Structure, Vol. 10, 465–472, April, 2002, ©2002 Elsevier Science Ltd. All rights reserved. PII S0969-2126(02)00743-8

Homo Crystallographicus—Quo Vadis?

Ways & Means

Gerard J. Kleywegt[1] and T. Alwyn Jones
Department of Cell and Molecular Biology
Uppsala University
Biomedical Centre
Box 596
SE-751 24 Uppsala
Sweden

Summary

As macromolecular crystal structures are determined and refined in an increasingly automated fashion, careful assessment of the reliability and quality of the resulting models becomes increasingly important. Here, we analyze various issues related to the reliability and quality of macromolecular crystal structures deposited between 1991 and 2000. We find that the average resolution at which these structures are determined is essentially constant. In line with this observation, the average quality as measured by Ramachandran analysis does not improve as a function of time. On the other hand, an observed decrease of the average discrepancy between free and conventional R values suggests that the fit of model and data is improving. Finally, we present a surprising correlation between the tendency of crystallographers to deposit their experimental data and the free R values of their models.

Introduction

In the mid-1990s, a series of "angry young men" articles were published in this journal that described the results of an assessment of a number of aspects related to quality control and model validation in protein crystallographic structure determination [1–3]. In the first paper ("Where freedom is given, liberties are taken") [1], it was shown how contemporary model refinement and publishing practices could lead to the publication of seemingly correct models that are in actual fact completely wrong. The second paper ("Checking your imagination") [2] was a survey of the use and applications of the free R value and discussed a number of unresolved issues related to the use (and abuse) of the free R value in model refinement. In the third paper ("Phi/psi-chology: Ramachandran revisited") [3], the usefulness of the Ramachandran plot as a simple yet powerful means of gaining a quick impression of the quality of a protein model was emphasized. On that occasion, a binary division of the Ramachandran-plot area in favorable (or core) and unfavorable regions was also introduced.

Half a decade later, structural biology is rapidly moving into a phase where structures will be determined on an industrial scale, with a reduced level of intervention by human experts. Moreover, structural biologists who determine structures at a conveyer belt will not have the same affinity (dare we say "love affair"?) with their proteins as the crystallographers who had to toil for years to determine just a single structure. Although increased automation might result in a reduction of human errors during model building, it may equally well lead to an increase of errors if too much faith is put in results obtained with magical black boxes. Although much has happened in the area of protein model validation in the past decade [4], it remains to be seen whether automatic error detection and correction protocols (in particular at low resolution) can rival skilled crystallographers.

At the dawn of the new millennium, we have reinvestigated the use of the free R value and reassessed the quality of protein models as judged by their Ramachandran plots, and looked for trends.

Statistics

Using the March 2001 version of the Protein Data Bank (PDB) [5, 6], we extracted information regarding the resolution and R values of 10,888 entries that were deposited between 1991 and 2000 and that were determined to a resolution no worse than 4.0 Å. Entries for which unresolvable errors or ambiguities in the R-value statistics were encountered were omitted, as were peptide structures containing fewer than 20 amino acid residues (nucleic acid structures were included, however). A total of 10,674 entries remained, and these were subjected to our analyses. For each entry, we recorded the PDB identifier, year of deposition, resolution, conventional R value, free R value, the difference between the free and conventional R value, the number of amino acid residues (for entries containing one or more proteins), and the number and percentage of Ramachandran-plot outliers (using our definition of outliers) [3], although not all of these quantities are necessarily defined for each entry. Table 1 shows an overview of the results obtained as a function of the year of deposition. In the following, we shall refer to a number of so-called "TIE-fighter plots" (also known as "box plots" or "box-whiskers plots"); the meaning of the various markers in such plots is explained in Figure 1.

Resolution

Conventional wisdom has it that the average resolution at which macromolecular crystal structures are determined is improving all the time, not in the least due to the use of cryocooling techniques and synchrotron radiation sources. However, both Table 1 and Figure 2 show this assertion to be false—there is essentially no correlation between resolution and year of deposition, and the average resolution for the decade is ~2.2 Å. Although the average resolution has improved very slightly since 1996, Figure 2 reveals that the 10th and 90th percentiles have been almost constant since 1992. As Table 1 shows (the row marked "<Residues>," i.e., the average number of amino acid residues found in

Key words: X-ray crystallography; protein structure; model validation; Ramachandran analysis; protein model quality

[1] Correspondence: gerard@xray.bmc.uu.se

Table 1. Overview of Results

Year	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	Overall
Nr X-ray structures	133	386	600	774	879	1100	1432	1727	1981	1662	10674
<Resolution> (Å)	2.179	2.130	2.184	2.166	2.202	2.236	2.175	2.180	2.141	2.122	2.168
<R value>	0.183	0.178	0.181	0.182	0.184	0.189	0.193	0.195	0.198	0.202	0.192
Nr with map in EDS	18	36	108	178	309	439	625	696	736	109	3254
% of total	13.5	9.3	18.0	23.0	35.2	39.9	43.6	40.3	37.2	6.6	30.5
Nr with R_{free}	0	0	6	55	289	528	1022	1413	1713	1534	6560
% of total	0	0	1	7	33	48	71	82	86	92	61
<R_{free}>	—	—	0.263	0.265	0.256	0.258	0.257	0.255	0.251	0.250	0.254
<R_{free}−R value>	—	—	0.072	0.068	0.062	0.063	0.060	0.057	0.050	0.047	0.054
Nr Ramachandran	127	382	567	731	833	1043	1354	1660	1908	1610	10215
% of total	95	99	95	94	95	95	95	96	96	97	96
<Residues>	385	305	319	340	387	433	461	466	486	549	448
<% Rama outliers>	4.1	3.5	4.2	3.8	4.0	4.0	3.5	3.7	3.5	3.6	3.7
Max % Rama outl	39.9	24.0	31.0	30.2	26.6	36.8	35.0	33.7	43.5	40.7	43.5

Various statistics are listed as a function of year of deposition of crystal structures in the PDB. The first five rows pertain to all crystal structures included in this survey. The subsequent set of four rows pertains to all structures for which a free R value could be extracted from the PDB entry. The final set of five rows pertain to all entries that contain one or more protein molecules. Average values are indicated by brackets.

PDB entries that contain one or more protein molecules), the improvements in the area of data collection appear to have been exploited to study larger and larger systems. The average number of amino acid residues of the protein-containing PDB entries included here has increased by roughly 40% when comparing the year 1991 (average 385 residues) and the year 2000 (average 549 residues).

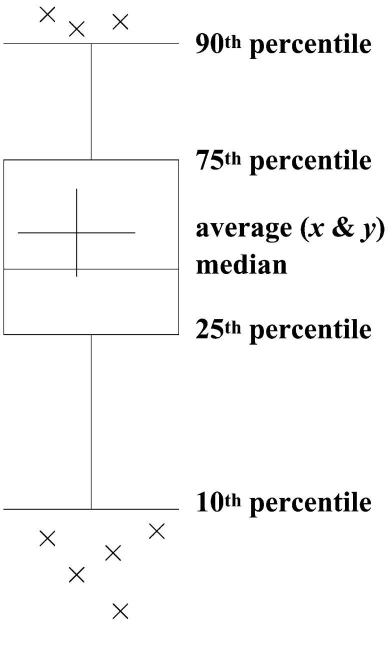

Figure 1. Explanation of the Markers Used in "TIE-Fighter Plots"

Rather than showing scatter plots containing more than 10,000 data points (one for each PDB entry in the survey), the data are binned (using the variable along the horizontal axis) and summarized for each bin using a box and several markers. The "whiskers" indicate the 10th and 90th percentile of the data in the bin, whereas the upper and lower boundaries of the box indicate the 25th and 75th percentile. The horizontal line inside the box indicates the median (i.e., the 50th percentile), and the crosshair inside the box indicates the average value (in both directions). Finally, in order to show the distribution of the "outliers," all individual data points outside the whiskers are shown. In addition, for bins containing fewer than ten data points, all points are shown and the box and whiskers are omitted. (Note: Star Wars aficionados will understand why these plots are called "TIE-fighter plots.")

R Values

Paradoxically, perhaps, the average R value has been increasing since the mid-1990s (Table 1). This is in agreement with our suggestion that many structures were over-fitted in the past, in particular in low-resolution studies and in cases involving the use (and sometimes abuse) of noncrystallographic symmetry (NCS) [1, 7, 8], leading to essentially meaningless low R values. Since the average resolution of the structures varies little over time, the increase of the average R values may be due to improved refinement protocols (e.g., restraining NCS, monitoring R_{free}) and refinement methodology (in particular, the use of maximum likelihood targets). Nevertheless, the distribution of R values versus resolution still shows an immense "plateau" of structures with R values close to 0.2 (Figure 3). In fact, 92% of all structures have an R value between 0.15 and 0.25.

It is interesting to note how rapidly the use of the free R value has become common practice since 1996. Initially, there were doubts as to the benefits and drawbacks of crossvalidation, but the work of Brunger et al., ourselves, and others (see, for instance, [2, 9–14] as well as the discussion described in [15]) appears to have convinced most crystallographers that the benefits far outweigh the drawbacks. In the year 2000, 92% of all crystal structure depositions reported a free R value, up from only 33% in 1995.

The average value of the free R value shows a slight tendency to drop with the progression of time (Table 1). We suspect that this too is due to the use of improved methods and, we hope, the adoption of more sensible refinement and rebuilding protocols by a large number of crystallographers.

Crystallographers often wonder what free R values to expect at a certain resolution. Although it doesn't answer that question, Figure 4 does reveal the distribution of free R values as a function of resolution. At present, only 0.7% of all entries have a free R value of less than

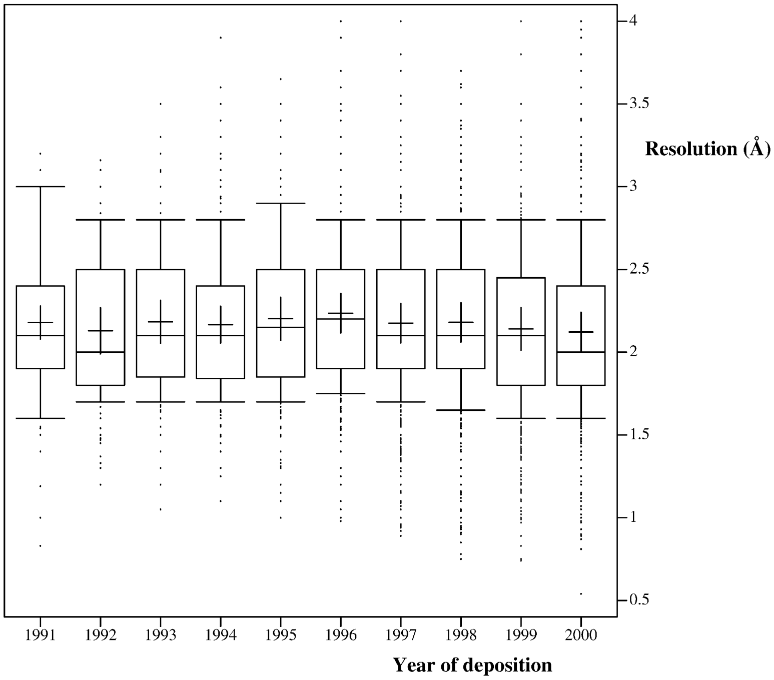

Figure 2. Average Resolution Remains Constant in Time

The average resolution of macromolecular crystal structures does not vary appreciably with time. The linear correlation coefficient of resolution and year of deposition is essentially zero (−0.04 using all data points; −0.21 if only the ten bin averages are taken into account).

0.15, and 1.2% have a free R value that exceeds 0.35. However, Figure 4 also shows that there are a number of remarkable outliers.

As one would hope, the free and conventional R values are very strongly correlated (linear correlation coefficient 0.77 if all entries are used, or even 0.99 if the data are binned and only the 17 bin averages are used). A linear fit using all data points yields the following regression line: $R_{free} = 1.065\,R + 0.036$. This implies that for a model with a conventional R value of 0.1, 0.2, or 0.3, one would expect to find a free R value of roughly 0.14, 0.25, and 0.36, respectively.

The slow but steady increase of the average R value (as a function of time) and the decrease of the average free R value conspire to reduce the average difference between the free and conventional R value (Figure 5;

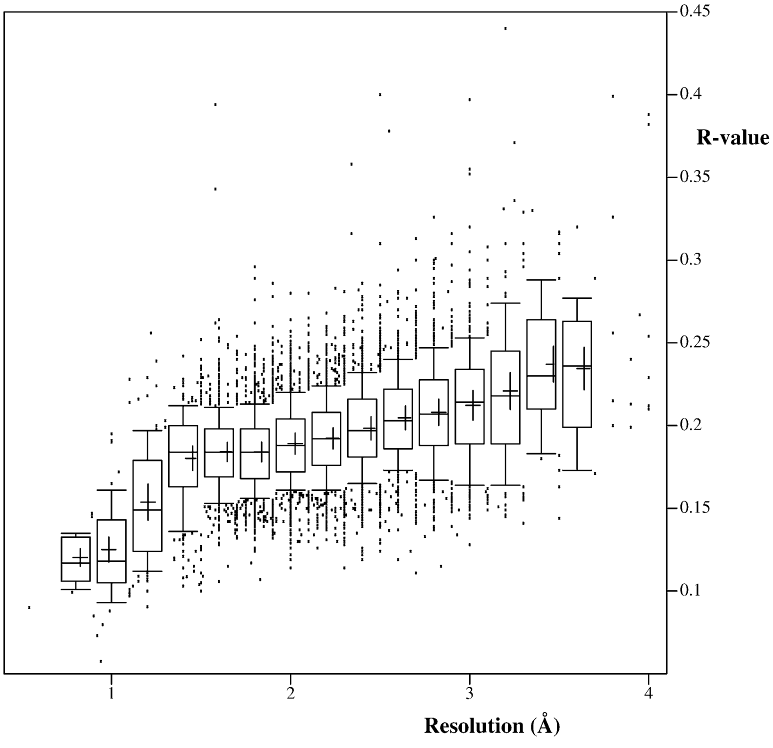

Figure 3. The Distribution of R Values versus Resolution

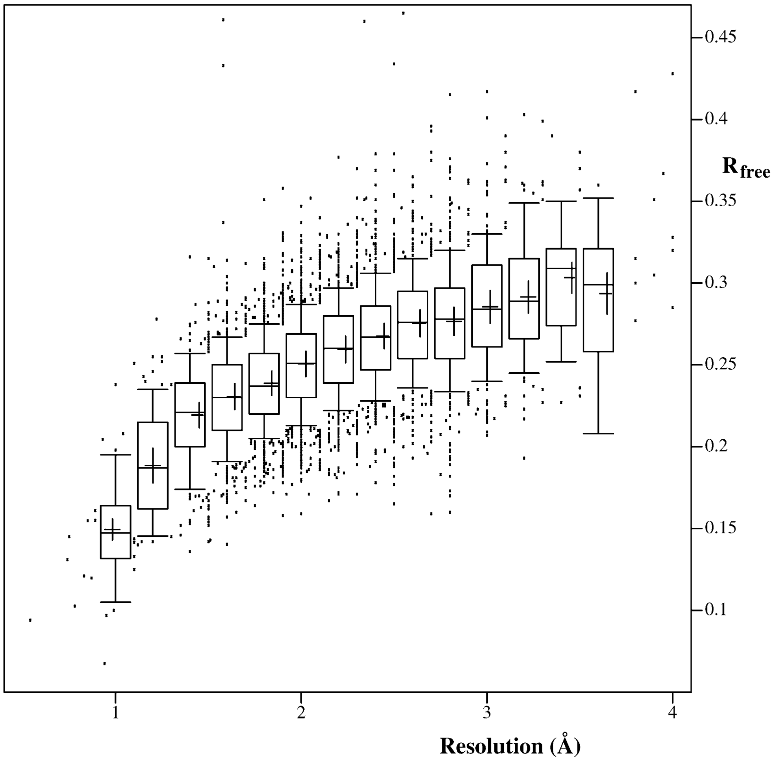

Figure 4. The Distribution of Free R Values versus Resolution

the linear correlation coefficient of the bin averages of this difference and the year of deposition is −0.97). The discrepancy between conventional and free R values can be attributed to two main factors (apart from noise in the data): incompleteness of the model (in particular, if no maximum likelihood target is used) and over-fitting (i.e., introducing more parameters in the model than is warranted by the information contained in the data). In other words, decreasing free R values (at roughly constant resolution) and decreasing differences between free and conventional R values both indicate that crystallographers (and their software) are on the whole

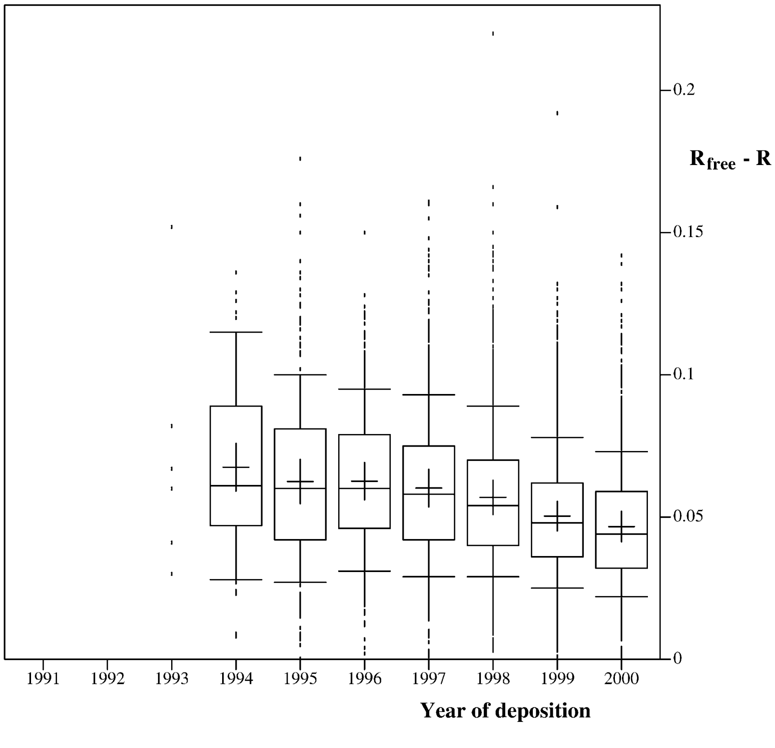

Figure 5. The Distribution of the Discrepancy between the Free and Conventional R Value versus Year of Deposition

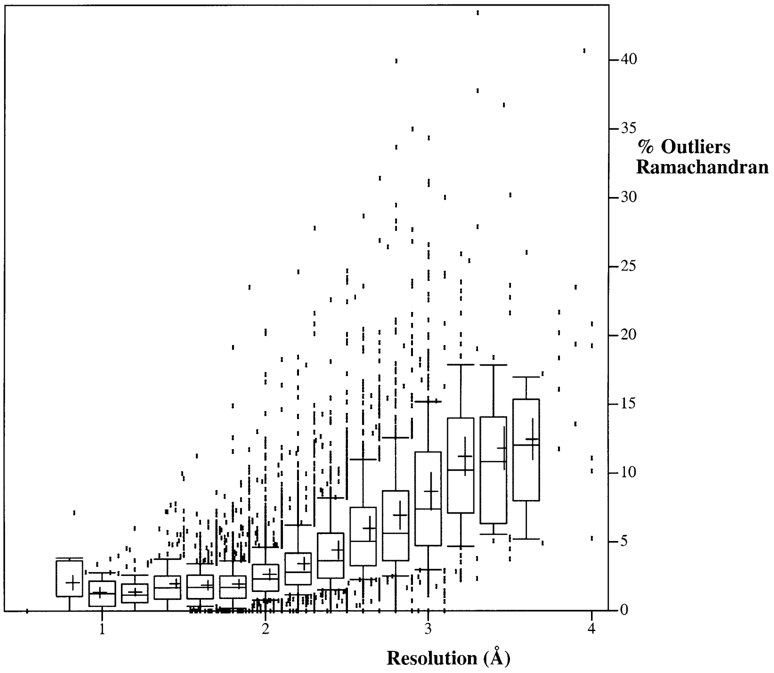

Figure 6. The Distribution of the Percentage of Ramachandran-Plot Outliers versus Resolution

doing a better job and producing better models (in terms of their explaining the experimental data).

Ramachandran Analysis

As Table 1 shows, the average percentage of Ramachandran-plot outliers (using our definition) [3] is essentially constant as a function of time, indicating that the average overall quality of the protein models that the community produces does not get any better. This may well be related to the observation that the average resolution of the studies has not changed appreciably during the past decade: only improved resolution can result in truly better models and thereby lower the fraction of Ramachandran-plot outliers.

The distribution of the percentage of Ramachandran-plot outliers versus resolution (Figure 6) shows that both the average and the 90th percentile are more or less constant up to about 2 Å resolution, and increase roughly linearly after that. When we delineated the core and noncore regions of the Ramachandran plot, we found that for structures refined to better than 2 Å resolution, we expected to find 0%–5% outliers in the Ramachandran plot. This rule of thumb is still valid today, as demonstrated by Figure 6.

Figure 7 shows the correlation between a statistic that measures the quality of a protein model (percentage of Ramachandran-plot outliers) and one that measures the quality of the fit of the model to the experimental data (free R value). Ideally, both values should be small, so that the bottom left area of this plot contains all the high-quality models. The top right area, on the other hand, probably includes models for which considerations other than protein crystallographic ones have been deemed crucially important in the refereeing process.

Of the 10,215 protein entries surveyed, 78.4% have fewer than 5% outliers in the Ramachandran plot (up from 76.5% in our earlier study) [3], and 6.2% have more than 10% outliers (down from 8.9%). In this study, the entry with the poorest Ramachandran plot (deposited in 1999) has no fewer than 43.5% outliers.

In this survey, we have calculated the percentage of Ramachandran-plot outliers for each PDB entry as a whole. However, there are cases where this can give misleading results. For example, a complex of botulinum neurotoxin with a target peptide [16] has recently been the source of some controversy [17, 18]. The overall fraction of Ramachandran-plot outliers for the PDB entry of the complex (1F83) is ~12%. Although this number is quite high in itself, especially when considering that the data extends to 2.0 Å resolution, it would nevertheless not have attracted anyone's attention in Figure 6, for instance. However, when each chain is analyzed separately, a completely different picture emerges: ~71% of the (nonterminal, nonglycine) residues in the peptide (chains B and C) are outliers.

To make it easier to recognize such cases, our web-based Ramachandran servers have been updated to display the results for each chain separately. We have used this facility to investigate how common large variations are in the Ramachandran-plot quality of multiple protein chains that are part of the same crystal structure. We identified 5421 protein crystal structures (deposited between 1991 and 2000, and with a resolution of 4.0 Å or better) that contained more than one protein or peptide chain (consisting of at least 10 residues). For each such entry, we calculated the percentage of Ramachandran-plot outliers of each individual chain, and from that, the difference between the "worst" (highest percentage outliers) and the "best" chain (data not shown). This analysis revealed that the difference is less than 5% for 87.4% of all entries, and between 5% and 10% for an additional 8.4%. Differences exceeding 30% do occur, but they are rare, accounting for 0.2% of all entries.

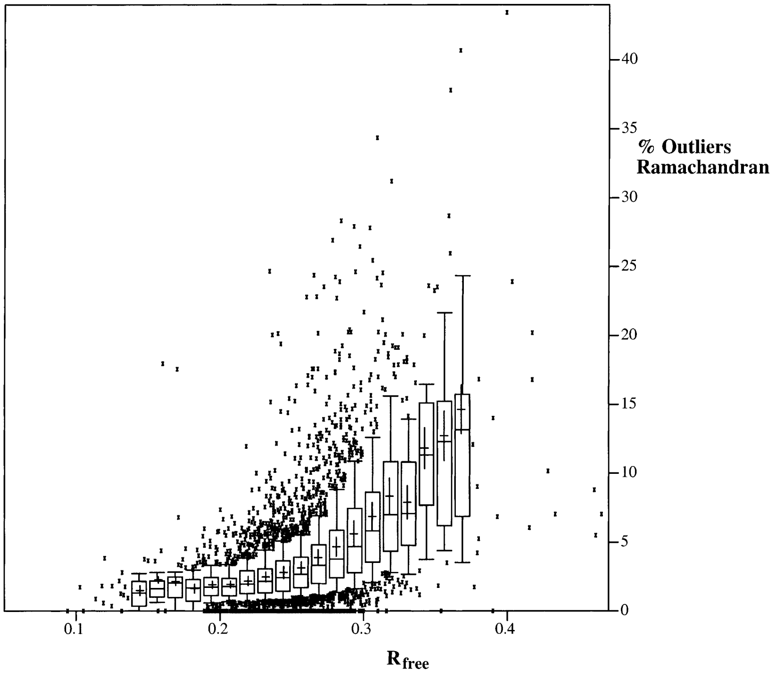

Figure 7. The Distribution of the Percentage of Ramachandran-Plot Outliers versus Free R Value

Data Deposition

In recent years, many journals have imposed (or paid lip service to; J.Y. Zou, M.R. Harris, T. Taylor, A. Wählby, G.J.K., and T.A.J., unpublished data) stricter conditions on the deposition of coordinates and experimental data in public data banks. Access to the experimental data is crucial for other workers to be able to properly assess the quality of (important aspects of) a model [19, 20]. If an electron density map is available, distinguishing unusual but genuine features of a model from model-building errors or refinement artifacts is made much easier. This observation has motivated us to develop the Uppsala Electron Density Server (EDS; J.Y. Zou, M.R. Harris, T. Taylor, A. Wählby, G.J.K., and T.A.J., unpublished data). This server provides access to various statistics, plots, and an electron density map for every PDB entry for which structure factors have been deposited and for which automatic map calculation succeeds (at present ~90% of all entries with structure factors).

Here, we have used the fraction of entries for which

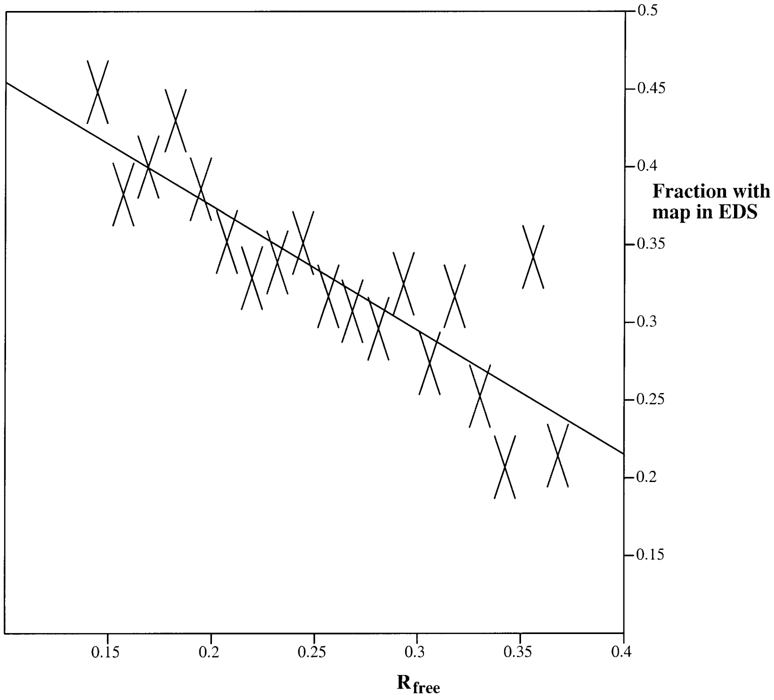

Figure 8. Model Quality Determines Whether Data Is Deposited

The tendency of macromolecular crystallographers to deposit their experimental data is strongly negatively correlated to the free R value of their models.

an electron density map file exists in EDS as an indicator of the extent to which crystallographers deposit their experimental data. As can be seen in Table 1, a map is available for about 30% of the crystal structures (the low fraction for the year 2000 is probably due to structure factor entries being "on hold" and possibly also to processing backlogs at the PDB and/or EDS). The fraction of entries for which EDS maps are available is not correlated with resolution (linear correlation coefficient −0.06 for 15 data points). Interestingly, however, said fraction is strongly correlated with the free R value of the study (linear correlation coefficient −0.86 for 19 data points; Figure 8). This means that the worse the free R value of a model, the less inclined the crystallographer is to deposit the experimental data which the model is supposed to explain! Structure factors are deposited for around 40% of the entries with free R values in the range of 0.15 to 0.2, but this number drops to about 20% for entries with free R values in excess of 0.3. The fraction of entries for which EDS maps are available is also correlated with the percentage of outliers in the Ramachandran plot, but weaker than with the free R value (linear correlation coefficient −0.68 for 14 data points).

Conclusions and Outlook

Looking back on the past decade, we see that the small set of quality indicators used here shows only modest improvements in structure quality. As our science adopts a more industrial approach, it will become even more important to improve methods in general and verification methods in particular.

It appears that improvements in experimental methodology are exploited to tackle larger and larger systems, leaving the average resolution essentially constant at ∼2.2 Å. However, this also means that poor, low-resolution models are here to stay, unless (1) crystallographers become significantly better at building, refining, and validating their models; (2) referees become significantly better at critically evaluating the models and data (and at requesting additional data if these are not provided with the manuscript); and (3) journal editors become significantly better at insisting on deposition of both model coordinates and experimental data, as well as on inclusion of "vital parameters" [21] in papers describing macromolecular crystal structures.

Unfortunately, there still exist a lot of misunderstandings concerning the nature of the crystallographic structure determination process. For example, a recent editorial in *Nature Structural Biology* [22] displayed a surprising naïveté concerning the structure-solving process. It notes that "…there still seems to be some degree of art and skill to solving structures…," despite a decade of literature on how errors can be made in that process (which, ironically, began with a commentary in the parent journal by Brändén and Jones 11 years earlier [23]). Obviously, there is still a need for continuous education within the structural biology community itself. Simultaneously, we need to teach the nonspecialist users of structures to critically assess the quality of models, both with respect to their overall reliability, and in their details (reliability of important residues, ligands, etc.).

Supplementary Material

Supplementary material (including the raw data file, plots, more extensive statistics, various tables, etc.) is available from our web server at http://xray.bmc.uu.se/gerard/supmat/rfree2000. The EDS server (J.Y. Zou, M.R. Harris, T. Taylor, A. Wählby, G.J.K., and T.A.J., unpublished data) can be accessed at http://portray.bmc.uu.se/eds. Crystallographers who wish to produce Ramachandran [3] and "Cα-Ramachandran" [24] plots for their own models can use the MOLEMAN2 server at http://xray.bmc.uu.se/cgi-bin/gerard/rama_server.pl. Ramachandran plots of existing PDB entries can be generated with the Ramachandran server at http://portray.bmc.uu.se/eds/ramachan.html (J.Y. Zou, unpublished data). A web-based tutorial that attempts to introduce nonexperts to the field of protein model validation can be found at http://xray.bmc.uu.se/embo2001/modval.

Acknowledgments

This work was supported by the Swedish Foundation for Strategic Research (SSF), and its Structural Biology Network (SBNet). G.J.K. is a Royal Swedish Academy of Sciences Research Fellow, supported through a grant from the Knut and Alice Wallenberg Foundation.

Received: December 4, 2001
Revised: February 13, 2002
Accepted: February 14, 2002

References

1. Kleywegt, G.J., and Jones, T.A. (1995). Where freedom is given, liberties are taken. Structure *3*, 535–540.
2. Kleywegt, G.J., and Brünger, A.T. (1996). Checking your imagination: applications of the free R value. Structure *4*, 897–904.
3. Kleywegt, G.J., and Jones, T.A. (1996). Phi/psi-chology: Ramachandran revisited. Structure *4*, 1395–1400.
4. Kleywegt, G.J. (2000). Validation of protein crystal structures. Acta Crystallogr. D *56*, 249–265.
5. Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Jr, Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol. *112*, 535–542.
6. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.
7. Kleywegt, G.J., and Jones, T.A. (1995). Braille for pugilists. In Making the Most of your Model, W.N. Hunter, J.M. Thornton, and S. Bailey, eds. (Daresbury, UK: SERC Daresbury Laboratory), pp. 11–24.
8. Kleywegt, G.J. (1996). Use of non-crystallographic symmetry in protein structure refinement. Acta Crystallogr. D *52*, 842–857.
9. Brünger, A.T. (1992). Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature *355*, 472–475.
10. Brünger, A.T. (1993). Assessment of phase accuracy by cross validation: the free R value. Methods and applications. Acta Crystallogr. D *49*, 24–36.
11. Brünger, A.T. (1997). The free R value: a more objective statistic for crystallography. Methods Enzymol. *277*, 366–396.
12. Tickle, I.J., Laskowski, R.A., and Moss, D.S. (1998). R_{free} and the R_{free} ratio. I. Derivation of expected values of cross-validation residuals used in macromolecular least-squares refinement. Acta Crystallogr. D *54*, 547–557.
13. Tickle, I.J., Laskowski, R.A., and Moss, D.S. (2000). R_{free} and the R_{free} ratio. II. Calculation of the expected values and variances of cross-validation statistics in macromolecular least-squares refinement. Acta Crystallogr. D *56*, 442–450.
14. Kleywegt, G.J., and Jones, T.A. (1997). Model-building and refinement practice. Methods Enzymol. *277*, 208–230.
15. Dodson, E., Kleywegt, G.J., and Wilson, K.S. (1996). Report of a workshop on the use of statistical validators in protein X-ray crystallography. Acta Crystallogr. D *52*, 228–234.
16. Hanson, M.A., and Stevens, R.C. (2000). Cocrystal structure of

synaptobrevin-II bound to botulinum neurotoxin type B at 2.0 Å resolution. Nat. Struct. Biol. *7*, 687–692.

17. Rupp, B.R., and Segelke, B. (2001). Questions about the structure of the botulinum neurotoxin B light chain in complex with a target peptide. Nat. Struct. Biol. *8*, 663–664.
18. Stevens, R., and Hanson, M. (2001). Response to Rupp and Segelke. Nat. Struct. Biol. *8*, 664.
19. Jones, T.A., Kleywegt, G.J., and Brünger, A.T. (1996). Storing diffraction data. Nature *381*, 18–19.
20. Abola, E.E., Bairoch, A., Barker, W.C., Beck, S., Benson, D.A., Berman, H., Cameron, G., Cantor, C., Doubet, S., Hubbard, T.J.P., et al. (2000). Quality control in databanks for molecular biology. Bioessays *22*, 1024–1034.
21. Borhani, D. (2001). Vital parameters need to be in print. Nature *413*, 249.
22. Editorial. (2001). Healthy debate of structural analysis. Nat. Struct. Biol. *8*, 729.
23. Brändén, C.I., and Jones, T.A. (1990). Between objectivity and subjectivity. Nature *343*, 687–689.
24. Kleywegt, G.J. (1997). Validation of protein models from Cα coordinates alone. J. Mol. Biol. *273*, 371–376.