CRISPR/Cas-9 Genetic Editing of ‘EGFR’ gene Using Computational tool

Manav Goenka
Department of Biotechnology
Techno India University, West Bengal
manav.gnk@gmail.com

Aniket De
Department of Biotechnology
Techno India University, West Bengal
aniketde9@gmail.com

Arup Ratan Biswas*
Corresponding Author:
Department of Chemistry,
Techno India University, West Bengal
hod.chemisty@technoindiaeducation.com

***Abstract*:**
CRISPR/Cas9 (Clustered Regularly Interspaced Short Palindromic Repeats/CRISPR associated protein-9) system, discovered in 2012 by the 2020 Nobel Prize-winning Laureates Jennifer Doudna and Emmanuelle Charpentier is a bacterial defensive mechanism that exhibits the cleavage of genomic DNA at the desired location, resulting in the exit of the old genes and the induction of a new set of genes. The accuracy, precision or fidelity of the genetic cut depends on the target and the proto-spacer adjacent motif (PAM) sequences. The Cas9 protein recognises the PAM sequence (5'-NGG-3') by selecting the correct location of base-pair bonds within the target sequence on the host genome. Assembling the nucleotide sequence related to PAM and target sequence into a plasmid and then transfecting the plasmid into a cell shows that Cas9 with the help of a crRNA detected the correct sequence within a host cell. This results in a single or double-stranded break at the appropriate location in the DNA, thereby working as a molecular scissor and performing a genetic cut. We choreographed this tool in achieving the information related to the generation of the PAM sequence and the off-target sites associated with the EGFR gene. We have identified the top 4 best gRNA sequences based on the highest SYNTHEGO scores range 0.98 to 0.99. Furthermore, we used CCTop to identify the 4 best targets and guide RNA sequences with the highest efficacy score. Our manuscript is aimed at showcasing the best target sequence utilizing model software like that of SYNTHEGO and CCTop.

Keywords: CRISPR/Cas9; gRNA; molecular scissor; EGFR gene; SYNTHEGO and CCTop computational tool.

I. INTRODUCTION

The clustered regularly interspaced short palindromic repeats (CRISPR) – CRISPR-associated protein 9 (Cas9) system is a bacterial defence mechanism against phage infection.The system is a component of the adaptive immunity in bacteria against viruses and plasmids. This method has had successful applications in biological systems ranging from yeasts to rodents and mammals and thus, has intentionally been used as a powerful RNA-guided DNA targeting platform for genome editing, transcriptional perturbation, epigenetic modulation, and genome imaging [1]. This technology allows precise manipulation of 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 Cas9 and effector domain [2, 3]. CRISPR based genome-wide screens can be leveraged using single-guide RNA (sgRNA) libraries for the identification of drug-target or disease-resistance genes, such as novel tumour suppressors or oncogenes, and to quickly assess drug targets [4, 5].

CRISPR/Cas9 endonuclease system is currently targeted as a molecular surgery tool to achieve success in cancer treatment. Cancers are mostly related with genetic alteration and mismatches in cell cycle checkpoints. The tumour suppressor genes and proteins play a crucial role in controlling the cell cycle. Mutation in any of the checkpoints or the tumour suppressor gene may change the scenario and may cause a chaotic situation to an individual’s life, causing a clinical manifestation leading to cancer. Briefly, Cas9 locates specific 20-base-pair (bp) target sequences within the genomes that are billions of base pairs long and subsequently induces sequence-specific double-stranded DNA (dsDNA) cleavage [3]. In this manuscript we are highlighting the genetic changes associated with the EGFR oncogene enhancement, also known as HER-1 or ERBB1, a transmembrane protein. Studies revealed mutation in the Ferroprotein tyrosine kinase domain as the majority source of 75% of non-small cell lung cancer (NSCLC) cases [6]. Amplification leads to higher rates of EGFR-mRNA and protein synthesis. Furthermore, as EGFR gene encodes a transmembrane protein of the same name, we envisaged the characteristic properties of the EGFR protein so as to get the holistic picture of the gene as well as the protein associated with lung carcinoma. Our main approach is to highlight the CRISPR/Cas9 technology as a novel genetic cutting tool that may be employed in editing the target region of the EGFR gene.

Despite of its full potential, this genetic cutting tool has its own demerits to the point that it cannot be used in full therapeutic measures since it is indeed a matter of further investigation regarding the precision and accuracy of cutting the target sequence and not generating the off targets, which may cause genetic damage instead of it achieving the goal. Even though of its demerits it is now the buzz of frontier research and had opened a new dimension in the field of cancer biology research in the name of a new discipline termed as CRISPR biology. This manuscript stands over the others since it highlights the latest tool ever developed in the treatment of any form of cancer and here to specify is the lung cancer.

II. EGFR ROLE AND PATHWAY IN LUNG CANCER

The EGFR (Epidermal Growth Factor Receptor) gene codes for making the epidermal growth factor receptor, which is a transmembrane glycoprotein undergoing conformational changes to assist in autophosphorylation and the MAPK pathway [7]. 90% of the known EGFR mutations occur as frame-shift mutations in Exon 19 or point mutations in Exon 21[8] resulting in the continuous activation of the various signal transduction pathways and thus leading to the various tumorigenic pathways. The EGFR pathway plays a crucial role in regulating growth, survival, proliferation, differentiation, and cell to cell communication in mammalian cells. 15 members belonging to the EGF ligand family have been identified as the input signals and induces the homodimerization (EGFR – EGFR) and heterodimerization (EGFR – HER2). This dimerization causes transphosphorylation on numerous tyrosine residues which lead to phosphotyrosine-binding adaptors linking to phosphorylated receptors. Various transcription factors are then activated when they translocate to the nucleus.

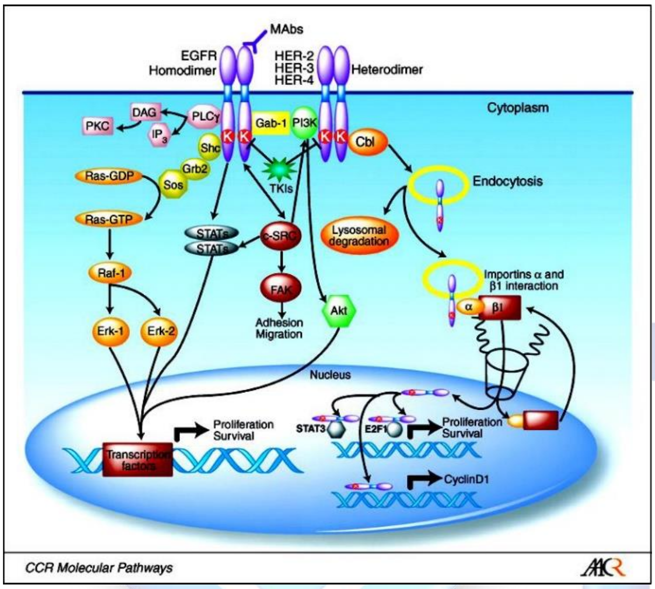


Figure 1: EGFR signalling pathways

III. MECHANISM OF CRISPR/CAS-9 ACTION

The CRISPR/Cas system is RNA mediated and relies on small RNA sequences (approx. 20-22 nucleotides long) for detection and silencing of foreign DNA in a site-specific manner. They use a non-specific endonuclease to cut a genomic sequence. A small guide RNA (gRNA) guides the Cas protein to a specific site [9]. The Cas protein in an endonuclease which by definition means that it cuts a specific stretch of nucleotides within the nucleic acid. It is guided by a short nucleotide guide RNA (or gRNA) which are approximately 20-22 nucleotides long, to locate the complementary protospacer DNA target in a genome. The defence mechanism of CRISPR/Cas9 involves 3 distinct steps [10] adaptation of the CRISPRs, genesis of crRNA and lastly, silencing of the foreign DNA. In the adaptation step, the Cas operon transcribes the Cas1-Cas2 complex which chooses a portion of the foreign DNA to integrate into the host's CRISPR arrangement. This copy is called the spacer sequence and the protospacer selected by the adaptation machinery is usually compatible with the PAM sequence of the silencing machinery. This sequence is integrated to the immediate downstream to the leader sequence with a record of the previous infections[11]. This is followed by the step of crRNA genesis where the crRNAs are transcribed and matured. In numerous organisms, continuous production of crRNA and Cas9 proteins takes place, operating in a 'surveillance mode'. In specific strains of E. coli, foreign presence also triggers an elevation in expression of the complex [10]. Finally, the crRNAs are loaded onto the final effector complex and guided to the invading DNA by the recognition of a PAM sequence. The Cas9 complex then cuts the double strand specifically 3 base pairs upstream to the PAM site.

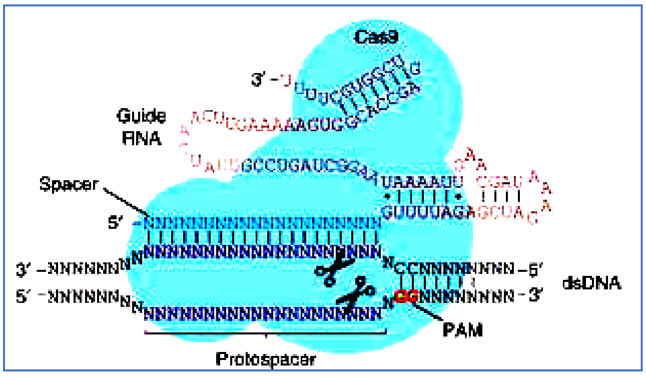


Figure 2: CRISPR/Cas-9 showing as molecular scissor

IV. RESEARCH OBJECTIVE

Our research is focused in elucidating the target sequence associated with the EGFR gene responsible for the clinical manifestation of lung cancer or more specifically, NSCLC that may be possible recognized by the PAM sequence corresponding to the target sequence and employing the computational tool we have focused to elucidate the corresponding PAM and target sequence related to the EGFR gene. At the same time, we also elucidated the protein properties in terms of its hydropathy index and polarity related to the EGFR protein. Our manuscript gives the necessary information regarding further work in synthesizing the software based generated PAM sequence and transplanting the same in the plasmid vector and to check out the interference of the generated PAM sequence with the single guide RNA for target identification and finally cessation of the target sequence and establishing the cut sequence with modified sequence which shall normalize the function of the EGFR gene or rather deactivate the over expression of the EGFR gene and thereby would diminish the cancer prognosis.

V. RESEARCH METHOD

The research method or the experimental set up for this manuscript is divided in to four parts.

Part-A mainly emphasizes the protein properties elucidated as protscale graphs related to hydropathy index and the polarity.

Part-B mainly emphasizes the computational tool-based approach undertaken to identify the target sequence, PAM sequence and SgRNA obtained from different online computational tool like that of CCTop and SYNTHEGO.

Part-C mainly emphasizes the output and experimental setup from Synthego

Part-D mainly emphasizes the output and experimental setup from Synthego

A. Elucidation of Protein Properties

A.1. Elucidation of Protscale Graphs

In order to properly understand and work with the EGFR protein, we first examined and analysed the chemical properties of the protein starting from its accession number, proteomics, properties such as hydropathy index and polarity

A.1.1. Hydropathy Index

An amino acid's hydropathy index is a number reflecting the sidechain's hydrophobic or hydrophilic properties [13]. The greater the number, the higher is the hydrophobic character of the amino acid. This property was proposed in 1982 by Jack Kyte and Russel F. Doolittle [14]. The most hydrophobic amino acids are considered to be Isoleucine and Valine. Arginine and lysine are the most hydrophilic ones. 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 [15]) while hydrophilic amino acids are more generally located on the protein surfaces. We have identified the amino acid sequences that show hydrophobic properties and it has been represented in **Figure 3**.

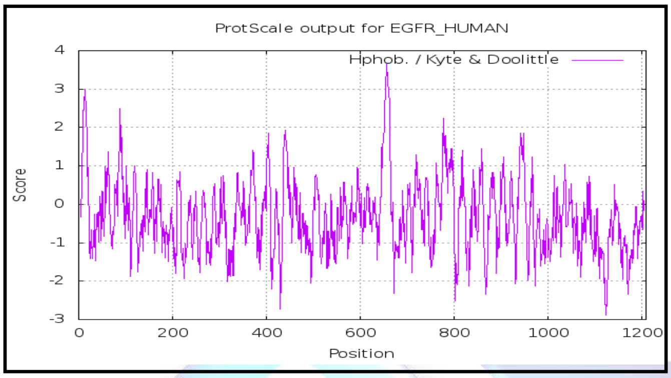


Figure 3: The hydrophobic property of EGFR protein

A.1.2. Polarity

Polarity of a protein is the resultant of the electronegativity difference between the bonded atoms. A protein may be termed as being polar or nonpolar depending on the distribution of charges in its amino acid sequence. It is generally observed that amino acids with polar side groups are present on the protein surface while the non-polar amino acids constitute the interior core of the proteins. Polar amino acids tend to be hydrophilic with their non-polar counterparts being generally hydrophobic. J M Zimmerman [16] in 1968 attempted to use statistical methods by taking into consideration polarity and bulkiness of the protein as factors to determine the individual amino acid role within the protein configuration. We have identified the amino acid sequences that show polar properties and it has been represented in **Figure 4.**

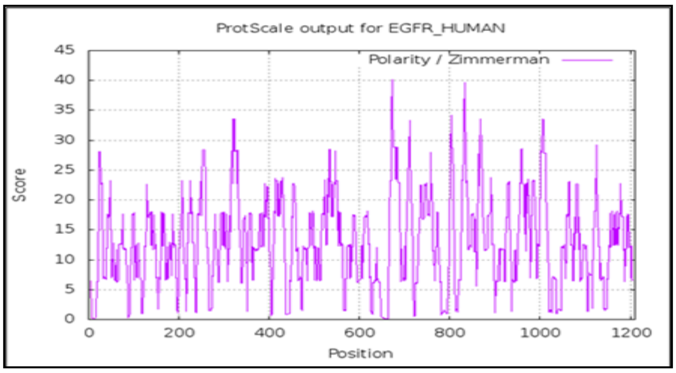


Figure 4: The polarity behaviour of EGFR protein

B. Computation tool-based approach to identify target sequence for CRISPR

At first, we worked out the amino acid sequence of the EGFR protein with its accession number from ExPASy after which we individually looked at the nucleotide sequence of each amino acid and converted that into a FASTA file. Now this FASTA file was scanned in batch mode in two different software – SYNTHEGO and CCTop and the following factors were common for both the software before analysis.

Genome – Homo sapiens – Ensembl GRCh38 (Genome Reference Consortium Human Build 38).

Nuclease – SpCas9 – Streptococcus pyrogenes.

The negative marks are a prime explanation representing the wastefulness of CRISPR in its space. Be that as it may, with present-day computational instruments, the system of activity of CRISPR was improved as well as its plausible results were likewise anticipated all the more precisely. The calculation is based on information that has been extracted through various sources and the amalgamation of all this information can be utilized by the AI to predict cleavage efficiencies. The essential bad mark of Cas-9 is that it divides askew DNA thus to counter that, analysts began executing AI calculations utilizing computational instruments to develop a progressively exact cleavage result and disposing of the off-target bad marks. They would breakdown a portion of the most important and reliable CRISPR AI systems that are eligible for usage and assess their validity by looking at their yields for our desired outcomes. Of the well-known analytical methods, SYNTHEGO [17] and CCTop [18, 19] are considered as the most innovative solutions because of their willingness to take into account DNA bulges, which are sometimes ignored by other devices. This has had a significant impact on improving accuracy because DNA bulges are verycommon phenomena that tend to hinder the desired result of our DNA manipulation.

B.1. SYNTHEGO

Synthego was created in 2012 with a vision for automating biological research. It uses machine learning with automation and genetic editing technology to integrate proprietary hardware, software, bioinformatics, chemistries, and molecular biology [20]. We have employed the freely available in-silico CRISPR design tool [17] from Synthego, an efficient tool in designing and validating guide RNA sequences. It suggests the user the best gRNA sequence according to the host genome and the gene of interest. The CRISPR design tool from Synthego ensures that the guide RNA sequences designed by the algorithm have the highest probability of generating a CRISPR knockout gene and minimizing potential off-target sites. The software commences by identifying the entire sequence of the entire gene within the genome of the host organism. It then proceeds to identify the primary transcript for all possible alternative transcripts or splice variants that might exist across numerous databases for that particular gene. The tool then identifies the exons and coding sequences for the primary transcript, narrowing the exons down to identify the guide RNAs in the target gene having the highest probability of producing a complete functional knockout through insertions or deletions (indels). Based on the available PAM sites on the exon, the software then identifies the prospective CRISPR-Cas9 target sites. Once the target sites and the exons are detected, the tool locates and determines all potential off-target sites and the mismatches between each target gRNA and prospective off-target sites. This information aids the algorithm to rank the sgRNAs based on the Azimuth 2.0 model from Doench et al. [21]. After collecting all the data, the algorithm finally recommends the guide RNAs that have the highest possibility to knock out the target gene within the host genome with the least number of off-target effects.

B.2. CCTop

CRISPR/Cas9 target online predictor (or CCTop) is an in-silico approach to provide the user with a set of easily adjusted reasonable default parameters. This tool can identify and rank the prospective sgRNA target sites as per their off-target quality. These sgRNA target sites are identified according to variable parameters like the type of PAM and the identity of the two most 5' or 3' nucleotides. CCTop also allows the user to make separate PAM type selections for off-target predictions. The results page generated provides a sophisticated range of information in the form of a graphical representation of the query sequence, with all identified sgRNA target sites. Detailed information such as the genomic coordinates, target sequences with highlighted mismatches along with the distance and position are provided for off-target sites. These results can be downloaded as .CSV or as a fasta file containing all sgRNA target sites [18].

C. Synthego Output

The output from Synthego gives us a clear comparative study of the possible guides after running them through (i) Knockout guide structure(ii) Verifying sgRNA plan and (iii) ICE Analysis. Thiseffective apparatus recommends to us the best gRNA grouping relying upon the genome of use and the quality that we are attempting to control and can be thus used to configure the data direct RNA. It likewise gives us a visual interface on each gRNAsuccessions on track versus the off-target score and positions themfrom the most noteworthy effectiveness to least for that specificquality. One can likewise arrange the gRNA groupings onlinefrom Synthego to be conveyed to their lab.

C.1. Experimental Setup in Synthego

We analysedthe gene EGFR, for lung cancer using Synthego's Knockout Guide Design with the following inputs:

(I) EGFR

1) Genome – Homo sapiens – Ensembl GRCh38 (Genome Reference Consortium Human Build 38).

2) Gene – EGFR – 1956 ENSG00000146648 ROS epidermal growth factor receptor.

3) Nuclease – SpCas9 – Streptococcus pyogenes.

D. CCTop Output

As for CCTop, it's not yet known to take into account the bulges and loops totally while analysing the sequence, however, the output presented by CCTop is more detailed and organized when it comes to actual experimentation. The output is given by breaking down the entire sequence into several target sequences and suggesting a guide RNA corresponding to that. It is sorted according to the target sequence and 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 control to the experimentation carrier [22].

VI. RESULT

TABLE- 1 attached below contains the list of all possible guide RNAs where there is the possibility of mimicking genetic editing. The best possible guide RNAs (gRNA) for maximum Cas-9 activity are as follows: UUACUCGUGCCUUGGCAAAC,CUUUUUCUUCCAGUUUGCCA,UGAGCUUGUUACUCGUGCCU,GAGUAACAAGCUCACGCAGU. The results show 4 top-rated guide RNAs for editing EGFR, the target sequences, the respective protein-codinggenes for that sequence, the chromosome number in parallelalong with the cute site and the PAM region. The identification of the gRNA sequence with possible off target sites located in a specific chromosome and with specific PAM region as obtained while running the Synthego software would help in designing the specific primer for the wet lab experimentation. The statistical analysis of the PAM ratio for the four gRNA sequences is shown in **Figure 5.**

TABLE 1: SYNTHEGO SOFTWARE ANALYSIS FOR EGFR

UUACUCGUGCCUUGGCAAAC		
Best off-target sites	***Chr no'***	***PAM***
TTGCTCGTTCCTGGGCAAAC	22	AGG
TTACTCATTCCTTGGCAGAC	18	TGG
TTATGCTTCCCTTGGCAAAC	3	AGG
TTAGTCCTGCCTAGGCAGAC	3	TGG
CUUUUUCUUCCAGUUUGCCA		
Best off-target sites	***Chr no'***	***PAM***
CTTTTTCTTCCAGTTTCCTA	8	AGG
CTTTTTCCTCCTGTTTGCCT	2	TGG
CTTTGTTCTCCAGTTTGCCA	2	TGG
TTTTGTCTTCTAGTTTGCCA	15	GGG
UGAGCUUGUUACUCGUGCCU		
Best off-target sites	***Chr no'***	***PAM***
TGAGCTTGTTACTCGTGCCT	7	TGG
AGAGCTTGTTACTTGTGCCC	9	TGG
TGATCTTGTTCCTCCTGACT	X	GGG
TGAGCTTGGCACTGGAGCCT	1	AGG
GAGUAACAAGCUCACGCAGU		
Best off-target sites	***Chr no'***	***PAM***
GAGTAACAAGCTCACGCAGT	7	TGG
GACTGACACGCTCACGCAGT	19	GGG
GAGCACCAAGCTCAAGCAGG	22	TGG
GAGTAAGAAGCTCAGGCTGA	HSCHR20_1_CTG3	AGG

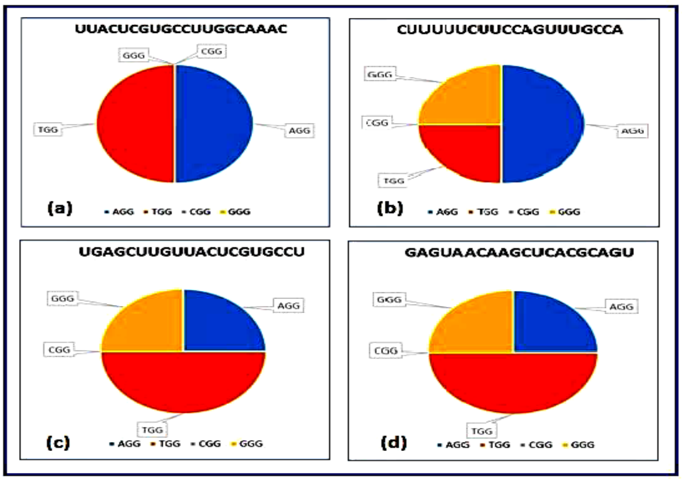
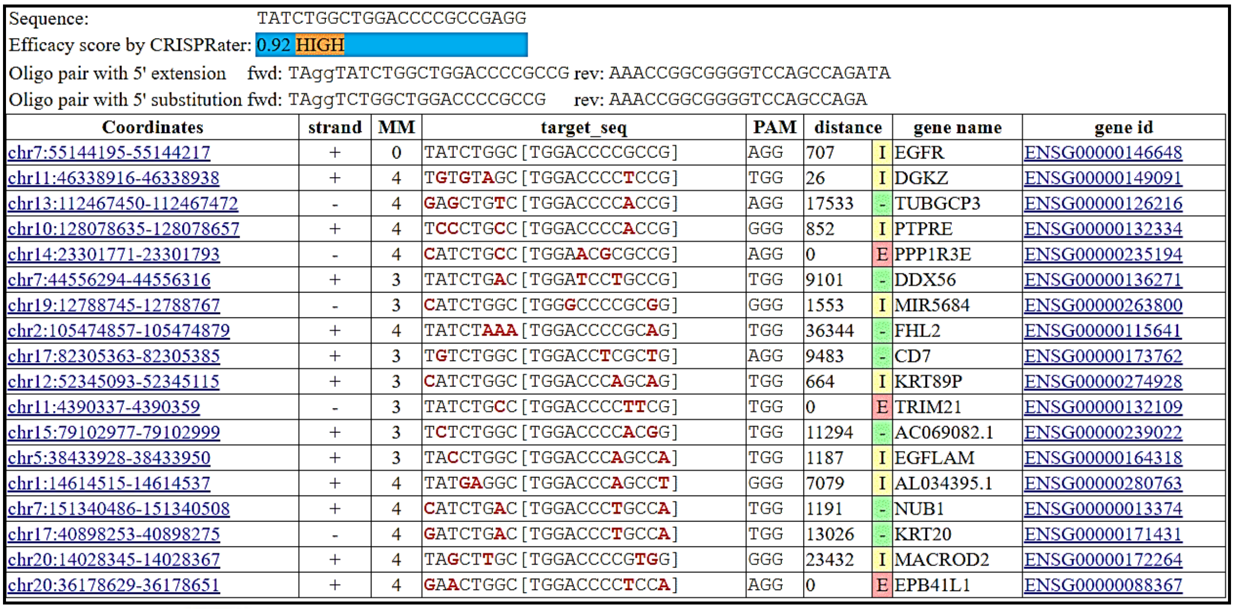


Figure 5: PAM ratio of the EGFR gene

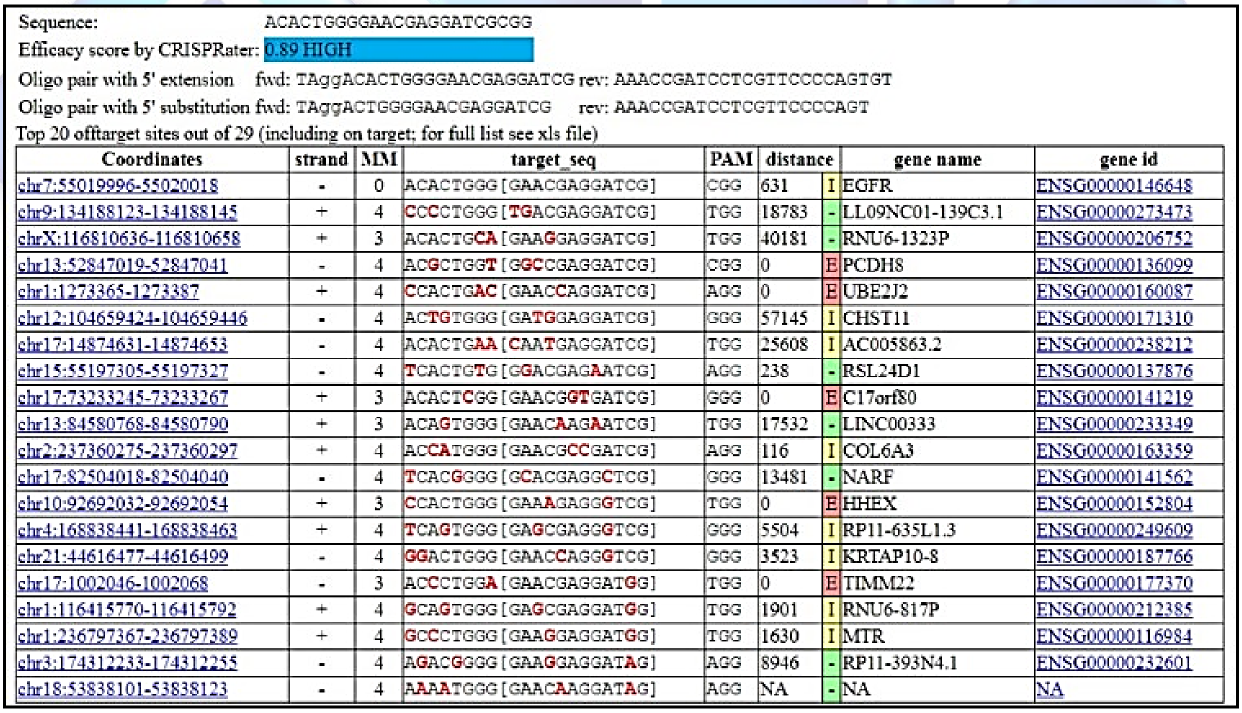
We have also identified the 4 best target sequences and guides based on the highest efficacy score as shown below as showcased in **Figures 6, 7, and 8**



Sequence: TATCTGGCTGGACCCCGCCGAGG
Efficacy score by CRISPRater: 0.92 HIGH
Oligo pair with 5' extension fwd: TAggTATCTGGCTGGACCCCGCCG rev: AAACCGGCGGGGTCCAGCCAGATA
Oligo pair with 5' substitution fwd: TAggTCTGGCTGGACCCCGCCG rev: AAACCGGCGGGGTCCAGCCAGA

Coordinates	strand	MM	target_seq	PAM	distance		gene name	gene id
chr7:55144195-55144217	+	0	TATCTGGC [TGGACCCCGCCG]	AGG	707	I	EGFR	ENSG00000146648
chr11:46338916-46338938	+	4	TGTGTAGC [TGGACCCCTCCG]	TGG	26	I	DGKZ	ENSG00000149091
chr13:112467450-112467472	-	4	GAGCTGTC [TGGACCCCACCG]	AGG	17533	-	TUBGCP3	ENSG00000126216
chr10:128078635-128078657	+	4	TCCCTGCC [TGGACCCCACCG]	GGG	852	I	PTPRE	ENSG00000132334
chr14:23301771-23301793	-	4	CATCTGCC [TGGAACGCGCCG]	AGG	0	E	PPP1R3E	ENSG00000235194
chr7:44556294-44556316	+	3	TATCTGAC [TGGATCCTGCCG]	TGG	9101	-	DDX56	ENSG00000136271
chr19:12788745-12788767	-	3	CATCTGGC [TGGGCCCCGCGG]	GGG	1553	I	MIR5684	ENSG00000263800
chr2:105474857-105474879	+	4	TATCTAAA [TGGACCCCGCAG]	TGG	36344	-	FHL2	ENSG00000115641
chr17:82305363-82305385	+	3	TGTCTGGC [TGGACCTCGCTG]	AGG	9483	-	CD7	ENSG00000173762
chr12:52345093-52345115	+	3	CATCTGGC [TGGACCCAGCAG]	TGG	664	I	KRT89P	ENSG00000274928
chr11:4390337-4390359	-	3	TATCTGCC [TGGACCCCTTCG]	TGG	0	E	TRIM21	ENSG00000132109
chr15:79102977-79102999	+	3	TCTCTGGC [TGGACCCCACGG]	TGG	11294	-	AC069082.1	ENSG00000239022
chr5:38433928-38433950	+	3	TACCTGGC [TGGACCCAGCCA]	TGG	1187	I	EGFLAM	ENSG00000164318
chr1:14614515-14614537	+	4	TATGAGGC [TGGACCCAGCCT]	GGG	7079	I	AL034395.1	ENSG00000280763
chr7:151340486-151340508	+	4	CATCTGAC [TGGACCCTGCCA]	TGG	1191	-	NUB1	ENSG00000013374
chr17:40898253-40898275	-	4	GATCTGAC [TGGACCCTGCCA]	TGG	13026	-	KRT20	ENSG00000171431
chr20:14028345-14028367	+	4	TAGCTTGC [TGGACCCCGTGG]	GGG	23432	I	MACROD2	ENSG00000172264
chr20:36178629-36178651	+	4	GAACTGGC [TGGACCCCTCCA]	AGG	0	E	EPB41L1	ENSG00000088367

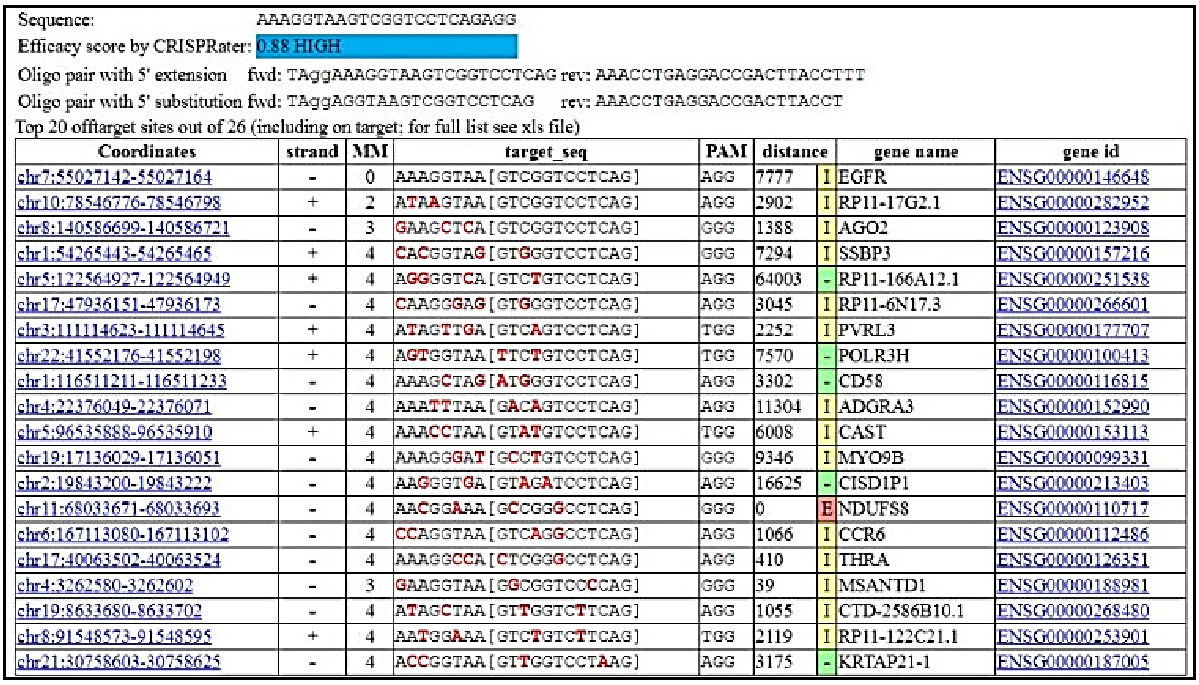
Figure 6: The identified target sequence for CRISPR/Cas-9 activity against EGFR gene with efficacy score of 0.92 – output obtained from using CCTop software



Sequence: ACACTGGGGAACGAGGATCGCGG
Efficacy score by CRISPRater: 0.89 HIGH
Oligo pair with 5' extension fwd: TAggACACTGGGGAACGAGGATCG rev: AAACCGATCCTCGTTCCCCAGTGT
Oligo pair with 5' substitution fwd: TAggACTGGGGAACGAGGATCG rev: AAACCGATCCTCGTTCCCCAGT
Top 20 offtarget sites out of 29 (including on target; for full list see xls file)

Coordinates	strand	MM	target_seq	PAM	distance		gene name	gene id
chr7:55019996-55020018	-	0	ACACTGGG [GAACGAGGATCG]	CGG	631	I	EGFR	ENSG00000146648
chr9:134188123-134188145	+	4	CCCCTGGG [TGACGAGGATCG]	TGG	18783	-	LL09NC01-139C3.1	ENSG00000273473
chrX:116810636-116810658	+	3	ACACTGCA [GAAGGAGGATCG]	TGG	40181	-	RNU6-1323P	ENSG00000206752
chr13:52847019-52847041	-	4	ACGCTGGT [GGCCGAGGATCG]	CGG	0	E	PCDH8	ENSG00000136099
chr1:1273365-1273387	+	4	CCACTGAC [GAACCAGGATCG]	AGG	0	E	UBE2J2	ENSG00000160087
chr12:104659424-104659446	-	4	ACTGTGGG [GATGGAGGATCG]	GGG	57145	I	CHST11	ENSG00000171310
chr17:14874631-14874653	-	4	ACACTGAA [CAATGAGGATCG]	TGG	25608	I	AC005863.2	ENSG00000238212
chr15:55197305-55197327	-	4	TCACTGTG [GGACGAGAATCG]	AGG	238	-	RSL24D1	ENSG00000137876
chr17:73233245-73233267	+	3	ACACTCGG [GAACGGTGATCG]	GGG	0	E	C17orf80	ENSG00000141219
chr13:84580768-84580790	+	3	ACAGTGGG [GAACAAGAATCG]	TGG	17532	-	LINC00333	ENSG00000233349
chr2:237360275-237360297	+	4	ACCATGGG [GAACGCCGATCG]	AGG	116	I	COL6A3	ENSG00000163359
chr17:82504018-82504040	-	4	TCACGGGG [GCACGAGGCTCG]	GGG	13481	-	NARF	ENSG00000141562
chr10:92692032-92692054	+	3	CCACTGGG [GAAAGAGGGTCG]	TGG	0	E	HHEX	ENSG00000152804
chr4:168838441-168838463	+	4	TCAGTGGG [GAGCGAGGGTCG]	GGG	5504	I	RP11-635L1.3	ENSG00000249609
chr21:44616477-44616499	-	4	GGACTGGG [GAACCAGGGTCG]	GGG	3523	I	KRTAP10-8	ENSG00000187766
chr17:1002046-1002068	-	3	ACCCTGGA [GAACGAGGATGG]	TGG	0	E	TIMM22	ENSG00000177370
chr1:116415770-116415792	+	4	GCAGTGGG [GAGCGAGGATGG]	TGG	1901	I	RNU6-817P	ENSG00000212385
chr1:236797367-236797389	+	4	GCCCTGGG [GAAGGAGGATGG]	TGG	1630	I	MTR	ENSG00000116984
chr3:174312233-174312255	-	4	AGACGGGG [GAAGGAGGATAG]	AGG	8946	-	RP11-393N4.1	ENSG00000232601
chr18:53838101-53838123	-	4	AAAATGGG [GAACAAGGATAG]	AGG	NA	-	NA	NA

Figure 7: The identified target sequence for CRISPR/Cas-9 activity against EGFR gene with efficacy score of 0.89 – output obtained from using CCTop software.



Sequence: AAAGGTAAGTCGGTCCTCAGAGG

Efficacy score by CRISPRater: 0.88 HIGH

Oligo pair with 5' extension fwd: TAggAAAGGTAAGTCGGTCCTCAG rev: AAACCTGAGGACCGACTTACCTTT

Oligo pair with 5' substitution fwd: TAggAGGTAAGTCGGTCCTCAG rev: AAACCTGAGGACCGACTTACCT

Top 20 offtarget sites out of 26 (including on target: for full list see xls file)

Coordinates	strand	MM	target_seq	PAM	distance		gene name	gene id
chr7:55027142-55027164	-	0	AAAGGTAA[GTCGGTCCTCAG]	AGG	7777	I	EGFR	ENSG00000146648
chr10:78546776-78546798	+	2	ATAAGTAA[GTCGGTCCTCAG]	AGG	2902	I	RP11-17G2.1	ENSG00000282952
chr8:140586699-140586721	-	3	GAAGCTCA[GTCGGTCCTCAG]	GGG	1388	I	AGO2	ENSG00000123908
chr1:54265443-54265465	+	4	CACGGTAG[GTGGGTCCTCAG]	GGG	7294	I	SSBP3	ENSG00000157216
chr5:122564927-122564949	+	4	AGGGGTCA[GTCTGTCCTCAG]	AGG	64003	-	RP11-166A12.1	ENSG00000251538
chr17:47936151-47936173	-	4	CAAGGGAG[GTGGGTCCTCAG]	AGG	3045	I	RP11-6N17.3	ENSG00000266601
chr3:111114623-111114645	+	4	ATAGTTGA[GTCAGTCCTCAG]	TGG	2252	I	PVRL3	ENSG00000177707
chr22:41552176-41552198	+	4	AGTGGTAA[TTCTGTCCTCAG]	TGG	7570	-	POLR3H	ENSG00000100413
chr1:116511211-116511233	-	4	AAAGCTAG[ATGGGTCCTCAG]	AGG	3302	-	CD58	ENSG00000116815
chr4:22376049-22376071	-	4	AAATTTAA[GACAGTCCTCAG]	AGG	11304	I	ADGRA3	ENSG00000152990
chr5:96535888-96535910	+	4	AAACCTAA[GTATGTCCTCAG]	TGG	6008	I	CAST	ENSG00000153113
chr19:17136029-17136051	-	4	AAAGGGAT[GCCTGTCCTCAG]	GGG	9346	I	MYO9B	ENSG00000099331
chr2:19843200-19843222	-	4	AAGGGTGA[GTAGATCCTCAG]	AGG	16625	-	CISD1P1	ENSG00000213403
chr11:68033671-68033693	-	4	AACGGAAA[GCCGGGCCTCAG]	GGG	0	E	NDUFS8	ENSG00000110717
chr6:167113080-167113102	-	4	CCAGGTAA[GTCAGGCCTCAG]	AGG	1066	I	CCR6	ENSG00000112486
chr17:40063502-40063524	-	4	AAAGGCCA[CTCGGGCCTCAG]	AGG	410	I	THRA	ENSG00000126351
chr4:3262580-3262602	-	3	GAAGGTAA[GGCGGTCCCCAG]	GGG	39	I	MSANTD1	ENSG00000188981
chr19:8633680-8633702	-	4	ATAGCTAA[GTTGGTCTTCAG]	AGG	1055	I	CTD-2586B10.1	ENSG00000268480
chr8:91548573-91548595	+	4	AATGGAAA[GTCTGTCTTCAG]	TGG	2119	I	RP11-122C21.1	ENSG00000253901
chr21:30758603-30758625	-	4	ACCGGTAA[GTTGGTCCTAAG]	AGG	3175	-	KRTAP21-1	ENSG00000187005

Figure 8: The identified target sequence for CRISPR/Cas-9 activity against EGFR gene with efficacy score of 0.88 – output obtained from using CCTop software.

VII. DISCUSSION

The manuscript aims at deciphering the latest molecular biology technique in terms of CRISPR/Cas-9 genetic alteration or modification or edition related to EGFR gene. At the same time, it also envisages the protein characteristics related to the EGFR protein, a biomarker in lung cancer prognosis. EGFR is a protein; overexpression relates to the lung carcinoma and has got clinical implication in diagnosing the lung carcinoma other than histochemical, histopathological or histo-immunological techniques. The EGFR protein is encoded by the gene termed as "EGFR", and its over expression translates it into EGFR protein. We tried to explore the concept of CRISPR/Cas-9 system to decipher the genetic alteration of the EGFR gene using computational tool like that of SYNTHEGO and CCTop to generate the target sequence as well as the PAM for each sequence of the EGFR gene. Our result in terms of EGFR protein characteristic like that of hydropathy index and the polarity envisages that the protein is a transmembrane spanning between the inner and outer domain of the membrane as evident from the Figures 1 and 2. The hydropathy index indicates that the most of the amino acid composition of the EGFR protein is hydrophobic in nature. Furthermore, we attempted to elucidate the single guide RNA sequence corresponding to the target sequence as referred to here as the DNA site sequence as evident from Table 1. For each SgRNA sequence we attempted to elucidated the possible target sequence and PAM sequence generated from the two computational tools (SYNTHEGO and CCTop) with highest efficacy score as shown in the Figure(s) 6, 7 and 8. The sequence reflected in the manuscript is needed to be processed in the wet lab by transfecting the designed SgRNA sequence in to plasmid and validating the same for precision and accuracy cutting or edition of the said target sequence asrecognized by the corresponding PAM sequence.

VIII. FUTURE DIRECTIONS

Our aim in this manuscript was to study the structure and the properties of the EGFR protein encoded by EGFR gene as well as to generate the target sequence with appropriate guide RNAs for genetic alteration or cutting or modification employing CRISPR/Cas9 genetic tool. The technique got more relevance and importance after being awarded with Noble prize by the investigators. Future research is needed to achieve the precision and accuracy of identifying the target location among the million-base pair of a gene and cutting the exact target sequence and not generating the off-target sequence is a million-dollar question that needs to be answered or investigated further to achieve its full potential in using as a therapeutic measure against all form of cancer and genetic disorders. It is well known fact that DNA bulges after certain base pairs that may get unrecognized by the Cas-9 system and therefore may generate off target sequence which shall be more detrimental rather than to be useful [23].

IX. CONCLUSION

Our paper identifies the best sequences of the EGFR 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 welfare [24]. We have attempted to contribute to the existing knowledge of science using some computational tools to aid the advances with limited resources since we are unable to visit our labs during this pandemic. We aspire to keep working using computational tools and work on the demerits of existing technology to make it desk to bed readily.

ACKNOWELEDGEMENT

The authors of this manuscript would like to shower their heartfelt acknowledgement to the Honourable Chancellor, Professor (Dr.) Goutam Roy Choudhury, of Techno India University, West Bengal for giving the opportunity to work and present the paper in the dynamic field of science that was acclaimed worldwide and awarded with Noble prize for the year 2020 in Chemistry.

N.B: *The manuscripts highlight the cutting edge technology of genetic editing as a promising field of discovery in molecular medicine by "***Jennifer Doudna and Emmanuelle Charpentier***" for which they together were awarded with* ***Noble Prize in Chemistry*** *for the year 2020. The manuscript is an interface between molecular biology and computer algorithm. The authors of the manuscript are thankful to the editors and the reviewers of the esteemed journal to provide the opportunity to present this unique research article of CRISPR/Cas-9 system interfaced with computer algorithm in order to promote interdisciplinary research as well as to highlight the latest knowhow in the field of molecular biology. The authors feel that few journals provide opportunities to the budding researcher to showcase their talent in the latest cutting edge technology of any discipline of science, for which the authors are thankful to the general administration of the esteemed journal.*

REFERENCES

[1] Zhang, J., Adikaram, P., Pandey, M., Genis, A. and Simonds, W., 2016. Optimization of genome editing through CRISPR-Cas9 engineering. Bioengineered, 7(3), pp.166-174.

[2] Jiang, F., & Doudna, J. A. (2017). CRISPR-Cas9 Structures and Mechanisms. Annual review of biophysics, 46, 505–529.

[3] De, A., & Biswas, A. (2020). CRISPR/Cas-9 Genetic Editing of 'Neu' gene in Breast Cancer Prognosis. International Journal for Innovative Research in Multidisciplinary Field.

[4] Dominguez, A. A., Lim, W. A., & Qi, L. S. (2016). Beyond editing: repurposing CRISPR-Cas9 for precision genome regulation and interrogation. Nature reviews. Molecular cell biology, 17(1), 5–15.

[5] Shalem, O., Sanjana, N. E., & Zhang, F. (2015). High-throughput functional genomics using CRISPR-Cas9. Nature reviews. Genetics, 16(5), 299–311.

[6] Bethune, G., Bethune, D., Ridgway, N., & Xu, Z. (2010). Epidermal growth factor receptor (EGFR) in lung cancer: an overview and update. Journal of thoracic disease, 2(1), 48–51.

[7] Ellison, G., Zhu, G., Moulis, A., Dearden, S., Speake, G. and McCormack, R., 2020. EGFR Mutation Testing In Lung Cancer: A Review Of Available Methods And Their Use For Analysis Of Tumour Tissue And Cytology Samples.

[8] Gupta, R., Dastane, A., Forozan, F., Riley-Portuguez, A., Chung, F., Lopategui, J. and Marchevsky, A., 2008. Evaluation of EGFR abnormalities in patients with pulmonary adenocarcinoma: the need to test neoplasms with more than one method. Modern Pathology, 22(1), pp.128-133.

[9] Goenka, M., De, A., & Biswas, A. R., Dr. (2020). Role of CRISPR/Cas9 in Genetic Manipulation of ROS1 and EGFR Genes using Synthego Platform. Volume 5 - 2020, Issue 9 - September International Journal of Innovative Science and Research Technology, 5(9), 1080-1085.

[10] McDade, J. (n.d.). Components of CRISPR/Cas9. Retrieved from https://blog.addgene.org/components-of-crispr/cas9-our-new-crispr-101-ebook

[11] Terns, Michael P., and Rebecca M. Terns. "CRISPR-Based Adaptive Immune Systems." Current Opinion in Microbiology, vol. 14, no. 3, June 2011, pp. 321–27.

[12] Hille, Frank, et al. "The Biology of CRISPR-Cas: Backward and Forward." Cell, vol. 172, no. 6, Mar. 2018, pp. 1239–59.

[13] Damodharan, L., &Pattabhi, V. (2004). Hydropathy analysis to correlate structure and function of proteins. Biochemical and biophysical research communications, 323(3), 996–1002.

[14] Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. Journal of molecular biology, 157(1), 105–132.

[15] Gallo, E., & Gellman, S.H. (1993). Hydrogen-bond-mediated folding in depsipeptide models of .beta.-turns and .alpha.-helical turns. Journal of the American Chemical Society, 115, 9774-9788.

[16] Zimmerman, J. M., Eliezer, N., &Simha, R. (1968). The characterization of amino acid sequences in proteins by statistical methods. Journal of theoretical biology, 21(2), 170–201.

[17] Design.synthego.com. 2020. Synthego. [online] Available at: https://design.synthego.com/#/

[18] Stemmer, M., Thumberger, T., Del Sol Keyer, M., Wittbrodt, J., & Mateo, J. L. (2015). CCTop: An Intuitive, Flexible and Reliable CRISPR/Cas9 Target Prediction Tool. PloS one, 10(4), e0124633.

[19] Abadi, S., Yan, W. X., Amar, D., &Mayrose, I. (2017). A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action. PLoS computational biology, 13(10), e1005807.

[20] Synthego.com. 2021. Synthego | Full Stack Genome Engineering. [online] Available at: <https://www.synthego.com/company>

[21] Synthego.com. 2021. Synthego | Full Stack Genome Engineering. [online] Available at: <https://www.synthego.com/guide/how-to-use-crispr/design-tool-tutorial#:~:text=How%20Does%20the%20Synthego%20CRISPR%20Design%20Tool%20Work%3F&text=The%20tool%20further%20narrows%20down,insertions%20or%20deletions%20(indels).>

[22] Labuhn, M., Adams, F. F., Ng, M., Knoess, S., Schambach, A., Charpentier, E. M., Schwarzer, A., Mateo, J. L., Klusmann, J. H., &Heckl, D. (2018). Refined sgRNA efficacy prediction improves large- and small-scale CRISPR-Cas9 applications. Nucleic acids research, 46(3), 1375–1385.

[23] De, A.,& Biswas, A. (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. 872-876.

[24] Annunziato, S., Lutz, C., Henneman, L., Bhin, J., Wong, K., Siteur, B., van Gerwen, B., de Korte-Grimmerink, R., Zafra, M. P., Schatoff, E. M., Drenth, A. P., van der Burg, E., Eijkman, T., Mukherjee, S., Boroviak, K., Wessels, L. F., van de Ven, M., Huijbers, I. J., Adams, D. J., Dow, L. E., … Jonkers, J. (2020). In situ CRISPR-Cas9 base editing for the development of genetically engineered mouse models of breast cancer. The EMBO journal, 39(5), e102169.