IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Computational, biochemical and genomic analyses for a novel gene on human chromosome 12, C12ORF24 and structural analyses of its cognate protein, FAM216A.

Vishnupriya J
Assisstant professor,
PMCHRI

Abstract:

The Human Genome Project was initiated to completely sequence our genome there still remain "uncharacterized crucial genes" in different chromosomes , under genetic changes lead to cancer prone diseases, solid tumors, leukemia, etc. This study represents a brief analysis particularly on chromosome 12 Open Reading Frame 24. The gene named as C12ORF24 and its novel protein as FAM216A, which is at the cytogenetic location q24.11.The C12ORF24 gene sequence retrived from the NCBI database and the BLAST (-N &-P) analysis performed within the species and also between the ortholog gene species. BLASTN result shows 100% nucleotide sequence similarity of Human gene C12ORF24 with the Human GPN3 transcript (reverse) and the genomic sequence of alternate assembly of Human C1 1.1. And shows 77% similarity over the species rat (C12) and mouse (C5) containing homolog gene. Likewise the amino acid sequence of human C12ORF24 compared with that of rat and mice showed 68% similarity in rat and 85% similarity in mice. So humans and mice , two species separated by more than million years of evolutionary history, have similar genes. Further Human gene study about the exon/intron organization revealed that three species containing homolog gene C12ORF24 has 7 exons and 6 introns and TSS, Poly(A) signal site, UTR's(5'&3'), epigenetics -were annotated using the respective public database tools. The cognate protein (FAM216A) study revealed about its post –translational modification sites especially phosphorylation which determines its functional state of activity via ExPASy; its primary and secondary structure was predicted and compared with that of rat and mice having homolog protein. Tissue distribution inferred from the whole human proteome project also revealed a prominent existence of FAM216A protein in brain and testis, thereby connecting psychiatric and hormone related functions. The STRING db extensive protein – protein interactions results has shown that FAM216A protein interacts with Ubiquitin, suggesting that its action may be pivotal in determining brain function(s).This pivotal information will certainly increase our future knowledge on this human gene and may lead to the development of novel treatment modalities in this genomic era.

Keywords: Chromosome 12 open reading frame 24(C12ORF24), Family with sequence similarity 216, member A(FAM216A),Basic Local Alignment Search Tool for -Nucleotide and –Protein(BLAST-N&-P),GPN-loop gtpase 3(GPN3),Expert Protein Analysis System(ExPASy), Search Tool for the Retrieval of Interacting Genes/Proteins (STRING db),Signal peptide site(Sp site).

Introduction:

The advent of human genome project and its successful analysis of major "genome shores" revealed that there still exists innumerable number of "uncharacterized" and "fully functional" genes (ORF). These genes are capable of expressing their genes through mRNAs and produce their cognate proteins/enzymes (Lander, *et al.,* 2001). The ORFs on different chromosomes for various species can be obtained in a cDNA format from various institutes of gene collection such as Mammalian Gene Collection(Gerhard, 2004).Likewise the ORF of interest in particular chromosome obtained via human ORFeome data base (Lamesch, *et al.*,2007). Analyses of the human chromosomes and extensive analyses of ORFs made to reveal the existence of classical genes with selective exons within these genes which are regularly intervened by introns strictly following "GT/AG splicing rules" (M. Burset,*et al*,2000) , they are fully capable of transcribing to functional mRNAs and synthesize active protein products that have to be identified, analyzed and studied to an extent relevant to the normal cellular function and under disease conditions.

Our careful search on all 23pairs of chromosomes and later focused on one strange name designated by the HGP called "C12ORF24" on chromosome 12,which is fully capable of giving rise to a functional protein, FAM216A .Our extensive analyses of billions of Human Genome Sequences both curated, non-curated and expressed sequence tag (ESTs) databases using many computational and bioinformatics tools revealed the conspicuous existence of this gene, C12ORF24 and its protein product, FAM216A.

The chromosome12 is being annotated by an International team consisting of members from USA, India, Taiwan, Singapore and Thailand .Chromosome 12 has a length of approximately 134 million bases (MB) (133,275,309 bp).The total number of genes in chromosome 12 is 6,696. It is an autosomal chromosome and accounts for 4.0-4.5% of the entire DNA content of a cell and also contains the Homeobox C gene cluster. The Chromosome-Centric Human Proteome Project (C-HPP) identifies less defined proteins lacking sufficient protein evidence in the human genome (Paik , 2012 and Legrain ,2011). The neXtProt release 2013-08-17 contains approximately 5,000 genes in the human genome with no experimental evidence at the protein level (Lydie, 2012). Integration of high level proteomic evidence from four major databases (neXtProt , Global Proteome Machine (GPMdb) , PeptideAtlas , and Human Protein Atlas (HPA)) along with Ensembl databases, it was observed that there are 1066 protein coding genes on chromosome 12, which are located in different sub-cellular locations and has different molecular functions (Prasad, *et al*, 2009), of which 171 are defined as "missing proteins" based on the weak or complete absence of experimental evidence, which include the protein - FAM216A(Manda, *et al.*, 2014).

The C12ORF24 gene produces FAM216A protein (HSU79274) which has its cytogenetic location at 12q24.11. FAM216A protein belongs to the protein family "FAM216"and was expected to localize in nucleus, cytoskeleton and mitochondria, given under the Gene Card of Human Gene database. There are 92 organisms having orthologs with human protein FAM216A. The phylogenetic tree for the gene C12ORF24 in human, Norway rat and house mouse were given under the family TF337546 (N/A, N/A). The sequence of this gene is defined by 210 GenBank accessions from 186 cDNA clones, some from brain (17 times), testis (16), prostate (9), carcinoma, cell line (7), lung (6), skin (6), head and neck (5) and 74 other tissues.

In accordance with the goals of the Human Genome Project, our primary aim was to analyze the sequence of the euchromatic gene containing in this sub-metacentric chromosomal region. In this study, C12ORF24 gene sequence retrived from the NCBI gene bank and BLASTN analysis carried out for the nucleotide sequence similarity alignment within the species *Homo sapiens* and also compared with the other homolog gene species especially *Rattus norvegicus* (C12) and *Mus musculus* (C5) for the sequence similarity. Since the evolutionary rate for this gene in all the species is same as ~1.16 .The BLASTP analysis also carried out for the amino acid sequence similarity between these three species. The studies on Epigenetics, Post Translation modification of protein FAM216A such as acetylation, phosphorylation, ubiquitination, etc. and the Protein-protein interaction (PPI) (NCBI, EMBL-EBI) and the 3D structural analysis (PDB, MMDB, *Sybyl* and *RaptorX*) of the protein in three different species were studied to determine its functional activity.

Methods and materials:

From the NCBI database the nucleotide sequence of the gene C12ORF24 was withdrawn for analysis of its features and for further protein sequence analysis.

For the prediction of the transcripts of the genes from this region , both the nucleotide sequence of the transcripts and their predicted conceptual translations were analyzed with various computer software programs, including BLASTN, BLASTP, and PROSITE(smith et al 1996).

One of the computational program we used to predict exons in gene C12ORF24 via Vertebrate and Genome Annotation (VEGA) ,which was developed by the Welcome Trust Sanger Institute and is in close association with other annotation databases, such as ZFIN (The Zebrafish Information Network), the Havana Group and GenBank (Benson D.A.,etal,2016).

The DNA *functional site miner* (DNAFS Miner) have two software tools (1) *TIS Miner* (ATG) for translation initiation site (TIS), (2) *Poly(A) Signal Miner* to predict polyadenylation (poly(A)) signal (Figure 12) (http://dnafsminer.bic.nus.edu.sg/) for the C12ORF24 Gene sequence.

The GP Miner (Lee, et al, 2012) web server for mining transcriptional regulatory elements in mammalian gene promoter regions, annotations for the regulatory features such as the transcription factor binding sites, CpG islands, tandem repeats, the TATA box, the CCAAT box, the GC box, over-represented oligonucleotides, DNA stability and GC-content are done for the gene sequence C12ORF24.

The CDS region in the genomic sequence (C12ORF24) can be identified by the CCDS databases(Pruitt, *et al*,2009; Harte, *et al.*,2012; Farrell , *et la.*,2014).

From ExPASy we can find the Isoelectric point and molecular weight of protein FAM216A (http://web.expasy.org/compute_pi/) and its modification sites can be determined by PhosphoSitePlus (PSP),for the study of protein post-translational modifications (PTMs) including phosphorylation , ubiquitination , acetylation and methylation .

The protein primary and secondary structure can be identified by submitting the protein sequence in the Protein Data Base (PDB).Using *RaptorX* program(Wang, *et al*,2011;Jianzhu Ma,etal,2013; Jian Peng,2011) we can get the 3D structure of the protein Fam216A and GOR method (Kloczkowski , *et al*,2002 and Sen , *et al*,2005) for secondary structure prediction.

To get idea about function of this protein FAM216A, the interaction of this one with other proteins' was predicted using ExPASy-STRING database (Franceschini, *et al.*, 2013).

Cancer Cell Metabolism Gene Database (ccmGDB) used for the annotation of cell metabolism of this gene C12ORF24 (FAM216A) in cancer (Kim, *et al*, 2015) and its tissue expression were studied using available antibodies.

U.S. Patents for C12ORF24:

Since future drugs already under the pipe were mentioned we searched the inexhaustible repository of US Patent Office for C12ORF24 gene yielded nine patents which are mentioned below. Therefore there is a good possibility that this gene and its products is a target for cancer research. FAM216A as a search word in US Patent Office yielded the following patents for future pharmaceutical purposes.

1. Patent # 9,139,867 - Rationale, methods, and assays for identifying human and non-human primate taste specific genes and use thereof in taste modulator and therapeutic screening assays

2. Patent # 8,669,066 - Rationale, methods, and assays for identifying human and non-human primate taste specific genes and use thereof in taste modulator and therapeutic screening assays

3. Patent # 8,580,757 - Methods of modulating mesenchymal stem cell differentiation

4. Patent # 8,492,328 - Biomarkers and methods for determining sensitivity to insulin growth factor-1 receptor modulators

5. Patent # 7,932,058 - Rationale, methods, and assays for identifying human and non-human primate taste specific genes and use thereof in taste modulator and therapeutic screening assays

6. Patent # 7,615,349 - Melanoma gene signature.

Results:

Initial library Search against all the available databases in the NCBI web site from "Assembly" to "Unigene" about this gene C12ORF24 resulted in the following information- PubMed Central retrieved 9 full-text based journal articles but none of them are direct relevant to this gene.GEO database gave us 1,460 data. "Unigene" analysis gave 4 sets of information on this gene, under Health related information did not yield any meaningful information. PubMed databases ranging from "Gene" (7); "Clone", (0); "Conserved Domain" (0); Epigenomics (0); "EST" (36 mRNAs under specific search); under "Genome" database (even 0); dbVar which is genomic structural variation studies yielded 55 segments; under "Probe" section 17 primer

sets were retrieved which are sequences specific probes and primers to analyze this gene. Other databases gave me the following information: "Homologene" (7) similar to "Gene", "OMIM" (0); "Nucleotide" (36)'' "Protein" (17) and "SNP" database gave 2,169 variants . Finally analyzed for chemicals (pharmaceuticals) possible modulating this gene and its protein products. Under the heading of "PubChem Bioassays 3 compounds retrieved under "Bioactivity Screening Studies" and under "PubChem Substances which are deposited chemicals (possible drugs modulating this gene) and other chemical informations 9 of them were retrieved.Under chromosomal Map Viewer many species encompass this gene on various chromosomes underscores its positive existence. The various species having C12ORF24 homolog gene(Table 1) .

Under evolutionary descriptions (OrthoDB) the phyletic profile resulting as nine genes of single copy in nine species and there is no multi-copies available in any species under "Primates" .The evolutionary rate for this gene in all the species is same as ~1.16 .The phylogenetic tree for the gene C12ORF24 in human, Norway rat and house mouse are available under family TF337546(N/A,N/A) .

Under Gene Map - C12ORF24 gene location in various species

- *Homo sapiens (Human); 12*
- *Gorilla gorilla (Western gorilla); 12*
- *Pan troglodytes (Chimpanzee); 12*
- *Pan paniscus (Pygmy chimpanzee); 12*
- *Pongo abelii (Sumatran orangutan);12*
- *Bos taurus (Cow); 17*
- *Bison bison (Bison); Undesignated*
- *Bubalus bubalis (Water buffalo); Undesignated*
- *Callithrix jacchus (White-tufted-ear marmoset); 9*
- *Chlorocebus sabaeus (Green monkey); 11*
- *Macaca fascicularis (Crab-eating macaque); 11*
- *Macaca mulatta (Rhesus monkey); 11*
- *Mus musculus (Laboratory mouse); 5*
- *Nomascus leucogenys (Northern white cheeked gibbon); 10*
- *Otolemur garnetti (Small-eared galago); Undesignated*
- *Oryctolagus cuniculus (Rabbit); 21*
- *Papio anubis (Olive baboon); 11*
- *Rattus norvegicus (Norway rat); 12*
- *Rhinopithecus roxellana (golden snub-nosed monkey); Undesignated*
- *Saimiri boliviensis (Bolivian squirrel monkey); Undesignated*
- *Tarsius syrichta (Philippine tarsier); Undesignated*

Genomic sequence: The chromosome 12 specific genomic sequence was obtained from the NCBI-Gene bank (GI:568815586) as a Reference Sequence: NC_000012.12.From this we obtained C12ORF24 specific genomic

sequence of about >29,757 bases were used and curated(Figure 22). The curated genomic location numbering starts from 110,468,427 bp from pter and ending at 110,490,387 bp from pter and these are oriented in plus strand.

BLASTN RESULTS: The results of blast for query nucleotide sequence of human C12ORF24 mRNA (cDNA form) containing 1529bp in length. Undergoing BLASTN analysis their found some similarity between the nucleotide sequence of query over other available sequences in the database (nr). On undergoing mega blast for the human C12ORF24 against Human G+T shows 100% identity with the *Homo sapiens* GPN loop GTPase 3 (XM_005253896.4 and NM_001164372.1) at transcript level in reverse and at genomic level shows 100% identity with the *Homo sapiens* C12 alternate assembly CHM 1 1.1(NC 018923.2) and Homo sapiens C12 GRCh38.p7 primary assembly(NC 000012.12), predicted one.

The query nucleotide sequence alignment of C12ORF24 gene is **100%** identical to the subject *Homo sapiens* C12 alternate assembly CHM 1 1.1(NC 018923.2) in order of

query range1-592 matches the subject range 110874062-110874653;
query range 1150-1521 matches the subject range 110895633-110896004;
query range 883-1070 matches the subject range 110892148-110892335;
query range754-886 matches the subject range 110891940-110892072;
query range 633-755 matches the subject range 110890694-110890816;
query range1069-1152 matches the subject range 110893477-110893560;
query range 592-633 matches the subject range 110878712-110878753.

Within the same species (Human) the gene sequence in the C12 is identical to the alternate assembly of sequence in the C1.Comparing over other species containing ortholog of the gene C12ORF24, there observed that the human C12ORF24 nucleotide sequence shows 77% similarity over other two species, *Rattus norvegicus* (C12) and *Mus musculus* (C5).

Messenger RNA(mRNA) of Gene C12ORF24 and organization of exons and introns: The sequences of coding and non coding one of the gene C12ORF24 are obtained from the NCBI Reference sequence NM_013300.The Human gene C12ORF24 contains seven exons and six distinct introns strictly following the GT/AG splicing rule. On Undergoing transcription process this gene produces 10 alternative spliced mRNA that appear to differ by truncation of 5' and 3' end ;possibly alternative - 2 promoters ; 2 polyadenylation sites and overlapping and non-overlapping exons. 434 bp of this gene are antisense to spliced gene GPN3, 639 to VPS29, raising the possibility of regulated alternate expression. Firstly transcribed to give Pre-matured mRNA and then to matured form of mRNA containing 1529bp in *Homo sapiens* ,then translated to give mature protein of 273 amino acid in length named ,FAM216A.The 5'-UTR contains about 512 bases and the 3'-UTR contains about 248 bases. The Transcription start site (ATG) for this gene C12ORF24 in the species *Homo sapiens* was notified at two regions one at the position 5533b and other at the 5584b.There is a stop codon "TGA" at the position 5506b.

By comparing the ortholog of this gene with other species such as *Rattus norvegicus* (mRNA-1270bp) and *Mus musculus* (mRNA-1014bp), the result obtained was very important that these two species have 7 exons and 6 introns. The first exon in the human C12ORF24 was observed to be lengthy than compared to other two species such as rat and mice (Figure 2) that result in their more alkaline nature. But the exons 1 to 6 are in same length in the two species rat and mice, while in human

exon 3 and exon 6 are same in length with the other two species.For this gene C12ORF24,the CCDS data set output containing seven CCDS (CCDS ID:31899.1) of about 822nucleotides in *Homo sapiens* was obtained.

Results of Identification of PolyA signal Site-DNAminer Functional Site: At the 3'end of mRNA there is a polyA tail. There are two polyA signal site detected by running the tool DNAminer-PolyA signal recognition site. The DNA PolyA signal minor output for the PolyA signal "AATAAA" was found to be at the position 1518b and 1521b. Using the other type of PolyA signal site "ATTAAA" we got the result at the position 1362b and at 1497b for *Homosapiens.* The PolyA signal site was also analysed in other species containing homolog of this gene such as *Rattus norvegicus* and *Mus musculus*.In rat the PolyA site was found at the position 1240b(for "AATAAA") and1221b(for "ATTAAA").And in mice at the position 962b and 982b for the PolyA signal "ATTAAA" exist.

The **protein sequences** for the gene C12ORF24 are obtained from the NCBI Gene bank containing Protein Id NP_037432.2 . The BLASTP was carried out and the result shows that the amino acid sequence of Human FAM216A (273aa) was 68% identical with *Rattus norvegicus,* and 70% identical with *Mus musculus.* Since Human FAM216A contains extra peptide sequence when compared with other two species, rat and mice(Figure 3). The ExPASy **Compute pI/Mw** for the protein FAM216A in the species *Homo sapiens* is 9.39/30791.85 which is very alkaline. Upon further analysis, found one other protein/enzyme which is human tartrate resistant acid phosphatase (ACP5, NM_001111035) is somewhat larger (M.W. 36,599 or 36.6 kDa) and has a pI of 9.8.Likewise analysed for the other species such as *Rattus norvegicus* (NM_001008290) and *Mus musculus* (NM_026883) which has the ortholog gene C12ORF24 and also Protein FAM216A but the length varies, the rat has the protein length of about 261 amino acid and the mice has the length of about 251 amino acid. The protein sequence was obtained in the same way as done for *Homo sapiens* from the NCBI-Gene bank. The theoretical pI/Mw for the species *Rattus norvegicus* was found to be 9.56 / 29482.68 and for the species *Mus musculus* was found to be 9.74 / 28339.42.The human FAM216A has extra peptide sequence

Protein 3D structure was predicted by the *RaptorX* program using the sequence available.Using this prediction was done for the three dimensional structure of protein FAM216A but it is considered less confidence until the structure is predicted using the X-Ray crystallography method. From this Raptor X programme output we get to know that the protein FAM216A in human contains "Four "domains (Figure 4).The domain 1 has a five alpha helixes and the domain 3 has a two alpha helixes .The domain 4 has a beta pleated sheets .The domain four has a "beta hairpin turn" which helps in structure stability in tertiary and quaternary structure of protein. Since the hairpin bend helps in the formation of the nucleation site for the folding of protein .The protein active site under any domain were tried to find but the table work must needed for the analysis. Using Prosite (PSS) scanning we got the structure as shown in the figure 5. It gives an idea that apart from four cysteine present in the sequence of protein FAM216A in human only two cysteine involved in the formation of single disulphide bond, thus other two cysteine which is not involving in the -S-S- bond formation will be helping in maintaining the stability of protein structure.

The three dimensional structure for the other species such as *Rattus norvegicus* and *Mus musculus* was obtained from the Raptor X programme for comparison towards *Homo sapiens*.The output given by *Raptor X* programme was observed keenly.

There were only “three domains” was noticed in the 3D structure of protein FAM216A in both the species rat and mice. The mouse consists of four alpha helixes in domain 1 and two alpha helixes in domain 2. The mice protein FAM216A doesn’t contain beta pleated sheet in their 3D protein structure. Rat protein contains same as three domains in its 3D protein structure and there are four alpha helixes in domain 1. Domain 2 contains three alpha helixes and one beta pleated sheet, while domain 3 contains one alpha helixes (Figure 6).

The**G**arnier-**O**sguthorpe-**R**obson method used to determine the secondary structure of Protein and explains with the scores of individual aminoacid involved in the particular structure formation such as beta sheet, helix and coil. The results of GOR method for protein FAM216A sequence was shown in the table (Figure 7).

The Chemical modification of proteins is an important tool for probing natural systems, creating therapeutic conjugates and generating novel protein constructs. Site-selective reactions require exquisite control over both chemo- and region-selectivity, under ambient, aqueous conditions. Expasy-Netphos predicted the phosphorylation sites in the Protein sequence FAM216A.The three forms of amino acid was specially checked for modification such as serine, threonine and tyrosine. There was about 19-serine, 5-threonine and 2-tyrosine Phospho-modifications were found (Figure 8). As the matter of function the phosphorylation or dephosphorylation may activate or inactivate the proteins function.

The other modification like acetylation , glycosylation and myristolylation was also found.The N-terminal glycosylation and N-terminal myristoylation were found so that the sequence of the protein at N-terminal was protected from cleavages and other proteosomal action. Since the protein FAM216A has N-terminal myristoylation it can be concluded that this protein is not a membrane bounded protein.The SignalP 4.1 server predicts the presence and location of signal peptide cleavage sites in amino acid sequence,from this server we get to know that there is no SP site in the amino acid sequence of protein FAM216A for all the three species *Homo sapiens* ,*Mus musculus* and *Rattus norvegicus*,which contain ortholog of this protein

Protein-protein interaction maps provide a valuable framework for a better understanding of the functional organization of the proteome.To determine Protein-protein Interactions by biochemical methods such as protein affinity chromatography, affinity blotting, co-immunoprecipitation, and cross-linking; molecular biological methods such as protein probing, the two-hybrid system, and phage display: and genetic methods such as the isolation of extragenic suppressors, synthetic mutants, and unlinked non-complementing mutants. The STRING database, where the protein interactions include direct (physical) and indirect (functional)associations. From this we get the proteins hat are interacting with the novel protein FAM216A are UBC, C16ORF62, UGT3A2, LYSMD2, STK31, PDXDC1 nd MKI67IP. Through these interacting known protein we can find the function of the novel protein FAM216A.

The two antibodies for the protein FAM216A provided by the Sigma Aldrich Company ,derived from a host species “Rabbit”.These are affinity purified using the PrEST-antigen as affinity ligand and the product name for these two antibodies are given as Antibody HPA038286 (AB-1) and Antibody HPA038287 (AB-2), but the general term is Anti-C12ORF24. The antigen length is about 74 amino acid for AB-1 and 76 amino acid for AB-2 (Figure 38).The AB-1 bounds at the position 120-193aa and the AB-2 bounds at the position 193-268aa in their respective antigens.Each antibody is tested by immunohistochemistry against hundreds of normal and disease tissues. These images can be viewed on the Human Protein Atlas (HPA) site by clicking on the Image Gallery link on the website

"proteinatlas.org".From the Human protein atlas we can get the immunohistochemical and the immunofluorescent studies of this protein FAM216A.

Tissue expressions were reported in GenBank/dbEST (tissue, stage, pathological or normal).The protein FAM216A was expressing most abundant in testis of about "69.0"-given by the Gene card database . The FAM216A protein acts as a co-expressor in protein-protein interaction network plot analysis as studied by the Cancer Cell Metabolism Gene Database.FAM216A protein is co-expressed in COAD- and BRCA- normal cells but its expression was reduced significantly in COAD- and BRCA- tumor cells (Figures 9). But the condition of expression of FAM216A protein occurs vice versa in case of LUAD –normal and tumor cells. The rate of expression was given here in different types of cancer in ascending order such as Breast cancer < Lymphoma < Urothelial cancer < Thyroid cancer < Pancreatic cancer.

Discussion:

The study identified a functional gene (C12ORF24) on human chromosome 12 from so called "junk DNA" (ORFs) and characterized the cognate protein using several high end Bioinformatic tools. Our findings on this interesting gene triggered the serendipitous identification of its structural protein (FAM216A), and its copious presence in specific two tissues. Its conspicuous existence in brain and (male) testis strongly suggests some specific neuronal and hormonal functions respectively. Existence of FAM216A protein from its *C12ORF24* gene in female ovaries is not yet confirmed. FAM216A is of particular interest because it is predominantly expressed only in these two tissues and needs further studies to assess its functional aspects. Further our pertinent observations also raise pivotal questions about the function of FAM216A protein in these two above tissues and how it is regulated when these tissues located in different part of the human body which is knotted with complex cognitive and hormonal functions. Incidence of insignificant amounts of this protein in other tissues also poses an interesting question of how this protein at the genome level is suppressed, probably by methylation of its promoter and enhancer elements certainly deserves further research. Epigenetic aspects also certainly elicit a central role in these two tissues for its regulated expression. Cross-talk between these two tissues via FAM216A protein as a circulating "hormone-like" function also warrants further analysis.

Although native FAM216A protein may exist as multiple forms and may also be modified by post-translational modifications, such as phosphorylation, ubiquitinylation and acetylation, unequivocal demonstration of these aspects in test tubes is essential at a later date. This suggests that cellular regulation of FAM216A is very complex since this protein exists in two tissues abundantly (brain and testis); this dictates that it is tailored to carry-out tissue specific physiological functions. Physiological cross-talk between brain and testis must exert a specific cognitive function and it must also exist in blood as a circulating protein with targeted function in brain/testes biochemical communication.Hence its first unequivocal identification from our laboratory using high end bioinformatics tools in several species, we would like to call this FAM216A protein as "*Moodin*" because brain may control testis or vice versa.

The genomic organization of *C12ORF24* gene encoding FAM216A was confirmed by several criteria: (i) we have identified the exact exon/intron structures manually and compared human and other rodent specific mRNA sequences (rat and mouse); (ii) Exon/intron organization fit perfectly between these three species almost verbatim and *C12ORF24* gene is not just "junk

DNA" but a fully functional unit to generate protein producing hereditable entity in humans; (iii) Comparison with mRNA and genomic sequences with other species indicate similar exon/intron organization following strict GT/AG splicing rules; (iv) The highly conserved third exon encoding amino acids in all these three species must have a specific enzymatic function as these residues are fully conserved and guarded by a large intron in their genomes. All these aspects suggest that these amino acids encoded from third exon must participate in the formation of the architecture of the active site and have some functional involvement in the enzymatic process. These aspects certainly need imminent studies .Detection of similar pI and molecular weights of the protein between species as carried out by "*Expasy-Compute pI*" indicates that FAM216A is highly alkaline in nature and possess similar molecular weights, except for human (as having extra amino acids derived from unique first exon). All these protein characteristics of FAM216A were extensively searched and compared with several other EST data base derived information for many species including primates encompassing C12ORF24 in their 12th chromosome akin to humans.

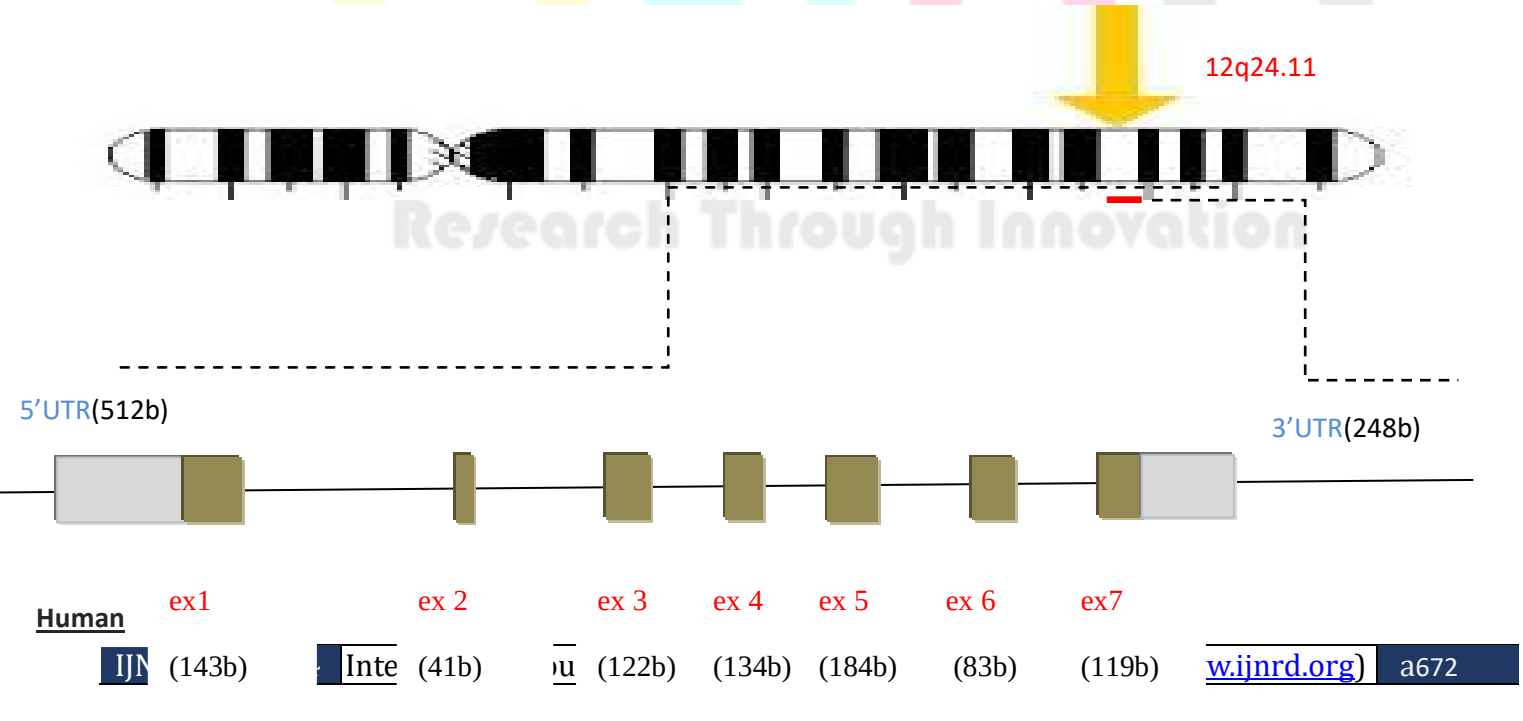
Further, human FAM216A was identified by two specific antibodies (high quality Sigma Company "Prestige" antibodies) by the International Human Global Proteome. These works confirm its positive occurrence in these two tissues (brain and testis). Certainly through evolution acquiring of an extra exon in human FAM216A surprisingly did not change its alkaline nature suggests its existence in alkaline milieu is primarily essential for some specific physiological function(s) for brain and testis "cross-talk". The full repertoire of human FAM216A characteristics is not fully understood at this point the future studies will undoubtedly unravel its ultimate cellular undermined secret(s). We have only identified its basic characteristics both at the DNA and protein levels using several Bioinformatic tools. Functional aspects at the cellular protein level and genomic regulation at the DNA level need to be identified in test tubes and for this work isolation of its promoter is a required bench work in future.

Human genes typically have shorter exons interspersed by large introns, a feature that appears to facilitate intron splice-site recognition through "exon-definition" model (Berget, 1995). *C12ORF24* posit as a prime candidate for this feature where almost all introns are found to be larger than exons. These genomic features of *C12ORF24* also pose a good possibility that it can generate multiple mRNAs and multiple isoforms. Intron retention (IR) can also occur with *C12ORF24* because of the failure to recognize weak splice sites flanking large introns by the "intron definition" mechanism. Probably by this selective process, human *C12ORF24* would have acquired the extra first exon retaining still its alkaline nature is intriguing. My observation on human *C12ORF24* gene may have a profound impact in understanding brain functions which are complex and in its relation to hormonal changes that occur in testes during adolescence. Especially this aspect may have a profound behavioral effect with respect to autism, a recently evolved disorder during which they become very aggressive in adolescence. This aspect needs certainly need to be analyzed. This *C12ORF24* gene definitely suggests several interesting questions for my future studies on the bench with biochemical techniques.

Figures and tables:

Table 1: This table shows the different chromosomal location of FAM216A in various species.

Name/gene ID	Description	Location	Aliases
FAM216A ID: 29902	family with sequence similarity 216 member A [Homo sapiens (human)]	Chromosome 12, NC_000012.12 (110468427..110490387)	C12orf24, HSU79274
FAM216A ID: 616613	family with sequence similarity 216 member A [*Bostaurus* (cattle)]	Chromosome 17, AC_000174.1 (56607627..56616415)	C12orf24, C17H12orf24
FAM216A ID: 100513180	family with sequence similarity 216, member A [*Susscrofa* (pig)]	Chromosome 14, NC_010456.4 (33771559..33781210)	C12orf24, C14H12orf24
FAM216A ID: 68948	family with sequence similarity 216, member A [*Mus musculus* (house mouse)]	Chromosome 5, NC_000071.6 (122364584..122371963, complement)	1500011H22Rik
C26H12orf24 ID: 477476	chromosome 26 open reading frame, human C12orf24 [*Canis lupus familiaris*(dog)]	Chromosome 26, NC_006608.2 (11310222..11319158)	RGD1310861
FAM216A ID: 288667	family with sequence similarity 216, member A [*Rattusnorvegicus* (Norway rat)]	Chromosome 12, NC_005111.4 (39679688..39688831)	
FAM216A ID: 100751529	family with sequence similarity 216, member A [*Cricetulusgriseus* (Chinese hamster)]	NW_003613692.1 1442509..1457449, complement	



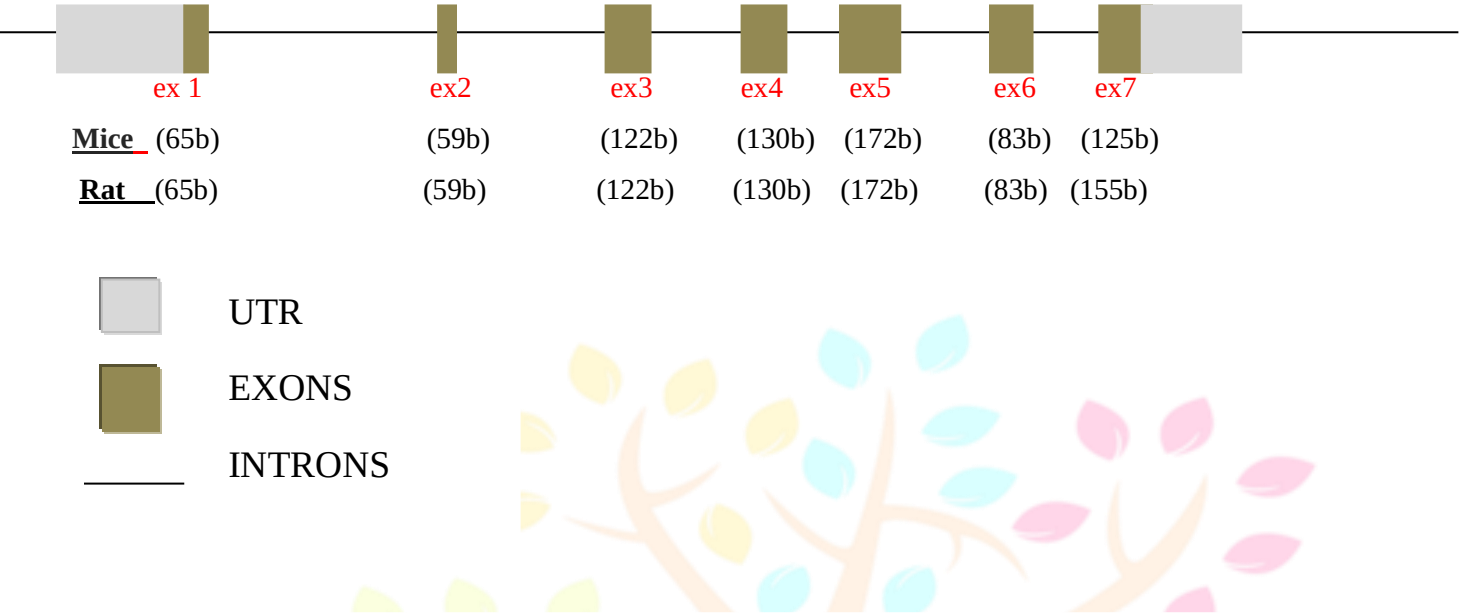


Figure 2: The gene structure of C12ORF24 (shown-arrow), at the cytogenetic location 12q24.11, representing 7 exons and 6 introns along with their 5' and 3' prime end. The number of bases in each exon (ex) and UTR are given respectively in parenthesis for three species Human, mice and rat.

```
>NP_037432.2 protein FAM216A [Homo sapiens]
MLGQLLPHTARGLGAAEMPGQGPGSDWTERSSSAEPPAVAGTEGGGGGSAGYSCYQNSKGSDRIKDGYKV
NSHIAKLQELWKTPQNQTIHLSKSMMEASFFKHPDLTTGQKRYLCSIAKIYNANYLKMLMKRQYMHVLQH
SSQKPGVLTHHRSRLSSRYSQKQHYPCTTWRHQLEREDSGSSDIAAASAPEMLIQHSLWRPVRNKEGIKT
GYASKTRCKSLKIFRRPRKLFMQTVSSDDSESHMSEEKKEEDLLNNFMQSMSIEEQGEHLMLT

>NP_001008291.1 protein FAM216A [Rattus norvegicus]
MPSRCPGVAGPPALARTEGSEGSAGQSYHQNSKGTGEQHKAERIKEGHRMSSHTAKLQELWRTTQIQTIH
IPKSMTDASFLKHPELTSGQKRYLCSIAKICNSSYLRTLMKRQYMHLFHHGLQKPGVLTHHRSHISSRYS
QKQHSPCTTWRHHLEREDSLGIAAEAPEMIIHALWRPLRHKEGLKIGYASKTRCKSLKIFRRPGRLFLLP
VSSKDYQPCMNDETKEEDLLNECMQSMSIQEQGSSHASLTVSCPTPSSAIP

>NP_081159.1 protein FAM216A [Mus musculus]
MPSRWPGVAGPPALARTEGGEGSAGHSYPQNSKGTGEQHKADRIKEGHRVYAHIAKLQELWKTTQIQTIH
IPKSMTDASFLKHPELTLGQKRYLCSVAKICNSSYLRTLMKRQYMHIFHHGSQKTGVLTHHRGHMSSRYS
QKQHSPCTAWRHHLEREDSLSIAAGAPEMIIHSLWRPLRHKEGLKIGYASKTRCKSLKIFRRPGRLFLLP
VPSNDSQSCPSEETQEEDLLNKCMQSMSIQEQGPAHASLTV
```

Figure 3: The FASTA format of protein sequence FAM216A in three different species human, rat and mice were shown and the extra peptide sequence in the human was highlighted in green.

Figure 4: The image **"A"** shows the colorfull three dimensional structure of protein FAM216A of the species *Homo sapien* and the image **"B"** shows the simple 3D structure with single colour coated on the structure.The 3D structurural domains of the protein FAM216A**.**The domain 1 containing four alpha helixs (**D1**),domain 2 (**D2**),domain 3 contains 2 alpha helixs (**D3**)and the domain 4 contains the beta sheets(**D4**) of Protein FAM216A for

human.

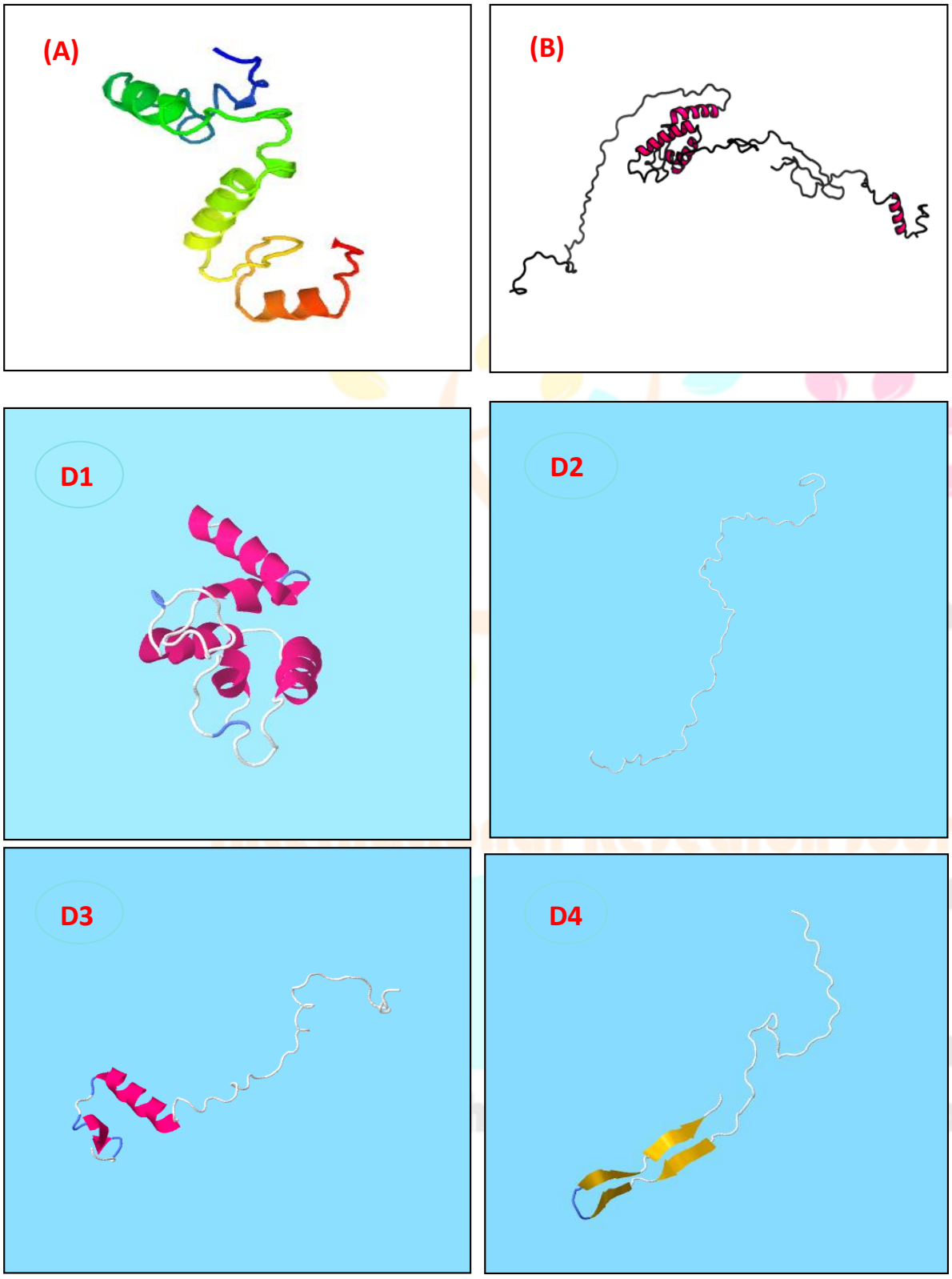


Figure 5: The image got from Prosite scanning (PSS) by placing the human FAM216A protein sequence, here the blue arrow represents the disulphide bond –S-S- and the black arrow represents the active site on the domain.

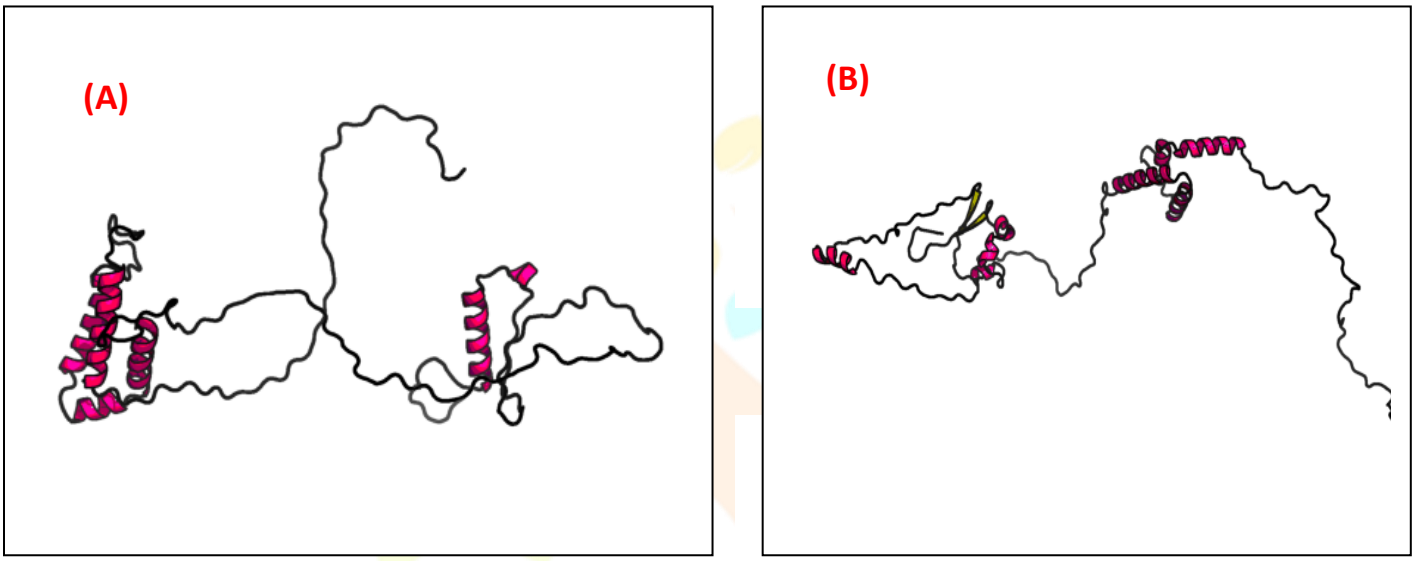
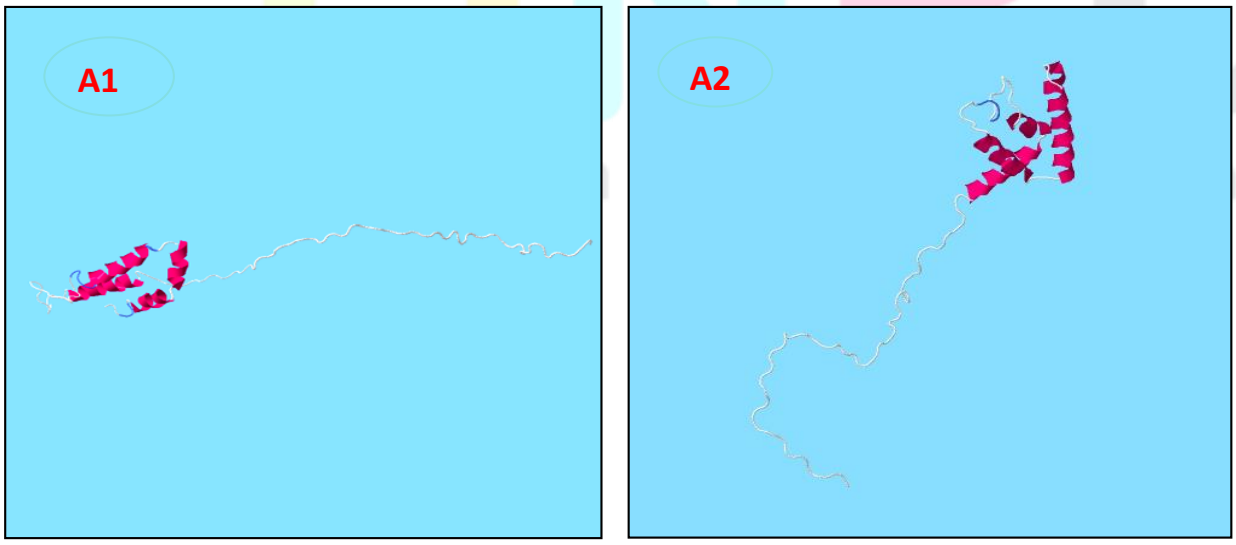
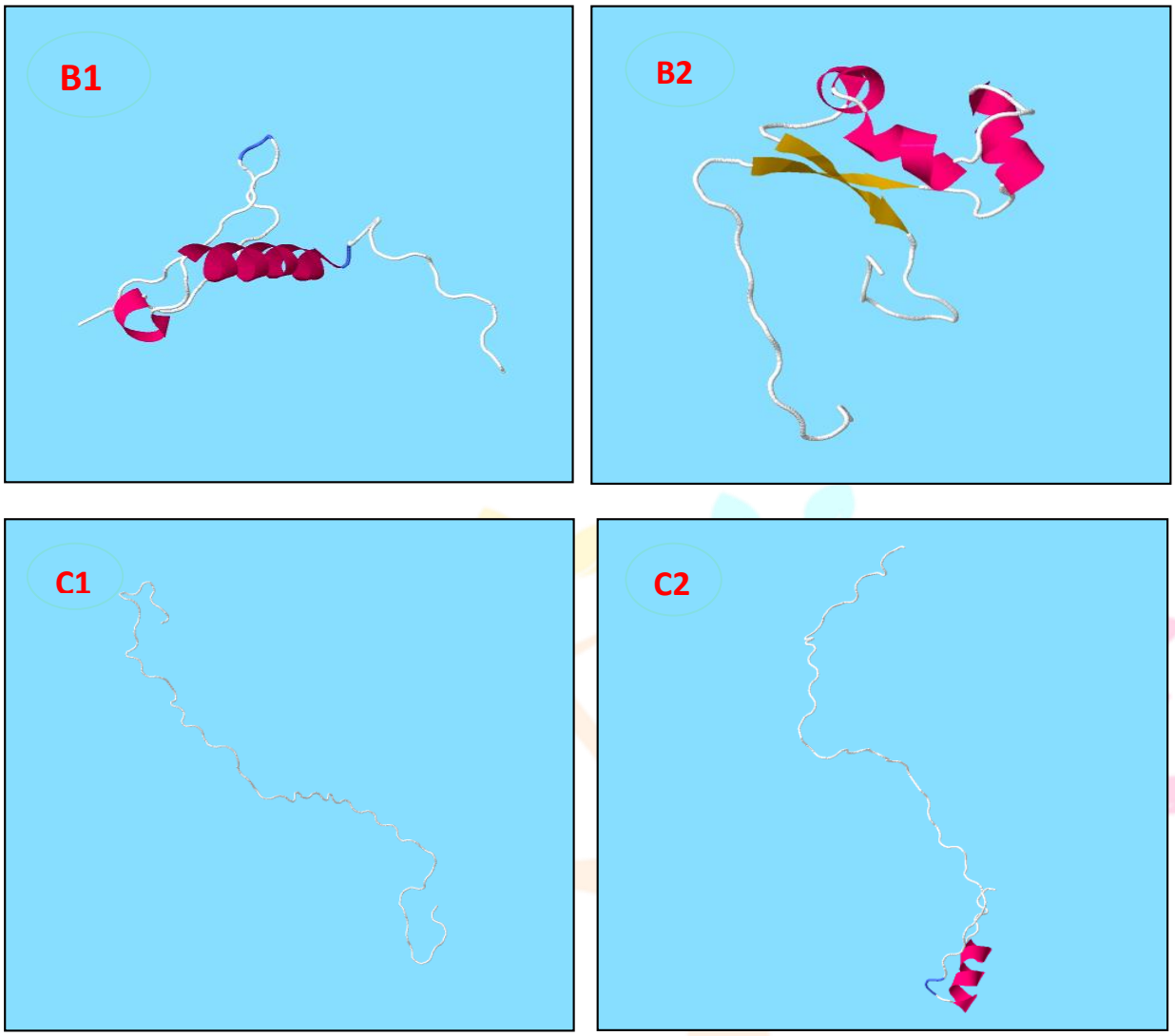


Figure 6: The images shows the three dimensional structure of protein FAM216A obtained from the Raptor X programme software for the species *Mus musculus*(**A**) and *Rattus norvegicus*(**B**).The domains for the species are shown on next page.*Mus musculus* has three domain such as domain 1(A1),domain 2(B1) and domain 3(C1).*Rattus norvegicus* has three domains such as domain 1(A2), domain 2(B2) and domain (C2).




B1
B2
C1
C2

1	M	0.061	0.123	0.816	C	72	S	0.505	0.292	0.203	H	143	Q	0.177	0.228	0.594	C	213	A	0.190	0.402	0.408	C
2	L	0.069	0.129	0.802	C	73	H	0.557	0.257	0.186	H	144	K	0.202	0.262	0.535	C	214	S	0.162	0.292	0.545	C
3	G	0.077	0.139	0.784	C	74	I	0.683	0.181	0.136	H	145	P	0.208	0.286	0.506	C	215	K	0.264	0.283	0.453	C
4	Q	0.092	0.210	0.699	C	75	A	0.655	0.175	0.170	H	146	G	0.195	0.357	0.448	C	216	T	0.337	0.291	0.373	C
5	L	0.159	0.202	0.639	C	76	K	0.706	0.165	0.129	H	147	V	0.234	0.539	0.227	E	217	R	0.339	0.242	0.419	C
6	L	0.167	0.206	0.627	C	77	L	0.737	0.170	0.093	H	148	L	0.243	0.585	0.172	E	218	C	0.342	0.216	0.442	C
7	P	0.420	0.154	0.426	C	78	Q	0.562	0.182	0.256	H	149	T	0.232	0.468	0.301	E	219	K	0.440	0.246	0.314	H
8	H	0.414	0.189	0.397	H	79	E	0.575	0.175	0.250	H	150	H	0.172	0.440	0.388	E	220	S	0.528	0.229	0.243	H
9	T	0.390	0.202	0.408	H	80	L	0.531	0.157	0.313	H	151	H	0.140	0.308	0.552	C	221	L	0.474	0.269	0.257	H
10	A	0.503	0.176	0.321	H	81	W	0.514	0.180	0.306	H	152	R	0.129	0.214	0.656	C	222	K	0.375	0.391	0.235	H
11	R	0.443	0.193	0.365	H	82	K	0.363	0.204	0.433	C	153	S	0.124	0.184	0.691	C	223	I	0.331	0.467	0.202	H
12	G	0.449	0.230	0.321	H	83	T	0.311	0.169	0.521	C	154	R	0.167	0.314	0.519	C	224	F	0.299	0.372	0.329	H
13	L	0.355	0.268	0.377	C	84	P	0.301	0.187	0.512	C	155	L	0.180	0.346	0.475	C	225	R	0.241	0.307	0.452	C
14	G	0.279	0.323	0.398	C	85	Q	0.272	0.191	0.537	C	156	S	0.207	0.274	0.518	C	226	R	0.195	0.184	0.621	C
15	A	0.357	0.257	0.386	C	86	N	0.234	0.244	0.522	C	157	S	0.262	0.255	0.483	C	227	P	0.311	0.144	0.545	C
16	A	0.257	0.340	0.404	C	87	Q	0.246	0.353	0.401	E	158	R	0.347	0.249	0.404	C	228	R	0.260	0.172	0.568	C
17	E	0.157	0.333	0.510	C	88	T	0.193	0.555	0.252	E	159	Y	0.382	0.249	0.369	C	229	K	0.284	0.257	0.459	C
18	M	0.074	0.160	0.766	C	89	I	0.215	0.506	0.279	E	160	S	0.332	0.219	0.449	C	230	L	0.210	0.517	0.273	E
19	P	0.069	0.138	0.793	C	90	H	0.191	0.516	0.293	E	161	Q	0.266	0.218	0.516	C	231	F	0.144	0.686	0.170	E
20	G	0.064	0.171	0.765	C	91	L	0.196	0.336	0.469	C	162	K	0.164	0.263	0.573	C	232	M	0.123	0.694	0.183	E
21	Q	0.066	0.246	0.688	C	92	S	0.244	0.196	0.560	C	163	Q	0.123	0.239	0.638	C	233	Q	0.099	0.585	0.316	E
22	G	0.065	0.239	0.696	C	93	K	0.280	0.215	0.505	C	164	H	0.092	0.174	0.733	C	234	T	0.115	0.401	0.484	C
23	P	0.071	0.257	0.672	C	94	S	0.292	0.219	0.489	C	165	Y	0.084	0.157	0.760	C	235	V	0.116	0.267	0.617	C
24	G	0.073	0.254	0.674	C	95	M	0.304	0.195	0.501	C	166	P	0.104	0.172	0.724	C	236	S	0.094	0.191	0.714	C
25	S	0.080	0.173	0.747	C	96	M	0.325	0.206	0.469	C	167	C	0.117	0.265	0.618	C	237	S	0.094	0.162	0.744	C
26	D	0.098	0.263	0.639	C	97	E	0.399	0.192	0.409	H	168	T	0.175	0.382	0.443	C	238	D	0.104	0.171	0.725	C
27	W	0.125	0.413	0.462	C	98	A	0.445	0.205	0.350	H	169	T	0.286	0.396	0.318	H	239	D	0.100	0.166	0.735	C
28	T	0.140	0.566	0.294	C	99	S	0.440	0.233	0.327	H	170	W	0.348	0.439	0.213	H	240	S	0.125	0.217	0.658	C
29	E	0.192	0.483	0.325	C	100	F	0.497	0.241	0.263	H	171	R	0.479	0.299	0.222	H	241	E	0.175	0.229	0.596	C
30	R	0.175	0.418	0.407	C	101	F	0.379	0.249	0.372	H	172	H	0.449	0.274	0.276	H	242	S	0.188	0.291	0.522	C
31	S	0.144	0.270	0.586	C	102	K	0.352	0.239	0.409	C	173	Q	0.443	0.281	0.277	H	243	H	0.207	0.256	0.537	C
32	S	0.144	0.217	0.639	C	103	H	0.168	0.232	0.600	C	174	L	0.486	0.228	0.286	H	244	M	0.250	0.210	0.540	C
33	S	0.081	0.246	0.674	C	104	P	0.192	0.182	0.626	C	175	E	0.409	0.245	0.346	H	245	S	0.291	0.191	0.517	C
34	A	0.084	0.191	0.724	C	105	D	0.363	0.179	0.458	C	176	R	0.327	0.204	0.469	C	246	E	0.378	0.198	0.423	C
35	E	0.074	0.155	0.771	C	106	L	0.387	0.178	0.434	C	177	E	0.335	0.197	0.468	C	247	E	0.433	0.164	0.403	H
36	P	0.090	0.139	0.771	C	107	T	0.349	0.202	0.449	C	178	D	0.315	0.163	0.522	C	248	K	0.493	0.161	0.346	H
37	P	0.196	0.157	0.647	C	108	T	0.438	0.174	0.389	C	179	S	0.397	0.195	0.408	C	249	K	0.605	0.154	0.241	H
38	A	0.312	0.254	0.434	C	109	G	0.325	0.158	0.517	C	180	G	0.486	0.190	0.324	C	250	E	0.748	0.144	0.108	H
39	V	0.339	0.371	0.290	C	110	Q	0.337	0.197	0.466	C	181	S	0.428	0.268	0.304	C	251	E	0.730	0.149	0.121	H
40	A	0.350	0.394	0.256	C	111	K	0.415	0.241	0.344	C	182	S	0.268	0.243	0.489	C	252	D	0.718	0.142	0.140	H
41	G	0.296	0.439	0.265	C	112	R	0.425	0.304	0.271	C	183	D	0.482	0.230	0.287	C	253	L	0.725	0.157	0.118	H
42	T	0.242	0.367	0.391	C	113	Y	0.410	0.433	0.157	H	184	I	0.431	0.364	0.204	C	254	L	0.683	0.168	0.149	H
43	E	0.241	0.250	0.509	C	114	L	0.471	0.370	0.159	H	185	A	0.525	0.287	0.187	C	255	N	0.652	0.157	0.191	H
44	G	0.086	0.233	0.682	C	115	C	0.468	0.367	0.165	H	186	A	0.576	0.226	0.198	C	256	N	0.600	0.179	0.221	H
45	G	0.143	0.174	0.683	C	116	S	0.464	0.303	0.234	H	187	A	0.485	0.169	0.346	C	257	F	0.525	0.187	0.288	H
46	G	0.127	0.177	0.696	C	117	I	0.444	0.312	0.244	H	188	S	0.310	0.141	0.550	C	258	M	0.448	0.193	0.359	H
47	G	0.142	0.258	0.600	C	118	A	0.371	0.361	0.268	H	189	A	0.265	0.135	0.600	C	259	Q	0.445	0.204	0.351	H
48	G	0.106	0.250	0.644	C	119	K	0.318	0.399	0.284	H	190	P	0.407	0.146	0.447	C	260	S	0.469	0.178	0.353	H
49	S	0.110	0.274	0.617	C	120	I	0.326	0.389	0.286	H	191	E	0.572	0.179	0.248	H	261	M	0.458	0.257	0.285	H
50	A	0.127	0.270	0.603	C	121	Y	0.237	0.317	0.446	C	192	M	0.554	0.270	0.176	H	262	S	0.432	0.239	0.328	H
51	G	0.121	0.305	0.574	C	122	N	0.241	0.197	0.562	C	193	L	0.553	0.294	0.154	H	263	I	0.588	0.197	0.215	H
52	Y	0.120	0.362	0.519	C	123	A	0.295	0.149	0.556	C	194	I	0.521	0.340	0.138	H	264	E	0.667	0.176	0.157	H
53	S	0.138	0.433	0.429	C	124	N	0.295	0.153	0.552	C	195	Q	0.543	0.247	0.210	H	265	E	0.627	0.180	0.192	H
54	C	0.138	0.503	0.359	C	125	Y	0.441	0.178	0.381	H	196	H	0.502	0.279	0.219	H	266	Q	0.561	0.161	0.278	H
55	Y	0.132	0.403	0.465	C	126	L	0.564	0.185	0.251	H	197	S	0.470	0.252	0.277	H	267	G	0.421	0.155	0.424	C
56	Q	0.134	0.313	0.554	C	127	K	0.683	0.189	0.128	H	198	L	0.354	0.225	0.421	C	268	E	0.389	0.176	0.435	C
57	N	0.123	0.246	0.631	C	128	M	0.735	0.185	0.080	H	199	W	0.263	0.244	0.493	C	269	H	0.342	0.212	0.446	C
58	S	0.143	0.169	0.688	C	129	L	0.749	0.159	0.092	H	200	R	0.279	0.229	0.492	C	270	L	0.241	0.237	0.522	C
59	K	0.141	0.178	0.682	C	130	M	0.733	0.152	0.115	H	201	P	0.394	0.227	0.380	C	271	M	0.103	0.203	0.694	C
60	G	0.122	0.169	0.708	C	131	K	0.670	0.160	0.170	H	202	V	0.431	0.228	0.340	C	272	L	0.078	0.164	0.758	C
61	S	0.204	0.275	0.522	C	132	R	0.693	0.154	0.154	H	203	R	0.425	0.187	0.388	C	273	T	0.062	0.130	0.808	C
62	D	0.238	0.379	0.384	C	133	Q	0.558	0.227	0.216	H	204	N	0.398	0.159	0.443	C						
63	R	0.315	0.459	0.225	C	134	Y	0.502	0.308	0.190	H	205	K	0.408	0.190	0.402	C						
64	I	0.391	0.350	0.259	C	135	M	0.484	0.333	0.183	H	206	E	0.318	0.212	0.470	C						
65	K	0.416	0.295	0.289	C	136	H	0.381	0.409	0.211	H	207	G	0.255	0.253	0.491	C						
66	D	0.381	0.243	0.376	C	137	V	0.317	0.433	0.250	H	208	I	0.279	0.294	0.427	C						
67	G	0.257	0.210	0.533	C	138	L	0.310	0.314	0.377	C	209	K	0.282	0.455	0.263	E						
68	Y	0.271	0.297	0.432	C	139	Q	0.251	0.281	0.468	C	210	T	0.211	0.517	0.272	E						
69	K	0.261	0.439	0.301	H	140	H	0.227	0.235	0.538	C	211	G	0.199	0.409	0.392	E						
70	V	0.281	0.493	0.225	H	141	S	0.120	0.188	0.692	C	212	Y	0.209	0.390	0.401	C						
71	N	0.383	0.400	0.217	H	142	S	0.218	0.188	0.594	C												

Figure 7: The output of the GOR method showing the participation of individual amino acid in forming the structures for the protein such as "C"-coil, "E"-sheet and "H"-helix. Column information:1) Sequence index, 2) Amino acid type, 3) Helix probability,4) Sheet probability, 5) Coil probability and 6) GOR V prediction.

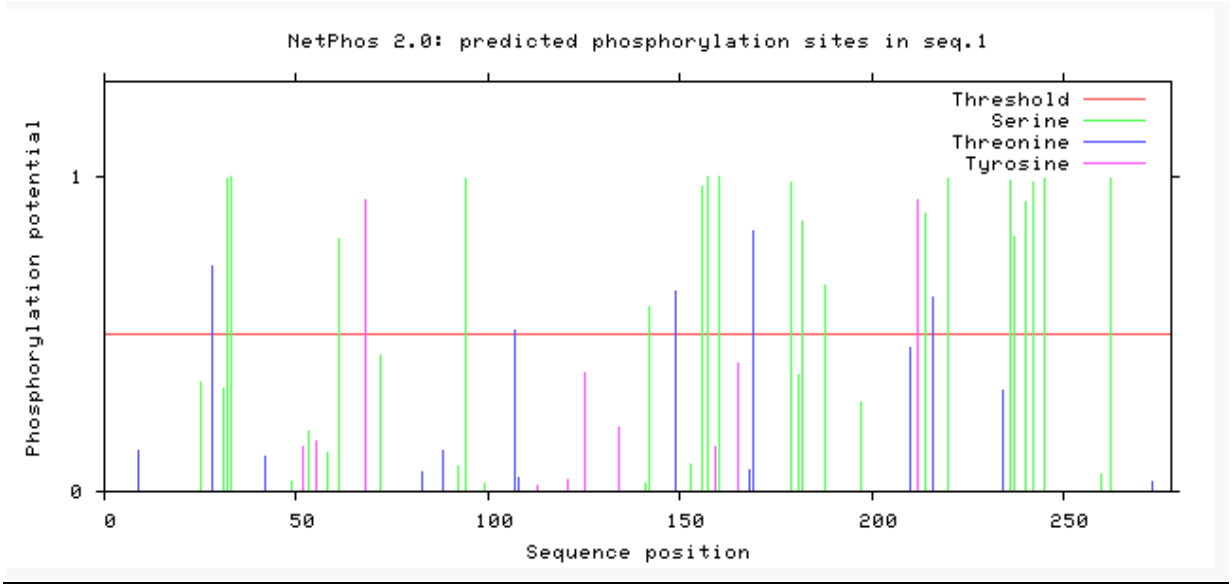


Figure 8: Graphical representation of my query protein FAM216A obtained from the Expasy-Netphos for the amino acid shown at top corner.

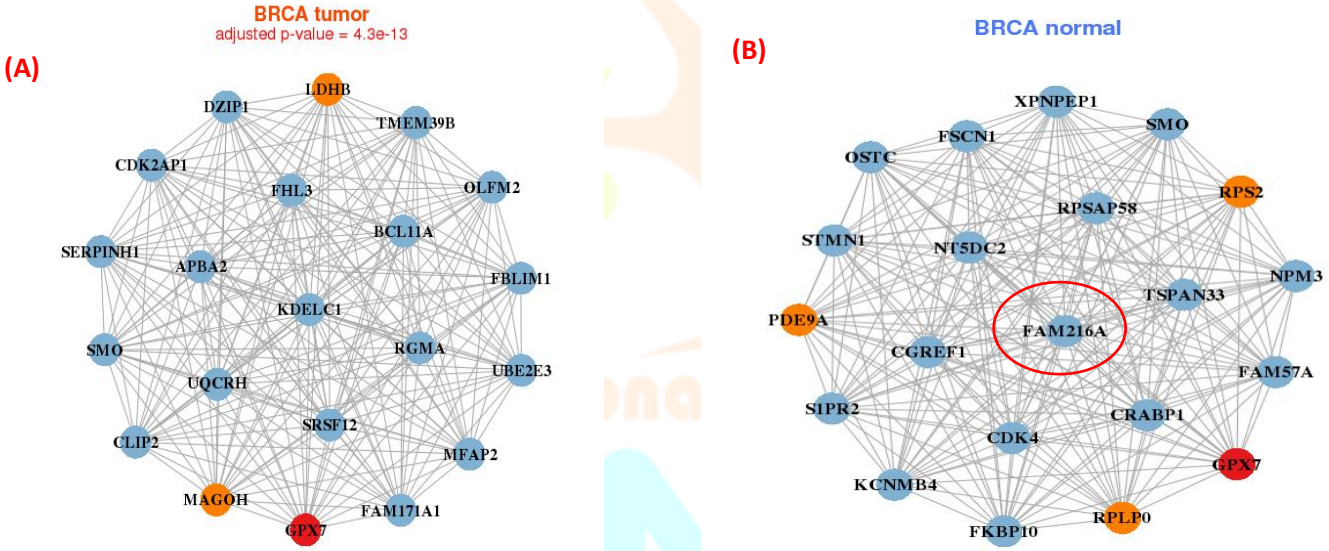


Figure 9: The co-expression of the protein FAM216A in the BRCA normal (A) and the abbsence of FAM216A expression in the BRCA tumor cell (B) obtained from the cancer cell metabolism gene database.

Reference:

1. Lander, E. S., Linton, L.M. et al., (2001) Human Genome, Nature 409, 860-921.

2. Gerhard D.S. (2004) The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC) Genome Res., 14:2121

3. M. Burset, I. A. Seledtsov, and V. V. Solovyev ; Analysis of canonical and non-canonical splice sites in mammalian genomes: Nucleic Acids Res. 2000 Nov 1; 28(21): 4364–4375

4. hORFeome 3.1:a resource of human open reading frames covering over 10,000 human genes.Genomics 2007:89:307-15

5. Paik, Y. K.; Jeong, S. K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Cho, S. Y.; Lee, H. J.; Na, K.; Choi, E. Y.; Yan, F.; Zhang, F.; Zhang, Y.; Snyder, M.; Cheng, Y.; Chen, R.; Marko-Varga, G.; Deutsch, E. W.; Kim, H.; Kwon, J. Y.; Aebersold, R.; Bairoch, A.; Taylor, A. D.; Kim, K. Y.; Lee, E. Y.; Hochstrasser, D.; Legrain, P.; Hancock, W. S. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. Nat. Biotechnol. 2012, 30 (3), 221−3.

6. Lane Lydie (2012). "neXtProt: a knowledge platform for human proteins". Nucleic Acids Research 40 (Database issue): D76–D83.

7. Srikanth Srinivas Manda, Raja Sekhar Nirujogi, Sneha Maria Pinto, Min-Sik Kim, Keshava K. Datta,Ravi Sirdeshmukh, T. S. Keshava Prasad, Visith Thongboonkerd, Akhilesh Pandey, and Harsha Gowda; Identification and Characterization of Proteins Encoded by Chromosome 12 as Part of Chromosome-centric Human Proteome Project;J. Proteome Res., 2014, 13 (7), pp 3166–3177

8. Keshava Prasad, T. S.; Goel, R.; Kandasamy, K.; Keerthikumar, S.; Kumar, S.; Mathivanan, S.; Telikicherla, D.; Raju, R.; Shafreen, B.; Venugopal, A.; Balakrishnan, L.; Marimuthu, A.; Banerjee, S.; Somanathan, D. S.; Sebastian, A.; Rani, S.; Ray, S.; Harrys Kishore, C. J.; Kanth, S.; Ahmed, M.; Kashyap, M. K.; Mohmood, R.; Ramachandra, Y. L.; Krishna, V.; Rahiman, B. A.; Mohan, S.; Ranganathan, P.; Ramabadran, S.; Chaerkady, R.; Pandey, A. Human Protein Reference Database–2009 update. Nucleic Acids Res. 2009, 37 (Database issue), D767−72.

9. ccmGDB: a database for cancer cell metabolism genes , Pora Kim1, Feixiong Cheng1 ,Junfei Zhao1,Zhongming Zhao1,2,3.;1.Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN

37203, USA;2.Department of Cancer Biology, Vanderbilt University School of Medicine, Nashville, TN 37232, USA;3.Department of Psychiatry, Vanderbilt University School of Medicine, Nashville, TN 37212, USA ; Nucl. Acids Res. (2015)

10. Kitts P.A., Church D.M., Choi J., Hem V., Smith R., Tatusova T., Thibaud-Nissen F., DiCuccio M., Murphy T.D., Pruitt K.D., et al. Assembly: a resource for assembled genomes at NCBI. Nucleic Acids Res.2016 doi:10.1093/nar/gkv1226.

11. Benson D.A., Cavanaugh M., Clark K., Karsch-Mizrachi I., Lipman D.J., Ostell J., Sayers E.W. GenBank. Nucleic Acids Res. 2016 doi:10.1093/nar/gkv1276.

12. Tzong-Yi Lee, Wen-Chi Chang, Justin Bo-Kai Hsu, Tzu-Hao Chang, and Dray-Ming Shien ; GPMiner: an integrated system for mining combinatorial cis-regulatory elements in mammalian gene group. BMC Genomics. 2012; 13(Suppl 1): S3. Published online 2012 Jan 17.

13. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, Hart E, Suner MM, Landrum MJ, Aken B, Ayling S, Baertsch R, Fernandez-Banet J, Cherry JL, Curwen V, Dicuccio M, Kellis M, Lee J, Lin MF, Schuster M, Shkeda A, Amid C, Brown G, Dukhanina O, Frankish A, Hart J, Maidak BL, Mudge J, Murphy MR, Murphy T, Rajan J, Rajput B, Riddick LD, Snow C, Steward C, Webb D, Weber JA, Wilming L, Wu W, Birney E, Haussler D, Hubbard T, Ostell J, Durbin R, Lipman D; The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Res. 2009 Jul; 19(7):1316-23.

14. Kloczkowski, A., Ting, K.-L, Jernigan, R.L., Garnier, J., "Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence", Proteins, 49, 154-166, 2002.

15. Sen, T.Z., Jernigan, R.L., Garnier, J., Kloczkowski, A., "GOR V server for protein secondary structure prediction", Bioinformatics, 21(11), 2787-2788, 2005

16. Wang Z, Zhao F, Peng J, Xu J; Protein 8-class secondary structure prediction using conditional neural fields. Proteomics 2011 Oct;11(19):3786-92. doi: 10.1002/pmic.201100196. Epub 2011 Aug 31.

17. Franceschini, A et al. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. In:'Nucleic Acids Res. 2013 Jan;41(Database issue):D808-15. doi: 10.1093/nar/gks1094.Epub 2012 Nov 29.