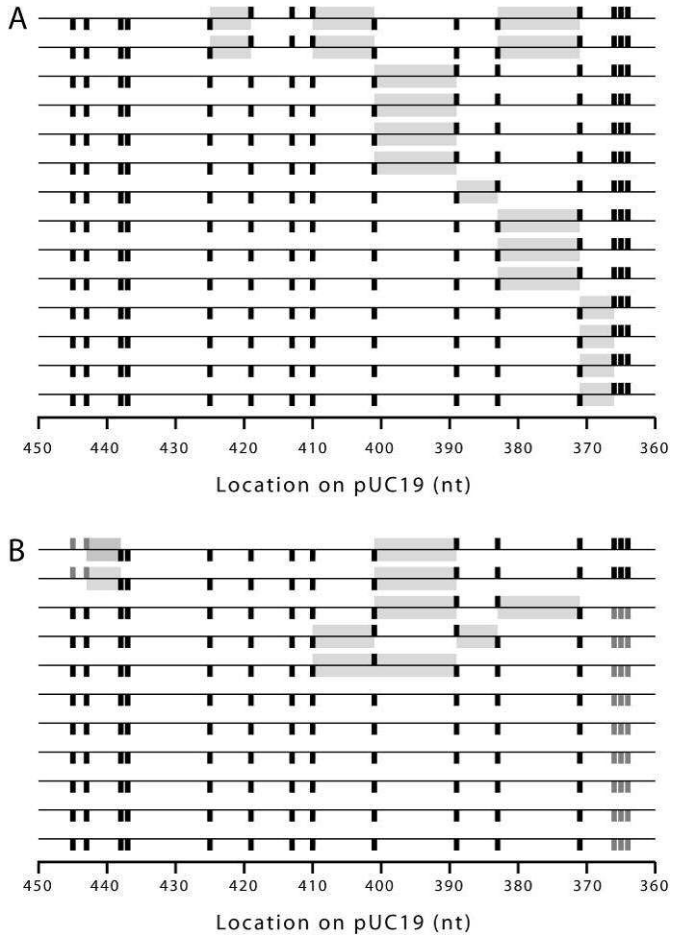


Figure 7. Recombinant Sequences Obtained From Six Rounds of OR-PCR. Assortments of single nucleotide polymorphisms in set B recombinants derived from blue (**A**) and white (**B**) colonies after six consecutive rounds of OR-PCR. Purified plasmid samples were retransformed into *E. coli* before being prepared for sequencing in order to avoid multiple alleles in a single sample. pUC19-06 derived markers are indicated by upward ticks while pUC19-07 derived markers are depicted with downward ticks. Nucleotide mutations that lead to the creation of ochre stop codons and loss of the *lacZ* phenotype are represented with gray ticks. The locations of crossover events are marked by shaded boxes.

Even after six rounds of PCR, crossover events continue to show a distinct preference for the end of the recombinant region closest to P1. However, crossover events do occur further into the heterologous region and are found as far as the 4 bp region between the two *lacZ* N-terminal stop codons. As a testimony to the *Pfu* fidelity, none of the sequences exhibited point or frame shift mutations within the recombinant region, even after six consecutive rounds of OR-PCR.

Discussion

These results indicate that the location of markers on an amplification product, polymerase elongation time, and number of PCR cycles have a discernable impact on recombination frequencies in amplifications reactions. When each of these conditions were optimized for the *lacZ* assay, recombination frequencies obtained for OR-PCR were competitive with those achieved by more complex cycling reactions (Judo, Wedel et al. 1998; Zhao, Giver et al. 1998; Ninkovic, Dietrich et al. 2001), especially when one considers that the 82 nt recombinant region studied in this chapter is only a fraction of the size of the recombinant regions studied elsewhere (**Table 1**). Using a standard PCR protocol, the offset recombination strategy explored here represents a simple but effective technique for generating high recombination rates among DNA homologs.

In contrast to other studies reported in **Table 1**, this chapter probed the limits of recombination within a compact heterologous region. DNA sequence data reveals that the optimized OR-PCR described in this chapter is capable of generating cross-over events among closely-spaced nucleotide polymorphisms found in the 40 nt stretch of

Table 1. Recombination Frequencies for Several PCR-Based Techniques.

Strategy	Size of Recombinant Region (nt) [a]	Homology Within Region [a]	Chimeric Recombination Frequency [b]	Polymerase	Thermo-cycles	Special Cycling Conditions
OR-PCR	82	91%	0.42	*Pfu*	30 x 6	Serial reactions with optimized elongation time.
	82	82%	0.42	*Pfu*	30 x 6	Serial reactions with optimized elongation time.
	82	91%	0.13	*Pfu*	30	Optimized elongation time.
	82	82%	0.08	*Pfu*	30	Optimized elongation time.
	82	91%	0.08	*Pfu*	30	None.
Centered PCR	287	99%	0.21	*Taq*	30	High-temperature annealing phase (Judo, Wedel et al. 1998).
	287	99%	0.19	*Vent*	25	High-temperature annealing phase (Judo, Wedel et al. 1998).
	287	99%	0.14	*Taq*	25	High-temperature annealing phase (Judo, Wedel et al. 1998).
	287	99%	0.07	*Vent*	25	None (Judo, Wedel et al. 1998).
	287	99%	0.01	*Taq*	25	None (Judo, Wedel et al. 1998).
StEP	113	96%	0.39	*Taq*	80	Highly abbreviated elongation phase (Zhao, Giver et al. 1998).
	260	95%	0.18	*Vent*	95	Highly abbreviated elongation phase (Ninkovic, Dietrich et al. 2001).

[a] This region is inclusive of the two observable genetic markers at either end.

[b] The chance that a product will contain an assortment of genetic markers at the ends of the recombinant region different from those found in the original population.

DNA closest to the P1 primer. However, cross-over events drop off precipitously after this, possibly due to an accumulation of point mutations at the primer's 3' end and a corresponding drop in the local melting temperature of the DNA heteroduplex. This result is mitigated by serial passage of the recombinant library through multiple OR-PCR amplifications. Given the low incidence of point mutations and other signs

of template degradation after six consecutive rounds of amplification, serial OR-PCR may be an effective means of enhancing recombination among heterologous alleles.

Perhaps the most intriguing observation to come out of this study is the apparent shift in recombination frequencies that resulted from serial amplifications of the heterologous pUC19-06 and pUC19-07 constructs. Sequence data suggests that this phenomena may result from modest increases in the overall homogeneity of the template populations and highlights an underlying advantage of PCR-based recombination over competing techniques. The chain reaction phenomena permits the rapid accumulation of recombinant templates in the mix, which provide a diverse supply of substrates to support the binding and extension heterologous primers. In fact, under the exponential amplification conditions, even the primer pool grows more diverse as primers are partially elongated on recombinant templates generated during earlier cycles. The net result is a homogenized population of heterologues, which have a better chance of forming extendable duplexes with one another from one cycle to the next.

Materials and Methods

pUC19 Mutants

The QuickChange™ Site-Directed Mutagenesis Kit (Stratagene, La Jolla, CA) was used to insert ochre codons and silent point mutations into the N-terminal region of the *lacZ* ORF on the pUC19 plasmid (GenBank Accession Number L09137 X02514). Where necessary, we followed the protocol outlined by Wang and Malcolm (Wang and Malcolm 1999) for QuickChange™ mutagenesis reactions involving

primers that exceeded the 40 nt limit recommended by Stratagene. XL-10 Gold Super Competent Cells (Stratagene, La Jolla, CA) were transformed with pUC19 mutants and spread along with 1.2 mg X-Gal and 5 μmol IPTG on LB agar plates containing 100 μg/mL ampicillin to confirm the absence of blue colonies containing *lacZ*$^+$ genes. pUC19 mutagenesis was also confirmed by DNA sequencing.

Polymerase Chain Reaction

All PCR recombination experiments were performed under the same general reaction conditions with variations noted in the Results section. Reaction mixes consisted of 2.5 units cloned *Pfu* polymerase (Stratagene, La Jolla, CA), 200 μM each dNTP, 0.5 μmol each primer, and a 100 ng equal mix of pUC19 mutants in 50 μL of the recommended reaction buffer. Each of the amplification reactions used the P1 primer combined with either P2 or P3 to create 329 bp or 511 bp products respectively (P1: 5'-TAA CTA TGC GGC ATC AGA GC-3'; P2: 5'-GAC CAT GAT TAC GCC AAG C-3'; P3: 5'-GCG TTG GCC GAT TCA TTA-3'). Thermocycling began with 30 s at 95°C followed by 30 cycles of 30 s at 95°C, 30 s at 55°C, and 1 min at 75°C. PCR products were purified using the QIAquick® PCR Purification Kit (Qiagen, Valencia, CA) and concentrations were determined via UV absorption at 260 nm.

Transformation and Screening

Escherichia coli transformation and screening began with the restriction digest of recovered PCR products and subsequent ligation back into the pUC19 vector. 75 ng of the purified PCR product was cut with NdeI and HindIII before being repurified

with the QIAquick® PCR Kit or QIAquick® Gel Extraction Kit (Qiagen, Valencia, CA). When the pUC19 mutants had been amplified using P1 and P2 the QIAquick® PCR Kit was used to efficiently remove the short terminal fragments cleaved from the PCR product during the digest reaction. However, when pUC19 mutants had been amplified using the P1 and P3 primers, an agarose gel was used to isolate the 264 bp recombinant restriction fragment for ligation into pUC19. In this case, the corresponding band was excised from the gel and cleaned up with the QIAquick® Gel Extraction Kit.

A 50 ng sample of a pUC19 *lacZ*′ mutant was cut with NdeI and HindIII and treated with CIP prior to purification with the QIAquick® PCR Kit. A pUC19 *lacZ*′ mutant was used as a cloning vector to avoid the possibility that small amounts of undigested plasmids could contribute to an inflated estimate of the phenotype rescue frequency. Digested plasmids were viewed on agarose gel and transformed into *E. coli* to confirm that the large majority of vectors were being digested and were therefore unlikely to significantly contribute to the numbers of white colonies observed in phenotype rescue titers. Digested PCR products and pUC19 vectors were mixed and incubated for 2 h at 25°C with T4 DNA ligase before a 1 μL aliquot was used to transform 20 μL XL-10 Gold Super Competent Cells with a 30 s heat shock at 42°C. After one hour at 37°C, the cells were spread along with X-Gal and IPTG on LB plates containing ampicillin and grown overnight at 37°C to observe the fraction of blue colonies on the plate—described throughout this paper as the phenotype rescue frequency. A minimum of 1,000 colonies were sampled for each data point. Experimental errors reported throughout this paper represent the statistical

uncertainty inherent in the size of the sampled population. These errors, which represent a 95.45% statistical confidence in the fraction of blue colonies (*b*) derived from a sampled population of *n* plated colonies, were computed using the equation:

$$\sigma = \pm 2\sqrt{\frac{b(1-b)}{n}} \tag{7}$$

DNA Sequencing

DNA samples were prepared for sequencing by growing selected colonies overnight at 37°C in LB with 100 μg/mL ampicillin. Plasmid DNA was extracted from the cell cultures using the Wizard® Plus SV Minipreps (Promega Corporation, Madison, WI). The P2 primer was used to amplify target DNA using Perkin-Elmer/Applied Biosystem's AmpliTaq-FS DNA polymerase and Big Dye terminators with dITP. Dye-terminated products were then run on an Applied Biosystems model 3100 DNA sequencer to produce sequence chromatographs.

Chapter 3: The Third Albumin Binding Domain of Streptococcal Protein G Exhibits Atypical Thermodynamics of Folding*

Introduction

Many gram-positive bacterial pathogens display surface receptors that bind common host proteins to support infection (Navarre and Schneewind 1999; Ingham, Brew et al. 2004). As one of the more common plasma proteins, serum albumin is bound to the surface of human group C and G streptococci (Myhre and Kronvall 1980) and some strains of *F. magna* (Myhre 1984), presumably in support of bacterial pathogenesis. *In vitro* studies have shown that albumin-binding bacterial strains exhibit increased growth rates in the presence of human serum albumin compared to those which were grown in the absence the ligand (de Chateau, Holst et al. 1996)—possibly benefiting from access to nutrients bound by the albumin. Furthermore, although only a fraction of *F. magna* strains bind albumin, the phenotype is predominantly associated with those isolated from deep wounds (de Chateau and Bjorck 1994).

Albumin binding has been localized to three N-terminal domains of streptococcal protein G (Akerstrom, Nielsen et al. 1987) and a two domains in the *F. magna* protein, PAB (de Chateau, Holst et al. 1996). The high degree of homology between the protein G and PAB albumin binding domains lead to the first

* The contents of this chapter were largely derived from a paper by the author and his colleagues, which has been accepted for publication in *Biochimica et Biophysica Acta*.

documented case of module shuffling in prokaryotes (de Chateau and Bjorck 1994). As a phenomenon that was until recently exclusively associated with eukaryotic exons, shuffled modules are identifiable as distinct functional and structural units of at least 25 amino acids that display a high degree of homology amidst a heterogeneous background. The prokaryotic protein G-related albumin binding (GA) module is a three-helix domain that spans about 46 amino acids and exhibits high interspecies homology against a heterogeneous protein scaffolding. As many as 16 GA modules have been identified in six proteins and four bacterial species (Johansson, de Chateau et al. 1995).

Cloned from the opportunistic streptococcal bacteria strain G148, the GA module G148-GA3 exhibits a broader range of affinities for non-primate albumins than ALB8-GA, which was isolated from pathogenic strains of the human commensal bacteria *F. magna* (Johansson, Frick et al. 2002). Johansson suggested the more dynamic G148-GA3 backbone, as observed by comparing nuclear magnetic resonance (NMR) hydrogen-deuterium (H-D) exchange data for both domains, may be responsible for the relaxed species specificity of the protein G albumin binding domain (Johansson, Nilsson et al. 2002).

Although data exists on the absolute and relative albumin binding affinities of protein G and several GA modules (Sjobring, Bjorck et al. 1991; Falkenberg, Bjorck et al. 1992; Johansson, Frick et al. 2002; Linhult, Binz et al. 2002) the full set of thermodynamic state functions for folding and albumin binding have not been defined for any members of this medically significant bacterial module. In this chapter I use differential scanning calorimetry and isothermal titration calorimetry to study the

thermodynamics of folding and human and guinea pig serum albumin binding for a histidine tagged G148-GA3 domain referred to as A002HC.

Experimental Results

A002HC Protein Design

The A002HC protein construct used in this study consists of a 46 amino acid albumin binding domain G148-GA3 flanked by a total of 16 additional amino acids on either end. The N-terminal flanking sequence contains a methionine and seven amino acids from the cloning artifact described elsewhere (Kraulis, Jonasson et al. 1996; Johansson, Frick et al. 2002). We included the cloning artifact to remain consistent with the version of the GA module used for NMR structural studies (PDB # 1GJT). The C-terminal flanking sequence consists of a two amino acid linker and six histidines to permit affinity purification on a nickel column. The complete 62 amino acid A002HC sequence reads: MEAVDANSLA EAKVLANREL DKYGVSDYYK NLINNAKTVE GVKALIDEIL AALPTEHHHH HH.

Differential Scanning Calorimetry

DSC was used to measure the heat capacity of A002HC as a function of temperature. When the temperature-dependent heat capacity of the buffer is subtracted from that of the protein in buffer, one is left with the total heat capacity contribution of the protein. Both calorimetric and van't Hoff enthalpies for unfolding can be computed from this data. A protein's calorimetric enthalpy is equal to the excess heat generated by the protein during the transition from folded to unfolded

states (Privalov and Potekhin 1986). Van't Hoff enthalpy assumes a two-state unfolding reaction and is obtained by computing the equilibrium constant from transition data and solving the van't Hoff equation for ΔH. If the calorimetric and van't Hoff enthalpies are in agreement the unfolding reaction represents a two-state process. In order to accurately determine the enthalpy of unfolding for A002HC we measured the temperature dependent heat capacity for A002HC and subtracted the buffer contributions. The Exam computer program developed by Schwarz and Kirchhoff (Schwarz and Kirchhoff 1988) was used to fit the baseline-subtracted data to a two-state model for which the calorimetric and van't Hoff enthalpies were required to be equal. These measurements and computations were repeated for A002HC under a range of buffer conditions as reported in **Table 2**. As appropriate,

glycine or acetate buffers were used to cancel the heat of ionization produced by buried carboxylate groups as the protein unfolded at extremes of pH. Analysis using a two-state thermodynamic model produced tight fits to the calorimetric data and consistent results across a range of buffers. **Figure 8** provides an example of the Exam output for one set of calorimetric data.

Table 2. Thermodynamic Data for A002HC (G148-GA3) Unfolding.

pH	No. of Independent Measurements	T_m [°C]	ΔH [kJ/mol]	ΔS [J/(deg·mol)]	ΔC_p [kJ/(deg·mol)]
11.0	1	55.3	149	454	
2.7	3	71.6 ± 0.4	168 ± 1	488	
7.0	1	72.1	170	493	1.1 ± 0.1
4.0	3	80.0 ± 0.5	177 ± 2	501	

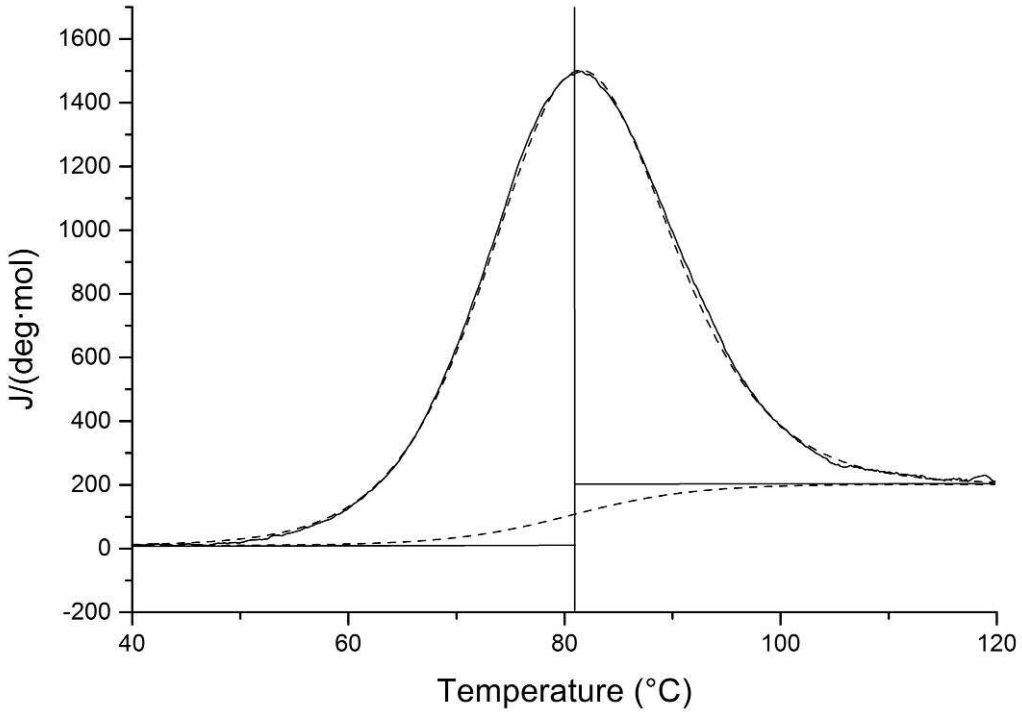


Figure 8. Analysis of A002HC (G148-GA3) Unfolding at pH 4.0. After subtracting out buffer contributions, DSC data (solid line) was fit using a two-state model (dotted lines for fit and baseline). Values for ΔC_p obtained from the baseline displacement in this and other exemplary calorimetric runs (data not shown) are within 0.1 kJ/(deg·mol) of the ΔC_p value derived by measuring ΔH over a range of buffer-induced melting temperatures (see **Figure 10A**).

Two-State Model for A002HC Unfolding

CD measurements were used to support the assumption that the A002HC reaction conforms to the two-state model by demonstrating that the protein does not adopt an intermediate conformation. CD spectra obtained from 42 μM A002HC in 100 mM $KHPO_4$ pH 7.0 as it was heated from 25°C to 100°C show the presence of a single isodichroic point at 204 nm which is indicative of a two-state reaction (**Figure 9A**).

CD spectra obtained near room temperature exhibited alpha-helical profiles consistent with similar data obtained elsewhere (Kraulis, Jonasson et al. 1996; Gulich, Linhult et al. 2000) Furthermore, the midpoint of the 222 nm ellipticity curve as the protein undergoes the transition from its folded to unfolded state places the melting point at 72°C (**Figure 9B**), which is consistent with the melting point derived from DSC measurements under the same buffer conditions but at much higher protein concentrations. The fact that T_m remains unaffected by significant changes in protein

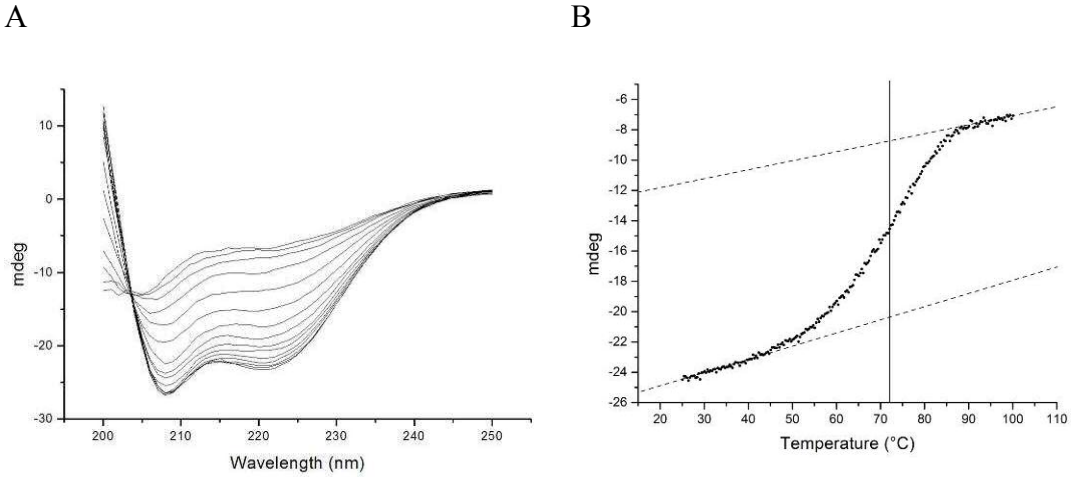


Figure 9. CD Analysis of A002HC (G148-GA3) Melting. Spectra were obtained by melting A002HC in 100 mM $KHPO_4$ pH 7.0 at 0.5 deg/min in a 1.0 cm cell. (**A**) Spectral scans of 42 µM A002HC taken every 5 minutes show the protein's secondary structure transition from alpha helix (lowest curve) to random coil (highest curve) as the temperature rises from 25°C to 95°C. The single isodichroic point at 204 nm is indicative of a two-state reaction. (**B**) A plot of the ellipticity for 840 pM A002HC at 222 nm as a function of temperature has a midpoint near 72°C, which is in close agreement with calorimetric data for 534 µM A002HC folding in the same buffer with the temperature increasing at a rate of 1.0 deg/min.

concentration supports the conclusion that A002HC melting involves a monomer to monomer transition. A002HC was also confirmed to be monomeric at 25°C by gel filtration.

Thermodynamic State Functions for A002HC Unfolding

Changes in enthalpy, entropy, and Gibbs free energy during protein unfolding can be related to the difference in heat capacities between the folded and unfolded states by the equations:

$$\Delta H = \Delta H_0 + \Delta C_p (T - T_0) \quad (8)$$

$$\Delta S = \Delta S_0 + \Delta C_p \ln(T / T_0) \quad (9)$$

$$\Delta G = \Delta H_0 - T \Delta S_0 + \Delta C_p [T - T_0 - T \ln(T / T_0)] \quad (10)$$

where ΔH_0 and ΔS_0 are values obtained at a reference temperature T_0 (Brandts 1964; Pace and Tanford 1968; Privalov and Khechinashvili 1974; Privalov 1979; Becktel and Schellman 1987). According to **Equation 8**, the change in enthalpy of a two-state reaction at equilibrium is proportionally related to temperature by ΔC_p, which is assumed to remain constant within the temperature range of this experiment. The temperature-dependent values of ΔH reported in **Table 2** were fit to a linear function in which the slope of 1.1 ± 0.1 kJ/mol is equal to ΔC_p per **Equation 8**. Experimentally derived values for ΔH and ΔC_p were used to calculate the remaining state functions

for the unfolding reaction and plot the temperature dependence of free energy in **Figure 10**.

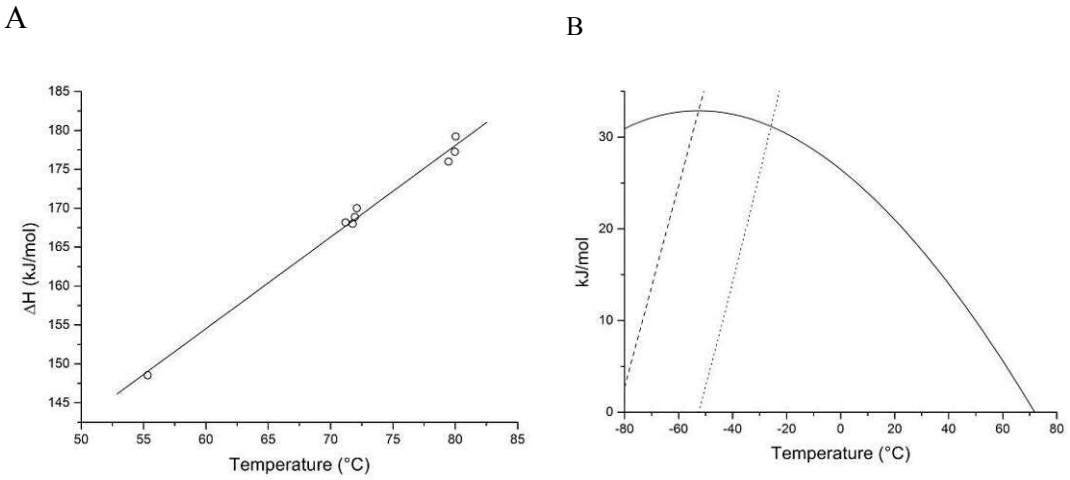


Figure 10. Thermodynamic Analysis of A002HC (G148-GA3) Unfolding. (**A**) Linear fit of A002HC enthalpies of unfolding plotted from data in **Table 2**. ΔC_p corresponds to the slope (1.1 ± 0.1 kJ/(deg·mol)) of the linear fit. (**B**) Thermodynamic profile for A002HC unfolding in 50 mM NaAc, pH 5.7. Temperature dependent energies are shown for ΔG (solid line), ΔH (dotted line), and TΔS (dashed line).

These results were verified by comparing DSC-derived values for ΔG to the free energies of transient opening (ΔG_{op}) computed from previously reported G148-GA3 H-D exchange data (Johansson, Nilsson et al. 2002). H-D exchange in unbound G148-GA3 can be described by the reaction path

$$F(H) \underset{k_f}{\overset{k_u}{\rightleftharpoons}} U(H) \xrightarrow[D_2O]{k_c} U(D) \underset{k_u}{\overset{k_f}{\rightleftharpoons}} F(D) \qquad (11)$$

where k_u, k_f, and k_c are the respective unfolding, folding, and intrinsic exchange rates, F is the folded state, and U is the unfolded state (Hvidt and Nielsen 1966). Because

the H-D exchange experiment was conducted at temperatures well below T_m ($k_f >> k_u$) with an intrinsic exchange rate significantly less than the rate of protein folding ($k_f >> k_c$), the measured exchange rate k_{ex} can be given by

$$k_{ex} = k_u k_c / k_f = K_{op} k_c \tag{12}$$

The free energy required for transient opening can then be expressed as

$$\Delta G_{op} = -RT \ln K_{op} \tag{13}$$

where K_{op} is the equilibrium constant for transient opening and ΔG_{op} is the free energy difference between locally or globally folded and unfolded states. Applying this analysis to H-D exchange data for G148-GA3 (Johansson, Nilsson et al. 2002) produces a histogram in which most amides have values for ΔG_{op} clustered near 5 or 20 kJ/mol (**Figure 11**). Amides exhibiting small and large ΔG_{op} represent those respectively involved in local and global unfolding reactions. As expected, the large free energies of transient opening, which represent global unfolding events, are comparable to the DSC-derived free energy value of 19 kJ/mol for A002HC unfolding under similar conditions.

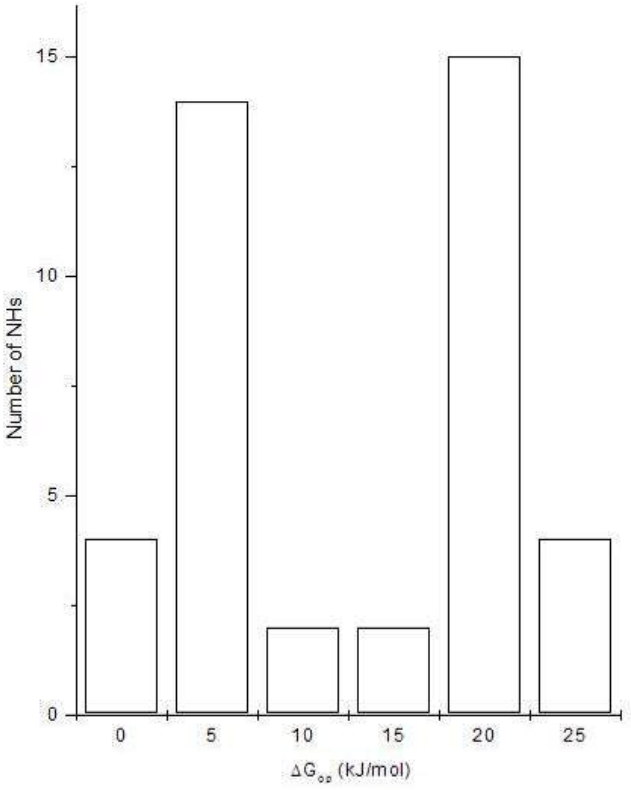
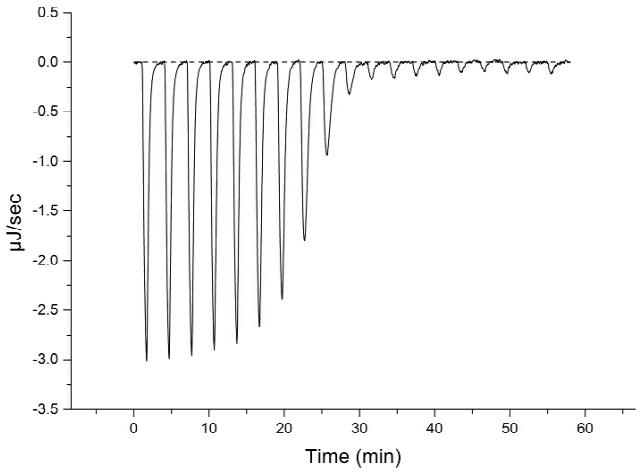


Figure 11. Distribution of ΔG_{op} Values for Folded G148-GA3. Values were derived from H-D exchange data collected by Johansson (Johansson, Nilsson et al. 2002) for unbound G148-GA3 at 27°C. Higher ΔG_{op} values, which correspond to global unfolding events, are comparable to the ΔG value for A002HC unfolding at 27°C (19 kJ/mol), which was computed from DSC data.

Albumin Binding Constants and Associated Thermodynamic State Functions

ITC was used to determine binding constants and thermodynamic state functions for A002HC interactions with HSA and GPSA. ITC measures the heat produced when small amounts of protein bind to an excess of ligand. As with DSC measurements, the enthalpy involved in each binding reaction is equal to the area under the calorimetric curve. The enthalpies of successive binding reactions can be plotted as a function of the protein-ligand molar ratios to produce a transition curve and compute the binding constant *K*. An example of the calorimetric data obtained for A002HC/HSA binding at 25°C is given in **Figure 12**. The midpoint for each of the binding reactions occurred when equal amounts of protein and ligand were present,

A



B

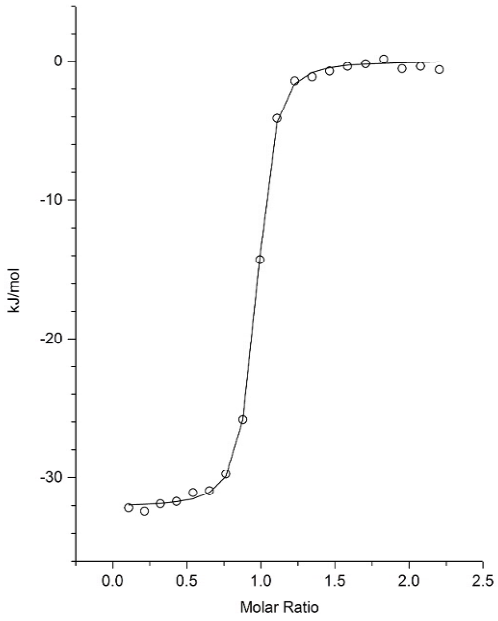


Figure 12. ITC Data for A002HC(G148-GA3)/HSA Binding. Measurements were made at 25°C in 100 mM $KHPO_4$, pH 7.0. (**A**) Instantaneous heat generated by adding 246 nM aliquots of A002HC to 25 µM HSA in three-minute intervals. (**B**) Total heat generated per mole of injectant as a function of the molar ratio of protein to ligand. The transition curve yields a binding constant of 1.2 (± 0.1) x 10^7 M^{-1}, a van't Hoff enthalpy of -31.89 ± 0.13 kJ/mol , and entropy of 28.61 J/(deg·mol). The midpoint of the transition curve occurs at equivalent concentrations of protein and ligand, indicating one-to-one stoichiometry.

Table 3. Thermodynamic Data for A002HC (G148-GA3) Binding to HSA and GPSA.

Ligand	Temp. [°C]	K [1/mol]	ΔG [kJ/mol]	ΔH [kJ/mol]	ΔS [J/(deg·mol)]	ΔC_p [kJ/(deg·mol)]
HSA	25	1.2 (± 0.1) x 10^7	-40.4	-31.89 ± 0.13	28.61	
	30	1.4 (± 0.1) x 10^7	-41.4	-40.53 ± 0.15	2.85	-1.75 ± 0.08
	35	1.4 (± 0.1) x 10^7	-42.1	-49.41 ± 0.18	-23.41	
GPSA	25	0.9 (± 0.8) x 10^7	-39.8	-3.69 ± 0.17	120.75	N/A

indicating one-to-one stoichiometry. I observed A002HC as it bound to HSA at 25°C, 30°C, and 35°C and determined ΔC_p from the slope of the linear fit to the binding enthalpies plotted against temperature. Experimental and computational results for ITC measurements are reported in **Table 3** and describe a free energy profile with its minimum near the physiologically significant 37°C (**Figure 13**). Differences in

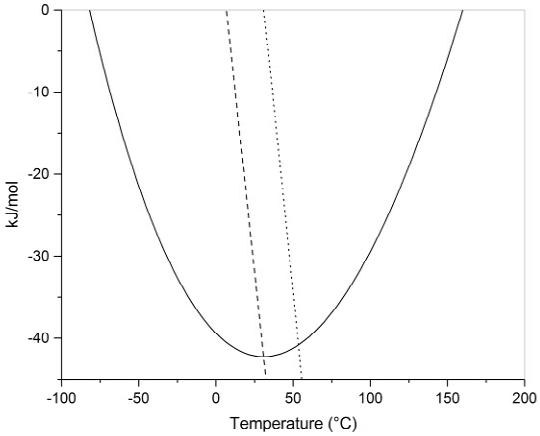


Figure 13. Free Energy profile for A002HC (G148-GA3)/HSA Binding. Data is presented for 100 mM KHPO4, pH 7.0. Temperature dependent energies are shown for ΔG (solid line), ΔH (dashed line), and TΔS (dotted line).

experimental conditions may account for the fact that the binding constant obtained for HSA in solution is one twentieth that observed during surface plasmon resonance experiments involving a G148-GA3 fusion protein and immobilized ligand (Linhult, Binz et al.).

ITC measurements were also obtained for GPSA binding at 25°C. Although the low enthalpy associated with GPSA binding made it difficult to accurately determine a binding constant, the values reported in **Table 3** are consistent, within the range of experimental error, with the relative affinities of G148-GA3 for HSA and GPSA obtained from competitive binding experiments (Johansson, Frick et al. 2002).

Discussion

Analysis of A002HC thermodynamics suggests that the three helix domain exhibits unusually low changes in protein mobility and heat capacity on a per residue basis during the transition between folded and unfolded states. According to Baldwin (Baldwin 1986), the hydrophobic (hϕ) contribution to the Gibbs free energy can be approximated by the expression

$$\Delta G_{h\phi} = \Delta C_p(T-T_h) + \Delta C_p\, T \ln (T_s /T) \qquad (14)$$

where T_h = 22°C and T_s = 113°C are the experimentally determined temperatures at which $\Delta H_{h\phi}$ and $\Delta S_{h\phi}$ are respectively equal to zero (Gill, Nichols et al. 1976; Sturtevant 1977).

Baldwin's liquid hydrocarbon model provides a convenient mechanism for comparing the impact of non-hydrophobic or residual (r) interactions on protein

folding and assessing their contributions to ΔH and ΔS. Since $\Delta H_{h\phi}$ is negligible at 22°C and the enthalpy is dominated by residual forces near this temperature, ΔH/(mol·residue) at this temperature provides a measure of relative contributions by non-hydrophobic forces to enthalpy. Likewise, $\Delta S_{h\phi}$ disappears at 113°C making this temperature ideal for assessing the relative contribution of residual forces to entropy.

As shown in **Table 4**, studies suggest that per residue values for ΔS_r and ΔC_p remain remarkably constant for globular proteins ranging in size from 7-25 kDa. (Privalov and Gill 1988; Alexander, Fahnestock et al. 1992) When normalized values were computed for A002HC's 46 amino acid structural core, they fell significantly below the mean.

In the absence of hydrophobic contributions to entropy, which result from differences in the ordering of solvent molecules around the folded and unfolded protein, ΔS_r reflects the amount of order achieved by the folded protein relative to its unfolded state. A002HC's low ΔS_r suggests that G148-GA3 is subject to fewer motion constraints as a result of folding than most globular proteins.

A002HC also exhibits a ΔC_p/residue well below that of the other proteins. Since C_p is generally proportional to the non-polar surface area of the solute, changes in C_p associated with protein folding are taken as an indication of how well the folded protein buries hydrophobic residues in its core. Indeed, when the hydrophobic surface areas for folded and unfolded proteins are computed, A002HC appears to bury a smaller fraction of hydrophobic residues (0.46) than the other globular proteins considered in **Table 4** (0.68 ± 0.06) and its homologue ALB8-GA (0.57).

Table 4. Thermodynamic Parameters of Globular Proteins. [a]

Protein	Molecular Weight [g/mol]	Fraction of Hydrophobic Surface Area Buried on Folding [b]	Residual Forces		Hydrophobic Forces
			ΔH_r (25°C) [kJ/(mol•res)]	ΔS_r (110°C) [J/(deg·mol·res)]	ΔC_p [J/(deg·mol·res)]
A002HC	6919	0.46	2.5	13.0	24
protein G B1	7179	0.55	1.4	16.1	53
parvalbumin	11500	0.68	1.4	16.8	46
cytochrome c	12400	N/A	0.65	17.8	76
ribonuculese A	13600	0.63	2.4	17.8	44
hen lysozyme	14300	0.70	2.0	17.6	52
staphylococcal nuclease	16800	0.68	0.85	17.5	61
myoglobin	17900	0.69	0.04	17.9	75
papain	23400	0.75	0.93	17.0	60
β-trypsin	23800	0.74	1.3	17.9	58
α-chymotrypsin	25200	0.74	1.1	18.0	58
mean ± std. dev.		0.68 ± 0.06	1.3 ± 0.7	17.0 ± 1.5	55 ± 14

[a] Data for proteins other than A002HC was obtained from Privalov (Privalov and Gill 1988) and Alexander (Alexander, Fahnestock et al. 1992)

[b] Fractions were determined by comparing the hydrophobic surface areas computed for random chains (Karplus 1997) and folded proteins (Fraczkiewicz and Braun 1998).

Materials and Methods

A002HC Assembly

An ORF encoding the 62 amino acid A002HC protein was constructed using two sequential polymerase chain reactions to assemble four overlapping oligonucleotides. The A002HC ORF was inserted into the pG5 vector as described by Alexander (Alexander, Fahnestock et al. 1992) via the NdeI and BamHI restriction sites, cloned in XL-10 Gold® Ultracompetent Cells (Stratagene, La Jolla, CA), and extracted with Wizard® Plus SV Minipreps (Promega Corporation, Madison, WI). Proper assembly and insertion of the A002HC ORF into pG5 was confirmed by DNA sequencing.

A002HC Expression and Purification

Rosetta™(DE3) Competent Cells (Novogen, Madison, WI) were transformed with the pG5/A002HC construct and grown to late log phase in 2 L phosphate buffered broth (10 g/L tryptone, 5 g/L yeast extract, 10 g/L NaCl, 100 mM $KHPO_4$ pH 7.0) with 100 μg/mL ampicillin. A002HC expression was induced by adding 1 mM IPTG and incubating four hours at 37°C. Cells were pelleted, resuspended in 100 mL 100 mM $KHPO_4$ pH 7.0 and sonicated before being centrifuged at 12,000 xg for 30 min. A column packed with 5 mL Ni-NTA Agarose (Qiagen, Valencia, CA) was used to purify A002HC from the supernatant. Purity of the A002HC protein was confirmed via SDS-PAGE and mass spectroscopy using the Voyager-DE™ BioSpectometry™ Workstation (PerSeptive Biosystems, Framingham, MA).

Extinction Coefficients

Since A002HC lacks tryptophan, it has a UV peak at 278 nm. The Edelhoch method, as described by Pace (Pace, Vajdos et al. 1995), was used to determine the A002HC extinction coefficients at 278 nm under a range of buffer conditions. Twice the absorbance at 331 nm was consistently subtracted from that obtained at 278 nm to account for the effects of light-scattering. This set of experimentally determined extinction coefficients was used to determine A002HC concentrations throughout the study.

Albumin Preparations

Dried HSA and GPSA samples were obtained from Sigma (St. Louis, MO). All albumin samples were dialyzed along side A002HC samples prior to ITC experiments in order to ensure ligands and proteins were suspended in the identical 100 mM $KHPO_4$ pH 7.0 buffers. As with A002HC, final albumin concentrations were determined using UV spectroscopy.

Circular Dichroism

CD experiments were performed on a J-720 Spectropolarimeter (Jasco Spectroscopic Co., LTD., Tokyo, Japan). The melting temperature for 42 μM A002HC in 100 mM $KHPO_4$ pH 7.0 was determined by measuring the ellipticity of the sample at 222 nm as it was heated in a 1.0 cm cell from 25°C to 70°C at 0.5 degrees per minute. The isodichroic point was observed for 840 pM A002HC in 100 mM $KHPO_4$ pH 7.0 by measuring the ellipticity of the sample in a 0.1 cm cell from

250 nm to 200 nm as the cell was heated from 25°C to 100°C at 0.5 degrees per minute.

Differential Scanning Calorimetry

DSC measurements were taken on a VP-DSC Micro Calorimeter (MicroCal Incorporated, Northampton, MA). During each run, the cell temperatures were increased from 15°C to 125°C at a rate of one degree per minute. Three buffers were used to observe unfolding at different pHs: 50 mM glycine at pH 2.7 and 11.0; 50 mM NaOAc at pH 4.0; 100 mM $KHPO_4$ at pH 7.0. Scans were run with the buffer in both the sample and reference cells. Then a quantity of A002HC, which had been dialyzed into the same buffer, was introduced into the sample cell and at least two more scans were performed. A002HC concentrations ranged from 50-350 μM as determined by UV spectra. Buffer effects were canceled by subtracting the melting curves of known protein concentrations to determine the heat contribution of their molar difference.

Isothermal Titration Calorimetry

ITC measurements were performed on the MicroCal VP-ITC Micro Calorimeter. For each experiment A002HC and the albumin were dialyzed side-by-side into 100 mM $KHPO_4$ pH 7.0 to ensure identical buffer conditions. Each run involved nineteen 15 μL injections of approximately 250 μM A002HC into a sample cell containing around 25 μM HSA or GPSA. Injections lasted 30 s each and were spaced 180 s apart. Precise A002HC and albumin concentrations were determined by UV spectra.

Jacket temperatures of 25°C, 30°C, and 35°C were used with HSA. GPSA measurements were taken only at 25°C.

Chapter 4: Using OR-PCR to Identify Functional Determinants in a Family of Albumin Binding Domains*

Introduction

In vitro experiments suggest that the albumin binding domain may support bacterial growth by scavenging albumin-bound nutrients from the blood (de Chateau, Holst et al. 1996). Variations in the abilities of these domains to bind albumin from different species (Johansson, Frick et al. 2002) may help to define the host ranges for certain bacterial pathogens. Native or engineered versions of the module could be used to support affinity-purification of albumin and other fusion proteins (Hammarberg, Nygren et al. 1989), or increase vaccine serum stability (Nygren, Flodby et al. 1991; Makrides, Nygren et al. 1996) and immunogenicity (Sjolander, Nygren et al. 1997; Libon, Corvaia et al. 1999). However, the rich array of albumin binding domains also offers opportunities for structural biologists wishing to exploit the well defined sequence space to study the impact of select polymorphisms on protein function and structure.

The functional and structural diversity exhibited by members of the module is evident in the *F. magna* ALB8-GA and streptococcal G148-GA3 albumin binding domains, which display variable affinities for different species of albumins and significantly different backbone dynamics. Researchers have observed that ALB8-GA demonstrates a distinct preference for primate serum albumins compared to the

* The contents of this chapter were largely derived from a paper that the author and colleagues have submitted for publication in *Biochemistry*.

broader range of affinities for albumin species exhibited by G148-GA3 (Johansson, Frick et al. 2002). Furthermore, comparative hydrogen-deuterium exchange data reveals that G148-GA3 maintains a more dynamic backbone than ALB8-GA, a feature that the researchers suggest may be associated with the former's ability to bind a broader range of albumins (Johansson, Nilsson et al. 2002).

Unfortunately, the identity and impact of functional determinants contained within the family of albumin binding domains remains largely unexplored. As is the case for many protein families, thermodynamic, kinetic, and structural data is unavailable for most members. Despite the availability of extensive biochemical data on two distinct members of the GA module (Sjobring 1992; Johansson, de Chateau et al. 1995; de Chateau, Holst et al. 1996; Kraulis, Jonasson et al. 1996; Johansson, de Chateau et al. 1997; Gulich, Linhult et al. 2000; Johansson, Frick et al. 2002; Johansson, Nilsson et al. 2002; Linhult, Binz et al. 2002), including a recently published crystal structure of ALB8-GA complexed with human serum albumin (Lejon, Frick et al. 2004), much remains unknown about the impact of module polymorphisms on domain structure and function.

One promising technique for deciphering the manners in which the GA module and other protein families encode species-specific traits involves creating a library of recombinant homologs that can be probed with phage display for variants that accommodate specific selection criteria. Analysis of the phage-selected mutants could provide significant insights into the natural mechanisms behind phenotypic diversity, permit researchers to predict the behavior of unexamined homologs, and help guide subsequent research.

Zhao and Arnold initially demonstrated the value of phage-displayed recombinant libraries in evaluating the impact of specific polymorphisms on a pair of subtilisin E mutants (Zhao and Arnold 1997). However, an accumulation of experimental and computational data (Kikuchi, Ohnishi et al. 1999; Moore, Maranas et al. 2001; Moore and Maranas 2002; Maheshri and Schaffer 2003) suggests that the traditional DNA shuffling strategy developed by Stemmer (Stemmer 1994; Stemmer 1994) and used by Zhao and Arnold in their landmark study becomes ineffective when applied to coding regions of decreasing size and homology. Although other strategies have been developed specifically to promote recombination among heterologous sequences (Kikuchi, Ohnishi et al. 1999; Ostermeier, Nixon et al. 1999; Kikuchi, Ohnishi et al. 2000; Gibbs, Nevalainen et al. 2001; Lutz, Ostermeier et al. 2001), few readily produce the density of crossover events needed to efficiently shuffle families of small globular proteins like that of the GA module.

OR-PCR is a novel strategy that appears to be capable of creating recombinant libraries from compact heterologous domains. Chapter 2 characterization of the technique, which exploits elevated recombination frequencies near template ends and the exponential accumulation of recombinant templates during PCR, suggests that OR-PCR can generate multiple recombination events among compact heterologous domains similar in size and complexity to those defined by the GA module.

This chapter describes the use of OR-PCR to create a library of recombinant GA modules, which is probed by phage display in an attempt to uncover differences between GA mutants required to bind one or two distinct albumin species. The two most prominent phage-selected domains were subjected to circular dichroic,

calorimetric, and limited structural analysis in order to identify structural and functional determinants within the GA module and possibly determine whether backbone dynamics do in fact contribute to the broad affinity for different albumins observed in G148-GA3.

Experimental Results

Reconstructing the Native GA Sequence Space.

The 16 known members of the bacterial albumin binding module describe a finite sequence space, which encodes a range of three-helix domains with varied stabilities and albumin binding potentials. I sought to identify some of the biochemical determinants that specify phenotypic variation in these domains by shuffling representatives of the natural sequence space and selecting for broad or narrow albumin binding affinities.

This effort began by assembling the 56 amino acid protein (A002), which contained the 46 amino acid streptococcal albumin binding domain, G148-GA3, surrounded by unstructured flanking sequences. For consistency, the complete A002 amino acid sequence shown in **Table 5** is largely identical to that described in the previous chapter and used in earlier structural studies of the streptococcal domain (Kraulis, Jonasson et al. 1996; Johansson, de Chateau et al. 1997). NotI and PstI restriction sites were used to insert A002 into pHEN1, an amp^R-tagged phagemid whose multiple cloning site supports the display of protein or polypeptide libraries on

Table 5. Native, Template, and Phage-Selected Albumin Binding Domains.

		10 20 30 40 50
	A002	MEAVDANSLAEAKVLANRELDKYGV-SDYYKNLINNAKTVEGVKALIDEILAALPTE
Native Domains [a, b]	*F. magna* L3316-GA1	.KN..EE.IK..KEA.IT..L.FS...K.......E..KN...K.
	F. magna L3316-GA2	.KN..ED.IK..KEA.IS..I.FDA..K.......E..KN...K.
	F. magna L3316-GA3	.KN..EA.IK..KEA.ITAE.LF....K.......ES.KN...K.
	F. magna L3316-GA4	.KN..ED.IK..KEA.IT..I.FDA..K...I...E..KN...K.
	F. magna ALB1-GA	.KN..ED.IA..K.A.IT..F.F.A..K......AN..KN...K.
	F. magna ALB8-GA	.KN..ED.IA..K.A.IT..F.F.A..K.....E.N..KN...K.
	F. magna ALB1B-uGA	.Q...DK.IQ.AKAN.LT.KLLKN.E....P.SA.SFAE.LIKS
	F. magna ALB8-uGA	.KLT.EE.EKA.K.L.IT.EFIL.Q.DK.TSR..LES.VQT.KQS
	S. dysgalactiae MAG-GA1	..KLAADTDLD..VAKIIN.-.TTKVE....A.D..KIFE.--SQ
	S. dysgalactiae MAG-GA2	..K..AD.IEI.K...I-G...IK....G..A...T..K.....S
	S. equi ZAG-GA	.L...EA.IN..KQ..I-....VT...K.......N..KA...S.
	S. canis DG12-GA2	.S...EM.I....AQ..-..F...K.........V..K.L..NS
	S. canis DG12-GA1	.DQ..QA.LK.F.R...-.N.......K......IME.QAQVVES
	S. Streptococcus G148-GA1	..K..AD.LK.FN....-..................D.QAQVVES
	S. Streptococcus G148-GA2	.................-...H.............D.QAQVVES
	S. Streptococcus G148-GA3	.................-..........................
Template Domains [b, c]	TD-1	AA LK G A T .....KN..EE.IR.....-.....N..KA.
	TD-2	A L ...ED.IEI.K...I-G...IK.....ES.KN...K.
	TD-3	A LS Q ...EA.IR..KK..I-....VT.....D.QAQVVES...
	TD-4	A N NFH ...EM.I....AQ..-.....K.....D.QAQVVES...
	TD-5	AQ Q LK LNR G A T ....EK..EA.IR.F.K...-.....N..KA.
	TD-6	SSE S .....KEA.ITA...F....K.....D.QAQVVES...
	TD-7	S L D .....KEA.IT..T.F.A..K.....ES.KN...K.....
Phage-Selected Domains [b, d, e]	PSD-1 (8, 10, 9, 9)	.....Q..EA.IK..KQ..I-G...IK.....ES.KN...K.....
	PSD-2 (0, 0, 1, 0)	.....TQ..EA.IK..KQ..I-G...IK.....ES.KN...K.....
	PSD-3 (0, 0, 1, 0)	.....AD.IEI.K...I-G...IK.....ES.KN...K.....
	PSD-4 (0, 0, 1, 0)	.....ED.LEI.K...I-G...IK.....ES.KN.....
	PSD-5 (0, 0, 0, 1)	.....ED.LEI.K...I-G...IK.....ES.KN.....
	PSD-6 (1, 0, 0, 0)	.....EA.LS..KQ..I-....VT.....ES.KN...K.....
	PSD-7 (2, 0, 0, 1)	.....EM.....-.....ES.KN...K.....
	PSD-8 (0, 1, 0, 0)	.....-EM.....-.....ES.KN...K.....
	PSD-9 (0, 0, 0, 1)	.....QM.....R...-.....ES.KN...K.....
	PSD-10 (1, 0, 0, 0)	.....V......KEA.IT..L.FDA..K...A...N..KA.....
	A002	MEAVDANSLAEAKVLANRELDKYGV-SDYYKNLINNAKTVEGVKALIDEILAALPTE
		10 20 30 40 50

[a] Sequences of the native domain were previously compiled by Johansson (Johansson, de Chateau et al. 1997).

[b] Gray characters are represented in only one of the three sequence sets defined by the native, template, or phage selected domains.

[c] Underlines reveal the extent of primers used to construct the templates from A002 via site directed mutagenesis. Stacked characters represent degeneracies introduced by the use a randomized nucleotide in the corresponding codon.

[d] Gray blocks describe the regions in which crossover events likely occurred during OR-PCR.

[e] The values in parenthesis indicate the number of times each domain was identified in sequences obtained from phage-selected challenge sets A, B, C, and D.

the surface of M13 phage by fusing the cloned fragments to the N-terminal of the gIII capsid protein (Hoogenboom, Griffiths et al. 1991).

Rather than reconstructing each of the sixteen homologs shown in **Table 5**, I used PCR site-directed mutagenesis to create seven variants of A002, which cumulatively represented much of the natural diversity found among members of the GA module. These variants have been labeled template domains (TD) one through seven in **Table 5**. Randomized bases were used to further increase TD coverage for the naturally defined GA sequence space. Although there is a slight disparity between the sequence spaces represented by the seven templates and the native GA domains, the approach significantly reduced the number of primers and reactions required to produce a starting library. On average, each of the seven templates exhibits an 83% homology to one another within the 132 nt variable region subject to recombination.

Shuffling Template Domains Via OR-PCR.

Chapter 2 experiments suggest that OR-PCR offers an effective strategy for promoting recombination among compact heterologous domains similar in size and complexity to the members of the GA module studied here. The technique significantly elevates recombination frequency during standard PCR by locating the recombinant region near one end of the amplicon. Since the technique appears to benefit from the accumulation of shuffled templates in the reaction mix, recombination rates can be further increased by passing the products through multiple iterations of the OR-PCR process.

I applied OR-PCR to an equal mixture of the seven pHEN1/TD constructs by designing primers to produce a 766 bp amplicon with the start of the 177 nt GA coding region located only 24 nt from one end. The optimal elongation time for generating crossover events within the offset GA coding region was determined to be approximately one minute—a value obtained by creating a similarly-sized amplicon on the pUC19-03 and pUC19-05 constructs described in Chapter 2 and measuring the impact of elongation time on *lacZ* phenotype rescue rates.

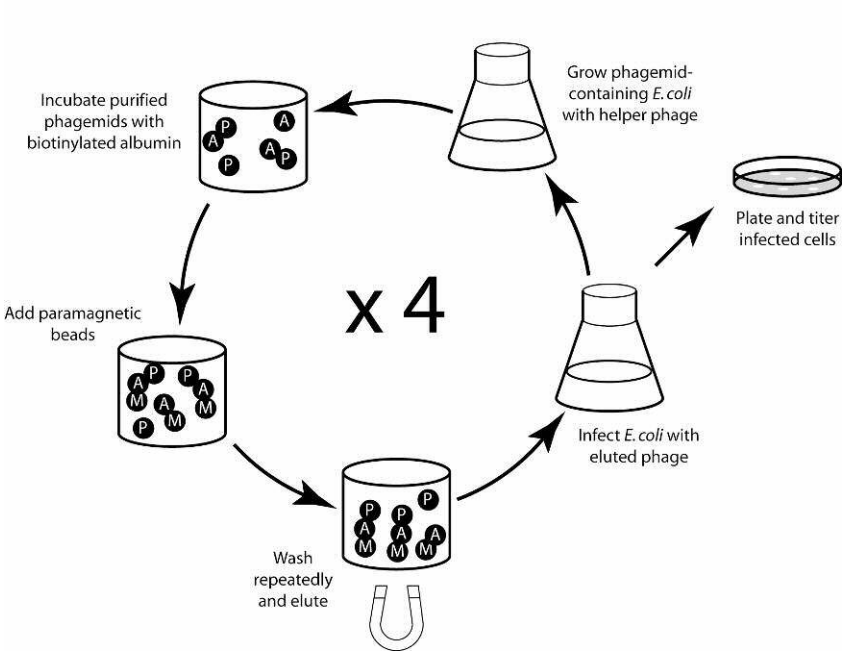
An equal mixture of the pHEN1/TD variants was shuffled via six consecutive rounds of OR-PCR. During this process, 2 μL samples from each 30-cycle reaction were transferred to fresh reaction mixes. After six rounds the final product was reintroduced into pHEN1 and cloned in *E. coli* for infection with M13KO7 phage. Sequence analysis (data not shown) revealed that twelve of fifteen pHEN1 samples obtained from phage-infected *E. coli* colonies had undergone recombination within the GA coding region with five of the sequences containing two to three crossover events. Eight of the fifteen sequences exhibited frame shift mutations that were likely to destroy the integrity of the albumin binding module. The presence of frame shift mutations contrasted sharply with Chapter 2 OR-PCR experiments, which showed no evidence of insertions, deletions, or point mutations after similar treatment. Although the frame shift mutations reduced the number of viable species in the recombinant library by as much as half, there was no evidence that any of the deleterious mutations were propagated through phage selection process described below.

Selecting Phage-Displayed Mutants.

It has been proposed that the relatively dynamic G148-GA3 structure may contribute to the domain's broad affinity for albumin from different species (Johansson, Nilsson et al. 2002). I sought to explore the notion that backbone dynamics were somehow tied to albumin specificity by designing a selection protocol in which phage displayed GA recombinants were required to bind HSA, GPSA, or both albumins.

HSA and GPSA were chosen as target ligands because they appear to represent the diverse range of albumins bound by G148-GA3. Competitive binding assays show only a ten-fold difference in the abilities of the streptococcal domain to bind HSA and GPSA (Johansson, Frick et al. 2002). At the same time, the *F. magna* ALB8-GA albumin binding domain was found to be a thousand fold less capable at binding GPSA than HSA. Sequence analysis also recommended HSA and GPSA as ideal candidates for a study of GA binding specificities because the two albumins are on opposite ends of a phylogenetic tree depicting albumins from eleven different species (not shown here).

Four identical aliquots (challenge sets A-D) of phage displaying the GA recombinant library were each subjected to four consecutive rounds of selection, amplification, and precipitation as depicted in **Figure 14**. Challenge sets A and D were passed respectively over only HSA or GPSA labeled beads during each of the four rounds. Sets B and C alternated between the two albumins after each round in an effort to select mutants that could efficiently bind both species of ligand. Sets B and



Round	Challenge Set A Ligand	Challenge Set A Output [pfu/mL]	Challenge Set B Ligand	Challenge Set B Output [pfu/mL]	Challenge Set C Ligand	Challenge Set C Output [pfu/mL]	Challenge Set D Ligand	Challenge Set D Output [pfu/mL]
1	HSA	2.0×10^5	HSA	1.2×10^5	GPSA	7.6×10^5	GPSA	8.0×10^5
2	HSA	5.3×10^6	GPSA	2.3×10^7	HSA	2.0×10^6	GPSA	2.5×10^7
3	HSA	2.1×10^6	HSA	1.2×10^6	GPSA	1.4×10^7	GPSA	1.2×10^7
4	HSA	6.1×10^7	GPSA	7.4×10^7	HSA	3.4×10^7	GPSA	7.1×10^7
Isolates	PSD-1 (x8), PSD-6, PSD-7 (x2), PSD-10		PSD-1 (x10), PSD-8		PSD-1 (x9), PSD-2, PSD-3, PSD-4		PSD-1 (x9), PSD-5, PSD-7, PSD-9	

Figure 14. Panning for HSA- and GPSA- Binding Mutants. Four rounds of selection, amplification, and purification were carried out on challenge sets A-D. During selection, purified phage (circle-P) displaying GA recombinants were incubated with biotinilated albumin (circle-A), exposed to streptavidin-coated paramagnetic beads (circle-M), and washed repeatedly with a magnetic manifold to remove unbound phage. Infected *E. coli* were grown in the presence of helper phage to amplify selected phage and complete the cycle. Phage-infected *E. coli* were also plated to obtain the output titers reported in the accompanying table. Depending on the challenge set, albumins were either varied or maintained from one round to the next. Eleven to twelve phagemids were isolated from titer plates and sequenced at the end of round four to identify PSD-1-10 reported in **Table 5**.

C differed only in their starting ligands in order to observe whether the initial selection criteria proved critical in determining which mutants were enriched.

After four rounds of amplification and selection challenge sets A-D showed significant signs of enrichment based upon elevated titers of phage in the eluant (**Figure 14**). For each of the four sets, eleven to twelve colonies of *E. coli* infected with phage derived from the fourth elution were isolated and sequenced. Ten distinct phage-selected domains (PSD 1-10) were identified in the 47 sequences obtained from the challenge sets. More than half of these sequences, which are listed in **Figure 14** and displayed in **Table 5**, exhibited two distinct cross over events. Two point mutations and a single codon deletion were also found among the selected mutants. Each of the biopanned challenge sets revealed a clear preference for PSD-1, which was represented by 36 of the 47 samples. The second most common mutant, PSD-7, was identified in 3 of the sequenced samples. All other mutants were found only once and tended to be close variants of PSD-1 or PSD-7.

Significantly, there was no discernable difference in the types or distributions of mutants appearing in each of the four challenge sets—suggesting that sequence polymorphisms in the human and guinea pig serum albumins had little effect on their respective abilities to enrich the dominant PSD-1 mutant.

Circular Dichroic Analysis of Selected Mutants.

Circular dichroism was used to assess the structural and thermodynamic properties of folded PSD-1 and PSD-7. **Figure 15** presents normalized CD data for both mutants alongside similar data obtained for G148-GA3. Per **Figure 15A**, the

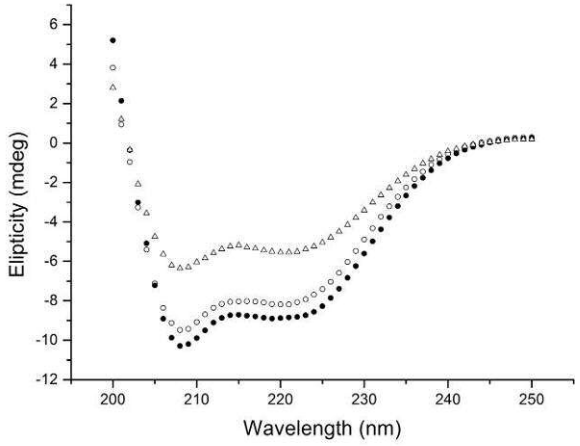
CD profiles for both mutants are similar in shape to those previously observed for G148-GA3 (Kraulis, Jonasson et al. 1996; Gulich, Linhult et al. 2000) and indicative of peptides with a predominantly alpha-helical content. The G148-GA3 sample studied in Chapter 3 uniquely contained a disordered six-histidine tag on its C-terminal, which likely contributed to the molecule's weaker signal in the **Figure 15A** normalized plot.

When mutant and wild-type domains were heated from 25°C to 100°C, they showed clean transitions from folded to unfolded states as indicated by the ellipticity at 222 nm. These transition curves were fit with upper and lower baselines in **Figure 15B** to determine the ratio of unfolded to folded proteins (K) and derive temperature dependent values of ΔG using the equation

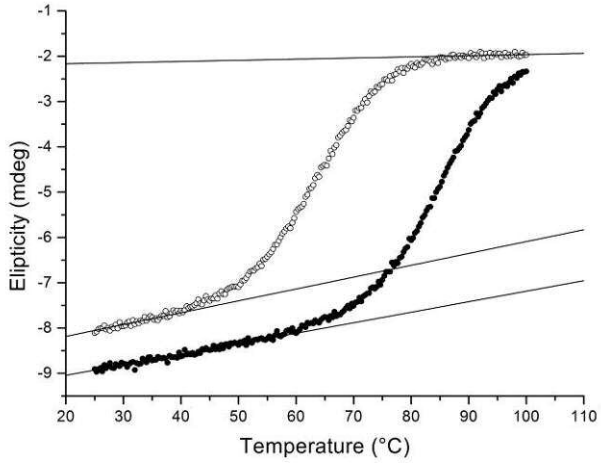
$$\Delta G = -RT \cdot \ln(K) \tag{15}$$

where R is the gas constant. This transformation, which is plotted in **Figure 15C**, reveals that PSD-1 is more stable than G148-GA3 in the measured temperature range with a T_m approximately 13°C above the 72°C value obtained for the wild-type domain. The opposite is the case for the less stable PSD-7, which reveals a T_m of 65°C. The relative stabilities of the three domains are probably maintained as the solution reaches room temperature. However, the CD data does not permit us to accurately determine ΔC_P and predict the behavior of the free energy curves at lower temperatures.

A



B



C

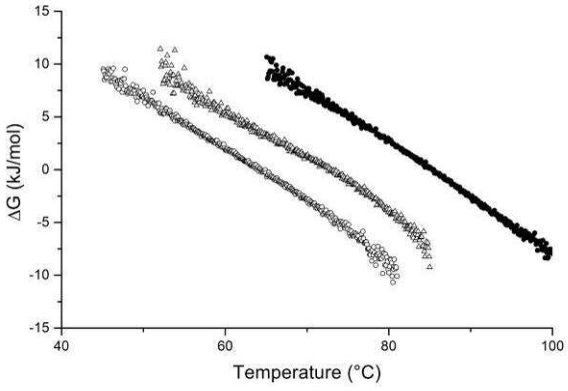


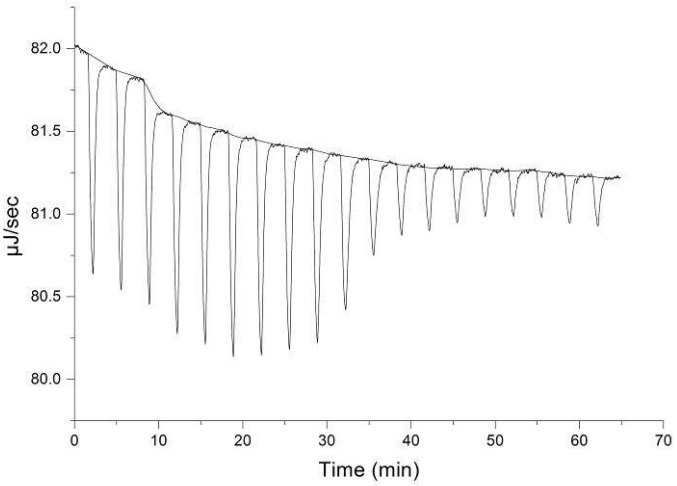
Figure 15. CD Analysis of Phage-Selected Mutants. Values are given for PSD-1 (solid circles), PSD-7 (open circles), and G148-GA3 (open triangles) in 50-100 mM $KHPO_4$ pH 7.0. G148-GA3 data were obtained from experiments described in the previous chapter. (**A**) Spectral scans normalized to 1.0 μM for each mutant at 25°C suggest that the folded proteins adopt alpha helical structures similar to those observed for the wild-type domain. (**B**) The temperature-dependent fraction of unfolded proteins was determined by fitting baselines to the normalized 222 nm transition curves generated by heating PSD-1 and PSD-7 from 25-100°C. The upper baseline for PSD-1 was approximated as being roughly equivalent to that of the normalized PSD-7 in its unfolded state. Baselines are fit to G148-GA3 in **Figure 9B**. (**C**) $\Delta G_{\text{unfolding}}$ was computed for each domain from the unfolded fraction.

Characterizing the Albumin Binding Reactions of Selected Mutants.

ITC was used to determine thermodynamic state functions and binding constants for PSD-1 and PSD-7 interactions with both HSA and GPSA at 25°C. The calorimetric technique measures heat produced when small amounts of protein bind an excess of ligand. The enthalpy involved in each binding reaction is equal to the area under the calorimetric curve. Enthalpies from successive reactions can be plotted as a function of the protein-ligand molar ratios to produce a transition curve and compute the binding constant *K*. Values for ΔG, ΔH, and ΔS are easily derived form this analysis. An example of the calorimetric data obtained for PSD-1/HSA binding at 25°C is given in **Figure 16**. The midpoints for each of the binding reactions occurred when equal molar concentrations of protein and ligand are present, indicating one-to-one stoichiometry.

According to the ITC data reported in **Table 6**, the PSD-1/HSA binding constant is nearly five times that of the native streptococcal domain. Furthermore, PSD-1's GPSA binding constant is twice the value obtained for HSA. This is significant given the observation that competitive binding assays for G148-GA3 and ALB8-GA show ten and thousand fold decreases in their abilities to bind GPSA as compared to HSA (Johansson, Frick et al. 2002). Remarkably, PSD-7, which is less stable than G148-GA3 at higher temperatures, yields modest gains over the native domain in binding HSA while maintaining an equivalent affinity for GPSA.

A



B

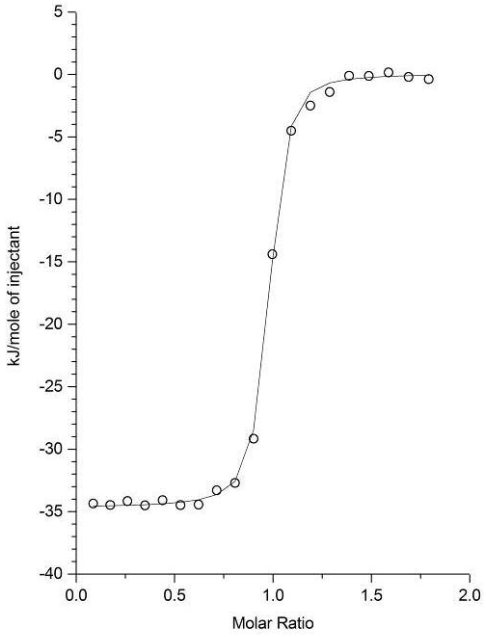


Figure 16. ITC Analysis of the PSD-1/HSA binding at 25°C. Measurements were made in 50 mM $KHPO_4$ pH 7.0. (**A**) Instantaneous heat generated by adding 1.2 nmol aliquots of PSD-1 to 12.6 nmol HSA at three-minute intervals. (**B**) Total heat produced per mole of injectant as a function of the molar ratio of protein to ligand. The binding constant and van't Hoff enthalpy corresponding to the transition curve are reported in **Table 6**. The midpoint of the transition occurs at equivalent concentrations of protein and ligand, indicating one-to-one stoichiometry.

Table 6. Thermodynamic Data for G148-GA3 and Phage-Selected Mutants. [a]

		G148-GA3[b]	PSD-1	PSD-7
HSA Binding	**K (mol^{-1})**	**1.2 (± 0.1) x 10^7**	**5.4 (± 0.6) x 10^7**	**3.5 (± 0.6) x 10^7**
	ΔG (kJ/mol)	**-40.4 ± 0.2**	**-44.1 ± 0.3**	**-43.0 ± 0.4**
	ΔH (kJ/mol)	**-31.9 ± 0.1**	**-8.27 ± 0. 04**	**-11.5 ± 0.1**
	ΔS (J/mol)	**29 ± 1**	**120 ± 1**	**106 ± 2**
GPSA Binding	**K (mol^{-1})**	**0.9 (± 0.8) x 10^7**	**1.1 (± 0.2) x 10^8**	**0.7 (± 0.2) x 10^7**
	ΔG (kJ/mol)	**-40 ± 2**	**-45.9 ± 0.4**	**-39.1 ± 0.6**
	ΔH (kJ/mol)	**-3.7 ± 0.2**	**-6.41 ± 0.05**	**-3.4 ± 0.1**
	ΔS (J/mol)	**121 ± 6**	**132 ± 2**	**129 ± 2**

[a] All thermodynamic state functions are reported at 25°C for samples in 50 mM $KHPO_4$ pH 7.0.

[b] Data obtained from Chapter 3 analysis of A002HC.

Identifying Potential Functional Determinants in Selected Mutants.

In an effort to associate phage-selected sequence polymorphisms with observed changes in protein thermodynamics for folding and binding, the spatial arrangements of PSD-1 and PSD-7 mutations were considered by aligning an unpublished NMR structure for PSD-1* with an NMR structure for G148-GA3 (Johansson, Frick et al. 2002) (PDB #1GJS) and a crystal structure of the HSA-bound ALB8-GA (Lejon, Frick et al. 2004) (PDB #1TFO). The structural alignments shown in **Figure 17** were made using the National Center for Biotechnology Information's (NCBI) Vector Alignment Search Tool (VAST).

* The NMR structure for PSD-1, which was recently obtained by Yanan He, Yihong Chen, and John Orban, will likely be the subject of a separate publication.

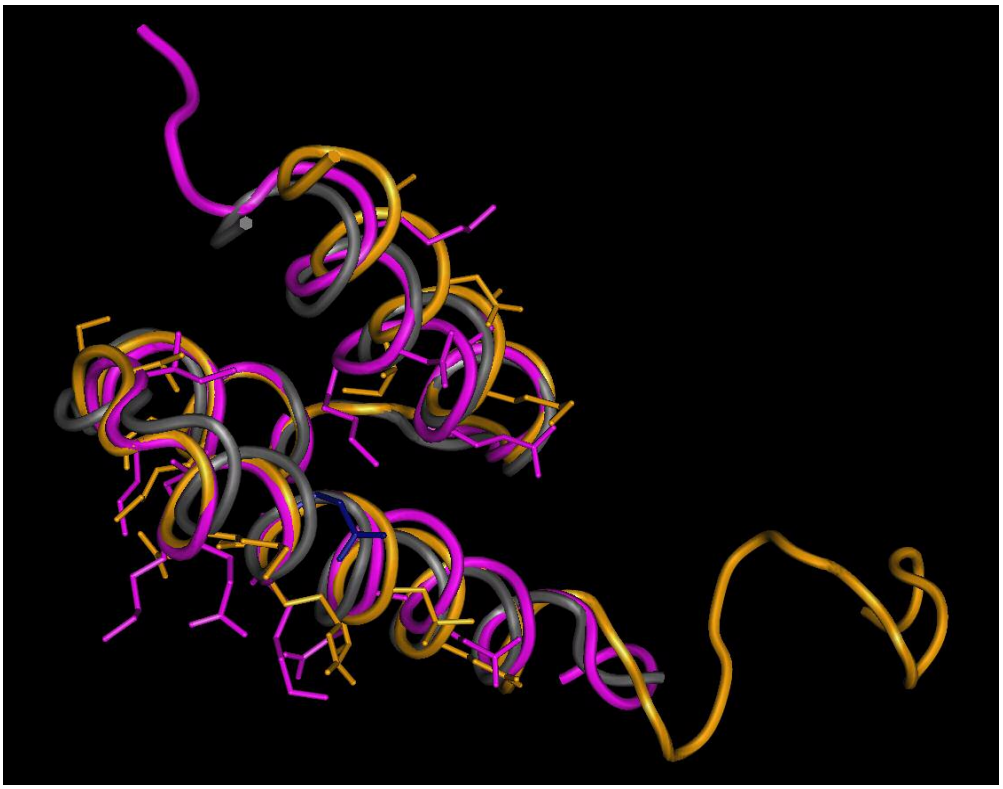


Figure 17. Structural Comparison of Mutant and Wild-Type Albumin Binding Domains. The recently-obtained, unpublished NMR structure for PSD-1 (purple) is aligned with the NMR structure for G148-GA3 (orange) and albumin-bound crystal structure of ALB8-GA (gray) using NCBI's VAST algorithm. The C-termini for all three structures are at the top of the diagram. The HSA, which is bound to the ALB8-GA surface defined by the second and third helices is not shown. Side chains are displayed only for PSD-1/G148-GA3 polymorphisms. The glutamic acid (position 19), which is located on all three domains and appears to interact with lysine (position 47) in PSD-1 and ALB8-GA, is rendered in blue. The PSD-1 structure obtained by Yanan He, Yihong Chen, and John Orban, will likely be the subject of a separate publication.

Although the PSD-1 and G148-GA3 NMR structures do not identify the albumin binding epitope, they share sufficient sequence identity and structural homology with ALB8-GA at the albumin interface to conclude that all three molecules encode a common binding epitope spanning the surface defined by the second and third helices of the module. Of the two wild-type albumin binding domains considered in this paper, PSD-1 and PSD-7 share the greatest sequence identity with G148-GA3

ITC analysis indicates that PSD-7 exhibits elevated and equivalent affinities for HSA and GPSA respectively when compared to G148-GA3, despite the recombinant's lower stability at high temperatures. Improvement to PSD-7 albumin binding is likely supported by A45S, which appears by comparison with the crystallized ALB8-GA/HSA complex to hydrogen bond with T260 on HSA (Accession No. P02768) and GPSA (Accession No. AY294645). This interaction is not supported by G148-GA3 and could give the less stable PSD-7 an added advantage in binding the two albumins. While the same mutation exists in PSD-1, its impact on albumin binding likely is mitigated by the simultaneous loss of a native G148-GA3 interaction with HSA and GPSA N342 through S27G.

Absent substantial enhancements to the binding epitope, PSD-1's superior ability to bind HSA and GPSA compared to PSD-7 and G148-GA3 might be driven by the gains in the mutant's stability observed during CD melts. The ten polymorphisms that differentiate PSD-1 from PSD-7 mostly are located on solvent-exposed surface residues of the first two helices. Being confined to the surface of the domain, these polymorphisms are unlikely to significantly destabilize PSD-7 through steric conflicts with other residues. Furthermore, these polymorphisms do not appear to encode the

changes in protein stability via discernable differences in their hydrophobic natures. In fact, the three PSD-1 mutations that do impact protein hydrophobicity (N17I, S27G, and K31I) collectively increase the hydrophobic surface area of PSD-1—changes that are likely to reduce rather than enhance the overall stability of the domain by promoting thermodynamically unfavorable order in the surrounding solvent.

Fortunately, the recently obtained PSD-1 NMR structure offers further insights into the domain's enhanced stability and binding affinity. While all three domains in **Figure 17** show a tight alignment of the first and second helices, the third helix of unbound G148-GA3 appears to reside further from the protein's core than the bound ALB8-GA. This is not the case with the corresponding PSD-1 helix, which remains closely aligned with the ALB8-GA helix despite the fact that the former molecule is unassociated with its ligand. Provided experimental error does not contribute significantly to the structural variations found in the third helix, the unpublished NMR data obtained by Yanan He *et. al.* suggests that PSD-1 achieves further stability over G148-GA3 by more tightly associating the third helix with the domain's core. Furthermore, observed improvements to PDS-1 HSA and GPSA binding may result from the closer conformational alignment the unbound PSD-1 helix shows with the bound ALB8-GA helix.

Only two G148-GA3/PSD-1 polymorphisms appear to be capable of directly impacting the relative stabilities of these domains and position of the third helix, although other polymorphisms could have less obvious effects on the structures. The first of these mutations (V25I) may contribute to PSD-1 stability by replacing a valine

at the intersection of the first and second helices with a more bulky isoleucine, which could help to fill a solvent accessible gap observed in the G148-GA structure. The second mutation (I47K) appears to bring the third helix closer to the core by removing a bulky isoleucine from this position and causing the helix to rotate slightly relative to its neighbors. An ionic interaction between PSD-1's E19 and K47, which is more accurately rendered in the ALB8-GA crystal structure may, further support the positioning of the third helix with respect the core of the domain and the albumin binding interface.

Discussion

Families of homologous proteins such as the GA albumin binding module provide opportunities for biologists to study the mechanisms underlying observed phenotypic variations. Unfortunately thermodynamic, kinetic, and structural data is rarely available for more than a few members of a family. Even when multiple homologs have been characterized, the diverse evolutionary pressures placed upon native proteins by their environments make it difficult to predict how polymorphisms support discrete biological functions.

Phage display and selection of recombinant libraries offer a promising strategy for unraveling the complexities of natural sequence spaces by permitting researchers to efficiently pan for functional determinants under well-defined conditions. In this chapter I described the use of OR-PCR to shuffle a library of seven homologs defined by the natural sequence space of a medically significant family of small globular proteins. Despite the occurrence of frame shift mutations in some members of the

recombinant population, OR-PCR proved to be highly effective at generating double or triple crossover events in half the sequenced samples, probably due to the accumulation of recombinant templates and primers in the mix.

Far from revealing a distinct set of determinants for recombinants required to bind either one or two species of albumin, each of the four challenge sets exhibited an overwhelming preference for the PSD-1 mutant. These results fail to support the possibility that G148-GA3's dynamic backbone enhances its ability to efficiently bind phylogenetically diverse human and guinea pig serum albumins (Johansson, Nilsson et al. 2002) However, it is certainly possible that the streptococcal domain's dynamic structure supports binding to albumins not considered here.

The recently obtained NMR structure for PSD-1 suggests that the domain might achieve substantial increases in stability and albumin binding affinity over the wild-type G148-GA3 by stabilizing the third helix against its core in a conformation that closely resembles that of the albumin-bound ALB8-GA domain. This transformation appears to be driven in part by the acquisition of a lysine at position 47, which extends from the mutant's core in a manner that is reproduced in the ALB8-GA structure and may be maintained in more than half the known GA homologs based on sequence analysis. These results suggest a course for future research and underscore the value of OR-PCR and phage display in uncovering novel and potentially predictive insights from the complex array of natural polymorphisms found in many protein families.

Materials and Methods

Primers

P4: 5’ - CGA AGG TGT TGG CGA ACA GGG AGT TGG ACA AAT ACG GCG TGT CGG ATT ATT ACA AGA ATC TGA TCA ACA ACG CC - 3’; P5: 5’ - CGA TTA ACG CCT TCA CGC CTT CCA CGG TTT TGG CGT TGT TGA TCA GAT TCT TGT AAT AAT CCG - 3’; P6: 5’ - CAA GCG ATC CTG CAG CAT ATG GAG GCC GTG GAC GCC AAC AGC CTG GCG GAG GCG AAG GTG TTG GCG AAC AGG - 3’; P7: 5’ - GCT CAC GGC AGT CGC GGC CGC GAA TTC CGT CGG CAA GGC CGC CAA GAT CTC GTC GAT TAA CGC CTT CAC GCC - 3’; P8: 5’- TTT TTG TGA TGC TCG TCA GG -3’; P9: 5’- TTC TGA GAT GAG TTT TTG TTC TGC -3’; P10: 5’ - CCG CTG GAT TGT TAT TAC TCG - 3’; P11: 5’ - AAA AAG GAT CCG AGC GTC GCT TAC GTT GAA GAA GAC AAA GTA TTT AAA GCG ATG ATG GAG GCG GTG GAC GC - 3’; P12: 5’ - ACG TTC AAG CTT GGC CGC TTA TTC CGT CGG - 3’.

A002 Assembly

The A002 G148-GA3 construct used in this study was assembled in two consecutive polymerase chain reactions using the contiguous primers P4-P7. The final product was purified with QIAquick® PCR Purification Kit (Qiagen, Valencia, CA) before and after restriction digest with PstI and NotI for insertion into pHEN1. The pHEN1 phage display vector is described by Hoogenboom *et al.* (Hoogenboom, Griffiths et al. 1991). Correct assembly of pHEN1/A002 was confirmed by DNA sequencing as described below.

Template Construction

Seven variants of A002 (TD-1 through TD-7) were produced by introducing point mutations into pHEN1/A002 using QuickChange™ Site-Directed Mutagenesis Kit (Stratagene, La Jolla, CA) per the protocol described in Wang and Malcolm (Wang and Malcolm 1999). Each construct was generated from two consecutive QuickChange™ mutagenesis reactions in which changes were made separately to the 5'- and 3'- ends of the GA coding region. Stretches of amino acids that correspond to the complementary forward and reverse primers used during mutagenesis are underlined in **Table 5**. Where necessary, primers were constructed with randomized nucleotides to produce some of the amino acid polymorphisms shown in the table. Since these primers can be derived from the information presented here, they are not listed above. DNA sequencing of pHEN1 variants confirmed the accurate assembly of all seven templates.

Offset Recombinant PCR

OR-PCR was performed per the protocol described in Chapter 2. An equal mixture by mass of pHEN1 vectors containing TD-1 through TD-7 was subjected to six consecutive rounds of OR-PCR. The first of these reactions consisted of 2.5 U cloned *Pfu* polymerase (Stratagene, La Jolla, CA), 200 μM each dNTP, 0.5 μmol each of P8 and P9, and 100 ng of the mixed pHEN1 templates in 50 μL of the recommended reaction buffer. Subsequent reactions substituted a 2 μL aliquot from the previous reaction for the 100 ng template mix described above. Thermocycling began with 30 s at 95°C followed by 30 cycles of 30 s at 95°C, 30 s at 55°C, and 1

min at 75°C. PCR products were purified using the QIAquick® PCR Purification Kit and concentrations determined via ultra violet (UV) absorption at 260 nm.

The purified product was further amplified using a standard PCR protocol to remove heteroduplexes formed during OR-PCR and generate a smaller amplicon that was conducive to cleavage and religation into pHEN1. During amplification, 100 ng of the recombinant product was added to a PCR mix similar to the one described above and thermocycled for only 8 cycles. P9 and P10 were used to generate a product with ends extending slightly beyond the indicated restriction sites. After purification with the QIAquick® PCR Purification Kit, the product was cut with PstI and NotI for ligation into pHEN1. The recombinant plasmids were transformed into XL-10 Gold Super Competent Cells (Stratagene, La Jolla, CA), plated on LB agar containing 100 μg/mL ampicillin, and incubated overnight at 37°C.

Phage Production

Transformed XL-10 Gold Super Competent Cells were grown in 20 mL YT with 100 μg/mL ampicillin until mid log phase. A 1 mL aliquot of the culture was then transferred to a fresh stock of 20 mL YT containing ampicillin and 10^7 pfu M13KO7 helper phage (New England Biolabs). After 1 h at 37°C, 80 μg/mL kanamycin was added to the culture. Incubation continued at 37°C with vigorous shaking for 16 h.

Phage Precipitation

Phage were precipitated by spinning the cell culture twice at 10k xg for 30 min and discarding the pelleted cells. 4 mL PEG/NaCl (20% PEG 8000, 2.5 M NaCl) was added to the supernatant and the solution was placed on ice for 20 min before being

centrifuged at 10k xg for 30 min. The supernatant was discarded and the pellet resuspended in 1 mL TE buffer (100 mM Tris pH 8.0, 0.1 mM EDTA). After adding 200 μL PEG/NaCl the sample was returned to ice for 20 min and centrifuged at 14k xg for 15 min. The supernatant was discarded and the pellet resuspended in 1 mL TE buffer for storage.

Biopanning

A 50 μL solution of Dynabeads® M-280 Streptavidin paramagnetic beads (Dynal Biotech) was pelleted on a magnetic manifold and resuspended in TBS Tween (50 mM Tris pH 7.4, 150 mM NaCl, 0.5% Tween 20) and 0.1% dried milk. This solution was rocked overnight at 4°C before repelleting the beads on a magnetic manifold. Meanwhile, 10^8 pfu pHEN1-containing phage were mixed with 1 μL 10 nM biotinylated albumin and 1 mL TBS Tween before rocking at room temperature for 3 h. The dried essentially fatty acid free HSA and GPSA samples used in this study were obtained from Sigma. The pelleted streptavidin beads were resuspended in the 1 mL phage solution and rocked for 30 min. Afterwards, the beads were returned to the magnetic manifold to remove the supernatant before being washed seven times with 1 mL TBS—rocking for five min at room temperature during each wash. Finally, the phage were eluted by resuspending the beads in 200 μL 0.1 M glycine pH 2.1 with 1 mg/mL bovine serum albumin (Sigma) and rocked for 20 min. The beads were pelleted and discarded before 10 μL 2 M Tris base was added to the supernatant to neutralize the acid. Titers of the selected phage were obtained by creating serial dilutions of the neutralized solution, mixing 1 μL of each dilution with 100 μL stationary phase TG-1 cells, and plating on LB agar laced with 100 μg/mL ampicillin.

Phagemid-containing colonies were counted after incubating the plates overnight at 37°C.

The remaining eluted phage were mixed with 20 mL YT, 100 μg/mL ampicillin, and 200 μL stationary phase TG-1 before shaking vigorously overnight at 37°C. 10^7 pfu M13KO7 helper phage were added to the culture 1 h into incubation to support production of GA-labeled phage. The resulting phage were precipitated as described above and the entire biopanning process repeated three more times before colonies were isolated from the titer plates for DNA sequencing and analysis.

Protein Production and Purification

Two of the phage-selected mutants (PSD-1 and PSD-7) were prepared for production, purification, and analysis by PCR amplification using P11 and P12. The PCR products were cut with BamHI and HindIII for ligation into the pG58 vector. P11 and P12 were used specifically to add restriction sites and an ochre stop codon to the ends of the GA coding region. The pG58 vector enables expression of a subtilisin pro domain fusion protein, which permits the fused protein to be purified and cut with subtilisin in a one-step reaction (Ruan, Fisher et al. 2004). XL-10 Gold Super Competent Cells were transform with the pG58/PSD construct and grown in 5 L LB with 100 μg/mL ampicillin. At log phase the cells were induced with 1 mM IPTG for 3 h before harvesting. Cells were pelleted by spinning at 8k xg for 30 min and resuspended in 100 mL 100 mM $KHPO_4$ pH 7.0, 30 μg/mL DNase I, and 0.1 mM PMSF for sonication on ice. Cellular debris was removed by centrifugation at 10k xg for 30 min and 100k xg for 1 h. Purification was carried out on a 5 mL S189 column

essentially as described by Ruan *et al.* (Ruan, Fisher et al. 2004) and confirmed by SDS-PAGE.

Extinction Coefficients

The Edelhoch method, as described by Pace (Pace, Vajdos et al. 1995), was used to precisely determine extinction coefficients at 278 nm for each of the thermodynamically characterized GA mutants and albumins. This set of empirically determined extinction coefficients was used to compute protein concentrations from UV spectra throughout the study. Twice the absorbance at 331 nm was consistently subtracted from that obtained at 278 nm to account for the effects of light-scattering.

Circular Dichroism

CD experiments were performed on a J-720 Spectropolarimeter (Jasco Spectroscopic Co., LTD.). Spectra for 2.96 μM PSD-1 and 2.67 μM PSD-7 in 50 mM $KHPO_4$ pH 7.0 were obtained by measuring the ellipticity from 250 nm to 200 nm of the samples in a 1.0 cm cell at 25°C. Melting temperatures for the same protein solutions were determined by measuring the ellipticity of the sample at 222 nm as it was heated in a 1.0 cm cell from 25°C to 70°C at 0.5 degrees per minute.

Isothermal Titration Calorimetry

ITC measurements were performed on the VP-ITC Micro Calorimeter (MicroCal). For each experiment the selected GA mutant and albumin were dialyzed side-by-side into 50 mM $KHPO_4$ pH 7.0 to ensure identical buffer conditions. Each run involved nineteen 15 μL injections of approximately 250 μM GA domain into a sample cell containing around 25 μM albumin. Injections lasted 30 s each and were

spaced 180 s apart. Precise protein concentrations were determined by UV spectra as described above. The jacket temperature was maintained at 25°C throughout.

DNA Sequencing

DNA samples were prepared for sequencing by growing selected colonies overnight at 37°C in LB with 100 μg/mL ampicillin. Plasmid DNA was extracted from the cell cultures using the Wizard® Plus SV Minipreps (Promega Corporation, Madison, WI). The P5 primer was used to amplify target DNA using Perkin-Elmer/Applied Biosystem's AmpliTaq-FS DNA polymerase and Big Dye terminators with dITP. Dye-terminated products were then run on an Applied Biosystems model 3100 DNA sequencer to produce sequence chromatographs.

Chapter 5: Outlook for OR-PCR, the Albumin Binding Module, and the Recombinant Analysis of Compact Heterologous Domains

OR-PCR Offers Substantial Advantages Over Existing Recombinogenic Techniques

The preceding chapters have laid the groundwork for recombinant-based analysis of compact heterologous domains by developing and characterizing a simple but effective PCR-based recombination technique and successfully applying it to the identification and analysis of functional polymorphisms found among members of the GA albumin binding module.

Experiments show that OR-PCR overcomes many of the problems associate with existing *in vitro* recombination techniques to efficiently shuffle compact heterologous domains of the complexity necessary for mutational analysis of entire protein families. By locating the recombinant region near one end of a largely homologous amplicon it is possible simultaneously promote premature termination of primer extensions within the recombinant region and subsequent reannealing of partially elongated primers to significantly elevate recombination frequencies during standard PCR. This simple mechanism appears to offer a number of advantages over competing techniques.

First, homologous parental recombination and out-of-sequence assembly are significantly reduced or eliminated by the availability of long stretches of identical sequence, which permit different members of the population to anneal without

sequence bias and support the proper alignment of heterologous regions. These stretches of identical overlapping sequences do not guarantee that the polymerase will bind and extend from heterologous regions. This is apparent from the relative scarcity of crossover events near the N-terminal of *lacZ* recombinants shown in **Figure 5**. Nor do the common regions of the amplicons guarantee proper alignment of heterologous regions as revealed by the insertions and deletions observed after GA recombination in Chapter 4. However, these results are far better than the 0.001% of the DNA shuffling fragments predicted to form full length products by computer modeling (Maheshri and Schaffer 2003).

OR-PCR also appears to benefit from the exponential accumulation of recombinant templates, which is not achieved by StEP, primerless PCR, or the other techniques described in Chapter 1, with the likely exception of Judo's PCR-based strategy (Judo, Wedel et al. 1998). The accumulation of recombinant templates over multiple rounds of OR-PCR appears to be responsible for the high concentration of crossover sites produced by the technique and the convergence of recombination frequencies observed for high and low homology *lacZ* sequences in **Figure 6.** The frequent occurrence of two or more crossover events within the 82 nt recombinant *lacZ* region is far superior to the estimated 2-3 crossover events per 1 kb predicted by DNA shuffling models (Maheshri and Schaffer 2003) and observed in experiments (Zhao and Arnold 1997).

Finally, OR-PCR offers a simplicity that many of the other techniques do not achieve. **Figure 4** shows that the technique can be optimized for high recombination rates over a broad range of elongation times, reducing the need to tinker with

experimental parameters—a challenge faced by researchers attempting to achieve reasonable performance with DNA shuffling (Moore and Maranas 2000). Furthermore, OR-PCR eliminates the need for the distinct fragmentation, assembly, and agarose gel purification steps required for many existing techniques.

Each of these factors combine to make OR-PCR an effective technique for generating a high density of recombination events among multiple members of homologous protein families. Without OR-PCR the creation, selection, and analysis of recombinant albumin binding domains addressed in the previous chapter would not have been possible.

Recombination and Phage Selection Provides Insights into GA Polymorphisms

Analysis of phage-selected recombinant domains proved to be a profitable strategy for analyzing GA sequence space and likely is generalizable to other protein families as well. Only one round of recombination and selection was required to generate clear and unambiguous preferences for a primary (PSD-1) and secondary (PSD-7) sequence that were distinct from any of the wild-type sequences and showed marked improvements in their abilities to bind both human and guinea pig serum albumins. By comparing the thermodynamic data obtained for each of these sequences with Chapter 3 analysis of the G148-GA3 domain and accumulated structural information for PSD-1, G148-GA3, and ALB8-GA (Johansson, de Chateau et al. 1995; Johansson, de Chateau et al. 1997; Johansson, Frick et al. 2002; Lejon,

Frick et al. 2004) it was possible to enhance our understanding of how sequence polymorphisms impact the performance of common albumin binding domains.

Specifically, analysis of phage-selected mutants revealed that G148-GA3 backbone flexibility is not required to support binding to the phylogeneticly diverse human and guinea pig albumins. Rather, domain stabilizing mutations serve to enhance binding to both albumins and, in the case of PSD-1, eliminate G148-GA3's ten-fold preference for HSA (**Table 6**). It is certainly possible that the dynamic G148-GA3 backbone structure reported by Johansson is instrumental in binding other species of albumin, such as horse albumin, which appears to share fewer of the polar interactions with G148-GA3 and ALB8-GA than GPSA based on analysis of the crystallized ALB8-GA/HSA complex (PDB #1TFO). Repetition of the phage-selection protocol using horse rather than guinea pig serum albumin would be easy to accomplish and could offer further insights into the potential roles of backbone dynamics and other module polymorphism on ligand affinity and species specificity.

Finally, comparative analysis of the available structures for phage-selected and wild-type domains suggest the possible impact of a partially buried lysine on enhancing GA stability and albumin binding. This observation and others like it may offer valuable insights into GA dynamics and help researchers understand the behavior of other members of the GA module, which appear to encode similar structural motifs. It is hoped that ongoing NMR studies of PSD-1 will provide more conclusive insights into how the lysine and other phage-selected mutations support the domain's enhanced stability and binding.

Anticipating a Broader Role for OR-PCR in Recombinogenic Studies

Finally, OR-PCR may provide an effective tool for tackling the more challenging problem of identifying a limited number of structural determinants in two similarly sized heteromorphs—proteins that adopt dramatically different structures. Inspired by a growing body of evidence to suggest radical changes in secondary and tertiary structure can be driven by a limited number of globally distributed mutations (Minor and Kim 1994; Mezei 1998; Cregut, Civera et al. 1999), researchers have been challenged to engineer heteromorphic domains that adopt different structures despite high sequence homology to one another (Rose and Creamer 1994). Small globular domains similar to the GA module are natural targets for this research because of their abilities to encode stable alpha and beta folds in a relatively small polypeptide. Two independent efforts used rational engineering in isolation (Dalal and Regan 2000) or combined with randomized mutagenesis and phage display (Alexander, Rozak et al. 2005) to achieve 60% and 59% sequence identities between stable alpha helical and beta sheet forming derivatives of a streptococcal protein G IgG binding domain. Using a strategy similar to that applied in Chapter 4, OR-PCR could prove instrumental in shuffling heteromorphic sequences to achieve greater sequence identity and isolate structural determinants among the remaining sequence polymorphisms.

Bibliography

Akerstrom, B., E. Nielsen, et al. (1987). "Definition of IgG- and albumin-binding regions of streptococcal protein G." J Biol Chem **262**(28): 13388-91.

Alexander, P., S. Fahnestock, et al. (1992). "Thermodynamic analysis of the folding of the streptococcal protein G IgG-binding domains B1 and B2: why small proteins tend to have high denaturation temperatures." Biochemistry **31**(14): 3597-603.

Alexander, P. A., D. A. Rozak, et al. (2005). "Directed evolution of highly homologous proteins with different folds by phage display: implications for the protein folding code." Biochemistry **44**(43): 14045-54.

Baldwin, R. L. (1986). "Temperature dependence of the hydrophobic interaction in protein folding." Proc Natl Acad Sci U S A **83**(21): 8069-72.

Becktel, W. J. and J. A. Schellman (1987). "Protein stability curves." Biopolymers **26**(11): 1859-77.

Bradley, R. D. and D. M. Hillis (1997). "Recombinant DNA sequences generated by PCR amplification." Mol Biol Evol **14**(5): 592-3.

Brandts, J. F. (1964). "The Thermodynamics of Protein Denaturation. I. The Denaturation of Chymotrypsinogen." J. Am. Chem. Soc. **86**: 4291-4301.

Cregut, D., C. Civera, et al. (1999). "A tale of two secondary structure elements: when a beta-hairpin becomes an alpha-helix." J Mol Biol **292**(2): 389-401.

Dalal, S. and L. Regan (2000). "Understanding the sequence determinants of conformational switching using protein design." Protein Sci **9**(9): 1651-9.

de Chateau, M. and L. Bjorck (1994). "Protein PAB, a mosaic albumin-binding bacterial protein representing the first contemporary example of module shuffling." J Biol Chem **269**(16): 12147-51.

de Chateau, M., E. Holst, et al. (1996). "Protein PAB, an albumin-binding bacterial surface protein promoting growth and virulence." J Biol Chem **271**(43): 26609-15.

Falkenberg, C., L. Bjorck, et al. (1992). "Localization of the binding site for streptococcal protein G on human serum albumin. Identification of a 5.5-kilodalton protein G binding albumin fragment." Biochemistry **31**(5): 1451-7.

Fraczkiewicz, R. and W. Braun (1998). "Exact and Efficient Analytical Calculation of the Accessible Surface Areas and Their Gradients." J. Comp. Chem. **19**(3): 319-333.

Gibbs, M. D., K. M. Nevalainen, et al. (2001). "Degenerate oligonucleotide gene shuffling (DOGS): a method for enhancing the frequency of recombination with family shuffling." Gene **271**(1): 13-20.

Gill, S. J., N. F. Nichols, et al. (1976). "Calorimetric determination of enthalpies of solution of slightly soluble liquids II. Enthalpy of solution of some hydrocarbons in water and their use in establishing the temperature dependence of their solubilities." J. Chem. Thermodyn. **8**: 445-452.

Gulich, S., M. Linhult, et al. (2000). "Stability towards alkaline conditions can be engineered into a protein ligand." J Biotechnol **80**(2): 169-78.

Hammarberg, B., P. A. Nygren, et al. (1989). "Dual affinity fusion approach and its use to express recombinant human insulin-like growth factor II." Proc Natl Acad Sci U S A **86**(12): 4367-71.

Hoogenboom, H. R., A. D. Griffiths, et al. (1991). "Multi-subunit proteins on the surface of filamentous phage: methodologies for displaying antibody (Fab) heavy and light chains." Nucleic Acids Res **19**(15): 4133-7.

Hvidt, A. and S. O. Nielsen (1966). "Hydrogen exchange in proteins." Adv Protein Chem **21**: 287-386.

Ingham, K. C., S. Brew, et al. (2004). "Interaction of Staphylococcus aureus fibronectin-binding protein with fibronectin: affinity, stoichiometry, and modular requirements." J Biol Chem **279**(41): 42945-53.

Johansson, M. U., M. de Chateau, et al. (1995). "The GA module, a mobile albumin-binding bacterial domain, adopts a three-helix-bundle structure." FEBS Lett **374**(2): 257-61.

Johansson, M. U., M. de Chateau, et al. (1997). "Solution structure of the albumin-binding GA module: a versatile bacterial protein domain." J Mol Biol **266**(5): 859-65.

Johansson, M. U., I. M. Frick, et al. (2002). "Structure, specificity, and mode of interaction for bacterial albumin-binding modules." J Biol Chem **277**(10): 8114-20.

Johansson, M. U., H. Nilsson, et al. (2002). "Differences in backbone dynamics of two homologous bacterial albumin-binding modules: implications for binding specificity and bacterial adaptation." J Mol Biol **316**(5): 1083-99.

Judo, M. S., A. B. Wedel, et al. (1998). "Stimulation and suppression of PCR-mediated recombination." Nucleic Acids Res **26**(7): 1819-25.

Karplus, P. A. (1997). "Hydrophobicity regained." Protein Sci **6**(6): 1302-7.

Kikuchi, M., K. Ohnishi, et al. (1999). "Novel family shuffling methods for the in vitro evolution of enzymes." Gene **236**(1): 159-67.

Kikuchi, M., K. Ohnishi, et al. (2000). "An effective family shuffling method using single-stranded DNA." Gene **243**(1-2): 133-7.

Kraulis, P. J., P. Jonasson, et al. (1996). "The serum albumin-binding domain of streptococcal protein G is a three-helical bundle: a heteronuclear NMR study." FEBS Lett **378**(2): 190-4.

Kwok, S., S.-Y. Chang, et al. (1995). Design and Use of Mismatched and Degenerate Primers. PCR Primer: A Laboratory Manual. C. W. Dieffenbach and G. S. Dveksler. Plainview, NY, Cold Spring Harbor Laboratory Press**:** 143-155.

Lejon, S., I. M. Frick, et al. (2004). "Crystal structure and biological implications of a bacterial albumin binding module in complex with human serum albumin." J Biol Chem **279**(41): 42924-8.

Libon, C., N. Corvaia, et al. (1999). "The serum albumin-binding region of streptococcal protein G (BB) potentiates the immunogenicity of the G130-230 RSV-A protein." Vaccine **17**(5): 406-14.

Linhult, M., H. K. Binz, et al. (2002). "Mutational analysis of the interaction between albumin-binding domain from streptococcal protein G and human serum albumin." Protein Sci **11**(2): 206-13.

Lutz, S., M. Ostermeier, et al. (2001). "Creating multiple-crossover DNA libraries independent of sequence identity." Proc Natl Acad Sci U S A **98**(20): 11248-53.

Maheshri, N. and D. V. Schaffer (2003). "Computational and experimental analysis of DNA shuffling." Proc Natl Acad Sci U S A **100**(6): 3071-6.

Makrides, S. C., P. A. Nygren, et al. (1996). "Extended in vivo half-life of human soluble complement receptor type 1 fused to a serum albumin-binding receptor." J Pharmacol Exp Ther **277**(1): 534-42.

Meyerhans, A., J. P. Vartanian, et al. (1990). "DNA recombination during PCR." Nucleic Acids Res **18**(7): 1687-91.

Mezei, M. (1998). "Chameleon sequences in the PDB." Protein Eng **11**(6): 411-4.

Minor, D. L., Jr. and P. S. Kim (1994). "Measurement of the beta-sheet-forming propensities of amino acids." Nature **367**(6464): 660-3.

Moore, G. L. and C. D. Maranas (2000). "Modeling DNA mutation and recombination for directed evolution experiments." J Theor Biol **205**(3): 483-503.

Moore, G. L. and C. D. Maranas (2002). "Predicting out-of-sequence reassembly in DNA shuffling." J Theor Biol **219**(1): 9-17.

Moore, G. L., C. D. Maranas, et al. (2001). "Predicting crossover generation in DNA shuffling." Proc Natl Acad Sci U S A **98**(6): 3226-31.

Myhre, E. B. (1984). "Surface receptors for human serum albumin in Peptococcus magnus strains." J Med Microbiol **18**(2): 189-95.

Myhre, E. B. and G. Kronvall (1980). "Demonstration of specific binding sites for human serum albumin in group C and G streptococci." Infect Immun **27**(1): 6-14.

Navarre, W. W. and O. Schneewind (1999). "Surface proteins of gram-positive bacteria and mechanisms of their targeting to the cell wall envelope." Microbiol Mol Biol Rev **63**(1): 174-229.

Ninkovic, M., R. Dietrich, et al. (2001). "High-fidelity in vitro recombination using a proofreading polymerase." Biotechniques **30**(3): 530-536.

Nygren, P. A., P. Flodby, et al. (1991). "In vivo stabilization of a human recombinant CD4- derivate by fusion to a serum-albumin-binding receptor." Vaccines **96**: 363-368.

Ostermeier, M., A. E. Nixon, et al. (1999). "Incremental truncation as a strategy in the engineering of novel biocatalysts." Bioorg Med Chem **7**(10): 2139-44.

Pace, C. N., F. Vajdos, et al. (1995). "How to measure and predict the molar absorption coefficient of a protein." Protein Sci **4**(11): 2411-23.

Pace, N. C. and C. Tanford (1968). "Thermodynamics of the unfolding of beta-lactoglobulin A in aqueous urea solutions between 5 and 55 degrees." Biochemistry **7**(1): 198-208.

Privalov, P. L. (1979). "Stability of proteins: small globular proteins." Adv Protein Chem **33**: 167-241.

Privalov, P. L. and S. J. Gill (1988). "Stability of protein structure and hydrophobic interaction." Adv Protein Chem **39**: 191-234.

Privalov, P. L. and N. N. Khechinashvili (1974). "A thermodynamic approach to the problem of stabilization of globular protein structure: a calorimetric study." J Mol Biol **86**(3): 665-84.

Privalov, P. L. and S. A. Potekhin (1986). "Scanning microcalorimetry in studying temperature-induced changes in proteins." Methods Enzymol **131**: 4-51.

Rose, G. D. and T. P. Creamer (1994). "Protein folding: predicting predicting." Proteins **19**(1): 1-3.

Ruan, B., K. E. Fisher, et al. (2004). "Engineering subtilisin into a fluoride-triggered processing protease useful for one-step protein purification." Biochemistry **43**(46): 14539-46.

Saiki, R. K., D. H. Gelfand, et al. (1988). "Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase." Science **239**(4839): 487-91.

Schwarz, F. P. and W. H. Kirchhoff (1988). "Biological thermodynamic data for the calibration of differential scanning calorimeters: heat capacity data on the unfolding transition of ribonuclease a in solution." Thermochim. Acta. **128**: 267-295.

Sjobring, U. (1992). "Isolation and molecular characterization of a novel albumin-binding protein from group G streptococci." Infect Immun **60**(9): 3601-8.

Sjobring, U., L. Bjorck, et al. (1991). "Streptococcal protein G. Gene structure and protein binding properties." J Biol Chem **266**(1): 399-405.

Sjolander, A., P. A. Nygren, et al. (1997). "The serum albumin-binding region of streptococcal protein G: a bacterial fusion partner with carrier-related properties." J Immunol Methods **201**(1): 115-23.

Stemmer, W. P. (1994). "DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution." Proc Natl Acad Sci U S A **91**(22): 10747-51.

Stemmer, W. P. (1994). "Rapid evolution of a protein in vitro by DNA shuffling." Nature **370**(6488): 389-91.

Sturtevant, J. M. (1977). "Heat capacity and entropy changes in processes involving proteins." Proc Natl Acad Sci U S A **74**(6): 2236-40.

Takagi, M., M. Nishioka, et al. (1997). "Characterization of DNA polymerase from Pyrococcus sp. strain KOD1 and its application to PCR." Appl Environ Microbiol **63**(11): 4504-10.

Wang, W. and B. A. Malcolm (1999). "Two-stage PCR protocol allowing introduction of multiple mutations, deletions and insertions using QuikChange Site-Directed Mutagenesis." Biotechniques **26**(4): 680-2.

Yang, Y. L., G. Wang, et al. (1996). "Long polymerase chain reaction amplification of heterogeneous HIV type 1 templates produces recombination at a relatively high frequency." AIDS Res Hum Retroviruses **12**(4): 303-6.

Zhao, H. and F. H. Arnold (1997). "Functional and nonfunctional mutations distinguished by random recombination of homologous genes." Proc Natl Acad Sci U S A **94**(15): 7997-8000.

Zhao, H., L. Giver, et al. (1998). "Molecular evolution by staggered extension process (StEP) in vitro recombination." Nat Biotechnol **16**(3): 258-61.