Biochemical and Biophysical Research Communications 404 (2011) 313–320

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Development of a software tool and criteria evaluation for efficient design of small interfering RNA

Aparna Chaudhary, Sonam Srivastava, Sanjeev Garg *

Department of Chemical Engineering, IIT Kanpur, UP 208 016, India

ARTICLE INFO

Article history:
Received 27 October 2010
Available online 8 December 2010

Keywords:
RNAi
siRNA
Software tool
Reynolds's rules

ABSTRACT

RNA interference can be used as a tool for gene silencing mediated by small interfering RNAs (siRNA). The critical step in effective and specific RNAi processing is the selection of suitable constructs. Major design criteria, i.e., Reynolds's design rules, thermodynamic stability, internal repeats, immunostimulatory motifs were emphasized and implemented in the siRNA design tool. The tool provides thermodynamic stability score, GC content and a total score based on other design criteria in the output. The viability of the tool was established with different datasets. In general, the siRNA constructs produced by the tool had better thermodynamic score and positional properties. Comparable thermodynamic scores and better total scores were observed with the existing tools. Moreover, the results generated had comparable off-target silencing effect. Criteria evaluations with additional criteria were achieved in WEKA.

1. Introduction

RNA interference (RNAi) a conserved mechanism in most eukaryotic cells [1], involves small double stranded RNA (dsRNA) molecules to direct sequence-dependent gene silencing [2,3]. Short dsRNA molecules are noncoding RNAs involved in regulatory cascades by controlling the gene activity [4,5]. In plants and lower organisms RNAi protects the genome from viruses and insertion of rogue genetic elements [6–8].

In RNAi mechanism, generation of short dsRNA in the form of microRNA (miRNA) or small interfering RNA (siRNA) triggers the process. siRNAs are intracellularly generated from long dsRNA cleaved by Dicer [9]. Dicer cuts the long dsRNA into short 21–23 nucleotide duplexes, each with two complementary strands [2]. These have two nucleotide overhang at the 3′-end of each strand, known as siRNA [10,11]. These strands asymmetrically incorporate into RNA-induced silencing complex (RISC) [12,13]. Antisense strand is thermodynamically favored and selectively binds with RISC for efficient silencing while the sense strand is cleaved by RISC assembly. Thereafter, the whole complex binds to target mRNA with the help of antisense strand. The Ago2 protein, catalytic component of RISC assembly, cleaves the target at one specific site within 10–11 bases [14]. Thereafter, these are degraded by nucleases thus no further protein expression takes place and complete gene silencing occurs [13].

RNA silencing is a useful technique with many applications [15]. Selection of suitable constructs is the critical step in effective and specific RNAi processing. This is achieved by a reliable computational design tools. Various rules based on experimental observations [16–19] as well as mathematical modeling are reported in the open literature [20].

Use of siRNA as a therapeutic tool in mammalian system is limited due to interferon response (IR), saturating endogenous RNAi pathways and sequence-dependent off-target silencing [21]. These need to be addressed while designing siRNA for mammalian system. siRNA design tools are available based on previously discussed criteria while addressing these difficulties. Criteria addressing these difficulties are either optional or inbuilt [22–27]. However, there is still a need to improve the design criteria for efficient silencing. In this work, empirical and rational characteristics from literature were selected and implemented and evaluated in regression framework. The immunostimulatory effects and cytotoxicity included in the present tool are previously reported in siDRM [28] design tool.

Criteria evaluation was implemented in regression framework using experimental siRNA silencing efficiency values reported [20,29]. Regression models with Waikato Environment for Knowledge Analysis (WEKA) 3.6.2 [30] were developed. Weight values were obtained by minimizing the errors between model predicted and experimental efficiency of the constructs. Regression models with *k*-mer [31] criteria were also developed to identify the role of *k*-mers in siRNA design.

2. Materials and methods

2.1. Databases

Curated and validated siRNA datasets are available online. One of such datasets is MIT/ICBP siRNA database at *National Cancer*

* Corresponding author. Fax: +91 512 259 0104.
E-mail address: sgarg@iitk.ac.in (S. Garg).

0006-291X/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2010.11.114

Institute (http://web.mit.edu/sirna/). This database contains more than hundred (112) human and mice genes. siRNA associated 66 genes were identified with 94 reported siRNAs. The parent gene sequences were collected from NCBI database (http://www.ncbi.nlm.nih.gov/) and used for benchmarking. The reported siRNA constructs matched with 62 parent gene sequences (for 89 siRNAs) at NCBI BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Katoh and Suzuki's [29] dataset was used for criteria evaluation. SNP (single-nucleotide polymorphisms) positions collected from NCBI dbSNP (http://www.ncbi.nlm.nih.gov/snp/) were specified in the tool. Their presence was penalized.

2.2. Algorithm

Primary sequence based design criteria reported for highly functional siRNAs were identified from literature. These criteria and their scores are listed in Table 1 and a schematic of the algorithm developed is depicted in Fig. 1. These criteria were used in regression framework to obtain the important criteria and corresponding weights in place of empirically assigned scores.

23 nucleotide long sequences were generated by one of the two ways – motif selection NA(N21), based on Tuschl rules [32], or complete enumeration (N23). Thermodynamic stability score, GC content and total fitness score were then evaluated based on all other criteria and their scores. AA(N19)TT was preferred as per Tuschl rules. The silencing efficiency is higher with Reynolds's rational design rules as compared to other rules [16]. Therefore, Reynolds's rational design rules were implemented with the previously reported empirical scores and with minor modification in the folding algorithm. The GC stretch rule by Ui-Tei [17] was used besides the Reynold's design rules. Holen et al. [33] reported that siRNA activity is highly position dependent.

Immunostimulation against the siRNA occurs if certain motifs are present. Thus, two immunostimulatory motifs (GUCCUUCAA and UGUGU) [34,35] with negative scores were included along with a cytotoxic motif (UGGC) [36] in the criteria set. Immunostimulation criteria score can be reversed for desired cases [37].

Poly-G are undesirable for synthesis, purification [38] and activity [39]. Poly-C in one of the strands will give rise to poly-G in the complementary strand. Polynucleotides also create low complex regions which may cause internal fold back structures or palindromic repeats. These structures may hinder in the binding of siRNAs. Thus, poly-A/T criterion was also included.

Table 1
Design criteria for the algorithm (all the positions from 5′-end of sense strand).

S. No.	Criteria	Empirical score	
		Presence	Absence
1	Thermo (thermodynamic) score < 0	+1	−1.5
2	G/C score (30–52%)	+1	−1
3	Motif: AA (N19) TT	+1	−1
4	A at 19	+1	−1
5	A at 3	+1	−1
6	T at 10	+1	−1
7	G/C at 19 (it should not be present)	−1	+1
8	G at 13 (it should not be present)	−1	+1
9	Folding (internal repeats/palindrome)	−1	+1
10	3–5 A/T at 15–19	+1	−1
11	Cytotoxic motif (UGGC)	−1	0
12	Immunostimulatory 1 (GUCCUUCAA)	−1	0
13	Immunostimulatory 2 (UGUGU)	−1	0
14	Poly-A/T (more than four)	−1	0
15	Poly-G/C (more than three)	−1	0
16	GC stretch > 9	−1	0
17	Contiguous AT/TA	−0.1	0
18	SNPs	−1	0

Contiguous AT/TA stimulates hydrolysis by ribonuclease-A [40] which is detrimental. Chemical modifications, if used, lead to inefficient siRNAs due to higher stability of siRNAs which affect binding with RISC complex. Small penalty was given to contiguous AT/TA as recommended by Patzel [41].

SNPs should be avoided in the designed siRNAs. SNP positions were specified and their presence was penalized. SNPs can be selected for target specificity [42] and need not be penalized in some other applications.

The presence of secondary structure in a siRNA strand was considered detrimental for silencing efficiency [43,44] and was penalized. The base pairing of RNA secondary structure was previously considered a biological palindrome [44]. The folding algorithm was adapted from general rules implemented in other secondary structure predicting algorithm [45]. The scoring system was modified to +HB_bp i, j (the number of hydrogen bonds between i and j base pairs) for a base pair and 0 for anything else. This scoring was iteratively implemented on the siRNA strand to get the maximum folding score. If bases at given positions i, j in a strand pair then the score for pairing between i and j was summation of the score for $(i-1)$,$(j-1)$ and HB_bp i, j (case 1). If i, j do not pair the pairing for $(i+1)$ and j (case 2) and i and $(j-1)$ (case 3) were checked. The pairing of i with k $(i<k<j)$ and $(k+1)$ with j was also checked for multiple pairing, and the score of sequences i, k and $(k+1)$, j could be the maximum score for a secondary structure between i and j (case 4). Thus the scoring was implemented as:

$$S(i,j)=\max\left\{\begin{array}{ll} S(i-1,j-1)+\text{HB\_bp } i,j & \text{for } i,j \text{ bp} \\ S(i+1,j) & \text{for } i+1,j \text{ bp} \\ S(i,j-1) & \text{for } i,j-1 \text{ bp} \\ \max\limits_{i<k<j}[S(i,k)+S(k+1,j)] & \text{for } i,k \text{ and } k+1,j \text{ bp} \end{array}\right\}$$

A 19×19 array was defined to monitor the scores neglecting the end overhangs in the folding algorithm. Only Watson and Crick base pairs were considered. For each G–C base pair and A–T base pair, HB_bp values of +3 and +2 were added, respectively. The scores S(i, i) and S(i, $i+1$) were initialized to zero for the impossibility of folding between these pairs. The calculation of scores was then done iteratively for the upper triangular matrix starting from the bottom right corner and moving right to left in each row from bottom. The algorithm resulted in a maximum value at the top right corner based on the primary sequence and the possible hydrogen bonds. A maximum HB_bp value of 27 corresponding to nine G–C base pairs was used to calculate the probability for secondary/hairpin-like structure formation. Fitness of the siRNA sequence was penalized if its folding probability was more than 0.5.

Thermodynamic stability score refers to the asymmetrical binding of antisense strand to RISC which is a crucial step in RNA silencing. Negative value of the difference in Gibbs free energy between antisense and sense strands [46,47] is recommended for entry of RISC from 5′-end of antisense rather than its 3′-end [48]. This difference was used as the thermodynamic design criterion and a high score was used (as shown in Table 1). The difference was calculated based on the nearest neighbourhood model [49] for five nucleotides [22] at the ends of strands and one adjacent nucleotide on the dangling end assumed to be nucleotide "T" based on cost factors.

2.3. Off-target silencing

Designed siRNA may bind to off-targets and prevent the expression of other essential genes [21]. Off-target searches were

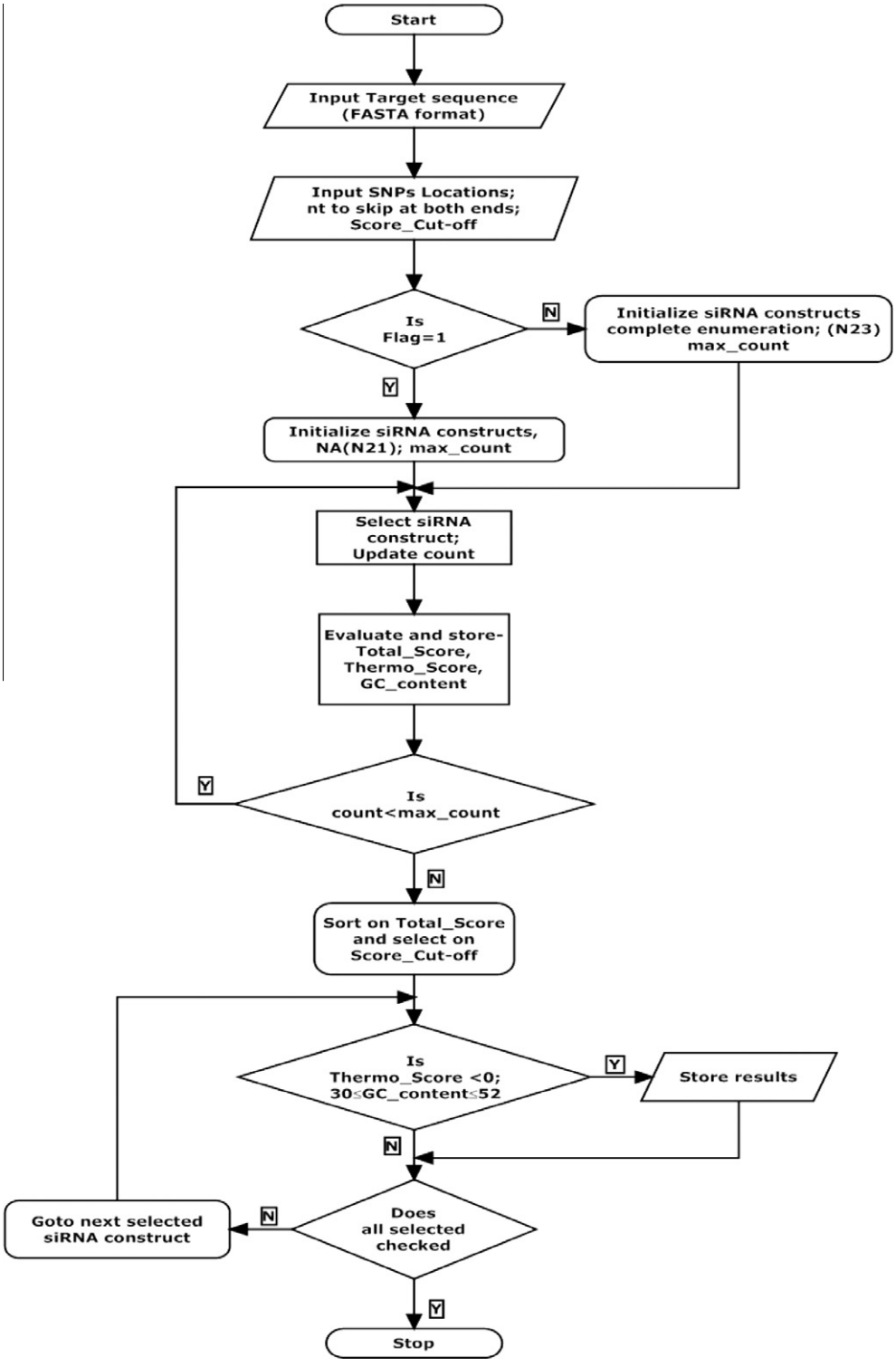

Fig. 1. The workflow for siRNA prediction with a new design algorithm (nt is any nucleotide, N – No and Y – Yes).

performed at NCBI BLAST (–S 2 –W 7 –e 1000) with RefSeq mRNA sequence set.

A different dataset [29] was used for criteria evaluation (silencing efficiencies for MIT/ICBP gene are reported in a range). Kataoh's dataset [29] consisted of 19 nt long sequences. "TT" overhangs were added to calculate the thermodynamic criterion. WEKA was employed for siRNA silencing efficiency prediction. Regression models with *k*-mers were generated.

3. Results

Gene sequences identified from the MIT/ICBP database were retrieved from NCBI database as discussed in Section 2. The mean value and the standard deviation were calculated for thermodynamic score, GC content and total score. These were compared for the two motifs and few values are shown in Table 2 (complete results: worksheet ST-1, Supplementary information).

Table 2
Comparison of NA(N21)* and (N23) motifs.

Gene (Accession no.)	Statistics	Complete enumeration (N23)			Tuschl motif (NA-N19)		
		Thermo-dynamic score	GC content	Total score	Thermodynamic score	GC content	Total score
ABCB1	Average	−2.14	39.17	4.07	−2.25	37.33	4.11
NM_000927	Std dev	1.54	6.85	0.71	1.61	6.53	0.8
CDH1	Average	−2.12	41.38	4.01	−1.95	43.43	4.06
NM_004360	Std dev	1.38	7.05	0.62	1.38	6.54	0.57
FH	Average	−2.36	37.37	4.02	−2.24	37.29	4.1
NM_000143	Std dev	1.65	6.28	0.74	1.59	5.06	0.81
GSK3B	Average	−2.19	38.99	3.96	−2.11	37.29	3.81
NM_002093	Std dev	1.58	6.78	0.61	1.85	5.87	0.41
IGF2R	Average	−2.22	42.55	4.03	−2.08	42.53	4.13
NM_000876	Std dev	1.56	6.55	0.65	1.4	7.11	0.73
AKT3	Average	−2.26	38.53	3.85	−2.18	38.94	3.85
NM_005465	Std dev	1.65	6.4	0.53	1.5	6.76	0.52
RHOA	Average	−1.81	40.14	3.91	−1.52	41.56	4.01
NM_001664	Std dev	1.17	7.1	0.55	0.75	7.52	0.65
STAT1	Average	−1.84	37.38	3.95	−1.81	36.65	3.95
NM_007315	Std dev	1.47	5.9	0.64	1.37	4.2	0.66
YWHAZ	Average	−2.24	39.02	3.89	−2.37	40.81	3.99
NM_003406	Std dev	1.61	6.69	0.52	1.71	7.57	0.65

* Tushl motif is for siRNA construct while proposed work used the parent gene sequence (target) and motif as NA(N21).

Table 3
Comparison between tool's and MIT/ICBP results.

Gene (Accession no.)		Position	Target sequence	Thermodynamic score	GC content	Total score
ABCB1 NM_000927	N23	2509–2531	ttggaggattatgaagctaaatt	−6.18	36	5.6
		4441–4463	gtggagagaaatcatagtttaaa	−5.11	31	5.6
		2876–2898	aatgatgctgctcaagttaaagg	−4.05	36	5.8
	NA(N21)	2876–2898	aatgatgctgctcaagttaaagg	−4.05	36	5.8
		4330–4352	aagcaaacacttacagaattatg	−3.6	31	5.7
		1954–1976	tatcatgaaactgcctcataaat	−2.7	36	5.7
	MIT/ICBP	0–22	gccgaacacattggaaggaaatg	−0.91	42	5.9
CDH1 NM_004360	N23	1711–1733	tgccaactggctggagattaatc	−4.75	47	5.8
		454–476	ttccaccaaagtcacgctgaata	−3.9	52	6
		1360–1382	atacaccatattgaatgatgatg	−2.44	31	5.5
	NA(N21)	4268–4290	gatccgtggtttgtactcaaagc	−4.39	42	3.9
		362–384	gatggtgtgattacagtcaaaag	−3.9	36	3.8
		1280–1302	gagaacgaggctaacgtcgtaat	−0.1	52	5.8
	MIT/ICBP	0–22	tcggcctgaagtgactcgtaacg	−5.66	52	1.9
FH NM_000143	N23	405–427	gatgaggtagctgaaggtaaatt	−4.87	42	5.8
		1465–1487	atggatcaaccttaaaggaaact	−2.09	36	5.8
		1364–1386	gatcaacaagctgatgaatgagt	−1.64	36	5.8
	NA(N21)	405–427	gatgaggtagctgaaggtaaatt	−4.87	42	5.8
		463–485	gatcaggaactcagacaaatatg	−6.01	36	3.8
		492–514	aatgaagtcattagcaatagagc	−1.11	31	5.6
	MIT/ICBP	0–22	gagatctacgatgaactttaaga	−4.26	31	3.6
		0–22	cccaacgatcatgttaataaaag	−3.85	26	5.5
GSK3B NM_002093	N23	3264–3286	tgagaggacattgtagttaaata	−6.15	31	5.7
		4511–4533	aggcaggtctgtgatagagaaat	−3.66	47	5.8
		5381–5403	agtgagaccggtttcatttcata	−3.07	42	5.9
	NA(N21)	790–812	aaggagaaatatatcgcttgttt	−3.18	36	5.5
		7099–7121	tagcctggaaatgaaattaaaaa	−7.25	31	3.7
		1823–1845	aatcagagaaatgaacccaaact	−2.16	36	4.9
	MIT/ICBP	0–22	ctgcatttatcgttaacctaaca	−0.46	31	3.5
		0–22	aacactggtcacgtttggaaaga	−1.57	47	2
IGF2R NM_000876	N23	2965–2987	caacagtggatttgtgtttaatc	−5.05	31	5.8
		6207–6229	acggagtctcgtactatataaat	−5.74	36	5.4
		1426–1448	gatgagcgtcataaactttgagt	−4.14	36	5.8
	NA(N21)	2965–2987	caacagtggatttgtgtttaatc	−5.05	31	5.8
		1426–1448	gatgagcgtcataaactttgagt	−4.14	36	5.8
		1145–1167	gagcagcaggatgtctccataga	−3.73	52	5.7
	MIT/ICBP	0–22	accgcaggtaacgatgggaaagg	−3	52	0.8
		0–22	gggagtctcgtactatataaatc	−5.23	31	1.4
AKT3 NM_005465	N23	606–628	tatgaagattctgaagaaagaag	−1.68	31	5.9
		330–352	agagagaacatttcatgtagata	−1.36	31	5.7
		2579–2601	acgcagctccaacttatataaaa	−6.84	36	3.5
	NA(N21)	606–628	tatgaagattctgaagaaagaag	−1.68	31	5.9
		2425–2447	aaacaggtgtttgccttatttaa	−5.9	36	3.8
		1449–1471	tatggactgcatggacaatgaga	−3.97	47	3.8

Table 3 (*continued*)

Gene (Accession no.)		Position	Target sequence	Thermodynamic score	GC content	Total score
	MIT/ICBP	0–22	acgcagctccaacttatataaaa	−6.84	36	3.5
		0–22	cggagtgatcatggaaatgtatt	−2.5	36	3.6
RHOA NM_001664	N23	840–862	tgggtgccttgtcttgtgaaacc	−5.21	47	4
		1490–1512	taatactgtcatcctcaaagaaa	−0.96	36	5.7
		1095–1117	acacaccaggcgctaattcaagg	−4.03	47	3.8
	NA(N21)	1490–1512	taatactgtcatcctcaaagaaa	−0.96	36	5.7
		1098–1120	caccaggcgctaattcaaggaat	−2.59	52	3.8
		313–335	gatggagcctgtggaaagacatg	−1.52	52	4
	MIT/ICBP	0–22	ttcggaatgatgagcacacaagg	−2.01	47	1.8
		0–22	caagctagacgtgggaagaaaaa	−3.12	47	0.9
STAT1 NM_007315	N23	259–281	ctcgagagctgtctaggttaacg	−4.21	47	5.8
		2831–2853	atgcatcttactgaaggtaaaat	−4.05	36	5.7
		778–800	tcagagcacagtgatgttagaca	−3.74	42	5.8
	NA(N21)	781–803	gagcacagtgatgttagacaaac	−2.75	42	5.8
		2297–2319	aaagaactttctgctgttacttt	−1.52	36	5.9
		803–825	cagaaagagcttgacagtaaagt	−1.36	36	5.9
	MIT/ICBP	0–22	cacaactatattatcatgcaaat	0.71	26	2.8
		0–22	ctgcttgacgtaggaacggtaaa	0.15	52	−2.7
YWHAZ NM_003406	N23	2142–2164	atgtagtgtgttccatttaaaat	−4.27	31	5.7
		503–525	aaggagattactaccgttacttg	−3.96	42	5.6
		226–248	actgagcaaggagctgaattatc	−6.17	42	3.8
	NA(N21)	503–525	aaggagattactaccgttacttg	−3.96	42	5.6
		777–799	cagcacgctaataatgcaattac	−5.56	36	3.5
		340–362	aagacggaaggtgctgagaaaaa	−4.42	52	4
	MIT/ICBP	0–22	caggtttatgttacttctatttg	−2.42	26	1.5

Table 4
Comparison between present and existing design tools (for ABCB1 (NM_000927) gene).

Tool	Position	Target sequence	Thermo-dynamic Score	GC content	Total Score
WI siRNA Selection Program	2592–2614	aaccagcatttgcaataatattt	−6.91	31	3.4
	3219–3241	aagcacacatctttggaattaca	−6.09	36	5.7
	1182–1204	aagaggtcttggcagcaattaga	−6.14	47	1.8
Ambion's siRNA Target finder	25–47	aaagattagagatcatttctcat	0.75	31	−2.9
	86–108	aagagaggtgcaacggaagccag	0.33	63	−6.5
	97–119	aacggaagccagaacattcctcc	−0.78	52	−1.1
siExplorer	3499–3521	gccgaacacattggaaggaaatg	−0.91	42	5.9
	444–466	gaggagcaaagaagaagaacttt	−4.01	42	2
	3913–3935	tagcactaaagtaggagacaaag	−2.72	42	3.8
Gene Script Corp siRNA Target Finder	3931–3953	aaggaactcagctctctggtgtg	0.44	57	−3.5
	4250–4271	aagcgccagtgaactctgactct	−4.18	57	−1
	312–332	aagaagcagaggccgctgttctc	−1.65	63	0
Design Tool	2509–2531	ttggaggattatgaagctaaatt	−6.18	36	5.6
	4441–4463	gtggagagaaatcatagtttaaa	−5.11	31	5.6
	2876–2898	aatgatgctgctcaagttaaagg	−4.05	36	5.8

The best three constructs were identified on the basis of scaled scores (summation of normalized thermodynamic scores and total scores between −1 and +1) and off-target silencing. Comparisons were made between thermodynamic score, GC content and total score. Table 3 shows a few results generated with the designed tool and the MIT/ICBP predictions (complete results: worksheet ST-2).

Comparisons for ABCB1 gene (NM_000927.3) with the exiting design tools – Whitehead Institute's siRNA Selection Program (http://jura.wi.mit.edu/bioc/siRNAext/), Ambion's siRNA Target finder (http://www.ambion.com/techlib/misc/siRNA_finder.html), siExplorer (http://rna.chem.t.u-tokyo.ac.jp/cgi/siexplorer.htm) and Gene Script Crop siRNA Target Finder (https://www.genscript.com/ssl-bin/app/rnai) were made. The best three results from each are outlined in Table 4.

Predominant criteria were identified by regression models. Three different regression models were generated. These were for a total of 18 criteria (model 1), 18 + 3-mer (model 2) and 18 + *k*-mers (model 3). Regression was done in WEKA 3.6.2. Model 1 had systematic errors and large scatter as compared to models 2 and 3 (Fig. 2). Model 2 was preferred over model 3 due to similar predictions with lesser criteria.

4. Discussions

In Table 2, the statistical analyses show that almost similar results were obtained by both motifs. Common siRNAs were generated for both motifs, e.g., for FH gene, the siRNAs generated with NA(N21) and (N23) were similar (Table 3). With NA(N21), some effective siRNAs could be missed as compared to (N23) although the search effort is much less for NA(N21). Thus, NA(N21) motif was not preferred.

Table 3 shows that the proposed design tool predicts better thermodynamic score as well as total score in comparison with MIT/ICBP dataset. Moreover, 12 siRNA out of 90 siRNAs in MIT/ICBP datasets had positive thermodynamic stability score that

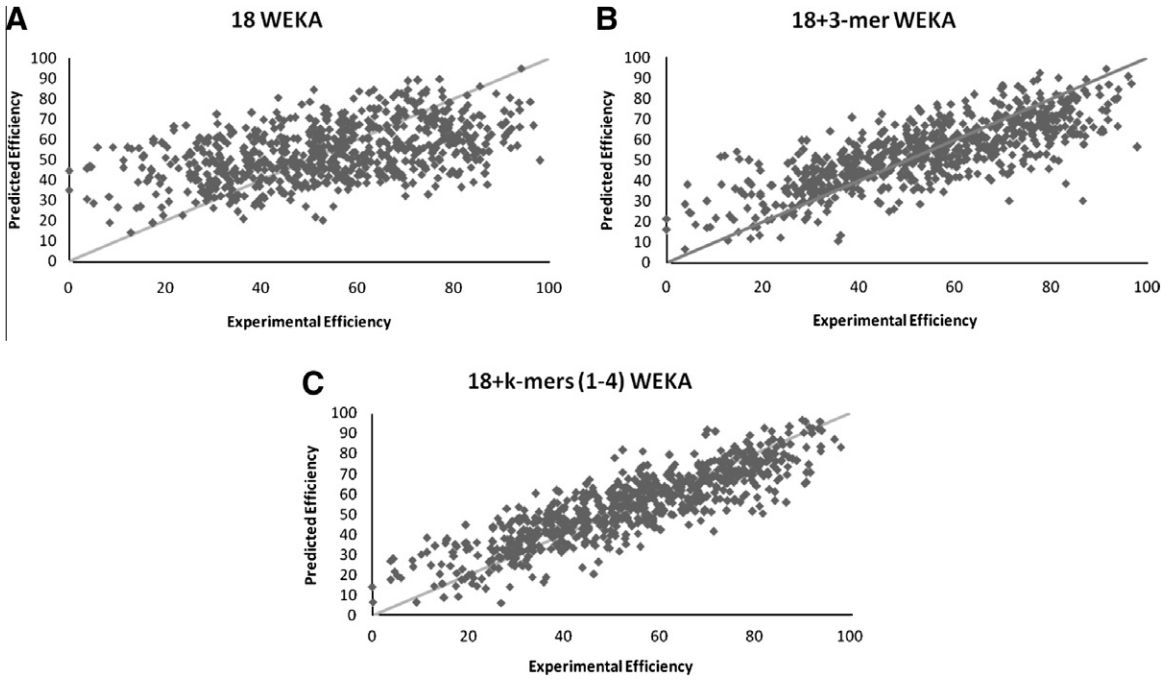

Fig. 2. A comparison between the predicted and experimental efficiencies (a) for 18 criteria (model 1); (b) for 18 + 3-mer criteria (3 nucleotide combination, model 2); and (c) for 18 + *k*-mer criteria (1, 2, 3,4 nucleotide combinations, model 3).

could produce off-target binding, (e.g., STAT1 gene, Table 3). Moreover, 42 siRNAs in the dataset had total score close to zero.

Finally, the results were compared on the basis of percentage failure of design criteria. The percentage failure was calculated only for fifteen out of eighteen criteria. Criteria 1, 16 and SNP values were not included. The criterion 1 was evaluated in a different way (Table 2) while criterion 16 resulted in negative values in most cases. SNP values were not listed for all genes at the NCBI dbSNP site and hence SNP criterion was not considered. The percentage failure was defined as the percentage of criteria failure out of the fifteen criteria (Column difference in worksheet ST-2). Percentage failure for the predicted constructs was in the range 0.0–13.33%. This was much better as compared to the reported results for MIT/ICBP predictions with 0.0–33.33% percentage failure. Few (23 out of 90) predicted siRNAs were similar to the reported siRNAs at the MIT/ICBP website. Though, comparable results for a few genes were obtained with the designed tool, most of these similar siRNAs were below the 10 best predicted sequences based on scaled score. Thus, better siRNAs were predicted as compared to those of MIT/ICBP dataset. This was further substantiated on comparing the best predictions and matching the design criteria (worksheet ST-2). Furthermore, off-target silencing effect was verified at NCBI BLAST search with a cut-off score of 30 [50]. Comparable results for most of the siRNAs were observed. MIT/ICBP dataset resulted in off-target silencing for a few siRNAs with this cut-off score, e.g., ARHGDIB, BAG4, HPRT1, NME1, etc. (worksheet ST-2). However, siRNAs for NME2 gene were showing high similarity with a closely related family gene, i.e., NME1–NME2 co-transcribed product of genes NME1 and NME2, hence, an effective siRNA could not be designed for this gene. MIT/ICBP results were also showing off-target effect for this gene.

Comparable thermodynamic scores and better total scores were observed with the designed tool as compared to the siRNA Selection Program. Better thermodynamic scores as well as total scores were observed with the designed tool as compared to the siRNA target finder, siExplorer and siRNA Target Finder (Table 4).

The common criteria exhibiting good correlation with experimental siRNA silencing efficiency were GC content, criteria 6, 7 and 8 (Table 1) in all the three regression models developed. Besides these defined criteria, some of the other criteria were also identified from models 2 and 3 (worksheet ST-3). Thermodynamic score and the GC content were the predominant criteria in model 1. The observed weights for these criteria were much less, as expected, in models 2 and 3 due to increase in number of criteria.

Table 5
Criteria evaluation and their corresponding weights.

S. no.	Criteria	Weights		
		Model 1	Model 2	Model 3
1	Thermo (thermodynamic) score < 0	−21.16	−8.59	0
2	G/C score (30–52%)	−55.58	−34.64	−46.48
3	Motif: AA (N19) TT	0	0	0
4	A at 19	3.8	0	0
5	A at 3	3.99	0	0
6	T at 10	5.43	6.89	9.03
7	G/C at 19 (it should not be present)	3.48	6.27	3.65
8	G at 13 (it should not be present)	2.79	3.34	3.16
9	Folding (internal repeats/ palindrome)	0	0	0
10	3–5 A/T at 15–19	0	0	10.71
11	Cytotoxic motif (UGGC)	0	0	0
12	Immunostimulatory 1 (GUCCUUCAA)	0	0	0
13	Immunostimulatory 2 (UGUGU)	0	0	0
14	Poly-A/T (more than four)	0	0	0
15	Poly-G/C (more than three)	7.5	0	−16.25
16	GC stretch > 9	0	0	−26.69
17	Contiguous AT/TA	17.94	21.79	0
18	SNPs	0	0	0
	Bias	66.63	55.23	126.77

A total of 11 criteria (1, 2, 4–8, 15 and 17) predominantly appeared in one of the three models while the remaining criteria did not appear in any of the models (Table 5). These 11 criteria are reported to be significant in the literature [16–19]. Khvorova et al. [46], Poliseno et al. [51] and Chalk et al. [52] suggested negative value for the difference in Gibbs free energy between antisense and sense strands. Criterion 1 (thermodynamic score, based on the difference in Gibbs free energy between antisense and sense strands) was thus desirable for effective binding with RISC. Various studies have reported G/C content (criterion 2) as an important criterion which had been also included in almost all the existing tools [16,18,53]. Criterion 4 (A at position 19) [16] and Criterion 7 (No G/C at position 19) [16,18,54,55] were also important criteria. Criterion 5 (A at position 3) was previously reported by Reynolds et al. [16] and was not reported in any further studies. Reynolds et al. [16] and Jagla et al. [19] reported criterion 6 (U at position 10) as a preferred criterion while Amarzguioui and Prydz [18] suggested that U should not be present at position 10. Criterion 10 (3–5 A/T at 15–19 positions) was included in different ways in different studies. These were either specific positions of A/T between 15 and 19 positions [54] or varying length from 13 to 19 [19] or 15 to 19 [16]. Ui-Tei et al. have defined this as one third of 3′-end sense strand [17], which was same as varying length from 13 to 19. Criterion 15 (poly-G/C) was included in various algorithm as an important criterion [22–24]. Criterion 16 (G/C stretch) was implemented in this study similar to Ui-Tei et al. [17], i.e., GC stretch should not be not more than 9. This was different than the more stringent criterion by Yuan et al. [22] (GC stretch should not be more than 7). Criterion 17 (contiguous AT) was observed to be important. Though the criterion is reported in literature, it has not been implemented in any of the existing design tools. Thus, it was implemented as a new criterion in the developed tool.

Some common nucleotides were observed in 3-mer and *k*-mers. 1-mer and 2-mer were not observed to be as significant as 3-mer. Thus, only these common nucleotides (worksheet ST-3) are recommended to be included as additional criteria along with the basic 18 criteria.

5. Conclusions

A siRNA design tool was developed and implemented. The viability of the tool was established by predicting and comparing the siRNA constructs for genes in the MIT/ICBP database. The tool was further used to predict and compare the results with some other existing design tools. The predictions, in general, were observed to have better values for thermodynamic score and other design criteria with almost similar or better off-target silencing. It is noted that the results are *in silico* and *in vitro* verification should be performed to further validate the claims.

6. Conflict of interest statement

None declared.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2010.11.114.

References

[1] A. Fire, RNA-triggered gene silencing, Trends Genet. 15 (1999) 358–363.
[2] P.D. Zamore, T. Tuschl, P.A. Sharp, D.P. Bartel, RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals, Cell 101 (2000) 25–33.
[3] T. Tuschl, M.M. Ng, W. Pieken, F. Benseler, F. Eckstein, Importance of exocyclic base functional groups of central core guanosines for hammerhead ribozyme activity, Biochemistry 32 (1993) 11658–11668.
[4] R. Almeida, R.C. Allshire, RNA silencing and genome regulation, Trends Cell Biol. 15 (2005) 251–258.
[5] M. Lagos-Quintana, R. Rauhut, W. Lendeckel, T. Tuschl, Identification of novel genes coding for small expressed RNAs, Science 294 (2001) 853–858.
[6] D. Baulcombe, RNA silencing in plants, Nature 431 (2004) 356–363.
[7] H. Vaucheret, C. Beclin, M. Fagard, Post-transcriptional gene silencing in plants, J. Cell Sci. 114 (2001) 3083–3091.
[8] T. Sijen, R.H.A. Plasterk, Transposon silencing in the *Caenorhabditis elegans* germ line by natural RNAi, Nature 426 (2003) 310–314.
[9] E. Bernstein, A.A. Caudy, S.M. Hammond, G.J. Hannon, Role for a bidentate ribonuclease in the initiation step of RNA interference, Nature 409 (2001) 363–366.
[10] S.M. Elbashir, W. Lendeckel, T. Tuschl, RNA interference is mediated by 21- and 22-nucleotide RNAs, Genes Dev. 15 (2001) 188–200.
[11] A. Nykanen, B. Haley, P.D. Zamore, ATP requirements and small interfering RNA structure in the RNA interference pathway, Cell 107 (2001) 309–321.
[12] D.S. Schwarz, Asymmetry in the assembly of the RNAi enzyme complex, Cell 115 (2003) 199–208.
[13] G. Hutvagner, Small RNA asymmetry in RNAi: function in RISC assembly and gene regulation, FEBS Lett. 579 (2005) 5850–5857.
[14] F.V. Rivas, Purified Argonaute2 and an siRNA form recombinant human RISC, Nat. Struct. Mol. Biol. 12 (2005) 340–349.
[15] Y. Dorsett, T. Tuschl, siRNAs: applications in functional genomics and potential as therapeutics, Nat. Rev. Drug Discovery 3 (2004) 318–329.
[16] A. Reynolds, D. Leake, Q. Boese, S. Scaringe, W.S. Marshall, A. Khvorova, Rational siRNA design for RNA interference, Nat. Biotechnol. 22 (2004) 326–330.
[17] K. Ui-Tei, Y. Naito, F. Takahashi, T. Haraguchi, H. Ohki-Hamazaki, A. Juni, R. Ueda, K. Saigo, Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference, Nucleic Acids Res. 32 (2004) 936–948.
[18] M. Amarzguioui, H. Prydz, An algorithm for selection of functional siRNA sequences, Biochem. Biophys. Res. Commun. 316 (2004) 1050–1058.
[19] B. Jagla, N. Aulner, P.D. Kelly, D. Song, A. Volchuk, A. Zatorski, D. Shum, T. Mayer, D.A. De Angelis, O. Ouerfelli, U. Rutishauser, J.E. Rothman, Sequence characteristics of functional siRNAs, RNA 11 (2005) 864–872.
[20] D. Huesken, J. Lange, C. Mickanin, J. Weiler, F. Asselbergs, J. Warner, B. Meloon, S. Engel, A. Rosenberg, D. Cohen, M. Labow, M. Reinhardt, F. Natt, J. Hall, Design of a genome-wide siRNA library using an artificial neural network, Nat. Biotechnol. 23 (2005) 995–1001.
[21] A.L. Jackson, P.S. Linsley, Recognizing and avoiding siRNA off-target effects for target identification and therapeutic application, Nat. Rev. Drug Discovery 9 (2010) 57–67.
[22] B. Yuan, R. Latek, M. Hossbach, T. Tuschl, F. Lewitter, siRNA selection server: an automated siRNA oligonucleotide prediction server, Nucleic Acids Res. 32 (2004) W130–W134.
[23] W. Cui, J. Ning, U.P. Naik, M.K. Duncan, OptiRNAi, an RNAi design tool, Comp. Meth. Prog. Biomed 75 (2004) 67–73.
[24] A. Henschel, F. Buchholz, B. Habermann, DEQOR: A web-based tool for the design and quality control of siRNAs, Nucleic Acids Res. 32 (2004) W113–W120.
[25] I. Bradác, R. Svobodová Vareková, M. Wacenovsky, M. Skrdla, M. Plchút, M. Polcík, SiRNA selection criteria – statistical analyses of applicability and significance, Biochem. Biophys. Res. Commun. 359 (2007) 83–87.
[26] H. Zhou, X. Zeng, Y. Wang, B.R. Seyfarth, A three-phase algorithm for computer aided siRNA design, Informatica (Ljubljana) 30 (2006) 357–364.
[27] Y. Naito, J. Yoshimura, S. Morishita, K. Ui-Tei, siDirect 2.0: updated software for designing functional siRNA with reduced seed-dependent off-target effect, BMC Bioinf. 10 (2009).
[28] W. Gong, Y. Ren, H. Zhou, Y. Wang, S. Kang, T. Li, SiDRM: an effective and generally applicable online siRNA design tool, Bioinformatics 24 (2008) 2405–2406.
[29] T. Katoh, T. Suzuki, Specific residues at every third position of siRNA shape its efficient RNAi activity, Nucleic Acids Res. 35 (2007) e27.
[30] E.F. Mark Hall, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten, The WEKA data mining software: an update, SIGKDD Explor. 11 (2009).
[31] R. Teramoto, M. Aoki, T. Kimura, M. Kanaoka, Prediction of siRNA functionality using generalized string kernel and support vector machine, FEBS Lett. 579 (2005) 2878–2882.
[32] T. Tuschl, P.D. Zamore, R. Lehmann, D.P. Bartel, P.A. Sharp, Targeted mRNA degradation by double-stranded RNA *in vitro*, Genes Dev. 13 (1999) 3191–3197.
[33] T. Holen, M. Amarzguioui, M.T. Wiiger, E. Babaie, H. Prydz, Positional effects of short interfering RNAs targeting the human coagulation trigger tissue factor, Nucleic Acids Res. 30 (2002) 1757–1766.
[34] V. Hornung, M. Guenthner-Biller, C. Bourquin, A. Ablasser, M. Schlee, S. Uematsu, A. Noronha, M. Manoharan, S. Akira, A. de Fougerolles, S. Endres, G. Hartmann, Sequence-specific potent induction of IFN-alpha by short interfering RNA in plasmacytoid dendritic cells through TLR7, Nat Med. 11 (2005) 263–270.

[35] A.D. Judge, V. Sood, J.R. Shaw, D. Fang, K. McClintock, I. MacLachlan, Sequence-dependent stimulation of the mammalian innate immune response by synthetic siRNA, Nat. Biotechnol. 23 (2005) 457–462.
[36] Y. Fedorov, E.M. Anderson, A. Birmingham, A. Reynolds, J. Karpilow, K. Robinson, D. Leake, W.S. Marshall, A. Khvorova, Off-target effects by siRNA can induce toxic phenotype, RNA 12 (2006) 1188–1196.
[37] M. Schlee, V. Hornung, G. Hartmann, SiRNA and isRNA: two edges of one sword, Mol. Ther. 14 (2006) 463–470.
[38] B. Pan, K. Shi, M. Sundaralingam, Synthesis, purification and crystallization of guanine-rich RNA oligonucleotides, Biol. Proced. Online 6 (2004) 257–262.
[39] R.H. Shafer, I. Smirnov, Biological aspects of DNA/RNA quadruplexes, Biopolymers 56 (2001) 209–227.
[40] R. Kierzek, Hydrolysis of oligoribonucleotides: influence of sequence and length, Nucleic Acids Res. 20 (1992) 5073–5077.
[41] V. Patzel, In silico selection of active siRNA, Drug Discovery Today 12 (2007) 139–148.
[42] D.S. Schwarz, H. Ding, L. Kennington, J.T. Moore, J. Schelter, J. Burchard, P.S. Linsley, N. Aronin, Z. Xu, P.D. Zamore, Designing siRNA that distinguish between genes that differ by a single nucleotide, PLoS Genet. 2 (2006) 1307–1318.
[43] T.A. Vickers, J.R. Wyatt, S.M. Freier, Effects of RNA secondary structure on cellular antisense activity, Nucleic Acids Res. 28 (2000) 1340–1347.
[44] V. Patzel, S. Rutz, I. Dietrich, C. Koberle, A. Scheffold, S.H.E. Kaufmann, Design of siRNAs producing unstructured guide-RNAs results in improved RNA interference efficiency, Nat. Biotech. 23 (2005) 1440–1444.
[45] S.R. Eddy, How do RNA folding algorithms work?, Nat Biotechnol. 22 (2004) 1457–1458.
[46] A. Khvorova, A. Reynolds, S.D. Jayasena, Functional siRNAs and miRNAs exhibit strand bias, Cell 115 (2003) 209–216.
[47] D.S. Schwarz, G. Hutvágner, T. Du, Z. Xu, N. Aronin, P.D. Zamore, Asymmetry in the assembly of the RNAi enzyme complex, Cell 115 (2003) 199–208.
[48] S.M. Elbashir, J. Martinez, A. Patkaniowska, W. Lendeckel, T. Tuschl, Functional anatomy of siRNAs for mediating efficient RNAi in *Drosophila melanogaster* embryo lysate, EMBO J. 20 (2001) 6877–6888.
[49] T. Xia, J. SantaLucia Jr., M.E. Burkard, R. Kierzek, S.J. Schroeder, X. Jiao, C. Cox, D.H. Turner, Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick base pairs, Biochemistry 37 (1998) 14719–14735.
[50] X. Wang, B. Seed, Selection of oligonucleotide probes for protein coding sequences, Bioinformatics 19 (2003) 796–802.
[51] L. Poliseno, M. Evangelista, A. Mercatanti, L. Mariani, L. Citti, G. Rainaldi, The energy profiling of short interfering RNAs is highly predictive of their activity, Oligonucleotides 14 (2004) 227–232.
[52] A.M. Chalk, C. Wahlestedt, E.L.L. Sonnhammer, Improved and automated prediction of effective siRNA, Biochem. Biophys. Res. Commun. 319 (2004) 264–274.
[53] S.M. Elbashir, J. Harborth, K. Weber, T. Tuschl, Analysis of gene function in somatic mammalian cells using small interfering RNAs, Methods 26 (2002) 199–213.
[54] S. Takasaki, S. Kotani, A. Konagaya, An effective method for selecting siRNA target sequences in mammalian cells, Cell Cycle 3 (2004) 790–795.
[55] A.C. Hsieh, R. Bo, J. Manola, F. Vazquez, O. Bare, A. Khvorova, S. Scaringe, W.R. Sellers, A library of siRNA duplexes targeting the phosphoinositide 3-kinase pathway: determinants of gene silencing for use in cell-based screens, Nucleic Acids Res. 32 (2004) 893–901.

Web references

[1] <http://web.mit.edu/sirna/> (18.09.10).
[2] <http://www.ncbi.nlm.nih.gov/> (20.10.10).
[3] <http://blast.ncbi.nlm.nih.gov/Blast.cgi> (18.09.10).
[4] <http://www.ncbi.nlm.nih.gov/snp/> (18.09.10).
[5] <http://jura.wi.mit.edu/bioc/siRNAext/> (18.09.10).
[6] <http://www.ambion.com/techlib/misc/siRNA_finder.html> (18.09.10).
[7] <http://rna.chem.t.u-tokyo.ac.jp/cgi/siexplorer.htm> (18.09.10).
[8] <https://www.genscript.com/ssl-bin/app/rnai> (18.09.10).