ARTICLE TITLE:

Combining protein conformational diversity and phylogenetic information using CoDNaS and CoDNaS-Q

AUTHOR(S) AND CONTACT INFORMATION:

Nahuel Escobedo,[1,3,#] Alexander Miguel Monzon,[2,#] María Silvina Fornasari,[1,3] Nicolas Palopoli[1,3,*] and Gustavo Parisi[1,3,*]

[1]Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Bernal, Argentina.

[2]Department of Biomedical Sciences, University of Padova, Padova, Italy.

[3]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina.

[#]Authors contributed equally to this work

[*]Corresponding authors: nicopalo@gmail.com and gusparisi@gmail.com

ABSTRACT:

CoDNaS (http://ufq.unq.edu.ar/codnas/) and CoDNaS-Q (http://ufq.unq.edu.ar/codnasq) are repositories of proteins with different degrees of conformational diversity. Following the ensemble nature of the native state, conformational diversity represents the structural differences between the conformers in the ensemble. Each entry in CoDNaS and CoDNaS-Q contains a redundant collection of experimentally determined conformers obtained under different conditions. These conformers represent snapshots of the protein dynamism. While CoDNaS contains examples of conformational diversity at the tertiary level, a recent development, CoDNaS-Q, contains examples at the quaternary level. In the emerging age of accurate protein structure prediction, many questions remain open regarding the characterization of protein dynamism. In this context, most bioinformatics resources take advantage of distinct features derived from protein alignments, however, the complexity and heterogeneity of information makes it difficult to recover reliable biological signatures. Here we present four protocols to explore tertiary and quaternary conformational diversity at the individual protein level as well as for the characterization of the distribution of conformational diversity at the protein family level in a phylogenetic context. These protocols can provide curated protein families with experimentally known conformational diversity, facilitating the exploration of sequence determinants of protein dynamism.

Basic Protocol 1: Performing a search on CoDNaS

Basic Protocol 2: Performing a search on CoDNaS-Q

Basic Protocol 3: Exploring conformational diversity in a protein family

Basic Protocol 4: Representing conformational diversity in a phylogenetic context

KEYWORDS:

Conformational diversity, protein family, phylogeny.

INTRODUCTION

The ensemble nature of a protein describes its native state as a collection of conformers in a dynamic equilibrium (Wei et al., 2016). This concept is central for the understanding of protein biology and gives mechanistic explanations in several process such as allosterism and cooperativism (Motlagh et al., 2014; del Sol et al., 2009), protein-protein interactions (Goh et al., 2004), protein signaling and information propagation (Wei et al., 2016; Boehr et al., 2009), enzyme catalysis (Ma and Nussinov, 2010; Hammes, 2002), promiscuity (Honaker et al., 2011), sequence substitution pattern during evolution (Parisi et al., 2015; Zea et al., 2013), protein evolvability (Tokuriki and Tawfik, 2009), drug design (Lin, 2011), sequence variants effects (Juritz et al., 2012), etc. Structural differences between conformers can be as tiny as those observed in the 'rigid' proteins, where slight residue movements or rotations allow the transit of ligands to the binding site of the protein, opening tunnels or enlarging cavities without backbone translations (Monzon et al., 2017). Movements of flexible regions or loops, displacements of secondary structure elements and relative rotations of domains commonly contribute to an increased structural difference between conformers (Gerstein and Krebs, 1998; Gerstein et al., 1994).

To better understand the relationship between conformational diversity and protein biology we developed the databases CoDNaS (Monzon et al., 2016) and CoDNaS-Q (Escobedo et al., 2022). Currently, CoDNaS contains ~30000 polypeptides with different degrees of conformational diversity. Each protein is represented by a collection of redundant, experimentally obtained conformers (~14 on average per protein). In the same way, CoDNaS-Q includes clusters of conformers for ~3600 high-confidence homooligomeric proteins (~5 conformers on average per protein). Conformers in these databases are taken as snapshots of protein dynamism, as shown by several studies by comparing collections of alternative crystallographic and NMR structures (Kondrashov et al., 2008; Best et al., 2006).

As conformational diversity and the underlying dynamic behavior between different conformations offer mechanistic explanations of protein function, a large number of scientific works explored the conservation of protein dynamics during evolution (Leo-Macias et al., 2005b, 2005a; Maguid et al., 2005, 2006). Most of these works found a common dynamic behavior in protein families. However, it is also known that functional divergence during evolution is almost necessarily sustained by dynamical and conformational changes (Narayanan et al., 2017; Mitchell-White et al., 2021; Glembo et al., 2012; Liu and Bahar, 2012). Furthermore, a more detailed analysis of backbone flexibility, inter-domain movements and ligand-binding motions revealed a poor conservation of the dynamic

behavior in homologous families (Marino-Buslje et al., 2019). An interesting example of these adaptations is represented in the hexokinase family, a large family with at least four isoenzymes (I-IV) essential in the management of glucose in mammals (Irwin and Tan, 2014). Glucokinase (hexokinase IV or D) is a cytoplasmic enzyme that phosphorylates glucose in liver and pancreas, maintaining the glucose homeostasis in mammals. Its unique kinetic behavior contrasts with the rest of the isoenzymes present in other tissues. Glucokinase activity relies in the presence of a specific conformer, called super-open form, and a particular flexibility pattern (Larion et al., 2012), which as a whole confers the ability to display positive cooperativity effects as a function of blood glucose concentration (Kamata et al., 2004; Whittington et al., 2015).

Experimentally-based exploration of conformational diversity in a phylogenetic context can provide biological insights about functional divergence and its mechanistic explanations. Furthermore, in a practical context, easy access to this information could provide gold standards in the light of new resources to predict protein structure using evolutionary information, like AlphaFold2 (Jumper et al., 2021), and to further develop promising protocols for sequence-based prediction of conformational diversity (del Alamo et al., 2021; Heo and Feig, 2021). Furthermore, it was recently found how distribution of conformational diversity among protein families included in sequence alignments used by AlphaFold2 predictions affects its predictive performance (Saldaño et al., 2022). In this article we provide detailed protocols on how to obtain information about the conformational diversity of a protein at the tertiary structural level using CoDNaS (Basic Protocol 1) and at the quaternary level using CoDNaS-Q (Basic Protocol 2). We also cover how to use this information to analyze the conformational diversity distribution in a given protein family (Basic Protocol 3) and display it in a phylogenetic context (Basic Protocol 4).

BASIC PROTOCOL 1

Basic protocol title

Performing a search in CoDNaS

Introductory paragraph

CoDNaS is freely accessible as a website at http://ufq.unq.edu.ar/codnas. This protocol describes how to perform basic and advanced searches in CoDNaS, explore the results and download useful data.

Necessary resources

- Hardware

CoDNaS can be displayed in different devices such as laptops and desktop computers. An active and stable internet connection is required.

- Software

An up-to-date web browser, such as Google Chrome (http://www.google.com/chrome/) or Firefox (http://www.mozilla.org/firefox/).

Protocol steps with step annotations

Basic search

1. Open a browser and navigate to http://ufq.unq.edu.ar/codnas.
2. Perform a search by typing a PDB or UniProt identifier, or the name of the protein, in the input field and then pressing the 'Search' button (Figure 1).
   Alternatively, use the example terms (Example 1, Example 2 or Example 3) provided below the text box.

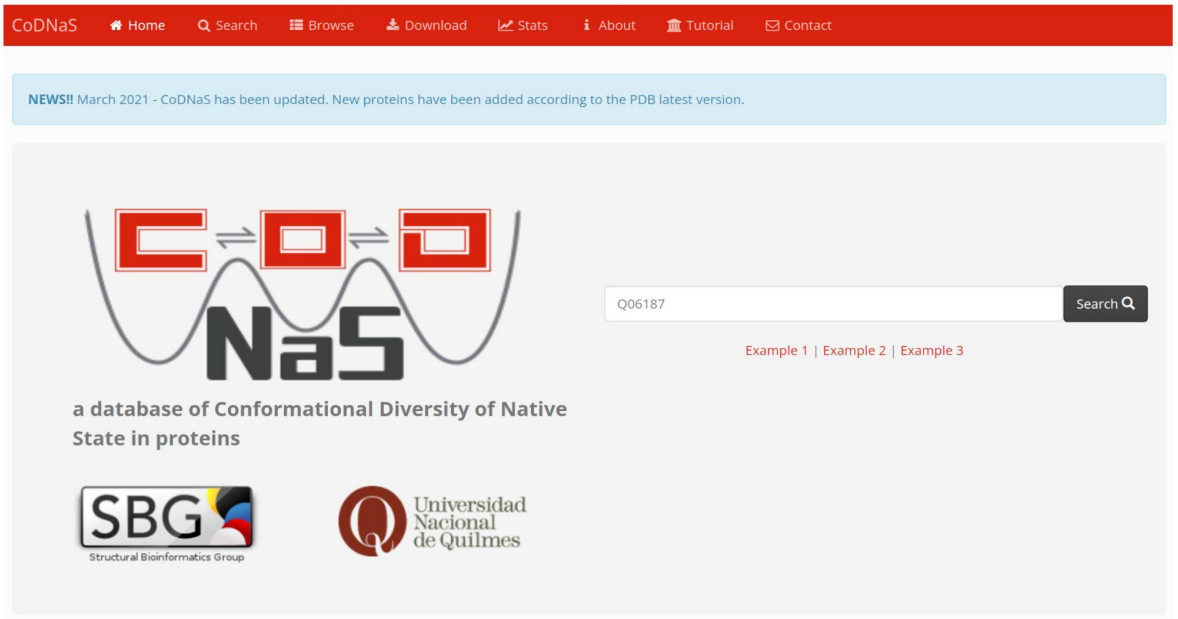


Figure 1. CoDNaS home page.

Advanced search

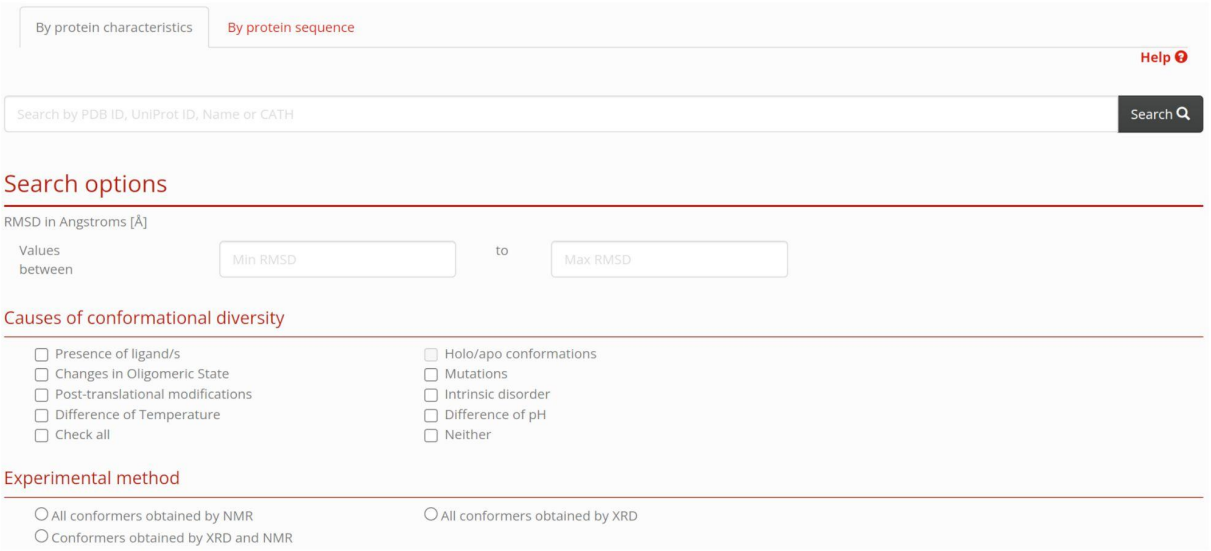
1. Access the advanced search interface by clicking on the 'Search' button in the navigation bar at the top of the website, or directly at http://ufq.unq.edu.ar/codnas/search.php. Make an advanced search by protein characteristics (Figure 2A) or by protein sequence (Figure 2B).
2. The protein characteristics available for searching include:
   a. PDB identification code (e.g. 1AWX), UniProt identifier (e.g. Q06187), protein name (e.g. tyrosine kinase) or CATH (Orengo et al., 1997) superfamily identifier (e.g. 2.30.30.40).
   b. Extension of conformational diversity.
      Define a range of RMSD values in order to retrieve proteins within a given extension of conformational diversity.
   c. Causes of conformational diversity.
      Tick the checkboxes to retrieve clusters with pairs of conformers that differ in the

presence/absence of a ligand, a change in oligomeric state or differences in pH or temperature in the experimental setup, among other possible causes of conformational diversity.

d. Experimental method.
Filter by the experimental method used to determine the structure of conformers from a given protein.

3. When searching with a protein sequence as input, paste a protein (as plain text or in Fasta format) in the text box. CoDNaS runs a BLASTp search against all protein sequences in the database. Alternatively, use a DNA coding sequence and CoDNaS will run BLASTx to search with the translated protein sequence as a query.
The retrieved list of clusters has a BLAST E-value cutoff of 1E-04.

A.



B.

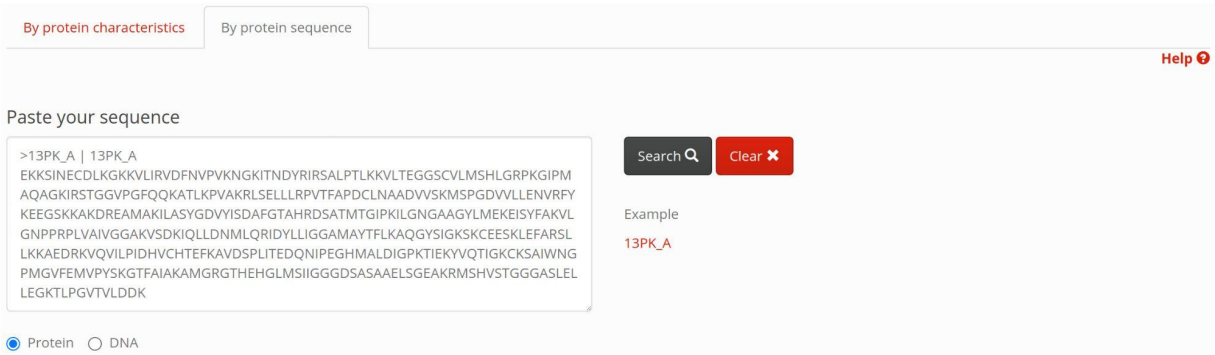


Figure 2. Advanced search options. Search by protein characteristics (**A**). Search by protein or DNA sequence (**B**).

Browse all entries

1. Another alternative is to browse CoDNaS according to the CATH hierarchy (Orengo et al., 1997). Hit the 'Browse' button in the navigation bar at the top to access a hierarchical tree diagram of CATH.
   Expand the tree with the 'plus' icon to show the category of interest. Click the 'magnifying glass' icon to get all the clusters with at least one conformer that belongs to that category.

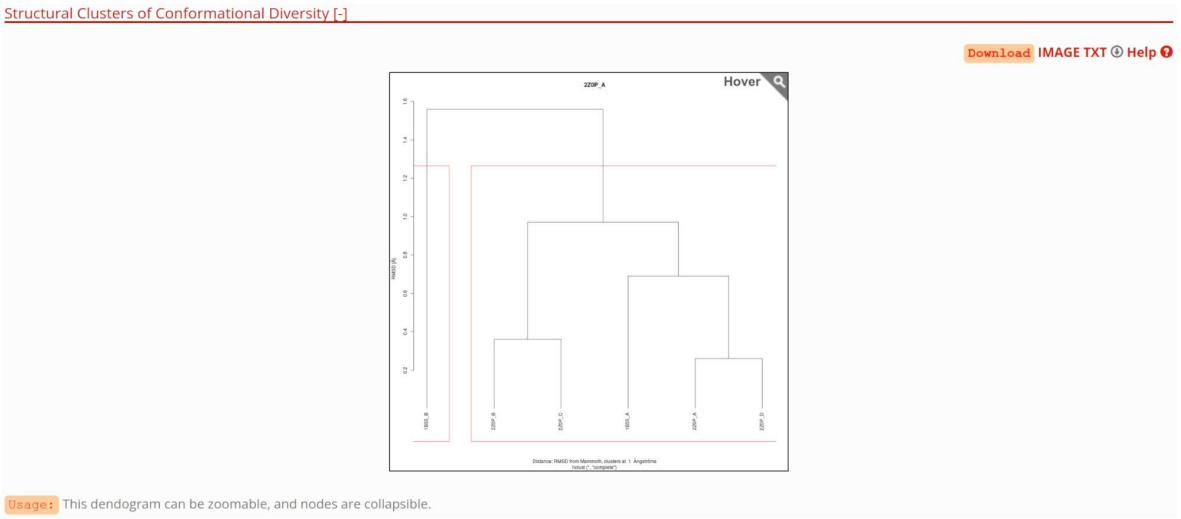
Search results

1. After a search, the results page displays a table of matching CoDNaS clusters. Click on a row to access the protein description page.
   This table lists the clusters by CoDNaS identifiers ('ID_POOL_CoDNaS' column, with each identifier based on the PDB code from a representative conformer). The columns include the UniProt identifier and protein name, along with the number of conformers in the cluster ('#CONF') and its minimum, maximum and average RMSD values. Navigate with the pagination buttons below the table or insert the desired value in the 'Search' space to filter the results.

View a protein entry in CoDNaS

1. This first section is 'General information'. It gives an overview of the biological information of the selected protein, including source organism, molecular function and subcellular localization. It also provides structural information of the entire cluster such as the number of conformers, the percentage of conformers obtained by Nuclear Magnetic Resonance (NMR) or X-ray diffraction (XRD) techniques, the range of pairwise sequence identity between conformers and the range of RMSD and TM-score values.
   Users can download the information in JSON or TSV (tab-separated values) formats by using the links in the top-right corner of this section.
   The website displays this section by default. To expand the contents of the following sections, click on the 'plus' sign next to the section title.
2. The section 'Conformers' lists all protein structure conformations. It provides specific information on the experimental method, resolution (only in X-ray structures), UniProt identifier, presence of ligands or post-translational modifications and missing residues in the structure file.
   Select two or more conformers using the checkboxes and click the 'Submit' button to get their structural comparison. See section 'Comparison of selected conformers' below for details.
3. The third section is 'Structural Clusters of Conformational Diversity'. It provides a hierarchical clustering of conformers by RMSD. This section has a dendrogram on top (Figure 3A) and an interactive flexible force-directed graph layout (Figure 3B) that allows the user to identify similar conformers according to the RMSD value in *all vs all* comparisons.
   Hover the mouse pointer over the dendrogram to zoom in. Use the mouse wheel on the graph to zoom in and out and click on a node to collapse it for a simpler display.

A.



B.

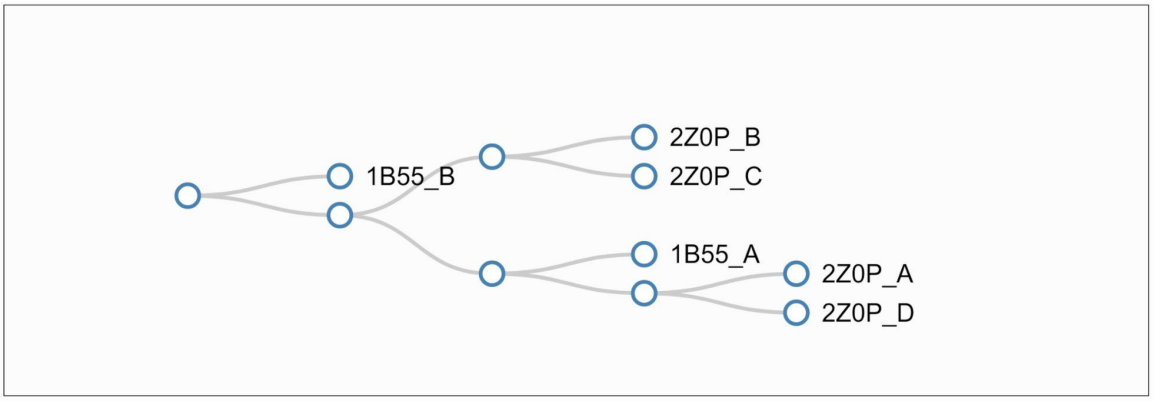


Figure 3. RMSD-based hierarchical clustering dendrogram between conformers of a cluster, exemplified by cluster 2Z0P_A (**A**). Same Relationship shown as an interactive graph (**B**).

4. The last section gives 'Information about pair maximum of conformational diversity'. The table shows the differences between the experimental conditions and other characteristics of the protein's pair of conformers with maximum RMSD between them (Figure 4).
   Press the 'View details' button to get more information on the structural comparison of this pair of conformers. Click on a row for a deeper comparison. See section 'Comparison of selected conformers' below for details.

Information about pair maximum of conformational diversity [-]

Help

View details	1B55_A	1B55_B	1B55_A / 1B55_B
RMSD [Å] calculated by Mammoth			1.56
TM value			0.9729
GDT-HA			0.9043
GDT-TS			0.9769
Sequence identity [%]			100.0
Experimental method	X-RAY DIFFRACTION	X-RAY DIFFRACTION	
Resolution [Å]	2.4	2.4	
Source	man	man	
pH	4.9	4.9	
Difference pH			0
Temperature [k]	100	100	
Difference Temperature [k]			0
Oligomeric state by PISA	4	4	
Oligomeric state by author	TETRAMERIC	TETRAMERIC	
Ligands HETATM	ZN,4IP	ZN,4IP	
Ligands Biolip Database	ZN,4IP	ZN,4IP	
Mutation	No	No	
Factors of conformational diversity			LIGAND,LIGAND BIOLIP,DISORDER
Accessible surface area of protein [Å2]	10052.4	9626.21	
Difference Global ASA			426.190000000001
Intrinsic disorder (REMARK 465)	6	7	

Figure 4. Detailed information about the maximum RMSD pair, i.e. the pair of conformers with maximum RMSD between them, exemplified by cluster 2Z0P_A.

Comparison of selected conformers

1. As previously mentioned, the 'Conformers' section of a CoDNaS entry page contains individual information about each conformer and allows their structural comparison (Figure 5). Select different conformers to compare by clicking the checkboxes and press the 'Submit' button.
   Two or more conformers should be selected. Use 'Check all' for an all-vs-all comparison.

Conformers [-]

Usage: Select conformers for structural comparison Help

Submit

Check all	PDB ID	Length	Resolution	UniProt	Experimental method	Ligand	Post. Mod.	Miss. Res.
☑	1B55_A	169	2.4	Q06187	X-RAY DIFFRACTION	Yes	No	Yes
☑	1B55_B	169	2.4	Q06187	X-RAY DIFFRACTION	Yes	No	Yes
☐	2Z0P_A	169	2.58	Q06187	X-RAY DIFFRACTION	Yes	No	Yes
☐	2Z0P_B	169	2.58	Q06187	X-RAY DIFFRACTION	Yes	No	Yes
☐	2Z0P_C	169	2.58	Q06187	X-RAY DIFFRACTION	Yes	No	Yes
☐	2Z0P_D	169	2.58	Q06187	X-RAY DIFFRACTION	Yes	No	Yes

FIgure 5. List of conformers that are available to compare *all-vs-all*.

2. The section 'Comparison of selected conformers' provides a table of the selected structural comparisons between conformers (Figure 6A). The table presents (dis)similarity measures like the pairwise RMSD, the number of aligned residues, TM-score, GDT-TS/GDT-HA values, etc.

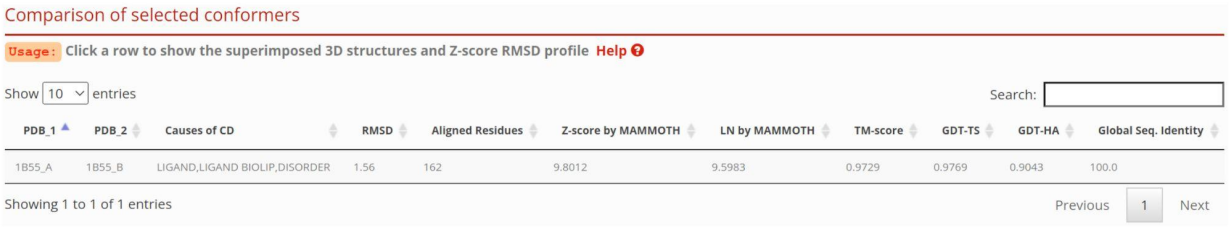
   Click on a row for a thorough comparison. The table remains visible and can be clicked again to show details of a different structural comparison.

3. After selecting a pair, the 'Structures' section will provide details on their structural comparison. Interact with the JSmol 3D visualizer that shows the selected superposed structures (Figure 6B). On the right side are different 'Display' and 'Color Options' available to customize the view.

   Click the 'Z-score RMS' button to see the structures colored according to the extent of C-alpha RMS by position, normalized in Z distribution. In addition, click the 'Z-score B-factor' button to see the superposed structures colored according to the extent of C-alpha B-factors, also normalized in Z distribution.

   In order to check the structural alignment between conformers, users can download the superposed structures in PDB format by pressing the links in the top-right corner of this section.

A.



B.

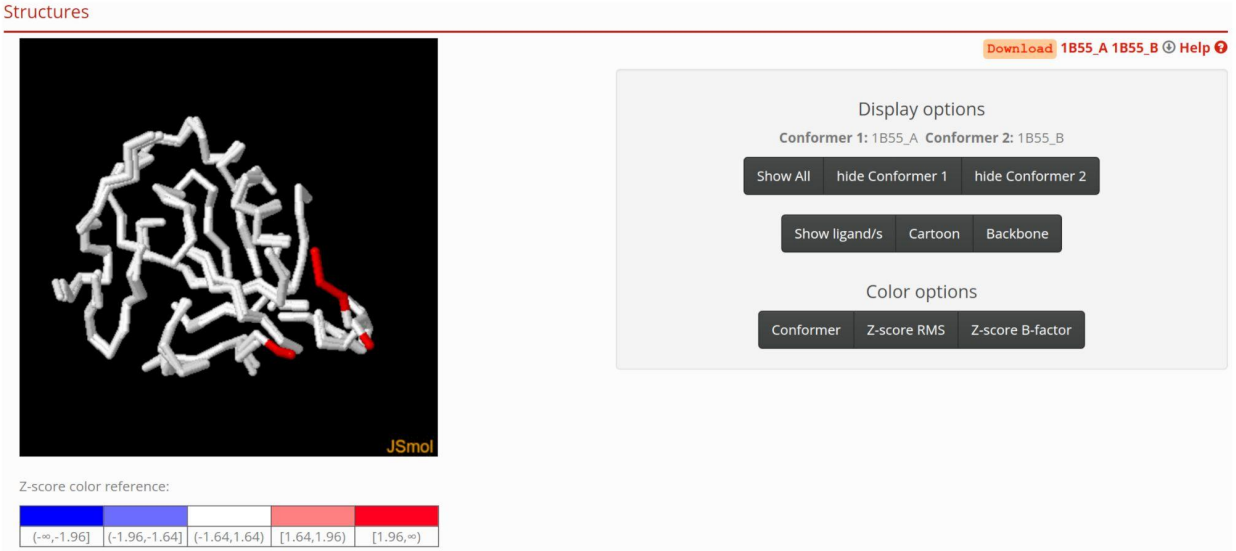


Figure 6. Structure section for the comparison of a selected pair of conformers. Table with all comparisons between the selected conformers (**A**). Interactive display of the structural superposition of the conformers(**B**).

4. The section 'Z-score profile by position' includes a plot that shows the C-alpha RMSD and C-alpha B-factor RMSD by position, both normalized in Z distribution (Figure 7A).
   Hover the plot to see the values at each position. The C-alpha B-factor RMSD are only available if structures were obtained by X-ray crystallography.
5. The 'Sequences alignment' section provides a structure-based sequence alignment that is derived from the same structural superposition presented in the previous step (Figure 7B).
   The residues in the alignment are colored by groups sharing the same physicochemical properties.
6. At the bottom, the section 'Information about pair of conformers' provides information about each of the selected conformers and their structural comparison.
   The contents of this section are identical to those presented in 'View a protein entry in CoDNaS' above. Again, users can download the information in JSON or TSV formats by using the provided links in the top-right corner.

Figure 7. Z-score profile by position (**A**). Sequences alignment (**B**).

Download

1. You can build a custom download. Hit the 'Download' button in the navigation bar at the top to access the Download builder.
2. On the text box to the left, paste a list of PDB codes corresponding to the conformers of interest (one code per row).
   The list of conformers should have one conformer per row. A valid format is 'PDB_CHAIN' (e.g. 1A22_B).
3. Choose the information you want to retrieve by using the checkboxes on the right side.
   A 'Check all' button allows retrieval of all available data for the specified entries at once.
4. Hit the 'Download' button to get all the pairwise comparisons for the selected PDB identifiers along with the information you choose to download presented as different columns.
   The website will provide the requested information as a file with tab-separated values.

BASIC PROTOCOL 2

Basic protocol title

Performing a search in CoDNaS-Q

Introductory paragraph

CoDNaS-Q is a database that follows the concept of CoDNaS but instead of single polypeptide chains, it clusters conformers of homo-oligomeric proteins. Their oligomeric states were assigned with high confidence by QSbio using a combination of several prediction methods (Dey et al., 2018).

Necessary resources

- Hardware

CoDNaS-Q can be displayed in different devices such as laptops, desktop computers, tablets and smartphones. An active and stable internet connection is required.

- Software

An up-to-date web browser, such as Google Chrome (http://www.google.com/chrome/) or Firefox (http://www.mozilla.org/firefox/).

Basic search

1. Open your web browser and navigate to the CoDNaS-Q landing page at https://codnas-q.bioinformatica.org.
2. Perform a search by first entering the query text in the input box and then selecting from the dropdown menu whether to match the name of a protein or UniProt identifier, the source

organism, a specific term in the description of an entry or the PDB code used as cluster identifier (Figure 8). Press the 'Search' button to access the 'Advanced Search' page with the results.
The website provides search term examples below the text box.
Also notice that if only one cluster is retrieved, the website displays the corresponding 'Cluster Information' page directly.

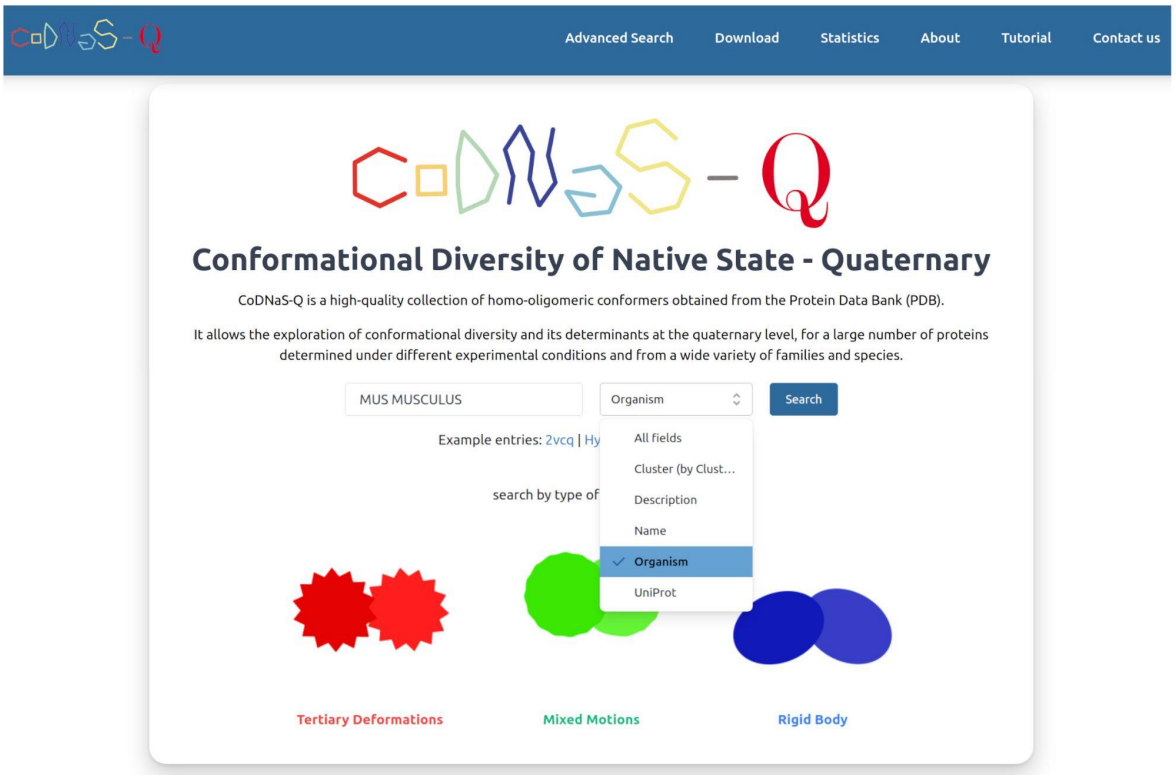


Figure 8. CoDNaS-Q home page. The dropdown menu lists different search criteria, e.g. 'Organism'.

3. Alternatively, search by the type of tertiary and quaternary movements that characterize the proteins by clicking on the animated images depicting 'Tertiary Deformations', 'Mixed Motions' or 'Rigid Body'.
   Click the question mark icon next to 'search by type of movement' or visit https://codnas-q.bioinformatica.org/about for more information about these categories.

Advanced search and browse all entries

1. Click the 'Advanced Search' button in the navigation bar at the top of the website or visit https://codnas-q.bioinformatica.org/adv-search to access the advanced search interface.
2. CoDNaS-Q shows the entire list of clusters available in the database by default. Each entry is displayed as a card with descriptive information of the cluster.
   Browse through all entries in CoDNaS-Q using the arrows located above and below the cards describing the entries.
3. Use the panels on the left to refine the search by selected characteristics of the protein. The website displays the search criteria organized in two different levels:

a. 'Cluster Properties': Search for clusters by intrinsic features such as oligomeric state, type of movement and RMSD range (Figure 9A).
b. 'Conformer Properties': Search for clusters by general properties of their individual conformers such as protein name or description (as registered in the PDB entry), source organism, X-ray crystallography resolution (if applicable) and length range (Figure 9B).

When applying multiple search terms, use the dropdown menu next to 'Cluster Properties' to specify if the results must satisfy all search fields simultaneously ('AND') or at least one of them ('OR').

4. After entering the parameters, click the blue 'Search' button to get the results on the left panel.

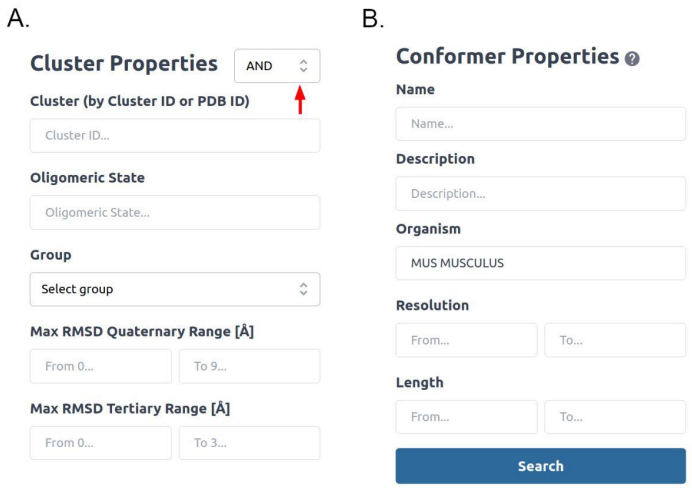


Figure 9. Left panel of the 'Advanced Search' page of CoDNaS-Q, allowing to search for 'Cluster Properties' (**A**) or 'Conformer Properties' (**B**). The red arrow indicates the dropdown menu to search by matching one ('OR') or all ('AND') the terms.

Search results

1. When a search returns multiple entries, the right panel of the Advanced Search interface lists the matching clusters as different cards that contain summarized and descriptive information for the entry (Figure 10).
   Each card presents the identifier of the cluster, its type of movement and oligomeric state, the number of conformers included in the cluster and the maximum pairwise RMSD at the quaternary and tertiary levels. It also shows an image of the protein's oligomeric structure. Browse through the results using the arrows above and below the cards.
2. Click on a card to open the respective cluster entry page.

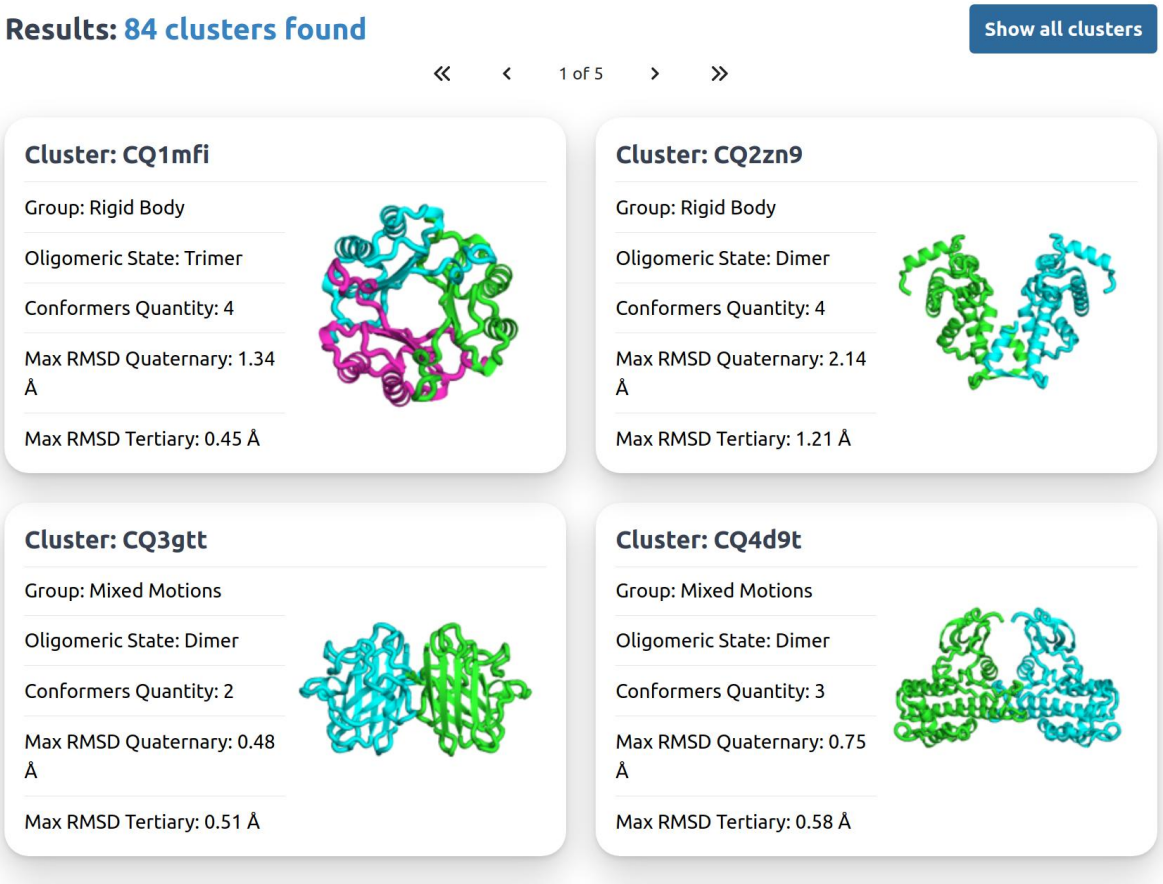


Figure 10. Right panel of the 'Advanced Search' page of CoDNaS-Q showing the first four cards retrieved as results from a sample search.

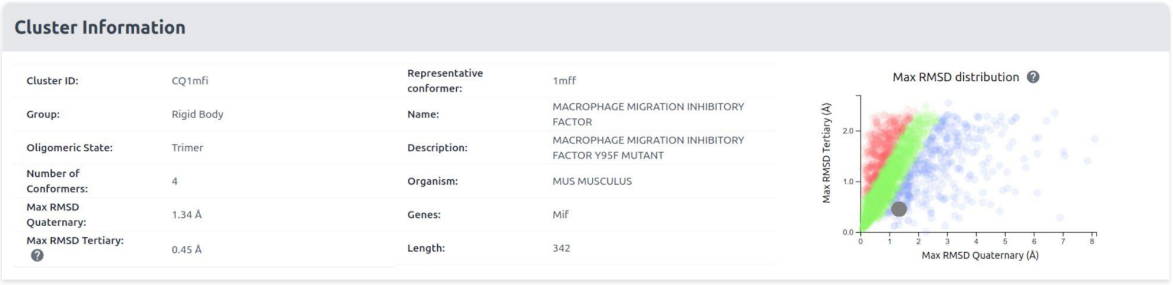
View a protein entry in CoDNaS-Q

1. The top section of a cluster entry is 'Cluster Information'. Along with the general information about the cluster already provided as a card in the search results page (see above), it gives an overview of a randomly chosen conformer as representative of the given cluster. On the right panel there is a plot comparing the distributions of maximum RMSD values at the tertiary and quaternary levels, with the points corresponding to the three types of movements shown with different colors and a gray point highlighting the position of the selected cluster in the plot (Figure 11A).
   The relationship between the maximum pairwise RMSD at the tertiary level ('maxRMSD-T') and the quaternary level ('maxRMSD-Q') is shown in green for proteins with mixed motions; in red for those mainly with tertiary deformations; and in blue for proteins behaving as rigid bodies.
2. The section 'Maximum RMSD Quaternary pair Comparison' describes the dissimilarities between the most different conformers of the cluster, as determined by their maximum pairwise RMSD (Figure 11B). The table on top summarizes precalculated data of their structural comparison obtained with TopMatch (Wiederstein and Sippl, 2020). Below, the website shows an interactive superposition of both conformers.
   The table's columns include comparative values such as sequence identity, quaternary RMSD

and number of structurally equivalent residue pairs. It also provides conformer-based values like alignment coverage, length, resolution, pH, temperature and presence of ligands in the PDB file.

3. The same section also provides a dendrogram and a heatmap based on the hierarchical clustering of pairwise RMSD values between all conformers in the cluster.
   Each cell of the heatmap matrix shows a pairwise RMSD value between two conformers. The color scale on the upper triangle heatmap is based on the range of RMSD values observed in the cluster, while the color scale on the lower triangle is based on RMSD values observed in the whole database.
   These resources are helpful to identify alternative pairs of conformers that may be of interest for better understanding the conformational diversity of the protein.
4. The last section is 'Conformers' and lists all conformers of the protein in the cluster. It provides information about the experimental details when solving the structure of each conformer.
   Use the checkboxes to select two or more conformers and click 'Compare' to get their structural comparison. For details see section 'Comparison of selected conformers' below.

A.



B.

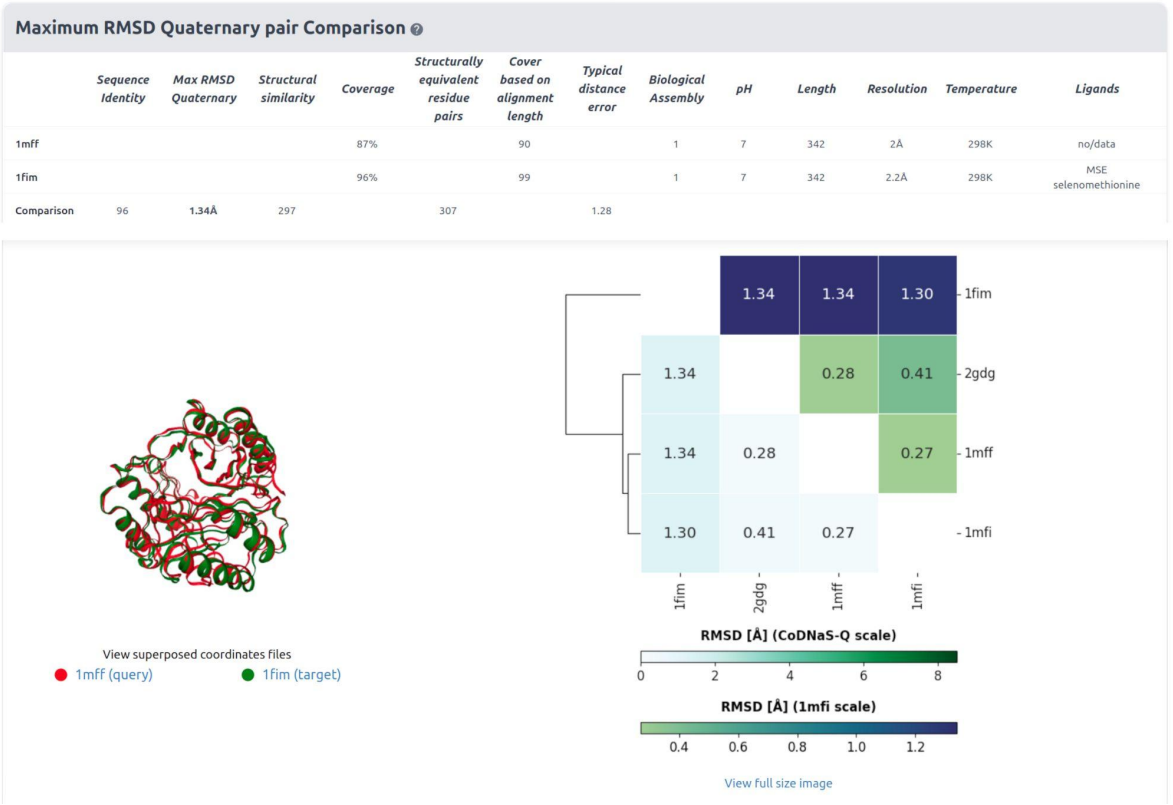


Figure 11. Top sections of a cluster entry page of CoDNaS-Q, providing general information about the cluster (**A**). Below, it gives details on the structural comparison of the pair of conformers with the highest RMSD between them, followed by their structural superposition (bottom left) and a heatmap and dendrogram of all pairwise RMSD comparisons (bottom right) (**B**).

Comparison of selected conformers

1. The 'Conformers' section of a cluster page allows their structural comparison. The website then shows the section 'Selected Pairs of Conformers'. Each row in its table corresponds to a different pair of conformers and provides (dis)similarity values such as sequence identity, pairwise RMSD and number of structurally equivalent residues.
   Click on a row to display a detailed comparison of the selected pair of conformers.
2. When a pair of conformers is selected, the section 'Comparison of selected Pair of Conformers' provides information about each conformer and their structural comparison.

The layout of this section is identical to the equivalent table presented in 'View a protein entry in CoDNaS-Q' above.

BASIC PROTOCOL 3

Basic protocol title

Exploring conformational diversity in a protein family

Introductory paragraph

This protocol describes how to obtain conformational diversity information for the members of a given protein family using BLAST searches.

Necessary resources

- Hardware

CoDNaS, a text editor and a spreadsheet can be displayed in different devices such as laptops and desktop computers. For CoDNaS an active and stable internet connection is required.

- Software

An up-to-date Web browser, such as Chrome or Firefox, is required. A text editor, like Notepad in Windows or gedit in Linux, is also needed. For data analysis and visualization use a spreadsheet program (like Microsoft Excel or Google Sheets) and/or a programming language like R or python.

- Files

This basic protocol requires a Fasta-formatted sequence of interest to be studied in a family context. We use as example the sequence of the murine catalytic subunit alpha of cAMP-dependent protein kinase, a type of serine/threonine kinase that is expressed in eukaryotic cells (Songyang et al., 1996) (Sample File 1).

Protocol steps with step annotations

1. Open a browser and navigate to the advanced search interface of CoDNaS at http://ufq.unq.edu.ar/codnas/search.php.
2. In the 'Search' page, select the tab 'By protein sequence'.
   Notice that by default, CoDNaS allows searching 'By protein characteristics'.
3. Paste the sequence of interest in the text box and hit the 'Search' button (Figure 12). CoDNaS searches by sequence similarity among all protein sequences in its database using BLAST tools.
   Provide a protein sequence to run BLASTp. If a nucleotide sequence is given, CoDNaS will use BLASTx to translate it before doing a protein similarity search.

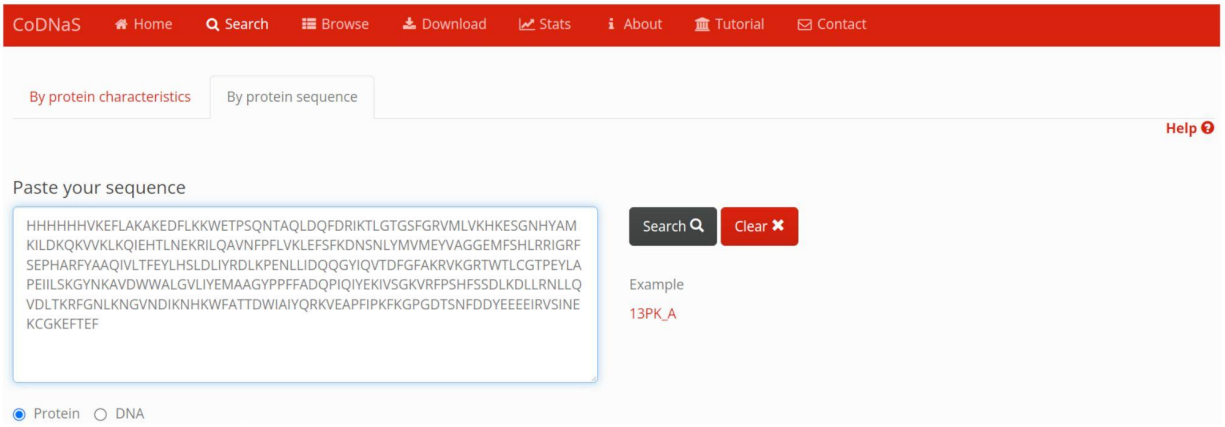


Figure 12. The interface for retrieving CoDNaS entries by running sequence similarity searches using BLAST (UniProt ID P05132) .

4. The main result of the search is a collection of putative homologous proteins detected by BLAST with E-values below 1E-04. From the dropdown menu, select the number of results to display as a table in a single page.

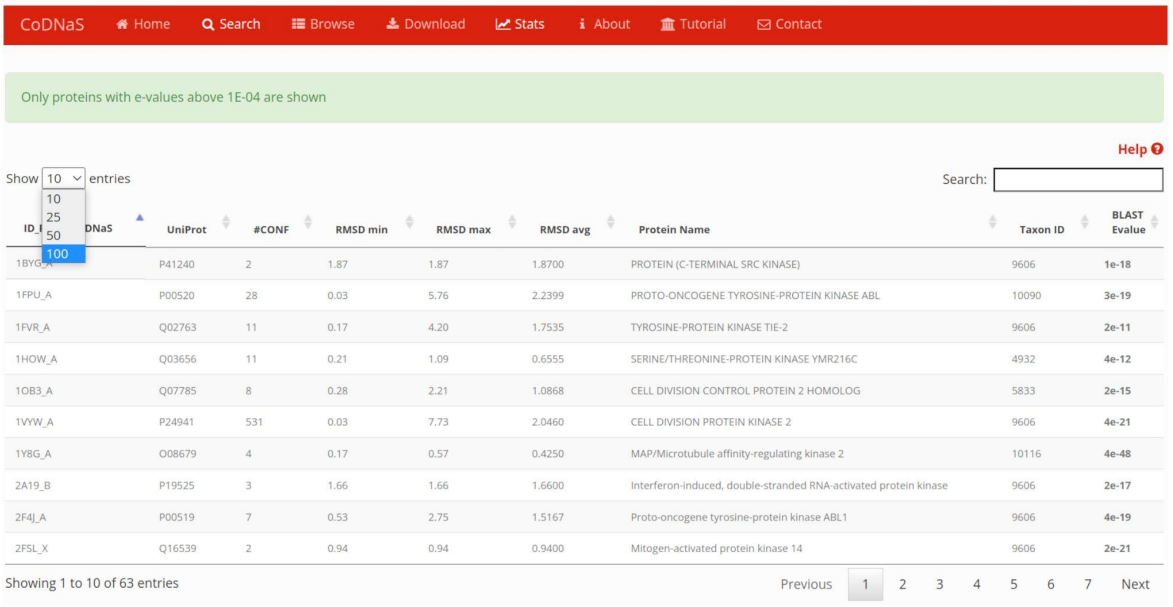


Figure 13. Output of BLAST search showing the putative homologous sequences with known conformational diversity in CoDNaS. The option 'Show 100 entries' is being set in this example.

5. Use the mouse pointer to highlight and copy the rows of interest from the results table.
6. Open a spreadsheet program (Microsoft Excel or LibreOffice Calc, for example) and paste the copied data as a new sheet.
   Alternatively, paste the data into a text editor, save it as a plain text file and import the data into the spreadsheet.
7. Create a box plot of the distribution of maximum conformational diversity per cluster, measured by the maximum RMSD (the column 'RMSD max'), for each of the species (i.e.,

grouping data by the column 'Taxon ID') (Figure 14). This allows exploration of the heterogeneity in conformational diversity across species within the protein family and helps to identify which of those species show the largest diversity.

A box plot is a type of graphic that provides a visual comparison of data distribution in different but related groups. Consider alternative charts like a scatter plot or a density plot to explore the data in different ways.

Instead of the distribution of maximum RMSD values, use the columns 'RMSD min' or 'RMSD avg' to inspect the minimum or the average conformational diversity, respectively, as derived from RMSD values between conformers.

8. Back in the spreadsheet with the BLAST results table (see point 6), select and copy the column of CoDNaS entry identifiers ('ID_POOL_CoDNaS').
9. Go to the 'Download' page in CoDNaS by clicking on the link at the navigation bar at the top of the website or visiting http://ufq.unq.edu.ar/codnas/download.php.

   The 'Download builder' interface is an utility to customize the retrieval of CoDNaS data as a tab-separated file. It allows retrieval of selected information that is presented divided into 6 sections: 'Structural Information', 'Experimental conditions', 'General Information', 'Biological Information' and 'Other'.
10. Paste the CoDNaS identifiers copied before in the 'Paste PDB conformers codes' text box.
11. Tick 'Taxonomy_PDB1_PDB2' in the 'General Information' panel to include the NCBI Taxonomy identifiers (from https://www.ncbi.nlm.nih.gov/taxonomy) of each protein found in the family being studied. In the same panel tick 'Source_PDB1' and 'Source_PDB2' to also retrieve the scientific name of the source species of each conformer. Leave all other checkboxes checked or unchecked as they appear by default.

    Select all other pieces of information that may be of interest for the analysis as desired.
12. Click 'Download' to obtain a file with detailed information of all pairwise structural comparisons between the available conformers in the protein family, including their taxonomic description.
13. Import the downloaded file into your preferred graphing software (e.g. Excel, Calc or the R programming language), splitting fields by tabulators. Create a box plot of the distribution of conformational diversity measured by the RMSD (i.e., the column 'Mammoth_RMS') as a function of the species (either the column 'NCBI taxonomic ID' or any of 'Source_PDB1' or 'Source_PDB2') (Figure 14). This provides a comprehensive overview of the whole extent of conformational diversity observed within the protein family.

    The column 'Taxonomy_PDB1_PDB2' contains taxon IDs for each pair of conformers, separated by an underscore. Split the column before plotting and select one of the resulting columns to avoid repetition.

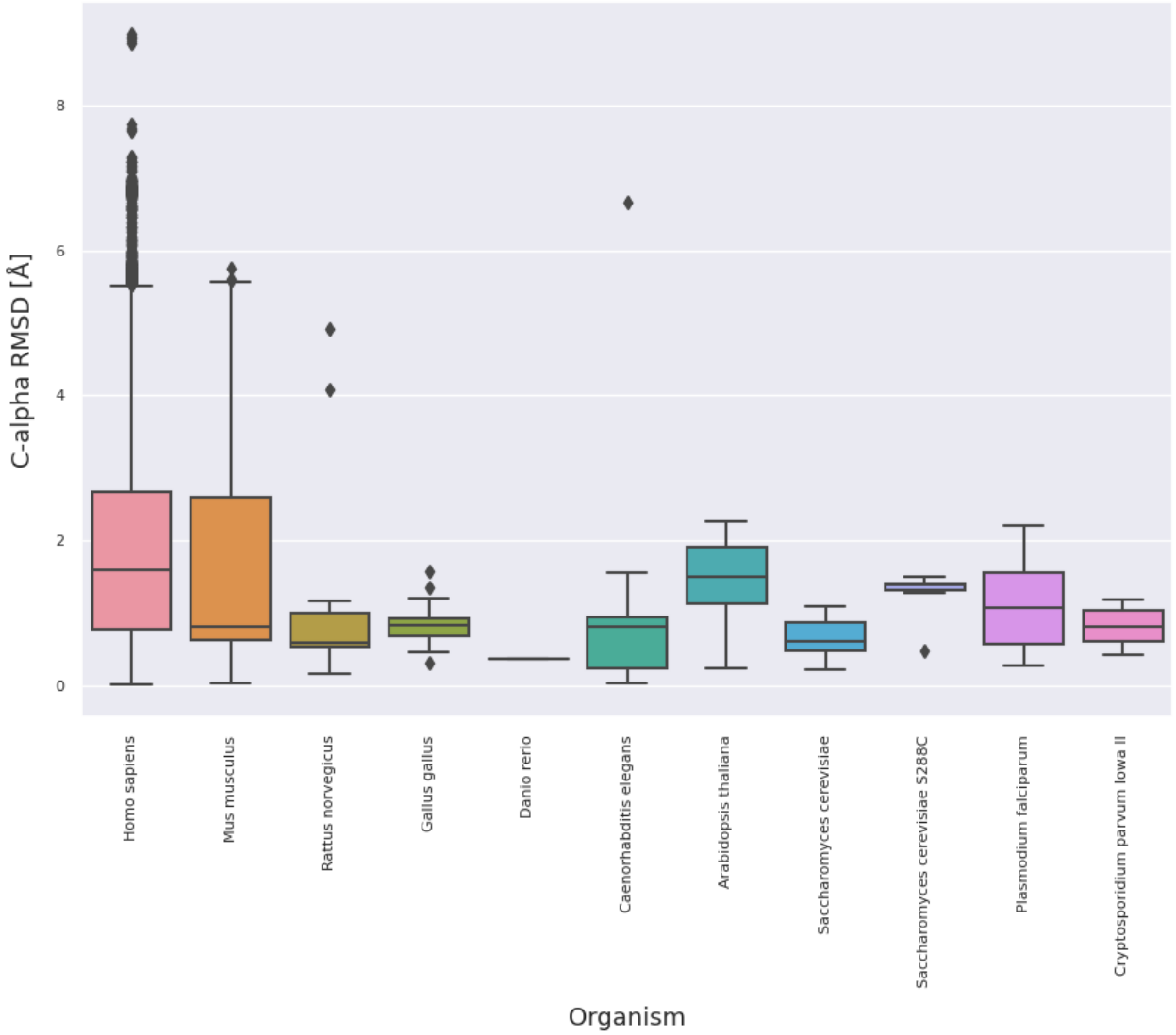


Figure 14. Box plot showing the distribution of conformational diversity measured by the maximum RMSD per species of a given protein family. In this example, the family of a cAMP-dependent protein kinase (catalytic subunit alpha) shows a wide range of conformational diversity among different species. The plot was obtained with Python3 and Seaborn library, the axis labels, graphic style and scientific names have been customized for clarity.

BASIC PROTOCOL 4

Basic protocol title

Representing conformational diversity in a phylogenetic context

Introductory paragraph

This protocol describes how to obtain conformational diversity information for the members of a given protein family in a phylogenetic context. This Basic Protocol will link taxonomic information coming from NCBI Taxonomy Browser with conformational diversity information from CoDNaS. Finally, visualization of a phylogenetic tree representing the taxonomic references obtained, will be displayed using iTOL (the Interactive Tree of Life website).

Necessary Resources

- Hardware

A computer with an active and stable internet connection is required.

- Software

An internet browser, a text editor (as Notepad in Windows or Gedit in Linux) and a spreadsheet processor (such as Microsoft Excel or Google Sheets). Access to the freely available online Taxonomic Browser (https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi) and iTOL Interactive Tree of LIfe (https://itol.embl.de/).

- Files

Tables containing taxonomic identifiers for each protein in a given family and measurements of conformational diversity derived from CoDNaS as shown in Basic Protocol 3.

Protocol steps with step annotations

1. Repeat steps 1-6 of the Basic Protocol 3 to obtain the taxonomic identifiers of putative homologs to a protein of interest (in this example, the cAMP-dependent protein kinase from *Mus musculus*) which have known conformational diversity annotated in CoDNaS.
2. Copy and paste the column of taxonomic identifiers ('Taxon ID') into a new text document (use for example Notepad in Windows or gedit in Linux). Save it as a plain text file.
3. Open an internet browser and visit the Taxonomy Browser at the NCBI website (https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi).
4. Upload the text file with taxonomic identifiers by clicking the button next to 'Add from file:' in the top menu (Figure 15).
5. Click on 'Choose subset' and the Taxonomy Browser will display an expandable tree with taxonomic descriptions of the proteins contained in the family under study.

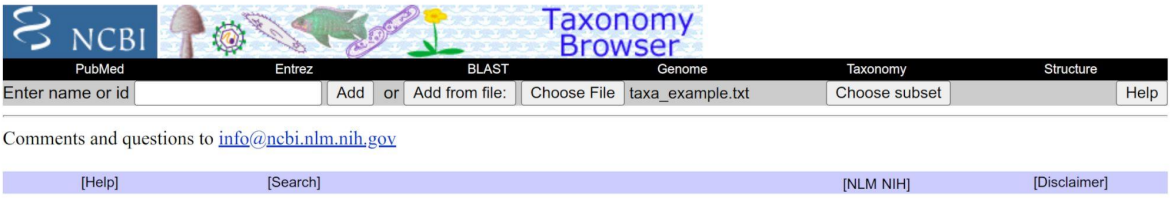


Figure 15. NCBI Taxonomy Browser, useful for obtaining the complete taxonomic lineage annotated for each taxonomic identifier retrieved in Basic Protocol 3.

6. Select the option 'select ranks to show' and the Taxonomy Browser will show different ranks to include in the output (Figure 16). This helps to explore the distribution of conformational diversity in different groups of organisms.
   The default selection will include all the following ranks: Superkingdom, Kingdom, Phylum, Class, Order, Family, Genus and Species.

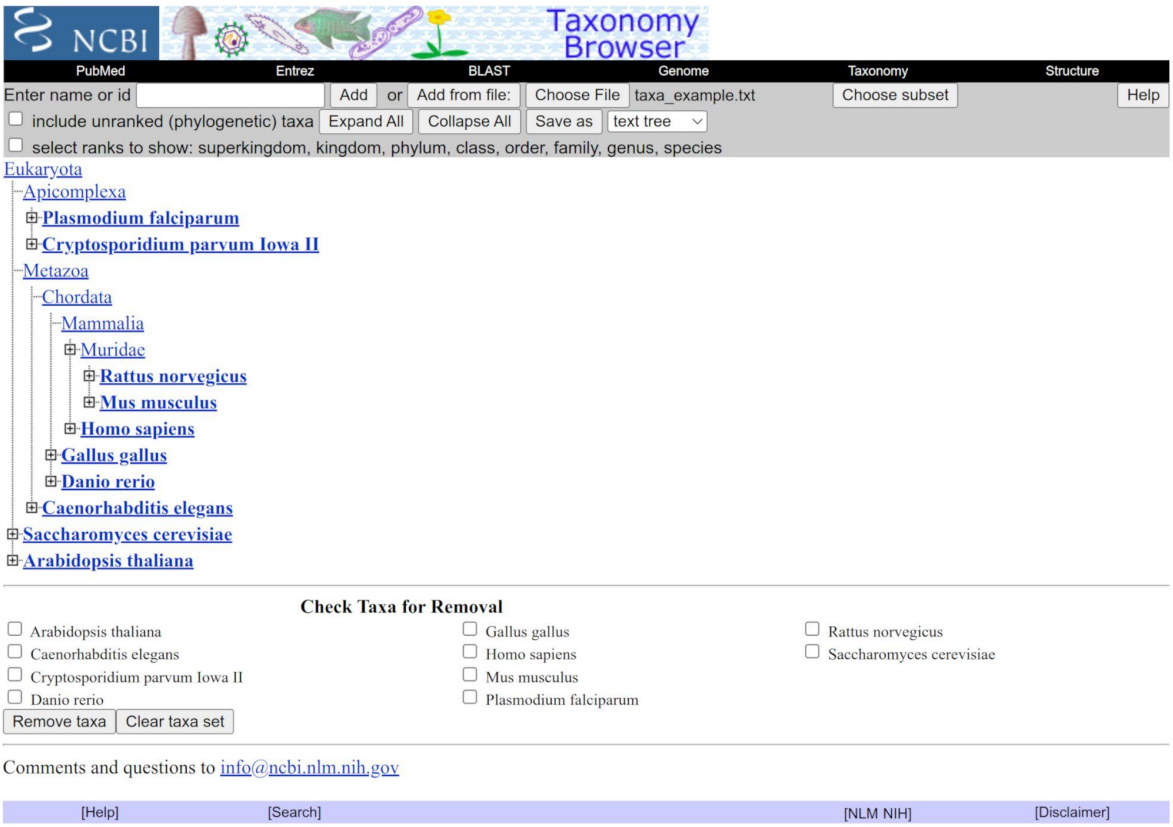


Figure 16. Full taxonomic description of putative homologous proteins to cAMP-dependent protein kinase from *Mus musculus*, retrieved as described in the main text.

7. Once the desired ranks have been chosen, click on the 'Save as' button and choose the option 'phylip tree' to download a text file containing the taxonomic information.
   This file is saved in the Phylip file format, also known as Newick format.
8. Open an internet browser and visit iTOL (Interactive Tree Of Life), an online resource to display and annotate phylogenetic trees, located at https://itol.embl.de/.
9. On the top menu, click on 'Upload' and provide the tree file downloaded from the NCBI Taxonomy browser.
10. The uploaded tree can be represented under different forms. Use the 'Control panel' window on the upper right corner to customize the display (Figure 17).

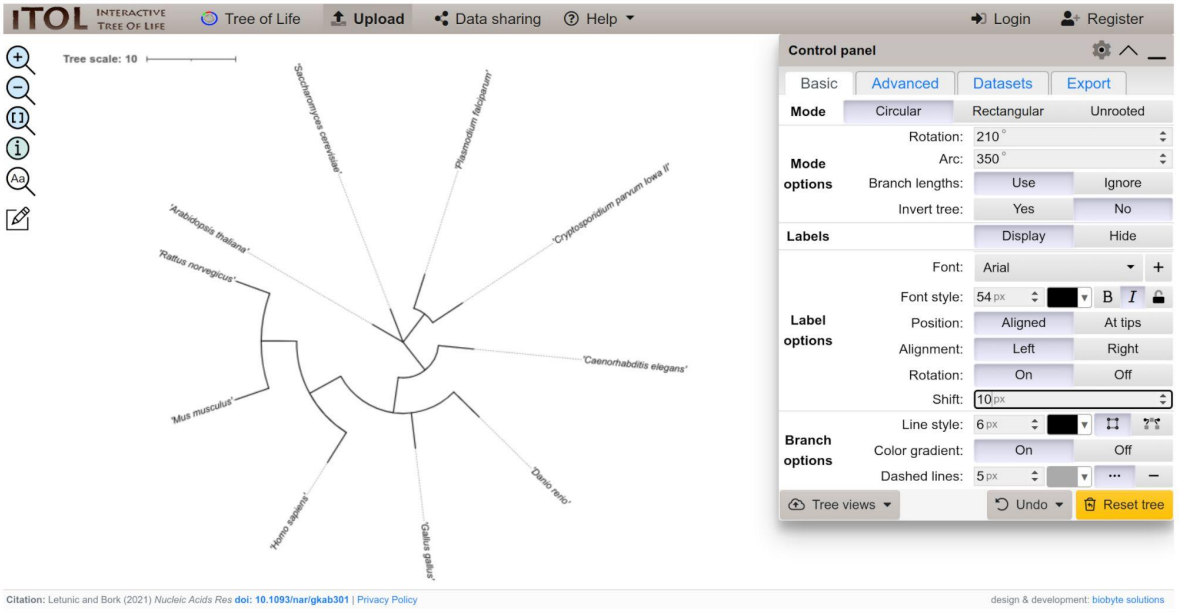


Figure 17. Main window in iTOL for displaying and formatting a phylogenetic tree. Use the 'Control panel' window to change font size and color, among other parameters. The sample tree in the figure is displayed with a circular representation.

11. Using a text editor or spreadsheet processor, create and format a text file as displayed in Sample File 2 (Figure 18), with the data corresponding to the family under study. Save the file to the local computer.

    The text file must be readable for iTOL. contain key commands for iTOL in the first 5 lines: from top to bottom, these lines indicate the type of figure, the character used to separate between fields, the label and color, plus the keyword 'DATA' indicating the beginning of the data to display. The following lines are organized in two columns and show the species name and the maximum RMSD value observed for the species.

```
DATASET_SIMPLEBAR
SEPARATOR COMMA
DATASET_LABEL, RMSDmax
COLOR,#ff0000
DATA
'Arabidopsis thaliana',2.27
'Caenorhabditis elegans',6.67
'Cryptosporidium parvum Iowa II',1.18
'Danio rerio',0.37
'Gallus gallus',1.58
'Homo sapiens',8.98
'Mus musculus',5.76
'Plasmodium falciparum',2.21
'Rattus norvegicus',4.91
'Saccharomyces cerevisiae',1.09
'Saccharomyces cerevisiae S288C',1.50
```

Figure 18. Information on the pair of conformers with maximum RMSD in a CoDNaS entry, related to each studied species. The data in this format can be uploaded into iTOL to visualize the distribution of conformational diversity displayed in a phylogenetic context.

12. In the 'Control panel window' identify the 'Datasets' section (Figure 17). There, upload a text file like Sample File 2 to iTOL. Explore the maximum RMSD values linked to the terminal nodes of the tree to assess the phylogenetic distribution of conformational diversity in the family under study (Figure 19).
    The lengths of the red bars next to the RMSD are proportional to these values. The average RMSD is also available to make the figure.

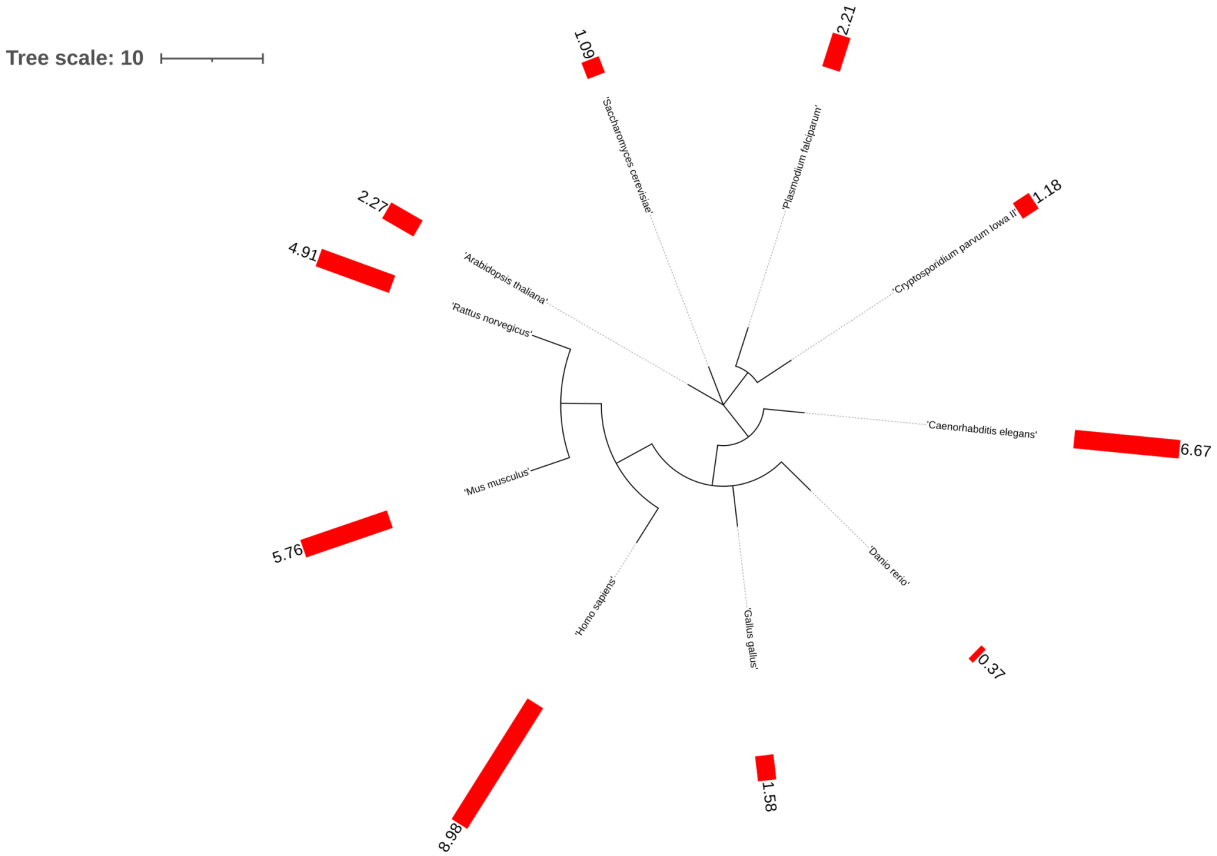


Figure 19. Phylogenetic tree representing the distribution of conformational diversity in the homologous family of the alpha subunit of the cAMP-dependent protein kinase from *Mus musculus*. Red bars and the values next to them indicate the maximum pairwise RMSD derived from CoDNaS for each species in the family.

13. iTOL is a powerful tool to format and enrich tree representations. For example, left-click on the tree to display a menu to highlight colored ranges that can include major clades (Figure 20). This facilitates to identify and focus on a given cluster of organisms (e.g., mammals).

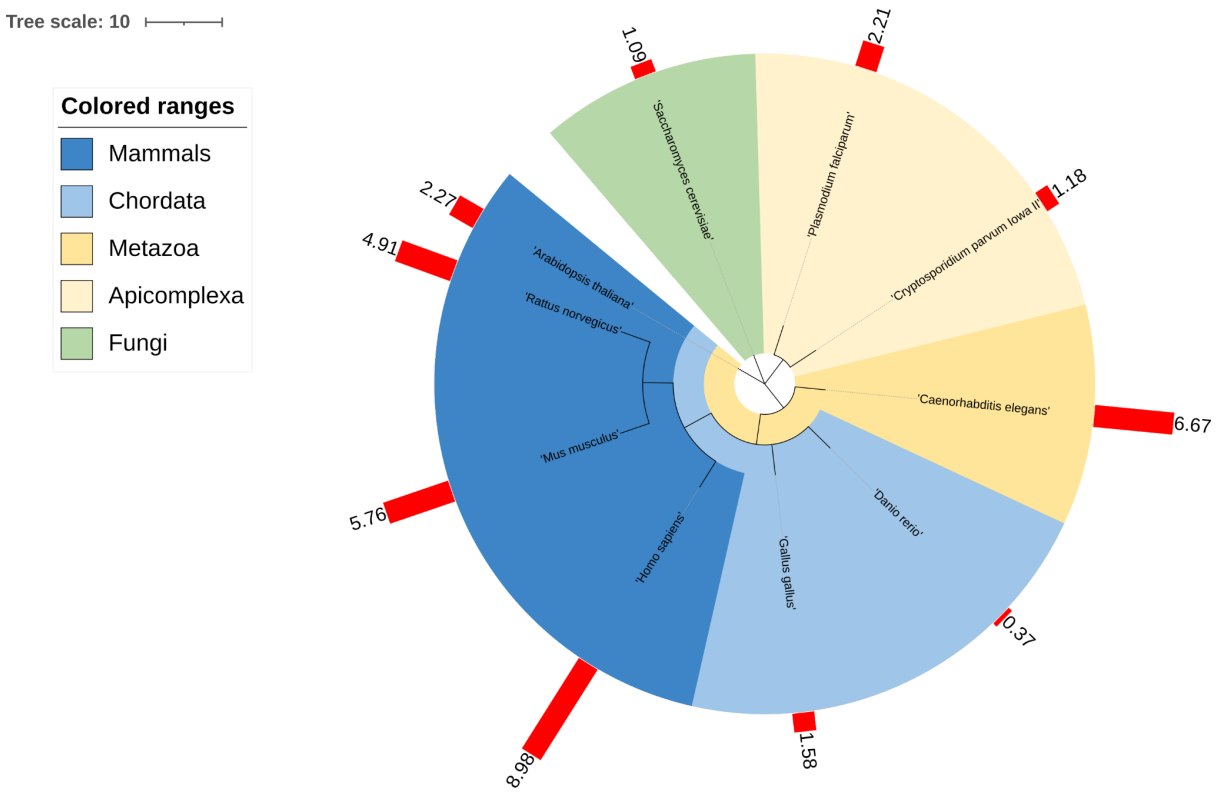


Figure 20. Phylogenetic distribution of maximum pairwise RMSD values (red bars) per protein and per species. The colors in the tree facilitate the identification of different taxonomic ranges. In this example it is clear that the conformational diversity is higher within the cluster of mammals than in the rest of the ranges, with the possible exception of *Caenorhabditis elegans*.

GUIDELINES FOR UNDERSTANDING RESULTS

CoDNaS and CoDNaS-Q are web-based tools that allow the exploration of the conformational diversity of proteins and multimeric complexes using a redundant collection of different structures for the same sequence. For demonstrative purposes, using Basic Protocols 3 and 4 we recognize a wide populated group of CoDNaS clusters defined using CD-HIT at 40% identity and 70% coverage as a threshold. Following this classification, a member of the family of cAMP-dependent protein kinase subunit alpha was chosen, characterized by a high degree of conformational diversity across different species (reaching 8.98Å in humans) (Figure 19 and 20).

Although several scores have been developed to estimate similarities between protein structures (for example GDT (Zemla, 2003) and TMscore (Zhang and Skolnick, 2004)), in CoDNaS and CoDNaS-Q we prefer to use the more classical RMSD value. Contrary to the RMSD, GDT and TMscore were designed to highlight similarities rather than differences. GDT and TMscore, like many other structural similarity scores, optimize the similarities to find the best protein alignments. In CoDNaS and CoDNaS-Q the different entries represent conformers, that is, different forms in which proteins can be arranged by simple rotations or stretchings of bonds. As the sequences of the different conformers are identical, the equivalent residues are already defined and their alignment is trivial. Then, the main result obtained from simple RMSD calculations indicates the true similarity between two different conformers.

In order to better understand the RMSD values obtained it is wise to analyze them in a relative way rather than in an absolute way. Distributions of conformational diversity measured using RMSD in known protein structure space have indicated that most proteins have low RMSD values (Monzon et al., 2017; Burra et al., 2009). A high RMSD observed in a globular protein in a phylogenetic context as described above, can indicate an interesting example of a functional adaptation during evolution. An RMSD of up to 1 Å can be found in "rigid" proteins as their backbones are almost identical (Monzon et al., 2017). Higher RMSD values showing monotonically increasing structural differences between conformers. However, extremely high RMSD values require a careful analysis as they might be coming from proteins containing highly flexible or disordered regions.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Data is available as the article Supplementary Material.

ACKNOWLEDGEMENTS

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 programme under grant agreement No 778247 and No 823886. NP, MSF and GP are researchers and NE is a PhD fellow from CONICET. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

Nahuel Escobedo: formal analysis; general writing and edition. **Alexander Monzon:** software; manuscript revision. **María Silvina Fornasari:** funding acquisition; manuscript revision. **Nicolas Palopoli:** supervision; manuscript writing. **Gustavo Parisi:** conceptualization; project administration; writing original draft.

BIBLIOGRAPHY

del Alamo, D., Sala, D., Mchaourab, H., and Meiler, J. 2021. Sampling the conformational landscapes of transporters and receptors with AlphaFold2. *BioRxiv*.

Best, R. B., Lindorff-Larsen, K., DePristo, M. A., and Vendruscolo, M. 2006. Relation between native ensembles and experimental structures of proteins. *Proceedings of the National Academy of Sciences of the United States of America* 103:10901–10906.

Boehr, D. D., Nussinov, R., and Wright, P. E. 2009. The role of dynamic conformational ensembles in biomolecular recognition. *Nature Chemical Biology* 5:789–796.

Burra, P. V., Zhang, Y., Godzik, A., and Stec, B. 2009. Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure.

Proceedings of the National Academy of Sciences of the United States of America 106:10505–10510.

Dey, S., Ritchie, D. W., and Levy, E. D. 2018. PDB-wide identification of biological assemblies from conserved quaternary structure geometry. *Nature Methods* 15:67–72.

Escobedo, N., Tunque Cahui, R. R., Caruso, G., Garcia Rios, E., Hirsh, L., Monzon, A. M., Parisi, G., and Palopoli, N. 2022. CoDNaS-Q: a database of conformational diversity of the native state of proteins with quaternary structure. *BioRxiv*.

Gerstein, M., and Krebs, W. 1998. A database of macromolecular motions. *Nucleic Acids Research* 26:4280–4290.

Gerstein, M., Lesk, A. M., and Chothia, C. 1994. Structural mechanisms for domain movements in proteins. *Biochemistry* 33:6739–6749.

Glembo, T. J., Farrell, D. W., Gerek, Z. N., Thorpe, M. F., and Ozkan, S. B. 2012. Collective dynamics differentiates functional divergence in protein evolution. *PLoS Computational Biology* 8:e1002428.

Goh, C.-S., Milburn, D., and Gerstein, M. 2004. Conformational changes associated with protein-protein interactions. *Current Opinion in Structural Biology* 14:104–109.

Hammes, G. G. 2002. Multiple conformational changes in enzyme catalysis. *Biochemistry* 41:8221–8228.

Heo, L., and Feig, M. 2021. Multi-state Modeling of G-protein Coupled Receptors at Experimental Accuracy. *BioRxiv*.

Honaker, M. T., Acchione, M., Sumida, J. P., and Atkins, W. M. 2011. Ensemble perspective for catalytic promiscuity: calorimetric analysis of the active site conformational landscape of a detoxification enzyme. *The Journal of Biological Chemistry* 286:42770–42776.

Irwin, D. M., and Tan, H. 2014. Evolution of glucose utilization: glucokinase and glucokinase regulator protein. *Molecular Phylogenetics and Evolution* 70:195–203.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596:583–589.

Juritz, E., Fornasari, M. S., Martelli, P. L., Fariselli, P., Casadio, R., and Parisi, G. 2012. On the effect of protein conformation diversity in discriminating among neutral and disease related single amino acid substitutions. *BMC Genomics* 13 Suppl 4:S5.

Kamata, K., Mitsuya, M., Nishimura, T., Eiki, J.-I., and Nagata, Y. 2004. Structural basis for allosteric regulation of the monomeric allosteric enzyme human glucokinase. *Structure* 12:429–438.

Kondrashov, D. A., Zhang, W., Aranda, R., Stec, B., and Phillips, G. N. 2008. Sampling of the native conformational ensemble of myoglobin via structures in different crystalline environments. *Proteins* 70:353–362.

Larion, M., Salinas, R. K., Bruschweiler-Li, L., Miller, B. G., and Brüschweiler, R. 2012. Order-disorder transitions govern kinetic cooperativity and allostery of monomeric human glucokinase. *PLoS Biology* 10:e1001452.

Leo-Macias, A., Lopez-Romero, P., Lupyan, D., Zerbino, D., and Ortiz, A. R. 2005a. An analysis of core deformations in protein superfamilies. *Biophysical Journal* 88:1291–1299.

Leo-Macias, A., Lopez-Romero, P., Lupyan, D., Zerbino, D., and Ortiz, A. R. 2005b. Core deformations in protein families: a physical perspective. *Biophysical Chemistry* 115:125–128.

Lin, J.-H. 2011. Accommodating protein flexibility for structure-based drug design. *Current Topics in Medicinal Chemistry* 11:171–178.

Liu, Y., and Bahar, I. 2012. Sequence evolution correlates with structural dynamics. *Molecular Biology and Evolution* 29:2253–2263.

Maguid, S., Fernandez-Alberti, S., Ferrelli, L., and Echave, J. 2005. Exploring the common dynamics of homologous proteins. Application to the globin family. *Biophysical Journal* 89:3–13.

Ma, B., and Nussinov, R. 2010. Enzyme dynamics point to stepwise conformational selection in catalysis. *Current Opinion in Chemical Biology* 14:652–659.

Maguid, S., Fernández-Alberti, S., Parisi, G., and Echave, J. 2006. Evolutionary conservation of protein backbone flexibility. *Journal of Molecular Evolution* 63:448–457.

Marino-Buslje, C., Monzon, A. M., Zea, D. J., Fornasari, M. S., and Parisi, G. 2019. On the dynamical incompleteness of the Protein Data Bank. *Briefings in Bioinformatics* 20:356–359.

Mitchell-White, J. I., Stockner, T., Holliday, N., Briddon, S. J., and Kerr, I. D. 2021. Analysis of sequence divergence in mammalian abcgs predicts a structural network of residues that underlies functional divergence. *International Journal of Molecular Sciences* 22.

Monzon, A. M., Rohr, C. O., Fornasari, M. S., and Parisi, G. 2016. CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state. *Database: the Journal of Biological Databases and Curation* 2016.

Monzon, A. M., Zea, D. J., Fornasari, M. S., Saldaño, T. E., Fernandez-Alberti, S., Tosatto, S. C. E., and Parisi, G. 2017. Conformational diversity analysis reveals three functional mechanisms in proteins. *PLoS Computational Biology* 13:e1005398.

Motlagh, H. N., Wrabl, J. O., Li, J., and Hilser, V. J. 2014. The ensemble nature of allostery. *Nature* 508:331–339.

Narayanan, C., Gagné, D., Reynolds, K. A., and Doucet, N. 2017. Conserved amino acid networks modulate discrete functional properties in an enzyme superfamily. *Scientific Reports* 7:3207.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. 1997. CATH--a hierarchic classification of protein domain structures. *Structure* 5:1093–1108.

Parisi, G., Zea, D. J., Monzon, A. M., and Marino-Buslje, C. 2015. Conformational diversity and the emergence of sequence signatures during evolution. *Current Opinion in Structural Biology* 32:58–65.

Saldaño, T., Escobedo, N., Marchetti, J., Zea, D. J., Mac Donagh, J., Velez Rueda, A. J., Gonik, E., García Melani, A., Novomisky Nechcoff, J., Salas, M. N., et al. 2022. Impact of protein conformational diversity on AlphaFold predictions. *Bioinformatics*.

del Sol, A., Tsai, C.-J., Ma, B., and Nussinov, R. 2009. The origin of allosteric functional modulation: multiple pre-existing pathways. *Structure* 17:1042–1050.

Songyang, Z., Lu, K. P., Kwon, Y. T., Tsai, L. H., Filhol, O., Cochet, C., Brickey, D. A., Soderling, T. R., Bartleson, C., Graves, D. J., et al. 1996. A structural basis for substrate specificities of protein Ser/Thr kinases: primary sequence preference of casein kinases I and II, NIMA, phosphorylase kinase, calmodulin-dependent kinase II, CDK5, and Erk1. *Molecular and Cellular Biology* 16:6486–6493.

Tokuriki, N., and Tawfik, D. S. 2009. Protein dynamism and evolvability. *Science* 324:203–207.

Wei, G., Xi, W., Nussinov, R., and Ma, B. 2016. Protein ensembles: how does nature harness thermodynamic fluctuations for life? the diverse functional roles of conformational ensembles in the cell. *Chemical Reviews* 116:6516–6551.

Whittington, A. C., Larion, M., Bowler, J. M., Ramsey, K. M., Brüschweiler, R., and Miller, B. G. 2015. Dual allosteric activation mechanisms in monomeric human glucokinase. *Proceedings of the National Academy of Sciences of the United States of America* 112:11553–11558.

Wiederstein, M., and Sippl, M. J. 2020. TopMatch-web: pairwise matching of large assemblies of protein and nucleic acid chains in 3D. *Nucleic Acids Research* 48:W31–W35.

Zea, D. J., Miguel Monzon, A., Fornasari, M. S., Marino-Buslje, C., and Parisi, G. 2013. Protein conformational diversity correlates with evolutionary rate. *Molecular Biology and Evolution* 30:1500–1503.

Zemla, A. 2003. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Research* 31:3370–3374.

Zhang, Y., and Skolnick, J. 2004. Scoring function for automated assessment of protein structure template quality. *Proteins* 57:702–710.