



| [Useful links](#) | [Glossary of terms](#) | [Google](#) |



Now it's time to get acquainted with some of the validation tools that are available on the world-wide web. We shall use the PDB entry [1CBS](#) as an example.

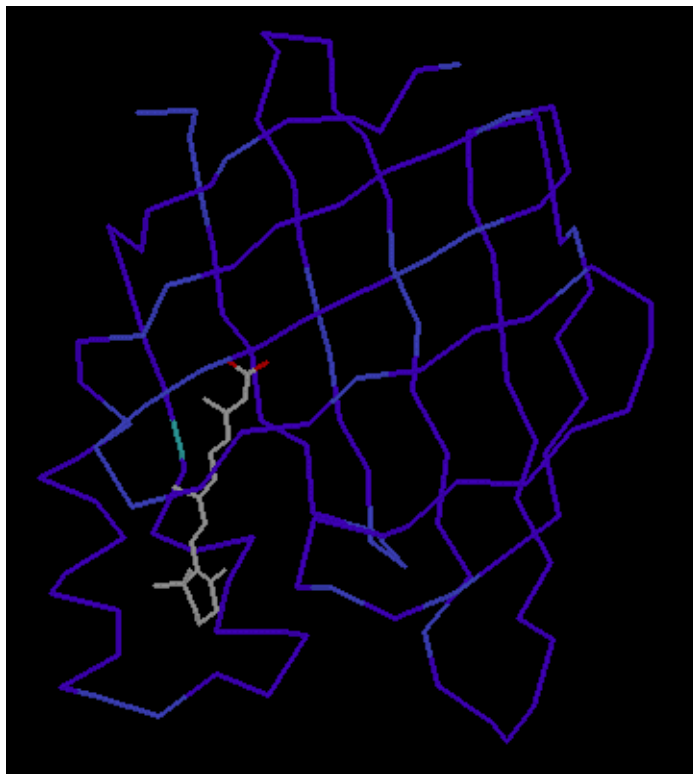
RCSB

Go to the [RCSB page for entry 1CBS](#). Some of the information on this page can be used for some "validation-lite":

- the resolution of the study
- the conventional and free R-value
- if there appears an item called "Structure factors" in the menu on the left, this means that the authors have deposited their experimental crystallographic data, which inspires some confidence, but more importantly enables us to correlate the model and the data ourselves
- the author list, journal and date :-)

If you follow the link to "Download/Display File" and look at the header of the PDB file, you may sometimes find a bit more detail about the model, the data, and the refinement process.

If you follow the link to "Geometry", you find more information about the covalent geometry of the model. There is information about bond lengths, angles and dihedrals (lots of blue entries in the tables are good news; lots of red entries are not). You can also inspect the geometric quality of the model graphically by clicking on the link "Fold Deviation Score". Before you do this, you must define yet another MINE-type, namely for `chemical/x-ras`, which should be set up identically to `application/x-rasmol`, *i.e.* launching Rasmol with a script as input (if you have forgotten how to do this, [check here](#)). Again - red is bad, m'kay ?



Fold Deviation Score display.

Q. 10. What are the values of R_{free} and of $R_{\text{free}} - R$ for 1CBS ? Are these good or bad ?

MSD

Basic validation-related information is also available from the [MSD page for entry 1CBS](#):

- the resolution of the study
- the conventional and free R-value (for most entries)
- if there is a link to download structure factors then that is reassuring (see above)
- the author list, journal and date again :-)

PDBj

Basic validation-related information is also available from the third partner of the wwPDB, the Protein databank of Japan (PDBj). Check the [PDBj page for entry 1CBS](#). For instance, under "Experimental details" you once again find the resolution limits and the values of the conventional and free R-values.

Note that PDBj also provides a mirror of the RCSB pages, *e.g.*, the page for 1CBS lives [here](#).

PDBsum

From the RCSB entry (under "Other Sources") and the MSD entry (under "Similarity") there are links to the PDBsum pages for 1CBS.

If you use the old PDBsum site (at UCL), you get to [this page](#) where you find:

- information about resolution, R and R_{free}
- a diagram of the secondary structure. Usually, around 60% of all residues in a protein is part of a regular secondary-structure element

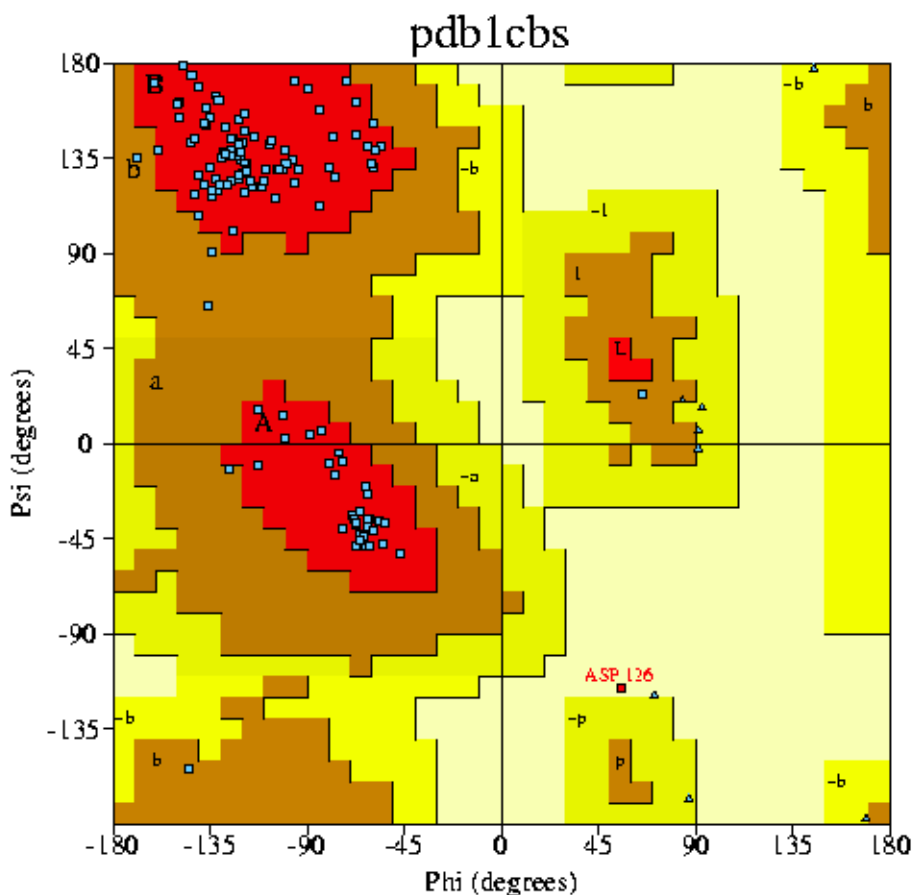
(although there are exceptions)

- a diagram ("LIGPLOT") of the interactions between the protein and a ligand, retinoic acid. If the fold of a protein model is correct, we expect that "sensible" residues will interact with ligands, substrates, ions, *etc.*
- at the top of the page there is a link to "PROCHECK" (both in the list of links on the right, and through the miniature Ramachandran plot icon) which provides a summary of the results of the PROCHECK validation program

The PROCHECK results come in three parts:

- a Ramachandran plot
- statistics pertaining to the Ramachandran plot. A good model would be expected to have > 90% residues in the core allowed areas, and no more than 1-2% in the disallowed ones
- a list of G-factors (or Geometry-factors, so named in analogy to crystallographic R-factors) that provide information as to how unusual various aspects of the model are. Positive values are good, but remember that bond lengths and angles are usually restrained during model refinement (as opposed to the phi, psi distribution)

If you use the new site of PDBsum (at EBI), you get to [this page](#) where you find essentially the same information, but spread out over multiple webpages. The top page contains information about resolution, R-values, authors *etc.*, the protein page contains information about the secondary structure *etc.*, the ligands page shows the LIGPLOT diagram, and both the top and the links page provide a link to the PROCHECK results.



PROCHECK Ramachandran plot. The red regions are the core allowed regions. Additional allowed (by PROCHECK, that is) regions are in brown, and generously allowed regions in dark yellow. The disallowed regions are in a lighter shade of yellow.

Q. 11. Are the interactions between the protein in 1CBS and its ligand "sensible" ?

Q. 12. Based on the PROCHECK output, what do you think of the quality of 1CBS so far ?

PDBreport

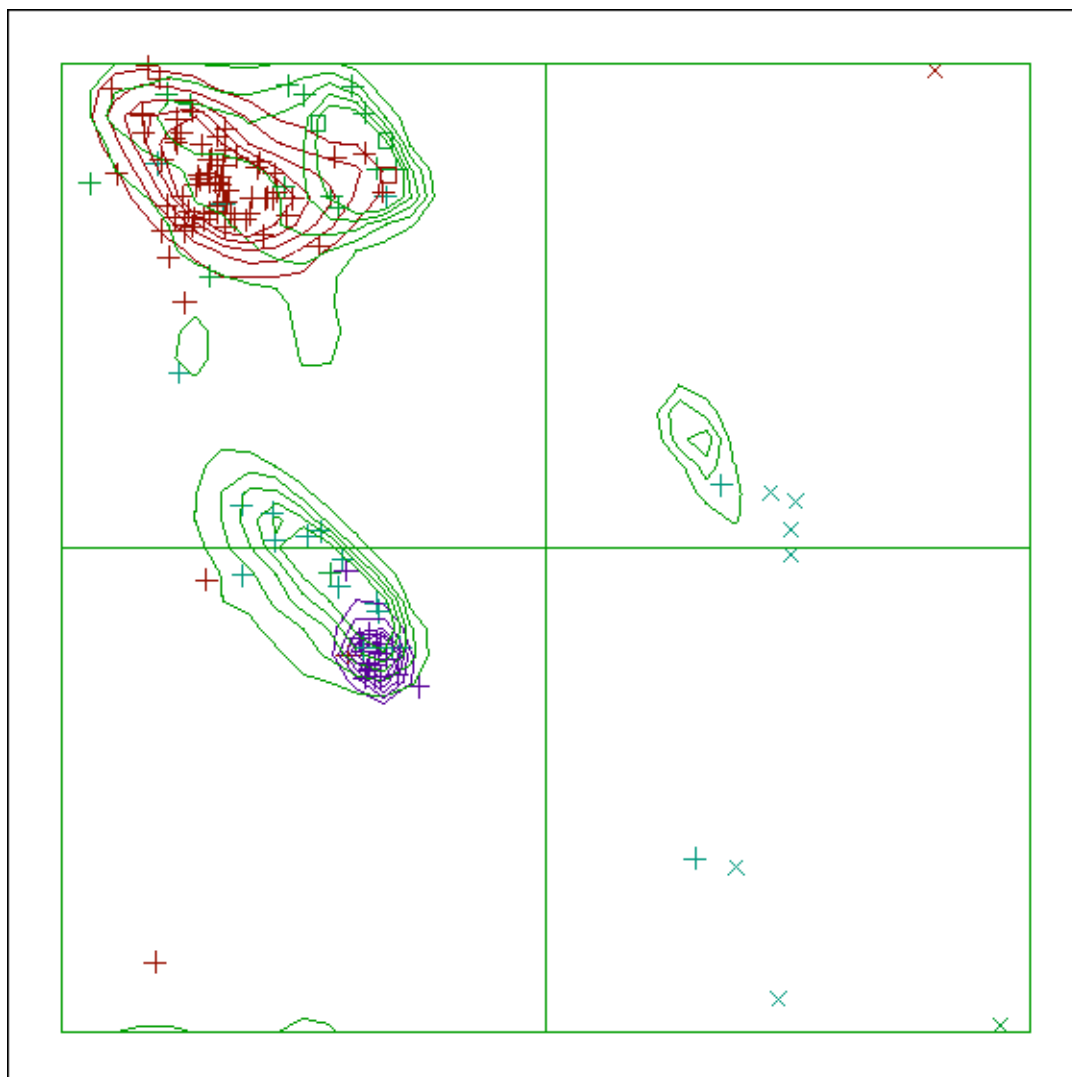
The PDBreport database contains quality analyses carried out with the program WHAT IF (or, rather, a subset called WHATCHECK). It can be reached via the "Other Sources" page of the RCSB entry, or via the WHATCHECK link from PDBsum. At the top of the PDBreport page for 1CBS there is a link to the "Full report". On that page you will be presented with the complete report from WHAT IF. This includes a large number of checks and tests. For a description, look [here](#), and for a discussion, look [here](#).

The diagnostics come in three classes of severity:

- "note" - no problem, mon !
- "warning" - something requires your attention
- "error" - something appears to be seriously amiss

The most useful checks are:

- the Ramachandran plot
- several of the 3D-database-related checks (including the packing scores, quality value plot, and rotamers)
- one or both summaries



WHAT IF Ramachandran plot.

At the bottom of the output, there are two summaries, one for users of a model (comparing the quality of this model to a set of high-resolution, reliable models), and one for the person who deposits the model (comparing the quality of this model to a set of structures solved at similar resolution). In particular the list of "Structure Z-scores" of the first summary provides a quick overview of the overall quality of the model.

Final summary

Note: Summary report for users of a structure

This is an overall summary of the quality of the structure as compared with current reliable structures. This summary is most useful for biologists seeking a good structure to use for modelling calculations.

The second part of the table mostly gives an impression of how well the model conforms to common refinement constraint values. The first part of the table shows a number of constraint-independent quality indicators.

Structure Z-scores, positive is better than average:

```
1st generation packing quality : 1.331
2nd generation packing quality : 2.436
Ramachandran plot appearance  : 1.673
chi-1/chi-2 rotamer normality : 0.866
Backbone conformation         : 1.547
```

RMS Z-scores, should be close to 1.0:

```
Bond lengths           : 0.433 (tight)
Bond angles            : 0.676
Omega angle restraints : 0.343 (tight)
Side chain planarity   : 0.653 (tight)
Improper dihedral distribution : 0.660
B-factor distribution  : 0.626
Inside/Outside distribution : 1.010
```

WHAT IF summary intended for users of a model.

Q. 13. Does WHAT IF detect any serious problems in 1CBS ? What do you think of 1CBS now ?

EDS

Finally, we shall have a look at [EDS](#), the Uppsala Electron-Density Server. This facility provides information about the model and its fit to the experimental data.

Go to the EDS page for PDB entry 1CBS. Information about this entry includes (click on the question mark images for more information about individual items):

- in the summary on the right, the resolution *etc.* is listed again. In addition, the R-value calculated by REFMAC (the program used to calculate the electron-density map) is listed. In this case, this "R factor for map" is in fact lower than the R-value reported (this may be due to a different selection of reflections to include in the calculations, different degrees of sophistication of the programs used now and then, *etc.*). If the map R-factor is considerably higher than the reported R-value, there may be reason for concern, though. This box also contains some statistics that summarise the results of the real-space fit calculations. The box below it contains a few selected records from the header of the PDB file that usually enable you to identify both the nature of the molecule(s) in the PDB entry, and the authors.

- plots of the real-space (RS) fit (R-factor and correlation coefficient) as a function of residue number. If JavaScript is enabled in your browser, when you move the cursor over the graph, the identity and real-space fit value of each residue will be displayed. If you have a sufficiently advanced Java plug-in, a mouse-click will start up an interactive viewer that will show you the residue you clicked on, its environment, and the electron density in the neighbourhood. This is extremely useful when you want to check that a particular residue or part of the model that is important for your work has good electron density (and is not a figment of the crystallographer's imagination ...).

[Note: if you are familiar with a density-viewing program such as O or SwissPDBViewer, you don't need to use the EDS Viewer. Instead, you can download the electron-density map and structure from the main page of each PDB entry (follow the "All files (.tar.gz)" link), and inspect them with your own viewer.]

- a plot of the occupancy-weighted average temperature factor as a function of residue number.
- the link "Significant regions" gives you a plot that shows residues that have considerably worse density than the average for that residue type in structures at similar resolution. This is derived from the "Z-scores" plot by only showing residues that differ more than two sigma from the sample average.
- the Ramachandran plot is included using yet another definition of "good, bad and ugly", namely that of MOLEMAN2.
- Some files for downloading onto your own computer (only of use if you know what to do with them, of course).
- A set of links specific for this PDB entry, to resources that provide information about the entry or its quality.

Q. 14. Which amino-acid residue in 1CBS has the worst RS-fit value ?

Q. 15. How good is the electron density for the ligand in 1CBS, compared to that of the protein ?

Your own models

All the tools described so far contain pre-cooked information (although it is sometimes generated on-the-fly) and only about models that have already been deposited in the PDB. If you want to assess the quality of a model that is not yet in the PDB, you can use a number of web-based servers (also if you want to assess quality aspects that are not covered by the resources discussed above). For more information, see the [Useful links](#) page.



Practical "Model Validation" - EMBO Bioinformatics Course - Uppsala 2001 - © 2001-2004 [Gerard Kleywegt](#)

Latest update at 3 September, 2004.

000785