Computational Biology and Chemistry 46 (2013) 8–15

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research article

Estimating structure quality trends in the Protein Data Bank by equivalent resolution

Anurag Bagaria, Victor Jaravine, Peter Güntert*

Institute of Biophysical Chemistry, Center for Biomolecular Magnetic Resonance, and Frankfurt Institute of Advanced Studies, Goethe University Frankfurt am Main, 60438 Frankfurt am Main, Germany

ARTICLE INFO

Article history:
Received 27 February 2013
Accepted 29 April 2013

Keywords:
Protein structure validation
Multiple linear regression
X-ray and NMR
PDB
Structure quality
Equivalent resolution

ABSTRACT

The quality of protein structures obtained by different experimental and ab-initio calculation methods varies considerably. The methods have been evolving over time by improving both experimental designs and computational techniques, and since the primary aim of these developments is the procurement of reliable and high-quality data, better techniques resulted on average in an evolution toward higher quality structures in the Protein Data Bank (PDB). Each method leaves a specific quantitative and qualitative "trace" in the PDB entry. Certain information relevant to one method (e.g. dynamics for NMR) may be lacking for another method. Furthermore, some standard measures of quality for one method cannot be calculated for other experimental methods, e.g. crystal resolution or NMR bundle RMSD. Consequently, structures are classified in the PDB by the method used. Here we introduce a method to estimate a measure of equivalent X-ray resolution (e-resolution), expressed in units of Å, to assess the quality of any type of monomeric, single-chain protein structure, irrespective of the experimental structure determination method. We showed and compared the trends in the quality of structures in the Protein Data Bank over the last two decades for five different experimental techniques, excluding theoretical structure predictions. We observed that as new methods are introduced, they undergo a rapid method development evolution: within several years the e-resolution score becomes similar for structures obtained from the five methods and they improve from initially poor performance to acceptable quality, comparable with previously established methods, the performance of which is essentially stable.

1. Introduction

The accuracy and quality of a three-dimensional protein structure are important factors deciding its utility. Knowledge of the three-dimensional structure is important for studying a protein's biological role, molecular mechanism, and molecular interactions. The closer an experimentally determined or theoretically calculated structure is to its native structure, the more useful it is for research. For example, it would be nearly meaningless to use a target protein structure for structure-based drug design if we are unsure about the quality of the target protein model. The importance of protein structures and their association with biological regulation has been known for over half a century (Tomkins et al., 1963). This interest was the driving force behind developing and improving methods of structure calculation (Kuntz et al., 1976; Floudas et al., 2006; Güntert P, 2009; Kelley and Sternberg, 2009). With structures coming from many different experimental and theoretical methods, several methods to assess protein structure quality have been developed, e.g. (Floudas et al., 2006; Sanchez and Sali, 1997; Bhattacharya et al., 2007; Rosato et al., 2012; Moult et al., 2009; Janin et al., 2003; Laskowski et al., 1993, 1996; Chen et al., 2010; Davis et al., 2004, 2007; Sippl, 1993; Vriend G, 1990; Cristobal et al., 2001; Siew et al., 2000; Eisenberg et al., 1997; Lovell et al., 2003; Huang et al., 2005; Bagaria et al., 2012; Berjanskii et al., 2012). Validation methods using the experimental data generally utilize their own set of specially designed scores. In most cases, these scores cannot be computed for structures obtained by other structure determination methods because the required experimental data are not available. Here we attempt to solve this problem by extending the notion of X-ray resolution toward a more generalized definition based on the linear combination of different coordinate-based validation scores. Restricting the input scores to only those coming from molecular coordinates implies that the e-resolution can be computed for any given protein structure.

About 81,700 protein structures have been deposited in the Protein Data Bank as of February 25, 2013. Most of these structures were determined by the two most popular experimental

* Corresponding author at: Institute of Biophysical Chemistry, Max-von-Laue Str. 9, 60438 Frankfurt am Main, Germany. Tel.: +49 69 79829621; fax: +49 69 79829632.
E-mail addresses: anurag.bagaria@bpc.uni-frankfurt.de (A. Bagaria), zharavin@em.uni-frankfurt.de (V. Jaravine), guentert@em.uni-frankfurt.de (P. Güntert).

1476-9271/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiolchem.2013.04.004

techniques: X-ray diffraction (88.8% of the structures) and solution NMR (10.5%). The deposition of structures calculated via "new methods" started only recently: electron crystallography (1991), fiber diffraction (1994), solid state NMR (1997), electron microscopy (1997), and solution scattering (1999). Experimental techniques have evolved over time for all these methods (DiMaio et al., 2011; Joosten et al., 2009; Chen et al., 2012). Sophistication of instruments has allowed performing more complicated experiments with improved efficiency and effectiveness. Comparative modeling of proteins including theoretical modeling have grown and improved rapidly (Sanchez and Sali, 1997; Pantazes et al., 2011). However, even when a structure is determined using the most accurate and established experimental method, we still have the obvious question of whether the structure is correct overall and in all its parts.

For X-ray structures, the most accepted criterion to assess the amount of experimental data is the crystal resolution. However, many other measures like R-factor, B-factors, stereo-chemical parameters etc. are also used for more detailed analyses of the structure. The resolution is formally defined as the smallest distance between structural features that still provide measurable X-ray diffraction and, as a result, can be distinguished from each other in electron density maps (Wlodawer et al., 2008). High-resolution structures have a resolution below 1.8 Å and the ones above 2.7 Å are considered to be of low resolution. The intermediary ones are classified as medium resolution (Minor, 2007). In case of structures obtained via NMR spectroscopy, structural quality is described by the bundle RMSD, the amount of experimental restraints, the number of distance and angle violations, RPF scores (Huang et al., 2012), etc. (Bhattacharya et al., 2007; Doreleijers et al., 2012).

Structure quality assessment criteria using experimental data are different for the major experimental methods. This makes it difficult to compare the quality of structures obtained by the two methods. Even comparing local and global features in a non-redundant dataset containing NMR and X-ray structure pairs of the same proteins revealed systematic (including method-related) differences (Sikic et al., 2010). Similar systematic differences were pointed out (Bagaria et al., 2012) while comparing the structure quality of proteins in the CASP8 (Moult et al., 2009) and CASD-NMR (Rosato et al., 2012, 2009) projects.

There exists a large range of software tools with their respective scores designed to evaluate different quality aspects of protein structures. These are based on the various elements and properties of molecular structure: torsion angles, bond lengths, atom clashes, van der Waals violations, stereo-chemical violations etc. Most of these tools do not provide an obvious scale to easily comprehend the overall goodness of a structure in question. For instance, there have been a number of approaches to convert various measures of NMR protein structure bundles into a single resolution score, using Ramachandran plot quality, ensemble precision, or numbers of NOEs per residue (Kwan et al., 2011). An attempt has been made to develop an intuitive score for the quality of a protein structure in terms of predicting its RMSD from the native structure (Bagaria et al., 2012). Several attempts were also made to develop "equivalent" X-ray resolution for structures based only on coordinate information (Laskowski et al., 1996; Chen et al., 2010), including a "resolution-by-proxy" measure (Berjanskii et al., 2012) incorporating 25 protein structure features.

Here, we propose another definition of "equivalent" resolution that is generally applicable to any protein structure regardless of the method that was used to determine it. We estimate and report the quality of structures obtained over the last two decades by five popular methods: X-ray, solution state NMR, neutron diffraction, solid state NMR, and hybrid methods.

2. Materials and methods

2.1. Molecular coordinate data sets

For this study we used all PDB entries that contained a single chain of a protein determined by the authors to be monomeric, irrespective of the experimental method they were obtained with. We grouped them by their experimental technique, and divided into sub-groups by year of submission to the PDB. This constituted 22,016 X-ray, 3777 NMR, 18 neutron diffraction, 12 solid-state NMR and 7 hybrid method protein structures. Non-monomeric biological units of proteins, complexes, and some structures for which certain validation scores could not be computed for technical reasons were excluded. Regarding structures solved by electron microscopy, it must be noted that though over 340 structures have been solved by this technique, we had to exclude this method because over 95% of these structures are either not single-chain or not monomeric. Although the restriction to monomeric, single chain proteins excluded many PDB entries, such a large scale study of protein structure quality trends across different experimental fields has not been performed so far. Structures from electron crystallography, solution scattering, and fiber diffraction were omitted because fewer than 4 single chained monomeric structures by these techniques were deposited in the last 5 years. More details regarding data composition may be found in the Discussion section. The time range selected for this study is from the year 1995 to 2012.

Based on earlier reports about the dependence of structure quality on protein size (Bagaria et al., 2012), this fact was presumed and the protein structures were divided into 3 size groups: "S" or "Small" (<100 amino acid residues), "M" or "Medium" (100–400 amino acid residues), and "L" or "Large" (>400 amino acid residues). This choice of segregating structures into size-dependent bins resulted in a clear-cut improvement of their resolution predictions manifested by a reduced mean absolute error (MAE) (see Section 3). For each protein structure, 17 score values from the software tools listed below were obtained. In the case of NMR structure bundles, the scores were calculated separately for each of the top 10 conformers sorted by increasing root-mean-squared-deviation (RMSD) to the mean of the structure bundle.

2.2. Validation scores

The following coordinate based validation scores were used to assess the quality of the protein structures. These scores were selected based on their popularity as indicated by the number of publication citations, and the possibility to implement and evaluate them for structures obtained by any method.

The ProsaII scores (Sippl, 1993) are based on the probability for two residues to be at a specific distance from each other. In this validation score the amino acid types, the distance, as well as the sequence separations are used. We used three of the ProsaII scores. Zp-Pair (Z score for pair potential energy), Ep-Pair (pair potential energy based on atom–atom interaction) and Ep-Surf (Surface energy based on atom-solvent interaction).

ProQ is a neural network based predictor (Wallner and Elofsson A, 2003). Based on a number of structural features, it predicts the quality of a protein model. ProQ is optimized to find correct models in contrast to methods which are optimized to find native structures. Two quality measures are predicted, the LGscore (Cristobal et al., 2001) and MaxSub (Siew et al., 2000).

The Procheck software (Laskowski et al., 1993, 1996) takes into account the number of residues in allowed/disallowed areas of Ramachandran plot, the number of unusual bond lengths or bond angles, and so forth. Here we choose two scores from the Procheck software: Core and Gener. The former represents the percentage of