CoDNas-Q: a database of conformational diversity of the native state of proteins with quaternary structure

Nahuel Escobedo^{1,4}, Ronaldo Romario Tunque Cahui², Gastón Caruso¹, Emilio García Ríos², Layla Hirsh², Alexander Miguel Monzon³, Gustavo Parisi^{1,4}, Nicolas Palopoli^{1,4,*}.

¹ Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Buenos Aires, Argentina.

² Departamento de Ingeniería, Pontificia Universidad Católica del Perú, Lima, Perú.

³ Department of Biomedical Sciences, University of Padova, Padova, Italy.

⁴ Consejo Nacional de Investigaciones Científicas y Técnicas, Buenos Aires, Argentina.

*To whom correspondence should be addressed.

Abstract

Summary: A collection of conformers that exist in a dynamical equilibrium defines the native state of a protein. The structural differences between them describe their conformational diversity, a defining characteristic of the protein with an essential role in multiple cellular processes. Since most proteins carry out their functions by assembling into complexes, we have developed CoDNas-Q, the first online resource to explore conformational diversity in homooligomeric proteins. It features a curated collection of redundant protein structures with known quaternary structure. CoDNas-Q integrates relevant annotations that allow researchers to identify and explore the extent and possible reasons of conformational diversity in homooligomeric protein complexes.

Availability and implementation: CoDNas-Q is freely accessible at <http://ufq.unq.edu.ar/codnasq/>. The data can be retrieved from the website. The source code of the database can be downloaded from <https://github.com/SfrRonaldo/codnas-q>.

Contact: npalopoli@unq.edu.ar

Introduction

The native state of a protein is represented by an ensemble of alternative conformers in equilibrium. Structural differences between these conformers define the extent of the protein's conformational diversity, from local loop rearrangements to large displacements of entire subunits. These conformational changes are usually of biological relevance, accounting for diverse functional mechanisms (Monzon *et al.*, 2017) and related with processes such as enzyme catalysis (Henzler-Wildman *et al.*, 2007), promiscuity (James and Tawfik, 2003), binding specificity (Yogurtcu *et al.*, 2008), allostery (Changeux, 2012) and cell signaling (Nussinov and Ma, 2012).

Most proteins carry out their functions by assembling into complexes, forming obligate or non-obligate (permanent or transient) interactions between two or more polypeptide chains (Nooren and Thornton, 2003). This is reflected in structural databases, where more than half of the known structures are multimeric, generally forming homooligomers (Bertoni *et al.*, 2017); (Levy *et al.*, 2006). Conformational diversity is functionally relevant at the quaternary structure (QS) level too. It has been shown that the conformational diversity of proteins upon complex formation is strongly dependent on the flexibility of their free monomers (Marsh *et al.*, 2012). The degree of protein flexibility is also associated with different quaternary topologies, reflecting the structural constraints on evolutionary history imposed by characteristic solvent accessible surface areas and interface contacts between subunits (Marsh and Teichmann, 2014). Addressing conformational diversity at all structural levels appears essential in pursuing a better understanding of protein function.

Our group has previously developed CoDNas, a database of Conformational Diversity on the Native State of proteins (Monzon *et al.*, 2016). It has allowed us to explore the structural and functional implications of conformational diversity of individual protein chains. (We recently published CoDNas-RNA with the similar goal of addressing conformational diversity of single RNA chains (Buitrón *et al.*, 2021)). In this work we extend our approach by presenting CoDNas-Q, a database of conformational diversity of the native state of proteins with QS. Each entry in CoDNas-Q presents a cluster of known structures of the same complex available from the Protein Data Bank (PDB) (Berman *et al.*, 2000), along with all-vs-all structural comparisons and annotations to facilitate the inspection of its conformational diversity. CoDNas-Q is currently limited to homooligomers, following their overrepresentation in the PDB.

Dataset

Experimental-based evidence of conformational diversity comes from the analysis of known protein structures in the Protein Data Bank (wwPDB consortium, 2019). A collection of crystallographically determined structures of a protein can be interpreted as snapshots of its native conformational ensemble. Thereby, conformational diversity can be studied using redundant structures of the same protein obtained in different experimental conditions (with or without ligands, varying pH or temperature, etc). We used sequence clusters of all individual chains from the PDB obtained with BLASTClust at 95% sequence identity, and only kept clusters with two or more chains from two or more PDB files (see Figure 1, panel A for an overview of the development scheme). Each cluster can be interpreted as a set of native conformers of a homooligomeric protein.

Figure 1. Panel A: CoDNas-Q development scheme. **Panel B:** Cluster Information section from a sample entry in CoDNas-Q. The table presents general data about the cluster and selected annotations from a representative conformer. **Panel C:** Comparison of conformers from a sample entry in CoDNas-Q. The table highlights data from the pair of conformers with maximum RMSD at the quaternary level that is useful to understand the extent and possible reasons of the observed conformational diversity in the cluster. An interactive window provides the structural superposition of the maximum RMSD conformers. A dendrogram and a heatmap of the RMSD-based hierarchical clustering of the available conformers facilitates inspection of structural differences between any pair of conformers.

The preliminary set of clusters was filtered to retain only those with a reliably assigned oligomeric state. To this end, the conformers with a biologically relevant QS were selected with QSBio, a method to annotate the QS of biological assemblies with high confidence by combining prediction tools and databases, and thus achieve accuracy close to manual curation (Dey *et al.*, 2018). CoDNas-Q is limited to conformers with two or more subunits and High or Very High confidence on QSBio annotations of their QS. We also discard entries with ambiguous QS assignments. The first release of CoDNas-Q comprises 3649 clusters and 18790 conformers. In the near future we plan to update CoDNas-Q with more clusters comprising homo- and hetero-oligomers, with an API REST service for data exchange with users and other data resources.

The RMSD (Root Mean Square Deviation) is a common metric used in structural biology to measure the dissimilarity between two biological structures that also provides a valid estimation of a protein's conformational diversity. CoDNas-Q uses the RMSD calculated with TopMatch (Sippl and Wiederstein, 2012) to compare multimeric complexes and individual chains. We compute the Ca-RMSD between all pairs of multimeric conformers within every cluster and identify the maximum RMSD at the quaternary level (maxRMSD-Q) to quantify the degree of conformational diversity in the ensemble. TopMatch also provides other parameters useful to characterize an ensemble of conformers such as typical distance error, alignment coverage, number of structurally equivalent residue pairs, etc. Each cluster is annotated with additional data like the maximum RMSD between any two individual chains (maxRMSD-T) and features extracted from the PDB (source organism, experimental method, pH and temperature, ligands, etc), UniProt (cross-referenced gene name and entry ID), and Pfam (family ID).

Website

The website allows users to search CoDNas-Q by Cluster ID, PDB code, name and description of an entry, source organism, related UniProt accession ID, type of movement and oligomeric state, among other criteria.

At the top of each CoDNas-Q entry, the Cluster Information section (Figure 1, panel B) presents general information about the structural properties and the conformational diversity related with the cluster, along with selected annotations from a representative conformer. Below, a table compares relevant features from the maximum pair of conformers, i.e. those with the maximum RMSD value between them (Figure 1, panel C). It also includes the interactive superposition of the structures (available for download). A summary of the hierarchical clustering of conformers based on their pairwise RMSD values is shown as a dendrogram, with coloring schemes that reflect the extent of conformational diversity among conformers and in comparison with the entire database. The bottom of the page displays a sortable list of all conformers in the cluster and their structural properties. Clicking on a row provides extended information about the conformer, cross-referenced from other databases, together with its 3D structure. Two or more conformers can be selected to access a dedicated page to compare all possible pairs among the selected conformers.

Conclusions

CoDNas-Q is a unique resource that helps the community to learn more about the structural nature of protein complexes and their conformational diversity patterns. It helps us to understand the complexity of studying the QS in

proteins of biological interest, and raises questions about the dynamics that a protein can present in its native state, as well as the factors that condition this ensemble. It provides valuable insights about the extension of quaternary conformational diversity and its relationship with protein function. We think CoDNaS-Q can be an essential resource for both computational biologists and life science researchers alike.

Funding

This work has been supported by grants from Universidad Nacional de Quilmes (PUNQ 1309/19), Agencia Nacional de Promoción de la Investigación, el Desarrollo Tecnológico y la Innovación (PICT-2018 3457) and Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) (PIP-2015-2017 11220150100853CO) from Argentina, and the European Union's Horizon 2020 Research and Innovation Staff Exchange program (grant agreements 778247 and 823886).. N.E. is a PhD Fellow and G.P. and N.P. are Researchers from CONICET.

Acknowledgements

The authors would like to thank Martin Gonzalez Buitrón for his help in designing and reviewing CoDNaS-Q.

References

- Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
- Bertoni,M. *et al.* (2017) Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Sci. Rep.*, **7**, 10480.
- Buitrón,M.G. *et al.* (2021) CoDNaS-RNA: a database of Conformational Diversity in the Native State of RNA. *Bioinformatics*.
- Changeux,J.-P. (2012) Allostery and the Monod-Wyman-Changeux model after 50 years. *Annu. Rev. Biophys.*, **41**, 103–133.
- Dey,S. *et al.* (2018) PDB-wide identification of biological assemblies from conserved quaternary structure geometry. *Nat. Methods*, **15**, 67–72.
- Henzler-Wildman,K.A. *et al.* (2007) Intrinsic motions along an enzymatic reaction trajectory. *Nature*, **450**, 838–844.
- James,L.C. and Tawfik,D.S. (2003) Conformational diversity and protein evolution--a 60-year-old hypothesis revisited. *Trends Biochem. Sci.*, **28**, 361–368.
- Levy,E.D. *et al.* (2006) 3D complex: a structural classification of protein complexes. *PLoS Comput. Biol.*, **2**, e155.
- Marsh,J.A. and Teichmann,S.A. (2014) Protein flexibility facilitates quaternary structure assembly and evolution. *PLoS Biol.*, **12**, e1001870.
- Marsh,J.A. *et al.* (2012) Probing the diverse landscape of protein flexibility and binding. *Curr. Opin. Struct. Biol.*, **22**, 643–650.
- Monzon,A.M. *et al.* (2016) CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state. *Database (Oxford)*, **2016**.
- Monzon,A.M. *et al.* (2017) Conformational diversity analysis reveals three functional mechanisms in proteins. *PLoS Comput. Biol.*, **13**, e1005398.
- Nooren,I.M.A. and Thornton,J.M. (2003) Diversity of protein-protein interactions. *EMBO J.*, **22**, 3486–3492.
- Nussinov,R. and Ma,B. (2012) Protein dynamics and conformational selection in bidirectional signal transduction. *BMC Biol.*, **10**, 2.
- Sippl,M.J. and Wiederstein,M. (2012) Detection of spatial correlations in protein structures and molecular complexes. *Structure*, **20**, 718–728.
- wwPDB consortium (2019) Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.*, **47**, D520–D528.
- Yogurtcu,O.N. *et al.* (2008) Restricted mobility of conserved residues in protein-protein interfaces in molecular simulations. *Biophys. J.*, **94**, 3475–3485.

A

B

Cluster Information

Cluster ID:	CQ1yl6	Name:	Dihydrodipicolinate reductase
Group:	Rigid Body	Description:	crystal structure of Mycobacterium tuberculosis dihydrodipicolinate reductase (Rv2773c) (crystal form B)
Oligomeric State:	Tetramer	Organism:	MYCOBACTERIUM TUBERCULOSIS
Number of Conformers:	5	Genes:	dapB Rv2773c MTV002.38c
Max RMSD Quaternary:	7.04 Å	Length:	980
Max RMSD Tertiary:	1.58 Å		

C

Maximum RMSD Quaternary pair Comparison

	Max RMSD Quaternary	Coverage	Biological Assembly	pH	Length	Resolution	Ligands
1yl5	7.04Å	59	1	8.5	988	2.3	MG magnesium ion
1p9l		60	1	8.5	980	2.3	NAI 1,4-dihyronicotinamide adenine dinucleotide, PDC pyridine-2,6-licarboxylic acid PG4 tetraethylene

● 1yl5 (query) ● 1p9l (target)