Briefings in Bioinformatics, 2017, 1–4

doi: 10.1093/bib/bbx084
Opinion Note

On the dynamical incompleteness of the Protein Data Bank

Cristina Marino-Buslje*, Alexander Miguel Monzon*, Diego Javier Zea, María Silvina Fornasari and Gustavo Parisi

Corresponding author: Gustavo Parisi, Roque Sáez Peña 352, B1876BXD Bernal, Buenos Aires, Argentina. Tel.: +54 (11) 4365-7100 ext. 5659, E-mail: gusparisi@gmail.com
*These authors contributed equally to this work.

Abstract

Major scientific challenges that are beyond the capability of individuals need to be addressed by multi-disciplinary and multi-institutional consortia. Examples of these endeavours include the Human Genome Project, and more recently, the Structural Genomics (SG) initiative.

The SG initiative pursues the expansion of structural coverage to include at least one structural representative for each protein family to derive the remaining structures using homology modelling. However, biological function is inherently connected with protein dynamics that can be studied by knowing different structures of the same protein. This ensemble of structures provides snapshots of protein conformational diversity under native conditions. Thus, sequence redundancy in the Protein Data Bank (PDB) (i.e. crystallization of the same protein under different conditions) is therefore an essential input contributing to experimentally based studies of protein dynamics and providing insights into protein function.

In this work, we show that sequence redundancy, a key concept for exploring protein dynamics, is highly biased and fundamentally incomplete in the PDB. Additionally, our results show that dynamical behaviour of proteins cannot be inferred using homologous proteins. Minor to moderate changes in sequence can produce great differences in dynamical behaviour.

Nonetheless, the structural and dynamical incompleteness of the PDB is apparently unrelated concepts in SG. While the first could be reversed by promoting the extension of the structural coverage, we would like to emphasize that further focused efforts will be needed to amend the incompleteness of the PDB in terms of dynamical information content, essential to fully understand protein function.

Key words: protein dynamic; protein function; Protein Data Bank; protein structure

Introduction

Major scientific challenges that are beyond the capability of individuals need to be addressed by multi-disciplinary and multi-institutional consortia. Examples of these endeavours include the Human Genome Project [1], and more recently, the Structural Genomics (SG) initiative [2]. Among the scientific goals of the SG initiative are the increase in the structural coverage of the sequence space to expand the possibility of protein structure prediction using template-based modelling, the identification of new folds to reach fold space convergence and a structural study of protein given superfamilies (i.e. poorly characterized, high-priority proteins). Although SG centres account

Cristina Marino-Buslje is the Director of the Bioinformatics Unit at Leloir Institute Foundation. Her research interest is protein function, structure, classification and evolution. Likewise, protein–protein interaction and the study of protein mutations in cancer.
Alexander Miguel Monzon is PhD student in the Department of Science and Technology at Universidad Nacional de Quilmes. His research is focused on the study of conformational diversity of proteins in their native state and development of the CoDNaS database.
Diego Javier Zea is a postdoctoral researcher at the Bioinformatics Unit at Leloir Institute Foundation. His research interests are protein structure and evolution.
María Silvina Fornasari is a researcher at Universidad Nacional de Quilmes. Her research interest involves protein molecular evolution and disease-related mutation analysis.
Gustavo Parisi is the head of the Bioinformatics Unit at Universidad Nacional de Quilmes. His research interest is protein function, structure and evolution.
Submitted: 4 May 2017; **Received (in revised form):** 14 June 2017

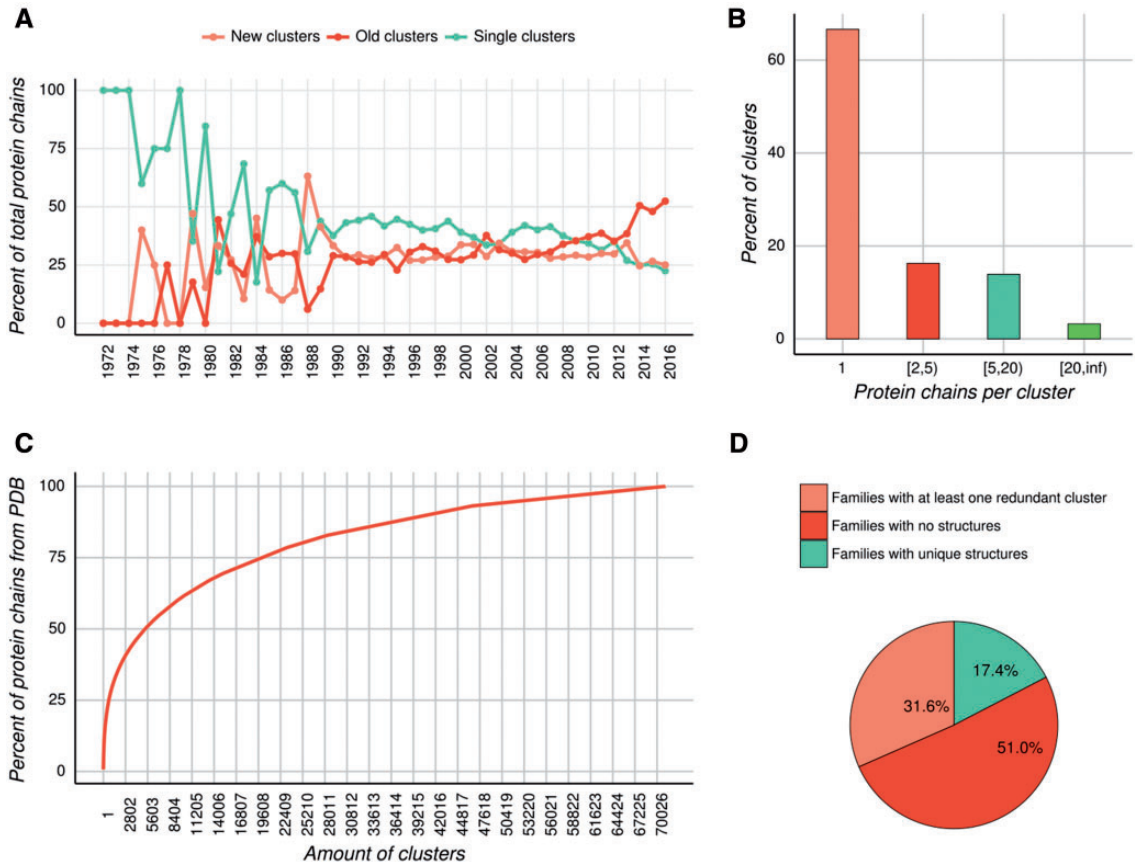

Figure 1. Protein sequence redundancy in the PDB. The redundancy was analysed for all protein chains in the PDB deposited between 8 November 1972 (first structure deposited) and 31 December 2016. A redundant cluster is composed of at least two identical protein chains from different PDB entries (**A**) The percentage of protein chains released by year in the PDB. The new cluster line represents the structures deposited in a particular year, but identical to a previous unique chain already deposited (a new cluster is formed with this addition). Old clusters correspond to the sequences deposited in a particular year that are identical to others already in a cluster (with at least two structures), and single clusters are formed by a new sequence structure (unique in PDB). (**B**) Percent of clusters with unique, 2–5, 5–20 or > 20 structures. (**C**) Cumulative percent of protein chains from total in the PDB per number of clusters. (**D**) Percent of Pfam families (version 31.0) with no associated structures (structurally unsolved), families with unique structures and families with at least one redundant cluster.

for only ∼10% of the total structures determined per year, this fraction (2000–2006) was 3.8 times less redundant than structures determined by traditional structural biology teams (TSBT), and contributed to ∼50% of novel folds [3], showing the efficiency of the discovery-driven efforts to uncover structural biology issues. Despite those efforts, the structural coverage of the sequence universe is still low [4], illustrating the need for further activities from the SG initiatives.

The goal of the SG and TSBT is the understanding of protein function in terms of structure–function relationships. However, as biological function is inherently connected with protein dynamics, we would like to use this communication to emphasize the incompleteness of the Protein Data Bank (PDB) in terms of providing the dynamical information required to fully understand protein function. Our results suggest the need of new focused efforts to complement those already implemented in the SG initiatives as a means of addressing the structural incompleteness.

Sequence redundancy helps in dynamics understanding

The study of protein dynamics using crystallographic data relies on the knowledge of different structures of the same protein. This ensemble of structures provides snapshots of protein conformational diversity under native conditions [5]. These alternative conformations arise from local flexibility (e.g. rotations and small displacement of residues), to global flexibility (e.g. large inter-domain movements or loop rearrangements), with a continuum of states between these extremes. In that way, proteins crystallized in different states provide experimentally determined information about protein flexibility at the atomic level [6]. Pioneering work in this respect came from Gerstein [7], describing the inter-domain mechanisms of hinge and shear movements. It is for these reasons that sequence redundancy in the PDB (crystallization of the same protein under different conditions), is an essential input to study protein dynamics and to gain insights into protein function.

Is the PDB redundant?

Redundancy can be described as the number of sequence-based clusters at 100% identity, where a redundant cluster is a group of at least two protein chains coming from different PDB entries (Supplementary Methods and Supplementary Figure S1). Figure 1A shows the fraction of total chains in the PDB generating new clusters per year. This is a measure of how a new PDB entry increases redundancy by reducing the count of single-member clusters. This fraction has been more or less constant

(average ~29%) over the past 10 years but shows a decrease of ~5% over the past 5 years. The fraction of total chains included in old clusters (a measure of how a new PDB entry increases already redundant datasets) has been steadily increasing. This result is not surprising, as it is well known that the PDB is highly redundant for a few sequences, as TSBT studies are primarily focused on the functional analysis of well-studied systems. Indeed, TSBT have contributed to increase the redundancy of the PDB (old and new clusters) with the 68% of structures, while SG initiatives just with ~29% (Supplementary Figure S2).

We obtained 71 057 clusters, which is the number of non-redundant sequences in the PDB. The number of protein chains per cluster is shown in Figure 1B. Among these clusters, 66% contain only one chain, while the rest (34%) has at least two chains coming from different PDB entries. As we mentioned, two structures are the minimum required to study protein movements from crystallization experiments; however, Best and co-workers [5] showed that five different structures are required per protein to reach convergence in the observed flexibility at the backbone level, and >20 structures per protein are needed for the study of residue structural heterogeneity. Considering these findings, 15% of proteins in the PDB have more than five different structures, and only 3.2% have >20. Figure 1C shows that the cumulated percent of structures in the PDB follows a steep ascent. Therefore, 50% of the chains lay in 5600 clusters representing only the 7% of the total number of non-redundant sequences in the PDB, revealing that the redundancy is dominated by only a few proteins. The top five repetitively crystallized proteins are lysozyme from chicken, human β-2-microglobulin and carbonic anhydrase-2, endothiapepsin from *Cryphonectria parasitica* and cationic trypsin from *Bos taurus* (all well >500 structures). Also, we found that this redundancy is unequally distributed in the sequence space. Figure 1D shows that only 31% of PFAM families have at least one protein with one redundant cluster, 17% have only one structure and around 50% of the Pfam families have no reported structure.

Redundancy in the PDB could also be narrowed considering additional information besides having 100% sequence identity among crystallized protein chains. Structure determination techniques [such as X-Ray crystallography, nuclear magnetic resonance (NMR) spectroscopy and cryo-electron microscopy (EM)] could provide different dynamical information as well as crystals obtained in different space groups or resolution. The consideration of these additional information over sequence redundancy would certainly increase even more the incompleteness of available data to infer dynamical information from the PDB. We found that the percentage of redundant clusters containing unique determination techniques is 81.4, 4.33 and 5.26% for X-Ray crystallography, NMR spectroscopy and cryo-EM, respectively. In addition, about ~9% of the redundant clusters contain structures obtained with different techniques (Supplementary Figure S3). Similarly, in 47% of the clusters obtained with X-Ray diffraction, their structures have been solved in the same space group (Supplementary Figure S4) and ~69% have structures with a better resolution than 2.5 Å (Supplementary Figure S5).

Can protein dynamical behaviour be inferred by homology?

The SG initiatives pursue the expansion of structural coverage to include at least one structural representative per each family, and to derive the remaining structures using homology modelling. Is this purpose still valid for 'dynamical' coverage? Could the dynamical behaviour of proteins be inferred using homology? To answer this question, we first considered the dynamical descriptions of rigid body inter-domain movements recently extended to describe up to 16 classes in the Dyndom database [8]. All the proteins with >30% of local sequence identity and with a coverage of 90% included in Dyndom were clustered to obtain groups of putative homologous proteins. We found that 81.8% of these families contain proteins with at least two different dynamical classes (Figure 2A). Interestingly, in proteins showing different dynamical behaviour, 60% of the studied families have maximum sequence identity above 40% (~15% of the protein families >80% identity). This finding reflects that minor to moderate changes in sequence can produce great differences in dynamical behaviour (Figure 2B). Similar results were obtained using the PSCDB database [9], which describes seven different classes of local and global motions coupled with ligand binding. Again, clustering proteins from this database with at least 30% sequence identity, we found that 60% of the families have different motions (Figure 2). Finally, we measured if the backbone dynamics are conserved among proteins of the same family using proteins from the CoDNaS database [10]. CoDNaS provides experimentally based backbone flexibility per protein expressed as the maximum C-alpha root mean square deviation (maxRMSD) among corresponding conformers. In total, 93 families from CoDNaS were analysed, considering that a family have the same backbone flexibility if the difference between the two

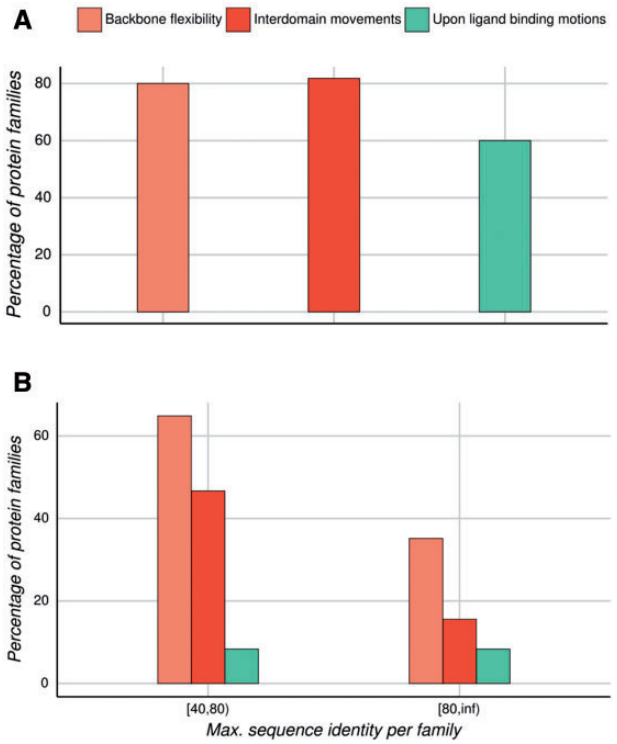

Figure 2. Dynamical behaviour within homologous protein families. Proteins with known motion classes describing three different dynamical behaviour (backbone flexibility, inter-domain movements and on ligand-binding motions) were compared to explore the class conservation within each family. Panel (**A**) shows percentages of protein families showing different motion classes among their proteins, while Panel (**B**) shows percentages of proteins per family showing different motion classes but being above a given cut-off (40 and 80% maximum sequence identity).

more extreme maxRMSD values within the family was <0.7 Å (propagated crystallographic error) [6]. We found that ∼80% of the families have larger maxRMSD differences than this threshold, suggesting variation in their backbone movements (Figure 2) (Supplementary Methods).

Conclusions

We have shown that sequence redundancy, a key concept for exploring protein dynamics using alternative structures of the same protein, is highly biased and fundamentally incomplete in the PDB. Nonetheless, the structural and dynamical incompleteness of the PDB is apparently unrelated concepts in SG. While the first could be reversed by promoting extension of the structural coverage, dynamical incompleteness could require further efforts. The apparent uniqueness of dynamical patterns allows to suggest that inference by homology is not a promising strategy to infer protein dynamics, even between proteins with small sequence divergence. Further studies and developments to improve dynamical behaviour prediction, based on sequence and structural data, are thus required to obtain a dynamical representation of entire protein families.

Key Points

- Protein alternative conformations under native conditions can be studied using crystallographic data of different structures of the same protein. Consequently, redundancy in the PDB is an essential ingredient for studying protein dynamics.
- Protein dynamical behaviour is apparently unique, making difficult to infer it by homology, even between proteins with small sequence divergence.
- Further efforts are needed to improve dynamical behaviour prediction based on sequence and structure data.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Funding

The Agencia de Ciencia y Tecnología (grant number PICT-2014-3430) and Universidad Nacional de Quilmes (grant number 1402/15). G.P., C.M.B. and S.M.F. are researchers of CONICET, and A.M.M. and D.Z. are PhD and postdoctoral fellows of the same institution.

References

1. Lander ES, Linton LM, Birren B, *et al*. Initial sequencing and analysis of the human genome. *Nature* 2001;**409**:860–921.
2. Chandonia J-M, Brenner SE. The impact of structural genomics: expectations and outcomes. *Science* 2006;**311**:347–51.
3. Levitt M. Growth of novel protein structural data. *Proc Natl Acad Sci USA* 2007;**104**:3183–8.
4. Khafizov K, Madrid-Aliste C, Almo SC, *et al*. Trends in structural coverage of the protein universe and the impact of the protein structure initiative. *Proc Natl Acad Sci USA* 2014;**111**:3733–8.
5. Best RB, Lindorff-Larsen K, DePristo MA, *et al*. Relation between native ensembles and experimental structures of proteins. *Proc Natl Acad Sci USA* 2006;**103**:10901–6.
6. Burra PV, Zhang Y, Godzik A, *et al*. Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure. *Proc Natl Acad Sci USA* 2009;**106**:10505–10.
7. Gerstein M, Lesk AM, Chothia C. Structural mechanisms for domain movements in proteins. *Biochemistry* 1994;**33**:6739–49.
8. Taylor D, Cawley G, Hayward S. Classification of domain movements in proteins using dynamic contact graphs. *PLoS One* 2013;**8**:e81224.
9. Amemiya T, Koike R, Kidera A, *et al*. PSCDB: a database for protein structural change upon ligand binding. *Nucleic Acids Res* 2012;**40**:D554–8.
10. Monzon AM, Rohr CO, Fornasari MS, *et al*. CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state. *Database* 2016, doi:10.1093/database/baw038.