International Journal of Biological, Life and Agricultural Sciences
ISSN: 2415-6612
Vol:9, No:1, 2015

Computational Model for Predicting Effective siRNA Sequences Using Whole Stacking Energy (Δ G) for Gene Silencing

Reena Murali, David Peter S.

***Abstract*—**The small interfering RNA (siRNA) alters the regulatory role of mRNA during gene expression by translational inhibition. Recent studies show that upregulation of mRNA because serious diseases like cancer. So designing effective siRNA with good knockdown effects plays an important role in gene silencing. Various siRNA design tools had been developed earlier. In this work, we are trying to analyze the existing good scoring second generation siRNA predicting tools and to optimize the efficiency of siRNA prediction by designing a computational model using Artificial Neural Network and whole stacking energy (ΔG), which may help in gene silencing and drug design in cancer therapy. Our model is trained and tested against a large data set of siRNA sequences. Validation of our results is done by finding correlation coefficient of experimental versus observed inhibition efficacy of siRNA. We achieved a correlation coefficient of 0.727 in our previous computational model and we could improve the correlation coefficient up to 0.753 when the threshold of whole tacking energy is greater than or equal to -32.5 kcal/mol

***Keywords*—**Artificial Neural Network, Double Stranded RNA, RNA Interference, Short Interfering RNA.

I. Introduction

IN central dogma of molecular biology, DNA is first transcribed into messenger RNA (mRNA). The information for a particular gene is encoded in mRNA, and mRNA acts as a template for protein production. A gene is expressed meaning that the information encoded in the mRNA of that gene is converted into amino acid sequences. This reveals the regulatory role of mRNA in gene expression. The normal role of gene regulation of mRNA may be altered which leads up or down regulations. This up and down regulation of mRNA may cause several diseases like Cancer. Gene silencing is a mechanism to control the regulatory role of mRNA. Recent studies show that non-protein coding RNAs such as microRNA (miRNA) and short interfering RNA (siRNA) play an important role in gene silencing, cancer diagnosis and therapy.

RNA interference (RNAi) is biological process by which selective gene silencing can be done by inducing exogenous siRNA capable of degrading the target mRNA. Selective gene silencing is widely useful in gene expression analysis and functional genomics. The short RNA species called siRNAs are formed naturally from double stranded RNA (dsRNA) or are synthesized externally and then introduced into the cell. siRNA, when activated with RNA induced silencing complex (RISC) degrades complementary mRNA sequences. This is called mRNA knockdown by siRNA. This knockdown prevents mRNA from producing amino acid sequences which are responsible for gene expression. Thus gene expression can be altered by siRNAs which are efficient enough to do translational inhibition. Here in designing siRNA with good knockdown efficacy play an important role in cancer detection and diagnosis.

Numerous siRNA design tools had been introduced earlier to synthesize possible siRNAs targeting the mRNAs. But different studies indicate that out of the possible siRNAs that can be synthesized against a particular target, only a fraction of these are successful in causing any degradation and all siRNAs do not result in equal knockdown effects [1]. The efficacy of the siRNAs differed among different target sites in the same target mRNA. Therefore, it is important to select effective siRNA sequences that are highly functional in causing more than a certain percentage of the target mRNA sequence to degrade. In most studies, siRNAs causing knockdown of more than 75 percentage of the target mRNA are considered highly efficient but the threshold varies depending on the level of silencing required. Thus the goal of siRNA efficacy prediction is to aid in designing siRNA sequences that are highly efficient against their target mRNA sequences.

A. RNAi Pathway

The RNAi pathway was discovered by Fire and Mello in 1998 [2]. RNAi is a biological process of post-transcriptional gene silencing mechanism [3]. It helps in developing various therapeutic applications because of its ability to do specific target silencing [4]. Genes causing diseases can be controlled during gene expression by transcriptional, post-transcriptional, and post translational intervention. Drugs for disease control have been targeted towards proteins, which occurs in the post translational phase. RNAi mainly targets the protein producing mRNA and can thereby control disease earlier in the transcription phase. RNAi has been successfully used to target diseases such as AIDS [5], neurodegenerative diseases [6], cholesterol [7] and cancer [8] on mice with the hope of extending these approaches to treat humans.

Reena Murali is Associate Professor with the Department of Computer Science and Engineering, Rajiv Gandhi Institute of Technology, Kerala, India (corresponding author e-mail: reena.rajesh@rit.ac.in).

David Peter S. is Professor with the Department of Computer Science and Engineering, Cochin University of Science and Technology, Kerala, India.

II. Existing Rules for siRNA Design

Even though several algorithms and methods have been developed to predict efficiency of siRNA, only a few of them have achieved an acceptable level of specificity and sensitivity. These algorithms are classified into two groups; first generation and second generation methods.

A. First Generation Tools

The first generation tools [9]-[16] select the most efficient siRNAs based on secondary structure, thermodynamic properties, target positions, and so on. But the results shown that they have a low prediction accuracy of only up to 65%, compared with experimentally proven data with 90% inhibition capacity. Also nearly 20% of the sequences were found inactive [17]. The following sections briefly summarize the results of several studies.

1) Amarzguioui Method - Study by Amarzguioui et al. [9] follows a scoring method identified by different set of rules. They studied 46 siRNAs, and identified some important features of the 19 nt siRNA that correlates with knockdown of more than 70. In this study, functionality is indicated by a knockdown of 70.
2) Tuschl Rules - This technique is widely used for designing effective siRNAs. According to this algorithm [12], synthesizing siRNA duplexes of lengths 21 nt with 19 nt base-paired sequence with 2 nt 3' overhang at both ends mediates efficient cleavage of target mRNA. According to this study, target sequence should have a GC content of around 50 percent.
3) Reynolds Rules - Reynolds et al. [13] analyzed a set of 180 siRNAs. They divided the siRNAs in to different groups based on their functionality to find properties with high correlation to functionality. Also they described a set of eight rules governing the siRNA sequence that are highly indicative in determining the extent of mRNA knockdown. This algorithm assigns a score based on the number of rules satisfied and siRNAs satisfying 6 or more rules are predicted to be functional.
4) Stockholm Rules - This prediction algorithm by Chalk et al. [14] incorporates the thermodynamic properties of the siRNA. Using a scoring scheme that adds 1 for each rule satisfied, and a cutoff score of 6, efficient siRNAs can be detected. They further analyzed the siRNAs using the regression tree technique, but the energy parameters which were found to be statistically significant in their study did not get chosen as important features by this method.
5) Ui-Tei Rules - Ui-Tei et al. [15] analyzed 62 targets in mammalian cells and Drosophila cells and came up with four features which siRNAs should simultaneously satisfy to cause efficient silencing. These rules were found applicable to mammalian cells but did not apply to Drosophila cells.
6) Hseih Rules - Hseih et al. [16] identify the following features which distinguish effective and ineffective RNAi.

- Target sequences that are in the middle of the coding sequence resulted in significantly less silencing.
- Silencing by duplexes targeting the 3 untranslated region (UTR) is comparable with duplexes targeting the coding sequence.
- Pooling of four or five duplexes per gene results in highly efficient silencing.
- siRNA sequences seen to produce more than 70G or C in position 11 and T in position 19.

B. Second Generation Tools

Because of some limitations in siRNA efficacy prediction of first generation tools, there was a need to develop techniques to improve the efficiency of predicted siRNA. These second generation models are based on either artificial neural network or linear regression model. Some of the good scoring second generation tools like Biopredsi [18], DSIR [19], ThermoComposition21 [20], i-Score [21], Scales [22], My siRNA-Designer [23], MysiRNA [24] were developed by introducing data mining techniques to improve the efficiency of siRNA with their experimental inhibition. Biopredsi, ThermoComposition21 MysiRNA-Designer package and MysiRNA used the Artificial Neural Network model. DSIR, i-Score and Scales used simplified linear regression model. In Biopredsi a reasonable amount of accuracy is obtained using 'Huesken' (Novartis) dataset [18]. ThermoComposition21 improved the prediction accuracy by combined position dependent features together with thermodynamic features in one artificial neural network model. The prediction accuracy is improved in DSIR, i-Score and Scales using linear regression model. Further the MysiRNA-Designer package and MysiRNA much improved the prediction accuracy by artificial neural network model.

III. Materials and Methods

A. Data Sets

To train our model we have used the Huesken dataset (Data Set A), which consists of 2431 siRNAs with their experimental inhibition efficiency. Many good scoring second generation tools like Biopredsi, DSIR, ThermoComposition21, i-Score, MysiRNA-Designer and MysiRNA used this data set to train their data. Another dataset (Data Set B) used as test data set consisted of 419 siRNA which was collected by Ichihara [21] from five different publications: Reynolds [13], Ui-Tie (15), Khovorova [25], Haborth [26] and Vickers [27] in the development and evaluation of the i-Score software. Both these data sets were used by us to validate the results.

B. Parameter Selection

All the existing siRNA design tools use different features and weights in their model design. We have used an attempt to combine these features for improving the design. In our previous model [28], we have considered Biopredsi, ThermoComposition21, i-Score, DSIR, MysiRNA as parameters to our model. We have extended our previous model by considering one of the important thermodynamic properties called Whole Stacking Energy (ΔG). When Whole Stacking Energy is combined with these parameters using

Artificial Neural Network Model, a considerable improvement in the prediction accuracy was obtained.

C. Neural Network Model

A multi-layer perceptron, feed-forward neural network trained using the Resilient Propagation (RProp) algorithm is used for computing the final score. The neural network which we use has 6 neurons in the input layer; three hidden layers of 8 neurons each and 1 neuron in the output layer (Fig. 1). The neural network was built and trained using Neuroph Studio and integrated into our siRNA designer tool. The serialized neural network model, and the normalization parameters which were used, are provided along with our designer tool. The Neuroph library for Java is used to create and use the siRNA designer neural network model. Neuroph is a lightweight Java neural network framework to develop common neural network architectures. The Neuroph Studio IDE provided by Neuroph was used to easily design and test the model. The IDE provides an easy-to-use graphical interface to design various neural network configurations, and train/test the network using various neural network training algorithms. It is available under version 2.0 of the Apache License. Working Model of the algorithm is shown in Fig. 2.

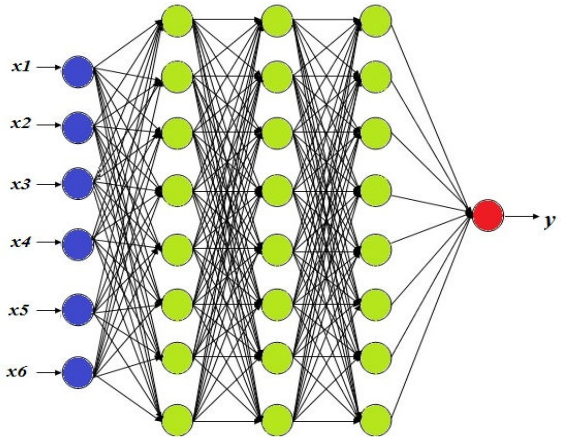

Fig. 1 Neural Network Model

IV. siRNA Designer Workflow

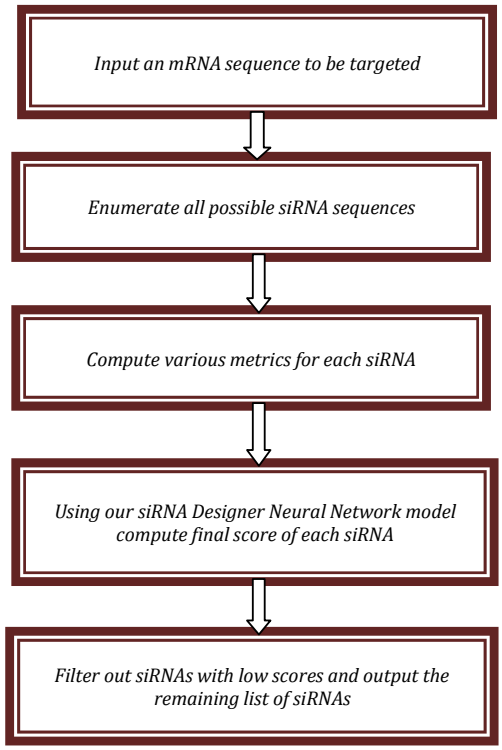

Fig. 2 Working Model of our Algorithm

V. Results

Inhibition capacity of siRNA for a targeted mRNA has been observed with our predicted model (Fig. 3). Also comparison between inhibition activities (Experimental versus Observed) for Huesken dataset (Data Set A) by each of the five good scoring tools (Biopredsi, DSIR,ThermoComposition21, i-Score and MysiRNA) with our model has been done. Pearson correlation coefficient (R) was calculated for each of the six scoring tools. We got a Pearson correlation coefficient of R= 0.727 for Data Set A in [28] and we could achieve an improved correlation coefficient of R =0.753 when the threshold of whole tacking energy is greater than or equal to -32.5 kcal/mol, which shows improvement in the performance compared to the other five models (Table I, Figs. 4, 5).

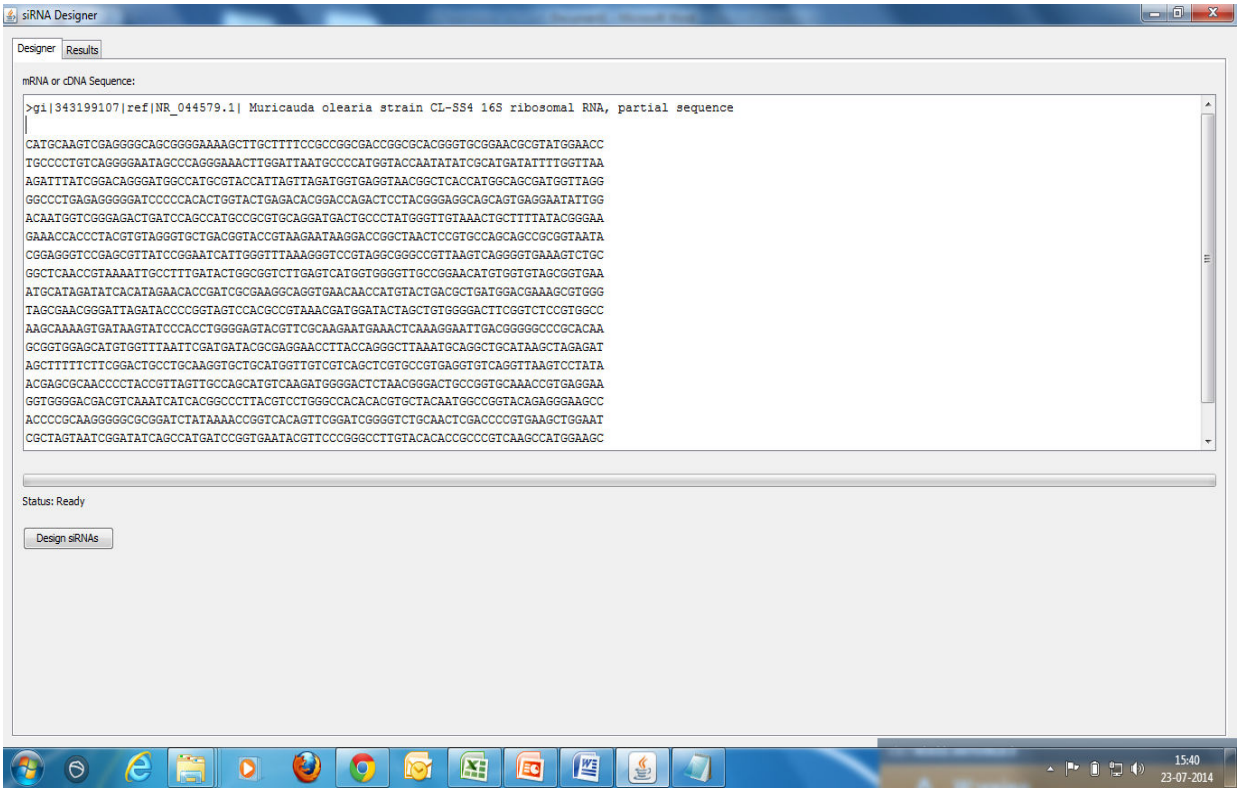

Fig. 3 Sample Screen Shot

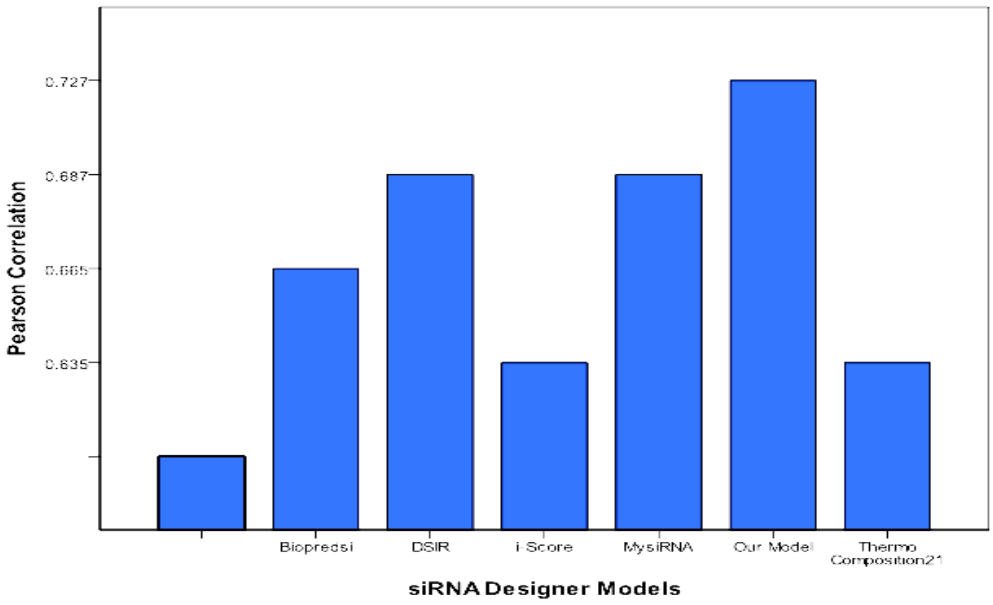

Fig. 4 Comparison between the Second Generation Models and Our Model using Pearson Correlation Analysis. Pearson Correlation Coefficient of Our model showed improvement in the performance compared to the other five models

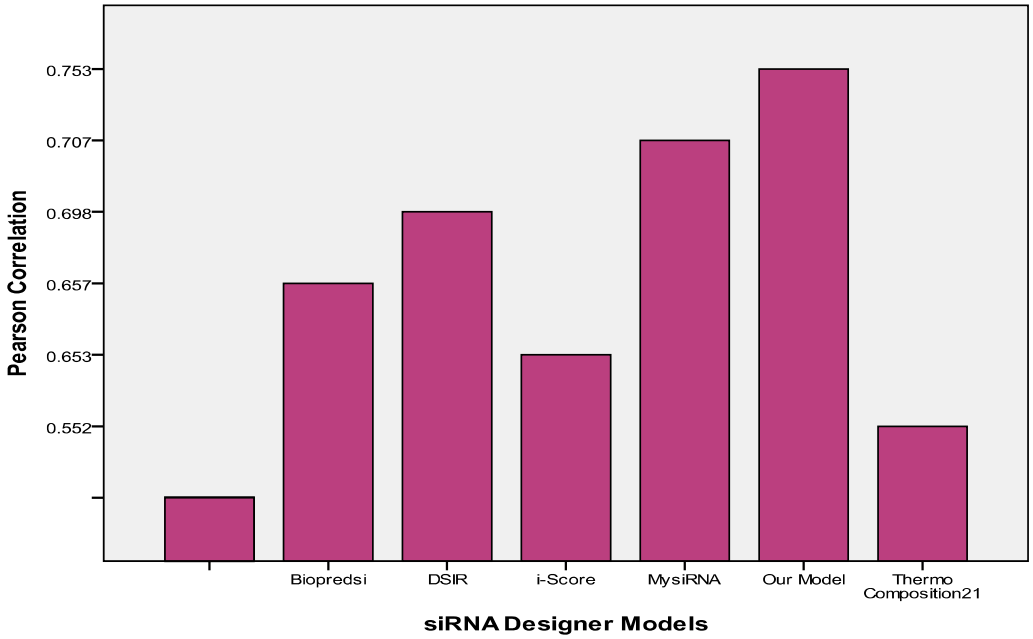

Fig. 5 Comparison between the Second Generation Models and Our Model designed with whole stacking energy, delta G >=-32.5 kcal/mol, using Pearson Correlation Analysis. Pearson Correlation Coefficient of Our model showed improvement in the performance compared to the other five models

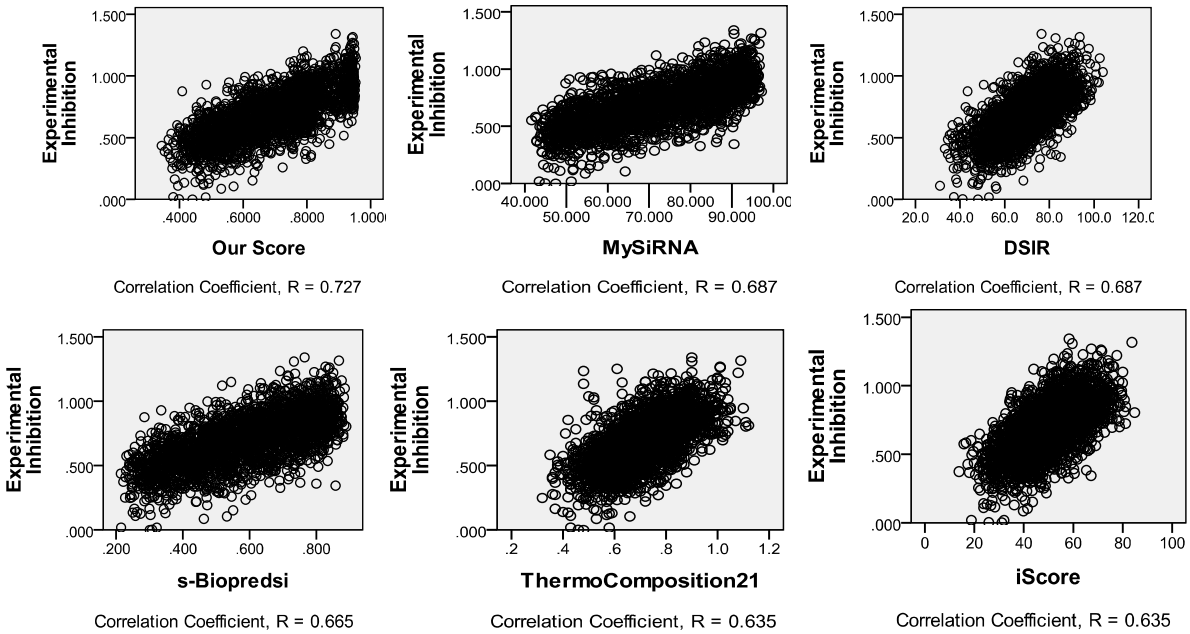

Fig. 6 Experimental siRNAs activities of Dataset A were plotted against the predicted siRNAs activities by each of the second generation tools (s-Biopredsi, DSIR, ThermoComposition21, i-Score and MysiRNA) together with Our model. Pearson correlation coefficient (R) was also shown for each of the six scoring tools

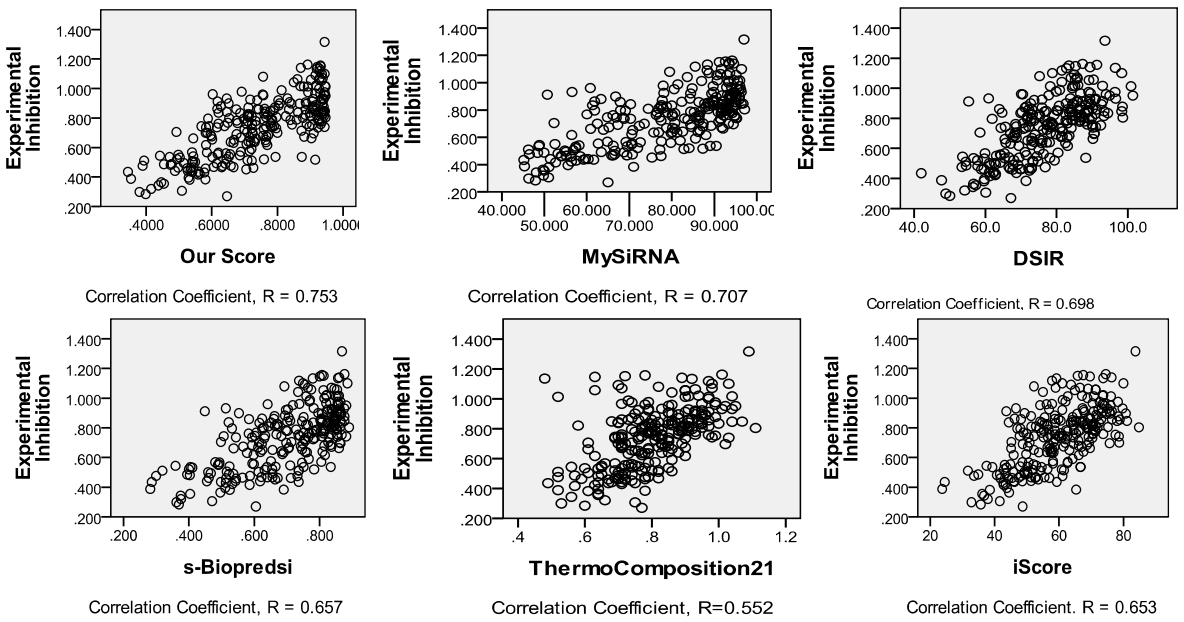

Fig. 7 Experimental siRNAs activities of Dataset A with whole stacking energy Delta G >=-32.5 kcal/mol were plotted against the predicted siRNAs activities by each of the second generation tools (s-Biopredsi, DSIR, ThermoComposition21, i-Score and MysiRNA) together with Our model. Pearson correlation coefficient (R) was also shown for each of the six scoring tools

VI. Validation

The efficiency of the developed model is tested with Huesken dataset (Data Set A) consisted of 2431 siRNA and the Test Data set (Data Set B) consisted of 419 siRNA, mentioned in data set description. Also we have done a comparative analysis with other second generation algorithms like Biopredsi, DSIR, ThermoComposition21, i-Score and MysiRNA. The experimentally proven siRNA activity was plotted against the predicted activity by all these five previous techniques along with our model (Figs. 6 and 7). Accuracy of siRNA prediction was validated using Pearson Correlation Coefficient.

TABLE I
Pearson Correlation Coefficient of siRNA Designer Models

Designer Model	Correlation Coefficient (R)	Correlation Coefficient (R) when ΔG > = -32.5 kcal/mol
Our Model	0.727	0.753
MysiRNA	0.687	0.707
DSIR	0.687	0.698
Biopredsi	0.665	0.657
ThermoComposition21	0.635	0.552
i-Score	0.635	0.653

VII. Conclusion and Future Work

In this work, an improved computational model is designed using one of the important thermodynamic property of siRNA called whole stacking energy (Delta G) and Artificial Neural Network model to predict siRNA inhibition activity based on five previous second generation models s-Biopredsi, DSIR, ThermoComposition21, i-Score and MysiRNA. The prediction accuracy is improved compared to all these previous models. The improvement in Pearson correlation coefficient shows better performance of our model with previous good scoring siRNA design models. This improvement in performance may help in gene silencing and there by cancer diagnosis and drug design. In our future work, we are trying to further improve and optimize the sensitivity and specificity which can address the off target effects of siRNA.

Reference

[1] T. Holen, M. Amarzguioui, M.T. Wiiger, E. Babaie, and H. Prydz, "Positional effects of short interfering RNAs targeting the human coagulation trigger tissue factor", in Nucleic Acids Res, vol. 30(8), 2002, pp. 1757 - 1766.

[2] A. Fire, S. Xu, M.K. Montgomery, S.A. Kostas, S.E. Driver, and C.C. Mello, "Potent and specific genetic interference by double stranded RNA in c. elegans", in Nature, vol. 391, 1998, pp. 806 - 811.

[3] C. Cogoni, and G. Macino, "Post-transcriptional gene silencing across kingdoms", in Genes Dev, vol. 10, 2000, pp. 638 – 643.

[4] H. Shi, A. Djikeng, T. Mark, E. Wirtz, C. Tschudi, and E. Ullu, "Genetic interference in trypanosoma brucei by heritable and in-ducible double-stranded RNA", in RNA, vol. 6(7), 2000, pp. 1069 - 1076.

[5] M.A. Martinez, A. Gutierrez, M. Armand-Ugon, J. Blanco, M. Parera, J. Gomez, B. Clotet, and J.A. Este, "Suppression of chemokine receptor expression by RNA interference allows for inhibition of HIV-1 replication ", in AIDS, vol. 16(18), 2002, pp. 2385 - 2390.

[6] H. Xia, Q. Mao, S.L. Eliason, S.Q. Harper, I.H. Martins, H.T. Orr, H.L. Paulson, L. Yang, R.M. Kotin, and B.L. Davidson, "Rnai suppresses polyglutamine induced neurodegeneration in a model of spinocerebellar ataxia", in Nature Medicine, vol. 10, 2004, pp. 816 - 820.

[7] J. Soutschek, A. Akinc, B. Bramlage, K. Charisse, R. Constien, M. Donoghue, S. El-bashir, A. Geick, P. Hadwiger, J. Harborth, M. John, V. Kesavan, G. Lavine, R.K. Pandey, T. Racie, K.G. Rajeev, I. Rohl, I. Toudjarska, G. Wang, S. Wuschko, D. Bumcrot, V. Koteliansky, S. Limmer, M. Manoharan, and H.P. Vornlocher, "Therapeutic silencing of an endogenous gene by systemic ad¬ministration of modified sirnas ", in Nature, vol. 432, 2004, pp. 173 -178.

[8] A. Borkhardt, "Blocking oncogenes in malignant cells by RNA interference new hope for a highly specific cancer treatment? ", in Cancer Cell, vol. 2(3), 2002, pp. 167 – 168.

[9] M. Amarzguioui, and H. Prydz, "An algorithm, for selection of functional sirna sequences", in Biochem Biophys Res Commun, vol. 316(4), 2004, pp. 1050 - 1058.
[10] T.Tuschl, "RNA interference and small interfering RNAs", in Chembiochem, . vol. 2(4), 2001, pp. 239 - 241.
[11] J. Martinez, A. Patkaniowska, H. Urlaub, R. Luhrmann, and T. Tuschl, "Single-stranded antisense sirnas guide target rna cleavage in rnai", in Cell, vol. 110(5), 2002, pp. 563 - 574.
[12] S.M. Elbashir, W. Lendeckel, and T. Tuschl, "RNA interfernce is mediated by 21 and 22nucleotide RNAs", in Genes and Development, vol. 15, 2001, pp. 188 - 200.
[13] A. Reynolds, D. Leake, Q. Boese, S. Scaring, W. Marshall, and A. Khvorova, "Rational siRNA design for RNA interference", in Nature Biotechnology, vol. 22(3), 2004, pp. 326 – 330.
[14] A.M. Chalk, C. Wahlestedt, and E.L. Sonnhammer, " Improved and automated prediction of effective sirna",. in Biochem. Biophys.Res. Commun, vol. 319, 2004, pp. 264 - 274.
[15] K. Ui-Tei, Y. Naito, F. Takahashi, T. Haraguchi, H. Ohki-Hamazaki, A. Juni, R. Ueda, R. and K Saigo, "Guidelines for the selection of highly effective sirna sequences for mammalian and chick rna interference", in Nucleic Acids Res., vol. 32, 2004, pp. 936 -948.
[16] A.C. Hsieh, R. Bo, J. Manola, F. Vazquez, O. Bare, A. Khvorova, S. Scaringe, and W.R.Sellers, " A library of siRNA duplexes targeting the phosphoinositide 3-kinase pathway: determinants of gene silencing for use in cell-based screens", in Nucleic Acids Res., vol. 32(3), 2004, pp.893 - 901.
[17] Y.Ren, W. Gong, Q. Xu, and X. Zheng, "siRecords: an extensive database of mammalian siRNAs with efficacy ratings", in Access, vol. 22(8), 2006, pp. 1–10
[18] D. Huesken, J. Lange, C. Mickanin, J. Weiler, F. Asselbergs, J. Warner, B. Meloon, S. Enge, A. Rosenberg, D. Cohen, M. Labow, M.Reinhardt, F.Natt and J.Hall, "Design of a genome-wide siRNA library using an artificial neural network", in Nat Biotechnology, vol. 23, 2006, pp. :995–1002.
[19] J.P. Vert, N. Foveau, C. Lajaunie, and Y. Vandenbrouck, "An accurate and interpretable model for siRNA efficacy prediction", in BMC Bioinformatics, vol. 7(520), 2006, pp. 1-17.
[20] S.A. Shabalina, A.N. Spiridonov, and A.Y. Ogurtsov, " Computational models with thermodynamic and composition features improve siRNA design", in BMC Bioinformatics, vol. 7(65), 2006, pp.1-16
[21] M.Ichihara, Y. Murakumo, A. Masuda, T. Matsuura, N. Asai, M. Jijiwa, M. Ishida, J. Shinmi, H.Yatsuya, S. Qiao, M. Takahashi and K. Ohno, "Thermodynamic instability of siRNA duplex is a prerequisite for dependable prediction of siRNA activities", in Nucleic Acids Research, vol. 35(18), 2007, pp. 1–10.
[22] O.Matveeva, Y. Nechipurenko, L. Rossi, B. Moore, A.Y. Ogurtsov, J.F. Atkins, P.Saetrom and S.A. Shabalina, "Comparison of approaches for rational siRNA design leading to a new efficient and transparent method", in Access, vol. 35, 2007, pp.1–10.
[23] M.Mysara, J.M. Garibaldi, and M. Elhefnawi, "MysiRNA-Designer a workflow for efficient siRNA design", in PLoS One, vol. 6(10), 2011, pp 1-10.
[24] M.Mysara, M. Elhefnawi, and J. M. Garibaldi., "MysiRNA: Improving siRNA efficacy prediction using a machine-learning model combining multi-tools and whole stacking energy (DG)", in Journal of Biomedical Informatics, vol. 45, 2012, pp. 528–534
[25] A. Khvorova, A. Reynolds, and S.D. Jayasena, " Functional siRNAs and miRNAs exhibit strand bias", in Cell, vol. 115, 2003, pp. 209–216..
[26] J. Harborth, S.M. Elbashir, K. Vandenburgh, H. Manninga, S.A. Scaringe, K. Weber, and T. Tuschl, "Sequence, chemical, and structural variation of small interfering RNAs and short hairpin RNAs and the effect on mammalian gene silencing", in Antisense Nucleic Acid Drug Dev, vol. 13, 2003, pp. 83–105.
[27] T.A. Vickers, S. Koo, C.F. Bennett, S.T. Crooke, N.M. Dean, and B.F. Baker, " Efficient reduction of target RNAs by small interfering RNA and RNase H-dependent antisense agents: A comparative analysis", in Journal of Biology and Chemistry, vol. 278, 2003, pp. 7108–7118,
[28] M. Reena, S. P. David, "Computational Model for Predicting Effective siRNA Sequences for Gene Silencing", Eight International Conference on Data Mining and Warehousing (ICDMW-2014),2014,pp138-142.