2017 International Conference on circuits Power and Computing Technologies [ICCPCT]

Off-target possibility of efficient siRNA using String matching

Prameela S.
Department of Computer Science
Cochin University of Science And Technology
Cochin-22, India
prameela@cusat.ac.in

Jereesh A. S.
Department of Computer Science
Cochin University of Science And Technology
Cochin-22, India
jereesh@cusat.ac.in

Abstract— **RNA interference (RNAi) is a gene silencing mechanism mediated by small interfering RNA (siRNA). Gene silencing can successfully be used for drug design. The inhibition capacity of the target gene and off-target effect of all other genes are the two key features that we need to consider for the designing of efficient exogenous siRNA. First generation rules and thermodynamic properties are used to calculate the inhibition score. Calculating the off-target possibility is a challenging issue in siRNA design. Most of the works on the design of siRNA were based on a BLAST search for off-target effect. The proposed system uses string matching algorithm in GPU instead of BLAST. The original experimental value of Huesken data set is used to evaluate the current work.**

Keywords— Messenger RNA, Small Interfering RNA, inhibition capacity, off-target possibility,RISC,GPU,BLAST.

I. INTRODUCTION

The human genome contains roughly 30,000 genes, spread out over 46 chromosomes. Just 1.5 percent of the DNA in the genome really codes for genes. One important function of RNA is coding of genetic materials into proteins. Normally RNAs are single stranded. Some non-coding RNAs are double stranded. The double stranded RNA (dsRNA) used in gene regulation. Small Interfering RNA (siRNA) and Micro RNA (miRNA) are examples of double stranded Non-coding RNA [1]. The central dogma of molecular biology describes how proteins are created from the genetic information. The genetic information in DNA is transcripted to messenger RNA (mRNA), then this mRNA is translated to protein [2].

Gene expression or protein production can be controlled by gene regulation. Gene regulation is achieved by a process of turning genes on and off. Gene silencing is classified as transcriptional gene silencing and post-transcriptional gene silencing. In transcriptional gene silencing, the messenger RNA is not formed and further activities of protein formation are stopped. In post-transcriptional gene silencing, the targeted messenger RNA is lost or degraded [3]. An active mRNA level may be controlled by splicing or by silencing with some of the non-coding RNAs like miRNA (micro RNA) and siRNA (short interfering RNA). Genes can be either up regulated or down regulated. Using down regulation, the expression of a particular gene may be prevented. Gene silencing is done by preventing the expression of a particular gene and thereby turning "off" gene expression.

RNA interference (RNAi) is a gene silencing mechanism. RNAi was discovered in 1998 by Andrew Fire and Craig Mello [4]. RNAi targets the protein producing mRNA and controls disease in the transcription phase by generating a non-coding RNA called siRNA. The process of gene silencing mechanism is divided into the following steps. First, dsRNA(Double Stranded) is processed by Dicer (Ribonuclease III protein) into siRNA which is loaded into the RISC (RNA Induced Silencing Complex). Then, due to RISC activation, siRNA will unwind and separate into a sense (Passenger) and an antisense (guide) strands. After that RISC activates antisense strand. Finally, the binding between full complementary antisense (guide) strand and target mRNA leads to the cleavage of mRNA [1].

In this paper, proposed a methodology for identifying efficient siRNA for any target mRNAs. The inhibition efficiency is used to reduce the search space and the string matching algorithm is used for reducing off target possibility

A. Motivation and Scope

In the case of cancer treatment, siRNA is very helpful to complete removal of the tumor without doing damage to any other parts of the body. Current treatment for cancer are radiation and chemotherapy, but it affects adjacent non-cancer cells and cause various side effects. siRNA based drugs have no side effect compared to traditional drugs. In current treatment, drugs temporarily degrade the protein expression. Drugs based on siRNA can cleave the target mRNA and stop the unwanted protein from that mRNA [5]. siRNAs developed for the diseases such as AIDS, neurodegenerative diseases, cholesterol, and cancer on mice. It can be extended to treatment on human through this approach [6].

978-1- 5090-4967- 7/17/$31.00 © 2017 IEEE

II. Literature Review

Several rules and methods were used in siRNA design.This mainly classified as first generation method and second generation method.

A. First Generation Method.

Reynolds et al.[7], Ui-Ti[8], chalk et al.[9], Hseih et al.[10] , **Amarzguioui et al.[11] and** Takasaki et al.[12] were explaining some first generation rules for siRNA design. Some common rules from this method are explained below.

1. Presence of G/C at position 1
2. Absence of U at position 1
3. Presence of A at position 6
4. Absence of G/C at position 19
5. Presence of A/U at position 19
6. Guanine-Cytosine (G-C) content of around 50%.

B. Second Generation Method.

BIOPREDsi [13], DSIR[14], ThermoComposition21 [15], i-Score [16], MysiRNA [17] are some second generation siRNA design method. BIOPREDsi, ThermoComposition21, MysiRNA Designer package and MysiRNA used the artificial neural network model for siRNA design. DSIR, i-Score and Scales were based on linear regression models. The ThermoComposition21 model used some first generation rules together with thermodynamic properties in Artifitial Nueral Network. TABLE I show the comparison of different siRNA designer models [6].

TABLE I. Comparision of siRNA Design Models.

Designer Model	Methods	correlation coefficient
MysiRNA	ANN models	0.687
DSIR	Linear regression models	0.687
Biopredsi	ANN models	0.665
ThermoComposition21	ANN models	0.635
OpsiD	ANN models	0.727

Jyoti K Shahe et al.[18] consider some rules to design of efficient siRNA. This paper explained a scoring system to calculate the inhibition capacity of each siRNA and Blast used to find the off target possibility of siRNA. String matching algorithm in GPU is used for finding off-target possibility of siRNA.

III. Materials and Methods

A. Dataset.

The data set used for this work was selected from various articles of siRNA design methods. Huesken[12] data set contains the experimental inhibition value of 2431 siRNAs targeting 32 genes.

B. Evaluation Metrics

G-C (Guanine-Cytosine) value: G-C value indicates the stability of siRNA. siRNA with high G-C value is more stable than siRNA with low G-C value.The equation(1) is used to calculate the G-C value.

$$G-C\ Value = \frac{G+C}{A+G+T+C} * 100 \quad (1)$$

Where A(Adenine), T(Thymine), G(Guanine) and C(Cytosine) are bases of DNA. The total number of G and C divides the total length of the sequence.Select the siRNA have a G-C value between 50-75 percentage.

Thermodynamic Property(ΔG): Thermodynamic property indicates the thermal stability of siRNA. TABLE II Nearest Neighbour method used to calculate the value of ΔG. TABLE II shows the ΔG value of the sequence in kcal/Mol. The equation(2) used to calculate the ΔG [19].

$$Whole\ \Delta G = \sum_{k=1}^{K-1} \Delta G37\,(Seq[k]Seq[K+1]) \quad (2)$$

TABLE II. ΔG_{37} Value of Nearest Neighbour Pair.

Nearest Neighbour Pair	ΔG_{37} *(kcal/mol)*	*Nearest Neighbour Pair*	ΔG_{37} *(kcal/mol)*
GA	-2.4	AA	-0.9
GG	-3.3	AG	-2.1
GC	-3.4	AC	-2.2
GU	-2.2	AU	-1.1
CA	-2.1	UA	-1.3
CC	-3.3	UC	-0.9
CG	-2.4	UG	-2.1
CU	-2.1	UU	-2.4

Inhibition Score: The score of siRNAs sequence is depending on some rules and weights [18].The TABLE III shows the weights of each rule. The inhibition score of siRNA is defined by (3)

$$Inhibition\ Score = \sum_{k=1}^{15} W_k A_k \quad (3)$$

Where, W_k are the weights of each rule obtained from the TABLE III and A_K are the binary entries for each rule. If the rule is satisfied, then A_k take the value '1'and it takes '0' if rules are not satisfied.Inhibition score were normalized to the range[0, 100].

TABLE III. WEIGHTS OF SCORING RULES

No.	*Rules for scoring System*	*Weights(W)*
A1	A/U at position 1 of sense strand	-1.4
A2	G/C at position 1 of sense strand	1.11
A3	A at position 6 of sense strand	0.70
A4	U at position 10 of sense strand	0.25
A5	G at position 13 of sense strand	-1.66
A6	U/T at position 13 of sense strand	0.31
A7	A/U at position 16 of sense strand	0.74
A8	A/U at position 17 of sense strand	1.20
A9	A/U at position 18 of sense strand	1.44
A10	A/U at position 19 of sense strand	0.87
A11	G/C at position 19 of sense strand	-1.02
A12	GC content	0.42

Off Target Possibility Prediction: siRNA should only silent target gene and not silent non-target genes. Most of the siRNA design tools used BLAST for off-target prediction. BLAST score of siRNA indicates the total number of nucleotide matches of the siRNA with the database. BLAST score of 19nt siRNA is 19, that means the siRNA is perfect match with any of the genes in the database. That siRNA is not good for gene silencing . The proposed system used string matching algorithm for finding off target prediction. Select the siRNAs with high inhibition capacity and finding matches to all genes in the database. Filter out the siRNA with more than seven continuous match to all genes in the database. The efficient siRNA will have high inhibition score and low off-target effect.

IV. PROPOSED APPROACH

The proposed system takes the cDNA sequence of 32 genes that specified in Huesken dataset as input. First calculates each 19-mer siRNA from 32 genes. Then, filter the siRNA based on some parameters, such as G-C content, ΔG etc. Some first generation rules is also used to filter the siRNA.Then calculate the score of each siRNA and remove the siRNA which does not have the score value in the specified range. That score value is known as the inhibition value of siRNA. String matching algorithms are used to find off-target possibility of siRNA. First, select an siRNA and find matches to all other genes.If the match is greater than or equal to seven, then remove that siRNA from output list.The Fig 1. Shows the system design of proposed system

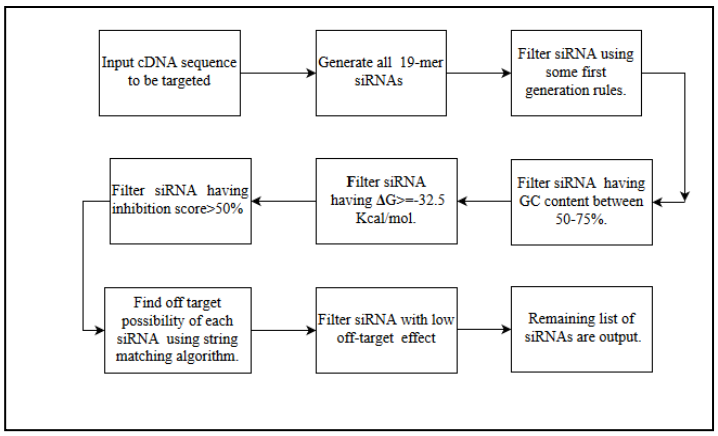

Fig. 1. System Design of Proposed System.

A. Algorithm.

The outline of the algorithm used in proposed approach is given below.

Input: Target Genes.

Output: Effective siRNAs.

1. Start
2. Input cDNA(complementary DNA) sequence of target gene.
3. Generate all 19-mer sequence of siRNAs of target gene.
4. Filter out the siRNA with patterns 'AAAA', 'CCCC', 'GGGG' and 'UUUU'.
5. Calculate the G-C value of siRNAs obtained from step 4 using (1) and select the siRNA having the G-C content between 50-75%.
6. Calculate ΔG value of siRNAs obtained from step 5 using (2) and select the siRNA having ΔG>=-32.5 Kcal/mol.
7. Calculate the inhibition score of siRNAs obtained from step 6 using (3) and select the siRNA having score>=50%.
8. Select effective siRNAs from step7 with match<=6 to all genes using string matching.
9. Stop

First, select the input CDNA sequence according to the disease. In this work 32 genes are used as input. Then generate all 19-mer siRNAs and filter the siRNA using some common first generation rules. For example,to remove the siRNA with patterns 'AAAA', 'CCCC', 'GGGG' and 'UUUU'. Calculate the G-C content of siRNAs and filter the siRNA having the G-C content between 50-75%.Then compute the ΔG value using the nearest neighbor method and select the siRNA having ΔG>=-32.5 Kcal/mol. Then calculate the inhibition score of each siRNA in the remaining list and remove the siRNA having inhibition score less than 50%. Select each siRNA from the remaining list and find off-target possibility with all genes in the database.String matching algorithm is used for the purpose. If the match is greater than or equal to six , then remove the siRNA from the output list.The remaining list of siRNAs are efficient for silencing the target gene.

V. RESULTS AND DISCUSSION

The performance of our proposed system is evaluated with the experimental inhibition values of Huesken dataset1.In this work ,we got 390 siRNA have inhibition score greater than or equal to 80%.The high inhibition score of siRNA means it has the ability to bind the target gene. Most of the siRNA design work has used BLAST for finding off-target effect.This work uses a string matching algorithm in GPU for finding off-target effect. Select the siRNA with less than seven matches to all other human genes, that means it could not bind to other genes.

Many siRNA design tools have already been developed. s-Biopredsi and i-score was used to compare the inhibition score of our method. The Fig 2. Shows the comparative analysis of inhibition scores of s-Biopredsi, i-score, experimental score and our score.X-axis indicates five different siRNAs and Y-axis indicates the inhibition score of that siRNAs. All the effective siRNAs selected through this approach pocesses an inhibition score greater than 80%. It is more accurate to experimental inhibition score.

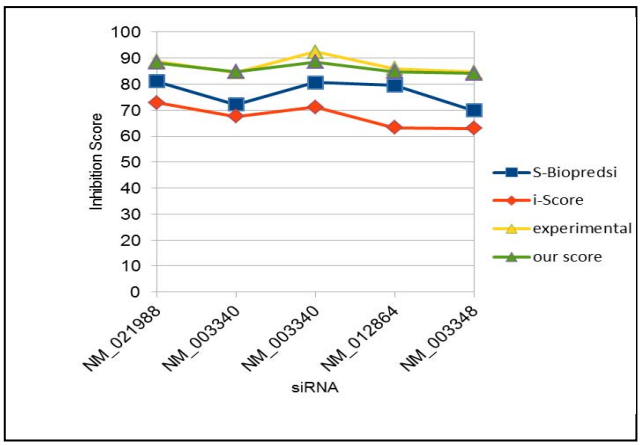

Fig. 2. Comparative analysis of inhibition scores of S-Biopredsi, i-Score, experimental score and our score.

VI. CONCLUSION AND FUTURE WORK

Most of the referred work used BLAST search for finding off-target possibility of siRNA. Proposed system uses string matching algorithm in GPU for finding off- target effect.This model can be used to design exogenous siRNA which can silence the disease causing target gene. The string matching algorithm takes more time to find off-target effect. Parallel computing can improve the performance.

REFERENCES

[1] Lam, J. K., Chow, M. Y., Zhang, Y., & Leung, S. W.. siRNA versus miRNA as therapeutics for gene silencing. Molecular Therapy-Nucleic Acids, 4, e252. 2015.

[2] Crick, Francis. "Central dogma of molecular biology." Nature 227.5258, 561-563, 1970.

[3] Jana, S., Chakraborty, C., Nandi, S., & Deb, J. K.. RNA interference: potential therapeutic targets. Applied microbiology and biotechnology, 65(6), 649-657, 2004.

[4] Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S.E., & Mello, C. C. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. nature, 391(6669), 806-811, 1998.

[5] Borkhardt, A. Blocking oncogenes in malignant cells by RNA interference - New hope for a highly specific cancer treatment?. Cancer cell, 2(3), 167-168, 2002.

[6] Reena Murali , Philips George John , David Peter S. Soft computing model for optimized siRNA design by identifying off target possibilities using artificial neural network model. Gene, 562(2), 152-158,2015

[7] Reynolds, A., Leake, D., Boese, Q., Scaringe, S., Marshall, W. S., & Khvorova, A. Rational siRNA design for RNA interference. Nature biotechnology, 22(3), 326-330, 2004.

[8] Ui-Tei, K., Naito, Y., Takahashi, F., Haraguchi, T., Ohki- Hamazaki, H., Juni, A., Ueda, R., & Saigo, K. Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. Nucleic Acids Research, 32(3), 936–948, 2004.

[9] Chalk, A.M., Wahlestedt, C., & Sonnhammer, E.L.L. Improved and automated prediction of effective siRNA. Biochemical and Biophysical Research Communications,319(1), 264–274, 2004.

[10] Hsieh A.C., Ronghai B., Manola J., Vazquez F., Bare O., Khvorova A., Scaringe S., and Sellers W.R. A library of siRNA duplexes targeting the phosphoinositide 3-kinase pathway: determinants of gene silencing for use in cell-based screens. Nucleic Acids Research, 32:893–901, 2004.

[11] Amarzguioui, M., & Prydz, H. An algorithm for selection of functional siRNA sequences. Biochemical andbiophysical research communications, 316(4), 1050-1058,2004.

[12] Takasaki, S., Kotani S., and Konagaya A. An Effective Method for Selecting siRNA Target Sequences in Mammalian Cells. Cell Cycle, 3, 790–795, 2004.

[13] Huesken, D., Lange, J., Mickanin, C., Weiler, J., Asselbergs, F., Warner, J., Hall, J. Design of a genome-wide siRNA library using an artificial neural network. Nature biotechnology, 23(8), 995-1001, 2005.

[14] Vert, J. P., Foveau, N., Lajaunie, C., & Vandenbrouck, Y. An accurate and interpretable model for siRNA efficacy prediction. BMC bioinformatics, 7(1), 520, 2006.

[15] Shabalina, S. A., Spiridonov, A. N., & Ogurtsov, A. Y. Computational models with thermodynamic and composition features improve siRNA design. BMC bioinformatics, 7(1), 65, 2006.

[16] Ichihara, M., Murakumo, Y., Masuda, A., Matsuura, T., Asai, N., Jijiwa, M., Ishida, M., Ishida, M.M., Shinmi, J., Yatsuya, H., Qiao, S., Takahashi, M., & Ohno, K. Thermodynamic instability of siRNA duplex is a prerequisite for dependable prediction of siRNA activities. Nucleic acids research, 35(18), e123, 2007.

[17] Mysara, M., Elhefnawi, M., & Garibaldi, J. M. MysiRNA: Improving siRNA efficacy prediction using a machine-learning model combining

multi-tools and whole stacking energy (ΔG). Journal of biomedical informatics, 45(3), 528-534, 2012.

[18] Shah, J. K., Garner, H. R., White, M. A., Shames, D. S., & Minna, J. D. sIR: siRNA Information Resource, a web-based tool for siRNA sequence design and analysis and an open access siRNA database. BMC bioinformatics, 8(1), 178,2007.

[19] Murali, R., & David Peter, S. Computational Model for Predicting Effective siRNA Sequences Using Whole Stacking Energy (% G) for Gene Silencing. World Academy of Science, Engineering and Technology, International Journal of Biological, Biomolecular,Agricultural, Food and Biotechnological Engineering, 9(1), 6-12,2015.