PROCHECK v.3.0

Programs to check the Stereochemical Quality of Protein Structures

OPERATING INSTRUCTIONS

Roman A Laskowski[†], **Malcolm W MacArthur**[†], **David K Smith**[§], **David T Jones**[†], **E Gail Hutchinson**[†], **A Louise Morris**[†], **Dorica Naylor**, **David S Moss**[‡], **& Janet M Thornton**[†]

[†] *Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, England.*

[‡] *Crystallography Department, Birkbeck College, Malet Street, London WC1E 7HX, England.*

[§] *NMR Laboratory, Biomolecular Research Institute, 381 Royal Parade, Parkeville, Victoria 3052, Australia.*

The preparation of this suite of programs was partly funded by **Oxford Molecular Ltd**.

March 1994

PROCHECK v.3.0

Operating Instructions

1. Introduction

These Operating Instructions describe how to run the PROCHECK suite of programs (Laskowski *et al.*, 1993) for assessing the "stereochemical quality" of a given protein structure. The instructions assume that the programs have already been installed on your computer system. If this is not the case, please refer to the separate *Installation Guide* which deals with the installation procedures for your particular system.

The aim of PROCHECK is to assess both the overall stereochemical quality of a given protein structure, as compared with well-refined structures at the same resolution, and to give an indication of its local, residue-by-residue reliability.

To assess a structure, the program makes use of a number of parameters that have been found to be good indicators of stereochemical quality, described in detail in Morris *et al.* (1992). These parameters, which are for the most part not included in standard refinement procedures (and so are less likely to be biased by them), are listed in Table 1 of Appendix A.

The checks also make use of "ideal" bond lengths and bond angles, as derived from a recent and comprehensive analysis (Engh & Huber, 1991) of small molecule structures in the Cambridge Structural Database, CSD (Allen *et al.*, 1979) - now numbering over 100,000 structures. These "ideal" values are listed in Table 2 of Appendix A.

The PROCHECK programs produce a number of plots, together with a detailed residue-by-residue listing. These are described in Appendices D and E. The plots are output in PostScript format (Adobe Systems Inc., 1985), and so can be printed off on a PostScript laser printer, or displayed on a graphics screen using suitable software (*eg* GHOSTSCRIPT on Sun workstations, or PSVIEW on Silicon graphics IRIS-4D systems). The plots can be in colour or black-and-white.

The input to PROCHECK is a single file containing the coordinates of your protein structure. This must be in Brookhaven file format (see Appendix B). One of the by-products of running PROCHECK is that your coordinates file will be "cleaned up" by the first of the programs. The cleaning-up process corrects any mislabelled atoms and creates a new coordinates file which has a file-extension of **.new**. The **.new** file will have the atoms labelled in accordance with the IUPAC naming conventions (IUPAC-IUB Commission of Biochemical Nomenclature, 1970).

The suite comprises 7 programs which are described in Appendix C. Below are given instructions on running and customizing PROCHECK.

2. How to run PROCHECK

To run PROCHECK, type the following:-

procheck **filename** [**chain-ID**] **resolution**

where: **filename** is the name of the file containing your protein structure and should include the full path unless the file is in the default directory; [**chain-ID**] is the single-letter chain ID of the chain to be analysed (this parameter is optional and can be left out); and **resolution** is the resolution at which the structure has been determined.

For example, to run the check on the Brookhaven file **1gcr**, which was solved to **1.6Å** you might enter:

procheck /data/pdb/p1gcr.pdb 1.6

For structures solved using NMR techniques, for which the term "resolution" does not apply, use some nominal value such as 2.0Å. For model-built structures, use the resolution of the structure(s) on which the model has been based. Note that, in this case, it is also useful to run PROCHECK on the starting structure beforehand to assess its likely reliability.

As mentioned above, the input file must be in Brookhaven format (Appendix B).

Running in batch-mode on a VAX

On VAX computers, you can submit the entire process to a batch queue by entering:-

prosub **filename** [**chain-ID**] **resolution** **queue-name** **default directory**

where **queue-name** is the name of the batch-queue to which you want the job submitted, and **default directory** is the directory in which the files created by PROCHECK are to be put. Note, this need not be the same directory as the one containing the structure file, but it must be one to which you have write access.

Alternatively, if you just type **prosub** you will be prompted for each of the 5 parameters in turn. If you do not want to specify a chain ID, leave this entry blank.

3. Outputs produced by PROCHECK

PROCHECK generates a number of output files in the default directory which have the same name as the original PDB file, but with different extensions, as follows:-

PostScript plot files and listing

_01.ps	Plot files (numbered **01** to **nn**) in PostScript format
⋮	⋮
_nn.ps	⋮
.out	The residue-by-residue listing (*ie* text file for printing)

Other files

.lan	Main-chain bond lengths and bond angles used by the plotting programs
.nb	List of atom-pairs making near-neighbour contacts
.new	"Cleaned-up" version of the original coordinates file
.pln	Coordinates of atoms in planar groups
.rin	Residue information used by the plotting programs
.sco	Main-chain and side-chain properties
.sdh	Residue-by-residue G-factors

The plot files (**.ps**) can be sent directly to a PostScript printer or viewed on a graphics or X Windows terminal. The **.out** listing file is a text file which can be sent directly to a line-printer.

Of the others, most are used internally by the PROCHECK suite, and the only one that is likely to be of use to you is the **.new** file. This holds the 'cleaned-up' version of the original PDB file, with any wrong atom-labels corrected in accordance with the IUPAC naming conventions (IUPAC-IUB Commission of Biochemical Nomenclature, 1970).

Each program in the suite also produces its own log file. Should the PROCHECK suite crash, or give strange-looking results, these log files should be the first place you look for a reason for the problem. The 7 files are:

anglen.log	**bplot.log**	**clean.log**	**nb.log**
pplot.log	**secstr.log**	**tplot.log**	

4. Running just the plotting programs

The plots and residue-by-residue listing are produced by the final 3 programs in the suite: TPLOT.F, PPLOT.F and BPLOT.F. Most of the calculations are performed by the 4 programs that are run beforehand, and these constitute the most CPU-intensive part of the process. If you need to produce plots of the same structure several times (perhaps using different parameters each time), you can save time by not running these other programs each time; you can run just the plotting programs on their own by using the command **proplot**. This has the same two, or three, parameters as before, namely:

proplot **filename** [**chain-ID**] **resolution**

Running just the plotting programs can save a lot of time but, of course, requires that the entire suite has been run at least once on the current structure of interest so that the files listed in section 3 above have been generated.

5. Customizing PROCHECK

The plots and listing produced by PROCHECK can be customized to a limited extent by amending the parameters in a file called **procheck.prm**. The file is created in the default directory when you first run PROCHECK. You can then amend it using any text editor.

In the file there is a separate section for each of the plots produced, as well as sections governing general parameters. These will be described below. The first line of the file gives the PROCHECK version number. If a subsequent version has altered the format in any way, the programs will detect this as an error and you will need to delete your existing copies of **procheck.prm** so that the programs can create new versions in the correct format.

5.1. General options

The first few options are general options covering all the plots.

Option	*Description*
Colour all plots?	
Y	<- Produce all plots in colour (Y/N)?
Which plots to produce?	
Y	<- 1. Ramachandran plot (Y/N)?
Y	<- 2. Gly & Pro Ramachandran plots (Y/N)?
Y	<- 3. Chi1-Chi2 plots (Y/N)?
Y	<- 4. Main-chain parameters (Y/N)?
Y	<- 5. Side-chain parameters (Y/N)?
Y	<- 6. Residue properties (Y/N)?
Y	<- 7. Main-chain bond length distributions (Y/N)?
Y	<- 8. Main-chain bond angle distributions (Y/N)?
Y	<- 9. RMS distances from planarity (Y/N)?
Y	<- 10. Distorted geometry plots (Y/N)?

Produce all plots in colour. This option determines whether all the plots are to be produced as colour or black-and-white PostScript files. If the option is set to **Y**, then colour files will be generated for all plots. If it is set to **N**, each plot will be generated in colour or black-and-white depending on the colour option for that particular plot (see below). The colours themselves can be altered as will be described below.

Which plots to produce. Each of the 10 plots can be individually switched on or off by entering **Y** or **N** here, respectively.

5.2. Plot options

The next 10 sets of options apply specifically to each of the 10 plots generated by PROCHECK. The plots are described in detail in Appendix D.

5.2.1. Ramachandran plot

```
1. Ramachandran plot
Y          <- Shade in the different regions (Y/N)?
Y          <- Print the letter-codes identifying the different regions (Y/N)?
Y          <- Draw line-borders around the regions (Y/N)?
N          <- Show only the core region (Y/N)?
1          <- Label residues in: 0=disallow,1=generous,2=allow,3=core regions
N          <- Produce a COLOUR PostScript file (Y/N)?
WHITE      <- Background colour
WHITE      <- Region 0: Disallowed
CREAM      <- Region 1: Generous
YELLOW     <- Region 2: Allowed
RED        <- Region 3: Most favourable, core, region
BLACK      <- Colour of markers in favourable regions
RED        <- Colour of markers in unfavourable regions
```

Shade in the different regions. Determines whether the different regions of the Ramachandran plot are to be shaded in. Without shading, the regions can still be made out if their borders are drawn in (see below). In black-and-white, the shading shows the most favourable regions in the darkest grey, with the less favourable regions being shown in progressively lighter tones. If you are only interested in the core region, then set the option **Show only the core region** (see below) to **Y**.

Print the letter-codes. Allows you to label the different regions of the Ramachandran plot using the appropriate codes: **A**, **B**, **L**, *etc*.

Draw line-borders around the regions. Determines whether borders are drawn round the different regions of the Ramachandran plot. If you are only interested in the core region, then set the option **Show only the core region** below to **Y**.

Show only the core region. Allows you to show only the core regions (*ie* the most favoured regions: **A**, **B**, and **L**) of the Ramachandran plot. This gives an easy-to-see guide as to how many residues are inside and outside the cores.

Label: 0=disallow, 1=generous, 2=allowed, 3=core. Determines which residues are labelled on the Ramachandran the plot. Option **0** labels only the residues in the "disallowed" regions. Option **1** labels residues in the disallowed and "generous" regions. And so on. In all cases, Gly residues are left unlabelled and are identified by a triangular marker.

Produce a COLOUR PostScript file. Determines whether a colour or black-and-white plot is required. Note that if the **Colour all plots** option is set to **Y** above, a colour PostScript file will be produced irrespective of the setting here.

The colour definitions on the following lines use the 'names' of each colour as defined in the colour table at the bottom of the file (see section 5.4 below). If the name of a colour is mis-spelt, or does not appear in the colour table, then white will be taken instead. Each colour can be altered to suit your taste and aesthetic judgement, as described in section 5.4 below.

The first of the defined colours, the **Background colour**, defines the background colour of the page on which the plots are drawn.

Colours. The various additional colours on the lines that follow define the colour of each of the 4 different regions on the Ramachandran plot, and the colours of the markers that are located in the favourable and unfavourable regions.

5.2.2. Gly & Pro Ramachandran plots

2. Gly & Pro Ramachandran plots	
-4.0	<- Cut-off value for labelling of residues
N	<- Plot all 20 Ramachandran plots (Y/N)?
N	<- Produce a COLOUR PostScript file (Y/N)?
WHITE	<- Background colour
CREAM	<- Lightest shaded regions on plots
GREEN	<- Darkest shaded regions on plots
BLACK	<- Colour of crosses in favourable regions
RED	<- Colour of crosses in unfavourable regions

Cut-off value for labelling of residues. G-factor values below which residues on the Ramachandran plot will be labelled. The G-factor gives a measure of how far from the 'normal' regions of the plot each residue lies. Low G-factors indicate residues in unlikely conformations.

Plot all 20 Ramachandran plots. Allows you to show a separate ϕ–ψ Ramachandran plot for each of the 20 different residue types, rather than just for Gly and Pro.

Produce a COLOUR PostScript file. See section 5.2.1 above.

Colours. Define the colours for the lightest- and darkest-shaded regions of the plot, and also the colours of the residue markers in the favourable and unfavourable regions.

5.2.3. χ_1–χ_2 plots

3. Chi1-Chi2 plots	
-4.0	<- Cut-off value for labelling of residues
N	<- Produce a COLOUR PostScript file (Y/N)?
WHITE	<- Background colour
CREAM	<- Lightest shaded regions on plots
GREEN	<- Darkest shaded regions on plots
BLACK	<- Colour of crosses in favourable regions
RED	<- Colour of crosses in unfavourable regions

Cut-off value for labelling of residues. G-factor values below which residues on the χ_1–χ_2 plot will be labelled. The G-factor gives a measure of how far from the 'normal' regions of the plot each residue lies. Low G-factors indicate residues in unlikely conformations.

Produce a COLOUR PostScript file. See section 5.2.1 above.

Colours. Define the colours for the lightest- and darkest-shaded regions of the plot, and also the colours of the residue markers in the favourable and unfavourable regions.

5.2.4. Main-chain parameters

```
4. Main-chain parameters
Y          <- Background shading (Y/N)?
N          <- Produce a COLOUR PostScript file (Y/N)?
WHITE      <- Background colour
CREAM      <- Background shading on each graph
PURPLE     <- Colour of band
```

Background shading. Defines whether the background of each plot is to be lightly shaded when the plot is in black-and-white.

Produce a COLOUR PostScript file. See section 5.2.1 above.

Colours. Define the colours for the background of each plot, and the band representing the range of values expected at each resolution.

5.2.5. Side-chain parameters

```
5. Side-chain parameters
Y          <- Background shading (Y/N)?
N          <- Produce a COLOUR PostScript file (Y/N)?
WHITE      <- Background colour
CREAM      <- Background shading on each graph
PURPLE     <- Colour of band
```

Background shading. Defines whether the background of each plot is to be lightly shaded when the plot is in black-and-white.

Produce a COLOUR PostScript file. See section 5.2.1 above.

Colours. Define the colours for the background of each plot, and the band representing the range of values expected at each resolution.

5.2.6. Residue properties

```
6. Residue properties
1 2 3       <- Which 3 main graphs to be printed (see Note 1 for full list)
Y           <- Background shading on main graphs (Y/N)?
2.0         <- Number of standard deviations for highlighting
Y           <- Show shading representing estimated accessibility (Y/N)?
N           <- Produce a COLOUR PostScript file (Y/N)?
WHITE       <- Background colour
CREAM       <- Background shading on main graphs
PURPLE      <- Colour of histogram bars on main graphs
RED         <- Colour of highlighted histogram bars
BLUE        <- Minimum accessibility colour (buried regions)
WHITE       <- Maximum accessibility colour
RED         <- Region 0: Disallowed
PINK        <- Region 1: Generous
GREEN       <- Region 2: Allowed
SKY BLUE    <- Region 3: Most favourable, core, region
YELLOW      <- Colour for favourable G-factor scores
RED         <- Colour for unfavourable G-factor scores
YELLOW      <- Colour for schematic of the secondary structure
```

Which 3 main graphs to be printed. These three options define which 3 out of the 11 possible graphs are to appear as the main graphs on the plot. The eleven possibilities are:

1. Absolute deviation of χ_1 torsion angle from the "ideal", excluding Pro.

2. Absolute deviation of ω torsion angle from the "ideal".

3. Absolute deviation of ζ "virtual" torsion angle (defined by the atoms Cα–N–C–Cβ) from the "ideal". This torsion angle provides a measure of the Cα chirality and, if negative, signifies a D-amino acid.

4. Absolute deviation of main-chain hydrogen bond energy from the "ideal".

5. B-value of γ atom (O, C, or S - being whichever is used in definition of the χ_1 torsion angle).

6. Average B-value of main-chain atoms.

7. Average B-value of side-chain atoms.

8. G-factors for ϕ–ψ distributions. Each residue's $G_{\phi-\psi}$-factor provides a measure of how 'normal' the ϕ–ψ values for this residue type are. Highlighted bars indicate low-probability conformations.

9. G-factors for χ_1–χ_2 distributions. Each residue's $G_{\chi_1-\chi_2}$-factor provides a measure of how 'normal' the χ_1–χ_2 values for this residue type are. Highlighted bars indicate low-probability conformations.

10. Residue-by-residue G-factor. The G-factor plotted here combines the contributions from the dihedral angle G-factor, G_{dih}, and the main-chain bond lengths and bond angles G-factor, G_{cov}.

11. Approximate accessibility, as estimated by each residue's Ooi number. An Ooi number is a count of the number of other Cα atoms within a radius of, in this case, 14Å of the given residue's own Cα (Nishikawa & Ooi, 1986).

Entering **0** for any of the 3 options will leave a blank region in place of a graph at the corresponding position on the page.

Background shading. Defines whether the background of each of the 3 main plots is to be lightly shaded when the plot is in black-and-white.

Number of standard deviations for highlighting. Determines which histogram bars in the 3 main plots are to highlighted. Residues whose plotted parameter deviates by more than the number of standard deviations specified here will be highlighted.

Show shading for accessibility. Determines whether the secondary structure diagram is to be plotted on a shaded background in which the shading shows the buried and accessible regions of the structure. At present the accessibility is estimated from each residue's Ooi number (see above), but a proper accessibility calculation will be included in PROCHECK very soon.

Produce a COLOUR PostScript file. See section 5.2.1 above.

Colours. Define the various colours to be used for the different parts of the plot. These include the colouring of the 3 main graphs at the top of the plot, the accessibility shading and colour of the secondary structure diagram, the markers representing the different regions of the Ramachandran plot, and the range of colours used on the chequer-board of the various G-factors.

5.2.7. Main-chain bond length distributions

```
7. Main-chain bond length distributions
Y          <- Background shading on graphs (Y/N)?
2.0        <- Number of standard deviations for highlighting
N          <- Produce a COLOUR PostScript file (Y/N)?
WHITE      <- Background colour of page
CREAM      <- Background shading on graphs
PURPLE     <- Colour of histogram bars
RED        <- Colour of highlighted histogram bars
```

Background shading. Defines whether the background of each plot is to be lightly shaded when the plot is in black-and-white.

Number of standard deviations for highlighting. Determines which histogram bars in the plot are to highlighted. Residues whose plotted parameter deviates by more than the number of standard deviations specified here will be highlighted.

Produce a COLOUR PostScript file. See section 5.2.1 above.

Colours. Define the colours for the highlighted and unhighlighted histogram bars, and the background colour of each graph.

5.2.8. Main-chain bond angle distributions

```
8. Main-chain bond angle distributions
Y          <- Background shading on graphs (Y/N)?
2.0        <- Number of standard deviations for highlighting
N          <- Produce a COLOUR PostScript file (Y/N)?
WHITE      <- Background colour of page
CREAM      <- Background shading on graphs
PURPLE     <- Colour of histogram bars
RED        <- Colour of highlighted histogram bars
```

Background shading. Defines whether the background of each plot is to be lightly shaded when the plot is in black-and-white.

Number of standard deviations for highlighting. Determines which histogram bars in the plot are to highlighted. Residues whose plotted parameter deviates by more than the number of standard deviations specified here will be highlighted.

Produce a COLOUR PostScript file. See section 5.2.1 above.

Colours. Define the colours for the highlighted and unhighlighted histogram bars, and the background colour of each graph.

5.2.9. RMS distances from planarity

```
9. RMS distances from planarity
Y          <- Background shading on graphs (Y/N)?
0.03       <- RMS distance for highlighting for ring groups
0.02       <- " " " " " other groups
N          <- Produce a COLOUR PostScript file (Y/N)?
WHITE      <- Background colour of page
CREAM      <- Background shading on graphs
PURPLE     <- Colour of histogram bars
RED        <- Colour of highlighted histogram bars
```

Background shading. Defines whether the background of each plot is to be lightly shaded when the plot is in black-and-white.

RMS distance for highlighting. Determines which histogram bars in the plot are to be highlighted. Bars corresponding to RMS distances from planarity greater than the values specified here will be highlighted. The two different values apply separately to ring groups (Phe, Tyr, Trp and His) and to planar end-groups (Arg, Asn, Asp, Gln and Glu).

Produce a COLOUR PostScript file. See section 5.2.1 above.

Colours. Define the colours for the highlighted and unhighlighted histogram bars, and the background colour of each graph.

5.2.10. Distorted geometry plots

```
10. Distorted geometry plots
0.05     <- Deviation from "ideal" bond-length (A)
10.0     <- Deviation from "ideal" bond-angle (degrees)
0.04     <- RMS distance from plane for ring atoms
0.03     <- " " " " for other atoms
N        <- Produce a COLOUR PostScript file (Y/N)?
CREAM    <- Background colour of page
BLUE     <- Colour of "ideal" bond-lengths and angles
RED      <- Colour of actual bond-lengths/angles/planes
GREEN    <- Colour for lettering showing differences from ideals
```

Deviation from 'ideal' bond length (Å). Defines which main-chain bonds are to be plotted as distorted. All bond-lengths that deviate by more than this amount from the Engh & Huber 'ideal' values will be plotted.

Deviation from 'ideal' bond angle (°). Defines which main-chain bond angles are to be plotted as distorted. All bond-angles that deviate by more than this amount from the Engh & Huber 'ideal' values will be plotted.

RMS distance from plane. Defines which planar groups are to be plotted as distorted. All planar groups that deviate by more than this amount from planarity will be plotted. The two different values apply separately to ring groups (Phe, Tyr, Trp and His) and to planar end-groups (Arg, Asn, Asp, Gln and Glu).

Produce a COLOUR PostScript file. See section 5.2.1 above.

Colours. Define the colours for the 'ideal' and actual bond lengths and angles and for the lettering showing the differences from the ideals.

5.3. Listing options

The listing options affect the output in the **.out** file which gives the residue-by-residue listing.

```
Listing options
Y     <- Print explanatory text at head of parameters listing (Y/N)
N     <- Print only asterisks denoting deviations, and not the values (Y/N)
N     <- Print only highlighted residues (Y/N)
0.0   <- Min. deviation from ideal required for parameter to be printed
66    <- Number of lines per page on parameters listing
Y     <- List the bad contacts (Y/N)
```

Print explanatory text. Allows you to include or suppress the printing of the explanatory text which gives details such as the tables of "ideal" values against which your structure has been compared, and the keys to the various codes used in the print-out.

Print only asterisks denoting deviations and not values. Answering **Y** to this option shows only the asterisks on the listing. No values are printed.

Print only highlighted residues. Allows you to only print those residues which have any asterisked parameters.

Min. deviation from ideal for parameter to be printed. Defines the minimum deviation required in a parameter before it is printed. The default value is 0.0, meaning that all parameters are printed. If the value is set to, say, **2.0**, only those parameters that deviate by 2.0 or more standard deviations from their ideal value will be shown.

Number of lines per page. Defines the number of lines to be printed before a page is thrown. It tells the program when to start a new page and hence when to print a new set of column headings. Normally, the standard is 66 lines per page, but your printer may print more or less per page, so amend this figure accordingly.

List the bad contacts. Determines whether the bad contacts are to be listed at the end of the print-out or not.

5.4. Colours

The colour table allows you to modify any of the default colour definitions and to set up new colours of your own (up to 20 colours are allowed). Each entry contains 3 numbers, between 0.0 and 1.0, giving the ratios of red, green and blue, respectively, making up the given colour. Each colour also has a 'name', in single quotes, by which the colour is referred to in the plot options above. The default colours are as shown below.

Colours				
0.0000	0.0000	0.0000	'BLACK'	<- Colour 1
1.0000	1.0000	1.0000	'WHITE'	<- Colour 2
1.0000	0.0000	0.0000	'RED'	<- Colour 3
0.0000	1.0000	0.0000	'GREEN'	<- Colour 4
0.0000	0.0000	1.0000	'BLUE'	<- Colour 5
1.0000	1.0000	0.0000	'YELLOW'	<- Colour 6
0.8000	0.5000	0.0000	'ORANGE'	<- Colour 7
0.5000	1.0000	0.0000	'LIME GREEN'	<- Colour 8
0.5000	0.0000	1.0000	'PURPLE'	<- Colour 9
0.5000	1.0000	1.0000	'CYAN'	<- Colour 10
1.0000	0.5000	1.0000	'PINK'	<- Colour 11
0.3000	0.8000	1.0000	'SKY BLUE'	<- Colour 12
1.0000	1.0000	0.7000	'CREAM'	<- Colour 13
0.0000	1.0000	1.0000	'TURQUOISE'	<- Colour 14
1.0000	0.0000	1.0000	'LILAC'	<- Colour 15
0.8000	0.0000	0.0000	'BRICK RED'	<- Colour 16
0.5000	0.0000	0.0000	'BROWN'	<- Colour 17
0.9700	0.9700	0.9700	'LIGHT GREY'	<- Colour 18
1.0000	1.0000	1.0000	'WHITE'	<- Colour 19
1.0000	1.0000	1.0000	'WHITE'	<- Colour 20

5.5. *G*-factors

```
G-factors
Y   <- Use Engh & Huber means for bond length/angle G-factors(Y/N)?
```

The final option determines whether the Engh & Huber mean values are to be used for the main-chain bond lengths and bond angles when calculating the *G*-factors. The alternative is to use the structure's own mean values.

Acknowledgements

We would like to thank Eleanor Dodson, Andrej Šali, Keith Wilson, Victor Lamzin, Mike Sutcliffe, George Sheldrick, Paula Kuser, Helen Stirk, V Dhanaraj and Geoff Barton for immensely helpful comments, criticisms and suggestions.

Appendix A

Stereochemical parameters

The two tables in this appendix list the stereochemical parameters used in the PROCHECK programs.

TABLE 1

Stereochemical parameter	Mean value	Standard deviation
ϕ–ψ in most favoured regions of Ramachandran plot	$> 90\%$	
χ_1 dihedral angle:		
gauche minus	64.1°	15.7°
trans	183.6°	16.8°
gauche plus	−66.7°	15.0°
χ_2 dihedral angle	177.4°	18.5°
Proline ϕ torsion angle	−65.4°	11.2°
Helix ϕ torsion angle	−65.3°	11.9°
Helix ψ torsion angle	−39.4°	11.3°
χ_3 (S–S bridge):		
right-handed	96.8°	14.8°
left-handed	−85.8°	10.7°
Disulphide bond separation	2.0Å	0.1Å
ω dihedral angle	180.0°	5.8°
Main-chain hydrogen bond energy (kcal/mol)†	-2.03	0.75
Cα chirality: ζ "virtual" torsion angle (Cα–N–C–Cβ	33.9°	3.5°

†Evaluated using the Kabsch & Sander (1983) method.

Table 1. Stereochemical parameters of Morris *et al.* (1992), derived from high-resolution protein structures.

TABLE 2

a. Bond lengths

Bond	X-PLOR labelling		Value (Å)	σ
C–N	C–NH1	(except Pro)	1.329	0.014
	C–N	(Pro)	1.341	0.016
C–O	C–O		1.231	0.020
Cα–C	CH1E–C	(except Gly)	1.525	0.021
	CH2G*–C	(Gly)	1.516	0.018
Cα–Cβ	CH1E–CH3E	(Ala)	1.521	0.033
	CH1E–CH1E	(Ile,Thr,Val)	1.540	0.027
	CH1E–CH2E	(the rest)	1.530	0.020
N–Cα	NH1–CH1E	(except Gly,Pro)	1.458	0.019
	NH1–CH2G*	(Gly)	1.451	0.016
	N–CH1E	(Pro)	1.466	0.015

b. Bond angles

Angle	X-PLOR labelling		Value (°)	σ
C–N–Cα	C–NH1–CH1E	(except Gly,Pro)	121.7	1.8
	C–NH1–CH2G*	(Gly)	120.6	1.7
	C–N–CH1E	(Pro)	122.6	5.0
Cα–C–N	CH1E–C–NH1	(except Gly,Pro)	116.2	2.0
	CH2G*–C–NH1	(Gly)	116.4	2.1
	CH1E–C–N	(Pro)	116.9	1.5
Cα–C–O	CH1E–C–O	(except Gly)	120.8	1.7
	CH2G*–C–O	(Gly)	120.8	2.1
Cβ–Cα–C	CH3E–CH1E–C	(Ala)	110.5	1.5
	CH1E–CH1E–C	(Ile,Thr,Val)	109.1	2.2
	CH2E–CH1E–C	(the rest)	110.1	1.9
N–Cα–C	NH1–CH1E–C	(except Gly,Pro)	111.2	2.8
	NH1–CH2G*–C	(Gly)	112.5	2.9
	N–CH1E–C	(Pro)	111.8	2.5
N–Cα–Cβ	NH1–CH1E–CH3E	(Ala)	110.4	1.5
	NH1–CH1E–CH1E	(Ile,Thr,Val)	111.5	1.7
	N–CH1E–CH2E	(Pro)	103.0	1.1
	NH1–CH1E–CH2E	(the rest)	110.5	1.7
O–C–N	O–C–NH1	(except Pro)	123.0	1.6
	O–C–N	(Pro)	122.0	1.4

Table 2. Means and standard deviations for main-chain bond lengths and bond angles, as observed in small molecules (Engh & Huber, 1991). The atom-labelling follows that used in the X-PLOR dictionary, with some additional atom types (marked with an asterisk) as defined by Engh & Huber (1991).

Appendix B

Brookhaven file format

The table below shows the Brookhaven file format for the coordinate records (*ie* ATOM and HETATM) of your PDB file. Each record holds the coordinates and other details of a single atom.

Field No.	Column range	FORTRAN format	Description
1.	1 – 6	A6	Record ID (*eg* ATOM, HETATM)
2.	7 – 11	I5	Atom serial number
–	12 – 12	1X	Blank
3.	13 – 16	A4	Atom name (*eg* ␣CA␣, ␣ND1)
4.	17 – 17	A1	Alternative location code (if any)
5.	18 – 20	A3	Standard 3-letter amino acid code for residue
–	21 – 21	1X	Blank
6.	22 – 22	A1	Chain identifier code
7.	23 – 26	I4	Residue sequence number
8.	27 – 27	A1	Insertion code (if any)
–	28 – 30	3X	Blank
9.	31 – 38	F8.3	Atom's x-coordinate
10.	39 – 46	F8.3	Atom's y-coordinate
11.	47 – 54	F8.3	Atom's z-coordinate
12.	55 – 60	F6.2	Occupancy value for atom
13.	61 – 66	F6.2	B-value (thermal factor)
–	67 – 67	1X	Blank
14.	68 – 68	I3	Footnote number

Appendix C

Program descriptions

This appendix describes the 7 programs making up the PROCHECK suite. The relationship of the programs is shown schematically in Figure 1.

Program 1. CLEAN.F

Written by: D K Smith & E G Hutchinson. Modified: R A Laskowski

The first of the programs is called CLEAN. Its purpose is to produce a cleaned-up version of your coordinates file. The new file is given a **.new** extension and is used by the other programs in the suite for the calculation of the stereochemical parameters.

The cleaning-up process involves a number of checks. The first one ensures that the atoms in your structure have been correctly labelled in accordance with the IUPAC naming conventions (IUPAC-IUB Commission on Biochemical Nomenclature, 1970).

A typical error is that the Nη_1 and Nη_2 atoms of arginine are labelled the wrong way round. This error is corrected by the program. Also corrected are atom labels for Phe, Tyr, Asp and Glu residues. Thus, for example, the carbon atom in tyrosine labelled Cδ_1 is the one that gives the lowest χ_2 torsion angle.

The program switches atom labels when they are in error and shows the old atom names in a column at the right of the **.new** file.

The program also checks that the correct L/D stereochemical assignments have been made on individual residues, and corrects any errors found.

All other errors are merely flagged in the **.new** file and described in the **clean.log** file. Also listed in **clean.log** are: chain breaks, unknown residues, and missing atoms. Chain breaks are defined as locations where the distance between two adjacent Cα atoms is greater than 5.0Å. Residue atoms on one side of such a break are marked with a '>' in the **.new** file while the atoms of the residue on the other side are marked with a '<'.

Atoms with alternate locations have only the location with the highest occupancy retained in the **.new** file. The alternate positions are written to the **.new** file, but are flagged: ATOM records are marked as ATALT and HETATM records as HEALT. Atoms with zero occupancy are flagged as ATZERO and HEZERO. None of the flagged atoms are included in the stereochemical checks.

Finally, residues with unusual values of the "notional" ζ dihedral angle (defined by the four atoms: Cα–N–C–Cβ) are marked with an asterisk in the **.new** file. The value of this angle should be $> 23°$ and $< 45°$.

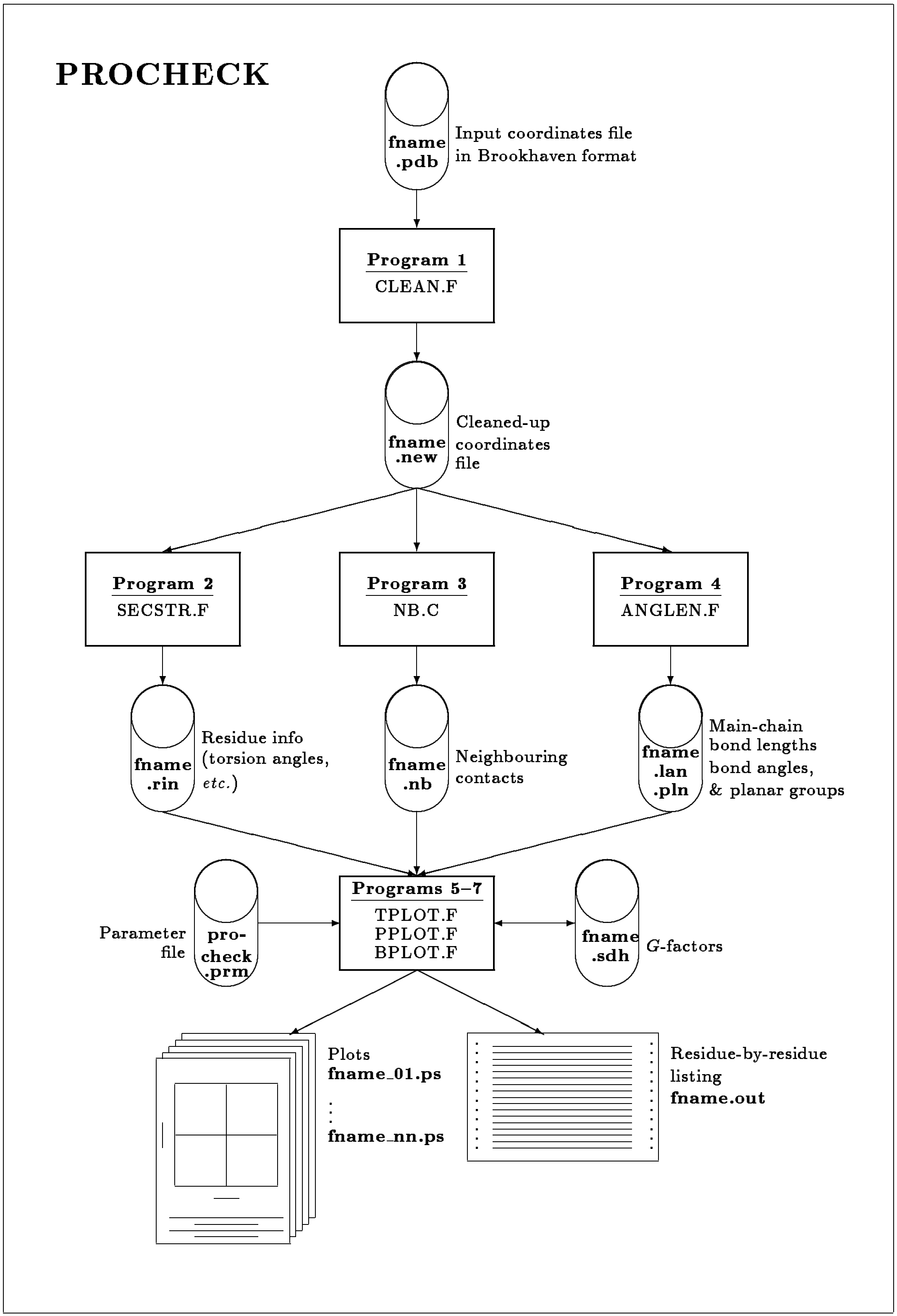

Figure 1. Schematic diagram of the programs in the PROCHECK suite. Each program is represented by a box, and each data file by a cylinder.

Program 2. SECSTR.F

Written by: D K Smith. Modified: R A Laskowski

The second program in the suite is SECSTR which calculates all the required torsion angles, main-chain hydrogen-bond energies, and secondary structure assignments. This information is written out to the **.rin** file for use by the plotting programs.

The hydrogen-bond energies and the secondary structure assignments are calculated using the method of Kabsch & Sander (1983).

Program 3. NB.C

Written by: D T Jones & D Naylor

The third program in the suite is called NB. It identifies all non-bonded interactions between different pairs of residues in the protein structure. These are defined where the closest atom–atom contact between two residues is less than 2.0Å and the atoms concerned are 4 or more bonds apart.

All such closest contacts are written out to the **.nb** file for use by the plotting programs. Note that, only intra-chain contacts are stored - inter-chain contacts are not considered.

Program 4. ANGLEN.F

Written by: R A Laskowski

The fourth program in the suite is called ANGLEN. It calculates all the main-chain bond lengths and bond angles in the structure. These are written out to the **.lan** file. Also calculated are the RMS distances from a best-fit plane for all planar sidechain groups. These comprise the aromatic rings of Phe, Tyr, Trp and His, and the end-groups of Arg, Asn, Asp, Gln and Glu. The coordinates of the atoms in the planar groups are written out to the **.pln** file.

Programs 5–7. TPLOT.F, PPLOT.F and BPLOT.F

Written by: R A Laskowski & M W MacArthur

The last 3 programs in the suite are the plotting programs which create the plots and the residue-by-residue listing for your structure. The plots and listing have a number of user-definable options which are held in the parameter file **procheck.prm** and which can be changed simply by editing the file using any text editor (see section 5 in the main body of these Operating Instructions). The plots are described in Appendix D, and the print-out is described in Appendix E.

Appendix D

Plot descriptions

This appendix describes the 10 plots produced by PROCHECK.

Plot 1. Ramachandran plot

The first of the plots is a Ramachandran plot of the ϕ–ψ torsion angles for all residues in the structure, except those at the chain termini. Glycine residues are separately identified by triangles as these are not restricted to any particular region of the plot.

The shading represents the different regions of the plot: the darker the area the more favourable the ϕ–ψ combination. The regions are labelled as follows:

A	–	Core alpha	**L**	–	Core left-handed alpha
a	–	Allowed alpha	**l**	–	Allowed left-handed alpha
~a	–	Generous alpha	**~l**	–	Generous left-handed alpha
B	–	Core beta	**p**	–	Allowed epsilon
b	–	Allowed beta	**~p**	–	Generous epsilon
~b	–	Generous beta			

The lightest of the regions is the "disallowed" region, and any residues found in this region may need careful investigation.

The different regions are those described in Morris *et al.* (1992) and were taken from the observed ϕ–ψ distribution for 121,870 residues from 463 known protein structures. The two most favoured regions are the "core" and "allowed" regions which correspond to $10^\circ \times 10^\circ$ pixels having more than 100 and 8 residues in them, respectively. The "generous" regions were defined by Morris *et al.* (1992) by extending out by 20° (two pixels) all round the "allowed" regions. In fact, the authors found very few residues in these "generous" regions, so they can probably be treated much like the "disallowed" region and any residues in them investigated more closely.

Ideally, one would hope to have over 90% of the residues in the "core" regions. The percentage of residues in the "core" regions is one of the better guides to stereochemical quality.

The appearance of the plot itself can be modified to some extent by amending some of the parameters in the file **procheck.prm** (see section 5.2.1 of the main part of these Operating Instructions). For example, you can choose in which regions residues are labelled, whether the plot's regions are shaded and lettered, and whether only the "core" regions are highlighted.

Plot 2. Gly & Pro Ramachandran plot

The second plot gives separate ϕ–ψ plots for the Gly and Pro residues only. Both these residue types have very different favoured- and unfavoured-regions from the other residues. The darker the shaded area on the plots, the more favourable the region. The data on which the shading is based has come from a data set of 163 non-homologous, high-resolution protein chains chosen from structures solved by X-ray crystallography to a resolution of 2.0Å or better and an R-factor no greater than 20%.

It is also possible to show all 20 individual Ramachandran plots, one for each residue type, by amending the appropriate parameter in the **procheck.prm** file (see section 5.2.2 in the main section of these Operating Instructions).

Plot 3. Plot of χ_1 *vs* χ_2

The third of the plots shows the χ_1–χ_2 plots for all residue types that have both these torsion angles. The shading on each plot indicates how favourable each region on the plot is. The data for the shading has come from a data set of high-resolution crystal structures (see Plot 2 above).

Plot 4. Main-chain properties

The six graphs on the main-chain properties plot show how the structure (represented by the solid square) compares with well-refined structures at a similar resolution. The dark band in each graph represents the results from the well-refined structures; the central line is a least-squares fit to the mean trend as a function of resolution, while the width of the band on either side of it corresponds to a variation of one standard deviation about the mean. In some cases, the trend is dependent on the resolution, and in other cases it is not.

The 6 properties plotted are:

a. Ramachandran plot quality. This property is measured by the percentage of the protein's residues that are in the most favoured, or "core", regions of the Ramachandran plot (Plot 1 above). For a good model structure, obtained at high resolution, one would expect this percentage to be over 90%. However, as the resolution gets poorer, so this figure decreases - as might be expected. The shaded region reflects this expected decrease with worsening resolution.

b. Peptide bond planarity. This property is measured by calculating the standard deviation of the protein structure's ω torsion angles. The smaller the value the tighter the clustering around the ideal of 180° (which represents a perfectly planar peptide bond).

c. Bad non-bonded interactions. This property is measured by the number of bad contacts per 100 residues. Bad contacts are selected from the list of non-bonded interactions found by program NB (see Appendix C). They are defined as contacts where the distance of closest approach is ≤ 2.6Å.

d. Cα tetrahedral distortion. This property is measured by calculating the standard deviation of the "ζ torsion angle". This is a *notional* torsion angle in that it is not defined about any actual *bond* in the structure. Rather, it is defined by the following four atoms within a given residue: Cα, N, C, and Cβ.

e. Main-chain hydrogen bond energy. This property is measured by the standard deviation of the hydrogen bond energies for main-chain hydrogen bonds. The energies are calculated using the method of Kabsch & Sander (1983).

f. Overall G-factor. The overall G-factor is a measure of the overall 'normality' of the structure (see Plot 6 below for a description of the various G-factors). The overall G is obtained from an average of all the different G-factors for each residue in the structure.

Plot 5. Side-chain properties

These five graphs show five different side-chain properties. Like the graphs in Plot 4, dealing with main-chain properties, they also show how the structure (represented by the solid square) compares with well-refined structures at a similar resolution. Again, the dark band in each graph represents the results from the well-refined structures, giving one standard deviation about a mean trend.

The 5 properties plotted are:

a. Standard deviation of the χ_1 *gauche minus* torsion angles.

b. Standard deviation of the χ_1 *trans* torsion angles.

c. Standard deviation of the χ_1 *gauche plus* torsion angles.

d. Pooled standard deviation of all χ_1 torsion angles.

e. Standard deviation of the χ_2 *trans* torsion angles.

Plot 6. Residue properties

The three main histographs

The various graphs and diagrams on this plot illustrate different properties of the residues in the structure. The three main histographs at the top, can be selected from 11 possibles by the user. The three default graphs, which are plotted when you first run PROCHECK, are the first three of the full list of:

1. Absolute deviation of χ_1 torsion angle from the "ideal".
2. Absolute deviation of ω torsion angle from the "ideal".
3. Absolute deviation of ζ "virtual" torsion angle (defined by the atoms Cα–N–C–Cβ) from the "ideal".
4. Absolute deviation of main-chain hydrogen bond energy from the "ideal".
5. B-value of γ atom (O, C, or S - being whichever is used in definition of the χ_1 torsion angle).
6. Average B-value of main-chain atoms.
7. Average B-value of side-chain atoms.
8. G-factors for ϕ–ψ distributions.
9. G-factors for χ_1–χ_2 distributions.
10. Overall residue-by-residue G-factor.
11. Approximate accessibility, as estimated by each residue's Ooi number.

Which three of these 11 are plotted can be amended by editing the parameter file **procheck.prm** (see section 5.2.6 in the main body of these Operating Instructions).

Secondary structure & estimated accessibility

Below the three main graphs is a schematic picture of the protein's secondary structure, as defined using the Kabsch & Sander (1983) assignments. The key just below the picture shows which structure is which. Beta strands are taken to include all residues with a Kabsch & Sander assignment of **E**, helices corresponds to both **H** and **G** assignments, while everything else is taken to be random coil.

The shading behind the schematic picture gives an approximation to the residue accessibilities. The approximation is a fairly crude one, being based on each residue's Ooi number (Nishikawa & Ooi, 1986). An Ooi number is a count of the number of other Cα atoms within a radius of, in this case, 14Å of the given residue's own Cα. Although crude, this does give a good impression of which parts of the structure are buried and which are exposed on the surface. Future versions of PROCHECK will include an accurate calculation of residue accessibility.

Sequence & Ramachandran regions

The next section shows the sequence of the structure (using the 20 standard amino-acid codes) and a set of markers that identify the region of the Ramachandran plot in which each residue is located in Plot 1 above. There are four marker types, one for each of the four different types of region, and the key explains which is which.

Max. deviation

The small histogram of asterisks and plus-signs shows each residue's "maximum deviation" from one of the ideal values. The asterisk scores are the same as those on the **.out** listing, and in fact correspond to the final column of that listing. Refer to the **.out** file to see which is the parameter that deviates by the amount shown here. (See also Part 1 of Appendix E).

G-factors

The final part of the plot shows a shaded chequer-board of the G-factors for various residue properties (the PROCHECKer board). The darker the shading the more unusual the value of that property. Where several of a residue's properties are unusual, the overall G-factor for that residue will reflect this and identify residues that may need closer scrutiny.

Each G-factor is a measure of the 'normality' of a particular property. For the dihedral angle G-factors, G_{dih}, the standard distribution of each property, for each residue type, has been obtained from a non-homologous, high-resolution data set. For the main-chain bond lengths and bond angles, the Engh & Huber (1991) small-molecule means and standard deviations are used.

Plot 7. Main-chain bond length distributions

The histograms on this plot show the distributions of each of the different main-chain bond lengths in the structure. The solid line corresponds to the small-molecule mean value, while the dashed lines either side show the small-molecule standard deviation, the data coming from Engh & Huber (1991). Highlighted bars correspond to values more than 2.0 standard deviations from the mean, though the value of 2.0 can be changed by editing the **procheck.prm** file (see section 5.2.6 of the main part of these Operating Instructions).

If any of the histogram bars lie off the graph, to the left or to the right, a large arrow indicates the number of these outliers.

Plot 8. Main-chain bond angle distributions

Like Plot 7, but for the main-chain bond angles.

Plot 9. RMS distances from planarity

These histograms show the RMS distances from planarity for the different planar groups in the structure. The dashed lines indicate different ideal values for aromatic rings (Phe, Tyr, Trp, His) and for planar end-groups (Arg, Asn, Asp, Gln, Glu). The default values are 0.03Å and 0.02Å, respectively, but these values can be altered by editing the **procheck.prm** file (see section 5.2.9 of the main part of these Operating Instructions).

Histogram bars beyond the dashed lines are shown as highlighted.

Plot 10. Distorted geometry plots

The final set of plots shows all distorted main-chain bond lengths, main-chain bond angles, and planar groups.

The parameters defining the degree of distortion for plotting here are given in the **procheck.prm** file which can be edited as described in section 5.2.10 of the main part of these Operating Instructions.

For each main-chain bond length and angle plotted, the plot shows the 'ideal' value (as defined by the Engh & Huber small-molecule data), the actual value, and the difference between the two.

For each distorted planar groups, three orthogonal projections are plotted and the value shown is the RMS distance of the atoms from the best-fit plane.

Appendix E

Residue-by-residue listing

Part 1. Residue information

The first part of the residue-by-residue listing (the **.out** file) deals with a number of stereochemical parameters, as will be described below.

Explanatory notes

The first page gives some explanatory notes about the stereochemical parameters used. These notes include the "ideal" values, and corresponding standard deviations, against which the values calculated for your structure are compared. The "ideals" used here come from an analysis of 118 high-resolution structures performed by Morris *et al.* (1992), and are listed in Table 1 of Appendix A.

Note that, the printing of this explanatory text can be suppressed, if required, by amending the parameter file **procheck.prm** (see section 5.3 in the main body of these Operating Instructions).

Residue-by-residue information

The explanatory text is followed by an analysis of each of the stereochemical parameters for each residue in the structure. Each value is highlighted by asterisks and plus-signs if it deviates from the "ideal" by more than 1 standard deviation. An asterisk represents one standard deviation, and a plus-sign represents half a standard deviation. So, a highlight such as +*** indicates that the value of the parameter is between 3.5 and 4.0 standard deviations from the ideal. Where the deviation is more than 4.5 standard deviations, its numerical value is shown instead: for example, ***5.5*** represents 5.5 standard deviations.

The appearance of the listing can be altered to some extent by editing the parameter file **procheck.prm**. This allows you to show, say, only the asterisks and not have the values themselves printed. You can also include only those values that are more than a given number of standard deviations from the ideal. For more information, see section 5.3 in the main body of these Operating Instructions.

The information shown for each residue is as follows:

1. **Residue number** - as given in the original coordinates file.

2. **Chain identifier** - where relevant, picked up from the original coordinates file.

3. **Sequential number** - starting at 1 for the first residue and numbering the residues sequentially from then on. This may differ from the residue numbering given in the original coordinates file.

4. **Kabsch & Sander secondary structure assignment** - assignment of secondary structure according to the method of Kabsch & Sander (1983). The codes used are as follows:

B	–	residue in isolated beta-bridge	**S**	–	bend
E	–	extended strand, participates in beta-ladder	**T**	–	hydrogen-bonded turn
			e	–	extension of beta-strand
G	–	3-helix (3/10 helix)	**g**	–	extension of 3/10 helix
H	–	4-helix (alpha-helix)	**h**	–	extension of alpha-helix
I	–	5-helix (pi-helix)			

The lower-case assignments are our extensions of the Kabsch & Sander definition and are obtained by slightly relaxing their criteria.

5. **Region of Ramachandran plot** - a single letter code identifies which region of the Ramachandran plot the residue is in. For end residues and glycines this assignment does not apply, so is shown by a hyphen, '–'. The other codes are as follows:

A	–	Core alpha	**L**	–	Core left-handed alpha
a	–	Allowed alpha	**l**	–	Allowed left-handed alpha
~a	–	Generous alpha	**~l**	–	Generous left-handed alpha
B	–	Core beta	**p**	–	Allowed epsilon
b	–	Allowed beta	**~p**	–	Generous epsilon
~b	–	Generous beta	**XX**	–	Outside major areas - *ie* disallowed

6. **Chi-1 dihedral angle** - three separate columns are given for the three possible conformations of χ_1: *gauche minus, trans,* and *gauche plus.*

7. **Chi-2 dihedral angle** - only the values for the χ_2 dihedral angles in the *trans* conformation are shown.

8. **Proline phi** - the ϕ torsion angle for proline residues only.

9. **Phi helix** - the ϕ torsion angle for all residues identified as being in an alpha-helix by the **H** of the Kabsch & Sander secondary structure assignment code.

10. **Helix psi** - as above, but for the ψ torsion angle.

11. **Chi-3 dihedral angle** - being the torsion angle defined by the S–S bridge in a disulphide bond, with separate columns for the right- and left-handed conformations.

12. **Disulph bond** - sulphur-sulphur distance, in Å, between paired cysteine residues.

13. **H-bond en.** - estimated strength of the main-chain hydrogen bond (in kcal/mol), where applicable, calculated using the method of Kabsch & Sander (1983).

14. **Chirality C-alpha** - value of the ζ "virtual" torsion angle, defined by the atoms Cα, N, C, and Cβ. This is a "virtual" torsion angle as it is not defined along an actual *bond.*

15. **Bad contacts** - number of bad contacts for this residue, as defined by non-bonded atoms at a distance of ≤ 2.6Å. The bad contacts are listed at the end of the print-out (see Part 3 below).

16. Max dev - this shows the maximum deviation (in terms of asterisks, *etc*) of all the columns in the current row.

At the end of this print-out, the column totals show the maximum deviation in each column, the column's mean value and standard deviation, and number of values it contains. If the mean values themselves deviate significantly from the "ideals", they too are highlighted by asterisks.

Part 2. Main-chain bond lengths and bond angles

The second part of the listing analyses the main-chain bond lengths and bond angles of your protein structure. As before, any deviations in the actual bond-lengths and bond angles from the "ideal" values are highlighted with asterisks and plus signs.

At the end of this print-out, the different bond lengths and bond angles are summarised in two tables giving the minimum, maximum, and mean values of each type, together with their standard deviations.

The "ideal" values used are given at the head of the listing (though the printing of these can be suppressed by amending the parameter file **procheck.prm**). The ideals are as determined from the analysis of small-molecule data by Engh & Huber (1991) and are shown in Table 2 of Appendix A.

Part 3. Bad contacts listing

The bad contacts listing shows the atom-pairs involved, the type of contact, and the separation between the two atoms. As already mentioned, bad contacts are defined here as any pair of non-bonded atoms that are at a distance of $\leq$ 2.6Å from one another.

Part 4. Summary statistics and quality assessment

The final part of the print-out reproduces the statistics printed on Plots 1, 4 and 5 (Appendix D). It also gives an overall assessment of the structure's quality using the Morris *et al.* (1992) stereochemical classification scheme. Here a number from 1 to 4 is assigned to the structure for each of three separate stereochemical parameters (1 being the best and 4 the worst score). Finally, it prints an analysis of the various overall G-factors calculated for the structure. Any G-factors below **-1.0** may indicate properties that need to be investigated more closely.

References

Adobe Systems Inc. (1985). *PostScript Language Reference Manual.* Addison-Wesley, Reading, MA.

Engh R A & Huber R (1991). Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Cryst.*, **A47**, 392–400.

IUPAC-IUB Commission on Biochemical Nomenclature (1970). Abbreviations and symbols for the description of the conformation of polypeptide chains. *J. Mol. Biol.*, **52**, 1–17.

Kabsch W & Sander C (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Laskowski R A, MacArthur M W, Moss D S & Thornton J M (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, **26**, 283–291.

Morris A L, MacArthur M W, Hutchinson E G & Thornton J M (1992). Stereochemical quality of protein structure coordinates. *Proteins*, **12**, 345–364.

Nishikawa K & Ooi T (1986). Radial locations of amino-acid residues in a globular protein - correlation with the sequence. *J. Biochem.*, **100**, 1043–1047.