3D-PP: Examples of use

3D-PP requires a list of 3D structures or homology models to make a comparison among them. There is no need to know any specific aspect about their ligands, binding sites, or additional pockets/motifs that may have functional relevance (e.g. allosteric binding sites, protein-protein interaction motifs, etc.). Thus, 3D-PP will compare and search for 3D conserved structural patterns in any set of proteins selected by the user. The criteria on which the selection is based will depend on the interests of the user. For example, one might want to search for conserved 3D patterns across a set of proteins that are over/under expressed after a pharmacological treatment, or that are involved sequentially in a specific metabolic pathway. In these and other cases, certainly, the user will require some knowledge about the involvement of the proteins in the phenomena he is interested in.

I) Target proteins of the drug "Acepromazine": In this first example, we are interested in to explore possibles conserved 3D patterns among different target proteins of the drug Acepromazine. To do this, we searched in the protein data bank using the word "Acepromazine" (Figure 1).

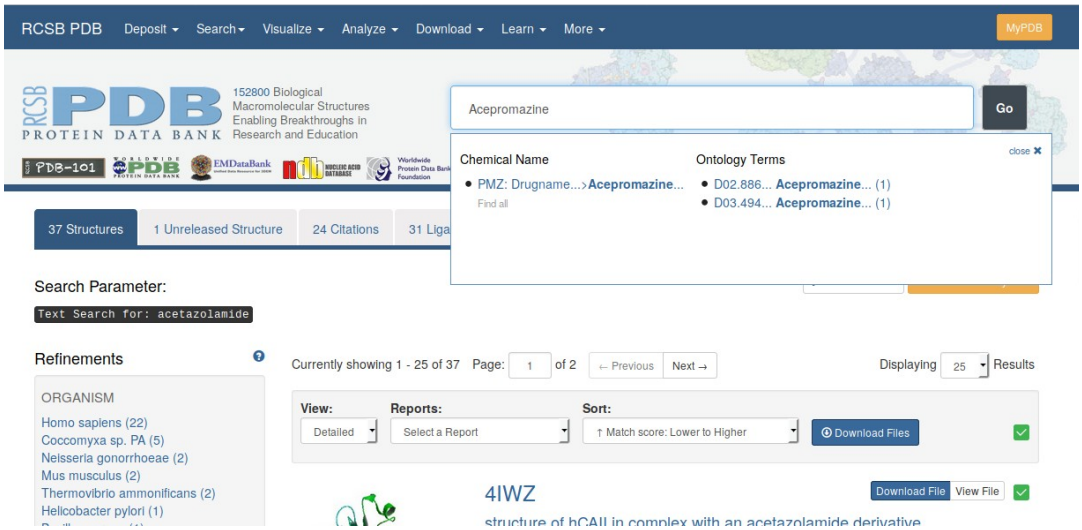


Figure 1 Protein data bank main page

The result of this search gave 15 structures with a redundancy lower than 95%. These structures have the PDB ids shown in Figure 2.

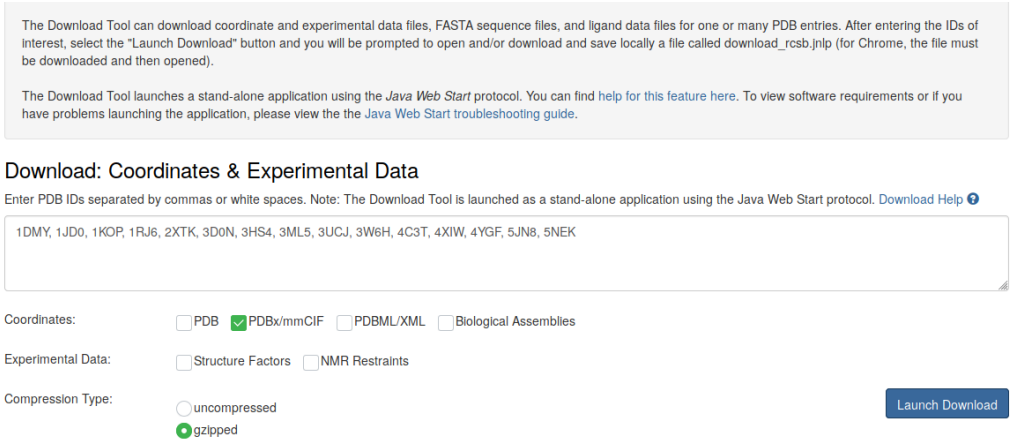


Figure 2 List of PDB ids to for the evaluation

The next step is to copy and past all the PDB ids into a text file named by the user (Figure 3). In our case, the file name is PDBids_Acepromazine.

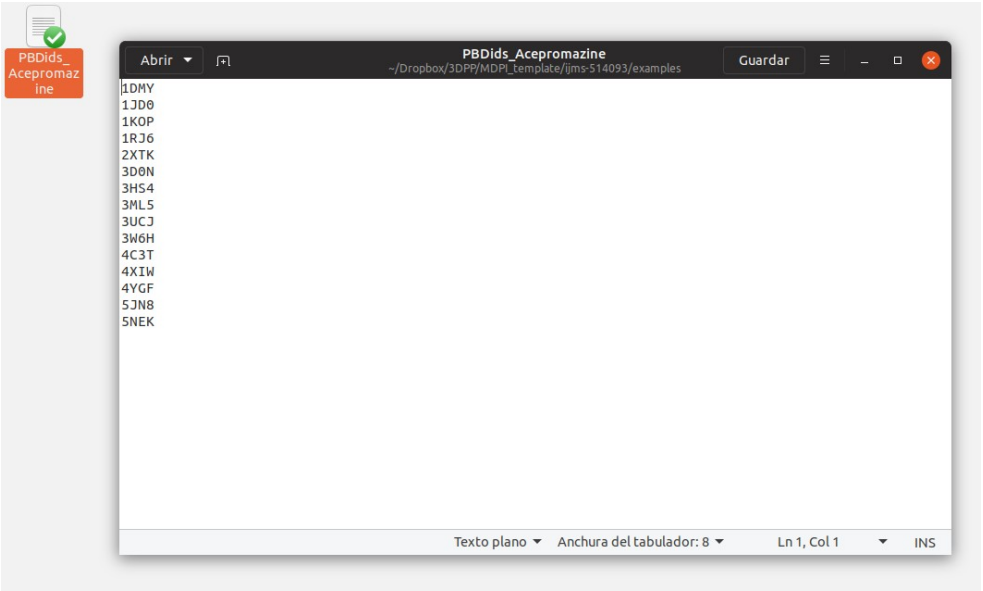


Figure 3 Text file with the PDB ids of interest

Once the user has the text file with the PDB ids of interest, it must be uploaded at the main page of 3D-PP (https://appsbio.utalca.cl/3d-pp/ Figure 4)

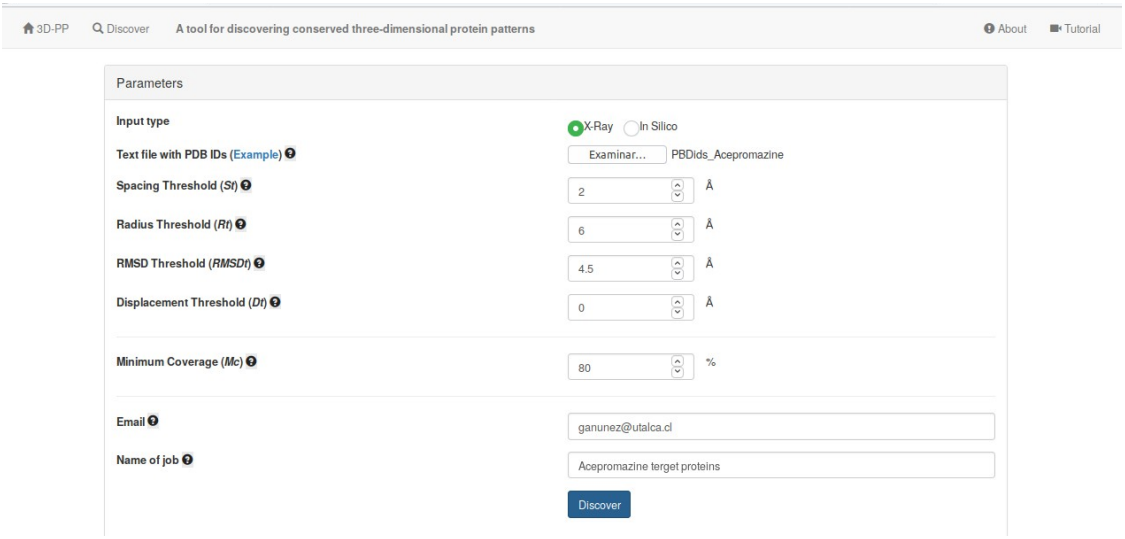


Figure 4 Input parameters of 3D-PP (main page)

As is stated in the manuscript, the user can modify some threshold parameters according to the research question:

Spacing Threshold (St): This value is used to create the Virtual Grid of Coordinates and defines, how broad and rigorous will be the exploration of 3D-patterns. For instance, a St = 0.5, means that every 0.5 Å in the 3D space of each protein structure, a new virtual coordinate of reference will be created. *In this case, we define a St value of 2 Å* .

Radius Threshold (Rt): This term represents the limits of the size of the 3D-patterns searched. Low Rt values are used to detect small binding sites (e.g. 3 Å), whereas high values allow to identify bigger sites (e.g. 7 Å). *In this case we used a Rt value of 6 Å.*

Displacement Threshold (Dt): This value is used to expand the size and shape for the exploration of the 3D-patterns. By default, this value is set in 0, which means that only the spherical 3D-patterns are searched. If the user changes this value; for example, Dt =2, two new virtual centers will be considered for the searching of 3D-patterns. This option allows to obtain seven new elliptical/oval zones that will be explored to detect non spherical 3D-patterns. *In this case, we set in 0 this value.*

RMSD Threshold (RMSDt): This value is used for clustering the 3D-patterns detected and represents a measure of structural variability for the sites composing each 3D-pattern. This parameter allows the comparison of a 3D-pattern with those, containing the same components (i.e. amino acid residues), previously found by 3D-PP. Thus, if the new site exceeds the threshold values defined by the user (RMSDt) when comparing it with the previously found site, a new cluster of the same 3D-pattern is created. Otherwise, the new 3D-pattern is included in the same cluster as the one previously found. Therefore this parameter is crucial

for 3D-PP accuracy, since it allows to discriminate between 3D-patterns that contain similar components but exhibit a different topological conformations (i.e. amino acid residues which are not in the same spatial localization/order). *In this example, we define a RMDSt value of 4.5 Å.*

Minimum Coverage (Mc): This value allows to show only 3D-patterns with a coverage value equal to or higher than Mc. *In this case, we used 80%.*

Once the measure has done, the results are sent to the e-mail indicated in the input parameters. For this data set, the first page of results is shown in Figure 5.

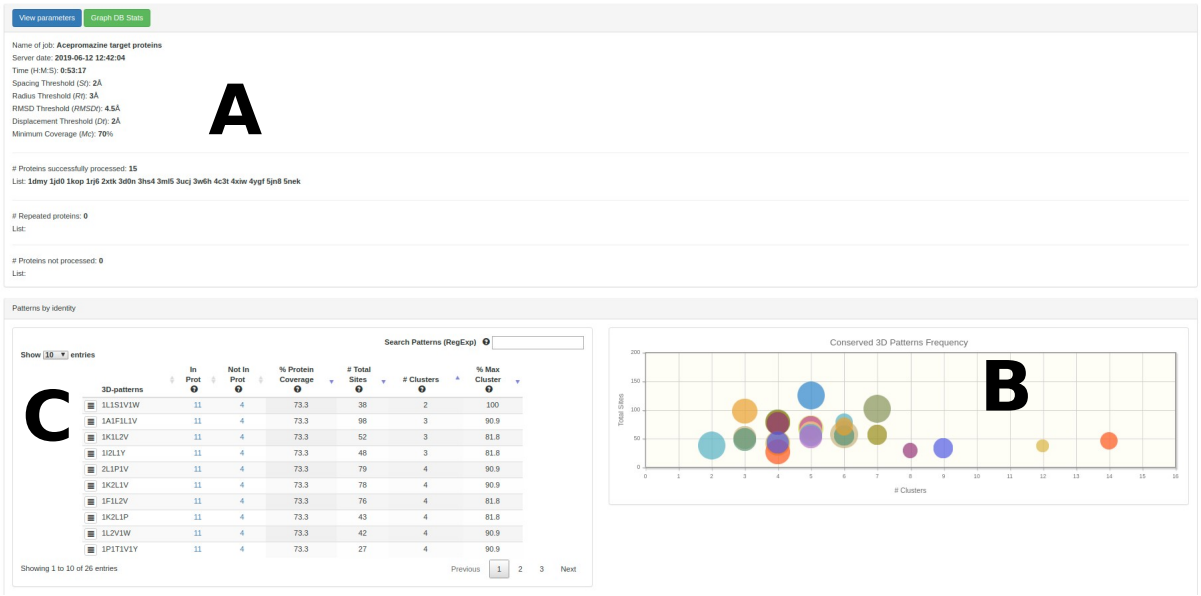


Figure 5 Fist page of results of 3D-PP

In the section denoted by Figure 5A are displayed all the input parameters defined by the user, the list of the PDB processed and the repeated or corrupted files. Additionally, the user might obtain all the data related to the database generated (green button “Graph DB Stats”).

In the section denoted by Figure 5B is displayed a graphics of circles that represent all the 3D-patterns identified in the set of proteins submitted. In this graphic, the X-axis denotes the number of cluster of each 3D-pattern and in the Y-axis are represented the number of sites belonging to each 3D-pattern. The size of the circles depict the significance of the 3D-patterns and its diameter is given by the factor between the protein coverage and the value of coverage of the cluster with the highest cluster coverage of each 3D-pattern.

In the section denoted by Figure 5C are displayed all the 3D-patterns detected. Each 3D-pattern has six features that can be used to filter or sort the results:

In Prot: The number of proteins in which a specific 3D-pattern was detected.

Not In: The number of proteins in which a specific 3D-pattern was not detected.

% Protein Coverage (PCv): Level of conservation of a 3D-pattern in the set of proteins evaluated. The PCv is calculated as: *In Prot / (amount of proteins submitted)*

 # Total Sites: Amount of sites (arrangement of residues) which are part of a specific 3D-pattern.

 # Clusters: This value represents the structural variability of a 3D-pattern. Thus, a low number of clusters denotes low variability, and on the contrary, a high number of clusters is indicative of several structural conformations (with different topologies) of sites forming a 3D-pattern.

% Max Cluster: Represents the cluster with the highest coverage on each 3D-pattern.

Additionally, the user can search a sub-3D-pattern of interest, using a simple regular expression. For example, if the user defines ^2L.*1Y as a regular expression, all the 3D-patterns that begin with 2L and ends 1Y will be showed (Figure 6).

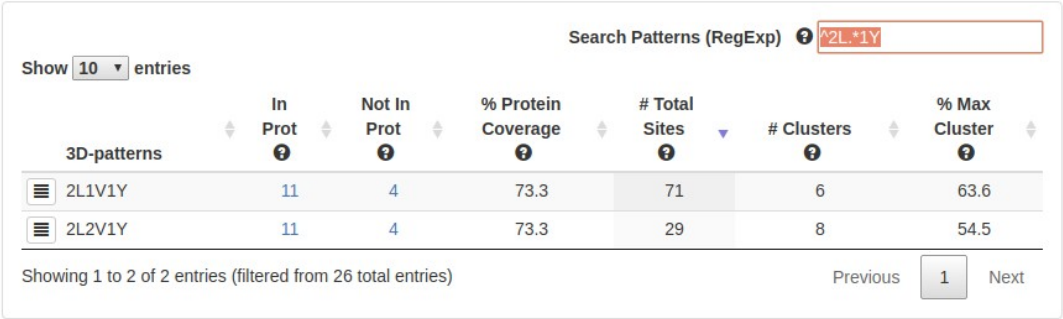


Figure 6 The search for sub-3D-patterns among the 3D-patterns detected

This searching shows two 3D-patterns with the same protein coverage (73.3 %) but different cluster coverages. Selecting the 3D-pattern with the highest cluster coverage (2L1V1Y), the next step of results is displayed as is showed in Figure 7.

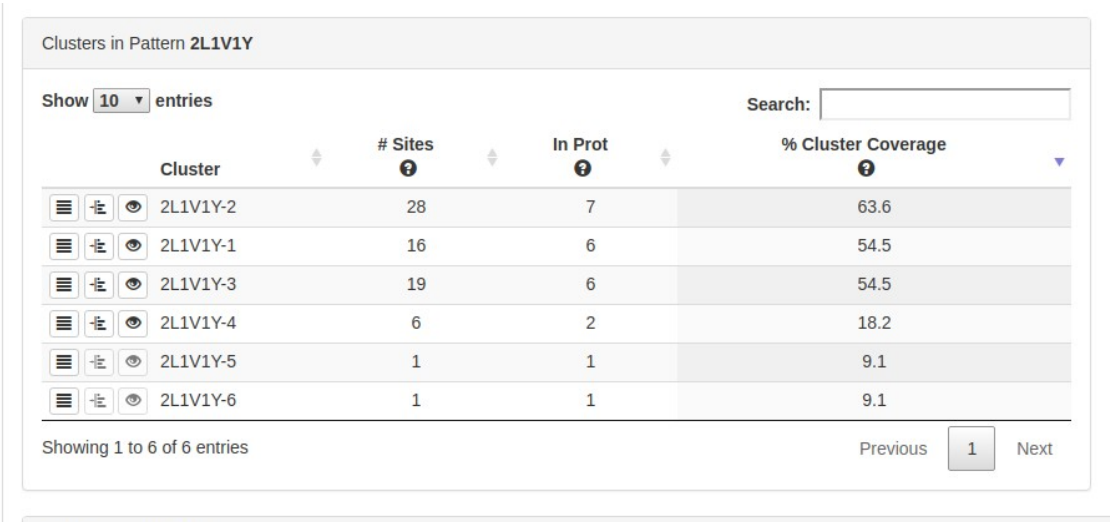


Figure 7 Clusters of the 3D-pattern 2L1V1Y

IIn this second level of results, each cluster has 3 features that can be used for filter or sort the data. Additionally, 3 action buttons are available:

 # Sites: Amount of sites (arrangement of residues) which are part of a specific cluster.

In Prot: The number of proteins that contain a particular 3D-pattern.

% Cluster Coverage (CCv): Level of conservation of a cluster in the set of proteins belonging to a particular 3D-pattern.

This button will show all the sequences of the sites that are part of each cluster. If we click on this button, for example, in the cluster 2L1V1Y-2, the following list of results will be displayed (Figure 8):

Sites in Cluster **2L1V1Y-2**

Show 10 entries Search:

Site ID	Site	Chain	Protein	RMSD	Base
4766-2	LEU141:LEU198:TYR131:VAL121	B	1dmy	0	yes
1527-3	LEU124:LEU79:TYR54:VAL52	A	1kop	2.3	no
827-1	LEU122:LEU124:TYR86:VAL81	B	1kop	2.7	no
612-1	LEU122:LEU188:TYR86:VAL81	A	1kop	2.9	no
553-1	LEU122:LEU188:TYR86:VAL81	B	1kop	2.9	no
2707-1	LEU185:LEU79:TYR51:VAL49	A	3w6h	3.2	no
531-1	LEU144:LEU185:TYR51:VAL49	A	3hs4	3.3	no
3583-3	LEU86:LEU87:TYR82:VAL237	D	5nek	3.4	no
1872-2	LEU163:LEU88:TYR54:VAL52	A	1kop	3.5	no
3365-2	LEU127:LEU162:TYR128:VAL126	A	1kop	3.6	no

Showing 1 to 10 of 28 entries Previous 1 2 3 Next

Figure 8 Sites belonging to the cluster 2L1V1Y-2

The sites listed in Figure 8 can be filtered and ordered by its chain, the name of the protein and the value of RMSD. Also, the user can visualize the site clicking on the eye button  In this step, the user can explore the site through a Jsmol viewer, take a picture or download the PDB file of the site (Figure 9).

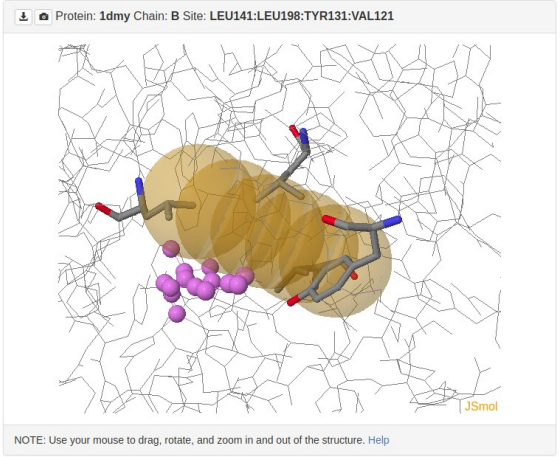


Figure 9 Visualization of a site

This button will show a sequence alignment of all the sites that are part of each cluster (Figure 10).

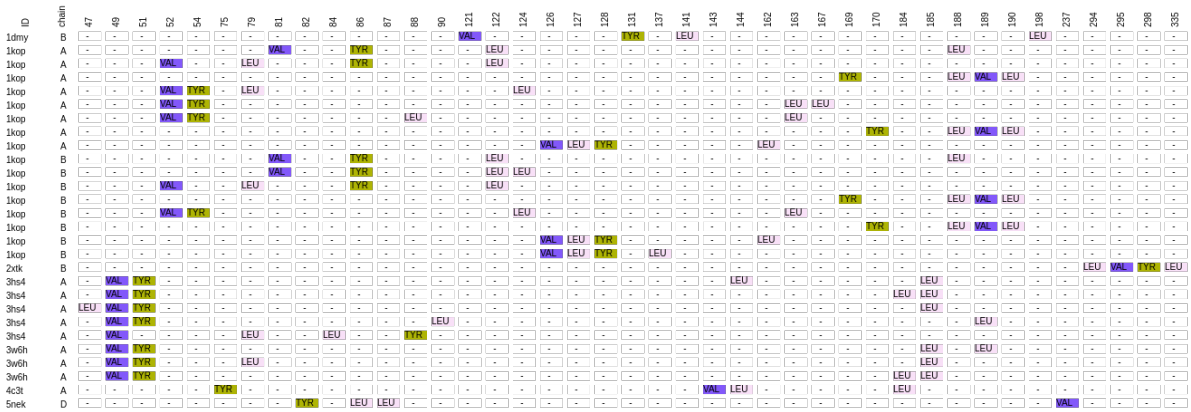


Figure 10 Sequence alignment of the sites belonging to the cluster 2L1V1Y-2

This button will show a structural alignment of all the sites that are part of each cluster (Figure 11).

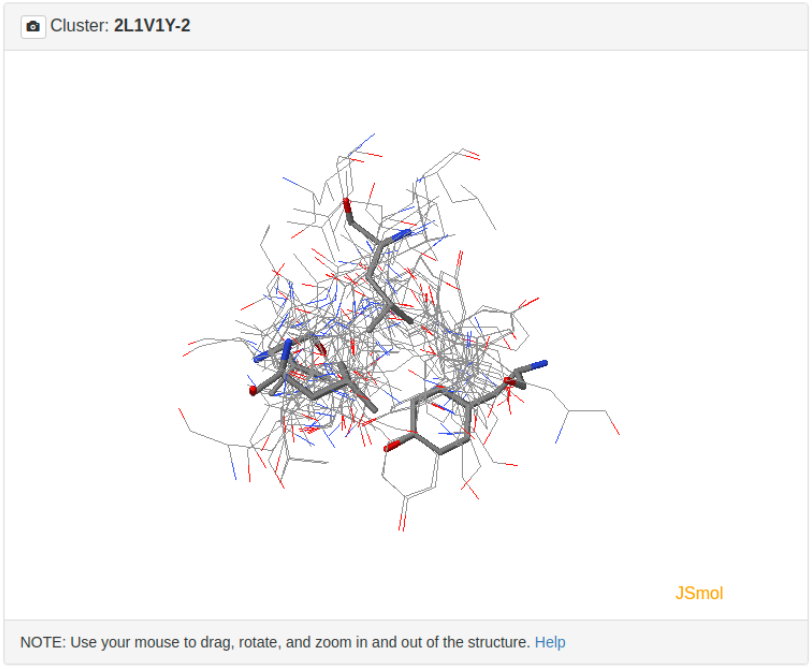


Figure 11 Structural alignment of the sites belonging to the cluster 2L1V1Y-2

In this step, the user can explore the sites through a Jsmol viewer, take a picture or download the PDB files of the sites. The remarked site is the site defined as the base for the measures of RMSD on a particular cluster.

II) Dopamine target proteins: In this second example, we are interested in looking for similar 3D patterns in four different proteins that are natural targets of the endogenous neurotransmitter Dopamine. We use the PDB ids 5WIU (dopamine receptor), 4m48 (dopamine transporter), 6FVZ (monoamine oxidase B) and 2Z5X (monoamine oxidase A).

In this example, the same as the previous example, we included a number 2 (Å) in the field "Displacement Threshold (Dt)". That means that in addition to the spherical searching for the sites, 3D-PP will search sites with elliptical shapes. This method is described in Figure 12 A and B.

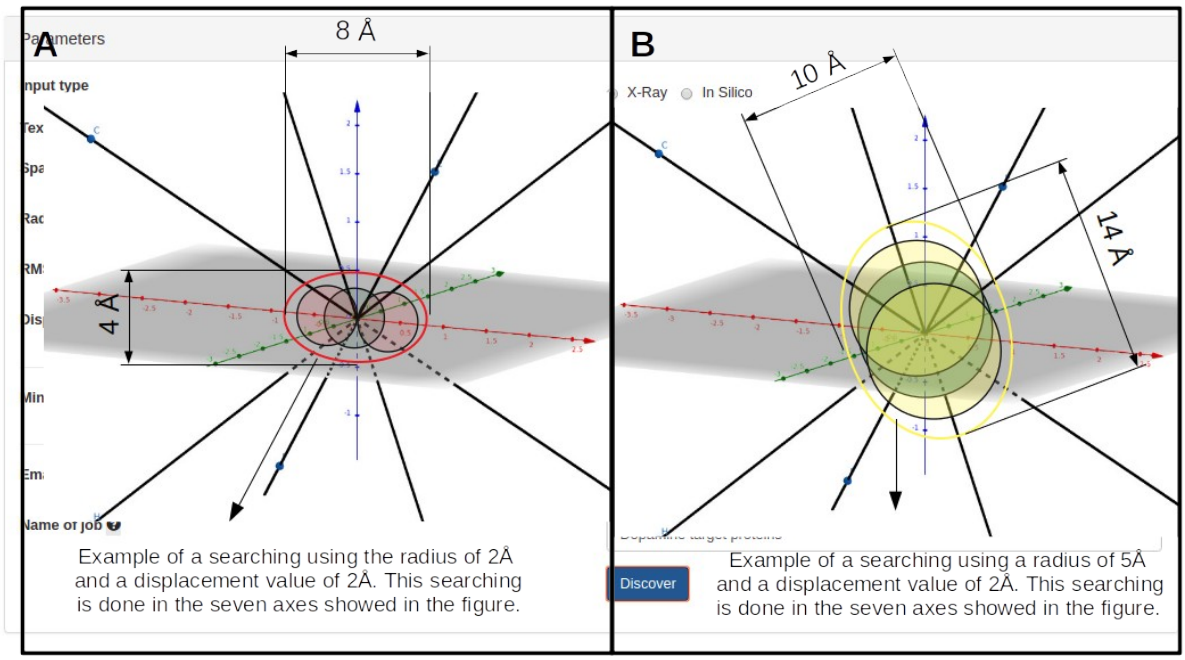


Figure 12 Input paremers for the Dopamine target proteins

The initial results are described in Figure 13. All the features shown in the first example can be applied to this or any other set of protein structures evaluated.

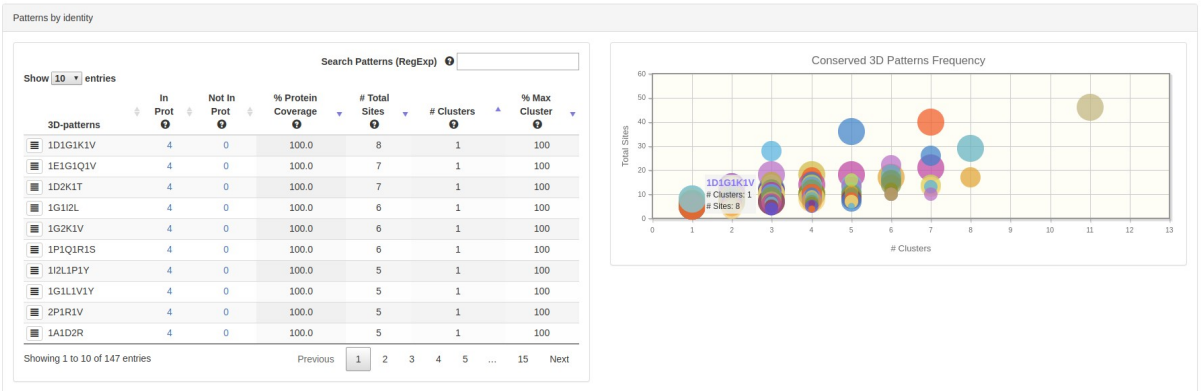
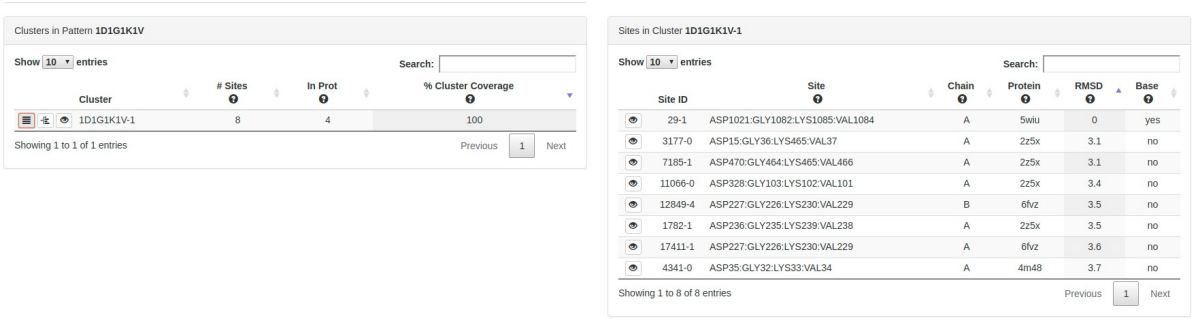


Figure 13 First level of results of the Dopamine target proteins

As is possible to observe in Figure 13, many 3D-patterns with 100% of coverage (protein and cluster) were detected. In this case, we might select by those with a higher number of sites;

that would be the 3D-pattern 1D1G1K1V. Exploring this 3D-pattern (Figure 14), we can see 1 cluster in where some proteins present more than one site (e.g. Site ID 3177-0, 7185-1, 11066-0 and 1782-1) in a different localization on the protein (Figure 15).



Clusters in Pattern **1D1G1K1V**

Show 10 entries — Search:

Cluster	# Sites	In Prot	% Cluster Coverage
1D1G1K1V-1	8	4	100

Showing 1 to 1 of 1 entries — Previous 1 Next

Sites in Cluster **1D1G1K1V-1**

Show 10 entries — Search:

Site ID	Site	Chain	Protein	RMSD	Base
29-1	ASP1021:GLY1082:LYS1085:VAL1084	A	5wiu	0	yes
3177-0	ASP15:GLY36:LYS465:VAL37	A	2z5x	3.1	no
7185-1	ASP470:GLY464:LYS465:VAL466	A	2z5x	3.1	no
11066-0	ASP328:GLY103:LYS102:VAL101	A	2z5x	3.4	no
12849-4	ASP227:GLY226:LYS230:VAL229	B	6fvz	3.5	no
1782-1	ASP236:GLY235:LYS239:VAL238	A	2z5x	3.5	no
17411-1	ASP227:GLY226:LYS230:VAL229	A	6fvz	3.6	no
4341-0	ASP35:GLY32:LYS33:VAL34	A	4m48	3.7	no

Showing 1 to 8 of 8 entries — Previous 1 Next

Figure 14 One cluster detected on the 3D-pattern 1D1G1K1V

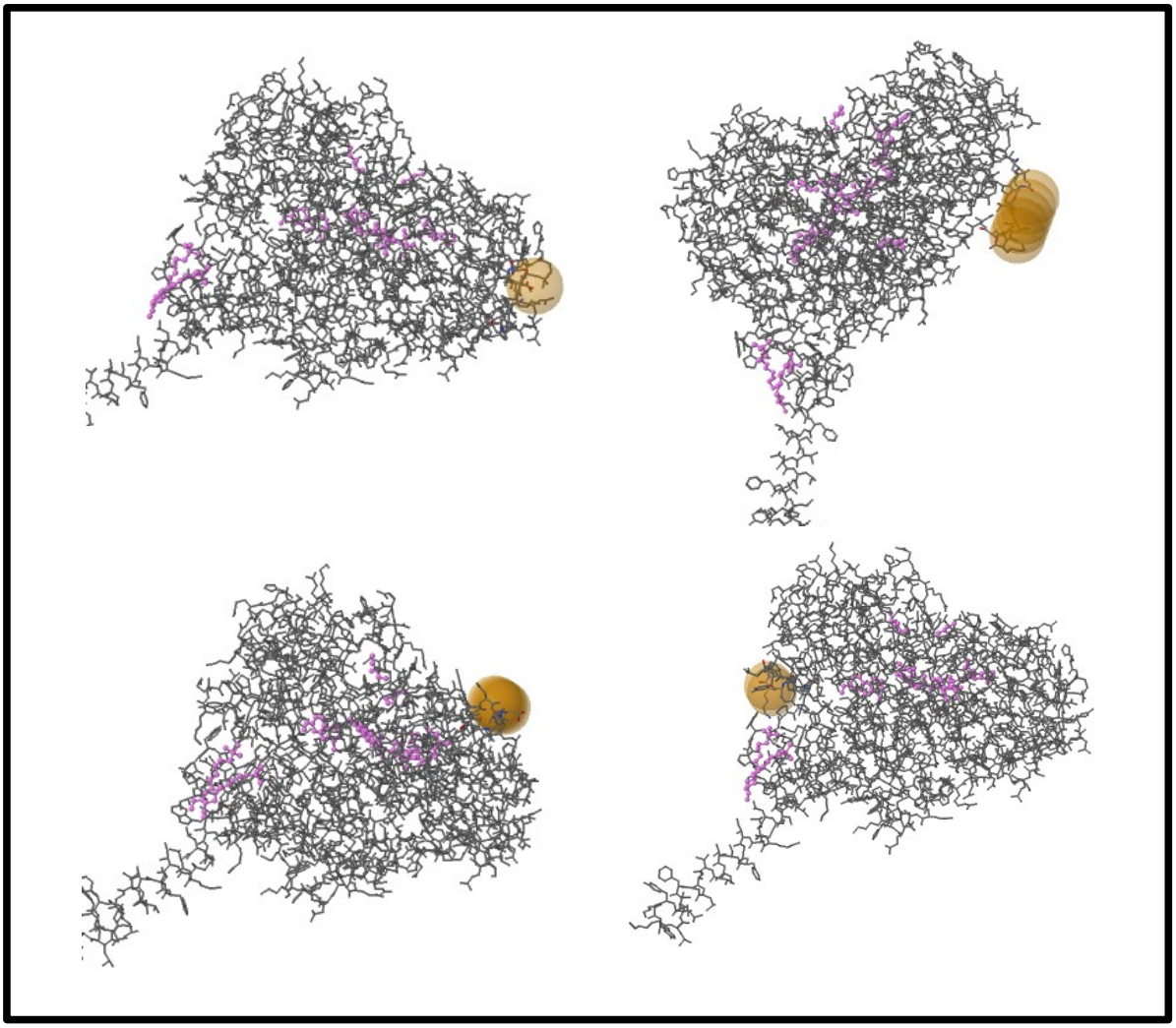


Figure 15 Four different sites of the same protein (PDBid 2z5x) in the cluster 1D1G1K1V-1.

Finally, in Figures 16 and 17 are shown the sequence and structural alignments for the sites of the cluster 1D1G1K1V-1.

ID	chain	15	32	33	34	35	36	37	101	102	103	226	227	229	230	235	236	238	239	328	464	465	466	470	1021	1082	1084	1085
2z5x	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	GLY	ASP	VAL	LYS	-	-	-	-	-	-	-	-	-
2z5x	A	ASP	-	-	-	-	GLY	VAL	-	-	-	-	-	-	-	-	-	-	-	-	-	LYS	-	-	-	-	-	-
2z5x	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	GLY	LYS	VAL	ASP	-	-	-	-
2z5x	A	-	-	-	-	-	-	-	VAL	LYS	GLY	-	-	-	-	-	-	-	-	ASP	-	-	-	-	-	-	-	-
4m48	A	-	GLY	LYS	VAL	ASP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5wiu	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	ASP	GLY	VAL	LYS
6fvz	A	-	-	-	-	-	-	-	-	-	-	GLY	ASP	VAL	LYS	-	-	-	-	-	-	-	-	-	-	-	-	-
6fvz	B	-	-	-	-	-	-	-	-	-	-	GLY	ASP	VAL	LYS	-	-	-	-	-	-	-	-	-	-	-	-	-

Figure 16 Sequence alignment of the sites belonging to the cluster 1D1G1K1V-1.

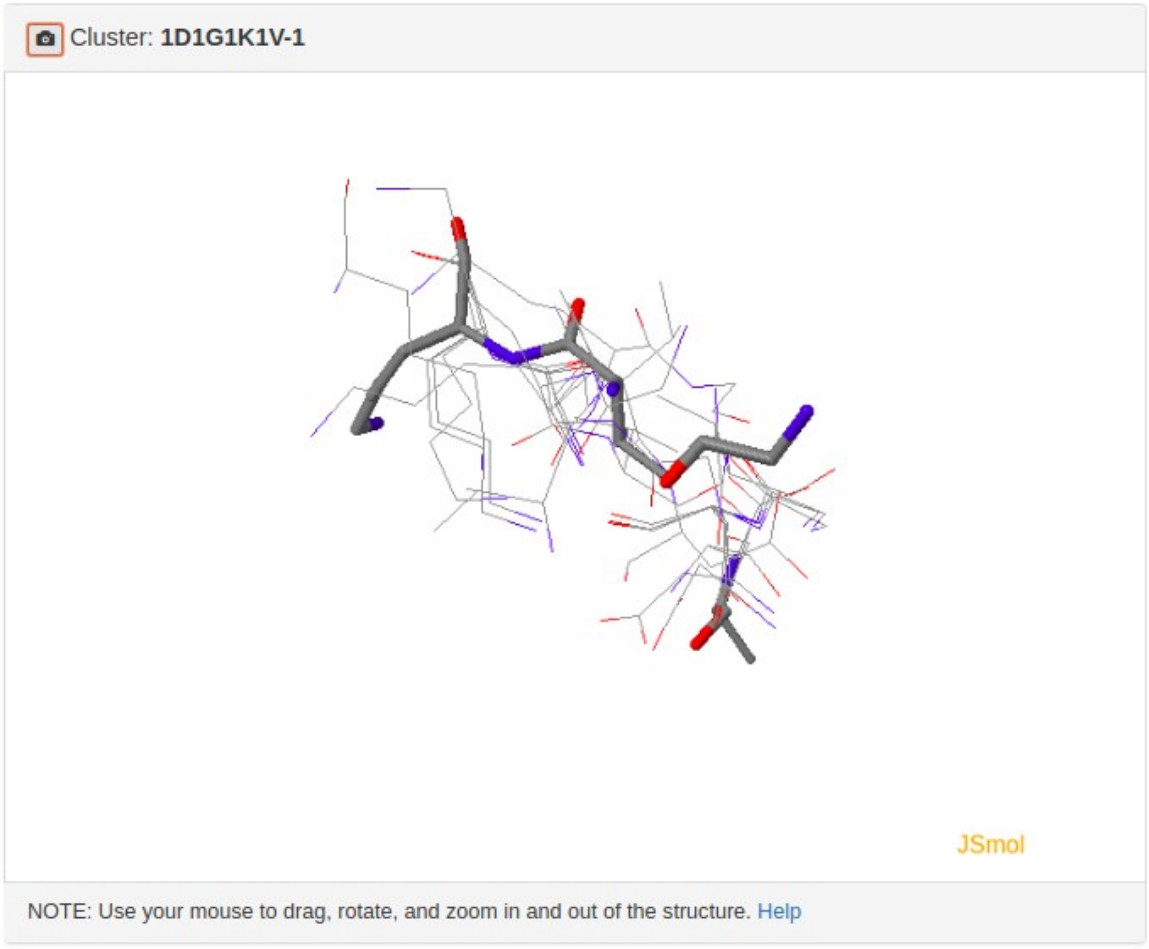


Figure 17 Structural alignment of the sites belonging to the cluster 1D1G1K1V-1.

III) Using homology models

The user can submit to 3DPP a set of protein structures generated through in silico methodologies. For this example, we obtained some structures deposited on the Swiss Model Repository (https://swissmodel.expasy.org/repository). In this website, we searched by the name of a protein of interest, for example, "Superoxide dismutase". After that, we selected four homology models of the same structure but different organisms:

1) P20379: Superoxide dismutase [Cu-Zn] of *Caulobacter vibrioides.*

2) Q9WU84 : Superoxide dismutase [Cu-Zn] of *Mus musculus.*

3) P00441: Superoxide dismutase [Cu-Zn] of *Homo sapiens.*

4) P61851: Superoxide dismutase [Cu-Zn] of *Drosophila melanogaster.*

After download (or generates your own) the models, the PDB files must be compressed as a ZIP format (Figure 18) and uploaded in the input page of 3DPP (Figure 19).

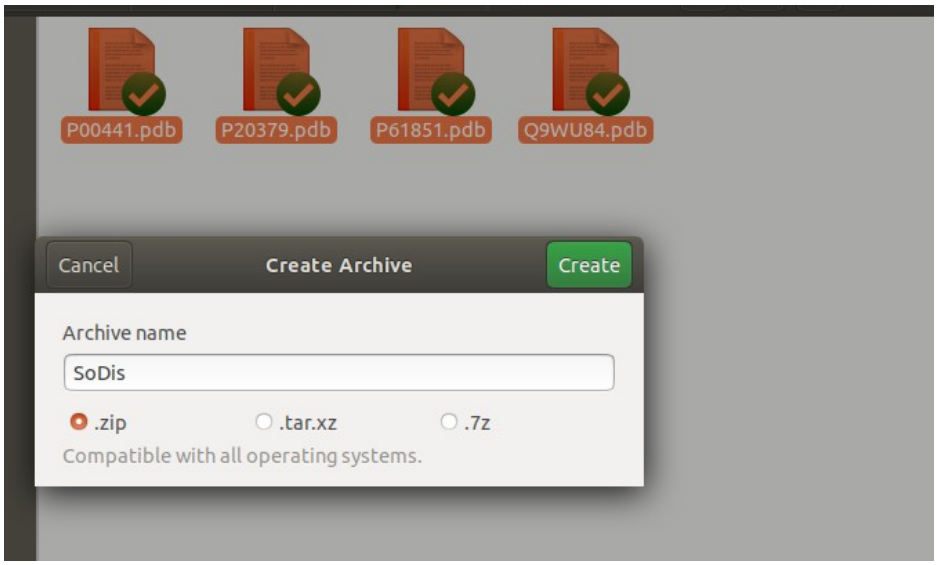


Figure 18 Compression of the four homology models used.

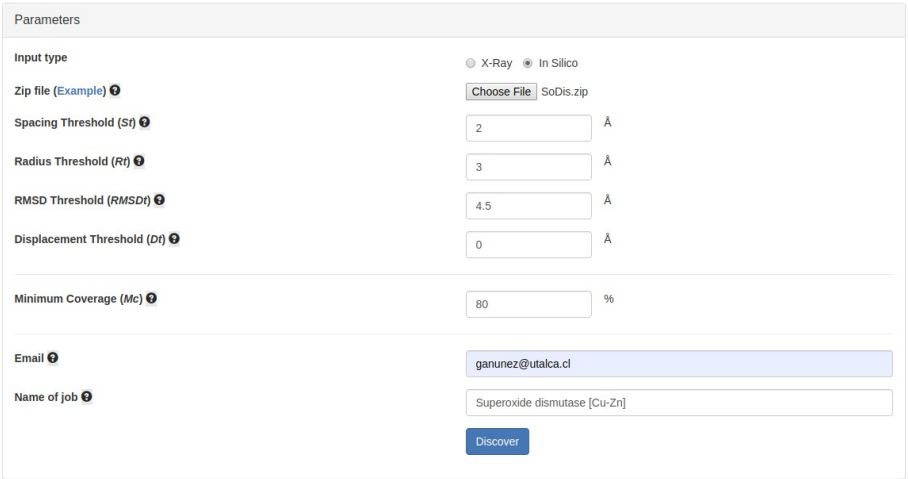


Figure 19 For homology models, the user must to select "In Silico" and upload the compressed file.

For the case of in silico structures, the results will be the same that would be generated using PDBids coming from the Protein Data Bank. In the case of the “Superoxide dismutase”, only two 3D patterns were detected (Figure 20).

Figure 20 3D-patterns detected for the structures of Superoxide dismutase of different organisms.

The first 3D-pattern on the list has a 100% of protein coverage and a 100% of cluster coverage. In this case, for example, the unique cluster of the 3D-pattern 1A1G2I shows an unsorted localisation of the residues on the sequences (Figure 21), but a conserved structural alignment (Figure 22). This type of results gives relevance to our software because it is capable of discovering similar sites that can not be identified with sequence-based methods.

ID	chain	5	29	112	147	148	149	171	173	185	186	194	195	214	215	230	231
p00441	B	-	-	-	-	-	-	-	-	ILE	GLY	-	-	ILE	ALA	-	-
p20379	B	-	ALA	-	ILE	-	-	GLY	ILE	-	-	-	-	-	-	-	-
p61851	A	ALA	-	ILE	-	ILE	GLY	-	-	-	-	-	-	-	-	-	-
p61851	B	ALA	-	ILE	-	ILE	GLY	-	-	-	-	-	-	-	-	-	-
q9wu84	B	-	-	-	-	-	-	-	-	-	-	ILE	GLY	-	-	ILE	ALA

Figure 21 Sequence alignment of the sites belonging to the cluster 1A1G2I-1

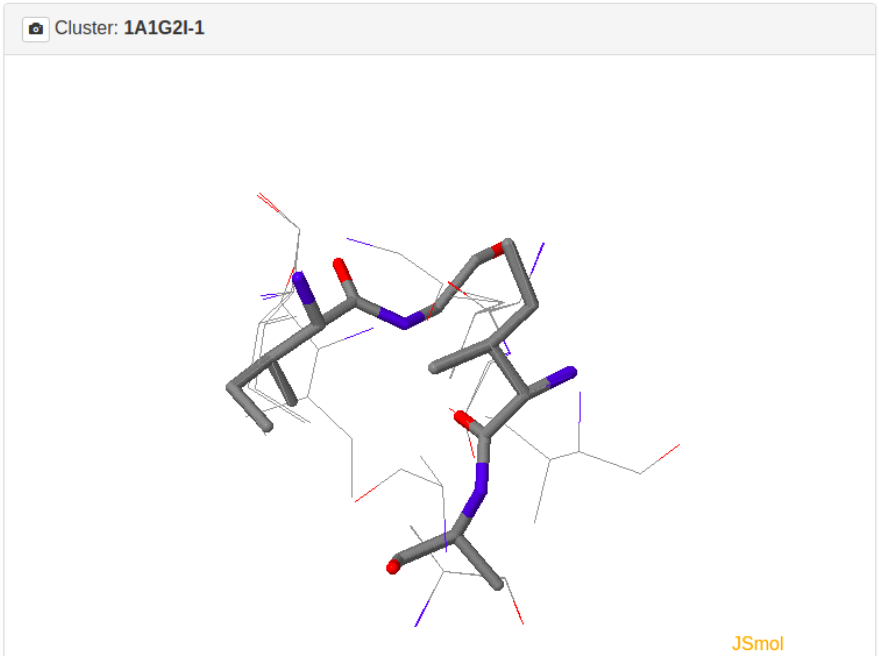


Figure 23 Structural alignment of the sites belonging to the cluster 1A1G2I-1