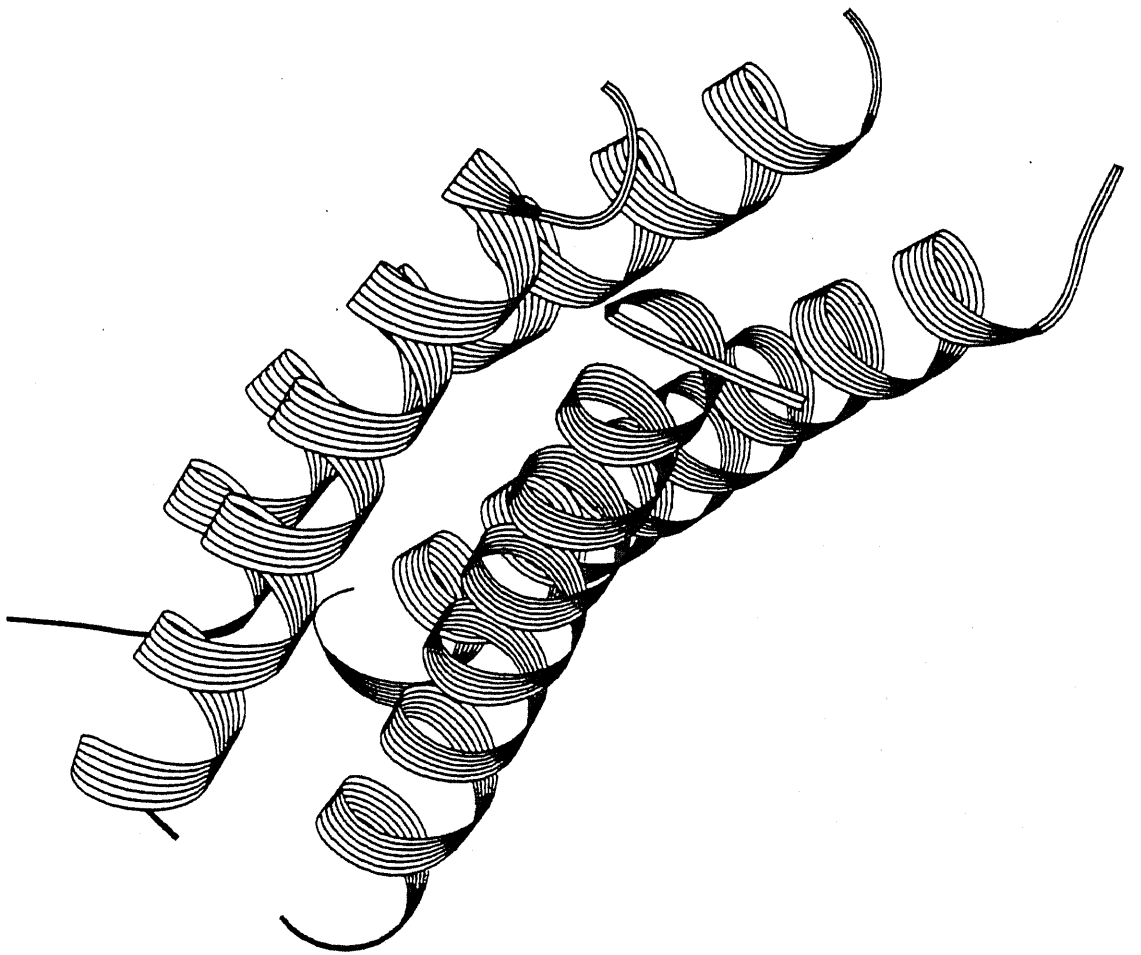
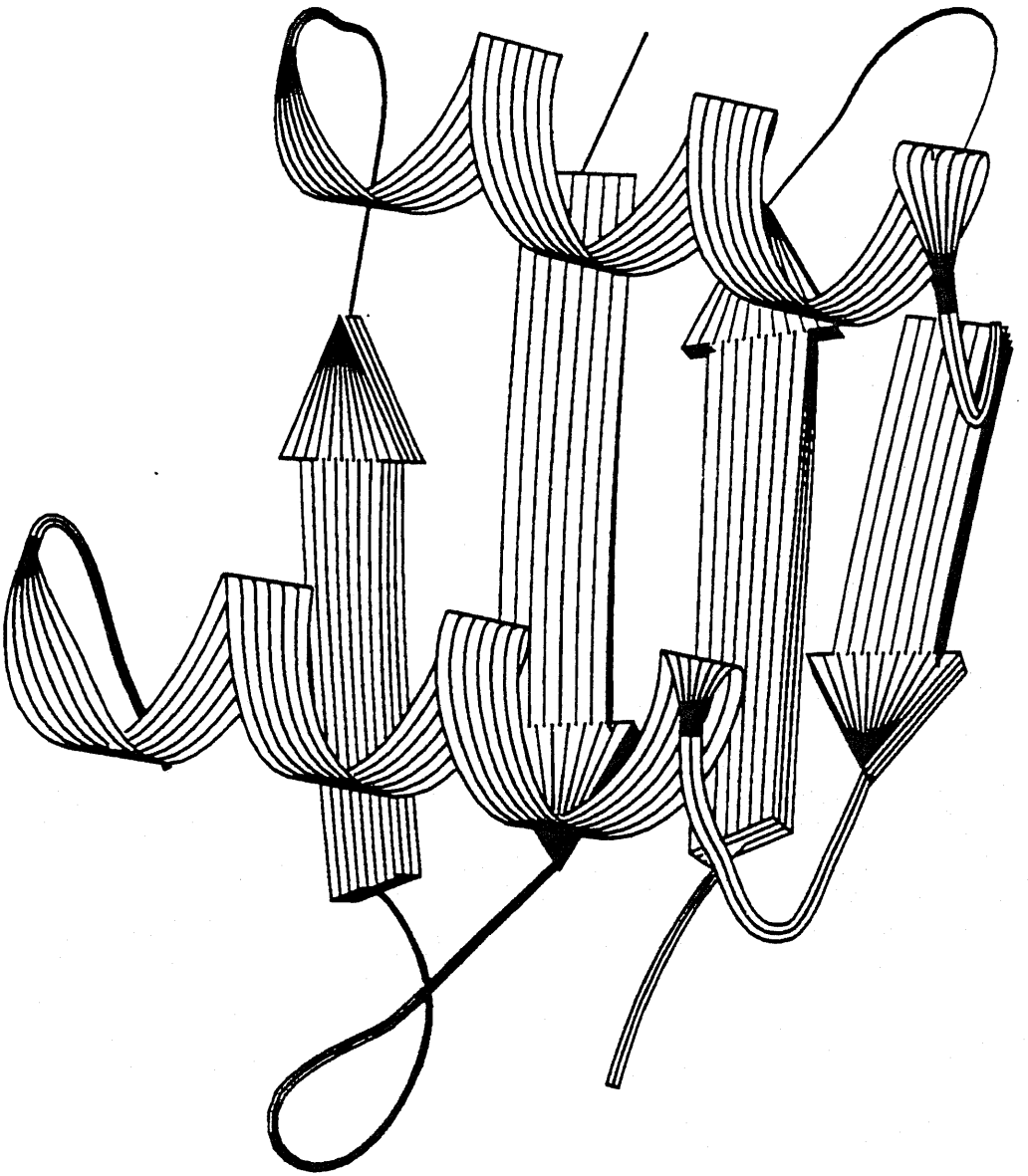
# PRODES90

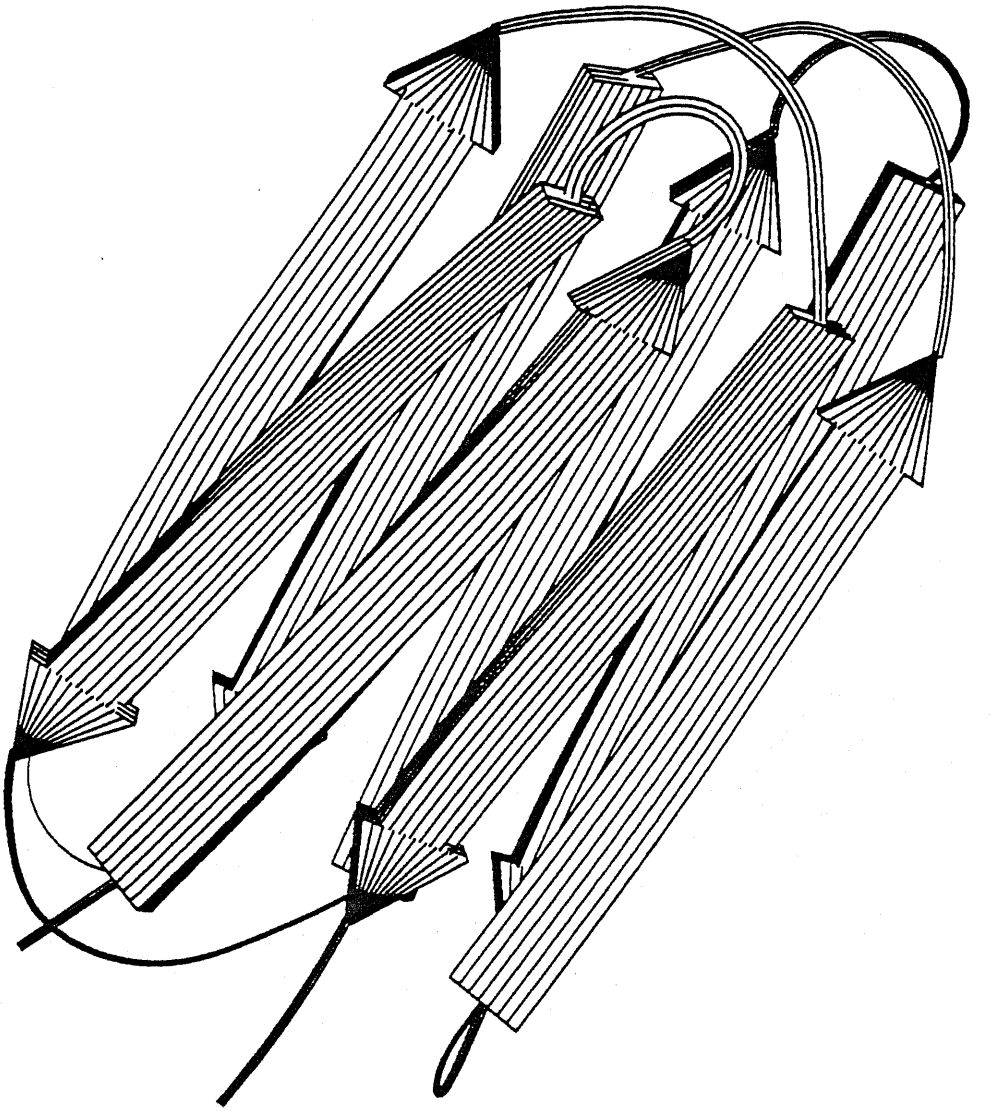
# Protein Design on Computers

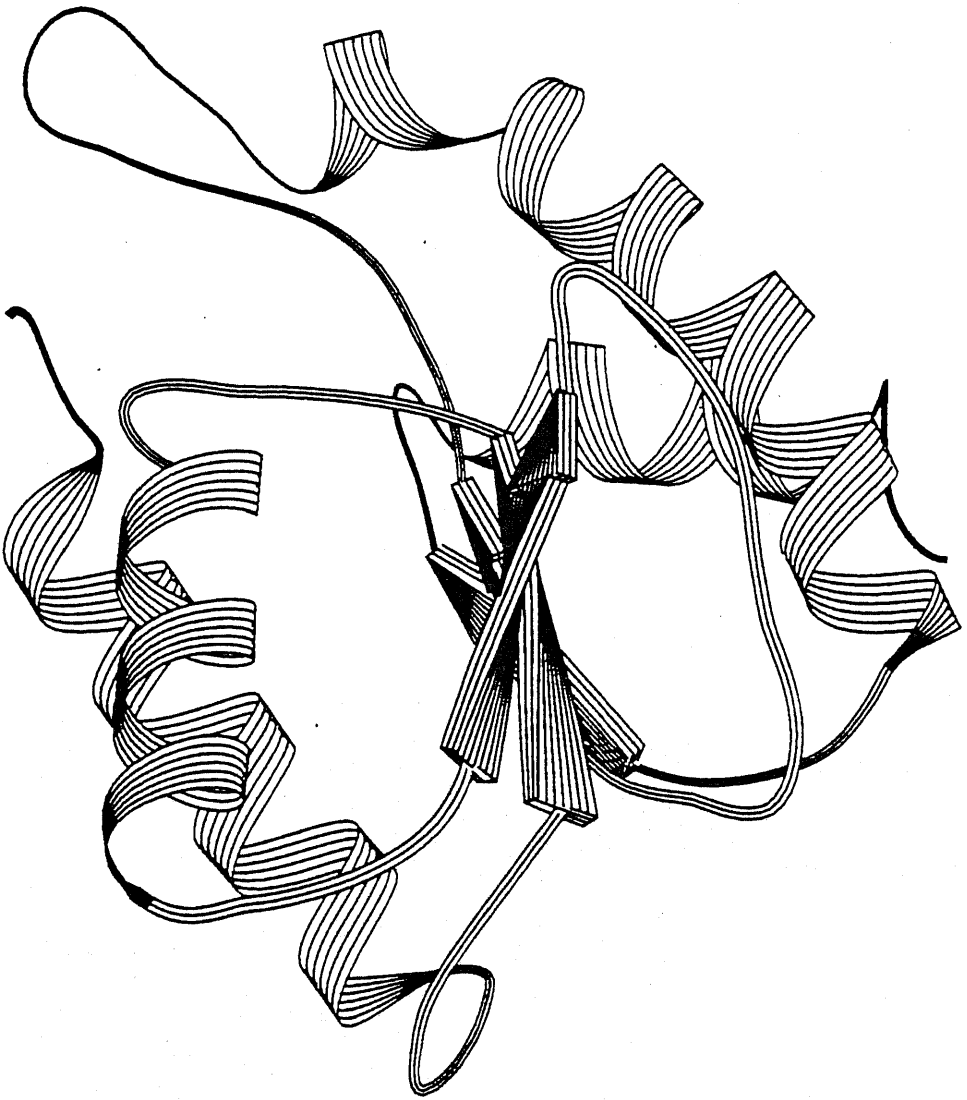
*C. Sander and G Vriend, eds.*

EMBL BIOcomputing Technical Document 6 1991









# Foreword

by Chris Sander and Gerrit Vriend

The material in this report was developed during the EMBO Practical Course 'Protein Design on Computers' held at EMBL Sept. 30 - Oct. 13, 1990

What is the current state of the art in protein design? This question was approached in a two-week protein design workshop sponsored by EMBO and held at the EMBL in Heidelberg. The goals were to test available design tools and to explore new design strategies. Five novel proteins were designed: Shpilka, a sandwich of two four-stranded β-sheets, a scaffold on which to explore variations in loop topology; Grendel, a four-helical membrane anchor, ready for fusion to water-soluble functional domains; Fingerclasp, a dimer of interdigitating β–β–α units, the simplest variant of the 'handshake' structural class; Aida, an antibody binding surface intended to be specific for flavodoxin; Leather - a minimal NAD binding domain, extracted from a larger protein. Each design is available as a set of three-dimensional coordinates, the corresponding amino acid sequence and a set of analytical results. The designs are placed in the public domain for scrutiny, improvement and possible experimental verification.

**From protein folding to protein design**

The forward protein folding problem, that of calculating structure from sequence, remains basically unsolved. The inverse problem, that of designing sequences to achieve desired structural properties, has been the subject of considerable effort over the last few years . We have attempted to promote the development of new techniques and to encourage the design of new types of proteins by organizing intensive workshops, first in 1986 and, as reported here, in 1990.

The problem posed to the participants was this: specify a model protein structure (or structural property) and then invent a protein sequence that will lead to the desired structure. In the same spirit, in 1986 two idealized 4*βαβα barrel structure scaffolds (Babarallin, Tiny Tim), two β/α folds (Betaalphacin, Idealized Flavodoxin), a bundle of four α-helices (Bundle) and a Cu-binding variant of a natural protein (CuRop) were constructed, and the corresponding protein sequences invented. Since then, Babarellin and CuRop have been synthesized and purified (G.Nyakatura, H.-J.

Fritz, S.C.Emery, personal communications) and structural tests are in progress (F.X.Schmid, W.Eberle, J. Richardson, M. Sagermann personal communications). The five new proteins designed in the recent 1990 workshop are described below by each of the working groups.

**Techniques used**

Typically, the design procedure followed these steps: (i) analyze known protein structures and identify structural units that might be used as building blocks, e.g. αβ units; (ii) sketch out the secondary structure elements, their relative orientations and the topology of loop connections, e.g. a four-stranded antiparallel β-sheet packed against two a-helices crossing the strands; (iii) construct the protein scaffold by building explicit backbone coordinates, first for the structural core, consisting primarily of elements of secondary structure, then for loops; (iv) choose an appropriate amino acid sequence in interior and surface regions, e.g. Glu on the surface of a helix near the N-terminus, Val or Ile at a β–β or β–α interface and so on; (v) optimize the model in interactive mode using visual inspection or in automatic mode using molecular mechanics software. i.e. vary backbone and side chain degrees of freedom, with simple energetics as a guide; primary goals are to regularize covalent geometry, remove clashes, avoid holes, optimize hydrogen bonding as well as charge-charge and protein-solvent interactions; (vi) check the quality of the model by analyzing , e.g., solvation, electrostatic or interior packing in more detail than can be done during construction; (vii) mutate the sequence where necessary. Steps (v) to (vii) are iterated until a satisfactory model is reached.

**The state of the art**

In practice the design process is driven by general understanding of protein structure and energetics, backed up by data base searches and simulation techniques. An extremely useful empirical guide to the design of both the scaffold and the conformational details are observations extracted from the database of experimentally known structures. Design software tools for modifying structures, patching loops, searching databases, for constrained energy minimization, systematic conformational exploration by dynamics or Monte Carlo procedures have improved greatly over the last few years, partly in the form of commercial software. Several new algorithms have been introduced, including database techniques for building complete protein coordinates from a $C_\alpha$ trace and for selecting sidechain environments; and

evaluation of given models by empirical criteria like internal packing and external solvation.

What is missing and required in the near future ? There is a need for more powerful constructive tools for going from a raw sketch to a first model and for globally modifying structures, e.g. increasing the overall twist of a b-sheet while maintaining optimal packing and good loop geometry. Two key technical problems remain to be solved. We need more precise energetics that can reliably distinguish between correct and incorrect structures. And, as it is still very difficult to explore a significant fraction of conformational space, we need more efficient simulation tools. Finally, the design principles for functional interfaces and for enzymatically active sites are still in their infancy.

**On to experiment**

The designs will have to be tested by gene synthesis, cloning, protein production, purification and, subsequently, structure determination by x-ray crystallography or NMR, functional assays, and folding studies. We hope that experiments like these will teach us more about protein folding and about the methods for producing proteins with desired structural properties. Experimental work is in progress for a variant of Shpilka, for Fingerclasp and for Leather.

## Work groups

Each group carries out its own design project and is responsible for submitting a final report, including coordinate data sets and amino acid sequences of the designed proteins.

1. Moult: Christophe Verlinde, Christine Gaboriaud, SteveEmery, Cara Marks
2. Nakamura: Lluis Ribas, Fernando Bazan, Amnon Horovitz
3. Finkelstein: Andrew Lockhart, Rainer Merkl, Jeanne Perry
4. Hubbard: Lynne Regan, Arne Elofsson, Marc Eberhard
5. Lesk/Tramontano: Roberto Japelli, Jay Banks

## Support

*Protein Design Group and Guests:*

Brigitte Altenberg
Wolfgang Eberle
Ulrike Goebel
Adam Godzik
Uwe Hobohm
Liisa Holm
Christos Ouzounis
Martin Sagermann
Michael Scharf
Reinhard Schneider
Pieter Stouten
Georg Tuparev
Alfonso Valencia

*Computer Group:*

Erich Schechinger
Wolfgang Winkler
Dominique Fedronic
Peter Rice
Roy Omond

*Other Groups:*

Johan Postma
Raimund Schnobel

*Secretaries:*

Ingeborg Fatscher
Christine Raulfs

## Calendar of Events

| | |
|---|---|
| Sep. 30 Sun | 17:00 Registration opens. ISG. |
| | 19:00 Welcome reception/buffet. ISG. |
| Oct. 1 Mon | 9:00 Course work starts in room 202 at EMBL |
| | 20:00 Round table: Status of current design projects |
| Oct. 2 Tue | Evening: design project outline done |
| Oct. 3 Wed | |
| Oct. 4 Thu | |
| Oct. 5 Fri | |
| Oct. 6 Sat | Morning: free to shop etc. |
| | 14:00 work resumes |
| Oct. 7 Sun | 10:00 leave for excursion; return evening. |
| Oct. 8 Mon | |
| Oct. 9 Tue | |
| Oct. 10 Wed | |
| Oct. 11 Thu | Evening: draft reports are due |
| Oct. 12 Fri | 20:00 Closing party |
| Oct. 13 Sat | Morning: departure |

## Daily events

| | |
|---|---|
| 9:00 | Course meeting: summary of previous day, announcements. |
| ~9:30 | Work starts. |
| 12:45 | Lunch |
| 14:00 | Daily lecture |
| 15:00 | Work resumes |
| 19:00 | Dinner |
| ~20:00 | Work resumes |
| 22:00 | bus leaves for guest house |

## Schedule of lectures

Lectures daily at 14:00 in the Small Operon.

| | |
|---|---|
| Oct. 1 Monday | Arthur Lesk / Anna Tramontano |
| Oct. 2 Tuesday | Haruki Nakamura |
| Oct. 3 Wednesday | John Moult |
| Oct. 4 Thursday | Alexei Finkelstein |
| Oct. 5 Friday | Tim Hubbard |

| | |
|---|---|
| Oct. 8 Monday | Jean Garnier |
| Oct. 9 Tuesday | Lynne Regan |
| Oct. 10 Wednesday | Steve Emery and Wolfgang Eberle |
| Oct. 11 Thursday | Paul Bash |
| Oct. 12 Friday | Chris Sander |

**Lecture titles**

TimHubbard, PERI, Osaka and MRC, Cambridge
Heat shock proteins and protein folding.

Haruki Nakamura, PERI, Osaka
Roles of Protein Designs at the Protein Engineering Research Institute.

John Moult, CARB, Maryland
Analysis of Protein Folding Pathways.

Alexei Finkelstein, Instutue of Protein Research,
Poushchino, USSR
Theory of Protein Structure.

Arthur Lesk, MRC Cambridge
Predizione delle anse degli anticorpi

Anna Tramontano, IRBM Rome
Prediction of antigen-binding sites of immunoglobulins

Jean Garnier, INRA, Jouy-en-Josas
Protein secondary structure prediction : an
aid for protein design and for understanding protein folding.

Lynne Regan, Yale University
Designed disulfide bonds to probe the structure of a model four helix bundle and engineering a tetrahedal zinc binding site.

Steve Emery
Design and construction of synthetic genes.

Wolfgang Eberle
NMR experiments on Rop proteins.

Paul Bash
Computer Simulation of the Enzyme Reaction in Triosephosphate Isomerase .

Chris Sander
Protein design studies using Rop.

## Computing Facilities

Course work will be done about equally under both the UNIX and VMS operating systems. Graphics workstations are all UNIX machines. Text processing can be done on Apple Macs. Laser printers speak Postscript. Protein Design software includes commercial packages as well as academic software, some of it supplied by the teachers and students. The EMBL has an excellent network of computing facilites maintained by the Computer Group: Roy Omond, Dominique Fedronic, Peter Rice, Erich Schechinger, Wolfgang Winkler.

## Networks

Local area network (EMBL) between VMS, UNIX and Mac worlds via ethernet. Global connections via X.25 packet switching network (using the German reseach network WIN) into Internet for downloading of files (ftp) and remote login exists (see Roy Omond).
Electronic mail is via bitnet and from there via gateways to other domains. Addresses are composed as e.g. IN%"JOHN@MSMFVM.bitnet" or IN%"nakamura@pes4.peri.co.jp" under VAX/VMS mail. EMBL adresses are e.g. VRIEND@EMBL.bitnet.

## Hardware

| | |
|---|---|
| Vax/VMS Cluster: | |
| Vax 6420 | sirius |
| Vax 8650 | vega |
| | |
| UNIX Cluster: | |
| SUN-4 | gold, black |
| SUN-3 | yellow, green, etc. |
| DECStations 3000 | tiger, felix, etc. |

Graphics workstations, part of UNIX cluster:

| | |
|---|---|
| IRIS 4D /20 | iris |
| IRIS 4D /210 | jo |
| Stellar GS2000 | stellar |

Graphics terminals: E+S PS390 (pollux, castor), attached to the VAX cluster.

In addition, several graphics workstations will be supplied by companies for exclusive use during the course:

| | |
|---|---|
| IRIS 4D/210 | somename |
| ESV | miraculix |

## Selected Software and databases

| | |
|---|---|
| Betasheet | Draws 2-D diagrams of beta sheets. By Shneior Lifson and Chris Sander. |
| Biogromos | User interface to Gromos. By Biostructure. |
| Composer | Builds protein models by homology. By Mike Sutcliffe. |
| Conan | Contact analysis for proteins. By Michael Scharf. |
| Constrictor | Distance geometry. By Andrew Smellie. |
| Convertcoor | Converts protein coordinates from many formats to PDB format. By Chris Sander |
| Delphi | Careful electrostatics calculations. By Barry Honig. |
| Disgeo | Distance geometry. By Tim Havel. |
| DSSP | Extracts secondary structure and surfaces from protein coordinates. By Wolfgang Kabsch and Chris Sander. |
| DssptoInsight | Display of secondary structure elements. By Brigitte Altenberg. |
| FindSite | Find metal etc. binding sites. By Adam Godzik. |
| GCG | Sequence Analysis Package. By John Devereux et al. |
| Gromos | One of the classic molecular dynamics programs. By Wilfred van Gunsteren and Herman Berendsen. |
| HmDisp | Display of hydrophobic moments. By Tim Hubbard. |
| Hssptoinsight | Colors proteins by variability. By R. Schneider. |
| I nsight | 3-D graphics. By Anna Tramontano et al. |
| MaxComp | Alignment of 3-D protein structures. By Georg Tuparev. |
| MaxHom | Pairwise and multiple sequence alignment and database sequence searches. By Chris Sander and Reinhard Schneider. |
| MaxSprout | Builds full coordinates from C(alpha) trace; Monte Carlo optimization of side chains. By Liisa Holm. |
| MaxTwist | Minimize side chains only, etc. By Chris Sander. |
| Micryfon | Conversion of protein coordinates from any format. By Arthur Lesk. There is also a set of molecular drawing programs by the same author. |

| | |
|---|---|
| nseqtool | Display of multiple sequence alignment coupled to 3-D graphics. By Raimund Schnobel. |
| PDB | Database of Protein Structures. By Brookhaven Labs |
| Pluto | B/w protein structure plots. By Sam Motherwell. |
| Polarh | Adds polar hydrogen atoms. By John Moult. |
| ProteinInfo | Collated protein analysis. By many authors. |
| Qpack | Assess residue packing in proteins. By Lydia Gregoret. |
| Quanta | 3D graphics, dynamics and stuff. By Polygen. |
| Quest | Find small molecules in database. By CCD. |
| Staden | Sequence analysis. By Rodger Staden. |
| Swissprot | Protein sequence data base. By Amos Bairoch et al. |
| WHATIF | Molecular modelling, database handling, drug design. Many options, you name it. By Gerrit Vriend. |

**Other software and databases available at EMBL**

| | |
|---|---|
| Discover | EM and MD. By Biosym, |
| Charmm | EM and MD. By Polygen. |
| Biograf | Graphics and model building and EM. By Biograf |
| Bioexplore | Graphics and model bulding. By Biostructure. |
| Frodo/Tom | Graphics and model bulding. By Alwyn Jones and Christian Cambilleau. |
| O | Graphics, model building. By Alwyn Jones. |
| Xplor | Molecular dynamics and refinement. By Axel Brünger, derived from Charmm. |
| UCSF Midas | - Second edition of the classic graphics program from UCSF |
| MaxTwist | Energy minimization in selected internal coordinates. By Chris Sander. |
| Mali/Prali | Multiple sequence alignment. By Martin Vingron. |
| Predict | Suite of secondary structure prediction programs. |

| | |
|---|---|
| PIR/PSQ | Protein sequence data base |
| EMBL | Database of nucleotide sequences |
| DSSP | Database of protein secondary structures |
| HSSP | Database of protein family alignments |
| CCD | Cambridge Crystallographic Database |

## Other Manuals available during the course

EMBL Software Catalogue
Delphi
Xplor
Composer
Scrutineer
WhatIf
Quanta
Insight
GCG

# BETASHEET

## FUNCTION

**Betasheet** makes two dimensional beta sheet diagrams of proteins. It also defines residue subsets for manipulation and colouring by the **Insight** 3-D graphics program.
Needs protein coordinates (Brookhaven format).

## AVAILABILITY

On the Vax-cluster **Betasheet** is made available by **Prepare proteins**; on the Unix-cluster by **Prepare_proteins**.

## DESCRIPTION

**Betasheet** produces two-dimensional diagrams of the beta sheets in a protein. The diagrams contain residue numbers and hydrogen bonding patterns separately for an entire sheet and for each face of the sheet, e.g. the solvent exposed and interior face.
The diagrams are useful for the analysis of beta sheets and their contacts with secondary structure elements and with solvent. The sheets and sheet faces can be displayed and colored selectively in 3-D by the graphics program **Insight**. Communication with **Insight** is via a file of residue subset definitions.

## AUTHORS

Shneior Lifson, Brigitte Altenberg and Chris Sander 1980 and 1990.

## EXAMPLE

$ betasheet **$pdb:4pti.brk**

$ prepare Insight

$ Insight

$ @4pti-sheets.x

### file name convention

input - proteinid.brk (Brookhaven format)

output - betasheet.x

proteinid-sheets.x

## OUTPUT

file betasheet.x contains:

```
 COMPND TRYPSIN INHIBITOR                                        4PTI    4
 NUMBER OF DOUBLE STRANDS FOR THIS PROTEIN    2

    JS                  strand pair number
    PAR                 parallel (+1) or antiparallel (-1) pair
    RES  AA             residue number and type
    TYP                 hydrogen bonding type
    DOT                 dot product between two  C(alfa)-C(beta)
    DCA DCB             C(alfa,i)-C(alfa,j) distance, same for C(beta)
    DNO DON NHNO NHNO N to O distance, O to N distance, Hbond angles (cos)
    BAAA                cosine of angle [C(beta) to C(alfa), C(alfa) to C(alfa)]
```

```
 JS PAR RES AA RES  AA TYP DOT DCA DCB   DNO DON NHNO NHNO   DNO DON NHNO NHNO

  1 -1  18 ILE  35 TYR  -1 0.5 5.1 4.3   3.1 2.8  1.0  1.0   6.1 5.3  0.8 -0.5
  1 -1  19 ILE  34 VAL  -3 0.2 4.8 6.4   6.9 7.1 -0.9 -1.0   4.5 6.4 -0.9  1.0
  1 -1  20 ARG  33 PHE  -1 0.9 5.6 6.0   3.0 2.8  1.0  1.0   6.7 5.3  0.7 -0.6
                                  etc.

     COMPND    TRYPSIN INHIBITOR

       0-0, 2-2 etc: number of NH or CO groups participating in the Hbond pair
    between residues I and J on adjacent  strands

           SHEET NUMBER    1

20          ||| TYR  35 2-2 ILE  18 |||
21          ||| VAL  34 0-0 ILE  19 |||
22          ||| PHE  33 2-2 ARG  20 0-0 LYS  46 |||
23          ||| THR  32 0-0 TYR  21 2-2 PHE  45 |||
24          ||| GLN  31 2-2 PHE  22 0-0 ASN  44 |||
25          ||| CYS  30 0-0 TYR  23 |||
26          ||| LEU  29 2-2 ASN  24 |||

  SIDE 1, SHEET NUMBER    1

20          ||| TYR  35 2-2 ILE  18 |||
22          ||| PHE  33 2-2 ARG  20 0-0 LYS  46 |||
24          ||| GLN  31 2-2 PHE  22 0-0 ASN  44 |||
                              etc
```

file 4pti-sheets.x contains **Insight** (Ps300) input:

```
GET PDB $PDB:4pti.BRK AS pti
ABBR ptisheet1ALL1pti:29-35,18-24,44-46
ABBR ptisheet1UPS1pti:35,33,31,29,18,20,22,24,46,44
ABBR ptisheet1DWN1pti:34,32,30,19,21,23,45
```

On the unix cluster 4pti-sheets.x contains **Insight** (IRIS) input:

```
get pdb /data/pdb/4pti.brk pti
define subset ptisheet1ALL1 pti:29-35,18-24,44-46
define subset ptisheet1UPS1 pti:35,33,31,29,18,20,22,24,46,44
define subset ptisheet1DWN1 pti:34,32,30,19,21,23,45
```

## RELATED PROGRAMS

**DSSP_TO_INSIGHT** also produces residue subsets for the graphic program **Insight**, for helices and strands.

## RESTRICTIONS

Only complete (up to C-beta) structures in Brookhaven format are valid input.

# BIO-GROMOS

## FUNCTION

**BioGromos** is an interactive graphics program for EM/MD on macromolecules.

## AVAILABILITY

**BioGromos** is available on the Iris Workstations.

## DESCRIPTION

**BioGromos** is a software tool that incorporates the **GROMOS** (Groningen Molecular Simulation) package. It allows one to intuitively create the required **GROMOS** files, eg for imposing distance restraints, and run the **GROMOS** programs.

## AUTHOR

**BioGromos** was developed by **BioStructure S.A.**, Strasbourg - France.



**GROMOS** was developed by W.F. van Gunsteren and H.J.C. Berendsen, at the University of Groningen, The Netherlands.

For help and advice see Johan Postma, Pieter Stouten or Christos Ouzounis.

## EXAMPLE

Here is a sample session with **BioGromos**:

```
(Opening the corresponding window on the Iris workstation, just respond
to the prompts of the program):

(a window-driven menu allows interactive use of BioGromos)

(see the manual and the on-line documentation for help)

(it is also recommended for first-time GROMOS users to read
 the GROMOS documentation before using BioGromos)
```

## OUTPUT

Coordinate files in PDB/BIO*(BioStructure)/Gromos formats.

## RELATED PROGRAMS

Other interfaces to the **GROMOS** core program are **PreGromos** by Johan Postma or the **GROMOS** menu in **WhatIf** by Gerrit Vriend.

Other EM/MD simulation programs include **Xplor** (on VAXcluster) and **Quanta/Charmm** (on Iris workstation) (documentation and manuals available).

## RESTRICTIONS

All general restrictions for EM/MD simulations apply, eg. lack of solvent and of careful electrostatics. Please consult the available User Guide.

## INPUT FILE

Input files of molecules should be in one of three formats (PDB/BIO*/Gromos). Click LOAD in the **BioGromos** window. For details, please consult the User Guide.

# COMPOSER

## FUNCTION

**COMPOSER** is a program that can build complete protein models from a protein sequence, by using 'homologous' protein(s) as a 'framework'.

## AVAILABILITY

**COMPOSER** is available on the VAX cluster. Please contact G. Vriend or C. Sander if you want to use **COMPOSER**. The program is designed to be used in batch.

## DESCRIPTION

**COMPOSER** is a series of four programs which run sequentially: The first generates a 'framework' of ca fragments either extracted from a single structure or an average of several similar structures superposed. The second aligns the sequence of these fragments with the sequence of the unknown sequence to be modelled. The third builds the unknown sequence where it aligns to a 'framework region'. The fourth searches for loops which may connect the framework regions together with the number of residues found in the unknown sequence.

## AUTHOR

**COMPOSER** was written by Mike Sutcliffe, Birkbeck College, UK.

## EXAMPLE

See full manual for examples (comp$man:new.man)

## INPUT FILE

Minimum input to **COMPOSER** is: - unknown sequence in PIR format (MODEL.SEQ) - at least one BRK format file for framework generation - instream file specifying framework regions from BRK file

## OUTPUT

The program creates intermediate files between each of the four stages and finally a full BRK format coordinate file (model.dat). Each intermediate file can be modified to select various modes and correct for incorrect automatic sequence alignment or loop selection. See manual.

## RESTRICTIONS

**COMPOSER** running in automatic mode will almost always produce a complete coordinate file, but unless you are modelling a sequence highly homologous that of a known structure, it may produce rubbish. Use graphics to check and be ready to think and intervene manually.

# CONAN

## FUNCTION

**Conan** is a program for protein **con**tact **an**alysis. **Conan** produces two-dimensional **contact maps** (generalisations of C-alpha distance plots). The concept of **contact masks** enables the user to select and color an interesting subset of all contacts.

## AVAILABILITY

**Conan** is available on the SUN3, SUN4 and the SPARC-station. To display the contact maps interactively you need to be in **SunTools.**

## DESCRIPTION

Some of the features of **conan** are **comparison of contact maps**; selection of different definitions of contacts; usage of masks to restrict the atoms, residues and secondary structure elements used for the calculation; display of atom-atom contacts as well as residue-residue contacts.

## AUTHORS

Chris Sander and Michael Scharf. Expierienced user: Alfonso Valencia.

## EXAMPLE

Here is a sample session with **conan.**

```
black% conan
==> conan > read-protein /data/pdb/1crn.brk
==> conan > calculate-contact-map
==> conan > set-display display -equal window
==> conan > write-contact-map -to 1ppt.bcm
==> conan > set-display display -equal postscript
==> conan > write-contact-map -to 1ppt.ps
==> conan > quit
black% lpr 1ppt.ps
black% conwin 1ppt.bcm &
```

If you don't like typing long commands try ^\TAB to complete a word and **Control-v** to show all possible completions.

## INPUT FILE

Generally, **conan** recognizes the data format of a file automatically if you use the **read filename** command.

A **view** is a one dimensional mask applicable to proteins to select a certain part of the protein :

```
>BEGIN-VIEW = sidechains
  VIEW
    atom      = N,C,CA,O
    structure = h
    brknumber = 10..200
    residue   = ala,gly
>END VIEW
```

A **mask** specifies the two contacting objects (atoms) :

```
>BEGIN-MASK = test-mask
  SELECT CONTACTS
```

```
            WITH ONE ATOM IN
              Chain      = A
              Structure  = E
              BRKNumber  =
              Residue    = D ,E, H, K, R
              Access     = 50%..100%
# means unequal
              Atom       # n,ca,c,o
          AND OTHER ATOM IN
            Structure    = H
        >END-MASK
```

## COMMAND LINE SUMMARY

Commands contain a verb and its object, e.g. read-protein. The most important verbs are **read, write, calculate, set, show, list, kill, change** and **help**. The objects are **protein, dssp, contact-map, view, mask, alignment, contact-param, io-prot-param, display-param** and **current**. Options usually start with a minus sign as in unix. **help-usage** shows the available commands and options.

## OUTPUT

You can get a **PostScript** file (the .ps file in the example) or a binary (.bcm) file that can be displayed with **conwin**.

## RELATED PROGRAMS

To display and query interactively a binary contact map use **conwin** under SunTools.

## RESTRICTIONS

Binary files can only be displayed on the workstations (SUN3 or SUN4/SPARC-Station) where they have been created.

Some PostScript files cause overflow errors on the small laserwriters. They can be printed on the PS40, however.

# CONVERTCOOR

## FUNCTION

**CONVERTCOOR** converts protein coordinates to standard **PDB** format.

## AVAILABILITY

**CONVERTCOOR** is made available on the VAX cluster by the command **Prepare proteins** followed by **CONVERTCOOR**.

## DESCRIPTION

**CONVERTCOOR** reads about ten different formats of protein structure coordinates, one line per atom, and outputs Brookhaven-PDB formatted data sets.

## AUTHOR

Chris Sander, 1987-1990.

## EXAMPLE

Here is a sample session with **CONVERTCOOR**.

```
|SIRIUS> prepare proteins

|SIRIUS> convertcoor       ! call up the program
 input format options:
 DIAMOND    D
 FREIBURG   F
 EDMONTON   E
 POLARH     P
 WAH        W
 RDI        R
 SHIFTEDBROOK  S
 BROOK      B
 Three Reals  3

                    Enter one letter !
                    Default: D
                    [CRdefault]:
                    echo: D
 output options:
 renumber residues    R
 no content changes   N

                    Enter one letter !
                    Default: N
                    [CRdefault]:
                    echo: N

 Format used is LDIAMOND
                  Enter input file name:
myprotein.dia

          Please enter output file name:
myprotein.pdb
```

```
--- interpret atom line:
     2.296    -9.636    18.253     1.000    1    1    1          GLY   1   N
--- interpret atom line:
     1.470    -9.017    17.255     1.000    1    1    2          GLY   1   CA
--- interpret atom line:
     0.448    -9.983    16.703     1.000    1    1    3          GLY   1   C
--- interpret atom line:
     0.208   -11.066    17.345     1.000    1    1    4          GLY   1   O
--- interpret atom line:
    -0.170    -9.672    15.624     1.000    1    2    5          PRO   2   N
    etc...

          19  lines processed.
          16  atoms converted.
```

## INPUT FORMATS

```
(D) DIAMOND;FORMAT(4F10.5,I5,1X,A4,I5,10X,A3,I4,3X,A3)
x...;....1....;....2....;....3....;....4....;....5....;....6....;....7....;....8
 -30.33598   0.12814   3.40177              1    2    5          LYS   2   N

(F) FREIBURG;FORMAT(9X,3F10.4,2X,2F5.2,1X,A3,2X,A4,1X,A3)

(E) EDMONTON;FORMAT(4X,A1,A3,A3,2X,5F10.4)

(P) POLARH;FORMAT(A1,A3,1X,A3,2X,5F10.5)
x...;....1....;....2....;....3....;....4....;....5....;....6....;....7....;....8
R  1 N       26.46500  27.45200  -2.49000

(W) WAH;  FORMAT(3X,A3,1X,A4,A3,2X,5F10.4)

(R) RDI;  FORMAT(3F10.5,33X,A3,1X,A4,3X,A3)

(S) SHIFTEDBROOK;  FORMAT(6X,5X,1X,A3,2X,A3,2X,A4,4X,3F8.3,2F6.2)
x...;....1....;....2....;....3....;....4....;....5....;....6....;....7....;....8
ATOM       1 N    ALA     1      24.714  16.360  23.605  1.00  1.00
             **
(B) BROOK;  FORMAT (6X,5X,2X,A3,1X,A3,2X,A4,4X,3F8.3,2F6.2)
x...;....1....;....2....;....3....;....4....;....5....;....6....;....7..
ATOM       1  N   ALA     1      24.714  16.360  23.605  1.00  1.00

(3) Three Reals; Almost free FORMAT - three real numbers first in each line
    which are read by READ(LINE,*,END90) X,Y,Z
```

## OUTPUT

PDB coordinates.

```
|SIRIUS> typ myprotein.pdb       ! look at output file

ATOM       1  N   GLY     1       2.296  -9.636  18.253  1.00  1.00    0
ATOM       2  CA  GLY     1       1.470  -9.017  17.255  1.00  1.00    0
ATOM       3  C   GLY     1       0.448  -9.983  16.703  1.00  1.00    0
```

```
ATOM      4  O   GLY     1       0.208 -11.066  17.345  1.00  1.00    0
ATOM      5  N   PRO     2      -0.170  -9.672  15.624  1.00  1.00    0
etc...
TER
END
```

## RELATED PROGRAMS

**Micryfon** also reformats protein coordinates. It can read free formatted input so try it if CONVERTCOOR does not read your format.

## RESTRICTIONS

**CONVERTCOOR** only understands a limited set of protein coordinate formats.

# DGEOM

## FUNCTION

**Dgeom** is a Distance Geometry program for molecular model-building, docking and receptor modeling, and conformational analysis.

## AVAILABILITY

**Dgeom** is available on the SGI 4D series (Jo) at /usr/local/distrib/dgeom/590sgrw.

## DESCRIPTION

**Dgeom** is a program for distance geometry calculations; utility programs distributed with Dgeom: **Compare** and **Dbond**.

## AUTHOR



## EXAMPLE

Here are three examples that show how to use **Dgeom**

```
(Run the program at /usr/local/distrib/dgeom/590sgrw/source/dgeom):

$ dgeom < fit.com (or ANY .com file: see 'examples' directory)
$ dgeom < fit.com > out.dat (re-direct screen output into a file)
$ nice dgeom < fit.com & (run dgeom in the background)
```

## INPUT FILE

Input to **Dgeom** is a PDB-format coordinate file. All input specified in the .com file.

## OUTPUT

The output is a number of structure file(s) in PDB (Brookhaven) format.

## RELATED PROGRAMS

A graphics program such as **Insight** is required, to examine the final structure file(s) visually. Utility programs include **Compare** which reads a series of PDB-format files generated by **Dgeom**, calculates the rms deviation values for all pairwise combinations of structures, and writes out the rms correlation matrix, for further processing (cluster analysis). Other distance geometry programs: **DisGeo**, **Disman**, **Constrictor**.

## RESTRICTIONS

Please consult /usr/local/distrib/dgeom/590sgrw/help/dgeom.mem. The size limitations can be set in the file dgeom.siz (see 'source' directory). The input file of the constraints must be in the given standard format (see Christos Ouzounis for input conversion programs for distance constraints).

## FUNCTION

**DSSP** defines protein secondary structure given a set of atomic coordinates.

## AVAILABILITY

**DSSP** is made available on the VAX cluster by the command **Prepare Proteins**.

## DESCRIPTION

**DSSP** assigns secondary structure elements and calculates solvent exposure of proteins given atomic coordinates in Brookhaven Protein Data Bank format.

**DSSP** is not a structure prediction program.

## AUTHOR

Wolfgang Kabsch and Chris Sander, MPI MF, Heidelberg (1983). Reference: Kabsch W., Sander C. (1983); Biopolymers 22:2577-2637.

See Peter Rice or Chris Sander for help.

## EXAMPLE

Here is a sample session with **DSSP** on the VAX.

```
$ Prepare proteins
$ DSSP $Pdb:1Ppt.Brk


 !!! RESIDUE TYR   36   HAS  9 INSTEAD OF EXPECTED   8 SIDECHAIN ATOMS.
     LAST SIDECHAIN ATOM NAME IS  OXT
     CALCULATED SOLVENT ACCESSIBILITY INCLUDES EXTRA ATOMS !!!

INPUTCOORDINATES DONE         36
FLAGSSBONDS DONE
FLAGCHIRALITY DONE
FLAGHYDROGENBONDS DONE
FLAGBRIDGE DONE
FLAGTURN DONE
FLAGACCESS DONE
PRINTOUT DONE

 your output file is 1PPT.DSSP

$
```

## INPUT FILE

The input file for **DSSP** is a coordinate file in PDB format. To convert other formats to PDB, use programs **ConvertCoor** or **ConvertToPdb**.

## OUTPUT

The output from **DSSP** in file protein.DSSP contains secondary structure assignments and other information, with one line per residue. The following is a simplified extract from the output file in the example.

## RELATED PROGRAMS

**ConvertSeq** converts **DSSP** output to a number of standard sequence formats for use with **FASTA** and **Predict**.

**Predict**, when applied to a **DSSP** output file, will assess the prediction accuracy of various secondary structure prediction methods.

**ConvertCoor** and **ConvertToPdb** convert coordinates in DIAMOND or other formats to the PDB format used by **DSSP**. Contact Chris Sander for details. **DSSPtoInsight** interfaces DSSP output for selective display of secondary structure elements with the INSIGHT graphics program.

Use the **WhatIf** options SHOHST, COLOUR, COLHST and SHOALL to list the DSSP secondary structure assignments, and to display the structure coloured as function of secondary structure.

## RESTRICTIONS

The maximum number of residues is 1500. If this number is exceeded, please contact Peter Rice or Chris Sander.

## ALGORITHM

The values for solvent exposure may not mean what you think!

(a) effects leading to larger than expected values: solvent exposure calculation ignores unusual residues, like ACE, or residues with incomplete backbone, like ALA 1 of data set 1CPA. it also ignores HETATOMS, like a heme or metal ligands. Also, side chains may be incomplete (an error message is written).

(b) effects leading to smaller than expected values: if you apply this program to protein data bank data sets containing oligomers, solvent exposure is for the entire assembly, not for the monomer. Also, atom OXT of c-terminal residues is treated like a side chain atom if it is listed as part of the last residue. also, peptide substrates, when listed as atoms rather than hetatoms, are treated as part of the protein, e.g. residues 499 s and 500 s in 1CPA.

Unknown or unusual residues are named X on output and are not checked for standard number of sidechain atoms. All explicit water molecules, like other hetatoms, are ignored.

# DSSPTOINSIGHT

## FUNCTION

Defines residue subsets for selective display and manipulation of secondary structure elements in 3-D using the molecular graphics program **Insight.**

## AVAILABILITY

On the Vax-cluster **DSSPtoInsight** is made available by Prepare proteins; on Unix-cluster by Prepare_proteins.

## DESCRIPTION

**DSSPtoInsight** produces subset definitions for the graphic program **INSIGHT.**

## AUTHOR

Brigitte Altenberg 1990.

## EXAMPLE

$ DSSPtoInsight $pdb:4pti.dssp

(Then there is some screen output, and several output files are created)

file name convention:

input : proteinid.dssp (Must be in DSSP format)

output : proteinid-dssp.x (Input file for **Insight**)

## OUTPUT

file 4pti-dssp.x for INSIGHT1 on PS390 contains: (In **Insight** use '@4pti-dssp.x' to execute this file)

get pdb $PDB:4pti.brk pti

**insight 1**: get pdb $pdb:5pti.brk pti

ABBR ptiH1 pti:3-6

ABBR ptiE1 pti:18-24

ABBR ptiE2 pti:29-35

ABBR ptiH2 pti:48-55

ABBR ptiHall1pti:3-6,48-55

ABBR ptiEall1pti:18-24,29-35

ABBR ptiTall1pti:25-28

ABBR ptiSall1pti:7,13,37-39,42-43,46-47

**insight 2**:

get pdb /data/pdb/5pti.brk pti

define subset ptiH1 pti:3-6

define subset ptiE1 pti:18-24

define subset ptiE2 pti:29-35

define subset ptiH2 pti:48-55

define subset ptiHall1 pti:3-6,48-55

define subset ptiEall1 pti:18-24,29-35

define subset ptiTall1 pti:25-28

define subset ptiSall1 pti:7,13,37-39,42-43,46-47

File 4pti-dssp.x for INSIGHT2 on the Iris workstation contains: (In **Insight** use 'source 4pti-dssp.x' to execute this file)

get pdb /data/pdb/4pti.brk pti

define subset ptidsspH1 pti:48-55

define subset ptidsspB1 pti:45

define subset ptidsspE1 pti:18-24,29-35

define subset ptidsspG1 pti:3-6

define subset ptidsspT1 pti:25-28

define subset ptidsspS1 pti:7,13,37-39,42-43,46-47,56

**RELATED PROGRAMS**

**BETASHEET** also produces subset definitions for the graphics program **INSIGHT**.

# FindSite

## FUNCTION

**FindSite** looks for positions where a metal binding site could be engineered into a protein.

## AVAILABILITY

**FindSite** is available on the VAX cluster. Contact C. Sander if you want to use it.

## DESCRIPTION

**Findsite** scans the target protein coordinates for sites that satisfy certain distance constraints angle constraints and that are similar to the metal binding site in the query protein.

**FindSite** reads an input file specifying the names of target protein and the query protein, the number of residues involved in metal binding and the atom number of the metal in the Brookhaven file. It expects a **.BRK**, **.DSSP** and **.HSSP** file on **$PDB** or in your own directory. **FindSite** creates a list of possible mutation sites.

## AUTHOR

Adam Godzik and Chris Sander, 1990.

## EXAMPLE

run $1:[godzik.public]findsite

sample input file $1:[godzik.public]findsite.dat

## OUTPUT

The output from **FindSite** is the file **findsite.out**, which gives a list of possible mutation sites in the order of increasing RMS deviation from the original site.

## RELATED PROGRAMS

**FindSite** is the only exportable program from a series (**KONTACTS, MATRIX, SPHERE, SEARCH**). All of these are simple Fortran programs which analyse protein structure.

## RESTRICTIONS

Input file with the name **FINDSITE.DAT** must be present in you local directory. Template and target **.BRK**, **.DSSP** and **.HSSP** files must be present either in **$PDB** or in your local directory.

## ALGORITHM

The program successively tests sets of four residues according to distance and angle criteria.

## INPUT FILE

Sample input file (FindSite.Dat) is listed below. Comments in input line are optional.

```
   (A test example is: $PDB:1REI.BRK, $PDB:1REI.DSSP and $PDB:1REI.HSSP,
    versus $PDB:1PCY.BRK, $PDB:1PCY.DSSP and $PDB:1PCY.HSSP.

1rei                    !    name of target protein
1pcy                    !    name of template (query) protein
4                       !    number of residues constituting the binding site
```

```
37 84 87 92              !    (max 5) and their .brk numbers
740                      !    .brk number of metal (Cu in this case)
5 1.0                    !    ntest (for debug purposes: set to 0) and cutoff -
                         !    pair of atoms with distances larger then cutoff
                         !    will not be considered.
```

## FUNCTION

**GCG** is a software package for sequence analysis in molecular biology.

## AVAILABILITY

**GCG** is made available on the VAXcluster by the command **prep gcg**. Local extensions are made available on the VAXcluster by the command **prep gcgembl**.

## DESCRIPTION

**GCG** is an integrated package of sequence analysis. Programs available for sequence comparison, database searches, multiple sequence alignments, secondary structure prediction, analysis of protein sequences, mapping, fragment assembly, pattern recognition in DNA and protein sequences and sequence editing/conversion. Also file and graphics utilities.

## AUTHOR

**GCG** was written by the Genetics Computer Group (1982-1990), directed by J.Devereux. For more information contact Peter Rice.

## EXAMPLE

Here is a sample session with **GCG**:

```
$ GCG
 $ Genmanual (on-line documentation by-category)
 $ Genhelp (on-line documentation by-program)
 $ (program-name)

(For further details see the GCG manual)

$
```

## OUTPUT

Sequence files; all databases made available by **gcg**. Postscript files for graphical output.

## RELATED PROGRAMS

Other graphics programs/packages in the same class are **PSQ/NAQ** and **Staden**. Accessible on the VAXcluster with the command **prep**.

## RESTRICTIONS

Please refer to the Program Manual.

## INPUT FILE

Generallly, sequence data (formatted to GCG style using the program **Reformat**).

## FUNCTION

**GROMOS** is a set of programs to simulate the dynamical trajectory of molecular systems in time, primarily aimed at macromolecules.

## AVAILABILITY

**GROMOS** is available on the VAX cluster (interfaced by **Whatif** and **Pregromos**) and on the Iris Workstations (interfaced by **BioGromos**).

## DESCRIPTION

**GROMOS** is a complete program package for Energy Minimizations (EM) and Molecular Dynamics (MD). It has its main use in the simulation of atomic motion of macromolecules. It solves the equations of motion numerically. Simulations of solutions, of crystals and of molecules in vacuo can be carried out. **GROMOS** has been successfully used e.g. for the calculation of free energies and for the refinement of crystal structures. The program package also contains programs to build the system of interest from prefab building blocks and a set of analysis programs to reduce the vast amounts of data produced by the simulations.

## AUTHOR

**GROMOS** was developed by Wilfred van Gusteren (now at ETH, Zuerich, Switzerland) and Herman Berendsen at the University of Groningen, The Netherlands.
The **GROMOS** package can be obtained from **BIOMOS BV**, Groningen, The Netherlands.

## EXAMPLE

The **GROMOS** programs are normally accessed through an interface from other programs (**Whatif**, **Pregromos** or **BioGromos**). These interfacing programs generate the necessary files substituting defaults for most parameters, unless the user specifies differently. They also generate script or command files to run the **GROMOS** programs. The script or command files can be executed immediately from the interfacing programs or later as standalone jobs. Standard input, script and command files are also available. Here follows an example of a VAX command file. The command file itself and the files it needs for proper execution can be found in $course:[vriend.gromos]):

```
$! Procedure tri-md.com
$! PFW Stouten  26-SEP-1990 15:31:12
$! Only input files FOR005, FOR020 and FOR021 must be present.
$!
$  prepare gromos
$  assign/user tri-run-parameters.md  FOR005
$  assign/user tri-186.mtp            FOR020
$  assign/user tri-in.xyz             FOR021
$  assign/user tri-out.xyz            FOR031
$! optional restraint files (not present):
$! assign/user tri-x-re.xyz           FOR024
$! assign/user tri-x-re.seq           FOR025
$! assign/user tri-dist-re.pairs      FOR026
$! assign/user tri-tors-re.quart      FOR028
$  run gromos$exe:promdl
$!
```

## INPUT FILES

Input files for the simulation programs are coordinates, molecular topology and a file describing the exact simulation procedure and conditions.

## OUTPUT FILES

The output files of the EM and MD programs contain coordinates, velocities and/or energies. The data contained in these files serve as input for the analysis programs and for picture display programs. A synopsis of results is written to standard output.
Logicals in the script or command files define the actual file names. Default file names on the VAX cluster are FOR0??.DAT in the current directory.

## RELATED PROGRAMS

Other EM/MD simulation programs include **Xplor**, **Discover** and **MaxTwist** (on the VAX cluster) and **Quanta/Charmm** (on the Iris Workstations).
For graphical display of minimized structures or MD runs (as a continuous film) **BioGromos** and **Whatif** can be used immediately and many other display programs (**Hydra**, **Insight** etc.) after the conversion of file formats. Use **ConvertCoor** to convert coordinates from Gromos to PDB format.

## RESTRICTIONS

The MD simulations can produce immense files. Simulations with large proteins and/or many water molecules can be quite time consuming. No **GROMOS** program on the VAX closes its input files after reading them. These files are therefore locked and cannot be accessed by other processes until the program terminates.

## CONSIDERATIONS

It is highly recommended for first-time **GROMOS** users to read the **GROMOS** documentation before using any of the programs.
If any of the programs reports errors (or crashes) while reading formatted input files, check the formats carefully. In case of cryptic error messages, consult the documentation.
For help and advice see Pieter Stouten or Johan Postma.

## FUNCTION

**HSSPtoinsight** is a utility program to convert information about conserved residues in a protein structure to an **Insight** script. The script file colours residues according to their variability values.

## AVAILABILITY

**HSSPtoinsight** is available on the VAX after typing the command **prepare proteins**. On the eunuchs cluster type **Prepare_proteins**.

## DESCRIPTION

**HSSPtoinsight** reads an HSSP-file and produces an **Insight** command file. The user has to specify two numbers, which influence the colouring of the residues. Residues with a HSSP-variability value smaller than the first number given are coloured red (conserved residues). If the variability value is in the range between the first and the second number the residues are coloured yellow. All other residues are green (high variability).

## AUTHOR

Reinhard Schneider, 1990.

## EXAMPLE

Here is a sample session with **HSSPtoinsight**.
(type the following responses to the prompts from **HSSPtoinsight**):

```
$pdb:1crn.hssp   (on UNIX machines: /data/hssp/1crn.hssp)
insight.com      (Insight command file)
10 25            (variability cutoffs; values given are the defaults)

$ prepare Insight (on UNIX machines prepare_Insight )
$ Insight
   source insight.com (IRIS)
   @ insight.com (PS300)
```

## INPUT FILE

Input to **HSSPtoinsight** is an HSSP-file; eg. $PDB:1CRN.HSSP, or /data/hssp/1crn.hssp.

## OUTPUT

The output is a file which contains Insight specific commands (get coordinate-file , define subset , colour subset)

## RELATED PROGRAMS

In **WhatIf** use the following commands to get a similar result:

**GETHSP**,

**COLOUR**,

**COLPRP**.

## RESTRICTIONS

Continuous HUE according to variability is not possible.

## FUNCTION

**Insight** is an interactive program to aid molecular modellers, drug designers and crystallographers.

## AVAILABILITY

**Insight** is made available on the VAX cluster by the command **PREPARE INSIGHT**. On the UNIX cluster type **prepare_insight**.

## DESCRIPTION

**Insight** is an integrated package of database handlers, protein analysis tools and high quality graphics.

## AUTHOR

Anna Tramontano and Henry Dehringer 1984 and later.

## EXAMPLE

Here is a sample session with **Insight**.
(Working on a terminal next to the POLLUX PS390, type the following responses to the prompts from **Insight**):

```
Sample session with Insight on the VAX

$ prepare insight
$ insight
INSIGHT> get pdb $pdb:1crn.brk as crambin
INSIGHT> color red ::O*
INSIGHT> color blue ::N*
INSIGHT> color green ::C*
INSIGHT> display only ::CA,C,N,O
INSIGHT> exit
```

(You now have a picture on the screen which can be rotated and translated. For further details see the **Insight** manual)

On an IRIS workstations, click **Insight**. Use the drop down menus to execute these commands: get molecules; pdb; 1crn.brk; color by atom type; ribbon.

## INPUT FILE

Input to **Insight** is typically PDB-format coordinate files.

## OUTPUT

Everything displayed on the PS390 screen can be plotted. Everything displayed on the terminal screen can be printed.

## RELATED PROGRAMS

Other graphics programs/packages in the same class are **HYDRA**, **FRODO**, **Whatlf**, **Midas**, **O**, **BioExplore** and **BioGraf**.

## RESTRICTIONS

None, except for bugs.

## FUNCTION

**MaxComp** compares the 3-D structures of two proteins and identifies similar substructures. It performs an automatic superposition of two homologous proteins.

## AVAILABILITY

**MaxComp** is available on the UNIX cluster (only on hosts black, iris and jo). Type **Prepare_proteins** and **MaxComp**.

## DESCRIPTION

**MaxComp** reads default constants and input data for two proteins in dssp (or brk) format, and creates a list of pairs with maximum structural similarity.

## AUTHOR

Georg Tuparev and Chris Sander.

## EXAMPLE

Here is a sample session with **MaxComp**.

```
1. Login to black, jo or iris.
2. Interactive use:
   Prepare_proteins
   MaxComp
3. Non interactive use:
   $maxcomp filename [-d default_file_name]
4. Get some short help:
   MaxComp -h

Self-explanatory examples for 'filename' and 'default_filename' are
/felix2/pub/lib/maxcomp_input.data and /felix2/bub/lib/maxcomp_defaults.data
```

## INPUT FILE

- Files in brk or/and dssp format

- Command file(s). (See example for more)

command line options

-h : Print a short help information.

-i : Interactive execution

-d file : New default values

file : File including input information and defaults.

local data files

If no -d command line option is given, **maxcomp** automaticaly reads the content of /felix/pup/lib/maxcomp_defaults.data.

If the program reports errors while reading the input files, check the input format carefully.

## OUTPUT

1. *.res file : Contains the pairs that were found and information about secondary structure, water accessibility, CA-CA distances, overall and local RMS deviations, as well as an echo of the default parameters and file names.

2. Optionally one can obtain two files (*_brk1.brk and *_brk2.brk) in **.BRK** format, containing atomic coordinates data for the identified similar substructures.

## RELATED PROGRAMS

In **Whatlf**, use **SUPPOS** and **MOTIV** to find similar substructures. Use the **3SSP** menu to find all similar substructures in the whole structure database.

## RESTRICTIONS

1. The program uses C-alpha atoms only;

2. It is currently impossible to run the program on only parts of molecules.

3. CPU times can be up to 30 minutes to 1 hour.

## ALGORITHM

The alignment algorithm, used in **MaxComp** is described in [T.F.Smith and M.S.Waterman, J.Mol.Biol.(1981),147,195-197]. For similarity measurements the RMS value is used. (For calculating the RMS values see [W.Kabsch, Acta Cryst. (1978),A34,827-828]).

## FUNCTION

**MAXHOM** finds the best alignment(s) between one or more sequences

## AVAILABILITY

**MAXHOM** is available on the VAX and UNIX machines. Please ask Chris Sander or Reinhard Schneider if you want to use it.

## DESCRIPTION

**MAXHOM** uses the dynamic sequence alignment algorithm of Smith and Waterman; local sequence similarity is normally given by the 20 by 20 matrix of amino acid similarities of McLachlan scaled to a minimum value of **smin** (usually negative) and a maximum value **smax** of 1.0 (for the top single residue identity). Maximum length of a deletion is **maxdel** (e.g. 10) residues, an initial gap costs **gap opening penalty** (e.g. 3.0 similarity units) and gap elongation costs **gap elongation penalty** (e.g. 0.1 units) per residue. This program generates a multiple sequence alignment (**HSSP**) file and can do a superposition of the aligned sequences in 3-D (if the structures of both proteins are known).

## AUTHOR

Chris Sander and Reinhard Schneider 1984-1991.

## EXAMPLE

Here is a sample session with **MAXHOM** on the VAX.

```
$ MAXHOM
VMS                          (UNIX or VMS machine)
5                            (NBEST per protein)
300                          (NGLOB all proteins)
-0.7,1.0                     (MIN,SMAX  worst and best amino acid similarity)
10                           (MAXDEL)
3.0                          (gap opening penalty)
0.1                          (gap elongation penalty)
ALIGNMENT.X                  (outputfile)
TRACE.X                      (interface file for postscript output)
METRIC-ADM.DATA              (metric-file)
NO                           (newmetric file)
NO                           (compare 3-D-struc: YES/NO)
YES                          (HSSP: NO/YES)
formula                      (threshold criterion)
$PDB:4HHB.DSSP               (test sequence)
$PDB:1MBS.DSSP               (comparison sequence)
METRIC-GCG.DATA              (metric for HSSP-variability)
SWISSPROT$RELEASE:RELNOTES.DOC     (release notes SWISSPROT)
```

## INPUT FILE

Input to **MAXHOM** are a test sequence, a comparison sequence (list of sequences) and 20*20 matrices of amino acid similarities.

## OUTPUT

The most important output files are: alignment.x (each pair alignment listed separately), strip.x (collection of alignments relative to the absolute position in the master sequence), protein.HSSP (multiple alignment with secondary structure information, variability, and sequence profile)

## RESTRICTIONS

VAX: the user needs a page file quota of at least 39441

## FUNCTION

**MaxSprout** builds backbone coordinates given C-alpha coordinates, and sidechain coordinates given backbone coordinates.

## AVAILABILITY

**MaxSprout** is available on the VAX cluster.

## DESCRIPTION

**MaxSprout** constructs a backbone by searching the database of known structures for appropriate fragments, using distance criteria. It builds side chains by using a set of preferred side chain rotamers. Internal packing is limited by a Monte Carlo procedure with simulated annealing.

MaxSprout reads PDB files as input. Output coordinates are also in PDB format

## AUTHOR

Liisa Holm and Chris Sander 1987-1990.

## EXAMPLE

Here is a sample session with **MaxSprout.**

logicals used by the programs:

$ define csdata $2:[sander.l.data]

$ define rotlib $2:[holm.rotamerlibrary.rotlib]

convert PDB files for backbone builder (xxxx.main and xxxx.side):

$run $2:[holm.export.sprout-90]readbrk

build backbones:

$run $c:[holm.export.sprout-90]runmaxd

build sidechains:

```
$run $c:[holm.export.sprout-90]newbuild
an example command file for building the backbone from ubiquitin
(1ubq) CA coordinates is in $c:[holm.export.sprout-90]example1.com

  enter readdirectory (e.g. cs:) *^<return>
  enter writedirectory (e.g. work$disk:) <return>
  enter filename with extension [.brk] (e.g. 1sn3.brkmod) 1ubq.brk
  ///
  enter name of testprotein (e.g. 1sn3)
        76 residues in test-protein
  enter name of database protein names file (e.g. $2:[SANDER.L]dglp.list)
  ////
  use default parameters [Y]/N ? Y
  ///
  gap moved from          76 to          77
```

an example command file for building sidechains to ubiquitin
(1ubq) is in $c:[holm.export.sprout-90]example2.com

```
enter brk file name protein to build sidechains
1fdx.brk
///
bestemin for core    115.639
core done
enter brk file name protein to build sidechains <return>
```

## OUTPUT

output from runmaxd (create backbone) is in files:

1ubq.brkmod (backbone+CB coordinates)

1ubq.stat (some statistics on fragment lengths)

1ubq.log (log of which fragments are used in the reconstructed model)

frag.pdb (coordinates of all fragments fetched)

output from newbuild (build side chains) is in files:

1ubq_sc.brk (backbone with sidechain coordinates added)

## RELATED PROGRAMS

**WhatIf** does roughly the same with the **CATOAL** option in the **DGLOOP** menu. Also **Insight**, **FRODO**, and **O** have similar options.

## RESTRICTIONS

Non-standard amino acid names in the input file can create problems.

Do not run newbuild of poly-alanine.

## INPUT FILE

Runmaxd and newbuild read input coordinates in PDB format

## ALGORITHM

The algorithms are described in Holm&Sander, JMB, 1991.

## LOCAL DATA FILES

The fragment database for backbone construction is defined by logical **csdata**. The rotamer library is defined by logical **rotlib**.

## FUNCTION

**MaxTwist** performs energy minimization in selected internal degress of freedom. Any subset of degrees of freedom can be frozen or set free.

## AVAILABILITY

**MaxTwist** is made available on the VAX cluster by the command **Prepare proteins.**

## DESCRIPTION

**MaxTwist** works in terms of bond lengths, bond angles dihedral angles and torsion angles. A convenient input language allows complete freedom in choosing any subset of the 3N-6 degrees of freedom in a protein. At one extreme, there can be three degrees of freedom per atom, equivalent in principle to cartesian coordinate minimization. At the other extreme, a few side chain torsion angles can be optimized or a local zone can be regularized in terms of bond lengths etc. Here, only a simple example is given.

## AUTHORS

Chris Sander, 1981-1990. Original core code from Weizmann Institute via Peter Stern.

## EXAMPLE

Here is a sample session with **MaxTwist.**

```
( answer all questions in response to prompts ):

  prepare proteins
  maxtwist
  prepare proteins
  minsidechains  $pdb:1ppt.brk

( this will run on screen and produce output on files, see  OUTPUT  )
( preferable, a batch job should be run. This is done by putting
  the above lines into a file twistjob.com, starting each line with a $,
  and typing, e.g.: )

  submit twistjob.com /queue=soon /noprint /notify

( here is an example of interactive minimization )

 < watch this space >
```

## OUTPUT

Results are initial and final coordinates, an energy log and information about intital and final geometry and energies.

```
TWIST-COOR-INI.X < initial coordinates
TWIST-COOR-MID.X        < intermediate coordinates
TWIST-COOR-FIN.X        < final coordinates
TWIST-LOG.X             < energy and clash log
TWIST-OUTPUT.X< intial and final state information
TWIST-INSIGHT.COM (soon)< command file to view coordinates in INSIGHT
```

## RELATED PROGRAMS

**Gromos, Charmm, Discover** all do energy minimization and molecular dynamics in cartesian coordinates. Their advantage is generality. However, freezing parts of the proteins molecule selectively can be laborious or impossible. **MaxSprout** can explore conformational space by a Monte Carlo procedure where mere energy minimizers would end up in a local minimum. The recently developed program **BKS** by R. Abagyan and colleagues has functionallity similar to **Max Twist**.

## RESTRICTIONS

**MaxTwist** currently has only very simple electrostatics and no solvent terms. The minimization algorithm will only find the next nearest local minimum.

## INPUT FILES

Background data input to **MaxTwist** are residue libraries and definitions of interatomic potentials.

# MICRYFON

## FUNCTION

**MICRYFON** converts protein coordinates to two standard formats.

## AVAILABILITY

**MICRYFON** is made available on the VAX cluster by the command **Prepare proteins** followed by **MICRYFON**.

## DESCRIPTION

**MICRYFON** reads free format protein structure coordinates, one line per atom, and outputs Brookhaven-PDB and Diamond formatted data sets. "May be I Can Read Your Format Or Not".

## AUTHOR

Arthur M. Lesk, 1986.

## EXAMPLE

Here is a sample session with **MICRYFON**.

```
|SIRIUS> prepare proteins

  (first, echo the input file)

|SIRIUS> type myprotein.data
test data set in some format
1.0 5.0 9.5
various remarks ***
      1  N   GLY     1       2.296  -9.636  18.253  1.00  0.00        65
      2  CA  GLY     1       1.470  -9.017  17.255  1.00  0.00        66
      3  C   GLY     1        .448  -9.983  16.703  1.00  0.00        67
      4  O   GLY     1        .208 -11.066  17.345  1.00  0.00        68
      5  N   PRO     2       -.170  -9.672  15.624  1.00  0.00        69
      etc.

  (next, call up the program and specify an inputfile and two output files)

|SIRIUS> micryfon  myprotein.data  myprotein.brk  myprotein.dia

HISTOGRAM OF TYPES OF LINES FOUND ...

              TYPE                      AT LEAST
              CODE       NUMBER      THREE REALS

         1         0          2          F
         2        63          1          T
         3    524137         16          T

EXAMPLES OF DIFFERENT APPARENT TYPES OF ATOM COORDINATE LINES ...

TYPE            EXAMPLE

 3       1  N   GLY     1       2.296  -9.636  18.253  1.00  0.00        65
```

```
3       2  CA  GLY     1       1.470  -9.017  17.255  1.00  0.00        66
3       3  C   GLY     1        .448  -9.983  16.703  1.00  0.00        67
3       4  O   GLY     1        .208 -11.066  17.345  1.00  0.00        68
3       5  N   PRO     2       -.170  -9.672  15.624  1.00  0.00        69
etc.
```

## OUTPUT

```
PDB coordinates.
|SIRIUS> type myprotein.brk
test data set in some format
1.0 5.0 9.5
various remarks ***
ATOM      1  N   GLY     1       2.296  -9.636  18.253          1.00
ATOM      2  CA  GLY     1       1.470  -9.017  17.255          1.00
ATOM      3  C   GLY     1       0.448  -9.983  16.703          1.00
ATOM      4  O   GLY     1       0.208 -11.066  17.345          1.00
ATOM      5  N   PRO     2      -0.170  -9.672  15.624          1.00
etc.
DIAMOND coordinates.
|SIRIUS> type myprotein.dia
test data set in some format
1.0 5.0 9.5
various remarks ***
   2.296    -9.636    18.253     1.000    1    1    1          GLY   1   N
   1.470    -9.017    17.255     1.000    1    1    2          GLY   1   CA
   0.448    -9.983    16.703     1.000    1    1    3          GLY   1   C
   0.208   -11.066    17.345     1.000    1    1    4          GLY   1   O
  -0.170    -9.672    15.624     1.000    1    2    5          PRO   2   N
  .........
```

**RELATED PROGRAMS**

**ConvertCoor** also reformats protein coordinates. It is not as general but may be more stable on standard formats.

**RESTRICTIONS**

**MICRYFON** attempts to interpret an atom line, but may fail to do so in strange cases.

## FUNCTION

**nseqtool** is a colour browser for multiple aligned sequences combined with 3D view of a backbone.

## AVAILABILITY

**nseqtool** runs under UNIX on a SUN-4 or SPARC workstation. It is made available by typing **prepare_proteins**.

## DESCRIPTION

The display of a set of aligned sequences is combined with a stick representation of a protein backbone. The user may select one protein out of the alignment whose sequential and structural views will get intimately coupled. The interactive graphical environment provides for direct control of objects: Picking and highlighting residues, scrolling the sequence display, permuting the colour grouping, rotation and scaling are performed with a mouse click. The program is also useful for the analysis of multiple sequence alignments in colour.

## AUTHOR

Raimund Schnobel and Arthur Lesk 1989-1991.

## EXAMPLE

Here is a sample session with **nseqtool**.

```
Suntools
(open a shell window)
nseqtool
CR /home/schnobel/projects/align/files/cd4
BRK 2rhe
ATTACH 1
QUIT

(This will read the alignment from a file supplied by the user and
the coordinates from BROOKHAVEN file 2RHE.BRK. The stick display is
attached to the first sequence.)
(Note that the upper-case commands should be typed in the command field
near the bottom of the screen. Click the left mouse to refresh the picture.)
```

## INPUT FILE

The coordinate file is in BROOKHAVEN or DSSP format.

The sequence file format looks like this:

```
2RHE
ESVLTQP-PSASGTPGQRVTISCTGSATDIGS----NSVIWYQQVPGKA
PKLLIYYNDLLPSGVSDRFSASKSGTSASLAISGLESEDEADYYCAA
WND-SLDEPGFGGGTKLTVLGQPK*
KOL VL
ESVLTQP-PSASGTPGQRVTISCTGTSSNIGS----ITVNWYQQLPGMA
PKLLIYRDAMRPSGVPTRFSGSKSGTSASLAISGLEAEDESDYYCAS
WNS-SDNSYVFGTGTKVTVLGQPKANPTVT
LFPPSSEELQANKATLVCLI*
NEW VL
-SVLTQP-PSVSGAPGQRVTISCTGSSSNIGA---GNHVKWYQQLPGTA
PKLLIF-------HNNARFSVSKSGSSATLAITGLQAEDEADYYCQS
```

```
YDR---SLRVFGGGTKLTVLRQPKAAPSVTLFPPSSEELQANKAT*
```
Alternatively, sequences can be read from a HSSP file.

## RELATED PROGRAMS

The program Cameleon by G.M. Morris, Oxford, has a rich set of features, but handles only two proteins, and has no structure display.

## RESTRICTIONS

The upper limit of the number of sequences is set to 100 each having at most 2000 residues. No side chains can be displayed.

## DEVICES REQUIRED

**nseqtool** requires a colour graphics screen with a mouse. It is implemented on SUN-4 and SPARC workstations under Suntools.

# PDB

## FUNCTION

The protein Data Bank (**PDB** ) contains 3-D coordinates and bibiographic entries of proteins, DNA, etc. as determined by X-ray crystallography and NMR spectroscopy.

## AVAILABILITY

**PDB** files are available on the VAX cluster in the directory **$PDB**. On the UNIX cluster in directory **/data/pdb/...**

## RESTRICTIONS

You may not edit or write in the **$PDB** and **/data/pdb** directories.

## FILE FORMAT

The official format description for PDB files can be obtained from G.Vriend.

## RELATED DATABASES

Derived databases are DSSP (files: $PDB:pid.DSSP; /data/dssp/pid.dssp) containing secondary structure assignments, as derived from 3-D coordinates. HSSP (files: $PDB:pid.HSSP; /data/hssp/pid.hssp) containing the family sequence alignment for each protein of known structure. Many other programs provide derived databases. **O, FRODO, Insight, WhatIf,** etc. all have structure fragment databases (so-called DGLOOP fragment databases). **WhatIf, BIPED** and **SESAM** provide relational structure, sequence and function databases with attached query systems.

## FUNCTION

**PLUTO** draws mono and stereo views of small and larger molecules in a variety of styles.

## AVAILABILITY

**PLUTO** is available on the VAX cluster. Type **prepare CCP**.

## AUTHOR

Sam Motherwell, 197*.

## EXAMPLE

The following example plots a stereo plot of 1ppt.brk.

```
$ prepare ccp
$ PLUTOPS FROM  $PDB:1ppt.BRK PLOT embl$scratch:PLOT.ps
$ GO
DATA
CELL 1.0 1.0 1.0 90. 90. 90.
SYMM X,Y,Z
INPUT BROOK
TITLE 1PPT SIZE 0 SCALE 2.5
LINK CA CA 4.0
JOIN RADII RESIDU CA 0.9 C 0.9 N 0.9 S 0.9 O 1.0 P 1.2 A 0.9
RADII BONDS ALL 0.15 LINES 8
SIZE 0 SCALE 2.5
STEREO
VIEW XO
PLOT
$ print /quelwv133 /parapostscript  embl$scratch:PLOT.ps
```

**INPUT** The user should specify: XYZ coordinates and display specifications like line type, symmetry viewing position, sphere radii, atomic bonds,etc.

## OUTPUT

Output is a postscript file eg. plot.ps.

## RELATED PROGRAMS

Several other programs (eg. **Whatif, Insight**) can also make hardcopy plots, but perhaps not as fancy as **PLUTO**.

## SUGGESTIONS

PostScript **PLUTO** plots can be generated using **QUANTA** (for stereo pictures the left-eye and right-eye views have to be plotted separately).

## FUNCTION

**Polarh** adds polar hydrogen atoms to a set of protein coordinates.

## AVAILABILITY

**Polarh** is made available on the VAX cluster by the command **Prepare proteins**.

## DESCRIPTION

**Polarh** takes a BRK/PDB format set of protein coordinates as input (e.g. $pdb:1crn.brk). It then adds polar hydrogen atoms as specified in a residue library. Output is a new set of coordinates. Polar hydrogen atoms are required for electrostatics and detailed energy calculations.

## AUTHORS

John Moult. About 1985. Jiffied at EMBL by Chris Sander.

## EXAMPLE

Here is a sample session with **Polarh.**

```
prepare proteins
polarh 1crn.brk

POLARH by John Moult will add polar hydrogen atoms
   to a PDB/BRK protein coordinate file

usage:
   (a) $ polarh xxxx              ! [e.g.  xxxx  4PTI ]
or (b) $ polarh myprotein.BRK     ! [with extension]

Output coordinates in xxxx.BRKH or myprotein.BRKH

...converting to john's format....
...adding polar hydrogens....
...converting back to BRK format...

       395  lines processed.
       395  atoms converted.

Polarh done.
Your new coordinates are in file 1CRN.brkh
```

## INPUT FILES

Background data input to **Polarh** is the residue library $2:[sander.proteins]polarh-reslib.dat; contains atoms names and geometry specifications.

## OUTPUT

Coordinates with polar hydrogens added are in file xxx.brkh Example:

```
ATOM     12  N   THR   2       15.115  11.555   5.265  1.00  0.00   0
ATOM     13  HN  THR   2       15.516  10.742   4.843  1.00  0.00   0
```

```
ATOM     14  CA  THR     2      13.856  11.469   6.066  1.00  0.00     0
ATOM     15  C   THR     2      14.164  10.785   7.379  1.00  0.00     0
ATOM     16  O   THR     2      14.993   9.862   7.443  1.00  0.00     0
ATOM     17  CB  THR     2      12.732  10.711   5.261  1.00  0.00     0
ATOM     18  OG1 THR     2      13.308   9.439   4.926  1.00  0.00     0
ATOM     19  HOG THR     2      12.650   8.892   4.408  1.00  0.00     0
ATOM     20  CG2 THR     2      12.484  11.442   3.895  1.00  0.00     0
```

## RELATED PROGRAMS

Use convertcoor to convert coordinates to BRK format required by **Polarh**. Most energy minimization and molecular dynamics programs also have utilities for adding polar hydrogens.

## RESTRICTIONS

Input must be in PDB/BRK format. **Polarh** follows its own residue library which may not be compatible in details with that expected by MD programs.

## FUNCTION

**PolDiagnostics** performs diagnostic checks on proteins 3D structures. It asks: "does this protein have a polar/nonpolar surface composition typical of globular proteins ?".

## AVAILABILITY

**PolDiagnostics** is made available on the VAX cluster by the command **Prepare proteins.**

## DESCRIPTION

**PolDiagnostics** calculates the polar and nonpolar fractions of the solvent accessible and interior faces of a protein. It compares the calculated values with the distribution of such values in the database of known crystal proteins structures. It then makes an estimate as to whether the input protein is OK (compatible with the database distribution) or WRONG (unusual values relative to the database).

## AUTHORS

Cornelius Froemmel and Chris Sander, 1988.

## EXAMPLE

Here is a sample session with **PolDiagnostics.**

```
  ( answer all questions in response to prompts ):

    prepare proteins
    poldiagnostics

Welcome to POLDIAGNOSTICS88

POLDIAGNOSTICS may identify your protein as typical (OK)
  or as atypical (WRONG) compared to the database of
  known globular proteins.

Enter protein identifier, for example
Coordinates in MYPROT.BRK ::              Enter MYPROT
Precalculated surface areas in PROT.SRF:: Enter PROT
Protein data bank protein 4PTI::          Enter 4PTI

 [DefaultCR4PTI] : lcrn

   ( here you might be asked to submit a batch job for surface
     calculation )

 POLDIAG results are also in file  1CRN.STP !
```

## OUTPUT

Here is sample output for a misfolded proteins, from file mis-heme.stp:

```
**** POLAR diagnostics list BY PROGRAM POlDIAGN, VERSION march  88 ***.
HEADER MCPTHEC30 Misfolded hemerythrin. Side chains generated by C
**** Reference C. Froemmel J.theor. Biol.  111 (1984) 247-260 .
**** Reference Baumann, Froemmel, Sander 1989 Prot.Eng.2 p329-334.
```

```
* ------------------------------------------------------------------------ *
** The chain length                  113
** The molecular weight            13319.0
*  *
                    in this prot|      in 150 proteins
                       observed | lowest  mean  highest value  decision
* ------------------------------------------------------------------------ *
             MW          117.867  97.000 109.000 115.000          WRONG ! (1)
         MAXPOL            0.169   0.163   0.173   0.185             OK  ! (2)
         ACCSRF            0.401   0.320   0.400   0.520             OK  ! (3)
         INTPOL            0.189   0.158   0.172   0.190             OK  ! (4)
         OUTPOL            0.131   0.147   0.172   0.204          WRONG ! (5)
          SCINT            0.126   0.074   0.098   0.118          WRONG ! (6)
          SCOUT            0.100   0.115   0.143   0.183          WRONG ! (7)

1  molecular weight per residue
2  value of polar fraction of the virtual extended chain of the given protein
3  normalized accessible surface area
4  normalized internal polar fraction
5  normalized external polar fraction
6  normalized internal polar fraction of side chains
7  normalized external polar fraction of side chains
```

## RELATED PROGRAMS

**WhatIf** does quality control on structures using contact distributions.

## RESTRICTIONS

**PolDiagnostics** currently requires that your own (non-$PDB:) coordinate files are in your home directory. File extension must be like *.BRK.

## INPUT FILE

Background data input to **PolDiagnostics** is a residue library with atomic partial charges and a list of maximal atomic surfaces.

## FUNCTION

**ProteinInfo** runs a set of programs which provide information on a protein, given its amino acid sequence. A summary report provides hypotheses about the structure and function of the protein.

## AVAILABILITY

Made available on the Vax cluster by typing **prepare proteins**. Best run as a batch job.

## DESCRIPTION

**ProteinInfo** runs these programs: predict (ALB, GORIII, Segment83), predictclass, fasta, prosite, pepplot) on one protein or a set of proteins.

## AUTHORS

Many. Contact Brigitte Altenberg or Chris Sander for help.

## EXAMPLE

(user provides a file which contains file names of protein

sequences; name the file ProteinInfo.lifi)

$ ProteinInfo

Batch (batch) or interactive job (int)?[batch]:

batch

(this will start a batch job)

## INPUT FILE

ProteinInfo.LIFI contains sequence file names or swissprot identifiers, e.g.:

tegu$ebv

tat$hiv21

myprotein.dat etc.

## OUTPUT

The following output files end up in the directory embl$scratch:[username] and are also printed on the laser printer PS40 :

proteinid.dat (protein sequence)

proteinid.x (result of secondary structure predictions)

predictclass.s (result of predictclass)

proteinid.fasta (result of fasta database search)

proteinid.pros (result of prosite pattern search)

gcgplotfile.lis (GCG's pepplot with hydrophobicity and other profiles)

proteinid.hssp (family sequence alignment, possible structural homology)

proteinid.prop (sequences with similar global properties)

## RELATED PROGRAMS

Suite of programs in the GCG or STADEN sequence analysis packages. PC/Gene package by Amos Bairoch.

## RESTRICTIONS

Protein input sequences must be in one of the standard formats (e.g. EMBL, GCG, PIR, NBRF). In input file ProteinInfo.lifi, do not mix file names (e.g. myprot.seq) with naked swissprot identifiers (e.g. tat$hiv21)

## SUGGESTIONS

Run **CONVERTSEQ** or **REFORMAT** prior to batch job to check your input format. Job should be run in batch. May take more than an hour of CPU time.

# QPACK

## FUNCTION

**QPACK** assesses how well residues are packed together in a protein structure.

## AVAILABILITY

**QPACK** is made available on the VAX cluster by the command **Prepare BBK** followed by **USE QPACK** and **QPA**.

## AUTHOR

Lydia Gregoret, Dept Pharmaceutical Chemistry, University of California, San Francisco. CA 94143-0446.

## DESCRIPTION

**QPACK** calculates packing on the basis of a sphere growth algorithm about (in most cases) the Cb atom of each residue. The packing is calculated according to when the first and second contact to another residue occurs. The command QPA executes a VAX command file to give the following options.

Select analysis option:

(1) D - create PDB file with packing information suitable for Display

(2) [P] - create a pair list for automatic pairpotential analysis

(3) A - create a full residue interaction list for analysis

(1) If D selected filename.QCM is created

-this file can be displayed and coloured according to BVALs (which are packing results)

Files for display on HYDRA are available in PROG$DIR:[QPACK]. These are:

QPHYD.DAT -hydra control file

QPM.SEL -hydra main select

QPM.COL -hydra main colour

QPS.SEL -hydra second select

QPS.COL -hydra second colour

-the above define 5 colour ranges as follows (as in paper)

overpacked (too tight) OK underpacked (too loose)

| Blue | Purple | Green | Yellow | Red |

-the paper suggests that a purple/green colour good packing

NOTES: Hydra does NOT shade the BVALs continuously, as done in the paper. I suggest using a different display program that can do this would be an improvement. The select files here only display backbone. This gives no loss of information since BVALs change by RESIDUE. If all atoms are

displayed then CB and CG will be detached from CA's since they are psuedo atoms. They will also be incorrectly coloured.

(2) If P selected filename.PA is created

-this is further processed automatically by PPQP to produce file filename.PAI suitable for SUMPAIRS which is also run creating file filename.PP (*.PAI is subsequently deleted)

-*.PP file contains a single value. This will also be output to the screen as the program runs, and represents the "pair potential" of the list of pairs defined by QPACK. Generally if -ve (10-20) this is a good indication. See paper for details.

(3) If A selected filename.QP is created

**INPUT FILE**

Input to **QPACK** is a Brookhaven coordinate file.

## FUNCTION

**QUANTA** is a program to aid molecular modellers, drug designers and crystallographers.

## AVAILABILITY

**QUANTA** is made available on the VAX cluster by the command **Prepare QUANTA**. It needs the PS 390 in order to use the graphic options. It is also available on the Iris workstations (Silicon Graphics).

## DESCRIPTION

**QUANTA** is an integrated package of database handlers, protein analysis tools and high quality graphics.

```
a brief summary of some of the options:
- display and colour facilities for multiple molecular structures
- 3D molecular construction
- energy minimization and molecular modelling (+ optional constraints)
- molecular modeling
- protein modeling
- comparison of molecules
- drawing modes like : ribbons
                       brick maps
                whole map
                grid contour meshes
                symmetry related copies
                unit cell edges
                molecular surfaces
                Buried surfaces
                        etc...
 - Conformational search and analysis like:
                random sampling
                hydrogen bonding
                        etc...

there is extensive descriptions available through menus.
```

## AUTHOR

The graphics part of **QUANTA, HYDRA** was written by Rod Hubbard. **QUANTA** is sold by Polygen. For help see Johan Postma.

## EXAMPLE

Here is a sample session with **QUANTA**.

$prepare QUANTA

$QUANTA.nf

```
pull down the data menu

pick 'read molecule into msf' and pick your options in the menus
```

## INPUT FILE

Input to **QUANTA** is typically PDB-format coordinate files.

## OUTPUT

Everything displayed on the PS390 screen can be plotted. Everything displayed on the terminal screen can be printed.

## RELATED PROGRAMS

Other graphics programs/packages in the same class are **Frodo, Insight, WhatIf, Discover** and **BioGraf.**

The MD is based on the CHARMM and optionally XPLOR packages.

## RESTRICTIONS

Protein Data Bank files must have the extension '.pdb'

## FUNCTION

**SWISS-PROT** is a protein sequence data bank where all the data are easily retrievable by computer programs and are stored in a format similar to that of the EMBL Nucleotide Sequence Data Library. **SWISS-PROT** contains all PIR sequence data converted into a EMBL-like format, as well as additional sequences either entered in house or translated from entries in the EMBL Nucleotide Sequence Data Library. **If you need a user manual please ask Peter Rice or Reinhard Schneider.**

## AVAILABILITY

**SWISS-PROT** is available on the VAX (GCG-package) and UNIX machines. See also the programs listed under **RELATED PROGRAMS.**

VAX:

$DL:[Seqlib.Swissprot.15]

flat files for each entry are in Data_lib$Swissprot_Current

UNIX:

/data/swissprot/

flat files for each entry are in /data/swissprot/current/a..z/

## DESCRIPTION

see user manual

## AUTHOR

Amos Bairoch (Medical Biochemistry Department, Centre Medical Universitaire, 1211 Geneva, Switzerland) and EMBL Data Library

## EXAMPLE

Here is a example of a **SWISS-PROT** entry.

```
ID   CRAM$CRAAB      STANDARD;      PRT;    46 AA.
AC   P01542;
DT   21-JUL-1986  (REL. 01, CREATED)
DT   21-JUL-1986  (REL. 01, LAST SEQUENCE UPDATE)
DT   01-NOV-1988  (REL. 09, LAST ANNOTATION UPDATE)
DE   CRAMBIN.
OS   ABYSSINIAN CRAMBE (CRAMBE ABYSSINICA).
OC   EUKARYOTA; PLANTA; SPERMATOPHYTA; ANGIOSPERMAE.
RN   [1] (SEQUENCE)
RA   TEETER M.M., MAZER J.A., L'ITALIEN J.J.;
RL   BIOCHEMISTRY 20:5437-5443(1981).
RN   [2] (X-RAY CRYSTALLOGRAPHY, 1.5 ANGSTROMS, AND DISULFIDE BONDS)
RA   HENDRICKSON W.A., TEETER M.M.;
RL   NATURE 290:107-113(1981).
CC   -!- FUNCTION: THE FUNCTION OF THIS HYDROPHOBIC PLANT SEED PROTEIN
CC       IS NOT KNOWN.
CC   -!- SIMILARITY: TO PLANT THIONINS.
DR   PIR; A01805; KECX.
```

```
DR   PDB; 1CRN; 16-APR-87.
DR   PROSITE; PS00271; THIONIN.
KW   THIONIN; 3D-STRUCTURE.
FT   DISULFID       3      40
FT   DISULFID       4      32
FT   DISULFID      16      26
FT   VARIANT       22      22        P -> S.
FT   VARIANT       25      25        I -> L.
SQ   SEQUENCE   46 AA;  4736 MW;  10132 CN;
     TTCCPSIVAR SNFNVCRLPG TPEAICATYT GCIIIPGATC PGDYAN
//
```

## RELATED PROGRAMS

**GCG-package, findprotein, findspecies, copyprotein**

## FUNCTION

**WhatIf** is a program to aid molecular modellers, drug designers and crystallographers.

## AVAILABILITY

**WhatIf** is made available on the VAX cluster by the command **Prepare WhatIf**. On the IRIS systems, click the **WhatIf** button. On the Evans & Sutherland workstation, ask Gert Vriend.

## DESCRIPTION

**WhatIf** is an integrated package of database handlers, protein analysis tools and high quality graphics. The options include:

**Interfaces to:**

GRID
GROMOS
FASTA
PSQ database
DSSP

**Many options in the fields of:**

Modeling by homology
Loop transplantations
Single or multiple mutations
Correction of (re-)designed structures
Evaluation of the quality of protein model

**Databases:**

Fragment database
Relational structure sequence database
Water position database

**Crystallographic tools:**

Knows all spacegroups and settings
10 maps in memory
Property coloured contour lines
Real space refinement
Crude and fine regularization
Graphical analysis of reciprocal space
Envelope editor
Automatic structure building from alpha carbons

**Graphics:**

Every atom can be coloured as function of every property
Smooth splines for main chain tracing
Mobile arrows or other objects
Draw cylinders through helices
Add text, and move text around

**Summary:** In principle it boils down to the following: You use your own favourite program. But one moment there is something you can not do with it. An now you will have to RTFM, and do it with **WhatIf**.

## AUTHOR

Gerrit Vriend. Period: 6-12-1987 - now.

## EXAMPLE

Here is a sample session with **WhatIf**.

```
Working on a terminal next to the POLLUX PS390, type:

   WhatIf
   1
Type the following responses to the prompts from WhatIf:
   getmol
   $PDB:1crn.brk
   crn
   Graphic
   DownLd
   ShoAll
   1
   Crambin
   Center

(You now have a picture on the screen which can be rotated and translated.
For further details see the WhatIf manual)
```

## OUTPUT

Everything displayed on the PS390 screen can be plotted. Everything displayed on the terminal screen can be printed.

## RELATED PROGRAMS

Other graphics programs/packages in the same class are **Frodo**, **Insight**, **Discover**, **O** and **BioGraf**.

## RESTRICTIONS

You must have a pagefile quota of at least 32000 on the VAX. The scratch files created by **WhatIf** can take up to 10,000 blocks.

**WhatIf** can work with up to 100 molecules, 4,000 residues and 32,000 atoms at a time. Up to 800 graphics items can be manipulated.

## INPUT FILE

**Input to WhatIf** is typically PDB-format coordinate files.

GRENDEL

# The design of GRENDEL,

# a four helix membrane protein

designed by:

John Moult — CARB, University of Maryland, 9600 Gudelsky drive, MD 20850, U.S.A.
Tel: 301-251-2241, FAX: 301-251-2255
e-mail: jmoult@iris4.carb.nist.gov

Cara Marks — LMB, Hills road, Cambrige CB12QH, U.K.,
Tel: 044-223-402 359/402, FAX: 044-223-412 178
e-mail: cara@mrc-lmb.cam.ac.uk

Christine Gaboriaud — L.I.P. CENG, BP85x 38041 Grenoble cedex, France
France
FAX ++33-77 88 59 18

Steve Emery — Institut fur Molekulare Genetik, Univ. Gottingen, Grisebachstr. 8, 3400 Gottingen, Germany
e-mail: u1098@dgogwdg5.bitnet

Christophe Verlinde — Lab. Chem. Physics, Univ. Groningen, Nijenborgh 16, 9747 AG Groningen, The Netherlands
Tel: 50-47 42 59
e-mail: verlinde@hgrrug52.bitnet

## 1.1 SUMMARY.

The recent determination of two membrane protein structures by X-ray crystallography (the photoreaction center (1)) and electron microscopy using 2 dimensional crystals (bacteriorhodopsin BRH (2)), together with progress in understanding the basis of their stability and folding (3), makes it feasible to attempt to design a membrane protein.
A four helix membrane protein has been designed. The design is conservative, in that it includes helices A and B from BRH, which are known to fold in the membrane in the absence of the rest of the molecule. These helices were found to be very similar in topology to the two helix monomer of the rop protein (4). Rop forms a 4 helix bundle dimer, and this motif was used as a basis for the design. Thus the the helix-turn-helix at the beginning the BRH structure was duplicated and positioned relative to the original motif in the same way as the components of the rop dimer, making a conventional 4 helix bundle. The sequence of the second pair of helices, referred to as 2′ and 1′, was modified so as to produce a side-chain side-chain polar interaction between the each of the original helices and its symmetry mate, and to guide and stabilize the fold. Further sequence changes were made to the second pair of helices to build a hydrophobic core in the centre of the bundle. A few conservative substitutions in the original helices were necessary to achieve this, but they do not appear critical for the stability of the unit. The loop between helix 2 and helix 2′ will be the 13 residue epitope from Sendai virus (5), and between helix 2′ and helix 1′, the 11 residue loop from myc (6), thus providing a means of monitoring the presence of the protein. No attempt to model the conformation of these loops was made. The helices are capped at the surface in the manner found in soluble proteins (7).
The resulting model conforms to the known rules for membrane proteins. The structure with-in the lipid region of the membrane is helical, satisfying the hydrogen bonding requirements of the backbone. There is approximately one polar interaction linking each pair of helices in the lipid region, and the core is reasonably packed. There are however a number of uncertainies remaining. In particular, the principles for structure in the head group region of the membrane are unknown, and there is no concensus on the mechanism of insertion of such structures in to the membrane. The energetics appear to be less finally balanced than for soluble proteins, but this has yet to be tested. Still, we consider the structure worth developing experimentally.

## 1.2 REFERENCES

(1) J.Deisenhofer and H.Michel Science vol 245 1463 (1989).

(2) R.Henderson, J.M.Baldwin, T.A. Ceska, F.Zemlin, F.Beckmann and K.H.Downing J.Mol.Biol. vol 213 899-929 (1990).

(3) D.W.Banner and M.Kokkinidis J.Mol.Biol. vol 196 657-675 (1987).

(4) J.-L.Popot and D.M.Engelman Biochemistry vol 29 4031-4037 (1990).

(5) K.Friedrich Dissertation, University of Munich, (1990).

(6) S.Munro and H.R.B.Pelham Cell vol 46 291-300 (1986).

(7) J.S.Richardson and D.C.Richardson vol 240 1648-1652 (1988).

## 1.3 INTRODUCTION

Why not a de novo membrane protein design? There may be some good answers to that question. First, there are few structures known. Second, there have been fewer studies of the stability and folding of these molecules, compared with globular proteins. Never-the-less, there seemed to us to be enough data and concepts to allow a design to be made. One of the objectives in doing this design was to learn more about these structures and to confront the difficulties head on. We also were serious about producing a viable sequence as well, though bearing in mind there might be unanticipated difficulties ahead.

## 1.4 CHRONOLOGICAL REPORT

### 1.4.1 Day 1: Choice Of Project.

Initial discussions concerned the introduction of either a binding site or catalytic function in an existing protein (such as an immuno- globulin or Ferritin) and in the design of novel membrane protein.

We decided to look at various structures in the PDB to assess the possibilities. Carefully studied the photo reaction centre (PRC) which took most of the day. Looked for key interactions that could be responsible for the the association of the intramembrane helices. Not so easy to determine, because of the large cofactor groups that clearly have a major role in the final structure. There were some polar interactions between and within helices, involving serines and asparagines, but no nice neat pattern of two key polar interactions between each pair, which would have made life simple. There was one place where a gly in one helix has a phe from a neighbouring helix packed up against it. Someone (Lynne) says that in general there are more glys in transmembrane helices than those found in soluble proteins. Is this because the resulting grooves are keys for the helix association??

Also started to look at REI, and its loops. In general, the loop design problem still seems difficult, with no good rules for doing it. Probably need more tool development before that's an attractive design project.

### 1.4.2 Day 2: Decision And Plan.

Started looking at bacteriorhodopsin. Saw nice examples of helix packing in this one! Anyway, after reading the Engelman mini-review, there was wide spread interest to attempt a membrane protein design. It all looks so easy!!!! The theory asserts that individual helices can fold stably in the membrane, even though there are some unsatisfied polar groups, because of the large number of hydrogen bonds within the backbone. In addition, all the hydrophobic groups ensure they stay in the membrane. The making of the polar contacts between pairs of helices is believed to provide enough energy to overcome the entropy of fixing the helices relative to each other (estimated as 1 to 10Kcal/mole). All in all, seems a much more robust energetic picture than in solution, assuming all this is true.
Of course, when we actually started looking at the bacteriorhodopsin structure, things were not quite so simple:
1) There seems to be about one polar inter helix interaction per helix.
2) One inter helix salt bridge was observed.
3) Also an unsatisfied asp, (no Cleopatra present !) but apparently that is well known to be involved in the function.
4) The packing does not look all that great, but it is hard to know whether this is due to the poor resolution of this structure.
5) In general, non-polar inter helix contacts appear to be similar to those found in soluble protein helix bundles.
The plan (when feasible in the time available):
Build various 4 helix membrane proteins, ranging from as conservative as possible to fairly wild. Then Steve will have something he hopefully believes in enough to make, and can work up until something fails. Thus, the first one will take helices A,B,C and G which pack against each other) remove the retinol, and repack as little as possible. We will need 3 external loops (and may be a bit more internal, we dont know yet). Then make a de novo bundle backbone, using some sort of ideal motif. Build the side chains of helix A on one helix, and re-pack the others. (Steve wants to keep the first helix plus the N-terminal sequence, if possible, to ease expression).
A few of the things we know we dont know:
1) What do the outsides look like? Are there any rules?
2) Do we have all the intra membrane bits?
3) What happens at the surface? Are charges needed to get through the head group zone?
4) How does one determine which way round the thing goes in?
5) Is very careful tuning of the amount of unsatisfied polar groups during assembly needed?

### 1.4.3 Day 3:

During the discussion last night, it was pointed out that helix G packs parallel to helix A i.e. our conservative initial bundle will not work. Instead, we decided try and and swing around helix D to complete the bundle. Spent most of the day collecting analysis data, and looking at the helix interactions using WHATIF with Gert as pilot. Confirmed the earlier impression of generally poor

packing, but there are some main chain side chain contacts. We have not yet absorbed all the amassed information. In particular, we still have no clear picture of electrostatics for side chain interactions in the helices.

1.4.4 Day 4:

Discussed the conclusions from the information gathering of yesterday. Conclusions were:
1) The helices seem a bit smoother on the outer surfaces than the inner ones. This has been suggested as an assembly mechanism: The lipid groups cannot follow the contours of the rough inside surfaces, so cavities exist when the helices are separated. We are trying to quantitate this effect by doing accessibility calculations with different size probes on the helices.
2) There are some close approaches of the helices, with only one side chain thick interfaces between them. For example H1 to H2 closest approach is 8.0 Angstroms. D-E 8.6, A-G 8.9. (data from Arthur Lesk's helix analysis program). Of course, this happens in soluble proteins too, for instance the famous Gly-Gly contact in hemoglobin.
3) Most (all?) buried charges are involved in function (i.e. R82, E 204, D212, D96, R235).
4) There are very few conserved residues (particularly those not obviously involved in function) over the 3 sequences we have:
H1 A18 (non-function related) H2 P50 (non-functional, but probably creating a kink to accomodate the retinol H3 P91 (as P50) and T90, W86, Y83, R82 (functional) H4 A114 (NF) and D115,M118, G12 (F). H5 L152 (NF) and S141 (F). H6 F171,L174, L181, V188 (NF) and W182, Y185, P186, W189 (F). H7 L211, V213, A215 (NF) and D212, K216 (F).
(These three sequences are all from halobacteria).
So again one has the impression of a rather unfussy structure - many substitutions can be tolerated.
5) Polar side chain contacts between helices.
H1-H2 T17 OG1 F54 O, M20 O, P50 O1, T24 OG1 A5 N, L28 N T47 OG1
i.e., 3 T OG to main chain, in a line on the two helices (i.e. roughly one turn apart 17, 20, 24, 28. But 20 is an O-O contact. We have to check this further, but it looks like a one side chain thick region stabilized by polar contacts. The specificity of the interactions must be determined by the contour of the other side chains around as well.

H2-H3 T46 N D96 OD1, Y57 OH M209 O, Y57 OH D212 OD1
Again 2 side to main chain contacts, but this time involving asps. As observed above, these are probably here for function, but can also have a role in determining assembly arrangements.

H3-H4,H7 L87 N D115 OD2, T90 OG1 D115 OD2, L87 O D115 OD2, Y79 OH E204 OE1,
R82 NH2 T205 OG1, R82 NH2 T205 N, R82 NH1 E205 O,
T89 OG1 LYR NZ (joined to retinol, +ve charge), T90 OG1 D115 OD1,
W86 NE1 D212 OD2.
Again side chain-main chain predominates, again with charges involved in function. Also some side chain-side chain, involving charges.

H6-H7 E116 OE2 R225 O, E116 OE2 R25 NH1,NH2, Y185 OH D212 OD2

More side-side chain interactions (with charges), one side-main chain. So there are a surpising number of side chain main chain contacts. Surpising because that does not to seem in itself to give specificity of interaction. But presumably these could be involved together with the shape considerations. Although a lot of charges are found here, we think that it would be safer to stick to polar interactions in the design: Its not clear how easy it is to bury the charges during assembly, and in fact the polar groups would probably should provide enough energy. In the conservative design we will try to preserve a layer thick side to main chain interaction motif.
6) Non-polar packing. Looks fairly normal, except that again there are a number of single layer arrangements.
7) Still not clear whether there are particular concentrations of charges at the surfaces: certainly no conservation seen. Checking the pattern further.
8) There are a large number of W and Y residues near the surfaces (1Y+6W on the N terminal side, 2Y+1W and 1Y+5W in the other sequences). The reaction centre also shows an astonishing concentration of these residues at both ends of the helices, even though they are largely adjacent to the soluble domain there and more commonly solvent here. We dont know what the significance of this is.
9) There are a lot of Glys: H1 4, H2 0, H3 0, H4 5, H5 1, H6 1, H7 2. Similar numbers in the other sequences BUT almost all in different places. May be involved in close approach, but again we dont know yet, and the variation in position is puzzling.

### 1.4.5 Day 5

Yesterday afternoon, we started building a 4 helix bundle by duplicating helices 1 and 2 of BRH and docking them to the original ones. A double two fold rotation performed to (a) place the faces to be buried opposite each other and effect a close approach and to (b) put the loop at the opposite end. (is there something about 2 such two fold rotations must be reducable to 1?). Marked up the residues that get buried on helices 1 and 2 (assessed using WHATIF surface accessibility for single helices and the whole molecule) on forming the complete BRH molecule. Then, using Insight II, tried to dock the two surfaces. Hopeless. Later, with Gert, looked at the same thing with WHATIF. Looked difficult still, so Gert suggested using the Rop structure as a guide for docking. He used a program feature to get the best structural alignment of two proteins (allowing insertions and deletions). Surprise! a very good fit. (Looks like less than 1 Angstrom RMS on the helix regions, though we do not have a number). But the loop is at the opposite end compared with the Rop subunit (see alignment picture). Gert thinks this is the best fit of a pair of helices in the database. Chris says there is a kink in one of the Rop helices. Is this matching the Pro kink in H2 of BRH ?.
Anyway, because of this fit, the main chain alignment of the two monomers of Rop provides a wonderful starting situation for the BRH 4 helix bundle, and we have these co-ordinates in the bag/computer. The thing is right way in/out. I.e residues that should be on the outside are there (it could just as easily have been the other way round). There are clashes in the core, and we now have to consider our

repacking options. Gert will use his bump remover first. This changes the torsions of clashing residues in a systematic way by steps of 120 degrees, starting at the end chi, until a staisfactory fit is obtained. We could also rebuild it as a pseudo Rop core, with a couple of polar interactions added.
Another interesting thing yesterday was the result of a homology search done by Reinhart. He found an 8 of 10 match for part of BRH H3 (87-96) with the inner membrane protein CET from E.coli. Seraching with that 10 residue stretch from the CET sequence, and accepting everything 7 or more correct out of 10 gave a lot of hits. Essentially all were membrane segements or signal sequences.

```
BRH   LFTTPLLLLD
CET   LFTLALLLLD
```

This is of course L rich, but a lot of the hits have the FT or F followed by something polar. Does this mean there really is a signal sequence in the third helix?

### 1.4.6 Day 6:

Friday and half Saturday have slipped by, plus all that goofing off on Sunday of course. In the cold light of the graphics, there were some problems with our 4 helix bundle based on the Rop packing. The pairs of helices are displaced approximately one turn out of register from each other. i.e. a turn sticks out the end of the bundle at each end. Also, the basic 2 helix motif from BRH has the second helix about a turn longer than the first. Consulting with Tom Ceska, we have decided to shorten helix 2 to be the same length as helix 1 in the second copy, on the belief that the extra piece sticks out of the membrane in BRH.
Thus, the bundle ends should be more or less flush with the surface for the second pair of helices (about 45 Angsroms of membrane including the head groups ?). This entails lengthening of the helices at one end and shortening them at the other to remove the undesirable stagger. Therefore:-

delete 38-41 from helix 4
delete 28-32 from helix 3
add 5 residues to helix 4
add 3 residues to helix 3

Note the new connectivity will be
1 - 2 - 2′ - 1′
Loops:
Going to use the Sendai virus epitope for the connecting loop for the 2 - 2′ helices (i.e. external side of membrane), and the myc antigenic site for the 2′ - 1′ connection (cytoplasmic side). The original 1 - 2 loop will be preserved. Remember our strategy is that as 1-2 is an independent folding unit, we should leave it as unaltered as possible, so that we can get maybe get away with mistakes elsewhere.
The Sendai virus sequence is:

DGSLGDIEFYDSS
Quite long for the available space, as is the myc sequence:
EQKLISEEDLN

1.4.7 Day 9

There was a set back yesterday morning, when we discovered that at some point we had started to use an incorrectly assembled bundle. This problem was noticed after a considerable amount of time had been spent engineering the helix ends. Captain Gert manually docked the helices to the correct bundle.
As a consequence, we adopted a new method for book keeping (called notes !).
Decided to finish off the helices with conventional end caps. This may be silly given that the helices (all but 2) end in the head region of the membrane. Still, it is likely to do more good than harm.
Used the data in the Richardsons' Science letter to design the caps. For the N cap:

```
GLY ASN PRO ASP
N'  CAP N1  N2
```

appears to be a consenus.
The idea here is that the lack of an NH on the PRO terminates the helix, and the ASN side chain interacts with helix NHs. The ASP compensates for the helix dipole. Used WHATIF database to search for the above sequence, with no hits. For GLY ASN PRO, 5 hits, only one of which is an N cap, in human hemoglobin:

```
56 -> GNPKVKA
```

Docked this to the start of helices 4 and 3.
For the C cap, consenus:

```
LYS GLY XXX THR
C1  CAP C'  C"
```

3 hits with this sequence. Checked some more by relaxing the LYS. A lot of hits. Turns out the THR does not usually H-bond to the helix backbone but to the C=O of the GLY, apparently preventing the helix starting again. The GLY NH does make H-bonds to 2 of the helix C=Os, in the prescribed manner. Gert extended H2 by LYS GLY ALA THR, then built in the epitope sequence in an extended conformation (-150,150). (Christine had used FASTA to search the PDB for homologies to the sequence, but there are none).
Same procedure for the loop between H2' and H1'. CRUDE in WHATIF was used to loop the extended regions round, and nearly joins them. Still have to finish that off. New complete structure saved as GRENDEL 5
So we have the protein backbone. Now for the core. We plan to add one polar interaction between helices 1 and 1' and 2 and 2'. Then we plan to leave the sequence on helices 1 and 2 unaltered, and change the sequence of the core on helices 2' and 1', trying to find a combination which packs satisfactorily. Several approaches to this

will be tried, including running propak.

### 1.4.8 Days 10 And 11

Building core, and polar interactions between helices 1 with 1' and 2 with 2'. Tried to use PROPAK extensively to help with this, but with no significant success: still seems to be a major tool missing here! So done painfully by hand. In the end, fairly satistfactory (see later).

### 1.4.9 Day 12

Energy minimization of the membrane spanning part of GRENDEL:

#### 1.4.9.1 Remove Loops Containing Epitopes -

Residues 62-72 and 104-114 were removed from the model. This leaves us with three polypeptide stretches: Pro 1 - Gly 61, Gly 73 - Gly 103, and Gly 115 - Val 141. The original loop from bacteriorhodopsin, now residues 25-32, was kept.

#### 1.4.9.2 Choice Of Force-field -

As there is not such a thing as the best force-field and as energy minimization (EM) of a model should be merely considered as a cosmetic clean-up procedure, a force-field with a user-friendly interface was chosen: Dreiding I from the Biograf package version 2.10 . This is a general molecular force-field not specifically dedicated to protein modelling. Nevertheless, it has a big advantage for a cosmetic clean-up procedure: it uses explicit geometrical potential energy functions for hydrogen bonds. This allows running EM without using electrostatics, which is advantageous in the current situation because of two reasons: (1) our molecule is not electrically not neutral, (2) the difficulties involved in using a reasonable dielectric constant in the membrane, out of the membrane and at the interface between them.

#### 1.4.9.3 Setting Up Force-field Calculations -

General settings: no electrostatics, default settings for all other potentials

Run 1: steps 1 - .50: steepest descent EM, with C alpha's of all residues fixed in order to avoid debumping by backbone deformation; in order to avoid large deviations from our original packing ideas the atoms were not allowed to move by more than 0.1 A in each cycle.

Run 2: steps 51 - 100: conjugate gradient EM, identical settings as for run 1

Run 3: steps 101 - 200: conjugate gradient EM , but with the C alpha

constraints of the residues in the loop between helix 1 and 2 removed, i.e. for residues 24 - 32 (this includes also some of the capping residues.

Run 4: steps 201 - 805: conjugate gradient EM with all constraints removed; refinement converged because the convergence criterion of rms gradient < 0.1 kcal/A was reached.

Table: Refinement evolution (all energies in kcal/mol):

| | before run1 | after run2 | run3 | run4 |
|---|---|---|---|---|
| INTERNAL: | | | | |
| Bonds : | 1293 | 199 | 194 | 154 |
| Angles : | 556 | 642 | 612 | 529 |
| Torsions : | 168 | 174 | 170 | 191 |
| Inversion: | 827 | 75 | 89 | 58 |
| NONBONDED: | | | | |
| van der Waals: | 22488 | 126 | 18 | -49 |
| Hydrogen Bonds: | 1230 | -634 | -660 | -923 |

Note: the rather high internal energy is merely a consequence of the fact that this force field has not been pararameterized especially for proteins

1.4.9.4 EM Flaws And Model Correction -

After convergence of the EM the model was checked for the correctness of its geometry. At this stage two right-handed residues were spotted, namely Val 42 and Ala 87. Checking against intermediate stages of the EM procedure showed that these residues changed their chirality during the first stages of the steepest descent EM, and this in spite of the movement damping applied. It should however be mentioned that the inversion potentials in the Dreiding I force field are rather weak.

The model was corrected by manual repositioning of the side chains of the two corrupted residues and WHATIF debumping. After this crude procedure another Biograf Dreiding I EM was executed using the same protocol as during run 4 of step 3. The refinement converged after 335 steps. From a comparison of the energies between the model with the wrong chiralities at positions 42 and 87 and the model with the correct handedness it can be seen that the latter model packs significantly better. While the internal energy changes marginally (2 kcal/mol) the nonbonded energy decreases by -26 kcal/mol.

Table: Refinement evolution (all energies in kcal/mol):

| | before run | after |
|---|---|---|
| INTERNAL: | | |
| Bonds : | 202 | 153 |
| Angles : | 593 | 528 |
| Torsions : | 259 | 197 |
| Inversion: | 92 | 57 |
| NONBONDED: | | |
| van der Waals: | 9 | -60 |
| Hydrogen Bonds: | -878 | -937 |

1.4.9.5 Geometry Analysis -
After the refinement the model was compared to the starting model. Most shifts are rather large, which is not so surprising in view of the limited resolution of the bacteriorhodopsin structure we used as a template. The geometry of the final model of GRENDEL is satisfactory as assessed using the program GEOANA (E. Dodson): rms deviation (bond lengths) = 0.01 A, rms deviation (bond angles) = 2 degrees.

Table: Comparison of the model before and after the 1140 steps of EM refinement (rms in Angstrom):

| | | |
|---|---|---|
| Alpha carbons | = | 1.1 |
| Centers of residues | = | 1.2 |
| Side chain atoms | = | 1.8 |
| Backbone atoms | = | 1.1 |
| All atoms | = | 1.5 |

Table: Differences larger than 1.5 A (centers of residues):

| Helix 1 | | Helix 2 | | Helix 2′ | | Helix 1′ | |
|---|---|---|---|---|---|---|---|
| 1 | 2.6 | 33 | 2.7 | 73 | 3.0 | 115 | 2.6 |
| 2 | 1.9 | | | 77 | 1.5 | 116 | 1.7 |
| 4 | 2.2 | | | 95 | 1.5 | 117 | 1.7 |
| 5 | 1.8 | | | 101 | 1.6 | 119 | 1.6 |
| 6 | 1.6 | | | | | 120 | 1.7 |
| | | | | | | 121 | 1.6 |
| | | | | | | 122 | 1.5 |
| | | | | | | 123 | 1.8 |
| | | | | | | 124 | 1.8 |

## 1.5 FINAL STRUCTURE ANALYSIS

Figure 1 shows a schematic of the final structure as it is intended to be positioned in the membrane. Figure 2 shows a ribbon diagram of the final model. Only the 4 helical regions are included in the co-ordinate set. The helices are in the lipid region, with the caps in the head zone. The sequence and its relationship to the first two helices of BRH is shown in figure 3.
The features which distinguish transmembrane proteins can be described as follows:
1) Charged residues are on the extra cellular or cytoplasmic side of the membrane.
2) A large number of tryptophans and tyrosines are observed at these membrane interfaces.
3) The alpha helices are hydrophobic.
4) There is one polar inter-helix interaction between each pair of

helices.
5) The non-polar inter helix contacts are similar to those in soluble proteins.
All of these features were taken into account during the design. In particular, there are only 2 sequence changes between helices H1 and H2 and BRH. This was done to preserve the independent folding unit. There are 13 differences between H1′ and H2′ and BRH, introduced to allow core packing and additional polar interactions. The loop between H2 and H1′ is the epitope from Sendai virus, and the one between helices H1′ and H2′ is the epitope from c myc. No attempt has been made to model the loops or the short N terminal extra helical region. A C terminal region after helix 1′ is also needed.
Figure 4 shows the contact map of GRENDEL (4(a)), compared with that of ROP(4(b)) and BRH(4(c)), produce by CONAN (M.Schraf).
Polar interaction between different helices:
GLU 2 EO2 ... SER 139 OG (H1 <> H1′)
THR 17 OG1 ... THR 40 OG1 (H1 <> H2)
THR 39 OG1 ... THR 98 OG1 (H2 <> H2′)
THR 81 OG1 ... THR 136 OG1 (H2′ <> H1′)
THR 81 O ... THR136 OG1 (H2′ <> H1′)

### 1.5.1 Packing Analysis

Packing of the final structure was assessed using PACANA (J. Moult), using the ratio of the volume inaccessible to solvent to the volume of the van der Waals envelope. This packing efficiency is 0.43, compared with the ROP value of 0.47. Thus, by this criterion the core is well packed. There are are also no large water accessible cavities in the core.

### 1.5.2 Electrostatic Analysis

Weak and unfavourable electrostatic interactions were examined using ENEANA (J.Moult). There are 7 backbone peptide groups with a total electro- static energy of greater than -2 Kcal/mole. I.e. they make clearly inadequate interactions for a membrane environment. These are probably a consequence of the imperfect starting BRH structure, which has low effective resolution. To some extent, the energy minimization procedure succeeded in improving the electrostatics, but further attention in this area is required.

### 1.5.3 Surface Burial Analysis

The change in surface accessibility resulting from assembly of the four helices into a bundle is shown in Table 1. The largest contributors to burial are leu, met, thr, phe and tyr. The two hydrophilic residues are involved in the polar interactions stabilizing the helix interactions. The high burial of met and leu partly reflects the choice of core packing residues made in the

design.

### 1.5.4 Other Analysis

A homology search revealed no significant homology, other than that of the first pair of helices with BRH and the epitope sequences. Secondary structure prediction (figure 5) primarily predicts helix for the helical regions of the sequence, with the usual variations between methods.

## 1.6 COMMENTS ON TOOLS

Compared with the course of 4 years ago, there are have been large improvements in the software. The commercial graphics packages running on Irises provide much of the support needed for passive inspection of structures, and a convenient interface for energy minimization. They do not provide the tools necessary to build structure (motif construction, loop connection, core packing and so on). These seem to be specialized tools for protein design. Analysis tools are still primitive as well, with a need for more flexiable and friendly area, volume, and electrostatic analysis. We used primarily WHATIF, BIOGRAF and INSIGHT. We are most appreciative of Gert Vriend's help with WHATIF.

,hl 2 Conclusions

In many respects the design appears satisfactory. The helices contain appropriate tyeps of residue, there is good core packing, most hydrogen bonding requirements are fulfilled, and the geometry of the structure is adequate. Some further minor work on electrostatic interactions is required, but appears straightforward. More serious concerns are the ability of the structure to insert into the membrane, and whether the interactions in the head group regions are satisfactory. In spite of these reservations, we believe the structure does provide a basis for proceeding to experiment.

## BACTERIORHODOPSIN

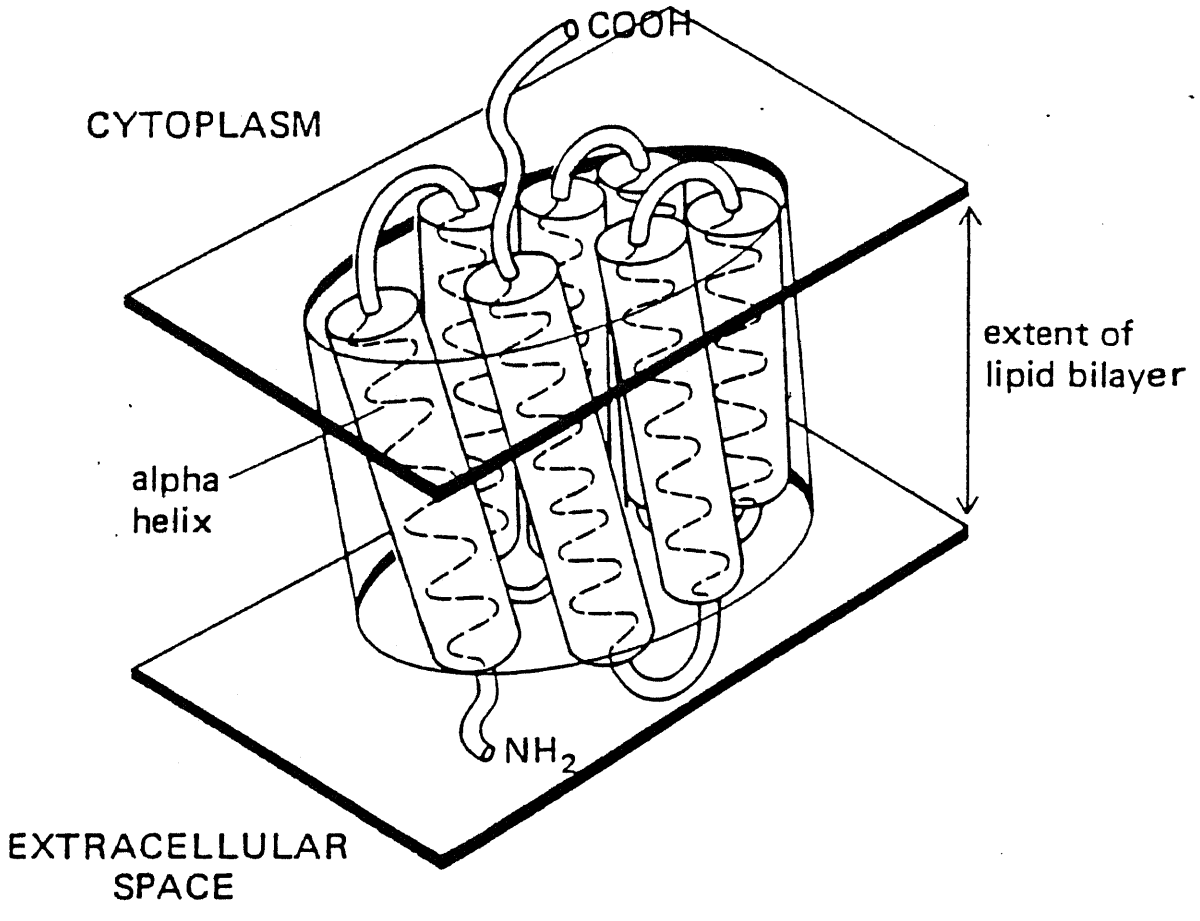



FIGURE LEGEND: Ia
Schematic drawing of one bacteriorhodopsin molecule and its relationship to the lipid bilayer. The polypeptide chain crosses the bilayer as seven α-helices.

## GRENDEL

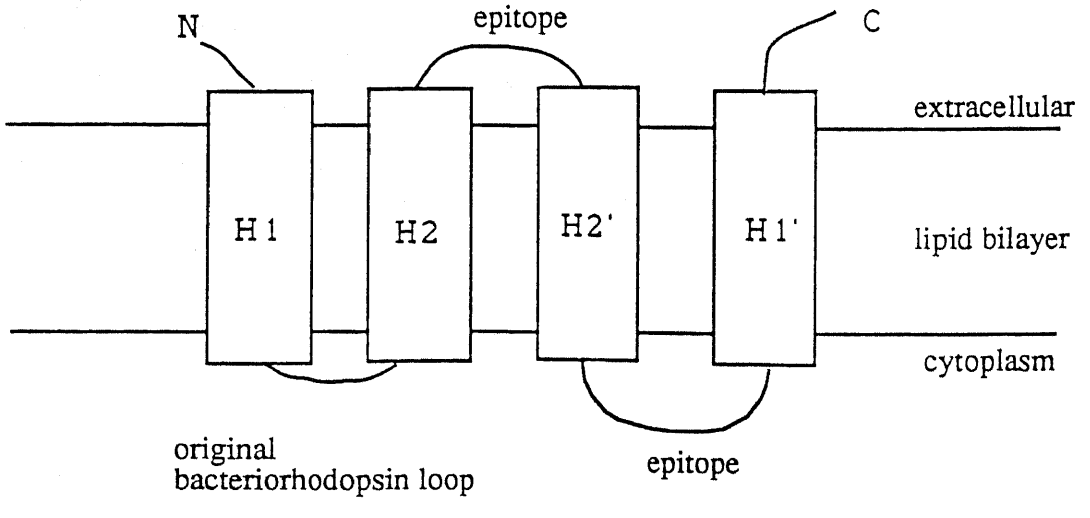
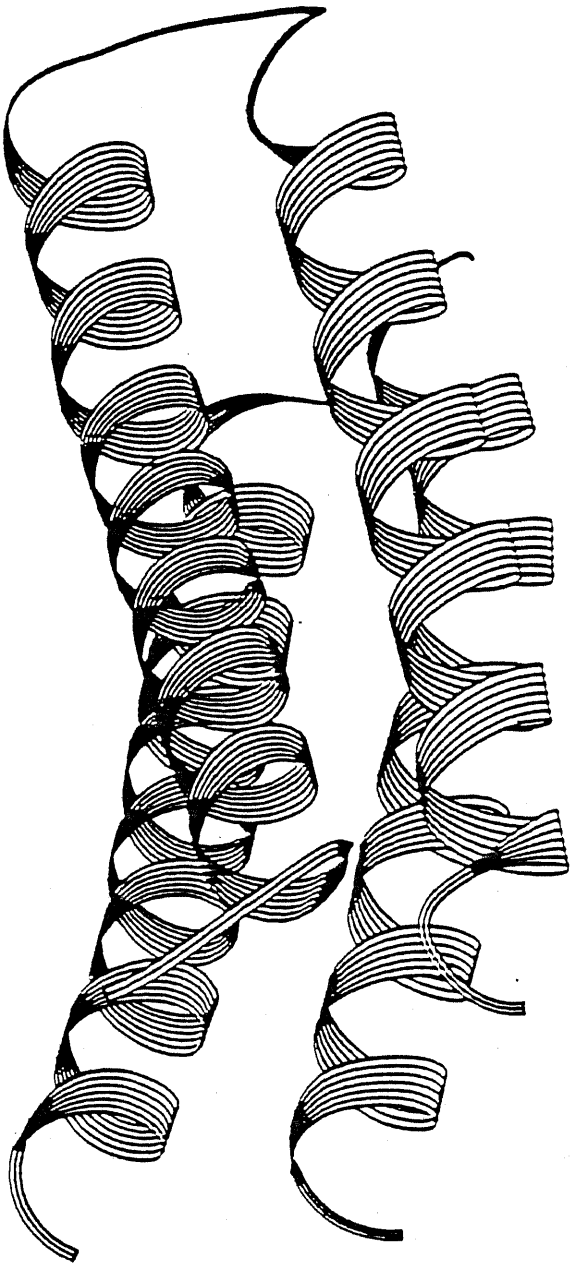



Figure legend: Ib
Schematic drawing of the GRENDEL design illustrating the 4 transmembrane helices

# GRENDEL DESIGN



# GRENDEL DESIGN

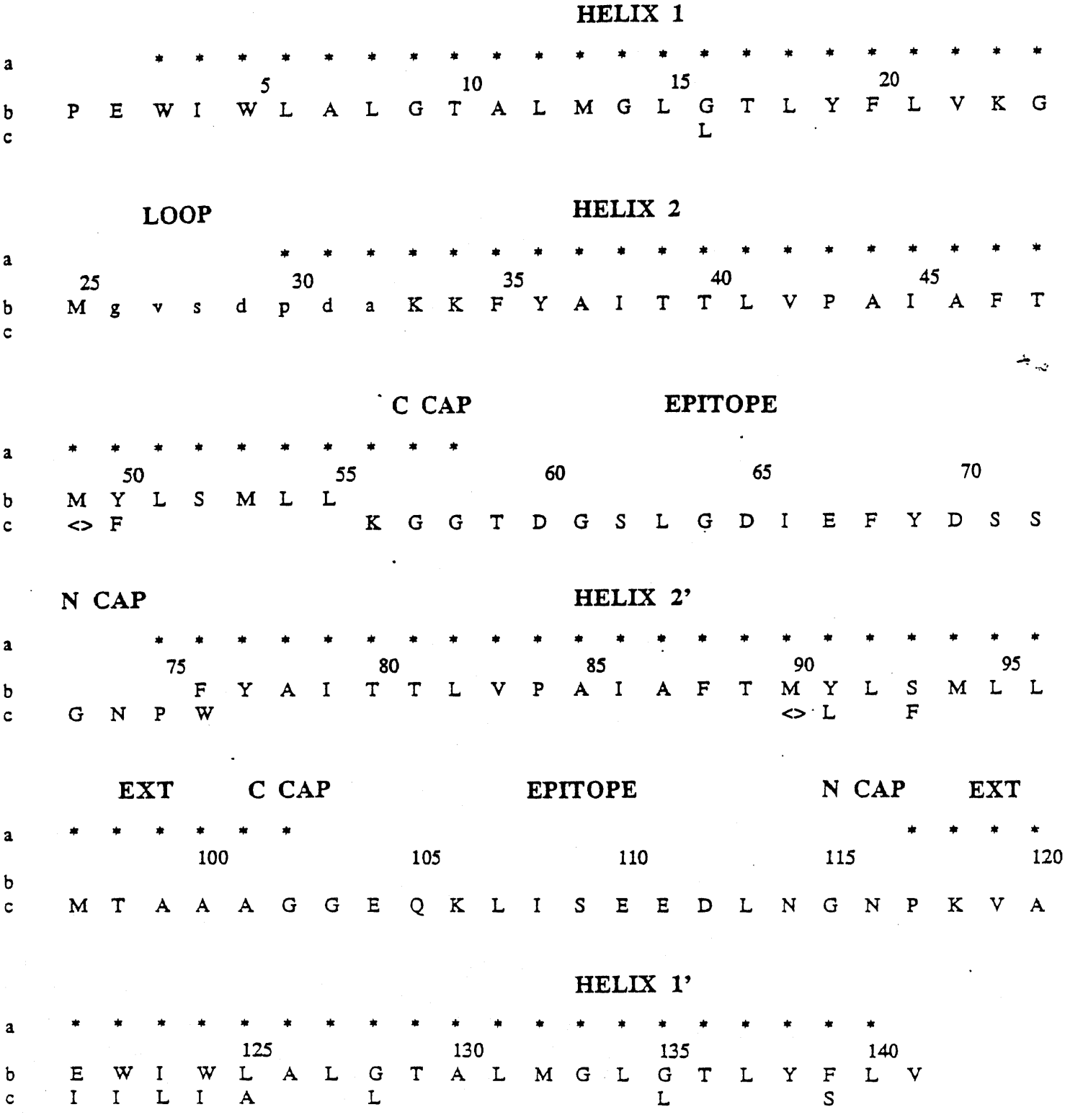
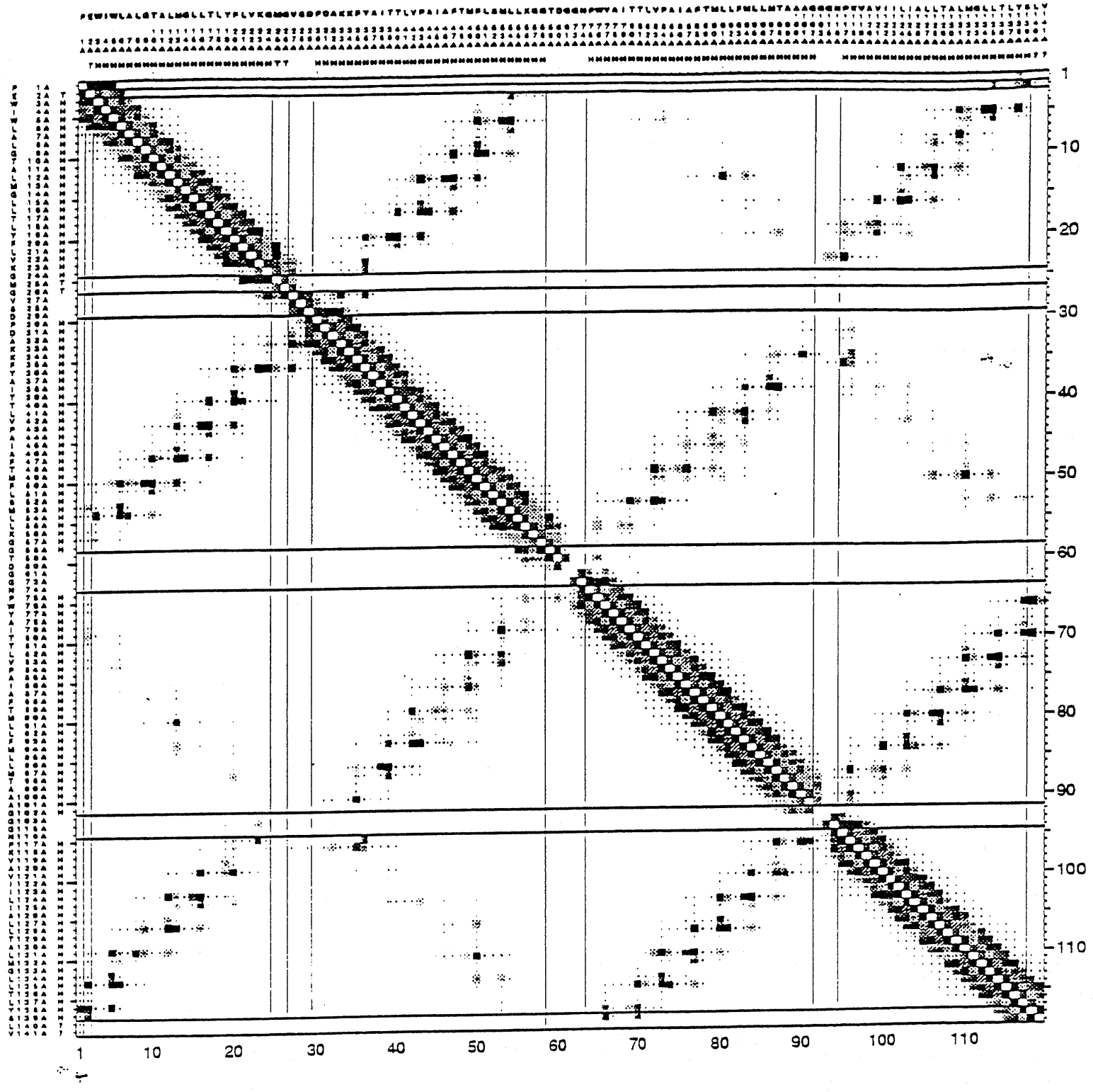



**Figure legend:** 3

GRENDEL DESIGN: Helices 1 and 2 are helices A and B of bacteriorhodopsin. Helices 1'and 2' were generated by duplicating helices A and B to create a four helix bundle. Residues on 1' and 2' were mutated to provide polar interactions between 1' and 1, 2' and 2, and to pack efficiently as a four helix bundle.

line a: helices (*), line b: Bacteriorhodopsin sequence (helices A & B), line c: substitutions, insertions and torsion angle changes (<>)

# RESIDUE CONTACTS IN GRENDEL

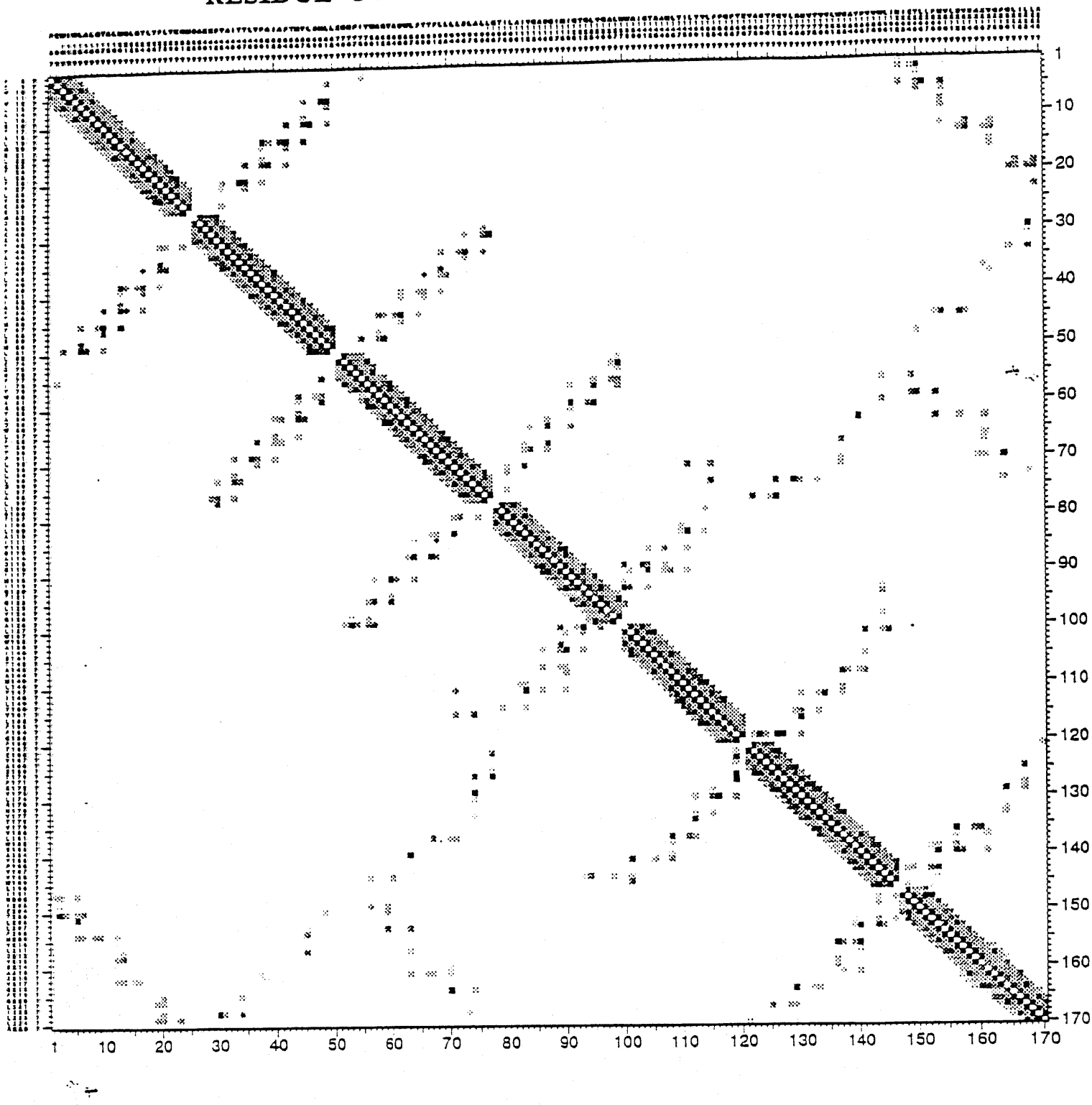



Van der Waals energy, n-conts = 3034

FIGURE LEGEND: **4a**

2 Dimensional residue contact map generated with CONAN (M. Scharf & C. Sander).
Top and left margins list the sequence, the corresponding numbers, the secondary structure (H=helix, T=turn)

# RESIDUE CONTACTS IN BACTERIORHODOPSIN

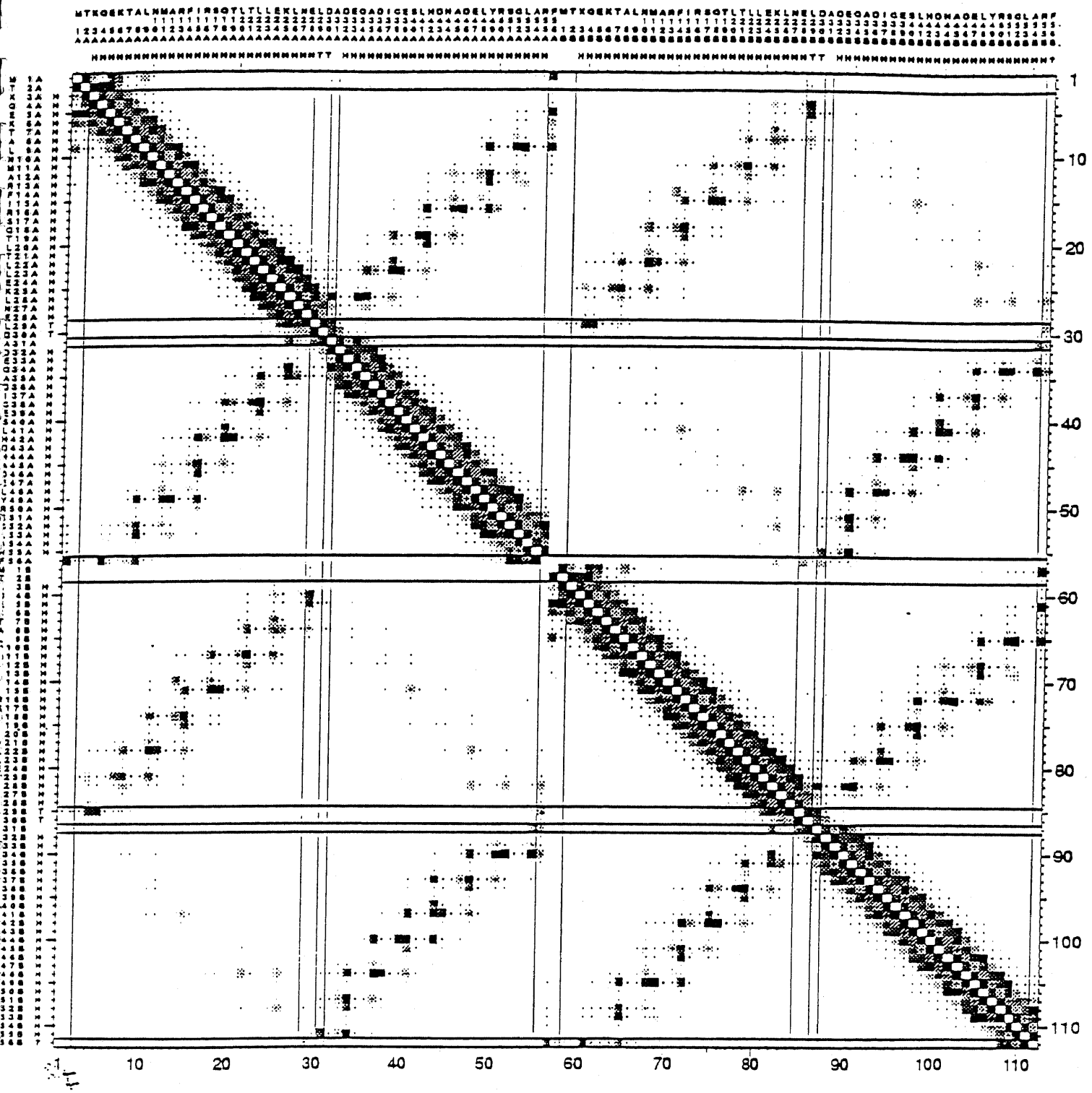



Van der Waals energy, n-conts = 2626

FIGURE LEGEND: **4b**
2 Dimensional residue contact map generated with CONAN (M. Scharf & C. Sander).
Top and left margins list the sequence, the corresponding numbers, the secondary structure (H=helix, T=turn)

kcal/mol

# RESIDUE CONTACTS IN ROP




Van der Waals energy, n-conts = 2886 .

FIGURE LEGEND: 4c

Dimensional residue contact map generated with CONAN (M. Scharf & C. Sander).
op and left margins list the sequence, the corresponding numbers, the secondary structure
=helix, T=turn)

kcal/mol

Secondary structure predictions by several different methods and comparison with secondary structures (DSSP) present in GRENDEL model.

```
GRENDEL.SEQ          LENGTH          141

   1  50  ....,....1....,....2....,....3....,....4....,....5
SEQ         PEWIWLALGTALMGLLTLYFLVKGMGVSDPDAKKFYAITTLVPAIAFTMF
DSSP         THHHHHHHHHHHHHHHHHHHHTT     HHHHHHHHHHHHHHHHHHHH
ROBSON       THHHHHH   HHEEHHHEEEEETT     HHHHHEEEEEEEEEHHHHHH
SEGMENT83    HHHHHHHHHHHHHHHHHHHHH            EEEE      HHHHHH
GOR          HHHHHHHHHHHHHHHHHEEE          HHHHHHHHH  HHHHHH
GORCLASS   2322334423353343233111242434431233333323312123353 3
ALB          HHHHHHHHHHHHHHHHHHHHH      TTT  EEEEEE  HHHHHHH
ALB*CLASS*2233333333333333333333212223222112222221123333333
HPHRUN        PPPP
CHARGE        -                  +     - - ++

  51 100  ....,....1....,....2....,....3....,....4....,....5
SEQ         LSMLLKGGTDGSLGDIEFYDSSGNPWYAITTLVPAIAFTMLLFMLLMTAA
DSSP        HHHHHHHH   ...........  HHHHHHHHHHHHHHHHHHHHHHHH
ROBSON      HHHH TT       EEEEETTT   TEEEEEEEEHHHHHHHHHHHHHH
SEGMENT83   HHHHH          EEEE       EEEE     HHHHHHHHHHHHHH
GOR         HHHH           EEEE        EEEEEH  HHHHHHHHHHHHHH
GORCLASS   4322125544542313353145554121221112334545554543333
ALB         HHHHH   TTTTT EEEE TTT HHHHHHHHHHHHHHHHHHHHHHHH
ALB*CLASS*3333332232222222222222221111233333333333333333333
HPHRUN                                                PPPPPPPP
CHARGE          +   -    - -  -

 101 141  ....,....1....,....2....,....3....,....4....,....5
SEQ         AGGEQKLISEEDLNGNPKVAVIILIALLTALMGLLTLYSLV
DSSP        HH ..........   HHHHHHHHHHHHHHHHHHHHHH
ROBSON      HHHHHHEEHHHTTT   HHEEEEEEHHHHHHHHHHEEEEEE
SEGMENT83   H  HHHHH          HHHHHHHHHHHHHHHHHHHHH
GOR           EEE HHHH     HHHEHHHHHHHHHHHHHHHH EH
GORCLASS   3421111113311355323311244555445324212111 1
ALB         HTT          TTT HHHHHHHHHHHHHHHHHHHHH
ALB*CLASS*2223332233332222221111133333333333333213
HPHRUN                        PPPPP
CHARGE        - +   ---     +
```

FIGURE LEGEND:5
programs by W.Kabsch,J.Lenstra,C.Oefner,B.Robson,J.Garnier,
D.J.Osguthorpe,J.-F.Gibrat,O.B.Ptitsyn,A.V.Finkelstein,
B.Altenberg,C.Sander

ALPHA STRUCTURE: H=HELIX
BETA STRUCTURE: E=EXTENDED
A,M,P=SURFACE,HALF BURIED,BURIED
LOOP STRUCTURE: T=TURN
C,BLANK=COIL OR UNDEFINED

SURFACE ACCESSIBILITY OF RESIDUES INVOLVED IN THE 4 HELIX BUNDLE

| Residue type | Number of residues involved | Av. accessibility individual helix | Av. accessibility complete 4 helix bundle | Total surface area buried per residue type |
|---|---|---|---|---|
| ALA | 15 | 16 | 12 | 77 |
| ASP | 1 | 23 | 24 | 0 |
| GLU | 1 | 34 | 11 | 24 |
| PHE | 6 | 43 | 18 | 156 |
| GLY | 8 | 14 | 9 | 46 |
| ILE | 8 | 31 | 24 | 63 |
| LYS | 5 | 39 | 24 | 81 |
| LEU | 24 | 34 | 20 | 326 |
| MET | 8 | 39 | 11 | 224 |
| PRO | 5 | 26 | 12 | 70 |
| SER | 2 | 19 | 8 | 22 |
| THR | 11 | 24 | 7 | 186 |
| VAL | 5 | 27 | 14 | 68 |
| TYR | 4 | 45 | 18 | 106 |

# DESIGN AND ANALYSIS OF NOVEL HANDSHAKE STRUCTURES

EMBL Protein Design Course 1990 - Course 2 Group

H. Nakamura*
A. Horovitz^
Ll. Ribas+
F. Bazan~

* Protein Engineering Research Institute,
Furuedai, Suita, Osaka, JPN
Tel: Japan-6-872-8212; Fax:Japan-6-872-8210; Email: nakamura@pes4.peri.co.jp.

^ MRC Unit for Protein Function and Design,
Univ. Chemical Lab.
Lensfield Rd., Cambridge CB1-2EW, UK.
Tel: ; Fax: ; Email: .

+ Dept. of Biochemistry, Medical School,
George Square, Edinburgh EH89XD,
Scotland, UK.
Tel. 6671011 (x2492). Email: LLuis@Biovax.ed.ac.uk

~ Dept. of Biochemistry and Biophysics,
Univ. of California, San Francisco,
CA 94143-0448, USA.
Tel: 415-476-5051; Fax: 415-476-1902; Email: Bazan@msg.ucsf.edu.

## 2.1 ABSTRACT DESIGN AND ANALYSIS OF NOVEL HANDSHAKE STRUCTURES

The recent observation of structural similarity between the protein folds of the MHC class I antigen headpiece domain (HLA), the platelet factor-4 (PF4) and interleukin-8 (IL8) cytokine structures and the MS2 bacteriophage coat protein has been the inspiration for the design of a family of novel alpha-beta proteins. This group of model proteins share a common motif of an antiparallel beta sheet that cradles a set of antiparallel alpha helices; typically this fold derives from the symmetric association of monomeric subunits. Because of its similarity to the action of two hands shaking, this fold has been termed the 'handshake' motif. The model handshake proteins differ among themselves and from the parent group of native proteins in their economy of sequence and variant strand-strand and strand-helix topologies.
The simplest structure (FingerClasp) is proposed to be a symmetrical dimer composed of two 37 amino acid subunits that individually display an antiparallel beta hairpin linked to a C-terminal helix. Packing side-by-side, the FingerClasp structure forms a minimal version of the HLA structure; however, the distance between helices is insufficient to form a deep groove as seen in the MHC molecule. The modeled tertiary structure was predicted to be similar to native protein structures in the packing of the hydrophobic core, electrostatic contacts and surface polarity. The ability of FingerClasp to fold into a stable protein will be tested by chemical synthesis of the amino acid chain; reporter groups of aromatic residues will indicate whether the protein has successfully folded in solution. Two other model structures derived from Fingerclasp, Lambada and a larger (beta)3-alpha Handshake, also exhibit this protein fold but will serve as scaffolds for the design of metal or peptide binding sites.

## 2.2 INTRODUCTION:

Analogous structures formed from unrelated sequences are particularly valuable guides to the design of novel proteins. We propose studying a rather simple folding motif formed by a pair of helices nestled in the curl of a single beta sheet. This fold is observed in the crystal and solution NMR structures of the headpiece (alpha1-alpha2) domain of the HLA class I molecule, the platelet factor-4 (PF4) and interleukin-8 (IL8) cytokine structures and the MS2 bacteriophage coat protein.
Nature has utilized symmetry to great advantage by forming these similar structures from two intercalated subunits that individually feature a four- stranded beta sheet and a single alpha helix. The "handshake" moniker for these disparate proteins describes the manner in which their subunits pack together, sheets side by side with the helix of one subunit overlaying the sheet of its symmetry-related partner. When viewed from the top, the antiparallel helices define a groo[Ave or trough on the surface of the beta sheet. In the case of the HLA structure (the best studied of the lot), the groove is the peptide antigen binding site; to this effect, the headpiece domain has followed the lead of antibody combining sites and lined the groove with aromatic residues that are well suited to packing around a

ligand.
The "function" of the HLA-like groove in the cytokine structures is not known. These small factors bind to specific cell-surface receptors; in addition, they uniformly display a great affinity for heparin. The MS2 fold is unique as the sole non-"jellyroll" viral capsid protein. In the capsid assembly, the dimer helix pairs (and the groove) face outward while the interior of the sheet is thought to interact with viral RNA.
Note: the handshake fold can be manually reproduced by positioning hands flat aside each other in such a manner that each index fingers packs underneath the thumb of the other hand. The fingers (eight in a neat row) form the beta sheet while the thumbs form the antiparallel helices atop the sheet. Keep trying until you get it, but be careful where you perform these digital contortions or your actions may be mistaken for a colorful local insult.

A list of original references:

1. Bjorkman et al. (1987) Structure of the human class I histocompatibility antigen, HLA-A2. Nature 329, 506
2. Bjorkman et al. (1987) The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. Nature 329,512
3. Brown et al. (1988) A hypothetical model of the foreign antigen binding site of class II histocompatibility molecules. Nature 332, 845
4. Garrett et al. (1989) Specificity pockets for the side chains of peptide antigens in HLA-Aw68. Nature 342, 692
5. St Charles, Walz and Edwards (1989) The three dimensional structure of bovine platelet factor 4 at 3.0 A resolution. J. Biol. Chem. 264, 2092
6. Clore et al. (1989) Determination of the secondary structure of interleukin-8 by nuclear magnetic resonance spectroscopy. J. Biol. Chem. 264, 18907
7. Clore et al. (1990) Three dimensional structure of interleukin-8 in solution. Biochem. 29, 1689
8. Wolpe and Cerami (1989) Macrophage inflammatory proteins 1 and 2: members of a novel superfamily of cytokines. FASEB J. 3, 2565
9. Valegard et al. (1990) The three dimensional structure for the bacterial virus MS2. Nature 345, 36

A further note on the five available crystal and NMR structures that serve as an initial dataset for this study:
HLA: Two X-ray structures are in the PDB, 2HLA and 3HLA. These are nearly identical structures of two MHC class I antigens, HLA-A2 and HLA-Aw68, that were solved in Don Wiley's laboratory at Harvard (see refs. 1-4 above).
PF4: The X-ray structure of the bovine PF4 dimer was solved in Brian Edwards laboratory at Wayne St. Univ. Medical School. Coordinates must be directly requested from Dr. Edwards.
IL8: The solution NMR structure of this cytokine was solved in Marius Clore's and Angela Gronenborn's laboratory at the NIH. Coordinates have been submitted to the PDB.
MS2: X-ray studies of the MS2 capsid are continuing at Uppsala in Lars Liljas' laboratory. Coordinates must be directly requested from Dr. Liljas.
We acknowledge with many thanks the kind support and interest of Drs. Edwards, Clore and Liljas in releasing their coordinates for this course, and for their continued interest in the analysis, methods and

results that have been generated in these hectic two weeks.

## 2.3 METHODS, TOOLS AND APPROACH: INITIAL ANALYSIS

The initial observation of gross three-dimensional similarity in the protein motifs of the HLA, PF4/IL8 and MS2 folds was the driving force for a more detailed analysis that would search for common and nonequivalent structural elements. It was obvious from a cursory examination of protein topologies that these diverse structures formed three distinct (i.e., non- homologous) classes. The HLA class I headpiece domain features a linked dimer of similar subunits and is the only protein in the dataset that has a pseudotwofold internal symmetry. The related HLA class II domain (for which no X-ray structure exists) is formed from non-linked subunits. The cytokine structures of PF4 and IL8 are closely superimposable, as expected from proteins that are over 50% identical. The MS2 protein forms the third group.

We can use homology to our advantage by first establishing a library of family sequences for each 'handshake' class. These aligned sequences will underscore the conserved and variable features of each structural motif mapped to the representative X-ray framework (with the aid of a DSSP roadmap). For this purpose, we utilized the compiled alignments of HLA class I and II subunits in the Brown et al. article (see above, ref. 4) and a cytokine alignment in the Walz and Cerami review (ref.8). An alignment of MS2 homologs was compiled with the aid of FASTA and LINEUP programs in the GCG suite.

Fig. 1: (A) Subunits of 2HLA alignment

note symbols: coni_1= residues conserved with other classI members where capital=absol conserved, @=polymorphic, small lett=not polymorphic, but replacement
so_ac1= solvent accessibility by DSSP, +=under 22
gr_face= residue sidechains facing groove - binding?

```
          1                                                  49
   A2_1   GSHSMRYFFT SVSRPGRGEP RFIAVGYVD. DTQFVRFDSD AASQRMEPRA
  Aw68_1  GSHSMRYFYT SVSRPGRGEP RFIAVGYVD. DTQFVRFDSD AASQRMEPRA
   Dssp1  ..eeeeeeee ee..ssss.. eeeeeeeet. teeeeeeett stt.s.ee.s
  Coni_1  ..H.....@. .........P ...@...VD. D.@F.RF... .enp.@....
  so_ac1  ....+.+.+. +.+......+ .+.+.+.+.. ...++.++.. +....+...+
 gr_face  ....+.+.+. .......... .+.+.+.... .......... ..........

         91                                                 140
   A2_2   GSHTVQRMYG CDVGSDWRFL RGYHQYAYDG KDYIALKEDL RSWTAADMAA
  Aw68_2  GSHTIQMMYG CDVGSDGRFL RGYRQDAYDG KDYIALKEDL RSWTAADMAA
    Dssp  ...eeeeeee eee.ttssee eeeeeeeett eeeee.ttss .eee.shhhh
  Coni_2  ....@.@.@. C...p.g.lL R.h@Q@.... cD..A.nEDL kt..AA.t.A
  so_ac2  ....+.+..+ ++++..+... .+.+..++.. ..++++...+ ..+.+...++
 gr_face  ....+.+..+ .......... ...+.+.... .......... ..........

          50                                                 90
   A2_1   PWIEQEGPEY WDGETRKVKA HSQTHRVDLG TLRGYYNQSE A
```

```
    Aw68_1  PWIEQEGPEY WDRNTRNVKA QSQTDRVDLG TLRGYYNQSE A
     Dssp1  gggtts.hhh hhhhhhhhhh hhhhhhhhhh hhhhtt..t t
     Coni_1 .W...@.... .er..qiaKg neQs@ResLr @@l@Y..... .
     so_ac1 ..+..+...+ ....+..+.. .+..+...+. .+........ .
    gr_face .........+ ...+..++.. +..++..+.. ++..+..... .

            141                                     177
      A2_2  QTTKHKWEAA .HVAEQLRAY LEGTCVEWLR RYLENGKE.. .
     Aw68_2 QTTKHKWEAA .HVAEQWRAY LEGTCVEWLR RYLENGKE.. .
       Dssp hhhhhhhhhh .thhhhhhhh hhthhhhhhh hhhhht.... .
     Coni_2 li.qr.WEq. .r@AE@@RA. .E@.CV..L. ...k..na.. .
     so_ac2 ..+..... ...+..... +.+.++..+. .+...+.... .
    gr_face ..+..++... .+..++..+. ..+...+... +......... .
```

(B) PF4/1L8 alignment

```
            1                                                   50
       PF4      QCVCLK TTS-GINPRH ISSLEVIGAG THCPSPQLLA TKKTGRKICL
      Dssp      ..s.s. ...-...ttt eeeeeee..s ss.ss.eeee eetts.eee.
       IL8  AKELRCQCIK TYSKPFHPKF IKELRVIESG PHCANTEIIV KLSDGRELCL
      Dssp  ......s.ss ...s...ggg eeeeeee.s. ss.ss.eeee eetttsssss
       Con      rCqCik t.s.gihpk. i.sl.vi..g phC..pevia tlk.grkiCl
     so_ac  .....+.... .......+.. +.++.+++.+ ..+...++++ .+......++
    gr_face  ......... ......... +..+.+..+. .......+.+ .........

            51                    71
       PF4  DQQR----PL YKKILKKLLD G
      Dssp  .sss----tt hhhhhhhhht .
       IL8  DPKENWVQRV VEKFLKRAEN S
      Dssp  .ttshhhhhh hhhhhhhhh. .
       Con  dp..pwvq.i v..llk.... .
     so_ac  ......+... +..+...+.. .
    gr_face  ......... .+..++..+. .
```

Walz and Cerami cytokine alignment:

```
HPF4     DGDLQCLCVKTTS-QVRPRHITSLEVIKAGPHCPTAQLIATLKN-GRKICLDLQA----PLYKKIIKKLLES
RPF4     DGDLSCVCVKTSSSRIHLKRITSLEVIKAGPHCAVPQLIATLKN-GSKICLDRQV----PLYKKIIKKLLES
mip1-1a  DTPTAC-CFSYQS-RKIPRQFIVDY-FETSSLCSQPGVIFLTKR-NRQICADSKETWVQEYITDLELNA
miAp-1B   DPPTSC-CFSYTS-RQLHRSFVMDY-YETSSLCSKPAVVFLTSR-GRQICANPSEPWVTEYMSDLELN
hLD78    DIPTAC-CFSYTS-RQIPQNFIADY-FETSSQCSKPGVIFLTKR-SRQVCADPSEEWVQKYVSDLELSA
hRANTES  SDTTPC-CFAYIA-RPLPRAHIKEY-FYTSGKCSNPAVVFVTRK-NRQVCANPEKKWVREYINSLEMS
mTCA3    TVSNSC-CLNTLK-KELPLKFIQCYRKMGSS-CDPPAVVFRLNK-GRESCASTNKTWVQNNLKKVNPC
hJE       APLTC-CYSFTS-KMIPMSRLESYKRITSSRCPKEAVVFVTKL-KREVCADPKKEWVQTYILNLDRN
HPBP     YAELRCMCIKTTS-GIHPKNIQSLEVIGKGTHCNQVEVIATLKD-GRKICLDPDAPRIKKIVQKKLAGDESAD
HIP10    SRTVRCTCISISNQPVNPRSLEKLEIIPASQFCPRVEIIATMKKKGEKRCLNPESKAIKNLLKAVSKEMSKRSP
CPE3     GNELRCQCISTHSKFIHPKSIQDVKLTPSGPHCKNVEIIATLKD-GREVCLDPTAPWVQLIVKALMAKAQLN
HGRO     ATELRCQCLQTL-QGIHPKNIQSVNVKSPGPHCAQTEVIATLKN-GRKACLNPASPIVKKIIEKMLNSDKSN
HAMGRO   ANELRCQCLQTM-TGVHLKNIESLKVTPPGPHCTQTEVIATLKN-GQEACLNPEAPMVQKIVQKMLKSGIRK
MKC      ANELRCQCLQTM-AGIHLKNIQSLKVLPSGPHCTQTEVIATLKN-GREACLDPEAPLVQKIVQKMLKGVPK
cons           C C                       C               C
```

(C) Alignment of MS2 with two homologs

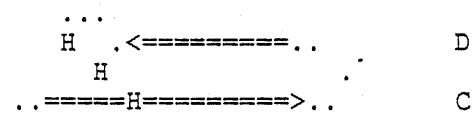
```
          1                                                        49
Prr1cp  .AQLQNLVLK DREATPNDHT FVPRDIRDNV GEVVESTGVP IGESRFTISL
  Gacp  MATLRSFVLV DNGGTGNVT. VVPVSNANGV AEWLSNNS.R SQAYRVTASY
 Ms2cp  ASNFTQFVLV NDGGTGNVT. VAPSNFANGV AEWISSNS.R SQAYKVTCSV
  DSSP  ......ees. ..ssss..e. eeeee..ss. eeeeess. g gg..eeeeee
  con   .a.l..fVLv d.ggTgNvt. vvP...angV aEw.ssns.r sqayrvT.S.

          50                                                       99
Prr1cp  RKTSNGRYKS TLKLVVPVVQ SQTVNGIVTP VVVRTSYVTV DFDYDARSTT
  Gacp  RASGADKRKY AIKLEVPKIV TQVVNGVELP GSAWKAYASI DLTIPIFAAT
 Ms2cp  RQSSAQNRKY TIKVEVPKVA TQTVGGVELP VAAWRSYLNL ELTIPIFATN
  DSSP  eeettteeee eeeeeeeeee eeeetteeee eeeeeeeeee eeeeess..h
  con  R.ssa..rKy tiKleVPkv. tQtVnGvelP v.aw.sY... dltipifatt

          100                  [B            132
Prr1cp  KERNNFVGMI ADALKADLML VHDTIVNLQG VY.
  Gacp  DDVTVISKSL AGLFK.VGNP IAEAISSQSG FYA
 Ms2cp  SDCELIVKAM QGLLK.DGNP IPSAIAANSG IY.
  DSSP  hhhhhhhhhh ttss..ttsh hhhhhttt.. ...
  con   .d...ivk.. agllK.dgnp i..aI...sG .Y.
```

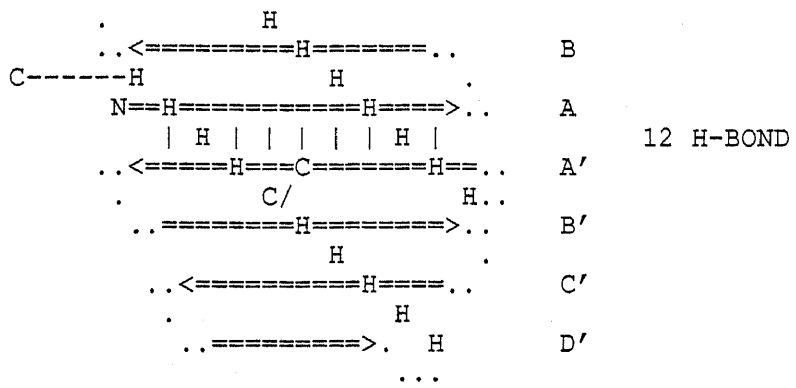
DSSP not only provides a map of the secondary structure of these proteins but also gives information about solvent accessibility of residues.
As the salient protein scaffold for these proteins is an eight-stranded, antiparallel beta sheet, the program BETASHEET is an excellent initial screen of the structures. The output details the rich hydrogen bonding interactions that characteristically stitch the sheet together, as well as providing a flat representation of sheet residues ('up' and 'down' facing).
CONAN was also part of the initial tests performed on the test group of structures (with the kind aid of M. Scharf). This program provides a very visual reference of distance and contact relationships in the folded structures by the use of residue contact maps. These can be tailored (by using DSSP files) to reveal special contacts between residues in packed secondary structures. Of interest to our efforts were the maps of the sheet-helix all-atom VDW contacts (see appendix).

## 2.4 STRUCTURAL FEATURES OF THE HANDSHAKE PROTEINS

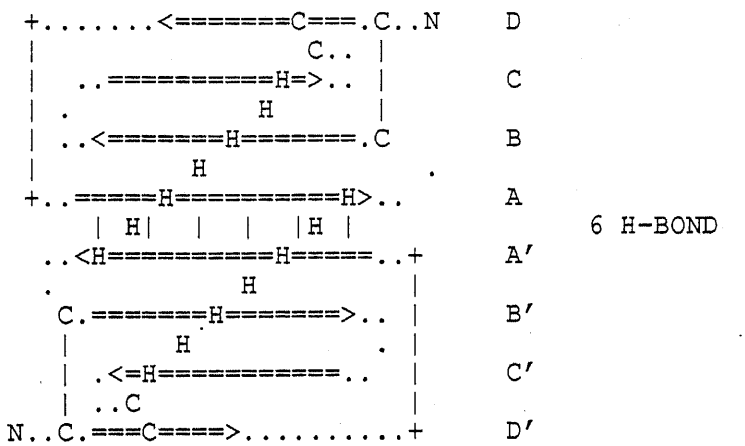
Representative folds of the three classes of eight-beta-stranded handshake structures are drawn below. Note that although they all share a common beta sheet, the strand topologies (i.e., the manner in which the strands are linked with each other and the helix) are quite different:

[ HLA-A2 and HLA-Aw68 ]

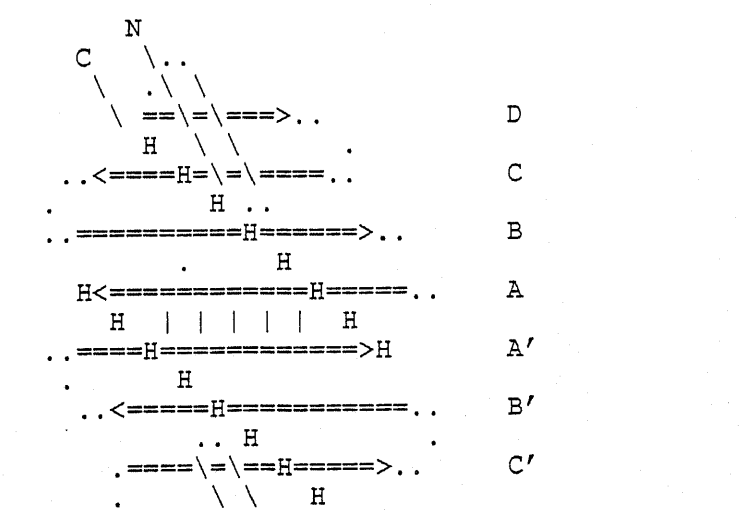


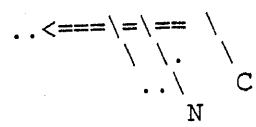



[ PF4 and IL8 ]




[ MS2 Coat Protein ]






Note that the order of strands in the protein chains for the three molecules is:
HLA: --A--B--C--D--helix ; PF4/IL8: --D--A--B--C--helix ;

MS2: --(-D)--(-C)--(-B)--(-A)--helix' *
The apparent reverse order of MS2 strands just indicates that the helix is packing on the reverse side of the sheet (in comparison to the HLA or PF4/IL8 structures.
The architecture of this handshake motif can be best described as an 'open beta-sandwich'. Jane Richardson describes a number of protein structures that are composed of helical structure packed on top of a beta sheet; however, the set of HLA, PF4/IL8 and MS2 structures distinguish themselves by a rather flat beta sheet. This sheet does conserve the characteristic right-handed curl of beta sheets, but is much less twisted than analogous structures in proteins such as T4 lysozyme domain 1, Strep. subtilisin inhibitor, hemmagluttinin, glutathione reductase domain 3, thermolysin domain 1 and glyceraldehyde-phosphate dehydrogenase domain 2.
An important feature of the handshake structures is their intrinsic symmetry. Of the PDB structures surveyed, only one 'open beta-sandwich' protein exhibited a flat sheet formed of a symmetric composite of simple alpha-beta units. This last structure, the regulatory domain 1 of aspartate transcarbamylase (ATC), will be fully described in a later section of this report.
Superposition of the HLA, PF4 and IL8 dimer structures shows that the center 4 strands, formed by strands A and B of individual subunits (see figure above), are closely matched. This section of the beta sheet is quite flat and contributes mostly to forming the floor of the groove. The edge two strands in PF4/IL8 are twisted away from the plane of the sheet in such a way that two things occur: (a) edge strand D (first in sequence in the cytokine structures) hangs off the plane of the sheet; in the case of PF4, this overhang forms the principal dimer contact. (b) Twisted strand D also connects to a loop that links to strand A; this loop curls up to form a lip to the flat beta sheet, effectively stabilizing the helix with a series of hydrophobic interactions.
The sheet is notably larger in the HLA structure because of longer component beta strands. The extent of H-bonding between strands establishes a more extensive network. As seen in PF4/IL8, the center two strands of each (nonidentical) subunit are flatter and contribute exclusively to forming the floor of the groove (see notation in above alignment). The edge two strands are distinctively curled away from the sheet plane. In analogy to the case of the cytokine loop, the helix is stabilized by an interaction with raised edges of the HLA sheet that are created by deformities in the edge strands (beta bulges, etc).
The helices of the handshake molecules rest on top of the flat beta sheets with their axes at an approximate 45 degree angle with respect to the direction of the beta strands. These helices are uniformly amphiphilic in order to pack against the hydrophobic sheet surface. Because of the size of the sheet and the degree of curl, the long HLA

helices must accomodate by kinking at their center. Reflecting their functional imperative, the HLA helices are lined with aromatic groups that point inwards toward the groove; in contrast, the PF4/IL8 helices feature exposed surfaces that display numerous basic residues (perhaps used to bind heparin).

## 2.5 INITIAL DESIGN CONCLUSIONS

A principal goal of this exercise in protein design is to understand the protein folding principles for a restricted set of structures that form analogous folding motifs with distinct sequences and chain topologies. What are the folding determinants for this class of structures? Sequence analysis of handshake homologs in light of the exisiting structural models sheds some light on specific structural roles of conserved residues. The variations in strand topology that we observe may indicate a relaxed requirement for inter- strand links outside of the stable protein scaffold. But each fold does have its peculiarities, small but significant structural differences in the sheet periphery and packing helix that stabilize that particular topological variant. An invariant feature remains the strong H-bond network between the central buried strands (one from each subunit) that forms the principal dimer interaction.

In addition, stabilization of the helix packing onto the beta sheet is imperative in order to guarantee the correct folding of a handshake protein. Following the examples of the HLA and cytokine structures, we must think of engineering deformities in the edge strands or strand-connecting loops (in addition to the placement of correct hydrophobic patches on the sheet surface) in order to tether the alpha helix. Things to consider include: (a) beta bulges that very reproducibly change the orientation of beta strands and may be useful in curling (forcibly) an edge strand to form a lip. (b) Consider having an edge parallel strand; this latter structure tends to form more curled structures than the antiparallel variety. This idea has topological repercussions.

In summary, the group decided to pursue the design of novel handshake structures whose assembly relies strongly on the principle of symmetry. A target size of ~70 residues is planned for a final model consisting of four beta strands and one alpha helix per subunit of the dimer. The topological restraints of the known handshake structures should not impose a strong barrier to the testing of alternative strand-strand and sheet-helix connectivities. The design should proceed in a modular fashion with the creation of an antiparallel beta sheet preceeding that of the helix and the various turn or loop segments. The assembly of these separate structural elements will require an understanding of long-range interactions between amino acids far apart in sequence.

The main attraction of this motif for protein engineering lies in its great versatility as a scaffold for the potential design of novel binding (receptor) or catalytic (enzyme) activity. The structure is uniquely suited to forming an extended cavity; in HLA, the groove between antiparallel helices is utilized by the MHC molecule to bind peptide ligands. In a similar fashion, a handshake groove could be designed be shape-complementary to special molecules like a drugs or

lipids. In a simplifying vein, this latter function may not require perfect packing around a prospective ligand structure, but rather, it may be sufficient to have a surrounding framework of aromatic, Ser and Arg residues (drawing from studies of amino acids commonly found in antibody combining sites and MHC grooves). These residues may form a plastic, hydrophobic and H-bonding pocket for a population of related ligands. Specificity may be engineered in by altering key topographical features of the groove as is observed for the HLA molecules (see Garrett et al. paper above).
Alternatively, an activity of some sort could be engineered into a handshake scaffold by constructing a catalytic site at one end or lip of the groove. Metal binding sites or geometrically precise groupings of amino acid residues would be required for this activity. A less ambitious goal would be just to produce a molecule that folds correctly and remain soluble in solution. Reporter groups could be cleverly placed in the protein so as to divulge (by spectroscopic analysis) the presence of correctly folded dimers in solution. Once this end has been achieved, the design of a receptor or enzyme could proceed with more confidence.

## 2.6 DESIGN QUESTIONS

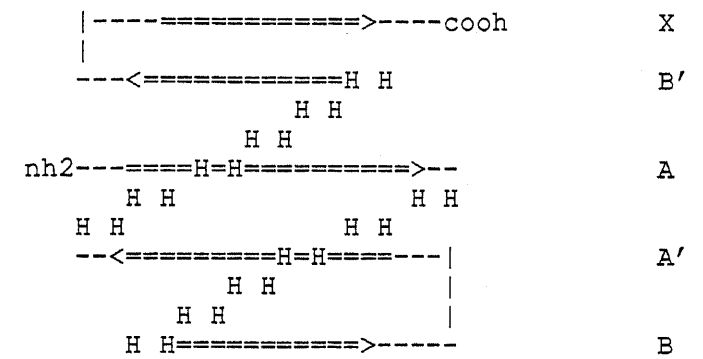
The building of a handshake structure (or structures) will require the design of several secondary structural frameworks. Do we want to use 'idealized' sheets and helices or construct templates from single or composite homolog/analog skeletons? It is most likely acceptable to utilize ideal alpha helices as these structures are very regular across a wide variety of protein structures. For the beta sheet, we have decided to utilize the best fit superposition of (beta)4-alpha subunits to generate 'average' coordinates for a rather flat sheet that seems to characterize this group of proteins. Turns will be located from loop libraries extracted from the PDB dataset.
As discussed before, the key design feature is to accurately design the helix-docking site for the sheets. The variant ways in which the helices and sheets pack in the HLA and PF4/IL8 structures (basically at 90 degrees with each other) illustrates how permissive a flat scaffold is to the positioning of packing structure. More highly curved sheets (as typically seen in Rossmann fold type structures or on the sides of beta barrels) would severely restrict the packing of helices to a few, well catalogued modes (see refs. below).
Refs.:

1. Cohen, Sternberg and Taylor (1982) Analysis and prediction of the packing of alpha helices against a beta-sheet in the tertiary structure of globular proteins. JMB 156, 821
2. Janin and Chothia (1980) Packing of alpha helices onto beta-pleated sheets and the anatomy of alpha/beta proteins. JMB 143, 95
3. Jane's magnum opus and '89 update in G. Fasmans structure prediction book.

A more directed survey of the PDB was made in order to analyze helix-(flat)-sheet interactions in existing protein structures. Several potentially interesting structures were found:
- Aspartate transcarbamylase regulatory subunit - domain 1 (ATC; 7atc.brk in PDB)

- Isocitrate dehydrogenase clasp domain (ICDH; submitted to PDB)
- Incredulase (Jane Richardson, in refinement and on the ProDes 90 T-shirts)

The ATC domain appears to have a very interesting twofold internal symmetry that is not evident in the amino acid sequence:

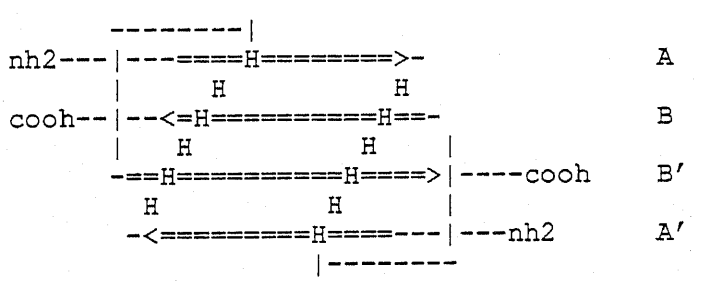



A few points to note: (a) The topology of the ATCase reg. domain 1 of PDB file 7ATC is a corrected version in which the chain topology has been changed. The present structure is a intercalated dimer of beta-alpha-beta (bab) subunits followed by an extra ('X'), edge strand. (b) In contrast to the HLA/PF4/IL8 group of structures, the ATCase structure features helices that are quite parallel to the the direction of the strands, similar to the disposition of similar elements in traditional Rossmann fold-type bab structures. (c) The helices lie on the same side of the sheet as in HLA/PF4/IL8. (d) True to its function as an allosteric effector domain, the ATCase reg. domain binds CTP in a wide cavity formed by the 'free' side of the beta sheet. (e) The entire ATCase domain consists of about 100 amino acids; however, if we remove the N- and C-terminal extensions and strand X, we are left with a minimal 80 amino acid structure.

Refs.:

1. Kim et al. (1987) Structural asymmetry in the CTP-liganded form of aspartate carbamoyltransferase from Escherichia coli. J. Mol. Biol. 196, 853.
2. Kantrowitz and Lipscomb (1988) Escherichia coli aspartate transcarbamylase: the relation between structure and function. Science 241, 669.

The ICDH 'clasp' domain is formed by an intercalated set of small protein motifs that are extrusions from much larger, packed catalytic subunits. This motif appears to form the smallest 'handshake'-type structure and features a novel strand/helix topology:



A few points to consider: (a) The helices of the ICDH domain are distinct from those seen in HLA/PF4/IL8 and most recently in ATCase: they are nested perpendicular to the strand direction. (b) In addition, the location of the helices on the sheet differs from that in the above structures: the ICDH helices are packed on the 'opposite' side of the sheet, reminiscent of MS2. (c) The domain subunits consist of about 40 amino acids. (d) The helices are close packed together.

Ref.:
1. Hurley et al. (1989) Structure of a bacterial enzyme regulated by phosphorylation, isocitrate dehydrogenase. Proc. Natl. Acad. Sci. USA 86, 8635.

## 2.7 FINAL STRATEGY FOR DESIGN

A three-stage approach was chosen as the most reasonable strategy. Analyses of the various families of handshake proteins define a hierarchy of structures:

| Structure Class | Member | Subunit Descrip* | Symmetry | Build | Symmetry | Variation |
|---|---|---|---|---|---|---|
| Complex | HLA<br>PF4/IL8<br>MS2 | 4-beta<br>helix<br>@ 45 | Str<br>Seq/Str<br>Seq/Str | beta4-a | Seq/Str | beta3-A |
| Intermediate | ATC | 2-beta<br>helix<br>@ 15 | Str | (bab)2 | Seq/Str | no |
| Simple | ICDH | 2-beta<br>helix<br>@ 90 | Seq/Str | beta2-a | Seq/Str | sheet up/<br>down link |

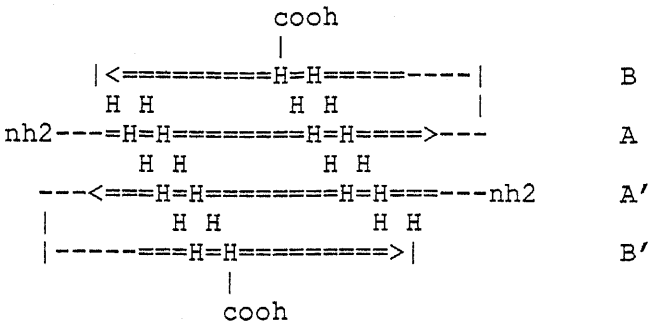
* Note that the numbers in this column refer to the approximate angles that the helices make with respect to strand direction.

Our hope is to learn more about this unusual and beautiful folding motif by first constructing simple examples, then topological variants and finally an 'elaborate' structure ripe with possibilities for the design of ligand-binding or catalytic sites.

## 2.8 RESULTS: DESIGN OF NOVEL PROTEIN SCAFFOLDS

### 2.8.1 Simple Structures: The Fingerclasp

Taking ICDH as the inspiration, we have constructed a simple 4-stranded dimer structure from two identical, 36 amino acid subunits that are beta2- alpha motifs. The chosen topology is (contrast with above ICDH connections):




## 2.8.2 Amino Acid Sequence

An amino acid sequence was chosen for the fingerclasp structure in such a manner as to provide a hydrophobic core between helices and sheet, good packing between helices and with the raised beta-loop. Again, ICDH provides a good example in filling the inter-helix groove with two stacked, symmetry- related Phe residues and size- and charge-complementary residues. In addition, the 'back' side of the sheet must be suitably hydrophilic for exposure to solvent and hindrance to unwanted sheet-sheet interactions. Hydrophobic patches on the sheet surface must be complementary to those on the hydrophobic side of the amphiphilic helix. The resultant sequence with a structural map is:

```
                    1         2         3
SEQUENCE. NDVKVDVRSNGGTKTEWKADSADLEKFAKQLKEKHGA
SUMMARY..   EEEEE STT EEEEE SS HHHHHHHHHHHHHHT
3-TURN...         >33<
4-TURN...                      >>>>XXXXXXX<<<<
5-TURN...                               >>555<<
BEND.....         SSS       SS SSSSSSSSSSSSS
CHIRALITY  ------++----+-----+-++++++++++++-
```

## 2.8.3 The Sheet

In order to design the strand, we made use of the superposition of the beta sheets of HLA, PF4 and IL8. Over a central region composed of four strands (two from each pair of subunits), all three structures showed very similar, slightly curled beta strands. An average c-alpha framework of this central sheet was then constructed using Composer (Mynfit) and a backbone with Ala sidechains was subsequently formed with the aid of MaxSprout. The H-bonding pattern of this sheet was strong and regular, comparing favorably with the more curled sheet of ICDH.
The composition of residues in the sheet was determined by four criteria: (i) Beta-sheet secondary structure preference parameters; (ii) solvent exposure of "bottom" facing side-chains in the sheet;

(iii) packing of the helices on top of the sheet; (iv) hydrophobic residues in the subunit interface to ensure correct formation of the dimer.
Residues 3,5 and 7 are hydrophobic and have high preference parameters for beta-sheet formation. Residues 2,4,6 and 8 are polar ("bottom" facing). Asp-2 is preferred at the start of a beta-strand. The identity of residues in strand A is determined mostly by criteria (i),(ii) and (iv). These residues are therefore predicted by secondary structure prediction algorithms to be in a beta-sheet. The identity of residues in strand B is also determined by criteria(iii). The packing of the two subunits is between two hydrophobic strands. A Trp residue at position 17 in the interface is introduced also to monitor folding as well as 'plug' the ends of the interhelix groove. Residues Thr-13 and Thr-15 were chosen because of their beta-sheet preference parameters. Threonines at positions 13 and 15 were preferred to valines because of solvent exposure. Residues 14 and 18 are polar since they are solvent exposed.

### 2.8.4 Linker Chains

The helix could lie on the sheet in two distinct ways because of the geometric constraints of curled beta sheets (resembling a saddle with diagonally opposite raised ends and dropped corners, respectively) diagonally Aopposite . In the case of HLA, the helix N-terminus is connected to an edge-located, downward-pointing strand D. The protein chain must then travel upwards before bending sideways to meet the helix N-terminus. PF4/IL8 utilizes a different strategy by connecting to helix using (effective) strand C whose end is 'up facing'. The resultant strand-helix connection is shorter (albeit somewhat unique given the disulfide-bridging pattern). A two-stranded form of the HLA topology on an ICDH-like platform was designed to make use of the short strand-up to helix link. If one examines the resultant structure, it is easy to note that the helix ends up in the opposite side of the sheet from that in HLA or PF4/IL8!
The choice of the strand-to-helix linker was influenced by the surveys of the Birkbeck group which identified four main types of b-a connections. The one we are interested in, labeled the beta-alpha-0 type loop, exhibits a typical phobic-polar-(Thr/Ser)-X-polar-X-phobic residue pattern (note the lack of glycine residues) and marks an approximate 90 degree turn from the strand direction to the helix axis. A turn such as this will enable the helix to lie approximately perpendicular to the strands in the sheet much like the conformation seen in ICDH. A quick survey of the ICDH structure reveals that the analogous loop is effectively of this type, with matching sequence ADSADL. It is easy to see that the Ser residue plays a critical role in bridging the turn with two H-bonds. As most of these ba0 turns are very superimposable, the ICDH loop is chosen as the 'donor' structure. Following an excision from the mother ICDH structure, the loop is quickly grafted onto our Fourbala framework. The remaining loop (beta hairpin between strands A-B) is found with the use of a conventional databank loop-search algorithm (FRGMNT, H. Nakamura). A best fit loop is chosen from cytochrome P450 (2CPP; this loop had the lowest rms deviation, - 0.7A - relative to the beta sheet framework) with the

sequence R-CNGG-H. We have slightly altered this sequence to avoid the possible formation of undesired disulfide bonds in the final model by changing the cysteine for a serine residue. This loop has the added advantage of furnishing a raised lip to the somewhat flat beta sheet (as observed in naturally occuring handshake structures) to further stabilize the helices.

### 2.8.5 Helix

The construction of a suitable helix is somewhat hindered by the inability of Insight to construct an ideal helix with good H-bonding distances. We finally settled on using a rather straight, well-shaped helix from Rop, grafting this onto the ba0-type loops on top of the sheet. The helix is terminated after 3 and a half turns by stitching on a suitable C-cap (see J.Richardson's reviews). Notable residues include His-35 (chosen to make a favorable interaction with Trp-17 in the hydrophobic core) and Gly-36, -136 that makes two H-bonds to the residues in the last turn of the helix. The N-cap residue is Asp-23 in the the loop. Glu-25, Lys-29 and Glu-33 are solvent exposed and might make i -> i+4 electrostatic interactions. Asp-23 and Lys-26 are also positioned so that they might form a solvent exposed electrostatic interaction. The remaining residues in the helix which pack against the beta-sheet (27,28,30 and 31) are hydrophobic with the exception of Gln-30 which is to some extent solvent exposed.
Following the above choices, each amino acid side chain was generated on the graphics screen using INSIGHT's replace option. Torsion angles of Trp17 and Phe27 in the both subunits were manually set to parameters observed for analogous residues in ICDH.

### 2.8.6 Energy Minimization Procedure

In order to get rid of bad contacts between atoms in the initial structures, the conformation energy of the model protein was minimized by the program PRESTO (H.Nakamura). All of the calculations were carried out in vacuo under the AMBER-united atom-force field with the dielectric constant of 2r(ij). r(ij) is here defined as the distance between atoms. The process of minimization was not straight forward because of several problems encountered in the course of the calculation.
(i) At first, 100 steps of steepest descent minimization was carried out, fixing all backbone atoms at their intial positions, with the repulsive forces proportional to squares of ( r(ij) - r0(ij) ) instead of the Lennard-Jones or Coulombic interaction forces. Here, r0(ij) is the equilibrium distance for two nonbonded atoms.
(ii) Starting from the final structure of the last step, 100 steps of steepest descent minimization was carried out, without any fixed atoms, with the Lennard-Jones and Coulombic interaction forces. By this calculation, most of bad contacts were repaired. Not only the total energy, but also individual energy terms were significantly decreased. The total energy decreased from 27,804 kcal/mol to -442 kcal/mol.

(iii) As the next step, a further 500 steps of conjugate gradient minimization was carried out, without any fixed atoms and with the LennardJones and Coulombic interaction forces. The total energy decreased from -442 kcal/mol to -699 kcal/mol.
(iv) Inspection of the final structure of the last step revealed that the orientation of Thr15 side chain of one subunit differed from the other. The OG1 atom was pointing inside to the protein core while the CG2 atom was outside. Therefore, the Chi-1 angle value was manually repaired to match the angle of the other subunit residue. Starting from this modified structure, 100 steps of steepest descent minimization was carried out, fixing all the backbone atoms, with the Lennard-Jones and Coulombic interaction forces.
(v) As the next step, a further 500 steps of conjugate gradient energy minimization was carried out, without any fixed atoms, with the Lennard-Jones and Coulombic interaction forces. The total energy then decreased from -754 kcal/mol to -918 kcal/mol.
(vi) The hydrogen bond networks at the N-caps of the alpha-helices are considered to be very important for maintainin the specific 'perpendicular' turns from the beta strands to the alpha-helices. These networks were broken in the structure resulting from the last minimization step. This is considered to be due to strong ionic interactions between the side chains of the two helices. Restraining distances were established between Ser21-OG and Lys18-N, and between Ser21-OG and Leu24-N as 2.9+-0.25 A; a further 500 steps of conjugate gradient minimization was then carried out without any fixed atoms and with the Lennard-Jones and Coulombic potentials. The total energy decreased from -889 kcal/mol to -940 kcal/mol.
(vii) Backbone dihedral angles were calculated for the final structure of the last step; one peculiar value (phi-angle= +79 degree ) was found at the N-terminus of one subunit. It was a careless mistake at the initial model building stage on the graphics screen. The torsion angle was repaired following the local structure of the other subunit (residues from 1 to 3).
(viii) Another 100 steps of conjugate gradient minimization was carried out, fixing all the atoms except those in the residues 1 to 3 of the 'repaired' subunit, with the Lennard-Jones and Coulombic potentials and the same distance restraints as used in the above step (vi).
(ix) Finally, 500 steps of conjugate gradient minimization was carried out, without any fixed atoms, with the Lennard-Jones and Coulombic potentials and with the distance restraints. The total energy decreased from -934 kcal/mol to -954 kcal/mol. There are abundant hydrogen bonds and salt-bridges found in this final structure, as indicated in Table 2. in the appendix.

### 2.8.7 Computational Analysis Of The Designed Sequence

Once the final chain for the FingerClasp molecule was decided, we subjected the sequence to a series of traditional sequence analyses. Secondary structure prediction, amino acid sequence alignments and hydrophobicity measurements were performed. The secondary structure prediction programs included in the package PREDICT suggested results slightly different from our model (recall that these algorithms are at

best 65% accurate). While all the algorithms used agreed in the prediction of the N-terminal strand A, they also failed in predicting any secondary structure in the center region of the sequence where the second strand is modelled (perhaps due to its structural role as a polar, edge strand). The C-terminal helix is strongly predicted in all cases. Loops were less well located since only two of the 4 algorithms used managed to predict the first loop ( Robson and ALB ) and none picked the second one (see appendix). The similarity search carried out by FASTA did not pick out any analogous sequences save for the loop fragment excised from ICDH.

### 2.8.8 Computational Analysis Of The Designed Structure

The diagnostic programs CONAN, QPACK ,POLDIAGNOSTICS and CONANA were run on the final structure. The analysis of the contacts between the secon- dary structures and residues of the model with CONAN (driven by Alfonso Valencia) revealed the expected pattern of linear contacts between adjacent helices in the model and a perpendicular pattern between the helices and the beta sheet. The intensity of the contacts appeared normal between the strands (no close, unfavorable contacts) but seemed to show that the helix- strand contacts were 'light'. The figure also reveals the intersubunit contacts between the central beta strands, as well as lighter contacts between these strands and the edge structures.Those contacts were not expected initially but we think that they might correspond to interactions between the lysine sidechains that are regularly distributed over the solvent-facing side of the sheet. QPACK confirmed that the general packing of the residues is correct although it showed a small number of cases where the volume covered was slightly small (underpacked) and others where there were slightly unfavorable contacts (overpacked). In general the second monomer showed a better packing of protein sidechains as assessed by QPACK. Our QPACK quality value was -12.6; normal structures display values from -5 to -50.
POLDIAGNOSTICS was used to analize the polarity of the protein, the values for the exterior side chains and maximum polarity being just over the values considered as top levels. The rest of parameters fall in the correct range of polarity according to this program. PACANA ( courtesy of J. Moult )gave again quite encouraging results. The program produces a final value based in the ratio :

$$\frac{\text{Vol not accessible to water and not in VDW radius}}{\text{total VDW radius}}$$

The value for the Fingerclasp model is 0.4 which is very close to values derived from naturally occurring proteins. The same program detected a cavity in the hydrophobic core of aproximately 1.5 water molecule volumes. In general, the results of those diagnostic programs are highly encouraging and seem to validate the proposed model. Only local and relatively small structural problems are found; these were largely corrected in the course of the minimization cycles.
Refs.
1. Edwards et al. (1987) Structural and sequence patterns in the loops of

BAB units. Prot. Engin. 1, 173.
2. Thornton et al. (1988) Analysis, design and modification of loop regions in proteins. Bioessays 8, 63.
3. Wilmot and Thornton (1988) Analysis and prediction of the different types of beta-turn in proteins. J. Mol. Biol. 203, 221.
4. Sibanda et al. (1989) Conformation of beta-hairpins in protein structures: A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. J. Mol. Biol. 206, 759.

## 2.9 INTERMEDIATE CLASS: THE "LAMBADA" PROTEIN

At the risk of sounding flippant, the name is actually quite descriptive of the chain contortions and close contacts that this protein must make in order to fold. An unusual intercalation of beta-alpha-beta units produces a final structure that is quite different from the typical Rossmann fold-type proteins and yet utilizes many of the same construction principles. The parent topology of the ATCase regulatory domain 1, after close examination of sequence, superposition of symmetry-related strands, etc, is revealed to be quite amenable to modeling with the existing framework used in the simple fingerclasp structure. The framework sheet and the ATCase sheet are almost perfect matches in curvature and amino acid length. The beta sheet scaffold is exactly the same that is utilized in the fingerclasp protein. The revised nature of the strand helix connections make necessary the design of new strand-helix and helix-strand linkers. For this we will follow the recomendations of the Birkbeck study on bab loops. A short beta-hairpin turn that connects the c-terminal end of subunit 1 to the N-terminal residue of subunit 2 was obtained by the FRGMNT program from human lysozyme (1LZ1). A schematic of the final topology can be seen from the cartoon of ATC if strand X is deleted. Refinement of this structure is proceeding.

## 2.10 COMPLEX CLASS: THE HANDSHAKE PROTEIN

Construction of a beta-3-alpha framework is proceeding using the FingerClasp structure as a point of departure.

## 2.11 CODA

For the purpose of studying the principles of protein folding, we have chosen a simple motif as a parent model for the design of a family of novel molecules. This report presents a detailed description of FingerClasp, a symmetrical dimer of 74 amino acids that is predicted to form a four-stranded beta sheet on which rest two antiparallel alpha helices. However, the process by which we arrived at this structure is perhaps of greater interest to prospective designers who may read these reports and judge our models. Our sample group of native folds (HLA, PF4, IL8 and MS2), found in a range of disparate proteins, served as the inspiration for the design of an initially elaborate eight-stranded sheet - helix protein (with beta-4-alpha handshake subunits). The pace of the course, adjustments to the computers and varied software packages, soon made it clear that a more methodical approach to the design of our protein was necessary. As a result, FingerClasp was designed as the simplest test of our understanding of the handshake fold. It directly utilized structural elements derived from its more elaborate cousins, but differed in its economy of chain and strand/helix topology. Two additional protein were designed that would slowly approach the elaborate eight-stranded, grooved model that we first discussed: Lambada utilizes the same

sheet and helix elements of FingerClasp but reconnects these in a novel way. This protein has the added property of being particularly suited to the addition of symmetric metal ion binding sites located at the C-terminal ends of the linked beta-alpha-beta subunits. The beta-3-alpha handshake structure would add an edge strand to the FingerClasp/Lambada framework (lengthening all strands by one to two residues)in an extension of the beta-meander topology seen in the simple FingerClasp model. This protein would also utilize the same linkers and helices of the FingerClasp structure but position these latter elements further apart so as to generate a real groove on the sheet surface. These latter pair of structures were not totally completed by course's end but will be finished (and added to the EMBL database) in the following months.

The FingerClasp structure appears to be a robust model as judged by a variety of diagnostic tests that analyze its hydrophobic core packing, electrostatic contacts and distribution, H-bonding and salt-bridging networks and surface polarity. While recognizable as a model structure, it appears to closely approach values found for native protein structures, in particular those which constitute its parent group of folds. Experimental verification of the folding of FingerClasp will take place at PERI by synthesizing the protein subunit chain and assaying for correct dimerization and solubility.

We entered this course with ambitious dreams of becoming real protein designers and engineers (afer all, we have a certificate to prove it!). Alyosha correctly reminded us of our limitations halfway through the workshop and suggested that a better description of our vocation was that of protein carpenters. In retrospect, and suitably humbled by the experience of designing a de novo protein structure, we would like to modify this allegory to place ourselves as sandbox protein architects. As we leave the EMBL, it is perhaps fitting to liken our situation to that of children awaiting a summer rain,wondering with some apprehension how their sandcastles will fare when doused with water. As these protein models move to the experimental stage, we are certain that the same jittery feelings (confidence, confidence!) will emerge as our first engineered genes are expressed, or the first chains synthesized. Our elegantly displayed,electronically massaged and brilliantly colored proteins must fold for themselves.

## 2.12 APPENDICES:

Table 1. DSSP output for the final structure of the fingerclasp model

```
**** SECONDARY STRUCTURE DEFINITION BY THE PROGRAM DSSP, VERSION OCT. 1988 **** DAT
REFERENCE W. KABSCH AND C.SANDER, BIOPOLYMERS 22 (1983) 2577-2637
HEADER    FINGER CLASP MODEL                          11-OCT-90    1CL
COMPND    FINGERCLASP
SOURCE    HUMAN BRAINS
AUTHOR    H.NAKAMURA,A.HOROVITZ,Ll.RIBAS,J.F.BAZAN
   74  2  0  0  0 TOTAL NUMBER OF RESIDUES, NUMBER OF CHAINS, NUMBER OF SS-BRIDGES(
 4376.0   ACCESSIBLE SURFACE OF PROTEIN (ANGSTROM**2)
   57 77.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(J)  , SAME NUMBER PE
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS IN     PARALLEL BRIDGES, SAME NUMBER PE
   19 25.7   TOTAL NUMBER OF HYDROGEN BONDS IN ANTIPARALLEL BRIDGES, SAME NUMBER PE
```

```
   0   0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-5), SAME NUMBER PE
   0   0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-4), SAME NUMBER PE
   0   0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-3), SAME NUMBER PE
   0   0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-2), SAME NUMBER PE
   0   0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-1), SAME NUMBER PE
   0   0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+0), SAME NUMBER PE
   0   0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+1), SAME NUMBER PE
   6   8.1   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+2), SAME NUMBER PE
   2   2.7   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+3), SAME NUMBER PE
  22  29.7   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+4), SAME NUMBER PE
   5   6.8   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+5), SAME NUMBER PE
1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 2
0  0  0  0  0  0  0  0  0  0  0  0  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  2  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
#  RESIDUE AA STRUCTURE BP1 BP2  ACC    N-H-->O    O-->H-N    N-H-->O    O-->H-N     TCO
  1    1   N              0   0   86     0, 0.0   19,-2.2     0, 0.0   46,-0.3    0.000
  2    2   D  E     -A    46  0A   17    44,-1.4   44,-1.2    17,-0.3    2,-0.4   -0.812
  3    3   V  E     -AB   45 17A    0    14,-2.3   14,-2.0    -2,-0.2    2,-0.5   -0.986
  4    4   K  E     -AB   44 16A   82    40,-2.1   40,-2.6    -2,-0.4    2,-0.4   -0.982
  5    5   V  E     -AB   43 15A    0    10,-2.7   10,-2.2    -2,-0.5    2,-0.5   -0.942
  6    6   D  E     -AB   42 14A   47    36,-2.7   36,-2.8    -2,-0.4    2,-0.6   -0.939
  7    7   V  E     -AB   41 13A    0     6,-2.8    6,-2.5    -2,-0.5    2,-2.4   -0.907
  8    8   R              +     0   0  141    32,-2.5    2,-0.3    -2,-0.6   32,-0.3   -0.418
  9    9   S  S >> S+     0   0    3    -2,-2.4    3,-1.8    30,-0.2    4,-0.5   -0.971
 10   10   N  T 34 S-     0   0  112    -2,-0.3   -1,-0.2     1,-0.3    2,-0.1    0.923
 11   11   G  T 34 S-     0   0   89    -3,-0.2   -1,-0.3     2,-0.1   -3,-0.1   -0.280
 12   12   G  T <4  -     0   0   36    -3,-1.8    2,-0.4    -5,-0.2   -4,-0.3    0.341
 13   13   T  E  <  -B    7  0A   37    -6,-2.5   -6,-2.8    -4,-0.5    2,-0.3   -0.896
 14   14   K  E     +B    6  0A  114    -2,-0.4    2,-0.3    -8,-0.2   -8,-0.2   -0.839
 15   15   T  E     -B    5  0A   37   -10,-2.2  -10,-2.7    -2,-0.3    2,-0.3   -0.911
 16   16   E  E     -B    4  0A   75    -2,-0.3    2,-0.5   -12,-0.2  -12,-0.2   -0.948
 17   17   W  E     -B    3  0A   17   -14,-2.0  -14,-2.3    -2,-0.3    3,-0.1   -0.806
 18   18   K              -     0   0  175    -2,-0.5    6,-0.3   -16,-0.2    5,-0.1   -0.352
 19   19   A  S     S+    0   0   52     1,-0.1  -17,-0.3     5,-0.1   -1,-0.2   -0.031
 20   20   D  S     S+    0   0  113   -19,-2.2    2,-0.3     1,-0.2   -1,-0.1    0.999
 21   21   S  S  > S-     0   0   34   -20,-0.1    4,-2.8     1,-0.1    5,-0.4   -0.951
 22   22   A  H  > S+     0   0   83    -2,-0.3    4,-2.3     1,-0.2    5,-0.2    0.933
 23   23   D  H  > S+     0   0   24     2,-0.2    4,-2.7     3,-0.2    5,-0.3    0.878
 24   24   L  H  > S+     0   0    1    -6,-0.3    4,-2.8     2,-0.2    5,-0.2    0.978
 25   25   E  H  X S+     0   0   92    -4,-2.8    4,-2.6     1,-0.2    5,-0.2    0.877
 26   26   K  H  X S+     0   0  102    -4,-2.3    4,-2.4    -5,-0.4   -1,-0.2    0.942
 27   27   F  H  X S+     0   0    3    -4,-2.7    4,-2.7     2,-0.2   -2,-0.2    0.951
 28   28   A  H  X S+     0   0    4    -4,-2.8    4,-2.7    -5,-0.3    5,-0.2    0.873
 29   29   K  H  X S+     0   0  112    -4,-2.6    4,-2.9    -5,-0.2    5,-0.4    0.956
 30   30   Q  H  X>S+     0   0   26    -4,-2.4    4,-2.2     1,-0.2    5,-1.2    0.886
 31   31   L  H  X>S+     0   0    0    -4,-2.7    5,-2.1     3,-0.2    6,-1.6    0.970
 32   32   K  H  <5S+     0   0   81    -4,-2.7    4,-0.3     4,-0.3   -2,-0.2    0.932
 33   33   E  H  <5S+     0   0  135    -4,-2.9   -3,-0.2    -5,-0.2   -2,-0.2    0.865
 34   34   K  H  <5S+     0   0  120    -4,-2.2   -3,-0.2    -5,-0.4   -2,-0.2    0.867
 35   35   H  T  <<S-     0   0   52    -5,-1.2   -3,-0.2    -4,-1.1   -4,-0.2    0.797
 36   36   G     <        0   0   68    -5,-2.1   -4,-0.3    -4,-0.3   -3,-0.1    0.798
 37   37   A              0   0   19    -6,-1.6   16,-0.1    -9,-0.2   15,-0.1    0.031
 38        !              0   0    0     0, 0.0    0, 0.0     0, 0.0    0, 0.0    0.000
```

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 39 | 101 | N | | | 0 | 0 | 89 | 0, 0.0 | 2,-0.2 | 0, 0.0 | -30,-0.2 | 0.000 |
| 40 | 102 | D | | - | 0 | 0 | 22 | -32,-0.3 | -32,-2.5 | 17,-0.2 | 2,-0.4 | -0.850 |
| 41 | 103 | V | E | -AC | 7 | 55A | 0 | 14,-2.7 | 14,-2.2 | -34,-0.2 | 2,-0.4 | -0.987 |
| 42 | 104 | K | E | -AC | 6 | 54A | 103 | -36,-2.8 | -36,-2.7 | -2,-0.4 | 2,-0.4 | -0.994 |
| 43 | 105 | V | E | -AC | 5 | 53A | 2 | 10,-2.7 | 10,-2.2 | -2,-0.4 | 2,-0.5 | -0.983 |
| 44 | 106 | D | E | -AC | 4 | 52A | 67 | -40,-2.6 | -40,-2.1 | -2,-0.4 | 2,-0.6 | -0.965 |
| 45 | 107 | V | E | -AC | 3 | 51A | 0 | 6,-2.4 | 2,-1.8 | -2,-0.5 | 6,-1.6 | -0.922 |
| 46 | 108 | R | E | +A | 2 | 0A | 129 | -44,-1.2 | -44,-1.4 | -2,-0.6 | 2,-0.4 | -0.324 |
| 47 | 109 | S | S | >> S+ | 0 | 0 | 1 | -2,-1.8 | 3,-1.9 | -46,-0.3 | 4,-0.5 | -0.967 |
| 48 | 110 | N | T | 34 S- | 0 | 0 | 98 | -2,-0.4 | 2,-0.3 | 1,-0.3 | -1,-0.2 | 0.939 |
| 49 | 111 | G | T | 34 S- | 0 | 0 | 89 | -3,-0.2 | -1,-0.3 | 2,-0.1 | -3, 0.0 | -0.278 |
| 50 | 112 | G | T | <4 - | 0 | 0 | 43 | -3,-1.9 | 2,-0.4 | -2,-0.3 | -4,-0.3 | 0.290 |
| 51 | 113 | T | E | < -C | 45 | 0A | 30 | -6,-1.6 | -6,-2.4 | -4,-0.5 | 2,-0.4 | -0.871 |
| 52 | 114 | K | E | -C | 44 | 0A | 120 | -2,-0.4 | 2,-0.3 | -8,-0.2 | -8,-0.2 | -0.818 |
| 53 | 115 | T | E | +C | 43 | 0A | 8 | -10,-2.2 | -10,-2.7 | -2,-0.4 | 2,-0.3 | -0.940 |
| 54 | 116 | E | E | -C | 42 | 0A | 103 | -2,-0.3 | 2,-0.4 | -12,-0.2 | -12,-0.2 | -0.959 |
| 55 | 117 | W | E | -C | 41 | 0A | 16 | -14,-2.2 | -14,-2.7 | -2,-0.3 | 2,-0.2 | -0.970 |
| 56 | 118 | K | | - | 0 | 0 | 175 | -2,-0.4 | 6,-0.3 | -16,-0.2 | -16,-0.2 | -0.455 |
| 57 | 119 | A | S | S- | 0 | 0 | 54 | -18,-0.2 | -17,-0.2 | 1,-0.2 | -1,-0.2 | 0.096 |
| 58 | 120 | D | S | S+ | 0 | 0 | 100 | 1,-0.2 | 2,-0.3 | -3,-0.1 | -1,-0.2 | 0.982 |
| 59 | 121 | S | | > - | 0 | 0 | 37 | 1,-0.1 | 4,-2.9 | -4,-0.1 | 5,-0.4 | -0.914 |
| 60 | 122 | A | H | > S+ | 0 | 0 | 81 | -2,-0.3 | 4,-2.0 | 1,-0.2 | 5,-0.1 | 0.915 |
| 61 | 123 | D | H | > S+ | 0 | 0 | 18 | 2,-0.2 | 4,-2.7 | 3,-0.2 | 5,-0.3 | 0.824 |
| 62 | 124 | L | H | > S+ | 0 | 0 | 0 | -6,-0.3 | 4,-2.7 | 2,-0.2 | -2,-0.2 | 0.986 |
| 63 | 125 | E | H | X S+ | 0 | 0 | 82 | -4,-2.9 | 4,-2.8 | 1,-0.2 | 5,-0.2 | 0.843 |
| 64 | 126 | K | H | X S+ | 0 | 0 | 103 | -4,-2.0 | 4,-2.4 | -5,-0.4 | -1,-0.2 | 0.962 |
| 65 | 127 | F | H | X S+ | 0 | 0 | 2 | -4,-2.7 | 4,-2.5 | 2,-0.2 | -2,-0.2 | 0.908 |
| 66 | 128 | A | H | X S+ | 0 | 0 | 6 | -4,-2.7 | 4,-3.0 | -5,-0.3 | -1,-0.2 | 0.946 |
| 67 | 129 | K | H | X>S+ | 0 | 0 | 120 | -4,-2.8 | 4,-3.0 | 1,-0.2 | 5,-0.5 | 0.902 |
| 68 | 130 | Q | H | X>S+ | 0 | 0 | 30 | -4,-2.4 | 4,-2.3 | -5,-0.2 | 5,-1.2 | 0.941 |
| 69 | 131 | L | H | X>S+ | 0 | 0 | 0 | -4,-2.5 | 5,-1.9 | 3,-0.2 | 6,-1.7 | 0.944 |
| 70 | 132 | K | H | <5S+ | 0 | 0 | 65 | -4,-3.0 | -2,-0.2 | 4,-0.3 | -1,-0.2 | 0.893 |
| 71 | 133 | E | H | <5S+ | 0 | 0 | 124 | -4,-3.0 | -3,-0.2 | -5,-0.2 | -2,-0.2 | 0.710 |
| 72 | 134 | K | H | <<S+ | 0 | 0 | 120 | -4,-2.3 | -3,-0.2 | -5,-0.5 | -2,-0.1 | 0.844 |
| 73 | 135 | H | T | <<S- | 0 | 0 | 55 | -5,-1.2 | -3,-0.2 | -4,-0.8 | -4,-0.2 | 0.785 |
| 74 | 136 | G | | < | 0 | 0 | 69 | -5,-1.9 | -4,-0.3 | -6,-0.3 | -3,-0.2 | 0.828 |
| 75 | 137 | A | | | 0 | 0 | 26 | -6,-1.7 | -1,-0.2 | -9,-0.2 | -2,-0.2 | -0.917 |

Table 2. Hydrogen bonds and salt-bridges found in the final refined structure.

| atom (i) | | atom (j) | atom (i) | | atom (j) |
|---|---|---|---|---|---|
| Asn1A-ND2 | ... | Arg8B-O | Asn1B-N | ... | Asp20B-OD1 |
| Asn1A-N | ... | Asp20A-OD1 | Asp2B-OD2 | ... | Ala19B-N |
| Asp2A-O | ... | Arg8B-N | Val3B-N | ... | Trp17B-O |
| Asp2A-OD1 | ... | Arg8B-NH2 | Val3B-O | ... | Trp17B-N |
| Asp2A-OD2 | ... | Ala19A-N | Val5B-N | ... | Thr15B-O |
| Val3A-N | ... | Trp17A-O | Val5B-O | ... | Thr15B-N |
| Val3A-O | ... | Trp17A-N | Val7B-N | ... | Trp13B-O |
| Lys4A-N | ... | Asp6B-O | Val7B-O | ... | Trp13B-N |
| Lys4A-NZ | ... | Asp6B-OD1 | Val7B-O | ... | Ser9B-N |
| Lys4A-O | ... | Asp6B-N | Ser9B-O | ... | Gly12B-N |
| Val5A-N | ... | Thr15A-O | Lys14B-NZ | ... | Glu16B-OE1 |
| Val5A-O | ... | Thr15A-N | Lys14B-NZ | ... | Glu16B-OE2 |
| Asp6A-N | ... | Lys4B-O | Thr15B-OG1 | ... | Glu16B-N |

```
Asp6A-OD1  ... Lys4B-NZ
Asp6A-O    ... Lys4B-N
Val7A-N    ... Thr13A-O
Val7A-O    ... Thr13A-N
Val7A-O    ... Ser9A-N
Arg8A-N    ... Asp2B-O
Arg8A-NH2  ... Asp2B-OD1
Arg8A-O    ... Asn1B-ND2
Ser9A-O    ... Gly12A-N
Asn10A-OD1 ... Lys32B-NZ
Thr13A-OG1 ... Ala37B-OXT
Lys14A-NZ  ... Glu16A-OE1
Thr15A-OG1 ... Glu16A-N
Lys18A-O   ... Ser21A-OG1
Asp20A-N   ... Asp20A-OD1
Ser21A-OG  ... Leu24A-N
Ser21A-O   ... Glu25A-N
Ala22A-O   ... Lys26A-N
Asp23A-OD1 ... Lys34B-NZ
Asp23A-OD2 ... Lys34B-NZ
Asp23A-O   ... Phe27A-N
Leu24A-O   ... Ala28A-N
Glu25A-N   ... Glu25A-OE2
Glu25A-OE1 ... Ser21A-N
Glu25A-OE2 ... Ser21A-N
Glu25A-O   ... Lys29A-N
Lys26A-NZ  ... Gln30B-OE1
Lys26A-O   ... Gln30A-N
Phe27A-O   ... Asn31A-N
Ala28A-O   ... Lys32A-N
Lys29A-O   ... Glu33A-N
Gln30A-OE1 ... Lys26B-NZ
Gln30A-O   ... Lys34A-N
Leu31A-O   ... His35A-N
Leu31A-O   ... Gly36A-N
Leu31A-O   ... Ala37A-N
Lys32A-NZ  ... Asn10B-OD1
Lys34A-NZ  ... Asp23B-OD2
Ala37A-OXT ... Thr13B-OG1
Lys18B-O   ... Ser21B-OG
Asp20B-N   ... Asp20B-OD1
Ser21B-N   ... Glu25B-OE1
Ser21B-N   ... GLu25B-OE2
Ser21B-OG  ... Leu24B-N
Ser21B-O   ... Glu25B-N
Ala22B-O   ... Lys26B-N
Asp23B-O   ... Phe27B-N
Leu24B-O   ... Ala28B-N
Glu25B-O   ... Lys29B-N
Lys26B-O   ... Glu30B-N
Phe27B-O   ... Leu31B-N
Ala28B-O   ... Lys32B-N
Lys29B-O   ... Glu33B-N
Gln30B-O   ... Lys34B-N
Leu31B-O   ... His35B-N
Leu31B-O   ... Gly36B-N
Leu31B-O   ... Ala37B-N
Lys32B-N   ... Ala37B-O
```

Table 3. Output of POLDIAGNOSTICS

```
**** POLAR diagnostics list BY PROGRAM Pol_DIAGN, VERSION march  88 ***.
HEADER    FINGER CLASP MODEL                          11-OCT-90    1CLS
**** Reference C. Froemmel J.theor. Biol.  111 (1984) 247-260 .
**** Reference Baumann, Froemmel, Sander 1989 Prot.Eng.2 p329-334.
* ------------------------------------------------------------------------ *
** The chain length          =       74
** The molecular weight       =   8184.0
* ======================================================================== *
                in this prot|      in 150 proteins
                   observed | lowest  mean  highest value  decision
* ------------------------------------------------------------------------ *
            MW        110.595  97.000 109.000 115.000          OK   !
       MAX_POL          0.186   0.163   0.173   0.185        WRONG  !
       ACC_SRF          0.389   0.320   0.400   0.520          OK   !
```

| | | | | | |
|---|---|---|---|---|---|
| INT_POL | 0.186 | 0.158 | 0.172 | 0.190 | OK ! |
| OUT_POL | 0.185 | 0.147 | 0.172 | 0.204 | OK ! |
| SC_INT | 0.124 | 0.074 | 0.098 | 0.118 | WRONG ! |
| SC_OUT | 0.152 | 0.115 | 0.143 | 0.183 | OK ! |

Table 4. Output of QPACK ( Only CA atom information is extracted. )

```
HEADER    FINGER CLASP MODEL                          10-OCT-90    1CLS
COMPND    FINGERCLASP
SOURCE    HUMAN BRAINS
ATOM      2  CA  ASN     1      63.003  58.255  13.444  1.80  0.75
ATOM      7  CA  ASP     2      61.226  60.071  10.557  1.87  0.79
ATOM     12  CA  VAL     3      61.197  60.231   6.711  2.05  0.82
ATOM     17  CA  LYS     4      61.154  63.196   4.279  2.12  0.80
ATOM     22  CA  VAL     5      61.095  62.326   0.552  2.18  0.87
ATOM     27  CA  ASP     6      61.482  64.405  -2.643  1.89  0.80
ATOM     32  CA  VAL     7      60.654  62.508  -5.858  1.98  0.79
ATOM     37  CA  ARG     8      60.974  64.344  -9.184  2.81  1.11
ATOM     42  CA  SER     9      58.074  62.399 -10.727  1.83  0.80
ATOM     47  CA  ASN    10      54.680  63.722 -11.926  2.55  1.05
ATOM     52  CA  GLY    11      56.116  67.241 -11.596  1.91  0.94
ATOM     57  CA  GLY    12      58.336  66.536  -8.575  1.91  0.94
ATOM     62  CA  THR    13      56.425  65.887  -5.359  2.33  0.97
ATOM     67  CA  LYS    14      57.197  65.386  -1.669  2.67  1.01
ATOM     72  CA  THR    15      55.694  63.268   0.998  2.58  1.07
ATOM     77  CA  GLU    16      56.806  62.815   4.617  2.56  1.01
ATOM     82  CA  TRP    17      56.204  59.837   6.895  2.43  1.07
ATOM     89  CA  LYS    18      55.854  60.489  10.603  3.42  1.29
ATOM     94  CA  ALA    19      58.166  59.157  13.299  2.16  0.96
ATOM     99  CA  ASP    20      58.134  55.382  13.829  2.29  0.96
ATOM    104  CA  SER    21      55.535  54.008  11.395  2.07  0.90
ATOM    109  CA  ALA    22      55.062  51.096   8.966  2.55  1.13
ATOM    114  CA  ASP    23      56.262  53.338   6.100  2.18  0.92
ATOM    119  CA  LEU    24      58.941  54.946   8.327  2.30  0.90
ATOM    124  CA  GLU    25      60.659  51.672   9.303  2.31  0.91
ATOM    129  CA  LYS    26      59.904  49.985   5.931  3.12  1.18
ATOM    134  CA  PHE    27      61.583  52.761   3.958  2.34  1.03
ATOM    140  CA  ALA    28      64.521  53.295   6.358  2.49  1.11
ATOM    145  CA  LYS    29      65.075  49.506   6.279  2.15  0.81
ATOM    150  CA  GLN    30      65.062  49.490   2.444  2.34  0.93
ATOM    155  CA  LEU    31      67.488  52.449   2.477  2.30  0.91
ATOM    160  CA  LYS    32      70.056  50.898   4.822  2.72  1.03
ATOM    165  CA  GLU    33      69.790  47.245   3.782  2.07  0.81
ATOM    170  CA  LYS    34      69.343  47.602   0.007  2.27  0.86
ATOM    175  CA  HIS    35      70.674  51.047  -0.909  2.50  1.07
ATOM    181  CA  GLY    36      73.351  51.639   1.760  1.98  0.98
ATOM    186  CA  ALA    37      71.593  54.565   3.485  2.20  0.98
TER
ATOM    191  CA  ASN   101      62.747  61.599 -12.896  1.94  0.80
ATOM    196  CA  ASP   102      65.496  61.085 -10.206  1.95  0.82
ATOM    201  CA  VAL   103      64.815  61.122  -6.451  1.98  0.79
ATOM    206  CA  LYS   104      66.310  62.765  -3.342  2.11  0.80
ATOM    211  CA  VAL   105      65.407  61.691   0.222  2.18  0.87
ATOM    216  CA  ASP   106      66.286  62.950   3.723  1.91  0.80
ATOM    221  CA  VAL   107      65.500  60.416   6.467  2.05  0.82
ATOM    226  CA  ARG   108      65.807  61.086  10.213  1.99  0.79
ATOM    231  CA  SER   109      66.629  57.402  10.823  1.71  0.75
```

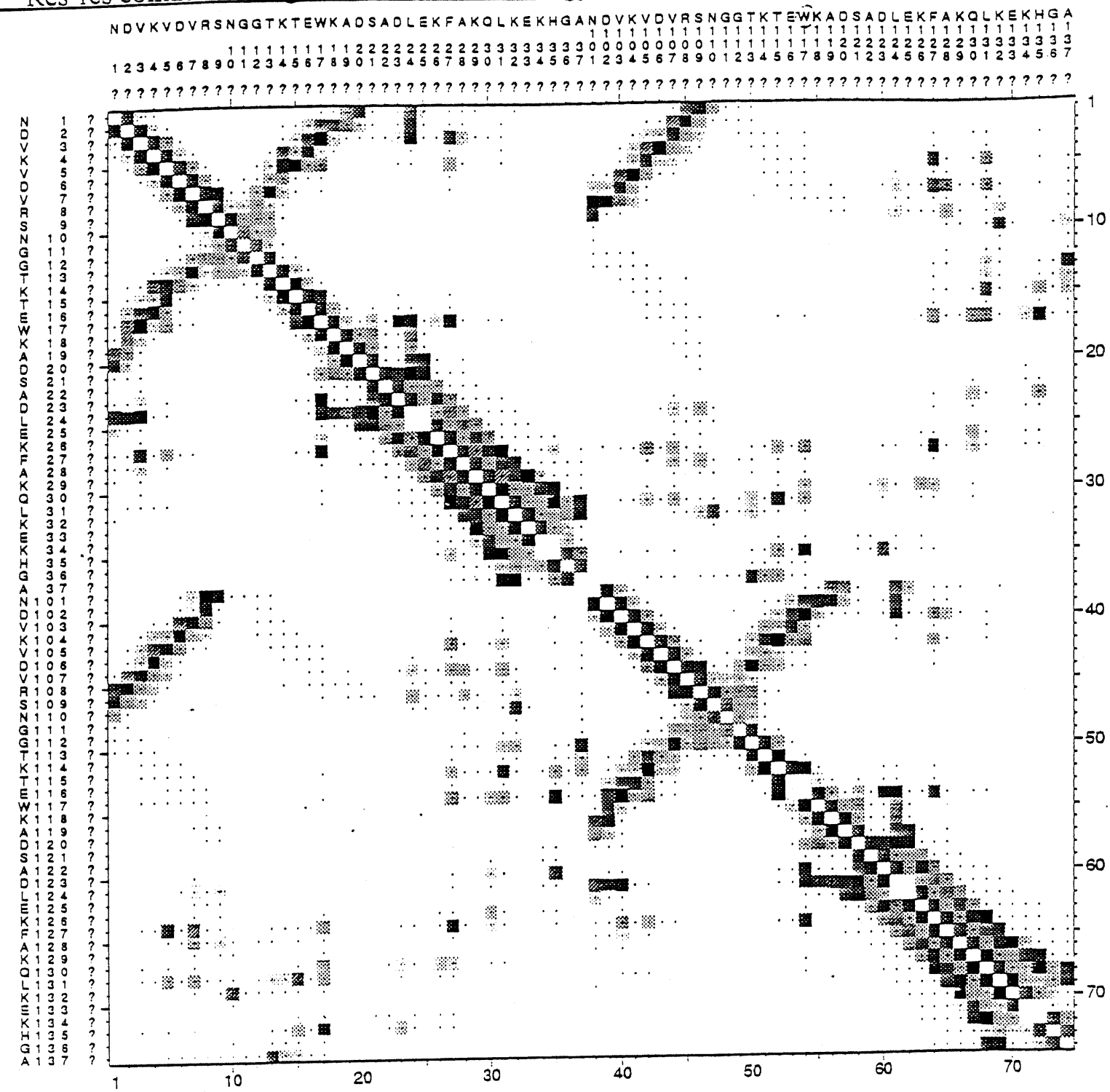
```
ATOM    236  CA  ASN   110      69.722  55.766  12.380  2.49  1.03
ATOM    241  CA  GLY   111      71.401  59.148  12.893  1.91  0.94
ATOM    246  CA  GLY   112      69.776  60.752   9.836  1.91  0.94
ATOM    251  CA  THR   113      70.746  59.418   6.410  2.48  1.03
ATOM    256  CA  LYS   114      70.271  60.795   2.923  2.47  0.93
ATOM    261  CA  THR   115      69.246  58.404   0.180  2.18  0.91
ATOM    266  CA  GLU   116      69.034  59.259  -3.529  2.36  0.93
ATOM    271  CA  TRP   117      68.316  57.372  -6.746  2.15  0.95
ATOM    278  CA  LYS   118      69.311  58.153 -10.307  4.09  1.55
ATOM    283  CA  ALA   119      66.478  59.072 -12.665  1.85  0.82
ATOM    288  CA  ASP   120      64.624  56.139 -14.200  2.34  0.99
ATOM    293  CA  SER   121      65.775  53.133 -12.164  2.10  0.92
ATOM    298  CA  ALA   122      64.183  50.137 -10.393  2.52  1.12
ATOM    303  CA  ASP   123      64.176  51.973  -7.039  2.28  0.96
ATOM    308  CA  LEU   124      63.414  55.328  -8.740  2.31  0.91
ATOM    313  CA  GLU   125      60.162  54.104 -10.340  2.31  0.91
ATOM    318  CA  LYS   126      59.404  51.800  -7.362  3.07  1.16
ATOM    323  CA  PHE   127      59.574  54.639  -4.829  2.46  1.08
ATOM    329  CA  ALA   128      57.894  57.148  -7.169  2.44  1.08
ATOM    334  CA  LYS   129      55.036  54.659  -7.755  2.12  0.80
ATOM    339  CA  GLN   130      54.738  54.038  -3.988  2.82  1.12
ATOM    344  CA  LEU   131      54.273  57.769  -3.325  2.55  1.01
ATOM    349  CA  LYS   132      51.999  58.239  -6.371  2.52  0.95
ATOM    354  CA  GLU   133      49.642  55.351  -5.851  2.04  0.80
ATOM    359  CA  LYS   134      49.758  54.541  -2.127  1.98  0.75
ATOM    364  CA  HIS   135      50.284  58.054  -0.750  2.36  1.01
ATOM    370  CA  GLY   136      48.917  60.325  -3.518  1.98  0.98
ATOM    375  CA  ALA   137      52.168  62.159  -4.334  2.21  0.98
TER
END
```

```
 LENGTH            37

   1  37  ....,....1....,....2....,....3....,....4....,....5
SEQ        NDVKLDVRDNGGVKTEWKADSADLEKFAKALKEKHGA

ROBSON     HHEEEEEETT     EHHHHHHHHHHHHHHHHHHHHH
SEGMENT83    EEEEE                  HHHHHHHHHH
GOR          HEE E        EHHHH HHHHHHHHHHHHHHH
GORCLASS   331131113344311233213545555555553212
ALB          EEEEE  TTT              HHHHHHHH
ALB*CLASS*3322222232222222233222333333322223 3
```

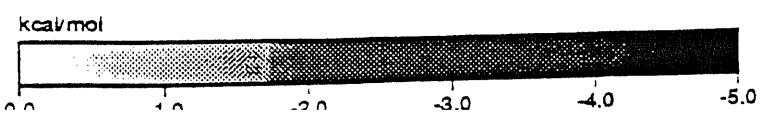
1

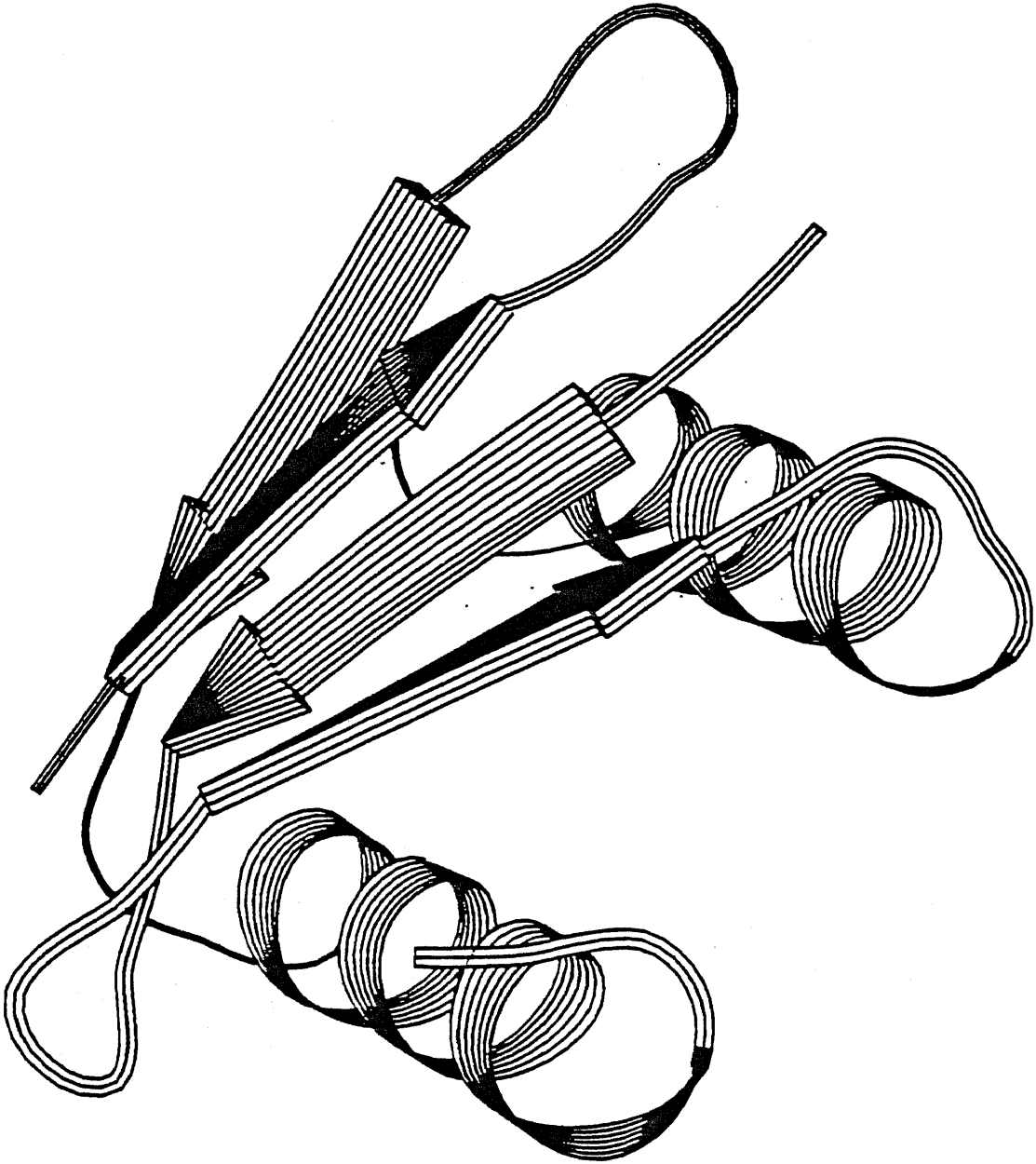
```
CLAS ROBSON          37

                          P R E D I C T E D
                           EAPM. TCLSB HIU..
                               7     6    24
                        %     19    16    65

CLAS SEGMENT83       37

                          P R E D I C T E D
                           EAPM. TCLSB HIU..
                               5    22    10
                        %     14    59    27

CLAS GOR             37

                          P R E D I C T E D
                           EAPM. TCLSB HIU..
                               4    11    22
                        %     11    30    59
```

Res-res contacts of finger: Van der Waals energy, n-conts = 1808 . 10 October 1990 16:38:05




FINGER_CLASP Contact analysis by CONAN.





# SHPILKA DESIGN: An eight-stranded anti-parallel beta-sandwich.

Alexei V. Finkelstein, Andrew Lockhart, Rainer Merkl, Jeanne Perry

Alexei V. Finkelstein
Institute of Protein Research,
Academy of Sciences of USSS,
Pushchino, Moscow Region,
142292, USSR
cdp!protres@labrea.stanford.edu
ATTN: FINKELSTEIN

Andrew Lockhart
Dept. of Crystallography
Birkbeck College
Malet Street
London, WC1E 7HX
England
A_LOCKHART@ICRF.AC.UK
UBCG13D@CR.BBK.AC.UK

Rainer Merkl
Institut fuer Molekulare Genetik
Grisebachstr. 8
3400 Goettingen
(0551)39-3823
U0908@DGOGWDG5.BITNET

L. Jeanne Perry
Molecular Biology Institute
UCLA
Los Angeles, CA 90024-1570
(213)825-8901
PERRY@UCLAUE.BITNET

## 3.1 ABSTRACT

We have designed an eight-stranded anti-parallel beta-sandwich called "Shpilka". Shpilka has a folding pattern that has not yet been found in the protein data base; the sequence is not homologous to any known protein. The design emphasizes the importance of the edge strands in defining the topology of the beta sandwich and carefully partitions hydrophobic and hydrophilic residues to the inside and outside of the sheet, respectively. The design was modular. We started from a single beta hairpin which was repeated four times creating an initial beta-sandwich with three two-fold symmetry axes. This symmetry was created to aid in the first steps of the design. Once the protein had been modeled and shown to have an appropriate packing density, the symmetry was broken: first to connect all the beta-strands by a single chain; second to achieve a better dense packing; third for the sake of future NMR studies. A buried tryptophan is included for experimental verification of the predicted structure. Shpilka contains 96 residues.

## 3.2 REPORT OF PROGRESS

I. A list of possible folding patterns for 8-stranded beta proteins has been prepared previously (A. Finkelstein and B. A. Reva) from physical considerations which have been discussed in: A.V.Finkelstein, O.B.Ptitsyn (1987) Progr. Biophys. Molec. Biol., 50, 171-190 (Figure 1).

### 3.2.1 PDB Searches

A modified BETASHEET and DSSP database-program was used to search for proteins containing at least one beta sheet. We extracted,a subset of proteins that contain at least two sheets of at least 4 strands each, that are predominately antiparallel. These proteins were sorted by their sheet configuration in the sandwich: Colinear or Orthogonal. We analyzed structures fitting the above criteria using INSIGHT.

### 3.2.2 Ananlysis Of Beta Sandwiches.

We have been partiularly interested in the orientations of the edge strands of the beta sheets. In general there are two kinds of orientations of the edge strands formed by L-amino acids. In the first class ("++"), residues where the NH and CO groups are orientated the center of the sheet and C-beta groups are orientated towards the inside the sandwich, are alternating with residues where all these groups are simultaneously oriented outward. In the second possible orientation of the strand ("+-"), residues whose NH and CO groups are orientated towards the inside of the sheet and C-beta groups are oriented outward, are alternating with residues where NH and CO are oriented outward while C-beta groups are oriented inward (Figure 2).

We found five combinations for the edge strands and these are shown in table 1. The most predominant folding pattern is all four edge strands in a "++" orientation (++,++/++,++). This folding class included gamma crystallin, tomato bushy stunt virus and satellite tobacco virus coat protein. Other folding classes were also seen and are summarized in Figure 3. Sheets of type "++" are three times more often than "+-,-+" strands. We decided to use them for our design.

3.2.3 The Design

3.2.3.1 Design Constraints - 1. The position of the edge strands in the primary sequence must define a unique folding topology.
2. This must be a topology not yet found in proteins.

3.2.3.2 Pattern Choice - Figure 1 shows a set of potentially stable folding patterns for eight-stranded anti-parallel beta-sandwiches. Only patterns 24 (2STV, 2TBV and 3GAP), 36 (HPL, F.K.Winkler, A.D'Arcy, W.Hunziker (1990) Nature, 343, 771-774) and 47 (1GCR) were found during this analysis.
The patterns of the edge strands are presented in figure 4. Each folding pattern corresponds to the same alteration of internal and edge ("++" and/or "+-") strands along a protein chain. (Fig. 4). For different folding patterns these alterations are usually different. For example, 22 topologies marked with an asterix (*) in figs. 1 and 4 have "unique" alternations of internal and "++"-edge beta strands : The pattern corresponds only to a given topology. Only these 22 unique folding topologies were considered for the design. From these, pattern "54" was chosen. This fold has a connectivity requiring only three kinds of loops: intra-sheet loops (1, 3 and 6), diagonal inter-sheet loops (2, 5 and 7) and a direct inter-sheet loop (4). Our aim was to design a sequence which codes for a unique 3D-structure. As we wanted to design a 8-stranded beta-sandwich protein, this sequence must contain 8 regions appropriate for a stable beta structure. Four of these regions must be more hydrophobic to form the internal strands; four must be good for edges, but bad for internal positions. Moreover, these four must be more appropriate for "++" positions, which we are going to use in out sandwich, than for "+-" ones. The ideas for the choice of the sequences for the internal and external strands are illustrated by fig. 2. The alternation of these sequences along the chain must correspond with the desired folding patterns (fig. 4).

3.2.3.3 Primary Sequence Of Strands For The Beta-sandwich Structure - At first, the sequence was constructed from four modular units consisting of two antiparallel beta strands connected by a loop to form a beta hairpin. This unit was then modified to allow proper packing between the sheets. The primary structures for the beta regions was enriched by valines and isoleucines. The edge strands were less hydrophobic and punctuated by prolines (to prohibit these

strands from becoming internalized).
The following sequence was the starting point:

```
Strand 1 (inner strand)
N-tyr-LYS-ile-THR-ile-THR-tyr-GLU-C
  ||      ||      ||      ||
C-tyr-GLU-val-PRO-val-SER-his-LYS-N
Strand 2 (outer strand)
```

Residues in upper case are those facing outside the sheet
|| = hydrogen bonds, N and C: N and C termini, respectively

3.2.3.4 The Search For "Perfect" Strand-Conformation - We wanted a proline in the edge strand. Proline has no NH group and therefore it cannot form a H-bond inside a beta sheet. Thus, proline can mark an edge: it can enter beta-sheets if its ring looks outside. We wanted the strand to be of "++"-type. Thus, this proline must be at the outer surface of a sheet (fig. 2). The database was again searched (WHATIF) for beta strands containing prolines with a solvent accesibility greater than 25 square Angstroms. The search found 54 examples. We looked at four proteins where the proline was more centrally located in the edge strand. These proteins were 2CAB, 1FC1, 2STV and 1TNF. The only protein which did not contain large distortions or beta bulges was TNF. We selected the outer two strands of beta-sheet 1 of TNF composed of residues C60 to C67 and C114 to C121 for our template strands upon which to fit the hairpin sequence.

3.2.3.5 The Search For The Hairpin-Loop - To form the hairpins, the strands of the sandwich were connected by short hydrophilic loops. INSIGHT was used to search for a hairpin turn of 5 residues, capable to connect residue C67 and C114, with 5 Preflex and Postflex residues. Seven turns were found. The primary sequence we planned to use for the turn was ser-gly-gly-ser. The loop that most closely corresponded to this sequence was a loop from 2RHE (68 to 71):

ser-gly-thr-ser.

The turn was grafted onto the two strands using WHATIF. This model, STRAND12LOOP.BRK, was refined to ensure that the bond angles, torsion angles and the distances were all within reasonable limits. Molecular dynamics (GROMOS) revealed only very minor changes in the structure.

The primary sequence for the beta-turn-beta module is shown below:

```
Strand 1 (inner strand)
N-tyr-LYS-ile-THR-ile-THR-tyr-GLU-ser-gly
  ||      ||      ||      ||      ||   |
C-gly-tyr-GLU-val-PRO-val-SER-his-LYS-ser-thr
Strand 2 (edge strand)
```

3.2.3.6 The Search For The "Perfect Sandwich" - We searched for a beta sandwich among our extracted lists of proteins, to use as a framework for our design. The beta sandwich structure which was most similar to our proposed design was STV. It contains one sheet that is essentially "perfect" however the second sheet in the sandwich is shorter and more twisted. Two copies of the beta hairpin module was superimposed onto the framework of the less twisted sheet of STV (INSIGHT). In order to complete the sandwich, the sheet was copied (WHATIF) to form another sheet and rotated into place using the less perfect sheet of STV for positioning. One of us (A.Finkelstein) wants to stress, that computers, which are very convinient for analysis of protein structures, are (at present) rather inconvinient (as compared to CPK models, for example) for protein design de novo. He has a filing as if he is drawing using xerox and ceasars instead of a good old pencil.

3.2.3.7 Preliminary Sandwich Packing - Bad contacts between sidechains were adjusted manually and the sandwich was energy minimized (GROMOS). The results of the minimization showed that threonines (4 and 6) of the internal strands were overlapping. The side chains were moved manually to avoid clashes.

3.2.4 Design Evolution

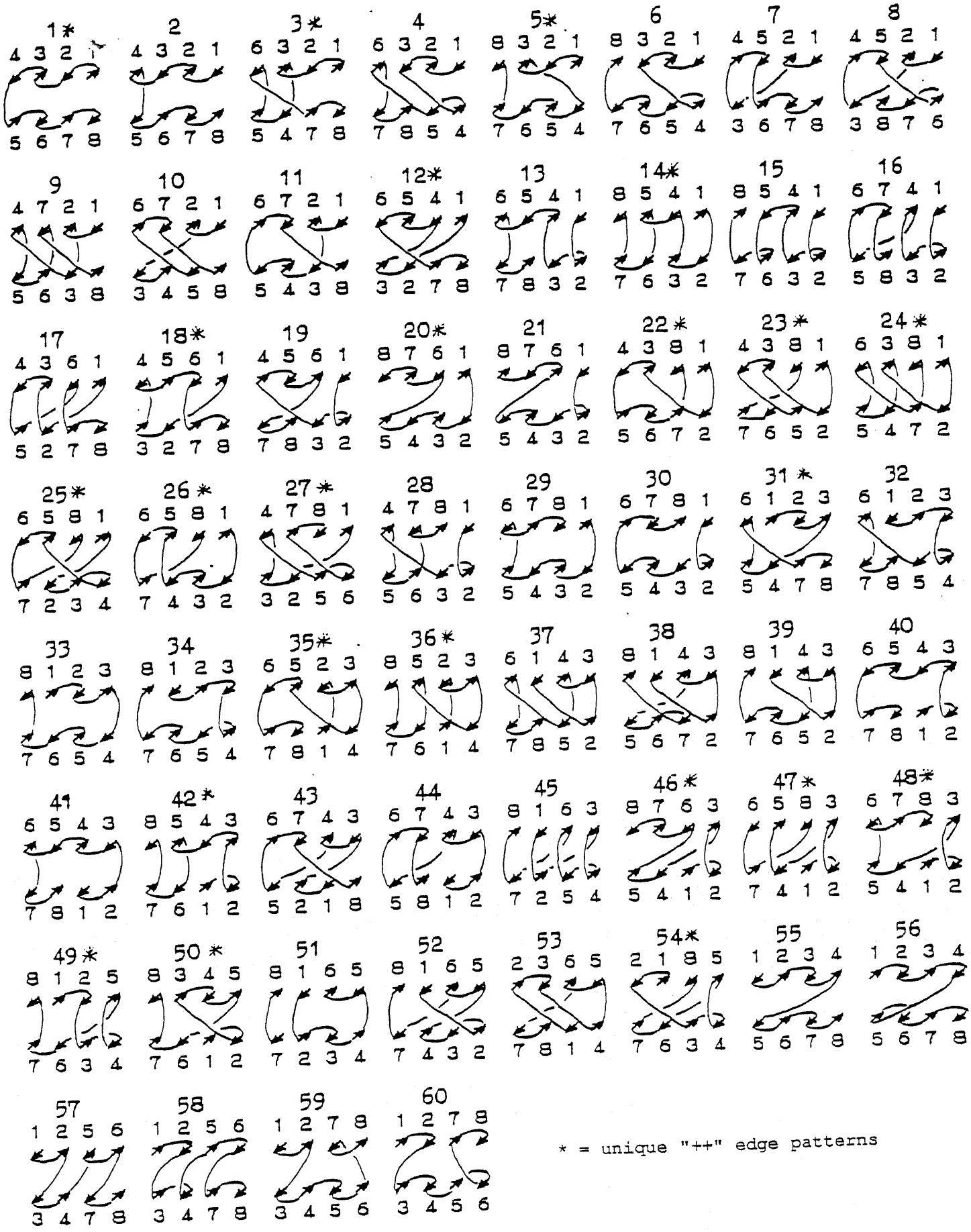
In order to connect the sheets in the correct topology it was necessary to break one of the beta hairpins and construct four new connections between the strands: three long loops and one short loop. Polyglycines were used initially for all these loops so we could perform molecular dynamics followed by energy minimization.
The results of the energy minimization pointed out the poor internal packing of the molecule. The regions of the large hydrophobic residues at the ends of the sheets were too densely packed whereas the regions of the small hydrophobic residues in the core of the sandwich were too sparse.
We tackled the packing problem at the ends of the sheets by mutating Tyr7 of the inner strands (nummeration according to the hairpin shown above) to Phe. However, these mutations produced no improvement. These residues were mutated to Leu. These mutations improved the packing between the sheets and alleviated steric clashes with the histidines in the edge strands. Due to the sparse packing in the middle of the sandwich, Ile 3 was mutated to Phe. These changes did not improve the packing and were subsequently mutated to Met as methionines are "the good for all stick them ins" (G. Vriend, personnal communication). In addition, Ser 9 and 12 were replaced by threonines, as the larger bulk of the threonines would force the methionines to pack internally.
It was obvious from packing analyses that the glycine loops were not

well packed. After energy minimization (WHATIF/GROMOS) it was alse noted, that several glycines in the loops were found in a cis conformation. Attempts were made to correct the conformations to trans. Following these attempts, a search was made of the database (WHATIF) to find loops, that could be used to connect these strands. At the same time, the chi angles were measured. It was found that about half of the internal residues had unfavored chi angles. The torsion angles were adjusted manually to fit proper rotomers. In addition several mutations were introduced, to fill the holes which remained and to remove bad contacts. We also changed residues on the surface to make it more hydrophilic. Now we broke the symmetry of the beta-sandwich (before only the loops were not symmetrical) and then we changed some lysines and some glutamic acids at the surface to arginines and aspartic acids, respectively, to make the strands distinguishable by NMR.
A tryptophan was inserted into the interior of the sheet to serve as an experimental marker for folding of SHPILKA. The final amino acid sequence of our protein, SHPILKA, is shown in table 2.

### 3.2.5 Plans

Alosha will stay at the EMBL until he improves close packing. Andrew will construct this protein from synthesized oligonucleotides. He will express, purify, characterize, crystallize, do NMR and X-ray study and then solve the structure of our designed protein. Once he has acheived these goals (to 1.5 Angstom resolution), the SHPILKA design group will once again meet (this time in Tahiti) to design the metal binding defect into the protein, a defect which can bind metal (e.g. platinum) or do something more interesting.

FIGURE 1: A SET OF "POTENTIALLY STABLE" FOLDING PATTERNS
FOR A BETA-SANDWICH WITH FOUR ANTIPARALLEL STRANDS IN EACH BETA-SHEET



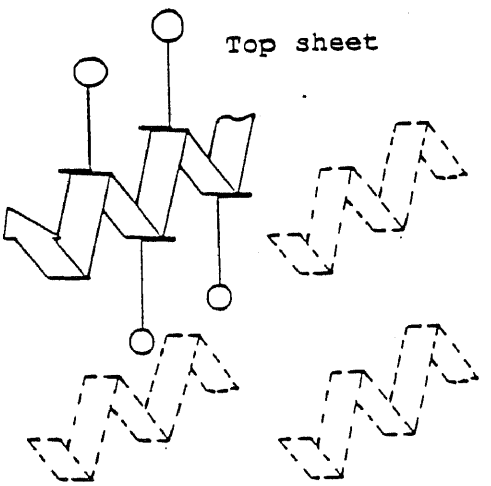

* = unique "++" edge patterns

FIGURE 2: EDGE STRAND ORIENTATION OF BETA SANDWICHES

A. "++" Orientation:

NH and CO groups oriented outward; C-beta groups oriented outward

NH and CO groups oriented inward; C-beta groups oriented inward



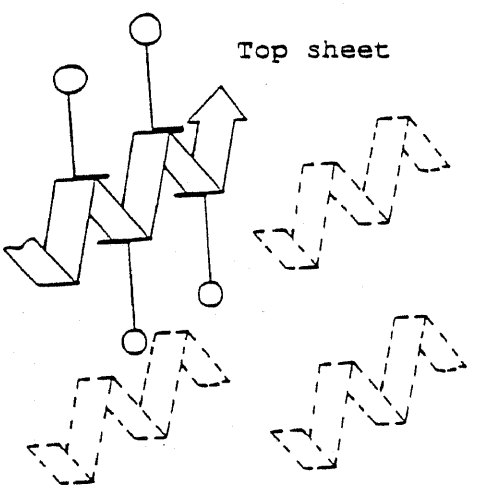

Proposed sequence for the ++ edge strand:

val
|
PRO
|
val
|
X

B. "+-" Orientation.

NH and CO groups oriented inward; C-beta groups oriented outward

NH and CO groups oriented outward; C-beta groups oriented inward




Proposed sequence for the +- edge strand:

X
|
pro
|
X
|
val
|
X
|
val

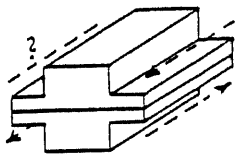
Proposed sequence for an internal strand:

X-val-X-val-..

X = Polar residue
Capitilised residue = residue outside beta sandwich
Lower case residue = residue inside beta sandwich

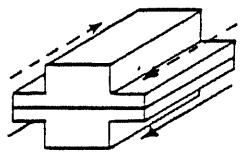
FIGURE 3: SUMMARY OF SHEET ANALYSIS OF BETA SANDWICH PROTEINS

I. ??,+-/+-,+-



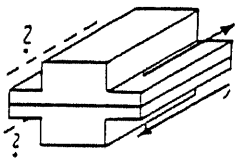
Proteins found in this folding class: 2TBV

II. +-,+-/++,++



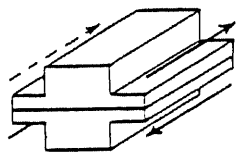
Proteins found in this folding class: 1FAB, 2FB4, 3HFM

III. ??,++/??,++



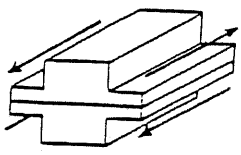
Proteins found in this folding class: 2SOD, 2CNA

IV. +-,++/++,++



Proteins found in this folding class: 2PAB, 2AZA, 2PCY

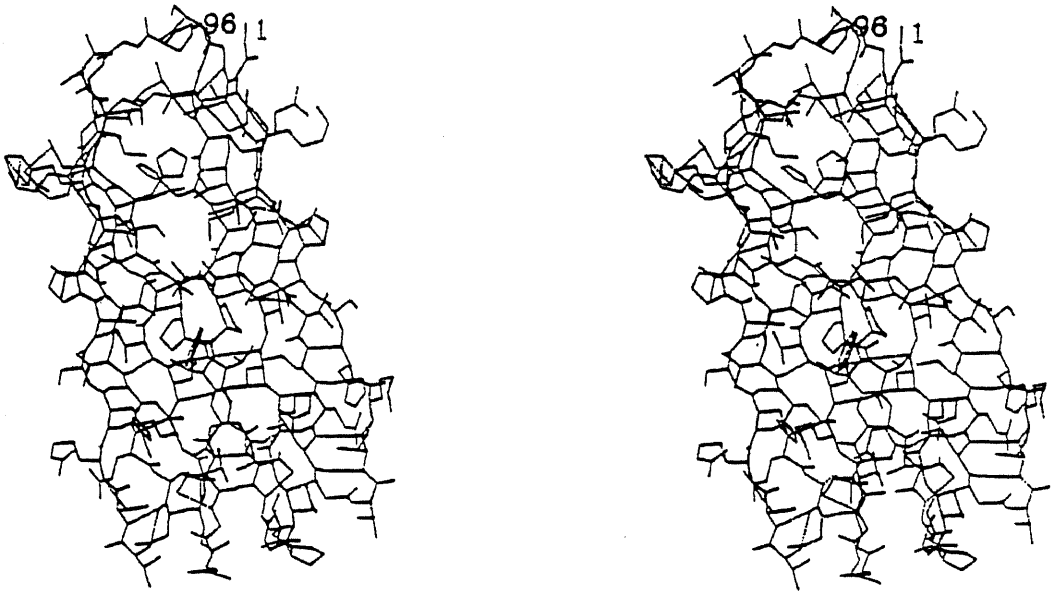
V. ++,++/++,++



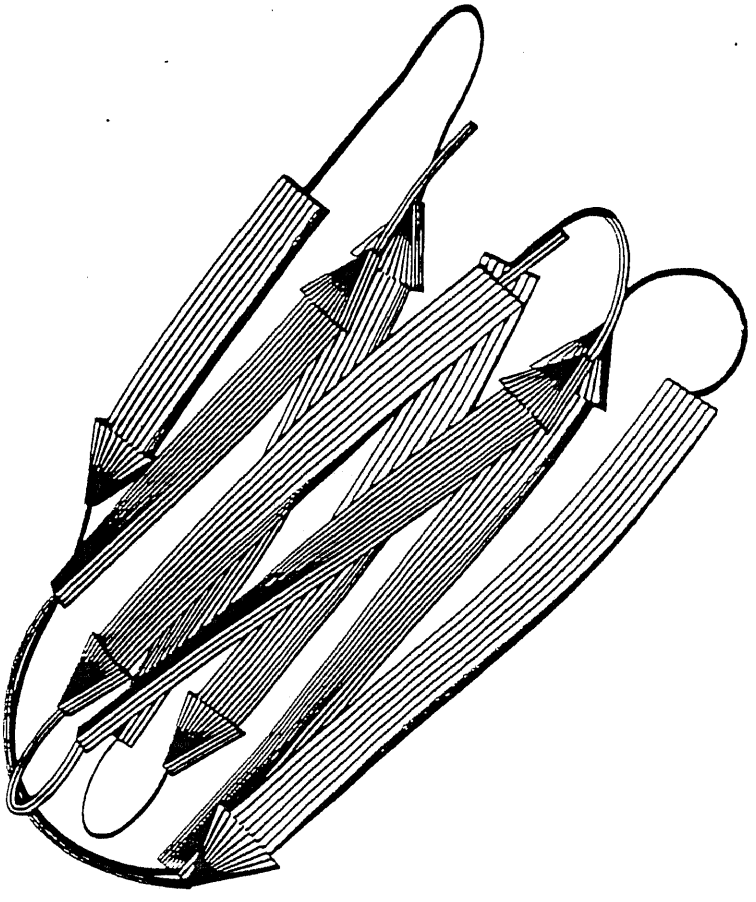
Proteins found in this folding class: 1GCR, 2TBV, 2STV

FIGURE 4: Edge pattern strands of antiparallel beta-sheets.

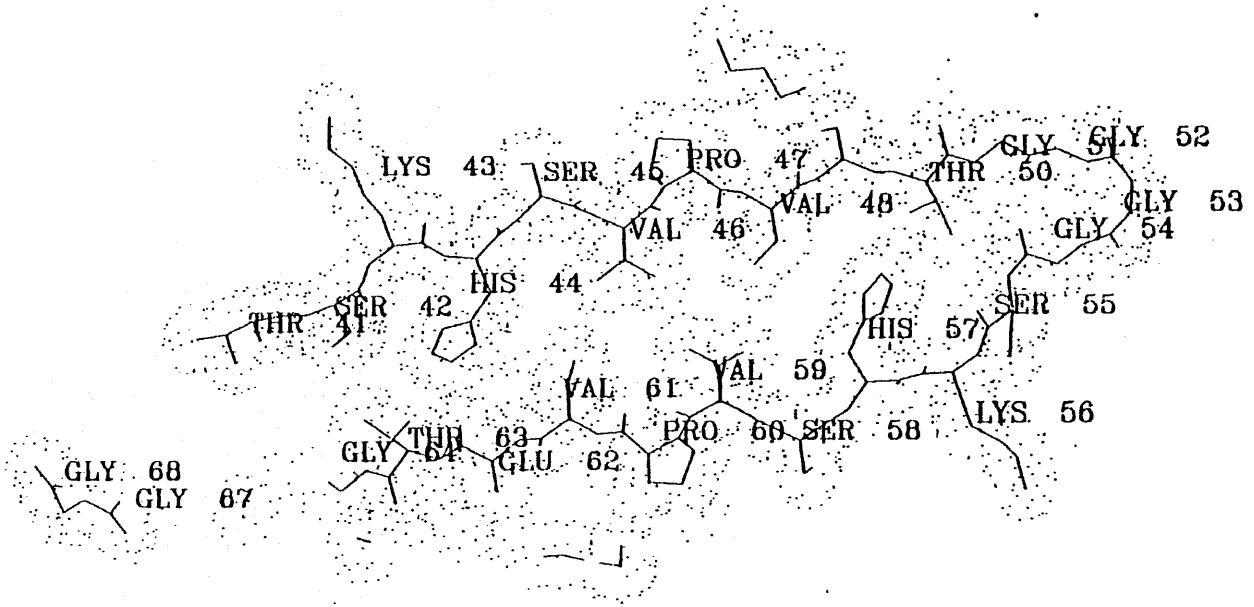
| PATTERN # | PATTERN | EQUIVALENT PATTERN # | NOTES |
|---|---|---|---|
| 1 | 1A - - 4A 5A - - 8A | 17 | A = ALTOGETHER = ++ |
| 2 | 1E - - 4E 5E - - 8E | 9 | E = EITHER/OR = +- |
| 3* | 1A - - - 5A 6A - 8A | | |
| 4 | 1E - - 4E - 6E 7E - | | |
| 5* | 1A - - 4A - - 7A 8A | | |
| 6 | 1E - - 4E - - 7E 8E | | |
| 7 | 1E - 3E 4E - - - 8E | | |
| 8 | 1E - 3E 4E - 6E - - | | |
| 9 | 1E - - 4E 5E - - 8E | 2 | |
| 10 | 1E - 3E - - 6E - 8E | | |
| 11 | 1E - - - 5E 6E - 8E | | |
| 12* | 1A - 3A - - 6A - 8A | | |
| 13 | 1E 2E - - - 6E 7E - | | |
| 14* | 1A 2A - - - - 7A 8A | | |
| 15 | 1E 2E - - - - 7E 8E | | |
| 16 | 1E 2E - - 5E 6E - - | 30 | |
| 17 | 1A - - 4A 5A - - 8A | 1 | |
| 18* | 1A - 3A 4A - - - 8A | | |
| 19 | 1E 2E - 4E - - 7E - | | |
| 20* | 1A 2A - - 5A - - 8A | | |
| 21 | 1E 2E - - 5E - - 8E | | |
| 22* | 1A 2A - 4A 5A - - - | | |
| 23* | 1A 2A - 4A - - 7A - | | |
| 24 | 1A 2A - - 5A 6A - - | 29 | |
| 25* | 1A - - 4A - 6A 7A - | | |
| 26* | 1A 2A - - - 6A 7A - | | |
| 27* | 1A - 3A 4A - 6A - - | | |
| 28 | 1E 2E - 4E 5E - - - | | |
| 29 | 1A 2A - - 5A 6A - - | 24 | |
| 30 | 1E 2E - - 5E 6E - - | 16 | |
| 31* | - - 3A - 5A 6A - 8A | | |
| 32 | - - 3E 4E - 6E 7E - | | |
| 33 | - - 3A 4A - - 7A 8A | 36 | |
| 34 | - - 3E 4E - - 7E 8E | 45 | |
| 35* | - - 3A 4A - 6A 7A - | | |
| 36 | - - 3A 4A - - 7A 8A | 33 | |
| 37 | - 2E 3E - - 6E 7E - | 41 | |
| 38 | - 2E 3E - 5E - - 8E | | |
| 39 | - 2E 3E - - - 7E 8E | | |
| 40 | - 2A 3A - - 6A 7A - | 47 | |
| 41 | - 2E 3E - - 6E 7E - | 37 | |
| 42* | - 2A 3A - - - 7A 8A | | |
| 43 | - - 3E - 5E 6E - 8E | | |
| 44 | - 2E 3E - 5E 6E - - | | |
| 45 | - - 3E 4E - - 7E 8E | 34 | |
| 46* | - 2A 3A - 5A - - 8A | | |
| 47 | - 2A 3A - - 6A 7A - | 40 | |
| 48* | - 2A 3A - 5A 6A - - | | |
| 49* | - - - 4A 5A - 7A 8A | | |
| 50* | - 2A - - 5A - 7A 8A | | |
| 51 | - - - 4E 5E - 7E 8E | | |
| 52 | - 2E - - 5E - 7E 8E | | |
| 53 | - 2E - 4E 5E - 7E - | | |
| 54* | - 2A - 4A 5A - 7A - | PATTERN CHOSEN FOR BETA SANDWICH DESIGN | |
| 55 | 1A - - 4A 5E - - 8E | | |
| 56 | 1E - - 4E 5A - - 8A | | |
| 57 | 1A - 3E - - 6A - 8E | | |
| 58 | 1E - 3A - - 6E - 8A | | |
| 59 | 1A - 3E - - 6E - 8A | | |
| 60 | 1E - 3A - - 6A - 8E | *UNIQUE ++EDGE PATTERNS | |




SHPILKA



Stereodrawing (above) and a scheme of beta-sandwich (below) for the designed protein.




A slice throug the SHPILKA globule. The dotts show the Vanderwaals surfaces of atoms.

TABLE 1: Sheet Analysis of Beta Sandwich Proteins

| PROTEIN | SHEET | RESIDUES | EDGE PATTERN | | SANDWICH PATTERN |
|---|---|---|---|---|---|
| GAMMA-II CRYSTALLIN (CALF) | | | | | |
| 1GCR | 1 | 60 - 64 | ++ | | ++,++/++,++ |
| | | 14 - 19 | ++ | | |
| | 2 | 21 - 24 | ++ | | |
| | | 53 - 58 | ++ | | |
| | 3 | 101-106 | ++ | | ++,++/++,++ |
| | | 149-153 | ++ | | |
| | 4 | 108-111 | ++ | | |
| | | 142-147 | ++ | | |
| IMMUNOGLOBULIN BENCE-JONES PROTEIN, V-DIMER: 2 SHEETS/MONOMER, 4 STRANDS/SHEET | | | | | |
| 1REI | 1 | 3 - 7 | +- | | |
| | | 62 - 67 | +- | | |
| | 2 | 10 - 13 | ++ | | |
| | | 44 - 47 | ++ | 5 STRANDS, 2 PARALLEL | |
| IMMUNOGLOBULIN: 4 BETA SANDWICHES THAT FORM 2 BARRELS BETWEEN EACH | | | | | |
| 1FAB | 1 | 4 - 7 | +- | | +-,+-/++,++ |
| | | 62 - 66 | +- | | |
| | 2 | 9 - 13 | ++ | | |
| | | 43 - 46 | ++ | | |
| | 3 | 160-169 | +- | | +-,+-/++,++ |
| | | 116-121 | +- | | |
| | 4 | 155-157 | ++ | | |
| | | 202-209 | ++ | | |
| PREALBUMIN, HUMAN PLASMA | | | | | |
| 2PAB | 1 | 54 - 56 | +- | | +-,++/++,++ |
| | | 115-122 | ++ | | |
| | 2 | 40 - 49 | ++ | | |
| | | 90 - 97 | ++ | | |
| AZURIN (ALCALIGENES DENITRIF.): DIMER, COLINEAR, 4 STRANDS/SHEET | | | | | |
| 1AZA | 1 | 14- 16 | +- | PARALLEL STRANDS | +-,++/++,++ |
| | | 92 - 98 | ++ | | |
| | 3 | 82 - 88 | ++ | MOSTLY BURIED | |
| | | 121-123 | ++ | | |
| CONCANAVALIN A: 2 LARGE BETA SHEETS/SANDWICH, 6 AND 7 STRANDS EACH | | | | | |
| 2CNA | 1 | 141-143 | ?? | | ??,++/??,++ |
| | | 37 - 39 | ++ | | |
| | 2 | 146-148 | ?? | | |
| | | 72 - 78 | ++ | | |
| IMMUNOGLOBULIN (FAB): HUMAN, MYELOMA | | | | | |
| 2FB4 | 1 | 4 - 7 | +- | | +-,+-/++,++ |
| | | 61 - 66 | +- | | |
| | 2 | 9 - 15 | ++ | | |
| | | 43 - 46 | ++ | | |
| | 3 | 161-169 | +- | MOSTLY BURIED | +-,+-/++,++ |
| | | 116-121 | +- | | |
| | 4 | 202-209 | ++ | | |
| | | 155-157 | ++ | | |
| SUPEROXIDE DISMUTASE | | | | | |
| 2SOD | 1 | 95-100 | ++ | 8-STRANDED SHEET | ??,++/??,++ |
| | | 83 - 88 | ++ | ALMOST A BARREL | |

TABLE 1 (Continued)

| | | | | | |
|---|---|---|---|---|---|
| APO-PLASTACYANIN: CUPROPROTIEN, 4 STRANDS/SHEET, NO OTHER SECONDARY STRUCTURE | | | | | |
| 2PCY | 1 | 14 - 16 | +- | PARALLEL SHEET | +-,++/++,++ |
| | | 68 - 74 | ++ | | |
| | 2 | 18 - 21 | ++ | 2 PARALLEL STRANDS | |
| | | 62 - 64 | ++ | | |
| TOMATO BUSHY STUNT VIRUS: TIMER; MANY, MANY ANTIPARALLEL BETA SHEETS | | | | | |
| 2TBV | 1 | 103-118 | ++ | | ++,++/++,++ |
| | | 208-222 | ++ | | |
| | 2 | 121-123 | ++ | | |
| | | 200-204 | ++ | | |
| | 3 | 286-288 | ?? | | ??,+-/+-,+- |
| | | 331-341 | +- | | |
| | 4 | 293-298 | +- | | |
| | | 324-330 | +- | | |
| SATELLITE TOBACCO VIRUS: COAT PROTEIN | | | | | |
| 2STV | 1 | 142-155 | ++ | | ++,++/++,++ |
| | | 26 - 38 | ++ | | |
| | 2 | 125-136 | ++ | | |
| | | 43 - 47 | ++ | | |
| IgG1 FAB FRAGMENT AND LYSOZYME | | | | | |
| 3HFM | 3 | 159-164 | +- | | +-,+-/++,++ |
| | | 114-119 | +- | | |
| | 4 | 204-210 | ++ | | |
| | | 151-155 | ++ | | |

? = the pattern of these strands were unable to be determined

TABLE 2. AMINO ACID SEQUENCE AND TORSIONAL ANGLES OF SHPILKA.

| | RESIDUE | PHI | PSI | OMEGA | CHI1 | CHI2 | CHI3 | CHI4 | CHI5 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | GLY (1 ) | | 146.6 | -179.3 | | | | | |
| 2 | ILE (2 ) | -114.0 | 143.6 | -176.4 | 51.6 | 119.9 | | | |
| 3 | LYS (3 ) | -126.7 | 140.3 | 171.6 | -166.3 | -179.1 | -78.7 | 65.6 | |
| 4 | MET (4 ) | -135.0 | 140.2 | 173.5 | -176.7 | 176.1 | -172.4 | | |
| 5 | THR (5 ) | -128.5 | 125.7 | 175.2 | -55.9 | | | | |
| 6 | ILE (6 ) | -112.6 | 119.0 | -179.7 | -49.0 | 150.6 | | | |
| 7 | THR (7 ) | -104.5 | 129.9 | -173.5 | -53.8 | | | | |
| 8 | LEU (8 ) | -125.2 | 105.7 | 166.9 | -52.7 | 161.5 | | | |
| 9 | GLU (9 ) | -101.7 | 102.5 | -170.5 | -173.4 | -178.0 | -42.5 | | |
| 10 | LEU (10 ) | -121.3 | 87.5 | -179.3 | -57.6 | 171.0 | | | |
| 11 | GLY (11 ) | 55.6 | -103.4 | 179.3 | | | | | |
| 12 | THR (12 ) | -80.8 | -17.0 | -177.8 | 44.9 | | | | |
| 13 | THR (13 ) | -119.0 | 150.6 | 174.2 | -163.4 | | | | |
| 14 | LYS (14 ) | -129.1 | 110.6 | -179.6 | -162.9 | 170.6 | -80.4 | 64.8 | |
| 15 | HIS (15 ) | -124.6 | 154.8 | -179.3 | -77.0 | 120.8 | | | |
| 16 | SER (16 ) | -135.6 | 140.9 | 175.4 | -63.5 | | | | |
| 17 | ILE (17 ) | -131.2 | 111.6 | 167.6 | 52.0 | 157.8 | | | |
| 18 | PRO (18 ) | -94.7 | 145.7 | 177.1 | | | | | |
| 19 | ILE (19 ) | -135.5 | 126.8 | -179.0 | 47.7 | 170.3 | | | |
| 20 | GLU (20 ) | -134.4 | 147.4 | 170.2 | -168.6 | 179.3 | -37.9 | | |
| 21 | THR (21 ) | -144.3 | 160.6 | 179.4 | -168.7 | | | | |
| 22 | GLY (22 ) | -70.7 | -142.9 | 177.6 | | | | | |
| 23 | SER (23 ) | -15.4 | 98.9 | -175.8 | 40.7 | | | | |
| 24 | PRO (24 ) | -97.0 | 171.8 | 170.8 | | | | | |
| 25 | GLY (25 ) | 149.6 | 117.6 | -179.2 | | | | | |
| 26 | GLY (26 ) | 69.9 | 120.8 | -177.2 | | | | | |
| 27 | ILE (27 ) | -124.4 | 131.1 | 171.3 | -166.7 | 159.2 | | | |
| 28 | ARG (28 ) | -129.2 | 104.4 | -177.8 | -178.8 | 176.9 | -62.1 | -59.4 | -6.5 |
| 29 | MET (29 ) | -124.6 | 135.5 | 176.3 | -170.9 | 175.7 | -156.8 | | |
| 30 | THR (30 ) | -122.4 | 121.1 | 177.0 | -56.9 | | | | |
| 31 | ILE (31 ) | -121.4 | 123.5 | 178.3 | -42.9 | 155.9 | | | |
| 32 | THR (32 ) | -126.1 | 129.2 | 177.0 | -50.5 | | | | |
| 33 | LEU (33 ) | -116.6 | 129.7 | 167.3 | 176.3 | 63.1 | | | |
| 34 | GLU (34 ) | -116.0 | 100.4 | -178.1 | -168.5 | 178.2 | -49.2 | | |
| 35 | LEU (35 ) | -110.0 | 98.0 | 175.8 | -170.9 | -74.9 | | | |
| 36 | GLY (36 ) | 47.0 | -90.9 | -177.9 | | | | | |
| 37 | THR (37 ) | -84.9 | -25.1 | -169.0 | 40.6 | | | | |
| 38 | THR (38 ) | -136.1 | 162.6 | -174.8 | -174.2 | | | | |
| 39 | ARG (39 ) | -118.1 | 142.2 | 171.0 | -175.9 | 167.9 | -57.3 | -53.6 | 0.2 |
| 40 | HIS (40 ) | -134.1 | 117.2 | 174.8 | -78.4 | 108.6 | | | |
| 41 | SER (41 ) | -119.6 | 113.7 | -178.4 | -61.4 | | | | |
| 42 | VAL (42 ) | -124.2 | 106.0 | 166.4 | 60.1 | | | | |
| 43 | PRO (43 ) | -88.3 | 127.6 | 177.9 | | | | | |
| 44 | VAL (44 ) | -124.2 | 121.1 | -170.9 | -74.7 | | | | |
| 45 | GLU (45 ) | -143.9 | 162.1 | 168.4 | -164.9 | -161.7 | -32.0 | | |
| 46 | GLY (46 ) | 123.0 | -124.8 | 176.6 | | | | | |
| 47 | GLY (47 ) | -129.2 | -161.4 | 179.8 | | | | | |
| 48 | THR (48 ) | 46.8 | -151.2 | 178.0 | -165.7 | | | | |
| 49 | LYS (49 ) | -100.3 | 142.3 | 176.1 | -167.5 | 177.1 | -77.4 | 69.3 | |
| 50 | HIS (50 ) | -126.9 | 116.8 | 170.5 | -42.4 | 109.9 | | | |
| 51 | SER (51 ) | -117.4 | 109.3 | -173.5 | -64.1 | | | | |
| 52 | VAL (52 ) | -122.4 | 115.3 | 166.3 | 174.9 | | | | |
| 53 | PRO (53 ) | -91.4 | 145.5 | 172.9 | | | | | |
| 54 | VAL (54 ) | -136.8 | 133.0 | -177.4 | -78.0 | | | | |
| 55 | ASP (55 ) | -147.6 | 150.7 | 173.2 | 178.5 | 100.8 | | | |
| 56 | THR (56 ) | -161.6 | -168.8 | 176.8 | -170.4 | | | | |
| 57 | GLY (57 ) | -146.8 | -160.2 | -175.2 | | | | | |
| 58 | GLY (58 ) | 109.4 | -115.2 | 175.1 | | | | | |
| 59 | GLY (59 ) | 166.7 | 163.4 | 177.5 | | | | | |
| 60 | GLY (60 ) | 124.8 | -154.1 | -179.1 | | | | | |

```
61 GLY (61  )  -77.5 155.6-173.6
62 VAL (62  ) -128.3 121.9 169.1-169.0
63 LYS (63  ) -134.0 127.2 176.6-165.6 178.0 -81.0  69.3
64 TRP (64  ) -125.6 129.6 175.8 169.7 -91.1
65 THR (65  ) -119.9 119.2 175.3 -54.5
66 VAL (66  ) -106.3 105.9-178.9-111.6
67 THR (67  ) -119.5 121.5-178.3 -52.4
68 MET (68  ) -132.1 134.9 167.9  60.5 177.3-165.4
69 ASP (69  ) -116.5 118.2 174.5-175.7  69.3
70 LEU (70  ) -129.9 113.2-179.9 176.1 123.4
71 GLY (71  )   48.5-103.5-179.7
72 THR (72  )  -74.7 -31.6-166.1  39.8
73 THR (73  ) -116.9 136.9 176.6-165.2
74 ARG (74  ) -110.7 112.8 178.5-176.8 171.5 -59.8 -55.1  -2.2
75 HIS (75  ) -114.8 141.0 175.3 -79.0  69.4
76 SER (76  ) -142.2 124.1 177.7 -61.5
77 THR (77  ) -135.6 123.3 170.7 175.4
78 PRO (78  )  -98.6 149.6 165.3
79 VAL (79  ) -128.9 139.4-173.2 -77.3
80 ASP (80  ) -151.4 146.1 172.4-178.2  93.0
81 THR (81  ) -163.8-168.3-172.1-168.8
82 SER (82  ) -109.0 176.9 175.3  59.0
83 GLY (83  )    2.3  98.8 170.4
84 SER (84  ) -110.2 125.4 167.3 -74.7
85 PRO (85  )  -76.9 -38.1-177.1
86 ASN (86  ) -130.3 151.4-179.3-157.4 -62.3
87 VAL (87  ) -152.3 117.7 174.6  56.0
88 ARG (88  ) -129.0 123.0 179.7-172.1 168.6 -60.8 -57.1  -1.9
89 MET (89  ) -129.3 144.7 173.6 -64.2 168.4 164.3
90 THR (90  ) -134.6 133.0 178.7 -55.0
91 VAL (91  ) -110.0 115.4 174.4 155.3
92 THR (92  ) -120.4 110.8 177.6 -53.1
93 VAL (93  ) -128.8 137.2 175.3  58.9
94 ASP (94  ) -126.5 141.6 168.5 179.8  55.3
95 LEU (95  ) -118.9 123.0 175.0-165.9 149.2
96 GLY (96  ) -120.5 999.9 999.9
```

# *LEATHER*

## A SMALL NAD-BINDING PROTEIN

1LEA.PDB
2LEA.PDB

TIM HUBBARD
ARNE ELOFSSON
MARC EBERHARD
LYNNE REGAN

TM: IRC, MRC, Hills Road, Cambridge, CBQ 2QH, U.K.:
AE: Dept. Mol. Biophys., Karolinska Inst. Box 60400, S-10401,
Stockholm, Sweden : ARNE@MEDFYS.KI.SE
ME: Abt. Biophysik. Chemie. Biozentrum der Univ, Klingelbergstr. 70
CH-4056, Basel, Switzerland : EBERHARD@URZ.UNIBAS.
LR: Yale University, Dept. Mol. Biophys. & Biochem., 260, Whitney Ave.
P.O. Box 6666, New Haven, CT 06511, U.S.A. :
REGAN%HHVMS8@VENUS.YCC.YALE.EDU

# LEATHER

[L]ynne, [E]lofsson, [A]rne, [T]im [H]ubbard, [E]berhard, [R]egan, [M]arc.

## 4.1 ABSTRACT

A small NAD binding protein has been designed, starting from the x-ray crystal structure of Lactate Dehydrogenase. The aim was to design the smallest protein fragment consistant with retaining the maximum number of interactions between protein and NAD. The C-terminal (catalytic) domain and one beta-alpha unit of the N-terminal (nucleotide-binding) domain of the natural protein have been removed. To obtain this minimal structure, the order of the secondary structural units has been rearranged. An essential C-terminal alpha-helix has been included by re-connecting it to the N-terminus using a new loop, obtained from a data-base search. A disulphide bond has been introduced, which should link the end of this helix to a nearby loop, if the fragment assumes the structure we propose. The other key changes were to make surface residue substitutions to "solubilize" the protein. A unique Tryptophan residue was incorporated into the hydrophobic core to provide a spectophotometric probe of folding. The binding site of the NAD was modified as little as was consistent with the other changes. The main loss was a specific H-bonding contact to the nicotinamide, which was replaced by an interaction with a side chain from elsewhere in the protein. The designed protein contains features to allow experimental assessment by a variety of assays for both structure and function. It's small size (131 amino acids) is appropriate for structural evaluation by NMR methods.

## 4.2 DESIGN PHILOSOPHY.

The common interest of our group was in designs involving ligand-binding sites. Our initial plan was to transfer a fairly discretely defined site from one protein into a similar secondary structural environment, but in a different tertiary structural context.

We scanned the Brookhaven structural database (PDB) for all non-protein molecules bound and then looked at the following binding motifs to find a simple motif that could be transfered into a new structural background:
Key: 1XXX is the filename used in the Brookhaven structural database (PDB) () shows molecule bound.
1) Ribonucleases:
5RSA, 6RSA (vanadate), 1RNT (2′ gmp), 1RN3 (so4)
The binding site is composed of several disconnected secondary structural elements and non-regular loops.
2) Alpha-Beta barrels:
1TIM (nothing bound, but later looked at structures with inhibitors bound [thanks to M.Noble, EMBL]), 1GOX (fmn) and IGPS:PRAI [thanks to M.Wilmanns and J.N.Jansonius]
In a similar fashion to 1), the binding site is at end of the 8 strand beta-barrel and involves residues from many of the beta-strands and their loops.
3) Nucleotide binding proteins:
1LDM (nad,oxamate), 1PFK (adp,fructose 16 bis-posphate), 1ETU (gdp)
The NAD binding has a simple binding site compared to the others looked at, involving few secondary structural elements despite its extended nature.
We were not able to identify a binding site in any protein we examined which was suitably discretely defined to enable transfer in the manner we had initially envisioned. To generalize, the sites were formed from multiple loops, derived from areas of the protein far distant in the primary sequence.
We decided that it was too difficult to transfer binding sites involving loops with conformation specific to a particular protein family or sites involving large numbers of secondary structural elements or precise charge interactions. We revised our plans and began to investigate designs based on an NAD binding protein.
Key reasons which influenced this choice were the realization that several regions of the site were mainly hydrophobic and site residues were from relativily few secondary structural elements.

## 4.3 DESIGN DIARY.

### 4.3.1 EXPLORING THE NATURE OF NAD BINDING PROTEINS

The PDB database was searched for files containing NAD molecules. The following were found:
Key: [] similar datasets, linked by HSSP (1) files
* similar dataset without NAD bound

1GD1 (glyceraldehydePDH) [1GPD, 3GPD, 4GPD*]
1LDM (lactateDH) [2LDH, 3LDH, 1LLC*, 1LDX*, 5LDH*]
4MDH (malate dehydrogenase)
6ADH (alcoholDH) [8ADH*]

Contacts between NAD and atoms within 4.0 A were listed for 5 NAD binding structures using CONTACT (1LDM, 1GD1, 1GPD, 4MDH, 6ADH). The

five structures 1GD1 ,1GDP, 1LDM, 4MDH, and 6ADH were aligned in 3-D using the program COMPOSER (Sutcliffe et al, 1987ab). This showed broad structural variations in the NAD binding family. About 50 residues were identified by COMPOSER as being structurally equivalent in a 3-dimensional superposition. These included the 3 conserved Glycine residues (GxGxxG) in the loop interacting with the adenosine ring (Wierenga et al, 1986).
Structures were manually superimposed according to the position of NAD (with the program INSIGHT II). We could divide the structures into two groups. In one group (1LDM and 4MDH) the binding residues are in the C-terminal part of the molecule but in other groups several residues distant in sequence were involved in the binding. In LDH and MDH there is a long loop for the nicotinamide part of the binding cleft. In the other group an alpha-helix from the substrate binding domain of the enzyme is involved in binding the nicotinamide. We decided that we would use 1LDM (Zapatero et al, 1987) as a template for building NAD-binding protein. A minimal NAD binding fragment was defined from 1LDM to be residues 21-152. This was used for subsequent analysis.

### 4.3.2 LACTATE DEHYDROGENASE

Lactate dehydrogenase catalyses the reaction shown below (see Holbrook et al, 1975):
The enzyme is a tetramer of 4 identical subunits, each of which can be divided into a catalytic and functional domain. The NAD binding site can be considered as composed of five sections : Adenosine, ribose1 di-Phosphate, ribose2 and Nicotinamide binding regions:
The Adenosine is mainly bound by a loop between a beta-strand and an alpha helix. This interaction is essentially hydrophobic and so less specific than for other parts of the molecule: it has been demonstrated to bind aromatic dyes competitively (e.g. 5-iodosalicylate and iodofluorescein). This part of the designed binding site could therefore be tested by similar experiments.
Ribose1 2 OH makes two hydrogen bonds to Asp 52.
The di-Phosphate is bound by water and Arg 99. The binding of phosphate contributes to the binding energy since a modified substrate with only one phosphate replaced binds less tightly.
Ribose2 hydrogen bonds to mainchain C=O Ala 98 and side chain C=O of Asn 138.
The nicotinamide portion binds in a cavity which is hydrophobic on one side and hydrophilic on the other, substrate-binding face. In addition there are specific H-bonding interactions from a Ser to the amide of the nicotinamide ring, to the backbone C=O of Val 136 and to water molecules.

### 4.3.3 DEFINING THE MINIMAL NAD BINDING FRAGMENTS

The accessibility of native LDH was calculated without NAD and compared to that with NAD. Removing the NAD increased the exposed surface of the molecule by about 160 A2 from 4840 A2. Little of this

area came from exposed mainchain (20 A2) and most from non-polar side chain (100 A2). This defined the full list of residues interacting with the NAD in LDM.
The solvent accessibilities of LDM and LDM:21-152 were compared. This revealed the residues involved in the contact between the NAD binding domain and the substrate binding domain of LDM. The main changes in solvent accessibility between the wildtype and the 21-152 fragment were situated along helix 130-152 (exposed side chains).
We decided that residues 53 - 90 in LDM probably could be removed as they not were involved in the binding of NAD. However we also discovered a hydrophobic patch on the surface of the beta-sheet which makes interactions with a helix (H0) from the part that we had excluded. We decided that this helix should be included in our model and made the new N-terminus, with a connecting loop from Residues 263 to 22 in LDM.
Several attempts were made to replace LDM:54-90 (loop-helix-loop-beta-loop-helix-loop) by some other piece of structure (loop-helix-loop) using COMPOSER and DMSCAN (Hubbard, unpublished program). We searched for all loops in the database that had four CA atoms before the loop and four after it in the same configuration as LDM. We could not find any structures that filled this criteria and did not collide with other parts of our protein. We therefore decided to keep the 3-10 turn part of loop 3T (although it looked very strange) and search for loops connecting only residues 52 and 81. We used the same method (DMSCAN) as above for searching for loops connecting residues 242 to 22.
The long loop linking strand E4 and helix H3 that makes the binding of the nicotinamide region of NAD seemed to be too flexible in our model. To decrease the flexibility we introduced a disulphide bond between residue 243 (the N-terminal residue in the new fragment) residue 102.
After checking the change in surface accessibility between LDM and our fragment, for individual residues we tried to decrease the hydrophobicity of the surface that resulted from the removal of the rest of LDM by introducing appropiate point mutations.

### 4.3.4 CONNECTION OF FRAGMENTS AND BUILDING OF THE FIRST MODEL:

From now numbering used will be according to the first model of LEATHER:
Two loops were finally identified by DMSCAN and added:
Loop 1 connecting H0 to E1 from 2ENL.BRK (this loop was identified by DMSCAN). Since only CA's were in the PDB file, MAXSPROUT (Holm, in Press) was used to rebuild the mainchain and side chains to make the loop usable.
Loop 2 connecting E2 to 3T (deleting H3 and E4) originated from 2PLV 1. Similar loops were found in 1R08 1, 2MEV 2. All had GLY at second position. We decided to introduce a double Gly loop to make this short turn. Asp 57 makes two hydrogen bonds to one of the riboses in NAD with removing H2 and E3 will ASP 57 be flexible. We try to restrain Asp 57 by changing Val 33 to an Arg.
We removed an existing Trp, which had become surface exposed and was not desirable (This also enabled us to introduce a unique Trp in the interior for flourescence studies in model 2). Similarly, we removed

two Cys residues so that our designed Cys pair would form a unique intra-molecular disulfide.
The model resulting from this first round of changes was energy minimized with GROMOS for 300 steps Steepest descent and 100 steps conjugent gradient minimization.

### 4.3.5 TESTING OF THE FIRST MODEL

We used several different programs for testing the structure.
One of the best ways to study a structure is to try to look at it in all possible views. This is very much simplified by using stereo however access to stereo was too limited for us to use this approach extensively. The course as a whole realized that displaying surfaces and then slicing through the molecule with a very small slab thickness can be a very informative way to discover faults in the structure. By such inspections we found some hydrophobic surfaces which were solvent exposed and also a buried cleft. Also we noticed that the loop we selected to conect H0 and E1 looked very strange by visual inspection.
POLDIAGNOSTICS (Baumann et al, 1989) was used to judge the polarity of the protein both interior and on the surface of the fragment with and without the proposed mutations. The polarity was improved and "OK" (meaning that it was within the limits calculated for all natural proteins in PDB), while the old fragment was apparently too hydrophobic. However the overall polarity of the model was more than that of LDM.
A folding analysis by Moult and Anger (Biochemistry in press) was carried out on both the 1LDM and our model. This showed that a possible nucleation site in the region around res 30-57 has been lost, and that the buried hydrophobical surface area per residue might be slightly too small to guarantee a stable fold. A Monty Carlo folding simulation using this burial data failed to fold the protein by propagation, although good propagation is observed for simulations on natural alpha/beta protein structures..
In WHATIF (Vriend, 1990) there is an option for calculating the probability to find each part of every residue (divided in one or several parts) in the particular surrounding that it is a protein. If a residue's probability is more than five standard deviations from the mean value, is it unlikely to be found it in this particular environment. We found that Arg 33 that we had hoped would cover Asp 57 was in a bad environment, probably due to the aliphatic part of the side chain not being well placed.
We used HMDISP (Method of Eisenberg, 1982: Hubbard, unpublished program) to check the orientation of the hydrophobic moments of secondary structural elements. The moments looked resonable, pointed mostly inwards, as for the natural 1LDM structure.
QPACK (Lydia Gregoret, UCSF) was used for checking the volume that each residue had to move within. A few side chains were discovered that had either too much or too little room. The program also reported a side chain Pairwise interaction sum. The value (-5) was higher (worse) than the normal value of (-10 to -20). PACANA (John Moult, CARB) indicated a similar 'packing coefficient' to that found for natural proteins.
ELEANA (John Moult, CARB) analysed electrostatic group interactions

and discovered some bad contacts between charged groups.
Secondary structure predictions showed that the regions that were strongly predicted for 1LDM were similarly strongly predicted for the sequence of our model.

### 4.3.6 MAKING THE SECOND AND FINAL MODEL

We found that our protein had a slightly too hydrophillic surface, so we replaced several Arginine with less hydrophillic residues.
The loop that connects H0 and E1 looked very strange and had a bad scoring in several of our tests. We decided to try to find a loop that started earlier on our N-terminal helix (H0). We found one in a similar way as described above that had four residues connecting residues 19 and 24. This loop come from 8ADH (res 259 - 263). This loop looked better so we decided to use it.
We wanted to put in a tryptophan in the interior of our structure for fluorescent studies. If this tryptohan shows to be interior (by spectral or time resolved studies) could this be taken as an indication for that the fold was as we wanted it. We searched all residues that were a tyrosine or a phenylalanine for finding a suitable internal place for a tryptophan. Phe 96 was found to be suitable for this change. We used PROPAK to check if we could mutate this tryptophan without moving the backbone fill the empty space to about 85% of the volume that was occupied by possible mutated side chains near Phe 96. We could not find anything that satisfied our demands. After visual studies and similar analysis with MAXSPROUT (modified by Liisa Holm) we decided to replace ALA 126 by a TRP and change Phe 96 to an Ala and Ile 100 to an ALA. The TRP was placed in a suitable position using INSIGHT.
By looking at the suface did we found some "holes" that we replaced by replacing VAL 114 with an ILE and VAL 104 with an ILE. We hoped that a smoother surface with these mutations.
A putative nucleation site found in the N-terminus of LDM was absent from our model. We tried to alter this region to increase the local hydrophobic burial by replacing two of the changes we had made in model 1 for hydrophobics.
We mutated residues 2 (GLY) and 81 (GLU) to form a disulphide bond. The N-terminal backbone was moved to a suitable possition. Since it had been recommended not to make Cys one of the first 2 residues, we included an extra N-terminal residue (MET). The first residue was mutated from Lys to Gly to remove one of the electrostatical collisions discovered by ELEANA.
The Nicotine amide part of NAD seemed not to be tightly bound so we tried to make an additional hydrogen bound between the Nitrate group on the end of NAD and a SER in position 119 (VAL in Model 1)

### 4.3.7 TESTING OF THE SECOND MODEL

This model was minimized for 300 step by Steepest Descents.
After the minimization was the structure checked according to WHATIF probabilities for each residue's environment. The overall structure

was better than model 1, but we still found some bad regions:
N-termini: We hoped that it would adopt a better position after the MD-run. The Met 1 is probably no good but we nded it for expression.
Loop1: Residue 21 and 22. This was the best loop we could find so we hope that the rest of the structure is good enough to fold even if this loop not is perfect.
Tyr 33: We hoped that this will flip around during the MD.
Tyr 55: Bad contact with Tyr 33.
106 and 109 : Outside charged group that could be investigated further.
C-termini :Will probably improve during MD. Might have to look further in this region.
The polar surfaces according to POLDIAGNOSTICS had been improved compared to the previous model, but the mutations to increase hydrophobic burial did not seem to make much difference. The pair-pair residue interactions sum (QPACK) did however improve to -10.5.
Molecular Dynamics was subsequently applied for 1ps with a timestep of 0.5 fs without SHAKE. We checked the regions that had moved most during this simulation. The overall RMS was about 0.4 A. A further simulation of 10ps was applied with SHAKE and a step size of 2fs. This produced large movements (2.7 A side chain, 2.4 A mainchain) including a substantial rearrangement of the N-terminus and disulphide into a perhaps better confomation. However, the results of such a simulation should be treated with caution due to the exclusion of water. Two final coordinate sets are therefore deposited: 1LEA (after minimisation only) and 2LEA (after molecular dynamics and subsequent minimisation)

## 4.4 CONCLUSIONS AND PROPOSED TESTING OF LEATHER

The designed protein fragment is 131 amino acids long. It will be possible to conveniently synthesise the gene encoding the protein for expression in bacteria.
1) All Trp residues were removed from the protein. We then designed a unique Trp at an interior position. Thererfore, it will be possible to use Trp fluorescence to monitor protein folding and to give an indication of the residue's burrial and motion in the folded protein.
2) A specific disulfide has been introduced which will link a flexible N-terminal extension of a helix to a loop, if the overall fold of the protein is as we imagine. Therefore this will provide an additional test of the structure.
3) The NAD binding function can be conveniently assayed by monitoring changes in optical absorption and fluorescence when NAD or its derivatives bind to the protein. The different regions of the NAD have essentially independent binding and we will be able to monitor the different fragments independently. For example, AMP, ADP and NAD binding can be followed. This will allow us to determine if different binding regions have been maintained in the fragment.
4) Additional techniques are obviously CD, NMR (the fragment is small enough) and x-ray crystallography.

## 4.5 ASSESSMENT OF TOOLS:

### 4.5.1 DREAMS AND REALITIES.

Ideally this workshop would be an exploration of possible protein design concepts using a computer interface. Computers are needed to manipulate the numbers representing the positions of thousands of atoms in a single structure. Many interfaces are available frequently providing hundreds of possible options as tools for the designer. Unfortunately all have deficiencies and bugs. A greater fraction of time in this course has therefore been spent pursuading programs (by all kinds of cheating, fixing and deviousness) to carry out the calculation required, than actually assessing the programs results. Some calculations did not get carried out, not because of the unavailability of the technique, but the difficulty of running the program.
This is partly inevitable, but does show that easy to use, reliable and well explained programs are frequently the most generally helpful.

### 4.5.2 BUILDING A MODEL

4.5.2.1 Loop Search - What is the best way to find a suitable loop to connect two pieces of secondary structure? There is two general different methods to do this:
A) Clever guesses. It is no problem to place the CA atoms in suitable positions and then fill it up with nice side chains. But, with this method do you really get a loop that both looks good and can exist? One way of checking this method would be to place the CA's and then search the database for loops with this conformation, taking the side chains from that loop.
B) Searching for the structural information that surrounds the loop. This method has the problem that even if a loop fits the backbone conformation at both ends, it might not come from similar surroundings. One might not discover what it is that makes this particular loop to be in this conformation.
The main difference between these methods arises from whether you believe that loops are structural stable by themselves or if they adopt this particular configuration due to what constrains their ends. A problem with both these methods is that there might not be any suitable loops in the database. It is then uncertain if a particular connection is impossible or that nature just has not have used it in structures in the present database. As an example it seems reasonable that poly GLY loops should work in a protein with an stable secondary structure but to our knowledge has this never been seen.

4.5.2.2 Molecular Graphics - With the new Molecular Graphical programs is it very easy to build proteins on the screen. There is no problem to twist dihedral angles. The only problem is sometimes the lack of stereo (see above). Sometimes it would probably be as easy to run some Molecular Dynamics just to move atoms close together, but the MD programs are still slow and can be difficult to use correctly.

## 4.6 CONCLUSIONS ABOUT PROGRAMS THAT EXIST (OUGHT TO EXIST)

There ought to be a more efficient way to redesign a protein than just looking at the structure and testing it and performing several point mutations at a time. One mutation that improves the results in one test might decrease the score in another. At the moment most reliance is placed on visual inspection as testing programs mostly can only act as guide to where major problems might be.

## 4.7 REFERENCES

Baumann, G., Frommel, C. and Sander, C. (1989) Prot. Eng., 2, 329-334.

Holbrook, J. J., Liljas, A., Steindel, S. J. and Rossmann, M. G. (1975) In The Enzymes Volume XI Oxidation-Reduction Part 4, 3rd ed., Academic Press, New York, San Franscisco, London.

Sutcliffe, M.J., Haneef, I., Carney, D. and Blundell, T.L. (1987a) Knowledge based modelling of homologous proteins, part I: three-dimensional frameworks derived from the simultaneous superposition of multiple structures. Prot. Eng., 1, 377-384.

Sutcliffe, M.J., Hayes, F.R.F. and Blundell, T.L. (1987b) Knowledge based modelling of homologous proteins, part II: rules for the conformations of substituted side chains. Prot. Eng., 1, 385-392.

Wierenga, R.K., Terpstra, P., and Hol, W. G. J. (1986) J. Mol. Biol. 187, 101-107.

Zapatero, C. A., Griffith, J. P., Sussman, J. L., and Rossmann, M. G. (1987) J. Mol. Biol. 198, 445.

NOTES

(1) HSSP files contain alignments for each brookhaven PDB file sequence to any homologous protein sequence.

## 4.8 APPENDICES

1) DSSP of Final model
2) Nomenclature in the model

```
In 1LDM:            E1-H1-E2-H2-E3-3T-E4-H3-E5-H4
In Model:        H0-E1-H1-E2-------3T-E4-H3-E5-H4
```

| Number in model 1 | LEATHER | Number in LDM | Comment |
|---|---|---|---|
| 0 | 1 | | N-termini Met |
| 1 - 21 | 2 - 19 | 240 - 261 | H0 |
| 21 - 26 | 20 - 26 | | Loop |

| | | | |
|---|---|---|---|
| 27 - 57 | 27 - 57 | 22 - 72 | E1, H1, E2, |
| 58 - 59 | 58 - 59 | | Loop |
| 60 - 131 | 60 - 131 | 81 - 152 | 3T, E4, H3, E5, H4 |

# APPENDICES

## 1) DSSP of Final model

```
**** SECONDARY STRUCTURE DEFINITION BY THE PROGRAM DSSP, VERSION OCT. 1985 **** MONTH=12 DAY=24 YEAR=2001
REFERENCE W. KABSCH AND C.SANDER, BIOPOLYMERS 22 (1983) 2577-2637
  131  1  0  0  0 TOTAL NUMBER OF RESIDUES, NUMBER OF CHAINS, NUMBER OF SS-BRIDGES(TOTAL,INTRACHAIN,INTERCHAIN)
 7500.0   ACCESSIBLE SURFACE OF PROTEIN (ANGSTROM**2)
   97 74.0    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(J)  , SAME NUMBER PER 100 RESIDUES
   16 12.2    TOTAL NUMBER OF HYDROGEN BONDS IN     PARALLEL BRIDGES, SAME NUMBER PER 100 RESIDUES
    0  0.0    TOTAL NUMBER OF HYDROGEN BONDS IN ANTIPARALLEL BRIDGES, SAME NUMBER PER 100 RESIDUES
    0  0.0    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-5), SAME NUMBER PER 100 RESIDUES
    0  0.0    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-4), SAME NUMBER PER 100 RESIDUES
    0  0.0    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-3), SAME NUMBER PER 100 RESIDUES
    0  0.0    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-2), SAME NUMBER PER 100 RESIDUES
    0  0.0    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-1), SAME NUMBER PER 100 RESIDUES
    0  0.0    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+0), SAME NUMBER PER 100 RESIDUES
    0  0.0    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+1), SAME NUMBER PER 100 RESIDUES
    0  0.0    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+2), SAME NUMBER PER 100 RESIDUES
   11  8.4    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+3), SAME NUMBER PER 100 RESIDUES
   14 10.7    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+4), SAME NUMBER PER 100 RESIDUES
   48 36.6    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+5), SAME NUMBER PER 100 RESIDUES
    2  1.5    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(J)  , SAME NUMBER PER 100 RESIDUES

                                                                      *** HISTOGRAMS OF ***
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30      RESIDUES PER ALPHA HELIX
 0  0  0  0  0  0  0  0  0  0  0  2  0  0  1  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0      PARALLEL BRIDGES PER LADDER
 0  0  0  2  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0      ANTIPARALLEL BRIDGES PER LADDER
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0      LADDERS PER SHEET
 0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0

  #  RESIDUE AA STRUCTURE BP1 BP2  ACC    N-H-->O    O-->H-N    N-H-->O    O-->H-N      TCO  KAPPA ALPHA   PHI     PSI    X-CA   Y-CA   Z-CA
  1    1  M              0   0  207    0, 0.0     2,-0.7     0, 0.0     3, 0.0     0.000 360.0 360.0 360.0   128.5  -3.2   14.6  -17.3
  2    2  G            + 0   0   81    1,-0.1     0, 0.0     0, 0.0     0, 0.0    -0.905 360.0 102.2-108.7    99.1  -0.9   12.0  -18.7
  3    3  C            + 0   0   51   -2,-0.7     2,-0.3    77, 0.0    -1,-0.1    -0.032  61.2  79.8-169.9    50.5   2.5   11.9  -16.7
  4    4  G            - 0   0   51   76, 0.0     2,-0.4    77, 0.0    76, 0.0    -0.973  65.3-118.1-158.4   160.7   2.4    8.9  -14.3
  5    5  T            + 0   0  110   -2,-0.3    31,-0.3     1,-0.1    30, 0.0    -0.875  43.6 134.9-112.1   133.9   2.8    5.1  -14.0
  6    6  S          >> + 0   0   33   -2,-0.4     4,-1.9    29,-0.1     3,-0.9     0.303  59.6  52.3-131.4   -74.6   0.1    2.7  -12.9
  7    7  D  H 3> S+   0   0  127    1,-0.2     4,-2.3     2,-0.2     5,-0.2     0.742 105.5  54.1 -49.1   -48.4  -0.5   -0.6  -14.8
  8    8  A  H 3> S+   0   0   86    1,-0.2     4,-2.3     2,-0.2    -1,-0.2     0.913 113.7  43.2 -54.6   -42.9   3.1   -2.0  -14.7
  9    9  G  H <> S+   0   0   45   -3,-0.9     4,-3.1     2,-0.2     5,-0.3     0.858 107.2  60.3 -73.8   -40.0   3.3   -1.7  -10.9
 10   10  G  H  X S+   0   0    0   -4,-1.9     4,-2.2     2,-0.2    -1,-0.2     0.949 109.9  43.6 -46.8   -53.0  -0.3   -3.1  -10.5
 11   11  A  H  X S+   0   0   54   -4,-2.3     4,-2.7     2,-0.2    -2,-0.2     0.944 113.0  51.8 -58.8   -50.3   1.0   -6.3  -12.2
 12   12  K  H  X S+   0   0  137   -4,-2.3     4,-2.3     2,-0.2    -2,-0.2     0.920 112.7  43.5 -54.0   -52.6   4.3   -6.2  -10.1
 13   13  V  H  X S+   0   0   10   -4,-3.1     4,-2.9     2,-0.2    -1,-0.2     0.894 112.0  53.7 -67.0   -37.3   2.5   -5.9   -6.7
 14   14  A  H  X S+   0   0    1   -4,-2.2     4,-2.9    -5,-0.3    -2,-0.2     0.931 107.6  51.8 -59.7   -44.3  -0.1   -8.6   -7.7
 15   15  D  H  X S+   0   0   77   -4,-2.7     4,-2.4     1,-0.2    -2,-0.2     0.959 111.4  46.9 -54.1   -50.9   2.8  -10.9   -8.5
 16   16  D  H  X S+   0   0   58   -4,-2.3     4,-2.0     2,-0.2     5,-0.2     0.883 112.4  49.4 -60.2   -45.4   4.3  -10.3   -5.0
 17   17  A  H  X S+   0   0    3   -4,-2.9     4,-1.3     1,-0.2    -2,-0.2     0.976 111.2  49.5 -57.9   -50.4   0.8  -10.8   -3.3
 18   18  E  H  X S+   0   0   82   -4,-2.9     4,-1.5     1,-0.2     5,-1.3     0.856 113.4  47.6 -54.8   -48.3   0.3  -14.1   -5.3
 19   19  T  H  X S+   0   0   87   -4,-2.4     3,-0.9     1,-0.2    -2,-0.2     0.976 106.7  51.1 -57.0   -47.4   3.8  -15.4   -4.3
 20   20  K  H  X S+   0   0  129   -4,-2.0    -1,-0.2     1,-0.3    -2,-0.2     0.672 119.3  39.0 -49.8   -35.8   4.0  -14.7   -0.5
 21   21  M  H  X S+   0   0   51   -4,-1.3    -1,-0.3    -5,-0.2    -2,-0.2     0.865  97.8-124.8 -64.6   -54.4   0.6  -16.5    0.0
 22   22  N  T  <5+   0   0  157   -4,-1.5     2,-0.3    -3,-0.9    -3,-0.2     0.889  96.5   -4.6  64.9    71.3   0.5  -19.4   -2.2
 23   23  G  S    <5-   0   0   48   -5,-1.3     2,-0.2    -4, 0.0    -5,-0.1    -0.417 121.7 -41.8 127.9   -56.8  -2.7  -18.2   -3.9
 24   24  G            -   0   0   25   -2,-0.3    26,-0.1    -7,-0.1    25,-0.1    -0.663  56.3  -93.4 156.2   141.0  -4.2  -15.0   -2.3
 25   25  V            -   0   0    9   -2,-0.2    26,-1.1     1,-0.1    27,-0.2    -0.225  24.4-126.4 -64.5   156.9  -4.9  -13.4    1.1
 26   26  N  S    S+   0   0   57    1,-0.3    26,-2.1    24,-0.2    25,-1.1     0.978  99.8   8.6 -56.4   -49.5  -8.1  -13.5    3.2
 27   27  K  E    S-a 52   0A  19   40,-0.3    42,-1.7    24,-0.2    43,-1.4    -0.935 71.1-167.3-126.2   112.5  -8.4   -9.7    3.4
 28   28  I  E     -ab 53  70A   0   24,-2.2    26,-2.8    -2,-0.5     2,-0.5    -0.835  7.8-154.4 -92.1   135.8  -6.1   -7.3    1.4
 29   29  T  E     -ab 54  71A   2   41,-2.2    43,-3.1    -2,-0.4     2,-0.7    -0.982  6.3-169.6-109.0   121.6  -6.0   -3.6    2.3
 30   30  V  E     -ab 55  72A   1   24,-1.8     2,-0.9    -2,-0.5    26,-0.9    -0.910  9.4-163.7-108.4    97.7  -5.1   -1.2   -0.5
 31   31  V  E     +ab 56  73A   2   41,-2.8    43,-2.6    -2,-0.7    44,-1.4    -0.799  63.2  43.3 -90.1   107.1  -4.5    2.2    1.1
 32   32  G  +     -   0   0    7   24,-1.5     2,-1.3    -2,-0.9    25,-0.2     0.472  64.0-149.5 117.4    99.1  -4.6    5.0   -1.5
 33   33  Y  +     +   0   0  121    4,-0.2     2,-0.1     3,-0.1    24,-0.1    -0.239  57.4 118.4 -90.3    49.7  -7.2    5.0   -4.3
 34   34  G  S  >  S-   0   0   39   -2,-1.3     4,-2.9    40,-0.1     5,-0.3    -0.247  84.9 -66.6 -96.2-165.8  -5.0    6.7   -6.9
 35   35  A  H  >  S+   0   0   48    1,-0.2     4,-1.8     2,-0.2   -29,-0.1     0.916 137.9  34.5 -49.8   -59.7  -3.6    5.7  -10.3
 36   36  V  H  >  S+   0   0   28  -31,-0.3     4,-3.1     2,-0.2    -1,-0.2     0.898 115.2  60.0 -65.6   -34.4  -1.4    2.9   -8.9
 37   37  G  H  X  S+   0   0    1    2,-0.2     4,-2.8     1,-0.2    -2,-0.2     0.932 107.9  42.5 -59.4   -55.3  -3.9    2.1   -6.1
 38   38  A  H  X  S+   0   0   58   -4,-2.9     4,-3.1     2,-0.2    -1,-0.2     0.904 114.1  53.6 -57.5   -42.2  -4.7    1.2   -8.6
 39   39  A  H  X  S+   0   0   11   -4,-1.8     4,-2.5    -5,-0.3    -2,-0.2     0.955 111.1  44.9 -57.9   -53.2  -4.1   -0.7  -10.7
 40   40  S  H  X  S+   0   0    0   -4,-3.1     4,-2.7     2,-0.2    -2,-0.2     0.947 115.5  48.3 -52.8   -53.1  -3.0   -2.8   -7.7
 41   41  A  H  X  S+   0   0    6   -4,-2.8     4,-2.7     1,-0.2    -2,-0.2     0.903 111.7  48.3 -55.5   -51.8  -4.6   -3.4   -6.7
 42   42  A  H  X  S+   0   0   57   -4,-3.1     4,-3.1     2,-0.2    -1,-0.2     0.912 112.9  48.1 -58.9   -46.8  -7.8   -4.5  -10.2

 43   43  S  H  X  S+   0   0   31   -4,-2.5     4,-2.6     2,-0.2    -2,-0.2     0.949 112.2  49.1 -59.1   -49.3  -4.8   -6.9  -10.6
 44   44  I  H  <  S+   0   0    2   -4,-2.7     5,-2.2     2,-0.2    -2,-0.2     0.904 114.2  46.8 -53.6   -50.6  -5.5   -8.4   -7.2
 45   45  L  H  X  S+   0   0   75   -4,-2.7     3,-2.2    -5,-0.2    -2,-0.2     0.966 111.4  49.7 -55.8   -54.4  -9.2   -8.8   -8.1
 46   46  A  H  X  S+   0   0   81   -4,-3.1    -2,-0.2     1,-0.3    -1,-0.2     0.848 109.0  53.5 -56.7   -38.8  -8.5  -10.3  -11.5
 47   47  K  T  3  S+   0   0   82   -4,-2.6    -1,-0.3    -5,-0.2    -2,-0.2     0.323 110.5-122.5 -79.9    8.2   -6.0  -12.9   -9.9
 48   48  D  T  <  S+   0   0  146   -3,-2.2    -3,-0.2    -5,-0.1    -2,-0.1     0.744  63.4 146.6  48.7    38.3  -8.8  -14.0   -7.4
 49   49  L  A    +   0   0    6   -5,-2.2     2,-0.3    -6,-0.2    -4,-0.1     0.980  47.9  40.6 -67.5   -66.5  -6.4  -13.0   -4.6
 50   50  A  A    -   0   0   21   -6,-0.1   -24,-0.2   -26,-0.1   -23,-0.2    -0.679  53.6-153.6 -97.9   137.9  -8.5  -11.8   -1.7
 51   51  D  S    S+   0   0  128  -26,-1.1     2,-0.3   -25,-1.1   -24,-0.2     0.673  99.6  21.9 -66.8   -26.0 -11.8  -13.1   -0.3
 52   52  Q  E     +a  27   0A  97  -26,-2.1   -24,-2.2   -27,-0.2     2,-0.7    -0.960  69.5-160.6-141.7   127.5 -12.2   -9.5    0.8
 53   53  V  E     -a  28   0A  32   -2,-0.3     2,-0.4   -26,-0.2   -24,-0.2    -0.948  11.5-174.9-106.9   103.1 -10.5   -6.3   -0.5
 54   54  A  E     -a  29   0A   9  -26,-2.8   -24,-1.8    -2,-0.7     2,-0.3    -0.762   6.8-176.2 -85.3   134.1 -10.7   -3.5    2.0
 55   55  Y  E     -a  30   0A  76   -2,-0.4     2,-0.4   -26,-0.3   -24,-0.2    -0.967   4.5-164.8-134.5   141.9  -9.4   -0.3    0.6
 56   56  V  E     -a  31   0A   5  -26,-0.9   -24,-1.5    -2,-0.3     2,-0.3    -0.950   2.1-166.6-135.0   135.1  -8.9    3.0    2.5
 57   57  D  >     -   0   0   80   -2,-0.4     3,-1.5   -26,-0.2   -26, 0.0    -0.900  49.0 -34.9-121.2   154.5  -8.4    6.7    1.4
 58   58  G  T  3  S-   0   0   71   -2,-0.3     2,-0.5     1,-0.2    -1,-0.1     0.393 117.6 -20.6  -3.0    95.3  -7.3   10.0    3.0
 59   59  G  T  3  S+   0   0   79    1,-0.2    -1,-0.2    -3,-0.1     3,-0.2    -0.054  97.4 127.2  93.0   -29.9  -4.7   10.4    4.6
 60   60  S  S  <  S-   0   0   82   -3,-1.5     2,-0.2    -2,-0.5    -1,-0.2    -0.305  73.3 -61.8 -71.5   153.9 -11.6    7.9    6.5
 61   61  D  >  >  S+   0   0  107    1,-0.1     3,-2.0    -5,-0.1    -1,-0.2    -0.017  34.7-163.9 -52.9   107.3 -12.1    5.1    9.1
 62   62  Y  G  >  S+   0   0   19    1,-0.3     3,-2.5    -2,-0.2    -1,-0.1     0.585  77.6  90.0 -61.0   -14.4  -9.3    2.5    9.2
 63   63  S  G  >  S+   0   0   55    1,-0.3     3,-1.8     2,-0.2    -1,-0.3     0.912  76.6  70.1 -48.1   -32.8 -11.7    0.2   11.2
 64   64  D  G  X  S+   0   0  100   -3,-2.0    -1,-0.3     1,-0.3    -2,-0.2     0.656  93.3  54.8 -57.2   -24.3 -12.5   -1.0    7.6
 65   65  S  G  X  S+   0   0    3   -3,-2.5     3,-1.0    -4,-0.1    -1,-0.3     0.062  78.2 137.7-104.2    24.3  -9.0   -2.6    7.3
 66   66  A  T  <  S+   0   0   34   -3,-1.8    41,-0.2     1,-0.2    40,-0.1    -0.502  68.4  13.4 -73.4   131.8  -9.3   -4.8   10.5
 67   67  G  T  3  S+   0   0   43   39,-1.2   -40,-0.3     1,-0.3    -1,-0.2     0.583  85.4 155.5  76.9    14.8  -8.0   -8.4   10.4
 68   68  S  S  <  -   0   0    4   -3,-1.0    -1,-0.3     1,-0.1   -40,-0.2    -0.438  35.2-161.6 -65.0   139.6  -6.1   -7.9    7.0
 69   69  K  S     S+   0   0  105   -42,-1.7    42,-2.8     1,-0.3     2,-0.4     0.720  81.9  35.3 -90.6   -29.1  -3.2  -10.3    6.3
 70   70  L  E     -bd 28 111A   7   -43,-1.4   -41,-2.2    40,-0.2     2,-0.4    -0.983  61.5-174.7-130.0   129.9  -1.8   -7.9    3.7
 71   71  V  E     -bd 29 112A   3   40,-2.8    42,-3.4    -2,-0.4     2,-0.5    -0.934   7.4-167.7-124.6   112.9  -1.8   -4.1    3.6
 72   72  V  E     -bd 30 113A   4   -43,-3.1   -41,-2.8    -2,-0.4     2,-0.6    -0.836  11.1-150.2-100.9   121.1  -0.4   -2.4    0.4
 73   73  I  E     -bd 31 114A   0   40,-2.9    42,-2.8    -2,-0.5    43,-1.6    -0.919  33.8 154.4-106.5   119.5   0.2    1.3    0.7
 74   74  T  +         0   0   16   -43,-2.6   -42,-0.2    -2,-0.6   -37,-0.2     0.203  35.6 129.9-115.9   -3.9  -0.1    3.6   -2.3
 75   75  A  +         0   0   17   -44,-1.4     2,-0.2    41,-0.1    41,-0.1    -0.433  43.4  55.3 -63.1   126.2  -0.9    6.8   -0.3
 76   76  G  S    S-   0   0   44    -2,-0.1     2,-0.1     19,-0.1    41,-0.1    -0.637  81.6 -43.4 142.0   174.3   1.2    9.9   -1.1
 77   77  A  -         0   0   32    -2,-0.2    40,-0.3    11,-0.2     2,-0.2    -0.373  49.5-144.2 -70.3   143.4   2.6   12.4   -3.6
 78   78  R  -         0   0  144    -2,-0.1     2,-0.2    38,-0.1    -1,-0.1    -0.704  28.0 -80.7-109.4   162.3   3.9   11.4   -7.0
 79   79  Q  -         0   0   58    -2,-0.2     2,-0.3     4,-0.1     6,-0.1    -0.420  34.0-141.3 -63.1   119.4   6.7   12.6   -9.3
 80   80  S  >  -     0   0   70    -2,-0.2     3,-2.2     4,-0.1     2,-0.1    -0.645  52.6 -78.0 -69.6   140.4   6.1   15.7  -11.4
 81   81  C  T  3  S+   0   0   90    -2,-0.3    -1,-0.1     1,-0.3     3,-0.1    -0.336 125.2  19.4 -57.0   108.5   7.7   15.0  -14.8
 82   82  G  T  3  S+   0   0   87     1,-0.3    -1,-0.3    -2,-0.1     2,-0.1    -0.238  98.0 115.2 120.9   -34.1  11.4   15.6  -14.2
 83   83  E  <  -      0   0   35    -3,-2.2    -1,-0.3     1,-0.1     2,-0.2    -0.377  63.8-124.4 -71.4   135.3  11.4   15.3  -10.3
 84   84  S  -         0   0   60    -2,-0.1     3,-1.2     1,-0.1     4,-0.4    -0.528  18.8-118.0 -78.3   148.2  13.3   12.4   -8.6
 85   85  R  G  >  S+   0   0  140     1,-0.2     3,-1.6    -2,-0.2     4,-0.3     0.865 112.6  65.2 -47.6   -37.8  11.5    9.9   -6.3
 86   86  E  G  >  S+   0   0  150     1,-0.3     3,-1.7     2,-0.2    -1,-0.2     0.830  88.0  66.5 -61.3   -31.7  13.9   11.1   -3.5
 87   87  N  G  <  S+   0   0  100    -3,-1.2     4,-0.3     1,-0.3    -1,-0.3     0.727  90.7  57.0 -65.0   -19.6  12.3   14.6   -3.5
 88   88  L  G  <  S+   0   0    0    -3,-1.6     4,-2.2    -4,-0.4     5,-0.3     0.670  71.5 100.1 -74.7   -17.8   9.0   13.1   -2.2
 89   89  V  H  <  S+   0   0   59    -3,-1.7     4,-3.2    -4,-0.3     5,-0.3     0.789  80.5  43.2 -39.3   -65.1  10.4   12.1    1.3
 90   90  Q  H  >  S+   0   0  124    -3,-0.3     4,-2.4     2,-0.2    -1,-0.2     0.884 116.6  46.8 -55.8   -48.6   9.1   14.9    3.5
 91   91  R  H  >  S+   0   0  100    -4,-0.3     4,-2.0     1,-0.2    -1,-0.2     0.947 118.9  40.9 -56.0   -49.2   5.6   15.0    2.0
 92   92  N  H  X  S+   0   0    2    -4,-2.2     4,-3.0     2,-0.2    -2,-0.2     0.796 111.0  54.4 -80.0   -27.8   5.2   11.2    2.2
 93   93  V  H  X  S+   0   0   31    -4,-3.2     4,-2.8    -5,-0.3    -1,-0.2     0.944 108.6  51.8 -67.3   -40.2   6.8   10.7    5.6
 94   94  N  H  X  S+   0   0   71    -4,-2.4     4,-0.8    -5,-0.3    -2,-0.2     0.958 112.8  44.8 -55.7   -48.9   4.3   13.2    7.0
 95   95  I  H  >X S+   0   0   58    -4,-2.0     4,-1.9     1,-0.2     3,-1.2     0.936 112.2  52.1 -60.2   -49.3   1.5   11.1    5.3
 96   96  A  H  3X S+   0   0    0    -4,-3.0     4,-2.5     1,-0.3    -1,-0.2     0.867  98.4  66.0 -55.4   -42.8   3.1    7.8    6.7
 97   97  K  H  3X S+   0   0  137    -4,-2.8    -1,-0.3     1,-0.2    -2,-0.2     0.749 109.6  36.8 -49.1   -39.5   3.2    9.3   10.3
 98   98  Y  H  <X S+   0   0  181    -3,-1.2     4,-1.0    -4,-0.8    -1,-0.2     0.918 120.4  46.7 -79.2   -46.5  -0.7    9.3   10.4
 99   99  I  H  X  S+   0   0   14    -4,-1.9     4,-2.1     1,-0.2    -2,-0.2     0.926 101.6  59.4 -64.1   -56.8  -1.3    6.1    8.4
100  100  A  H  X  S+   0   0    3    -4,-2.5     4,-1.7     2,-0.2     5,-0.3     0.820 103.7  49.6 -53.5   -45.1   1.2    3.5   10.0
101  101  P  H  >  S+   0   0   58     0, 0.0     4,-2.4     0, 0.0    -1,-0.2     0.967 113.3  48.1 -54.5   -49.7  -0.2    3.6   13.6
102  102  N  H  X  S+   0   0   62    -4,-1.0     4,-2.1     1,-0.2     5,-0.2     0.851 108.1  55.2 -61.7   -42.9  -3.8    3.1   12.3
103  103  I  H  X  S+   0   0    4    -4,-2.1     4,-2.3     1,-0.2     3,-0.3     0.952 115.0  35.9 -50.6   -65.0  -2.8    0.2   10.1
104  104  Q  H  <  S+   0   0   47    -4,-1.7    -1,-0.2     1,-0.2    -2,-0.2     0.778 111.0  62.0 -69.3   -30.0  -1.2   -2.0   12.8
105  105  K  H  <  S+   0   0  148    -4,-2.4    -1,-0.2    -5,-0.3    -2,-0.2     0.932 115.3  34.7 -54.2   -46.9  -3.8   -0.9   15.4
106  106  Q  H  <  S+   0   0   66    -4,-2.1   -39,-1.2    -3,-0.3    -2,-0.2     0.760 133.5  27.7 -79.3   -35.6  -6.4   -2.5   13.2
107  107  S     >< +   0   0    5    -4,-2.3     3,-1.0   -41,-0.2    -1,-0.3    -0.786  64.9 174.0-129.0    88.8  -4.3   -5.4   11.9
108  108  P  T 3  S+   0   0  103     0, 0.0    -1,-0.1     0, 0.0    -4,-0.1     0.868  81.4  45.3 -61.7   -35.7  -1.5   -6.3   14.5
```

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 109 | 109 | E | T | 3 | S+ | 0 | 0 | 187 | -41,-0.1 | 2,-0.2 | 2,-0.1 | -5, 0.0 | 0.043 | 90.3 | 134.1 | -102.1 | 32.6 | -0.3 | -9.4 | 12.6 |
| 110 | 110 | V | | < | - | 0 | 0 | 17 | -3,-1.0 | 2,-0.7 | -6,-0.1 | -40,-0.2 | -0.493 | 57.8 | -126.3 | -96.9 | 147.0 | -0.2 | -7.8 | 9.1 |
| 111 | 111 | T | E | | +c | 70 | 0A | 67 | -42,-2.8 | -40,-2.8 | -2,-0.2 | 2,-0.6 | -0.779 | 28.0 | -148.3 | -84.3 | 109.9 | 2.3 | -7.8 | 6.2 |
| 112 | 112 | I | E | | +c | 71 | 0A | 45 | -2,-0.7 | 2,-0.5 | -42,-0.2 | -40,-0.2 | -0.800 | 14.1 | -171.9 | -82.4 | 119.7 | 2.9 | -4.1 | 5.3 |
| 113 | 113 | L | E | | +c | 72 | 0A | 28 | -42,-3.4 | -40,-2.9 | -2,-0.6 | 2,-0.5 | -0.937 | 5.6 | -163.4 | -119.7 | 102.8 | 3.7 | -3.7 | 1.6 |
| 114 | 114 | I | E | | +c | 73 | 0A | 22 | -2,-0.5 | -40,-0.2 | -42,-0.2 | 9,-0.1 | -0.794 | 24.9 | 158.8 | -95.8 | 122.2 | 4.8 | -0.1 | 0.8 |
| 115 | 115 | V | | | + | 0 | 0 | 14 | -42,-2.8 | 4,-0.4 | -2,-0.5 | -41,-0.2 | 0.692 | 39.1 | 114.2 | -109.6 | -28.3 | 4.8 | 1.1 | -2.8 |
| 116 | 116 | S | S | | S- | 0 | 0 | 10 | -43,-1.6 | 6,-0.2 | 2,-0.1 | 7,-0.2 | -0.065 | 75.8 | -70.1 | -47.0 | 133.4 | 4.7 | 4.9 | -2.3 |
| 117 | 117 | N | S | | S+ | 0 | 0 | 44 | 1,-0.3 | -25,-0.1 | -40,-0.3 | -1,-0.1 | -0.512 | 121.8 | 35.9 | -78.1 | 140.4 | 7.8 | 6.8 | -3.5 |
| 118 | 118 | P | S | > | S+ | 0 | 0 | 11 | 0, 0.0 | 4,-2.9 | 0, 0.0 | 3,-0.4 | -0.019 | 76.9 | 179.6 | -89.0 | 99.4 | 10.6 | 6.7 | -2.4 |
| 119 | 119 | S | H | > | S+ | 0 | 0 | 37 | -4,-0.4 | 4,-3.7 | 1,-0.2 | 5,-0.2 | 0.703 | 76.3 | 47.9 | -17.3 | -62.5 | 9.9 | 3.0 | -1.6 |
| 120 | 120 | D | H | > | S+ | 0 | 0 | 133 | 2,-0.2 | 4,-1.6 | 1,-0.2 | -1,-0.2 | 0.935 | 119.0 | 37.3 | -59.9 | -54.1 | 13.4 | 2.2 | -0.1 |
| 121 | 121 | A | H | > | S+ | 0 | 0 | 31 | -3,-0.4 | 4,-1.7 | 2,-0.2 | -1,-0.2 | 0.924 | 119.9 | 51.8 | -63.1 | -40.6 | 13.6 | 5.3 | 2.3 |
| 122 | 122 | L | H | X | S+ | 0 | 0 | 1 | -4,-2.9 | 4,-2.1 | 2,-0.2 | 3,-0.5 | 0.961 | 103.1 | 54.2 | -63.5 | -51.8 | 9.8 | 4.9 | 2.9 |
| 123 | 123 | T | H | X | S+ | 0 | 0 | 41 | -4,-3.7 | 4,-2.6 | 1,-0.3 | 5,-0.2 | 0.864 | 107.8 | 54.3 | -52.2 | -40.0 | 9.8 | 1.1 | 3.9 |
| 124 | 124 | E | H | X | S+ | 0 | 0 | 128 | -4,-1.6 | 4,-2.7 | -5,-0.2 | -1,-0.3 | 0.917 | 108.1 | 47.4 | -55.0 | -51.1 | 12.5 | 2.1 | 6.5 |
| 125 | 125 | V | H | X | S+ | 0 | 0 | 40 | -4,-1.7 | 4,-3.3 | -3,-0.5 | 5,-0.2 | 0.906 | 111.9 | 50.1 | -62.9 | -41.3 | 10.2 | 4.8 | 8.0 |
| 126 | 126 | W | H | X | S+ | 0 | 0 | 7 | -4,-2.1 | 4,-2.4 | 2,-0.2 | 5,-0.3 | 0.978 | 113.1 | 46.2 | -59.4 | -54.0 | 7.2 | 2.4 | 8.1 |
| 127 | 127 | K | H | X | S+ | 0 | 0 | 143 | -4,-2.6 | 4,-2.5 | 1,-0.2 | -2,-0.2 | 0.808 | 116.1 | 46.1 | -53.7 | -44.3 | 9.3 | -0.3 | 9.9 |
| 128 | 128 | K | H | < | S+ | 0 | 0 | 185 | -4,-2.7 | -1,-0.2 | -5,-0.2 | -2,-0.2 | 0.943 | 117.0 | 42.6 | -64.6 | -52.0 | 10.7 | 2.3 | 12.3 |
| 129 | 129 | N | H | < | S+ | 0 | 0 | 82 | -4,-3.3 | -2,-0.2 | 1,-0.2 | -1,-0.2 | 0.719 | 122.6 | 40.8 | -62.3 | -32.2 | 7.2 | 3.9 | 13.0 |
| 130 | 130 | S | H | < | | 0 | 0 | 30 | -4,-2.4 | -2,-0.2 | -5,-0.2 | -1,-0.2 | 0.753 | 360.0 | 360.0 | -90.9 | -38.4 | 5.6 | 0.4 | 13.3 |
| 131 | 131 | G | | < | | 0 | 0 | 118 | -4,-2.5 | -3,-0.1 | -5,-0.3 | -4,-0.1 | 0.780 | 360.0 | 360.0 | 64.8 | 360.0 | 8.1 | -1.7 | 15.2 |

2) Nomenclature in the model

```
In 1LDM:          E1-H1-E2-H2-E3-3T-E4-H3-E5-H4
In Model:       H0-E1-H1-E2-------3T-E4-H3-E5-H4
```

| Number in model 1 | LEATHER | Number in LDM | Comment |
|---|---|---|---|
| 0 | 1 | | N-termini Met |
| 1 - 21 | 2 - 19 | 240 - 261 | H0 |
| 21 - 26 | 20 - 26 | | Loop |
| 27 - 57 | 27 - 57 | 22 - 72 | E1, H1, E2, |
| 58 - 59 | 58 - 59 | | Loop |
| 60 - 131 | 60 - 131 | 81 - 152 | 3T, E4, H3, E5, H4 |

SEQUENCE NUMBERING SCHEME ACCORDING TO LEATHER

```
1LDM A T L K D K L I G H L A T S Q E P R S Y N


     1       5        10        15        20        25        30
1LEA M G C G T S D A Q G A K V A D D A E T K M   N G G V N K I T V
Mod1   K G G T S D A Q G A K V A D D A E T K M K R Y P R N K I T V
1LDM * K G Y T S W A I G L S V A D L A E T I M  *         K I T V


    31      35        40        45        50      55
1LEA V G Y G A V G A A S A A A S I L A K D L A D Q V A Y V D
MOD1 V G R G A V G A A S A A A S I N A K D N A D Q V A Y V D
1LDM V G V G A V G M A C A I S I L M K D L A D E V A L V D

1LDM V M E D K L K G E M M D L Q H G S L F L H T A K I V S G K D Y S V S A G S


       60        65        70        75        80        85        90
1LEA G G S D Y S D S A G S K L V V I T A G A R Q S C G E S R E N L V Q
MOD1 G G K D Y S D S A G S K L V V I T A G A R Q Q E G E S R E N L V Q
1LDM     K D Y S V S A G S K L V V I T A G A R Q Q E G E S R L N L V Q


    91      95       100       105       110       115       120
1LEA R N V N I A K Y I A P N I Q K Q S P E V T I L I V S N P S D
MOD1 R N V N I F K Q I I P N I V K H S P D V T I R V V S N P V D
1LDM R N V N I F K F I I P N I V K H S P D C I I L V V S N P V D


   121     125       130
1LEA A L T E V W K K N S G
MOD1 A L T E V A K K L S G
1LDM V L T Y V A W K L S G

1LDM :
     152
     L P M H R I I G S G C N L D S A R F R Y L M G E R L G V
     H S C S C H G W V I G E H G D S V P S V W S G M N V A S I K
     L H P L D G T N K D K Q D W K K L H K D V V D S A Y E V I K L 241

        * A SEGMENT KEPT AND ALIGNED WITH RES 2 - 21 IN 1LEA *

     262
     K N L C R V H P V
     S T M V K D F Y G I K D N V F L S L P C V L N D H G I S N I
     V K M K L K P N E E Q Q L Q K S A T T L W D I Q K D L K F  329
```

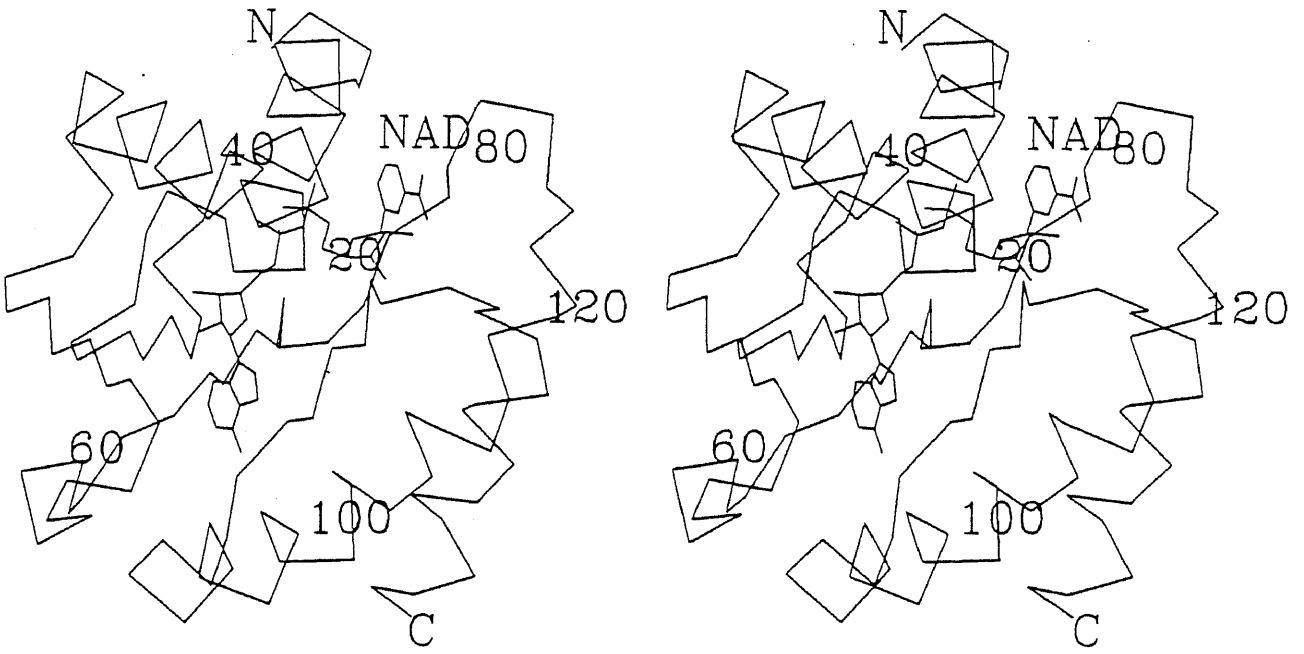



Fig 4. Alpha-carbon trace of LEATHER, with NAD bound. Numbering according to the final model.

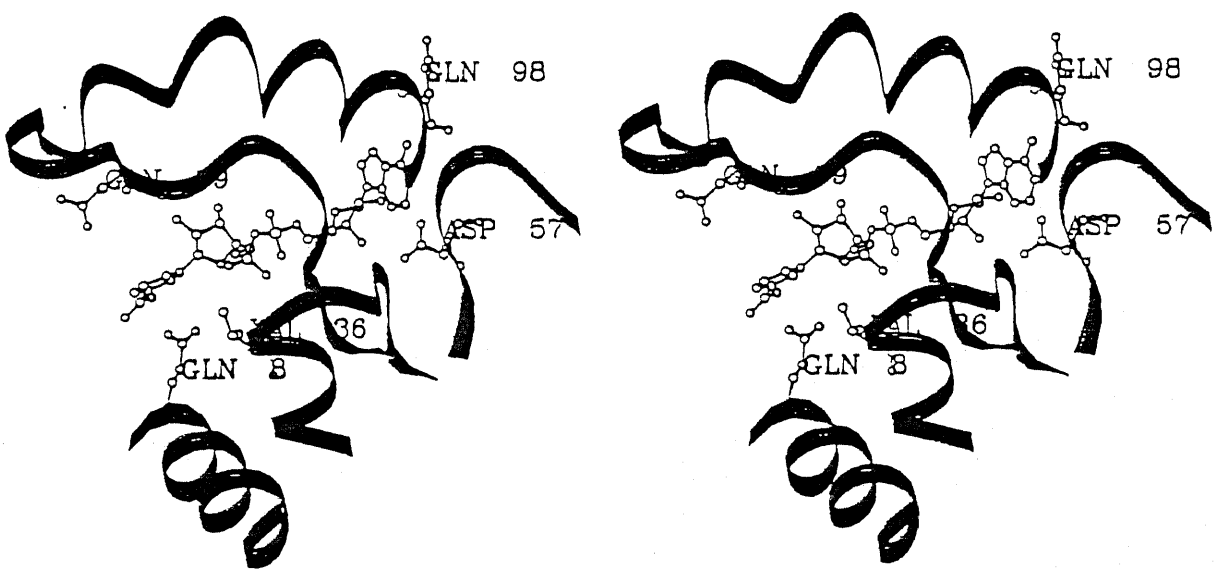



Fig 5. Schematic outline of the NAD binding site of LEATHER.

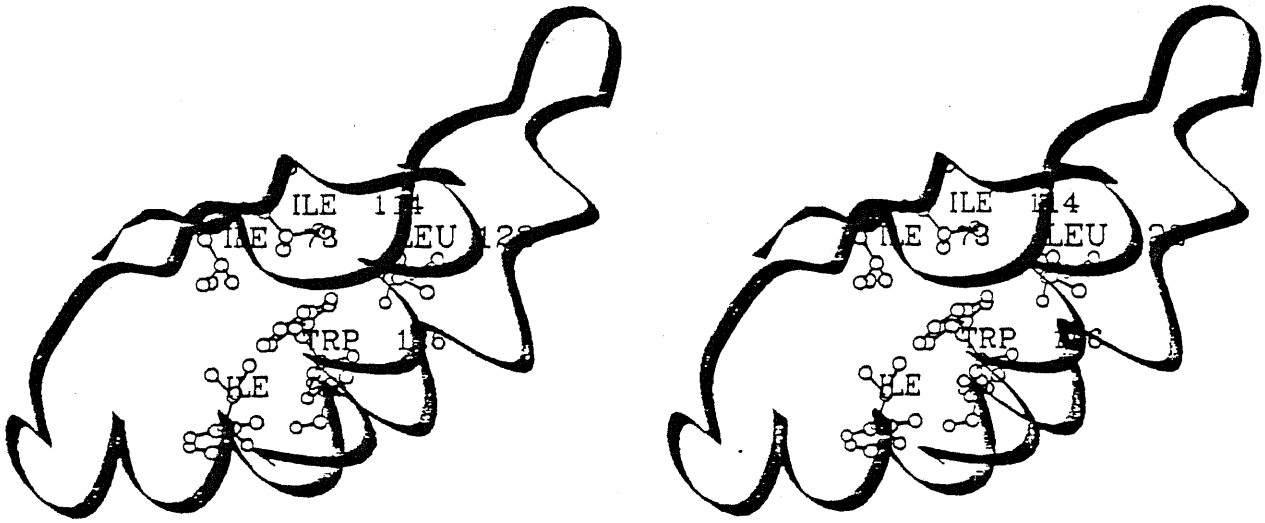



Fig 6. Stereo drawing showing the environment of the unique Trp residue (126).

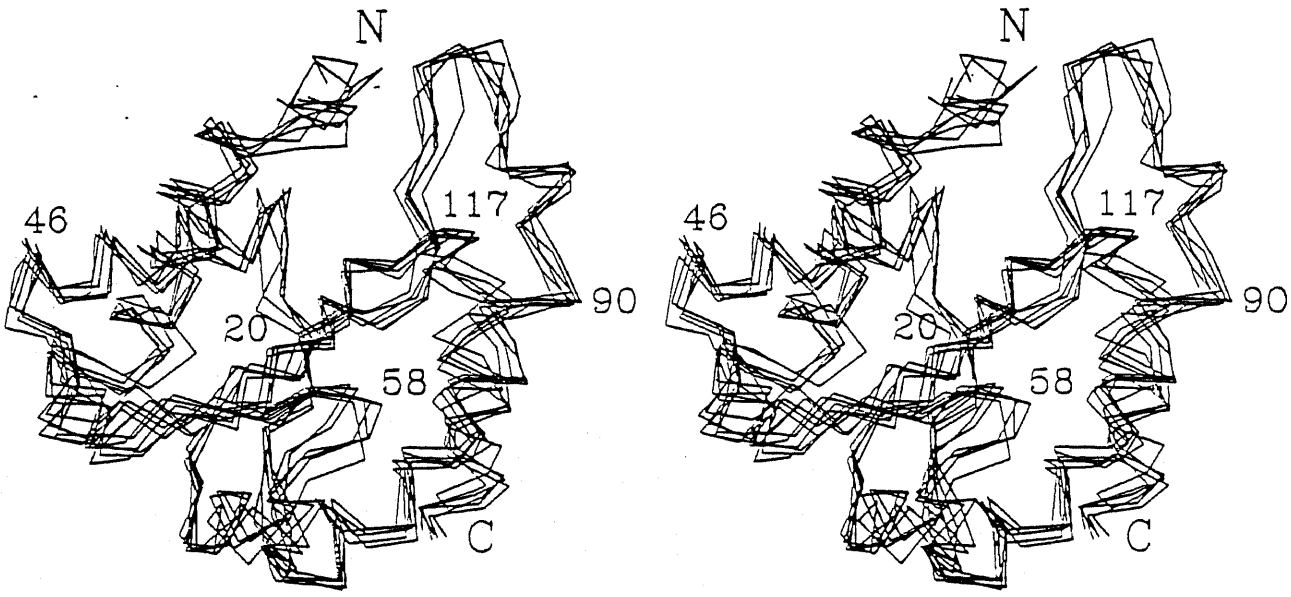



Fig 7. Molecular dynamics simulation of the final model. The starting model and the model after 3, 5, 7, 9, and 11 ps simulation are superimposed. The MD simulation were performed in vacuo using the standard GROMOS force field, a temperature of 300 K and the "SHAKE" option.

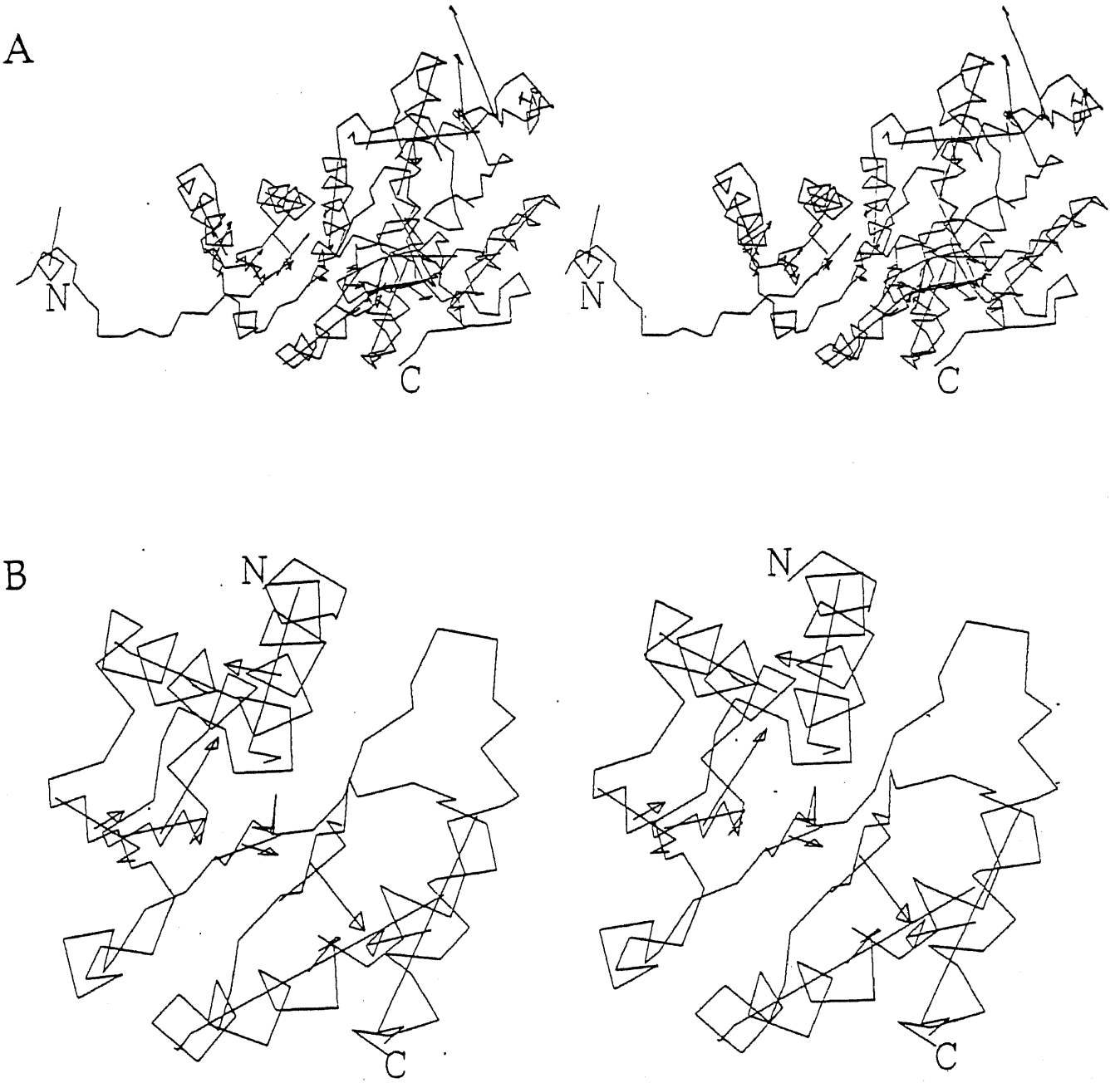



Fig 8. Stereo drawing showing the hydrophobic moments of secondary structure elements. The arrows represent vectors whose length indicates the magnitude of the hydrophobic moment. (defined according to Eisenberg).

A. Lactate dehydrogenase. Note that the arrows from seconday structures involved in the subunit interface point outwards.

B. LEATHER. The arrows point inwards, reflecting the hydrophobic packing of the model.

# AIDA : AMERICAN - ITALIAN DESIGNER ANTIBODY

Jay Banks
Department of Biomedical Engineering
Boston University
44 Cummington Street
Boston, MA 02215
U. S. A.
e-mail: JAY@BUENGA.BITNET or JAY@BUENGA.BU.EDU (Internet)

Roberto Jappelli
Dipartimento di Biologia
Universita' di Roma Tor Vergata
Via E. Carnevale
00173 ROMA, ITALY
e-mail: JNET%"MACINO@IPIVAXIN"

Arthur M. Lesk
Department of Hematology
Cambridge University
Cambridge CB2 2QH, UK
e-mail: aml2@mrc-lmb.cam.ac.uk

Anna Tramontano
I. R. B. M.
Via Pontina, Km 30.6
00040 Pomezia, ROMA, ITALY
e-mail: tramontano@EMBL

## 5.1 BACKGROUND

Immunoglobulins are multimeric, multidomain proteins, with domains of similar structure (Fig. 1). Each domain is formed by two beta-sheets packed face to face, pinned together by a disulphide bridge (see Fig. 2). Two of the domains - the variable domains of a light and a heavy chain (VL and VH) - are packed together to create a scaffolding of conserved structure, on which the antigen-binding site is formed by six hypervariable loops. The structural definition of the hypervariable loops [Chothia and Lesk, 1987] is similar but not identical to Kabat's sequence-based definition of complementarity determining regions (CDRs) [Kabat et al., 1987]. Three of the hypervariable loops are from the VL domain (L1, L2, and L3), and three from the VH domain (H1, H2, and H3) (Fig. 1).

At least five of the six hypervariable antigen binding loops can assume only a limited number of main chain conformations called canonical structures [Chothia and Lesk, 1987; Chothia et al., 1989; Tramontano et al., 1990]. These loop conformations are determined by particular residues, through their packing, hydrogen bonding, or ability to assume unusual mainchain conformations (see Table 1). Canonical structures and conformation-determining residues for the sixth loop, H3, have not been determined at present.

A procedure for predicting the structure of antigen-binding sites of immunoglobulins from their sequences, based on this understanding of the determinants of the canonical conformations, has recently been shown to be successful in "blind" tests [Chothia et al., 1986; Chothia et al., 1989]. In this project, we elected to test our understanding of the sequence-structure relationship in immunoglobulins in a different way: we decided to take a known antigen-antibody interface and redesign it to bind another protein.

The project is interesting for theoretical reasons, in that it allows us to investigate whether it is possible at least in principle to design an interface complementary to a putative antigen. It is also of obvious practical interest because, if successful, it opens the possibility of designing an antibody against a specific selected epitope.

For several reasons, we are more confident in the consequences of our substitutions than we would be in a more general interface design. For antibodies, we think we know which residues can be changed without affecting the general fold of the antibody and the main-chain conformation of the hypervariable loops. Additional restrictions on possible changes result from insisting that the main chains of both antibody and antigen be kept fixed, that newly introduced or rotated side chains remain in reasonable conformations [Ponder and Richards, 1987], and that the conformation of the free and bound antigen be similar, as is the case in known natural complexes [Davies et al., 1990].

## 5.2 PROCEDURE

The first step in this project was to select an antibody-antigen complex of known structure to serve as the starting point for our design. Examples of such complexes in the Brookhaven Protein Data

Bank include HyHEL-5 (2HFL) and HyHEL-10 (3HFM), with the antigen in both cases being hen egg-white lysozyme. We selected HyHEL-5 because its structure has been determined to better resolution (2.54 A).
The interface between HyHEL-5 and lysozyme has been studied in detail [Sheriff et al., 1987; Davies et al. 1990]. The surface area in contact between the two molecules is about 750 A**2. Both molecules contain residues that are buried by the formation of the interface, but that do not make direct contacts. In the antigen, there are 24 such buried residues, plus 14 that make direct contact, within an epitope formed by three separated regions. In the antibody, there are 28 buried residues plus 17 in contact. These residues are distributed among all six hypervariable loops. In addition, residue Trp 47 of the heavy chain, although not in the hypervariable loops, is in contact with the antigen. The interface includes 74 van der Waals contacts, 10 hydrogen bonds, and 3 salt bridges. There are two water molecules in the interface [Davies et al., 1990], but they are not present in the PDB file as originally deposited. Superposition of the bound and unbound lysozyme shows an r.m.s. displacement of 0.48 A for the backbone atoms and 1.22 A for the side chains.
We used several rather broad criteria in determining an antigen to use for our designed complex. The molecule should be monomeric so that aggregation of subunits will not compete with binding to the antibody. It should be soluble, and otherwise readily available and easily handled for experimental purposes. And it should have a well-determined, high-resolution structure, to provide a solid basis for design of the antibody-antigen interface.
We surveyed a number of molecules as potential antigens. These included ROP, Ubiquitin (1UBQ), crambin (1CRN), ferredoxin (4FD1), flavodoxin (3FXN), interleukin-1-b (2I1B), and cytochrome c (5CYT). We attempted to dock each protein visually to the antibody structure, using the INSIGHT molecular graphics package. For this purpose, we replaced the side chains of the antibody with alanines, and put Connolly surfaces on all the molecules, in order to visualize the maximal "pocket" into which an antigen might fit. We also left the native lysozyme structure on the screen, and attempted to match prospective antigens to its surface. To get a sensitive picture of the surface complementarity, we displayed a thin slab of the Connolly surfaces on the screen, and moved the position of the slab through the interface.
One of the molecules we inspected, ferredoxin, matched well, but on examination of the epitope we found that it would be necessary to bury an INTRAMOLECULAR salt bridge in the interface. Although this is not impossible, we wished to pick a subject that had as few obvious handicaps as possible. Among other molecules, we observed the best complementarity with flavodoxin and cytochrome c. On closer inspection, using the native antibody, we observed that the cytochrome interface had more steric clashes than flavodoxin. In addition, the chosen docking configuration for flavodoxin had a lysine residue (Lys 76) projecting into the groove between the heavy and light chains of the antibody, thus increasing the complementary surface area, and potentially providing a stabilizing salt bridge. For these reasons, we decided to proceed with flavodoxin.
The bulk of the interface design consisted of iteration of three steps. First, we inspected the current interface on the INSIGHT graphics screen. Second, we decided which residues in the antibody

should be changed (mutated and/or rotated) in order to improve the interface. Third, we performed a short energy refinement (usually 50 or 100 steps) on the side chains of the resulting structure.
We attempted to be conservative in our design. For one thing, we did not make substitutions in the antigen. In addition, we did not insert or delete residues in the antibody, and we did not make changes in the main chain angles of either molecule. When performing mutations, we were careful not to change residues in the antibody that are believed to be critical to the main-chain conformation of the hypervariable loops [Chothia and Lesk, 1987; Chothia et al., 1989; Tramontano et al., 1990]. For one hypervariable loop, H3, it is not known which residues are critical, and we had hoped to change as little as possible in this loop. Unfortunately, H3 is significantly involved in the interface, and two changes were unavoidable. We did not, however, change a glycine in H3 (Gly H99) that in the native complex is in a "++" conformation (both backbone dihedral angles positive) unlikely to be assumed by another residue. When we changed side-chain dihedral angles in the antibody, we chose values known to occur in the database for a given residue, in most cases using the set with the highest or second-highest frequency [Ponder and Richards, 1987]. Finally, we held the backbones of both molecules fixed during energy refinement.
Early in the process, we observed a cavity in the interface, which we had difficulty filling by mutating and rotating residues. A similar cavity is present in the native HyHEL-5 - lysozyme interface. We believe this cavity is filled by a water molecule mentioned by the authors who determined this structure [Davies et al., 1990]. We therefore decided to place a water molecule in our cavity, in such a position as to have potential hydrogen bonds to both antibody and antigen. Energy refinement, and calculations using potential energy grids in the interface, resulted in slight movement of the water.
After seven rounds of mutation and refinement, we arrived at a structure for AIDA (lower left color plate). The most striking features of its interface with flavodoxin are the water molecule (lower right color plate), and the salt bridge between Lys 76 of the antigen and Asp H35 of the antibody. A total of 18 residues have been changed from the original HyHEL-5, as shown in Table 2.

## 5.3 EVALUATION

Visual examination of the interface reveals good surface complementarity (see upper left color plate) with the exception of the aforementioned cavity, which is not completely filled by the combination of water and changed side chains. We used the program QPACK to evaluate the packing density of the interface, and found no major deviations from the average packing found in natural proteins for each residue. We obtained the approximate surface area of the interface by calculating the accessible surface area of each molecule in the presence and absence of the other, and subtracting. The result is about 650 A**2, which is comparable to known native antibody-antigen interfaces. We determined the contacts between the antigen and antibody using the programs PINQ and CONAN. The contacts include one salt bridge, five hydrogen bonds, and 61 van der Waals contacts (Tables 3 and 4, and Fig. 3). All six hypervariable loops

are involved in the contacts. By design, the Lys 76 - Asp H35 salt bridge is well shielded from competing polar interactions, with a suitable environment for the hydrophobic portion of the lysine side chain.

In its current position, the water molecule makes two hydrogen bonds, one each with the antibody and antigen (see Table 3 and lower right color plate). It is also fairly close to a hydrophobic side chain, Leu 44, in the flavodoxin molecule. We have used the program WHATIF to find out what positions are acceptable for water in the vicinity of leucine, by surveying positions where it is found in natural proteins of well-determined structure. As shown in Fig. 4, our water molecule is in an acceptable position.

In general, our interface is not as good a "fit" as that in the native HyHEL-5 - lysozyme complex. We believe that some of this deficiency can be remedied in a real AIDA - flavodoxin complex, as a result of motion of the flavodoxin side chains. The energy refinement we have performed has resulted in much less side-chain displacement than the 1.2 A r.m.s. that is observed between bound and unbound lysozyme. From analysis of our interface, we are confident that displacements of the latter order of magnitude can lead to improvements such as more hydrogen bonds, a better environment for the water molecule, and increased filling of the cavity.

TABLE 1: CATALOG OF CANONICAL STRUCTURES

The asterisks indicate the residues responsible for the observed conformation:

L1

| Kabat numbering: | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 2 | 25 | 33 | 71 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AIDA numbering : | L26 | L27 | L28 | L29 | L30 | | L31 | L2 | L25 | L32 | L70 |
| | | | | * | | | | * | * | * | * |
| HyHEL-5 : | S | S | S | V | N | - | Y | I | A | M | Y |
| AIDA : | S | S | S | V | S | - | Q | I | A | M | Y |

L2

| Kabat numbering: | 50 | 51 | 52 | 48 | 64 |
|---|---|---|---|---|---|
| AIDA numbering : | L49 | L50 | L51 | L47 | L63 |
| | | | | * | * |
| HyHEL-5 : | D | T | S | I | G |
| AIDA : | N | T | S | I | G |

L3

| Kabat numbering: | 91 | 92 | 93 | 94 | 95 | 96 | 90 |
|---|---|---|---|---|---|---|---|
| AIDA numbering : | L90 | L91 | L92 | L93 | L94 | | L89 |
| | | | | | * | | * |
| HyHEL-5 : | W | G | R | N | P | - | Q |
| AIDA : | W | S | S | N | P | - | Q |

H1

| Kabat numbering: | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 34 |
|---|---|---|---|---|---|---|---|---|
| AIDA numbering : | H26 | H27 | H28 | H29 | H30 | H31 | H32 | H34 |
| | * | * | | * | | | | * |
| HyHEL-5 : | G | Y | T | F | S | D | Y | I |
| AIDA : | G | Y | T | F | S | N | S | I |

H2

| Kabat numbering: | 52A | 53 | 54 | 55 | 71 |
|---|---|---|---|---|---|
| AIDA numbering : | H53 | H54 | H55 | H56 | H72 |
| | | | | | * |
| HyHEL-5 : | P | G | S | G | A |
| AIDA : | P | G | S | G | A |

TABLE 2
---

Sequence of the variable domains of AIDA (top line). The differences with the HyHEL-5 sequence are indicated (bottom line) (Variable domains only)

219 IDENTICAL RESIDUES OUT OF 238 TOTAL RESIDUES
92.02 PERCENT IDENTICAL

| L1 | L2 | L3 | L4 | L5 | L6 | L7 | L8 | L9 | L10 | L11 | L12 | L13 | L14 | L15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ASP | ILE | VAL | LEU | THR | GLN | SER | PRO | ALA | ILE | MET | SER | ALA | SER | PRO |
| | | | | | | | | | | | | | | |
| L16 | L17 | L18 | L19 | L20 | L21 | L22 | L23 | L24 | L25 | L26 | L27 | L28 | L29 | L30 |
| GLY | GLU | LYS | VAL | THR | MET | THR | CYS | SER | ALA | SER | SER | SER | VAL | SER |
| | | | | | | | | | | | | | | (ASN) |
| L31 | L32 | L33 | L34 | L35 | L36 | L37 | L38 | L39 | L40 | L41 | L42 | L43 | L44 | L45 |
| GLN | MET | ILE | TRP | PHE | GLN | GLN | LYS | SER | GLY | THR | SER | PRO | LYS | ARG |
| (TYR) | | (TYR) | | (TYR) | | | | | | | | | | |
| L46 | L47 | L48 | L49 | L50 | L51 | L52 | L53 | L54 | L55 | L56 | L57 | L58 | L59 | L60 |
| TRP | ILE | GLN | ASN | THR | SER | LYS | LEU | ALA | SER | GLY | VAL | PRO | VAL | ARG |
| | | (TYR) | (ASP) | | | | | | | | | | | |
| L61 | L62 | L63 | L64 | L65 | L66 | L67 | L68 | L69 | L70 | L71 | L72 | L73 | L74 | L75 |
| PHE | SER | GLY | SER | GLY | SER | GLY | THR | SER | TYR | SER | LEU | THR | ILE | SER |
| | | | | | | | | | | | | | | |
| L76 | L77 | L78 | L79 | L80 | L81 | L82 | L83 | L84 | L85 | L86 | L87 | L88 | L89 | L90 |
| SER | MET | GLU | THR | GLU | ASP | ALA | ALA | GLU | TYR | TYR | CYS | GLN | GLN | MET |
| | | | | | | | | | | | | | | (TRP) |
| L91 | L92 | L93 | L94 | L95 | L96 | L97 | L98 | L99 | L100 | L101 | L102 | L103 | L104 | L105 |
| SER | SER | ASN | PRO | THR | PHE | GLY | GLY | GLY | THR | LYS | LEU | GLU | ILE | LYS |
| (GLY) | (ARG ) | | | | | | | | | | | | | |
| L106 | L107 | L108 | L109 | L110 | L111 | L112 | L113 | L114 | L115 | L116 | L117 | L118 | L119 | |
| ARG | ALA | ASP | ALA | ALA | PRO | THR | VAL | SER | ILE | PHE | PRO | PRO | SER | |
| | | | | | | | | | | | | | | |
| H1 | H2 | H3 | H4 | H5 | H6 | H7 | H8 | H9 | H10 | H11 | H12 | H13 | H14 | H15 | H16 |
| GLN | VAL | GLN | LEU | GLN | GLN | SER | GLY | ALA | GLU | LEU | MET | LYS | PRO | GLY | ALA |
| (PCA) | | | | | | | | | | | | | | | |
| H17 | H18 | H19 | H20 | H21 | H22 | H23 | H24 | H25 | H26 | H27 | H28 | H29 | H30 | H311 | |
| SER | VAL | LYS | ILE | SER | CYS | LYS | ALA | SER | GLY | TYR | THR | PHE | SER | ASN | |
| | | | | | | | | | | | | | | | |
| H32 | H33 | H34 | H35 | H36 | H37 | H38 | H39 | H40 | H41 | H42 | H43 | H44 | H45 | H46 | |
| SER | HIS | ILE | ASP | TRP | VAL | LYS | GLN | ARG | PRO | GLY | HIS | GLY | LEU | GLU | |
| (TYR) | (TRP) | | (GLU) | | | | | | | | | | | | |
| H47 | H48 | H49 | H50 | H51 | H52 | H53 | H54 | H55 | H56 | H57 | H58 | H59 | H60 | H61 | |
| TRP | ILE | GLY | VAL | ILE | ASN | PRO | GLY | SER | GLY | SER | THR | ARG | TYR | HIS | |
| | | | (GLU) | | (LEU) | | | | | | | (ASN) | | | |

| H62 | H63 | H64 | H65 | H66 | H67 | H68 | H69 | H70 | H71 | H72 | H73 | H74 | H75 | H76 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GLU | ARG | PHE | LYS | GLY | LYS | ALA | THR | PHE | THR | ALA | ASP | THR | SER | SER |

| H77 | H78 | H79 | H80 | H81 | H82 | H83 | H84 | H85 | H86 | H87 | H88 | H89 | H90 | H91 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SER | THR | ALA | TYR | MET | GLN | LEU | ASN | SER | LEU | THR | SER | GLU | ASP | SER |

| H92 | H93 | H94 | H95 | H96 | H97 | H98 | H99 | H100 | H101 | H102 | H103 | H104 | H105 | H106 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GLY | VAL | TYR | TYR | CYS | LEU | HIS | GLY | HIS (ASN) | TYR | GLN (ASP) | PHE | ASP | GLY | TRP |

| H107 | H108 | H109 | H110 | H111 | H112 | H113 | H114 | H115 | H116 | H117 | H118 | H119 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GLY | GLN | GLY | THR | THR | LEU | THR | VAL | SER | SER | ALA | LYS | THR |

TABLE 3: HYDROGEN BONDS IN DESIGNED INTERFACE

3A: HYDROGEN BONDS AIDA VL - FLAVODOXIN

| | | DISTANCE | ANGLE |
|---|---|---|---|
| L91SER OG ... OD2 | Y41ASP | 3.33 | 112.8 |

3B: HYDROGEN BONDS AIDA VH - FLAVODOXIN

| | | DISTANCE | ANGLE |
|---|---|---|---|
| H31ASN ND2 ... O | Y104ASN | 3.10 | 130.7 |
| H32SER OG ... OG | Y78SER | 3.31 | 144.7 |
| H52ASN ND2 ... O | Y106TYR | 3.11 | 113.1 |
| H59ARG NH1 ... O | Y72GLU | 3.45 | 131.2 |

3C: HYDROGEN BONDS WATER - AIDA

| | | DISTANCE | ANGLE |
|---|---|---|---|
| L31GLN NE2 ... OH | 1WAT | 3.30 | |

3D: HYDROGEN BONDS WATER - FLAVODOXIN

| | | DISTANCE | ANGLE |
|---|---|---|---|
| Y45ASN OD1 ... OH | 1WAT | 2.81 | |

TABLE 4: CONTACTS IN DESIGNED INTERFACE

(Distance lower than sum of van der Waals radii + 0.60 A):

4A: CONTACTS BETWEEN AIDA VL and FLAVODOXIN:

| | | | | |
|---|---|---|---|---|
| L31GLN | OE1 | ... | OD1 | Y45ASN | 3.089 |
| L49ASN | ND2 | ... | OD1 | Y45ASN | 3.094 |
| L91SER | OG | ... | OD2 | Y41ASP | 3.326 |
| L90MET | CE | ... | NZ | Y76LYS | 3.414 |
| L49ASN | ND2 | ... | CG | Y45ASN | 3.446 |
| L31GLN | OE1 | ... | ND2 | Y45ASN | 3.487 |
| L31GLN | CD | ... | ND2 | Y45ASN | 3.635 |
| L90MET | CE | ... | CE | Y76LYS | 3.715 |
| L31GLN | OE1 | ... | CG | Y45ASN | 3.730 |
| L31GLN | NE2 | ... | ND2 | Y45ASN | 3.731 |
| L30SER | CB | ... | CG | Y42GLU | 3.820 |
| L90MET | SD | ... | NZ | Y76LYS | 3.843 |
| L49ASN | ND2 | ... | ND2 | Y45ASN | 3.853 |
| L90MET | CE | ... | CD | Y76LYS | 4.025 |
| L30SER | CB | ... | CD | Y42GLU | 4.153 |

4B: CONTACTS BETWEEN AIDA VH and FLAVODOXIN:

| | | | | |
|---|---|---|---|---|
| H31ASN | O | ... | OG | Y78SER | 2.672 |
| H100HIS | NE2 | ... | OG | Y78SER | 2.984 |
| H31ASN | ND2 | ... | O | Y104ASN | 3.101 |
| H52ASN | ND2 | ... | O | Y106TYR | 3.106 |
| H33HIS | NE2 | ... | OG1 | Y75THR | 3.120 |
| H100HIS | CE1 | ... | O | Y78SER | 3.125 |
| H35ASP | OD1 | ... | CE | Y76LYS | 3.137 |
| H33HIS | NE2 | ... | CB | Y75THR | 3.153 |
| H35ASP | OD1 | ... | NZ | Y76LYS | 3.237 |
| H32SER | OG | ... | OG | Y78SER | 3.306 |
| H31ASN | O | ... | CA | Y107GLY | 3.342 |
| H32SER | CA | ... | OG | Y78SER | 3.344 |
| H33HIS | NE2 | ... | CA | Y75THR | 3.405 |
| H59ARG | NH1 | ... | O | Y72GLU | 3.451 |
| H100HIS | NE2 | ... | CB | Y78SER | 3.483 |
| H31ASN | ND2 | ... | O | Y105GLY | 3.508 |
| H33HIS | CD2 | ... | CG | Y76LYS | 3.524 |
| H100HIS | NE2 | ... | O | Y78SER | 3.549 |
| H35ASP | OD2 | ... | CE | Y76LYS | 3.551 |
| H31ASN | O | ... | CB | Y78SER | 3.576 |
| H31ASN | CG | ... | O | Y104ASN | 3.585 |
| H52ASN | ND2 | ... | O | Y74SER | 3.587 |
| H100HIS | CE1 | ... | C | Y78SER | 3.590 |
| H31ASN | CB | ... | CA | Y107GLY | 3.610 |
| H31ASN | ND2 | ... | CA | Y105GLY | 3.621 |
| H33HIS | CE1 | ... | CA | Y75THR | 3.633 |
| H33HIS | CD2 | ... | N | Y76LYS | 3.638 |
| H99GLY | CA | ... | NZ | Y76LYS | 3.654 |
| H31ASN | C | ... | OG | Y78SER | 3.662 |
| H33HIS | CE1 | ... | OG1 | Y75THR | 3.688 |

| | | | | | |
|---|---|---|---|---|---|
| H31ASN | CB | ... | N | Y107GLY | 3.695 |
| H31ASN | ND2 | ... | C | Y105GLY | 3.706 |
| H33HIS | CG | ... | N | Y76LYS | 3.706 |
| H33HIS | CD2 | ... | CB | Y75THR | 3.709 |
| H35ASP | CG | ... | CE | Y76LYS | 3.724 |
| H52ASN | ND2 | ... | CA | Y106TYR | 3.724 |
| H33HIS | CD2 | ... | CA | Y75THR | 3.736 |
| H100HIS | NE2 | ... | C | Y78SER | 3.737 |
| H32SER | CB | ... | OG | Y78SER | 3.757 |
| H52ASN | ND2 | ... | C | Y106TYR | 3.766 |
| H99GLY | O | ... | CD | Y76LYS | 3.809 |
| H59ARG | NH1 | ... | CG2 | Y75THR | 3.814 |
| H33HIS | CE1 | ... | CB | Y75THR | 3.857 |
| H59ARG | CZ | ... | CG2 | Y75THR | 3.889 |
| H100HIS | CE1 | ... | N | Y79GLY | 3.910 |
| H32SER | CA | ... | CB | Y78SER | 3.941 |
| H33HIS | CD2 | ... | C | Y75THR | 3.966 |
| H33HIS | CG | ... | CG | Y76LYS | 4.013 |
| H33HIS | CB | ... | CG | Y76LYS | 4.044 |
| H100HIS | CE1 | ... | CA | Y79GLY | 4.062 |
| H31ASN | C | ... | CA | Y107GLY | 4.107 |
| H99GLY | CA | ... | CE | Y76LYS | 4.150 |
| H31ASN | CA | ... | CA | Y107GLY | 4.225 |

4C: CONTACTS BETWEEN WATER and the AIDA complex:

| | | | | | |
|---|---|---|---|---|---|
| 1WAT | OH | ... | OD1 | Y45ASN | 2.813 |
| 1WAT | OH | ... | O | Y44LEU | 3.179 |
| 1WAT | OH | ... | NE2 | L31GLN | 3.301 |
| 1WAT | OH | ... | CG | Y45ASN | 3.354 |
| 1WAT | OH | ... | C | Y44LEU | 3.444 |
| 1WAT | OH | ... | ND2 | Y45ASN | 3.638 |
| 1WAT | OH | ... | CB | Y44LEU | 3.669 |

TABLE 5: HYDROGEN BONDS IN NATIVE INTERFACE

5A: HYDROGEN BONDS HyHEL-5 VL - LYSOZYME

| | | DISTANCE | ANGLE |
|---|---|---|---|
| L91GLY O ... NE | Y45ARG | 3.08 | 174.7 |
| L92ARG NH1 ... O | Y45ARG | 3.05 | 142.8 |
| L92ARG NE ... O | Y45ARG | 2.88 | 141.5 |
| L92ARG NH2 ... OD1 | Y46ASN | 3.29 | 113.0 |
| L92ARG NH1 ... O | Y45ARG | 3.05 | 142.8 |
| L92ARG NE ... O | Y45ARG | 2.88 | 141.5 |
| L92ARG NH2 ... OD1 | Y46ASN | 3.29 | 113.0 |

5B: HYDROGEN BONDS HyHEL-5 VH - LYSOZYME

| | | DISTANCE | ANGLE |
|---|---|---|---|
| H33TRP NE1 ... OH | Y53TYR | 2.94 | 139.9 |
| H59ASN ND2 ... OG1 | Y43THR | 3.01 | 136.0 |
| H59ASN OD1 ... OG1 | Y43THR | 2.10 | 159.4 |

TABLE 6: CONTACTS IN NATIVE INTERFACE

(Distance lower than sum of van der Waals radii + 0.60 A):

6A: CONTACTS BETWEEN HyHEL-5 VL and LYSOZYME

| | | | | |
|---|---|---|---|---|
| L90TRP | CH2 | ... | NH1 | Y68ARG | 2.612 |
| L92ARG | CD | ... | O | Y45ARG | 2.718 |
| L90TRP | CZ3 | ... | NH2 | Y45ARG | 2.838 |
| L90TRP | CE3 | ... | NH2 | Y45ARG | 2.867 |
| L92ARG | NE | ... | O | Y45ARG | 2.876 |
| L92ARG | NH1 | ... | O | Y45ARG | 3.046 |
| L90TRP | CH2 | ... | NH2 | Y45ARG | 3.052 |
| L92ARG | CZ | ... | O | Y45ARG | 3.053 |
| L90TRP | CD2 | ... | NH2 | Y45ARG | 3.072 |
| L91GLY | O | ... | NE | Y45ARG | 3.077 |
| L91GLY | O | ... | CG | Y45ARG | 3.168 |
| L30ASN | OD1 | ... | CB | Y48ASP | 3.182 |
| L90TRP | CE2 | ... | NH2 | Y45ARG | 3.235 |
| L90TRP | CZ3 | ... | NH1 | Y68ARG | 3.244 |
| L90TRP | CZ2 | ... | NH2 | Y45ARG | 3.254 |
| L49ASP | OD1 | ... | CB | Y70PRO | 3.274 |
| L94PRO | CG | ... | NE | Y45ARG | 3.281 |
| L92ARG | NH2 | ... | OD1 | Y46ASN | 3.292 |
| L91GLY | O | ... | CD | Y45ARG | 3.426 |
| L30ASN | OD1 | ... | CA | Y48ASP | 3.502 |
| L91GLY | C | ... | O | Y46ASN | 3.522 |
| L90TRP | CH2 | ... | CZ | Y68ARG | 3.541 |
| L90TRP | CZ2 | ... | NH1 | Y68ARG | 3.608 |
| L90TRP | CZ3 | ... | O | Y49GLY | 3.697 |
| L92ARG | CG | ... | O | Y45ARG | 3.718 |
| L92ARG | CG | ... | CG2 | Y47THR | 3.723 |
| L90TRP | CZ3 | ... | CZ | Y68ARG | 3.730 |
| L92ARG | O | ... | CD | Y45ARG | 3.768 |
| L90TRP | CH2 | ... | CD | Y68ARG | 3.808 |
| L92ARG | CA | ... | O | Y45ARG | 3.839 |
| L90TRP | CZ3 | ... | CZ | Y45ARG | 3.842 |
| L91GLY | CA | ... | O | Y46ASN | 3.843 |
| L92ARG | NE | ... | CA | Y46ASN | 3.845 |
| L94PRO | CD | ... | NE | Y45ARG | 3.863 |
| L92ARG | CD | ... | C | Y45ARG | 3.882 |
| L31TYR | CZ | ... | CB | Y70PRO | 3.932 |
| L90TRP | CE3 | ... | CZ | Y65ARG | 3.974 |
| L31TYR | CE2 | ... | CB | Y70PRO | 3.982 |
| L92ARG | CZ | ... | C | Y45ARG | 4.054 |
| L90TRP | CE3 | ... | CA | Y49GLY | 4.062 |
| L31TYR | CE2 | ... | CG | Y70PRO | 4.084 |
| L91GLY | C | ... | CG | Y45ARG | 4.089 |
| L91GLY | CA | ... | CA | Y47THR | 4.096 |
| L94PRO | CG | ... | CZ | Y45ARG | 4.103 |
| L90TRP | CH2 | ... | CZ | Y45ARG | 4.118 |
| L49ASP | CG | ... | CB | Y70PRO | 4.132 |
| L92ARG | CA | ... | CG | Y45ARG | 4.190 |
| L90TRP | CZ3 | ... | CD | Y68ARG | 4.206 |

| | | | | |
|---|---|---|---|---|
| L31TYR | CD2 | ... | CG | Y70PRO | 4.219 |
| L94PRO | CG | ... | CD | Y45ARG | 4.238 |
| L94PRO | CD | ... | CD | Y45ARG | 4.335 |

6B: CONTACTS BETWEEN HyHEL-5 VH and LYSOZYME

| | | | | |
|---|---|---|---|---|
| H59ASN | OD1 | ... | OG1 | Y43THR | 2.100 |
| H50GLU | OE1 | ... | NH1 | Y45ARG | 2.568 |
| H101TYR | OH | ... | CA | Y70PRO | 2.679 |
| H57SER | CB | ... | O | Y41GLN | 2.882 |
| H59ASN | CG | ... | OG1 | Y43THR | 2.919 |
| H57SER | CB | ... | CG2 | Y43THR | 2.933 |
| H33TRP | NE1 | ... | OH | Y53TYR | 2.937 |
| H59ASN | ND2 | ... | OG1 | Y43THR | 3.010 |
| H35GLU | OE1 | ... | NH1 | Y68ARG | 3.064 |
| H99GLY | O | ... | CB | Y68ARG | 3.089 |
| H101TYR | OH | ... | N | Y70PRO | 3.165 |
| H101TYR | OH | ... | O | Y70PRO | 3.179 |
| H50GLU | OE1 | ... | NH2 | Y68ARG | 3.223 |
| H57SER | OG | ... | CG2 | Y43THR | 3.256 |
| H101TYR | CD1 | ... | O | Y67GLY | 3.300 |
| H33TRP | CD1 | ... | NE | Y68ARG | 3.359 |
| H50GLU | OE1 | ... | CZ | Y45ARG | 3.371 |
| H33TRP | NE1 | ... | NH2 | Y68ARG | 3.374 |
| H101TYR | OH | ... | C | Y70PRO | 3.391 |
| H59ASN | ND2 | ... | O | Y43THR | 3.392 |
| H99GLY | O | ... | CD | Y68ARG | 3.427 |
| H33TRP | CD1 | ... | CZ | Y68ARG | 3.475 |
| H55SER | C | ... | OE1 | Y41GLN | 3.481 |
| H33TRP | CZ2 | ... | OH | Y53TYR | 3.485 |
| H59ASN | OD1 | ... | CB | Y43THR | 3.505 |
| H57SER | C | ... | CG2 | Y43THR | 3.520 |
| H33TRP | NE1 | ... | NE | Y68ARG | 3.532 |
| H101TYR | CE1 | ... | O | Y69THR | 3.533 |
| H33TRP | CE2 | ... | OH | Y53TYR | 3.535 |
| H59ASN | ND2 | ... | C | Y43THR | 3.561 |
| H101TYR | OH | ... | C | Y69THR | 3.572 |
| H33TRP | NE1 | ... | CZ | Y68ARG | 3.580 |
| H50GLU | CD | ... | NH1 | Y45ARG | 3.609 |
| H57SER | CA | ... | CG2 | Y43THR | 3.623 |
| H101TYR | N | ... | O | Y67GLY | 3.628 |
| H101TYR | CE1 | ... | C | Y68ARG | 3.638 |
| H101TYR | CD1 | ... | C | Y67GLY | 3.641 |
| H56GLY | N | ... | OE1 | Y41GLN | 3.646 |
| H101TYR | CB | ... | O | Y67GLY | 3.646 |
| H58THR | O | ... | CG2 | Y43THR | 3.657 |
| H33TRP | CH2 | ... | OG1 | Y43THR | 3.673 |
| H101TYR | CD1 | ... | CA | Y68ARG | 3.686 |
| H101TYR | OH | ... | CB | Y70PRO | 3.686 |
| H59ASN | ND2 | ... | N | Y44ASN | 3.693 |
| H101TYR | CE1 | ... | CA | Y68ARG | 3.697 |
| H33TRP | CE2 | ... | NH2 | Y68ARG | 3.700 |
| H33TRP | CD1 | ... | NH2 | Y68ARG | 3.706 |
| H99GLY | O | ... | CG | Y68ARG | 3.724 |

| | | | | |
|---|---|---|---|---|
| H59ASN | ND2 | ... | CA | Y44ASN | 3.741 |
| H101TYR | CG | ... | O | Y67GLY | 3.749 |
| H101TYR | CD1 | ... | N | Y68ARG | 3.791 |
| H57SER | O | ... | CG2 | Y43THR | 3.815 |
| H55SER | OG | ... | CB | Y84LEU | 3.818 |
| H57SER | OG | ... | CD1 | Y84LEU | 3.821 |
| H33TRP | CD1 | ... | CD | Y68ARG | 3.831 |
| H55SER | CA | ... | OE1 | Y41GLN | 3.833 |
| H101TYR | CE1 | ... | N | Y69THR | 3.834 |
| H58THR | N | ... | CG2 | Y43THR | 3.843 |
| H101TYR | CE1 | ... | C | Y69THR | 3.847 |
| H55SER | CB | ... | CB | Y84LEU | 3.847 |
| H33TRP | CZ2 | ... | CE1 | Y53TYR | 3.873 |
| H59ASN | ND2 | ... | N | Y45ARG | 3.896 |
| H101TYR | CZ | ... | CA | Y70PRO | 3.902 |
| H47TRP | CH2 | ... | CD | Y45ARG | 3.909 |
| H33TRP | CZ2 | ... | CZ | Y53TYR | 3.951 |
| H101TYR | CE1 | ... | N | Y68ARG | 4.009 |
| H57SER | CB | ... | C | Y41GLN | 4.009 |
| H55SER | CB | ... | CD1 | Y84LEU | 4.108 |
| H33TRP | CG | ... | CZ | Y68ARG | 4.114 |
| H99GLY | C | ... | CB | Y68ARG | 4.122 |
| H100ASN | CA | ... | CB | Y68ARG | 4.125 |
| H57SER | CB | ... | CB | Y43THR | 4.167 |
| H33TRP | CH2 | ... | CB | Y43THR | 4.189 |
| H33TRP | CZ2 | ... | CG2 | Y51THR | 4.220 |
| H52LEU | CD1 | ... | CD1 | Y84LEU | 4.297 |
| H52LEU | CD1 | ... | CB | Y81SER | 4.335 |

PROGRAMS USED

PINQ by Arthur Lesk
WHATIF by Gerrit Vriend
INSIGHT
DISCOVER
CONAN by Michael Scharf
QPACK by Lydia Gregoret
ACCESS by Tim Hubbard

HARDWARE

Digital Equipment VAX Cluster
Silicon Graphics Iris 4D/210
Evans and Sutherland PS390

COORDINATES

AIDA.BRK_MOD L1-L240 light chain, H1-H240 heavy chain, Y1-Y138 Flavodoxin
HETATM 1 water

ACKNOWLEDGEMENTS

G. Vriend
R. Schneider
A. Godzik
T. Hubbard
J. Moult
M. Scharf
and all the participants of the course.

REFERENCES

Chothia, C., Lesk, A. M., Levitt, M., Amit, A. G., Mariuzza, R. A., Phillips, S. E. V., and Poljak, R. J. (1986). Science 233, 755-758.

Chothia, C. and Lesk, A. M. (1987). J. Mol. Biol. 196, 901-917.

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

Davies, D. R., Padlan, E. A., and Sheriff, S. (1990). Annu. Rev. Biochem 59, 439-473.

Janin, J. and Chothia, C. (1990). "Structural Basis of Protein-Protein Recognition," Preprint.

Kabat, E. A., Wu, T. T., Reid-Miller, M., Perry, H. M., Gottesman, K. S. (1987). Sequences of proteins of immunological interest, 4th ed., National Institutes of Health, Bethesda, MD, USA.

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

Sheriff, S., Silverton, E. W., Padlan, E. A., Cohen, G. H., Smith-Gill, S. J., Finzel, B. C., and Davies, D. R. (1987). Proc. Natl. Acad. Sci. 84, 8075-8079.

Tramontano, A., Chothia, C. and Lesk, A. M. (1990) J. Mol. Biol., vol. 214.

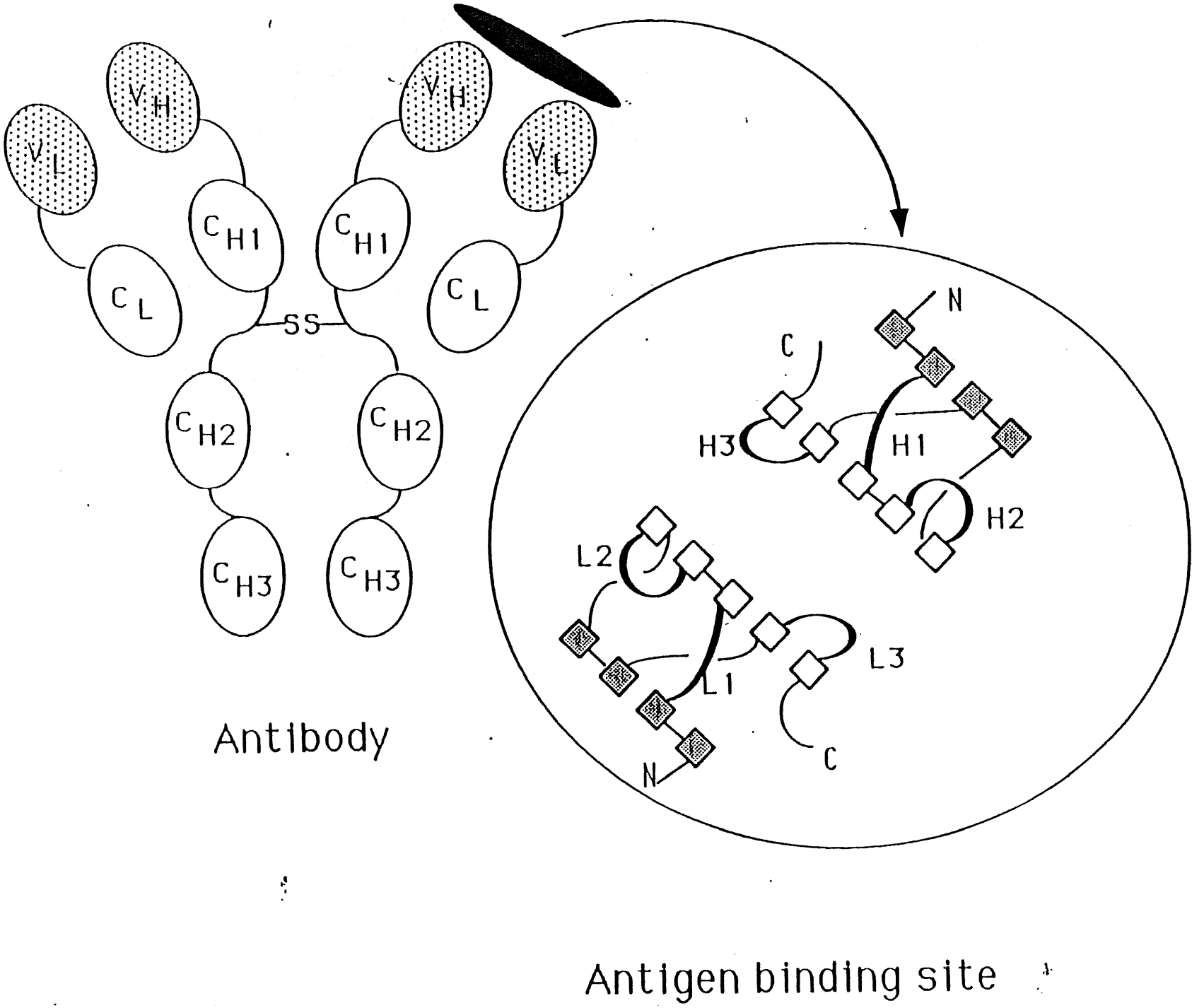

VH
VL
VH
VL
CH1
CH1
CL
CL
SS
CH2
CH2
CH3
CH3
Antibody
N
C
H3
H1
H2
L2
L1
L3
C
N
Antigen binding site

Figure 1.

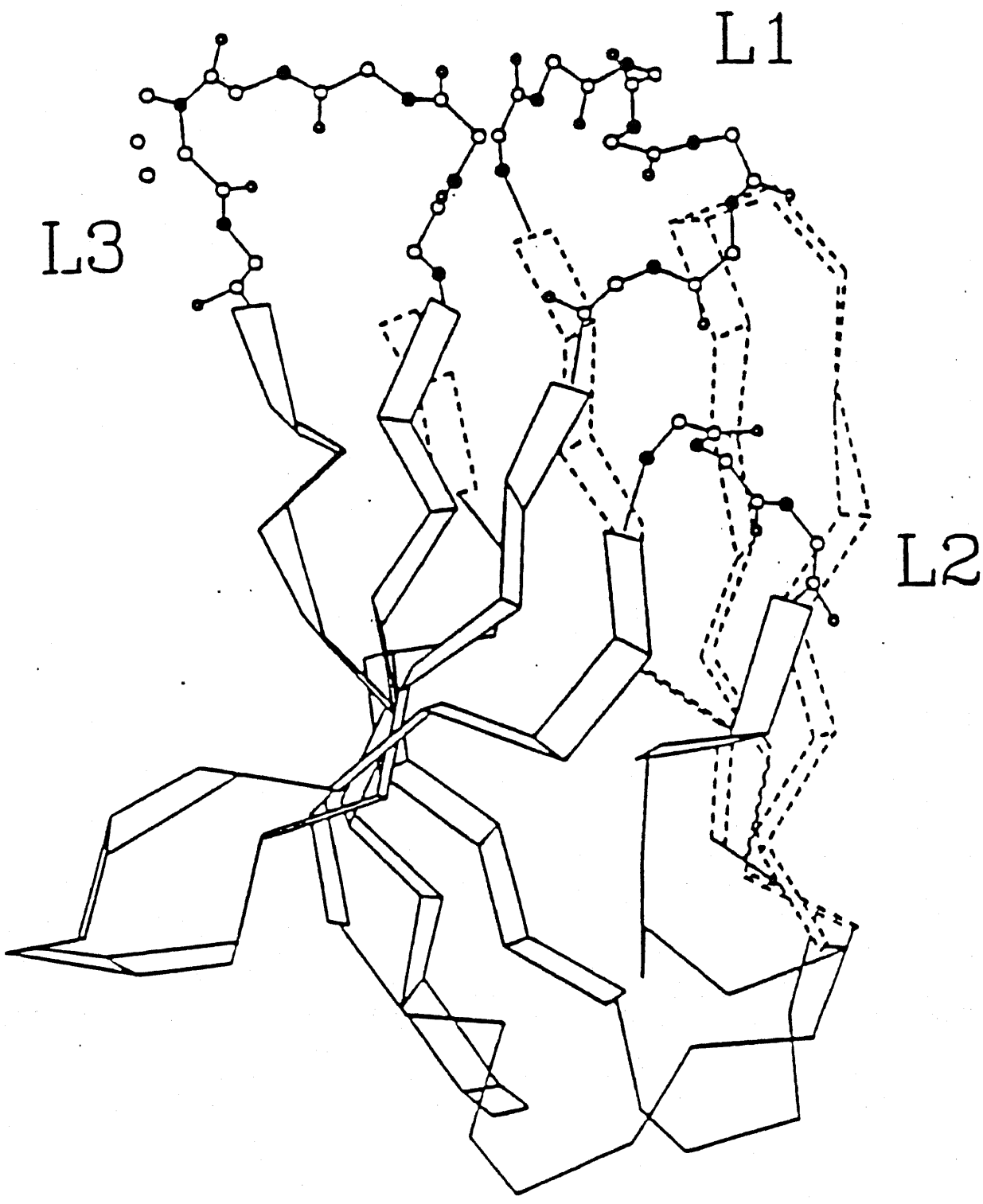



Figure 2.

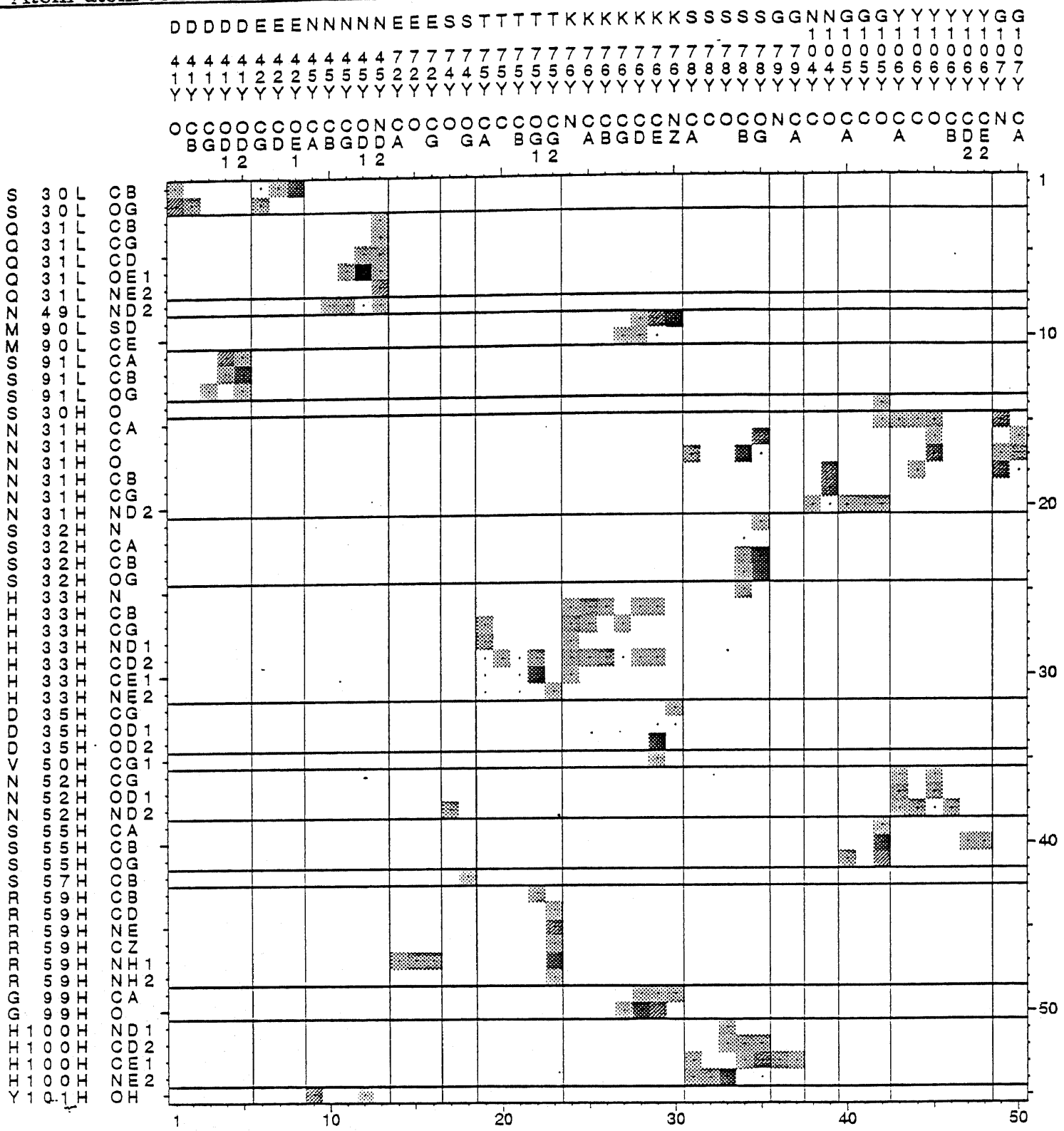



Figure 3a.

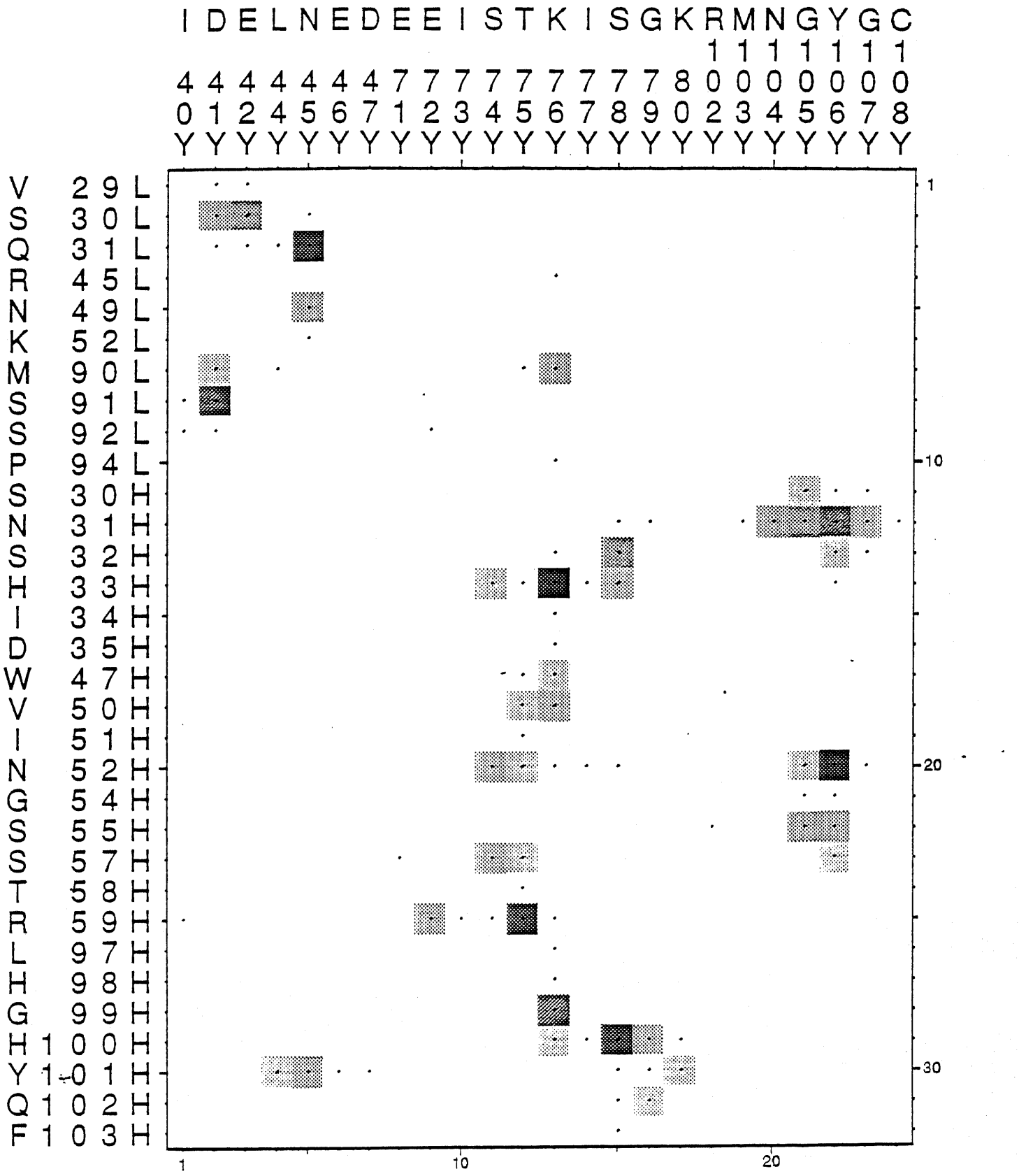

I D E L N E D E E I S T K I S G K R M N G Y G C
1 1 1 1 1 1 1 1
4 4 4 4 4 4 4 7 7 7 7 7 7 7 7 7 8 0 0 0 0 0 0 0
0 1 2 4 5 6 7 1 2 3 4 5 6 7 8 9 0 2 3 4 5 6 7 8
Y Y Y Y Y Y Y Y Y Y Y Y Y Y Y Y Y Y Y Y Y Y Y Y
V 29 L
S 30 L
Q 31 L
R 45 L
N 49 L
K 52 L
M 90 L
S 91 L
S 92 L
P 94 L
S 30 H
N 31 H
S 32 H
H 33 H
I 34 H
D 35 H
W 47 H
V 50 H
I 51 H
N 52 H
G 54 H
S 55 H
S 57 H
T 58 H
R 59 H
L 97 H
H 98 H
G 99 H
H 100 H
Y 101 H
Q 102 H
F 103 H
1
10
20
30
kcal/mol

Figure 3b.

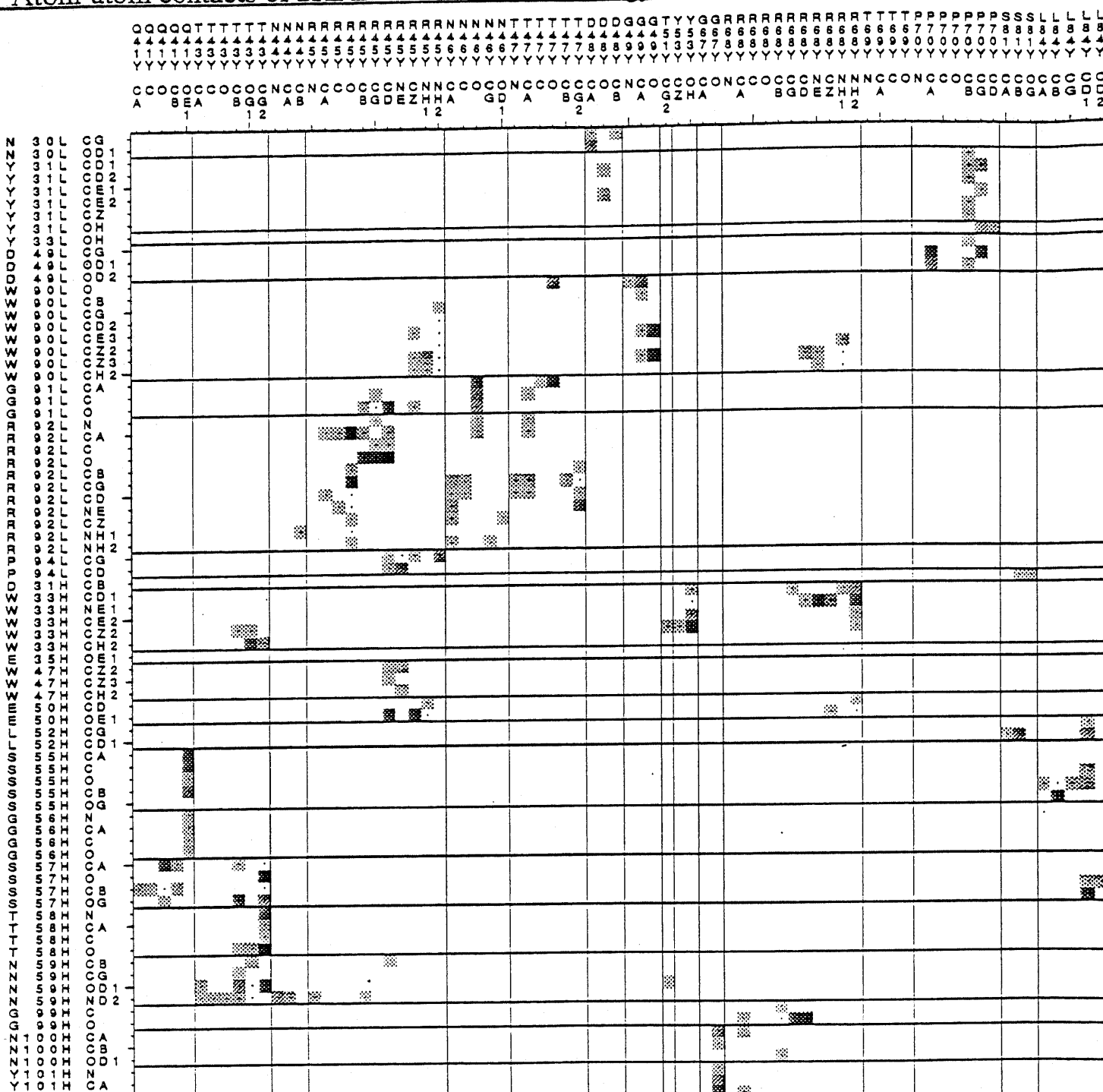
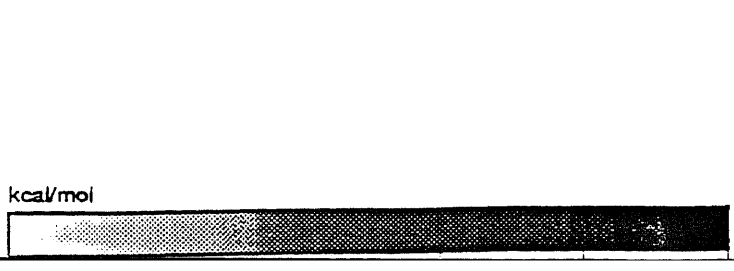



Figure 3c.



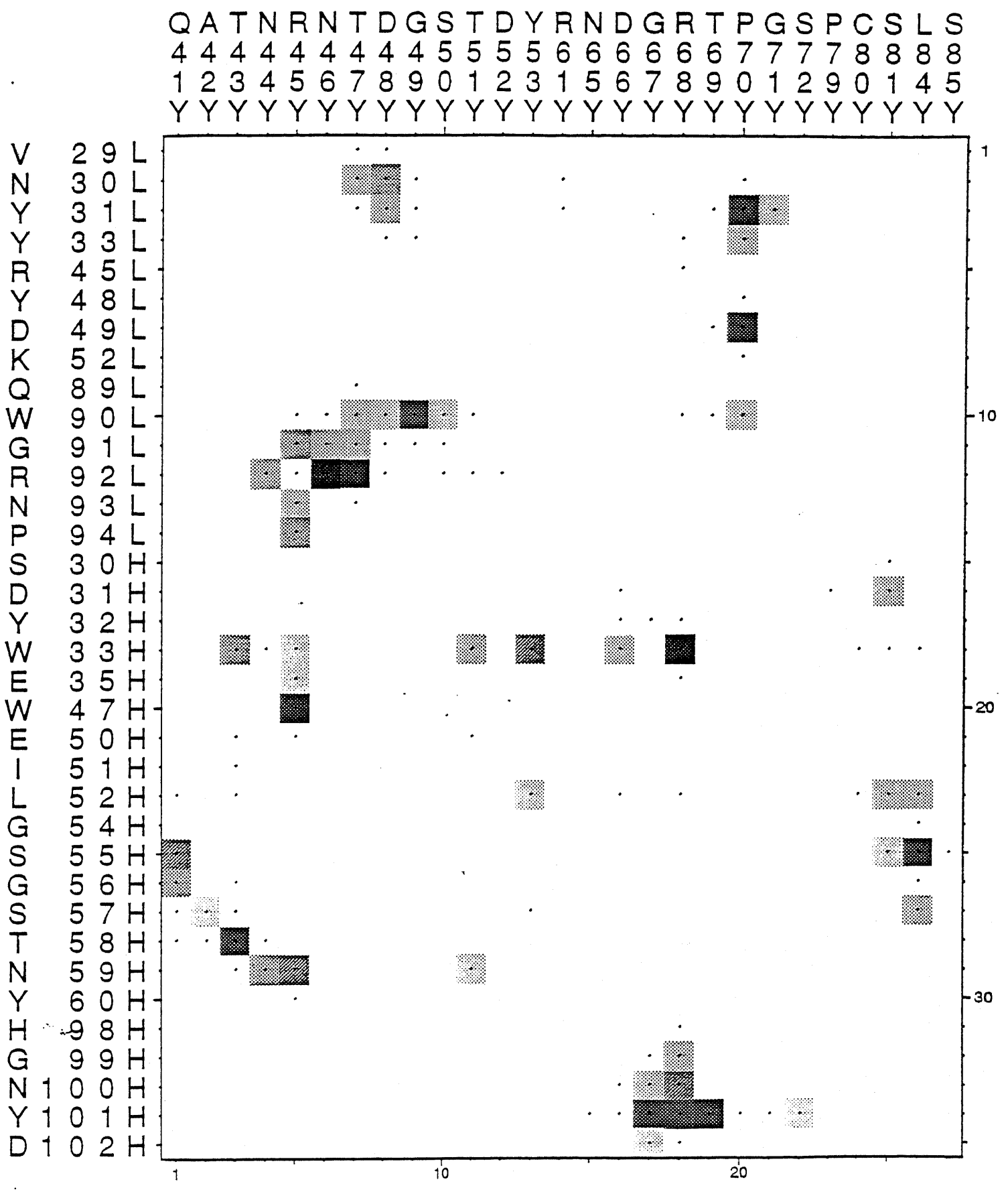

Q A T N R N T D G S T D Y R N D G R T P G S P C S L S
4 4 4 4 4 4 4 4 4 5 5 5 5 6 6 6 6 6 6 7 7 7 7 8 8 8 8
1 2 3 4 5 6 7 8 9 0 1 2 3 1 5 6 7 8 9 0 1 2 9 0 1 4 5
Y Y Y Y Y Y Y Y Y Y Y Y Y Y Y Y Y Y Y Y Y Y Y Y Y Y Y
V 2 9 L
N 3 0 L
Y 3 1 L
Y 3 3 L
R 4 5 L
Y 4 8 L
D 4 9 L
K 5 2 L
Q 8 9 L
W 9 0 L
G 9 1 L
R 9 2 L
N 9 3 L
P 9 4 L
S 3 0 H
D 3 1 H
Y 3 2 H
W 3 3 H
E 3 5 H
W 4 7 H
E 5 0 H
I 5 1 H
L 5 2 H
G 5 4 H
S 5 5 H
G 5 6 H
S 5 7 H
T 5 8 H
N 5 9 H
Y 6 0 H
H 9 8 H
G 9 9 H
N 1 0 0 H
Y 1 0 1 H
D 1 0 2 H
1
10
20
30
1
10
20

Figure 3d.


kcal/mol

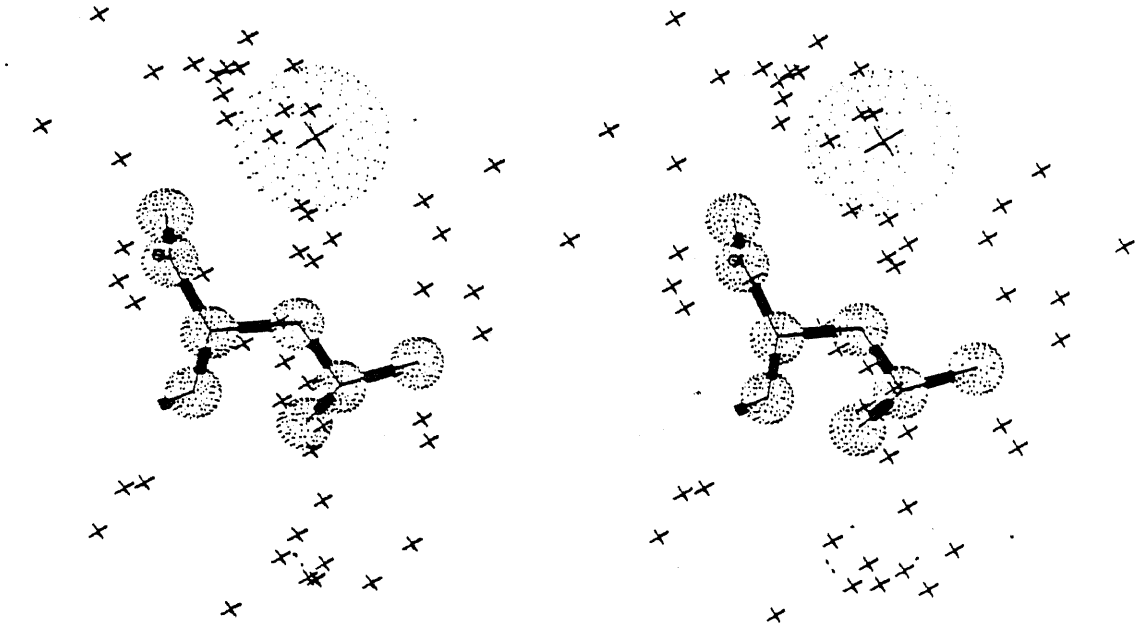


Figure 4. Observed distribution of water molecules (small crosses), with respect to the positions of leucine residues in well-determined natural protein structures. The position of the bound water molecule in AIDA, with respect to Leu Y44, is indicated by the large cross.

# Residue reference parameters

**A-1 Residues at specific segment positions**

**A-2 Residues in contact interfaces**

*Information of sequence about structure*
Both protein structure prediction and protein design require an evaluation of how well particular amino acids fit into a particular structural context. The main approaches to this problem are physical and statistical. The statistical approach is based on counting occurrences of particular sequence-structure combinations and then deriving information values or preference parameters. For example, the number of times N(Ala,H) one observes alanine in a helix is simply divided by the number N(Ala) of alanines and the number N(H) of residues in the helical state and multiplied by the total number N of residues in the database to yield

$$I(Ala,H) = \log( N(Ala,H)*N / N(Ala)*N(H))$$

which is, alternatively, (1) a preference parameter for Ala in state H, or (2) the information that Ala carries about the state H or (3) the logarithm of the ratio of observed (O) and expected (E) number of occurrences of Ala in state H. This can be expressed as

$$I(Ala,H) = \log( O(Ala,N) / E(Ala,H) )$$

or as

$$I(Ala,H)= \log( p(H \mid Ala) / p(H))$$

where $p(H \mid Ala)$ is the conditional probability of observing H given Ala.

*Choosing a simplified description of three-dimensional structure*
While the simple statistics or information theory of this problem have been worked out, the simplified description of protein structure that underlies most statistical approaches is clearly inadequate. The strongest evidence for this comes from the failure of secondary structure prediction schemes to predict, on

the average, more than 65% of all residues accurately in a three-state description of secondary structure. In the tables of preference parameters presented here, derived by Reinhard Schneider [RS89] and Michael Scharf [MS89] in their diploma theses, we go beyond the description of proteins merely in terms of secondary structure. In this we have built on the experience since the first protein design course at EMBL in 1986 [PD86]. At that time we had available for use of the participants position-specific preferences for residues that took account of the beginnings and ends of segments as well as the inside/outside faces of segments, where the outside face is defined as the solvent-accessible surface ([CO81] and [KS85]). Parameters of this type for α-helices were later published by two of the participants of the 1986 course [RR88].

*Buried/inside/outside and beginning/end of segments*
The first (A-1) description more refined than mere secondary structure elements distinguishes between various positions on these elements, e.g. positions near the ends or in the middle of a helix or strand, or positions on the solvent exposed or interior face of a segment [RS89]. The definition of what is removed from solvent ('buried') or exposed to solvent ('outside') or in between ('inside') is made separately for each type of secondary structure, based on a study of the actual distributions of solvent accessibilities in the database of known structures. The results justify the description in the sense that very strong preferences are observed for some residues near the ends of segments and, for others, on the solvent accessible face of segments. In fact, in protein design this type of parameter is primarily used for choosing residues on the outside face and near the ends of segments.

*Interfaces between secondary structure segments and with solvent*
A second (A-2) description of protein structure, more sophisticated than the standard helix-strand-turn-loop classification, is in terms of contact interfaces. This type of description is motivated by observations that residue preferences vary strongly on different sides of secondary structure segments. For example, the amino acid composition of solvent exposed faces of helices is very different compared to that of helix-sheet interfaces. The interface definition described in A-2 is in terms of the residue type and secondary structure of a central residue and

the secondary structure and chain distance of its contact partners. As protein-solvent interaction appears to be the physically dominant effect in the folding of globular proteins, water contacts are explicitly taken into account. In protein design, this type of parameter is extremely useful in choosing residues at helix-sheet, sheet-sheet, strand-strand and helix-helix interfaces.

## References


(CO81) C. Oefner, Diploma thesis, Universität Heidelberg, 1981

(KS85) W Kabsch and C Sander, unpublished results

PD86) Protein Design 86, EMBL BIOcomputing Technical Document 1, C.Sander, ed. (1987).

(RR88) Richardson, J.S. & Richardson, D.C. (1988). Amino acid preferences for specific locations at the ends of helices. Science 240, 1648-1652.

(RS89) R. Schneider, Diploma thesis, Universität Heidelberg, 1989

(MS89) M. Scharf, Diploma thesis, Universität Heidelberg, 1989

(SS92) C Sander, M Scharf and R Schneider, in Protein Engineering, A Practical Approach, eds. A. Rees and M. Sternberg, Oxford University Press, 1992.

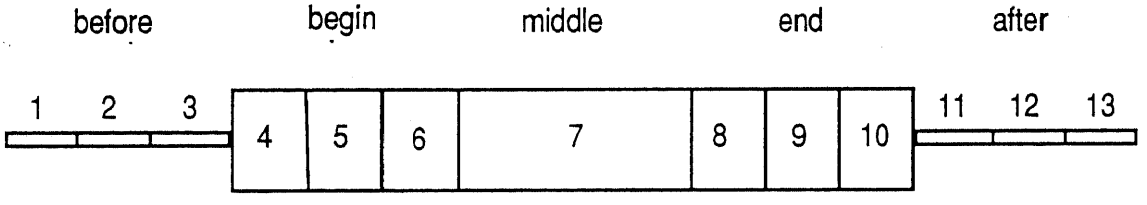
# Structural preference parameters of amino acids

(R. Schneider and C.Sander, manuscript in preparation and R.Schneider, Diplomarbeit)

**Here we only list single amino acid preferences; pair, triplet and pentapeptide preferences are available. They are stored on the VAX in the directory $course:[vriend.prefpar].**
**Please ask Reinhard Schneider if you have further questions.**

R: amino acid in one-letter code
S: secondary structure type according to DSSP
(H: α-helix ; E: β-strand ; L: not α and not β)
X: exposure to water
P: position in a secondary structure segment

**secondary structure segment and assignment of position number**




## Inside / outside assignment

The assignment depends on the maximal accessible surface area of an amino acid (Frömmel and Sander) and on the secondary structure in which the residue participates.

3 different states : B: buried
I: inside
O: outside

| secondary structure type | % of maximal surface area exposed | | |
|---|---|---|---|
| | buried | inside | outside |
| E | 5 | 15 | >15 |
| L | 35 | 60 | >60 |
| H | 25 | 50 | >50 |

## preference parameter of single amino acids for a secondary structure type

| S | V | L | I | M | F | W | Y | G | A | P | S | T | C | H | R | K | Q | E | N | D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| E | 0.6 | 0.4 | 0.6 | 0.5 | 0.5 | 0.1 | 0.3 | -0.5 | -0.1 | -0.6 | -0.1 | 0.2 | 0.0 | -0.2 | -0.1 | -0.4 | -0.1 | -0.4 | -0.6 | -0.7 |
| L | -0.4 | -0.4 | -0.5 | -0.5 | -0.3 | 0.0 | -0.1 | 0.3 | -0.1 | 0.4 | 0.1 | 0.0 | -0.1 | 0.1 | 0.1 | 0.1 | 0.0 | 0.0 | 0.3 | 0.2 |
| H | 0.0 | 0.1 | 0.0 | 0.2 | 0.0 | -0.2 | -0.3 | -0.4 | 0.3 | -0.5 | -0.2 | -0.2 | 0.2 | 0.0 | -0.1 | 0.2 | 0.1 | 0.3 | -0.1 | 0.0 |

## preference parameter of single amino acids for an inside/outside position in a secondary structure type

| S X | V | L | I | M | F | W | Y | G | A | P | S | T | C | H | R | K | Q | E | N | D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| E B | 1.0 | 0.9 | 1.2 | 0.9 | 0.9 | 0.4 | 0.2 | -0.2 | 0.2 | -0.9 | -0.4 | -0.1 | 0.6 | -1.0 | -1.4 | -2.2 | -1.0 | -1.7 | -1.6 | -1.8 |
| E I | 0.6 | 0.5 | 0.7 | 0.6 | 0.8 | 0.3 | 0.8 | -0.4 | -0.4 | -0.8 | -0.3 | -0.1 | 0.0 | 0.3 | -0.1 | -0.6 | -0.3 | -0.7 | -0.7 | -1.1 |
| E O | 0.0 | -0.2 | 0.0 | 0.1 | -0.1 | -0.3 | 0.2 | -0.8 | -0.3 | -0.3 | 0.1 | 0.3 | -1.0 | 0.1 | 0.4 | 0.3 | 0.3 | 0.1 | -0.1 | -0.2 |
| L B | 0.0 | 0.0 | 0.0 | -0.1 | 0.2 | 0.3 | 0.3 | 0.3 | -0.1 | 0.2 | 0.1 | -0.1 | 0.5 | 0.2 | -0.2 | -0.6 | -0.2 | -0.4 | 0.0 | 0.0 |
| L I | -0.5 | -0.6 | -0.8 | -0.7 | -0.8 | -0.4 | -0.3 | 0.2 | -0.2 | 0.5 | 0.1 | 0.2 | -0.3 | 0.0 | 0.3 | 0.3 | 0.1 | 0.2 | 0.4 | 0.3 |
| L O | -1.0 | -0.9 | -1.2 | -0.9 | -0.9 | -0.1 | -0.5 | 0.5 | -0.1 | 0.6 | 0.3 | 0.0 | -1.4 | 0.0 | 0.2 | 0.4 | 0.1 | 0.3 | 0.5 | 0.4 |
| H B | 0.6 | 0.8 | 0.7 | 0.8 | 0.6 | 0.5 | 0.1 | -0.5 | 0.4 | -1.2 | -0.4 | -0.2 | 0.9 | -0.2 | -0.8 | -0.8 | -0.6 | -0.6 | -0.7 | -1.0 |
| H I | -0.4 | -0.4 | -0.6 | -0.3 | -0.6 | -0.6 | -0.1 | -0.5 | 0.2 | -0.4 | -0.3 | -0.3 | -0.6 | 0.1 | 0.4 | 0.6 | 0.5 | 0.6 | 0.1 | 0.3 |
| H O | -0.8 | -1.1 | -1.4 | -0.7 | -1.1 | -1.3 | -1.1 | -0.3 | 0.3 | 0.0 | 0.1 | 0.0 | -1.2 | 0.1 | 0.1 | 0.6 | 0.4 | 0.8 | 0.4 | 0.6 |

# position preference parameter of a single amino acid in a secondary structure segment

## β-strand

| S | P | V | L | I | M | F | W | Y | G | A | P | S | T | C | H | R | K | Q | E | N | D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| E | 1 | -0.5 | -0.4 | -0.5 | -1.0 | -0.3 | -0.1 | 0.1 | 0.2 | 0.0 | 0.3 | 0.2 | -0.2 | -0.6 | 0.3 | 0.3 | 0.1 | -0.1 | 0.0 | 0.4 | 0.3 |
| E | 2 | -0.6 | -0.6 | -0.9 | -0.5 | -0.4 | 0.1 | -0.7 | 0.6 | -0.2 | 0.4 | 0.1 | -0.2 | -0.7 | 0.1 | 0.0 | 0.0 | 0.2 | 0.2 | 0.4 | 0.6 |
| E | 3 | -0.5 | -0.4 | -0.4 | 0.1 | 0.1 | 0.4 | 0.2 | 0.4 | -0.3 | 0.3 | 0.1 | -0.2 | -0.3 | 0.2 | 0.0 | 0.1 | 0.1 | -0.2 | 0.2 | 0.1 |
| E | 4 | 0.2 | 0.1 | 0.5 | 0.6 | 0.4 | 0.0 | 0.4 | -0.7 | -0.2 | -0.5 | 0.0 | 0.2 | 0.0 | -0.1 | 0.0 | -0.1 | 0.3 | -0.3 | -0.7 | -0.9 |
| E | 5 | 0.9 | 0.4 | 0.9 | 0.7 | 0.5 | 0.7 | 0.2 | -0.9 | -0.1 | -0.9 | -0.4 | 0.1 | 0.1 | -0.5 | -0.3 | -0.5 | -0.4 | -0.5 | -0.6 | -1.0 |
| E | 6 | 0.7 | 0.6 | 0.6 | 0.4 | 0.6 | 0.2 | 0.7 | -0.8 | 0.0 | -1.1 | -0.2 | 0.1 | 0.3 | 0.2 | 0.3 | -0.7 | -0.3 | -0.7 | -1.1 | -1.6 |
| E | 7 | 0.4 | 0.3 | 0.6 | 0.4 | 0.2 | -0.2 | -0.1 | -0.1 | 0.0 | -0.7 | 0.0 | 0.2 | -0.4 | -0.2 | -0.3 | -0.3 | 0.0 | -0.3 | -0.5 | -0.7 |
| E | 8 | 0.7 | 0.6 | 0.9 | 0.0 | 0.7 | -0.7 | 0.2 | -0.2 | 0.2 | -1.0 | -0.3 | 0.3 | 0.0 | -0.6 | -0.5 | -0.7 | -0.5 | -0.7 | -0.8 | -0.9 |
| E | 9 | 0.7 | 0.5 | 0.7 | 0.6 | 0.5 | -0.1 | 0.2 | -0.9 | -0.1 | -0.8 | -0.2 | 0.0 | -0.3 | -0.2 | 0.0 | -0.4 | -0.1 | -0.5 | -0.7 | -0.9 |
| E | 10 | 0.3 | 0.6 | 0.5 | 0.2 | 0.2 | 0.0 | 0.2 | -0.4 | -0.5 | 0.0 | -0.1 | 0.0 | 0.0 | -0.2 | 0.0 | -0.4 | -0.4 | -0.3 | -0.3 | 0.0 |
| E | 11 | 0.0 | -0.2 | -0.3 | -0.7 | -0.3 | 0.0 | -0.1 | 0.3 | 0.0 | 0.2 | 0.2 | 0.2 | -0.2 | 0.1 | -0.1 | -0.3 | -0.4 | -0.1 | 0.3 | 0.1 |
| E | 12 | -0.7 | -0.7 | -0.6 | -0.9 | -0.5 | 0.1 | -0.3 | 0.5 | -0.1 | 0.5 | 0.3 | 0.2 | -0.8 | -0.1 | 0.1 | 0.1 | 0.0 | 0.1 | 0.1 | 0.3 |
| E | 13 | -0.4 | -0.6 | -0.5 | -0.5 | -0.2 | 0.2 | 0.2 | 0.2 | -0.2 | 0.3 | 0.1 | 0.0 | -0.1 | 0.2 | 0.1 | 0.1 | -0.1 | 0.2 | 0.0 | 0.4 |

position preference parameter of a single amino acid in a secondary structure segment

loop

| S | P | V | L | I | M | F | W | Y | G | A | P | S | T | C | H | R | K | Q | E | N | D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| L | 1 | 0.4 | 0.4 | 0.5 | 0.1 | 0.3 | 0.3 | 0.1 | -0.3 | 0.1 | -0.4 | -0.2 | 0.0 | -0.2 | 0.0 | -0.1 | -0.2 | -0.3 | -0.2 | -0.5 | -0.6 |
| L | 2 | 0.4 | 0.2 | 0.4 | 0.5 | 0.2 | -0.3 | 0.0 | -0.7 | 0.0 | -1.0 | -0.1 | 0.0 | 0.4 | -0.2 | 0.0 | 0.0 | 0.1 | 0.0 | -0.3 | -0.5 |
| L | 3 | 0.1 | 0.4 | 0.2 | 0.0 | 0.2 | -0.1 | 0.3 | -0.3 | -0.1 | -0.7 | -0.1 | -0.1 | 0.3 | 0.0 | -0.1 | -0.1 | 0.0 | -0.2 | -0.1 | -0.1 |
| L | 4 | -0.1 | -0.1 | -0.4 | -0.1 | -0.1 | -0.1 | -0.1 | 0.5 | 0.0 | -0.2 | 0.1 | 0.0 | 0.1 | 0.1 | 0.0 | -0.2 | -0.3 | -0.1 | 0.3 | 0.0 |
| L | 5 | -0.5 | -0.5 | -0.5 | -0.9 | -0.4 | -0.2 | -0.3 | 0.5 | 0.0 | 0.6 | 0.1 | 0.1 | -1.4 | 0.0 | 0.2 | 0.2 | 0.2 | 0.0 | 0.1 | 0.2 |
| L | 6 | -0.5 | -0.3 | -0.6 | -0.4 | -0.6 | 0.2 | 0.0 | 0.3 | -0.2 | 0.4 | -0.1 | 0.0 | -0.2 | 0.1 | 0.3 | 0.3 | 0.0 | 0.1 | 0.2 | 0.3 |
| L | 7 | -0.3 | -0.4 | -0.3 | -0.6 | -0.4 | 0.1 | 0.1 | 0.1 | -0.1 | 0.4 | 0.0 | 0.0 | 0.2 | -0.1 | 0.2 | 0.0 | 0.1 | 0.0 | 0.1 | 0.2 |
| L | 8 | -0.4 | -0.3 | -0.5 | -1.5 | -0.3 | -0.4 | -0.1 | 0.1 | -0.2 | 0.5 | 0.1 | -0.4 | 0.0 | 0.3 | 0.4 | 0.2 | 0.0 | 0.1 | 0.4 | 0.4 |
| L | 9 | -0.4 | -0.3 | -0.4 | -0.3 | -0.3 | 0.3 | -0.4 | 0.4 | -0.2 | 0.4 | 0.0 | -0.1 | -0.1 | 0.0 | 0.1 | 0.1 | 0.0 | 0.0 | 0.2 | 0.4 |
| L | 10 | -0.9 | -0.5 | -0.7 | -0.3 | -0.1 | 0.0 | -0.1 | 0.3 | -0.4 | 0.4 | 0.4 | 0.1 | -0.1 | 0.2 | -0.1 | 0.1 | 0.0 | -0.1 | 0.5 | 0.3 |
| L | 11 | 0.1 | 0.0 | 0.2 | 0.4 | 0.2 | 0.0 | 0.2 | -0.5 | 0.0 | 0.4 | 0.0 | 0.1 | 0.0 | -0.2 | -0.1 | 0.0 | 0.2 | -0.1 | -0.6 | -0.5 |
| L | 12 | 0.4 | -0.1 | 0.3 | 0.2 | 0.0 | 0.2 | -0.2 | -0.4 | 0.1 | -0.2 | -0.1 | 0.0 | 0.1 | -0.3 | -0.3 | -0.2 | -0.1 | 0.3 | -0.2 | 0.1 |
| L | 13 | 0.4 | 0.3 | 0.3 | 0.3 | 0.4 | 0.1 | 0.4 | -0.5 | 0.0 | -0.8 | -0.1 | -0.1 | 0.2 | 0.0 | -0.2 | -0.4 | -0.1 | 0.0 | -0.2 | -0.1 |

## position preference parameter of a single amino acid in a secondary structure segment

α-helix

| S | P | V | L | I | M | F | W | Y | G | A | P | S | T | C | H | R | K | Q | E | N | D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| H | 1 | 0.0 | -0.2 | -0.1 | 0.0 | 0.1 | -1.5 | 0.1 | 0.3 | 0.1 | 0.0 | -0.1 | -0.1 | 0.3 | -0.1 | -0.1 | 0.3 | -0.1 | -0.2 | 0.0 | -0.3 |
| H | 2 | 0.2 | 0.4 | 0.3 | 0.3 | 0.1 | -0.4 | -0.2 | -0.1 | 0.2 | 0.2 | 0.0 | 0.2 | -0.1 | -0.2 | -0.2 | -0.2 | -0.4 | -0.7 | -0.4 | -0.1 |
| H | 3 | -1.7 | -0.8 | -1.7 | -0.9 | -0.9 | -1.4 | -0.7 | 0.2 | -0.3 | 0.7 | 0.7 | 0.4 | 0.4 | 0.2 | -0.2 | -0.2 | -0.4 | -0.1 | 0.7 | 0.7 |
| H | 4 | 0.0 | 0.0 | -0.1 | -0.3 | -0.2 | 0.1 | -0.2 | -0.3 | 0.2 | 1.2 | 0.0 | -0.1 | 0.0 | -0.5 | -0.4 | 0.1 | -0.2 | 0.2 | -0.4 | -0.1 |
| H | 5 | -0.6 | -1.3 | -1.0 | -0.5 | -0.8 | -0.1 | -1.0 | 0.1 | 0.4 | 0.1 | 0.1 | 0.0 | 0.3 | -0.1 | -0.4 | 0.0 | 0.1 | 0.8 | 0.2 | 0.7 |
| H | 6 | 0.0 | -0.3 | -0.4 | 0.4 | -0.1 | -0.5 | 0.0 | -1.2 | -0.1 | -0.3 | -0.4 | -0.2 | -0.7 | 0.0 | -0.7 | -0.2 | 0.6 | 1.0 | 0.1 | 0.8 |
| H | 7 | 0.2 | 0.4 | 0.3 | 0.5 | 0.1 | -0.2 | -0.4 | -0.5 | 0.5 | -1.8 | -0.4 | -0.2 | 0.3 | -0.1 | 0.0 | 0.2 | 0.0 | 0.2 | -0.2 | -0.4 |
| H | 8 | 0.2 | 0.6 | 0.5 | 0.5 | 0.0 | 0.2 | -0.8 | -0.6 | 0.3 | -1.9 | -0.5 | -0.3 | -0.8 | 0.4 | 0.2 | 0.4 | -0.1 | 0.0 | -0.2 | -0.5 |
| H | 9 | 0.1 | 0.3 | 0.2 | 0.6 | 0.1 | -0.6 | -0.4 | -0.8 | 0.2 | -1.0 | -0.2 | -0.2 | 0.0 | 0.0 | -0.1 | 0.4 | 0.3 | 0.2 | 0.0 | -0.4 |
| H | 10 | -0.3 | 0.1 | -0.6 | -0.2 | 0.1 | -0.1 | 0.5 | -0.3 | 0.3 | -2.2 | -0.2 | -0.3 | 0.6 | 0.2 | 0.1 | 0.3 | 0.3 | 0.0 | 0.2 | -0.1 |
| H | 11 | -0.4 | 0.2 | -0.4 | 0.2 | 0.3 | -0.2 | 0.2 | 0.6 | -0.2 | -0.9 | 0.0 | -0.3 | 0.4 | 0.1 | 0.1 | 0.0 | -0.1 | -0.2 | 0.4 | -0.1 |
| H | 12 | -0.6 | -0.3 | -0.5 | -0.5 | -0.3 | -1.1 | -0.3 | 0.6 | 0.1 | 0.5 | 0.0 | -0.1 | -0.4 | 0.0 | -0.1 | 0.4 | 0.1 | -0.3 | 0.3 | 0.0 |
| H | 13 | -0.1 | -0.1 | -0.5 | 0.0 | -0.6 | -0.2 | -0.3 | 0.0 | -0.2 | 0.5 | -0.2 | 0.1 | 0.0 | 0.1 | 0.0 | 0.4 | 0.1 | 0.0 | 0.2 | 0.2 |

# position preference parameter of a single amino acid in an inside/outside position of a secondary structure segment

## β-strand

| S | X | P | V | L | I | M | F | W | Y | G | A | P | S | T | C | H | R | K | Q | E | N | D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| E | B | 1 | 0.2 | 0.7 | 0.5 | 0.0 | 0.8 | 1.0 | 0.2 | -0.1 | 0.3 | -2.2 | 0.3 | -0.5 | 1.4 | 0.1 | -0.2 | -2.3 | -0.4 | -1.5 | -0.7 | -1.5 |
| E | B | 2 | 0.5 | 0.2 | 0.3 | 0.8 | 0.8 | 0.3 | -0.8 | 0.8 | 0.2 | 0.0 | -0.1 | -0.6 | 0.1 | -2.0 | -1.4 | -1.5 | -3.6 | -1.4 | 0.4 | 0.4 |
| E | B | 3 | 0.0 | 0.6 | 0.5 | 0.8 | 0.5 | 1.8 | 0.4 | 0.1 | -0.3 | 0.7 | 0.1 | -0.3 | 0.4 | -0.1 | -1.9 | -1.8 | -0.8 | -1.4 | -1.0 | -0.5 |
| E | B | 4 | 0.2 | 0.5 | 1.2 | 1.0 | 0.9 | 0.5 | 0.3 | -0.2 | -0.1 | -0.4 | 0.0 | -0.1 | 0.9 | -0.9 | -1.4 | -1.3 | 0.1 | -1.3 | -1.4 | -2.2 |
| E | B | 5 | 1.3 | 0.9 | 1.3 | 1.1 | 0.9 | 0.9 | 0.1 | -0.8 | 0.2 | -1.8 | -0.9 | -0.4 | 0.7 | -0.6 | -1.8 | -1.9 | -1.1 | -1.7 | -1.7 | -2.4 |
| E | B | 6 | 1.1 | 1.0 | 0.9 | 0.7 | 1.0 | 0.4 | 0.7 | -0.3 | 0.4 | -1.2 | -0.5 | -0.5 | 0.5 | -0.6 | -0.7 | -2.4 | -1.4 | -2.6 | -2.0 | -2.5 |
| E | B | 7 | 1.0 | 0.9 | 1.1 | 0.7 | 0.8 | 0.3 | 0.0 | 0.2 | 0.2 | -1.9 | -0.4 | 0.2 | 0.4 | -0.9 | -1.3 | -2.5 | -0.8 | -1.6 | -2.4 | -2.4 |
| E | B | 8 | 1.1 | 1.0 | 1.3 | 0.1 | 0.8 | -1.0 | -0.2 | 0.0 | 0.4 | -0.7 | -0.9 | 0.2 | -0.2 | -0.8 | -1.1 | -3.1 | -1.8 | -1.4 | -1.4 | -1.6 |
| E | B | 9 | 1.2 | 0.9 | 1.1 | 0.9 | 0.7 | 0.4 | 0.2 | -0.5 | 0.3 | -0.7 | -0.4 | 0.1 | 0.4 | -2.0 | -1.8 | -2.8 | -1.0 | -2.1 | -1.5 | -2.0 |
| E | B | 0 | 0.9 | 1.1 | 1.1 | 0.7 | 0.5 | 0.1 | 0.0 | -0.2 | -0.4 | -0.4 | -0.3 | -0.1 | 0.9 | -1.6 | -1.2 | -1.9 | -1.5 | -1.4 | -1.2 | -0.5 |
| E | B | 1 | 0.7 | 0.6 | 0.5 | 0.0 | 0.5 | 1.1 | 0.2 | 0.8 | 0.3 | -1.0 | 0.1 | 0.2 | 0.8 | -0.5 | -1.7 | -2.4 | -2.1 | -1.6 | -2.1 | -0.9 |
| E | B | 2 | 0.3 | -0.2 | -0.4 | -0.9 | 0.2 | -1.5 | -0.9 | 1.1 | 0.5 | 0.1 | 0.4 | 0.3 | 0.1 | -1.1 | -0.9 | -0.4 | -1.3 | -1.5 | -0.7 | -0.1 |
| E | B | 3 | 0.5 | 0.2 | 0.8 | 0.1 | 1.0 | 1.7 | 0.8 | 0.2 | 0.0 | -0.1 | -0.8 | 0.1 | 1.1 | -0.7 | -1.6 | -2.0 | -1.3 | -1.2 | -1.5 | -0.3 |
| E | I | 1 | 0.0 | 0.6 | 0.6 | -0.4 | 0.2 | 0.0 | 0.4 | -0.3 | 0.0 | -0.1 | 0.2 | -0.7 | 0.4 | 0.6 | -0.1 | -0.8 | -0.3 | -0.8 | -0.3 | -0.5 |
| E | I | 2 | 0.4 | 0.4 | 0.5 | -1.1 | 0.1 | 0.0 | 0.4 | 0.6 | 0.3 | -1.5 | -0.1 | 0.0 | 1.1 | -1.3 | -1.2 | -0.8 | 0.3 | -2.0 | -0.5 | -0.2 |
| E | I | 3 | -0.1 | 0.1 | -0.3 | 0.2 | 0.6 | 0.8 | 0.7 | 0.4 | -0.1 | -0.7 | -0.3 | -0.4 | 0.0 | 0.7 | -0.4 | -0.5 | 0.3 | -0.6 | -0.1 | -0.5 |
| E | I | 4 | 0.4 | 0.5 | 0.5 | 0.6 | 0.9 | 0.3 | 1.0 | -0.5 | -0.4 | -0.1 | -0.6 | 0.0 | -0.1 | 0.4 | -0.1 | -0.8 | -0.2 | -0.7 | -1.7 | -0.7 |
| E | I | 5 | 0.7 | 0.3 | 0.4 | 0.7 | 1.0 | 0.6 | 0.5 | -0.6 | -0.5 | -0.3 | -0.2 | -0.4 | 0.8 | -0.6 | -0.1 | -0.8 | -0.2 | -0.8 | -0.1 | -1.0 |
| E | I | 6 | 0.3 | 0.2 | 0.6 | 0.9 | 0.1 | 0.0 | 0.8 | -1.2 | -0.9 | -1.5 | -0.3 | 0.4 | 0.8 | 0.8 | 0.9 | 0.0 | -0.4 | -0.5 | -1.3 | -3.6 |
| E | I | 7 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | -0.2 | -0.1 | -0.5 | -0.4 | -0.4 | 0.3 | 0.2 | -0.7 | -0.2 | -0.6 | -0.6 | 0.5 | 0.1 | -0.5 | -0.9 |
| E | I | 8 | 0.4 | 0.4 | 0.4 | 0.0 | 1.4 | 1.1 | 0.7 | -0.1 | -0.2 | -1.9 | -0.7 | -0.5 | 0.7 | 0.6 | -0.2 | -0.6 | -0.7 | -1.5 | -0.2 | -0.9 |
| E | I | 9 | 0.9 | 0.8 | 1.0 | 0.9 | 0.9 | -1.1 | 1.0 | -1.3 | -0.2 | -2.5 | -0.6 | -0.3 | -0.9 | 0.1 | 0.0 | -0.7 | -1.1 | -1.1 | -0.5 | -1.5 |
| E | I | 0 | 0.2 | 0.6 | 0.7 | 0.3 | 0.6 | 0.9 | 0.8 | 0.2 | -0.4 | -0.8 | -0.4 | -0.5 | 0.0 | 0.3 | -0.5 | -0.7 | -0.4 | -0.4 | -0.7 | -0.7 |
| E | I | 1 | 0.4 | 0.2 | 0.2 | -0.1 | 0.2 | 0.7 | 0.8 | 0.2 | 0.1 | 0.3 | -0.1 | -0.1 | 1.1 | 0.1 | -0.6 | -1.7 | -1.6 | -0.7 | -0.4 | -0.3 |
| E | I | 2 | -0.4 | -0.8 | 0.3 | -1.3 | -0.5 | -0.9 | 0.4 | 0.2 | 0.3 | 0.7 | -0.5 | 0.2 | 0.3 | 0.6 | 0.0 | -0.3 | 0.1 | -0.2 | -0.4 | 0.0 |
| E | I | 3 | -0.1 | 0.3 | 0.4 | -0.8 | 0.4 | -0.1 | 0.7 | -0.4 | -0.4 | 0.0 | -0.3 | -0.3 | 0.5 | 0.6 | -0.3 | -0.5 | 0.0 | 0.0 | -0.2 | 0.1 |
| E | O | 1 | -0.7 | -0.8 | -1.0 | -1.3 | -0.7 | -0.3 | -0.1 | 0.3 | -0.1 | 0.4 | 0.2 | -0.2 | -1.4 | 0.2 | 0.3 | 0.2 | 0.0 | 0.1 | 0.5 | 0.4 |
| E | O | 2 | -0.9 | -0.8 | -1.2 | -0.8 | -0.7 | 0.1 | -0.8 | 0.6 | -0.3 | 0.4 | 0.1 | -0.2 | -1.1 | 0.3 | 0.1 | 0.1 | 0.3 | 0.4 | 0.4 | 0.6 |
| E | O | 3 | -0.8 | -0.9 | -0.7 | -0.2 | -0.2 | -0.6 | 0.0 | 0.4 | -0.4 | 0.3 | 0.1 | -0.1 | -0.7 | 0.2 | 0.2 | 0.3 | 0.1 | 0.0 | 0.4 | 0.2 |
| E | O | 4 | 0.1 | -0.4 | -0.1 | 0.4 | -0.2 | -0.4 | 0.2 | -1.0 | -0.2 | -0.7 | 0.1 | 0.4 | -0.8 | 0.0 | 0.3 | 0.4 | 0.5 | 0.1 | -0.4 | -0.7 |
| E | O | 5 | 0.3 | -0.2 | 0.2 | 0.1 | -0.4 | 0.2 | 0.2 | -1.2 | -0.4 | -0.5 | -0.1 | 0.6 | -1.6 | -0.3 | 0.4 | 0.2 | -0.1 | 0.2 | -0.1 | -0.3 |
| E | O | 6 | 0.0 | 0.0 | -0.2 | -0.9 | 0.1 | 0.0 | 0.7 | -1.3 | -0.3 | -0.8 | 0.0 | 0.3 | -0.4 | 0.4 | 0.7 | 0.0 | 0.5 | 0.1 | -0.3 | -0.6 |
| E | O | 7 | -0.4 | -0.5 | 0.0 | 0.0 | -0.6 | -0.5 | -0.2 | -0.3 | -0.2 | -0.2 | 0.1 | 0.2 | -1.7 | 0.1 | 0.3 | 0.4 | 0.3 | 0.2 | 0.2 | -0.1 |
| E | O | 8 | 0.1 | -0.1 | 0.3 | -0.1 | 0.1 | -0.9 | 0.3 | -0.4 | -0.2 | -0.9 | 0.1 | 0.5 | 0.1 | -0.7 | -0.1 | 0.2 | 0.2 | -0.1 | -0.4 | -0.4 |
| E | O | 9 | -0.1 | -0.1 | -0.1 | 0.0 | 0.0 | -0.3 | -0.2 | -1.2 | -0.5 | -0.5 | 0.1 | 0.0 | -0.9 | 0.4 | 0.7 | 0.4 | 0.5 | 0.3 | -0.3 | -0.3 |
| E | O | 0 | -0.1 | 0.1 | -0.1 | -0.2 | -0.2 | -0.8 | 0.0 | -0.8 | -0.6 | 0.3 | 0.0 | 0.1 | -0.9 | 0.1 | 0.4 | 0.2 | 0.0 | 0.1 | 0.0 | 0.3 |
| E | O | 1 | -0.4 | -0.8 | -0.8 | -1.3 | -0.9 | -0.8 | -0.6 | 0.2 | -0.2 | 0.4 | 0.3 | 0.2 | -1.5 | 0.2 | 0.2 | 0.1 | -0.1 | 0.2 | 0.7 | 0.3 |
| E | O | 2 | -0.9 | -0.7 | -0.8 | -0.9 | -0.6 | 0.2 | -0.4 | 0.5 | -0.2 | 0.5 | 0.3 | 0.2 | -1.1 | -0.1 | 0.1 | 0.1 | 0.0 | 0.2 | 0.1 | 0.4 |
| E | O | 3 | -0.6 | -0.9 | -1.0 | -0.5 | -0.5 | -0.1 | 0.0 | 0.2 | -0.2 | 0.4 | 0.2 | 0.1 | -0.4 | 0.2 | 0.2 | 0.2 | 0.0 | 0.3 | 0.1 | 0.4 |

# position preference parameter of a single amino acid in an inside/outside position of a secondary structure segment

## loop

| S X P | V | L | I | M | F | W | Y | G | A | P | S | T | C | H | R | K | Q | E | N | D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| L B 1 | 0.8 | 0.7 | 0.9 | 0.5 | 0.7 | 0.5 | 0.2 | -0.4 | 0.1 | -0.9 | -0.6 | 0.0 | 0.0 | 0.1 | -0.5 | -0.9 | -0.6 | -1.0 | -1.1 | -1.2 |
| L B 2 | 0.8 | 0.7 | 0.9 | 0.8 | 0.7 | -0.2 | 0.3 | -0.7 | -0.1 | -1.9 | -0.4 | -0.1 | 0.8 | -0.4 | -0.4 | -0.9 | -0.4 | -0.9 | -0.9 | -1.2 |
| L B 3 | 0.4 | 0.8 | 0.6 | 0.2 | 0.5 | 0.4 | 0.5 | -0.1 | -0.1 | -0.7 | -0.2 | -0.2 | 0.8 | -0.2 | -0.6 | -0.9 | -0.6 | -0.8 | -0.7 | -0.5 |
| L B 4 | 0.1 | 0.3 | -0.1 | 0.1 | 0.2 | 0.3 | 0.4 | 0.4 | -0.1 | -0.4 | 0.1 | 0.0 | 0.6 | 0.1 | -0.3 | -1.0 | -0.8 | -0.5 | -0.2 | -0.2 |
| L B 5 | 0.1 | 0.0 | 0.1 | -0.7 | 0.1 | -0.9 | 0.4 | 0.4 | 0.1 | 0.3 | -0.1 | 0.0 | -1.4 | 0.3 | -0.3 | -0.5 | 0.0 | -0.4 | -0.5 | 0.1 |
| L B 6 | -0.3 | 0.1 | -0.1 | 0.3 | 0.4 | 0.2 | 0.3 | -0.1 | -0.6 | 0.0 | -0.2 | -0.3 | 0.4 | 0.4 | 0.2 | -0.3 | 0.0 | -0.1 | 0.0 | 0.3 |
| L B 7 | 0.0 | -0.1 | 0.1 | -0.4 | 0.0 | 0.4 | 0.2 | 0.3 | -0.1 | 0.1 | -0.1 | -0.1 | 0.7 | 0.1 | 0.0 | -0.6 | -0.1 | -0.3 | -0.1 | 0.0 |
| L B 8 | -0.2 | 0.1 | 0.0 | -1.4 | 0.2 | -0.2 | 0.4 | -0.2 | -0.2 | 0.5 | 0.0 | -0.7 | 0.9 | 0.0 | 0.3 | -0.4 | 0.1 | -0.2 | 0.2 | -0.1 |
| L B 9 | 0.2 | 0.3 | 0.5 | 0.2 | 0.2 | 0.6 | -0.3 | 0.3 | -0.1 | 0.2 | -0.2 | -0.1 | 0.6 | -0.5 | -0.5 | -0.7 | -0.8 | -0.5 | -0.1 | 0.1 |
| L B 0 | -0.6 | -0.2 | -0.3 | 0.0 | 0.2 | 0.4 | 0.2 | 0.2 | -0.3 | 0.4 | 0.3 | -0.1 | 0.2 | 0.3 | -0.4 | -0.3 | 0.0 | -0.5 | 0.2 | 0.0 |
| L B 1 | 0.2 | 0.3 | 0.6 | 0.5 | 0.5 | 0.2 | 0.4 | -0.4 | 0.0 | 0.1 | -0.1 | 0.0 | 0.2 | -0.1 | -0.4 | -0.4 | 0.0 | -0.5 | -1.0 | -1.0 |
| L B 2 | 0.9 | 0.5 | 0.9 | 0.7 | 0.6 | 0.5 | 0.3 | -0.4 | 0.0 | -1.0 | -0.4 | -0.1 | 0.8 | -0.4 | -0.8 | -1.3 | -0.7 | -0.8 | -1.0 | -0.9 |
| L B 3 | 0.6 | 0.5 | 0.6 | 0.6 | 0.6 | 0.4 | 0.5 | -0.5 | 0.0 | -1.3 | -0.3 | -0.2 | 0.5 | 0.0 | -0.4 | -0.9 | -0.4 | -0.6 | -0.7 | -0.6 |
| L I 1 | -0.5 | -0.3 | -0.4 | -0.9 | -0.4 | 0.0 | 0.0 | -0.2 | -0.1 | -0.3 | 0.2 | 0.1 | -1.0 | -0.3 | 0.4 | 0.5 | 0.2 | 0.4 | 0.1 | 0.0 |
| L I 2 | -0.1 | -0.7 | -0.7 | 0.1 | -0.7 | 0.1 | -0.1 | -1.5 | -0.1 | -0.8 | 0.2 | 0.2 | 0.0 | 0.2 | 0.6 | 0.6 | 0.5 | 0.4 | -0.1 | -0.5 |
| L I 3 | -0.5 | -0.2 | -0.8 | -0.1 | -0.6 | -1.5 | 0.0 | -0.5 | -0.2 | -0.6 | -0.2 | 0.1 | -0.9 | 0.4 | 0.4 | 0.5 | 0.4 | 0.3 | 0.3 | 0.2 |
| L I 4 | -0.4 | -0.6 | -0.6 | -0.6 | -0.7 | -0.7 | -0.8 | 0.2 | -0.1 | 0.1 | 0.1 | 0.1 | -1.1 | 0.0 | 0.4 | 0.5 | -0.1 | 0.1 | 0.5 | 0.1 |
| L I 5 | -1.0 | -0.5 | -0.9 | -0.8 | -0.9 | 0.1 | -0.6 | 0.4 | -0.3 | 0.7 | 0.1 | 0.1 | -0.7 | -0.4 | 0.4 | 0.4 | 0.3 | 0.2 | 0.1 | 0.3 |
| L I 6 | -0.5 | -0.4 | -0.8 | -0.7 | -2.1 | 0.2 | -0.5 | 0.3 | -0.1 | 0.6 | -0.4 | 0.2 | -0.3 | -0.2 | 0.5 | 0.5 | 0.1 | 0.1 | 0.2 | 0.2 |
| L I 7 | -0.2 | -0.5 | -0.5 | -0.6 | -0.8 | -0.9 | 0.0 | -0.1 | -0.2 | 0.4 | 0.0 | 0.1 | 0.1 | -0.1 | 0.3 | 0.2 | 0.2 | 0.2 | 0.2 | 0.1 |
| L I 8 | -0.2 | -0.5 | -0.8 | -1.5 | -0.3 | -0.2 | -0.9 | 0.0 | -0.2 | 0.0 | 0.1 | -0.2 | -1.4 | 0.5 | 0.6 | 0.2 | -0.2 | 0.1 | 0.4 | 0.6 |
| L I 9 | -0.6 | -0.6 | -1.1 | -0.4 | -0.5 | 0.1 | -0.3 | 0.4 | -0.4 | 0.2 | 0.0 | -0.2 | 0.2 | 0.0 | 0.4 | 0.3 | 0.4 | 0.2 | 0.2 | 0.3 |
| L I 0 | -1.3 | -1.1 | -1.6 | -1.1 | -0.9 | -0.6 | -0.4 | 0.1 | -0.6 | 0.6 | 0.4 | 0.4 | -0.5 | 0.0 | 0.0 | 0.4 | -0.1 | 0.1 | 0.7 | 0.7 |
| L I 1 | -0.1 | -0.4 | -0.6 | 0.2 | -0.1 | -0.2 | -0.1 | -0.9 | 0.0 | 0.2 | 0.2 | 0.2 | 0.1 | -0.5 | 0.3 | 0.4 | 0.3 | 0.3 | -0.3 | -0.4 |
| L I 2 | -0.4 | -0.9 | -0.3 | -0.7 | -0.9 | 0.1 | -0.8 | -0.6 | 0.0 | 0.0 | 0.1 | 0.1 | -1.4 | -0.1 | 0.3 | 0.5 | 0.4 | 0.6 | 0.2 | 0.3 |
| L I 3 | -0.5 | -0.5 | -1.1 | -0.9 | -0.6 | -1.5 | 0.0 | -0.8 | -0.3 | -0.4 | 0.0 | 0.2 | -1.6 | 0.0 | 0.1 | 0.4 | 0.5 | 0.9 | 0.5 | 0.5 |
| L O 1 | -0.6 | -1.3 | -0.9 | -2.2 | -1.9 | -1.7 | -1.8 | -0.1 | 0.2 | 0.5 | 0.2 | -0.1 | -0.8 | -0.4 | 0.2 | 0.4 | 0.2 | 1.1 | 0.4 | 0.4 |
| L O 2 | -0.7 | -0.7 | -1.7 | 0.0 | -0.8 | -1.6 | -1.4 | -0.2 | 0.2 | -0.1 | 0.2 | -0.2 | -1.7 | -0.1 | 0.0 | 0.6 | 0.4 | 0.7 | 0.5 | 0.4 |
| L O 3 | -0.7 | -0.7 | -1.2 | -1.1 | -0.6 | -1.0 | -0.4 | -0.8 | 0.2 | -0.8 | 0.1 | -0.1 | -1.5 | -0.5 | 0.5 | 0.7 | 0.5 | 0.5 | 0.6 | 0.4 |
| L O 4 | -0.9 | -0.9 | -1.5 | -0.3 | -0.9 | -1.6 | -1.5 | 0.7 | 0.2 | -0.2 | 0.2 | -0.2 | -0.8 | 0.1 | 0.0 | 0.2 | 0.3 | 0.3 | 0.8 | 0.3 |
| L O 5 | -1.1 | -1.1 | -1.1 | -1.1 | -0.9 | 0.0 | -1.0 | 0.6 | 0.0 | 0.7 | 0.2 | 0.0 | -2.2 | -0.2 | 0.2 | 0.5 | 0.2 | 0.2 | 0.4 | 0.3 |
| L O 6 | -0.8 | -0.9 | -1.0 | -1.1 | -1.0 | 0.1 | 0.1 | 0.5 | -0.1 | 0.4 | 0.3 | 0.0 | -0.8 | -0.1 | -0.1 | 0.5 | -0.2 | 0.2 | 0.3 | 0.3 |
| L O 7 | -0.8 | -0.7 | -0.9 | -0.9 | -0.9 | 0.2 | -0.3 | 0.1 | -0.1 | 0.7 | 0.2 | -0.1 | -1.1 | -0.3 | 0.3 | 0.4 | 0.1 | 0.2 | 0.2 | 0.4 |
| L O 8 | -1.4 | -1.1 | -1.7 | -1.3 | -1.5 | -1.0 | -0.4 | 0.3 | -0.1 | 0.7 | 0.0 | -0.2 | -1.5 | 0.5 | 0.1 | 0.7 | 0.1 | 0.2 | 0.6 | 0.6 |
| L O 9 | -1.2 | -1.0 | -1.5 | -0.9 | -0.9 | 0.0 | -0.7 | 0.4 | -0.2 | 0.6 | 0.1 | -0.1 | -2.5 | 0.4 | 0.2 | 0.4 | 0.2 | 0.3 | 0.4 | 0.6 |
| L O 0 | -1.6 | -0.9 | -1.3 | -0.4 | -0.6 | -1.7 | -1.1 | 0.8 | -0.4 | 0.0 | 0.5 | 0.0 | -1.1 | -0.5 | 0.2 | 0.3 | 0.0 | 0.3 | 0.7 | 0.4 |
| L O 1 | -0.5 | -0.8 | -0.7 | -0.3 | -1.0 | -0.8 | -1.1 | -0.4 | 0.1 | 1.4 | 0.2 | 0.1 | -1.6 | -0.3 | -0.2 | 0.4 | 0.4 | 0.5 | -0.2 | 0.2 |
| L O 2 | -0.7 | -1.4 | -1.6 | -0.6 | -1.3 | -0.9 | -1.7 | -0.3 | 0.4 | 0.4 | 0.0 | 0.1 | -1.2 | -0.2 | -0.3 | 0.0 | 0.2 | 1.1 | 0.3 | 0.9 |
| L O 3 | -1.1 | -1.8 | -1.6 | -1.0 | -0.9 | -1.0 | -0.4 | 0.2 | 0.3 | 0.6 | 0.3 | -0.1 | -0.8 | -0.2 | -0.4 | 0.3 | -0.2 | 0.5 | 0.5 | 0.8 |

# position preference parameter of a single amino acid in an inside/outside position of a secondary structure segment

## α–helix

| S | X | P | V | L | I | M | F | W | Y | G | A | P | S | T | C | H | R | K | Q | E | N | D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| H | B | 1 | 0.5 | 0.5 | 0.6 | 0.6 | 1.0 | -0.2 | 1.1 | 0.1 | 0.1 | 0.0 | -0.3 | -0.5 | 0.8 | -0.5 | -0.9 | -1.0 | -1.7 | -1.4 | -0.7 | -1.7 |
| H | B | 2 | 0.8 | 0.9 | 0.9 | 0.5 | 0.6 | 0.2 | 0.0 | -0.3 | -0.1 | -0.3 | -0.2 | 0.3 | 0.2 | -0.5 | -1.3 | -1.8 | -0.9 | -1.2 | -1.3 | -0.8 |
| H | B | 3 | -0.9 | -0.4 | -0.9 | -0.2 | -0.2 | -0.9 | 0.0 | 0.5 | 0.1 | 0.4 | 0.6 | -0.3 | 0.8 | 0.7 | -0.6 | -1.0 | -1.0 | -0.5 | 0.6 | 0.2 |
| H | B | 4 | 0.4 | 0.6 | 0.5 | -0.1 | 0.2 | 0.8 | 0.1 | -0.2 | 0.1 | 0.7 | -0.2 | 0.0 | 0.5 | -0.3 | -0.8 | -0.4 | -0.5 | -0.4 | -1.1 | -1.3 |
| H | B | 5 | 0.1 | -0.4 | 0.3 | 0.3 | 0.3 | 0.8 | 0.0 | 0.3 | 0.0 | -0.7 | 0.0 | 0.0 | 1.9 | 0.1 | -1.3 | -0.7 | -0.7 | -0.3 | 0.0 | -0.6 |
| H | B | 6 | 0.4 | 0.1 | 0.1 | 0.9 | 0.3 | -0.2 | 0.1 | -0.9 | -0.1 | -1.0 | -0.4 | -0.3 | 0.1 | -0.2 | -1.0 | -1.0 | 0.1 | 0.4 | -0.2 | 0.7 |
| H | B | 7 | 0.6 | 0.8 | 0.8 | 0.9 | 0.6 | 0.3 | -0.2 | -0.6 | 0.6 | -2.4 | -0.6 | -0.3 | 0.7 | -0.3 | -0.7 | -0.7 | -0.7 | -0.6 | -0.9 | -1.4 |
| H | B | 8 | 0.7 | 1.2 | 1.0 | 1.2 | 0.7 | 0.8 | -0.2 | -0.9 | 0.4 | -2.4 | -0.8 | -0.3 | -0.7 | -0.1 | -0.7 | -0.6 | -1.2 | -1.8 | -0.7 | -3.0 |
| H | B | 9 | 0.6 | 1.0 | 1.0 | 1.1 | 0.7 | -0.3 | -0.3 | -0.7 | 0.3 | -2.1 | -0.3 | -0.2 | 1.1 | -1.1 | -1.0 | -1.1 | -1.4 | -1.0 | -0.8 | -1.6 |
| H | B | 0 | 0.3 | 0.7 | 0.4 | 0.2 | 0.8 | 1.1 | 0.8 | -0.3 | 0.5 | -1.7 | -0.5 | -0.7 | 1.8 | 0.1 | -1.4 | -1.2 | -0.6 | -1.3 | -0.6 | -1.2 |
| H | B | 1 | 0.1 | 0.6 | 0.0 | 0.6 | 0.8 | -0.2 | 0.7 | 0.5 | -0.4 | -1.7 | -0.2 | -0.3 | 1.1 | 0.2 | -0.3 | -1.2 | -0.6 | -1.1 | -0.3 | -0.3 |
| H | B | 2 | 0.0 | 0.3 | 0.2 | 0.3 | 0.6 | -0.4 | 0.3 | 0.0 | 0.2 | -0.2 | -0.1 | -0.1 | 0.3 | -0.4 | -0.2 | -0.4 | -0.2 | -0.9 | -0.2 | -0.2 |
| H | B | 3 | 0.7 | 0.7 | 0.5 | 0.8 | 0.4 | 0.9 | 0.1 | -0.5 | -0.3 | -0.1 | -0.8 | -0.4 | 1.3 | 0.0 | -0.4 | -0.5 | -0.7 | -0.9 | -0.4 | -0.6 |
| H | I | 1 | -0.3 | -0.6 | -0.6 | -0.3 | -0.6 | -2.6 | -0.8 | 0.1 | -0.1 | -0.6 | -0.1 | 0.1 | 0.6 | -0.1 | 0.3 | 0.7 | 0.5 | 0.2 | 0.2 | 0.0 |
| H | I | 2 | -0.3 | -0.1 | -0.6 | 0.3 | -0.7 | -0.7 | -0.1 | -0.1 | 0.4 | 0.5 | -0.3 | -0.1 | 0.1 | 0.0 | 0.5 | 0.3 | -0.1 | -0.1 | -0.4 | -0.1 |
| H | I | 3 | -2.2 | -1.2 | -2.4 | -1.6 | -1.4 | -1.1 | -1.0 | -0.1 | -0.7 | 0.7 | 0.7 | 0.8 | 0.1 | -0.2 | -0.1 | 0.0 | 0.0 | 0.1 | 0.7 | 0.8 |
| H | I | 4 | -0.1 | -0.5 | -0.6 | -0.4 | -0.3 | -0.2 | -0.3 | -0.7 | 0.2 | 1.2 | 0.2 | -0.2 | 0.0 | -0.6 | -0.2 | 0.3 | -0.1 | 0.2 | -0.1 | 0.2 |
| H | I | 5 | -0.9 | -1.9 | -0.7 | -0.9 | -1.5 | 0.0 | -0.6 | 0.2 | 0.4 | 0.3 | 0.3 | 0.0 | -0.2 | 0.1 | -0.4 | 0.2 | 0.3 | 0.7 | -0.1 | 0.5 |
| H | I | 6 | -1.0 | -0.7 | -1.3 | -0.3 | -0.7 | -0.2 | -0.1 | -2.1 | -0.5 | 0.1 | -1.0 | -0.2 | -1.7 | 0.2 | -0.5 | 0.3 | 1.1 | 1.4 | 0.3 | 0.9 |
| H | I | 7 | -0.5 | -0.3 | -0.6 | -0.5 | -0.9 | -1.5 | -0.7 | -0.5 | 0.2 | -1.8 | -0.5 | -0.4 | -0.8 | -0.2 | 0.7 | 0.9 | 0.6 | 0.8 | 0.2 | 0.2 |
| H | I | 8 | -0.1 | 0.3 | 0.0 | 0.2 | -0.8 | 0.4 | -1.0 | -0.5 | -0.1 | -1.4 | -0.8 | -0.9 | -0.3 | 1.0 | 0.8 | 0.7 | 0.4 | 0.5 | -0.6 | -0.3 |
| H | I | 9 | 0.3 | 0.1 | 0.2 | 0.3 | 0.2 | -0.2 | 0.4 | -1.4 | -0.2 | -3.6 | -0.6 | -0.3 | -0.6 | 0.0 | 0.5 | 0.6 | 0.4 | 0.1 | -0.1 | -0.3 |
| H | I | 0 | -0.6 | -0.1 | -1.1 | 0.0 | -0.1 | -0.7 | 0.8 | -0.3 | 0.3 | -1.8 | -0.3 | -0.3 | -0.6 | 0.4 | 0.4 | 0.4 | 0.6 | -0.1 | -0.2 | 0.0 |
| H | I | 1 | -0.8 | -0.2 | -0.7 | -0.4 | -0.3 | 0.0 | 0.4 | 0.5 | -0.4 | -0.7 | -0.1 | -0.4 | -0.1 | -0.1 | 0.5 | 0.4 | 0.5 | 0.2 | 0.1 | -0.1 |
| H | I | 2 | -0.8 | -0.3 | -0.6 | -1.1 | -0.6 | -1.6 | 0.0 | 0.5 | -0.1 | 0.2 | 0.1 | -0.1 | 0.0 | 0.3 | 0.1 | 0.6 | 0.0 | -0.2 | 0.4 | -0.2 |
| H | I | 3 | -0.3 | 0.0 | -0.6 | -0.5 | -0.9 | 0.0 | -0.2 | -0.2 | -0.2 | 0.6 | -0.1 | 0.0 | -0.5 | 0.0 | 0.1 | 0.7 | -0.1 | 0.1 | 0.0 | 0.0 |
| H | O | 1 | -0.7 | -1.2 | -1.5 | -0.8 | -1.3 | -2.9 | -1.3 | 0.6 | 0.1 | 0.3 | 0.1 | -0.1 | -1.3 | 0.3 | 0.0 | 0.8 | 0.2 | 0.3 | 0.4 | 0.1 |
| H | O | 2 | -1.1 | -0.5 | -1.3 | -0.5 | -0.7 | -1.6 | -1.2 | 0.1 | 0.3 | 0.4 | 0.3 | 0.1 | -1.0 | 0.2 | 0.3 | 0.6 | -0.1 | -0.5 | 0.4 | 0.5 |
| H | O | 3 | -3.0 | -0.9 | -2.0 | -1.2 | -1.6 | -1.7 | -3.2 | -0.5 | -0.2 | 0.9 | 0.8 | 0.4 | -0.6 | -0.2 | 0.0 | 0.4 | -0.4 | 0.2 | 0.8 | 1.0 |
| H | O | 4 | -0.8 | -0.8 | -1.1 | -0.5 | -0.9 | -1.3 | -0.8 | -0.1 | 0.2 | 1.5 | 0.2 | -0.3 | -1.1 | -0.7 | -0.3 | 0.2 | 0.1 | 0.6 | -0.1 | 0.5 |
| H | O | 5 | -0.9 | -1.6 | -2.4 | -0.8 | -1.4 | -0.6 | -2.1 | -0.1 | 0.4 | 0.2 | 0.1 | -0.1 | -1.1 | -0.3 | -0.2 | 0.1 | 0.3 | 1.1 | 0.3 | 1.0 |
| H | O | 6 | 0.1 | -2.5 | -0.9 | -1.4 | -0.4 | -1.3 | -0.4 | -0.7 | 0.5 | 0.1 | 0.0 | 0.0 | -1.8 | -0.2 | -0.4 | -0.1 | 0.2 | 1.0 | 0.4 | 0.8 |
| H | O | 7 | -0.8 | -0.9 | -1.1 | -0.7 | -1.1 | -0.9 | -1.1 | -0.4 | 0.3 | -0.8 | 0.0 | 0.1 | -0.9 | 0.4 | 0.4 | 0.8 | 0.5 | 0.6 | 0.3 | 0.2 |
| H | O | 8 | -1.1 | -1.2 | -0.4 | -2.1 | -1.5 | -2.0 | -1.8 | -0.4 | 0.2 | -1.5 | 0.0 | 0.0 | -0.9 | 0.3 | 0.5 | 1.0 | 0.3 | 0.7 | 0.5 | 0.6 |
| H | O | 9 | -0.7 | -0.6 | -1.2 | 0.1 | -1.0 | -1.0 | -1.1 | -0.6 | 0.2 | 0.0 | 0.0 | -0.2 | -1.4 | 0.4 | -0.1 | 0.8 | 0.5 | 0.7 | 0.3 | 0.1 |
| H | O | 0 | -0.9 | -0.8 | -1.9 | -0.9 | -0.7 | -2.2 | -0.4 | -0.5 | 0.2 | -2.6 | 0.0 | 0.0 | -1.0 | 0.0 | 0.4 | 0.8 | 0.6 | 0.6 | 0.8 | 0.3 |
| H | O | 1 | -1.1 | -0.6 | -1.2 | -0.2 | -0.6 | -0.1 | -0.9 | 0.8 | 0.0 | -0.4 | 0.1 | -0.3 | -1.4 | -0.1 | 0.0 | 0.4 | 0.0 | 0.2 | 0.9 | 0.0 |
| H | O | 2 | -1.2 | -1.1 | -1.2 | -1.0 | -1.0 | -1.0 | -1.2 | 0.8 | 0.0 | 0.9 | 0.1 | -0.3 | -1.5 | 0.0 | -0.1 | 0.7 | 0.2 | -0.1 | 0.4 | 0.2 |
| H | O | 3 | -1.0 | -1.0 | -1.5 | -0.4 | -1.6 | -2.3 | -0.8 | 0.3 | -0.1 | 0.6 | -0.1 | 0.3 | -2.2 | 0.2 | 0.0 | 0.6 | 0.4 | 0.3 | 0.5 | 0.5 |

# Preference Parameters: Residues in contact interfaces

## How where the preferences derived?

To treat protein-water contacts in the same way as protein-protein contacts it was necessary to define protein-water contacts so that they are comparable to protein-protein contacts. Based on the conservation of the total number of atomic contacts a simple empirical formula has been derived (Francois Colonna-Cesari and Chris Sander, Biophys.J Vol 57, pp 1103-1107 (1990)) to estimate the number of solvent contacts from the residue solvent accessibility

$$C_{iw} = 0.31S$$

where $C_{iw}$ is the number of solvent contacts a residue $i$ makes and $S$ is the accessible surface in square Angstrom.

The contacts strength $C_{ij}$ between atom $i$ and $j$ is 1.0 if the atoms overlap or just touch, i.e if the atom-atom distance $d_{ij} \leq r_i + r_j$ (where $r_i$ and $r_j$ are hard sphere or van-der Waals radii); the contact strength decreases linearly down to zero with distance until a water molecule can just fit between the two atoms :

$$C_{ij} = \begin{cases} 1 & d_{ij} \leq r_i + r_j \\ (d_{ij} - r_i - r_j)/2r_{H_2O} & 2r \leq d_{ij} < r_i + r_j + 2r_{H_2O} \\ 0 & d_{ij} > r_i + r_j + 2r_{H_2O} \end{cases}$$

All atoms (side chain and backbone atoms) where used in the calculation of the contact strength between two residues.

### Preference calculation

The preferences where calculated as follows:

- For each observed contact $C_{ij}$ (i.e. residue-residue or residue-water) find its contact type $t$ and the aminoacid $a$ of the contact partner $i$. Add the contact strength to the *contact_statistic* matrix $s_{at}$ and add 1 to the *contact_occurrences* matrix $o_{at}$.
- Scale $s_{at}$ so that the sum $\sum_{at} s_{at} = 1$
- Build the marginal sums $m_a = \sum_t s_{at}$ and $m_t = \sum_a s_{at}$
- The preference $p_{at}$ for aminoacid $a$ being involved in a contact of type $t$ is

$$p_{ai} = \log_2 \left( \frac{s_{at}}{m_a m_t} \right)$$

## Contact Types

A contact type classifies a contact of an amino acid with another amino acid or water molecule by the following features:

- the residue type of the contact partner (i.e. one of the 20 amino acids or water)
- the secondary structure of the partner
- the secondary structure of the actual residue
- the distance of the two contact partners in the sequence

### AM: Contacts of amino acids with protein or water

20 × 2 Preferences.

### AS: Contacts of amino acids with a secondary structure element or water

20 × 5 Preferences. Here is a list of a partner classes:

- **H** = Helix : H,G or I (taken from the DSSP summary column)
- **E** = Beta Strand: E
- **T** = Turn: T
- **X** = Other: ' ', S or B
- **%** = Water

### AA: Contacts of amino acids with an other aminoacid or water

20 × 21 Preferences.

### AInt: Contacts of amino acids in a secondary structure element with another secondary structure element or water

20 × 29 Preferences. The name of the preferences consists of tree parts:
The *first* letter describes the secondary structure element in which the actual amino acid is located:

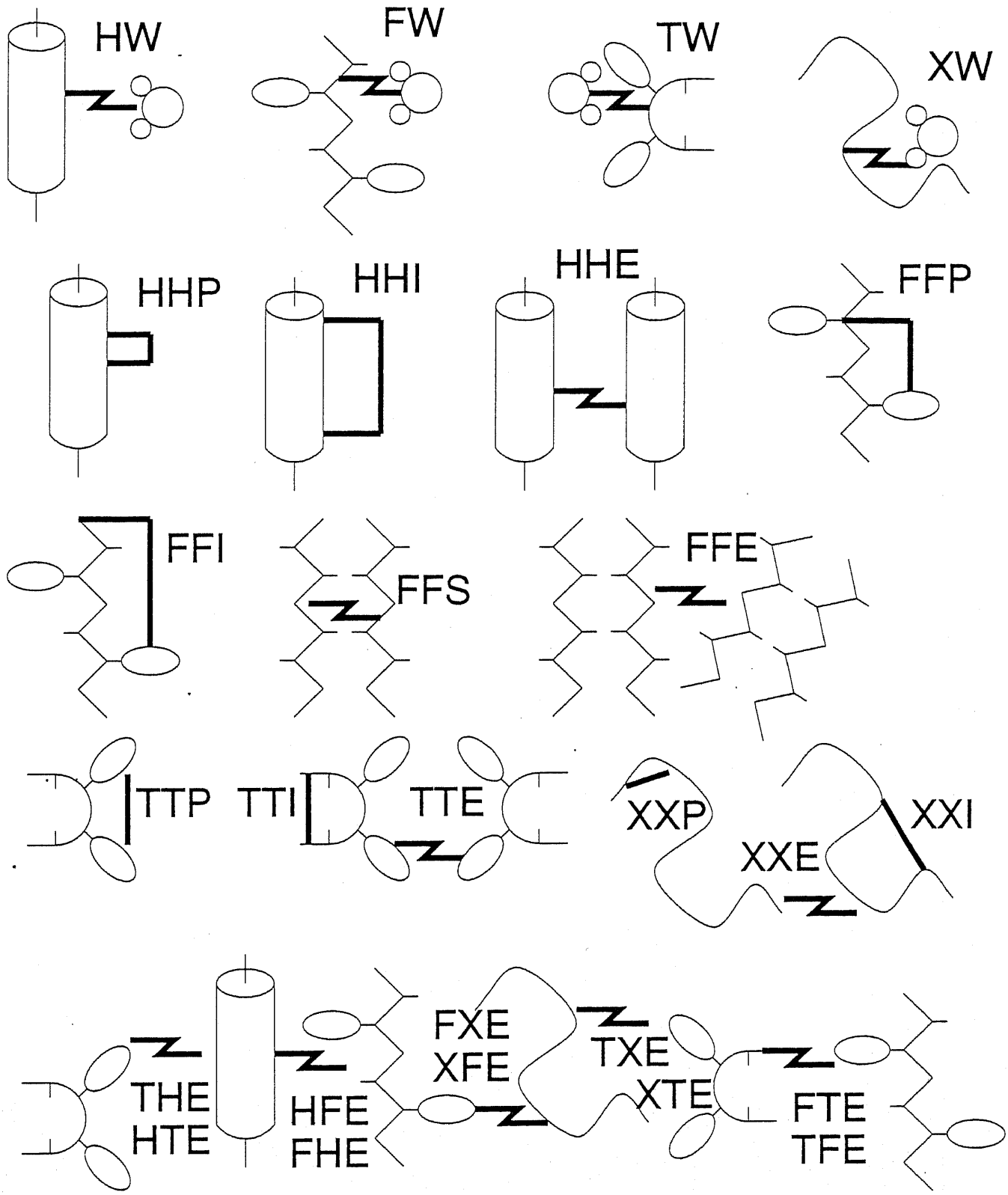
- **H** = H,G or I
- **F** = E
- **T** = T
- **X** = ' ', S or B

The *second* letter describes the secondary structure element in which the contact partner is located.

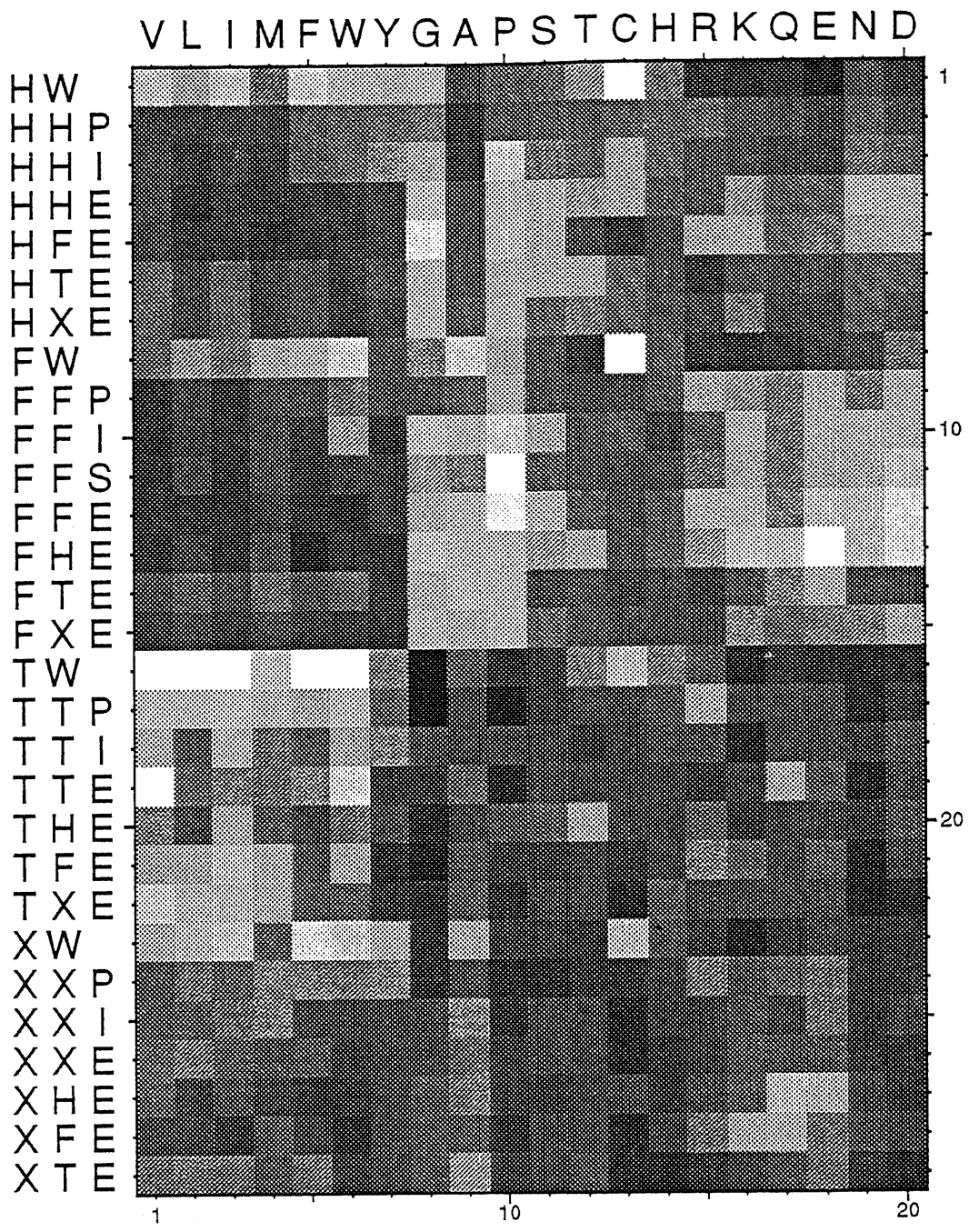
- **H** = H,G or I
- **F** = E
- **T** = T
- **X** = ' ', S or B
- **W** = Water

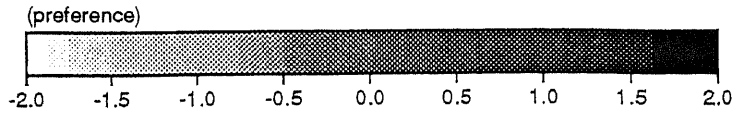
The *third* letter describes the relation of the two contacting partners in the sequence:

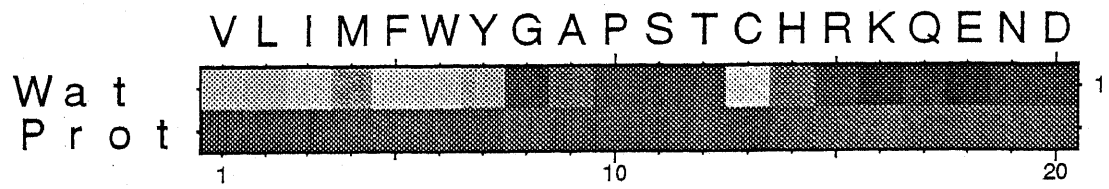
- **P** = The two contacting amino acids are just neighbors in the sequence. (Proximate)
- **I** = The two amino acids belong to the same secondary structure element. (Internal)
- **S** = Both partners are in two different extended strands: The two strands belong to the *same sheet.*
- **E** = The two amino acids belong to different secondary structure elements. In the case of two contacting strands this means the strands belong to different sheets. (External)

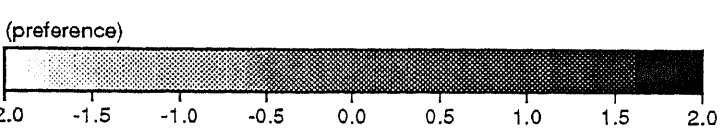

HW
FW
TW
XW
HHP
HHI
HHE
FFP
FFI
FFS
FFE
TTP
TTI
TTE
XXP
XXE
XXI
THE
HTE
HFE
FHE
FXE
XFE
TXE
XTE
FTE
TFE

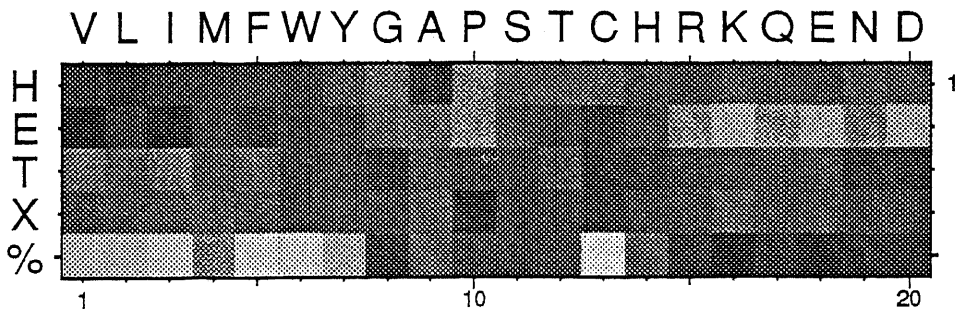
| #PID | C | SIZ | RES | %H | %B | SID | PROTEIN_NAME |
|---|---|---|---|---|---|---|---|
| 351C | | 82 | 1.6 | 50 | 4 | C551$PSEAE | CYTOCHROME $C=551= (OXIDIZED) |
| 256B | A | 106 | 1.4 | 79 | 0 | C562$ECOLI | CYTROCHROME $B562 (OXIDIZED) |
| 8ADH | | 374 | 2.4 | 28 | 24 | ADHE$HORSE | APO-LIVER ALCOHOL DEHYDROGENASE |
| 8ATC | A | 310 | 2.5 | 40 | 15 | PYRI$ECOLI | ASPARTATE CARBAMOYLTRANSFERASE (ASPARTATE TRANSCARBAMYLASE) (R STATE) COMPLEX WITH N-PHOSPHONACETYL-L-ASPARTATE (/PALA$) |
| 8ATC | B | 146 | 2.5 | 15 | 34 | PYRI$ECOLI | ASPARTATE CARBAMOYLTRANSFERASE (ASPARTATE TRANSCARBAMYLASE) (R STATE) COMPLEX WITH N-PHOSPHONACETYL-L-ASPARTATE (/PALA$) |
| 2AZA | A | 129 | 1.8 | 16 | 35 | AZUR$ALCDE | AZURIN (OXIDIZED) |
| 3B5C | | 85 | 1.5 | 31 | 23 | CYB5$BOVIN | CYTOCHROME $B5 (OXIDIZED) |
| 3BLM | | 257 | 2.0 | 42 | 17 | BLAC$STAAU | BETA-*LACTAMASE |
| 2CA2 | | 256 | 1.9 | 16 | 30 | CAH2$HUMAN | CARBONIC ANHYDRASE /II$ (CARBONATE DEHYDRATASE) (/HCA II$) COMPLEX WITH THIOCYANATE ION |
| 1CCR | | 111 | 1.5 | 42 | 1 | CYC$ORYSA | CYTOCHROME $C |
| 2CCY | A | 127 | 1.7 | 74 | 1 | CYCP$RHOMO | CYTOCHROME $C(PRIME) |
| 1CD4 | | 173 | 2.3 | 5 | 41 | CD4$HUMAN* | /CD4$ (1 - 183 PLUS ASP - THR) (/D1D2$) (N-TERMINAL FRAGMENT OF /CD4$ PLUS TWO MIS-SENSE RESIDUES (ASP - THR)) |
| 3CLA | | 213 | 1.8 | 29 | 28 | CAT3$ECOLI | TYPE /III$ CHLORAMPHENICOL ACETYLTRANSFERASE (/CAT=III=$) COMPLEX WITH CHLORAMPHENICOL |
| 5CPA | | 307 | 1.5 | 38 | 16 | CBPA$BOVIN | CARBOXYPEPTIDASE A=ALPHA= (COX) |
| 2CPP | | 405 | 1.6 | 51 | 10 | CPXA$PSEPU | CYTOCHROME P450CAM (CAMPHOR MONOOXYGENASE) WITH BOUND CAMPHOR |
| 4CPV | | 108 | 1.5 | 56 | 1 | PRVB$CYPCA* | CALCIUM-BINDING PARVALBUMIN ($P*I=4.25) |
| 1CSE | E | 274 | 1.2 | 30 | 20 | SUBT$BACLI | SUBTILISIN CARLSBERG (COMMERCIAL PRODUCT FROM SERRA, HEIDELBERG CALLED SUBTILISIN NAGARSE) COMPLEX WITH EGLIN-C |
| 1CSE | I | 63 | 1.2 | 22 | 33 | SUBT$BACLI | SUBTILISIN CARLSBERG (COMMERCIAL PRODUCT FROM SERRA, HEIDELBERG CALLED SUBTILISIN NAGARSE) COMPLEX WITH EGLIN-C |
| 1CTF | | 68 | 1.7 | 55 | 26 | RL7$ECOLI | L7(SLASH)*L12 50 S RIBOSOMAL PROTEIN (C-TERMINAL DOMAIN) |
| 2CYP | | 293 | 1.7 | 50 | 7 | CCPR$YEAST | CYTOCHROME $C PEROXIDASE (FERROCYTOCHROME $C (COLON) H2*O2 REDUCTASE) |
| 8DFR | | 186 | 1.7 | 23 | 33 | DYR$CHICK | DIHYDROFOLATE REDUCTASE |
| 1ECN | | 136 | 1.4 | 75 | 0 | GLB3$CHITH* | HEMOGLOBIN (ERYTHROCRUORIN, CYANO MET) |
| 2ER7 | E | 330 | 1.6 | 11 | 45 | | ENDOTHIA ASPARTIC PROTEINASE (ENDOTHIAPEPSIN) COMPLEX WITH H-261 |
| 4FD1 | | 106 | 1.9 | 33 | 14 | FER1$AZOVI* | FERREDOXIN |
| 4FXN | | 138 | 1.8 | 36 | 22 | FLAV$CLOSP | FLAVODOXIN (SEMIQUINONE FORM) |
| 3GAP | A | 208 | 2.5 | 30 | 14 | CRP$ECOLI | CATABOLITE GENE ACTIVATOR PROTEIN - CYCLIC /AMP$ COMPLEX (/CAP$) |
| 2GBP | | 309 | 1.9 | 43 | 19 | DGAL$ECOLI* | D-*GALACTOSE/D-*GLUCOSE BINDING PROTEIN (/GGBP$) |
| 1GCR | | 174 | 1.6 | 7 | 46 | CRGB$BOVIN | GAMMA-/II$ CRYSTALLIN |
| 1GD1 | O | 334 | 1.8 | 29 | 29 | G3P$BACST | $HOLO-*D-*GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE |
| 1GOX | | 350 | 2.0 | 44 | 13 | 2HAO$SPIOL | GLYCOLATE OXIDASE |
| 1GP1 | A | 183 | 2.0 | 32 | 18 | GSHP$BOVIN | GLUTATHIONE PEROXIDASE |
| 2HLA | B | 99 | 2.6 | 0 | 49 | B2MG$HUMAN | HUMAN CLASS I HISTOCOMPATIBILITY ANTIGEN AW 68.1 (/HLA-AW 68.1$, HUMAN LEUCOCYTE ANTIGEN) |
| 1HOE | | 74 | 2.0 | 0 | 48 | IAA$STRTE* | ALPHA-*AMYLASE INHIBITOR HOE-467*A |
| 1I1B | | 151 | 2.0 | 5 | 47 | IL1B$HUMAN | INTERLEUKIN-1*BETA (/IL$-1*BETA) |
| 4ICD | | 414 | 2.5 | 39 | 18 | IDH$ECOLI* | PHOSPHORYLATED ISOCITRATE DEHYDROGENASE |
| 1IL8 | A | 71 | NMR | 26 | 25 | IL8$HUMAN* | INTERLEUKIN 8 (IL-8) (NEUTROPHIL ACTIVATION PROTEIN) /NAP$ (/NMR$, MINIMIZED MEAN STRUCTURE) |
| 1L13 | | 164 | 1.7 | 64 | 9 | LYCV$BPT4* | LYSOZYME (MUTANT WITH THR 157 REPLACED BY ARG) (T157R) |
| 6LDH | | 329 | 2.0 | 43 | 17 | LDHM$SQUAC* | M=4= APO-*LACTATE DEHYDROGENASE |
| 2LIV | | 344 | 2.4 | 44 | 19 | LIVJ$ECOLI | LEUCINE(SLASH)*ISOLEUCINE(SLASH)*VALINE-BINDING PROTEIN (/LIVBP$) |
| 2LTN | A | 181 | 1.7 | 1 | 43 | LEC$PEA | PEA LECTIN |
| 2LTN | B | 47 | 1.7 | 8 | 63 | LEC$PEA | PEA LECTIN |
| 1LZ1 | | 130 | 1.5 | 39 | 12 | LYC$HUMAN | LYSOZYME |
| 1MBD | | 153 | 1.4 | 77 | 0 | MYG$PHYCA* | MYOGLOBIN (DEOXY, $P*H 8.4) |
| 2MHR | | 118 | 1.7 | 70 | 0 | HEMM$THEZO | MYOHEMERYTHRIN |
| 2PAB | A | 114 | 1.8 | 7 | 51 | TTHY$HUMAN | PREALBUMIN (HUMAN PLASMA) |
| 1PAZ | | 120 | 1.6 | 16 | 37 | AZUP$ALCFA | PSEUDOAZURIN (OXIDIZED CU ++ AT $P*H 6.8) |
| 4PTP | | 223 | 1.3 | 10 | 34 | TRYP$BOVIN* | BETA TRYPSIN, DIISOPROPYLPHOSPHORYL INHIBITED |
| 1R69 | | 63 | 2.0 | 63 | 0 | RPC1$BP434 | 434 REPRESSOR (AMINO-TERMINAL DOMAIN) (R1-69) |
| 1RHD | | 293 | 2.5 | 29 | 13 | THTR$BOVIN | RHODANESE |
| 7RSA | | 124 | 1.3 | 20 | 35 | RNP$BOVIN* | RIBONUCLEASE A (PHOSPHATE-FREE) |
| 2RSP | A | 115 | 2.0 | 5 | 41 | GAG$RSVP | ROUS SARCOMA VIRUS PROTEASE (/RSV PR$) |
| 5RXN | | 54 | 1.2 | 16 | 22 | RUBR$CLOPA | RUBREDOXIN (OXIDIZED, FE(/III$)) (CONSTRAINED MODEL) |
| 2SGA | | 181 | 1.5 | 9 | 55 | PRTA$STRGR | PROTEINASE A (COMPONENT OF THE EXTRACELLULAR FILTRATE PRONASE) (/SGPA$) |
| 4SGB | I | 51 | 2.1 | 0 | 29 | PRTB$STRGR | SERINE PROTEINASE B COMPLEX WITH THE POTATO INHIBITOR /PCI-1$ |
| 2SNS | | 141 | 1.5 | 20 | 22 | NUC$STAAU | STAPHYLOCOCCAL NUCLEASE COMPLEX WITH 2(PRIME)-DEOXY-3(PRIME)-5(PRIME)-DIPHOSPHOTHYMIDINE |
| 2SOD | O | 151 | 2.0 | 1 | 42 | SODC$BOVIN | CU,ZN SUPEROXIDE DISMUTASE |
| 2SSI | | 107 | 2.6 | 15 | 28 | ISUB$STRAO* | STREPTOMYCES SUBTILISIN INHIBITOR |
| 2STV | | 184 | 2.5 | 11 | 47 | COAT$STNV | SATELLITE TOBACCO NECROSIS VIRUS |
| 2TMN | E | 316 | 1.6 | 40 | 17 | THER$BACTH | THERMOLYSIN COMPLEX WITH N-PHOSPHORYL-*L-LEUCINAMIDE (P-*LEU-/NH2$) |
| 1TNF | A | 152 | 2.6 | 1 | 44 | TNFA$HUMAN | TUMOR NECROSIS FACTOR-ALPHA (CACHECTIN) |
| 2TS1 | | 317 | 2.3 | 54 | 10 | SYY$BACST | TYROSYL-TRANSFER /RNA$ SYNTHETASE |
| 1UBQ | | 76 | 1.8 | 23 | 34 | UBIQ$HUMAN | UBIQUITIN |
| 1UTG | | 70 | 1.3 | 75 | 0 | UTER$RABIT | UTEROGLOBIN (OXIDIZED) |
| 2WRP | R | 104 | 1.6 | 78 | 0 | TRPR$ECOLI | $TRP REPRESSOR (ORTHORHOMBIC FORM) |
| 1WSY | A | 248 | 2.5 | 50 | 13 | TRPA$SALTY | TRYPTOPHAN SYNTHASE |
| 4XIA | A | 393 | 2.3 | 47 | 10 | XYLA$ARTS7 | D-*XYLOSE ISOMERASE , D-*SORBITOL COMPLEX |
| 1YPI | A | 247 | 1.9 | 43 | 17 | TPIS$YEAST* | TRIOSE PHOSPHATE ISOMERASE (/TIM$) |

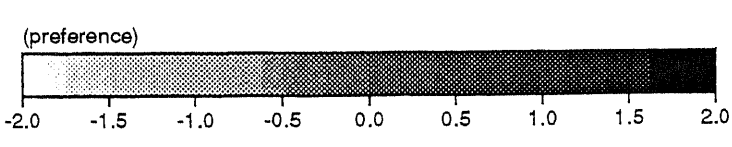

V L I M F W Y G A P S T C H R K Q E N D
H W
H H P
H H I
H H E
H F E
H T E
H X E
F W
F F P
F F I
F F S
F F E
F H E
F T E
F X E
T W
T T P
T T I
T T E
T H E
T F E
T X E
X W
X X P
X X I
X X E
X H E
X F E
X T E
1
10
20
1
10
20

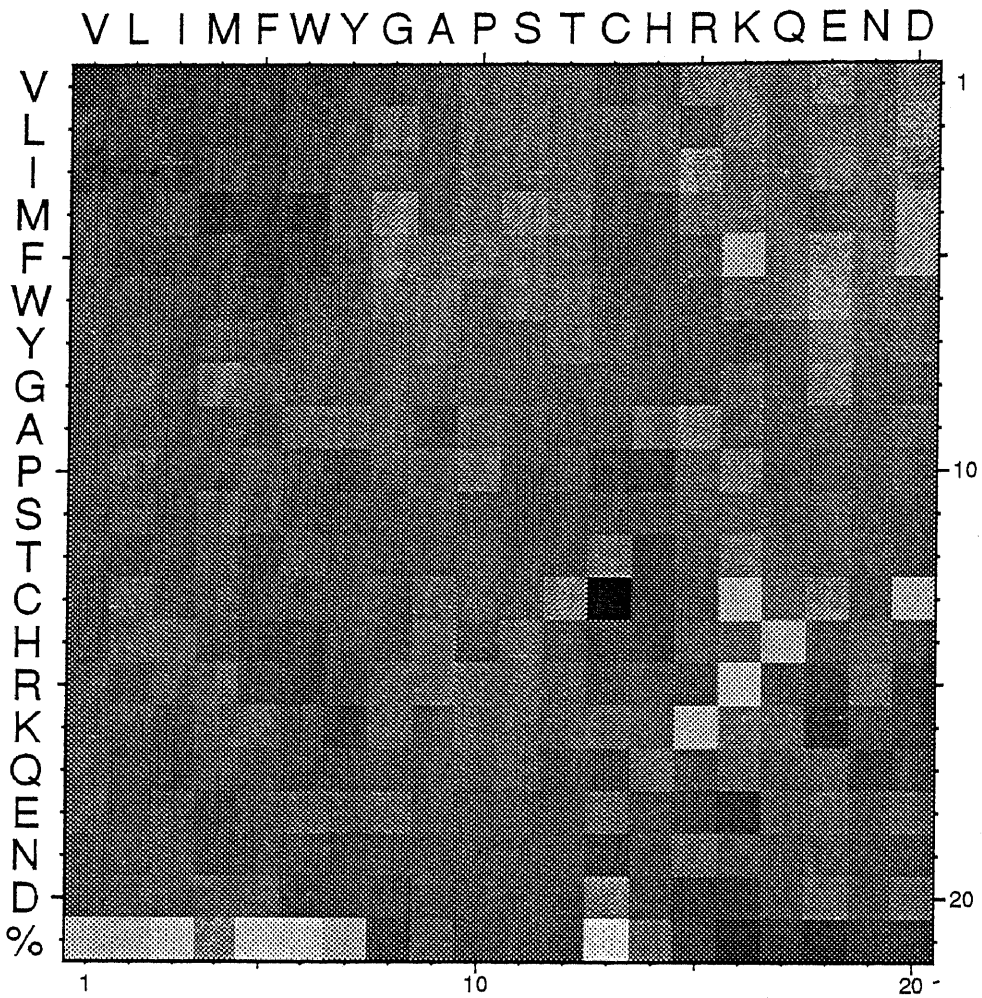

(preference)
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0

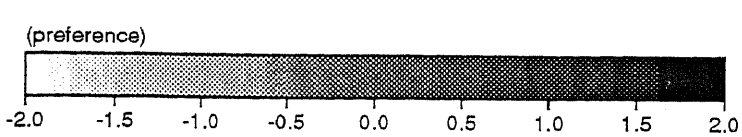

V L I M F W Y G A P S T C H R K Q E N D
Wat
Prot
1
1
10
20


(preference)
2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0


V L I M F W Y G A P S T C H R K Q E N D
H
E
T
X
%
1
1
10
20


(preference)
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0


V L I M F W Y G A P S T C H R K Q E N D
V
L
I
M
F
W
Y
G
A
P
S
T
C
H
R
K
Q
E
N
D
%
1
10
20
1
10
20


(preference)
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0

Contact Preferences AM:

| | V | L | I | M | F | W | Y | G | A | P | S | T | C | H | R | K | Q | E | N | D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Wat | -1.19 | -1.09 | -1.45 | -0.62 | -1.57 | -1.48 | -0.73 | 0.42 | -0.33 | 0.27 | 0.17 | 0.03 | -1.72 | -0.33 | 0.50 | 1.07 | 0.59 | 0.77 | 0.35 | 0.48 |
| Prot | 0.13 | 0.12 | 0.14 | 0.08 | 0.15 | 0.14 | 0.09 | -0.08 | 0.05 | -0.05 | -0.03 | -0.00 | 0.16 | 0.05 | -0.10 | -0.29 | -0.13 | -0.18 | -0.07 | -0.10 |

Contact Preferences AS:

| | V | L | I | M | F | W | Y | G | A | P | S | T | C | H | R | K | Q | E | N | D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| H | 0.01 | 0.39 | 0.18 | 0.32 | 0.21 | 0.04 | -0.13 | -0.42 | 0.47 | -0.60 | -0.23 | -0.29 | -0.42 | -0.10 | -0.04 | -0.13 | -0.01 | 0.01 | -0.29 | -0.13 |
| E | 0.71 | 0.29 | 0.65 | 0.15 | 0.50 | 0.31 | 0.29 | -0.24 | -0.30 | -0.62 | -0.09 | 0.23 | 0.48 | 0.06 | -0.62 | -0.80 | -0.55 | -0.71 | -0.47 | -0.75 |
| T | -0.49 | -0.39 | -0.54 | -0.09 | -0.31 | 0.04 | 0.08 | 0.42 | -0.12 | 0.08 | 0.14 | -0.14 | 0.40 | 0.36 | 0.25 | 0.06 | -0.03 | -0.11 | 0.47 | 0.43 |
| X | -0.17 | -0.28 | -0.24 | -0.27 | -0.13 | 0.16 | 0.17 | 0.17 | -0.27 | 0.63 | 0.17 | 0.14 | 0.33 | 0.10 | 0.03 | -0.26 | -0.02 | -0.10 | 0.22 | 0.13 |
| % | -1.19 | -1.09 | -1.45 | -0.62 | -1.57 | -1.48 | -0.73 | 0.42 | -0.33 | 0.27 | 0.17 | 0.03 | -1.72 | -0.33 | 0.50 | 1.07 | 0.59 | 0.77 | 0.35 | 0.48 |

Contact Preferences AA:

| | V | L | I | M | F | W | Y | G | A | P | S | T | C | H | R | K | Q | E | N | D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| V | 0.39 | 0.23 | 0.33 | 0.20 | 0.26 | 0.20 | -0.02 | -0.04 | 0.18 | -0.08 | -0.14 | -0.07 | 0.12 | -0.06 | -0.36 | -0.36 | -0.30 | -0.41 | -0.20 | -0.35 |
| L | 0.22 | 0.39 | 0.32 | 0.21 | 0.28 | 0.09 | -0.03 | -0.22 | 0.09 | -0.18 | -0.19 | -0.09 | -0.16 | -0.19 | -0.09 | -0.37 | -0.16 | -0.27 | -0.20 | -0.40 |
| I | 0.33 | 0.33 | 0.40 | 0.18 | 0.23 | 0.09 | 0.01 | -0.05 | 0.06 | -0.12 | -0.12 | -0.07 | 0.01 | -0.25 | -0.37 | -0.30 | -0.18 | -0.33 | -0.30 | -0.35 |
| M | 0.20 | 0.23 | 0.13 | 0.61 | 0.42 | 0.40 | -0.05 | -0.23 | 0.07 | -0.08 | -0.34 | -0.23 | 0.12 | 0.05 | -0.30 | -0.36 | -0.18 | -0.15 | -0.14 | -0.42 |
| F | 0.20 | 0.27 | 0.23 | 0.35 | 0.55 | 0.28 | 0.06 | -0.25 | -0.08 | -0.21 | -0.22 | -0.06 | 0.10 | 0.14 | -0.09 | -0.54 | -0.30 | -0.38 | -0.27 | -0.48 |
| W | 0.19 | 0.10 | 0.11 | 0.40 | 0.31 | 0.08 | 0.14 | -0.01 | -0.16 | 0.01 | -0.17 | -0.18 | 0.15 | 0.21 | 0.06 | -0.24 | -0.19 | -0.41 | -0.22 | -0.18 |
| Y | 0.00 | 0.00 | 0.09 | 0.03 | 0.14 | 0.18 | 0.12 | -0.03 | -0.11 | 0.15 | 0.10 | -0.15 | 0.01 | 0.04 | -0.02 | -0.04 | 0.02 | -0.31 | -0.07 | -0.10 |
| G | 0.12 | -0.08 | 0.08 | -0.25 | -0.06 | 0.09 | 0.09 | -0.02 | 0.00 | 0.02 | 0.20 | 0.18 | 0.14 | 0.09 | -0.04 | -0.24 | -0.03 | -0.32 | 0.02 | 0.06 |
| A | 0.23 | 0.16 | 0.16 | 0.08 | -0.04 | -0.14 | -0.07 | -0.02 | 0.35 | -0.11 | -0.02 | 0.05 | -0.03 | -0.24 | -0.34 | -0.11 | -0.04 | -0.15 | -0.04 | -0.11 |
| P | 0.03 | -0.10 | 0.01 | 0.04 | -0.08 | 0.19 | 0.25 | -0.01 | -0.02 | -0.26 | 0.06 | -0.02 | 0.29 | 0.32 | -0.06 | -0.32 | 0.01 | -0.00 | 0.06 | -0.02 |
| S | -0.03 | -0.13 | -0.03 | -0.24 | -0.08 | 0.03 | 0.21 | 0.16 | -0.00 | 0.06 | 0.20 | 0.20 | 0.19 | 0.02 | -0.12 | -0.21 | -0.09 | -0.13 | 0.05 | 0.09 |
| T | -0.00 | 0.01 | 0.01 | -0.13 | 0.07 | -0.06 | -0.04 | 0.08 | 0.07 | -0.03 | 0.17 | 0.19 | -0.18 | 0.24 | -0.08 | -0.33 | -0.11 | -0.13 | 0.03 | 0.12 |
| C | 0.04 | -0.18 | 0.01 | 0.06 | 0.13 | 0.15 | -0.00 | 0.09 | -0.09 | 0.19 | 0.10 | -0.31 | 1.84 | 0.29 | 0.00 | -0.61 | -0.14 | -0.34 | 0.09 | -0.50 |
| H | -0.03 | -0.15 | -0.16 | 0.07 | 0.21 | 0.31 | 0.04 | 0.02 | -0.24 | 0.27 | -0.10 | 0.13 | 0.29 | 0.35 | -0.09 | -0.31 | -0.44 | 0.12 | -0.05 | 0.06 |
| R | -0.16 | 0.05 | -0.11 | -0.20 | 0.12 | 0.23 | 0.14 | -0.09 | -0.24 | -0.11 | -0.12 | -0.06 | 0.10 | 0.04 | 0.04 | -0.89 | 0.18 | 0.33 | -0.18 | 0.40 |
| K | -0.08 | -0.09 | -0.01 | -0.18 | -0.25 | 0.04 | 0.25 | -0.12 | 0.05 | -0.07 | -0.07 | -0.13 | -0.26 | -0.06 | -0.63 | -0.30 | -0.07 | 0.50 | 0.08 | 0.41 |
| Q | -0.11 | 0.00 | -0.01 | -0.05 | -0.10 | 0.05 | 0.12 | 0.05 | 0.04 | 0.00 | -0.09 | -0.04 | 0.18 | -0.27 | 0.20 | -0.17 | -0.03 | -0.07 | 0.24 | 0.07 |
| E | -0.19 | -0.05 | -0.13 | 0.04 | -0.13 | -0.22 | -0.11 | -0.19 | -0.02 | 0.08 | -0.03 | -0.05 | -0.14 | 0.18 | 0.35 | 0.43 | -0.05 | -0.17 | -0.01 | -0.16 |
| N | -0.08 | -0.11 | -0.17 | 0.05 | -0.10 | 0.02 | 0.09 | 0.04 | 0.05 | 0.08 | 0.05 | 0.04 | 0.25 | -0.00 | -0.17 | -0.03 | 0.17 | -0.12 | 0.12 | 0.16 |
| D | -0.17 | -0.23 | -0.17 | -0.28 | -0.29 | 0.09 | 0.03 | 0.04 | 0.00 | 0.03 | 0.15 | 0.16 | -0.34 | 0.12 | 0.41 | 0.26 | 0.05 | -0.23 | 0.16 | -0.29 |
| % | -0.81 | -0.73 | -0.98 | -0.45 | -1.06 | -1.00 | -0.49 | 0.25 | -0.22 | 0.18 | 0.12 | 0.02 | -1.15 | -0.24 | 0.35 | 0.75 | 0.40 | 0.54 | 0.24 | 0.33 |

Contact Preferences AInt:

| | V | L | I | M | F | W | Y | G | A | P | S | T | C | H | R | K | Q | E | N | D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HW | -1.48 | -0.74 | -1.56 | -0.49 | -1.60 | -1.10 | -1.06 | -1.25 | 0.20 | -0.28 | -0.41 | -0.66 | -2.59 | -0.55 | 0.73 | 1.14 | 0.83 | 1.24 | -0.14 | 0.55 |
| HHP | 0.01 | 0.27 | -0.02 | 0.17 | -0.30 | -0.36 | -0.46 | -0.55 | 0.89 | -0.10 | -0.15 | -0.20 | -0.43 | -0.35 | -0.24 | -0.05 | 0.16 | 0.41 | -0.37 | 0.05 |
| HHI | -0.02 | 0.47 | 0.17 | 0.50 | -0.27 | -0.34 | -0.62 | -0.80 | 0.82 | -1.59 | -0.49 | -0.36 | -0.83 | -0.42 | 0.24 | 0.16 | 0.46 | 0.36 | -0.53 | -0.28 |
| HHE | -0.01 | 0.74 | 0.50 | 0.52 | 0.79 | 0.68 | 0.43 | -1.40 | 0.25 | -1.15 | -1.01 | -0.68 | -0.85 | -0.07 | 0.21 | -0.66 | -0.20 | -0.36 | -1.10 | -1.03 |
| HFE | 0.13 | 0.54 | 0.39 | 1.07 | 0.65 | 1.02 | 0.37 | -1.81 | 0.26 | -1.33 | -0.85 | 0.12 | 0.87 | 0.27 | -0.71 | -0.92 | -0.32 | -0.54 | -0.72 | -0.98 |
| HTE | -0.38 | 0.19 | -0.42 | 0.56 | 0.12 | 0.84 | 0.19 | -1.19 | 0.10 | -0.91 | -0.84 | -0.79 | -0.44 | 0.04 | 0.97 | -0.06 | 0.18 | 0.22 | -0.38 | -0.30 |
| HXE | -0.34 | 0.11 | -0.37 | 0.37 | 0.18 | 0.61 | 0.38 | -1.23 | -0.12 | -1.03 | -0.37 | -0.49 | -0.42 | 0.02 | 0.44 | -0.54 | 0.53 | 0.61 | -0.33 | -0.05 |
| FW | -0.44 | -0.67 | -0.48 | -0.73 | -0.98 | -1.68 | 0.02 | -0.59 | -1.62 | -0.74 | 0.23 | 0.67 | -2.41 | -0.20 | 0.79 | 1.02 | 0.64 | 0.46 | 0.08 | -0.54 |
| FFP | 1.02 | 0.30 | 0.81 | -0.02 | 0.25 | -0.49 | 0.10 | -0.04 | -0.29 | -0.95 | 0.11 | 0.47 | 0.32 | -0.13 | -0.76 | -0.66 | -0.52 | -0.81 | -0.56 | -0.98 |
| FFI | 0.93 | 0.52 | 0.71 | 0.26 | 0.84 | -0.68 | 0.71 | -1.33 | -1.02 | -1.66 | -0.75 | 0.09 | -0.08 | 0.51 | -0.32 | -1.07 | -0.55 | -0.92 | -0.97 | -1.07 |
| FFS | 0.89 | 0.34 | 0.84 | 0.12 | 0.65 | 0.36 | 0.50 | -0.65 | -0.61 | -2.05 | -0.20 | 0.23 | 0.24 | -0.03 | -0.54 | -0.85 | -0.47 | -0.76 | -0.80 | -1.38 |
| FFE | 0.75 | 0.78 | 0.87 | 0.61 | 1.12 | 1.04 | 0.43 | -0.87 | -1.19 | -1.88 | -1.06 | -0.15 | 0.21 | -0.27 | -1.05 | -1.14 | -0.59 | -1.29 | -1.25 | -1.62 |
| FHE | 1.04 | 0.51 | 1.15 | 0.34 | 1.19 | 0.64 | 0.67 | -1.37 | -1.13 | -1.14 | -0.70 | -0.73 | 0.16 | -0.24 | -0.71 | -1.35 | -1.33 | -2.15 | -1.15 | -1.60 |
| FTE | 0.24 | -0.33 | 0.21 | -0.46 | -0.05 | -0.36 | 0.80 | -0.99 | -1.20 | -1.11 | 0.36 | 0.23 | 0.40 | 0.09 | 0.58 | -0.12 | -0.58 | -0.88 | 0.32 | 0.09 |
| FXE | 0.59 | 0.15 | 0.55 | -0.07 | 0.45 | 0.44 | 0.74 | -1.15 | -0.86 | -1.12 | -0.24 | 0.00 | 0.15 | 0.21 | 0.19 | -0.63 | -0.42 | -0.53 | -0.48 | -0.73 |
| TW | -2.79 | -2.53 | -2.36 | -1.42 | -1.95 | -2.60 | -0.63 | 1.72 | -0.32 | 0.81 | 0.44 | -0.49 | -1.34 | -0.52 | -0.18 | 1.07 | 0.43 | 0.74 | 1.06 | 0.68 |
| TTP | -1.34 | -1.19 | -1.45 | -1.09 | -0.81 | -1.29 | -0.33 | 1.70 | 0.16 | 1.25 | 0.49 | -0.07 | 0.03 | -0.22 | -0.81 | 0.26 | -0.05 | 0.24 | 0.95 | 0.52 |
| TTI | -0.93 | -0.34 | -1.17 | -0.52 | -0.24 | -0.73 | -0.63 | 0.34 | 0.20 | 0.01 | 0.07 | 0.51 | 0.07 | -0.09 | -0.12 | 0.76 | -0.05 | 0.34 | 0.51 | -0.14 |
| TTE | -1.96 | -0.46 | -0.59 | -0.55 | -0.58 | -1.59 | 0.46 | 0.57 | -0.40 | 1.00 | -0.06 | 0.07 | 0.09 | 0.24 | 0.54 | 0.11 | -0.85 | -0.17 | 1.15 | 0.12 |
| THE | -0.57 | -0.04 | -0.93 | -0.39 | 0.31 | -0.49 | 0.09 | 0.99 | 0.04 | -0.11 | -0.30 | -0.78 | 0.11 | 0.12 | -0.39 | 0.45 | 0.31 | 0.14 | 0.63 | -0.09 |
| TFE | -0.78 | -0.79 | -1.39 | -0.97 | -0.13 | -0.86 | 0.83 | 1.32 | -0.06 | 0.50 | 0.46 | 0.11 | 0.51 | -0.05 | -0.61 | -0.35 | 0.13 | -0.27 | 0.94 | 0.15 |
| TXE | -1.48 | -1.06 | -1.19 | -0.91 | -0.28 | -0.42 | 0.43 | 0.63 | -0.29 | 0.67 | 0.16 | -0.03 | 0.82 | -0.38 | 0.17 | 0.09 | -0.02 | -0.02 | 1.08 | 0.60 |
| XW | -0.85 | -1.21 | -1.50 | -0.44 | -1.63 | -1.46 | -0.82 | 0.62 | -0.71 | 0.62 | 0.46 | 0.44 | -1.19 | -0.11 | 0.41 | 1.01 | 0.38 | 0.26 | 0.41 | 0.56 |
| XXP | -0.16 | -0.49 | -0.40 | -0.60 | -0.57 | -0.53 | -0.58 | 0.79 | -0.03 | 1.13 | 0.64 | 0.51 | 0.48 | 0.16 | -0.53 | -0.11 | -0.35 | -0.48 | 0.32 | 0.44 |
| XXI | -0.24 | -0.47 | -0.53 | -0.59 | -0.06 | 0.19 | 0.09 | 0.02 | -0.61 | 0.70 | 0.29 | 0.25 | 0.74 | 0.19 | 0.11 | -0.18 | 0.27 | -0.47 | 0.46 | 0.19 |
| XXE | -0.42 | -0.47 | -0.40 | -0.16 | -0.28 | 0.43 | 0.21 | 0.22 | -0.44 | 0.33 | 0.16 | 0.20 | 0.74 | 0.55 | 0.12 | -0.30 | -0.19 | -0.31 | 0.50 | 0.39 |
| XHE | -0.34 | 0.17 | -0.15 | 0.23 | 0.35 | 0.45 | -0.03 | -0.15 | -0.52 | 0.59 | 0.38 | 0.10 | -0.02 | 0.38 | -0.10 | -0.40 | -1.19 | -0.93 | 0.14 | 0.42 |
| XFE | 0.17 | -0.07 | 0.41 | -0.30 | 0.26 | 0.87 | -0.07 | -0.21 | -0.11 | 0.64 | 0.21 | -0.07 | 0.86 | 0.29 | -0.53 | -0.83 | -0.93 | -0.48 | -0.16 | 0.13 |
| XTE | -0.54 | -0.62 | -0.48 | -0.24 | -0.51 | 0.40 | 0.03 | -0.18 | -0.69 | 0.34 | 0.12 | -0.18 | 0.87 | 0.88 | 0.38 | -0.02 | -0.17 | -0.53 | 0.53 | 0.76 |

Contact Statistics AM:

| | V | L | I | M | F | W | Y | G | A | P | S | T | C | H | R | K | Q | E | N | D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Wat | 4483 | 6191 | 3273 | 2039 | 3180 | 1494 | 5493 | 4588 | 5584 | 6903 | 8140 | 8331 | 514 | 3836 | 12407 | 23936 | 9544 | 17042 | 9417 | 13105 |
| Prot | 68059 | 87015 | 60066 | 20188 | 63726 | 27968 | 58921 | 19761 | 44043 | 33709 | 43098 | 49515 | 11487 | 30380 | 49719 | 56908 | 35302 | 53573 | 42974 | 53582 |

Contact Statistics AS:

| | V | L | I | M | F | W | Y | G | A | P | S | T | C | H | R | K | Q | E | N | D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| H | 22503 | 37653 | 22077 | 8569 | 23864 | 9383 | 18191 | 5601 | 21219 | 8287 | 13463 | 14558 | 2773 | 9844 | 18600 | 22736 | 13749 | 21967 | 13226 | 18845 |
| E | 24344 | 23346 | 20318 | 5044 | 19340 | 7465 | 16114 | 4231 | 8242 | 5411 | 9860 | 13858 | 3426 | 7288 | 8274 | 9481 | 6267 | 8866 | 7732 | 8106 |
| T | 4392 | 6085 | 3716 | 1778 | 4589 | 2578 | 5808 | 2787 | 3897 | 3652 | 4818 | 4489 | 1354 | 3747 | 6317 | 7176 | 3753 | 5567 | 6185 | 7662 |
| X | 16820 | 19930 | 13944 | 4790 | 15929 | 8542 | 18797 | 7141 | 10659 | 16353 | 14954 | 16606 | 3933 | 9501 | 16520 | 17505 | 11533 | 17156 | 15831 | 18962 |
| % | 4483 | 6191 | 3273 | 2039 | 3180 | 1494 | 5493 | 4588 | 5584 | 6903 | 8140 | 8331 | 514 | 3836 | 12407 | 23936 | 9544 | 17042 | 9417 | 13105 |

Contact Statistics AA:

| | V | L | I | M | F | W | Y | G | A | P | S | T | C | H | R | K | Q | E | N | D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| V | 6766 | 7355 | 5557 | 1746 | 5514 | 2306 | 3966 | 1494 | 3778 | 2352 | 2789 | 3397 | 888 | 2036 | 2712 | 3529 | 2067 | 2917 | 2684 | 2929 |
| L | 7115 | 10876 | 6868 | 2224 | 6975 | 2575 | 4918 | 1545 | 4305 | 2676 | 3337 | 4158 | 828 | 2234 | 4465 | 4357 | 3018 | 4189 | 3359 | 3497 |
| I | 5396 | 6911 | 5078 | 1435 | 4497 | 1723 | 3466 | 1227 | 2791 | 1902 | 2392 | 2861 | 657 | 1397 | 2242 | 3159 | 1984 | 2662 | 2038 | 2459 |
| M | 1707 | 2270 | 1371 | 783 | 1985 | 833 | 1150 | 318 | 996 | 683 | 650 | 845 | 214 | 655 | 846 | 1045 | 674 | 1138 | 836 | 795 |
| F | 4695 | 6464 | 4228 | 1704 | 6190 | 2093 | 3586 | 984 | 2397 | 1718 | 2137 | 2860 | 717 | 2099 | 2971 | 2441 | 1721 | 2500 | 2083 | 2128 |
| W | 1952 | 2277 | 1550 | 724 | 2019 | 677 | 1631 | 492 | 901 | 878 | 916 | 1036 | 272 | 909 | 1448 | 1374 | 780 | 987 | 894 | 1201 |
| Y | 3399 | 4362 | 3257 | 1066 | 3592 | 1660 | 3415 | 1095 | 2058 | 2213 | 2664 | 2319 | 562 | 1665 | 2835 | 3627 | 2143 | 2390 | 2269 | 2816 |
| G | 3338 | 3492 | 2791 | 675 | 2562 | 1311 | 2888 | 964 | 2028 | 1684 | 2559 | 2835 | 565 | 1534 | 2431 | 2584 | 1771 | 2065 | 2170 | 2895 |
| A | 5029 | 6051 | 4095 | 1341 | 3523 | 1402 | 3279 | 1315 | 3933 | 1987 | 2749 | 3357 | 646 | 1470 | 2397 | 3964 | 2363 | 3342 | 2769 | 3273 |
| P | 2455 | 2762 | 2089 | 739 | 2019 | 1163 | 2726 | 772 | 1583 | 991 | 1780 | 1849 | 525 | 1560 | 1907 | 1891 | 1478 | 2310 | 1828 | 2142 |
| S | 3229 | 3750 | 2839 | 784 | 2829 | 1392 | 3682 | 1324 | 2278 | 1996 | 2916 | 3282 | 685 | 1605 | 2531 | 3010 | 1871 | 2836 | 2550 | 3372 |
| T | 3395 | 4400 | 2997 | 899 | 3363 | 1295 | 2895 | 1240 | 2496 | 1848 | 2880 | 3307 | 451 | 2064 | 2674 | 2707 | 1866 | 2886 | 2541 | 3543 |
| C | 933 | 957 | 785 | 244 | 948 | 390 | 785 | 284 | 528 | 590 | 682 | 483 | 1082 | 546 | 759 | 501 | 445 | 590 | 691 | 456 |
| H | 1815 | 2080 | 1380 | 589 | 2156 | 1039 | 1728 | 612 | 979 | 1375 | 1182 | 1713 | 398 | 1267 | 1455 | 1507 | 700 | 2078 | 1267 | 1826 |
| R | 2621 | 4170 | 2400 | 746 | 3233 | 1591 | 3150 | 928 | 1629 | 1528 | 1918 | 2299 | 564 | 1514 | 2748 | 1356 | 2303 | 4199 | 1848 | 4259 |
| K | 3558 | 4523 | 3336 | 976 | 2736 | 1632 | 4405 | 1140 | 2785 | 2008 | 2537 | 2675 | 475 | 1714 | 1698 | 3161 | 2216 | 6221 | 3033 | 5369 |
| Q | 2042 | 2958 | 1964 | 639 | 1894 | 961 | 2312 | 787 | 1623 | 1273 | 1460 | 1748 | 439 | 785 | 2422 | 2131 | 1356 | 2076 | 2121 | 2267 |
| E | 2990 | 4429 | 2770 | 1159 | 2905 | 1156 | 2872 | 992 | 2430 | 2211 | 2472 | 2753 | 509 | 2071 | 4447 | 6235 | 2114 | 2955 | 2589 | 2821 |
| N | 2583 | 3225 | 2044 | 887 | 2308 | 1145 | 2722 | 966 | 2000 | 1699 | 2082 | 2311 | 603 | 1304 | 1996 | 3011 | 2052 | 2411 | 2292 | 3040 |
| D | 3043 | 3701 | 2657 | 818 | 2475 | 1623 | 3334 | 1282 | 2499 | 2095 | 2995 | 3422 | 404 | 1950 | 4725 | 5306 | 2378 | 2804 | 3113 | 2488 |
| % | 4483 | 6191 | 3273 | 2039 | 3180 | 1494 | 5493 | 4588 | 5584 | 6903 | 8140 | 8331 | 514 | 3836 | 12407 | 23936 | 9544 | 17042 | 9417 | 13105 |

Contact Statistics AInt:

| | V | L | I | M | F | W | Y | G | A | P | S | T | C | H | R | K | Q | E | N | D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HW | 1293 | 2776 | 1074 | 791 | 1098 | 685 | 1543 | 511 | 2836 | 1672 | 1924 | 1826 | 99 | 1169 | 5121 | 8884 | 3979 | 8316 | 2373 | 4863 |
| HHP | 6466 | 9937 | 5516 | 2215 | 4812 | 2026 | 4138 | 1472 | 8106 | 3353 | 4091 | 4464 | 785 | 2380 | 4640 | 6897 | 4419 | 8284 | 3587 | 6118 |
| HHI | 6176 | 11129 | 6164 | 2725 | 4811 | 2015 | 3610 | 1207 | 7556 | 1166 | 3152 | 3895 | 583 | 2217 | 6324 | 7785 | 5350 | 7806 | 3135 | 4754 |
| HHE | 3640 | 7874 | 4524 | 1617 | 5851 | 2384 | 4383 | 467 | 2995 | 924 | 1286 | 1824 | 337 | 1648 | 3631 | 2597 | 1978 | 2775 | 1233 | 1655 |
| HFE | 1500 | 2559 | 1574 | 883 | 1992 | 1129 | 1574 | 131 | 1121 | 305 | 536 | 1190 | 415 | 779 | 720 | 809 | 679 | 921 | 601 | 639 |
| HTE | 923 | 1759 | 781 | 542 | 1196 | 873 | 1210 | 177 | 876 | 358 | 471 | 551 | 146 | 580 | 2012 | 1275 | 841 | 1354 | 663 | 892 |
| HXE | 2880 | 5086 | 2477 | 1451 | 3817 | 2261 | 4236 | 522 | 2306 | 1005 | 1994 | 2070 | 452 | 1754 | 4239 | 2796 | 3260 | 5418 | 2102 | 3254 |
| FW | 936 | 1025 | 795 | 234 | 592 | 161 | 1140 | 283 | 281 | 424 | 1050 | 1611 | 39 | 520 | 1878 | 2862 | 1218 | 1695 | 969 | 800 |
| FFP | 8666 | 6764 | 6531 | 1294 | 4703 | 1236 | 4063 | 1394 | 2393 | 1242 | 3270 | 4726 | 885 | 1848 | 2169 | 3011 | 1850 | 2379 | 2095 | 1988 |
| FFI | 1533 | 1479 | 1146 | 295 | 1324 | 203 | 1168 | 107 | 271 | 142 | 337 | 683 | 126 | 540 | 550 | 428 | 339 | 412 | 295 | 351 |
| FFS | 8882 | 7804 | 7477 | 1597 | 6948 | 2489 | 6008 | 1021 | 2146 | 647 | 2936 | 4477 | 938 | 2207 | 2824 | 2952 | 2131 | 2743 | 1981 | 1689 |
| FFE | 1468 | 1936 | 1399 | 410 | 1758 | 728 | 1048 | 160 | 263 | 133 | 297 | 630 | 168 | 342 | 363 | 442 | 360 | 348 | 266 | 262 |
| FHE | 2848 | 2540 | 2688 | 539 | 2928 | 879 | 1964 | 180 | 434 | 353 | 603 | 666 | 257 | 555 | 729 | 606 | 342 | 304 | 451 | 420 |
| FTE | 736 | 637 | 632 | 139 | 558 | 198 | 968 | 106 | 186 | 162 | 566 | 583 | 137 | 313 | 797 | 639 | 258 | 330 | 565 | 612 |
| FXE | 3781 | 3590 | 3216 | 734 | 3173 | 1391 | 3734 | 380 | 951 | 648 | 1507 | 2009 | 461 | 1375 | 2457 | 1818 | 1165 | 1703 | 1307 | 1396 |
| TW | 252 | 388 | 296 | 200 | 417 | 117 | 1000 | 1938 | 958 | 1709 | 1675 | 993 | 114 | 574 | 1316 | 4089 | 1455 | 2831 | 2622 | 2564 |
| TTP | 907 | 1288 | 732 | 329 | 1208 | 380 | 1615 | 2500 | 1749 | 3059 | 2271 | 1747 | 386 | 928 | 1118 | 3058 | 1374 | 2629 | 3194 | 3017 |
| TTI | 53 | 103 | 36 | 10 | 79 | 12 | 56 | 68 | 85 | 58 | 77 | 125 | 23 | 42 | 79 | 193 | 60 | 125 | 119 | 85 |
| TTE | 79 | 285 | 177 | 62 | 188 | 41 | 372 | 152 | 158 | 342 | 207 | 255 | 57 | 169 | 381 | 367 | 105 | 264 | 489 | 305 |
| THE | 876 | 1630 | 596 | 304 | 1487 | 376 | 1229 | 868 | 919 | 677 | 749 | 607 | 232 | 666 | 850 | 1982 | 999 | 1397 | 1452 | 1128 |
| TFE | 295 | 376 | 169 | 79 | 427 | 114 | 800 | 424 | 332 | 403 | 493 | 436 | 128 | 231 | 284 | 445 | 342 | 410 | 702 | 517 |
| TXE | 511 | 878 | 545 | 232 | 1087 | 433 | 1709 | 742 | 797 | 1269 | 1129 | 1117 | 418 | 518 | 1375 | 1694 | 874 | 1376 | 2175 | 1990 |
| XW | 2003 | 2002 | 1109 | 814 | 1073 | 532 | 1810 | 1857 | 1509 | 3098 | 3491 | 3901 | 261 | 1574 | 4092 | 8101 | 2892 | 4200 | 3452 | 4878 |
| XXP | 5623 | 5745 | 4164 | 1272 | 3905 | 1762 | 3727 | 3649 | 4211 | 7703 | 6912 | 7121 | 1452 | 3298 | 3719 | 6472 | 3055 | 4379 | 5675 | 7831 |
| XXI | 1171 | 1289 | 838 | 281 | 1229 | 645 | 1314 | 473 | 621 | 1258 | 1200 | 1312 | 383 | 746 | 1286 | 1366 | 1034 | 972 | 1376 | 1458 |
| XXE | 2730 | 3372 | 2402 | 996 | 2764 | 1997 | 3733 | 1421 | 1833 | 2572 | 2871 | 3331 | 1006 | 2514 | 3395 | 3293 | 1977 | 2864 | 3732 | 4397 |
| XHE | 2616 | 4784 | 2605 | 1188 | 3890 | 1832 | 2864 | 997 | 1570 | 2786 | 3027 | 2816 | 540 | 2028 | 2641 | 2783 | 896 | 1684 | 2637 | 4048 |
| XFE | 2621 | 2855 | 2689 | 578 | 2572 | 1723 | 1963 | 677 | 1473 | 2032 | 1906 | 1769 | 699 | 1344 | 1379 | 1462 | 755 | 1622 | 1508 | 2335 |
| XTE | 1077 | 1315 | 978 | 406 | 1013 | 842 | 1424 | 465 | 663 | 1108 | 1205 | 1107 | 473 | 1357 | 1753 | 1729 | 860 | 1057 | 1634 | 2434 |

Contact Occurrences AM:

| | V | L | I | M | F | W | Y | G | A | P | S | T | C | H | R | K | Q | E | N | D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Wat | 917 | 1037 | 684 | 240 | 489 | 162 | 421 | 994 | 1099 | 540 | 819 | 737 | 182 | 276 | 491 | 854 | 451 | 742 | 577 | 748 |
| Prot | 10107 | 12177 | 8256 | 2781 | 6422 | 2274 | 5219 | 6360 | 9315 | 4331 | 6460 | 6527 | 1897 | 2902 | 5057 | 6906 | 3972 | 6090 | 4982 | 6015 |

Contact Occurrences AS:

| | V | L | I | M | F | W | Y | G | A | P | S | T | C | H | R | K | Q | E | N | D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| H | 3443 | 5310 | 3055 | 1206 | 2505 | 783 | 1656 | 2066 | 4606 | 1218 | 2092 | 2039 | 487 | 982 | 1982 | 2894 | 1654 | 2689 | 1645 | 2261 |
| E | 3526 | 3220 | 2757 | 684 | 1899 | 609 | 1432 | 1318 | 1753 | 794 | 1526 | 1851 | 562 | 688 | 866 | 1142 | 690 | 979 | 912 | 966 |
| T | 672 | 904 | 552 | 246 | 452 | 211 | 483 | 835 | 832 | 515 | 714 | 602 | 214 | 350 | 619 | 858 | 425 | 651 | 671 | 776 |
| X | 2466 | 2742 | 1890 | 644 | 1564 | 671 | 1647 | 2141 | 2118 | 1803 | 2127 | 2034 | 634 | 882 | 1588 | 2010 | 1203 | 1768 | 1754 | 2011 |
| % | 917 | 1037 | 684 | 240 | 489 | 162 | 421 | 994 | 1099 | 540 | 819 | 737 | 182 | 276 | 491 | 854 | 451 | 742 | 577 | 748 |

Contact Occurrences AA:

| | V | L | I | M | F | W | Y | G | A | P | S | T | C | H | R | K | Q | E | N | D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| V | 1144 | 1159 | 868 | 263 | 595 | 202 | 427 | 499 | 861 | 349 | 477 | 501 | 168 | 221 | 332 | 495 | 268 | 406 | 353 | 377 |
| L | 1163 | 1665 | 1054 | 332 | 777 | 230 | 467 | 558 | 966 | 366 | 537 | 599 | 153 | 239 | 509 | 636 | 375 | 569 | 424 | 473 |
| I | 870 | 1057 | 731 | 203 | 509 | 157 | 347 | 393 | 625 | 252 | 394 | 418 | 127 | 152 | 265 | 434 | 238 | 358 | 255 | 315 |
| M | 257 | 329 | 212 | 107 | 195 | 71 | 114 | 109 | 222 | 102 | 116 | 121 | 37 | 73 | 106 | 149 | 91 | 140 | 111 | 116 |
| F | 574 | 776 | 493 | 189 | 510 | 145 | 261 | 286 | 447 | 200 | 313 | 361 | 104 | 170 | 249 | 280 | 153 | 266 | 226 | 241 |
| W | 195 | 230 | 157 | 69 | 142 | 40 | 103 | 121 | 152 | 83 | 125 | 117 | 42 | 62 | 97 | 110 | 67 | 95 | 81 | 122 |
| Y | 423 | 461 | 341 | 121 | 272 | 100 | 225 | 303 | 349 | 215 | 351 | 267 | 76 | 121 | 205 | 304 | 164 | 234 | 226 | 294 |
| G | 569 | 598 | 440 | 122 | 330 | 138 | 345 | 407 | 538 | 301 | 479 | 473 | 116 | 202 | 324 | 416 | 275 | 329 | 335 | 429 |
| A | 901 | 1007 | 662 | 230 | 453 | 161 | 373 | 498 | 966 | 318 | 481 | 532 | 118 | 193 | 345 | 595 | 333 | 483 | 383 | 449 |
| P | 329 | 368 | 258 | 98 | 199 | 85 | 221 | 236 | 301 | 141 | 258 | 223 | 80 | 127 | 185 | 228 | 160 | 227 | 186 | 221 |
| S | 501 | 545 | 408 | 116 | 323 | 137 | 377 | 425 | 485 | 279 | 430 | 435 | 124 | 157 | 296 | 373 | 208 | 311 | 314 | 364 |
| T | 506 | 606 | 423 | 129 | 379 | 118 | 281 | 399 | 528 | 233 | 416 | 442 | 82 | 196 | 280 | 338 | 212 | 311 | 289 | 362 |
| C | 173 | 153 | 126 | 35 | 106 | 42 | 78 | 95 | 118 | 82 | 116 | 80 | 146 | 49 | 85 | 75 | 57 | 84 | 86 | 60 |
| H | 213 | 236 | 156 | 73 | 176 | 61 | 124 | 171 | 199 | 127 | 153 | 190 | 51 | 105 | 131 | 148 | 79 | 177 | 123 | 162 |
| R | 335 | 495 | 281 | 105 | 254 | 95 | 215 | 290 | 330 | 180 | 287 | 270 | 84 | 129 | 241 | 193 | 217 | 370 | 206 | 371 |
| K | 512 | 632 | 452 | 146 | 282 | 113 | 315 | 356 | 583 | 253 | 363 | 338 | 81 | 161 | 209 | 434 | 251 | 595 | 338 | 507 |
| Q | 279 | 388 | 245 | 84 | 161 | 69 | 174 | 243 | 321 | 166 | 206 | 209 | 65 | 79 | 222 | 239 | 143 | 224 | 215 | 243 |
| E | 401 | 568 | 352 | 138 | 278 | 100 | 238 | 306 | 477 | 258 | 313 | 301 | 88 | 175 | 386 | 602 | 229 | 346 | 268 | 306 |
| N | 356 | 418 | 262 | 106 | 231 | 85 | 235 | 287 | 378 | 190 | 293 | 281 | 91 | 121 | 212 | 338 | 211 | 268 | 238 | 322 |
| D | 406 | 485 | 333 | 114 | 248 | 125 | 298 | 378 | 463 | 235 | 351 | 368 | 64 | 170 | 376 | 517 | 241 | 294 | 325 | 280 |
| % | 917 | 1037 | 684 | 240 | 489 | 162 | 421 | 994 | 1099 | 540 | 819 | 737 | 182 | 276 | 491 | 854 | 451 | 742 | 577 | 748 |

Contact Occurrences AInt:

| | V | L | I | M | F | W | Y | G | A | P | S | T | C | H | R | K | Q | E | N | D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HW | 281 | 457 | 232 | 108 | 176 | 64 | 136 | 167 | 545 | 112 | 206 | 191 | 42 | 85 | 207 | 321 | 199 | 362 | 150 | 249 |
| HHP | 562 | 914 | 464 | 216 | 352 | 128 | 272 | 334 | 1090 | 224 | 412 | 382 | 84 | 170 | 414 | 642 | 398 | 724 | 300 | 498 |
| HHI | 1219 | 1866 | 997 | 464 | 739 | 249 | 512 | 569 | 2160 | 291 | 684 | 748 | 144 | 323 | 852 | 1253 | 792 | 1276 | 548 | 829 |
| HHE | 602 | 1257 | 681 | 239 | 586 | 178 | 353 | 161 | 680 | 144 | 213 | 275 | 60 | 148 | 332 | 299 | 228 | 315 | 160 | 191 |
| HFE | 300 | 482 | 291 | 141 | 235 | 106 | 150 | 54 | 301 | 47 | 105 | 184 | 77 | 75 | 84 | 98 | 81 | 118 | 75 | 80 |
| HTE | 177 | 334 | 156 | 87 | 137 | 75 | 123 | 81 | 265 | 67 | 110 | 94 | 33 | 73 | 216 | 206 | 130 | 216 | 95 | 134 |
| HXE | 491 | 795 | 388 | 211 | 403 | 185 | 390 | 186 | 535 | 172 | 306 | 292 | 82 | 165 | 415 | 358 | 317 | 519 | 254 | 350 |
| FW | 361 | 284 | 258 | 55 | 154 | 37 | 124 | 149 | 154 | 48 | 161 | 190 | 44 | 59 | 89 | 128 | 73 | 99 | 82 | 77 |
| FFP | 722 | 568 | 516 | 110 | 308 | 74 | 248 | 298 | 308 | 96 | 322 | 380 | 88 | 118 | 178 | 256 | 146 | 198 | 164 | 154 |
| FFI | 355 | 241 | 234 | 46 | 143 | 20 | 108 | 46 | 93 | 33 | 90 | 160 | 30 | 58 | 66 | 78 | 51 | 68 | 54 | 53 |
| FFS | 1428 | 1161 | 1073 | 232 | 734 | 191 | 539 | 357 | 496 | 125 | 481 | 636 | 163 | 219 | 300 | 356 | 230 | 294 | 233 | 205 |
| FFE | 266 | 304 | 227 | 64 | 198 | 60 | 116 | 63 | 66 | 24 | 60 | 110 | 34 | 41 | 35 | 58 | 47 | 45 | 34 | 44 |
| FHE | 512 | 451 | 459 | 91 | 304 | 63 | 164 | 74 | 121 | 50 | 107 | 117 | 46 | 58 | 77 | 84 | 38 | 39 | 54 | 67 |
| FTE | 153 | 114 | 116 | 22 | 73 | 24 | 78 | 48 | 50 | 32 | 96 | 107 | 21 | 40 | 77 | 85 | 42 | 46 | 76 | 65 |
| FXE | 751 | 629 | 571 | 121 | 379 | 129 | 399 | 143 | 242 | 113 | 277 | 346 | 99 | 154 | 235 | 256 | 151 | 201 | 178 | 178 |
| TW | 38 | 56 | 29 | 14 | 37 | 9 | 47 | 279 | 118 | 103 | 113 | 70 | 21 | 28 | 45 | 130 | 57 | 102 | 127 | 119 |
| TTP | 76 | 112 | 58 | 28 | 74 | 18 | 94 | 558 | 236 | 206 | 226 | 140 | 42 | 56 | 90 | 260 | 114 | 204 | 254 | 238 |
| TTI | 14 | 22 | 12 | 5 | 17 | 5 | 15 | 35 | 33 | 24 | 26 | 31 | 10 | 10 | 19 | 43 | 11 | 34 | 30 | 22 |
| TTE | 15 | 41 | 24 | 13 | 19 | 5 | 33 | 63 | 38 | 61 | 37 | 34 | 10 | 16 | 37 | 48 | 14 | 44 | 57 | 41 |
| THE | 140 | 223 | 89 | 39 | 157 | 32 | 110 | 417 | 228 | 118 | 139 | 90 | 50 | 65 | 83 | 252 | 114 | 170 | 204 | 162 |
| TFE | 49 | 65 | 27 | 17 | 34 | 11 | 73 | 174 | 77 | 76 | 95 | 69 | 29 | 26 | 36 | 68 | 35 | 56 | 105 | 81 |
| TXE | 83 | 152 | 85 | 43 | 120 | 33 | 157 | 323 | 206 | 236 | 201 | 154 | 70 | 55 | 139 | 248 | 116 | 194 | 285 | 242 |
| XW | 237 | 240 | 165 | 63 | 122 | 52 | 114 | 399 | 282 | 277 | 339 | 286 | 75 | 104 | 150 | 275 | 122 | 179 | 218 | 303 |
| XXP | 463 | 466 | 324 | 114 | 239 | 104 | 226 | 786 | 542 | 552 | 665 | 566 | 149 | 207 | 293 | 537 | 242 | 353 | 429 | 601 |
| XXI | 231 | 195 | 145 | 46 | 123 | 50 | 110 | 209 | 185 | 268 | 251 | 233 | 78 | 86 | 144 | 214 | 131 | 144 | 192 | 216 |
| XXE | 436 | 482 | 349 | 121 | 280 | 173 | 341 | 491 | 414 | 397 | 488 | 469 | 178 | 246 | 357 | 383 | 221 | 314 | 432 | 508 |
| XHE | 422 | 643 | 374 | 163 | 379 | 143 | 262 | 418 | 378 | 424 | 489 | 405 | 100 | 207 | 242 | 362 | 110 | 196 | 333 | 468 |
| XFE | 458 | 443 | 438 | 82 | 270 | 150 | 218 | 262 | 384 | 358 | 362 | 315 | 134 | 147 | 164 | 229 | 111 | 192 | 225 | 317 |
| XTE | 182 | 216 | 156 | 65 | 117 | 68 | 127 | 210 | 181 | 192 | 217 | 189 | 86 | 139 | 170 | 231 | 102 | 127 | 211 | 270 |

```
//*****************************************************************************
//*                    C++ CODE FOR THE INTERFACE
//*****************************************************************************

enum ContIntSet{
   eHW,  //  water
   eHHP, //  dist =  1
   eHHI, //  dist >  1 within the same helix
   eHHE, //  two different helicees
   eHFE, //  any dist
   eHTE, //  any dist
   eHXE, //  any dist
   eFW,  //  water
   eFFP, //  dist =  1
   eFFI, //  dist >  1 within a strand
   eFFS, //  two strands of the same sheet
   eFFE, //  strands in different sheets
   eFHE, //
   eFTE, //  any dist
   eFXE, //  any dist
   eTW,  //  water
   eTTP, //  dist =  1
   eTTI, //  dist >  1 within the turn
   eTTE, //  two different turns
   eTHE, //
   eTFE, //
   eTXE, //  any dist
   eXW,  //  water
   eXXP, //  dist = 1
   eXXI, //  dist > 1
   eXXE, //  two different elements
   eXHE, //
   eXFE, //
   eXTE, //
   eUNK = -1
 };

char
FindHFTX(char secStruc)
{
  static char *DsspH =  "HGI";
  static char *DsspF =  "E";
  static char *DsspT =  "T";
  if(strchr(DsspH,secStruc))
    return 'h';
  else if (strchr(DsspF,secStruc))
    return 'f';
  else if (strchr(DsspT,secStruc))
    return 't';
  else
    return 'x';
}

ContIntSet
ContAIntStat::SpecifyInterfKind (SeqObjPtr seq1, SeqObjPtr seq2)
{
  char HFTX1;
  char HFTX2;
  if( seq1->Seq()=='%' || seq1->IsOfType(eAcc) ||
      seq2->Seq()=='%' || seq2->IsOfType(eAcc)   )
    {
      // --- We are dealing with water contacts....
      char HFTX;
      if(seq1->Seq()=='%' || seq1->IsOfType(eAcc) )
        {
          HFTX=Guard(seq2,Res)->SecStruc();
        }
      else
        {
          HFTX=Guard(seq1,Res)->SecStruc();
        }
```



```
      switch(HFTX)
        {
        case 'h':
          return eHW;
        case 'f':
          return eFW;
        case 't':
          return eTW;
        default: // 'x'
          return eXW;
        }
    }
  ResPtr res1=Guard(seq1,Res);
  ResPtr res2=Guard(seq2,Res);
  HFTX1=res1->SecStruc();
  HFTX2=res2->SecStruc();
  if(HFTX1=='?' || HFTX2=='?')
    return eUNK;                    // Ignore this contact!

  char relation='?';                // Relation of sec struc elements in which the two
                                    // residues are.

  if (ABS(res2->DsspNo() - res1->DsspNo()) == 1) // Distance in Sequence
    {
      relation='P';
      HFTX2=HFTX1; // just a little hack...
    }
  else if(res1->SecStrucSegmentNo() != res2->SecStrucSegmentNo()) // EXTERNAL
    relation='E';
  else
    relation='I';

  switch(HFTX1)
    {
    case 'h':
      switch(HFTX2)
        {
        case 'h':
          switch(relation)
            {
            case 'P':
              return eHHP;
            case 'I':
              return eHHI;
            case 'E':
              return eHHE;
            default:
              BUG;
            }
        case 'f':
          return eHFE;
        case 't':
          return eHTE;
        case 'x':
          return eHXE;
        default:
          BUG;
        }
      break;
    case 'f':
      switch(HFTX2)
        {
        case 'h':
          return eFHE;
        case 'f':
          switch(relation)
            {
            case 'P';
              return eFFP;
            case 'I':
              return eFFI;
```

```
          case 'E':
            if( res1->DSSP()->Sheet() == res2->DSSP()->Sheet() )
              return eFFS;
            else
              return eFFE;
          default:
            BUG;
          }
        break;
      case 't':
        return eFTE;
      case 'x':
        return eFXE;
      default:
        BUG;
      }
  case 't':
    switch(HFTX2)
      {
      case 'h':
        return eTHE;
      case 'f':
        return eTFE;
      case 't':
        switch(relation)
          {
          case 'P':
            return eTTP;
          case 'I':
            return eTTI;
          case 'E':
            return eTTE;
          default:
            BUG;
          }
      case 'x':
        return eTXE;
      default:
        BUG;
      }
  case 'x':
    switch(HFTX2)
      {
      case 'h':
        return eXHE;
      case 'f':
        return eXFE;
      case 't':
        return eXTE;
      case 'x':
        switch(relation)
          {
          case 'P':
            return eXXP;
          case 'I':
            return eXXI;
          case 'E':
            return eXXE;
          default:
            BUG;
          }
      default:
        BUG;
      }
  default:
    BUG;
  }
 return eUNK;
}
```



```
//*****************************************************************************
//*                    C++ CODE FOR ENERGY CALCULATION
//*****************************************************************************

float
LinSquare::Stren(AtomPtr a1,AtomPtr a2,float dist)
{
  // ignore the mainchain N,C,O atoms ( but not for he contact partner!).
  if(a1->Name()== " N  " ||
     a1->Name()== " C  " ||
     a1->Name()== " O  ")
    return 0;
  if (dist < 3.6)
    return 1;
  else if (dist >= 6.4)
    return 0;
  else
    return ((6.4-dist) / 2.8);
}

float
Access::Stren(ResPtr res)
{
  // The accessibility was generated with a dssp version which ignores
  // the mainchain N,C,O atoms.

  // The factor 0.31:
  // Francois Collona-Cesari and Chris Sander
  // Biophys.J Vol 57, pp 1103-1107 (1990)
  return res->Access()*0.31;

}
```