**International Journal of Research in Chemistry and Environment**
***Vol. 2 Issue 1 January 2012(97-104)***
**ISSN 2248-9649**

***Research Paper***

# Three Novel Breast Cancer Genes: A Proteomic and Genomic Study

**Ashokan K. V.**
Department of Zoology, P. V. P. College, Kavathe Mahankal, Sangli, (MH), INDIA

**Available online at: www.ijrce.org**

**(Received 14th October 2011, Accepted 01th November 2011)**

***Abstract -*** *Breast cancer is a heterogeneous disease. Many genes are involved in breast cancer and most of them are express the cancer property by mutation. Recently three novel genes have been discovered in relation to breast cancer. The three genes-C6ORF96, C6ORF97 and C6ORF211 has not been studied widely so far. Hence, in this paper I try to annotate the physicochemical properties, and other parameters of the genes translational proteins. The physicochemical properties like isoelectric point, GRAVY, instability index, hydrophobicity, antigenicity, extinction coefficient and charges were studied. It will help to study the protein further in the cancer research. The paper also exposed the amino acid percentage, which will assists in further research in cancer proteomics. The study also concentrated on multiple sequence alignment and gene interaction with other genes of the vicinity. The gene functional interaction revealed interesting factors like C6ORF96 has no physical interacting gene related. While the c6ORF97 and C6ORF211 shows well marked physical interacting genes, hence a further study concentrating on the interacting genes along with C6ORF97 and C6ORF211 genes should be promoted to understand more about the breast cancer therapy.*

**Keywords*:*** Breast cancer, Novel gens, physicochemical parameters, MSA.

## Introduction

Breast cancer is a disease in which certain cells in the breast become abnormal and multiply without control or order to form a tumor. Breast cancer is a heterogeneous disease[1]. It is now accepted that breast cancer is not a single disease, but instead it is composed of a spectrum of tumor subtypes with distinct cellular origins, somatic changes, and etiologies. Gene expression profiling using DNA microarrays has contributed significantly to our understanding of the molecular heterogeneity of breast tumor formation, progression, and recurrence[2]. Many genes are involved in breast cancer and most of them are express the cancer property by mutation. For example, mutations in the BRCA1 and BRCA2 genes are inherited in an autosomal dominant pattern but not all people who inherit mutations in these genes will develop cancer. Variations of the BRCA1, BRCA2, CDH1, PTEN, STK11, and TP53 genes increase the risk of developing breast cancer. The AR, ATM, BARD1, BRIP1, CHEK2, DIRAS3, ERBB2, NBN, PALB2, RAD50, and RAD51 genes are associated with breast cancer. Inherited changes in several other genes, including CDH1, PTEN, STK11, and TP53, have been found to increase the risk of developing breast cancer. Of these genes, ATM and CHEK2 have the strongest evidence of being related to the risk of developing breast cancer. Somatic mutations also have been identified in breast tumors. For example, somatic mutations in the ERBB2 (also called Her-2/neu), DIRAS3, and TP53 genes have been associated with some cases of breast cancer.

Recently discovered three genes on the chromosome number six viz. C6ORF86, C6ORF97 and C6ORF112 were found to be linked to the estrogen receptor, but working separately from it. Approximately 80% of human breast carcinomas present as estrogen receptor alpha-positive (ER+ve) disease, and ER status is a critical factor in treatment decision-making. Multiple-testing corrected Spearman correlation revealed that three previously uncharacterized open reading frames (ORFs) located immediately upstream of ESR1, C6ORF96, C6ORF97, and C6ORF211 were highly correlated with ESR1. Publicly available datasets confirmed this relationship in other groups of ER+ve tumours. DNA copy number changes did not account for the correlations. The correlations were maintained in cultured cells. An ER alpha antagonist did not affect the ORFs' expression or their correlation with ESR1, suggesting their transcriptional co-activation is not directly mediated by ER alpha. siRNA inhibition of C6ORF211 suppressed proliferation in MCF7 cells, and C6ORF211 positively correlated with a proliferation metagene in tumours. In contrast, C6ORF97 expression correlated

negatively with the metagene and predicted for improved disease-free survival in a tamoxifen-treated published dataset, independently of ESR1. Study suggest that some of the biological effects previously attributed to ER could be mediated and/or modified by these co-expressed genes [3].

In the present investigation we explored the possible role of these genes in development of breast cancer in human. The gens were analyzed by various bioinformatics techniques available to date. Bioinformatics techniques are now days become inseparable tolls in the analysis of genes and proteins. Most of the predictions are enlighten the right direction to explore the unknown facts in genomics and proteomics.

## Material and Methods

The nucleotide sequence of the three genes C6ORF97, C6ORF98 and C6ORF211 dig out from NCBI data bank. The primary NCBI source accession number and the sequence length both genes and proteins are shown in the Table 1.

The gene sequence was translated into their corresponding protein sequences with the help of CLC Main workbench nucleotide analysis option. The protein sequences were subjected to various parameters like protein charge, hydrophobicity and antigenicity. The protein sequence was analyzed for physical parameters like half-life period, extinction coefficient, frequency of hydrophobic and hydrophilic residues, number of amino acids, amino acid frequencies and frequency of charged residues. Alpha helix, beta strands and regions were also analyzed. The secondary structure was nalysed by **CFSSP** - Chou & Fasman Secondary Structure Prediction Server. The Chou–Fasman methods are an empirical technique for the prediction of secondary structures in proteins, originally developed in the 1970s [4,5,6,7]. The method is based on analyses of the relative frequencies of each amino acid in alpha helices, beta sheets, and turns based on known protein structures solved with X-ray crystallography. From these frequencies a set of probability parameters were derived for the appearance of each amino acid in each secondary structure type, and these parameters are used to predict the probability that a given sequence of amino acids would form a helix, a beta strand, or a turn in a protein. The method is at most about 50–60% accurate in identifying correct secondary structures[7]. Which is significantly less accurate than the modern machine learning–based techniques.[The protein sequence of the genes was subjected to multiple sequence alignment (MSA) using CLC workbench. The possible interaction of the three genes with other genes also predicted and plotted using GeneMANIA [9].

GeneMANIA (http://www.genemania.org) is a flexible, user-friendly web interface for generating hypotheses about gene function, analyzing gene lists and prioritizing genes for functional assays. Given a query list, GeneMANIA extends the list with functionally similar genes that it identifies using available genomics and proteomics data. GeneMANIA also reports weights that indicate the predictive value of each selected data set for the query. Six organisms are currently supported (*Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus*, *Homo sapiens* and *Saccharomyces cerevisiae*) and hundreds of data sets have been collected from GEO, BioGRID, Pathway Commons and I2D, as well as organism-specific functional genomics data sets. Users can select arbitrary subsets of the data sets associated with an organism to perform their analyses and can upload their own data sets to analyze. The GeneMANIA algorithm performs as well or better than other gene function prediction methods on yeast and mouse benchmarks. The high accuracy of the GeneMANIA prediction algorithm, an intuitive user interface and large database make GeneMANIA a useful tool for any biologist.

## Results and Discussion

The secondary structure prediction (Figure1) reveals that helix is more predominant than sheets in C6ORF96, but in C6ORF97 coils and sheets are more predominant, while in C6ORF211 sheet turns and coils are negligible. A helix has an overall dipole moment caused by the aggregate effect of all the individual dipoles from the carbonyl groups of the peptide bond pointing along the helix axis. This can lead to destabilization of the helix through entropic effect. The predominant helices in C6ORF96 and 211 shows that it is more stable than C6ORF97. The higher-level association of β sheets has been implicated in formation of the protein aggregates and fibrils observed in many human diseases, notably the amyloidosis such as Alzheimer's disease. Sheet pattern is observed more in C6ORF97, a probable role in the formation of breast cancer tumor than C6ORF96 and 211. The presence of low level turns in the secondary structure in all the three protein support the prediction of more sheet and helices in the protein.

The MSA (Figure 2) shows that the proteins are more conserved at the middle and 3` end and close to the 5`. This prediction proved that the protein is more stable and not influence easy mutagens. The first 10-50 amino acids showed zero consensus sequences.

The three ORF sequence of protein shows relation to many genes (Figure 3)[9]. The physical interaction is shown by C6ORF96 are genes- LRRC-40 (Leucine rich repeat containing40), PRPF3 (Pre-mRNA processing factor 3) and SUCLG1 (Succinate-CoA Ligase, alpha subunit), PSMB1 (Proteosome subunit) and PHP (prohibitin). The physical interaction is shown by C6ORF211 are genes- and SAP18 (Sn3A-associuated protein), DARS (Aspartyle-tTRNA synthetase) and PCMT1 (Protein-L isoapsertate), and predicted interaction of SNX3 (Sorting nexin3) and CCNC (CyclinC). The C6ORF97 showed any physical interacting genes.

Histone acetylation plays a key role in the regulation of eukaryotic gene expression. Histone acetylation and deacetylation are catalyzed by multisubunit complexes. The protein encoded by this gene is a component of the histone deacetylase complex, which includes SIN3, SAP30, HDAC1, HDAC2, RbAp46, RbAp48, and other polypeptides. This protein directly interacts with SIN3 and enhances SIN3-mediated transcriptional repression when tethered to the promoter. A pseudo gene has been identified on chromosome 2. The interaction of these genes should be studied in detail to get clear picture of the three novel genes involved in the breast cancer.

The physic-chemical parameters analysis of the three genes shows that $p^I$ is minimum in C6ORF96 and high in C6ORF211 (Table 2). The isoelectric point (pI) of a protein molecule is the pH at which there is no electric charge on that protein. It is often the point of lowest solubility for the protein, probably because it is the point at which there are less intermolecular repulsions so that molecules tend to form aggregates. Recently, a significant relationship between the theoretical pI of a protein and the difference between the reported pI and pH for successfully crystallized proteins was also established [10]. The isoelectric point of a protein can be estimated by adding the number of positively charged residues (i.e., protonated lysine, arginine and histidine), minus the number of negatively charged residues (deprotonated tyrosine, cysteine, glutamate and aspartate), plus the number of protonated amino-termini, minus the number of deprotonated carboxyl-termini. This calculation does not take into account any ionization perturbations incurred through electrostatic interactions, which can be very significant. The calculated $p^I$ shows that the electrophoretic separation of C6ORF96 and C6ORF97 proteins will show migration zone towards the acidic or (-) ve electrode and C6ORF211 towards alkaline or (+) ve electrode. Aliphatic index (AI) is maximum for C6ORF96 and lower for C6ORF97 and C6ORF211 proteins. Hence the protein of C6ORF997 and 211 is more soluble in water, but C6ORF96 require organic solvent to extract. Extinction coefficient (EQ) and instability index (II) shows that C6ORF96 and C6ORF87 protein is less stable (II >40) than and C6ORF211 proteins (II<40). Generally, stable proteins were found to have instability indices smaller than 40, whereas unstable proteins had instability indices larger than 40 [11] .This measure cannot take into account higher-order properties that also affect the stability of proteins (e.g. the degree of cross-linking), hence exceptions to this threshold are likely to occur. All the three proteins show maximum (+) ve charge at $P^H$ ~4.5. Amino acid percentage (Fig 5) shows maximum glycine, alanine, methionine seen in C6ORF97 protein, but asparagine shows maximum in C6ORF96 protein. Antigenicity property shows in all the three proteins (Figure 6), but it is more in C6ORF96 and C6ORF211 proteins. Hydrophobicity is more or less (-) ve in all the three proteins studied (Figure 7).

The comparative study of three proteins corresponding to the gene C6ORF96 and C6ORF97 and C6ORf211 shows that purification of the proteins is favorable in the acidic $P^H$ except for C6ORF211. The charge distribution (Figure 4) and II also confirm this observation.

## References


1. Cummings M.C, Chambers R., Simpson P.T., Lakhani S.R., Molecular classification of breast cancer: is it time to pack up our microscopes? Pathology. 43(1):1-8 **(2011)**
2. Perou C.M., Børresen-Dale A.L., Systems biology and genomics of breast cancer. Cold Spring Harb Perspect Biol.1;3(2) **( 2011).**
3. Dunbier A.K.,Anderson H., Ghazoui Z.,Lopez-Knowles E., Pancholi S.,Ribas R.,, Drury Z., Sidhu K.,Leary A.,Martin L., Dowsett M. *ESR1* Is Co-Expressed with Closely Adjacent Uncharacterised Genes Spanning a Breast Cancer Susceptibility Locus at 6q25.1, PLoS Genet 7(4): e1001382 **(2011).**
4. Chou P.Y. and Gerald D. Fasman., Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. *Biochemistry*. **13**(2): 211–222 **(1974).**
5. Chou P.Y and Gerald D. Fasman., Prediction of protein conformation. *Biochemistry*. **13**(2), 222–245 **(1974b).**
6. Chou P.Y and Fasman GD., "Empirical predictions of protein conformation". *Annu Rev Biochem*, 47: 251–276 **(1978).**
7. Chou P.Y and Fasman G.D., "Prediction of the secondary structure of proteins from their amino acid sequence". *Adv Enzymol Relat Areas Mol Biol*, 47: 45–148 **(1978b)**.
8. Kabsch W. and Sander C., "How good are predictions of protein secondary structure?". *FEBS Lett*, 155 (2): 179–82 **(1983).**
9. Warde-Farley D., Donaldson SL., Comes O., Zuberi K., Badrawi R., Chao P., Franz M., Grouios C., Kazi F., Lopes C.T., Maitland A., Mostafavi S., Montojo J., Shao Q., Wright G., Bader G.D., Morris Q., The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function, Nucleic Acids Res. 2010 Jul 1;38 Suppl:W214-20 **(2010).**
10. Kantardjieff K. and Rupp B.T.B., Structural bioinformatic approaches to the discovery of new antimycobacterial drugs. *Current Pharmaceutical Design*, 10(26), 3195-211 **(2004).**
11. Guruprasad K., Reddy B.V.P., Pandit M.W., Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. Prot Eng 4: 155-164 **(1990).**


**Table 1: Gene accession number**

| S. No | Gene | NCBIPrimary source | Nucleotide seq.length | Translational protein Seq.length |
|---|---|---|---|---|
| 1 | C6ORF97 | HGNC:21177 | 2160 | 720 |
| 2 | C6ORF96 | HGNC:21176 | 1340 | 448 |
| 3 | C6ORF211 | HGNC:17872 | 1320 | 440 |

**Table 2: Physico-chemical characters of breast cance gene 1:C6ORF96, 2:C6ORF211 and 3: C6ORF96**

| Gene | IP | AI | Half life In hours | EQ at 280nm | Gravy | II | Mol.Wt |
|---|---|---|---|---|---|---|---|
| **C6ORF96** | 5.5 | 91.1 | 30 | 102705 | -0.322 | 44.5 | 51213 |
| **C6ORF97** | 6.2 | 87.4 | 30 | 20565 | -0.760 | 47.4 | 83089.6 |
| **C6ORF211** | 9.36 | 70.9 | 30 | 19855 | -0.211 | 33.7 | 39771.2 |

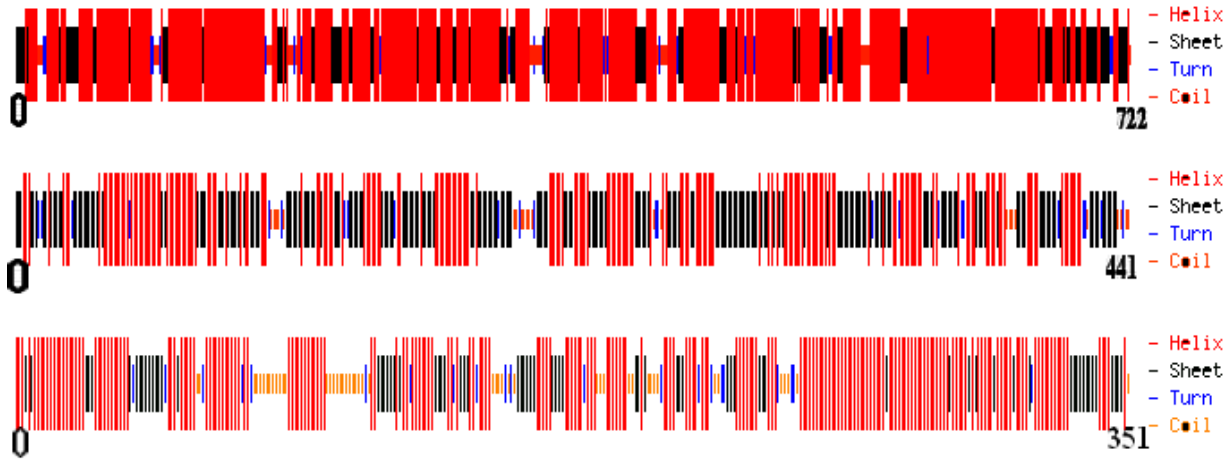


**Figure 1: Secondary structure predicted by CFSSP - Chou & Fasman Secondary Structure Prediction Server**

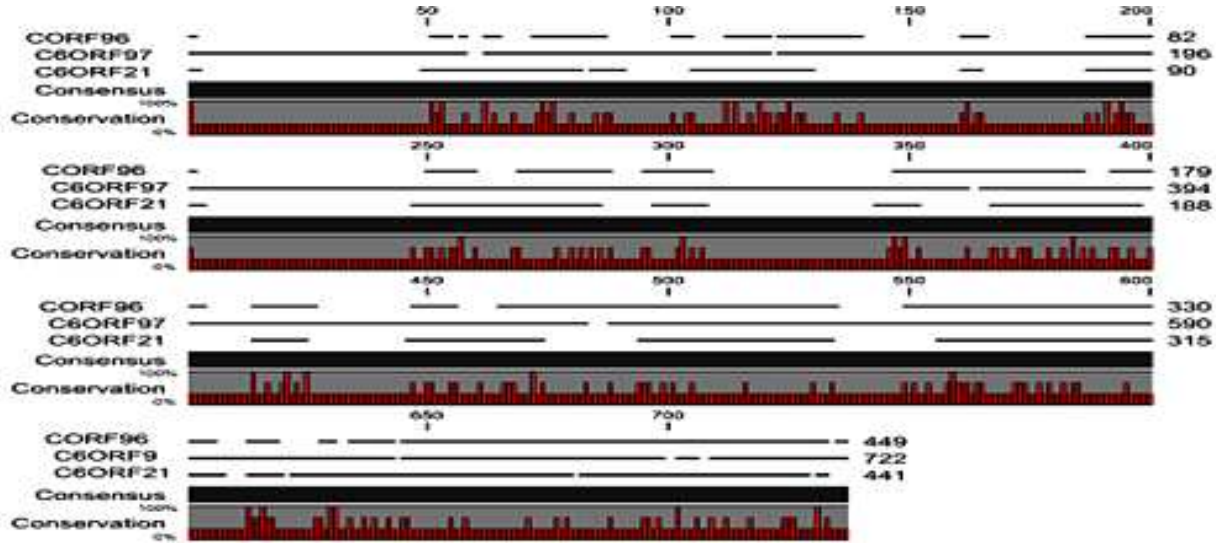


**Figure 2: Alignment of protein sequences of C6ORF96, C6ORf97 and C6ORF112 genes**

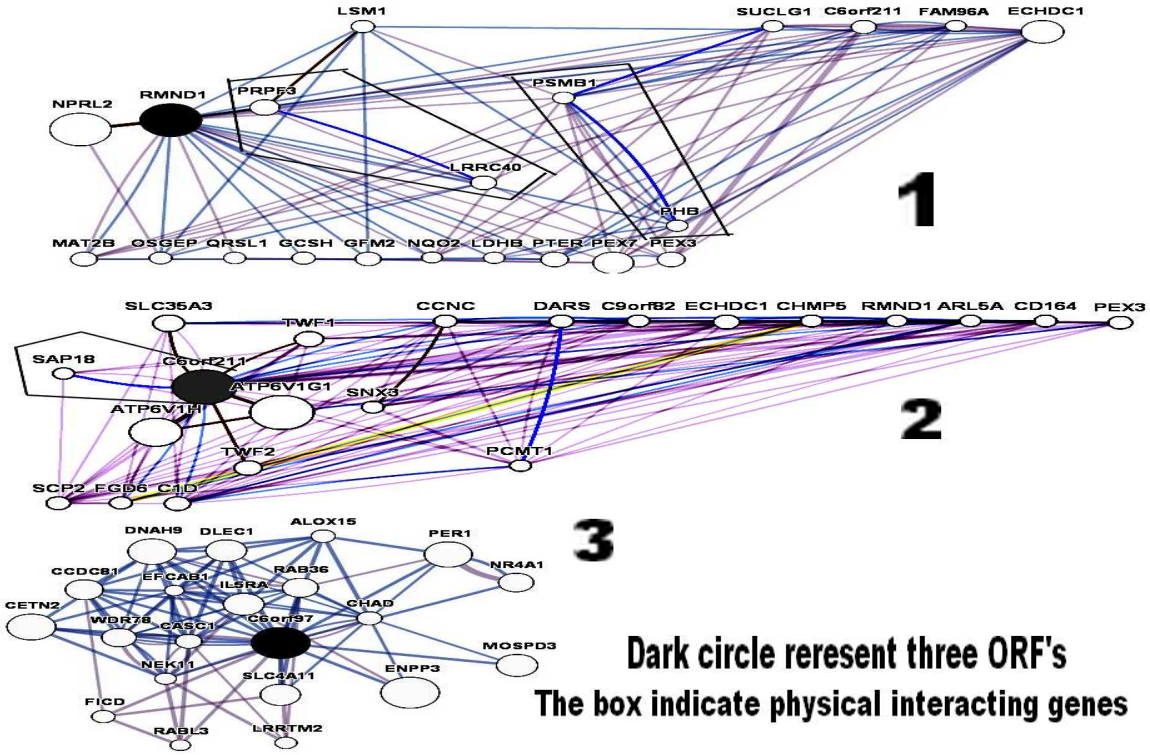


**Figure 3: Related genes of 1: C6ORF96; 2: C6ORf97; 3: C6ORF112**

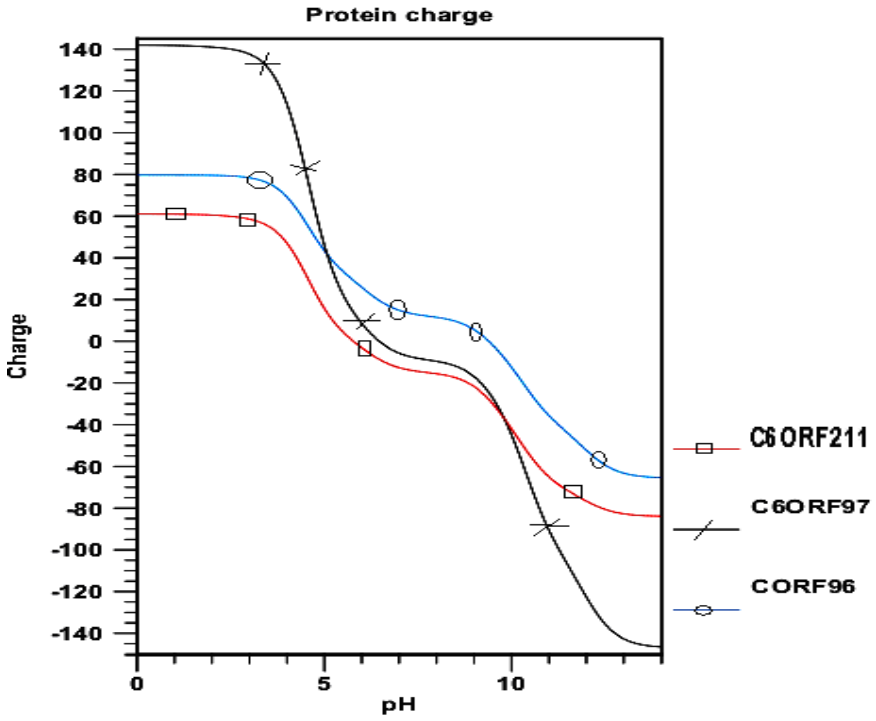


**Figure 4: Protein cahrges of C6ORF96, C6ORf97 and C6ORF112 genes translation**

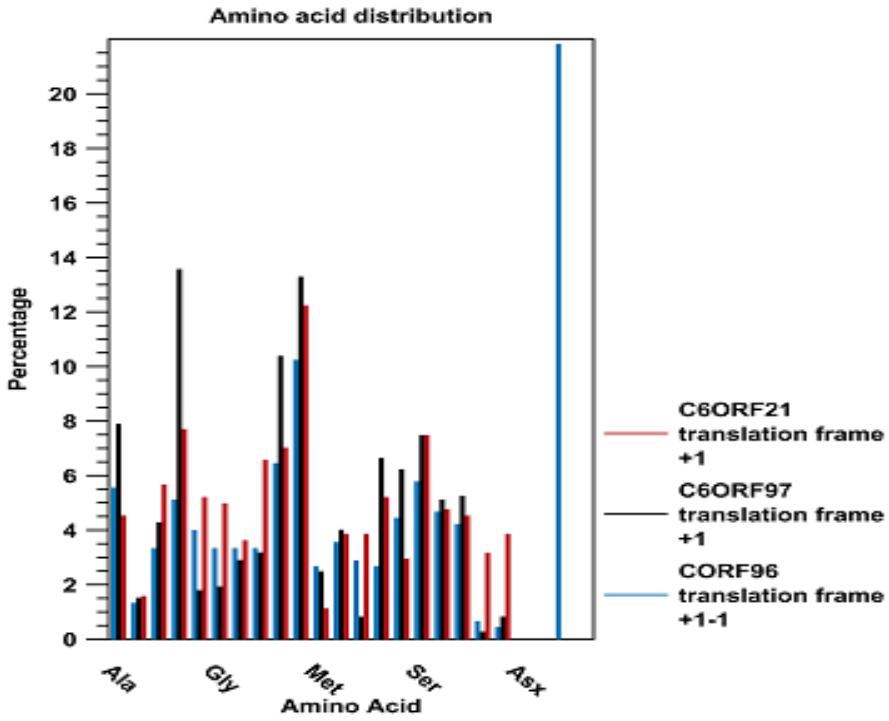


**Figure 5:Percentage of mino acid distribution in C6ORF96, C6ORf97 and C6ORF112 genes translation**

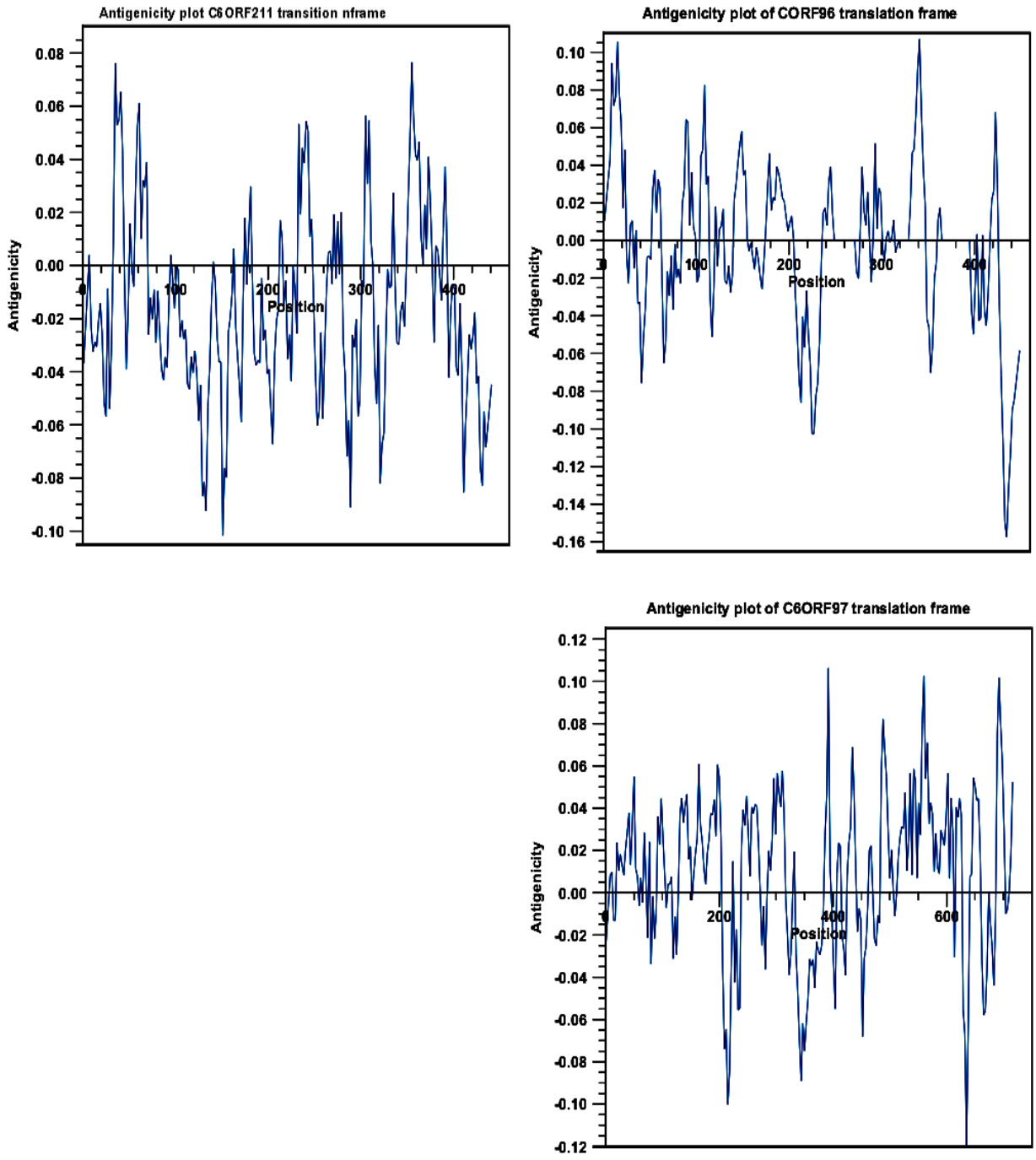


**Figure 6: Antigenicity plot of C6ORF96, C6ORf97 and C6ORF112 genes translation**

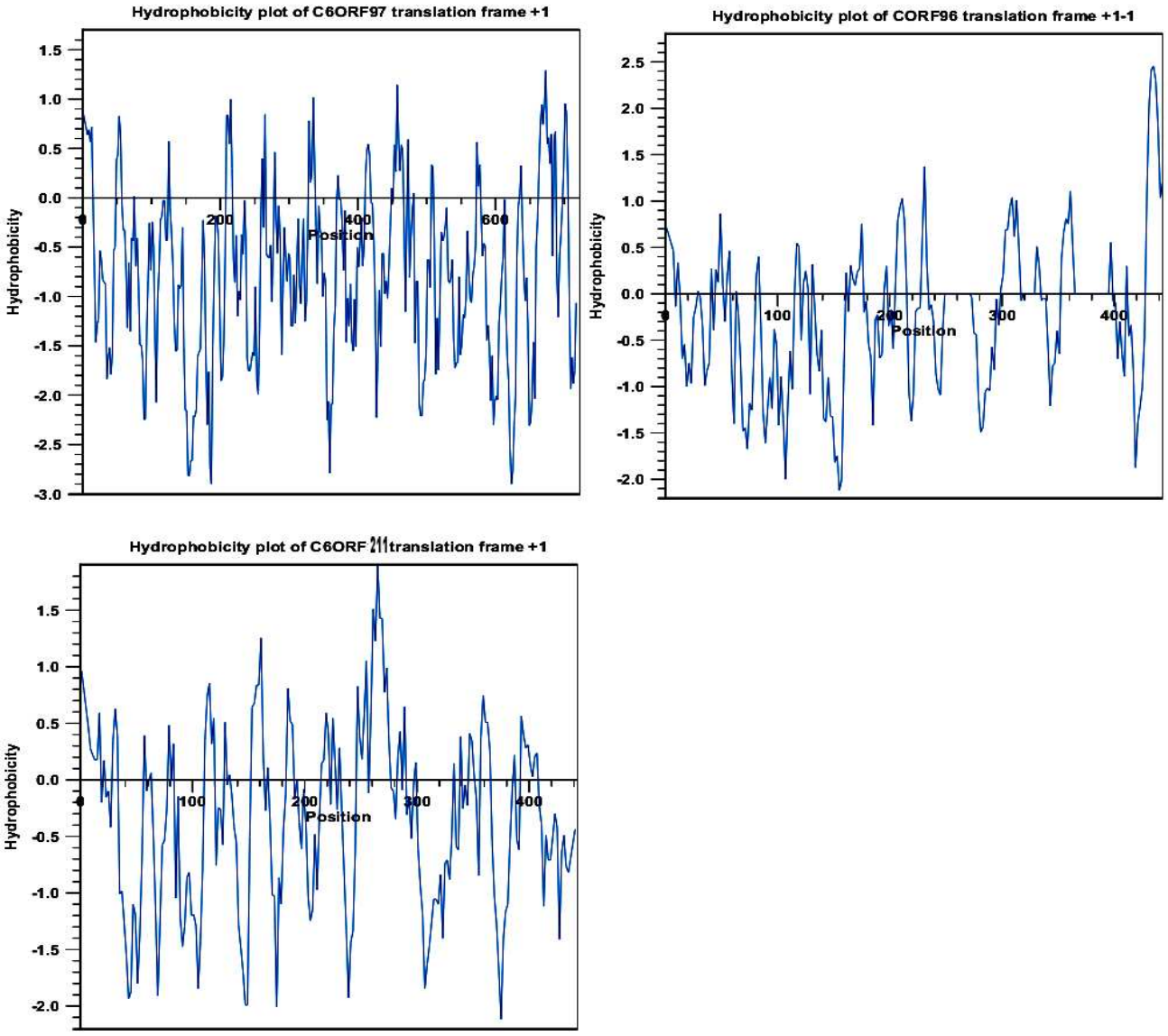


**Figure 7: Hydrophobicity of C6ORF96, C6ORf97 and C6ORF112 genes translation**