Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Notes of a protein crystallographer: on the high-resolution structure of the PDB growth rate

Cele Abad-Zapatero

Center for Pharmaceutical Biotechnology, University of Illinois at Chicago, MBRB 3020 MC 870, Chicago, IL 60607, USA

Correspondence e-mail: caz@uic.edu

Dedicated to the people that during the past 40 years have worked indefatigably at the PDB, making it an invaluable resource of structural information to the biomedical community at large.

On 28–30 October 2011, a meeting organized by the World Wide Protein Data Bank (wwPDB) to commemorate the birth of the PDB was held at Cold Spring Harbor Laboratory. It has been 40 years since the incipient community of macromolecular crystallographers came together at this same location to discuss the latest developments in the young field of protein crystallography. On that occasion, macromolecular crystallographers from all over the world felt confident to discuss their latest results, originally impelled by the breakthroughs of myoglobin and hemoglobin, and declared that protein crystallography had its 'Coming of Age' (Phillips, 1971). Although there had been discussions at earlier meetings (namely the ACA meeting in 1971) about the interest in creating a worldwide depository of structural data, the consensus is that it was at the Cold Spring Harbor meeting in 1971 that the seeds for what is now the wwPDB were planted. I have written before about this seminal meeting and my personal memories and recollections connected to my graduate student training and my initiation as a protein crystallographer (Abad-Zapatero, 2011).

This time, I would like to use these notes with a different focus. Normally, these notes have been reflective and even philosophically so. This time, I would like to make them 'analytically reflective' in the sense that I am going to base them on numerical experimental data (not anecdotes or personal insights) and that my reflections will be based on these data. Why not write a formal scientific paper? I have not for two reasons. Firstly, the data do not permit a conclusive and definitive proof of my inferences as to the causes of the different rates observed: establishing a cause–effect relationship in deposition rates is very difficult. Secondly, the conclusions will depend on the future. I will only be analyzing retrospective changes and do not intend to predict the future, only glimpse at it.

The history of the PDB has been superbly reviewed by Berman (2008), providing dates, events and data to provide a comprehensive view from its timid beginnings as a community effort to the current achievements and challenges of the wwPDB. She reviews the impact of the technology, the evolution of the content and the challenges and presents a perspective about the future. From this work and other publications on the PDB website, and from the record of structure depositions, the iconic image of the exponential growth of the PDB has developed, beautifully illustrating the growth in content and complexity of the PDB (Fig. 1).

These data can be downloaded from the PDB server in a tabulated form containing year, yearly entry and cumulative content, and are current up to the date of access. Other forms of examining the data are also accessible, searching the content by the method of structure determination (X-ray, NMR or EM) or by resolution, enzyme class *etc*. This time, I would like to take a closer look (a 'higher resolution look') at the growth of the PDB during the past forty years, relate it to the beginnings of the field, examine it in relation to the technical and sociological developments of the field and possibly take a glimpse into the future. One further clarification about the data used should be made. The results presented here correspond to all of the available entries in the PDB as of November 2011. Using only the structures obtained by crystallographic methods (X-ray entries) did not affect the general inferences presented here for two reasons. Firstly, structures solved by alternative methods (namely NMR and EM) only began to be deposited in the late 1990s (5% for NMR, corresponding to 190 structures of the 3816 in total, and 13 EM structures in 2000) and the total number of entries solved by these two methods is only approximately 16% of the total number of entries. Secondly, the number of depositions for structures obtained by NMR and EM experienced a brief surge at the beginning but later slowed to the same deposition rates as those obtained by crystallography or even lower. An independent analysis of deposition rates of structures

determined by NMR and EM, similar to that presented here, could be performed with the data available from the PDB.

Because of the scale and the effort required to convey the complexity and beauty of the structures present, the conventional way of looking at the growth of the PDB (Fig. 1) does not permit a detailed view of its growth. The growth in the number of entries per year is shown in Fig. 2 with an expanded view (inset) of the initial growth rate as shown in the historical letter that Professor Richard Dickerson sent to the community in 1978. A more analytical way is to plot the natural logarithm (ln) of the cumulative number of entries (N) *versus* the year [ln(N) *versus* year]. This representation is more effective at showing the overall rate of growth of the PDB as the slope of a line. Since the dominant feature is a straight line giving the growth rate as ln(N)/year, departures from this overall line are easier to detect (Fig. 3) and thus trends are easier to extract. Are these deviations significant? What do they mean? Can we relate them to technical or sociological developments in the field?

Following the historical perspective of Berman (2008), I have annotated this graph with arrows and numbers referring to technological or sociological events in the field on the lower part of the graph below the line. On the upper part (above the line), I have marked with thicker lines the points where there is a visually appreciable change in the slope of the line: up arrows indicate an increase in the slope (higher deposition rate) and down arrows mark what appears to be a decrease in the slope (lower deposition rate). The reason why I did not write these notes as a scientific paper is because I could not find ways to demonstrate that these changes were 'statistically significant' and thus I have to leave them as solid 'visual trends'. They appear to be correct but not proven.

Before I go into a discussion of the sequence of technological and sociological events that could have affected the deposition rate (growth rate) of the PDB, I wish to compare the overall growth rate of the PDB as indicated by the straight line in Fig. 3 with the growth rate envisioned (or rather 'feared'!) by the pioneers in the field as reflected in the letter that Richard Dickerson (of early cytochrome fame) sent to the crystallographic community in 1978 (Fig. 2, inset). In this typed letter, one can see that based on the deposition of early protein structures up to 1978 the exponential growth was given by $N = \exp(0.19 \times \text{year})$. One can easily compute that with this exponential growth the duplication time for the number of entries should be around 3.65 years [(ln 2)/0.19]. This early growth rate can be seen in the lower left part of Fig. 3 (dashed line), approximately corresponding to the period labeled *A* (1978–1980) and giving some confidence in the validity of the visual inferences drawn from this representation.

From a linear least-squares fit of all the data represented in Fig. 3, one can derive that the exponential growth corresponds to $N = \exp(0.252 \times \text{year})$. One of the interesting conclusions from this way of analyzing the data is that the growth rate (or deposition rate) of the PDB at the beginning and during the last 20 years or so is not that much different (0.252 *versus* 0.190). This translates to a duplication time for the number of entries of approximately 2.75 years *versus* 3.65 years for the mature PDB *versus* the timid beginnings, respectively. In view of the dramatic changes in technology that the field of macromolecular crystallography has experienced, I found this perplexing and thought-provoking. Given the advances in protein crystallization methodology, computational advances in hardware and software, together with computer graphics, synchrotron radiation and the larger number of researchers involved in macromolecular crystallography, it is sobering to think that the deposition rate has only increased from 0.190 to 0.252. Given the limited data of the early years, one wonders if this difference is even significant.

An important study of the impact of synchrotron sources on the deposition of structures in the PDB was published by Jiang & Sweet (2004). The work analyzes in detail the effect that access to second- and third-generation sources has had on the number of structures deposited and also on their size and possibly the quality of the data. An important observation is also the delay observed between data collection at in-house sources *versus* synchrotron

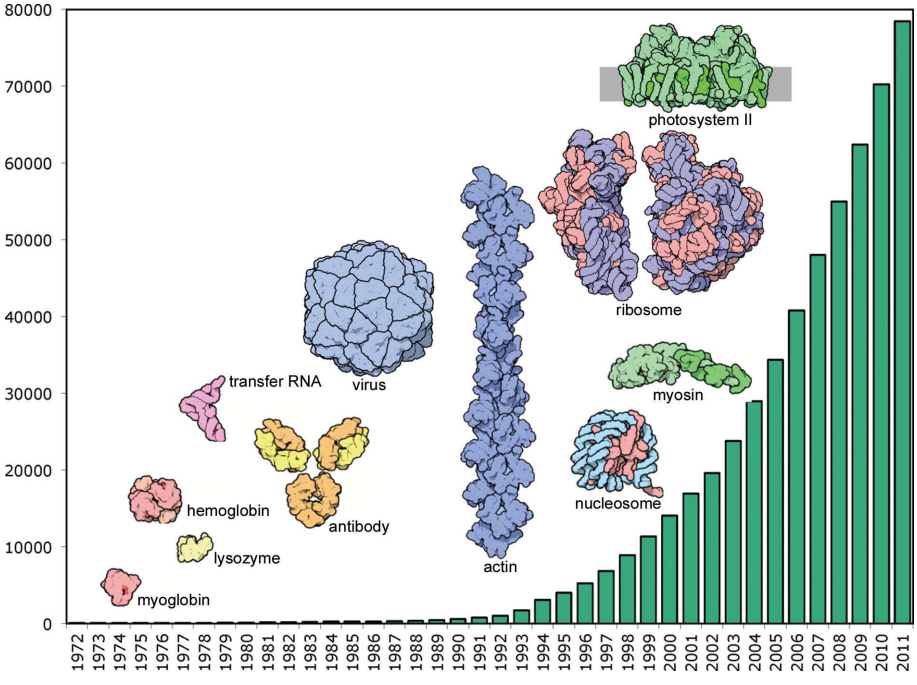

Figure 1
Iconic representation of the growth in content and complexity of the PDB. The image is adapted from that presented by Berman (2008), including the PDB content as of October 2011. The abscissa corresponds to the year and the ordinate is the total number of deposited entries for that year. Courtesy of the PDB (C. Zardecki).

beamlines (25 *versus* 18 months). In this study it is clearly demonstrated that the major impact of access to synchrotrons has been on the 'mean' (or shall we say typical) structure deposited. Based on their statistics, the size of the entries in the PDB solved using synchrotron radiation is almost double that of those solved using in-house sources (6071 atoms and 57 928 reflections compared with 3231 atoms and 27 720 reflections, respectively). There also indicators suggesting that the quality of the data, and therefore also of the structures, is better for the depositions from synchrotrons than those from in-house sources. However, the effect of synchrotron sources on the deposition rates of the PDB was not examined, except for documenting that the number of structures (as a percentage of the total) solved using synchrotron data has continued to increase over the years (Jiang & Sweet, 2004; BIOSYNC website; http://biosync.sdsc.edu).

There may be many reasons why the overall growth rate of entries in the PDB is not more dramatically different from what was suggested from the early deposition rates. Naturally nowadays many more structures are solved, refined, analyzed and studied than are published and deposited in the PDB. This is certainly the case in the industrial laboratories (pharmaceutical industries and others), where hundreds of target–ligand complexes are solved in a typical structure-based drug-design project. Nonetheless, at the academic level I would have personally expected to see a larger growth rate after so many years of technical innovation. Is this deposition limit reflecting the maximum rate at which the community can clone, purify, crystallize, solve, refine, 'understand', write and publish the biological results of interest for the biomedical community? As opposed to the early revolutionary days of protein crystallography, are we feeling a sense of *déjà vu* when

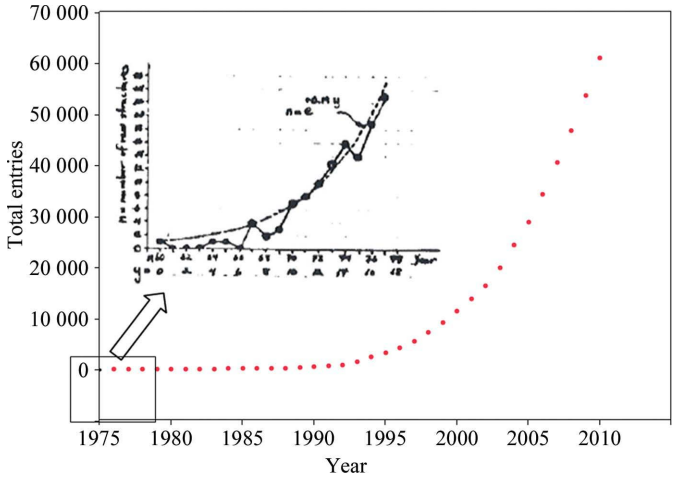

Figure 2
Detailed view of the number of PDB entries *versus* year. The graph was constructed using the data available from the PDB for the yearly number of entries and the cumulative content. The inset shows an excerpt from the letter that Richard Dickerson (California Institute of Technology) sent to the community on September 1, 1978 discussing the growth in the number of structures available. The inset plot shows the exponential fitting for the available entries as of 1978 (1960–1978). The actual text of the letter referring to the 'alarming' growth in the number of structures reads: 'The number of new structures appearing per year is rising exponentially, as shown in the plot at the right. It can be fitted well by the expression n = exp(0.19y). The last four years have averaged one structure every three weeks! If this exponential growth were to continue, by 1991 we would see one new protein structure every day and Geis and I would give up in despair. Even today, with 132 structures, it is a surprise to realise that the 1968 'Structure and Action of Proteins' had the benefit of only eight high-resolution structures'. The text of the letter was extracted from a copy of the original letter available from the PDB website. Deposition data are available from the statistics at the PDB website.

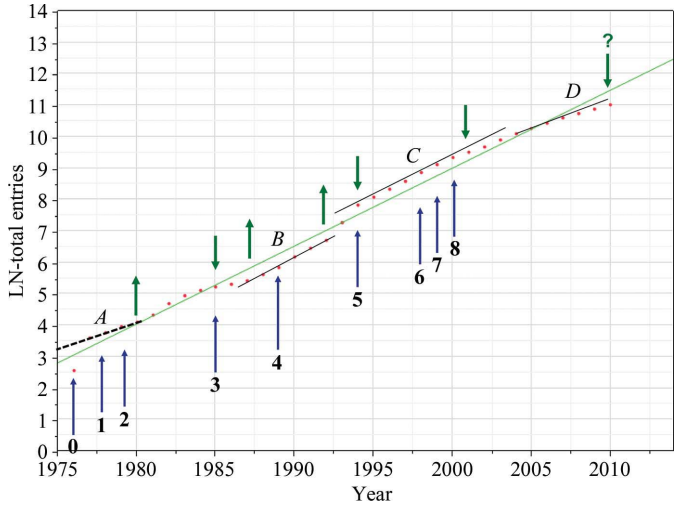

Figure 3
Overall least-squares linear fit of the growth rate of the PDB. The graph was constructed from the tabulated form of the PDB content. Instead of representing the cumulative number of entries per year, another column was created with the natural logarithm of the total number of entries [ln(N), labeled as LN-total entries; y axis] and was plotted against the year number (x axis). The numbers below the blue arrows in the lower part of the diagram (below the least-squares fit) correspond to the historical or sociological events described by Berman (2008) as follows. **0**, January 1976. Report to ACA Council. **1**, 1978. R. Dickerson's letter to the community reporting an initial deposition growth rate (exponent) of 0.19. *FRODO* and computerized molecular graphics. **2**, 1979. *PROLSQ* restraint refinement available (Hendrickson & Konnert, 1981). **3**, 1985. Structure of rhinovirus. **4**, 1989. Petition from the F. M. Richards committee; formal guidelines for data deposition published by the IUCr. These were deposition of the coordinates at publication and release no more than year after publication. Major journals adopted these guidelines and the NIGMS (as well as other funding agencies such as HHMI) set a policy saying that funding would be contingent on the open sharing of structure data. **5**, ESRF user access. **6**, APS user access. **7**, SPring-8 user access. The exact user-access dates for the three main third-generation synchrotron-radiation sources are only approximate as they are difficult to specify for the different beamlines within each facility. The dates of significant impact on the number of structures have been documented by Jiang & Sweet (2004). **8**, Ribosome structure (Ban *et al.*, 2000). The partial yearly data for 2011 have been removed so as not to bias the total number of entries in 2011. The green arrows above the least-squares line indicate (up and down) the points where there visually appears to be a change in the slope of the line increasing or decreasing (respectively) the overall deposition rate through the years. A statistical analysis was performed with the program *JMP*9. The overall least-squares line fitting corresponds to the following statistical parameters: R^2 = 0.986; root-mean-square error = 0.312; N = 35 (1976–2010); slope = 0.252. Entries for 1972–1975 (no structures) and the partial data corresponding to 2011 have been removed from the analysis and are not shown. The deposition rates of periods A (1976–1980), B (1987–1992), C (1994–2002) and D (2006–2010) appear to be visually distinct and preliminary analysis of the covariance of the slopes of these periods (ANCOVA) suggest that they should be considered as different 'components' and together provide a slightly superior model of the available data (see Fig. 4) to the simple least-squares fit for all the data (this figure).

our most intriguing enzymes turn out to have an already pre-existing fold? Are there any other limiting factors such funding or synchrotron access that limit the 'production' rates of structures as measured by the deposition rate? Does the PDB have enough resources to expedite the deposition rates? Is the number of entries, independent of the complexity (or size), the right metric to monitor the deposition rates? These are issues that the community should ponder and take action accordingly.

The soft undulations of the data points above and below the overall deposition rate of 0.252 are also intriguing, particularly as they may relate to the historical or sociological events annotated following the historical events described by Berman (2008). The change in the overall rate between the initial rate of 0.19 and the actual rate of 0.252 has already been noted and was possibly fueled by the model-refinement tools of computer graphics such as *FRODO* (Jones, 1978) and restraint refinement (*PROLSQ*; Hendrickson & Konnert, 1981). There is an appreciable decrease in the rate of deposition after 1985 until 1987. Before and during this period of time more structures were solved, but some members of the community were reluctant to deposit the hard-earned fruit of their labors in view of the relevance of the results for the pharmaceutical industry and the beginning of structure-based drug design in several public and private laboratories. Owing to this reluctance to deposit coordinates after the structures had been reported in the scientific literature, a method was reported in 1980 describing an algorithm to extract the coordinates from the stereo diagrams presented in scientific articles (Rossmann & Argos, 1980). The argument was made that withholding the

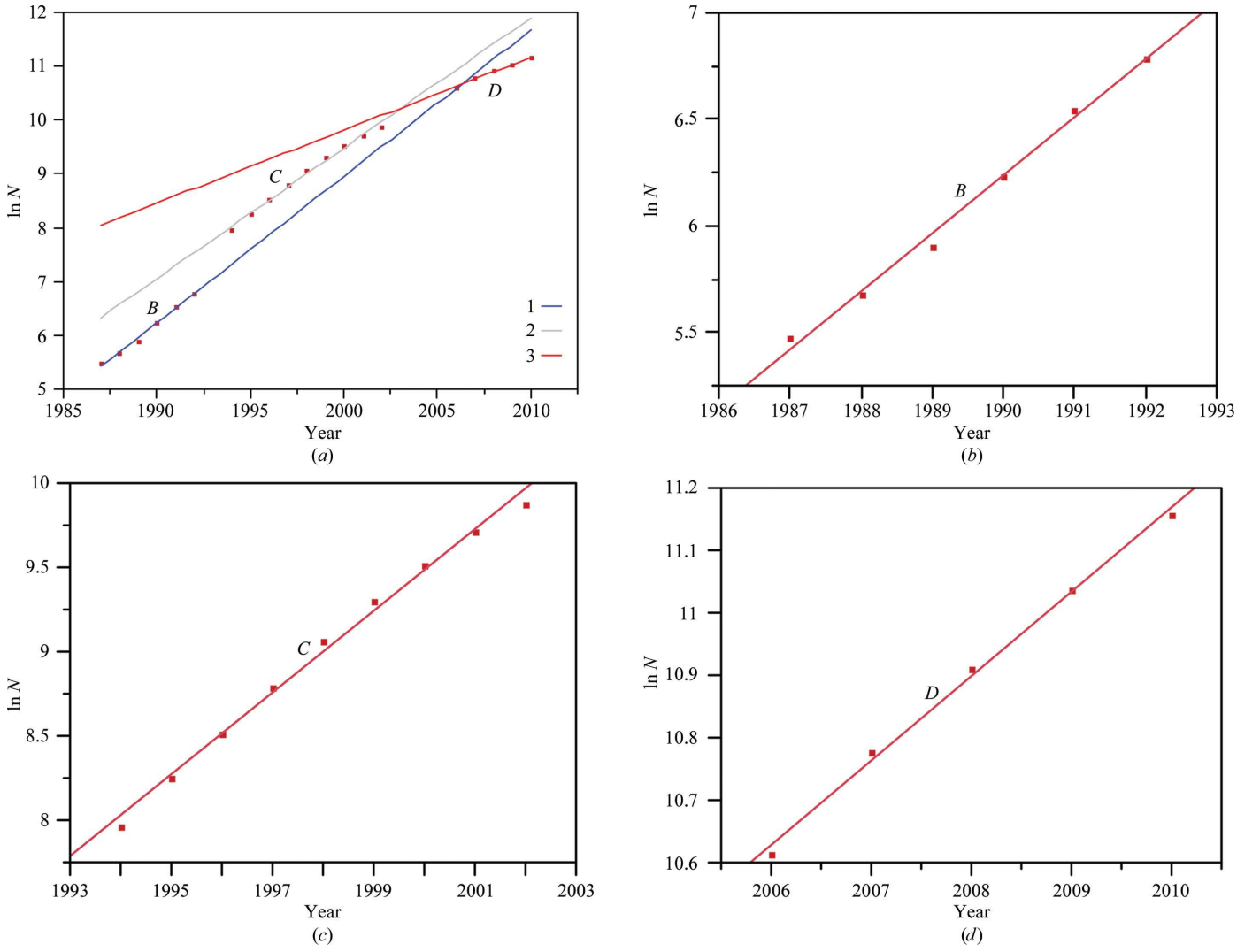

Figure 4
(*a*) Overall ANCOVA analysis of the PDB deposition rates for periods *B*, *C* and *D* as indicated in Fig. 3. The color-coded lines 1, 2 and 3 are for periods *B*, *C* and *D*, respectively. The corresponding statistical results for the entire model, including distinct slopes for the three periods, are $R^2 = 0.999$, root-mean-square error = 0.049, $N = 20$. The parameters for the model suggest that the least-squares fit is slightly superior to the overall fit presented in Fig. 3. (*b*)–(*d*) show the least-squares fits of the individual periods to obtain individual growth rates for the different periods. (*b*) Least-squares fit for period *B* (1987–1992). The least-squares fit corresponds to a growth rate of 0.272 [$\ln(N) = \exp(0.272 \times \text{year})$; $R^2 = 0.993$, $N = 6$]. (*c*) Least-squares fit for period *C* (1994–2002). The least-squares fit corresponds to a growth rate of 0.242 [$\ln(N) = \exp(0.242 \times \text{year})$; $R^2 = 0.993$, $N = 9$]. (*d*) Least-squares fit for period *D* (2006–2010). The least-squares fit corresponds to a growth rate of 0.135 [$\ln(N) = \exp(0.135 \times \text{year})$; $R^2 = 0.996$, $N = 5$].

atomic coordinates of the proteins solved was at odds with the long scientific tradition of open communication and data exchange.

Surprisingly, it seems that the actions of the committee headed by Fred M. Richards in the ensuing years, the community discussions, the letters to the editors of scientific journals and the other actions culminating in the guidelines issued in 1989 by the IUCr, all had a delayed effect and by 1992 increased the rate of deposition considerably (Fig. 3). This rate of deposition was sustained until the end of the century. A little disconcerting is the appearance of what appears to be a definitive downward trend after 2001. It is still uncertain what will happen next. The indicated decrease in the deposition rate may be temporary as has happened in the past, but it is too early to say. It is interesting to note that the most noticeable increase in the growth rate (1993–1995) was caused not by a major technological advance such as access to synchrotron radiation (the commissioning of the ESRF, APS or SPring-8) but rather by the requirements imposed by the major journals in the field and the funding agencies (National Institute of General and Medical Sciences, NIGMS and Howard Hughes Medical Institute, among others) to make results publicly available. It is quite reasonable to assume though that the access to third-generation sources (ESRF, APS and SPring-8) beginning in 1994 has had an impact on sustaining the PDB deposition rate (Fig. 3, period *C*) even though the size of the typical entries has approximately doubled, as indicated by the Jiang and Sweet study. This should make the community of structural biologists proud and ready to take up other challenges.

The visual trends indicated in Fig. 3 within the full timespan of the available data can be used to extract four periods (*A*–*D*) that have been marked by an extended linear growth and that can be used statistically to examine the different rates using ANCOVA (analysis of covariance). Period *A*, corresponding approximately to the initial rate discussed in Dickerson's letter, does not have enough data points in the PDB statistics to run a robust least-squares fit. Fig. 4(*a*) shows three different rates for periods *B* (1987–1992), *C* (1994–2002) and *D* (2006–2010) as characterized by the ANCOVA analysis. The R^2 of the combined model is only slightly superior to the overall least-squares fit presented in Fig. 3 for the entire (full year) statistics provided at the PDB website (1976–2010) (0.999 *versus* 0.986), preventing us from drawing any definitive conclusions. Statistically, the root-mean-square error of the two models is dramatically different: 0.049 (N = 20) *versus* 0.312 (N = 35), respectively.

These caveats aside, one cannot escape the observation that the deposition rate for the most recent period *D* (2006–2010) is much smaller than (almost half!) the overall growth rate (0.135 *versus* 0.252). If the trends suggested in the study by Jiang & Sweet (2004) continue, perhaps it is the fact that we (the community of structural biologists) are taking up more challenging biological problems that is causing the slowing down of deposition (or structure-solution) rates? Hopefully, upcoming technical innovations at synchrotron beamlines such as microfocusing and others will be able to restore the deposition rates while continuing to tackle more technically challenging problems. I am leaving this observation for the community to ponder as to the possible reasons and potential solutions.

In conclusion, analysis of the PDB deposition rates in some detail may prove to be a valuable exercise for the community of structural biologists at large. It could be argued that the available data are not adequate to provide statistical rigor to some of the above inferences. This may be so and this is why I decided not to write a formal scientific paper to present these observations and ideas and to maintain the format of a reflective essay. However, I do think that the visual inferences obtained from analysis of the deposition rates of the PDB in fine detail can be valuable guides to suggest trends and possible connections among events related to the growth of the PDB. This is particularly relevant on this 40th birthday of the PDB. Is the PDB going to continue to grow at the constant rate suggested by previous years, or is it beginning to face a 'middle-age crisis'? Naturally, only the future can tell. We as a community should think about the above issues and coordinate any actions necessary to maintain this invaluable resource for us and for the biomedical and scientific communities at large. The main lesson learned might be that in the future we should look in more detail at the content and growth of the PDB, not only by its visual appearances, as shown in Fig. 1, but also by a rigorous analysis of its data-deposition rates.

Suggestions, comments and insights on earlier versions of this manuscript by Professor John R. Helliwell are greatly appreciated; however, the views expressed are only my own. I am indebted to my colleague J. Wass for his valuable assistance in the statistical analysis. Constructive comments by the reviewers are also appreciated.

References

Abad-Zapatero, C. (2011). *ACA RefleXions*, **4**, 24–25.
Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. (2000). *Science*, **289**, 905–920.
Berman, H. M. (2008). *Acta Cryst.* A**64**, 88–95.
Hendrickson, W. A. & Konnert, J. H. (1981). *Biomolecular Structure, Conformation, Function and Evolution*, Vol. 1, edited by R. Srinivasan, E. Subramanian & N. Yathindra, pp. 43–57. Oxford: Pergamon Press.
Jiang, J. & Sweet, R. M. (2004). *J. Synchrotron Rad.* **11**, 319–327.
Jones, T. A. (1978). *J. Appl. Cryst.* **11**, 268–272.
Phillips, D. C. (1971). *Cold Spring Harb. Symp. Quant. Biol.* **36**, 589–592.
Rossmann, M. G. & Argos, P. (1980). *Acta Cryst.* B**36**, 819–823.