Acta Crystallographica Section D
Biological Crystallography
ISSN 1399-0047

How precise are reported protein coordinate data?

Arun S. Konagurthu,[a] Lloyd Allison,[a] David Abramson,[a,b] Peter J. Stuckey[c] and Arthur M. Lesk[d]*

[a]Clayton School of Computer Science and Information Technology, Monash University, Clayton, VIC 3800, Australia, [b]Research Computing Center, University of Queensland, St Lucia, QLD 4072, Australia, [c]Department of Computing and Information Systems, University of Melbourne, Parkville, VIC 3010, Australia, and [d]The Huck Institute of Genomics, Proteomics and Bioinformatics and the Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA 16802, USA

Correspondence e-mail: aml25@psu.edu

Received 13 October 2013
Accepted 21 November 2013

Atomic coordinates in the Worldwide Protein Data Bank (wwPDB) are generally reported to greater precision than the experimental structure determinations have actually achieved. By using information theory and data compression to study the compressibility of protein atomic coordinates, it is possible to quantify the amount of randomness in the coordinate data and thereby to determine the realistic precision of the reported coordinates. On average, the value of each C^{α} coordinate in a set of selected protein structures solved at a variety of resolutions is good to about 0.1 Å.

1. Introduction

Worldwide Protein Data Bank (wwPDB) entries for protein structures report coordinate data in ångström units (Å) to three places after the decimal point. There is consensus that for most, if not all, protein structures, the stated precision is significantly greater than can be inferred from X-ray crystallography or any other structure-determination method. However, estimating *quantitatively* the precision of protein coordinate data continues to be a thorny problem (Luzzati, 1952; Read, 1990; Murshudov & Dodson, 1997; Cruickshank, 1999; Ten Eyck, 2003).

For protein crystal structures, the distribution of the *B* factors gives a qualitative index of the precision of the atomic coordinates. Classic papers by Luzzati (1952) and Read (1990) produce statistical distributions of coordinate errors based on *R* factors. The work of Ten Eyck (2003) comes perhaps closest to the information that one wants, in allowing assignment of errors to individual atoms; in particular, in identifying *where* a model is inconsistent with the data. For NMR, Nabuurs *et al.* (2003) have studied the root-mean-square deviation of structures consistent with satisfying the experimental distance constraints.

All of these methods depend on analysis of structures together with the experimental data on which they are based.

In this paper, we present a different approach, in which we attempt to derive the precision of the C^{α} atoms of a protein data set *without* reference to experimental data. The method is therefore applicable, without change, to structures determined experimentally by X-ray crystallography, NMR or electron microscopy, and structure prediction as well.

Stating protein coordinates to a higher precision than is derivable from experiment (or theory) introduces randomness into their reported digits. Information theory and data compression provide a rigorous way to quantify the randomness in any data, and can thereby evaluate quantitatively the true precision of the structure determinations.

Intuitively, compressibility and predictability go hand in hand. That is, the more predictable any data are, the more compressible they become (Shannon, 1948; Solomonoff, 1960; Kolmogorov, 1965). Conversely, the more random any data are, the less compressible. Therefore, an approach to quantify the extent of randomness in PDB entries is the investigation of the compressibility of the coordinate values.

In information theory, the framework of minimum message length encoding (Wallace & Boulton, 1968; Wallace, 2005) gives a quantitative estimate of the compressibility of data. This framework relies on explaining any observed set of data as a *two-part message*. The first part describes a 'theory' (or 'signal') implicit in the observed data. The second part provides the 'details' (or 'noise') of the data not explained by the theory. This two-part message is encoded in the shortest possible way from which the original data can be recovered exactly. It can be observed from this framework that the more random the data are, the longer the encoded message becomes, dominated by the explanation of noise in the second part of the message. Conversely, the more predictable the data are, the more compressible is the encoded message, as the second part becomes extremely concise.

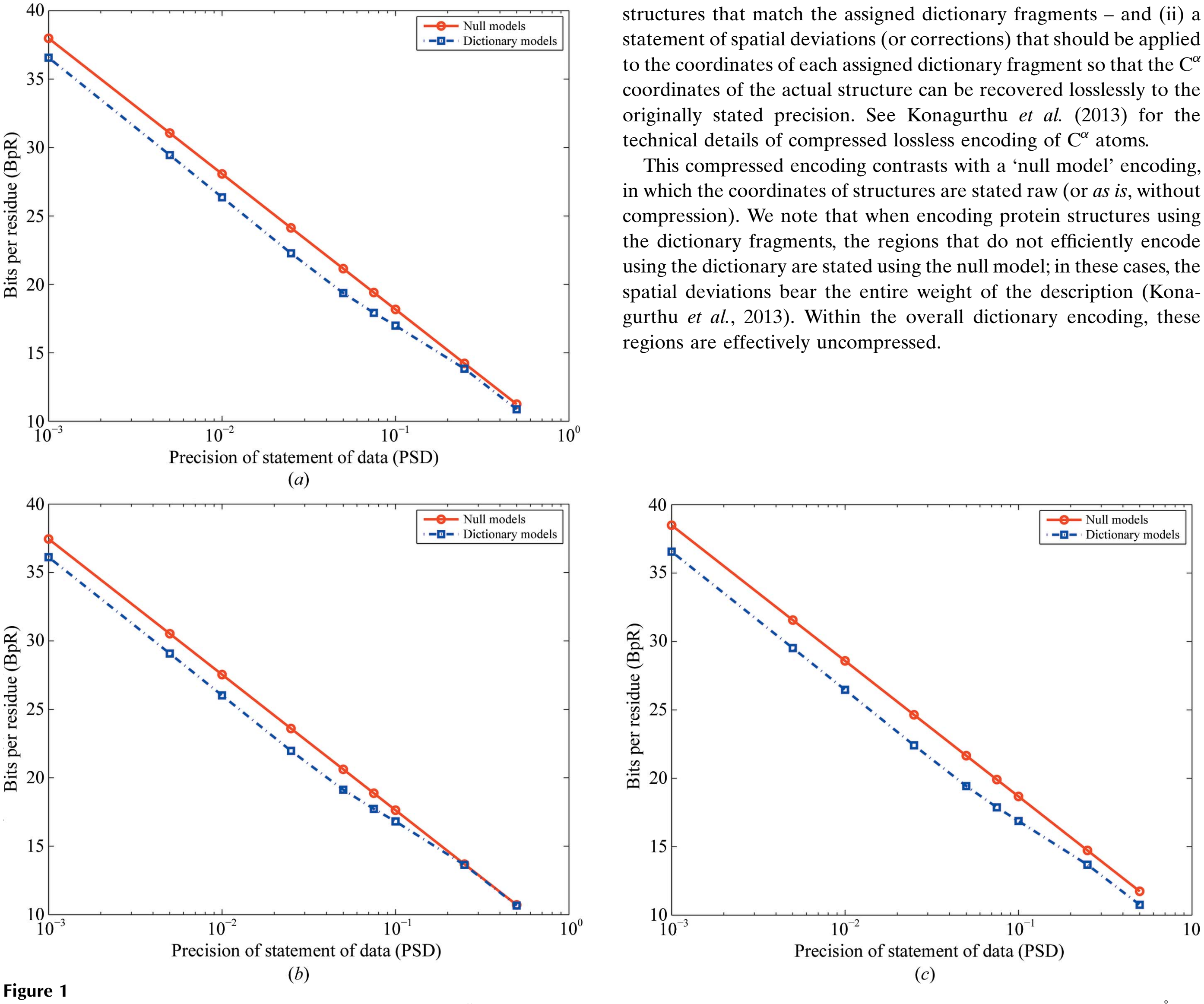
2. Compression of protein C$^\alpha$ coordinate data

Our measure of the compressibility of the data reported here depends on our recently introduced information-based technique to infer a dictionary of recurrent protein fragments for an entire collection of protein structures representative of the nonredundant wwPDB (Konagurthu *et al.*, 2013). This approach relies on the Bayesian method of minimum message length inference (Wallace & Boulton, 1968; Wallace, 2005). In our application of this method (Konagurthu *et al.*, 2013), the optimal fragment dictionary is defined as that which permits the most concise explanation (technically, the shortest lossless encoding) of the coordinates of the source structures in the collection.

This dictionary allows the efficient, lossless representation (or encoding) of the positions of the C$^\alpha$ atoms in any given protein coordinate set. By lossless, we mean that the encoded coordinates should be decodable to the same precision at which they were encoded: the encoded digits should be reproduced literally; that is, exactly, not approximately. In the work of Konagurthu *et al.* (2013), we used the full precision, 0.001 Å, reported in the wwPDB entries.

How do we use the dictionary to encode the C$^\alpha$ coordinates of a set of protein structures? The optimal lossless encoding of any particular protein structure comprises (i) a dissection (or segmentation) – that is, a designation of successive non-overlapping regions in the protein structures that match the assigned dictionary fragments – and (ii) a statement of spatial deviations (or corrections) that should be applied to the coordinates of each assigned dictionary fragment so that the C$^\alpha$ coordinates of the actual structure can be recovered losslessly to the originally stated precision. See Konagurthu *et al.* (2013) for the technical details of compressed lossless encoding of C$^\alpha$ atoms.

This compressed encoding contrasts with a 'null model' encoding, in which the coordinates of structures are stated raw (or *as is*, without compression). We note that when encoding protein structures using the dictionary fragments, the regions that do not efficiently encode using the dictionary are stated using the null model; in these cases, the spatial deviations bear the entire weight of the description (Konagurthu *et al.*, 2013). Within the overall dictionary encoding, these regions are effectively uncompressed.



(*a*)

(*b*)

(*c*)

Figure 1

Comparison of the average number of bits required to state each C$^\alpha$ coordinate using null and dictionary models, with varying values of the PSD between 0.001 and 0.5 Å. (*a*) Plot corresponding to all 8992 source structures from the Protein Data Bank. (*b*) Plot corresponding to a subset of high-resolution (better than 1.7 Å resolution) structures in the collection. (*c*) Plot corresponding to a subset of low-resolution (worse than 2.8 Å resolution) structures in the collection.

Acta Cryst. (2014). D**70**, 904–906 Konagurthu *et al.* • Precision of protein coordinate data **905**

It is possible to encode the coordinates at different statements of precision. For, given any data reported to some precision we can arbitrarily restate them to lower precision. For example, truncating or rounding 654.123 to 654.12 decreases the precision from 0.001 to 0.01. By examining the compressibility of the coordinate data as a function of the expressed precision, we can determine at what precision the data lose compressibility and, therefore, non-randomness. This reveals the true information content of the data, as opposed to the putative significance implied by the stated precision.

To implement these ideas, we define the 'precision of statement of data' (PSD) to mean the precision to which each individual x, y or z coordinate of protein structures is reported. Three places after the decimal point corresponds to a PSD of 0.001 Å. Although it is convenient to think of the PSD as discrete in terms of numbers of decimal digits, the PSD is a continuous variable[1]. A smaller value of the PSD implies a more precise statement of coordinates compared with a larger value of the PSD.

Our information-theoretic measure of compressibility thereby provides a way to measure quantitatively the precision of coordinate data. Suppose, for the sake of argument, that the first two figures after the decimal point reflect experimental precision, but the third figure after the decimal point is effectively random. The third digit is then by definition incompressible. The dictionary model message is the compressed statement of the coordinate data; the null model message encoding the same data corresponds to an uncompressed statement of the same information. Dictionary encoding the coordinate data (that is, compressing them) will *not* produce a more concise message than the null model, at least with respect to the random third digit after the decimal.

To apply these ideas, we compared the lengths of the null and dictionary model messages for the C^{α} coordinates of a collection of protein structures as a function of the PSD. A lower message length implies greater compression. The dictionary and null model message lengths are measured in *bits*. We measure bits per residue (BpR) for each of the two models, measuring the average number of bits required to encode the C^{α} coordinates of all residues in the collection of protein structures under that model. The difference between the message lengths of null-model encoding and dictionary encoding (ΔML), as function of the PSD, reflects the nonrandomness of successive digits of the data.

3. Results and conclusions

We considered a collection of 8992 experimentally determined structures from the wwPDB which were dissimilar in amino-acid sequence to avoid experimental and selection bias. Fig. 1(*a*) shows both compressed (dictionary-based) and uncompressed (null model-based) message lengths for varying values of the PSD. Notice that while the BpR for both models diminishes when the PSD is increased in value from 0.001 Å upwards, their average difference in message lengths, ΔML, remains roughly constant for PSD values in the range 0.001–0.075 Å. Upon further increasing the PSD from 0.075 to 0.1 Å, the difference starts to decrease in comparison.

We therefore conclude that the last place value after the decimal point in the C^{α} coordinate data in our source set is random because the dictionary model does not find any compressible information in these digits.

Our method reports the *average* precision of the C^{α} coordinates in the chosen set of proteins. Clearly, there must be a large subset of C^{α} atoms for which the precision is better than the average and, correspondingly, another large subset for which the precision is worse. In many cases, the region around the active site is the best determined portion of a protein, and this is also the region of greatest interest in interpreting function. Therefore, our results should not be taken to engender undue pessimism for structural studies.

Notice that varying the PSD from 0.001 to 0.01 Å changes the BpR for the null model from 37.97 to 28.07 bits, a drop of roughly 10 bits. For the dictionary model, the BpR changes by approximately the same amount, from 36.56 to 26.36 bits over the same PSD values, again a drop of about 10 bits, leaving ΔML roughly constant. Why 10 bits? It takes $\log_2(10) = 3.32$ bits to encode, optimally, random integers in the range 0–9. Therefore, to encode three random integers in the range 0–9 (for the x, y and z coordinates) takes $3 \times \log_2(10) \simeq 10$ bits.

The constancy of ΔML at a PSD of up to (arguably) 0.1 Å suggests that, on average, the experimental measurement precision for each x, y and z component of the C^{α} coordinates is no better than 0.1 Å. ΔML diminishes for values of the PSD beyond 0.1 Å. This suggests that beyond 0.1 Å the dictionary model loses compression because valid compressible information is being discarded at larger values of the PSD, and hence the dictionary model converges to the null model message length in the absence of compressible information.

We studied separately the dependence of the PSD on message lengths for high-resolution and low-resolution data sets (as defined in the caption to Fig. 1). As expected, the precision of the coordinates varies with the resolution of the X-ray structure determinations. Figs. 1(*b*) and 1(*c*) show that the coordinates of the high-resolution data set retain compressibility to a PSD of 0.075 Å, but that for the low-resolution structures the data lose compressibility at a PSD of >0.25 Å. Interestingly, the gap between null and dictionary models for the high-resolution data set is narrower than that for the low-resolution data set. This is the result of a ~1 bit per residue saving for the null model with respect to the dictionary model arising from the fact that the distance between successive C^{α} coordinates is more tightly centered around the mean of 3.8 Å for high-resolution structures than for low-resolution structures, a fact which the null model exploits. That high-resolution structures are usually more precisely determined than low-resolution structures comes as no surprise, but our method permits a quantitative description of the difference.

References

Cruickshank, D. W. J. (1999). *Acta Cryst.* D**55**, 583–601.

Kolmogorov, A. (1965). *Probl. Inf. Transm.* **1**, 1–7.

Konagurthu, A. S., Lesk, A. M., Abramson, D., Stuckey, P. J. & Allison, L. (2013). *ICDM13: 13th IEEE International Conference on Data Mining*. http://arxiv.org/abs/1310.1462.

Luzzati, V. (1952). *Acta Cryst.* **5**, 802–810.

Murshudov, G. & Dodson, E. (1997). *CCP4 Newsl. Protein Crystallogr.* **33**, 31–39.

Nabuurs, S. B., Spronk, C. A., Krieger, E., Maassen, H., Vriend, G. & Vuister, G. W. (2003). *J. Am. Chem. Soc.* **125**, 12026–12034.

Read, R. J. (1990). *Acta Cryst.* A**46**, 900–912.

Shannon, C. E. (1948). *Bell Syst. Tech. J.* **27**, 379–423.

Solomonoff, R. (1960). *A Preliminary Report on a General Theory of Inductive Inference*. Report V-131. Cambridge, MA: Zator Co. http://world.std.com/~rjs/rayfeb60.pdf.

Ten Eyck, L. (2003). *Methods Enzymol.* **374**, 345–369.

Wallace, C. (2005). *Statistical and Inductive Inference by Minimum Message Length*. Berlin: Springer.

Wallace, C. & Boulton, D. (1968). *Comput. J.* **11**, 185–194.

[1] A traditional 'rule of thumb': number of significant figures $\simeq \log_{10}$(relative error) = pRE (in analogy with pH).