

Combinatorial Structure and Function Studies

by

Simon Delagrave

B.Sc. First Class Honours in Biochemistry  
McGill University, Montréal, Canada (1991)

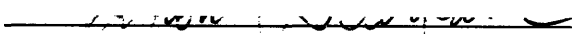
Submitted to the Department of Chemistry  
in partial fulfillment of the requirements for the Degree of


Doctor of Philosophy in Chemistry

at the

Massachusetts Institute of Technology  
June, 1995

© Massachusetts Institute of Technology  
All rights reserved

Signature of Author   
February 8, 1995

Certified by   
Professor Douglas C. Youvan, Thesis Supervisor

Accepted by   
Professor Dietmar Seyferth  
Chairman, Departmental Committee on Graduate Students

Science

MASSACHUSETTS INSTITUTE  
OF TECHNOLOGY

JUN 12 1995

LIBRARIES

This Doctoral thesis has been examined by a Committee of the Department of Chemistry as follows:

Professor Alexander M. Klibanov \_\_\_\_\_ Chairman

Professor Douglas C. Youvan \_\_\_\_\_ Thesis Supervisor

Professor William H. Orme-Johnson III \_\_\_\_\_ Committee Member

Professor James R. Williamson \_\_\_\_\_ Committee Member

# Combinatorial Structure and Function Studies

by

Simon Delagrave

Submitted to the Department of Chemistry  
on February 8 1995 in partial fulfillment of the requirements for the  
Degree of Doctor of Philosophy in Chemistry

## Abstract

Through billions of years of evolution, Nature has developed an enormously versatile array of intricate "machines" which are both the product and an essential component of what we call Life. These molecular machines, or proteins, are all composed of the same twenty building blocks (i.e., amino acids) arranged in different ways, like beads on a string. In recent years much effort has been spent to understand how a chain of amino acids in solution achieves a single conformation allowing it to bind other molecules or catalyze chemical reactions. Strategies which attempt to mimic and accelerate the evolutionary process which has given rise to the proteins we observe in Nature have successfully been used to engineer proteins. These strategies all have in common the construction of an ensemble of different proteins from which a functional protein with a desired property is selected. Through molecular genetic techniques, the ensemble of proteins is made diverse by introducing different combinations of amino acids at a few key positions in the polypeptide chain. For example, combinatorial cassette mutagenesis (CCM) can be used to randomize codons of the gene encoding a protein to be altered.

Typically, these ensembles of altered proteins contain *random* combinations of all twenty amino acids at a few positions along the polypeptide chain. Based on theoretical considerations and computer simulations, it has been proposed (Arkin and Youvan, *Proc. Natl. Acad. Sci. USA*, **89**, 7811-7815, 1992) that by determining which amino acids occur in the functional proteins selected from a random ensemble, a new combinatorial ensemble could be designed where a bias is introduced towards the expression of these functional amino acids. This new ensemble would contain a high number of functional sequences diverse enough to show phenotype diversity. In this thesis, the combinatorial optimization scheme described above is first experimentally shown to be valid and is then refined as a protein engineering methodology. Finally, the design and construction of an optimized library of antibodies, using many of the features of the scheme described above, are also discussed.

We used the Light Harvesting II (LHII) protein of *Rhodobacter capsulatus* as an initial target for combinatorial mutagenesis studies. This integral membrane protein, composed of  $\alpha$ ,  $\beta$  and  $\gamma$  subunits, binds chromophores (bacteriochlorophylls) which act as reporter groups of LHII structural integrity and expression. Combinatorial ensembles (or libraries) of altered proteins

expressed in *R. capsulatus* were spectroscopically characterized by Digital Imaging Spectroscopy (DIS) of colonies on petri dishes. Absorption spectra of hundreds of colonies can be acquired rapidly by DIS, allowing an efficient isolation and characterization of functional LHII mutants. Recursive Ensemble Mutagenesis (REM) is the designation of the first combinatorial optimization scheme to be investigated experimentally. REM was used to construct an optimized combinatorial library where one in 300 mutants screened expressed a functional LHII protein. This is a 30-fold improvement over a purely random library where only one in  $10^4$  mutants screened showed significant LHII expression as determined by DIS. While only six residues were mutated in the first experiment, an extension of REM called Exponential Ensemble Mutagenesis (EEM) was used to simultaneously alter 16 amino acid residues of LHII. Approximately one percent of the mutants in the optimized combinatorial library expressed functional LHII proteins, corresponding to an estimated  $10^7$ -fold increase in the frequency of functional mutants compared to a random library. The EEM method requires, at first, that small groups of residues be randomized independently and that functional mutants be isolated from each library. After the sequences of functional proteins are determined, a new combinatorial ensemble is designed in which the information from different "regions" of the protein sequence are pooled into a single library. We also describe experiments where these residues, if grouped differently, will produce different results. A specific mutated position of the LHII protein sequence was associated with a switch between two phenotypes, LHI and LHII. At that position, the amino acids which lead to phenotype LHI in one library did not have the same effect in another library. These results are based on correlations of sequence and phenotype as well as the observation of different proportions of the two phenotypes in each library. Our conclusion is that phenotype prediction from sequence is dependent on the experimental context but that elements of primary structure responsible for phenotype change can be identified despite changes in this context.

Finally, an application of these optimization algorithms is described. Using phylogenetic information on antibody sequences, an optimized combinatorial library was designed in which mutations are introduced in the heavy chain Complementarity Determining Region 3 (CDR3). These mutations should be complex enough to constitute a naive ensemble of antibodies, capable of recognizing any antigen. Moreover, the optimization should increase the proportion of functional antibody sequences in the ensemble, leading to an efficient search of possible sequences. Such an efficient search could produce a variety of antibodies with higher affinities than those produced by the initial immune response of an individual ( $10^{-6}$  M). Theoretically, the immune response could be fully reproduced *in vitro* by recombining (through REM or EEM) initial isolates from the naive library, thus mimicking affinity maturation to obtain very high affinity antibodies ( $< 10^{-9}$  M).

Thesis supervisor: Dr. Douglas C. Youvan  
Title: Adjunct Associate Professor of Pharmaceutical Chemistry,  
University of California San Francisco

*Pour mes parents,  
à qui je dois tout.*

## **Acknowledgements**

My years at MIT were *very* instructive. Perhaps the area where I learned the most was in understanding other people and how to deal with their behaviours in the course of trying to do your work. There is clearly more to leading a research programme than just being a good scientist: my hope is that I can reach that stage and make an honest living, evolving in a stimulating and gratifying field. I am not sure, however, that I could ever assume the roles of Thesis Advisor and Principal Investigator with quite as much kindness, effectiveness and originality as Douglas Youvan. As I "leave the nest", I can only hope that he will take pride in any successes I might have in the future and for which he can have some of the credit. If there should be failures, however, Doug is solely to blame!

My life has become much more interesting since I entered graduate school. It is also more complicated and difficult, which is why I am grateful for the friends I made during this time. In particular, Christine Goddard has become a beloved companion who adds dimension to my otherwise lopsided existence. She has brought me happiness through her innumerable qualities which I admire and love. In the Youvan lab and in the rest of the department, there were enough friends that enumerating them would be taking the risk of omission. Since I don't like to take risks, I'll only say to all those who are (or were) my friends that I thank them for the shared dinners, conversations, drinks and good times!

## **A day at MIT.**

It was a typical late November day in Cambridge Massachusetts. The air was cool, the sky was cloudy and gray and the trees had abandoned any hope of getting warmth from a sun which was past its prime, until next spring. We were sitting in the empty shell of what used to be our lab. Most of our energies had been spent that day, packing everything up for the move. I was perched on a desk by the window, my left arm wrapped around Christine's waist. On my right was the telephone which was still plugged in. Doug was sitting in one of our old beat up chairs which wasn't going to make the trip to the "other side". Opposite me and to the left of Doug was Ellen, also sitting in an old chair. Mary was present too, standing behind Doug.

I don't remember what we were talking about that day, but we were probably having one of our cynical but up-beat conversations, and laughing a little in the process. The telephone rang so I picked up the receiver because I was closest to it. "Hello?" I said. A woman's voice on the other end : "Hello, is this the MIT public relations office?". I answered politely that no, this wasn't in fact, the MIT public relations office and hung up the phone. Everyone was staring at me. I looked at them, they looked at me, and I immediately understood my mistake. We laughed a lot.

## Table of Contents

Title page .....	1
Committee page .....	2
Abstract .....	3
Dedication .....	5
Acknowledgements .....	6
A day at MIT .....	7
Table of contents .....	8
List of figures .....	10
List of tables .....	12

## Chapters

Zeroeth Iteration.	Introduction	
	0.0 Overview .....	13
	0.1 Recursive Ensemble Mutagenesis .....	13
	0.2 Exponential Ensemble Mutagenesis .....	17
	0.2.1 The Effect of the Experimental Context...	19
	0.3 Light Harvesting II protein .....	20
	0.4 Applications .....	25
	0.5 References .....	26
First Iteration.	Recursive Ensemble Mutagenesis	
	1.0 Summary .....	29
	1.1 Introduction .....	30
	1.2 Materials and methods .....	33
	1.3 Results .....	35
	1.4 Discussion .....	39
	1.5 Notes and acknowledgements .....	41
	1.6 References .....	42
Second Iteration.	Exponential Ensemble Mutagenesis	
	2.0 Summary .....	44
	2.1 Introduction .....	45
	2.2 Results .....	47



	2.3 Discussion .....	54
	2.4 Experimental protocol .....	58
	2.5 Notes and acknowledgements .....	59
	2.7 References .....	60
<b>Third Iteration.</b>	<b>Context Dependence of Phenotype Prediction and Diversity in Combinatorial Mutagenesis</b>	
	3.0 Summary .....	64
	3.1 Introduction .....	65
	3.2 Materials and Methods .....	69
	3.3 Results .....	72
	3.4 Discussion .....	81
	3.5 Notes and acknowledgements .....	84
	3.6 References .....	85
<b>Fourth Iteration.</b>	<b>Optimized Combinatorial Libraries of Antibodies, an Application.</b>	
	4.0 Summary .....	88
	4.1 Introduction .....	89
	4.2 Materials and Methods .....	92
	4.3 Results and Discussion .....	97
	4.4 References .....	101

## List of figures

	Figure	Page
Figure 0.1	Genetic Algorithm schematic diagram .....	16
Figure 0.2	Possible grouping schemes in EEM .....	19
Figure 0.3	Photons are funneled by LHII to the Reaction Center.	21
Figure 0.4	Schematic representation of LHII $\alpha$ and $\beta$ .....	23
Figure 0.5	Absorption spectrum of <i>Rb. capsulatus</i> LHII .....	24
Figure 1.1	REM flowchart .....	32
Figure 1.2	Results of REM experiment summarized by DIS .....	36
Figure 2.1	EEM strategy for mutagenesis of LHII $\beta$ subunit .....	46
Figure 2.2	EEM combinatorial cassette and nucleotide mixtures.	49
Figure 2.3	Results of EEM experiment summarized by DIS .....	51
Figure 2.4	Plot of throughput vs residues mutated .....	56
Figure 3.1	Comparison of random and phylogenetic libraries ...	67
Figure 3.2	Colour contour plot of TSM LH1 and LH2 mutants ....	74
Figure 3.3	LHII wild-type and pseudo-LH1 mutant spectra .....	77
Figure 3.4	Spectra of ten mutants of the group 1.4 random library	80
Figure 4.1	Schematic diagram of an IgG .....	89

Figure 4.2	Plasmid pDU1 .....	94
Figure 4.3	Construction of plasmid pDU1 .....	95

## List of tables

	Table	Page
Table 1.1	Sequences and phenotypes of REM mutants .....	38
Table 3.1	Sequences of TSM library mutants .....	73
Table 3.2	Sequences of pseudo-LH1 mutants from the group 1.4 random library .....	79
Table 4.1	Sample results of the optimization of the KABAT library of antibodies .....	99

## Zeroeth Iteration. Introduction

*"I have called this principle, by which  
each slight variation, if useful, is  
preserved, Natural Selection."*

Charles Darwin,  
On the Origin of the Species

### 0.0 Overview

Can a problem solving technique borrowed from the field of computational optimizations be applied to molecular genetics, potentially to achieve a protein engineering goal? This thesis will discuss the results of an investigation centered around this question. In this introductory chapter, we will look back at the origins and consider some important features of various fields of study which are the groundwork of the investigations presented in this thesis.

The next section of the introduction (0.1), leaves out much of the scientific jargon associated with biochemistry and takes an informal look at the conceptual underpinning of this thesis. In the following chapters, the experimental work is described and discussed (jargon and all). The chapters are in chronological order, reproducing the sequence of experiments that were carried out in the laboratory. Finally, the last chapter will look ahead, using what was learned as a starting point, to highlight how fertile the field of combinatorial mutagenesis has become in recent years and how it may provide a rich source of materials and insights with which to fight disease and study biomolecular phenomena.

### 0.1 Recursive Ensemble Mutagenesis.

In a science fiction story entitled "The Universal Library", written at the turn of the century by Kurd Lasswitz (reprinted in 1985), the characters describe a "Universal Library" which contains all the books that ever could be written. In this library, the complete works of Shakespeare, all editions of the New York Times as well as all theses written by graduate students of this and any other

planet of the universe could be found. This library can be constructed by randomly picking a character from a basic set (the alphabet, the digits and some punctuation marks) at regular intervals. By applying this procedure a sufficient number of times, an entire book of, say, 500 pages can be written without the usual requirements imposed on authors. Unfortunately, along with all the treasures of literature and science, the Universal library would contain staggering amounts of nonsense. So much nonsense in fact, that the impossibly large amounts of paper required to contain the entire library would make the probability of finding any intelligible text vanishingly small.

An entirely analogous situation to the one described above arises in molecular biology. Proteins are composed of an "alphabet" of 20 amino acids and three translation stop signals. Like the texts in the Universal library, proteins are of varied length and composition, within the constraints imposed by the translation apparatus of a cell. In a Universal library of protein sequences, you have your masterpieces such as the photosynthetic Reaction Center (RC) or, say, chymotrypsin and you have garbage: sequences which do not yield a molecule with a single, stable conformation and therefore could not carry out a chemical function. The daunting odds of finding "meaningful" sequences in a Universal library can actually be approached at the molecular level. Here is a realm where large numbers are readily manipulated using standard chemical techniques and the powerful property of living cells (or replicable molecules such as RNA and DNA) to make copies of themselves. RNA molecules with desired chemical properties (in our analogy, meaningful text) have already been isolated from completely random pools of RNA molecules which represent subsets of the Universal library of RNA sequences (Ellington and Szostak, 1990). Part of the reason why RNA has yielded functional sequences from random libraries has to do with the small set of "characters" with which one spells out an RNA sequence. Only four nucleotides occur in RNA. Applying a similar approach to proteins, however, is somewhat more complicated. Aside from the larger alphabet, which presents a very real statistical problem, proteins cannot be directly replicated (or copied) like RNA. It is this ease of replication which allows for the amplification of minute amounts of functional RNA sequences from a large random pool of nonfunctional molecules.

Recursive Ensemble Mutagenesis (REM) was conceived (Arkin and Youvan, 1992) as a way to isolate from a random library many different functional proteins by a more efficient sampling of the possible permutations of sequences. REM would be more efficient because it would eliminate the bulk of the sequences in the library which do not lead to a functional protein. REM would be valuable because it would provide diversity remaining inaccessible otherwise. Diverse sequences can lead to similar protein function, but occasionally some mutations will be found which produce a significant and possibly desirable departure from the original phenotype.

An interesting feature of REM is that it is reminiscent of Genetic Algorithms (GAs). For illustration purposes, it is worth describing GAs in some detail. The algorithm is similar to the procedure followed in REM and gives insight into the mechanism by which REM succeeds. GAs were invented by John Holland in 1975 (for a good overview of GAs, see Goldberg, 1989). They are meant to be efficient and adaptable optimization methods which could be used to find optimal solutions to complex, non-linear problems which defy analysis by commonly used techniques such as calculus. The remarkable adaptability of life on Earth has been attributed to Natural Selection, as first described by Charles Darwin (1859). Many features of Natural Selection can be found in Genetic Algorithms.

Let us imagine the following optimization task in which the global maximum of a function  $f(x)$  must be found in some interval (fig. 0.1). The first step is to encode potential solutions to the optimization problem (i.e., candidate values of  $x$ ) into what are sometimes called "chromosomes". These chromosomes are often binary representations of ordinary decimal numbers, in this particular example, values of  $x$ . A population of chromosomes of a convenient size is constructed. The "fitness" of each chromosome is then evaluated by determining the value of  $f(x)$ . The fitness values of individual chromosomes are used to compute the probability that each chromosome will reproduce. Individuals with higher fitnesses (higher values of  $f(x)$ ) will have higher probabilities of producing "offspring". Once a new generation of chromosomes has been produced, these binary strings are recombined. Chunks (or substrings) of zeroes and ones from two binary strings are swapped to produce offspring which differ from their parents. The last two operations

described (reproduction and recombination) are analogous to sexual reproduction in animals. The operations also introduce into the algorithm the element of randomness. Repeating the scenario described above a number of times will eventually lead to solutions to the problem which are very close to the "real" answer. In essence, a population of different potential solutions is submitted to conditions which "evolve" the population towards an ideal solution.

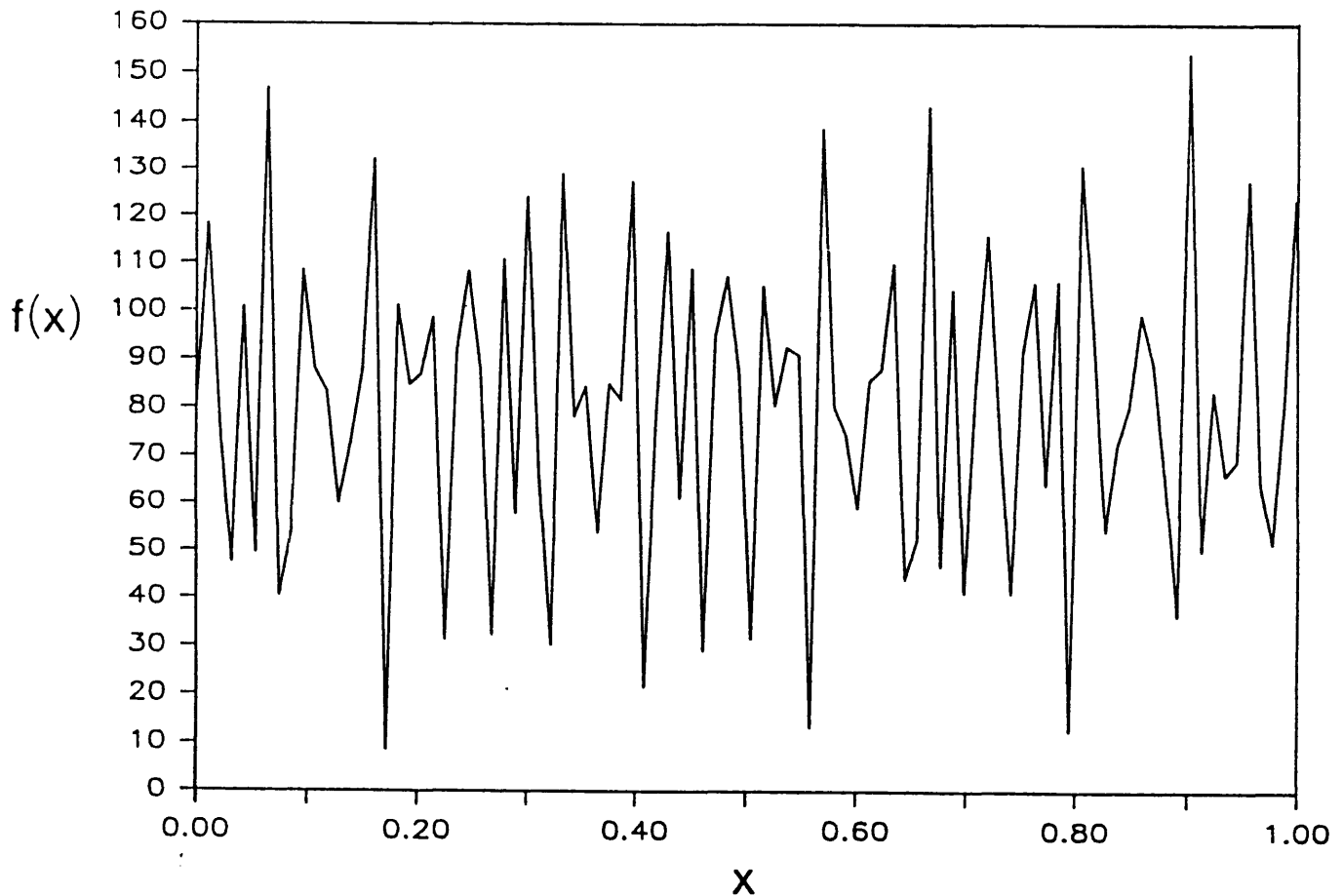


Figure 0.1. An illustration of Genetic Algorithms (GAs). It is difficult to find the global maximum ( $x_{max}$ ) of the function  $f(x)$  by conventional methods such as calculus. Representing candidate solutions to  $f(x)$  as binary numbers allows the GA method to rapidly find good approximations of



$x_{\max}$ . These binary numbers are allowed to "reproduce" if they are "fit" (higher values of  $f(x)$ ). The selected sequences are recombined to produce new recombinant sequences. These can be evaluated for fitness and allowed to reproduce. The process is repeated iteratively until a satisfactory  $x_{\max}$  is found.

Recursive Ensemble Mutagenesis depends on the construction, by molecular biological methods, of large populations of alleles of a single gene on which a selection is applied. This selection identifies individuals (or, as in the previous paragraph, "chromosomes") which produce a protein with desired properties. Determination of the sequences of some "fit" individuals then allows a recombination to be carried out chemically with a DNA synthesizer. The new generation of "offspring" are again exposed to selection. Large numbers of unique new proteins are obtained with different sequences and consequently, different phenotypes (see Iteration 1 and fig. 1.1).

REM comes full circle in that it is a biological application of a mathematical optimization method, itself inspired by a biological phenomenon, namely, Natural Selection. This is reminiscent of artificial neural networks being applied for instance, to studies of protein tertiary structure prediction (Hirst and Sternberg, 1992; Goldman et al., 1994) or used as models of cognition.

## 0.2 Exponential Ensemble Mutagenesis

Beyond the initial formulation of the method (REM) lies the implementation in a form which will be useful to the protein engineer. We are now forced to consider in greater detail some of the parameters which must be "tweaked" for this approach to work effectively. Basic biochemistry jargon is used from now on, to better convey the finer points of this discussion.

The principal requirement of REM is that functional (or fit) sequences must first be isolated from a combinatorial population of randomly altered proteins. This means that the initial population has not been "optimized" by REM to produce many functional mutants. Therefore, the initial experiment is limited in the number of amino acid residues of a protein which can be

simultaneously randomized. [By "randomized", I mean that a random codon is introduced in the nucleotide sequence such that any of the 20 amino acids and the stop codons can be encoded at that position.] Depending on the type of experiment, the exact number of residues randomized may vary, but in all cases it must be small (upper limit of 10 to 20 residues), and typically is a small fraction of the total number which make up a protein. Otherwise, the functional mutants necessary to design an optimized combinatorial library will not be obtained.

Since many residues usually contribute to the properties of a protein, it behooves a protein engineer to alter as many residues as possible simultaneously. The aspect of *simultaneous* mutation is important. For instance, imagine that we wish to improve the binding affinity of 'protein X' for some target molecule. Let us assume that the binding site in protein X spans residues 20 to 55 in its sequence and that, unbeknownst to the experimenter, residues 23, 33 and 43 must all be mutated to achieve the affinity change. Ideally we would randomize simultaneously all 36 residues, but this is not technically feasible because the odds of finding a protein X mutant with the desired phenotype are vanishingly small. A compromise can be reached by first carrying out random mutagenesis on small groups of amino acids in the binding site, independently. After these different libraries have yielded functional sequences (i.e., mutants which bind to the target molecule), a new combinatorial library can be designed where all amino acids, which were first mutated in small groups, are now allowed to change simultaneously, in a single, large group. These changes, however, are not random but limited to those mutations which have been observed to lead to functional proteins in the first round of experiments.

In essence, Exponential Ensemble Mutagenesis (EEM) expands on REM by allowing more residues to be simultaneously mutated in a combinatorial fashion. This allows the optimization described above to be applied to complex protein engineering tasks where many residues must be "tuned" to achieve a desired effect such as enhanced binding affinity.

## 0.2.1 The Effect of the Experimental Context.

Continuing with our protein engineering experiment described above, we note that the 36 residues spanning the binding site of protein X can be grouped in different ways to carry out EEM. For example, six groups of six contiguous residues (i.e., residues 20-25, 26-31, 32-37, etc...) could be mutagenized independently by constructing library 'A' where residues 20 to 25 are randomized, library 'B' where residues 26 to 31 are randomized, and so on. Alternatively, residues could be grouped based on a structural rationale such as alignment along the face of an  $\alpha$ -helix. Thus, the following four groups of 10 residues  $[(20+3.6n), (21+3.6n), (22+3.6n), (23+3.6n)]$ , where  $n$  is an integer between 0 and 9, would also cover all of the binding surface but in a different grouping scheme than the previous example (Fig. 0.2).

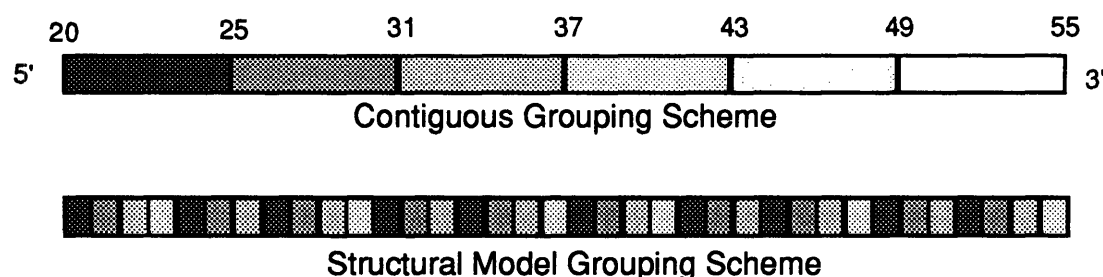


Figure 0.2. Any sequence can be broken into smaller groups of residues to be mutagenized independently. This partitioning can be done in many different ways. Two schemes (Contiguous grouping and Structural Model grouping) are shown above. Each residue is shaded according to the group to which it belongs. Residues in a group are simultaneously mutagenized.

Given these alternatives, one may ask whether the two experiments will give different results. The answer, discussed in the Third Iteration, appears to be both yes and no. First, we found that two grouping schemes may require different subsets of amino acids, at a given mutated position, to produce a

functional sequence. Second, the data suggest that "regions" of the sequence can be linked to a phenotype change, such that different schemes bring about the same phenotype change as long as these schemes span the same region of the sequence. These findings may guide the choice of grouping schemes in the future. Since a 'contiguous grouping' strategy is as capable of yielding altered phenotypes as a 'structural model' strategy, the former is entirely satisfactory when structural information is unavailable. In addition to these findings, our work presents some interesting phenotype changes in the protein used in our studies. The mechanisms of phenotype change can also be investigated by combinatorial mutagenesis.

### 0.3 The Light Harvesting II Protein.

The first goal of the work described in these pages was to experimentally verify that REM improves combinatorial libraries as predicted. The choice of an appropriate target for combinatorial mutagenesis studies was shaped by the availability, in our laboratory, of a well characterized genetic system. The investigation of the molecular events of bacterial photosynthesis had provided us with a chromogenic protein which could be easily mutagenized by molecular genetics techniques and rapidly assayed by spectroscopic means developed in our laboratory.

The Light Harvesting II protein (LHII, also referred to as light harvesting antenna or complex) of *Rhodobacter capsulatus* is an integral membrane protein which efficiently captures and funnels photons to the photosynthetic reaction center (RC). A single RC is surrounded by a small number (six) of LHI proteins and this aggregate is itself surrounded by hundreds of LHII proteins. LHII and LHI are structurally related but have different spectral properties (Youvan and Ismail, 1985). Photons captured by LHII are funneled, via LHI, to the RC where photoinduced charge separation occurs and is used to generate energy for the bacterial cell, ultimately in the form of ATP. This 'funneling' occurs by providing an energy gradient for photons to 'follow' towards the RC (fig. 0.3). To achieve this effect, different absorption wavelengths are required by the molecular components of the photosynthetic apparatus (Zuber, 1986). Within an LHII protein, carotenoids can transfer energy from shorter

wavelengths (e.g., 510 nm) to the bchls absorbing in the near infra-red (NIR) at 800 nm and 855 nm.

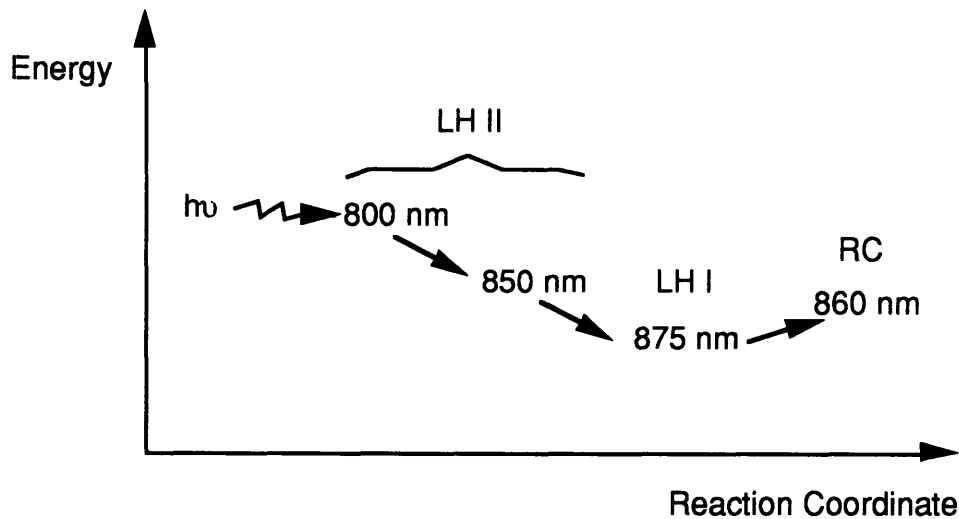


Figure 0.3. The photons absorbed by the LHII protein follow an energy gradient to the RC. Different absorption wavelengths are necessary to achieve this gradient.

The atomic structure of LHII is not known despite continuing work (Cogdell and Hawthornthwaite, 1993) on this subject. However, many details have been inferred by spectroscopic, biochemical and computational analyses (Kramer et al., 1984; Donnelly and Cogdell, 1993). These data are summarized graphically in figure 0.4. Much information has been gained by studying spectroscopically the bacteriochlorophylls (bchls) and carotenoids bound by LHII. For instance, linear dichroism data and energy transfer rates between these molecules were used to infer the distances separating them. These chromophores have also been used in the experiments described in this thesis as 'reporter groups' to rapidly assess the level of expression and structural integrity of genetically altered LHII proteins. The characteristic bands of LHII in

the NIR (fig. 0.5) were particularly useful, as described in the following chapters. Both radical and subtle changes in absorption bands in the NIR could be followed and correlated with sequence (Youvan, 1994).

Efforts are now being made (R. Niederman, Rutgers University) to understand the exact molecular mechanisms through which the mutations introduced during this work produce the observed phenotypes. For example, spectroscopic methods (e.g., Raman spectroscopy) could be used to compare the hydrogen-bonding patterns of wild-type and mutant LHII proteins in the vicinity of the chromophores. Circular dichroism and linear dichroism measurements may show changes in the orientations of the chromophores which could be correlated with certain types of mutations. Particularly, the nature of the differences between LHI and LHII may be clarified by studying LHII mutants which show spectra characteristic of LHI (see Third Iteration).

Figure 0.4. Schematic representation of LHII, composed of  $\alpha$  and  $\beta$  subunits. The core LHII protein is thought to be an  $\alpha_2\beta_2$  tetramer. This tetramer forms higher order multimers in the bacterial membrane. The subunits are thought to be  $\alpha$ -helices spanning the membrane. Wavy lines represent carotenoid pigments while squares represent bacteriochlorophylls (bchls). The arrows represent transition dipoles of the bchls ( $Q_x$  and  $Q_y$ ). The bchls responsible for the dimer band are in the upper part of the figure while the monomer band bchls are at the bottom. [Taken from Kramer *et al.*, 1984.]

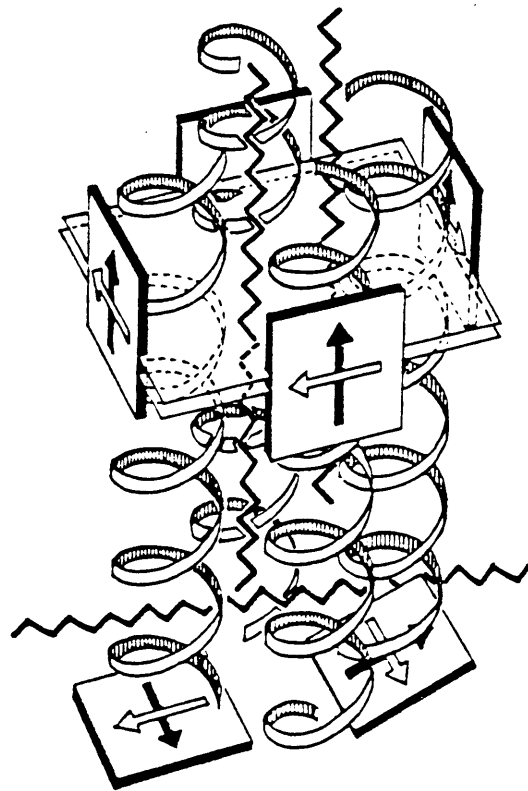
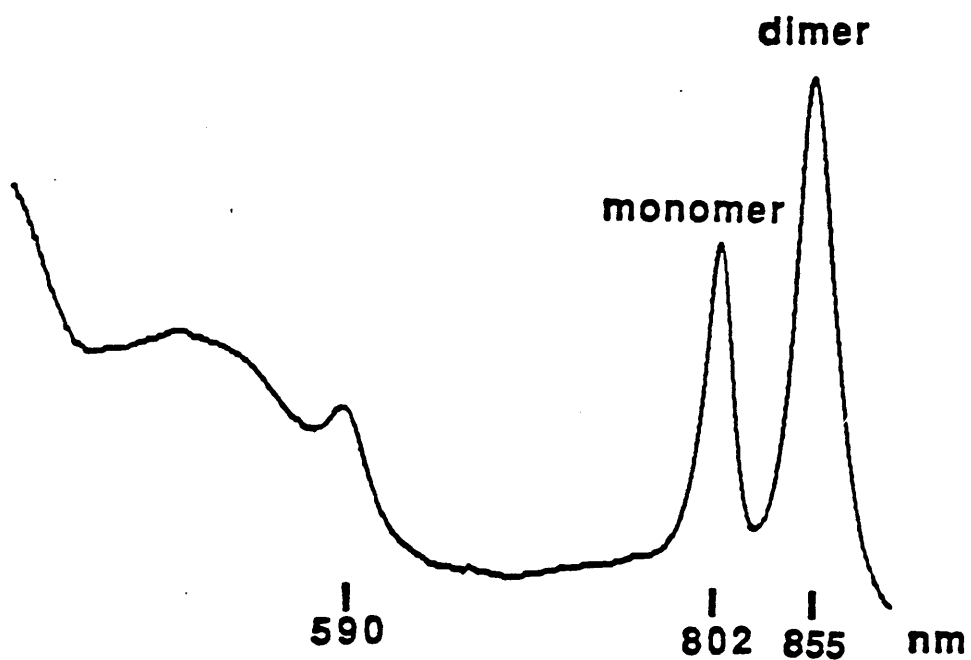


Figure 0.5. Ground state absorption spectrum of wild-type LHII. The LHII protein is intensely colored in the visible due to carotenoids. It also has two characteristic bacteriochlorophyll absorption peaks in the near infrared (NIR) at 800 nm and 855 nm. These are of great value in assaying for level of expression and structural integrity of mutant LH proteins.





## 0.4 Applications

*"A man of genius makes no mistakes. His errors are volitional and are the portals of discovery."*  
James Joyce, *Scylla and Charybdis*.

Where does all this work take us? The field of protein engineering is like an ecosystem where different methodologies evolve and fight for survival. The fittest strategies will be used by more scientists and, with time, may even recombine their best features to produce fitter 'progeny', better able to achieve the goals of protein engineering. It is possible that future offspring of the protein engineering ecosystem will have some features of the methodologies described in these pages. Our approach allows fit sequences to be recombined at the amino acid level. This search of possible sequences is efficient and can be applied to specific engineering tasks.

One such task is the engineering of antibodies, to improve affinity or change specificity. Antibodies are like our hypothetical 'protein X', described above, in that they have a binding surface we wish to alter. This surface is made of six loops with, on average, 10 amino acids per loop. The binding surface can be altered by protein engineering strategies. For example, antibodies have been isolated which neutralize HIV-1 (Barbas et al., 1992). One of these antibodies was mutagenized by an approach, similar to EEM, which recombines variant loops of high fitness (Barbas et al., 1994). The resulting library of mutant antibodies was selected for improved affinity and altered binding specificities. This demonstration of the feasibility of engineering antibodies by molecular biological methods shows the potential rewards that an improvement in mutagenesis strategies could bring.

An example of antibody engineering methodology is given in the last chapter (Fourth Iteration). It involves creating a population of different antibodies where the sequences in this population recapitulate the sequences present in a typical human's immune system. This optimized library would be

naive in that it could recognize any antigen, but it would also have a higher proportion of clones capable of binding antigens than a simple random library.

Strategies which involve introducing combinatorial diversity in protein sequences are now used often to study or engineer proteins (Glaser et al, 1992; Cormack and Struhl, 1993). The use of recombination in protein engineering is an additional dimension of combinatorial mutagenesis which has been studied by us (Goldman and Youvan, 1992; Delagrave et al., 1993; Delagrave and Youvan, 1993) and others (Lowman and Wells, 1993; Caren et al., 1994; Stemmer, 1994; Barbas et al., 1994). *De novo* design of proteins has had some success (Robertson et al., 1994) but has been limiting itself to *in vitro* synthesis for reasons of simplicity. Combinatorial mutagenesis and *de novo* design have recently joined forces to construct new proteins conveniently expressed in a biological system (Kantekar et al., 1993). In the particular case of the light-harvesting protein, combining the eventual solution of its structure with our combinatorial data and methodology could provide a basis for the *de novo* design of transmembrane proteins with new biochemical activities. This convergence of facts and ideas will almost certainly lead us to a new and unprecedented technology wherein molecules can be manipulated with exquisite precision to construct novel and useful molecular devices.

## 0.5 References

Arkin, A.P. and Youvan, D.C. (1992) *Proc. Natl. Acad. Sci. USA.*, **89**, 7811-7815.

Barbas, C.F., Bjorling, E., Chiodi, F., Dunlop, N., Cababa, D., Jones, T.M., Zebedee, S.L., Persson, M.A.A., Nara, P.L., Norrby, E. & Burton, D.R. (1992) *Proc. Natl. Acad. Sci. USA.*, **89**, 9339-9343.

Barbas, C.F. Hu, D., Dunlop, N., Sawyer, L., Cababa, D., Jones, T.M., Hendry, R.M., Nara, P.L. & Burton, D.R. (1994) *Proc. Natl. Acad. Sci. USA.*, **91**, 3809-3813.

Caren, R., Mørkeberg, R. and Khosla, C. (1994) *Bio/Technology*, **12**, 517-520.

Cogdell, R.J. and Hawthornthwaite, A.M. (1993) in *The Photosynthetic Reaction Center*. eds. Deisenhofer, J. and Norris, J.R. (Academic Press, San Diego, CA), Vol. 1, pp.23-42.

Cormack, B.P. and Struhl, K. (1993) *Science*, **262**, 244-248.

Darwin, C.E. (1966) *On the Origin of Species*. Facsimile ed. Harvard University Press, Cambridge, Mass.

Delagrave, S., Goldman, E.R. and Youvan, D.C. (1993) *Protein Eng.* **6**, 327-331.

Delagrave, S. and Youvan, D.C. (1993) *Bio/Technology* **11**, 1548-1552.

Donnelly, D. and Cogdell, R.J. (1993) *Prot. Eng.* **6**, 629-635.

Ellington, A.D. and Szostak, J.W. (1990)*Nature*, **346**, 818-822.

Glaser, S.M, Yelton, D.E. and Huse, W.D. (1992) *J. Immunol.*, **149**, 3903.

Goldman, E.R. and Youvan, D.C. (1992) *Bio/Technology*, **10**, 1557-1561.

Goldman, E.R., Füllen, G. and Youvan, D.C. (1994) *Drug devel. res.*, **33**, 125-132.

Hirst, J.D. and Sternbeg, M.J.E. (1992) *Biochemistry*, **31**, 7211-7218.

Kantekar, S., Schiffer, J.M., Xiong, H., Babik, J.M. and Hecht, M.H. (1993) *Science*, **262**, 1680-1685.

Kramer, H.J.M., Van Grondelle, R., Hunter, C.N., Westerhuis, W.H.J. and Amesz, J. (1984) *Biochem. Biophys. Acta*, **765**, 156-165.

Lasswitz, K. (1985) *Great science fiction stories by the world's greatest scientists*. Asimov, I, Greenberg, M.H. and Waugh, C.G. (Eds.) pp.73-81

Lowman, H.B. and Wells, J.A. (1993) *J. Mol. Biol.* **234**, 564-578.

Robertson, Farid, Moser, Urbauer, Mulholland, Pidikiti, Lear, Wand, DeGrado and Dutton (1994) *Nature*, **368**, 425-432.

Stemmer, W.P.C. (1994) *Nature*, **370**, 389-391.

Youvan, D.C. (1994) *Nature* **369**, 79-80.

Youvan, D.C. and Ismail, S. (1985) *Proc. Natl. Acad. Sci. USA.*, **82**, 58-62.

Zuber, H. (1986) *TIBS* **11**, 414-419.

## **First Iteration. Recursive Ensemble Mutagenesis**

### **1.0 Summary**

**We have developed a generally applicable experimental procedure to find functional proteins that are many mutational steps from wild-type. Optimization algorithms, which are typically used to search for solutions to certain combinatorial problems, have been adapted to the problem of searching the "sequence space" of proteins. Many of the steps normally performed by a digital computer are embodied in this new molecular genetics technique, termed Recursive Ensemble Mutagenesis (REM). REM uses information gained from previous iterations of combinatorial cassette mutagenesis (CCM) to search sequence space more efficiently. We have used REM to simultaneously mutagenize six amino acid residues in a model protein. As compared to conventional CCM, one iteration of REM yielded a 30-fold increase in the frequency of "positive" mutants. Since a multiplicative factor of similar magnitude is expected for the mutagenesis of additional sets of six residues, performing REM on 18 sites is expected to yield an exponential (30,000-fold) increase in the throughput of positive mutants as compared to random [NN(G/C)]<sub>18</sub> mutagenesis.**

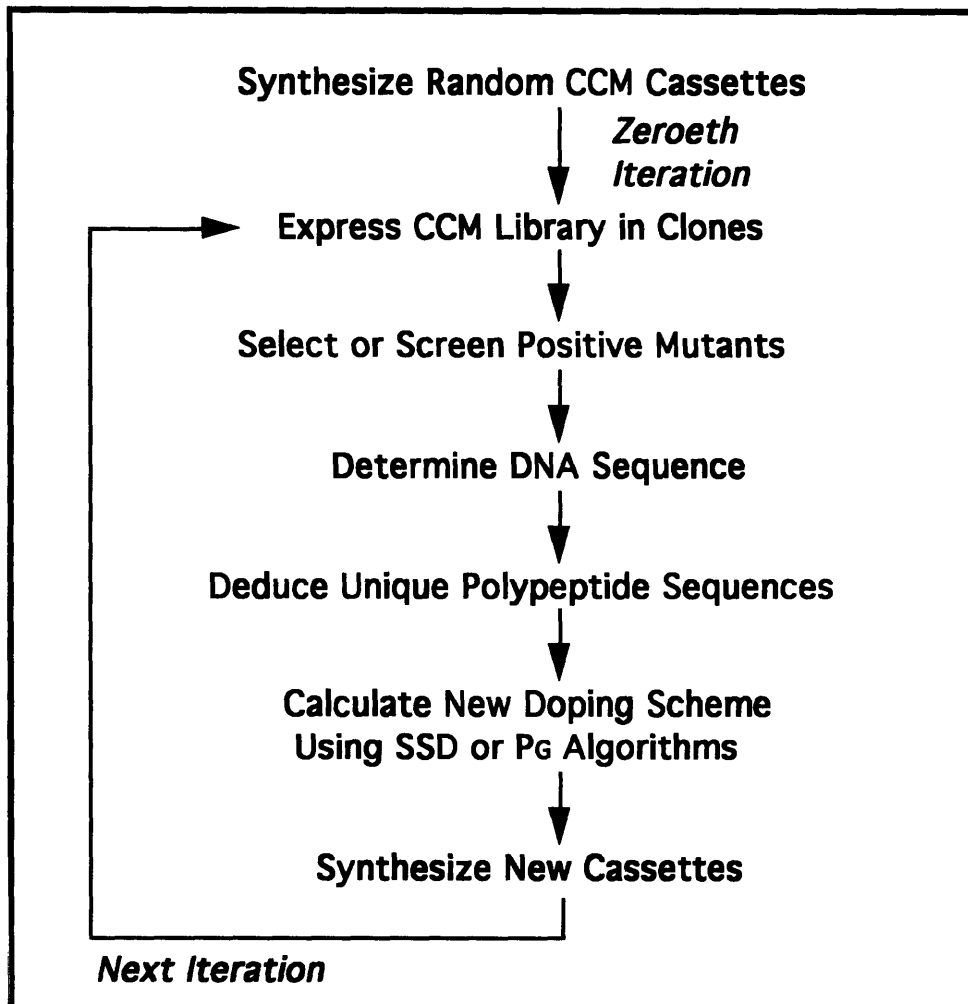
## 1.1 Introduction

Current endeavors to engineer new specificities in antibodies and their derivatives hold the promise of new therapeutic and diagnostic tools. The generation of new and informative mutant proteins is necessary to our understanding of protein structure and function relationships. Such tasks are made difficult by our inability to predict structure from primary sequence or even to predict function from structure. One strategy circumventing the gaps in our understanding involves the selection of desired phenotypes from a large pool of different genotypes, in a manner analogous to natural selection. A limitation of this is the combinatorial explosion problem: as the number of randomized (mutagenized with all 20 amino acids) sites in a protein increases, the number of possible combinations which must be evaluated to identify "positives" grows exponentially as  $20^n$ , where  $n$  is the number of sites altered. Ingenious methods have been devised to allow screening of increasingly complex libraries of mutant proteins, peptides and oligonucleotides. Phage display libraries (Smith, 1985; Hoogenboom *et al.*, 1991; Kang *et al.*, 1991) and mutagenized ribozyme populations (Beaudry and Joyce, 1992) are instances of "systems" where the genotypes and phenotypes are physically linked to allow for rapid selection and amplification of extremely complex ensembles of mutants. To completely screen a library of mutant proteins with 20 randomized amino acid residues ( $n=20$ ), the synthesis of  $20^{20}$ , or  $10^{26}$  different protein molecules is required. Obviously, this will challenge our technical capabilities for some time. It may be desirable to avoid the very high proportion of non-functional proteins in a random library and simply enhance the frequency of functional proteins, thus decreasing the complexity required to achieve a useful sampling of sequence space. Recursive Ensemble Mutagenesis (REM) is an algorithm which enhances the frequency of functional mutants in a library when an appropriate selection or screening method is employed (Arkin and Youvan, 1992a; Youvan *et al.*, 1992).

REM uses successive rounds of CCM (Oliphant *et al.*, 1986; Reidhaar-Olson *et al.*, 1991) to generate a diverse library of genetically altered proteins that fit certain selection criteria (Figure 1.1). Amino acids are retained in the library if they are found in an altered protein fitting the selection criteria. Lists of

all amino acids that are acceptable at each mutagenized position (i.e., "target sets" of amino acids) are compiled. In the next iteration of REM, combinatorial cassettes are resynthesized according to mathematical functions that bias the nucleotide mixtures (Arkin and Youvan, 1992b; Youvan *et al.*, 1992) at each mutagenized position in the protein to encode these target sets of amino acids. For example, if Ala, Ser and Thr occur at a given position in different selected mutants, these amino acids constitute the target set at that position. A mathematical function is used to select the best "dope" that maximizes the probabilities of the amino acids in the target set. The next cassette is then designed such that this target set is encoded by a simple mixture of nucleotides at that codon (e.g.: [(G,A,T)(C)(G,C)] ). In certain cases, where there is a good match between selection criteria and structure inherent in the genetic code (Sjostrom and Wold, 1985; Youvan, 1991) like hydrophathy and molar volume, computer simulations predict that multiple iterations of REM will yield thousands of times more mutants than conventional CCM (Arkin and Youvan, 1992b; Youvan *et al.*, 1992).

Fig. 1.1 (Next page) REM involves the recursive use of combinatorial cassette mutagenesis (CCM). The first step of REM begins by expressing and screening a CCM library. Two or more "positive" mutants are then picked and sequenced. [Positive mutants are defined in the current experiment as binding significant levels of red-shifted Bchl which is characteristic of LHII assembly.] Next, a list of unique protein sequences is determined by translating these DNA sequences. A "unique sequence" is defined at the protein level. If more than one protein has the same sequence, only the first occurrence of this sequence is retained and counted as unique. For each mutagenized position in the protein, a target set of acceptable amino acids is compiled and the most appropriate dope is determined by a mathematical function such as group probability (PG). The next iteration of REM proceeds by using these "intelligent" dopes to generate a combinatorial cassette of lower complexity. In order to take advantage of the properties of REM, the complexity of the possible peptide sequences arising from CCM should be shown to be in vast excess of the screening size (Youvan *et al.*, 1992).



As a model system to experimentally verify the computer predicted amplification by REM, the light harvesting II (LHII)  $\beta$  subunit gene (Youvan and Ismail, 1985) of *Rhodobacter capsulatus* was chosen. The LH II protein has two characteristic absorption bands in the near infrared (800 and 858 nm) that are red shifted relative to protein-free bacteriochlorophyll (Bchl) absorption at 770 nm. These prosthetic groups serve as colorimetric indicators of protein expression and subunit assembly. Six carboxy-terminal residues of the  $\beta$  subunit were initially mutagenized by construction of a combinatorial cassette containing the sequence  $[NN(G,C)]_6$ , where 'N' designates an equiprobable



mixture of all four nucleotides. This CCM library was conjugated into a strain of *Rb. capsulatus* (U71) totally deficient in Bchl-binding proteins, or any other compounds with significant absorption in the near infrared (Youvan *et al.*, 1985). This deletion background facilitates the use of Digital Imaging Spectroscopy (DIS) (Arkin *et al.*, 1990; Arkin and Youvan, in press) to screen thousands of colonies directly on petri dishes for LHII expression. We then sequenced five functional mutants and used this limited data to construct a new CCM library. The frequency of positives was increased 30-fold relative to the original library.

## 1.2 Materials and Methods

### *Plasmids and strains*

Plasmid pU4b is a shuttle vector used for cassette mutagenesis as well as expression of the mutant LHII genes (Goldman and Youvan, in press). M13 was our vector for single-stranded sequencing and was propagated in *E. coli* MV1190. *E. coli* strain S17-1 was used for library construction and conjugation with *Rb. capsulatus* U71. For expression of the libraries, *Rb. capsulatus* U71, an LHII chromosomal deletion background (LHI and RC expression inactivated by a point mutation) was used.

### *Materials and DNA manipulations*

DNA manipulations were essentially performed as described by Sambrook *et al.* (1989). Restriction enzymes were obtained from New England Biolabs, T4 DNA Ligase was from Bethesda Research Labs as was Taq polymerase. Sequencing was carried out using a Sequenase kit from United States Biochemicals. Electroporation was carried out in 0.2 cm cuvettes on 0.45 mL of competent cells using a Bio-Rad electroporator according to instructions provided. All oligonucleotides were synthesized on an Applied Biosystems model 381 DNA synthesizer using commercially available reagents.

### *Library construction*

The unique KpnI and XhoI sites of pU4b flank the region encoding the dimer Bchl binding site and the carboxy-terminus of the  $\beta$  subunit LHII gene.

These restriction sites were engineered to allow double-stranded combinatorial cassettes to be subcloned in place of the wild-type sequence.

The sense strand of the 113mers, which included the KpnI-XhoI sites, as well as two PCR primers (20mers each spanning a restriction site) were synthesized. The doped sequence within the cassette used in the zeroth iteration was: [NN(G,C)]<sub>6</sub>. The purified 113mer was amplified by PCR. Amplified double-stranded cassette was then purified by phenol extraction and ethanol precipitation. Complete digestion of cassette with KpnI and XhoI is carried out in a single incubation. The digested cassette is then purified by phenol and ether extractions and ultrafiltration in a Centricon 30 device (Amicon).

Ligation is carried out for 24 hours at 16 °C in 20 µL with approximately 0.1 µg of pU4b similarly digested with KpnI and XhoI. The resulting pU4b derivatives (an aliquot of the ligation) is directly electroporated into S17-1 *E. coli*. Aliquots of the transformation are plated on LB-tetracycline plates (after allowing one hour for resistance expression) for complexity estimation and the remainder of the transformation is incubated over night in 60 mL of LB-tetracycline. Plasmid pU4b derivatives were conjugated from *E. coli* S17-1 donors into *Rb. capsulatus* strain U71. The library is expressed by U71 transconjugants selected for by growth on RCV-tetracycline plates at 32°C.

#### *Dope optimization*

In computer simulations, various functions were used to optimize the "nucleotide mixtures". In this work, only five functional mutant sequences were obtained in the zeroth iteration. Given this small number of sequences, and in order to conserve diversity, we elected to use the group probability (P<sub>G</sub>) function because it retains all amino acids in the target set. When presented with a target set at one position, the program "CyberDope" (provided courtesy of KAIROS, Cambridge, Ma, USA) goes through all integer nucleotide mixtures possible for a codon and evaluates for each mixture the value of P<sub>G</sub> :

$$P_G = \prod_i P_D[i] \quad \text{Eq. 1}$$

where  $P_D[i]$  is the frequency of occurrence of the  $i$ th amino acid (in a target set of  $i$  amino acids) as encoded by a specific triplet dope. For the hypothetical target set mentioned above (Ala, Ser, Thr), any mixture not encoding a member of the target set (e.g.:  $P_D[\text{Ala}] = 0$ ) will cause  $P_G$  to be zero. The mixture with the highest value of  $P_G$  will be selected for the dope at that position. The doped sequence within the cassette used in the first iteration of REM was:

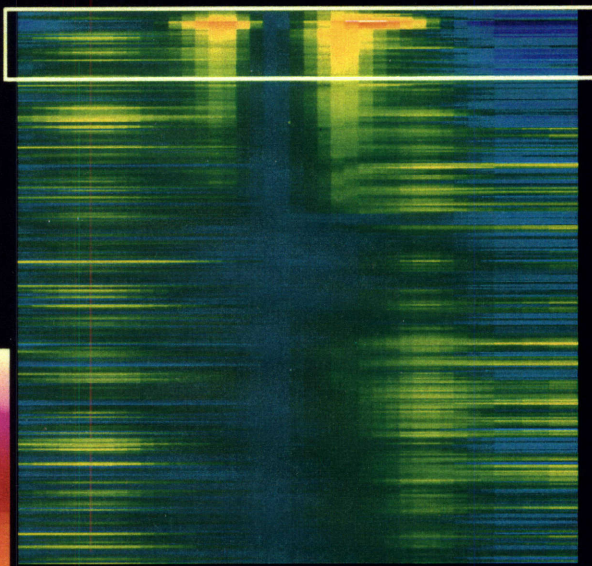
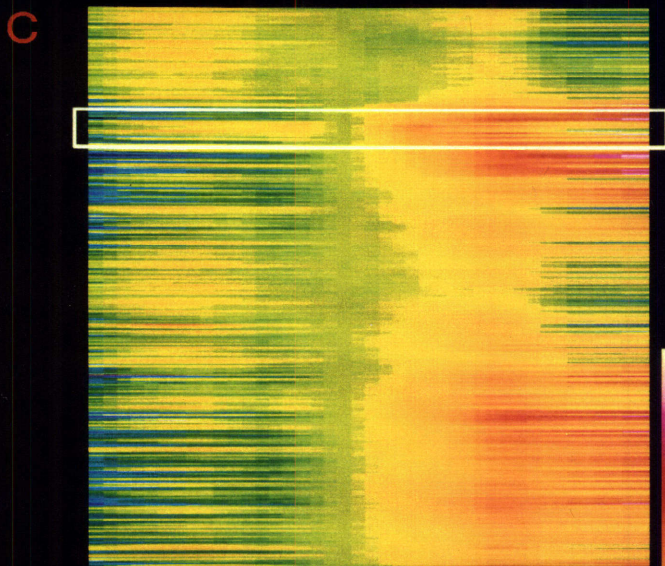
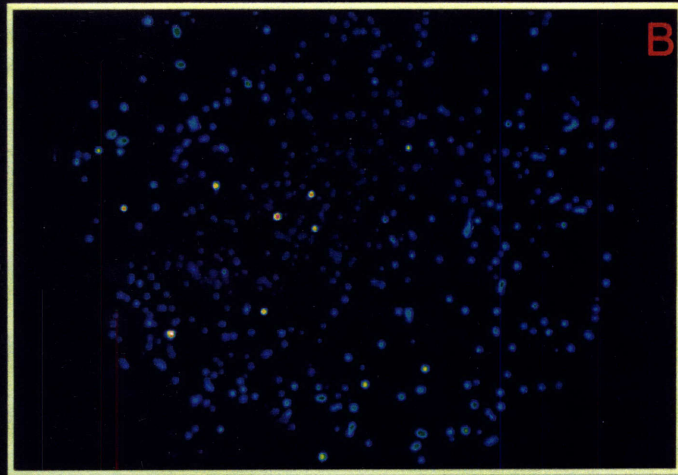
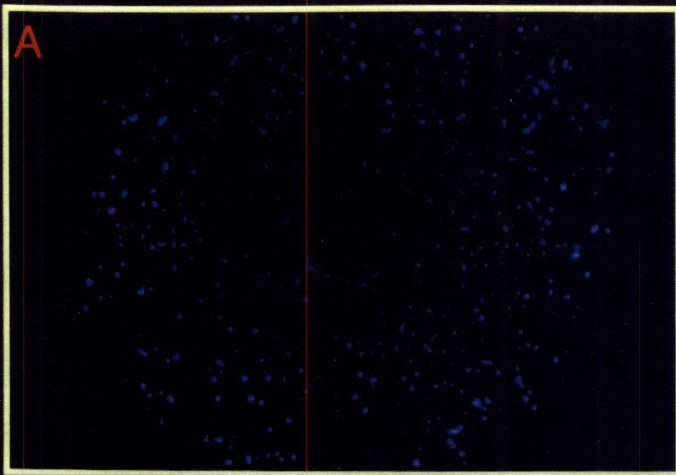
[(G,T)(C,T)(C,G)][(A,G,T)(C,T)(C,G)][(C,T)(C,G)G][(C,T)(C,G)G][(A,G,T)(G,T)G] [(C,T,G)(C,T,G)C].

### *Imaging spectroscopy*

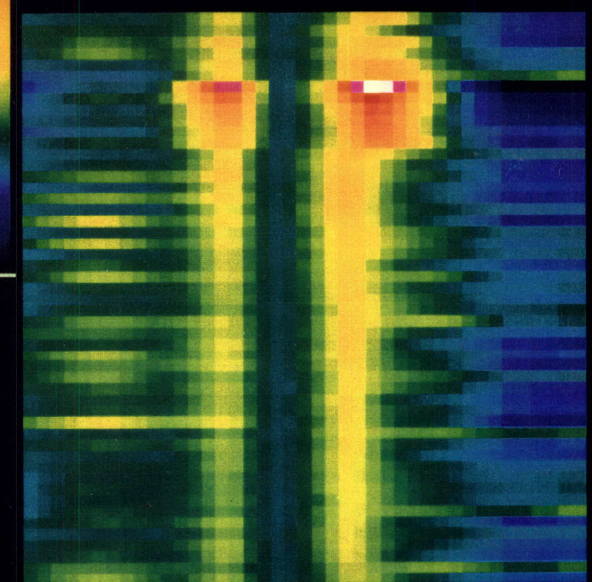
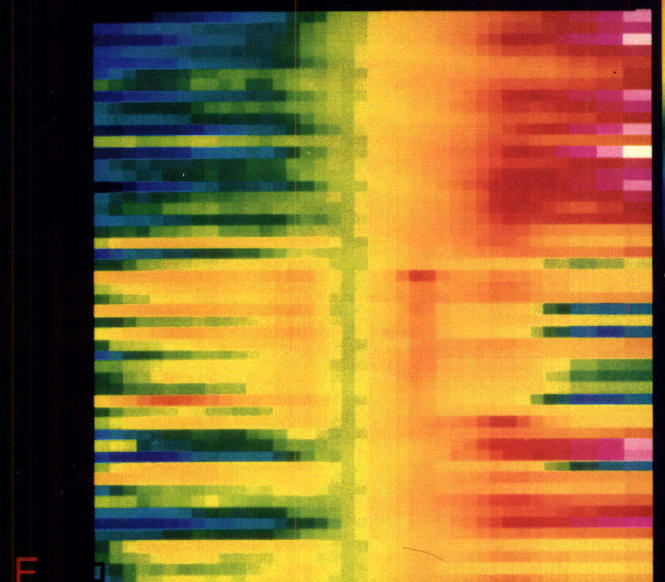
Colonies were imaged as spreads on RCV-tetracycline plates from the bacteria resuspended after conjugation. The most recent configuration of the digital imaging spectrophotometer has been described (Arkin and Youvan, in press). For the fluorescence images, the petri dishes were illuminated with broad-band blue-green light and an 830 nm longpass filter was placed in front of the CCD lens to obtain radiometrically calibrated monochrome images which were linearly mapped to pseudocolors after establishing the low and high grayscale values for both images.

## **1.3 Results**

The experimental complexity (i.e., number of independently generated clones) of the "zeroeth iteration"  $[\text{NN}(\text{G,C})]_6$  library was approximately 45,000. The theoretical complexity of such a library at the nucleotide level is calculated as  $32^6$  ( $= 1.1 \times 10^9$ ) because there are 32 possible  $[\text{NN}(\text{G,C})]$  codons; the experimental complexity is only a small fraction of this number. Preliminary screening used fluorescence, (Yang and Youvan, 1988) which is indicative of LHII assembly, to rapidly identify mutants expressing LHII. Mutants are then more closely evaluated by ground state absorption measurements using DIS. We observed a low frequency of highly fluorescent colonies in the zeroeth iteration of REM (ca. one positive mutant in 10,000 colonies screened). Relative to wild-type absorption, DIS showed a decrease in the optical density at 800 and 858 nm for these few positives.



C  
O  
L  
O  
N  
Y  
  
N  
U  
M  
B  
E  
R



730 800 850 930 730 800 850 930

WAVELENGTH (nm)

Fig. 1.2. (Previous page) Fluorescence screening and DIS contour maps showing REM amplification of mutants expressing LHII. Panels A, C, and E (left side) and panels B, D, and F (right side) correspond to the zeroeth and first iterations of REM, respectively. Fluorescence assays (panels A and B) were pseudocolored to aid in the visualization of relative levels of fluorescence according to the color bar shown below. These show brighter and more numerous fluorescent colonies in the first iteration of REM (panel B) than in the zeroeth iteration of REM (panel A). DIS color contour maps (panels C and D) of the ground state absorption spectra of the same petri dishes shown in panels A and B confirm the presence of mutants expressing Bchl-binding proteins. Each horizontal line represents the color coded absorption spectrum of a single colony; the color bar spans OD values of 0.0876 to 0.1177 for panels C and E and 0.0927 to 0.2088 for panels D and F. The spectra were sorted (Arkin and Youvan, in press) according to their absorption at 850 nm. 897 spectra from the zeroeth iteration plate and 451 spectra from the first iteration plate are displayed in panels C and D, respectively. Panels E and F highlight the spectra of 50 colonies from panels C and D (within the white boxes), respectively. Panel F shows many more colonies with spectra characteristic of LHII than found in panel E.

Because of their rarity, only five positives were obtained from the zeroeth iteration of REM. Four of these five mutants fit the selection criterion of displaying significant absorbance at 858 nm and another, REM0.10, had an interesting phenotype. The five positives were repurified and sequenced (Table I). The composition of a first iteration cassette was calculated by the computer program "CyberDope", which generates DNA dopes that maximize the overall probability of the target set. To add diversity to the target set, the wild-type sequence was also included. Therefore, while not taking frequency of occurrence into account because of the small sample size, for the first doped position the target set is F, S, A, L. The output of "CyberDope" at the nucleotide level gave the codon [(G,T)(C,T)(C,G)], which encodes amino acids A, S, V

(0.25 probability of occurrence for each) and F, L (0.12 probability of occurrence for each). Valine is unavoidably encoded by this dope because of the structure of the genetic code.

TABLE 1.1 Sequences and corresponding phenotypes of mutants isolated from the zeroeth and first iterations of REM

Mutant	Deduced sequence	OD 800nm	OD 855nm	$\frac{OD855}{OD800}$
Wild-Type	ATPWLG	0.26	0.39	1.5
REM0.6	LTPWVA	0.15	0.24	1.6
REM0.7	LTPWVP	0.13	0.20	1.5
REM0.8	ASPWMS	0.09	0.15	1.7
REM0.9	SSPWLP	0.15	0.22	1.5
REM0.10	FVWPGF	0.05	0.09*	1.8
REM1.1	STPWVF	0.11	0.17	1.5
REM1.2	FTPWVG	0.11	0.18	1.6
REM1.3	ATPWLA	0.10	0.15	1.5
REM1.4	STPWLA	0.33	0.48	1.5
REM1.5	LTPWGR	0.09	0.13	1.4
REM1.6	VTPWLF	0.11	0.18	1.6
REM1.7	VTPWLG	0.13	0.21	1.8
REM1.8	LIWPVL	0.05	0.09*	1.8
REM1.9	ALWPLV	0.05	0.09*	1.8
REM1.10	LTPWGG	0.20	0.29	1.5
REM1.11	VTPWVR	0.06	0.11	1.8
REM1.12	VTPWGL	0.12	0.21	1.8

\*Peak shifted to 845nm.

Figure 1.2 demonstrates the amplification properties of the REM methodology as assayed by digital imaging spectroscopy using both fluorescence emission and ground state absorption imagery. The first iteration of REM yields a 30-fold increase in the frequency of enhanced fluorescence mutants (Figure 1.2, panels A and B). As compared to zeroeth iteration REM data, DIS analysis of the first iteration library shows both an increase in the percentage of positive mutants (i.e. throughput) and an increase in protein levels as determined by the intensity of the Bchl absorption bands (Figure 1.2 and Table 1).

Twelve positive mutants were sequenced from the first iteration REM library. All of these mutants express unique peptide sequences that differ from wild-type. Two mutants (REM1.8 and REM1.9) show a 10 nm blue shift in the 858 nm band. These blue-shifted mutants have an inversion of the Pro-Trp motif found in all 29 sequences in the known phylogeny (Zuber, 1990) of  $\beta$  subunits. This phenotype was first observed in the zeroeth iteration library (mutant REM0.10) but now finds itself amplified in the first iteration. Note also that REM0.10 contains the same Pro-Trp motif inversion.

#### **1.4 Discussion**

To show that computer simulations were accurate in their prediction of an increased throughput of positives, an LHII gene was iteratively mutagenized at its six carboxy terminal residues. From the zeroeth iteration (CCM) data, target sets of amino acids were defined. A computer encoded algorithm generated a doped oligonucleotide which best represented the target set at each mutagenized position. Expression of this new library (the first iteration of REM) revealed a substantial amplification in the throughput of pseudo wild-type mutants. From the zeroeth iteration library where roughly 10,000 colonies were screened to identify one positive, we can now conveniently identify a new positive by screening only about 300 colonies. This corresponds to a 30-fold increase in overall throughput, suggesting that mutagenizing 18 sites of similar stringency would yield a  $30^3$  or 27,000-fold increase in throughput over random mutagenesis using [NN(G,C)]<sub>18</sub>.

The altered proteins obtained by combinatorial mutagenesis are not necessarily trivial variations of the wild-type sequence. An inversion of a completely conserved motif was observed in some mutants. Therefore, the sequence data indicate that REM does not recapitulate the known phylogeny. Mechanistically, the simultaneous (experimental) randomization of six sites in a protein may have no analogy in Nature.

In this work, experimental evidence is given that REM allows an efficient search of sequence space by producing mutant libraries with increased frequencies of selected "positives". Due to the high stringency of the region chosen for mutagenesis, only a small sequence database was available for the construction of the first iteration dope. In systems where large complexities can be achieved easily (e.g.: phage display libraries), more sites can be mutagenized at once and more positives isolated, giving a more complex sequence database. As a consequence, other dope optimizing equations (Youvan *et al.*, 1992) could be used which would be better suited to yield large increases in throughput. Alternatively, different short stretches of amino acids could be randomized and the zeroeth iteration data from these libraries pooled to produce a first iteration dope mutagenizing many more sites than ordinarily possible with CCM.

It is important to make the connection between our algorithmically-based doping schemes and protein engineering projects where CCM is currently being used. REM decreases the fraction of null mutants in the population, therefore more sites can be simultaneously mutagenized. Model experiments on LHII can be used to optimize REM methodology, including the nucleotide doping equations. While DIS is limited to screening about  $10^6$  colonies, phage display libraries (Smith, 1985; Hoogenboom *et al.*, 1991; Kang *et al.*, 1991) can be used to select mutants from libraries with complexities exceeding  $10^9$ . Based on our preliminary experiments, we expect greater phenotypic diversity after one iteration of REM. This means that stronger "binders" can be isolated, which is the fundamental goal of the phage display methodology. The use of CCM to introduce additional diversity in antibody libraries has already proven a useful approach (Barbas *et al.*, 1992) and may well be enhanced by the use of our mutagenesis scheme. REM is the first optimization technique that can be



used to address this problem and explore sequence space in a mathematically rigorous fashion.

### **1.5 Notes and acknowledgments**

This iteration was published in Protein Engineering (Delagrave, S., Goldman, E.R. and Youvan, D.C., **6**, 327-331, 1993).

## 1.6 References

Arkin, A., Goldman, E., Robles, S., Coleman, W., Goddard, C., Yang, M. and Youvan, D.C. (1990) *Bio/Technology*, **8**:746-749.

Arkin, A.P. and Youvan, D.C. (1992a) *Proc. Natl. Acad. Sci. U.S.A.*, **89**:7811-7815.

Arkin, A.P. and Youvan, D.C. (1992b) *Bio/Technology*, **10**:297-300.

Arkin, A.P. and Youvan, D.C. (In press) In Deisenhofer, J. and Norris, J.R. (eds.) *The Photosynthetic Reaction Center*. Academic Press, New York.

Barbas, C.F., Bain, J.D., Hoekstra, D.M. and Lerner, R.A. (1992) *Proc. Natl. Acad. Sci. U.S.A.*, **89**:4457-4461.

Beaudry, A.A. and Joyce, G.F. (1992) *Science*, **257**:635-641.

Hoogenboom, H. R., Griffiths, A. D., Johnson, K. S., Chiswell, D. J., Hudson, P. and Winter, G. (1991) *Nucleic Acids Res.*, **19**:4133-4137.

Kang, A.S., Barbas, C.F., Janda, K.D., Benkovic, S.J., and Lerner, R.A. (1991) *Proc Natl. Acad. Sci. U.S.A.*, **88**:4363-4366.

Oliphant, A. R., Nussbaum, A. L. and Struhl, K. (1986) *Gene*, **44**:177-183.

Reidhaar-Olson J. F., Bowie J. U., Breyer, R. M., Hu, J. C., Knight K. L., Lim W. A., Mossing M. C., Parsell D. A., Shoemaker K. R. and Sauer R. (1991) *Methods in Enzymology*, **208**:564-587 .

Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) *Molecular Cloning, a Laboratory Manual*. Cold Spring Harbour Laboratory Press, Cold Spring Harbour, NY.

Sjostrom, M. and Wold, A. (1985) *J. Mol. Evol.*, **22**:272-277.

Smith, G. P. (1985) *Science* **228**:1315-1317.

Yang, M. M. and Youvan, D. C. (1988) *Bio/Technology* , **6**:939-942.

Youvan, D.C., Arkin, A.P. and Yang, M.M. (1992) Maenner,R. and Manderick, B. (eds.) *Parallel Problem Solving from Nature, 2*, Elsevier Publishing Co, Amsterdam, 401-410.

Youvan, D.C. (1991)*Trends in Biochemical Sciences*, **16**:145-149.

Youvan, D.C. and Ismail, S. (1985) *Proc Natl. Acad. Sci. U.S.A.*, **82**:63-67.

Youvan, D.C., Ismail, S. and Bylina, E.J. (1985) *Gene*, **38**:19-30.

Zuber, H. (1990) In Drews, G. and Dawes,E.A. (eds.) *Molecular Biology of Membrane-Bound Complexes in Phototrophic Bacteria*. Plenum Press, New York, pp .161-180.

## **Second Iteration. Exponential Ensemble Mutagenesis**

### **2.0 Summary**

**An efficient method for generating combinatorial libraries with a high percentage of unique and functional mutants is described. Combinatorial libraries have been successfully used in the past to express ensembles of mutant proteins in which all possible amino acids are encoded at a few positions in the sequence. However, as more positions are mutagenized the proportion of functional mutants is expected to decrease exponentially. Small groups of residues were randomized in parallel to identify, at each altered position, amino acids which lead to functional proteins. By using optimized nucleotide mixtures deduced from the sequences selected from the random libraries, we have simultaneously altered 16 sites in a model pigment binding protein: approximately one percent of the observed mutants were functional. Mathematical formalization and extrapolation of our experimental data suggests that a  $10^7$ -fold increase in the throughput of functional mutants has been obtained relative to the expected frequency from a random combinatorial library. Exponential ensemble mutagenesis should be advantageous in cases where many residues must be changed simultaneously to achieve a specific engineering goal, as in the combinatorial mutagenesis of phage displayed antibodies. With the enhanced functional mutant frequencies obtained by this method, entire proteins could be mutagenized combinatorially.**

## 2.1 Introduction

Rather than attempt to rationally engineer new properties in biopolymers in a serial fashion, a parallel approach can be taken which circumvents the gap in our understanding of structure-function relationships. In the parallel approach, large populations of molecules are constructed by random mutagenesis or by chemical synthesis. A selection or screen can be employed to search for polymers with desired properties. This approach has taken various experimental forms and is conceptually analogous to biological processes such as natural selection or the immune response. Large populations of genetically altered proteins can be produced by cassette mutagenesis (Oliphant *et al.*, 1986; Reidhaar-Olson *et al.*, 1991) and have been used, for example, to isolate antibody fragments (Fab's) with high affinities for arbitrary compounds (Gram *et al.*, 1992; Barbas *et al.*, 1992). Similarly, libraries of peptides and other proteins have been screened, yielding new molecules with desirable properties (Lam *et al.*, 1991; Houghten *et al.*, 1991; Roberts *et al.*, 1992; Lowman *et al.*, 1991). The association of recombinant proteins to filamentous phage (phage display) is currently one of the most efficient methods for generating large combinatorial libraries, facilitating the parallel synthesis of up to  $10^9$  different proteins (Smith, 1985; Hoogenboom *et al.*, 1991; Kang *et al.*, 1991).

The simultaneous mutagenesis of a larger number of amino acid residues may increase the probability of generating a useful altered protein, either because of the need for an extensive reorganization of binding sites with large surface areas (as in antibody complementarity determining regions, CDRs) or because the exact point mutation necessary to obtain the desired phenotype is difficult to predict. However, as more positions in a protein are randomized, the number of possible mutants increases exponentially as  $20^n$ , where  $n$  is the number of positions altered. As a zeroeth order approximation, we propose that the frequency ( $f$ ) of functional mutants decreases exponentially as:

$$f = x^n \qquad \text{Eq. 1}$$

where  $x$ , a fractional value, is an overall measure of the stringency of the mutagenized residues (defined below). A consequence of our "exponential hypothesis" is that after some number of altered sites is exceeded, it becomes



Figure 2.1. (Previous page) By partitioning the protein into smaller segments, EEM facilitates the mutagenesis of more sites than would be possible using conventional combinatorial schemes. Shown at the bottom is the amino acid sequence of the LHII b subunit which binds monomeric and dimeric bchls through histidine ligands. Spectroscopic signals from these bchl molecules can be used as reporters to assay mutant colonies for functional proteins expressed, *in situ*. Each line above the protein sequence represents one mutagenesis experiment of a group of residues. Underlined are the experiments which were carried out and are discussed in this paper, the others are shown for completeness. Dots represent sites which are not mutagenized and letters represent mutagenized sites. The letter X represents a randomized site, where all amino acids are encoded (32 codons: [NN(G,C)] ). Level 1 includes the nine random mutagenesis experiments (groups 1.1 to 1.9) necessary to span the entire polypeptide. The letter Y represents a site where a target set (subset of all amino acids) derived from the results of level 1 experiments is expressed. Similarly, level 2 is used to design the group 3.1 library, wherein the combinatorial complexity per site is further restricted. The letter Z represents a site where a target set was derived from the results of level 2 experiments. In terms of the number of amino acids allowed in each target set,  $X \geq Y \geq Z$ . Nevertheless, the complexity (possible number of different sequences) of each library remains sufficiently high that the diversity in the population is greater than what can be constructed and assayed.

## 2.2 Results

Figure 2.1 shows how an EEM strategy can be applied to a small protein. The first "level" of mutagenesis (groups 1.1 to 1.9) can be carried out in parallel using randomized codons (i.e., [NN(G,C)] ) over manageable groups of 5 or 6 amino acids each. In the work described here, libraries corresponding to groups 1.1, 1.2 and 1.3 have been constructed, and then combined in group 2.1, the EEM library. Continuing with the EEM strategy, the entire polypeptide

could, in theory, be mutagenized to produce a group 3.1 library, in which all the sites have been subjected to random combinatorial mutagenesis.

Three contiguous groups, spanning a total of 16 residues, at the carboxy terminal end of LHII  $\beta$  were randomized independently (Fig.2.1). All the steps involved in the production of a library have been described (Goldman and Youvan, 1992; Delagrave *et al.*, 1993) and are summarized in the *experimental protocol* section. From the group 1.2 and group 1.3 libraries, the frequency of "positive" mutants observed by DIS was approximately  $10^{-2}$  and  $10^{-3}$ , respectively. Randomizing group 1.1 resulted in a lower frequency of positive mutants (ca.  $10^{-4}$ ). This group was carried through one iteration of recursive ensemble mutagenesis [or REM, see Delagrave *et al.*, 1993 (the first iteration) and Arkin and Youvan, 1992a] which yielded a library with a positive frequency of approximately one in 300.

Sequences of positive mutants isolated from the level-1 libraries (i.e., groups 1.1, 1.2 and 1.3) were used to design the optimized nucleotide mixtures (i.e., the reduced complexity "dopes") shown in Figure 2.2. Twenty-three group 1.1 mutants, 11 group 1.2 mutants, and 12 group 1.3 mutants were used to construct 16 different target sets (i.e., the sets of amino acids found to occur in functional mutants at each position). These target sets can be thought of as an artificial phylogeny which provides information as to the amino acid requirements of each site in the protein (Goldman and Youvan, 1992).



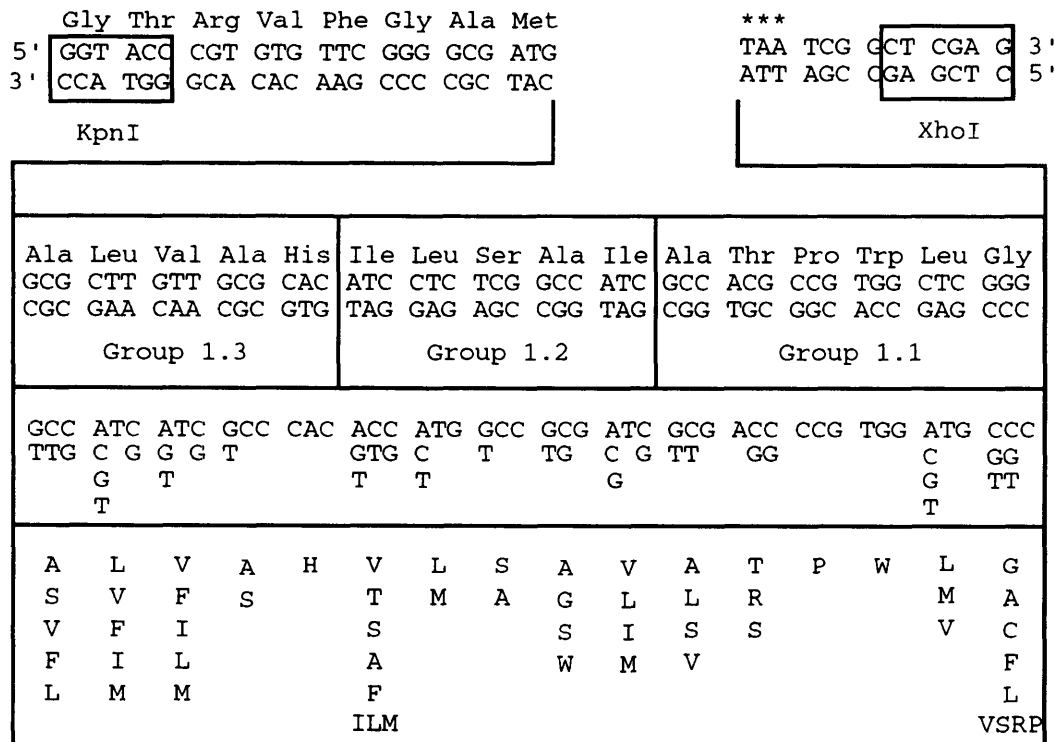


Figure 2.2. An optimized combinatorial cassette was designed for the construction of the group 2.1 EEM library. Shown at the top of the figure is the wild-type nucleotide sequence of the double-stranded cassette, with the restriction sites used for cassette insertion (Kpn I and Xho I). Above the nucleotide sequence are the encoded amino acids. The three groups mutagenized in this work have been expanded for clarity. Below each amino acid position mutagenized are the nucleotide dopes of the group 2.1 library and the amino acids which are encoded by such a mixture. For instance, the second codon of group 1.3 encodes leu in the wild-type protein but in the doped cassette the first codon position is allowed to be any nucleotide while the second and third positions of the codon are restricted to (T) and (G,C), respectively. As a result L, V, F, I, or M are encoded at this position by the EEM cassette.

Using a commercially available DNA synthesizer, integer mixtures of nucleotides can be easily prepared. Software has been written which evaluates all possible (3375 unique) integer nucleotide mixtures for a codon. A target set from the sequence database is then compared by the program

"CyberDope" with the amino acids encoded by each codon mixture. Two mathematical functions have been developed (Arkin and Youvan, 1992a; Arkin and Youvan, 1992b) to quantify how well a codon mixture represents the target set: 1) group probability ( $P_G$ ), or 2) sum-of-squares-of-differences (SSD). The use of these functions is critically important as they have been designed to provide nucleotide mixtures directly derived from the data with minimal intervention from the investigator, thereby reducing chances of inconsistencies in the formulation of these mixtures. Given the large sequence database at our disposal, SSD was used because it weights the frequency of occurrence of amino acids found in the functional mutants, which contributes to the increase in throughput of positive mutants (Youvan *et al.*, 1992).

The expression of the 16-site, group 2.1 library showed a high frequency of positive mutants; more than one percent of these mutants assembled a functional protein despite the number of sites altered (Fig. 2.3). Twenty-one of these positives were isolated and their sequences in the altered region were determined (Fig. 2.3 C). On average, six amino acids are found to be conserved out of the 13 mutagenized. Some mutants have as few as 4 amino acids conserved (# 5) and some, as many as 8 (#31). The spectral characteristics of the mutants are typical of LHII, with substantial diversity in the intensities of the absorption bands. However, the ratio of absorbance between the long wavelength and short wavelength bands is relatively constant. Small but reproducible shifts (5 to 10 nm) are observed in the 858 nm peaks of many mutants (Fig. 2.3 C).



Figure 2.3. (Previous page) Digital imaging spectroscopy of the group 2.1 library shows a high frequency of unique functional LHII mutants despite the number of sites altered. Panel A shows an absorption image at 855 nm of a typical petri plate on which an aliquot of the library was spread. The colonies expressing higher amounts of LHII appear darker (bottom of color bar) in the image. Panel B is a fluorescence image of the same plate as in panel A, showing the level of LHII expression; false coloring encodes higher fluorescence as a brighter color (top of color bar). In panel C, spectra between 730 nm and 930 nm of spots of LHII mutants were obtained and shown on the right as contour maps where each horizontal line represents the color coded absorption spectrum of a single mutant. The color bar spans OD values of 0.091 (black) to 0.575 (white). Each contour map of a mutant is aligned with its respective amino acid sequence deduced from DNA sequencing. All the mutants sequenced were unique at the amino acid and DNA levels. The asterisks designate mutants which include a spontaneous A to C transversion observed in the first nucleotide of the codon immediately upstream of group 1.3.

An important parameter for the construction and analysis of combinatorial libraries is the frequency of functional mutants, expected or observed. We offer here a mathematical context for the study and discussion of combinatorial mutagenesis data. Stringency ( $\mathbf{x}$ ) is defined as the ratio of codons leading to functional proteins to the total number of codons. As a corollary of eq.1, if  $n$  amino acid residues are mutagenized and the throughput of positive mutants is a fraction ( $\mathbf{f}$ ) of the total library, then the overall stringency ( $\mathbf{x}$ ) for  $n$  residues can be calculated as the geometric mean of the individual stringencies ( $\mathbf{x}_i$ ) of these mutagenized residues:

$$\mathbf{x} = \left( \prod_i^n \mathbf{x}_i \right)^{\frac{1}{n}} \quad \text{Eq. 2}$$

This reasoning can be extended to groups of residues where each  $\mathbf{x}_i$  is the stringency of a group  $i$ , and  $\mathbf{x}$  is an overall stringency. For a hypothetical 16 site *random* [NN(G/C)] library, we must use the frequencies obtained experimentally from groups 1.1, 1.2 and 1.3 (respectively, 6 + 5 + 5 = 16 mutagenized residues). By grouping the  $\mathbf{x}_i$  variables in these three groups, we have:

$$\mathbf{x}_{\text{random}} = (10^{-4} \times 10^{-2} \times 10^{-3})^{1/16} = 0.274 \quad \text{Eq. 3}$$

Since 32 codons are used in a random [NN(G/C)] nucleotide mixture, then  $0.274 \times 32 = 8.8$  of the codons are functional (geometric mean per amino acid site). This corresponds to approximately  $20 \times 8.8/32 = 5.5$  functional amino acids per site. This can be contrasted with the 16 site *EEM* library (group 2.1) that has an experimentally measured throughput of  $10^{-2}$ , therefore:

$$\mathbf{x}_{\text{EEM}} = (10^{-2})^{1/16} = 0.750 \quad \text{Eq. 4}$$

This suggests that three quarters of the optimized nucleotide dopes used in the group 2.1 EEM library are functional. This is almost a three-fold improvement over the random library and is the basis of the experimental gain ( $\mathbf{Q}$ ) of EEM

over random mutagenesis. According to the exponential hypothesis, the two  $x$  values are ratioed:

$$Q = (x_{EEM} / x_{random})^n \quad \text{Eq. 5}$$

For the 16-site group 2.1 library, the gain is:

$$Q = (0.750 / 0.274)^{16} = 10^7 \quad \text{Eq. 6}$$

In Figure 2.4, values are plotted and extrapolated according to the exponential hypothesis as a function of  $n$ . Certainly, the  $x$  values will change depending on the specific sites and specific proteins chosen for mutagenesis. However, for proteins with stringencies similar to LHII, these plots suggest that mutagenesis of 30 residues will result in a throughput of positive mutants of approximately  $2 \times 10^{-4}$  for EEM but only  $1 \times 10^{-17}$  for random [NN(G/C)] mutagenesis. In the latter case, no positive mutants would be observed in actual experiments since the reciprocal of this number exceeds cloning efficiencies by at least 5 orders of magnitude.

The group 2.1 EEM optimized nucleotide dopes (Fig. 2.2) reflect the frequency of occurrence of functional amino acids in the three random libraries, highlighting the functional importance of particular amino acids. The histidine ligand of the dimer bchl is completely conserved in this mutagenesis scheme because group 1.3 mutants show no diversity at this position. Similarly, SSD forced the optimized doping scheme to encode only Pro and Trp in group 1.1 because these two residues are highly conserved (Delagrave *et al.*, 1993). Overall, the amino acids encoded by the optimized nucleotide mixtures are hydrophobic residues or small amphiphilic residues, consistent with their position inside the membrane or near the membrane surface (Zuber, 1990).

### 2.3 Discussion

Decrease in absorbance of both the 800 nm and 858 nm bands in a given mutant LHII spectrum is consistent with a change in the level of expression of the protein. Shifts in maximal wavelength of absorption of the 858 nm band probably reflect a modification of the environment surrounding the

dimer bchl. This type of phenotypic diversity suggests that mutants with interesting properties may be present in these libraries and that the development of new screening or selection techniques could lead to their isolation. The possibility of obtaining mutants with more radical departures from the original phenotype may be limited by the structural constraints which the remainder of the protein (the  $\alpha$  and  $\gamma$  subunits) imposes on the region mutagenized. Nevertheless, random mutagenesis of group 1.4 (Fig. 2.1) has recently yielded LHI-like phenotypes (single absorption peak ca. 875 nm; Delagrave, unpublished results) similar to those observed previously by Goldman and Youvan (1992).

The usefulness of this method depends on how efficiently desired phenotypes may be recovered. In this study, a wild-type phenotype was very successfully amplified by the optimization employed, with some observed phenotypic variability (e.g., absorption peak shifts). Clearly, the subset of functional variants is better sampled using EEM than by random mutagenesis, for the same number of sites altered. A potential limitation to the approach would be the narrowing of the sequence space search down a genetic "dead end" through the elimination of some amino acid residues necessary for a desired phenotype. The large number of functional altered proteins apparently present in sequence space makes unlikely the chance elimination of a desirable variant. Therefore, one should look for systematic biases inherent to the search strategy. For instance, it is possible that two amino acids (at different sites in the sequence) which are deleterious when expressed alone, are fully functional when expressed together. The grouping and parallel optimization of amino acid residues can prevent such a cooperative interaction from taking place and therefore may eliminate some functional double mutants from an optimized library. This brings into question the grouping strategy taken here. Although the scheme depicted in Figure 2.1 shows mutagenesis carried out on contiguous segments of protein sequence, the experiments need not be done in such a fashion. As structural and other data warrant, it may become desirable to alter various sites interspersed throughout the sequence. This might be called a semi-rational approach to protein engineering, as opposed to the nearest neighbor strategy taken here, which avoids assumptions about the actual structure of the protein. Analogies can be drawn with artificial intelligence approaches wherein "strong" methods incorporate expert data. In

this case, a "weak" method would group residues randomly. Grouping contiguous residues could be considered as intermediate between these two extremes since adjacent residues are likely to participate in the same structural motif (e.g., a hydrophobic alpha helix).

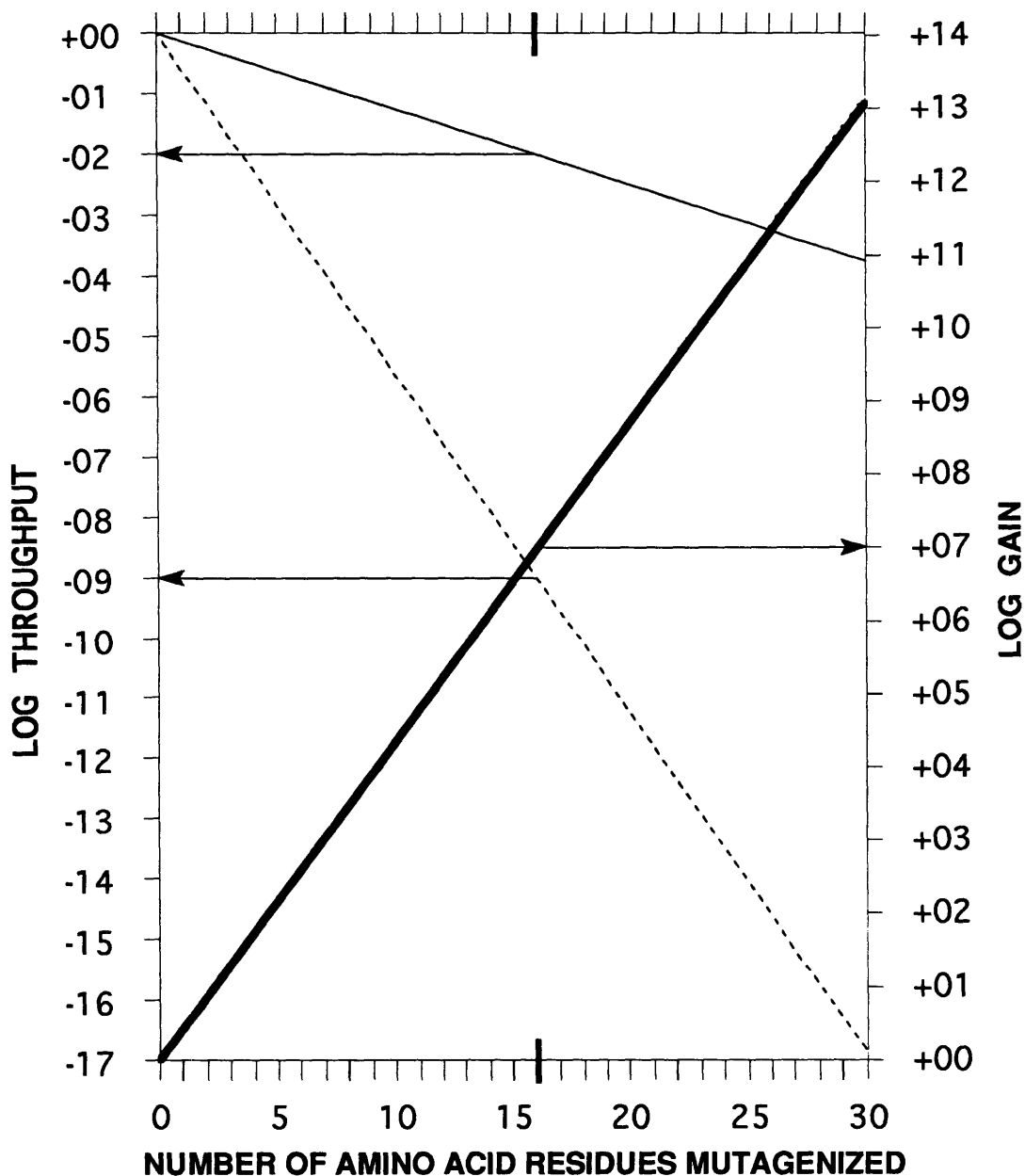


Figure 2.4. Extrapolation of experimental data according to the exponential hypothesis shows the efficiency of EEM as the number of



mutagenized amino acid residues is increased. These plots represent the zeroeth order approximations given by equations 1 and 5. Experimental data are indicated by the large tick marks at 16 residues. Arrows to the left side y-axis indicate a  $10^{-9}$  throughput for random mutagenesis (- - -) and a  $10^{-2}$  throughput of positive mutants for EEM (—) on 16 sites. The ratio of these two numbers yields a  $10^7$  fold gain (——) of EEM over random mutagenesis, as indicated by an arrow to the right side y-axis. The slopes of these lines are based on the measured stringency for the 16 site mutagenesis of the LHII proteins: 75% of the codons encoded by the group 2.1 EEM library are functional, whereas only 27% of the randomly generated codons lead to positive mutants. Further extrapolation of the EEM data suggests that >500 positive mutants could be obtained after mutagenizing 50 residues of similar stringency using a library of  $10^9$  mutants. This library size is feasible using phage display. These data and extrapolations are based on one "level" of EEM (see text); however, multiple levels of EEM might be used to mutagenize more residues.

As more experimental data are accumulated for a variety of proteins, deviations from the exponential hypothesis might cause curvature of the gain (in Figure 2.4, plotted semilogarithmically as a function of the number of residues mutagenized). A more detailed theory should account for possible compensatory interactions between amino acids (cooperativity). Interactions between distant amino acid residues in the primary structure imposes restrictions on which combinations of mutations will be functional. In some cases, these combinations can be explained by direct contact between amino acids, as in the hydrophobic core of  $\lambda$  repressor, where the independence of each altered residue is proposed as a model of stringency (Lim and Sauer, 1991). The nature of "site-to-site" interaction, however, is not necessarily limited to direct physical contact as evidenced, among other examples, by the observation of compensatory mutations in photosynthetic reaction center mutagenesis at unexpected locations (Robles *et al.*, 1992). The simplification that cooperative interactions can be neglected may be reasonable if "site-to-site" interactions are both of a negative or positive nature and if these

interactions roughly cancel out. Moreover, this simplification is supported by mutagenesis and structural data reviewed by Wells (Wells, 1990). The multiplicative property (i.e., additivity in exponents) of functional mutant frequencies have been observed in other combinatorial mutagenesis experiments, despite the demonstration of cooperative interactions between the regions mutagenized (Robles and Youvan, 1993).

Certain generalizations from this and other work can be made which could help future efforts in protein engineering. The availability of naturally versatile molecules such as antibodies makes them stand out as possible progenitors of a population of mutants on which a selection can be made to extract many interesting phenotypes. A starting point to this concept requires the definition of amino acid sequence requirements for functional antibodies (Chothia *et al.*, 1992). EEM is well suited in this regard because of the variety of functional mutants it can generate. Presumably because they are generally exposed to the solvent and independent from their framework, altering numerous residues in CDRs may be more likely to lead to functional molecules. Lower stringency implies that a diverse ensemble of mutagenized antibodies could be constructed by employing EEM at a greater number of sites. *In vitro*, EEM coupled with the phage display methodology may out perform the immune system in the formation of high affinity antibodies. Combinatorial libraries from which valuable antibodies (Fab's) can be selected (Burton *et al.*, 1992) may benefit from the capacity that EEM has to simultaneously adjust many residues.

## 2.4 Experimental protocol

**Cassette construction.** DNA manipulations were carried out as described by Sambrook *et al.* (1990). All restriction enzymes were from New England Biolabs. T4 DNA ligase and Taq DNA polymerase were from Bethesda Research Labs. Oligonucleotides were synthesized on an Applied Biosystems model 381 DNA synthesizer. Sequencing of M13 subclones was done using a kit from United States Biochemicals; in some cases, pU4b plasmids were sequenced directly using a cycle sequencing kit from New England Biolabs. As described (see Appendages, as well as First Iteration and Delagrave *et al.*, 1993 and Goldman and Youvan, 1992), cassettes were constructed by synthesizing 113mers with some degenerate nucleotide

positions. These were subsequently amplified by PCR and digested with appropriate restriction enzymes (see Fig.2.2). A similarly digested expression vector (pU4b) was then ligated to the cut cassettes for 24h at 16°C. Electroporation of the ligation products into S17-1 *E.coli* allowed the production of libraries with complexities of up to 10<sup>5</sup> transformants. Expression of these libraries required the conjugation of S17-1 transformants with strain U71 of *Rb. capsulatus*. By this procedure, the libraries described were 80% to 100% free of Wild-Type contaminants.

**Imaging spectroscopy.** The simultaneous imaging of thousands of colonies (i.e., 25 petri dishes, simultaneously) is achieved using a 4 megapixel Photometrics CCD camera. This camera is interfaced to a Silicon Graphics Crimson Elan computer which processes the images using the program "CyberPeeper" (obtained from KAIROS Inc.). Fluorescence images as well as absorption images of a fixed wavelength can be taken, manipulated and stored for analysis (Fig.2.3, panels A and B). Imaging of spots (Fig.2.3, panel C) was carried out by first growing duplicate liquid cultures of the purified group 2.1 EEM mutants for 24 hours at 32°C in RM-tetracycline medium. Each liquid culture was spotted (3 µL) on two separate RM-tetracycline plates and incubated for 24 hours at 32°C.

**Selection criteria.** In the search of LHII combinatorial libraries, primary screening based on fluorescence was used to rapidly identify mutants expressing significant levels of LHII. This method was also used to count positive frequencies and is the basis of our comparison of different libraries. For the purposes of establishing a sequence database, we have defined a "positive" as an LHII mutant showing the characteristic 800 nm and 858 nm absorption bands, where the absorbance at 858 nm is 10% or more of wild-type as determined by DIS.

## 2.5 Notes and acknowledgements

Many thanks to Jennifer Reyna for technical assistance and the occasional batch of brownies. CyberDope and CyberPeeper were provided by KAIROS Inc. This iteration was published in Bio/Technology (Delagrave,S. and Youvan,D.C., 1993, 11, 1548-1552).

## 2.6 References

Arkin, A.P., Goldman, E.R., Robles, S.J., Coleman, W., Goddard, C.A., Yang, M.M. and Youvan, D.C. (1990). Applications of imaging spectroscopy in molecular biology: Colony screening based on absorption spectra. *Bio/Technology* **8**:746-749.

Arkin, A.P. and Youvan, D.C. (1992). A combinatorial optimization procedure for protein engineering: Simulation of recursive ensemble mutagenesis. *Proc. Natl. Acad. Sci. U.S.A.* **89**:7811-7815.

Arkin, A.P. and Youvan, D.C. (1992). Optimizing nucleotide mixtures to encode specific subsets of amino acids for semi-random mutagenesis. *Bio/Technology* **10**:297-300.

Barbas, C.F., Bain, J.D., Hoekstra, D.M. and Lerner, R.A. (1992). Semisynthetic combinatorial antibody libraries: A chemical solution to the diversity problem. *Proc. Natl. Acad. Sci. U.S.A.* **89**:4457-4461.

Chothia, C., Lesk, A.M., Gherardi, E., Tomlinson, I.M., Walter, G., Marks, J.D., Llewelyn, M.B. and Winter, G. (1992). Structural repertoire of the human Vh segments. *J. Mol. Biol.* **227**:799-817.

Delagrave, S., Goldman, E.R. and Youvan, D.C. (1993). Recursive ensemble mutagenesis. *Prot. Engng.* **6**:327-331.

Goldman, E.R. and Youvan, D.C. (1992). An algorithmically optimized combinatorial library screened by digital imaging spectroscopy. *Bio/Technology* **10**:1557-1561.

Gram, H., Marconi, L.-A., Barbas, C.F., Collet, T.A., Lerner, R.A., and Kang, A.S. (1992). In vitro selection and affinity maturation of antibodies from a naive combinatorial immunoglobulin library. *Proc. Natl. Acad. Sci. U.S.A.* **89**:3576-3580.

Hoogenboom, H.R., Griffiths, A.D., Johnson, K.S., Chiswell, D.J., Hudson, P. and Winter, G. 1991. Multi-subunit proteins on the surface of filamentous phage: Methodologies for displaying antibody (Fab) heavy and light chains. *Nucleic Acids Res.* **19**: 4133-4137.

Houghten, R.A., Pinilla, C., Blondelle, S.E., Appel, J.R., Dooley, C.T. and Cuervo, J.H. (1991). Generation and use of synthetic peptide combinatorial libraries for basic research and drug discovery. *Nature* **354**: 84-86 .

Kang, A.S., Barbas, C.F., Janda, K.D., Benkovic, S.J., and Lerner, R.A. (1991). Linkage of recognition and replication functions by assembling combinatorial antibody Fab libraries on phage surfaces. *Proc. Natl. Acad. Sci. U.S.A.* **88**:4363-4366.

Lam, K.S., Salmon, S.E., Hersh, E.M., Hruby, V.J., Kazmiersky, W.M., Knapp, R.J. (1991). A new type of synthetic peptide library for identifying ligand-binding activity. *Nature* **354**: 82-84.

Lim, W.A. and Sauer, R.T. (1991). The role of internal packing interactions in determining the structure and stability of a protein. *J. Mol. Biol.* **219**:359-376.

Lowman, H.B., Bass, S.H., Simpson, N. and Wells, J.A. (1991). Selecting high-affinity binding proteins by monovalent phage display. *Biochemistry* **30**:10832-10838.

Oliphant, A. R., Nussbaum, A. L. and Struhl, K. (1986). Cloning of random sequence oligonucleotides. *Gene* **44**:177-183.

Reidhaar-Olson J. F., Bowie J. U., Breyer, R. M., Hu, J. C., Knight K. L., Lim W. A., Mossing M. C., Parsell D. A., Shoemaker K. R. and Sauer R.T. (1991). Random mutagenesis of protein sequences using oligonucleotide cassettes. *Methods in Enzymology* **208**:564-587.

Roberts, B.L., Markland, W., Ley, A.C., Kent, R.B., White, D.W., Guterman, S.K. and Ladner, R.C. (1992). Directed evolution of a protein: Selection of potent

neutrophil elastase inhibitors displayed on M13 fusion phage. *Proc. Natl. Acad. Sci. U.S.A.* **89**:2429-2433.

Robles, S.J., Ranck, T. and Youvan, D.C. (1992). Symmetrical intragenic suppressors of the bacterial reaction center cd-helix exchange mutants. *In: Structure of the Bacterial Photosynthetic Reaction Center(II)*. J. Breton and A. Vermeglio (Eds.). Plenum Press, New York.

Robles, S.J. and Youvan, D.C. (1993). Hydropathy and molar volume constraints on combinatorial mutants of the photosynthetic reaction center. *J. Mol. Biol.* In press.

Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989). *Molecular cloning: a laboratory manual* (2nd Edition). Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Smith, G.P. 1985. Filamentous fusion phage: Novel expression vectors that display cloned antigens on the virion surface. *Science* **228**:1315-1317.

Wells, J.A. (1990). Additivity of mutational effects in proteins. *Biochemistry* **29**:8509-8516.

Yang, M.M. and Youvan, D.C. (1988). Applications of imaging spectroscopy in molecular biology. I. Screening photosynthetic bacteria. *Bio/Technology* **6**:939-942.

Youvan, D.C., Arkin, A.P. and Yang, M.M. (1992). Recursive ensemble mutagenesis: A combinatorial optimization technique for protein engineering. *In: Parallel Problem Solving from Nature*, 2, pp. 401-410. R. Maenner and B. Manderick (Eds.) Elsevier Publishing Co., Amsterdam.

Youvan, D.C., Goldman, E.R., Delagrave, S. and Yang, M.M. (1993). Digital imaging spectroscopy for massively parallel screening of mutants. *Methods in Enzymology.* In press.

Youvan, D.C. and Ismail, S. (1985). Light-harvesting II (B800-B850 complex) structural genes from *Rhodospseudomonas capsulata*. *Proc. Natl. Acad. Sci. U.S.A.* **82**:63-67.

Zuber, H. (1990). Consideration on the structural principles of the antenna complexes of phototrophic bacteria, pp.161-180. *In: Molecular Biology of Membrane-Bound Complexes in Phototrophic Bacteria.* G. Drews and E.A. Dawes (Eds.) Plenum Press, New York.

## **Third Iteration. Context Dependence of Phenotype Prediction and Diversity in Combinatorial Mutagenesis**

### **3.0 Summary**

Two different combinatorial mutagenesis experiments on the Light Harvesting II (LH2) protein of *Rhodobacter capsulatus* indicate that heuristic rules relating sequence directly to phenotype are dependent on which sets, or groups, of residues are mutagenized simultaneously. Previously reported combinatorial mutagenesis of this chromogenic protein (based on both phylogenetic and structural models) showed that substituting amino acids with large molar volumes at Gly<sup>β31</sup> caused the mutagenized protein to have a spectrum characteristic of Light Harvesting I (LH1). The six residues that underwent combinatorial mutagenesis were modeled to lie on one side of a transmembrane  $\alpha$ -helix that binds bacteriochlorophyll. In a second CCM experiment described here, we have not used structural models or phylogeny in choosing mutagenesis sites. Instead, a set of six contiguous residues was selected for combinatorial mutagenesis. In this latter experiment, the residue substituted at Gly<sup>β31</sup> was not a determining factor in whether LH2 or LH1 spectra were obtained; therefore, we conclude that the heuristic rules for phenotype prediction are context dependent. While phenotype prediction is context dependent, the ability to identify elements of primary structure causing phenotype diversity appears not to be. This strengthens the argument for performing combinatorial mutagenesis with an arbitrary grouping of residues if structural models are unavailable.



### 3.1 Introduction

Combinatorial mutagenesis is the simultaneous introduction of mutations at a number of different positions in a sequence, such that various combinations of amino acids can be found in different mutants. Molecular genetic techniques are commonly used to construct such libraries of altered sequences (1,2). Typically, these experiments give information on requirements at the structural level to obtain stable, functional proteins (3). Simultaneously randomizing (i.e., including all combinations of the 20 amino acids and a stop codon) as few as six amino acids in a protein leads to  $20^6$ , or  $6.4 \times 10^7$  different possible amino acid sequences. Whether the goal is to engineer a protein or study structure-function relationships, one is usually interested in isolating *functional* mutants and determining their sequences. Unfortunately, these make up a smaller proportion of the library of combinatorial mutants as more sites are randomized.

To overcome this combinatorial complexity problem, one option is to divide the sequence into subsets (i.e., groups) of amino acids which will be altered independently from each other (4,5). The division of the sequence into groups of amino acids can be limited by experimental constraints but is essentially arbitrary. For instance, one can choose contiguous amino acids to form a small group of residues that will be mutagenized simultaneously. Alternatively, residues thought to belong to some structural element (e.g., the same face of an  $\alpha$ -helix, or a complementarity determining region of an antibody) can be mutated together (6,7). The approach of dividing the sequence into groups, however, leads to the interesting possibility that underlying interactions between amino acids in the protein will give different results if residues are grouped differently. Using combinatorial mutagenesis data from a model protein, we demonstrate that this is indeed the case.

Our model system is the Light Harvesting II (LH2) protein of *Rb. capsulatus*. The LH2 protein is an integral membrane protein found in certain phototrophic purple bacteria at the periphery of a core complex composed of the LH1 protein and the photosynthetic reaction center (RC) (8). LH1 is a protein closely related to LH2, with homologous  $\alpha$  and  $\beta$  subunits (9). The degree of sequence identity suggests that LH1 and LH2 are the result of a gene duplication. Because these proteins bind bchls, their level of expression can be estimated by ground state

absorption spectroscopy (10). The LH1 absorption spectrum shows a single peak in the NIR at 870 nm while LH2 has two bands, one at 800 nm and another at 855 nm. The  $\beta$  subunit of LH2 has two histidines, His $^{\beta 20}$  and His $^{\beta 38}$ . The first is thought to be the axial ligand of the bchl responsible for the 800 nm band, called the monomer band. His $^{\alpha 31}$  and His $^{\beta 38}$  are modeled to bind one bchl each; together these bchls are excitonically coupled and cause the 855 nm band, called the dimer band. Changes in the protein environment of these chromophores can lead to changes in peak intensity or shifts in the maximal wavelength of absorption (6,11). Non-specific mutagenesis has been used in the past to introduce a change in the phenotype of LH2 to a pseudo-LH1 phenotype in *Rb. capsulatus* (12). At that time, the exact genetic basis of this transformation was not clear.

Two different types of combinatorial libraries are discussed in this communication. Both of these are made in the same way, by combinatorial cassette mutagenesis (CCM). One of the two types of libraries is a Target Set Mutagenesis (TSM) library; the other is a fully random combinatorial library (both are described below). TSM was first used (6) as a means of investigating structure-function relationships in LH2 and combinatorial library optimization methods. The extensive phylogeny of Light Harvesting proteins (13) was used to identify which amino acids, at a given position in the sequence, were most likely to preserve the protein's structural integrity. At each position, a few different amino acids identified in this way were encoded, constituting a subset (or Target Set) of all possible amino acids. This strategy increases the proportion of functional mutants contained in the library. The mutations were introduced simultaneously at six residues in the LH2  $\beta$  subunit gene (Fig. 3.1, top line of figure). These residues were chosen because they are modeled to be on the same face of the putative  $\alpha$ -helix as His $^{\beta 38}$  which binds one of the dimer bchls. Expression of this library produced different phenotypes, including numerous LH1-like mutants displaying a single absorption peak between 860 nm and 870 nm.

**Groups**

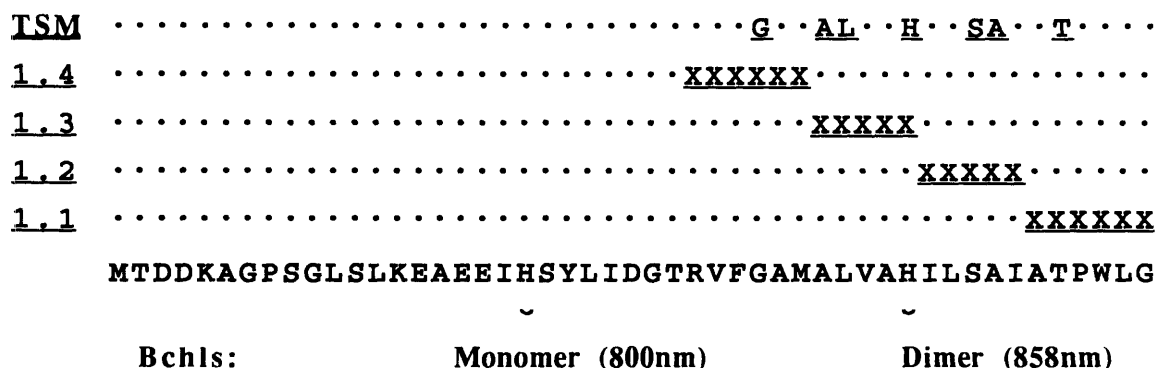


Figure 3.1. The Light Harvesting II  $\beta$  subunit amino acid sequence (bottom line of figure, single letter code) has been genetically altered in five different mutagenesis experiments. The His ligands of the two bchIs are indicated, along with the assigned NIR bands of these chromophores. Each combinatorial library contained different groups of mutated residues. Group 1.1, for instance, spans the six amino acids at the carboxy terminal end of the subunit. In this library, as well as groups 1.2 to 1.4, all mutated sites were randomized (i.e., 20 amino acids and one stop codon allowed at each altered position). The Target Set Mutagenesis (TSM) library simultaneously mutated six residues (the His ligand was not altered) but allowed only subsets of the 20 amino acids to be encoded at each one, according to the phylogeny of LH1 and LH2  $\beta$  subunits. Note how Gly $\beta^{31}$  (the first altered residue in the TSM library) is the only site overlapping with group 1.4.

In parallel with the work described above, another type of combinatorial mutagenesis was carried out where five or six contiguous amino acids were simultaneously randomized (4,14). Surprisingly, these random mutagenesis experiments spanning 16 codons at the carboxy terminal end of LH2  $\beta$  (groups 1.1 to 1.3 in Fig.3.1), in the same region of the  $\beta$  gene, were never found to yield LH1 phenotypes. Optimized libraries of this entire region, where many different functional mutants could be spectroscopically assayed, did not reproduce this phenotype (4,14). Analysis of the sequences of the pseudo-LH1 mutants from

the TSM library revealed that the molar volume of the substitution at residue Gly<sup>β31</sup> could be used to predict whether an LH1-like or LH2-like phenotype would be observed (Table 3.1). Examination of figure 3.1 will show that this particular residue is the only one which does not overlap with the region altered by the three random mutagenesis experiments, groups 1.1, 1.2 and 1.3. This observation prompted us to make the group 1.4 library, analogous to the other three but involving a different, adjacent group of residues which included Gly<sup>β31</sup>. All functional (i.e., pigment binding) mutants isolated from the group 1.4 random library displayed LH1-like spectra. Some mutants from this library were characterized in order to compare their sequences with the sequences of previously isolated pseudo-LH1 mutants. The implications of these results on sequence-phenotype relationships and underlying structural events are discussed.

### 3.2 Materials and Methods

*Construction of group 1.4 combinatorial library.* This library of mutants was constructed by combinatorial cassette mutagenesis (1-4,6,14). A 112-mer oligonucleotide was synthesized on an Applied Biosystems model 381 DNA synthesizer. This oligo, which encodes part of the sense strand, had the following sequence (Kpn I and Xho I restriction sites underlined) : 5' GC TAC CTG ATC GAT GGT ACC (NNS)<sub>6</sub> GCG CTT GTT GCG CAC ATC CTC TCG GCC ATC GCC ACG CCG TGG CTC GGG TAA TCG GCT CGA GGA GAA ATA CAA TG 3'. Where N is an equimolar mixture of A, C, G and T and S is an equimolar mixture of G and C. G and C were chosen at the wobble position because they do not lead to forbidden codons with respect to *Rb. capsulatus* codon usage. Also, using G and C rather than N at this position leads to a more equiprobable representation (i.e., closer to 5% each) of the amino acids in the library. Two primers were used to make a double stranded "mutagenic cassette" by the PCR reaction: the Kpn I oligo, 5' GC TAC CTG ATC GAT GGT ACC 3', and the Xho I oligo, 5' CA TTG TAT TTC TCC TCG AGC 3'. In both cases, the restriction sites are underlined. Plasmid K is a construct (described previously, ref. 6) in which the  $\beta$  subunit gene of LH2 is interrupted by a non-coding Kpn I-Xho I fragment. Substitution of this "dead cassette" with a mutagenic cassette was carried out by separately digesting the double stranded cassette produced by PCR and plasmid K with Xho I and Kpn I. The digestion products were purified by phenol:chloroform extractions and ethanol precipitation, mixed together in a roughly 8:1 molar ratio of cassette to plasmid, with 0.2 pmol of cut plasmid used. The ligation was carried out overnight at 16°C in a volume of 25  $\mu$ L with 2 units of T4 DNA ligase from Gibco-BRL. Half of the ligation was transformed into *E. coli* S17-1 by electroporation. Plasmid K is a derivative of the broad host range plasmid pRK404 (14), it is stable and confers tetracycline resistance in both *E. coli* and *Rb. capsulatus*. The S17-1 electrocompetent cells were prepared essentially as described (15). Immediately after electroporation of 0.25 mL of competent cells in a 0.2 cm cuvette, the cells were resuspended in LB and incubated with shaking at 37°C for one hour. Aliquots were then spread on LB<sup>tet</sup> plates and incubated overnight. The number of colonies on these plates are a measure of the complexity, i.e., the number of unique transformants.

To express the library of altered LH2 genes, the library was conjugated from S17-1 to *Rb. capsulatus*. The strain of *Rb. capsulatus* used was U71 (RC<sup>-</sup>/LH1<sup>-</sup>/LH2<sup>-</sup>), a deletion background which has no absorbing species in the NIR (17). Early log phase cultures of the donor (S17-1, 0.3 mL) and the recipient (U71, 0.7 mL) are mixed in a sterile microfuge tube and aliquots are spotted eight to ten times on a dry MPYE plate. The conjugation occurs at 32°C overnight. The spots are then picked off the plate with a sterile loop and resuspended in 1 mL of RCV in a sterile microfuge tube. Centrifugation for 30 seconds will produce a biphasic pellet with *Rb. capsulatus* in the upper portion of the bacterial pellet. The *Rb. capsulatus* pellet is removed with a pipette and resuspended in RCV. This library of transconjugates can be spread on RCV<sup>tet</sup> plates on which contaminating *E. coli* will not grow. Mutants, identified as described below, can be picked and restreaked on RM<sup>tet</sup> to isolate them from contaminants.

*Digital imaging spectroscopy.* Characterization of the libraries of mutant LH2 genes was carried out by Digital Imaging Spectroscopy, or DIS (10,18,19). Briefly, a charge coupled device camera acquires images of petri plates mounted on the exit port of an integrating sphere. The light source uses a 1/8 meter monochromator to illuminate the plate at different wavelengths at 5 nm intervals. This resolution allows the detection of 2 nm band shifts. The libraries were expressed in *Rb. capsulatus* by spreading aliquots on RCV<sup>tet</sup> plates and growing at 32°C for at least 48 hours. Alternatively, after mutants found were repurified by streaking, liquid cultures were grown to early log phase in RM<sup>tet</sup> for 24 hours at 32°C and 3 µL of the cultures were spotted on RM<sup>tet</sup> plates and grown under the same conditions. This procedure gives reproducible measurements of maximal absorption wavelengths using DIS (14). The phenotypes of a few of the mutants appeared to be growth dependent, thus there is the potential for a 10% error in the classification of spectra as LH1-like or LH2-like. Spectra of the group 1.4 mutants in the NIR, from 730 nm to 930 nm, were determined using the first implementation of DIS (19), while spectra of the TSM library mutants were obtained using a ColonyImager (KAIROS Inc., Mountain View, CA).

*Sequencing of mutants.* After DIS identification of *Rb. capsulatus* colonies showing interesting spectra, mutants were picked and restreaked twice on

RM<sup>tet</sup> plates. Liquid cultures of the mutants are used to prepare plasmid DNA by standard methods (20). The plasmid DNA isolated from *Rb. capsulatus* cannot be digested readily and so must be transformed into *E. coli* MV1190. From this transformant, plasmid DNA is again isolated which can be manipulated by conventional techniques. For each mutant to be sequenced, a 1.5kb Pst I-Hind III fragment was then subcloned from the mutant plasmid K into M13mp19. Single stranded DNA was sequenced in the altered region of the LH2  $\beta$  subunit gene, using the Sequenase 2.0 kit from US Biochemicals.

### 3.3 Results

New functional mutant spectra and sequences were obtained from the TSM library previously described (6). Figure 3.2 shows the results of DIS characterization of 62 mutants and triplicate WT controls. Because of the large number of spectra generated by DIS, the spectral information must be displayed in a more efficient way than the standard OD vs. wavelength plots. In figure 3.2, each spectrum, obtained from a single colony, is displayed as a horizontal row: the position along the length of the row corresponds to the wavelength while the color at any point encodes the absorption at that wavelength. Algorithms have been designed to sort spectra according to similarity (10). These sorting algorithms were applied to a spectral data set which, at first, was arranged randomly. As a result of the sorting of the spectral data, two groups are clearly visible in the right hand panel of figure 3.2 : LH2-like spectra occupy the first 39 rows while LH1-like spectra take the bottom 26 rows. Imaging of aliquots of the TSM library, as previously reported, revealed that 6% of the mutants bind bchl and that these functional mutants belong to either one of the two spectroscopic classes described. The pseudo-LH1 phenotype is defined as showing a red shift of the 855 nm band and a substantial decrease or disappearance of the 800 nm band. Sequencing of each of the mutants imaged in figure 3.2 is summarized in Table 3.1, where the amino acids found at each mutated position are given. Each mutant is numbered according to the position of its spectrum in figure 3.2. Inspection of the Table shows that amino acids with molar volumes larger than Thr at Gly<sup>β31</sup> (labeled in Table 3.1 as the -7 position, relative to His<sup>β38</sup>) are expected to cause a pseudo-LH1 phenotype (left column in Table 3.1). Smaller amino acids lead to an LH2 phenotype, as shown in the right-hand column of Table 3.1. There are 6 sequences which do not obey this rule: mutants number 40, 41, 50, 63, 7 and 29. This corresponds to a 90% rate of accurate prediction using the described rule. A Chi-square test applied to this data (21,22) confirms that the observed discrepancy is not statistically significant at the 95% confidence level ( $\chi^2 = 0.2649$ ,  $p > 0.5$ ).



Table 3.1 Sequences of TSM library mutants

Pseudo-LH1 mutants    Pseudo-LH2 mutants

	-7 -4 -3 0 3 4 7		-7 -4 -3 0 3 4 7
40	C G I H A G T	1	S G T H A M T
41	T A A H S A T	2	G A A H S I T
42	I V T H S A T	3	G A L H A G T
43	M G V H A M T	4	A V I H A M T
44	V A A H A Y T	5	T A T H V A T
45	M A L H S C T	6	A A I H V A S
46	L A V H A M T	7	F A V H V A T
47	M A L H A A T	8	G V I H A G T
48	L V L H S A T	9	G A T H S F T
49	I A T H A T T	10	G G L H A V T
50	C G L H A A T	11	G G A H A W T
51	V A L H A W T	12	A A A H A W T
52	L G L H A F T	13	G A V H A K T
53	V A V H S W T	14	A A A H A W T
54	V A L H S G T	15	G G A H A V T
55	L A A H A G T	16	G G I H A V T
56	L A V H S A T	17	A A I H A A T
57	L V I H A G T	18	G A V H A F S
58	V G A H S A T	19	G A A H S F S
59	L A V H S A T	20	G V L H S G T
60	V G I H S A T	21	G A A H A K T
61	L A A H A G T	22	G G L H A V T
62	L G L H A A T	23	A V L H S A T
63	A A L H S T N	24	G G A H V A T
64	I V T H A A T	25	A G A H S A T
65	V G V H A G T	26	G A A H S A T
		WT 27	G A L H S A T
		WT 28	G A L H S A T
		29	W G V H A F T
		30	A A V H S G T
		31	T A V H A Y T
		32	G A L H S I T
		33	S G I H A G T
		34	T A I H S M T
		35	G G I H S Y T
		36	A G V H S A T
		37	T A T H A V T
		38	G G I H S Y T
		WT 39	G A L H S A T

Amino acid sequences of the LH2 mutants of the TSM library were deduced from nucleotide sequences. Only the sequence positions which were altered are included. These positions are numbered relative to the histidine ligand (0). All mutants had unique nucleotide sequences. The number of each mutant indicates its row number in Figure 1.

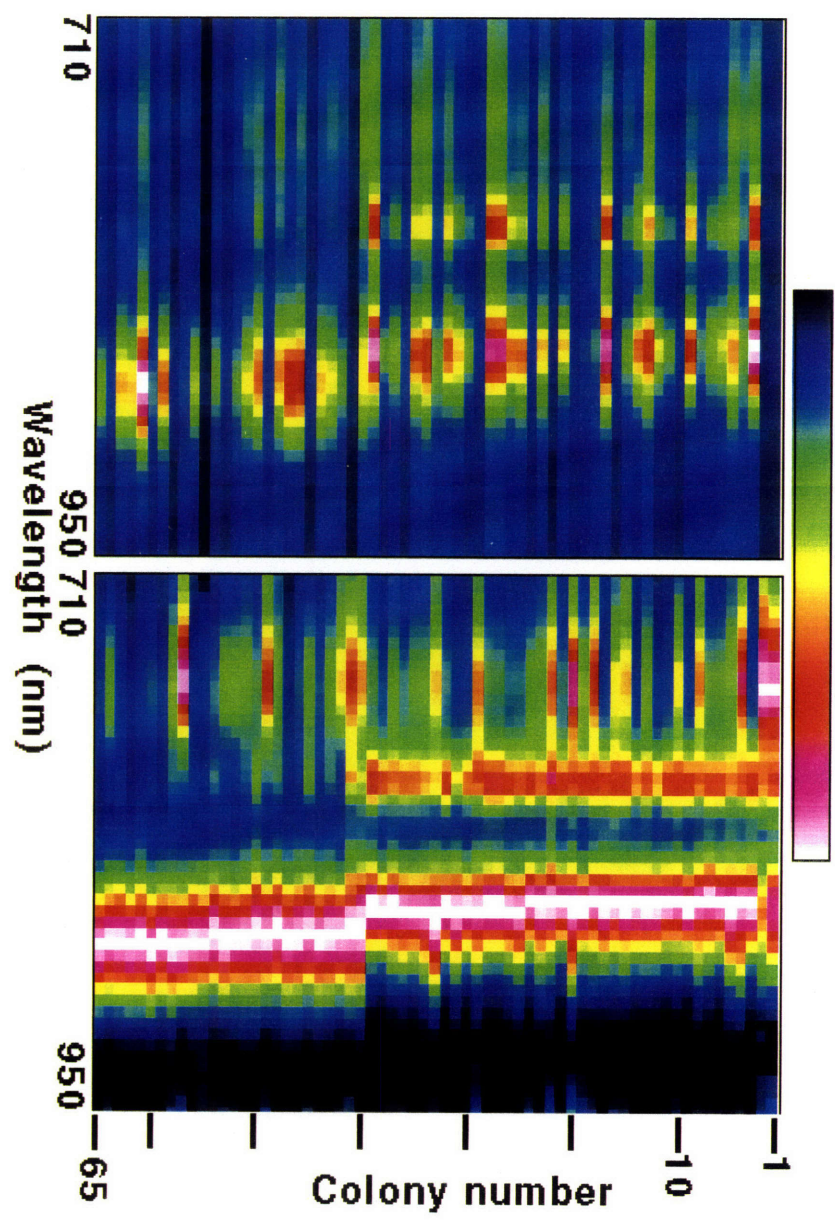


Figure 3.2. (Previous page) Color contour plots generated by the DIS ColonyImager showing the spectrum of each sequenced mutant. The horizontal axis corresponds to wavelength (710-950 nm) and the vertical axis to colony row number. Each row represents the spectrum of a mutant, encoded by pseudocolor. The color bar shows the range of optical density (OD) from low (black) to high (white). The left panel is in 'absolute mode' (highest OD in the entire spectral data set mapped to white, lowest OD mapped to black); this shows the range of expression levels. The right panel is in 'full deflection mode' (highest OD for each spectrum mapped to white and the lowest OD mapped to black); this representation enables a more direct comparison of the spectral bands. The row number can be used to find the corresponding amino acid sequence in Table 3.1. Raw spectral data were sorted according to maximum absorption in the absolute mode, and then by wavelength of maximum absorbance in the full deflection mode. Rows 27, 28 and 39 are all WT LH2 colonies.

CCM was used to construct another library of altered LH2  $\beta$  genes expressed in *Rb. capsulatus*. Six contiguous residues (residues 28-33 towards the carboxy terminus end of the  $\beta$  polypeptide) were simultaneously randomized, allowing any of the 20 amino acids and one stop codon to be expressed at each position. This group of six amino acids is designated as group 1.4 in figure 3.1. The experimental complexity of this library was approximately  $2.5 \times 10^4$  unique transformants. DIS examination of 3882 colonies, of which greater than 50% were properly ligated recombinants, revealed 18 mutants showing significant absorption peaks in the NIR. This represents a proportion of functional mutants of  $9 \times 10^{-3}$  [ =  $18 / (0.50 \times 3882)$ ].

It is informative to compare frequencies of functional mutants from different libraries to reach conclusions about which regions of a sequence are more sensitive to mutations (i.e., have more stringent structural requirements). To be able to do this, we must normalize for the number of sites,  $n$ , which were altered. For each site mutated, only a fraction of all 20 amino acids will usually be functional. We have called this fraction ( $x_i$ ) the *stringency* at position  $i$  (14).

If six residues are randomized, each of these mutated sites changes the frequency of functional mutants by a factor  $x_j$ . From the observed frequency of functional mutants ( $f$ ) for a library of  $n$  altered residues, one can calculate a value  $x$ , the geometric mean of the stringencies for all six sites, with the following equation.

$$x = (f)^{1/n} \quad (\text{eq. 1})$$

A mean stringency ( $x$ ) of 0.46 [ =  $(9 \times 10^{-3})^{1/6}$  ] is obtained for the residues of the group 1.4 combinatorial library. By comparison, if a particular sequence position accepted only one amino acid, it would have a stringency value of 1/20 (= 0.05).

Ten mutants of the group 1.4 library were isolated and characterized further. All of the isolated mutants displayed LH1-like spectra. Five LH2-like colonies were isolated but were found not to be CCM recombinants. Figure 3.3 shows the spectral differences between wild-type LH2 and mutant #1 isolated from the group 1.4 library. This mutant displays the characteristic absorption spectrum of an LH1 protein: a single absorption peak at 870 nm. Moreover, the measured OD of this mutant is about 90% of the value observed in WT LH2. The deduced sequence of mutant #1 in the altered region is also given in figure 3.3.

Figure 3.3. (Next page) The ground state absorption spectra of wild-type (WT) LH2 and mutant #1 were determined by examining spots of *Rb. capsulatus* liquid cultures on  $\text{RM}^{\text{tet}}$  plates by DIS from 730 nm to 930 nm. Both the optical density and wavelength scales are the same in both panels. Note the disappearance, in mutant #1, of the 800 nm band as well as the red shift from 855 nm to 870 nm. The sequences resulting in each phenotype are shown in the single letter amino acid code. Only the amino acids which were altered and, consequently, differ between the two proteins are shown. The group of residues mutated cover residues 28-33 of the LH2  $\beta$  subunit gene.

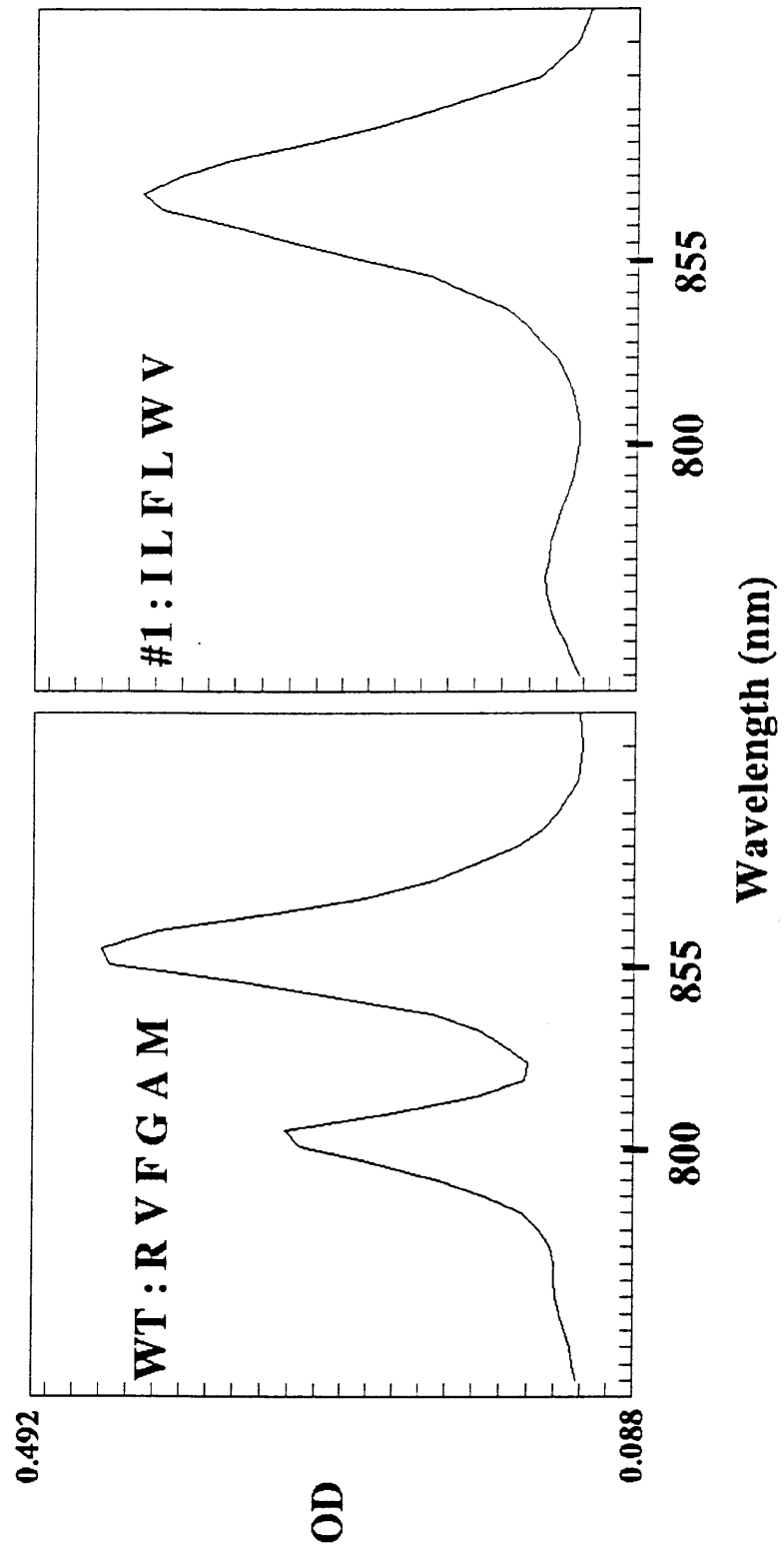


Figure 3.4 shows ten different mutants isolated from the group 1.4 library and sequenced in the mutated region. Note that mutant #2 shows a diminished 800 nm peak and a large peak at 860 nm. The ratio  $OD_{800}/OD_{855}$ , typically equal to 0.6 for WT LH2, is roughly 0.16 in mutant #2. All other mutants in figure 3.4 have a single peak absorbing between 850 nm (#8) and 870 nm (# 1 and #12). The corresponding sequences of all mutants shown in figure 3.4 are given in Table 3.2, along with the Hamming distances (number of mutations; ref. 23) relative to 1) the WT LH2 sequence, and 2) the homologous region of WT LH1. Thus, according to Table 3.2, there are five mutations between LH2 and mutant #1 and four mutations between the mutagenized region of mutant #1 and the homologous region of LH1. The comparison with the homologous LH1  $\beta$  sequence is based on the alignment proposed by Zuber (13). Mutant #1 has the greatest absorption at 870 nm of the ten mutants and the fewest mutations compared to LH1. Also, mutant #2 shows the closest phenotype to WT LH2 in this group, and shows the fewest mutations relative to LH2. Although too few sequences are known from the group 1.4 database to reach conclusions about the validity of Hamming distance for phenotype predictions, it is interesting that this parameter, instead of molar volume change, may be useful to solve the same sequence-phenotype problem under different mutagenesis circumstances.

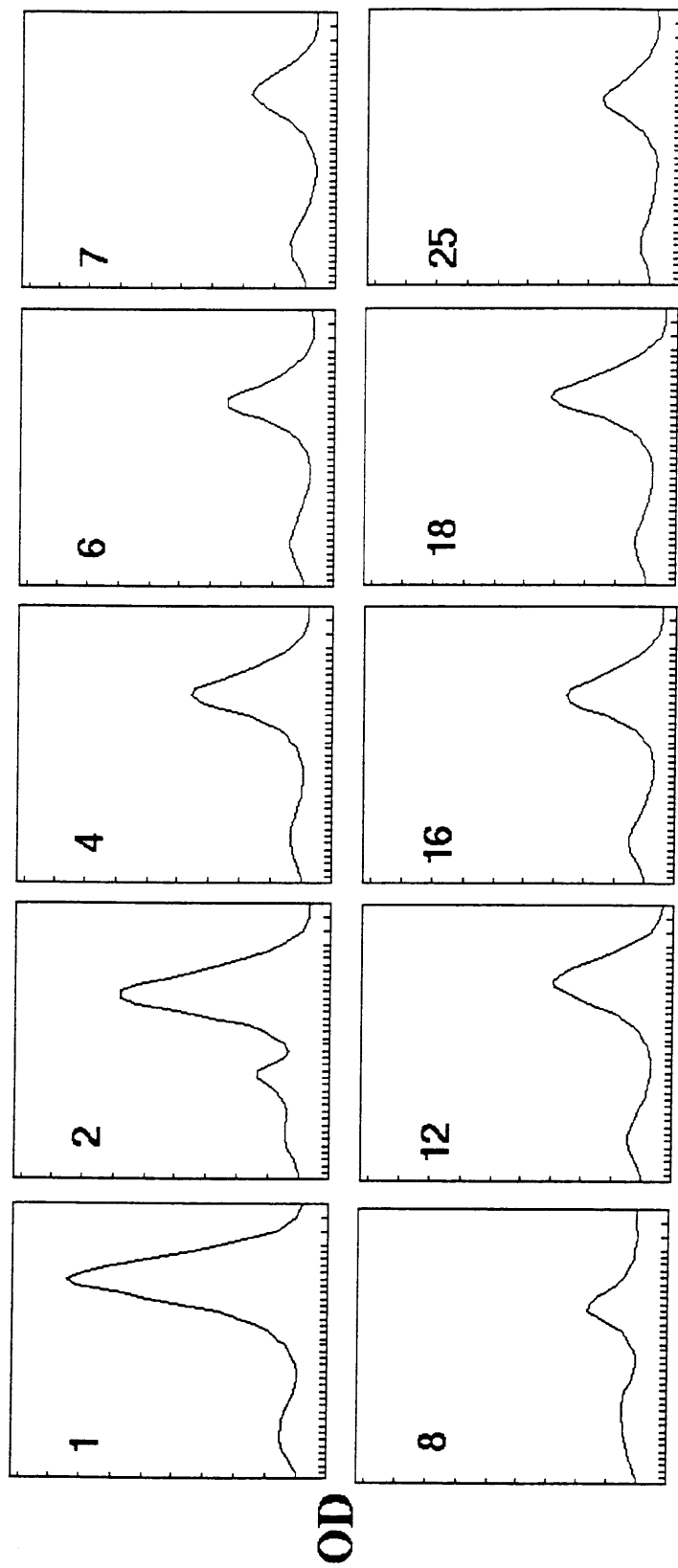
TABLE 3.2

Sequences of pseudo-LH1 mutants isolated from the group 1.4 random library.

Mutant #	Deduced amino acid sequence*	Hamming distance† to LH2	to LH1
WT LH2	R V F G A M	-	4
WT LH1	S A F I A V	4	-
1	I L F L W V	5	4
2	R V W A A G	3	5
4	Y Q W L L W	6	6
6	Y T F A C W	5	5
7	F Y W C G L	6	6
8	L M Y A W L	6	6
12	R W M L A T	4	5
16	K T W L Q L	6	6
18	F K L L G V	6	5
25	Q C Y G T Y	5	6

\* The amino acid sequences were deduced from nucleotide sequences determined in the region of the LH2  $\beta$ -subunit where mutations were introduced.

† The Hamming distance is the number of mutations in the sequence, relative to the WT (wild-type LH2) sequence or the homologous region in wild-type LH1.



Wavelength

OD



Figure 3.4. (Previous page) Ground state absorption spectra of ten mutants of the group 1.4 library. Spectra were obtained as in figure 3.3 and the scales of optical density and wavelength are the same, for all spectra, as the scales in figure 3.3. Note the variability in optical density and slight peak shifts. Also, mutant #2 shows a diminished band at 800 nm. The sequences of these mutants are shown in Table 3.2.

### 3.4 Discussion

The remarkable spectral changes obtained by specific mutagenesis of the group of amino acids shown in figure 3.1 were first observed by using a phylogenetically based mutagenesis strategy (6). For completeness, it is necessary to recount the relevant results of this study. Six amino acids were modeled to be on the same face of the putative  $\alpha$ -helix as the histidine ligand of the dimer (855 nm) bchl. Using phylogenetically derived target sets, these residues were altered simultaneously by CCM (Fig. 3.1). The resulting combinatorial library (termed TSM library) showed both LH1-like and LH2-like phenotypes in roughly equal numbers (Fig. 3.2). The sequences of numerous unique mutants were determined and analyzed to see if phenotype could be correlated somehow with genotype (Table 3.1). A comparison by simple inspection of the LH1-like and LH2-like sequences revealed that the size of the amino acid at a single position (Gly $\beta$ 31, seven residues to the amino terminal side of the histidine ligand) seemed to determine the observed phenotype. Specifically, the presence at this position of an amino acid larger than Thr was sufficient to cause a pseudo-LH1 spectrum in 90% of cases. It can be seen in Table 3.2 that this heuristic "rule" used to predict the phenotype of a sequence from the phylogenetic TSM library is not applicable to the sequences of the random group 1.4 library. Mutants #2, #6, #7, #8 and #25 have amino acids smaller than Thr at position  $\beta$ 31 but show LH1-like phenotypes, in contradiction with the rule. A Chi-square test applied to this data shows that statistically, the observed discrepancy is not consistent with the heuristic rule ( $\chi^2 = 10$ ,  $p > 0.01$ ). Therefore, the rule is context dependent. Gly $\beta$ 31 is the only position mutated in the TSM library which overlaps with group 1.4 (Fig.3.1). Of all the *random* libraries discussed, group 1.4 is the only one which yielded pseudo-LH1 phenotypes. Optimized libraries previously described (4,14) spanning this

region were not found to contain pseudo-LH1 mutants. For these reasons, we conclude that alterations in this particular *region* (i.e., group 1.4) of the primary structure are sufficient to eliminate an absorption peak (800 nm) and cause a red-shift in another (855 nm to 870 nm).

As more information becomes available on the atomic structure of LH2 (24,25), it will be very informative to attempt to find the exact structural basis of our observations. However, it is possible to speculate on the properties of the region of LH2 mutagenized in our experiments. In the TSM library, the main determinant of the phenotype change is Gly $\beta$ 31. One might hypothesize that specific interactions required for the LH2 spectral characteristics are disrupted by steric interference produced by large substitutions at Gly $\beta$ 31. In addition, the fact that information at this position is essentially sufficient to predict phenotype in Table 3.1 suggests that the other five amino acids which were also mutated in the TSM library could not interact with position  $\beta$ 31. While the sequence data from the group 1.4 library are not so simply analyzed, they are not inconsistent with the above results but merely reflect a more complex situation. Residues adjacent to  $\beta$ 31 were altered in the group 1.4 library which apparently interact with this position. Consistent with such interactions, no single residue was seen as a determinant of phenotype. This observation and the fact that the TSM library rule is not observed in the group 1.4 data may reflect a complex summation of contributions to the phenotype and stability of the protein instead of a single important contribution. Another aspect of the data adds to the picture we are attempting to sketch. In the group 1.4 library, we observed a greater propensity for LH1 phenotypes to arise rather than LH2 phenotypes. If the requirements for an LH1 phenotype are less stringent, the protein will assemble more readily into an LH1-like structure. This would make the occurrence of an LH2-like mutant a less likely event than that of a pseudo-LH1 mutant. The observed frequency of functional mutants was relatively high for the number of sites randomized. A stringency value of 0.46 was calculated for the six amino acid positions of group 1.4. In comparison, the mean stringency of the residues in groups 1.1, 1.2 and 1.3 were 0.22, 0.40 and 0.25, respectively (14). According to Table 3.2, the phylogenetically conserved Phe $\beta$ 30 (26 Phe and 3 Tyr out of 29 LH  $\beta$  subunit sequences; ref. 13) can be replaced by Leu, Met or Trp. These data suggest that the structural requirements for maintaining a stable structure are less rigorous than in the Ala-x-x-x-His motif (group 1.3,

stringency of 0.25) or the carboxy terminal six residues (group 1.1, stringency of 0.22). In sum, specific, sensitive interactions necessary to achieve an LH2 phenotype and maintain a stable protein are provided by residues such as Gly<sup>β31</sup> as well as other nearby residues. These interactions can be sterically hindered by bulky substitutions at β31 and are not readily affected by residues to the carboxy terminal end of β33. If the residues in this region merely achieve a stable structure, an LH1 phenotype is achieved instead.

An important difference between the TSM library and the group 1.4 library is that the former expressed a subset of all possible amino acids at its mutagenized positions. Because we are comparing sequence-phenotype data *after* selection for functional mutants, we expect our conclusions to be unaffected by this difference. Moreover, the expected frequencies of occurrence at the Gly<sup>β31</sup> position of the amino acids encoded in both libraries were similar. For instance, the relative probabilities of small amino acids over large amino acids in the TSM library was 1.37 (= 0.56/0.41), while this same ratio was 1.11 (= 0.30/0.27) in the group 1.4 library. From the latter ratio, the heuristic rule would predict similar numbers of pseudo-LH1 and pseudo-LH2 mutants in the group 1.4 library, in disagreement with observation. Nevertheless, it should be understood that the "context" of the TSM experiment includes the fact that target sets were encoded at the mutagenized positions.

We expect that the results of this work will generalize to other proteins and that phenotypic diversity will be achieved even in cases where amino acid residues are grouped arbitrarily. While groupings based on structural models that involve interacting amino acids might be preferred over random groupings of amino acids, we have shown that a simple, colinear grouping of six residues could still be useful in obtaining diverse phenotypes and identifying the sequence elements responsible for the observed changes. We have identified residues in the sequence which will trigger a phenotype change using a given grouping scheme and we have found that this information is still relevant in other grouping schemes. Our observations also indicate that numerous sequences that are very different, but none the less isofunctional, can be found by combinatorial mutagenesis of different groups of amino acid residues. This suggests that "pockets" of isofunctional proteins exist within a highly convoluted sequence space.

### **3.5 Notes and acknowledgments**

Effusions of gratitude go to Ellen Goldman who contributed the TSM library data (figure 3.2 and table 3.1). We wish to thank Jennifer Reyna for technical and culinary assistance. We also wish to thank KAIROS Inc. for software support. This iteration will be published in Delagrave,S., Goldman, E.R. and Youvan, D.C.,*Protein Engineering*, **8** (3), (1995).

### 3.6 References

1. Oliphant, A. R., Nussbaum, A. L. and Struhl, K. (1986) *Gene* **44**, 177-183.
2. Reidhaar-Olson J. F., Bowie J. U., Breyer, R. M., Hu, J. C., Knight K. L., Lim W. A., Mossing M. C., Parsell D. A., Shoemaker K. R. and Sauer, R.T. (1991) *Methods Enzymol.* **208**, 564-587.
3. Reidhaar-Olson, J.F. and Sauer, R.T. (1988) *Science* **241**, 53-57.
4. Delagrave, S., Goldman, E.R. and Youvan, D.C. (1993) *Protein Eng.* **6**, 327-331.
5. Lowman, H.B. and Wells, J.A. (1993) *J. Mol. Biol.* **234**, 564-578.
6. Goldman, E.R. and Youvan, D.C. (1992) *Bio/Technology* **10**, 1557-1561.
7. Barbas, C.F., Bain, J.D., Hoekstra, D.M. and Lerner, R.A. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 4457-4461.
8. Zuber, H. (1986) *Trends Biochem. Sci.* **11**, 414-419.
9. Youvan, D.C. and Ismail, S. (1985) *Proc. Natl. Acad. Sci. U.S.A.* **82**, 63-67.
10. Youvan, D.C. (1994) *Nature* **369**, 79-80.
11. Fowler, G.J.S., Visschers, R.W., Grief, G.G., van Grondelle, R. and Hunter, C.N. (1992) *Nature* **355**, 848-850.
12. Tadros, M. H., Garcia, A., F., Gad'on, N. and Drews, G. (1989) *Biochim. Biophys. Acta* **976**, 161-167.
13. Zuber, H. (1990) in *Molecular Biology of Membrane-Bound Complexes in Phototrophic Bacteria*. eds. Drews, G. and Dawes, E.A. (Plenum Press, New York, NY) pp. 161-180.

14. Delagrave, S. and Youvan, D.C. (1993) *Bio/Technology* **11**, 1548-1552.
15. Ditta, G., Schmidhauser, T., Yakobson, E., Lu, P., Liang, X.W., Finlay, D.R., Guiney, D. and Helinsky, D.R. (1985) *Plasmid* **13**, 149-153.
16. Dower, W.J., Miller, J.F. and Raysdale, C. (1988) *Nucl. Acids Res.* **16**, 6127-6145.
17. Bylina, E.J., Jovine, R.V.M. and Youvan, D.C. (1989) *Bio/Technology* **7**, 69-74.
18. Arkin, A.P., Goldman, E.R., Robles, S.J., Coleman, W., Goddard, C.A., Yang, M.M. and Youvan, D.C. (1990) *Bio/Technology* **8**, 746-749.
19. Youvan, D.C., Goldman, E.R., Delagrave, S. and Yang, M.M. (1994) *Methods Enzymol.* **276**, In press.
20. Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) *Molecular cloning: a laboratory manual* (2nd Edition). Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
21. Suzuki, D.T., Griffiths, A.J.F., Miller, J.H., and Lewontin, R.C. (1986) *Introduction to Genetic Analysis* (W.H. Freeman and Company, New York), p.92.
22. Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1988) *Numerical Recipes in C* (Cambridge University Press, Cambridge), pp. 487-490.
23. Schuster, P. and Swetina, J. (1988) *Bull. Math. Biol.* **50**, 635-660.
24. Donnelly, D. and Cogdell, R.J. (1993) *Protein Eng.* **6**, 629-635.

25. Cogdell, R.J. and Hawthornthwaite, A.M. (1993) in *The Photosynthetic Reaction Center*. eds. Deisenhofer, J. and Norris, J.R. (Academic Press, San Diego, CA), Vol. 1, pp.23-42.

## **Fourth Iteration. Optimized Combinatorial Libraries of Antibodies, an Application.**

### **4.0 Summary**

The immune system is endowed with a large repertoire of antibodies (about  $10^7$ ) capable of binding with moderate affinity almost any molecule it encounters. Recent efforts have been made to reproduce this naive repertoire *in vitro* by combinatorial mutagenesis of antigen-binding loops in antibody sequences. These libraries are limited to roughly the same number ( $10^7$ ) of different sequences, but will contain a large proportion of non-functional sequences because the introduced mutations are random. Phylogenetic information and our optimization algorithms can be used to direct libraries towards the expression of functional sequences. Because of the large number of possible sequences in this optimized library, the resulting ensemble of antibodies is expected to remain naive. We discuss the design and construction of such an optimized library which more closely duplicates the naive repertoire available in mammals, while offering the flexibility, speed and low cost of *in vitro* manipulations.



#### 4.1 Introduction.

The effectiveness of antibodies lies in their ability to recognize and bind essentially any molecule with very high specificity and affinity. This capability resides both in their three dimensional structure and in the genetics of the immune system. Figure 4.1 shows a schematic diagram of the structural organization of the heavy (H) and light (L) subunits of an antibody.

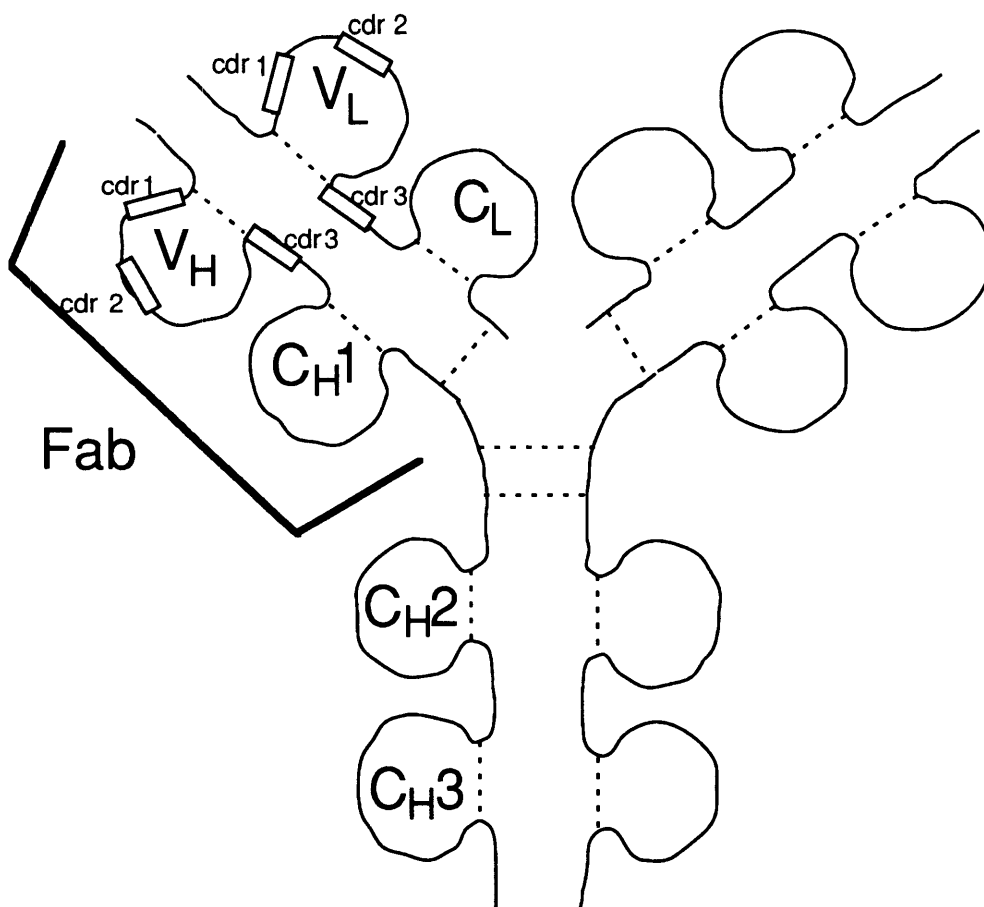


Figure 4.1. Schematic diagram of an IgG antibody structure. Two identical copies of the heavy (H) and light (L) polypeptide chains are present. The variable domains (V) as well as their constant domains (C) are shown. The six Complementarity Determining Regions (CDRs, numbered 1 to 3 for each chain) constitute regions of the sequence where most of the variability occurs and also contain the majority of

antigen binding residues. The Fab domain is bracketed and comprises a truncated H chain and a full length L chain. Disulfide linkages are shown as dashed lines.

Antibody fragments with binding activities nearly identical to those of the original antibody have been expressed successfully in *E. coli* (Winter and Milstein, 1991). One type of fragment called an Fab (see figure 4.1) can be obtained by proteolytic (papain) digestion of an antibody, or by co-expression of an Fd (truncated heavy chain gene) and an L gene. Fabs have been found to crystallize more readily than whole antibodies, presumably because the latter have too much flexibility at the hinge region, where the two Fab domains join with the Fc domain. Three-dimensional crystal structures have been solved for several Fabs, but the experimental strategies described in this thesis require no accurate modeling of three-dimensional structure.

To engineer antibodies, one may draw from a wealth of information on their genetics and structure. Kabat *et al.* (1991) have compiled hundreds of antibody sequences from various sources. Homology alignments and the analysis of sequence variability have identified certain regions of primary structure associated with antigen recognition. These Complementarity Determining Regions (CDRs) were found to be highly variable in sequence and to hold key positions in the tertiary structure of antibodies, as determined by X-ray crystallography. Crystal structures of antigen-antibody complexes exist and show the importance of CDR residue contacts with the antigen (Colman *et al.*, 1989; Davies *et al.*, 1988; Padlan *et al.*, 1989). Mutagenesis studies of antigen-antibody interactions have been made which confirm the importance of CDR residues in binding (Glockshuber *et al.*, 1991) but also point to unexpected contributions from residues not directly interacting with the antigen (Chien *et al.*, 1989).

Extensive analysis by Chothia and Lesk (Chothia and Lesk, 1987; Chothia *et al.*, 1989; Chothia *et al.*, 1992) of structural and genetic data on immunoglobulin variable domains has led them to propose a "canonical structure" model of the backbone conformations of CDRs. According to this model, a limited repertoire of structures can be adopted by antigen binding loops. Which conformation is taken depends mostly on the amino acid residues

present at a few key sites within, and in the vicinity of, each CDR. Altering residues at key sites is proposed to change CDR backbone conformation while changing other CDR residues modifies the surface exposed to the antigen.

The immune system of an individual (human or mouse) is expected to have on the order of  $10^7$  to  $10^8$  different antibody sequences. This repertoire is called *naive* because almost any antigen can be recognized with moderate ( $K_d \sim 10^{-6}$  M) affinity by one or more of these antibodies. Subsequent to the initial recognition event of a new antigen by an antibody, affinity maturation is achieved by the immune system through somatic mutation of the antibody sequence. Improved variants of the initial antibody are generated in a few days to better bind and neutralize the invading antigen.

A naive library of antibodies can be constructed artificially by randomizing one or more CDRs, using molecular genetics techniques (Barbas, 1992a). However, the resulting library will contain a majority of sequences which will not fold into functional antibodies. One possible course of action to reduce the proportion of non-functional sequences present in the library is to introduce a bias towards the expression of amino acids which are known to lead to functional sequences. This can be done using phylogenetic information and optimization functions, such as PG and SSD, described in the previous chapters and elsewhere (Goldman and Youvan, 1992). The purpose of this chapter is to discuss the design and construction of a naive combinatorial library of antibodies which is optimized, using known antibody sequences, to express functional antibodies. This optimization, however, should not take away the ability of this library to produce antibodies recognizing almost any antigen. In other words, despite the optimization, the library should remain naive. Such a library could be advantageous to isolate rapidly a number of antibodies with different specificities and moderate to good affinities.

## 4.2 Materials and Methods

### *Library Design*

Two types of libraries were designed. Library NNK16 fully randomized the 16 residues of the heavy chain CDR3 (HCDR3) of Fab F22. Library KABAT is an optimized combinatorial library introducing mutations in the 16 residues of the HCDR3 of Fab F22. A phylogeny of sequences was obtained from the Kabat compilation of antibody sequences (Kabat, 1991). There are 79 full-length and fragmentary sequences in the subgroup called "human heavy chains subgroup I". Only unique sequences better than 90% complete were used to design the library. Specifically, 27 sequences from the "human heavy chains subgroup I" section were used: 1\* LS2'CL, 7 1B9/F2'CL, 10 21/28'CL, 11 8E10'CL, 14 51P1'CL, 15 AND'CL, 16 NEI'CL, 17 HP1'CL, 21 783c'CL, 22 X17115'CL, 23 TH9'CL, 24 WIL2'CL, 25 EVI-15'CL, 26\* KAS, 27\* BOR', 28 RF-TS1 'CL, 30 ND'CL, 32 EU, 33 RF-TS3 'CL, 38 83P2'CL, 39 MOT, 40 WS1 'CL, 41\* Ab2022'CL, 42\* SIE, 43 lambda IGD-1'CL, 44 Ab2'CL, 45\* WOL. An important assumption in this choice of sequences is that they will lead to antibodies which fold into stable conformations capable of recognizing antigens. Six of the sequences (asterisk following their number) represent antibodies with known binding activities, clearly stable. Moreover, since most sequences were either obtained from direct protein sequencing or from rearranged genes, many of which show evidence of a stable antibody being expressed, this assumption is not unreasonable.

The other important step is to align the sequence of Fab F22 with the phylogeny. In the framework regions, the alignment is straightforward as the sequence of F22 is better than 80% identical to the consensus sequence. The task is made difficult by the inherent variability of the sequences in the HCDR3. However, the large number of residues (16) in HCDR3 of F22 dictates that positions 95 to 100 and 101 and 102 should all be occupied by an amino acid. Positions 100A to 100K (between 100 and 101) correspond to residues where gaps can be introduced somewhat arbitrarily to improve the alignment with the

consensus sequence. The final alignment used for the design of the library was:

```
95 96 97 98 99 100 A B C D E F G H I J K 101 102
G V N L F R V R N S R P H L D M.
```

We can now look at the frequency of occurrence of amino acids at each position of HCDR3 and enter the data in a computer program called CyberDope (KAIROS Scientific Inc.). As described in "Results and Discussion", the program calculates a nucleotide mixture for each codon to be mutagenized. These mixtures are then used to design the oligonucleotides necessary to construct the combinatorial library. The sequence of oligonucleotide NNK16 was: 5'p-GGT GAC CGT GGT CCC TTG GCC CCA (MNN)<sub>16</sub> TCT CGC ACA ATA ATA CAC GGC-3'. The sequence of oligonucleotide KABAT was: 5'p-GGT GAC CGT GGT CCC TTG GCC CCA AWH ATC MAW ABM ADV ANV MHH ANV ANH ANY ANN MYY MHV MNN MBV MBY ACG CGC ACA ATA ATA CAC GGC-3'. The sequences of these oligonucleotides are antisense. The underlined codons both encode arginine but were chosen as a means of determining whether a given mutant is from one or the other library. Both oligonucleotides were 5' phosphorylated. Integer mixtures of nucleotides are denoted by the following single letter code:

M=A:C (1:1) R=A:G (1:1) W=A:T (1:1) S=C:G (1:1) Y=C:T (1:1) K=G:T (1:1)  
V=A:C:G (1:1:1) H=A:C:T (1:1:1) D=A:G:T (1:1:1) B=C:G:T (1:1:1)  
N=A:C:G:T (1:1:1:1).

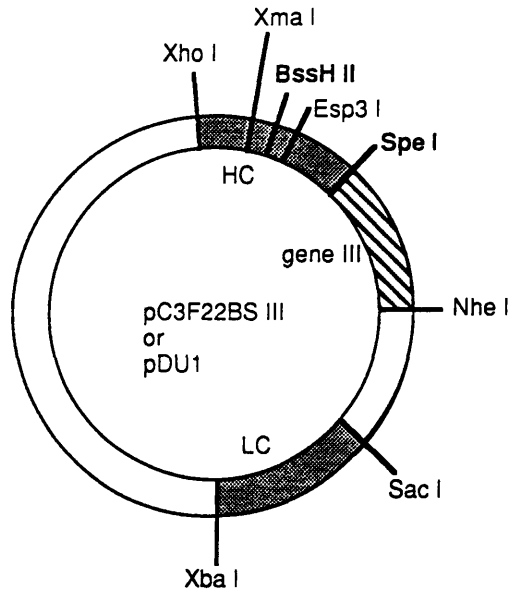


Figure 4.2. Plasmid pC3F22BSIII (also referred to as pDU1) is an expression vector with two lac promoters directing the expression of the light chain (LC) and heavy chain (HC). The HC is in frame with the carboxy terminal half of gene III. Bss HII and Spe I (in bold) were engineered as described.

#### *Phage-Display Vector Construction*

Standard DNA manipulations were carried out as described (Sambrook et al., 1989). DNA sequencing was carried out with a kit from US Biochemicals (double-stranded templates required plasmid purification with a Qiagen spin-prep kit). The Fabs were expressed in *E. coli* using a phage-display vector reconstructed from plasmids pC3F22 and pComb3 (Burton et al., 1991; Barbas et al, 1992a; see Figs 4.2 and 4.3). Reconstruction of pDU1 involved many steps. First, it was necessary to subclone a Xho I - Sac I fragment of pC3F22 into M13mp19; the resulting construct was called M13Ab1.3. BssHII and SpeI restriction sites were engineered by site-directed mutagenesis using a BioRad Mutagen Kit. The mutant Xho I - Sac I fragment was then excised from M13Ab1.3 and used to replace the wild-type Xho I - Sac I fragment of pC3F22 to create plasmid pC3F22BS. With the Spe I restriction site in place, both the LC and HC genes could be inserted into pC3 to yield the phage-display vector pC3F22BSIII, also abbreviated pDU1.

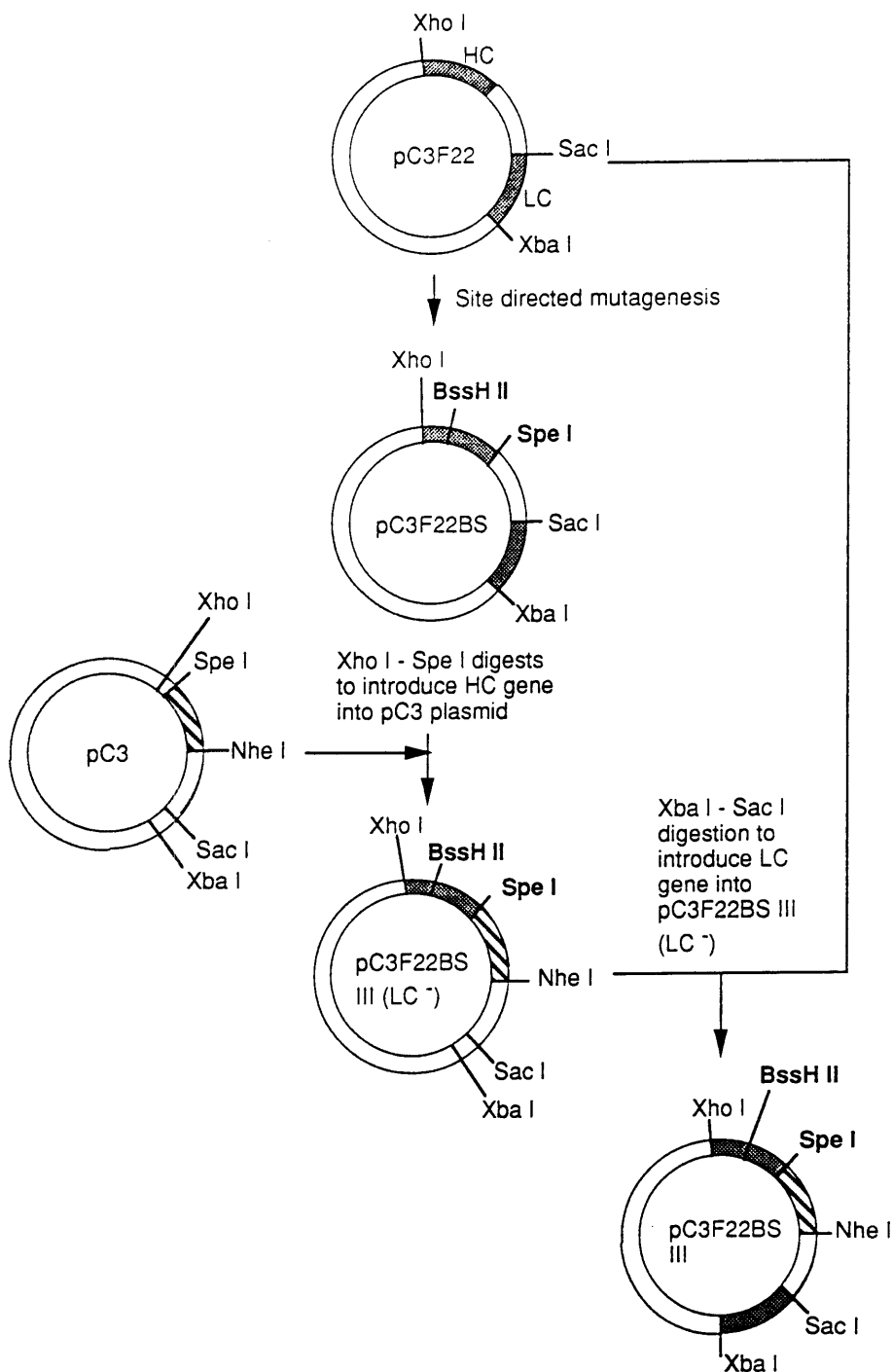


Figure 4.3. Construction of pC3F22BSIII (or pDU1).

Construction of pC3F22BS was achieved by digesting pDU1 with *Nhe I* and *Spe I* (leaving mutually cohesive ends) and self-ligating the digested, gel purified vector. This eliminates gene III (*Spe I* - *Nhe I* fragment) to yield a vector expressing the soluble Fab F22.

### *Library Construction*

Construct M13Ab1.3 containing the engineered Bss HII restriction site was used as template of two site-directed mutagenesis reactions (BioRad mutagene kit, reaction as specified in kit protocol but scaled up 8-fold). The oligonucleotides KABAT and NNK16 were used in separate reactions. The extension products were phenol:chloroform and chloroform extracted. The aqueous fractions were subjected to ultrafiltration (Amicon) to remove salts and traces of chloroform and to reduce their volumes to about 10 $\mu$ L. The resulting samples were then electroporated into XL1-Blue competent cells (Stratagene). Aliquots of the transformations were plated to determine the complexity of the resulting library of altered genes: both libraries contained approximately 10<sup>7</sup> transformants.

The libraries of altered genes must then be introduced into the phage-display vector pDU1. This is done by excision of the Xho I - Sac I mutant fragments from M13Ab1.3 and ligation into the Xho I - Sac I digested vector. Both DNA fragments were gel purified prior to ligation. The resulting libraries were electroporated again and the complexities were determined for both: pDUKABAT = 10<sup>7</sup>, pDUNNK16 = 10<sup>6</sup>. In order to compare both libraries, an aliquot of the pDUKABAT library was taken to reduce its complexity to 10<sup>6</sup>.

### *Affinity Selection*

Once the libraries were synthesized, selection of functional mutants was carried out as described (Burton *et al.*, 1991; Barbas *et al.*, 1992b; Collet *et al.*, 1992), by affinity panning against BSA covalently linked to Rhodamine B-isothiocyanate (Sigma, St-Louis, MO). The antigen is adsorbed onto a solid support (polystyrene microtitre tray, Costar) and exposed to a suspension of recombinant M13 phage expressing the mutant Fabs on their surfaces. Other antigens such as recombinant gp120 from HIV-1 (Intracel), hydrocortisone-BSA and  $\beta$ -estradiol-BSA (Sigma) were also used in additional panning experiments. After repeated washes of the titer tray, remaining clones were eluted from the tray with a 10  $\mu$ M rhodamine B solution (or with pH 2.2 glycine-HCl solution, as in the case of gp120 panning) and used to infect a fresh XL1-Blue culture. The overnight culture amplified the eluted phage which were then isolated from the culture and reapplied to a microtiter tray for another round of



affinity selection. After the fourth round of selection, individual phage were isolated by plating and their DNA was purified to allow deletion of gene III from the plasmid, using convenient restriction sites (Spe I and Nhe I). This construct expresses free Fab in *E.coli* XL1-Blue. A rapid analysis of mutant Fab affinity can be carried out by releasing Fabs from the periplasmic space of bacterial cells using a simple freeze-thaw protocol. The Fab extract is then used in an ELISA against BSA-rhodamine B conjugate (or other appropriate hapten such as gp120). Absorbance at 405 nm is determined using an ELISA tray reader (BioRad).

### **4.3 Results and Discussion**

To develop their theory of canonical structures, Chothia and Lesk relied in part on the observation that certain residues remained unchanged in antibody sequences for which a structure was known. The heavy chain CDR3, however, is the most variable of the six CDRs and no canonical structures could be proposed (Chothia and Lesk, 1987). Examination of the sequences in the human group I phylogeny identifies some positions where a clear trend in hydrophathy or molar volume constraints can be found. It seems likely that these trends reflect some structural requirements of the protein, independent from the exact composition of a particular HCDR3.

The observed trends are revealed by the task of designing an optimized combinatorial library. Table 4.1 shows, for three of the 16 positions in the sequence, the set of observed amino acids (target set) and the set of amino acids encoded by the nucleotide mixture which best reproduces the frequency of amino acids observed in the target set. These amino acids, with their frequencies of occurrence, were used as the data necessary to define a codon nucleotide mixture which would optimally reproduce the observed frequencies of occurrence. To do this, the program CyberDope establishes a table of all the possible codon nucleotide mixtures and determines which amino acids are encoded by each mixture, and with what frequency. This table of amino acid frequencies is then compared with the frequencies observed at some position in the sequence (e.g., position 97). The codon nucleotide mixture which best matches the observed frequencies of amino acids is chosen for the

combinatorial library. To carry out this comparison, a Sum-of-the-Squares-of-the-Differences (SSD) formula was used (Youvan et al., 1992; Goldman and Youvan, 1992; Delagrave and Youvan, 1993).

Let us consider the three examples shown in table 4.1. In the phylogeny, position 97 displays a wide variety of amino acids, each occurring at a fairly low frequency. The algorithm found that such diversity was best reproduced by simply using a random codon (NNK). In other words, no trends in hydropathy or molar volume were seen at this position, possibly because it makes no contribution to the overall structural stability of the antibody. At position 100K, two trends are readily noticed. Large hydrophobic residues are preferred and, among these, Phe is most favoured. The algorithm reflects part of this trend by encoding four large hydrophobic residues equiprobably. Finally, position 101 shows that an Asp is strongly preferred among a fairly diverse group of residues. This preference is considered sufficiently strong by the algorithm to only encode Asp at that position. Asp101 is known in some structures (e.g., MCPC603) to form a salt bridge with strongly conserved framework residue Arg94.

An important consequence of the optimization of a combinatorial library is its decrease in complexity compared with a random library of the same size. Theoretically, the random NNK16 library encodes  $20^{16}$  ( $= 7 \times 10^{20}$ ) possible amino acid sequences. In contrast  $7 \times 10^{14}$  sequences are encoded by the KABAT optimized library. This corresponds to a million-fold decrease in complexity. Since deleterious amino acids are discarded preferentially from the KABAT library, a large increase in the proportion of functional mutants is expected. In fact, a part of the increase in this ratio can be calculated based on the elimination of numerous stop codons from the mutant sequences: the probability of not finding a stop codon in the random library is 60% while this number is 80% in the KABAT library.

**Table 4.1 Optimization of three positions in the sequence of Fab F22.**

<u>Position #</u>	<u>Observed Amino Acids</u>	<u>SSD Optimization Result</u>
97	C(.04) E(.07) G(.15) H(.04) I(.15) K(.04) L(.07) P(.07) R(.11) S(.11) T(.04) Y(.11)	A(.06) C(.03) D(.03) E(.03) F(.03) G(.06) H(.03) I(.03) K(.03) L(.09) M(.03) N(.03) P(.06) Q(.03) R(.09) S(.09) T(.06) V(.03) W(.03) Y(.03) X(.03)
100K	F(.61) L(.06) M(.28) Y(.06)	F(.25) I(.25) L(.25) M(.25)
101	A(.04) D(.74) G(.07) N(.04) P(.04) Y(.07)	D(1.0)

The sequence of Fab F22 was aligned with a phylogeny of human antibody sequences. At a given position in the phylogeny, certain amino acids were found to occur at some frequency (shown in brackets) and are listed as "Observed amino acids". Optimized nucleotide mixtures were found which result in the expression of amino acids at a certain frequency of occurrence ("SSD Optimization Result").

The experiments yielded some clones with engineered specificity. Panning the pDUKABAT library against rhodamine B-BSA produced three clones out of 29 which gave an ELISA signal (OD<sub>405</sub> after 30 minutes) of greater than 0.4. This is approximately 5-fold stronger than the parent clone (wild-type F22) gives in an ELISA against rhodamine B. Panning of pDUNNK16 did not yield clones with similar affinities to rhodamine B. This result is consistent with the idea that library pDUKABAT is optimized to yield higher proportions of functional antibodies. Unfortunately, panning of either the random or optimized libraries against gp120,  $\beta$ -estradiol, and hydrocortisone

did not produce any clones with significant ELISA signals. Moreover, the clones with affinity for rhodamine B give much weaker ELISA signals than does clone F22 against fluorescein (OD<sub>405</sub> = 2.5, after 30 minutes).

The results of the affinity selection were unfortunately inconclusive despite a careful examination of potential problems. Combinatorial mutants randomly picked for sequencing did not reveal any problem with the libraries such as strong unwanted biases in the relative amounts of nucleotides or incorrect incorporation of an unwanted nucleotide at some residue position. A test of affinity selection was carried out on a random library. This test involved affinity panning of the library against fluorescein-BSA conjugate. Because a small, undetermined amount of wild-type F22 antibody clone is present in the library, it is expected that this clone would be selectively amplified along with any other mutant with significant affinity for fluorescein. Pools of selected clones, after each of four rounds of affinity selection, were sequenced and showed a gradual decrease in random sequence and concomitant increase in the wild-type sequence. That is to say, in the sequencing gel, the sequence of the first pool of clones showed four equally intense bands at the random nucleotide positions while the sequence of the fourth pool of clones was essentially that of wild-type F22. This shows that affinity selection was indeed functioning, although the extent of amplification is not known because the initial amount of wild-type F22 clone in the random library is undetermined.

One possible source of difficulty is the small size of the combinatorial libraries examined. Libraries 50-fold larger were used in the past (Barbas et al., 1992a) to isolate clone F22 by panning against fluorescein. Future efforts to achieve the goals described in this chapter may also focus on the affinity selection procedure. Although it was demonstrated to work to some extent, it may be that other methods are better suited to carry out affinity selection reproducibly (Dueñas and Borrebaeck, 1994; Lowman and Wells, 1993). New approaches are being investigated by members of the Youvan laboratory (Rob Mitra and Rachael Hawtin).

#### 4.4 References

Barbas, C.F., Bain, J.D., Hoekstra, D.M. and Lerner, R.A. (1992a). Semisynthetic combinatorial antibody libraries: A chemical solution to the diversity problem. *Proc Natl. Acad. Sci. U.S.A.*, **89**, 4457-4461.

Barbas, C.F.III, Björling, E., Chiodi, F., Dunlop, N., Cababa, D., Jones, T.M., Zebedee, S.L, Persson, M.A.A., Nara, P.L., Erling, N. and Burton, D.R. (1992b) Recombinant human Fab fragments neutralize human type 1 immunodeficiency virus *in vitro*. *Proc Natl. Acad. Sci. U.S.A.*, **89**, 9339-9343.

Burton, D.R., Barbas, C.F.III, Persson, M.A.A., Koenig, S., Chanock, R.M., Lerner, R.A. (1991). A large array of human monoclonal antibodies to type 1 human immunodeficiency virus from combinatorial libraries of asymptomatic seropositive individuals. *Proc Natl. Acad. Sci. U.S.A.*, **88**, 10134-10137.

Chien, N.C., Roberts, V.A., Giusti, A.M., Scharff, M.D. and Getzoff, E.D. (1989) Significant structural and functional change of an antigen-binding site by a distant amino acid substitution: Proposal of a structural mechanism. *Proc. Natl. Acad. Sci. USA.* **86**, 5532-5536.

Chothia, C. and Lesk, A.M. (1987) Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* **196**, 901-917.

Chothia, C., Lesk, A.M., Tramontano, A., Levitt, M., Smith-Gill, S.J., Air, G., Sheriff, S., Padlan, E.A., Davies, D., Tulip, W.R., Colman, P.M., Spinelli, S., Alzari, P.M., and Poljak R.J. (1989) Conformations of immunoglobulin hypervariable regions. *Nature* **342**, 877-883.

Chothia, C., Lesk, A.M., Gherardi, E., Tomlinson, I.M., Walter, G., Marks, J.D., Llewelyn, M.B. and Winter, G. (1992) Structural repertoire of the human VH segments. *J. Mol. Biol.* **227**, 799-817.

Collet, T.A., Roben, P., O'Kennedy, R., Barbas, C.F.III, Burton, D.R. and Lerner, R.A. (1992) A binary plasmid system for shuffling combinatorial antibody libraries. *Proc Natl. Acad. Sci. U.S.A.*, **89**, 10026-10030.

Colman, P.M., Tulip, W.R., Varghese, J.N., Tulloch, P.A., Baker, A.T., Laver, W.G., Air, G.M. and Webster, R.G. (1989) *Phil.Trans. R. Soc. Lond. B.* , **323**, 511-518.

Davies, D.R., Sheriff, S. and Padlan, E.A. (1988) Antibody-antigen complexes. *J. Biol. Chem.* **263**, 10541.

Delagrave, S. and Youvan, D.C. (1993) Searching sequence space to engineer proteins: exponential ensemble mutagenesis. *Bio/Technology.* **11**, 1548-1552.

Dueñas, M. and Borrebaeck, C.A.K. (1994) Clonal selection and amplification of phage displayed antibodies by linking antigen recognition and phage replication. *Bio/Technology*, **12**, 999-1002.

Glockshuber, R., Stadlmüller, J. and Plückthun, A. (1991) Mapping and modification of an antibody hapten binding site: a site-directed mutagenesis study of McPC603. *Biochemistry*, **30**, 3049

Goldman, E.R. and Youvan, D.C. (1992). An algorithmically optimized combinatorial library screened by digital imaging spectroscopy. *Bio/Technology*, **10**, 1557-1561.

Kabat, E.A., Wu, T.T., Perry, H.M., Gottesman, K.S. and Foeller, C. (1991) *Sequences of proteins of immunological interest*. U.S. Dept. of Health and Human Services, Washington, D.C.

Lowman, H.B. and Wells, J.A. (1993) *J. Mol. Biol.* **234**, 564-578.

Padlan, E.A., Silverton, E.W., Sheriff, S., Cohen, G.H., Smith-Gill, S.J. and Davies, D.R. (1989) Structure of an antibody-antigen complex: crystal structure of the HyHel-10 Fab-lysozyme complex. *Proc. Natl. Acad. Sci. USA* **86**, 5938.

Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) *Molecular cloning: a laboratory manual* (2nd Edition). Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Winter, G. and Milstein, C. (1991) *Nature* **349**, 293-299.

Youvan, D.C., Arkin, A.P. and Yang, M.M. (1992). Recursive ensemble mutagenesis: A combinatorial optimization technique for protein engineering. *In: Parallel Problem Solving from Nature*, 2, pp. 401-410. R. Maenner and B. Manderick (Eds.) Elsevier Publishing Co., Amsterdam.