Combining Protein Conformational Diversity and Phylogenetic Information Using CoDNaS and CoDNaS-Q

Nahuel Escobedo,[1,2,4] Alexander Miguel Monzon,[3,4]
María Silvina Fornasari,[1,2] Nicolas Palopoli,[1,2,5] and Gustavo Parisi[1,2,5]

[1]Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Bernal, Argentina
[2]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina
[3]Department of Information Engineering University of Padova, Padova, Italy
[4]These authors contributed equally to this work.
[5]Corresponding authors: *nicopalo@gmail.com*; *gusparisi@gmail.com*

Published in the Bioinformatics section

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 by machine learning approaches, 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 five 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. © 2023 Wiley Periodicals LLC.

Basic Protocol 1: Assessing conformational diversity with CoDNaS
Alternate Protocol 1: Assessing conformational diversity at the quaternary level with CoDNaS-Q
Basic Protocol 2: Exploring conformational diversity in a protein family
Alternate Protocol 2: Exploring quaternary conformational diversity in a protein family
Basic Protocol 3: Representing conformational diversity in a phylogenetic context

Keywords: conformational diversity • databases • phylogeny • protein family • quaternary structure

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 (del Sol et al., 2009; Motlagh et al., 2014), protein-protein interactions (Goh et al., 2004), protein signaling and information propagation (Boehr et al., 2009; Wei et al., 2016), enzyme catalysis (Hammes, 2002; Ma & Nussinov, 2010), promiscuity (Honaker et al., 2011), sequence substitution pattern during evolution (Parisi et al., 2015; Zea et al., 2013), protein evolvability (Tokuriki & 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., 2017a). 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 & Krebs, 1998; Gerstein et al., 1994; Parisi et al., 2015).

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 ∼30,000 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 comparing collections of alternative crystallographic and NMR structures (Best et al., 2006; Kondrashov et al., 2008).

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., 2005a; Leo-Macias et al., 2005b; 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 (Glembo et al., 2012; Liu & Bahar, 2012; Mitchell-White et al., 2021; Narayanan et al., 2017). 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 & Tan, 2014). Glucokinase (hexokinase IV or D) is a cytoplasmic enzyme that phosphorylates glucose in the liver and pancreas, maintaining glucose homeostasis in mammals. Its unique kinetic behavior contrasts with the rest of the isoenzymes present in other tissues. Glucokinase activity relies on 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 mechanistic explanations at the structural level. 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 (Heo & 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). Derived from this basic protocol, we also provide an alternative way to explore conformational diversity using the quaternary structure of proteins with the information contained in CoDNaS-Q (Alternate Protocol 1). We also cover how to use this information to analyze the conformational diversity distribution in a protein family (Basic Protocol 2 and Alternate Protocol 2 to explore conformational diversity at the tertiary and quaternary levels, respectively) and display it into a phylogenetic context (Basic Protocol 3).

ASSESSING CONFORMATIONAL DIVERSITY WITH CoDNaS

BASIC PROTOCOL 1

CoDNaS contains clusters with a redundant collection of alternative protein structures for the same protein, for example, apo and holo forms of the protein, containing post-translational modifications, or structures obtained at different experimental pH (Monzon et al., 2016, 2013). Within each cluster, all protein structures share identical or highly similar sequences (to account for mutations) and can collectively be taken as snapshots representing the conformational ensemble of the protein (Best et al., 2006). For each protein in CoDNaS, an extensive structural comparison is provided among all conformers. The comparison includes different structural similarity scores, including RMSD (Root Mean Square Deviation), TM-score (Template Modeling score) (Zhang & Skolnick, 2004), and GDT (Global Distance Test) presented as GDT-TS (total score) (Zemla, 2003) and GDT-HA (high accuracy) (Cozzetto et al., 2009). TM-score, GDT-TS, and GDT-HA were developed to highlight structural similarities between protein structures, using different criteria to detect structurally equivalent residues. As protein structures are generally more conserved than sequences during evolution (Chothia & Lesk, 1986; Monzon et al., 2017b), detecting regions of structural similarity could be convenient to obtain reliable alignments. However, the alignment between conformers is straightforward because their sequences are identical. Unlike GDT variations, RMSD uses all the positions of the proteins being compared, in general after a sequence alignment, to define equivalent residues. While TM-score assesses similarity between global topologies, RMSD can better capture local variations. Therefore, in most of the analyses using conformers, the use of RMSD is preferred. CoDNaS also offers a side-by-side display of the crystallization conditions for each conformer (ligands, temperature, presence of disorder, solvent accessibility, pH) to provide mechanistic causes of the conformational diversity, measured using the scores mentioned above, observed between a pair of conformers. 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.

Escobedo et al.

Figure 1 Search options in CoDNaS. Basic search on home page **(A)**. Search by protein characteristics **(B)**. Search by protein or DNA sequence **(C)**.

Necessary resources

Hardware

CoDNaS can be displayed on 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/*).

Searching and browsing

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 press the 'Search' button (Fig. 1A).

   Alternatively, use the example terms (Example 1, Example 2, or Example 3) provided below the text box.

3. To search through more fields you can use the advanced search interface by clicking on the ꞌSearchꞌ 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 (Fig. 1B) or by protein sequence (Fig. 1C). 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 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.

4. If you are going to use a protein sequence as input, paste a protein (as plain text or in FastA format) in the text box (Fig. 1C). 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.

5. Another alternative is to browse CoDNaS according to the CATH hierarchy (Orengo et al., 1997). If you prefer this option, 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.

6. 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.
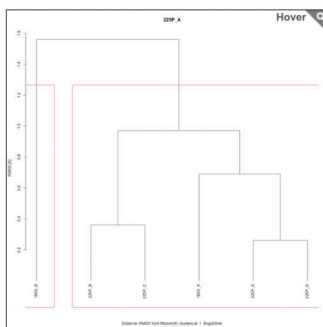
   Users can download the information in JSON or TSV (tab-separated values) formats by using the links in the top-right corner of this section.

   The website displays this section by default. To expand the contents of the following sections, click on the ꞌplusꞌ sign next to the section title.
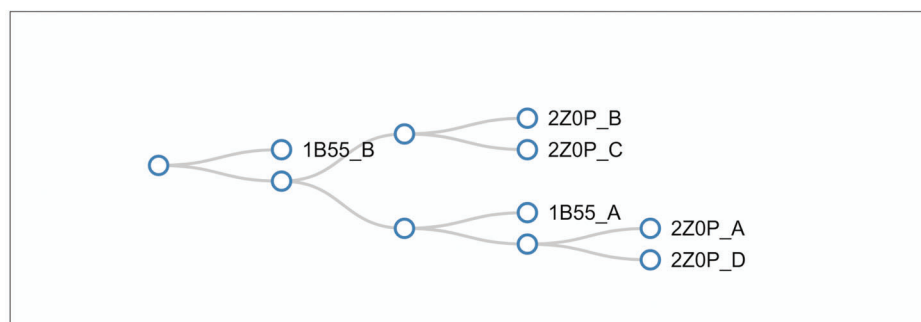
Escobedo et al.

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

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 (Fig. 2A) and an interactive flexible force-directed graph layout (Fig. 2B) 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.

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 (Fig. 3).

   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.

Escobedo et al.

Figure 3 Detailed information about the maximum RMSD pair, i.e. the pair of conformers with maximum RMSD between them, exemplified by cluster 2Z0P_A.

Figure 4 List of conformers that are available to compare all-vs-all.
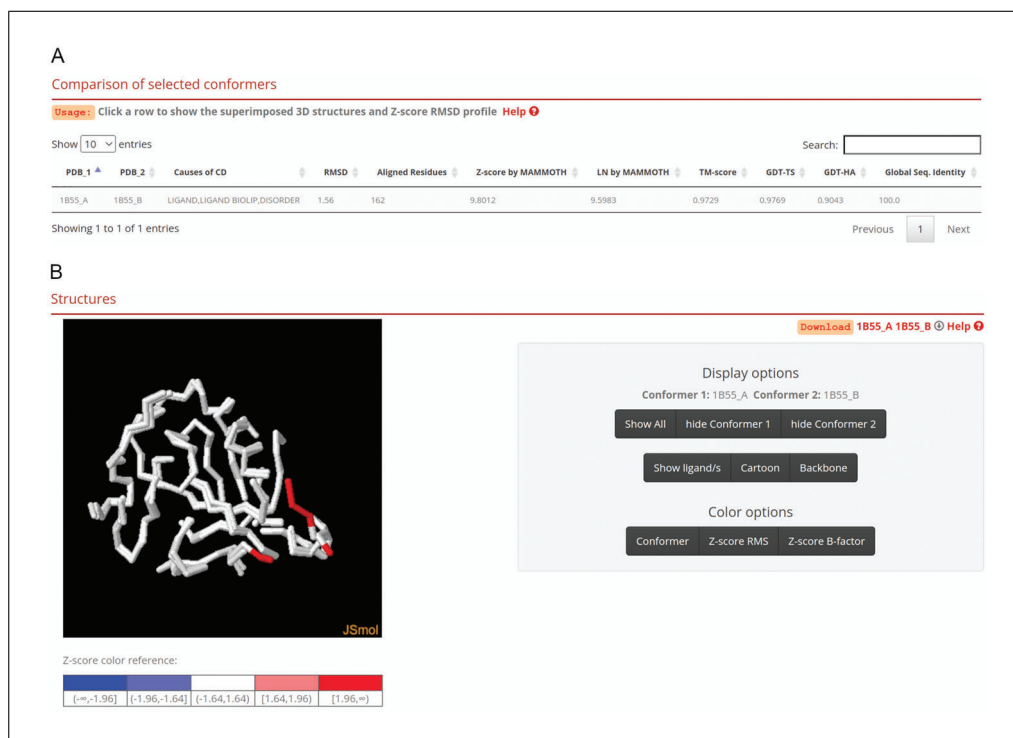
Comparison of selected conformers

1. As previously mentioned, the 'Conformers' section of a CoDNaS entry page contains individual information about each conformer and allows their structural comparison (Fig. 4). 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.

2. The section 'Comparison of selected conformers' provides a table of the selected structural comparisons between conformers (Fig. 5A). The table presents (dis)similarity measures like the pairwise RMSD, the number of aligned residues, TM-score, GDT-TS/GDT-HA values, etc. (see Guidelines for Understanding Results for more details about these metrics).

   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 (Fig. 5B). On the right side are different 'Display' and 'Color Options' available to customize the view.

Escobedo et al.

Figure 5 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)**.

Click the 'Z-score RMSD' button to see the structures colored according to the extent of C-alpha RMS by position, normalized in Z distribution. The Z-score is a way to measure the deviation of a value from the average of the population and is calculated as the difference of the RMSD with respect to the average of the RMSD for all the positions, normalized with the standard deviation. 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.

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.

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 (Fig. 6A).

   Hover over 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 (Fig. 6B). *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.

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).

Figure 6 Z-score profile by position **(A)**. Sequences alignment **(B)**.

> *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.*

ASSESSING CONFORMATIONAL DIVERSITY AT THE QUATERNARY LEVEL WITH CoDNaS-Q

Most protein functions are carried out by complexes of two or more polypeptides (Nooren & Thornton, 2003a, 2003b). This aspect is represented in structural databases, where a larger proportion of the entries were reported to constitute homooligomers (Bertoni et al., 2017; Levy et al., 2006). Conformational diversity is also functionally relevant at the quaternary structure level. The conformational diversity of protein oligomers is mainly associated with the flexibility of the monomers (Marsh et al., 2012) and the structural constraints on solvent accessibility and contact interface imposed along evolution on different quaternary topologies (Marsh & Teichmann, 2014). We have developed CoDNaS-Q to better understand conformational diversity at all structural levels. This database follows the general approach of CoDNaS, but lists clusters of conformers of homooligomeric proteins. Their oligomeric states were assigned with high confidence by QSbio using a combination of several prediction methods (Dey et al.,

Escobedo et al.

9 of 27

2018). CoDNaS-Q is freely accessible as a website at *http://ufq.unq.edu.ar/codnasq/* or *https://codnas-q.bioinformatica.org/home*.

This alternative protocol briefly describes how to perform basic and advanced searches for conformational diversity of homooligomeric proteins in CoDNaS-Q, explore the results, and download useful data.

Necessary resources

Hardware

CoDNaS-Q can be displayed on 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/*).

Searching and browsing

1. Open your web browser and navigate to the CoDNaS-Q landing page at *http://ufq.unq.edu.ar/codnasq/* or *https://codnas-q.bioinformatica.org/home*.

   The URL http://ufq.unq.edu.ar/codnasq/ redirects users to https://codnas-q.bioinformatica.org within Bioinformatica.org, a portal listing different bioinformatics resources that currently hosts CoDNaS-Q.

2. Perform a search by first entering the query text in the input box and then selecting the correct search field (Fig. 7A). Press the 'Search' button to access the 'Advanced Search' page with the results.

   The website provides search term examples below the text box.

   Click on the animated images depicting 'Tertiary Deformations', 'Mixed Motions', or 'Rigid Body' to search by the type of tertiary and quaternary movements that characterize the proteins, as described in https://codnas-q.bioinformatica.org/about.

3. For a more specific search, 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. Use the panels on the left to refine the search by selected characteristics of the cluster (Fig. 7B) or the individual conformers (Fig. 7C). After entering the parameters, click the blue 'Search' button to get the results on the left panel.

4. 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 (Fig. 7D). Click on a card to open the respective cluster entry page.

   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.

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

Figure 7 CoDNaS-Q webpage. The dropdown menu lists different search criteria on the home page, e.g. ˈOrganismˈ **(A)**. The left panel of the ˈAdvanced Searchˈ page of CoDNaS-Q, allows searching for ˈCluster Propertiesˈ **(B)** or ˈConformer Propertiesˈ **(C)**. The red arrow indicates the dropdown menu to search by matching one (ˈORˈ) or all (ˈANDˈ) the terms. The right panel of the ˈAdvanced Searchˈ page of CoDNaS-Q shows the first two cards retrieved as results from a sample search **(D)**.

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 (Fig. 8A).

The relationship between the maximum pairwise RMSD at the tertiary level (ˈmaxRMSD-Tˈ) and the quaternary level (ˈmaxRMSD-Qˈ) 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 ˈMaximum RMSD Quaternary pair Comparisonˈ describes the dissimilarities between the most different conformers of the cluster, as determined by their maximum pairwise RMSD (Fig. 8B). The table on top summarizes precalculated data of their structural comparison obtained with TopMatch (Wiederstein & Sippl, 2020). Below, the website shows an interactive superposition of both conformers.

The table 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.
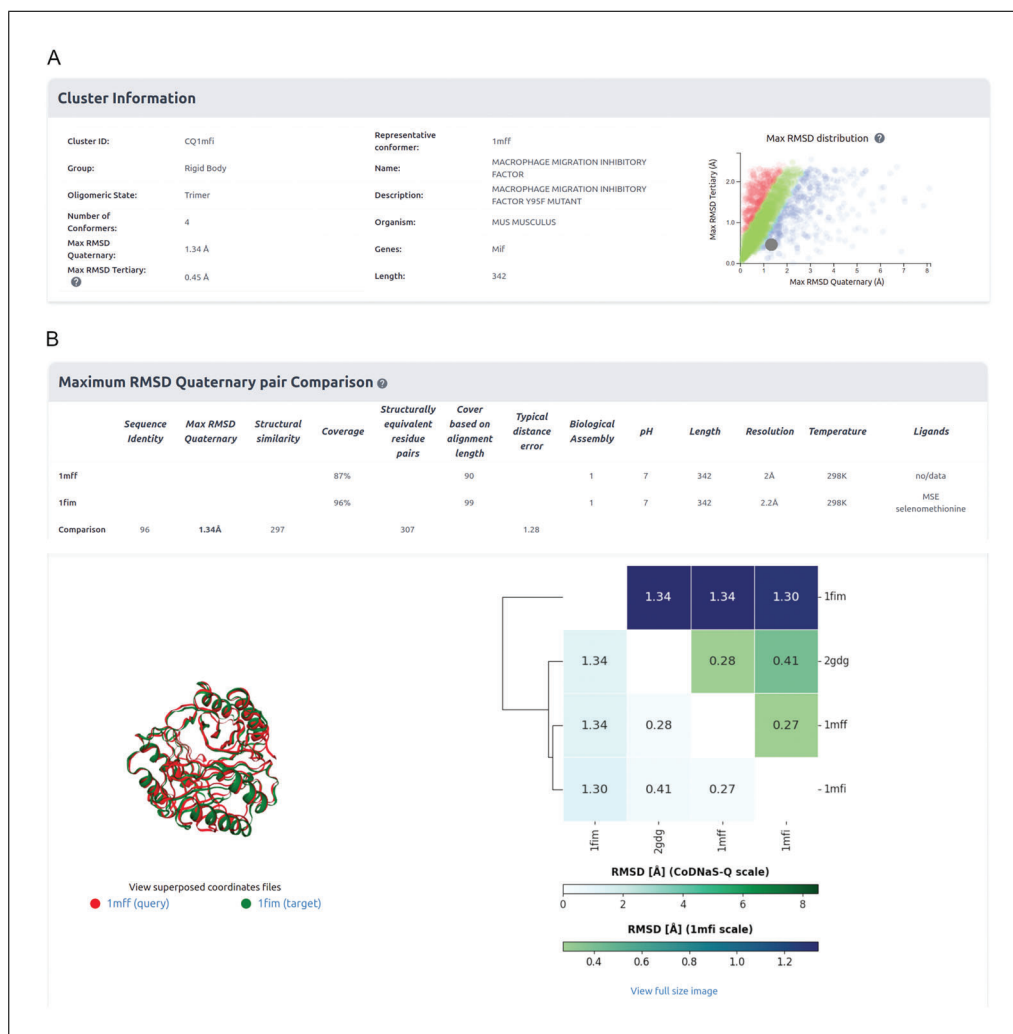
Figure 8 Top sections of a cluster entry page of CoDNaS-Q, providing general information about the cluster **(A)**. Below, it gives details on the structural comparison of the pair of conformers with the highest RMSD between them, followed by their structural superposition (bottom left) and a heatmap and dendrogram of all pairwise RMSD comparisons (bottom right) **(B)**.

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. It also allows structural comparison of any selected pair of conformers.

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.

EXPLORING CONFORMATIONAL DIVERSITY IN A PROTEIN FAMILY

This protocol describes how to obtain conformational diversity information for the members of a given protein family using BLAST searches. Using this resource, the user can search within the proteins stored in CoDNaS database to detect possible members of a

family for the desired query protein, and associate a given conformational diversity (measured as the maximum RMSD, as explained in Basic Protocol 1), to each found species. This procedure will open the possibility of studying the conformational variation of homologous proteins present in different organisms. As conformational diversity is closely related with protein function, it is useful to explore conformational diversity in an evolutionary context both to infer structural and dynamical properties (Bertoni et al., 2017; Odokonyero et al., 2014) as well as functional behavior and the origin of new protein functions (Campbell et al., 2016; Fornasari et al., 2004; Gilson et al., 2017; Jiao et al., 2017).

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).

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 (Fig. 9). 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 9 The interface for retrieving CoDNaS entries by running sequence similarity searches using BLAST (UniProt ID P05132).

Figure 10 Output of BLAST search showing the putative homologous sequences with known conformational diversity in CoDNaS. The option ˈShow 100 entriesˈ is being set in this example.

4. The main result of the search is a collection of putative homologous proteins detected by BLAST with E-values below 1E-04 (Fig. 10). From the dropdown menu, select the number of results to display as a table in a single page.

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) 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. In the spreadsheet with the BLAST results table, select and copy the column of CoDNaS entry identifiers (ˈID_POOL_CoDNaSˈ).

8. Go to the ˈDownloadˈ 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 ˈDownload builderˈ 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: ˈStructural Informationˈ, ˈExperimental conditionsˈ, ˈGeneral Informationˈ, ˈBiological Informationˈ, and ˈOtherˈ.

9. Paste the CoDNaS identifiers copied before in the ˈPaste PDB conformers codesˈ text box.

10. Tick ˈTaxonomy_PDB1_PDB2ˈ in the ˈGeneral Informationˈ 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 ˈSource_PDB1ˈ and ˈSource_PDB2ˈ 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.

11. Click ˈDownloadˈ to obtain a file with the detailed information of all pairwise structural comparisons between the available conformers in the protein family, including their taxonomic description.

Escobedo et al.

14 of 27
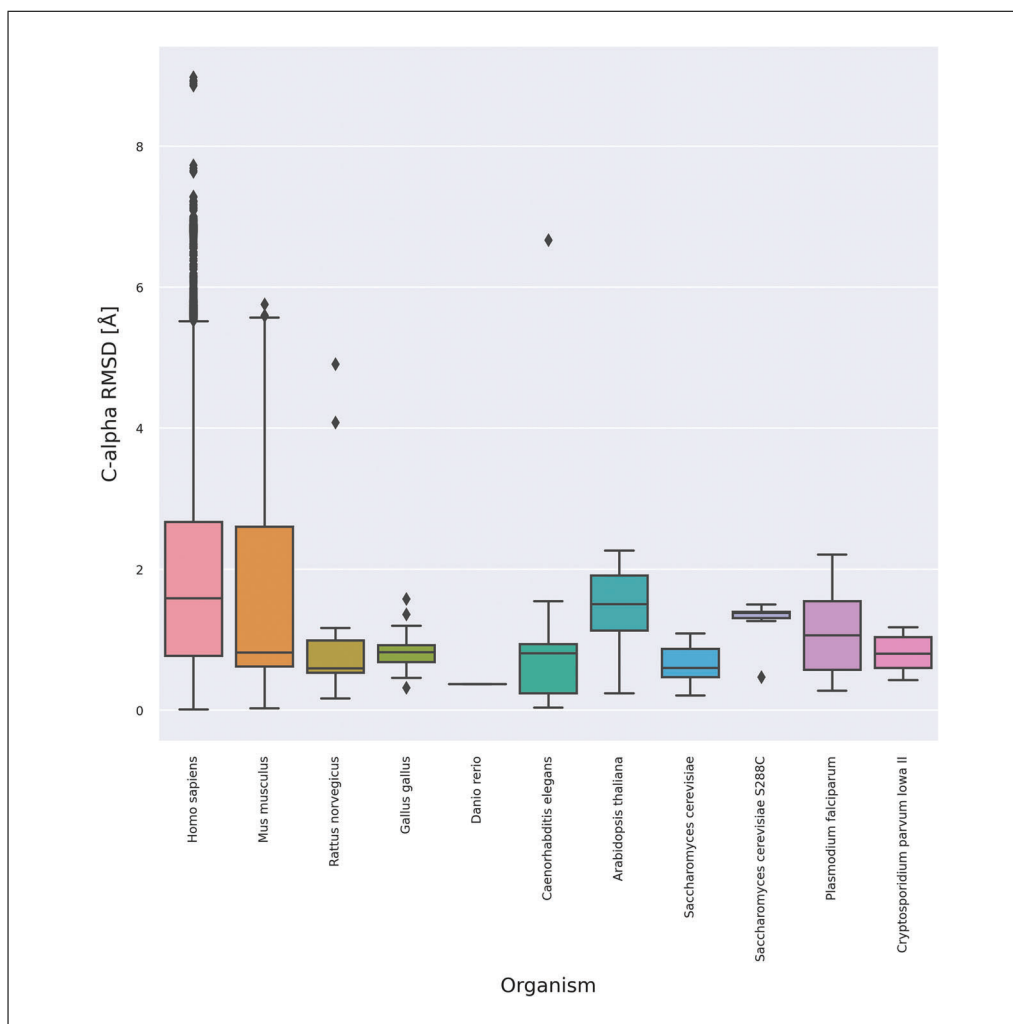
Figure 11 Box plot showing the distribution of conformational diversity ranges measured by the pairwise RMSD calculated between all conformers per protein per species. In this example, the family of a cAMP-dependent protein kinase (catalytic subunit alpha) shows a wide range of conformational diversity degrees across different species. The plot was obtained with Python3 and the Seaborn library. The axis labels, graphic style, and scientific names have been customized for clarity.

12. Import the downloaded file into your preferred graphing software (e.g., Excel, Calc, or the R programming language), splitting fields by tabulators.

13. 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') (Fig. 11). This provides a comprehensive overview of the 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.

   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.

14. Alternatively, determine the maximum RMSD value for each organism from the column 'Mammoth_RMSD'. With these values, plot them by species as in Figure 12.

Escobedo et al.
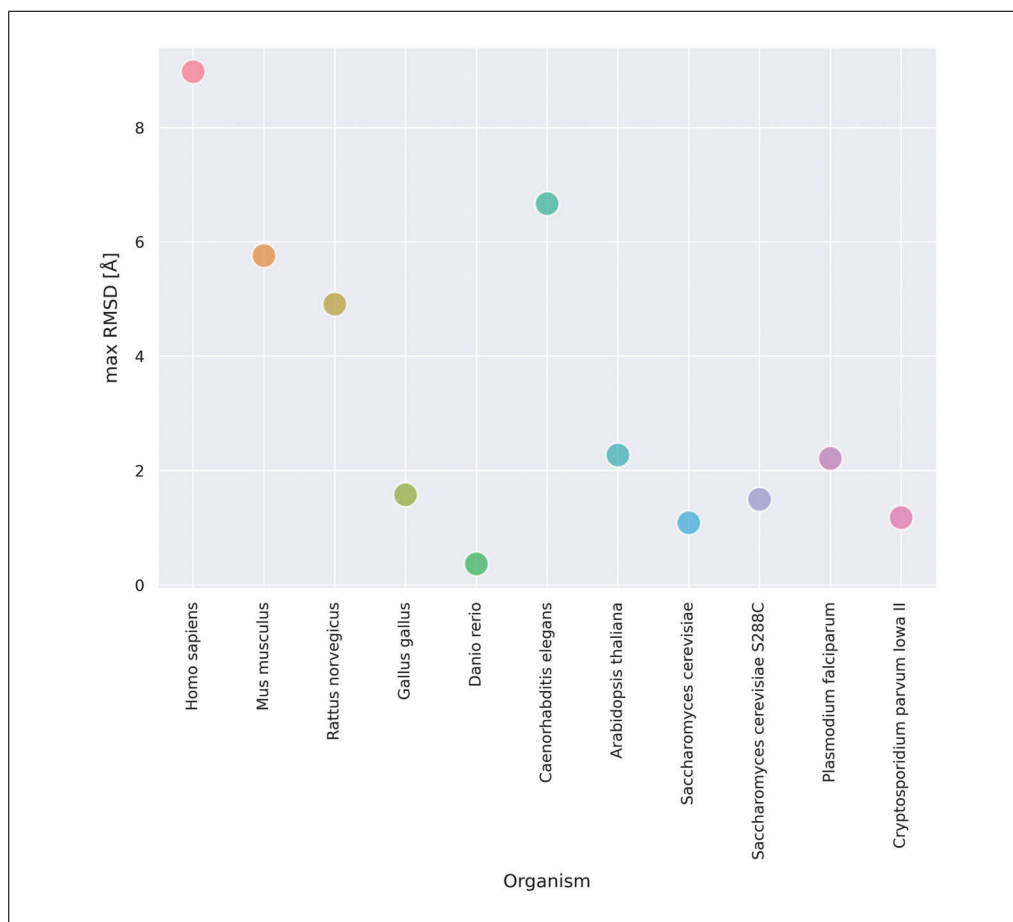
Figure 12 Plot showing the maximum pairwise RMSD values per species.

This allows exploration of the heterogeneity in conformational diversity across the species within the protein family and helps to identify which of those species show the largest diversity.

You can also take the RMSD max values directly from CoDNaS using the table retrieved in step 4 of this protocol. This table can also be used to inspect the minimum ('RMSD min') or the average ('RMSD avg') conformational diversity, as derived from RMSD values between conformers.

ALTERNATE PROTOCOL 2

EXPLORING QUATERNARY CONFORMATIONAL DIVERSITY IN A PROTEIN FAMILY

CoDNaS and CoDNaS-Q facilitate exploration of the conformational diversity of proteins at the tertiary and quaternary level, respectively. In Alternate Protocol 2, we use the conformational diversity at the quaternary level to explore differences in a protein family. It was demonstrated that protein family members can display changes in their quaternary structural arrangements (e.g., lumazine synthase (Fornasari et al., 2004)), or can be represented by different isoforms, built from paralogous genes (like in the case of lactate dehydrogenase (Eventoff et al., 1977), malate dehydrogenase (Madern et al., 2001), or plant carbonic anhydrase (Perales et al., 2004)). Lactate dehydrogenase is an enzyme responsible for the reduction of pyruvate to lactate, yielding L or D-lactate showing homo or heterooligomeric isoforms (Garvie, 1980). Members of this family also show biological behaviors such as cooperativism and allosterism that are linked to specific conformational dynamics (Arai et al., 2010; Iorio et al., 2021; Pasti et al., 2022).

Escobedo et al.

In this protocol we will cover how CoDNaS-Q can provide useful information about the conformational diversity of lactate dehydrogenase, and other proteins with quaternary structure, that is useful to explore biological adaptations in different species.

Necessary resources

Hardware

CoDNaS-Q can be displayed on 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/*).

Searching and browsing

1. Open your web browser and navigate to the CoDNaS-Q landing page at *http://ufq.unq.edu.ar/codnasq/* or *https://codnas-q.bioinformatica.org/home*.

   The URL http://ufq.unq.edu.ar/codnasq/ redirects users to https://codnas-q.bioinformatica.org within Bioinformatica.org, a portal listing different bioinformatics resources that currently hosts CoDNaS-Q.

2. Perform a text search for ˈlactate dehydrogenaseˈ in all fields to collect all the available entries of this enzyme in the database.

   Search in 'All fields' to get the most comprehensive results, or specify which index to query (e.g., Name or Description).

3. The Results page will show 21 different entries, each containing different information about the conformational diversity of the protein as explained in Alternate Protocol 1.

4. Browse the retrieved protein entries and inspect the taxonomic information contained in each of them to build a collection of entries from species of interest. For this example, choose entries for species belonging to bacteria and species belonging to mammals.

5. Inspect the maximum RMSD for the quaternary structure of each protein belonging to bacteria and mammals.

   Select other parameters to explore different aspects of biological relevance.

6. Create a box plot as in Basic Protocol 2 to explore differences between both groups of organisms.

7. Observe that in this case, the average in the maximum RMSD on a quaternary level for mammals is 1.05 Å, while for bacteria it doubles to 2.35 Å. The evidence shows that lactate dehydrogenase oligomers from bacteria are more conformationally diverse than mammalian homologs, suggesting that their flexibility might have reduced along evolution. Additionally, it is possible to explore conformational differences at the tertiary level. In mammals, the average for the maximum RMSD at the tertiary level is 1.20 Å, while for bacteria is 1.40 Å, just slightly above mammals' conformational diversity. Therefore, the individual monomers that combine into lactate dehydrogenases show a significant but similar degree of conformational diversity at the tertiary level that seems conserved along evolution, however they significatively differ at the quaternary level.

Escobedo et al.

17 of 27

REPRESENTING CONFORMATIONAL DIVERSITY IN A PHYLOGENETIC CONTEXT

This protocol describes how to obtain conformational diversity information for the members of a given protein family in a phylogenetic context. Basic Protocol 3 will link taxonomic information coming from NCBI Taxonomy Browser with conformational diversity information from CoDNaS. Finally, the obtained visualization of a phylogenetic tree representing the taxonomic references will be displayed using iTOL (the Interactive Tree of Life website).

What are the main differences between this and the previous protocol? Basic Protocol 2 studies the distribution of conformational diversity per protein in different species (Figs. 11 and 12). However, it is not possible to infer from that analysis how evolutionarily close are the species and therefore, the proteins being compared. As protein structure and protein function conservation show an inverse relationship with sequence divergence, it is then expected that both will be more conserved in homologous proteins in closely related organisms (Chothia & Lesk, 1986; Pascual-García et al., 2010; Reeves et al., 2006; Sangar et al., 2007). Evolutionary relationships between organisms are better represented using phylogenetic trees built with different measures of similarity between organisms (see *https://www.ncbi.nlm.nih.gov/taxonomy* and (Kumar et al., 2022)). Therefore, the phylogenetic approach provides a better representation of how conformational diversity is distributed in a protein family, as described in the present protocol. Moreover, recent machine learning approaches to predict conformational diversity, as in the recent breakthrough of AlphaFold protein structure predictions (Jumper et al., 2021; Tunyasuvunakool et al., 2021), relay in sequence subsampling of protein families to better detect sequence features related with protein dynamics (Del Alamo et al., 2022; Stein & Mchaourab, 2022; Wayment-Steele et al., 2022). Despite those advances, conformational diversity predictions are still difficult (Saldaño et al., 2022). This protocol can possibly help to explore the relationship between conformational diversity in a phylogenetic context and prove to be a useful resource to better detect evolutionary signatures related with different conformations.

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 2.

1. Repeat steps 1 to 7 of the Basic Protocol 2 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.
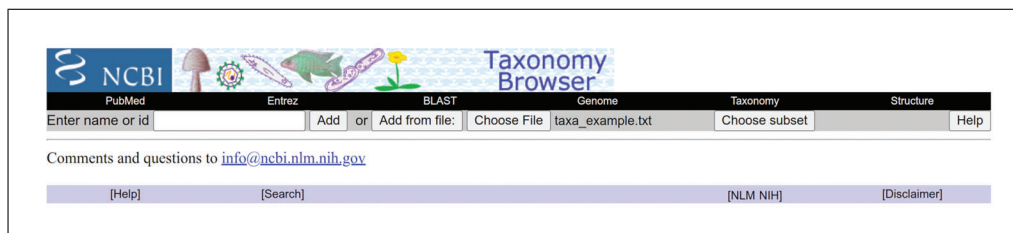
Figure 13   NCBI Taxonomy Browser, useful for obtaining the complete taxonomic lineage annotated for each taxonomic identifier retrieved in Basic Protocol 2.
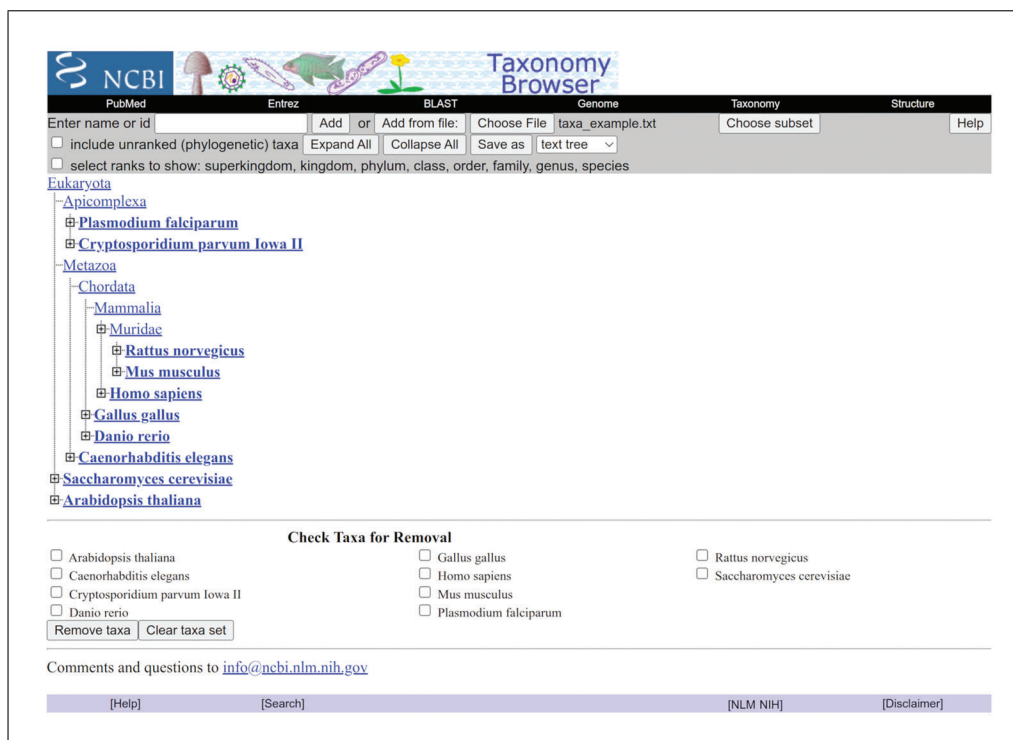
Figure 14   Full taxonomic description of putative homologous proteins to cAMP-dependent protein kinase from *Mus musculus*, retrieved as described in the main text.

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 ʹAdd from fileʹ: in the top menu (Fig. 13).

5. Click on ʹChoose subsetʹ and the Taxonomy Browser will display an expandable tree with taxonomic descriptions of the proteins contained in the family under study.

6. Select the option ʹselect ranks to showʹ and the Taxonomy Browser will show different ranks to include in the output (Fig. 14). 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.

7. Once the desired ranks have been chosen, click on the ʹSave asʹ button and choose the option ʹphylip treeʹ 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/*.

Escobedo et al.
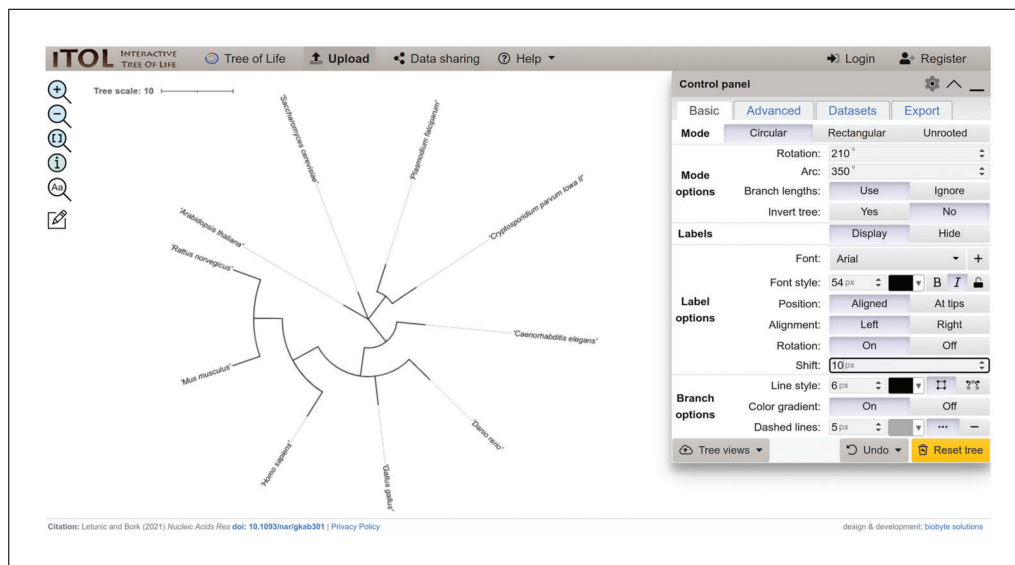
Figure 15 Main window in iTOL for displaying and formatting a phylogenetic tree. Use the 'Control panel' window to change font size and color, among other parameters. The sample tree in the figure is displayed with a circular representation.

```
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 16 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.

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 (Fig. 15).

11. Using a text editor or spreadsheet processor, create and format a text file as displayed in Sample File 2 (Fig. 16), 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
```
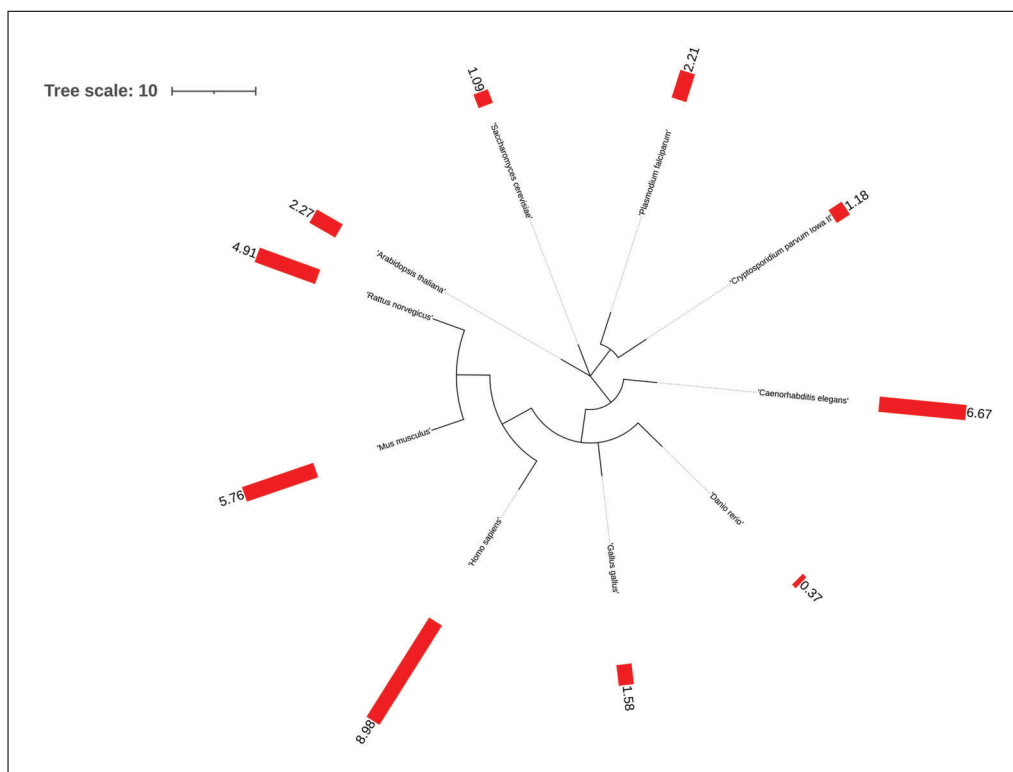
Figure 17 Phylogenetic tree representing the distribution of conformational diversity in the homologous family of the alpha subunit of the cAMP-dependent protein kinase from *Mus musculus*. Red bars and the values next to them indicate the maximum pairwise RMSD derived from CoDNaS for each species in the family.

```
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
```

12. In the 'Control panel window' identify the 'Datasets' section (Fig. 15). 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 (Fig. 17).

    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.

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 (Fig. 18). This facilitates to identify and focus on a given cluster of organisms (e.g., mammals).

Escobedo et al.
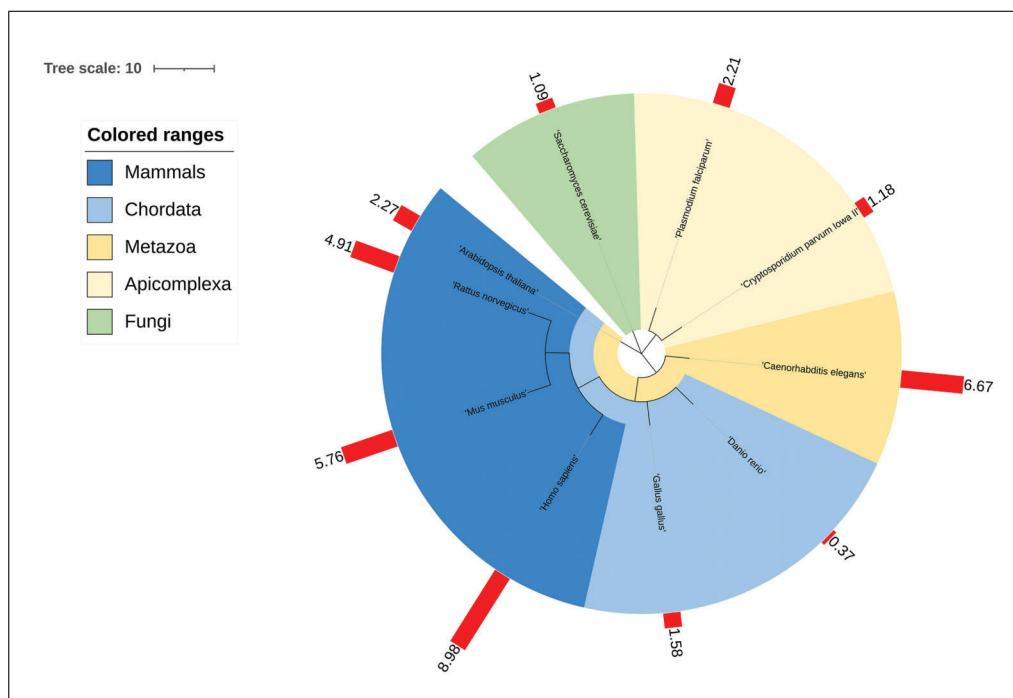
Figure 18 Phylogenetic distribution of maximum pairwise RMSD values (red bars) per protein and per species. The colors in the tree facilitate the identification of different taxonomic ranges. In this example, it is clear that the conformational diversity is higher within the cluster of mammals than in 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. This is an experimental based approach to estimate conformational diversity in contrast to other computational based methods like normal mode analysis or molecular dynamics (Chennubhotla et al., 2005; Hollingsworth & Dror, 2018). Although conformational diversity is a key concept for the understanding of protein biology, its incorporation into bioinformatic tools and protein characterization has been delayed due to several difficulties associated with its estimation (Marino-Buslje et al., 2019; Monzon et al., 2017b; Saldaño et al., 2022). In the coming age of accurate structural prediction, we expect that the presented protocols promote the development of new bioinformatics methods to reach a better understanding of protein biology.

Conformational diversity in both databases is estimated as a measure of the structural differences between the conformers of each protein. To capture these differences, we use several of the scores that have been developed to estimate similarities between protein structures. For example we included in the databases some of the most frequently used, RMSD (Root Mean Square Deviation), the GDT (Global DIstance Test) (Zemla, 2003), or the TM-score (Template Modeling Score) (Zhang & Skolnick, 2004). GDT finds the largest set of residues in common between two structures that fit under a predefined distance threshold to measure the whole structure's similarity.

Alternatively, TM-score was designed as size-independent and indicates with a value between 0 to 1 how similar two structures are (in general, it is accepted that TM-scores lower than 0.2 correspond to random alignments, while TM-scores higher than 0.5 indicate a conserved fold). However, in CoDNaS and CoDNaS-Q we prefer to use the more classical RMSD value. Contrary to the RMSD, GDT and TM-score were designed to highlight similarities rather than differences. GDT and TM-score, 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, which are different forms a protein can adopt 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.

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 (Burra et al., 2009; Monzon, et al., 2017a). 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. A RMSD of up to 1 Å can be found in 'rigid' proteins as their backbones are almost identical (Monzon, et al., 2017a). Higher RMSD values are typical of flexible proteins that change their conformations as a whole, such as those with flexible loops or those with domain movements. When distributions of RMSD were performed in large datasets, the RMSD is observed to monotonically increase up to large values (>15 Å). However, extremely high RMSD values require careful analysis as they might be coming from disordered proteins or with disordered regions or alignments errors.

As described in the introduction, conformational diversity has been associated with a myriad of biological processes sustained by proteins. Their exploration for a given protein (Basic Protocol 1 and Alternate Protocol 1) can provide information about the flexible regions of the protein under study. In general, those movements can be related with protein function, such as those involved in the opening or closing of a binding site, adoption of active conformations by relative domain movements or movements involved in the docking between protein interfaces (Gora et al., 2013; Monzon, et al., 2017a; Qi & Hayward, 2009). Additionally, at the quaternary level, proteins can move as rigid bodies but also showing tertiary deformations opening the possibility to classify proteins with different sort of quaternary structural diversity (Escobedo et al., 2022). The other protocols described (Basic Protocols 2 and 3, and Alternate Protocol 2) offer examples of how to explore variations in the extension of conformational diversity in the taxonomic and phylogenetic context. These procedures allow the user to detect functional adaptations or functional diversification that might have appeared during evolution.

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.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

Data is available in Supplementary Material.

Literature Cited

Arai, K., Ishimitsu, T., Fushinobu, S., Uchikoba, H., Matsuzawa, H., & Taguchi, H. (2010). Active and inactive state structures of unliganded

Escobedo et al.

Lactobacillus casei allosteric L-lactate dehydrogenase. *Proteins*, *78*(3), 681–694. https://doi.org/10.1002/prot.22597

Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L., & Schwede, T. (2017). Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Scientific Reports*, *7*(1), 10480. https://doi.org/10.1038/s41598-017-09654-8

Best, R. B., Lindorff-Larsen, K., DePristo, M. A., & 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*(29), 10901–10906. https://doi.org/10.1073/pnas.0511156103

Boehr, D. D., Nussinov, R., & Wright, P. E. (2009). The role of dynamic conformational ensembles in biomolecular recognition. *Nature Chemical Biology*, *5*(11), 789–796. https://doi.org/10.1038/nchembio.232

Burra, P. V., Zhang, Y., Godzik, A., & 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*(26), 10505–10510.

Campbell, E., Kaltenbach, M., Correy, G. J., Carr, P. D., Porebski, B. T., Livingstone, E. K., Afriat-Jurnou, L., Buckle, A. M., Weik, M., Hollfelder, F., Tokuriki, N., & Jackson, C. J. (2016). The role of protein dynamics in the evolution of new enzyme function. *Nature Chemical Biology*, *12*(11), 944–950. https://doi.org/10.1038/nchembio.2175

Chennubhotla, C., Rader, A. J., Yang, L.-W., & Bahar, I. (2005). Elastic network models for understanding biomolecular machinery: From enzymes to supramolecular assemblies. *Physical Biology*, *2*(4), S173–80. https://doi.org/10.1088/1478-3975/2/4/S12

Chothia, C., & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, *5*(4), 823–826. https://doi.org/10.1002/j.1460-2075.1986.tb04288.x

Cozzetto, D., Kryshtafovych, A., Fidelis, K., Moult, J., Rost, B., & Tramontano, A. (2009). Evaluation of template-based models in CASP8 with standard measures. *Proteins*, *77*(Suppl 9), 18–28. https://doi.org/10.1002/prot.22561

Del Alamo, D., Sala, D., Mchaourab, H. S., & Meiler, J. (2022). Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife*, *11*, e75751. https://doi.org/10.7554/eLife.75751

del Sol, A., Tsai, C.-J., Ma, B., & Nussinov, R. (2009). The origin of allosteric functional modulation: Multiple pre-existing pathways. *Structure*, *17*(8), 1042–1050. https://doi.org/10.1016/j.str.2009.06.008

Dey, S., Ritchie, D. W., & Levy, E. D. (2018). PDB-wide identification of biological assemblies from conserved quaternary structure ge-

ometry. *Nature Methods*, *15*(1), 67–72. https://10.1038/nmeth.4510

Escobedo, N., Tunque Cahui, R. R., Caruso, G., García Ríos, E., Hirsh, L., Monzon, A. M., Parisi, G., & Palopoli, N. (2022). CoDNaS-Q: A database of conformational diversity of the native state of proteins with quaternary structure. *Bioinformatics*, *38*(21), 4959–4961. https://doi.org/10.1093/bioinformatics/btac627

Eventoff, W., Rossmann, M. G., Taylor, S. S., Torff, H. J., Meyer, H., Keil, W., & Kiltz, H. H. (1977). Structural adaptations of lactate dehydrogenase isozymes. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(7), 2677–2681. https://doi.org/10.1073/pnas.74.7.2677

Fornasari, M. S., Laplagne, D. A., Frankel, N., Cauerhff, A. A., Goldbaum, F. A., & Echave, J. (2004). Sequence determinants of quaternary structure in lumazine synthase. *Molecular Biology and Evolution*, *21*(1), 97–107. https://doi.org/10.1093/molbev/msg244

Garvie, E. I. (1980). Bacterial lactate dehydrogenases. *Microbiological Reviews*, *44*(1), 106–139. https://doi.org/10.1128/mr.44.1.106-139.1980

Gerstein, M., & Krebs, W. (1998). A database of macromolecular motions. *Nucleic Acids Research*, *26*(18), 4280–4290. https://doi.org/10.1093/nar/26.18.4280

Gerstein, M., Lesk, A. M., & Chothia, C. (1994). Structural mechanisms for domain movements in proteins. *Biochemistry*, *33*(22), 6739–6749. https://doi.org/10.1021/bi00188a001

Gilson, A. I., Marshall-Christensen, A., Choi, J.-M., & Shakhnovich, E. I. (2017). The role of evolutionary selection in the dynamics of protein structure evolution. *Biophysical Journal*, *112*(7), 1350–1365. https://doi.org/10.1016/j.bpj.2017.02.029

Glembo, T. J., Farrell, D. W., Gerek, Z. N., Thorpe, M. F., & Ozkan, S. B. (2012). Collective dynamics differentiates functional divergence in protein evolution. *PLoS Computational Biology*, *8*(3), e1002428. https://doi.org/10.1371/journal.pcbi.1002428

Goh, C.-S., Milburn, D., & Gerstein, M. (2004). Conformational changes associated with protein-protein interactions. *Current Opinion in Structural Biology*, *14*(1), 104–109. https://doi.org/10.1016/j.sbi.2004.01.005

Gora, A., Brezovsky, J., & Damborsky, J. (2013). Gates of enzymes. *Chemical Reviews*, *113*(8), 5871–5923. https://doi.org/10.1021/cr300384w

Hammes, G. G. (2002). Multiple conformational changes in enzyme catalysis. *Biochemistry*, *41*(26), 8221–8228. https://doi.org/10.1021/bi0260839

Heo, L., & Feig, M. (2021). Multi-state modeling of G-protein coupled receptors at experimental accuracy. *Proteins*, *90*(11), 1873–1885. https://doi.org/10.1002/prot.26382

Hollingsworth, S. A., & Dror, R. O. (2018). Molecular dynamics simulation for all.

Neuron, *99*(6), 1129–1143. https://doi.org/10.1016/j.neuron.2018.08.011

Honaker, M. T., Acchione, M., Sumida, J. P., & 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*(49), 42770–42776. https://doi.org/10.1074/jbc.M111.304386

Iorio, A., Roche, J., Engilberge, S., Coquelle, N., Girard, E., Sterpone, F., & Madern, D. (2021). Biochemical, structural and dynamical studies reveal strong differences in thethermal-dependent allosteric behavior of two extremophilic lactate dehydrogenases. *Journal of Structural Biology*, *213*(3), 107769. https://doi.org/10.1016/j.jsb.2021.107769

Irwin, D. M., & Tan, H. (2014). Evolution of glucose utilization: Glucokinase and glucokinase regulator protein. *Molecular Phylogenetics and Evolution*, *70*, 195–203. https://doi.org/10.1016/j.ympev.2013.09.016

Jiao, W., Blackmore, N. J., Nazmi, A. R., & Parker, E. J. (2017). Quaternary structure is an essential component that contributes to the sophisticated allosteric regulation mechanism in a key enzyme from Mycobacterium tuberculosis. *Plos One*, *12*(6), e0180052. https://doi.org/10.1371/journal.pone.0180052

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Juritz, E., Fornasari, M. S., Martelli, P. L., Fariselli, P., Casadio, R., & 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. https://doi.org/10.1186/1471-2164-13-S4-S5

Kamata, K., Mitsuya, M., Nishimura, T., Eiki, J.-I., & Nagata, Y. (2004). Structural basis for allosteric regulation of the monomeric allosteric enzyme human glucokinase. *Structure*, *12*(3), 429–438. https://doi.org/10.1016/j.str.2004.02.005

Kondrashov, D. A., Zhang, W., Aranda, R., Stec, B., & Phillips, G. N. (2008). Sampling of the native conformational ensemble of myoglobin via structures in different crystalline environments. *Proteins*, *70*(2), 353–362. https://doi.org/10.1002/prot.21499

Kumar, S., Suleski, M., Craig, J. M., Kasprowicz, A. E., Sanderford, M., Li, M., Stecher, G., & Hedges, S. B. (2022). Timetree 5: An expanded resource for species divergence times. *Molecular Biology and Evolution*, *39*(8), msac174. https://doi.org/10.1093/molbev/msac174

Larion, M., Salinas, R. K., Bruschweiler-Li, L., Miller, B. G., & Brüschweiler, R. (2012). Order-disorder transitions govern kinetic cooperativity and allostery of monomeric human glucokinase. *PLoS Biology*, *10*(12), e1001452. https://doi.org/10.1371/journal.pbio.1001452

Leo-Macias, A., Lopez-Romero, P., Lupyan, D., Zerbino, D., & Ortiz, A. R. (2005a). An analysis of core deformations in protein superfamilies. *Biophysical Journal*, *88*(2), 1291–1299. https://doi.org/10.1529/biophysj.104.052449

Leo-Macias, A., Lopez-Romero, P., Lupyan, D., Zerbino, D., & Ortiz, A. R. (2005b). Core deformations in protein families: A physical perspective. *Biophysical Chemistry*, *115*(2-3), 125–128. https://doi.org/10.1016/j.bpc.2004.12.016

Levy, E. D., Pereira-Leal, J. B., Chothia, C., & Teichmann, S. A. (2006). 3D complex: A structural classification of protein complexes. *PLoS Computational Biology*, *2*(11), e155. https://doi.org/10.1371/journal.pcbi.0020155

Lin, J.-H. (2011). Accommodating protein flexibility for structure-based drug design. *Current Topics in Medicinal Chemistry*, *11*(2), 171–178. https://doi.org/10.2174/156802611794863580

Liu, Y., & Bahar, I. (2012). Sequence evolution correlates with structural dynamics. *Molecular Biology and Evolution*, *29*(9), 2253–2263. https://doi.org/10.1093/molbev/mss097

Ma, B., & Nussinov, R. (2010). Enzyme dynamics point to stepwise conformational selection in catalysis. *Current Opinion in Chemical Biology*, *14*(5), 652–659. https://doi.org/10.1016/j.cbpa.2010.08.012

Madern, D., Ebel, C., Dale, H. A., Lien, T., Steen, I. H., Birkeland, N. K., & Zaccai, G. (2001). Differences in the oligomeric states of the LDH-like L-MalDH from the hyperthermophilic archaea Methanococcus jannaschii and Archaeoglobus fulgidus. *Biochemistry*, *40*(34), 10310–10316. https://doi.org/10.1021/bi010168c

Maguid, S., Fernandez-Alberti, S., Ferrelli, L., & Echave, J. (2005). Exploring the common dynamics of homologous proteins. Application to the globin family. *Biophysical Journal*, *89*(1), 3–13. https://doi.org/10.1529/biophysj.104.053041

Maguid, S., Fernández-Alberti, S., Parisi, G., & Echave, J. (2006). Evolutionary conservation of protein backbone flexibility. *Journal of Molecular Evolution*, *63*(4), 448–457. https://doi.org/10.1007/s00239-005-0209-x

Marino-Buslje, C., Monzon, A. M., Zea, D. J., Fornasari, M. S., & Parisi, G. (2019). On the dynamical incompleteness of the Protein Data Bank. *Briefings in Bioinformatics*, *20*(1), 356–359. https://doi.org/10.1093/bib/bbx084

Marsh, J. A., Teichmann, S. A., & Forman-Kay, J. D. (2012). Probing the diverse landscape of protein flexibility and binding. *Current Opinion in Structural Biology*, *22*(5), 643–650. https://doi.org/10.1016/j.sbi.2012.08.008

Marsh, J. A., & Teichmann, S. A. (2014). Protein flexibility facilitates quaternary structure

Escobedo et al.

assembly and evolution. *PLoS Biology*, *12*(5), e1001870. https://doi.org/10.1371/journal.pbio.1001870

Mitchell-White, J. I., Stockner, T., Holliday, N., Briddon, S. J., & 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*(6), 3012. https://doi.org/10.3390/ijms22063012

Monzon, A. M., Juritz, E., Fornasari, M. S., & Parisi, G. (2013). CoDNaS: A database of conformational diversity in the native state of proteins. *Bioinformatics*, *29*(19), 2512–2514. https://doi.org/10.1093/bioinformatics/btt405

Monzon, A. M., Rohr, C. O., Fornasari, M. S., & 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*, baw038. https://doi.org/10.1093/database/baw038

Monzon, A. M., Zea, D. J., Fornasari, M. S., Saldaño, T. E., Fernandez-Alberti, S., Tosatto, S. C. E., & Parisi, G. (2017a). Conformational diversity analysis reveals three functional mechanisms in proteins. *PLoS Computational Biology*, *13*(2), e1005398. https://doi.org/10.1371/journal.pcbi.1005398

Monzon, A. M., Zea, D. J., Marino-Buslje, C., & Parisi, G. (2017b). Homology modeling in a dynamical world. *Protein Science*, *26*(11), 2195–2206. https://doi.org/10.1002/pro.3274

Motlagh, H. N., Wrabl, J. O., Li, J., & Hilser, V. J. (2014). The ensemble nature of allostery. *Nature*, *508*(7496), 331–339. https://doi.org/10.1038/nature13001

Narayanan, C., Gagné, D., Reynolds, K. A., & Doucet, N. (2017). Conserved amino acid networks modulate discrete functional properties in an enzyme superfamily. *Scientific Reports*, *7*(1), 3207. https://doi.org/10.1038/s41598-017-03298-4

Nooren, I. M. A., & Thornton, J. M. (2003a). Diversity of protein-protein interactions. *The EMBO Journal*, *22*(14), 3486–3492. https://doi.org/10.1093/emboj/cdg359

Nooren, I. M. A., & Thornton, J. M. (2003b). Structural characterisation and functional significance of transient protein-protein interactions. *Journal of Molecular Biology*, *325*(5), 991–1018. https://doi.org/10.1016/S0022-2836(02)01281-0

Odokonyero, D., Sakai, A., Patskovsky, Y., Malashkevich, V. N., Fedorov, A. A., Bonanno, J. B., Fedorov, E. V., Toro, R., Agarwal, R., Wang, C., Ozerova, N. D. S., Yew, W. S., Sauder, J. M., Swaminathan, S., Burley, S. K., Almo, S. C., & Glasner, M. E. (2014). Loss of quaternary structure is associated with rapid sequence divergence in the OSBS family. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8535–8540. https://doi.org/10.1073/pnas.1318703111

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., & Thornton, J. M. (1997). CATH–a hierarchic classification of protein domain structures. *Structure*, *5*(8), 1093–1108. https://doi.org/10.1016/S0969-2126(97)00260-8

Parisi, G., Zea, D. J., Monzon, A. M., & Marino-Buslje, C. (2015). Conformational diversity and the emergence of sequence signatures during evolution. *Current Opinion in Structural Biology*, *32*, 58–65. https://doi.org/10.1016/j.sbi.2015.02.005

Pascual-García, A., Abia, D., Méndez, R., Nido, G. S., & Bastolla, U. (2010). Quantifying the evolutionary divergence of protein structures: The role of function change and function conservation. *Proteins*, *78*(1), 181–196. https://doi.org/10.1002/prot.22616

Pasti, A. P., Rossi, V., Di Stefano, G., Brigotti, M., & Hochkoeppler, A. (2022). Human lactate dehydrogenase A undergoes allosteric transitions under pH conditions inducing the dissociation of the tetrameric enzyme. *Bioscience Reports*, *42*(1) BSR20212654. https://doi.org/10.1042/BSR20212654

Perales, M., Parisi, G., Fornasari, M. S., Colaneri, A., Villarreal, F., González-Schain, N., Echave, J., Gómez-Casati, D., Braun, H.-P., Araya, A., & Zabaleta, E. (2004). Gamma carbonic anhydrase like complex interact with plant mitochondrial complex I. *Plant Molecular Biology*, *56*(6), 947–957. https://doi.org/10.1007/s11103-004-6324-z

Qi, G., & Hayward, S. (2009). Database of ligand-induced domain movements in enzymes. *BMC Structural Biology*, *9*, 13. https://doi.org/10.1186/1472-6807-9-13

Reeves, G. A., Dallman, T. J., Redfern, O. C., Akpor, A., & Orengo, C. A. (2006). Structural diversity of domain superfamilies in the CATH database. *Journal of Molecular Biology*, *360*(3), 725–741. https://doi.org/10.1016/j.jmb.2006.05.035

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., Peters, T., Demitroff, N., Alberti, S. F., Palopoli, N., Fornasari, M. S., & Parisi, G. (2022). Impact of protein conformational diversity on AlphaFold predictions. *Bioinformatics*, *38*(10), 2742–2748. https://doi.org/10.1093/bioinformatics/btac202

Sangar, V., Blankenberg, D. J., Altman, N., & Lesk, A. M. (2007). Quantitative sequence-function relationships in proteins based on gene ontology. *BMC Bioinformatics*, *8*, 294. https://doi.org/10.1186/1471-2105-8-294

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., DeMaggio, A. J., Hoekstra, M. F., Blenis, J., Hunter, T., & Cantley, L. C. (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*(11), 6486–6493. https://doi.org/10.1128/MCB.16.11.6486

Stein, R. A., & Mchaourab, H. S. (2022). SPEACH_AF: Sampling protein ensembles and conformational heterogeneity with Alphafold2. *PLoS Computational Biology*, *18*(8), e1010483. https://doi.org/10.1371/journal.pcbi.1010483

Tokuriki, N., & Tawfik, D. S. (2009). Protein dynamism and evolvability. *Science*, *324*(5924), 203–207. https://doi.org/10.1126/science.1169375

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., … Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature*, *596*(7873), 590–596. https://doi.org/10.1038/s41586-021-03828-1

Wayment-Steele, H. K., Ovchinnikov, S., Colwell, L., & Kern, D. (2022). Prediction of multiple conformational states by combining sequence clustering with AlphaFold2. *BioRxiv*, 2022.10.17.512570. https://doi.org/10.1101/2022.10.17.512570

Wei, G., Xi, W., Nussinov, R., & 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. https://doi.org/10.1021/acs.chemrev.5b00562

Whittington, A. C., Larion, M., Bowler, J. M., Ramsey, K. M., Brüschweiler, R., & 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*(37), 11553–11558. https://doi.org/10.1073/pnas.1506664112

Wiederstein, M., & Sippl, M. J. (2020). TopMatch-web: Pairwise matching of large assemblies of protein and nucleic acid chains in 3D. *Nucleic Acids Research*, *48*(W1), W31–W35. https://doi.org/10.1093/nar/gkaa366

Zea, D. J., Miguel Monzon, A., Fornasari, M. S., Marino-Buslje, C., & Parisi, G. (2013). Protein conformational diversity correlates with evolutionary rate. *Molecular Biology and Evolution*, *30*(7), 1500–1503.

Zemla, A. (2003). LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Research*, *31*(13), 3370–3374. https://doi.org/10.1093/nar/gkg571

Zhang, Y., & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins*, *57*(4), 702–710. https://doi.org/10.1002/prot.20264

Escobedo et al.