PROTEIN DESIGN

EXERCISES

EMBL, 1986

EMBL BIOcomputing Technical Document 1 1987

Report on the

EMBO Practical Course

Protein Design on Computers

Proteinbasteln / Bricolage des Proteins / Protein Tinkering

Heidelberg, September 1 - 19, 1986 at EMBL

*Protein Cartoons by **Arthur Lesk***

*Contact Maps by **Michael Scharf***

*Special Appendix by **G. Baumann and Cornelius Froemmel***

*Editorial Assistance by **Miri Hirshberg and Christine Raulfs***

*Edited by **Chris Sander***

This document is a work report of design exercises performed at a three-week workshop and not a presentation of finished scientific results nor a polished publication.

••••••••••••••CAUTION : USE ONLY AS DIRECTED••••••••••••••

Paper copies of this report are available from:
Christine Raulfs, BIOcomputing, EMBL, 6900 Heidelberg, Germany.

Coordinate data sets and related files can be obtained as electronic bitnet mail from the EMBL fileserver by sending the one-line command message "send filename.ext".

INTRODUCTION :

Inverting the protein folding problem

Choice of design exercises

Narrative of the course

Names, relatives and files of designed proteins

EMBL file server

People

E-mail addresses

Software

Inverting the protein folding problem by chris sander

The protein folding problem has remained essentially unsolved in spite of many years of effort. Physical simulations at present take too much computer time or fail to correctly simulate the essential interactions. Statistical or empirical rules for structure prediction are generally weak. Three principal difficulties lie in the **cooperativity** of the folding process, in the **delicate energetic balance** between enthalpy and entropy and in the fact that protein sequences have evolved to satisfy **diverse functional constraints** and are not optimized to merely fold up correctly.

In the face of these difficulties new approaches are needed. Here, we decided to break new ground by simply turning the question upside down. "Ask not to calculate the structure from the sequence, but ask instead to calculate a sequence from the structure". So, protein design in this context is protein folding turned upside down and structure prediction becomes **sequence prediction.**

Inverting the problem leads to **new ways** of thinking about protein folding. Consider these questions: "Enumerate all protein sequences compatible with a given fold!", or "Design a sequence that will fold up like lysozyme!"; or "How much sequence variation at the surface of a globular protein is compatible with maintaining the chain fold?". In addition, the inverted problem may actually be **easier to solve** than the forward problem. First, the forward problem has only one essentially correct solution, namely the right structure. The inverse problem, if one is satisfied with *one* sequence that folds up as planned, has many solutions. We know that in nature a protein family typically has many different sequences of essentially identical three-dimensional structue. Second, and more importantly, sequence prediction needs pay no heed to functional requirement; the sequence can be **optimized** for the sole purpose of yielding the correct **structure** in vitro.

To illustrate the possible simplicity of a designed sequence compared to a natural sequence, consider the following example. Suppose you want to construct an all-α-helical protein with a certain given topology. You have at your disposal tables of single residue preferences for certain positions in secondary structures, in loops, in interfaces between secondary structures and on the surface; also, you have tools to optimize packing by removing steric clashes. Suppose you then simple-mindedly choose for each position residues optimal merely from the point of view of single residue preferences, naively ignoring one of the principal rules of protein folding *"local sequence information is insufficient to fully determine the native fold"* or *"the structure of oligopeptides as long as five residues is sometimes determined by the surrounding tertiary structure interactions"*. Is it not conceivable that such a protein sequence, although clearly less complicated than natural sequences, may well do the job ? Is it not conceivable that natural sequences are more complicated because they have been subjected to additional functional selective pressure and randomizing mutational events ? We shall see in due time.

Protein design as sequence prediction is a new approach to the old problem of protein folding. To the extent that designed proteins are not subject to the same evolutionary constraints as are natural proteins, the protein design problem transcends the classical protein folding problem.

Choice of Protein Design Exercises

by chris sander

The exercises in this course were designed to get started exploring the methods for *de novo* design. One cannot hope to produce carefully evaluated designs in three weeks, but one can assess the state of the art and determine the weaknesses of current methods and current thinking.

The problem in each case (with two exceptions, FXNM and RCU1) **was to start from an idealized form of a well-known protein fold and to design one sequence that would produce that fold.**

Accordingly there were three main steps:

A: choose a typical protein fold (say, a βαβ nucleotide binding fold) and idealize (simplify) its architecture. Incorporate if possible a binding site for a ligand the binding of which can be detected experimentally. The result is a backbone coordinate set that clearly belongs to a well-known structural class yet is unlike any of the natural members of the class in detail.

This first step is not necessary for the sequence design problem, but very useful as an exercise to familiarize students with structural principles.

B: invent, calculate, predict an appropriate amino acid sequence.The result is a full coordinate set (all atoms) of the designed protein and a list of its amino acid sequence.

C: critically evaluate the designed protein by as many intelligent methods as possible Most of the evaluation is done by a group not involved in its design. The result is a statement about the quality of the design and an assessment of the chances for a successful experiment.

Narrative of the protein design course

In an intensive three-week workshop held at EMBL in Heidelberg six groups each designed at least one protein sequence and its corresponding three-dimensional model structure by computer graphics and protein folding techniques. The designs were aimed either at a particular binding function or a particular structure. Five of the designs were completely de novo, echoing a natural protein structure motif, two were point-mutation variants of natural proteins, one was a topological rearrangement of a natural protein with some sequence changes.

The results of each group were shared and discussed with the other five groups and this report places them in the public domain, for use in teaching and research. One of the designed protein sequences has been synthesized, cloned and expressed by Steve Emery and Helmut Bloecker at the GFB, Braunschweig. Structural studies are planned.

The following protein sequence-structure pairs were designed:

(1) Two TIM barrels.

TINY	**Tiny Tim**, a small TIM barrel
BABA	**Babarellin**, a fourfold symmetric α/β/α/β barrel

(2) Two alternating α/β proteins.

FXNM, FXNI	**FXNMut** and **FXNIdeal**, mutated and idealized flavodoxin
BEAL	**Betalphacin**

(3) Two α-helical bundles.

BUND	**Bundle**, an idealized four-α-helix bundle that binds calcium
RCU1, RCU2	**CopRop1** and **CopRop2**, two α-helical bundles that bind copper

State-of-the-art software was used during the course: for molecular graphics and model-building, sequence searches, structure prediction, structure evaluation, energy minimization and molecular dynamics: FRODO, UCSF-MIDAS, HYDRA, INSIGHT, BRUGEL, UWGCG, GROMOS, ROBSON-PRED etc. Hardware available at the European Molecular Biology Laboratory or on loan were a VAX cluster (8600/785), a microVax II, an Evans and Sutherland MPS and PS350, and a Silicon Graphics IRIS. On-line data banks were EMBL/GenBank DNA sequence libraries, PIR/NBRF protein sequences, PDB protein structures, DSSP secondary structures.

Organizer: Chris Sander
Secretaries: Anke Hennemann and Christine Raulfs
Bitnet/Earn: SANDER@EMBL, RAULFS@EMBL
Telephone: 0049-6221-387361

Abteilung Biophysik, Max Planck Institute of Medical Research, Heidelberg and BIOcomputing Programme, EMBL, Heidelberg

Designed Protein Id / Group

TINY	A.Jones, J.de Vlieg, M.Eliasson, M.Hirshberg, C.Sander
BABA	J.Richardson, P. Argos, D.Kneller, D.Osguthorpe, M.Scharf
FXNI, FXNM	A.Lesk, O.Herzberg, M.Schaefer, Z.Wassermann
BEAL	D.Richardson, F.Colonna-Cesari, E.v.Cutsem, K.K. Mortensen
BUND	J.Moult, C.Froemmel, J.Postma, A.Skerra, A.Valencia
RCU2, RCU1	B.DeGrado, T.Hubbard, J.Reichelt, C.Woodward

Closest natural relatives of the designed proteins

TINY Tiny Tim	Triose Phosphate Isomerase (TIM)
	Pyruvate Kinase (PYK)
	Glycolate Oxidase (GAO)
	Taka Amylase (TAA)
BABA Babarellin	TIM, PYK, GAO, TAA
FXNI Idealized Flavodoxin	Flavodoxin (FXN)
FXNM Mutated Flavodoxin	Flavodoxin (FXN)
BEAL Betalphacin	Flavodoxin (FXN)
BUND Bundle	repressor of primer (ROP), hemerythrin
RCU2 CopRop monomer	ROP
RCU1 CopRop dimer	ROP

Current and Original File Names

$PDB:TINY.BRK_MODEL	TINY_TIM.BRK;4
$PDB:BABA.BRK_MODEL	BABARELLIN_JDDMP.BRK
$PDB:FXNI.BRK_MODEL	FXN_IDEAL19.BRK;1
$PDB:FXNM.BRK_MODEL	FXN_MUT9EM.BRK;1
$PDB:BEAL.BRK_MODEL	BETALPHACIN.BRK;2
$PDB:BUND.BRK_MODEL	BUNDLE.BRK
$PDB:RCU2.BRK_MODEL	ROPCU2.BRK;1

Bitnet fileserver requests should use the first file name.

How to Obtain Files from the EMBL Fileserver

On VAX/VMS systems connected to bitnet do:

```
$ send /remote DHDEMBL5 NETSERV send $pdb:xxxx.BRK_MOD
$ send /remote DHDEMBL5 NETSERV send $pdb:xxxx.DSSP_MOD
```

where xxxx = Protein Id, like xxxx = TINY.

People

Teachers:

Jane Richardson	Duke U	US
Alwyn Jones	BMC, Uppsala	S/GB
Arthur Lesk	MRC Cambridge	GB/US
John Moult	U of Alberta	CND/GB
Bill DeGrado	DuPont, Wilmington	US
Dave Richardson	Duke U	US

Participants:

Francois Colonna	CNRS, U at Orsay	F
Cornelius Froemmel	Humboldt U	GDR
Clare Woodward	U of Minnesota	US
David Osguthorpe	U of Bath	GB
Zelda Wassermann	DuPont, Wilmington	US
Don Kneller	UCSF	US
Jacob de Vlieg	U of Groningen	NL
Miriam Hirshberg	Weizmann Institute	IS
Arne Skerra	Uni Muenchen	D
Kim Kusk Mortensen	U of Aarhus	DK
Joachim Reichelt	GBF, Braunschweig	D
Eric van Cutsem	ULB, Bruxelles	B
Alfonso Valencia	CSIC, Madrid	ES
Margareta Eliasson	RIT, Stockholm	S
Timothy Hubbard	Birkbeck C, London	GB
Osnat Herzberg	U of Alberta	CND/IS

EMBL/MPI

	Pat Argos	US
	Johan Postma	NL
	Heinz Bosshard	CH/US
	Michael Scharf	D
	Michael Schaefer	D

Secretaries:

	Anke Hennemann	D
	Christine Raulfs	D

Organizer:

	Chris Sander	D

Bitnet addresses

$	Arne Skerra	SKERRA@DM0MPB51
$	Alwyn Jones	ALWYN@SEMAX51
$	Kim Kusk Mortensen	KEMBIOS@DKARH01
$	Miri Hirshberg	CFHIRSH@WEIZMANN
$	John Moult	USERJMCP@UALTAMTS
$	Osnat Herzberg	USEROSNT@UALTAMTS
$	Don Kneller	KNELLER@UCSFCGL
$	Jakob de Vlieg	VLIEG@HGRRUG5
$	Jane, Dave Richardson	DXRAY@TUCC
$	Joachim Reichelt	JRE@VENUS.BRAUNSCHWEIG-GBF.DFN
$	David Osguthorpe	CH_DJO%UK.AC.BATH.UX63@AC.UK
$	Arthur Lesk	LESK@EMBL
$	Chris Sander	SANDER@EMBL.BITNET
$		SANDER@DHDEMBL5.BITNET

Software

Software	Authors	Local Experts (if different)
AMBER	Peter Kollmann +	Zelda Wasserman, Joachim Reichelt
AMBRUN	Zelda Wasserman	
BRUGEL	Philip Delhaise +	Eric van Cutsem
DSSP	W. Kabsch, C.Sander	
FRODO	Alwyn Jones	Heinz Bosshard, Johan Postma
GROMOS	van Gunsteren	Jacob de Vlieg, Johan Postma
INSIGHT		David Osguthorpe
HYDRA	Rod Hubbard	
MAXSURFACE	Cornelius Froemmel	
PACANA	John Moult	
UCSF-MIDAS	UCSF/Langridge	Heinz Bosshard, Johan Postma

Lectures

September

1. Mon	Jane Richardson: Duke Univ., Durham, NC	Anatomy of protein structure - a survey
2. Tue	Arthur Lesk: MRC Cambridge, EMBL	Structural changes in protein evolution
3. Wed	Alwyn Jones: BMC, Uppsala	Use of known substructures in protein design
3. Wed	John Moult Univ. of Alberta Edmonton, Canada	Systematic Conformational Search: The Determination of the Structure of Loops in Proteins
4. Thu	Dave Richardson Duke University Durham,N.C.	Betabellin
4. Thu	Osnat Herzberg Univ. of Alberta Edmonton, Canada	Structure of troponin C and muscle function
5. Fri	Bill DeGrado: DuPont, Wilmington, DE	Design of alpha-helical proteins
9. Tue	Pat Argos	Detecting weak homologies among proteins
12. Fri	Clare Woodward	Hydrogen exchange kinetics and the dynamic struture of proteins
15. Mon	Rainer Frank EMBL, Heidelberg	Chemical synthesis of peptides
16. Tue	Jacob de Vlieg	Restrained molecular dynamics procedure for protein structure determination from NMR data; a lac repressor headpiece structure based on constraints from the presence and absence of NOE's and J-coupling
17. Wed	Helmut Bloecker GBF, Braunschweig	Chemical synthesis of genes

GROUP REPORTS

TINY tiny tim
BABA babarellin

FXNI, FXNM shades of flavodoxin
BEAL betaalphacin

BUND bundle
RCU2 coprop

Alwyn Jones
Jacob de Vlieg
Margareta Eliasson
Chris Sander
Miriam Hirshberg

* Tiny-TIM. *
(TINY)

Design Task

We decided to design a simplified symmetrical TIM barrel. Inspection of TIM (triose phosphate isomerase) showed TIM to be much less symmetrical than any of us thought (this was because our knowledge of the protein was based on simplified sheet and helix drawings). As the central beta strands are not symmetric (they are more like two 4-stranded sheets packed on top of each other) we decided on a construction strategy that would maintain TIM's beta barrel shape.

Steps in Design

1. The basic unit was constructed from residues 89 to 127 in TIM. In the loop region 93 to 107, we replaced 9 residues by the loop from 41-46. This loop is like the one described by Schellman [1]. It connects an alpha helix to a beta strand and terminates the helix with a glycine in a left handed helix conformation. The beta to helix loop was kept since it had a moderate cluster of database loops with FRODO's DGNL. This gave us a 30 residue beta-loop-helix-loop-beta unit (89-127). We found that TIM's 108-126 (19 residues) matched FXN3_15 to within .80 Å rms deviation.

2. We constructed the first model by taking this unit and applying 7 least squares fits to the central beta strand residues in TIM. We used a displacement scheme of forward 2,0,2,0,2,0,2 along the beta strands to arrive at a total displacement of 8 residues around the whole barrel as follows:

Tim	Our Fragment	Rms	Shift
123-128,162-166	89-94,123-127	0.77	+2
160-165,206-210	·	1.27	0
206-211,229-233	·	1.15	+2
227-232, 8- 12	·	0.51	0
8- 13, 41- 45	·	1.53	+2
39- 44, 61- 65	·	1.60	0
61- 66, 91- 95		2.4	+2

The resulting structure has each beta strand overlaid by an extra strand. The extra ones were deleted in Frodo. The helix-helix distances in this structure showed 4-fold symmetry, but were uneven. We tried to even them out by measuring the distances between the tops, bottoms of helices, helix center to beta strand, across the top of the barrel, averaging them and using the averages as restraints for a dynamics run in GROMOS. The result was considered not suitable for continuing modeling. We therefore made the structure more even by moving each helix manually on the display.

3. The internal barrel residues were added in eight layers, with two staggered rings of four residues in each layer, trying to keep 4-fold symmetry. We had checked TIM,GAO,TAA and found that the barrel residues are mostly Val, Ile, Gly while the barrel/helix residues are branched hydrophobics. The top and bottom of the barrel were filled with hydrophilic residues without the 4-fold symmetry. The barrel/helix was randomly sprinkled with forked hydrophobics.

4. Bad contacts were checked by calculating residue energies (MaxTwist, Chris Sander) and coloring residues according to their energies as blue-to-red. Worst contacts were removed by hand.

5. Energy minimization was performed using the GROMOS package. To compensate for the shielding effect of the solvent we neutralized all charged atom groups. We used a cutoff of R_C=8 Å beyond which no interactions were included. The neighbour list of the nonbonded atom pairs was updated every ten cycles of minimization. The strain in the molecule was relaxed by performing a few hundred steepest descent steps, followed by a few hundred conjugate gradient steps. The overall rms deviation was 0.4 Å. Comparing to the initial structure only the side chains moved during the minimization.

6. We then modified the second beta-loop-helix unit, picking a pro-gly turn between the beta and alpha and sprinkling the outer helix surface with side chains, guided by the lists of single residue secondary structure preferences (Palau and Argos [2], Kabsch and Sander [3]).

7. Careful checking showed an error. Using the Gly after the helix as the count end point gave lengths of 23,25,23,24,23,24,23,23 residues for each unit. The unit just modified had one residue too much in the beta/alpha loop. We therefore checked another beta/alpha loop, the one before helix 4 and found res 77-86 similar to CRN1_3 over (11 residues). Judging by the way in which the two alternative loops merge into the pleat of the beta sheet, the crambin loop turned out to be the better one. We decided to use this turn on each unit, such that each unit is 24 residues long. We deleted one residue in the loop before helix 2 and added one in the loops before helices 1,3,5,7. The first beta-helix unit was excised and duplicated onto the three other odd helices (3,5,7); and beta-helix 4 was duplicated onto the other even helices (2,6,8). Magically we made no mistakes. This model has 48 residues in its repeat, has the same turns at each beta-alpha, and for each alpha-beta.

8. The changes in the framework required refining of our barrel residues. Then we completed replacing the side-chains of the loops and of the surface of the helices. The beta-helix and outer helix surfaces have 8-fold sequence symmetry. This structure was energy minimized again. After 150 steps of minimization the total energy dropped from +0.3E+5 to -0.8E+5 Kjoule/mol and the overall rms was 0.3 Å.

Final Model

Our tiny-TIM consists of 192 residues, each loop-beta-turn-helix unit is 24 residues long.

Appendix: Helix-Helix Distances. Since the barrel is egg-shaped the distances between neighboring helices differ as follows:

* Distances between the nitrogen of proline at the N-terminus of the helices.

Residue numbers	Distance in Å
10- 34	10.3
34- 58	12.2
58- 82	16.2
82-106	13.0
106-130	11.3
130-154	13.9
154-178	14.7
178- 10	13.0

* Distances between the nitrogen of glycine at the C-terminus of the helices

Residue number	Distance in Å
1- 24	13.3
24- 48	12.8
48- 72	11.1
72- 96	11.3
96-120	10.1
120-144	10.3
144-168	10.9
168- 1	10.7

Appendix: Surface and Volume

* Molecular surface accessibility: 10022.0 Å^2

* Free volume residue packing analysis; Volume change by group type

		TIM	TINY.TIM
MAIN	N	3.96	3.72
MAIN	CA	2.91	2.43
MAIN	C	1.21	1.05
MAIN	O	10.43	10.84
SIDE	CB	6.18	5.55
SIDE	CG	7.55	7.31
SIDE	CD	8.47	7.22
SIDE	CE	7.18	8.91
SIDE	CZ	6.54	6.22
SIDE	CH	9.96	0.00
SIDE	O	10.93	9.79
SIDE	N	10.05	9.42
SIDE	S	11.95	0.00

References

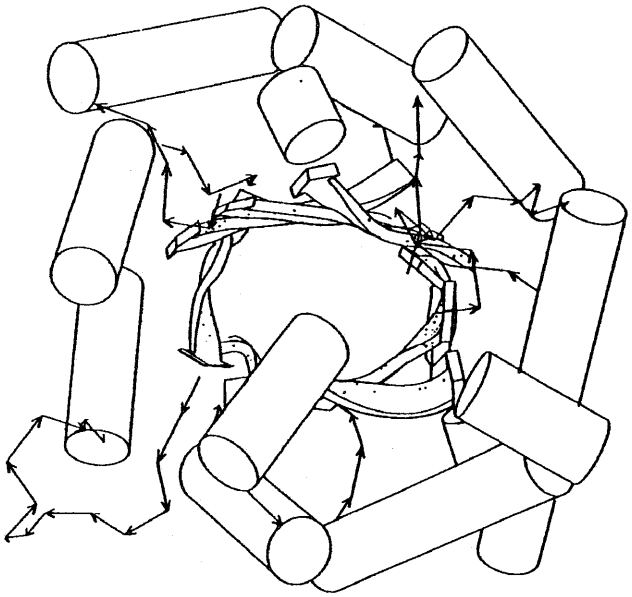
[1] C. Schellman, Protein Folding, Rainer Jaenicke Ed., 1980, Elsevier/North-Holland Biomedical Press 53-61.

[2] P. Argos and J. Palau, Int. J. Peptide Protein Res., 19, (1982) 380-393.

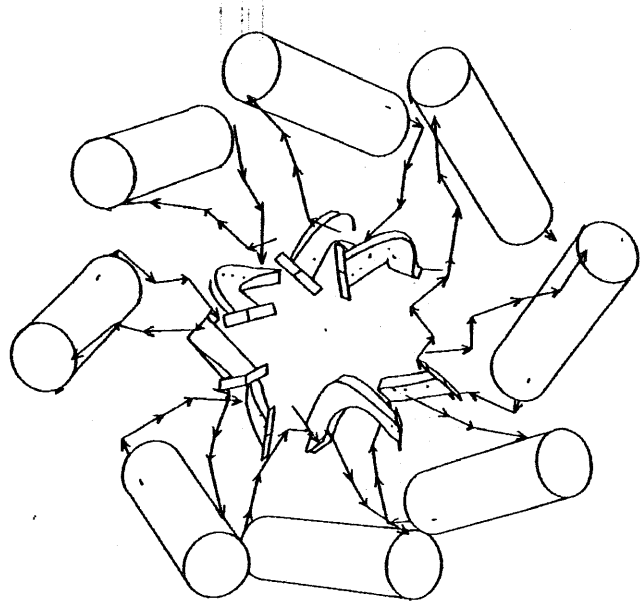
[3] W. Kabsch and C. Sander (private communication).

```
1)  Tiny Tim.....TINY
    SHEET....   AAAA                          AAAA                          A
    BRIDGE2..    b                             cc
    BRIDGE1..   aaaa                          aaaa                          c
    CHIRALITY   --+----+-+ +++++++++ ++-+-----+- -+++++++++ ++++++-+--
    BEND.....         SSS SSSSSSSSSS SSSS         SSSSSSSSS SSSSSSS
    5-TURN...                       > 5555<
    4-TURN...          >> >>XXXXXXX< <<<           >>>>XXXX XXXX<<<<
    3-TURN...          >3 3<        >33<
    SUMMARY..   EEEE  SSH HHHHHHHHHH HHTT EEEE   SSHHHHHHH HHHHHHH  E
    EXPOSURE.  4000291962 550**00840 7*47191000 1868751**0 0*503*7422
  1 SEQUENCE.  AGVVITVSNP AEFQKALDQA LKDGARVIIQ VSNPAEFQKA LNQALKNGAV

    SHEET....  AAA                           AAA                           AAA
    BRIDGE2..  ddd                           ee                            ff
    BRIDGE1..  c                             ddd                           ee
    CHIRALITY  ++---+-+++ +++++++++ -+++++++-+ -+++++++++ ++++-+----
    BEND.....       SSSSS SSSSSSSSSS S       S SSSSSSSSSS SSSSSS
    5-TURN...             >5555< >55 55<                    >5555<
    4-TURN...        >>>> XXXXXXX<<< <          >>>>XXXXXX X<<<<
    3-TURN...        >33<         >3 3<         >33<          >33<
    SUMMARY..  EEE  SSHHH HHHHHHHHHH TT  EEE  S SHHHHHHHHH HHHHTT EEE
    EXPOSURE.  00195*8*62 3**009707* 971700053* 787818*217 *02*751000
 51 SEQUENCE.  LVIEVSNPAE FQKALNQALK DGAEVAVYVS NPAEFQKALN QALKNGAVVV

    SHEET....                        AAAA                          AAA
    BRIDGE2..                         ggg                           h
    BRIDGE1..                        ff                            ggg
    CHIRALITY  +--+-+++++ +++++++++ --------+-+ ++++++++++ ++-+------
    BEND.....     SSSSSSS SSSSSSSSSS        SSS SSSSSSSSSS SSSS
    5-TURN...                  >5555 <                    > 5555<
    4-TURN...       >>>>XX XXXXXX<<<<          >> >>XXXXXXX< <<<
    3-TURN...                  >33<            >3 3<        >33<
    SUMMARY..     SSHHHHH HHHHHHHHHT   EEEE SSH HHHHHHHHHH HHTT EEE
    EXPOSURE.  083*765318 *108603*86 1601032*7* 822**11*60 3*5611001*
101 SEQUENCE.  ITVSNPAEFQ KALNQALKDG ATVIVEVSNP AEFQKALNQA LKNGAVVAIQ

    SHEET....                        A A
    BRIDGE2..
    BRIDGE1..                        h b
    CHIRALITY  ---++++++ ++++++-+-- ---+-+-+++ +++++++++
    BEND.....   SSSSSSSSS SSSSSSSS         SSSS SSSSSSSSS
    5-TURN...        >555 5< >5555<
    4-TURN...    >>>>XXXX XXX<<<<           >>>> XXXXXXX<<< <
    3-TURN...                  >33<
    SUMMARY..   SSHHHHHHH HHHHHHTT   B B   SHHH HHHHHHHHHH
    EXPOSURE.  4*986509*0 05*06**815 00042*826* 07*106903* *
151 SEQUENCE.  VSNPAEFQKA LNQALKDGAT VAVYVSNPAE FQKALNQALK N
```

Notation: SUMMARY H=ALPHA-HELIX
E=BETA-STRAND
B=BETA-BRIDGE G=3-HELIX I=5-HELIX
T=3-, 4-, OR 5-TURN
S=BEND

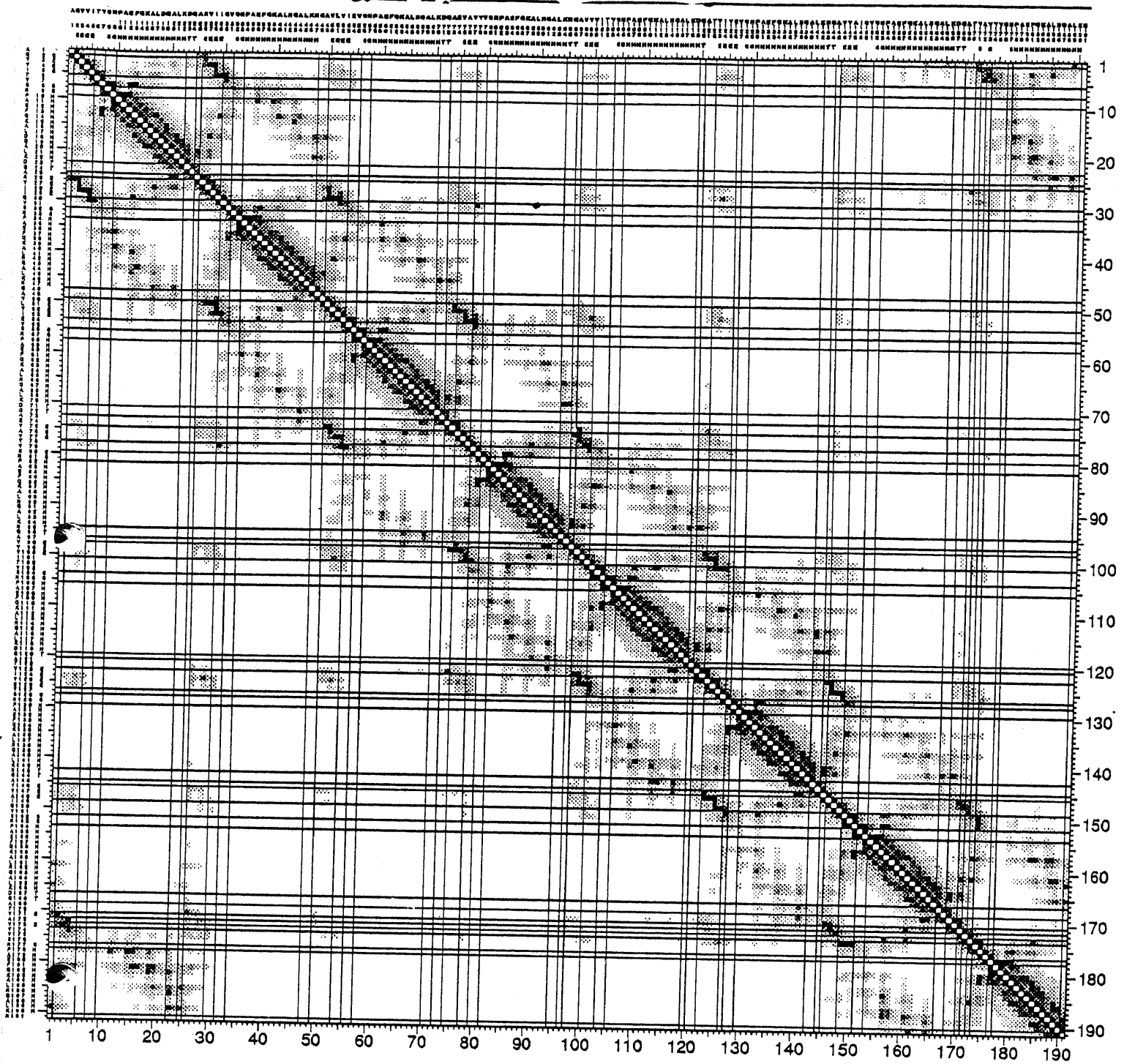


Chicken TIM



tiny TIM

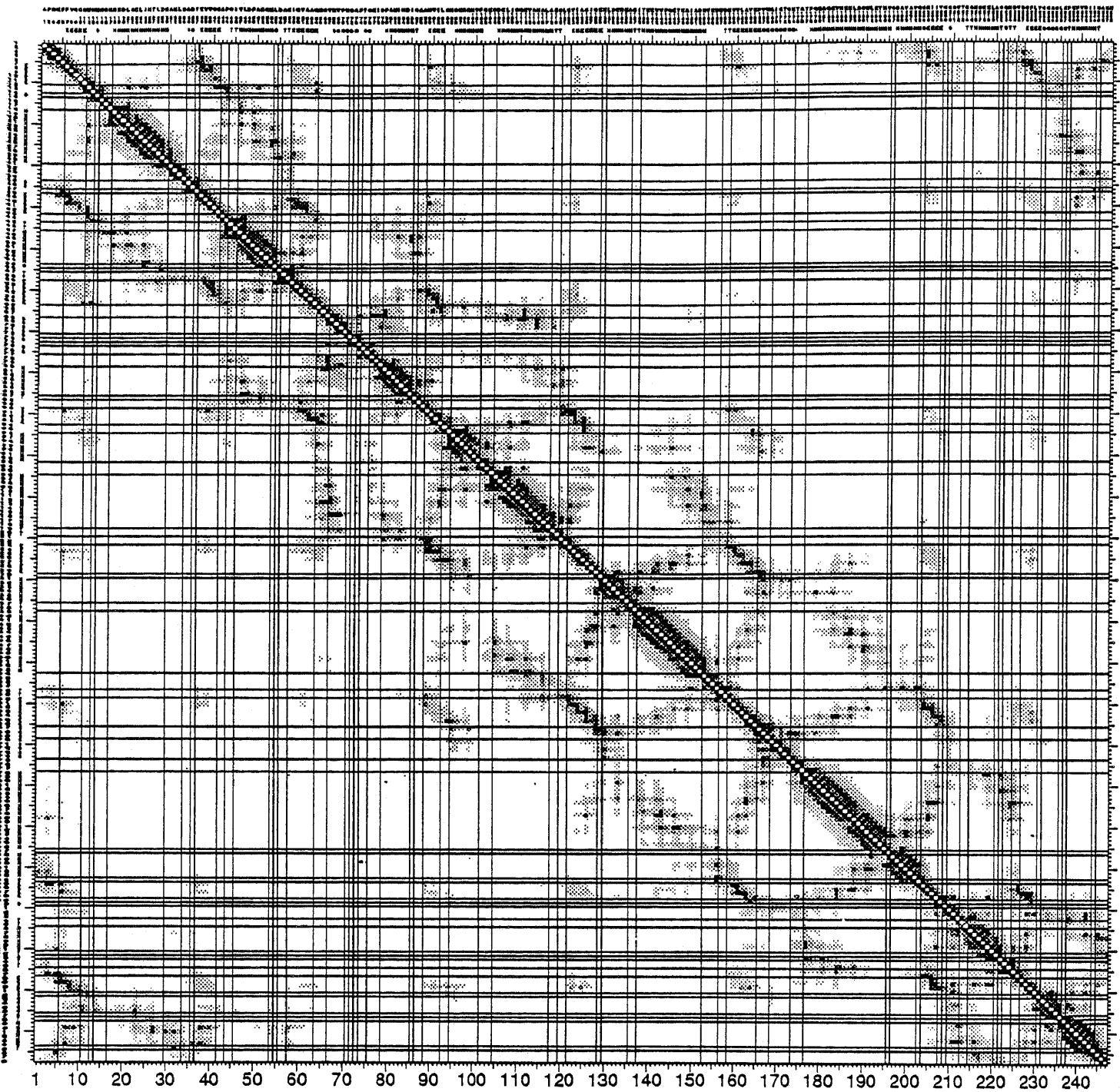
TINY: Van der Waals energy.



kcal/mol

0 -0.5 -1.0 -1.5 -2.0 -2.5 -3.0

TIM: Van der Waals energy.



al/mol

-0.5 -1.0 -1.5 -2.0 -2.5 -3.0

Jane Richardson
Patrick Argos
Don Kneller
David Osguthorpe
Michael Scharf

* BABARELLIN *
(BABA)

A DESIGN FOR A 4-FOLD SYMMETRIC TIM BARREL

Design Criteria

*8-strand, 8-helix, right-handed crossover, singly-wound parallel alpha/beta barrel.

*Small size (as small as consistent with presumed structure).

*Simplicity.

*Fourfold symmetry:
Exact in the sequence, although possibly with one or two extra N and/or C terminal residues. Exact in conformation for the starting model, but allowed to relax whenever necessary, and presumed inexact in the real protein.

*Geometry of strands and helices and their relations, side chain occurrence and conformation, internal packing, and surface properties either within the range or idealized versions of those observed in known proteins of this type (TIM, TAA, GAO, PYK, and KGA).

*Reasonable secondary-structure prediction by several methods.

*No statistically significant homology with known native sequences.

*Reasonable behavior in energy minimization; elimination of bad contacts.

*Prevention of alternative folding as doubly-wound parallel alpha/beta structure - primarily by uniform strand hydrophobicity and steric effects of helix packing.

*Feasible synthesis by genetic methods.

Computer Tools Used

*INSIGHT	display, change sequence, fit side chains, superimpose fragments.
*FRODO	especially DGLP (loop building from data base)
*Universe (DR)	display. (PS300)
*MIDAS	(on the Iris) coupled rotation of symmetrically related torsion angles
*FRODO	(on the MPS) display of native structures when PS300's not available.
*ARPLOT (AL)	stereo plots
*VFF (DO)	energy minimization, building ideal structures, finding bad contacts.
*Pacana (JM)	finding holes in native-protein TIM barrels and in our models.
*Angana (JM)	lists of phi, psi, chis
*Mutate (JM)	changing residues off graphics
*Contac (FC)	lists alpha/beta contact residues
*SGEOM (DK)	angles and distances between lines rms-fit to strands and helices.
*LGEOM (DK)	lists loops by length and gives crossing angle, twist, and tilt.
*DSSP (CS)	secondary structure assignments, exposure, phi psis.
*Prediction(MS)	secondary structure prediction, hydrophobicities; interactive; response to sequence changes.
*FASTP	sequence homology search

Analysis of known Structures

The first stage was an examination of known native proteins with TIM barrels:

Complete coordinates	File name
Chicken Triose Phosphate Isomerase	1TIM.brk
Taka-amylase	2TAA.brk
Glycolate oxidase (courtesy of Alwyn Jones)	GAOT.brk
Alpha-carbon coordinates	
KDPG aldolase (courtesy of Al Tulinsky)	KGA.dat
Pyruvate kinase (not aligned to new sequence)	1PYK.brk

The structures were examined and the following features extracted:

*The assignment of beta strands 1-8 and helices 1-8 (using graphics & DSSP).

*The assignment of residues in the interior and exterior of the barrel (graphics).

*H-bond diagrams: Offset of hydrogen-bonding around the barrel (due to twist): must be an even number if H-bonding is perfect; must be a multiple of 4 in order to make a 4-fold repeat in our structure. PYK is hard to assign because one strand is turned and doesn't H-bond on one side; TIM, GAO, TAA, and KGA all have an offset of 8.

*Analysis of strand twist, barrel flattening, etc., and which proteins are most nearly ideal for which properties: TIM is most flattened; TAA quite round and well H-bonded all around, but one side flared out conically; PYK has most regular helix arrangement; crossing angle of opposite strands typically around -65°; structures do not usually superimpose best with matched strand numbers; geometry varies a lot, some of it explainable by effects of additional domains.

*Tabulation of distance out and crossing angle between strand-helix pairs in the native structures, using SGEOM.

*Analysis of hydrogen bonding patterns using H-bond diagrams, showing which residues form spirals down barrel interior, which sets of 4 residues are symmetry-related in horizontal layers.

*Stereo plots of barrel interiors.

*Analysis of the internal packing in the barrels: Average side chain volume in the central 4 or central 5 internal layer is 59 +/- 10 Å3; i.e., a bit less than Val. Layers interpenetrate; typically 3 side chains and a Gly per layer. Graphics and PACANA both find large holes inside barrels, some opposite. Gly's that look as though they could be larger and have normal phi and psi's - question: What are those Gly's doing? Last layer at top or bottom typically has one or more saltbridges. Lists of residues that make up barrel interior, barrel outside that contact helices, and helix inside that contact sheet - used for coloring on display and for tabulating residue frequencies in each class.

*Average phi,psi backbone dihedral angles for barrel strands - huge scatter, center around phi=-114, psi=+124 -. Plot of CA(n-2)-CA-CA(n+2) bend angles (kappa in DSSP); for well-behaved strands very small in central region. Analysis of symmetry requirements and possibilities - 8-fold completely impossible, 2-fold clearly possible, 4-fold dubious but interesting.

*Note - in spite of possible utility as a probe of structure formation, no SS will be used, because there are none in any alpha/beta proteins except at active sites, and because of problems expressing the protein.

*Secondary structure prediction for the five native sequences.

Collection of Relevant Information from the Literature or Personal Communications

*Crystallographic structure reports for the five proteins
*Description of beta sheet geometries - Salemne
*Analysis of beta/alpha and alpha/alpha packing - Chothia; Sternberg; Richards
*Analysis of beta/alpha and alpha/beta loops - Jones; Thornton; Efimov
*Residue-pair frequencies on adjacent strands; antipar. vs par. - Lifson and Sander
*Chi angle preferences - Janin; later tabulations
*Helix dipole - Hol
*Charge effects on helix formation in isolated peptides - Baldwin, Kim
*Residue preferences for N & C breaker, first turn, last turn, and middle of helices - J & D Richardson

Building a Symmetrical Barrel Model.

We started the model using the TAA barrel as a template, which was taken as that most nearly approaching 8-fold symmetry in structure. Result: 2TAA.cur

We then created 3 different strands of approx. 6 residues that were more-or-less ideal strands:

a) GAO beta strand 5 GAOTb5.brk
b) Twisted strand phi=-113,psi=129 from Salemme model barrel
c) Twisted strand phi=-114,psi=124 from center of native phi,psi
BRLBLDA.brk

(The last 2 were generated as poly-ala chains using a rigid geometry builder and applying Ramchandran's geometry as Salenme did (Ramachandran et al., Biochem. Biophys. Acta. 359, 298, 1974)).

By visual inspection and RMS superposition (using an INSIGHT inbuilt procedure, which takes into account CA,N,C,O) the first two fit badly onto TAA (the real strands have different, as well as large, variations in phi,psi; Ray's strand is much too twisted).

The average ideal strand fits surprisingly well, with a fit to each strand of TAA from .5 to 1.3 Å RMS. Yet it also reproduced the conical, somewhat lop-sided shape of TAA barrel. Result: BRLA8.brk

Since one side of the barrel looked very regular, it was fitted onto itself in an orientation rotated by 4 strands (superimposed using a horizontal ring of 8 residues at the end with regular H-bonding). Then the more regular 4 strands from each copy were kept:
BRL4.brk

One strand still leans out at the top, so this new structure was fitted onto itself rotated by 2 strands and the bad strands removed. Result: BRL2.brk

The H-bonding relation between strands was still a bit bad at one junction between copies, so every other strand was rotated slightly by hand. Result: BRL1.brk

The resulting strand still appeared conical, as though sliced high (above the narrowest neck) from the hyperboloid of revolution. Therefore we deleted one residue from each strand at the top end and extended each strand by one residue at the bottom (by overlapping with the ideal strand again). It now looked extremely symmetrical. The barrel diameter was approx. 14 Å in the middle, from the centres of lines fit to opposing strands. The crossing angle between opposite strands was 65°.

The next stage of the model building was to add on the inside and outside barrel side chains. A major question was whether a 4-fold sequence or conformational symmetry is possible.

There are six layers in our barrel interior, as defined by the C-beta of the residues, which we shall label from 1-6, with 1 and 6 being the top and bottom exposed surfaces and 3-4 being in the middle and totally buried. Note that the six layers are made of 4 residues each from alternate strands. We started with the central 2 layers, layers 3 and 4. The barrel diameter is too small for TRP with 4-fold symmetry (which requires 4 Trp's at the same level), so Phe was the first choice. We expected to have to break symmetry, but surprisingly, the symmetrical conformations were the best. There was one possible symmetric chi1 choice if the Phe's were in layer 3 and 2 choices if they were in layer 4. However, for one of the layer 4 chi1's the Phe rings were too close to the surface. So we tried for each remaining chi1-layer combination to find a good residue and conformation for the neighboring layer from Ile, Leu, Met. We had originally wanted Ile because its wedge shape would seem the best for inside a cylinder. However, it was not good for either alternative. Met also didn't work.

The two possibilities surviving in the central two layers were F L and L F. Both are 4-fold symmetric, and the Phe rings have an excellent herringbone packing. For each of these central two possibilities, a complete inside was fit with 6 layers, consisting of the central two big hydrophobic layers, the next 2 layers Gly or Ala, with the outer two layers being big hydrophilic residues. This was done by using the REPLACE residue option of INSIGHT to mutate the residues followed by manual setting of the chi1, chi2 angles etc. to the required values (again using the torsion rotation option of INSIGHT).

The outside residues of the strands of both of these barrels were also fitted with side chains, based on the observed distribution of residues in the 3 known barrels.(**) Thus, the 4 middle outside residues of a pair of strands were set to 2 Val, 1 Ile, and 1 Leu.

Again, this was done by mutation with INSIGHT or MIDAS followed by hand setting of the torsion angles to preferred values (which neither program does correctly initially). Note also that Midas showed arbitrary chi values (which was later fixed) and INSIGHT had the sign convention backwards.

The first barrel was packed on Iris MIDAS, taking advantage of its ability to move equivalent chis on 4 related residues simultaneously. (However, 3-D perception is relatively poor on this display.) Result: `GLYQFA.brk`

The second barrel was packed on PS300 INSIGHT, starting from a slightly better initial Phe version. In both cases, the outer layers of Q or Y needed to break conformational symmetry in pairs, either because of bad contacts or H-bonding. Result: `BGFQYLG.brk`

Note - as well as some possible bad side chain contacts (such as the Tyr ring to OH), very serious clashes of the CO on one strand with the inward pointing CB of the next strand were observed all around barrel, particularly for the Leu side chain. This is presumably the worst problem with a symmetrical, evenly twisted barrel. It probably explains a great deal of the extreme scatter in phi,psi and certainly explains the need for Gly at internal positions.

The 2 barrels (denoted GLYQFA and GFQYLG according to sequence) were both run through an energy program (which used a flexible geometry force field and an all-hydrogen potential, with hydrogens added arbitrarily by a hydrogen builder - the VFF program) to check for bad non-bonded contacts. This found all the above-noted clashes, plus a few more. GFQYLG was the clear winner, and was chosen for continuing. Files ending in H have hydrogens added. Result: `BGFQYLGH.brk`

In order to attempt to improve the LEU side chain clash, the chi angles of the 4 internal leu residues were modified slightly by hand. Result: `BRLLEUA.brk`

Choosing the beta-alpha connections

The structure now consisted of an isolated initial beta strand, 3 helix-connection-strand segments, and an isolated C-terminal helix. These need to be joined with beta-alpha connections. The first survey was done on FRODO using DGLP. Loop searches were done on each of the two non-equivalent positions, using gaps of 3 to 7 residues, match points of two residues on the strand and 3 on the helix, and moving the position of the match up and down along the helix. Best fit loops were in the range of 1.1 - 1.5 Å rms on C(alpha)'s, but all of them had serious drawbacks in the tilt or rotation relative to the strand, or in the pleat or peptide plane direction of the strand. 4 or 5 of the most reasonable examples were brought into INSIGHT from their Brookhaven files, along with half a dozen short, low-crossing-angle examples found by LGEOM. All were hand-fit to the possible locations, and superimposed using the CA-N-C-O-CB rms fit when plausible alignments were identified. Three possible connections were kept - ADH269 for the b2- a2 connection, and a choice for the b1-a1 of either LDH27 or a two-part connection stitched together from FXN10 and TIM15. After some side chains had been added, the choice was made to keep the LDH27 version. Result: `PROTDT.brk`

At this stage the chain was essentially complete, with the repeat fixed at 44 residues. Since we have been told that "The Answer is Forty Two", there may be a flaw in our design, but we couldn't see a way of shortening it by 2 residues.

Choice and Fitting of Remaining Side Chains

Side chains were already in place for the barrel strands, and some of the loop positions had retained critical side chains (usually Gly). The rest were chosen, starting with the connections and working then from both ends toward the middle of the helices.

The top and bottom layers of the barrel interior were hydrophilic but uncharged (Gln or Tyr), and it was felt there should be a charge pair as soon as pos-sible on the connections, both top and bottom. This was easy at the beta-alpha end, but not at the alpha-beta end, because one of the connections was completely occupied with required side chains (Tyr, Gly, hydrophobic) until the helix end. At the alpha-beta end, therefore, there are 3 Lys and the N terminus, but the paired charges are on the last turns of the two helices. Each repeating 44-residue unit has 10 charges (which is within the normal range of charges/residue for small-to-medium

proteins), approx. Three at each end of each helix, with the extra charge of the correct sign to interact with the helix dipole. In one case there is a Glu-Arg pair two turns apart in the same relationship as the pair implicated in S-peptide helix formation.

The remaining side chains on the helices were chosen primarily by two types of criteria: 1) from compilations of residue preferences for breakers, end turns, and middles of helices (**) and 2) suitability for their local environment - that is, good hydrophobic packing for the alpha-beta contact, helix-helix packing at those interfaces, and hydrophilics outside.

The chain termini needed additional consideration. The + on the N-terminus takes the place of the Lys; a Ser and an Asn get the chain out free of the barrel and are hydrophilic, and a Met is added at the end for initiation which may or may not be cleaved off after synthesis. In synthesizing such a repeating gene, it is easy to add extra residues outside the exact repeat, but it is much more difficult to delete residues from the ends of the outer repeats. Therefore GAK was added on the C-terminus to complete the fourth repeat. Either all 3 could be "coil", or the GA might continue the helix and the K wave about at the end.

Average Side Chain Volume Inside alpha/beta Barrels

Side chain volumes (from Richards, JMB 82, 1-14) in Å^3

Gly	0	Gln	72.2
Ala	26.0	His	76.8
Ser	31.8	Leu	82.0
Cys	38.2	Ile	82.0
Asp	49.4	Met	82.4
Thr	51.4	Phe	93.4
Asn	54.0	Lys	93.9
Pro	54.7	Tyr	99.9
Val	63.8	Arg	106.6
Glu	67.7	Trp	121.4

Barrel interiors - average volume of central residues:

TIM	GAA	TAA	grand averagee	
49.7	54.6	69.2	57.8	(using central 20 residues)
51.8	59.6	67.7	59.7	(using central 16 residues)

Side Chain Composition - Inside vs Outside of Barrels

Inside	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
TAA	1		1	1	3	3		4		2	1				1		2		1	3
GAO	3		1		2	2		3	1	1	1		1	1		2	2			
TIM	3	1	1	1	1	6		3		2		1			1			2	1	
total	7	1	3	2	6	11		10	1	5	2	1	1	1	2	2	4	2	2	3

Outside	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
TAA	1	2				2		3		4		1	2	1		2	1	6		1
GAO					1			6		4	1				1		2	5		
TIM	3	1			4			3		3							1	12	1	2
total	4	3			5	2		9		11	1	1	2	1	1	2	4	23	1	3

Positional Occurrence Frequencies in alpha helix

(from about 100 helices in Brookhaven examined on graphics - total number at left, relative preference at right)

NH breaker			CO breaker		
9	Asn	3.25	12	Gly	1.9
9	Asp	2.6			
11	Gly	1.7			
8	Ser	2.2	7	Leu	1.4
7	Lys	1.4	6	Asp	1.6
			6	Lys	1.14
4	His	3.0	4	Arg	2.0
5	Ala	.9	4	Asn	1.5
4	Glu	1.2	4	Gln	2.0
3	Pro	1.0	4	Glu	1.2
3	Tyr	1.0	3	His	2.0
			5	Ser	.8
2	Thr	.4	4	Thr	.9
2	Val	.4	4	Ala	.75
1	Trp	.5			
1	Cys	.5	2	Tyr	.75
1	Phe	.3	2	Ile	.6
1	Glu	.3	1	Met	1.0
			1	Phe	.3
			1	Pro	.3
0	Arg	0			
0	Ile	0	1	Val	.1
0	Leu	0	0	Trp	0
0	Met	0	0	Cys	0

NH first turn			CO last turn		
34	Asp	3.2	29	Lys	2.0
32	Glu	3.0	31	Ala	1.9
29	Ala	1.8			
			17	Leu	1.14
8	Gln	1.3	11	Phe	1.7
			9	Arg	1.3
10	Ile	1.0	9	Gln	1.3
12	Val	.75	7	His	1.5
10	Thr	.7	15	Val	.9
10	Lys	.7	11	Ile	1.0
9	Pro	1.0	12	Thr	.9
7	Phe	1.0	13	Ser	.7
7	Arg	1.0	9	Asn	1.0
6	Tyr	.75	8	Tyr	1.0
9	Leu	.6	9	Asp	.8
5	His	1.0	8	Glu	.8
			5	Cys	1.0
9	Gly	.4	2	Met	1.0
7	Ser	.3			
3	Trp	.5	0	Trp	0
3	Asn	.3	1	Pro	.1
2	Cys	.5	3	Gly	.1
1	Met	0			

Middle of helix

99	Ala	1.9
81	Leu	1.7
17	Met	3.0
30	Arg	1.7
19	His	1.5
40	Ile	1.2
28	Gln	1.3
27	Phe	1.2
54	Lys	1.1
52	Val	1.0
35	Asp	1.0
30	Glu	1.0
26	Asn	1.0
13	Trp	1.0
12	Cys	1.0
23	Tyr	.75
26	Thr	.57
27	Gly	.44
18	Ser	.33
9	Pro	.25

Sequence Homology Search by FASTP

(done on the 44-residue repeat plus 5 at each end)

We found a usual distribution of noise matches with scores up to 44, plus one with a score of 57 and one with 56 which are somewhat above the top of the noise. These are histidine permease membrane gene Q protein, with 35% identity for a 40 aa overlap, and fruit fly heat shock protein 83, with 33% identity for a 27 aa overlap. The first of these is shown below.

```
LARYGAKGVFLQGMDAEATQAAANIAKEGLYVLIGTGQEALQNLARY          (babarellin)
   ::  :: :::       :         :  : : :    ::
 MLYGFSGVILQGAIVTLELALSSVVLAVLIGLVGAGAKLSQNRVTG          (permease)
```

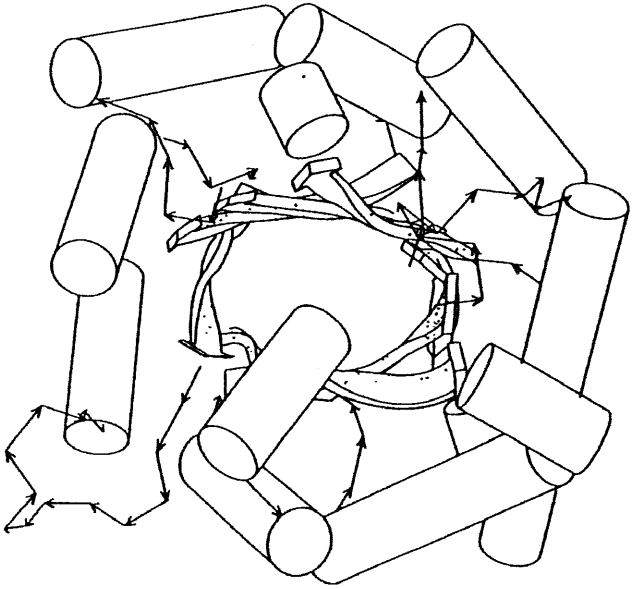
Total Amino Acid Composition for the Three Barrel Proteins: KGO, TAA, and TIM

A	96	10.11	1.15
C	19	2.00	0.88
D	68	7.16	1.24
E	44	4.63	0.96
F	29	3.05	0.83
G	89	9.37	1.00
H	15	1.58	0.72
I	63	6.63	1.33
K	49	5.16	0.78
L	72	7.58	1.05
M	18	1.89	1.25
N	38	4.00	0.89
P	42	4.42	1.06
Q	32	3.37	0.94
R	32	3.37	1.10
S	59	6.21	0.78
T	61	6.42	0.98
V	66	6.95	0.87
W	17	1.79	1.22
Y	41	4.32	1.21
total:	950	100.0	

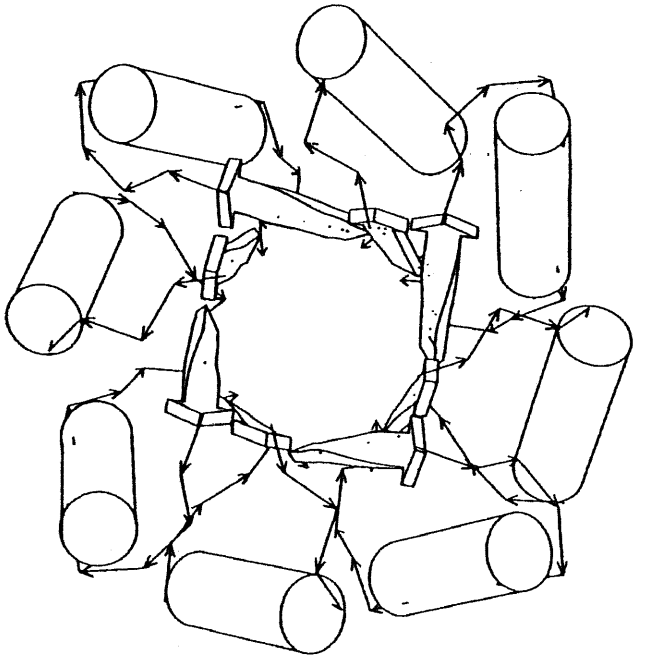
Amino Acid Composition of Babarellin

	number	%	preference
A	36	20.11	2.29
C	0	0.00	0.00
D	5	2.79	0.48
E	12	6.70	1.40
F	4	2.23	0.61
G	24	13.41	1.44
H	0	0.00	0.00
I	8	4.47	0.90
K	12	6.70	1.02
L	20	11.17	1.55
M	5	2.79	1.84
N	8	4.47	0.99
P	0	0.00	0.00
Q	16	8.94	2.50
R	4	2.23	0.73
S	1	0.56	0.07
T	8	4.47	0.68
V	8	4.47	0.56
W	0	0.00	0.00
Y	8	4.47	1.26
total:	179		

```
  2)  Babarellin.....BABA
      SHEET....      AAAAA                          AA A                    AA
      BRIDGE2..        bbb                           c c                     d
      BRIDGE1..      aaa                            bb b                    cc
      CHIRALITY    ----+-+-+ ++++++++++ +++-+----- +-+-++++++ +++-+-+--+
      BEND.....              SSSSSSSSSS SSSSS       S  SSSSSS SSSSS
      5-TURN...                         >5555<                  >5555<
      4-TURN...               >>>>XXXXX <<<<           >>>>XXX X<<<<
      3-TURN...              >33<        >33<
      SUMMARY..     EEEEE    STHHHHHHHHH HHHTT    EE ES  HHHHHH HHHHT    EE
      EXPOSURE.   *730012*41 88*1380015 03**723001 063***00*5 13**31*001
    1 SEQUENCE.   MDSGVFLQGM DAEATQAAAN IAKEGLYVLI GTGKQEALQN LARYGAKGVF

      SHEET....   AA                      AAA                     AAAA
      BRIDGE2..   dd                       ee                      fff
      BRIDGE1..                            ddd                     ee
      CHIRALITY   -+-++++++ +++++++-+- ----+-+-++ +++++++-+- +--+-+-+++
      BEND.....       SSSSSS SSSSSSSSS       S  SS SSSSSSSSS          SS
      5-TURN...                    >5555<                 >5555<
      4-TURN...        >>>>X XXXX<<<<          >>> >XXXX<<<<           >
      3-TURN...       >33<        >33<                                >3
      SUMMARY..   EE   STHHHH HHHHHHHTT    EEES  HH HHHHHHHHT    EEEE  ST
      EXPOSURE.   2*41*7*128 001503**73 8001063*** 109513**31 *0012*4188
     51 SEQUENCE.  LQGMDAEATQ AAANIAKEGL YVLIGTGKQE ALQNLARYGA KGVFLQGMDA

      SHEET....                     AAA                     AAAA
      BRIDGE2..                      gg                      hhh
      BRIDGE1..                     fff                     gg
      CHIRALITY   +++++++++ +-+-----+- +-++++++++ +-+-+--+-+ -+++++++++
      BEND.....   SSSSSSSSS SSS      S   SSSSSSSS SSS          SSSSSSSS
      5-TURN...          >5 555<                 >5 555<
      4-TURN...   >>>XXXXX<< <<           >>>>XXXX< <<<           >>>>XXX
      3-TURN...   3<       > 33<                                  >33<
      SUMMARY..   HHHHHHHHHH HTT   EEES   HHHHHHHH HHT   EEEE   STHHHHHH
      EXPOSURE.   *128001503 **73300106 39**01*513 **31*0012* 4188*13802
    101 SEQUENCE.  EATQAAANIA KEGLYVLIGT GKQEALQNLA RYGAKGVFLQ GMDAEATQAA

      SHEET....             AAA
      BRIDGE2..             hhh
      BRIDGE1..             aaa
      CHIRALITY   ++++-+---- --+-+-++++ ++++++
      BEND.....   SSSSSS       S  SSSS SSSSSS
      5-TURN...     >5555<
      4-TURN...   XX<<<<          >>>>X XXXX<<<<
      3-TURN...      >33<
      SUMMARY..   HHHHHTT    EEES  HHHH HHHHHHH
      EXPOSURE.   4501**5250 01063***01 7210**2*
    151 SEQUENCE.  ANIAKEGLYV LIGTGKQEAL QNLARYGA

Notation:  SUMMARY H=ALPHA-HELIX
                   E=BETA-STRAND
                   B=BETA-BRIDGE G=3-HELIX I=5-HELIX
                   T=3-,4-, OR 5-TURN
                   S=BEND
```

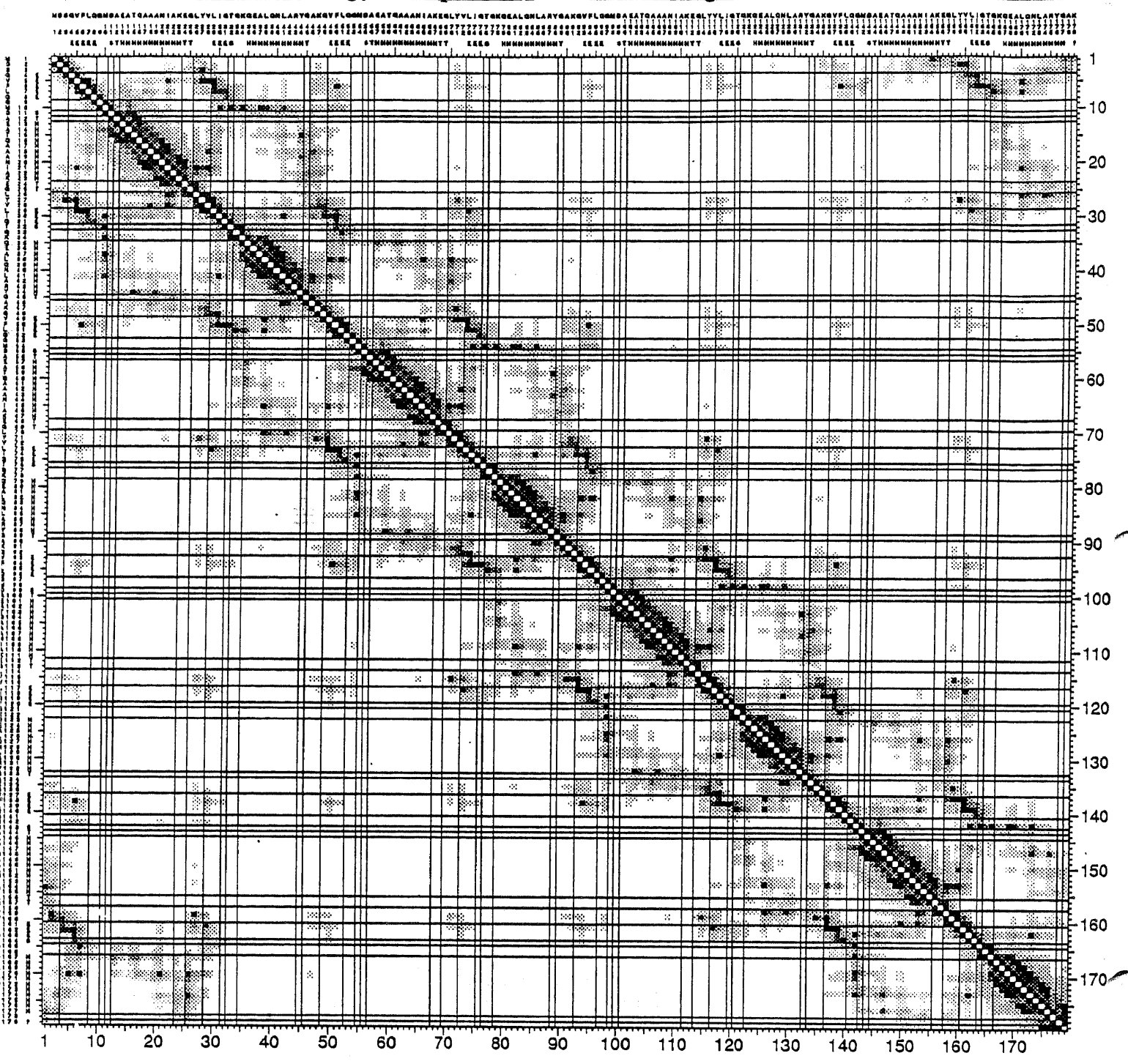


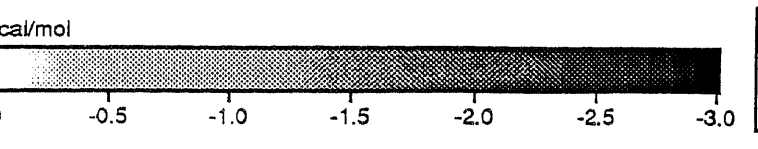
Chicken TIM



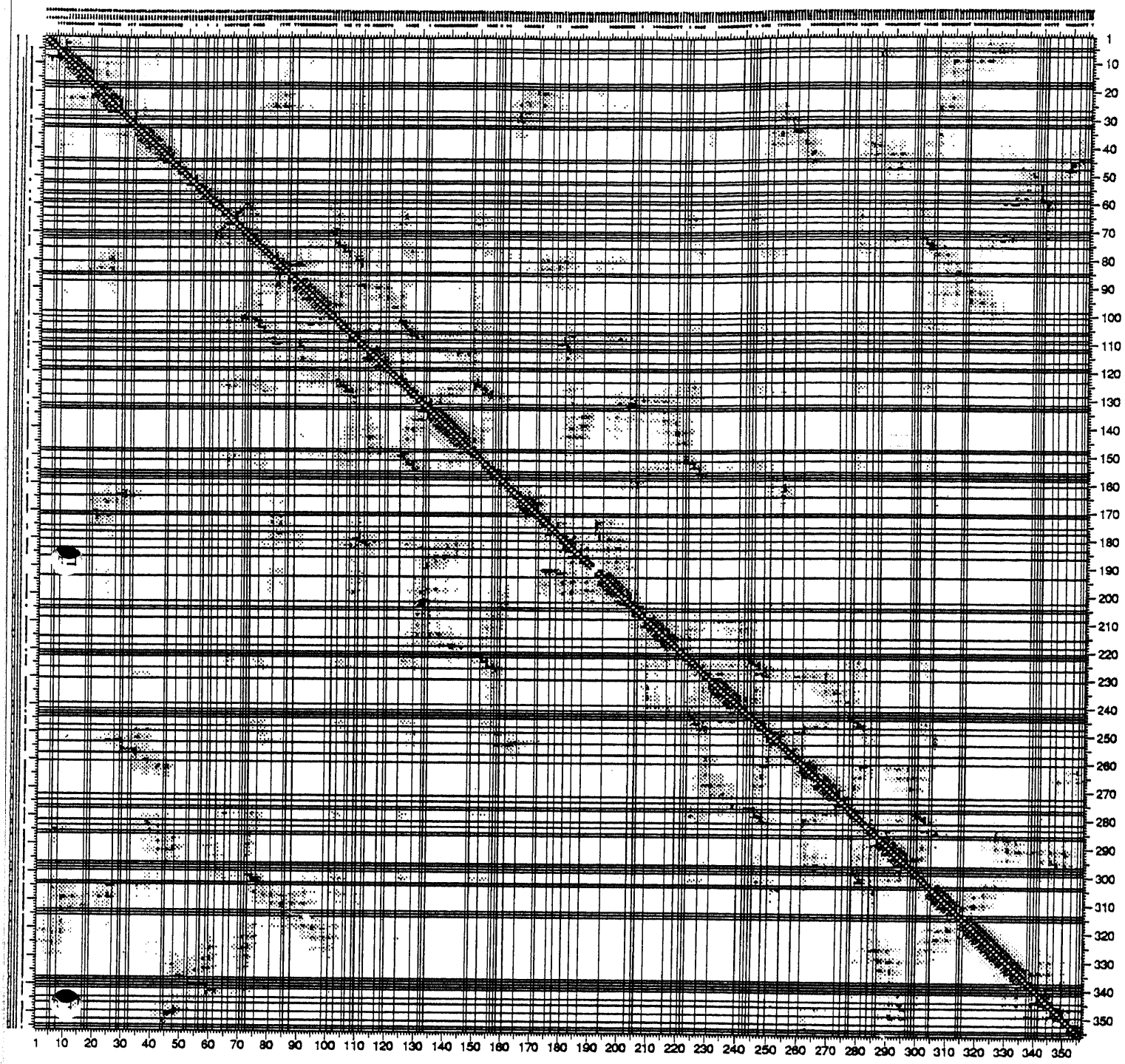
BARABELLIN

BABA: Van der Waals energy.





GAO0: Van der Waals energy. (a TIM barrel protein)



kcal/mol

.0 -0.5 -1.0 -1.5 -2.0 -2.5 -3.0

Arthur Lesk
Osnat Herzberg
Michael Schaefer
Zelda Wasserman

* MUTATED AND IDEALIZED FLAVODOXIN *
(FXNM, FXNI)

** BUILDING-STEPS FOR AN IDEALISED A/B-PROTEIN WITH BOTH THE TOPOLOGICAL FRAMEWORK AND THE BINDING SITE MODELED AFTER FLAVODOXIN **

Mutating Flavodoxin to bind 3,6-dioxyxanthone

1) The conformation and charge of 3,6-dioxyxanthone were determined using the molecular modeling package SYBYL. Charges were calculated by the method of Gasteiger and Marsili, and the geometry determined using the simplex method.

2) Changing the binding specificity of Flavodoxin to bind 3,6-dioxy-xanthone

```
Mut0 ---> Mut1 :  Residue   9 THR ---> ASN
                  Residue  54 SER ---> THR
                  Residue  56 MET ---> ARG
                  Residue  87 SER ---> THR
                  Residue 119 ASN ---> GLN

Mut1 ---> Mut2 :  Rotate ARG 56

Mut2 ---> Mut3 :  Superposition of SUB

Mut3 ---> Mut4 :  Residue   9 ASN ---> GLU   (second)
                  Residue 119 GLN ---> ASN   (back!)

Mut4 ---> Mut5 :  Residue  54 THR ---> GLU   (second)
                  Residue  87 THR ---> ASN   (second)

Mut5 ---> Mut6 :  Residue  11 ASN ---> MET

Mut6 ---> Mut8 :  Residue  11 MET ---> ASP   (second)
                  Delete 8 - 10
                  Fit imported loop of onto residues 6 - 12
                  Adjust 5TYR 6TRP
                  Renumber residues

Mut8 ---> Mut9 :  Residue 51 (was 54) GLU ---> THR ---> SER  (back!)
                  Adjust sidechain SER 51
```

```
Net number of mutations: 3 residues;
                         deletions: 3 residues.

                      11
fxn    MKIVYWSGTGNTEKMAELIAKGIIESGKDVNTINVSDVNIDELLNEDILI
mut      IVYWS   DTEKMAELIAKGIIESGKDVNTINVSDVNIDELLNEDILI
       **     ****


            56                                      87
fxn    LGCSAMGDEVLEESEFEPFIEEISTKISGKKVALFGSYGWGDGKWMRDFE
mut    LGCSARGDEVLEESEFEPFIEEISTKISGKKVALFGNYGWGDGKWMRDFE
            *                                       *

fxn    ERMNGYGCVVVETPLIVQNEPDEAEQDCIEFGKKIANI
mut    ERMNGYGCVVVETPLIVQNEPDEAEQDCIEFGKKIANI
```

3) To relax the model-built conformation, an energy minimization of the new binding site was carried out with the program AMBER. The ligand, all residues which had undergone mutation and the residues of the shortened loop, as well as all residues in contact with any of the above, were allowed to move; the rest of the molecule was held fixed. Eight hundred conjugate gradient minimization steps were performed, during which the energy fell from the initial strained 635 to a relaxed -651 kcal/mol. Atoms which moved the most were in Gly 54, Arg 53 and Asn 8. The carbonyl oxygen of Gly 54, which had been positioned too close to the ligand, tilted upwards and the attached carbon moved away .77 Å. Arg 53 and Asn 8 form a salt bridge. The rms deviation from initial position of the backbone atoms which were allowed to move was .33 Å.

The FXN modified coordinates: FXNM.BRK (alias FXN_MUTGEM.PDB)

Building Idealized Flavodoxin

4) Building of an ideal parallel beta-sheet:
Construction of a single beta-strand using following the set of phi/psi angles:

```
phi1 = -110.5 ; psi1 = 125.5
phi2 = -115.0 ; psi2 = 126.0
```

*Literature: F.R. Salemme, D.W. Weatherford, JMB**146** (1981), 101-117.

Duplication and translation perpendicular to previous strand by 5 Å, repeated four times.

Twist angle between each pair of strands (to make propeller sheet) -19°, giving a right handed twist of the sheet.

*Literature: J. Janin, C. Chothia, JMB **143** (1980), 95-128.

```
Strand No : 1, 2, 3, 4, 5
No of res.: 5, 6, 7, 8, 8
Name      : B, A, C, D, E
```

5) The relative postions of helices a and e to strand A and E, respectively, were maintained as in FXN. This altered the orientation of these helices such that their geometry is as seen in sperm-myoglobin (helices F,H). Therefore the sperm-myoglobin helices were fitted to the ideal sheet. The rms deviation of all main-chain atoms of these helices from ideal helix with (phi,psi;-62,-45) was 0.26 Å and 0.54 Å, respectively. The inter-helix angle is -68.4°, and the minimum distance between their axes is 9.3 Å.

For comparison, note the values for helices a and e in FXN: The rms-fit to ideal helix is for helix a 0.229 Å and for helix e 0.275 Å. The inter-helix angle is -72°, and the distance between the axes is 10.3 Å. Helix a was extended by 10 residues and helix e by 7 residues using ideal helix spare parts to cover the sheet.

Both c- and d-helix are taken from FXN itself to assure the integrity of the binding-site loops. The rms deviations from ideal helix are 0.298 Å and 0.288 Å, respectively. The inter-helix angle is -18° and the minimal axes distance is 0.582 Å. Note that these helices abut at their ends and do not cross each other as in the paradigm cases of helix-helix interactions.

```
Helix No  :  1,  2,  3,  4
No of res.: 16,  8, 12, 16
Name      :  a,  c,  d,  e
```

6) Building of the loops using A. Jones' procedure (DGLP in FRODO). About three or four markers were used on each secondary structure motif to fish out only loops characteristic to alpha/beta connectivity.

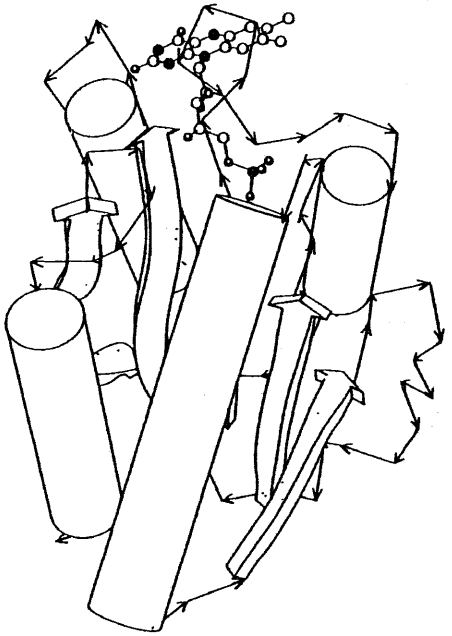
*Loop A/a picked from FXN3 (!) YGWGDGKWMR rms = 0.792 Å
*Loop a/B picked from RHD1 FRVFGHRTV rms = 0.719 Å
Helix a was shortened by 4 residues
*Loop B/C picked from FXN3 (!) TINVSDVNIDELLNEDILI rms = 0.666
13 residues were inserted between the strands B and C.
*Loop C/c was taken as is from FXN (active site loop). The gap between the strand and the loop that was created due to the deviation of the ideal sheet from the real one was closed by aligning 4 residues of TLN3 VVGI rms = 0.68.
*Loop c/D picked from FXN3 (!) IEEISTKISGKKV rms = 1.34. This loop is identical to the native one.
*Loop D/d was taken as is from FXN (active site loop). The gap between the strand and the loop was closed by using FRODO's REFI.
*Loop d/E was picked from CPA5 YNQGIKYSF rms 0.23.
*Loop E/e was picked from REI1 QASQDIIK rms = 0.869.

7) Analysis of the sequence: We used a genetic sequence analysis program developed by the University of Wisconsin Genetics Computer Group to examine the designed sequence. The UWGCG-package includes information on hydropathy, acidity, alpha- and beta-moment and the secondary structure prediction scheme of Chou & Fasman.

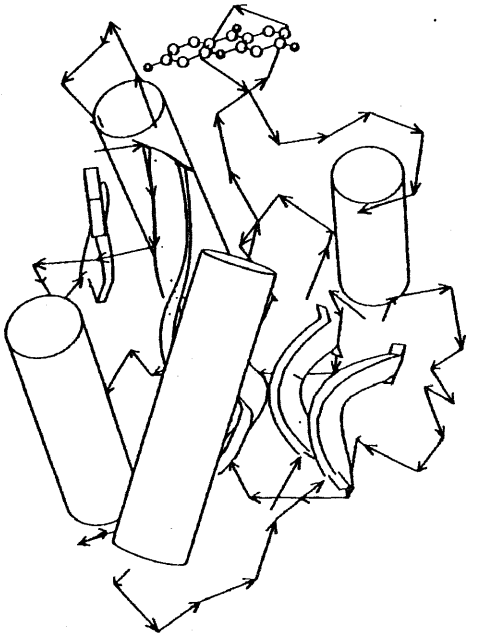
For comparison with properties of the FXN-sequence, we ran the PEPPLOT-program on the sequence of both FXN and FXN_ideal. The score of well predicted secondary structure elements was essentially the same in both cases, with the e-helix (taken from myoglobin) not correctly predicted. The shape of the curves of alpha- and beta probability and hydropathy of FXN_ideal resembles the ones of FXN. Noting that the sequence of FXN_ideal is to a high proportion taken from FXN, this is not a surprising result.

```
4)  Mutated Flavodoxin.....FXNM
    SHEET....   AAAA                          AAA A               AAAAA
    BRIDGE2..   bbbb                                              ccccc
    BRIDGE1..   aaaa                          aaa a                bbbb
    CHIRALITY   --+--++++ ++++++++++ ++-+-+--+- -+++--++++ ++-+--+-+-
    BEND.....       SSSS SSSSSSSSSS SSSS         SSS  SSSS SS S
    5-TURN...                      > 5555<
    4-TURN...         >> >>XXXXXXX< <<<          >444< >444 <
    3-TURN...      >>3X< 3<       > >3<<         >33<      > 33<
    SUMMARY..   EEEE GGGH HHHHHHHHHH HHTT   EEE ETTT  STTT TT SEEEEE
    EXPOSURE.  4500136409 9009101*00 9*57**1952 517*484**0 3*5*300000
  1 SEQUENCE.  MKIVYWSDTE KMAELIAKGI IESGKDVNTI NVSDVNIDEL LNEDILILGC

    SHEET....    B   B                        AAA AAAAA
    BRIDGE2..                                 dd   eeee
    BRIDGE1..    F   F                        ccc cc
    CHIRALITY  -++--+--+ ++++++++++ +++++---- ++-+-++--+ ++++++++++
    BEND.....    SSSS  S SSSSSSSSSS SSS SS          SSS S SSSSSSSSS
    5-TURN...       >5 555<                                      >
    4-TURN...   >444<     >>>>XX<<< <                    > >>>XXXXXX<
    3-TURN...                       >33X33<
    SUMMARY..   BTTTB  T TTHHHHHHHHH STT TT EEE EEEEESSS S HHHHHHHHHH
    EXPOSURE.  13*7*3512* 9805610**0 2**0863900 000252*2*0 901*607*50
 51 SEQUENCE.  SARGDEVLEE SEFEPFIEEI STKISGKKVA LFGNYGWGDG KWMRDFEERM

    SHEET....      AA     AAAA
    BRIDGE2..
    BRIDGE1..      dd     eeee
    CHIRALITY  ++-+---+-- -+---+-+++ ++++++++++ ++
    BEND.....  SSSS    S      SSS SS S SSSSSSSS SS
    5-TURN...  5555<
    4-TURN...  <<<                   >>>>XXXXXX <<<<
    3-TURN...  >33<               >>3 <<         >33<
    SUMMARY..  HHTT EE S   EEEESS GG GHHHHHHHHH HHH
    EXPOSURE.  976317364* 2361***19* 28**018107 *11*
101 SEQUENCE.  NGYGCVVVET PLIVQNEPDE AEQDCIEFGK KIAN
```

Notation: SUMMARY H=ALPHA-HELIX
E=BETA-STRAND
B=BETA-BRIDGE G=3-HELIX I=5-HELIX
T=3-,4-, OR 5-TURN
S=BEND

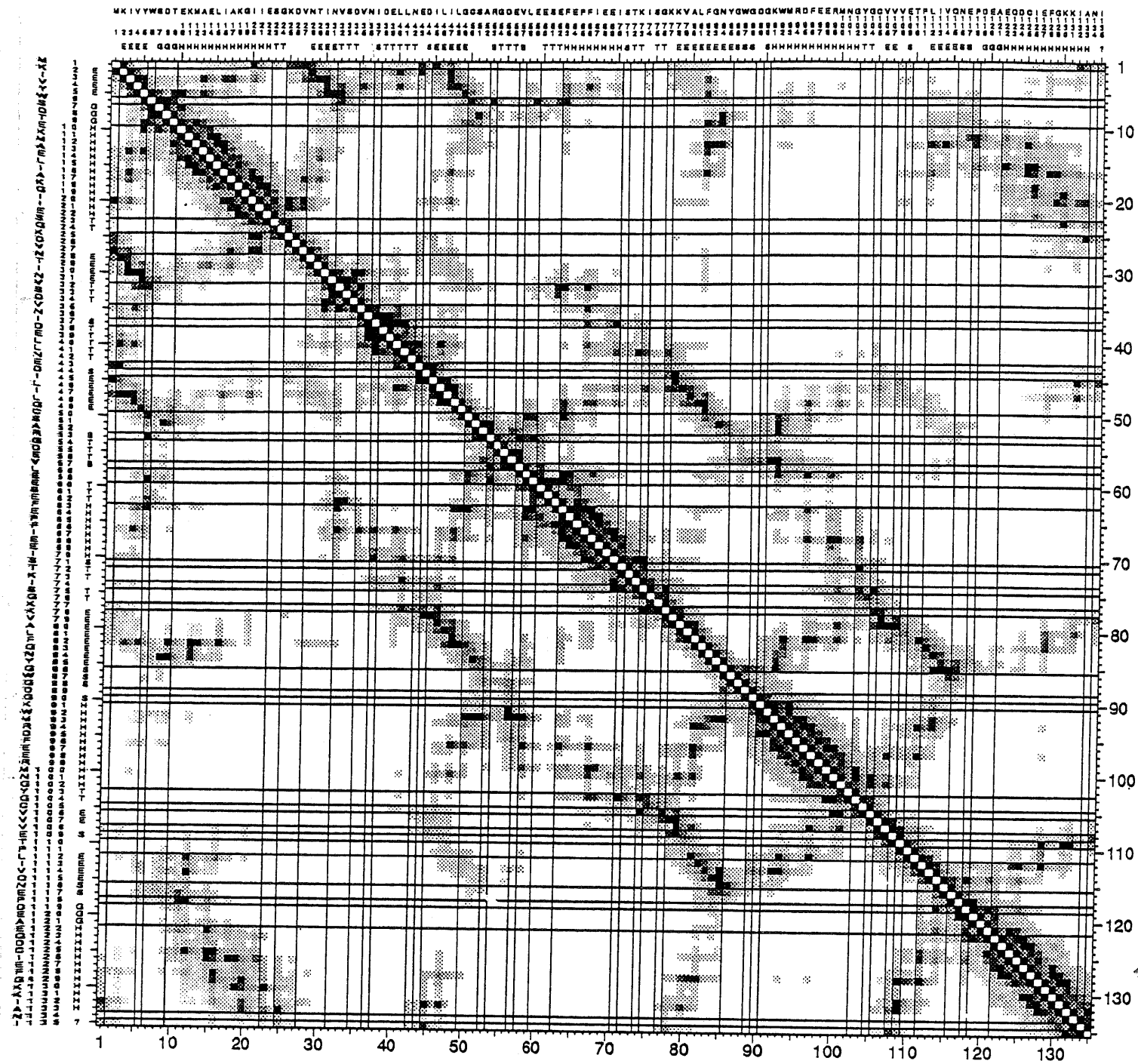


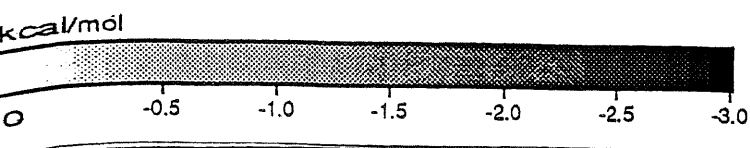
Native Flavodoxin



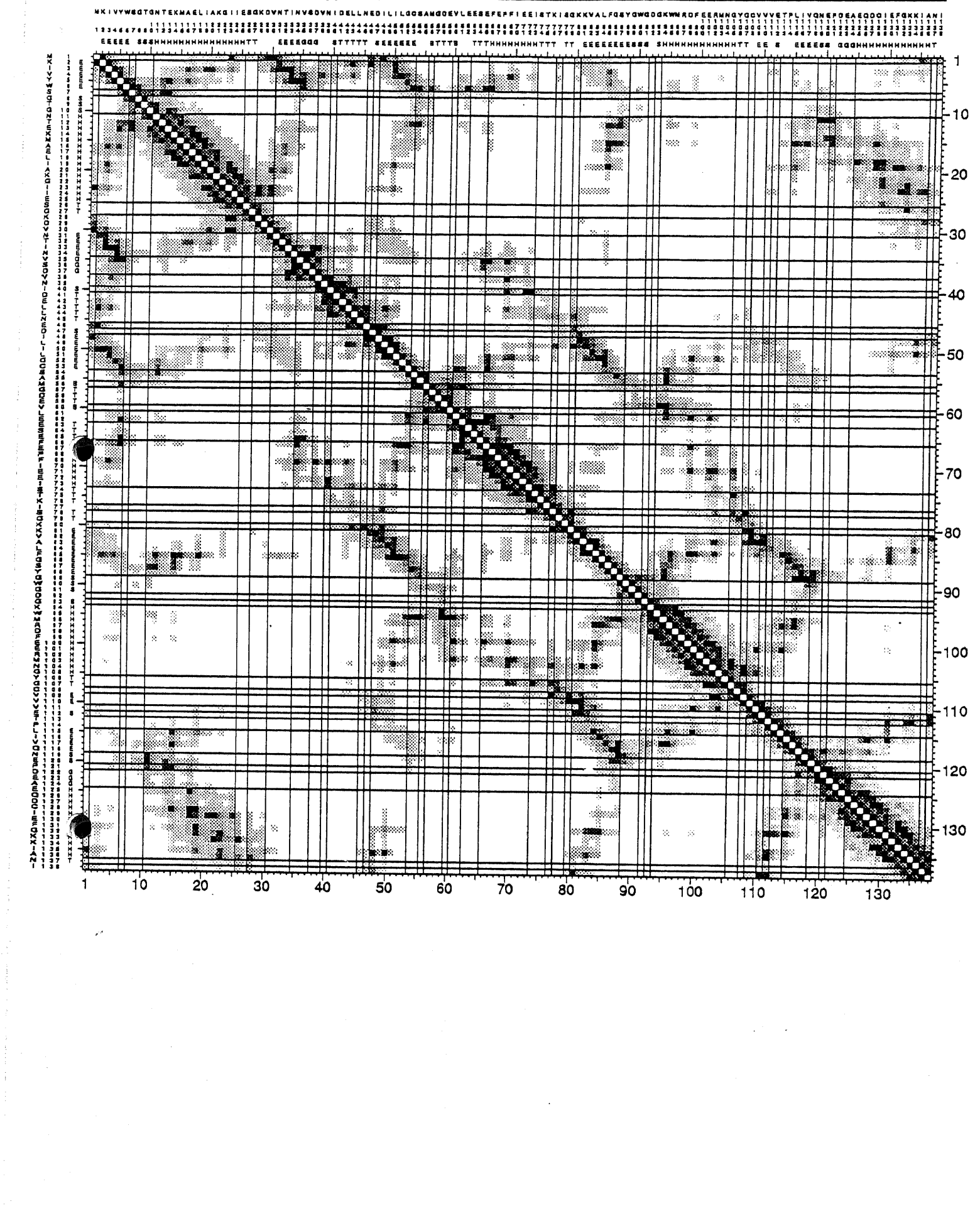
Model Flavodoxin

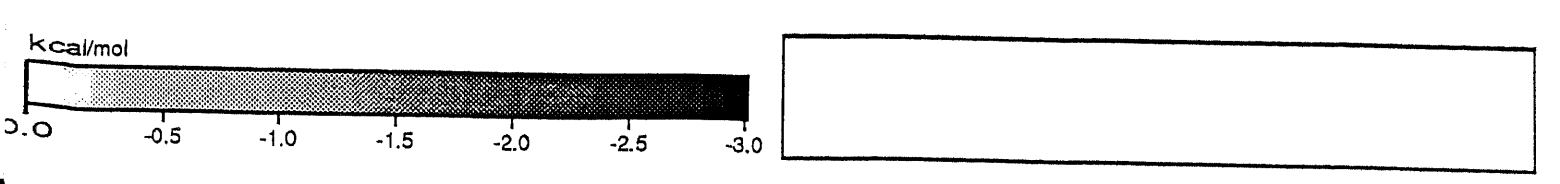
FxNM


1
10
20
30
40
50
60
70
80
90
100
110
120
130

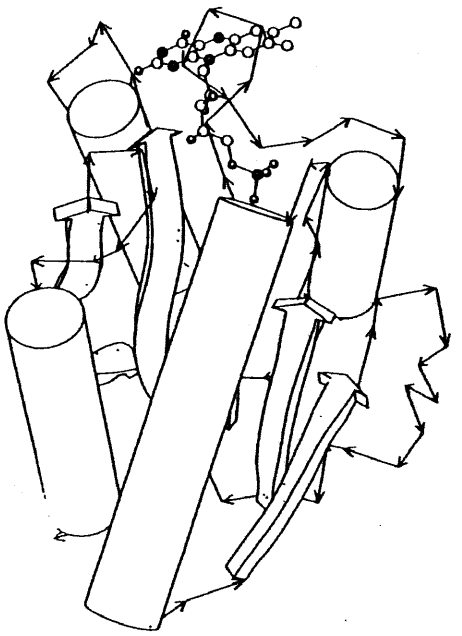

kcal/mol
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0

3FXN: Van der Waals energy.

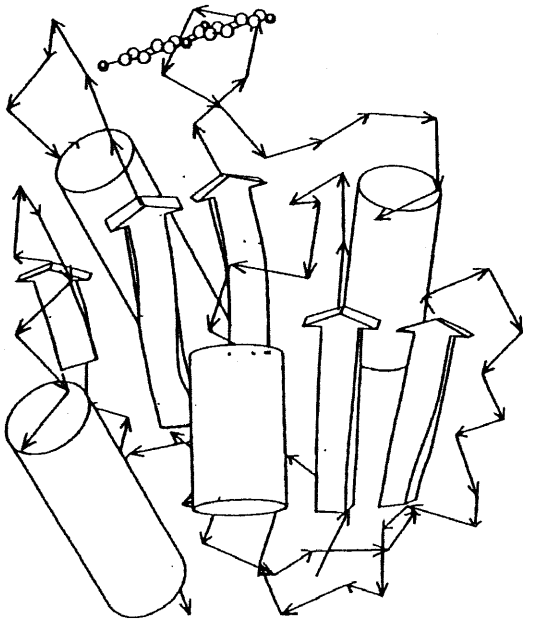




```
3)  Idealized Flavodoxin.....FXNI
    SHEET....   AAAA                    AAAA                  AAAAAA B
    BRIDGE2..   bbbb                                          ccccc
    BRIDGE1..   aaaa                    aaaa                  bbbb   E
    CHIRALITY   -----+--+ +++++++++-+ -++---+++- -++++++-++ ---+--++--
    BEND.....        SS S SSSSSSSSS        SSS   SSSSSS S          SS
    5-TURN...                  >5555 <
    4-TURN...             > 4>>X><<<<              >444<             >44
    3-TURN...              >33< >33<       >>3<<   >3><3<
    SUMMARY..   EEEE SS S TTHHHHHHTT    EEEEGGG   STTTTT S  EEEEEE BTT
    EXPOSURE.  72000061*7 401*007*59 2*791707*3 91**04*7*1 011102375*
  1 SEQUENCE.  ALVVVASGTE AADLAMSHQA TRTKVNVSDV NIDELLNEDG LIMIISAMGD

    SHEET....   B                     AAAAA
    BRIDGE2..                          ddd
    BRIDGE1..   E                      ccccc
    CHIRALITY  +---+-++++ ++++++-+++ +-----+-++ --+++++++ +++++-++++
    BEND.....  S   SSSSSS SSSSS SS   S        SS S S SSSSSS SSSSSSS  S
    5-TURN...     >5555<                                    >5555<
    4-TURN...  4<    >>>> XX<<<<                  >>>>XXXX XX<<<<
    3-TURN...                 >3><3<                 >33<      >>3<<>33
    SUMMARY..  TB  TTTHHH HHHHHTTT    S EEEEE SS S SHHHHHHHH HHHHHTT TT
    EXPOSURE.  2612**7087 10**038*17 510000438* 3*0900*606 *9187661**
 51 SEQUENCE.  EVLEESEFEP FSEEINTKIS ASVVIAYSGW GDGKWMRDFE ERMNGTGIKY

    SHEET....    AAA
    BRIDGE2..
    BRIDGE1..    ddd
    CHIRALITY  ---++-+-+- +++++++++
    BEND.....       SS   SSSSSSSSS
    5-TURN...
    4-TURN...            >>>>XXX<<< <
    3-TURN...  <                 >33 <
    SUMMARY..    EEE SS   SHHHHHHHHH
    EXPOSURE.  2443537*9* 1*706812** 4
101 SEQUENCE.  SNSMNNGRKD GKDMYEIGKK I

Notation:  SUMMARY H=ALPHA-HELIX
                   E=BETA-STRAND
                   B=BETA-BRIDGE G=3-HELIX I=5-HELIX
                   T=3-,4-, OR 5-TURN
                   S=BEND
```

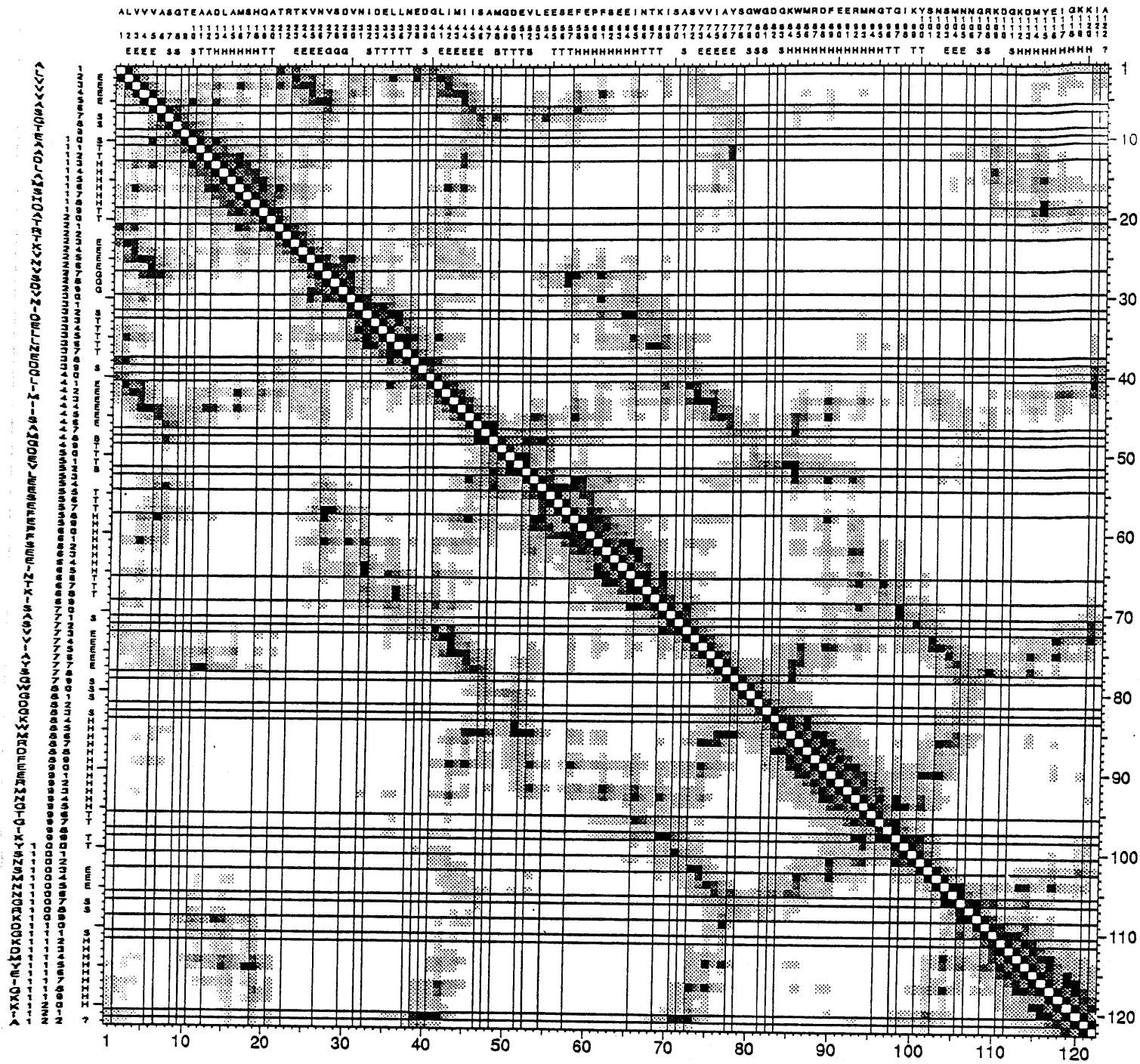


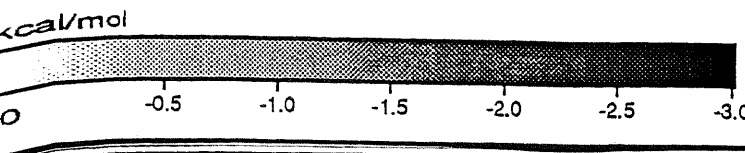
Native Flavodoxin



Ideal Flavodoxin

FXNI


1
10
20
30
40
50
60
70
80
90
100
110
120


kcal/mol
0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0

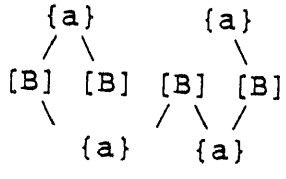
Dave Richardson
François Colonna
Eric van Cutsem
Kim Kusk Mortensen

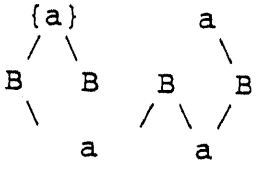
* BETALPHACIN *
(BEAL)

Choice of the Design Paradigm

4 beta strands, 4 alpha helices, minimal nucleotide binding domain. Preservation of a PO4 binding site homologous with that of FXN flavodoxin). Real nucleotide binding domains have 3 beta strands winding out on a side that binds a base. Flavodoxin winds 3 out on the flavin side, LDH winds out 3 in both directions and binds a dinucleotide across the middle. So it is unclear that a 4 strand domain, winding out 2 on each side, is sufficient to bind even a mononucleotide (though in terms of actual contacts it might be possible).

Paradigm:





Initial Construction of the Model

Beta Sheet

3 Strands taken from LDH:
[B]2: [47-53]beta, [B]1: [22-28]beta, [B]3: [92-97]beta. 4th strand ([B]4) generated by superimposing [22-28] onto [134-138]beta of LDH (this actually flipped [22-28] with respect to its fit as [B]1).

Fit: rms on C(alpha) : 0.946 Å
Now beta sheet with 4 strands:

A22-A28		as [B]4		
92	-	97	[B]3	
22	-	28		[B]1
47	-	53	[B]2	

28 and A28 deleted to even ends of sheet and remove poor beta conformation.

Now we had the first part of Betalphacin, BAP (with artifactual numbering):

[B]2	[B]1	[B]3	[B]4
47	22		
48	23	92	a22
49	24	93	a23
50	25	94	a24
51	26	95	a25
52	27	96	a26
53		97	a27

Brugel (Energy Minimization Program): steepest descent (500 steps) to regularize the structure, with no constraints.

Result: BETAST.pdb

Alpha Helices

MacLachlan fit of 92-95 (BAP) N,C,0,CA onto B82-B85 (FXN) N,C,0,CA. RMS : 0.59 Å (these are Brugel operations).

Application of the resultant matrix onto the 4 strands idealized from LDH. Resulted in a model based on the 4 strands idealized from LDH with 4 helices surrounding the beta sheet. The 4 helices were very much like the FXN helices, but the fit to the sheet appeared faulty on graphics: Poor fit of 47 - 53 & 22 - 27 (LDH # on model BAP). Not twisted enough, not cupped enough, not in good relation to {a}1.

Brugel again: MacLachlan fit of 47 - 53 (model) N,C,0,CA (N term strand) onto B1 - B5 (FXN) N,C,0,CA RMS = 0.71 Å and fit of 23 - 27 (model) N,C,0,CA ([B]2 outside strand) onto B49-B53 (FXN) N,C,O,CA RMS = 0.89 Å.

Result: Model2.pdb

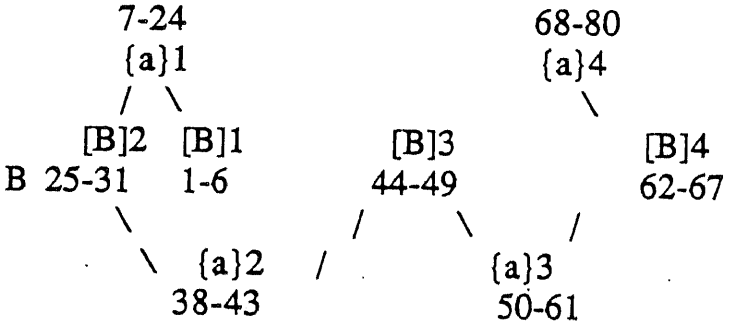
Now, of course, the model looks very much like the center of FXN. However, helix {a}2 was too short and in a bad shape, so it was replaced with a duplicate of helix {a}3, which was fitted in place using the Maclachlan fit. Helix {a}2 was now of a good length and quality, but was too far away from the beta sheet. The helix was therefore handfitted to a "good" position using FRODO FBRT. A good position means in this case something which resembles the position in Flavodoxin.

Idealization (Minimization) of the Model

Steepest descent (500 steps) by Brugel to obtain secondary structure framework on which to build the loops.

Result: Model3.pdb

Frame design, numbering scheme:



Loop Building

Extra residues were tacked onto the model on one side of each of the gaps to be joined so that FRODO/DGLP would have the expected number of residues that would give a good connection. Usually 3 extra aa (amino acids) were used, several loops were cut back to 2 aa. Generally, the connection search gave a lot of scatter, most acted as if the model beta-alpha distance was less than native beta-alpha, though some just connected beta-alpha at an angle rather than the parallel design. Targets for the loop search were 2 or 3 ca in beta and 5 or 6 in alpha. Thus is was not surprising that the alpha side usually looked better, but often the fit was surprisingly poor. Sometimes the beta looked good, but crossed the model beta at an angle. Occasionally an alpha-alpha connection was found. Loops were picked to match the design geometry, usually one in the top 2 rms choices. When the picked loops were joined to the model, there were bad distortions in both beta and alpha structure. 5 cycles of FRODO (Herman's) idealization in the join region made things look only a little bit better. After annealing, several loops were checked again by

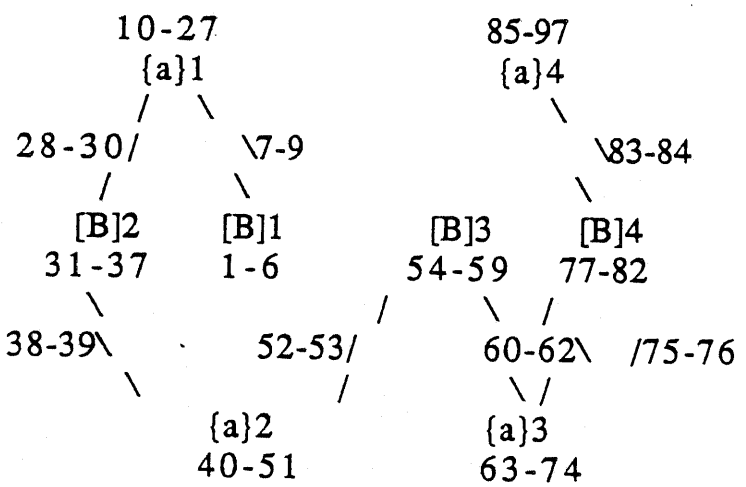
searches, and found what seemed like different populations but still with considerable scatter. However, the chosen loop always returned as the best RMS fit so the annealing distortions were not so severe as to prevent that. 10 to 14 length segments of poly-ala in the conformation of the found sequences were placed in the model, replacing the gap and its flanking chain. The fitting as well as the search is based on C(alpha) positions.

Loop	in model	from protein in PDB		replacement over gap	C(alpha) rms [Å]
1	7- 9	3FXN	5- 17	Y-WSGT-GDTEKMAE	0.9
2	28-30	1RHD	258-269	LAAYLCG-KPDV-A	0.7
3	38-39	2ACT	115-225	NVP-YN-NEWALE	1.35
4	52-53	5CPA	183-192	VKNGH-NF-KAF	1.39
5	60-62	1ECD	40- 50	EF-AGK-DLESIK	0.89*
6	75-77	5CPA	179-193	IVEFVKNH-GNF-KAFL	1.2
7	84-86	1BP2	84- 95	CS-SEN-NACEAFI	0.96

*Loop 5 taken from 1ECD was originally an alpha-alpha connection, and required hand reworking later.

Result: Model4.pdb

Now renumbered to include the loops:



Model Refinement of the Connected Whole: Starting with Model5

	RMS to set #1 for sc & mc	RMS to set #1
Set #1 model5		
a) Constraint forces on helices, loops and PO4 binding site; refinement of 500 steps on strands.		
Set #2 model5_bet	1.16 (beta only)	1.16 (beta only)
b) Constraint forces on betas, loops and PO4 binding site; refinement of 500 steps on helices.		
Set #3 model5_hel	0.97 (alpha only)	0.97 (alpha only)
c) Constraint forces on betas, helices. PO4 binding site; refinement of 250 steps on loops.		
Set #4 model5_loop	0.36 (loops only)	0.30 (loops only)
d) Constraint forces on PO4 binding site only; refinement of 250 steps on all the rest.		
Set #5 model5st	0.95 (all)	0.95 (all)

For side chain and main chain total RMS = 0.95 relative to #1

Comments: Brugel seems to respect alpha, doesn't seem to know about 3-10 ends on helices, rather capricious on beta: often weird phi-psi's, and omega was allowed to go as far as 170. This is within the possibilities in real proteins but for this case of constructing an idealized framework, we probably should have set a tighter constraint on omega in the initial refinements. Looking down the helix axes, the helix wheel view showed irregularities in the pattern of beta carbon positions. This probably is the result of variations in the dihedral angles which were, at least in part, introduced by the process of fitting in the loop using DGLP in FRODO. Only steepest descent minimization was used since we didn't want shifts in bulk secondary structure position, and so Brugel also left some bad contacts between loops. It should be noted that it was working on poly ALA, and did, in general, idealize the structure without irrepairable damage. Result: model6

Repair before Side Chain Addition

Refitting by "hand" on FRODO: just pretended one was a crystallographer faced with good helix density, moderate beta density and poor loop density. By breaking out fragments, successive frames of torsion angles, and reiterated attempts, rebuilt two of the beta strands and two loops to look a lot better. Loop (C)5 between [B]3 and {A}3 was finally trimmed in with a gly-pro coming out of the [B]. The phi-psi angles were adjusted to be near the center of the observed distribution in FXN and LDH of -120,130 in [B]3 and [B]4 which gives a nicely pleated, twisted sheet. Loop (C)7 was rebuilt as well. This is a very easy process and can often get the connections redone within about1 Å distance and 10 deg angle of desired geometry. The two helices 1 and 4 are very close together with gly on each side at the crossing. Helix 4 was rotated by 5° around a convenient bond on the connecting loop to [B]4 in order to relieve this close contact somewhat. (We think is was psi of GLU 83.) (Approximate measuring estimated the axis-axis distance to be about 7-8 Å.) Superposition of FXN ca on model C(alpha) showed those sets of helices to be of similar spacing, though there is a slight rotation of each - which could either be a good way to avoid direct homology with FXN, or a source of bad contacts with the beta sheet.

Side Chain Addition

Side chains were put on using several levels of attention. First, a set of them around the PO4 binding site and on the helix that points into that site were taken directly from FXN. (7-12, SGTGNT, where the gly's are probably important to allow a sharply bent loop which presents most main chain N's to the PO4.) The design is to preserve that binding and both that special sequence in immediate contact with PO4 was kept homologous and residues that seem important to tilt the helix with respect to that loop were retained. The general size and nature of the residues in the interior were retained in a more inexact way since this not only may be important for packing of these alpha-beta structures, but the relative closer spacing near the c-terminal end of the beta sheet may be important for functional orientation of the helices. In part, this was accomplished by the original spacing of the secondary structure, and in part by remembering the general distribution of side chain types. A few VALs, ILEs, etc. on the beta sheet were put in just on this criterion of internal position. At the immediate level, the side chains were put on only really paying attention to the local needs of packing, charge distribution, and

general preference of side chains with respect to secondary structure.

Amino Acid Distribution References:

*General preference with respect to beta:
Lifson, S. & Sander, C. (1979). Nature, 282, 109-111
*Beta strand pair preference:
Lifson, S. & Sander, C. (1980). J. Mol. Biol., 139, 627-639
*Alpha helix preference and distribution in regions:
Feldmann, R. (1976). Microfiche Atlas
(and a tabulation made by the class that designed FELIX)

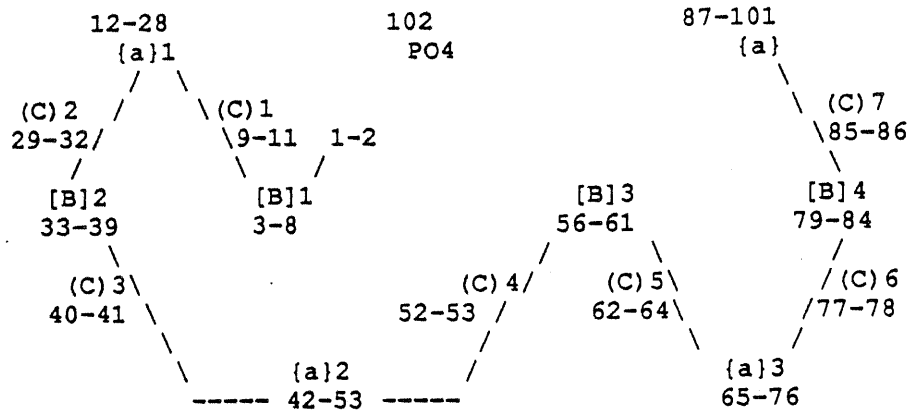
Then one hopes one can say :
"After all side chains were in place and adjusted to look reasonable in terms of angles and packing, the structure was minimized as a whole, and general problems of distribution and specific packing were solved."

But:
As of Sept 15, 1986:
So far, the only real fitting of side chains was done around the PO4 site using MIDAS with van der Waals surfaces on the IRIS.

And as of Sept 16, 1986:
Two residues were built onto the N-term and onto the C-term to make sure the ends really came out to the surface, these were roughly fitted by hand. This means, of course, that the whole molecule should be renumbered, a trivial task as far as the coordinates are concerned, but a real pain for the write-up and notes. The lesson: settle a viable n-term early on. A few of the really bad obvious clashes of side chains were relieved by bond twisting, but most were left to be adjusted by Brugel running with the main chain constrained (this seemed to work well on the earlier partial set). It was noteable that Brugel steepest descents on an almost fully assigned interior holding the main chain constrained did relieve all significant collisions. So for the next round of side chain assignments almost all of the preassigned sequence was maintained.

Brugel cannot do anything with the outside side chains without solvent, so tends to put them in straight, extended form - unsightly, but there is no reason to adjust them until after doing the internal Brugel work.

The first fully assigned sequence, renumbered for the new ends now looks something like this:



a = alpha helices
B = beta helices
C = connecting loops

The first whole sequence with justification, before overall minimization and reexamination:

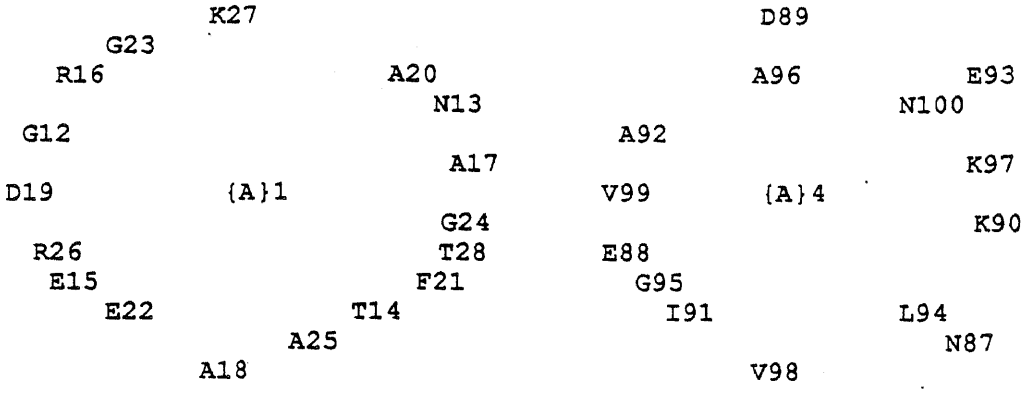
BETALPHACIN, model7 as of Tuesday, 16th Sept., 1986, renumbered model6_8 produced by defk:

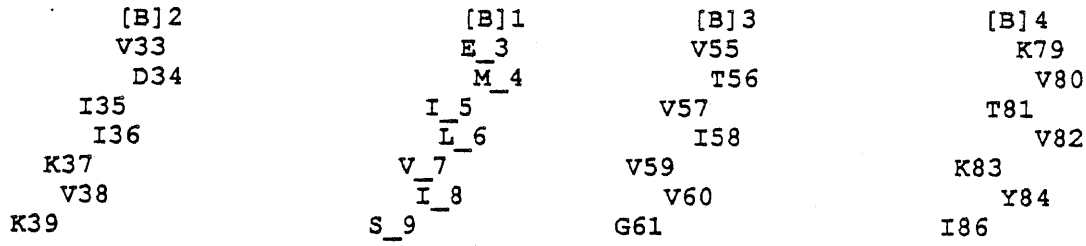
```
   1 MET          starting residue for expresion, may be processed
                  off
   2 ASP          something hydrophilic
   3 ALA    [B]1  fairly exposed, but restricted room
   4 MET    [B]1  along (C)4 toward outside of {A}2
   5 ILE    [B]1  in toward PHE on {A}1
   6 LEU    [B]1  in toward {A}2 PHE 47
   7 VAL    [B]1  inside under {A}1
   8 ILE    [B]1  against loop (C)3
   9 SER    (C)1  under PO4 inding loop, which was made
                        homologous to FXN
  10 GLY    (C)1  PO4 binding loop, N binds PO4
  11 THR    (C)1  PO4 binding loop, OG,N
  12 GLY    {A}1  PO4 binding loop, N
  13 ASN    {A}1  PO4 binding loop, N, (OE?, NE?)
  14 THR    {A}1  PO4 binding loop, OG?, N?
  15 GLU    {A}1  N term in one alpha, outside
  16  ARG   {A}1 outside
? 17  ALA   {A}1 toward (C)7, not much room, see ALA 90 on {A}4
  18 ALA    {A}1  toward [B]1,[B]2; helix orientor, homologous to
                  FXN
  19 ASP    {A}1  outside
? 20 ALA    {A}1  points toward GLU 86 {A}4, maybe could be bigger?
  21 PHE    {A}1  toward [B], helix orientor, homo. to FXN
  22 GLU    {A}1  outside
  23 GLU    {A}1  outside, on the corner of the molecule
  24 GLY    {A}1  close {A}1-{A}4 contact
  25 ALA    {A}1  inside toward [B], possible space problem,
```

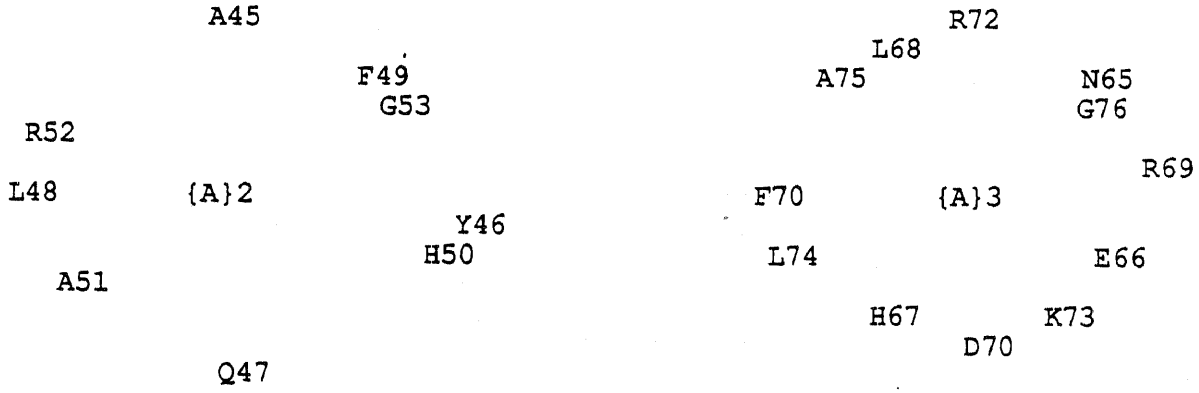
			keep small
	26	ARG	{A}1 last turn C-term, points out past loop (C)1
	27	LYS	{A}1 C-term {A}, outside
	28	THR	{A}1 C-term breaker, points between {A}1-{A}4, should be small.
	29	GLY	(C)2 really in loop, L alpha conf.
	30	THR	(C)2 tucks inside from loop, but not much room, had been ALA
	31	PRO	(C)2 looks like a good fit
?	32	PRO	(C)2 also looks like a good fit for PRO, is PRO,PRO ok?
	33	VAL	[B]2 inside toward {A}
?	34	ASP	[B]2 outside is ASP ok on parallel beta?
	35	ILE	[B]2 inside toward {A}1
	36	ILE	[B]2 needs fairly large group packed on outside toward {A}2
	37	LYS	[B]2 inside, points outside
	38	VAL	[B]2 inside toward {A}2
	39	LYS	[B]2 points outside toward ASP on {A}2
	40	ASN	(C)3 points inside toward PO4 loop, could be part of h-bond network
	41	GLU	(C)3 points outside vaguely toward HIS of {A}3
	42	GLY	{A}2 n-term breaker phi,psi: 131,145 must be GLY
	43	GLU	{A}2 n-first-turn
	44	ASP	{A}2 outside near n-term
	45	ALA	{A}2 pack against main-chain and Ile, needs to be small
	46	TYR	{A}2 between {A}2,{A}3 and out to surface
	47	GLN	{A}2 outside
	48	LEU	{A}2 points inside
	49	PHE	{A}2 into large space between {A},[B]
	50	HIS	{A}2 outside toward {A}3, maybe could tuck inside a bit
?	51	ALA	{A}2 keep small because of possible packing problem with ARG 50
	52	ARG	{A}2 c-first-turn, starts slightly in then turns out
	53	GLY	{A}2 c-term breaker in L alpha conf.
	54	ASP	(C)4 points out
	55	VAL	(C)4 points in, not much room
	56	THR	[B]3 points out
	57	VAL	[B]3 inside toward {A}1,{A}4 esp. PHE on {A}1
	58	ILE	[B]3 inside toward {A}3
	59	VAL	[B]3 inside toward {A}1
	60	VAL	[B]3 inside toward {A}3, LEU 66
	61	GLY	[B]3 part of the neat GLY-PRO connection
	62	PRO	(C)5 part of the neat GLY-PRO connection
	63	THR	(C)5 near PO4, could go in or out
	64	ALA	(C)5
	65	ASN	{A}3 N-term breaker
	66	GLU	{A}3 N-first-turn, outside
	67	HIS	{A}3 points toward 39 on (C)3, could go inside
	68	LEU	{A}3 inside
	69	GLU	{A}3 outside
	70	ASP	{A}3 outside
	71	PHE	{A}3 inside, to fill interior
	72	ARG	{A}3 mid helix, maybe cover over sheet edge
	73	LYS	{A}3 near C-term, outside
	74	LEU	{A}3 could go inside
	75	ALA	{A}3 close to phe 47, needs to be small
	76	GLY	{A}3 C-term breaker in L alpha conf.
	77	ASN	(C)6 points sort of outside
	78	LEU	(C)6 points in, room under ARG70
	79	LYS	[B]4 points out
	80	VAL	[B]4 toward {A}3 too close to VAL 55

	81	THR	[B]4	edge beta sheet
	82	VAL	[B]4	in toward {A}4
	83	LYS	[B]4	edge beta, but outside
	84	TYR	[B]4	points between PRO 60, THR 12
	85	GLU	(C)7	points out
	86	ILE	(C)7	points in
	87	ASN	{A}4	N-term breaker
	88	GLU	{A}4	N-first-turn, outside
	89	ASP	{A}4	outside
	90	LYS	{A}4	outside
	91	ILE	{A}4	inside
?	92	ALA	{A}4	points out next to {A}1, same space problem as ALA 15
	93	GLU	{A}4	outside
	94	LEU	{A}4	should be enough room
	95	GLY	{A}4	close {A}1,{A}4 contact
?	96	ALA	{A}4	outside
	97	LYS	{A}4	outside
	98	VAL	{A}4	between {A}4 and [B] sheet
	99	VAL	{A}4	points down toward ALA 23
	100	ASN	{A}4	something to extend the helix in a non-critical way
	101	GLY	{A}4	C-term end, COO- with no side chain
	102	H1	PO4	

Betalphacin helix wheels and beta wiggles, view point from C-terminal end of beta:







Distortions in the helix wheels are in part from not working out the circle very well on the printed page, and in part because the model helices are actually twisted a bit poorly in terms of ideal helix wheel geometry.

Revision of Full Sequence Model

(starts with Model7, Wed. Sept. 17, 1986)

Brugel steepest descents on all side chains with main chain constrained.

Brugel output comments:

```
SET NUMBER    1

6-12:       Evdw= 62066.477
10-12:      Ehbo= -13.182     Eelec=-60.637
TO=   62500.736         BO=   99.109      TH=   176.348     PH=   201.230
PT=   31.391            NB=   61992.658   LH=   -13.182     ELEC= -60.637
VDW=  62066.478


SET NUMBER    15, final step reported

[.MODFIL]MODEL7                                Sept. 17, 1986    10:53:56

IRST=  0 ICY=   0 ITN= 250 ITH= 250 STEP=  0.760E-03      ALPHA=  0.733E-03
DCA= 3804.        DMO= 1.288        ERMS= 33.02       DMAX= 15.08
ETO=    1.995     EBO=   29.578     ETH=  173.253     EPH=  191.515
EPL=   38.765     ENB= -464.132     ELH= -304.713     ELEC=  -98.542
EVDW=  -60.877    MNBCT=  0.029     RMSORI=  0.237    BIGMVT=  0.256

RMS between first set and last set of energy minimization on main
chain

Set   1 - Set  15  RMS=          0.02021

RMS between first set and last set of energy minimization on side
chains

Set   1 - Set  15  RMS=          0.34483
```

Chou-Fasman prediction of secondary structure, run from UWGCG package at EMBL using routine ChouFas (and Plotchou for a fancy but misleading [ed.] plot!); Brugel display facility was used to visualize the Chou-Fasman results. Different colors represented different levels of prediction in the model. This was very powerful, several residues that could affect the prediction were easily identified and new predictions calculated. Surface residues were chosen since we wished to avoid disturbing the internal packing in this change.

Model7 sequence as stated above:

```
          1         2         3         4         5         6
12345678901234567890123456789012345678901234567890123456789012345678901234567890
  bbbbbb    aaaaaaaaaaaaaaaaa     bbbbbbb   aaaaaaaaaaaa  bbbbb  <model

MDAMILVISGTGNTERAADAFEEGARKTGTPPVDIIKVKNEGEDAYQLFHARGDVTVIVV
  BBBBBBttTTtHHHHHHHHHHHHHHHH     hhhhhhh tt  BBBBBBTTt BBBBBB  <prediction

6         7         8         9         0
12345678901234567890123456789012345678901
b   aaaaaaaaaaaa  bbbbbb  aaaaaaaaaaaaaaa <model
GPTANEHLGDFRKLAGNLKVTVKYEINEDKIAELGAKVVNG
 tt  hhhhhhhhhh     BBBBB    TTHHHHHHHHHH   <prediction
```

Notice that [B]2 predicts weakly as helix, and {A}2 strongly as beta.

Model7 modified sequence:

```
        1         2         3         4         5         6
12345678901234567890123456789012345678901234567890123456789012345678901234567890
  bbbbbb    aaaaaaaaaaaaaaaaa     bbbbbbb   aaaaaaaaaaaa  bbbbb <model
                                   D  K          QL              <old sequence
MDAMILVISGTGNTERAADAFEEGARKTGTPPVTIIYVKNEGEDAYEKFHARGDVTVIVV
  BBBBBBttTTtHHHHHHHHHHHHHHH     BBBBBB  ttHHHHHHHHHH  BBBBBB <prediction

6         7         8         9         0
12345678901234567890123456789012345678901
b   aaaaaaaaaaaa  bbbbbb  aaaaaaaaaaaaaaa <model
GPTANEHLGDFRKLAGNLKVTVKYEINEDKIAELGAKVVNG
 tt  hhhhhhhhhh     BBBBB    TTHHHHHHHHHH  <prediction
```

Robson prediction on original model7:

```
          1         2         3         4         5         6
12345678901234567890123456789012345678901234567890123456789012345678901234567890
  bbbbbb    aaaaaaaaaaaaaaaaa     bbbbbbb   aaaaaaaaaaaa  bbbbb <model
MDAMILVISGTGNTERAADAFEEGARKTGTPPVDIIKVKNEGEDAYQLFHARGDVTVIVV
HHHHEEEECCCCCCCHHHHHHHHHHTTTCCCEEEEEEECCCCHHHHHHHHCTTTEEEEEE  <prediction
```

Robson prediction on MODIFIED model7:

```
          1         2         3         4         5         6
12345678901234567890123456789012345678901234567890123456789012345678901234567890
  bbbbbb    aaaaaaaaaaaaaaaaa     bbbbbbb   aaaaaaaaaaaa  bbbbb <model
                                   D  K          QL              <old
MDAMILVISGTGNTERAADAFEEGARKTGTPPVTIIYVKNEGEDAYEKFHARGDVTVIVV
HHHHEEEECCCCCCCHHHHHHHHHHTTTCCCEEEEEEECCCHHHHHHHHHHHTTEEEEEE
                                        *        **         <diff

6         7         8         9         0
12345678901234567890123456789012345678901
b   aaaaaaaaaaaa  bbbbbb  aaaaaaaaaaaaaaa <model
GPTANEHLGDFRKLAGNLKVTVKYEINEDKIAELGAKVVNG
CCCCHHHHHHHHHHHHHHHEEEEHHHHHHHHHHHHHHHEEE
```

and this {A}3 prediction is lower than the other {A} predictions.

Although the sequence change does not affect the prediction so dramatically in the Robson program, the graphed potentials for the beta state does improve in the [B]2 region while the helix potential drops. Similarly, the helix potential rises and the beta falls in the {A}2 region.

Changes made to improve secondary structure prediction:

```
34 ASP          ->   THR   [B]2  on surface, has an indented region it
                                  might go in
37 LYS          ->   TYR   [B]2  and might better fill space in this
                           [B]{A} contact
47 GLN          ->   GLU   {A}1  outside
48 LEU          ->   LYS   {A}1  outside
                           Result: -> Model7_2, sequence modified.
```

Display of Predicted Forces Remaining in the Brugel Minimized Model

Note that the main chain was constrained since we wished to avoid having bad side chain contacts initially relieved by main chain motion. Now, we needed an evaluation of packing to see if side chain torsions or replacements had to be made before global minimization. Hots spots of the remaining predicted forces were displayed by Brugel as red for real strained, yellow for less. Observed that the C-term end of {A}4 is strained, as well as [B]4 and (C)7 up to {A}4. There is still a little unrelieved strain in the nearly stacked PHEs in the bottom rear under {A}2,3. And some problem still in the N-term region of [B]2, [B]1, (C)6, and a few other places.

Connolly surfaces were calculated and displayed with Brugel. The selection was buried surface in the model, each helix separately, the beta structure as a unit, and the loops as a unit. This leaves some ambiguous regions at the edge where it is unclear whether a lack of dots is a poorly packed region, or a sensible undulation in the accessible surface. However, one must balance information displayed with information that is digestible. Helices, beta, and loops all colored differently, usually visualized with both main and side chains, gave a most dramatic, beautiful, and revealing presentation of the packing. Setting and moderate clipping depth and adjusting the z-translation while using the other rotations to view selected regions was very effective in isolating views.

The really disturbing aspect was how well this model seemed to be packed! The helix-helix packing looked excellent. The beta-

helix interface was remarkably well-fit - though there were a few holes. For example, a region about a methyl in size between PHE 71 and the main chain of residue 7. There is a lot of narrow spaces between {A}3 and [B]3, [B]4; one wonders if a free refinement would bring them closer together (but note the earlier observation from the loop fitting process that secondary structure elements might be unusually close together already). There is an odd sort of hole behind HIS 67 over toward 71. One wonders if 68 could be bigger, or if 60 were bigger it might push over toward 8 and in the end result in better packing. There were no obvious, simple replacements that would improve the packing. Since the refinement changes of position as visualized by Brugel's movie feature were as much as an Å or so with rotations of about 50° in extreme, it seems that Brugel had accomplished what one would have wanted to do with hand bond tweaking. With a nod to the skill of the group in selecting and initially putting in the residues, and to Brugel for cleverly positioning them, we proceeded to the next step of full minimization of the model.

It was decided to put hydrogens explicitly on the model. This not only would improve the results of packing refinement, but, very importantly, would allow Brugel to better repair the somewhat distorted hydrogen bonding in both the beta sheet and alpha helix.

Successive steps:

Model7	refined, all main chain constrained.
Model7_2	modified sequence.
Model7_2H	hydrogens put on with François' program.
Model7_2HCREF	refined again, constraints on all C(alpha), 500 steps.
Model8	refined yet again, constraints on C(alpha) of beta strands alone, 200 steps.

Abandoned steps:

Model7_2H -> Model7_2HREF:
refined with no constraints from first model with hydrogens: beta sheet torn in the middle, whole structure expanded; 500 steps.

Model7_2HCREF -> MOD7:
refined with no constraints from model first refined with constraints on all ca. 500 steps. Beta sheet torn in the middle, whole structure expanded. One begins to suspect that we had packed too much stuff in the interior!

Model7_2HCREF:
A lot of local tweaking occurred, as seen by Brugel animation. In general, things look even better packed and better h-bonded. Connally surfaces look even better than before.

Model8 : (vs model7_2H, rms = .96)
Tendency for expansion, helices moved about 1 Å apart. Helices 1,2,4 moved out from beta, at least in the middle, so actually bow out making them curved. Helix 3 did not seem to move out from beta sheet, in line with the observation on the earlier model that there was some slight space there. Perhaps the lesson is that a beautifully packed interior, as seen with the matched surfaces, is too tightly packed. In this model, many surface h-bonds were made, including one from the TYR on [B]2 to GLU on {A}1 with the TYR finally packing in the depression behind it.

Analysis of polar/nonpolar surface (Cornelius Frömmel):

	1FXN	Betalphacin	
Fudged surface polar	38%	20%	
" " nonpolar	22%	66%	
Total accessible surface	32%	39%	
Hydrophobic stabilization energy	-79	-11	
Solvation energy	+362	+454	(destabilizing)

At this point there are two obvious ways to proceed. One line of work would be to go back a step, using the expansion as a guide to volume analysis and try to understand how to change side groups and by using these tools, obtain a different packing.

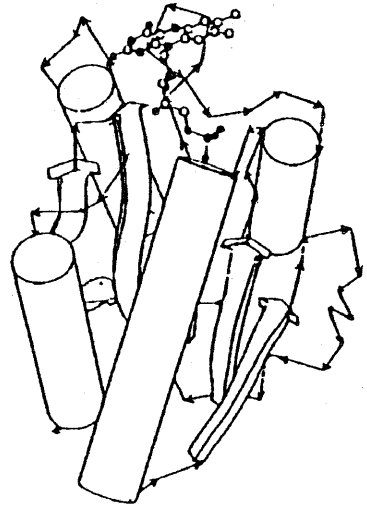
Another approach is to use this last step as a valid way of changing the model, that is, declare that this fitting of the main chain alpha and loops around the interior results in a model acceptible as a design. It would be then desirable to refine and refit on this model. There is a little unrelieved force on some of the atoms as shown by Brugel. We had observed that FRODO found loops indicated that the secondary structure elements had been too close together anyway. (One should check FXN to see what the spacing really is there!) On this basis, we deposit these coordinates.

Model8 (with hydrogens removed from the file) deposited as Betalphacin.pdb [Sept. 18, 1986]

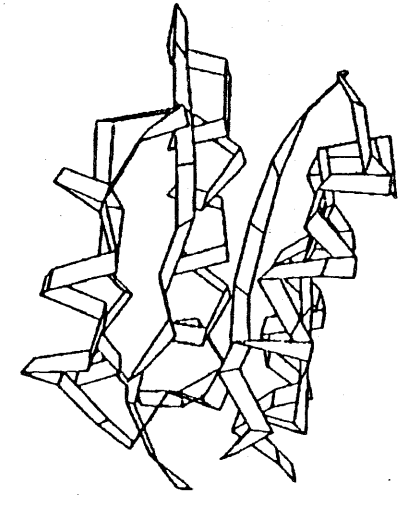
Appendix: Key to Files in [cutsem.pdb]

model2	:	Helices + betas
model3	:	Helices + betas after refinement
model4	:	
model5	:	Helices, betas and loops
model6	:	idem 5 but after refinement
model7	:	side chains on all
model7_2	:	4 substitutions to satisfy sec.struct. pred.
model7_2H	:	all hydrogens on
mod7_2Href	:	refinement from model7_2H, without constr. (500 steps)
mod7_2Hcref	:	constraint ref. on C(alphas), refin. of entire structure (500 steps)
model8	:	constraint ref. on betas C(alphas), ref. on all structure (200 steps)

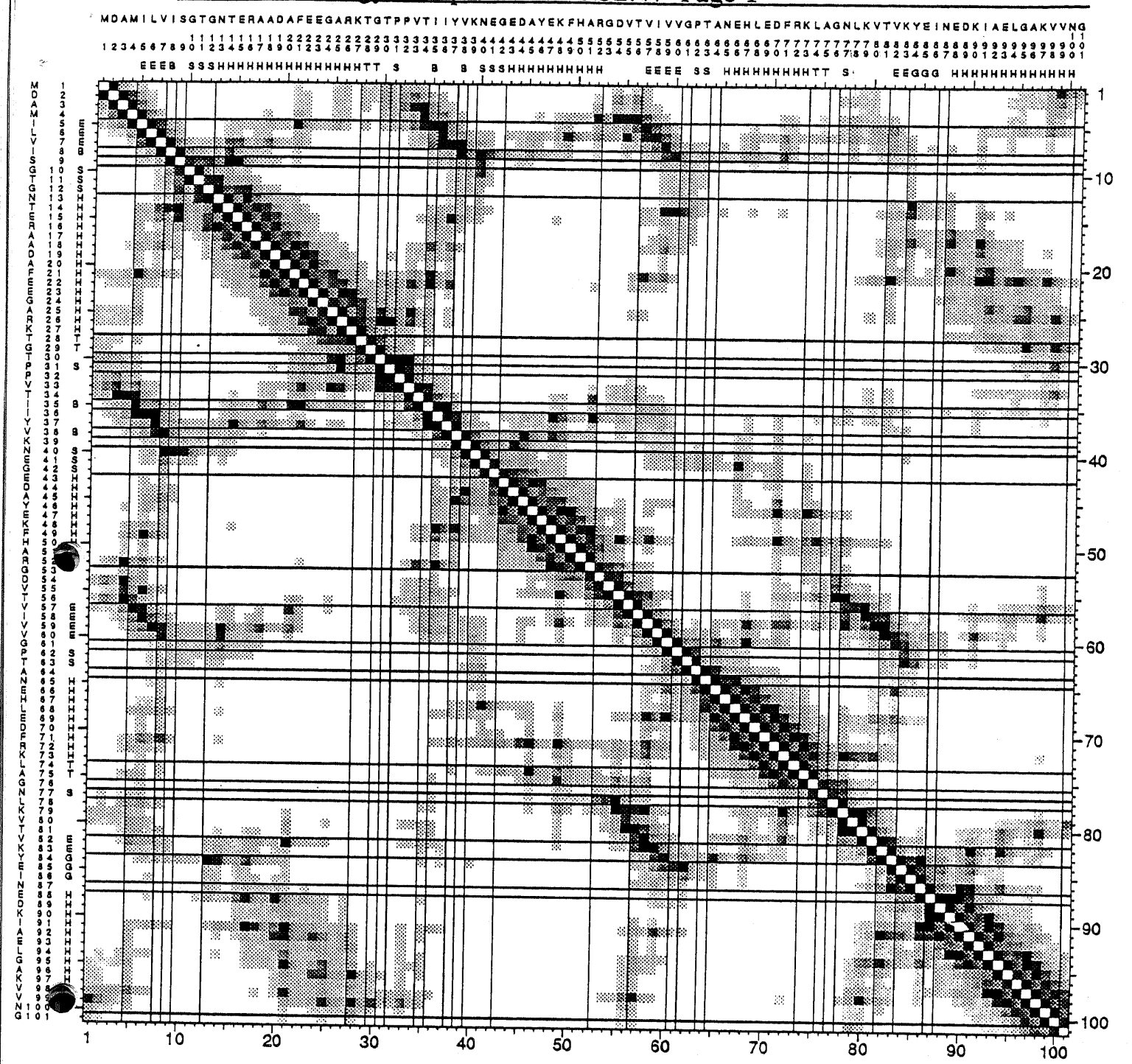
```
5)  Betalphacin.....BEAL
    SHEET....      AAAB                                  A  B
    BRIDGE2..      bbb
    BRIDGE1..      a  d                                  a  d
    CHIRALITY   -++-+-+++ -+++++++++ +++++++-+- ++----+--+ +-+++++++++
    BEND.....           S SSSSSSSSSS SSSSSSSSS  S        S SSSSSSSSSS
    5-TURN...
    4-TURN...               >>>>XXXXX XXXX<X<<4<                >>>>XXXX<
    3-TURN...               >33<>33<       >33<                 >33<   >3
    SUMMARY..      EEEB S SSHHHHHHHHH HHHHHHHTT  S   B  B S SSHHHHHHHH
    EXPOSURE.   *732000225 *1609*00*1 06*00**674 9*161172*5 *4**05**1*
  1 SEQUENCE.   MDAMILVISG TGNTERAADA FEEGARKTGT PPVTIIYVKN EGEDAYEKFH

    SHEET....        AAAA                          AA
    BRIDGE2..          cc
    BRIDGE1..        bbb                           cc
    CHIRALITY   +++++++--+- -+++++++++ +++++++++- ---++--+++ +++++++++
    BEND.....   SS           SS SSSSSS SSSS  S       SSS SSS SSSSSSSSS
    5-TURN...
    4-TURN...   <<<            >>>>XXX <X<<4<           >>>> XXXXXXX<<<
    3-TURN...   3<             >33<     >33<         >>3<X33<     >33<
    SUMMARY..   HH    EEEE  SS HHHHHH HHHTT S     EEGGG HHH HHHHHHHHHH
    EXPOSURE.   9848030001 2*755*71*9 19*726*2*3 3089*77*9* 02*402*21*
 51 SEQUENCE.   ARGDVTVIVV GPTANEHLED FRKLAGNLKV TVKYEINEDK IAELGAKVVN

    SHEET....
    BRIDGE2..
    BRIDGE1..
    CHIRALITY
    BEND.....
    5-TURN...
    4-TURN...   <
    3-TURN...
    SUMMARY..
    EXPOSURE.   6
101 SEQUENCE.   G

Notation:  SUMMARY H=ALPHA-HELIX
                   E=BETA-STRAND
                   B=BETA-BRIDGE G=3-HELIX I=5-HELIX
                   T=3-,4-, OR 5-TURN
                   S=BEND
```

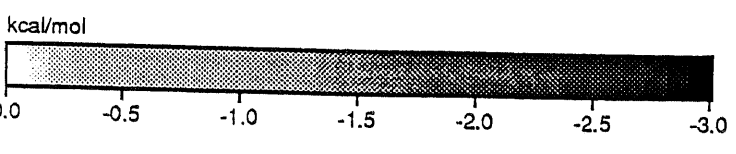


Native Flavodoxin

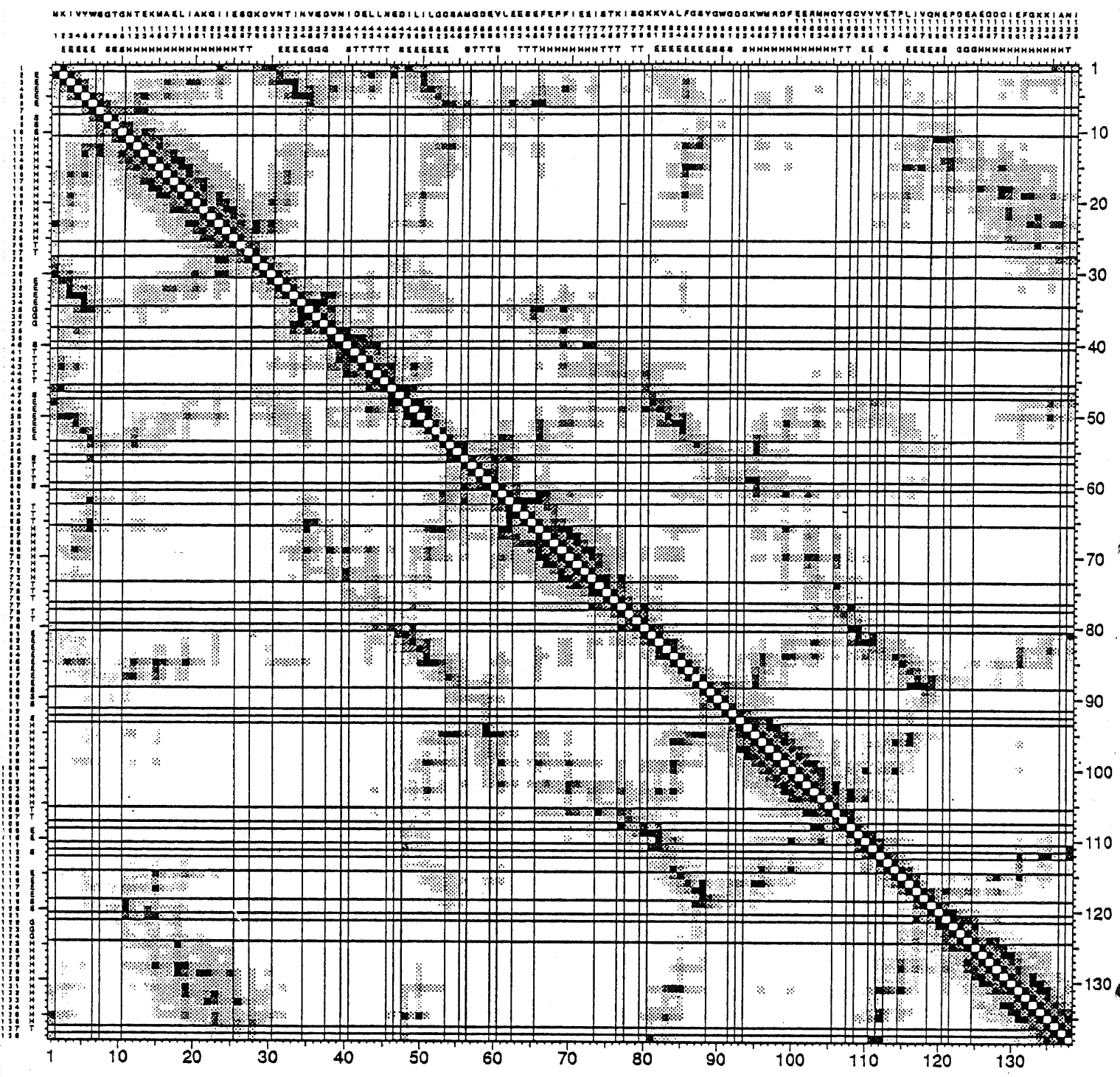


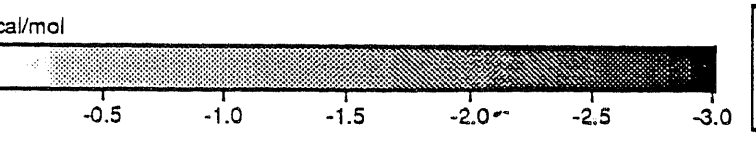
BEAL MODEL


MDAMILVISGTGNTERAADAFEEGARKTGTPPVTIIYVKNEGEDAYEKFHARGDVTVIVVGPTANEHLEDFRKLAGNLKVTVKYEINEDKIAELGAKVVNG
1 10 20 30 40 50 60 70 80 90 100


kcal/mol
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0

3FXN: Van der Waals energy.





John Moult
Cornelius Frömmel
Johan Postma
Arne Skerra
Alfonso Valencia

* DESIGN OF A HELIX BUNDLE *
(BUND)

General Considerations

Why a helix bundle?

1) It seems necessary to imitate a known structural motif: We do not understand the rules that govern them well, and therefore cannot say what unobserved ones are possible. Much might be learned by trying to build a new one, but not in this time frame.

2) Four motifs were considered: the alpha helix bundle, the anti-parallel beta sandwich, the alpha/beta slab, and the alpha/beta barrel. Of these, the simplest is the alpha bundle. It is found in nature with a relatively small size, and a fairly regular structure.

3) The locally (EMBL) solved structure ROP is the simplest and neatest structure so far determined. The local crystallographers (David Banner, Michael Kokkinidis) were very, very helpful both in passing on their insight, and in providing co-ordinates.

Choice of Function

It is desirable that the protein should do something, both as test of the design, and as a more pleasing outcome. On the other hand, the departures from regularity in accommodating a substrate, and the precision probably required, greatly complicate the design. A common interest of the group was in calcium binding sites, and as

these have been relatively well characterized, and require relatively simple arrangements to form a binding site, this function was selected. A desirable extension of this is to have the binding event cause a conformational change. However, it was decided that the first design should not have this ability but should as far as possible leave room for its subsequent incorporation.

Design Strategy

Basic Bundle

The structures of ROP, hemerythrin and cytochrome C562 were used as examples upon which to base the design. The initial concept was to assemble four straight helices with a crossing angle of 18°, an approximate concensus value. At the point of doing so it was realised that real helix bundles do not contain straight helices, but are supercoiled around each other. The reason for the supercoiling is unclear (to us, anyway). It does not appear to be a consequence of any "knobs into holes" type of core packing. It does imply some systematic variation of the backbone geometry as a function of the position on the alpha helix, and so may reflect the effects of solvent accessibility. A possibly related feature is a 3.5 residue per turn pitch, which is necessary to produce a repeating pattern of Ca positions with respect to the bundle axis, with a repeat every two turns.

All structures depart from this ideal somewhat: ROP has a varying curvature of its helices, cytochrome C562 has one pair splayed apart to fit in the heme group, and hemerythrin has a slight widening towards one end of the bundle to accommadate the iron binding sites. Richard Bryan has written a program which fits superhelix parameters to bundles, and we used the results of this on ROP to generate a set of four spirals around a common axis.

Four super helices were generated with a pitch of 172.5 Å, and a radius of 7 Å. The alpha helix has a pitch of 3.5 residues per turn, with 1.48 Å rise per residue, and a radius of 2.3 Å. The helices were generated using the parametric description of Crick (1953). Comparison of the resulting Ca positions with those of the ROP structure verified the parameters were appropriate. Further details of this step are given below.

The other main chain atoms were added to the helices by first producing an ideal straight helix with dihedral angles -60, -42. This helix was distorted so that its Ca atoms coincided with those of each spiral in turn by setting the Ca coordinates as targets in GROMOS and performing an energy minimization. This worked well. Further details of all the applications of GROMOS are given below. Initial rough alignments were made on a graphics system. Helices opposite each other run parallel, and helices adjacent to each other, anti parallel. Apart from the difference in helix direction with respect to the bundle axis, the helices are very similar, and are equi-distant from the bundle axis, with a four-fold symmetry between the local helix axes about any point on the bundle axis. Helices running in opposite directions are one turn out of phase from each other. The phase of the helices relative to the bundle axis was adjusted to obtain the orientation of the inward pointing Ca atoms required for the packing scene (see below).

Loops between the Helices

Four helices in one chain implies three connecting regions. Two of these, at one end of the bundle, are symmetry related to each other by a two fold-axis along the bundle axis. This end was chosen to be the position for calcium binding, with the objective being to position the calcium roughly at the end of the helices, in the center of the bundle.

Thus, loops of the miniumum possible length were required at that end. The ROP structure contains such a loop, with only two residues not in the flanking helices. The ROP helices are also similarly oriented to those in the design structure at this position. The main chain atoms of the ROP loop were thus aligned with the design structure, using MIDAS. The loop at the other end of the bundle is not equivalent to those at the calcium binding end. An attempt was made to use the same ROP loop, but it was clearly inappropriate. A suitable conformation was sought using the loop matching facility of FRODO, specifying the Ca atoms of the last 3 residues of each of the flanking helices as match points, and leaving 5 residues in between. (Three of these residues were helical, and 2 were the inappropriate ROP loop). Loops ranging from 1.1 Å RMS were obtained, with no strong tendency for clustering of their conformations. The best fit one was selected, residues 26 to 36 in carboxypeptidase.

Key features of this loop are a proline residue, and an across loop hydrogen bond. All three loops were fitted more properly to the helices using using GROMOS with the helix Ca atoms as target co-ordinates to prevent distortion of the bulk of the structure.

Calcium Binding Site

There are a number of calcium binding sites in proteins of known structure. Those in the Brookhaven data bank were surveyed to gain some insight into typical liganding arrangements, particularly for cases where the site is formed from ligands from distant parts of the sequence. However, the best ordered sites were clearly those from the proteins that bind calcium for specific functional reasons, namely parvalbumin and intestinal calcium binding protein. The four sites in these proteins display distorted octahedral coordination, with 3 or 4 charged ligands out of the total of 6 or 7. The site between the A and B helices of parvalbumin was selected as a model for imitation. This has four carboxylate groups, a main chain carbonyl group, and a serine hydroxyl group as ligands. Although the structure is nominally refined it was clear from the bond angles and bond lengths that it is not very reliable. Nevertheless it seems the most appropriate.

The relevant residues were removed as a unit and superposed onto two loop ends of the helix bundle using MIDAS. Three glutamate residues, an aspartate and a glutamine were introduced on the bundle so that the ends of the side chains were close to those of the calcium ligands in the parvalbumin site (the glutamine side chain carbonyl group replaces the main chain carbonyl of parvalbumin in this scheme). This alignment was achieved using side chain dihedral angles close to staggered values. More precise liganding was obtained using distance constraints in GROMOS amongst the ligand atoms and to the calcium. This arrangement left the site occupied by the serine hydroxyl in parvalbumin exposed to solvent roughly on the bundle axis, and it is intended that a water molecule should ligand at this position.

It is further intended that in the absence of calcium, the liganding side chains may rotate away from each other, as is the case in the empty calcium binding sites of the X-ray structure of troponin C. Some further tuning of this site is desirable.

Core Packing

Examination of the core inside the helix bundles of known structures showed a variety of arrangements and apparent packing efficiencies, as judged from van der Waals surface representations obtained by using FRODO. A particularly simple looking arrangement is one in which residues from each of the four helices lie in a plane approximately perpendicular to the bundle axis. ROP provides the most striking example of this arrangement, though it does occur elsewhere. We would here again like to acknowledge particularly the role of David Banner and Michael Kokkinidis in bringing this arrangement to our attention.

Typically, two relatively large hydrophobic side chains (LEU or ILE) point in towards each other from one pair of helices on opposite sides of the bundle, and two smaller side chains (ALA, THR, CYS) point to neighboring helices in the bundle so that these residues lie on either side of the large residue pair. When THR is involved, the OH group is directed out from the bundle core. An appropriate total volume seems to be a total of three methyl groups contributed by these smaller residues.

The helix bundle backbone was constructed so that helices running in the same direction were in register. That is, Ca atoms occur at the same angle on the alpha helices and at the same height on the super helix axis. The pair of helices running in one direction were then translated along the super helix axis by an amount of nearly one alpha helix turn, 3.5 residues. This arrangement, with an appropriate initial position for the first residues, results in planes with the 'a' positions on one pair of helices approximately in the same plane as the 'd' positions of the other pair. Large side chains are then attached to the 'd' positions, small ones to the 'a' positions. There are 5 such planes in the design core, with 2 LEU residues in each layer, originating from alternate pairs of helices. Each layer has one flanking ALA and one THR in the 'a' positions. A sixth layer at the chain termini end of the bundle has ASN residues instead of LEU at the interface with solvent. At the calcium binding end, an ILE and a THR residue pad the space at the back of the calcium ligands.

Interhelix Packing

The space between adjacent residues is partly filled with GLN residues, with the hydrophobic portion of their side chains stacked against the inward pointing LEUs. The other helix-helix contacts have THR hydroxyl groups protruding, hydrogen bonding to ASN side chains on the outside of the helix.

Salt Bridges

Each helix has one salt bridge spanning two turns at the outermost positions, with an ALA residue on the bridged turn. There is also one salt bridge between two helices.

Markers

There are only one TYR and no TRP residues. The TYR is adjacent to the calcium binding site, so that changes in fluoresence can be used to monitor binding. Helix formation can be monitored by CD.

The Rest of the Surface

No careful design was executed here, just spattering of charges, polar side chain and a few ALA residues; appropriate charged residues to compensate for the helix dipoles.

Checking the Structure

Packing Efficiency

The overall volume inaccessible to solvent and not inside the protein van der Waals envelope, i.e. the volume of holes, was calculated using PACANA. Before energy minimization of the model, this amounted to 3900 Å3, after minimization, 4100 cubic Å3. The same quantity for ROP, the best packed looking helix bundle, is 4900 Å3 for the same number of residues. Thus the model is, by this criterion, better packed than the best natural structures. Examination of a

break down by atom type of the packing showed this efficiency to be distributed over all sorts of groups.

Secondary Structure Prediction

The method of Robson et al predicted 60% of the helical regions of the model to have this structure.

Secondary Structure Type

The DSSP program applied to the atomic coordinates found the following number of residues to be in helical conformation:

	before minimization	after minimization of bundle8_40 (weak constrain
helix (plan: 90)	87	83
extended (plan: 0)	0	0

Surface Properties

The program MAXSURF found the following amounts of polar and non-polar accessible surface (% of maximal values in extended conformation):

	before minimization	after minimization of bundle8_40 (weak constraints)
polar surface : (normal: 37%)	33%	29%
apolar surface : (normal: 33%)	40%	44%

These numbers indicate too much apolar surface and not enough polar surface, a surprising result for this structure. The cause of this maybe the proximity of the polar side chains to the underlying surface. A more realistic conformation might be to have the side chain ends turned away from the surface. Such a change would

certainly improve the numbers, but would not be realistic. Such questions could probably be settled by an extensive molecular dynamics simulation.

Amino Acid Composition

```
# AA number

  A     21
  R      2
  N     15
  D      3
  C      0
  E     10
  Q     13
  G      2
  H      1
  I      3
  L     11
  K      8
  M      0
  F      0
  P      1
  S      9
  T     11
  W      0
  Y      1
  V      1
============
      112
```

A rather odd looking composition, but so what?

Homology to other proteins sequences (PIR/NBRF identifiers)

```
>P1;TVBYR2  :     Transforming protein homolog (H-ras-2) - yeast
                  17.9% identity in 78 aa overlap

>P1;FOFV1R  :     gag polyprotein - Rous sarcoma virus
                  16.4% identity in 67 aa overlap

>F1;GNFVFP  :     gag-fps polyprotein - Avian sarcoma virus
                  14.9% identity in 67 aa overlap

>P1;DFPG    :     Beta-neoendorphin-dynorphin precursor - Pig
                  20.3% identity in 69 aa overlap

>P1;TMHOBP  :     Tropomyosin beta chain, platelet - Horse
                  21.1% identity in 76 aa overlap
```

```
>P1;TMRBA    :    Tropomyosin alpha chain, skeletal and cardiac muscle
                  20.2% identity in 99 aa overlap

>F1;TMCHA    :    Tropomyosin alpha chain, skeletal muscle
                  18.9% identity in 95 aa overlap

>P1;HMXRS3   :    Hemagglutinin - Reovirus (type 3)
                  18.0% identity in 111 aa overlap

>P1;LPRTA4   :    Apolipoprotein A-IV precursor - Rat
                  17.6% identity in 85 aa overlap
```

Nothing significant here. ROP was not included in the data base used.

Conclusions

1) The core is nice.
2) The calcium binding site is doubtful since the non-calcium bound state may not allow the liganding side chains to get far enough away from each other. This is correctable though.
3) The loop at the chain termini end is not well stabilized, but this should not be important.
4) The surface 'design' is unsatisfactory, and tests of overall charge distribution and water structure supporting properties are desirable.
5) Apart from the loose ends, I like it.

Appendix: Generation of the Backbone-Coordinates for the Coiled-Coil Structure

The coiled-coil formed by each of the four helices making up the bundle was produced by application of the parametric equation for a continuous coiled-coil given by Crick (F. H. C. Crick, The Fourier Transform of a Coiled-Coil, Acta Cryst. 6, 685 (1953)). The x-, y- and z-coordinates (in Å) for the alpha-carbon atoms of one helix are given as follows:

$$x = r_0 * \cos(w_0 * t) + r_1 * \cos(w_0 * t) * \cos(w_1 * t + w_{10}) \; r_1 * \cos(a) * \sin(w_0 * t) * \sin(w_1 * t + w_{10})$$

$$y = r_0 * \sin(w_0 * t) + r_1 * \sin(w_0 * t) * \cos(w_1 * t + w_{10}) + r_1 * \cos(a) * \cos(w_0 *) * \sin(w_1 * t + w_{10})$$

$$z = p * w_0 * t / (2. * pi) - r_1 * \sin(a) * \sin(w_1 * t + w_{10})$$

Parameters:

Radius of the major helix: $r_0 = 7.0$ Å
Pitch of the major helix: $p = -172.5$ Å
Radius of the minor helix: $r_1 = 2.3$ Å
Pitch of the minor helix: $t_1 = 3.5 * 1.48$ Å

(pitch is 3.5 residues per turn, with a separation of 1.48 A in the direction of the alpha helix axis.)

Offset-angle of the minor helix:

helix #1 and #3: $w_{10} = (180. - 25. - 2. * 360. / 3.5) * pi / 180$
helix #2 and #4: $w_{10} = (180. + 25. - 2. * 360. / 3.5) * pi / 180$

Derived parameters:

$a = \text{atan}(2. * pi * r_0 / p)$
$t_0 = p / \cos(a)$
$t_{10} = t_1 * w_{10} / (2. * pi)$
$w_0 = 2. * pi / t_0$
$w_1 = 2. * pi / t_1$

The coordinates for the first carbon atom in each helix were computed for t=0 and all succeeding positions with an increment of delta-t=1.48. The final coordinates for the second, third and fourth helix were obtained by rotating the positions derived from the above equations around the z-axis by 90°, 180° and 270° respectively. The first carbon atoms in the first and the third helix correspond to the N-terminus, whereas in the second and the fourth helix they represent the C-terminus. All parameters are averaged values from the original ROP1 structure.The positions for the atoms of the peptide bonds were then obtained by fitting of ideal alpha helices to the computed coordinates. This step was carried out using the GROMOS energy minimization routine with position constraints.

Appendix: Applications of GROMOS for Bundle Construction

GROMOS = Program to minimize energy potential (molecular dynamics part was not used).

Used to pull C(ALPHA) coordinates of an ideal helix to geometrically derived C(ALPHA) positions of an idealized supercoil (about 90 position constraints are fixed, loops are not fixed); weight of the position constraints 9000; harmonic oscillator force.

Used to fix the positions of polar side chain atoms relative to each other in order to obtain local geometry comparable to the known calcium ion binding site taken from CPA (position constraints to fix the helical conformation, distance geometry to maintain the 'internal' geometry of the binding site); the following atoms were included in the distance geometry calculation: both strong polar atoms of the side chains of GLU, GLN, ASP and the connected carbon atom. As the main chain carbonyl oxygen in the original Ca++ binding site was replaced by GLN, the position of the nitrogen was unknown, so for this side chain no distances were used. In all 98 distance contraints were used; force constant 1000 and 4000 was used, the assumed error of each distance was +/-0.5 Å.

Results:

Step 1:
*4 straight ideal helices ----> supercoiled helices;
*error in postions between 0.1 and 0.8 Å;
*number of constraints 88 of 117 amino acids;
*starting potential energy +1.0 E+6;
*final potential energy -4.0 E+3 after 100 cycles
(keep in mind: no [or only five] side chains).

Step 2:
*regularisation of the graphically roughly built loops while fixing the alpha carbon positions of the supercoiled helices; alternatively, one could use FRODO for regularisation, but in this case there are some difficulties with close contacts, so regularisation including non-bonded interaction is necessary (GROMOS).

Step 3: *supercoiled helix bundle ---> supercoiled bundle, but superposition helix1 ---> helix2 and so on;
*error in position up to 1.5 Å;
*number of CA constraints 88 of 117 amino acids;
*only 5 side chains;
*minimization stops in a local minimum, all peptide bonds in the constrained helix positions are in the cis conformation (starting energy +E9; final +E4);
*after refinement using FRODO, another 100 cycles GROMOS with final energy -0.2E+4.

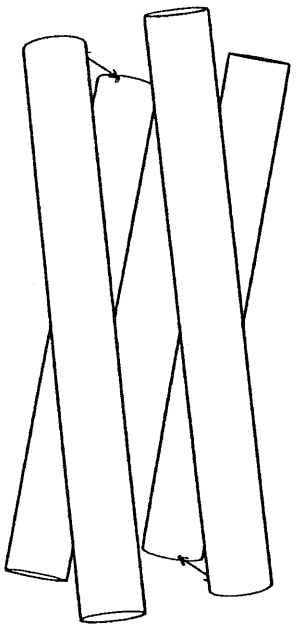
Step 4: *energy minimization of bundle8_40 including all side chains (total 112 amino acids: with CA constraints (94) (force constant 9000) and distance constraints (98) (force constant 4000) for the calcium ion binding site;
*starting potential energy about 1.0 E+6 kcal/mol, after 550 cycles -1.5 E+3 kcal/mol;
*the final energy of the contraints is about 200 kcal/mol.Main chain movements up to 0.2 Å.

Distance constraints for calcium binding site ligand geometry.

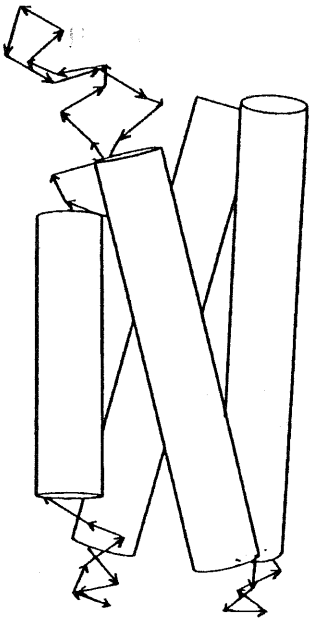
OE1,OE2,CD	===>	GLU	22,30,89
OD1,OD2,CG	===>	ASP	86
OE1,CD	===>	GLN	27

Assumed error: +/- .5 Å

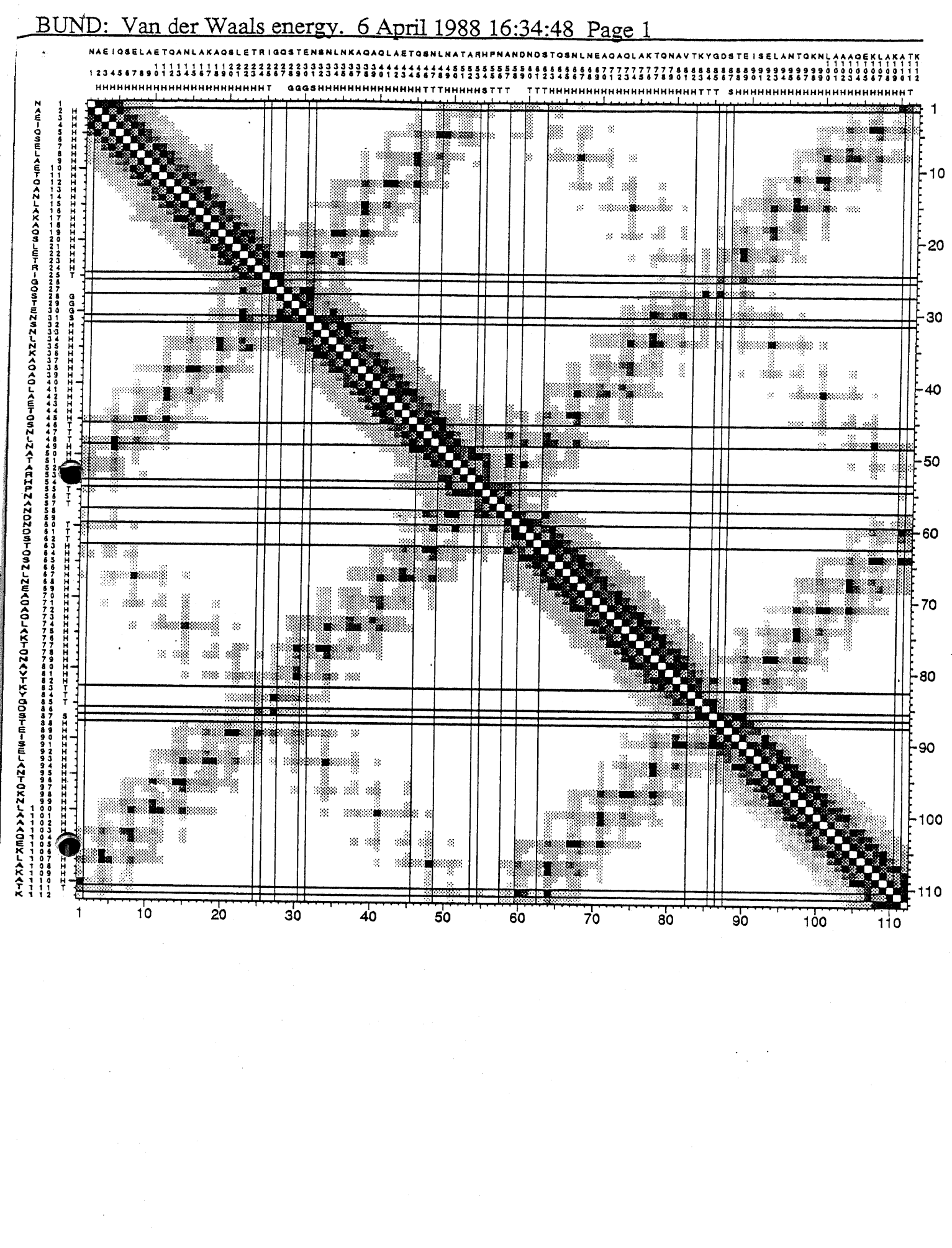
```
6)  Bundle.....BUND
      SHEET....
      BRIDGE2..
      BRIDGE1..
      CHIRALITY   +++++++++ +++++++++ ++++-+-+++ +++++++++ +++++++++
      BEND.....    SSSSSSSS SSSSSSSSSS SSSSS  SSS SSSSSSSSSS SSSSSSSSSS
      5-TURN...
      4-TURN...  >>44XX4>X< >XX>X<XXX> X<<<<       >>>>XXXX<< XX4><X4X>X
      3-TURN...  >33<   >33 <  >33X33<  >>3<<>>3< X33< >>3<< >33<   >>3
      SUMMARY..   HHHHHHHHH HHHHHHHHHH HHHHT  GGG SHHHHHHHHH HHHHHTTTHH
      EXPOSURE.  96*126602* 013603*117 706*8629*0 17903*0258 03*1156050
    1 SEQUENCE.  NAEIQSELAE TQANLAKAQS LETRIGQSTE NSNLNKAQAQ LAETQSNLNA

      SHEET....
      BRIDGE2..
      BRIDGE1..
      CHIRALITY  ++++++++-+ +++++++++ +++++++++ +++-+-++++ +++++++++
      BEND.....  SSSSSSS  S SSSSSSSSSS SSSSSSSSS SSSS  SSSS SSSSSSSSSS
      5-TURN...                                  >5555<
      4-TURN...  4<<X444<>4 4>X4>XX4XX X>XXXXXXX< <<<   >>4> X<>XX>X<XX
      3-TURN...  << >33< >3 3X33<>33X3 3<   >33<  >>3<< >33X 33< >>3<<
      SUMMARY..  HHHSTTT  T TTHHHHHHHH HHHHHHHHH HHTTT SHHH HHHHHHHHH
      EXPOSURE.  36*97*91*1 78106704*0 14402*1095 07**729*01 6801900*70
   51 SEQUENCE.  TARHPNANDN DSTQSNLNEA QAQLAKTQNA VTKYGDSTEI SELANTQKNL

      SHEET....
      BRIDGE2..
      BRIDGE1..
      CHIRALITY  ++++++++++
      BEND....:  SSSSSSSSSS
      5-TURN...        >555 5<
      4-TURN...  <4XX4>X<4< <
      3-TURN...       >>3<<
      SUMMARY..  HHHHHHHHHH T
      EXPOSURE.  35118*01*4 3*
    101 SEQUENCE.  AAAQEKLAKA TK

otation:  SUMMARY H=ALPHA-HELIX
                  E=BETA-STRAND
                  B=BETA-BRIDGE G=3-HELIX I=5-HELIX
                  T=3-,4-, OR 5-TURN
                  S=BEND
```

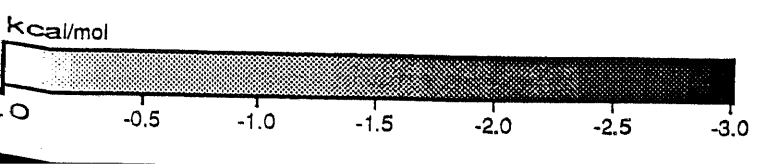


ROP

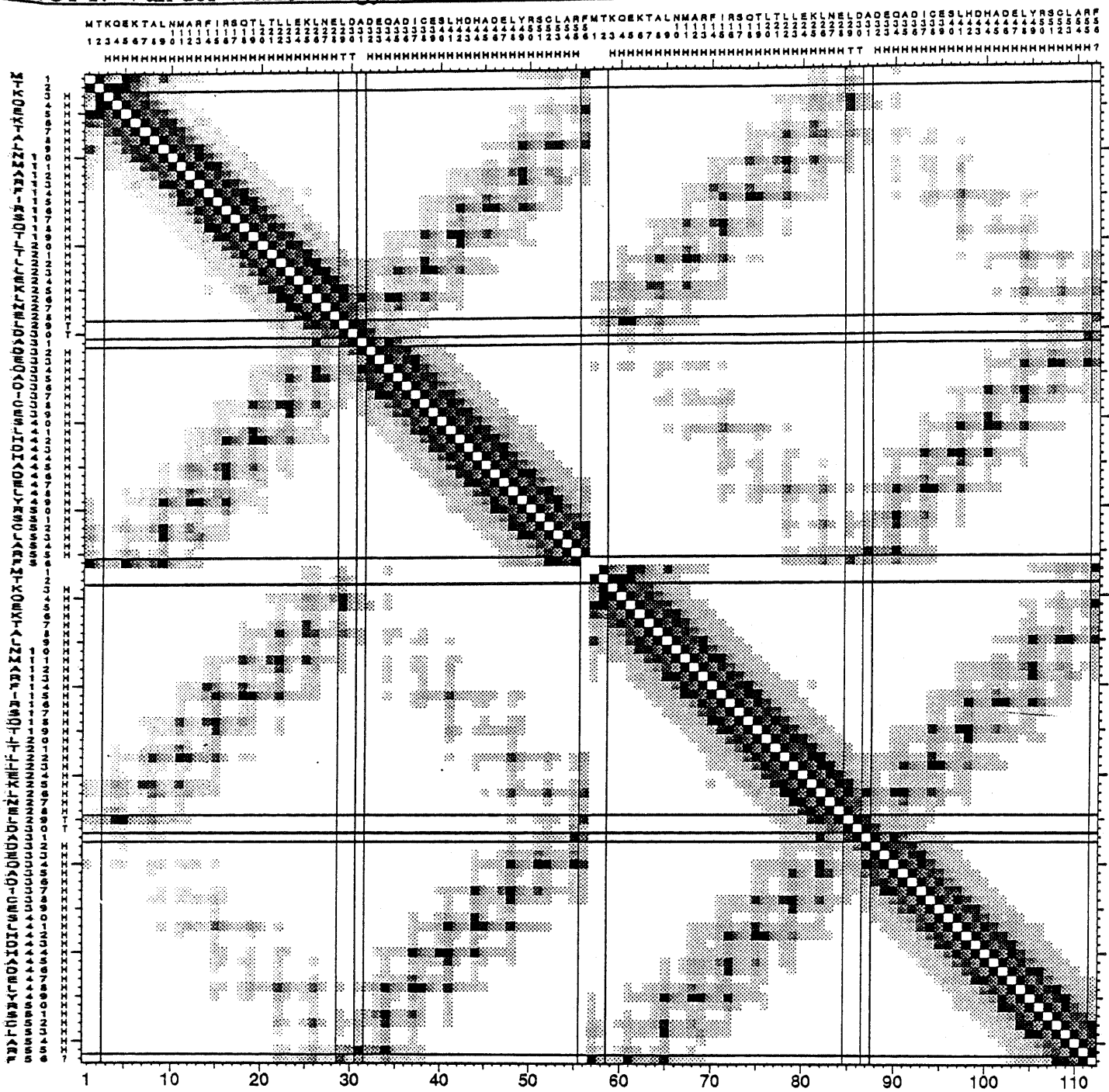


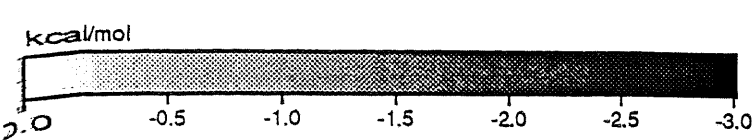
Bundle


1
10
20
30
40
50
60
70
80
90
100
110


kcal/mol
0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0

ROP1: Van der Waals energy.





Bill DeGrado
Tim Hubbard
Joachim Reichelt
Clare Woodward

* INSERTION OF A COPPER-BINDING SITE INTO THE STRUCTURE OF ROP *

(RCU1, RCU2)

This project investigates whether a metal-binding site such as that found in plastocyanin and azurin (blue copper proteins) can be accomodated within a four-helix bundle. These proteins use four ligands to bind copper in a distorted tetrahedral geometry. These are two histidines (delta-N to Cu distance 2.1 Å), one cysteine (S to Cu distance 2.1 Å), and one methionine (S to Cu distance 3.0 Å).

We hoped to mimic the ligand environment of the blue copper proteins as closely as possible. An examination of the structure of ROP (solved by David Banner and Michael Kokkinidis; EMBL) suggested that this might be accomplished by changing Phe(56B) with a His residue, Ala(31A) with a Cys, Leu(29A) with a His, and Leu(26A) with a Met. The rationale for choosing these changes were 1) these positions are near the top of the bundle; changes at such positions should leave the well packed hydrophobic core invariant; 2) the total volume of the four sidechains in the native and altered structures should be the same (273 vs. 274 $Å^3$). 3) it should be possible to form a geometrically reasonable copper-binding site by applying acceptable chi angles to the changed side chains, and only minor modifications to the phi and psi angles of the main chain.

Also important in the choice of the positions to change was the fact that Phe(56B) has been mutated to Gln and the resulting protein appeared to fold, albeit with a lowered temperature stability results of Gianni Cesarini et al.; EMBL). This suggested that a His might be located there. This residue formed a pivotal point about which the rest of the mutations were determined. The Ala(31A) to Cys mutation fills in some of the space which was vacated when Phe was converted to His. The Leu(29A) which was changed to a His is located in the connecting loop and points in an appropriate direction. The

final methionine ligand could be located at Leu(26A). In addition to its interaction with the metal it was well positioned to undergo apolar interactions with the hydrophobic side chains in the protein's apolar interior.

A major clash between His(56B) and Cys(52B) was observed and remedied by exchanging the Cys(52B) for Ala. This opened a small hole which was large enough to allow entry of a water molecule and hydrogen bond to the epsilon nitrogen of His(56B). Leu(9B) was also changed to Ile to improve its vdW interaction with His(56B).

Result: RCU1.BRK

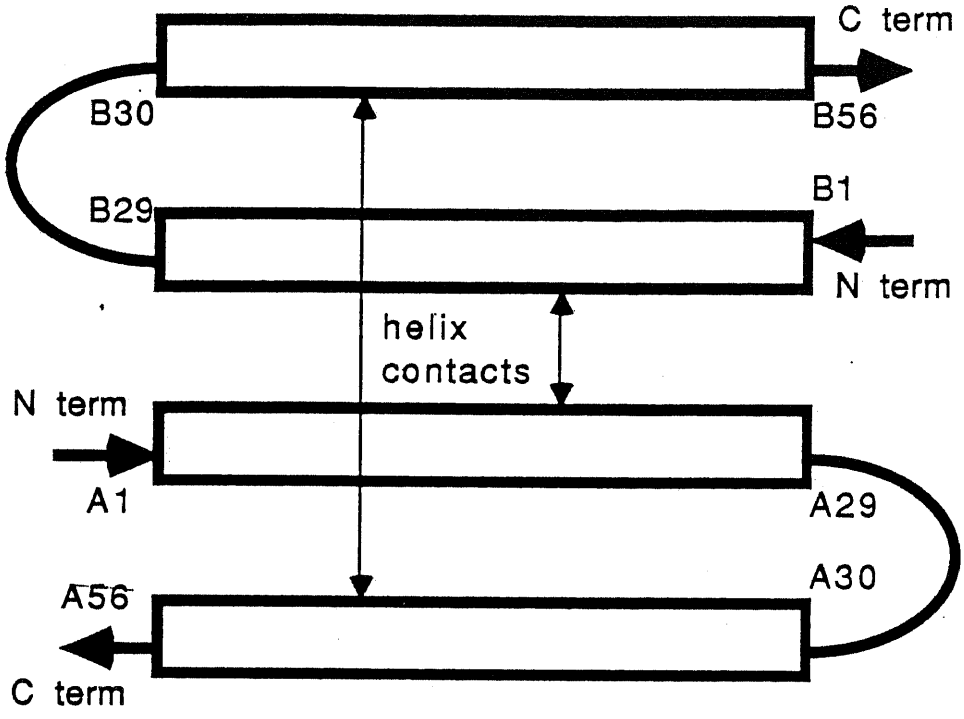
The geometry of this model was optimized using AMBER. The geometry of the Cu ligands was constrained to the geometry observed in plastocyanin by applying distance and angle constraints (force constant for the ligand distance was taken to be half that for a S-S bond). The structure was refined using 100 cycles of AMBER. No bad contacts were detected, and the final energy was reasonable. Further, analysis of the energies of the groups in contact with the copper failed to show any major bad contacts. Result: not kept

RCU1 is a dimer like ROP with symmetrical ends so the binding site could be modelled into each end. We changed this by breaking the B chain between residues 29 and 30 and designing two new loops to link residues A1 to B30 and residues B56 to B1 These loops are designed to conform with the pattern found in the single-chain four-helix bundles such as hemerythryn, myohemerythryn, and cytochrome c'. Two loops were placed near the binding site in an effort to stabilize the geometry of the ligands. An initial attempt was made to model these loops using the loop generating option in Frodo. However, no loop was found with a fit better than 1.5 Å rms. Therefore the best of these loops were used as a rough templates, and loops which roughly followed the chain of the template were fitted in by hand. Residues which would either stabilize loop formation (e.g., Pro), or which would stabilize the geometry of the ligand-binding sites, were finally added. After minimization of the resulting model, some further changes in the side chains of existing residues were made to help stabilize the binding site in its desired conformation.

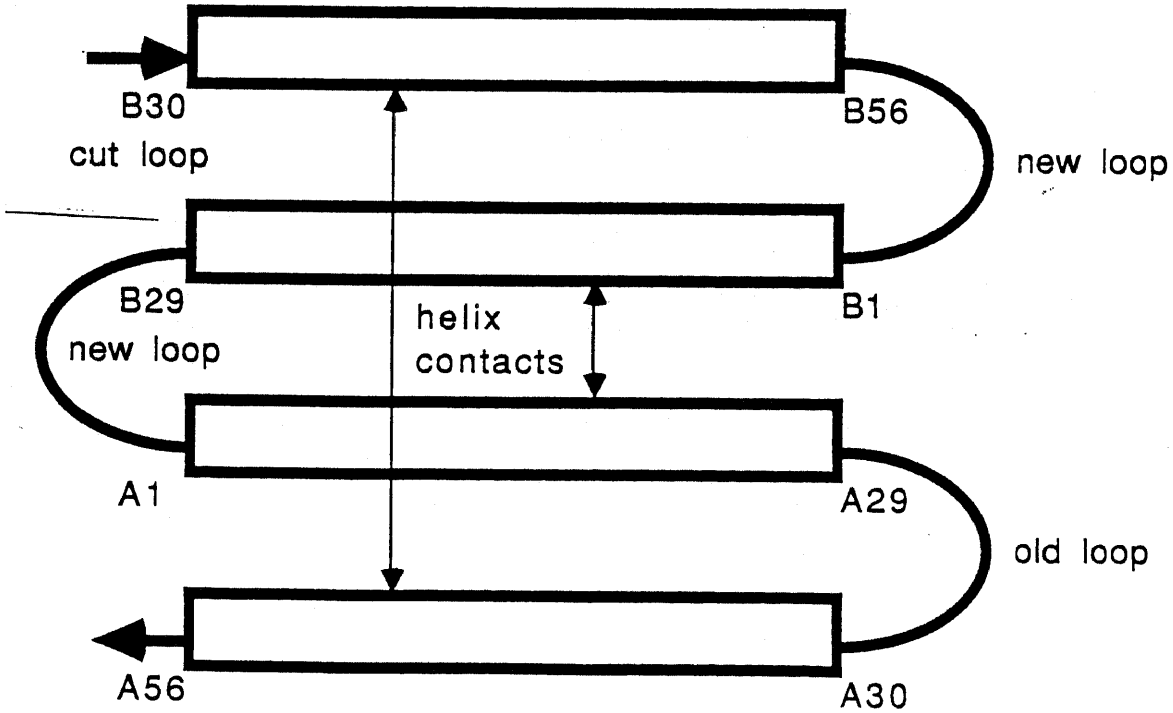
Result: RCU2.BRK

The following page shows diagramatically the changes we have made in RCU2 relative to ROP.

Native ROP, a two-chain double-two-helix bundle



CopRop (RCU2), the new one-chain four-helix bundle



Summary of Changes from Original ROP Dimer (ROP) to Redesigned ROP Dimer (RCU1)

Total number of sequence changes: 2*6 / 112 (only one chain shown)
blank = chain continues

```
      1.........................29  30.......................56
rop   MTKQEKTALNMARFIRSQTLTLLEKLNEL  DADEQADICESLHDHADELYRSCLARF GDDGENL
               *                  * *   *                    *   *
rcu1  MTKQEKTAINMARFIRSQTLTLLEKMNEH  DCDEQADICESLHDHADELYRSALARH GDDGENL
```

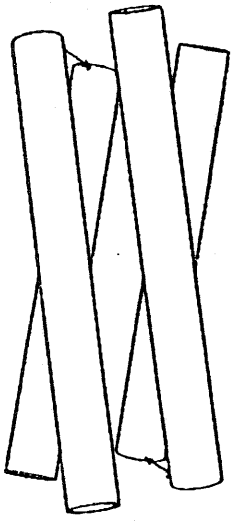
Summary of Changes from Original ROP Dimer (ROP) to Redesigned ROP Monomer (RCU2)

(Chris Sander, Dec. 1, 1986)

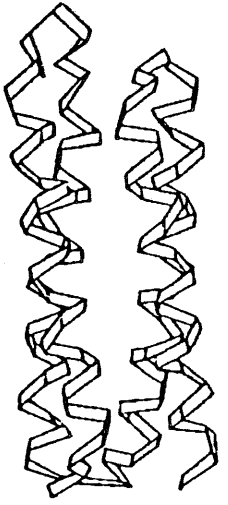
Total number of sequence changes: 16 / 112
blank = chain continues
/ = fragment end
i = inserted
* = sequence change
: = points to a residue

```
rop   MTKQEKTALNMARFIRSQTLTLLEKLNEL  DADEQADICESLHDHADELYRSCLARF GDDGENL
      :                                                        :
      A1.......................A29  A30.....................A56

rop   MTKQEKTALNMARFIRSQTLTLLEKLNEL  DADEQADICESLHDHADELYRSCLARF GDDGENL
      :                                                  :
      B1.......................B29  B30..................... B56


rop   DADEQADICESLHDHADELYRSCLARF / MTKQEKTALNMARFIRSQTLTLLEKLNEL /
rcu2  DADEQADICESLHDHADELYRSALARH   PGTAQKTAINMARFIRSQTLTLLEKLNEL
      :                        *   *   *****   *                  :
      B30.....................B56   B1.......................B29   old name
      1.....................................................56   new numbers

rop   MTKQEKTALNMARFIRSQTLTLLEKLNEL  DADEQADICESLHDHADELYRSCLARF
rcu2 GQAPKQIKTALNMARFIRSQTLTLLEKMNEH  DCDEQADICESLHDHADELYRSCLARF
    ii**  *                        *  *    *                     :
      :                                                          :
      A1.....................................................A56   old name
     57.....................................................114   new numbers
```

```
8)  RopCop2.....RCU2
      SHEET....
      BRIDGE2..
      BRIDGE1..
      CHIRALITY   +++++++++ +++++++++ +++++++++ +-++++++++ +++++++++
      BEND.....     SSSSSSS SSSSSSSSS SSSSSS SSS   SSSSSSSS SSSSSSSSS
      5-TURN...
      4-TURN...   >>>>XXXXX XXXXXXXXXX XXX<<<<     >>>>XXXXX XXXXXXXXXX
      3-TURN...   >>3<<    > 33<>33<>33 <     >>><< <>33<              >
      SUMMARY..    HHHHHHHH HHHHHHHHHH HHHHHHGGGG   HHHHHHHH HHHHHHHHHH
      EXPOSURE.  *09*01*40* 602*303*13 *2098*0*84 4**902630* 9075408702
    1 SEQUENCE.  DADEQADICE SLHDHADELY RSALARHPGT AQKTAINMAR FIRSQTLTLL


      SHEET....
      BRIDGE2..
      BRIDGE1..
      CHIRALITY  +++++++++-+ +++++++++ +++++++++ ++++++-+-+ +++++++++
      BEND.....  SSSSS SS S SSSSSSSSS SSSSSSSSS SSSSSSS  S SSSSSSSSS
      5-TURN...                                    >5555<
      4-TURN...  X<<<<    >4 >>X>XXXXXX XXXXXXXXXX XX<X<<4<>> >>XXXXXXXX
      3-TURN...  33<         >33<                 >33<>33<>> 3<<    >33<
      SUMMARY..  HHHHS SS T THHHHHHHHHH HHHHHHHHHH HHHHHTTT H HHHHHHHHHH
      EXPOSURE.  *607*16*2* *74*703730 9907540870 2*717*4*09 *01*40*602
     51 SEQUENCE.  EKLNELGQAP KQIKTALNMA RFIRSQTLTL LEKMNEHDCD EQADICESLH

      SHEET....
      BRIDGE2..
      BRIDGE1..
      CHIRALITY  ++++++++++ ++
      BEND.....  SSSSSSSSSS SS
      5-TURN...
      4-TURN...  XXXXXXXX<X <<4<
      3-TURN...  >33<>33<>> 3<<
      SUMMARY..  HHHHHHHHHH HTT
      EXPOSURE.  *303*03*20 89*7
    101 SEQUENCE.  DHADELYRSC LARF

Notation:  SUMMARY H=ALPHA-HELIX
                   E=BETA-STRAND
                   B=BETA-BRIDGE G=3-HELIX I=5-HELIX
                   T=3-,4-, OR 5-TURN
                   S=BEND
```

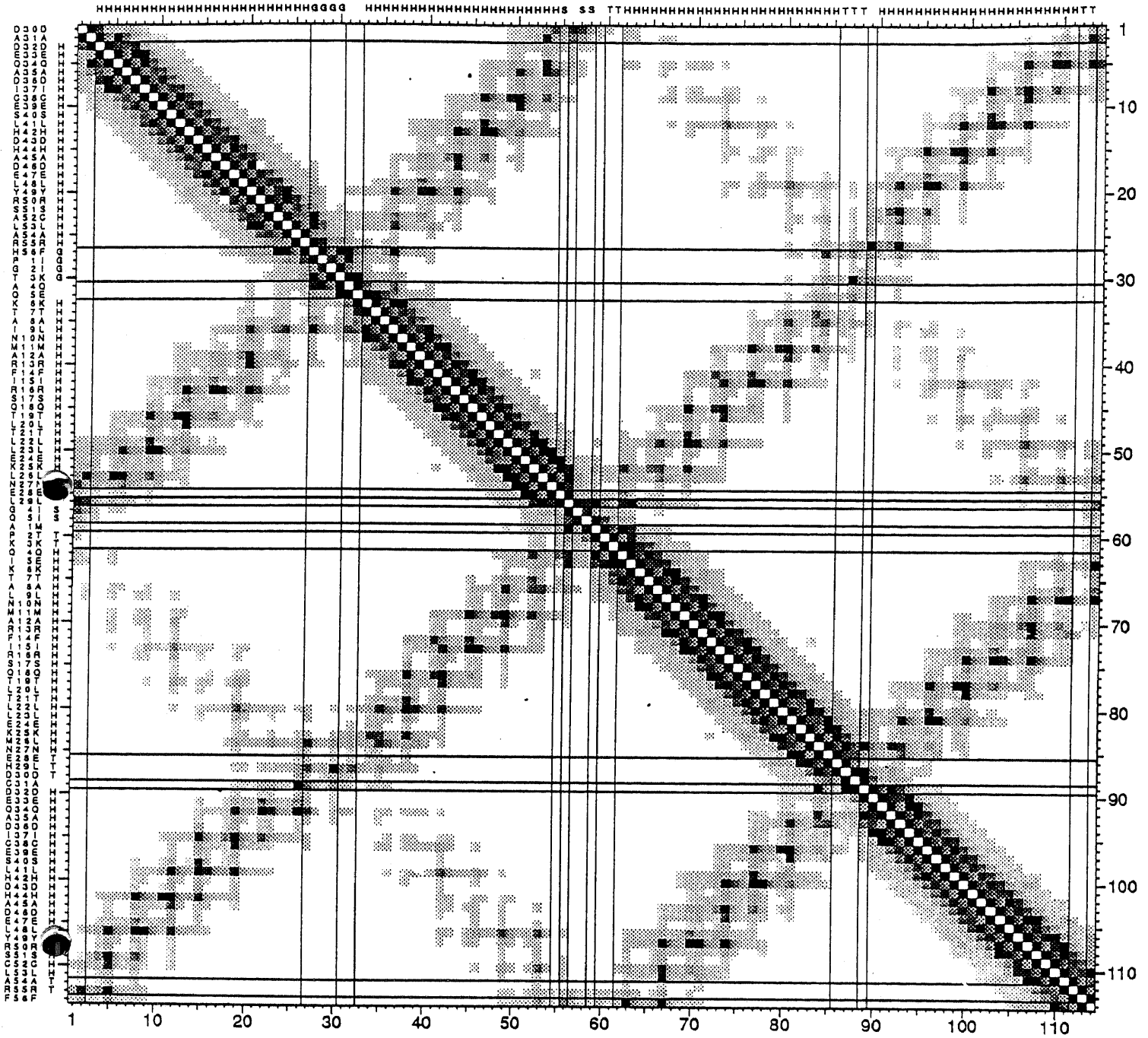


ROP



RCU2 MODEL

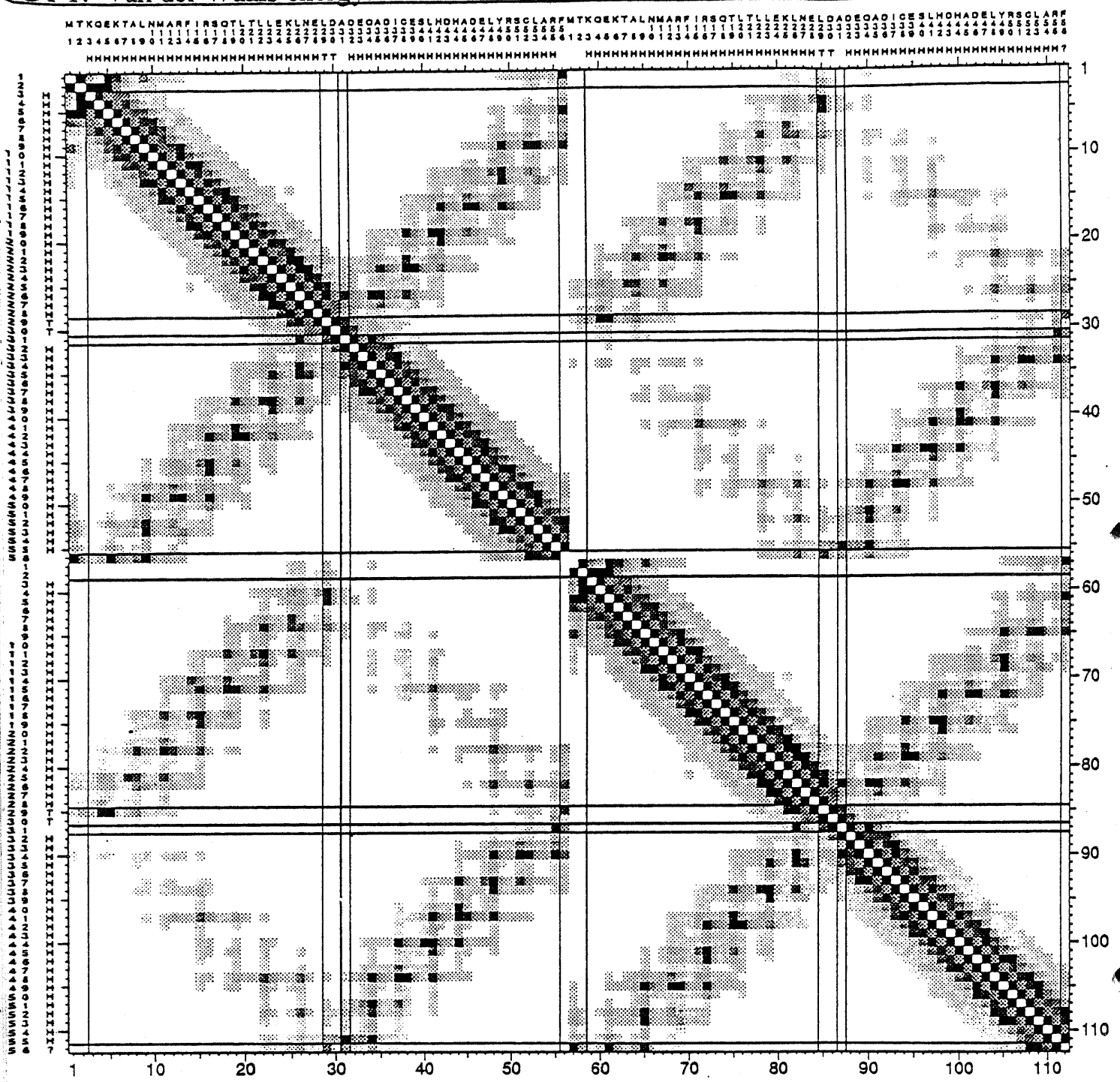
RCU2: Van der Waals energy.



kcal/mol

0 -0.5 -1.0 -1.5 -2.0 -2.5 -3.0

OP1: Van der Waals energy.



cal/mol

-0.5 -1.0 -1.5 -2.0 -2.5 -3.0

APPENDIX

POLARITY CRITERIA FOR DESIGN EVALUATION

by
C. Frömmel and G. Baumann

The accessible total and polar surface area of globular proteins - a useful criterion in protein design

C. Froemmel* and G. Baumann+

* Institut of Biochemistry, Humboldt-University, Berlin, GDR

+ Central Institute of Molecular Biology, Acadamy of Science GDR

Abstract: Today many tools for protein tertiary structure prediction are available: protein graphics including constructing tools, homology search, energy minimization, molecular dynamics etc.. The use of these powerful methods can result in a large number of designed structures which has to be evaluated for reliability. To do this there are several method (e.g. packing analysis, torsion angle distribution, calculation of polar interactions, estimation of the exposed surface area etc.). Using the concept of polar and apolar surfaces areas (Froemmel, C., 1984), the classification in protein-like and non-protein-like spatial structures by accessible surface area calculations is improved. There is a strong functional relationship between molecular weight and accessible polar and apolar surface area, respectively. According to this criterion two of six predicted structures constructed during the EMBO practical course 'Protein Tinkering' exhibit design flaws. Also the benchmarks of misfolded proteins designed by NOVOTNY and coworker show significant deviation of their polar and apolar accessible surface area from comparable proteins, whereby the analysis of the total surface area shows no significant deviations.

1. Introduction

The concept of the apolar bond in proteins was introduced more as a qualitative than a quantitative model. However, the most important practical and theoretical questions of protein stability are concerned with the magnitude of thermodynamic data associated with apolar (or hydrophobic) bonding. Attempts at a quantitative assignment was made by partitioning experiments (e.g. Nozaki, Y. and Tanford, C., 1971). Later linear correlations between the surface area of molecules and free energy of transfer from water to organic solvent were used to calculate 'hydrophobic energy' of proteins and its contribution to protein stability (see Chothia, C., 1975, 1976; Richards, F.M., 1977; Gelles, J. and Klapper, M.H., 1978, Eisenberg, D.A. , and McLachlan, A.D., 1986). The gain of about 25 cal/mol hydrophobic free energy by each square Angstrom removed from contact with water has been widely used to estimate the free energy of interacting molecules or parts of them. For calculating the hydrophobic contribution of any kind of residue it was assumed that as long as uncharged polar groups retain their hydrogen bonds inside the protein their hydrophobicity as pair is similar to that of non-polar groups with comparable change of water- accessible surface area. Thus the term 'hydrophopbic' was used to describe surfaces that are formed by non-polar atoms and /or polar atoms which show intramolecular hydrogen bonds. In a more detailed analysis of the accessible surfaces of proteins (Chothia, C., 1976, recently Miller, S. et al., 1987) polar and apolar atoms

were classified in a qualitative manner: the nitrogen, oxygen and sulphur atoms are classed as polar, all carbon atoms are considered as apolar. To gain a more quantitative description of this problem it has to be included polarity measures of each atom type. We used the partial atom charge (Froemmel, C., 1984), EISENBERG and McLACHLAN (Eisenberg, D.A. , and McLachlan, A.D., 1986) have applied the atomic contribution to transfer energy (transfer octanol ---> water) as polarity scale (Eisenberg, D. and McLachlan, A.D., 1986). Another useful measure can be the solvation energy of different atom types in water (Ooi, T. et al.. 1987).

In this report the characteristics of accessible polar and apolar area of proteins are analysed. We estimate standard values over a broad range of molecular weight of buried polar and apolar surface area, respetcively, for globular proteins. It is shown that the derived relationship between molecular weight and surface areas of different types is useful for the evaluation of designed protein structures.

2. Methods

The accessible surface area defined according to RICHARDS, F.M., 1977 was calculated by the method proposed by SHRAKE, A. and RUPLEY, J.A. (1973). Hydrogen atoms are included in the heavy atoms using larger values of atom radii. The surface area of sphere representing an atom is described by a set of 320 test points that are uniformely distributed. Their distance to the center is the sum of the atom radius (see Tab. 1) and the assumed water radius (1.4 Angstrom). The program used is an extended version of DSSP (Kabsch, W. and Sander, C., 1983; extended by C.F. 1986). As a measure of the polarity of single constituents of a molecule we chose the absolute value of the partial charge. The partial charge of united atoms was calculated as the mean of atomic partial charges of the heavy atom and the connetced hydrogens, where the partial charges are weighted by the accessibility of each of the constituents (eq. 7 in Froemmel, C., 1984). The atomic partial charges were taken from MOMANY, F.A. et al., 1975. The polar surface area A(pol) of an atom is defined as the product of partial charge and accessible surface area (eq. 1). The apolar surface area A(apol) is the difference of the total accessible surface area minus the the weighted polar surface area (eq. 2).

$$A(pol) = |n| * A(tot) \qquad \text{eq. 1}$$

$$A(apol) = A(tot) - w * A(pol) \qquad \text{eq. 2}$$

where n is the atomic partial charge; w is the weighting factor, estimated from least square analysis of the relationship between apolar surface and apolar surface (Froemmel, C., 1984). With w = 3.4 a linear function between apolar surface A(apol) and transfer energy of amino acid residues (transfer between water and organic solvent) is observed. The polar accessible surface area of amino acids is directly related to the solvation energy of amino acid side chains in water (Tab. 2,3; see also Froemmel, C., 1984). As standard state of the amino acid residue X we used the extended structure of the pentapeptide Gly-Gly-X-Gly-Gly.

3. Results

According to the results of MILLER, CHOTHIA, JANIN, and TELLER, respectively, (Miller, S. et al. 1987; Chothia, C., 1975; Janin, C., 1976, Teller, D.C., 1976) there should be a nonlinear relationship between the total buried surface and molecular weight: the total surface of the unfolded protein (standard state) increases linearely with molecular weight (see also Fig. 1A), the increase of the accessible surface area could be fitted with a function owing the 2/3 power up to 3/4 power of molecular weight. Using an identic data set of protein structures we can reproduce their results but we have checked also a simple linear relation between molecular weight and accessible surface area. The resulting correlation does not differ significantly in the r values. Consequently there is no statistical reason to prefer one of the both models. The linear function does not cross the origin like the logaritmic function does. The excess value of accessibilty at low molecular weight indicated by the linear model is based on the fact that smaller peptides are not able to bury large surface area by itself. Including an offset in the logarithmic model (in the linear relation about 1000) and fitting the accessible surface area in dependence on the molecular weight according eq. 3

$$A(acc) = A(0) + MG^{**}x \quad \text{eq.3}$$

we get for A(0)= 1300 and x= 0.95 the largest r value. But with the data base given in the Brookhaven protein data bank a discrimination between the three models (linear, logarithmic, extended logaritmic) is not possible. Due to simplicity we prefer the linear model. The x value close to one in the extended logaritmic model and the validity of the linear model acrross the whole range of protein molecular weight including monomeric and polymeric proteins support our model choice. The linear relationship is probably caused by the form and roughness of protein structure including subunit structure, domains, secondary structure elements and at least atoms (Gates, R.E., 1979).

All types of surface area are linearely related to the molecular weight both types it is (Fig. 1 A,B,C). In proteins owing a molecular weight > 8 000 the extent of accessible area is about 35 % for all types considered. (Tab. 4). Due to the assumption of linear relationship between molecular weight and different types of surfaces we get a simple evaluation method for protein structure (The constant term becomes negligible at higher molecular weight, Tab. 3). According to the strong dependence of total and polar surface area on molecular weight we can check the amino acid content of a sequence if it does fullfill the general requirements on surface area and their polarity. (Exclusion e.g. of poly alanin as primary structure for folded protein). If the sequence behaves like one of a folded proten we can evaluate easely the folded structure by the relation (%polar accessible surface area/%apolar accessible surface area surface area. The relations between percentage of polar and apolar surface area (accessible or buried), respectively, are scattered arround 1.0 (Fig. 2). For the ratio of polar/apolar of accessible surface (= evaluation of the external surface) the lowest value of an enzyme is observed for staphylococal nuclease (2SNS, 0.750. Low molecular weight peptides have lower ratios (mellitin, 1MLT; and crambin, 1CRN 0.69, not included in the graph).

One of the EMBO-course designed proteins shows a borderline value equal to that of the staphylococcal nuclease. It is designed as a 4-helix bundle. All proteins given in the protein data bank of this folding type show ratios significant larger than 1.0. So we would guess, that this protein may be worse designed. The proteins contructed by NOVOTNY and Bruccoleri are recognized by the evalution of external surface without doubt as misfolded one. The evaluation of the buried surface area leads to similar results. Two of 6 designed proteins (EMBO-course) are incorrect folded as the given sequence it demands. Also the misfolded and energy minimized protein structures derived by NOVOTNY et. al. show a ratio %polar buried surface area to %apolar buried surface area higher than 1.2 (Fig. 2). No known protein structure has such high ratio.

4. Discussion and conclusions

Using modern computer tools it is possible to get a large number of protein models for one sequence. Only few (or no) of them are probably folded in a correct fashion. One evaluation of the model structure can be done by estimating the accessible surface area. In a very rough manner this method considers the water protein interaction at the surface of proteins. Because the interaction energy between water and any protein constituent depends on the polarity of the latter this approach is improved considering seperately accessible polar and apolar surface area. Recently several authors (Eisenberg, D. and McLachlan, A.D., 1986, Ooi, T. et al., 1987) have proposed a direct method to assess this energy using atom contribution to hydrophobic and hydrophilic energy, respectively. Also these approaches can be used in evaluation of protein structure. Due to the known correlation between polar and apolatr accessible surface area and transfer energy and solvation energy (Froemmel, C. 1984; Tab. 3) the basic principle of our approach is comparable to Eisenberg, MacLachlan and Ooi, et al. Another possibility is to count the apolar contacts in proteins (Bryant, S.H. and Amzel, L.M., 1987) but it does not yield to such linear expressions given in this paper. Using the procedure described above we get a simple evaluation method over a very broad range of molecular weight without necessity to distinct monomeric and polymeric proteins.

Tab. 1 Radii in Angstrom used for surface area calculation

N	1.65
C	1.80
O	1.55
S	1.90
water	1.40

Tab. 2

Compilation of the reference values of the accessible surface area and derived values of amino acids in an extended structure owing Gly-Gly neighbours in the sequence

used in evaluation the change of area upon folding. The accessible surface was calculated using an extended version (C.F.) of DSSP (Kabsch, W. and Sander, C. , 1983). The polar and apolar accessible surface was calculated according eq. 1,2. In the main chain the apolar part is about zero.

	main chain		side chain			residue			CHOTHIA
	total	polar	total	polar	apolar	total	polar	apolar	total(a)
	square Angstrom								
A	43.6	13.9	62.7	2.6	53.8	106.3	16.5	50.1	115
R	40.1	13.5	199.3	45.5	44.5	239.5	59.0	38.5	225
N	40.5	13.8	109.2	28.9	11.1	149.8	42.7	4.7	160
D	41.1	13.8	108.4	30.4	4.9	149.5	44.3	-1.0	150
C	41.1	13.8	97.4	1.9	91.0	138.5	15.7	85.8	135
E	42.0	13.8	140.8	37.5	13.2	182.8	51.3	18.4	190
Q	43.4	14.1	143.1	31.0	37.6	186.6	45.2	33.0	180
G	83.6	19.6	0.0	0.0	0.0	83.6	19.6	16.8	75
H	41.1	13.8	141.1	13.8	94.1	182.2	27.6	87.7	195
I	39.7	13.4	131.8	3.8	119.0	171.5	17.2	113.0	175
L	40.2	12.9	123.4	3.7	110.7	163.6	16.7	106.9	170
K	40.4	13.5	160.5	15.0	109.4	200.8	28.5	104.0	200
M	42.1	13.9	151.5	7.6	125.5	193.7	21.5	120.4	185
F	40.5	13.5	160.4	2.2	152.8	200.9	15.7	147.4	210
P	39.6	11.4	96.3	1.8	90.2	135.9	13.2	90.9	145
S	42.6	14.0	80.7	11.3	42.3	123.1	25.2	37.4	115
T	41.3	13.8	100.4	9.7	67.4	141.7	23.5	61.7	140
W	40.5	13.5	204.6	10.7	168.2	245.4	24.7	162.8	255
Y	40.5	13.5	171.9	14.7	122.1	212.4	28.2	116.7	230
V	39.5	13.5	108.9	3.2	98.0	148.4	16.7	91.8	155
Z	40.7	13.5	42.0	0.9	39.0	82.7	14.4	33.6	135

Z = CYS in disulfide bridges (1/2 cystine)
a Chothia, C., 1976.
The amino acids are given in the one-letter code.

Tab. 3 Some linear relationships and their parameters connected with protein surface areas.

y	x	y= a*x + b	a	b	r	
amino acids						
1 solvation energy	polar surface		-0.413	6.262	0.93	a
2 transfer energy	apolar surface		-0.0176	0.344	0.90	b
proteins						
maximum, observed in hypothetical extended structure						
3 total maximal surface	molecular weight		1.434	-290	1.00	c
4 total polar surface	molecular weight		0.240	63	1.00	c
5 total apolar surface	molecular weight		0.621	-490	1.00	c
surface areas of proteins buried during folding						
6 total surface	molecular weight		1.031	-1470	1.00	c
7 polar surface	molecular weight		0.174	-196	1.00	c

8 apolar surface	molecular weight	0.444	-788	0.99	c

surface areas in square Angstrom, energy data in kcal/mol.
a solvation energy values of amino acid side chains (transfer vacuo --> water (Wolfenden, R.V. et al., 1981)
b compilation of transfer energy (transfer water --> organic solvent) in Froemmel, C., 1984.
c linear regression coefients corresponding to straight lines in Fig. 1

Tab. 4 The extent of accessibility of different types of protein surface areas (including small proteins) (reference structure 100% extended chain).

type	mean	standard deviation
total	35.1	6.3
apolar	35.6	9.7
polar	34.5	6.3

The tendency is decreasing until from low to molecular weight of 10 000. After this the accesible part remains quite constant.

Fig. 1 The linear relationship between the molecular weight and several types of surface area (in square Angstrom).
A: The increase of total and apolar surface area with molecular weight.
B: The increase of buried total and apolar surface area with molecular weight.
C: The increase of buried polar surface area with molecular weight.

Fig. 2 Quality control of designed proteins. Proteins designed during the practical EMBO course 'Protein Tinkering' 1986 in Heidelberg, FRG, and (1) misfolded proteins (NOVOTNY). Plotted is the ratio ratio %polar buried surface area to %apolar buried surface area ('internal surface'; top) and the ratio ratio %polar accessible surface area to %apolar accessible surface area ('external surface'; bottom). All proteins show optimized nonbonded interactions and similar fraction of buried total surface area.

References

Bryant, S.H. and Amzel, L.M. (1987) Correctly folded proteins make twice as many hydrophobic contacts. Int. J. Peptide Protein Res. 29: 46 - 52.

Chothia, C. (1975) Structural invariants in protein folding. Nature 254: 304 - 308.

Chothia, C. (1976) The nature of the accessible and buried surfaces in proteins. J. Mol. Biol. 105: 1 - 14.

Eisenberg,D. and McLachlan, A.D. (1986) Solvation energy in protein folding and binding. Nature 319: 199 - 203.

Froemmel, C. (1984) The apolar surface area of amino acids and its empirical correlation with hydrophobic energy. J. theor. Biol. (1984) 111: 247 - 260.

Gates, R.E. (1979) Shape and accessible surface area of globular proteins. J. Mol. Biol. 127: 345 - 351.

Gelles, J. and Klapper, M.H. (1978) Pseudo-dynamic contact surface areas:estimation of apolar bonding. Biochem. Biophys. Acta 533: 465 - 477.

Janin, J. (1976) Surface area of globular proteins. J. Mol. Biol. 105: 15 - 37.

Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structur: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577 - 2637.

Miller, S., Janin, J., Lesk, A.M. and Chothia, C. (1987) Interior and surface of monomeric proteins. J. Mol. Biol. 196: 641-656.

Nozaki, Y. and Tanford, C. (1971) The solubility of amino acids and glycin peptides in aques ethanol and dioxan solutions. J. Biol. Chem. 246, 2211 - 2217.

Ooi, T., Oobatake, M., Nemethy, and Scheraga, H.A. (1987) Accessible sutface areas as a measure of the thermodynamic parameters of hydration of peptides. Proc. Natl. Acad. Sci. USA 84: 3086 - 3090.

Richards, F.M. (1977) Areas, volumes, packing, and protein structure. Ann. Rev. Biophys. Bioeng. 6: 151 - 176.

Shrake, A. and Rupley, J.A. (1973) Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J. Mol. Biol. 79: 351 - 371.

Teller, D.C. (1976) Accessible area, packing volumes and interaction surfaces of globular proteins. Nature 260: 729 - 731.

Wolfenden, R.V., Anderson, L., Cullis, P.N. and Southgate, C.B. (1981) Affinities of amino acids side chains for solvent water. Biochemistry 20: 849 - 855.

Data base:

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977) J. Mol. Biol. 112: 535 - 542.

List of protein structures used (and their Brookhaven file names):

155C, 156B, 1ABP, 1APR, 1AZU, 1BP2, 1CAC, 1CPV, 1ECD, 1EST, 1FAB, 1FDX, 1GPD, 1HIP, 1INS, 1LH1, 1LHB, 1LZM, 1MBN, 1MBS, 1OVO, 1P2P, 1PCY, 1PTN, 1PYP, 1REI, 1RNS, 1SBT, 1TIM, 2ACT, 2ADK, 2ALP, 2APE, 2APP, 2B5C, 2CAB, 2GCH, 2GRS, 2PAB, 2PGK, 2RHE, 2SBV, 2SGA, 2SNS, 2SOD, 2SSI, 35IC, 3BP2, 3C2C, 3CNA, 3CYT, 3FXC, 3FXN, 3PGK, 3PGM, 3TLN, 4ADH, 4ATC, 4DFR, 4LDH, 4PTI, 5CPA, 7LYZ, 8PAP

(Bernstein, F.C. et al., 1977)

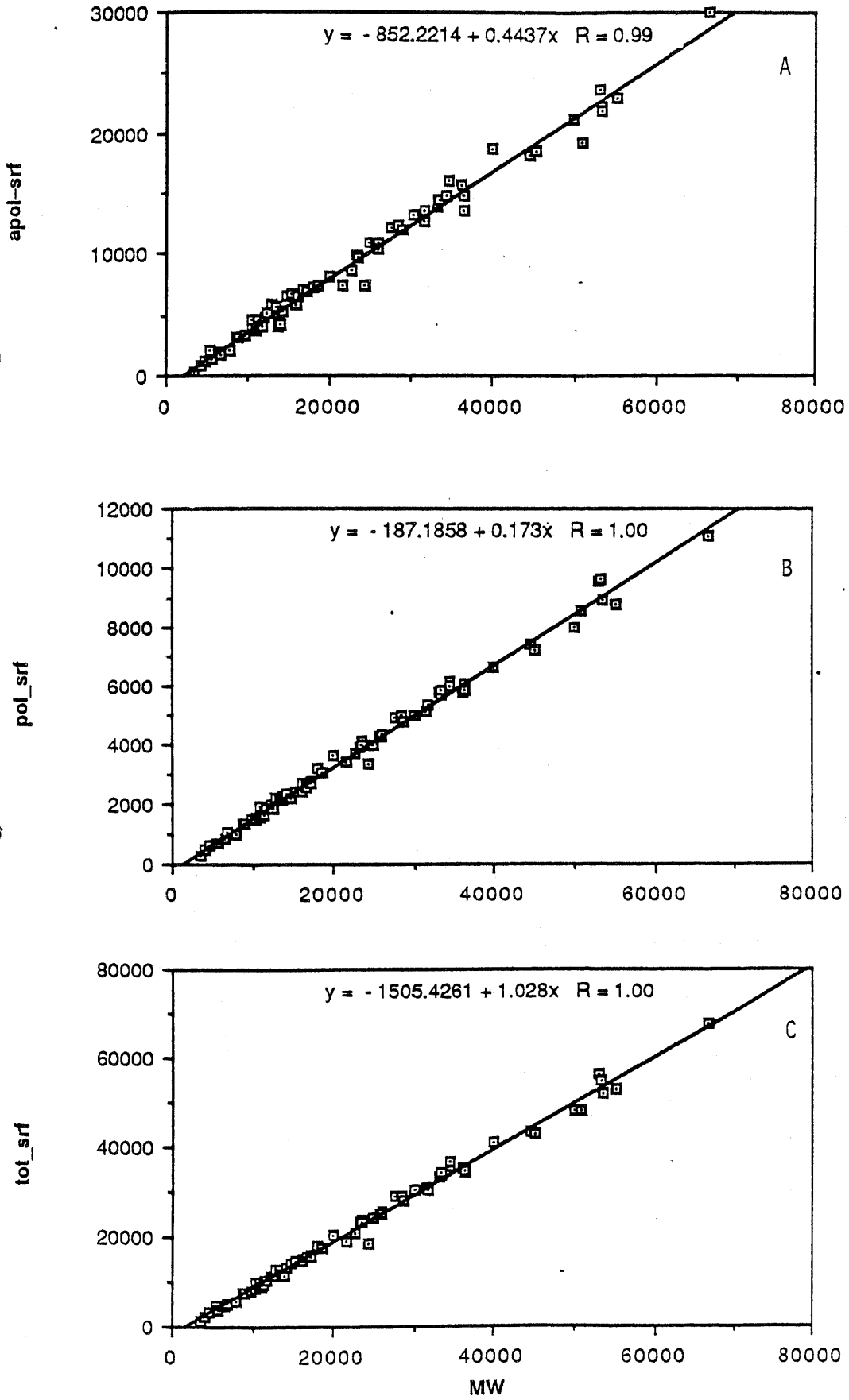
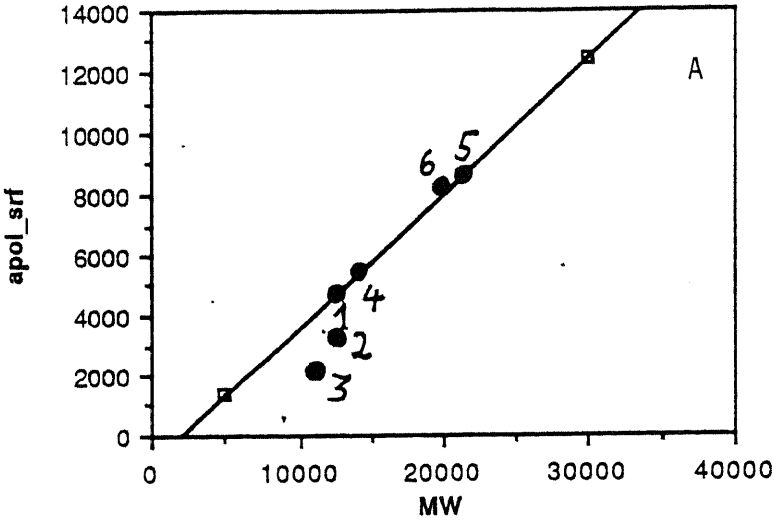
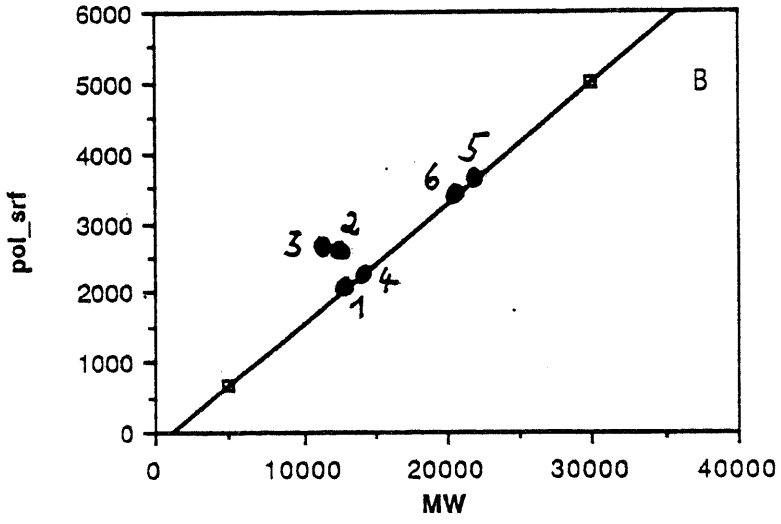

y = - 852.2214 + 0.4437x R = 0.99
A
apol-srf
30000
20000
10000
0
0
20000
40000
60000
80000
y = - 187.1858 + 0.173x R = 1.00
B
pol_srf
12000
10000
8000
6000
4000
2000
0
0
20000
40000
60000
80000
y = - 1505.4261 + 1.028x R = 1.00
C
tot_srf
80000
60000
40000
20000
0
0
20000
40000
60000
80000
MW

Fig. 1


14000
12000
10000
8000
6000
4000
2000
0
apol_srf
0
10000
20000
30000
40000
MW
A
6
5
4
1
2
3


6000
5000
4000
3000
2000
1000
0
pol_srf
0
10000
20000
30000
40000
MW
B
5
6
3
2
4
1

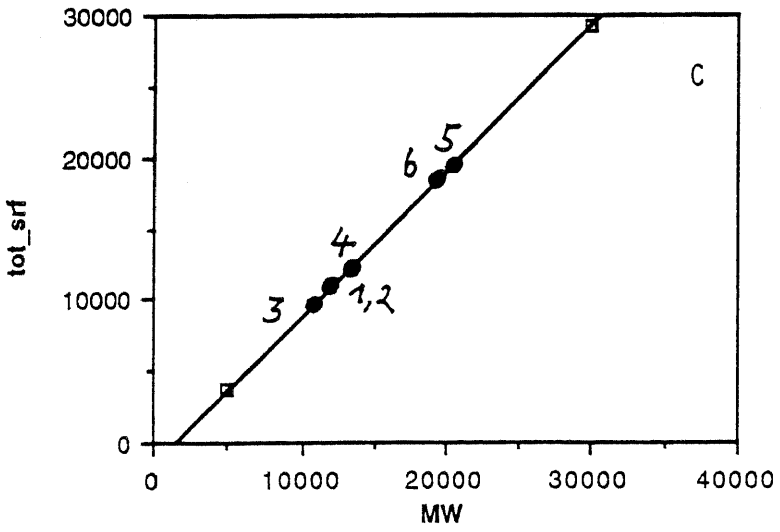

30000
20000
10000
0
tot_srf
0
10000
20000
30000
40000
MW
C
5
6
4
3
1,2

Fig. 2

Secondary structure predictions of designed proteins with three methods

Robson *by Garnier, Osguthorpe, Robson, Suzuki*
HD-Segment (Orakel) *by Kabsch and Sander*
GOR (GORIII) *by Gibrat, Garnier, Robson*

Secondary structure prediction was not used directly in the design process. Postdesign evaluation included calculation of likely secondary structure. As most groups used intuition of or tables of single residue secondary structure preferences, it is not surprising that the secondary structure predictions by various algorithms agrees quite well with the designed structure.

It may well be possible to overdesign a protein by giving it much more clearcut secondary structure preferences than natural proteins have, provided that packing and other energetic considerations are not violated.

notation:

ALPHA STRUCTURE:	H=HELIX
BETA STRUCTURE:	E=EXTENDED
LOOP STRUCTURE:	T=TURN
	C,BLANK=COIL/UNDEFINED

Reliability (GORIII) classes: 1 to 5 (5=best)

```
$PDB:TINY.DSSP_MODEL ! TINY TIM
 LENGTH        191

   1  50  ....,....1....,....2....,....3....,....4....,....5
SEQ        AGVVITVSNPAEFQKALDQALKDGARVIIQVSNPAEFQKALNQALKNGAV
STRUC       EEEE  SSHHHHHHHHHHHHHTT EEEE  SSHHHHHHHHHHHHHH  E

ROBSON     EEEEEEE   HHHHHHHHHHHHHHEEEEEE     HHHHHHHHHHT  EE
ORAKEL       EEEE    HHHHHHHHHHHH   EEEEEE   HHHHHHHHHHHH  EE
GOR        EEEEEEE   HHHHHHHHHHHHH HEEEEEE    HHHHHHHHHHHH  E
GORCLASS   12344424531434545544443311123214521445555344332313

  51 100  ....,....1....,....2....,....3....,....4....,....5
SEQ        LVIEVSNPAEFQKALNQALKDGAEVAVYVSNPAEFQKALNQALKNGAVVV
STRUC      EEE  SSHHHHHHHHHHHHHTT  EEE  SSHHHHHHHHHHHHHTT EEE

ROBSON     EEEEE  HHHHHHHHHHHHHHHHHHEEEEE  HHHHHHHHHH T  EEEE
ORAKEL     EEEEE   HHHHHHHHHHHH   EEEEEE   HHHHHHHHHHH  EEEE
GOR        EEEEE   HHHHHHHHHHHHH HHHEEEE   HHHHHHHHHHHH  EEE
GORCLASS   23211451244555534432221121311351354555534332141434

 101 150  ....,....1....,....2....,....3....,....4....,....5
SEQ        ITVSNPAEFQKALNQALKDGATVIVEVSNPAEFQKALNQALKNGAVVAIQ
STRUC        SSHHHHHHHHHHHHHHT  EEEE SSHHHHHHHHHHHHHTT EEE

ROBSON     EEE    HHHHHHHHHHH EEEEEEE  HHHHHHHHHHHHT  EEEEEE
ORAKEL     EE    HHHHHHHHHHHH    EEEEE   HHHHHHHHHHH  EEEEEE
GOR        EEE   HHHHHHHHHHHHH  EEEEEE   HHHHHHHHHHHHH HEHHEE
GORCLASS   44245314455553433214133342145214455553443313131111

 151 191  ....,....1....,....2....,....3....,....4....,....5
SEQ        VSNPAEFQKALNQALKDGATVAVYVSNPAEFQKALNQALKN
STRUC       SSHHHHHHHHHHHHHTT  B B   SHHHHHHHHHHHHH

ROBSON       HHHHHHHHHHT EEEEEEE  HHHHHHHHHHHTT
ORAKEL     E   HHHHHHHHHHH    EEEEE   HHHHHHHHHH
GOR        E   HHHHHHHHHHHH  EEEEEE   HHHHHHHHHHHH.
GORCLASS   14522445555344322313124223513545555444331

TINY ROBSON      191 75.4%
                                  P R E D I C T E D
                                   EBAPM TCLS  HGI..
                          OBS EBAPM    27     0     0     27
                          OBS TCLS     25    26     7     58
                          OBS HGI..     0    15    91    106
                                       52    41    98
                                  %    27    21    51

TINY ORAKEL      191 83.2%
                                  P R E D I C T E D
                                   EBAPM TCLS  HGI..
                          OBS EBAPM    26     1     0     27
                          OBS TCLS     20    38     0     58
                          OBS HGI..     0    11    95    106
                                       46    50    95
                                  %    24    26    50

TINY GOR         191 77.5%
                                  P R E D I C T E D
                                   EBAPM TCLS  HGI..
                          OBS EBAPM    24     0     3     27
                          OBS TCLS     21    27    10     58
                          OBS HGI..     0     9    97    106
                                       45    36   110
                                  %    24    19    58
```

```
$PDB:BABA.DSSP_MODEL ! model
 LENGTH         178

   1  50  ....,....1....,....2....,....3....,....4....,....5
SEQ        MDSGVFLQGMDAEATQAAANIAKEGLYVLIGTGKQEALQNLARYGAKGVF
STRUC         EEEEE  STHHHHHHHHHHHTT   EEES  HHHHHHHHHHT   EE

ROBSON     TT  HHHH  HHHHHHHHHHHHHHHHEEEEE    HHHHHHHHHTTT EEE
ORAKEL         EEEE   HHHHHHHHHHHH  EEEEE   HHHHHHHHHH     EE
GOR            HHHH HHHHHHHHHHHHHHHHHEEHHH  HHHHHHHHHHH    EH
GORCLASS   13331211123444554443444413232111323434334322332211

  51 100  ....,....1....,....2....,....3....,....4....,....5
SEQ        LQGMDAEATQAAANIAKEGLYVLIGTGKQEALQNLARYGAKGVFLQGMDA
STRUC      EE  STHHHHHHHHHHTT   EEES  HHHHHHHHHT   EEEE  ST

ROBSON     EE  HHHHHHHHHHHHHHHHEEEEE    HHHHHHHHHTTT EEEEE  HH
ORAKEL     EE   HHHHHHHHHHHH  EEEEE   HHHHHHHHHH     EEEE   H
GOR        EEHHHHHHHHHHHHHHHHHHEEHHH  HHHHHHHHHHH    EHEEHHHH
GORCLASS   11123444554443444413232111323434334322332211111234

 101 150  ....,....1....,....2....,....3....,....4....,....5
SEQ        EATQAAANIAKEGLYVLIGTGKQEALQNLARYGAKGVFLQGMDAEATQAA
STRUC      HHHHHHHHHHTT   EEES  HHHHHHHHHT   EEEE  STHHHHHH

ROBSON     HHHHHHHHHHHHHEEEEE    HHHHHHHHHTTT EEEEE  HHHHHHHH
ORAKEL     HHHHHHHHHH  EEEEE   HHHHHHHHHH     EEEE   HHHHHHH
GOR        HHHHHHHHHHHHHHEEHHH  HHHHHHHHHH    EHEEHHHHHHHHH
GORCLASS   44554443444413232111323434334322332211111234445544

 151 178  ....,....1....,....2....,....3....,....4....,....5
SEQ        ANIAKEGLYVLIGTGKQEALQNLARYGA
STRUC      HHHHHTT   EEES  HHHHHHHHHH

ROBSON     HHHHHHEEEEE    HHHHHHHHTTT
ORAKEL     HHHHH  EEEEE   HHHHHHHHHH
GOR        HHHHHHHEEHHH  HHHHHHHHHH H
GORCLASS   434444132321113234343343213 1
```

BABA ROBSON 178 74.2%

		P R E D I C T E D EBAPM	TCLS	HGI..	
OBS EBAPM		20	5	4	29
OBS TCLS		15	33	16	64
OBS HGI..		0	6	79	85
		35	44	99	
	%	20	25	56	

BABA ORAKEL 178 83.1%

		P R E D I C T E D EBAPM	TCLS	HGI..	
OBS EBAPM		24	5	0	29
OBS TCLS		12	44	8	64
OBS HGI..		0	5	80	85
		36	54	88	
	%	20	30	49	

BABA GOR 178 65.7%

		P R E D I C T E D EBAPM	TCLS	HGI..	
OBS EBAPM		9	1	19	29
OBS TCLS		8	24	32	64
OBS HGI..		0	1	84	85
		17	26	135	
	%	10	15	76	

```
$PDB:FXNI.DSSP_MODEL ! model
 LENGTH        121

    1  50  ....,....1....,....2....,....3....,....4....,....5
SEQ         ALVVVASGTEAADLAMSHQATRTKVNVSDVNIDELLNEDGLIMIISAMGD
STRUC        EEEE SS STTHHHHHHTT  EEEEGGG  STTTTT S EEEEEE BTT

ROBSON      EEEEEEHHHHHHHHHHHHHHHTEEEEEEEEEEHHHHHHHHHHEEEH  HH
ORAKEL       EEEEE     EEEEE       EEEE    HHHHH    HHHHHHHHHH
GOR         EEEEEH   HHHHHHHHHHH  EEEE     HHHHH H  HHEEHH
GORCLASS    13332113234445333231222231232331233112124211221354

   51 100   ....,....1....,....2....,....3....,....4....,....5
SEQ         EVLEESEFEPFSEEINTKISASVVIAYSGWGDGKWMRDFEERMNGTGIKY
STRUC       TB  TTTHHHHHHHHTTT  S EEEEE SSS SHHHHHHHHHHHHTT TT

ROBSON      HHHHHHHHHHHHHHHHHHHHEEEEEEE TT T T  HHHHHHTTTT  EE
ORAKEL      HHHH                   EEEEE      HHHHHHHHHH   EEEE
GOR             H  H HHHHHH HH  EEEEEEE       EHHHHHH        EE
GORCLASS    23211311122133321312134443223245521113332113333121

  101 121   ....,....1....,....2....,....3....,....4....,....5
SEQ         SNSMNNGRKDGKDMYEIGKKI
STRUC         EEE SS  SHHHHHHHHH

ROBSON        TTTTTTT HHHHHHHHHH
ORAKEL              EEEE
GOR                 HHHH HHH
GORCLASS    13314454455332233212 1

FXNI ROBSON       121 52.9%
                                    P R E D I C T E D
                                     EBAPM TCLS  HGI..
                          OBS EBAPM    16     4     4     24
                          OBS TCLS      9    19    31     59
                          OBS HGI..     3     6    29     38
                                       28    29    64
                                  %    23    24    53
FXNI ORAKEL       121 52.9%
                                    P R E D I C T E D
                                     EBAPM TCLS  HGI..
                          OBS EBAPM    12     4     8     24
                          OBS TCLS      6    42    11     59
                          OBS HGI..     9    19    10     38
                                       27    65    29
                                  %    22    54    24
FXNI GOR          121 64.5%
                                    P R E D I C T E D
                                     EBAPM TCLS  HGI..
                          OBS EBAPM    15     5     4     24
                          OBS TCLS      5    38    16     59
                          OBS HGI..     1    12    25     38
                                       21    55    45
                                  %    17    45    37
```

```
$PDB:FXNM.DSSP_MODEL

$PDB:FXNM.DSSP_MODEL ! model
 LENGTH          134

   1  50  ....,....1....,....2....,....3....,....4....,....5
SEQ        MKIVYWSDTEKMAELIAKGIIESGKDVNTINVSDVNIDELLNEDILILGC
STRUC       EEEE GGGHHHHHHHHHHHHTT   EEEETTT  STTTTT SEEEEE

ROBSON     EEEEEHHHHHHHHHHHHHHHEETTT  EEEEE EEEHHHHHHHHHHEEET
ORAKEL       EEEE   HHHHHHHHHHHHH       EEEE    HHHHH   EEEE
GOR        EEEEE   HHHHHHHHHH HEH       EEE      HHHH H HHEE
GORCLASS   13331131333455444222123433312223333412332211311312

  51 100  ....,....1....,....2....,....3....,....4....,....5
SEQ        SARGDEVLEESEFEPFIEEISTKISGKKVALFGNYGWGDGKWMRDFEERM
STRUC       BTTTB  TTTHHHHHHHHSTT TT EEEEEEEESSS SHHHHHHHHHH

ROBSON           HHHHHHHHHHHHHHHHHHHHTEEEEEEETTTT T T  HHHHHTT
ORAKEL           EEEE       HHHH          EEEE         HHHHHHHHHH
GOR                 HH HHHHHHHHH HHH   EEEEE          EHHHHH
GORCLASS   44555313122111233334122223134332133424552111231232

 101 134  ....,....1....,....2....,....3....,....4....,....5
SEQ        NGYGCVVVETPLIVQNEPDEAEQDCIEFGKKIAN
STRUC      HHTT EE S  EEEESS GGGHHHHHHHHHHHH

ROBSON     TTTTEEEEEEEEE    THHHHHHHHHHHH HH
ORAKEL         EEEE   EEEE         EEEE
GOR             EEE              HHHHHHH HHH
GORCLASS   44242331334111355432233334431111 32


FXNM ROBSON       134 62.7%
                                    P R E D I C T E D
                                     EBAPM TCLS  HGI..
                            OBS EBAPM    22     4     3     29
                            OBS TCLS     11    22    21     54
                            OBS HGI..     2     9    40     51
                                         35    35    64
                                    %    26    26    48
FXNM ORAKEL       134 67.2%
                                    P R E D I C T E D
                                     EBAPM TCLS  HGI..
                            OBS EBAPM    21     8     0     29
                            OBS TCLS      7    42     5     54
                            OBS HGI..     4    20    27     51
                                         32    70    32
                                    %    24    52    24
FXNM GOR          134 68.7%
                                    P R E D I C T E D
                                     EBAPM TCLS  HGI..
                            OBS EBAPM    16    11     2     29
                            OBS TCLS      2    41    11     54
                            OBS HGI..     2    14    35     51
                                         20    66    48
                                    %    15    49    36
```

```
$PDB:BEAL.DSSP_MODEL ! model
 LENGTH        101

   1  50  ....,....1....,....2....,....3....,....4....,....5
SEQ        MDAMILVISGTGNTERAADAFEEGARKTGTPPVTIIYVKNEGEDAYEKFH
STRUC          EEEB SSSHHHHHHHHHHHHHHHTT S   B  B SSSHHHHHHHH

ROBSON     HHHHEEEE        HHHHHHHHHHTTT   EEEEEEE   HHHHHHHHH
ORAKEL       EEEEEE      HHHHHHHH           EEEEEE      HHHHH
GOR        HHHHEEEE          HHHHHH  H      EEEEEE     HHHHHHH
GORCLASS   22313443255543121324443111335554144433133521344322

  51 100  ....,....1....,....2....,....3....,....4....,....5
SEQ        ARGDVTVIVVGPTANEHLEDFRKLAGNLKVTVKYEINEDKIAELGAKVVN
STRUC      HH    EEEE SS HHHHHHHHHTT S    EEGGG HHHHHHHHHHHHH

ROBSON     HHTTEEEEEE    HHHHHHHHHHHHHHHHEEEEHHHHHHHHHHHHHHHEE
ORAKEL         EEEEEE       HHHHHHHH  EEEEEEEEE HHHHHHHHHHHH
GOR        H   EEEEEE     HHHHHHHHHH   HHHHHHHH HHHHHHHHHHEE
GORCLASS   13443355534444311444545531311111111123345453132223

 101 101  ....,....1....,....2....,....3....,....4....,....5
SEQ        G
STRUC

ROBSON     E
ORAKEL
GOR
GORCLASS   4
```

BEAL ROBSON 101 71.3%

		P R E D I C T E D EBAPM	TCLS	HGI..	
OBS EBAPM		12	0	0	12
OBS TCLS		10	17	12	39
OBS HGI..		2	5	43	50
		24	22	55	
	%	24	22	54	

BEAL ORAKEL 101 67.3%

		P R E D I C T E D EBAPM	TCLS	HGI..	
OBS EBAPM		12	0	0	12
OBS TCLS		12	25	2	39
OBS HGI..		3	16	31	50
		27	41	33	
	%	27	41	33	

BEAL GOR 101 69.3%

		P R E D I C T E D EBAPM	TCLS	HGI..	
OBS EBAPM		10	0	2	12
OBS TCLS		6	24	9	39
OBS HGI..		2	12	36	50
		18	36	47	
	%	18	36	47	

```
$PDB:BUND.DSSP_MODEL ! model - helical bundle designed from scratch
 LENGTH          112

   1  50  ....,....1....,....2....,....3....,....4....,....5
SEQ         NAEIQSELAETQANLAKAQSLETRIGQSTENSNLNKAQAQLAETQSNLNA
STRUC        HHHHHHHHHHHHHHHHHHHHHHHT  GGGSHHHHHHHHHHHHHTTTHH

ROBSON      HHHHHHHHHHHHHHHHHHHHHEETT      T T  HHHHHHHHHT
ORAKEL       HHHHHHHHHHHHHHHHHHHHHHH .          HHHHHHH
GOR          HHHHHHHHHHHHHHHHHH HEEEEE        HHHHHHHHHHHHH H H
GORCLASS    22333345555553543221212231222333313334344444331213

  51 100  ....,....1....,....2....,....3....,....4....,....5
SEQ         TARHPNANDNDSTQSNLNEAQAQLAKTQNAVTKYGDSTEISELANTQKNL
STRUC       HHHSTTT  TTTHHHHHHHHHHHHHHHHHHHTTT SHHHHHHHHHHHH

ROBSON       EE  T TTTT        HHHHHHHHHHEEEETT    EHHHHHHHHHH
ORAKEL                         HHHHHHH              HHHHH     H
GOR         HH             H  HHHHHHHHHHHH HEEE    EHHHHHHHHHH
GORCLASS    11133455554311133354555544232133124542211343333434

 101 112  ....,....1....,....2....,....3....,....4....,....5
SEQ         AAAQEKLAKATK
STRUC       HHHHHHHHHHT

ROBSON      HHHHHHHHHHH
ORAKEL      HHHHHHHHH
GOR         HHHHHHHHHHH
GORCLASS    555555554424

BUND ROBSON       112 72.3%
                                P R E D I C T E D
                                 EBAPM TCLS  HGI..
                      OBS EBAPM      0     0     0      0
                      OBS TCLS       0    21     3     24
                      OBS HGI..      9    19    60     88
                                     9    40    63
                                 %   8    36    56

BUND ORAKEL       112 67.0%
                                P R E D I C T E D
                                 EBAPM TCLS  HGI..
                      OBS EBAPM      0     0     0      0
                      OBS TCLS       0    23     1     24
                      OBS HGI..      0    36    52     88
                                     0    59    53
                                 %   0    53    47

BUND GOR          112 76.8%
                                P R E D I C T E D
                                 EBAPM TCLS  HGI..
                      OBS EBAPM      0     0     0      0
                      OBS TCLS       3    17     4     24
                      OBS HGI..      6    13    69     88
                                     9    30    73
                                 %   8    27    65
```

```
$PDB:RCU2.DSSP_MODEL ! model
 LENGTH          114

    1  50  ....,....1....,....2....,....3....,....4....,....5
SEQ         DADEQADICESLHDHADELYRSALARHPGTAQKTAINMARFIRSQTLTLL
STRUC         HHHHHHHHHHHHHHHHHHHHHHHHGGGG  HHHHHHHHHHHHHHHHHH

ROBSON      HHHHHHHHHHHHHHHHHHHHHHHHH    T HHHHHHHHHEEETT    EHH
ORAKEL         HHHHHHHH     HHHHHHHHHHH     HHHHHHHHHHHHHHHHHHHH
GOR          H HHHHHHH H     HHHHHHHHH      H HHHHHHHEEEHHHHHHHH
GORCLASS    3122122433121332144333432134221122412231111132123

   51 100   ....,....1....,....2....,....3....,....4....,....5
SEQ         EKLNELGQAPKQIKTALNMARFIRSQTLTLLEKMNEHDCDEQADICESLH
STRUC       HHHHS SS TTHHHHHHHHHHHHHHHHHHHHHHHTTT HHHHHHHHHHH

ROBSON      HHHHHHHHHHHHHEEHHHHHHHHET    HHHHHHHHHHHHHHHHHHHHHH
ORAKEL      HHHHHH     HHHHHHHHHHHHHHHHHHHHHHH        HHHHHHHH
GOR         HHHHHH H HHHHHHHHHHHHHHHH HHHHHHHH           HHH H
GORCLASS    4242331221333334323211122123133311312333223132212 1

 101 114   ....,....1....,....2....,....3....,....4....,....5
SEQ         DHADELYRSCLARF
STRUC       HHHHHHHHHHHTT
ROBSON      HHHHHHHHHHHHE
ORAKEL        HHHHHHHHHH
GOR             HHHHHHH HE
GORCLASS    34423323123111
```

RCU2 ROBSON 114 65.8%

		P R E D I C T E D EBAPM	TCLS	HGI..	
OBS EBAPM		0	0	0	0
OBS TCLS		1	0	17	18
OBS HGI..		7	14	75	96
		8	14	92	
	%	7	12	81	

RCU2 ORAKEL 114 80.7%

		P R E D I C T E D EBAPM	TCLS	HGI..	
OBS EBAPM		0	0	0	0
OBS TCLS		0	11	7	18
OBS HGI..		0	15	81	96
		0	26	88	
	%	0	23	77	

RCU2 GOR 114 68.4%

		P R E D I C T E D EBAPM	TCLS	HGI..	
OBS EBAPM		0	0	0	0
OBS TCLS		1	9	8	18
OBS HGI..		3	24	69	96
		4	33	77	
	%	4	29	68	

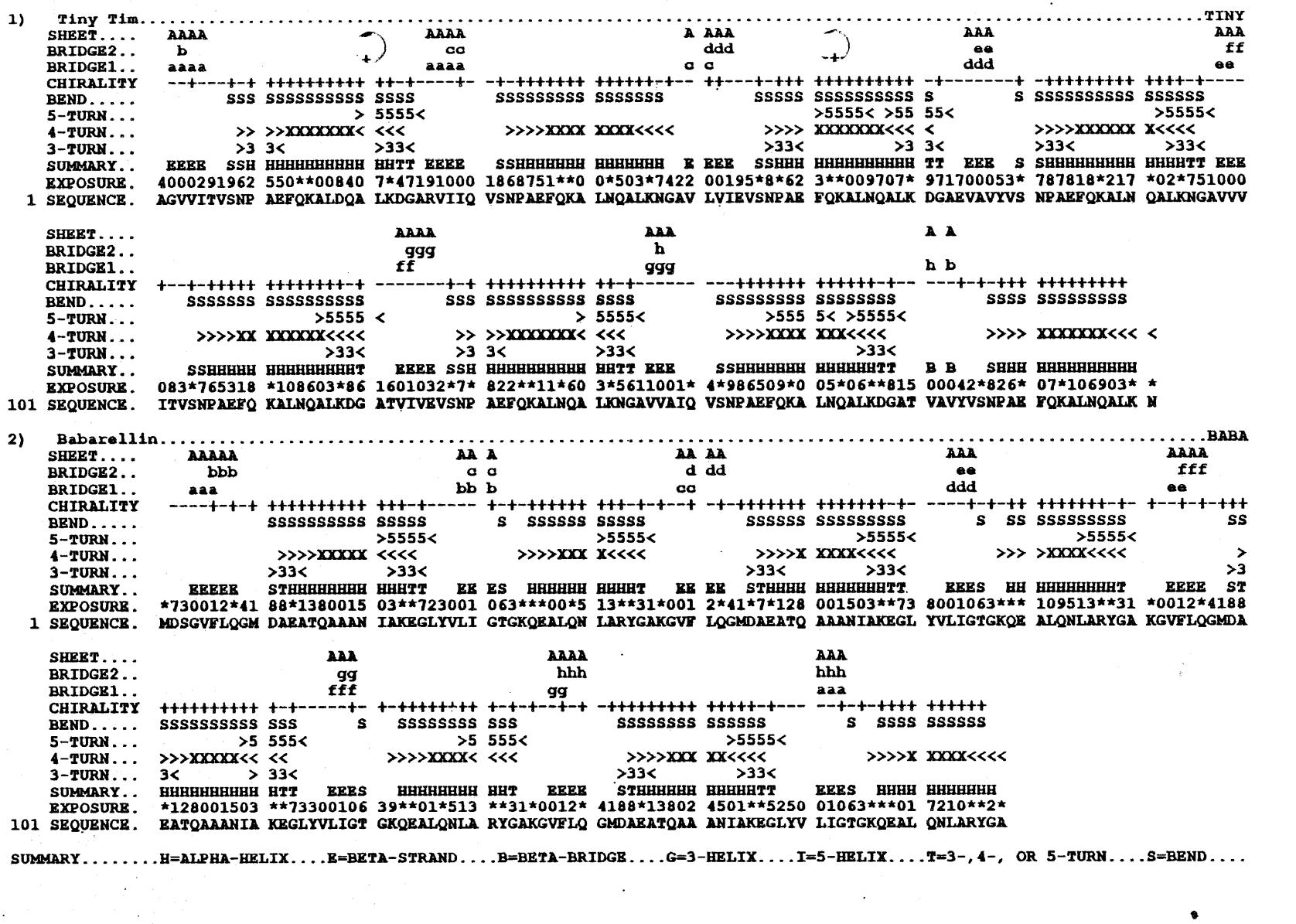
DSSP DIGEST OF DESIGNED PROTEINS

Sequences
Secondary Structure
Solvent Accessibility
Chirality

Legend

TABLE AII
Structure Notation Used in Table AIII

First line:	running number 1–62, data set identifier (3PTI,4LDH . . .), protein name, [function], [source]
SHEET. . .	One-character name of β-sheet ("A," "B," "C" . . .) in which residue i participates.
BRIDGE2 . . .	One-character name of β-ladders in which residue i participates,
BRIDGE1 . . .	"A," "B," "C" . . . = antiparallel, "a," "b," "c" . . . = parallel. Ladders are named sequentially from N- to C-terminus. A β-strand can be part of two ladders, one to each side, so there are two lines for the possible ladder partners. Each ladder name appears twice, once for each participating strand. Partner strands can thus be easily identified by identical letters. The sheet topology can be reconstructed by starting from a β-strand and tracing all partners and their partners.
CHIRALITY	"+" or "−" Chirality at residue i is the sign of the dihedral angle defined by C^α $i-1$ to $i+2$. Thus, a right-handed α-helix has "+," an ideal twisted β-strand "−."
BEND . . .	"S" = five-residue bend centered at residue i.
5-TURN . . .	Hydrogen-bonding pattern for turns and helices:
4-TURN . . .	")" = backbone CO of this residue makes H bond $(i,i+n)$
3-TURN . . .	"(" = backbone NH of this residue makes H bond $(i-n,i)$ "X" = both CO and NH make H bond "3," "4," "5" = residues bracketed by H bond
SUMMARY . . .	Structure summary: "H" = 4-helix (α-helix) "B" = residue in isolated β-bridge "E" - extended strand, participates in β-ladder "G" = 3-helix (3_{10}-helix) "I" = 5-helix (π-helix) "T" = H-bonded turn "S" = bend In case of structural overlaps, priority is given to the structure first in this list.
EXPOSURE . . .	Solvent exposure is the estimated number of water molecules in contact with residue i. The scale is 0–9; "*" = more than 9 water molecules. Exposure can be read as solvated surface area in units of 10 Å^2.
SEQUENCE . . .	Amino acid sequence in one letter code: "a," "b," "c" . . . are Cys residues labeled by their SS-bond name. "!" = chain break (peptide bond length exceeds 2.5 Å). Residues including chain breaks are numbered sequentially within the coordinate data set, irrespective of the residue identifier given there. Thus, the total number of residues is equal to the total number of print positions minus the number of chain breaks.


1) Tiny Tim.....TINY
SHEET.... AAAA AAAA A AAA AAA AAA
BRIDGE2.. b cc ddd ee ff
BRIDGE1.. aaaa aaaa c c ddd ee
CHIRALITY --+---+-+ ++++++++++ ++-+----+- -+-+++++++ ++++++-+-- ++---+-+++ ++++++++++ -+-------+ -+++++++++ ++++-+----
BEND..... SSS SSSSSSSSSS SSSS SSSSSSSSS SSSSSSS SSSSS SSSSSSSSSS S S SSSSSSSSSS SSSSSS
5-TURN... > 5555< >5555< >55 55< >5555<
4-TURN... >> >>XXXXXXX< <<< >>>>XXXX XXXX<<<< >>>> XXXXXXX<<< < >>>>XXXXXX X<<<<
3-TURN... >3 3< >33< >33< >3 3< >33< >33<
SUMMARY.. EEEE SSH HHHHHHHHHH HHTT EEEE SSHHHHHHHH HHHHHHH E EEE SSHHH HHHHHHHHHH TT EEE S SHHHHHHHHH HHHHTT EEE
EXPOSURE. 4000291962 550**00840 7*47191000 1868751**0 0*503*7422 00195*8*62 3**009707* 971700053* 787818*217 *02*751000
1 SEQUENCE. AGVVITVSNP AEFQKALDQA LKDGARVIIQ VSNPAEFQKA LNQALKNGAV LVIEVSNPAE FQKALNQALK DGAEVAVYVS NPAEFQKALN QALKNGAVVV
SHEET.... AAAA AAA A A
BRIDGE2.. ggg h
BRIDGE1.. ff ggg h b
CHIRALITY +--+-+++++ ++++++++-+ -------+-+ ++++++++++ ++-+------ ---+++++++ ++++++-+-- ---+-+-+++ +++++++++
BEND..... SSSSSSS SSSSSSSSSS SSS SSSSSSSSSS SSSS SSSSSSSSS SSSSSSSS SSSS SSSSSSSSS
5-TURN... >5555 < > 5555< >555 5< >5555<
4-TURN... >>>>XX XXXXXX<<<< >> >>XXXXXXX< <<< >>>>XXXX XXX<<<< >>>> XXXXXXX<<< <
3-TURN... >33< >3 3< >33< >33<
SUMMARY.. SSHHHHH HHHHHHHHHT EEEE SSH HHHHHHHHHH HHTT EEE SSHHHHHHH HHHHHHTT B B SHHH HHHHHHHHHH
EXPOSURE. 083*765318 *108603*86 1601032*7* 822**11*60 3*5611001* 4*986509*0 05*06**815 00042*826* 07*106903* *
101 SEQUENCE. ITVSNPAEFQ KALNQALKDG ATVIVEVSNP AEFQKALNQA LKNGAVVAIQ VSNPAEFQKA LNQALKDGAT VAVYVSNPAE FQKALNQALK N
2) Babarellin.....BABA
SHEET.... AAAAA AA A AA AA AAA AAAA
BRIDGE2.. bbb c c d dd ee fff
BRIDGE1.. aaa bb b cc ddd ee
CHIRALITY ----+-+-+ ++++++++++ +++-+----- +-+-++++++ +++-+-+--+ -+-+++++++ +++++++-+- ----+-+-++ +++++++-+- +--+-+-+++
BEND..... SSSSSSSSS SSSSS S SSSSSS SSSSS SSSSSS SSSSSSSSS S SS SSSSSSSSS SS
5-TURN... >5555< >5555< >5555< >5555<
4-TURN... >>>>XXXXX <<<< >>>>XXX X<<<< >>>>X XXXX<<<< >>> >XXXX<<<< >
3-TURN... >33< >33< >33< >33< >33<
SUMMARY.. EEEEE STHHHHHHHH HHHTT EE ES HHHHHH HHHHT EE EE STHHHH HHHHHHHTT EEES HH HHHHHHHHT EEEE ST
EXPOSURE. *730012*41 88*1380015 03**723001 063***00*5 13**31*001 2*41*7*128 001503**73 8001063*** 109513**31 *0012*4188
1 SEQUENCE. MDSGVFLQGM DAEATQAAAN IAKEGLYVLI GTGKQEALQN LARYGAKGVF LQGMDAEATQ AAANIAKEGL YVLIGTGKQE ALQNLARYGA KGVFLQGMDA
SHEET.... AAA AAAA AAA
BRIDGE2.. gg hhh hhh
BRIDGE1.. fff gg aaa
CHIRALITY ++++++++++ +-+-----+- +-++++++++ +-+-+--+-+ -+++++++++ +++++-+--- --+-+-++++ ++++++
BEND..... SSSSSSSSSS SSS S SSSSSSSS SSS SSSSSSSS SSSSSS S SSSS SSSSSS
5-TURN... >5 555< >5 555< >5555<
4-TURN... >>>XXXXX<< << >>>>XXXX< <<< >>>>XXX XX<<<< >>>>X XXXX<<<<
3-TURN... 3< > 33< >33< >33<
SUMMARY.. HHHHHHHHHH HTT EEES HHHHHHHH HHT EEEE STHHHHHH HHHHHTT EEES HHHH HHHHHHH
EXPOSURE. *128001503 **73300106 39**01*513 **31*0012* 4188*13802 4501**5250 01063***01 7210**2*
101 SEQUENCE. EATQAAANIA KEGLYVLIGT GKQEALQNLA RYGAKGVFLQ GMDAEATQAA ANIAKEGLYV LIGTGKQEAL QNLARYGA
SUMMARY.....H=ALPHA-HELIX....E=BETA-STRAND....B=BETA-BRIDGE....G=3-HELIX....I=5-HELIX....T=3-,4-, OR 5-TURN....S=BEND....

```
3)   Mutated Flavodoxin.........................................................................................FXNM
    SHEET....   AAAA                        AAA A              AAAAA    B   B                      AAA AAAAA
    BRIDGE2..   bbbb                                           ccccc                               dd   eeee
    BRIDGE1..   aaaa                        aaa a               bbbb    F   F                      ccc cc
    CHIRALITY   --+--++++ ++++++++++ ++-+-+--+- -+++--++++ ++-+--+-+- -++--+---+ ++++++++++ ++++++---- ++-+-++--+ ++++++++++
    BEND.....         SSSS SSSSSSSSSS SSSS        SSS  SSSS SS S          SSSS  S SSSSSSSSSS SSS SS          SSS S SSSSSSSSSS
    5-TURN...                       > 5555<                                    >5 555<                                      >
    4-TURN...           >> >>XXXXXXX< <<<        >444< >444 <            >444<     >>>>XX<<< <                   > >>>XXXXXX<
    3-TURN...        >>3X< 3<       > >3<<       >33<     > 33<                              >33X33<
    SUMMARY..   EEEE GGGH HHHHHHHHHH HHTT   EEE ETTT  STTT TT SEEEEE    BTTTB  T TTHHHHHHHH STT TT EEE EEEEESSS S HHHHHHHHHH
    EXPOSURE.  4500136409 9009101*00 9*57**1952 517*484**0 3*5*300000 13*7*3512* 9805610**0 2**0863900 000252*2*0 901*607*50
  1 SEQUENCE.  MKIVYWSDTE KMAELIAKGI IESGKDVNTI NVSDVNIDEL LNEDILILGC SARGDEVLEE SEFEPFIEEI STKISGKKVA LFGNYGWGDG KWMRDFEERM

    SHEET....       AA     AAAA
    BRIDGE2..
    BRIDGE1..       dd     eeee
    CHIRALITY  ++-+---+-- -+---+-+++ ++++++++++ ++
    BEND.....  SSSS    S      SSS SS S SSSSSSSS SS
    5-TURN...  5555<
    4-TURN...  <<<                   >>>>XXXXXX <<<<
    3-TURN...  >33<              >>3 <<         >33<
    SUMMARY..  HHTT EE S   EEEESS GG GHHHHHHHHH HHH
    EXPOSURE.  976317364* 2361***19* 28**018107 *11*
101 SEQUENCE.  NGYGCVVVET PLIVQNEPDE AEQDCIEFGK KIAN

4)   Idealized Flavodoxin.......................................................................................FXNI
    SHEET....   AAAA                   AAAA                AAAAAA B    B                      AAAAA
    BRIDGE2..   bbbb                                        ccccc                              ddd
    BRIDGE1..   aaaa                   aaaa                bbbb   E    E                      ccccc
    CHIRALITY   -----+--+ ++++++++-+ -++---+++- -++++++-++ ---+--++-- +---+-++++ ++++++-+++ +-----+-++ --++++++++ +++++-++++
    BEND.....        SS S SSSSSSSSS        SSS   SSSSSS S          SS S   SSSSSS SSSSS SS   S       SS  S S SSSSSS SSSSSSS  S
    5-TURN...                  >5555 <                                   >5555<                                     >5555<
    4-TURN...           > 4>>X><<<<              >444<            >44 4<    >>>> XX<<<<                   >>>>XXXX XX<<<<
    3-TURN...              >33< >33<       >>3<<   >3><3<                            >3><3<                >33<      >>3<<>33
    SUMMARY..   EEEE SS S TTHHHHHHTT   EEEEGGG   STTTTT S  EEEEEE BTT TB  TTTHHH HHHHHTTT   S EEEEE SS  S SHHHHHHH HHHHHTT TT
    EXPOSURE.  72000061*7 401*007*59 2*791707*3 91**04*7*1 011102375* 2612**7087 10**038*17 510000438* 3*0900*606 *9187661**
  1 SEQUENCE.  ALVVVASGTE AADLAMSHQA TRTKVNVSDV NIDELLNEDG LIMIISAMGD EVLEESEFEP FSEEINTKIS ASVVIAYSGW GDGKWMRDFE ERMNGTGIKY

    SHEET....    AAA
    BRIDGE2..
    BRIDGE1..    ddd
    CHIRALITY  ---++-+-+- +++++++++
    BEND.....        SS   SSSSSSSSS
    5-TURN...
    4-TURN...             >>>>XXX<<< <
    3-TURN...  <                 >33 <
    SUMMARY..    EEE SS   SHHHHHHHHH
    EXPOSURE.  2443537*9* 1*706812** 4
101 SEQUENCE.  SNSMNNGRKD GKDMYEIGKK I

SUMMARY.......H=ALPHA-HELIX....E=BETA-STRAND....B=BETA-BRIDGE....G=3-HELIX....I=5-HELIX....T=3-,4-, OR 5-TURN....S=BEND....
```

```
5)   Betalphacin.....BEAL
     SHEET....      AAAB                  A  B                 AAAA                    AA
     BRIDGE2..      bbb                                        cc
     BRIDGE1..      a  d                  a  d                 bbb                     cc
     CHIRALITY   -++-+-+++ -+++++++++ +++++++-+- ++----+--+ +-++++++++ ++++++--+- -+-+++++++ +++++++++- ---++--+++ +++++++++
     BEND.....           S SSSSSSSSSS SSSSSSSSS  S        S SSSSSSSSSS SS          SS SSSSSS SSSS  S       SSS SSS SSSSSSSSS
     5-TURN...
     4-TURN...             >>>>XXXXX XXXX<X<<4<              >>>>XXXX< <<<           >>>>XXX <X<<4<           >>>> XXXXXXX<<<
     3-TURN...             >33<>33<       >33<               >33<   >3 3<            >33<     >33<         >>3<X33<      >33<
     SUMMARY..      EEEB S SSHHHHHHHH HHHHHHHTT  S   B  B S SSHHHHHHHH HH    EEEE  SS HHHHHH HHHTT S      EEGGG HHH HHHHHHHHHH
     EXPOSURE.   *732000225 *1609*00*1 06*00**674 9*161172*5 *4**05**1* 9848030001 2*755*71*9 19*726*2*3 3089*77*9* 02*402*21*
   1 SEQUENCE.   MDAMILVISG TGNTERAADA FEEGARKTGT PPVTIIYVKN EGEDAYEKFH ARGDVTVIVV GPTANEHLED FRKLAGNLKV TVKYEINEDK IAELGAKVVN

     SHEET....
     BRIDGE2..
     BRIDGE1..
     CHIRALITY
     BEND.....
     5-TURN...
     4-TURN...  <
     3-TURN...
     SUMMARY..
     EXPOSURE.  6
   101 SEQUENCE.  G

6)   Bundle.....BUND
     SHEET....
     BRIDGE2..
     BRIDGE1..
     CHIRALITY   +++++++++ ++++++++++ ++++-+-+++ ++++++++++ ++++++++++ +++++++--+ ++++++++++ ++++++++++ +++-+-++++ ++++++++++
     BEND.....    SSSSSSSS SSSSSSSSSS SSSSS  SSS SSSSSSSSSS SSSSSSSSSS SSSSSSS  S SSSSSSSSSS SSSSSSSSSS SSSS  SSSS SSSSSSSSSS
     5-TURN...                                                                                              >5555<
     4-TURN...  >>44XX4>X< >XX>X<XXX> X<<<<      >>>>XXXX<< XX4><X4X>X 4<<X444<>4 4>X4>XX4XX X>XXXXXXX< <<<   >>4> X<>XX>X<XX
     3-TURN...  >33<   >33 <  >33X33<  >>3<<>>3< X33< >>3<< >33<   >>3 << >33< >3 3X33<>33X3 3<   >33<  >>3<< >33X 33< >>3<<
     SUMMARY..   HHHHHHHHHH HHHHHHHHHH HHHHT  GGG SHHHHHHHHH HHHHHTTTHH HHHSTTT  T TTHHHHHHHH HHHHHHHHHH HHTTT SHHH HHHHHHHHHH
     EXPOSURE.  96*126602* 013603*117 706*8629*0 17903*0258 03*1156050 36*97*91*1 78106704*0 14402*1095 07**729*01 6801900*70
   1 SEQUENCE.  NAEIQSELAE TQANLAKAQS LETRIGQSTE NSNLNKAQAQ LAETQSNLNA TARHPNANDN DSTQSNLNEA QAQLAKTQNA VTKYGDSTEI SELANTQKNL

     SHEET....
     BRIDGE2..
     BRIDGE1..
     CHIRALITY  ++++++++++
     BEND.....  SSSSSSSSSS
     5-TURN...        >555 5<
     4-TURN...  <4XX4>X<4< <
     3-TURN...       >>3<<
     SUMMARY..  HHHHHHHHHH T
     EXPOSURE.  35118*01*4 3*
   101 SEQUENCE.  AAAQEKLAKA TK

SUMMARY.....H=ALPHA-HELIX....E=BETA-STRAND....B=BETA-BRIDGE....G=3-HELIX....I=5-HELIX....T=3-,4-, OR 5-TURN....S=BEND....
```

```
7)   Copper Rop dimer.....RCU1
    SHEET....
    BRIDGE2..
    BRIDGE1..
    CHIRALITY   -++++++++ ++++++++++ ++++++++-+ -+++++++++ ++++++++++ ++++     -+ ++++++++++ ++++++++++ +++++-+-++ ++++++++++
    BEND.....    SSSSSSSS SSSSSSSSSS SSSSSSSSS   SSSSSSSSS SSSSSSSSSS SSSS     S SSSSSSSSSS SSSSSSSSSS SSSSSS  SS SSSSSSSSSS
    5-TURN...                        >5555 <                                                                 >5555<
    4-TURN...   >>>>XXXXX XXXXXXXXXX XXXXX<<<<  >>>>XXXXXX XXXXXXXXXX XX<<<<  >> >>XXXXXXXX XXXXXXXXXX XX<<<< >>> >XXXXXXXXX
    3-TURN...                  >33<       >33<                                                  >33<       >33<
    SUMMARY..     HHHHHHHHH HHHHHHHHHH HHHHHHHHTT  HHHHHHHHH HHHHHHHHHH HHHHH    H HHHHHHHHHH HHHHHHHHH HHHHHTT HH HHHHHHHHHH
    EXPOSURE.   *7**5*7007 20*9076407 602*608*4* 1**11*40*6 028204*03* 2079*40*7* *5*700720* 9076407602 *608*4*1** 11*40*6028
  1 SEQUENCE.   MTKQEKTAIN MARFIRSQTL TLLEKMNEHD CDEQADICES LHDHADELYR SALARH|MTK QEKTAINMAR FIRSQTLTLL EKMNEHDCDE QADICESLHD

    SHEET....
    BRIDGE2..
    BRIDGE1..
    CHIRALITY  ++++++++++ +
    BEND.....  SSSSSSSSSS S
    5-TURN...
    4-TURN...  XXXXXXXXX< <<<
    3-TURN...
    SUMMARY..  HHHHHHHHHH HH
    EXPOSURE.  204*03*207 9*4
101 SEQUENCE.  HADELYRSAL ARH

8)   Copper Rop monomer.....RCU2
    SHEET....
    BRIDGE2..
    BRIDGE1..
    CHIRALITY   +++++++++ ++++++++++ ++++++++++ +-++++++++ ++++++++++ ++++++++-+ ++++++++++ ++++++++++ ++++++-+-+ ++++++++++
    BEND.....     SSSSSSS SSSSSSSSSS SSSSSS SSS   SSSSSSSS SSSSSSSSSS SSSSS SS S SSSSSSSSSS SSSSSSSSSS SSSSSSS  S SSSSSSSSSS
    5-TURN...                                                                                                >5555<
    4-TURN...   >>>>XXXXX XXXXXXXXXX XXX<<<<     >>>>XXXXX XXXXXXXXXX X<<<<   >4 >>X>XXXXXX XXXXXXXXXX XX<X<<4<>> >>XXXXXXXX
    3-TURN...   >>3<<   > 33<>33<>33 <    >>><< <>33<               > 33<       >33<                    >33<>33<>> 3<<   >33<
    SUMMARY..    HHHHHHHH HHHHHHHHHH HHHHHHGGGG   HHHHHHHH HHHHHHHHHH HHHHS SS T THHHHHHHHH HHHHHHHHHH HHHHHTTT H HHHHHHHHH
    EXPOSURE.   *09*01*40* 602*303*13 *2098*0*84 4**902630* 9075408702 *607*16*2* *74*703730 9907540870 2*717*4*09 *01*40*602
  1 SEQUENCE.   DADEQADICE SLHDHADELY RSALARHPGT AQKTAINMAR FIRSQTLTLL EKLNELGQAP KQIKTALNMA RFIRSQTLTL LEKMNEHDCD EQADICESLH

    SHEET....
    BRIDGE2..
    BRIDGE1..
    CHIRALITY  ++++++++++ ++
    BEND.....  SSSSSSSSSS SS
    5-TURN...
    4-TURN...  XXXXXXXX<X <<4<
    3-TURN...  >33<>33<>> 3<<
    SUMMARY..  HHHHHHHHHH HTT
    EXPOSURE.  *303*03*20 89*7
101 SEQUENCE.  DHADELYRSC LARF

SUMMARY.....H=ALPHA-HELIX....E=BETA-STRAND....B=BETA-BRIDGE....G=3-HELIX....I=5-HELIX....T=3-,4-, OR 5-TURN....S=BEND....
```

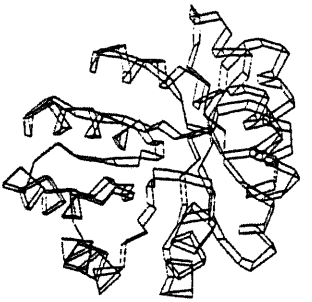
STEREO VIEWS OF DESIGNED PROTEINS

C-alpha ribbon drawings using a program by Arthur Lesk

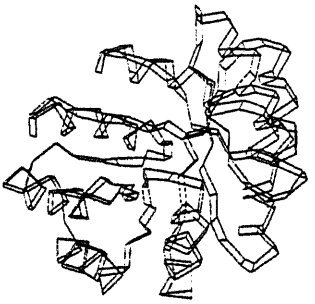
TINY
BABA

FXNI
FXNM
BEAL

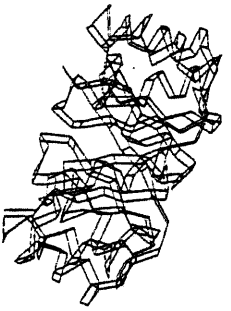
BUND
RCU2



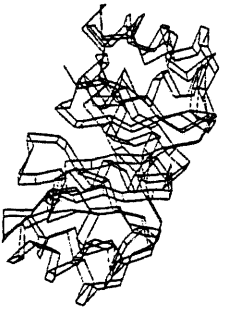
TINY MODEL



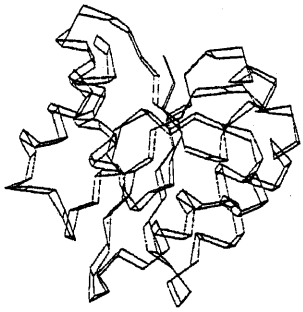
TINY MODEL



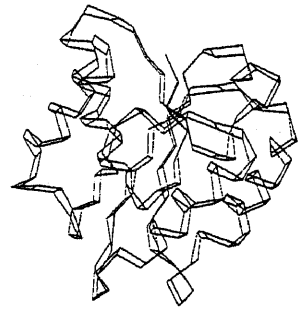
BABA MODEL



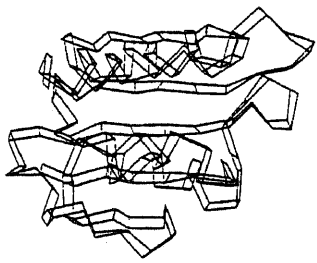
BABA MODEL



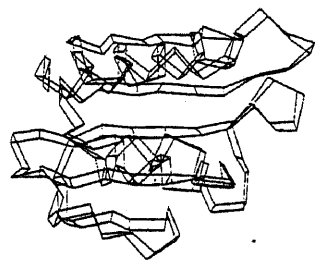
FXNM MODEL



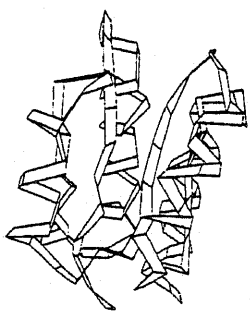
FXNM MODEL



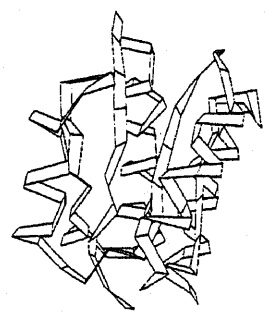
FXNI MODEL



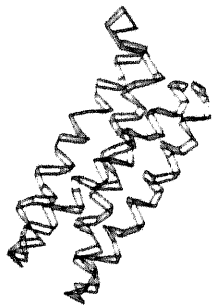
FXNI MODEL



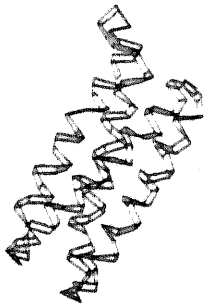
BEAL MODEL



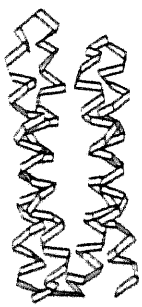
BEAL MODEL



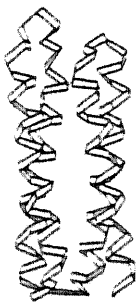
BUND MODEL



BUND MODEL



RCL2 MODEL



RCL2 MODEL

fax sent
(today!)
CS