

Proteus

documentation v1.1

5 DE DEZEMBRO, 2019

Department of Computer Science

Laboratory of Bioinformatics and Systems

In the Greek mythology, Proteus (Πρωτεύς) was Poseidon's son (god of the seas). According to the legend, Proteus had the ability to predict the future. However, whenever someone asked him to make a prediction, he would turn into a monster. Only the bravest were worthy of using knowledge of Proteus. Now, we have brought the Proteus myth to structural bioinformatics, not to predict the future, but to predict pairs of mutations that stabilize regions in enzymes.

Sumário

What is Proteus?	4
http://proteus.dcc.ufmg.br	4
1. Introduction	5
2. Proteus SUITE	6
2.1 ProteusDB	7
2.2 ProteusWEB	10
2.3. PSE (Proteus Search Engine)	17
2.3.1 Defining target triad pairs	17
2.3.2 Search	18
2.3.3 Using structural signatures to reduce the structural comparisons	19
2.3.4 $\Delta\Delta G$ and clash	22
3. Case Studies	23
3.1 β -glucosidase	23
3.2 Protease	26
3.3 NAR (new antigen receptor)	27
3.4 Lipase	28
4. Discussion	29
4.1 Comparison to other tools	29
5. References	32

What is Proteus?

<http://proteus.dcc.ufmg.br>

Proteus is a method, database and a webtool to propose mutations for proteins used in industrial applications. Proteus uses the hypothesis of mutation transference of residue pairs in contact detected in all PDB database to suggest mutations for a target protein.

1. Introduction

Proteus (an acronym for Protein Engineering Supporter) is based on the assumption that if a pair of non-interacting amino acid residues are changed to a pair of interacting amino acids, this could improve protein stability. Furthermore, these mutations only would be allowed if the main-chain conformation of two sets of three amino acids (herein called triad pair) were conserved between the target and the proposed mutations. This triad pair is composed of the amino acid residues in close contact, and the previous and posterior residues. Amino acids that could be mutated are called in our notation n and n' (Figure 1).

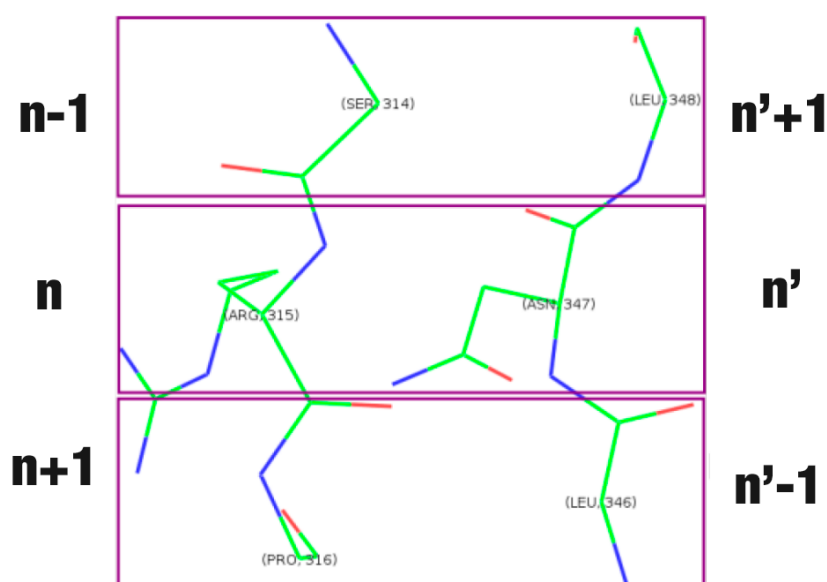


Figure 1. Composition of triad pairs. Amino acids that could make interactions are called n and n' . The anterior and posterior are called $n-1$ and $n+1$ for n , and $n'-1$ and $n'+1$ for n' . In this example, the amino acid R315 is not performing interactions with N347, but they are in a cutoff distance position that could allow interactions if they were changed. The anterior and posterior are S314 and P316 (R315), and L346 and L348 (N347). The Proteus basic assumption is that it is possible to trade the amino acids n and n' to another amino acid pair if the main chain of triad pairs of target and template protein were conserved.

We hypothesized that if the main-chain conformation of triad pairs is not changed upon mutations, *i.e.*, the proposed mutations introduce a pair of interacting residues without modifying the main chain trace of the native triad pairs; then, the substitution would be allowed. We believe that this is possible because of this interaction is a real and possible conformation already observed for a specific pair of interacting residues. Thus, the

interaction might occur in the mutated protein, improving its stability when compared to the native macromolecule (Figure 2).

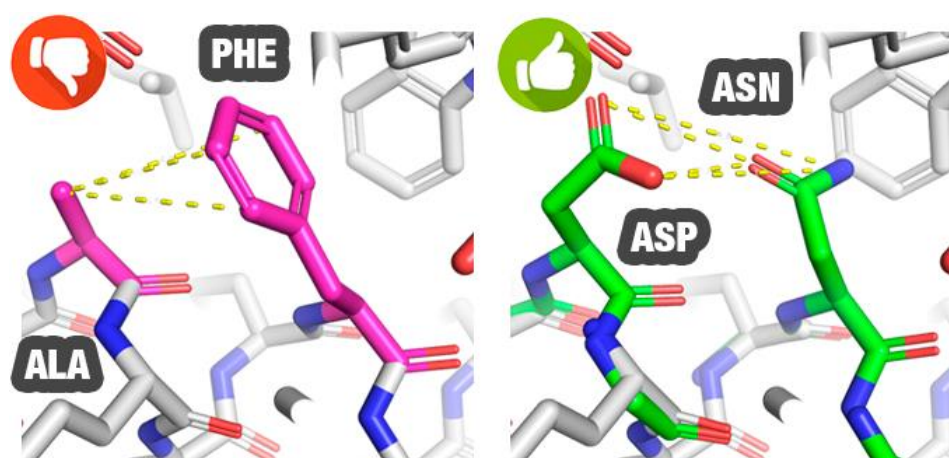


Figure 2. Mutation pair A174D - F151N suggested by Proteus for the beta-glucosidase enzyme (PDB ID: 1BGA).

Aiming to evaluate this strategy, we developed Proteus, a structure-based algorithm. The software receives as input a protein 3D target, and then it selects every two amino acids in close distance to each other (not in direct contact) and their four neighboring residues (a pair for possible mutations within the triad pair; refer to Methods in “ProteusDB” section). A comparison between each selected triad pairs in the target protein and the triad pairs in ProteusDB (main-chain conformation comparison) allows the identification of potential mutation pairs that could be introduced into the target protein, improving its stability without significant conformational changes. This is possible because each triad pairs in ProteusDB is formed by two interacting residues (N and N') and their respective neighboring residues, as collected from all available structures to date at the PDB (Protein Data Bank, available at <<http://www.rcsb.org/pdb>>).

2. Proteus SUITE

Proteus SUITE is divided into three parts (Figure 3):

- (i) **ProteusDB**, a database containing the coordinates of triad pairs formed by pairs of interacting amino acids (n and n') and their neighboring residues (n-1, n+1, n'-1 and n'+1) extracted from known structures deposited at the PDB to date;
- (ii) **ProteusWEB**, a web tool with a user-friendly interface available for running the referred algorithm and which is accessible at <<http://proteus.dcc.ufmg.br>>;
- (iii) **PSE**, method used by Proteus to search for stabilizing mutation pairs in a target protein.

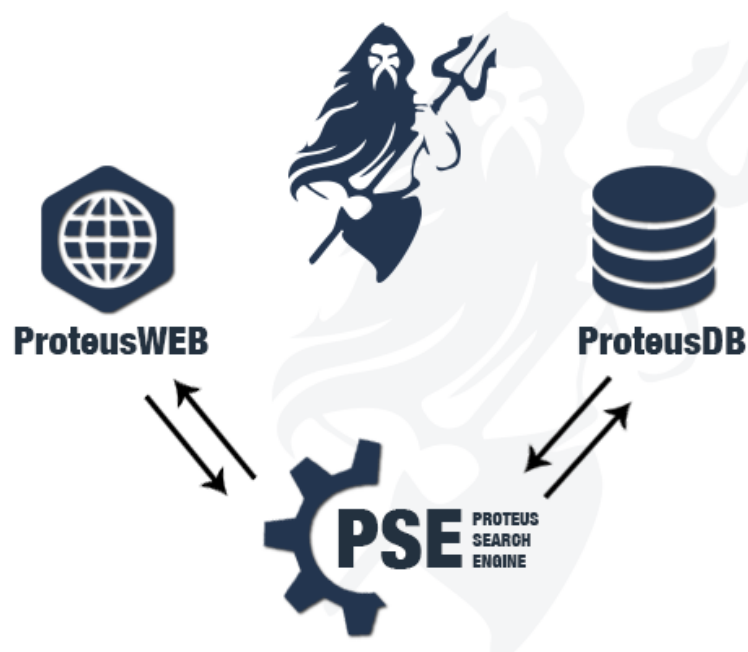


Figure 3. Proteus SUITE. ProteusWEB connects to ProteusDB through the PSE algorithm.

2.1 ProteusDB

To construct the ProteusDB (Proteus DataBank), we used X-ray crystallography derived structures from the PDB with resolution equal or better than 2.0 Å (approximately 52,000 PDB files). Then, each PDB file was processed to extract every triad pairs containing an interacting pair of residues (n and n'). The triad pairs were clustered to remove redundancy and stored in a relational MySQL database (Figure 4).

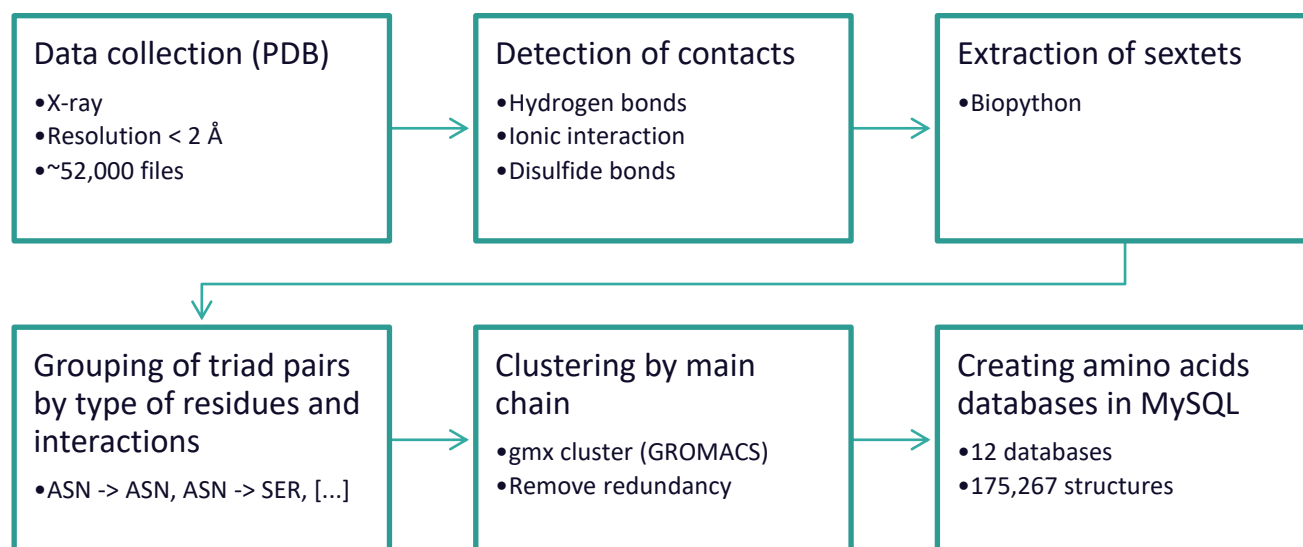


Figure 4. ProteusDB's pipeline.

We used Biopython (Hamelryck and Manderick, 2003; Cock *et al.*, 2009) and in-house scripts to detect the contacts in all three-dimensional structures collected. The list of atoms that could perform interactions, and their respective distance cutoffs, were established by Bickerton *et al.*, (2011) when classifying hydrogen bonds, ionic interactions, and disulfide bonds (Table 1). In the present form, we only collected interactions between atoms from amino acid residues side chains.

Table 1. List of atoms that could perform interactions and distance cutoffs. Adapted from (Bickerton *et al.*, 2011).

Interaction	Distance	Atoms
Hydrogen bond	< 3.50 Å	<ul style="list-style-type: none"> • ARG: NE, NH1, NH2 • ASN: ND2, OD1 • ASP: OD1, OD2 • GLN: NE2, OE1 • LYS: NZ • GLU: OE1, OE2 • HIS: ND1, NE2 • SER: OG • THR: OG1 • TRP: NE1 • TYR: OH
Ionic	< 6.00 Å	<ul style="list-style-type: none"> • ARG: CZ • ASP: CG, OD1, OD2 • GLU: CD • HIS: CD2, CE1, CG
Disulfide bond	< 2.08 Å	<ul style="list-style-type: none"> • CYS: S

For every two amino acid residues involved in the interaction, we collected the coordinates of all their atoms and, in addition, the coordinates of main-chain atoms of their anterior and posterior residues, generating the triad pair. We saved each triad pairs (a file in PDB format) in folders named according to the amino acid pair involved in the interaction (for example, an interaction D100-K200 would be saved in the folder ASP, while an interaction K300-D400 would be saved in the folder LYS). Among all twenty amino acids, twelve of them were considered in this work due to their potential to perform specific interactions as previously described (Table 1): arginine (ARG/R), asparagine (ASN/N), aspartate (ASP/D), glutamine (GLN/Q), lysine (LYS/K), glutamate (GLU/E), histidine (HIS/H), serine (SER/S), threonine (THR/T), tryptophan (TRP/W), tyrosine (TYR/Y), and cysteine

(CYS/C). Therefore, we constructed 122 folders (the combination of the first 11 amino acids, plus one folder to contain disulfide bonds).

To reduce redundancy of triad pairs in ProteusDB, we clustered them. For this step, we used gmx cluster 9 tool from GROMACS 10 software (Berendsen *et al.*, 1995; Van Der Spoel *et al.*, 2005). We used a single linkage algorithm at gmx tool for clustering structures based on the RMSD (Root Mean Square deviation) cutoff ranging from 0.3-0.9 Å (values defined empirically).

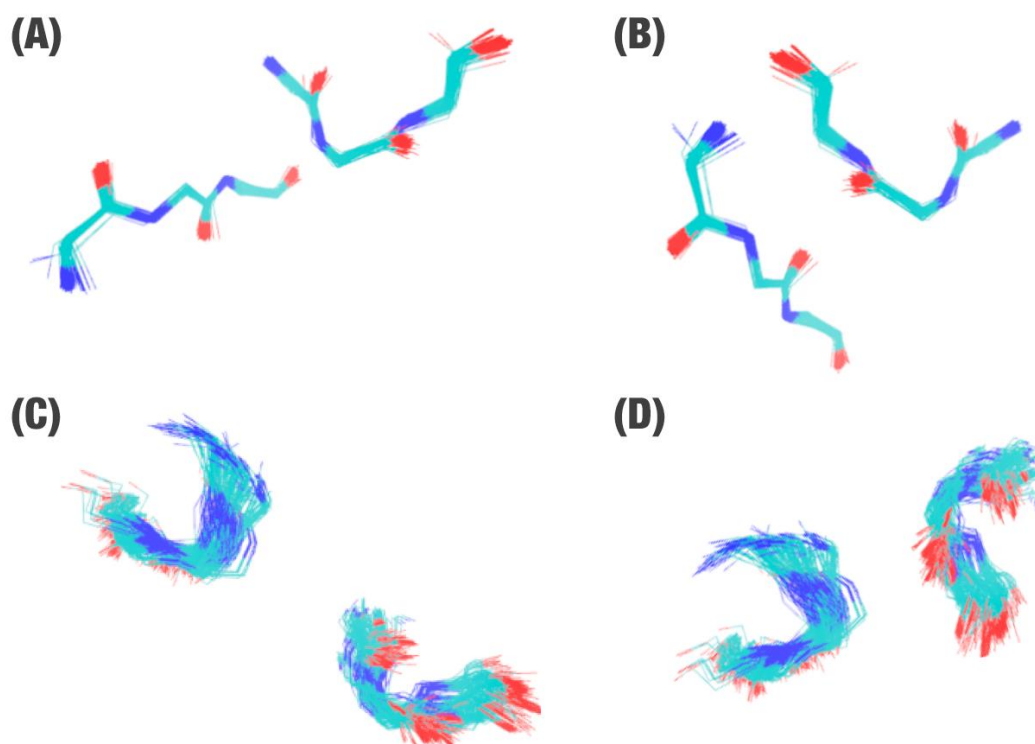


Figure 5. Example of cysteine cluster. In this example, we can see two CYS-CYS clusters (A-B) and two ASP-ARG clusters (C-D), which were obtained from the structural alignment of all triad pairs with an RMSD score of 0.5 Å.

After the clustering step, the first element in each cluster was defined as its representant. To improve the search performance (in Proteus Search Engine; PSE), clusters were regrouped in 12 databases based on the first amino acid in contact (N) and stored in a MySQL database management system. The 12 databases, which together were called the ProteusDB, have a total of 175,267 representants of triad pair structures, being CYS (database of cysteine contacts) the smallest database with 2,552 structures, and ASP (a database with aspartate contacts) the biggest with 28,782 structures (Table 2).

Table 2. The number of representants of triad pair structures in each database that forms ProteusDB.

Amino acid database (ProteusDB)	Number of structures
---------------------------------	----------------------

<i>ARG</i>	28,785
<i>ASN</i>	12,123
<i>ASP</i>	35,980
<i>CYS</i>	2,552
<i>GLN</i>	8,800
<i>GLU</i>	24,828
<i>HIS</i>	15,440
<i>LYS</i>	11,348
<i>SER</i>	13,494
<i>THR</i>	9,947
<i>TRP</i>	2,458
<i>TYR</i>	9,512
<i>Total</i>	175,267

It is important to highlight that the CYS database contains only contacts between CYS-CYS amino acid residues, while all other databases contain contacts among one specific residue and all the other amino acids. For example, the ARG database contains all the following contacts: ARG-ARG, ARG-ASN, ARG-ASP, ARG-GLN, ARG-GLU, ARG-HIS, ARG-LYS, ARG-SER, ARG-THR, ARG-TRP, and ARG-TYR. In the future, we intend to include the contact ARG-CYS.

2.2 ProteusWEB

We constructed a user-friendly interface, herein called ProteusWEB, to provide an easy method to run the Proteus algorithm. ProteusWEB was constructed using the PHP, HTML, and JavaScript languages, and the frameworks CodeIgniter, JQuery, Bootstrap, and 3Dmol.js for visualization of protein structures (Rego and Koes, 2015).

Proteus' homepage shows the main menu with "New project", "Documentation", "About", and "Help" buttons (Figure 6). At the center of the page, there is a brief description of the Proteus Suite with a "Run now!" button that is linked to the "New project" section. Proteus' homepage is available at <<http://proteus.dcc.ufmg.br>>.

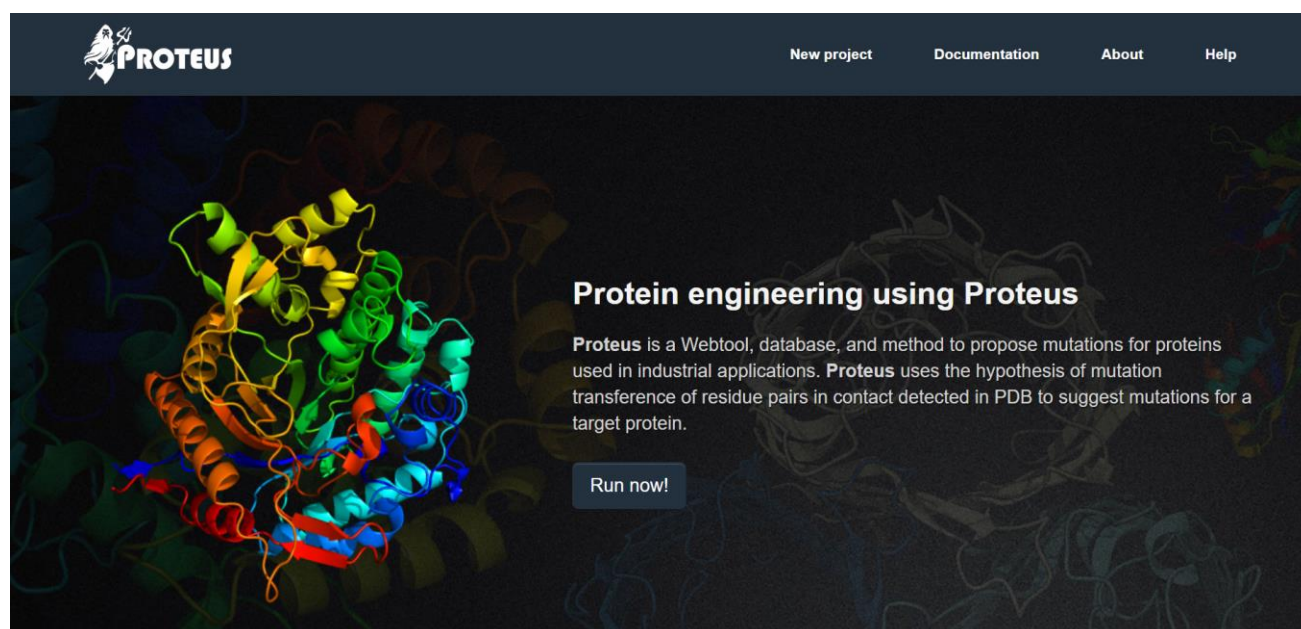
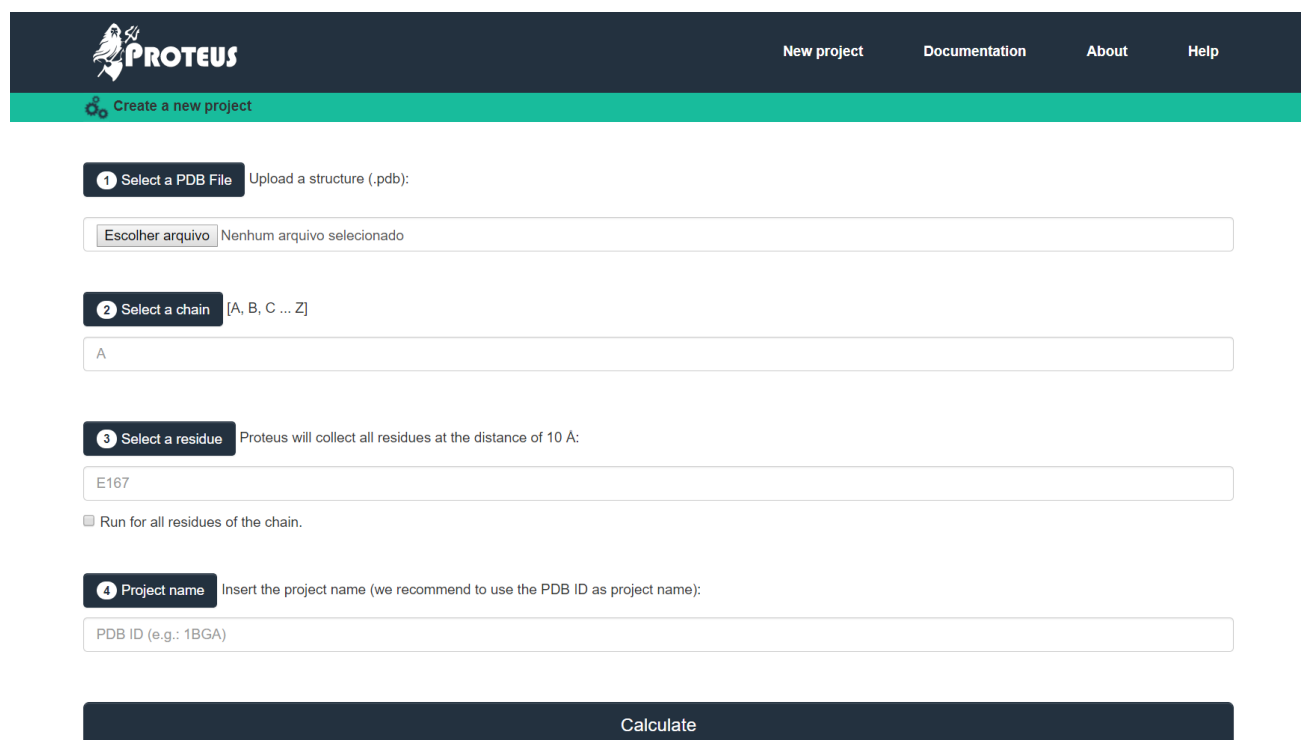


Figure 6. Proteus' homepage. Using the main menu, the users can create a new project (or click on "Run now!"), see the documentation, read more about the creators of Proteus Suite, or click on Help to see Proteus running details. Available at <<http://proteus.dcc.ufmg.br>>.

A new project can receive up to four input parameters to start: (i) an uploaded PDB file (mandatory; files in PDB format can be obtained from the PDB website <<http://www.rcsb.org/pdb>> or prepared by the users); (ii) a specific chain in the uploaded PDB file (mandatory; due to the high computational cost, Proteus analyzes only one chain per run); (iii) by default, Proteus performs its search for all residues in the previously specified chain or, optionally, users can provide an amino acid name (one-letter code) followed by the residue number (e.g. E167) for fast searches (in this case, only residues at a maximum distance of 10 Å from the informed amino acid position will be considered in PSE); and (iv) a project name that will be used to identify the project (Figure 7). After clicking on the "Calculate" button, Proteus creates a new project in the system and executes Python scripts for detecting target triad pairs. The triad pairs are separated in PDB files and PSE is used to compare them to the triad pairs in ProteusDB, aiming to identify potential mutations at residues N and N' that could introduce a new interaction in the engineered protein triad pairs (details will be presented in the PSE section). While the search runs, ProteusWEB

shows a warning message with the link for the project page (that will be available when the process is finished).



The screenshot displays the Proteus web application interface. At the top, there is a dark blue header with the Proteus logo on the left and navigation links for 'New project', 'Documentation', 'About', and 'Help' on the right. Below the header is a teal bar with the text 'Create a new project' and a plus icon. The main content area contains a form with four numbered steps: 1. 'Select a PDB File' with a sub-label 'Upload a structure (.pdb):' and a file selection button labeled 'Escolher arquivo' and the text 'Nenhum arquivo selecionado'. 2. 'Select a chain' with a sub-label '[A, B, C ... Z]' and a text input field containing 'A'. 3. 'Select a residue' with a sub-label 'Proteus will collect all residues at the distance of 10 Å:' and a text input field containing 'E167'. Below this is a checkbox labeled 'Run for all residues of the chain.' which is currently unchecked. 4. 'Project name' with a sub-label 'Insert the project name (we recommend to use the PDB ID as project name):' and a text input field containing 'PDB ID (e.g.: 1BGA)'. At the bottom of the form is a large dark blue button labeled 'Calculate'.

Figure 7. Proteus' new project section. Four input parameters are required as input: (i) a PDB file (mandatory); (ii) an amino acid chain (mandatory; character, e.g. A); (iii) a residue identifier (optional; for fast search; e.g. E167); and (iv) a project name (optional).

At the PSE routine, Proteus redirects the user to the project page (Figure 8). The project page is divided into two sections: (i) the header, that includes the project name (in gray background) and the number of mutation pairs suggested by the algorithm (in green background); and (ii) the body, that includes the list of suggested mutations, a summary button, a set of filters, free search box and, at the right side, a 3D visualization window which shows the target protein structure (we also included a button to allow downloading wild and mutant PDB structures).

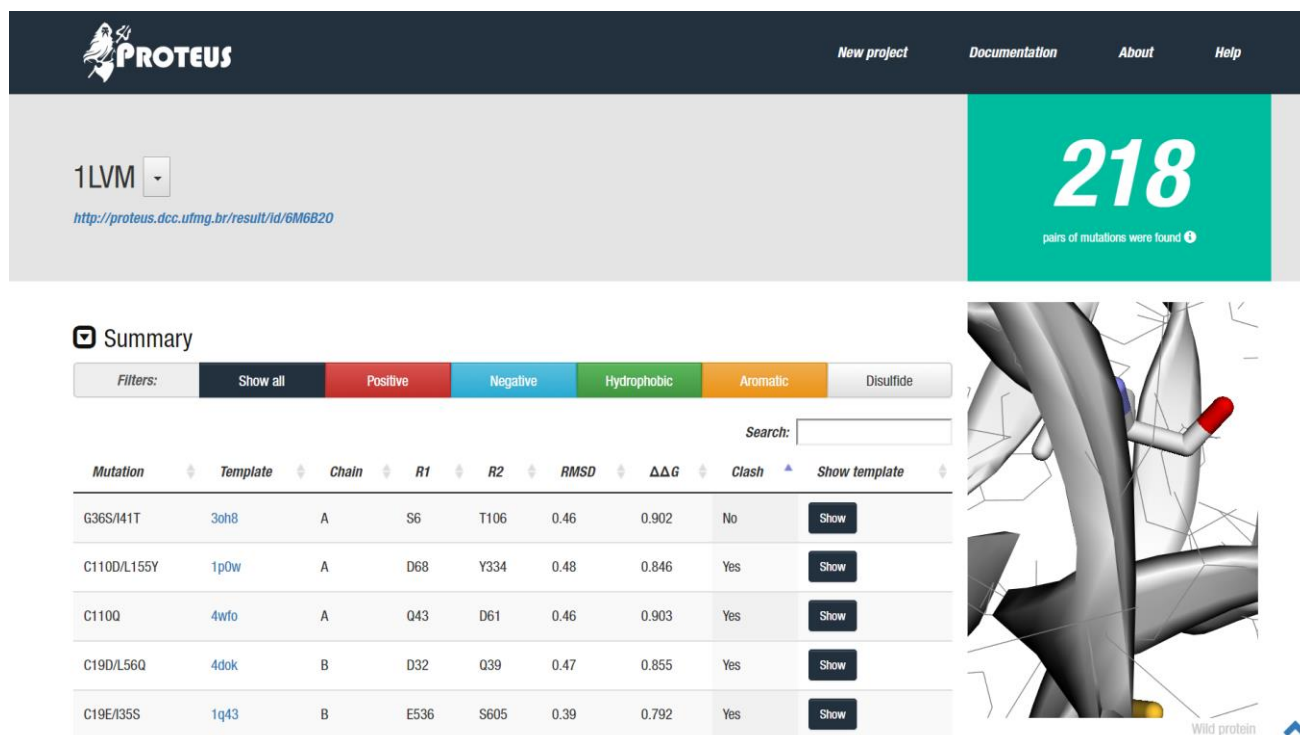


Figure 8. Project's page. At the top of the page, users can see the project's name with a link to the page, and the number of mutation pairs suggested by the algorithm. Users can utilize a list of filters ("Show all", "Positive", "Negative", "Hydrophobic", "Aromatic", and "Disulfide") and routine parameters ("Mutation", "Template", "Chain", " $n = R1$ ", " $n = R2$ ", RMSD, $\Delta\Delta G$, and Clash) to sort the proposed mutations according to their own interests. On the right side, users can use the 3D-visualization window to analyze the three-dimensional structure of the target protein (zoom in, zoom out, rotate and translate). If users want to have a closer view of a specific suggested mutation pair, he/she can click on the respective "Show" button, and both the native and suggested triad pairs (with the mutation) are shown superimposed to each other in a separate window.

In addition, next to the project's name, a dropdown button allows the users to download the results, such as the list of proposed mutation pairs (in CSV format), the mutation pairs grouped by site, and the PDB file used in the project (Figure 9).

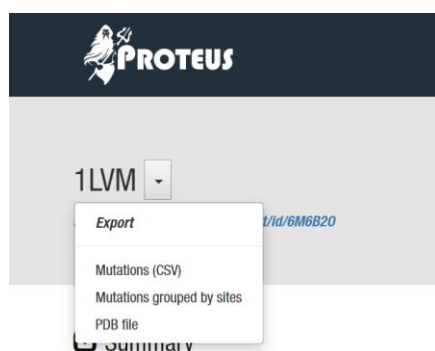
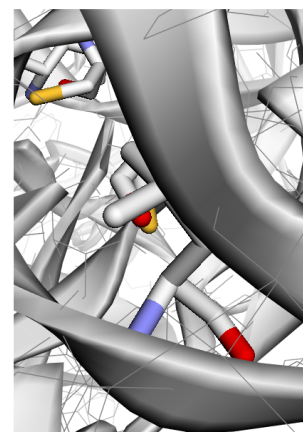


Figure 9. Export section. The user can export all mutations suggested as a CSV file or mutations grouped by sites. Also, they can export the PDB file used in the project.

Below, the summary dropdown button, when clicked on, shows the list of proposed mutations grouped by sites (Figure 10).

Summary

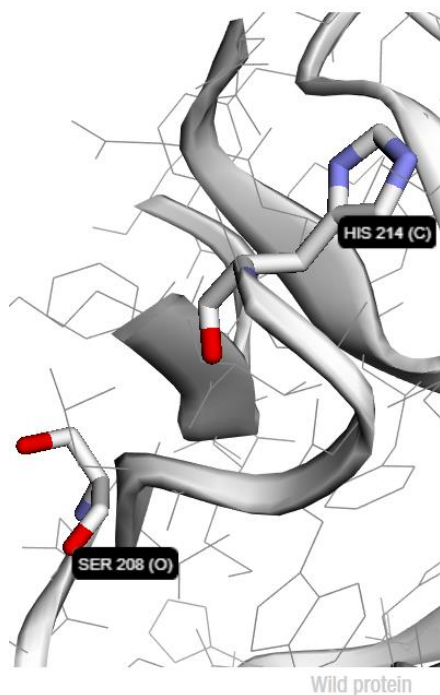
Sites	Mutation suggested
C110/L155	C110D/L155Y
E24/H28	E24N/H28S
T17/G34	T17D/G34Y
I14/T17	I14W/T17E I14Y/T17E
G165/S181	G165D/S181Q
W140/W140	W140D/W140Y W140T/W140K
V164/T180	V164D/T180E V164K/T180R
L21/T29	L21Q/T29E
L21/Y33	L21H/Y33N
N23/L56	N23Q/L56Y N23Y/L56W N23Y/L56Y
V156/D160	V156T
H20/L21	H20W/L21H
S208/H214	S208N/H214T



Wild protein

Figure 10. The summary button shows a list of mutations summarized by sites. At the right side, 3Dmol zooms the amino acid pairs into the three-dimensional protein structure,

When the user hovers over each row, at the right side is shown a three-dimensional visualization of the protein structure, highlighting the amino acid residues target for the mutations (Figure 11).



Wild protein

Figure 11. 3Dmol's visualization of the three-dimensional structure. The user can click on atoms to obtain a label with the amino acid name (three-letter code), number, and atom name.

In the filter panel, the user can filter the results based on amino acid types, such as positively charged, negatively charged, hydrophobic, aromatics, and disulfide bonds (Figure 12).

Summary

Filters:		Show all	Positive	Negative	Hydrophobic	Aromatic	Disulfide		
								Search:	<input type="text"/>
Mutation	Template	Chain	R1	R2	RMSD	$\Delta\Delta G$	Clash	Show template	
C19K/L60K	3os7	B	K127	K161	0.47	0.867	Yes	Show	
C19T/Y33K	2p3j	A	T176	K187	0.38	0.842	Yes	Show	

Figure 12. Filter functions based on the amino acid type. Clicking on “Positive” for example, ProteusWEB will show only mutations for positive amino acids (such as the lysine showed in the example).

If no mutation attends to the filter requirements, a warning message informs “no matching records found” (Figure 13).

Summary

Filters:		Show all	Positive	Negative	Hydrophobic	Aromatic	Disulfide										
							Search:	<input type="text"/>									
Mutation	⬆	Template	⬆	Chain	⬆	R1	⬆	R2	⬆	RMSD	⬆	ΔΔG	⬆	Clash	⬆	Show template	⬆
No matching records found																	
Showing 0 to 0 of 0 entries (filtered from 218 total entries)																	

Figure 13. When clicked on a filter without mutations proposed, ProteusWEB returns the warning “No matching records found”. In this example, Proteus did not found mutations for disulfide bonds.

Also, ProteusWEB provides a free search box, which could be used to search for residue numbers, amino acids one code letter, template PDB ID, chain, amino acid residue positions for suggesting mutations, numerical values of RMSD and $\Delta\Delta G$, and if the mutations insert a stereochemistry clash (Figure 14).

Filters:		Show all	Positive	Negative	Hydrophobic	Aromatic	Disulfide		
								Search:	186
Mutation	Template	Chain	R1	R2	RMSD	$\Delta\Delta G$	Clash	Show template	
F186K/M187H	2ifq	A	K39	H43	0.43	0.763	Yes	Show	
F186R/M187H	4xv8	A	R48	H52	0.48	0.77	Yes	Show	
F186R/M187R	1qtx	B	R12	R16	0.41	0.762	Yes	Show	
F186T/M187H	3vpc	C	T148	H152	0.49	0.688	Yes	Show	

Showing 1 to 4 of 4 entries (filtered from 218 total entries)

Figure 14. Free search method. This section allows free searches, for example, for a residue number (as the example), amino acid one-letter code, template code or any other field.

Stereochemistry clash is a limitation of the Proteus method (Figure 15). This occurs because of mutations for bulkier amino acids could insert atoms in regions already occupied by atoms of neighbor's amino acids, which would require conformational changes in the region. This contraries Proteus' principles, once we hypothesized that the conservation in the main chain conformation in known structures is which allows the amino acids to be changed. However, a better indication of the mutation impact only could be inferred by methods with higher computational costs, like molecular dynamics, which hardly could be executed on a large scale.

								Search:	
Mutation	Template	Chain	R1	R2	RMSD	$\Delta\Delta G$	Clash	Show template	
G36S/I41T	3oh8	A	S6	T106	0.46	0.902	No	Show	
C110D/L155Y	1p0w	A	D68	Y334	0.48	0.846	Yes	Show	
C110Q	4wfo	A	Q43	D61	0.46	0.903	Yes	Show	

Figure 15. Clicking on the column name, Proteus orders the outcome (ascending or descending). In the example, we ordered the column clash to separate the results without stereochemistry clash.

Lastly, Proteus allows users to visualize the alignment between the triad pairs wild (target protein) and the triad pairs obtained from the known structure (Figure 16). The alignment demonstrates possible differences caused by the substitution of both amino acids. Also, on the right side, Proteus shows the complete structure of the target protein. Allied to other

information showed by Proteus, this visualization could help users to decide what mutation would be unusual for an experimental test.

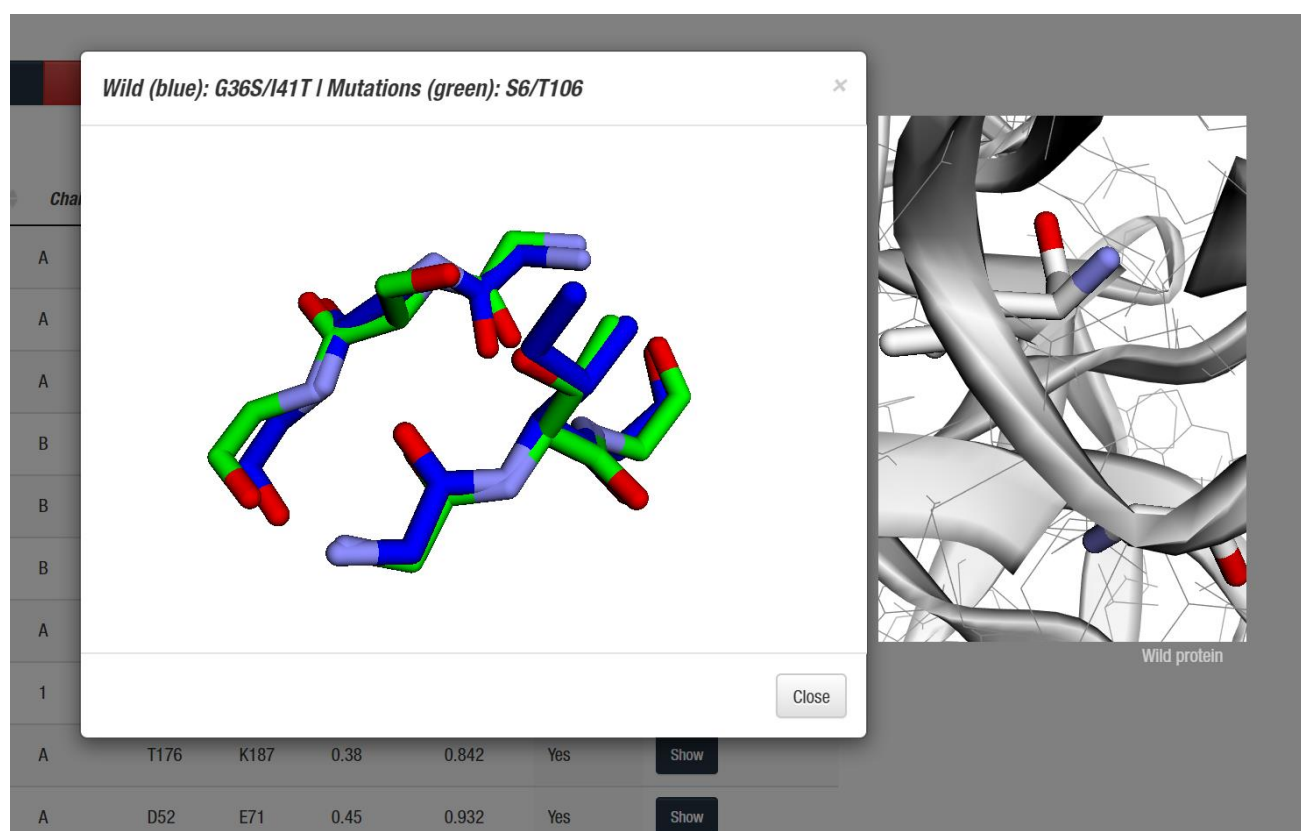


Figure 16. Structural alignment modal. It shows the alignment between the triad pairs of wild and mutant (obtained from the template). On the right side, it is available the complete structure of the target protein.

2.3. PSE (Proteus Search Engine)

2.3.1 Defining target triad pairs

To define target triad pairs for starting the search, Proteus analyzes the distance between all alpha carbon atoms of the protein. Proteus defines amino acid pairs, which are not neighbors and have alpha carbons at a distance between 3.35 and 16.4Å are a target for mutations (Figure 17). This distance range was defined based on the minimum and maximum distance value of alpha carbons of residues interacting found in ProteusDB.

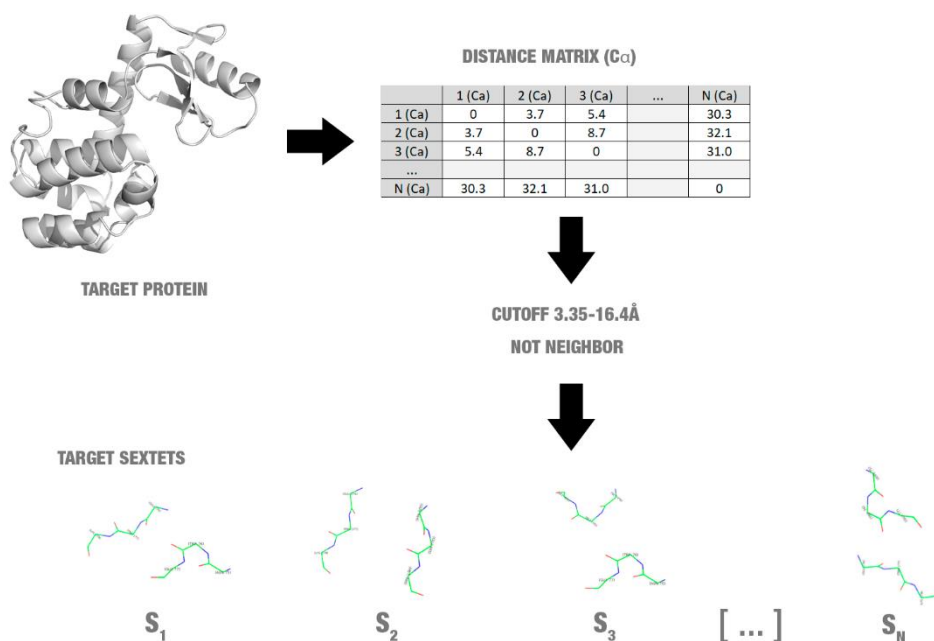


Figure 17. Detection of target triad pairs.

2.3.2 Search

To determine if a mutation could be suggested, Proteus performs structural alignment between triad pairs of a target protein and the triad pairs of ProteusDB. Structural alignment allows the comparisons between the shape of macromolecules. We hypothesized that if the main chain conformation of six residues is conserved, a double mutation is possible (Figure 18).

Wild (blue): C19S/I35Y | Mutations (green): S94/Y190

×

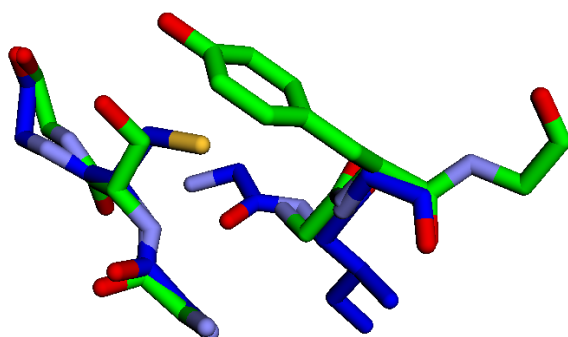


Figure 18. Example of structural alignment between triad pairs. In this case, Proteus suggested mutations for the sites C19-I35 (blue sticks) of the crystal structure of tobacco etch virus protease (PDB ID: 1q31). Proteus detected overlap with RMSD of 0.46 between the C19-I35 residues and triad pairs formed between S94-Y190 (green sticks) of a membrane protein (PDB ID: 1iiw). For this reason, Proteus suggested the mutation C19S-I35Y.

The differences between structures are evaluated using RMSD (Root Mean Square deviation). RMSD is the average distance between the atoms of a structure when superposed to another (equation 1). The lower this value, the most similar are both structures.

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^N \delta_i^2}$$

(1)

Where δ_i is the difference between coordinates of an atom i and the equivalent in another structure for N atoms. Proteus algorithm compares the 24 atoms of each triad pairs (C, N, O, and C α from each main chain of six amino acids).

However, for each target triad pairs, Proteus had to perform 175,267 comparisons between structures (total of ProteusDB's structures), which requires high computational costs. The structural alignment is a vital step for Proteus to suggest mutations. Therefore, this step cannot be removed, but some strategies with lower computational costs could be used to reduce the number of structural alignments removing for this comparison, triad pairs with low possibilities to present similar structures. For this step, we used structural signatures (also called fingerprints).

2.3.3 Using structural signatures to reduce the structural comparisons

After preliminary tests, we observed that the structural alignment between a target triad pair and all structures from ProteusDB presented a high computational cost. For instance, the structure of bacteriophage T4 lysozyme (PDB ID: 2LZM) presents a sequence with 164 amino acids and approximately 1,000 amino acid pairs target for mutation suggestions based on a preliminary Proteus analysis. The structural alignment between 2LZM's target triad pairs and all structures of ProteusDB taken almost one week using 32 CPUs (data not shown). However, we noticed that some comparisons performed by structural alignment were unnecessary. For instance, the distance between the main chain atoms for the same

amino acids is almost conserved (distances between C, N, O, and C α), but the structural alignment promotes these comparisons.

Hence, to reduce the computational costs, we introduced a filter step before the structural alignment using the SSV (Structural Signature Variation) method (Mariano *et al.*, 2019). SSV is a graph-based methodology for three-dimensional structure comparisons using structural signatures, linear algebra techniques, and the variation of vector distances. The SSV algorithm suggests that molecules with similar structures have similar structural signatures. Therefore, the Euclidian distance between signatures of different macromolecules could be used to define if a macromolecule is more like a model macromolecule than another one (Mariano *et al.*, 2019). Hence, we introduced the SSV filter before the structural alignments step (Figure 19).

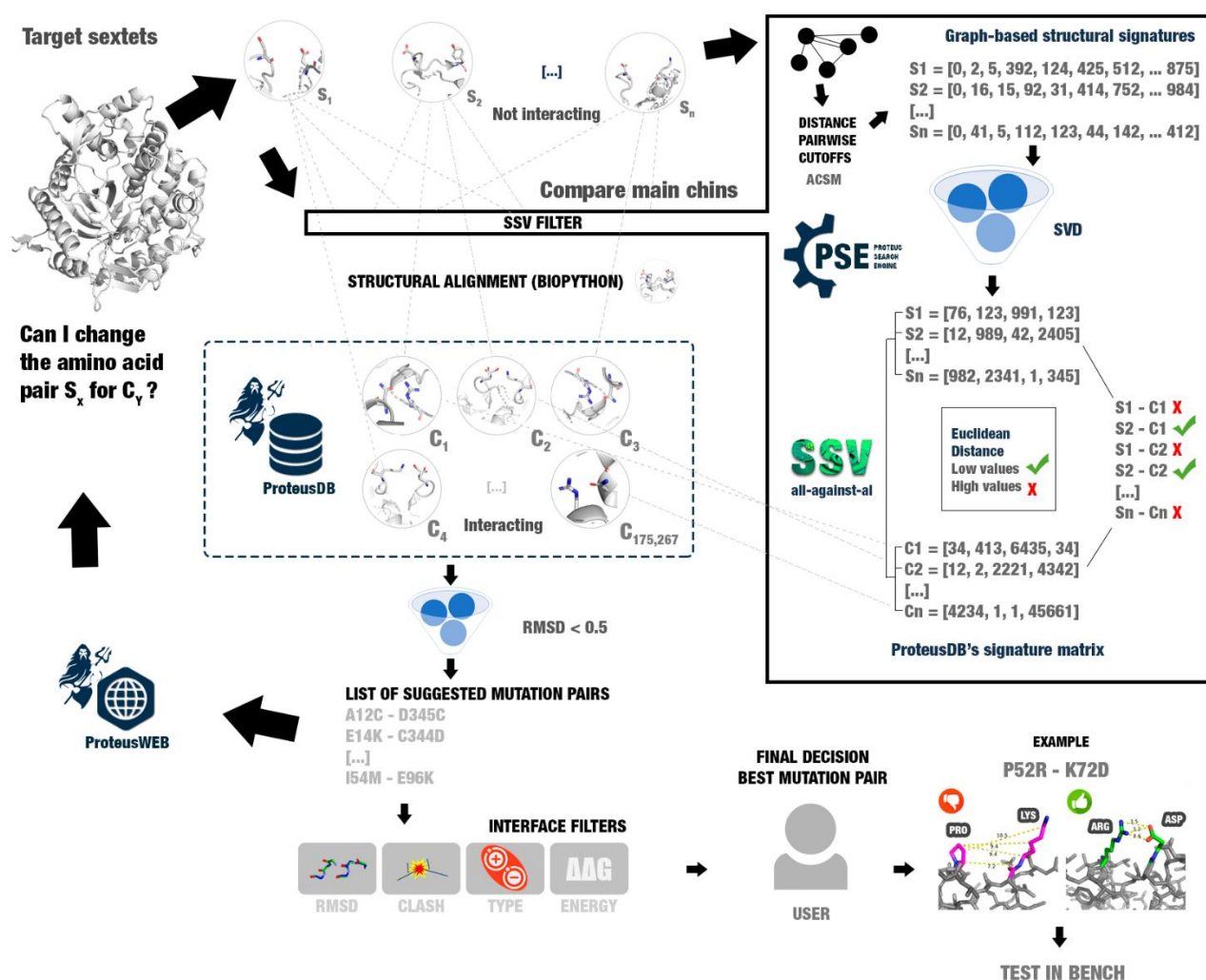


Figure 19. Proteus complete pipeline. Structural alignment has a high computational cost. Hence, we used a filter step to reduce the number of structural alignments. The filter step uses the variation of structural signatures (fingerprints) to remove structures from ProteusDB with

a low possibility to be like target triad pairs. SSV (structural signature variation) uses Euclidean distance between four-dimensional vectors, which has lower computational cost than structural alignment. This accelerates the search for similar structures considerably.

SSV uses aCSM (atomic Cutoff Scanning Matrix) to construct structural signatures (Pires *et al.*, 2013). In aCSM method, a protein structure is converted in a graph, where the atoms are the vertices and edges are defined by a set of distance cutoffs between atoms. The pairwise atoms are calculated, analyzing also the pharmacophore atom properties. Lastly, a signature vector (fingerprint) that represents the protein structure is constructed.

To construct the ProteusDB triad pairs signature, we used aCSM with parameters cutoff minimum distance of 0 Å, a maximum cutoff distance of 10 Å, and a cutoff distance step of 0.1 Å – parameters defined in (Mariano *et al.*, 2019). This generated a matrix of 175,267 lines (structures) and 576 columns (features).

We constructed signature vectors for each triad pair of ProteusDB. Besides, to reduce noise and boost the search process, we performed dimensional reduction using SVD (Singular Value Decomposition). SVD is a technique from linear algebra for noise-reducing based on analysis of multivariate data and rearrangement of the vector space for retrieve non-evident relationships between the matrix elements (Silvério-Machado *et al.*, 2015). In SVD, a real or complex matrix A (of length $m \times n$) is factored in three other matrixes, where the matrix U is an $m \times m$ unitary matrix, S is an $m \times n$ rectangular diagonal matrix with the singular values, and V is a $n \times n$ transposed unitary matrix (equation 2).

$$A = USV^T$$

(2)

The plot of matrix S was used to analyze the number of singular values. We calculated that four dimensions were necessary to represent the ProteusDB's signature matrix (70% of the singular values presented in matrix S), *i.e.* the 576 representative values for each triad pair were reduced for just four. Also, we used SVD matrixes to calculate the reduced vectors for each new target triad pairs (see methodology details in Silvério-Machado *et al.*, 2015; Pires *et al.*, 2013).

To calculate similarity, SSV used Euclidean distance between target triad pairs reduced signatures and all reduced signatures of ProteusDB. Despite SSV presented a fast comparison method with a high number of true positives, we observed a high number of false positives (data not shown). Therefore, we used SSV only as a filter to reduce the number of structural comparisons. For example, in the case study with 2LZM protein, the

SSV filter reduced the number of probable structures with considerable signature differences. For each triad pairs defined as a possible target for mutations, Proteus did not need to perform a structural alignment with each element of ProteusDB. Using a simple Euclidean distance between the reduced signature vector, we were able to define if a structure has a chance to be like another. We defined that a maximum distance of cutoff was 13 (maximum distance used to include all possible mutations suggested for 2LZM). An Euclidean distance between a reduced signature and all ProteusDB could be executed in less than one second and reduced the number of structural alignments from 700 million to 2 million. Thus, we obtained the same result in less than one day using only one CPU (data not shown).

Although the search for mutations using only structural alignment and using SSV followed of a reduced search for similar main chain structures have presented the same result in our tests, with considerable difference of running time, we cannot affirm that the SSV filter could remove a true positive result for the structural alignment step in another protein. However, we believe that the SSV filter turns viable the large-scale use of Proteus as a required for a web server application.

2.3.4 $\Delta\Delta G$ and clash

Proteus shows a Gibbs free energy estimation calculated by the MAESTRO command-line tool (Laimer *et al.*, 2015). MAESTRO receives as input a PDB and a pair of mutations. MAESTRO returns a $\Delta\Delta G$ predicted value, which represents the total predicted change of stability (kcal/mol). $\Delta\Delta G$ values lower than zero indicates stabilizing mutations and $\Delta\Delta G$ values higher than zero destabilizing.

Furthermore, stereochemistry clash is detected when the performed changing of the side chain atoms for the suggested mutations, inserts another atom a non-allowed position. To detect this, we calculated the Euclidean distance between the coordinates of the new atoms and the neighborhood. We defined that when an inserted atom crosses a cutoff distance of 2 Å of any one atom from the neighborhood, a possible stereochemistry clash is reported.

Despite the main objective of Proteus is to suggest stabilizing mutations and without clash, we expect most of the predicted mutations would present destabilizing changes or stereochemistry clash. However, we decided to show these results to give the users a higher

decision power. Mutations that insert stereochemistry clashes could be better evaluated using high computational cost strategies, like molecular dynamics.

3. Case Studies

To evaluate Proteus, we performed four case studies: (i) *Bacillus polymyxa* GH1 β -glucosidase (PDB ID: 1BGA); (ii) tobacco etch virus protease (PDB ID: 1LVB); (iii) antigen receptor variable domain from sharks (PDB ID: 2YWY); and (iv) lipase II from *Rhizopus niveus* (PDB ID: 1LGY). These structures were submitted to Proteus for suggesting mutations (Table 3).

Table 3. Proteus results for the case studies.

#	Protein	PDB ID	ProteusID	Sequence length	Number of mutations suggested	Number of mutations with $\Delta\Delta G < 0$	Number of mutations without clash
1	β -glucosidase (hydrolase)	1BGA	XZVFYD	447	2065	874	16
2	Protease	1LVB	P359JG	243	178	19	1
3	NAR	2YWY	KT1ORV	113	204	5	2
4	Lipase	1LGY	7DN8R8	269	961	116	29

3.1 β -glucosidase

β -glucosidases are vital enzymes in the second-generation biofuel production. They act cleaving cellobiose in two molecules of glucose, which will be used in fermentation for producing bioethanol (Mariano *et al.*, 2017). The design of thermostable and more efficient

β -glucosidases enzymes has excellent value for the industry. For this case study, we run Proteus for the *Bacillus polymyxa* GH1 β -glucosidase (PDB ID: 1BGA).

For 1BGA, Proteus suggested 2065 mutations pairs, being 874 with predicted $\Delta\Delta G$ lower than zero, and 16 without stereochemistry clash (Table 3; Table 4).

Table 4. List of mutations suggested for 1BGA (filtered by no clash).

Mutation	Template	Chain	R1	R2	RMSD	$\Delta\Delta G$	Clash
A147D/F151N	4nac	B	D65	N69	0.43	0.619	No
A147H/F151N	4hmm	B	H177	N181	0.23	0.665	No
A147H/F151S	1p7g	H	H87	S91	0.38	0.738	No
A147N/F151E	4dno	B	N233	E237	0.39	0.137	No
A147S/F151D	4iaq	B	S37	D41	0.49	0.821	No
A147S/F151H	4toh	B	S126	H130	0.25	0.714	No
A147S/F151K	4kn8	B	S26	K30	0.16	1.17	No
A147S/F151Y	3r3r	A	S159	Y163	0.43	0.167	No
A237H/R241N	4hmm	B	H177	N181	0.29	-0.367	No
A240N/S244T	3bol	B	N190	T194	0.39	-0.348	No
G194S/L198H	4toh	B	S126	H130	0.26	0.369	No

G194T/L198R	3mux	A	T233	R237	0.49	0.321	No
G279S/I283R	4ise	A	S418	R422	0.26	-0.082	No
G332S/V336D	4iqg	B	S37	D41	0.42	0.153	No
G97S/Y101H	4toh	B	S126	H130	0.26	0.574	No
G97T/Y101H	3vpc	C	T148	H152	0.28	0.279	No

Full table available at <<http://proteus.dcc.ufmg.br/result/id/XZVFYD>>.

We highlighted the mutation A237H/R241N with a predicted $\Delta\Delta G$ of -0.367 and without clashes (Figure 20). The insertion of positive amino acids in the surface of *Bacillus polymyxa* β -glucosidase has been described in the literature as positive for thermostability increasing (Lopez-Camacho *et al.*, 1996). The mutant E96K of *B. polymyxa* β -glucosidase is reported to keep kinetic parameters with increased resistance to denaturing agents, such as pH and urea. In addition, Proteus predicted four mutations for positive amino acids in this site: E96R/D99W, E96R/D99R, E96Q/D99R, and E96N/D99Y.

Wild (blue): A237H/R241N | Mutations (green): H177/N181

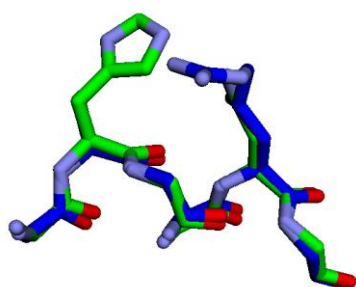


Figure 20. Mutation A237H-R241N for 1BGA.

3.2 Protease

Also, we run Proteus for a catalytically inactive tobacco etch virus protease (PDB ID: 1LVB). Proteus found 178 possible sites for mutation, being 19 with $\Delta\Delta G$ lower than zero, but only one without clash (Table 5).

Table 5. List of mutations suggested for 1LVB (filtered by no clash or negative $\Delta\Delta G$).

Mutation	Template	Chain	R1	R2	RMSD	$\Delta\Delta G$	Clash
G36S/I41T	3oh8	A	S6	T106	0.48	0.93	No
S170H/T173Y	4hq8	B	H200	Y204	0.35	-0.458	Yes
Q58W/G62H	2qqz	A	W49	H58	0.42	-0.371	Yes
L56H/G62K	5h8t	A	H116	K127	0.46	-0.338	Yes
K99R/G161R	2hhc	A	R5	R34	0.49	-0.313	Yes
Q58K/G62R	2amu	A	K57	R73	0.38	-0.262	Yes
Q58W/G62Y	2ny2	A	W375	Y384	0.41	-0.256	Yes
P13W/S16H	1i5z	A	W13	H17	0.46	-0.218	Yes
K141Q	2jjo	A	Q292	N313	0.47	-0.187	Yes
Q58Y/F64Y	4s1l	A	Y105	Y164	0.45	-0.168	Yes

Q58S/G62Y	1k8k	D	S166	Y222	0.5	-0.132	Yes
H167Y/S168K	3i7u	B	Y61	K81	0.41	-0.13	Yes
N23Y/L56Y	4s1l	A	Y105	Y164	0.37	-0.109	Yes
E24N/H28S	3o41	L	N161	S177	0.5	-0.105	Yes
Q58Y/F64W	4h2a	A	Y23	W65	0.5	-0.076	Yes
T22S/T30Y	1k8k	D	S166	Y222	0.42	-0.047	Yes
Q58W/G62D	2v8h	C	W251	D320	0.35	-0.034	Yes
N23Y/L56W	4h2a	A	Y23	W65	0.4	-0.022	Yes
L56H/G62Y	3t8x	C	H207	Y244	0.48	-0.015	Yes
L56W/G62H	3i24	B	W34	H93	0.45	-0.013	Yes

Full table available at <<http://proteus.dcc.ufmg.br/result/id/P359JG>>.

3.3 NAR (new antigen receptor)

We run Proteus for a new antigen receptor variable domain from sharks (PDB ID: 2YWY). Proteus found 204 possible sites for mutation, being five with $\Delta\Delta G$ lower than zero, but only two without clash (Table 6).

Table 6. List of mutations suggested for 2YWY (filtered by no clash or negative $\Delta\Delta G$).

Mutation	Template	Chain	R1	R2	RMSD	$\Delta\Delta G$	Clash
G15S/R74Q	3v6o	D	S17	Q88	0.38	0.746	No
Y86S/F87T	3oh8	A	S6	T106	0.45	4.813	No
N21R/E57R	5a02	D	R251	R273	0.5	-0.472	Yes
T34W/G84Y	2ny2	A	W375	Y384	0.48	-0.333	Yes
T34Y/G84W	1o26	D	Y84	W160	0.29	-0.333	Yes
N60R	5a03	E	E209	R213	0.48	-0.209	Yes
Q5R/S65R	5a02	D	R251	R273	0.49	-0.195	Yes

Full table available at: <<http://proteus.dcc.ufmg.br/result/id/KT1ORV>>.

3.4 Lipase

We run Proteus for the lipase II from *Rhizopus niveus* (PDB ID: 1LGY). Proteus found 961 possible sites for mutation, being 116 with $\Delta\Delta G$ lower than zero and 29 without clash (Table 7).

Table 7. List of mutations suggested for 1LGY (filtered by no clash or negative $\Delta\Delta G$).

Mutation	Template	Chain	R1	R2	RMSD	$\Delta\Delta G$	Clash
G111R/S115W	1w35	A	R163	W167	0.36	-0.788	No
G111Y/S115T	4cae	C	Y88	T92	0.3	-0.605	No

G111Q/S115Y	1c27	A	Q130	Y134	0.23	-0.468	No
G111Q/S115R	2x1c	B	Q196	R200	0.24	-0.282	No
G111K/S115H	2ifq	A	K39	H43	0.28	-0.191	No
G111T/S115R	3mux	A	T233	R237	0.48	-0.13	No

Full table available at: <<http://proteus.dcc.ufmg.br/result/id/7DN8R8>>.

4. Discussion

In this paper, we proposed a platform to support protein engineering, called Proteus (PROTein Engineering Supporter). Proteus intended to become the first step in the mutation prediction process for proteins that do not have previous mutation studies, and which aim to introduce new intra-chain contacts without alteration of protein conformation. Our tool evaluates residues in a wild protein and suggests, based on the alignments between the pair residues and contacts in query databases, possible mutations in the residue pair in order to introduce a new (intra-chain) interaction between the proposed residues without altering the main conformation of the polypeptide chain.

Proteus suggests mutations in residue pairs if the main chain of these amino acids and their posterior and anterior neighbors are overlapped (cutoff below the set value of 0.5 Å), ensuring the conformation. The proposed mutations are based on other high-resolution three-dimensional structures, experimentally resolved and available in the PDB. Hence, if a given conformation of two amino acid residues and their target protein neighbors is found in the database, so it is suggested that these residues may be replaced.

4.1 Comparison to other tools

Several methods, tools, and web servers have been developed to propose mutations in target proteins. The first computational strategies aimed to insert new disulfide bonds in such macromolecules, aiming to improve their thermostability. For instance, SSBOUND (Hazes and Dijkstra, 1988) and MODIP software (Sowdhamini *et al.*, 1989) use a classification

system based on conformation parameters for disulfide bonds (as distances of 1.87 Å for C β -S, 2.04 Å for S-S, and angles of 114° for C α C β S and 104° for C β SS) to suggest locations where the introduction of a disulfide cross-link could lead to protein stabilization.

With the advent of the new sequencing technologies, an abundant number of primary structures from several proteins in different organisms allowed the emergence of new sequence-based strategies for the prediction of beneficial mutations in target proteins. For instance, MuStab is a sequence-based tool that uses machine learning to predict protein stability changes upon amino acid substitutions (Teng *et al.*, 2010). Another example is iPTREE-STAB web server that aims to predict protein stability changes upon single amino acid substitutions (Huang *et al.*, 2007).

Structure-based approaches have also been used to predict mutation impacts. PopMuSiC webtool predicts thermodynamic stability changes using a linear combination of statistical potentials based on the solvent accessibility of the mutated residue (Dehouck *et al.*, 2011). SDM web server uses substitution probability tables obtained from known 3-D structures to analyze the variation of amino acid replacements tolerated within the family of homologous proteins (Pandurangan *et al.*, 2017). mCSM uses graph-based signatures and pharmacophore properties in a machine learning approach to predict stabilizing and destabilizing mutations (Pires *et al.*, 2014b). Also, SDM and mCSM methods were combined in a hybrid approach called DUET, which tries to obtain a more accurate prediction of the free energy change; $\Delta\Delta G$ (Pires *et al.*, 2014a). SSV is another method that uses graph-based structural signatures and compares Euclidian distances between vectors to predict if mutations introduce similar characteristics to reference proteins (Mariano *et al.*, 2019). Lastly, MAESTRO is a web server that combines high-throughput scanning for multi-point mutations, prediction of free energy change, and stabilizing disulfide bonds (Laimer *et al.*, 2015).

Proteus presents a new strategy to suggest mutations using known structures from PDB. For this reason, we compared the functionalities of Proteus and other mutation tools (Table 8). Proteus' primary objective is suggesting mutations based on known structures. For this, Proteus searches for target sites in the protein structure for possible mutations, which differ Proteus from sequence-based methods, such as MODIP and Mustab. To predict the variation of Gibbs free energy, Proteus uses the MAESTRO command line. Hence, Proteus' objective is different from tools in which the main objective is to predict $\Delta\Delta G$, such as SDM,

mCSM, DUET, and even, MAESTRO webtool. Also, Proteus uses SSV to provide faster searches. In conclusion, Proteus can be used together with other tools for suggesting more accurate mutations.

Table 8. Comparison of mutation tools.

TOOL	INPUT	DESCRIPTION	REFERENCE
PROTEUS	Structures	Suggests mutation pairs based on known structures found in PDB.	-
SSBOUND	Structures	Inserts new disulfide bonds, which could improve the thermostability.	(Hazes and Dijkstra, 1988)
MODIP	Structures	It uses a classification system based on conformation parameters for disulfide bonds.	(Sowdhamini <i>et al.</i> , 1989)
MUSTAB	Sequence	Uses machine learning to predict protein stability changes upon amino acid substitutions.	(Teng <i>et al.</i> , 2010)
IPTREE-STAB	Sequence	Predicts protein stability changes ($\Delta\Delta G$) upon single amino acid substitutions.	(Huang <i>et al.</i> , 2007)
POPMUSIC	Structures	Predicts thermodynamic stability changes using a linear combination of statistical potentials based on the solvent accessibility of the mutated residue.	(Dehouck <i>et al.</i> , 2011)
SDM	Structures	uses substitution probability tables obtained from known 3-D structures to analyze the variation of amino acid replacements tolerated within the family of homologous proteins.	(Pandurangan <i>et al.</i> , 2017)
DUET	Structures	Combines the methods SDM and mCSM in a hybrid approach to try to obtain a more accurate prediction of $\Delta\Delta G$.	(Pires <i>et al.</i> , 2014a)

MCSM	Structures	mCSM uses graph-based signatures (Pires <i>et al.</i> , 2014b) and pharmacophore properties in a machine learning approach to predict stabilizing and destabilizing mutations.
SSV	Structures	SSV uses graph-based structural signatures and compares Euclidian distances between vectors to predict if mutations insert similar characteristics to reference proteins.
MAESTRO	Structures	Combines high throughput scanning for multi-point mutations, prediction of free energy change ($\Delta\Delta G$) values, and stabilizing disulfide bonds. (Laimer <i>et al.</i> , 2015)

5. References

- Berendsen, H.J.C. *et al.* (1995) GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, 91, 43–56.
- Bickerton, G.R. *et al.* (2011) Comprehensive, atomic-level characterization of structurally characterized protein-protein interactions: the PICCOLO database. *BMC Bioinformatics*, 12, 313.
- Cock, P.J.A. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25, 1422–1423.
- Dehouck, Y. *et al.* (2011) PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinformatics*, 12, 151.
- Hamelryck, T. and Manderick, B. (2003) PDB file parser and structure class implemented in Python. *Bioinformatics*, 19, 2308–2310.
- Hazes, B. and Dijkstra, B.W. (1988) Model building of disulfide bonds in proteins with known three-dimensional structure. *Protein Eng.*, 2, 119–125.
- Huang, L.-T. *et al.* (2007) iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. *Bioinformatics*, 23, 1292–1293.
- Laimer, J. *et al.* (2015) MAESTRO--multi agent stability prediction upon point mutations. *BMC Bioinformatics*, 16, 116.

-
- Lopez-Camacho,C. *et al.* (1996) Amino acid substitutions enhancing thermostability of *Bacillus polymyxa* beta-glucosidase A. *Biochem. J.*, 314 (Pt 3), 833–838.
- Mariano,D. *et al.* (2019) A Computational Method to Propose Mutations in Enzymes Based on Structural Signature Variation (SSV). *Int J Mol Sci*, 20.
- Mariano,D.C.B. *et al.* (2017) Characterization of glucose-tolerant β -glucosidases used in biofuel production under the bioinformatics perspective: a systematic review. *Genet. Mol. Res.*, 16.
- Pandurangan,A.P. *et al.* (2017) SDM: a server for predicting effects of mutations on protein stability. *Nucleic Acids Res*, 45, W229–W235.
- Pires,D.E.V. *et al.* (2013) aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction. *Bioinformatics*, 29, 855–861.
- Pires,D.E.V. *et al.* (2014a) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res.*, 42, W314-319.
- Pires,D.E.V. *et al.* (2014b) mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, 30, 335–342.
- Rego,N. and Koes,D. (2015) 3Dmol.js: molecular visualization with WebGL. *Bioinformatics*, 31, 1322–1324.
- Silvério-Machado,R. *et al.* (2015) Retrieval of Enterobacteriaceae drug targets using singular value decomposition. *Bioinformatics*, 31, 1267–1273.
- Sowdhamini,R. *et al.* (1989) Stereochemical modeling of disulfide bridges. Criteria for introduction into proteins by site-directed mutagenesis. *Protein Eng.*, 3, 95–103.
- Teng,S. *et al.* (2010) Sequence feature-based prediction of protein stability changes upon amino acid substitutions. *BMC Genomics*, 11, S5.
- Van Der Spoel,D. *et al.* (2005) GROMACS: fast, flexible, and free. *J Comput Chem*, 26, 1701–1718.

