Imaging Spectroscopy and Combinatorial Mutagenesis of Light Harvesting II Antenna from *Rhodobacter capsulatus*

by

Ellen R. Goldman

B.S. *summa cum laude* in Chemistry
University of Massachusetts at Amherst (1988)

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
February, 1994

Signature of Author ______________________________
November, 29 1993

Certified by ______________________________
Professor Douglas C. Youvan, Thesis Supervisor

Accepted by ______________________________
Professor Glenn A. Berchtold
Chairman, Departmental Committee on Graduate Students

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

Professor Keith A. Nelson ______________________________
Chairman

Professor Douglas C. Youvan ______________________________
Thesis Supervisor

Professor Joanne Stubbe ______________________________
Committee Member

Professor James R. Williamson ______________________________
Committee Member

Imaging Spectroscopy and Combinatorial Mutagenesis of
Light Harvesting II Antenna from *Rhodobacter capsulatus*

by

Ellen R. Goldman

Submitted to the Department of Chemistry
on November 29, 1993 in partial fulfillment of the requirements for the
Degree of Doctor of Philosophy in Chemistry

Abstract

The light harvesting II (LHII) antenna of *Rhodobacter capsulatus* provides a model system for developing complex mutagenesis schemes, because ground state absorption spectroscopy can be used to assay protein expression, structure, and function. LHII is modeled to include two transmembrane alpha helical polypeptides (α,β) that bind both a dimer and monomer of bacteriochlorophyll (Bchl). These Bchls can be specifically differentiated by their ground state near infrared absorption bands. A genetic system was designed to facilitate combinatorial cassette mutagenesis in several regions of the β subunit structural gene.

The first experimental implementation of target set mutagenesis (TSM) was performed in the LHII model system. Combinatorial cassettes based on phylogenetic "target sets" were used to simultaneously mutagenize seven amino acid residues on one face of a transmembrane alpha helix comprising a Bchl binding site. Colony screening by digital imaging spectroscopy (DIS) showed that 6% of the optimized library bound Bchl in two distinct spectroscopic classes. This is approximately 200 times the throughput (ca. 0.03%) of conventional combinatorial cassette mutagenesis using NN(G/C) at the same 7 sites (where N=25% A,T,G, and C).

Two schemes for formulating nucleotide mixtures from target sets of amino acids were compared experimentally. Phylogenetic data were used to construct target sets for seventeen sequence positions flanking the monomer Bchl binding site of LHII. Cassettes were constructed using both the group probability (P_G) algorithm, which encodes every amino acid in the target set no matter how infrequently it occurs and the sum of the squares of the differences algorithm (SSD), which may drop less used amino acids from the nucleotide mix. The SSD based library showed 2% pigment binding positives, while no positives were found in the P_G library with 10,000 mutants screened. This gives the SSD method at least a 200 fold throughput advantage over P_G in this experiment. Experiments using SSD to construct nucleotide mixtures show phenotypic diversity when using this formulation to construct libraries.

Correlations between protein sequence and phenotype were examined by inspecting databases of sequence and corresponding spectral information gathered from combinatorial cassette mutagenesis (CCM) experiments. Simple experimenter formulated decision algorithms achieved 80-84% accuracy in crude classifications of the data (protein assembly versus no assembly as judged by the screening or selection criteria). Neural network analysis of the sequence-phenotype data takes into account possible nonlinear interactions between the sequence positions, and shows approximately an 8% increase in correct classification.

The "doping" algorithms evaluated in the LHII model system as well as the methods for analysis of sequence and phenotype correlations should be generally applicable to other proteins. These techniques should lead to new advances in photosynthesis research as well as being applicable to protein engineering applications, especially phage-display libraries.

Thesis supervisor: Dr. Douglas C. Youvan
Title: Associate Professor of Chemistry

To my family (immediate and extended) who have been so loving and supportive. I am so lucky; you all are great!

And in memory of Vic and Scott, my two P chem buddies from UMass.

Acknowledgements

There are so many people to thank it is hard to figure out where to start! Thanks to Doug Youvan, my advisor for all his help, guidance, support, and advice in my thesis work. My graduate studies at MIT have certainly been interesting, and taught me alot more about university politics and the tenure process than I had bargained for.

Thanks to all the members of the Youvan lab for their friendship and scientific interaction. Mary provided me with software, hardware, and company during my impulse rodent shopping. Steve taught me the hands on basics of molecular biology; he is a great teacher, a fine dart partner, and a good friend. Adam started in the lab at the same time as I did, and my thesis work involved the experimental implementation of some of his theoretical studies. Simon came to the lab a few years after I did and his first project involved working with me on CCM of LHII. Georg joined the lab a year ago and collaborated with me on the final project of my thesis work. I have learned from the postdocs: Bill, with his wacky sense of humor, and then Kai, who offered alot of helpful advise during the "trauma" of thesis writing. I have worked with some fantastic undergrads (Christine, Joy, and Matt) I wish all of them well in their future scientific careers.

Of course a big thanks goes to my family and friends. My family has been wonderful in supporting me emotionally throughout the ups and downs of graduate school. My grandpa always wrote me encouraging letters, my uncle Manny gave me help and advice in my search for a postdoctoral position, my parents and siblings have been terrific throughout these 5.5 years. All my extended family is great! I am glad that I was a member of the P chem entering class of 1988, my classmates have been good friends. I was also lucky to be on the second floor of 56, I learned alot from the graduate students and postdocs on the hallway. Thanks to my friend Paul for his trips to Boston to rescue me from the lab! Again thanks to my labmate Steve for his friendship and introducing me to cactus collecting, and for reading my thesis even after he left for his postdoc in Texas. Steve helped me to keep my perspective and "sanity" during the first four years of my research, I hope we contiune to keep in contact with each other. Mark deserves a huge thanks (and a big hug!) for dealing with me during the last year of my research. He gave me emotional support through the craziness of finishing my research, writing and defending my thesis, writing a paper based on the last part of my research, and finding a postdoctoral position. Mark was understanding even when I was a bear for days at a time.

It feels good to finally be done with my graduate work! Thanks to everyone! As I finish revising this thesis we are packing the laboratory for its move out west. I wish Doug all the best in the future.

Table of contents

Chapters

Appendices

List of abbreviations

ANN	artificial neural network
Bchl	bacteriochlorophyll
bp	base pair
Bphe	bacteriopheophytin
CCD	charged coupled device
CCM	combinatorial cassette mutagenesis
CD	circular dichroism
Da	dalton
DA	decision algorithm
DIS	digital imaging spectroscopy
EEM	exponential ensemble mutagenesis
kb	kilo base
LH	light harvesting
LHI	light harvesting I
LHII	light harvesting II
ml	milliliter
µl	microliter
NIR	near infrared
nm	nanometer
P_G	group probability
pLH1	pseudo light harvesting I phenotype
pLH2	pseudo light harvesting II phenotype
Q_x	high energy bacteriochlorin ground state transition
Qy	low energy bacteriochlorin ground state transition
RC	reaction center
REM	recursive ensemble mutagenesis
SDM	site directed mutagenesis
SSD	sum of the squares of the differences
TSM	target set mutagenesis
WT	wild type

List of figures

List of tables

Chapter 1: Background and significance

The pigment-protein light harvesting II (LHII) from *Rhodobacter capsulatus*, in conjunction with digital imaging spectroscopy (DIS), provides a model system for developing and testing algorithmically optimized combinatorial cassette mutagenesis schemes. An introduction to LHII structure and genetics, a review of the role of mutagenesis in photosynthesis research, and an overview of massively parallel screening of mutants follows.

1.1 An introduction to LHII structure and genetics

The focus of this thesis is the design and implementation of combinatorial cassette mutagenesis techniques. LHII is used as a test system to characterize methods that can be generalized to any protein system. This introduction to LHII provides some background information about the LHII antenna to explain why it is an ideal model system to study complex mutagenesis schemes.

1.1.1 LHII in the membrane

The light reactions of photosynthesis are mediated by membrane incorporated pigment-protein complexes. Figure 1 shows a cartoon of the photosynthetic apparatus in the membrane. Light is absorbed by the bacteriochlorophylls (Bchls) and carotenoids of light harvesting (LH) proteins which direct the energy to the bacterial reaction center (RC) where charge separation takes place. There are two types of antenna complexes in *Rb. capsulatus*. The core antenna (LHI) is in direct proximity with the RC, while the peripheral antenna (LHII) surrounds this RC/core complex (Drews, 1985). The pigments in the antennae have varying absorption maxima, and energy is transferred in a directed way from the peripheral antennae to the core antennae and RC with increasing wavelength of maximum absorption (Zuber & Brunisholz, 1991, and references therein).

The LH complexes serve to greatly increase the absorbance cross section of the photosynthetic apparatus. The quantity of LHII is dependent on growth conditions such as light and temperature (Hawthornthwaite & Cogdell, 1991). In cells of *Rb. capsulatus* grown under low light, the relative amount of

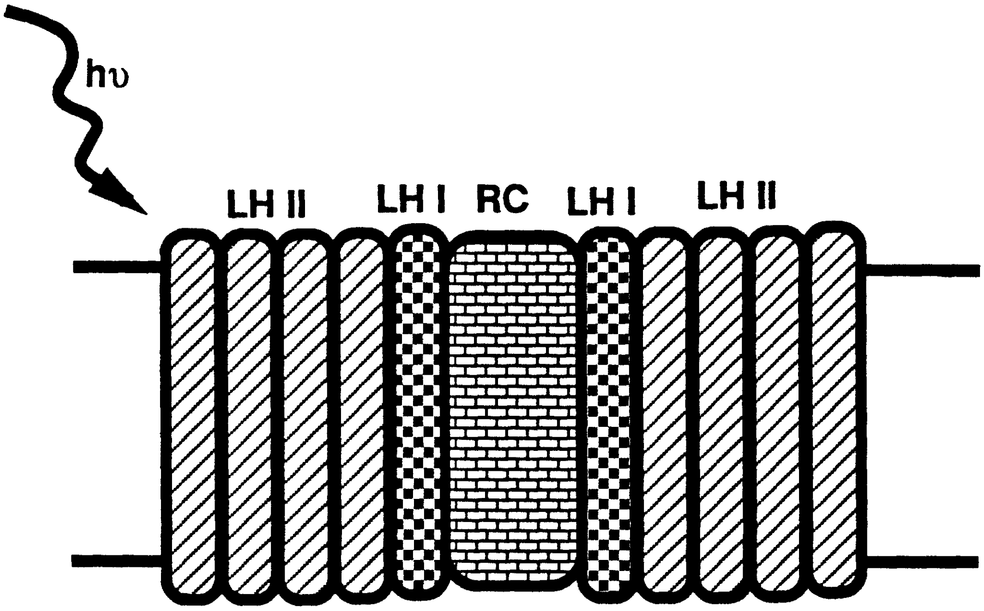


Figure 1. Cartoon showing the photosynthetic apparatus in the membrane. LHII (denoted by diagonal lines) is the peripheral light harvesting antenna, while the core antenna, LHI (denoted by checkers) is in direct contact with the RC (filled with brick). Light is captured by light harvesting complexes (I and II) and the excitation energy migrates to the RC where a charge separation event occurs.

RC:LHI:LHII is 1:20:(40-80) (Drews, 1985). The synthesis of LHII is also regulated by oxygen partial pressure (Marrs, 1981). Highly pigmented cultures can be obtained by growing *Rb. capsulatus* non-photosynthetically, semi-aerobically in the dark (Schumacher & Drews, 1978). LHII is present even in mutants where RC and LHI synthesis have been eliminated, as in Y142 (Drews *et al.*, 1976; Marrs, 1981).

1.1.2 LHII structure

1.1.2.1 General structural features

Much has been determined about the secondary and tertiary structure of peripheral LH proteins through conventional biochemical and spectroscopic techniques and examination of their primary structures. LHII is composed of three polypeptide subunits, three Bchls, and several carotenoids. Two subunits, α and β (180 and 147 bp respectively in *Rb. capsulatus* (Youvan & Ismail, 1985)), each contain three regions: an amino terminal hydrophilic segment on the cytoplasmic side of the membrane, a membrane spanning hydrophobic region, and a carboxy-terminal hydrophilic segment on the periplasmic side of the membrane (Zuber & Brunisholz, 1991). Ultraviolet circular dichroism, polarized infrared spectroscopy, and hydropathy plots of the amino acid sequence give evidence that the membrane spanning stretch of the LH α and β polypeptides has an alpha-helical structure (Drews, 1985). The α β dimer binds both a dimer and monomer of Bchl (Zuber, 1986; Figure 2). LHII possesses two strong absorption bands in the near infrared (NIR) at 800 nm and 855 nm (Figure 3). Circular dichroism (CD) analysis of the absorption bands has been interpreted to show that the 855 nm band has a strong dimeric character, while the 800 nm band is monomeric (Cogdell & Scheer, 1985). A third subunit, γ (14,000 Da), does not bind any pigments, but appears to be necessary for the assembly and stability of the complex (Welte *et al.*, 1985; Youvan & Ismail, 1985). The LHII proteins form aggregates of the basic unit (α,β,γ in *Rb. capsulatus*) (Zuber & Brunisholz, 1991, and references therein).

The orientations of the Q_X (low energy ground state transition) and Q_Y (high energy ground state transition) transitions of the Bchls in LHII were determined by linear dichroism spectroscopy (Kramer *et al.*, 1984; see Figure 2). The Q_Y transitions of the Bchl 800 (monmer) and Bchl 850 (dimer) were

Figure 2. (From Kramer *et al.*, 1984) Schematic representation of LHII α and β subunits. This model shows the membrane-spanning alpha helices and the pigments for a dimer of LHII (two α β units each binding a dimer and monomer of Bchl). The dimer Bchls are shown by the upper boxes and the monomers are represented by the lower boxes. The Q_y transitions are shown with the open arrows, and the Q_x transitions are shown in solid arrows. The zig-zag lines represent carotenoids.

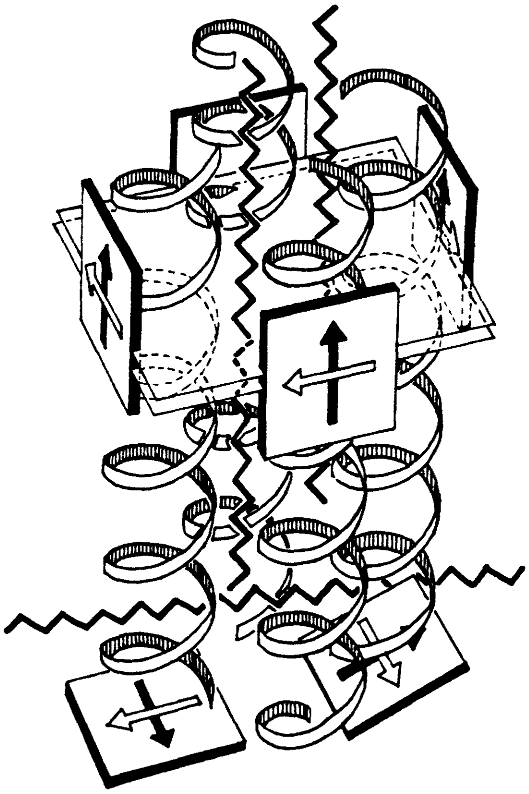
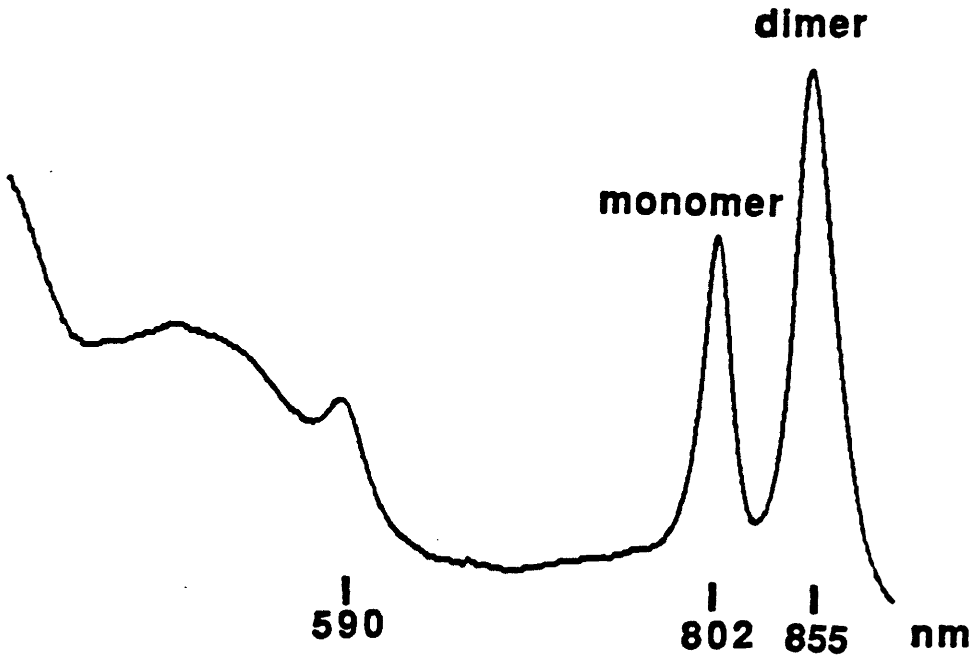


Figure 3. NIR spectrum of *Rb. capsulatus* LHII. The 855 nm band is attributed to the dimers of Bchl while the 800 nm band is assigned to the monomer Bchl.



determined to be perpendicular to the membrane. The Q_x transition of Bchl 800 was found to be approximately perpendicular to the membrane, while the Q_x transition of Bchl 855 is nearly parallel to the membrane. This LH data indicates that the porphyrin rings of the dimer Bchls are approximately perpendicular to the membrane, while the porphyrin rings of the monomer Bchls are nearly parallel to the membrane.

The distance between the Bchl monomer and dimer was calculated to be 2.1 nm based on the energy transfer efficiency (Kramer *et al.*, 1984). This distance is consistent with models determined from the primary structure. Homologous His residues in the alpha helical stretch of the α and β polypeptides are the likely candidates as the Mg ligands for the dimer Bchls. These His residues are conserved in the antennae of purple bacteria (Zuber & Brunisholz, 1991). Additionally, resonance Raman spectroscopy supports the role of these His residues as the binding sites for the Bchls of the dimer (Robert & Lutz, 1984). The most probable binding site for the monomer Bchl is in the β subunit at the transition of the alpha helical section and the polar amino terminal side, where a second His (conserved across the LHII antennae of purple bacteria) is found (Zuber & Brunisholz, 1991).

1.1.2.1 Crystallization of LHII

The LHII complex from *Rb. capsulatus* is stable in the presence of high detergent concentrations, and can be purified by a single DEAE anion exchange column. Crystals have been obtained from *Rb. capsulatus* LHII as early as 1985 (Welte *et al.*, 1985) as well as from a number of other organisms (Cogdell & Hawthornthwaite, 1993 and references therein). However there is a problem in reproducibly obtaining crystals that display significant X-ray diffraction. Cogdell & Hawthornthwaite (1993) summarize the current state of LH crystallization trials from a number of species and present a general protocol for the crystallization of antennae complexes from purple bacteria. Currently, Donnelly & Cogdell (1993) report crystals of LHII from *Rhodopseudomonas acidophila* that diffract beyond 3.5 Å. Cogdell states that work is in progress analyzing heavy atom derivatives to solve the phases. Work by Christine Goddard in the Youvan laboratory (in preparation) shows that a wide variety of crystallization conditions can be used to obtain crystals of *Rb. capsulatus* LHII.

Crystals were obtained that diffracted down to 4.5 Å, however reproducibly producing these crystals was a problem (Goddard & Youvan, in preparation).

The problems with producing crystals that diffract X-rays are probably due to heterogeneity in the protein preparation. One source of the heterogeneity that disorders crystals could be variation of lipid content in the protein preparation (Cogdell & Hawthornthwaite, 1993). However, when antenna complexes were washed on a column to remove excess lipids, some crystals diffracted well, but reproducibility was still a problem. Other sources of heterogeneity include aggregation of LHII, and the slow denaturation of the protein during the crystallization trials.

1.1.3 LHII spectra

LHII has an intense, characteristic, NIR absorption spectrum. The dimer and monomer Bchls can be specifically detected, with the dimer absorbing at 855 nm and the monomer at 800 nm. The dimer and monomer bands have a peak ratio of 1.5 to 1, respectively. These absorption bands are red-shifted with respect to monomeric Bchl in organic solvents. The spectral properties of the protein bound pigments make them excellent reporters for protein assembly, structure, and function. Structural changes could lead to spectral shifts, or a change in the ratio of the absorption bands. Changes in the protein that result in a decrease of LHII function can be monitored by fluorescence, as LHII has a high fluorescence when it can not efficiently transfer energy to the RC (Youvan *et al.*, 1983).

The cause of the red-shifts of pigment-bound Bchls may be due to pigment-protein, or pigment-pigment interactions. Aromatic amino acids in the vicinity of the Bchl have been postulated to be determining factors for the spectral properties of the pigment-proteins (Brunisholz & Zuber, 1988). In antenna complexes, aromatic amino acid residues are located within the hydrophobic stretches of the transmembrane region, as well as in the carboxy-terminal domain. Charged amino acids near the Bchl binding site have also been hypothesized to produce the spectroscopic red-shift (Eccles & Honig, 1983). A second model suggests that the protein provides the frame work for the pigments, but that pigment-pigment interactions control the spectral features of the Bchl. The red-shift is attributed to exciton interactions between the electronic transitions of neighboring pigments (Scherz & Parson, 1984;

Pearlstein, 1991). The red-shifts of these bands are probably a function of both pigment-protein interactions, and excitonic coupling between pigments.

LH complexes exist as aggregates of subunit-pigment units. The minimal size of isolated LH complex required for retention of its *in vivo* absorbance is not known. Work has been done to reconstitute LH complexes from their separately isolated component parts (Parkes-Loach *et al.*, 1988; Chang *et al.*, 1990). Core LH antennae from *Rhodospirillum rubrum* have been successfully reconstituted to a detergent-isolated form absorbing about 50 nm to the blue of the *in vivo* antennae. This blue-shifted form can be reassociated to regain a WT-like absorbance spectra. Chang and coworkers report similar blue-shifted forms have been found for other core antennae from other species, including *Rb. capsulatus*. Dimers of Bchl a absorbing at 853 nm were prepared in a solution of formamide and water what contained micelles of triton X-100 (Scherz & Rosenbach-Belkin, 1989). Scherz & Rosenbach-Belkin suggest that in LHII, the wavelength shift is explained by dimers of Bchl whose excited states are strongly coupled, but other spectral properties of the Bchl 855 (i.e., CD spectra) are influenced by weak exciton coupling of the red-shifted Q_Y transitions.

1.1.4 LHII Genetics

1.1.4.1 Isolating the genes for the photosynthetic apparatus

Spontaneous mutants were used to isolate and/or confirm the identity of the genes for the photosynthetic apparatus. An R-factor with enhanced chromosomal mobilizing ability in *Rb. capsulatus* was isolated and used to generate R-prime derivatives that carry a photosynthetic gene cluster (Marrs, 1981). The R-prime plasmids carrying parts of the chromosome specifying the photosynthetic apparatus were confirmed by their ability to complement photosynthetic growth *in trans* when mated with photosynthetically defective mutants. Mutants incapable of photosynthetic growth were isolated by subjecting them to a tetracycline suicide procedure. Mutants isolated from the tetracycline suicide were screened for enhanced fluorescence in the NIR; enhanced fluorescence mutants are typically defective in LHI or RC but have functional LHII (Youvan *et al.*, 1983). An R-prime plasmid, pRPS404, complemented all the fluorescent mutants. The location of the LHI and the RC genes were further narrowed down by complementation with pBR322 derivatives containing smaller fragments of pRPS404.

The LHII structural genes are expressed from the *puc* operon and are not within the photosynthetic gene cluster. They were isolated first in *Rb. capsulatus* using synthetic probes based on the amino-terminal sequences of α and β LHII which were used in Southern hybridizations (Youvan & Ismail, 1985). A 6 kb Eco RI fragment that hybridized to the probes was subcloned into pBR322 (to form pRPSLH2). This construction was able to complement, *in trans*, the strain MW422, which has a chromosomal mutation that inactivates LHII.

Recently a high resolution physical and genetic map of the the *Rb. capsulatus* (strain SB1003) genome was reported (Fonstein & Haselkorn, 1993). The *puc* operon (LHII genes) was mapped on the opposite side of the chromosome from the *puf* operon which contains the RC (L and M subunits) and LHI genes. To facilitate high resolution gene mapping of *Rb. capsulatus*, blots were made of the minimal set of cosmids covering the chromosome (192) to be used along with the plasmid with the known map position of each cosmid.

1.1.4.2 LH phylogeny

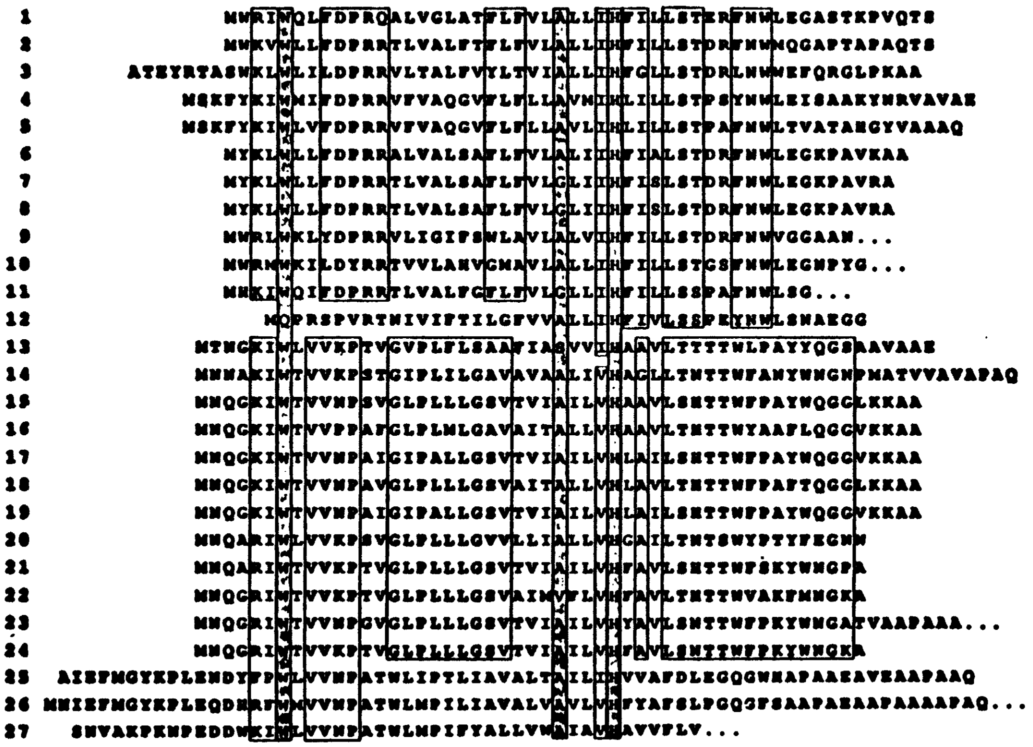
The primary structure of core and peripheral antennae from a number of purple bacteria have been determined (Zuber, 1990; Brunisholz & Zuber, 1991; Table 1). Because the three dimensional structure has not been solved, examination of the amino acid sequences of the homologous LH antennae can give insights into important regions of the protein. What is currently known about LH structure is based largely on biochemical characterization and spectroscopy in conjunction with analysis of the sequence database. Recently Donnelly & Cogdell (1993) used the database in predictions of the point at which the transmembrane helix leaves the bilayer. The experiments in this thesis incorporated phylogenetic information in combinatorial cassette design.

1.1.4.3 LHII deletion backgrounds and expression plasmids

In the mid 1980's a series of genomic deletion backgrounds and expression plasmids were constructed (Youvan *et al.*, 1985; Bylina *et al.*, 1986) for both the *puc* (LHII) and *puf* (RC and LHI) operons. Both the RC and LHI were the focus of many site directed mutagenesis (SDM) experiments, but prior

Table 1. (From Zuber, 1990) Primary structure of the α (panel A) and β (panel B) subunits of core and peripheral antennae complexes of purple bacteria.

A.


1 MWRIWQLFDPRQALVGLATFLFVLALLIHFILLSTERFNWLEGASTKPVQTS
2 MWKVWLLFDPRRTLVALFTFLFVLALLIHFILLSTDRFNWMQGAPTAPAQTS
3 ATEYRTASWKLWLILDPRRVLTALFVYLTVIALLIHFGLLSTDRLNWWEFQRGLPKAA
4 MSKFYKIWMIFDPRRVFVAQGVFLFLLAVMIHLILLSTPSYNWLEISAAKYNRVAVAE
5 MSKFYKIWLVFDPRRVFVAQGVFLFLLAVLIHLILLSTPAFNWLTVATAEGYVAAAQ
6 MYKLWLLFDPRRALVALSAFLFVLALIIHFIALSTDRFNWLEGKPAVKAA
7 MYKLWLLFDPRRTLVALSAFLFVLGLIIHFISLSTDRFNWLEGKPAVRA
8 MYKLWLLFDPRRTLVALSAFLFVLGLIIHFISLSTDRFNWLEGKPAVRA
9 MWRLWKLYDPRRVLIGIFSWLAVLALVIHFILLSTDRFNWVGGAAN...
10 MWRMWKILDYRRTVVLAHVGMAVLALLIHFILLSTGSFNWLEGNPYG...
11 MNKIWQIFDPRRTLVALFGFLFVLGLLIHFILLSSPAFNWLSG...
12 MQPRSPVRTNIVIFTILGFVVALLIHFIVLSSPEYNWLSNAEGG
13 MTNGKIWLVVKPTVGVPLFLSAAFIASVVIHAAVLTTTTWLPAYYQGSAAVAAE
14 MNNAKIWTVVKPSTGIPLILGAVAVAALIVHAGLLTNTTWFANYWNGNPMATVVAVAPAQ
15 MNQGKIWTVVNPSVGLPLLLGSVTVIAILVHAAVLSHTTWFPAYWQGGLKKAA
16 MNQGKIWTVVPPAFGLPLMLGAVAITALLVHAAVLTHTTWYAAFLQGGVKKAA
17 MNQGKIWTVVNPAIGIPALLGSVTVIAILVHLAILSHTTWFPAYWQGGVKKAA
18 MNQGKIWTVVNPAVGLPLLLGSVAITALLVHLAVLTHTTWFPAFTQGGLKKAA
19 MNQGKIWTVVNPAIGIPALLGSVTVIAILVHLAILSHTTWFPAYWQGGVKKAA
20 MNQARIWLVVKPSVGLPLLLGVVLLIALLVHGAILTNTSWYPTYFEGNW
21 MNQARIWTVVKPTVGLPLLLGSVTVIAILVHFAVLSHTTWFSKYWNGPA
22 MNQGRIWTVVKPTVGLPLLLGSVAIMVFLVHFAVLTHTTWVAKFMNGKA
23 MNQGRIWTVVNPGVGLPLLLGSVTVIAILVHYAVLSNTTWFPKYWNGATVAAPAAA...
24 MNQGRIWTVVKPTVGLPLLLGSVTVIAILVHFAVLSNTTWFPKYWNGKA
25 AIEFMGYKPLENDYFPWLVVNPATWLIPTLIAVALTAILIHVVAFDLEGQGWHAPAAEAVEAAPAAQ
26 MNIEFMGYKPLEQDHRFWMVVNPATWLMPILIAVALVAVLVHFYAFSLPGQSFSAAPAEAAPAAAAPAQ...
27 SNVAKPKNPEDDWKIWLVVNPATWLMPIFYALLVWAIAVHAVVFLV...

1 Rhodospirillum rubrum B870-α
2 Rhodopseudomonas marina B880-α
3 Rhodopseudomonas viridis 1015-α
4 Rhodobacter sphaeroides B870-α
5 Rhodobacter capsulatus B870-α
6 Rp. acidophila Ac7050 B880-α
7 Rp. acidophila Ac7750 B880-α
8 Rp. acidophila Ac10050 B880-α
9 Ectothiorhodospira halophila B890$_1$-α
10 Ectothiorhodospira halophila B890$_2$-α
11 Chromatium vinosum B890-α
12 Chloroflexus aurantiacus J-10-fl B808-866-α
13 Rhodobacter sphaeroides B800-850-α
14 Rhodobacter capsulatus B800-850-α
15 Rp. acidophila Ac7050 B800-850-α
16 Rp. acidophila Ac7050 B800-820-α
17 Rp. acidophila Ac7750 B800-850-α
18 Rp. acidophila Ac7750 B800-820-α
19 Rp. acidophila Ac10050 B800-850-α
20 Ectothiorhodospira halophila B800-850-α
21 Rp. palustris 2.6.1 B800-850-α$_1$
22 Rp. palustris 2.6.1 B800-850-α$_2$
23 Rp. palustris 2.6.1 B800-850-α$_3$
24 Rp. palustris 2.6.1 B800-850-α$_4$
25 Chromatium vinosum B800-850-α$_1$
26 Chromatium vinosum B800-820-α
27 Chromatium vinosum B800-850-α$_2$

B.

```
 1      EVKQESLSGITEGEAKEFHKIFTSSILVFFGVAAFAHLLVWIWRPWVPGPNGYS
 2    AEIDRPVSLSGLTEGEAREFHGVFMTSFMVFIAVAIVAHILAWMWRPWIPGPEGYA
 3      ADLKPSLTGLTEEEAKEFHGIFVTSTVLYLATAVIVHYLVWTARPWIAPIPKGWV
 4     ADKSDLGYTGLTDEQAQELHSVYMSGLWPFSAVAIVAHLAVYIWRPWF
 5    MADKNDLSFTGLTDEQAQELHAVYMSGLSAFIAVAVLAHLAVMIWRPWF
 6      AEDRSSLSGVSDAEAKEFHALFVSSFMGFMVVAVLAHVLAWAWRPWIPGPKGWA
 7      AEDRSSLSGVSDAEAKEFHALFVSSFTAFIVIAVLAHVLAWAWRPWIPGPKGWA
 8      AEDRSSLSGVSDAEAKEFHALFVSSFTAFIVIAVLAHVLAWAWRPWIPGPKGWA
 9         TDIRTGLTDEECQEIHEMNMLGMHAYWSIGLIANALAYAWRPFEQGRAGNRLEDEAPDYVRSALT
10       ADNMSLTGLSDEEAKEFHSIFMQSFLIFTAVAVVAHFLAWAWRPWIPGAEGYG...
11         ADEMRNVSDEEAKEFHAMFSQAFTVYVGVAVVAHILAWAWRPWIPGDEGFG...
12        ANSSMTGLTEQEAQEAHGIFVQSMTAFFGIVVIAHILAWLWRPWL...
13        DQKSMTGLTEEEAKEFHGIFTQSMTMFFGIVIIAHILAWLWRPWL...
14 MRDDDDLVPPKWRPLFNNQDWNLHDIVVKSFYGFGVIAAIAHLLVYLWKPWLP
15    TDDLNKVWPSGLTVAEAEEVHKQLILGTRVFGGMALIAHFLAAAATPWLG
16     MTDDKAGPSGLSLKEAEEIHSYLIDGTRVFGAMALVAHILSAIATPWLG
17      ADDVKGLTGLTAAESEELHKHVIDGTRVFFVIAIFAHVLAFAFSPWLH
18            AEVLTSEQAEELHKHVIDGTRVFLVIAAIAHFLAFTLTPWLH
19             ATLTAEQSEELHKYVIDGTRVFLGLALVAHFLAFSATPWLH
20             AVLSPEQSEELHKYVIDGARAFLGIALVAHFLAFSMTPWLH
21           ADKPLTADQAEELHKYVIDGARAFVAIAAFAHVLAYSLTPWLH
22             ATLTAEQSEELHKYVIDGTRVFLGLALVAHFLAFSATPWLH
23         DKTLTGLTVEESEELHKHVIDGTRIFGAIAIVAHFLAYVYSPWLH
24    ADDPNKVWPTGLTIAESEELHKHVIDGTRIFGAIAIVAHFLAYVYSPWLH
25   MVDDPNKVWPTGLTIAESEELHKHVIDGSRIFVAIAIVAHFLAYVYSPWLH
26       ADMKSLSGLTEQQAKEFHEQFKVTYTAFVGLAALAHLFVIAANPW...
27          AELSGLTDQQAKEFHEQFKVTYTAFVGLAALAHLFVIAANPW...
28        ASLLLSGLTEQQAKEFHEQFKVTYTAFVGLAALAHLFVIAANPW...
29            MNGLTEQQAKEFHAQFKVTYTAFVGLAALAHLMVLANNPWF...
```

1 Rhodospirillum rubrum B890-β
2 Rhodopseudomonas marina B880-β
3 Rhodopseudomonas viridis B1015-β
4 Rhodobacter sphaeroides B870-β
5 Rhodobacter capsulatus B870-β
6 Rp. acidophila Ac7050 B890-β
7 Rp. acidophila Ac7750 B890-β
8 Rp. acidophila Ac10050 B890-β
9 Ectothiorhodospira halochloris β-Polypeptid
10 Ectothiorhodospira halophila $B890_1$-β
11 Ectothiorhodospira halophila $B890_2$-β
12 Chromatium vinosum $B890_1$-β
13 Chromatium vinosum $B890_2$-β
14 Chloroflexus aurantiacus J-10-fl B806-866-β
15 Rhodobacter sphaeroides B800-850-β
16 Rhodobacter capsulata B800-850-β
17 Rp. acidophila Ac7050 B800-850-β
18 Rp. acidophila Ac7050 B800-820-β
19 Rp. acidophila Ac7750 B800-850-β
20 Rp. acidophila Ac7750 B800-820-$β_2$
21 Rp. acidophila Ac7750 B800-820-$β_1$
22 Rp. acidophila Ac10050 B800-850-β
23 Rp. palustris 2.6.1 B800-850-$β_1$
24 Rp. palustris 2.6.1 B800-850-$β_2$
25 Rp. palustris 2.6.1 B800-850-$β_3$
26 Chromatium vinosum B800-850-β
27 Chromatium vinosum B800-820-$β_1$
28 Chromatium vinosum B800-820-$β_2$
29 Chromatium vinosum B800-820-$β_3$

to this thesis, mutagenesis of LHII was not examined in *Rb. capsulatus*. For this thesis work, a set of LHII expression plasmids was designed that allow for easier manipulation of the LHII genes than the original LHII expression plasmid.

1.2 A review of the role of mutagenesis in photosynthesis research

Historically many biophysically important mutants have been isolated through advances in screening and mutagenesis. This section gives a review of some of the mutants constructed through SDM and structural motif rearrangements.

For the purposes of discussing the role of mutagenesis in photosynthetic research, it is interesting to look at the work done on the RC and LHI as well as that on LHII. The photosynthetic reaction center contains a Bchl dimer, two Bchl monomers, two bacteriopheophytin (Bphe) monomers, two quinones, and a ferrous non-heme iron atom attached to two quasi-symmetrically arranged protein subunits (L and M). The structure of the purple bacteria *Rhodopseudomonas viridis* RC has been determined through X-ray crystallography (Deisenhofer *et al.*, 1985). LHI is modeled to contain two transmembrane alpha helical polypeptides (α and β) which bind a Bchl dimer (Zuber, 1986). The bound pigments (of LHII, LHI and the RC) can be exploited as reporter groups; NIR and visible spectral information can be used to assay for protein assembly, structure, and function.

1.2.1 Site directed mutagenesis

Introducing specific alterations in the protein sequence has allowed biophysicists to study some of the fundamental mechanisms of the light reactions of photosynthesis. Molecular biological systems have been developed for the manipulation of both LH and RC proteins. Both LHI and LHII antennae have been altered by SDM in attempts to correlate amino acid sequence with protein structure and function. Sequence modifications have been made in the vicinity of all the prosthetic groups in the RC, and at a variety of non-symmetric amino acid residues in the RC (see Coleman & Youvan, 1990 and references therein). Many of these mutations lead to changes in spectral properties or differences in the rate of electron transfer. A few examples of the

SDM experiments that have been performed on both LH antennae and RCs are summarized below.

To study Bchl-peptide interactions in LH antennae, mutations were made (Bylina *et al.*, 1988b) in the putative Bchl binding sequence, Ala-X-X-X-His (Theiler & Zuber, 1984; Theiler *et al.*, 1984), of the α subunit in LHI from *Rb. capsulatus*. No LHI assembly, as judged by absorption spectroscopy, was detected when this His is changed to any other amino acid residue, suggesting that in LH, no other residue can function in place of His as a ligand to the Mg of the Bchl. Mutagenesis of the Ala residue showed protein assembly occurred only when it is changed to an amino acid with a small molar volume (Gly, Ser, Cys) indicating that there are molar volume constraints for Bchl binding at this position.

In an investigation of residues at the N-terminus of the α and β subunits of LHI, charged amino acids were exchanged with those of opposite charge (Stiehle *et al.*, 1990). It was believed that the positively charged α subunit N-terminal segment helped influence the protein's orientation with respect to the membrane, while the negative charges on the β subunit N-terminal were thought to help stabilize the complex by interacting with the α subunit. When four positively charged amino acid residues in the α subunit were exchanged with negatively charged amino acids no formation of the LH complex was observed. However, when four negative amino acids in the β subunit were changed to positive amino acids, LHI assembly was impaired but not completely blocked. Stiehle and coworkers claim that charged amino acids in α and β influence LHI formation in different ways, with β being much more tolerant of the mutations than the α subunit.

Extensive point mutations in LHI at amino acid positions which are highly conserved across species were performed to obtain information about their structural and functional role (Babst *et al.*, 1991). Eighteen mutants with single amino acid substitutions were constructed. All the substitutions resulted in structural effects judged from characterization based on quantification of core complexes and the LHI spectral characteristics. For example, exchanges at α8 ($Trp^{\alpha 8}$→Leu, Ala, Tyr) resulted in the absence of core complex, the absence of core antennae and a reduction in the carotenoid content of the complex, respectively. Substitutions at α43 ($Trp^{\alpha 43}$→ Ala, Leu, Tyr) resulted in 8-11 nm blue shifts of the absorption peak.

Genomic deletion backgrounds and expression plasmids for LHII from a related strain (*Rhodobacter sphaeroides*) were constructed (Burgess *et. al.*, 1989), and several amino acid positions modified. SDM on aromatic (Fowler *et al.*, 1992) and charged (Fowler *et al.*, 1993) residues of LHII from *Rb. sphaeroides* resulted in blue-shifts of the dimer band. A correlation was found between two Tyr residues in the α subunit of LHII and the dimer absorption band. By constructing the single (Tyr$^{\alpha 44}$→Phe) and double (Tyr$^{\alpha 44}$→Phe and Tyr$^{\alpha 45}$→Leu) mutants, the band shifted 11 and 24 nm, respectively. This effect is interpreted as being due to direct interactions of the Bchls with the aromatic amino acids. Changing Lys$^{\beta 23}$→Gln resulted in an 18 nm blue shift of the dimer band. This amino acid is near the putative Bchl monomer binding site and is conserved in a number of LHII complexes. The monomer band was not affected in this mutant, and the interpretation given to the spectral shift of the dimer peak is that the relative orientation of the polypeptide backbone in the membrane might have been altered.

The RC heterodimer mutants are examples of SDM experiments which resulted in significant advances in our understanding of the molecular mechanisms of photosynthesis. Replacement of either His ligand to the Bchls in the special pair (His^{L173}, His^{M200}) with an aliphatic amino acid residue (Leu) leads to a Bchl/Bphe heterodimer (Bylina & Youvan, 1988a, 1990). The His^{M200}→Leu heterodimer mutant has been extensively characterized and gave insights into the very early steps involved in charge separation. The overall quantum yield is 50% (as opposed to 100% for wild type (WT)); the decrease in quantum yield is attributed to unproductive decay of the excited dimer to the ground state (Bylina *et al.*, 1989). Picosecond transient absorption spectroscopy (Kirmaier *et al.*,1988, 1989), low temperature ground state and linear dichroism absorption spectra (Breton *et al.*,1989), electron paramagnetic resonance (Bylina *et al.*, 1990), and Stark spectroscopy (DiMango *et al.*, 1990) suggest the existence of an intradimer charge transfer state that mixes with the excited singlet state. A charge transfer state has been postulated to exist in WT, but at a higher energy, making it more difficult to observe. The observation of the charge transfer state in the heterodimer mutant suggests that these states probably facilitate efficient charge separation in the RC (Youvan, 1991).

In a similar type of pigment switching experiment, the Bphe acceptor was replaced with a Bchl by substituting a Leu with a His which can act as a metal coordinating ligand for the Mg in Bchl (Kirmaier *et al.*, 1991). The resulting RCs from this Leu^{M214}→His mutant undergoes charge separation from the special pair through the Bchl intermediate acceptor to the quinone, with a reduced quantum yield of 60%. The electron transfer from the acceptor to the quinone is responsible for this effect because charge recombination is much faster in the mutant and competes with electron transfer to the quinone. The heterodimer and Leu^{M214}→His mutants suggest that the high quantum yield of the WT RC is accomplished by lack of significant electronic coupling to the ground state, therefore limiting participation of states where the charge is separated between strongly interacting chromophores (Kirmaier *et al.*, 1991).

Although the RC contains two symmetrical branches of pigments, only one branch is active in electron transfer. A strategy to determine which residues are responsible for the unidirectionality of electron transfer has been to symmetrize the RC by changing amino acid residues that differ between the quasi symmetrical L and M subunits. The first example of this type of experiment was performed in the Youvan laboratory by Ed Bylina who focused on Glu^{L104}, because this protonated amino acid is thought to hydrogen bond to the ring V C-9 keto group of the primary acceptor, which would be expected to lower the energy of the anion, thus facilitating unidirectionality (Bylina *et al.*, 1988c). There is no analogous residue on the M side. The Glu^{L104}→Leu mutant resulted in only minor changes in the kinetics of electron transfer. The initial electron transfer step is less than two-fold slower in the mutant and the second step is only marginally slower. However, Glu^{L104} is responsible for the spectroscopic red-shift of the active branch Bphe relative to the inactive branch Bphe.

1.2.2 Structural motif rearrangements

Large scale rearrangement of LH and RC sequences has been necessary to engineer phenotypes that could not be found by spontaneous or site directed mutagenesis. Global changes have been made in both the LH antennae and RC.

Structural motif mutagenesis has been used by Adam Arkin, in our laboratory, to investigate pigment binding and the criteria for proper helix formation in the LHI antenna. Five amino acid residues in the α subunit of LHI (flanking the His thought to be the ligand to a Bchl of the dimer) were changed to Leu (Arkin & Youvan, 1993). This resulted in a stretch of fourteen residues all of which are Leu except the Ala and the His in the Ala-X-X-X-His Bchl binding motif. This 'poly Leu' mutant rapidly mutates to loose either carotenoid or LH expression. The α subunit of LHI seems to have only a limited tolerance to change in the vicinity of the Bchl binding site.

Structurally important segments of the RC have been duplicated and exchanged between the L and M subunits. The D helices are good candidates for this type of manipulation as they have interactions with all the prosthetic groups. Steven Robles, in our group, constructed mutants with two M subunit D helices (D_{MM}), two L subunit D helices (D_{LL}), and with the helices exchanged (D_{LM}) in the hope of identifying the asymmetric region responsible for the unidirectionality of electron transfer (Robles *et al.*, 1990a, 1990b). The D_{MM} mutant resulted in a photosynthetically functional, but severely impaired RC. The D_{LL} mutant is photosynthetically inactive, and was found to be missing the primary electron acceptor. The D_{LL} RC has a total side chain molar volume greater than WT, which probably interferes with the Bphe binding pocket. Photosynthetically competent revertants of both D helix duplications have been isolated. All the D_{LL} revertants decrease the molar volume near the Bphe binding site, and all bind the Bphe. The positions at which reversions or secondary site compensatory mutations take place under photosynthetic selection point to what may be critical amino acid asymmetries. The extended lifetime of the excited dimer in the D_{LL} mutant allowed the observation of oscillations in the decay of this state which suggests that vibrational coherence may modulate electron transfer (Vos *et al.*, 1991).

A second series of helix exchange and duplication experiments involved the amphipathic cd helixes which provide the axial ligands to the accessory Bchls and form part of the special pair binding pocket (Robles *et al.*, 1992). None of the constructions resulted in photosynthetically competent mutants; however, functional revertants were obtained from both helix duplications. Compensatory mutations were found in both the mutagenized and non-mutagenized subunits, and several pairs of symmetrical suppressors (compensatory mutations that occur at homologous positions in either subunit) were isolated. A speculative possibility is that the symmetrization of the RC caused the two chromophore branches to have the same potential for electron transfer, and that either branch could be activated by a secondary mutation. Further experiments will be needed to search for aberrant electron transfer in these mutants.

1.3 Massively parallel screening of mutants

Screening technologies and mutagenesis techniques have undergone a dramatic increase in sophistication with the development of digital imaging spectroscopy and combinatorial cassette mutagenesis. The newest techniques promise to yield more interesting and informative phenotypes than site directed mutagenesis or structural motif rearrangements. Optimization of combinatorial cassettes, and digital imaging spectrophotometry are described in this section.

1.3.1 Searching protein sequence space

Combinatorial mutagenesis has a greater potential to generate novel phenotypes that can not be found using either SDM or structural motif mutagenesis, because of the inability to accurately predict the structure and function of proteins from their primary amino acid sequence. Combinatorial cassette mutagenesis (CCM) (Oliphant *et al.*, 1986; Reidhar-Olson *et al.*, 1991) allows the exploration of a large number of mutations in a protein. Multiple amino acid residues are simultaneously mutagenized in a random fashion resulting in a library of proteins, some of which may exhibit desired phenotypes. The complexity of these libraries grows exponentially with the number of sites mutagenized.

It is advantageous to reduce the total number of proteins in a combinatorial mutagenesis library to increase the efficiency of searching 'sequence space' for mutants with the desired phenotypes. Instead of using the codon NNN (i.e., N=25% each A,T,G,C) at each mutagenized position, the nucleotide composition can be restricted. A simple example is NN(G/C), which is an alternate way to randomize a sequence position using all 20 amino acids but only half as many codons. In more sophisticated doping schemes, criteria such as physicochemical parameters, expert rules, structural, and phylogenetic data can be used to construct 'target sets' of amino acids to be used at each mutagenized sequence position (Arkin & Youvan, 1992b). One of several algorithms can be used to convert these target sets into nucleotide mixtures amenable to DNA synthesis (Youvan *et al.*, 1992).

Two equations for adjusting nucleotide dopes have been used experimentally:

1) Group Probability P_G:

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

where $P_D[i]$ is the probability of the ith amino acid in a target set occurring based on a specific doping scheme, and

2) Sum of the Squares of the Differences (SSD):

$$SSD = \sum_i (P_D[i] - P_T[i])^2 \qquad \text{Eq. 2}$$

where $P_T[i]$ is the fractional representation of the ith amino acid in the target set. The best dope is found by either maximizing P_G or minimizing SSD. The P_G function forces every amino acid in the target set, regardless of the frequency at which they occur, to be encoded by the cassette. In contrast, the SSD algorithm takes into consideration how many times each amino acid is found in the target set and may drop infrequently used amino acids from the dope. Computer simulations show SSD generates a higher throughput of positive mutants than P_G (Youvan *et al.*, 1992), but this is potentially at the expense of phenotypic diversity.

1.3.2 Digital imaging spectroscopy

Too many mutants are generated by combinatorial mutagenesis to screen individual isolates by conventional spectroscopy. Digital imaging spectroscopy (DIS) provides a method for the parallel screening of the ground state (visible and NIR) spectra from up to five hundred colonies directly on a petri dish (Yang & Youvan, 1988; Arkin *et al.*, 1990; Arkin & Youvan, 1993; Youvan *et al.*, 1993). Spectra acquired from DIS are often superior to spectra recorded from purified chromatophore membranes taken with a conventional spectrophotometer because DIS is less sensitive to light scattering from these turbid samples.

1.3.2.1 Instrumentation

All current configurations of DIS use a charged coupled device (CCD) camera as a detector to image petri dishes mounted on the exit port of an integrating sphere. The sphere provides uniform light across the petri dish. In a second generation imager, up to 25 petri dishes can be illuminated simultaneously. The light source uses Fabry-Perot interference filters or an 1/8 meter monochromator to illuminate the target at different wavelengths. Typically, spectra can be obtained at 5-10 nm resolution which can detect 2 nm band shifts. Images are transferred to a Silicon Graphics Personal Iris computer where they are stored, manipulated, and compiled into absorption spectra.

Fluorescence images and images taken at a single wavelength are useful for rapid pre-screening of colonies. Fluorescence images can be obtained by illuminating with broad band blue-green light, and placing an 830 nm long pass filter in front of the CCD camera. This is analogous to a photographic technique developed 10 years ago (Youvan *et al.* 1983). Screening at a single wavelength is a fast method to select mutants with strong absorption characteristics at an important wavelength. For example, a library of light harvesting mutants can be scanned by absorbance at 860 nm before acquiring the full DIS spectra.

1.3.2.2 Spectral display

Spectra acquired by DIS can be displayed in either 'tile mode' or as 'color contour maps'. In tile mode, each spectrum is displayed as a conventional two dimensional absorption spectrum. However, after a few hundred colonies have been imaged, this type of display is not suitable for rapid inspection of the data. In the color contour mode, all the spectral data from a single petri dish are presented as a two dimensional display. The vertical axis corresponds to different colonies and the spectra of each colony is represented by a horizontal row. Absorption is color-coded at each wavelength along the row (white= high absorption, black= zero absorption). Spectra can be sorted according to similarity or absorption at various wavelengths. This type of display makes it easier to identify and compare spectral classes. In either mode, the spectra can be scaled relative to the lowest and highest absorption anywhere in the image, or each spectra can be scaled between its own minimum and maximum absorption (Arkin & Youvan, 1993; Youvan *et al.*, 1993).

Images of petri dishes screened by absorption at a single wavelength or by fluorescence can be displayed as either monochrome or pseudocolored images. Single wavelength images are divided by a blank image at the same wavelength to correct for uneven light intensity from the monochrome illumination. Grayscale values from the ratioed single wavelength absorption (or fluorescence) image are rescaled to enhance contrast. After establishing the low and high grayscale values, such monochrome images can be linearly mapped by pseudocoloring schemes to enhance differences between mutants.

1.4 LHII as a model system

The experiments performed for this thesis were done in a LHII chromosomal deletion strain in which RC and LHI expression are prevented by point mutation on the chromosome. A system of plasmids for LHII expression was designed that facilitates cassette mutagenesis of LHII. This combination of deletion background and plasmid allows the examination of only LHII protein expressed from the plasmid.

Light harvesting II antennae from *Rb. capsulatus* provides a model system for implementing complex mutagenesis schemes. The prosthetic groups serve as colorimetric indicators of protein expression and subunit assembly. *Rb. capsulatus* synthesizes large quantities of LHII, and since it is a stable protein and has a very intense characteristic absorption spectrum, the effects of engineered mutations can be easily assayed. For the work detailed in this thesis, a system of plasmids for LHII cassette mutagenesis was designed (See Appendix B). DIS can be employed to rapidly screen thousands of colonies from libraries resulting from CCM experiments.

Chapter 2: Target set mutagenesis

2.1 Summary

Combinatorial cassettes based on a phylogenetic "target set" were used to simultaneously mutagenize seven amino acid residues on one face of a transmembrane alpha helix comprising a bacteriochlorophyll binding site in the light harvesting II antenna of *Rhodobacter capsulatus*. This pigmented protein provides a model system for developing complex mutagenesis schemes, because simple absorption spectroscopy can be used to assay protein expression, structure, and function. Colony screening by digital imaging spectroscopy showed that 6% of the optimized library bound bacteriochlorophyll in two distinct spectroscopic classes. This is approximately 200 times the throughput (ca. 0.03%) of conventional combinatorial cassette mutagenesis using $[NN(G/C)]_7$. "Doping" algorithms evaluated in this model system are generally applicable and should enable engineers to simultaneously mutagenize more positions in a protein than currently possible, or alternatively, to decrease the screening size of combinatorial libraries.

2.2 Introduction

In order to increase one's chances of finding mutants with desired properties in "sequence space", it is advantageous to formulate nucleotide mixtures that restrict the sequence complexity of combinatorial cassettes. This is especially true when the experimenter's screening size is far less than the theoretical complexity of the combinatorial library. According to this scheme, various physicochemical parameters, expert rules, structural or phylogenetic data can be used to limit the actual "target set" of amino acids used at each sequence position in the protein. This should be contrasted with conventional combinatorial cassette mutagenesis (CCM) (Oliphant *et al.*, 1986; Reidhaar-Olson *et al.*, 1991; Robles & Youvan, 1993) which uses all 20 amino acids at each position. Recently, an algorithm has been described which converts target sets of amino acids into nucleotide mixtures amenable to DNA synthesis (Arkin & Youvan 1992; Youvan *et al.*, 1992). This chapter presents the first experimental implementation of target set mutagenesis (TSM).

Phylogenetic data (Zuber, 1990) from 29 homologous light harvesting proteins were used as a database to construct target sets for seven sequence positions. Light harvesting antennae serve to funnel energy to the photosynthetic reaction center where charge separation takes place. The light harvesting II (LHII) protein is modeled to include two transmembrane alpha-helical polypeptides (α, β) that bind a bacteriochlorophyll (Bchl) dimer and a Bchl monomer (Zuber, 1986). These Bchls can be specifically differentiated by their ground state near infrared (NIR) absorption bands at 860 nm (dimer) and 800 nm (monomer). Free Bchl in membranes absorbs at 760 nm, hence red-shifted Bchl absorption bands serve as reporters for LHII expression, structure, and function. Since digital imaging spectroscopy (DIS) can be used to acquire hundreds of colony spectra from a single petri dish (Yang & Youvan, 1988; Arkin *et al.*, 1990; Arkin & Youvan, 1993, Youvan *et al.*, 1993), these structurally sensitive colored reporters make LHII a model protein to test new mutagenesis techniques.

2.3 RESULTS

2.3.1 Calculations for a phylogenetic TSM cassette

Seven amino acid residues were simultaneously mutagenized (Figure 4) at positions -7, -4, -3, 0, +3, +4, +7 relative to the β subunit histidine that is modeled to bind one of the Bchls of the LHII Bchl dimer. These positions were chosen because they are one and two turns of the alpha helix away from the ligand site. Specific nucleotide mixtures were calculated using the program "PHYLO" to maximize the probabilities of occurrence of amino acids in the known phylogeny (i.e. the target set) according to the equation (Arkin & Youvan, 1992b; Youvan *et al.*, 1992) for group probability P_G:

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

where $P_D[i]$ is the probability of the *ith* amino acid in a target set (i.e. a subset of the 20 amino acids) occurring based on a specific doping scheme. These calculations correspond exactly to tabular data on optimized nucleotide

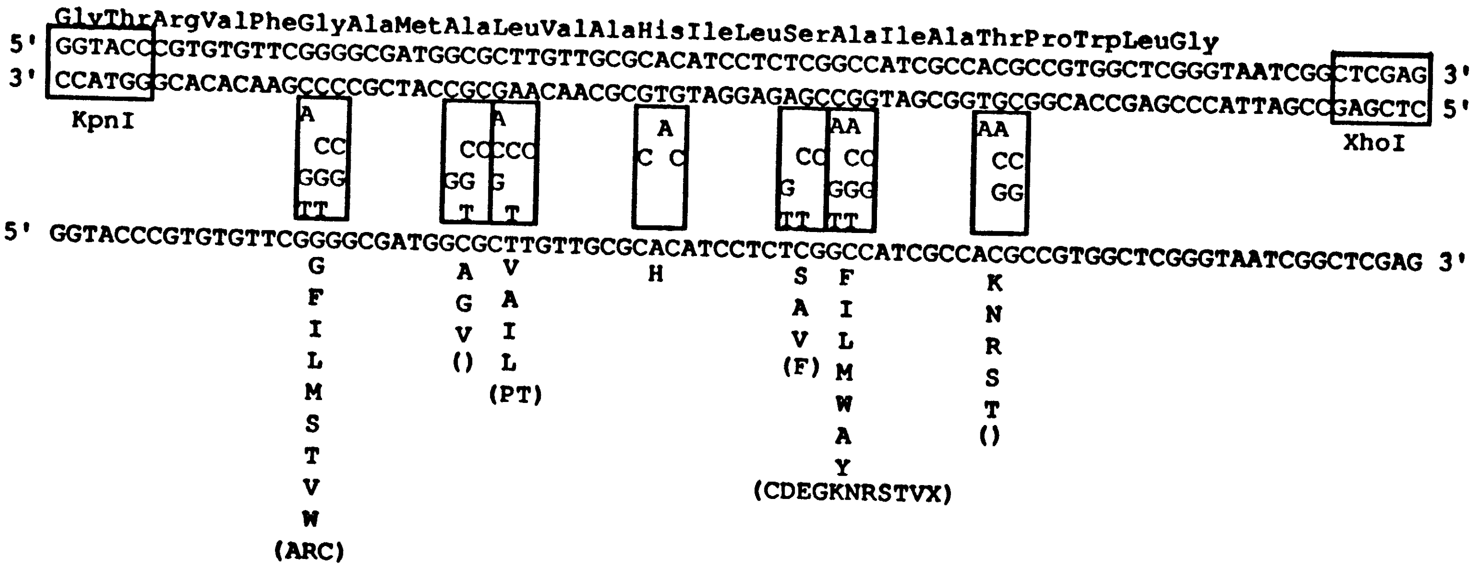

GlyThrArgValPheGlyAlaMetAlaLeuValAlaHisIleLeuSerAlaIleAlaThrProTrpLeuGly
5' GGTACCCGTGTGTTCGGGGCGATGGCGCTTGTTGCGCACATCCTCTCGGCCATCGCCACGCCGTGGCTCGGGTAATCGGCTCGAG 3'
3' CCATGGGCACACAAGCCCCGCTACCGCGAACAACGCGTGTAGGAGAGCCGGTAGCGGTGCGGCACCGAGCCCATTAGCCGAGCTC 5'
KpnI
XhoI
A CC GGG TT
CC GG T
A CC G T
A C C
CC G TT
AA CC GGG TT
AA CC GG
5' GGTACCCGTGTGTTCGGGGCGATGGCGCTTGTTGCGCACATCCTCTCGGCCATCGCCACGCCGTGGCTCGGGTAATCGGCTCGAG 3'
G F I L M S T V W (ARC)
A G V ()
V A I L (PT)
H
S A V (F)
F I L M W A Y (CDEGKNRSTVX)
K N R S T ()

Figure 4. Nucleotide and encoded polypeptide sequences of the combinatorial cassettes used in the phylogenetic TSM experiment. The boxes show the combination of nucleotides used at each codon position. The amino acids encoded by these optimized nucleotide mixtures are listed under the sequence. Amino acids in parentheses are unavoidably encoded by the dope, i.e. these latter residues are not found in the known phylogeny. Empty parenthesis indicate that no amino acid residues are encoded by the nucleotide mixture that are outside the target set (i.e. the known phylogeny).

mixtures (previously published in Arkin & Youvan 1992) that exemplify the properties of the P_G equation.

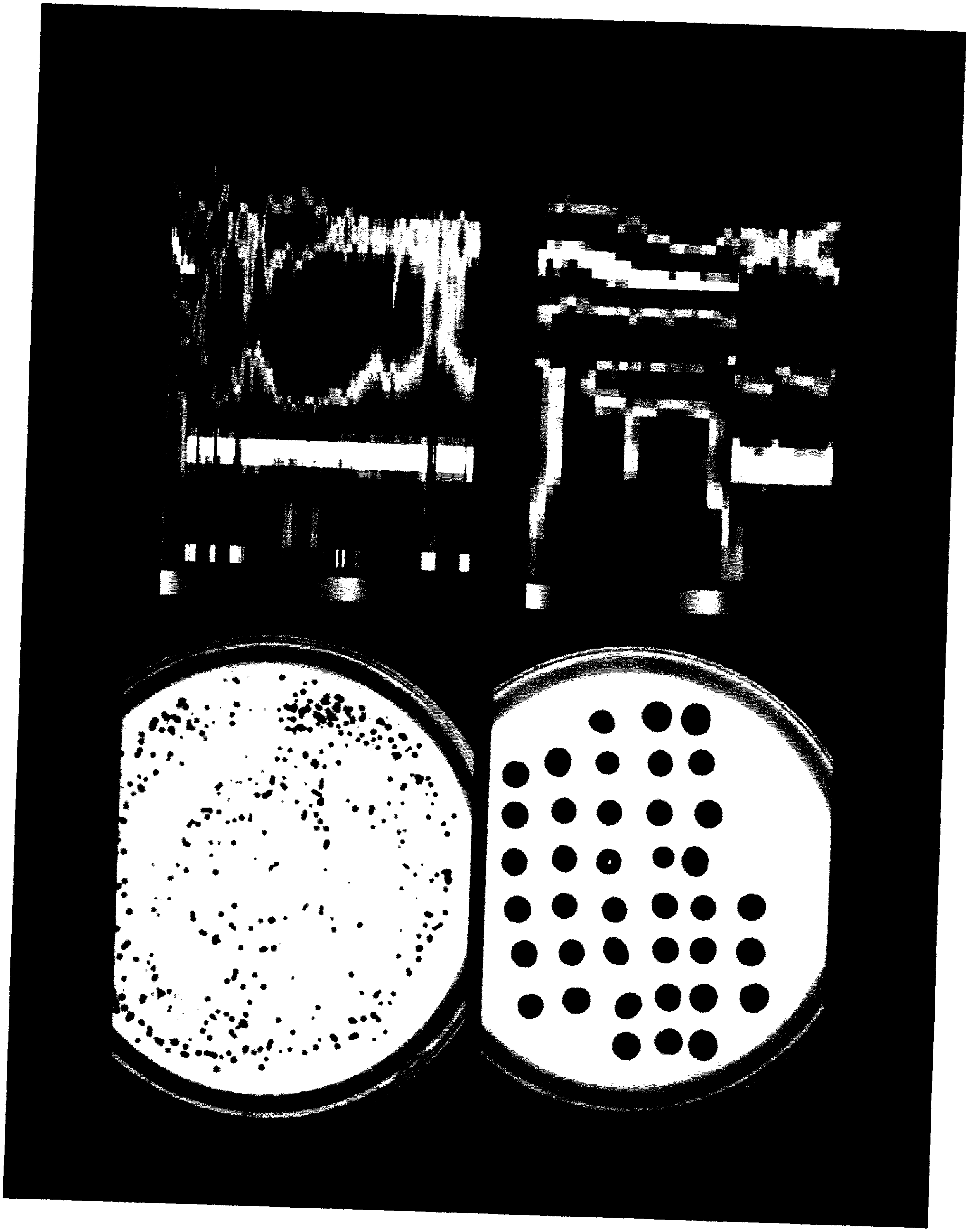
2.3.2 Spectroscopic screening of the TSM library

The TSM library was conjugated into strain U71 (an LHII$^-$ deletion background of *Rb. capsulatus* (Youvan *et al.*, 1985)) and approximately 10,000 transconjugants were screened (Figure 5) directly from petri dishes using DIS. Three major spectroscopic classes of mutants were observed: 1) pseudo wild type (pLH2) absorb at 800 nm and 860 nm, 2) pseudo LHI (pLH1) show reduced absorbance at 800 nm and maximal absorbance near 875 nm, and 3) "nulls" have absorption spectra characteristic of membranes bearing free Bchl. Approximately six percent of the library bound Bchl (4% pLH2 and 2% pLH1). To ensure that the spectra of the pLH1 mutants were not due to alpha complementation by LHI β, the libraries were also expressed in an LHI$^-$ LHII$^-$ chromosomal deletion background (Udd4, to be described in Appendix B). The spectra of the mutants were unchanged in this double deletion background, showing that the pLH1 proteins were encoded by the mutagenized LHII gene. In addition, spontaneous mutants of LHII have been isolated that express the pLH1 phenotype (Tadros *et al.*, 1989).

2.2.3 DNA sequencing of selected mutants

The DNA sequence of the LHII β subunit was determined for 38 mutants representative of the different spectroscopic classes generated by the phylogenetic dope (Table 2). The major sequence difference between pLH2 and pLH1 mutants occurred at the -7 position (relative to the histidine ligand). This amino acid is a glycine in wild-type, and a small residue in the pLH2 mutants. Out of the 13 pLH2 mutants sequenced, 7 glycines, 3 alanines, 1 cysteine, 1 threonine, and 1 serine were observed. All but one of the pLH1 mutants had an amino acid with a molar volume larger than threonine at this position (e.g. valine, isoleucine, leucine, methionine). The larger amino acid residues found at position -7 in the pLH1 mutants are interpreted as causing a change in the conformation of the helices, or a change in LHII aggregation that leads to the loss of the monomer Bchl absorption band and a red-shift in the

Figure 5. TSM mutants of the LHII gene screened by digital imaging spectroscopy (DIS). Panels A and C show monochrome images of the petri dishes taken at 400 nm. Panel A shows a typical spread of transconjugants resulting from the phylogenetically based combinatorial mutagenesis of seven sequence positions in the β subunit of LHII. Panel C shows "spots" of repurified mutants grown under similar (i.e. aerobic) conditions. Panels B and D are color contour maps (Arkin & Youvan, 1993) generated by DIS, where the horizontal axis corresponds to wavelength (700 nm - 930 nm) and the vertical axis is colony number. Each horizontal line represents a spectrum encoded by a pseudo color scheme that enables one to rapidly identify spectrally distinct mutant classes. DIS is radiometrically calibrated such that the color bar to the left of each contour map shows the range of optical densities for a particular colony or spot highlighted in the monochrome image. Spectra are sorted by similarity using a least squares criterion, and each colony spectrum is scaled to full deflection. Panel B shows 209 spectra from a spread of colonies. Colonies C1-C13 assembled protein as indicated by the spectra: pLH1 mutants are in the top 4 rows and pLH2 mutants in the next 9 rows. Panel D displays the spectra of typical mutants isolated and repurified after screening 10,000 colonies: S1-S12 are pLH1, P1-P13 are pLH2, and N1-N12 are nulls. The pLH1 mutants have an intense absorption band at 870 nm, while the 800 nm band is either diminished or absent. The pLH2 mutants display spectra similar to wild-type (805 nm and 860 nm, labeled WT in this figure). Null mutants absorb mainly at 760 nm due to free Bchl in the membrane. U71 is the chromosomal LHII deletion strain which serves as the genetic background for all constructions described herein.



	SEQUENCE POSITION						
	-7	-4	-3	0	+3	+4	+7
WT	G	A	L	H	S	A	T
S1	L	G	L	H	A	F	T
S2	L	G	L	H	A	A	T
S3	I	V	T	H	A	A	T
S4	V	G	V	H	A	G	T
S5	L	A	A	H	A	G	T
S6	M	A	L	H	A	A	T
S7	L	A	V	H	S	A	T
S8	V	A	V	H	S	W	T
S9	A	A	L	H	S	T	N
S10	L	A	V	H	S	A	T
S11	L	V	L	H	S	A	T
S12	V	A	L	H	S	G	T
P1	S	G	I	H	A	G	T
P2	C	G	I	H	A	G	T
P3	A	G	V	H	S	A	T
P4	T	A	I	H	S	M	T
P5	A	A	A	H	A	W	T
P6	G	A	L	H	A	G	T
P7	G	A	T	H	S	F	T
P8	G	G	I	H	S	Y	T
P9	A	V	L	H	S	A	T
P10	G	A	V	H	A	K	T
P11	G	G	A	H	A	V	T
P12	G	G	A	H	V	A	T
P13	G	A	V	H	A	F	S
N1	G	G	T	H	V	V	N
N2	L	V	I	H	A	G	K
N4	V	G	P	H	V	T	S
N5	F	G	L	H	F	T	T
N6	L	A	A	H	S	W	T
N7	W	G	I	H	A	A	R
N8	A	G	I	H	A	V	N
N9	G	A	V	H	A	S	R
N10	W	V	I	H	F	R	T
N11	W	V	A	H	S	L	T
N13	V	G	P	H	V	G	S
N14	V	V	L	H	F	S	K

Table 2. Amino acid sequences of the mutagenized positions of the three classes of mutants shown in Figure 5 (panels C and D). The column headings refer to the sequence position of the mutagenized residues relative to the Bchl-binding histidine (0). Each mutant and sequence is designated by the notation used in the first column which is consistent with the labeling in Figure 5.

dimer band. TSM shows that the +7 position is also important: threonine is conserved in 23 out of 25 of the Bchl-binding mutants (i.e. both pLH2 and pLH1). Position +4 is extremely variable: the phylogeny displays amino acid residues ranging in size from alanine to tryptophan, and 9 out of 25 of the Bchl-binding mutants used residues not found in the phylogeny (including one lysine).

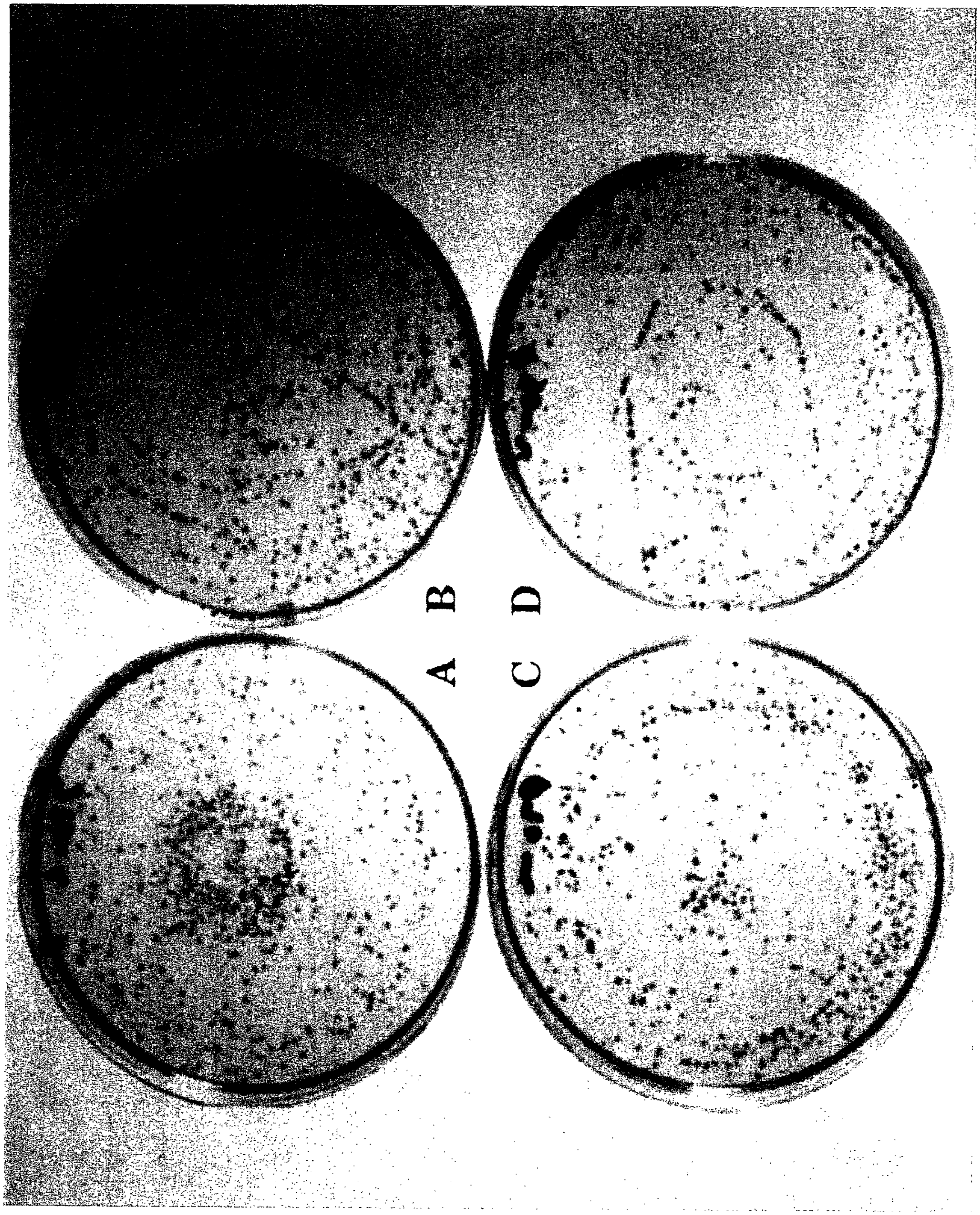
2.3.4 Comparison to random CCM

A direct comparison of the TSM and conventional CCM libraries is made in Figure 6. Based on DIS analysis, there is an excellent correlation between the "dark" colonies observed in Figure 6 and the expression of pLH1 or pLH2. This figure shows 860 nm absorption images of four petri dishes. Hundreds of times more dark colonies are observed in the TSM plates than in the conventional CCM plates. We observed only three positive mutants out of 10,000 screened in the conventional $[NN(G/C)]_7$ library compared to 6% "positives" in the TSM library. [In this doping nomenclature, N represents a 25% mixture of all four nucleotides.] Because of the small number of positive mutants observed in the TSM library, the $N^{1/2}$ law limits the accuracy of comparison of TSM over conventional CCM to approximately a 100 - 600 fold improvement in throughput. Most of this difference in throughput (or gain) is probably due to restricting the TSM dope to encode only histidine at the 0 position. Site-saturation mutagenesis of a Bchl-binding histidine residue in LHI (Bylina *et al.*, 1988) suggests that histidine is required for binding Bchl. Because histidine is encoded by only 1/32 of the NN(G/C) dope, we estimate that a six site random CCM library (at positions -7, -4, -3, +3, +4, +7) would yield about 1% throughput for Bchl binding mutants. This has been confirmed by the construction and expression of a conventional six site CCM library wherein the axial histidine residue was not mutagenized (data not shown).

2.4 DISCUSSION

These experiments demonstrate the use of LHII as a model protein for studying new mutagenesis techniques that are guided by computer algorithms. Phylogenetically based TSM experiments, in which seven sequence positions

Figure 6. Comparison of the TSM library and conventional CCM library by absorption at 860 nm. Panels A and B show spreads of colonies from the conventional CCM library. Panel A shows an atypical plate with one positive (dark) colony that is indicated by an arrow near the center of the plate. [Note that this was the only positive colony observed in 10 similar spreads totaling approximately 4000 colonies.] Panel B is a more typical plate from conventional CCM which shows no positive colonies. Panels C and D show typical spreads of the analogous TSM library. Approximately 6% of these colonies are categorized as positive. Images were recorded using a CCD camera (f/5.6; 10 second exposure; 860 nm illumination). The original image of the petri dishes was divided by a blank image at 860 nm to correct for uneven light intensity. Grayscale values from the ratioed image were rescaled to enhance contrast.


A
B
C
D

were simultaneously mutagenized, yielded fewer null mutants than conventional CCM by a factor of several hundred times.

The probability of finding "positive" mutants in a random CCM library is greatly diminished when critical sequence positions accept only a few amino acid substitutions. As a first approximation, the stringency of a specific sequence position in a protein can be estimated from phylogenetic and single-site saturation mutagenesis data. In the worse case, if only one amino acid residue is functional at a given sequence position, and if this residue is represented by only one codon in the NN(G/C) mixture, 31 out of 32 CCM mutants are nullified per position. TSM throughputs will be highest in comparison with conventional CCM under these conditions.

Using TSM and extrapolating the current data, we expect a throughput of $(0.06)^3$ or 22 positives for every 100,000 screened in a 21 site mutagenesis experiment of comparable stringency. In contrast, we expect a throughput of only $(0.0003)^3$ for a 21 site random CCM library. This corresponds to only 3 positive mutants in 10^{11}, which is beyond current cloning and screening capabilities. TSM is essential in such experiments.

The P_G error function which was used to design the TSM cassette, forces the nucleotide mixtures to encode all amino acids present in the target set regardless of the frequency they are found in the phylogeny. A second function (Arkin & Youvan 1992b; Youvan *et al.*, 1992) that can be used to adjust nucleotide concentrations in a combinatorial cassette uses a sum of the squares of differences criterion:

$$SSD = \sum_{i} (P_D[i] - P_T[i])^2 \qquad \text{Eq. 2}$$

$P_D[i]$ is defined as in Eq.1 and $P_{T[i]}$ is the fractional representation of the *ith* amino acid in the target set. Unlike P_G, the SSD sum is taken over all twenty amino acids rather than a restricted target set. SSD takes into account the relative frequency of occurrence of all amino acids within the target set and may omit infrequently used amino acids from the dope. The SSD error function is expected to yield a higher throughput of positive mutants than P_G, possibly at the expense of phenotypic diversity. Computer simulations (Youvan *et al.*,1992) have been used to compare P_G and SSD. Experimental comparisons using the LHII protein as a model system are presented in Chapter 3.

Combinatorial mutagenesis schemes should not be dependent upon the properties of any one protein. Algorithms should be tested on a variety of proteins and structural motifs. Myoglobin expressed in *E. coli* (Springer & Sligar, 1987) is amenable to spectroscopic screening by DIS, wherein the visible heme spectrum is analogous to the NIR Bchl spectrum as a reporter. Studies on this globular protein would complement current studies on the LHII membrane protein.

In cases where phylogenetic data are not available, TSM can be based on the results of one iteration of conventional CCM using NN(G/C) triplets. This process defines recursive ensemble mutagenesis (REM), which has been studied by computer simulation (Youvan *et al.*, 1992; Arkin & Youvan 1992a) and recently tested in LHII mutagenesis (Delagrave *et al.*, 1993). Once perfected in simple model systems, TSM and REM technology can be transferred to other areas of protein engineering, including the expression of peptides and proteins in phage display libraries (Smith, 1985; Roberts *et al.*, 1992; Hoogenboom *et al.*, 1991; Kang *et al.*, 1991; Gherardi & Milstein, 1992). This new technology should be very useful in the mutagenesis of antibodies (Hoogenboom *et al.*, 1991; Kang *et al.*, 1991), where combinatorial complexity based on conventional CCM rapidly exceeds screening or selection capacities (Gherardi & Milstein, 1992). It is noteworthy that the spectral diversity achieved in these LHII TSM experiments should be analogous to catalytic and/or binding site diversity in other proteins.

2.5 EXPERIMENTAL PROTOCOL

2.5.1 Strains and plasmid vectors

An LHII expression plasmid (pU4) was constructed by ligating the PstI-BamHI fragment of pU2 into the conjugative plasmid pRK404 (Ditta *et al.*, 1985). A unique Hind III restriction endonuclease site was engineered for shuttling fragments into M13mp18 for DNA sequencing (M13 phage derivatives were maintained in *E. coli* strain MV1190). In addition, unique Kpn I and Xho I sites were incorporated in pU4 around the region encoding the dimer Bchl binding site in the β subunit to facilitate cassette mutagenesis (pU4b). Derivatives of pU4 were maintained in *E. coli* strain S17-1 (Simon *et al.*, 1983). Plasmid pU4 derivatives were conjugated into *Rb. capsulatus* (Bylina *et al.*, 1989) strain U71 (LHII chromosomal deletion background; LHI and RC expression inactivated by a point mutation) from *E. coli* S17-1 donors.

2.5.2 Cassette construction

The sense strand of 113mers including the Kpn I - Xho I fragment (as shown in Figure 3) was polymerized on an Applied Biosystems model 381 DNA synthesizer. At each of the mutagenized positions of the 113mer, the nucleotide ratios shown in Figure 3 were used. In the conventional CCM experiment, NN(G/C) triplets were used at positions -7, -4, -3, 0, +3, +4, and +7. Twenty nucleotide long sense and antisense primers were synthesized for PCR of the combinatorial cassette. PCR reactions were performed in a total volume of 100 µl, and contained 0.5 pmol template, 100 pmol each primer, 100 mM each dNTP, 50 mM KCl, 10 mM Tris-HCl pH 8.3, 1.5 mM $MgCl_2$, 0.01% gelatin, 10 mg BSA, 5 units Taq polymerase (Perkin-Elmer Cetus). Samples were amplified for 12 cycles in a Coy Tempcycler: 96° C denature (1 min), 55° C annealing (30 sec), 72° C elongation (1.5 min). The fragment was cut with both restriction enzymes (Kpn I and Xho I), purified by gel electrophoresis, and ligated into the unique Kpn I and Xho I sites of the LHII expression plasmid pU4b.

2.5.3 Screening by digital imaging spectroscopy

Colonies were imaged as spreads on RCV-tetracycline (2.5 µg/ml) plates from bacteria resuspended after conjugation. Representatives from each spectral class of mutants (pLH2, pLH1, and nulls) were repurified several times on selective media, spotted on RM plates, and sequenced. The number of TSM transformants and transconjugants exceeds the number of mutants screened by DIS by a factor of four, thus insuring the sequence complexity of the library. The configuration of the digital imaging spectrophotometer used for these assays is described in Youvan & Arkin 1993 and in Youvan *et al.*, 1993.

2.5.4 Screening monochrome images

A second generation digital imaging spectrophotometer has recently been constructed that uses a 36 inch hemisphere with an 18 inch by 18 inch exit port to illuminate up to 25 petri dishes, simultaneously (Youvan *et al.*, 1993). Monochrome images (Figure 6) were generated by capturing a 4 megapixel image of the exit port with a Photometrics CCD camera. A "blank" image was also recorded. Both images were transferred to a Silicon Graphics Crimson Elan computer. The ratioed image was rescaled, modified by a gamma function, and displayed. Monochrome images used Fabry-Perot filtered light at 860 nm (5 nm bandpass) in a transmission mode.

2.6 Acknowledgements

The information in this chapter was published in Bio/Technology Vol 10, pp. 1557-1561, 1992.

2.7 Note

At the time we designed this experiment we did not realize that an asparagine substituted for histidine at the putative Bchl binding site in one example from the known phylogeny. Furthermore site specific mutagenesis (Bylina *et al.*,1988b), and random CCM (Delagrave & Youvan, 1993) indicates that asparagine never substitutes for histidine in *Rb. capsulatus* LH at a putative Bchl binding site. His was the only amino acid found in the "0" position when it was randomized.

Chapter 3: Nucleotide formulations

3.1 Introduction

The complexity of a protein library resulting from randomization of multiple amino acid residues rapidly grows too large to screen as additional positions are mutagenized. As demonstrated in the previous chapter, target set mutagenesis (TSM) can be used to reduce the total number of proteins in a library while enriching for functional mutants by restricting each position to accept only a subset of amino acids based on some criterion such as physicochemical properties, expert rules, or phylogenetic data (Goldman & Youvan, 1992). There are many ways to calculate nucleotide mixtures based on a target set of desired amino acid residues (Arkin & Youvan, 1992b; Youvan *et al.*, 1992). We used the LHII genetic system from *Rb. capsulatus* in conjunction with DIS to experimentally compare two schemes for formulating nucleotide dopes.

Phylogenetic data from 29 homologous light harvesting proteins (Zuber, 1990) were used to construct target sets for seventeen sequence positions on the β subunit flanking the monomer Bchl binding site. Two cassettes were constructed: 1) using the group probability (P_G) algorithm which includes each member of the target set in the dope no matter how rarely it occurs in the phylogeny, 2) using the sum of the squares of the differences (SSD) formulation which takes into account the frequency of each amino acid in the phylogeny and may drop infrequently occurring members of the target set.

As predicted by computer simulation (Youvan *et al.*, 1992), the throughput of positives (pigment binding mutants) was much higher using the SSD calculation. We found 2% pigment binding recombinants in the SSD formulated library while less than .01% of the P_G based library bound Bchl. The SSD calculation seems to be the appropriate method to use when mutagenizing a large number of sequence positions.

In a second experiment, the SSD algorithm was used to construct nucleotide dopes from phylogenetic data on seventeen positions flanking the Bchl of the dimer associated with the β subunit. The resulting library had high throughput (5-10% positives) and spectrally diverse phenotypes.

3.2 Results

3.2.1 Determining nucleotide dopes

The two methods used to adjust the codons at each mutagenized amino acid site are:

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

where $P_D[i]$ is the probability of the ith amino acid in a target set occurring based on a specific doping scheme

$$SSD = \sum_i (P_D[i] - P_T[i])^2 \qquad \text{Eq. 2}$$

where $P_T[i]$ is the fractional representation of the ith amino acid in the target set. The best dope is found by either maximizing P_G or minimizing SSD. The P_G function forces every amino acid in the target set to be coded for in the cassette. In contrast, the SSD algorithm takes into consideration how many times each amino acid occurs in the target set and may drop infrequently used amino acids from the dope. A computer program (CYBERDOPE) was used to convert target sets of amino acids into nucleotide mixtures using either the P_G or SSD algorithm.

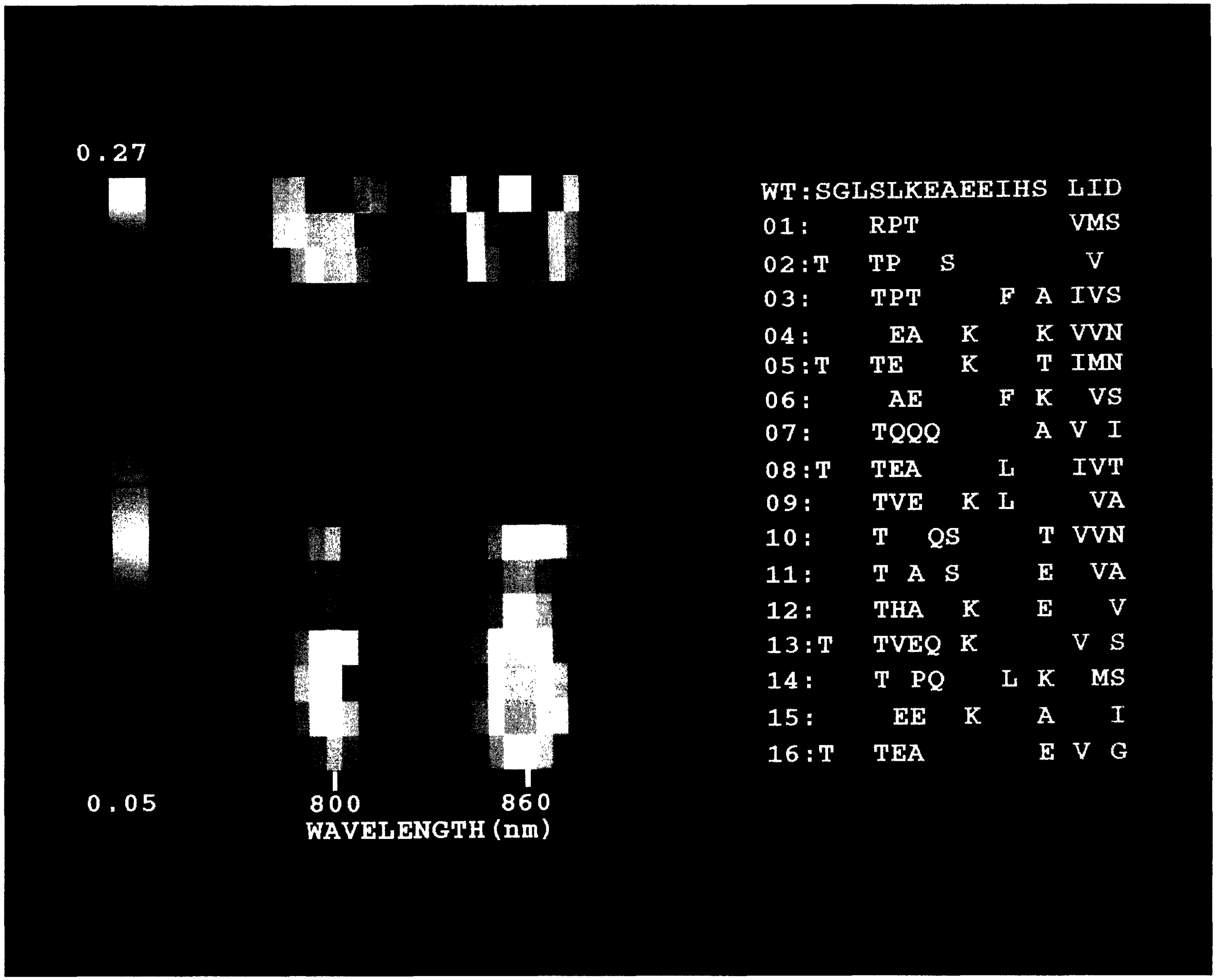
3.2.2 Throughput comparison SSD vs P_G

Seventeen amino acid positions in the β subunit in the vicinity of the Bchl monomer binding site were mutagenized simultaneously. SSD was used with phylogenetically based target sets except at position β21 where a dope was used that omitted the WT amino acid (Tyr) but still encoded the other two aromatic amino acids and other amino acids found in the core antenna phylogeny (Table 3). Ten thousand colonies were screened by DIS, and even with the inclusion of a non-WT position, 2% of the library was found to assemble LH antennae. Most of the 'positives' showed a WT like absorption spectrum in the NIR, albeit with lower optical density (Figure 7). Several mutants showed

Site	Dope(SSD)	Amino Acid Residues
β8	(AT)C(GC)	ST
β9	GG(GC)	G
β10	(TC)T G	L
β11	A(GC)(GC)	ST(R)
β12	(GC)(ATC)(GC)	L*DEPAV(QH)
β13	(GAC)(AC)G	K*AEQ(PT)
β14	(GC)AG	EQ
β15	(GT)C(GC)	AS
β16	(GA)AG	EK
β17	GAG	E
β18	(ATC)TC	I*LF
β19	CAC	H
β20	(GAT)(AC)G	S*AEK(TX)
β21	**(GATC)(GT)(GC)**	**LIMVFW(RSCG)**
β22	(GAC)TC	L*FV
β23	(GA)(AT)(GC)	IKMV(DEN)
β24	(GA)(GATC)C	DSTV(AGIN)

Table 3. Nucleotide dopes used for the SSD based mutagenesis of the Bchl monomer binding site of LHII. Amino acids listed in parenthesis are not in the target set, but are unavoidably coded for in the dope, because of the structure of the genetic code. The asterisks indicates that the (preceding) WT amino acid had to be entered into the algorithm more often than it occurred in the phylogeny to ensure that it was encoded in the dope. Target sets were based on phylogeny except at position β21 (in bold).

Figure 7. Correlation of sequence data with spectra from a combinatorial mutagenesis experiment on LHII near the β-subunit Bchl (monomer) binding site. This color contour map is displayed in 'absolute' mode with the highest absorbance on the image set to white (OD=0.72) and the lowest absorbance on the image set to black (OD= 0.05). The top spectra is WT which shows stronger absorption than these mutants. The amino acid sequence for each mutant is displayed to the right of its spectrum. Only amino acids different from the WT sequence are listed. The position indicated in red did not include the WT Tyr in the target set.


0.27
0.05
800
860
WAVELENGTH (nm)
WT:SGLSLKEAEEIHS LID
01: RPT VMS
02:T TP S V
03: TPT F A IVS
04: EA K K VVN
05:T TE K T IMN
06: AE F K VS
07: TQQQ A V I
08:T TEA L IVT
09: TVE K L VA
10: T QS T VVN
11: T A S E VA
12: THA K E V
13:T TVEQ K V S
14: T PQ L K MS
15: EE K A I
16:T TEA E V G

small (5nm) blue shifts of the monomer peak. Although these shifts are small they were reproducible and were also seen in chromatophore preparations of the LHII mutants. Sixteen of these pigment binding mutants were sequenced, and each one was found to have a unique sequence.

A second cassette was constructed (using the same seventeen sites) according to the equation for P_G. This library used a phylogenetically based target set at all of the mutagenized positions (Table 4). Out of 10,000 colonies screened no positive recombinants were found as judged by imaging spectroscopy.

As a rough comparison of the theoretical complexity of the P_G and SSD libraries, the amino acid complexities were calculated from the number of amino acids coded for at each position. The amino acid complexity is 1 x 10^{14} for the P_G and 7 x 10^7 for the SSD libraries. Using the SSD algorithm serves to reduce the protein sequence possibilities by a factor of 10^6 at the coding level.

3.2.3 SSD Cassette mutagenesis in the vicinity of the dimer Bchl

The SSD equation was used to construct a phylogenetically based cassette to mutagenize seventeen amino acid residues (β27-β43) in the region of the binding site of the Bchl dimer associated with the β subunit of LHII. The amino acid complexity of the resulting library is 7x10^7 as opposed to 3x10^{13} if the P_G algorithm had been used to construct the cassette. DIS was used to screen 5000 colonies; the library showed an extremely high throughput (5-10%) of mutants expressing LH, with good phenotypic diversity (Figure 8). Extrapolating from the results of the seven site phylogenetically based P_G experiment in this region, a dope which used the P_G algorithm on seventeen sites would generate a 0.03% throughput of positives. Thus, while SSD reduced the total number of possible protein sequences in the library by orders of magnitude, different classes of Bchl binding mutants were still observed.

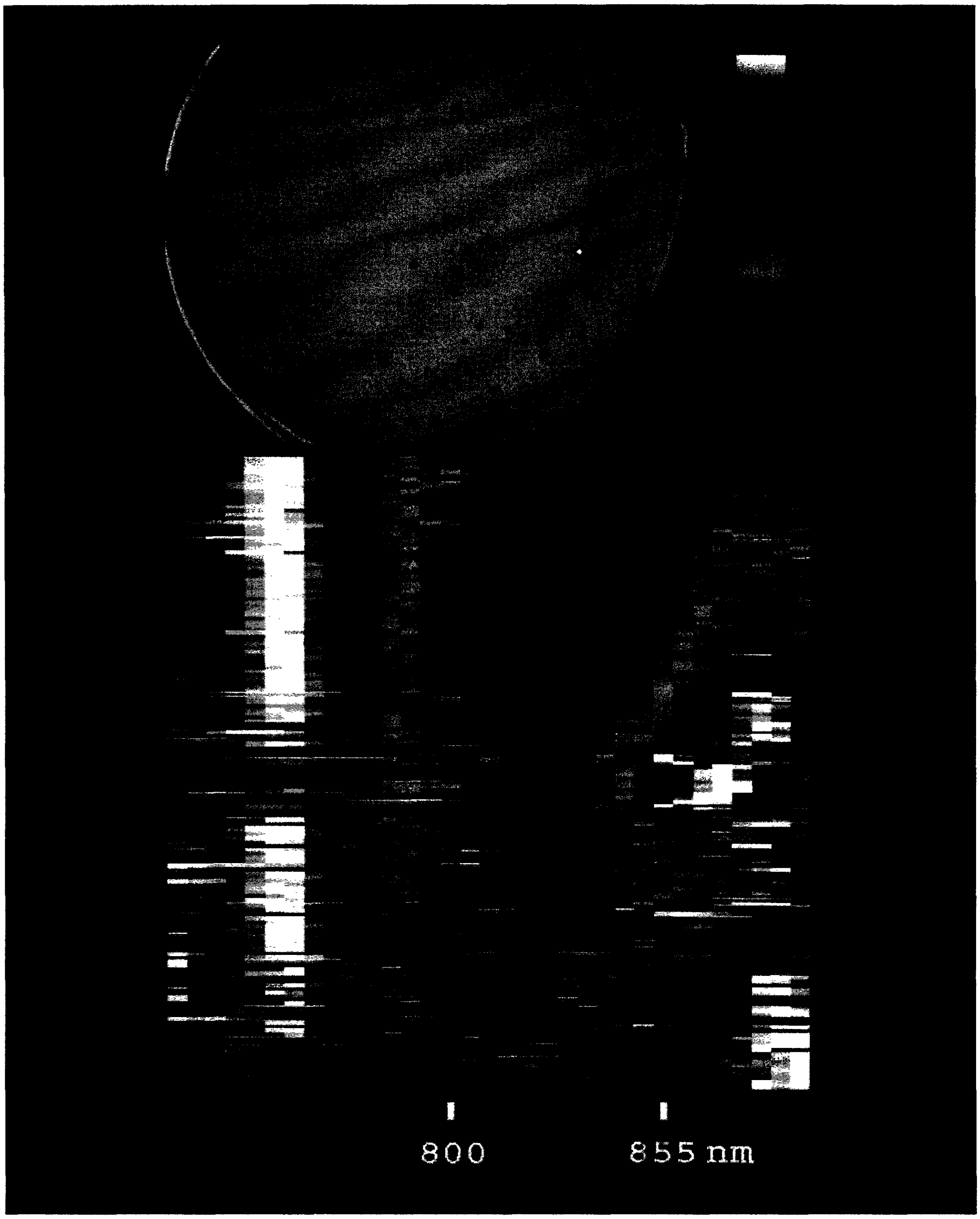
3.3 Discussion

We experimentally confirmed that SSD gives a higher throughput than P_G when a large number of amino acid sites (i.e., 17) are simultaneously mutagenized. Finding no positives in the P_G library in conjunction with the 2% throughput from the modified SSD library indicates that SSD gives at least a

Site	Dope(PG)	Amino Acid Residues
β8	(GA)(GAC)(GC)	STEAKNR(DG)
β9	(GAC)(GACT)C	GNPTV(ADHILRS)
β10	(GAC)TC	LVI
β11	(TA)(CT)C	STF(I)
β12	(TACT)(ACT)C	LAVIEDPNS(FHKMQSYX)
β13	(GAC)(GAC)(GC)	KAEQGNS(HPRST)
β14	(GC)AG	EQ
β15	(GT)(GAC)C	ASCD(GY)
β16	(GATC)(GA)G	EKWQR(GX)
β17	(GA)A(GC)	EN(DK)
β18	(GATC)(TC)C	ILFVA(LPST)
β19	CAC	H
β20	(GA)(GAC)(GC)	SKEAGP(NRT)
β21	(GATC)(AT)(GC)	YQHIMVL(DEFKNX)
β22	(GATC)(AT)C	LVFNY(DHIL)
β23	(GA)(ATC)(GC)	IKVTM(ADEN)
β24	(GATC)(ATC)(GC)	DVLKQST(AEFHIMNPYX)

Table 4. Nucleotide dopes used for the phylogenetic P_G based experiment in the vicinity of the Bchl monomer binding site of LHII. This algorithm forces each amino acid in the target set to be coded for in the dope regardless of its frequency. Amino acid residues in parenthesis are not in the target set, but were unavoidably encoded because of the genetic code.

Figure 8. Color contour map showing the spectral diversity of a seventeen site combinatorial library affecting amino acid residues near the β-subunit Bchl (dimer) binding site. The upper left hand panel shows a monochrome image taken at 400 nm of a typical spread of colonies. The upper right hand panel shows a color bar; black corresponds to the lowest absorbance and white corresponds to the highest absorbance. The lower panel is a color contour map generated by DIS where the horizontal axis corresponds to wavelength (730-890 nm) and the vertical axis is colony number. Each horizontal row represents a spectrum encoded by pseudocolor. Spectra of Bchl binding mutants are enclosed in the gray box. Mutants are displayed in 'full deflection' mode so each row scales from black to white. Nine percent of the colonies (15 out of 168) were judged to assemble LH antennae. Several categories of spectra are observed: 1) pseudo WT (i.e., dimer peak at 855 nm, monomer peak at 800 nm), 2) pseudo-WT mutants with a 5 nm red-shift of the dimer band, 3) pseudo-LHI (i.e., 10-15 nm shift of the dimer band and reduced or missing monomer peak), 4) a mutant showing a single peak at 855 nm, and 5) mutants absorbing mainly at 760 nm (due to free Bchl in the membrane) which are classified as 'nulls'.


800
855 nm

200 fold enhancement in the number of Bchl binding proteins. When non-stringent sequence positions are mutagenized, P_G has the tendency to go towards a random (NNG/C) dope as each amino acid in the target set is required to be coded for by the dope. The complexities of P_G based libraries can be orders of magnitude higher that those of their SSD formulated counterparts in these experiments. When using TSM on a large number of sites (> seven), it is advantageous to use the SSD algorithm to construct cassettes because it increases the throughput of the library over P_G.

It was anticipated that phenotypic diversity might be lost through the SSD algorithm. However both the SSD libraries still show variable phenotypes. The library in the dimer region showed mutants with a few different spectral classes. P_G might lead to more unusual phenotypic variation, but at the expense of having to increase the screening size by orders of magnitude. Although DIS allows rapid screening from petri dishes, screening greater than 100,000 mutants per library may be impractical.

When using TSM on more than six sites, or with permissive sequence positions it is advantageous to use the SSD algorithm to construct cassettes. SSD drastically increases the throughput of the library over PG while maintaining phenotypic diversity.

3.4 Experimental protocol

3.4.1 Strains and plasmid vectors

A unique Nsi I site was incorporated into the LHII expression plasmid (along with the Kpn I and Xho I sites in pU4b) to facilitate cassette mutagenesis of the β subunit in the monomer Bchl binding region as an Nsi I - Kpn I fragment (pU4c). Derivatives of pU4c were maintained in *E. coli* strain S17 and conjugated into *Rb. capsulatus* strain U71 (LHII chromosomal deletion background, LHI and RC inactivated by point mutation) from *E. coli* donors. Individual positive mutants were repurified by streaking several times on selective media (RCV-tetracycline); the LHII genes were then cloned into M13 for single strand sequencing. TSM of 17 sites in the Bchl dimer binding region of β was done as described in the previous chapter using the Kpn I and Xho I sites of pU4c.

3.4.2 Cassette construction

For each mutagenized codon the amino acids in the target set were entered into the computer program CYBERDOPE, and nucleotide mixtures were calculated both to maximize P_G and minimize SSD (Tables 3 and 4). If the WT sequence was not included in the SSD dope, we forced the SSD algorithm to include the WT codon by increasing the frequency of the WT amino acid in the target set. Target sets were composed of phylogenetic data except in position β21 of the β monomer SSD based experiment, where the WT amino acid (Tyr) was not included in the target set, but the other two aromatic amino acids were, as well as phylogenetic data from core antennae.

The CYBERDOPE output for the β monomer region was used to construct two cassettes; one based on the P_G information, and a second based on the SSD results. The sense strands of 121mers which include the Nsi I and Kpn I sites as well as the sequences corresponding to two PCR primers (21mers each spanning a restriction site) were synthesized. The purified 121mers were amplified by PCR using the conditions outlined in section 2.5.2. The amplified double strand cassette was purified by phenol extraction and ethanol precipitation. Complete digestion of the cassette with Nsi I and Kpn I was carried out in a single incubation. The digested cassette was purified by polyacrylamide gel electrophoresis, and ligated into the Nsi I and Kpn I sites of the LHII expression plasmid pU4c, and electroporated into *E. coli* strain S17.

CYBERDOPE output using the SSD formulation was used to construct a cassette for the β dimer region experiment. A 113mer including the Kpn I and Xho I sites was polymerized and amplified as in section 2.5.2. Digested cassette was ligated into the Kpn I and Xho I sites of pU4c and electroporated into *E. coli* strain S17.

3.4.3 Library constructions

Complexities for each transformation were estimated by plating aliquots of the transformation on L-tetracycline (12.5 μg/ml) plates (after allowing two hours for resistance expression). The remainder of the transformation was incubated over night (12-14 hours) in 60 ml L-tetracycline. Plasmid pU4c derivatives were conjugated from *E. coli* S17 donors into *Rb. capsulatus* strain

U71. The libraries were expressed by U71 transconjugants and grown for screening on RCV-tetracycline plates at 32 degrees C.

The β monomer region libraries had complexities on the order of 10^5. Several independent ligation/transformations were done for each library. A total of 10,000 mutants were screened for each library. The β dimer region libraries had complexities on the order of 10^4. Several independent ligation/transformations were performed and 5,000 colonies were screened.

3.4.4 Screening

Colonies were screened by DIS. Representative positives were chosen for sequencing.

3.5 Acknowledgements

The program CYBERDOPE was made available by Dr. Mary Yang.

Chapter 4: Phenotypic estimation
Collaborator: Georg Fuellen

4.1 Summary

Correlations between protein sequences and phenotypes were explored using databases of sequence and corresponding spectral information established from combinatorial cassette mutants of pigment-proteins (LHII and RC). Heuristically formulated decision algorithms and computer implemented neural networks were tested to determine their accuracy in spectroscopic (i.e., phenotypic) classifications of the data. Simple single-site-based decision algorithms were able to separate spectral classes based on amino acid sequence 80-84% of the time, due to the stringency of critical amino acid positions. Neural networks scored up to 10% higher than the primitive rules on the same sequence databases. If the critical sites for the decision algorithm are omitted, the efficiency of the neural network is still much better than random: 74% of the members of the LHII library, and 87% of the RC library are correctly sorted. This implies that there are some determining factors in the sequence of these proteins outside the highly stringent sites used by the decision algorithms. A linear perceptron scores 6% lower than a more sophisticated three-layer network on the RC data (1% lower than the best performing network for the LH data), indicating some of the factors important for sorting sequences may involve nonlinear (site-to-site) interactions. However, the success of the primitive decision algorithms and perceptrons at sorting sequences into categories suggests that to a first approximation simple features predominate in the determination of a phenotype.

4.2 Introduction

Parallel construction of large populations of molecules by random mutagenesis or chemical synthesis has become an important approach for biopolymer engineering. The advances in combinatorial biological and chemical techniques make it necessary to develop methods for analyzing large databases of sequence-function correlations. This is particularly crucial when reiterative methods are planned to improve the library.

Combinatorial cassette mutagenesis (CCM) provides molecular biologists with a powerful method of exploring mutations in a protein (Oliphant *et al.*, 1986; Reidhaar-Olson *et al.*, 1991). Entire segments of a gene can be replaced with cassettes of synthetic DNA in which multiple codons have been changed randomly or semi-randomly. Selection or screening criteria are established by the experimenter to classify the CCM mutants as "positives" or "nulls". Positive mutants may be pseudo wild type (phenotypically indistinguishable from wild type) or may express novel phenotypes, while nulls have no functional protein assembly as judged from the selection or screening criterion. Currently, a major challenge lies in interpreting the massive amount of sequence information from a CCM experiment; correlations between the amino acid sequence of a mutant versus its phenotype may not be obvious.

The efficiency of simple decision algorithms (DAs) constructed by the experimenter (as previously described, Arkin & Youvan, 1992; Youvan *et al.*, 1992) and artificial neural networks (ANNs) was compared in the analysis of amino acid sequence data from CCM experiments. The DAs considered in this chapter are in the form of simple decision trees based on amino acid positions judged to be critical in phenotypic determination. ANNs allow one to model nonlinear, almost arbitrary interactions between the input variables (in this case, the amino acid sequence of the mutants). Unlike DAs, ANNs considers that the determination of protein phenotype is influenced by non linear interactions between amino acids in the chain.

We examined sequence data from a phylogenetically based CCM experiment of the light harvesting II (LHII) antenna (Goldman & Youvan, 1992) and random CCM of the bacterial reaction center (RC) (Robles & Youvan, 1993) from *Rb. capsulatus.* LHII is a peripheral LH protein which binds both a dimer and monomer of bacteriochlorophyll (Bchl) (Zuber, 1986). These pigments can be specifically detected by their near infrared absorption spectra (dimer absorbing at 855 nm; monomer at 800 nm) and provide a colorimetric indicator of protein expression and assembly. Digital imaging spectroscopy (Yang & Youvan, 1988; Arkin *et al.*, 1990; Arkin & Youvan, 1993; Youvan *et al.*, 1993) was used to screen the LHII CCM library. Two classes of mutants were observed in addition to nulls: pseudo wild type (spectrally similar to wild type), and pseudo LHI (spectrally resembling the LHI core antennae). Sixty two of the positives comprising both of the classes were sequenced. The RC is the pigment-protein responsible for charge separation, and without a functional RC, the bacteria cannot grow photosynthetically. Mutants were selected for photosynthetic growth, and sequences for 25 functional RC mutants were analyzed.

4.3 Materials and methods

4.3.1 Digital imaging spectroscopy

Digital imaging spectroscopy (DIS) facilitates the parallel screening of the ground state visible and NIR spectra from hundreds of colonies directly on a petri dish. A charged coupled device (CCD) acts as a detector to image petri plates mounted on the exit port of an integrating sphere. The light source uses an 1/8 meter monochromator to illuminate the dish at different wavelengths, and the integrating sphere produces uniform illumination. Spectra are obtained at 5-10 nm resolution; 2 nm band shifts can be detected. For quick analysis of the data, all the spectra from a single petri dish can be presented as a two dimensional color contour map display. Each colony is represented by a horizontal row; absorption is color coded at each wavelength along the row (white = highest absorption, black = lowest absorption). The spectra can be sorted according to similarity, maximal absorption at various wavelengths, or wavelength of maximum absorption. Different display modes allow the spectra to be scaled relative to the lowest and highest absorption anywhere in the image, or to have each spectrum scaled between its own maximum and minimum absorption.

4.3.2 LHII sequence and spectral database

Additional "positive" mutants were spectrally characterized and sequenced from the previously constructed library described in Chapter 2. Seven amino acid residues in the β subunit were simultaneously mutagenized using combinatorial cassettes based on phylogenetic target sets. The mutagenized positions were chosen to be on one face of a transmembrane alpha helix that comprise the binding site for one of the Bchls of the dimer. DIS showed that 6% of the library bound Bchl in two spectroscopic classes: 1) pLH2 mutants have wild type like absorption characteristics with an 855 nm dimer peak and 800 nm monomer peak, 2) pLH1 mutants have the dimer band red-shifted to 865 nm and the monomer band reduced or absent. Mutants which showed only absorption characteristic of free pigments in the membrane (760 nm absorption) were classified as nulls. The complete sequence data are shown in Table 5. Each mutant was classified according to its spectral characteristics. Figure 9 shows a color contour map representing the spectra of each sequenced positive. The phenotypes of a few of the mutants appeared to be growth dependent, therefore there is the potential for a 10% experimental misclassification based on how old the colony was when imaged.

4.3.3 Reaction center mutants

We used the sequence information from the nine-site library described by Robles and Youvan, 1993. Nine amino acids in the vicinity of the monomeric Bchl in the active branch of the RC were randomized simultaneously using CCM. These positions were both in the L and M subunits of the RC, and according to the X-ray structure of the *Rhodopseudomonas viridis* RC (Deisenhofer *et al.*, 1985), all are in van der Waals contact with the active branch monomer Bchl. Functional mutants were selected by photosynthetic growth selection: 1/50,000 colonies plated was found to be functional. Representative positives were sequenced and spectrally characterized.

4.3.4 Formulating and evaluating decision algorithms

DAs are a protein grammar that serves to evaluate each sequence and either classify the protein as a positive, or reject it as a null. Each DA is a set of rules that can be based on experimentally observed stringent sequence positions, or a combination of criteria such as heuristic rules for protein folding, energy minimization, or overall change in hydropathy and/or molar volume relative to WT.

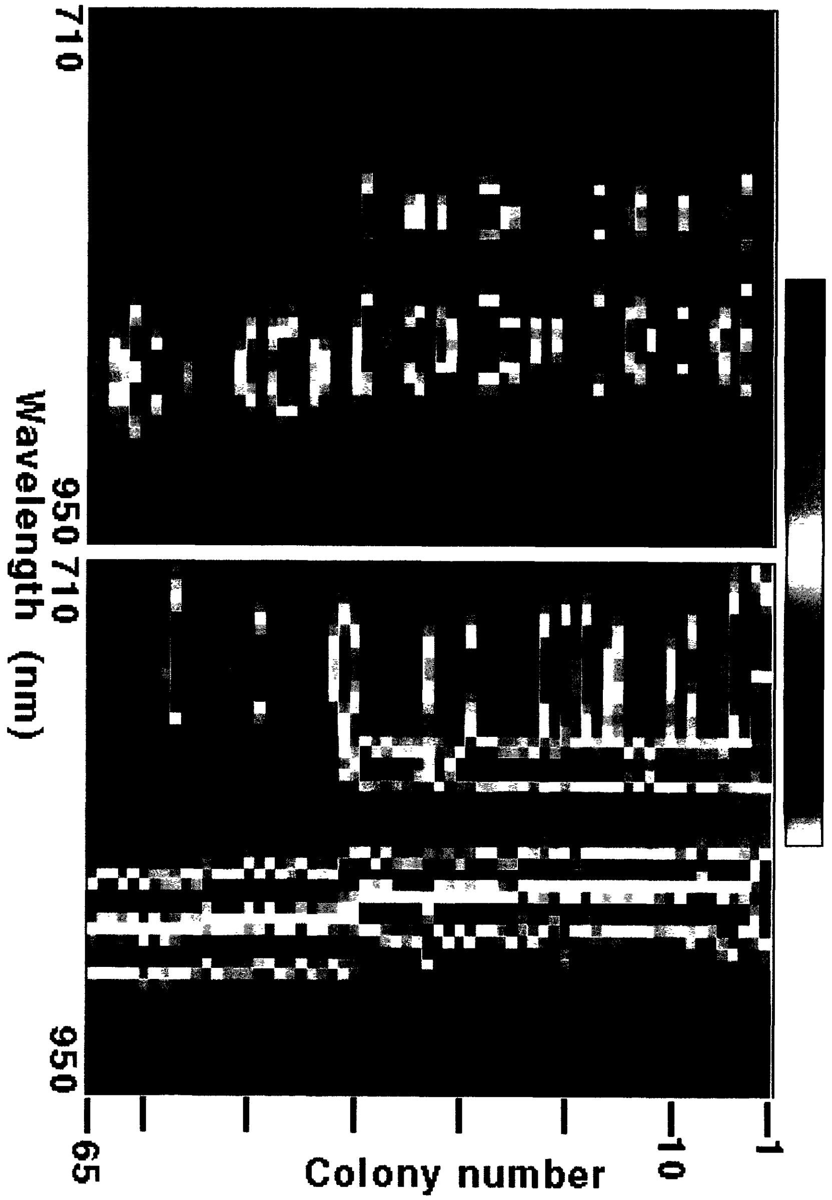
For both the LHII and RC sequences it was possible to formulate simple rules for determining positives based on visual inspection of the sequence databases. In the case of LHII one can generalize that if there is not a Thr in the +7 position, the sequence leads to a null. Additionally, a rule can be formulated to separate the two Bchl binding phenotypes: if the sequence is

	sequence							row
	-7	-4	-3	0	3	4	7	
WT	G	A	L	H	S	A	T	27,28,39
S1	L	G	L	H	A	F	T	52
S2	L	G	L	H	A	A	T	62
S3	I	V	T	H	A	A	T	64
S4	V	G	V	H	A	G	T	65
S5	L	A	A	H	A	G	T	61
S6	M	A	L	H	A	A	T	47
S7	L	A	V	H	S	A	T	59
S8	V	A	V	H	S	W	T	53
S9	A	A	L	H	S	T	N	63
S10	L	A	V	H	S	A	T	56
S11	L	V	L	H	S	A	T	48
S12	V	A	L	H	S	G	T	54
S13	V	G	A	H	S	A	T	58
S14	V	A	L	H	A	W	T	51
S15	I	A	T	H	A	T	T	49
S16	M	A	L	H	S	C	T	45
S17	L	A	A	H	A	G	T	55
S18	V	A	A	H	A	Y	T	44
S19	L	V	I	H	A	G	T	57
S20	I	V	T	H	S	A	T	42
S21	V	G	I	H	S	A	T	60
S22	L	A	V	H	A	M	T	46
S23	M	G	V	H	A	M	T	43

	sequence							row
	-7	-4	-3	0	3	4	7	
WT	G	A	L	H	S	A	T	27,28,39
P1	S	G	I	H	A	G	T	33
P2	C	G	I	H	A	G	T	40
P3	A	G	V	H	S	A	T	36
P4	T	A	I	H	S	M	T	34
P5	A	A	A	H	A	W	T	12
P6	G	A	L	H	A	G	T	3
P7	G	A	T	H	S	F	T	9
P8	G	G	I	H	S	Y	T	35
P9	A	V	L	H	S	A	T	23
P10	G	A	V	H	A	K	T	13
P11	G	G	A	H	A	V	T	15
P12	G	G	A	H	V	A	T	24
P13	G	A	V	H	A	F	S	18
P14	G	A	L	H	S	I	T	32
P15	G	G	I	H	S	Y	T	38
P16	A	A	V	H	S	G	T	30
P17	G	G	L	H	A	V	T	10
P18	A	A	A	H	A	W	T	14
P19	C	G	L	H	A	A	T	50
P20	T	A	A	H	S	A	T	41
P21	G	V	I	H	A	G	T	8
P22	G	G	L	H	A	V	T	22
P23	T	A	V	H	A	Y	T	31
P24	G	A	A	H	S	I	T	2
P25	G	A	A	H	S	A	T	26
P26	A	A	I	H	A	A	T	17
P27	T	A	T	H	A	V	T	37
P28	G	G	I	H	A	V	T	16
P29	A	A	I	H	V	A	S	6
P30	T	A	T	H	V	A	T	5
P31	G	A	A	H	A	K	T	21
P33	F	A	V	H	V	A	T	7
P34	S	G	T	H	A	M	T	1
P35	A	G	A	H	S	A	T	25
P36	G	A	A	H	S	F	S	19
P37	G	G	A	H	A	W	T	11
P38	W	G	V	H	A	F	T	29
P39	G	V	L	H	S	G	T	20
P40	A	V	I	H	A	M	T	4

Table 5. Amino acid sequences of the LHII mutants experimentally classified as positives. The sequence positions are relative to the Bchl-binding His (0). The 'P' mutants show the pLH2 phenotype, while the 'S' mutants are classified as pLH1. The row number indicates the position of the spectra in Figure 9. Although there are some duplications at the amino acid level, each mutant had a unique nucleotide sequence.

Figure 9. Color contour maps generated by DIS showing the spectra of sequenced positives. 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, and is encoded by pseudocolor. The color bar shows the range of optical density from low (black) to high (white). The left panel is in 'absolute mode' (highest value in the image mapped to white, lowest value in the image mapped to black) and shows the range of expression levels. The expression level was dependent on growth (time and temperature) conditions. The right panel is in full deflection mode (highest value for each spectrum mapped to white, lowest value in each spectrum mapped to black) and allows a comparison of spectral peaks. The row number can be used to find the corresponding amino acid sequence in Table 5. Raw spectral data was sorted first according to maximum absorption while in absolute mode, then according to wavelength of maximum absorbance in full deflection mode.


710
950 710
950
Wavelength (nm)
65
10
1
Colony number

classified as positive, amino acids with molar volumes larger than Thr (116.1 Å^3) in the -7 position give a pLH1 phenotype, while amino acids of smaller or equal molar volume yield pLH2 type spectra. In the case of the RC sequences, the criterion for classifying a mutant as positive requires L154 to be a Leu; otherwise the mutant is classified as null.

The percent of incorrect phenotypic classification can be evaluated for these simple DAs. The average number of null sequences which the DA wrongly labels as positives can be calculated based on the frequency of critical amino acids in the nucleotide mixture used to construct the cassette. The number of nulls considered is set equal to the number of unique positive amino acid sequences. The experimentally determined positives that the DA would mistake for nulls are determined by counting from the compiled sequences.

4.3.5 Randomly generated nulls for ANN evaluations

Since only a few nulls were sequenced, we generated sequences randomly obeying the 'doping' scheme of the libraries. The throughput of the library (6% and 0.0002% for the LHII and RC libraries, respectively) gives us an estimate on how many of these "pseudo-nulls" are false negatives. Adding the few true nulls did not change significantly the outcome of our ANN simulations. We always averaged several experiments with different random sequences.

4.3.6 Neural network construction and training

We employed standard backpropagation neural networks (Hertz *et al.,* 1991) consisting of an input layer, a varying number of hidden units, and a binary output layer. In most experiments, the input layer consisted of 12 units for the LHII mutants (2 features for all 6 sequence positions, excluding the His), and 18 units for the RC mutants (9 sites, each with 2 features). We considered two kinds of features: (1) molar volume and hydropathy, and (2) artificial letter encoding. For the latter, each one-letter amino acid abbreviation was viewed as a binary number, split into high and low significant bits, and scaled to lie in the interval (0,1). The physical property values were normalized to lie in the same interval.

The class labels ("0" for nulls, and "1" for positives) were used as target values. To train the network, we used the backpropagation of errors method, employing a conjugate gradient descent with a sophisticated line search, implemented as "Rudi Mathon's conjugate gradient with Ray Watrous' line search" by the back propagation module of the Xerion Neural Network Simulator. The sum of all incoming activations was transferred to the units of the next layer using the logistic (standard), linear (perceptron), or exponential functions. Target values and calculated output activations of the network were considered to agree if they deviated by less than 0.1. One half the square of any excess difference was added to the overall error to be minimized.

4.3.7 Evaluation of neural network performance

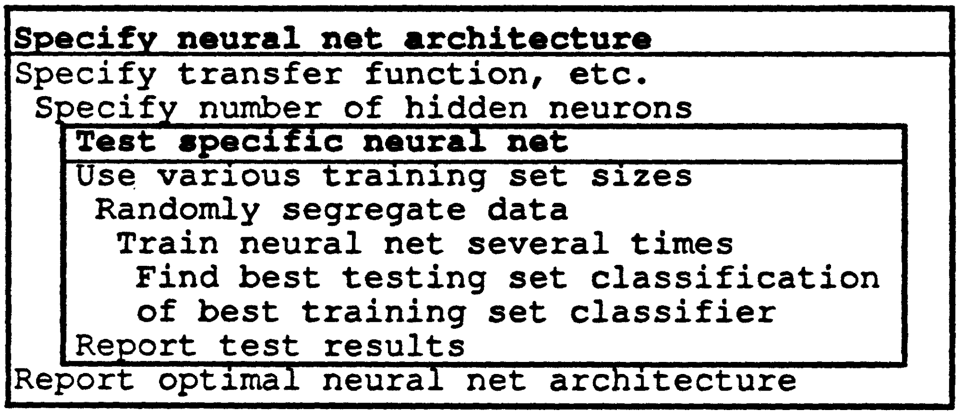
We were not interested in constructing any specific ANN, but in investigating the general usefulness of ANNs for phenotypic estimation of combinatorial libraries. In this case crossvalidation is a valid technique for accessing the quality of various neural network architectures. In most cases otherwise, we did 16 partitions of the data into randomly ordered training and testing sets. For each division of data we trained the network 8 times using different random starting weights, and recorded the testing set classification of the best training set classifier. Often, *this* classifier has overlearned and does not generalize well. However, it would not be fair to record the best or average testing set classification accuracy. Figure 10 shows a strategy for constructing, and testing networks.

4.4 Results

4.4.1 LH versus nulls

In the separation of LH positives from nulls, the DA required positive mutants to have a Thr in the +7 position. Four of the sequenced positives would be wrongly classified according to this rule. Further, 19 nulls would be misclassified as positives (this was calculated based on 57 null sequences, so that there is an equal number of unique positive and null amino acid sequences, and the fact that the original doping scheme coded for 33% Thr). This leads to an overall rate of 23 wrong classifications per 114 mutants, or 80% correct categorization.

Figure 10. Exploration and testing of neural networks. We recommend such an extensive evaluation to minimize assumptions on the model, and to maximize confidence in the network's accuracy. In the inner box entitled "Test specific neural net", every indentation specifies a repetition over all possible values specified one block above, similar to indentation in a C program. The outer box gives an idea of different neural network architectures that can be explored; one may also change the error measure, the training algorithm, etc.



```
Specify neural net architecture
Specify transfer function, etc.
 Specify number of hidden neurons
   Test specific neural net
   Use various training set sizes
    Randomly segregate data
     Train neural net several times
      Find best testing set classification
      of best training set classifier
   Report test results
Report optimal neural net architecture
```

As shown in Figure 11A, a neural network with one hidden neuron and an exponential transfer function scored, about 86% correct classification. We assume that using an exponential transfer function works as a guard against overfitting since the target values are "0" and "1". In this case, large weights memorizing particular aspects of the training set are discouraged because their influence on the activations would be amplified exponentially and the small target values could no longer be met. Indeed, the corresponding network with a logistic transfer function performs less well (82%) within a 95% confidence interval. If the network has more weights than there are training cases, memorization becomes dominating, explaining the very bad performance for small training set sizes. The same effect has been observed for networks with more hidden neurons. Using the artificial letter encoding of the amino acids, results in 84% accuracy regardless of the transfer function. This result indicates that using physically realistic features aids the learning process, but may lead to unwanted memorization.

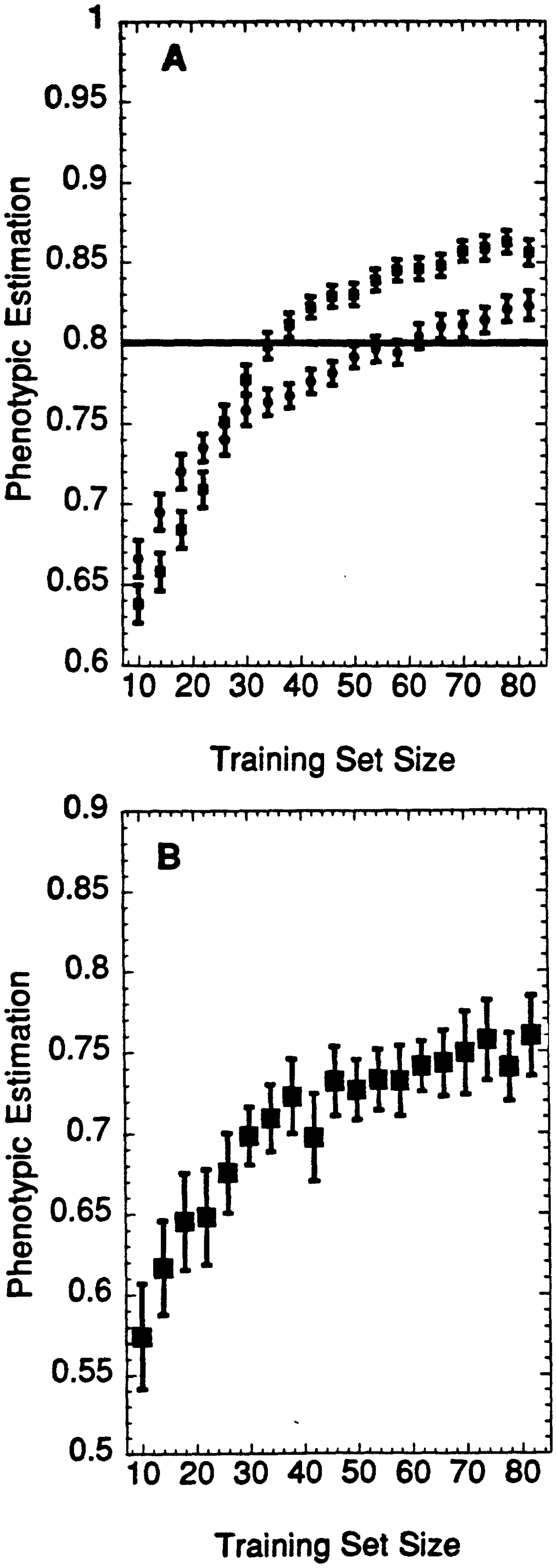
Slightly suboptimal performance (85%) has been observed for perceptron-like networks with no hidden neuron and a linear transfer function. This could suggest the presence of some nonlinear interrelationships between the amino acids.

Figure 11B shows the network accuracy if it is not presented the amino acid position on which the DA is based (+7 position). The 74% accuracy proves that the determination of phenotype is influenced by the remaining positions. Using a direct letter encoding in this case, we observe an accuracy of 73%. Presenting only molar volume or only hydropathy values of the +7 position, we observe 79% and 85% accuracy, respectively. In these cases, the network's ability to distinguish between Thr and residues with similar physical properties is diminished. Indeed, Thr is more readily distinguishable from the other amino acids doped at this position (i.e. Lys, Asp, Arg, Ser) with respect to its hydropathy value.

4.4.2 Three way separation of LH data

Further separation of the positives into pLH1 and pLH2 was based on the rule that if the mutant is positive and has an amino acid of large molar volume at the -7 position, it will be pLH1 otherwise it will be pLH2. Three positives will be

Figure 11. Neural network performance, for the LH two class separation between positives and nulls, as a function of training set size and transfer function. Panel A shows the performance of networks with one hidden neuron, and both logistic (filled circles) and exponential (filled squares) transfer functions. The horizontal line at .8 indicates 80% correct classification by a simple DA approach. Panel B shows the accuracy of networks with one hidden neuron and an exponential transfer function that was confronted with data missing the most important residue as judged by the DA (+7 position). All error bars indicate the standard error of the mean for a 95% confidence level.


A
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
Phenotypic Estimation
10 20 30 40 50 60 70 80
Training Set Size
B
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Phenotypic Estimation
10 20 30 40 50 60 70 80
Training Set Size

misclassified as to their phenotype leading to 26/114 (23%) overall wrong categorization.

Figure 12 indicates a very slight advantage for neural networks (20% misclassification), obtained by the network architecture less vulnerable to memorization. The network with no hidden neuron performs significantly better than the network with one hidden neuron.

4.4.3 RC versus nulls

Separation of functional versus non functional RC mutants was based on the premise that Leu must be in position L154 for the protein to be positive. There would be 6 errors out of the 25 sequenced functional mutants, and an additional 2 out of 25 nulls misclassified as positives because of the frequency of Leu in an NN(G/C) dope. This results in 8 errors out of 50 mutants and therefore 84% accuracy for this simple DA approach.

Networks do not seem to have a problem with memorization for the RC database. Figure 13A shows that the network with a logistic transfer function performs better than the one with an exponential transfer function; we achieve 91% versus 84% accuracy (logistic versus exponential transfer function).

ANN accuracy remains slightly above 90% for networks with up to 40 hidden neurons. No enhanced performance was observed if the number of restarts with new random weights is increased from 8 to 128 or decreased from 8 to 2. However, adding a cost term (i.e. 10% of the sum of all weights) to the error, produces predictions of up to 93%. Networks with 20 hidden neurons obtained almost 94% accuracy if an appropriate weight cost was selected.

Network performance drops to 88% for a perceptron-like architecture, indicating the existence of possible non-linerar interrelationships between amino acids. When artificial letter encoding is used, accuracy drops to 70-78%. Learning is diminished by using physically meaningless features.

If the network, is not given information about residue L154 (the position at which the DA is based), performance drops to 87% (Figure 13B). This indicates the presence of strong determining factors in the remaining residues. Dropping other sites, we observe accuracies between 88 and 91%. In particular, leaving out residue L146 does not impair accuracy. At this site we observe a range of amino acids widely scattered in molar volume and hydropathy space contributing no valuable information to the decision process.

Figure 12. Neural network performance for the LH three class separation between pLH1, pLH2, and null phenotypes, as a function of training set size and number of hidden neurons. An exponential transfer function was used in both cases. Open circles represent 0 hidden neurons; filled circles represent 1 hidden neuron. All error bars indicate the standard error of the mean for a 95% confidence level. The horizontal line at .77 indicates 77% correct classification by a simple DA approach.

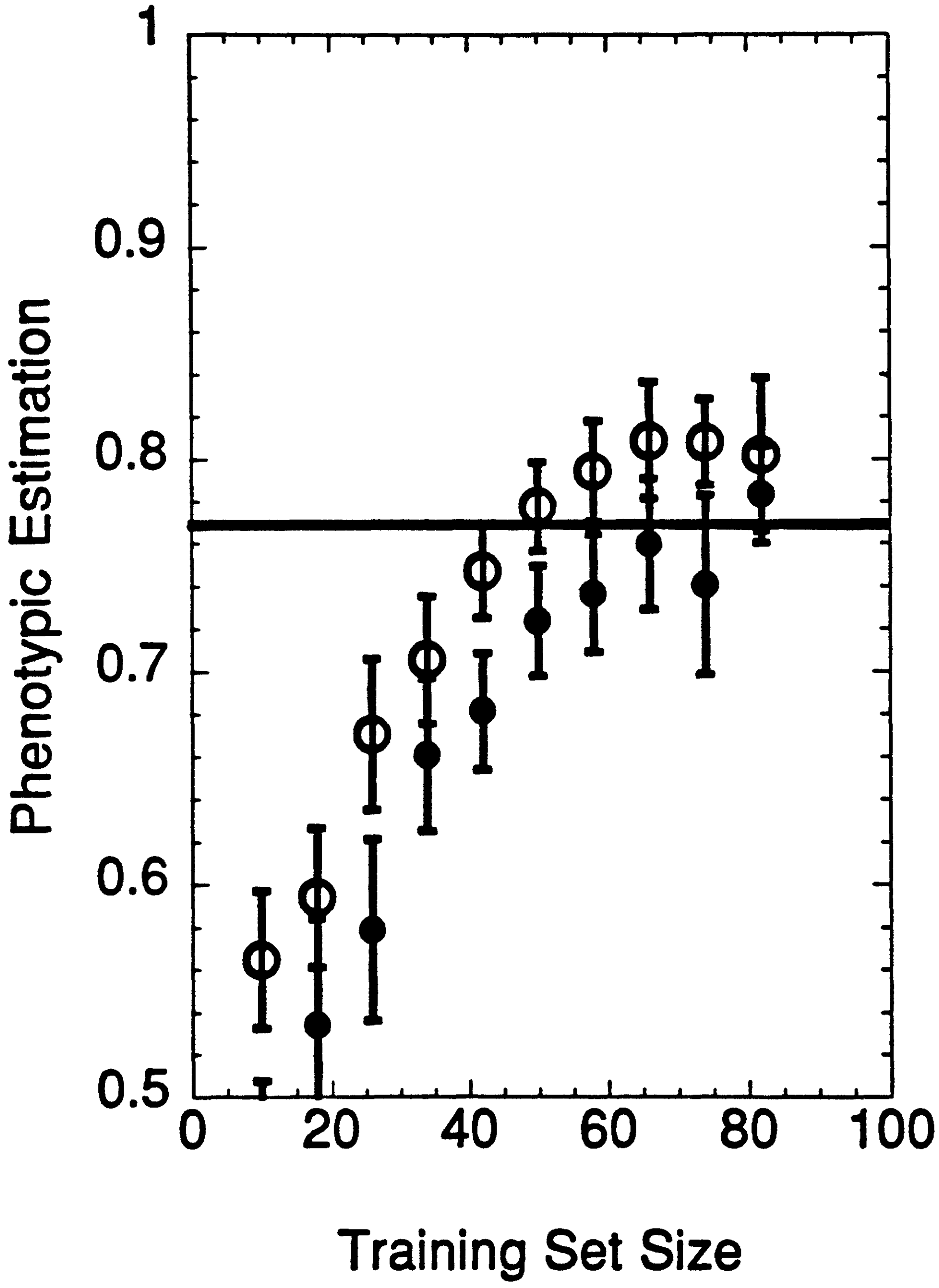
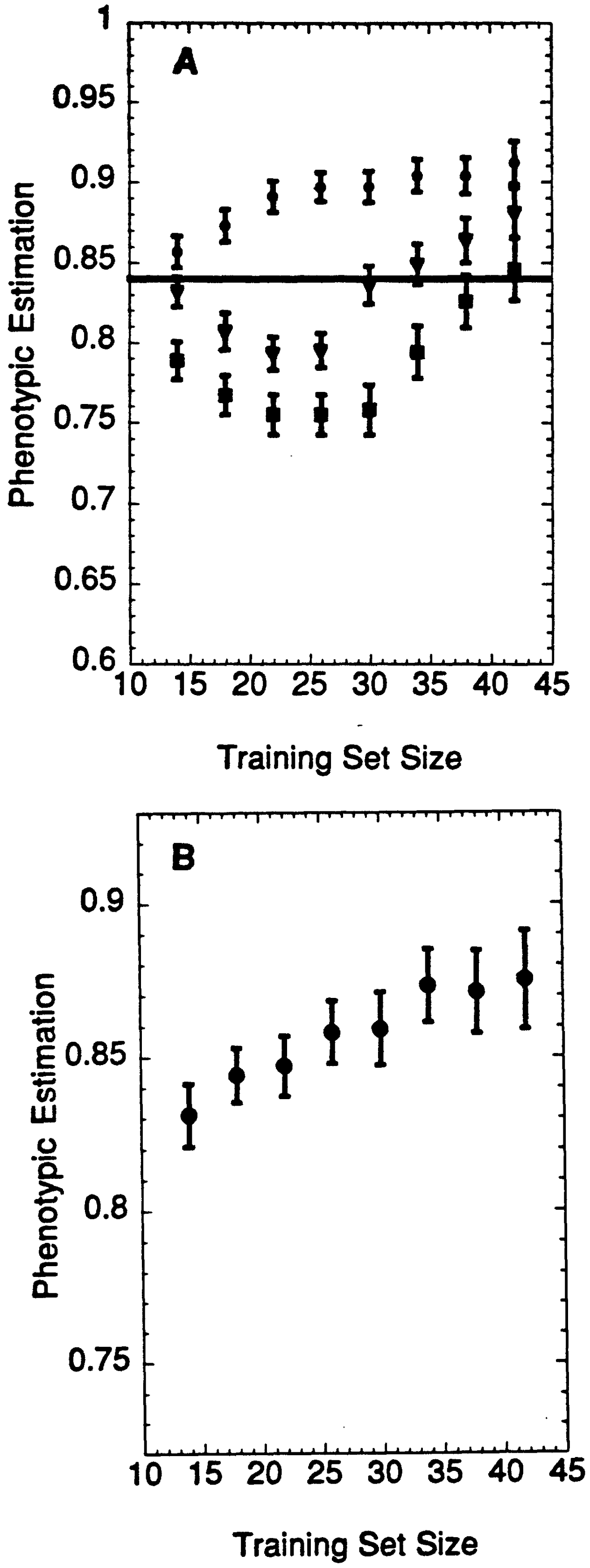

1
0.9
0.8
0.7
0.6
0.5
Phenotypic Estimation
0
20
40
60
80
100
Training Set Size

Figure 13. Neural network performance, for the RC two class separation between functional and nulls, as a function of training set size and transfer function. Panel A shows network performance with one hidden neuron, and both logistic (filled circles) and exponential (filled squares) transfer functions. The performance of a perceptron network (no hidden neurons, linear transfer function) is shown with filled triangles. The horizontal line at .84 indicates the 84% accuracy achieved by the DA approach. Panel B shows the accuracy of networks with one hidden neuron and a logistic transfer function that was confronted with data missing the residue (L154) on which the DA is based. All error bars indicate the standard error of the mean for a 95% confidence level.


A
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
Phenotypic Estimation
10 15 20 25 30 35 40 45
Training Set Size
B
0.9
0.85
0.8
0.75
Phenotypic Estimation
10 15 20 25 30 35 40 45
Training Set Size

4.5 Discussion

There are many types of combinatorial biological and chemical experiments which should be amenable to analysis of sequence-phenotype data by simple DAs and ANNs. Phage display technology can be used to generate libraries of up to 10^9 different proteins (Smith, 1985; Hoogenbaum *et al.*, 1991), that can be screened by affinity for an arbitrary compound. Libraries of synthetic random peptides (Lam *et al.*, 1991; Houghten *et al.*, 1991) have also been constructed and screened by binding to acceptors. As combinatorial chemistry becomes more feasible, the databases will become even more complex with a larger repertoire of building blocks which will include thousands of organic chemicals.

We chose to use pigment binding proteins in our experiments because DIS can be used to rapidly assay phenotypes off petri dishes. Genomic RC and LHII deletion backgrounds (Youvan *et al.*, 1985) and plasmids that facilitate CCM have been developed for both the LHII (Goldman & Youvan, 1992) and RC (Robles & Youvan, 1993) systems.

The selection of our machine learning method was preceded by a review of comparisons between various symbolic and connectionist paradigms. On the theoretical side, Pao (1989) points out that the essence of pattern recognition is the replacement of an "opaque" mapping from examples to attributes by a similar but "transparent" computer-encoded mapping. In our case, we try to map amino acid sequences to estimates of functionality. Neural networks can learn an exceptionally rich class of mappings (Cybenko, 1989; Hornik *et al.*, 1989; Hornik, 1991). Exploration of different architectures minimizes our assumptions on the underlying opaque mapping.

Many publications (for reviews, see Presnell & Cohen, 1993; Hirst & Sternberg, 1992) report encouraging results using ANNs in protein research. ANNs were observed to have an advantage over other machine learning methods in two cases (Shavlik *et al*, 1991): 1) small amounts of training data, and 2) numerical training data. Our experimental data meet both these criteria.

4.5.1 When to use Neural Networks

In cases where there is a high amount of non-linear site-by-site interactions ANNs can be expected to perform better than DAs. This could include examples where long stretches of amino acids were mutagenized, or in the region of an enzyme active site. ANNs might also be expected to give better predictions than simple DAs when differentiating between more subtle phenotypic differences than the "on" and "off" type problems we examined.

4.5.2 Factors affecting decision algorithms

For both the LH and RC experiments there is some misclassification inherent to the experiment. Up to 10% of the mutants may vary in phenotype depending on their growth time and conditions. The 90% categorization efficiency achieved by the ANN may be close to the highest level that can be expected taking into consideration the inaccuracy of the the experimental data. For each library some of the experimentally determined positives will be wrongly classified by the DA. The number of these misclassified positives could vary if a second set of positives were isolated for each experiment.

It is possible that DAs could change based on the context of the mutagenesis. A sequence position that is decisive when non continuous amino acid residues are mutagenized might be more flexible when a contiguous stretch is mutagenized. Preliminary results show that the molar volume of the -7 position might not be responsible for pLH1 versus pLH2 phenotype when six amino acids (-10, -9, -8, -7, -6, -5) are randomized simultaneously (S. Delagrave personal communication). However, preliminary sequences in a library where 17 amino acids (-10 through +6) were mutagenized with phylogenetic target sets (library from chapter 3, unpublished results) showed that the phenotypes of the mutants followed the -7 rule.

4.5.3 Decision algorithms and phylogeny

The LH deduced DAs do not recapitulate the phylogenetic data, but rather appear to be specific for *Rb. capsulatus* LHII. The rules fail when applied to the sequences of homologous light harvesting antennae compiled by Zuber, 1990. The +7 sequence position (the determiner between nulls and LH in the library) is a conserved Arg in the β subunits of core antennae (LH I). Among the peripheral antennae (LHII), there is a division among Thr:Asn:Ser of 7:4:4, with a single sequence having a Lys at the +7 position. In the TSM LHII positives the molar volume of the -7 position determines the type of spectral phenotype.

However, many of the LHII type antennae from different species (69%) have amino acids with larger molar volume (Val, Leu, Phe) in their -7 position. Although most of the LHI type antennae also have amino acids with large molar volume in their -7 position, there are a few exceptions.

4.5.3 Success of primitive decision algorithm

Originally DAs formulated based on expert rules were used in computer simulations (Arkin & Youvan, 1992; Youvan *et al.* 1992); we postulated that they might be too simple to be applicable to experimental data. However, LHII antennae and RCs were found to have critical amino acid positions that are basic phenotypic determiners (positive vs. nulls). The success of the simple rules suggests that as a first approximation the correlation of sequence and phenotype can be examined on a site-by-site basis. This implies that CCM can be based on evaluation per site, and justifies the use of phylogenetic target sets in target set mutagenesis (TSM) as well as the construction of target sets per position from positives using random mutagenesis in the recursive ensemble mutagenesis (REM) (Delagrave *et al.* 1993). One should be able to randomize arbitrary regions of these proteins and then combine the sequence information to formulate target sets for larger cassettes (exponential ensemble mutagenesis (EEM) (Delagrave & Youvan, 1993)) without excess concern about selecting the regions for randomization.

4.6 Acknowledgements

Georg Fuellen performed the neural network analysis as part of the work for his masters thesis. Our neural network simulation scripts were based on the Xerion neural network simulator developed at the University of Toronto by Drew van Camp, Tony Plate and Geoffrey Hinton.

Chapter 5: Conclusions and future research

5.1 Conclusions

My thesis research has concentrated on the construction and analysis of phylogenetically based TSM libraries of *Rb. capsulatus* LHII antenna. LHII is an ideal system for implementing mutagenesis strategies; it is synthesized in large quantities, is stable, and has an intense characteristic absorption spectrum. The effects of engineered mutations can be easily assayed by changes in absorption peaks. Peak heights, and the ratio of the peak heights can be examined in addition to wavelength shifts allowing for the exploration of more than just on/off expression. I designed plasmids for performing cassette mutagenesis in several regions of the structural genes for LHII. The LHII genetic system in conjunction with DIS provides a perfect combination to evaluate and test novel mutagenesis schemes.

The techniques which are explored using the LHII model should be applicable to any protein. In other protein systems, binding or enzymatic activity could be used as a screen for the desired phenotype. Combinatorial mutagenesis methods have much potential for intelligent protein design.

The work performed for this thesis has important implications to the photosynthesis community as well as the protein engineering field. The mutants that were generated in these experiments have the potential to enhance the understanding of pigment binding and energy transfer in LH proteins. Representative mutants from the different libraries with different spectral characteristics should be analyzed for pigment content and protein composition. It is possible that the unstable mutants may have lost the γ subunit. The mutants should also be screened for their ability to be crystallized. Crystals of WT LHII have been obtained in our lab (C. Goddard pers. comm.) but did not diffract well enough for structure determination. Perhaps one of the mutants may have a change that affects LH aggregation and would lead to a crystal that would diffract to high resolution. The mutants constructed in these TSM experiments should be screened for their ability to perform energy transfer. Mutants with WT-like spectra, but diminished energy transfer ability might be interesting to characterize. Future CCM experiments on photosynthetic proteins may uncover a mutant that performs wrong way electron transfer or a LH antenna that is capable of charge transfer.

5.2 Recursive ensemble mutagenesis

The CCM experiments I reported on in chapters 2 and 3 utilized sequences of homologous LH β subunits to formulate target sets. However, phylogenetic data are not always available. Recursive ensemble mutagenesis (REM) (Arkin & Youvan, 1992a; Youvan *et al.*, 1992) uses sequence information acquired from random CCM to construct target sets for subsequent cassettes (Figure 14). LHII was used in a first experimental implementation of this technique. Six amino acid positions on the carboxyl-terminus of the β subunit of LHII were mutagenized using REM (Delagrave, Goldman & Youvan, 1993). Sites were randomized using NN(G/C) codons; mutants were pre-screened by fluorescence before evaluation by DIS. Only one out of 10,000 mutants screened had the desired fluorescence characteristics. Five positives recombinants were isolated and sequenced. This sequence data showed that the mutants were not merely trivial variations of the WT sequence. The sequences did not recapitulate the known phylogeny; one mutant showed an inversion in a completely conserved sequence motif that resulted in a 10 nm blue shift of the dimer peak. The next iteration of REM cassettes were designed using the P_G algorithm. REM returned a thirty-fold increase in the number of positive mutants over random CCM. Twelve mutants were sequenced and all were found to have unique peptide sequences that differed from both WT and the set of mutants used to generate the first target sets. This is a useful method when there is not an extensive phylogenetic data base. When mutagenizing larger stretches of amino acids, short segments can be randomized independently and these sequences can then be combined. This defines exponential ensemble mutagenesis (EEM) (Delagrave & Youvan, 1993).

5.3 Combinatorial mutagenesis of photosynthetic proteins

At present, combinatorial cassette mutagenesis of LH antennae and the bacterial reaction center has only been accomplished in *Rb. capsulatus*. Although there is currently no crystal structure for the *Rb. capsulatus* RC, a chemoheterotrophic *Rp. viridis* genetic system is being developed in the Bylina

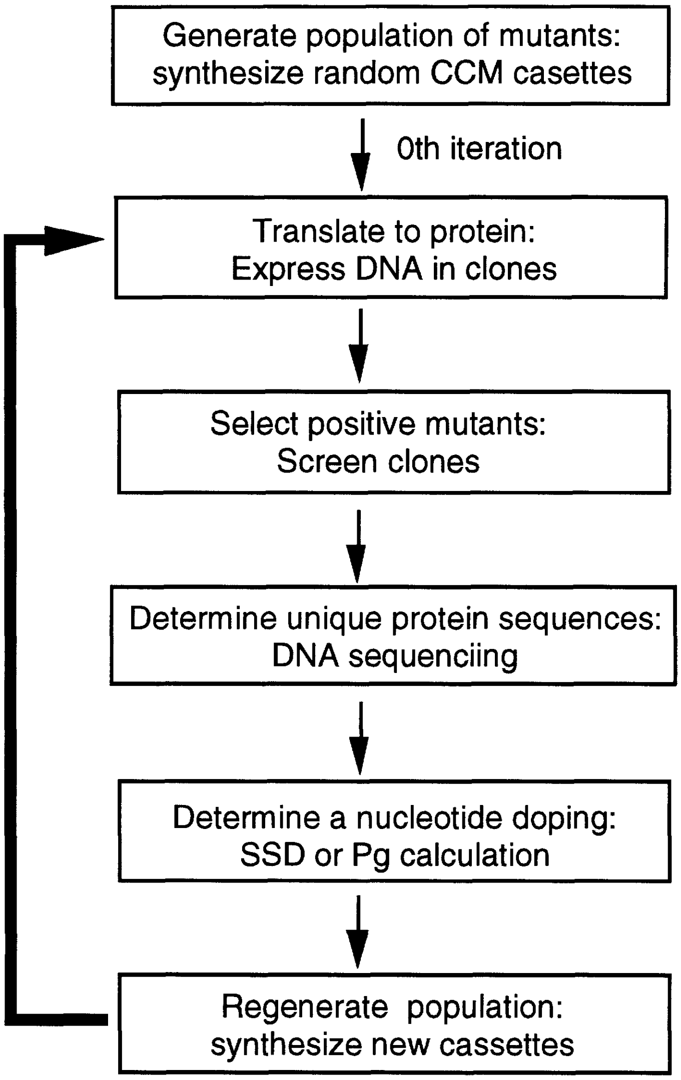

Generate population of mutants: synthesize random CCM casettes
0th iteration
Translate to protein: Express DNA in clones
Select positive mutants: Screen clones
Determine unique protein sequences: DNA sequenciing
Determine a nucleotide doping: SSD or Pg calculation
Regenerate population: synthesize new cassettes

Figure 14. Flowchart for the implementation of recursive ensemble mutagenesis. On the zeroth iteration, a random CCM library is expressed and screened. Two or more positives (clones that meet the phenotypic criteria) are selected and sequenced. The protein sequence of the positive mutants are used to define target sets for the construction of subsequent cassettes, and either the PG or SSD algorithm can be used in the formulation of the nucleotide dopes.

laboratory (E. J. Bylina, pers. comm.). This could lead to a potentially useful combination of experiments in these two species: sequence space can be screened by combinatorial mutagenesis in *Rb. capsulatus*, and when single interesting mutants are found they can be re-constructed in the *Rp. viridis* strain for crystallography.

Future goals include using cassette mutagenesis to engineer a RC mutant that performs wrong-way electron transfer. Instead of randomly mutagenizing sections of the RC, regions of the protein in contact with the pigments could be mutagenized using TSM. Phylogenetically based target sets are being constructed using the amino acid residues at homologous sites in both the L and M subunits, which are thought to be the product of a gene duplication (Youvan & Marrs, 1984).

A second goal is to alter LH antennae to perform charge separation. Combinatorial cassette mutagenesis could be used to change the protein environment (i.e., the electronic environment) around the pigments. An antenna that was modified to perform charge separation would allow biophysicists to study electron transfer in a much simpler protein system than the RC.

LHII is an excellent model system in which to attempt to establish a correlation between sequence and phenotype. Computer programs can be used to correlate sequence data with phenotypic information to generate phenotypic estimators. This obliterates the need for a solution to the protein folding problem, and facilitates various engineering projects on pigment-protein complexes.

Mutant LH and RC complexes have helped biophysicists better understand the events in photosynthesis. In the future, increasingly complex mutagenesis techniques should contribute still more to our understanding of the molecular mechanisms of photosynthesis and more general problems in protein folding and design.

Appendices

A. Materials and methods

1. Standard techniques of molecular biology were performed as detailed in Sambrook *et al.*, 1989. Most enzymes were purchased from New England Biolabs, with the exception of T4 DNA Ligase, which was purchased from GIBCO BRL.

2. Site directed mutagenesis (Kunkel *et al.*, 1987, Zoller & Smith, 1984) was carried out using the muta-gene M13 *in vitro* mutagenesis kits from BioRad. Non-mutagenic restriction sites were engineered in several of the positions shown in Figure 15 using the following oligonucleotides.

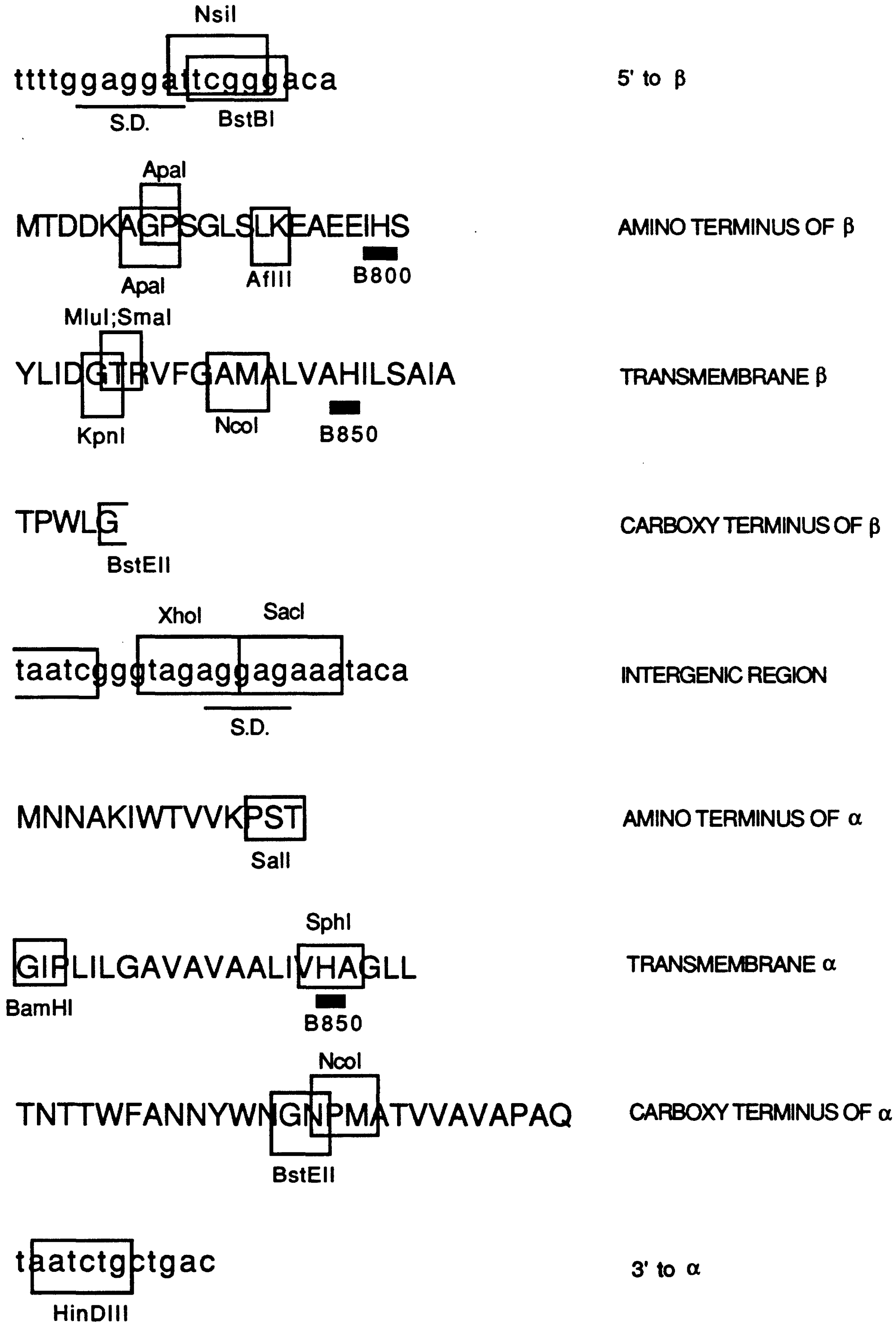
Hind III oligo: Sense oligonucleotide to engineer a Hind III site for convenient shuttling between pU4 derivatives and M13.
5' CAGTAAGCTTCTGACCTT 3'

Kpn I oligo: Antisense oligonucleotide to engineer the Kpn I site for cassette mutagenesis.
5' CGCCCCGAACACACGGGTACCATCGA 3'

Xho I oligo: Antisense oligonucleotide to engineer the Xho I site for cassette mutagenesis.
5' CATTGTATTTCTCCTCGAGCCGATTAC 3'

Nsi I oligo: Antisense oligonucleotide to engineer Nsi I site for cassette mutagenesis:
5' CATTGTCATGCATCCTCCAAAC 3'

Figure 15. Location of potentially engineered restriction sites in the vicinity of the LHII β and α subunits. The Hind III site down stream of the α subunit (3' to α) was engineered to facilitate shuttling a 1 kb fragment containing the genes for α and β into M13. The Kpn I site in the transmembrane β region, the Xho I site in the inter-subunit region, and the Nsi I site 5' to β were engineered to enable CCM of the regions of the β subunit.


NsiI
ttttggaggattcgggaca
S.D.
BstBI
5' to β
ApaI
MTDDKAGPSGLSLKEAEEIHS
ApaI
AflII
B800
AMINO TERMINUS OF β
MluI;SmaI
YLIDGTRVFGAMALVAHILSAIA
KpnI
NcoI
B850
TRANSMEMBRANE β
TPWLG
BstEII
CARBOXY TERMINUS OF β
XhoI
SacI
taatcggtagaggagaaataca
S.D.
INTERGENIC REGION
MNNAKIWTVVKPST
SalI
AMINO TERMINUS OF α
SphI
GIPLILGAVAVAALIVHAGLL
BamHI
B850
TRANSMEMBRANE α
NcoI
TNTTWFANNYWNGNPMATVVAVAPAQ
BstEII
CARBOXY TERMINUS OF α
taatctgctgac
HinDIII
3' to α

3. DNA sequencing was performed using the sequenase version 2.0 system from United States Biochemical. The following sequencing primers were used to sequence the α and β subunit genes of mutants from the combinatorial cassette libraries.

Bam HI primer: Antisense oligonucleotide used to sequence the β dimer region. This primer could also be used to engineer a Bam HI site in the α subunit
5' GATCAGCGGGATCCCGGTCGA 3'

Xho I oligo (sequence shown in section 2) was used to sequence the β monomer region.

The Universal -40 sequencing primer was used to sequence the α subunit .

3. Cassette mutagenesis was carried out using the scheme outlined in Figure 16. The sequences of the cassettes and PCR primers are listed below. In these sequences: N = 25% each A,T,G,C. parenthesis enclose various combinations of nucleotides incorporated at a single position.

PCR primers for β dimer cassettes:
5' GCTACCTGATCGATGGTACC 3'
5' TCATTGTATTTCTCCTCGAG 3'

7 site phylogenetically based cassette:
5' GC TAC CTG ATC GAT GGT ACC CGT GTG TTC (AGT)(CGT)(GC) GCG ATG G(CGT)C (ACG)(CT)C GTT GCG CAC ATC CTC (GT)(CT)C (AGT)N(GC) ATC GCC A(ACG)(CG) CCG TGG CTC GGG TAA TCG GCT CGA GGA GAA ATA CAA TGA 3'

6 site random cassette:
5' GC TAC CTG ATC GAT GGT ACC CGT GTG TTC NN(GC) GCG ATG NN(CG) NN(CG) GTT GCG CAC ATC CTC NN(GC) NN(GC) ATC GCC NN(CG) CCG TGG CTC GGG TAA TCG GCT CGA GGA GAA ATA CAA TGA 3'

7 site random cassette:
5' GC TAC CTG ATC GAT GGT ACC CGT GTG TTC NN(GC) GCG ATG NN(CG) NN(CG) GTT GCG NN(GC) ATC CTC NN(GC) NN(GC) ATC GCC NN(CG) CCG TGG CTC GGG TAA TCG GCT CGA GGA GAA ATA CAA TGA 3'

17 site SSD formulated cassette:
5' GC TAC CTG ATC GAT GGT ACC A(GC)G G(TC)(GC) TTC (GAC)(GT)(GC) G(GTC)(GC) G(TC)(GC) GC(CG) (GAC)(TC)C (GAC)TC GC(GC) CAC NTC (TC)T(GC) NN(GC) ((GT)(TC)(GC) (GT)N(GC) (GT)(GC)G ACG CCG TGG CTC GGG TAA TCG GCT CGA GGA GAA ATA CAA TGA 3'

PCR primers for β monomer cassettes:
5' TCCCAGTTTTGGAGGATGCAT 3'
5' CGCCCCGAACACACGGGTACC 3'

17 site P_G formulated cassette:
T CCC AGT TTT GGA GGA TGC ATG ACA ATG ACT GAC GAT AAA GCT GGG CCG (GA)(GAC)(GC) (GAC)NC (GAC)TC (AT)(CT)C N(ACT)(GC) (GAC)(GAC)(GC) (GC)AG (GT)(GAC)C N(GA)G (GA)A(GC) N(CT)C CAC (GA)(GAC)(GC) N(AT)(GC) N(AT)C (GA)(ACT)C N(ACT)(GC) GGT ACC CGT GTG TTC GGG GCG

17 site SSD formulated cassette:
T CCC AGT TTT GGA GGA TGC ATG ACA ATG ACT GAC GAT AAA GCT GGG CCG (AT)C(GC) GG(GC) (TC)TG A(GC)(GC) (GC)(ACT)(GC) (GAC)(AC)G (GC)AG (GT)C(GC) (GA)AG GAG (ACT)TC CAC (GAT)(AC)G N(GT)(GC) (GAC)TC (GA)(AT)(GC) (GA)NC GGT ACC CGT GTG TTC GGG GCG

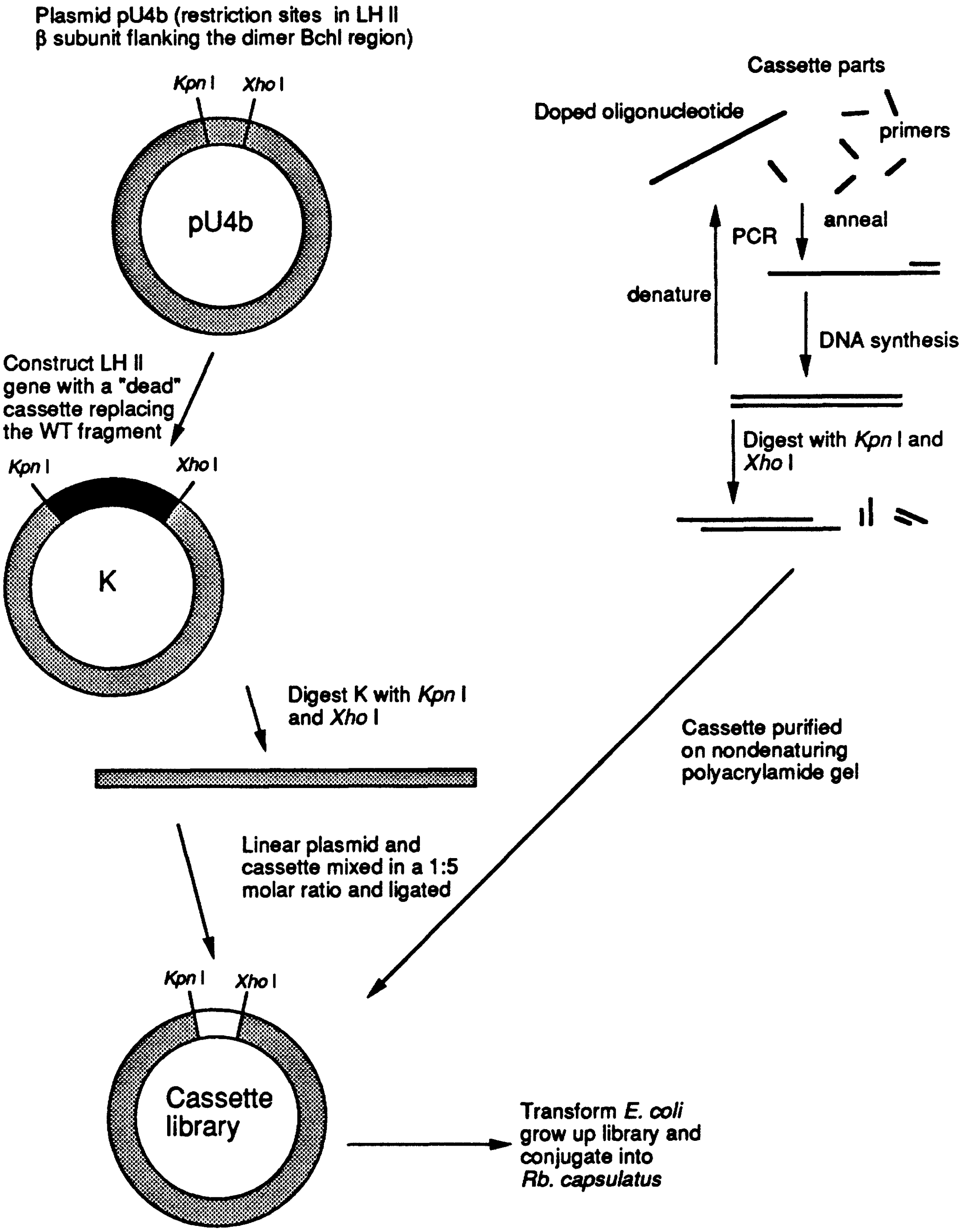

Plasmid pU4b (restriction sites in LH II
β subunit flanking the dimer Bchl region)
Kpn I
Xho I
pU4b
Construct LH II gene with a "dead" cassette replacing the WT fragment
Kpn I
Xho I
K
Digest K with *Kpn* I and *Xho* I
Linear plasmid and cassette mixed in a 1:5 molar ratio and ligated
Kpn I
Xho I
Cassette library
Transform *E. coli* grow up library and conjugate into *Rb. capsulatus*
Cassette parts
Doped oligonucleotide
primers
PCR
anneal
denature
DNA synthesis
Digest with *Kpn* I and *Xho* I
Cassette purified on nondenaturing polyacrylamide gel

Figure 16. Cloning scheme for combinatorial cassette mutagenesis in the region of the β dimer of LHII. A similar scheme is used for the β monomer region using *Nsi* I *Kpn* I sites.

After deprotection with NH_4OH, the large oligonucleotides were purified on a 6% polyacrylamide gel containing 8M urea. The conditions for PCR are detailed in section 2.5.2. After PCR, the concentration of the cassettes was estimated from agarose gels. The PCR products were phenol:chloroform (1:1) extracted, and chloroform extracted (each extraction in this protocol was back extracted), then concentrated by ethanol precipitation or centrifuging in centricon-10 microconcentrators (Amicon). Cassettes were digested with 1.5 times the amount of enzyme needed for complete digestion (calculated assuming no loss of DNA during the extractions and concentration). Digested cassette was phenol:chloroform extracted, chloroform extracted, de-salted and concentrated by centrifuging in centricon-10 microconcentrators before purifying on a 12% non denaturing polyacrylamide gel.

Vector with dead cassette was digested with the appropriate restriction enzymes. Digested vector was phenol:chloroform extracted, chloroform extracted, and ethanol precipitated. DNA concentration was estimated from an agarose gel.

Digested cassette and linear plasmid were mixed in approximately a 5:1 molar ratio, respectively. Ligations were done in 30-50 μl. Before adding ligase and ligase buffer, the solution was heated to 70° C for 10 min and allowed to cool slowly. Ligations were incubated for 24 hours at 16 C before ethanol precipitating or centrifuging in centricon-30 concentrators (Amicon).

The DNA for the library constructions was electroporated into *E. coli* strain S17. Cell preparation of S17 for electroporation were performed according to Dower *et. al.*, 1988. The harvested cells from one liter of culture were resuspended to a final volume of 8 ml in sterile 10% glycerol water v/v. Aliquots were then frozen (0.5 ml) in liquid nitrogen and stored at -80 C. Prior to transformation, DNA was mixed with 400 μl of thawed cells kept on ice and transferred to cooled 0.2 cm cuvettes (BioRad). Electroporation was performed using a BioRad Gene pulser and pulse controller set at 25 μF and 1,000 Ω. Immediately after pulsing, the cells were transferred to 10 ml of L media equilibrated at 37 C.

For pU4 libraries in S17, the transformation was incubated between one and two hours by shaking at 200 rpm at 37 C. Aliquots of serial dilutions of the transformations were plated on L media with 12.5 μg/ml tetracycline. The remainder of the transformed cells were inoculated into 60 ml of liquid L media with tetracycline and incubated for 12 hours with shaking at 37 C. Stocks were prepared from the final culture by mixing with an equal volume of L media containing 30% glycerol. DNA from the transformations was analyzed on agarose gels to estimate the fraction of mutagenic cassette to dead cassette in the library. Usually the ratio was 10-15 times the amount of mutagenic casette to dead cassette, but it was always at least 3:1, mutagenic:dead in the transformations that were conjugated into *Rb. capsulatus* and screened.

Conjugations were performed as described by Bylina *et al.*, 1989 with a few modifications. The recipient strain (U71) was grown shaking at 32 C to early log-phase in liquid RM media. For the libraries in the S17 donor strain, 200 µl of cell stock was inoculated into 10 ml of L media without antibiotics and grown shaking at 200 rpm for 6-7 hours at 37 C. For each library, 300 µl of S17 culture and 700 µl U71 culture were mixed by gentle vortexing in a sterile 2059 tube (Falcon). Each mixture was then spotted on separate MPYE plates. For controls, 1 ml of U71 and each S17 donor culture were spotted onto separate plates. The MPYE plates were incubated at 32 C for 24 hours. By this time the cells had formed a thick spot on the plate. The cells on each plate were resuspended in 1 ml RCV media using a spreader, and transferred to sterile microcentrifuge tube. The tubes were vortexed for 30 sec and then the cells were pelleted in a microcentrifuge for 5 min. A biphasic pellet resulted with *Rb. capsulatus* on the top and *E. coli* cells on the bottom. The U71 pellets were resuspended in RCV, transferred to a fresh sterile tube and pelleted a second time. The top portion of the pellet was suspended in 1 ml RCV and transferred to a fresh tube. Serial dilutions were made in RCV from these suspensions and spread on RCV tetracycline plates for growth in the dark. The remaining portion of the suspensions were mixed with an equal volume of sterile filtered RM media with 30% glycerol v/v to maintain stocks of the library.

Screening was performed by DIS as described in the previous chapters.

B. Plasmids and deletion backgrounds

Prior to my thesis work LHII was expressed from plasmid pU2 (Youvan *et al.*,1985) a pBR322 derivative that did not replicate efficiently in *Rb. capsulatus.* Triparental matings were required to transfer the plasmid from *E. coli*, where the genetic manipulation is done, to *Rb. capsulatus* where the protein is expressed. This genetic system would be inadequate for passing large libraries with a high complexity of sequences. The genes coding for LHII were cloned into a derivative of the broad host range plasmid, pRK404 (Ditta *et al.*, 1985), which replicates in *Rb. capsulatus* (pBW). This conjugative plasmid complements LHII in the genomic LHII deletion background. Unique restriction sites were engineered into the new plasmid to facilitate subcloning into M13 and CCM of the β subunit (Figure 15).

Starting plasmids:

pRK404: broad host range plasmid (Ditta *et al.*, 1985)

pU2: Original LHII expression plasmid (Youvan *et al.*,1985). Figure 17 shows a restriction map of this plasmid.

Intermediates:

D1: Pst I - Bam HI fragment from pU2 cloned into pRK404. Constructed by Steve Robles, a former graduate student in the Youvan lab.

pRK404-HT: Hind III site removed from plasmid pRK404. pRK404 was digested with Hind III, the 3' recessed ends were filled using T4 DNA polymerase and a mixture of all four nucleotides, and blunt end ligated in a minimal volume.

pBW: Pst I - Bam HI fragment from pU2 cloned into pRK404-HT

pBW-H4: pBW with the Hind III site from the pU2 DNA removed by cutting with Hind III, filling in the ends using T4 DNA polymerase and a mixture of all four nucleotides, and blunt end ligating.

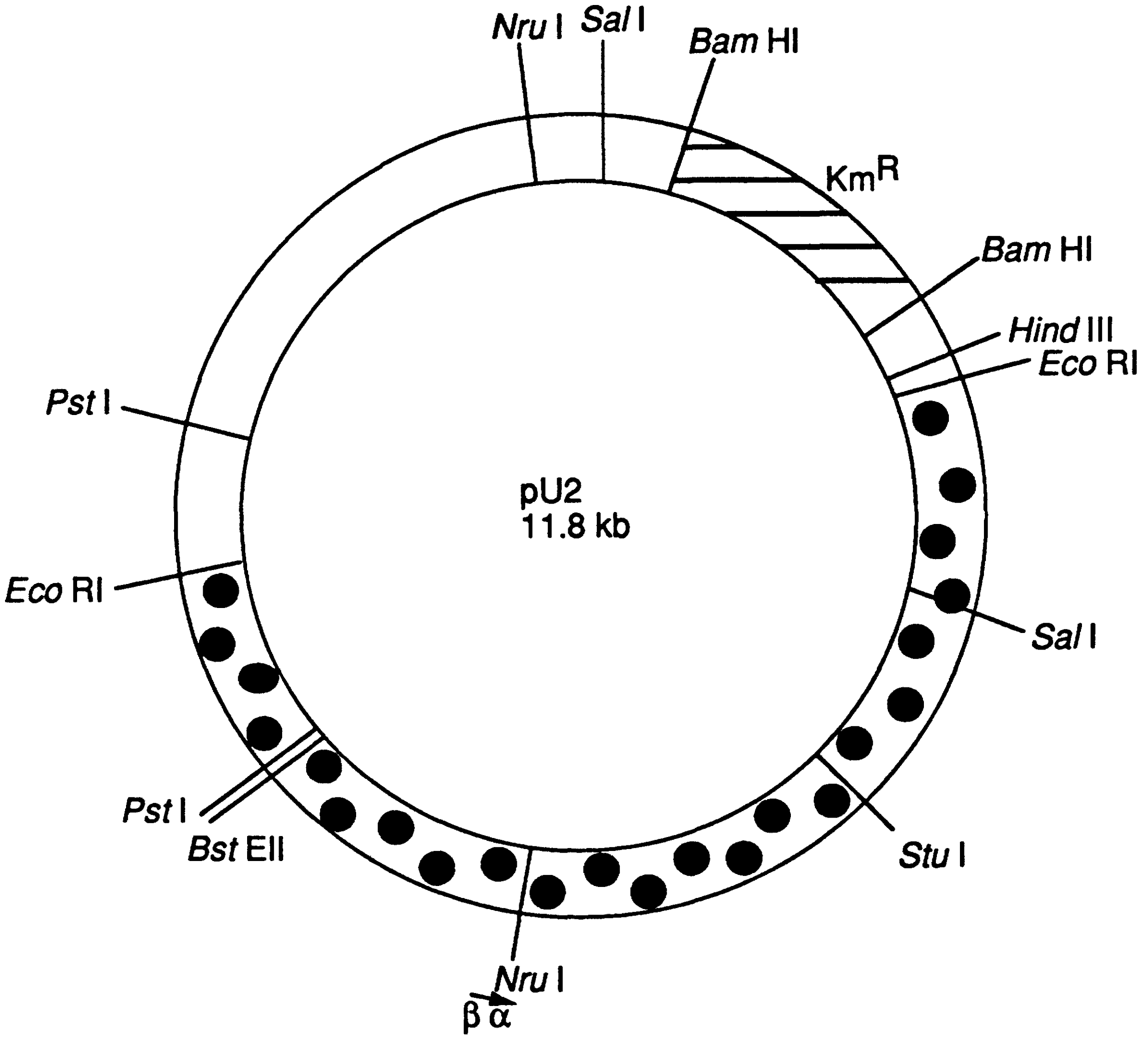


Figure 17. Map of LHII expression plasmid pU2. The LH II genes are contained on a 5.75 kb Eco RI fragment of DNA from *Rb. capsulatus* denoted by circles. The LH II coding transcript is shown by a small arrow. The smallest Pst I - Bam HI fragment containing the LH II genes was used in the construction of the new LH II expression plasmids.

D118H: 3.8 kb Pst I - Sal I fragment from D1 containing the LH α and β structural genes was cloned into mp18. D1 and mp18 were digested with both enzymes. The 3.8 kb Pst I - Sal I fragment from D1 was isolated from a low melting point point agarose gel and ligated with double cut mp18.

9F: D118H with Hind III site engineered just down stream of α (see Figure 15 for the location of the site).

pU4: (Figure 18; also called 4Z in lab notebook) Pst I - Sal I fragment from 9F (with Hind III site for shuttling into M13) cloned into pBW-H4. pBW-H4 was digested with Pst I followed by a partial digest with Sal I. The desired fragment (11.2 Kb) was purified on a low melting point agarose gel and ligated with Pst I-Sal I digest of 9F.

Final LH II expression plasmids used in CCM experiments

pU4b: Unique Xho I and Kpn I sites engineered into pU4 to facilitate CCM of dimer Bchl associated with the β subunit.

pU4c: Unique Nsi I, Xho I, and Kpn I site engineered into pU4 to facilitate CCM of monomer Bchl associated with the β subunit as Nsi I- Kpn I cassette.

Dead cassettes:

K: 4 kb Kpn I-Xho I fragment from pU2925 (Robles, 1993) cloned into the Kpn I - Xho I sites of pU4b used to prevent WT contamination of the library. The dead cassette is later replaced by the mutagenic cassette in the construction of the library.

TC: 6.5 kb Nsi I- Kpn I fragment from bacteriophage lambda purified from a low melting point agarose gel and ligated into the Nsi I - Kpn I sites of pU4c.

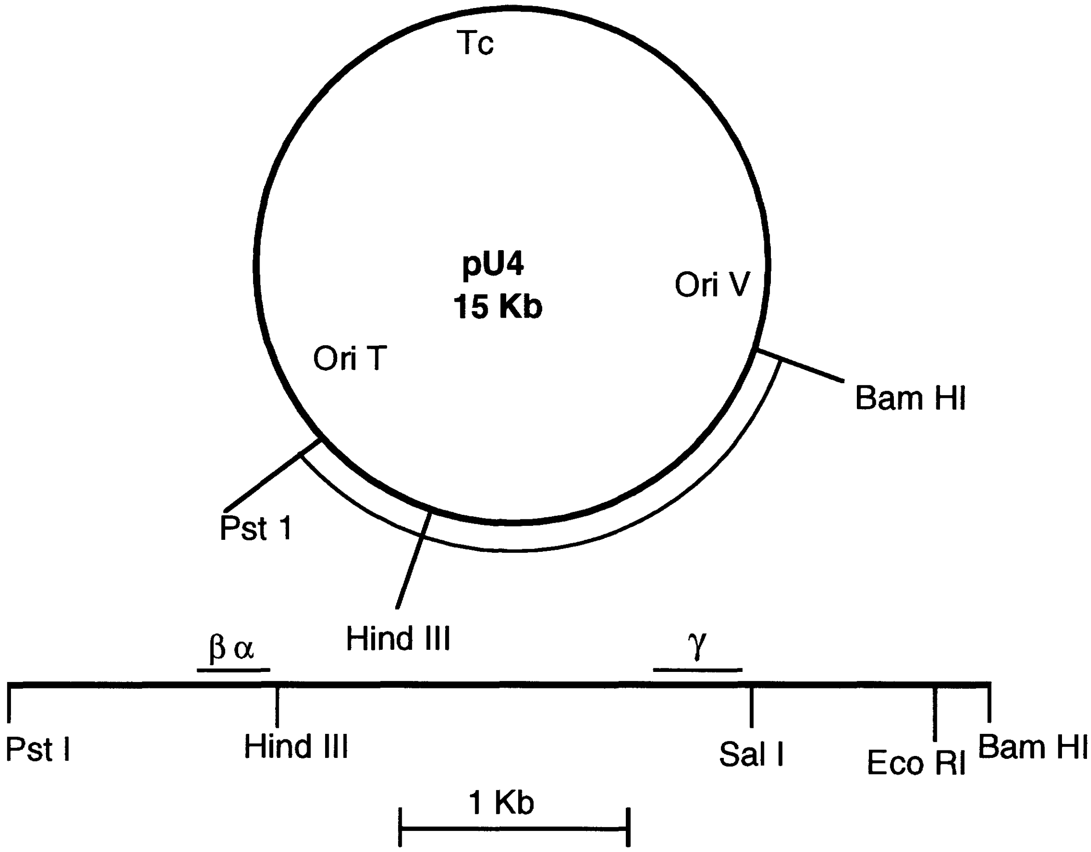


Figure 18. Map of LH II expression plasmid pU4.

Deletion backgrounds:

U71: Light harvesting II Bst E II- Stu I deletion, spectinomycin insertion, with LHI and RC expression inactivated by a point mutation (Youvan *et al.*, 1985).

Udd4: Double chromosomal deletion background. Starting with strain U72 (interposon mediated insertion/deletion of LHII genes, WT LHI and RC genes) an Eco RI - Bam HI deletion (LHI and RC L and M subunits) was added. The Eco RI - Bam HI fragment from pU2924 (Robles *et al.*, 1990; Robles, 1993; Figure 19), was cut out and the ends filled with T4 polymerase and the plasmid blunt end ligated to get a construction missing the fragment with the genes to be deleted. This plasmid was then conjugated into strain U72 and subcultured 10 times with RM kanamycin and at least 10 times in RM without kanamycin. Aliquots were plated from the 10 onward subcultures with out kanamycin on RM agar plates. The plates were visually inspected for light colored colonies. Pale colonies were re-streaked and subcultured in RM. Each potential deletion background was tested for growth on kanamycin, assayed spectrally, and put in high and low light conditions to look for possible revertants. Deletion strains should show no kanamycin resistance, their absorption spectra should be of free pigments in the membrane, and they should not revert to photosynthetic growth under any conditions.

Four potential deletion backgrounds were found that did not grow on kanamycin, had the spectra characteristic of only free Bchl in the membrane, and never reverted to photosynthetic growth under any light conditions. Southern hybridizations were done on two of these backgrounds to confirm the genomic deletion. Two strategies were used. First, chromosomal DNA from WT, the potential deletion backgrounds, and U43 (a LHI, RC insertion/deletion background Youvan *et al.*, 1985) was digested with Apa I, as a three kb fragment should vanish in the deletion strains. The blots were probed with the WT L subunit of the RC (Hind III - Kpn I fragment see Figure 19) which was radioactively labeled using sequenase and a sequencing primer. WT "lit up" with this probe while the potential double deletion background and the established deletion background did not. Next, chromosomal DNA from WT, the potential deletion backgrounds, and U43 was digested with Bam HI to look for a 2 kb deletion in the potential deletion backgrounds. The blots were probed with a Kpn I-Sac I fragment that extends past the M subunit (see Figure 19).

The probe was labeled using sequenase and the universal primer. The southern showed the two potential deletion backgrounds had the band shifted to lower molecular weight by approximately 2kb, confirming that they had the desired chromosomal deletion.

When the LHII expression plasmids were conjugated into the deletion background Udd4, the cultures lost their carotenoids under certain growth conditions (liquid RM media). This indicated that there may be a mutation in genes associated with the carotenoids. The decision was made not to further characterize the deletion strains, but to do the experimental work for this thesis in the established U71 deletion background.

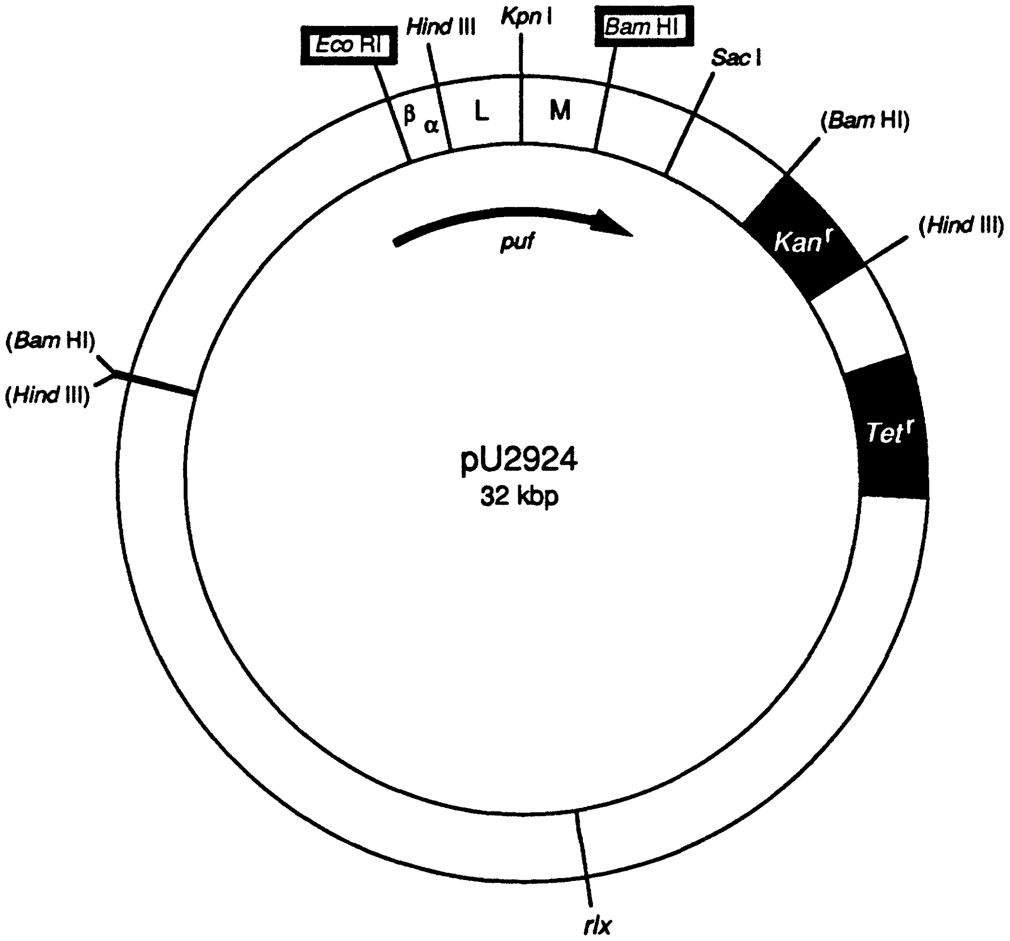


Figure 19. Map of plasmid pU2924 containing the *puf* operon the Eco RI and Bam HI sites bounding the fragment of the *puf* operon deleted in the construction of the deletion background are enclosed in boxes.

C. Strains:

Translation table
stock name --> publication name

stock name	publication name	stock name	publication name
2-D	S1	2-B	P1
1-b1	S2	1-a2	P2
2-G	S3	2-c	P3
2-E	S4	4-S	P4
2-A	S5	1-2	P5
870.1A	S6	1-3	P6
1-C2	S7	C21.Z	P7
870.2A	S8	C11.Z	P8
1-C3	S8	3-W	P9
5-LH1	S10	C21.Y	P10
870.2E	S11	4-T	P11
2-H	S12	1-1	P12
20	S13	2-D?	P13
5	S14	21	P14
3	S15	24	P15
23	S16	10	P16
15	S17	11	P17
38	S18	2	P18
29	S19	9	P19
17	S20	6	P20
50	S21	7	P21
51	S22	22	P22
27	S23	8	P23
		16	P24
N11	N1	31	P25
N12	N2	34	P26
N19	N3	37	P27
N18	N4	19	P28
N9	N5	35	P29
N2	N6	25	P30
N8	N7	33	P31
N8	N8	36	P33
N16	N9	45	P34
N15	N10	54	P35
N13	N11	58	P36
N23	N12	59	P37
		52	P38
		53	P39
		55	P40

D. Media

L broth:
10 g Bacto-tryptone
5 g yeast extract
5 g NaCl
1 l water, autoclave 20 min

2xYT:
16 g Bacto-tryptone
10 g yeast extract
5 g NaCl
1 l water, autoclave

MPYE:
3 g Bacto-peptone
3 g yeast extract
1.6 ml 1 M $MgCl_2$
1.0 ml 1 M $CaCl_2$
1 l water, autoclave 20 min

RCV:
10 ml 10% $(NH_4)_2SO_4$
40 ml Na malate
50 ml supersalts
15 ml 0.64 M potassium phosphate
880 ml water, autoclave 20 min

RM:
1 g Bacto-peptone
1 g yeast extract
7 ml 10% $(NH_4)_2SO_4$
27 ml Na malate
33 ml supersalts
10 ml 0.64 M potassium phosphate
0.53 ml 1M $MgCl_2$
0.33 ml 1M $CaCl_2$
930 ml water, autoclave 20 min

RCV+ or RM+
Add to 1 l RCV or RM
before autoclaving:
5 g Na pyruvate
6 g glucose
3.6 ml DMSO

For plates add 15 g agar before autoclaving. For RCV or RM plates, add sterile potassium phosphate after autoclaving.

0.64 M potassium phosphate:
40 g KH_2PO_4 (anhydrous)
60 g K_2HPO_4 (anhydrous)
bring volume to 1 l with water
pH should be 6.8,
autoclave 20 min

10% sodium malate:
100 g DL-malic acid
60 g NaOH (add 20 g at a time)
bring volume to 1 l with water
pH should be 6.8
autoclave 20 min

super salts:
0.4 g Na_2EDTA
4.0 g $MgSO_4 \cdot 7H_2O$
1.5 g $CaCl_2 \cdot H_2O$
0.24 g $FeSO_4 \cdot 7H_2O$
20 ml trace elements
bring to 1 l with water
autoclave 20 min
add 10 ml 2 mg/ml sterile filtered thiamine

trace elements:
0.4 g $MnSO_4 \cdot H_2O$
0.7 g H_3BO_3
10 mg $Cu(NO_3)_2 \cdot 3H_2O$
60 mg $ZnSO_4 \cdot 7H_20$
187 mg $NaMoO_4 \cdot 2H_2O$
bring to 250 ml with water

References

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

Arkin, A. P., & Youvan, D. C. (1993). Digital imaging spectroscopy. In Deisenhofer H. & Norris JR. (eds) *The Photosynthetic Reaction Center, Vol. 1* (pp. 133-155) Academic Press, New York.

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

Babst, M., Albrecht, H., Wegmann, I., Brunisholz, R., & Zuber, H. (1991). Single amino acid substitutions in the B870 α and β light-harvesting polypeptides of *Rhodobacter capsulatus. Eur. J. Biochem* **202**:277-284.

Breton, J., Bylina, E. J., & Youvan, D. C. (1989). Pigment orientation in genetically modified reaction centers of *Rhodobacter capsulatus. Biochemistry* **28**:6423-6430.

Brunisholz, R. A., & Zuber, H (1988). Primary structure analysis of bacterial antennae polypeptides: Correlation of aromatic amino acids with spectral properties, Structural similarities with reaction center polypeptides. In Scheer H, Schneider S (eds.) *Photosynthetic light-harvesting systems organization and function* (pp. 103-114) Walter de Gruyter & Co., New York.

Burgess, J. G., Ashby, M. K. & Hunter, N. (1989). Characterization and complementation of a mutant of *Rhodobacter sphaeroides* with a chromosomal deletion in the light-harvesting (LH2) genes. *J. Gen. Microbiol.* **135**:1809-1816.

Bylina, E. J., & Youvan, D. C. (1988a). Directed mutations affecting spectroscopic and electron transfer properties of the primary donor in the photosynthetic reaction center. *Proc. Natl. Acad. Sci U.S.A.* **85**:7226-7230.

Bylina, E. J., & Youvan, D. C. (1989). Mutagenesis of reaction center histidine L173 yields an L-side heterodimer. In Baltscheffsky M (ed.)*Current Research in Photosynthesis* (pp. 53-59) Klumer Academic Press, Boston.

Bylina, E. J., Ismail, S., & Youvan, D. C. (1986). Plasmid pU29, a vehicle for mutagenesis of the photosynthetic *puf* operon in *Rhodopseudomonas capsulatus*. *Plasmid* **16**:175-181.

Bylina, E.J., Jovine, R.V.M. & Youvan, D.C. (1989). A genetic system for rapidly assessing herbicides that compete for the quinone binding site of photosynthetic reaction centers. *Bio/Technology* **7**:69-74.

Bylina, E. J., Robles, S., & Youvan, D. C. (1988b). Directed mutations affecting the putative bacteriochlorophyll-binding sites in the light-harvesting antenna of *Rhodobacter capsulatus*. *Israel J. Chem.* **28**:73-78.

Bylina, E. J., Kolaczkowski, S. V., Norris, J. R., & Youvan, D. C. (1990). EPR characterization of genetically modified reaction centers of *Rhodobacter capsulatus*. *Biochemistry* **29**:6203-6210.

Bylina, E. J., Kirmaier, C., McDowell, L., Holten, D., & Youvan, D. C. (1988c). Influence of an amino acid residue on the optical properties and electron transfer dynamics of a photosynthetic reaction center complex. *Nature* **336**:182-184.

Chang, M. C., Callahan, P. M., Parkes-Loach, P. S., Cotton, T. M. & Loach, P. A. (1990). Spectroscopic characterization of the light harvesting complex of *Rhodospirillum rubrum* and its structural subunit. *Biochemistry* **29**:421-429.

Coleman, W. J., & Youvan, D. C. (1990). Spectroscopic analysis of genetically modified photosynthetic reaction centers. *Ann. Rev. Biophys. Biophys. Chem.* **19**:333-367.

Cogdell, R. J. & Scheer, H. (1985). Circular dichroism of light-harvesting complexes from purple photosynthetic bacteria, *Photochem. Photo biol.* **42**:669-678.

Cogdell, R & Hawthornthwaite, A. M. (1993). Preparation, and crystallization of purple bacteria antenna complexes. In Deisenhofer J Norris JR (eds). *The photosynthetic reaction center volume 1.* (pp 23-42). Academic Press, Inc. Boston.

Cybenko, G. (1989). *Math. Contr. Signals, Syst* **2**:303-314.

Deisenhofer, J., Epp, O., Miki, K., Huber, R., & Michel, H. (1985). Structure of the protein subunits in the photosynthetic reaction center of *Rhodopseudomonas viridis* at 3Å resolution. *Nature* **318**:618-624.

Delagrave, S. & Youvan, D. C. (1993). Searching sequence space to engineer proteins: Exponential ensemble mutagenesis. *Bio/Technology* in press.

Delagrave, S., Goldman, E. R. & Youvan, D. C. (1993). Recursive Ensemble Mutagenesis. *Protein Eng.* **6**:327-331.

DiMango, T. J., Bylina, E. J. Angerhofer, A., Youvan, D. C., & Norris, J. R. (1990). The stark effect in wild-type and heterodimer reaction centers from *Rhodobacter capsulatus. Biochemistry* **29**:6201-6210.

Ditta, G., Schmidhauser, T., Yakobson, E., Lu, P., Liang, X.W., Finlay, D.R., Guiney, D. & Helinsky, D.R. (1985). Plasmids related to the broad host range vector pRK290, useful for gene cloning and for monitoring gene expression. *Plasmid* **13**:149-153.

Donnelly, D. & Cogdell, R. J. (1993). Predicting the point at which transmembrane helices protrude from the bilayer: a model of the antenna complexes from photosynthetic bacteria. *Protein Eng.* **6**:629-635.

Dower, W. J., Miller, J. F. & Raysdale, C. (1988). High efficiency transformation of *E. coli* by high voltage electroporation. *Nuc. Acids. Res.* **16**:6127-6145.

Dorge, B., Klug, G. & Drews, G. (1987). Formation of the B800-850 antenna pigment-protein complex in the strain GK2 of *Rhodobacter capsulatus* defective in carotenoid synthesis. *Biochem. Biophys. Acta.* **892**:68-74.

Drews, G. (1985). Structure and functional organization of light-harvesting complexes and photochemical reaction centers in membranes of phototropic bacteria. *Microbiol. rev.* **49**:59-70.

Drews, G, Dierstein, R. & Schaumacher, A. (1976). Genetic transfer of the capacity to form bacteriochlorophyll-protein complexes in *Rhodopseudomonas capsulata*. *FEBS Lett.* **68**:132-135.

Eccles, J. & Honig, B. (1983). Charged amino acids as spectroscopic determinants for chlorophyll *in vivo.* *Proc. Natl. Acad. Sci. USA* **80**:4959-4962.

Fonstein, M. & Haselkorn, R. (1993). Chromosomal structure of *Rhodobacter capsulatus* strain SB1003: Cosmid encyclopedia and high-resolution physical and genetic map. *Proc. Natl. Acad. Sci. USA* **90**:2522-2526.

Fowler, G. J. S., Visschers, R. W., Grief, G. G., van Grondell, R., & Hunter, C. N. (1992). Genetically modified photosynthetic antenna complex with blue shifted absorption bands. *Nature* **355**:848-850.

Fowler, G. J. S., Crielaard, W., Visschers, R. W., van Grondell, R., & Hunter, C. N. (1993). Site directed mutagenesis of the LH2 light-harvesting complexes of *Rhodobacter sphaeroides*: Changing βlys23 to gln results in a shift in the 850 nm absorption peak. *Photochem. and Photobiol.* **57**:2-5.

Fuellen, G. (1994). Master's thesis. Massachusetts Institute of Technology.

Gherardi, E. & Milstein, C. (1992). Original and artificial antibodies. *Nature* **357**:201-202.

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

Hertz, J., Krogh, A. & Palmer, R. (1991). *Introduction to the Theory of Neural Computation*. Addison-Wesley: Redwood City, California.

Hirst, J.D. & Sternberg M.J.E. (1992). *Biochemistry* **31**:7211-7218.

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

Hornik, K. (1991). *Neural Networks* **4**:251-257.

Hornik, K., Stinchcombe, M. & White, H. (1989). *Neural Networks* **2:**359-366.

Houghten, R. A., Pinilla, C., Blondelle, S. E., Appel, J. R., Dooley, C.T. & 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. & Lerner, R.A. (1991). Linkage of recognition and replication functions by assembling combinatorial antibody Fab libraries on phage surfaces. *Proc. Natl. Acad. Sci. USA* **88**:4363-4366.

Kirmaier, C., Holten, D., Bylina, E. J. & Youvan, D. C. (1988). Electron transfer in a genetically modified reaction center containing a heterodimer. *Proc. Natl. Acad. Sci. USA*. **85**:7562-7566.

Kirmaier, C., Bylina, E. J., Youvan, D. C. & Holten, D. (1989). Subpicosecond formation of the intradimer charge transfer state [$\mathrm{BChl_{LP}{}^{+}BPh_{MP}{}^{-}}$] in reaction centers from the His^{M200}→Leu mutant of Rhodobacter capsulatus. *Chem. Phys. Lett.* **30**:609-613.

Kirmaier, C., Gaul, D., Debey, R., Holten, D. & Schenck, C. (1991). Charge separation in a reaction center incorporating bacteriochlorophyll for photoactive bacteriopheophytin. *Science* **251**:922-927.

Kramer, H. J. M., Van Grondelle, R, Hunter, C. N., Westerhuis, W. H. J. & Amesz, J. (1984). Pigment organization of the B800-B850 antenna complex of *Rhodopseudomonas sphaeroides*. *Biochem Biophys Acta* **765**:156-165.

Kunkel, T. A., Roberts, J. D. & Zabour, R. A. (1987). Rapid efficient site specific mutagenesis with out phenotypic expression. *Methods Enzymol.* **154**: 367-382.

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

Marrs, B. (1981). Mobilization of the genes for photosynthesis from *Rhodopseudomonas capsulata* by a promiscuous plasmid. *J. Bacteriol.* **146**:1003-1012.

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

Pao, Y. H. (1989). *Adaptive Pattern Recognition and Neural Networks,* (Addison-Wesley, Reading, Massachusetts), p.8.

Parkes-Loach, P. A., Sprinkle, J. R. & Loach, P. A. (1988). Reconstitution of the B873 light-harvesting complex of *Rhodosprillum rubrum* from the separately isolated α and β polypeptides and bacteriochlorophyll a. *Biochemistry* **27**:2718-2727.

Pearlstein, R. M. (1991). Theoretical interpretation of antenna spectra. In Scheer H (ed.) *Chlorophylls.* (pp 1042-1078) CRC Press, Boston.

Presnell, S. R. & Cohen, F. E. (1993). Artificial neural networks for pattern recognition in biochemical sequences. *Annu. Rev. Biophys. Biomol. Struct.* **22**:2283-98.

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. & Sauer, R. T. (1991). Random mutagenesis of protein sequences using oligonucleotide cassettes. *Meth. Enzym.* **208**:564-587.

Robert, B. & Lutz, M. (1985). Structures of antenna complexes of several *Rhodospirillales* from their resonance Raman spectra. *Biochem. Biophys. Acta.* **807**:10-23.

Roberts, B. L., Markland, W., Ley, A. C., Kent, R. B., White, D. W., Guterman, S. K. & Ladner, R. C. (1992). Directed evolution of a protein: Selection of potent neutrophil elastase inhibitors displayed on M13 fusion phage. *Proc. Natl. Acad. Sci. USA* **89**:2429-2433.

Robles. S. J. (1993). Ph. D. Dissertation, Massachusetts Institute of Technology.

Robles, S. J. & Youvan, D. C. (1993). Hydropathy and molar volume constraints on combinatorial mutants of the photosynthetic reaction center. *J. Mol. Biol.* **232**:242-252.

Robles, S. J., Breton, J. & Youvan, D. C. (1990a). Transmembrane helix exchange between quasi-symmetric subunits of the photosynthetic reaction center. In Michel-Beyerle M-E (ed.)*Reaction Centers of Photosynthetic Bacteria* (pp. 283-291) Springer Verlag, Berlin Heidelberg.

Robles, S. J., Breton, J. & Youvan, D. C. (1990b). Partial symmetrization of the photosynthetic reaction center. *Science* **248**:1402-1405.

Robles S.J., Ranck, T. & Youvan, D. C. (1992). Symmetrical intragenic suppressors of the bacterial reaction center cd-helix exchange mutants. In Breton J & Vermeglio A (eds) *Structure of the bacterial photosynthetic reaction center (II).* Plenum press, New York.

Rost, B. & Sander, C. (1993). *Proc. Natl. Acad. Sci. USA* **90**:7558-62.

Sambrook, J., Fritsh, E. F. & Maniatis, T. (1989) . *Molecular cloning: A laboratory manual 2nd edition*, Cold spring harbor laboratory, Cold Spring Harbor, New York.

Scherz, A. & Parson, W. W. (1984). Exciton interactions in dimers of bacteriochlorophyll and related molecules. *Biochem. Biophys. Acta.* **766**:666-678.

Scherz, A. & Rosenbach-Belkin (1989). Comparative study of optical absorption and circular dichroism of bacteriochlorophyll oligomers in triton X-100, the antenna pigment B850, and the primary donor P-860 of photosynthetic bacteria indicates that all are similar dimers of bacteriochlorophyll a. *Proc. Natl. Acad. Sci. USA* **86**:1505-1509.

Schumacher, A., & Drews, G. (1978). *Biochem. Biophys. Acta.* **501**:183-194.

Shavlik, J. W., Mooney R. J. & Towell G. G. (1991). *Machine Learning* **6:**111-43.

Simon, R., Priefer, U. & Puhler, A. (1983). A broad host range mobilization system for *in vitro* genetic engineering: transposon mutagenesis in gram negative bacteria. *Bio/Technology* **1**:784-791.

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

Springer, B.A. & Sligar, S.G. (1987). High-level expression of sperm whale myoglobin in *E. coli. Proc. Natl. Acad. Sci. USA* **84**:8961-8965.

Stiehle, H., Cortez, N., Klug, G. & Drews, G. (1990). A negatively charged N terminus in the α polypeptide inhibits formation of light-harvesting complex in *Rhodobacter capsulatus*. *Eur. J. Biochem* **202**:277-284.

Tadros M.H., Garcia, A.F., Gad'on, N. & Drews,G. (1989). Characterization of a pseudo-B870 light-harvesting complex isolated from the mutant strain Ala$^+$ Pho$^-$ of *Rhodobacter capsulatus* which contains B800-850-type polypeptides. *Bioch. Biophys. Acta* **976**:161-167.

Theiler, R. & Zuber, H. (1984). The light-harvesting polypeptides of *Rp. sphaeroides* R-26.1: II. Conformational analysis by attenuated total reflection infrared spectroscopy and the possible molecular structure of the hydrophobic domain of the B850 complex. *Hoppe-Seyler's Z. Physiol. Chem.* **365**:721-729.

Theiler, R., Suter, F., Wiemken, V. & Zuber, H. (1984). The light-harvesting polypeptides of *Rp. sphaeroides* R-26.1:I. Isolation, purification and sequence analyses. *Hoppe-Seyler's Z. Physiol. Chem.* **365**:703-719.

Vos, M. H., Lambry, J.-C., Robles, S. J., Youvan, D. C., Breton, J. & Martin, J.-L. (1991). Direct observation of vibrational coherence in bacterial reaction centers using femtosecond absorption spectroscopy. *Proc. Natl. Acad. Sci. U.S.A.* **88**:8885-8889.

Welte, W., Wacker, T., Lewis, M., Kreutz, W., Shiozawa, J., Gad'on, N. & Drews, G. (1985). Crystallization of the photosynthetic light-harvesting pigment-protein complex B800-850 of *Rhodopseudomonas capsulata*. *FEBS* **182**:260-263.

Yang M. M. & Youvan, D. C. (1988). Applications of imaging spectroscopy in molecular biology. I. Screening photosynthetic bacteria. *Bio/Technology* **6**:746-749.

Youvan, D. C. (1991). Photosynthetic reaction centers: interfacing molecular genetics and optical spectroscopy. *TIBS* **16**: 145-149.

Youvan, D. C. & Ismail S. (1985). Light-harvesting II (B800-B850 complex) structural genes from Rhodopseudomonas capsulata. *Proc. Natl. Acad. Sci. U.S.A.* **82**:58-62.

Youvan, D. C. & Mars B. L. (1984). Molecular genetics and the light reactions of photosynthesis *Cell* **39**:1-3

Youvan, D. C., Hearst, J. E. & Marrs, B. L. (1983). Isolation and characterization of enhanced fluorescence mutants of *Rhodopseudomonas capsulata*. *J. Bacteriol.* **154**:748-755.

Youvan, D. C., Ismail, S. & Bylina, E. J. (1985). Chromosomal deletion and plasmid complementation of the photosynthetic reaction center and light harvesting genes from *Rhodopseudomonas capsulatus*. *Gene* **38**:19-30.

Youvan, D. C., Arkin, A. P. & Yang M. M. (1992). Recursive ensemble mutagenesis: A combinatorial optimization technique for protein engineering. In: Manverik B (ed) *Parallel problem solving from Nature, 2* (pp 401-410) Elsevier publishing Co. Amsterdam.

Youvan, D. C., Goldman, E., Delagrave, S. & Yang, M. M. (1993). Digital imaging spectroscopy for massively parallel screening of mutants. *Meth. Enzym.* in press.

Zoller, M. J. & Smith, M. (1984). Oligonucleotide-directed mutagenesis: a simple method using two oligonucleotide primers and a single strand DNA template. *DNA* **3**:479-488.

Zuber, H. (1986). Structure of light harvesting antenna complexes of photosynthetic bacteria, cyanobacteria, and red algae. *TIBS* **11**:414-419.

Zuber, H. (1990). Consideration on the structural principles of the antenna complexes of phototrophic bacteria. In: Drews G & Dawes EA (Eds.) *Molecular biology of membrane-bound complexes in phototrophic bacteria.* (pp 161-180) Plenum press, New York.

Zuber, H. & Brunisholz, R. A. (1991). Structure and function of antenna polypeptides and chlorophyll-protein complexes: Principles and variability. In Scheer H (ed.) *Chlorophylls*. (pp 627-703) CRC Press, Boston.