UNIVERSITY OF ILLINOIS

May 9 1989

THIS IS TO CERTIFY THAT THE THESIS PREPARED UNDER MY SUPERVISION BY

Alec Hodel

ENTITLED The Active Site of Cytochrome P-450cam Through Static Computer Modelling

IS APPROVED BY ME AS FULFILLING THIS PART OF THE REQUIREMENTS FOR THE

DEGREE OF Bachelor of Science in Engineering Physics

S Sligar

Instructor in Charge

APPROVED: Robert L Switzer

HEAD OF DEPARTMENT OF Biochemistry

O-1364

THE ACTIVE SITE OF CYTOCHROME P-450$_{CAM}$ THROUGH STATIC COMPUTER MODELLING

BY

ALEC HODEL

THESIS

for the

DEGREE OF BACHELOR OF SCIENCE

in

BIOCHEMISTRY

College of Engineering

University of Illinois

Urbana, Illinois

1989

Acknowledgements

I sincerely thank the entire Sligar lab for giving much more than they received.

Table of Contents

I. Abstract

We describe herein modelling efforts at the re-engineering of a protein to a low-order approximation. The specific problem addressed is the closing of the active site of cytochrome $P\text{-}450_{cam}$ by filling the volume that the substrate previously occupied with the side chains of the surrounding residues. This may exclude water from the heme pocket creating a high spin, substrate-free protein which is intrinsically interesting to study and characterize. It is also a necessary first step toward the popular goal of *de novo* active site design. The site-directed mutagenesis is modelled by a computer algorithm based on the work of Ponder and Richards[1]. This algorithm applies basic rules of steric hinderance to a static view of the protein while also making use of the rotamer library concept of side chain conformations. The results yield a strategy for mutagenesis which may be further examined through dynamic simulation or executed through molecular biology.

The algorithm developed as my thesis is neither specific to this problem nor to this protein and further studies in protein engineering are discussed.

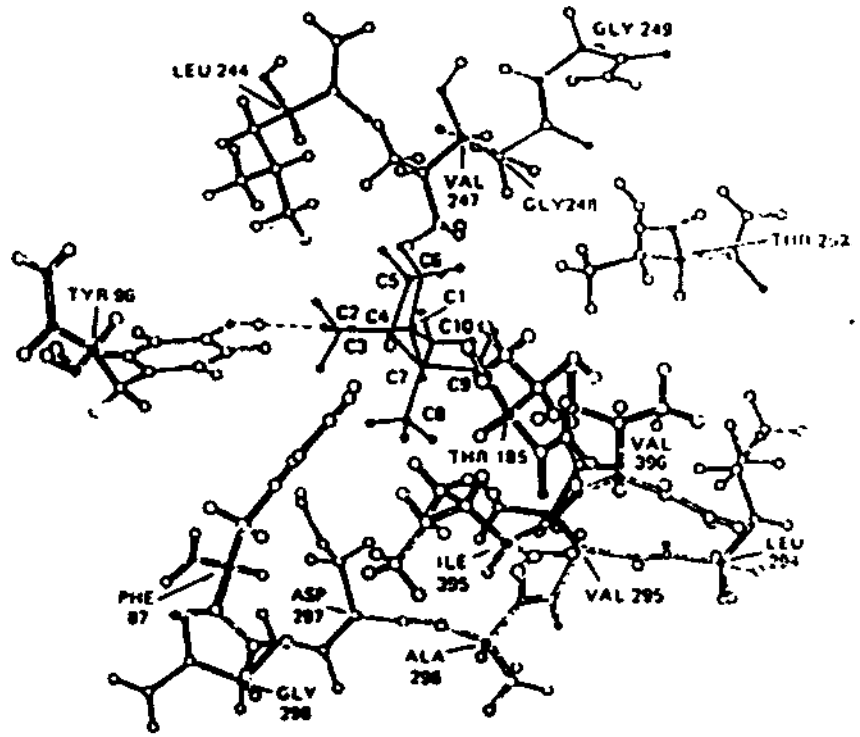
II. Introduction

The problem of protein folding and design is one of the most interesting and most frustrating fields in modern biochemistry. Every short step toward the understanding of protein-substrate recognition and the engineering of specificity promises great leaps in the application of biotechnology toward man-made catalysts with biological specificity and efficiency. The oxidative enzyme cytochrome $P\text{-}450_{cam}$ has been the subject of studies in molecular recognition and enzyme redesign by Atkins and Sligar[2]. In their paper, they commented on the utility and possibility of re-engineering cytochrome P-450 monooxygenases into oxidative catalysts of unactivated alkanes and alkenes. The cytochromes P-450 are a class of heme bearing proteins whose thiolate axial ligand distinguishes the class spectroscopically and functionally. Presently, cytochrome $P\text{-}450_{cam}$ is one of the best characterized of this class, with both its gene and its x-ray structure published[3,4]. It is a soluable bacterial protein originally isolated from Pseudomonas putida but has been cloned into E coli on a plasmid vector[3]. As it is easy to genetically manipulate, express, and purify in large quantities and has a certain future importance to biotechnology, $P\text{-}450_{cam}$ provides a perfect subject for a study in protein engineering.

$P\text{-}450_{cam}$ catalyzes the 100 percent regio- and stereospecific hydroxylation of d-camphor into 5-exo-hydroxycamphor. As evidenced form the x-ray structure, the active site in the heme pocket provides numerous Van der Waals contacts with camphor through ten different primarily hydrophobic residues[4] (see Fig. 1). A hydrogen bond is also evident in the structure between one of the residues, tyrosine-96, and the carbonyl of camphor; however, it has been shown that this hydrogen bond does not contribute significantly to the binding energy of the molecule, but serves as an additional anchor for the regiospecificity of the reaction[4]. Thus the binding energy and much of the specificity of $P\text{-}450_{cam}$ are provided by the ten residues which surround the camphor and provide a snug, well-shaped cocoon above the heme plane. As a starting point in the quest for re-engineering the specificity of the protein, one could consider the protein minus the ten aforementioned residues as a static hydroxylating machine whose substrate is dictated by the ten heme pocket residues. As the exact role of each residue in the hydroxylation reaction or in the stability of the protein has not been determined, this may be a rather strong approximation; but if considered carefully, it may provide a fitting foundation on which to begin.

Thus the problem of redesigning the active site may now be concentrated on how these ten residues should be

RESIDUES CONTACTING CAMPHOR



Phe-87
Tyr-96
Thr-185
Leu-244
Val-247
Thr-252
Val-295
Asp-297
Ile-395
Val-396

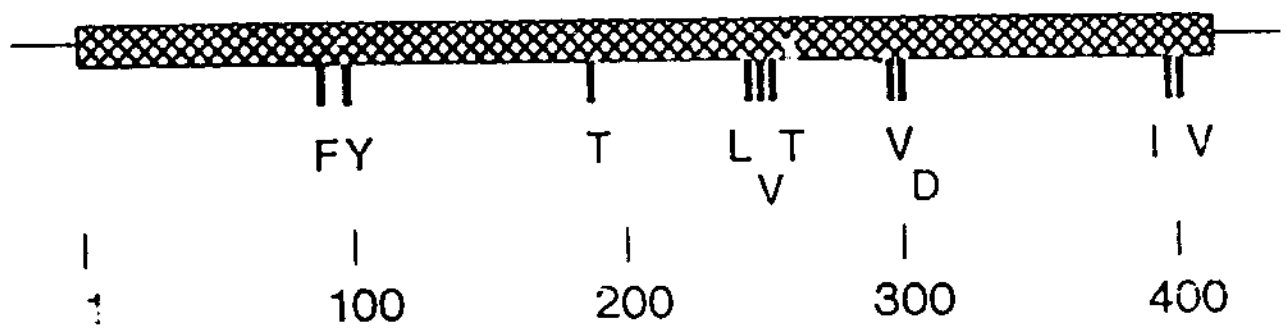


Figure 1

mutated to produce the desired result, such as specificity for a molecule other than camphor. Atkins had made three mutations in the P-450$_{cam}$ active site, including Val-247 to Ala, Val-295 to Ile, and Tyr-96 to Phe[2]. All of these consist of single residue mutations, and their effect on camphor hydroxylation was minimal. It is my hope to design large-scale mutations involving several residues and, with luck, produce more dramatic results.

As an initial goal, I chose simply to fill as much of the heme pocket volume as possible with polypeptide side chains. The effect of such a mutation can be considered only in conjecture. When it is substrate free, the wild type P-450$_{cam}$ active site is occupied by a small hydrogen bond lattice of six water molecules. One of these molecules serves as the sixth ligand for the heme iron atom, causing a primarily low spin state where the electron affinity of P-450$_{cam}$ is too low to accept electrons from its electron donor putidaredoxin. The binding of camphor excludes water, causing a spin state shift to high spin while also raising P-450$_{cam}$'s electron affinity to a reducible level[5]. The reduction of P-450$_{cam}$ is a necessary step in the hydroxylation event which is only possible if water, as the sixth ligand to the heme iron, is excluded from the active site. If the proposed mutation scheme of filling the active site of P-450$_{cam}$ with side chain residues were to exclude water from the heme pocket, the result would be a substrate-

free, high spin protein. If a stable substrate-free, high spin protein could be produced, it would be quite an interesting subject for characterization as well as being a minor achievement in protein engineering.

The design strategy for such a feat, which was suggested by the work of Ponder and Richards, may be stated quite simply. First, one must define a cavity which is to be filled, such as the heme pocket where camphor and the water lattice bind. Next, one defines which residues of the protein one wishes to mutate, such as the ten heme pocket residues. Then one must find all possible combinations of amino acids which can substitute for the native residues and still fit in the given cavity volume. Finally, one chooses the sequence with the desired properties, or, in this case, the one with the greatest volume. The difficult step in this strategy is the determination of all possible combinations of amino acid substitutions which could fit in the cavity space. This problem is tackled through the approximation and algorithm of Ponder and Richards which is fully discussed in the next two chapters.

III. The Concepts and Approximations

Software to find all possible combinations of amino acid residues to fill a defined volume has been developed based on the algorithm and concepts of Ponder and Richards[1]. In their paper they presented several broad statements concerning protein structure whose validity has been well established through the years:

1) The covalent geometry of the various bonds in a polypeptide are identical to that found in the relevant small molecules.

2) No atomic overlaps are allowed.

These two statements are explicitly present in the model. Two more rules were set forth in their paper; yet, these must be applied as the user sees fit. These are not included in the current computer model:

3) Buried hydrogen bonding donors and acceptors occur in pairs.

4) Groups with formal charges occur only in contact with the solvent.

Next, Ponder and Richards set forth a concept which drastically reduces the computation problem at hand. Through statistical analysis of the conformations of polypeptide side chains in proteins, they found the

residues, rather than taking on any random distribution of conformations, all settled into discrete, preferred conformations. These conformations, or rotamers, represent configurations of minimum energy. Ponder and Richards sampled 19 structures from the Brookhaven Protein Data Bank, all of which were resolved to 1.8 A or better. From this sample, they found that 17 of the amino acids conformed to only 69 rotamers. These rotamers were found and described through their chi (bond torsion) angles. For example, leucine requires two chi angles to fully specify its conformation. Four rotamers or combinations of those chi angles were found, describing four discrete conformations for leucine. And thus, if one is faithful to the concept that all side chains in proteins remain in states of minimum energy (i.e, no energy is stored in the torsion angles), then every leucine in every protein should exist in one of those four conformations. These rotamers then form the three-dimensional puzzle pieces from which proteins may be constructed.

The model presented here utilizes the rotamer concept as one of its central dogmas. The present rotamer library used here contains 67 rotamers of 17 amino acids (see Table 1). Methionine, lysine, and arginine were not well characterized in their study, thus Ponder and Richards did not include these three in their library. For achnical

Table 1
The Rotamer Library of Side Chain Chi Angles[1]

Rotamer	Amino Acid	Chi 1	Chi 2	Chi 3
0	Ala			
1	Asn	-68.3	-36.8	
2		-177.1	1.3	
3		-67.2	128.8	
4		63.9	-6.8	
5		-174.9	-156.8	
6		63.6	53.8	
7	Asp	-68.3	-25.7	
8		-169.1	3.9	
9		63.7	2.4	
10	Cys	-65.2		
11		-179.6		
12		63.5		
13	Gln	-66.7	-178.5	------
14		-174.6	-177.7	------
15		-58.7	-63.8	-46.3
16		-179.4	67.3	26.8
17		70.8	-165.6	------
18		-51.3	-90.4	165.0
19		167.5	70.9	174.2
20	Glu	-69.6	-177.2	-11.4
21		-176.2	175.4	-6.7
22		-64.6	-69.1	-33.4
23		-55.6	77.0	25.3
24		69.8	-179.0	6.6
25		-173.6	70.6	14.0
26		63.0	-80.4	16.3
27	Gly			
28	His	-62.8	-74.3	
29		-175.2	-87.7	
30		-69.8	96.1	
31		67.9	-80.5	
32		-177.3	100.5	
33		48.0	85.9	

Table 1 (continued)

Rotamer	Amino Acid	Chi 1	Chi 2	Chi 3
34	Ile	-60.9	168.7	
35		-59.6	-64.1	
36		61.7	163.8	
37		-166.6	166.0	
38		-174.8	72.1	
39	Leu	-64.9	176.0	
40		-176.4	63.1	
41		-165.3	168.2	
42		44.3	60.4	
43	Phe	-66.3	94.3	
44		-179.2	78.9	
45		66.0	90.7	
46		-71.9	-0.4	
47	Pro			
48	Ser	64.7		
49		-69.7		
50		-176.1		
51	Thr	62.7		
52		-59.7		
53		-169.5		
54	Trp	-70.4	100.5	
55		64.8	-88.9	
56		-177.3	-95.1	
57		-179.5	87.5	
58		-73.3	-87.7	
59		62.2	112.5	
60	Tyr	-66.5	96.6	
61		-179.7	71.9	
62		63.3	89.1	
63		-67.2	-1.0	
64	Val	173.5		
65		-63.4		
66		69.3		

reasons, only one rotamer of proline could be formed. Thus these 67 rotamers, describing all possible conformations of the 17 amino acids, form the geometric units which may be utilized to fill the volume in question.

The final, most important, and most limiting of the approximations used in this model of protein structure is that the main chain of the protein is fixed in space. No relaxation of any kind of the protein infra-structure is taken into account. Whatever shape the protein took in its crystalline state is the permanent structure assumed by this algorithm and defines a static volume in which to work. The static model reduces the computation of this problem by leaps and bounds; however, this approximation limits the validity of the results. Comments on the implications of such an approximation are presented in the discussion.

IV. The Algorithm

The following is the algorithm used to re-engineer cytochrome P-450$_{cam}$. Again, it was primarily suggested by the Ponder and Richards paper. As an example, the problem of excluding camphor and water from the active site is presented; however, bear in mind that the algorithm is very general and may be applied to many other such problems. As a visual aid, two-dimensional cartoons of the procedure are presented in Figure 2.

1) Consider the x-ray structure of the protein static in time and space. Figure 2a shows an unrealistic schematic of P-450$_{cam}$ with the active site magnified.

2) Remove the atoms from the space which shall be filled. For example, camphor is removed from the active site leaving a cavity above the heme plane (see Fig. 2b).

3) The amino acid residues which are targeted for mutagenesis are chosen and their side chains are removed (their main chain atoms remain part of the infrastructure). For example, one could try to fill the space that camphor left by mutating leucine 244 and valine 247. This is just an example; any other combination of residues may be chosen.

In Figure 2c two residues are chosen and their side chains are removed.

The remaining atoms of the protein define the cavity left by camphor and the two amino acid residue side chains. It can be considered a puzzle with a hole of defined shape in the middle (see Fig. 2d). Using the members of the rotamer library as puzzle pieces (see Fig. 2e), the program will now attempt to fill the gap in the protein.

4) The 'main chain' check:

First, every rotamer in the library is fitted in the position of each one of the chosen mutant amino acid residues, one at a time, substituting for the native residue. The program notes which rotamers overlap the sides of the cavity and which ones fit within the gap in the framework (see Fig. 2f). This is done by simple pairwise distance checks on each and every atom of the framework protein to each atom of the rotamer.

Each atom is considered a hard sphere of radii defined by the Ramachandran 'normal limit'[6] (see Table 2). Here, a fudge factor is introduced in order to accommodate the inflexibility of the protein and side chains. The summation of the overlap of all the atoms of the rotamer with the protein is compared to the fudge factor which I call the 'overlap criterion'. If the summation of the overlaps is less than the criterion, then that rotamer is considered

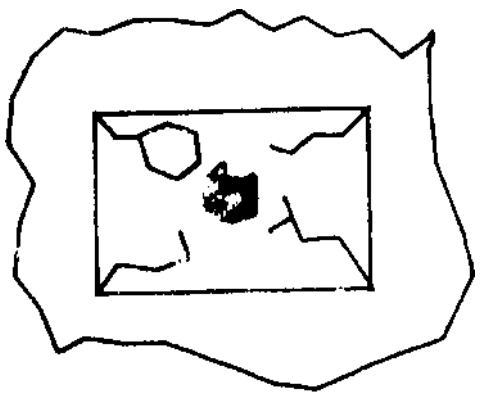


FIG. 2a

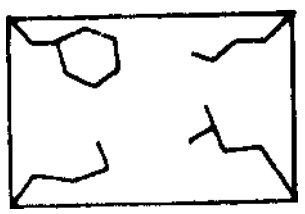


FIG. 2b

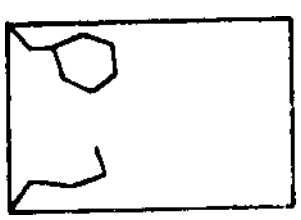


FIG. 2c

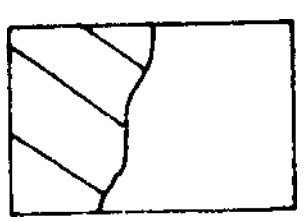


FIG. 2d

Rotamer Library (abridged)

Ala –

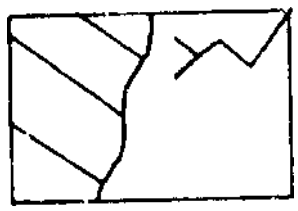
Asn 1

Asn 2

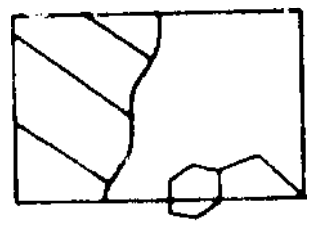
Phe 1

Phe 2

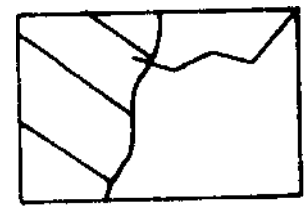
FIG. 2e



GOOD

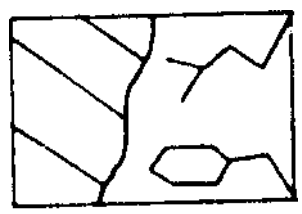


BAD

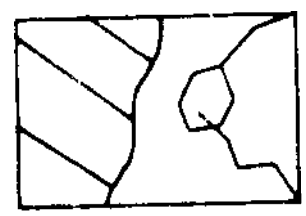


BAD

FIG. 2f



GOOD



BAD

FIG. 2g

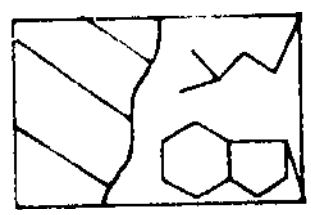


FIG. 2h

Table 2

Contact distances (in angstroms) used for overlap checks

	H	C	N	O	S	Fe
H	2.0	2.4	2.4	2.4	2.5	2.6
C		3.0	2.9	2.8	3.05	3.15
N			2.7	2.7	2.95	3.05
O				2.7	2.9	3.0
S					3.1	3.2
Fe						3.3

compatible with the framework. This overlap criterion is entered by the user and should be small (< 1 A). It has the effect of allowing one small overlap or several very small overlaps. It was found to be a necessary addition to the program so as not to make the model too strict. The determination of the overlap criterion is discussed further in results.

This 'main chain' check results in a table of rotamers which are compatible with the cavity as it is defined. For an example of such a table, see Table 4.

5) The 'side chain' check:

Here the rotamers allowed in the main chain check are substituted in the positions of the mutant residues and are pairwise checked for overlaps with each other (see Fig. 2g). The overlap criterion is again introduced to loosen the restrictions of inflexibility as in the main chain check.

The result of this test is an enumeration of allowed sequences which are compatible with the cavity and with each other. In the example, it would produce a list of all amino acid combinations which could replace the native leucine and valine without disrupting the structure of the protein. From this list one must choose the sequences which are useful to the problem at hand. In general, and in the example, the sequences with the largest volume are the most

useful as they fill the space in the shape desired (see Fig. 2h).

The limits of modern computation cause one more step to be necessary in this process. The side chain check can only handle running the permutation on four to five residues at a time, and thus larger mutation schemes must be broken into packing units. These units should be made of residues which are geographically close so as to reduce the eventual number of sequences which are produced. The program must then combine each packing unit combination together in a manner similar to the side chain check to see which packing unit combinations are compatible with each other. This final procedure results in the true list of allowed sequences for large mutations schemes. Again, from this list, the high-volume sequences should be chosen for further study. How one interprets the data from such a model is subject to whimsy. The discussion will delve into this topic in some depth.

V. Results

In all the following exercises, the crystal structure of the camphor-bound cytochrome P-450$_{cam}$, labelled 2cpp in the Brookhaven Protein Data Bank, was used.

1) Determination of Native Rotamers.

The torsion angles of all ten active site residues were analyzed and compared to the chi angles represented in the rotamer library. For each residue there was one rotamer which held a very similar conformation. For example, valine 247, which can be described by one chi angle, has a chi angle of 178.3 degrees. In the rotamer library (see Table 1), there are three rotamers for valine with chi angles of 173.5, -63.4, and 69.3. The first rotamer, or rotamer number 64, is very close in its conformation to the native valine 247, and thus rotamer 64 is defined as the 'native' rotamer. The native rotamers for all ten residues are shown in Table 3.

2) Determination of the Overlap Criterion.

As a first experiment, camphor was removed and all ten active site residues were chosen for mutation. Only the main chain check was run to determine how well the rotamers fit into the rather large cavity left by removing all ten

Table 3

The active site native rotamers and their overlap with the cavity wall

Residue	Native Rotamer	Cavity Overlap (angstroms)
Phe 87	46	0.015
Tyr 96	61	1.63
Thr 185	51	0.13
Leu 244	40	0
Val 247	64	0.04
Thr 252	51	0.13
Val 295	64	0
Asp 297	9	0
Ile 395	35	0
Val 396	64	0.20

Table 4

The main chain check for all ten active site residues with no overlap criterion correction

Residue	Allowed Rotamers
	1111111111222222222333333333
	0123456789012345678901234567890123456789
	ANNNNNNDDDCCCQQQQQQQEEEEEEEGHHHHHHIIIIIL
87	AN N D C Q Q EEE GH H L
96	A D C G H H
185	AN N C Q Q E G H
244	A N N D C Q G
247	A N DD CC Q E G H H
252	A D C E G H
295	A CC Q Q E E G H I
297	A N DC CQ E GH H
395	ANNN DD CC Q EEE G H H I
396	AN N D CC Q EE GH H

Residue	Allowed Rotamers
	4444444445555555555666666
	0123456789012345678901234
	LLLFFFFPSSSTTTWWWWWWYYYYVVV
87	F * SS W Y
96	S *
185	S *
244	L S
247	L SSS T W *
252	S *
295	SSSTT V
297	F S T W Y V
395	L F SSS T W V
396	F SS W Y *

Underlined = Native Rotamer

* - Native Rotamer Absent

side chains. With no overlap criterion correction, the map in Table 4 was produced. This map shows every rotamer allowed at each position which fit within the cavity walls. The results show that minor overlaps occur in six of the ten 'native' rotamers with the cavity walls (see Table 3). Except for Tyr-96, the largest of these overlaps was Val-396 which overlapped about 0.2 angstroms with the peptide oxygen of Asp-251. As these overlaps were minor (< 10% of the normal limit), the overlap criterion mentioned above was created to compensate for the inflexibility of the model used. Using the .2 A overlap of Val-396, I arbitrarily set the overlap criterion for this problem at 0.3 angstroms. With this criterion, all native rotamers were allowed in the main chain check except Tyr-96, which overlaps over an angstrom with the Thr-101 gamma oxygen. Again, for rather arbitrary reasons, I did not want to increase the criterion high enough to accommodate this large of an overlap. As Tyr-96 barely enters the active site and since tyrosine is a large volume amino acid, I decided not to consider the mutation of Tyr-96 any further. The only mutation I could suggest is simply to remove the hydrogen bond donor by mutating Tyr-96 to Phe as Atkins did. This would be the subject of a different kind of analysis than the one presented here and thus shall be ignored for the rest of this discussion.

3) A Trial Three Amino Acid Mutation.

As a first try, and also to present an easily executed mutation, Leu-244, Val-247, and Thr-252 were considered for mutation. Camphor and the three amino acid side chains were removed creating a cavity. Note: all the other active site residues were left in place as part of the infra-structure and cavity walls. With the overlap criterion at 0.3A, the main chain check and side chain check were run. The results of the main chain check are shown in Table 5. Using the volume values from Richards[1,7], the native volume of Leu, Val, and Thr adds up to 432 cubic angstroms. 2376 total combinations of mutations were allowed including the native sequence. About 300 of these sequences had volumes greater than the native volume, thus reducing the cavity volume. The highest volume sequences are presented in Table 6 along with Atkins's mutants for comparison of the residue volume change. As shown, the highest volume increase would result from leaving Leu-244 be, mutating Val-247 to Leu or Ile, and mutating Thr-252 to His. This would increase the residue volume by 70 A^3, partially closing the active site. These volume changes may be misleading, depending on the flexibility of the active site. Interpretation of these results appears in the discussion.

Table 5

The main chain check for the three residue mutation

Residue	Allowed Rotamers
	1111111111222222222233333333333
	01234567890123456789012345678901234567890
	ANNNNNNDDDCCCQQQQQQQEEEEEEEGHHHHHHHIIIIIL
244	A N N D C Q Q E G
247	ANNN DD CC Q E G I
252	AN N D C Q E G H
	44444444445555555555666666666
	01234567890123456789012345678901234567890
	LLLFFFFPSSSTTTWWWWWWYYYYVVV
244	L S
247	L SSS T V
252	SS T

Underlined = Native Rotamer

Table 6

The high volume sequences for the three resid utation

Amino Acid Sequence			Rotamer Sequence			Residue Volume Increase (A^3)
244	247	252	244	247	252	
Native:						
L	V	T	40	64	51	0
L	I	H	40	35	30	72
L	L	H	40	41	30	71
L	L	Q	40	41	18	65
L	L	E	40	41	22	59
L	Q	Q	40	14	18	58
L	Q	H	40	14	30	64
L	E	H	40	21	30	58
L	I	Q	40	35	18	66
L	I	E	40	35	22	60
Q	I	Q	16	35	18	59
			19	35	18	
Q	I	H	16	35	30	65
			19	35	30	
Q	L	Q	16	41	18	58
			19	41	18	
Q	L	H	16	41	30	64
			19	41	30	
E	I	H	25	35	30	59
E	L	H	25	41	30	58

Atkins's Mutants	Residue Volume Change (A^3)
247: Val -> Ala	-50
295: Val -> Ile	27

4) A Two Packing Unit Mutation Scheme.

Because of time constraints, I wanted to run a mutation scheme which was smaller than the entire nine residue active site. Two packing units of three residues each were chosen. The first packing unit was composed of the three residues run above: Leu-244, Val-247, and Thr-252. The second packing unit was composed of Thr-185, Val-295, and Val-396. These residues were chosen because they formed a self-contained, close-knit packing volume. The other three remaining active site residues, Phe-87, Asp-297, and Ile-395, each present some reason why they should not be chosen. Phe-87 is already a high-volume residue. Whether reducing its volume could allow a net residue volume increase would be the subject of later runs. Ile-396 tends to point away from the active site, and Asp-297 forms a hydrogen bond with the heme propionic acids. Although weak, these are some reasons to exclude these residues and concentrate on mutating the six remaining.

The program was run on these two packing units. The results of the main chain check are shown in Table 7. The program produced 2882 sequences for the 244/247/252 packing unit and 3513 sequences for the 185/295/396 packing unit. These sequences were then cross checked, resulting in the final allowed sequences. Six million allowed rotamer sequences were produced, 100,000 of which had greater net residue volume than the native sequence. The highest volume

sequences, which ran up to 120 cubic angstroms larger than the native sequence, are presented in Table 8. This large of a volume increase would certainly have a marked effect on the protein and should produce some interesting experimental results. Interpretation of these results appears in the discussion.

Table 7

The main chain check for the six residue mutation

Residue	Allowed Rotamers
	1111111111222222222333333333
	0123456789012345678901234567890123456789
	ANNNNNNDDDCCCQQQQQQQEEEEEEEGHHHHHHIIIIIL
185	AN N D C E GH H
244	A N N D C Q Q E G
247	ANNN DD CC Q E G H H I
252	AN N D C Q E G H
295	A N DD CCC Q Q E E G H I
396	AN N D CC Q Q EE GH H I

	4444444444555555555566666666
	0123456789012345678901234567
	LLLFFFFPSSSTTTWWWWWWYYYYVVV
185	F SST Y
244	L S
247	L F SSS T W V
252	SS T
295	SSSTTT VVV
396	F SSSTT W Y V

Underlined = Native Rotamer

Table 8

The high volume sequences for the six residue mutation

Amino Acid Sequence 244 252 396 247 185 295	Rotamer Sequence 244 252 396 247 185 295	Residue Volume Increase (A^3)
Native: L V T T V V	40 64 51 51 64 64	0
L N H F T H	40 1 30 43 52 29 40 3 30 43 52 29	124
L N Q F T H	40 1 18 43 52 29 40 3 18 43 52 29	118
Q N H F T H	16 1 30 43 52 29 16 3 30 43 52 29	117
L D H F T H	40 7 30 43 52 29	114
L N H F T E	40 1 30 43 52 25 40 3 30 43 52 25	112
L N E F T H	40 1 22 43 52 29 40 3 22 43 52 29	112
Q N Q F T H	16 1 18 43 52 29 16 3 18 43 52 29 19 1 18 43 52 29 19 3 18 43 52 29	111
E N H F T H	25 1 30 43 52 29 25 3 30 43 52 29	111
L N H F C H	40 1 30 43 10 29 40 1 30 43 11 29 40 3 30 43 10 29 40 3 30 43 11 29	108
L D Q F T H	40 7 18 43 52 29	108
Q D H F T H	16 7 30 43 52 29 19 7 30 43 52 29	107

VI. Discussion

Presented above is a model and some initial results for protein re-engineering. The model used is certainly not the optimum method with regard to accuracy or reality, but it provides a starting point and yields a low-order approximation to protein engineering problems. The results were only given for one protein and one problem; however, the model and the software are easily expandable to other proteins and other problems where space filling is a primary concern.

1) The Approximation.

Two main approximations were used, both of which reduced the computational problem immeasurably. The first was the concept of discrete side chain conformations or a rotamer library. Although somewhat young in origin, this concept has held fairly well under various scrutiny by Ponder and Richards. This approximation could be made better by further advances in the construction of the rotamer library, such as the addition of methionine, arginine, lysine, or other rotamers; and these are easily implemented in the software.

The strongest and most limiting of the approximations is the inflexibility of the main chain and the protein structure as a whole. Not even the slightest relaxation,

which would easily be accommodated in a real protein, can be considered by this software. This approximation most probably results in an overly strict model. In other words, this program will tend to produce only a subset of the true substitutions which would result in a stable protein. Small relaxations or movements in the main chain of the protein would allow larger volume sequences to be substituted into the native protein than predicted by the model in this software. And thus, the highest volume sequences given by this procedure used here may provide only a lower, softer limit to the mutation schemes which would truly be accommodated by the protein structure.

The strictness of the model inflicts both desirable and unfortunate effects on the outcome of the software. Being strict in the inflexibility of the main chain, the software will most probably produce loosely packed volume instead of the snug packing sequences found in nature. On the other hand, the severity of the approximation all but insures that the protein will fold without large distortions of the main chain infra-structure. Thus, the mutations suggested will most probably produce stable proteins, but the desired effects will not be as strong as they could be with more closely packed sequences.

2) Interpretation of the Data.

All these conjectures and 'most probables' look good on paper; yet they say nothing very strong about the true value of the suggested mutations above. How should the sequences produced be interpreted? If one were very faithful to the model, one could just go on to perform the mutation and characterization of the proteins. However, it would seem more intelligent to check each of the suggested sequences through a molecular graphics terminal or even molecular dynamics software. One should also consider the two other rules of protein structure put forth by Ponder and Richards:

3) Hydrogen bond donors and acceptors occur in donor-acceptor pairs within the interior of the protein.

4) Charged residues only occur on the exterior of the protein.

In the problem studied above, it is difficult to say how these rules should be applied to the heme pocket; however, these thoughts should be considered when deciding what mutations to perform.

I, personally, would be satisfied if any of the mutations suggested in the results produced a folding, stable protein with the heme intact. This would add some validity to the model used to design the molecule and would

also provide a basis for further modification of the active site of cytochrome P-450$_{cam}$. The question of whether a high spin, substrate-free protein can be produced is best answered by the construction of the proteins.

3) Further Experiments.

The most interesting possibility for this algorithm is in the construction of novel specificity in the active site. If one considers that most of the specificity of camphor binding is in the shape of the Van der Waals contacts surrounding camphor, then one may possibly construct a cavity geometry specific to many other molecules through mutagenesis. One might construct various geometries in the active site by providing the right rotamer sequence in the active site residues. However, this idea brings up several questions. If one chooses an arbitrary molecule, for example, ethyl benzene, is it possible to construct an active site of the right shape through changing only nine to ten residues, or does the entire main chain geometry of P-450$_{cam}$ have to be altered? It 's very probable that the main chain of P-450$_{cam}$ severely restricts the number of useful cavity geometries allowed. However, as one looks at the number of configurations allowed by mutating N different residues, the number would be on the order of ten to the N^{th} power. Changing ten amino acids allows nearly a trillion different cavity geometries; thus hope should be maintained.

To implem[illegible]h a problem using the current algorithm is simp[illegible] concept. If the exact position of the molecule for [illegible] the specificity was to be designed was known, then t[illegible]omic coordinates could be entered as part of the protein infra-structure. Then the program could be run exactly as with the cavity closing problem, except the cavity would be very irregularly shaped. The cavity to be filled would be defined as the space left by removing the chosen mutant side chains which surround the substrate molecule. Then the highest volume sequence would be found which fits in the cavity. In the ideal case, this sequence would then fill almost all the space surrounding the substrate providing a Van der Waals cocoon as the native sequence does for camphor.

This simple concept has a few technical problems. First, the exact orientation of the substrate molecule is probably not known, and various orientations should be checked to optimize the fit. This adds a dimension to the computation which may prove difficult. With a successfully designed sequence, the flexibility of the active site may ruin the beautiful geometry of the design in a real protein. Here, the static approximation is most stringent, and energy minimization simulations should be run to complement this somewhat simplistic model.

All cautions aside, this model provides a very handy stepping stone into an otherwise very difficult task. The

utility of this algorithm is just now being studied, and there are a multitude of possibilities which have not been exhausted.

VII. References

1. Ponder, J., and Richards, F. M. (1987) J. Mol. Biol. 193, 775-791

2. Atkins, W., and Sligar, S. (1989) J. Amer. Chem. Soc. 111, 2715

3. Unger, B., Gunsalus, I. C., and Sligar, S. (1986) J.Biological Chemistry 261, 1158-1163

4. Poulos, T., Finzel, B., and Howard, A. (1987) J. Mol. Biol. 195, 697-700

5. Sligar, S., and Gunsalus, I. C. (1976) Proc. Natl. Acad. Sci. 73, 1078-1082

6. Ramachandran, G. N., and Sasisekharan, V. (1968) Advan. Protein Chem. 23, 283-438

7. Richards, F.M. (1977) Annu. Rev. Biophys. Bioeng. 6, 151-175