Journal of Information Systems and Telecommunication
Vol.10, No.3, July-September 2022, 161-168
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

Membrane Cholesterol Prediction from Human Receptor using Rough Set based Mean-Shift Approach

Rudra Kalyan Nayak[1*], Ramamani Tripathy[2], Hitesh Mohapatra[3], Amiya Kumar Rath[4], Debahuti Mishra[5]

[1]. School of Computing Science and Engineering,VIT Bhopal University, Bhopal-Indore Highway, Kothrikalan, Sehore, MP, India
[2]. Department of Computer Science and Engineering, Chitkara University Himachal Pradesh Campus, Pinjore-Nalagarh National Highway, Dist-Baddi, Himachal Pradesh, India
[3]. School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar 751024, Odisha, India
[4]. Department of Computer Science and Engineering,Veer Surendra Sai University of Technology, Burla, Odisha 768018, India
[5]. Department of Computer Science and Engineering, Siksha 'O' Anusandhan (Deemed to be) University, Bhubaneswar, Odisha, India

Received: 15 Sep 2021/ Revised: 20 Dec 2021/ Accepted: 17 Jan 2022

Abstract

In human physiology, cholesterol plays an imperative part in membrane cells which regulates the function of G-protein-coupled receptors (GPCR) family. Cholesterol is an individual type of lipid structure and about 90 percent of cellular cholesterol is present at plasma membrane region. Cholesterol Recognition/interaction Amino acid Consensus (CRAC) sequence, generally referred as the CRAC (L/V)-X1−5-(Y)-X1−5-(K/R) and the new cholesterol-binding domain is similar to the CRAC sequence, but exhibits the inverse orientation along the polypeptide chain i.e. CARC (K/R)-X1−5-(Y/F)-X1−5-(L/V). GPCR is treated as a biggest super family in human physiology and probably more than 900 protein genes included in this family. Among all membrane proteins GPCR is responsible for novel drug discovery in all pharmaceuticals industry. In earlier researches the researchers did not find the required number of valid motifs in terms of helices and motif types so they were lacking clinical relevance. The research gap here is that they were not able to predict the motifs effectively which are belonging to multiple motif types. To find out better motif sequences from human GPCR, we explored a hybrid computational model consisting of hybridization of Rough Set with Mean-Shift algorithm. In this paper we made comparison among our resulted output with other techniques such as fuzzy C-means (FCM), FCM with spectral clustering and we concluded that our proposed method targeted well on CRAC region in comparison to CARC region which have higher biological relevance in medicine industry and drug discovery.

Keywords: GPCR; CRAC; CARC; ANN; Decision Tree; Rough Set; Mean Shift.

1- Introduction

In cell biology, cholesterol acts as a major component in cell membrane and has a modulatory role in integral membrane protein like GPCRs. Cholesterol is an individual type of lipid structure and about 90 percent of cellular cholesterol is present at plasma membrane region. GPCR is treated as a biggest super family in human physiology and probably more than 900 protein genes included in this family. As GPCR super family is responsible for novel drug discovery in pharmaceuticals area, so it is an emerging field for all researchers. Normally, GPCR family is arranged by lengthy protein sequences which include three basic regions like N-terminus known as external portion, C- terminus is known as internal portion and another middle segment is their which containing seven transmembrane domains shown in figure 1. A long protein sequence is the combination of amino acid which starting from extracellular part to intracellular region through the cell membrane surface. Once a GPCR binds a ligand in the meantime ligand triggers a conformational varies in the 7-TM region of the receptor. These things stimulate the C-terminus, which subsequently recruits a substance that in order activates the G protein linked with the GPCR.

✉ **Rudra Kalyan Nayak**
rudrakalyannayak@gmail.com

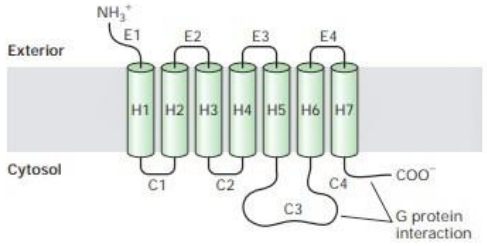

Fig. 1 Structure of GPCR Receptor

GPCRs contain seven helices such as H1-H7. Each helix contains individual protein chain which is the combination of all amino acids. Generally 20 amino acids are named as A, R,N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V have used for protein strain construction which is shown in Table 1. GPCRs family is a superfamily comprising of various subfamilies, including Class A rhodopsin-like, Class B secretin-like, Class C metabotropic glutamate/pheromone, Class F frizzled (FZD), Taste receptors, Vomeronasal receptors and 7TM orphan receptors. They are categorized into 7 subfamily based on the character of stimuli that stimulates the GPCRs and sequence similarity. Cholesterol is a 27 carbon compound with a distinctive structure with a hydrocarbon tail, a central sterol nucleus made of four hydrocarbon rings, and a hydroxyl group. The center sterol nucleus or ring is a feature of all steroid hormones. The hydrocarbon tail and the central ring are non-polar and therefore do not mix with water. Therefore cholesterol (lipid) is packaged together with apoproteins (protein) in order to be carried through the blood circulation as a lipoprotein. Lipid satisfies numerous biological functions and its presence is very essential for successful cellular homeostasis.

Table 1: List of Amino acid

Amino Acid	ABBREVIATION 3-Letter	1-Letter	Amino Acid	ABBREVIATION 3-Letter	1-Letter
Alanine	Ala	A	Leucine	Leu	L
Arginine	Arg	R	Lysine	Lys	K
Asparagine	Asn	N	Methionine	Met	M
Aspartic acid	Asp	D	Phenylalanine	Phe	F
Cysteine	Cys	C	Proline	Pro	P
Glutamic acid	Glu	E	Serine	Ser	S
Glutamine	Gln	Q	Threonine	Thr	T
Glycine	Gly	G	Tryptophan	Trp	W
Histidine	His	H	Tyrosine	Tyr	Y
Isoleucine	Ile	I	Valine	Val	V

Cholesterol is known as organic molecules and is biosynthesized by every animal cells and also very much essential structural component of animal cell membrane. Cholesterol plays a pivotal function in vitamin D production, bile secretion and hormone production. Cell membrane cholesterol acts as a significant role in modulating the function of numerous membrane proteins and from these proteins a special cholesterol binding motif is reported to which the membrane cholesterol binds and modulates their movement. This consensus motif is either seen as CRAC or CARC, [1-10] which correspond to the mixture of amino acid and its structure is shown in figure 2.

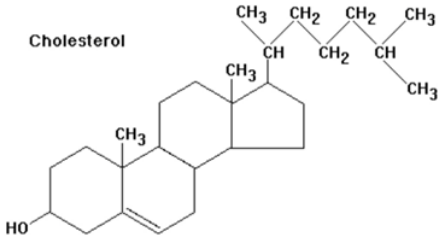

Fig. 2 Structure of Cholesterol

This work we highlights on prediction of membrane cholesterol from human integral protein such as GPCR family. Due to much more involvement of GPCR in cell biology it is very much important to identify novel signature motifs which have biological significance in the entire medical area. In modern times, many relevance areas in biomedical science and bioinformatics are introduced where most advanced soft computing techniques were applied successively. Main intention is to boost the correctness of clustering algorithms with least number of iterations. So a lot of scientists have explored diversity of classifiers algorithm like artificial neural network (ANN), logistic regression, fuzzy C-means and decision tree etc. to forecast the cholesterol signature motifs [11-19].

To investigate the signature motif of cholesterol using the dataset is our focal objectives. The research gap in earlier researches is that they were not able to predict the motifs which were belonging to multiple motif types. This dilemma can be removed by introducing rough set method which classifies the motif types to one or more regions based on their belongingness. Subsequently mean shift clustering algorithm gets the valid clusters which exhibits our proposed rough set based mean shift algorithm as prominent. Therefore, our present work is concentrating to uncover valid cholesterol motif signature from large helical protein sequences of human GPCR.

Gao, Q et al. [20] presented an ensemble method for G-protein coupled receptors classification in four stages to discuss about GPCRs and non-GPCRs. Gu, Q et al. [21] introduced a prediction model of Adaboost using a data base of low homology based on pseudo amino acid composition with close entropy and property of

hydrophobic forms to guess GPCR classes and this gave best consequence of the whole experiment. Furthermore Peng, Z et al. [22] gave a broad view of an improved classification model for prediction of GPCRs based on different characteristics.

The draft of the paper is as follows: in section 2, materials and methods, the results and discussion is instanced in section 3 plus in last section conclusion part is portrayed.

2- Research Method

All helical files namely H1 to H7 of GPCR family contains totals of 900 protein sequences. Generally, Protein dataset is the combination of unlike amino acid sequences retrieved sequentially from uniprot database using text (.txt) format [23]. The primary aim of our present work is to obtain and discover the valid cholesterol signature motif using different window size using the formula of CRAC/CARC. CRAC is a tiny linear motif which carries out a very simple algorithm, which is N-terminus to C-terminus direction and the formula is used here is (L/V)-X1−5-(Y)-X1−5-(K/R). Similarly CARC is the opposite orientation of CRAC which is formulated by (K/R)-X1−5-(Y/F)-X1−5-(L/V).

Table 2 shows that L is length of cholesterol motif. The length of window size range varies from minimum five to maximum thirteen for both forward and backward pattern recognition methods. A cholesterol dictionary d is constructed according to their motif type and window size. For example, considering $d9$ having length $L = 9$ can have motif types MT={15, 24, 34, 42, 51} and cholesterol motif signatures can be in the form of Leu/Val-X1-Y-X5-Lys/Arg, Leu/Val-X2-Y-X4-Lys/Arg, Leu/Val-X3-Y-X4-Lys/Arg, Leu/Val-X4-Y-X2-Lys/Arg and Leu/Val-X5-Y-X1-Lys/Arg, Lys/Arg-X1-Y/F-X5-Leu/Val, Lys/Arg-X2-Y/F-X4-Leu/Val, Lys/Arg-X3-Y/F-X4-Leu/Val, Lys/Arg-X4-Y/F- X2-Leu/Val and Lys/Arg-X5-Y/F-X1-Leu/Val for both CRAC and CARC algorithms respectively. And also X position can be varying from any residue among 20 amino acids. From above cholesterol formula, it is revealed that motif length remains stable but membrane cholesterol motif sequences fluctuate depending upon the 'Y and F' positions.

The probable combination of cholesterol sequence found in Table 3 and Table 4 for different motif lengths is show below. Whole numbers of uncovered cholesterol subsequence for CRAC (forward) and CARC (backward) after mapping are 2003 and 4013 respectively. From the result, it is perceptible that the combination backward sequence has more targets over forward. The combinations of N-terminus to C-terminus and vice versa are calculated for both CRAC and CARC. The motif sequence combination of CARC are calculated using{Arg-X(1-5)-Y/F-X(1-5)-Leu=1799, Lys-X(1-5)-Y/F-X(1-5)-Leu = 978, Arg-X(1-5)-Y/F-X(1-5)-Val = 779, Lys-X(1-5)-Y/F-X(1-5)-Val= 457} are to be found for unlike motif types. Likewise, for forward (CRAC) motif sequence combinations are {Leu-X(1-5)-Y-X(1-5)-Arg = 764, Leu-X(1-5)-Y-X(1-5)-Lys= 527, Val-X(1-5)-Y-X(1-5)-Arg =330, Val-X(1-5)-Y-X(1-5)-Lys = 382} are to be found.

Table 2. Depiction of all possible motif types with different combination using 20 amino acids

CHOLESTEROL DICTIONARY FORMULA	MOTIF TYPE/ CHOLESTEROL MOTIF LENGTH	SIGNATURE MOTIF SEQUENCE
BACKWARD (CARC) (R/K-$X_{(1-5)}$-Y/F-$X_{(1-5)}$-L/V)	11, ...15, =5,6,7,8,9 21,... 25 = 6,7,8,9,10 31,... 35 = 7,8,9,10,11 41.... 45 = 8,9,10,11,12 51,... 55 = 9,10,11,12,13	K/R-X1-Y/F-X1-L/V, ...K/R-X1-Y/F-X5-L/V K/R-X2-Y/F-X1-L/V, ...K/R-X2-Y/F-X5-L/V K/R-X3-Y/F-X1-L/V, ...K/R-X3-Y/F-X5-L/V K/R-X4-Y/F-X1-L/V, ...K/R-X4-Y/F-X5-L/V K/R-X5-Y/F-X1-L/V... .....K/R-X5-Y/F-X5-L/V
FORWARD (CRAC) (L/V-$X_{(1-5)}$-Y-$X_{(1-5)}$-R/K)	11, ...15, =5,6,7,8,9 21,... 25 = 6,7,8,9,10 31,... 35 = 7,8,9,10,11 41.... 45 = 8,9,10,11,12 51,... 55 = 9,10,11,12,13	L/V-X1-Y-X1-K/R.....L/V-X1-Y-X5-K/R L/V-X2-Y-X1-K/R.....L/V-X2-Y-X5-K/R L/V-X3-Y-X1-K/R.....L/V-X3-Y-X5-K/R L/V-X4-Y-X1-K/R.....L/V-X4-Y-X5-K/R L/V-X5-Y-X1-K/R.....L/V-X5-Y-X5-K/R

Table 3. Dissimilar Motif Types (MT) detected in GPCRs for Backward (CARC) cholesterol sequences

Motif-Length	Arg-$_{(1-5)}$-Y/F-$X_{(1-5)}$-Leu	Lys-$X_{(1-5)}$-Y/F-$X_{(1-5)}$-Leu	Arg-$X_{(1-5)}$-Y/F-$X_{(1-5)}$-Val	Lys-$X_{(1-5)}$-Y/F-$X_{(1-5)}$-Val	Total
5	49	25	35	23	**132**
6	249	56	177	27	**509**
7	79	46	94	45	**264**
8	377	219	100	95	**791**
9	630	112	146	89	**977**

10	191	70	91	63	**415**
11	125	297	56	50	**528**
12	65	122	62	54	**303**
13	34	31	18	11	**94**
Total	**1799**	**978**	**779**	**457**	**4013**

Table 4.Dissimilar Motif Types (MT) detected in GPCRs for forward (CRAC) cholesterol sequences

Motif-Length	**Leu- $X_{(1-5)}$-Y- $X_{(1-5)}$-Arg**	**Leu- $X_{(1-5)}$-Y- $X_{(1-5)}$-Lys**	**Val- $X_{(1-5)}$-Y- $X_{(1-5)}$-Arg**	**Val- $X_{(1-5)}$-Y- $X_{(1-5)}$-Lys**	**Total**
5	29	21	19	11	**71**
6	51	35	38	26	**150**
7	80	52	33	22	**187**
8	107	57	48	28	**240**
9	126	87	70	112	**395**
10	160	107	56	98	**421**
11	70	63	26	29	**188**
12	38	29	17	27	**111**
13	103	76	23	29	**231**
Total	**764**	**527**	**330**	**382**	**2003**

Figure 3 which is given below depicts about the architecture of proposed model which stores all helical files in text format and also stores cholesterol dictionary whose data are retrieved utilizing the technique known as sliding window which has considered the length of motif and it is denoted as $L = \{5, 6, 7, 8, 9, 10, 11, 12, 13\}$. Dictionary of cholesterol can be calculated using motif length/ number of sequence.

Our aim is to investigate the signature motif of cholesterol using above dataset. In the first stage of our proposed model entire helical data are retrieved sequentially and cholesterol data are retrieved according to their motif type from dictionary. Then we are mapping our two datasets. Once mapping is over we go for next step for calculation of CRAC and CARC motif. Then we apply our hybrid technique Rough set with Mean-Shift algorithm for computing all cluster centers. After finding the cluster centers we can sort, merge the motifs using weight value. We have a cut-off on calculation of motif according to their weight. Finally we reconstruct the data points and obtained our motifs which have biological relevance.

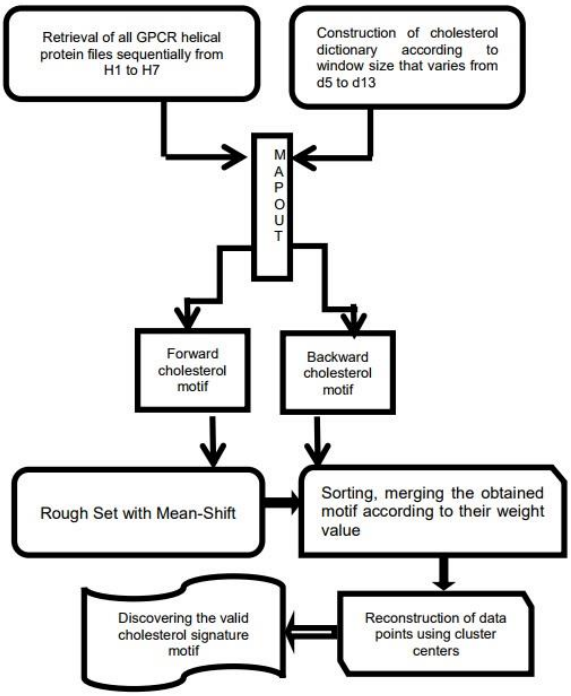

Fig.3Architecture of proposed Rough Set with Mean-Shift cholesterol model for identification of valid motif signature

2-1- Rough Set Theory (RS)

Rough set theory is a technique for dealing with imperfect knowledge, in particular with vague concepts. Rough set theory has gained interest of many researchers and practitioners from all over the world. This theory has been implemented in many areas like Bioinformatics, acoustics, business and finance, chemistry, computer engineering and electrical engineering decision analysis and systems, economics, digital image processing, informatics, medicine, molecular biology, neurology, robotics, social science, software engineering, spatial visualization, Web engineering, and Web mining. That's why we are motivated to implement this theory within our paper for prediction purpose. Generally, biological data are very

complex and sensitive. To handle this complex data set and to discover structural relationship within expected imprecise and noisy data, rough set approach is used here.
Our both data such as cholesterol and GPCR protein were taken for prediction purpose. After mapping is over we have categorized the forward motif and backward motif. We have proposed a hybrid approach Rough set with Mean shift for prediction of valid motif sequences. Basically this procedure clearly dealt with finding hidden pattern motif sequences and evaluation of significance of motifs without any overlapping.
RS theory is characterized as a system S = < U , V, R, F>by Z. Pawlak, with R = D∪C, [24] at which U be a non-void bounded collection of items and R be a non-void bounded collection of properties, D and C be the subsets known as decision and condition feature collection. Where V = $\bigcup_{a \in R} Va$, Vais represented as a gathering of feature values of a, also cardinality of (Va)>1 and f: R → V is a fact or characterization mapping. Indiscernible relation [24]: For a known subclass of feature collection B⊆R, an imperceptible relationship imp(B) in the space of discussion U which could be characterized in equation (1) as below,

$$imp(B) = \{(m,n) \mid (m,n) \in U^2, \forall_{b \in B}(b(m) = b(n))\} \quad (1)$$

So similarity relationship here is nothing but an imperceptible relationship. [n]imp(B) or [n]B and [n] refers equality family unit of an piece. And then (U, [n]imp(B)) pair off is addressed as an guesstimate space. Lower and Upper approximation sets [25]: For a given system S = <U, V, R, F>,Considering Y⊆ U where Y is a subset, upper and lower bound sets respectively be determined in equation (2) plus (3) by

$$\overline{appr}(Y) = \{n \in U \mid [n] \cap Y \neq \phi\}, \quad (2)$$

$$\underline{appr}(Y) = \{n \in U \mid [n] \subseteq Y\}, \quad (3)$$

Where [n] refers equality family of n.
So collection of every equality families is received as the ratio collection of U, and U/R = {[n] | n∈U} refers it. Here three disjoint parts of space are like the negative, positive plus boundary approximation sets are given in equations (4-6) below [24-28].

$$P(Y) = \underline{appr}(Y), \quad (4)$$

$$B(Y) = \overline{appr}(Y) - \underline{appr}(Y), \quad (5)$$

$$N(Y) = U - \overline{appr}(Y), \quad (6)$$

When object n∈ P(Y), then, it is counted on target set Y positively. When object n∈ B(Y), then, it would not be counted as goal set Y positively. When n∈ N(Y), so it is not easy to decide whether n would be a part of goal set Y.

2-2- Mean-Shift Approach

In human pathogen every amino acid sequence has some biological significance. To compute valid cholesterol motif from human GPCR we implemented Mean-Shift algorithm which cluster each data points employing window across it and calculates the average of the data point. Then it shifts the center of the window to the average and reiterates the algorithm till it converges. After each iteration, the window shifts to a denser region of the dataset. In case of time and space complexity Rough Set with Mean-Shift does well on GPCR data [29-34]. Now the steps of Rough Set with Mean-shift algorithm for a collection of different amino acid sequences S are given below:

Rough Set with Mean-Shift Algorithm:

Step-1: Construct cholesterol dictionary according to window size that varies from $d_5$to d_{13}.

Step-2: Apply Rough Set based method over the targeted motif sequences from constructed dataset consisting of sequence of amino acids.

Step-3: For each amino acid sequence $s \in S$, discover the neighboring points *N(s)* of *s*.

Step-4: For each amino acid sequence $s \in S$, calculate the *mean shift m(s)* from the equation (7):

$$m(s) = \frac{\sum_{s_i \in N(s)} K(s_i - s) s_i}{\sum_{s_i \in N(s)} K(s_i - s) s_i} \quad (7)$$

Step-5: For each amino acid sequence $s \in S$, update $s \leftarrow m(s)$.

Step-6: Iterate Step-1 for *n times* or until *m(s)* converges.

Here N(s) represents the function to evaluate the neighbors of a data sequence s ∈ S. The distances of neighboring points are calculated by the Euclidean distance metric [35]. Again K(d) represents the kernel used in Mean-Shift, where K denotes a Gaussian Kernel [36-37] and d denotes the distance between the two data sequences. The algorithm runs with time complexity $O(KN^2)$. Clustering results are depicted in Figure 4 where prediction of both membrane cholesterol and GPCRs are done.

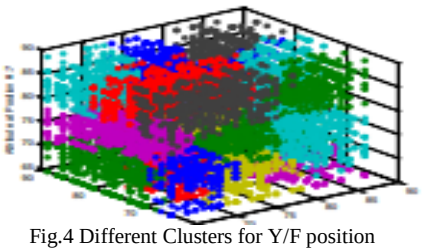

Fig.4 Different Clusters for Y/F position

3- Results and Discussion

All the transmembrane proteins data were collected from uniprot database and cholesterol dictionary is constructed using CARC /CRAC algorithm. According to helix number protein sequences are maintained. Each helix includes ~820 proteins. The most important objective of this work is to identify and fetch the entire cholesterol consensus motif available in the protein primary sequences and develop a signature motif for receptors belonging to GPCR superfamily. Dissimilar motif types are considered using parameters. For example motif sequence for forward region is LAAMAYDR. It means first position is L/V and last position is combination of K/R and Y position is fixed. Here motif type is 41, i.e. L4Y1R and the window size is 8. Using CRAC algorithm we have calculated the motif types.

The main intention of this paper is to uncover the most important motif signatures utilizing both forward plus backward formula from human GPCR. After mapping a large volume of data set, we are looking for an enhanced data mining method for obtaining effective signature motif of cholesterol. Through one example we explain this as; a given motif: K/RXY/FXL/V of length 9 can be in more than one motif types like {15, 24, 34, 42, 51} given in Table 1. We found Rough Set technique [24-28] is better to handle the data belonging to more than one class and Mean-Shift technique [29-34] as an efficient one to find the cluster centers. Hence, to mine this category of information wherever data can fit in to more than one cluster, we employed Rough set with Mean-Shift algorithm in our projected model.

Table 5 expressed about the signature motifs of cholesterol for GPCRs using the forward (CRAC) and backward formula CARC. In these tables we have elaborated helix number, motif types plus its valid signature. Among all helices like H1-H7 only targeted helix for forward region is h3, h5, h6 and h7 and similarly for backward region the targeted helices are h2, h7. Each helix target includes both motif types with corresponding valid signature.

Table 5. Signature motifs of cholesterol for GPCRs using the forward formula (CRAC) and backward formula (CARC)

CRAC			CARC		
HELIX	**CRAC MOTIF TYPE**	**SIGNATURE**	**HELIX**	**CARC MOTIF TYPE**	**SIGNATURE**
H3	L4Y1R	LAAMAYDR	H2	R2F2L	RPMFFLL
H3	L5Y1R	LLAVMAYDR	H2	R5F1L	RLHTPMFFL
H3	L3Y2K	LSIFYCLK	H2	R5Y2L	RLHTPMYFFL
H3	V3Y1R	VLMAYDR	H2	R4Y4L	RTVTNYFILNL
H5	L4Y4R	LNPFIYSLRNR	H2	R5Y5L	RLHTPMYFFLSNL
H5	L5Y4K	LLNPIIYSLRMK	H2	R4Y2V	RTVTNYFIV
H5	V4Y4K	VNPLVYTLRMK	H2	K2Y2L	KPMYIFL
H6	L2Y1R	LMSYDR	H2	K3Y1L	KAMYYFL
H6	L3Y1R	LVFMYLR	H2	K5Y1L	KTASVFYTL
H6	V5Y1R	VPCIYAYLR	H2	K2Y5L	KPMYFFLSML
H7	L1Y4R	LSYTRINR	H2	K5Y2L	KLHTPMYFFL
H7	L4Y2R	LNPLIYSL	H2	K5F3L	KTATNIFLLNL
H7	L5Y2R	LLNPFIYSLR	H2	K5Y5L	KLLTPMYFFLTPL
H7	L4Y4R	LNPLIYTLRNR	H7	K5Y1V	KVASVFYTV
H7	L3Y2K	LSIFYLLK	H7	K5Y2V	KVASVFYTVV
H7	L4Y2K	LNPLIYSLK	H7	K5Y1L	KLLTVIYSL
H7	L4Y4K	LNPLIYSLRNK	H7	K5Y2L	KLHTPMYTFL
H7	L4Y5K	LNPILYFLRNEK	H7	K5Y5L	KLLTLFYFFLTPL
H7	L5Y4K	LLNPFIYTLRNK			
H7	V4Y2R	VNPLIYSLR			
H7	V5Y2R	VLNPLIYSLR			
H7	V1Y4K	VIYTLRNK			
H7	V4Y4K	VNPLVYSLRNK			

In our present work Rough Set deals data with uncertainty as very well and Mean-Shift being a non-parametric clustering technique with the strengths of capable of handling arbitrary feature spaces with no pre-specified number of clusters analyzes the real GPCR data in a fair manner.

3-1- Comparison among the Methods

Table 6 shows the contrast among all methods with respect to helix name and motif type for both forward

with backward region. Helix name means it represented the targeted helix name of GPCR. Each time, membrane cholesterol is bound with N-C terminus region of membrane proteins and also with their corresponding helix. Another important part is the motif type which is denoted as forward and backward position. With the help of algorithm CRAC and CARC we choose motif type. If motif type is written as 55 means for forward position the formula as: L/V-X5-Y-X5-K/R. Here X5 is any combination of five amino acid which is residing in between L/V and Y and in next part X5 is represented as any five amino acid combination reside within Y and K/R. Likewise it is represented for 11,12,--15, 21---25, 31…35, 41…45, 51…55 etc. From comparison table (Table 6) we conclude that our proposed model Rough with mean shift target on higher priority motif types such as 55, 52, 53, 51, 31, and 32 in comparison with other existing methods [11,12].

Table 6. Motif type comparison by different methods

METHODS	HELIX NAME	MOTIF TYPE (FORWARD/ BACKWARD)
FCM [11]	h2,h5,h7	11,12,21,54,34
FCM with Spectral [12]	h6,h3,,h7,h5	44,42,32,22.21
Rough Set with Mean Shift (Proposed)	h3,h5, h6,h2,h7	55, 52,53,51, 31,22

4- Conclusion and Future Scope

G-Protein-Coupled-Receptor is one of the compelling fields which is mostly involved for cell signaling and more frequently targeted by the entire pharmaceutical domain. In cell membrane, among all integral membrane protein, GPCR is treated as an important super family. Each time, membrane cholesterol targets with this GPCR family to find out the best motif sequences which has biological relevance. Our aim is to investigate the signature motif of cholesterol using above dataset. Due to high dimensionality of the data, this paper projected a hybrid model Rough Set and Mean-Shift based method for cholesterol prediction from human GPCR. Our proposed Rough Set with Mean-Shift based model yielded satisfactory result as discussed in experimental section. The best motifs we found can have reliable clinical treatment and can also be used in drug discovery for diseases. Based on the weight value for each motif type we calculated sub-motifs from huge amount of protein which gave better results. In our analysis we conclude that most of the target sites are included in helix 2 and 7 in addition of motif types 55, 53, 52, 31, 32 etc. which have greater biological relevance. Further study can be extended considering other disease databases which can be used for membrane cholesterol prediction with higher biological relevance and could be helpful for drug designers.

References

[1] D. M. Rosenbaum, S. G. Rasmussen, and B. K. Kobilka, "The structure and function of G-protein-coupled receptors," Nature, vol. 459, no. 7245, pp. 356-363, May 2009.

[2] C. Ellis, "The state of GPCR research in 2004," Nature Reviews Drug Discovery, vol. 3, no. 7, pp. 577-626, 2004.

[3] C. J. Baier, J. Fantini, and F. J. Barrantes, "Disclosure of cholesterol recognition motifs in transmembrane domains of the human nicotinic acetylcholine receptor," Scientific reports, vol. 1, no. 1, pp. 1-7, 2011.

[4] X. Sun and G. R. Whittaker, "Role for influenza virus envelope cholesterol in virus entry and infection," Journal of virology, vol. 77, no. 3, pp. 12543-12551, 2003.

[5] T. J. Pucadyil, A. Chattopadhyay, "Role of cholesterol in the function and organization of G-protein coupled receptors," Progress in lipid research, vol. 45, no. 4, pp. 295-333, 2006.

[6] S. Putluri, M. Z. Rahman, C. S. Amara, and N. Putluri, "New exon prediction techniques using adaptive signal processing algorithms for genomic analysis, IEEE Access, vol. no. 7, pp. 80800-80812, 2019.

[7] T. A. Masoodi, N. A. Shaik, S. Burhan, Q. Hasan, G. Shafi, and V. R. Talluri, "Structural prediction, whole exome sequencing and molecular dynamics simulation confirms p. G118D somatic mutation of PIK3CA as functionally important in breast cancer patients," Computational biology and chemistry, vol. 80, no. 2, pp. 472-479, 2019.

[8] A. Ahilan, G. Manogaran, C. Raja, S. Kadry, S. N. Kumar, C. A. Kumar, T. Jarin, S. Krishnamoorthy, P. M. Kumar, G. C. Babu, and N. S. Murugan, "Segmentation by fractional order darwinian particle swarm optimization based multilevel thresholding and improved lossless prediction based compression algorithm for medical images," IEEE Access, vol. 7, pp. 89570-89580, 2019.

[9] N. Jayanthi, B. V. Babu, and N. S. Rao, "Survey on clinical prediction models for diabetes prediction," Journal of Big Data, vol. 4, no. 1, pp. 1-5, 2017.

[10] M. Anila, and G. Pradeepini, "Study of prediction algorithms for selecting appropriate classifier in machine learning," Journal of Advanced Research in Dynamical and Control Systems, vol. 9, pp. 257-268, 2017.

[11] R. Tripathy, D. Mishra, and V. B. Konkimalla, "A novel fuzzy C-means approach for uncovering cholesterol consensus motif from human G-protein coupled receptors (GPCR)," Karbala International Journal of Modern Science, vol. 1, no. 4, pp. 212-224, 2015.

[12] R. Tripathy, D. Mishra, V. B. Konkimalla, and R. K. Nayak, "A computational approach for mining cholesterol and their potential target against GPCR seven helices based on spectral clustering and fuzzy c-means algorithms," Journal of Intelligent & Fuzzy Systems, vol. 35, no. 1, pp. 305-314, 2018.

[13] R. M. Epand, A. Thomas, R. Brasseur, and R. F. Epand, "Cholesterol interaction with proteins that partition into membrane domains: an overview," Cholesterol Binding and Cholesterol Transport Proteins, pp. 253-278, 2010.

[14] R. M. Epand, "Cholesterol and the interaction of proteins with membrane domains. Progress in lipid research," vol. 45, no. 4, pp. 279-294, 2006.
[15] D. Gurram, and M. N. Rao, 'A comparative study of support vector machine and logistic regression for the diagnosis of thyroid dysfunction," International Journal of Engineering & Technology, vol. 7, no. 1.1, pp. 326-328, 2018.
[16] H. Jyothula, S. K. Rao, and V. Vallikumari, "Two phase active counter mechanism embedded with particle swarm optimization technique for segmentation of bio-medical images," Journal of Advanced Research in Dynamical and Control Systems, vol. 9, no. 6, pp. 232-242, 2017.
[17] S. Razia, M. R. Narasingarao, and P. Bojja, "Development and analysis of support vector machine techniques for early prediction of breast cancer and thyroid," Journal of Advanced Research in Dynamical and Control Systems, vol. 9, no. 6, pp. 869-878, 2017.
[18] P. Siva Kumar, V. Sarvani, P. Prudhvi Raj, K. Suma, and D. Nandu, "Prediction of heart disease using multiple regression analysis and support vector machines," Journal of Advanced Research in Dynamical and Control Systems, vol. 9, no. 18, pp. 675-682, 2017.
[19] N. Rajesh, T. Maneesha, S. Hafeez, and H. Krishna, "Prediction of heart disease using machine learning algorithms," International Journal of Engineering & Technology(UAE), vol. 7, no. 2.32, pp. 363-366, 2018.
[20] Q. B. Gao QB, Z. Z. Wang, "Classification of G-protein coupled receptors at four levels," Protein Engineering, Design and Selection, vol. 19, no. 11, pp. 511-516, 2006.
[21] Q. Gu, Y. S. Ding, and T. L. Zhang, "Prediction of G-protein-coupled receptor classes in low homology using Chou's pseudo amino acid composition with approximate entropy and hydrophobicity patterns," Protein and peptide letters, vol. 17, no. 5, pp. 559-567, 2010.
[22] M. Bhasin, and G. P. Raghava, "GPCRsclass: a web tool for the classification of amine type of G-protein-coupled receptors," Nucleic acids research, vol. 33, no. 2, pp. W143-147, 2005.
[23] http://www.uniprot.org
[24] Z. Pawlak, "Rough sets," International journal of computer & information sciences, vol. 11, no. 5, pp. 341-356, 1982.
[25] A. Skowron, J. Komorowski, Z. Pawlak, and L. Polkowski, "Rough sets perspective on data and knowledge," In Handbook of data mining and knowledge discovery, pp. 134-149, 2002.
[26] Z. Pawlak, and A. Skowron, "Rough membership functions," In Advances in the Dempster-Shafer theory of evidence, pp. 251-271, 1994.
[27] L. Polkowski, "Rough sets," Heidelberg: Physica-Verlag, 2002.
[28] L. Polkowski, and A. Skowron, "Rough mereological calculi of granules: A rough set approach to computation," Computational Intelligence, vol. 17, no. 3, pp. 472-492, 2001.
[29] M. A. Carreira-Perpinan, "Gaussian mean-shift is an EM algorithm," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 5, pp. 767-776, 2007.
[30] H. E. Cetingul, and R. Vidal, "Intrinsic mean shift for clustering on Stiefel and Grassmann manifolds" In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1896-1902, 2009, IEEE.
[31] C. C. Chang, and C. J. Lin, "IJCNN 2001 challenge: Generalization ability and text decoding," In IJCNN'01 International Joint Conference on Neural Networks, Proceedings (Cat. No. 01CH37222), vol. 2, pp. 1031-1036, 2001, IEEE.
[32] Y. Cheng, "Mean shift, mode seeking, and clustering," IEEE transactions on pattern analysis and machine intelligence, vol. 17, no. 8, pp. 790-799, 1995.
[33] R. T. Collins, "Mean-shift blob tracking through scale space," In 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Proceedings, vol. 2, pp. II-234, 2003, IEEE.
[34] D. Comaniciu, and P. Meer, "Mean shift: A robust approach toward feature space analysis," IEEE Transactions on pattern analysis and machine intelligence, vol. 24, no. 5, pp. 603-619, 2002.
[35] D. S. Wilks, "Statistical methods in the atmospheric sciences," Academic press, 2011.
[36] M. A. Carreira-Perpinán, "Fast nonparametric clustering with Gaussian blurring mean-shift" In Proceedings of the 23rd international conference on Machine learning, pp. 153-160, 2006.
[37] M. A. Carreira-Perpinan, "Acceleration strategies for Gaussian mean-shift image segmentation," In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 1, pp. 1160-1167, 2006, IEEE.