###
#
MSKCC Document Delivery Services
#
Friday, November 25, 2005
#
##

Request ID: DDS36697
User: Gangi-Dino, Rita
Location: MSK
Requested on: 11/23/2005
Needed by: 11/28/2005
Book Title: Oxford University Press
ISBN:
Book Author:
Chapter Author: Scharf, M, Schneider, R
Chapter Title: Design of Protein structures
Year: 1992
Pages:
Edition:
Publisher:
User's Comments: In color, if available
Staff Notes: Protein Engineering a Practical Approach 89-115.

R., and Goldstein, C. (1988).
age handling, pp. 840–8.
pl. Math., **48**, 1073–82.
D. (1989). *J. Mol. Biol.*, **86**,

(1990). *Proteins*, **9**, 180–90.
–14.

Biologist, **2**, 197–206.
Gaber, B. P. (1988). *J. Mol.*

Gaber, B. P. (1990). *Protein*

823–6.
.
Rayment, I. (1987). *EMBO*

3.
–4.

(1990). *Proteins Struct. Func.*

4

Design of protein structures

CHRIS SANDER, MICHAEL SCHARF,
and REINHARD SCHNEIDER

1. Introduction

1.1 The protein design cycle

The design of protein structures, whether by variation of natural proteins or from scratch, is an essential first step in the development of newly engineered proteins. Designs based on theoretical and computer methods enter the protein design cycle of repeated steps of experimental testing and improvement, much as in other engineering disciplines. Success in protein design depends on how well we make use of this cycle and on how well we understand the principles of protein folding, both in terms of the underlying molecular physics and in terms of knowledge extracted from the data-bases of macromolecular sequences and structures.

1.2 Two examples: redesign, *de novo* design

Here we illustrate two basic approaches to protein design, one incremental, a redesign project, the other bolder, a *de novo* design. The first example is the re-engineering of the loop connections in a naturally occurring protein of simple architecture, a bundle of four α-helices, Rop protein. The second example is the design from scratch of a geometrically regular sandwich of two four-stranded β-sheets, called Shpilka. We describe some of the methods used in these designs, but the selection is far from exhaustive.

1.3 Two tools: secondary structure and interface preference parameters

Of the many theoretical, data-based, and statistical tools, we present two in detail, namely two types of statistical preference tables for amino acid residue types in specific structural positions in a protein. The first uses structural states in terms of secondary structure type, position in a secondary structure segment, and solvent accessibility. The second, in terms of contacts between secondary structure segments and with water. These parameters can be used to help answer questions such as 'which residue types are preferred in solvent-exposed positions at the C-terminal end of an α-helix?' or 'what is the typical residue composition of a helix–sheet interface?'

1.4 Aspects not covered here

This chapter focuses on purely structural aspects, leaving out the important questions of biological function or activity. Also not covered here are: protein design using non-natural scaffolds, design of metal binding sites, the design of enzymatically active sites, and the design of membrane channels. Readers interested in further details of the theoretical and computer methods used in protein design are referred to recent reviews (1, 2) and may request from the authors detailed written reports from protein design workshops held in 1986 and 1990, under the auspices of the European Molecular Biology Organization (3, 4, 5).

2. Redesigning natural proteins: example, an α-helix bundle

2.1 The effect of mutations on protein structure

As the problem of designing protein structures from first principles is a very difficult one, we are well advised to devote some effort to a simpler problem, that of understanding the effects of amino acid sequence changes on a known protein structure. In which way does replacement of certain amino acids affect the conformation of the native (time-averaged) structure? In which way do point mutations affect the stability of the native protein structure relative to an unfolded or denatured state?

2.2 Nature's experience

Nature has accumulated vast experience, apparently by trial and error, of testing point mutations. Just think of the many variations in sequence in immunoglobulin molecules tested daily. Natural selection operates by selecting for functional properties, but structural integrity is a prerequisite for the proper function of many proteins. So the thousands of natural sequences known to be homologous to proteins of known structure can be considered the result of a giant series of evolutionary experiments in testing the effect of sequence changes on protein structure. We merely have to draw on the databases of protein sequences and structures to learn from these experiments (6, 7, 8, 9).

2.3 The protein engineer's experience

In comparison, the number of deliberately engineered mutations tested in the laboratory is very small, and often very little direct information about the details of structural changes is available. However, the data-base of experi-

n α-helix?' or 'what is the typical
ce?'

pects, leaving out the important
lso not covered here are: protein
metal binding sites, the design of
of membrane channels. Readers
l and computer methods used in
(1, 2) and may request from the
n design workshops held in 1986
an Molecular Biology Organiza-

s: example, an

otein structure

res from first principles is a very
some effort to a simpler problem,
cid sequence changes on a known
acement of certain amino acids
veraged) structure? In which way
native protein structure relative

apparently by trial and error, of
many variations in sequence in
tural selection operates by select-
integrity is a prerequisite for the
thousands of natural sequences
own structure can be considered
xperiments in testing the effect of
merely have to draw on the data-
to learn from these experiments

ence

ngineered mutations tested in the
ttle direct information about the
lowever, the data-base of experi-

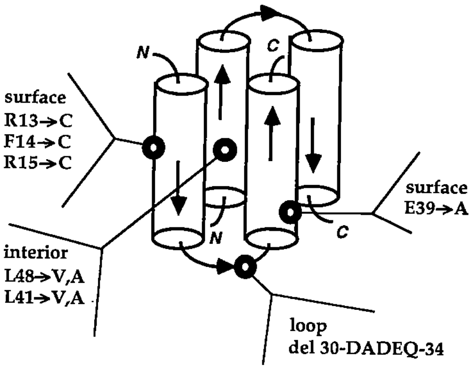


Figure 1. Cartoon of Rop protein based on the wild-type crystal and NMR structures (44, 45). Mutations on the protein *surface*, in the protein *interior*, and in *loop* regions have explored the stability of this fold with respect to sequence changes (12). Because of the dimeric symmetry of the helix bundle, each mutation appears twice in the three-dimensional structure, once in the front and once in the back subunit. In the figure, each mutation is only shown once, for the front subunit. The main result is that this four-helix bundle is very resistant to sequence changes on the solvent-exposed surface and in loop regions, and quite sensitive to mutations in the protein core (contact interface between the helices). Thus the bundle is an excellent scaffold for protein redesign experiments. The amino acid sequence of wild-type Rop is, in one-letter code, MTKQEKTALN MARFIRSQTL TLLEKLNELD ADEQADICES LHDHADELYR SCLARF.

mentally determined structures of mutant proteins is increasing steadily. One of the richest sets of mutant crystal structures, of phage T4 lysozyme, comes from the group of Brian Matthews (10, 11). Here, we illustrate some of the principles used in redesigning protein structure with a particularly simple example, that of the four-helix bundle, Rop (*Figure 1*).

2.4 Regions of protein structure

Just as a mechanical engineer would, protein engineers have developed certain concepts of the constructive elements of a protein, based on physical principles and on the analysis of natural proteins. A principal distinction is between the *framework*, which consists of hydrogen-bonded secondary structure segments, and *loops*, which connect these segments. In Rop protein, the framework consists of the four α-helices. Another important distinction is between the exposed surface, or protein *exterior*, defined as those parts of the protein accessible to direct contact by solvent molecules in the time-averaged native structure, and the protein *interior*, not in direct contact with solvent. In Rop protein, each helix has an interior and an exterior face, most of the backbone is in the interior and most of the loop residues are exposed to solvent. A third major distinction is made between the protein *core*, defined as those regions which are the most conserved in sequence and structure in the evolution of a protein family, and the *variable regions*, which are the most flexible in conformation and the most variable in sequence. In the Rop four-helix bundle, the mutually contacting interior faces of the helices constitute the core. These distinctions are useful not only in describing protein structure but also in formulating design rules. We now illustrate some of these rules by example.

2.5 Mutations in the protein core

In our example protein, Leu41 and Leu48 in the core of Rop protein were each changed to Val or Ala (*Figure 1*) (G. Cesareni and M. Kokkinidis, personal communication). The expectation was that replacing one non-polar side-chain by another would not lead to major structural changes but that the introduction of a hole in the hydrophobic protein core would reduce the stability of the native state. The crystal structure (M. Kokkinidis, personal communication) confirmed that the deletion of methyl groups in these interior side-chains simply leaves a hole in the protein interior without significant structural rearrangement. Calorimetric measurements (H. J. Hinz and P. Weber, personal communication) confirmed that the introduction of holes in the protein interior destabilizes the native structure relative to unfolded states.

2.6 Mutations on the exposed protein surface

As the primary interactions of exterior residues are with solvent molecules, their contribution to details of three-dimensional (3D) structure are expected to be minimal. On the other hand, solvation properties, such as a tendency for protein–protein aggregation, are expected to depend on surface residues, although perhaps not in a position-specific way. In the Rop example, replacement of many residues on the exposed surface of α-helices had no apparent effect of the native structure of the protein (detailed 3D structures were, however, not determined). Examples are the mutants D28A, D32F, R13C, R15C, F14C, E39A, D59H (*Figure 1*) (12) (although the function of some of these mutants was impaired).

2.7 Mutations in loops

Residues in loop regions also are primarily exposed to solvent and are observed to be variable in evolution, but they may also have a crucial role in determining the folding pathway. In one experiment, Cesareni and colleagues (12) deleted five amino acids (30-DADEQ-34; *Figure 1*) from the loop connecting the two helices of the Rop monomer in order to test the possibility of the formation of a single long helix. The deletion was chosen such that the long helix would have a continuous strip of hydrophobic residues on one helix face, as analysed in a helical wheel projection. The outcome of the experiment underscored the considerable conformational flexibility of loop regions

ble regions, which are the most
in sequence. In the Rop four-
faces of the helices constitute
in describing protein structure
llustrate some of these rules by

the core of Rop protein were
Cesareni and M. Kokkinidis,
as that replacing one non-polar
structural changes but that the
protein core would reduce the
cture (M. Kokkinidis, personal
of methyl groups in these in-
protein interior without signific-
measurements (H. J. Hinz and
d that the introduction of holes
structure relative to unfolded

ein surface

ues are with solvent molecules,
nal (3D) structure are expected
properties, such as a tendency
d to depend on surface residues,
y. In the Rop example, replace-
ce of α-helices had no apparent
(detailed 3D structures were,
e mutants D28A, D32F, R13C,
lthough the function of some of

ly exposed to solvent and are
y may also have a crucial role in
riment, Cesareni and colleagues
34; *Figure 1*) from the loop con-
in order to test the possibility of
letion was chosen such that the
ydrophobic residues on one helix
ion. The outcome of the experi-
ational flexibility of loop regions

and the persistence of the core structure formed by the interface between the four helices of the bundle: in the crystal structure, the helical bundle was well formed, except that a few residues at the ends of the two helices simply unravelled and formed a new type of loop in place of the deleted five residues (12).

2.8 Re-engineering protein topology

Mutation experiments on Rop and many other globular proteins have established the fact that the protein surface as well as loop regions are, in general, the most tolerant of mutations while the core is much more sensitive. Physically, the difference is primarily due to the fact that interactions involving sidechains in the protein core are highly specific and vary little in time, while interactions involving solvent are mostly non-specific and fluctuate strongly in time.

These results suggest the simple hypothesis that protein folding is dominated by interactions in the core of the protein. Let us call this the 'core hypothesis of protein folding'. According to this hypothesis, both the formation of the correct tertiary structure and the stability of folded proteins may be explained in terms of tight packing and specific interactions in the protein interior; the surface of the protein merely needs to have average solvation properties and loops merely need to provide connections of the appropriate length and flexibility.

To test one aspect of this hypothesis, the question asked was: to what extent could the loops connecting the four helices of the Rop bundle be rearranged without negatively affecting the correct fold? In a first experiment, the chain threading of this small protein was modified in a simple but radical fashion: deletion of one loop and addition of two new loops would turn the wild-type dimeric bundle into an artificial left-handed monomeric bundle, with an identically packed hydrophobic core but rather different topology of loop connections (*Figure 2*).

The synthetic gene for the correspondingly redesigned left-handed monomeric bundle (LmRop) was overexpressed and subsequent nuclear magnetic resonance spectroscopy (NMR) experiments yielded sufficient distance constraints (NOEs) to prove that helix–helix packing in the topologically rearranged bundle was preserved, in complete agreement with the core hypothesis of protein folding (ref. 13 and S. C. Emery *et al.*, submitted). The results of this type of experiment indicate that the protein engineer has considerable latitude in re-engineering loop regions in this type of protein, not only in length and amino acid sequence, but also in topology (14).

The successful modifications of Rop protein illustrate the fact that simple rules can work well. However, many more protein engineering experiments of this type are required to elucidate fully the role of loops, surface, and core regions of globular proteins in general.

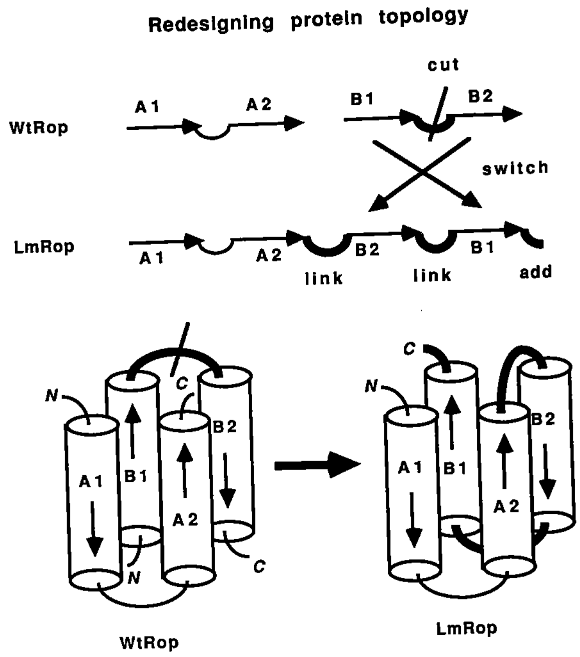


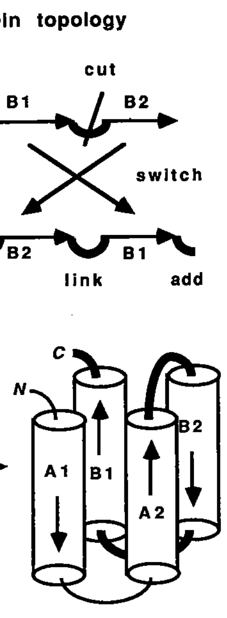
Figure 2. Re-engineering the topology of loop connections. Design steps from wild-type four-helix Rop dimer to the LmRop monomer involved deletion of one loop, addition of two new loops, and a C-terminal extension. Top: The orientation in space and packing of the four helices remains intact. Bottom: Two helical segments are switched at the gene level. In detail: wild-type Rop has two identical subunits, say A and B, and a total of four helices, say A1, A2 and B1, B2 (cylinders/arrows). To achieve the desired topology of a left-handed four-helix bundle which starts with helix A1, delete the loop between helices B1 and B2; switch the order (in the chain sense) of these helices from B1,B2 to B2,B1; connect by a new loop helices A2, B2 and by another loop helices B2, B1; complete helix B1 by addition of a few residues. Loops involved in these changes are emphasized (thick lines).

3. Designing proteins from scratch: example, a β-sheet sandwich

Designing proteins from scratch, rather than modifying naturally evolved ones, is a fascinating and challenging key problem of molecular engineering. Given a desired function and/or structure, the task is to design or calculate a suitable amino acid sequence that fulfils the design specifications. The goal in *de novo* protein design is to gain total control over the design process and to overcome the constraint of having to use existing amino acid sequences.

3.1 Build backbone model

The methods and tools used in the design process are rapidly evolving, so only an outline of some of the current techniques can be given here, illustrated



LmRop

ınnections. Design steps from wild-type olved deletion of one loop, addition of The orientation in space and packing of cal segments are switched at the gene bunits, say A and B, and a total of four o achieve the desired topology of a left- , delete the loop between helices B1 and e helices from B1,B2 to B2,B1; connect p helices B2, B1; complete helix B1 by e changes are emphasized (thick lines).

ratch: example, a

than modifying naturally evolved roblem of molecular engineering. the task is to design or calculate a e design specifications. The goal in rol over the design process and to existing amino acid sequences.

rocess are rapidly evolving, so only ues can be given here, illustrated

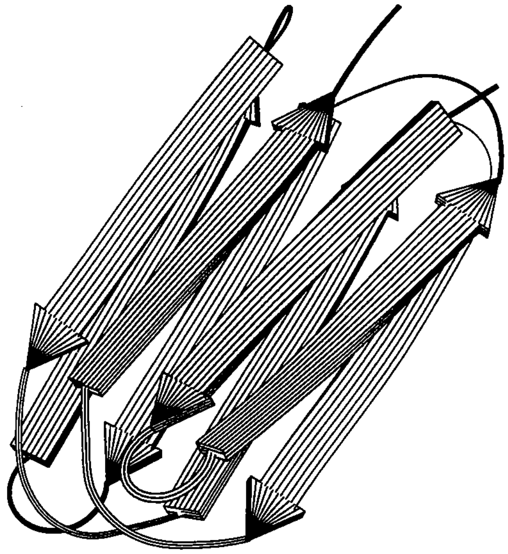


Figure 3. Ribbon cartoon of Shpilka, an antiparallel sandwich of two four-stranded β-sheets designed *de novo* as a scaffold on which to explore variations in the topology of loop connections (4, 5). The 96 residues of its amino acid sequence are: G {IKMTITLEL} GT {TKHSIPIET} GSPGG {IRMTITLEL} GT {TRHSVPVEG} GT {KHSVPVDT} GGGGG {VKWTVTMDL} GT {TRHSTPVDT} SGSPN {VRMTVTVD} LG, with β-strands enclosed in curly brackets. This sequence was designed in 2 weeks as an exercise. The design has recently been improved by A. V. Finkelstein (personal communication). The atomic co-ordinates of the corresponding model can be obtained from the electronic mail file server *netserv@embl-heidelberg.de* by sending the message *send proteindata:shpilka.brk.*

with a simple example (4, 5), a hypothetical protein called Shpilka (*Figure 3*). Suppose the design specification is for a globular protein consisting of a single-chain β-sheet sandwich with loops that can be varied flexibly. First, one makes a rough sketch of the structure: how many strands and which approximate strand lengths are to be used? Are the β-sheets to cross at approximately right angles or to run essentially parallel to each other? How strongly are the sheets twisted? Then, using molecular modelling software (15, 16, 17, 18) one builds a backbone model, best achieved by picking appropriate fragments from the data-base of known 3D structures so as not to deviate too strongly from well-tested structural units.

3.2 Choose sequence to fit structure

In the crucial next step, one chooses a sequence to fit the desired structure (see also Sections 6 and 7). This step is not yet automated, but instead is performed interactively, using empirical criteria based on physics, statistics, or data-base analysis (3, 4, 5, 19, 20). For choosing individual residues, statistical tables of residue preferences are available (21, 22, 23, 24). A very useful parameter set gives preferences for specific positions in secondary structure segments (see *Figures 5a, 6, 7*), e.g. near the beginning, middle, and end of strands or helices (21, 25, 26, 27, 28), turns (29, 30, 31), or loops (32). An additional useful distinction is between solvent-exposed and interior positions on helices and strands (see *Figures 4b, 5b*) (28, 33). Statistical parameters of a new type presented here are preferences for interfaces (see *Figure 11*), e.g. between two β-strands, between two β-sheets, between a β-sheet and an α-helix, or between a β-sheet and solvent (34). Specific examples of parameter usage are shown in *Figures 8* and *12*.

Particular physical properties of individual residues can also be exploited directly, without recourse to statistics. For example, Pro can be used to lock in local backbone conformation (the dihedral ϕ angle) and to exclude the formation of a backbone hydrogen bond (to nitrogen); Gly can be used for maximum backbone flexibility; and Trp may be chosen as a spectroscopic probe of folding. The full set of all types of rules for choosing amino acid types for a desired structural purpose cannot be given here.

For choosing pairs and clusters of residues, criteria include local complementarity of shape, size, and polarity, optimization of packing, hydrogen bonding, and charge–charge interactions. In this part of the design process frequent comparison of the current model with similar 3D constellations in the data-base of known structures is extremely useful (ref. 17 and K. Robson and C. Sander, unpublished). To assure good solvation properties, the overall amino acid composition of the apolar cores and solvent-exposed surfaces of globular proteins can be used as a guide.

3.3 Example: choosing a sequence for Shpilka

In the example of Shpilka (*Figure 3*), the following criteria were considered particularly important for the formation of the desired eight-stranded β-sandwich:

- no possibility to form α-helices
- eight regions with the capability to form β-strands
- stronger hydrophobicity in internal strands compared to edge strands
- well-defined non-polar and polar surfaces in all β-sheet regions
- edge strands that are restricted in backbone H-bond formation on one side
- close packing of side-chains in the interior

The designed sequence of Shpilka (*Figure 3*) has eight regions enriched in β-forming residues, separated by β-breaking connections. Non-polar and polar residues alternate to form well-defined inner and outer surfaces for β-strands, not suitable for α-helices. The interior face of the strands contains Val, Ile, and Met residues, the best β-formers. The interior face of edge

e available (21, 22, 23, 24). A very
for specific positions in secondary
.g. near the beginning, middle, and
8), turns (29, 30, 31), or loops (32).
veen solvent-exposed and interior
igures 4b, 5b) (28, 33). Statistical
are preferences for interfaces (see
etween two β-sheets, between a β-
and solvent (34). Specific examples
8 and *12*.

lual residues can also be exploited
r example, Pro can be used to lock
edral ϕ angle) and to exclude the
(to nitrogen); Gly can be used for
may be chosen as a spectroscopic
s of rules for choosing amino acid
not be given here.

idues, criteria include local com-
optimization of packing, hydrogen
In this part of the design process
l with similar 3D constellations in
nely useful (ref. 17 and K. Robson
od solvation properties, the overall
es and solvent-exposed surfaces of

e for Shpilka

following criteria were considered
of the desired eight-stranded β-

β-strands

ds compared to edge strands

s in all β-sheet regions

ne H-bond formation on one side

or

e 3) has eight regions enriched in
ing connections. Non-polar and
d inner and outer surfaces for β-
erior face of the strands contains
mers. The interior face of edge

strands contains β-forming threonines, which are partly polar, making contact with the hydrophobic core as well as with solvent. The solvent-accessible outer surfaces of all strands include charged groups that can form salt bridges. Prolines, which have no backbone NH group, are incorporated into the edge strands so that only one side of an edge strand can form a continuous hydrogen-bond net. Turns and loops are made of β-breaking residues; to help ensure unambiguous folding, they have minimal length, just enough to connect the strands in the designed structure. A Trp was included on the interior face of the sheet to serve as an experimental marker for correct folding. At the interior face of internal strands bulky side-chains alternate with small alanines in order to form complementary surfaces for close packing.

3.4 Avoid alternate folds

Simply adapting a sequence to a desired static scaffold is, however, not sufficient. How does one guarantee that the sequence actually allows a sufficient number of folding pathways from an unfolded to the final fold? That it is stable with respect to conformational fluctuations? That it will not equally well, or better, fold up into another globular shape? The current theories of protein folding do not yet provide satisfactory answers to these questions. However, empirical rules are being developed for some of these aspects. For example, in the design of the four-helix bundle, 'Felix', the number of residues in a loop was chosen so as to disfavour formation of an uninterrupted helix of excessive length (1). Also, in a helix bundle design, DeGrado *et al.* (35) attempted to make use of repulsive ion-pair interactions to force loop formation and prevent incorrect helix association, by insertion of a Pro–Arg–Arg sequence in a loop.

3.5 Refine three-dimensional model and sequence

After a first sequence has been chosen, one fully optimizes the 3D model by Monte Carlo or molecular dynamics exploration, followed by energy minimization. In our experience, Monte Carlo methods are vastly preferable for this purpose (36). Next, one evaluates the overall quality of the designed protein according to general criteria that were perhaps not used in the construction process, e.g. estimates of solvation energy (37), transfer energy (38), packing rules (39), or distribution of contacts around residue types (47). Such analysis will invariably lead to suggestions for point mutations in the model. Before proceeding to experiment, it is advisable to subject the sequence and 3D model to several or many cycles of sequence change, conformational optimization, and analysis.

3.6 Evaluate by structural experiments

With a designed sequence in hand, the protein or corresponding gene can be synthesized (see other chapters in this volume). For *in vivo* expression of synthetic genes, the main problem in some cases has been insufficient recovery of overexpressed product and purification into a soluble fraction. *In vitro* expression systems side-step some of these problems (O. Ptitsyn, personal communication), but suffer from low yields. Finally, spectroscopic methods, such as circular dichroism, Trp fluorescence, and 1D NMR are used as first tests of structure. Alternatively, in the case of a designed function that is strongly dependent on a specific 3D structure, a functional assay is used. The free energy of unfolding can be estimated by following a spectral signal as a function of the concentration of denaturant. A full determination of correct 3D structure requires X-ray crystallography or higher-dimensional (at least 2D) NMR spectroscopy, but so far the crystal structure of only one *de novo* designed protein has been completed (40). In general, several or many cycles of design, verification, and redesign are required to evolve a correctly folded protein.

4. Limitations

4.1 Difficulty in designing well-packed protein core

In spite of some successes in *de novo* protein design (1, 2), recent experiments show that currently used rules of protein design are incomplete or insufficient. None of the experimentally tested *de novo* designs has yielded a crystal or NMR structure. There is one exception (40), but in that case the 3D arrangement of α-helices in the crystal structure did not conform with the designers' intentions. It appears that so far none of the *de novo* designed proteins has achieved a unique fold comparable in quality to that of natural proteins, in spite of considerable stability against unfolding in some cases (2, 41). Both the physical theory of protein folding and the statistical and data-derived rules need to be improved.

4.2 Need for more powerful modelling computer tools

What tools are required in the near future? For *de novo* design, there is a clear need for more powerful constructive tools for going from a raw sketch to a first model. Example: given a sketch of a four-on-four β-sandwich, build a first backbone model in which the strands have typical twist angles and the intersheet angle is approximately as specified. Then there is a need for globally modifying structures. Example: increase the overall twist of a β-sheet while maintaining optimal packing and good loop geometry.

4.3 Need for improved energy calculations

There are two key problems in energy calculations. With current potentials, even in simulations of molecular dynamics including water, there is in many cases little correlation between correctness of structure and low energy; and it is still extremely time consuming to explore a significant fraction of conforma-

ses has been insufficient recovery
into a soluble fraction. *In vitro*
problems (O. Ptitsyn, personal
Finally, spectroscopic methods,
, and 1D NMR are used as first
e of a designed function that is
e, a functional assay is used. The
y following a spectral signal as a
. A full determination of correct
or higher-dimensional (at least
al structure of only one *de novo*
n general, several or many cycles
ired to evolve a correctly folded

cked protein core

design (1, 2), recent experiments
gn are incomplete or insufficient.
designs has yielded a crystal or
but in that case the 3D arrange-
not conform with the designers'
e *de novo* designed proteins has
ty to that of natural proteins, in
ling in some cases (2, 41). Both
statistical and data-derived rules

lling computer tools

' For *de novo* design, there is a
ols for going from a raw sketch to
four-on-four β-sandwich, build a
ave typical twist angles and the
fied. Then there is a need for
ease the overall twist of a β-sheet
l loop geometry.

culations

lations. With current potentials,
ncluding water, there is in many
f structure and low energy; and it
significant fraction of conforma-

tional space. So we urgently need more precise energetics that can reliably distinguish between correct and incorrect structures, and more efficient simulation tools for exploring many alternate protein conformations.

4.4 Novel protein structure motifs? Enzyme design?

A principle limitation of current work is the fact that protein design tends to stay within the limits of the structural motifs discovered in natural proteins. For example, all *de novo* designed proteins so far have been helical bundles, β-sheet sandwiches, or regular types of β–α proteins (1, 2). Protein engineers are generally not yet bold (or foolish) enough to attempt designs that deviate significantly from the architecture of natural proteins. Finally, the design principles for functional interfaces and for enzymatically active sites are still in their infancy (2).

5. Outlook

5.1 More rapid protein design cycle

Protein design methods are evolving rapidly. In our opinion, key advances will be: the development of more efficient cycles of design and verification, with the current bottleneck being, in many cases, protein expression and purification; and the development of better methods for the evaluation of models by empirical or energetic criteria, with the current bottleneck being proper treatment of solvation effects, coupled to electrostatics.

5.2 *De novo* design of structures

It appears that *de novo* protein design is much more difficult than the re-engineering of natural proteins. However, there is one way in which *de novo* design may turn out to be very simple. The sequences of natural proteins contain a superposition of information about structure, function, protein transport, average lifetime before degradation, etc., and, in addition, have been randomized within the bounds of selective pressure by mutational events. Sequences of *de novo* designed proteins, however, can be very simple, if the design goal is limited solely to the attainment of a stable structure. So the simplest application of the type of rules described here may soon lead to the successful design of properly folded and well-packed protein, once the major deficiency of current designs, i.e. insufficiently optimized ıesidue–residue interactions in the protein core and excessive flexibility of alternate packings, have been overcome.

5.3 Learn from natural evolution and evolution experiments

The two goals of protein design are to achieve a complete understanding of the folding of natural proteins and to develop design principles to achieve any physically possible target structure, such as enzyme scaffolds, self-assemblying structures, catalytically active sites, and structural regulators of activity. On the way to achieving these goals, both the study of natural protein families (9) and the planning, execution, and analysis of evolution experiments in the laboratory complement activities in protein design. All these ways need to be pursued, with evolution experiments being the most underdeveloped. We are in the early stages of protein design, in which skills are actively being developed at a high rate. 'The future is bright, but the road is tortuous.'

Acknowledgements

We thank François Colonna-Cesari for his important contributions several years ago to the definition and analysis of contact interfaces and John Priestle for his ribbon drawing program. Research was supported in part by the BRIDGE program of the Commission of the European Communities and the Human Frontiers Science Program.

enzyme scaffolds, self-
structural regulators of
h the study of natural
nd analysis of evolution
s in protein design. All
eriments being the most
in design, in which skills
re is bright, but the road

ant contributions several
terfaces and John Priestle
upported in part by the
ean Communities and the

Appendix

6. Tool: choosing amino acid types in helices, strands, and loops

6.1 Analysis of known protein structures

The data-base of known protein structures (6) is the basis for many empirical rules of protein design. In order to derive such rules, the very complicated set of 3D coordinates of a protein is first reduced to a simplified representation in terms of structurally characteristic regions. Statistical analysis can then be performed and preference rules can be derived for which types of amino acids to use in which regions. We now introduce two different ways of representing protein structure. The first (this section) refers to particular positions in secondary structure segments, the second (Section 7) refers to interactions, or contacts, between such segments and with water.

6.2 Representing protein structure: positions in secondary structure segments

The classical representation of 3D structure is in terms of α-helices, β-strands, hydrogen-bonded turns, and loops (42). A more refined description, presented here, 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 (*Figures 4–8*) (28). In an even more refined statistical analysis (not presented here), preferences for pairs, triplets, tetrapeptides, and pentapeptides in certain structural states can be derived (28).

6.3 Definition of sequence–structure preference parameters

How are preference parameters defined? In a given structural state (S), the preference parameter for a residue type (R) is calculated from the number of observed cases of R in S, N(R,S), as

$$\mathrm{pref(R,S)} = \mathrm{ld}\left(\frac{N(\mathrm{R,S})\ N}{N(\mathrm{R})\ N(\mathrm{S})}\right)$$

where

$$N(\mathrm{R}) = \sum_{\mathrm{S}} N(\mathrm{R,S}),\ N(\mathrm{S}) = \sum_{\mathrm{R}} N(\mathrm{R,S}),\ N = \sum_{\mathrm{R,S}} N(\mathrm{R,S})$$

and ld is the logarithm base 2, so that pref(R,S) is in information units called bits. The expression in parentheses is the ratio of the number of observed

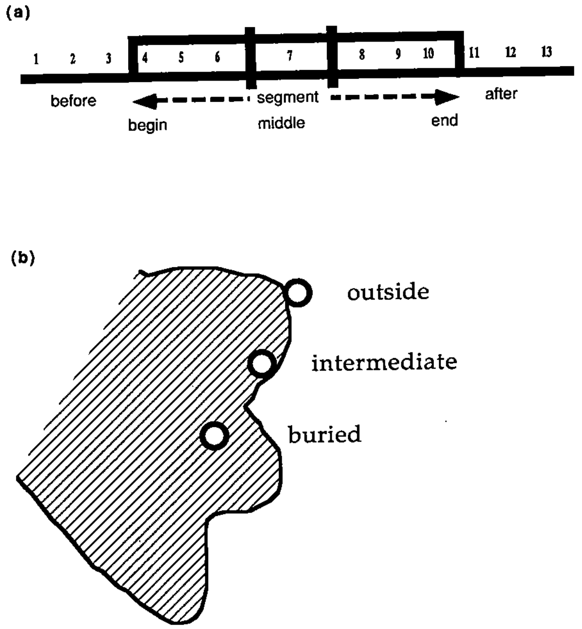
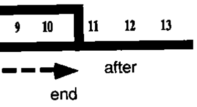


Figure 4. Definitions of positions on the secondary structure segment used in *Figures 5–8*. (a) Definition of linear positions along the secondary structure segment. Three positions each are distinguished on either side of the segment boundaries; positions in the middle of the segment are lumped together in one class. Positions are numbered 1–13 for convenience. Note that position 4 is at the beginning (N-terminus) of a segment and position 10 is at the segment end (C-terminus). (b) Definition of outside/inside/buried positions on the secondary structure segment. The solvent-accessible surface area is used as a basis for the classification. Residues with large accessibility are classified as outside, those with intermediate values as inside, and the rest as buried. The notation used is: 'exposed to solvent' (O, for outside); 'intermediate' (I, for intermediate or inside); 'removed from solvent' (B, for buried). Residue accessibility is defined as the ratio of the actual accessible surface area and its maximal value, in per cent, as tabulated by Baumann *et al.* (37). The cut-offs used in the definition vary with secondary structure type and were determined on the basis of residue accessibility histograms (28). In terms of the percentage of maximal surface exposed to solvent, the cut-offs are as follows, for all residue types:

		B		I		O
β-strand	E 0%	– buried	– 5%	– interm.–	15%–	outside– 100%
loop	L 0%	– buried	– 35%–	interm.–	60%–	outside– 100%
α-helix	H 0%	– buried	– 25%–	interm.–	50%–	outside– 100%

These cut-off values reflect the fact that the average exposure is largest for loops, followed by helices and strands (28). Such cut-offs are, of course, somewhat arbitrary. Note that solvent exposure can be quantified in terms of interatomic contacts without cut-offs (see *Figure 9*).



ːside

ıediate

ructure segment used in *Figures 5–8.* structure segment. Three positions boundaries; positions in the middle . Positions are numbered 1–13 for ing (N-terminus) of a segment and Definition of outside/inside/buried lvent-accessible surface area is used ccessibility are classified as outside, est as buried. The notation used is: ' (I, for intermediate or inside); 're- sibility is defined as the ratio of the alue, in per cent, as tabulated by n vary with secondary structure type bility histograms (28). In terms of the , the cut-offs are as follows, for all

O
.– 15%– outside– 100%
.– 60%– outside– 100%
.– 50%– outside– 100%

xposure is largest for loops, followed urse, somewhat arbitrary. Note that atomic contacts without cut-offs (see

(a) preference parameter of a single amino acid for a secondary structure type

S	V	L	I	M	F	W	Y	G	A	P	S	T	C	H	R	K	O	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

(b) preference parameter of a single amino acid 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	O	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

Figure 5. Preference parameters for amino acid types in various structural states, going beyond the simple description in terms of helices, strands, and loops. These parameters are useful for choosing residues for particular structural positions in protein design. They can also be used in protein structure prediction from sequence (28). The parameters are based on a non-redundant set of 38 selected high-resolution protein structures (list not shown). In order to improve the effective data-base size, for each of the 38 proteins all clearly homologous (10 percentage points above the threshold for structural homology (9), i.e. more than 35% identical residues for alignment lengths longer than 80 residues) protein sequences are also used, counting each distinct sequence–structure pair (R, S) only once at each sequence position, for a total of 52 426 unique residue occurrences in the 38 protein families. In this way, refined distinctions of structural states become amenable to statistical analysis. (a) Secondary structure preference parameters (set S3). Secondary structure can be S = E, H, L where E = β-strand; L = loop; H = α-helix (see *Figure 4b*). Parameters are information values in bits. For example, Val in β-strands has a preference of 0.6 bits, i.e. the odds are $2^{0.6}$ = 1.52 to 1 to observe Val in β-strands. (b) Combined accessibility and secondary structure preference parameters (set SX9). Secondary structure in S = E, L, H as above. The three accessibility states are: 'exposed to solvent' (X = O for outside), 'intermediate' (X = I for intermediate or inside), and 'removed from solvent' (X = B for buried). Numerical example: pref(Met,H,B) = 0.8, but pref(Met,H,O) = 0.7, showing that Met is preferentially found at the solvent-protected side of helices rather than on the surface. Note that these preferences are clearer than the average preference for the helical state, pref(Met,H) = 0.2. Averaging obviously results in loss of information and more refined distinctions of structural states enhance the information content of preference parameters. Notation is, for example, EB, β-strand, buried; LO, loop, outside; HI, helix, intermediate.

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	O	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

loop

S	P	V	L	I	M	F	W	Y	G	A	P	S	T	C	H	R	K	O	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

α-helix

S	P	V	L	I	M	F	W	Y	G	A	P	S	T	C	H	R	K	O	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

Figure 6. Combined positional and secondary structure preference parameters (set SP39). Positions P = 1–13 are defined in *Figure 4a*, secondary structure types S = E, L, H in *Figure 5a*. Notation is, for example, E1, third residue position before the beginning of a β-strand; L6, third residue position in a loop; H12, second residue after end of a helix.

a secondary structure segment

C	H	R	K	O	E	N	D
-0.6	0.3	0.3	0.1	-0.1	0.0	0.4	0.3
-0.7	0.1	0.0	0.0	0.2	0.2	0.4	0.6
-0.3	0.2	0.0	0.1	0.1	-0.2	0.2	0.1
0.0	-0.1	0.0	-0.1	0.3	-0.3	-0.7	-0.9
0.1	-0.5	-0.3	-0.5	-0.4	-0.5	-0.6	-1.0
0.3	0.2	0.3	-0.7	-0.3	-0.7	-1.1	-1.6
-0.4	-0.2	-0.3	-0.3	0.0	-0.3	-0.5	-0.7
0.0	-0.6	-0.5	-0.7	-0.5	-0.7	-0.8	-0.9
-0.3	-0.2	0.0	-0.4	-0.1	-0.5	-0.7	-0.9
0.0	-0.2	0.0	-0.4	-0.4	-0.3	-0.3	0.0
-0.2	0.1	-0.1	-0.3	-0.4	-0.1	0.3	0.1
-0.8	-0.1	0.1	0.1	0.0	0.1	0.1	0.3
-0.1	0.2	0.1	0.1	-0.1	0.2	0.0	0.4

	C	H	R	K	O	E	N	D
	-0.2	0.0	-0.1	-0.2	-0.3	-0.2	-0.5	-0.6
0	0.4	-0.2	0.0	0.0	0.1	0.0	-0.3	-0.5
1	0.3	0.0	-0.1	-0.1	0.0	-0.2	-0.1	-0.1
0	0.1	0.1	0.0	-0.2	-0.3	-0.1	0.3	0.0
1	-1.4	0.0	0.2	0.2	0.2	0.0	0.1	0.2
0	-0.2	0.1	0.3	0.3	0.0	0.1	0.2	0.3
0	0.2	-0.1	0.2	0.0	0.1	0.0	0.1	0.2
4	0.0	0.3	0.4	0.2	0.0	0.1	0.4	0.4
1	-0.1	0.0	0.1	0.1	0.0	0.0	0.2	0.4
1	-0.1	0.2	-0.1	0.1	0.0	-0.1	0.5	0.3
1	0.0	-0.2	-0.1	0.0	0.2	-0.1	-0.6	-0.5
0	0.1	-0.3	-0.3	-0.2	-0.1	0.3	-0.2	0.1
1	0.2	0.0	-0.2	-0.4	-0.1	0.0	-0.2	-0.1

T	C	H	R	K	O	E	N	D
.1	0.3	-0.1	-0.1	0.3	-0.1	-0.2	0.0	-0.3
.2	-0.1	-0.2	-0.2	-0.2	-0.4	-0.7	-0.4	-0.1
.4	0.4	0.2	-0.2	-0.2	-0.4	-0.1	0.7	0.7
.1	0.0	-0.5	-0.4	0.1	-0.2	0.2	-0.4	-0.1
.0	0.3	-0.1	-0.4	0.0	0.1	0.8	0.2	0.7
.2	-0.7	0.0	-0.7	-0.2	0.6	1.0	0.1	0.8
.2	0.3	-0.1	0.0	0.2	0.0	0.2	-0.2	-0.4
.3	-0.8	0.4	0.2	0.4	-0.1	0.0	-0.2	-0.5
.2	0.0	0.0	-0.1	0.4	0.3	0.2	0.0	-0.4
.3	0.6	0.2	0.1	0.3	0.3	0.0	0.2	-0.1
.3	0.4	0.1	0.1	0.0	-0.1	-0.2	0.4	-0.1
.1	-0.4	0.0	-0.1	0.4	0.1	-0.3	0.3	0.0
.1	0.0	0.1	0.0	0.4	0.1	0.0	0.2	0.2

ure preference parameters (set SP39).
ndary structure types S = E, L, H in
position before the beginning of a β-
cond residue after end of a helix.

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	O	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

loop

S X P	V	L	I	M	F	W	Y	G	A	P	S	T	C	H	R	K	O	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

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	-0.4	-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

Figure 7. Combined positional, accessibility, and secondary structure, preference parameters (set SXP117). This is the most sophisticated set of parameters reported here, using the finest distinction in structural states. Positions 10, 11, 12, 13 are labelled 0, 1, 2, 3 (bottom of lists). The many interesting and strong preferences and antipreferences have not yet been fully interpreted. Note, for example, that the tendency for Phe in solvent-protected positions on a helix increases from pref(Phe,H,B,4) = +0.2 at the N-terminus to pref(Phe,H,B,10) = +0.8 at the C-terminus; or, that the tendency for Pro in solvent-exposed positions on a helix decreases from pref(Pro,H,O,4) = +1.5 at the N-terminus to pref(Pro,H,O,10) = −2.6. Comparison with the collapsed tables (*Figures 5* and *6*) is an aid in interpreting the underlying effects. Notation is, for example, EB1, third residue position before the beginning of a β-strand, buried; LO6, third residue position in a loop, outside; HI12, second residue position after the end of a helix, intermediate solvent exposure.

C	H	R	K	Q	E	N	D
.8	-0.5	-0.9	-1.0	-1.7	-1.4	-0.7	-1.7
.2	-0.5	-1.3	-1.8	-0.9	-1.2	-1.3	-0.8
.8	0.7	-0.6	-1.0	-1.0	-0.5	0.6	0.2
.5	-0.3	-0.8	-0.4	-0.5	-0.4	-1.1	-1.3
.9	0.1	-1.3	-0.7	-0.7	-0.3	0.0	-0.6
.1	-0.2	-1.0	-1.0	0.1	0.4	-0.2	0.7
.7	-0.3	-0.7	-0.7	-0.7	-0.6	-0.9	-1.4
.7	-0.1	-0.7	-0.6	-1.2	-1.8	-0.7	-3.0
.1	-1.1	-1.0	-1.1	-0.4	-1.0	-0.8	-1.6
.8	0.1	-1.4	-1.2	-0.6	-1.3	-0.6	-1.2
.1	0.2	-0.3	-1.2	-0.6	-1.1	-0.3	-0.3
.3	-0.4	-0.2	-0.4	-0.2	-0.9	-0.2	-0.2
1.3	0.0	-0.4	-0.5	-0.7	-0.9	-0.4	-0.6
0.6	-0.1	0.3	0.7	0.5	0.2	0.2	0.0
0.1	0.0	0.5	0.3	-0.1	-0.1	-0.4	-0.1
0.1	-0.2	-0.1	0.0	0.0	0.1	0.7	0.8
0.0	-0.6	-0.2	0.3	-0.1	0.2	-0.1	0.2
-0.2	0.1	-0.4	0.2	0.3	0.7	-0.1	0.5
-1.7	0.2	-0.5	0.3	1.1	1.4	0.3	0.9
-0.8	-0.2	0.7	0.9	0.6	0.8	0.2	0.2
-0.3	1.0	0.8	0.7	0.4	0.5	-0.6	-0.3
-0.6	0.0	0.5	0.6	0.4	0.1	-0.1	-0.3
-0.6	0.4	0.4	0.4	0.6	-0.1	-0.2	0.0
-0.1	-0.1	0.5	0.4	0.5	0.2	0.1	-0.1
0.0	0.3	0.1	0.6	0.0	-0.2	0.4	-0.2
-0.5	0.0	0.1	0.7	-0.1	0.1	0.0	0.0
-1.3	0.3	0.0	0.8	0.2	0.3	0.4	0.1
-1.0	0.2	0.3	0.6	-0.1	-0.5	0.4	0.5
-0.6	-0.2	0.0	0.4	-0.4	0.2	0.8	1.0
-1.1	-0.7	-0.3	0.2	0.1	0.6	-0.1	0.5
-1.1	-0.3	-0.2	0.1	0.3	1.1	0.3	1.0
-1.8	-0.2	-0.4	-0.1	0.2	1.0	0.4	0.8
-0.9	0.4	0.4	0.8	0.5	0.6	0.3	0.2
-0.9	0.3	0.5	1.0	0.3	0.7	0.5	0.6
-1.4	0.4	-0.1	0.8	0.5	0.7	0.3	0.1
-1.0	0.0	0.4	0.8	0.6	0.6	0.8	0.3
-1.4	-0.1	0.0	0.4	0.0	0.2	0.9	0.0
-1.5	0.0	-0.1	0.7	0.2	-0.1	0.4	0.2
-2.2	0.2	0.0	0.6	0.4	0.3	0.5	0.5

econdary structure, preference par-
ed set of parameters reported here,
ons 10, 11, 12, 13 are labelled 0, 1, 2, 3
references and antipreferences have
hat the tendency for Phe in solvent-
ne,H,B,4) = +0.2 at the N-terminus to
at the tendency for Pro in solvent-
ro,H,O,4) = +1.5 at the N-terminus to
osed tables (*Figures 5* and *6*) is an aid
r example, EB1, third residue position
rd residue position in a loop, outside;
lix, intermediate solvent exposure.

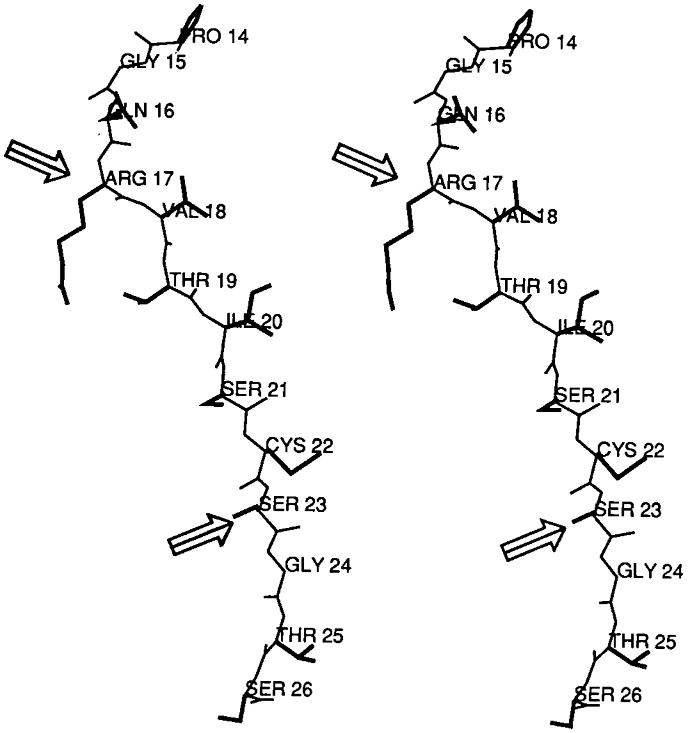


Figure 8. Stereo view. Example of how single residue preference parameters are combined to evaluate an amino acid sequence in a given structural context, using position (P) dependent and accessibility (X) dependent secondary structure (S) preference parameters from *Figures 5–7*. This 13-residue fragment from a known structure, contains a seven-residue β-strand (between arrows) and has the sequence PGQRVTISCSGTS (β-strand residues underlined). To evaluate the suitability of the sequence, look up the single residue preferences in, say, *Figure 7* and add them up along the fragment. For example, the first position in the fragment (labelled Pro14) has the structural state, as derived from the atomic coordinates, of S=E, X=O, P=1 (segment is β-strand; high solvent accessibility; position -3 before the segment) and Pro has a preference of +0.4 bits in this state; the fourth position in the fragment (labelled Arg17) is in the state S=E, X=O, P=4, with pref(Arg,E,O,4) = +0.3 bits; similarly pref(Val,E,I,5) = +0.7, pref(Ile,I,B,7) = +1.1, pref(Cys,E,O,9) = −0.9 (negative value—not preferred), and so on. Using 'PXS' preference parameters, the sum over the 13 residue is +3.3 bits, a relatively high value, corresponding to a probability ratio of observed/expected of $2^{3.3}$ = 9.85, i.e. the odds are 10 to 1 to find this sequence in this structural state. The fragment is residues 14–26 from an immunoglobulin Fab fragment, data-set 2FB4 in the Protein Data Bank. Cys22 is involved in an intramolecular disulphide bond. The entire strand has state S = (LLLEEEEEEELLL), X = (OOOOIOBOOOOOO), and P = (1–13).

cases, $N(\mathrm{R,S})$, to the number of expected cases, $E(\mathrm{R,S}) = N(\mathrm{R})N(\mathrm{S})/N$, assuming a random model. A ratio of 2.0 (pref = 1.0) means that the observation is made twice as frequently as the random model would suggest. A value of pref = −1.0 indicates that the number of observed cases is half as large as expected from random. A numerical example: N(Pro, helix, first residue) = 82, N(Pro) = 2037, N(helix, first residue) = 893, N(all) = 52 426 gives pref(Pro, helix, first residue) = +1.2.

7. Tool: choosing amino acid types in interfaces

7.1 Representing protein structure: interface states

Another description of protein structure, more sophisticated than the standard helix–strand–turn–loop classification, is in terms of contact interfaces. The contact view was developed in collaboration with F. Colonna-Cesari in 1984–85 and is published here for the first time. This type of description is motivated by observations that residue preferences vary strongly on different sides of secondary structure segments (24). For example, the amino acid composition of solvent-exposed faces of helices is very different compared to that of helix–sheet interfaces (*Figure 11a*). The interface definition described here 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 (*Figure 9*). As protein–solvent interaction appears to be the physically dominant effect in the folding of globular proteins, water contacts are explicitly taken into account.

7.2 Contacts between protein atoms

Contacts are best represented in terms of atom–atom interactions. A very simple form, fairly insensitive to errors in atomic coordinates, is that of a 'linear-square well'. A single contact event is counted with a strength of 1.0 when two atoms are 'touching' (interatomic distance less or equal to the sum of the van der Waals radii). With increasing interatomic distance the strength decreases linearly and reaches 0.0 when a water molecule just fits between two atoms. No contacts are counted when the interatomic distance is larger than the sum of the van der Waals radii plus the diameter of a water molecule. The contact strength of a residue is the sum of the contact strengths over all of its atoms.

7.3 Contacts of protein atoms with water molecules

In order to account for protein–water contacts, one is faced with the difficulty that the positions of (most) water molecules are not known and, in any event, vary strongly with time. In a reasonable hydration shell model the number of water molecules in contact with the protein is proportional to the volume of a single hydration shell. In terms of the classical definition of solvent-

cases, $E(R,S) = N(R)N(S)/N$,
0 (pref = 1.0) means that the
ne random model would suggest.
mber of observed cases is half as
cal example: N(Pro, helix, first
residue) = 893, N(all) = 52 426

ypes in interfaces

e: interface states

nore sophisticated than the stan-
is in terms of contact interfaces.
ration with F. Colonna-Cesari in
time. This type of description is
erences vary strongly on different
). For example, the amino acid
ices is very different compared to
The interface definition described
secondary structure of a central
in distance of its contact partners
pears to be the physically domin-
ins, water contacts are explicitly

ns

atom–atom interactions. A very
atomic coordinates, is that of a
is counted with a strength of 1.0
distance less or equal to the sum
interatomic distance the strength
water molecule just fits between
the interatomic distance is larger
the diameter of a water molecule.
of the contact strengths over all of

h water molecules

cts, one is faced with the difficulty
are not known and, in any event,
ydration shell model the number
ein is proportional to the volume
ne classical definition of solvent-

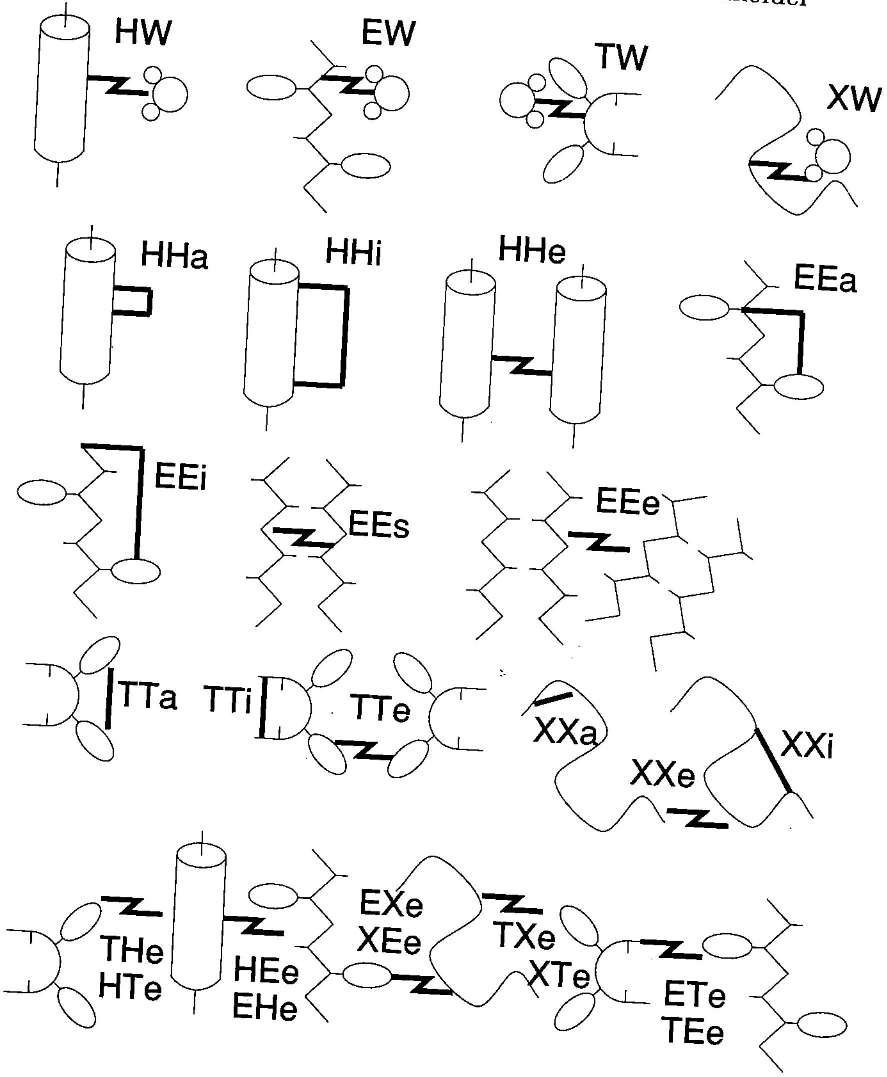


Figure 9. Schematic representation of the interface types used for the derivation of contact preferences given in *Figure 11*. Cylinders are helices (H), zig-zags are β-strands (E), semi-circles are hydrogen-bonded turns (T), smooth curves are loops (X), two small and one big circle is a water molecule (W), and ellipses are amino acid side-chains. Contacts are shown as bold face lines (squashed Z shapes or rectangular brackets). Interfaces are, for example, HHe for helix–helix, HEe for sheet–sheet. HW for helix–water, etc. For details of notation, see *Figure 11.*

#PID	C	SIZ	RES	%H	%B	%BP	%BA	SID	ORIGIN	PROTEIN_NAME
351C	_	82	1.6	50	4	0	100	C551$PSEAE	PSEUDOMONAS AERUGINOSA	CYTOCHROME C 551
256B	Ā	106	1.4	79	0	0	0	C562$ECOLI	ESCHERICHIA COLI	CYTOCHROME B 562
8ADH	_	374	2.4	28	24	45	55	ADHE$HORSE	EQUUS CABALLUS	ALCOHOL DEHYDROGENASE
8ATC	Ā	310	2.5	40	15	100	0	PYRB$ECOLI	ESCHERICHIA COLI	ASPARTATE CARBAMOYLTRANSFERASE (ASPARTATE TRANSCARBAMYLASE)
8ATC	B	146	2.5	15	34	1	98	PYRI$ECOLI	ESCHERICHIA COLI	ASPARTATE CARBAMOYLTRANSFERASE (ASPARTATE TRANSCARBAMYLASE)
2AZA	A	129	1.8	16	35	36	63	AZUR$ALCDE	ALCALIGENES DENITRIFICANS	AZURIN
3B5C	_	85	1.5	31	23	25	75	CYB5$BOVIN	BOS TAURUS	CYTOCHROME B 5
3BLM	_	257	2.0	42	17	0	100	BLAC$STAAU	STAPHYLOCOCCUS AUREUS	BETA-LACTAMASE
2CA2	_	256	1.9	16	30	23	76	CAH2$HUMAN	HOMO SAPIENS	CARBONIC ANHYDRASE II (CARBONATE DEHYDRATASE)
1CCR	_	111	1.5	42	1	0	100	CYC$ORYSA	ORYZA SATIVA	CYTOCHROME C
2CCY	Ā	127	1.7	74	1	0	100	CYCP$RHOMO	RHODOSPIRILLUM MOLISCHIANUM	CYTOCHROME C'
1CD4	_	173	2.3	5	41	11	88	CD4$HUMAN	HOMO SAPIENS, recombinant	T-CELL SURFACE GLYCOPROTEIN CD4 (N-TERMINAL FRAGMENT)
3CLA	_	213	1.8	29	28	23	76	CAT3$ECOLI	ESCHERICHIA COLI, engineered	CHLORAMPHENICOL ACETYLTRANSFERASE TYPE III
5CPA	_	307	1.5	38	16	63	36	CBPA$BOVIN	BOS TAURUS	CARBOXYPEPTIDASE A
2CPP	_	405	1.6	51	10	11	88	CPXA$PSEPU	PSEUDOMONAS PUTIDA	CYTOCHROME P450CAM (CAMPHOR MONOOXYGENASE)
4CPV	_	108	1.5	56	1	0	100	PRVB$CYPCA	CYPRINUS CARPIO	CALCIUM-BINDING PARVALBUMIN
1CSE	Ē	274	1.2	30	20	73	26	SUBT$BACLI	BACILLUS SUBTILIS	SUBTILISIN
1CSE	I	63	1.2	22	33	44	55	ICIC$HIRME	HIRUDO MEDICINALIS	EGLIN-C
1CTF	_	68	1.7	55	26	0	100	RL7$ECOLI	ESCHERICHIA COLI	50S RIBOSOMAL PROTEIN L7/L12 (C-TERMINAL DOMAIN)
2CYP	_	293	1.7	50	7	8	91	CCPR$YEAST	SACCHAROMYCES CEREVISIAE	CYTOCHROME C PEROXIDASE
8DFR	_	186	1.7	23	33	57	42	DYR$CHICK	GALLUS GALLUS	DIHYDROFOLATE REDUCTASE
1ECN	_	136	1.4	75	0	0	0	GLB3$CHITH	CHIRONOMOUS THUMMI)	HEMOGLOBIN (ERYTHROCRUORIN) (FRACTION III)
2ER7	Ē	330	1.6	11	45	13	86	CARP$CRYPA	ENDOTHIA PARASITICA	ASPARTIC PROTEINASE (ENDOTHIAPEPSIN)
4FD1	_	106	1.9	33	14	0	100	FER1$AZOVI	AZOTOBACTER VINELANDII	FERREDOXIN
4FXN	_	138	1.8	36	22	95	4	FLAV$CLOSP	CLOSTRIDIUM MP	FLAVODOXIN
3GAP	Ā	208	2.5	30	14	0	100	CRP$ECOLI	ESCHERICHIA COLI	CATABOLITE GENE ACTIVATOR PROTEIN
2GBP	_	309	1.9	43	19	90	10	DGAL$ECOLI	ESCHERICHIA COLI	D-GALACTOSE/D-GLUCOSE BINDING PROTEIN
1GCR	_	174	1.6	7	46	0	100	CRGB$BOVIN	BOS TAURUS	CRYSTALLIN GAMMA-II
1GD1	Ō	334	1.8	29	29	52	47	G3P$BACST	BACILLUS STEAROTHERMOPHILUS	D-GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE
1GOX	_	350	2.0	44	13	78	21	2HAO$SPIOL	SPINACIA OLERACEA	GLYCOLATE OXIDASE
1GP1	Ā	183	2.0	32	18	47	52	GSHP$BOVIN	BOS TAURUS	GLUTATHIONE PEROXIDASE
2HLA	B	99	2.6	0	49	0	100	HA1H$HUMAN	HOMO SAPIENS	HISTOCOMPATIBILITY CLASS I ANTIGEN
1HOE	_	74	2.0	0	48	0	100	IAA$STRTE	STREPTOMYCES TENDAE	ALPHA-AMYLASE INHIBITOR
1I1B	_	151	2.0	5	47	0	100	IL1B$HUMAN	HOMO SAPIENS, recombinant	INTERLEUKIN-1 BETA
4ICD	_	414	2.5	39	18	52	47	IDH$ECOLI	ESCHERICHIA COLI	ISOCITRATE DEHYDROGENASE
1IL8	Ā	71	NMR	26	25	0	100	IL8$HUMAN	HOMO SAPIENS, recombinant	INTERLEUKIN 8
1L13	_	164	1.7	64	9	0	100	LYCV$BPT4	BACTERIOPHAGE T4, mutant	LYSOZYME
6LDH	_	329	2.0	43	17	51	48	LDHM$SQUAC	SQUALUS ACANTHIAS	LACTATE DEHYDROGENASE
2LIV	_	344	2.4	44	19	73	26	LIVJ$ECOLI	ESCHERICHIA COLI	LEU/ILE/VAL-BINDING PROTEIN
2LTN	Ā	181	1.7	1	43	0	100	LEC$PEA	PISUM SATIVUM, recombinant	LECTIN
2LTN	B	47	1.7	8	63	0	100	LEC$PEA	PISUM SATIVUM, recombinant	LECTIN
1LZ1	_	130	1.5	39	12	11	88	LYC$HUMAN	HOMO SAPIENS	LYSOZYME
1MBD	_	153	1.4	77	0	0	0	MYG$PHYCA	PHYSETER CATODON	MYOGLOBIN
2MHR	_	118	1.7	70	0	0	0	HEMM$THEZO	THEMISTE ZOSTERICOLA	MYOHEMERYTHRIN
2PAB	Ā	114	1.8	7	51	16	83	TTHY$HUMAN	HOMO SAPIENS	PREALBUMIN
1PAZ	_	120	1.6	16	37	35	64	AZUP$ALCFA	ALCALIGENES FAECALIS	PSEUDOAZURIN
4PTP	_	223	1.3	10	34	2	97	TRYP$BOVIN	BOS TAURUS	BETA TRYPSIN
1R69	_	63	2.0	63	0	0	0	RPC1$BP434	PHAGE 434	434 REPRESSOR (N-TERMINAL DOMAIN)
1RHD	_	293	2.5	29	13	87	12	THTR$BOVIN	BOS TAURUS	RHODANESE
7RSA	_	124	1.3	20	35	3	96	RNP$BOVIN	BOS TAURUS	RIBONUCLEASE A
2RSP	Ā	115	2.0	5	41	17	82	GAG$RSVP	ROUS SARCOMA VIRUS	RSV PROTEASE
5RXN	_	54	1.2	16	22	0	100	RUBR$CLOPA	CLOSTRIDIUM PASTEURIANUM	RUBREDOXIN
2SGA	_	181	1.5	9	55	6	93	PRTA$STRGR	STREPTOMYCES GRISEUS	PROTEINASE A
4SGB	Ī	51	2.1	0	29	11	88	IPR2$SOLTU	SOLANUM TUBEROSUM	SERINE PROTEINASE B INHIBITOR PCI-I
2SNS	_	141	1.5	20	22	15	85	NUC$STAAU	STAPHYLOCOCCUS AUREUS	STAPHYLOCOCCAL NUCLEASE
2SOD	Ō	151	2.0	1	42	2	97	SODC$BOVIN	BOS TAURUS	CU,ZN SUPEROXIDE DISMUTASE
2SSI	_	107	2.6	15	28	5	95	ISUB$STRAO	STREPTOMYCES ALBOGRISEOLUS	SUBTILISIN INHIBITOR
2STV	_	184	2.5	11	47	1	98	COAT$STNV	SATELLITE TOBACCO NECROSIS VIRUS	COAT PROTEIN
2TMN	Ē	316	1.6	40	17	26	73	THER$BACTH	BACILLUS THERMOPROTEOLYTICUS	THERMOLYSIN
1TNF	A	152	2.6	1	44	0	100	TNFA$HUMAN	HOMO SAPIENS, recombinant	TUMOR NECROSIS FACTOR-ALPHA
2TS1	_	317	2.3	54	10	85	14	SYY$BACST	BACILLUS STEAROTHERMOPHILUS	TYROSYL-tRNA SYNTHETASE
1UBQ	_	76	1.8	23	34	25	75	UBIQ$HUMAN	HOMO SAPIENS	UBIQUITIN
1UTG	_	70	1.3	75	0	0	0	UTER$RABIT	ORYCTOLAGUS CUNICULUS	UTEROGLOBIN
2WRP	R̄	104	1.6	78	0	0	0	TRPR$ECOLI	ESCHERICHIA COLI	TRP REPRESSOR
1WSY	A	248	2.5	50	13	100	0	TRPB$SALTY	SALMONELLA TYPHIMURIUM	TRYPTOPHAN SYNTHASE
4XIA	A	393	2.3	47	10	85	14	XYLA$ARTS7	ARTHROBACTER SP	D-XYLOSE ISOMERASE
1YPI	A	247	1.9	43	17	96	3	TPIS$YEAST	SACCHAROMYCES CEREVISIAE	TRIOSE PHOSPHATE ISOMERASE

Figure 10. List of 67 protein chains used to derive the contact parameters given in *Figure 11*, with a total of 12 460 residues. The list was produced with the aid of an algorithm that selects a representative set of proteins from a protein data-bank, assuring that no two proteins have higher than 30% sequence identity after optimal alignment (46). Membrane proteins, very small proteins, proteins with a large number of disulphide bonds, proteins with many heteroatoms in the data-set, with a low percentage of secondary structure, or with nominal resolution not better than 2.7 Å were excluded. PID is the Protein Data Bank identifier (6); C, the chain index; SIZ, the number of residues in the chain; RES, the nominal crystallographic resolution (or 'NMR'); %H, the number of hydrogen bonds involved in (*i*,*i*+4) and (*i*,*i*+3) type H-bonds (helices and turns) per 100 residues (42); %B, the same in β-structure (parallel and antiparallel bridges); %BP and %BA, the percentage thereof in parallel or antiparallel bridges (%BA+%BP = 100 if there is any β-structure). SID is the corresponding SWISS-PROT sequence identifier (8) and ORIGIN the species origin.

```
ME
------
 C 551
 B 562
EHYDROGENASE
 CARBAMOYLTRANSFERASE (ASPARTATE TRANSCARBAMYLASE)
 CARBAMOYLTRANSFERASE (ASPARTATE TRANSCARBAMYLASE)

 B 5
AMASE
ANHYDRASE II (CARBONATE DEHYDRATASE)
 C
 C'
RFACE GLYCOPROTEIN CD4 (N-TERMINAL FRAGMENT)
ENICOL ACETYLTRANSFERASE TYPE III
PTIDASE A
 P450CAM (CAMPHOR MONOOXYGENASE)
INDING PARVALBUMIN
N

OMAL PROTEIN L7/L12 (C-TERMINAL DOMAIN)
E C PEROXIDASE
LATE REDUCTASE
N (ERYTHROCRUORIN) (FRACTION III)
PROTEINASE (ENDOTHIAPEPSIN)
N
N
E GENE ACTIVATOR PROTEIN
SE/D-GLUCOSE BINDING PROTEIN
N GAMMA-II
LDEHYDE-3-PHOSPHATE DEHYDROGENASE
 OXIDASE
NE PEROXIDASE
PATIBILITY CLASS I ANTIGEN
LASE INHIBITOR
KIN-1 BETA
TE DEHYDROGENASE
KIN 8

DEHYDROGENASE
VAL-BINDING PROTEIN



N
YTHRIN
IN
URIN
PSIN
ESSOR (N-TERMINAL DOMAIN)
E
EASE A
EASE
IN
SE A
ROTEINASE B INHIBITOR PCI-I
COCCAL NUCLEASE
PEROXIDE DISMUTASE
IN INHIBITOR
TEIN
SIN
CROSIS FACTOR-ALPHA
-tRNA SYNTHETASE
N
OBIN
RESSOR
HAN SYNTHASE
 ISOMERASE
PHOSPHATE ISOMERASE
```

contact parameters given in *Figure*
ed with the aid of an algorithm that
in data-bank, assuring that no two
optimal alignment (46). Membrane
mber of disulphide bonds, proteins
rcentage of secondary structure, or
cluded. PID is the Protein Data Bank
of residues in the chain; RES, the
, the number of hydrogen bonds
nd turns) per 100 residues (42); %B,
ges); %BP and %BA, the percentage
 = 100 if there is any β-structure).
entifier (8) and ORIGIN the species

(a) Contact Preferences AInt29:

	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
HHa	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
HEe	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
EW	-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
EEa	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
EEi	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
EEs	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
EEe	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
EHe	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
ETe	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
EXe	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
TTa	-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
TEe	-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
XXa	-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
XEe	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

(b) Contact Preferences AS5:

	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
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

(c) Contact Preferences AM2:

	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

(d) Contact Preferences AA21:

	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
Wat	-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

Figure 11. Preference parameters for amino acid types in 29 different contact interfaces. Contact preference parameters extracted from the non-redundant data-set of 67 protein chains. The parameters can be used to evaluate how well a particular sequence fits into a particular 3D structure, e.g. in sequence/structure alignment. For example, line 'HHe' contains the single-residue preferences for helix–helix interfaces, line 'HEe' those for helix–sheet interfaces. The 20 standard amino acids are given in one-letter code. Secondary structure notation is: 'H', helix; 'E', extended or β-sheet; 'T', hydrogen-bonded turn; 'X', everything else, called loop (42). Contacts with water are labelled by a 'W'. Chain distance (proximity) of two contacting residues is: 'a', adjacent—the two residues are adjacent in sequence; 'i', internal—the two residues are on the same element of secondary structure, i.e. on the same helix or strand, but not adjacent; 's', strand–strand—the two residues are on adjacent strands in the same β-sheet; 'ə', external—the two residues are on two different elements of secondary structure (for strands, in different sheets). (a) Preferences of amino acid side-chains for 29 interface types, e.g. helix–helix, helix–sheet (parameter set AInt29). These 'structure–structure' interface types are classified according to: the secondary structure of a specific residue type, S_1 = H,E,T, or X; the secondary structure of the contacting residue(s) of any type S_2 = H,E,T, or X (or S_2 = W for contacts with water); and the chain distance (proximity) of any two contacting residues, p_{12} = a,i,s,e. Notation for a contact type is $S_1S_2p_{12}$, e.g. HHa, HHi, HHe, HEe, etc. For example, Pro has a strong preference for TTa (Pro located in a turn makes contacts with other residue(s) in the same turn); Lys has clear preference for HW, EW, TW, and XW (Lys

Figure 11. *Continued*
located on any element of secondary structure make contacts with water); the strongest preferences for EHe are expressed by Ile and Phe (Ile, Phe in a β-strand make strand–helix contacts). Writing 'strand' for β-strand, the notation is, for example, HHa, intra-helix contact between residues *i* and *i*+1; XXi, contact between two residues in the same loop; EEi, intra-strand contact; EEs, strand–strand contact; EEe, sheet–sheet contact; HEe, helix–sheet contact, first residue in helix; EHe, helix–sheet contact, first residue in strand; EXe, contact between strand and loop; TW, contact between an H-bonded turn and water. (b) Preferences of amino acid side-chains for four interface types, e.g. contact with a helix, contact with a sheet (parameter set AS5). These simpler 'anything–structure' interface types are derived from the 29 interface types in (a) by summing over the secondary structure state 'S_1' of the first residue. So the four 'contact-with' interface types are: 'H', residue contact of the first residue (in any secondary structure state) with a helix residue; 'E', contact with a β-sheet residue; 'T', contact with a turn residue; 'X', contact with a loop residue; 'W', contact with water. For example, Val and Ile have a clear preference to be in contact with β-sheet residues; Ala, for making contacts with helix residues; Lys, for making contact with water; and so on. These preferences are dominated by the secondary structure state of the first residue, as residues in a helix are likely to make contacts with other helix residues. Notation is, for example, H, residue (e.g. Ala) makes contact with a helix; E, residue (e.g. Leu) makes contact with a strand; Wat, residue (e.g. Lys) makes contact with water. (c) Preferences of amino acid side-chains for two interface types, i.e. contact with protein atoms ('Prot') or contact with water molecules ('Wat') (parameter set AM2). These very simple 'protein–water' interface types are derived from the parameters in (a) by summing protein–protein contacts over the secondary structure states of both participating residues, or from those in (b) by summing over the secondary structure state 'S_2' of the second residue. The only remaining distinction is that between contacts of a protein atom with other protein atoms or with solvent atoms. These parameters resemble a hydrophobicity scale, e.g. Lys has the strongest preference for water contacts while Ile, Phe, Trp, and Cys have the weakest preference for contacts with protein atoms. The apparently weaker contrast in line 'Prot' is a numerical effect, due to the fact that, in the data-base used, the total number of contacts with protein atoms exceeds by far that with water atoms. Notation is: Wat, residue (e.g. Lys) makes contact with water; Prot, residue (e.g. Leu) makes contact with other protein atoms. (d) Preferences of amino acid side-chains for contacting another side-chain or water, irrespective of secondary structure (parameter set AA21). These parameters are analogous to effective residue–residue potential energy parameters for a polypeptide chain, but also include a residue–water contact term. Water contacts are in line 'Wat'. When these parameters are summed over one of the contact partners, the parameters in (c) result, for contacts of a side-chain with 'protein' or with 'water'. Notation is (column/row), for example, V/I, Val and Ile make contact (in any structural state); K/D, Lys and Asp make contact (in any structural state); K/Wat, Lys makes contact with water.

accessible surface area, this translates to 0.31 atom–water contact for each Å^2 of surface area (43). So water–protein contacts can be included in the definition of contact interfaces by using this simple equivalence.

7.4 Statistical evaluation of contacts

For the statistical evaluation, contacts are labelled with the residue type of the central residue and the interface type. The contact strengths of all residues in

e contacts with water); the strongest
, Phe in a β-strand make strand–helix
on is, for example, HHa, intra-helix
:ween two residues in the same loop;
act; EEe, sheet–sheet contact; HEe,
-sheet contact, first residue in strand;
between an H-bonded turn and water.
erface types, e.g. contact with a helix,
impler 'anything–structure' interface
a) by summing over the secondary
contact-with' interface types are: 'H',
structure state) with a helix residue;
turn residue; 'X', contact with a loop
nd Ile have a clear preference to be in
ntacts with helix residues; Lys, for
nces are dominated by the secondary
helix are likely to make contacts with
sidue (e.g. Ala) makes contact with a
rand; Wat, residue (e.g. Lys) makes
de-chains for two interface types, i.e.
ater molecules ('Wat') (parameter set
pes are derived from the parameters
e secondary structure states of both
ng over the secondary structure state
nction is that between contacts of a
t atoms. These parameters resemble
eference for water contacts while Ile,
r contacts with protein atoms. The
cal effect, due to the fact that, in the
otein atoms exceeds by far that with
kes contact with water; Prot, residue
(d) Preferences of amino acid side-
irrespective of secondary structure
ogous to effective residue–residue
n, but also include a residue–water
these parameters are summed over
ult, for contacts of a side-chain with
for example, V/I, Val and Ile make
ake contact (in any structural state);

atom–water contact for each Å^2
ts can be included in the defini-
equivalence.

ts

lled with the residue type of the
ntact strengths of all residues in

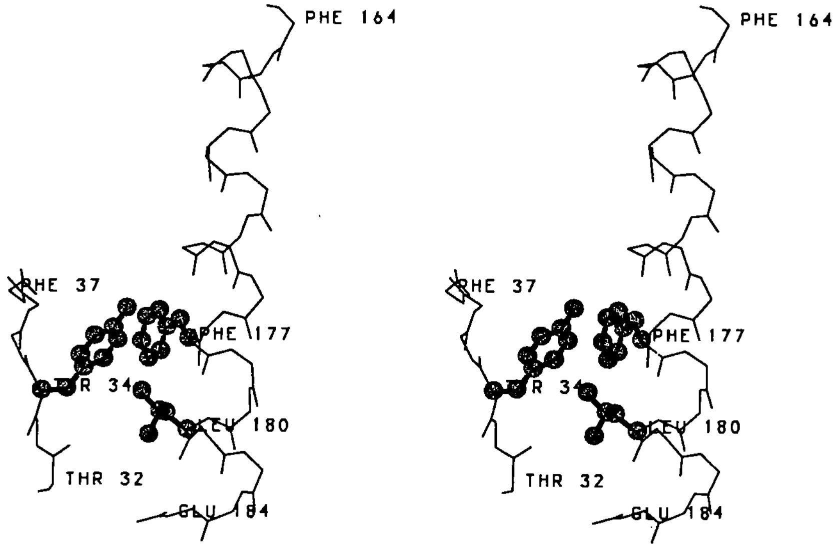


Figure 12. Stereo view. Example of how single-residue contact preference parameters are used to evaluate amino acid residues in a given structural context, using interface-dependent parameters from *Figure 11a*. The two contacting segments (backbone trace), a β-strand (left) and an α-helix (right) make a central contact involving Phe37, Phe177, and Leu180 (side-chains highlighted as balls and sticks). To evaluate the suitability of these residues in this interface sequence, look up the single residue preferences in *Figure 10.* Phe37 participates in the interface of type 'EHe' (residue in β-strand (=E) in contact with an α-helix (=H) in two different secondary structure segments (=e), with preference pref(Phe,EHe) = +1.19; similarly for Phe177 we have pref(Phe,HEe) = +0.65 and for Leu180 pref(Leu,HEe) = +0.54. So the combination has a combined preference of +2.38 bits or an observed/expected ratio of $2^{2.38}$ = 5.21. Note that residues in general participate in more than one interface so that all other contacts of each residue have to be evaluated in order to judge the entire structure. The fragments are residues 32–37 and 164–184 from tyrosyl tRNA synthase, Protein Data Bank entry 3TS1.

a data-base of selected high-resolution proteins (*Figure 10*) are summed in each category and are written as C(Ri,I), where R is the amino acid type and I the type of the contact interface. An example is C(Ala, HHe), i.e. the total contact strength of alanine in helix–helix contact interfaces. Preference parameters, pref(R,I), that express the over- or under-representation of certain residue types in certain interface types can then be calculated in a way completely analogous to those for residue occurrences in secondary structures, by simply replacing the index S by I in the equation for pref(R,S) above (see Section 6.3). However, the information contained in the contact interface parameters is much richer and reflects directly the important hydrophobic effect by including water–protein contacts in a natural way.

References

1. Richardson, J. and Richardson, D. C. (1989). *TIBS*, **13,** 304–9.
2. Sander, C. (1991). *Current Opinion in Structural Biology,* **1,** 630–8.
3. *Protein design exercises*, EMBL BIOcomputing Technical Document 1. European Molecular Biology Laboratory, Heidelberg.
4. Sander, C. and Vriend, G. (ed.) (1991). *ProDes90*, EMBL BIOcomputing Technical Document 6. European Molecular Biology Laboratory, Heidelberg.
5. Sander, C., *et al.* (1991). *Proteins,* **12,** 105–10.
6. Bernstein, F. C., *et al.* (1977). *J. Mol. Biol.,* **112,** 535–42.
7. Barker, W. C., George, D. G., and Hunt, L. T. (1990). In *Methods in enzymology* (ed. R. F. Doolittle), Vol. 183, pp. 31–49. Academic Press, San Diego.
8. Bairoch, A. and Boeckmann, B. (1991). *Nucl. Acids Res.,* **19,** 2247–50.
9. Schneider, R. and Sander, C. (1991). *Proteins,* **9,** 56–68.
10. Karpusas, M., Baase, W. A., Matsumura, M., and Matthews, B. W. (1989). *Proc. Natl Acad. Sci. USA,* **86,** 8237–41.
11. Zhang, X. J., Baase, W. A., and Matthews, B. W. (1991). *Biochemistry,* **30,** 2012–17.
12. Castagnoli, L., Scarpa, M., Kokkinidis, M., Banner, D. W., Tsernoglou, D., and Cesareni, G. (1989). *EMBO J.,* **8,** 621–9.
13. Sander, C. (1990). *Biochem. Soc. Symp.,* **57,** 25–33.
14. Luger, K., Hommel, U., Herold, M., Hofsteenge, J., and Kirschner, K. (1989). *Science,* **243,** 206–10.
15. Jones, T. A. (1978). *J. Appl. Cryst.,* **11,** 268–72.
16. Dayringer, H. E., Tramontano, A., Sprang, S. R., and Fletterick, R. J. (1986). *J. Mol. Graphics,* **4,** 82–7.
17. Vriend, G. (1990). *J. Mol. Graphics,* **8,** 52–6.
18. Vriend, G. and Sander, C. (1991). *Proteins,* **11,** 52–8.
19. Finkelstein, A. V. and Ptitsyn, O. B. (1987). *Prog. Biophys. Mol. Biol.,* **50,** 171–90.
20. Finkelstein, A. V. and Reva, B. A. (1990). *Biofisika (USSR),* **35,** 401–6.
21. Chou, P. Y. and Fasman, G. D. (1974). *Biochemistry,* **13,** 211.
22. Robson, B. and Suzuki, E. (1976). *J. Mol. Biol.,* **107,** 327–56.
23. Levitt, M. (1978). *Biochemistry,* **17,** 4277–85.
24. Lifson, S. and Sander, C. (1979). *Nature,* **282,** 109–11.
25. Oefner, C. (1981). Diplomarbeit (Master thesis), Universität Heidelberg.
26. Argos, P. and Palau, J. (1982). *Int. J. Pep. Prot. Res.,* **19,** 380.
27. Richardson, J. S. and Richardson, D. C. (1988). *Science,* **240,** 1648–52.
28. Schneider, R. (1989). Diplomarbeit (Master thesis), Universität Heidelberg.
29. Chou, P. Y. and Fasman, G. D. (1979). *Biophys. J.,* **26,** 367–73.
30. Isogai, Y., Nemethy, G., Rackovsky, S., Leach, S. J., and Scheraga, H. A. (1980). *Biopolymers,* **19,** 1183–210.
31. Wilmot, C. M. and Thornton, J. M. (1988). *J. Mol. Biol.,* **203,** 221–32.
32. Leszczynski, J. F. and Rose, G. D. (1986). *Science,* **234,** 849–55.
33. Ptitsyn, O. B. and Finkelstein, A. V. (1983). *Biopolymers,* **22,** 15–25.
34. Scharf, M. (1989). Diplomarbeit (Master thesis), Universität Heidelberg.
35. Regan, L. and DeGrado, W. F. (1988). *Science,* **241,** 976–8.
36. Holm, L. and Sander, C. (1991). *J. Mol. Biol.,* **218,** 183–94.

TIBS, **13,** 304–9.
ral Biology, **1,** 630–8.
Technical Document 1. European

es90, EMBL BIOcomputing Tech-
gy Laboratory, Heidelberg.

12, 535–42.
(1990). In *Methods in enzymology*
ademic Press, San Diego.
Acids Res., **19,** 2247–50.
, **9,** 56–68.
and Matthews, B. W. (1989). *Proc.*

B. W. (1991). *Biochemistry,* **30,**

nner, D. W., Tsernoglou, D., and

25–33.
nge, J., and Kirschner, K. (1989).

72.
R., and Fletterick, R. J. (1986). *J.*

1, 52–8.
rog. Biophys. Mol. Biol., **50,** 171–

iofisika (USSR), **35,** 401–6.
emistry, **13,** 211.
ol., **107,** 327–56.

109–11.
is), Universität Heidelberg.
ot. Res., **19,** 380.
8). *Science,* **240,** 1648–52.
esis), Universität Heidelberg.
ys. J., **26,** 367–73.
ach, S. J., and Scheraga, H. A.

Mol. Biol., **203,** 221–32.
ience, **234,** 849–55.
Biopolymers, **22,** 15–25.
s), Universität Heidelberg.
ce, **241,** 976–8.
., **218,** 183–94.

37. Baumann, G., Froemmel, C., and Sander, C. (1990). *Prot. Engng,* **2,** 329–34.
38. Eisenberg, D. and McLachlan, A. D. (1986). *Nature,* **319,** 199–203.
39. Gregoret, L. M. and Cohen, F. E. (1990). *J. Mol. Biol.,* **211,** 959–74.
40. Hill, C. P., Anderson, D. H., Wesson, L., DeGrado, W. F., and Eisenberg, D. (1990). *Science,* **249,** 543–5.
41. DeGrado, W. F., Wassermann, Z. R., and Lear, J. D. (1989). *Science,* **243,** 622–8.
42. Kabsch, W. and Sander, C. (1983). *Biopolymers,* **22,** 2577–637.
43. Colonna-Cesari, F. and Sander, C. (1990). *Biophys. J.,* **57,** 1103–7.
44. Banner, D. W., Kokkinidis, M., and Tsernoglou, D. (1987). *J. Mol. Biol.,* **196,** 657–75.
45. Eberle, W., Pastore, A., Sander, C., and Roesch, P. (1991). *Biomol. NMR,* **1,** 71–82.
46. Hobohm, U., Scharf, M., Schneider, R., and Sander, C. (1992). *Protein Science,* **1,** 409–17.
47. Vriend, G. and Sander, C. (1992). *Acta. Cryst.,* in press.