Initiated Exhaustive search for SRS-MOTIF detection

Primary Author
A. Nishanth

Primary Supervisor
Dr. C. H. Chu
Co-Supervisor
Dr. Rasiah Loganantharaj
November 26, 2020

Contents

Chapter 1

Initiated Exhaustive search for SRS-MOTIF detection

This report directly show the part(from my dessertation) related to this project.

Possible-SRS-MOTIFs: just suspecting these surface substructures may contribute to the functionality.

Figure 1.1 Exhaustive search for possible SRS-MOTIF detection (pseudo code)

1: **for** PDB in $Functionalitygroup of PDBs$ **do**
2: Find the surface $C\alpha$ atoms from PDB; {this create surface atoms coordinate file, corresponding polar-coordinate file, and corresponding amino acid details for those atoms}
3: Triangulate all the surface atoms; {as explained in *Appendix* B, class "***motif_group_intialize_pikle***" is used. That class calls the convex hull (uses coordinate file) to triangulate all.}
4: Find all the groups of Possible SRS-MOTIFs and save it with the detail of center amino acid and the corresponding PDB's name; hereinafter "$saved - PDB - pickle - group$". {when it saves, if the PDBs' Possible-SRS-MOTIFs contain same center residue then it is saved in one pickle file, with the ascending order of number of vertices in that Possible-SRS-MOTIF}
5: **end for**
6: **for** $aminoacid$ in $center(1to20)$ **do**
7: Take a $saved - PDB - pickle - group$ and go through all $saved - PDB - pickle - group$s with the same center amino acid number. Check the equality of the groups if same then it hit; if the hits are higher than 15 (the count 15 is chosen without losing the possibility of SRS-MOTIFs structure; and not consider all the structures; please check the *Appendix* C.1), then the Possible-SRS-MOTIF(taken $saved - PDB - pickle - group$) is saved with the hit number (class ***motif_group_find_pikle*** is used, please check the *Appendix* C.1.1 Pseudo Codes).
8: **end for**

Fig. 1.1 covers the Main pseudo code used for the Exhaustive search. The *Appendix* C covers the details of the Exhaustive search. *Appendix* C.7 shows the details of HashMap's used for this code. The groups (all possible SRS-MOTIFs as presented in *Appendix* C.3 and *Appendix* C.4 for the ONGO and TSG accordingly) are found in the exhaustive search, do not mean all these groups contribute to the functionality.

Table 1.1: Minimum coefficients of RV Vs RV_2 among different data representation for Possible MOTIF groups comparison

Comparison between the Unique groups of/ between	Amino-acids' presentation in the possible-SRS-MOTIF representation matrix with					
	Vertices information				Edges information	
	Positive property representation		Negative property representation		Negative property representation	
	RV	RV_2	RV	RV_2	RV	RV_2
ONGO	0.847	0.875	0.015	-0.298	0.002	-0.452
TSG	0.872	0.867	0.004	-0.362	0.001	-0.416
ONGO Vs TSG	0.92	0.866	-0.169	-0.395	-0.202	-0.464

The underlying hypothesis is the groups fell in ONGO group contributes to ONGO functionality, and the groups fell TSG group contributes to TSG functionality. Thus, the groups need to be had non overlapping groups. To do this first these groups are pooled together and only the non-overlapping groups between ONGO, and TSG is selected, and their details are presented *Appendix* C.5 and *Appendix* C.6 accordingly.

1.1 Modified RV coefficient for comparing the possible MOTIF groups

As mentioned in [1]

"The RV coefficient was introduced by Escoufier (1973, see also Robert & Escoufier, 1976) as a measure of similarity between squared symmetric matrices (specifically: positive semi-definite matrices) and as a theoretical tool to analyze multivariate techniques."

Another variation of RV coefficient as modified RV coefficient (RV_2) coefficient is proposed in [2]. That is more effective to compare the matrixes with they are similar or opposite by the values between +1 to -1. The definitions of RV coefficient and modified RV coefficient is shown in Fig. 1.2.

The functional similarity between these possible-SRS-MOTIFs can be measured by the RV coefficient or modified RV coefficient (RV_2). And these possible-SRS-MOTIFs biochemical matrix representations have more influence in these coefficient calculations. Table. 1.1 summarize the biochemical matrix representations along with the minimum coefficient value achieved, because the minimum coefficient value implies the negative connection between the possible-SRS-MOTIFs. From the results shown in Table. 1.1, RV_2 coefficient is better than RV coefficient.

Possible-SRS-MOTIF biochemical representation, can be majorly categorized into two ways, such as

1. **property representation:** Each amino acid in the group is represented by 16 biochemical properties (*Appendix* A)
   (a) **Positive property representation:** Just directly use the property as mentioned *Appendix* A, without normalization.
   (b) **Negative property representation:** If there are opposite properties, then one of them is inverted to negative to represent the opposite biochemical property. (subsection **??** the

Figure 1.2: Definition of RV coefficient

X = Possible SRS-MOTIF group property representation
Z = Another possible SRS-MOTIF group property representation
To find the positive semidefinite of the matrix X, and Z separately as, B, and U accordingly

$$B = XX^T; \text{where } X^T \text{ annotate the transpose of } X$$
$$U = ZZ^T$$

Definition to trace

$$Trace\ (B^TU) = \sum_{i,j} b_{ij}u_{ij}$$

Definition to RV coefficient

$$RV = \frac{Trace(B^TU)}{\sqrt{Trace(B^TB) \times Trace(U^TU)}} \tag{1.1}$$

Definition to modified RV coefficient

let us define find the positive semidefinite of the matrix without their diagonals X, and Z separately as $\widetilde{B}$ and $\widetilde{U}$ accordingly

$\widetilde{B} = XX^T - diag(XX^T)$; where the matrix $diag(XX^T)$ only contains the diagonal elements of XX^T

Like wise, $\widetilde{U} = ZZ^T - diag(ZZ^T)$

$$\mathrm{RV}_2 = \frac{\mathrm{Trace}\left(\tilde{B}^T\widetilde{U}\right)}{\sqrt{\mathrm{Trace}\left(\tilde{B}^T\tilde{B}\right) \times \mathrm{Trace}\left(\widetilde{U}^T\widetilde{U}\right)}} \tag{1.2}$$

negative included property representation, further *Appendix* D shows the negative representation)

2. **Amino-acids' presentation in the possible-SRS-MOTIF:**
   (a) **Vertices information:** Represent the matrix with the properties of vertices presented in the Possible-SRS-MOTIF; thus, all the amino acids presented in the possible-SRS-MOTIF get same weightage.
   (b) **Edges information:** Represent the matrix with the properties with edges presented in the Possible-SRS-MOTIF; thus, all the amino acids presented in the possible-SRS-MOTIF get different weightage depends on the way presented.

Note: Negative Normalized property representation scale used in the coefficient calculation, to represent all biochemical property as same weightage. None of these representation covers the distance

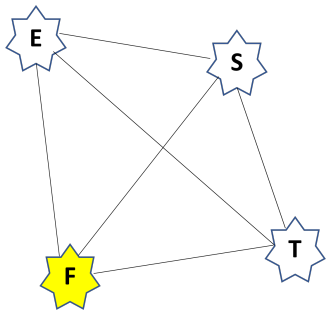
Table 1.2: Matrix representation example of first possible-SRS-MOTIF group of ONGO

property representation	Vertices Vs Edge	Amino acids	charged	Polar	Hydrophobic	Hydrophobic	Moderate	Hydrophilic	polar	Aromatic	Aliphatic	Acid	Basic	Negative charge	Neutral	Positive charge	$pK_a - NH_2$	$pK_a - COOH$
Positive	Vertices	F	0	0	1	1	0	0	0	1	0	0	0	0	1	0	9.24	2.58
		S	0	1	0	0	0	1	1	0	0	0	0	0	1	0	9.15	2.21
		E	1	0	0	0	0	1	0	0	0	1	0	1	0	0	9.67	2.19
		T	0	1	0	0	0	1	1	0	0	0	0	0	1	0	9.12	2.15
Negative normalized	Edge	FS	-2	-2	2	2	-2	-2	-2	2	-2	-2	-2	-2	2	-2	0.444	0.242
		FE	-2	-2	2	2	-2	-2	-2	2	-2	-2	-2	-2	2	-2	0.444	0.242
		FT	-2	-2	2	2	-2	-2	-2	2	-2	-2	-2	-2	2	-2	0.444	0.242
		ES	2	-2	-2	-2	-2	2	-2	-2	-2	2	-2	2	-2	-2	0.879	0.134
		ET	2	-2	-2	-2	-2	2	-2	-2	-2	2	-2	2	-2	-2	0.879	0.134
		ST	-2	2	-2	-2	-2	2	2	-2	-2	-2	-2	-2	2	-2	0.354	0.139

between the amino acids presented; because already the properties are in normalized scale, then more factorization due to distance may lead to poor coefficient calculation.

Since the "property representation" is straight forward just assigning the properties to the amino acid presented. Lets' take a closer look on "Amino-acids" presentation in the possible-SRS-MOTIF" via an example.

Figure 1.3: First possible-SRS-MOTIF group of ONGO's structure representation considering the center residue and the amino acids presented (not include the information of length of none of the edges).



E.g.: let us take the fist possible-SRS-MOTIF group of ONGO Which has center residue as "F" and edges as "ES", "ET" and "ST" with 40 hits (40 times presented in the ONGO's PDBs) For this possible-SRS-MOTIF group *"Vertices information"* represent the matrix properties of all the amino acids such as "F", "E", "S", and "T" presented in the group. For this possible-SRS-MOTIF group *"Edge information"* represented by the matrix with the edges' properties (edge property is obtained by the sum of the property of the vertices presented in the edge; lets' take a look on "FE" edge, which represented by sum of the properties of "F" and "E"), such as edges formed with center "FE", "FS", and "FT", and the other edges "ES", "ET", and "ST". And the Table. 1.2 presents the matrix of the representation with vertices with positive property representation, and Edge with negative representation.

From the results of ONGO Vs TSG possible-SRS-MOTIF groups' minimum RV_2 shown in Table. 1.1, the Negative normalized property with edge representation obtain minimum among all and the positive vertices representation got maximum (not intended). Thus, it is clear the RV_2 coefficient with

Negative normalized property with edge representation works well.

The Negative normalized property with edge and vertices representation results can be checked in the links,

$RV_2\_motif\_normalized\_negated\_coefficient\_edge$:
`https://drive.google.com/file/d/122gW6-v27eVcZfMOtF_nIeBET8bohfad/view?usp=sharing` and,

$RV_2\_motif\_normalized\_negated\_coefficient\_vertices$:
`https://drive.google.com/file/d/14bJnF3GbGlZxIW4Cz8nlK7icqsoXblse/view?usp=sharing`

and some samples is shown in *Appendixes* D.2 and explained.

Using the calculated ONGO possible-SRS-MOTIF groups with TSG possible-SRS-MOTIF groups' RV_2 coefficients, mean RV2 coefficient for each possible-SRS-MOTIF group is calculated. If the possible-SRS-MOTIF group's mean RV_2 coefficient is negative, then proposing that structure may SRS-MOTIF groups. Only in TSG's 108th SRS-MOTIF obtain 17 hits

With center residue as "H" and the edges as "MR", "MS", "RS", "ST", and "ST" obtain $RV_2 = -0.014$

Expected negative mean RV_2 coefficient for more groups, thus this approach not worked out. Thus, move on to check with the details of distance between the residues to find out possible-SRS-MOTIFs.

Bibliography

[1] Hervé Abdi. RV coefficient and congruence coefficient. page 10.

[2] A. K. Smilde, H. A. L. Kiers, S. Bijlsma, C. M. Rubingh, and M. J. van Erk. Matrix correlations for high-dimensional data: the modified rv-coefficient. 25(3):401–405.

Appendix A

Level 1 Property Tables

Table A.1: Amino acids with properties (first half)

Amino acid short form		Arg	Lys	Asp	Glu	Gln	Asn	His
Code		R	K	D	E	Q	N	H
Web reference	charged	1	1	1	1	0	0	0
	Polar	0	0	0	0	1	1	1
	Hydrophobic	0	0	0	0	0	0	0
Soluble reference	Hydrophobic	0	0	0	0	0	0	0
	Moderate	0	0	0	0	0	0	1
	Hydrophilic	1	1	1	1	1	1	0
	polar	0	0	0	0	1	1	0
	Aromatic	0	0	0	0	0	0	0
	Aliphatic	0	0	0	0	0	0	0
	Acid	0	0	1	1	0	0	0
	Basic	1	1	0	0	0	0	1
	Negative charge	0	0	1	1	0	0	0
	Neutral	0	0	0	0	1	1	0
	Positive charge	1	1	0	0	0	0	1
	$pK_a - NH_2$	9.09	10.28	9.6	9.67	9.13	8.8	8.97
	$pK_a - COOH$	2.18	8.9	1.88	2.19	2.17	2.02	1.78
Normalized	$pK_a - NH_2$	0.15	0.75	0.4	0.44	0.17	0	0.09
	$pK_a - COOH$	0.07	1	0.02	0.07	0.06	0.04	0.01

Table A.2: Amino acids with properties (second half)

Amino acid short form		Ser	Thr	Tyr	Cys	Met	Trp	Ala	Ile	Lue	Phe	Val	Pro	Gly
Code		S	T	Y	C	M	W	A	I	L	F	V	P	G
Web reference	charged	0	0	0	0	0	0	0	0	0	0	0	0	0
	Polar	1	1	1	1	1	1	0	0	0	0	0	0	0
	Hydrophobic	0	0	0	0	0	0	1	1	1	1	1	1	1
Soluble reference	Hydrophobic	0	0	1	0	0	1	1	1	1	1	1	1	1
	Moderate	0	0	0	1	1	0	0	0	0	0	0	0	0
	Hydrophilic	1	1	0	0	0	0	0	0	0	0	0	0	0
	polar	1	1	0	1	1	0	0	0	0	0	0	0	0
	Aromatic	0	0	1	0	0	1	0	0	0	1	0	0	0
	Aliphatic	0	0	0	0	0	0	1	1	1	0	1	0	0
	Acid	0	0	0	0	0	0	0	0	0	0	0	0	0
	Basic	0	0	0	0	0	0	0	0	0	0	0	0	0
	Negative charge	0	0	0	0	0	0	0	0	0	0	0	0	0
	Neutral	1	1	1	1	1	1	1	1	1	1	1	1	1
	Positive charge	0	0	0	0	0	0	0	0	0	0	0	0	0
	$pK_a - NH_2$	9.15	9.12	9.11	10.78	9.21	9.39	9.87	9.76	9.6	9.24	9.72	10.6	9.6
	$pK_a - COOH$	2.21	2.15	2.2	1.71	2.28	2.38	2.35	2.32	2.36	2.58	2.29	1.99	2.34
Normalized	$pK_a - NH_2$	0.18	0.16	0.16	1	0.21	0.3	0.54	0.48	0.4	0.22	0.46	0.91	0.4
	$pK_a - COOH$	0.07	0.06	0.07	0	0.08	0.09	0.09	0.08	0.09	0.12	0.08	0.04	0.09

Appendix B

SRS-MOTIF Excusive search Initial Code files

This appendix covers the detailed description of the pseudo code. First, the main classes' description is given for this coding in Fig. B.1.*Motif_group_intialize_pikle* classes' main functions description is given in Fig. B.2.(since the *motif_group_pikle_to_property_pikle* is just using the groups created by the *Motif_group_intialize_pikle* to properties).

Figure B.1 Main class discription

motif_group_intialize_pikle;
This class is basically written for creating the pickle of groups of amino acids with neighbors' details with ascending order of length in the group
motif_group_pikle_to_property_pikle;
This class is basically written for check the pickle of groups of amino acids PDB files, and make the property for neural network

Extend the surface triangulation to cover all surface points the $creating\_surface\_by\_triangle$ is used the loop structure is shown in Fig. B.3.

Angle condition consideration

The angles θ and ϕ angles are considered differently, from the Fig. **??** shown in below ϕ angle does not has any issues, Theta angle 180^o and -180^o degrees are same.

Here only consider the 15^o(or $15^o = 0.2618$ radians) in all way's such that

✓ ϕ considers the range between 15^o, and -15^o

✓ θ considers the range between 15^o, and -15^o

check which triangles middle(point's) angle fell into these ranges

These loosing conditions as metioned below one after another; if no traingle satisfied the given condition, only then loose condition to the next level. The condtions are loose(15^o apart) in the order as,

Figure B.2 Important functions in class *motif_group_intialize_pikle*

1: *distance_surface* (self, coordinates, *triangle_points*, *missed_point_index*); {This function calculates the distance from the missed point where,

- ✓ coordinates: cartesian coordinates of the surface $C\alpha$ atoms
- ✓ *triangle_points*: contains the, *vertices_position_triangle* Which triangle from the vertices want to be considered
- ✓ *missed_point_index*: The point which need to calculate the distance

}

2: *creating_surface_by_triangle* (self); {This function creates the surface with triangles from,

1. create the convex hull using the library
2. Extend the surface triangulation which should cover all surface points

This function does

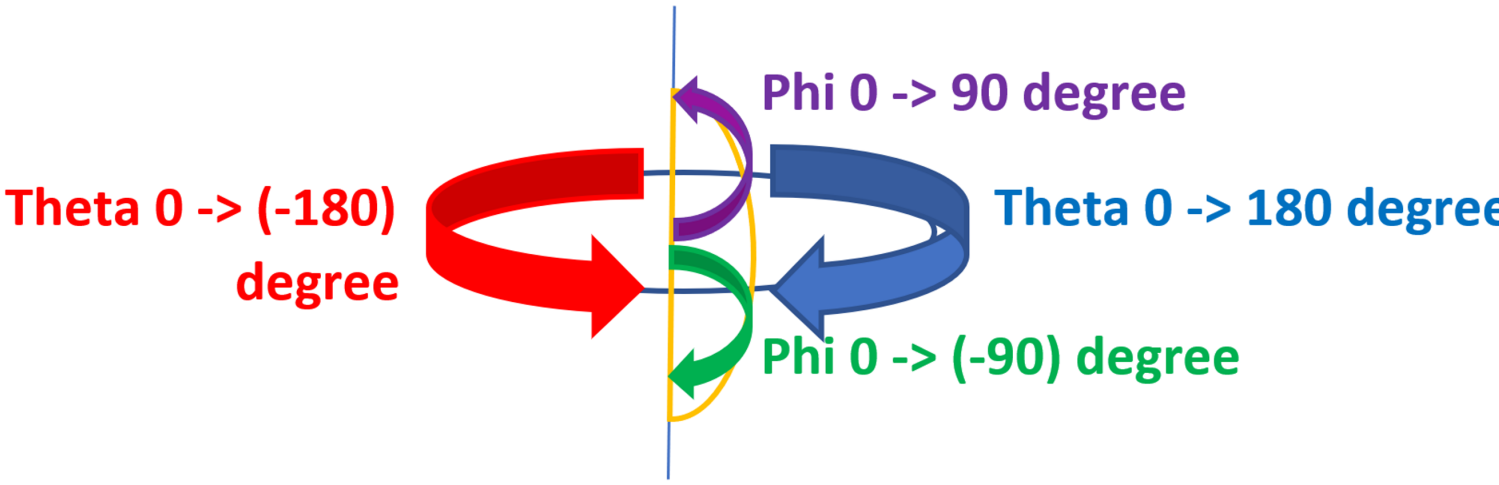
- ✓ *checking_points_order*: keep the missing points checking order in the descending order of the radius
- ✓ the triangles points are selected one by one and choose the minimum and maximum ϕ angle and place with the triangle position as HashMap triangle position to ϕ angles (xy_z) minimum and maximum

}

Figure B.3 Functions *creating_surface_by_triangle*(Pseudocode)

1: **while** *checking_points_order* is non-empty **do**
2: *missed_point_index* = *checking_points_order*[0] {choose first point form the *checking_points_order*}
3: *sat_traingles* are chosen from the surface triangles fall, in the range of θ and ϕ of the missing coordinate {missing coordinate (found by coordinate[*missed_point_index*])}
4: **while** *sat_traingles* is empty **do**
5: *sat_traingles* are chosen from the surface triangles fall, in the range of θ and ϕ of the missing coordinate
6: If none of the surface triangles satisfied, loose the condition even further, increase the θ and ϕ angle of missing coordinate. {Detailed explonation is found in hypothetical example on the section "Angle condition consideration"}
7: **end while**
8: **for** *triangle_points* in *sat_traingles* **do**
9: calculate the distance from the point(via *distance_surface*) to *sat_traingles* and choose the minimum.
10: **end for**
11: Break the selected minimum distance triangle with the *missed_point_index*, and update the surface triangles with the broken updated set; {selected minimum distance triangle is also removed from the list of surface triangles}
12: remove the *missed_point_index* from *checking_points_order*
13: **end while**

Figure B.4: PDB files atom in surface polar coordinates; the relation between the figure annotation and the text are, Theta = θ, Phi = ϕ, and degree in angle = ^{o}.



1. range between 30^o, and -30^o
2. range between 45^o, and -45^o
3. range between 60^o, and -60^o
4. range between 75^o, and -75^o

Breaking into triangles hypothetical example

Consider the breaking triangle contain (a, b, c) as points and added point going to be d (missed point), then the set of new triangles are,

1. (a, b, d)
2. (b, c, d)
3. (a, c, d)

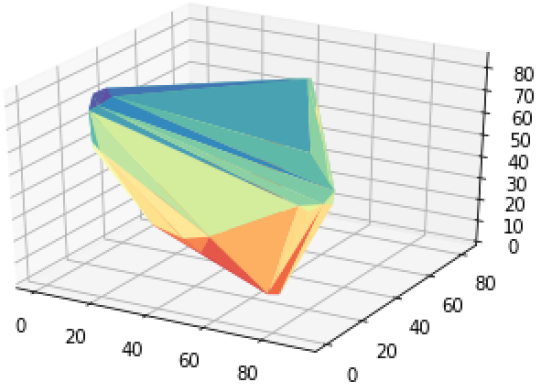
Since these "*triangle_break*" and θ and ϕ angles are used again and again after updated also thus those are made as deep copies

Remove the (a, b, c) triangle and add the new triangles such as, (a, b, d), (b, c, d), and (a, c, d). Further these new triangles details (phi, theta) are also added; as mentioned in the Fig. B.3.

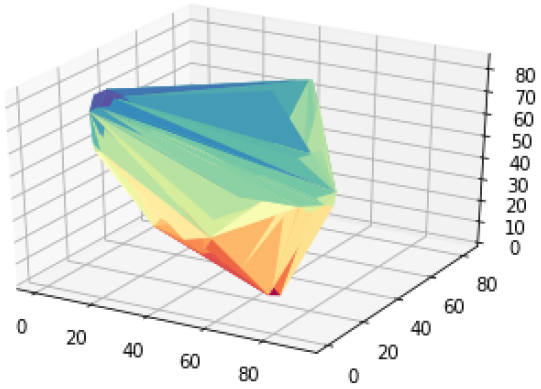
Effectiveness of the algorithm/ visualize the algorithms' outcome is shown in the Fig. B.5.

Figure B.5: Examples of surface visualize before and after the surface algorithm applied with convex hull surface algorithm to find the surface

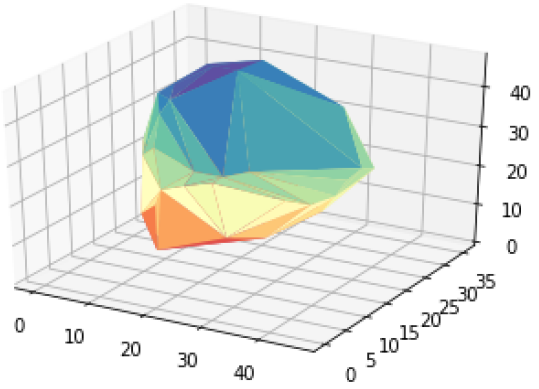
(a) 1BQU's surface visualized the surface only with convex hull



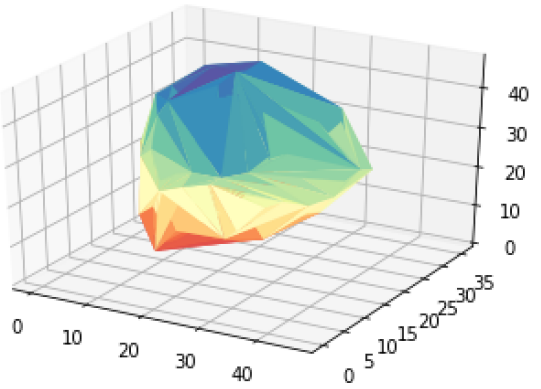
(b) 1BQU's surface visualized after the surface algorithm applied



(c) 1E3G's surface visualized the surface only with convex hull



(d) 1E3G's surface visualized after the surface algorithm applied



Appendix C

SRS-MOTIF Excusive search

C.1 Possible SRS-MOTIF hit selection define threshold

The search is executive, along the 1270 ONGO and 1159 TSG non overlapping PDBs from COSMIC V82 gene and Uniprot mapping. Searches end up with unique small group of amino acids numbers as ONGO: 108,346 and TSG: 107,343. The uniqueness of the small groups is only among those classes. Here overlapping of small groups can be occur between ONGO and TSG.

However, among these $\approx$ .2 Million, unique small group of surface amino acids, at least $\approx$ 1000 small substructures (Surface-Residues ($C\alpha$)-Structural-MOTIF (SRS-MOTIF)) contribute to the functionality. To find those, instead of checking matching (exact number and exact group of amino acids) frequency; find out the frequency of same kind of amino acids groupings per unique small groups on surface. This may capture the same kind of amino acids groups' biochemical property distributions on the surfaces' contribution on the functionality.

Table. C.1 results represent the frequency of such unique small group of same kind of surface amino acids, occurring for ONGO and TSG accordingly. From this threshold 15 is chosen. Because in both cases of ONGO and TSG (as shown in Table. C.1) will make sure to reduce the search space of the possible SRS-MOTIF among the top 10 % highest hits per unique groups of surface amino acids(from the statistics of the same kind of amino acids occurrence frequency among the surface group).

And the histograms shown in Fig. C.1 and Fig. O2 visualize frequency of occurrence among the ONGO and TSG accordingly.

To illustrate let's take an example some unique groups from the ONGO is shown below all these groups fell into same category having 'M', 'K', 'E', 'T', 'Y' as amino acids in the group.

- ✓ $group_1$: 'M', 'K', 'E', 'T', 'Y'
- ✓ $group_2$: 'M', 'K', 'E', 'T', 'Y', 'Y'
- ✓ $group_3$: 'M', 'K', 'E', 'T', 'T', 'Y'
- ✓ $group_4$: 'M', 'K', 'E', 'T', 'T', 'Y', 'Y'
- ✓ $group_5$: 'M', 'K', 'K', 'E', 'T', 'Y', 'Y'

  so, when the $group_1$ is taken to check along the PDBs all the PDBs having any of these groups got hit.

Table C.1: Presents the frequency of hits per unique groups of surface amino acids, containing the same kind of amino acids per group in all

(a) over the ONGO class

Bins		Frequency	General frequency / %	Cumulative frequency/ %
1	5	65674	60.62	60.62
5	10	23489	21.68	82.29
10	15	8466	7.81	90.11
15	20	4113	3.80	93.90
20	25	2378	2.19	96.10
25	30	1361	1.26	97.36
30	35	1021	0.94	98.30
35	40	556	0.51	98.81
40	45	435	0.40	99.21
45	50	310	0.29	99.50
50	55	172	0.16	99.66
55	60	127	0.12	99.77
60	65	69	0.06	99.84
65	70	50	0.05	99.88
70	75	28	0.03	99.91
75	80	11	0.01	99.92
80	85	0	0.00	99.92
85	90	43	0.04	99.96
90	95	0	0.00	99.96
95	100	0	0.00	99.96
100	105	0	0.00	99.96
105	110	25	0.02	99.98
110	115	0	0.00	99.98
115	120	0	0.00	99.98
120	125	18	0.02	100.00

[a] Here Bin 1-5 mean, the number of unique groups of surface amino acids, occurred between one time to five times (not including five)

(b) over the TSG class

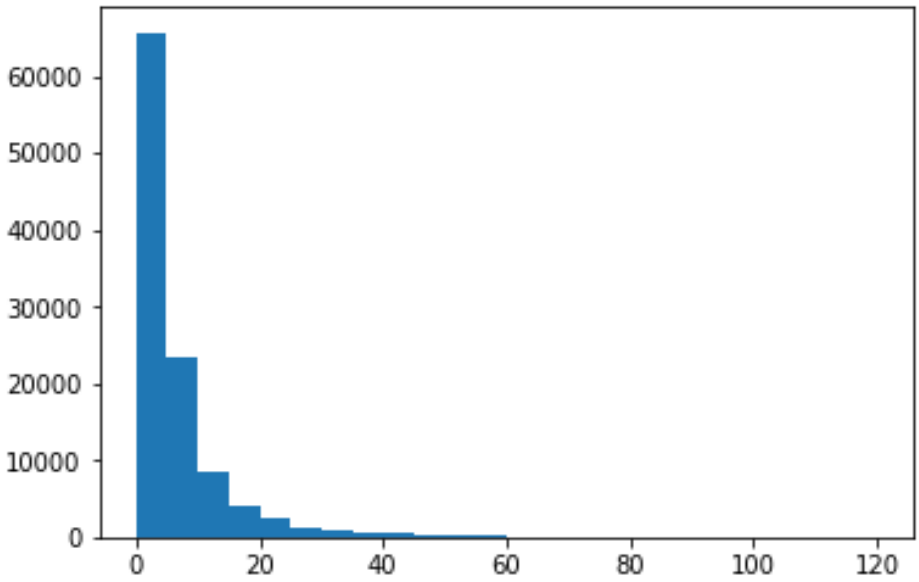
Bins		Frequency	General frequency / %	Cumulative frequency/ %
1	5	67707	63.08	63.08
5	10	21249	19.80	82.87
10	15	8014	7.47	90.34
15	20	3908	3.64	93.98
20	25	2371	2.21	96.19
25	30	1490	1.39	97.57
30	35	727	0.68	98.25
35	40	603	0.56	98.81
40	45	513	0.48	99.29
45	50	235	0.22	99.51
50	55	154	0.14	99.65
55	60	100	0.09	99.75
60	65	99	0.09	99.84
65	70	27	0.03	99.86
70	75	6	0.01	99.87
75	80	18	0.02	99.89
80	85	24	0.02	99.91
85	90	16	0.01	99.92
90	95	15	0.01	99.94
95	100	0	0.00	99.94
100	105	2	0.00	99.94
105	110	7	0.01	99.95
110	115	6	0.01	99.95
115	120	5	0.00	99.96
120	335	47	0.04	100.00

C.1.1 SRS-MOTIF detection initial Code files exhaustive search

This section covers the detailed description of the pseudo code. First, the main classes' description is given for this coding; then for each main classes' main functions description is given.

Figure C.1: Histogram of, unique groups of surface amino acids, containing the same kind of amino acids per group in all

(a) over the ONGO class



(b) over the TSG class

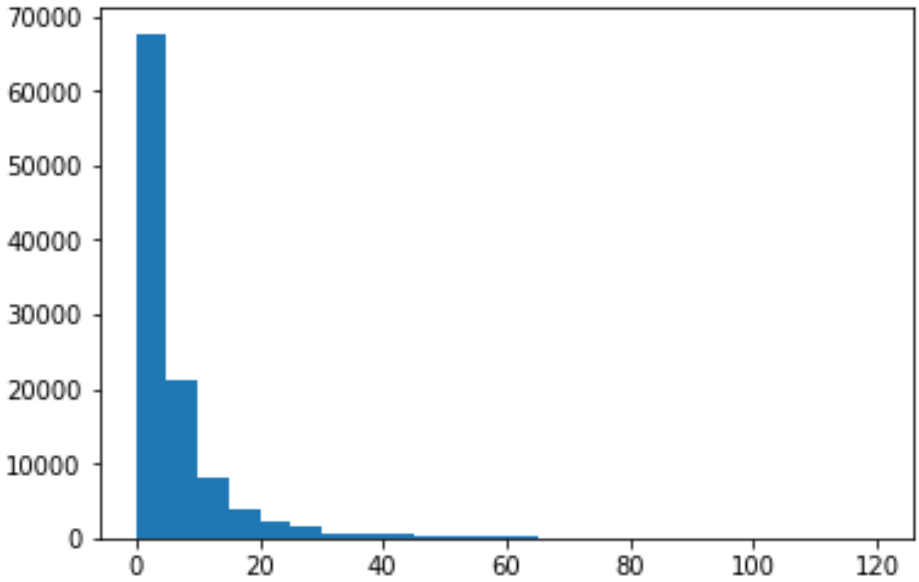


Figure C.2 Important functions in class ***motif_group_intialize_pikle*** (PseudoCode)

group_amino_acid_pikle(self) {This function create the pickle of the group of triangle points and center go through the surface atoms and find out the hit map of triangles(that coordinate included triangle) in triangles.
From the hit matrix select the triangles and select the edges which does not have center atom as one set.
#first group the center amino acids together with ascending order of number of elements in the group (this part is done for make easier to checking the two subgroups are same) }

Figure C.3 Important functions in class ***motif_group_find_pikle*** (PseudoCode)

checking_the_groups_same(self, $group_i$, $group_m$)
This function takes two groups and check whether those groups are same or not

if mean($group_i$) == mean($group_m$) **then**
  Check the Each elements in $group_i$ and $group_m$ are same.
  if all elements are identical **then**
    return 1
  else
    return 0
  end if
else
  return 0
end if

This function check if both groups are exactly same

find_unique_group(self)(self, $group_i$, $group_m$)
This function will find the unique motifs and their hits points are more than self. $hits_{satisfy}(15)$

name convention:

- ✓ *group* is the set elements made by neighbors
- ✓ *set* represents the PDB groups

then take the list from the sets and cluster them according to their length And the minimum number in the group allowed is 3,

while set_a in *sets* **do**
  while $group_i$ in set_a **do**
    while set_c in *sets* **do**
      while $group_m$ in set_c **do**
        if $group_i$==$group_m$ **then**
          $count_{group_i}$=$count_{group_i}$+1; {how many times $group_i$ hit in each PDBs; keep an array with the $group_i$ and $count_{group_i}$ to keep the count}
          remove($group_m$, set_c); {remove the $group_i$ from all the PDB sets/ pickles to avoid rerun to find the hit}
        end if
      end while
    end while
  end while
end while

Table C.3: ONGO's uncleaned all possible-SRS-MOTIF.

Index	Hits	Centre	Group		
0	40	F	ES	ET	ST
1	39	M	RT	RA	TA
2	32	D	RE	RT	ET
3	29	L	KE	KL	EL
4	28	K	NS	NS	SS
5	28	L	KI	KL	IL
6	28	L	GA	GL	AL
7	28	L	KL	KV	LV
8	27	E	RT	RL	TL
9	27	L	KL	KL	LL
10	27	L	SL	SL	LL
11	26	G	DE	DV	EV
12	25	L	GL	GV	LV
13	24	D	KN	KS	NS
14	24	K	KE	KL	EL
15	24	L	RQ	RL	QL
16	24	L	RE	RL	EL
17	24	L	SL	SV	LV
18	24	L	IL	IL	LL
19	24	L	EI	EL	IL
20	23	L	KS	KL	SL
21	23	L	DE	DL	EL
22	22	E	TL	TV	LV
23	22	L	EA	EL	AL
24	22	L	RK	RL	KL
25	22	L	DA	DL	AL
26	22	L	AL	AV	LV
27	22	I	KE	KL	EL
28	21	K	KA	KL	AL
29	21	F	TA	TI	AI
30	21	L	EL	EV	LV
31	21	L	NL	NL	LL
32	21	L	GK	GL	KL
33	21	L	LL	LP	LP
34	21	L	EL	EL	LL
35	21	L	AL	AL	LL
36	21	A	RI	RV	IV
37	20	S	KL	KV	LV
38	20	E	GK	GE	KE
39	20	E	KI	KL	IL
40	20	E	KE	KA	EA

Table C.3 ONGO's uncleaned all possible-SRS-MOTIF (continued)

Index	Hits	Centre	Group		
41	20	R	KL	KV	LV
42	20	V	RC	RI	CI
43	20	L	KE	KI	EI
44	20	L	GK	GV	KV
45	20	L	ES	EL	SL
46	20	L	RL	RV	LV
47	20	L	KA	KV	AV
48	20	L	QL	QL	LL
49	20	L	RL	RL	LL
50	20	A	GD	GL	DL
51	20	A	GE	GL	EL
52	19	D	GE	GL	EL
53	19	K	GE	GL	EL
54	19	K	ET	EF	TF
55	19	K	KQ	KL	QL
56	19	V	IL	IV	LV
57	19	V	KL	KV	LV
58	19	V	KE	KL	EL
59	19	V	GL	GV	LV
60	19	V	DL	DV	LV
61	19	L	TI	TL	IL
62	19	L	GD	GV	DV
63	19	L	GI	GL	IL
64	19	L	KL	KP	LP
65	19	L	GE	GL	EL
66	19	L	ES	EV	SV
67	19	L	DS	DL	SL
68	19	L	DL	DL	LL
69	19	L	GR	GL	RL
70	19	L	KE	KP	EP
71	19	L	KQ	KL	QL
72	19	L	SI	SL	IL
73	19	L	KA	KL	AL
74	19	I	GL	GV	LV
75	18	E	KL	KV	LV
76	18	E	RT	RV	TV
77	18	E	DE	DI	EI
78	18	E	TI	TL	IL
79	18	K	EA	EL	AL
80	18	R	EQ	EA	QA

Table C.4: TSG's uncleaned all possible-SRS-MOTIF.

Index	Hits	Centre	Group				
0	54	R	MH	MS	HS		
1	53	L	KD	KA	DA		
2	45	S	KD	KW	DW		
3	31	L	KD	KL	DL		
4	28	L	KL	KV	LV		
5	28	L	SA	SL	AL		
6	27	E	KD	KL	DL		
7	27	L	KA	KL	AL		
8	27	S	DS	DL	SL		
9	26	L	DE	DL	EL		
10	25	G	GS	GP	SP		
11	24	D	RE	RL	EL		
12	24	L	DA	DL	AL		
13	24	E	DL	DV	LV		
14	24	A	KD	KA	DA		
15	24	A	LV	LP	VP		
16	24	L	GD	GL	DL		
17	24	K	EL	EV	LV		
18	24	F	GS	GT	ST		
19	24	L	RS	RL	SL		
20	24	A	RS	RP	SP		
21	23	L	AL	AV	LV		
22	23	L	KE	KL	EL		
23	23	Y	GK	GD	KD		
24	23	E	DE	DL	EL		
25	23	M	RH	RS	HS	SS	
26	23	D	EA	EL	AL		
27	23	V	SA	SL	AL		
28	23	R	GD	GS	DS		
29	23	E	KE	KL	EL		
30	23	D	EN	EL	NL		
31	22	R	GD	GL	DL		
32	22	L	DE	DV	EV		
33	22	L	DY	DL	YL		
34	22	L	EL	EV	LV		
35	22	L	RD	RS	DS		
36	22	S	ES	EL	SL		
37	22	Y	GD	GL	DL		
38	22	L	GK	GL	KL		
39	22	L	GL	GV	LV		
40	22	R	GD	GE	DE		

Table C.4 TSG's uncleaned all possible-SRS-MOTIF (continued)

Index	Hits	Centre	Group				
41	22	E	AL	AL	LL		
42	21	E	GS	GL	SL		
43	21	S	SL	SV	LV		
44	21	A	DY	DI	YI		
45	21	A	DA	DL	AL		
46	21	A	DL	DV	LV		
47	21	E	DS	DL	SL		
48	21	D	GK	GL	KL		
49	21	L	ET	EA	TA		
50	21	D	RA	RL	AL		
51	21	E	KL	KV	LV		
52	21	R	EA	EL	AL		
53	21	A	EL	EV	LV		
54	21	S	DE	DS	ES		
55	21	L	ST	SL	TL		
56	21	L	ET	EL	TL		
57	20	G	EN	EL	NL		
58	20	L	ET	EV	TV		
59	20	E	DT	DL	TL		
60	20	D	EL	EV	LV		
61	20	R	DE	DL	EL		
62	20	V	GL	GV	LV		
63	20	L	LL	LV	LV		
64	20	D	KE	KL	EL		
65	19	R	ES	EL	SL		
66	19	E	GR	GP	RP		
67	19	E	GT	GL	TL		
68	19	L	ES	EL	SL		
69	19	T	KE	KA	EA		
70	19	S	GG	GT	GT		
71	19	E	SA	SP	AP		
72	19	V	GS	GL	SL		
73	19	E	EA	EL	AL		
74	19	R	KA	KL	AL		
75	19	L	KE	KS	ES		
76	19	S	ST	SL	TL		
77	19	A	KA	KL	AL		
78	19	E	SA	SL	AL		
79	19	E	KA	KL	AL		
80	19	D	TA	TV	AV		
81	19	V	DS	DL	SL		

Table C.4 TSG's uncleaned all possible-SRS-MOTIF (continued)

Index	Hits	Centre	Group				
82	19	K	DE	DS	ES		
83	19	A	AI	AL	IL		
84	19	R	RE	RA	EA		
85	19	R	RD	RS	DS		
86	19	L	TL	TV	LV		
87	19	E	RE	RL	EL		
88	18	S	NS	NL	SL		
89	18	L	DD	DL	DL		
90	18	S	KD	KS	DS		
91	18	E	ET	EL	TL		
92	18	A	ST	SL	TL		
93	18	R	RS	RL	SL		
94	18	S	KS	KL	SL		
95	18	L	DS	DL	SL		
96	18	L	KE	KV	EV		
97	18	L	RT	RL	TL		
98	18	L	DE	DF	EF		
99	18	R	GE	GA	EA		
100	18	A	RQ	RS	QS		
101	18	L	RD	RL	DL		
102	18	L	NA	NL	AL		
103	18	K	DE	DY	EY		
104	18	R	SL	SV	LV		
105	18	G	KA	KL	AL		
106	18	G	RA	RL	AL		
107	18	L	KS	KV	SV		
108	18	R	DE	DT	ET		
109	18	R	GE	GL	EL		
110	18	K	DE	DL	EL		
111	18	L	KD	KE	DE		
112	18	E	KS	KL	SL		
113	18	E	ES	EL	SL		
114	18	L	KE	KI	EI		
115	18	L	SL	SV	LV		
116	18	F	DE	DS	ES		
117	18	L	RL	RL	LL		
118	18	S	DE	DL	EL		
119	18	T	KS	KL	SL		
120	18	S	ES	EV	SV		
121	17	V	DT	DL	TL		
122	17	L	RE	RL	EL		

Table C.4 TSG's uncleaned all possible-SRS-MOTIF (continued)

Index	Hits	Centre	Group				
123	17	L	YA	YL	AL		
124	17	T	EL	EP	LP		
125	17	G	DT	DL	TL		
126	17	R	TL	TV	LV		
127	17	S	DL	DV	LV		
128	17	V	RS	RT	ST		
129	17	E	GS	GT	ST		
130	17	G	RE	RV	EV		
131	17	L	RK	RT	KT		
132	17	L	RS	RP	SP		
133	17	S	SL	SP	LP		
134	17	Y	ET	EL	TL		
135	17	E	RD	RL	DL		
136	17	E	ES	EA	SA		
137	17	R	DL	DL	LL		
138	17	E	KY	KA	YA		
139	17	L	RT	RA	TA		
140	17	D	AL	AV	LV		
141	17	S	DA	DV	AV		
142	17	H	MR	MS	RS	ST	ST
143	17	K	KD	KL	DL		
144	17	S	GE	GL	EL		
145	17	L	SL	SL	LL		
146	17	Y	MS	MF	SF		
147	17	V	TA	TL	AL		
148	17	N	KI	KL	IL		
149	17	G	GS	GV	SV		
150	17	P	ES	ET	ST		
151	17	L	KD	KW	DW		
152	17	A	DY	DL	YL		
153	17	D	SA	SL	AL		
154	17	G	KS	KL	SL		
155	17	S	DA	DL	AL		
156	17	L	RL	RV	LV		
157	17	T	KD	KE	DE		
158	17	L	RS	RV	SV		
159	17	S	SS	SL	SL		
160	17	F	DE	DV	EV		
161	17	L	RE	RS	ES		
162	17	P	ET	EF	TF		
163	17	D	RT	RV	TV		

Table C.4 TSG's uncleaned all possible-SRS-MOTIF (continued)

Index	Hits	Centre	Group				
164	17	R	GY	GI	YI		
165	17	L	DT	DA	TA		
166	17	P	GS	GL	SL		
167	17	S	RE	RP	EP		
168	17	R	DL	DV	LV		
169	17	E	DE	DS	ES		
170	16	V	SL	SF	LF		
171	16	V	SV	SP	VP		
172	16	A	SA	SL	AL		
173	16	V	KE	KT	ET		
174	16	L	LF	LV	FV		
175	16	R	KD	KA	DA		
176	16	S	QL	QV	LV		
177	16	S	KD	KE	DE		
178	16	S	KL	KV	LV		
179	16	V	RD	RL	DL		
180	16	T	KD	KS	DS		
181	16	T	GQ	GS	QS		
182	16	L	GA	GL	AL		
183	16	L	EA	EL	AL		
184	16	L	RT	RV	TV		
185	16	S	EE	EL	EL		
186	16	T	RN	RL	NL		
187	16	K	KI	KL	IL		
188	16	S	GD	GS	DS		
189	16	L	KE	KA	EA		
190	16	S	KA	KL	AL		
191	16	E	KD	KE	DE		
192	16	E	AL	AV	LV		
193	16	V	ST	SV	TV		
194	16	V	DE	DL	EL		
195	16	A	GE	GF	EF		
196	16	M	HS	HS	SS		
197	16	Y	KD	KS	DS		
198	16	V	KE	KL	EL		
199	16	V	EL	EL	LL		
200	16	D	EE	EL	EL		
201	16	L	KD	KS	DS		
202	16	E	EA	EV	AV		
203	16	A	ES	EP	SP		
204	16	T	KT	KA	TA		

Table C.4 TSG's uncleaned all possible-SRS-MOTIF (continued)

Index	Hits	Centre	Group				
205	16	E	RK	RL	KL		
206	16	K	GE	GL	EL		
207	16	L	NS	NL	SL		
208	16	L	QT	QL	TL		
209	16	Y	DA	DL	AL		
210	16	G	EL	EV	LV		
211	16	E	RL	RP	LP		
212	16	I	KE	KL	EL		
213	16	K	ES	EL	SL		
214	16	A	EA	EV	AV		
215	16	G	DL	DL	LL		
216	16	G	DS	DL	SL		
217	16	V	DS	DA	SA		
218	16	R	SA	SV	AV		
219	16	A	KE	KL	EL		
220	16	V	RT	RV	TV		
221	16	A	GE	GA	EA		
222	16	L	ES	EA	SA		
223	16	L	GQ	GV	QV		
224	16	D	KE	KS	ES		
225	16	E	RA	RV	AV		
226	16	P	ES	EL	SL		
227	16	S	GT	GV	TV		
228	16	N	EL	EV	LV		
229	16	T	ES	EA	SA		
230	16	L	AL	AL	LL		
231	16	L	EL	EL	LL		
232	16	A	ES	ET	ST		
233	16	L	TA	TL	AL		
234	16	E	EE	EL	EL		
235	16	G	SL	SV	LV		
236	16	A	EA	EL	AL		
237	16	K	DS	DL	SL		

Table C.5: ONGO's unique possible-SRS-MOTIFs.

New Index	Old Index	Hits	Centre	Group		
0	0	40	F	ES	ET	ST
1	1	39	M	RT	RA	TA
2	2	32	D	RE	RT	ET
3	4	28	K	NS	NS	SS

Table C.5 ONGO's unique possible-SRS-MOTIFs. (continued)

New Index	Old Index	Hits	Centre	Group		
4	5	28	L	KI	KL	IL
5	8	27	E	RT	RL	TL
7	11	26	G	DE	DV	EV
8	13	24	D	KN	KS	NS
9	14	24	K	KE	KL	EL
10	15	24	L	RQ	RL	QL
11	18	24	L	IL	IL	LL
12	19	24	L	EI	EL	IL
13	20	23	L	KS	KL	SL
14	22	22	E	TL	TV	LV
15	24	22	L	RK	RL	KL
16	28	21	K	KA	KL	AL
17	29	21	F	TA	TI	AI
18	31	21	L	NL	NL	LL
19	33	21	L	LL	LP	LP
20	36	21	A	RI	RV	IV
21	38	20	E	GK	GE	KE
22	39	20	E	KI	KL	IL
23	40	20	E	KE	KA	EA
24	41	20	R	KL	KV	LV
25	42	20	V	RC	RI	CI
26	44	20	L	GK	GV	KV
27	47	20	L	KA	KV	AV
28	48	20	L	QL	QL	LL
29	50	20	A	GD	GL	DL
30	51	20	A	GE	GL	EL
31	52	19	D	GE	GL	EL
32	54	19	K	ET	EF	TF
33	55	19	K	KQ	KL	QL
34	56	19	V	IL	IV	LV
35	57	19	V	KL	KV	LV
36	60	19	V	DL	DV	LV
37	61	19	L	TI	TL	IL
38	62	19	L	GD	GV	DV
39	63	19	L	GI	GL	IL
40	64	19	L	KL	KP	LP
41	65	19	L	GE	GL	EL
43	68	19	L	DL	DL	LL
44	69	19	L	GR	GL	RL
45	70	19	L	KE	KP	EP
46	71	19	L	KQ	KL	QL

Table C.5 ONGO's unique possible-SRS-MOTIFs. (continued)

New Index	Old Index	Hits	Centre	Group		
47	72	19	L	SI	SL	IL
48	74	19	I	GL	GV	LV
49	76	18	E	RT	RV	TV
50	77	18	E	DE	DI	EI
51	78	18	E	TI	TL	IL
52	79	18	K	EA	EL	AL
53	80	18	R	EQ	EA	QA
54	81	18	V	ML	MV	LV
55	84	18	L	RI	RL	IL
56	85	18	L	TL	TL	LL
57	87	18	L	LV	LP	VP
58	88	18	L	IL	IV	LV
59	90	18	I	EL	EP	LP
60	91	18	A	GL	GV	LV
61	92	18	A	EI	EL	IL
62	93	18	T	KD	KL	DL
63	94	18	T	KL	KP	LP
64	95	18	G	EL	EL	LL
65	96	18	G	AL	AV	LV
66	98	17	N	DL	DV	LV
67	99	17	E	GK	GL	KL
68	100	17	E	IL	IV	LV
69	101	17	D	DL	DV	LV
70	102	17	D	TL	TV	LV
71	105	17	D	GT	GA	TA
72	106	17	D	QL	QP	LP
73	108	17	K	DL	DV	LV
74	110	17	K	GL	GV	LV
75	111	17	R	IL	IV	LV
76	113	17	V	RC	RA	CA
77	114	17	V	LL	LP	LP
78	115	17	V	GD	GE	DE
79	116	17	V	EL	EV	LV
80	119	17	L	GS	GV	SV
81	120	17	L	GE	GV	EV
83	122	17	L	SY	SL	YL
84	123	17	L	GR	GD	RD
85	124	17	L	SI	SV	IV
86	125	17	L	KA	KI	AI
87	126	17	L	NL	NV	LV
88	127	17	L	DQ	DL	QL

Table C.5 ONGO's unique possible-SRS-MOTIFs. (continued)

New Index	Old Index	Hits	Centre	Group		
89	128	17	L	GE	GP	EP
90	129	17	L	DS	DV	SV
91	130	17	I	EL	EV	LV
92	131	17	C	SY	SL	YL
93	132	17	Y	QA	QL	AL
94	134	17	G	KL	KV	LV
95	135	16	S	GL	GL	LL
96	136	16	Q	KA	KL	AL
97	137	16	Q	EA	EL	AL
98	139	16	E	GA	GL	AL
99	140	16	E	KL	KL	LL
100	141	16	E	RI	RL	IL
101	142	16	E	RQ	RV	QV
102	143	16	D	DS	DL	SL
103	145	16	D	QS	QL	SL
104	146	16	D	RS	RL	SL
105	147	16	K	KN	KS	NS
106	148	16	K	EL	EP	LP
107	149	16	K	KD	KA	DA
108	151	16	K	DA	DL	AL
109	152	16	R	QL	QV	LV
110	155	16	R	RQ	RL	QL
111	156	16	V	RE	RI	EI
112	158	16	V	AL	AV	LV
113	159	16	V	SL	SV	LV
114	160	16	V	KA	KL	AL
115	161	16	V	RE	RL	EL
116	162	16	V	EI	EL	IL
117	163	16	F	GK	GL	KL
118	164	16	L	KT	KY	TY
119	165	16	L	DT	DL	TL
120	166	16	L	KK	KL	KL
121	167	16	L	AI	AV	IV
122	169	16	L	DL	DP	LP
123	170	16	L	ES	EI	SI
124	173	16	L	KA	KF	AF
125	175	16	L	DA	DV	AV
126	176	16	L	GE	GA	EA
127	177	16	L	TL	TP	LP
128	178	16	L	RD	RV	DV
129	179	16	I	KS	KI	SI

Table C.5 ONGO's unique possible-SRS-MOTIFs. (continued)

New Index	Old Index	Hits	Centre	Group		
130	180	16	I	RK	RE	KE
131	181	16	I	IL	IV	LV
132	182	16	I	QS	QV	SV
133	183	16	A	GR	GT	RT
134	184	16	Y	ET	EV	TV
135	185	16	Y	EQ	EI	QI
136	186	16	T	MK	MV	KV
137	187	16	T	EL	EV	LV
138	188	16	G	ES	EL	SL
Total hits		2403	N/A			

Table C.6: TSG's unique possible-SRS-MOTIFs.

New Index	Old Index	Hits	Center	Group				
0	0	54	R	MH	MS	HS		
1	1	53	L	KD	KA	DA		
2	2	45	S	KD	KW	DW		
3	6	27	E	KD	KL	DL		
4	10	25	G	GS	GP	SP		
5	11	24	D	RE	RL	EL		
6	13	24	E	DL	DV	LV		
7	14	24	A	KD	KA	DA		
8	15	24	A	LV	LP	VP		
9	18	24	F	GS	GT	ST		
10	19	24	L	RS	RL	SL		
11	20	24	A	RS	RP	SP		
12	23	23	Y	GK	GD	KD		
13	24	23	E	DE	DL	EL		
14	25	23	M	RH	RS	HS	SS	
15	27	23	V	SA	SL	AL		
16	28	23	R	GD	GS	DS		
17	29	23	E	KE	KL	EL		
18	30	23	D	EN	EL	NL		
19	31	22	R	GD	GL	DL		
20	33	22	L	DY	DL	YL		
21	35	22	L	RD	RS	DS		
22	36	22	S	ES	EL	SL		
23	37	22	Y	GD	GL	DL		
24	40	22	R	GD	GE	DE		
25	41	22	E	AL	AL	LL		
26	42	21	E	GS	GL	SL		

Table C.6 TSG's unique possible-SRS-MOTIFs. (continued)

New Index	Old Index	Hits	Center	Group				
27	43	21	S	SL	SV	LV		
28	44	21	A	DY	DI	YI		
29	45	21	A	DA	DL	AL		
30	46	21	A	DL	DV	LV		
31	47	21	E	DS	DL	SL		
32	49	21	L	ET	EA	TA		
33	53	21	A	EL	EV	LV		
34	54	21	S	DE	DS	ES		
35	55	21	L	ST	SL	TL		
36	57	20	G	EN	EL	NL		
37	58	20	L	ET	EV	TV		
38	59	20	E	DT	DL	TL		
39	60	20	D	EL	EV	LV		
40	61	20	R	DE	DL	EL		
42	65	19	R	ES	EL	SL		
43	66	19	E	GR	GP	RP		
44	67	19	E	GT	GL	TL		
45	69	19	T	KE	KA	EA		
46	70	19	S	GG	GT	GT		
47	71	19	E	SA	SP	AP		
48	73	19	E	EA	EL	AL		
49	74	19	R	KA	KL	AL		
50	75	19	L	KE	KS	ES		
51	76	19	S	ST	SL	TL		
52	77	19	A	KA	KL	AL		
53	78	19	E	SA	SL	AL		
54	79	19	E	KA	KL	AL		
55	80	19	D	TA	TV	AV		
56	81	19	V	DS	DL	SL		
57	82	19	K	DE	DS	ES		
58	83	19	A	AI	AL	IL		
59	84	19	R	RE	RA	EA		
60	85	19	R	RD	RS	DS		
61	86	19	L	TL	TV	LV		
62	87	19	E	RE	RL	EL		
63	88	18	S	NS	NL	SL		
64	89	18	L	DD	DL	DL		
65	90	18	S	KD	KS	DS		
66	91	18	E	ET	EL	TL		
67	92	18	A	ST	SL	TL		
68	93	18	R	RS	RL	SL		

Table C.6 TSG's unique possible-SRS-MOTIFs. (continued)

New Index	Old Index	Hits	Center	Group				
69	94	18	S	KS	KL	SL		
70	97	18	L	RT	RL	TL		
71	98	18	L	DE	DF	EF		
72	99	18	R	GE	GA	EA		
73	100	18	A	RQ	RS	QS		
74	102	18	L	NA	NL	AL		
75	103	18	K	DE	DY	EY		
76	105	18	G	KA	KL	AL		
77	106	18	G	RA	RL	AL		
78	107	18	L	KS	KV	SV		
79	108	18	R	DE	DT	ET		
80	109	18	R	GE	GL	EL		
81	111	18	L	KD	KE	DE		
82	112	18	E	KS	KL	SL		
83	113	18	E	ES	EL	SL		
84	116	18	F	DE	DS	ES		
85	118	18	S	DE	DL	EL		
86	119	18	T	KS	KL	SL		
87	120	18	S	ES	EV	SV		
88	121	17	V	DT	DL	TL		
89	123	17	L	YA	YL	AL		
90	124	17	T	EL	EP	LP		
91	125	17	G	DT	DL	TL		
92	126	17	R	TL	TV	LV		
93	127	17	S	DL	DV	LV		
94	128	17	V	RS	RT	ST		
95	129	17	E	GS	GT	ST		
96	130	17	G	RE	RV	EV		
97	131	17	L	RK	RT	KT		
98	132	17	L	RS	RP	SP		
99	133	17	S	SL	SP	LP		
100	134	17	Y	ET	EL	TL		
101	135	17	E	RD	RL	DL		
102	136	17	E	ES	EA	SA		
103	137	17	R	DL	DL	LL		
104	138	17	E	KY	KA	YA		
105	139	17	L	RT	RA	TA		
106	140	17	D	AL	AV	LV		
107	141	17	S	DA	DV	AV		
108	142	17	H	MR	MS	RS	ST	ST
109	144	17	S	GE	GL	EL		

Table C.6 TSG's unique possible-SRS-MOTIFs. (continued)

New Index	Old Index	Hits	Center	Group				
110	146	17	Y	MS	MF	SF		
111	147	17	V	TA	TL	AL		
112	148	17	N	KI	KL	IL		
113	149	17	G	GS	GV	SV		
114	150	17	P	ES	ET	ST		
115	151	17	L	KD	KW	DW		
116	152	17	A	DY	DL	YL		
117	153	17	D	SA	SL	AL		
118	154	17	G	KS	KL	SL		
119	155	17	S	DA	DL	AL		
120	157	17	T	KD	KE	DE		
121	158	17	L	RS	RV	SV		
122	159	17	S	SS	SL	SL		
123	160	17	F	DE	DV	EV		
124	161	17	L	RE	RS	ES		
125	162	17	P	ET	EF	TF		
126	163	17	D	RT	RV	TV		
127	164	17	R	GY	GI	YI		
128	165	17	L	DT	DA	TA		
129	166	17	P	GS	GL	SL		
130	167	17	S	RE	RP	EP		
131	169	17	E	DE	DS	ES		
132	170	16	V	SL	SF	LF		
133	171	16	V	SV	SP	VP		
134	172	16	A	SA	SL	AL		
135	173	16	V	KE	KT	ET		
136	174	16	L	LF	LV	FV		
137	175	16	R	KD	KA	DA		
138	176	16	S	QL	QV	LV		
139	177	16	S	KD	KE	DE		
140	179	16	V	RD	RL	DL		
141	180	16	T	KD	KS	DS		
142	181	16	T	GQ	GS	QS		
143	184	16	L	RT	RV	TV		
144	185	16	S	EE	EL	EL		
145	186	16	T	RN	RL	NL		
146	187	16	K	KI	KL	IL		
147	188	16	S	GD	GS	DS		
148	189	16	L	KE	KA	EA		
149	190	16	S	KA	KL	AL		
150	191	16	E	KD	KE	DE		

Table C.6 TSG's unique possible-SRS-MOTIFs. (continued)

New Index	Old Index	Hits	Center	Group				
151	193	16	V	ST	SV	TV		
152	194	16	V	DE	DL	EL		
153	195	16	A	GE	GF	EF		
154	196	16	M	HS	HS	SS		
155	197	16	Y	KD	KS	DS		
156	199	16	V	EL	EL	LL		
157	200	16	D	EE	EL	EL		
158	201	16	L	KD	KS	DS		
159	202	16	E	EA	EV	AV		
160	203	16	A	ES	EP	SP		
161	204	16	T	KT	KA	TA		
162	205	16	E	RK	RL	KL		
163	207	16	L	NS	NL	SL		
164	208	16	L	QT	QL	TL		
165	209	16	Y	DA	DL	AL		
166	211	16	E	RL	RP	LP		
167	213	16	K	ES	EL	SL		
168	214	16	A	EA	EV	AV		
169	215	16	G	DL	DL	LL		
170	216	16	G	DS	DL	SL		
171	217	16	V	DS	DA	SA		
172	218	16	R	SA	SV	AV		
173	219	16	A	KE	KL	EL		
174	220	16	V	RT	RV	TV		
175	221	16	A	GE	GA	EA		
176	224	16	D	KE	KS	ES		
177	225	16	E	RA	RV	AV		
178	226	16	P	ES	EL	SL		
179	227	16	S	GT	GV	TV		
180	228	16	N	EL	EV	LV		
181	229	16	T	ES	EA	SA		
182	232	16	A	ES	ET	ST		
183	233	16	L	TA	TL	AL		
184	234	16	E	EE	EL	EL		
185	235	16	G	SL	SV	LV		
186	236	16	A	EA	EL	AL		
187	237	16	K	DS	DL	SL		
Total hits		3280	N/A					

Table C.7: Excaustive earch of Possible SRS-MOTIF detection Hashmap

Residue	Indices	Amino acids (Residue)																			
		G	M	R	K	D	E	Q	N	H	S	T	Y	C	W	A	I	L	F	V	P
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
G		1	2	4	7	11	16	22	29	37	46	56	67	79	92	106	121	137	154	172	191
M	1	2	3	5	8	12	17	23	30	38	47	57	68	80	93	107	122	138	155	173	192
R	2	4	5	6	9	13	18	24	31	39	48	58	69	81	94	108	123	139	156	174	193
K	3	7	8	9	10	14	19	25	32	40	49	59	70	82	95	109	124	140	157	175	194
D	4	11	12	13	14	15	20	26	33	41	50	60	71	83	96	110	125	141	158	176	195
E	5	16	17	18	19	20	21	27	34	42	51	61	72	84	97	111	126	142	159	177	196
Q	6	22	23	24	25	26	27	28	35	43	52	62	73	85	98	112	127	143	160	178	197
N	7	29	30	31	32	33	34	35	36	44	53	63	74	86	99	113	128	144	161	179	198
H	8	37	38	39	40	41	42	43	44	45	54	64	75	87	100	114	129	145	162	180	199
S	9	46	47	48	49	50	51	52	53	54	55	65	76	88	101	115	130	146	163	181	200
T	10	56	57	58	59	60	61	62	63	64	65	66	77	89	102	116	131	147	164	182	201
Y	11	67	68	69	70	71	72	73	74	75	76	77	78	90	103	117	132	148	165	183	202
C	12	79	80	81	82	83	84	85	86	87	88	89	90	91	104	118	133	149	166	184	203
W	13	92	93	94	95	96	97	98	99	100	101	102	103	104	105	119	134	150	167	185	204
A	14	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	135	151	168	186	205
I	15	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	152	169	187	206
L	16	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152	153	170	188	207
F	17	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	189	208
V	18	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	209
P	19	191	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208	209	210

Example: If the center amino acid is D, then it is represented by 15

Vertices NF by 161

Vertices FN are represented by 161

Both are represented by 161

Appendix D

Level 3 RV coefficient calculation

D.1 Properties with negative representation for RV coefficient calculation

Table D.1: Amino acids with negative property representation (first half)

Amino acid short form			Arg	Lys	Asp	Glu	Gln	Asn	His
Amino acid code			R	K	D	E	Q	N	H
Properties	Web reference	charged	1	1	1	1	-1	-1	-1
		Polar	-1	-1	-1	-1	1	1	1
		Hydrophobic	-1	-1	-1	-1	-1	-1	-1
	Soluable reference	Hydrophobic	-1	-1	-1	-1	-1	-1	-1
		Moderate	-1	-1	-1	-1	-1	-1	1
		Hydrophillic	1	1	1	1	1	1	-1
		polar	-1	-1	-1	-1	1	1	-1
		Aromatic	-1	-1	-1	-1	-1	-1	-1
		Aliphatic	-1	-1	-1	-1	-1	-1	-1
		Acidic	-1	-1	1	1	-1	-1	-1
		Basic	1	1	-1	-1	-1	-1	1
		Negative charge	-1	-1	1	1	-1	-1	-1
		Neutral	-1	-1	-1	-1	1	1	-1
		Positive charge	1	1	-1	-1	-1	-1	1
		Pka_NH2	0.15	0.75	0.40	0.44	0.17	-1.00	0.09
		P_ka_COOH	0.07	1.00	0.02	0.07	0.06	0.04	0.01

Table D.2: Amino acids with negative property representation (second half)

Amino acid short form			Ser	Thr	Tyr	Cys	Met	Trp	Ala	Ile	Lue	Phe	Val	Pro	Gly
Amino acid code			S	T	Y	C	M	W	A	I	L	F	V	P	G
Properties	Web reference	charged	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
		Polar	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1
		Hydrophobic	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1
	Soluable reference	Hydrophobic	-1	-1	1	-1	-1	1	1	1	1	1	1	1	1
		Moderate	-1	-1	-1	1	1	-1	-1	-1	-1	-1	-1	-1	-1
		Hydrophillic	1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
		polar	1	1	-1	1	1	-1	-1	-1	-1	-1	-1	-1	-1
		Aromatic	-1	-1	1	-1	-1	1	-1	-1	-1	1	-1	-1	-1
		Aliphatic	-1	-1	-1	-1	-1	-1	1	1	1	-1	1	-1	-1
		Acidic	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
		Basic	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
		Negative charge	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
		Neutral	1	1	1	1	1	1	1	1	1	1	1	1	1
		Positive charge	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
		Pka_NH2	0.18	0.16	0.16	1.00	0.21	0.30	0.54	0.48	0.40	0.22	0.46	0.91	0.40
		P_ka_COOH	0.07	0.06	0.07	-1.00	0.08	0.09	0.09	0.08	0.09	0.12	0.08	0.04	0.09

D.2 Properties of RV coefficients' Samples of Edge and Vertices representation

Table D.3: Properties of RV_2 coefficients' Samples of Edge and Vertices representation

(a) Portion of edge representations' RV_2 coefficients among ONGO possible-SRS-MOTIF groups ONGO possible-SRS-MOTIF groups Vs TSG possible-SRS-MOTIF groups

Groups		TSG possible-SRS-MOTIF									
ONGO possible -SRS-MOTIF	index	0	·	106	107	108	109	110	·	186	Average
	0	-0.122	·	0.667	0.596	-0.199	0.757	0.450	·	0.644	0.506
	1	0.796	·	-0.123	0.619	0.829	0.615	0.635	·	-0.134	0.286
	:	:	:	:	:	:	:	:	:	:	:
	137	0.271	·	0.523	1.000	0.225	0.870	0.497	·	0.395	0.561
	138	0.057		0.781	0.663	-0.011	0.864	0.412		0.794	0.601
	Average	0.100		0.588	0.449	-0.014	0.553	0.197		0.628	0.100

(b) Portion of vertices representations' RV_2 coefficients among ONGO possible-SRS-MOTIF groups

Group		ONGO possible-SRS-MOTIF						
ONGO possible -SRS-MOTIF	index	0	1	2	···	136	137	138
	0	1.000	0.759	0.565	···	0.762	0.710	0.767
	1	0.759	1.000	0.341	···	0.982	0.655	0.683
	2	0.565	0.341	1.000	···	0.261	0.304	0.330
	:	:	:	:	:	:	:	:
	136	0.762	0.982	0.261	···	1.000	0.682	0.701
	137	0.710	0.655	0.304	···	0.682	1.000	0.961
	138	0.767	0.683	0.330	···	0.701	0.961	1.000

(c) 108^{th} index of TSG's unique possible-SRS-MOTIF

New Index	Old Index	Hits	Centre	Group				
108	142	17	H	MR	MS	RS	ST	ST

Here the group index directly related new indexes of the groups presented in the Table. C.5 (ONGO's possible-MOTIF groups) and Table. C.6 (TSG's possible-MOTIF groups). Such that from Table. C.6 TSG's 108^{th} index points to group as shown in Table. D.3c.

The Table. D.3 only represent portion of RV_2 coefficients; In Table D.3a the 0^{th} index of ONGO possible-SRS-MOTIF group and 0^{th} index of TSG possible-SRS-MOTIF group precents the RV_2 coefficient between those groups.

Table. D.3b presents the portion of vertices representations' RV_2 coefficients among ONGO possible-SRS-MOTIF groups. The diagonal elements of the matrix are 1.000, because the diagonal RV_2 coefficients present the coefficient calculation between the similar substructures. Likewise, 3 matrixes between the groups such as,

- ✓ Among only the ONGO groups
- ✓ Among only the TSG groups
- ✓ Among ONGO Vs TSG groups

are calculated for $RV2\_motif\_normalized\_negated\_coefficient\_edge$ and presented in
`https://drive.google.com/file/d/122gW6-v27eVcZfMOtF_nIeBET8bohfad/view?usp=sharing`

and,

$RV2\_motif\_normalized\_negated\_coefficient\_vertices$ presented in `https://drive.google.com/file/d/14bJnF3GbGlZxIW4Cz8nlK7icqsoXblse/view?usp=sharing`.