EVOLUTIONARY CONSTRAINTS SPECIFYING PROTEIN FOLDING AND FUNCTION

APPROVED BY SUPERVISORY COMMITTEE

Rama Ranganathan, M.D., Ph.D.

Philip Thomas, Ph.D.

Joseph Albanesi, Ph.D.

Alfred Gilman, M.D., Ph.D.

Michael Rosen, Ph.D.

ACKNOWLEDGEMENTS

I thank the members of my dissertation committee for their valuable guidance and advice. I greatly thank my principle collaborators: Alan Poole with the WW domain project, and Pilong Li and Mike Stiffler with the PDZ domain project. I also thank members of the Ranganathan lab for their help in learning not only scientific techniques, but the scientific thought process. In particular, Bill Russ for teaching me the fundamentals of molecular biology, Steve Lockless for giving me a better understanding of not only how SCA is calculated but how it is interpreted, and Mike Socolich for guidance on everything from protein expression to fly husbandry. Zifen Wang for instructions on *Drosophila* transformation, Alex Kiselev for his help with *Drosophila* biochemistry, and Kelly Liu for assistance with the GPCR project. Gürol Süel, Shan Mishra, and Stephen Helms for their help with elecrophysiology. And last but certainly not least, Rama Ranganathan who provided guidance from the most highly conceptual levels to hands-on instruction with patch-clamp recording.

EVOLUTIONARY CONSTRAINTS SPECIFYING PROTEIN FOLDING AND FUNCTION

by

CHRISTOPHER LARSON

DISSERTATION

Presented to the Faculty of the Graduate School of Biomedical Sciences

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas Southwestern Medical Center at Dallas

Dallas, Texas

June, 2007

EVOLUTIONARY CONSTRAINTS SPECIFYING PROTEIN FOLDING AND FUNCTION

Christopher Larson

The University of Texas Southwestern Medical Center at Dallas, 2007

Supervising Professor: Rama Ranganathan M.D., Ph.D.

Proteins are complex macromolecules that carry out biological functions while under constant mutational load and selective pressure during evolution. Consequently, evolution has generated protein families by exploring the set of sequences able to carry out a particular biological activity, maintaining sequence motifs critical for function while varying the rest of the protein. Statistical coupling analysis of a protein family examines an alignment of such sequences and detects the evolutionarily preserved interresidue interactions critical for the proteins' selective fitness. This set of information has provided a sufficiently detailed description of evolutionary design constraints to

allow the design of novel WW domain sequences that fold and function like natural proteins.

This work expands the initial investigations, and probes the minimal information content necessary to specify the WW domain fold, as well as the effect of increasing the coupling constraints in the design process. This work also evaluates the ability of coupling information to design larger and more complex protein folds and to specify their biological functions. Experimental expression and characterization of WW domains designed with varying levels of coupling information indicates that incorporating even small amounts of coupling information has a notable impact on these proteins' ability to fold. Moreover, different coupling-based design approaches produce results robust to details of how coupling information is incorporated. Similar experiments with designed PDZ domains and *in vivo* characterization of designed G-protein coupled receptors show that, to the extent studied, this design approach is successful with these larger and more complex proteins as well. This indicates that the typically sparse matrices of coupling values observed for a protein family capture the core evolutionary constraints on the proteins in sufficient detail to generate even complex proteins with natural-like folds and functions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

CHAPTER ONE
Introduction

STATISTICAL COUPLING ANALYSIS

Proteins are complex polymers whose constituent amino acids interact to fold into a three dimensional structure and carry out a biological activity. In general, there is no way of determining the importance of any particular interaction between two residues *a priori*. For example, even with a three dimensional structure of a protein, it is often not obvious how binding of an allosteric regulator at one site will transmit a signal across the protein to regulate activity at a distant catalytic site.

Any interaction between two residues in a protein that is critical for its function and evolutionary selectability may manifest as non-independence in the amino acid types found at those positions in related proteins. Statistical coupling analysis (SCA) is a method of quantifying such non-independence, and thereby detecting evolutionarily interacting residues within a protein family (calculations described in detail later, Lockless and Ranganathan, 1999, Sharma, 2006). The original description of this approach examined the family of PDZ domains, and found that most residues are statistically coupled to only a few other residues in their immediate vicinity while a subset of sites show statistical coupling to a network of contiguous residues that span to distant positions in the three dimensional structure. Site directed mutagenesis in one family member (PDZ3 of PSD-95) showed that perturbing residues in this coupled network leads to non-additive effects on the energetics of the domain's canonical function of binding a target peptide (Lockless and Ranganathan, 1999).

The theme of finding sparse networks of statistically coupled residues amid a majority of sites that show no or only local coupling appears to be a general property of all protein families. In the case of serine proteases, a coupled network around the catalytic site and S1 binding pocket comprises most of the residues known to be involved in determining the enzyme's specificity. In hemoglobin, a statistically coupled network links pairs of hemes through the F-helix and FG corner and across the $\alpha 1 \beta 2$ tetramerization interface – all of which are involved in structural rearrangements between the T and R states (Suel et al., 2003).

Forward experiments have validated the importance of statistically coupled networks in complex protein families. The statistically coupled network in G proteins links the nucleotide binding pocket to the switch loops, forming a contiguous network in the three dimensional structure of the GTP bound state that becomes fragmented in the GDP state. Alanine mutations of the statistically coupled residues alter the nucleotide-dependent change in affinities for adenylyl cyclase and $G_{\beta\gamma}$ more so than do control non-coupled mutations in the same vicinity, showing the importance of these residues in nucleotide gating (Hatley et al., 2003). In the nuclear receptor family, SCA of the ligand binding domain reveals a network of residues connecting the ligand binding pocket, dimerization interface, and AF-2 helix. Mutation of the coupled residues in the LXR receptor alters the ligand responsiveness of RXR/LXR heterodimers from the wild-type permissive behavior (activating transcription in response to agonist for either receptor) to conditional behavior (requiring agonist for both receptors) while mutations in uncoupled positions with similar conservation that directly pack against the coupled network do not affect permissivity (Shulman et al., 2004). These studies show that SCA of a number of

protein families identifies networks of interacting residues that are critical for allosteric activity.

SCA-based protein design

The strongest interactions revealed by SCA have been shown to mediate the activity of proteins in many different families, establishing the importance of these strong coupling values. Furthermore, most protein families exhibit sparse coupling patterns in which one or a few networks of residues (generally about 20% of the protein) show large coupling values and are strongly coupled primarily to other residues within their network (Suel et al., 2003). This suggests that coupling within these small networks is critical for the protein family, but that residues outside of the networks do not need to coevolve in order to maintain evolutionary fitness. Alternatively, another explanation for the sparseness of coupling interactions is that this analysis only detects evolutionarily conserved interactions across the entire protein family, but many of the key interactions within each specific protein are idiosyncratic rather than conserved and are undetectable with this approach. In such a model, proteins conserve a small core of interactions, but may achieve evolutionary fitness in a more individualized and non-conserved way.

To test the hypothesis that the core set of interactions detected with SCA is an essentially complete description of the interactions necessary for protein function, Socolich et al. (2005) used a sequence design algorithm that generates alignments of novel sequences using only the coupling and conservation patterns of a natural protein family. If this set of information is really a complete description of the evolutionary

constraints on a protein family, then the designed sequences should also satisfy evolution's constraints – they should behave like naturally evolved proteins. Socolich et al. used this SCA-based design algorithm on an alignment of WW domains: proteins ~36 residues in length that form 3 stranded antiparallel beta sheets and function by binding proline-containing targets. The SCA-designed sequence alignment has essentially the same coupling pattern as the natural alignment, although the designed sequences are quite diverse from the starting set with on average 59% sequence identity to their most similar natural sequences. To test whether SCA information is necessary, they also designed sequences using a similar algorithm that preserves the conservation pattern without imposing coupling information, generating sequences with an average of 56% identity to their most similar natural sequences. These protein sequences, as well as a set of natural WW domain sequences, were then experimentally expressed and assayed for folding. 67% of the natural sequences were folded, 28% of the sequences designed with coupling information were folded, and none with only conservation information were folded. Furthermore, Russ et al. (2005) assayed these designed domains for binding to proline-containing target peptides and found that sequences designed with SCA also function, exhibiting the same diversity of binding specificity and binding affinity as the natural WW sequences.

These results demonstrate the sufficiency of SCA information to specify the WW domain's fold and function. The goals of this dissertation work are to explore (1) the effect of changing the amount of SCA information used to design WW domains, and (2) the generality of this approach for other protein families. The first goal addresses the fact that the probability of folding for the SCA-designed WW domains was lower than that

for natural proteins, suggesting that some pertinent information may have been missing from the design process and might be specified by adding more coupling information. It also investigates what minimal information content is necessary to specify this protein fold. The second goal is necessary because WW domains are much smaller than most protein folds and lack a substantial hydrophobic core, so this work determines whether coupling information is sufficient to capture the design constraints of a sizable, well-packed protein core. Toward this end, one line of research applies SCA-based design to the PDZ domain family. Also, because SCA was originally developed to identify allosteric networks within proteins, this work addresses the application of SCA-based design to a complex, allosteric protein family: the G-protein coupled receptors.

Methodology of SCA calculations

SCA analysis requires a large and diverse protein sequence alignment that maintains an evolutionarily selected biological activity, but which has diverged sufficiently such that random mutation has eliminated any statistical correlations between residues that do not functionally interact (criteria are described in Suel et al., 2003). Given such an alignment, SCA begins by evaluating the probability of finding the observed amino acid distribution within each column of the sequence alignment, denoted P_i^x for amino acid *x* found in column *i*. This probability is given by the binomial function

$$P_i^x = \frac{N!}{n_i^x!\left(N - n_i^x\right)!}\left(p_x\right)^{n_i^x}\left(1 - p_x\right)^{N - n_i^x}$$

where N is the number of sequences in the alignment, n_i^x is the number of sequences with amino acid *x* in column *i*, and p_x is the probability of observing amino acid *x* in the absence of selective pressure, which is approximated as the frequency observed in the nonredundant protein database. Because the coupling calculations will compare P_i^x between a full sequence alignment and a smaller subalignment, the alignment size *N* for each is normalized to 100 and the number of sequences n_i^x scaled accordingly. The probability can be converted to a statistical energy using the Boltzman distribution

$$\Delta E_i^x = -\gamma^* \ln\left(\frac{P_i^x}{P_{ref}}\right)$$

where γ^* is an analog to temperature in physical systems and P_{ref} is the probability of observing amino acid *x* in a reference state. It will be shown later that the particular reference state chosen will not impact the coupling values.

The interaction between two sites can then be evaluated with a perturbation based approach (Lockless and Ranganathan, 1999) or a global approach (Sharma, 2006). First, the perturbation based approach makes a statistical perturbation to the sequence alignment by restricting it to include only sequences with a particular amino acid *y* at some site *j*. The response to this perturbation is then measured at another position *i* as the difference between ΔE_i^x for the full alignment and $\Delta E_{i,j}^x$ for the restricted alignment in the background of the perturbation, calculated for each amino acid *x* and denoted $\Delta\Delta E_{i,j}^x$. Thus,

$$\Delta\Delta E_{i,j}^{x} = -\gamma^{*} \ln\left(\frac{P_{i,j}^{x}}{P_{ref}}\right) + \gamma^{*} \ln\left(\frac{P_{i}^{x}}{P_{ref}}\right) = -\gamma^{*} \ln\left(\frac{P_{i,j}^{x}}{P_{i}^{x}}\right)$$

and the reference state does not affect the coupling values. To display these values as a two-dimensional matrix, coupling values between each position *i* and perturbation *j* are collapsed into a single statistical energy term over all amino acids:

$$\Delta\Delta E_{i,j} = \sqrt{\sum_{x} \left(\Delta\Delta E_{i,j}^{x}\right)^{2}}$$

The global approach instead uses perturbations independent of the sites being evaluated (Sharma, 2006). A perturbation is made by removing a subset of sequences from the alignment, and the response to this perturbation is measured for both an amino acid *x* at site *i* and another amino acid *y* at site *j*. If one performs many trials of making such perturbations with random sets of sequences, then the change in n_i^x and n_j^y will on average be zero, but for any particular trial they may have some small nonzero change. Furthermore, if these two residues interact, then these small responses to each perturbation trial should be correlated. The coupling between these two residues is then defined as the correlation between these responses, calculated as their mean product.

$$E_{i(x),j(y)} = \left\langle \Delta\Delta E_{i,pert}^{x} \cdot \Delta\Delta E_{j,pert}^{y} \right\rangle$$

This differs from a Pearson correlation coefficient only in that it includes the magnitude of the contributing $\Delta\Delta E$ values instead of implicitly normalizing them. As with perturbation-based coupling values, the global SCA values are compacted into a single term describing the coupling between two sites in order to display them as coupling matrices.

$$E_{i,j} = \sqrt{\sum_{x,y} \left(E_{i(x),j(y)}\right)^2}$$

In practice, rather than making a large number of random perturbations, coupling values are calculated by making subtle perturbations of removing a single sequence from the alignment and measuring the response, and evaluating the values calculated upon removal of each sequence in the alignment. This allows for an efficient, exact calculation of coupling values, and has been shown by Sharma (2006) to produce essentially the same values as making many random perturbations that exclude many sequences. It should be noted that whereas the perturbation-based calculation produces coupling values between each amino acid at each position in the alignment and each perturbation used (total number of terms = 20 x sequence length x number of perturbations), the global approach produces coupling values between all amino acids at all pairs of positions (total number of terms = (20 x sequence length)^2).

Interestingly, the statistical energy ΔE_i^x is related to the Kullback-Leibler divergence used in information theory – a measure of the distance between two probability distributions (Cover and Thomas, 1991), as noted by Olivier Rivoire. The Kullback-Leibler divergence between two probability distributions *p(n)* and *q(n)* is defined as

$$D(p \parallel q) = \sum_n p(n) \log \frac{p(n)}{q(n)}$$

For the purposes of coupling analysis, there are two possibilities for *n*: 0 if a sequence does not have a particular amino acid at a given position, and 1 if it does. Stating the Kullback-Leibler divergence in terms of the observed frequency of an amino acid at a

site, f_i^x, compared to the probability of its appearance in the absence of selective pressure, p_x

$$D\left(f_i^x \parallel p_x\right) = \left(1 - f_i^x\right)\log\frac{1 - f_i^x}{1 - p_x} + f_i^x \log\frac{f_i^x}{p_x}$$

ΔE_i^x can take the same form as derived below. By omitting the P_{ref} term which does not affect coupling values,

$$\Delta E_i^x = -\gamma^* \ln\left(P_i^x\right) = -\gamma^* \ln\left(\frac{N!}{n_i^x!\left(N - n_i^x\right)!}\left(p_x\right)^{n_i^x}\left(1 - p_x\right)^{N - n_i^x}\right)$$

$$\Delta E_i^x = \gamma^*\left[-\ln\left(N!\right) + \ln\left(n_i^x!\right) + \ln\left(\left(N - n_i^x\right)!\right) - \ln\left(p_x\right)^{n_i^x} - \ln\left(1 - p_x\right)^{N - n_i^x}\right]$$

If the number of sequences in the alignment and the number of sequences sampled by the statistical perturbation are both large, then using Sterling's approximation

$$\ln\left(x!\right) \cong x\ln x - x$$

allows the equation to be rewritten

$$\Delta E_i^x = \gamma^*\left[-N\ln N + N + n_i^x \ln n_i^x - n_i^x + \left(N - n_i^x\right)\ln\left(N - n_i^x\right) - \left(N - n_i^x\right) - n_i^x \ln\left(p_x\right) - \left(N - n_i^x\right)\ln\left(1 - p_x\right)\right]$$

$$\Delta E_i^x = \gamma^*\left[n_i^x \ln\frac{n_i^x}{p_x} + \left(N - n_i^x\right)\ln\frac{\left(N - n_i^x\right)}{\left(1 - p_x\right)} - N\ln N\right]$$

Converting the terms n_i^x from a number of sequences with the indicated amino acid to a frequency of amino acids f_i^x ranging from 0 to 1 gives

$$\Delta E_i^x = \gamma^* \left[N f_i^x \ln \frac{N f_i^x}{p_x} + N\left(1 - f_i^x\right) \ln \frac{N\left(1 - f_i^x\right)}{\left(1 - p_x\right)} - N \ln N \right]$$

$$\Delta E_i^x = N\gamma^* \left[f_i^x \ln \frac{f_i^x}{p_x} + f_i^x \ln N + \left(1 - f_i^x\right) \ln \frac{\left(1 - f_i^x\right)}{\left(1 - p_x\right)} + \left(1 - f_i^x\right) \ln N - \ln N \right]$$

$$\Delta E_i^x = N\gamma^* \left[f_i^x \ln \frac{f_i^x}{p_x} + \left(1 - f_i^x\right) \ln \frac{\left(1 - f_i^x\right)}{\left(1 - p_x\right)} \right]$$

So the final form of the equation is simply

$$\Delta E_i^x = \frac{N\gamma^*}{\log e} \left[f_i^x \log \frac{f_i^x}{p_x} + \left(1 - f_i^x\right) \log \frac{\left(1 - f_i^x\right)}{\left(1 - p_x\right)} \right] = \frac{N\gamma^*}{\log e} D\left(f_i^x \parallel p_x\right)$$

Because the alignment size *N* is normalized in the course of coupling calculations, ΔE_i^x and $D\left(f_i^x \parallel p_x\right)$ are proportional.

WW DOMAINS

The first of my design targets is the WW domain. This is an experimentally tractable, small (~36 residue) protein fold that includes a highly conserved buried tryptophan that can be used as a fluorescent reporter of folding. Because it was used in the design project of Socolich et al. (2005), it was an established model system in this laboratory, and it was known that SCA information could specify this fold.

Biology and structure

The WW domain was first identified as a module in YAP65 (Sudol et al., 1995) – a protein that binds to the Yes proto-oncogene product – and is found in a variety of other

signaling proteins including Nedd-4 (a ubiquitin ligase), ORF-1 (a Ras GTPase activator-related protein), utrophin (a dystrophin related protein), and FE65 (a transcriptional regulator), with a species distribution from humans to yeast (RSP5) (Bork and Sudol, 1994). The WW domain of YAP65 was characterized as a protein binding module with a target motif of xPPxY, similar to the functional role of the SH3 domains but with a different specificity than the PxxP motif of SH3 targets and without cross-binding of these two domains with each other's binding partners (Chen and Sudol, 1995). Other WW domains show distinct binding preferences; the WW domains of Pin1 and Nedd4, for example, bind specifically to the phosphoserine or phosphothreonine modified forms of their target proteins (Lu et al., 1999). Binding specificity within the WW family is classified into four groups: group I which recognizes PPxY containing targets, group II which binds PPLP, group III specific for PPR, and group IV targeting phospho-S/T flanked by proline (Sudol and Hunter, 2000).

The first structure of a WW domain shows YAP65 bound to a target PPPY containing peptide (Macias et al., 1996, see figure 1.1). The overall structure of this fold is a three stranded, antiparallel beta sheet. The peptide ligand lies across a hydrophobic concavity of the sheet, with its prolines packing against a conserved tryptophan of the WW domain. Other group I domains – dystrophin (Huang et al., 2000) and Nedd4 (Kanelis et al., 2001) – have essentially the same structure and mode of peptide binding. The group IV WW domain of Pin1 (Ranganathan et al., 1997, Verdecia et al., 2000) and the group II domain of FBP11 (Kato et al., 2006) both adopt a similar fold and mechanism of binding as group I domains, although their target peptides bind in the opposite N to C orientation as seen in group I domains. The overall similarity of the three

dimensional fold of these proteins, along with their conserved mode of binding (albeit with variable directionality of the target), makes these attractive proteins for SCA-based design – the WW domain family members likely share the same evolutionary constraints on folding and function.

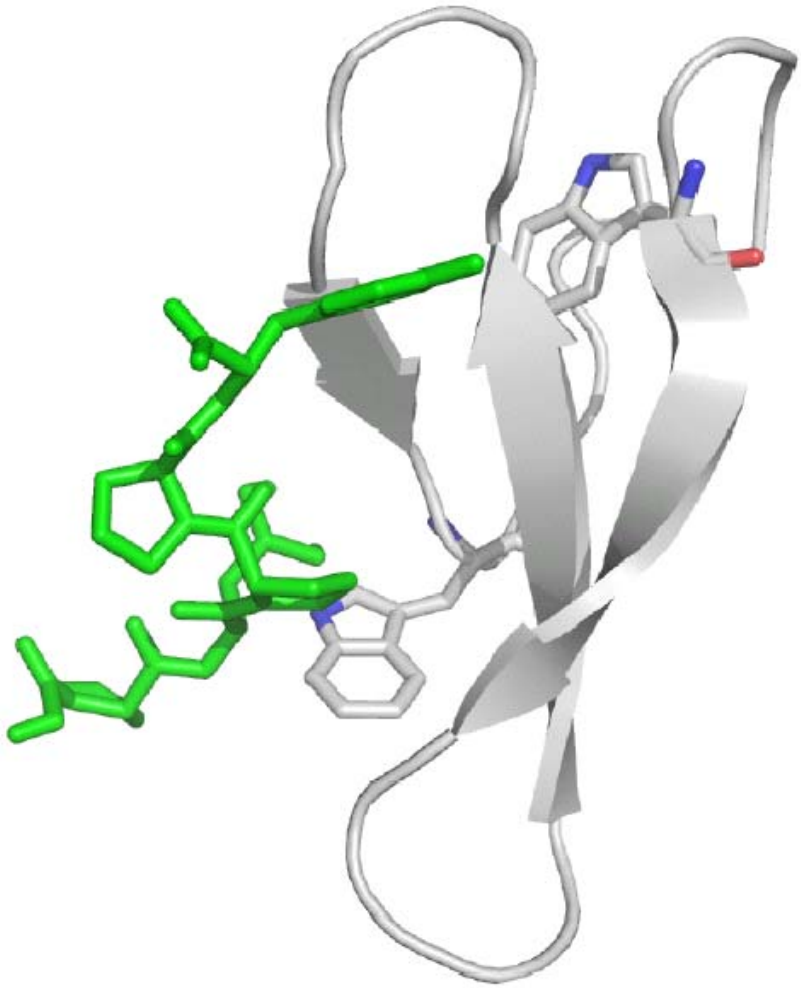


Figure 1.1: Structure of a WW domain
The WW domain of YAP65 is shown with its target ligand in green (Pires et al., 2001). The first (upper) tryptophan served as a fluorescent reporter for folding in Socolich et al. (2005) and in this dissertation work, and the second (lower) tryptophan contributes to peptide binding.

WW domain design

Because it is the smallest naturally occurring protein domain that spontaneously folds in the absence of disulfide formation or strong ionic interactions, the WW domain has received particular attention from the protein design community. The first reported effort at WW domain design is from Macias et al. (2000). Rather than a test of a specific design algorithm, this was a manual sequence design effort based on (and meant as a demonstration of) the body of knowledge gained about this domain. Much of the prototype sequence that they designed was generated by selecting the most commonly observed residue in the alignment. The other residues were manually selected and aimed toward increasing the contacts between the first and third beta strands, increasing solubility with polar residues at exposed positions, and introducing residues at the termini. While the designed WW domain does fold, this study was a test of an expert's knowledge, not of a design principle.

Kraemer-Pecore et al. (2003) have used a more automated approach to WW domain design. Their study used a design approach based on energy minimization, with the energy function consisting of an Amber/OPLS force field, added solvation potentials, and a term for amino acid baseline corrections used to maintain reasonable compositions in the designed sequences (Raha et al., 2000). The novelty in this particular design experiment came from the use of an ensemble of backbone structures in place of a single fixed conformation. The authors designed two sequences with this approach, and found that one of these sequences exhibits folded characteristics at 5 degrees. In particular, alpha proton resonances downfield of 4.7 ppm indicated that this likely has a beta strand

structure, and NOEs between amide and alpha protons indicate that they are arranged in the antiparallel orientation of a WW domain, although slow conformational exchange prevented a full structure determination. Unlike the SCA-based design of WW domains discussed earlier, this work focused on protein folding rather than function, and did not report whether the designed protein binds to any target sequences.

PDZ DOMAINS

My second design target is the PDZ domain family. This was chosen to serve as a larger protein fold that would test the ability of SCA-based design to specify a sizable, well-packed protein core. Because this was the first system with which SCA was developed and validated (Lockless and Ranganathan, 1999), the quality of this sequence alignment and the relevance of the resulting coupling analysis have been established.

Biology and structure

Named after PSD-95, Discs-large, and ZO-1, PDZ domains are 90-100 amino acid proteins whose primary known function is to bind the C-termini of target proteins. Their phylogenetic distribution stretches to *E coli*, but they are much more common in multicellular than unicellular organisms (Ponting, 1997). Although PDZ domains themselves do not span the membrane, the vast majority of PDZ containing proteins are associated with the plasma membrane to scaffold protein complexes (Fanning and Anderson, 1999). Scaffolding proteins often contain multiple different PDZ domains; for

example, InaD contains 5 PDZ domains involved in scaffolding various components of the phototransduction cascade in complexes at the membranes of *Drosophila* photoreceptor cells (Ranganathan and Ross, 1997).

The classification system for PDZ domain binding specificity has been pioneered by Songyang et al. (1997). Using an oriented peptide library screen, they characterized the site-specific binding preferences for each of several PDZ domains. Most PDZ domains show specificity for the four or so most C-terminal residues of their binding targets, with binding specificity extending to nine residues for a few PDZ domains. They found two major specificity classes: the first described as preferring E-S/T-X-V/I at the C-terminus, and the second binding to peptides with hydrophobic and aromatic residues at the last three positions. The classification system has since been expanded and refined to the following three classes: class I, X-S/T-X-Φ; class II, X-Φ-X-Φ; and class three, X-D/E/K/R-X-Φ (Jelen et al., 2003). In addition to these classifications based on the assumption that the PDZ domain recognizes the target protein's C-terminus, other binding modes clearly exist: phage display library screens have shown that some PDZ domains are capable of binding to internal sequences (Fuh et al., 2000).

The structure of the third PDZ domain from PSD-95 has been solved with and without bound ligand (figure 1.2, Doyle et al., 1996). The overall architecture of the fold is a six stranded beta sandwich with two alpha helices, and remains essentially unchanged in the presence and absence of peptide. The peptide is bound between the β2 strand and α2 helix, hydrogen bonding with and adding a strand to the beta sheet. The β1-β2 loop, also called the carboxylate binding loop, contributes to peptide binding as well. This interaction is of particular interest because residues in the carboxylate binding loop are

highly statistically coupled to histidine 76 (Lockless and Ranganathan, 1999), a major determinant of target specificity (Songyang et al., 1997). While a coupling analysis itself does not reveal a physical mechanism for the observed coupling, Sharma (2006) experimentally investigated this coupling interaction and found that the carboxylate binding loop is flexible in the peptide-free state but clamps onto the ligand in the bound state, and that a mutation locking the carboxylate binding loop into the closed conformation in the absence of peptide results in increased binding affinity by reducing the entropic cost of binding.

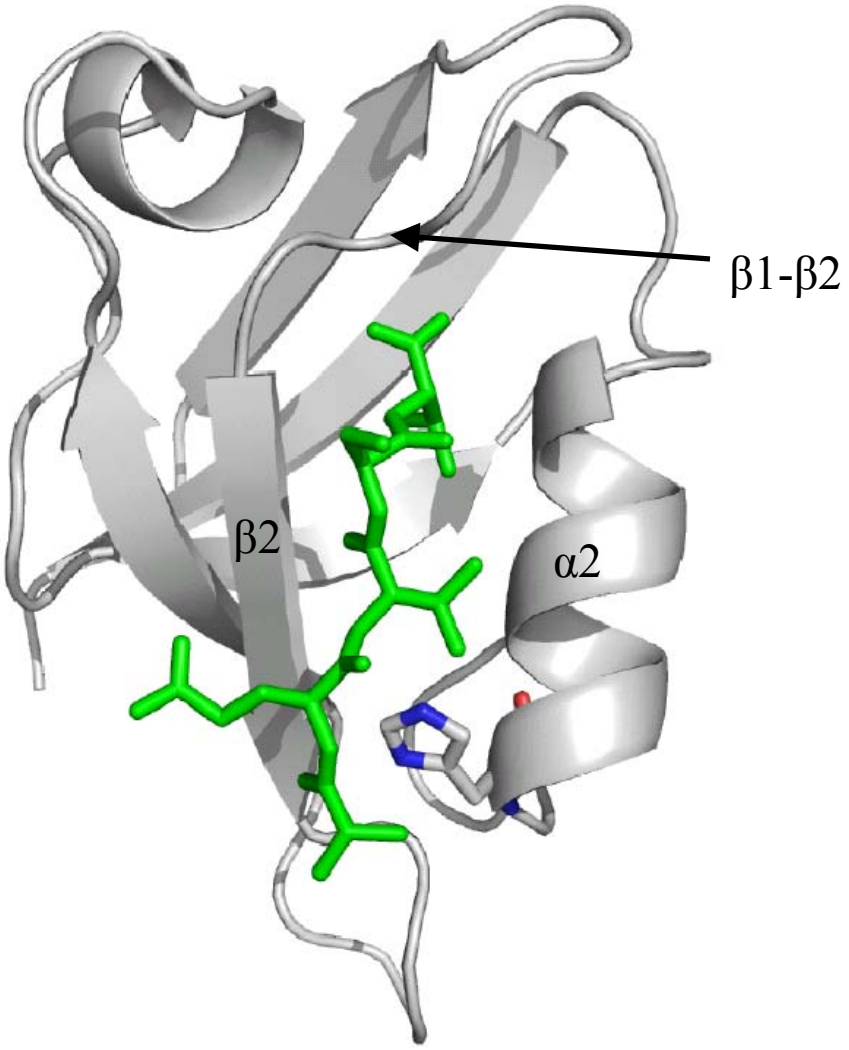


Figure 1.2: Structure of a PDZ domain
The third PDZ domain of PSD-95 is shown with its target peptide in green (Doyle et al., 1996). Histidine 76 on the α2 helix, the primary determinant of binding class specificity, is shown in sticks. The α2 helix, β2 strand, and β1-β2 loop involved in peptide binding are indicated.

Although Reina et al. (2002) have computationally redesigned a PDZ domain's binding pocket to modulate its target specificity, no designs of an entire PDZ domain have been described. However, Dantas et al. (2003) have successfully used RosettaDesign to design several target proteins of similar size, also with mixed alpha/beta folds. Unless there are unappreciated complexities of the PDZ domain in particular, it would not be surprising if current energy minimization based methodologies could be extended to this protein family.

G-PROTEIN COUPLED RECEPTORS AND RH1

My final design target is Rh1: a member of the G-protein coupled receptor family. This is an extreme test of the design algorithm; however, this system offers several unique experimental advantages. Rh1 is highly expressed in fly photoreceptors and the phototransduction activity of a single molecule can be detected electrophysiologically. In addition, Rh1 carries out other orthogonal functions that may be experimentally assayed, which will be described in detail in later sections.

G-protein coupled receptors

The G protein-coupled receptors (GPCRs) constitute a large and diverse protein superfamily of sensors responsive to a multitude of stimuli, from small molecules to glycoproteins, including stimuli not generally classified as "ligands" such as photoisomerization and their own partial proteolysis (Gether, 2000, Karnik et al., 2003).

The defining feature of these family members is their stimulus-dependent ability to catalyze the exchange of GTP for GDP on cognate G proteins, triggering dissociation of GTP-bound G_{α} subunits from $G_{\beta\gamma}$ complexes to begin signaling cascades (Gilman, 1987). In addition, agonist mediated desensitization by G protein-coupled receptor kinases and arrestins has been described for many GPCRs. Interaction with arrestins not only deactivates receptors, but also allows arrestin to initiate endocytosis of the activated receptor into clathrin-coated vesicles (Krupnick and Benovic, 1998). Endocytosis has been implicated as a mechanism through which GPCRs can activate noncanonical signaling avenues, feeding into pathways such as the MAP kinase cascade (Daaka et al., 1998).

Although GPCRs as a whole do not share any overall sequence homology, several subfamilies do appear to be related (Kolakowski, 1994). The class A GPCR family has rhodopsin as its most thoroughly studied member, and includes the adrenergic, prostanoid, chemokine, olfactory and several other subfamilies. This family has several conserved features, including an S-S bond linking the third transmembrane domain to the second extracellular loop, a (D/E)RY motif in the third and an (N/D)PxxY motif in the seventh transmembrane helices, and absence of an extended N-terminal extension found in most other GPCRs (Karnik et al., 2003). The crystal structure of bovine rhodopsin has been determined, making it a valuable model of GPCR activity (figure 1.3, Palczewski et al., 2000). The structure and much biochemical data suggests that chromophore isomerization leads to displacement of the third transmembrane helix, as well as disruption of the interactions between the kinked region of the sixth helix and the other transmembrane domains, culminating in rearrangement of the helices into a conformation

capable of activating transducin. A structure of activated receptor would of course provide much more detailed and definitive evidence as to the mechanism of G protein activation, and activity-dependent recognition by arrestins, phosphatases, and kinases.

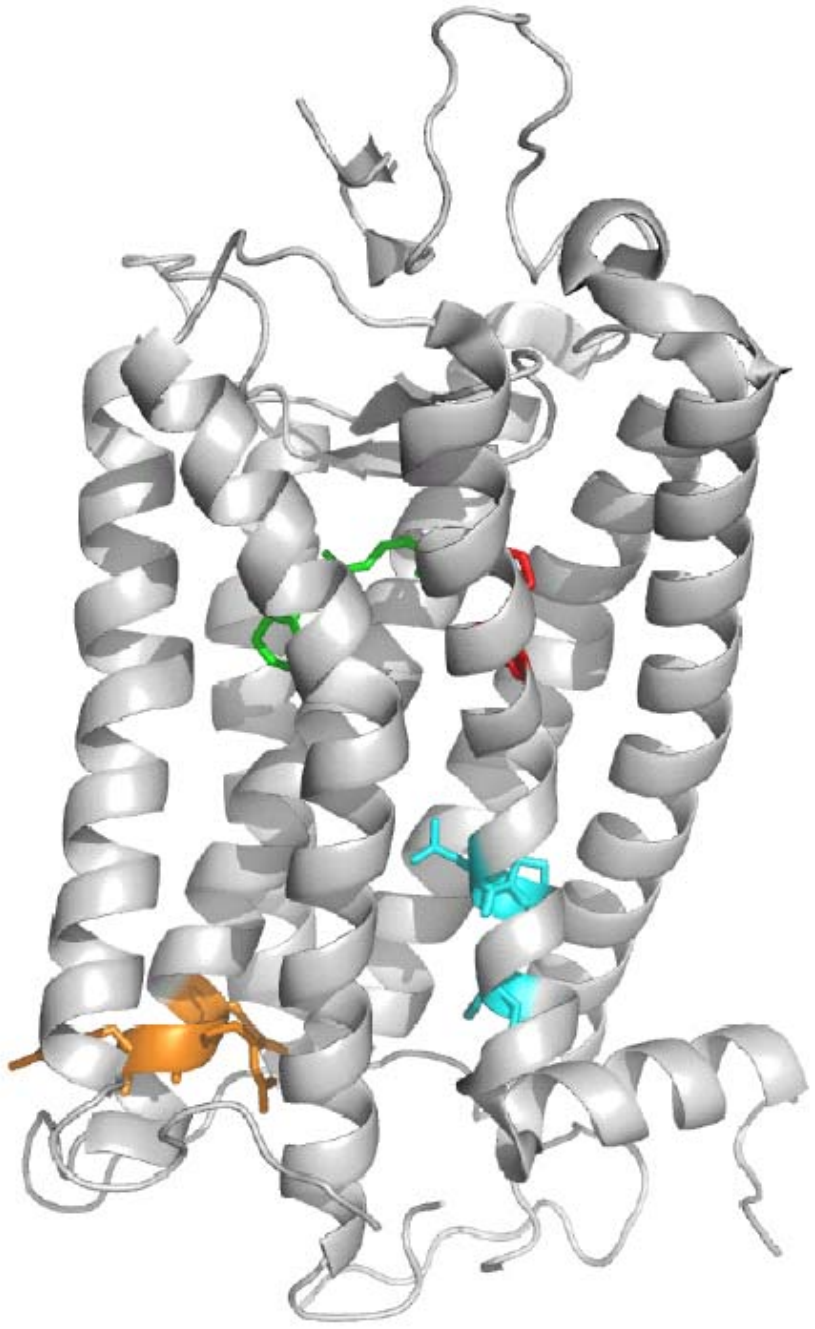


Figure 1.3: Structure of bovine rhodopsin
Rhodopsin is shown with its chromophore retinal in green, the covalently attached lysine in red, and the (D/E)RY and (N/D)PxxY motifs in orange and cyan respectively. (Palczewski et al., 2000)

Drosophila Rh1

The *Drosophila* compound eye consists of 800 ommatidia, or unit eyes, which each contain eight photoreceptor cells and other supporting cells. The six outer photoreceptor cells (R1-R6) are functionally equivalent and express the blue-light sensitive rhodopsin Rh1 encoded by the *ninaE* locus (O'Tousa et al., 1985, Zuker et al., 1985). The two central photoreceptor cells express other opsins with different spectral sensitivities (Feiler et al., 1992, Salcedo et al., 1999). The photoreceptor cells each contain a specialized organelle called the rhabdomere, a bundle of ~60,000 microvillar projections that provides a large membrane surface area to house the phototransduction machinery (Zuker, 1996).

Rh1 is synthesized at the endoplasmic reticulum where it is transiently glycosylated before reaching a mature, deglycosylated state on transport to the rhabdomere (Huber et al., 1990). Glycosylation plays an important role in maturation, as an N to I mutation of the glycosylation site drastically reduces protein expression levels (Katanosaka et al., 1998, Webel et al., 2000). The cyclophilin homolog NinaA is a chaperone required for efficient maturation. In the absence of NinaA, Rh1 levels are greatly reduced and the remaining protein is predominantly retained in the ER in its glycosylated state (Stamnes et al., 1991, Colley et al., 1991). Rh1 covalently binds its chromophore 3-hydroxyretinal as a required step during maturation; in the absence of retinoids, the protein fails to mature and is largely degraded in the ER (Ozaki et al., 1993).

Ground state rhodopsin is converted to the active metarhodopsin state by absorption of a photon of blue light, isomerizing the chromophore from 11-*cis* to all-*trans* (Ostroy et al., 1974). This triggers activation of a G_q protein (Lee et al., 1990) and subsequent signaling culminating with the opening of cation channels to generate a light response (Hardie and Raghu, 2001). Activated metarhodopsin does not bleach by releasing its chromophore like vertebrate rhodopsins, but stays in the metarhodopsin state until a photon of red light drives re-isomerization back to the ground state (Hardie, 1986). Unless it is reisomerized, metarhodopsin is phosphorylated and bound by arrestin to form stable complexes that no longer activate G protein (Byk et al., 1993, Kiselev and Subramaniam, 1994). These metarhodopsin-arrestin complexes are internalized and, unless photoconverted back to the ground state, will trigger photoreceptor apoptosis (Alloway et al., 2000, Kiselev et al., 2000).

In addition to its well-known activity as a photoreceptor, Rh1 is also necessary for normal development of the fly eye. In the absence of Rh1, the rhabdomeres degenerate at the end of pupal development and are severely abnormal in adults (Kumar and Ready, 1995). This is likely due to a signaling activity rather than mass action of protein: a single pulse of Rh1 expression in the correct developmental window can permanently rescue the rhabdomeres (Kumar et al., 1997), and degeneration can be prevented in the absence of Rh1 by activated *Drosophila* rac (Chang and Ready, 2000). Rhabdomere rescue is orthogonal to Rh1's phototransduction activity: mutants of the cognate Gq and phospholipase C have normal rhabdomeres (Kosloff et al., 2003), as do flies raised in total darkness. However, no immediate effector molecules directly interacting with Rh1 to preserve the rhabdomeres have been identified.

CHAPTER TWO
SCA information required to specify the WW domain

INTRODUCTION

The SCA-based sequence design algorithm described in Socolich et al. (2005) is able to design WW domains, but is not efficient enough to allow for the design of larger proteins. One of the goals of this project is to increase the computational efficiency of the design algorithm to reliably converge for larger sequence alignments. Also, the original design algorithm was written before the global SCA approach was developed, so all of its SCA information is based on site-specific statistical perturbations. The global SCA calculations produce a larger matrix of coupling values (one term for each pair of amino acids at all pairs of columns in the alignment) than perturbation-based calculations (one term for each amino acid in each column coupled to a limited number of statistically allowable perturbations). The larger matrix might in principle carry more evolutionarily relevant information, so a second goal of this project is to develop an algorithm that incorporates globally calculated coupling values. Finally, the ultimate goal of this project is to use these new algorithms to vary the amount of SCA information imposed in the design process, and see how the information content used in the design process affects the probability of folding for the designed sequences.

SCA-based sequence design

The SCA-based sequence design algorithms used in this dissertation are conceptually identical to that described in Socolich et al. (2005). One of the algorithms

used here incorporates information from perturbation-based SCA to design sequences, and the other uses global SCA. The algorithms accept a sequence alignment as input, and in the case of perturbation-based SCA a list of statistical perturbations as well. Both algorithms operate by taking the natural sequence alignment and making a series of random swaps of two residues within the same column of the alignment. Such swaps will not change the conservation pattern of the alignment, but may change the coupling pattern between the residues being swapped and residues at all other positions in the alignment. The degree to which the coupling pattern of the alignment being designed has deviated from the natural alignment's coupling pattern is quantified with an energy term, defined as the difference between the natural and designed alignments' coupling patterns. For perturbation-based SCA

$$Energy = \sum_{i,j,x} \left| \Delta\Delta E^{x}_{i,j}(natural) - \Delta\Delta E^{x}_{i,j}(designed) \right|$$

and for global SCA

$$Energy = \sum_{i,j,x,y} \left| E_{i(x),j(y)}(natural) - E_{i(x),j(y)}(designed) \right|$$

Making many such random swaps sequentially eventually destroys the coupling pattern of the designed alignment, making it very different from the natural alignment and producing a large energy value. In order to rebuild the coupling pattern, the algorithm uses a simulated annealing protocol. Before a swap is made, the change in energy that would result from the swap (*ΔEnergy*) is calculated. The swap is then either accepted or rejected with a probability based on Metropolis Monte Carlo criteria.

$$P(accept) = e^{-\Delta Energy / Temperature}$$

The "temperature" in this equation is a parameter that is initially set to a large value (large enough that swaps are accepted essentially regardless of their effect on energy) and gradually decreases over the course of the design. The effect of this equation is that all swaps that produce an energy change of zero or less (either rebuilding the natural coupling pattern or having no net effect) will be accepted. Swaps that increase the energy will be accepted with a probability that decreases as the energy change becomes large and as the temperature decreases. As the temperature decreases, this forces the energy to a low value (meaning that the coupling patterns of the natural and designed alignments become similar) while allowing occasional energetically unfavorable swaps in order to escape from local minima.

One difference between the design algorithms used in this work and that described in Socolich et al. (2005) is that one of the two design algorithms described here converges on information from a global SCA while the previous algorithm uses perturbation-based SCA. Also, both algorithms described here converge on precise statistical coupling values as described above, whereas the previous algorithm converges on a simplified function approximating SCA in order to achieve computational efficiency. And of practical importance, both new algorithms are much more computationally efficient than the original algorithm, making it possible to apply them to alignments of all sizes tested in a reasonable period of time. This efficiency is achieved primarily by pre-calculating a table of SCA values that are then accessed in each swap rather than calculating the needed values each time a swap is attempted.

RESULTS

Performance of the design algorithm

The criteria that I sought to meet with the new perturbation-based design algorithm are that (1) it initially destroys the natural alignment's coupling pattern, (2) it rebuilds an alignment using only SCA (and site specific conservation) information and successfully incorporates it into the alignment, (3) it does so in a reasonable amount of time, and (4) it performs reliably. I first tested the design algorithm with an alignment of WW domains containing 247 sequences and 37 columns, using 14 statistical perturbations. The progress of design runs for the WW domains are shown in figure 2.1. At the beginning of a design run, the system's energy initially rises, indicating loss of the original coupling pattern. The magnitude of this energy value is equal to that obtained by generating an alignment *de novo* using only site-specific conservation. As the temperature decreases, the energy first fluctuates about its initial high value. During this phase, the probability of accepting each attempted swap is near unity, indicating there is not yet sufficient selection of energetically favorable swaps to drive the simulated annealing process. As the algorithm progresses, the energy eventually begins to decrease, indicating that the coupling pattern is returning to that of the natural sequence alignment. Late in the run, the energy reaches another, lower plateau. In this second plateau phase, the probability of accepting each attempted swap is near zero, indicating that the algorithm has settled into an energy minimum. The fact that the energy is at a lower value indicates that the coupling pattern has become more like that of the natural alignment.

Comparison of the coupling patterns of the final designed alignment to the original natural sequence alignment shows that they are nearly identical, confirming that most of the coupling information has been successfully reincorporated. To evaluate the reliability of the design algorithm, 20 independent runs were performed with the WW domain alignment using different random seeds. Each run follows the same trajectory of energy versus temperature, and they converge to essentially the same final value. Finally, each run took about fifteen minutes on a 667 MHz alpha processor with parameters that led to acceptable convergence, which is significantly faster than the original algorithm.

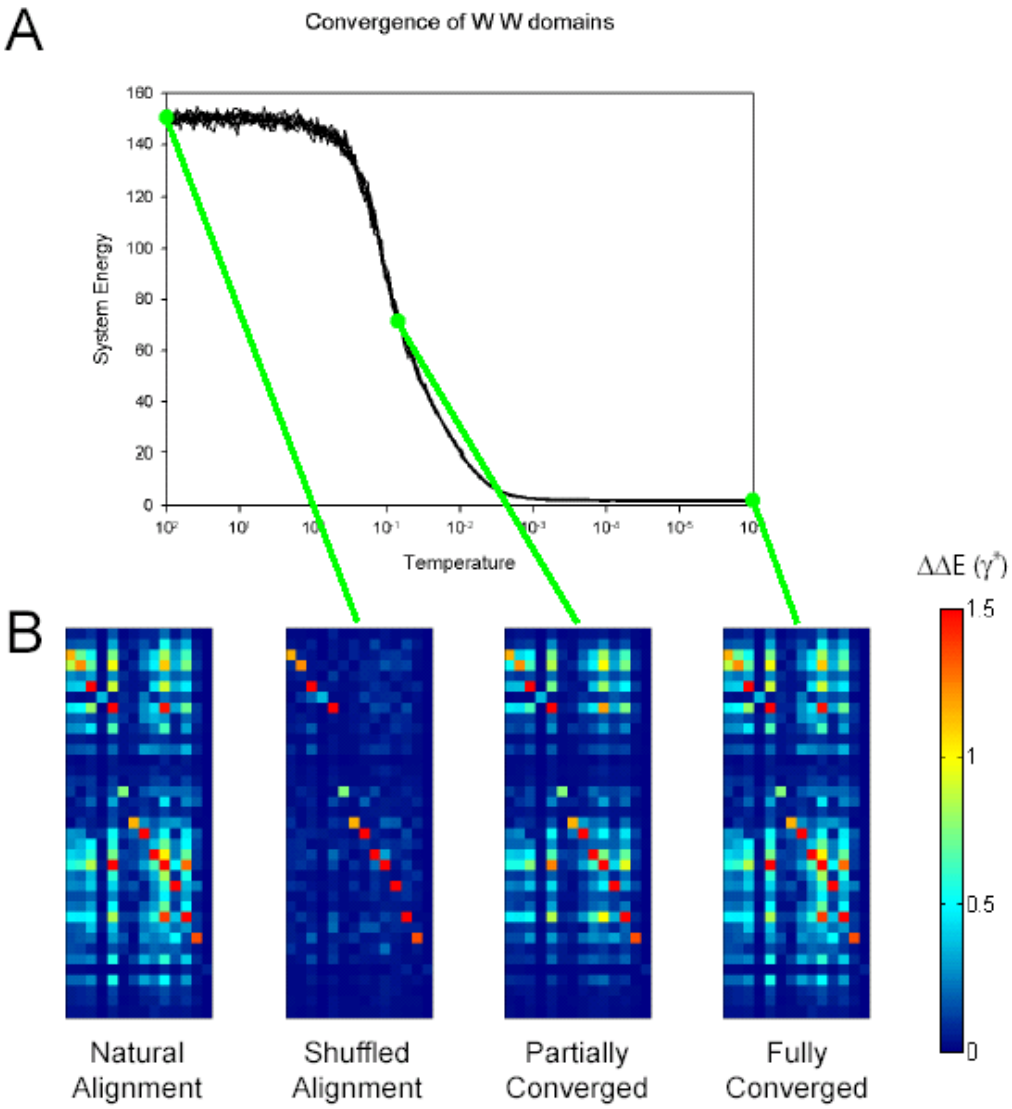


Figure 2.1: WW domain design with perturbation-based SCA

A) System energy versus temperature as the algorithm converges with the WW domain alignment. Initially, high temperatures allow almost all swaps to be made, destroying the coupling pattern and bringing the system to a high energy. As the temperature decreases, the algorithm converges to a low energy state, indicating that it has reconstructed the natural coupling pattern. The overlay of twenty trials shows that the algorithm converges reliably.

B) Comparison of the SCA patterns of the natural alignment and the redesigned alignment at various points on the convergence trajectory of one run. The SCA information is displayed as a matrix, with each cell showing ΔΔE in units of γ^* between the indicated perturbation (column) and position (row). The coupling pattern is initially lost, but gradually returns to that of the natural alignment as the algorithm converges.

The algorithm also meets these performance criteria with larger sequence alignments (figure 2.2). For alignments of PDZ domains, SH2 domains, globins, and G-protein coupled receptors (GPCRs), the algorithm shows the same initial rise to and plateau at a high energy equal to that of a shuffled alignment with a probability of accepting swaps that is near one. This is followed by a gradual decrease in energy to a lower final value, with a final probability of accepting swaps that is near zero. The coupling patterns of the designed alignments are in all cases nearly identical to that of the starting natural alignment. Even for the large GPCR alignment, a design run is completed in approximately two weeks, making this practical for sequence design. Furthermore, ten runs with the GPCR alignment show essentially the same pattern of convergence of energy as a function of temperature, indicating that convergence is reliable, while the sequences designed in separate runs are not similar: the mean sequence identity between each designed sequence and its most similar sequence from the natural alignment is 47%, while the mean sequence identity between each designed sequence and its most similar sequence from a separate design run is 44% for all pairs of runs.

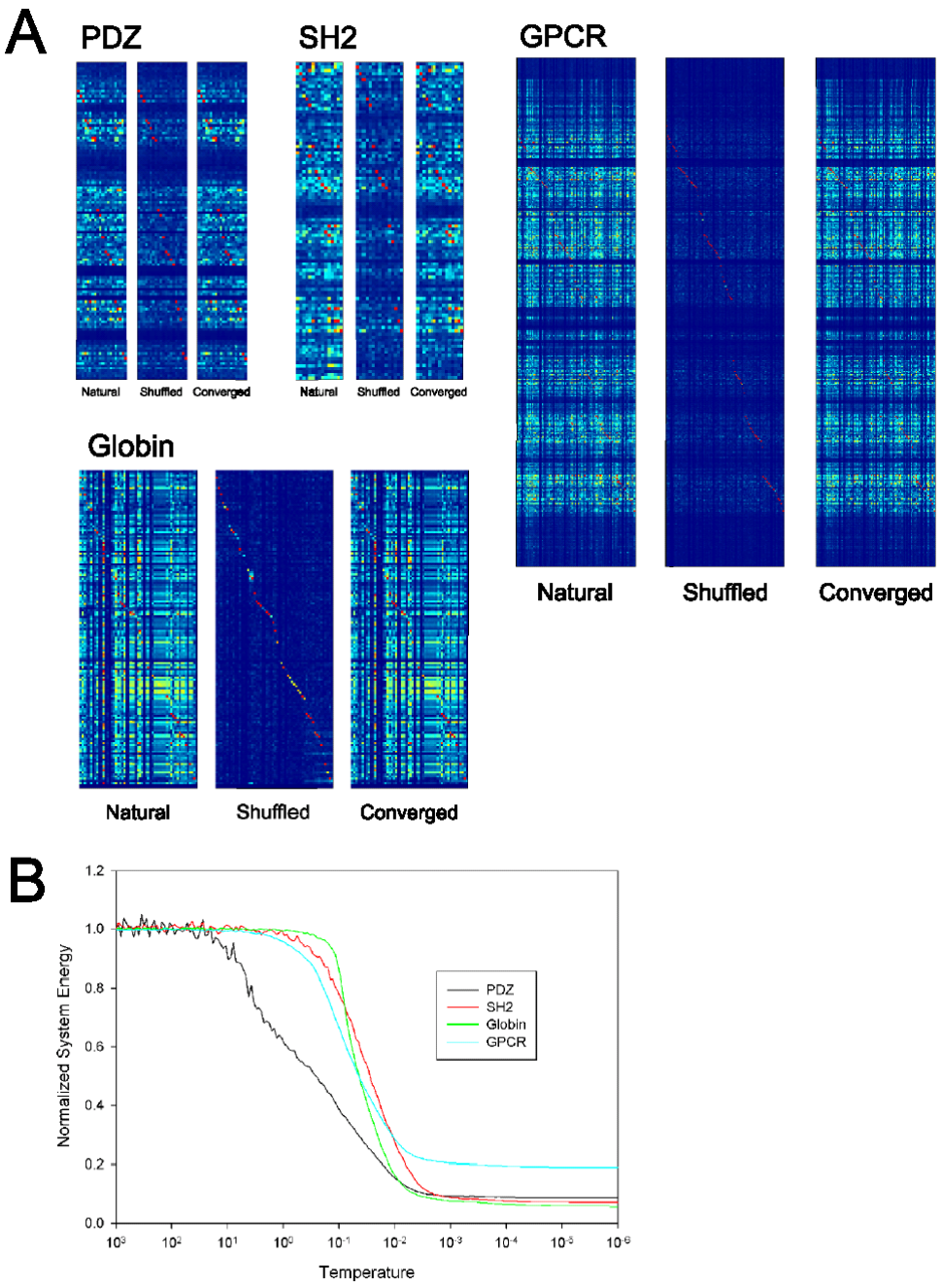


Figure 2.2: SCA-based design with several protein families

A) SCA patterns of the natural alignment, the alignment after it is initially shuffled by the algorithm, and the redesigned alignment after convergence for several protein families.

B) Energy traces for each of the protein families. In each case, the energy initially plateaus when the alignment is shuffled and decreases as the algorithm converges near the target SCA pattern.

Alignment	Size (sequences x positions)	Statistical perturbations	Mean ± σ percent identity to most similar natural sequence	
			Shuffled	Converged
WW	247 x 37	14	50% ± 5%	59% ± 9%
PDZ	233 x 128	20	36% ± 3%	47% ± 9%
SH2	316 x 100	15	45% ± 3%	51% ± 6%
Globin	859 x 151	56	60% ± 3%	79% ± 9%
GPCR	940 x 429	100	28% ± 2%	47% ± 14%

Table 2.1: Characteristics of the designed alignments

Increasing constraints on WW domain design

The original design experiment (Socolich et al., 2005) used information from a perturbation-based SCA of an alignment of 120 WW domain sequences using 5 perturbations. This provided sufficient information to specify sequences with a 28% probability of folding. While it successfully specified folded proteins, the observed probability of folding for the designed sequences is lower than the 67% observed for sequences from the natural alignment. The probability of folding may be lower because the information used did not completely capture all of the design constraints for this protein family, or could be due to errors in the protein sequence alignment, or errors in estimating the amino acid frequencies at sites given the limited number of sequences in the alignment and even smaller numbers of sequences sampled by statistical perturbations.

To address these possibilities, WW domains were designed using the perturbation-based design algorithm described here with more statistical perturbations and an alignment with more sequences (247 sequences here versus 120 in the original experiment). While increasing the number of statistical perturbations would increase the information content of the analysis, if any of the statistical perturbations lead to only a small number of sequences in the perturbed set, then undersampling the distribution of sequences consistent with that perturbation could lead to large errors in the coupling values. This has been addressed by Suel et al. (2003). To estimate the error introduced by undersampling with a perturbation containing *n* sequences, coupling values were calculated for many trials of randomly selecting *n* sequences from the alignment to serve as perturbations. These random "perturbations" would not actually be coupled to any residues, so the resulting coupling values are an estimate of the effect attributable purely to undersampling. A plot of coupling values arising from such random perturbations versus the number of sequences in the random perturbations facilitates determination of a minimal number of sequences that would avoid introducing large undersampling errors. In the case of the WW domain alignment, perturbations that included half of the alignment produced little error (figure 2.3), and fourteen statistical perturbations included this number of sequences (see Methods).

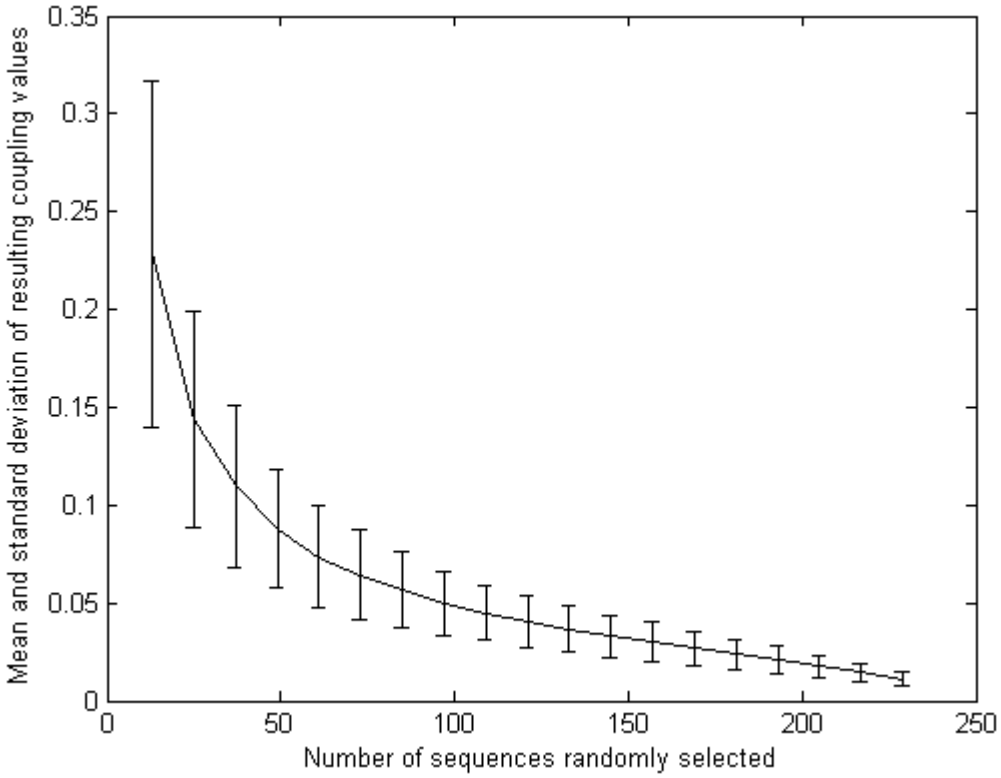


Figure 2.3: Estimated error of undersampling for the WW domain alignment
Subalignments with varying numbers of sequences were generated by randomly selecting a subset of sequences from the alignment, and coupling values for the resulting subalignments were calculated. Because these coupling values represent coupling to random perturbations, they provide an estimate of the effect of undersampling for perturbations that include small numbers of sequences.

SCA information from these fourteen perturbations against an alignment of 247 WW domain sequences were used in the perturbation-based design algorithm described here. Genes encoding the designed sequences were cloned and the protein expressed following the procedures described in Socolich et al. (2005) to compare with the previous set of sequences designed using SCA of five perturbations against an alignment of 120 sequences with the original design algorithm. Because Socolich et al. found that monitoring tryptophan fluorescence during thermal denaturation and ^{1}H NMR measurements were in good agreement, only tryptophan fluorescence was used to evaluate the proteins designed in this study.

Figure 2.4 shows example tryptophan fluorescence melts for two designed proteins. Panel A shows a well-folded protein, with a cooperative transition from a folded state at low temperature where a tryptophan buried in the core of the protein has high fluorescence in this hydrophobic environment, to an unfolded state at high temperature where the tryptophan fluorescence is quenched due to solvent exposure. Panel B shows an unfolded protein, which does not show such a transition between folded and unfolded states. Domains were scored as folded or unfolded based on the presence or absence of a cooperative transition in tryptophan fluorescence on heating and recooling, and thermodynamic parameters for the folded domains were determined by fitting the derivative of these melts to the van't Hoff equation (John and Weeks, 2000). Of 45 redesigned domains, 15 show cooperative denaturation and refolding (figure 2.5). Comparison of the melting temperature and the free enthalpy of unfolding for WW domains designed by the original design algorithm using 5 perturbations and the revised algorithm using 14 perturbations shows no difference between the two (figure 2.6).

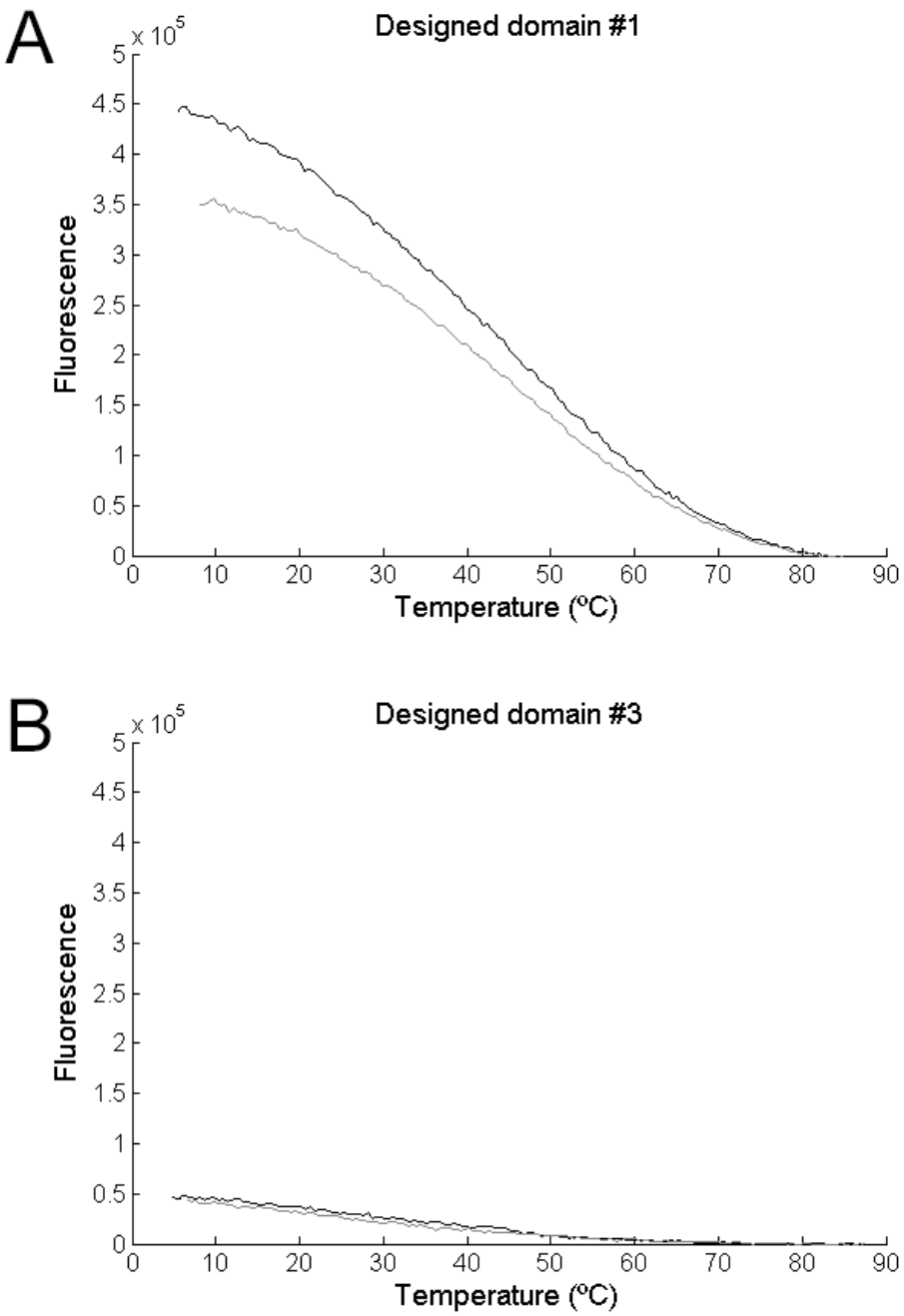


Figure 2.4: Fluorescence traces of folded and unfolded WW domains
Plots of fluorescence for a folded (A) and unfolded (B) domain as they are heated to 90 degrees (black) and recooled (gray), after subtracting the temperature dependence of fluorescence for free tryptophan. The folded domain shows a cooperative transition between the folded state with high fluorescence and the unfolded state where fluorescence is quenched, while the unfolded domain shows no cooperative transition.

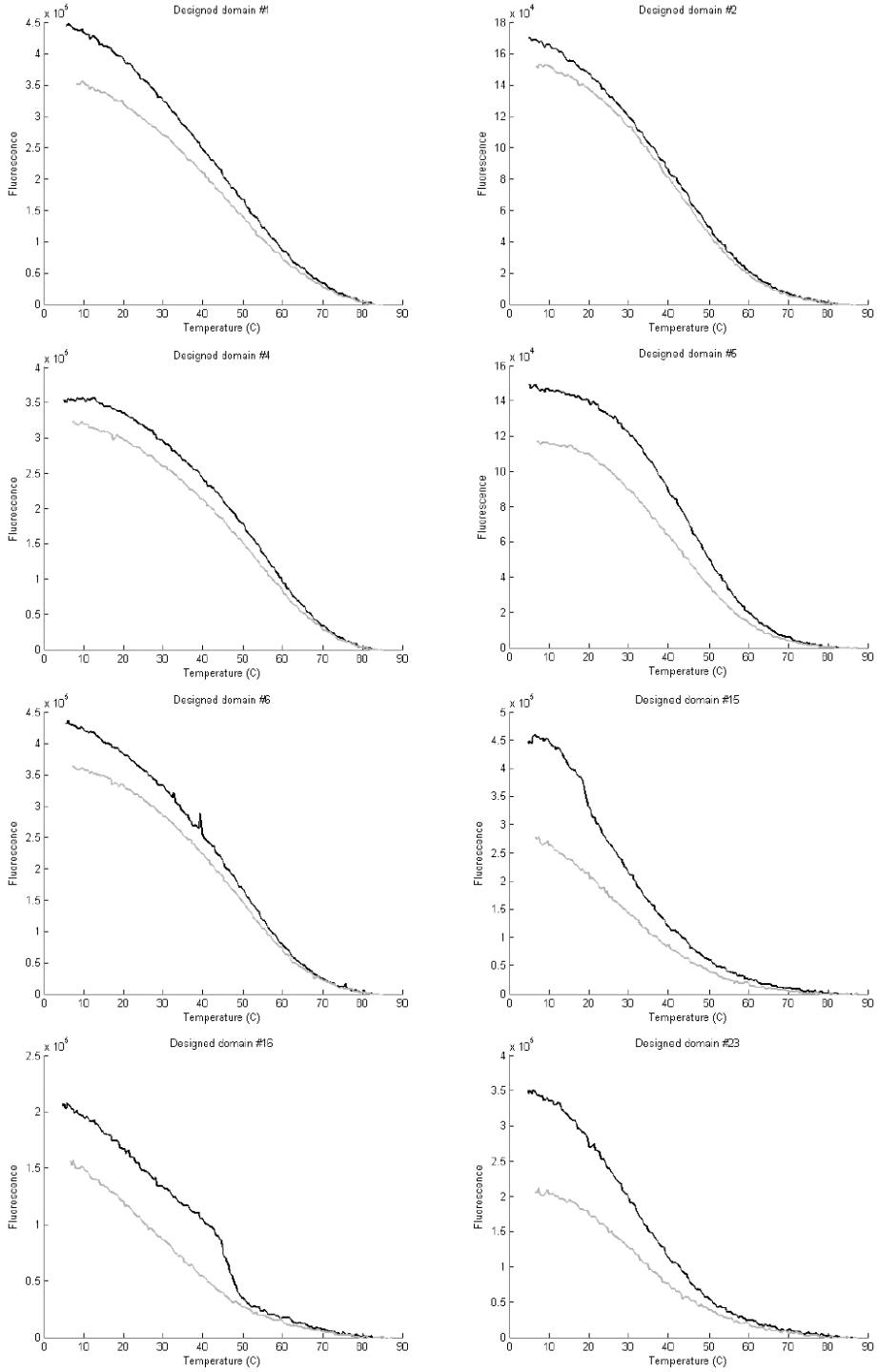


Figure 2.5: Melts of the 15 folded SCA-designed WW domains
Tryptophan fluorescence is monitored as a function of temperature as the domain is heated (black) and subsequently cooled (grey).

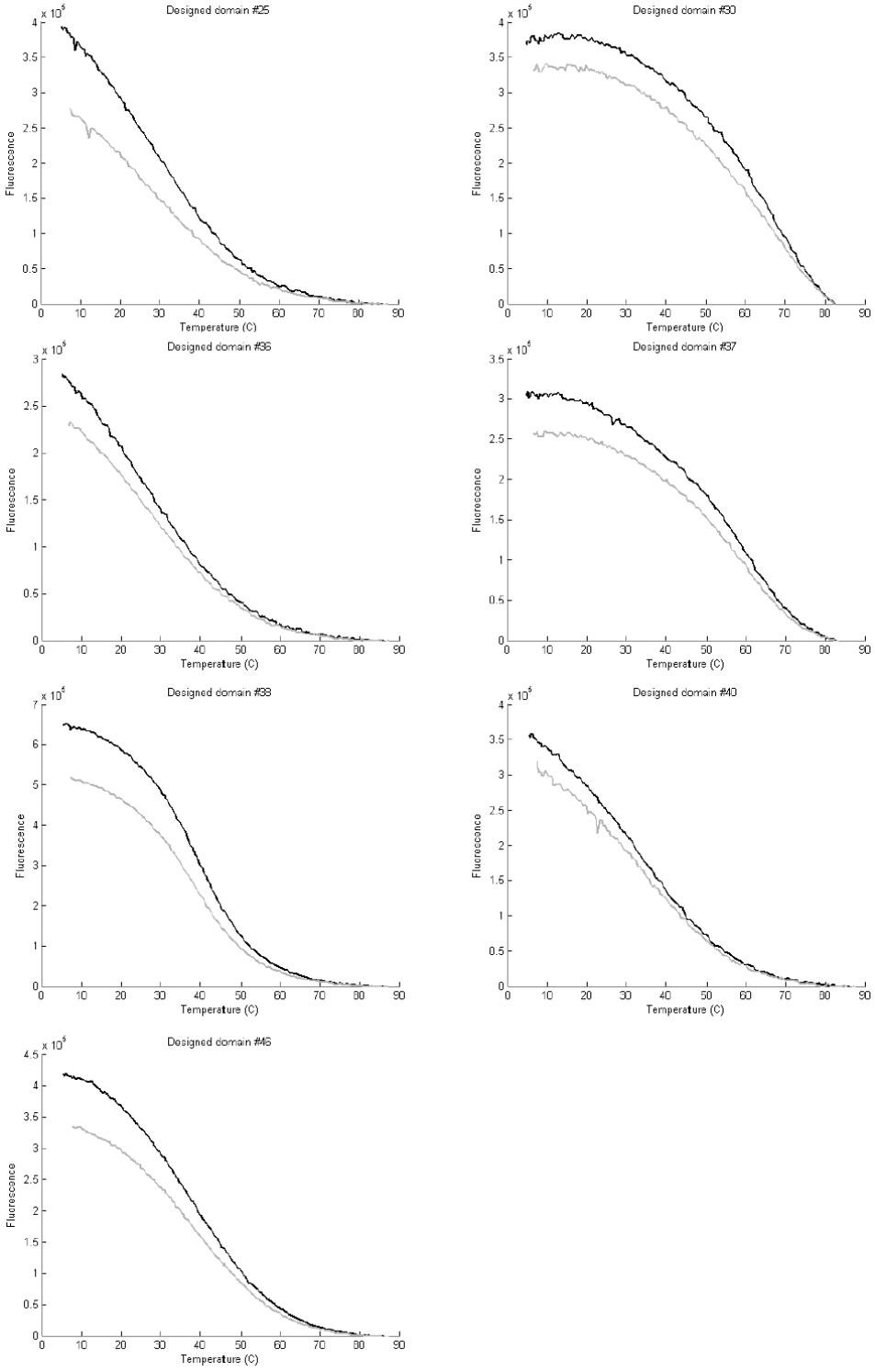


Figure 2.5 continued

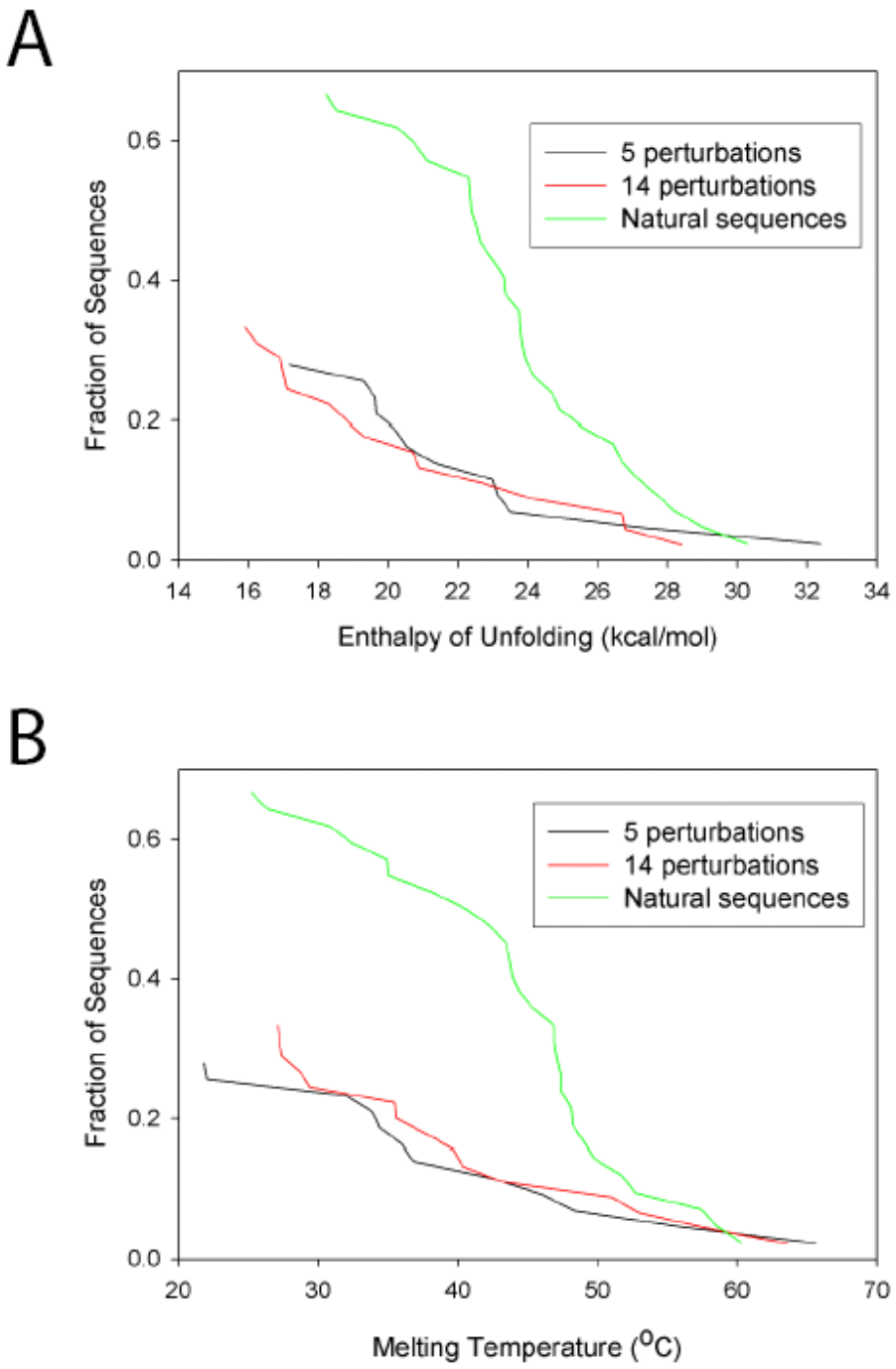


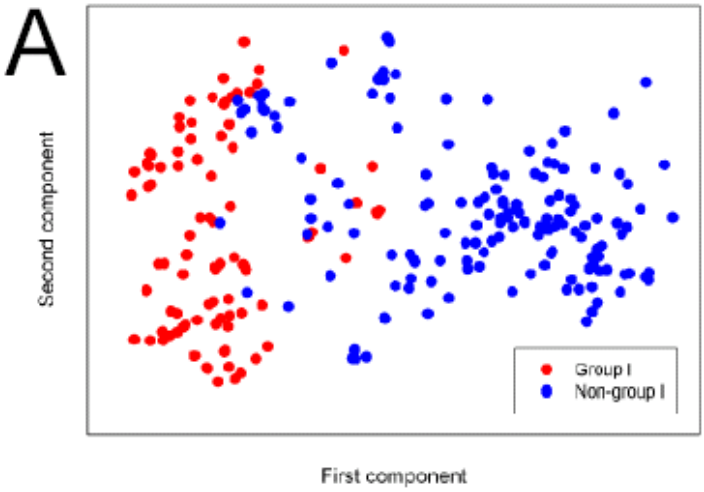
Figure 2.6: Thermodynamic characteristics of designed and natural WW domains
Plotted is the fraction of sequences that have at least the specified enthalpy of unfolding (A) or melting temperature (B). Data for natural sequences and sequences designed with 5 perturbations are from Socolich et al., 2005.

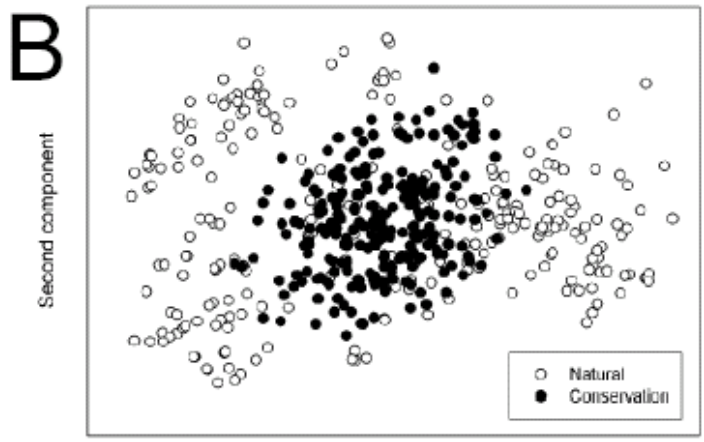
Sequence mapping

To visualize the distribution of the designed sequences in sequence space, I used the principal components analysis (PCA) based approach described by Casari et al. (1995). Conceptually, a "sequence space" can be imagined in which every protein sequence from an alignment is represented as one point in space, and the distance between any two points is equal to the number of residues that are different between the two sequences. Similar sequences would lie near each other, and dissimilar sequences would be far apart. In general there is no way of making such a mapping on a two- or even three-dimensional space that fulfills this for all sequences; doing so would require a large number of dimensions – up to the number of sequences in the alignment minus one. However, principal components based mapping of the sequences allows one to make a mapping onto a low dimensional space. Each principal component can be represented as a weighted sum of scores for the different amino acids at each position, and they have the property that the first principal component captures the greatest possible variance between the sequences and subsequent principal components capture the greatest possible amount of the remaining variance while being orthogonal to previous principal components. The utility of this approach for a biologist is that the first few principal components tend to capture most of the biologically important information in the sequences – mappings of G proteins, SH2 domains, and cyclins produce clusters of sequences with similar biological activities, and the residues with the greatest weight on the first principal components are those that have been shown to be important for biologic activity (Casari et al., 1995)

Such a mapping of the alignment of natural WW domains segregates these sequences according to their binding specificity, with sequences containing signature residues of group 1 binding described by Otte et al. (E 8, V/I 21, H 23, R/K 26) separated on the primary axis from those without the motif (figure 2.7). Applying the same projection to an alignment made with only the conservation pattern produces a map that is qualitatively different from that of the natural alignment, with the conservation sequences falling in a homogeneous distribution in the center of the map. In contrast, the sequences from an alignment made with coupling information fall in a distribution that resembles that of the natural alignment.

With the algorithm now capable of converging with large alignments, it is possible to similarly examine a more complex protein family such as the GPCRs (figure 2.8). The natural GPCR sequences are distributed nonuniformly, falling into clusters corresponding to their biological functions. Remarkably, this two-dimensional projection in which the sequence information underlying the receptors' specificity is reduced to two numbers is still able to convey much information about a complex protein. The previous results are even more pronounced with the designed sequences: sequences with only conservation again fall in a homogeneous central distribution, and coupling information shapes the distribution of the designed sequences to resemble the natural alignment.


A
Second component
Group I
Non-group I
First component


B
Second component
Natural
Conservation
First component

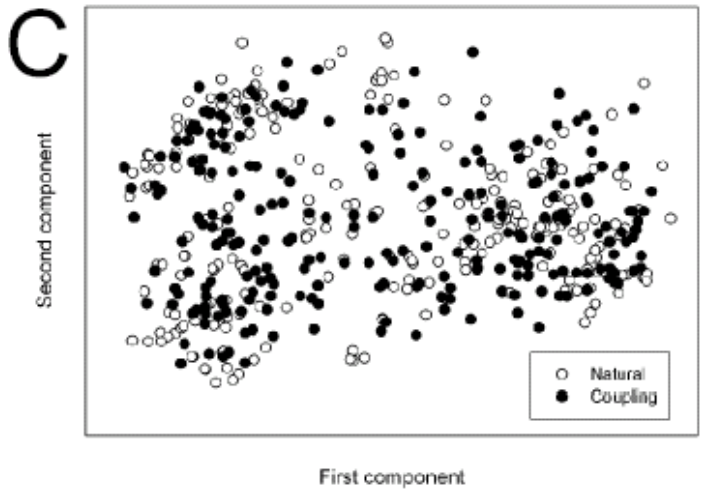
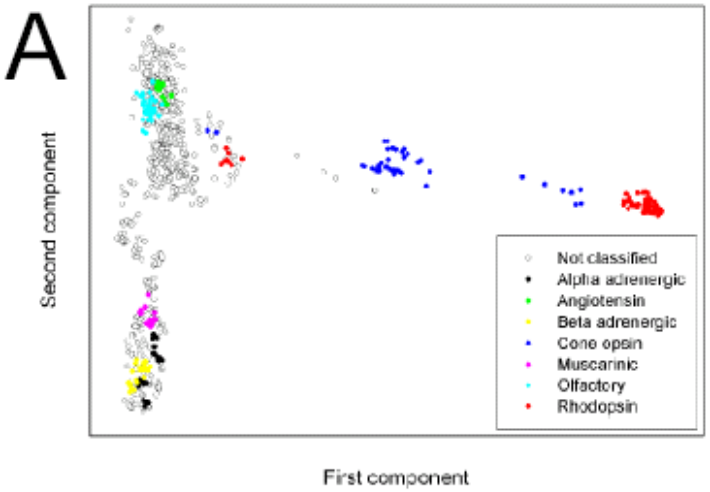
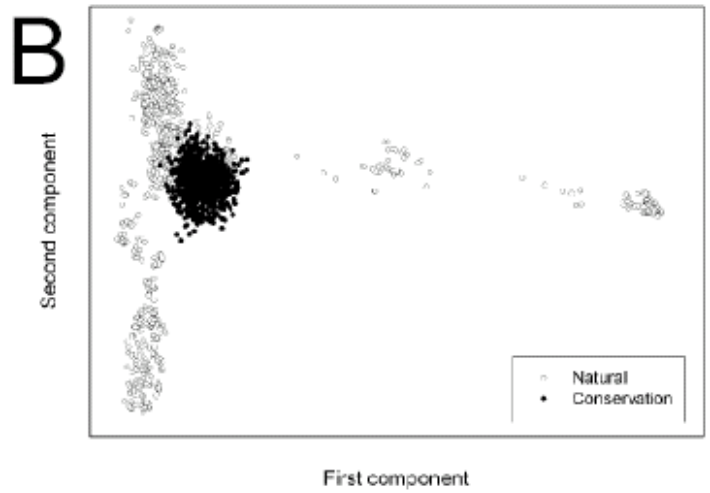

C
Second component
Natural
Coupling
First component

Figure 2.7: PCA mapping of natural and designed WW sequences
A) Map of natural WW domain sequences, colored according to expected binding specificity based on sequence motifs (see text). Each circle represents one sequence. Axes correspond to the first two principle components.
B, C) Maps of sequences designed using only conservation information (B) or coupling information (C), using the same projection as the map of natural sequences.


A
Second component
First component
Not classified
Alpha adrenergic
Angiotensin
Beta adrenergic
Cone opsin
Muscarinic
Olfactory
Rhodopsin


B
Second component
First component
Natural
Conservation

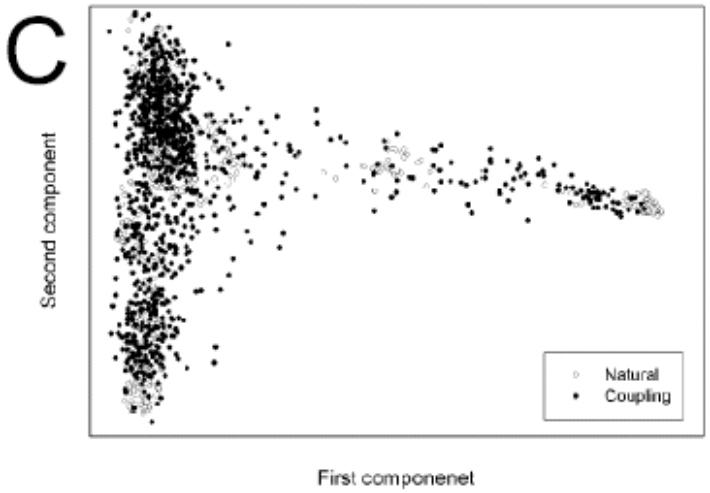

C
Second component
First componenet
Natural
Coupling

Figure 2.8: PCA mapping of natural and designed GPCR sequences
A) Map of natural GPCR sequences, colored according to their annotated function.
B, C) Maps of sequences designed using only conservation information (B) or coupling information (C), using the same projection as the map of natural sequences.

WW domain design using global coupling analysis

The global SCA can in principle provide much more information than perturbation-based SCA; the number of elements in a matrix of $E_{i(x),j(y)}$ where *i* and *j* cover all positions in the alignment and *x* and *y* cover all 20 amino acids is much larger than a matrix of $\Delta\Delta E_{i,j}^{x}$ where *i* covers all perturbations, *x* covers all positions, and *j* covers all perturbations. There is a coupling value between all pairs of amino acids at all positions, rather than between all amino acids at each position and a set of perturbations. The functional significance of these extra coupling values has not been evaluated because the coupling values unique to global SCA are weak when examined individually – they would not be expected to create an appreciable effect if probed by site-specific mutations. However, the entire set of data taken as a whole may comprise a set of constraints that places significant restrictions on which sequences are consistent with the coupling pattern. If these restrictions represent actual evolutionary pressure, then imposing them in an SCA-based sequence design algorithm would narrow the set of sequences consistent with the coupling pattern, excluding those that do not meet the added constraints and enriching the final set of designed sequences for those that meet all of the evolutionary constraints on the protein family.

When the global SCA-based sequence design algorithm is used with the alignment of 247 WW domains, a notable difference is observed compared to perturbation-based design (figure 2.9). The alignment is initially scrambled during a shuffling phase, losing the natural coupling pattern and reaching the same energy as that

of an alignment created with conservation alone. After a plateau in this phase, the algorithm then begins to converge toward a lower energy, gradually rebuilding the natural coupling pattern and eventually reaching a lower plateau at which the natural coupling pattern has been rebuilt and the probability of accepting further swaps is near zero. While this is similar to the behavior of the perturbation-based algorithm, the global SCA-based design algorithm converges to a final alignment that is much more similar to the starting natural alignment. The mean sequence identity between each designed sequence and its most similar sequence from the natural alignment is 50% for sequences designed using conservation alone, 56% for sequences designed in Socolich et al., and 59% for the WW domain sequences designed here with 14 statistical perturbations. With global SCA-based design, this reaches 87% at convergence.

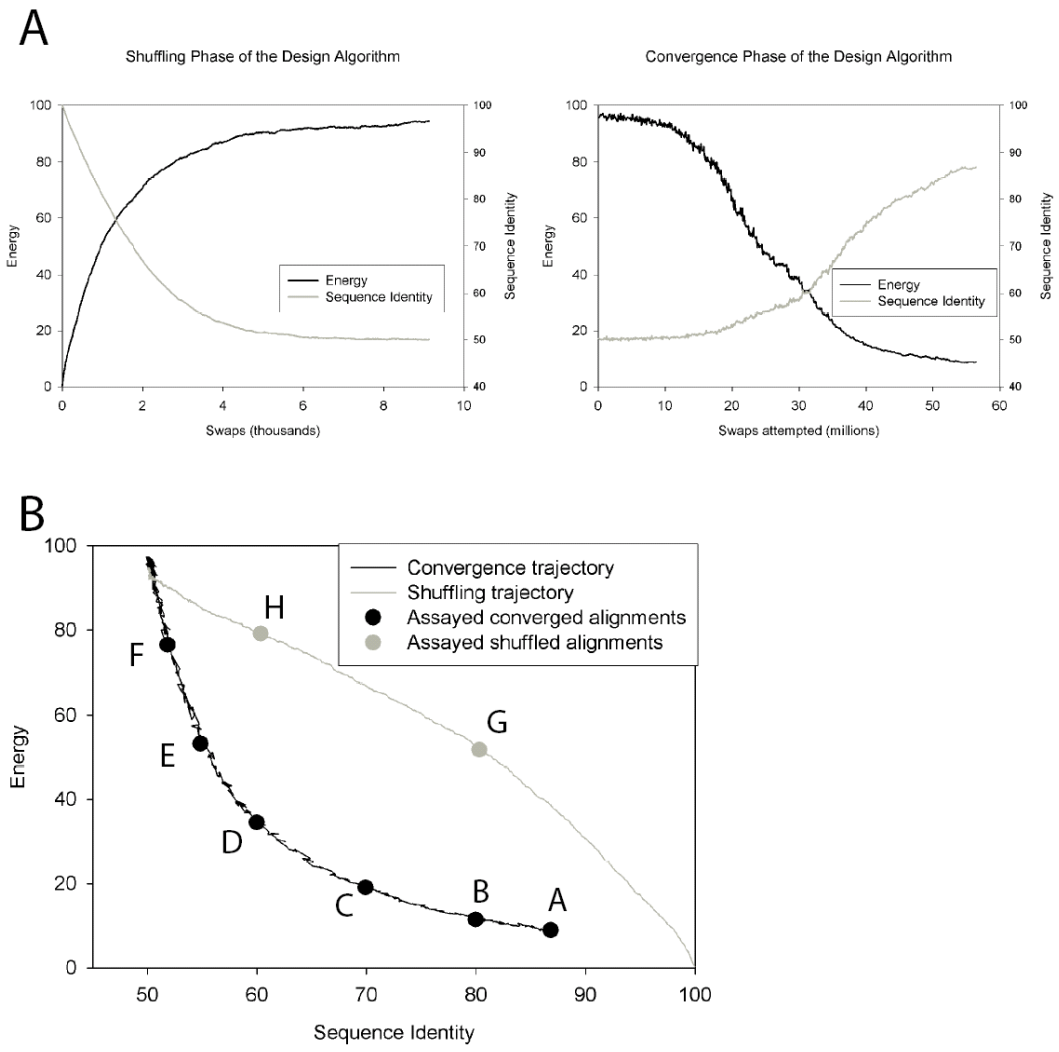


Figure 2.9: Global SCA-based design with the WW domain alignment

A) The algorithm begins with a shuffling phase, starting with the natural alignment and making random swaps of residues within the same column of the alignment, eventually destroying the coupling pattern. This is followed by a convergence phase, in which the coupling pattern is rebuilt.

B) A plot of energy as a function of sequence identity shows the different relationship between these parameters during the shuffling and convergence phases. Circles indicate the alignments from which sequences were drawn for experimental characterization.

The increased number of constraints imposed by global SCA clearly further restricts the available set of sequences consistent with the coupling pattern, and it would be no surprise to find a higher probability of folding for the sequences designed with the global algorithm simply because they are very similar to the natural sequences. While this design approach surely provides *sufficient* information to specify the restraints on the WW domain, it also provides a way to address how much SCA information is *necessary* to specify the protein family. All previous experiments have examined sequences after convergence of the design algorithm, and have shown that at this point enough SCA information has been incorporated to generate natural-like sequences. However, the status of the designed sequences before this point is unknown. Previously, experiments to relate the fraction of SCA information imposed at intermediate points during convergence to the probability of folding for these sequences were not undertaken because the probability of folding for the fully converged sequences was modest, meaning that assaying intermediate points that would be expected to have even lower probabilities of folding would require large libraries of sequences to generate meaningful statistics. With global SCA-based design, the sequences at complete convergence are expected to have a probability of folding near that of the natural sequences, so such an experiment now becomes feasible.

WW domains with varying amounts of SCA information

The primary goal of this project is to determine the amount of SCA information necessary to specify the WW domain, which is accomplished by evaluating the

probability of folding for sequences designed with varying amounts of information. Furthermore, because the design algorithm creates sequences with increasing levels of sequence identity to the natural alignment, which might in itself be an explanation for generating folded proteins, a second goal of this project is to dissect the degree to which folding is attributable to the incorporated SCA information versus increased sequence identity alone. In order to differentiate the effects of coupling and sequence identity, alignments must be compared in which these two values change orthogonally.

This can be done by using sequences generated during the shuffling phase of the design process. At the beginning of a design run, the algorithm accepts all attempted swaps of two randomly chosen residues within the same column of the alignment. Beginning with the natural sequence alignment at zero energy and 100% sequence identity, this causes the energy to increase and the sequence identity to decrease (figure 2.9 A). Importantly, it does not simply follow a mirror image of the trajectory in the convergence phase. This can be appreciated in a plot of energy as a function of sequence identity (figure 2.9 B). At any given level of sequence identity, the alignment has a lower (or equal) energy in the convergence phase than in the shuffling phase. Because these trajectories are different, this allows comparison of alignments that have the same level of sequence identity but different energies, thus showing the effect of the incorporated SCA information independent of the associated change in sequence identity. It should also be noted that the two phases of the design process operate very similarly – both phases use swaps of randomly selected pairs of residues within the same column, always maintaining the site-specific conservation pattern of the original alignment, and the only

difference between the two is whether or not SCA information is used as a selection criteria for the swaps.

Folding with varying amounts of SCA information

Alignments from the shuffling and convergence phases of the design process were taken at the points shown in figure 2.9 B. Twenty sequences were randomly selected from each of these alignments and assayed for folding as in the previous WW domain experiments: Alan Poole and I synthesized genes encoding the proteins, expressed and purified the protein, and assayed folding by monitoring tryptophan fluorescence as a function of temperature. Results are shown in table 2.2.

Eleven of twenty sequences chosen from the natural alignment are folded. The previously observed probability of folding for sequences from a natural sequence alignment of 67% (Socolich et al., 2005) is within the 95% confidence interval of this value assuming a binomial distribution. Alignments from both phases of the design algorithm with 80% or higher sequence identity (sets A, B, and G) show the same probability of folding. Thus, mutating up to 20% of the residues in these proteins, even ignoring coupling information but preserving site-specific conservation, has minimal effect on the stability of the fold.

Alignment	Alignment Energy	Sequence Identity	Number folded	T_m (ºC) Mean ± σ	ΔH_{unf} (kcal/mol) Mean ± σ
Natural	0.0	100%	11 (55%)	42.2 ± 8.1	23.0 ± 5.3
A	8.9	87%	14 (70%)	39.7 ± 6.4	18.8 ± 3.6
B	11.5	80%	11 (55%)	38.9 ± 9.9	18.6 ± 2.5
C	19.1	70%	6 (30%)	39.9 ± 9.4	19.0 ± 2.0
D	34.5	60%	6 (30%)	33.8 ± 6.5	18.6 ± 2.3
E	53.2	55%	6 (30%)	34.6 ± 6.9	18.0 ± 3.4
F	76.4	52%	4 (20%)	30.8 ± 7.4	17.3 ± 2.0
G	51.8	80%	9 (45%)	37.7 ± 7.7	18.2 ± 2.1
H	79.1	60%	1 (5%)	49.4	25.5

Table 2.2: Characteristics of the designed sequences
Alignments are labeled according to figure 2.8 B

At lower levels of sequence identity, the difference between the two design approaches becomes evident. Alignments from the convergence phase with mean sequence identity ranging from 70% to 55% (sets C, D, and E) all have six of twenty sequences that demonstrate reversible cooperative thermal denaturation, indicating a folded tertiary structure. In contrast, of twenty sequences from the shuffling phase at 60% sequence identity (set H), only one is folded. This demonstrates that sequences designed with SCA information have a significantly higher probability of folding than do sequences designed without coupling information, even at the same level of sequence identity. Coupling information, and not the associated increase in sequence identity, is the reason why the designed sequences are able to fold.

DISCUSSION

The algorithm presented here is an implementation of the concept described by Socolich et al. (2005) that converges on the SCA values directly rather than the frequency based function used previously, and achieves computational efficiency by using tables of precalculated SCA values. While the use of SCA values instead of the previous frequency function may be conceptually more elegant, this particular modification would not be expected to make a major difference in the outcome for cases in which both design algorithms are able to converge. This is because minimization of the previous algorithm's energy function does result in restoration of the SCA pattern, even if those were not the exact values used to drive convergence. Instead, the major value of the modified design algorithm lies in its increased computational efficiency and robustness in converging for many different alignments in a reasonable time frame. That makes it possible to use the design algorithm with the larger perturbation set described here, and allows for SCA-based design of larger protein families.

The WW domains designed with additional statistical perturbations beyond those used by Socolich et al. (2005) show behavior essentially indistinguishable from the set of sequences described previously. The five statistical perturbations in the initial study capture as much SCA information relevant for specifying the WW fold as the set of 14 perturbations used here. This is consistent with the fact that SCA analysis of many diverse protein families has revealed that each contains a network of mutually interacting residues: specifying the coupling patterns of a few perturbations within the group of

interacting positions would then be expected to constrain the entire network. Thus, the fundamental architecture of the WW domain is defined by a highly interacting network obeying a simple set of rules that are revealed by a few SCA perturbations, rather than by a large set of potentially complex interactions that would require several perturbations to be revealed in its entirety.

The global SCA based design algorithm provides another perspective on WW domain complexity. This algorithm converges on a larger information set and ultimately attains very high sequence identity with the natural alignment. Because the previous perturbation based design experiments did not generate sequences with the same probability of folding as the natural alignment, some degree of information relevant for folding must be missing from the design process. That information is recovered by global SCA in the late stages of convergence, as sequences at this stage have a probability of folding as high as the natural sequences. At lower levels of sequence identity to the natural alignment comparable to those reached in the perturbation-based design experiments, the sequences designed using either of the SCA based methodologies show similar probabilities of folding. In contrast, sequences with the same level of sequence identity but designed without SCA information are much less likely to fold. The implications of this can best be appreciated by considering where the coupling information arises in global SCA based design.

Figure 2.10 shows histograms of the coupling value that each residue pair has in the natural sequence alignment, and the bars of the histograms are colored according to the fraction of natural-like coupling that is found in the designed sequences for those residue pairs. From this figure, it can be appreciated that alignments designed using SCA

information (sets A-F) retain natural-like coupling at the originally strongly coupled positions and lose the coupling pattern preferentially at sites that were only weakly coupled in the starting alignment. In contrast, sequences generated without coupling information (sets G and H) show a more uniform loss of coupling for all pairs of residues, regardless of the strength that the coupling interactions originally had. This explains why the global and perturbation-based design algorithms perform similarly at comparable levels of sequence identity: the global SCA based design algorithm preferentially preserves the strongest coupling interactions, and the perturbations chosen for the perturbation-based design experiments were those with the highest coupling values.

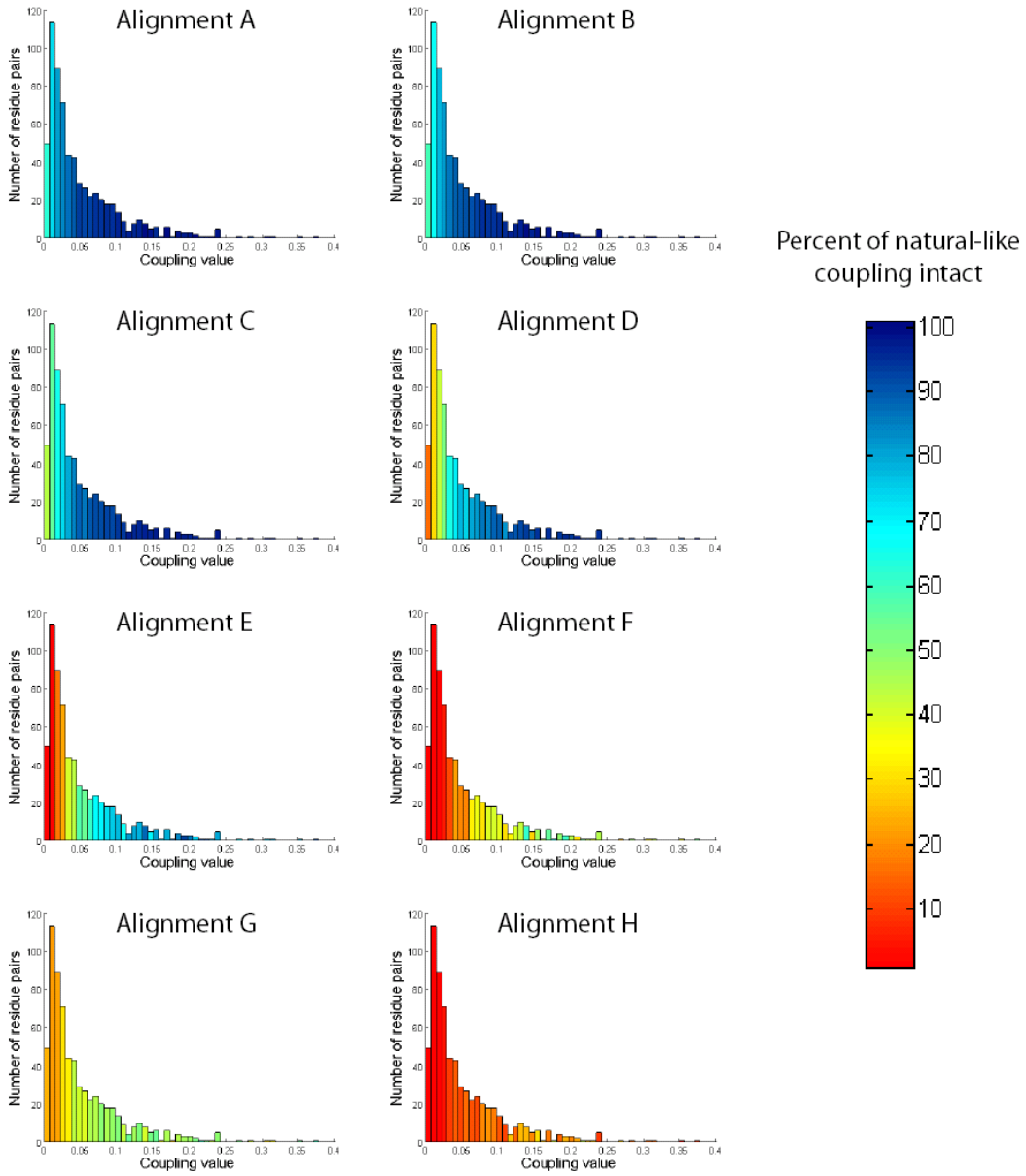


Figure 2.10: Restored coupling interactions in the design process
Histograms are shown for each of the designed alignments. Each histogram bin represents all of the coupling interactions that have the indicated strength in the natural alignment. Bar heights are the number of residue pairs in the natural alignment that have the indicated coupling value. Bars are colored according to the fraction of natural-like coupling that has been restored for the corresponding residue pairs in the designed alignments.

Furthermore, these experiments show that imposing even a small degree of coupling information can create surprisingly large effects. While sequences maintaining only the conservation pattern have 50% identity to their nearest natural sequences, they were shown previously to have no observable probability of folding (Socolich et al., 2005). In this study, incorporating a level of SCA information that changed the mean sequence identity only from 50% to 52% (set F) was sufficient to achieve an appreciable probability of folding. Put into perspective, for a WW domain with 36 residues, a single residue constitutes 2.7% of the protein. Part of the explanation for this surprising potency of SCA information may lie in inherent limitations of using sequence identity alone to compare sequences. For example, it is well known that not all positions in a protein have equal importance – the PSIBLAST algorithm uses alignments of sequences to deduce the site-specific importance of amino acids at different positions, improving the quality of database searches (Altschul et al., 1997). Because of the unequal importance of different sites, sequence identity scores can be confounded when the "signal" from important positions is masked by "noise" from unimportant ones. Regardless, the effect of incorporating even a small amount of SCA information remains quite dramatic.

METHODS

Perturbation-based design algorithm

The general approach of the design algorithm remains unaltered from the initial implementation (Socolich et al., 2005). The algorithm begins with a natural sequence

alignment and at each step it randomly selects one column of the alignment, chooses two sequences that do not have a gap in that column, and attempts to swap the corresponding residues. The energy of the system is defined as the sum of magnitudes of differences between the SCA values of the original alignment versus the alignment being designed:

$$Energy = \sum_{i,j,x} \left| \Delta\Delta E^{x}_{i,j,original} - \Delta\Delta E^{x}_{i,j,designed} \right|$$

where $\Delta\Delta E^{x}_{i,j}$ is the statistical energetic change of amino acid *x* at position *i* in response to perturbation *j*, *i* ranges over each column in the alignment and *j* covers each perturbation at which coupling is calculated. The perturbations to use for calculations are defined by the user in a parameter set. The algorithm converges via simulated annealing: an attempted swap is accepted or rejected based on a Metropolis Monte Carlo selection criterion, with probability $P(accept) = e^{-\Delta Energy / Temperature}$ as described in Results, initially allowing energetically large swaps to shuffle each column of the alignment and gradually increasing the stringency for acceptance and leading to convergence on the original coupling pattern. Design runs for this project used a starting temperature of 100 and an ending temperature of 10^{-6}, with 10% decreases in temperature between iterations of 45695 accepted swaps (5 x #sequences x #columns) or 456950 attempted swaps (50 x #sequences x #columns). The perturbation set used to calculate SCA values was: 3 L, 4 P, 6 G, 7 W, 8 E, 16 G, 19 Y, 20 Y, 22 N, 23 H, 25 T, 28 T, 30 W, 33 P.

Global SCA-based design algorithm

Coupling values were calculated using the global SCA calculation described in (Sharma, 2006) using omission of single sequences. The shuffling phase consisted of 91390 (10 x #sequences x #columns) random swaps. In the convergence phase, the temperature began at 100 and ended at 0.1. Iterations of 91390 swaps were performed until two successive iterations had no decrease in energy, at which point the temperature was decreased by 10% before the next iteration.

Protein folding assays

Assays were performed as described in (Socolich et al., 2005). Genes encoding the redesigned sequences were constructed with PCR of partially overlapping oligonucleotides encoding the genes with E. coli optimized codons and NcoI and XhoI restriction sites. These were cloned into the vector pHis8.3 with an N-terminal octahistidine tag and thrombin cleavage site. Protein was expressed in 50 ml cultures and purified on 75 μl nickel-nitrilotriacetic acid (Ni-NTA) agarose. The domains were transferred to 100 mM NaCl, 100 mM KPO_4, pH 7.0 to 5 μM final concentration and subjected to thermal denaturation and recooling while monitoring fluorescence quenching of the buried tryptophan (ex. 295 nm, em. 340 nm). Derivatives were fit to the van't Hoff equation to determine thermodynamic parameters (John and Weeks, 2000).

Sequence mapping

Natural sequences were mapped essentially as described by Casari et al. (1995). Calculations were performed in Matlab. A matrix of dissimilarity between sequences from the natural alignment **A** was calculated such that a_{ij} equals one minus the fraction sequence identity between sequences *i* and *j*. This was used in principal components analysis to generate a transformation matrix of principal component coefficients, **T**, and the product **A** × **T** produced a matrix of coordinates. Designed sequences were mapped using the projection calculated for the natural alignment by generating a dissimilarity matrix **B** where b_{ij} is one minus the fraction sequence identity between designed sequence *i* and natural sequence *j*, and using the same matrix of principal component coefficients from the natural alignment to calculate coordinates as **B** × **T**.

CHAPTER THREE
SCA-based PDZ domain design

INTRODUCTION

While SCA information is sufficient to specify the WW domain, this particular model system has the drawback that this protein fold is much smaller than most, and the hydrophobic core is comprised largely of a tryptophan residue that is highly conserved and therefore essentially invisible to coupling analyses. The control experiments comparing sequences designed with coupling information to sequences made with only conservation information show that conservation alone is insufficient to specify the protein, indicating that the success of SCA-based design cannot be attributed solely to the simplicity of this protein fold. Nevertheless, it remains an important outstanding question as to whether the sufficiency of SCA information to specify a protein extends to more complex folds with substantial hydrophobic cores – often considered to be the greatest challenge in protein design.

The PDZ domain is an excellent system to test the applicability of SCA-based design with a larger protein family. The PDZ fold comprises about 100 residues in a mixed alpha/beta fold with a substantial hydrophobic core, providing a sufficiently complex design target. In addition to folding, the PDZ domain functions as a protein scaffold by binding the C-termini of its targets with varying degrees of affinity and promiscuity (Stiffler et al., 2006). The PDZ evolutionary record is extensive (Ponting, 1997), providing sufficient information to serve as a basis for SCA. Furthermore, site-directed mutagenesis has validated the importance of the statistical couplings observed in this protein family (Lockless and Ranganathan, 1999). Therefore, the following project

tests the sufficiency of SCA information to specify the fold and function of the PDZ domain.

RESULTS

Sequence design

To test the sufficiency of SCA information to specify the PDZ domain, the global SCA-based design algorithm was used on an alignment of PDZ domains. Its progress is shown in figure 3.1. The algorithm begins with the alignment of natural PDZ domains and initially shuffles the columns of the alignment without regard to SCA information, eventually reaching the same energy and sequence identity as an alignment generated *de novo* using only the conservation pattern. This is followed by the convergence phase, in which the coupling pattern is rebuilt. The algorithm ultimately converges on a designed alignment with very high sequence identity to the natural alignment, where high sequence identity alone would likely be responsible for specifying folded proteins. In order to differentiate between the effect of coupling information and the associated increase in sequence identity, alignments generated in the shuffling and convergence phases were compared. The two phases show differing relationships between the amount of SCA information incorporated and the levels of sequence identity with the natural alignment (figure 3.1). Alignments at points C_A and C_B on the convergence trajectory examine sets of sequences with two differing levels of SCA information imposed and correspondingly different degrees of sequence diversity from the starting natural sequence alignment. The

alignment at point S_A from the shuffling trajectory tests whether the behavior of proteins from alignments C_A and C_B can be attributed to sequence identity alone, as this alignment has higher sequence identity to the natural alignment but less preservation of the coupling pattern than either of the two alignments on the convergence trajectory.

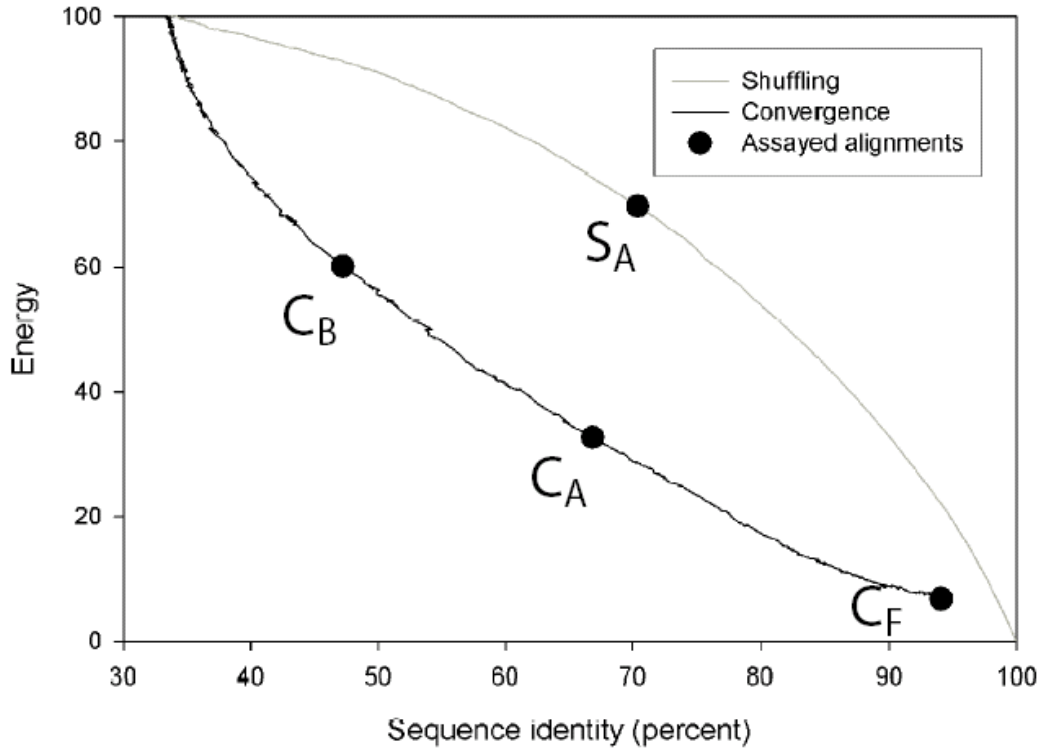


Figure 3.1: SCA-based design with the PDZ alignment
The design algorithm begins with the natural sequence alignment at 100% sequence identity and zero energy (the difference between the designed and natural alignments' coupling patterns). The initial shuffling phase randomly swaps pairs of residues in the same column of the alignment, reducing sequence identity to the natural alignment and increasing the energy as it destroys the coupling pattern. The convergence phase conditionally accepts swaps based on SCA information in a simulated annealing protocol to rebuild the natural alignment's coupling pattern. Importantly, at any given level of sequence identity, the alignments from the convergence phase have a lower energy (more natural-like coupling pattern) than those from the shuffling phase.

Compactness and folding designed sequences

Before undertaking work with alignments on the convergence and shuffling trajectories, I first evaluated the final alignment generated by completely converging on the coupling matrix. This would indicate the maximum probability of folding that should be expected from any of the alignments designed with coupling information. I drew twenty sequences from this final alignment (C_F) and expressed and assayed the proteins for compactness with size exclusion chromatography (SEC, figure 3.2). Seven of the twenty show a population of protein eluting at the mobility expected of a compact PDZ domain, setting the upper limit that should be expected from the designed sequences.

Similarly, twenty sequences were selected from alignment C_A and the proteins were expressed and assayed with size exclusion chromatography (figure 3.2). Five of the twenty show a detectable population at the mobility of a compact PDZ domain. However, gel filtration only demonstrates that these proteins are compact, not necessarily that they have a tertiary structure. To determine whether these proteins are folded in a stable tertiary structure, I evaluated their chemical shift dispersion with NMR. I expressed and purified ^{15}N-labeled samples and Pilong Li from Michael Rosen's laboratory recorded proton-nitrogen HSQC spectra. Folded proteins show characteristic peak dispersion on these spectra, whereas crosspeaks from backbone amides of unfolded proteins are generally limited to the 8.5-8.0 ppm range of the proton dimension. Three of the proteins from set C_A (C_A-2, C_A-9, and C_A-18) show widely dispersed chemical shifts indicating that they have a folded structure (figure 3.3), while one (C_A-15) was undetectable and one (C_A-19) was unstable and precipitated as the spectrum was acquired.

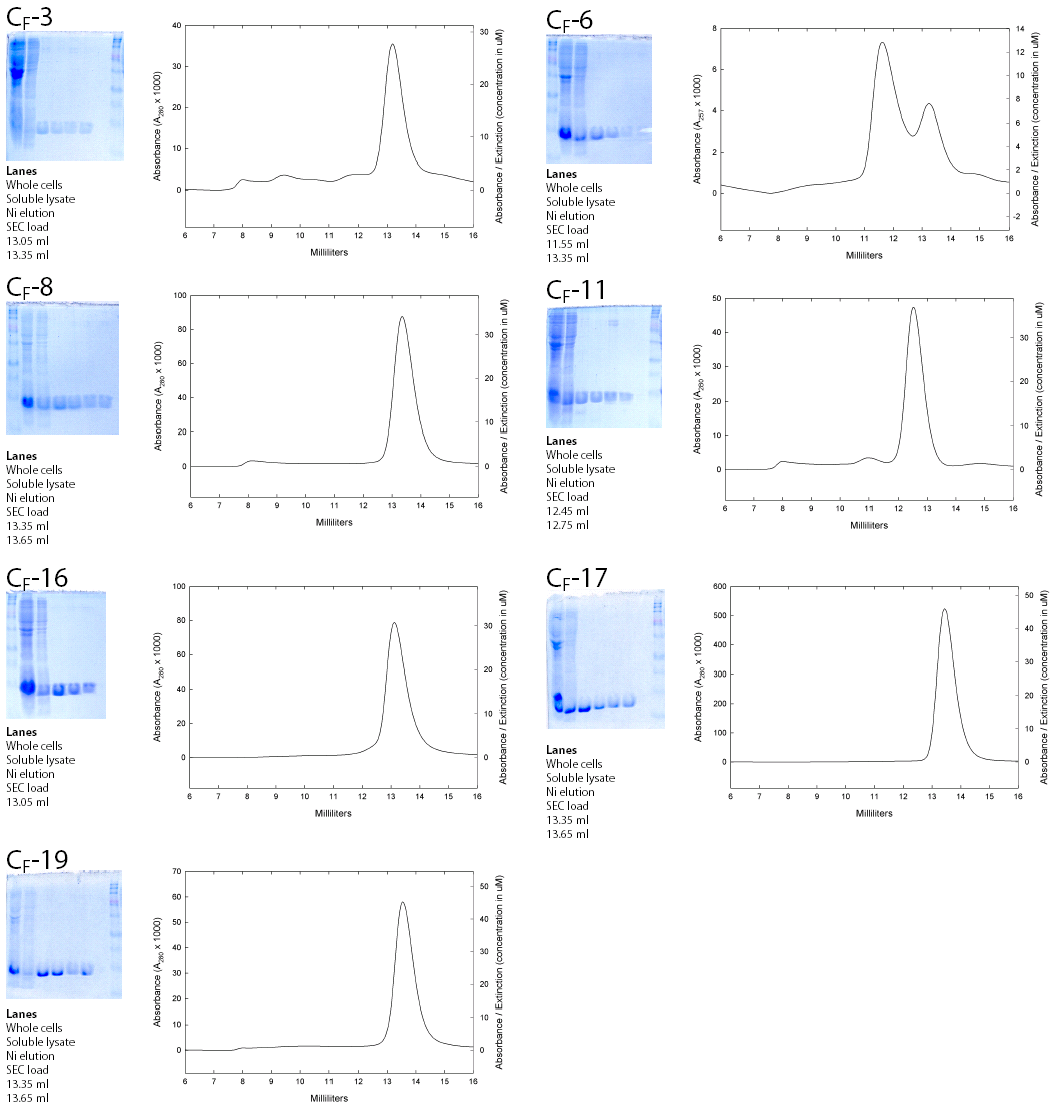


Figure 3.2: Purification and size exclusion chromatography of the designed proteins

Gel lanes are of whole cells, soluble lysate, elution, load for SEC, and 0.3 ml fractions with the center of the fraction indicated. Ladders: 180, 115, 82, 64 (pink), 49, 37, 26, 19, 15, 6 kDa. Chromatography absorbance traces are shown, with the right axis indicating the absorbance divided by the extinction coefficient of the designed protein to estimate concentrations of peaks containing the protein. Elution volumes of chymotrypsinogen (25 kDa) and RNAse A (13.7 kDa) standards are 11.4 ml and 12.7 ml, respectively. Proteins with detectable populations at the mobility of a compact PDZ domain are shown, results for all proteins are included in the appendix.

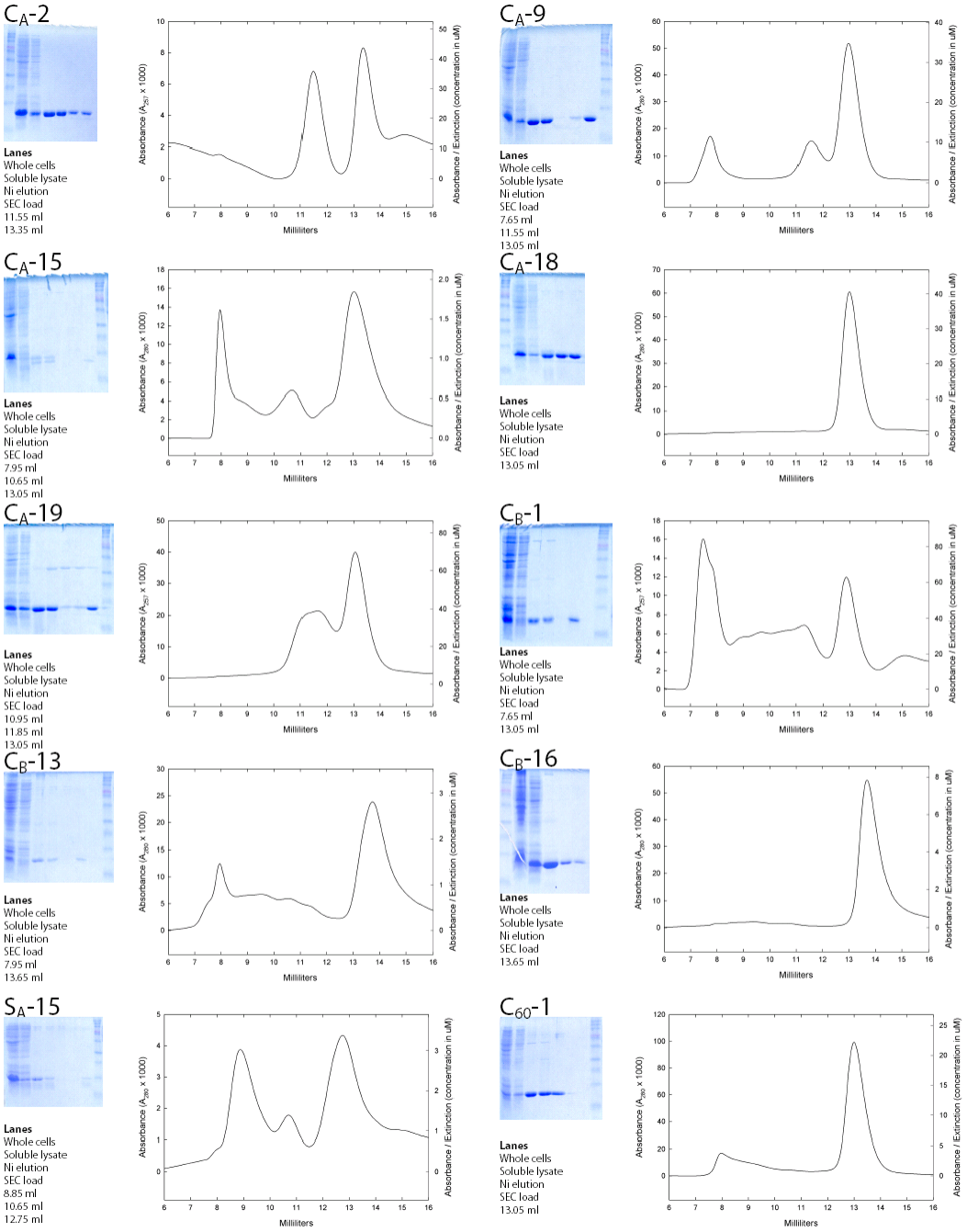


Figure 3.2 continued

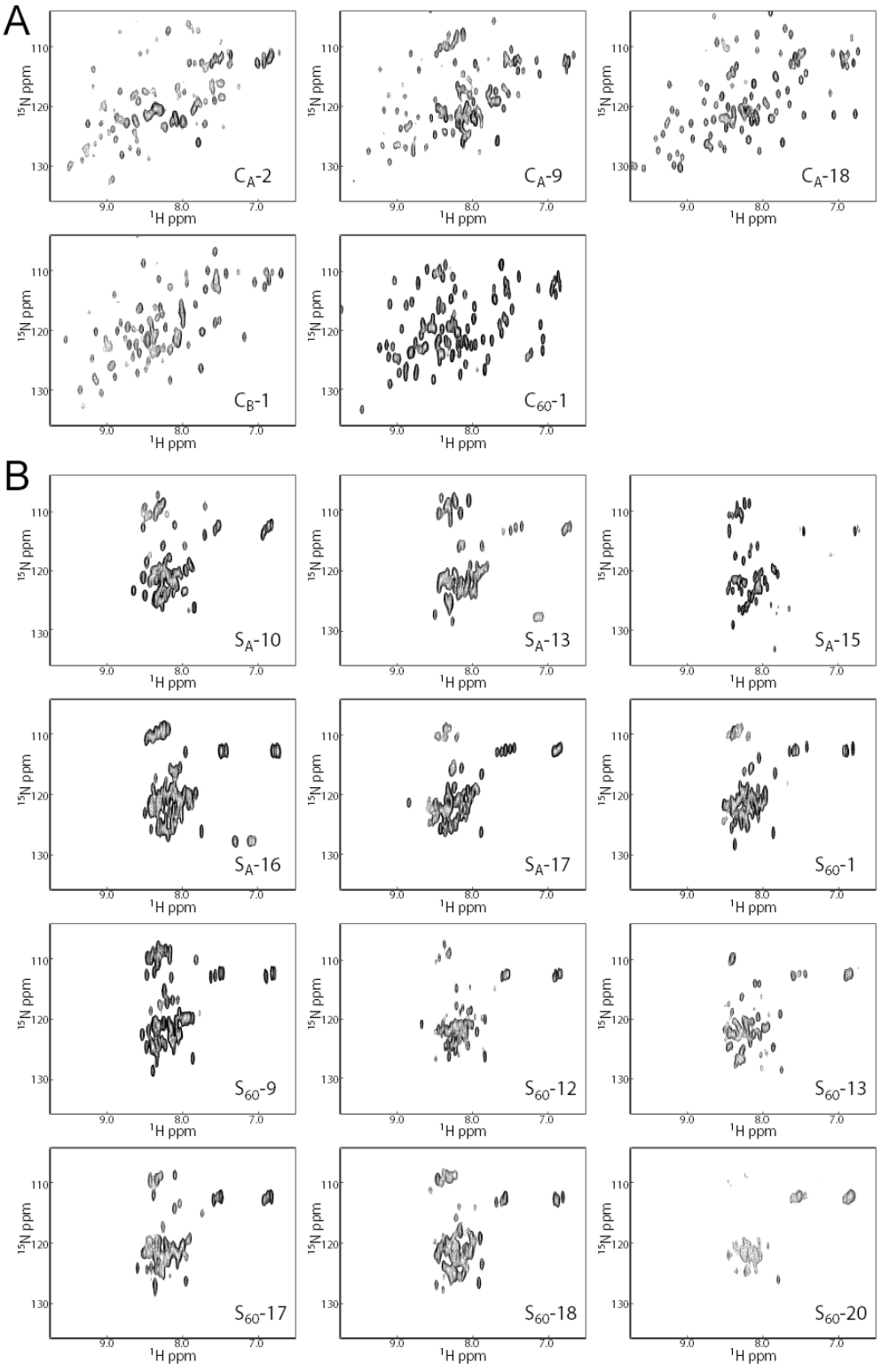

A
C_A-2
C_A-9
C_A-18
C_B-1
C_{60}-1
B
S_A-10
S_A-13
S_A-15
S_A-16
S_A-17
S_{60}-1
S_{60}-9
S_{60}-12
S_{60}-13
S_{60}-17
S_{60}-18
S_{60}-20
^{15}N ppm
^{1}H ppm

Figure 3.3: HSQC spectra of the designed proteins
Chemical shift dispersion downfield of 8.5 ppm and upfield of 8.0 ppm indicates that C_A-2, C_A-9, C_A-18, C_B-1, and C_{60}-1 are folded (A), while proteins from the shuffled alignments are not (B). Protein C_A-19 aggregated during HSQC acquisition and is classified as not folded. Proteins C_A-15, C_B-13, C_B-16, S_A-4, and S_A-14 had insufficient yield to produce detectable signal.

The fact that these folded proteins were found at point C_A indicates that sufficient coupling information has been imposed to achieve a probability of folding that is not drastically lower than in the fully converged alignment. I next tested whether greater levels of sequence diversity could be achieved by slightly relaxing the SCA constraints – using alignments with higher energy – without compromising folding. Twenty sequences were drawn from alignment C_B and assayed with SEC and NMR. Three of these show populations of protein with the mobility of a compact PDZ domain (figure 3.2), and one (C_B-1) shows proton chemical shifts downfield of 9 ppm on an HSQC indicating that it is folded (figure 3.3).

These experiments indicate that the alignment at C_A was created with enough information to achieve an appreciable probability of generating folded proteins, while the alignment at point C_B still has enough information for a low probability of folding. I next addressed what type of information is responsible for these results: the coupling information that the algorithm imposed, or the simultaneous increase in sequence identity to the natural alignment. I examined alignment S_A (figure 3.1), which has a higher energy (less natural-like coupling) and higher sequence identity to the natural alignment than either C_A or C_B. Twenty sequences from S_A were assayed using the same procedure as for the alignments designed with coupling information. Only one of these (S_A-15) shows a detectable population on SEC consistent with a folded monomer. In order to exclude the possibility that any proteins in this set are folded but do not show the expected mobility of a monomer due to constitutive dimerization (a phenomenon reported for some natural PDZ domains – Im et al. 2003a, Im et al. 2003b), HSQCs were collected for every protein with a population outside of the void volume. All of these show limited chemical shift

dispersion characteristic of unfolded proteins (figure 3.3). This indicates that none of the proteins designed without coupling information, even at relatively high levels of sequence identity, are able to fold.

Structure of a SCA-designed PDZ domain

I then sought to evaluate the quality of the fold for a SCA-designed PDZ domain with relatively low sequence identity to the natural alignment. Protein C_B-1 expresses in insufficient yield for structural analysis. Protein C_A-2 expresses well, but would not crystallize despite extensive efforts. It also exhibited slow interconversion between species with mobility on SEC and molecular mass on dynamic light scattering consistent with populations of a monomer and of a dimer that would severely complicate NMR structure determination.

I next assayed sets of 25 sequences generated in the convergence phase and the shuffling phase, restricting these to sequences with 60% ± 2% sequence identity to their nearest natural sequence from the starting alignment, designated C_{60} and S_{60}. Consistent with my observations for set S_A, no proteins from set S_{60} showed the chemical shift dispersion of a folded protein. One protein from set C_{60} was folded, and this protein expressed in high yield and did not exhibit interconversion between distinct populations. Protein C_{60}-1 did not crystallize, but it was amenable to structure determination through NMR spectroscopy. I prepared protein samples, Pilong Li collected the appropriate spectra, and I performed spectral analysis and structure calculations.

The structure of the SCA-designed protein C_{60}-1 is essentially identical to that of natural PDZ domains (figure 3.4 A and B), showing that the protein is not only folded, but folded in the expected conformation. Beyond recapitulating the global protein fold, every side-chain falls in the correct position in the three-dimensional structure expected based on its position in the sequence alignment compared to natural PDZ domains (see figure 3.5), indicating that not only the overall topology but also the interactions between specific residues are correctly specified. Interestingly, the β1-β2 and β2-β3 loops that show poor structural alignment with the natural proteins are also poorly aligned between the different natural PDZ domains. These regions lack long-range NOEs, suggesting that they might not be ordered in solution. Changes in the β1-β2 loop upon peptide binding have been demonstrated by Sharma (2006), and have been implicated as a potential physical mechanism linked to the SCA pattern of this protein family that served as the basis for this design project. Future work may further examine the dynamics of these regions.

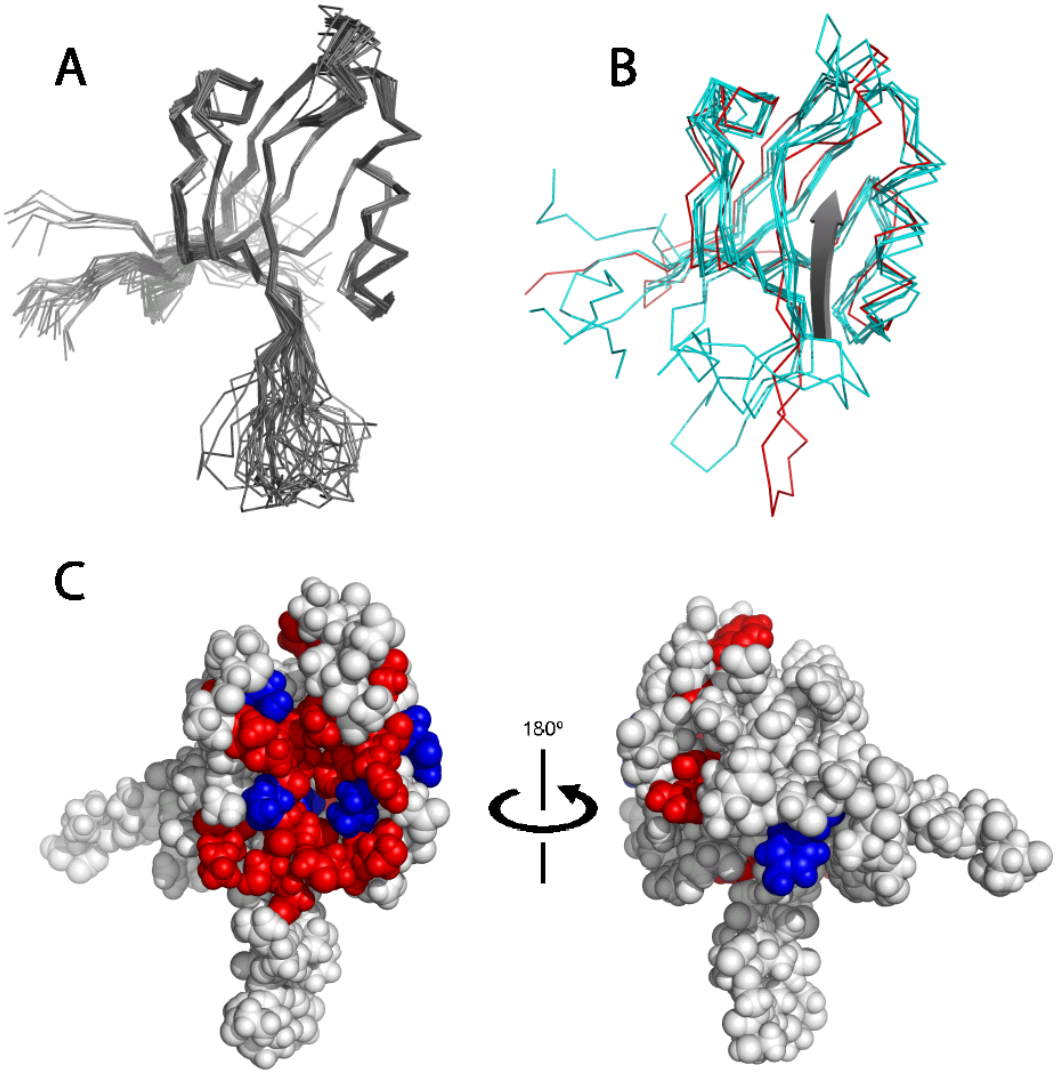


Figure 3.4: Solution structure of C_{60}-1
A) 25 lowest energy models of C_{60}-1.
B) Alignment of the energy minimized average structure of C_{60}-1 (red) and several natural PDZ domains (cyan, 1BE9 - Doyle et al., 1996, 1G9O - Karthikeyan et al., 2001, 1IHJ - Kimple et al., 2001, 1MFG - Birrane et al., 2003, 1NTE - Kang et al., 2003). The ligand of PSD-95 in the canonical binding site is shown as a grey arrow.
C) Residues with backbone amides in slow exchange on titration of TrpV6 ligand are shown in red. Residues that could not be classified due to spectral overlap are in blue. Orientation of the left figure is the same as in panels A and B.

RESTRAINTS

Unambiguous NOEs		
Total:	1635	
Intra-residue (*i*=*j*):	675	
Sequential (|*i*-*j*|=1):	378	
Short range (1<|*i*-*j*|<5):	152	
Long range (|*i*-*j*|≥5):	430	
Ambiguous NOEs:	83	
Hydrogen bonds:	32	(64 restraints)
Dihedrals:	102	(51 phi/psi pairs)

RAMACHANDRAN ANALYSIS (PROCHECK)

Entire protein

Most favored regions:	1781	77.4%
Additional allowed regions:	492	21.4%
Generously allowed regions:	27	1.2%
Disallowed regions:	0	0.0%

Ordered regions (res. 10-20, 26-33, 45-103; excludes termini, β1-β2 and β2-β3 loops)

Most favored regions:	1434	85.6%
Additional allowed regions:	235	14.0%
Generously allowed regions:	6	0.4%
Disallowed regions:	0	0.0%

ENERGIES

Bond:	2.27 (± 0.16)
Angle:	35.15 (± 0.98)
Improper:	3.57 (± 0.25)
Dihedral:	471.64 (± 3.47)
van der Waals:	-859.95 (± 9.10)
NOE:	4.28 (± 1.10)

Mean RMSD

Backbone, secondary structure:	0.309 Å
Heavy atoms, secondary structure:	0.658 Å
Backbone, all residues:	1.727 Å
Heavy atoms, all residues:	1.972 Å

Table 3.1: Structure statistics for C_{60}-1

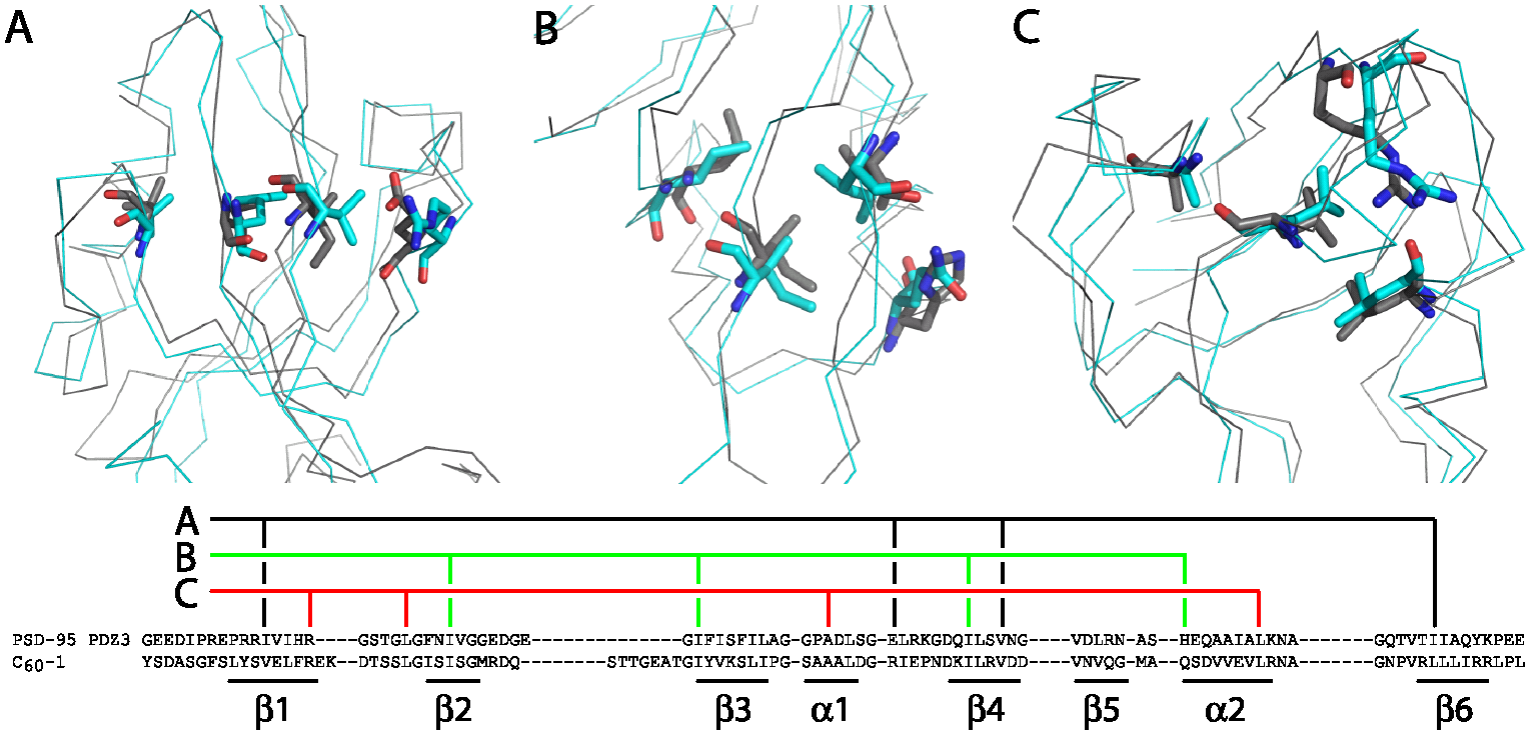


Figure 3.5: Structural comparison between C_{60}-1 and the third PDZ domain of PSD-95
Each panel shows several residues in C_{60}-1 (cyan) with unambiguous sidechain NOEs, and the corresponding residues in the canonical third PDZ domain of PSD-95 (grey, Doyle et al., 1996) based on the sequence alignment resulting from the design algorithm. Residues were chosen to highlight interactions between different secondary structural elements, and to confirm the alignment flanking the β2-β3 loop (panel B) and the β1-β2 loop (panel C). The residues shown in each panel are indicated in the sequence alignment.

Binding activity of SCA-designed PDZ domains

Since some SCA-designed proteins are folded, I next sought to determine whether or not these proteins also exhibit the function of natural PDZ domains: binding the C-terminus of target proteins. To answer this question, Michael Stiffler from Gavin McBeath's laboratory at Harvard University collaborated on this project. Their lab has in place a high throughput fluorescence polarization assay to screen for binding between a PDZ domain and 217 target peptides from the mouse genome. While it is possible that some of the designed PDZ domains might have target specificities that are not covered by this library of ligands, the relatively loose selection criteria found by Songyang et al. (1997) in which a few key residues are important for determining binding specificity, and the fact that PDZ domains generally fall into families with similar binding specificity indicates that such a screen would likely include ligands that are able to bind to the designed PDZ domains if they are capable of binding at all.

The results of the screen are summarized in table 3.2. Four of the five proteins express in sufficient yield to assay, and three of these show binding with low- to mid-micromolar affinities typical for PDZ domains. C_{60}-1 binds specifically to one peptide from the library: TrpV6, which has a tyrosine at the -2 (third from last) position and falls in the group II class of peptides. This specificity is consistent with the fact that C_{60}-1 has a glutamine at the first position of the second alpha helix – the primary determinant of binding specificity for the peptide's -2 position – and the natural PDZ domain AF-6 with a glutamine at this position also shows group II specificity (Songyang et al., 1997). C_A-9 binds to several peptides, which is not unusual for natural PDZ domains (Stiffler et al.

2006). Surprisingly, C_A-9 has a histidine at the α2 1 position, which typically indicates group I specificity, but most of its target peptides do not have a threonine or serine at the -2 position to be classified as group I peptides. However, there are natural PDZ domains that also defy the canonical classification system. For example, Shank3 binds class I peptides Cnksr2 and Dlgap1/2/3 as well as the class II peptide AN2 with tight affinities for a PDZ domain (Stiffler et al. 2006).

Protein	**Target peptide**	**Affinity**
C_A-2	QKNKDKEYYV (Neurexin 3)	68 µM
C_A-9	IDESKKEWLI (Caspr2)	39 µM
	VGENQKEYFF (Caspr4)	38 µM
	HTHSYIETHV (Cakrr2)	69 µM
	GDTSKKEYFI (Glycophorin C)	47 µM
	QKNKDKEYYV (Neurexin 3)	89 µM
	AHFSSLESEV (NMDAR2D)	97 µM
	SGAEDIIAWV (SSTR2)	51 µM
	EDGEGWEYQI (TrpV6)	70 µM
C_A-18	no hits	
C_B-1	insufficient expression	
C_{60}-1	EDGEGWEYQI (TrpV6)	15 µM

Table 3.2: Binding by SCA-designed PDZ domains

I next investigated the strongest of these interactions: C_{60}-1 binding to TrpV6. While C_{60}-1 binds specifically to this peptide, an important question is whether this peptide has a high probability of nonspecifically binding to many proteins. Michael Stiffler performed a screen of 158 natural PDZ domains against the TrpV6 peptide and found that only eighteen bind to TrpV6 with <50 μM affinity, indicating that this peptide does not have a high probability of nonspecific binding to PDZ domains.

To further characterize this binding interaction, Pilong Li performed titrations of TrpV6 with C_{60}-1 while recording ^{15}N-HSQCs. This shows a mixture of crosspeaks that are in fast exchange (crosspeaks that continuously move from the starting to ending coordinates over the course of the titration) and slow exchange (crosspeaks that shift intensity from a peak at the starting coordinates to a peak at the ending coordinates without peaks at intermediate coordinates). The crosspeaks in slow exchange are those with the largest chemical shift changes upon peptide binding, and would be expected to lie predominantly in the peptide binding pocket. Mapping the crosspeaks in slow exchange onto the structure of C_{60}-1 (figure 3.4 C) shows that they form a distinct pocket, consistent with specific binding rather than nonspecific aggregation. This pocket is the same binding site used by natural PDZ domains (figure 3.4 B and C). This indicates that not only is this specific binding, but binding is taking place with the same mechanism used by natural PDZ domains.

DISCUSSION

SCA based design can clearly be extended to protein families larger and more complex than the WW domain, as several of the SCA-designed PDZ domains are folded. The structure of one of these confirms that it not only recapitulates the overall PDZ fold, but the position and orientation of each sidechain is in agreement with expectations based on the sequence alignment with natural proteins whose structures have been solved. This is noteworthy, as no structural information is used in statistical coupling analysis. Three of the designed domains also exhibit binding activity, with affinity and specificity in the same range as natural PDZ domains. Furthermore, the success of SCA-based design can be attributed entirely to the coupling information incorporated in the design process rather than the associated increase in sequence identity, as proteins designed at the same level of sequence identity without coupling information fail to fold.

This work shows that complexity, as judged by intuitive measures such as protein size and the density of inter-residue interactions, is not a fundamental impediment to generating folded and functional proteins using a coupling matrix that is only sparsely populated by strong coupling interactions among many weak ones. In fact, many of the residues that participate in extensive packing interactions in the core of the protein are not highly coupled (figure 3.6). The lack of strong coupling interactions for many of the core positions does not indicate that these sites do not contribute to the protein's stability; they often show site-specific conservation patterns indicating that they are under evolutionary pressure. However, their lack of coevolution with other residues indicates that their contribution can be well approximated by considering it to be intrinsic to that site without

taking into account the identity of other residues. This surprising result is evidenced by the fact that the design algorithm succeeds while taking exactly that approach.

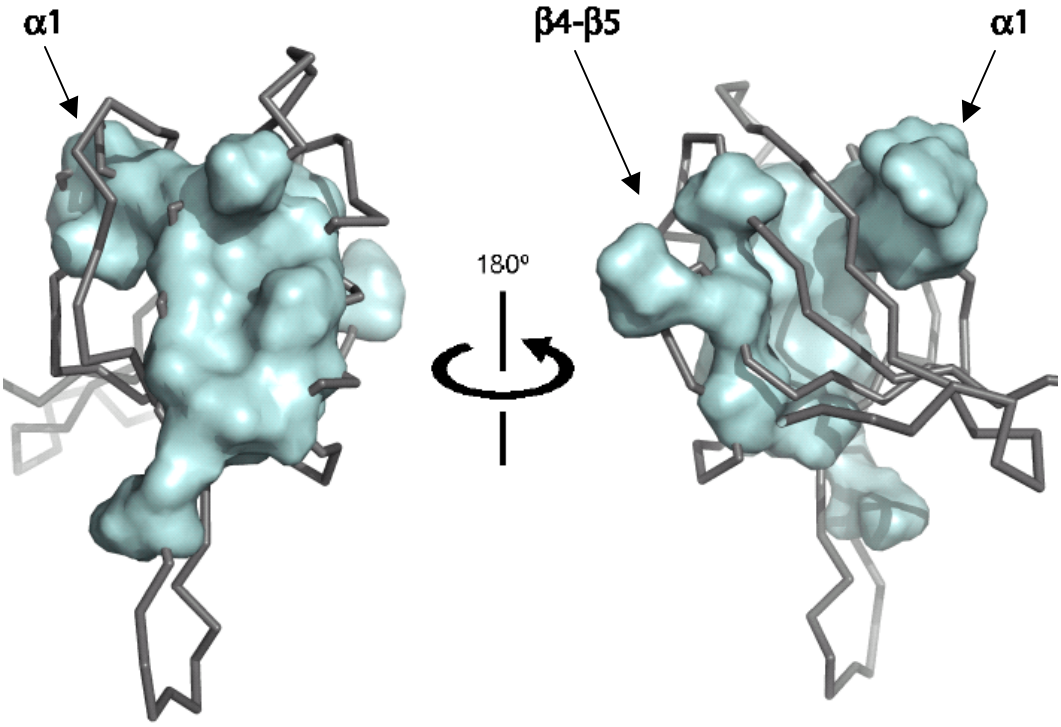


Figure 3.6: The coupled network of the PDZ domain
The highly coupled residues mapped onto the structure of C_{60}-1 lie primarily in the ligand binding pocket, with extensions to the α1 helix and the β4-β5 hairpin. Note that many residues in the core of the protein are not strongly coupled. Orientation on the left is the same as in figure 3.4 A and B.

METHODS

Sequence design

The SCA-based design algorithm operates as described previously (Socolich et al., 2005, Larson and Ranganathan, manuscript in preparation) except that SCA values are calculated as described in (Sharma, 2006). Briefly, the algorithm begins with the natural sequence alignment and takes steps by randomly selecting two residues from the same column in different sequences. During an initial shuffling phase, these two residues are swapped regardless of the impact on the alignment's coupling pattern, and the coupling pattern is destroyed after many such steps. During convergence, the algorithm follows a simulated annealing protocol in which swaps are accepted or rejected based on their effect on the system's energy (defined as the difference between the designed and natural alignments' coupling patterns) and an exponentially decreasing temperature parameter that scales the probability of accepting a step as $e^{-\Delta Energy / Temperature}$.

Folding assays

Genes encoding the designed sequences were constructed by PCR of partially overlapping oligonucleotides and cloned into an expression vector (pHis8.3, Socolich et al., 2005) adding an N-terminal His-tag and thrombin cleavage site and a C-terminal leucine and glutamate. For size exclusion chromatography screens, proteins were expressed in 100 ml cultures of TB induced at 18º C overnight with 250 μM IPTG. Cells

were sonicated and protein purified on 100 µl Ni-NTA (Qiagen) following the manufacturer's instructions, and dialyzed into PBS pH 7.4 with 1 mM DTT prior to size exclusion chromatography on a superdex 75 column (GE). For HSQCs, protein was expressed in 1 L cultures of ^{15}N labeled M9 induced as above, purified with 500 µl of Ni-NTA, dialyzed into 50 mM NaCl, 25 mM $NaPO_4$ pH 7.0, further purified with size exclusion chromatography, and concentrated to ~0.5 ml for ^{15}N-HSQCs.

Structure determination

Structure determination of C_{60}-1 was performed using 1 mM ^{15}N, ^{13}C double labeled samples in 50 mM NaCl, 25 mM $NaPO_4$, 0.02% azide pH 6.5 at 15º C with spectra recorded at 500 MHz and 600 MHz. Spectra used for chemical shift assignment were CBCA(CO)NH, HNCACB, HNCO, C(CO)NH-TOCSY, H(CCO)NH-TOCSY, HCCH-TOCSY, 2-D TOCSY, (HB)CB(CGCD)HD, (HB)CB(CGCDCE)HE, ^{15}N-NOESY-HSQC, and ^{13}C-HSQC-NOESY, plus a CT-^{13}C-HSQC of a sample labeled with 10% ^{13}C-glucose. Spectra were processed with nmrPipe (Delaglio et al., 1995) and analyzed with NMRView (Johnson and Blevins, 1994). Structures were calculated with aria 1.2 (Linge et al. 2003) and CNS (Brunger et al. 1998) using partially manually assigned ^{15}N-NOESY and ^{13}C-NOESY peaklists, dihedral restraints from talos (Cornilescu et al. 1999), and hydrogen bond restraints from an HSQC of an ^{15}N-labeled sample lyophilized and resuspended in D_2O.

CHAPTER FOUR
SCA-based GPCR design

INTRODUCTION

The sufficiency of SCA to specify the fold and function of the PDZ domain indicates that it is capable of capturing the design constraints of a well-packed hydrophobic core and the functional interactions required for specifically binding target proteins, and it raises the question of how much complexity can be encapsulated with coupling information. The next logical step is to test the information content provided by SCA on a more complex protein family, both in terms of packing a folded tertiary structure and in recapitulating a complex biological activity. Also, statistical coupling analysis was originally developed to examine not packing interactions, but energetic pathways mediating allosteric effects (Lockless and Ranganathan, 1999), suggesting that the methodology may be particularly well suited for this design goal. While SCA calculations are inherently free of physical mechanisms, success in SCA-based design of allostery could help direct mechanism-based design efforts.

An attractive design target is the major *Drosophila* rhodopsin, Rh1. This protein is a member of the class A GPCR family, interesting both from a basic science standpoint as a well-studied allosteric protein family and medically as members of this protein family are commonly targets of pharmaceuticals. Rh1 is an experimentally tractable member of this protein family. It is transiently modified with N-linked glycosylation while in the endoplasmic reticulum that is removed on export to the rhabdomeres (Huber et al., 1990), so the glycosylation state of the designed proteins can signal whether they have passed the endoplasmic reticulum's quality control checkpoints for folding. Rh1 is

expressed at extremely high concentrations in the photoreceptors, accounting for about 65% of total membrane protein (Paulsen and Schwemer, 1979), while the light response produced by activation of even a single molecule of Rh1 can be detected experimentally (Ranganathan et al., 1995). In addition to its canonical activity as a photoreceptor, Rh1 interacts with the cyclophilin homolog NinaA during maturation (Colley et al., 1991), binds its ligand 3-hydroxyretinal as a required step in maturation (Ozaki et al., 1993), and provides a stimulus that preserves the morphology of the rhabdomeres (Kumar and Ready, 1995). All of these can serve as further tests of which, if any, functions can be specified by coupling information.

RESULTS

Sequence design

In this project, sequences were generated using the information from an alignment of class A GPCRs using the perturbation-based design algorithm with 100 statistical perturbations. Figure 4.1 shows that the algorithm successfully converges for this protein family: the alignment is initially shuffled to destroy the coupling pattern, and it is gradually rebuilt with the original coupling pattern restored.

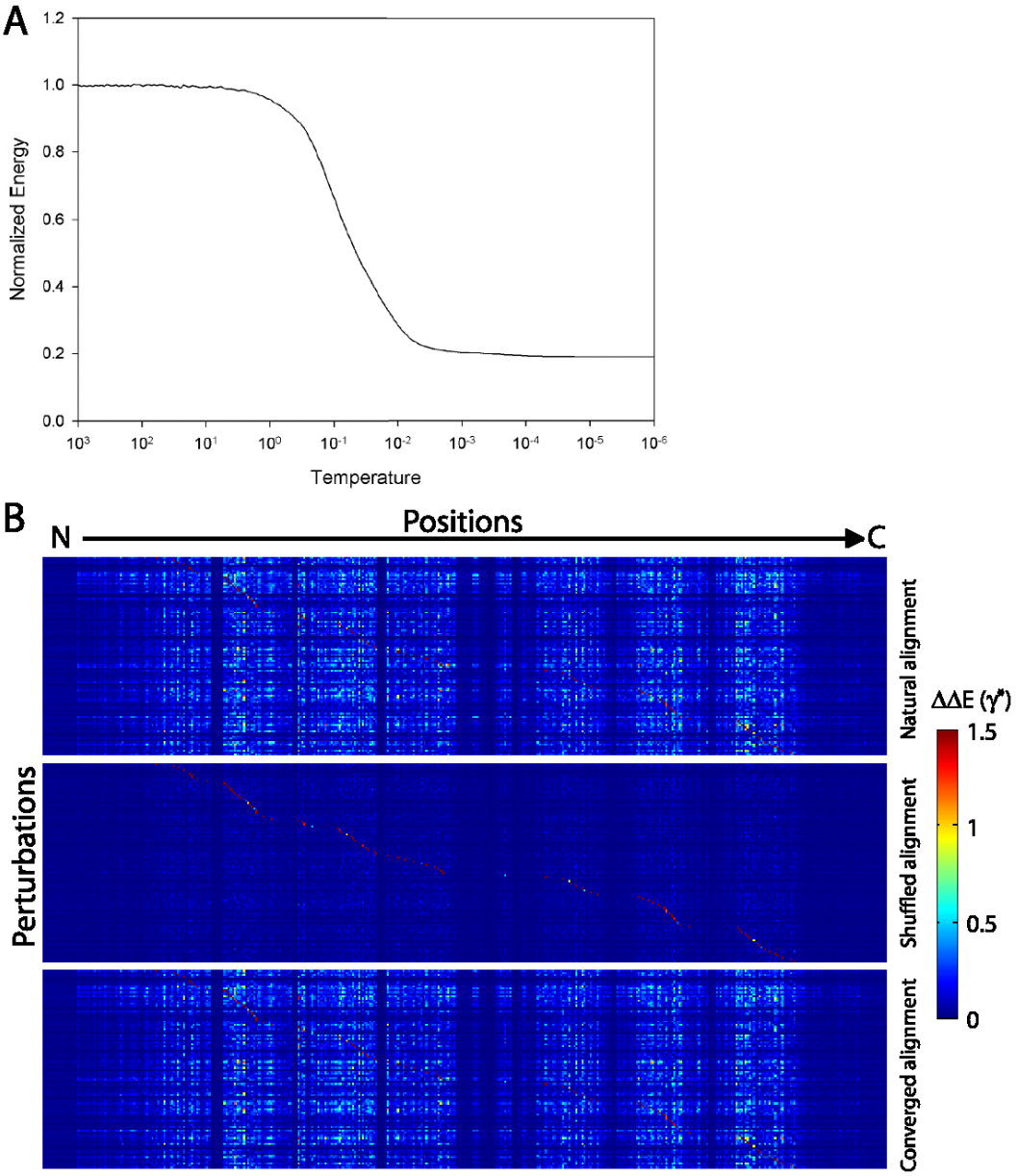


Figure 4.1: SCA-based design of a GPCR alignment

A) The energy, defined as the difference between the natural and designed alignments' coupling patterns, plotted against the temperature, a parameter that decreases exponentially as the algorithm progresses. Each column of the alignment is initially shuffled, destroying the coupling pattern and producing the same energy as an alignment made with only conservation information. The energy decreases as the simulated annealing protocol imposes the coupling pattern on the designed alignment.

B) Matrices of coupling values show that the coupling pattern of the natural GPCR alignment is destroyed when the algorithm initially shuffles each column, and is restored after the algorithm converges.

Because this design strategy uses an alignment of class A GPCRs of all types, but the goal of this project is to recapitulate the activity of a particular receptor, it was necessary to identify those sequences that were most likely to have been designed with Rh1-like characteristics. Based on the observed success of PCA in segregating WW domains (see chapter 2) and members of other protein families (Casari et al., 1995) into subfamilies based on their functional specificity, I applied the same approach to the GPCR sequences. The mapping of the natural GPCRs is shown in figure 4.2.

The most striking feature is the separation of the opsins from the other receptors along the first principal component. The vertebrate rhodopsins cluster at the extreme end of the map, the cone opsins fall between the rhodopsin cluster and the other receptors, and the fly and other invertebrate opsins lie near the large cluster of other receptors but slightly separated from them in the direction of the rhodopsins. The large difference between the vertebrate and invertebrate opsins is consistent with their differing activity: vertebrate opsins are coupled to G_t while invertebrate opsins are coupled to G_q (Lee et al., 1990), vertebrate opsins release photoisomerized retinal after activation while invertebrate opsins do not (Hardie, 1986), and vertebrates use retinal as their chromophore while *Drosophila* and some other flies use 3-hydroxyretinal (Goldsmith et al., 1986). The fact that the fly opsins form a distinct cluster makes this a potentially useful approach for identifying designed sequences with fly opsin-like properties. Furthermore, other receptor types also lie in clusters under this projection. The fact that many receptor subtypes cluster in ways that reflect their biological activities makes this an attractive approach for classifying sequences based on their probable activity.

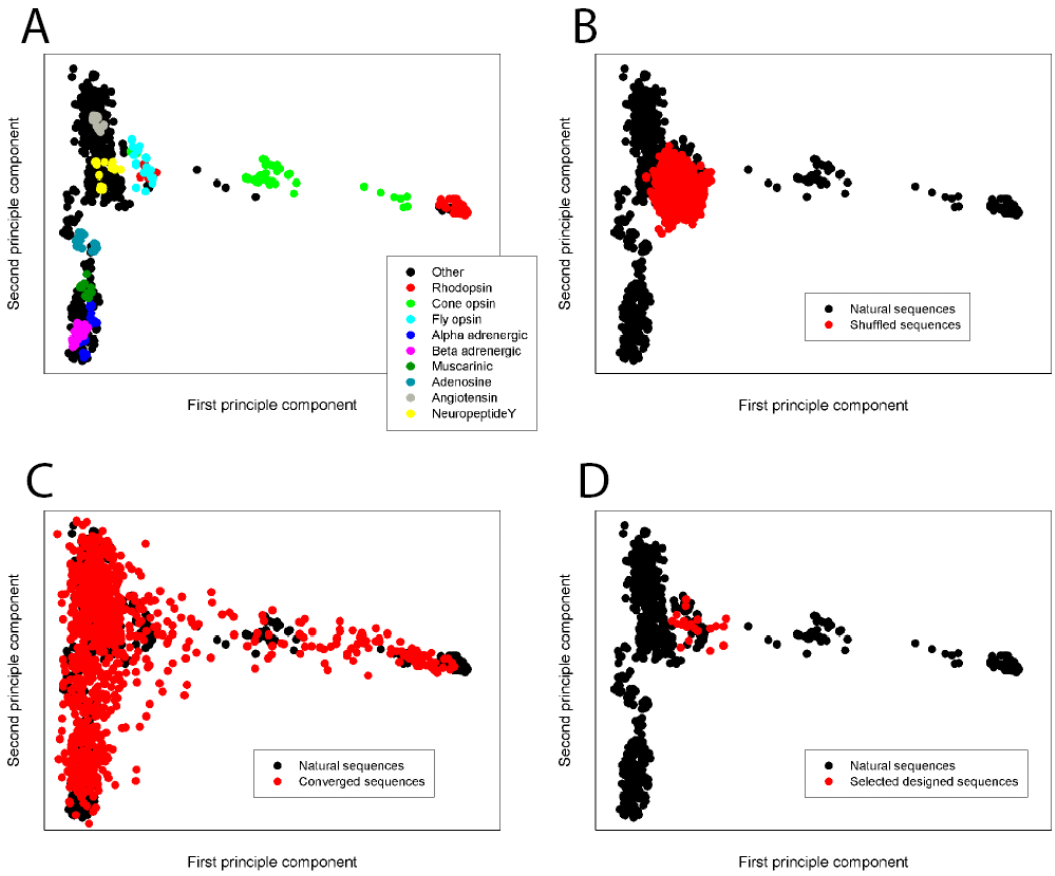


Figure 4.2: PCA mapping of GPCR sequences

A) PCA mapping of the natural GPCR sequences. Each circle represents a particular sequence from the alignment. Many are colored based on their annotated function, showing that functionally related proteins map to clusters.

B and C) PCA mapping of sequences after shuffling the alignment (B) or converging on SCA information (C), using the same projection as with the natural sequence alignment.

D) PCA mapping of the 20 selected designed sequences.

To classify the designed sequences, I applied the transformation matrix calculated from the natural GPCR alignment to project the designed sequences on the same map. Interestingly, at the beginning of the design process, after the alignment has been shuffled but before it has begun to converge on the coupling values, the designed sequences fall in a homogeneous zone in the center of the map without any distinct clustering (figure 4.2 B). As the algorithm converges, the designed sequences redistribute on the map to cover the same area in a similar pattern as the natural sequences, suggesting that this mapping shows how the sequences differentiate into their subfamilies during the design process (figure 4.2 C). Examination of the designed sequences that were ultimately selected from the fly opsin area confirms that functional elements such as the lysine that forms a Schiff base with the chromophore, and the (D/E)RY and (N/D)PxxY motifs involved in signal transduction and endocytosis are preserved in the designed constructs. Although the (D/E)RY and (N/D)PxxY motifs are highly conserved in the alignment, they are not perfectly conserved: a sequence made with conservation information alone would have only a 68% chance of having both elements. The Schiff base lysine is present in only 15% of sequences in the GPCR alignment.

Rh1	...MCMISL**DRY**QV...WGACFA**K**SAACY**NPIVY**GISH...	ID to Rh1
Construct 1	...MTAIAL**DRY**VV...LAAVFA**K**SSTMY**NPVIY**GILN...	44%
Construct 2	...MFIIAV**DRY**NV...WFALFA**K**LVAVF**NPIVY**GISH...	52%
Construct 3	...MVLIAI**DRY**NV...LCSLFA**K**LSTVY**NPIIY**VLSN...	34%
Construct 4	...MTMIAM**DRY**NV...WPAFFA**K**ALTVY**NPIIY**AISH...	48%
Construct 5	...MAIIAY**DRY**NV...VPALFA**K**GVTVY**DPILY**GVSN...	44%
Construct 6	...MTMVSF**DRY**MV...WPAVLA**K**SNTVY**NPIIY**VYSD...	36%
Construct 7	...NALITY**DRY**NV...IPALLA**K**LVACV**NPLVY**AVSH...	39%
Construct 8	...LAVIAF**ERY**NI...LPAYFA**K**STTVY**NPIVY**PLMN...	35%
Construct 9	...NTVIAI**ERY**IV...IPAVTC**K**SVACY**NPIIY**SLNH...	34%
Construct 10	...GAMISL**DRY**LS...LSACFA**K**WNSVF**DPIVY**VISH...	36%
Construct 11	...MSMIAV**DRY**NV...VPACFA**K**SNAVI**DPIVY**GIVH...	34%
Construct 12	...MVVMTV**DRY**LA...ILVLTA**K**ASTCF**NPIIY**GKLH...	36%
Construct 13	...MAVIAL**DRY**NG...FLALFT**K**LVAVY**NPLLY**AISH...	42%
Construct 14	...MTAMAF**DRY**NA...WIAFFA**K**SSAAY**NPCIY**GILH...	46%
Construct 15	...NTMIAW**DRY**NV...IPAVFA**K**ASAVY**NPIVY**GISH...	42%
Construct 16	...MTFIAC**DRY**NV...LPAVFA**K**SVAMY**NPIIY**VISH...	41%
Construct 17	...NTAIAL**DRY**NS...WPALPS**K**ANAVY**NPIIY**TISN...	35%
Construct 18	...MTFIAY**DRY**NV...IPALFC**K**ASAVY**DPIVY**AISH...	46%
Construct 19	...NTMIAI**DRY**NV...APALAA**K**FVACH**NPLVY**TISH...	42%
Construct 20	...MASIAL**DRY**NV...IFALFY**K**ISALV**NPIVY**TISN...	46%

Table 4.1: Functional elements in the SCA-designed sequences

Although N-linked glycosylation of the first loop of Rh1 is required for efficient receptor maturation (Katanosaka et al., 1998, Webel et al., 2000), the glycosylation sites of natural fly opsins do not fall in the same column of the alignment. Because this signal sequence is not an aligned feature, it would not be expected to be detectable with SCA. Consistent with this, many of the SCA-generated sequences do not have an N-linked glycosylation site. Because the goal of this project is to determine whether SCA can specify the fold and function of the protein, and because this is an unaligned feature that would not be expected to be specified with this analysis and will severely reduce expression of mature protein if absent, I included as a selection criteria that the designed sequence must contain an N-linked glycosylation site before the first transmembrane helix. Finally, a ten residue segment of the third cytoplasmic loop of Rh1 that is not

found in the other GPCRs could not be included in the sequence alignment. Those ten residues were maintained unaltered in the designed constructs.

Expression of SCA-designed proteins

The selected designed sequences were transformed into *ninaE*^{I17} (Rh1 null) *Drosophila* with an epitope tag consisting of the C-terminal 17 residues from Rh1 recognized by a widely used anti-Rh1 monoclonal antibody (4C5). Transformed genes will be denoted as *P[1t]* for the epitope-tagged construct #1, and later *P[1]* for the construct without an epitope tag. Many of these tagged constructs show visible expression on western blots (figure 4.3 A), and show a mobility consistent with either a monomer or dimer. The appearance of dimeric populations of some constructs under reducing SDS-PAGE conditions is not completely unexpected; natural Rh1 also shows discrete populations of protein that migrate as monomers, dimers, and higher-order oligomers. Constructs 7t, 10t, 12t, 13t, 14t, 16t, and 19t consistently show expression on multiple blots, while the other constructs cannot be confidently distinguished from background. Expression levels of these constructs are lower than for wild-type Rh1, but are comparable to or higher than Rh1 in the background of *ninaA*^{269} – a null allele for an Rh1 chaperone which reduces protein levels to about 1% of wild-type. Construct 7t, the highest expressing of the SCA-designed proteins, shows expression at about 5-10% of wild-type levels.

Construct 7t shows two visible bands near the expected molecular mass of a monomer, with most of the protein in the higher molecular mass population. Rh1 is

known to be transiently glycosylated during its maturation, and treatment of construct 7t with endoglycosidase H shifts the protein to the lower molecular mass species (figure 4.3 B). This indicates that most of the protein is glycosylated with a high-mannose oligosaccharide and is retained in the endoplasmic reticulum, while a population of about 10% of the protein is in a deglycosylated state. The major populations of all other constructs also show a mobility shift upon endoglycosidase H treatment (figure 4.3 B), indicating that they have high-mannose glycosylation.

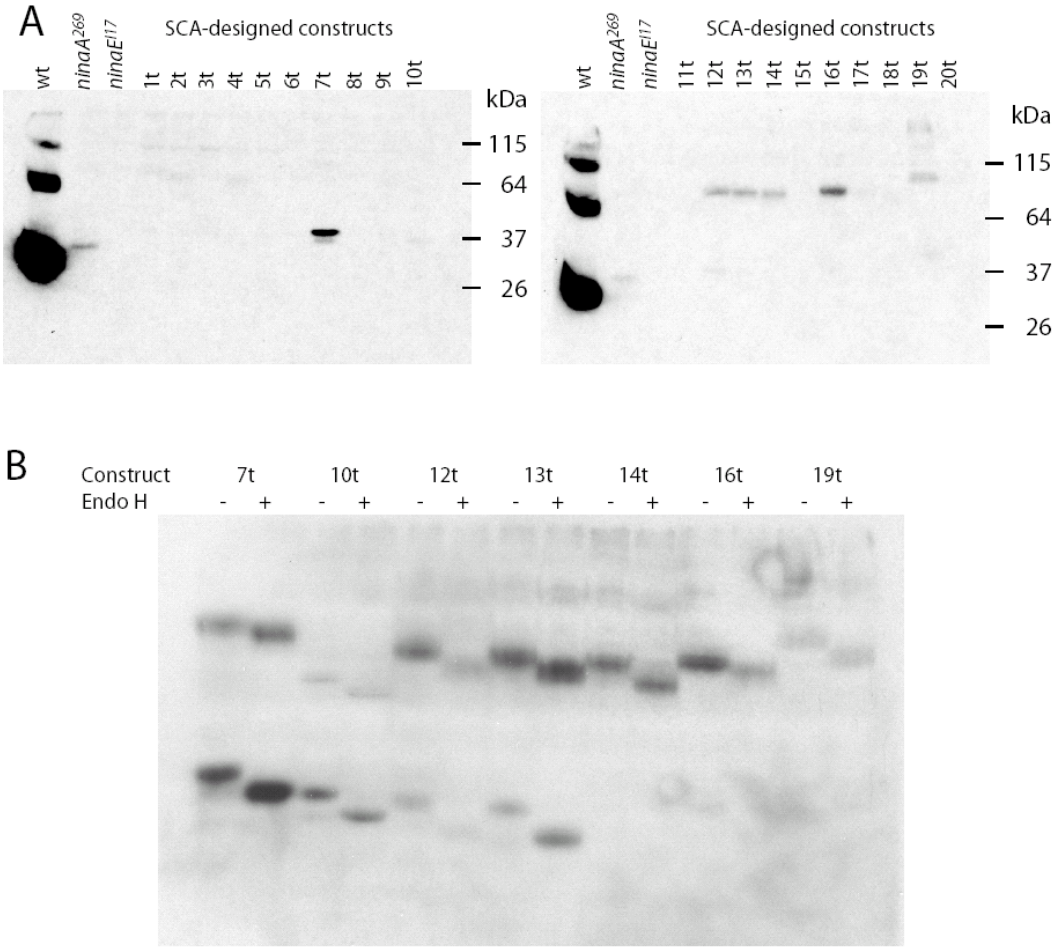


Figure 4.3: Expression of SCA-designed proteins

A) Fly head homogenates probed with anti-Rh1 antibody. Several epitope-tagged constructs show visible expression at levels lower than wild-type flies, but higher than flies expressing wild-type Rh1 in a *ninaA*^{269} background. Expected molecular masses of the designed constructs are 40-45 kDa. Like Rh1, many constructs show populations at a mobility consistent with dimerization under SDS-PAGE conditions.

B) The predominant populations of these constructs show a decrease in apparent molecular mass after treatment with endoglycosidase H, indicating that the proteins are mostly glycosylated.

Functions of SCA-designed proteins

NinaA interaction

NinaA is a cyclophilin homologue that is essential for efficient maturation of Rh1 to the deglycosylated state, and this interaction is specific – NinaA is required for maturation of Rh1 and Rh2 but not other *Drosophila* opsins (Stamnes et al., 1991), and has no other reported phenotypes. Because interaction with NinaA is a property specific for my design target, I examined whether this interaction is shared by the designed constructs. Construct 7t shows a consistently detectable subpopulation of deglycosylated protein, and I crossed these flies into a NinaA null ($ninaA^{269}$) background to determine if maturation is affected. There is a decrease in the levels of mature construct 7t in a NinaA null background (figure 4.4), indicating that NinaA promotes its maturation and that this protein-protein interaction is maintained in this designed construct.

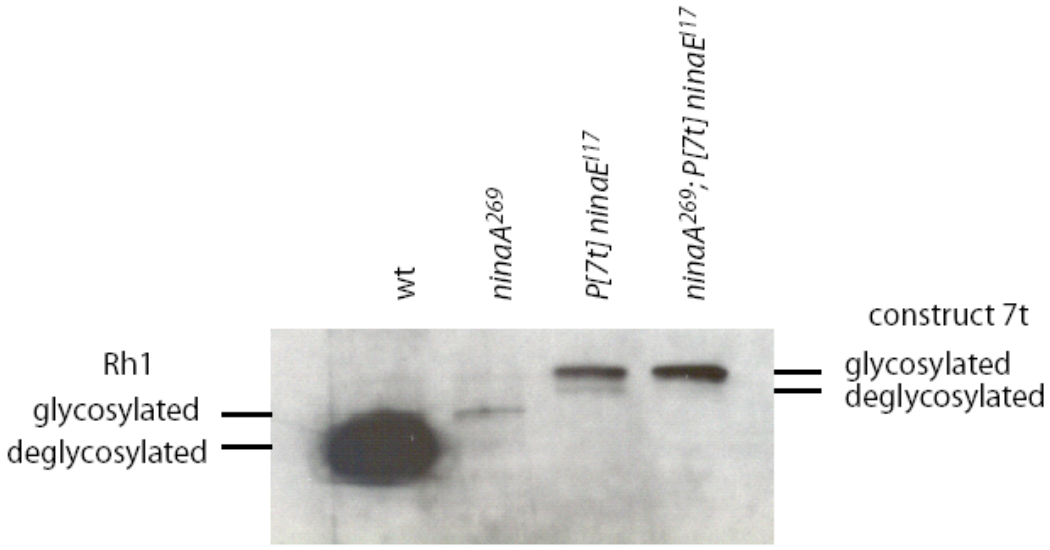


Figure 4.4: Construct 7t interaction with NinaA
As with Rh1, levels of deglycosylated construct 7t are dependent on NinaA. In the background of a NinaA null allele (*ninaA*^{269}), the deglycosylated species is reduced.

Ligand binding

Rh1 maturation is dependent on the presence of its ligand: 3-hydroxyretinal. In the absence of ligand, the protein is expressed but is not exported from the endoplasmic reticulum and is largely degraded, with the remaining population in the glycosylated state (Ozaki et al., 1993). To determine if this property is also seen in the designed proteins, flies were grown on either standard media or on yeast-sucrose media that lacks carotenoids. No constructs show a change in protein levels in response to retinoid deprivation (figure 4.5). This suggests that they do not bind chromophore, but does not rule out the possibility that they are able to bind retinal but are unaffected by its absence. I tested the possibility that construct 7t is able to bind retinal and initiate a light response with whole-cell voltage-clamped recordings of isolated photoreceptors. The light

responses from photoreceptors expressing construct 7t are indistinguishable from those seen in *ninaE*^{I17} flies when excited with very bright stimuli (figure 4.6; these responses presumably arise from expression of very low levels of the minor opsins, Hardie, 1996), indicating that this protein does not initiate a light response.

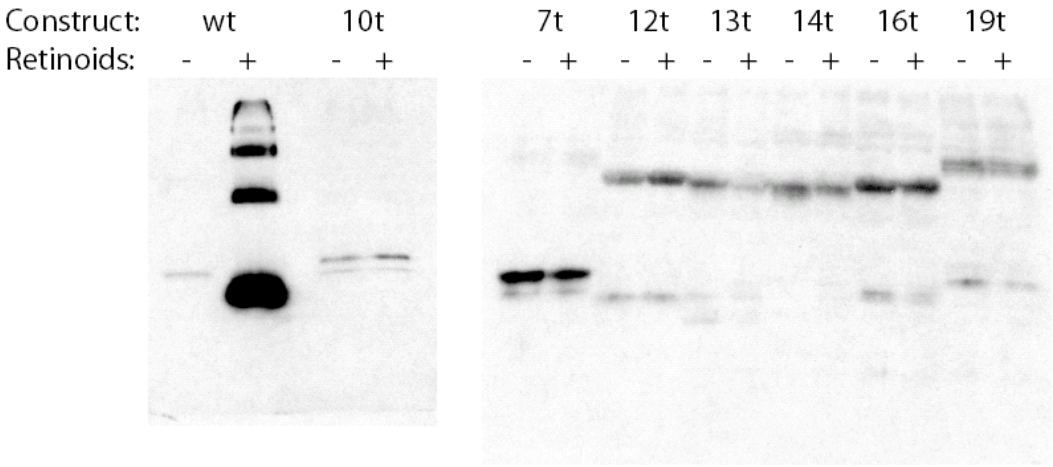


Figure 4.5: SCA-designed construct expression in retinoid deprived flies
None of the highly expressing SCA-designed constructs show the retinoid-dependent stability characteristic of Rh1.

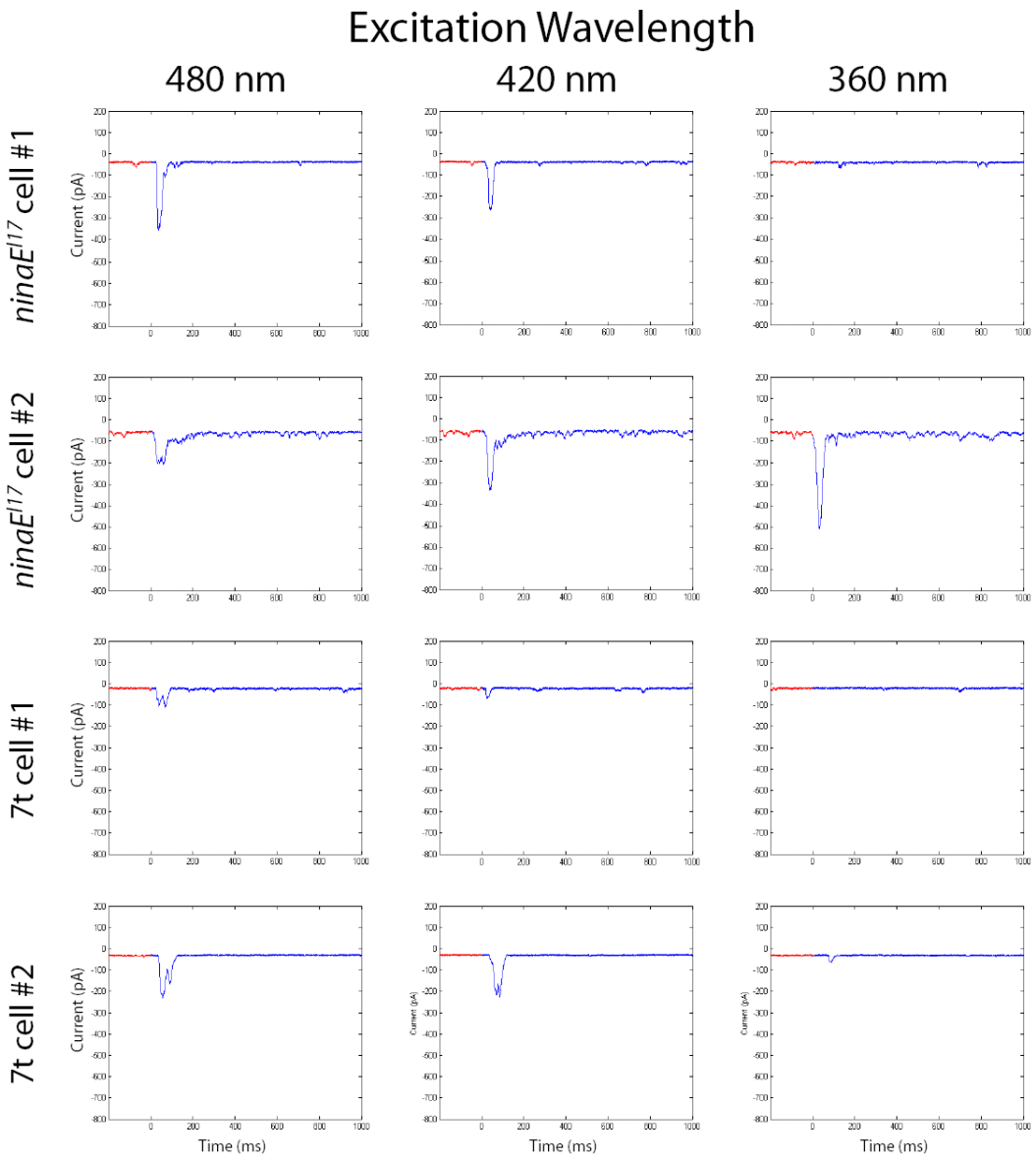


Figure 4.6: Light responses from *ninaE*^{l17} flies and flies expressing construct 7t. Two cells are shown per genotype. Traces show pre-stimulus baseline current in red followed by post-excitation responses in blue. Cells were excited with light from a xenon source attenuated only by the indicated wavelength filters. Downward deflections indicate inward cation current. Most but not all cells with normal resting potentials showed light responses to this intensity of stimulus. Responses were not seen when the stimulus was attenuated by more than one order of magnitude with neutral density filters.

Rhabdomere morphology

Rh1 expression is required for normal development of the rhabdomeres – organelles consisting of densely packed microvilli that house the proteins involved in phototransduction. Rh1 null flies show morphological abnormalities of the rhabdomeres that are apparent at eclosion (the transition from the pupal stage to the adult stage) (Kumar and Ready, 1995). I therefore tested the ability of the designed constructs to rescue rhabdomere morphology. Electron micrographs of wild-type and $ninaE^{I17}$ (Rh1 null) flies show that this degeneration phenotype is apparent within the first day post-eclosion (figure 4.7). At this early timepoint, degeneration of the rhabdomeres in $ninaE^{I17}$ flies is evident as many of the R1-R6 photoreceptors have much smaller rhabdomeres than wild-type, with a flattened, oblong shape and curtains of membrane extending from the subrhabdomeric border into the cell body. However, this phenotype can be rescued by expression of even very low levels of Rh1 (Leonard et al., 1992). Figure 4.7 shows that the low level of Rh1 expressed in $ninaA^{269}$ flies is sufficient to restore essentially normal rhabdomere morphology at eclosion. While the restored rhabdomeres are smaller than wild-type, their size is much larger than would be expected from a proportional decrease related to bulk protein expression (compare with figure 4.3).

Because the C-terminus of Rh1 that I have used as an epitope tag has itself been proposed to play a promoting role in maintaining rhabdomere morphology (Ahmad et al., 2007), I generated flies carrying many of the constructs without this tag. Electron micrographs of these flies show that all of the tested constructs rescue rhabdomere morphology at eclosion, demonstrating that these transgenes provide some degree of the rhabdomere-preserving function of Rh1 (figure 4.7).

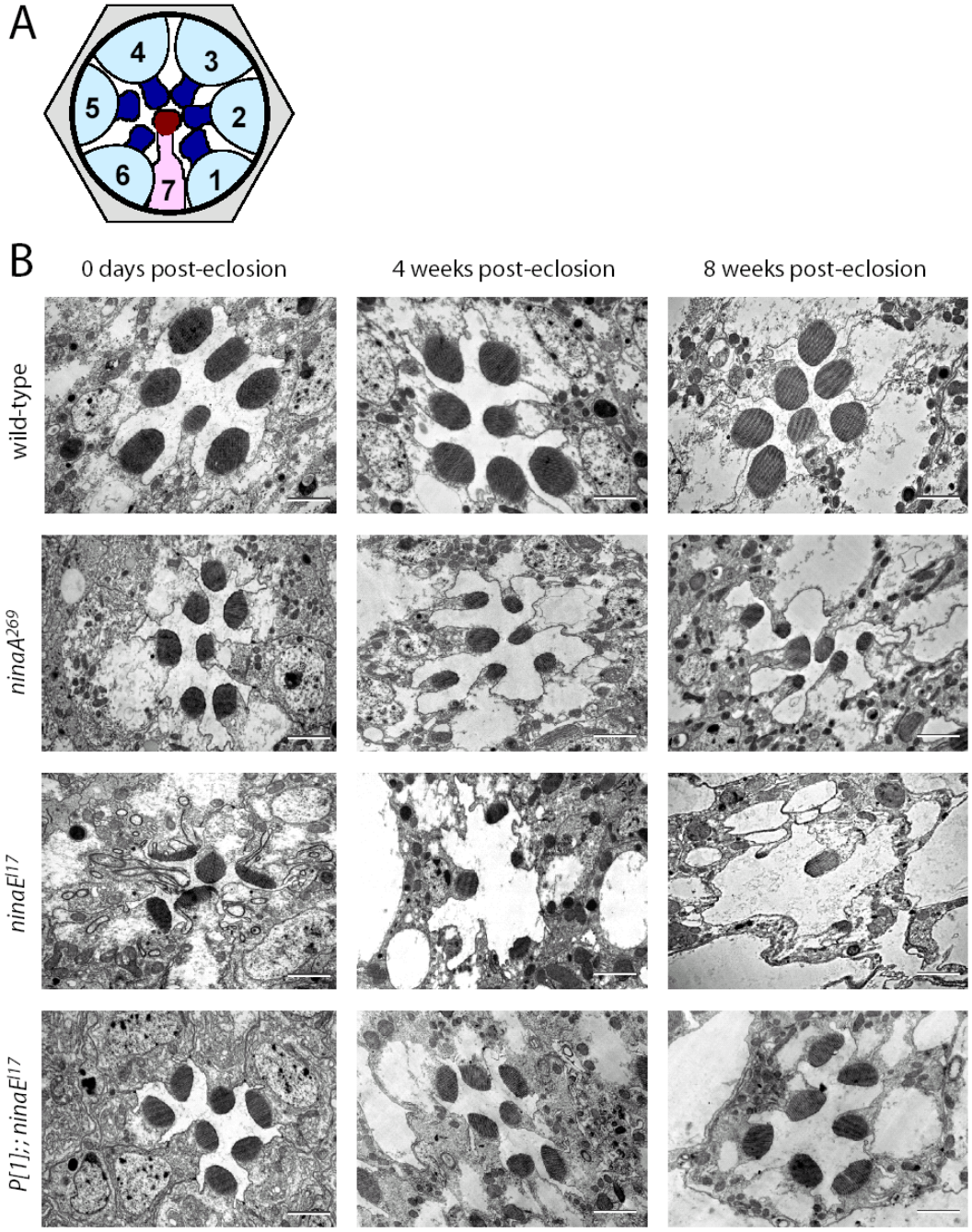


Figure 4.7: Rhabdomere rescue by SCA-designed constructs

A) Schematic of a cross-section of an ommatidium. Seven photoreceptor cells are visible per ommatidium. Cells R1-R6 express Rh1, or the designed construct in transgenic animals, while cell R7 expresses a different opsin. Rhabdomeres are shaded darkly.

B) Electron micrographs of experimental animals. Scale bar = 2 μm.

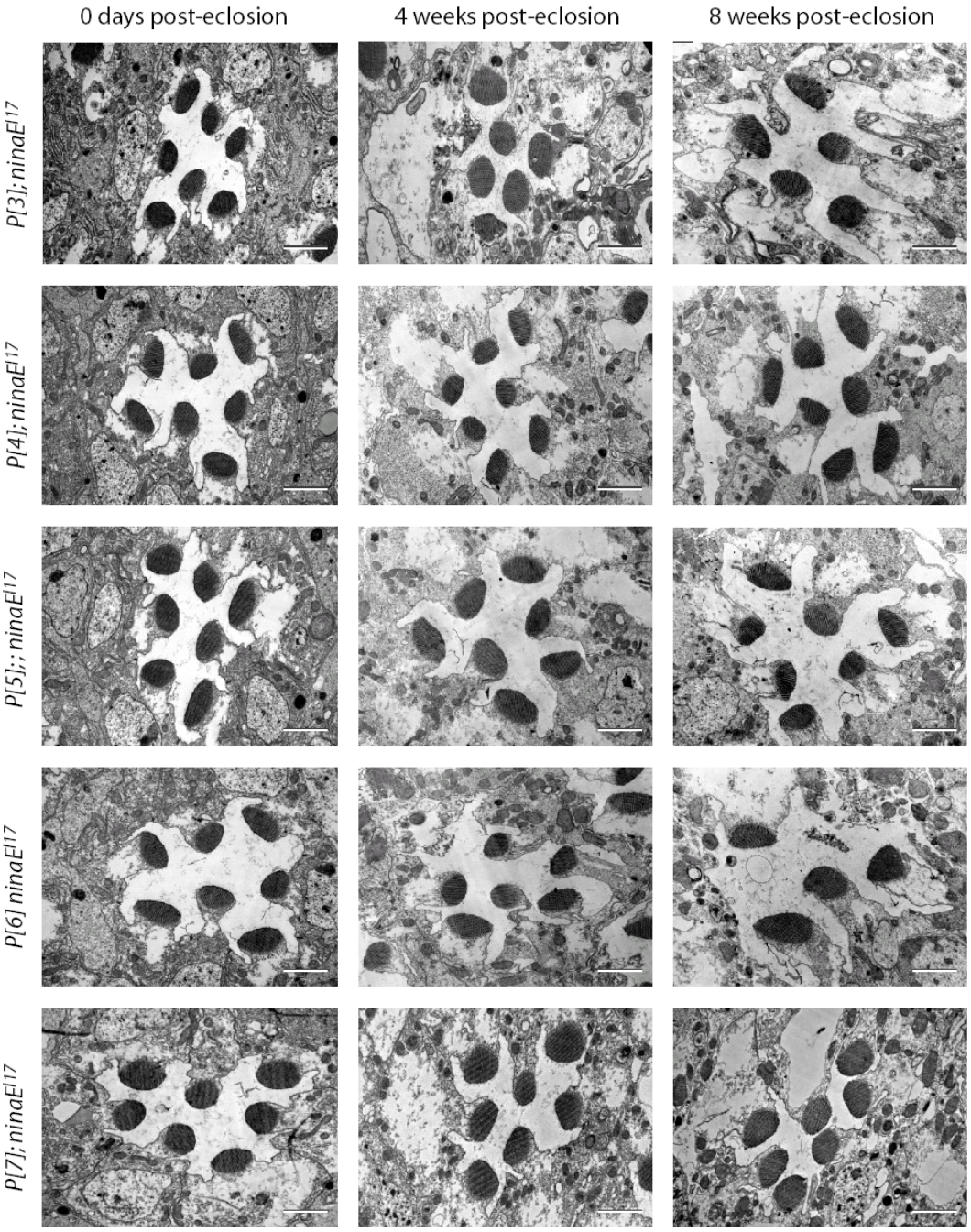


Figure 4.7 continued

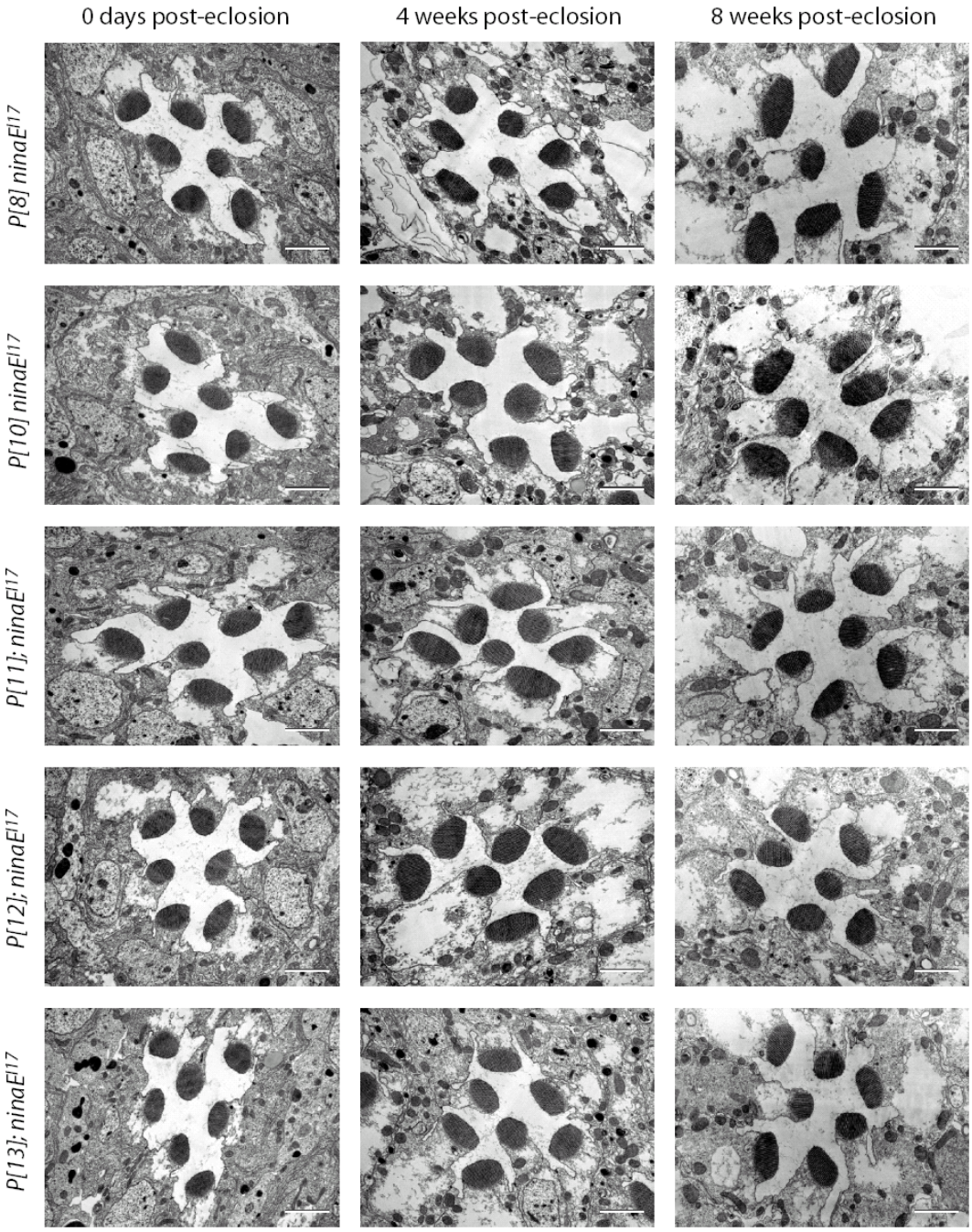


Figure 4.7 continued

In a study of hypomorphic *ninaE* alleles, Leonard et al. (1992) found that while all hypomorphs show some degree of rhabdomere rescue at eclosion, the size of their rhabdomeres and the persistence of rescue as the flies age both increase with increasing levels of Rh1. I therefore examined ommatidia of aged flies (figure 4.7). Wild-type flies maintain essentially normal size rhabdomeres to 8 weeks of age, near the end of the flies' lifespan. Rh1 null $ninaE^{I17}$ flies show essentially complete obliteration of the outer R1-R6 rhabdomeres that normally express Rh1, while only the R7 rhabdomere that expresses other opsins remains intact. $ninaA^{269}$ flies retain rhabdomeres even at 8 weeks of age, although like flies with hypomorphic alleles for Rh1 (Leonard et al., 1992) their rhabdomeres shrink with age. Flies expressing the SCA-designed constructs similarly show preserved rhabdomeres compared to $ninaE^{I17}$ even at these advanced ages, albeit to different levels (figure 4.7).

Because every SCA-designed construct surprisingly shows rhabdomere rescue at eclosion, I tested the possibility that the rescue is being mediated by the P-element rather than the designed sequences. I used flies with the same P-element carrying a nonspecific protein, InaD, rescuing an $inaD^{1}$ (InaD null) background. These flies express wild-type levels of InaD and have normal light responses (Mishra et al., submitted). In the presence of wild-type Rh1 ($ninaE^{+}$ allele), they show normal morphology at eclosion, indicating that the InaD rescue does not interfere with normal rhabdomere development (figure 4.8 A). In the presence of the Rh1-null allele $ninaE^{I17}$, they show the same degree of degeneration at eclosion as Rh1 null animals without the InaD P-element (figure 4.8 A), demonstrating that the P-element alone is not responsible for rhabdomere rescue.

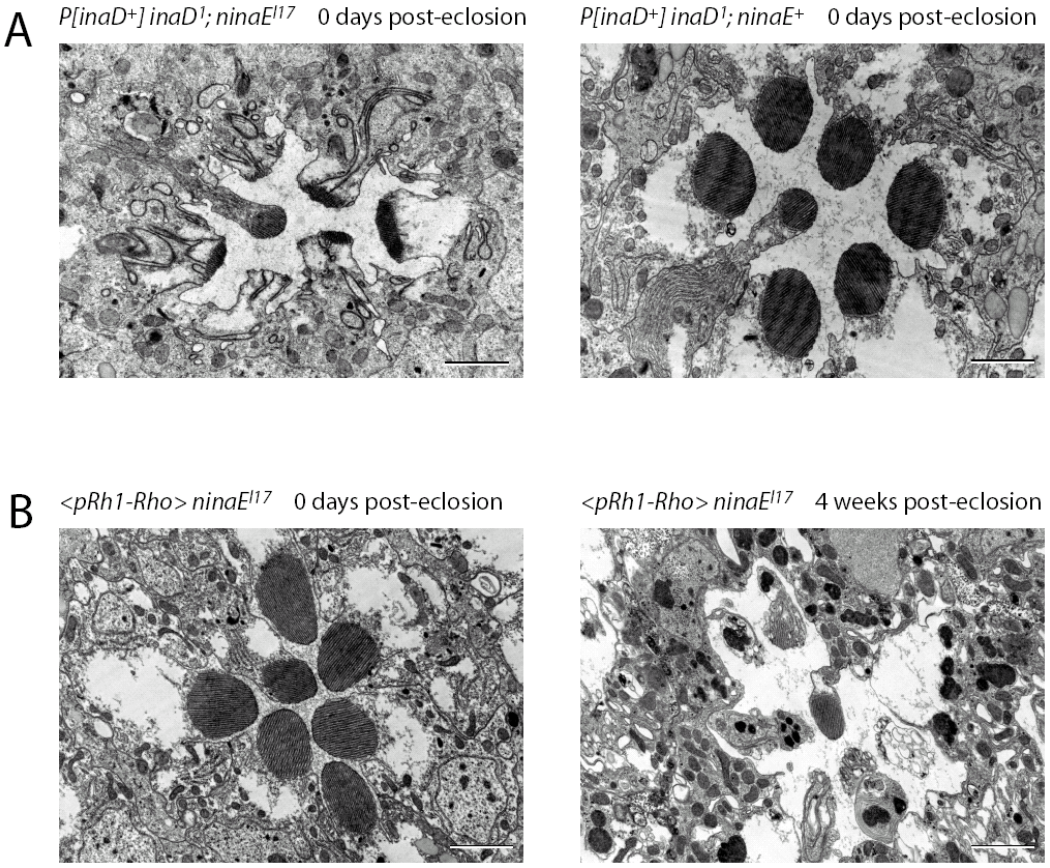


Figure 4.8: Rhabdomeres of flies expressing a nonspecific protein or bovine opsin
A) InaD carried on the same P-element as the SCA-designed constructs fails to rescue rhabdomere morphology at eclosion in Rh1 null flies (left). This is not due to a defect in the InaD rescue, as these flies have normal rhabdomeres in the presence of Rh1 (right).
B) Flies expressing bovine rhodopsin in an Rh1 null background show rhabdomere rescue at eclosion, but degeneration by 4 weeks post-eclosion. Scale bar = 2 μm.

While InaD does not rescue rhabdomere morphology, an interesting question is how similar a molecule must be to Rh1 in order to rescue the rhabdomeres. Ahmad et al. (2006) have addressed this by generating flies that express bovine rhodopsin in place of Rh1. Bovine rhodopsin is expressed in the *Drosophila* eye at levels comparable to Rh1, the protein is transported to the rhabdomeres, and the fly head homogenates of transformed but not wild-type animals show light dependent activation of G_t *in vitro* (Ahmad et al., 2006). All of these data indicate that bovine rhodopsin is expressed and natively folded. With regard to their rhabdomeres, these flies show normal morphology at eclosion, but the rescue is transient: Ahmad et al. (2007) found that the rhabdomeres subsequently degenerate within only two weeks, and I have independently examined rhabdomere rescue in these animals. While all SCA-designed proteins show substantial rhabdomere rescue at 4 weeks of age, bovine rhodopsin does not provide such a persistent morphological rescue (figures 4.7, 4.8 B).

Rhabdomere degeneration in constructs 3 and 6

Almost all flies with rescued rhabdomeres, both in this study and in the survey of Rh1 hypomorphs by Leonard et al. (1992), have a full complement of rhabdomeres for each of the R1-R6 photoreceptors that uniformly shrink until (in some hypomorphs) they disappear. Two exceptions to this are the SCA-designed constructs 3 and 6. Flies expressing these proteins show a full complement of R1-R6 rhabdomeres in each ommatidium at eclosion, indistinguishable from the other rescued flies. However, as these flies age, they seem to randomly lose some of their R1-R6 rhabdomeres (figure 4.9). At 4 weeks of age a minority of ommatidia contain a full complement of normal

rhabdomeres; most are missing at least one and many are missing several. Other SCA-designed constructs do not show such loss of rhabdomeres at this age, although some with low signal on western blotting have a milder loss of rhabdomeres at 8 weeks post-eclosion.

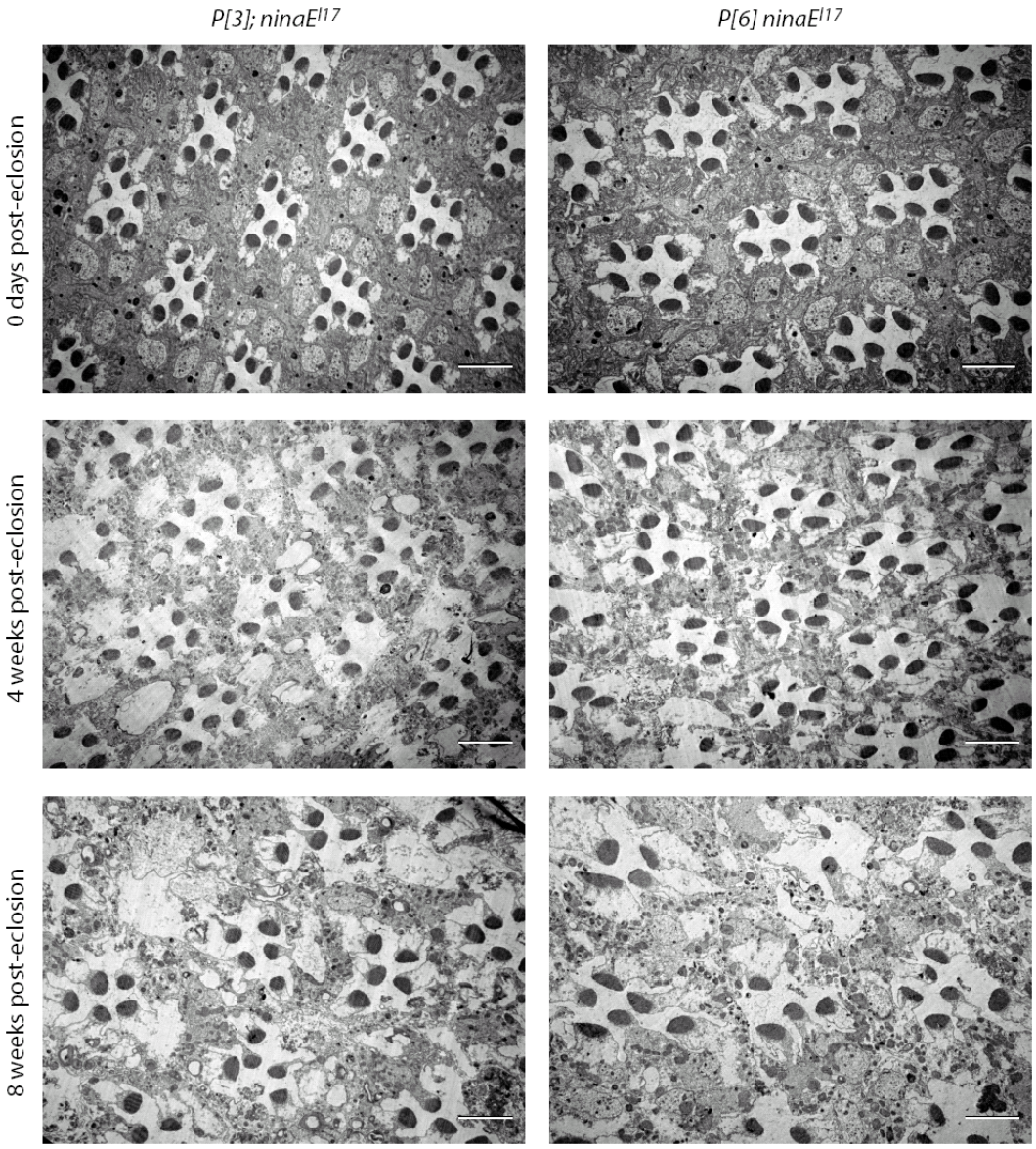


Figure 4.9: Degeneration in flies expressing constructs 3 and 6
While essentially all ommatidia have a full complement of rhabdomeres at eclosion, stochastic loss of some rhabdomeres is apparent by 4 weeks post-eclosion.
Scale bar = 5 μm.

This degenerative phenotype is unlike that of the Rh1 hypomorphs described by Leonard et al. (1992) that lose their rhabdomeres after gradual uniform shrinking until eventual disappearance. Constructs 3 and 6 instead show stochastic loss of some rhabdomeres which does not appear to be the culmination of gradual shrinking based on the size of the surviving rhabdomeres, suggesting that their degenerative process is different from the hypomorphs'. Two possible explanations for degeneration with these constructs are that they lack a trophic signal or they generate a toxic effect. To address these possibilities, I assayed heterozygous flies: if the constructs are deficient in trophic signaling then heterozygotes should show more pronounced degeneration than homozygotes, but if the constructs generate a toxic effect then degeneration should be less severe in heterozygotes. Figure 4.10 shows that degeneration is less severe in heterozygotes of construct 3, indicating that degeneration is likely due to a toxic activity. Construct 6 does not show a sufficiently clear phenotypic difference to draw conclusions about the effect of heterozygosity.

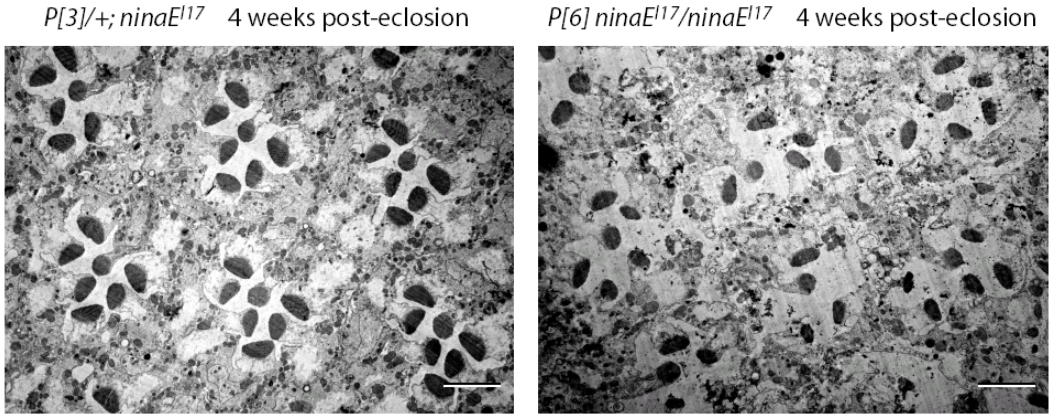


Figure 4.10: Rhabdomeres of construct 3 and 6 heterozygotes
Construct 3 heterozygotes show markedly reduced degeneration at 4 weeks post-eclosion compared to homozygotes (figure 4.9). Construct 6 does not show a clear difference between homozygotes and heterozygotes at this age. Scale bar = 5 μm.

Because degeneration is likely due to a toxic effect, at least in construct 3, I investigated known ways in which Rh1 activity can lead to photoreceptor degeneration. First, prolonged activation of the phototransduction cascade can lead to photoreceptor degeneration: a constitutively active TRP channel produces a phenotype similar to that seen with these constructs (Yoon et al., 2000). Second, even if the phototransduction cascade is blocked by eliminating the intermediate phospholipase C, persistently activated Rh1 will lead to photoreceptor apoptosis through a pathway requiring Arrestin2 interaction (Alloway et al., 2000, Kiselev et al., 2000) that is ameliorated by Arrestin1 (Satoh and Ready, 2005). While this work is ongoing (see chapter 5), the following results with arrestin are available.

In otherwise wild-type flies, the Arrestin1 mutation *arr1*^{1} leads to rhabdomere degeneration within 7 days that is accelerated by light exposure (Satoh and Ready, 2005). Figure 4.11 shows that at 7 days, degeneration is not apparent in flies expressing

constructs 3 and 6 regardless of whether *arr1*^{1} is present. At 4 weeks post-eclosion, flies expressing the designed constructs show degeneration of some central R7 rhabdomeres (which are also affected by the *arr1*^{1} mutation), but the outer rhabdomeres expressing the SCA-designed constructs do not show clearly accelerated degeneration in the background of *arr1*^{1}. These data do not indicate an Rh1-like interaction between the designed constructs and Arrestin1.

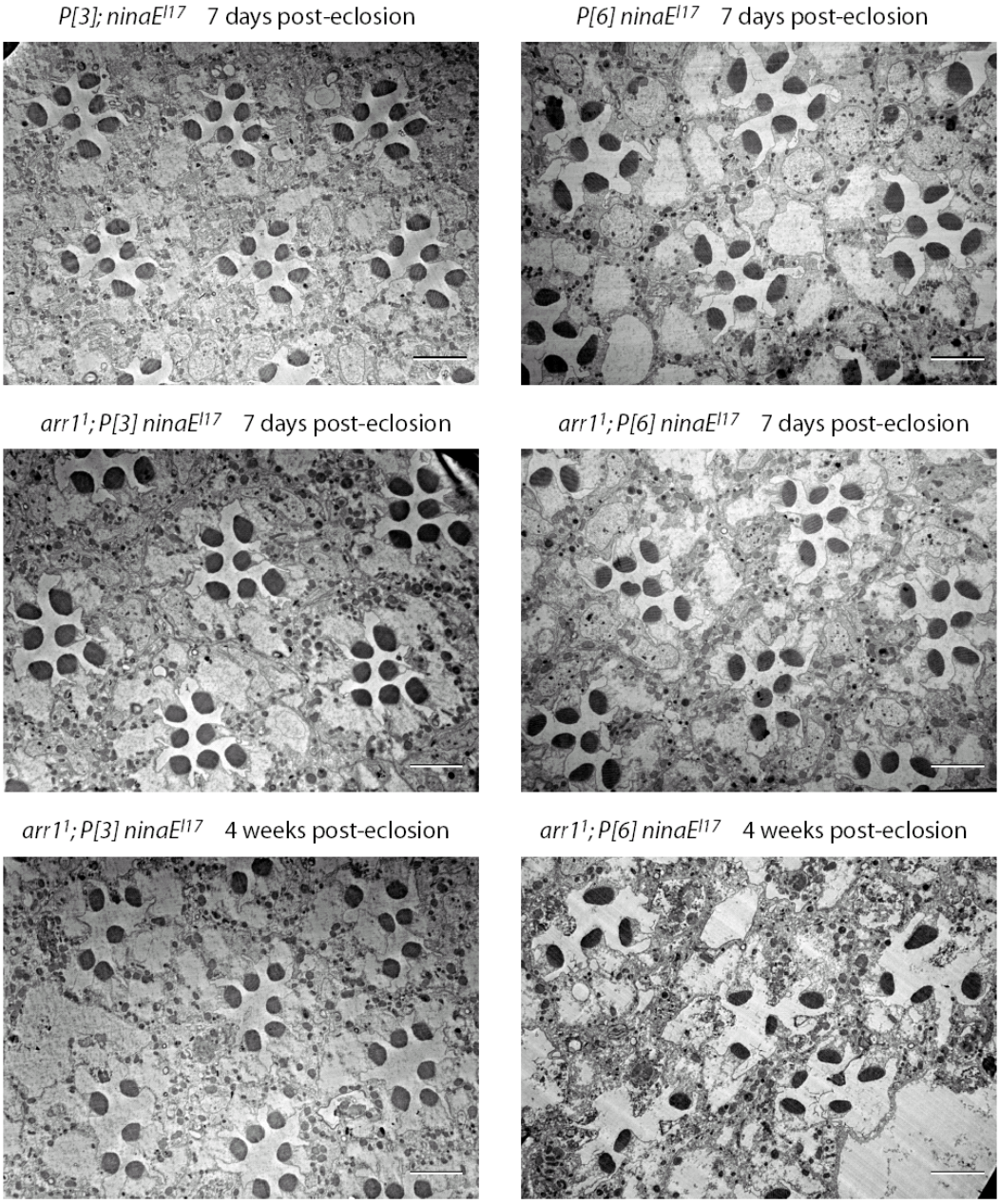


Figure 4.11: Rhabdomere degeneration in an Arrestin1 mutant background
Degeneration is neither accelerated to occur by 7 days post-eclosion, nor appreciably more severe at 4 weeks post-eclosion, in an Arrestin1 mutant background.
Scale bar = 5 μm.

DISCUSSION

This work shows that coupling information, even without the use of direct structural information, specifies the evolutionary constraints on the family of class A GPCRs in sufficient detail to generate proteins with many features of a natural member of this family. Several of the designed proteins express at appreciable levels in the fly eye. One of these shows a clear population of deglycosylated protein that has passed the quality control checkpoints of the ER, and interacts with NinaA. More work will be necessary to realize the goal of light dependent signaling, but these results are a promising start to the process.

Every construct tested rescues rhabdomere morphology at eclosion. Although no immediate effector molecules that interact with Rh1 to preserve rhabdomere morphology have been identified, current evidence indicates that this is due to a signaling pathway rather than mass action of protein. A pulse of Rh1 expression is able to permanently rescue the rhabdomeres if administered in the correct developmental window (Kumar et al., 1997), and degeneration can be prevented in the absence of Rh1 by activated *Drosophila* rac (Chang and Ready, 2000). Rhabdomere rescue is orthogonal to Rh1's phototransduction activity: mutants of the cognate Gq and phospholipase C have normal rhabdomeres (Kosloff et al., 2003), as do flies raised in total darkness. While the molecular mechanism of rhabdomere rescue remains unknown, it is striking that such a high probability of rescue was attained by the SCA-designed constructs while no rescue is observed with transformation of a nonspecific protein and only a temporary rescue is

afforded by a vertebrate rhodopsin that is expressed and folded in the fly eye (Ahmad et al., 2006, Ahmad et al., 2007).

The degenerative process in constructs 3 and 6 is also noteworthy. While their mechanism of degeneration is unknown, the fact that this phenotype is seen in only a subset of the designed proteins is significant. This indicates that some of the designed proteins exhibit yet another activity that cannot be attributed to expression of a nonspecific protein. If degeneration is due to aberrant activation of phototransduction or Arrestin2-mediated pathways, then this indicates that the design process further specifies these activities albeit not in the appropriately regulated manner. On the other hand, if those two constructs are driving degeneration by toxicity secondary to a design "flaw" such as an unfolded protein response, then it further indicates that the long-term preservation of the other constructs is significant.

The conclusion of this work is not simply that a molecule as complex as an integral membrane protein with the GPCR fold can be designed and exhibit functionality, but that the design rules sufficient for this task are completely encoded in the conservation and coupling pattern of this protein family. As is the case with all protein families studied with SCA, the coupling matrix is not dominated by strong interactions between all packing neighbors as one might expect to be necessary for protein design. Suel et al. (2003) have found that the strongest coupling values for the GPCR alignment arise from a network of 47 positions that form a physically connected unit from the ligand-binding pocket to the cytoplasmic face, and suggested that this represents a canonical basis for signal transduction and perhaps structural stability. Indeed, these results show that this basis of information specifies all interresidue interactions necessary

to supplement the conservation pattern in order to generate novel proteins with many natural-like functions.

METHODS

Sequence design

A sequence alignment of 940 class A GPCRs (Suel et al., 2003) was used as the starting alignment and source of SCA information for sequence design. The design algorithm used perturbation-based SCA calculations with the following set of perturbations: 58 L, 67 L, 69 G, 72 G, 73 N, 75 L, 76 V, 77 I, 79 V, 84 K, 93 L, 94 R, 95 T, 96 P, 97 T, 98 N, 100 F, 101 L, 103 N, 104 L, 105 A, 106 V, 107 A, 108 D, 109 L, 110 L, 113 L, 117 P, 130 W, 132 F, 133 G, 137 C, 150 A, 151 S, 152 I, 155 L, 157 A, 158 I, 159 S, 161 D, 162 R, 163 Y, 165 A, 166 I, 169 P, 170 L, 176 Y, 181 T, 185 A, 189 I, 193 W, 196 S, 200 S, 202 P, 203 P, 204 L, 235 C, 256 Y, 265 F, 268 P, 269 L, 271 I, 272 I, 275 C, 276 Y, 279 I, 283 L, 303 E, 305 K, 308 K, 310 L, 313 I, 314 V, 316 V, 317 F, 319 L, 320 C, 321 W, 322 L, 323 P, 324 Y, 329 L, 354 L, 355 A, 356 Y, 358 N, 359 S, 360 C, 362 N, 363 P, 364 I, 365 I, 366 Y, 368 F, 370 N, 373 F, 374 R, 377 F, 381 L, 383 C. The starting and ending values for the temperature parameter were 1000 and 10^{-7}, respectively, with a 10% decrease in temperature between iterations of either 2016300 accepted swaps or 20163000 attempted swaps (5 x #sequences x #columns and 50 x #sequences x #columns). The sequences from the designed alignment were sorted in decreasing order of identity to Rh1, and those that did not have a more similar sequence

that mapped away from the fly opsin cluster of the PCA map were selected if they also had an N-linked glycosylation site (N-X-S/T) before the first transmembrane helix. The 10 residue segement of the third cytoplasmic loop of Rh1 that is not included in the GPCR alignment was added unaltered. Fourteen design runs were performed to obtain the 20 selected sequences.

Principal components analysis

Principal components analysis was performed by calculating a matrix **A** of dissimilarity between each sequence in the alignment of natural GPCRs, such that element a_{ij} is one minus the sequence identity between sequences *i* and *j*. Principal components analysis was performed on this matrix, and the matrix **A** was multiplied by the resulting transformation **T** to obtain a set of coordinates representing the mapping of each sequence (Casari et al., 1995). To map designed sequences, a dissimilarity matrix for the designed sequences **B** was calculated with each element b_{ij} equal to one minus the sequence identity between designed sequence *i* and natural sequence *j*, and the product of **B** with the same transformation **T** calculated from the natural sequences produced the coordinates for the designed sequences.

Experimental animals

Genes encoding the designed proteins were synthesized by PCR of partially overlapping oligonucleotides of alternating sense encoding the designed proteins, with or

without a C-terminal epitope tag of the last seventeen residues of Rh1 (SSDAQSQATASEAESKA), using *Drosophila* optimized codons with flanking N-terminal SacI and C-terminal XhoI restriction sites. These were cloned into the P-element Yellow C4 carrying the y^+ marker and containing the 5' and 3' UTR of the Rh1 promoter. These were transformed into *y w; ; ninaE^{I17} sr* flies and bred to homozygosity using standard fly genetics. Flies expressing the epitope tagged construct 8t were only recovered with a homozygous lethal insertion; these flies were maintained over balancer chromosomes and assayed as heterozygotes. All flies were in a white-eyed (w^{1118}) background, except *w^{1118};; <pRh1-Rho> ninaE^{I17}* kindly provided by Joseph O'Tousa which carries a red eye marker on the P-element. Flies were maintained at 20º C with diurnal light/dark cycles unless otherwise indicated. Retinoid deprived flies were grown from egg to adult on media containing 10% yeast, 10% sucrose, 2% agar, and 0.02% cholesterol (Ozaki et al., 1993).

Western blotting

Standard SDS-PAGE was modified by using sample buffer containing 10% SDS and 5% 2-mercaptoethanol, and samples were heated to 40º C for one hour instead of boiled. These modifications are essential for high recovery of Rh1 as a monomeric species on SDS-PAGE. Rh1 and the designed proteins were detected with the monoclonal antibody 4C5 (DSHB, U Iowa). Treatment with endoglycosidase H was performed by homogenizing fly heads in 20 mM Tris pH 7.5, 1 mM PMSF with or

without 0.0025 U endoglycosidase H and incubating 1 hour at 37º C before proceeding with electrophoresis.

Whole cell voltage recording

Whole cell patch clamp recordings were performed as described in (Ranganathan et al., 1991). Briefly, heads were taken from dark-reared flies within one hour of eclosion, dissected under a solution of 120 mM NaCl, 4 mM KCl, 25 mM sucrose, 10 mM HEPES pH 7.1, gently triturated, and added to a solution with the above components plus 2 mM $MgCl_2$ and 0.5 mM $CaCl_2$. Patch pipettes were loaded with 95 mM potassium gluconate, 40 mM KCl, 2 mM $MgCl_2$, 2 mM EGTA, 3 mM Mg ATP, 0.5 mM Na GTP, 1 mM NAD, 10 mM HEPES pH 7.1. Cells were imaged under a Zeiss inverted microscope and whole cell access was achieved with standard techniques. All procedures were performed under dim red light. Responses to light from a xenon lamp filtered with neutral density and/or wavelength filters were recorded with an Axopatch 200B amplifier.

Electron microscopy

Flies were prepared for electron microscopy by perfusion with fixative (2% paraformaldehyde, 2% gluteraldehyde, 100 mM cacodylate pH 7.5) injected into the thorax and drained through a snip in the proboscis. Heads were then bisected under fixative and incubated 2 hours at room temperature before storage at 4º C, washed and postfixed with 2% osmium tetroxide followed by 2% uranyl acetate, dehydrated in

ethanol gradients and propylene oxide, and embedded in epon. Sections were cut at a depth of 40 microns below the lens, corresponding to the proximal edge of the R1-R6 nuclear level, and stained with uranyl acetate and lead citrate prior to imaging on a JEOL 1200 EX.

CHAPTER FIVE
Conclusions and Recommendations

WW domains designed using SCA information showed essentially the same probability of folding regardless of the details of the design process. Increasing the number of perturbations used, and using a global rather than perturbation-based algorithm, both had no discernable effect on the outcome. These findings are consistent with the concept that SCA reveals a highly coupled network of coevolving residues, and due to the redundancy of the information in the coevolving positions, the vast majority of relevant information can be revealed by examining a handful of sites if those sites are involved in the coupled network. This appears to be the case with the WW domains; extension of this concept to larger and more complex systems is likely to hold, with the caveat that in some cases such as the serine proteases (Suel et al., 2003) there are multiple coupled networks with potentially orthogonal functions. Any methodologies that effectively incorporate SCA information from all such networks seem likely to meet with similar degrees of success.

Sequences designed with the SCA information from an alignment of PDZ domains are able to fold and function like natural proteins albeit with a lower probability of folding, while sequences designed with conservation information alone – even at the same level of sequence identity to the natural alignment – do not. The physical constraints on building a well packed hydrophobic core and a functioning protein are substantial, evidenced by the challenge in computationally designing folded proteins (Dantas et al., 2003, Kuhlman et al., 2003) or re-specification of new functions on a naturally occurring scaffold (Looger et al., 2003, Dwyer et al., 2004) that have been met

only by quite sophisticated algorithms. The computational design of a membrane protein as complex as a GPCR based on physical principles will certainly be an even more formidable challenge for the future.

CONCLUDING GPCR DESIGN WORK

To investigate the possibility that constructs 3 and/or 6 are causing retinal degeneration by constitutive activity of the phototransduction cascade or interaction with Arrestin2, I have generated flies carrying these constructs in the background of a $norpA^{P41}$ mutant of the downstream phospholipase C and the $arr2^{3}$ mutant of Arrestin2. If retinal degeneration is blocked by either of these mutations, then the corresponding pathway would be implicated in the process. In addition, I have investigated whether the receptor's potential ligand modulates rhabdomere degeneration. Flies carrying the construct were raised in complete darkness to prevent photoisomerization of 3-hydroxyretinal, and were grown from egg to adult on yeast-sucrose media to eliminate retinoids. Flies carrying either construct show a similar degree of degeneration whether raised in a light/dark cycle or under constant darkness. However, preliminary results indicate that flies carrying construct 3, but not construct 6, have preserved rhabdomeres at 4 weeks post-eclosion when raised on yeast-sucrose media. I am preparing two more cohorts of flies raised on yeast-sucrose media, one with and one without supplementation with beta-carotene as a retinoid source, to confirm the retinoid dependence of this phenotype.

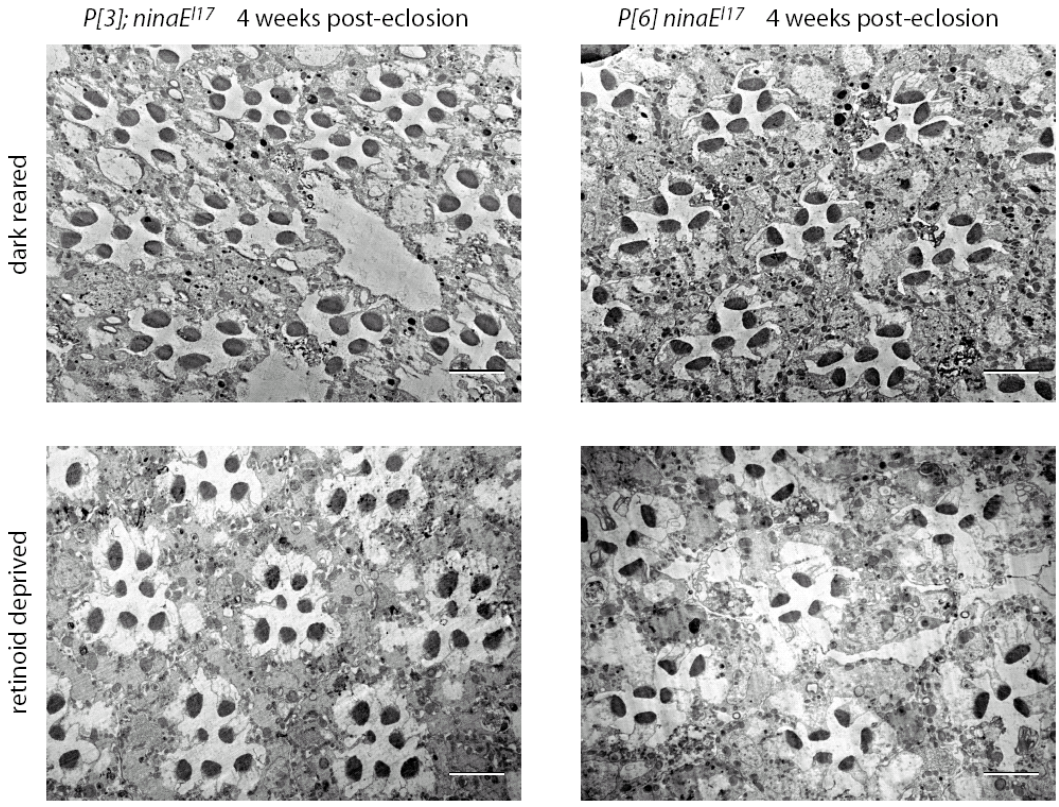


Figure 5.1: Retinoid effects on degeneration
Preventing retinoid isomerization by rearing in constant darkness does not affect degeneration. Degeneration is markedly reduced by retinoid deprivation on yeast-sucrose media with construct 3, but not construct 6. Scale bar = 5 μm.

One possible explanation for rhabdomere preservation with construct 3 on yeast-sucrose media is that expression of this protein may be suppressed by retinoid deprivation. Because the other SCA-designed constructs do not show a change in expression on yeast-sucrose media (figure 4.5), such a dependence would be a unique property of this construct and indicate that the encoded protein interacts with retinoids (possibly indirectly). Although the initial westerns did not reliably show expression of

construct 3t, I am growing larger numbers of these flies on normal and yeast-sucrose media in an attempt to obtain a direct measure of expression.

If rhabdomere rescue is not due to reduction of construct 3 expression levels, and it is not due to a nonspecific rhabdomere-preserving effect of yeast-sucrose media as evidenced by degeneration in construct 6, then construct 3 has some form of ligand-dependent, toxic activity. If it activates the phototransduction cascade in response to ligand, then degeneration should also be suppressed by introducing a *norpA* mutation. The photoreceptors should also show a calcium dependent current on whole-cell patch clamp recordings that is dietary retinoid dependent, and can possibly be induced in retinoid-deprived flies by stimulation with a saturated solution (~0.1 μM, Szuts and Harosi, 1991) of an appropriate retinoid.

RECOMMENDATIONS

This work indicates that an explanation of the physical basis for the observed coupling patterns would be an explanation of the mechanics sufficient to specify a functional protein, and developing such a physical explanation is the next great challenge. Prior work has investigated the degree to which coupling interactions may manifest as changes in protein structure in response to mutation. Structural studies of GFP (Jain and Ranganathan, 2004) and the third PDZ domain of PSD-95 (Sharma, 2006) have shown that in some cases a mechanistic response to mutation at a coupled position can be seen in crystal structures, and in the case of GFP these rearrangements indicate that packing interactions may be the physical constraint behind the interrogated coupling interactions.

But this is not a universal feature of coupled sites. In the case of PSD-95, structural rearrangements were not seen in response to mutation of the highly coupled residue H372Y in the ligand-free state. Only with ligand bound was a reorientation of the peptide in the binding cleft and reorientation of the carboxylate binding loop in response to H372Y evident. The G322A mutation of a coupled residue was found to induce a conformational change in the carboxylate binding loop, and comparing this structure to the structures of the wild-type free and ligand-bound states showed that this mutation locks the carboxylate binding loop into the position it adopts in the ligand-bound state. Ultimately, structures of the protein with several combinations of wild-type and mutant residues at positions 372 and 322, with and without the target peptide, were necessary to elucidate the physical basis of this coupling interaction. This illustrates the fact that although a statistical coupling matrix is a simple parameterization of the evolutionary pressure, the underlying process that constrains evolution to follow such simple rules may in principle (and in this real example) be quite complex.

Interrogation of the mechanics driving coupling may require a (minimally) two step approach. Because coupling interactions reflect evolutionary pressure on a protein family, and evolution selects sequences based on their ability to function (with folding generally being a prerequisite for efficient function), it is reasonable to expect that some coupling interactions may contribute primarily to folding while others contribute primarily to function, with the possibility that some interactions contribute to both. Naturally, coupling interactions responsible for specifying the protein's structure should be most amenable to structural analysis, while functional interactions will likely require sophisticated approaches and, as in the case for Sharma, assessment of the effect of

substrate – particularly difficult if the protein carries out functions that have not yet been identified. Two projects in this lab indicate such orthogonality of interactions in two different protein families, so segregation of interactions with structural versus functional roles and targeted interrogation of structural interactions is feasible. Presently there is no way of deducing the functional roles of particular interactions from the coupling matrix itself, so in general such an analysis of a protein family must begin with experimental characterization of the effects of mutations at different sites, followed by detailed structural studies of those mutations that affect fold stability and their mechanism of interaction.

While targeted structural studies like (Jain and Ranganathan, 2004) and (Sharma, 2006) would provide a fairly generalizable way of investigating the mechanistic basis of coupling interactions that contribute to folding, interactions involved in function will likely require experimental approaches tailored to the particular activity of each different protein family. However, some generalizations can be made. For the most part, protein families share some fundamental activity – a conserved mode of peptide binding for the PDZ domains, a conserved catalytic mechanism for the serine proteases, a conserved allosteric response to ligand binding for the GPCRs – with diversification not in the core functionality of the individual family members but in the specialized aspect of their mechanism, such as which particular substrate they interact with, what binding affinity they have, or how sensitive to allostery they are. Binding specificity could in principle be studied structurally, with the understanding that finer aspects such as maintaining negative selectivity for off-target interactions could be missed unless specifically tested. Dynamic processes such as coordinated motions along the reaction coordinate

(Eisenmesser et al., 2005, Hammes-Schiffer and Benkovic, 2006, Labeikovsky et al., 2007) or allostery mediated by dynamic rather than structural changes (Popovych et al., 2006) would be essentially invisible to static structural analysis and require specialized NMR experiments. If properties such as modulation of the response to allosteric signals are specified in the coupling matrix (as in Shulman et al., 2004), then teasing out the mechanics of how coupling interactions contribute to such processes would require prior knowledge of their likely effect and experiments designed specifically for the protein under study. Addressing such specialized roles for coupled networks would likely require collaboration with laboratories specifically geared for their study with expertise in the particular protein family, but could provide an enlightening perspective on the system under investigation.

With this in mind, the experimental search for the physical basis of the coupling patterns that are necessary and sufficient to specify protein folding and function must be as multifaceted as the evolutionary pressures that a protein must satisfy. The ongoing work indicating orthogonal functions of different subsets of coupling interactions already indicates that no single biophysical assay is likely to recapitulate the entire coupling pattern of most proteins. However, progress along this line of research would lend itself to many interesting tests of the design principles well before a full explanation is attained. For example, if structural studies reveal a mechanistic logic to the allowed modes of evolution that maintain fold stability, then it would in principle be possible to use the elucidated design rules to redesign a protein core by applying the design rules in ways that might lead to a diverse set of solutions like the ensembles of naturally occurring homologs, rather than the typical single solution or set of very similar solutions obtained

by energy minimization. The difference between such a project and the work presented here is that the proposed research would not involve performing statistical coupling analysis on a sequence alignment of the design target, but instead taking a single structure and anticipating what the evolutionary pressure for the fold would be. As functional studies advance, it would also be possible to incorporate functions into proteins in ways that go beyond the current computational approach of stabilizing a transition state with local interactions (Dwyer et al., 2004, Allert et al., 2007) similar to catalytic antibodies, particularly if the proposed slow scale motions along the reaction coordinate that contribute to activity and the mechanism by which they are encoded at the amino acid level are understood.

APPENDIX A
Chemical shift assignments for C_{60}-1

Assignments for methyl groups of leucine and valine residues are stereospecific.
Assignments for geminal protons are not.

2.CA	58.038	8.N	124.000	13.HG	1.387	18.CA	52.827	21.HG1	2.275
2.HA	4.351	8.HN	8.230	13.CD1	25.312	18.HA	4.773	21.C	176.798
2.CB	63.701	8.CA	52.505	13.HD11	0.711	18.CB	44.957	22.N	119.243
2.HB2	3.680	8.HA	4.262	13.CD2	25.707	18.HB2	1.116	22.HN	8.220
2.HB1	3.680	8.CB	19.165	13.HD21	0.677	18.HB1	1.509	22.CA	56.271
2.C	174.523	8.HB1	1.283	13.C	176.767	18.CG	27.369	22.HA	4.207
3.N	120.637	8.C	177.778	14.N	123.420	18.HG	1.392	22.CB	33.514
3.HN	8.577	9.N	114.684	14.HN	9.182	18.CD1	25.079	22.HB2	1.735
3.CA	55.950	9.HN	8.237	14.CA	56.665	18.HD11	0.688	22.HB1	1.735
3.HA	4.523	9.CA	59.045	14.HA	4.834	18.CD2	24.035	22.CG	24.610
3.CB	29.909	9.HA	4.285	14.CB	41.032	18.HD21	0.756	22.HG2	1.321
3.HB2	2.939	9.CB	63.672	14.HB2	2.813	18.C	175.248	22.HG1	1.321
3.HB1	3.064	9.HB2	3.782	14.HB1	2.813	19.N	124.353	22.CD	28.989
3.HD2	6.800	9.HB1	3.782	14.HD1	6.825	19.HN	9.109	22.HD2	1.571
3.HE1	7.113	9.C	175.174	14.HE1	6.327	19.CA	57.235	22.HD1	1.571
3.C	175.192	10.N	109.520	14.HE2	6.327	19.HA	4.558	22.CE	41.791
4.N	109.557	10.HN	8.373	14.HD2	6.825	19.CB	40.594	22.HE2	2.869
4.HN	8.328	10.CA	45.267	14.C	171.553	19.HB2	2.866	22.HE1	2.869
4.CA	45.119	10.HA2	3.815	15.N	114.785	19.HB1	2.866	22.C	177.114
4.HA2	3.776	10.HA1	3.815	15.HN	8.388	19.HD1	7.241	23.N	119.604
4.HA1	3.776	10.C	173.304	15.CA	55.585	19.HE1	7.299	23.HN	8.621
4.C	173.861	11.N	117.933	15.HA	5.454	19.HZ	6.990	23.CA	55.936
5.N	120.227	11.HN	7.952	15.CB	65.642	19.HE2	7.299	23.HA	4.261
5.HN	8.150	11.CA	56.549	15.HB2	3.618	19.HD2	7.241	23.CB	39.995
5.CA	57.789	11.HA	5.277	15.HB1	3.618	19.C	175.254	23.HB2	2.687
5.HA	4.483	11.CB	42.258	15.C	174.399	20.N	121.883	23.HB1	2.687
5.CB	38.944	11.HB2	2.789	16.N	118.298	20.HN	8.675	23.C	176.691
5.HB2	2.810	11.HB1	2.869	16.HN	8.591	20.CA	54.856	24.N	108.447
5.HB1	2.922	11.HD1	7.083	16.CA	60.023	20.HA	4.254	24.HN	7.508
5.HD1	7.077	11.HE1	7.299	16.HA	4.373	20.CB	32.157	24.CA	61.176
5.HE1	6.795	11.HZ	7.345	16.CB	36.434	20.HB2	1.627	24.HA	4.285
5.HE2	6.795	11.HE2	7.299	16.HB	1.836	20.HB1	1.390	24.CB	69.496
5.HD2	7.077	11.HD2	7.083	16.CG2	21.340	20.CG	27.340	24.HB	4.192
5.C	176.066	11.C	174.852	16.HG21	0.751	20.HG2	1.393	24.CG2	21.515
6.N	117.234	12.N	115.389	16.CG1	21.751	20.HG1	1.393	24.HG21	1.131
6.HN	8.241	12.HN	9.041	16.HG11	0.854	20.CD	43.513	24.C	175.321
6.CA	58.038	12.CA	56.782	16.C	173.731	20.HD2	2.876	25.N	118.295
6.HA	4.352	12.HA	4.725	17.N	124.705	20.HD1	2.876	25.HN	8.006
6.CB	63.716	12.CB	65.468	17.HN	7.740	20.NE	84.084	25.CA	59.017
6.HB2	3.758	12.HB2	3.664	17.CA	54.680	20.HE	7.160	25.HA	4.345
6.HB1	3.681	12.HB1	3.664	17.HA	5.075	20.C	176.480	25.CB	63.424
6.C	174.163	12.C	172.898	17.CB	31.252	21.N	123.226	25.HB2	3.848
7.N	122.386	13.N	128.549	17.HB2	1.751	21.HN	8.611	25.HB1	3.878
7.HN	8.236	13.HN	8.501	17.HB1	1.669	21.CA	57.235	25.C	174.237
7.CA	54.272	13.CA	54.111	17.CG	37.105	21.HA	4.045	26.N	116.847
7.HA	4.484	13.HA	5.183	17.HG2	1.787	21.CB	30.113	26.HN	8.364
7.CB	41.207	13.CB	44.257	17.HG1	1.914	21.HB2	2.013	26.CA	57.541
7.HB2	2.604	13.HB2	1.370	17.C	174.924	21.HB1	1.917	26.HA	4.433
7.HB1	2.604	13.HB1	1.556	18.N	123.301	21.CG	36.258	26.CB	63.862
7.C	176.255	13.CG	27.880	18.HN	8.711	21.HG2	2.198	26.HB2	3.692

26.HB1	3.875	32.N	115.601	37.CG	33.762	44.HN	8.209	49.CG	24.537
26.C	175.707	32.HN	8.829	37.HG2	2.240	44.CA	61.409	49.HG2	1.308
27.N	125.283	32.CA	56.768	37.HG1	2.240	44.HA	4.471	49.HG1	1.222
27.HN	8.783	32.HA	4.533	37.NE2	112.629	44.CB	70.094	49.CD	28.843
27.CA	56.841	32.CB	65.394	37.HE21	7.490	44.HB	4.237	49.HD2	1.497
27.HA	4.112	32.HB2	3.433	37.HE22	6.792	44.CG2	21.486	49.HD1	1.582
27.CB	43.075	32.HB1	3.664	37.C	176.161	44.HG21	1.118	49.CE	41.688
27.HB2	1.460	32.C	173.545	38.N	116.000	44.C	174.752	49.HE2	2.857
27.HB1	1.722	33.N	110.304	38.HN	8.306	45.N	110.785	49.HE1	2.857
27.CG	27.238	33.HN	8.446	38.CA	58.432	45.HN	8.317	49.C	176.099
27.HG	1.392	33.CA	44.841	38.HA	4.393	45.CA	44.885	50.N	108.718
27.CD1	25.734	33.HA2	3.712	38.CB	63.818	45.HA2	4.036	50.HN	7.526
27.HD11	0.816	33.HA1	4.949	38.HB2	3.761	45.HA1	3.752	50.CA	56.957
27.CD2	24.114	33.C	173.930	38.HB1	3.869	45.C	171.493	50.HA	4.267
27.HD21	0.732	34.N	119.352	38.C	175.299	46.N	121.360	50.CB	63.716
27.C	177.160	34.HN	8.391	39.N	115.225	46.HN	8.339	50.HB2	3.671
28.N	102.333	34.CA	54.928	39.HN	8.188	46.CA	57.804	50.HB1	3.779
28.HN	8.149	34.HA	4.491	39.CA	61.964	46.HA	4.484	50.C	170.864
28.CA	45.148	34.CB	34.375	39.HA	4.347	46.CB	36.346	51.N	118.925
28.HA2	4.717	34.HB2	1.792	39.CB	69.452	46.HB	1.976	51.HN	8.491
28.HA1	3.446	34.HB1	1.931	39.HB	4.212	46.CG1	26.085	51.CA	51.790
28.C	173.845	34.CG	31.806	39.CG2	21.486	46.HG12	1.154	51.HA	5.258
29.N	116.640	34.HG2	2.341	39.HG21	1.127	46.HG11	1.541	51.CB	43.352
29.HN	7.466	34.HG1	2.341	39.C	175.224	46.CD1	9.648	51.HB2	0.942
29.CA	60.344	34.CE	16.961	40.N	113.831	46.HD11	0.603	51.HB1	1.590
29.HA	4.218	34.HE1	1.849	40.HN	8.050	46.CG2	18.712	51.CG	27.442
29.CB	39.791	34.C	175.561	40.CA	61.730	46.HG21	0.679	51.HG	1.347
29.HB	1.553	35.N	122.291	40.HA	4.277	46.C	174.288	51.CD1	25.296
29.CG1	27.372	35.HN	8.388	40.CB	69.491	47.N	123.567	51.HD11	0.646
29.HG12	0.422	35.CA	56.052	40.HB	4.291	47.HN	8.726	51.CD2	24.610
29.HG11	1.320	35.HA	4.083	40.CG2	21.500	47.CA	55.760	51.HD21	0.569
29.CD1	13.910	35.CB	30.726	40.HG21	1.109	47.HA	5.123	51.C	177.524
29.HD11	0.538	35.HB2	1.574	40.C	175.399	47.CB	41.715	52.N	126.180
29.CG2	18.537	35.HB1	1.534	41.N	110.905	47.HB2	2.552	52.HN	8.855
29.HG21	0.600	35.CG	27.121	41.HN	8.222	47.HB1	2.621	52.CA	58.461
29.C	175.238	35.HG2	1.376	41.CA	45.133	47.HD1	6.971	52.HA	4.130
30.N	123.785	35.HG1	1.376	41.HA2	3.954	47.HE1	6.752	52.CB	38.084
30.HN	8.368	35.CD	43.090	41.HA1	3.791	47.HE2	6.752	52.HB	1.602
30.CA	56.768	35.HD2	3.037	41.C	174.077	47.HD2	6.971	52.CG1	27.384
30.HA	4.754	35.HD1	3.037	42.N	120.326	47.C	175.676	52.HG12	1.126
30.CB	63.730	35.NE	84.473	42.HN	8.095	48.N	119.512	52.HG11	1.394
30.HB2	3.767	35.HE	7.220	42.CA	56.067	48.HN	8.872	52.CD1	11.356
30.HB1	3.666	35.C	175.968	42.HA	4.210	48.CA	63.249	52.HD11	0.674
30.C	174.864	36.N	121.513	42.CB	30.464	48.HA	3.685	52.CG2	15.895
31.N	120.612	36.HN	8.282	42.HB2	1.949	48.CB	31.661	52.HG21	0.883
31.HN	8.652	36.CA	54.272	42.HB1	1.817	48.HB	2.128	53.CA	64.197
31.CA	59.410	36.HA	4.488	42.CG	36.185	48.CG2	20.902	53.HA	4.126
31.HA	5.196	36.CB	41.192	42.HG2	2.145	48.HG21	0.682	53.CB	31.602
31.CB	41.459	36.HB2	2.594	42.HG1	2.145	48.CG1	21.267	53.HB2	1.798
31.HB	1.762	36.HB1	2.614	42.C	176.201	48.HG11	0.559	53.HB1	2.227
31.CG1	25.238	36.C	176.309	43.N	125.198	48.C	175.604	53.CG	27.622
31.HG12	1.211	37.N	119.479	43.HN	8.410	49.N	109.425	53.HG2	2.048
31.HG11	1.011	37.HN	8.350	43.CA	52.287	49.HN	9.427	53.HG1	1.908
31.CD1	14.406	37.CA	56.140	43.HA	4.365	49.CA	57.308	53.CD	51.440
31.HD11	0.666	37.HA	4.190	43.CB	19.267	49.HA	4.281	53.HD2	3.573
31.CG2	18.610	37.CB	28.639	43.HB1	1.324	49.CB	33.552	53.HD1	4.076
31.HG21	0.589	37.HB2	2.087	43.C	178.030	49.HB2	1.393	53.C	177.951
31.C	175.269	37.HB1	1.910	44.N	112.924	49.HB1	1.482	54.N	111.626

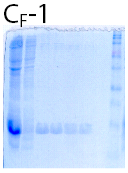
54.HN	9.015	61.CA	46.243	66.CA	55.979	70.CD2	22.508	76.HN	8.697
54.CA	45.148	61.HA2	3.730	66.HA	4.192	70.HD21	0.571	76.CA	53.060
54.HA2	4.112	61.HA1	3.956	66.CB	37.184	70.C	178.139	76.HA	4.714
54.HA1	3.471	61.C	175.104	66.HB2	2.934	71.N	119.055	76.CB	38.127
54.C	173.322	62.N	119.089	66.HB1	3.366	71.HN	7.554	76.HB2	2.676
55.N	114.533	62.HN	8.102	66.ND2	113.703	71.CA	55.088	76.HB1	2.937
55.HN	7.640	62.CA	58.417	66.HD21	7.532	71.HA	5.137	76.ND2	111.702
55.CA	57.557	62.HA	4.020	66.HD22	6.923	71.CB	35.923	76.HD21	7.519
55.HA	4.387	62.CB	32.317	66.C	175.297	71.HB2	1.492	76.HD22	6.844
55.CB	65.681	62.HB2	1.942	67.N	122.450	71.HB1	1.560	76.C	176.578
55.HB2	3.723	62.HB1	1.556	67.HN	7.933	71.CG	27.369	77.N	117.515
55.HB1	4.010	62.CG	27.442	67.CA	55.804	71.HG2	1.231	77.HN	7.477
55.C	174.318	62.HG2	1.577	67.HA	4.781	71.HG1	1.366	77.CA	60.504
56.N	122.289	62.HG1	1.577	67.CB	41.271	71.CD	43.907	77.HA	4.483
56.HN	9.109	62.CD	43.465	67.HB2	2.480	71.HD2	2.918	77.CB	31.062
56.CA	55.396	62.HD2	3.276	67.HB1	2.701	71.HD1	2.918	77.HB	2.369
56.HA	4.095	62.HD1	3.031	67.C	174.671	71.NE	84.063	77.CG2	19.267
56.CB	18.406	62.NE	86.770	68.N	122.677	71.HE	7.070	77.HG21	0.418
56.HB1	1.508	62.HE	8.786	68.HN	8.554	71.C	177.420	77.CG1	21.325
56.C	181.306	62.C	176.092	68.CA	54.797	72.N	124.030	77.HG11	0.768
57.N	116.159	63.N	119.123	68.HA	4.679	72.HN	8.322	77.C	176.509
57.HN	8.393	63.HN	8.287	68.CB	34.971	72.CA	60.753	78.N	121.965
57.CA	54.928	63.CA	61.147	68.HB2	1.698	72.HA	4.247	78.HN	8.087
57.HA	3.968	63.HA	4.273	68.HB1	1.439	72.CB	33.252	78.CA	58.154
57.CB	19.574	63.CB	39.076	68.CG	25.092	72.HB	1.658	78.HA	4.025
57.HB1	1.199	63.HB	1.404	68.HG2	0.771	72.CG2	19.953	78.CB	28.522
57.C	179.872	63.CG1	27.500	68.HG1	0.771	72.HG21	0.560	78.HB2	2.101
58.N	121.686	63.HG12	1.767	68.CD	29.632	72.CG1	21.121	78.HB1	1.961
58.HN	7.655	63.HG11	1.393	68.HD2	1.324	72.HG11	0.569	78.CG	34.230
58.CA	54.695	63.CD1	14.800	68.HD1	1.396	72.C	175.149	78.HG2	2.395
58.HA	3.954	63.HD11	0.724	68.CE	42.053	73.N	128.043	78.HG1	2.225
58.CB	18.990	63.CG2	19.194	68.HE2	2.644	73.HN	9.969	78.NE2	112.851
58.HB1	1.366	63.HG21	0.769	68.HE1	2.644	73.CA	57.060	78.HE21	7.501
58.C	180.041	63.C	175.066	69.N	123.785	73.HA	4.034	78.HE22	6.888
59.N	117.812	64.N	125.368	69.HN	8.368	73.CB	38.915	78.C	177.246
59.HN	8.492	64.HN	8.003	69.CA	61.876	73.HB2	2.649	79.N	112.784
59.CA	56.811	64.CA	52.710	69.HA	3.869	73.HB1	2.869	79.HN	8.684
59.HA	3.938	64.HA	4.719	69.CB	38.689	73.C	176.209	79.CA	45.102
59.CB	41.588	64.CB	31.320	69.HB	1.318	74.N	126.846	79.HA2	3.911
59.HB2	1.675	64.HB2	1.871	69.CG1	27.765	74.HN	7.833	79.HA1	3.793
59.HB1	1.426	64.HB1	1.871	69.HG12	1.426	74.CA	54.928	79.C	174.357
59.CG	27.106	64.CG	35.709	69.HG11	1.304	74.HA	3.984	80.N	120.321
59.HG	1.588	64.HG2	2.043	69.CD1	13.297	74.CB	43.922	80.HN	7.521
59.CD1	25.019	64.HG1	2.113	69.HD11	0.485	74.HB2	2.680	80.CA	56.115
59.HD11	0.704	65.CA	63.292	69.CG2	17.880	74.HB1	2.516	80.HA	4.197
59.CD2	22.505	65.HA	3.787	69.HG21	0.517	74.C	176.480	80.CB	34.609
59.HD21	0.781	65.CB	31.500	69.C	174.267	75.N	122.975	80.HB2	1.926
59.C	179.098	65.HB2	1.699	70.N	124.395	75.HN	8.608	80.HB1	1.819
60.N	118.220	65.HB1	2.141	70.HN	8.958	75.CA	62.387	80.CG	33.500
60.HN	7.437	65.CG	28.668	70.CA	55.177	75.HA	3.828	80.HG2	2.382
60.CA	57.322	65.HG2	1.585	70.HA	4.373	75.CB	32.580	80.HG1	2.483
60.HA	4.250	65.HG1	2.120	70.CB	43.557	75.HB	2.184	80.CE	17.048
60.CB	43.849	65.CD	50.477	70.HB2	1.331	75.CG2	21.033	80.HE1	1.884
60.HB2	2.528	65.HD2	3.521	70.HB1	1.239	75.HG21	0.846	80.C	175.444
60.HB1	2.673	65.HD1	3.677	70.CG	28.828	75.CG1	21.486	81.N	123.805
60.C	177.871	65.C	177.496	70.HG	1.312	75.HG11	0.682	81.HN	8.785
61.N	102.229	66.N	116.442	70.CD1	25.293	75.C	175.319	81.CA	51.396
61.HN	7.101	66.HN	9.698	70.HD11	0.511	76.N	126.370	81.HA	4.256

81.CB	19.077	87.HA	4.124	91.HD22	6.824	97.CD	43.148	101.HD11	0.430
81.HB1	1.411	87.CB	28.975	91.C	174.538	97.HD2	3.017	101.CG2	17.486
81.C	179.430	87.HB2	2.054	92.N	123.327	97.HD1	3.078	101.HG21	0.767
82.N	122.770	87.HB1	2.054	92.HN	6.973	97.NE	108.757	101.C	174.388
82.HN	9.003	87.CG	35.237	92.CA	52.491	97.HE	7.175	102.N	129.293
82.CA	60.607	87.HG2	2.281	92.HA	4.172	97.C	176.094	102.HN	9.054
82.HA	3.578	87.HG1	2.281	92.CB	19.369	98.N	124.831	102.CA	54.169
82.CB	28.070	87.C	178.130	92.HB1	1.416	98.HN	8.906	102.HA	4.758
82.HB2	2.016	88.N	121.535	92.C	177.721	98.CA	53.828	102.CB	34.011
82.HB1	2.016	88.HN	7.785	93.N	109.704	98.HA	5.145	102.HB2	1.826
82.CG	33.368	88.CA	66.226	93.HN	8.545	98.CB	45.878	102.HB1	1.826
82.HG2	2.158	88.HA	3.566	93.CA	44.374	98.HB2	1.159	102.CG	25.676
82.HG1	2.158	88.CB	31.485	93.HA2	3.589	98.HB1	1.535	102.HG2	1.542
82.NE2	111.006	88.HB	2.169	93.HA1	3.985	98.CG	27.544	102.HG1	1.542
82.HE21	7.342	88.CG2	23.180	93.C	171.322	98.HG	1.446	102.CD	42.783
82.HE22	6.786	88.HG21	0.961	94.N	115.682	98.CD1	26.450	102.HD2	2.857
82.C	177.816	88.CG1	20.916	94.HN	7.940	98.HD11	0.578	102.HD1	3.200
83.N	111.681	88.HG11	0.750	94.CA	49.586	98.CD2	26.391	102.NE	82.873
83.HN	8.712	88.C	179.584	94.HA	4.996	98.HD21	0.678	102.HE	8.524
83.CA	61.468	89.N	118.155	94.CB	41.148	98.C	174.613	102.C	175.330
83.HA	3.939	89.HN	8.201	94.HB2	2.455	99.N	123.658	103.N	127.698
83.CB	61.468	89.CA	57.935	94.HB1	2.811	99.HN	8.077	103.HN	8.822
83.HB2	3.767	89.HA	3.782	94.ND2	113.197	99.CA	53.863	103.CA	56.414
83.HB1	3.814	89.CB	41.455	94.HD21	7.696	99.HA	4.849	103.HA	4.235
83.C	177.300	89.HB2	1.420	94.HD22	7.090	99.CB	45.148	103.CB	30.639
84.N	121.325	89.HB1	1.886	95.CA	62.708	99.HB2	1.562	103.HB2	1.671
84.HN	6.964	89.CG	27.223	95.HA	4.269	99.HB1	1.209	103.HB1	1.671
84.CA	56.724	89.HG	1.578	95.CB	33.558	99.CG	27.165	103.CG	27.092
84.HA	4.527	89.CD1	25.559	95.HB2	1.305	99.HG	1.585	103.HG2	1.374
84.CB	40.565	89.HD11	0.694	95.HB1	1.482	99.CD1	23.019	103.HG1	1.374
84.HB2	2.697	89.CD2	23.837	95.CG	24.537	99.HD11	0.694	103.CD	43.119
84.HB1	2.830	89.HD21	0.642	95.HG2	1.220	99.CD2	25.559	103.HD2	3.003
84.C	178.811	89.C	179.039	95.HG1	1.305	99.HD21	0.664	103.HD1	3.003
85.N	122.588	90.N	119.728	95.CD	50.272	99.C	175.280	103.NE	84.288
85.HN	7.895	90.HN	7.962	95.HD2	3.671	100.N	126.964	103.HE	7.183
85.CA	66.708	90.CA	58.943	95.HD1	3.522	100.HN	8.883	103.C	175.854
85.HA	3.410	90.HA	3.957	95.C	175.412	100.CA	54.564	104.N	127.480
85.CB	30.726	90.CB	30.515	96.N	119.449	100.HA	5.333	104.HN	8.094
85.HB	1.985	90.HB2	1.841	96.HN	8.485	100.CB	46.067	104.CA	52.904
85.CG2	24.406	90.HB1	1.841	96.CA	60.315	100.HB2	1.236	104.HA	4.464
85.HG21	0.907	90.CG	27.282	96.HA	4.305	100.HB1	1.236	104.CB	41.002
85.CG1	21.529	90.HG2	1.643	96.CB	33.164	100.CG	28.362	104.HB2	1.309
85.HG11	0.736	90.HG1	1.534	96.HB	1.970	100.HG	1.342	104.HB1	1.392
85.C	177.540	90.CD	43.206	96.CG2	22.990	100.CD1	26.858	104.CG	27.106
86.N	120.418	90.HD2	3.050	96.HG21	0.869	100.HD11	0.526	104.HG	1.341
86.HN	8.168	90.HD1	3.144	96.CG1	20.916	100.CD2	27.909	104.CD1	25.371
86.CA	67.248	90.NE	83.754	96.HG11	0.755	100.HD21	0.679	104.HD11	0.721
86.HA	3.414	90.HE	7.665	96.C	174.078	100.C	173.907	104.CD2	22.654
86.CB	31.602	90.C	177.867	97.N	127.446	101.N	114.854	104.HD21	0.742
86.HB	2.116	91.N	114.533	97.HN	8.639	101.HN	8.745	105.CA	62.708
86.CG2	23.092	91.HN	7.640	97.CA	54.695	101.CA	58.972	105.HA	4.272
86.HG21	1.015	91.CA	52.929	97.HA	4.942	101.HA	5.223	105.CB	31.909
86.CG1	21.033	91.HA	4.641	97.CB	31.208	101.CB	42.068	105.HB2	2.133
86.HG11	0.877	91.CB	39.265	97.HB2	1.691	101.HB	1.491	105.HB1	1.730
86.C	179.041	91.HB2	2.539	97.HB1	1.691	101.CG1	28.362	105.CG	27.267
87.N	118.880	91.HB1	2.849	97.CG	28.099	101.HG12	0.849	105.HG2	1.891
87.HN	6.997	91.ND2	114.005	97.HG2	1.323	101.HG11	1.385	105.HG1	1.891
87.CA	58.621	91.HD21	7.435	97.HG1	1.437	101.CD1	13.662	105.CD	50.345

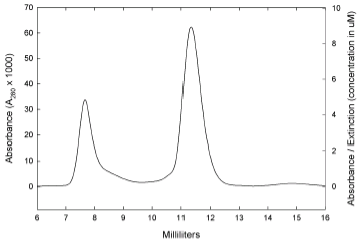
105.HD2	3.684	106.HB1	1.471	107.HN	8.260	107.HD11	0.784	108.HB2	1.755
105.HD1	3.521	106.CG	26.986	107.CA	54.724	107.CD2	23.326	108.HB1	1.912
105.C	176.468	106.HG	1.469	107.HA	4.283	107.HD21	0.724	108.CG	36.550
106.N	122.371	106.CD1	24.718	107.CB	42.141	107.C	176.305	108.HG2	2.044
106.HN	8.189	106.HD11	0.771	107.HB2	1.488	108.N	126.539	108.HG1	2.044
106.CA	54.724	106.CD2	23.804	107.HB1	1.488	108.HN	7.804		
106.HA	4.199	106.HD21	0.724	107.CG	26.931	108.CA	57.600		
106.CB	42.199	106.C	177.142	107.HG	1.482	108.HA	4.001		
106.HB2	1.471	107.N	124.505	107.CD1	24.844	108.CB	31.281		

APPENDIX B
Expression and size exclusion of PDZ constructs
(See figure 3.2 caption)

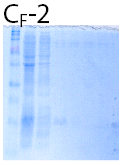
C_F-1



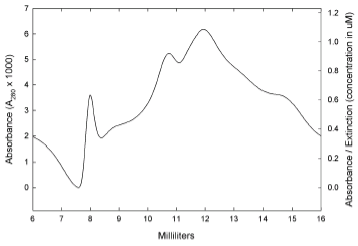
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
11.25 ml
11.55 ml



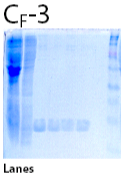
C_F-2



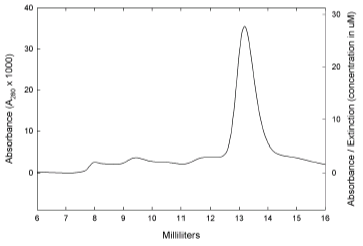
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
10.65 ml
11.85 ml



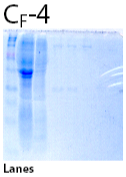
C_F-3



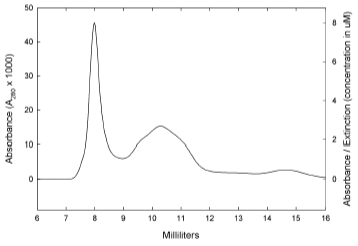
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
13.05 ml
13.35 ml



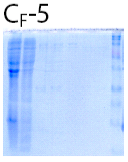
C_F-4



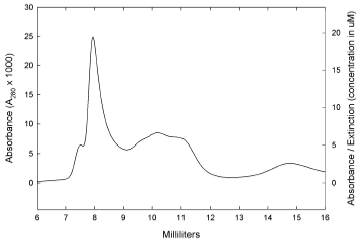
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.35 ml
11.25 ml



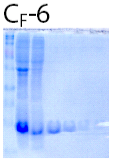
C_F-5



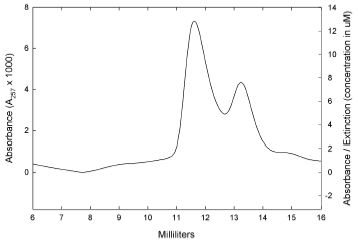
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.35 ml
11.25 ml



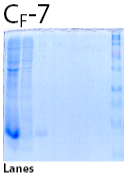
C_F-6



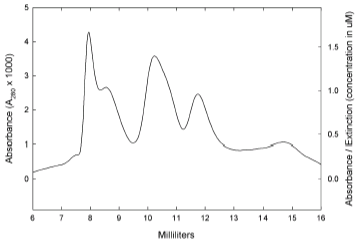
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
11.55 ml
13.35 ml



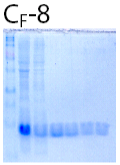
C_F-7



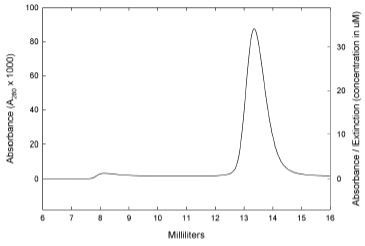
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.65 ml
10.35 ml
11.85 ml



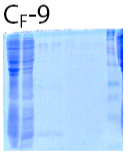
C_F-8



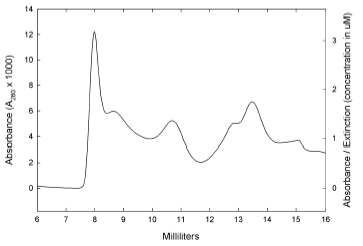
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
13.35 ml
13.65 ml



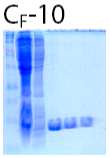
C_F-9



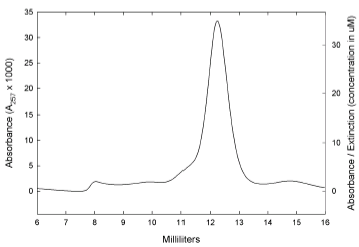
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.65 ml
13.65 ml



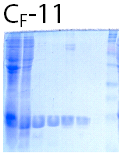
C_F-10



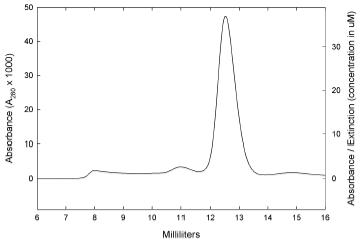
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
12.15 ml



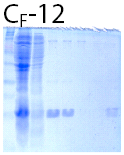
C_F-11



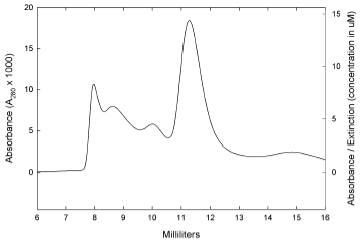
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
12.45 ml
12.75 ml



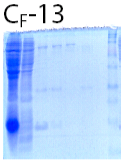
C_F-12



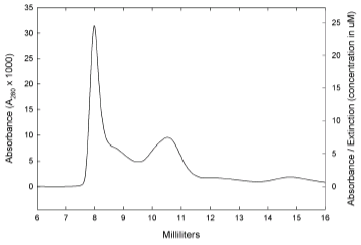
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
8.55 ml
10.05 ml
11.55 ml



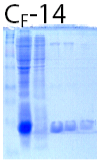
C_F-13



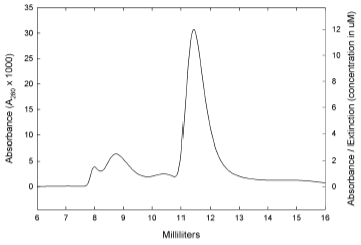
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.65 ml



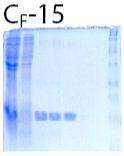
C_F-14



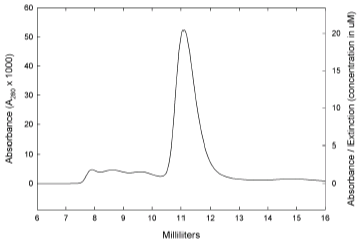
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
11.55 ml



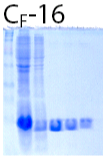
C_F-15



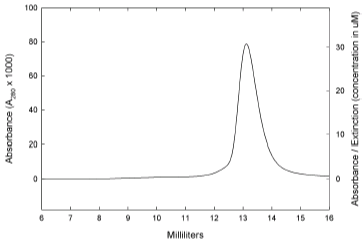
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
11.25 ml



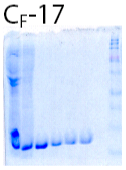
C_F-16



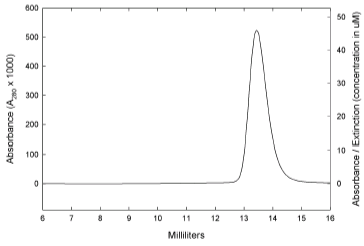
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
13.05 ml



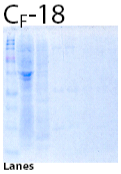
C_F-17



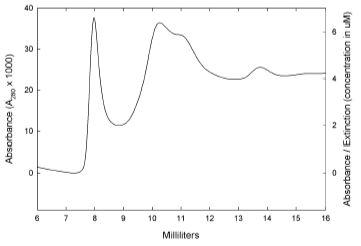
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
13.35 ml
13.65 ml



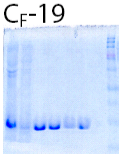
C_F-18



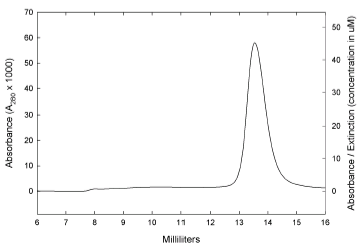
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
10.35 ml
11.25 ml
13.65 ml



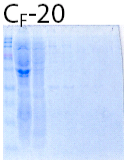
C_F-19



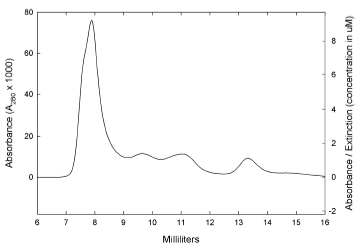
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
13.35 ml
13.65 ml



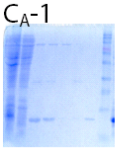
C_F-20



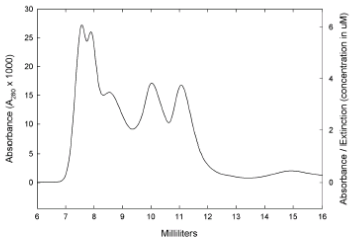
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
9.75 ml
11.25 ml
13.35 ml



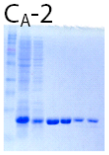
C_A-1



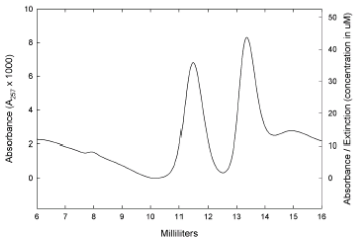
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
8.55 ml
10.05 ml
11.25 ml



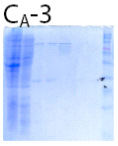
C_A-2



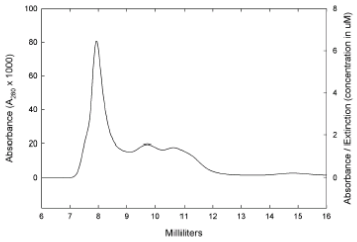
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
11.55 ml
13.35 ml



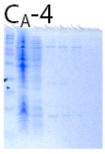
C_A-3



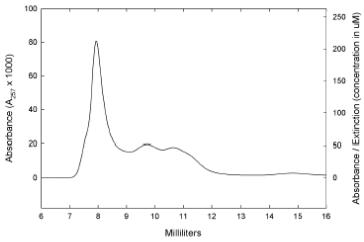
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
9.75 ml
10.65 ml



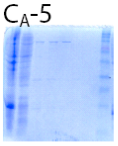
C_A-4



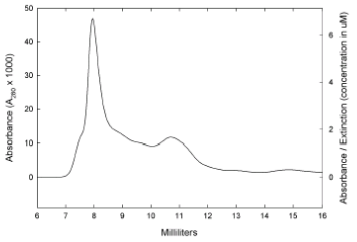
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.35 ml



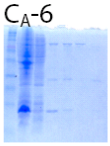
C_A-5



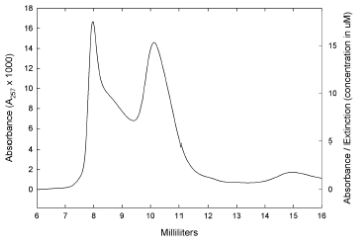
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.65 ml



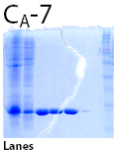
C_A-6



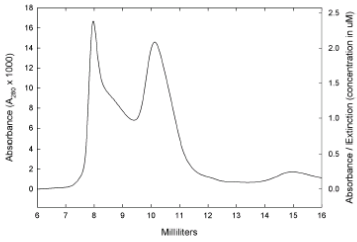
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.05 ml



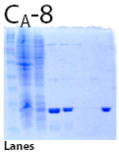
C_A-7



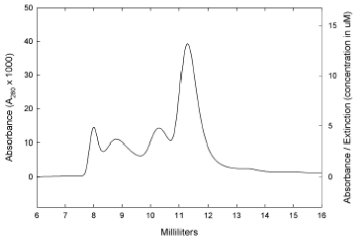
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
10.65 ml



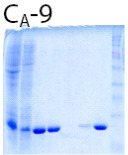
C_A-8



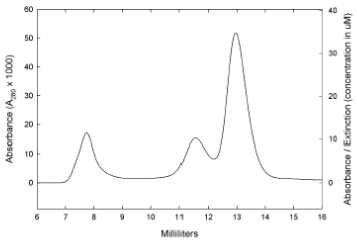
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
8.85 ml
10.35 ml
11.25 ml



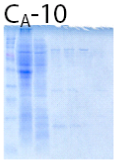
C_A-9



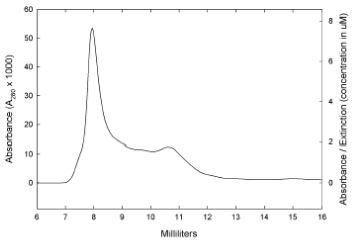
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.65 ml
11.55 ml
13.05 ml



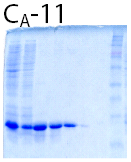
C_A-10



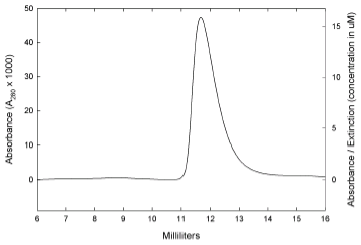
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
10.65 ml



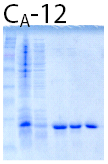
C_A-11



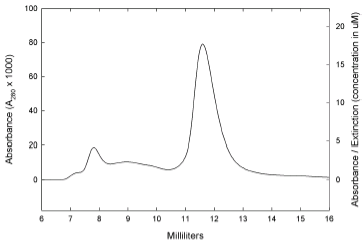
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
11.85 ml



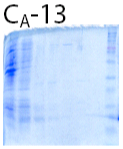
C_A-12



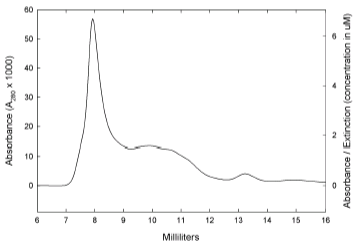
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
11.55 ml



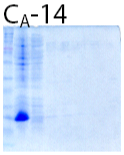
C_A-13



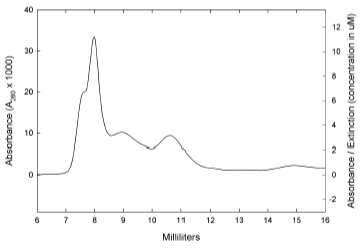
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.65 ml
13.35 ml



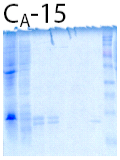
C_A-14



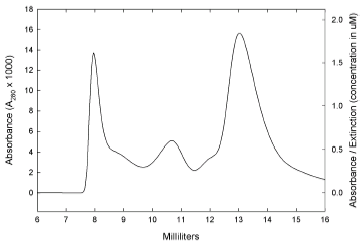
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.65 ml
8.85 ml
10.65 ml



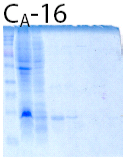
C_A-15



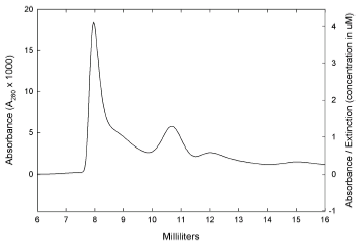
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.65 ml
13.05 ml



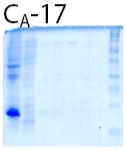
C_A-16



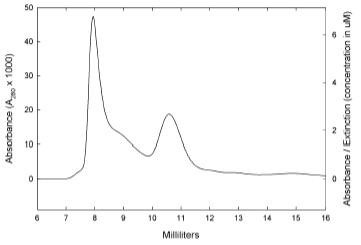
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.65 ml
12.15 ml



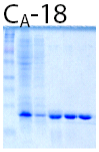
C_A-17



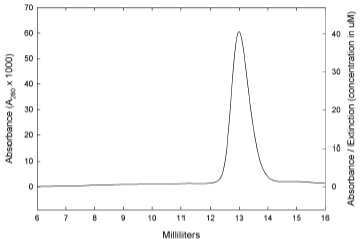
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.65 ml



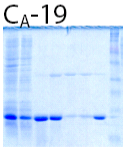
C_A-18



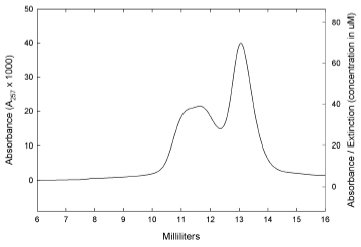
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
13.05 ml



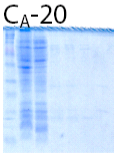
C_A-19



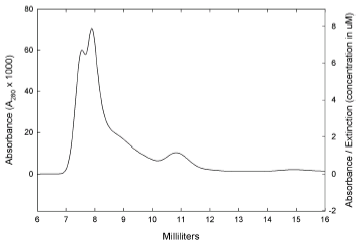
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
10.95 ml
11.85 ml
13.05 ml



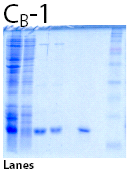
C_A-20



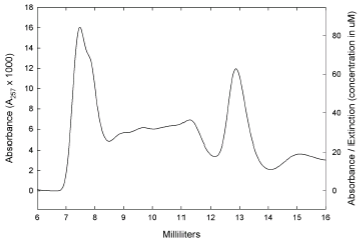
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.65 ml
10.95 ml



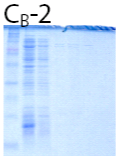
C_B-1



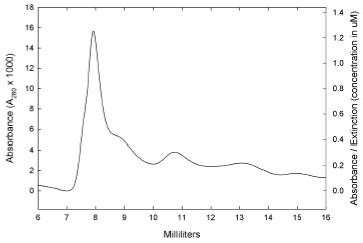
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.65 ml
13.05 ml



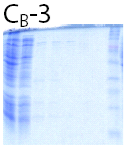
C_B-2



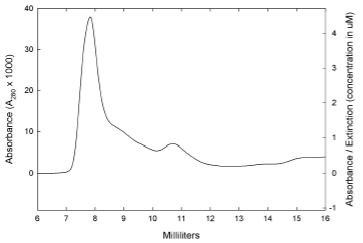
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.65 ml



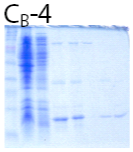
C_B-3



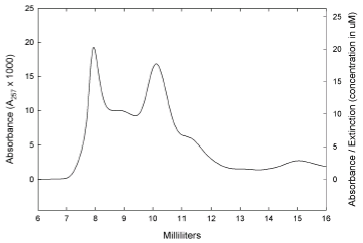
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.65 ml



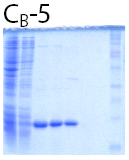
C_B-4



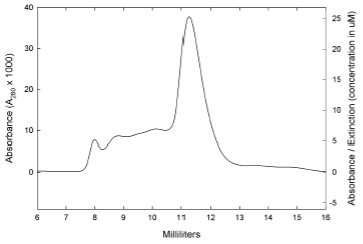
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.05 ml
11.25 ml



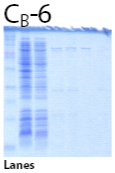
C_B-5



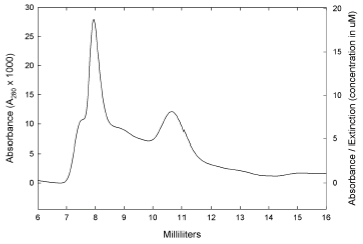
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
11.25 ml



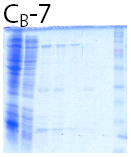
C_B-6



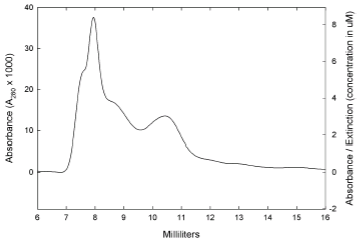
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.65 ml



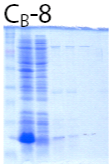
C_B-7



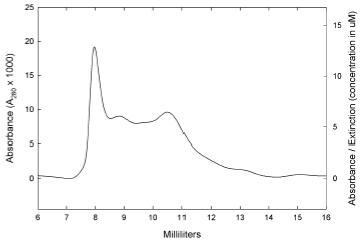
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.35 ml



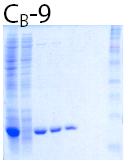
C_B-8



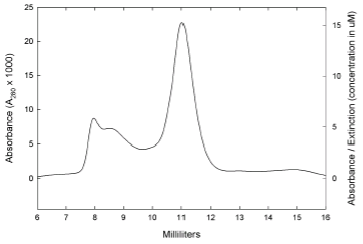
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.65 ml



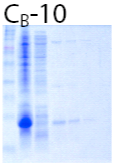
C_B-9



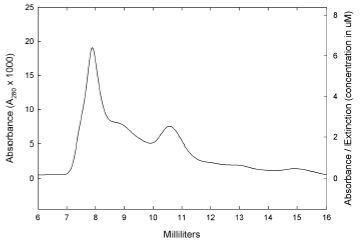
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
10.95 ml



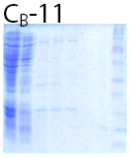
C_B-10



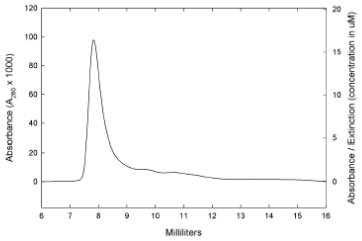
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.65 ml
10.65 ml



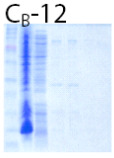
C_B-11



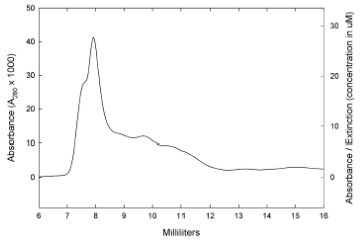
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml



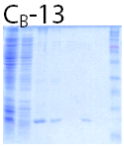
C_B-12



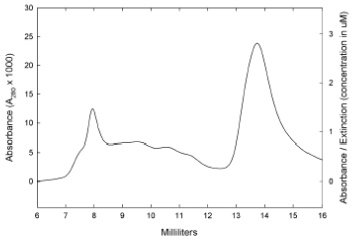
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.65 ml
9.75 ml



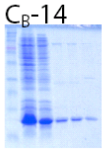
C_B-13



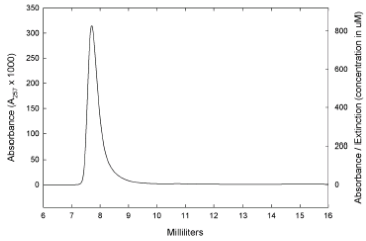
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
13.65 ml



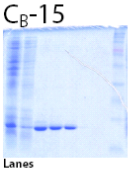
C_B-14



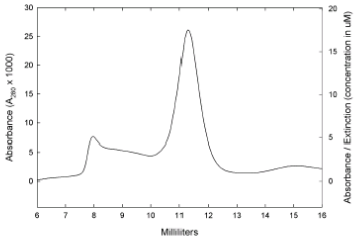
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.65 ml



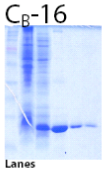
C_B-15



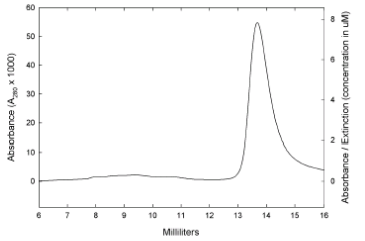
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
11.25 ml



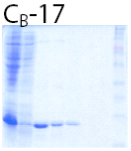
C_B-16



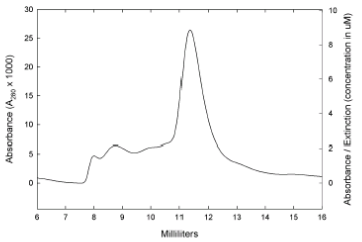
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
13.65 ml



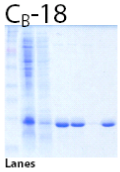
C_B-17



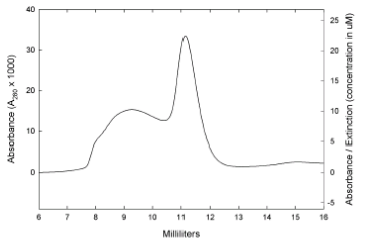
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
11.25 ml



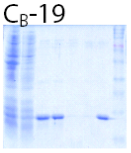
C_B-18



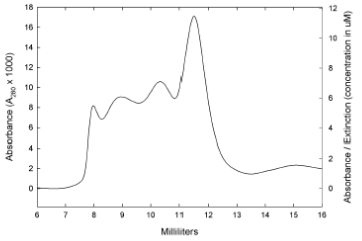
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
9.15 ml
11.25 ml



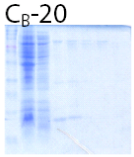
C_B-19



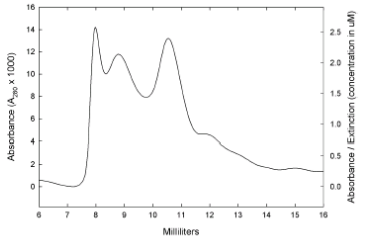
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
8.85 ml
10.35 ml
11.55 ml



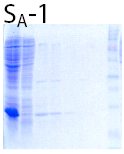
C_B-20



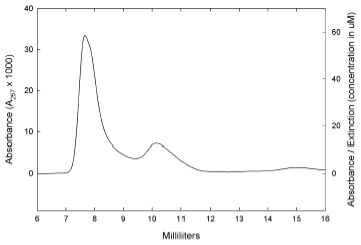
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
8.85 ml
10.65 ml



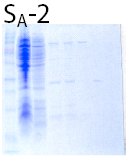
S_A-1



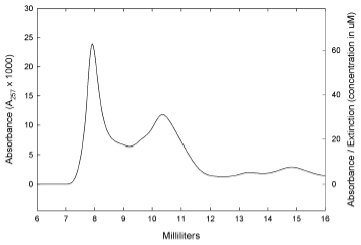
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.65 ml
10.35 ml



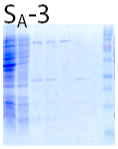
S_A-2



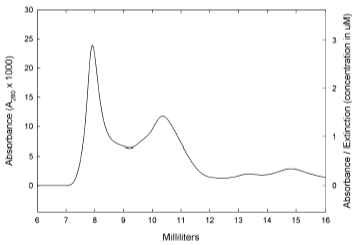
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.35 ml
13.35 ml



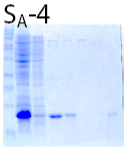
S_A-3



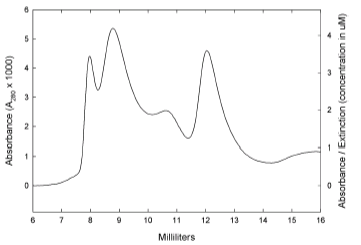
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.65 ml



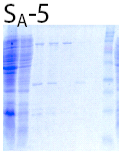
S_A-4



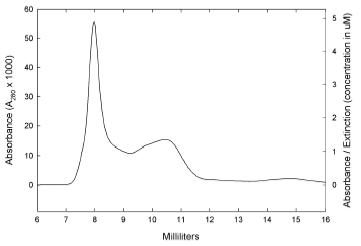
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
8.85 ml
10.65 ml
12.15 ml



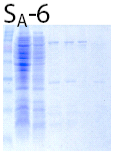
S_A-5



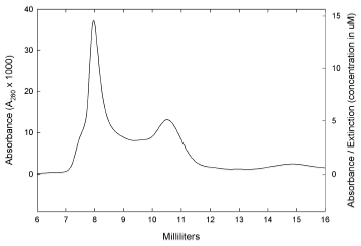
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.35 ml



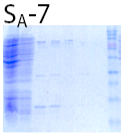
S_A-6



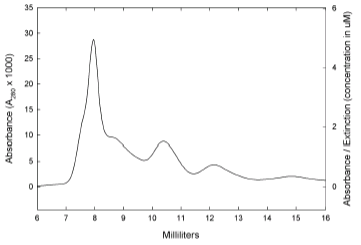
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.65 ml



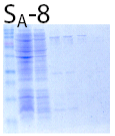
S_A-7



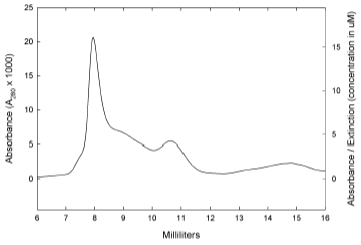
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.35 ml
12.15 ml



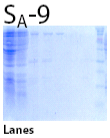
S_A-8



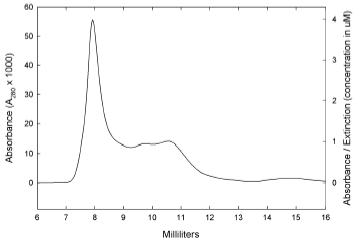
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.65 ml



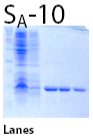
S_A-9



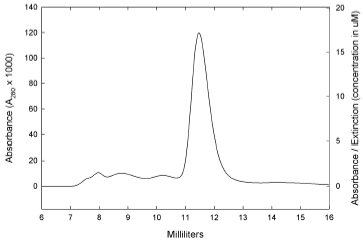
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
9.75 ml
10.65 ml



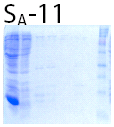
S_A-10



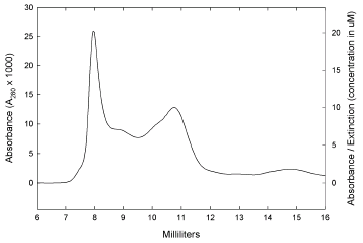
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
11.55 ml



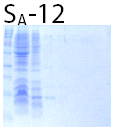
S_A-11



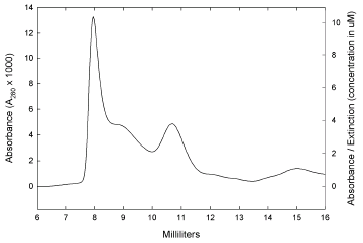
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.95 ml



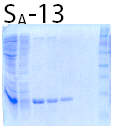
S_A-12



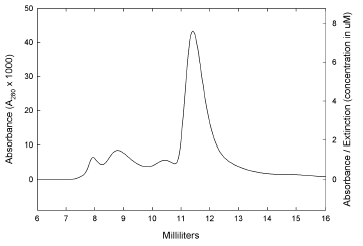
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
8.85 ml
10.65 ml



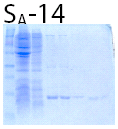
S_A-13



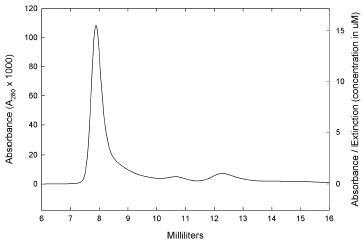
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
11.55 ml



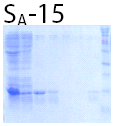
S_A-14



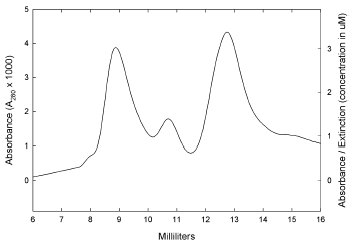
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
12.15 ml
12.45 ml



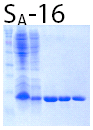
S_A-15



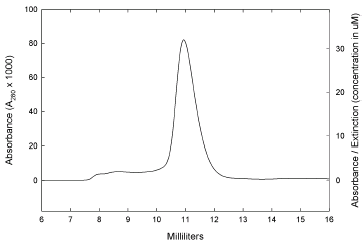
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
8.85 ml
10.65 ml
12.75 ml



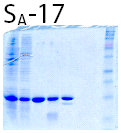
S_A-16



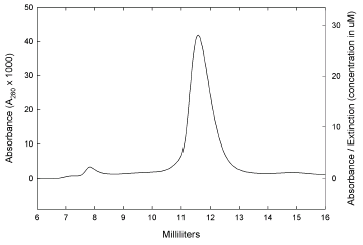
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
10.95 ml



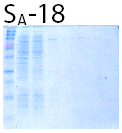
S_A-17



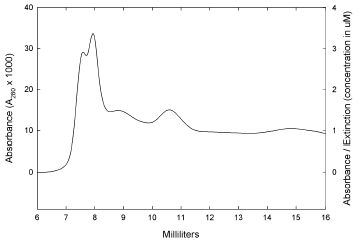
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
11.55 ml



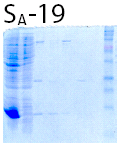
S_A-18



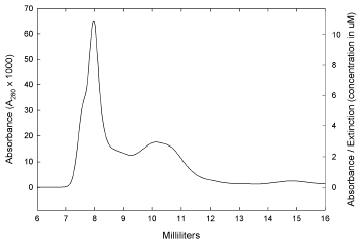
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
8.85 ml
10.65 ml



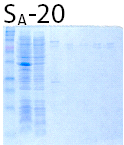
S_A-19



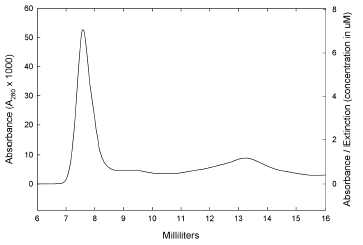
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.35 ml



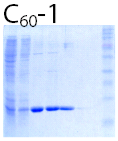
S_A-20



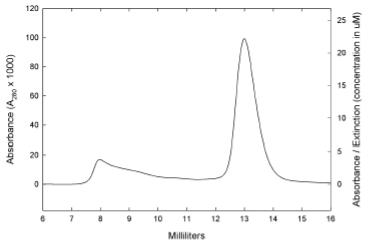
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.65 ml
13.05 ml
13.35 ml



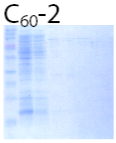
C_{60}-1



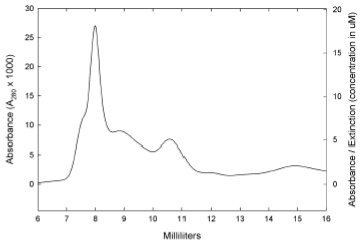
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
13.05 ml



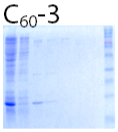
C_{60}-2



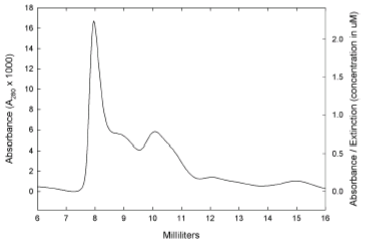
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.65 ml



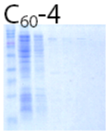
C_{60}-3



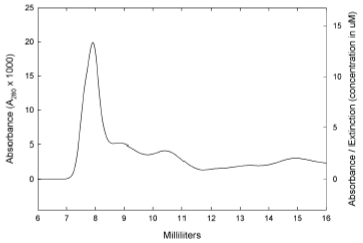
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.65 ml
12.15 ml



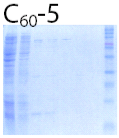
C_{60}-4



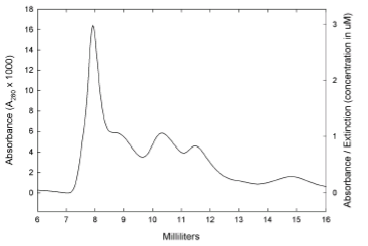
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.65 ml



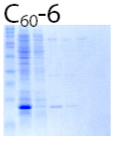
C_{60}-5



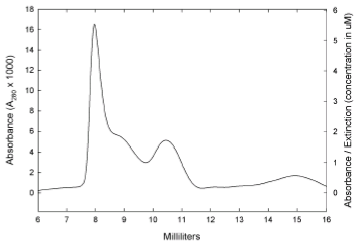
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.35 ml
11.55 ml



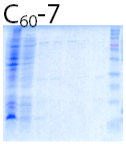
C_{60}-6



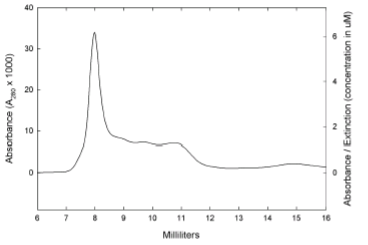
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.35 ml



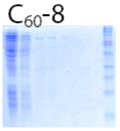
C_{60}-7



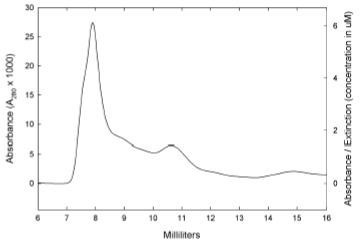
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
9.75 ml
10.95 ml



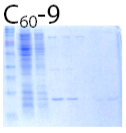
C_{60}-8



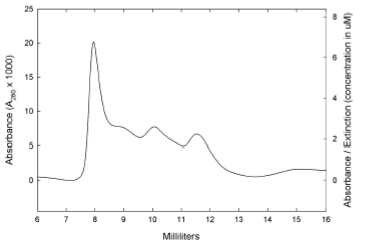
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.65 ml



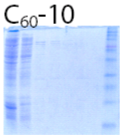
C_{60}-9



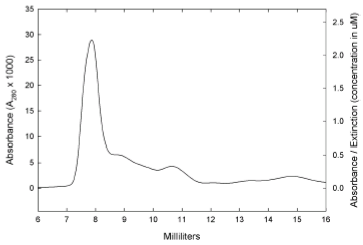
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.05 ml
11.55 ml



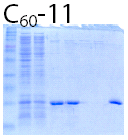
C_{60}-10



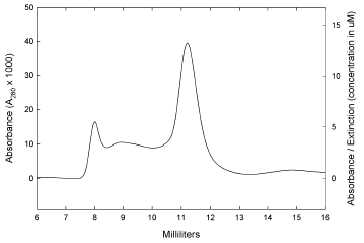
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.65 ml
10.65 ml



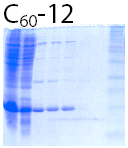
C_{60}-11



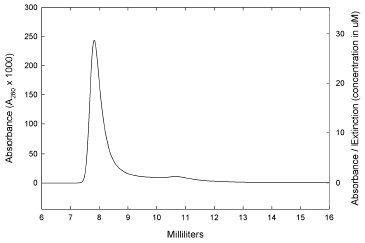
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
8.85 ml
11.25 ml



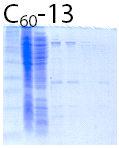
C_{60}-12



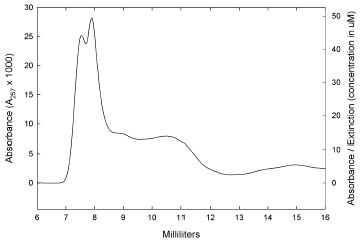
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
9.75 ml



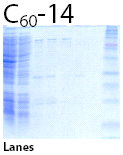
C_{60}-13



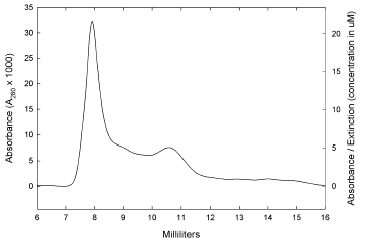
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.65 ml
10.65 ml



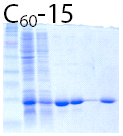
C_{60}-14



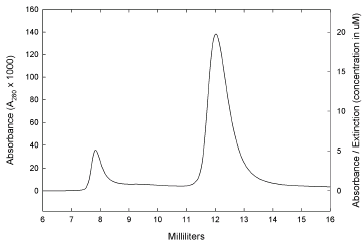
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.65 ml



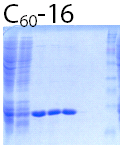
C_{60}-15



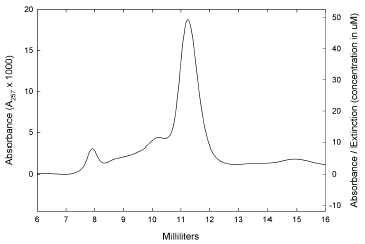
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
12.15 ml



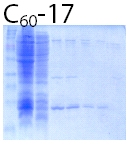
C_{60}-16



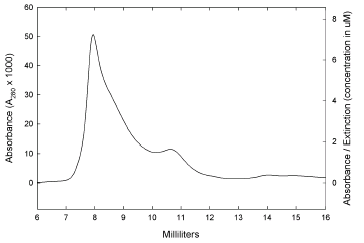
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
11.25 ml



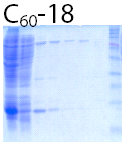
C_{60}-17



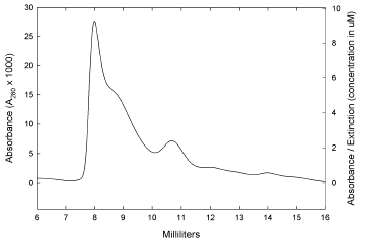
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
8.85 ml
10.65 ml



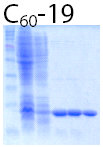
C_{60}-18



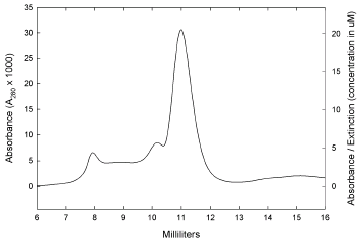
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
8.85 ml
10.65 ml



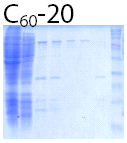
C_{60}-19



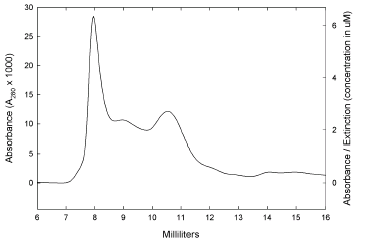
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
10.95 ml



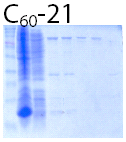
C_{60}-20



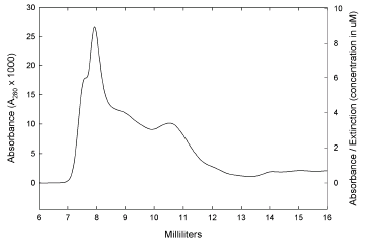
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
8.85 ml
10.65 ml



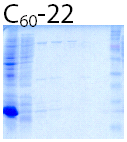
C_{60}-21



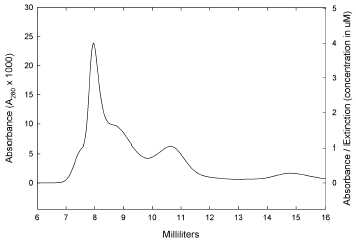
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
8.85 ml
10.65 ml



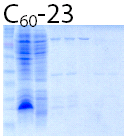
C_{60}-22



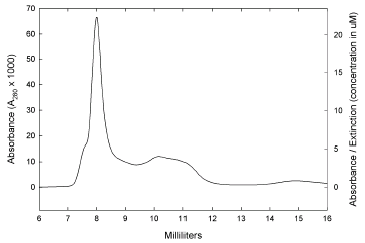
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.65 ml



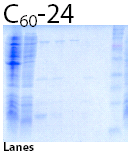
C_{60}-23



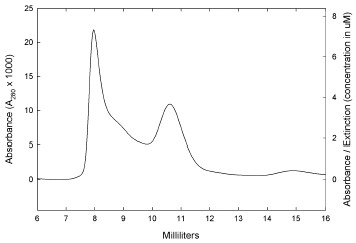
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.05 ml
11.25 ml



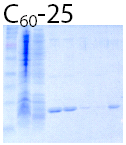
C_{60}-24



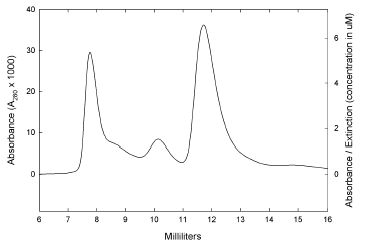
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.65 ml



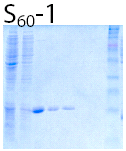
C_{60}-25



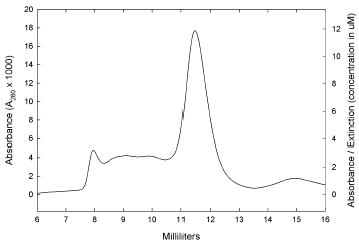
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.05 ml
11.85 ml



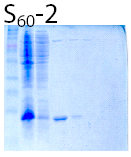
S_{60}-1



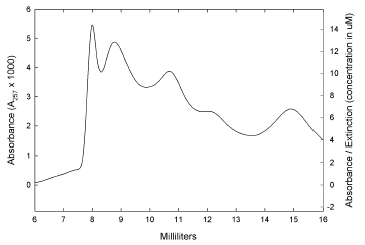
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
11.55 ml



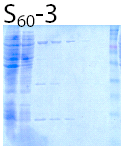
S_{60}-2



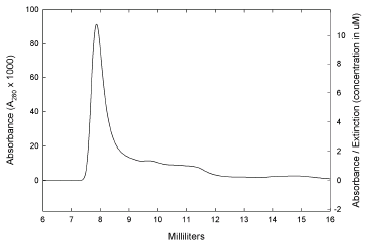
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
8.85 ml
10.65 ml
12.15 ml



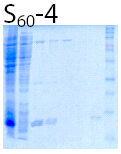
S_{60}-3



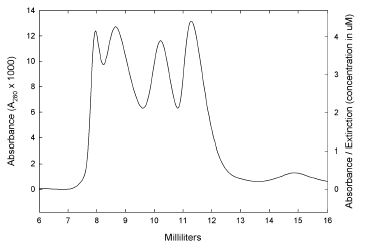
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
9.75 ml
11.25 ml



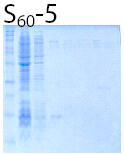
S_{60}-4



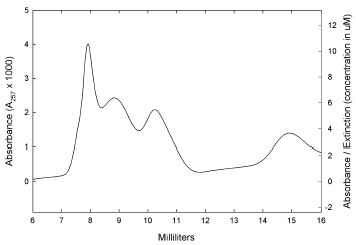
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
8.55 ml
10.35 ml
11.25 ml



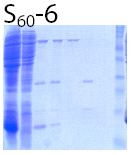
S_{60}-5



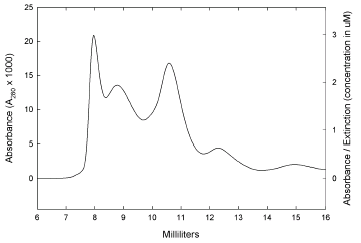
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
8.85 ml
10.35 ml



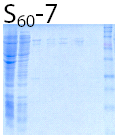
S_{60}-6



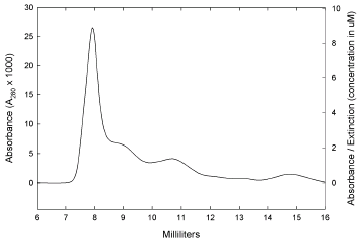
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
8.85 ml
10.65 ml
12.45 ml



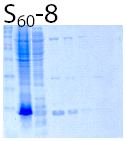
S_{60}-7



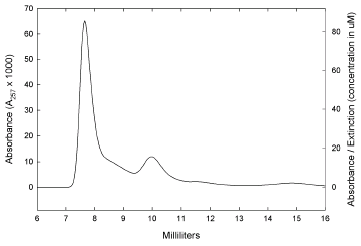
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
8.85 ml
10.65 ml



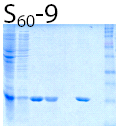
S_{60}-8



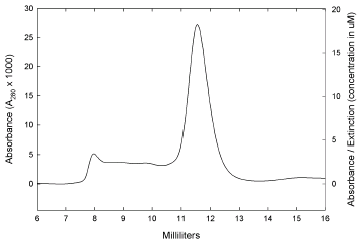
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.05 ml
11.55 ml



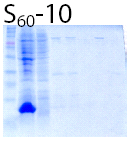
S_{60}-9



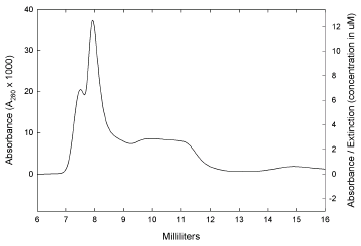
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
11.55 ml



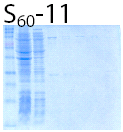
S_{60}-10



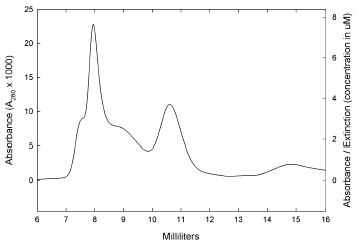
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.35 ml
7.95 ml
11.25 ml



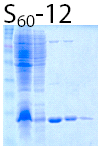
S_{60}-11



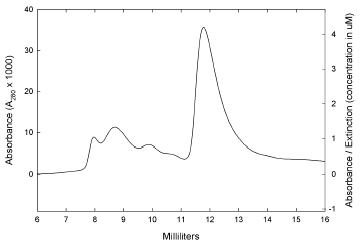
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
8.85 ml
10.65 ml



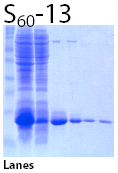
S_{60}-12



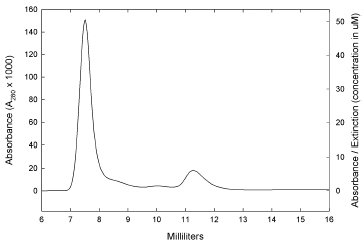
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
11.85 ml



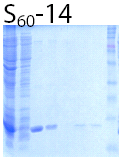
S_{60}-13



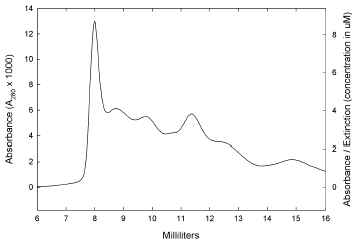
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.65 ml
11.25 ml



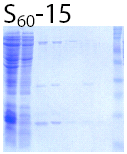
S_{60}-14



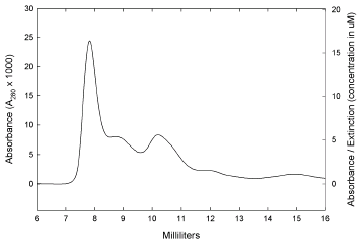
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
10.05 ml
11.55 ml
12.45 ml



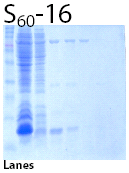
S_{60}-15



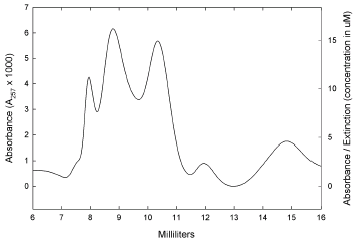
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.35 ml
12.15 ml



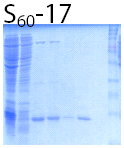
S_{60}-16



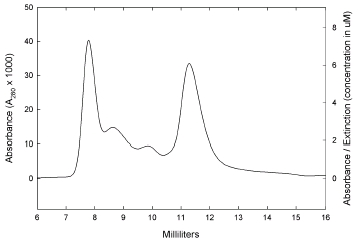
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
8.85 ml
10.35 ml
11.85 ml



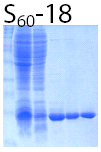
S_{60}-17



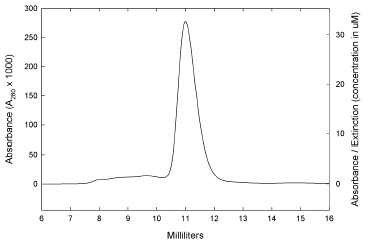
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.65 ml
11.25 ml



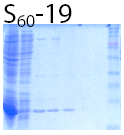
S_{60}-18



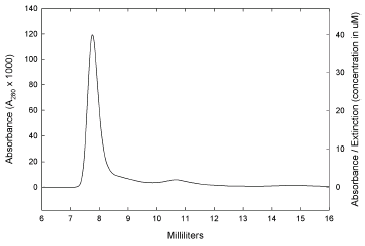
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
10.96 ml



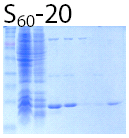
S_{60}-19



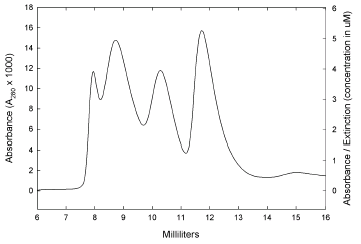
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.65 ml
10.65 ml



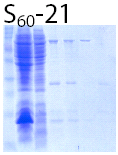
S_{60}-20



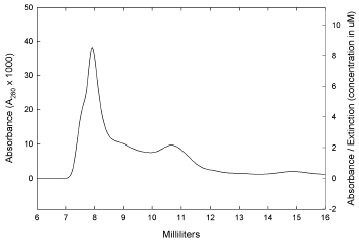
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
8.85 ml
10.35 ml
11.85 ml



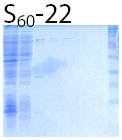
S_{60}-21



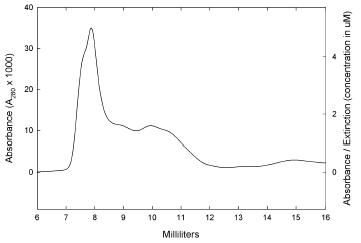
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.65 ml



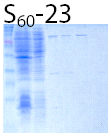
S_{60}-22



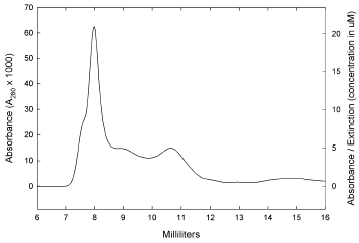
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.05 ml
10.65 ml



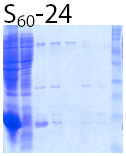
S_{60}-23



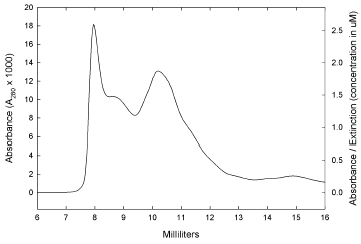
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
10.65 ml



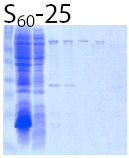
S_{60}-24



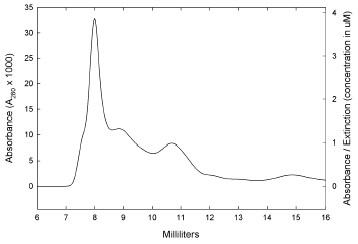
Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
8.85 ml
10.05 ml
11.25 ml



S_{60}-25



Lanes
Whole cells
Soluble lysate
Ni elution
SEC load
7.95 ml
8.85 ml
10.65 ml



BIBLIOGRAPHY

Ahmad ST, Natochin M, Barren B, Artemyev NO, O'Tousa JE. (2006) Heterologous expression of bovine rhodopsin in Drosophila photoreceptor cells. Invest Opthalmol Vis Sci. 47(9):3722-3728.

Ahmad ST, Natochin M, Artemyev NO, O'Tousa JE. (2007) The Drosophila rhodopsin cytoplasmic tail domain is required for maintenance of rhabdomere structure. FASEB J. 21(2):449-455.

Alloway PG, Howard L, Dolph PJ. (2000) The formation of stable rhodopsin-arrestin complexes induces apoptosis and photoreceptor cell degeneration. Neuron. 28(1):129-138.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402.

Baumann O. (2004) Spatial pattern of nonmuscle myosin-II distribution during the development of the Drosophila compound eye and implications for retinal morphogenesis. Dev Biol. 269(2)519-33.

Birrane G, Chung J, Ladias JA. (2003) Novel mode of ligand recognition by the erbin PDZ domain. J Biol Chem. 278(3):1399-1402.

Bork P, Sudol M. (1994) The WW domain: a signaling site in dystrophin? Trends Biochem Sci. 19(12):531-533.

Casari G, Sander C, Valencia A. (1995) A method to predict functional residues in proteins. Nat Struct Biol 2:171-178.

Chang HY, Ready DF. (2000) Rescue of photoreceptor degeneration in rhodopsin-null Drosophila mutants by activated Rac1. Science. 290(5498):1978-1980.

Chen HI, Sudol M. (1995) The WW domain of Yes-associated protein binds a proline-rich ligand that differs from the consensus established for Src homology 3-binding modules. PNAS. 92(17):7819-7823.

Colley NJ, Baker EK, Stamnes MA, Zuker CS. (1991) The cyclophilin homolog ninaA is required in the secretory pathway. Cell. 67(2):255-263.

Cover TM, Thomas JA. (1991) Elements of Information Theory. John Wiley and Sons, Inc.

Daaka Y, Luttrell LM, Ahn S, Della Rocca GJ, Ferguson SS, Caron MG, Lefkowitz RJ. (1998) Essential role for G protein-coupled receptor endocytosis in the activation of mitogen-activated protein kinase. J Biol Chem. 273(2):685-688.

Daniels DL, Cohen AR, Anderson JM, Brunger AT. (1998) Crystal structure of the hCASK PDZ domain reveals the structural basis of class II PDZ domain target recognition. Nat Struct Biol. 5(4):317-325.

Dantas G, Kuhlman B, Callender D, Wong M, Baker D. (2003) A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. J Mol Biol. 332(2):449-460.

Doyle DA, Lee A, Lewis J, Kim E, Sheng M, MacKinnon R. (1996) Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. Cell. 85(7):1067-1076.

Dwyer MA, Looger LL, Hellinga HW. (2004) Computational design of a biologically active enzyme. Science. 304(5679):1967-1971.

Eisenmesser EZ, Millet O, Labeikovsky W, Korzhnev DM, Wolf-Watz M, Bosco DA, Skalicky JJ, Kay LE, Kern D. (2005) Intrinsic dynamics of an enzyme underlies catalysis. Nature. 438(7064):117-121.

Elkins JM, Schoch GA, Smee CEA, Berridge G, Salah E, Sundstrom M, Edwards A, Arrowsmith C, Weigelt J, Doyle DA. (2005) Deposited to PDB, to be published.

Fanning AS, Anderson JM. (1999) PDZ domains: fundamental building blocks in the organization of protein complexes at the plasma membrane. J Clin Invest. 103(6):767-772.

Feiler R, Bjornson R, Kirschfeld K, Mismer D, Rubin GM, Smith DP, Socolich M, Zuker CS. (1992) Ectopic expression of ultraviolet-rhodopsins in the blue photoreceptor cells of Drosophila: visual physiology and photochemistry of transgenic animals. J Neurosci. 12(10):3862-3868.

Forster M, Heath A, Afzal M. (1999) Application of distance geometry to 3D visualization of sequence relationships. Bioinformatics. 15(1):89-90.

Friedland N, Hung LW, Cheyette B, Moon RT, Earnest TN. (2005) Deposited to PDB, to be published.

Fuh G, Pisabarro MT, Li Y, Quan C, Lasky LA, Sidhu SS. (2000) Analysis of PDZ domain-ligand interactions using carboxyl-terminal phage display. J Biol Chem. 275(28):21486-21491.

Gether U. (2000) Uncovering molecular mechanisms involved in activation of G protein-coupled receptors. Endocr Rev. 21(1):90-113.

Gilman AG. (1987) G proteins: transducers of receptor-generated signals. Annu Rev Biochem. 56:615-649.

Goldsmith TH, Marks BC, Bernard GD. (1986) Separation and identification of geometric isomers of 3-hydroxyretinoids and occurrence in the eyes of insects. Vision Res. 26(11):1763-1769.

Hammes-Schiffer S, Benkovic SJ. (2006) Relating protein motion to catalysis. Annu Rev Biochem. 75:519-541.

Hardie RC. (1986) The photoreceptor array of the dipteran retina. Trends Neurosci. 9:419-423.

Hardie RC. (1996) INDO-1 measurements of absolute resting and light-induced Ca2+ concentration in Drosophila photoreceptors. J Neurosci. 16(9):2924-2933.

Hardie RC, Raghu P. (2001) Visual transduction in Drosophila. Nature. 413(6852):186-193.

Hatley ME, Lockless SW, Gibson SK, Gilman AG, Ranganathan R. (2003) Allosteric determinants in guanine nucleotide-binding proteins. Proc Natl Acad Sci U S A. 100(24):14445-14450.

Huang X, Poy F, Zhang R, Joachimiak A, Sudol M, Eck MJ. (2000) Structure of a WW domain containing fragment of dystrophin in complex with beta-dystroglycan. Nat Struct Biol. 7(8):634-638.

Huber A, Smith DP, Zuker CS, Paulsen R. (1990) Opsin of Calliphora peripheral photoreceptors R1-6. Homology with Drosophila Rh1 and posttranslational processing. JBC. 265(29):17906-17910.

Im YJ, Park SH, Rho SH, Lee JH, Kang GB, Sheng M, Kim E, Eom SH. (2003) Crystal structure of GRIP1 PDZ6-peptide complex reveals the structural basis for class II PDZ target recognition and PDZ domain-mediated multimerization. J Biol Chem. 278(10):8501-8507.

Im YJ, Lee JH, Park SH, Park SJ, Rho SH, Kang GB, Kim E, Eom SH. (2003) Crystal structure of the Shank PDZ-ligand complex reveals a class I PDZ interaction and a novel PDZ-PDZ dimerization. J Biol Chem. 278(48):48099-48104.

Jain RK, Ranganathan R. (2004) Local complexity of amino acid interactions in a protein core. Proc Natl Acad Sci U S A. 101(1):111-116.

Jelen F, Oleksy A, Smietana K, Otlewski J. (2003) PDZ domains - common players in the cell signaling. Acta Biochim Pol. 50(4):985-1017.

John DM, Weeks KM. (2000) van't Hoff enthalpies without baselines. Protein Sci. 9(7):1416-1419.

Kanelis V, Rotin D, Forman-Kay JD. (2001) Solution structure of a Nedd4 WW domain-ENaC peptide complex. Nat Struct Biol. 8(5):407-412.

Kang BS, Cooper D, Devedjiev Y, Derewenda U, Derewenda ZS. (2003) Molecular Roots of Degenerate Specificity in Syntenin's PDZ2 Domain. Reassessment of the PDZ Recognition Paradigm. Structure. 11(7):845-853.

Karnik S, Gogonea C, Patil S, Saad Y, Takezako T. (2003) Activation of G-protein-coupled receptors: a common molecular mechanism. Trends Endocrinol Metab. 14(9):431-437.

Karthikeyan S, Leung T, Birrane G, Webster G, Ladias JA. (2001) Crystal structure of the PDZ1 domain of human Na(+)/H(+) exchanger regulatory factor provides insights into the mechanism of carboxyl-terminal leucine recognition by class I PDZ domains. J Mol Biol. 308(5):963-973.

Katanosaka K, Tokunaga F, Kawamura S, Ozaki K. (1998) N-linked glycosylation of Drosophila rhodopsin occurs exclusively in the amino-terminal domain and functions in rhodopsin maturation. FEBS Lett. 424(3):149-154.

Kato Y, Miyakawa T, Kurita J, Tanokura M. (2006) Structure of FBP11 WW1-PL ligand complex reveals the mechanism of proline-rich ligand recognition by group II/III WW domains. J Biol Chem. 281(52):40321-40329.

Kimple ME, Siderovski DP, Sondek J. (2001) Functional relevance of the disulfide-linked complex of the N-terminal PDZ domain of InaD with NorpA. EMBO J. 20(16):4414-4422.

Kiselev A, Subramaniam S. (1994) Activation and regeneration of rhodopsin in the insect visual cycle. Science. 266(5189):1369-73.

Kiselev A, Socolich M, Vinos J, Hardy RW, Zuker CS, Ranganathan R. (2000) A molecular pathway for light-dependent photoreceptor apoptosis in Drosophila. Neuron. 28(1):139-152.

Kolakowski LF. (1994) GCRDb: a G-protein-coupled receptor database. Receptors Channels. 2(1):1-7.

Kosloff M, Elia N, Joel-Almagor T, Timberg R, Zars TD, Hyde DR, Minke B, Selinger Z. (2003) Regulation of light-dependent Gqalpha translocation and morphological changes in fly photoreceptors. EMBO J. 22(3):459-468.

Krupnick JG, Benovic JL. (1998) The role of receptor kinases and arrestins in G protein-coupled receptor regulation. Annu Rev Pharmacol Toxicol. 38:289-319.

Kuhlman B, Dantas G, Ireton G, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. Science. 302(5649):1364-1368.

Kumar JP, Bowman J, O'Tousa JE, Ready DF. (1997) Rhodopsin replacement rescues photoreceptor structure during a critical developmental window. Dev Biol. 188(1):43-47.

Labeikovsky W, Eisenmesser EZ, Bosco DA, Kern D. (2007) Structure and dynamics of pin1 during catalysis by NMR. J Mol Biol. 367(5):1370-1381.

Lee SJ, Montell C. (2004) Suppression of constant-light-induced blindness but not retinal degeneration by inhibition of the rhodopsin degradation pathway. Curr Biol. 14(23):2076-2085.

Lee YJ, Dobbs MB, Verardi ML, Hyde DR. (1990) dgq: a Drosophila gene encoding a visual system-specific G alpha molecule. Neuron. 5(6):889-898.

Lehmann M, Kostrewa D, Wyss M, Brugger R, D'Arcy A, Pasamontes L, van Loon AP. (2000) From DNA sequence to improved functionality: using protein sequence comparisons to rapidly design a thermostable consensus phytase. Protein Eng. 13(1):49-57.

Leonard DS, Bowman VD, Ready DF, Pak WL. (1992) Degeneration of photoreceptors in rhodopsin mutants of Drosophila. J Neurobiol. 23(6):605-626.

Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. Science. 286(5438):295-299.

Looger LL, Dwyer MA, Smith JJ, Hellinga HW. (2003) Computational design of receptor and sensor proteins with novel functions. Nature. 423(6936):185-190.

Lu PJ, Zhou XZ, Shen M, Lu KP. (1999) Function of WW domains as phosphoserine- or phosphothreonine-binding modules. Science. 283(5406):1325-1328.

Macias MJ, Hyvonen M, Baraldi E, Schultz J, Sudol M, Saraste M, Oschkinat H. (1996) Structure of the WW domain of a kinase-associated protein complexed with a proline-rich peptide. Nature. 382(6592):646-649.

Ostroy SE, Wilson M, Pak WL. (1974) Drosophila rhodopsin: photochemistry, extraction and differences in the norpAP12 phototransduction mutant. Biochem Biophys Res Commun. 59(3):960-966.

O'Tousa JE, Baehr W, Martin RL, Hirsh J, Pak WL, Applebury ML. (1985) The Drosophila ninaE gene encodes an opsin. Cell. 40(4):839-850.

Otte L, Wiedemann U, Schlegel B, Pires JR, Beyermann M, Schmieder P, Krause G, Volkmer-Engert R, Schneider-Mergener J, Oschkinat H (2003) WW domain sequence activity relationships identified using ligand recognition propensities of 42 WW domains. Protein Sci. 12:491-500.

Ozaki K, Nagatani H, Ozaki M, Tokunaga F. (1993) Maturation of major Drosophila rhodopsin, ninaE, requires chromophore 3-hydroxyretinal. Neuron. 10(6):1113-1119.

Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Le Trong I, Teller DC, Okada T, Stenkamp RE, Yamamoto M, Miyano M. (2000) Crystal structure of rhodopsin: A G protein-coupled receptor. Science. 289(5480):739-745.

Papagrigoriou E, Berridge G, Johansson C, Colebrook S, Salah E, Burgess N, Smee C, Savitsky P, Bray J, Schoch G, Phillips C, Gileadi C, Soundarajan M, Yang X, Elkins J, Gorrec F, Turnbull A, Edwards A, Arrowsmith C, Weigelt J, Sundstrom M, Doyle D, Structural Genomics Consortium. (2005) Deposited to PDB, to be published.

Paulsen R, Schwemer J. (1979) Vitamin A deficiency reduces the concentration of visual pigment protein within blowfly photoreceptor membranes. Biochim Biophys Acta. 557(2):385-390.

Popovych N, Sun S, Ebright RH, Kalodimos CG. (2006) Dynamically driven protein allostery. Nat Struct Mol Biol. 13(9):831-838.

Ranganathan R, Harris GL, Stevens CF, Zuker CS. (1991) A Drosophila mutant defective in extracellular calcium-dependent photoreceptor deactivation and rapid desensitization. Nature. 354(6350):230-232.

Ranganathan R, Malicki DM, Zuker CS. (1995) Signal transduction in Drosophila photoreceptors. Annu Rev Neurosci. 18:283-317.

Ranganathan R, Lu KP, Hunter T, Noel JP. (1997) Structural and functional analysis of the mitotic rotamase Pin1 suggests substrate recognition is phosphorylation dependent. Cell. 89(6):875-886.

Ranganathan R, Ross EM. (1997) PDZ domain proteins: scaffolds for signaling complexes. Curr Biol. 7(12):R770-R773.

Reina J, Lacroix E, Hobson SD, Fernandez-Ballester G, Rybin V, Schwab MS, Serrano L, Gonzalez C. (2002) Computer-aided design of a PDZ domain to recognize new target sequences. Nat Struct Biol. 9(8):621-627.

Russ WP, Lowery DM, Mishra P, Yaffe MB, Ranganathan R. (2005) Natural-like function in artificial WW domains. Nature. 437(7058):579-583.

Salcedo E, Huber A, Henrich S, Chadwell LV, Chou WH, Paulsen R, Britt SG. (1999) Blue- and green-absorbing visual pigments of Drosophila: ectopic expression and physiological characterization of the R8 photoreceptor cell-specific Rh5 and Rh6 rhodopsins. J Neurosci. 19(24):10716-10726.

Satoh AK, Ready DF. (2005) Arrestin1 mediates light-dependent rhodopsin endocytosis and cell survival. Curr Biol. 15(19):1722-1733.

Sharma, R. (2006) Logic and mechanism of an evolutionarily conserved interaction in PDZ domains. Dissertation, University of Texas Southwestern Medical Center.

Shih P, Kirsch JF. (1995) Design and structural analysis of an engineered thermostable chicken lysozyme. Protein Sci. 4(10):2063-2072.

Shulman AI, Larson C, Mangelsdorf DJ, Ranganathan R. (2004) Structural determinants of allosteric ligand activation in RXR heterodimers. Cell. 116(3):417-429.

Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R. (2005) Evolutionary information for specifying a protein fold. Nature. 437(7058):512-518.

Songyang Z, Fanning AS, Fu C, Xu J, Marfatia SM, Chishti AH, Crompton A, Chan AC, Anderson JM, Cantley LC. (1997) Recognition of unique carboxy-terminal motifs by distinct PDZ domains. Science. 275(5296):73-77.

Stamnes MA, Shieh BH, Chuman L, Harris GL, Zuker CS. (1991) The cyclophilin homolog ninaA is a tissue-specific integral membrane protein required for the proper synthesis of a subset of Drosophila rhodopsins. Cell. 65(2):219-227.

Steipe B, Schiller B, Pluckthun A, Steinbacher S. (1994) Sequence statistics reliably predict stabilizing mutations in a protein domain. J Mol Biol. 240(3):188-192.

Stiffler MA, Grantcharova VP, Sevecka M, MacBeath G. (2006) Uncovering quantitative protein interaction networks for mouse PDZ domains using protein microarrays. J Am Chem Soc. 128(17)5913-5922.

Sudol M, Hunter T. (2000) NeW Wrinkles for an old domain. Cell. 103(7):1001-1004.

Suel GM, Lockless SW, Wall MA, Ranganathan R. (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat Struct Biol.10(1):59-69.

Szuts EZ, Harosi FI. (1991) Solubility of retinoids in water. Arch Biochem Biophys. 287(2):297-304.

Verdecia MA, Bowman ME, Lu KP, Hunter T, Noel JP. (2000) Structural basis for phosphoserine-proline recognition by group IV WW domains. Nat Struct Biol. 7(8):639-643.

Von Ossowski I, Oksanen E, Von Ossowski L, Cai C, Sundberg M, Goldman A, Keinanen K. (2006) Crystal structure of the second PDZ domain of SAP97 in complex with a GluR-A C-terminal peptide. Febs J. 273(22):5219-5229.

Wang Q, Buckle AM, Foster NW, Johnson CM, Fersht AR (1999) Design of highly stable functional GroEL minichaperones. Protein Sci 8(10):2186-2193.

Webel R, Menon I, O'Tousa J, Colley N. (2000) Role of asparagine-linked oligosaccharides in rhodopsin maturation and association with its molecular chaperone, NinaA. J Biol Chem. 275(32):24752-24759.

Yoon J, Ben-Ami HC, Hong YS, Park S, Strong LL, Bowman J, Geng C, Baek K, Minke B, Pak WL. (2000) Novel mechanism of massive photoreceptor degeneration caused by mutations in the trp gene of Drosophila. J Neurosci. 20(2):649-659.

Zuker CS, Cowman AF, Rubin GM. (1985) Isolation and structure of a rhodopsin gene from D. melanogaster. Cell. 40(4):851-858.

Zuker CS. (1996) The biology of vision in Drosophila. Proc Natl Acad Sci U S A. 93(2):571-576.

VITAE

Christopher Lee Larson was born in Houston, Texas, on March 1, 1978, the son of Harold Larson Jr. and Janice Larson. He graduated from the High School for Health Professions in Houston, Texas in 1996 and received a B.S. in Biochemical and Biophysical Sciences at the University of Houston in 2000.

Permanent Address: 5326 Fleetwood Oaks #248
Dallas, TX 75235