PROFEAT Tutorial

Protein 3D-Structure Descriptor

Table of Contents

(A) Computational Flowchart 2

(B) List of Protein 3D-Structure Descriptors 3

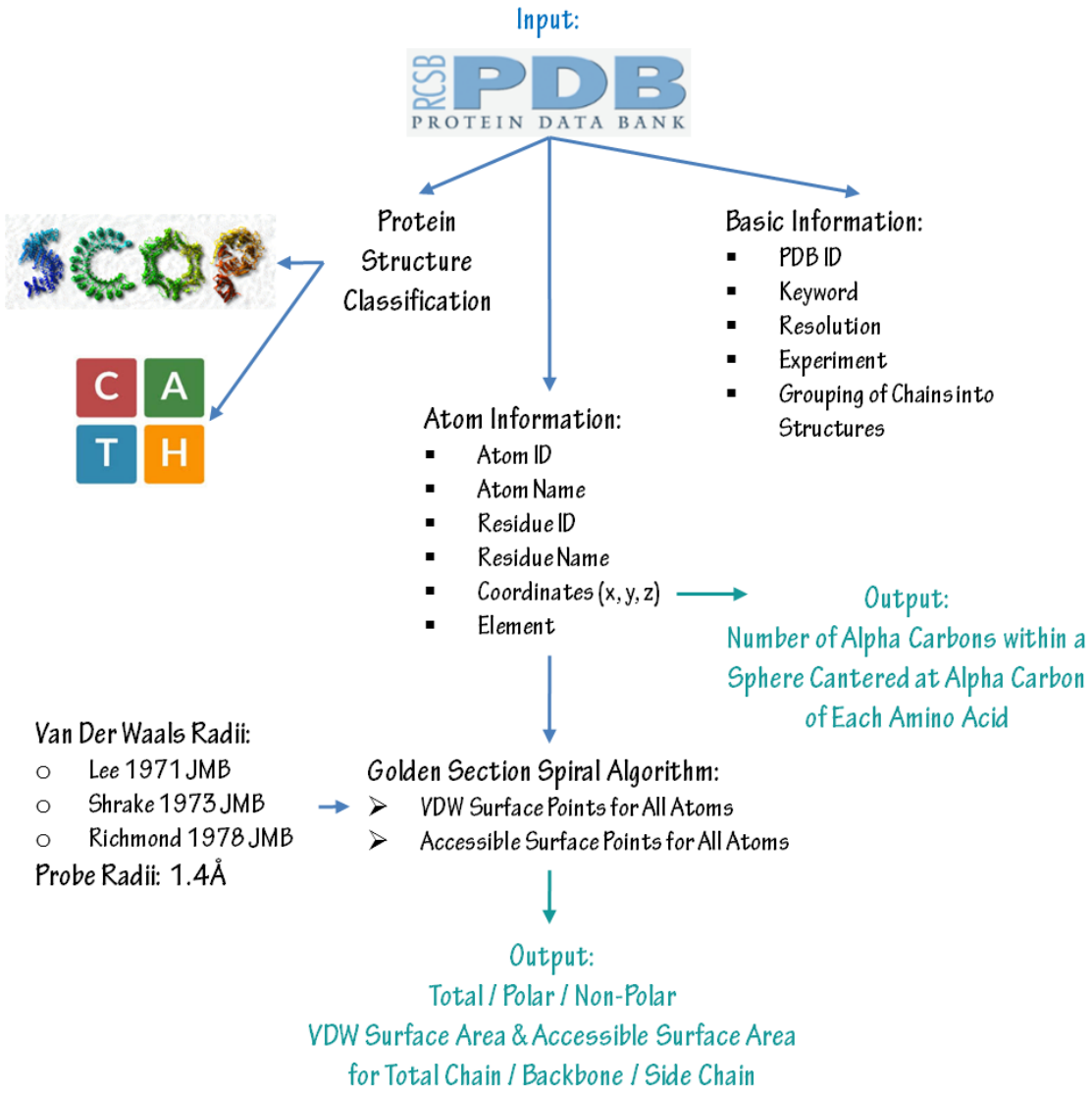
(C) Illustration of Computational Methodology 5

(D) Sample Input & Output 9

Prepared by: Zhang Peng (zhangpeng1202@gmail.com), 31 Dec 2016

(A) Computational Flowchart

Figure 1 | Computational flowchart for PROFEAT protein 3D structure descriptors



(B) List of Protein 3D-Structure Descriptors

Following the feature group indexing in PROFEAT web server, each protein 3D structure descriptor is indexed as (*X, Y, Z*). When *X = G13*, it stands for the global features of the protein structure, *Y* represent the adopted atomic van der Waals radii (**Table 1**) (*Y=1*: Lee 1971 JMB), (*Y=2*: Shrake 1973 JMB), and (*Y=3*: Richmond 1978 JMB), and Z represents each specific structural descriptors (**Table 2**). When *X = G14*, it stands for the vector of local features for each residue or each atom, such that the dimension of the vector is the number of residues. Currently, PROFEAT calculates the sphere-based descriptor: the number of alpha carbons within a sphere centered at alpha carbon of each residue, at the radii of 8Å, 10Å, 12Å, 14Å, 16Å respectively.

Table 1 | Three different sets of van der Waals radii adopted in PROFEAT

VDW radii	Lee 1971 JMB (Y=1)	Shrake 1973 JMB (Y=2)	Richmond 1978 JMB (Y=3)
C	1.7	2.0	1.9
N	1.55	1.5	1.7
O	1.52	1.4	1.4
S	1.8	1.85	1.8

Table 2 | The list of currently provided protein 3D-structure descriptors in PROFEAT

NO.	Index	Protein 3D-Structure Descriptor	Dimension
1	(G13.1.1)	(Lee's) Total Van Der Waals Surface Area	1
2	(G13.1.2)	(Lee's) Total Polar Van Der Waals Surface Area	1
3	(G13.1.3)	(Lee's) Total Non-Polar Van Der Waals Surface Area	1
4	(G13.1.4)	(Lee's) Backbone Van Der Waals Surface Area	1
5	(G13.1.5)	(Lee's) Backbone Polar Van Der Waals Surface Area	1
6	(G13.1.6)	(Lee's) Backbone Non-Polar Van Der Waals Surface Area	1
7	(G13.1.7)	(Lee's) Sidechain Van Der Waals Surface Area	1
8	(G13.1.8)	(Lee's) Sidechain Polar Van Der Waals Surface Area	1
9	(G13.1.9)	(Lee's) Sidechain Non-Polar Van Der Waals Surface Area	1
10	(G13.1.10)	(Lee's) Total Accessible Surface Area	1
11	(G13.1.11)	(Lee's) Total Polar Accessible Surface Area	1
12	(G13.1.12)	(Lee's) Total Non-Polar Accessible Surface Area	1
13	(G13.1.13)	(Lee's) Backbone Accessible Surface Area	1
14	(G13.1.14)	(Lee's) Backbone Polar Accessible Surface Area	1
15	(G13.1.15)	(Lee's) Backbone Non-Polar Accessible Surface Area	1
16	(G13.1.16)	(Lee's) Sidechain Van Der Waals Surface Area	1
17	(G13.1.17)	(Lee's) Sidechain Polar Van Der Waals Surface Area	1
18	(G13.1.18)	(Lee's) Sidechain Non-Polar Van Der Waals Surface Area	1

19	(G13.2.1)	(Shrake's) Total Van Der Waals Surface Area	1
20	(G13.2.2)	(Shrake's) Total Polar Van Der Waals Surface Area	1
21	(G13.2.3)	(Shrake's) Total Non-Polar Van Der Waals Surface Area	1
22	(G13.2.4)	(Shrake's) Backbone Van Der Waals Surface Area	1
23	(G13.2.5)	(Shrake's) Backbone Polar Van Der Waals Surface Area	1
24	(G13.2.6)	(Shrake's) Backbone Non-Polar Van Der Waals Surface Area	1
25	(G13.2.7)	(Shrake's) Sidechain Van Der Waals Surface Area	1
26	(G13.2.8)	(Shrake's) Sidechain Polar Van Der Waals Surface Area	1
27	(G13.2.9)	(Shrake's) Sidechain Non-Polar Van Der Waals Surface Area	1
28	(G13.2.10)	(Shrake's) Total Accessible Surface Area	1
29	(G13.2.11)	(Shrake's) Total Polar Accessible Surface Area	1
30	(G13.2.12)	(Shrake's) Total Non-Polar Accessible Surface Area	1
31	(G13.2.13)	(Shrake's) Backbone Accessible Surface Area	1
32	(G13.2.14)	(Shrake's) Backbone Polar Accessible Surface Area	1
33	(G13.2.15)	(Shrake's) Backbone Non-Polar Accessible Surface Area	1
34	(G13.2.16)	(Shrake's) Sidechain Accessible Surface Area	1
35	(G13.2.17)	(Shrake's) Sidechain Polar Accessible Surface Area	1
36	(G13.2.18)	(Shrake's) Sidechain Non-Polar Accessible Surface Area	1
37	(G13.3.1)	(Richmond's) Total Van Der Waals Surface Area	1
38	(G13.3.2)	(Richmond's) Total Polar Van Der Waals Surface Area	1
39	(G13.3.3)	(Richmond's) Total Non-Polar Van Der Waals Surface Area	1
40	(G13.3.4)	(Richmond's) Backbone Van Der Waals Surface Area	1
41	(G13.3.5)	(Richmond's) Backbone Polar Van Der Waals Surface Area	1
42	(G13.3.6)	(Richmond's) Backbone Non-Polar Van Der Waals Surface Area	1
43	(G13.3.7)	(Richmond's) Sidechain Van Der Waals Surface Area	1
44	(G13.3.8)	(Richmond's) Sidechain Polar Van Der Waals Surface Area	1
45	(G13.3.9)	(Richmond's) Sidechain Non-Polar Van Der Waals Surface Area	1
46	(G13.3.10)	(Richmond's) Total Accessible Surface Area	1
47	(G13.3.11)	(Richmond's) Total Polar Accessible Surface Area	1
48	(G13.3.12)	(Richmond's) Total Non-Polar Accessible Surface Area	1
49	(G13.3.13)	(Richmond's) Backbone Accessible Surface Area	1
50	(G13.3.14)	(Richmond's) Backbone Polar Accessible Surface Area	1
51	(G13.3.15)	(Richmond's) Backbone Non-Polar Accessible Surface Area	1
52	(G13.3.16)	(Richmond's) Sidechain Accessible Surface Area	1
53	(G13.3.17)	(Richmond's) Sidechain Polar Accessible Surface Area	1
54	(G13.3.18)	(Richmond's) Sidechain Non-Polar Accessible Surface Area	1
55	(G14.1.1)	Number of alpha carbons within a sphere centered at alpha carbon of each residue, at (r = 8 Å)	# residue
56	(G14.1.2)	Number of alpha carbons within a sphere centered at alpha carbon of each residue, at (r = 10 Å)	# residue
57	(G14.1.3)	Number of alpha carbons within a sphere centered at alpha carbon of each residue, at (r = 12 Å)	# residue
58	(G14.1.4)	Number of alpha carbons within a sphere centered at alpha carbon of each residue, at (r = 14 Å)	# residue
59	(G14.1.5)	Number of alpha carbons within a sphere centered at alpha carbon of each residue, at (r = 16 Å)	# residue

(C) Illustration of Computational Methodology

Based on the structure of PDB file format (http://www.wwpdb.org/documentation/file-format), PROFEAT reads the input PDB file by:

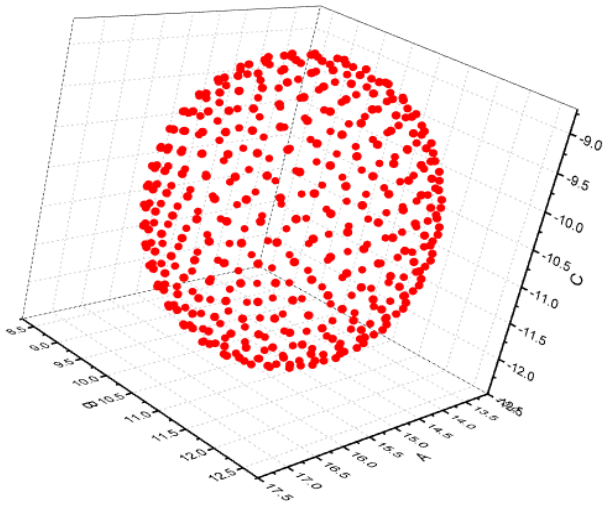
(1) Extracting the wanted information (e.g. resolution, experiment, keywords, etc.), and then creating a structure ID to represent the chains having the same structure.

(2) Providing the SCOP and CATH classifications by matching against the latest SCOP (http://scop.berkeley.edu/) and CATH (http://www.cathdb.info/) databases. SCOP classifies each protein structure into Class, Fold, Superfamily, Family, Protein Domain, Species, and Domain. CATH classifies each protein structure into Class, Architecture, Topology/Fold, and Homologous Superfamily.

(3) Computing the structure descriptors by reading the XYZ coordinates and the vdw radii of each atom, generating its vdw surface points evenly distributed over the sphere surface, by applying Golden Section Spiral algorithm (**Figure 2** illustrates the generation of vdw surface points for a single atom, and **Figure 3** illustrates the overlap of vdw surface between two adjacent atoms), and then followed by determining if each surface point is buried inside or exposed outside (**Figure 4**), such that vdw surface area is computed. Similarly, the accessible surface area (**Figure 5**) is calculated by taking account of the probe radius at 1.4Å.

(4) Computing the number of alpha carbons within a sphere centered at alpha carbon of each residue, only requires the XYZ coordinates of all alpha carbons and the sphere radii of interest. The pair-wise alpha carbon distance is calculated to generate this feature vector.

Figure 2 | Generation of vdw surface points for a single atom



For example, in 1MBS.pdb (Myoglobin, resolution=2.5Å), the NO.3 Carbon atom at is located at (16.284, 9.483, -9.998), and its vdw radius is 1.7 Å i n Lee's definition, such that its atomic surface area is $4 * \pi * r^2 = 36.3168$ Å^2. Then we set the number of surface points to 10 times of the integer part of the surface area, which is 360 in this case, such that each surface point represents 1.0088 Å^2. By applying the Golden Section Spiral algorithm, PROFEAT generates the atomic surface points evenly distributed over the sphere, as shown in **Figure 2**.

Figure 3 | Overlap of vdw surface between two adjacent atoms

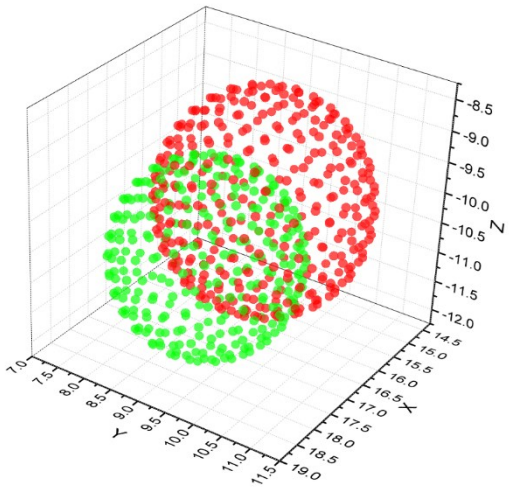
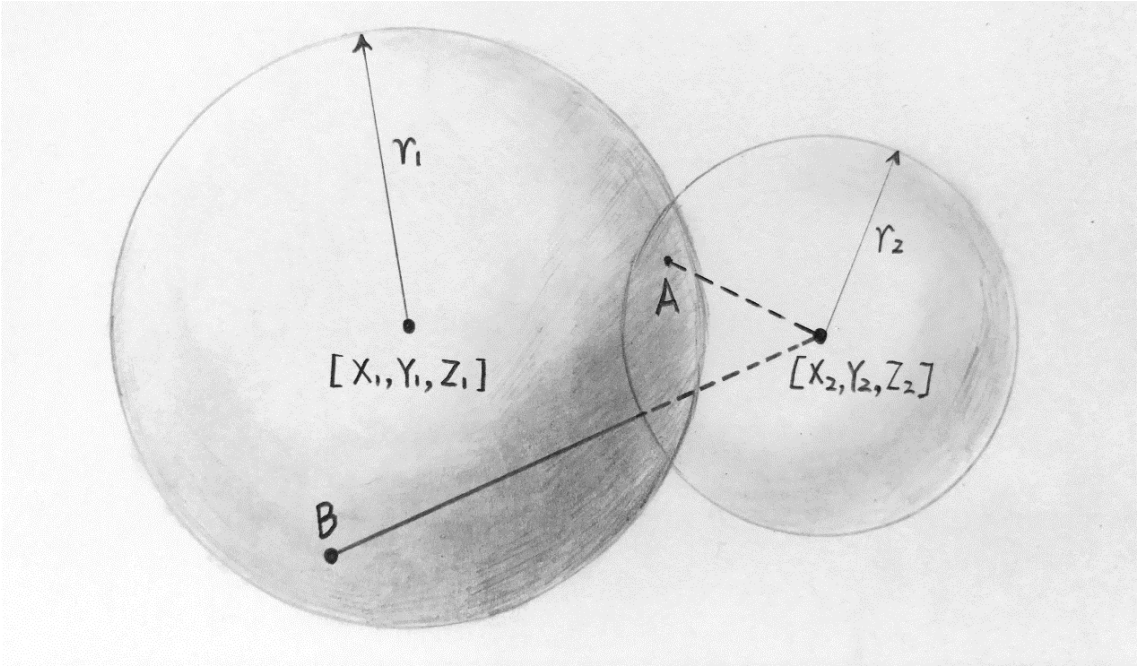


Figure 3 illustrates the overlap between the NO.3 Carbon atom with its adjacent NO.4 Oxygen atom, which is located at (17.150, 8.979, -10.709) with its vdw radius is 1.52 Å. The surface points buried inside another atoms will then be removed from the group of vdw surface points, and the calculation is shown in **Figure 3**.

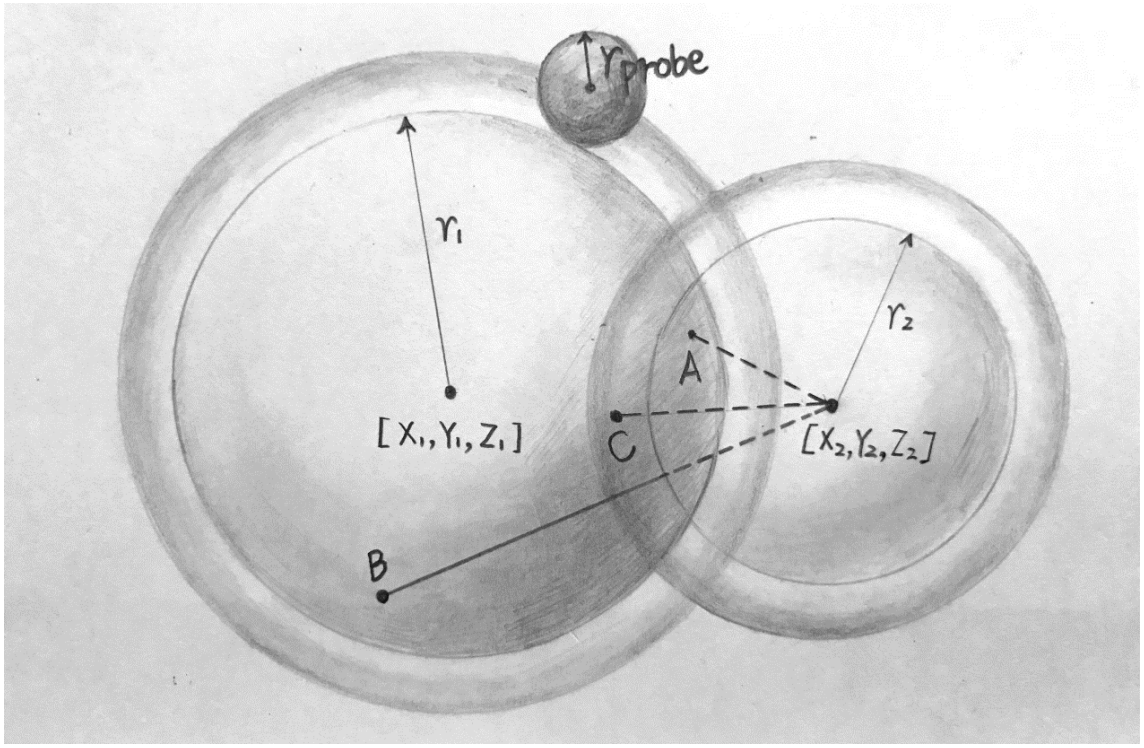
Figure 3 | Illustration of vdw surface area computation



To determine if a surface point of atom 1 is buried inside or exposed outside, the distances between this point and all other atoms' centers will be calculated. Once this surface point is buried inside any other atom's sphere, then this point will be labelled as 'buried' and never be compared with any atom center any more.

In the above **Figure 3**, suppose atom 1 is the atom of interest, and all its surface points will be measured against the center of atom 2, if the distance is less than r_2 (e.g. A), then this point is buried and will be no longer measured; if the distance is larger than r_2 (e.g. B), then this point will continue to be measured with other atoms' centers. Once atom 1 has been measured against all the other atoms, the left surface points will be considered as exposed outside. The number of left surface points multiply with the area represented by each point, gives the exposed vdw surface area of that particular atom.

Figure 4 | Illustration of accessible surface area computation



Similarly as vdw surface area calculation, accessible surface area take account of a probe with radius 1.4 Å, which is approximately the radius of a water molecule. Following the same method as above, by only enlarging the atomic radii by 1.4 Å for all atoms, such that the surface point C of atom 1 that should be considered as exposed, now is buried inside. Therefore, accessible surface area could be calculated.

(D) Sample Input & Output

Figure 5 | PROFEAT Home page with added module for protein 3D-structrure descriptors

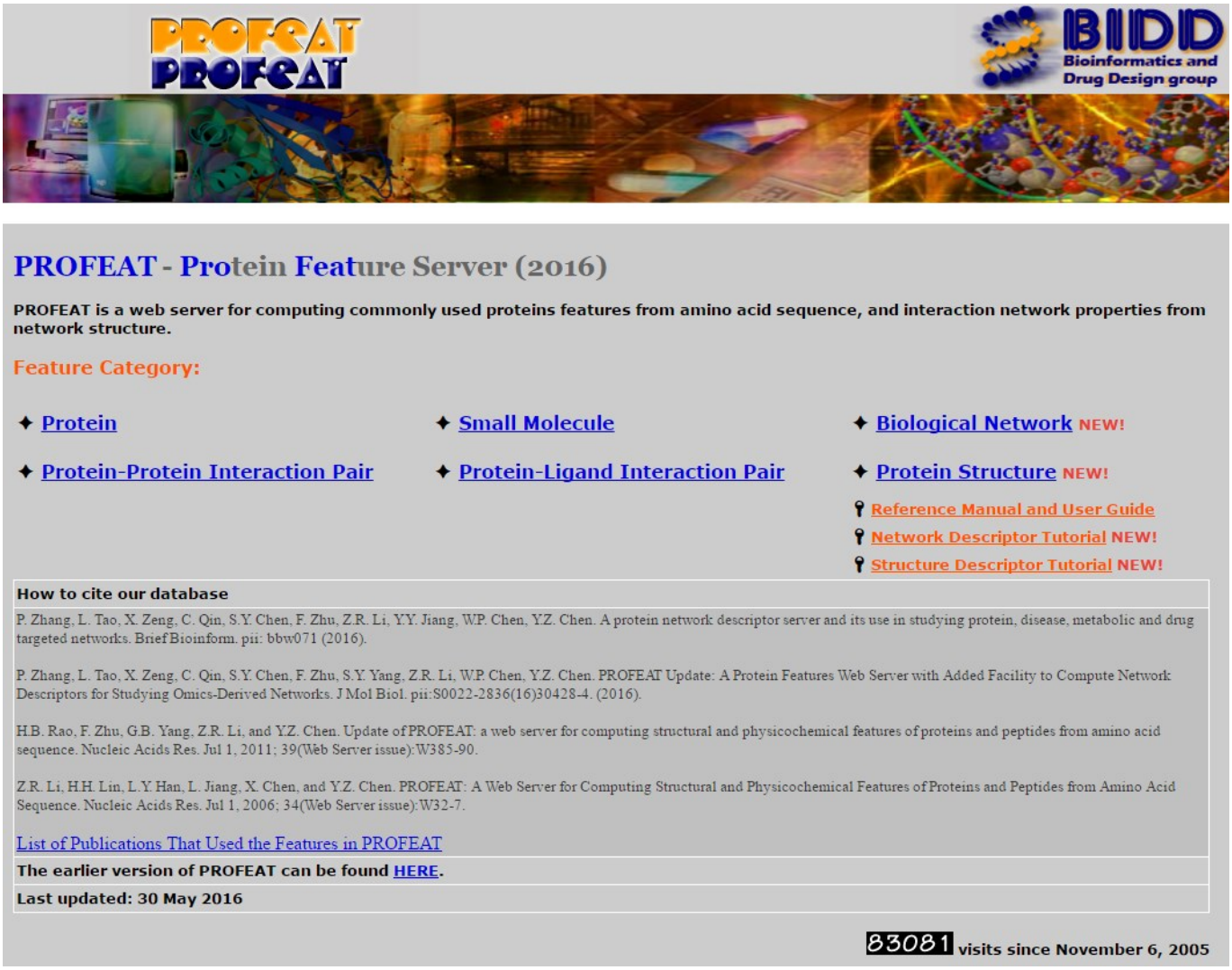


Figure 6 | Web page for computing protein 3D-structrure descriptors in PROFEAT

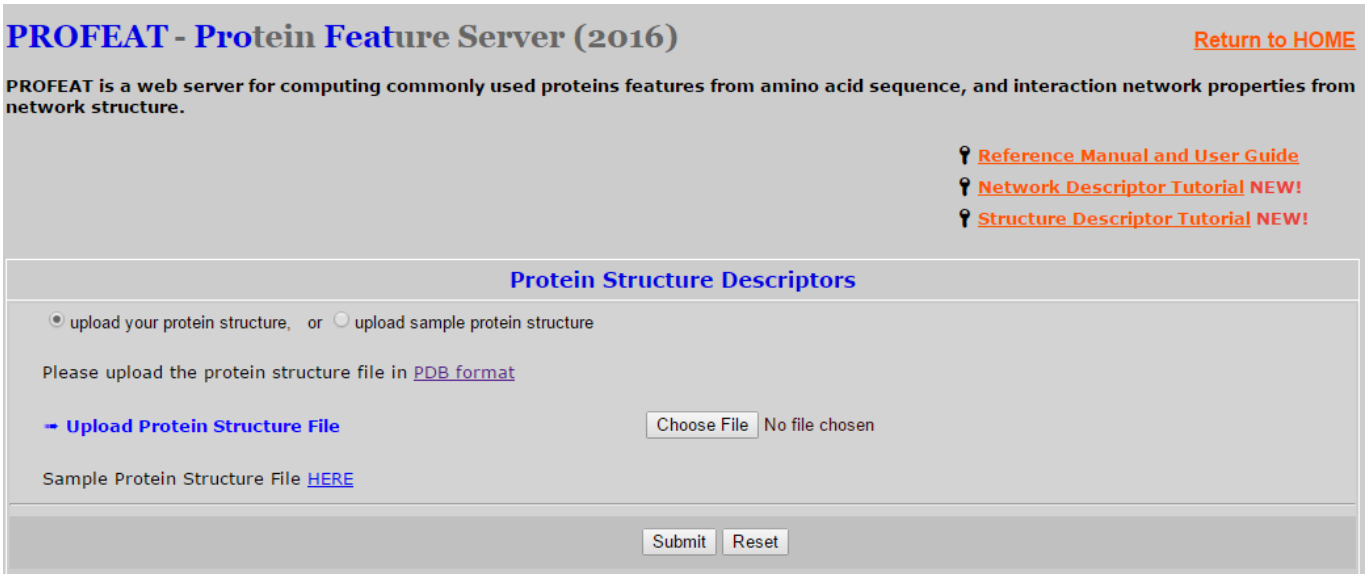


Figure 7 | Result page for the computed 3D-structure descriptors by the sample input *1nxb.pdb*

Protein structure descriptors in PROFEAT raw format

Download ALL Data

Back to PROFEAT Structure Descriptor Page

```
PDB ID: 1NXB
Classification: neurotoxin (post-synaptic)
Resolution: 1.38 Angstroms
Experiment: x-ray diffraction
Protein Structure (1): Chain A,


****** SCOP CLASSIFICATION ******
# PDB Chain ID: A:
# SCOP Code: g.7.1.1
# SCOP Class: Small proteins
# SCOP Fold: Snake toxin-like
# SCOP Superfamily: Snake toxin-like
# SCOP Family: Snake venom toxins
# SCOP Protein Domain: Erabutoxin B (also neurotoxin B)
# SCOP Species: Sea snake (Laticauda semifasciata) [TaxId: 8631]
# SCOP Domain: 1nxb A:


****** CATH CLASSIFICATION ******
# PDB Chain: 1nxb-A-00
# CATH Code ID: 2.10.60.10
# CATH Class: Mainly
# CATH Architecture: Ribbon
# CATH Topology: CD59
# CATH Homologous Superfamily: CD59


****** STRUCTURE DESCRIPTOR ******
# Protein Structure (1): Chain A,
# Protein Structure Descriptor for Chain A

# Based on the Atomic Van Der Waals Radii Published by Lee in JMB (1971) 55:379
# # Total Van Der Waals Surface Area (A^2): 5694.456
# # Total Polar Van Der Waals Surface Area (A^2): 2501.099
# # Total Non-Polar Van Der Waals Surface Area (A^2): 3193.357
# # Backbone Van Der Waals Surface Area (A^2): 2398.05
# # Backbone Polar Van Der Waals Surface Area (A^2): 1437.168
# # Backbone Non-Polar Van Der Waals Surface Area (A^2): 960.882
# # Sidechain Van Der Waals Surface Area (A^2): 3296.406
# # Sidechain Polar Van Der Waals Surface Area (A^2): 1063.931
# # Sidechain Non-Polar Van Der Waals Surface Area (A^2): 2232.475
# # Total Accessible Surface Area (A^2): 4035.343
# # Total Polar Accessible Surface Area (A^2): 2013.572
# # Total Non-Polar Accessible Surface Area (A^2): 2021.771
# # Backbone Accessible Surface Area (A^2): 1120.757
# # Backbone Polar Accessible Surface Area (A^2): 829.92
# # Backbone Non-Polar Accessible Surface Area (A^2): 290.837
# # Sidechain Van Der Waals Surface Area (A^2): 2914.586
# # Sidechain Polar Van Der Waals Surface Area (A^2): 1183.652
# # Sidechain Non-Polar Van Der Waals Surface Area (A^2): 1730.934

# Based on the Atomic Van Der Waals Radii Published by Shrake in JMB (1973) 79:351
# # Total Van Der Waals Surface Area (A^2): 5553.693
# # Total Polar Van Der Waals Surface Area (A^2): 1481.183
# # Total Non-Polar Van Der Waals Surface Area (A^2): 4072.509
# # Backbone Van Der Waals Surface Area (A^2): 2107.865
# # Backbone Polar Van Der Waals Surface Area (A^2): 787.391
# # Backbone Non-Polar Van Der Waals Surface Area (A^2): 1320.474
# # Sidechain Van Der Waals Surface Area (A^2): 3445.828
# # Sidechain Polar Van Der Waals Surface Area (A^2): 693.793
# # Sidechain Non-Polar Van Der Waals Surface Area (A^2): 2752.035
# # Total Accessible Surface Area (A^2): 4011.739
# # Total Polar Accessible Surface Area (A^2): 1274.203
# # Total Non-Polar Accessible Surface Area (A^2): 2737.536
# # Backbone Accessible Surface Area (A^2): 982.588
# # Backbone Polar Accessible Surface Area (A^2): 495.192
# # Backbone Non-Polar Accessible Surface Area (A^2): 487.397
# # Sidechain Accessible Surface Area (A^2): 3029.151
# # Sidechain Polar Accessible Surface Area (A^2): 779.011
# # Sidechain Non-Polar Accessible Surface Area (A^2): 2250.14

# Based on the Atomic Van Der Waals Radii Published by Richmond in JMB (1978) 119:537
# # Total Van Der Waals Surface Area (A^2): 5588.018
# # Total Polar Van Der Waals Surface Area (A^2): 1984.557
# # Total Non-Polar Van Der Waals Surface Area (A^2): 3603.461
# # Backbone Van Der Waals Surface Area (A^2): 2192.007
# # Backbone Polar Van Der Waals Surface Area (A^2): 1092.168
# # Backbone Non-Polar Van Der Waals Surface Area (A^2): 1099.839
# # Sidechain Van Der Waals Surface Area (A^2): 3396.011
# # Sidechain Polar Van Der Waals Surface Area (A^2): 892.389
# # Sidechain Non-Polar Van Der Waals Surface Area (A^2): 2503.622
# # Total Accessible Surface Area (A^2): 4006.502
# # Total Polar Accessible Surface Area (A^2): 1595.566
# # Total Non-Polar Accessible Surface Area (A^2): 2410.936
# # Backbone Accessible Surface Area (A^2): 1034.985
# # Backbone Polar Accessible Surface Area (A^2): 638.529
# # Backbone Non-Polar Accessible Surface Area (A^2): 396.456
# # Sidechain Accessible Surface Area (A^2): 2971.517
# # Sidechain Polar Accessible Surface Area (A^2): 957.037
# # Sidechain Non-Polar Accessible Surface Area (A^2): 2014.48

# Number of Alpha Carbons winthin a Sphere (r = 8A) Centered at Alpha Carbon of Each Amino Acid:
# # [001-050]: 11, 12, 17, 14, 12, 13, 9, 7, 7, 6, 5, 8, 7, 10, 10, 7, 9, 6, 5, 4, 9, 8, 12, 17, 17, 13, 13, 11, 10, 8, 6, 4, 5, 6, 8,
# # [051-062]: 9, 9, 9, 11, 9, 8, 10, 7, 9, 10, 12, 5,
# Number of Alpha Carbons winthin a Sphere (r = 10A) Centered at Alpha Carbon of Each Amino Acid:
# # [001-050]: 17, 21, 27, 23, 25, 17, 17, 15, 14, 8, 9, 11, 13, 16, 15, 14, 16, 11, 7, 7, 14, 17, 20, 27, 25, 26, 21, 16, 14, 12, 8,
# # [051-062]: 14, 17, 16, 19, 18, 17, 13, 16, 16, 15, 23, 14,
# Number of Alpha Carbons winthin a Sphere (r = 12A) Centered at Alpha Carbon of Each Amino Acid:
# # [001-050]: 23, 29, 33, 35, 29, 24, 23, 22, 20, 14, 10, 18, 20, 25, 24, 21, 22, 15, 12, 12, 22, 23, 30, 34, 38, 34, 34, 25, 23, 15,
# # [051-062]: 22, 25, 27, 30, 25, 22, 20, 24, 28, 24, 35, 28,
# Number of Alpha Carbons winthin a Sphere (r = 14A) Centered at Alpha Carbon of Each Amino Acid:
# # [001-050]: 28, 34, 41, 41, 39, 30, 30, 29, 27, 19, 17, 24, 28, 32, 32, 29, 30, 23, 17, 17, 26, 28, 37, 43, 48, 47, 40, 35, 26, 20,
# # [051-062]: 27, 33, 35, 36, 30, 31, 31, 32, 34, 36, 43, 36,
# Number of Alpha Carbons winthin a Sphere (r = 16A) Centered at Alpha Carbon of Each Amino Acid:
# # [001-050]: 33, 37, 45, 45, 46, 40, 40, 38, 35, 26, 21, 30, 32, 41, 41, 35, 36, 29, 21, 23, 33, 33, 39, 47, 51, 54, 47, 41, 33, 24,
# # [051-062]: 34, 40, 45, 45, 40, 35, 32, 35, 42, 42, 48, 43,
```