Using statistics of Patterns find the Possible-SRS-MOTIFs

Primary Author
A. Nishanth

Primary Supervisor
Dr. C. H. Chu

November 26, 2020

Contents

Chapter 1

Using statistics of Patterns find the Possible-SRS-MOTIFs

For finding a new algorithm for the SRS-MOTIF patterns detection. The algorithm should find the SRS-MOTIF patterns and their networks repetition (Network of SRS-MOTIFs) which contributes to the main function of the PDB structure.

First separate the training set and testing set, follow the [2] to do this.

1.1 Step 1

First extract the available $SITE$ patterns (substructures) on the surface (the surface detectionof the PDBs is done as mentioned in [2]), and pool them into separate classes like ONGO and TSG accordingly, the created dataset named as “Data-1”.

The substructures are extracted from V91 census data, the Tier 1's ONGO and TSG's all PDBs which have $SITE$ information is extracted (No filtering based on primary structural length > 81). Then the overlapping PDBs among the ONGO and TSG is extracted; these overlapping PDBs are left in whole $MOTIF$ experiment (since these PDB structure's functionality is undefinable/ no use in this experiment). The PDBs have at least one $C\alpha$ surface atom by MSMS tool is used; thus, PDBs such as “4MDQ” and “721P” (both belongs to ONGO class) were left. Rest of the selected (those has $SITE$ information, satisfied by $MSMS$, and at least has one $C\alpha$ surface atom) PDB_{SITE}s are used in the experiment.

1.2 Step 2

From the PDB_{SITE}s, their corresponding $SITE$ substructures' number of residues are extracted to form a statistic of number of residues per 3D-MOTIF (or $SITE$ information by $SOFTWARE$ or $Human$) group. The details are shown in Table. 1.1; and the histogram shown in Fig. 1.1 visualize the distribution of frequency (number of) residues per surface (soft threshold and All $C\alpha$) MOTIF group. From the Table. 1.1, it can be said, surface $C\alpha$ atom definition by $soft$ threshold reduces the

Table 1.1: Frequency of Number of residues per MOTIF groups summery

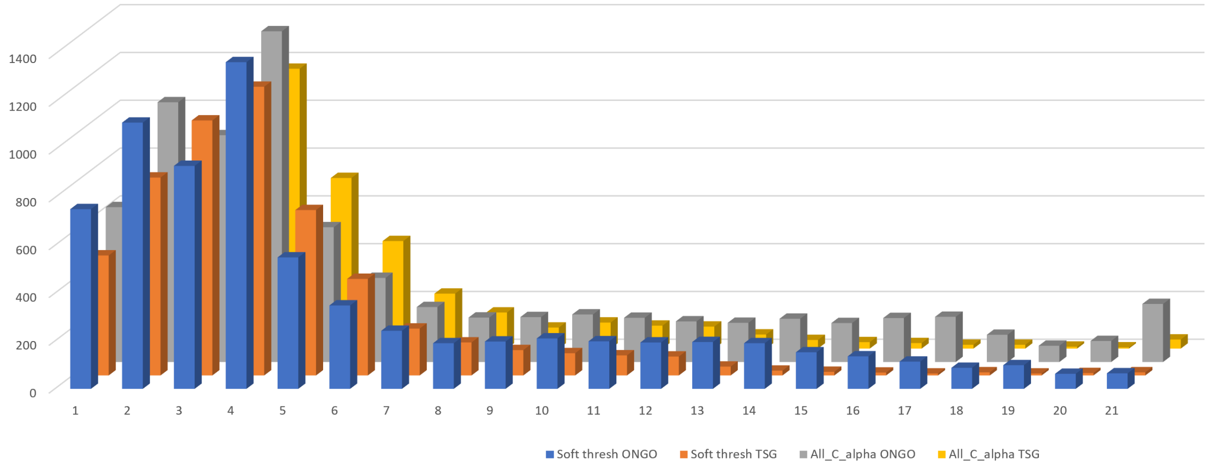
Number of residues in the group	Soft thresh		All $C\alpha$		Soft thresh				All $C\alpha$			
					ONGO		TSG		ONGO		TSG	
	ONGO	TSG	ONGO	TSG	Cumulative	Cumulative (%)	Cumulative	Cumulative (%)	Cumulative	Cumulative (%)	Cumulative	Cumulative (%)
1	752	502	648	459	752	10.15	502	9.07	648	8.60	459	8.21
2	1114	828	1087	763	1866	25.20	1330	24.04	1735	23.02	1222	21.87
3	933	1067	949	1037	2799	37.79	2397	43.32	2684	35.62	2259	40.43
4	1367	1208	1384	1171	4166	56.25	3605	65.15	4068	53.98	3430	61.38
5	550	692	564	714	4716	63.68	4297	77.66	4632	61.46	4144	74.16
6	349	404	352	450	5065	68.39	4701	84.96	4984	66.14	4594	82.21
7	243	196	230	230	5308	71.67	4897	88.51	5214	69.19	4824	86.33
8	191	138	186	151	5499	74.25	5035	91.00	5400	71.66	4975	89.03
9	198	106	188	88	5697	76.92	5141	92.92	5588	74.15	5063	90.60
10	211	93	199	109	5908	79.77	5234	94.60	5787	76.79	5172	92.56
11	199	84	185	96	6107	82.46	5318	96.11	5972	79.25	5268	94.27
12	194	79	170	92	6301	85.08	5397	97.54	6142	81.50	5360	95.92
13	196	37	164	59	6497	87.73	5434	98.21	6306	83.68	5419	96.98
14	191	20	181	36	6688	90.31	5454	98.57	6487	86.08	5455	97.62
15	153	15	163	27	6841	92.37	5469	98.84	6650	88.24	5482	98.10
16	136	12	184	23	6977	94.21	5481	99.06	6834	90.68	5505	98.51
17	114	8	189	16	7091	95.75	5489	99.20	7023	93.19	5521	98.80
18	88	13	114	16	7179	96.93	5502	99.44	7137	94.71	5537	99.09
19	99	8	68	8	7278	98.27	5510	99.58	7205	95.61	5545	99.23
20	63	10	88	5	7341	99.12	5520	99.77	7293	96.78	5550	99.32
21	65	13	243	38	7406	100	5533	100	7536	100	5588	100
All	7406	5533	7536	5588								

Note: In All $C\alpha$ group of ONGO PDB "4I51" has at-least one residue (that do not fall among these 21-amino acids) thus that is left for only All $C\alpha$ group

.

number of residues per group.

Figure 1.1: Histogram of number residues per surface (soft threshold)/ All $C\alpha$ MOTIF group.



Since the number of residues per MOTIF-substructure varies, thus these substructures' residues (amino acids; using the $C\alpha$ atoms of their corresponding amino acids) and the distances should be evaluated separately. For, initial evaluation small substructures (MOTIF group contain < 5 residues) are considered. There, center residue and distance from the center residue evaluated.

1.2.1 Define small substructures

If small substructures are clearly separated (showing statistical difference) depends on the class (they fall ONGO/ TSG). Then, these substructure groups' probabilities (using Bayes rule and Naïve Bayes theorem) can be used to identify the PDB structure is ONGO or TSG.

Available MOTIF substructures' statistics can be used to define the new possible-SRS-MOTIF substructures. To make statistics, initially the MOTIF groups are categorized based on the number of residues presented. These MOTIF substructures can be analyzed based on the number of residues presented in them. To analyze these MOTIF substructures (groups), different data structures are proposed, based on the number of residues presented in the MOTIF substructure as shown in Table. 1.2a. In these data structures the residues (amino acids) are presented by integer; and the integer value is obtained from the hash table as shown in Table. 1.2b.

Just taking the statistics of groups separately, groups one and two are not much interested in this context. Because the study is focused on finding possible SRS-MOTIF, at least three residues must be presented to form a surface.

Way-1

The initial evaluation of $way-1$ does not consider detailed information about the residues combinations presented in the groups at all; just give weightage for the center residue (for an **E.g.:** if the subgroup contain four residues and center atom “C” with “R”, “R”, “Q” residues given hits to each in the four group array, likewise another subgroup contain four residues and center atom “C” with , “Q”, “C”, “V” then the four group array give hits for those residues; thus this array representation (for each number of residues presented in the group, **E.g.:** group three array, group four array and group five array) not preserve the sub group information “R”, “R”, “Q” with “C” nor “Q”, “C”, “V” with “C”).

To show the naïve approach planned; Just took the counts of only the four residues group small substructures' center residue “C” (“CYS”) with the same residue. (results of that selected substructure) occurred higher in ONGO class higher than TSG.

The probability of substructure as ONGO or TSG can be calculated based on this substructure's number of occurrences among the ONGO or TSG PDBs accordingly. In order to do this just only consider these groups.(Just take this population; assume only ONGO and TSG is the classes exist; since define unknown function is not straightforward; for an **E.g.:** if the PDB structure contain ONGO or TSG related structure unless it is activated or not by enzyme).

Using the soft-threshold surface detection to define/ count the occurrence among the PDBs, Annotations/abbreviation used in probability:

✓ num_{res} = number of residues group

✓ CR= Center residue

✓ Occ = occurrence of given residue belongs to one of group residue

The probability is good enough to classify (as ONGO); and the occurrence of this kind of substructure is very good enough for define constrain as center residue as within the ADFC provided 8.46.

Table 1.2: Data structure

(a) Discription with number of residues presented in the group

group with		Main approach such the statistics based on	Implemented data structure (Arrays/matrixes presented here are updated based on hash table index mapping shown in Table. 1.2b); for each class, the matrix/array is implemented separately
1 res		their frequency of the residue occurrence	an array-of 21; and each time amino acids occurred in the specific class and the corresponding arrays' amino acid count is increased
2 res		their frequency of both occurred together	A diagonal array-of 21×21 array and update based on the indexes of both amino acids' indexes. To avoid delays the 21×21 is preferred; $(21 \times (21 + 1))/2$ elements; where diagonal can be there to increase if the residue occurred more (memory efficient); it almost like adjacency matrix representation.
3 or more residues (res)	$Way - 1$	This can be considered as it is; get the stat of center atom and their group of amino acid together	From Hash table get the index (center residue as key); from the hash table get the indexes of the rest of the residue's occurrence as well. Thus, the array of 21×21 contain the occurrence-table. And another array contains the distance addition and divided by their occurrence--table find the average distance.
	$Way - 2$	Any number of residue groups presented only with 3 residues: Any of the MOTIF groups (contain 3 residues) can be presented as center residue combined with other residues, or group in of three all	Data structure of occurrence table is same as above only difference here is each center residue has separate 21×21 occurrence-table. And if the MOTIF groups contain more than three residues, make all the possibilities of three groups and the occurrence table is filled.

(b) Hash table of residue to index

Residue/Key	R	K	D	E	Q	N	H	S	T	Y	C	M	W	A	I	L	F	V	P	G	U
Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21

On the other hand, if the other substructures have higher differentially of occurrence in ONGO Vs TSG, like the group mentioned above, then they can be combined to classify. Unfortunately other substructures does not show that much differentiation from that statistics (shown in *Appendix* A); they((P($num_{res} = n$ >2, CR = given center residue, Occ = given residue) always lie > near to 0.1. Thus, this way is not efficient confirm the classification ONGO Vs TSG.

Figure 1.2: Probability calculations of statistics including $Way - 1$

$$
\begin{aligned}
&P\,(ONGO) \;=\; \frac{7406}{(7406+5533)} && = \; 0.572\\
&P\,(num_{res} = 4) \;=\; \frac{(1367+1208)}{(7406+5533)} && = \; 0.199\\
&P\,(ONGO \mid num_{res} = 4) = \frac{P\;\,(ONGO,\; num_{res} = 4)}{P\,(num_{res} = 4)} &&\\
&= \frac{1367/(7406+5533)}{(1367+1208)/(7406+5533)} \;=\; \frac{1367}{(1367+1208)} && = \; 0.531\\
&P\,(ONGO, num_{res} = 4,\; CR = \text{“}C\text{”},\; Occ = \text{“}C\text{”}) \;=\; \frac{1092}{7406+5533} && = \; 0.084\\
&P\,(num_{res} = 4,\; CR = \text{“}C\text{”},\; Occ = \text{“}C\text{”}) \;=\; \frac{(1092+381)}{(7390+5531)} && = \; 0.113\\
&P(ONGO|\; num_{res} = 4,\; CR = \text{“}C\text{”},\; Occ = \text{“}C\text{”}) &&\\
&= \frac{P(ONGO,\; num_{res} = 4,\; CR = \text{“}C\text{”}, Occ = \text{“}C\text{”})}{P(num_{res} = 4,\; CR = \text{“}C\text{”},\; Occ = \text{“}C\text{”})} \;=\; \frac{0.084}{0.113} && = \; 0.743
\end{aligned}
$$

From the initial evaluation of $way-1$, those statistics reported in Fig. 1.2, does not show that much of information to move forward in the direction. Only one possibility is using select all possible left “C” as center residue and find the statistics between them. And there is already paper [1] published on 2019 supports this finding related to residue “C” in cancer. However, this is not enough (only one residue as constraint) for building a new dataset.

Way-1 to use the frequency of occurrence of center residue to define the center residue As the $way - 1$ consider each sub-group (MOTIF) separately; the frequency of the center (and the average distance of residue from center (ADFC)) is checked based on the number of residues presented in the groups separately. The Table. 1.3 shows the frequency of occurrence center residues along with their groups contain 3, 4, and 5 amino acids with center residue. (And the Tables in *Appendix* A presented the groups' average distance of residue from center).

From the Table.1.3 the statistics does not show significant among the occurrence of center residues, thus consider the structure presented in PDB like Pocket or dump must be checked. But from the Table. 1.3 that number of substructures are not enough to move forward in this direction.

Way-2

The hypothetical examples for updating the $way - 2$ presented in pseudo code in Fig. 1.3.

E.g.:

Table 1.3: Frequency of overall occurrence center residues along with their groups

Residues	3 groups				4 groups				5 groups			
	Soft thresh		All $C\alpha$		Soft thresh		All $C\alpha$		Soft thresh		All $C\alpha$	
	ONGO	TSG	ONGO	TSG	ONGO	TSG	ONGO	TSG	ONGO	TSG	ONGO	TSG
R	152	132	168	128	115	153	127	145	74	120	75	121
K	101	90	100	92	82	59	78	57	35	32	34	34
D	111	99	143	89	54	120	59	107	31	79	40	92
E	65	74	76	76	80	56	70	54	19	37	33	37
Q	37	39	34	33	44	42	35	44	38	34	28	26
N	63	44	66	46	63	42	62	40	47	26	48	22
H	51	109	70	108	128	158	140	163	23	44	19	46
S	42	54	32	53	55	60	50	57	47	63	46	59
T	33	54	36	50	33	45	29	44	27	20	26	24
Y	62	61	55	60	53	56	51	61	32	32	40	42
C	37	16	11	11	423	166	448	171	11	6	12	3
M	9	20	8	21	25	10	25	8	15	10	17	8
W	25	49	23	45	22	32	23	23	15	31	16	29
A	21	42	17	40	17	16	28	14	18	20	19	19
I	8	21	7	20	16	24	15	20	14	13	15	17
L	32	41	24	41	46	36	42	33	27	33	31	29
F	14	50	12	55	60	42	61	44	24	35	25	38
V	16	19	16	20	14	22	14	20	17	22	7	22
P	30	24	26	22	14	28	8	28	14	17	13	22
G	24	29	25	27	23	41	19	38	22	18	20	24

✓ 4-subgroup contain center is "C" other residues as D, E, Q then center residue "C" is array is chosen and DE, DQ, EQ are given hits

✓ 5-subgroup contain center is "C" other residues as D, E, Q, H then center residue "C" is array is chosen again and DE, DQ, EQ, DH, EH, QH are given hits.

Figure 1.3 PseudoCode for $Way - 2$

```
1: Center residue is popped from the selected MOTIF group then rest of the residues mapped using hash table presented in Table. 1.2b.
2: Mapped residue indexes are sorted from lowest to highest depends on the mapping.
3: Sorted group of step 3 is the selected list.
4: Then take the lowest (from selected list) and group them with rest of the selected list.
5: Then remove the lowest selected in step 5; assign the remaining group as selected list and repeat step 4 and step 5, until 1 residue exist in the selected list.
6: Using the groups of indexes to update the center reside occurrence table. E.g.: Like if one of the groups has [0,5] then occurrence table $[0, 5]^{th}$ place is increased by 1.
```

The overall hits of the center atom groups are shown in Table. 1.4. From the Table. 1.4 statistics, if the occurrence of the subgroup combination is more than factor of 10 in the overall occurrence as mentioned in Table. 1.4. Then the group is chosen for analysis. If the group is chosen either in ONGO or TSG, then the group is retried from the other class. In short words, from the results of $way - 2$;

only the higher occurrence group is chosen.

Table 1.4: Overall hits of center residue occurrence of $Way-2$

Center residue	Soft thresh		All $C\alpha$	
	ONGO	TSG	ONGO	TSG
R	941	1311	999	1289
K	557	459	538	467
D	459	933	560	962
E	419	464	484	460
Q	397	369	307	321
N	534	326	540	298
H	573	847	604	873
S	489	612	458	578
T	294	309	279	326
Y	413	421	448	495
C	1372	550	1427	542
M	174	110	185	93
W	181	331	188	288
A	180	210	215	196
I	140	171	142	182
L	332	347	336	314
F	338	386	345	415
V	160	217	100	212
P	156	210	128	238
G	225	260	202	285
U	0	0	0	0

Just considering $way-2$'s ONGO and TSG of soft threshold to calculate the conditional probability of the groups (which have higher than 100 hits) shown in Table. 1.5.

Table 1.5: Selected subgroups of $way-2$

Center residue	other residues in the group		$soft$ threshold		All $C\alpha$	
			ONGO	TSG	ONGO	TSG
A	S	G	-	-	26	0
C	H	C	286	116	306	126
C	C	C	994	328	1024	332
D	D	D	47	13	81	11
D	D	Y	9	134	10	109
F	D	E	-	-	73	4
F	D	D	36	10	39	10
H	H	C	54	168	54	170
H	C	C	162	132	186	131
M	H	H	-	-	2	10
N	N	T	70	1	70	1
P	R	R	17	5	17	2
R	R	Y	-	-	111	23
V	R	R	2	23	1	22
V	R	A	2	37	2	35

✓ P(ONGO | CR = “C”, Occ = “HC”) = 286/ (286+116) =0.711

✓ P(ONGO | CR = “C”, Occ = “CC”) = 994/ (994+328) =0.752

✓ P(TSG | CR = “D”, Occ = “DY”) = 134/ (134+9) = 0.94

✓ P(TSG | CR = “H”, Occ = “HC”) = 168/ (168+54) = 0.76

✓ P(ONGO | CR = “H”, Occ = “CC”) = 162/ (162+132) =0.551

From this result, it can be concluded the group of CR= “D”, and residue “D” and “Y” occurred together that may higher probable to TSG class.

This $way - 2$ statistics results of center residue “C” is also given more relative positivity with the finding of the recent paper [1] published on 2019.

Bibliography

[1] Joseph A. Combs and Gina M. DeNicola. The non-essential amino acid cysteine becomes essential for tumor proliferation and survival. 11(5):678.

[2] Anandanadarajah Nishanth, Chee Hung Chu, and Rasiah Loganantharaj. An integrated deep learning and dynamic programming method for predicting tumor suppressor genes, oncogenes, and fusion from pdb structures. Under review with Computers in Biology and Medicine; submitted on 17-Oct-2020 and manuscript number is CIBM-D-20-02914.

Appendix A

Level 3 statistics of Patterns find the Possible-SRS-MOTIFs

A.1 $Way - 1$ Results

Table A.1: Frequency/ hits of subgroup contain 1-residue

	Soft thresh		All_C_alpha	
	ONGO	TSG	ONGO	TSG
R	77	109	77	106
K	41	33	38	30
D	92	42	78	38
E	24	32	23	31
Q	19	15	18	14
N	111	36	85	39
H	57	62	39	42
S	193	51	184	47
T	42	17	23	16
Y	33	12	28	11
C	9	9	5	9
M	9	3	9	2
W	4	8	2	7
A	2	4	3	4
I	8	8	7	8
L	9	16	9	14
F	4	7	4	7
V	7	6	5	5
P	0	11	0	10
G	11	21	11	19
U	0	0	0	0

Table. A.1 present the single residue occurrence hits with classes. Table.A.2 presents the Average Distance From center residue (ADFC) along with their groups. This ADFC is used as threshold condition to find out the group residues, with the chosen center residue (as constraint).

Fig. A.1 shows the array representation of $Way-1$'s SOFT threshold condition with Four residues

Table A.2: Average Distance From center residue (ADFC) along with their groups

Residues	3 groups				4 groups				5 groups			
	Soft thresh		All $C\alpha$		Soft thresh		All $C\alpha$		Soft thresh		All $C\alpha$	
	ONGO	TSG	ONGO	TSG	ONGO	TSG	ONGO	TSG	ONGO	TSG	ONGO	TSG
R	11.89	13.92	11.62	14.31	12.49	16.52	12.02	17.06	13.57	13.41	13.65	13.26
K	14.23	12.66	13.98	12.44	12.71	13.37	12.49	13.46	11.37	16.32	10.84	14.57
D	12.07	13.75	10.76	12.71	13.58	13.77	12.67	14.75	10.39	12.60	9.98	12.44
E	16.22	18.64	15.52	18.42	17.36	12.77	18.14	12.19	11.04	14.56	11.33	14.76
Q	12.05	15.53	12.10	14.92	11.29	12.46	10.40	12.23	12.01	14.17	12.18	15.35
N	9.93	10.12	9.76	9.76	8.39	7.95	8.47	7.49	9.93	12.08	9.00	12.62
H	12.64	9.47	11.13	9.42	9.65	11.39	9.38	10.81	13.82	21.99	13.09	20.94
S	6.93	11.89	6.40	12.18	9.18	16.15	9.46	16.63	7.32	13.38	7.61	13.60
T	8.29	8.90	8.03	9.56	9.78	13.01	10.13	13.19	8.41	8.50	8.27	8.10
Y	9.45	10.28	9.44	9.28	12.83	17.84	13.07	17.01	16.53	14.78	14.57	14.50
C	8.46	7.95	8.45	7.76	7.48	7.68	7.49	7.77	9.06	13.49	9.00	18.53
M	8.38	11.29	8.61	11.17	12.00	14.52	12.08	15.25	14.01	10.41	13.79	9.65
W	13.89	12.74	13.87	11.26	6.88	10.41	6.97	11.10	8.90	11.24	10.68	13.29
A	20.18	13.35	22.36	13.76	14.50	12.01	11.06	12.70	13.13	9.30	11.83	9.27
I	6.90	10.53	6.50	9.45	9.32	9.22	9.08	9.01	12.59	17.37	12.41	18.05
L	8.15	9.62	8.22	10.03	9.35	11.99	9.10	13.34	11.48	13.81	9.80	13.84
F	13.52	23.18	13.59	25.16	11.25	18.94	11.34	18.39	12.51	11.28	12.43	11.53
V	15.71	7.48	15.80	7.40	8.68	11.84	10.58	12.57	10.64	9.74	8.79	9.79
P	23.35	8.28	24.55	8.42	17.83	8.97	26.07	8.74	8.91	10.48	9.00	9.22
G	9.80	14.47	9.61	15.01	8.72	12.31	8.71	12.49	9.64	13.15	9.77	12.05

in groups. In this manner $Way-1$ just give weightage for the center atom.

For an **E.g.:**

✓ Group “C” as center residue with “R”, “R”, “Q”

✓ Group “C” as center residue with “Q”, “C”, “V”

✓ Group “C” as center residue with “Q”, “C”, “C”

In all these instances the row “C” In the first instance column “R” get two hits in the last instance column “C” get two hits all instances column “Q” gets a hit, thus totally *3 hits for “Q” column

Figure A.1: Matrix representation of $Way-1$’s SOFT threshold condition with Four residues in groups

	Center residue row																				
	U	G	P	V	F	L	I	A	W	M	C	Y	T	S	H	N	Q	E	D	K	R	
R	0	8	8	4	12	20	1	3	11	7	6	18	10	11	20	28	29	15	7	36	60	R
K	0	4	4	5	12	14	8	4	14	2	1	7	8	7	26	12	17	9	11	16	32	K
D	0	4	2	4	73	3	1	3	2	2	2	10	8	21	18	15	6	22	18	13	24	D
E	0	7	3	1	36	7	0	2	3	6	2	8	8	3	16	15	7	36	9	35	13	E
Q	0	2	3	1	6	8	8	3	1	14	*3	12	3	2	16	10	6	21	13	12	15	Q
N	0	2	2	3	11	10	0	3	3	1	1	10	5	7	12	17	13	11	8	14	26	N
H	0	7	1	5	3	7	3	3	3	12	151	12	5	1	40	9	8	9	16	19	25	H
S	0	5	2	1	5	10	2	6	2	4	0	15	13	12	4	19	1	9	14	14	28	S
T	0	0	1	1	3	6	2	7	0	3	0	11	7	30	4	31	7	5	31	8	17	T
Y	0	5	2	5	1	7	4	1	2	1	1	8	2	12	4	3	11	27	8	7	20	Y
C	0	0	0	0	0	1	1	0	0	0	1092	0	1	0	190	3	0	1	2	0	5	C
M	0	0	1	0	2	3	1	0	0	12	2	1	0	0	11	2	8	15	2	1	5	M
W	0	0	3	0	0	3	1	1	1	0	0	3	0	0	1	2	0	6	3	1	2	W
A	0	5	1	1	0	2	2	1	2	4	0	7	10	16	2	0	5	14	1	4	11	A
I	0	0	2	1	3	2	1	2	8	0	1	7	0	6	6	0	2	6	2	6	4	I
L	0	2	4	6	2	12	2	1	0	0	6	8	2	3	1	4	5	3	1	2	8	L
F	0	5	0	2	2	5	8	0	1	0	0	3	1	2	2	11	1	9	2	11	12	F
V	0	3	1	1	0	6	2	2	0	0	1	11	3	5	0	2	1	4	3	7	13	V
P	0	1	1	1	8	6	1	2	11	2	0	5	2	9	8	3	2	2	7	20	11	P
G	0	9	1	0	1	6	0	7	2	5	0	3	11	18	3	3	3	16	4	20	14	G
U	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	U

A.2 $Way-2$ Results

Using the overall statistics and triangle hits, subgroup is assigned as either ONGO or TSG. Then the subgroup is cannot be assigned for the other class for MOTIF selection. Here both subgroups such as,

$$group_1 : center``C''with``H'', ``C'',$$

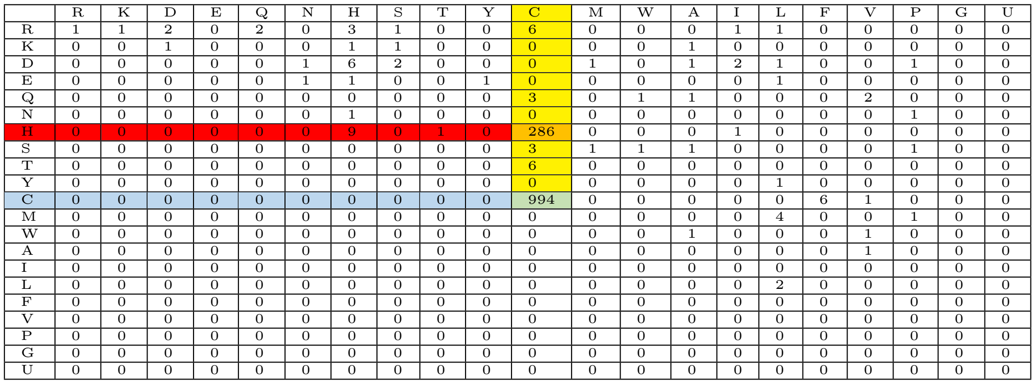
$$and$$

$$group_2 : center``C''with``C'', ``C''$$

are chosen for ONGO; because both hits are higher to ONGO class (from Fig. A.2 and Fig. A.3; 286 > 116 and 994 > 328).

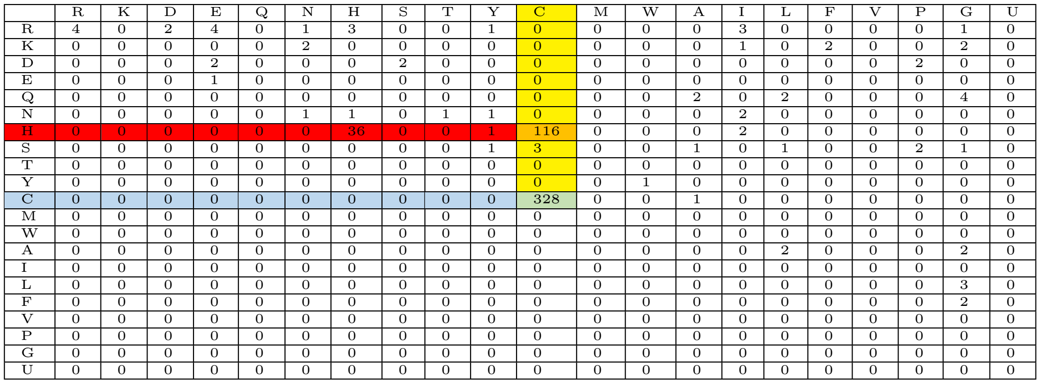
Even though TSGs' both groups ($group_1$ and $group_2$) hits satisfy the base conditions as explained in (paragraph 1.2.1 's Table. 1.4, TSG soft threshold's center residue C's overall occurrence 550, and the factor of 10 is 55), $group_1$'s hit 116 > 55 and, $group_2$ hit 328 > 55, these groups were not chosen to represent the TSG's SRS-MOTIF.

Figure A.2: ONGO soft threshold of center residue C's 21 × 21 $Way-2$'s triangle residue hit array.



	R	K	D	E	Q	N	H	S	T	Y	C	M	W	A	I	L	F	V	P	G	U
R	1	1	2	0	2	0	3	1	0	0	6	0	0	0	1	1	0	0	0	0	0
K	0	0	1	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0
D	0	0	0	0	0	1	6	2	0	0	0	1	0	1	2	1	0	0	1	0	0
E	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	3	0	1	1	0	0	0	2	0	0	0
N	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0
H	0	0	0	0	0	0	9	0	1	0	286	0	0	0	1	0	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0	3	1	1	1	0	0	0	0	1	0	0
T	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	994	0	0	0	0	0	6	1	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	1	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
U	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure A.3: TSG soft threshold of center residue C's 21 × 21 $Way-2$'s triangle residue hit array.



	R	K	D	E	Q	N	H	S	T	Y	C	M	W	A	I	L	F	V	P	G	U
R	4	0	2	4	0	1	3	0	0	1	0	0	0	0	3	0	0	0	0	1	0
K	0	0	0	0	0	2	0	0	0	0	0	0	0	0	1	0	2	0	0	2	0
D	0	0	0	2	0	0	0	2	0	0	0	0	0	0	0	0	0	0	2	0	0
E	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2	0	0	0	4	0
N	0	0	0	0	0	1	1	0	1	1	0	0	0	0	2	0	0	0	0	0	0
H	0	0	0	0	0	0	36	0	0	1	116	0	0	0	2	0	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	1	3	0	0	1	0	1	0	0	2	1	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	328	0	0	1	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	2	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
U	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0