TIU Transactions on Human sciences

# Software Analysis of CRISPR/Cas-9 Mediated Genetic Editing of 'Neu' Oncogene in Breast Cancer Manifestation

Aniket De[a], Manav Goenka[a], Arup Ratan Biswas[a*]

[a] *Department of Chemistry, Techno India University, EM-4, EM Block, Sector V, Bidhannagar, Kolkata, West Bengal 700091, India. Email: arup.b@technoindiaeducation.com*

**Abstract**

CRISPR/Cas9 exhibits the genetic modification or alteration or editing by nicking the desired genomic DNA at the target location. The accuracy modification depends on PAM (5'-NGG-3') sequence identified by Cas-9. The target sequence is 20 bases long on a crRNA array. The corrected PAM sequences to the target sequence are implanted in to a plasmid related to its accuracy and precision with no error in the genetic cut or modification. The given target sequence and PAM sequence are matched with the help of the software available like that of CRISTA and ccTOP. The amino acid sequence of the "Neu" (Herceptin) protein obtained from Expassy was subjected to CRISTA and ccTOP. The highest similarity of the SgRNA sequence showed a value of 0.98 to 0.99, counts up to 4 in number. Current manuscript an attempt to use software based correct identification of target sequence corresponding to the correct PAM sequence as recognized by Cas-9 for the generation of proper matched single-guided RNA sequence without the generation of the off targets formed either due to the repeated bulging of DNA. Furthermore, the characteristic properties of the Neu gene are also explored in order to assess the membrane position, polarity index, and hydropathy index. Thus the manuscript shall provide a glimpse of the software analysis for the corrected target sequence to the corresponding PAM sequence related to the precise genetic modification or alteration.

***Key words**: CRISPR/Cas9; genetic modification; SgRNA; Neu gene; PAM; CRISTA; and ccTOP*

## 1. Introduction

CRISPR/Cas-9 is precisely used as molecular nicking tools for the purpose of twinkering the genes of interest. This technology allows precisely manipulating any genomic sequence specified by a short stretch of guide RNA, allowing elucidation of gene function involved in disease development and progressions, correction of disease-causing mutations, and inactivation of activated oncogenes or activation of deactivated cancer suppressor genes when utilizing a fusion protein of nuclease-deficient Cas-9 and effector domain [1, 2, 3, 4] .

Moreover, this programmable endonuclease technology enables researchers to examine the function of multiple genes at once by simultaneously targeting multiple genomic loci in a single experiment [5], which markedly accelerates our understanding of pathological processes that involve large sets of genes or mutations, such as tumor development. Using single-guide RNA (SgRNA) libraries, CRISPR based genome-wide screens can be leveraged to identify drug-target or disease-resistance genes, such as novel tumor suppressors or oncogenes, and to quickly assess drug targets [6, 7].

In this manuscript we used software for detecting the most corrected target sequence corresponding to the PAM sequence related to the Neu oncogene responsible for mammary cells proliferation. Neu gene is related to herceptin protein that needs to be checked for its polarity, hydropathy, and its position within the cellular membrane. It is indeed a necessity to use the software before designing the PAM sequence; else it may result to generation off target sites and lead to false positive result. Here we employed computational software (ccTOP and CRISTA) for the precise designing of the single-guided RNA sequence, PAM recognition site, corresponding to correct target sequence. The facilitation of the computer aided analysis of the SgRNA sequence and the PAM recognition sites for a specific target sequence that needs to be genetically edited or modified were adopted. At the same time it is necessary to understand

the properties of the oncoprotein that is the "Herceptin" protein translated from the "Neu" mRNA sequence. Main properties associated are hydropathy index, the alpha helix, beta-turns, helix wheel, and the transmembrane tendency. Thus we emphasized up on software evaluation and characteristic determination of the Neu oncogene and its related herceptin protein for correct analysis and designing of PAM and SgRNA sequences.

## 2. Research Method

Adopted methodology involves characteristic determination of herceptin protein in terms of hydropathy using Kyte and Doolittle index, polarity degree, and membrane position utilizing helix-loop wheel approach. Furthermore, approach has been initiated to explore the correct PAM recognition site for the given target sequence so as to generate the corrected single-guided RNA sequence with no off targets generation aided by CRISTA and ccTOP software analysis.

### 2.1 Elucidation of Protein Properties

#### a) Elucidation of Protscale Graphs

In order to properly understand and work with the Herceptin protein, we first examined and analyzed the chemical properties of the protein starting from its accession number, and properties such as hydropathy index, structural composition [8] through beta turns, alpha-helix and its transmembrane tendencies.

#### b) Hydropathy Index

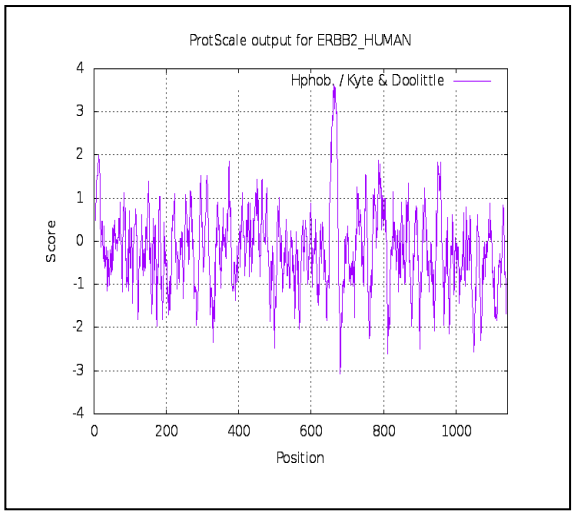



Figure 1: Hydropathy Index

An amino acids hydropathy index is a number reflecting the side chain's hydrophobic or hydrophilic properties [9]. General concept on hydrophobic and hydrophilic nature, involves the presence of specific amino acids that contributes to a protein hydrophobic and hydrophilic character, namely the presence of isoleucine and valine contributes to hydrophobic nature of the protein, whereas presence in number of arginine and lysine contributes to the hydrophilic nature of the given protein.

In this study of herceptin protein more presence of isoleucine and valine were observed suggesting its hydrophobic nature. It is very important in the composition of proteins; hydrophobic amino acids appear to be central (in terms of the 3-dimensional form of the Protein), while hydrophilic amino acids are located on the protein surfaces [10]. We have identified the amino acid sequences that show hydrophobic properties Figure 1.

#### c) Protein Structure

In a helix coiling, one amino acid carbonyl (C = O) [11] is hydrogen-bonded to the amino (-NH2) terminal of an amino acid that is 4 down the chain resulting to the formation of helical structure (Figures 2 and 3) as evident from the structure retrieved from Prot scale result. Furthermore, we have also found the helix wheel of the amino acid sequence using the Helixator of TCDB [12] and shown it in Figure 5.

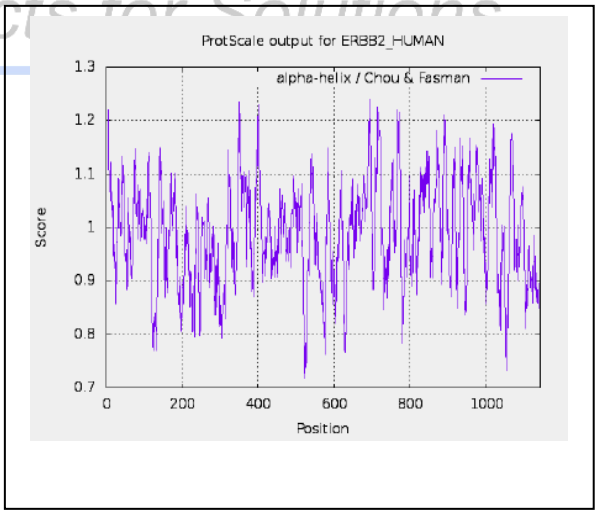



Figure 2: Alpha Helix Score

#### d) Transmembrane tendency

A transmembrane protein helps in and out of substances across the membrane and is hydrophobic in nature. To move a substance through the membrane

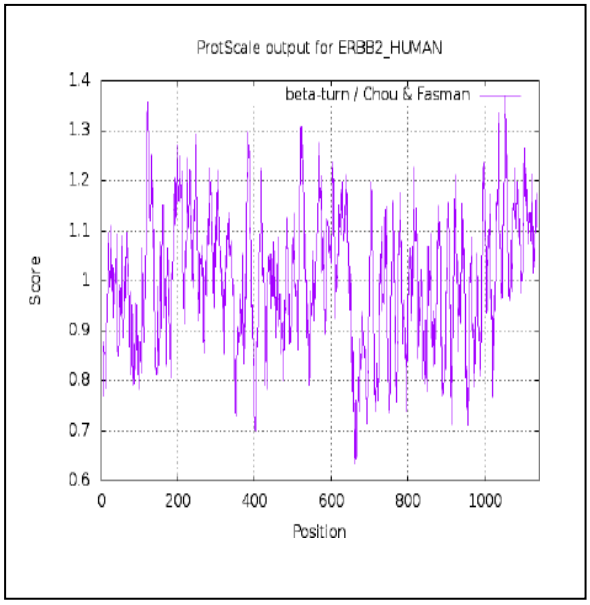



Figure 3: $\beta$ turn score

they frequently undergo significant conformation changes [13]. Doolittle and Kyte hydropathy index is used to investigate the transmembrane formation ability and the values plotted in Prot scale (Figure 4).

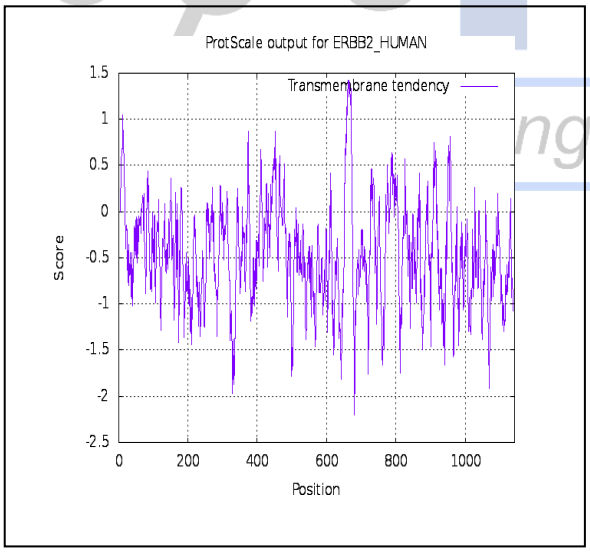



Figure 4: Spanning between membranes

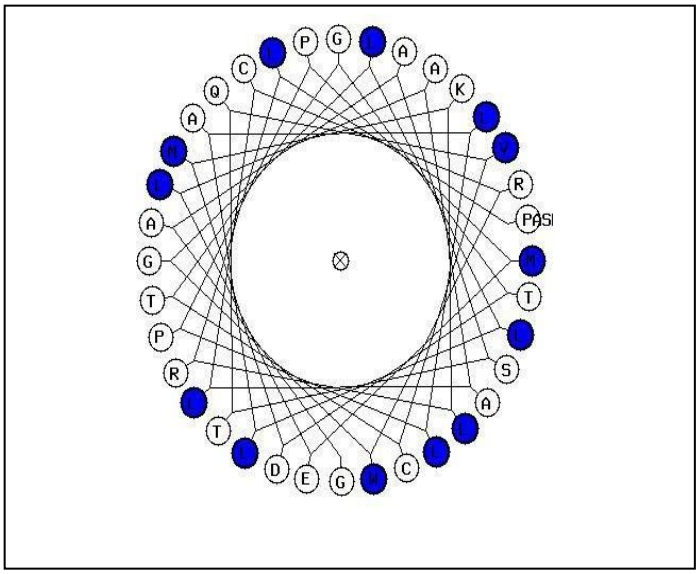


Figure 5: Helix Loop Wheel

### 2.2 Software Analysis

**(a) Processing of the FASTA File**

The FASTA file of the amino acid sequence of "Neu" gene obtained from Expassy, and the nucleotide sequence of homo sapiens genome under the name "Ensembl GRCh38" (Genome Reference Consortium Human Build 38) and the Cas-9 sequence obtained from Streptococcus Pyrogenes were considered and subjected to screening with CRISTA and ccTOP software for generating the SgRNA sequence corresponding to the target sequence.

**b) Computational analysis using ccTOP and CRISTA**

The computational analysis of the SgRNA and the target sequence were analyzed by the ccTOP and CRISTA software respectively. The reliability of the accuracy in identifying the SgRNA sequence and the target sequence are maximum using ccTOP and CRISTA. The other advantage is that both, takes into account of DNA bulge, since DNA bulge account into maximum number of misleading information. DNA after a period of sequence bulges and if the PAM recognition site falls within the bulging portion of the DNA, the SgRNA sequence and the target sequences are misleaded. Most other software contributes to that misleading information resulting to mismatch sequences.

## 3. Result

### a) CRISTA Output

The output from CRISTA, in the form of Single-guided RNA sequences [14] scoring is based on factors like DNA bulges, loops, the efficacy, off-target cleavage activity, and the final product is presented in the form of the highest CRISTA score sequence to lowest. Table 1 shows the matching pattern of each SgRNA sequences corresponding to the target DNA sequences with highest to lowest as 0.99 to 0.98.

Table 1: Matching scores obtained from CRISTA analysis

| | sgRNA | DNA site | strand | start position | end position | CRISTA score |
|---|---|---|---|---|---|---|
| 1 | AAACCCGGAATATCTGGGCCNGG | AAACCCGGAATATCTGGGCCTGG | + | 3732 | 3754 | 0.99 |
| 2 | GCGTATAGCCTGACCCTGCANGG | GCGTATAGCCTGACCCTGCAGGG | + | 1318 | 1340 | 0.99 |
| 3 | GACCTGCTTTGGCCCGGAAGNGG | GACCTGCTTTGGCCCGGAAGCGG | + | 1722 | 1744 | 0.98 |
| 4 | CCGGAATATCTGACCCCGCANGG | CCGGAATATCTGACCCCGCAGGG | + | 3580 | 3602 | 0.98 |

### b) ccTOP Output

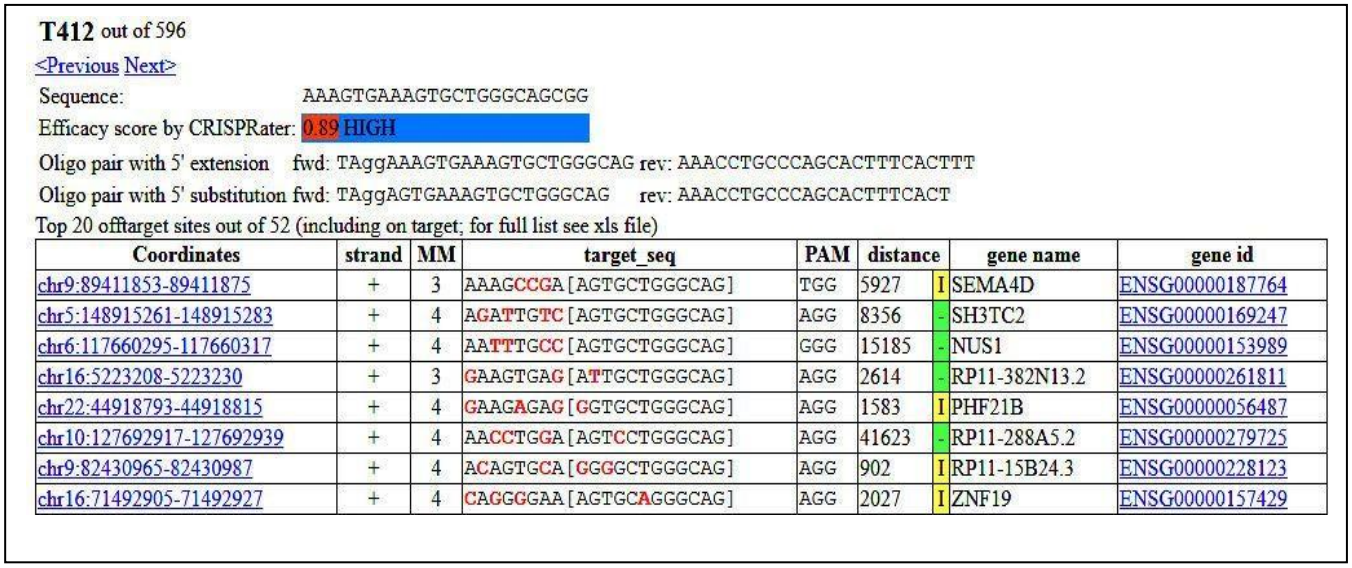
The nucleotide and the amino acid sequences are split into segments suggesting single guide RNA sequence. It is sorted according to the varying efficacy score that depends on their off-target activity and is presented with its oligo pair extension coordinates, PAM, gene name of the corresponding sequence, and the gene id giving a higher value to the control against the experiment [17]. Based on it analysis 4 best target sequences and corresponding SgRNA sequences have evolved (Figure 6, 7, 8, and 9).



**T102** out of 596
<Previous Next>
Sequence: GTTCGGCATCGCGCCGCTCGGGG
Efficacy score by CRISPRater: 0.90 HIGH
Oligo pair with 5' extension fwd: TAgGTTCGGCATCGCGCCGCTCG rev: AAACCGAGCGGCGCGATGCCGAA
Oligo pair with 5' substitution fwd: TAGgTCGGCATCGCGCCGCTCG rev: AAACCGAGCGGCGCGATGCCGA

| Coordinates | strand | MM | target_seq | PAM | distance | | gene name | gene id |
|---|---|---|---|---|---|---|---|---|
| chr3:123291256-123291278 | + | 4 | GCTCATCA[TTGCGCCGCTCG] | CGG | 0 | E | ADCY5 | ENSG00000173175 |
| chr19:1067075-1067097 | - | 4 | GCTCGGAG[CCGCGCCGCTCG] | CGG | 69 | I | HMHA1 | ENSG00000180448 |
| chr16:31459563-31459585 | + | 3 | GTTCTGCC[TCGCGCAGCTCG] | CGG | 0 | E | RP11-452L6.5 | ENSG00000260267 |
| chr20:17531128-17531150 | + | 4 | GCCCGGCA[TGGCGCTGCTCG] | AGG | 0 | E | BFSP1 | ENSG00000125864 |
| chr17:81961072-81961094 | + | 4 | GCTCGGCG[TCGAGGCGCTCG] | GGG | 0 | E | NOTUM | ENSG00000185269 |
| chr10:79347803-79347825 | - | 4 | GCCCGGCA[CCGCGCCGCCCG] | GGG | 60 | I | PPIF | ENSG00000108179 |
| chr5:32312954-32312976 | - | 4 | GGGCGGCA[TCGCGGCGGTCG] | CGG | 0 | E | MTMR12 | ENSG00000150712 |
| chr12:123533380-123533402 | - | 4 | GTACGACA[TCGCGTCGCTTG] | TGG | 0 | E | RILPL1 | ENSG00000188026 |

Figure 6: Matching score of 0.90



**T412** out of 596

<Previous Next>

Sequence: AAAGTGAAAGTGCTGGGCAGCGG

Efficacy score by CRISPRater: 0.89 HIGH

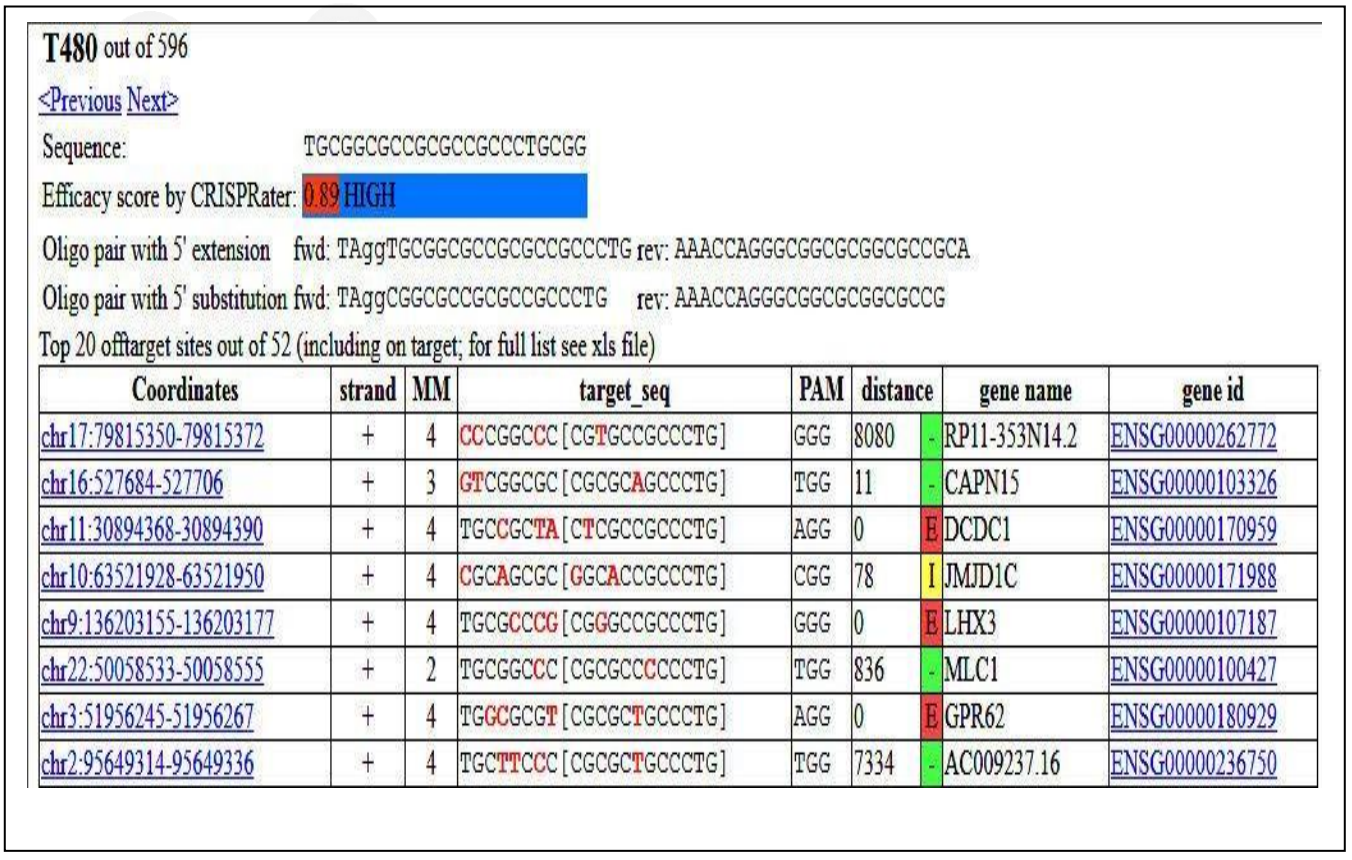
Oligo pair with 5' extension fwd: TAggAAAGTGAAAGTGCTGGGCAG rev: AAACCTGCCCAGCACTTTCACTTT

Oligo pair with 5' substitution fwd: TAggAGTGAAAGTGCTGGGCAG rev: AAACCTGCCCAGCACTTTCACT

Top 20 offtarget sites out of 52 (including on target; for full list see xls file)

| Coordinates | strand | MM | target_seq | PAM | distance | | gene name | gene id |
|---|---|---|---|---|---|---|---|---|
| chr9:89411853-89411875 | + | 3 | AAAGCCGA[AGTGCTGGGCAG] | TGG | 5927 | I | SEMA4D | ENSG00000187764 |
| chr5:148915261-148915283 | + | 4 | AGATTGTC[AGTGCTGGGCAG] | AGG | 8356 | - | SH3TC2 | ENSG00000169247 |
| chr6:117660295-117660317 | + | 4 | AATTTGCC[AGTGCTGGGCAG] | GGG | 15185 | - | NUS1 | ENSG00000153989 |
| chr16:5223208-5223230 | + | 3 | GAAGTGAG[ATTGCTGGGCAG] | AGG | 2614 | - | RP11-382N13.2 | ENSG00000261811 |
| chr22:44918793-44918815 | + | 4 | GAAGAGAG[GGTGCTGGGCAG] | AGG | 1583 | I | PHF21B | ENSG00000056487 |
| chr10:127692917-127692939 | + | 4 | AACCTGGA[AGTCCTGGGCAG] | AGG | 41623 | - | RP11-288A5.2 | ENSG00000279725 |
| chr9:82430965-82430987 | + | 4 | ACAGTGCA[GGGGCTGGGCAG] | AGG | 902 | I | RP11-15B24.3 | ENSG00000228123 |
| chr16:71492905-71492927 | + | 4 | CAGGGGAA[AGTGCAGGGCAG] | AGG | 2027 | I | ZNF19 | ENSG00000157429 |

Figure 7: Matching score of 0.89



**T480** out of 596

<Previous Next>

Sequence: TGCGGCGCCGCGCCGCCCTGCGG

Efficacy score by CRISPRater: 0.89 HIGH

Oligo pair with 5' extension fwd: TAggTGCGGCGCCGCGCCGCCCTG rev: AAACCAGGGCGGCGCGGCGCCGCA

Oligo pair with 5' substitution fwd: TAggCGGCGCCGCGCCGCCCTG rev: AAACCAGGGCGGCGCGGCGCCG

Top 20 offtarget sites out of 52 (including on target; for full list see xls file)

| Coordinates | strand | MM | target_seq | PAM | distance | | gene name | gene id |
|---|---|---|---|---|---|---|---|---|
| chr17:79815350-79815372 | + | 4 | CCCGGCCC[CGTGCCGCCCTG] | GGG | 8080 | - | RP11-353N14.2 | ENSG00000262772 |
| chr16:527684-527706 | + | 3 | GTCGGCGC[CGCGCAGCCCTG] | TGG | 11 | - | CAPN15 | ENSG00000103326 |
| chr11:30894368-30894390 | + | 4 | TGCCGCTA[CTCGCCGCCCTG] | AGG | 0 | E | DCDC1 | ENSG00000170959 |
| chr10:63521928-63521950 | + | 4 | CGCAGCGC[GGCACCGCCCTG] | CGG | 78 | I | JMJD1C | ENSG00000171988 |
| chr9:136203155-136203177 | + | 4 | TGCGCCCG[CGGGCCGCCCTG] | GGG | 0 | E | LHX3 | ENSG00000107187 |
| chr22:50058533-50058555 | + | 2 | TGCGGCCC[CGCGCCCCCCTG] | TGG | 836 | - | MLC1 | ENSG00000100427 |
| chr3:51956245-51956267 | + | 4 | TGGCGCGT[CGCGCTGCCCTG] | AGG | 0 | E | GPR62 | ENSG00000180929 |
| chr2:95649314-95649336 | + | 4 | TGCTTCCC[CGCGCTGCCCTG] | TGG | 7334 | - | AC009237.16 | ENSG00000236750 |

Figure 8: Matching score of 0.89

**T498** out of 596

<Previous Next>

Sequence: GCCGGGCGCGCCGCCGGCAGCGG

Efficacy score by CRISPRater: 0.89 HIGH

Oligo pair with 5' extension fwd: TAgGCCGGGCGCGCCGCCGGCAG rev: AAACCTGCCGGCGGCGCGCCCGG

Oligo pair with 5' substitution fwd: TAGgCGGGCGCGCCGCCGGCAG rev: AAACCTGCCGGCGGCGCGCCCG

Top 20 offtarget sites out of 52 (including on target; for full list see xls file)

| Coordinates | strand | MM | target_seq | PAM | distance | | gene name | gene id |
|---|---|---|---|---|---|---|---|---|
| chr18:35581722-35581744 | + | 2 | GCCGGGGG[AGCCGCCGGCAG] | TGG | 0 | E | GALNT1 | ENSG00000141429 |
| chr12:121467015-121467037 | + | 3 | CCCGGCCG[CCCCGCCGGCAG] | CGG | 121 | I | KDM2B | ENSG00000089094 |
| chr2:60881732-60881754 | + | 4 | ACTGGGGG[CCCCGCCGGCAG] | AGG | 0 | E | REL | ENSG00000162924 |
| chr17:51120680-51120702 | + | 3 | GCCCGGGG[CGTCGCCGGCAG] | AGG | 0 | E | SPAG9 | ENSG00000008294 |
| chr3:126380922-126380944 | + | 3 | GCTGGGCT[CGTCGCCGGCAG] | AGG | 12088 | - | CCDC37-AS1 | ENSG00000249833 |
| chrX:49250256-49250278 | + | 4 | GCCTGAGG[AGCCGCCGGCAG] | AGG | 0 | E | CCDC22 | ENSG00000101997 |
| chr4:2250558-2250580 | + | 4 | GGCGGCCC[AGCCGCCGGCAG] | TGG | 0 | E | MXD4 | ENSG00000123933 |
| chr4:1172555-1172577 | + | 4 | AGCGGGAG[CGCGGCCGGCAG] | CGG | 0 | E | SPON2 | ENSG00000159674 |
| chr19:41883360-41883382 | + | 4 | GGCTGGGG[CGCTGCCGGCAG] | CGG | 71 | I | ARHGEF1 | ENSG00000076928 |
| chr16:5239696-5239718 | + | 4 | TGCGGGCG[AGCGGCCGGCAG] | AGG | 84 | - | RP11-420N3.3 | ENSG00000279877 |

Figure 9: Matching score of 0.89

## 4. Discussion and Analysis

The research aims at elucidating the SgRNA sequence and the target sequence corresponding to the PAM sequence mediated by the software. We used two software's to validate the result of the SgRNA sequences and the target sequences based on the best fit matching scores. The result showed that score between 0.98 and 0.99 was obtained from both the software (CRISTA and ccTOP). Furthermore we also deciphered the properties of the herceptin protein in terms of its hydropathy index, polarity index, secondary structure, and transmembrane tendency.

Over expression of Neu gene leads to the proliferation of breast lymph nodes and other lymphatic system leading to the proliferation of the mammary duct. The herceptin protein, the biomarker of breast carcinoma has made clinical progress in term of the development of the drug "Tamoxifen" needs to be characterized in the line of the study. Accordingly, "Kyte and Dooliitle" Protscale graph (Figure-1) is used for the elucidation of hydropathy index, evidently showing the hydrophobic character of the protein due to its high content of isoleucine and valine. Polar characteristic of the protein is attributed to its hydrophobic nature.

In an helical turn, 3.6 amino acids are present in every turn of the protein secondary structure, (Figures 2 and 3) where the carbonyl carbon (C=O) is bonded to the amino (-NH2) terminal of the amino acids, that looks like a curled structure featuring a helical turn resembling the alpha-helix of the secondary structure of the protein as evident from the data delineated from the helix-loop wheel (Figure-5). The transmembrane tendency of the herceptin protein is due to its hydrophobic character (Figure-4). Presence of isoleucine and valine has made the protein to span between membranes thereby allowing the specific substance to pass across it.

The Homosapiens amino acid sequences and the nucleotide sequences obtained from Expassy were segmented and processed in both the software for the generation of the best fit SgRNA and the target sequence corresponding to PAM recognition site as recognized by the nuclease that is the Cas-9 protein obtained from the Streptococcus Pyrogenes. The score of 0.98 and 0.99 (Figures 6, 7, 8 and 9) provides the 4 best fit target and the SgRNA sequences that further needs to be transfected to a suitable plasmid for identification of the correct expression of the genetically edited Neu gene before going for the

final designing of the SgRNA sequences and the PAM sequence for the specific target sequence with a specific Cas-9 protein corresponding to the PAM recognition site.

## 5. Future Directions

The manuscript paves the way for further research in the field of CRISPR/Cas-9 for various oncogenes exploring different software studies before going for the final wet lab work. This study gives the glimpse of better understanding the designing of the PAM recognition site sequences, and SgRNA sequences for the corresponding target sequences and the Cas-9 protein. It minimizes the time of screening the proper sequences of SgRNA and PAM. Researchers working in the field of genetic editing with CRISPR/Cas-9 shall be beneficial by the approach undertaken as delineated by the manuscript.

## 6. Conclusion

Our paper identifies the best sequences of the "Neu" gene that can be targeted with the highest efficacy and lowest off-target cleavage with the CRISPR system and potentially pave way for higher oncological research for human's welfare [16].

## References

[1] Charpentier E, Marraffini L.A. 2014. "Harnessing CRISPR-Cas9 immunity for genetic engineering". Curr. Opin. Microbiol. 19:114–19

[2] Doudna JA, Charpentier E. 2014. "The new frontier of genome engineering with CRISPR-Cas9". Science 346(6213):1258096

[3] Hsu PD, Lander ES, Zhang F. 2014. "Development and applications of CRISPR-Cas9 for genome engineering". Cell 157(6):1262–78

[4] Wang H, La Russa M, Qi LS. 2016. "CRISPR/Cas9 in genome editing and beyond". Annual. Rev. Biochem.85:227–64

[5] Cong L, Ran FA, Cox D, Lin S, Barretto R, et al. 2013. "Multiplex genome engineering using CRISPR/Cas systems". Science 339(6121):819–23

[6] Dominguez A A, Lim WA, Qi LS. 2016. "Beyond editing: repurposing CRISPR–Cas9 for precision genome regulation and interrogation". Nat. Rev. Mol. Cell Biol. 17(1):5–15

[7] Shalem O, Sanjana NE, Zhang F. 2015. "High-throughput functional genomics using CRISPR–Cas9", Nat. Rev. Genet. 16(5):299–311

[8] Kyte J, Doolittle R F (May 1983). "A simple method for displaying the hydropathic character of a protein". J. Mol. Biol. 157 (1): 105–32. PMID 7108955

[9] Damodharan, Lakshminarasimhan, and Vasantha Pattabhi. "Hydropathy Analysis to Correlate Structure and Function of Proteins." Biochemical and Biophysical Research Communications, Academic Press, 17 Sept. 2004.

[10] Gallo, Elizabeth A., and Samuel H. Gellman. "Hydrogen-Bond-Mediated Folding in Depsipeptide Models of .Beta.-Turns and .Alpha.-Helical Turns." Journal of the American Chemical Society, vol. 115, no. 21, 1993, pp. 9774–9788.

[11] Saier, Milton H Jr et al. "The Transporter Classification Database (TCDB): recent advances." Nucleic acids research vol. 44,D1 (2016): D372-9.

[12] Zhao, Gang, and Erwin London. "An Amino Acid 'Transmembrane Tendency' Scale That Approaches the Theoretical Limit to Accuracy for Prediction of Transmembrane Helices: Relationship to Biological Hydrophobicity." Protein Science, vol. 15, no. 8, 2006, pp. 1987–2001.

[13] Abadi, Shiran, et al. "A Machine Learning Approach for Predicting CRISPR-Cas9 Cleavage Efficiencies and Patterns Underlying Its Mechanism of Action." PLOS Computational Biology, Public Library of Science.

[14] Stemmer, Manuel, et al. "CCTop: An Intuitive, Flexible and Reliable CRISPR/Cas9 Target Prediction Tool." PLOS ONE, Public Library of Science.

[15] Stemmer, Manuel et al. "CCTop: An Intuitive, Flexible and Reliable CRISPR/Cas9 Target Prediction Tool." PloS one vol. 10, 4 e0124633. 24 Apr. 2015.

[16] De, A., & Biswas, D. (2020). "Elucidative PAM/Target Sequence for CRISPR/Cas- 9 activity in Breast Cancer Using a Computational Approach". International Journal Of Innovative Science And Research Technology, 5(7), 872-876.

[17] Annunziato, Stefano, et al. "In Situ CRISPR-Cas9 Base Editing for the Development of Genetically Engineered Mouse Models of Breast Cancer." The EMBO Journal, John Wiley & Sons, Ltd, 13 Jan. 2020.