Online Protein Structure Analysis with the Bio3D WebApp

Lars Skjærven, Shashank Jariwala & Barry J. Grant

August 13, 2015

Bio3D[1] is an established R package for structural bioinformatics that provides interactive tools for biomolecular structure, sequence and simulation data analysis [1, 2]. The aim of this document is to introduce the new online WebApp interface to Bio3D functionality for **comparative protein structure analysis**[2].

Overview

Online analysis typical proceeds through five consecutive and dependent steps, (namely **SEARCH**, **ALIGN**, **FIT**, **PCA**, and **eNMA**). Each step is implemented as a consecutive navigation tab of the WebApp (see **Figure 1**) and described further below.

Figure 1. Main navigation tabs and analysis steps of the WebApp.

1. *Structure search and selection (**SEARCH**)*. This tab enables the identification and selection of PDB structures related to a user input PDB code or protein sequence. Identified structures are presented in rank order of decreasing sequence similarity to the query. Selected structures from this set will be subject to ensemble analysis of their sequence structure and conformational relationships in additional tabs.

2. *Multiple sequence alignment analysis (**ALIGN**)*. In this tab all previously selected structures are subject to multiple sequence alignment and initial sequence similarity and sequence conservation analysis.

3. *Structure fitting and analysis (**FIT**)*. In this tab aligned structures are superimposed on their invariant structural core. Analysis of pair-wise structural deviations (RMSD),

[1] The latest version of the package, full documentation and further vignettes (including detailed installation instructions) can be obtained from the main Bio3D website.

[2] WebApps provide an online interface to select Bio3D functionality and negate the need for package installation and configuration on a local user machine.

fluctuations (RMSF) and multiple structure visualization is also provided along with RMSD clustering heatmaps, dendrograms and histograms.

4. *Principal component analysis (**PCA**)*. In this tab PCA is performed on the coordinates of all superimposed structures to characterize inter-conformer relationships. This analysis effectively captures and summarizes the main conformational features and structural displacements of the ensemble. This tab also provide clustering of the structures based on the calculated principal components.

5. *Ensemble normal mode analysis (**eNMA**)*. In this tab normal mode analysis (NMA) of all structures is performed to predict large-scale motions. Here NMA is performed on all structures in the ensemble in a way that facilities the interoperation of structural similarity and dissimilarity trends. This tab also provide clustering of the structures based on the calculated normal modes and fluctuation profiles.

Example application to the aromatic amino acid hydroxylases

In the following example we will explore the available high resolution structures of the *aromatic amino acid hydroxylases* (AAAH). These important metabolic enzymes catalyze the hydroxylation of aromatic amino acids including L-Phe (phenylalanine hydroxylase; PAH), L-Tyr (tyrosine hydroxylase; TH) and L-Trp (tryptophan hydroxylase; TPH1 and TPH2). Family members assemble into a tetramer with each subunit containing an N-terminal regulatory domain, a catalytic domain, and a tetramerization domain. This study will focus on the common catalytic domain (**Figure 2**).

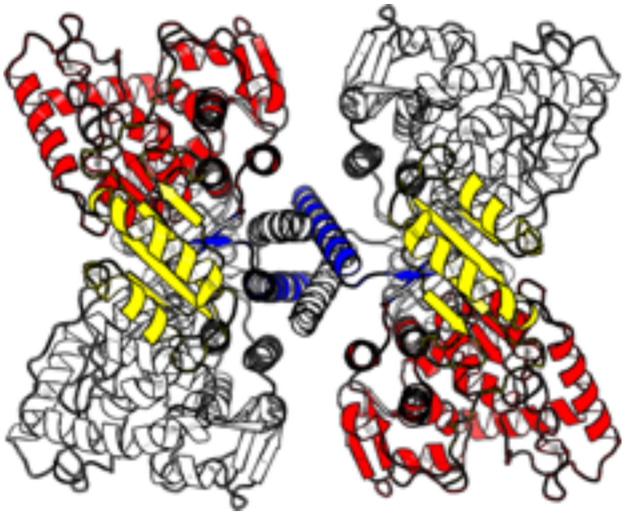


Figure 2: Structure model of phenylalanine hydroxylase (PAH) built from PDB IDs 1PHZ, 1J8U and 1PAH. The catalytic domains are shown in red and white to highlight the four subunits. The tetramerization domains are shown in blue and white, while the regulatory N-terminal domains are colored yellow (only two of four N-terminal domains are visible in this orientation).

1 SEARCH: Structure search and selection

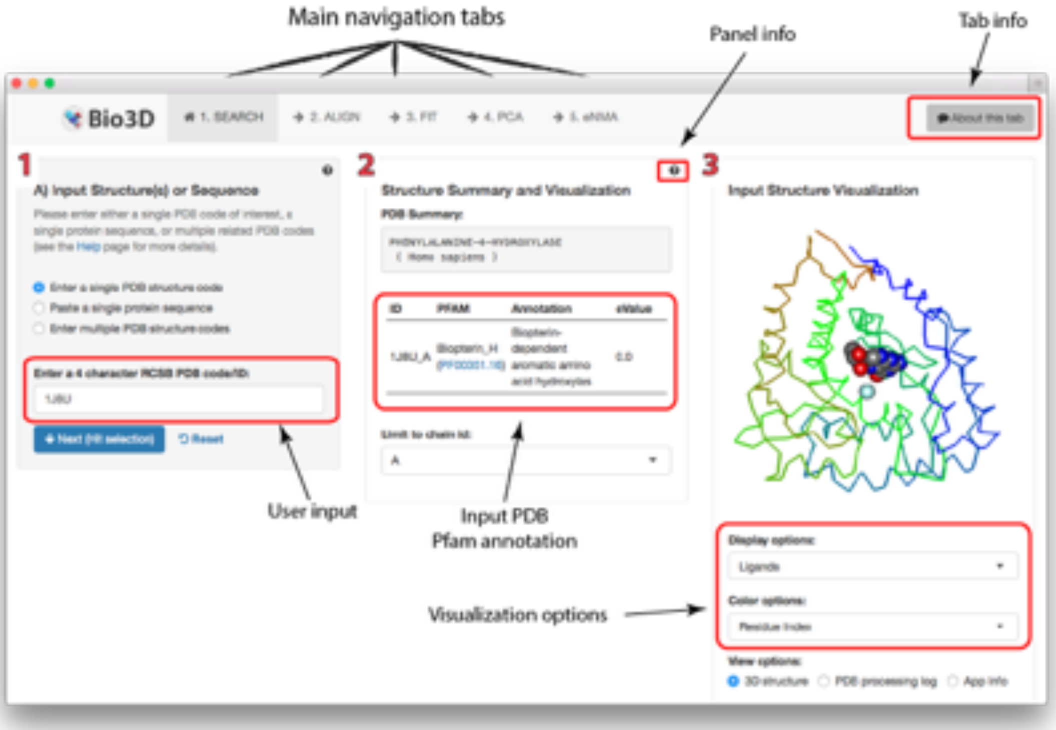
To start the analysis, open a web browser and go to the Bio3D PCA WebApp (dcmb-grant-shiny. umms.med.umich.edu/pca-app). This will bring you to the first part of the application – the **SEARCH** tab. This tab contains a total of three sub-steps (labeled **A-C**).

A) Input Structure(s) or Sequence

Here we will use a single_PDB structure code as input and type a phenylalanine hydroxylase PDB code 1J8U into the input text box (**Figure 3**).

Note: When the four characters have been entered the search will automatically start with a progress bar appearing at the very top of the screen to indicate that the server is working.

When the search is complete you will find a short summary of your query protein in the middle panel of the first row (**Structure Summary**, see **Figure 3**). This includes the protein name and species, as well as Pfam annotation data with a link to the Pfam database. You can use this link to learn more about this family of hydroxylases.



A) Input Structure(s) or Sequence

Please enter either a single PDB code of interest, a single protein sequence, or multiple related PDB codes (see the Help page for more details).

● Enter a single PDB structure code
○ Paste a single protein sequence
○ Enter multiple PDB structure codes

Enter a 4 character RCSB PDB code/ID: 1J8U

Structure Summary and Visualization

PDB Summary:

PHENYLALANINE-4-HYDROXYLASE
(Homo sapiens)

ID	PFAM	Annotation	eValue
1J8U_A	Biopterin_H (PF00351.16)	Biopterin-dependent aromatic amino acid hydroxylas	0.0

Limit to chain id: A

Input Structure Visualization

Display options: Ligands

Color options: Residue Index

View options: ● 3D structure ○ PDB processing log ○ App info

Figure 3: The Bio3D PCA WebApp is divided into five major steps, each represented by consecutive navigation tabs. Each tab is divided into multiple panels representing a task or analysis. Additional info on each tab and panel can be found by clicking the **About this tab** button on the page and the small question mark in panels.

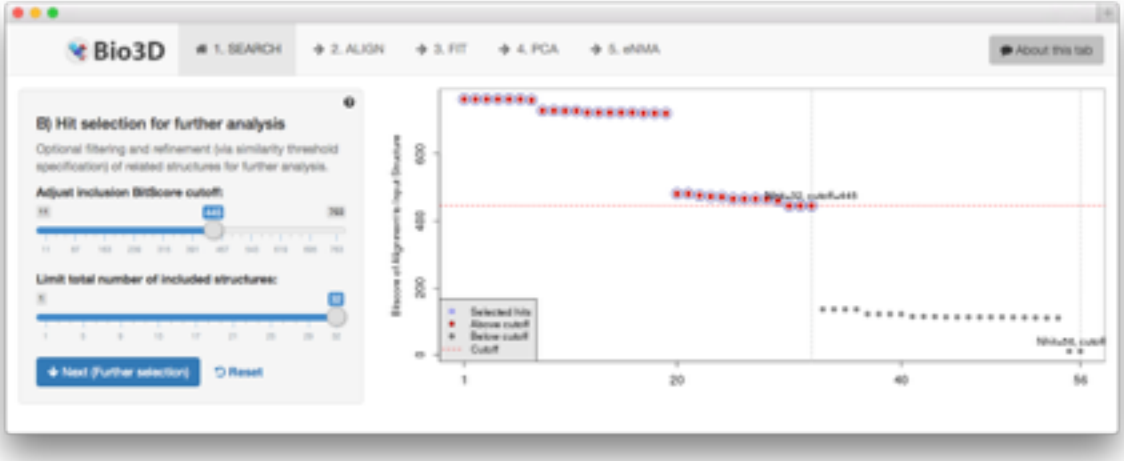
*Note the **Limit to chain ID** dropdown selector in this section, which enables selecting the chain ID (in multi-chain PDB structures). In this particular case, the input PDB contains only a single chain "A" representing the PAH catalytic domain, and thus only one option appears.*

The third and final panel of the first row (right hand side) provides a simple interactive **Input Structure Visualization**. Click and drag the mouse pointer over the protein to rotate, scroll to zoom, and select from the dropdown to apply different display and coloring options (Figure 3).

B) Hit selection for further analysis

To proceed click the blue **Next (Hit selection)** button in the first panel or simply scroll down to panel **B) Hit selection for further analysis** (see Figure 4).

As the title indicates, this panel controls the selection of hits to be analyzed in subsequent steps. It includes setting a similarity threshold cutoff value (**Adjust inclusion BitScore cutoff** slider), in which structures above this cutoff can be chosen for further analysis, and the **Limit total number of included structures** slider to set the maximum number of structures to be used. Next to this panel a plot provides a schematic representation of the search results. In this plot, each dot represents a particular hit in the PDB. Dots above the red dashed line are hits above the cutoff, while blue circles indicate selected hits. Slide the **Limit total number of included structures** to the maximum of value of 32 and notice the blue circles around the selected dots. For now, we will leave the cutoff value as it is (with a value of 445), and click the blue **Next (Further selection)** button or scroll down to the next panel.



B) Hit selection for further analysis

Optional filtering and refinement (via similarity threshold specification) of related structures for further analysis.

Adjust inclusion BitScore cutoff: 445

Limit total number of included structures: 32

Figure 4: Hit selection for further analysis. The plot shows a summary of the search results where each dot represents one particular PDB ID. Dots above the dashed red line are above the specified threshold and can be selected for further analysis. Only hits marked with a blue circle are currently selected.

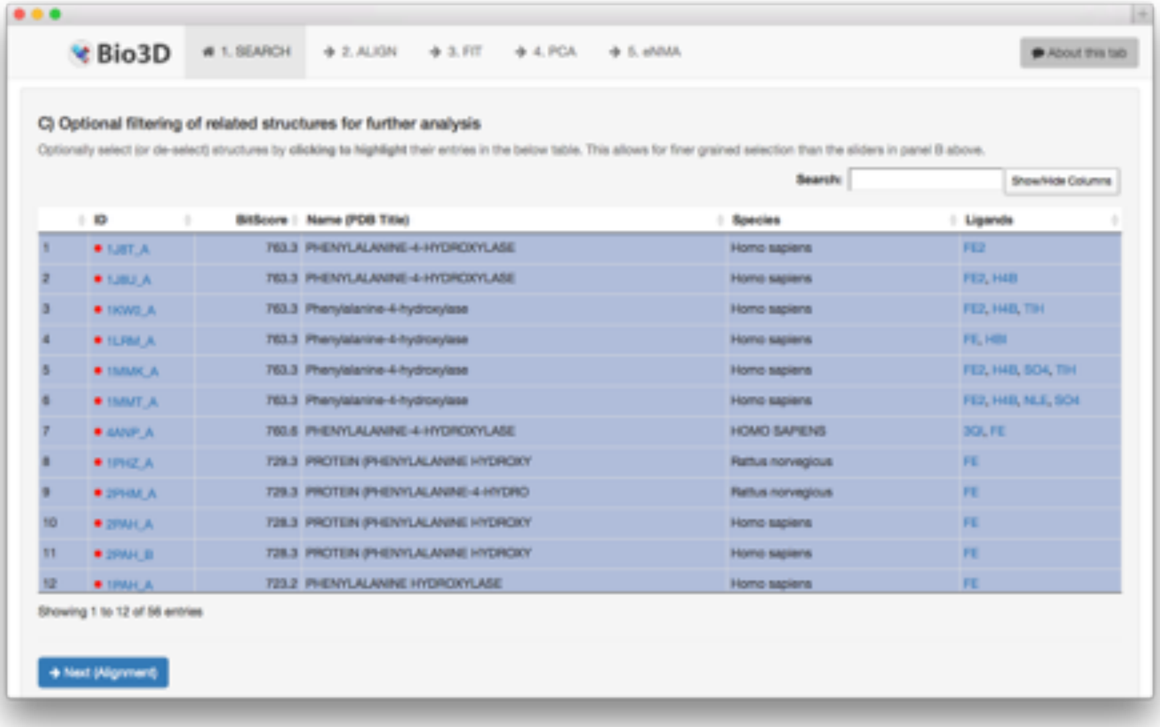
C) Optional filtering of related structures for further analysis

This panel allows for optional finer grained selection of structures and provides annotation data to help in this process. This table also links directly to the PDB if you want to explore the individual PDB entries, as well as their bound ligands. In the row of PDB ID 1J8U click the link to **Ligands** H4B to explore this ligand (**Figure 5**).

Note that 32 of the entries (rows) in this table are colored blue. These are the PDBs selected for further analysis. You can (de)select any PDB entry in this table to include/exclude it for further analysis. Deselect the first entry by clicking and notice that the blue circle in the panel B plot above also disappears.

Question I: Scroll down in the table and investigate how many PAH structures are identified. Where does TH (tyrosine hydroxylase) and TPH (tryptophan hydroxylase) appear in the list? Which PDB entries are below the specified cutoff (look for hits marked with a black dot), and which species do these entries come from?

Verify that all eukaryotic PAH, TH and TPH structures are selected, while leaving the bacterial PAH structures unselected. Proceed to the next main step by clicking on the **ALIGN** tab on the top of the page.



C) Optional filtering of related structures for further analysis

Optionally select (or de-select) structures by clicking to highlight their entries in the below table. This allows for finer grained selection than the sliders in panel B above.

Search:

	ID	BitScore	Name (PDB Title)	Species	Ligands
1	● 1J8T_A	763.3	PHENYLALANINE-4-HYDROXYLASE	Homo sapiens	FE2
2	● 1J8U_A	763.3	PHENYLALANINE-4-HYDROXYLASE	Homo sapiens	FE2, H4B
3	● 1KW0_A	763.3	Phenylalanine-4-hydroxylase	Homo sapiens	FE2, H4B, TIH
4	● 1LRM_A	763.3	Phenylalanine-4-hydroxylase	Homo sapiens	FE, HBI
5	● 1MMK_A	763.3	Phenylalanine-4-hydroxylase	Homo sapiens	FE2, H4B, SO4, TIH
6	● 1MMT_A	763.3	Phenylalanine-4-hydroxylase	Homo sapiens	FE2, H4B, NLE, SO4
7	● 4ANP_A	760.6	PHENYLALANINE-4-HYDROXYLASE	HOMO SAPIENS	3QI, FE
8	● 1PHZ_A	729.3	PROTEIN (PHENYLALANINE HYDROXY	Rattus norvegicus	FE
9	● 2PHM_A	729.3	PROTEIN (PHENYLALANINE-4-HYDRO	Rattus norvegicus	FE
10	● 2PAH_A	728.3	PROTEIN (PHENYLALANINE HYDROXY	Homo sapiens	FE
11	● 2PAH_B	728.3	PROTEIN (PHENYLALANINE HYDROXY	Homo sapiens	FE
12	● 1PAH_A	723.2	PHENYLALANINE HYDROXYLASE	Homo sapiens	FE

Showing 1 to 12 of 56 entries

Figure 5: Annotation of hits and optional filtering. The data table on bottom of the SEARCH tab provides a list of all hits found in the PDB. Click a row to (de)select a PDB ID for further analysis.

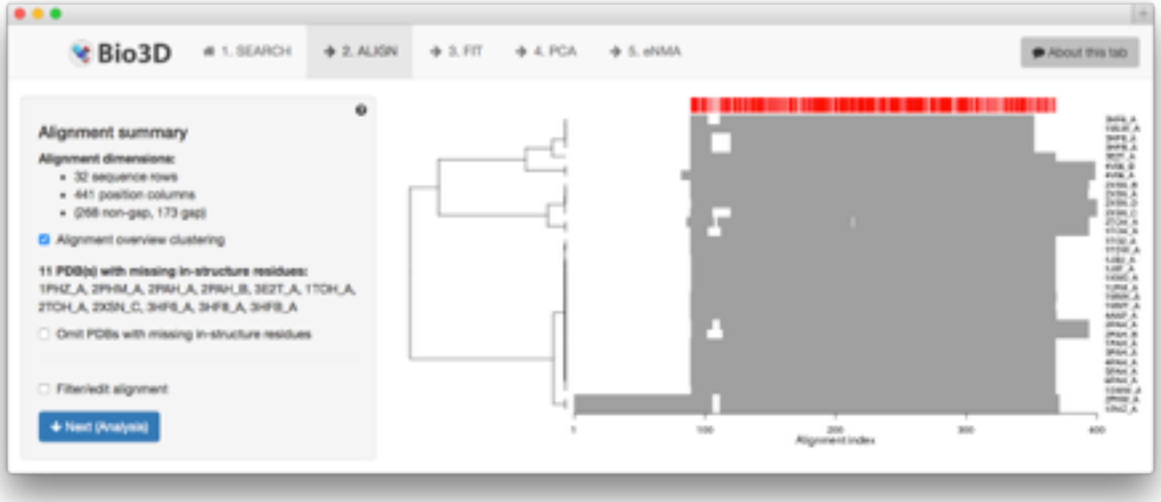
2 ALIGN: Sequence alignment and analysis

The next step in the application includes reading all PDB structures selected in the previous (**SEARCH**) tab and alignment of their sequences. This is automatically performed upon entering the **ALIGN** tab. When the sequence alignment has been completed panels providing (A) a summary of the sequence alignment, and (B) basic analyses of the sequence alignment.

A) Alignment summary

This panel shows a short summary of the alignment providing details on the number of sequence rows (equivalent to the number of PDB structures), as well as the number of position columns including a specification of the number of gap and non-gap containing columns. This panel also shows which PDB structures (if any) contain missing in-structure residues (e.g. amino acid residues which have not been resolved in the X-ray crystallography experiment).

The figure on the right hand side in the first row provides a schematic representation of the sequence alignment (Figure 6). Here the grey areas represents non-gap positions, while white areas in the alignment correspond to gaps. A representation of the sequence conservation is shown above the alignment with red areas indicating conserved positions, and white indicating less conserved. Note that the sequences in this figure are ordered based on their similarity provided by the clustering dendrogram on the left hand side (more on this below).



Alignment summary

Alignment dimensions:

- 32 sequence rows
- 441 position columns
- (268 non-gap, 173 gap)

☑ Alignment overview clustering

11 PDB(s) with missing in-structure residues:
1PHZ_A, 2PHM_A, 2PAH_A, 2PAH_B, 3E2T_A, 1TOH_A, 2TOH_A, 2XSN_C, 3HF6_A, 3HF8_A, 3HFB_A

☐ Omit PDBs with missing in-structure residues

☐ Filter/edit alignment

Next (Analysis)

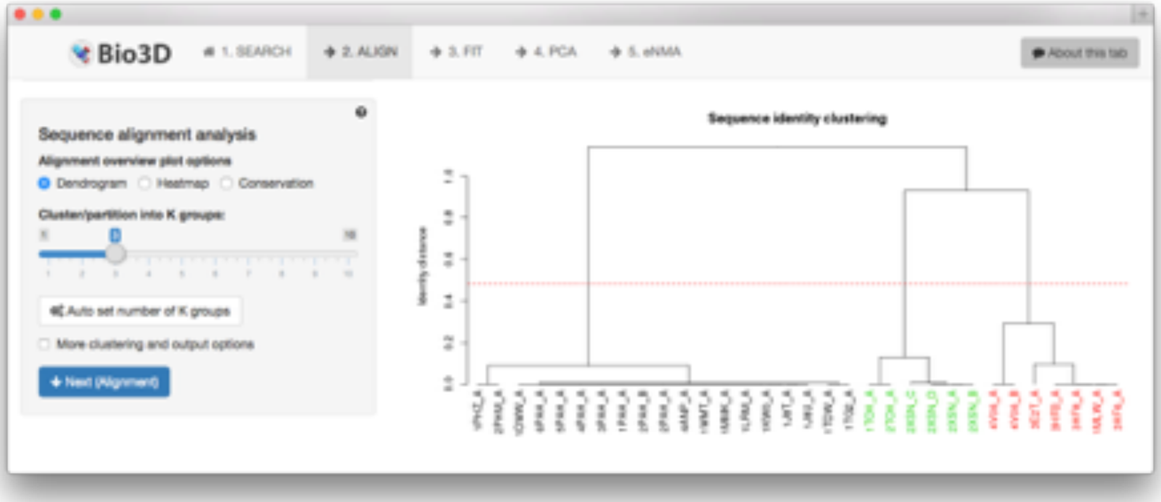
Figure 6: Alignment summary. The first row of the ALIGN tab provides an alignment summary, as well as the option to include/exclude structures for further analysis. Note the information on the missing in-structure residues.

Side note: The **Filter/edit alignment** checkbox enables custom filtering of the structures. To remove a particular PDB from the sequence alignment and further analysis, click on it with the mouse pointer and hit the delete button on your keyboard. Notice that the alignment regenerates once your edit is performed. To include it again, click an empty space within the boarders of the **Include / exclude hits** box. This panel also contains the option to upload a manually corrected sequence alignment file.

B) Sequence alignment analysis

This panel presents the results of structure clustering based on pairwise sequence identity as well as analysis of sequence conservation (see Figure 7). By default, a **dendrogram** (or tree diagram) representing the arrangement of clusters is shown. The y-axis of the dendrogram represents the distance (in terms of sequence identity) between the clusters. The cluster analysis shows that the structures can be divided into three major groups (indicated by black, red and green labels). Here, the largest cluster (colored black) corresponds to the PAH PDBs, while the red and the green clusters corresponds to TH and TPH, respectively. You can set the number of cluster groups using the slider and optionally change the clustering method and obtain PDF outputs using the "More clustering and output options".

Chose the **Heatmap** option to display a clustering dendrogram with an associated heatmap representing the pairwise sequence identity (red color corresponds to high identity and white to low identity). The colored boxes between the heatmap and the dendrogram correspond to the cluster membership of the structures.



Sequence alignment analysis

Alignment overview plot options

● Dendrogram ○ Heatmap ○ Conservation

Cluster/partition into K groups: 3

Auto set number of K groups

□ More clustering and output options

Next (Alignment)

Figure 7: Basic sequence analysis is provided in the ALIGN tab. Here the heatmap of the pairwise sequence identities are shown with cluster dendrogram in marginal regions.

Question II: Adjust the number of clusters to 4 (use the slider **Cluster by pairwise sequence identity**). What changes in the dendrogram/heatmap plot? Which hydroxylase family is divided into a new cluster? (**Hint**: Go back to the SEARCH tab and investigate the name of a few of these structures).

Question III: Switch to option **Dengrogram**. Both the PAH and TH structures seem to be contain at least two main branches. What do these branches correspond to?

C) Optional alignment display

Below these two first panels, the final sequence alignment is optionally shown with amino acid residues colored according to their physicochemical properties (**Figure 8**). Also note that conserved columns are depicted with an asterisk (*), while columns containing similar amino acid are marked with a hat (^) below the alignment.

In the sequence alignment of the hydroxylases you will find two PAH structures with a long N- terminal segment (PDB IDs: 1PHZ and 1PHM). This corresponds to the regulatory domain which is only determined for rat PAH. The remaining structures lack this region and therefore are marked with gaps (–) in the corresponding columns. The catalytic domain starts at position 100, with the conserved VPWFP motif.

When you are done inspecting the sequence analysis, proceed to the next step of the application by clicking on the **FIT** tab on the top of the page.

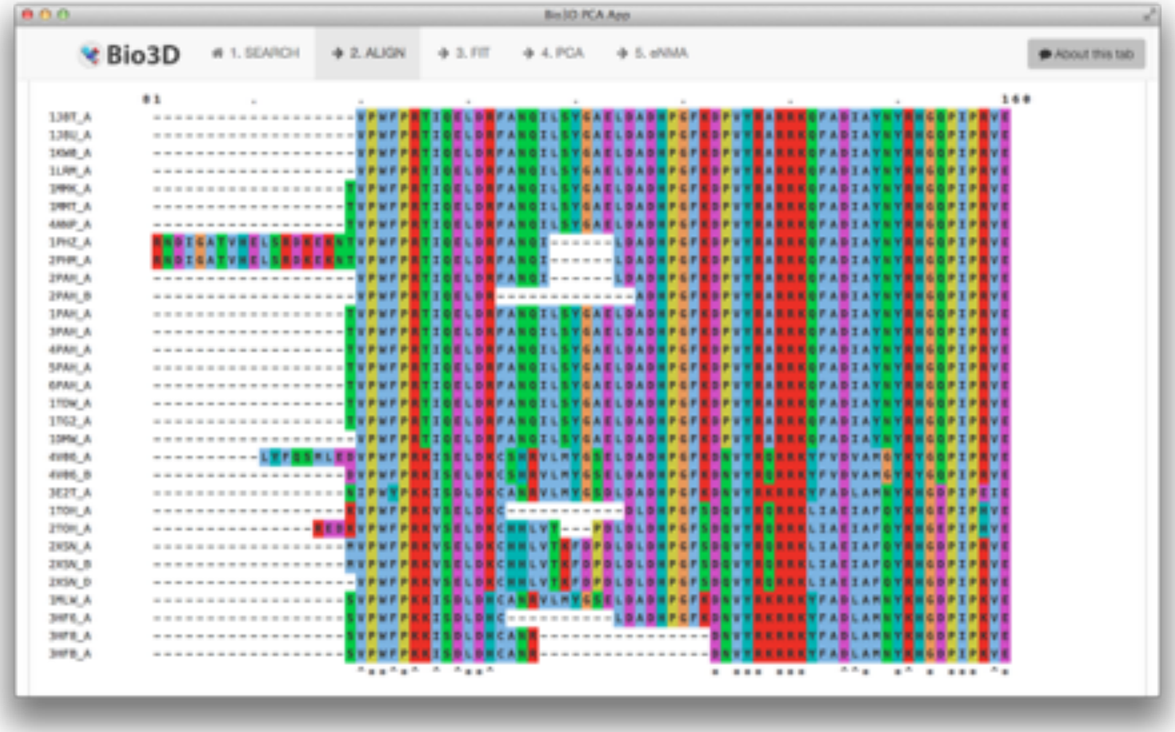


Figure 8: The ALIGN tab provides the full sequence alignment of the selected PDB structures.

Question IV: Based on the sequence alignment alone, can you identify which structures belong to the three different hydroxylases? (Scroll down to position 161-240).

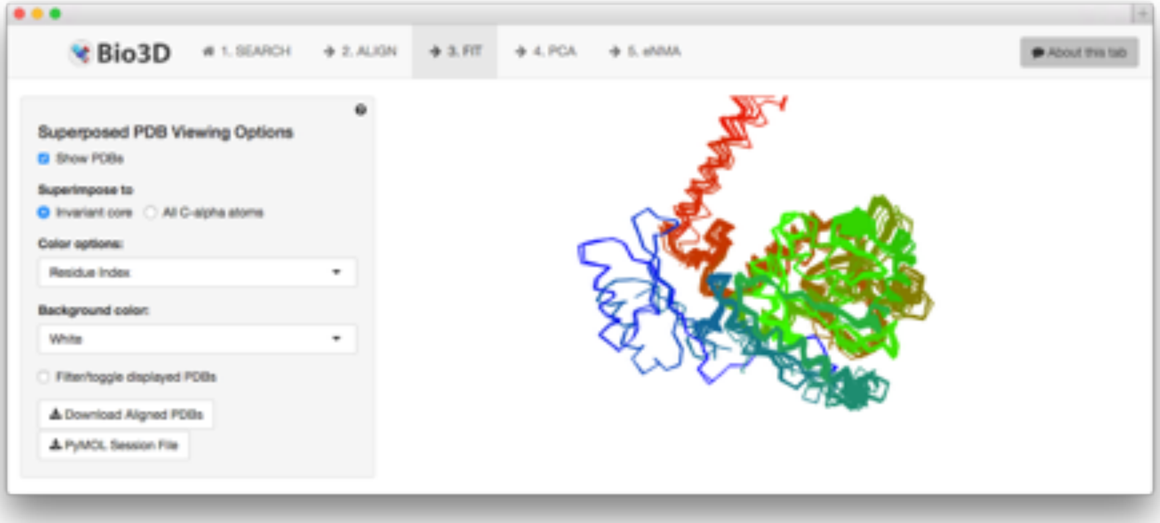
3 FIT: Structure superposition and analysis

When entering the **FIT** tab the server will automatically start the process of superimposing all structures onto each other. By default the program will identify the *invariant core* – a region with low structural variability within the ensemble – and superimpose this region.

A) Superposed PDB viewing options

The superimposed structures are shown in the first row of the **FIT** tab. Click and drag the mouse over the structures to rotate, and scroll to zoom. By default the structures are colored according to the Residue index. Another useful colouring option is by their cluster membership (RMSD Cluster: calculated based on pairwise structural deviations, which discussed further below), i.e. structures with the same color share a similar conformation.

Task: Explore the colouring options RMSF/Variability and RMSD Cluster to identify the observed conformational change. Rotate the structures to for a better visualisation angle.



Superposed PDB Viewing Options

☑ Show PDBs

Superimpose to

● Invariant core ○ All C-alpha atoms

Color options:

Residue Index

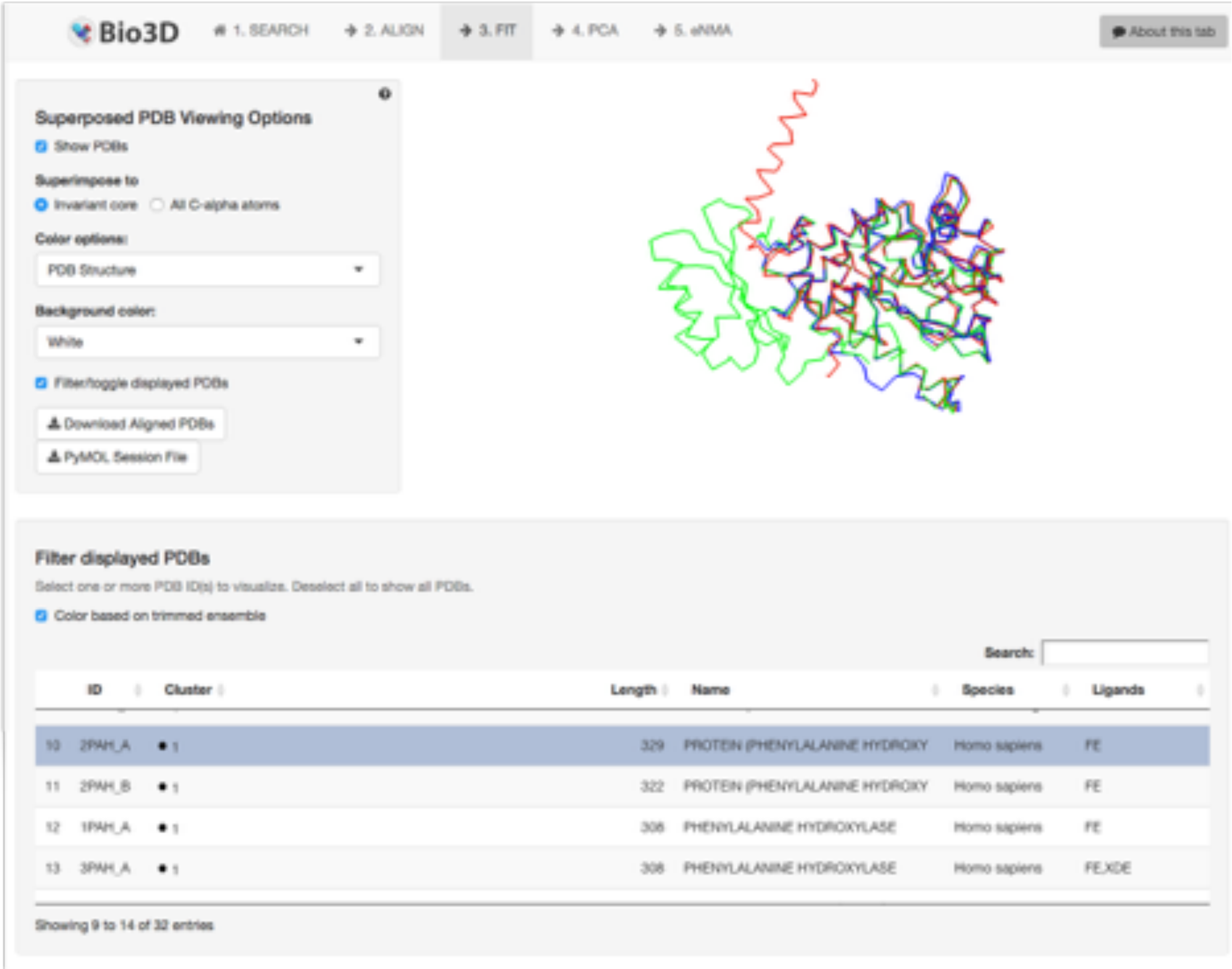
Background color:

White

○ Filter/toggle displayed PDBs

Figure 9: Structure superposition of all collected PDBs in the browser window. Here the structures are colored according to their cluster membership. Notice that a couple of structures contain an additional domain (blue) in addition to catalytic domain. This is the regulatory domain which has only been resolved for rat PAH. The additional long helix (red) pointing outwards from the catalytic domain is the tetramerization domain responsible for binding to other subunits of the tetramer. This is also only resolved in a few of the PDB IDs.

To investigate only a subset of the structures, toggle the **Filter/toggle displayed PDBs** check box in the **PDBs Viewing Options** panel. A table of all the hits will now appear below the visualization window. Select PDB IDs 1J8U, 2PAH and 1PHZ, and choose to color the structures by **PDB structure** (Figure 10). The visualizer will now only show these three selected PDB IDs, making it more easy to visualize the catalytic domain (PDB ID 1J8U; colored red), catalytic and tetramerization domain (PDB ID 2PAH; colored blue), and catalytic and regulatory domain (PDB ID 1PHZ; colored green). Deselect all in the table to show all PDBs again.



Superposed PDB Viewing Options

☑ Show PDBs

Superimpose to

● Invariant core ○ All C-alpha atoms

Color options: PDB Structure

Background color: White

☑ Filter/toggle displayed PDBs

Filter displayed PDBs

Select one or more PDB ID(s) to visualize. Deselect all to show all PDBs.

☑ Color based on trimmed ensemble

Search:

	ID	Cluster	Length	Name	Species	Ligands
10	2PAH_A	● 1	329	PROTEIN (PHENYLALANINE HYDROXY	Homo sapiens	FE
11	2PAH_B	● 1	322	PROTEIN (PHENYLALANINE HYDROXY	Homo sapiens	FE
12	1PAH_A	● 1	308	PHENYLALANINE HYDROXYLASE	Homo sapiens	FE
13	3PAH_A	● 1	308	PHENYLALANINE HYDROXYLASE	Homo sapiens	FE,XDE

Showing 9 to 14 of 32 entries

Figure 10: Three selected PDB IDs are shown in the visualizer: 1J8U (blue), 1PHZ (red), and 2PAH (green). Table (shown only when the **Filter/toggle displayed PDBs** checkbox is toggled in the **Superposed PDB Viewing Options** panel) below the structure visualizer enables selecting one or more PDB IDs for visualization.

To visualize the invariant core (in which all structures are superimposed to), toggle the **Invariant core** radio button under **Structure color** list. This will color the region defined as the invariant core red, and all other residues are colored black. As you can see the region colored red show very little structural variability. Next, color by **Gap regions**. This will color all residues placed in a gap containing column (in the ALIGN tab) red.

The ensemble of aligned PDBs can also be visualized in your favorite molecular viewer program (e.g. PyMOL or VMD) by downloading the aligned PDBs with the **Download Aligned PDBs** or **Download PyMOL session file** button. The latter option will generate a PyMOL session file with the structures aligned and colored according to the structure color options chosen in the panel. Click the button to download the file. Note that the file is zipped and you will have to unzip it before opening it in PyMOL. Open the unzipped .pse file in PyMOL and make a nice figure (**Figure 11**).

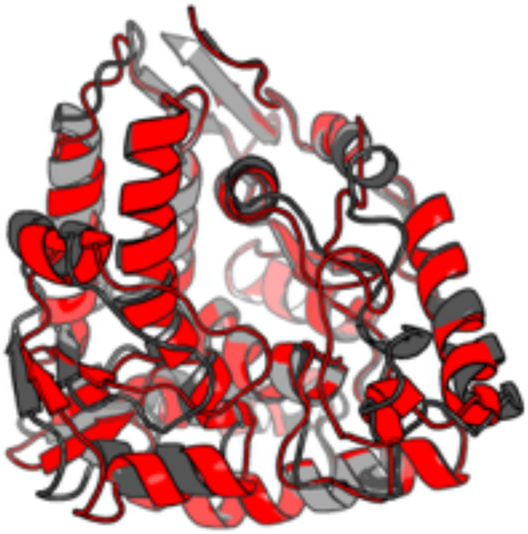
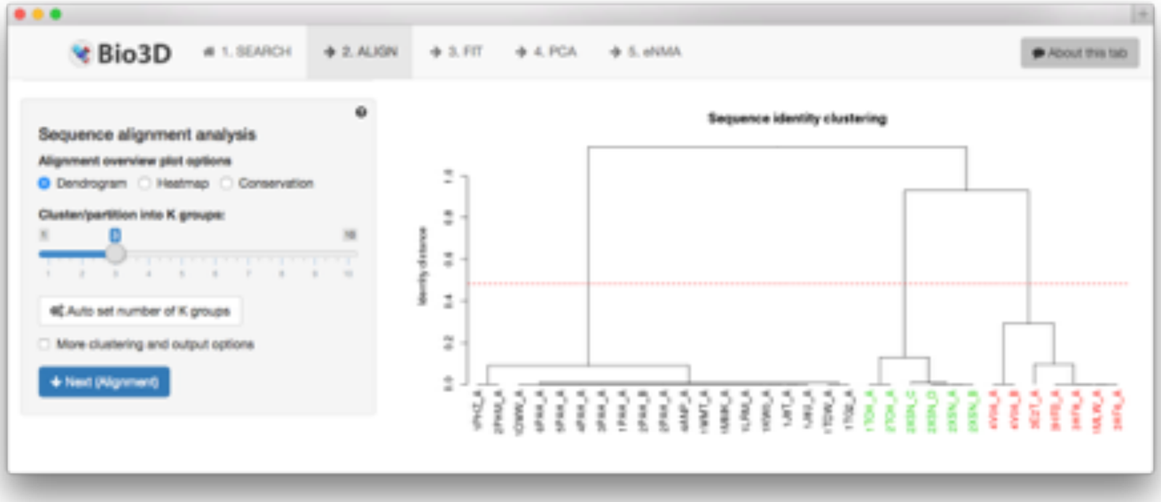


Figure 11: Visualization of two selected PDB IDs in PyMOL using the **Download PyMOL session file** button. The figure shows a comparison of PDB IDs 1J8U (black; PAH) and 3E2T (red; TPH).

B) Initial structure analysis

Scroll further down to the **Initial structure analysis** panel. This panel provides basic analyses and plotting options of the structure data. This includes clustering of the structures based on all pair wise RMSD (root mean square deviation) values. By default, a **dendrogram** (or tree diagram) representing the arrangement of clusters is shown (**Figure 12**). The y-axis of the dendrogram shows the distance (in Å) between the clusters. The cluster analysis shows that the structures can be divided into two major groups (indicated by black and red labels).

The full matrix of pair-wise RMSD values can be visualised as a **heatmap** representing the structural deviations using a color scale from white (dissimilar) to red (similar). Toggle the **Color rows by sequence identity clusters** checkbox to compare the clusters from the sequence and RMSD analysis. Recall that the first (black) cluster from the sequence analysis corresponds to PAH structures, the second (red) to TPH, and third (green) to TH. Notice that the red cluster from the RMSD analysis contains both PAH and TPH structures indicating the difference in sequence is not reflected in the structure.



Sequence alignment analysis

Alignment overview plot options

● Dendrogram ○ Heatmap ○ Conservation

Cluster/partition into K groups: 3

Auto set number of K groups

□ More clustering and output options

Next (Alignment)

Figure 12: Basic structure analyses of the collected PDBs. This plot shows the a clustering dendrogram based on he pairwise RMSD values. The label colors depict the three cluster membership obtained by RMSD clustering.

The final plot option of this panel (**RMSF Profile;** root mean square fluctuation) shows the structural variation per residue positions in the structure ensemble (Figure 13). Secondary structure elements (SSEs) are depicted in the marginal regions of the plot as black (helices) and grey (sheets) boxes. The residue numbers provided are obtained from the first structure in the ensemble (PDB ID 1J8T_A).

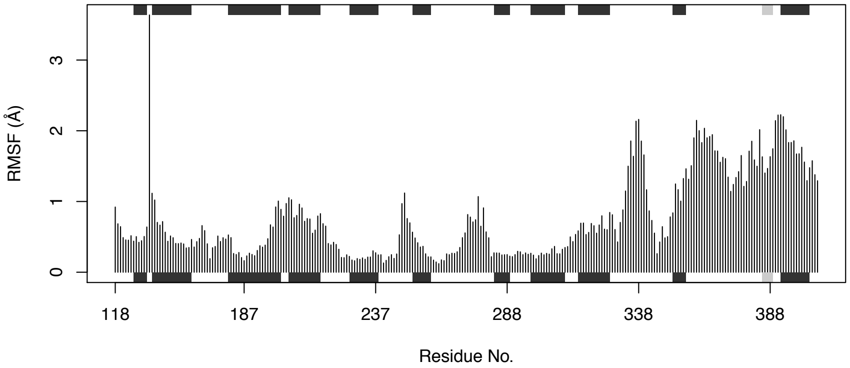
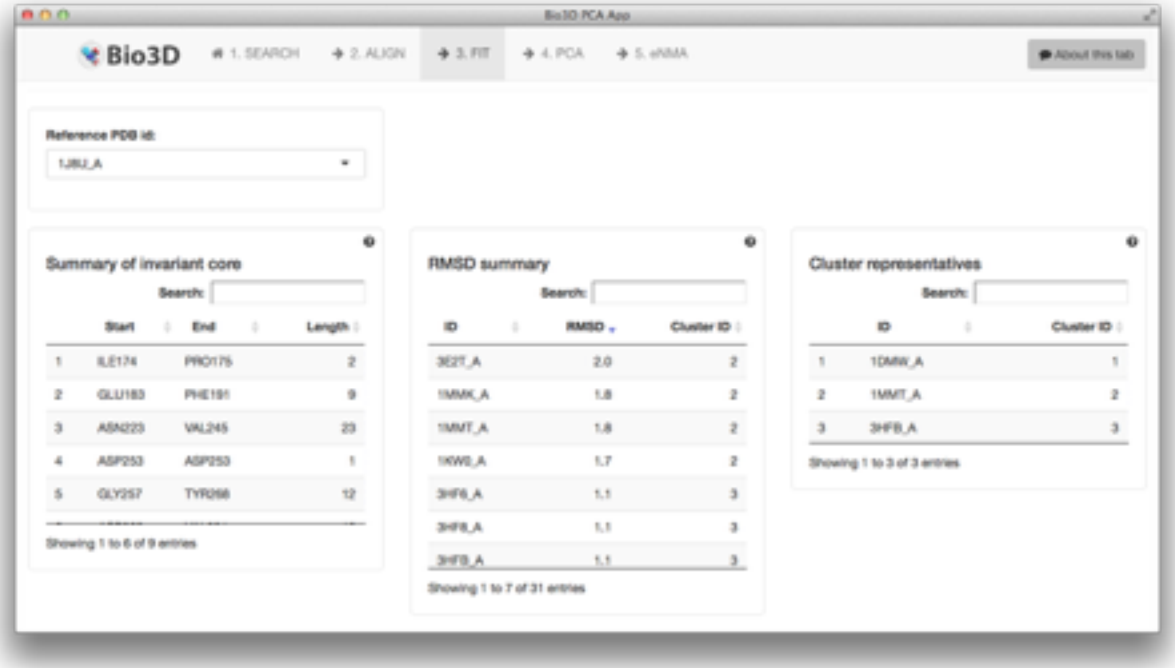


Figure 13: RMSF profile of the catalytic domains of the family of AAAHs.

C) Structural analysis summary

At the bottom of the **FIT** tab a section with three panels provides additional data of the structure analyses (**Figure 14**). The first panel gives an overview of the residues comprising the invariant core (with residues identifiers belonging to he **Reference PDB**). The second panel (**RMSD summary**) displays the RMSD values between the reference PDB to each PDBs in the ensemble. You can sort this RMSD table by clicking on either of the column titles – try e.g. to sort by RMSD. Finally, the third panel shows the list of cluster representatives – one structure from each cluster with the minimal distance to all the other cluster members.

Question: Which PDB has the highest RMSD value towards PDB ID 1J8U? What is the RMSD value between the TPH structure 3E2T and the PAH structure 1KW0?



Reference PDB id:

1J8U_A

Summary of invariant core

Search:

	Start	End	Length
1	ILE174	PRO175	2
2	GLU183	PHE191	9
3	ASN223	VAL245	23
4	ASP253	ASP253	1
5	GLY257	TYR268	12

Showing 1 to 6 of 9 entries

RMSD summary

Search:

ID	RMSD	Cluster ID
3E2T_A	2.0	2
1MMK_A	1.8	2
1MMT_A	1.8	2
1KW0_A	1.7	2
3HF6_A	1.1	3
3HF8_A	1.1	3
3HFB_A	1.1	3

Showing 1 to 7 of 31 entries

Cluster representatives

Search:

	ID	Cluster ID
1	1DMW_A	1
2	1MMT_A	2
3	3HFB_A	3

Showing 1 to 3 of 3 entries

Figure 14: The final row of the FIT tab provides a summary of the invariant core (left panel), RMSD values between the PDBs (middle panel), and cluster representatives (right panel). The first two panels show information relative to the selected **Reference PDB id**.

4 PCA: principal component analysis

The **PCA** tab provides principal component analysis (PCA) of the structure data. PCA is a statistical approach used to transform a data set down to a few important components that describe the directions where there is most variance. In terms of protein structures PCA is used to capture major structural variations within an ensemble of structures.

A) Principal Component Visualization

The first panel of the PCA tab (**Principal Component Visualization)** provides visualization of the principal components (PCs). By default the PC describing the most of the structural variations (PC-1) is show in the visualization window. Use the mouse to rotate to get a better feeling of the motion described by the first PC. By exploring the see that the PC describes a closing motion of a sub-domain of the catalytic domain. Switch the **Structure color** option to **Magnitude**. In this view atoms are colored on a scale from blue to red, where red represents atoms showing large motion amplitudes, and blue are more rigid atoms.

A trajectory view of the motion described by the PCs can be obtained by clicking on the **Download PDB trajectory** button in the **Principal Component Visualization** panel. This gives you a multi model PDB file to be opened in your favorite molecular viewer, e.g. PyMOL or VMD (**Figure 15**). The button **Download PyMOL session file** generates a PyMOL session file giving the motions as a vector field. Use what you learned in the PyMOL tutorials yesterday to make a nice figure to describe the motions of PC1.

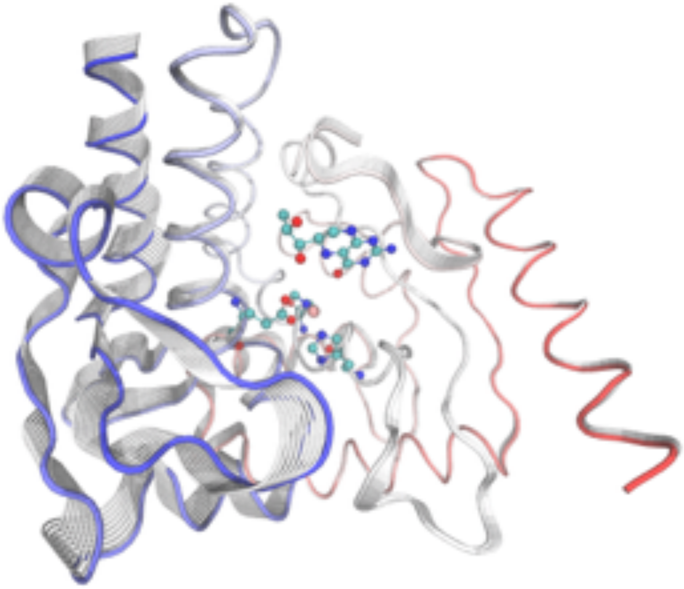


Figure 15: The first principal component of the AAAH structures reveals a closing motion. This figure is made with VMD.

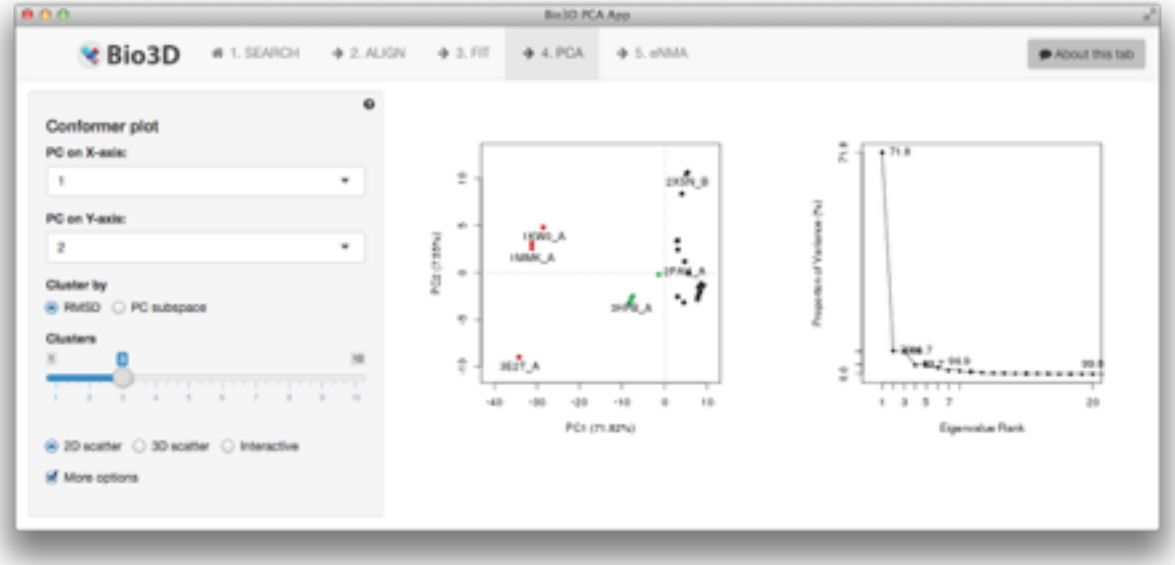
B) Conformer Plot

The second panel of the PCA tab shows a *conformer plot* – a low-dimensional representation of the conformational variability within the ensemble of PDB structures. The plot is obtained by projecting the individual structures onto two selected PCs (e.g. PC-1 and PC-2). These projections display the inter-conformer relationship in terms of the conformational differences described by the selected PCs.

The plot shows that the AAAH structures can be divided into two groups along the two first PCs (Figure 16). To inspect which PDB IDs corresponds to the different dots, scroll down to the **Hits annotation** panel. Click on the row of PDB ID 1J8U and 1KW0. This will highlight these two structures in the conformer plot above. Notice that both of these are PAH structures, but they fall into different conformational clusters. The difference seems to be ligand bound to them: 1J8U have the cofactors Fe2+ and BH4 bound, while 1KW0 also contains the substrate analogue TIH (thienylalanine).

To label the dots in the plot you can either toggle the **More options** → **Label plot** checkboxes. You should then see labels on each dot in the plot. Toggle the checkbox **Distribute labels** to avoid plotting labels on top of each other. Notice that you can label individual dots using the checkboxes with the PDB IDs next to them.

Toggle the **Interactive** plotting mode option in the **Conformer plot** panel. In this plot type you can hover over any dot to get information on which PDB ID the dot represents. In the **3D scatter** plot option you get a three dimensional representation of the same plot (but now for three PCs).



Conformer plot

PC on X-axis: 1

PC on Y-axis: 2

Cluster by

● RMSD ○ PC subspace

Clusters

3

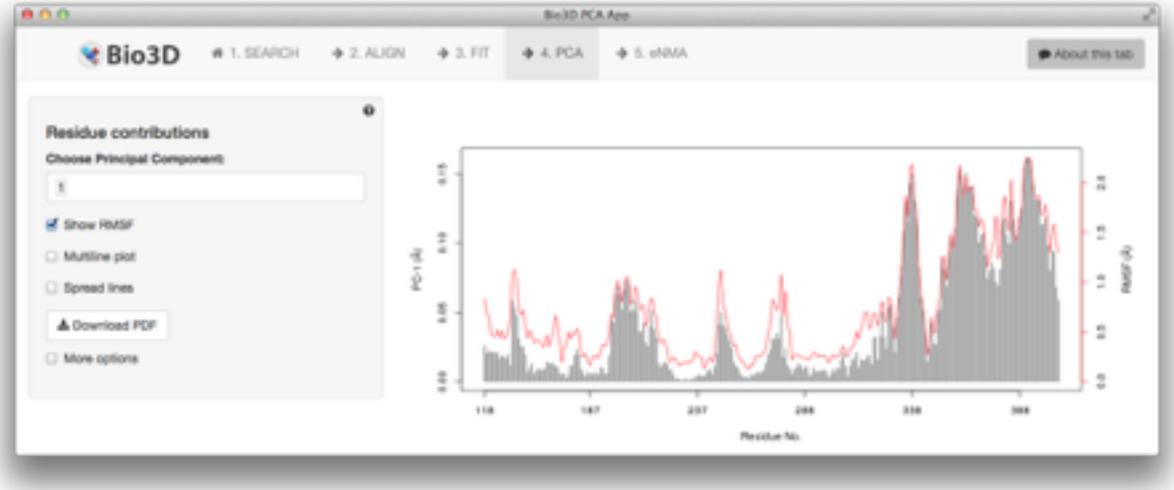
● 2D scatter ○ 3D scatter ○ Interactive

☑ More options

Figure 16: The conformer plot of all available AAAH structures. The conformer plot shows a two-dimensional representation of the conformational variability.

C) Residue contributions

The final panel of the PCA tab shows the contribution of each residue to the individual PCs (Figure 17). The height of each bar represents the relative displacement of each residue described by a given PC. Toggle the **Show RMSF** checkbox to plot the RMSF profile in the same plot, and use the **Choose Principal Component** input field to plot the contributions of additional PCs.



Residue contributions

Choose Principal Component: 1

☑ Show RMSF

□ Multiline plot

□ Spread lines

□ More options

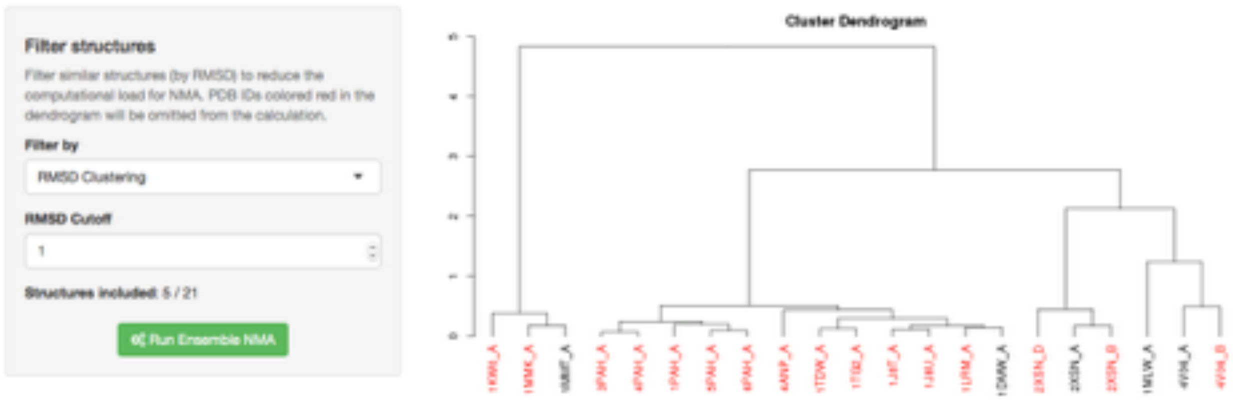
Figure 17: The contribution of each residue to the PC-1 is shown with the red line representing the RMSF profile of the aligned protein structures.

5 eNMA: Ensemble Normal Mode Analysis

The final tab of the app enables normal mode analysis (NMA) on selected structures of the ensemble. This facilitates characterising and comparing flexibility profiles of related protein structures.

A) Filter structures

Prior to calculating the normal modes we have added the option to reduce the size of the structure ensemble by filtering out structures of similar conformation (panel **Filter structures;** Figure 18). This is useful to reduce the computational load of the ensemble NMA approach. Set the **RMSD Cutoff** to 1. Observe that the a few of the labels in the cluster dendrogram on the right hand side are coloured black. These are the PDBs which will be used for the subsequent NMA, while the red will be omitted from the calculation.



Filter structures

Filter similar structures (by RMSD) to reduce the computational load for NMA. PDB IDs colored red in the dendrogram will be omitted from the calculation.

Filter by

RMSD Clustering

RMSD Cutoff

1

Structures included: 5 / 21

Figure 18: The first panel of the eNMA tab enables filtering out similar structures to reduce the computational load of the normal modes calculation.

Click the green **Run Ensemble NMA** button to start the calculation of the normal modes.

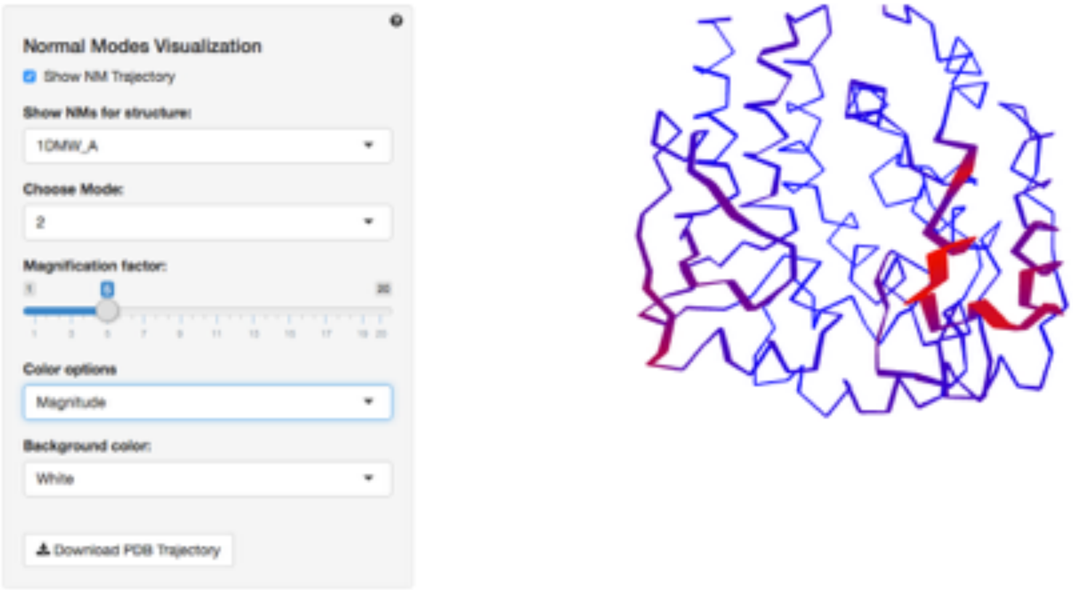
B) Normal Modes Visualisation

Once the calculation of the normal modes is complete multiple panels offering various types of analyses of the normal modes appears. The first panel (**Normal Modes Visualization**) offers visualization of the motions described by the normal modes. Click and drag the mouse in the visualization window to get a feeling on the motions. Change the **Color options** to Magnitude to color the structure by the magnitude of the motions, and increase the **Magnification factor** to amplify the motions.

Question: How does the predicted motions correspond to those observed in the PCA tab?

C) Residue fluctuations

The next panel offers plotting of the NMA derived fluctuation profiles. Here the lines in the plot are colored according to their cluster membership (see checkboxes **Cluster by**).



Normal Modes Visualization

☑ Show NM Trajectory

Show NMs for structure: 1DMW_A

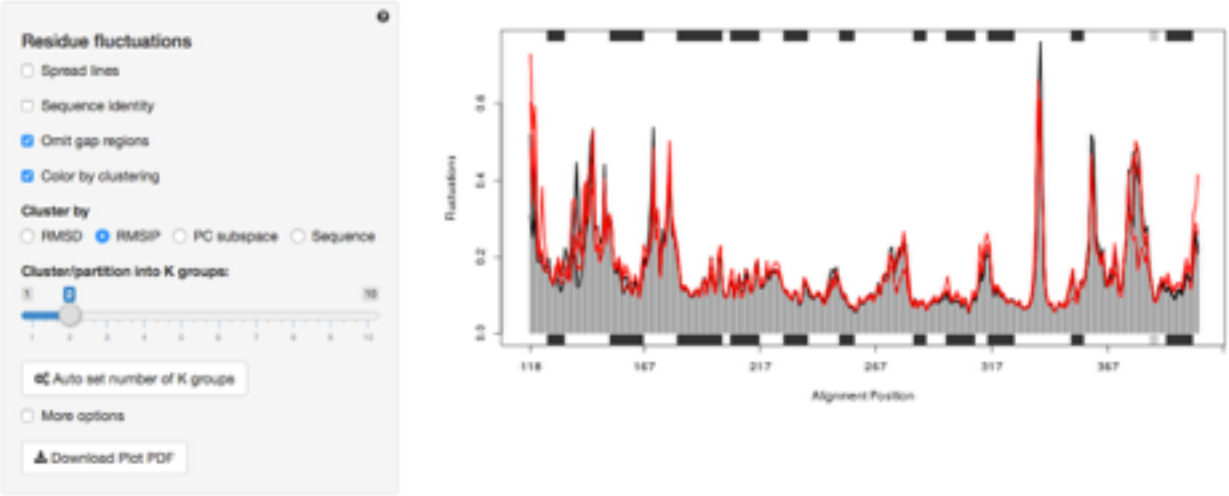
Choose Mode: 2

Magnification factor: 5

Color options: Magnitude

Background color: White

Figure 19: Visualisation of the normal mode #2 for structure 1DMW_A using the Magnitude colouring option.



Residue fluctuations

□ Spread lines

□ Sequence identity

☑ Omit gap regions

☑ Color by clustering

Cluster by

○ RMSD ● RMSIP ○ PC subspace ○ Sequence

Cluster/partition into K groups: 2

□ More options

Figure 20: NMA derived fluctuation profiles of 5 structures of the AAAH.

Summary

In this tutorial we have searched, identified and collected all available (32) structures of the aromatic amino acid hydroxylases (mammalian variants). Sequence analysis reveals that the structures can be divided into three groups which corresponds to the three families of hydroxylases: PAH, TH, and TPH. Structure analysis shows on the contrary that the three families share the same structural architecture at the catalytic domain level despite sequence identity down to 60%. The structural differences are rather explained by the presence or absence of substrate in the catalytic domain.

References

(1) Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD (2006) Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics 22: 2695–6.

(2) Skjærven L, Yao Xq, Scarabelli G, Grant BJ (2014) Integrating protein structural dynamics and evolutionary analysis with Bio3D. BMC Bioinformatics 15: 1–11. 17