Indian Journal of Science and Technology, Vol 9(22), DOI: 10.17485/ijst/2016/v9i22/94237, June 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Hybridized Clustering Approach based on Rough Set and Fuzzy c-Means to Mine Cholesterol Sequence from ABC Family

Ramamani Tripathy[1], Debahuti Mishra[2*] and V. Badireenath Konkimalla[3]

[1]Department of Computer Science and Engineering, Siksha O Anusandhan University, Bhubaneswar, Odisha, India; ramatripathy1978@gmail.com
[2]National Institute of Science Education and Research (NISER), Jatni - 752050, Odisha, India; mishradebahuti@gmail.com
[3]Department of Atomic Energy, Bhubaneswar, Odisha, India; badireenath@niser.ac.in

Abstract

Objectives: The current study is focused on design of a computational model for human ABC transporters; wherein the TM-sequences matching the CRAC/CARC motif are extracted. **Methods:** The postulation of cholesterol binding motif (CRAC/CARC), its presence in different proteins and validating its interaction with cholesterol has indeed established the importance of the motif in cholesterol-mediated modulation of protein/signaling pathway. Several viral proteins and membrane proteins (especially alpha-helical trans membrane proteins) such as GPCR transporters are reported to be modulated by cholesterol. The experimental studies are so far performed on only a few proteins in a family but based on an evolutionary conservation and consensus an exploration can be done confidently within a family. However, the representation of motif has a low consensus yielding several false positives thus reducing its reliability. **Findings:** A computational hybrid clustering method based on rough set with fuzzy c-means algorithm is used to mine the cholesterol sequence from ABC family. Higher weightage is given to those sequences based on the following parameters: motifs with more number of sub motifs, number of helices bearing the motif in a protein and compliance with the orientation of the cholesterol in the membrane for its interaction with the motif. **Improvement:** A detailed study in a given super family with an approach to reduce redundancy and enrichment can improve its predictability.

Keywords: ABC transporter, CRAC/CARC, Fuzzy c-Means, GPCR, Motif, Rough Set

1. Introduction

Maintenance of cholesterol homeostasis within the cell is critical for normal human physiology[1,2]. Cholesterol is reported to be a very significant constituent of cell membranes and several membrane proteins are reported to be modulated by cholesterol[3,4]. From a previous study on peripheral benzodiazepine receptor a low consensus cholesterol binding motif was reported. The forward pattern of the motif was referred to as CRAC (L/V-X(1-5)-Y-X(1-5)-R/K) and backward pattern as CARC (R/K-X(1-5)-Y/F-X(1-5)-L/V)[5]. Proteins belonging diverse family of microorganisms (especially, viruses) and humans are reported to contain this motif[6]. Viral envelope proteins bearing short CRAC/CARC motifs are reported to interact with cholesterol containing cell membranes acting as a dagger to gain entry the host cell[7,8]. These motifs are abundantly found in many super families of the human membrane proteins and its interaction with cholesterol well studied in helical membrane proteins such as GPCR[9-11].

Like GPCRs, ATP binding cassette (ABC) transporter is a super family of Trans Membrane (TM) helical proteins containing 48 well-characterized human ABC genes. Based on sequence similarity and phylogenetic analysis, they are divided into seven distinct subfamilies, which are

*Author for correspondence

represented as ABCA through ABCG. These transporters are localized in different sub-cellular components of a cell. These highly conserved multi span transmembrane helices (1-17 TMs) utilize the energy of ATP hydrolysis to translocate a broad spectrum of molecules across the cell membrane[4]. Majority of ABC genes are reported to be important for translocating cholesterol and maintaining cellular cholesterol homoeostasis[12]. Several human genetic disorders including cystic fibrosis, neurological disease, retinal degeneration, cholesterol and bile transport defects, anemia, and drug response are linked to ABC transporter malfunction[13,14].

Different mutations in these transporters are also reported in several diseases pertaining to cholesterol such as type 2 diabetes, tangier disease, atherosclerotic cardiovascular disease, premature CVD[15-19]. As ABC transporters are reported to be modulated by membrane cholesterol[20] and are involved in transport of cholesterol across the membrane one can signify the importance of cholesterol in such proteins.

Therefore, the current work is aimed at developing a computational approach to identify those ABC transporters with the consensus motif. As a continued work from our previous report, the following considerations are made: Presence of CRAC/CARC motif (CRAC (L/V-X(1-5)-Y-X(1-5)-R/K) and CARC (R/K-X(1-5)-Y/F-X(1-5)-L/V)), relating the motif in the upper/lower part of the helix with respect to the orientation of the cholesterol in the respective membrane leaflet. The following modifications are made in the current approach to improve the reliability of the prediction. Here, a motif in a given helix is enriched by giving weightage to the number of sub-motifs the main motif carry and overall the number of motifs in a given ABC transporter will determine the cholesterol modulatory activity the ABC transporter. Such an approach would help in predicting the cholesterol binding motif more reliably[10-14]. However, the low consensus cholesterol binding motif can give rise to several hits which might make it difficult to predict a potential or relevant motif[6]. The main objective of this research is to develop a model for cholesterol with ABC transporter in transmembrane region with Fuzzy C Means (FCM) Clustering algorithm using past databases to make intelligent scientific decisions. Several computer aided diagnosis methodologies have been proposed in the literature for the diagnosis of cholesterol prediction[20-22]. Sellappan Palaniappan et al proposed an intelligent heart disease prediction system built with the aid of data mining technique like decision trees, naïve bayes and neural network[23]. Therefore, the important point of our current work is to identify signature motifs that fulfill with the cholesterol binding in ABC and report on their sub motif, occurrence and their helices. Rest of the paper is organized as follows, materials and methods are discussed in section 2, methodology, experimental evaluation is explained in Section 3 and Conclusion is described in Section 4.

2. Materials and Methods

2.1 Rough Set Theory

Rough set theory introduced by Pawalk is defined as $S = \{U, f \cup A, D, I\}$, where U is the universe of all non-empty set of object, f and A are the non-empty finite set of feature and attributes which satisfy f$\cup A = K$, D refers to the domain of all attributes such that $D = U_{a \epsilon k} D$, where D_a is the set of the value of a, I is the information function for all attributes such that $I = U_{a \epsilon A} I_a$, where I_a is a total function $I_a : \cup \epsilon D_a$. Every subset of attribute $B \subseteq K$ can be associated with an indiscernibility relation $I(B)$ defined as (1). The two sets are key concepts in rough set theory and named as the lower and upper $X \subseteq \bigcup$ ions of $\mathbf{X}$, respectively. For a subset of objects and a subset of attributes $B \subseteq f$ the lower and upper approximations of X are defined as (2) and (3), respectively24-27.

$$I(B) = \{(x, y) \in \cup x \cup | \forall b \in B, I_b(x) = I_b($$ (1)

$$\underline{B}(X) = \{x | x_B \subseteq X, x \in \cup\},$$ (2)

$$\overline{B}(X) = \{x | x_B \cap X \neq \emptyset, x \in \cup\}.$$ (3)

The lower approximation set $\underline{B}(X)$ contains all objects which can be certainly classified as objects of X based on the set of attributes B . The upper approximation set $\overline{B}(X)$ is the set of objects which can be possibly classified as objects of X . The concepts of positive, negative and boundary regions are defined as (4), (5) and (6), respectively.

$$POS_B(X) = \underline{B}(X). POS_B(X) = \underline{B}(X). \quad (4)$$

$$NEG_B(X) = \cup - \overline{B}(X). \quad (5)$$

$$BN_B(X) = \overline{B}(X) - \underline{B}(X) \quad (6)$$

2.2 Fuzzy C Means

This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. Clearly, summation of membership of each data point should be equal to one[29-35]. The algorithm is based on minimization of the following equation (7) and (8):

$$Cluster_k = \sum_{i=1}^{No.of\ Motifs} \sum_{j=1}^{No.of\ Centers} membership_{ij}^k \left\| motif_i - center_j \right\|^2, 1 \le k < \infty \quad (7)$$

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m \left\| x_i - c_i \right\|^{2,1} \le m \le \infty \quad (8)$$

Where, k > 1, No. of centers = number of clusters, association of $motif_i$ in the j[th] cluster, $motif_i$ is the i[th] of d-dimensional measured data, $center_j$ is the d-dimensional center of the cluster. Fuzzy partitioning is carried out through an iterative optimization of the objective function as shown above, with updating the membership u_{ij} and the cluster centers c_i computed using (9). The algorithm stops when, $max_{ij}\left\{\left\| membership_{ij}^{p+1} - membership_{ij}^{p} \right\|\right\} < \varepsilon$, where ε is a termination criterion between 0 and 1, whereas; p are the iteration steps. This procedure converges to a local minimum or a saddle point $Cluster_k$.

$$membership_{ij} = \frac{1}{\sum_{p=1}^{No.of\ Center} \left(\frac{\left\| motif_i - center_j \right\|}{\left\| motif_i - center_p \right\|} \right)^{\frac{2}{k-1}}} \quad (9)$$

Where, $center_j$ is computed using (10),

$$center_j = \frac{\sum_{i=1}^{No.of\ Motifs} membership_{ij}^k . motif_i}{\sum_{i=1}^{No.of\ Motifs} membership_{ij}^k} \quad (10)$$

3. Methodology and Experimental Evaluation

Trans-membrane information of all protein sequence was downloaded from UniProt database[30]. Datasets downloaded contains 494 genes information with 6 attributes α= {Gene, Protein ID, Helix Name, Length, Position, Sequence (-7 from left and +7 from right}. Table 1 depicts the information about the attributes in the dataset.

The aim of this paper is to uncover all the cholesterol consensus motif sequences available in the protein primary sequences. The signature of cholesterol motifs are in the form as shown in Table 2. Looking at Table 2; X (1-5) can be a combination of protein primary residue of maximum length 5. Considering the signature of cholesterol motif and fixing the three residues at beginning, middle, and last position, the length of any cholesterol chain can be of length from 5 (being the minimum) to 13 (being the maximum) with CRAC and CARC recognition methods. In this paper, steps have been taken to uncover all the cholesterol motif sequences in ABC data files and to design a dictionary

$$\mathcal{D} = \{d_5, d_6, d_7, d_8, d_9, d_{10}, d_{11}, d_{12}, d_{13}\}$$

$$d_i \in dictionary\ of\ length\ L, L \in [5, 13]$$

. Looking at Table 2 it can be realized that, cholesterol motif sequence for any length L >5 will have multiple motif types. For example, considering d_8, having length $L = 8$ can have motif types MT={14, 23, 32, 41} and cholesterol motif signatures can be in the form of

$$\frac{L}{\frac{VXYXXXXK}{R}} \quad \frac{L}{\frac{VXXYXXXK}{R}} \quad \frac{L}{\frac{VXXXYXXK}{R}}$$

Table 1. Information about the attributes present in UniProt

Sl. No.	Attribute Name	Description
1	Gene	Contains the information about the gene, have the value in between ABCA1 to ABCG8.
2	Protein ID	It represents the unique identification of Protein in Alphanumeric characters.
3	Helix Name	Contains the file number of helical file 1 to 17.
4	Length	Length of the protein string. Here the length is varying.
5	Position	Position of the protein residue taken from the helical file. It contains the starting position – end position value.
6	Sequence	Consist of actual protein sequence of length from helix file of appropriate positions.

$\frac{L}{\frac{VXXXXYXK}{R}}$,and $\frac{\frac{K}{RXY}}{\frac{FXXXX_L}{V}}$, $\frac{\frac{K}{RXXY}}{\frac{FXXX_L}{V}}$, $\frac{\frac{K}{RXXXY}}{\frac{FXX_L}{V}}$ and $\frac{\frac{K}{RXXXXY}}{\frac{FX_L}{V}}$ for the sequences uncovered using both CRAC and CARC methods respectively. Where, X can be any residue from a set of 20 amino acid residues R= {"..."}. Here, it can be noticed that length remains constant but cholesterol signature motifs differs only in the position with respect to 'Y/F', beginning residue and ending residue.

Figure 1 is a schematic layout of the model that describes retrieving and processing of data. Data are read sequentially one by one and whose data are extracted one at a time using sliding window technique for the length $L = \{5, 6, 7, 8, 9, 10, 11, 12, 13\}$. Dictionary D_{L_i} can be formulated using (11).

$$D_{L_i} = \frac{Motif\,length_h}{No.of\,Sequence} = \left\{ \begin{matrix} \frac{5}{14349}, \frac{6}{13855,7} \\ \hline 13361,8 \\ \hline 12867,9 \\ \hline 12373,10 \\ \hline 11879,11 \\ \hline 11385,12 \\ \hline 10891,13 \\ \hline 10397 \end{matrix} \right\} \forall i \in [1\text{-}9] \quad (11)$$

Let D be the dataset set of dimension $\{NoOfSequence_i, MotifLength_{h_i}\}$ available in $D_{L_i} for\, i = 1, \ldots, 9.$ Let for i=3, size of the dataset D will be {13361, 7}. Our motive is to search that entire feature index which matches the cholesterol forward or backward signature as per table 1. As it can be noticed the motif starts with L/V, ends with K/R and residue 'Y' in between are treated as valid candidate forward signature for cholesterol. Similarly motif starts with K/R and end with L/V can be valid candidate for backward cholesterol with residue 'Y/F' in between. For the valid position of residue 'Y' or 'Y/F' at $MotifLength_h = 7$ is $pos_{\frac{Y}{F}}$ = {3, 4 and 5}. Hence we create secondary data D' from D by extracting the information on Attribute $\{1, pos_{\frac{Y}{F}}, MotifLength_{h_7}\}$. Now, for, we will get three secondary dataset D' for different $pos_{\frac{Y}{F}}$ each of dimension {13361, 3} using by (12).

$$D'_r = Extract\left(D_{1,pos_{\frac{Y}{F}},MotifLength_{h_i}}\right), \quad where\, r = 1, \ldots |pos_{\frac{Y}{F}}|, i = 1, \ldots, 9 \quad (12)$$

Extracted information is then processed through rough set theory. Objective of this work is to extract the index of sequence S = {'LYK', 'LYR', 'VYK', 'VYR', 'KYL', 'KYV', 'RYL', 'RYV', 'KFL','KFV','KFL','KFV'} present in D'_r. For which we implement rough set theory on D'_r to group sequence in different cluster. Figure 2 represents the clusters formed by using rough set on D'_r for $MotifLength_h = 7.$ Care is been taken to store the index of each feature present in cluster using (13).

$$featureindex = M_{ind}(C_j.S) \quad (13)$$

Where, ind is the index of feature of clusters C_j , $j = 1, \ldots k$ formed after implementation of rough

Table 2. 12 Cholesterol backward and forward sequences where X= {set of 20 amino acid residues}

Motif Type	FORWARD (CRAC) (L/V-$X_{(1-5)}$-Y-$X_{(1-5)}$-R/K)	BACKWARD (CARC) (R/K-$X_{(1-5)}$-Y/F-$X_{(1-5)}$-L/V)	L = Length of Cholesterol motif
11	L/V-X-Y-X-K/R	K/R-X- Y/F -X-L/V	5
12	L/V X Y XXK/R	K/R X Y/F XXL/V	6
13	L/V X Y XXXK/R	K/R X Y/F XXXL/V	7
14	L/V X Y XXXXK/R	K/R X Y/F XXXXL/V	8
15	L/V X Y XXXXXK/R	K/R X Y/F XXXXXL/V	9
21	L/V XX Y XK/R	K/R XX Y/F XL/V	6
22	L/V XX Y XXK/R	K/R XX Y/F XXL/V	7
23	L/V XX Y XXXK/R	K/R XX Y/F XXXL/V	8
24	L/V XX Y XXXXK/R	K/R XX Y/F XXXXL/V	9
25	L/V XX Y XXXXXK/R	K/R XX Y/F XXXXXL/V	10
31	L/V XXX Y XK/R	K/R XXX Y/F XL/V	7
32	L/V XXX Y XXK/R	K/R XXX Y/F XXL/V	8
33	L/V XXX Y XXXK/R	K/R XXX Y/F XXXL/V	9
34	L/V XXX Y XXXXK/R	K/R XXX Y/F XXXXL/V	10
35	L/V XXX Y XXXXXK/R	K/R XXX Y/F XXXXXL/V	11
41	L/V XXXX Y XK/R	K/R XXXX Y/F XL/V	8
42	L/V XXXX Y XXK/R	K/R XXXX Y/F XXL/V	9
43	L/V XXXX Y XXXK/R	K/R XXXX Y/F XXXL/V	10
44	L/V XXXX Y XXXXK/R	K/R XXXX Y/F XXXXL/V	11
45	L/V XXXX Y XXXXXK/R	K/R XXXX Y/F XXXXXL/V	12
51	L/V XXXXX Y XK/R	K/R XXXXX Y/F XL/V	9
52	L/V XXXXX Y XXK/R	K/R XXXXX Y/F XXL/V	10
53	L/V XXXXX Y XXXK/R	K/R XXXXX Y/F XXXL/V	11
54	L/V XXXXX Y XXXXK/R	K/R XXXXX Y/F XXXXL/V	12
55	L/V XXXXX Y XXXXXK/R	K/R XXXXX Y/F XXXXXL/V	13

set theory in D'_r, M(.) is the function which checks the mean of j^{th} cluster with mean of S and return the index of feature available in cluster j. Now the dataset consisting of valid cholesterol sequence can be retrieved based on $featureindex$ fom D using (14).

$$D'' = D_{featureindex} \tag{14}$$

D'' is the valid possible combination of cholesterol sequence. List of cholesterol found is tabulated in Table 3, Table 4 for different motif lengths. Total number of forward and backward subsequence uncovered after the proposed methods are 143 and 373 respectively. Details are shown in Table 3, Table 4.

The objective of this paper is to find most significant motifs signatures using both CARC and CRAC motif discovery methods. After filtration, looking at the large number of available motif sequences, a better data mining method for finding most significant motif structures is required. For example; a given motif: RCYYYAL of length 7 can belong to more than one motif types such as 13, 22 and 31 with respect to Table 1. Therefore, to mine such kind of information where data can belong to more than one clusters, FCM algorithm is used in this proposed work[28,29]. For example, motif length=7, we can have valid forward cholesterol of motif type {13, 22, 31}, as per Table 1. Sequence available is in form L/V $(X)_{1-3}$ Y $(X)_{1-3}$ K/R for motif length 7. In cholesterol sequence $(X)_{1-3}$ can be combination of amino acid from length 1 to 3. In order to

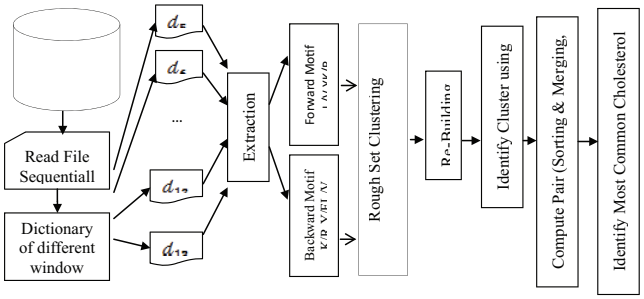

Figure 1. Schematic layout of cholesterol common signature identification model.

Table 3. Forward cholesterol sequences observed in ABC for different Motif Types (MT) matching CRAC

MT Length	5	6	7	8	9	10	11	12	13	Total
L-$X_{(1-5)}$-Y-$X_{(1-5)}$-R	3	2	5	7	9	9	2	8	0	45
L-$X_{(1-5)}$-Y-$X_{(1-5)}$-K	1	4	11	6	9	7	9	3	0	50
V-$X_{(1-5)}$-Y-$X_{(1-5)}$-R	1	0	1	8	4	4	2	0	0	20
V-$X_{(1-5)}$-Y-$X_{(1-5)}$-K	1	0	9	6	3	3	3	1	2	28
Total	6	6	26	27	25	23	16	12	2	143

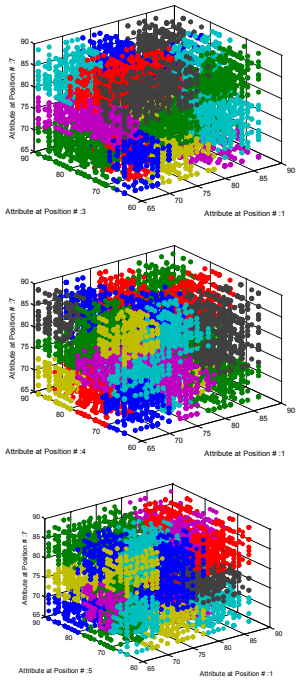

Figure 2. Scatter plot of the clusters at different position of 'Y' at (a) 3 (b) 4 and (c) 5.

Table 4. Backward cholesterol sequences observed in ABC for different Motif Types (MT) matching CARC

MT-Length	5	6	7	8	9	10	11	12	13	Total
R- X(1-5)-Y-X(1-5)-L	2	1	1	2	10	3	9	1	0	29
K- X(1-5)-Y-X(1-5)-L	3	2	2	8	7	7	2	0	0	31
R- X(1-5)-Y-X(1-5)-V	0	1	3	9	3	3	0	2	0	21
K- X(1-5)-Y-X(1-5)-V	2	3	2	9	4	1	3	3	1	28
R- X(1-5)-F-X(1-5)-L	0	0	7	14	16	14	9	8	0	68
K- X(1-5)-F-X(1-5)-L	3	4	14	16	25	11	10	10	2	95
R- X(1-5)-F-X(1-5)-V	0	2	4	10	6	9	7	2	1	41
K- X(1-5)-F-X(1-5)-V	2	2	5	18	15	10	4	3	1	60
Total	12	15	38	86	86	58	44	29	5	373

calculate the most common cholesterol sequence, weight of residue X_i, i=1,…, 20 for different position p=1,…, $MotifLength$ -1(excluding position of 'Y'), is calculated using (15). Weight represents the number of times a particular residue is present at position p.

$$W_{ij} = count_j(X_i) \quad (15)$$

Where, count (.) is the function which return number of times i^{th} residue present in position j of all sequence

Table 5(a). Forward signature motifs for ABC derived from L/V-X (1-5)-Y-X (1-5)-R/K

Protein Id	Gene Name	Helix	Sequence	Start / End	No of Sub Motifs	Total Motif	Motif Type
ABCA1	O95477	15	LIQYRFFIR	End	2	2	L2Y4R,V3Y4R
ABCA2	Q9BZC7	6	VPYMYVAIR	End	3	5	V1Y5R,V3Y3R,L4Y5R,L5Y3R,L5Y4R
ABCA2	Q9BZC7	14	LTIMCQYNFLRR	End	2		
ABCA5	Q8WWZ7	15	LLQYYEKK	End	8	8	L1Y2K,L1Y3K,L2Y1K,L2Y2K,L2Y3K,L3Y1K,L3Y2K
ABCA6	Q8N139	7	LLLALYFDK	End	3	6	L2Y2K,L3Y2K,L4Y2K,V2Y5K,L3Y5K,V4Y5K
ABCA6	Q8N139	13	VLVPSYTLLGFK	End	3		
ABCA8	O94911	7	LALAIYFEK	End	2	5	L2Y2K,L4Y2K,V1Y3K,L2Y3K,L3Y3K
ABCA8	O94911	8	LLVEYTMVK	End	3		
ABCA9	Q8IUA7	7	LVLTLYFDK	End	3	3	L2Y2K,V3Y2K,L4Y2K
ABCA12	Q86UK0	6	VENELSYVLK	End	2	8	L1Y2K,V5Y2K,V1Y5K,V1Y3K,L3Y3K,L3Y5K,V5Y3K,V5Y5K
ABCA12	Q86UK0	12	V5Y5K (55)	End	6		
ABCB2 (TAP1)	Q03518	7	LSLFLWYLVR	End	3	6	L1Y2R,L3Y2R,L5Y2R,L2Y1R,L3Y1R,V4Y1R
ABCB2 (TAP1)	Q03518	10	VLLSIYPR	End	3		
ABCB3 (TAP2)	Q03519	8	LERALYLLVRR	Start	2	2	L4Y3R,L4Y4R
ABCB6	Q9NP58	9	LLCAYFVTEQK	End	2	2	L2Y5K,L3Y5K
ABCB7	O75027	2	VLIGYGVSR	Start	2	2	L2Y3R,V3Y3R
ABCB9	Q9NP78	3	LFVGIYAMVK	Start	2	2	L4Y3K,V2Y3K
ABCB10	Q9NRK6	3	VIYGRYLRK	End	5	5	V1Y1R,V1Y4R,V1Y5K,V4Y1R,V4Y2K
ABCC1	P33527	17	LQVTTYLNWLVR	End	2	2	V2Y5R,L4Y5R
ABCC3	O15438	1	LPCYLLYLR	End	4	7	L1Y1R,L2Y4R,V4Y4R,L5Y1R,V1Y4R,L5Y4R,L3Y4R
ABCC3	O15438	15	LPLAVLYTLVQR	End	3		
ABCC4	O15439	6	VAVTLYGAVR	End	2	2	V2Y3R,V4Y3R
ABCC6	O95255	12	VHLAYLR	Start	2	2	V3Y1R,L1Y1R
ABCC7(CFTR)	P13569	5	LSVLPYALIK	End	3	3	L1Y3K,V2Y3K,L4Y3K
ABCC8	Q09428	5	VIRVRRYIFFK	End	2	4	V2Y3K,V5Y3K,V1Y4K,L4Y4K
ABCC8	Q09428	9	LAPVQYFVATK	End	2		
ABCC9	O60706	14	LGVAFYFIQK	End	2	2	V2Y3K,L4Y3K
ABCC10	Q5T3U5	1	VLSACYLGTPR	End	2	6	L3Y4R,V4Y4R,L2Y4K,V3Y1K,V4Y4K,L5Y1K
ABCC10	Q5T3U5	12	LHVYQAYWK	Start	4		
ABCD3	P28288	2	VNNFLKYGLNELK	End	2	5	L1Y5K,V5Y5K,V1Y4R,V2Y4R,L5Y4R
ABCD3	P28288	4	LATVVGYLVVSR	End	3		
ABCG1	P45844	6	LRLIAYFVLRYK	End	4	4	L2Y5K,L4Y3R,L4Y5K,L2Y3R
ABCG2	Q9UNQ0	1	LVIGAIYFGLK	End	2	2	V4Y3K,L5Y3K
ABCG4	Q9H172	6	LRLLAYLVLRYR	End	9	9	L1Y3R,L1Y2K,L1Y5R,V2Y2K,L2Y5R,L4Y5R,L2Y3R,L4Y3R,L3Y2K

Table 5(b). Backward signature motifs for ABC derived from K/R-X(1-5)-Y/F-X(1-5)-L/V

Protein Id	Gene Name	Helix	Sequence	Start /End	No Of Sub Motifs	Total Motif	Motif Type
ABCA1	O95477	9	RKGFFAQIVL	Start	10	13	R3F4L; R3F3V; R2F5L; R2F4V; K2F4L; K2F3V; K1F4V; K1F5L; K5Y5V; K5Y2L; K5Y1V
ABCA1	O95477	13	KIPSTAYVVLTSV	Start	3		
ABCA2	Q9BZC7	2	KEAFYTAAPL	Start	2	11	K3Y4L,R2F5L,K4Y4V,K4Y1V,K4Y1V,K1F1L,K1F4V,R5F4L,R5F3L,K2F4,K2F3L
ABCA2	Q9BZC7	7	KYFALYEVAGV	Start	5		
ABCA2	Q9BZC7	9	RNSKALFSQILL	Start	4		
ABCA3	Q99758	9	RKGFDIALNL	Start	4	6	R2F5L,R2F3L,K1F3L,K1F5L
ABCA3	Q99758	13	KTLDHVFLVL	Start	2		
ABCA4	P78363	1	KRQKIRFVVELV	Start	10	14	K5F4V,K5F3L,K5F1V,R4F4V,R4F3L,R4F1V,R4F1V,K2F4V,K2F3L,K2F1V,K1F4V,K1F5L,R5F4L
ABCA4	P78363	7	KDFLAQIVL	Start	2		
ABCA4	P78363	11	RKLLIVFPHFCL	Start	2		
ABCA5	Q8WWZ7	9	KDYVFAAV	Start	2	8	K3F2V,K1Y4V,K4F4L,K4F3L,K3Y5L,K3Y4L,K5F5V,K5F1L
ABCA5	Q8WWZ7	10	KIELYFQAALL	Start	4		
ABCA5	Q8WWZ7	12	KFLAVVFCLIGYV	Start	2		
ABCA7	Q8IZY2	7	RPTADVFSLAQV	Start	2	6	R5F4V,R5F1L,R3F4L,R3F3V,R2F4L,R2F3V
ABCA7	Q8IZY2	9	RRGLFAQIVL	Start	4		
ABCA8	O94911	3	RDSAFWLSWGL	Start	2	9	R3F5L,K3Y1V,K2F5V,K2F4V,K2F3L,K1F3L,K1F4V,K1F5V,R5F3L,R5F1V
ABCA8	O94911	5	KKSFLTGLVV	Start	6		
ABCA8	O94911	14	RMDVQPFLVFL	Start	2		
ABCA9	Q8IUA7	3	RESAFWLSWGL	Start	2	8	R3F5L,R3F1L,K2F5V,K2F4V,K2F3L,K1F3L,K1F4V,K1F5V
ABCA9	Q8IUA7	5	KKPFLTGLVV	Start	6		
ABCA10	Q8WWZ4	7	KMIATFFIL	Start	2	6	K5F1L,K4F2L,K4F2L,K4F1V,K3F2L,K3F1V
ABCA10	Q8WWZ4	9	KKLNCFPVL	Start	4		
ABCA12	Q86UK0	4	KTNGFILFL	Start	2	4	K3F3L,K3F1L,K4F3V,K4F2L
ABCA12	Q86UK0	13	KLGAMFVALV	Start	2		
ABCA13	Q86UQ4	10	RMYWFTNFL	Start	2	2	R3F3L,R1Y5L
ABCB1	P08183	12	KLMSFEDVLLV	Start	4	4	K3F5V,K3F4L,K3F3L,K3F2V
ABCB2 (TAP1)	Q03518	5	RRLSLFLVLVVL	Start	11	22	R4F5L,R4F4V,R4F3V,R4F2L,R4F1V,R4F1V,R3F5L,R3F4V,R3F3V,R3F2L,R3F1V,K5Y4V,K5Y2L,K5Y1L,K4Y4L,K4Y2L,K4Y1L,K1Y1L,K1Y2L,K1Y4V,K4Y5V,K4Y4L
ABCB2 (TAP1	Q03518	8	KKVGKWYQLLEV	End	9		
ABCB2 (TAP1	Q03518	9	KVGILYIGGQLV	Start	2		
ABCB3 (TAP2)	Q03519	2	RGLLGFVGTLLL	Start	3	6	R4F5L,R4F4L,R4F3L,R2Y5V,R2Y2V,R2Y1L
ABCB3 (TAP2)	Q03519	8	RALYLLVRRV	Start	3		
ABCB4	P21439	2	RYAYYYSGL	Start	3	12	R4Y2L,R3Y3L,R2Y4L,K4F2V,K3F3V,K5F4V,K5F1V,K4Y5V,K4Y2V,K4Y1V,K3F4L,K3F1L
ABCB4	P21439	3	KVGMFFQAV	Start	2		
ABCB4	P21439	7	KTEWPYFVVGTV	Start	5		
ABCB4	P21439	8	KCNIFSLIFL	Start	2		

ABCB5	Q2M3G0	7	KPEWPFVVLGTL	Start	3	3	K4F5L,K4F2L,K4F1V
ABCB6	Q9NP58	2	RISPYVLQLLL	Start	4	4	R3Y5L,R3Y4L,R3Y3L,R3Y1L
ABCB7	O75027	2	RAGAAFFNEV	Last	2	8	R5F2V,R4F3V,R3Y5V,R3F4L,R3F1L,K4F4L,K4F2V,K4F1L
ABCB7	O75027	3	RGISFVLSALV	Start	3		
ABCB7	O75027	4	KCGAQFALVTL	Start	3		
ABCB9	Q9NP78	4	RDPWFWALFV	Start	2	5	R3F4V,R3F2L,K4F1V,K4F4V
ABCB9	Q9NP78	5	KPDVAFLV	Start	1		
ABCB9	Q9NP78	6	KSMDQFSTAVV	Start	2		
ABCC1	P33527	1	KCFQNTVLV	Start	3	18	K1F3V,K1F4L,K1F5V,K4F4V,K4F1L,R4F5L,R4F3V,R2F5L,R2F3V,R4Y5L,R4Y4L,R3F5L,R3F1V,K5F4L,K412L,K4Y1L,K1F4L,K2Y3V,K4F3L
ABCC1	P33527	2	KTALGFLLWIV	Start	2		
ABCC1	P33527	3	RSRGIFLAPVFL	Start	4		
ABCC1	P33527	5	RDITFYVYFSLL	Start	4		
ABCC1	P33527	6	KVLYKTFGPYFL	Start	4		
ABCC1	P33527	12	KAIGLFISFL	Start	1		
ABCC2	Q92887	2	KQVFVGFLLIL	Start	5	12	K5F3L,K5F1L,K2F4L,K2F3L,K1Y4V,K5F4L,K5F3L,K5F2V,K2Y3L,K2Y2L,K2Y1V
ABCC2	Q92887	6	KALFKTFYMVLL	Start	7		
ABCC3	O15438	3	RAPAPVFFVTPLV	Start	3	9	R5F5V,R5F4L,R5F1V,R4Y5V,R4Y4L,R3F5L,K5F4L,K2F4L,K2F3L
ABCC3	O15438	5	RFTTFYIHFAL	Start	3		
ABCC3	O15438	6	KALLATFGSSFL	Start	3		
ABCC4	O15439	1	KCYWKSYLVL	Start	5	10	K5Y2L,K5Y1V,K1Y2L,K1Y4L,K1Y5V,R5Y2V,R5Y1L,R4F3V,K4Y2L,R4F1V
ABCC4	O15439	9	RSLLVFYVLV	Start	5		
ABCC5	O15440	5	KAGYFQSITV	Start	2	4	K3F4V,K2Y5V,K3F4L,K3F1V
ABCC5	O15440	6	KVTPFSVKSL	End	2		
ABCC6	O95255	2	KMVLGFALIVL	Start	3	11	K4F4L,K4F3V,K4F1L,R4Y4L,R4Y2L,K5F5L,K5F4L,K1Y3V,K1Y4L,R4Y4L,R4Y3L
ABCC6	O95255	5	RHLSTYLCLSL	Start	2		
ABCC6	O95255	6	KAIWQVFHSTFLL	Start	2		
ABCC6	O95255	7	KGYLLAVL	Start	2		
ABCC6	O95255	14	RSLLMYAFGLL	Start	2		
ABCC7 (CFTR)	P13569	1	RFMFYGIFLYL	Start	3	5	R3Y5L,R3Y3L,R2F4L,K3F1L,K3F5L
ABCC7 (CFTR)	P13569	7	KSLIFVLIWCL	Start	2		
ABCC8	Q09428	16	RMEYIGACVV	Start	2	2	R2Y5V,R2Y4V
ABCC9	O60706	2	RWILTFALLFV	Start	3	8	R4F4V,R4F2L,R4F1L,K4Y2L,K1F1L,K1F5L,K4F4L,K4F2L
ABCC9	O60706	10	KTFALYTSL	Start	3		
ABCC9	O60706	11	KPAEAFASLSL	Start	2		
ABCC10	Q5T3U5	2	RLAASFLLSV	Start	2	5	R4F3V,R4F1L,K1F2L,K1F4L,K1F5V
ABCC10	Q5T3U5	11	KVFTALAL	Start	3		
ABCC11	Q96J66	1	RTRLIFDALL	Start	4	7	R4F2L,R4F3L,R2F3L,R2F2L,K4F4V,R3Y5V,R3F1V
ABCC11	Q96J66	6	RLSVFFVPIAV	End	3		
ABCC12	Q96J65	2	KVFFWAL	End	2	4	K2F2L,K1F3L,K4Y3L,K4Y1L
ABCC12	Q96J65	7	KASGGYLLSL	Start	2		

ABCD1	P33897	1	RTFLSVYVARL	End	3	9	R5Y3L,R1F2V,R1F4V,R5F4L,R5F3L,K4F4L ,K4F3L,R1F3L,R1F4L
ABCD1	P33897	2	RKDPRAFGWQLL	Start	6		
ABCD2	Q9UBJ2	2	KKPRTFIIKL	Start	3	3	K4F3L,K3F3L,R1F3L
ABCD3	P28288	1	KETGYLVLIAV	Start	4	14	K3Y5V,K3Y2L,K3Y1V,R1Y2V,K5F5L,R5Y1 L,R4F5L,K4Y1L,K4Y1L,R2F4L,R2F3L,K1Y 1L,K1F3L,K1F4L
ABCD3	P28288	2	KRYLLNFIAAMPL	Start	10		
ABCD4	O14678	2	KDLEGFKTLTFL	Start	4	4	K4F5L,K4F2L,K3F4L,K3F2V
ABCG1	P45844	6	KLYLDFIVL	Start	7	7	K4F2L,K4F1V,R4F1L,R3Y2L,R3Y1V,K1Y4 V,K1Y5L
ABCG2	Q9UNQ0	4	KPKADAFFV	Start	3	3	K5F1V,K4F5L,K3F1V
ABCG4	Q9H172	6	KLYMDFLVL	Start	7	7	K4F2L,K4F1V,R3Y2L,R3Y1V,K1Y3L,K1Y 4V,K1Y5L
ABCG5	Q9H222	4	RFGYFSAALL	Start	4	4	R3F4L,R3F3L,R2Y5L,R2Y4L

Table 6. ABC transporters enriched with the motifs

Protein Id	Total motifs	Forward Motif			Reverse Motif		
		Total	Motif sequence	Motif Pattern	Total	Motif Sequence	Motif Pattern
ABCB2 (TAP1)	28	6	LSLFLWYLVR, VLLSIYPR	L1Y2R, L3Y2R, L5Y2R, L2Y1R, L3Y1R, V4Y1R	22	RRLSLFLVLVVL, KKVGKWYQLLE, KVGILYIGGQLV	R4F5L, R4F4V, R4F3V, R4F2L, R4F1V, R4F1V, R3F5L, R3F4V, R3F3V, R3F2L, R3F1V, K5Y4V, K5Y2L, K5Y1L, K4Y4L, K4Y2L, K4Y1L, K1Y1L, K1Y2L, K1Y4V, K4Y5V, K4Y4L
ABCC1	20	2	LQVTTYLNWLVR	V2Y5R, L4Y5R	18	KCFQNTVLV, KTALGFLLWIV, RSRGIFLAPVFL, RDITFYVYFSLL, KVLYKTFGPYFL, KAIGLFISFL	K1F3V, K1F4L, K1F5V, K4F4V, K4F1L, R4F5L, R4F3V, R2F5L, R2F3V, R4Y5L, R4Y4L, R3F5L, R3F1V, K5F4L, K412L, K4Y1L, K1F4L, K2Y3V, K4F3L
ABCD3	19	5	VNNFLKYGLNELK, LATVVGYLVVSR	L1Y5K, V5Y5K, V1Y4R, V2Y4R, L5Y4R	14	KETGYLVLIAV, KRYLLNFIAAMPL	K3Y5V, K3Y2L, K3Y1V, R1Y2V, K5F5L, R5Y1L, R4F5L, K4Y1L, K4Y1L, R2F4L, R2F3L, K1Y1L, K1F3L, K1F4L
ABCA2	16	5	VPYMYVAIR, LTIMCQYNFLRR	V1Y5R, V3Y3R, L4Y5R, L5Y3R, L5Y4R	11	KEAFYTAAPL, KYFALYEVAGV, RNSKALFSQILL	K3Y4L,R2F5L,K 4Y4V,K4Y1V,K4 Y1V,K1F1L,K1F4 V,R5F4L,R5F3L, K2F4,K2F3L

ABCA5	16	8	LLQYYEKK	L1Y2K, L1Y3K, L2Y1K, L2Y2K, L2Y3K, L3Y1K, L3Y2K	8	KDYVFAAV, KIELYFQAALL, KFLAVVFCLIGYV	K3F2V, K1Y4V, K4F4L, K4F3L, K3Y5L, K3Y4L, K5F5V, K5F1L
ABCC3	16	7	LPCYLLYLR, LPLAVLYTLVQR	L1Y1R, L2Y4R, V4Y4R, L5Y1R, V1Y4R, L5Y4R, L3Y4R	9	RAPAPVFFVTPLV, RFTTFYIHFAL, KALLATFGSSFL	R5F5V, R5F4L, R5F1V, R4Y5V, R4Y4L, R3F5L, K5F4L, K2F4L, K2F3L
ABCG4	16	9	LRLLAYLVLRYR	L1Y3R, L1Y2K, L1Y5R, V2Y2K, L2Y5R, L4Y5R, L2Y3R, L4Y3R, L3Y2K	7	KLYMDFLVL	K4F2L, K4F1V, R3Y2L, R3Y1V, K1Y3L, K1Y4V, K1Y5L
ABCA1	15	2	LIQYRFFIR	L2Y4R, V3Y4R	13	RKGFFAQIVL, KIPSTAYVVLTSV	R3F4L; R3F3V; R2F5L; R2F4V; K2F4L; K2F3V; K1F4V; K1F5L; K5Y5V; K5Y2L; K5Y1V
ABCA4	14	5	VPYMYVAIR, LTIMCQYNFLRR	V1Y5R, V3Y3R, L4Y5R, L5Y3R, L5Y4R	14	KRQKIRFVVELV, KDFLAQIVL, RKLLIVFPHFCL	K5F4V, K5F3L, K5F1V, R4F4V, R4F3L, R4F1V, R4F1V, K2F4V, K2F3L, K2F1V, K1F4V, K1F5L, R5F4L
ABCA8	14	5	LALAIYFEK, LLVEYTMVK	L2Y2K, L4Y2K, V1Y3K, L2Y3K, L3Y3K	9	RDSAFWLSWGL, KKSFLTGLVV, RMDVQPFLVFL	R3F5L, K3Y1V, K2F5V, K2F4V, K2F3L, K1F3L, K1F4V, K1F5V, R5F3L, R5F1V
ABCC6	13	2	VHLAYLR	V3Y1R,L1Y1R	11	KMVLGFALIVL, RHLSTYLCLSL, KAIWQVFHSTFLL, KGYLLAVL, RSLLMYAFGLL	K4F4L, K4F3V, K4F1L, R4Y4L, R4Y2L, K5F5L, K5F4L, K1Y3V, K1Y4L, R4Y4L, R4Y3L
ABCA12	12	8	VENELSYVLK, V5Y5K (55)	L1Y2K, V5Y2K, V1Y5K, V1Y3K, L3Y3K, L3Y5K, V5Y3K, V5Y5K	4	KTNGFILFL, KLGAMFVALV	K3F3L, K3F1L, K4F3V, K4F2L
ABCB4	12	2	LERALYLLVRR	L4Y3R, L4Y4R	12	RYAYYYSGL, KVGMFFQAV, KTEWPYFVVGTV, KCNIFSLIFL	R4Y2L, R3Y3L, R2Y4L, K4F2V, K3F3V, K5F4V, K5F1V, K4Y5V, K4Y2V, K4Y1V, K3F4L, K3F1L
ABCC2	12	6	VLSACYLGTPR, LHVYQAYWK	L3Y4R, V4Y4R, L2Y4K,V3Y1K,V4Y4K, L5Y1K	12	KQVFVGFLLIL, KALFKTFYMVLL	K5F3L, K5F1L, K2F4L, K2F3L, K1Y4V, K5F4L, K5F3L, K5F2V, K2Y3L, K2Y2L, K2Y1V

ABCC4	12	2	VAVTLYGAVR	V2Y3R,V4Y3R	10	KCYWKSYLVL, RSLLVFYVLV	K5Y2L, K5Y1V, K1Y2L, K1Y4L, K1Y5V, R5Y2V, R5Y1L, R4F3V, K4Y2L, R4F1V
ABCA9	11	3	LVLTLYFDK	L2Y2K,V3Y2K, L4Y2K	8	RESAFWLSWGL, KKPFLTGLVV	R3F5L,R3F1L,K2F5V,K2F4V,K2F3L,K1F3L,K1F4V, K1F5V
ABCC10	11	6	VLSACYLGTPR, LHVYQAYWK	L3Y4R, V4Y4R, L2Y4K, V3Y1K, V4Y4K, L5Y1K	5	RLAASFLLSV, KVFTALAL	R4F3V, R4F1L, K1F2L, K1F4L, K1F5V
ABCG1	11	4	LRLIAYFVLRYK	L2Y5K, L4Y3R, L4Y5K, L2Y3R	7	KLYLDFIVL	K4F2L, K4F1V, R4F1L, R3Y2L, R3Y1V, K1Y4V, K1Y5L
ABCB7	10	2	VLIGYGVSR	L2Y3R, V3Y3R	8	RAGAAFFNEV, RGISFVLSALV, KCGAQFALVTL	R5F2V, R4F3V, R3Y5V, R3F4L, R3F1L, K4F4L, K4F2V, K4F1L
ABCC9	10	2	LGVAFYFIQK	V2Y3K, L4Y3K	8	RWILTFALLFV, KTFALYTSL, KPAEAFASLSL	R4F4V, R4F2L, R4F1L, K4Y2L, K1F1L, K1F5L, K4F4L, K4F2L

found in a group. Now the sequence is sorted with respect to the *W* to arrange in descending order. Applying FCM on motifs sequence led to discovery of clusters with similar signatures that help to calculate the weight of individual residue with respect to their position.

Table 5 (a), Table 5 (b) is the summary of significant cholesterol signatures motifs discovered using the FCM algorithm with details regarding *Protein ID* (Column #1), *Gene name* (Column #2), *Helix* (Column #3), *Conservered motif signature* (Column #4), *Start/End* where cholesterol motif is found in ABC protein (Column #5), Number of sub motif (Column #6) and *Motif type* (Column #7). The results obtained clearly shows that the combinations one can obtain from the CRAC or CARC is indeed restricted and can be further developed as a signature motif depending on the family, helix or the location in the membrane leaflet. Applying the FCM algorithm to identify motifs in cholesterol sequence resulted in significantly more number of backward motifs 143 than the forward motif 373.

The CARC/CRAC being a low consensus motif can give rise to several possibilities. Therefore, in order to further improvise the prediction, motifs from a given helix were checked for number of submotifs it contains. Here the assumption is that greater the number of submotifs, higher is the chances for its interaction with cholesterol. A total number of forward and backward motifs in a given ABC transporter showed numbers ranging from 2 – 28 indicating that the motifs are not uniformly disturbed across all ABC transporters but there is a polarization depending on its function as many of them are sterol transporters.

With a cut-off of 10 as motifs observed in a transporter, 20 transporters were obtained as shown in Table 6. Many of these proteins are involved in sterol transport, accumulation or associated with rafts. ABCB2, ABCC1, ABCD3, ABCA1, ABCA4 has more backward motifs (13 to 22) rising from either one or two helices only. Here it is interesting to note that the length of the sequence containing the motif (motif sequence) has no relation with the number of submotifs it contained. ABCA12 contains a 13-aa peptide with only 8 motifs in it while ABCG1 and ABCG4 a single 9-amino acid motif long contained 7 submotifs in it emphasizing the importance of the motif in ABCG1 and ABCG4. Altogether it is observed that enrichment of the submotifs improved the predictability of the cholesterol consensus motif.

4. Conclusion

The low consensus cholesterol binding motif (CARC and CRAC) gives rise to several possible sequences matching the pattern of which only a few of them might have a biological relevance. In our previous study, the number of motifs that are present in the GPCRs along with some signature motifs in different helices and subclasses were reported. In the current study, apart from extraction of the CRAC/CARC motifs, the significance of the motif was weighted by the number of sub-motifs it harbored to predict its binding to cholesterol reliably. Our approach indeed correlates with the reports and activity of ABC transporters. It was observed that the CRAC/CARC motifs were highly enriched in the TM helices of those that were modulated by cholesterol and/or involved in cholesterol transport. Mutations in these transporters have also been reported to have impaired function due to deficiency in cholesterol transport. From this study we report a much reliable approach to predict the significance of cholesterol binding motif in ABC transporters.

5. References

1. Chauhan NB. Membrane dynamics, cholesterol homeostasis, and Alzheimer's disease. Journal of Lipid Research. 2003; 44(11):2019 –29.
2. Oram JF. Molecular basis of cholesterol homeostasis: lessons from Tangier disease and ABCA1. Trends in Molecular Medicine. 2002; 8(4):168 –73.
3. Paila YD, Chattopadhyay A. Membrane cholesterol in the function and organization of G-protein coupled receptors. Subcell Biochem. 2010; 51:439–66.
4. Schmitz G, Kaminski WE. Phospholipid transporters ABCA1 and ABCA7. In: Broer S, Wagner CA. editors. Membrane Transporter Diseases, Kluwer Academic, Plenum Publishers, New York, 2004; 291–99.
5. Tripathy R, Mishra D , Konkimalla VB. A novel fuzzy C-means approach for uncovering cholesterol consensus motif from human G-protein coupled receptors (GPCR). Karbala International Journal of Modern Science. 2015; 1(4):212–24.
6. Babu MM, Lee RV, de NS, Groot J, Gsponer G. Intrinsically disordered proteins: regulation and disease. Current opinion in structural Biology. 2011; 21(3):432–40.
7. Ahmad N. The vertical transmission of human immunodeficiency virus type 1: molecular and biological properties of the virus. Crit Rev Clin Lab Sci. 2005; 42(1):1–34 .
8. Abrams EJ, Wiener J, Carter R, Kuhn L, Palumbo P, Nesheim S, Lee F, Vink P, Bulterys M. Maternal health factors and early pediatric antiretroviral therapy influence the rate of perinatal HIV-1 disease progression in children. Aids. 2003; 17(6):867–77.
9. Ballesteros J, Weinstein H. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. Methods Neurosci, San Diego, CA: Academic Press, 1995; 25(1):366–428.
10. Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SG, Thian FS, Kobilka TS, Choi HJ, Kuhn P, Weis WI, Kobilka BK. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. Science. 2007; 318(5854):1258–65.
11. Zhang Y, Devries ME, Skolnick J. Structure modeling of all identified G protein-coupled receptors in the human genome. PLoS Comput Biol. 2006; 2(2):29.
12. Higgins CF. ABC transporters: from microorganisms to man. Annu Rev Cell Biol. 1992; 8(4):67–113.
13. Altschul S F, Madden TL, Schaffer AA, Zhang J , Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25(17):3389–402.
14. Hubbard TJ, Ailey B, Brenner SE, Murzin AG, Chothi C. SCOP: a Structural Classification of Proteins database. Nucleic Acids Res. 1999; 27(1CC):254–56.
15. Patel DC, Albrecht C, Pavitt D. Type 2 diabetes is associated with reduced ATP-binding cassette transporter A1 gene expression, protein and function. PLoS ONE. 2011; 6(7):1–8.
16. Wilson B, Angela A, Marcil M, Clee SMLH, Zhang K, Roomp M, Dam VL, Yu Y. Mutations in ABC1 in Tangier disease and familial high-density lipoprotein deficiency. Nature Genetics. 1999; 22(4):336–45.
17. Lawn RM, Wade DP, Garvin MR, Wang X, Schwartz K, Porter JG, Seilhamer JJ, Vaughan AM, Oram JF. The Tangier disease gene product ABC1 controls the cellular a polipoprotein-mediated lipid removal pathway. The Journal of Clinical Investigation. 1999; 104(8):R25-R31.
18. Mendez AJ. Cholesterol efflux mediated by apolipoproteins is an active cellular process distinct from efflux mediated by passive diffusion. Journal of Lipid Research. 1997; 38(9):1807–21.
19. Oram JF, Mendez AJ, Lymp J, Kavanagh TJ, Halbert CL. Reduction in apolipoprotein-mediated removal of cellular lipids by immortalization of human fibroblasts and its reversion by cAMP: lack of effect with Tangier disease cells. Journal of Lipid Research. 1999; 40(10):1769–81.
20. Gottesman MM, Fojo T, Bates SE. Multidrug resistance in cancer: role of ATP-dependent transporters. Nat Rev Cancer. 2002; 2:48–58.
21. Patil SB, Kumaraswamy YS. Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial

Neural Network. European Journal of Scientific Research. 2009; 31(04):642–56.
22. Ordonez C. Association rule discovery with the train and test approach for heart disease prediction. IEEE Transactions on Information Technology in Biomedicine. 2006; 10(2):334–43.
23. Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. IEEE/ACS International Conference on Computer Systems and Applications. 2008; 108–15.
24. Pawlak Z. Rough sets and data analysis, Fuzzy Systems Symposium. Soft Computing in Intelligent Systems and Information Processing. 1996; 1–6.
25. Skowron A, Rauszer C. The discernibility matrices and functions in information systems. Intelligent Decision Support. 1992; 331–62.
26. Midelfart H, Komorowski J, Nørsett K, Yadetie F, Sandvik A, Lægreid A. Learning rough set classifiers from gene expression and clinical data. Fundamenta Informaticae. 2002; 53(1):155–83.
27. Srimani PK, Koti MS. Knowledge Discovery in Medical Data by using Rough Set Rule Induction Algorithms. Indian Journal of Science and Technology. 2014; 7(7):905–15.
28. Anjaneyulu GSGN, Kaushika C, Kumar A. Content based Image Search using Rough Set and Representative Graph. Indian Journal of Science and Technology. 2015; 8(S2):257–61
29. Kumar DS, Ezhilarasu P, Prakash J, Kumar KBA. Assimilated Strong Fuzzy C-means in MR Images for Glioblastoma Multiforme. Indian Journal of Science and Technology. 2015; 8(31):1–8.
30. Revathy S, Parvaathavarthini B, Rajathi S. Futuristic validation method for rough fuzzy clustering. Indian Journal of Science and Technology. 2015; 8(2):120–27.
31. Venu N, Anuradha B. Multil-Kernels Integration for FCM Algorithm for Medical Image Segmentation using Histogram Analysis. Indian Journal of Science and Technology. 2015; 8(34):1–8.
32. Ravindraiah R, Reddy SCM, Prasad PR. Detection of Exudates in Diabetic Retinopathy Images using Laplacian Kernel Induced Spatial FCM Clustering Algorithm. Indian Journal of Science and Technology. 2016; 9(15):1–6.
33. Kavitha R, Christopher T. An Effective Classification of Heart Rate Data using PSO-FCM Clustering and Enhanced Support Vector Machine. Indian Journal of Science and Technology. 2015; 8(30):1–9.
34. Lin PL, Huang PW, Kuo CH, Lai YH. A size-insensitive integrity-based fuzzy c means method for data clustering. Pattern Recognition. 2014; 47(5):2042–56.
35. Consortium U. The Universal Protein Resource (UniProt). Nucleic Acids Re. 2007; 35(1):D193–D197.