Purdue University
Purdue e-Pubs

Open Access Dissertations | Theses and Dissertations

Fall 2013

Generation And Statistical Modeling Of Active Protein Chimeras: A Sequence Based Approach

Nicholas Fico
Purdue University

Follow this and additional works at: https://docs.lib.purdue.edu/open_access_dissertations

 Part of the Molecular Biology Commons, and the Statistics and Probability Commons

Recommended Citation
Fico, Nicholas, "Generation And Statistical Modeling Of Active Protein Chimeras: A Sequence Based Approach" (2013). *Open Access Dissertations*. 144.
https://docs.lib.purdue.edu/open_access_dissertations/144

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

Graduate School ETD Form 9
(Revised 12/07)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Nicholas J. Fico

Entitled
GENERATION AND STATISTICAL MODELING OF ACTIVE PROTEIN CHIMERAS: A SEQUENCE BASED APPROACH

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Daisuke Kihara
Chair

Alan M. Friedman

Cynthia Stauffacher

Clinton Chapple

To the best of my knowledge and as understood by the student in the *Research Integrity and Copyright Disclaimer (Graduate School Form 20)*, this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Approved by Major Professor(s): Alan M. Friedman

Approved by: Peter J. Hollenbeck 10/08/2013
Head of the Graduate Program Date

GENERATION AND STATISTICAL MODELING OF ACTIVE PROTEIN CHIMERAS: A SEQUENCE BASED APPROACH

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Nicholas Justin Fico

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2013

Purdue University

West Lafayette, Indiana

For my wife Mary and my parents Ron and Fran.

ACKNOWLEDGEMENTS

I would like to thank the Purdue Genomics Core for years of assistance in obtaining high quality sequencing data, without which this work would not be possible. In particular, I would like to thank Allison Sorg for her assistance and advice regarding efficient plate layouts, primer design and sequencing reactions. I would also like to thank Phillip San Miguel for his work aligning countless sequencing runs into meaningful contigs, and his efforts to allow our lab easy access to the data on the genomics core computer servers.

I would also like to thank Bruce Cooper and Amber Jannasch at the Purdue Metabolite Profiling Facility for advice and time spent preparing and maintaining critical LC/MS equipment, as well as many hours spent helping us identify particularly difficult metabolites in some very noisy samples.

All of the people in Clint Chapple's lab including Jo Cusumano, Nicholas Bonawitz, Shiva Hemmati, Yi Li, and Whitney Dolan for getting me started with many protocols, materials and practical advice for working with yeast. Thomas Sors for initating the AtC4H-SmC4H project and demonstrating its feasibility.

I would like to thank the members of my committee: for supporting me throughout my graduate career, and especially thank my advisor Alan Friedman

for endowing my PhD education with great breadth, as well as great depth.

All members, past and present, in Alan Friedman's lab have undoubtedly contributed to my success. In particular, I would like to thank Corinne Price and Samuel Schaffter for their experimental contributions for this work. And Patrick Dolan for invaluable perspective.

The Purdue Outing Club has kept me in good spirits and helped me maintain an even keel throughout the sometimes stressful work of graduate school.

Finally, I would like to thank my parents Ron and Fran for raising me and my wife Mary for supporting me and giving me guidance. Without you, this would not have been possible.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Figure Page

LIST OF ABBREVIATIONS

MSA	multiple sequence alignment
C4H	cinnamate 4-hydroxylase
AtC4H	*A. thaliana* cinnamate 4-hydroxylase
SmC4H	*S. moellendorffii* cinnamate 4-hydroxylase
CA	cinnamate
pCA	*para*-coumaric acid
SRS	substrate recognition sequence
P450	cytochrome P450-dependent monooxygenase
AtC4H N/C	AtC4H gene with 12 N-terminal and 5 C-terminal codons optimized for expression in *S. cerevisiae* (sp)
SmC4H N/C	SmC4H gene with 12 N-terminal and 5 C-terminal codons optimized for expression in *S. cerevisiae* (sp)
F5H	freulate 5-hydroxylase
AtF5H	*A. thaliana* ferulate 5-hydroxylase
SmF5H	*S. moellendorffii* ferulate 5-hydroxylase

ABSTRACT

Fico, Nicholas J. Ph.D., Purdue University, December 2013. Generation and Statistical Modeling of Active Protein Chimeras in the Absence of Structural Information. Major Professor: Alan Friedman.

Generation of active protein chimeras is a valuable tool to probe the functional space of proteins. Statistical modeling is the next logical step, allowing us to build a model of gene fragment replaceability between species. In this thesis I begin to develop the statistical tools that are needed to systematically describe combinatorial protein libraries. I present three sets of diverse chimeric protein libraries developed using sequence information. The statistical model of the human N-Ras and human K-Ras-4B genes reveal a set previously unidetifed surface residues on the N-Ras G-Domain that may be involved in cellular localization. Statistical modeling of a library of chimeric proteins between *A. thaliana* cinnamate 4-hydroxylase (AtC4H) and *S. moellendorffii* cinnamate 4-hydroxylase (SmC4H) reveal a possible stabilizing effect of the N-terminal amino acids from SmC4H and irreplaceable catalytic domains between AtC4H and SmC4H. I also show gene fragment replaceability on a small scale between functionally divergent AtC4H and *A. thaliana* ferulate 5-hyrdoxylase proteins. Finally, I show that commonly occurring residue pairs in the sequence record are effective covariates when modeling activity in the AtC4H-SmC4H chimeric library.

CHAPTER 1. INTRODUCTION

Protein engineering has enabled the drastic alteration of protein function, and unlocking of novel protein functions, in single step and iterative experiments. Directed evolution and informed residue modification are two established methods of protein engineering. Each method assumes that the desired function lies within the accessible sequence space from the starting protein sequence, and that new function can be selected by incremental alteration of amino acid residues (1,2,3,4). In contrast, shuffling gene fragments 10-100 amino acids long of related extant sequences introduces dozens of variant amino acid residues, enabling rapid exploration of a much greater diversity of sequence space (5).

Here, we present three protein libraries with active members. The first protein library consists of 16 novel proteins that are crosses of human N-Ras and human K-Ras-4B. All 16 chimeras are localized by COS-7 cells in a manner similar to

Directed evolution introduces one to a few point mutations in a protein, resulting in often modest changes to activity. When the process is iterated, a dozen or more beneficial mutations can be identified, resulting in protein that is significantly more active, more thermostable or more resistant to degradation,

depending on the selection mechanism. Since the mutations are introduced at random, investigators do not need to know or understand the underlying mechanisms driving protein function in the target protein (6).

Informed residue modification requires a detailed model of the protein and intimate knowledge of the residues contributing to functionality. Often, thermodynamic models are built on extensive structural models incorporating substrate binding (4). This level of knowledge is only available for the best studied protein systems, and involves modifying an already existing function or attribute of the protein.

To access novel protein functions, or create substantial changes in protein attributes in a single step, it becomes necessary to introduce large numbers of point mutations into the parental sequence simultaneously (5). The challenge to this approach is that introducing multiple divergent point mutations in a single protein is rarely beneficial, or even neutral. Homologous proteins provide a pool of sequences that exist in the same functional space (7,8). It follows that exchange of amino acids residues between two homologous sequences might be less detrimental than introduction of completely random mutations. Many methods exist that recombine homologous sequences in a stochastic fashion. Indeed, highly functional variants have been identified by these methods. However, most rely on powerful selection mechanisms to identify the few active recombinants out of a large pool of inactive recombinant proteins.

When constructing and testing a large number of novel proteins, experimental efficiency becomes a concern. Functional screening becomes much more efficient when the majority of targets are successes, rather than failures. Also, a recombinant library containing a high proportion of active proteins need not rely on any functional selection; the entire library can be screened and characterized with little wasted effort. This eliminates sequence bias that is present when testing a random sampling of a chimeric library.

Site-directed, homologous recombination guided by structure-based computation (SCHEMA) is the current protocol for creation of recombinant protein libraries enriched for active members. This protocol works by analyzing a protein structure and dividing the protein into fragments with minimal inter-fragment residue contacts (9). The idea is to maximize fragment structural independence, and thus interchangeability. This procedure has been successfully used to rapidly diversity the activity levels, themostability and functional space of target proteins (10,11).

However, structural information is not always available for a target protein system. Fortunately, this need not be an impediment to successful design of recombinant protein libraries. Evolution preserves fold, functional motifs, sequence, and catalytic activity of homologous proteins. Importantly, these homologous structures are readily identified in protein multiple sequence alignments (12,13).

In this thesis, I demonstrate that multiple sequence alignments of homologous proteins contain sufficient information to identify interchangeable gene fragments. The results are recombinant protein libraries with a high proportion of active members.

First, I present a small library of protein chimeras formed by recombining the hypervariable region between human N-Ras and human K-Ras-4B. The hypervariable region is unstructured; only a multiple sequence alignment (MSA) and previous deletion studies can be used as a guide to identify functionally equivalent regions. The resultant chimeric Ras proteins were constructed as N-terminal GFP fusion. Cellular localization of each chimera was visualized in COS-7 cells, with the chimeras displaying a mixture of K-Ras like localization to the inner plasma membrane and N-Ras like localization in the Golgi. This demonstrates that is possible to create functionally relevant protein chimeras where no structural information exists. I also present a statistical model of the Ras proteins, which explains localization in a gene-fragment dependent manner. The secondary cystine in N-Ras and the polybasic region in K-Ras are confirmed as necessary sequence motifs for N-Ras and K-Ras like localization, respectively. It is also shown that the G-Domain of the N-Ras protein contains a localization signal to the Golgi. This Golgi localization signal most likely involved the surface residues 91-95 of N-Ras (ADINL).

The second set of chimeric protein libraries I present has been constructed from phenypropanoid P450s. Here, we are recombining gene fragments that code for defined secondary and tertiary structure units. Similar to the Ras library, interchangeable gene fragments have been determined from MSA.

Again, we show that the information present in an MSA is sufficient to identify functionally interchangeable gene fragments without incorporating explicit structural information. I also present a set of functionally relevant statistical models that reveal interacting structure-function-sequences constraints between regions of the cinnamate 4-hydroxylase (C4H) protein. In particular, the N-terminal 90 amino acids for *S. moellendorffii* cinnamate 4-hydroxylase (SmC4H) stabilize C4H proteins to a greater degree than the N-terminal 91 amino acids of *A. thaliana* cinnamate 4-hydroxylase (AtC4H) protein. Structure-function-sequence constraints have diverged between these proteins around the catalytic site and heme domain.

Construction of a chimeric library between AtC4H and *A. thaliana* ferulic acid 5-hydroxylase (AtF5H) resulted in only one functionally active chimera indicating that structure-function-sequence constraints between AtC4H and AtF5H have diverged.

CHAPTER 2. PLASMA MEMBRANE LOCALIZATION OF CHIMERIC RAS PROTEINS

2.1 Introduction

K-Ras-4B and N-Ras are GTPase isoforms that signal cell proliferation as part of the Ras-Raf-MEK-ERK pathway. Ras family proteins are activated by a guanine nucleotide-exchange factor in response to an extracellular signal (growth factor). Activated Ras proteins have many immediate targets including PIK3CA, B-Raf and Raf-1. Ultimately, the initial extracellular signal is propagated into the nucleus resulting in transcription and translation of dozens of genes involved in progression of the cell cycle and apoptosis (14).

Ras Family proteins are oncogenes whose mutation and overexpression is associated with many human cancers (15). Cellular localization is essential to their function and the biological processes underlying their proper cellular localization has become a target for cancer therapies (16,17).

In activated cells, all Ras proteins are eventually localized to the inner plasma membrane, but arrive through different pathways (18,19). Deletion studies combined with residue substitutions have established minimal sequences necessary and sufficient for both N-Ras and K-Ras localization (Figure 1) (19,20).

Targeting of Ras proteins to the different pathways is largely determined by the 25 C-terminal amino acids (hypervariable region) of the proteins, with recent evidence also showing a contribution by the N-terminal catalytic domain (G-Domain) of N-Ras (18,21,19,22).

Common to all Ras proteins is a C-terminal CAAX motif. The CAAX motif is the four C-terminal amino acids of the protein consisting of a cysteine and two aliphatic amino acids followed by any amino acid. After protein translation, the cysteine is farnesylated or geranylgeranylated. The modified protein is then targeted to the surface of the endoplasmic reticulum where the three terminal amino acids, AAX, are cleaved (23,24). At the endoplasmic reticulum, post translational modification and cellular targeting diverge for N-Ras and K-Ras-4B. K-Ras-4B is localized through cytoskeletal transport and associates with the inner plasma membrane by a series of basic residues in the hypervariable region (22,25). In N-Ras, a secondary cysteine present in the N-Ras hypervariable region is palmitoylated in the endoplasmic reticulum. In activated cells, N-Ras is then localized to the inner plasma membrane by vesicular transport through the Golgi (18). In this study non-activated COS-7 cells are used; N-Ras will remain localized in the endoplasmic reticulum and Golgi (perinuclear) whereas K-Ras-4B will be transported to the inner plasma membrane (PM). This divergence in cellular localization is easily visualized using N-terminal GFP fusions (Figure 4).

Although informative, previous studies have not explored interactions among residues within the hypervariable region. Here, we present a set of N-Ras and K-Ras-4B chimeras designed to test localization signals within the context of a complete Ras protein. This strategy provides three advantages over traditional methods. First, localization signals are tested in the context of a complete Ras protein. An alanine scan may indicate whether a particular residue, or series of residues, is necessary for protein function. Exchanging gene fragments between known functional sequences tests subtle differences between these sequences. Second, swapping regions allows us to test interaction of localization signals between different gene fragments. Finally, a complete set of chimeras allows us to build a statistical model describing Ras localization based on identity of the gene fragments.

2.2 Results and Discussion

2.2.1 Selection of Gene Fragments

To test localization signals in the hypervariable region, Ras protein sequences were aligned and divided into four fragments. Alignment of the G-Domain can be considered reliable because of high sequence identity and homologous structures. Alignment of the hypervariable region is less certain due to low sequence identity in these unstructured regions. To address this, multiple alignments were evaluated and compared to develop a consensus (Figure 1).

Gene fragments for chimeragenesis were selected for their potential for contribution to membrane localization and to separate probable independent membrane localization signals. In defining breakpoints for the hypervariable region, only functional, and not structural, equivalence between N-Ras and K-Ras-4B was considered, as the regions lack defined structure.

The first fragment is the G-domain. It has been previously reported that the N-Ras G-Domain affects localization in activated HeLa cells (21). However, the main function of the G-Domain is cell signaling, and here we chose to separate any possible localization effects of the G-Domain from the hypervariable region by making it the first gene fragment.

The second through fourth gene fragments span the hypervariable region, which has been shown to contain the majority of localization signals in Ras isoforms

(20,22). The second gene fragment consists of part of the putative linker domain (Figure 1). In N-Ras, a motif in the linker domain was identified as contributing to the stability of N-Ras on the inner plasma membrane of activated HeLa cells (21). Despite strong conservation in the K-Ras-4B linker domain across species, and great divergence (presumably selected for) between the linker domains of N-Ras and K-Ras-4B isoforms, there is presently no known function for the linker domain in K-Ras-4B.

The third gene fragment contains a small remainder of the linker region along with the secondary cysteine in N-Ras and the polybasic region in K-Ras-4B, which have been classically determined to be responsible for N-Ras and K-Ras-4B localization, respectively (Figure 1) (22,21). Swapping this gene fragment tests the classic interpretation that these residues exclusively determine localization, while also testing the interaction of these localization signals with other parts of the protein.

The fourth fragment contains the two divergent amino acids prior to the CAAX motif, and the CAAX motif itself. We do not expect swapping the two (slightly divergent) CAAX motifs to affect localization, but swapping the earlier amino acids might affect localization.

The nomenclature used to refer to chimeric sequences is as follows. Each chimera is defined by a four letter code, one letter for each gene fragment. The

letters are either N, indicating that a particular gene fragment is from human N-Ras, or K, indicating that a particular gene fragment is from K-Ras-4B. For example, the reconstructed human K-Ras-4B gene is coded as KKKK, and the reconstructed human N-Ras gene is coded as NNNN. The chimera KNNN is the K-Ras-4B G domain (gene fragment 1) joined with the N-Ras hypervariable region (gene fragments 1, 2 and 3).

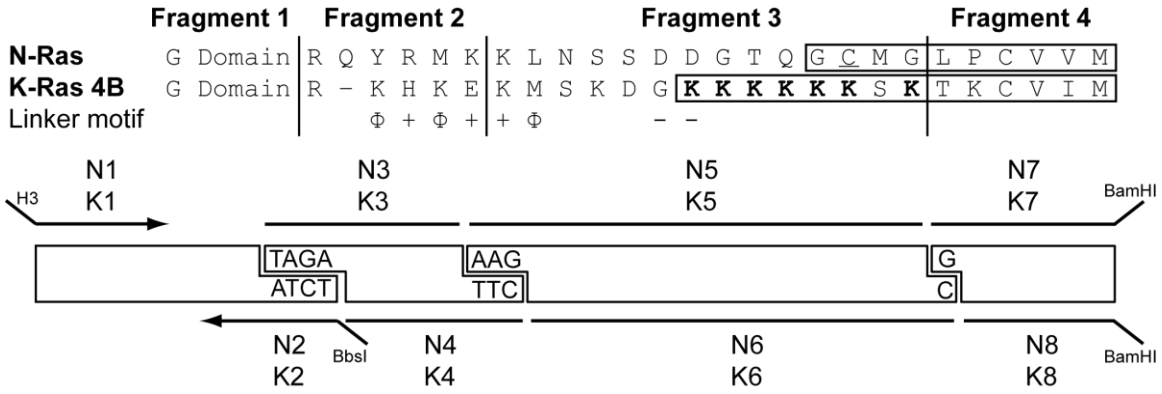


Figure 1 Gene sequences and Fragments Used to Construct Ras Chimeras. The classic primary and secondary localization signals are boxed. Multiple alignments of the Ras hypervariable region were performed and this consensus was developed. The linker motif found in palmitoylated Ras isoforms is also shown. Φ and + symbols represent the aliphatic (Φ) and positively charged (+) residues that comprise the linker motif (21).

.

2.2.2 Construction of Ras Chimeras Using Selective Overhangs

The four gene fragments were assembled into 16 complete chimeric genes by our SPLISO method (26) in independent parallel reactions (Figure 1). Ras constructs were fused to EGFP in pEGFP-C3 plasmids and transformed into DH5α. For each chimera, three to six colonies were selected for sequencing. A total of 51 genes were sequenced, yielding 40 correct chimeras; a 78% success rate.

Of the 21 incorrect chimeras sequenced, six incorrect chimeras shared an identical point mutation in fragment 2. Surprisingly, these belong to different chimeras and only share a common synthetic fragment, suggesting an error in oligo synthesis. Twelve incorrect chimeras contained fragment 4 directly ligated to fragment 1 with a variety of junctions. This may be due to fragment 4 containing just a single nucleotide overhang, which is known to ligate inefficiently. The remaining four incorrect chimeras contain unrelated point mutations or single or double nucleotide deletions. These are likely background errors from PCR. We expect future library construction using these methods to be much more efficient. Single base pair overhangs could be avoided and errors during oligo synthesis should be encountered rarely. Although PfuTurbo is a proof reading polymerase, higher fidelity polymerases are now available (e.g. Phusion) which may reduce PCR errors. Taken together, we expect future chimeric SPLISO libraries to be significantly improved in the proportion of correct chimeras.

2.2.3 Steady State Localization of EGFP- Ras Chimeras

EGFP-Ras Chimera fusions were tested for steady-state localization in COS-7 cells (Figure 2, Figure 4). N-terminal GFP-Ras fusions are an established method for investigation of localization of Ras variants (18,21,19) Individual COS-7 cells displayed three general patterns: Predominantly plasma membrane, predominantly perinuclear (reflecting Golgi/ER association) and mixed localization. These observations agree with earlier studies on Ras localization in COS-7 cells (27).

Localization of the set of chimeras displays a strong binary distribution. Chimeras are either strongly K-Ras-4B-like (>95% average plasma membrane localization) or N-Ras-like (< 40% average plasma membrane localization). Only chimera KNNK shows an intermediate distribution (68% average plasma membrane localization). (Figure 2)

Still, some chimeras show greater than expected variance within a group (Figure 5). These higher variances may be due to unknown variability in the COS-7 cell cultures potentially combined with difficulty in cellular processing of a chimeric localization signal. Determining the origins of the variability and any alteration in the localization pathways taken by the proteins await further studies. Indeed, these chimeras may prove to be very useful probes for testing detailed mechanisms for localization.

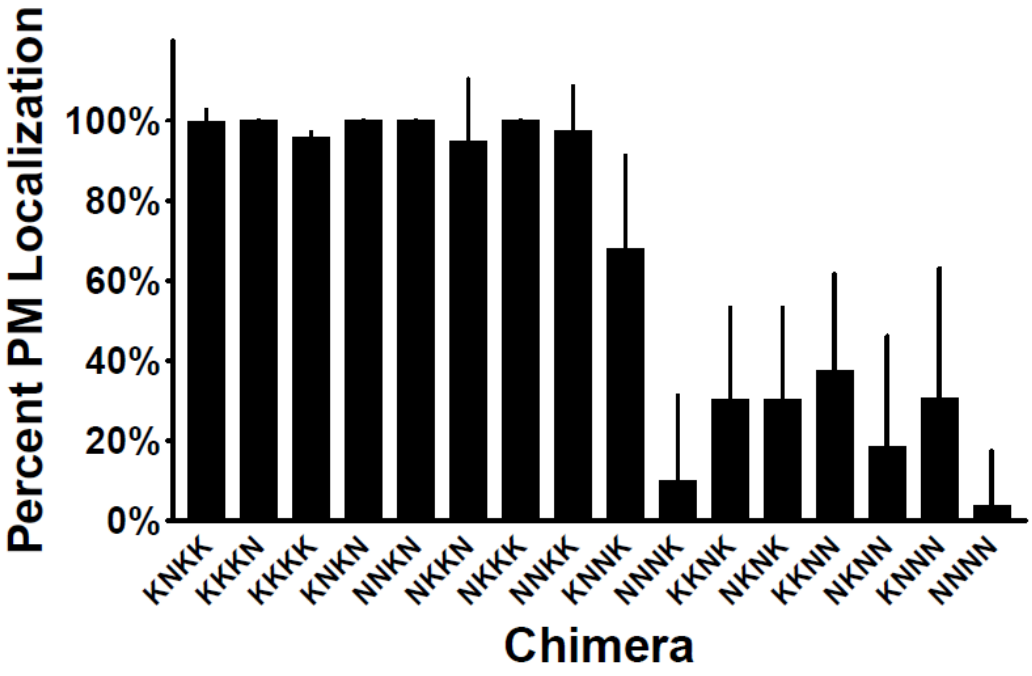


Figure 2 Plasma membrane localization by chimera.

2.2.4 Statistical Modeling of Ras Localization by Gene Fragment

Our primary goal is to create a statistical model of Ras localization. Gene fragments found to be significantly associated with differences in cellular localization will indicate a divergence of localization signals between N-Ras and K-Ras-4B. Once functional divergence has been narrowed down to one, or a few, gene fragments, careful comparison of the sequences should identify a limited set of amino acid residues responsible for the different cellular localization signals.

Only cells displaying strong plasma membrane localization or strong perinuclear localization were considered. Cells showing mixed localization of Ras isoforms did not appear to clearly contribute to the localization model and were not included in cell counts for the final statistical model. Investigation of cells displaying mixed localization patterns may require special consideration in future studies.

Since the localization data is binomial, a logistic model (Table 1A) is preferred in fitting the data using the four gene fragments and their interactions as explanatory variables (Table 1C,section 2.3.3). The data was also fit to a traditional linear ANOVA model (Table 1A and Table 1B), because this model is easier to interpret and more widely understood. In development of a model of each type, we tested all the individual and pairwise interactions, eliminating the insignificant interactions from the final models (Table 1B and Table 1C). A

comparison of the model sum of squares shows that both models explain the data equally well (Table 1A). While the logistic model is a better theoretical representation of the data, the fragment effects are strong enough to be significant in a traditional linear ANOVA model.

As expected, fragment 3 shows the greatest significance in determining plasma membrane localization (Table 1B and Table 1C). This agrees with extensive previous research using point mutants and Ras gene fragments to localize GFP fusions (22,18,28). An important secondary role for fragment 1 (the G-Domain) and the interaction between fragment 1 and 3 is also significant. (Figure 3 and Table 1B and Table 1C). The model parameter for the fragment 1-3 interaction effect is significant in the linear ANOVA model, but is subsumed into the logit link function of the logistic model and thus does not require a separate predictor. ($p=0.66$, not shown.) (Figure 3 and Table 1C). Neither fragment 2, fragment 4, nor any of their interactions, have any significant effect on localization.

Based on the linker motif identified by Laude 2008 (21), we might expect fragment 2 to also affect the localization of Ras proteins, however we find that neither fragment 2 (which contains part of the linker motif) nor the interaction between fragments 2 and 3 (which contains the complete linker motif) play a significant role. Possible explanations include: Laude used gene fragments to explore this effect, whereas the chimeras in the current study explore localization of a whole protein; differences between the boundaries of our fragments and

their domains; and differences between HeLa cells used by Laude, 2008 and COS-7 cells in this study to promote plasma membrane localization in N-Ras.

Fragment 4 contains two divergent amino acids followed by the CAAX motif: LPCVVM for N-Ras and TKCVIM for K-Ras-4B. Since the identity of fragment 4 does not affect localization, we conclude that the first two amino acid residues in fragment 4 (LP and TK) are interchangeable between N-Ras and K-Ras-4B.

The model also reveals that the G-Domain (fragment 1) has a significant effect on localization. Chimeras containing N-Ras Fragment 3 display more N-Ras like localization (e.g. lower percentage of plasma membrane localization) when combined with the N-Ras G-Domain. (Compare especially KNNK with NNNK and KNNN with NNNN in Figure 2). This result corroborates the observations of Laude, 2008 (21) who noted that the complete N-Ras protein displayed weaker plasma membrane localization than just the N-Ras hypervariable region. This work extends that result and shows the importance of testing protein chimeras, since we make the complementary observation that chimeras carrying the K-Ras-4B G-Domain and N-Ras fragment 3 have stronger plasma membrane localization compared to chimeras carrying the N-Ras G-Domain and N-Ras fragment 3.

However, we do not observe the identity of the G-Domain having any effect on chimeras containing K-Ras-4B fragment 3. Chimeras containing K-Ras-4B

fragment 3 localize to the plasma membrane at or near 100% of the time, regardless of the identity of their G-Domain. One possibility is that the localization signals contained in K-Ras-4B fragment 3 coupled with a CAAX motif are so strong that they overwhelm any alternate localization signals present in the G-Domain. Another explanation may be that the cytoskeletal transport pathway responsible for localizing K-Ras-4B simply does not perceive the Ras G-Domain. Alternatively, the Ras G-Domain may affect the kinetics (but not steady state) of K-Ras-4B-like localization, which have not been measured in the present study.

In determining the specific residues that are responsible for the effects of the Ras G-Domain on localization, we note that the K-Ras-4B and N-Ras G-Domains differ by only nine amino acids. There are six conservative substitutions on the protein surface: S87T, T122S, H131Q, E132D, K135R and E152D (listed N-Ras to K-Ras). In addition, three non-conservative amino acid substitutions on the protein surface between positions 91-95, ADINL (N-Ras) and EDIHH (K-Ras-4B) are of particular interest. Noting the difference in charge between these surface proteins, we propose that A91, N94 and L95 are largely responsible for the localization contributions of the N-Ras G-Domain. Future chimeras between the two catalytic domains will help reveal both any independent role of these polymorphisms, and any interactions with the hypervariable region, in determining membrane localization of Ras proteins.

Table 1 Linear ANOVA and logistic models of plasma membrane localization.

A

Comparison of Linear ANOVA and Logistic Regression of Ras Chimeras

Regression Model	Model Sum of Squares	Error Sum of Squares	Total Sum of Squares	Percent explained
Linear ANOVA	46921.74	13531.63	60453.38	77.6
Logistic model	45472.37	14981.01	60453.38	75.2

B

Linear ANOVA Model of Ras Chimeras

Model Significance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	46921.74	9384.35	24.97	<.0001
Error	36	13531.63	375.88		
Corrected Total	41	60453.38			

Estimate and Significance of Parameters

Parameter	Value	Estimate	Std. Error	T Value	Pr > |t|
Intercept		42.69	11.62	3.68	0.0008
Frag1	K	0.04	9.23	0.01	0.9960
Frag1	N	0.00			

Table 1 Continued

Frag3	K	53.19	8.05	6.61	<0.0001
Frag3	N	0.00			
Frag1*Frag3	N,N	-26.29	12.23	-2.15	0.0383
Frag1*Frag3	K,N	0.00			
Frag1*Frag3	K,K	0.00			
Frag1*Frag3	N,K	0.00			
Frag2	K	-1.97	6.06	-0.32	0.7474
Frag2	N	0.00			
Frag4	K	6.76	6.09	1.11	0.2745
Frag4	N	0.00			

C

Logistic Model of Ras Chimeras

Model Significance

Model	DF	-2 Log Pseudo-Likelihood	Chi-Squared	P > x
$Y=\mu_{..}+f_1+f_2+f_3+f_4$	5	171.67	50.38	<0.0001
Y=1	42	121.29		

Table 1 Continued

Estimate and Significance of Parameters

Effect	Value	Estimate	Std. Error	DF	t value	Pr > |t|
Intercept		-1.85	0.54	31.98	-3.45	0.0016
Frag1	K	1.54	0.45	30.66	3.46	0.0016
Frag1	N	0.00				
Frag3	K	6.71	0.43	11.38	15.67	<0.0001
Frag3	N	0.00				
Frag2	K	0.39	0.47	31.18	0.86	0.3980
Frag2	N	0.00				
Frag4	K	-0.30	0.51	34.70	-0.59	0.56
Frag4	N	0.00				

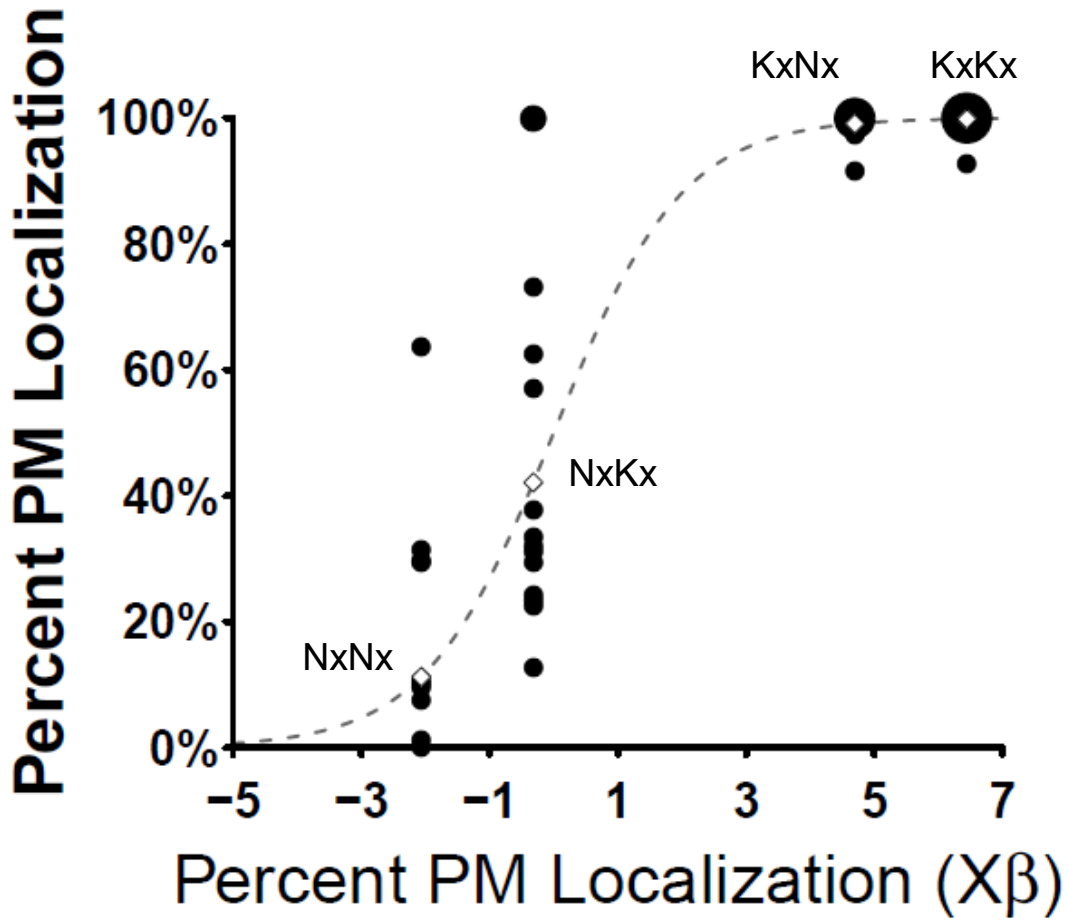


Figure 3 Observed versus predicted localization..
The sum of the predictors times their fitted parameters is plotted on the x-axis and is used to predict plasma membrane localization on the y-axis. The area of each black circle represents the relative number of observations with given percent plasma membrane localization. White diamonds represent expected plasma membrane localization for each group. The Grey curve is the logit link function. The 100% plasma membrane localization point in KxNx and the 60% plasma membrane localization in NxNx are possible experimental outliers. However, retaining or removing these data points does not alter the conclusions of the model. Therefore, they have been conservatively retained.

2.3 Materials and Methods

2.3.1 Construction of Ras Chimeras

A schematic for assembling the gene fragments including the selective overhangs is shown below the alignment in Figure 1. For this study, four N-Ras and K-Ras-4B gene fragments have been recombined, generating 16 chimeric genes. Ras chimeras were constructed by assembling DNA fragments as shown: N-Ras Fragment 1 and K-Ras Fragment 1 were PCR amplified with primers N1 and N2 or K1 and K2, respectively, and PCR products were digested with the typeIIS restriction endonuclease BbsI. All other fragments were formed by annealing synthetic the oligonucleotides in Figure 1(N3-8, K3-8) and listed in Table 2. Each chimera was constructed in a separate ligation reaction; the gene fragments for each chimera was mixed in equimolar amounts and ligated with T4 DNA ligase. Correct, simultaneous assembly of all four fragments was possible by employing unique selective overhangs. Full length ligation products were enriched by PCR and each chimera was ligated into pEGFP-C3 (Gift from Dr. Michael Phillips) using BamHI and HindIII to create a fusion with EGFP at the N-terminus. Surprisingly, direct cloning of chimera NKNN into pEGFP-C3 was not successful. Instead, NKNN was cloned into pET-30b in reverse orientation, and then transferred to pEGFP-C3. Inserts were confirmed by sequencing both strands using the pEGFP C-term primer and EBV rev primers.

Table 2 Oligonucleotides used to assemble and sequence Ras chimeras (See Figure 1).

Oligo Name	Sequence
K1	5’-GCCGCCAAGCTTATGACTGAATATAAACTTGTGG-3’
K2	5’-GGCGGCGAAGACAATCTAATTTCTCGAACTAATGTATAG-3’
K3	5’-pTAGAAAACATAAAGAA-3’
K4	5'-pCTTTTCTTTCTGTTT-3'
K5	5’-pAAGATGAGCAAAGATGGTAAAAAGAAGAAAAAGAAGTCAAA-3’
K6	5'-pCTTTGACTTCTTTTTCTTCTTTTTACCATCTTTGCTCAT-3'
K7	5’-pGACAAAGTGTGTAATTATGTAATAAGGATCCGCCGCC-3’
K8	5'-GGCGGCGGATCCTTATTACATAATTACACACTTTGT-3'
N1	5’-CGCGCGAAGCTTATGACTGAGTATAAACTGGTGGTG-3’

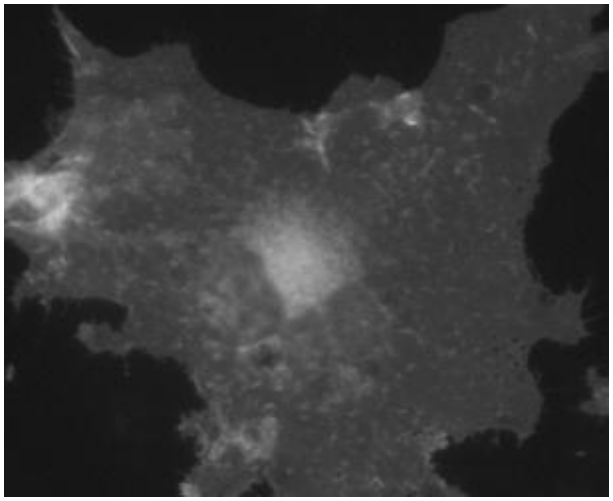
Table 2 Continued

Oligo Name	Sequence
N2	5’-GGCGGCGAAGACAATCTAATTTCTCTTACCAGTGTGTAAAAAGC-3’
N3	5’-pTAGACAGTACCGAATGAAA-3’
N4	5'-pCTTTTTCATTCGGTAGTC-3'
N5	5’-pAAGCTCAACAGCAGTGATGATGGGACTCAGGGTTGTATGGG-3’
N6	5'-pCCCCATACAACCCTGAGTCCCATCATCACTGCTGTTGAG-3'
N7	5’-pGTTGCCATGTGTGGTGATGTAATAAGGATCCGCCGCC-3’
N8	5'-GGCGGCGGATCCGTCTTACATCACCACACATGGCAA-3'
pEGFP C-term primer	5'-CATGGTCCTGCTGGAGTTCGT G-3'
EBV rev primer	5'-GATGAGTTTGGACAAACCCA-3'

2.3.2 Steady State Localization of Ras Chimeras

Two samples of each EGFP-chimera fusion plasmid were coded and assayed for localization in a single blind experiment. Chimeras displaying either high variance or unexpected localization were given new codes and again assayed in a single blind experiment. Chimeras KNKK, KKNK, KKNN, NKNK and KNNK were assayed four times. The remaining eleven chimeras were assayed twice. COS-7 cells were seeded onto coverslips in 6-well plates and grown to 80% confluency. Cells were transfected with 1 μg of pEGFP-C3 plasmid DNA bearing an EGFP-Ras chimera using Lipofectamine 2000 and allowed to recover overnight. Cells were harvested 24 hours post transfection for fluorescence microscopy analysis. Cells were fixed in 3.7% formaldehyde solution and washed three times with phosphate buffered saline. Imaging was performed on Olympus BH-2RFCA equipped with a Sony DXC-950 3CCD color video camera, using the 60x objective. For each chimera, at least 100 cells were manually scored for either plasma membrane or perinuclear localization (Figure 4). Cellular localization work was performed by Su-Sien Ong.

A



B

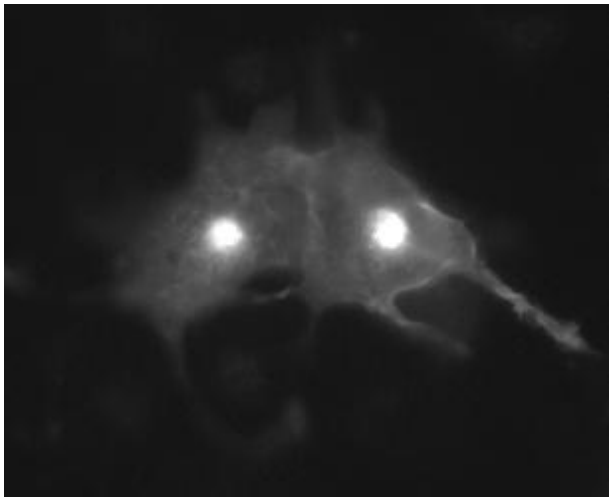


Figure 4 Cellular localization of EGFP-Ras chimera fusions expressed in COS-7 cells viewed under fluorescence microscopy.
A-EGFP-KKKK showing plasma membrane localization. B-EGFP-NNNN showing perinuclear localization.

2.3.3 Logistic Regression of Localization Data

Localization of Ras chimeras was modeled in SAS 9.2 using a logistic generalized linear model with random effects in three groups for chimeras of high, medium or low variance and fit using the Trust Region algorithm (29)
Count data has many features in common with a logistic predictor including bounded response and a correlation between predicted value and observed error. This correlation can be seen empirically (Figure 5). These criteria suggest that logistic model is a more appropriate choice, as explained more fully below. A traditional linear ANOVA model has also been presented for clarity (Table 1).

A logistic model is preferred because it is a more accurate representation of the data than a traditional linear model. In a traditional linear model, $Y=X\beta+\varepsilon$, the response term Y is unbound and can take on any value between $-\infty$ and ∞. This is undesirable as binomial data is bound between 0 and 1. The logistic model solves this problem by fitting to the logit function: $Y=e^{X\beta+\varepsilon} / 1+e^{X\beta+\varepsilon}$, Where $Y = n_{pm} / (n_{pm} + n_{pn})$, where n_{pm} is the number of observed COS-7 cells displaying plasma membrane localization for a given chimera and n_{pn} is the number of observed COS-7 cells displaying perinuclear localization for a given chimera. (The number of COS-7 cells showing a mixed localization pattern was not reflective of fragment composition of intact Ras chimeras). X is the matrix of predictors, β is the matrix of fitted model parameters, and ε is the error term for each observation. Here, as $X\beta+\varepsilon$ approaches $-\infty$, Y approaches 0 and as $X\beta+\varepsilon$

approaches +∞, Y approaches 1. Thus predictions are bound to the experimental observable values between 0 and 1.

Logistic models also include a more appropriate error term for binomial data. While a traditional linear model applies the same error term for all observations, the error term in binomial data is expected to vary based on the predicted value, as described by V[X]=p(p-1). A traditional linear model will thus overestimate the error for values with very high or very low probability, potentially leading to type II errors (false positive) for these observations. Similarly, a traditional linear model will underestimate the variance for values with moderate probability and may lead to type I error (false negative) for these observations.

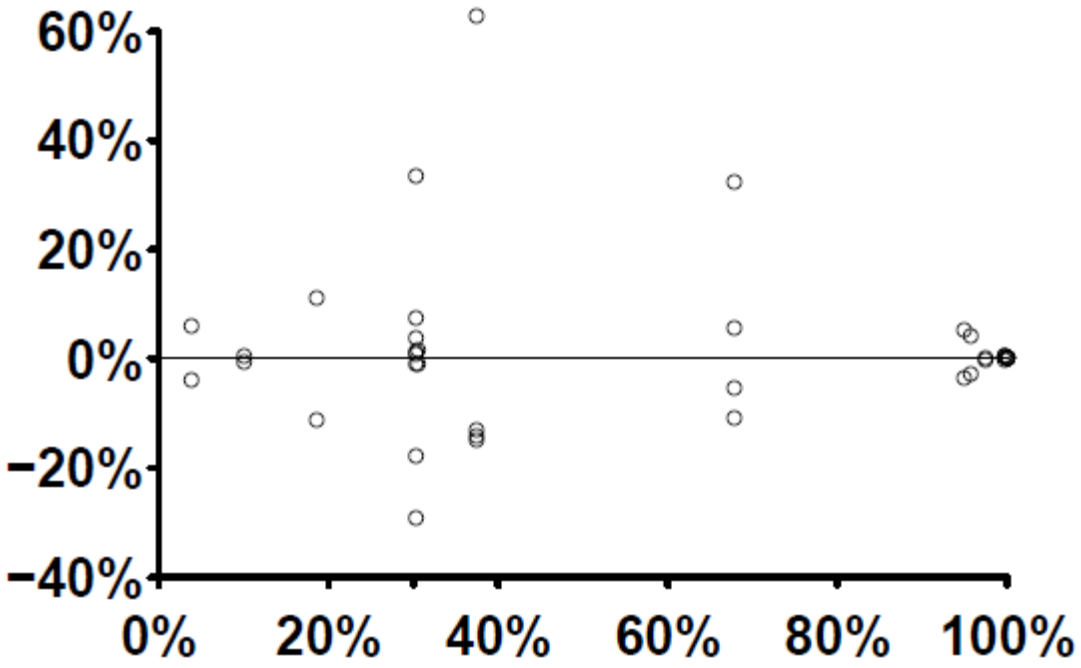


Figure 5 Residual plasma membrane localization vs. average plasma membrane localization by chimera. It can be seen that chimeras of either high or low average plasma membrane localization have very low variance, whereas chimeras of intermediate plasma membrane localization have high variance as predicted by V[x]=p(1-p).

CHAPTER 3. GENE FRAGMENT INTERCHANGEABILITY BETEWEEN CINNAMATE 4-HYDROXYLASE PROTEINS FROM *A. THALIANA* AND *S. MOELLENDORFFII*

3.1 Introduction

Cinnamate 4-hydroxylase (C4H) has been used for the synthesis of medically important compounds in both synthetic gene pathways in the lab (30,31), and in naturally occurring secondary metabolite pathways in plants (32,33). C4H is also an essential enzyme for lignin production in vascular plants (34).

Lignin provides vascular plants the structural support to grow tall and the vascular rigidity needed for water transport. Due to its insolubility and heterogeneous structure, lignin is difficult to degrade and inhibits conversion of cellulose to bioethanol (35,36), as well as inhibiting conversion of feedstock into energy for ruminants (37). For these reasons, C4H's role in lignin production has been directly investigated with potential economic impact for biofuel production (38,39,35), agricultural feedstock (40), and material properties of wood (41,42).

A better understanding of how the C4H gene itself can be altered will greatly aid synthetic biology efforts, either by increasing expression and activity levels, or

altering substrate specificity. Previous work has tested the limits of substrate recognition in *Arabidopsis thaliana* C4H (AtC4H) (43,44,45), as well as tested catalytically relevant amino acid residues with point mutations (46).

Currently, no published work has tested the sensitivity of large regions of the C4H protein to multiple amino acid substitutions. By exchanging large gene fragments between theC4H gene from two diverse species, we will gain insight into which regions of the protein have been conserved through evolution (functionally interchangeable gene fragments) and which regions of the protein have diverged through evolution (functionally non-interchangeable gene fragments). By identifying which regions of the C4H gene are sensitive to substitutions, we hope to inform where future protein engineering efforts should focus their attention.

Here, we present a library of chimeric C4H proteins containing gene fragments from *Arabidopsis thaliana* C4H (AtC4H) and *Selaginella moellendorffii* C4H (SmC4H). We find that the N-terminal region from SmC4H (but not AtC4H) appears to stabilize protein chimeras, while the catalytic region of AtC4H and SmC4H have diverged structurally, limiting gene recombination in this region. Overall, we find that large gene fragments (up to 161 amino acids, containing 58 residue polymorphisms) are generally interchangeable between AtC4H and SmC4H proteins and a model of the presence of gene fragments and their interactions that mediate activity levels has been developed.

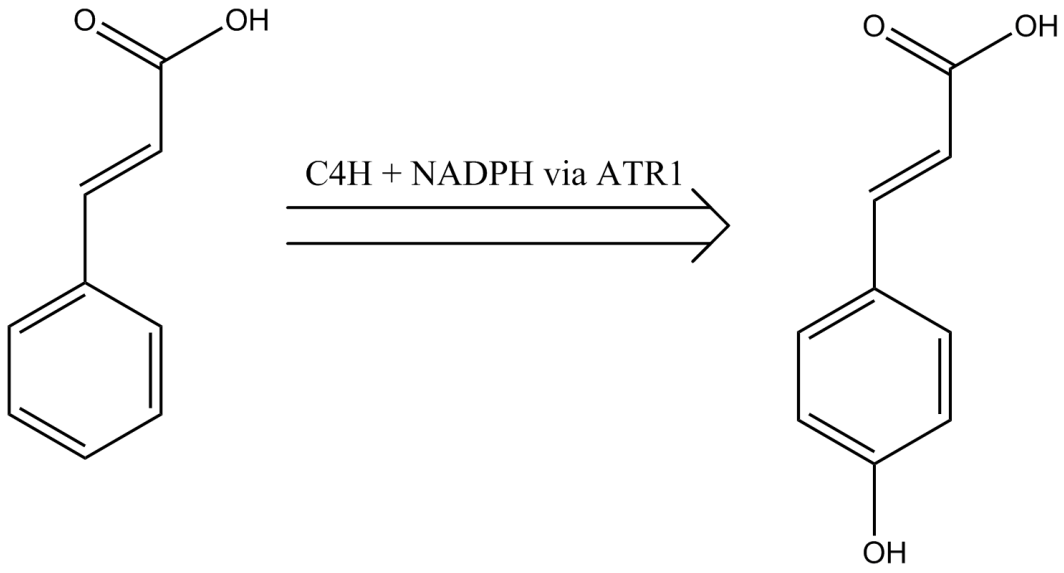


Figure 6 Enzymatic reaction of C4H

3.1.1 Selection of AtC4H and SmC4H as Parental Genes

Cinnamate 4-hydoxylase (C4H, EC 1.14.13.11) is a membrane bound cytochrome P450-dependent monooxygenase (P450) that belongs to the CYP73 protein family. *A. thaliana* C4H (AtC4H) has a calculated molecular weight of 57.8kD (505 amino acid residues) and *S. moellendorffii* C4H (SmC4H) has a calculated molecular weight of 58.7kDa (518 amino acid residues).
C4H is a type IV P450, requiring a reductase partner to supply a free electron during catalysis. In *A. thaliana*, the reductase partner is either ATR1 or ATR2.

C4H catalyzes the second step of the phenylpropanoid pathway. After deamination of phenylalanine by phenylalanine amonina lyase, C4H catalyzes the hydroxylation of cinnamate (CA) at the 4 ring position into *para*-coumaric acid (pCA). pCA is a substrate for a variety of enzymes in the phenylpropanoid pathway. The 4-coumarate:coenzyme A ligase catalyzes the conversion of pCA into 4-coumaroyl-CoA. In turn, 4-coumaroyl-CoA can be converted into many secondary metabolites including flavones, isoflavones and proanthocyanidins (47). Alternately, 4-coumaroyl-CoA can be directed into monolignol synthesis by either cinnamoyl CoA reductase or hyroxycinnamoyl-CoA shikimate/quinate hydroxoycinnamoyl transferase (48).

A. thaliana and *S. moellendorffii* represent two diverse branches of plant evolution. In the lab, *S. moellendorffii* is the representative species of lycophyptes, an ancient branch of plants that diverged from ancestors of extant

flowering plants shortly after the rise of vascular tissue 400 million years ago (49). *A. thaliania* is a well studied angiosperm that has been used as a model organism for decades. As a result of diverging 400 million years ago, AtC4H and SmC4H share 70% sequence identity. SmC4H has the lowest sequence identity to AtC4H from the set of 106 C4H sequences available at the beginning of this study. The known functional motifs of the substrate recognition sequences (SRS) were originally identified in C4H proteins computationally, and confirmed with point mutations (45). The hinge motif and heme binding domain are highly conserved among membrane bound P450s and readily visualized in sequence alignments of C4H proteins (Table 3).

3.1.2 Multiple Sequence Alignment of P450 Proteins

Assigning structurally and functionally equivalent elements between proteins can be accomplished by using domain matching algorithms (50), structural contact maps (51), or visual inspection of sequences and structures. In this study, motif identification from protein homology models followed by visual inspection of a multiple sequence alignment (MSA) guided the division of parental AtC4H and SmC4H proteins into six gene fragments (Table 3).

To identify the gene fragments of the C4H protein, we construct a protein MSA of plant P450s related to flavonoid and lignin metabolism. This includes the CYP73, CYP74, CYP75, CYP79, CYP788, CYP84A1, CYP90, CYP97 and CYP98 protein families. After redundant sequences were removed from the alignment,

465 sequences are available for study. The sequences were aligned using MUSCLE (52,53).

3.1.3 Selection of Gene Fragments

Gene fragment were selected to contain at least one known functional motif. The first gene fragment contains the transmembrane domain and the hinge motif. Gene fragments 2, 3, 4 and 6 each contain a single SRS motif. Gene fragment 5 contains both a SRS motif and the heme binding domain common to all P450s. Recombining these gene fragments will allow us to determine the interchangeability of the SRS motifs and adjacent sequences in phenylpropanoid P450s.

Having selected six functionally equivalent gene regions, we must next select the exact crossover points between adjacent gene fragments. Some ambiguity as to the exact crossover points to use between gene fragments is present in the MSA as gaps and mismatches. This ambiguity is present as divergence between AtC4H and SmC4H, and is compounded by a lack of crystal structures for C4H proteins.

Researchers have addressed this uncertainty with methods that stochastically generate crossover locations across the entire length of the protein, as in ITCHY (54), SCRATCHY (55), or SHIPREC (56,57), and around short, preselected crossover regions as in SCOPE (58). When coupled with powerful selection

mechanisms, the large number of non-functional chimeras generated by stochastic crossover locations can be culled, thus leaving a small population of functional chimeras for further analysis.

Where powerful selection assays are not possible, methods for selecting effective, fixed breakpoints are required. SCHEMA (59), uses a structure and MSA based algorithm to select fixed crossover locations, known as breakpoints. The breakpoints developed by SCHEMA are fixed (not stochastic), and chimeric libraries with a high proportion of active members have been generated (51,10).

In this study, we chose to fix breakpoints at regions of high sequence conservation away from known functional motifs. By using fixed breakpoints we have ensured that all chimeric proteins contain gene fragments of the same extent. This will prove important later because it means that all gene fragments are independent, allowing us to build a statistical model of their role in activity.

Studies have shown that breakpoints in conserved secondary structural elements of proteins are not disruptive (10). This is an important consideration because we have chosen breakpoints at highly conserved regions of the MSA. These breakpoints almost certainly occur in the middle of secondary structural elements. We have reasoned that crossing over between two different genes in an evolutionarily conserved region is much more likely to conserve the structure-

function of a protein, rather than crossing over in an evolutionarily un-conserved (and structurally uncertain) region.

In the nomenclature used to describe protein chimeras, each chimera is represented by six capital letters, each designating the parental identity of a particular gene fragment. The letter A refers to AtC4H, whereas the letter S refers to SmC4H. For example, the reconstructed wild-type AtC4H is written as AAAAAA. The chimera containing gene fragments 1, 2 and 3 from AtC4H and gene fragments 4, 5 and 6 from SmC4H is written as AAASSS.

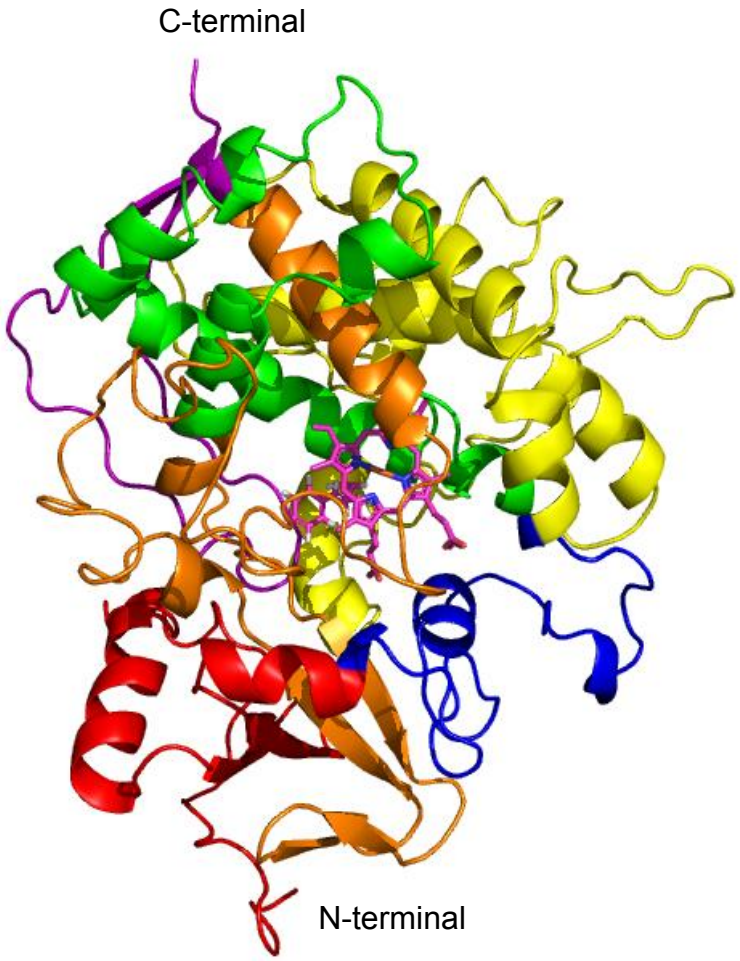


Figure 7 Colorized gene fragments of the AtC4H-SmC4H chimeric protein library with heme ligand.
Fragment 1 is red, fragment 2 blue, fragment 3 yellow, fragment 4 green, fragment 5 orange and fragment 6 purple. Heme ligand is shown as a stick model.

Table 3 Parental protein sequences with gene fragment breakpoints (vertical dotted lines) and functionally relevant sequence motifs (solid boxes)

Table 3 Continued

3.2 Results and Discussion

3.2.1 Codon Optimization of Parental Genes

To increase *in vivo* activity of both AtC4H and SmC4H, codons for each gene were optimized in several different ways and tested. Optimizing a length of N and C terminal codons has been shown to increase expression in yeast (60). The 12 N-teminal codons and the 5 C-terminal codons and all codons were changed to the most frequently occurring codons in *Saccharomyces Sp.* for each amino acid. This strategy was successful for improving measured *in vivo* activity for both AtC4H and SmC4H genes (Studies performed by Larisa Avramova, Bindley Bioscience Center, Purdue. Data not shown). The genes used in this study are AtC4H with optimized 12-N terminal codons and optimized 5 C-terminal codons (AtC4H N/C) and SmC4H with optimized 12-N terminal codons and optimized 5 C-terminal codons (SmC4H N/C). The AtC4H and SmC4H genes with all optimized codons did not have any measureable *in vivo* activity.

3.2.2 OLE PCR of Gene Fragments

In this study, we employed OLE-PCR to simultaneously recombine up to six gene fragments in a single PCR reaction. This method was developed by Dr. Thomas Sors and refined by the author. A two phase PCR scheme was employed to construct the chimeric genes. First, gene fragments (Table 4) were produced using hybrid oligonucleotide primers spanning breakpoints. These primers (Table 19) were used to synthesize gene fragments with homologous,

overlapping ends. These PCR derived gene fragments were purified by gel electrophoresis.

Gene fragments from round one PCR reactions were combined according to Table 5 by an automated BiomekFx workstation, and amplified with appropriate end primers containing restriction enzyme sites for cloning into the pYeDP60u shuttle vector (60).

OLE-PCR enables us to construct a specific, complete chimera per PCR reaction. This can be a distinct advantage over stochastic construction techniques conducted en masse. Dedicating one PCR reaction to construct each chimera allows us to deterministically recover a complete chimeric protein library. With the continual advancement in high throughput technologies, the additional cost of dedicating one PCR reaction to construct a single chimera is considerably reduced, especially when simultaneous stochastic construction of multiple chimeras may result in biased libraries. Recovery of a complete protein library greatly strengthens subsequent statistical analysis and modeling.

Simultaneous construction of all possible gene fragment combinations also has a distinct advantage over iterative techniques, since all possible chimeras are recovered at the same time, and with the same effort. Some iterative techniques also employ positive selection between rounds of crossing over. These techniques face the possibility of missing functional chimeras through erroneous

early round elimination of chimeras with simple crossovers. To use this library as an example, AAAASS is not functional, whereas the chimera SSAASS is highly active.

During early stages of our investigation, we found it necessary to optimize the second phase PCR reaction conditions to maximize target chimeric gene products. Conditions with the highest ratio of amplified full length chimeras in complete reactions, versus reactions missing one or more of the required gene fragments, are selected as final reaction conditions for generation of AtC4H-SmC4H chimeras (Figure 15). Utilizing annealing temperatures higher than suggested by the polymerase literature was also found to be helpful. Using the optimal PCR reaction conditions determined by this single experiment, we were able to amplify all 64 chimeras on a single 96-well PCR plate during a single PCR amplification

Table 4 Gene fragments used to construct the AtC4H-SmC4H and AtC4H-AtF5H protein libraries. Parent 2 is either SmC4H or AtF5H. Gene fragments are amplified from full length parental genes using hybrid primers listed in Table 19.

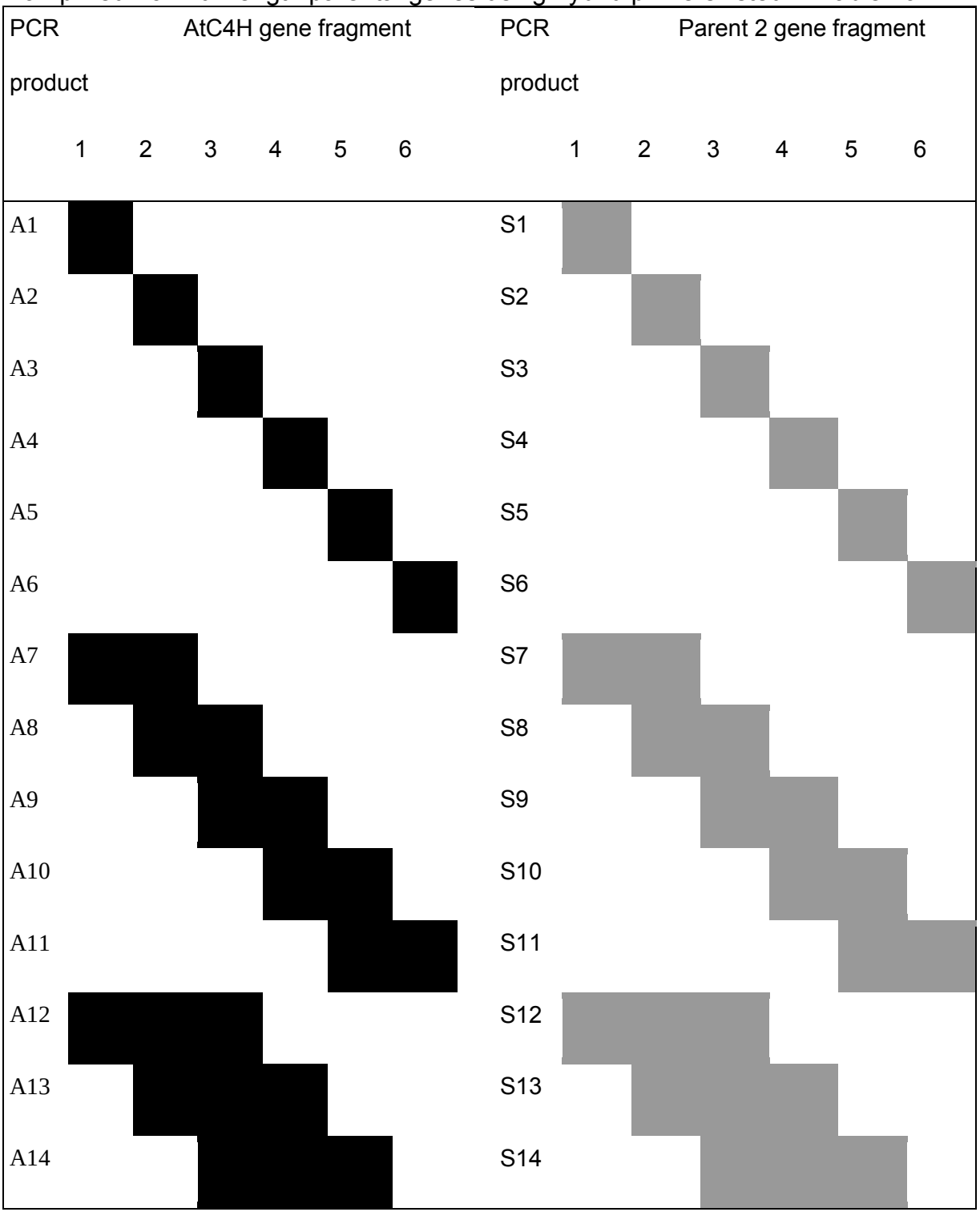


Table 4 Continued

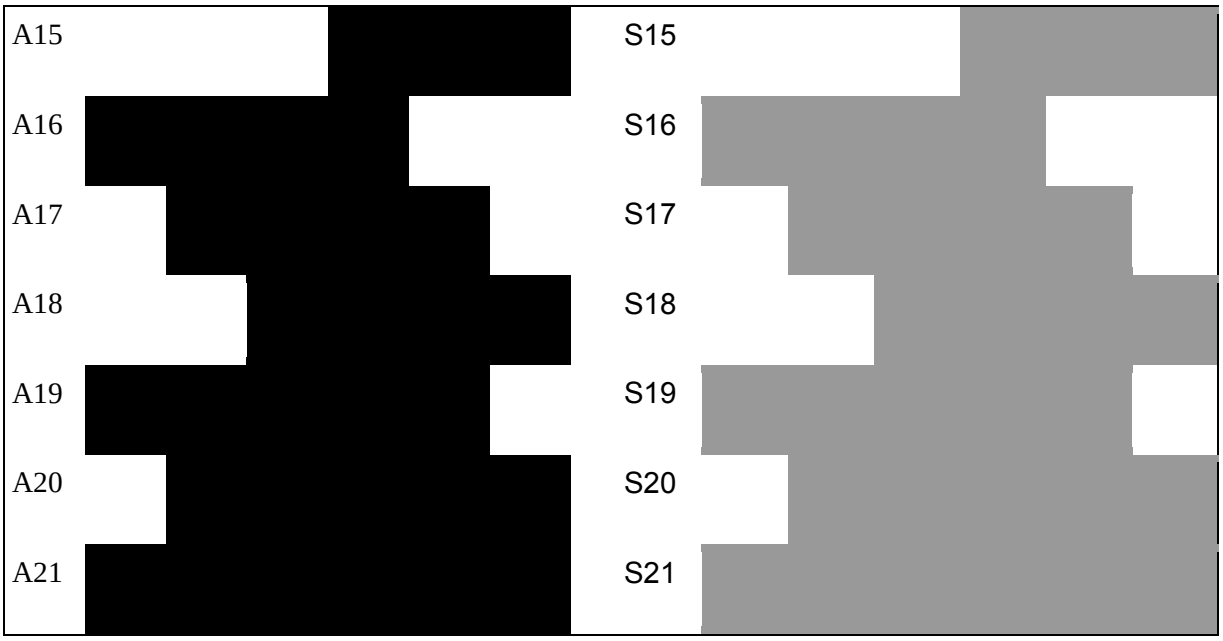


Table 5 Gene fragments used to construct each chimera in the AtC4H-SmC4H chimeric library. Gene fragments are coded as follows. Leading letter A indicates gene fragment from AtC4H. Leading letter S indicates gene fragment from SmC4H. Number corresponds to portion of each parental gene as described in Table 4

Chimera Number	Gene fragments	Chimera Number	Gene fragments
1	A20 S1	33	A12 S4 A11
2	S20 A1	34	S12 A4 S11
3	A19 S6	35	A12 S4 A6
4	S19 A6	36	S12 A4 S6
5	A18 S7	37	A7 S9 A11
6	S18 A7	38	S7 A9 S11
7	A18 S2 A1	39	S1 A2 S9 A11
8	S18 A2 S1	40	A1 S2 A9 S11
9	A17 S1 S6	41	S1 A8 S4 A11
10	S17 A1 A6	42	A1 S8 A4 S11
11	A16 S11	43	S7 A3 S4 A11
12	S16 A11	44	A7 S3 A4 S11
13	A16 S5 A6	45	A1 S2 A3 S4 A11
14	S16 A5 S6	46	S1 A2 S3 A4 S11
15	A15 S12	47	A1 S8 A10 S6
16	S15 A12	48	S1 A8 S10 A6
17	A15 S8 A1	49	A7 S3 A10 S6
18	S15 A8 S1	50	S7 A3 S10 A6

Table 5 Continued

Chimera Number	Gene fragments	Chimera Number	Gene fragments
19	A15 S3 A7	51	S1 A2 S3 A10 S6
20	S15 A3 S7	52	A1 S2 A3 S10 A6
21	A15 S3 A2 S1	53	S7 A9 S5 A6
22	S15 A3 S2 A1	54	A7 S9 A5 S6
23	A14 S6 S7	55	A1 S2 A9 S5 A6
24	S14 A6 A7	56	S1 A2 S9 A5 S6
25	A14 S6 S2 A1	57	S1 A8 S4 A5 S6
26	S14 A6 A2 S1	58	A1 S8 A4 S5 A6
27	A13 S1 S11	59	A7 S3 A4 S5 A6
28	S13 A1 A11	60	S7 A3 S4 A5 S6
29	A13 S1 S5 A6	61	S1 A2 S3 A4 S5 A6
30	S13 A1 A5 S6	62	A1 S2 A3 S4 A5 S6
31	A12 S10 A6	63	A21
32	S12 A10 S6	64	S21

3.2.3 Recovery and Sequencing of Chimeric Libraries

All 64 chimeras from the AtC4H-SmC4H chimeric libraries have been recovered. Chimeric genes were amplified in individual OLE-PCR reactions on a single 96 well PCR plate. Initial OLE-PCR reaction yielded bands for all chimeras except 43 and 46; OLE-PCR bands for chimeras 50, 53, 56, 59 and 62 were not stronger than a control reaction with one fragment not added (a “drop out” control) (Figure 8). Reactions for these seven chimeras were repeated by hand, and a single strong band at 1500bp was obtained in each reaction, indicating product (gel not shown). Failure of OLE-PCR reactions for these seven chimeras as observed in the above gel is assumed to be stochastic.

Sequences have been confirmed by sequencing both strands of each chimera at least once, and in the final library no point mutation, insertions or deletions are present. Multiple clones of some chimeras had to be sequenced before identifying a chimera without defects. The overall recovery rate was 60%. That is to say, for every chimera submitted for sequencing, there is a 60% chance of that chimera having no defects. Half of the observed defects are construction errors, with chimeras missing some gene fragments. A quarter of the defects are point mutations and the remaining quarter are single nucleotide deletions and insertions. No identifiable pattern was found among the observed defects. Sequence identity, chimera complexity, location of the PCR reaction on the 96 well PCR plate and other experimental factors were considered to try and explain these defects. We believe that the observed construction errors are inherent in

this construction procedure and the observed point mutations and insertions and deletions are the cost of subjecting DNA fragments to over forty cycles of PCR, and both of these errors arise stochastically.

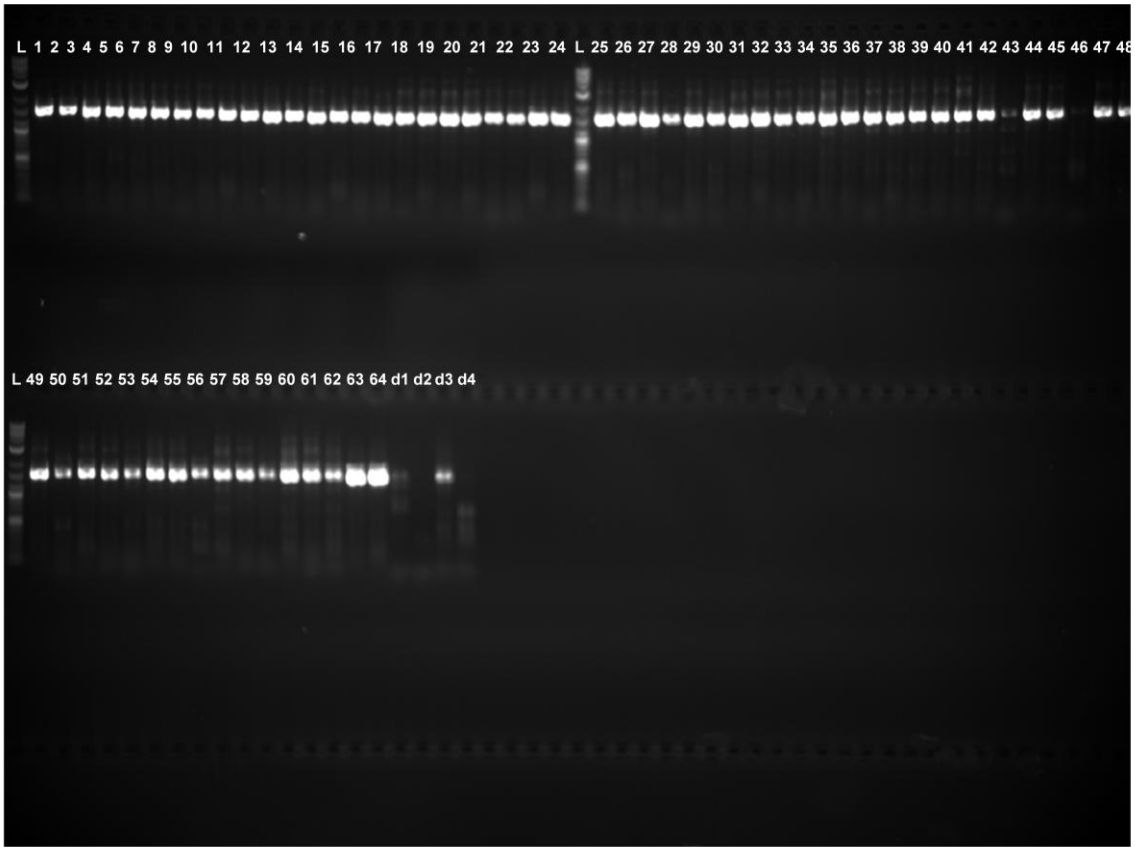


Figure 8 Representative gel of OLE-PCR products, including four “dropout” controls.

This gel shows the results of assembly of the AtC4H-AtF5H chimeric library Numbers above each lane indicate chimera numbers. d1- dropout control 1 containing only upstream fragment (S19) of chimera 4. d2- dropout control 2 containing only downstream fragment (A6) of chimera 4. d3- dropout control 3 missing the upstream fragment for chimera 59 (S3, A4, S5, A6). d4- dropout control 4 missing the middle gene fragment for chimera 59 (A7, S3, S5, A6). See section 3.3.6 for a more complete explanation of the dropout controls

3.2.4 Functional Screening of Chimeric Libraries

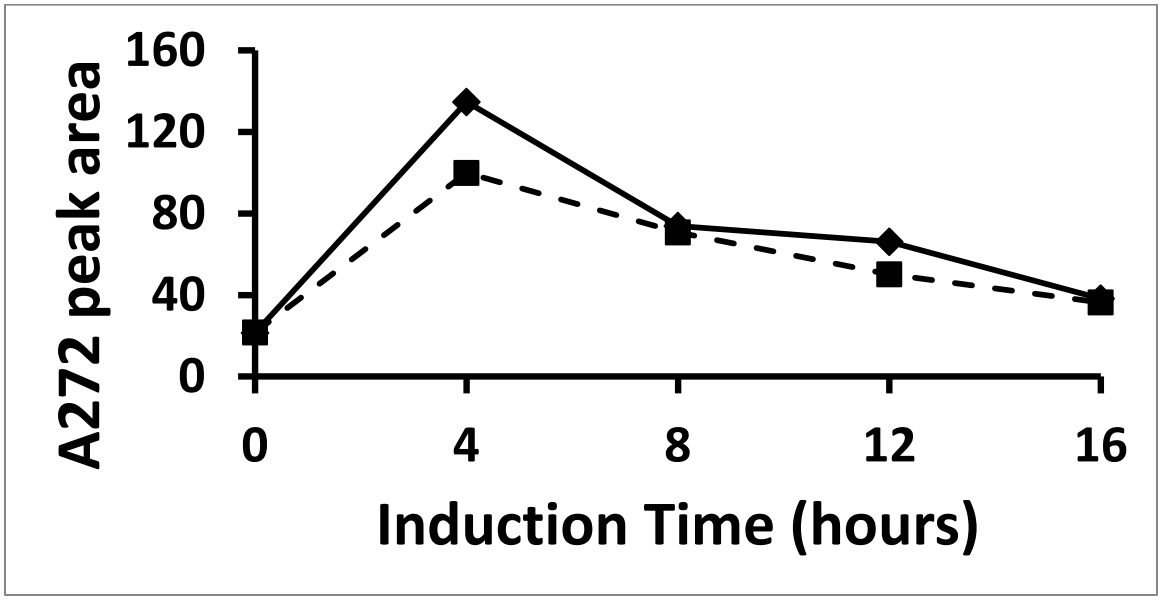
Expression and analysis of C4H and F5H proteins has already been established using the pYeDP60 shuttle vector in the *Saccharomyces cerevisiae* strain Wat11 (61,62). pYeDP60 is a shuttle vector capable of high copy number expression in *E. coli* and yeast. It is optimized for high expression levels in yeast by combining the *GAL10-CYC1* inducible promoter with the *ADE2* gene. Inducible promoters avoid possible toxic effects from constitutive gene expression, while the *ADE2* gene selects for the multiple required plasmid copies to fully complement lost purine synthesis in ade^- Wat11 yeast. A recent update to this expression system is pYeDP60u (60), which incorporates a Kozak sequence and TAAT stop codon for optimized gene expression in yeast. In this study, we use the updated pYeDP60u plasmid in Wat11 for expression of all genes. The updated pYeDP60u plasmid also incorporates the USER cloning system. After unsuccessful trials with the USER cloning system, we chose instead to rely on restriction enzyme mediated ligation for cloning into the pYeDP60u plasmid.

Measured activity for Wat11 carrying pYeDP60-AtC4H or pYeDP60-AtF5H was previously reported to peak between 16 to 20 hours after galactose induction (63,62). Induction times for Wat11 carrying the updated pYeDP60u plasmid have not been reported. To determine the optimal time to perform *in vivo* assays after galactose induction, Wat11 strains carrying AtC4H or AtF5H in pYeDP60u were tested for activity every 2 hours after galactose induction, for up to 20 hours. Measured *in vivo* activity peaks almost immediately and stays high for up to 4 to

6 hours, before beginning a roughly linear decrease (Figure 9). This is very different from the induction timing reported for Wat11 strains carrying the pYeDP60 plasmid (63,62). Wat11 carrying pYeDP60 and expressing either AtC4H or AtF5H have been previously shown to have very little measureable activity immediately after galactose induction, followed by a continual increase in measured activity, with maximum in vivo activity occurring 16-20 hours after induction.

As a result of the above tests, all protein chimeras were induced in 96-well deep well blocks and substrate (either CA or coniferyl alcohol) was added four hours after induction. Supernatant from in vivo reactions was analyzed on HPLC for either p-coumarate (pCA) after one hour after addition of CA, or 5-hydroxy coniferyl alcohol production four hours after addition of coniferyl alcohol. For each chimera, a minimum of four different isolated Wat11 colonies were selected from the transformation plates for assay. The selected colonies were all of intermediate size and were distributed across the transformation plate. All Wat11 colonies expressing a given chimera either showed activity, or not. No discordant colonies were ever observed, although there was variation in levels of observed activity dependent on observed cell density immediately prior to assay (section 3.2.5).

A



B

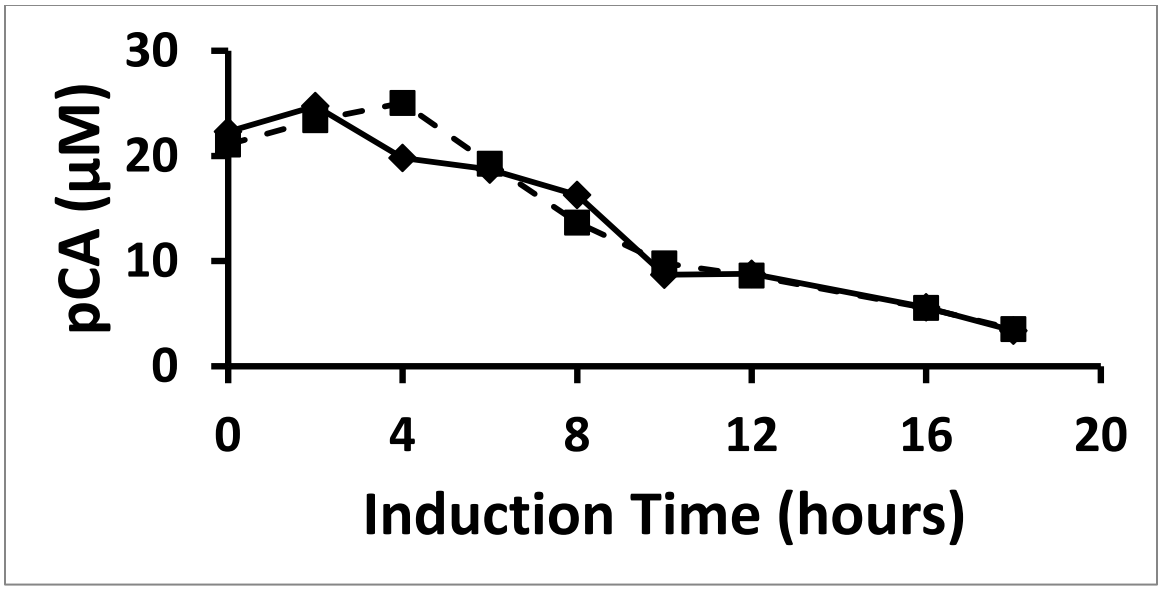


Figure 9 In vivo activity of Watt11 carrying pYeDP60u AtF5H (A) or pYeDP60u AtC4H(B). Times are hours after galactose induction.

3.2.5 AtC4H-SmC4H Chimeric Library Regression Analysis

All chimeras from the AtC4H-SmC4H chimeric library were tested for activity against CA. As expected, both reconstructed AtC4H and reconstructed SmC4H have wild-type activity against CA. Thirty four of the sixty two recombinant chimeras in the AtC4H-SmC4H library show activity against CA (Table 6). This result suggests that evolution has preserved key features among C4H proteins which allow for extensive gene fragment interchangeability.

To complete an analysis of chimera activities, multiple high throughput experiments were combined into a single dataset. Since differences between experiments may affect the overall measured response of each in vivo assay, the parents, the empty vector, and a set of overlapping chimeras with high, moderate, and low activities were repeated in different experiments. Comparing the activity of the repeated chimeras allows us to account for differences in average measured activity between experiments.

During data analysis, we noticed an inverse relationship between measured activity and OD_{550} of induced Wat11 yeast cultures immediately prior to assay (Figure 10). When evaluating the reproducibility of each chimera across multiple experiments, including OD_{550}^{-1} as a covariate explains 94 percent of the total observed variability in the dataset. The remaining 6 percent variability is due to experimental error. Note that this is without any additional scaling of the data between experiments. Including a scale factor between experiments, either with

or without, OD_{550}^{-1} as a covariate, actually results in a worse fit than using OD_{550}^{-1} as a covariate alone. Therefore, we conclude that differences between experiments are due to differences in OD_{550}, and that inherent differences between experiments do not significantly contribute to measured variability in pCA production by the chimeras. With the OD_{550}^{-1} correction, this experimental system then has high precision.

Currently, there is no theory to suggest a particular relationship between sequence and function of a given protein. Indeed, our work is directed towards beginning to understand this relationship. One simulation study (64) suggests that logistic models succeed at identifying underlying interaction terms, however, the study is limited and does not include linear data present in many activity studies.

Given the design of our experiment, it is natural to consider each gene fragment as an explanatory variable in a non-parametric model. Since there is no known relationship between the non-parametric gene fragment terms and protein chimera activity, I developed a set of logistic and linear models in order to form a consensus for significant terms. The first set of regression models we chose were logistic ANOVA models distinguishing different levels of activity. Here, we hope to learn if gene fragments and their two-way interactions make similar contributions to chimeras of high, medium or low activity. If they do, this would support the hypothesis that the functional contribution of gene fragments is

largely independent of their context. We also fit a linear ANOVA model with all gene fragments and their two-way interaction terms. Three and four way interaction terms were considered, but not significant. Here, we check for over fitting and compare to the logistic models. We expect the linear and logistic models to share significant explanatory variables.

Three logistic models were fit. For the active model, chimeras were either considered active if any pCA production was observed or not active if no pCA production was observed. In the geometric mean model (Geo), the active chimeras were considered high activity if their observed pCA production was greater than the geometric mean of the parental controls AtC4H and SmC4H (66 μM pCA hr^{-1}), adjusted for OD_{550} prior to assay, otherwise chimeras were considered low activity. In the median model, chimeras were considered high activity if their observed pCA production was greater than the median observed activity; otherwise chimeras are considered low activity (Table 7).

A consensus and complete linear model were developed. For the complete linear ANOVA model (Table 8), the significant terms were determined by iterative subtraction of insignificant terms by t-test. The consensus model involves those terms found to be significant in all models. The activity data used to fit the complete linear model is said to be “right censored”. In statistics, censored data means is when a value can only be accurately measured over a certain interval. In this study, we can only accurately measure the activity of chimeras greater

than 1μM pCA hr^{-1}. Chimeras with activity below this level cannot be distinguished from inactive chimeras. When only a lower bound exists on a measured response, it is referred to as right censored. Tobit models have been developed as an alternative to least squares regression to avoid bias when fitting right censored data (65). Fitting a two-way Tobit model to the activity does not alter which parameters are significant, nor were any changes in the magnitude of the fit parameters found.

Comparing all of the linear and logistic regression models, we see that a similar, but not identical set of gene fragments significantly contribute to activity. All regression models show that in general, SmC4H gene fragments negatively influence chimeric activity over their corresponding AtC4H gene fragments within the same order of magnitude (Table 8). This is not surprising since N/C optimized AtC4H is almost five times as active as N/C optimized SmC4H. Part of this difference in activity may be due to the expression system. C4H requires a reductase partner for enzymatic activity. In plants this is provided by ATR1 and ATR2. Wat11 has been engineered to express ATR1 from *A. thaliana.* The lower observed activity of SmC4H may be partly due to poor cross-species protein interaction between SmC4H and *A. thaliana* ATR1, expressed by Wat11. In this case, the effects of the ATR interaction may form a significant part of the underlying model.

Since there is no theory that suggests use of a particular regression model, or a particular way to evaluate a given regression model for inferring protein activity, we took a conservative approach. We considered a given gene fragment or interaction functionally relevant if it appeared in multiple models and the p-value was less than a Bonferroni correction for the model. For example, all six individual gene fragments and the 4:5 interaction are statistically significant in all of the models and have a p-value less than a Bonferroni correction for most of the fit models. We consider these factors to be real. In contrast, the interactions 3:4 and 3:5 are present in only half of the fit models, and are not always below Bonferroni correction when present. Therefore, we do not conclude that these interactions are fundamental to explaining the differences in observed activity of the AtC4H-SmC4H chimeras, and could be a statistical artifact. However, structural information does support the potential importance of these interactions and they are presented as a possibility for future investigation.

Using this conservative approach, we identified a consensus of gene fragments and interactions that significantly contribute to AtC4H-SmC4H activity (Table 9, Table 10, Table 11, Table 12, Table 13, Table 14). The consensus linear model (Table 8 and Figure 13) contains gene fragments 1, 2, 3, 4, 5, 6 and the interaction term 4:5. These terms explain most of the activity variation observed between chimeras with an adjusted R^2=0.8353.

Twenty-six active chimeras contain fragment 1 from SmC4H while only 9 active chimeras contain fragment 1 from AtC4H. Evaluated by Fischer's exact test, this difference is significant with p-value <0.01. One possible explanation is that SmC4H fragment 1 increases expression in Wat11 cells. However, SmC4H fragment 1 is only associated with moderately active chimeras. Of the 9 active chimeras containing AtC4H fragment 1, four are in the top five most active chimeras. This moderate activity observed for chimeras containing SmC4H fragment 1 suggests that SmC4H fragment 1 does not increase gene expression in yeast. Instead, it suggests that SmC4H fragment 1 may be stabilizing the tertiary structure of C4H proteins, leading to a greater fraction of active chimeras without an overall increase activity. Testing the thermostability of proteins carrying SmC4H fragment 1 would help answer this question.

SmC4H fragment 2 appears to reduce the catalytic activity of chimeras. The effect of SmC4H fragment 2 is seen when comparing AAAAAA to ASAAAA (171±10μM pCA Hr^{-1} and 144±6μM pCA Hr^{-1}, respectively) and SASSSS to SSSSSS (70±28μM pCA Hr^{-1} and 36±7μM pCA Hr^{-1}, respectively). Fragment 2 contains only 2 divergent residues; the fewest number of divergent residues out of all six gene fragments. They are L93H and K129R, both of which occur outside of the SRS motif. One possibility is that the SRS motifs in C4H proteins are larger than other cytochrome P450 proteins, or that these amino acids are important for positioning the SRS residues, or enhances protein stability during folding and/or catalysis.

Overall, SmC4H fragment 3 reduces activity of chimeras compared to AtC4H fragment 3, but this occurs in a context dependent manner. For example, SmC4H fragment 3 greatly reduces activity when comparing AAAAAA to AASAAA (171±10μM pCA Hr^{-1} and 45±3μM pCA Hr^{-1}, respectively), but SmC4H fragment 3 does not have a significant negative impact on activity when comparing SSASSS to SSSSSS (30±21μM pCA Hr^{-1} and 36±7μM pCA Hr^{-1}, respectively), or SSSASA to SSAASA (21±7μM pCA Hr^{-1} and 15±9μM pCA Hr^{-1}, respectively). Out of the 32 pairs of chimeras which differ only by fragment 3, only 4 of these pairs contain only one functional chimera (chimera (12,43), (6,16), (32, 23) and 34, 38)). The remaining 28 chimera pairs are either both functional, or both non-functional.

The context dependent nature that SmC4H fragment 3 has on activity implies structure-function interactions between fragment 3 and the surrounding gene fragments. Statistically significant interactions are present in a subset of the models between fragment 3 and fragments 4 and 5.

Fragment 3 is the largest gene fragment in this system. It is 161 amino acids long, comprising almost one third the entire length of the protein and it contains most of the core helicies including C, D, E, F and G. Therefore, we might think of fragment 3 as forming the core of the protein, around which the rest of the gene fragments are placed. It would be natural for interactions to arise in this

circumstance and future libraries may subdivide fragment 3, testing the constraints contained within this region, and between neighboring regions.

Gene fragment 4 is a significant parameter in all statistical models; the *S moellendorffii* variant having a strong negative effect on activity. Gene fragment 4 is not replaceable between SmC4H and AtC4H (Table 16). Both single fragment replacement chimeras AAASAA and SSSASS have no measured *in vivo* activity. This gene fragment is 65 amino acids long with 17 polymorphisms; fewer polymorphisms than either fragment 3 (58) or fragment 1 (35), both of which are replaceable between *A. thaliana* and *S. moellendorffii*. This gene fragment is predicted to be involved in substrate orientation during catalysis. Classically, we would expect this to the most conserved part of the protein. However, the lack of interchangeability suggests a structural divergence between AtC4H and SmC4H.

Gene fragment 5 is a significant parameter in all statistical models, with S. moellendorffii having a strong negative effect on activity. Gene fragment 5 is not replaceable between AtC4H and SmC4H. Single fragment replacement chimeras AAAASA and SSSAAS have no measured in vivo activity. This gene fragment is 103 amino acids long with 24 point mutations and 2 gaps. Similar to gene fragment 4, gene fragment 5 is also predicted to be involved in substrate orientation during catalysis. The non-replaceability of gene fragment 5 further

supports the hypothesis that the catalytic regions of AtC4H and SmC4H have structurally diverged.

All regression models support a strong fragment 4:5 interaction, favoring chimeras with fragment 4 and 5 from the same parent. Only 11 out of 34 active chimeras contain gene fragment 4 and 5 from different parents (Figure 12). The published homology model of C4H suggests that fragment 4 and 5 orient the CA substrate with the heme domain during catalysis. This supports the fragment 4-5 interactions observed in the regression models.

One might expect that the catalytic region of a protein would be highly conserved due to stringent functional constraints. The result would be interchangeable catalytic regions across species. However, these results suggest that the catalytic region in AtC4H and SmC4H have structurally diverged during their evolution. If this is the true, then we expect a clade based replacement pattern when testing gene fragments 4 and 5 from other extant C4H sequences. Alternately, future C4H chimeric libraries might combine gene fragments 4 and 5 into a single gene fragment to enrich the proportion of active chimeras.

SmC4H fragment 6 has an overall negative effect on activity compared to AtC4H fragment 6. There is evidence for a fragment 5:6 interaction, but it is not seen in all statistical models. Gene fragment 6 is considered to be interchangeable between *A. thaliana* and *S. moellendorffii*. Creation of new chimeric libraries with

a greater diversity of parents, or subdividing the gene fragments 5 and 6 would enable a more careful testing of any structure-function-sequences constraints present in this region of the protein.

AtC4H-SmC4H chimeras have been tested for catalytic activity on CA. Studying the function of chimeras against the natural substrate of the parental sequences has provided many insights. Future work testing alternate substrates against the AtC4H-SmC4H library will measure diversification of substrate recognition. It is known that AtC4H is sensitive to large substitutions on the 3-ring position of CA (44). It is not known if, or how, SmC4H is sensitive to substitutions on the 3-ring position of CA. If SmC4H is divergent from AtC4H regarding sensitivity to substitutions on the 3-ring position of CA, then changes in this functional constraint can be identified by testing for activity against 3-Cl-cinnamate, 3-Methyl-cinnamate and 3-Methoxy cinnamate. Otherwise, if AtC4H and SmC4H are not divergent regarding sensitivity to substitutions on the 3-ring position of CA, then the previously mentioned substrates can be used to probe perturbations or relaxations in substrate specificity of C4H chimeras, possibly revealing new functional constraints.

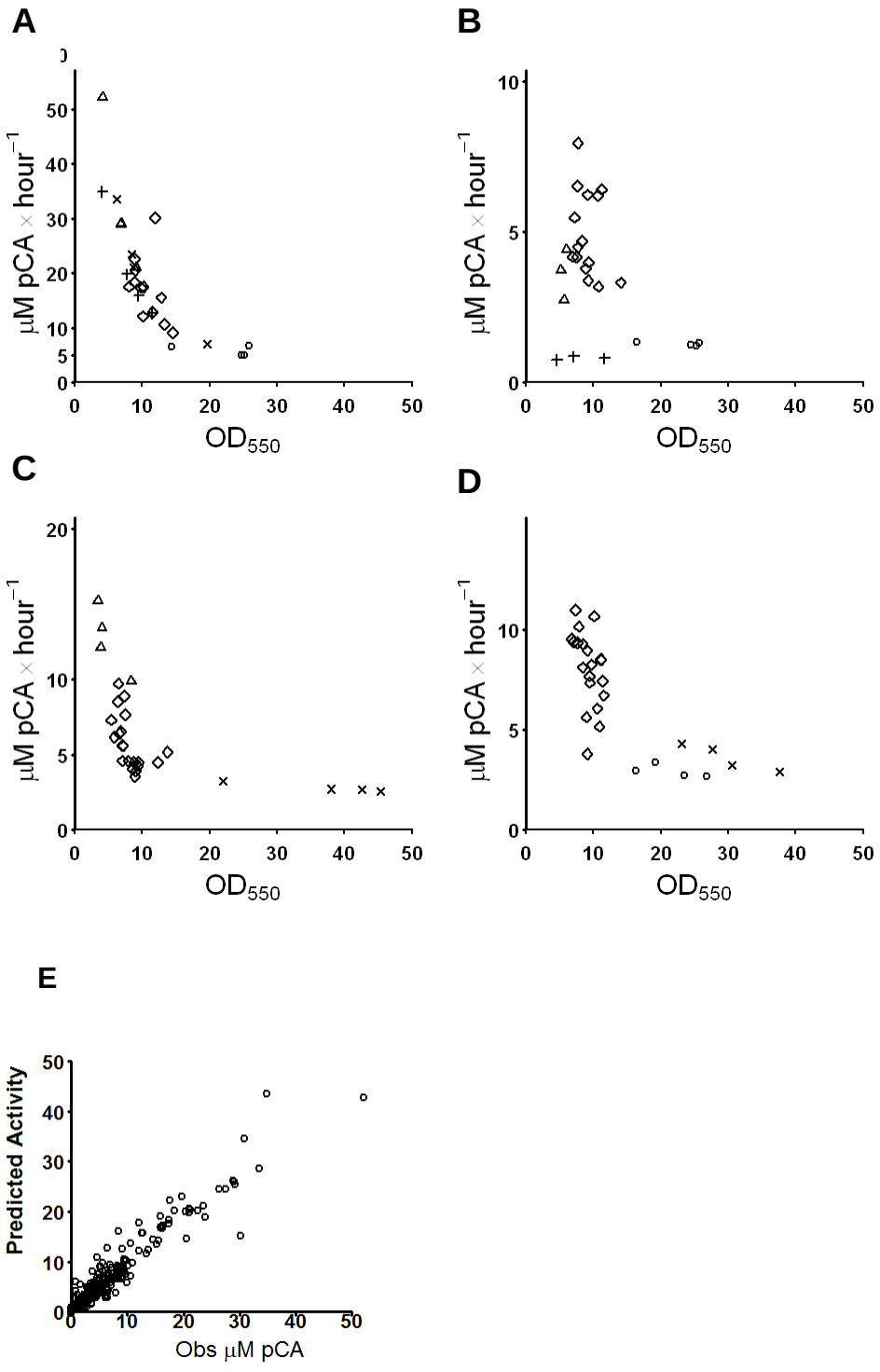


Figure 10 Inverse relationship between pCA production and OD_{550} of induced yeast cultures prior to assay. Each symbol represents one yeast culture, and

different symbols represent experiments preformed on different days. Open circles are experiment 1, open triangles are experiment 2, crosses are experiment 3, “x”s are experiment 4, and open diamonds are experiment 5. A is parental AtC4H, B is parental SmC4H, C is chimera 5 and D is chimera 19. Note that not all chimeras were tested in every experiment. E- Predicted vs. observed for $Y_{ij}=chimera_i*OD^{-1}_{ij}+\varepsilon_{ij}$ model

Table 6 Fit activity estimates of each chimera.
[a]-Chimera column indicates fragment identity of each chimera. 'A' indicates gene fragment from AtC4H and 'S' indicates gene fragment from SmC4H. [b]-units are μM pCA hr^{-1}. Fit values are based on the model: $Y_{ij}=chimera_{ij}:OD_{550ij}^{-1}\ \varepsilon_{ij}$, where OD_{550ij}^{-1} term is a covariate. This model contains 64 fit terms; one for each row of this table. Chimeras with zero activity never produced measureable amounts of pCA and were estimated to have small negative numbers by the full model. [C]-Standard error of the activity estimates were based on the full model.

Number	Chimera[a]	Activity[b]	Std. Error[c]	Number	Chimera[a]	Activity[b]	Std. Error[c]
1	SAAAAA	148.5	±6.5	39	SASSAA	4.4	±12.3
10	ASSSSA	0.0		4	SSSSSA	8.9	±15.4
11	AAAASS	0.0		40	ASAASS	0.0	
12	SSSSAA	0.0		41	SAASAA	37.8	±5.9
13	AAAASA	0.0		42	ASSASS	0.0	
14	SSSSAS	0.0		43	SSASAA	28.7	±8.2
15	AAASSS	16.2	±13.4	44	AASASS	0.0	
16	SSSAAA	1.9	±5.2	45	ASASAA	12.9	±7.7
17	ASSAAA	5.5	±11.7	46	SASASS	22.9	±8.0
18	SAASSS	54.3	±7.0	47	ASSAAS	15.5	±14.5
19	AASAAA	44.5	±3.0	48	SAASSA	45.5	±7.1
2	ASSSSS	0.0		49	AASAAS	7.0	±4.9
20	SSASSS	28.9	±21.2	5	SSAAAA	67.2	±4.5
21	SASAAA	29.8	±14.6	50	SSASSA	45.2	±8.9
22	ASASSS	0.0		51	SASAAS	10.5	±4.7
23	SSAAAS	40.2	±5.4	52	ASASSA	0.0	
24	AASSSA	0.0		53	SSAASA	15.3	±9.0

Table 6 continued

Number	Chimera[a]	Activity[b]	Std. Error[c]	Number	Chimera[a]	Activity[b]	Std. Error[c]
25	ASAAAS	36.6	±19.0	54	AASSAS	0.0	
26	SASSSA	23.7	±6.7	55	ASAASA	0.0	
27	SAAASS	19.9	±3.2	56	SASSAS	0.0	
28	ASSSAA	0.0		57	SAASAS	0.0	
29	SAAASA	26.0	±3.7	58	ASSASA	0.0	
3	AAAAAS	136.2	±33.3	59	AASASA	0.0	
30	ASSSAS	0.0		6	SSAAAA	0.0	
31	AAASSA	0.0		60	SSASAS	0.0	
32	SSSAAS	0.0		61	SASASA	39.7	±16.2
33	AAASAA	0.0		62	ASASAS	0.0	
34	SSSASS	0.0		63	AAAAAA	171.6	±10.1
35	AAASAS	0.0		64	SSSSSS	36.4	±6.9
36	SSSASA	20.7	±6.9	7	ASAAAA	144.2	±5.7
37	AASSAA	0.0		8	SASSSS	71.5	±28.2
38	SSAASS	47.0	±13.3	9	SAAAAS	55.3	±18.7

Table 7 Significant explanatory variables of fit for the logistic models. Individual parameters for all logistic models are fit by backwards selection. Initially, all single fragment and two body terms are included in the model. The term with the greatest p-value > 0.1000 is removed, and the remaining terms are refit. This process is repeated until all fit terms have a p-value < 0.1000. Asterisk indicates explanatory term is significant at the Bonferroni level (α≤0.0041).

Factor		Active			Hi:Lo geometric mean			Hi:Lo median		
		Estimate	P-value		Estimate	P-value		Estimate	P-value	
α		10.06	<0.0001	*	4.99	<0.0001	*	18.026	<0.0001	*
1	Sm	8.38	0.0073		-3.09	0.0011	*			
2	Sm	8.15	0.0075		-4.11	<0.0001	*	-3.27	0.0086	
3	Sm	-7.01	<0.0001	*	-5.95	<0.0001	*	-16.04	0.0003	*
4	Sm	-14.52	<0.0001	*	-15.88	0.0032	*	-17.71	0.0001	*
5	Sm	-18.50	<0.0001	*	1.63	<0.0001	*	-17.97	<0.0001	*
6	Sm							-14.99	0.0007	*
1:2	Sm:Sm	-8.45	0.0226		1.63	0.0905		2.51	0.0370	
1:3	Sm:Sm	19.95	<0.0001	*						
1:4	Sm:Sm									
1:5	Sm:Sm				9.72	0.0611				
1:6	Sm:Sm	-7.19	0.0067							
2:3	Sm:Sm							-1.82	0.0823	
2:4	Sm:Sm				-4.56	0.0053				
2:5	Sm:Sm	-7.78	0.0303							
2:6	Sm:Sm	-6.84	0.0049							
3:4	Sm:Sm							-6.40	0.0908	
3:5	Sm:Sm							16.53	0.0002	*
3:6	Sm:Sm							8.60	0.0228	

Table 7 Continued

Factor	Active			Hi:Lo geometric mean			Hi:Lo median		
	Estimate	P-value		Estimate	P-value		Estimate	P-value	
4:5 Sm:Sm	19.73	<0.0001	*	9.23	<0.0001	*	19.55	<0.0001	*
4:6 Sm:Sm				6.33	0.0002	*			
5:6 Sm:Sm	6.02	0.0158					11.95	0.0056	
Model Df		313			190			188	
Model Dev		30.5			91.4			104.2	

Table 8 –Significant fit terms for linear ANOVA models using one and two way gene fragments as factors. Terms for the complete linear model were fit as described in Table 7. Terms for the consensus linear model are terms and interactions significant at the Bonferroni level across all regression models. Asterisks indicate explanatory factor is significant at the Bonferroni level (α≤0.0041).

Factor		Consensus Linear Model			Complete Linear model		
		Estimate	P-value		Estimate	P-value	
Intercept		84.72	<0.0001	*	129.60	<0.0001	*
1	Sm	12.76	0.1204		-25.34	0.0299	
2	Sm	-14.41	0.0805		-14.40	0.0160	
3	Sm	-27.79	0.0011	*	-70.13	<0.0001	*
4	Sm	-62.70	<0.0001	*	-91.71	<0.0001	*
5	Sm	-55.35	<0.0001	*	-114.88	<0.0001	*
6	Sm	-14.14	0.0862		-30.28	0.0005	*
1:2	Sm:Sm						
1:3	Sm:Sm				2.44	0.8338	
1:4	Sm:Sm				26.23	0.0276	
1:5	Sm:Sm				47.55	0.0001	*
1:6	Sm:Sm						
2:3	Sm:Sm						
2:4	Sm:Sm						
2:5	Sm:Sm						
2:6	Sm:Sm						
3:4	Sm:Sm				37.41	0.0021	*
3:5	Sm:Sm				44.85	0.0003	*
3:6	Sm:Sm						

Table 8 Continued

4:5	Sm:Sm	73.78	<0.0001	*	68.17	<0.0001	*
4:6	Sm:Sm						
5:6	Sm:Sm				32.28	0.0073	

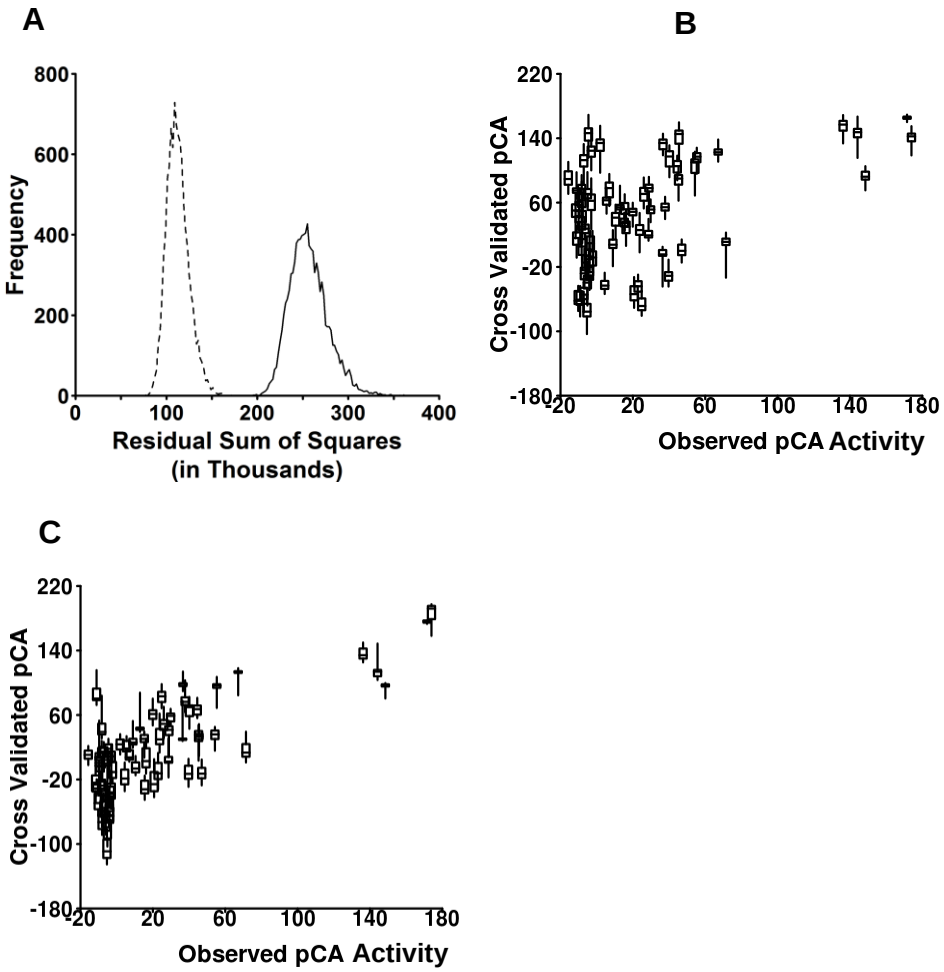


Figure 11 Cross validation of linear models.
Regression Models were cross validated by conducting 10000 repeated sub samplings of 8-fold cross-validations. What is plotted are the sum of square totals of the estimates on each subsample. A-The open bars are the consensus model and the hatched bars are the linear model. B and C box and whiskers plot of the residuals for each subsample by chimera. Box indicates 25 and 75 percentile. Whiskers indicate 10 and 90 percentile. B is for the consensus linear model and C is complete linear model.

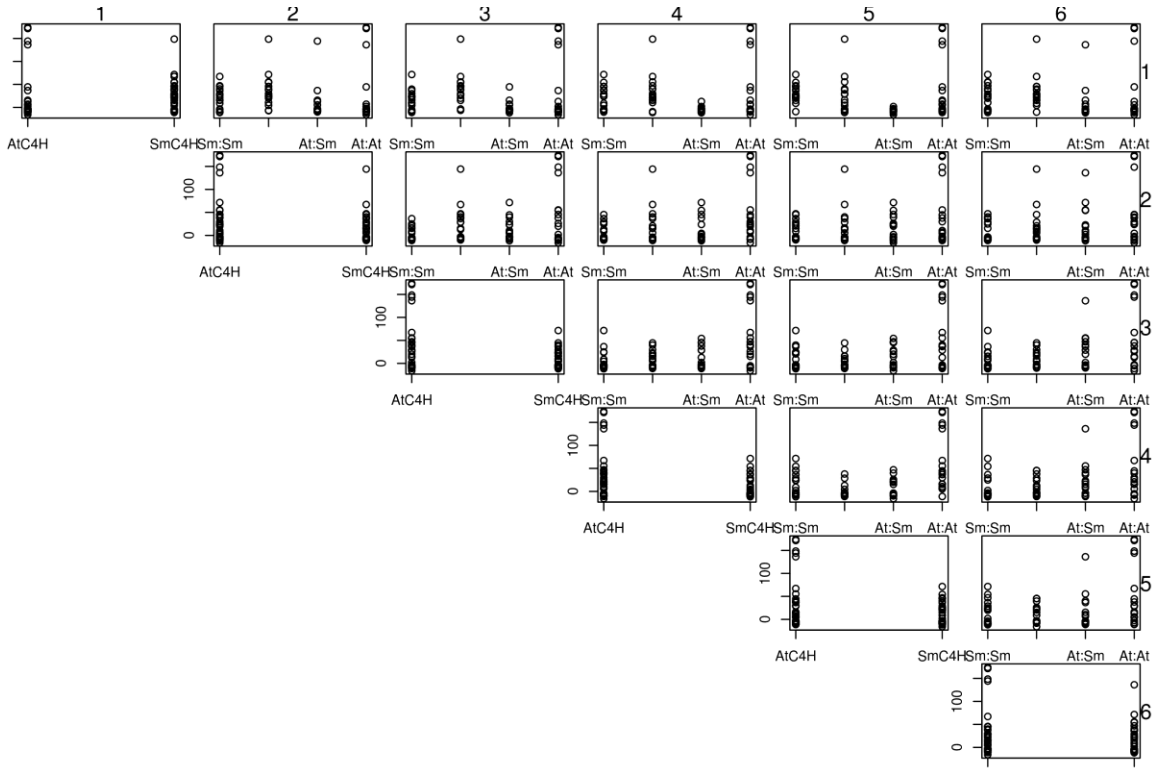


Figure 12 Two-Way interaction plots. Each graph shows activity of all 64 chimeras plotted against the identity of selected pair of gene fragments. Selected gene fragments are indicated by the column and row of each plot. For example: the plot in row 1, column 2 shows the activity of all 64 chimeras conditional on the identity of gene fragments 1 and 2. Plots along the diagonal are conditioned to a single gene fragment.

Fragment	1	2	3	4	5	6
				++		
AtC4H	++	++	++	-	-	++
SmC4H	+	+	+	-	-	+
				+		

Figure 13 Consensus model with relative activity by gene fragment. Consensus terms come from logistic and linear models for contributions to activity by gene fragment. + indicates small, positive effect on activity. ++ indicates large positive effect on activity. – indicates negative effect on activity.

Table 9 Significant terms involving gene fragment 1 in the regression models fit to AtC4H-SmC4H activity data. X indicates term is significant in a given model. + indicates fit term is significant and is positive in value. – indicates fit term is significant and is negative in value. Note the role of Sm fragment one changes from promoting activity in weakly active chimeras to reducing activity in highly active chimeras. This is the only explanatory variable that does this.

		Logistic Models			Linear Models
		Active	Median	Geo Mean	Complete Linear model
1	Sm	+		-	-
2:1	Sm:Sm	X	X	X	
3:1	Sm:Sm	X			X
4:1	Sm:Sm				X
5:1	Sm:Sm			X	X
6:1	Sm:Sm	X			

Table 10 Significant terms involving gene fragment 2 in the regression models fit to AtC4H-SmC4H activity data. X indicates term is significant in a given model

		Logistic Models			Linear Models
		Active	Median	Geo Mean	Complete Linear model
2	Sm	X	X	X	X
1:2	Sm:Sm	X		X	
3:2	Sm:Sm		X		
4:2	Sm:Sm			X	
5:2	Sm:Sm	X			
6:2	Sm:Sm	X			

Table 11 Significant terms involving gene fragment 3 in the regression models fit to AtC4H-SmC4H activity data. X indicates term is significant in a given model.

		Logistic Models			Linear Models
		Active	Median	Geo Mean	Complete Linear model
3	Sm	X	X	X	X
1:3	Sm:Sm	X			
2:3	Sm:Sm		X		X
4:3	Sm:Sm		X		X
5:3	Sm:Sm		X		X
6:3	Sm:Sm		X		

Table 12 Significant terms involving gene fragment 4 in the regression models fit to AtC4H-SmC4H activity data. X indicates term is significant in a given model.

		Logistic Models			Linear Models
		Active	Median	Geo Mean	Complete Linear model
4	Sm	X	X	X	X
1:4	Sm:Sm				X
2:4	Sm:Sm			X	
3:4	Sm:Sm				X
5:4	Sm:Sm	X	X	X	X
6:4	Sm:Sm				

Table 13 Significant terms involving gene fragment 5 in the regression models fit to AtC4H-SmC4H activity data. X indicates term is significant in a given model.

		Logistic Models			Linear Models
		Active	Median	Geo Mean	Complete Linear model
5	Sm	X	X	X	X
1:5	Sm:Sm			X	X
2:5	Sm:Sm	X			
3:5	Sm:Sm		X		X
4:5	Sm:Sm	X	X	X	X
6:5	Sm:Sm	X			

Table 14 Significant terms for gene fragments 4, 5 and the 4:5 interaction in the regression models fit to AtC4H-SmC4H activity data. X indicates term is significant in a given model.

		Logistic Models			Linear Models
		Active	Median	Geo Mean	Complete Linear Model
4	Sm	X	X	X	X
5	Sm	X	X	X	X
4:5	Sm:Sm	X	X	X	X

Table 15 Significant terms involving gene fragment 6 in the regression models fit to AtC4H-SmC4H activity data. X indicates term is significant in a given model.

	Active	Median	Geo Mean	Liner model
6		X		X
1:6	X			
2:6	X	X		
3:6		X		
4:6				
5:6	X		X	X

Table 16 Activity of single fragment replacement chimeras.
Activity levels are relative to AtC4H parent.

Chimera	Activity (%)		Chimera
AAAAAA	100	21	SSSSSS
SAAAAA	91	0	ASSSSS
ASAAAA	52	44	SASSSS
AASAAA	40	21	SSASSS
AAASAA	0	0	SSSASS
AAAASA	0	0	SSSSAS
AAAAAS	76	5	SSSSSA

3.3 Materials and Methods

3.3.1 Materials

Restriction enzymes and polymerases were purchased from NEB. Common reagents were purchased from Sigma-Aldrich. DNA oligos were purchased from IDT. Wat11 and pYeDP60 were kind gifts from Clint Chapple's lab in the department of biochemistry at Purdue.

3.3.2 Solutions and Media

As described in Table 17 and Table 18.

Table 17 All solutions are filter sterilized before use and stored at 20°C.

Solution	**Composition**
Tris Assay Buffer	20% Glycerol, 50mM Tris-HCl, 4mM EDTA (pH 7.4 @25°C)
High Salt Tris Assay Buffer	20% Glycerol,150mM NaCl, 50mM Tris-HCl, 4mM EDTA (pH 7.4 @25°C)
1M LiAc	1 M Lithium Acetate
50%PEG	50% PEG 3350 v/v
LiAc/PEG	0.1M Lithium Acetate, 40% PEG 3350
100x HLW	0.4g Histidine, 0.6g Leucine, 0.4g Tryptophan, H_2O to 100ml.

Table 18 Media used in study

Media	Composition
YPAD	Yeast Extract 10g/l, Peptone 20g/l, Adenine 30mg/l, Dextrose 20g/l; autoclaved
SGI	Glucose 20g/l, bactocasammino acids 5g/l, Yeast nitrogen Base without amino acids 3.4 g/l, L-tryprophan 40mg/l; autocalved
SLI	Galactose 20g/l, bactocasammino acids 5g/l, Yeast nitrogen Base without amino acids 3.4 g/l, L-tryprophan 40mg/l; autocalved
Yeast Minimal	Yeast nitrogen base without amino acids 6.7g/l, glucose 20g/l

3.3.3 Linearization of pYeDP60u Cloning Vector

5µg pYEDP60u vector was digested in a 200µl reaction containing 1x NEB buffer 1, 1x BSA and 4µl PacI. Reactions were incubated at 37C for 4 hours, then purified with Qiagen MinElute PCR purification kit (Column purified) and eluted with 50µl EB buffer. 45µl of the resulting pYEDP60-U PacI digested DNA was digested in a subsequent 200µl reaction containing 1x NEB EcoRI buffer, 1xBSA, 80U BamHI and 80U EcoRI. The digest was incubated at 37C for 4 hours, then column purified and eluted with 50µl EB buffer, yielding prepared cloning vector

3.3.4 High Efficiency Yeast Transformation

This protocol was adapted from *Molecular Genomics of Yeast,* 1994 and Clint Chapple's Lab, Department of Biochemistry, Purdue. Conditions were optimized in part by Corinne P. Price, Department of Biological Sciences, Purdue. A single Wat11 colony less than 1 month old was inoculated into 5ml YPAD media and grown for 24 hours at 30°C with shaking. 3µl were transferred into 10ml YPAD media and grown overnight at 30°C with shaking. In the morning, the overnight Wat11 culture was pelleted by centrifugation at 1000 rcf for 5 minutes. The pellet was resuspended in 10ml fresh 0.1M LiAc, and immediately centrifuged at 1000 rcf for 5 minutes. The pellet was resuspended in 1.6ml fresh 0.1M LiAc, yielding competent Wat11 cells.

For each transformation reaction, 50µl competent Wat11 cells were added to a 1.5ml microfuge tube containing 8µl (1.6µg-4µg) transforming DNA, 2µl carrier DNA and 350µl LiAc/PEG solution. Microfuge tubes were briefly, and gently, vortexed to mix contents, then incubated at 30°C for 30 minutes. Wat11 cells were heat shocked in a 42°C heat bath for 15minutes. Heat shocked cells were pelleted by centrifugation in a microfuge at 10,000 rcf for 5 seconds. Cell pellets were resuspended in 100µl sterile 200x HLW stock and plated on Minimal Yeast Plates. Plates were incubated at 30°C until the appearance of visible colonies (3-4 days).

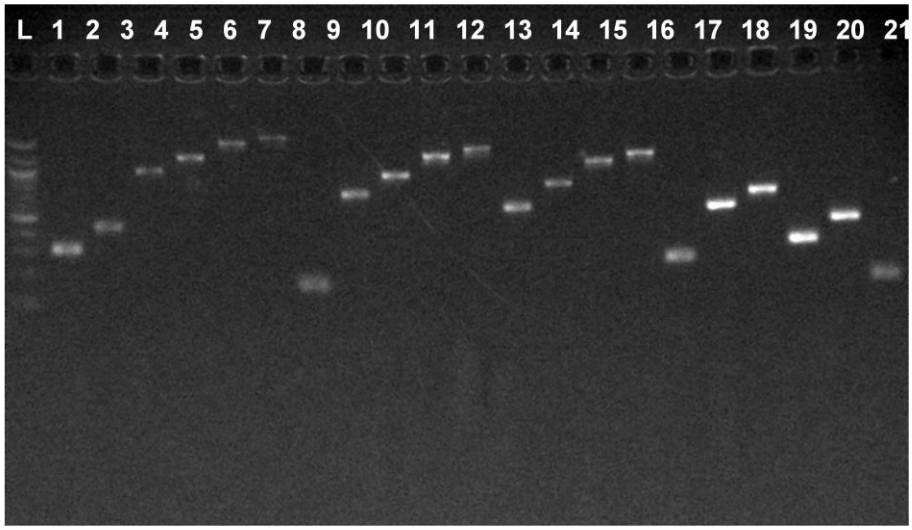
3.3.5 Preparation of Gene Fragments

1ug of Target DNA containing either AtC4H N/C optimized or SmC4H N/C optimized in pYEDP60u shuttle vector was linearized with 10U EcoRI in 1x NEB EcoRI buffer in a 50µl reaction for 16 hours at 37C. Digested plasmids were column purified and diluted in TE to 1ng/µl. Gene fragments were generated by using all possible upstream and downstream primer pairs in 42 separate reactions, yielding 21 fragments for each AtC4H and AtF5H (Table 4). Gene fragments were amplified in individual 50µl PCR reactions containing 0.2 mM DNTP mix, 0.2 µM upstream primer, 0.2 µM downstream primers, 1ng template DNA, 1U Phusion DNA polymerase in 1x Phusion HF buffer. The PCR protocol was: 95C for 2 minutes, followed by 25 cycles of 72C for 30 seconds, 54C for 30 seconds and 72C for 2 minutes. After the last cycle, samples were held at 72C for 3 minutes.

Amplified gene fragments were immediately gel purified in a large format 0.9% agarose gel prestained with EtBr, and run at 80V for 120 minutes. Orange band indicating EtBr stain for all gene fragments was visible in ambient light, indicating high yield. Viewing the gel under longwave UV while does not reveal minor bands or significant smearing (Figure 14). Bands were excised manually under longwave UV light.

Excised bands were transferred to sterile 1.5ml microfuge tubes and stored at 4°C until further processing. Gene fragments were extracted from excised gel bands using a Qiagen Gel Extraction kit. Each gel band was dissolved in 400µl QG buffer, washed with 0.5ml QG buffer, incubated with 0.75ml PE buffer for 5 minutes and eluted with 50µl sterile EB buffer after 5 minute incubation. Eluted fragments were stored at -20°C until further processing.

Gene fragments were digested with DpnI to remove potential remaining template plasmid: 1µl 10xNEB buffer 4 and 20U DpnI is added to each 50µl solution of purified gene fragment. Digests were incubated at 37C for 1 hour, followed by 80C for 20 minutes. 20µl of each digested, gel extracted sub fragment was diluted with 80µl sterile H_2O, yielding prepared gene fragments.



Lane	Gene Fragment	Lane	Gene Fragment	Lane	Gene Fragment
L	NEB 2-log ladder	8	8	16	4
1	1	9	13	17	10
2	7	10	17	18	15
3	12	11	20	19	5
4	16	12	3	20	11
5	19	13	9	21	6
6	21	14	14		
7	2	15	18		

Figure 14 Representative agarose gel of prepared gene fragments. This picture contains all 21 AtF5H gene fragments. 2.0% Agarose 0.5x TBE run for 90 minutes at 80 volts.

3.3.6 Optimization of OLE PCR Reaction Conditions

To reduce background of incorrect chimeras during OLE-PCR reaction, chimeras 2, 37, 59, and their associated dropout controls (see table 6) were amplified in parallel under the following conditions: 1:1, 1:5 fold dilutions of each prepared gene fragment was amplified for 15, 20 and 25 cycles for a total of 6 different reaction conditions. Aside from changes in cycle number, amplification conditions were identical to the conditions used to amplify individual gene fragments (Section 3.3.5). Conditions with the highest ratio of amplified full length chimeras to dropout controls were selected as final reaction conditions for generation of AtC4H/AtF5H chimeras. The final selected amplification conditions were 20 PCR cycles using gene fragments diluted five-fold.

This optimization was done as a single, self-contained experiment. For the optimization conditions, three chimeras were selected to represent the complexity of OLE-PCR reactions across the entire library. Chimera 2 consisting of 2 gene fragments, chimera 37 consisting of three gene fragments, and chimera 59 consisting of 5 gene fragments were selected. Each chimera is associated with two dropout controls. One is missing the upstream gene fragment and the second is missing a single internal gene fragment. For chimera 2, constructed from only two gene fragments, the second dropout control is missing the downstream gene fragment. The set of 9 PCR reactions were amplified for 15, 20 or 25 PCR cycles and 1-fold or 5-fold dilution of gene

fragments. Two PCE conditions show successful amplification of chimeras and no amplification of dropout controls. 15 cycles with undiluted gene fragments and 20 cycles with 5-fold diluted gene fragments. 20 cycles with 5-fold diluted gene fragments was selected as final amplification conditions for efficient use of gene fragments (Figure 15).

		No Dilution of Gene Fragments									5 Fold Dilution of Gene Fragments								
		Chimera 2			Chimera 37			Chimera 59			Chimera 2			Chimera 37			Chimera 59		
Gel Lane		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Gene		A1		A1	A7		A7	A7	A7		A1		A1	A7		A7	A7	A7	
Fragments		S20	S20		S9	S9		S3	S3	S3	S20	S20		S9	S9		S3	S3	S3
					A11	A11	A11	A4		A4				A11	A11	A11	A4		A4
								S5	S5	S5							S5	S5	S5
								A6	A6	A6							A6	A6	A6
	25	++			++	+		++	+	+	++	+	+	++	+	+	++	+	+
PCR Cycles	15	++			+			++			+			+					
	20	++			++			++	+		++			++			+		

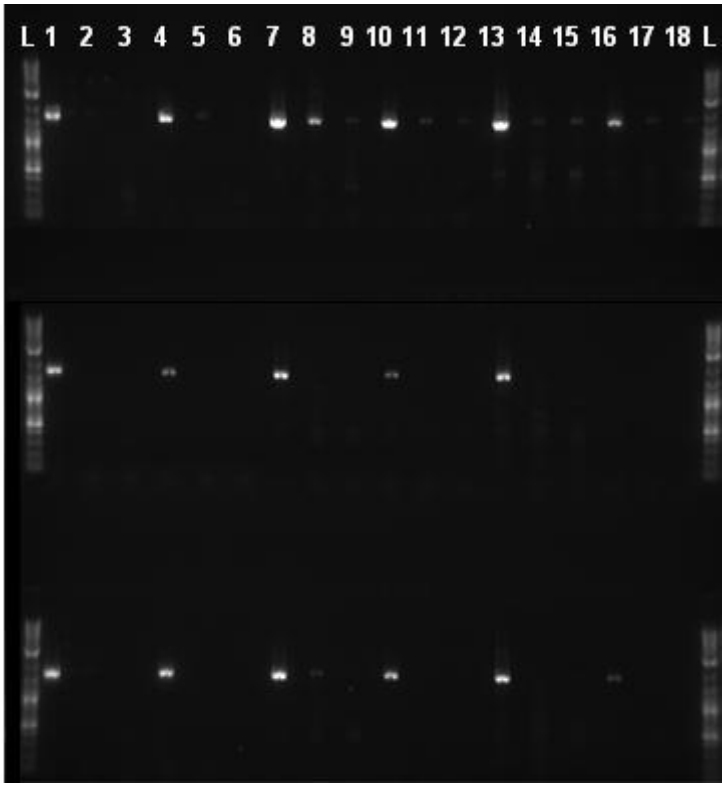


Figure 15 Previous Page, Figure of agarose gel of PCR “dropout” controls. ++ indicate strong PCR product, + indicates weak PCR product. No mark indicates no PCR product for a given reaction. Blue Boxes indicate reaction conditions favorable to amplification of chimera and disfavorable to amplification of “dropout” controls. Twenty PCR cycles and fivefold dilution of gene fragments was selected as final amplification conditions. This page, agarose gel photograph of PCR of “dropout” controls corresponding to previous page. 1.5% Agarose gel 0.5xTBE 80V 120 minutes. Lane L is NEB 2-log ladder.

3.3.7 Generation and Sequencing of AtC4H-SmC4H Chimeras

Generation of AtC4H-SmC4H chimeras proceeded as described in sections 4.3.5 using the OLE-PCR primers listed in Table 19. The following procedural changes were also made: The OLE-PCR was performed with an annealing temperature of 70C with 22 cycles using 5µl of 1:25 dilution for each gene fragment. Sequencing proceeded in the same manner as described in section 4.3.6 using the sequencing primers listed in Table 20.

Table 19 OLE PCR Oligos for the AtC4H-SmC4H library

Oligo Name	Sequence
SmAt282F.A2	5'- CGC CAA GGA GGT GCT CCT CAC TCA AGG C -3'
SmAt282R.A2	5'- GCC TTG AGT GAG GAG CAC CTC CTT GGC G -3'
SmAt324F.A2	5'- GCG GCG GAT GAG GAG AAT CAT GAC GGT TCC TTT CTT C -3'
SmAt324R.A2	5'- GAA GAA AGG AAC CGT CAT GAT TCT CCT CAT CCG CCG C -3'
SmAt559F.A2	5'- CTC AAA AGC AGG GAG AGA TCA ACG AGG ACA ATG TTC TTT ACA TC -3'
SmAt559R.A2	5'- GAT GTA AAG AAC ATT GTC CTC GTT GAT CTC TCC CTG CTT TTG AG -3'
SmAt559F.A2	5'- GCC GTG ATC AAG GAG ACG CTT CGT CTG AGA ATG GCG ATT C -3
SmAt559R.A2	5'- GAA TCG CCA TTC TCA GAC GAA GCG TCT CCT TGA TCA CGG C -3'
SmAt770F.A2	5'- GGT GAT TGG AAG CCT CGT CCA GAA CTT CGA GCT TCT TC -3'
SmAt770F.A2	5'- GAA GAA GCT CGA AGT TCT GGA CGA GGC TTC CAA TCA CC -3'
AtSm282F.A2	5'- CGG ATC TAA CAA AGG AAG TGC TCC ACA CGC AG -3'

Table 19 Continued

Oligo Name	Sequence
AtSm282R.A2	5'- CTG CGT GTG GAG CAC TTC CTT TGT TAG ATC CG -3'
AtSm324F.A2	5'- GAG CAT TGG AGG AAG ATG AGA AGG ATC ATG ACC GTC CCG -3'
AtSm324R.A2	5'- CGG GAC GGT CAT GAT CCT TCT CAT CTT CCT CCA ATG CTC -3'
AtSm559F.A2	5'- CTG AGC AGA AGG GAG AAA TCA CCG AAG CCA ACG TC -3'
AtSm559R.A2	5'- GAC GTT GGC TTC GGT GAT TTC TCC CTT CTG CTC AG -3'
AtSm648F.A2	5'- CAA GCT GTG GTT AAG GAG ACT TTC CGC TAC CAC ATG GC -3'
AtSm648R.A2	5'- GCC ATG TGG TAG CGG AAA GTC TCC TTA ACC ACA GCT TG -3'
AtSm770F.A2	5'- GAT CAC CAT TGG TAG GAT GGT TGC CAA GTT TGG GCT CTT G -3'
AtSm770R.A2	5'- CAA GAG CCC AAA CTT GGC AAC CAT CCT ACC AAT GGT GAT C -3'
SmC4H.UP	5'-GCCG CCG GAT CCG CTG A GG ATT A AT A ATG ATT AAT GTT GCT TC -3'
AtC4H.UP	5'-GCCG CCG GAT CCG CTG A GG ATT A AT A ATG GAT TTG TTA TTG TTA G-3'
SmC4H.DN	5’-GCC GCG GAA TTC GGG TTA AT TTA CAA AAC TCT TGG CTT C-3’
AtC4H.DN	5'-GCC GCG GAA TTC GGG TTA AT TTA ACA ATT TCT TGG TTT CAT AAC G-3'

Table 20 Sequencing primers for the AtC4H-SmC4H chimeric library. These primers were designed by Larisa Avramova.

Oligo Name	Sequence
P50	5'— CCK TTC TTC ACC AAC A -3'
P53	5'— GGA GGT TCR TGT GWG -3'
P70	5'— TGT TGG TGA AGA AMG G -3'
P71	5'— CWC ACA YGA ACC TCC -3'
P79	5'— CGT GTA TAT AGC GTG GAT GGC CAG -3'
P85	5'— AAT TCA ATT CAA TTT ATT TC -3'

3.3.8 *In Vivo* Activity Assay

A single Wat11 colony less than two weeks old carrying a pYeDP60u plasmid with insert was inoculated into 50ml SGI media and incubated for 24 hours at 30°C with shaking. Cells were pelleted by centrifugation at 1000rcf for 3 minutes, resuspended in 50ml SLI media, and then incubated for 4 hours at 30°C with shaking. For each in vivo assay, 250µl cells were spiked with substrate in DMSO for a final substrate concentration of 400µM. Spiked cultures were incubated at 30°C for one hour with shaking. Reactions were stopped at various time points by pelleting cells and decanting the spent media for analysis.

3.3.9 *In Vivo* Activity Analysis

Spent media from in vivo activity assays were analyzed in 96 well plate on an Agilent 1100 LC/MS fitted with a Shimadzu Shim-pack XDR-ODS 75Lx3.0 column. Products in spent media from assays spiked with CA were separated on a 59.95% H_2O: 39.95% ACN: 0.10% Formic Acid isocratic gradient. Products in spent media from assays spiked with coniferyl alcohol were separated on a 84.95%H_2O:14.95% ACN: 0.10% Formic Acid to 64.95%H_2O:34.95% ACN:0.10% Formic Acid gradient over 7.5 column volumes.

LC/MS software automatically calculated product peak area based on retention time. pCA product consistently eluted at 3.360±0.010 minutes on the isocratic separation. 5-hydroxy coniferyl alcohol product consistently eluted at 5.24±0.010

minutes. Data was exported into .csv file and imported into R 2.14.0 and analyzed using the standard packages.

3.3.10 High Throughput *In Vivo* Activity Assay

Protocol Developed in part by Larisa Avramova. A single wat11 colony less than two weeks old carrying a pYEDP60u plasmid with a chimeric insert was inoculated into 1ml SGI media and incubated for 24hours at 30°C with shaking. This culture is used to inoculate 1ml SGI to 0.1 OD_{550} and grown for 24 hours at 30°C with shaking. Cells were pelleted by centrifugation at 1000rcf for 5 minutes and resuspended in 1ml SLI media. Cells were then incubated for 4 hours at 30°C with shaking to induce gene expression. This induced culture is used to inoculate two wells containing 0.5ml SLI to 0.5 OD_{550}. This measured cell density immediately prior to assay was used as a covariate in linear regression models (section 3.2.5). Each well is a single assay and is spiked with either 200µM Coniferyl Alcohol or 200µM CA. The second well was a control well and was not spiked with any substrate. Cultures spiked with CA were incubated for 1 hour at 30°C with shaking, while cultures spiked with Coniferyl Alcohol were incubated for 4 hours at 30°C with shaking. Control wells were incubated under the same conditions as their corresponding assay well. Reactions were stopped by pelleting cells at 1000 rcf for 5 minutes. 160µl of supernatant was mixed with 40µl MeOH and 4µl 50mM Ascorbic Acid. Samples were stored at 4°C until further analysis. Multiple cultures were handled in parallel using 96 well deep well blocks and Beckman Coulter BioMek NXP workstation for all culture transfers.

3.3.11 Induction Timing of pYeDP60u in Wat11

Individual yeast colonies were inoculated into 7ml modified SGI media. Twelve 500µl aliquots of each yeast inoculate were each transferred into a 96 well deep well block, and incubated at 30°C and 350rpm for 24-30 hours. Cells were diluted to 0.1 OD_{550} in 500µl fresh SGI media and incubated at 30°C and 350rpm for 24 hours. Cells were pelleted at 1000 rcf for 5 minutes, decanted, and then resuspended in 500µl SLI media. At every time point, one well of induced cells was used to inoculate 500µl SLI media at 0.5 OD_{550}, spiked with substrate and incubated at 30C and 350rpm for 2 or 4 hours. Reactions were stopped by centrifugation at 17,000 rcf for 2 minutes. 160µl of supernatant was mixed with 40µl MeOH and 4µl 50mM Ascorbic Acid. Samples were stored at 4°C until further analysis on HPLC as described for In vivo activity analysis (section 3.3.9). Induction time of four hours after transfer of cells to SLI media was selected for all future *in vivo* assays.

CHAPTER 4. GENE FRAGMENT INTERCHANGEABILITY BETWEEN CINNAMATE 4-HYDROXYLASE AND FERULIC ACID 5-HYDROXYLASE PROTEINS FROM *A. THALIANA*

4.1 Introduction

Combinatorial interchange of large gene fragments among functionally identical, homologous sequences is an established tool in protein engineering for improving protein function, altering substrate recognition, and identifying functional constraints among related sequences (59,66)(section 3.2.5).

However, directed combinatorial interchange of multiple large gene fragments among functionally divergent proteins has not been explored. Stochastic construction techniques have been used to randomly cross over functionally divergent proteins. These techniques produce great numbers of non-functional proteins. Identifying the few functional proteins in stochastically constructed chimeric libraries requires powerful screening techniques, not available for all proteins.

Combining *A. thaliana* Ferulic acid 5-hydroxylase (AtF5H) with AtC4H will test the limits of interchangeability among functionally divergent, homologous proteins.

Like AtC4H, AtF5H is a membrane bound type IV cytochrome P450 monooxygenase in the phenylpropanoid pathway. F5H diverged from C4H about 400 million years ago and catalyzes hydroxylation of the 5-ring position of coniferyl alcohol. F5H is structurally similar to C4H, as well as catalyzing the hydroxylation on a similar substrate.

An AtC4H-AtF5H chimeric library consisting of gene fragments analogous to those for the AtC4H-SmC4H chimeric library presented in CHAPTER 3 are non-functional. This weakens the hypothesis that functionally divergent proteins are easily amenable to interchange of large gene fragments. However, a library of 10 AtC4H-AtF5H chimeras consisting of replacement of small gene fragments does contain functional proteins. The functional proteins do not swap residues inferred to be catalytically important. This finding supports the hypothesis that exchanging gene fragments among functionally diverse proteins is more disruptive than exchanging gene fragments among functionally identical proteins. Nevertheless, exchange of gene fragments among functionally divergent genes is possible on a small scale. This finding suggests that functionally divergent proteins are a rich source for highly divergent sequences. Introduction of short sequences from functionally divergent proteins may allow rapid diversification of enzyme function in future protein engineering experiments.

4.1.1 Selection of AtC4H and AtF5H as Parental Genes

Ferulic acid 5-hydroxylase (F5H; EC 1.14.-.-) is a P450 of the CYP84A1 protein family which catalyzes the hydroxylation of coniferyl alcohol at the 5 ring position into 5-hydroxy coniferyl alcohol. *A. thaliana* F5H (AtF5H) has a calculated molecular weight of 58.7kDa (520 amino acid residues).

Both C4H and F5H proteins share the P450 fold, which has strong structural conservation across families, despite low sequence conservation (67). C4H and F5H are both type IV P450s. AtF5H SRS have been inferred by alignment to C4H proteins (Section 3.1.2). The inferred SRS show high conservation within, but not between, the CYP73 and CYP84A1 protein families. Molecular modeling has shown a high degree of similarity among architecture of the catalytic site between AtC4H and AtF5H, including predicted substrate orientation and location of SRS residues in contact with the substrate (68,69,45).

C4H and F5H proteins recognize very similar substrates. CA differs from Coniferyl Alcohol by hydroxylation at the 4 ring position and methoxy at the 3 ring position. AtC4H will not recognize substrates with substitutions larger than methyl at the 3 ring position (44). AtC4H is also very specific for substrates with a terminal acid (43), whereas AtF5H will recognize substrates with a terminal alcohol or aldehyde, and much less efficiently, a terminal acid (70).

AtC4H and F5H proteins arose around 125 million years ago with syringyl monolignin precursor in the angiosperm lineage. AtF5H and AtC4H share 29% sequence identity.

4.1.2 Multiple Sequence Alignment of P450 Proteins

Selection and alignment of protein sequences is as described in section 3.1.2

4.1.3 Library Design

The breakpoints described in section 3.1.3 were derived using a MSA including F5H proteins. Therefore, the same breakpoints will be used for AtC4H-AtF5H chimeric library as used for the AtC4H-SmC4H library. This will also allow direct comparison of exchangeable gene fragments.

The nomenclature used to refer to chimeras in the AtC4H-AtF5H chimeric library is analogous to the nomenclature used to refer to chimeras in the AtC4H-SmC4H library (Section 3.1.3), with the addition that F is used to refer to gene fragments from AtF5H. For example, reconstructed AtF5H is written as FFFFFF. The chimera containing gene fragments 1, 2 and 3 from AtC4H and gene fragments 4, 5 and 6 from AtF5H is written as AAAFFF.

4.2 Results and Discussion

Failure of AtC4H-AtF5H library suggests against general interchangeability of gene fragments on a large scale limited success of protein subchimeras indicates much more limited interchangeability exists between AtC4H and AtF5H.

4.2.1 Codon Optimization of AtF5H

Wild-type AtF5H has low *in vivo* activity levels. We attempted to improve expression of AtF5H in Wat11 through codon optimization. Three codon optimized AtF5H variants were constructed and tested for activity: The 12 N-terminal and 5-C terminal codons are optimized for expression in *S. cerevisiae (sp)* (AtF5H N12/C5); A sequence of four Arg codons after the membrane insertion sequence (AtF5H Arg) is codon optimized for expression in *S. cerevisiae (sp)*; All codons in wild-type AtF5H are altered to the most commonly observed codons in *S. cerevisiae (sp)* (AtF5H synth). Codons, or combinations of codons that introduced one or more restriction site used for cloning were avoided. In these cases, the next most common Wat11 codon was used.

It was hypothesized that these arginine residues may play a critical role in proper membrane association of AtF5H during translation (Clint Chapple, personal communication). The primers used in the construction of AtF5H N12/C5 and AtF5H Arg are listed in Table 25. The AtF5H Arg variant was constructed by Samuel Schaffter. In AtF5H Synth all codons have been optimized for expression in Wat11.

Testing by CO difference spectroscopy showed that both AtF5H Arg and AtF5H N12/C5 successfully increased expression levels, but only AtF5H Arg has higher levels of in vivo activity than wild-type AtF5H (Figure 17). Both codon variants are associated with high levels of misfolded protein (Figure 16). Correctly folded P450 proteins are associated with an absorbance peak at 450nm, while misfolded P450 proteins still bound to heme ligand are associated with an absorbance peak at 420nm (71). Misfolded P450 proteins not bound to heme are not seen in this assay. AtF5H Synth shows no measureable levels of in vivo activity or folded protein, implying lack of synthesis, inability to bind heme during translation, and/or rapid degradation by the host cell.

The failure of codon optimization to increase activity levels of AtF5H while also minimizing accumulation of misfolded protein may be due to the observation that codons must be optimized in pairs (72). Rare codons are necessary for proper protein folding, causing translational pauses which allow folding of domains or subdomains (73). Due to time constraints, it was not feasible to redesign and test codon pair optimized AtF5H proteins. To avoid the possibility of deleterious codon pairs at cross over points in protein chimeras, it was decided that wild-type AtF5H would be paired with AtC4H N/C as parental sequences for the chimeric library.

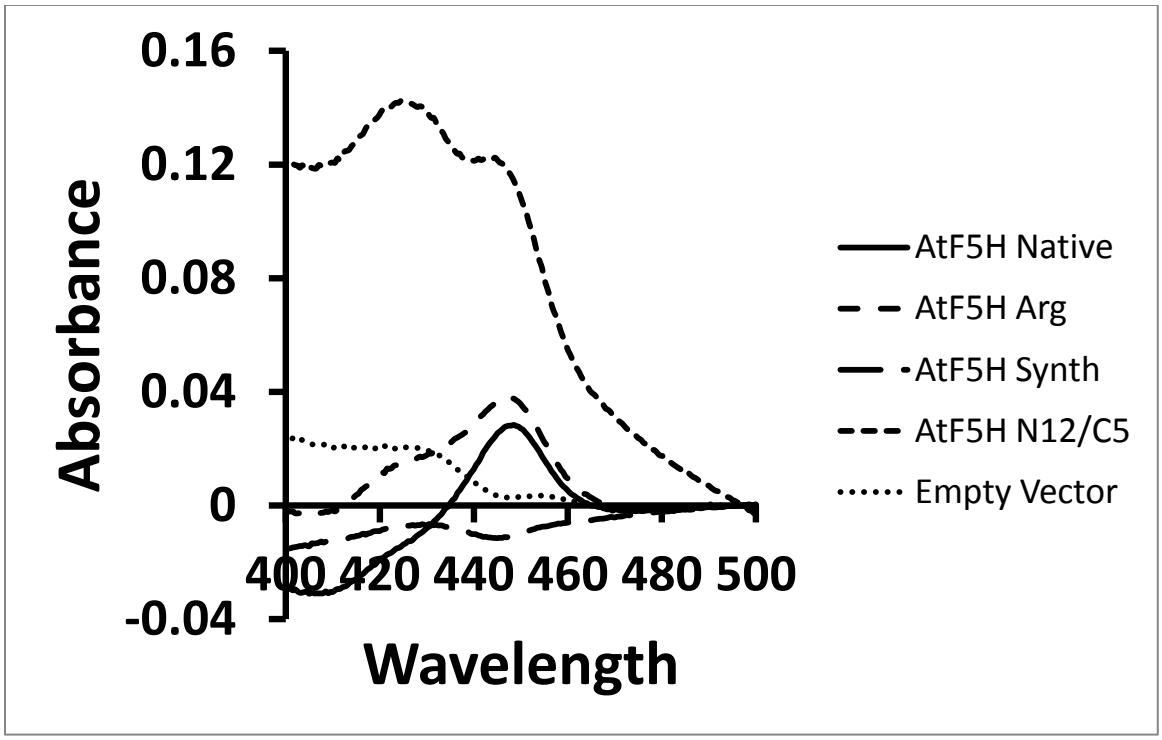


Figure 16 P450 CO difference spectra of codon optimized AtF5H variants.

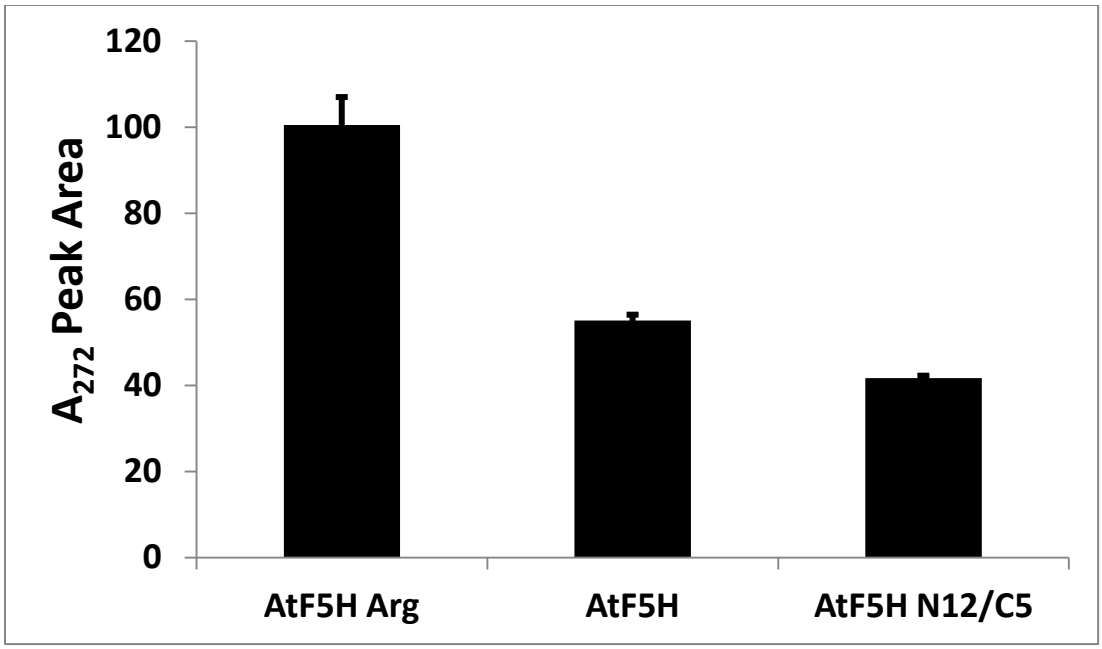


Figure 17 Average *in vivo* activity of codon optimized AtF5H variants tested against Coniferyl alcohol. Error bars are standard error of mean 5-hydroxy coniferyl alcohol production as measured by A_{272} peak area from four replicates. *In vivo* activity assay performed by Samuel Schaffter.

4.2.2 Recovery of the AtC4H-AtF5H Chimeric Library

Efficiency of constructing AtC4H-AtF5H chimeras with out nucleotide errors was 60%; the same as reported for the AtC4H-SmC4H chimeric library (section 3.2.3). A complete AtC4H-AtF5H chiemric library has been recovered. Sequences have been confirmed by sequencing both strands of every chimera at least once.

4.2.3 AtC4H-AtF5H Chimeric Library Analysis

All chimeras from the AtC4H-AtF5H chimeric library were tested for activity against CA and coniferyl alcohol. As expected, reconstructed AtC4H has wild-type activity against CA and no measurable activity against coniferyl alcohol. Similarly, reconstructed AtF5H shows wild-type activity against coniferyl alcohol and no measureable activity against CA. Surprisingly, only one chimera has any measurable activity against either CA or coniferyl alcohol. FAAAAA has activity against CA at 120% of wild-type activity. FAAAAA has no measurable activity against coniferyl alcohol (Table 21).

This apparent lack of combinatorial interchangeability between AtF5H and AtC4H was tested through a set of 10 protein sub-chimeras. Gene fragment 4, believed to be involved in orienting the ring portion of the substrate to the heme domain during catalysis (45), was sub-divided into 5 smaller region (Figure 18). The first sub-fragment is 5 amino acid residues long with 3 polymorphisms. It immediately follows the third breakpoint and is in a highly conserved region of the MSA, implying structural and functional conservation. The second sub-fragment is 19

amino acid residues long and comprises an SRS motif with 13 polymorphisms. This region is conserved within P450 families, but diverse between different P450 families. The third and fourth sub-fragments are 23 and 17 amino acid residues long, respectively. Sub-fragment 3 has 15 polymorphisms. Sub-fragment 4 has 9 polymorphisms and 2 gaps; the only sub-fragment with gaps in the sequence alignment between AtC4H and AtF5H. This region of the MSA shows greater diversity than the other gene sub-fragments. The fifth gene sub-fragment comprises the 9 amino acid residues immediately upstream of the fourth breakpoint. This region of the MSA is well aligned and contains 4 polymorphic amino acid residues.

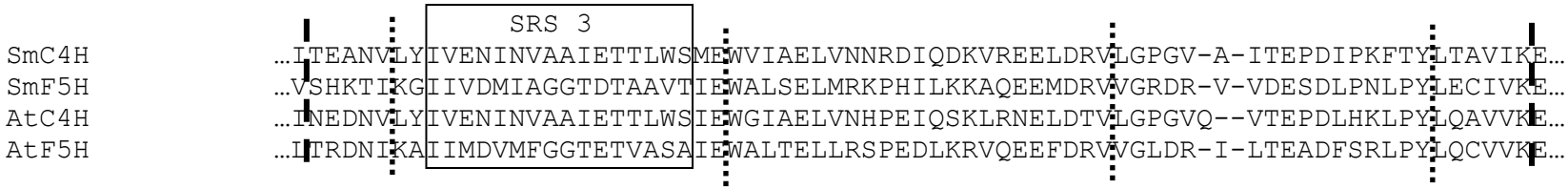
Single fragment replacements of AtC4H-AtF5H sub-chimeras were constructed, and tested for activity on CA and coniferyl alcohol. Only the four sub-chimeras containing recombinant fragments immediately proximate to the breakpoints were observed to be active. $AAA_{FAAAA}AA$ and $AAA_{AAAAF}AA$ are active on CA, whereas $FFF_{FFFFA}FF$ and $FFF_{AFFFF}FF$ are active on coniferyl alcohol. This finding supports the hypothesis that core structural constraints have been preserved between C4H and F5H proteins, whereas specific catalytic constraints have diverged beyond the point of interchangeability.

Since current homology models suggest that the C-terminal half of the C4H and F5H proteins orient the ring portion of the substrate during catalysis, this suggests that gene fragments 1, 2 and 3 are involved in recognition of the tail portion of CA and coniferyl alcohol. AtC4H has a strong preference for

substrates with a carboxylic acid moiety at the tail, whereas AtF5H has highest catalytic activity against substrates with either an alcohol or aldehyde moiety. It follows that xxxCCC chimeras can be tested for activity against cinnamic alchol or cinnamic aldehyde to determine if structure-function-sequence constraints for distal portions of the substrate are independent of structure-function-sequence constraints for recognition of the substrate proximal to the site of catalysis.

Future chimeric libraries between AtC4H and AtF5H can take smaller steps (i.e. smaller gene fragments) away from the parental sequences and more carefully probe the functional sequence space between these two proteins. Alternately, construction of an AtF5H-SmF5H chimeric library would test the importance of functional constraints over evolutionary history. The failure of the AtC4H-AtF5H library is from either a lack of functional constraints, low sequence identity, or both. SmF5H (CYP788A1) and AtF5H have a similar sequence identity to AtF5H and AtC4H, at 31% and 29%, respectively. However, SmF5H obtained F5H activity through convergent evolution towards other plant F5H proteins (48). If AtF5H-SmF5H chimeras were functional, this would greatly strengthen the hypothesis that functional constraints are most important to defining protein activity.

A



B

Subfragment	1	2	3	4	5
Length (amino acids)	5	20	23	17	6
Polymorphisms	3	13	15	9	1
gaps	0	0	0	2	0

Figure 18 Alignment and properties of AtC4H-AtF5H subchimeras

Table 21 Activity of AtC4H-AtF5H chimeras on CA and Conif OH. AtC4H-AtF5H Chimeras 2-62 are inactive on both CA and Coniferyl alcohol and have been omitted from this table for clarilty. AFAx and FAFx chimeras were constructed and tested for activity by Corinne P. Price, Biological Sciences, Purdue.

Number	Cimera	CA	Conif OH
1	$FAA_{AAAAA}AA$	++	-
63	$AAA_{AAAAA}AA$	+	-
AFA1	$AAA_{FAAAA}AA$	+	-
AFA2	$AAA_{AFAAA}AA$	-	-
AFA3	$AAA_{AAFAA}AA$	-	-
AFA4	$AAA_{AAAFA}AA$	-	-
AFA5	$AAA_{AAAAF}AA$	+	-
64	$FFF_{FFFFF}FF$	-	+
FAF1	$FFF_{AFFFF}FF$	-	+
FAF2	$FFF_{FAFFF}FF$	-	-
FAF3	$FFF_{FFAFF}FF$	-	-
FAF4	$FFF_{FFFAF}FF$	-	-
FAF5	$FFF_{FFFFA}FF$	-	+

4.3 Materials and Methods

4.3.1 CO Difference Spectra

Microsomes are diluted to about 1.0 OD_{450} in Tris Assay Buffer prior to analysis. One ml diluted microsomal fractions are aliquoted to a reference and sample cuvette. CO gas is bubbled through the sample cuvette for 1 minute. Cuvettes are placed in a Cary 4000 dual beam UV-Vis spectrophotometer and a baseline measurement is taken from 350 to 500nm. Approximately 1mg Sodium Dithionite is then added to each cuvette and allowed to react for one minute. Difference spectra are recorded from 350 to 500nm. Adapted from (71,74)

4.3.2 Construction of AtF5H Codon Optimized Variants

Purified pYeDP60u AtF5H was amplified with AtF5H.UP.12N and AtF5H.DN.5C to make AtF5H N12/C5, and amplified with AtF5H.PR.UP and AtF5H.PR.DN to make AtF5H Arg(Table 25). The amplified genes were column purified and digested with EcoRI and BamHI in 1xNEB EcoRI Buffer. Digested inserts are column purified eluted in 50µl EB buffer.

2µl of the column purified, digested inserts are ligated into 50ng linearized pYEDP60u vector with 0.25µl T4 DNA ligase in 1xNEB T4DNA ligase beffer. Total reaction volume is 10µl. Ligation reactions are incubated at 16°C for 16 hours. 1µl of ligation mixtures are used to transform competent DH5α cells by heat shock, and plasmids with inserts are selected for on LB Amp plates. Single isolated colonies are to the Purdue Geneomics Core and each strand is

sequenced at least once. AtF5H N12/C5 and AtF5H Arg are recovered and confirmed correct at each nucleotide.

The complete codon optimized sequence was calculated by Chris Bailey-Kellogg's lab at Dartmouth. The AtF5H Synth gene was synthesized by DNA2.0, and nucleotide sequence is confirmed by sequencing each strand at least once.

4.3.3 Preparation of Gene Fragments

Gene fragments for the AtC4H-AtF5H chimeric library were prepared as reported for the AtC4H-SmC4H chimeric library (section 3.3.5). PCR primers used for the AtC4H-AtF5H chimeric library are listed in Table 22

4.3.4 Optimization of OLE-PCR conditions

Optimization of primer concentration, cycle number, and annealing temperature for OLE-PCR reaction used to amplify all AtC4H-AtF5H chimeras was carried out in the same manner as for the AtC4H-SmC4H chimeric library (section 3.3.6).

4.3.5 Generation of AtC4H-AtF5H Chimeras

Each chimera is produced in a separate PCR reaction with the following conditions: 5µl of each prepared gene fragment required for the individual chimera, 0.2 mM DNTP mix 0.2 µM upstream primer, 0.2 µM downstream primers, 1ng template DNA, 1U Phusion DNA polymerase in 1x Phusion HF buffer and enough H_2O to bring the reaction volume to 50µl. Upstream and

downstream primers are matched to the N and C terminal identity of each chimera. OLE-PCR protocol is identical to amplification of individual gene fragments. The OLE-PCR is assembled on a Beckman Coulter BioMek FX workstation.

OLE-PCR chimeric gene fragments visualized on agarose gel shows a single high yield band of the correct length for each chimera. Based upon the gel, each PCR reaction is estimated to contain up to 2µg of band of interest. All PCR products are column purified, then digested with EcoRI and BamHI in a 100µl reaction containing 1x NEB EcoRI buffer, 1x BSA, 40U EcoRI and 40U BamHI. Digests are incubated for 4 hours at 37C, then column purified and stored at -20°C until further processing.

1µl of each digested, column purified OLE-PCR chimeras is added to a 10µl ligation reaction containing 25ng linearized YeDP60u cloning vector and 0.25µl T4 DNA Ligase. Ligation reactions are incubated at 16C for 16 hours in a thermocycler, and stored at -20°C. Approximately 0.1-0.25µl (P10 set to 0.0µl) of each ligation reaction is used to transform 16µl Invitrogen Library efficiency DH5alpha competent cells. Timing and buffers follow manufactures instructions, scaled down to accommodate 16µl competent cells. Transformants are selected for on LB-Amp plates. Each transformation reaction yielded >100 colonies with no detectable background (linearized vector incubated with T4 DNA ligase at 16C for 16 hours without insert did not yield and amp^R *E.coli* colonies; data not shown).

Table 22 OLE PCR primers for AtC4H-AtF5H chimeric library.

Oligo Name	Sequence
N.C4H.F5H.T.1	5'- CAC CGG ATC TAA CAA AGG AAG TGC TTC AAG TCC AAG ACA GCG TC -3'
N.C4H.F5H.B.1	5'- GAC GCT GTC TTG GAC TTG AAG CAC TTC CTT TGT TAG ATC CGG TG -3'
N.F5H.C4H.T.1	5'- GAG GTG GCT CGA CAA GTC CTC CTC ACT CAA GGC GTT G -3'
N.F5H.C4H.B.1	5'- CAA CGC CTT GAG TGA GGA GGA CTT GTC GAG CCA CCT C -3'
N.C4H.F5H.T.2	5'- GAG CAT TGG AGG AAG ATG AGA AAA GTG TGT GTC ATG AAG GTG TTT AG -3'
N.C4H.F5H.B.2	5'- CTA AAC ACC TTC ATG ACA CAC ACT TTT CTC ATC TTC CTC CAA TGC TC -3'
N.F5H.C4H.T.2	5'- CCG TTT TGG AGA CAG ATG AGA AGA ATC ATG ACG GTT CCT TTC TTC -3'
N.F5H.C4H.B.2	5'- GAA GAA AGG AAC CGT CAT GAT TCT TCT CAT CTG TCT CCA AAA CGG -3'
N.C4H.F5H.T.3	5'- GCT GAG CAG AAG GGA GAA ATC ACC CGT GAC AAT ATC AAA GCA ATC -3'
N.C4H.F5H.B.3	5'- GAT TGC TTT GAT ATT GTC ACG GGT GAT TTC TCC CTT CTG CTC AGC -3'
N.F5H.C4H.T.3	5'- CGG ATC TTC AAA ATT CCA TCA AAC TTA ACG AGG ACA ATG TTC TTT ACA TCG -3'

Table 22 Continued

Oligo Name	Sequence
N.F5H.C4H.B.3	5'- CGA TGT AAA GAA CAT TGT CCT CGT TAA GTT TGA TGG AAT TTT GAA GAT CCG -3'
N.C4H.F5H.T.4	5'- CAA GCT GTG GTT AAG GAG ACT CTA AGG ATG CAC CCA CCG -3'
N.C4H.F5H.B.4	5'- CGG TGG GTG CAT CCT TAG AGT CTC CTT AAC CAC AGC TTG -3'
N.F5H.C4H.T.4	5'- CTC AAA TGC ACA CTC AAA GAA ACC CTT CGT CTG AGA ATG GCG ATT C -3'
N.F5H.C4H.B.4	5'- GAA TCG CCA TTC TCA GAC GAA GGG TTT CTT TGA GTG TGC ATT TGA G -3'
N.C4H.F5H.T.5	5'- GGA TCA CCA TTG GTA GGA TGT TAC ATT GCT TCA CGT GGA AAT TAC C -3'
N.C4H.F5H.B.5	5'- GGT AAT TTC CAC GTG AAG CAA TGT AAC ATC CTA CCA ATG GTG ATC C -3'
A/F.UP.B	5'-GCC GCC AGA TCT GCT GAG GAT TAA TAA TGG-3'
N.F5H.C4H.B.5	5'- GAA GAA GCT CGA AGT TCT GGA CTA TAT GAG CCA CGG CTA AGT CAA G -3'
AtC4H.NC.DN	5'-GGC CGC GAA TTC GCT GAG GGT TAA ATT AAC AAT TTC-3'

Table 23 Primers used in construction of AtC4H-AtF5H subchimeras

Oligo Name	Sequence
AFA.4A.UP	5'- gAT gTA AAg gAT ATT gTC ACg ggT gAT TTC TCC CTT CTg CTC AgC -3'
AFA.4A.DN	5'- ACC CgT gAC AAT ATC CTT TAC ATC gTC gAg AAC ATC AAT gTC -3'
AFA.4B.UP	5'- CgT TCC TCC AAA CAT AAC gTC CAT gAT gAT TgC TTT AAC ATT gTC CTC gTT gAT TTC TCC -3'
AFA.4B.DN	5'- ggA CgT TAT gTT Tgg Agg AAC ggA AAC ggT AgC gTC ggC gAT AgA gTg ggg AAT TgC AgA gCT Ag -3'
AFA.4C.UP	5'- gAg TTC TTg TTg gAC CCg TTT TAg ATC CTC ggg gCT CCg TAA TAA CTC CgT TAA ggC CCA CTC gAT AgA CCA CAA Tg -3'
AFA.4C.DN	5'- Cgg gTC CAA CAA gAA CTC gCC gAA gTC CTT ggA CCg ggT gTg C -3'
AFA.4D.UP	5'- CgA TgT Cgg ATT CTT CAA CTC gTC TgT CAA gTC CAA CAA CTg TgT CgA gTT CgT TCC -3'
AFA.4D.DN	5'- CgA gTT gAA gAA TCC gAC ATC gAg AAg TTg ACT TAT CTT CAA gCT gTg gTT AAg gAg AC -3'

Table 23 Continued

Oligo Name	Sequence
AFA.4E.UP	5’- ggT TTC TTT gAg TgT gCA TTT gAg gTA Tgg AAg TTT gTg AAg ATc Agg -3’
AFA.4E.DN	5’-CTC AAA TgC ACA CTC AAA gAA ACC CTT CgT CTg AgA ATg gCg ATT C-3’
FAF.4A.DN	5'- AAC ATT GTC CTC GTT AAG TTT GAT GGA ATT TTG AAG ATC CG -3'
FAF.4A.UP	5'- CC ATC AAA CT AAC GAG GAC AAT GTT AAA GCA ATC ATC ATG GAC GTT ATG TTT G -3'
FAF.4B.DN	5'- CAA TCG CGG CGA CAT TGA TGT TCT CGA CGA TGT AAA GGA TAT TGT CAC GGG TAA GTT TGA TGG -3'
FAF.4B.UP	5'- CAA TGT CGC CGC GAT TGA GAC AAC ATT GTG GTC TAT CGA GTG GGC CTT AAC GGA GTT ATT AC -3'
FAF.4C.DN	5'- GCT TAC TCT GGA TTT CAG GAT GGT TCA CTA GCT CTG CAA TTC CCC ACT CTA TCG CCG ACG CTA C -3'

Table 23 Continued

Oligo Name	Sequence
FAF.4C.UP	5'- CCA TCC TGA AAT CCA GAG TAA GCT AAG GAA CGA ACT CGA CAC AGT TGT TGG ACT TGA CAG ACG AGT TG -3'
FAF.4D.DN	5'- GAA GAT CAG GCT CGG TGA CTT GCA CAC CCG GTC CAA GGA CTT CGG CGA GTT CTT GTT G -3'
FAF.4D.UP	5'- GCA AGT CAC CGA GCC TGA TCT TCA CAA ACT TCC ATA CCT CAA ATG CAC ACT CAA AGA AAC C -3'
FAF.4E.DN	5'- AGT CTC CTT AAC CAC AGC TTG AAG ATA AGT CAA CTT CTC GAT GTC GG -3'
FAF.4E.UP	5'- CTT CAA GCT GTG GTT AAG GAG ACT CTA AGG ATG CAC CCA CCG -3'

4.3.6 Sequencing of AtC4H-AtF5H chimeras

One single isolated colony from each transformation plate is inoculated into 6ml LB-amp liquid culture and grown overnight at 37C with shaking. 0.5ml from each culture is used to prepare glycerol stocks, with the remaining culture used for plasmid purification (Qiagen miniprep). Both DNA strands on every insert is sequenced at least once. Sequencing and assembled contigs are completed by the Purdue Genomics Core(Table 24). Assembled contigs are compared against expected sequences (based upon the parental genes). Clones without any nucleotide defects are retained. In this manner, a complete set of AtC4H-AtF5H chimeras have been recovered.

Table 24 Sequencing primers for the AtC4H-AtF5H library

Oligo Name	Sequence
AtC4H.Nat.450.fwd	5'- CCA ACA AAG TTG TTC AAC AGA ATC GTG AAG G-3'
AtF5H.Nat.450.fwd	5'- GAA GGT GTT TAG CCG TAA AAG AGC TGA GTC -3'
AtC4H.Nat.550.rev	5'- CTC AAA TCT TCT ATC GAA CAT GAT ACG GAA CAT ATT GTT ATA C -3'
AtF5H.Nat.550.rev	5'- CTG CCC GGT AAG TTA TGT TGC GG -3'
AtC4H.Nat.950.fwd	5'- CGC GAT TGA GAC AAC ATT GTG GTC TAT C-3'
AtF5H.Nat.950.fwd	5'- CAA AGC AAT CAT CAT GGA CGT TAT GTT TGG-3'
AtC4H.Nat.1050.rev	5'- GCT TGA AGG TAT GGA AGT TTG TGA AGA TCA G -3'
AtF5H.Nat.1050.rev	5'- CTT CAA CTC GTC TGT CAA GTC CAA CG -3'
AtC4H.Nat.725.fwd	5'- GAC CAT TCC TCA GAG GCT ATT TGA AGA TTT GTC -3'
AtF5H.Nat.725.fwd	5'- GGC TCG TGA AGG CCC GTA ATG -3'
AtC4H.Nat.775.rev	5'- GCT TCA AGG ATG TGA TCA ATG GCA CAT TTC -3'
AtF5H.775.rev	5'- GCC TCT TCA CTG TAA AAA GCA AGA AGA TCA TC -3'

Table 25 PCR primers used to construct AtF5H N12/C5 and AtF5H Arg.

Oligo Name	Sequence
AtF5H.UP.12N	GCC GCC GGAT CCG CTG AGG ATT AAT A ATG GAA TCT TCT ATT TCT CAA ACT TTG ACT AAA TTA GAT CCC ACG ACG TC
AtF5H.DN.5C	GGC CGC GAA TTC GCT GAG GGT TAAA TTA CAA AGC ACA TAT GAG GCG CGT GGT TGG
AtF5H.PR.UP	AGA AGA AGA AGA AGG CCT CCA TAT CCT CC
AtF5H.PR.DN	TCT TCT TCT TCT TGT GAT GAA GCT GAT GAA GAT G

CHAPTER 5. RESIDUE PAIR ANALYSIS

5.1 Introduction

Extant protein sequences represent a sampling of evolutions' successful traverse through the functional space of a given protein. A common element of functionally related extant sequences is preservation of common elements through conserved sequence motifs (75). Examples include DNA binding motifs (76), catalytic centers (77,78), and signals for cellular localization (79,80). This observation suggests that a limited sequence space describes the functional space of a given protein.

This limited, related 'functional island' within sequence space is partly be an evolutionary artifact. Since homologous sequences evolve from a common ancestor, we expect modern extant sequences to be similar. However, many examples of convergent evolution weaken the suggestion that 'functional islands' are just an evolutionary artifact (81).

Functional space existing within an 'island' of sequence space has been exploited throughout molecular biology to identify functionally relevant protein

sequences: Hidden Markov Models have been used to identify functional motifs (82,83,84); highly conserved residues in protein MSAs are assumed to be functionally important (85); many conservation models exist to identify protein secondary structure (86,87,88).

Sequence conservation is routinely used to classify protein sequences into known structural domains, families, and functions (89,90), with recent algorithms incorporating secondary structure information (91), pairwise protein similarity (92) or statistical weighting of evolutionary information (93). Although identification of the most common amino acid residues for a given protein family is a highly effective way of classifying proteins, conservation of of individual amino acid residues fails to produce divergent protein sequences that maintain their biological function (94).

In contrast, it has been shown that conserving the identity of amino acid residue pairs enables the creation of divergent protein sequences that maintain their intended biological function (94). The interaction of residues as pairs within a protein has been well documented as hydrogen bonds, ionic bonds and hydrophobic interactions in the core of a protein (95,96,97,98). Therefore, it makes sense that maintaining pairs of residues within a novel protein sequence may be necessary to increase the likelihood of producing biologically functional proteins.

We hypothesize that chimeras in the AtC4H-SmC4H chimeric library with the highest functional activity will also have the highest number of residue pairs common to the CYP73 protein family. In addition to the fragment based metric explored in the previous section, here we introduce a summation of commonly occurring residue pairs as another possible metric.

5.2 Results and Discussion

5.2.1 CYP73 Multiple Sequence Alignment

Aligned CYP73 sequences were extracted from the MSA described in section 3.1.2. A tree of the aligned CYP73 sequences (Figure 19) shows that more of the known CYP73 sequences are closer to AtC4H than to SmC4H. This is not surprising. As of this writing, the complete genomes sequence of 27 organisms from the eudicotyledons class (including *A. thaliana*) are available online from the NCBI genome database, whereas only one organism from the Isoetopsida class has been sequenced (*S. moellendorffii*). This difference in known CYP73 sequences may bias the residue pair analysis.

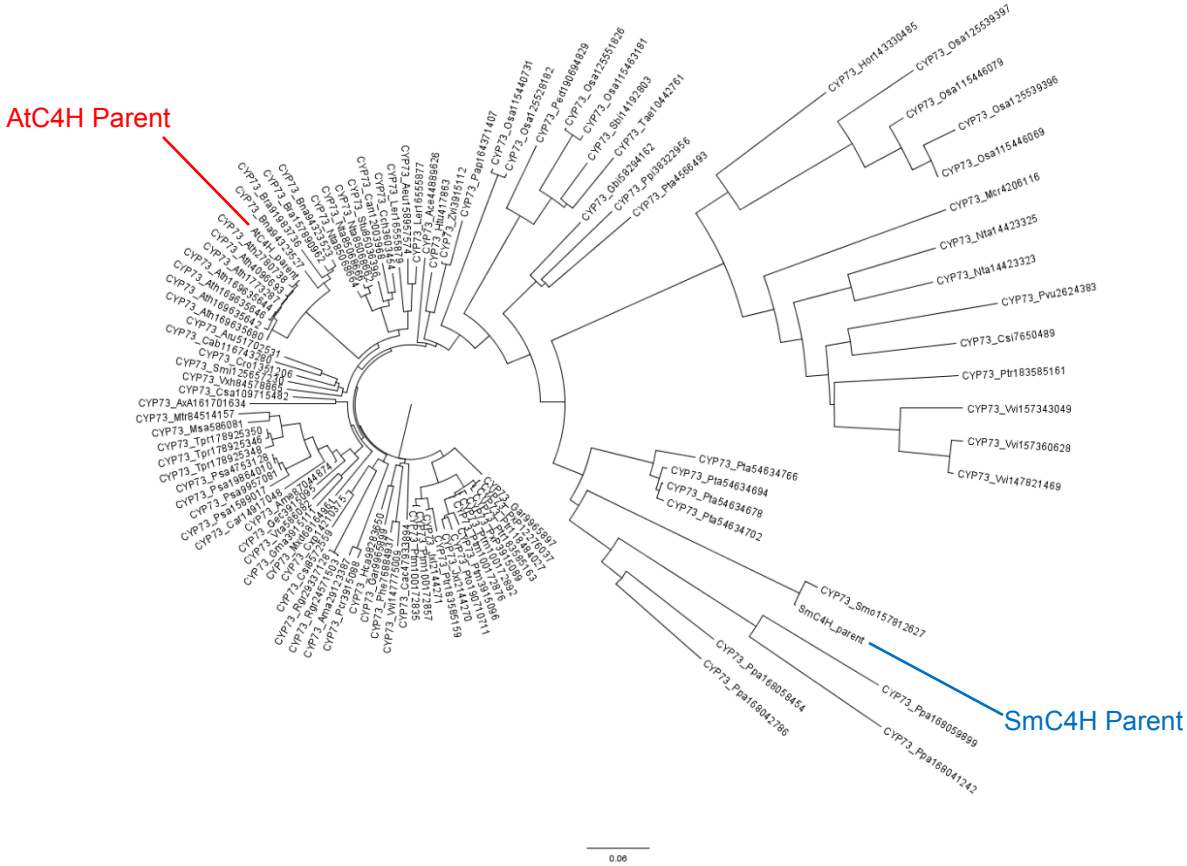


Figure 19 Tree of CYP73 sequences used in study

5.2.2 Column Pair Totals

The total number of polymorphic column pairs appearing the CYP73 MSA are grouped by distance and totaled for each chimera. Here, a polymorphic column pair is any residue pair that differs between AtC4H and SmC4H, and is associated with a non-gapped region of the AtC4H protein sequence in the CYP73 MSA. The term “column pair” is used instead of residue pair to emphasize that residue pairs are associated with their respective, fixed columns in the CYP73 MSA. The AtC4H sequence is used for a MODELLER homology model (99)(section 5.3.2). This structure is used to infer inter atomic distance between residue pairs. Only polymorphic column pairs are considered. Since non-polymorphic column pairs will contribute the same score for every chimera, no additional information will be added to the model. Therefore, non-polymorphic column pairs are not been included.

5.2.3 Column Pair Total Distance Groups are Highly Correlated

Based on the expected number of column pairs in Table 26, a large number of polymorphic column pairs do not appear between fragment 1 and all other fragments. This is because the 25 N-terminal amino acid residues of AtC4H were not fit by MODELLER. Gene fragment 1 has many polymorphic column pairs in physical contact with fragments 3 and 4. Some of these interactions may be the basis for the presumed stabilizing effect of SmC4H fragment 1 discussed in section 3.2.5.

Gene fragment 2 is highly conserved, with only two polymorphic residues between AtC4H and SmC4H. Neither of these residues occurs in the SRS motif. Therefore, it is not surprising to see that fragment 2 has few polymorphic column pairs overall, including zero polymorphic column pairs closer than 8 angstroms with gene fragment 3 and 4.

Gene fragments 4 and 5 display strong functional interaction; activity is greatly reduced or eliminated when chimeras contain gene fragments 4 and 5 from different species (Section 3.2.5). Interestingly, gene fragments 4 and 5 share 1843 divergent column pairs, but none of these column pairs are separated by less than 8 angstroms. This suggests a high degree of structural conservation within the 4-5 region, as we would expect since the homology model suggests that this region is involved in orienting the substrate during catalysis (69)

Polymorphic column pairs separated by 8 to 15 angstroms may also be in physical contact. These polymorphic residue pairs may be responsible for the non-replacability between AtC4H and SmC4H gene fragments 4 and 5.

Gene fragment 6 has the highest number of polymorphic column pairs less than 8 angstroms with every other gene fragment, except gene fragment 1. This is not surprising since gene fragment 6 and 1 are spatially distant (Figure 7). Taken together with the observation that gene fragment 6 is replaceable between

AtC4H and SmC4H (Section 3.2.5), this suggests that these polymorphic column pairs are not functionally and/ or structurally essential.

Table 26 Summary of divergent residue pairs grouped by gene fragments. [A]-residue pairs are grouped by distance. [B]-Total number of divergent residue pairs observed for a given gene fragment pair. [C]-Total number of residue pairs for a given pair of gene fragments. This is the product of the length of each gene fragment.

Fragment Pair	0 to 8Å[A]	8 to 15Å[A]	15 to 25Å[A]	25 to 99Å[A]	Total Divergent in MODELLER structure	Total divergent in sequence[B]	Total pairs in sequence[C]
1:2	0	15	64	441	520	1400	4136
1:3	37	76	278	1610	2001	5565	15134
1:4	33	98	122	682	935	2555	6862
1:5	0	5	164	1092	1261	3430	9400
1:6	0	11	54	442	507	1400	3854
2:3	0	0	1	307	308	318	7084
2:4	0	0	19	125	144	146	3212
2:5	5	41	58	90	194	196	4400
2:6	32	34	12	0	78	80	1804
3:4	15	292	1054	2167	3528	3796	11753
3:5	37	210	1189	3316	4752	5096	16100
3:6	38	95	331	1447	1911	2080	6601
4:5	0	51	338	1454	1843	1862	7300
4:6	6	32	56	647	741	760	2993
5:6	40	147	240	743	1170	1200	4100

5.2.4 Column Pair Totals are Effective Coveriates

The first model considered attempts to explain the activity of all 64 chimeras against the polymorphic column totals of all four distance bins. However, these predictors suffer from multicoliniearity as seen visually in Figure 20 and Table 28. After one or more distance bins are fit in the model, the marginal contribution of the remaining distance bins are not significant (Table 27). Furthermore, the fit parameters are either close to zero or have taken on negative values, opposite of what is expected from the visualization plot (Table 27). These are all classic indications of multicoliniarity and are not remedied by centering and scaling each predictor or increasing the number of distance bins used (data not shown). Combining the four distance bins into a single group does not result in a superior fit, nor does dropping terms from the model. The best fit is obtained when correlating the measured activity levels of all chimeras onto individual distance bins (Table 28). Here, the best correlation is found between chimera activity and column pair totals 8 to 15 Angstroms with ρ=0.58. This does not improve on any of the regression models presented in section 3.2.5 Future work may attempt a ridge regression to try and estimate more accurate, though biased, fit parameters.

On their own, column pair totals do not improve the predictive power of the consensus linear model (section 3.2.5). However, since the column pair totals pool information from the entire protein sequence, it was hypothesized that the sums may be capturing information not present in the simple linear regression analysis. The polymorphic column pair totals for each bin was tested as a

covariate with the consensus linear model (Table 29-32). All column pair totals added information above and beyond the simple linear model. Column pairs in the 0 to 8 angstrom group added the least information to the linear model, and the signs of fit parameters (negative or positive) is the opposite of what was observed for the conservation model presented in section 3.2.5 (Table 29).

Although we expect atoms in direct contact to carry the most relevant information about beneficial pairs, this does not seem to be the case. We note that this group contains the least number of polymorphic column pairs; including six gene fragment pairs that do not contain any polymorphic column pairs. Further, column pairs less than 8 angstroms apart may be subject to the most stringent selection, limiting extant residue pairs with negative effects on activity. Taken together, the limited number of observed column pairs and possible selection against discordant column pairs may limit the amount of recoverable information available from residue pairs less than 8 angstroms.

The column pair sums of the 8 to 15 angstrom group are also effective covariates, improving model fit above the simple linear model (Table 30). The adjusted R^2 is improved from 0.8353 in the linear consensus model to 0.8993 when including the 8 to 15 Å column pair totals as a covariate. Further, the fit parameters agree with the observed fragment based activity. This might be due to the fact that residue pairs in the 8 to 15 angstrom group are close enough to measure direct interactions, like the 0 to 8 angstrom group, while maintaining a large enough

diversity and number of column pairs to measure differences in activity, unlike residue pair sums in the 0 to 8 angstrom group. Compare that 13 out of 15 gene fragment pairs have at least one column pair 8 to 15 Å, versus only 9 out of 15 fragment pairs have at least one column pair less than 8 Å. Also, a total of 1107 column pairs have been identified in the 8 to 15 Å group, whereas only 237 column pair have been identified in the less than 8 Å group.

Column pair total in the 15 to 25 angstrom pair group are also effective covariates (Table 31). Fragment terms 3, 4, 5 and 6 have been successfully dropped from the model leaving only fragment 1, 2 and a 4:5 interaction. The R^2 for this model is 0.9048, indicating a very good fit. It appears that the residue pair sums have subsumed the functional relevance of the C-terminal half of the CYP73 protein, except for the fragment 4:5 interaction. Although initially encouraging, the biological interpretation of the fit parameters does not perfectly match observed functional activity. The fit parameters for fragment 1 and 2 show a negative impact on activity, agreeing with observed data. However, the fit parameter for At:At interaction in fragments 4 and 5 is highly negative, strongly contrasting with observed data. This complicates interpretation of this particular model and questions its ultimate utility.

Using column pair sums from the 25 to 99 angstrom group as a covariate to the simple linear model also improves model fit, and the parameters agree with the fragment based behavior of the chimeras (Table 32). Surprisingly, fragment 2

has been dropped from the model due to insignificance. This finding appears to suggest that the differences in activity attributed to fragment 2 have been subsumed by the column pair sum. Interestingly, the 25 to 99 covariate has lower correlation to chimera activity than 8 to 15 (ρ=0.046 vs ρ=0.58), but using 25 to 99 as a covariate results in a slightly better fitting model (R^2=0.9023 vs R^2=0.8993). This suggests that the 25 to 99 column pair score ads more information to the model independent of the fragment terms.

Removing the fragment 2 term from the model was not initially obvious. The ANOVA table for the simple linear model with 25 to 99 as a covariate shows fragment 2 significant by the F test, but not fragment 4. Removing fragment 4 from the model leads to the subsequent removal of fragments 3 and 5, but not 2, by failure of the F test. The final model has no significant parameters, and is discarded.

In the original consensus linear model with 25 to 99 as covariate, fragment 2 is significant by F test, but not the individual parameter by t-test due to high variance. Removing the fragment 2 term results in the model presented in Table 32, with adjusted R^2=0.9042 and all terms significant by t-test and F test.

The column pairs in the 25 to 99 distance group add the most information to the model beyond what is captured in the gene fragment terms. This is surprising because residue interactions are assumed to be indirect at this distance, if

existing at all. The significant contribution of column pairs in this group may be due in part to the large number of residues present, although combining distance bins did not improve model fit.

Attempts to combine or subdivide column pair bins result in worse fit of the models. This should not be surprising as different distance groups appear to explain different portions of the protein: 15 to 25 subsumes the activity explained by the gene fragments of the C-terminal half of the protein, whereas the 25 to 99 group subsumes the activity of fragment 2. Fragment 2 contains the fewest number of polymorphic column pairs at every distance group, so it is not clear why the effects of fragment 2 would be strongest. The C-terminal half of CYP73 contains the greatest number of column pair interactions, so it is perhaps not surprising that the effects of this dominate the 15 to 25 angstrom group. However, the C-terminal half of the protein has the greatest number of polymorphic column pairs at all distances, so it is not clear why we are unable to eliminate fragment terms 3, 4, 5 or 6 from other linear models using other distance groups as covariates.

Future work may assess the best covariate measure by testing a sliding window of angstrom distances. Based on the high level of multicolinearity observed between the different distance groups, the contribution of a single, optimal, distance group may be marginal. This approach should be taken with great care as the underlying relationship integrating column pair scores with fragment terms

is unknown, as are the exact selection criteria that should be used. Any optimization procedure that samples tens or hundreds of distance groups would need to consider many factors to select an appropriate linear model, including which fragment terms to include and how to select them.

In addition to finding an optimal group of column pairs to use as a covariate, improvements to the column pair sum metric itself may also be gained by improving the modeled C4H structure giving a more accurate measure of distance between column pairs, expanding the pool of extant sequences used, or weighting the contribution of each extant CYP73 sequences based on the evolutionary distance between sequences, or based on the number of known sequences in a given branch of the tree (i.e. down weighting the highly sampled angiosperm sequences).

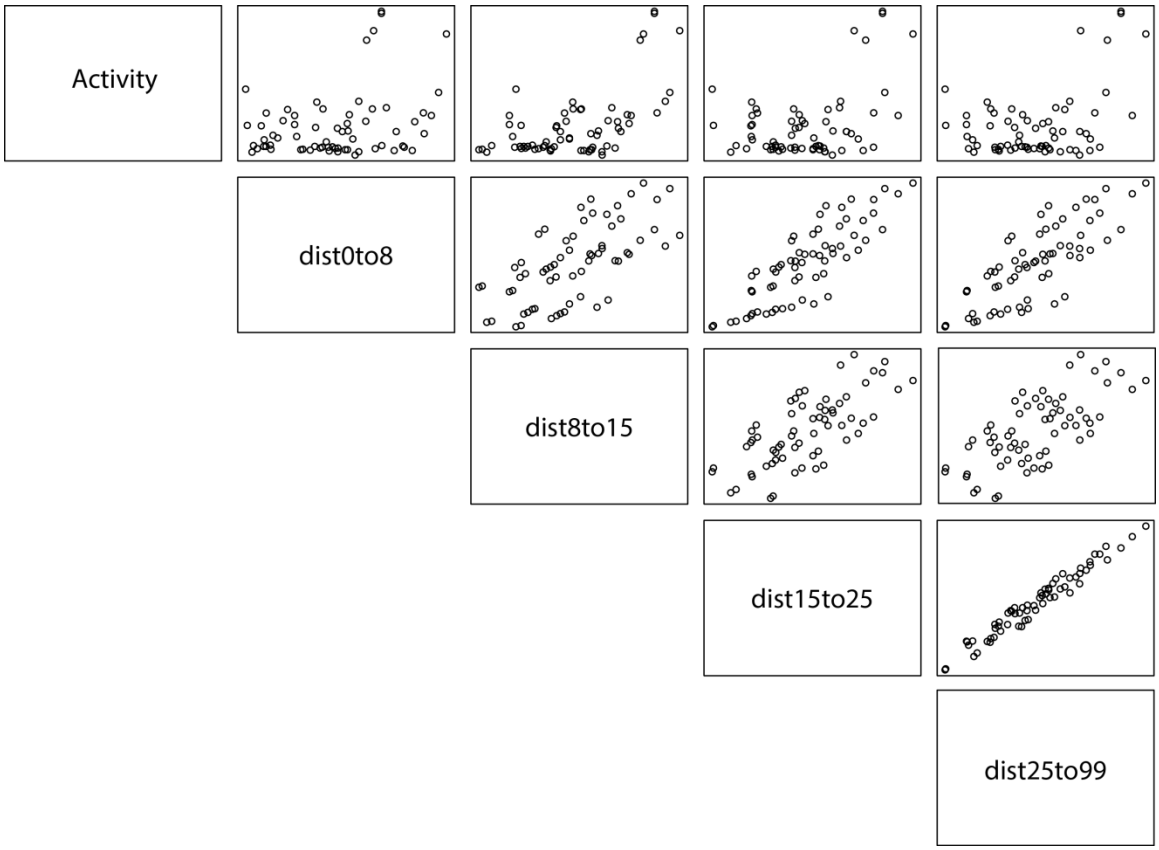


Figure 20 Coplot visually displaying the correlation between column pair sums in different distance groups. Each point in the graph is the column pair sum for a single chimera.

Table 27 ANOVA table and Type III Sums of Squares for activity by all distance groups.

Term	Coef	Df	Sum Sq	Mean Sq	F value	P-value	Type III SS	Type III F value
0to8	-0.16	1	100334	100334	76.6	<0.0001	2794	2.1
8to15	0.22	1	56704	56704	43.3	<0.0001	10308	7.8
15to25	-0.44	1	44237	44237	33.8	<0.0001	1191	0.9
25to99	0.83	1	7074	7074	5.4	0.02	7074	5.4
Error		61	79828					

Table 28 Covariance table showing linear relationship between chimera activity and column pair sums, and multicolinarity between different column pair sums.

	0to8Å	8to15Å	15to25Å	25to99Å	0to99Å
Activity	0.31	0.58	0.41	0.41	0.46
0to8Å	1	0.67	0.84	0.80	0.90
8to15Å		1	0.73	0.72	0.85
15to25Å			1	0.97	0.96
25to99Å				1	0.94
0to99Å					1

Table 29 ANOVA table for simple linear model with dist0to8 column pair sum as covariate. Adjusted R^2 is 0.8848

Term	Coefficient	Df	Sum Sq	Mean Sq	F value	P-value
Frag 1	At -1.6	2	154241	77121	150	<0.0001
	Sm 1.6					
Frag 2	Sm 1.6	1	70869	70869	138	<0.0001
Frag 3	Sm 0.21	1	5024	5024	9	0.0020
Frag 4	Sm 0.21	1	487	487	0.9	0.3000
Frag 5	Sm 0.027	1	6793	6793	13	0.0005
Frag 6	Sm 0.023	1	3636	3636	7	0.0090
Frag4:5	Sm 0.129	1	18010	18010	35	<0.0001
Error		57	29117	511		

Table 30 ANOVA table for simple linear model with dist8to15 column pair sum as covariate. Adjusted R^2 is 0.8993

Term	Coefficient	Df	Sum Sq	Mean Sq	F value	P-value
Frag 1	At -0.9	2	224167	112084	250.8	<0.0001
	Sm -0.89					
Frag 2	Sm -0.06	1	2957	2957	6.6	0.01
Frag 3	Sm -0.34	1	2390	2390	5.3	0.02
Frag 4	Sm -0.26	1	2377	2377	5.3	0.02
Frag 5	Sm -0.22	1	9619	9619	21.5	<0.0001
Frag 6	Sm -0.05	1	12029	12029	26.9	<0.0001
Frag4:5	Sm:Sm0.13	1	9163	9163	20.5	<0.0001
Error		57	25473	447		

Table 31 ANOVA table for simple linear model with dist15to25 column pair sum as covariate. Adjusted R^2 is 0.9048

Term	Coefficient	Df	Sum Sq	Mean Sq	F value	P-value
Frag 1	At 1.34	2	184032	92016	218.1	<0.0001
	Sm 1.36					
Frag 2	Sm -0.134	1	44706	44706	105.9	<0.0001
Frag 3 Term Not in Model						
Frag 4 Term Not in Model						
Frag 5 Term Not in Model						
Frag 6 Term Not in Model						
Frag4:5	At:At -0.326	3	34544	11515	27.3	<0.0001
	Sm:At -0.305					
	At:Sm -0.319					
	Sm:Sm 0.000					
Error		59	24894	422		

Table 32 ANOVA table for simple linear model with dist25to99 column pair sum as covariate. Adjusted R^2 is 0.9023

Term	Coefficient	Df	Sum Sq	Mean Sq	F value	P-value
Frag 1	At -0.49	2	190613	95307	223.6	<0.0001
	Sm-0.61					
Frag 2 Term Not in Model						
Frag 3	Sm -0.27	1	16577	16577	38.9	<0.0001
Frag 4	Sm -0.31	1	8784	8784	20.6	<0.0001
Frag 5	Sm -0.21	1	28490	28490	66.8	<0.0001
Frag 6	Sm -0.05	1	4584	4584	10.8	0.0017
Frag4:5	Sm 0.18	1	14409	14409	33.8	<0.0001
Error		58	426			

5.3 Materials and Methods

5.3.1 Column Pair Sums

For a given chimeric protein sequence, each residue is associated with a unique column in the CYP73 MSA. Therefore, each pair of residues in a given chimeric protein sequence is associated with a unique pair of columns in the CYP73 MSA. The number of times each residue pair in a chimeric protein sequence appeared in the CYP73 MSA at its respective column pair was summed. Each chimeric sequence has n(n-1)/2 column pair sums, where n is the length of the sequence in amino acids. Column pair sums were calculated using a custom Python script.

5.3.2 Sorting of Column Pairs into Distance Groups

Pairs of columns in the MSA were assigned α–carbon to α-carbon distances using a structural model of the AtC4H protein. The structural model of the AtC4H protein was made by threading the AtC4H amino acid sequence onto CYP7A1 (3dax) crystal structure by Thomas Sors using the MODELLER software package (99). The 25 N-terminal amino acids of AtC4H, corresponding to the transmembrane domain, were not included in the MODELLER results. The MSA columns corresponding to the 25 N-terminal amino acids and columns in the MSA that did not correspond to any AtC4H amino acid (i.e. gaps) were ignored. Based on the α–carbon to α-carbon distance in the threaded AtC4H structure, column pairs were sorted into one of four distance groups: 0 to 8 Angstroms, 8 to 15 Angstroms, 15 to 25 Angstroms or 25 to 99 Angstroms. No amino acid pair in

the model was greater than 99 Angstroms apart. Interatomic distances were calculated using a custom Python script.

5.3.3 Column Pair Totals

For a given chimera, column pair sums for each α–carbon to α-carbon distance group were added together. This results in four column pair totals for every protein chimera. Column pair totals were calculated using a custom Python script.

5.3.4 Statistical Modeling of Column Scores as a Covariate

Formatting of raw column pair totals and interatomic distance data was completed using custom Python scripts. Statistical analysis was done using R 3.0.1 with the standard packages. Column pair totals for each chimera were used as a covariate in ANCOVA regression models, with the parameters from the consensus model (Table 8 –Significant fit terms for linear ANOVA models using one and two way gene fragments as factors. Terms for the complete linear model were fit as described in Table 7. Terms for the consensus linear model are terms and interactions significant at the Bonferroni level across all regression models. Asterisks indicate explanatory factor is significant at the Bonferroni level ($\alpha \leq 0.0041$).). The activity terms fit by the full model for the AtC4H-SmC4H chimeric library were used as the response variable with the standard deviation as weights (Table 6).

CHAPTER 6. CONCLUSIONS

We have shown design and construction of a successful chimeric protein library in the absence of structural information. Structure-function-sequence constraints preserved by evolution, and associable through multiple sequence alignments, are sufficient for identifying interchangeable regions of homologous proteins. Combinatorial interchange of these regions yields a high proportion of function children chimeras. Every protein that could be a candidate for use as a parent in a chimeric library must have a known sequence. With ever increasing sequence information available, it becomes rarer and rarer to find a single protein that is not associated with a set of known homologues. We hope demonstrating that MSA contain all the information needed to generate functional protein chimeras makes research with chimeric proteins much more accessible to scientists everywhere.

We have also shown the importance of recovering a complete, unbiased chimeric library. Recovering all chimeras, as opposed to a subset of functional chimeras, biasing sequence data, allows us to construct a complete statistical model of the proteins. In each library, a complete model has given new insights to the underlying function and importance of different regions of the genes under study. These insights might not be available when testing a possibly biased subset of

chimeras. Fortunately, increasingly accessible high throughput technologies makes pursuing these kinds of studies easier than ever.

I have also shown that column pair totals are significantly associated with the activity levels of chimeric proteins. This relationship describes activity partly independent of a gene fragment relationship. The relationship between protein function and whole gene sequence has long been known to be complex. We are just beginning to describe some of the metrics that will enable predictable, off the shelf protein design.

6.1 Future Work

This work has shown that multiple sequence alignments are sufficient for identifying functionally equivalent gene fragments in related proteins. This implies, perhaps unsurprisingly, that structure-function-sequence constraints are preserved by evolution. Extending this work along the branches of an evolutionary tree is a logical next step. One method would be testing the interchangeability of selected gene fragments within, and between, gene families. If this hypothesis it true, then we expect gene fragments most important to function are replaceable in a clade based manner. That is to say, genes more closely related should be more readily swapped without negatively impacting function. If enough genes are tested in the manner, perhaps a metric describing the tradeoff between evolutionary distance and functional impact could be developed. Higher functional impact could result in fewer functional chimeras,

but also greater diversity of function through greater exploration of sequence space. If a gene fragment does not affect function, then it may be interpreted that that region of the protein is simply structurally relevant, and has not been subjected to functional evolutionary selection and divergence.

The secondary structure of short amino acid short chameleon sequences have been shown to be highly dependent on the amino acids they are in immediate contact with. Probing chimeras with novel chameleon sequences, or looking at the range of chameleon sequences present in a given multiple sequence alignment may also reveal valuable structural or functional insights.

The literature has not conclusively answered if non-functional chimeras are the result of bad breakpoints or bad proteins. Comparing the success of the AtC4H-SmC4H library with the failure of the AtC4H-AtF5H library suggests bad proteins. If a multiple sequence alignment truly captures functionally equivalent regions of a protein, then we would expect randomly selected breakpoints to perform as well as breakpoints selected in regions of high conservation.

Underlying all research into chimeric proteins is a poorly understood relationship between sequence space and functional space. Does functional space exist as a contiguous, unbroken volume within sequence space, does it contain gaps, or is functional space non-contiguous? Answering this question would grant valuable insight into the utility of specific, or groups of specific residue changes.

The well known existence of rescue mutations and plastic structural elements suggests a complex relationship between functional space and sequence space. This relationship may be directly probed with the advent of next gen sequencing technologies. First, use degenerate DNA oligos to construct a pool of genes that are a mosaic of homologous. Next, use rapid functional screening combined with next gen sequencing to classify hundreds of thousands, or millions, of sequences as either functional or non-functional. Clustering and distance algorithms can then be applied to this dataset to begin to describe the functional space that the selected gene occupies.

WORKS CITED

WORKS CITED

1. Pardo L, Vicente AI, Mate DM, Alcalde M, Camarero S. Development of Chimeric Laccases by Directed Evolution. *Biotechnol. Bioeng.* 2012: DOI: 10.1002/bit.24588.
2. Cobb RE, Si T, Zhao H. Directed evolution:an evolving and enabling synthetic biology tool. *Current Opinion in Chemical Biology.* 2012; **16**: 285-291.
3. Coelho PS, Wang ZJ, Ener ME, Baril SA, Kannan A, Arnold FH, et al. A serine-substituted P450 catalyzes highly efficient carbene transfer to olefins in vivo. *Nature Chemical Biology.* 2013; **9**: 485-490.
4. Chodera JD, Mobley DL. Entropy-Enthalpy compensation: Role and ramifications in biomolecular ligand and recognion and design. *Annual Reviews of Biophysics.* 2013; **42**: 121-142.
5. Brannigan JA, Wilkinson AJ. Protein engineering 20 years on. *Nature Reviews:molecular cell biology.* ; **3**: 964-970.
6. Yano T, Kagamiyama H. Directed Evolution of Ampicillin-Resistant Activity from a Functionally Unrelated DNA Fragment: A Laboratory Model of Molecular Evolution. *PNAS.* 2001; **98**: 903-907.
7. Kinch LN, Grishin NV. Evolution of protein structures and functions. *Current Opinion in Structural Biology.* 2002; **12**: 400–408.
8. Shenoy AR, Visweswariah SS. Mycobacterial adenylyl cyclases: Biochemical diversity and structural plasticity. *FEBS letters.* 2006; **580**: 3344–3352.
9. Hiraga K, Arnold FH. General Method for Sequence-independent Site-directed Chimeragenesis. *J. Mol. Biol.* 2003; **330**: 287-296.
10. Otey CR, Landwehr M, Endelman JB, Hiraga K, Bloom JD, Arnold FH. Structure-Guided recombination creates an artifical family of cytochromes P450. *PLOS biology.* 2006; **4**: 789-798.
11. Landwehr M, Carbone M, Otey CR, Li Y, Arnold FH. Diversification of catalytic function in a synthetic family of chimeric cytochrome P450s. *Chemistry and Biology.* 2007; **14**: 269-278.

12. Hasegama H, Holm L. Advances and pitfalls of protein structural alignment. *Curr Opin Struct Biol..* 2009; **19**: 341-348.
13. Csaba G, Birzele F, Zimmer R. Protein Structure Alignment Considering Phenotypic Plasticity. *Bioinfomatics.* 2008; **24**: i98-i104.
14. McCubrey JA, Steelman LS, Chappell WH, Abrams SL, Montalto G, Cervello M, et al. Mutations and Deregulation of Ras/Raf/MEK/ERK and PI3K/PTEN/Akt/mTOR Cascades Which Alter Therapy Response. *Oncotarget.* 2012 September; **3**: 954-987.
15. Bos JL. Ras Oncogenes in Human Cancer:A Review. *Cancer Res.* 1989; **49**: 4682-9.
16. Metro G, Crinò L. Novel molecular trends in the management of advanced non-small-cell lung cancer. *Expert Rev Anticancer Ther.* 2012; **12**: 729-32.
17. Konstantinopoulos PA, Karamouzis MV, Papavassiliou AG. Post-translational modifications and regulation of the Ras superfamily of GTPases as anticancer targets. *Nature Reviews.* 2007; **6**: 551-555.
18. Choy E, Chiu VK, Silletti J, Feoktistov M, Morimoto T, Michaelson D, et al. Endomembrane Trafficiking of Ras: The CAAX Motif Targets Proteins to the ER and Golgi. *Cell.* 1999; **98**: 69-80.
19. Silvius JR. Mechanisms of Ras Protein Targeting in Mammalian Cells. 2002; **190**: 83-92.
20. Hancock JF, Cadwallader K, Paterson H, Marshall CJ. A CAAX or a CAAL motif and a second signal are sufficient for plasma membrane targeting of ras proteins. *EMBO J.* 1991; **10**: 4033-4039.
21. Laude AJ, Prior IA. Palmitoylation and localization of PAS isofroms are moulated by the hypervairable linker domain. *J. Cell Sci.* 2008; **121**: 421-427.
22. Hancock JF, Paterson H, Marshall CJ. A polybasic domain or palmitoylation is required in addition to the CAAX motif to localize p21ras to the plasma membrane. *Cell.* 1990; **63**: 133-139.
23. Gutierrez L, Magee AI, Marshall CJ, Hancock JF. Post-translational processing of p21ras is two-step and involves carboyl-methylation and coroxy-terminal proterolysis. *EMBO J.* 1989; **8**: 1093-1098.
24. Ashby MN. CaaX converting enzymes. *Curr Opin Lipidol.* 1998; **9**: 99-102.
25. Hancock JF, Magee AI, Childs JE, Marshall CJ. All ras proteins are polyisoprenylated but only some are palmitoylated. *Cell.* 1989; **57**: 1167-1177.
26. Saftalov L, Smith PA, Friedman AM, Bailey-Kellogg C. Site-Directed Combinatorial Construction of Chimeric Genes: General Method for Optimizing Assembly of Gene Fragments. *Proteins.* 2006; **64**: 629-642.
27. Kranenburg O, Verlaan I, Moolenaar WH. Regulating c-Ras function: cholesterol depletion affects caveolin association, GTP loading, and signaling. *Curr Biol.* 2001; **11**: 1880-1884.

28. Quatela SE, Sung PJ, Ahearn IM, Bivona TG, Philips MR. Analysis of K-Ras phosphorylation, translocation and induction of apoptosis. *Methods Enzymol.* 2008; **439**: 87-102.
29. Nocedal J, Wright SJ. Numerical Optimization. *New York: Springer-Verlag.* 1999: 65-100.
30. Xiao Y, Zhang L, Gao S, Saechao S, Di P, Chen J, et al. The c4h, tat, hppr and hppd Genes Prompted Engineering of Rosmarinic Acid Biosynthetic Pathway in Salvia miltiorrhiza Hairy. *PLOS One.* 2011; **12**: e29713.
31. Shin SY, Jung SM, Kim MD, Han NS, Seo JH. Production of reveratrol from tyrosine in metabolically engineered Saccharomyces cerevisiae. *Enzyme and Mincrobial Technology.* 2012; **51**: 211-216.
32. Park NI, Park JH, Park SU. Overexpression of Cinnamate 4-Hydroxylase Gene Enhances Biosynthesis of Decursinol Angelate in Angelica gigas Hairy Roots. *Molecular Biotechnology.* 2012: 114-120.
33. Muñoz C, Sánchez-Sevilla JF, Botella MA, Hoffmann T, Schwab W, Valpuesta V. Polyphenol Composition in the Ripe Fruits of Fragaria Species and Transcriptional Analyses of Key Genes in the Pathway. *Journal of Agricultural and Food Chemistry.* 2011; **59**: 12598-12604.
34. Lewis N. A 20th century roller coaster ride: a short account of lignification. *Curr. Opin. Plant Biol.* 1999; **2**: 153-162.
35. Weng JK, Li X, Bonawitz ND, Chapple C. Emerging strategies of lignin engineering and deradation for cellulosic biofuel production. *Curr. Opin. Biotech.* 2008; **19**: 166-172.
36. Pandiyan K, Tiwari R, Rana S, Arora A, Singh S, Saxena AK, et al. Comparative efficiency of different pretreatment methods on enzymatic digestibility of Parthenium sp. *World Journal of Microbiology and Biotechnology.* 2013 July: ePub DOI 10.1007/s11274-013-1422-1.
37. Baucher M, Monties B, Van Montagu M, Boerjan W. Biosynthesis and Genetic Engineering of Lignin. *Critical Reviews in Plant Sciences.* 1998; **17**: 125-197.
38. Chen F, Dixon R. Lignin modification improves fermentable sugar yields for biofuel production. *Nat. Biotechnology.* 2007; **25**: 759-761.
39. Mainsfield D. Solutions for dissolution-engineering cell walls for deconstruction. *Curr Opin in Biotechnology.* 2009; **20**: 286-294.
40. Chen F, Srinivasa Reddy MS, Temple S, Jackson L, Shadle G, Dixon RA. Multi-site genetic modulation of monolignol biosynthesis suggests new routes for formation of syringyl lignin and wall bound ferulic acid in alfalfa(Medicago sativa L.). *The Plant Journal.* 2006; **48**: 113-124.
41. Bjurhager I, Olsson AM, Zhang B, Gerber L, Kumar M, Berglund LA, et al. Ultrastructure and Mechanical Properties of Populus Wood with Reduced Lignin Content Caused by Transgenic Down-Regulation of Cinnamate 4-Hydroxylase. *Biomacromolecules.* 2010; **11**: 2359-2365.

42. Kumar S, Omer S, Patel K. Cinnamate 4-Hydroxylase (C4H) genes from Leucaena leucocephala: a pulp-yielding leguminous tree. *Molecular Biology Reports.* 2013; **40**: 1265-1274.
43. Schalk M, Batard Y, Seyer A, Nedelkina S, Durst F, Werck-Reichart D. Design of Fluorescent Substrates and Potent Inhibitors of CYP73As, P450s That Catalyze 4-Hydroxylation of Cinnamic Acid in Higher Plants. *Biochemistry.* 1997; **36**: 15253-15261.
44. Chen H, Jiang H, Morgan JA. Non-natural cinnamic acid derivatives as substrates of cinnamate 4-hydroxylase. *Phytochemistry.* 2006; **68**: 306-311.
45. Schoch GA, Attias R, Le Ret M, Werck-Reichhart D. Key substrate recognition residues in the active site of a plany cytochrome P450, CYP73A1. *Eur. J. Biochem.* 2003; **270**: 3684-3695.
46. Schalk M, Nedekina M, Schoch G, Batard Y, Werck-Reichhart D. Role of unusual amino acid residues in the proximal and distal heme regions of a plant P450, CYP73A1. *Biochemistry.* 1999; **38**: 6093-6103.
47. Zabala G, Zou J, Tuteja J, Gonzalez DO, Clough SJ, Vodkin LO. Transcriptome changes in the phenylpropanoid pathway of Glycine max in response to Pseudomonas syringae infection. *BMC Plant Biology.* 2006; **6**: 26.
48. Weng JK, Akiyama T, Bonawitz ND, Li X, Ralph J, Chapple C. Conevergent Evolution of Syringyl Lignin Biosynthesis via distinct pathways in the lycophyte selaginella and flowering plants. *Plant Cell.* 2010; **4**: 1033-1045.
49. Banks JA. Selaginella and 400 Million Years of Separation. *Annual Reviews in Plant Biology.* 2009; **60**: 223-238.
50. Portugaly E, Linial N, Linal M. EVEREST: a collection of evolutionary conserved protein domains. *Nucleic Acids Research.* 2007; **35**: D241-D246.
51. Voight CA, Martinez C, Wang ZG, Mayo SL, Arnold FH. Protein building blocks preserved by recombination. *Nature Structural Biology.* 2002; **9**: 553-558.
52. Edgar CR. MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Research.* 2004; **32**: 1792-1797.
53. Edgar CR. MUSCLE: A Multiple Sequence Alignment Method with Reduced Time and Space Complexity. *BMC Bioinformatics.* 2004; **2004**: 113-131.
54. Lutz S, Ostermeier M, Moore GL, Maranas CD, Benkovic SJ. Creating multiple-crossover DNA libraries independent of sequence identity. *PNAS.* 2001; **98**: 11248-11253.

55. Kawarasaki Y, Griswold KE, Stevenson JD, Selzer T, Benkovic SJ, Iverson BL, et al. Enhanced crossover SCRATCHY: construction and high-throughput screening of a combinatorial library containing multiple non-homologous crossovers. *Nucleic Acids Research.* 2003; **31**: e126.
56. Udit AK, Silberg JJ, Sieber V. Sequence Homology-Independent Protein Recombination (SHIPREC). *Methods in Molecular Biology.* 2003; **231**: 153-163.
57. Sieber V, Martinez CA, Arnold FH. Libraries of hybrid proteins from distantly related sequences. *Nature Biotechnology.* 2001; **19**: 456-460.
58. O'Maille PE, Bakhtina M, Tsai MD. Structure-based Combinatorial Protein Engineering (SCOPE). *J. Mol. Biol.* 2002; **321**: 677-691.
59. Otey CR, Landwehr M, Endelman JB, Hiraga K, Bloom JD, Arnold FH. Structure Guided Recombination Creates an Artificial Family of Cytochromes P450. *PLOS Biology.* 2006; **4**: e112.
60. Hamann T, Lindberg Møller B. Improved cloning and expression of cytochrome P450s and cytochrome P450 reductase in yeast. *Protein Expression and Purification.* 2007; **56**: 121-127.
61. Pompon D, Louerat B, Bronine A, Urban P. Yeast expression of animal and plant P450s in optimized redox environments. *Methods in Enzymology.* 1996; **272**: 51-64.
62. Jiang H, Morgan JA. Optimization of an in vivo plant P450 monooxygenase system in Saccharomyces cerevisiae. *Biotechnology and Bioengineering.* 2004; **85**: 130-137.
63. Chen H, Morgan JA. High throughput screening of heterologous P450 whole cell activity. *Enzyme and Microbial Technology.* ; **38**: 760-764.
64. Endelman JB, Bloom JD, Otey RC, Landwehr M, Arnold FH. Inferring interactions from combinatorial protein libraries. *Arxiv.* 2005: 1-21.
65. Tobin J. Estimation of relationships for limited dependent variables. *Econometrica.* 1958; **26**: 24-36.
66. Heinzelman P, Komor R, Kanaan A, Romero P, Yu X, Mohler S, et al. Efficient Screening of Fungal Cellobiohydrolase Class I Enzymes for Thermostabilizing Sequence Blocks by SCHEMA Structure-Guided Recombination. *Protein Engineering, Design & Selection.* 2010; **23**: 871-880.
67. Graham SE, Peterson JA. How similar are P450s and what can their differences teach us? *Archives of Biochem and Biophys.* 1999; **369**: 24-29.
68. Gotoh O. Substrate recognition sites in cytochrome P450 family 2 (CYP2) proteins inferred from comparative analyses of amino acid and coding nucleotide sequences. *Journal of Biological Chemistry.* 1992; **267**: 83-90.
69. Rupasinghe S, Baudry J, Schuler MA. Common active site architecture and binding strategy of four phenylpropanoid P450s from Arabidopsis thalana as revealed by molecular modeling. *Protein Engineering.* 2003; **16**: 721-731.

70. Humphreys JM, Hemm MR, Chapple C. New Routes for Lignin Biosynthesis Defined by Biochemical Characterization of Recombinant Ferulate 5-Hydroxylase, a Multifunctional Cytochrome P450-dependent Monoxoygenase. *PNAS:Biochemistry.* 1999; **96**: 10045-10050.
71. Estabrook RW, Werringlower J. The Measurement of Difference Spectra: Application to the Cytocromes of Microsomes. *Methods in Enzymology.* 1978; **52**: 212-220.
72. Hatfield GW, Roth DA. Optimizing scaleup yield for protein production: Computationally Optimized DNA Assembly (CODA) and Translation Engineering (TM). *Biotechnology Annual Review.* 2007; **13**: 27-42.
73. Irwin B, Heck DJ, Hatfield WG. Codon Pair Utilization Biases Influence Translational Elongation Step Times. *The Journal of Biological Chemistry.* 1995: 22801-22806.
74. Phillips IR, Shepard EA. Spectral Analyses of Cytrochrome P450. 107th ed. Phillips IR, Shepard EA, editors.: Springer; 1998.
75. Fox-Erlich S, Schiller MR, Gryk MR. Structural conservation of a short, functional, peptide-sequence motif. *Fronteirs in Bioscience.* 2009; **14**: 1143-1151.
76. Sharif J, Endo TA, Ito S, Ohara O, Koseki H. Embracing change to remain the same:conservation of polycomb functions despite divergence of binding motifs among species. *Current Opinion in Cell Biology.* 2013; **25**: 305-313.
77. Fetzner S. Ring-Cleaving Dioxygenases with a Cupin Fold. *Applied Environemental Microbiology.* 2012; **78**: 2505-2514.
78. Hide WA, Chan L, Li WH. Structure and evolution of the lipase superfamily. *Journal of Lipid Research.* 1992; **33**: 167-178.
79. Hatayama M, Aruga J. Gli Protein Nuclear Localization Signal. *Vitamins & Hormones.* 2012; **88**: 73-89.
80. Davey NE, Edwards RJ, Shields DC. Computational identification and analysis of protein short linear motifs. *Fronteirs in Bioscience.* 2010; **15**: 801-25.
81. Galperin MY, Koonin EV. Evolution of Divergence and Convergence in Enzymes. *Journal of Biology and Chemistry.* 2011; **287**: 21-28.
82. Cane DE, Ikeda H. Exploration and Mining of the Bacterial Terpenome. *Acc Chem res.* 2012; **45**: 463-472.
83. Wu J, Xie J. Hidden Markov Models and its Application in Motif Findings. *Methods in Molecular Biology.* 2010; **620**: 405-416.
84. Brystoff C, Krogh A. Hidden Markov Models for prediction of protein features. *Methods in Molecular Biology.* 2008; **413**: 173-198.
85. Bentez-Paez A, Cardenas-Brito S, Gutierrez AJ. A Practical Guide For the Computational Selection of Residues to be Experimentally Characterized in Protein Families. *Briefings in Bioinformatics.* 2012; **13**: 329.

86. McGuffin LJ, Bryson K, Jones DT. The PSIPRED Protein Structure Prediction Server. *Bioinformatics.* 2000; **16**: 404-405.
87. Jones DT. Protein Secondary Structure Prediction Based on Psition-specific Scoring Matrices. *Journal of Molecular Biology.* 1999; **292**: 195-202.
88. Lin K, Simossis VA, Taylor WR, Heringa J. A Simple and Fast Secondary Structre Prediction Method using Hidden Neural Networks. *Bioinformatics.* 2005; **21**: 152-159.
89. Pierri CL, Parisi G, Porcelli V. Computational Approaches for Protein Function Prediction: A Combined Strategy From Multiple Sequence Alignment to Molecular Docking-Based Virtual Screening. *Biochemistry and Biophysics Acta.* 2010; **1804**: 1695-1712.
90. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI_BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Research.* 1997; **25**: 3389-3402.
91. Fieldhouse RJ, Merrill AR. Needle in the Haystack:Structure-Based Toxin Discovery. *Trends in Biochemical Science.* 2008; **33**: 546-556.
92. Nobel WS, Kuang R, Leslie C, Weston J. Identifying Remote Protein Homologs by Network Propagation. *The FEBS Journal.* 2005; **272**: 5119-5128.
93. Agrawal A, Huang X. PSIBLAST_PairwiseStatSig: Reordering PSI_BLAST Hits Using Pairwise Statistical Significance. *Bioinformatics.* 2009; **25**: 1082-1083.
94. Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R. Evolutionary information for specifying a protein fold. *Nature.* 2005; **437**: 512-518.
95. Gromiha MM, Pathak MC, Saraboji K, Ortund EA, Gaucher EA. Hydorphobic Environment is a Key Factor For the Stability of Thermophilic Proteins. *Proteins.* 2013; **81**: 715-721.
96. Kamiya K, Boero M, Shiraishi K, Oshiyama A, Shigeta Y. Energy Compensation Mechanism for Charge-Separated Protonation Sates in Aspartate-Histidine Amino Acid Residue pairs. *Journal of Physical Chemistry.* 2010; **114**: 6567-6578.
97. Clark AT, Smith K, Muhandiram R, Edmondson SP, Shriver JW. Carboxyl pK(a) Values, Ion Pairs, Hydrogen Bonding, and the pH-Dependence of Folding the Hyperthermophile Proteins Sac7d and Sso7d. *Journal of Molecular Biology.* 2007; **28**: 992-1008.

98. Shirota M, Kinoshita K. Analyses of the General Rule on Residue Pair Frequencies in Local Amino Acid Sequences of Soluble, Ordered Proteins. *Protein Science.* 2013; **22**: 725-733.
99. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen My, et al. Comparative Protein Structure Modeling With MODELLER. *Current Protocols in Bioinformatics, John Wiley & Sons, Inc.* 2006; **Supplement 15**: 5.6.1-5.6.30.

VITA

VITA

Nicholas J. Fico
Graduate School, Purdue University

Education

B.S., Microbiology, 2005, University of Maryland at College Park, Greenbelt, MD
M.S., Applied Statistics, 2013, Purdue University, West Lafayette, Indiana
Ph.D., Biology, 2013, Purdue University, West Lafayette, Indiana